

# DATA LITERACY CURRICULUM

---

This curriculum, developed by the Open Knowledge Foundation with the support of the International Republican Institute, aims to present the fundamentals of data literacy in a way that is approachable for civic actors in general, and democracy activists in particular. The curriculum's structure is based on the Data Pipeline, a methodology developed by School of Data, an Open Knowledge Foundation program.



# TABLE OF CONTENTS

---

<b>Module 1 – Thinking and working with data</b>	<b>4</b>
Introduction	4
Description	4
Skills	4
Prerequisites	4
Section 1: What is data?	5
Section 2: Working with tabular data	7
Section 3: Tidy data	11
Section 4: Summary table, data visualization	12
<b>Module 2 – Finding data</b>	<b>16</b>
Introduction	16
Description	16
Skills	16
Prerequisites	16
Section 1: Understanding data formats	17
Section 2: Where can I find data?	20
Section 3: Searching for data	24
External resources	25
<b>Module 3 – Getting data</b>	<b>26</b>
Introduction	26
Description	26
Skills	26
Prerequisites	26
Section 1: Extracting data from digital PDFs	27
Section 2: An introduction to web scraping	32
Section 3: Collecting data	33
<b>Module 4 – Verifying data</b>	<b>47</b>
Introduction	47
Description	47
Skills	47
Prerequisites	47
Section 1: What is good data?	48
Section 2: Four verification methods	49
Section 3: Common data issues	55

<b>Module 5 – Cleaning data</b>	<b>56</b>
Introduction	56
Description	56
Skills	56
Prerequisites	56
Section 1: Data tidying	57
Section 2: Data editing	64
Section 3: Data consolidation (or “merging”)	74
<b>Module 6 – Analyzing data</b>	<b>78</b>
Introduction	78
Description	78
Skills	78
Prerequisites	79
Section 1: From research question to hypothesis	79
Section 2: Creating your analysis plan	80
Section 3: Common techniques for data analysis	81
<b>Module 7 – Presenting data</b>	<b>90</b>
Introduction	90
Description	91
Skills	91
Prerequisites	91
Section 1: Choosing your emphasis	91
Section 2: Choosing your visualization	94
Section 3: The tools for data presentation	99
<b>Module 8 – Using geographical data</b>	<b>100</b>
Introduction	100
Description	101
Skills	101
Prerequisites	101
Section 1: The basics of GIS	102
Section 2: Creating an online map	103
<b>Module 9 – Going further with spreadsheets</b>	<b>107</b>
Introduction	107
Description	107
Skills	107
Prerequisites	107
Section 1: Creating a dashboard in Google Sheets	108
Section 2: Using custom scripts in Google Sheets	113

<b>Module 10 – Using data to evaluate your projects</b>	<b>115</b>
Introduction	115
Description	115
Skills	116
Prerequisites	116
Section 1: Planning	116
Section 2: Project monitoring	118
Section 3: Project evaluation	120



# MODULE 1 – THINKING AND WORKING WITH DATA

---

## Introduction

What is data? Many use the word but asking this question to ten people will probably give you ten answers. And that's because we all have an instinctive sense of what data is, although not having a clear definition inhibits us from properly making use of it. But in a context where most civic organizations are still far from being data savvy, the confusion about data leads to many missed opportunities for social change.

This module will cover the basic concepts needed to navigate the world of data and start your own data-driven projects.

## Description

The module will cover the following concepts:

- Data
- Data literacy
- Data ethics
- Data security
- Data privacy
- Data Pipeline
- Spreadsheets

## Skills

As part of this module, you will learn how to:

- Read a tabular dataset
- Navigate the Data Pipeline

## Prerequisites

- Module dataset: <http://bit.ly/2019TunisBudget>
- A Google account

There are many different forms of spreadsheets available online, free or paid. In order to improve the accessibility of the curriculum we have chosen to use Google Sheets as the software to be used throughout this curriculum. Google Sheets is accessible through any modern browser and doesn't require installation. Nonetheless, everything learned in the curriculum will be transferable to other spreadsheet software, with minor differences in functionalities.

- Spreadsheet software (Google Sheets)
- A working computer
- A modern internet browser
- Internet connection

## Section 1: What is Data?

Data is all around us. But what exactly is it? Data is a value assigned to a thing. Take, for example, the balls in the picture below.



*Photo by Cristina Anne Costello*

What can we say about these? They are golf balls, right? So, one of the first data points we have is that they are used for golf. Golf is a category of sport, so this helps us to put the balls in a taxonomy. But there is more to them. We have the color: "white," the condition "used." They all have a size, there is a certain number of them, and they probably have some monetary value, and so on.

Even unremarkable objects have a lot of data attached to them. You, too: you have a name (most people have given and family names), a date of birth, weight, height, nationality, etc. All these things are data.

In the example above, we can already see that there are different types of data.

Here are the categories that you will most likely encounter:

**Categorical data** puts the item you are describing into a category. In our example, the condition "used" would be categorical (with categories such as "new," "used," "broken," etc.). Categorical data may be unordered, as above, or ordered or ranked, as in "great," "good," "fair," and "poor" with reference to the quality of the ball.

**Discrete data** is numerical data that has gaps in it, e.g., the count of golf balls. There can only be whole numbers of golf balls (there is no such thing as 0.3 golf balls). Other examples are scores in tests (where you receive 7/10) or shoe sizes.

**Continuous data** is numerical data with a continuous range. For example, the size of the golf balls can be any value (e.g., 10.53 mm or 10.54 mm but also 10.536 mm), or the size of your foot (as opposed to your shoe size, which is discrete). In continuous data, all values are possible with no gaps in between.

## Qualitative vs Quantitative Data

The word “data” is used in multiple ways, both in casual conversation and across scientific disciplines. In casual conversation, data is often used interchangeably with “information.” In technical or scientific conversations, data also means “information collected in a structured way” – the emphasis is on the fact that there is a methodology behind the data collection. This explains the use of the term “qualitative data,” which can be anything from an interview transcript to a description of a golf ball.

**Qualitative data** is everything that refers to the quality of something: A description of colors, texture, and feel of an object, a description of experiences, and interviews are all qualitative data. Qualitative data can be unstructured (e.g., a poem) or structured (e.g., a dataset organizing famous French poems by genre, year, author, etc.).

**Quantitative data** is data that refers to a number, like the number of golf balls, their size, price, or a score on a test, etc.

So, what is data? In this curriculum, we will use the following definition:

*Data is a structured representation of the world. Data is:*

**A representation** of the world because data always tries to capture a part of reality. An important aspect of the process of verifying data is to check the choices made by the creator of the data about which part of reality they decided to capture. Verifying data also means verifying if the produced data correctly reflects reality, given those choices (that we will call methodology).

**Structured**, because to call it data, the result of that act of representing the world needs to follow a logical ordering. The structure differentiates a simple piece of information, such as a textual description of the golf balls, from a dataset, which presents various aspects of the golf balls in an ordered and logical way.

Non structured description	Structured description
There were 15 golf balls grouped together, and the white of their envelope shone brightly under the sun.	Type, sport, color, number of balls, white

Lastly, to differentiate the various elements of the data we will be working with, we will discuss data points which describe the individual elements collected about the reality, such as “white” being one data point about a golf ball and also dataset, to talk about a collection of data points.

## Section 2: Working with tabular data

Let's open the module dataset: <http://bit.ly/2019TunisBudget>

The software used to open the dataset is called spreadsheet software. It is designed to read and transform data organized with lines and columns it is called tabular data (as in, a clean slate). Nowadays spreadsheets are in common use, so many people are familiar with them already. A variety of spreadsheet programs and applications exist and depending on what you want to do you might consider using different spreadsheet software. Here are some of the considerations you might make when picking your weapon of choice:

Spreadsheet	Google Sheets	Libre Office	Microsoft Excel
Cost	Free	Free	Paid
Data storage	Google Drive	Your computer	Your computer
Needs internet	Yes	No	No
Installation required	No	Yes	Yes
Collaboration	Yes	No	No
Sharing results	Easy	Harder	Harder
Visualizations	Yes	Yes	Yes

Since it does not require you to purchase or install any additional software, we will be using Google Sheets for this course.

A spreadsheet is basically a table of cells in which you can input data. The cells are organized in rows and columns. Typically, rows are labelled by numbers, columns by letters. This also means cells can be addressed by their column and row coordinates. The cell A1 denotes the cell in the first row in the first column, A2 the one in the second row, B1 the one in the second column, and so on.

Let's try to identify the data that we have in front of us:

On row 1 we have the data headers, 'item' and 'amount'. They describe what type of information is contained in each column. Because we're talking about a budget, Column A, with the header 'item' includes the name of the budget categories. Column B, with the header 'amount' includes the amount linked to the budget category on the same line.

Beside the header row, notice that different rows contain different types of information:

- Row 2 is a budget section title (and not a budget item).
- Row 3 is also a budget subsection title, and the value in amount (B3) is the sum of the values of rows 4, 5 and 6. This is consequently a subtotal and not an individual item.
- Row 8 is the same as row 3, a subtotal of the budget, representing the sum of rows 9 and 10.

- Row 11 is the total of the revenue category of the budget and is equivalent to the sum of rows 3 and 8.
- Rows 4, 5, 6, 7, 9 and 10 are the only rows that are individual budget items.

If it is confusing, it's normal – this budget is set up to facilitate human reading at the cost of using a complex structure. Data tables are often formatted like this in official reports and other documents meant to be read by humans rather than analyzed by computers. The goal is to surface key information (the subtotals) rather than having a clean structure (which computers prefer). We'll see how to improve that structure in a minute, but first let's practice our spreadsheet skills a bit more.

## Functions and formulas

In order to verify that the section 1 and section 2 subtotals match the TOTAL CATEGORY 1 REVENUE, we want to use a sum. This is where the reference system of the spreadsheet (numbers for rows and letters for columns) become useful. To add the value of Section 1 and Section 2, you simply need to double click on an empty cell and enter

```
=B3+B8
```

We have three elements here:

- the = (equal) sign, which tells the spreadsheet that you want to do a calculation and not just write some text
- the + (plus) sign, which is used for sums (- for subtraction, \* for multiplication, / for division)
- B3 and B8 are the references to the values contained in the cells in column B and rows 3 and 8 respectively.

This referencing system is the core feature of spreadsheet software and allows you to run complex calculations without having to copy-paste the initial data over and over. Similarly, if you change the source data, the result of the calculation will change accordingly.

Pressing 'enter' on your keyboard allows you to complete the calculation. Comparing the results of our calculation and the value of TOTAL CATEGORY 1 REVENUE, we see that the number is the same, which confirms that TOTAL CATEGORY 1 REVENUE is indeed the sum of 'Section 1: Ordinary Tax Revenue and Section 2: Ordinary non-tax revenue.

As long as you start with the = sign, you can write mathematical **formulas** in a spreadsheet, just like you would do on a sheet of paper: with brackets, subtractions, exponents, etc. But what happens when you have a lot of values that you want to add together? Even with the referencing system, adding dozens of values by writing value1+value2+value3, etc. can be time-consuming and error prone. Which is why spreadsheet software includes a feature called **functions**. These are shortcuts allowing you to write functions faster. For example, instead of writing

```
=B3+B8
```

You can write

```
=SUM(B3,B8)
```

While a **formula** is an equation designed by the user of the spreadsheet, a **function** is a predefined calculation baked into the spreadsheet software. Functions exist for all the common mathematical operations, such as **DIVIDE()**, **MULTIPLY()**, **ROUND()**, **COUNT()**, and so on. The values that you put in the brackets of the function are called parameters. Each function has its own set of rules. For example, the **DIVIDE()** function will only take 2 parameters while the **SUM()** function can include as many as you want.

<p><b>SUM(value1, [value2, ...])</b> ^ X</p> <p><b>EXAMPLE</b> <b>SUM(A2:A100, 101)</b></p> <p><b>ABOUT</b> Returns the sum of a series of numbers and/or cells.</p> <hr/> <p><b>value1</b> The first number or range to add together.</p> <p><b>value2... - [optional] repeatable</b> Additional numbers or ranges to add to 'value1'.</p>	<p><b>DIVIDE(dividend, divisor)</b> ^ X</p> <p><b>EXAMPLE</b> <b>DIVIDE(4, 2)</b></p> <p><b>ABOUT</b> Returns one number divided by another. Equivalent to the '/' operator.</p> <hr/> <p><b>dividend</b> The number to be divided.</p> <p><b>divisor</b> The number to divide by.</p>
<i>The <b>SUM()</b> function can take many values as its parameters</i>	<i>The <b>DIVIDE()</b> function can only take two values as its parameters</i>

Some functions take text instead of numerical values as parameters and allow you to do text-based operations. For example, the **SUBSTITUTE()** function can be used to replace a letter in any cell that contains text. We will use several functions throughout this curriculum and will explain their uses whenever we introduce a new one. But if you want to learn more about functions, a simple search on the web will provide detailed lists of all functions used by various spreadsheet software.

## Shortcuts, navigation and filtering

As you practice writing functions and formulas in a spreadsheet, you'll notice that it's inconvenient to always use your mouse to select different cells or move your cursor. Most seasoned spreadsheet users rely on the keyboard to navigate and modify a spreadsheet, rather than a mouse. You can find below a list of common spreadsheet actions and their relevant keyboard shortcuts:

Key or combination	What it does
Tab	End input on the current cell and jump to the cell right to the current one
Enter	End input and jump to the next row (This is designed to be intelligent, so if you're entering multiple columns, it will jump to the first column you are entering)
Up	Move to the cell one row up
Down	Move to the cell one row down

Left	Move to the cell on the left
Right	Move to the cell on the right
Ctrl+<direction>	Move to the outermost cell in the direction given
Shift+<direction>	Select the current cell and the cell in <direction>
Ctrl+Shift+<direction>	Select all cells from the current to the outermost cell in <direction>
Ctrl+c	Copy - copies the selected cells into the clipboard
Ctrl+v	Paste - pastes the clipboard
Ctrl+x	Cut - copies the selected cells into the clipboard and removes them from their original position
Ctrl+z	Undo - undoes the last change you made
Ctrl+y	Redo - undoes an undo

A couple of other features are good to know:

- If you want to freeze the first row of the spreadsheet (which allows you to keep the first row in view even if you scroll down), you can do so in View → Freeze → 1 row. The same is possible for columns. When a row is frozen, it is unaffected by sorting actions.
- To select a full column or row, click on their letter or number, respectively.
- If you want to sort a column (in order to have the highest values at the top, for example), you need to select the whole column, then go to Data → Sort range. Alternatively, you can right click on the column letter and pick either “sort from A to Z” or “sort from Z to A.” These options work on numerical values as well, with the Z to A option corresponding to a descending sort.
- If you want to select a range of cells quickly, you can do it by clicking on the first cell of your range, then shift-clicking on the last cell of your range.
- To select the whole sheet, you just need to click on the gray cell at the top right corner, between the A and the 1.
- You have several options to format your text under the Format menu. Under Format → Text wrapping, you can choose how your text behaves in the cell-- it either spills over to the next cell (if the cell is empty), wraps over multiple lines (if it is longer than the length of the cell, that’s Wrap), or it is cut off at the end of the cell (Clip).

One of the most important tools to know about is the filter feature. It allows you to display only the rows that you want, making it easy to select specific rows even in a dataset with hundreds of rows.



### Walkthrough: Filtering data

1. Select the whole table.
2. Select "Filter" from the "Data" menu.
3. You now should see triangles next to the column names in the first row.
4. Click on the triangle next to 'Item.' A floating menu will appear.
5. Click on 'Clear' to unselect all the options.
6. In the same floating window, scroll the list until you find TOTAL CATEGORY 1 REVENUE.
7. Select it, then click 'OK.'

Voila! You have filtered the data and left only the total value visible in the spreadsheet. To show all the values again, you repeat the process, but click on "Select all" instead.

## Section 3: Tidy data

We mentioned before that the data was structured to facilitate reading, instead of a more complex structure that makes analysis more difficult. For example, filtering budget items corresponding to Section 1 requires selecting them individually in the filter window. This is not ideal. This also makes it harder to create data visualizations out of this data. So, what should a well-structured dataset look like?

When properly structuring tabular data, we want to distinguish:

- The types or categories of data points, with **one type of data point per column**. Each type of information is described across multiple observations.
- The individual observations, with **one observation per row**. An observation is the thing about which we are collecting and analyzing data. Observations can be individuals, organizations, countries, or golf balls. Therefore, each row should represent one observation, a single golf ball, for example, with each observation made of one or more types of information

			type/category of data
	colour	diameter (mm)	purchase date
	white	42.67	2020-1-8
observation	white	42.67	2020-3-11
	white	42.67	2020-2-1

*An example using a dataset about golf balls*

Restructuring your data to facilitate analysis (called data reshaping) can feel overwhelming. Where to start? What should I move? But the process is simple

## Walkthrough: Data reshaping

1. Identify the different categories of data that we have. For the budget dataset, those are
  1. The number of the budget category: here we have only 1, but we can guess that the full budget has multiple categories.
  2. The name of the category: here we have only 'revenue,' but we can guess that other categories can exist, such as 'expenditure.'
  3. The number of the section: here we have two sections with numbers 1 and 2.
  4. The name of the section: 'ordinary tax revenue' and 'ordinary non-tax revenue' are our two options in this data category.
  5. The budget item: this includes all the rows at the exception of the summary rows (section names, total).
  6. The item amount: these are the values corresponding to the item.
2. Select an empty row and create column headers that match the identified data categories.
3. Start by copy/pasting the budget items (without including the budget category/section rows!).
4. From there you can also copy past the budget values, making sure that they match the correct item.
5. Fill the section name, by making sure to match the right section name with the right budget item.
6. Finish the work for the other columns.

You should now have something like this:

category number	category name	section number	section name	budget item	amount
1	revenue	1	Ordinary tax revenue	Property taxes and taxes on activities	68,450,000.00
1	revenue	1	Ordinary tax revenue	Income from occupation and concession of public s	7,060,000.00
1	revenue	1	Ordinary tax revenue	fees for administrative formalities and fees collecte	10,710,000.00
1	revenue	1	Ordinary tax revenue	Other ordinary tax revenue	820.00
1	revenue	2	Ordinary non-tax revenue	Ordinary income from the municipal or regional don	3,440,000.00
1	revenue	2	Ordinary non-tax revenue	Ordinary financial income	30,770,000.00

Each row should contain the exact same type of information as all other rows, across all columns.

## Section 4: Summary table, data visualization

If this presentation looks strange to you, this is normal, after all, we have now a lot of repeat values. The previous format of our table needed only one row to show the category number and name, while this new format has a lot of repeat cells. And we haven't transferred the summary values (section values and total value)!

But the reason why someone who works with data will prefer this format becomes very clear when you try to filter the data. To try it yourself, switch to the 'machine readable' tab of the spreadsheet.

Now select the whole table, then Data → Create a filter. If you open the filter window on the 'section name' column, it becomes apparent that the new structure makes it very easy to filter the items belonging to a specific section!

As for the summary values that we did not transfer, this is because they are easy to regenerate:

### Walkthrough: Generating a summary table

1. In E9, write 'Ordinary tax revenue'. You can also write '=D2' and the same text as D2 will appear.
2. In F9, create a sum of the values of that category: '=SUM(F2:F5)'. Using ':' (colon) allows me to tell the software that I want to include all the cells between F2 and F5. It's the equivalent of writing 'F2,F3,F4,F5'.
3. In E10, write 'Ordinary non-tax revenue' or simply '=D6'.
4. In F10, create a sum similar to the one in F9, but for the correct category: '=SUM(F6,F7)'.
5. In E11, write 'TOTAL'.
6. In F11, create a sum of the two rows above.

You should obtain this:

Ordinary tax revenue		86,220,820.00
Ordinary non-tax revenue		34,210,000.00
Total		120,430,820.00

If you compare those numbers to the ones from the original table, you will notice that they are not the same! The problem is not in our reformatting but in the original data. The value in the summary rows did not match the value of the individual items. This is a very common issue with budget data: if you do not verify the 'total' line yourself, you may miss that there is a discrepancy between the total entered manually by the data producer and the actual value of the sum of all individual items. This is yet another reason to reshape the data--it makes your verification work easier.

And if you needed yet another reason, here is one more: properly structured tables also make data visualization easier!

### Walkthrough: Making a simple chart

1. Select the cells E9 to F10 in order to highlight the section names and their values
2. In the menu, go to Insert → Chart.

Google Sheets should automatically create a pie chart. You are also free to change the visualization, if you wish, by using the settings panel that opened on the right of your spreadsheet. Now look back at the initial dataset. Given the way it was structured, even a basic pie chart of the section values would have been challenging to do.

And we're done! With this module completed, you now have the fundamentals needed to be able to work with data. With the dataset you were given, you were able to:

- **Get** the data, by accessing the spreadsheet and making a copy of it
- **Verify** it by checking if the budget items matched the total
- **Clean** it by reshaping it to facilitate your analysis
- **Analyze** it by calculating the percentages
- **Visualize** it with a pie chart

Those steps are part of what we call the Data Pipeline.

## The Data Pipeline

The Data Pipeline is School of Data's approach to working with data from beginning to end. Once you understand your action cycle and your stakeholders, it will be time to work with the data. We have broken down this process in steps. The Data Pipeline is a work in progress. We started out by suggesting five steps, but our community is constantly experimenting and tweaking it to reflect the core steps that are present in every kind of data-driven project. The steps are:

**Define:** Data-driven projects always have a "define the problem you're trying to solve" component. It's in this stage you start asking questions and focus on the issues that will matter in the end. Defining your problem means going from a theme (e.g., air pollution) to one or multiple specific questions (has bike sharing reduced air pollution?). Being specific forces you to formulate your question in a way that hints at what kind of data will be needed. This, in turn, helps you scope your project: is the data needed easily available? Or does it sound like some key datasets will probably be hard to get?

**Find:** While the problem definition phase hints at what data is needed, finding the data is another step, of varying difficulty. There are a lot of tools and techniques to do that, ranging from a simple question on your social network, to using the tools provided by a search engine (such as Google search operators), open data portals or a Freedom of Information request querying about what data is available in that branch of government. This phase can make or break your project, as you can't do much if you can't find the data! But this is also where creativity can make a difference by using proxy indicators, searching in non-obvious locations... don't give up too soon!

**Get:** Getting the data from its initial location to your computer can be short and easy, or long and painful. Luckily, there're plenty of ways of doing that. You can crowdsource using online forms, you can perform offline data collection, you can use some crazy web scraping skills, or you could simply download the datasets from government sites, using their data portals or through a Freedom of Information request.

**Verify:** We got our hands on the data, but that doesn't mean it's the data we need. We have to verify that details are valid, such as the meta-data, the methodology of collection, if we know who organized the dataset, and that it's a credible source. We've heard a joke once, but it's only funny because it's true: all data is bad, we just need to find out how bad it is!



**Clean:** It's often the case that the data we get and validate is messy. Duplicated rows, column names that don't match the records, values that contain characters which make it difficult for a computer to process, and so on. In this step, we need skills and tools that will help us get the data into a machine-readable format so that we can analyze it. We're talking about tools like OpenRefine (<https://openrefine.org/>) or LibreOffice Calc (<https://www.libreoffice.org/discover/calc/>) and concepts like relational databases.

**Analyze:** This is it! It's here where we get insights about the problem we defined in the beginning. We're going to use our mad mathematical and statistical skills to interview a dataset like any good journalist. But we won't be using a recorder and a notebook. We can analyze datasets using many, many skills and tools. We can use visualizations to get insights of different variables, we can use programming languages packages, such as Pandas (Python) or R, we can use spreadsheet processors, such as LibreOffice Calc, or even statistical suites like IBM SPSS (<https://www.ibm.com/analytics/spss-statistics-software>).

**Present:** And, of course, you will need to present your data. Presenting it is all about thinking of your audience, the questions you set out to answer, and the medium you select to convey your message or start your conversation. You don't have to do it all by yourself. It's good practice to get support from professional designers and storytellers, who are experts at understanding the best ways to present data visually and with words

What you've done during this module is just a tiny fragment of what is possible, but you will use what you've learned here over and over during your data journey. Learning tools, including some presented in the follow-up modules, is important for your productivity, but they can only be mastered through a solid understanding of data fundamentals.

# MODULE 2 – FINDING DATA

---

## Introduction

Now that we know what data is and the questions we're interested in, we're ready to go out and hunt for it online. In this tutorial, you will learn where to start looking for data. In this module, we will look at different ways of getting data before setting you loose to find data yourselves!

We will also consider what types of data you create and/or work with, and what format those datasets use. Your data stewardship practices will be dictated by the types of data that you work with, and what format they are in.

## Description

The module will cover the following concepts:

- Search engines
- Search operators
- Data sources
- Data portals
- Data formats
- Data transformations
- Web, HTML, API, programming

## Skills

As part of this module, you will learn the following:

- How to use advanced search engine features
- How to identify different data formats

## Prerequisites

- Basic knowledge of operating a computer
- Internet access and connection
- A fair understanding of all previous modules
- Module dataset: [Municipal Finance Data](#): (see instructions below to find the dataset 'Cape Town budget')
- A working computer

## Section 1: Understanding data formats

Different types of digital files use different structures. For example, a text file is structured differently than a picture, which is structured differently than a web page. Consequently, most computer applications can only open a few file types; they are programmed to be able to read the specific structure of the files they were designed to operate with. To make it easy for your computer to know which file can be opened by which application, we use extensions. .txt (text file), .pdf (PDF file), .jpg (JPEG image file) are all extensions used to signal the structure of the file and, by extension (no pun intended), the type of software that can open them.

This setup holds true for data files. As we've seen in module 1, a dataset is "a structured representation of the world." But there are many possible structure types! To identify at a glance which structure a data file uses, you can look at its extension. Common data file extensions include:

**.txt** – TXT is the extension for basic text files. It is not a structured data format per se, but it is possible to write data in a text file and have the right software recognize the structure despite the .txt extension.

**.csv/tsv** – CSV stands for Comma Separated Value and is used for storing tabular data, which is data arranged in rows and columns (i.e., tables). An alternative is the TSV file format, which uses tabs instead of commas to separate the values. Both are simply text arranged in a structured way. The .csv extension simply indicates to the software how to read the file, but many applications can detect the csv structure in .txt files too.

Countries & regions	Abuse score	Underlying situation score	Global score	Diff. score 2020
<b>Norway</b>	0	6.72	6.72	-1.12
<b>Finland</b>	0	6.99	6.99	-0.94
<b>Sweden</b>	0	7.24	7.24	-2.01

Countries & Regions,Abuse score,Underlying situation score,Global score,Diff score 2020

Norway,0,6.72,6.72,-1.12

Finland,0,6.99,6.99,-0.94

Sweden,0,7.24,7.24,-2.01

If the Freedom of Press ([https://rsf.org/en/ranking\\_table](https://rsf.org/en/ranking_table)) ranking table was stored as a CSV, it would look like the above.

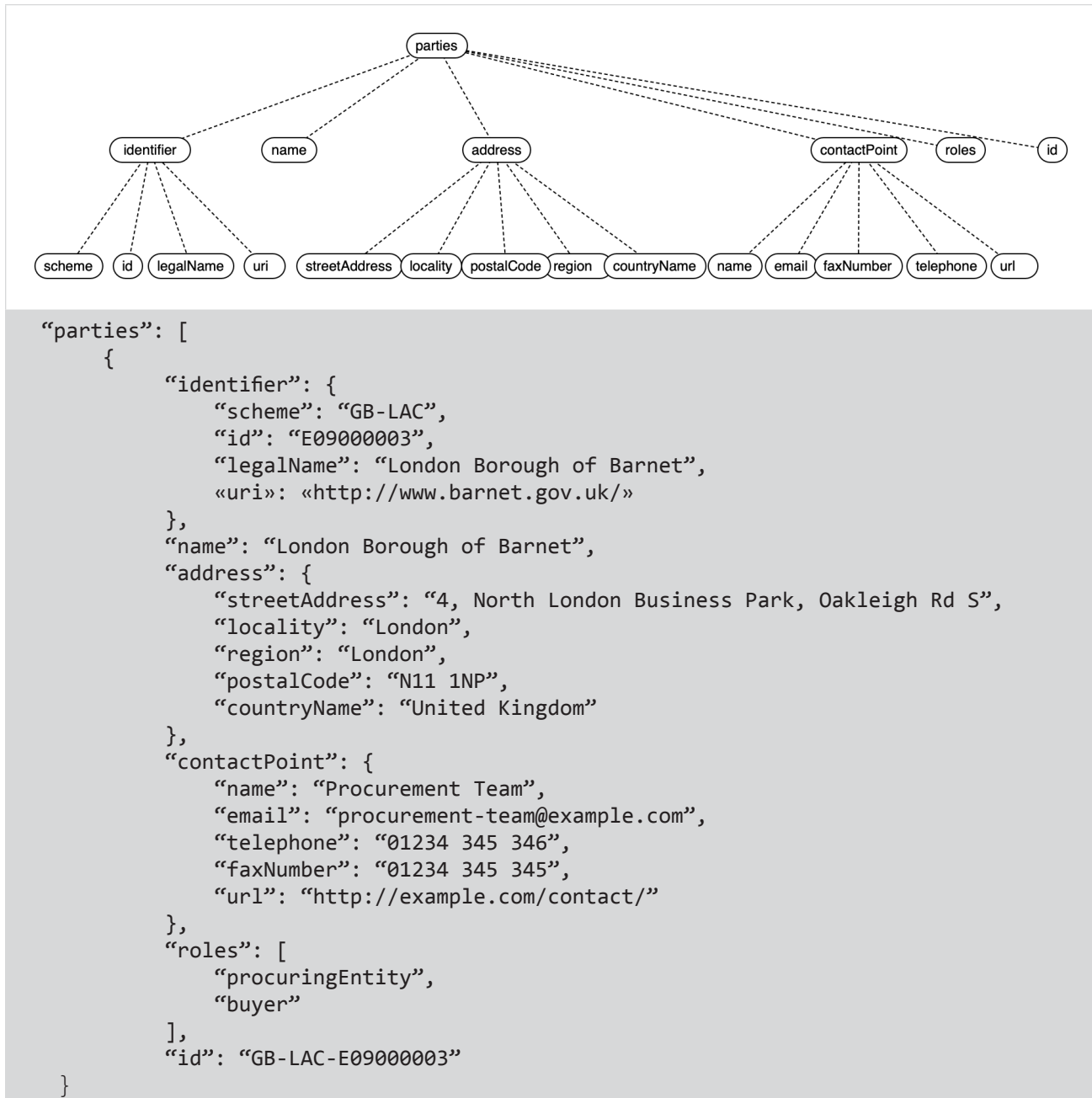
**.xls/.xlsx** – the format used by Excel to store its spreadsheets. Although Excel spreadsheets store tabular data similarly to CSV files, they also carry a lot of additional information which is not data (such as the code used to make the formulas work). The consequence is that XLS files are generally heavier (i.e., they require more computer memory storage) than CSV files and, although commonly used due to the widespread use of Excel, are less suitable for data sharing.

**.ods** – a format similar to xls, but specific to the LibreOffice Calc software, a free and open-source spreadsheet software.



**.doc/.docx** – Similar to .txt files, these are not primarily data files although data can be stored in them (often presented as a table). Like .xls files, they store more than the content they're known for (text), making them heavier than their basic counterparts (TXT files).

**.json/.geojson** – a simple file format designed to be lightweight and easy to read by programming languages as well as easy to share through APIs. Common data sources you may encounter that store data in JSON include Twitter or other social media data, the Open Contracting Data Standard, and data fetched from APIs. While JSON is an international standard like CSV, they store different types of data structures: CSV is designed for tabular data while JSON structures its data in a tree-like structure. GeoJSON is a format used for geographic data which is derived from JSON.



The Open Contracting Data Standard uses JSON to structure public procurement data. [https://standard.open-contracting.org/latest/en/getting\\_started/releases\\_and\\_records/](https://standard.open-contracting.org/latest/en/getting_started/releases_and_records/)

## Tabular vs. Tree Structure

When is each structure most appropriate? As a new data practitioner, most data that you will deal with can easily be stored in a tabular format. This means that you know in advance how many columns your dataset should have, and that all data fits into these columns neatly.

But what if you can't predict some part of your data? Take a list of public tenders; you may want to include, at minimum, the tender number, the tender name, the tender date, the issuing organization, and the bidders. But some tenders will have 2 bidders and others may have 30. To store this in a tabular format, you would have to compromise on the quality of your data.

- Either you store all bidders in one column, which will make analysis harder.
- Or you add a lot of columns in order to write one bidder per column, and hope that the number of bidders will never go over the number of columns you set. You end up with a dataset full of empty values/cells and a cumbersome number of columns (which may or may not be needed).

The tree format can associate an arbitrary number of values to each 'node.' If the number of bidders from your first tender differs from your second one, this will just change the number of tree branches linked to each tender. As a result, the structure of JSON files is slightly more complex for humans to read.

Beyond their differences, both CSV and JSON are international open standards that are widely adopted, designed to be readable by humans without a need for software, and should be preferred over their closed alternatives.

**.xml** – a predecessor to JSON, which also uses the tree data structure and is widely used as an export format for databases.

**.pdf** – PDF stands for Portable Document Format and was introduced at first to facilitate the sharing of image files between designers and printers. It gradually became a de facto standard for storing information that you do not want to be easily modified. Similarly, to a format like .docx, this is not a proper data format nor a simple format like .txt. But its wide usage means that many datasets are only available in PDF format. It is important to distinguish digital PDFs, which store content created by another software, and scanned PDFs, which stores pictures of content that was originally on paper.

**.html** – A web page is a simple document displayable by a browser. Web pages generally use HTML (HyperText Markup Language), which is a structured language that your web browser recognizes. You will often have to deal with data structured using HTML tables or lists.

**.shp** – is the .xls of geodata. Created by a company to store geographical information, it has become the de facto standard for sharing geodata.

Use the table below to find an appropriate and recommended format for preserving and sharing your data over the long term.

## Section 2: Where can I find data?

Concretely, a data source may be a database, a file, live measurements from physical devices (e.g., wind sensors), a table on a web page, a download link on an online portal, or an API.

There are three main questions related to the process of finding data:

1. Where is the data?
2. Which format is it stored under?
3. Do I need to transform it?

### Where is the data?

**On an open data portal:** Open data portals are web-based interfaces designed to make it easier to find re-usable information, for example, <https://municipaldata.treasury.gov.za/> or <https://africaopendata.org/dataset>.

**In the digital system of a public or private organization:** Most organizations store data in some form, and only a small portion of this is released publicly. Knowing which organization may have the data you need is an important skill.

**In the physical archives of a public or private organization:** Similarly, a lot of historical data is stored within physical archives.

**Behind an API:** Using an API (Application Programming Interface) is like using an administrative form used to request documents (like birth certificates). The form was designed by the administration to make it easy for them to fulfil your request. To do that, they set up the form in a way that forces you to formulate your request in a certain way (by checking specific boxes, for example) and input the information that they need to search for your document. APIs work like this, but instead of an administrative form, they're a piece of code that interfaces between you and the system's database, or between two different applications.

**Freely accessible on the web:** While data portals are convenient and becoming more common, a lot of data is still simply displayed on web pages using HTML.

**Nowhere:** Often the data that you're looking for does not exist. You may have to structure (if the information exists in an unstructured form) or collect (if the information does not exist in a single place) it yourself.

## Which format is it stored under?

Beside the list of common data formats shared above, your data could also be stored on more esoteric formats such as microfilms (found in archives), or more complex ones such as .sql (used by databases). It is also not rare to have to extract data from photos or printed pictures. Knowing what format your data is stored under will help you anticipate the steps needed to fetch it.

Localization	Possible formats	Retrieval
On an Open Data Portal	xls, csv, shp, pdf, ods, json, xml, sqlite...	Download
In a non-public digital system	xls, csv, shp, pdf, ods, json, xml, sqlite...	Access request
In a physical archive	Paper, photos, microfilm	Access request
Behind an API	CSV, JSON, XML	With a little bit of programming
On a web page	HTML, PDF	Varies based on the format
Nowhere	Depends on the data collection process	Data collection

## Walkthrough: Downloading data from a data portal

In this tutorial, we'll show how to access South African municipal money data from a data portal maintained by the South African government.

- Open the following South African municipal money data portal:  
<https://municipaldata.treasury.gov.za/>

The screenshot shows the landing page of the 'Municipal Finance Data' portal. The header features the South African national treasury logo and the text 'Municipal Finance Data' with a subtitle 'Current and historical Municipal budget and financial performance data from the National Treasury.' Below this, three key statistics are highlighted: '4 years of data' (Financial years 2012-2013 to 2015-2016), '278 municipalities' (8 metros, 44 district and 226 local municipalities), and '39 million facts' (Budgeted and actual figures for income and expenditure, cash flow and lots more). At the bottom, there is a section titled 'Explore a dataset' with four cards: 'Aged Creditor Analysis', 'Aged Creditor Analysis', 'Aged Debtor Analysis', and 'Aged Debtor Analysis'. Each card provides a brief description of the data and its update status.

Landing page for the portal

- You'll notice there are a range of datasets on the website. Select the money data you'd like to download. For example, we'll select 'Income and Expenditure' data for the City of Cape Town.

**Income and Expenditure**

City of Cape Town

Financial year end: 2019

Amount type: Audited Actual

Government function: All government functions

Category	Amount
6100 OPERATING REVENUE	19 762 688 021
6200 Property Rates	96 379 289
6300 Property Rates - Penalties And Collection Charges	230 144 449
6400 Service Charges	19 754 554 021
6500 Rent Of Facilities And Equipment	610 564 000
6600 Interest Earned - External Investments	1 142 983 906
6700 Interest Earned - Outstanding Debtors	367 402 226
6800 Dividends Received	1 409 180 381
6900 Loans and Permits	56 379 289
7000 Agency Services	230 144 449
7100 Transfers Received - Operating	7 027 548 753
7200 Transfers Received - Capital	2 079 449 346
7300 Other Revenue	790 676 176
7400 Loan On Disposal Of Property, Plant & Equipment	120 453 449
7500 Total Operating Revenue Generated	44 389 488 584
7600 Less Revenue Foregone	1 460 179 026
7700 Total Direct Operating Revenue	42 929 309 558
7800 INTERNAL TRANSFERS - (NET OUT WITH -)	1 302 134 763
7900 Internal Revenue Activity Based Costing Etc.	10 076 489 107
8000 Disbursements Received - Internal From Municipal Entities	-
8100 Total Indirect Operating Revenue	19 430 620 870
8200 Total Operating Revenue	62 370 930 428
8300 OPERATING EXPENDITURE	-
8400 Employee Related Costs - Wages & Salaries	9 916 279 589
8500 Employee Related Costs - Social Contributions	2 949 900 160
8600 Less Employee Costs Capitalised	-31 285 509
8700 Less Employee Costs Allocated To Other Operating...	-
8800 Remuneration Of Councilors	150 180 310
8900 Debt Impairment	1 969 391 291
9000 Collection Costs	200 200 200
9100 Depreciation and Asset Impairment	2 889 927 445
9200 Interest Expense - External Borrowings	703 889 491
9300 Redemption Payments - External Borrowings (Gains...)	-
9400 Bulk Purchases	6 864 620 646
9500 Other Materials	-
9600 Contracted Services	2 096 008 146
9700 Grants and Subsidies	171 583 156
9800 Other Expenditure	7 574 054 039

### Income and Expenditure for the City of Cape Town

- Click the download button. You'll be prompted to download the data as a CSV or Excel file or using the API. The CSV data format will look like the screenshot below. With this, it's now possible to verify the quality of the data as explained in the next module on 'Verifying data'

income\_aggregate\_2020-12-14T18:53:14.520337.csv (read-only) - LibreOffice Calc

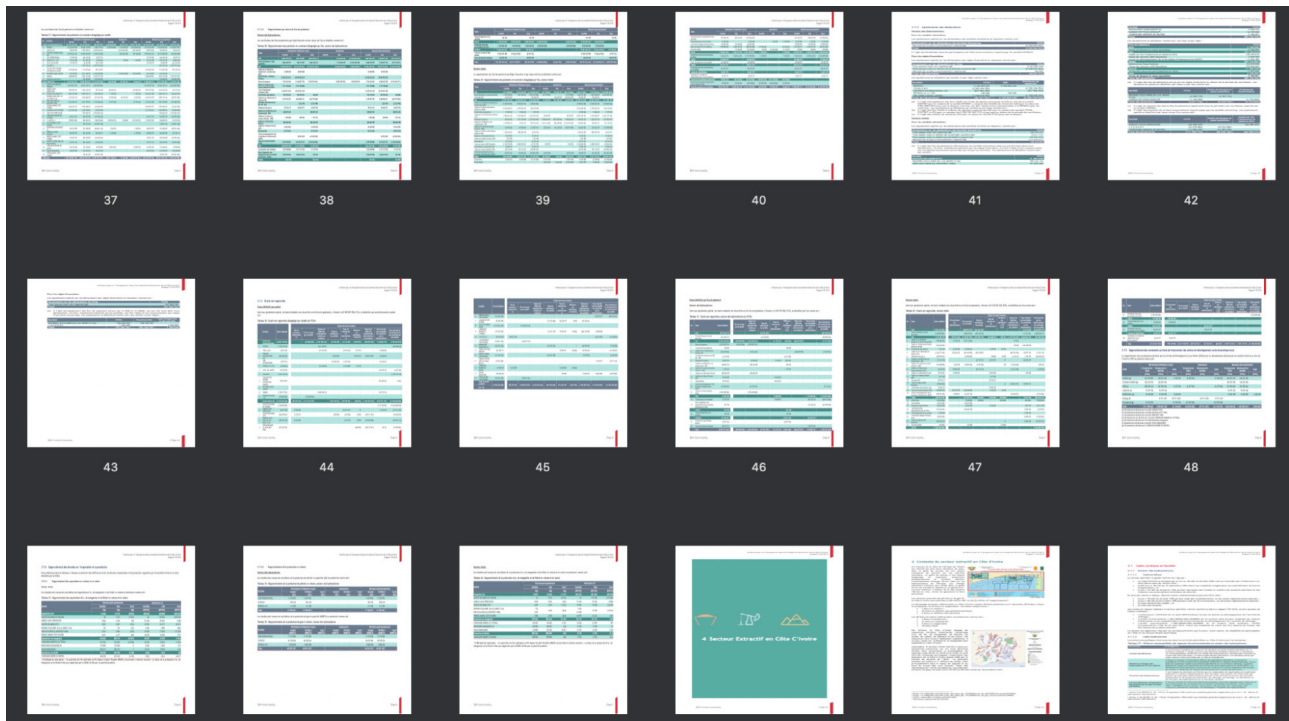
Financial year	Amount	Label
2019	19 762 688 021	6100 OPERATING REVENUE
2019	96 379 289	6200 Property Rates
2019	230 144 449	6300 Property Rates - Penalties And Collection Charges
2019	19 754 554 021	6400 Service Charges
2019	610 564 000	6500 Rent Of Facilities And Equipment
2019	1 142 983 906	6600 Interest Earned - External Investments
2019	367 402 226	6700 Interest Earned - Outstanding Debtors
2019	1 409 180 381	6800 Dividends Received
2019	56 379 289	6900 Loans and Permits
2019	230 144 449	7000 Agency Services
2019	7 027 548 753	7100 Transfers Received - Operating
2019	2 079 449 346	7200 Transfers Received - Capital
2019	790 676 176	7300 Other Revenue
2019	120 453 449	7400 Loan On Disposal Of Property, Plant & Equipment
2019	44 389 488 584	7500 Total Operating Revenue Generated
2019	1 460 179 026	7600 Less Revenue Foregone
2019	42 929 309 558	7700 Total Direct Operating Revenue
2019	1 302 134 763	7800 INTERNAL TRANSFERS - (NET OUT WITH -)
2019	10 076 489 107	7900 Internal Revenue Activity Based Costing Etc.
2019	-	8000 Disbursements Received - Internal From Municipal Entities
2019	19 430 620 870	8100 Total Indirect Operating Revenue
2019	62 370 930 428	8200 Total Operating Revenue
2019	-	8300 OPERATING EXPENDITURE
2019	9 916 279 589	8400 Employee Related Costs - Wages & Salaries
2019	2 949 900 160	8500 Employee Related Costs - Social Contributions
2019	-31 285 509	8600 Less Employee Costs Capitalised
2019	-	8700 Less Employee Costs Allocated To Other Operating...
2019	150 180 310	8800 Remuneration Of Councilors
2019	1 969 391 291	8900 Debt Impairment
2019	200 200 200	9000 Collection Costs
2019	2 889 927 445	9100 Depreciation and Asset Impairment
2019	703 889 491	9200 Interest Expense - External Borrowings
2019	-	9300 Redemption Payments - External Borrowings (Gains...)
2019	6 864 620 646	9400 Bulk Purchases
2019	-	9500 Other Materials
2019	2 096 008 146	9600 Contracted Services
2019	171 583 156	9700 Grants and Subsidies
2019	7 574 054 039	9800 Other Expenditure

### Downloaded data in CSV format

## Do I need to transform it?

As mentioned previously, many of the formats used to store data are not proper data formats. Making use of this data consequently requires you to transform or extract it first. The level of effort depends on the format, but you some common transformation steps include:

**Extracting data from digital PDFs:** although it is possible to copy paragraphs of text from digital PDFs and paste it somewhere else, trying to do the same for data tables is a recipe for disaster. What about when there are hundreds of pages of PDFs, such as in country reports done under the EITI (Extractive Industry Transparency Initiative) program? In this case, automation is the most reasonable answer. And luckily, the world of open source software has our back. Tabula (<https://tabula.technology/>), a free and open source piece of software initially developed by data journalists, serves this use case perfectly.



The 2018 Ivory Coast EITI report includes hundreds of pages of tables <https://eiti.org/files/documents/rapport-itie-ci-2018-version-finale-30-12-20.pdf>

**Using OCR:** OCR stands for Optical Character Recognition and is useful when data is stored in an image file such as a scanned PDF. Unlike a digital PDF, which allows you to highlight text and copy it, scanned PDFs cannot be interacted with, as their content is pictures. To extract data from pictures, you need a technology able to identify the letters in a document, just like a human would. This is the purpose of OCR software. Although the extraction process is often full of errors, using OCR can be a key time saver in your project. There is unfortunately no free and open-source OCR software that is reliable enough to be recommended at the time of writing. But a number of options are available for the motivated, at different price ranges.

Format	Retrieval
XLS, CSV, SHP, ODS, JSON, XML, SQL	No additional step
Digital PDF	Manual extraction or with the help of a software like Tabula ( <a href="https://tabula.pdf">https://tabula.pdf</a> )
Scanned PDF	Manual extraction or with the help of Optical Character Recognition (OCR) software
Paper, photograph, microfilm	Manual extraction or with the help of an OCR-equipped scan machine
HTML	Manual or automated (scraping) collection

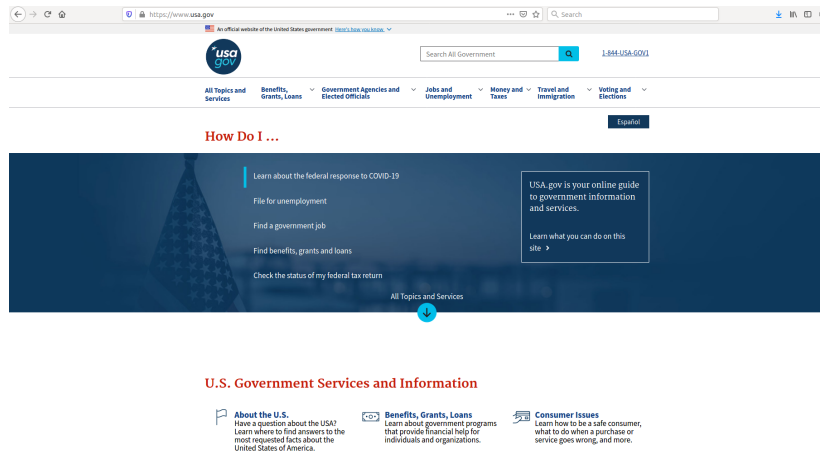
## Section 3: Searching for data

Now that a lot of data is available on the web, search skills have become essential to quickly find the specific dataset or document needed.

### Walkthrough: Using Google search operators

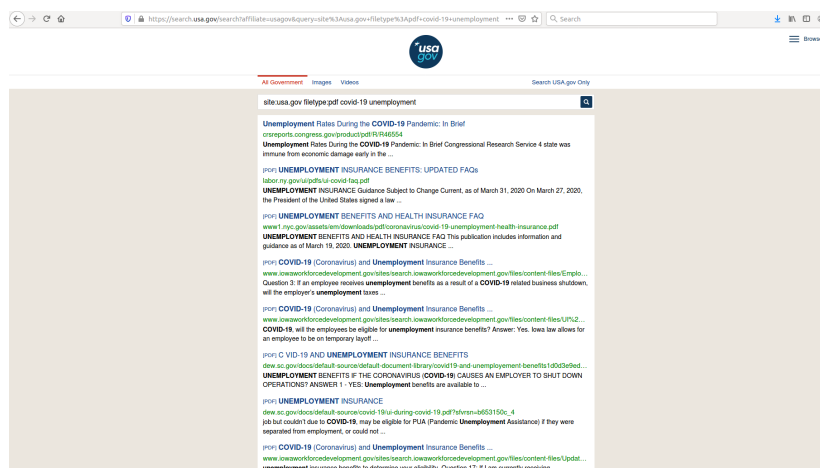
In this walkthrough, we will search for a report with data on a specific website. The example website is [usa.gov](https://www.usa.gov), the US government services website. The file type we're searching for is a .pdf with information on COVID-19 employment.

- Open the [usa.gov](https://www.usa.gov) website



The [usa.gov](https://www.usa.gov) landing website page

- You'll notice a search bar on the website. Type the following string 'site:usa.gov filetype:pdf covid-19 unemployment' in the search bar



Results from the search

- Based on the results, you can now select a pdf file and then verify if the data you're searching for is contained within.



## External resources

- An overview of data files: <http://opendatahandbook.org/guide/en/appendices/file-formats/>
- Google search operators <https://support.google.com/websearch/answer/2466433?hl=en>

# MODULE 3 – GETTING DATA

---

## Introduction

We saw in the previous modules the many forms data can take and the various locations where it could be located. Through that process, we started discussing methods and tools for the retrieval of data, which is the step that immediately follows the search for data.

Even after confirming that a dataset is available, getting it on your computer can be an arduous process or a test of patience. Getting data in a readable format is a specific step that requires dedicated planning, including the time needed to research and locate the appropriate tools for the task. Very rarely will data just be magically available to analyze. One of the most common roadblocks faced by journalists and civic organizations is the lack of disclosure of the data owned by a public administration. Your first step should always be to ask nicely, but that is often not enough. This is where FOI (Freedom of Information) requests become necessary. Provided that your country has a law guaranteeing this right (as of April 2021, 119 countries have a FOI law), you should be able to make a formal request to request the release of the data you seek. This is no guarantee that you'll get it, but it is an important step too often forgotten.

While FOI requests leave you at the mercy of public administrations, there are many ways to take matters in your own hands when it comes to getting data. We'll cover three of them: PDF extraction, web scraping, and field data collection.

## Description

The module will cover the following concepts:

- Web scraping
- PDF extraction
- Field data collection

## Skills

As part of this module, you will learn the following:

- Using Tabula to extract data from PDFs
- Using Google Sheets to scrape data from the web
- Using Kobo Toolbox for doing field surveys

## Prerequisites

- Basic knowledge of operating a computer
- Internet access and a working connection
- A fair understanding of all previous modules
- Module dataset: <https://eiti.org/files/documents/rapport-itie-ci-2018-version-finale-30-12-20.pdf>
- A Google account
- Excel (optional)
- An Android device (optional, for the end of the module)
- A working computer

## Section 1: Extracting data from digital PDFs

Extracting data from PDFs can be done with:

- PDF to Word/Excel converters which allow you to copy the information you need. But the result is often messy if there are tables in the pdf. Some free tools include Excel Online.
- OCR (Optical Character Recognition) which “reads” the PDF and then copies its content in a different format, usually simple text. OCR are most of the time not needed for digital PDFs, but for the rare case when the PDF has a unique formatting that other methods can’t make sense of, OCR can help.
- Programming, with some libraries existing for Python (PDFMiner), Java (Tika, PDFBoc), and your computer’s command line (pdftotext, pdftohtml).
- Crowdsourcing, which is not specifically for PDFs, but can be used when you have many documents to transcript.
- And Tabula, a free and open-source software specifically designed to get data out of PDF tables, which is often where the data you’re looking for lives.

### Introduction to Tabula

Tabula is an offline software, available under MIT open-source license, that allows you to upload a PDF file and extract a selection of rows and columns from any table it may contain.

Tabula is available for the three major operating systems – Windows, Mac, and Linux. It works in a Java environment so you will have to download the Java runtime environment if you don’t already have it (you will be prompted to do it when trying to run Tabula for the first time).

Note: Tabula for Mac OS X comes with Java.

### Walkthrough: Installing and starting Tabula

Once the program is downloaded, you are halfway toward your first table extraction.

- Your downloaded file would be a zip file, so extract the folder within.
- Go into the extracted folder to find the application.
  - If on a Mac, you can drag and drop the application file into your Applications folder.
- Run the Tabula program.
- It should automatically open a terminal window (this is normal, don’t panic) that will show the different processes starting up. Then a new tab should appear in your browser with the Tabula interface.
  - Note: although Tabula uses a browser tab to display its interface, it does not need internet to work.
- If it does not launch on your browser, you can try to directly enter the URL – <http://localhost:8080>
- You should now see the user interface of Tabula.

Import one or more PDFs

---

Imported PDFs

File Name	Size	Pages	Date Added	Remove	Process
-----------	------	-------	------------	--------	---------

---

If you have several PDFs with the same layout, you can select the appropriate regions once, then save the selections as a Tabula Template from the Select Tables page. If someone has shared a template with you, you can upload it to Tabula at the [My Templates page](#).

Tabula's user interface

## Extracting your data

Tabula is a pretty easy application to use once installed. The following steps should see you through the process:

### Walkthrough: Extracting PDF data with Tabula

Note: You can use any PDF that you have on hand, as long as it contains a data table. But if you want to follow along with the same PDF, download it from here: <https://eiti.org/files/documents/rapport-itie-ci-2018-version-finale-30-12-20.pdf>

Upload your PDF file: Run the application file in your extracted folder. Tabula should launch and show the interface in the picture below. Click on the Browse button as highlighted on the image to select among your documents the PDF you want to extract from.

Import one or more PDFs

---

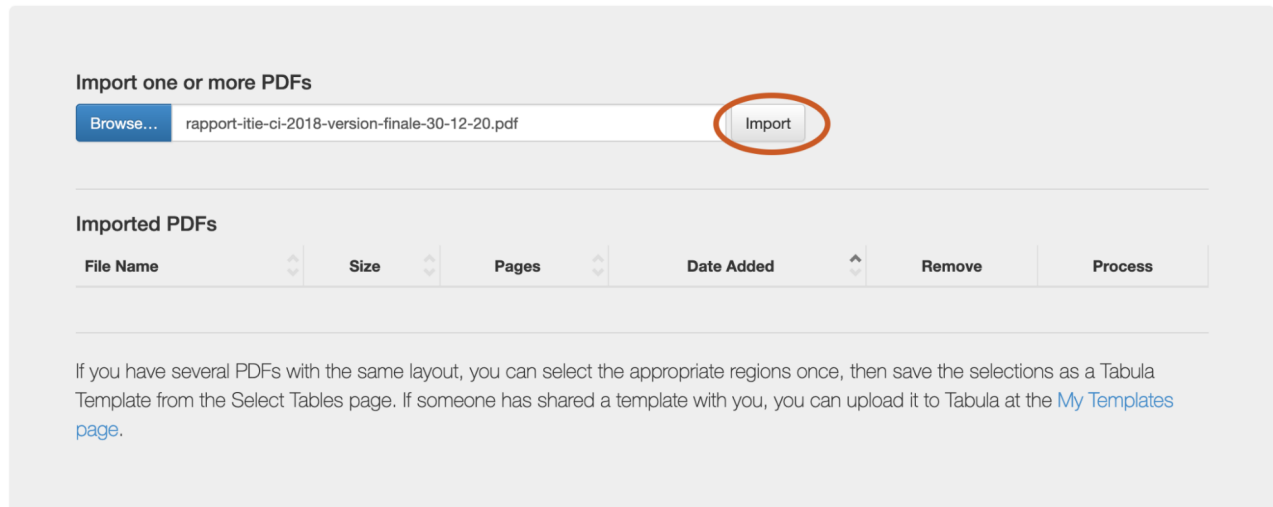
Imported PDFs

File Name	Size	Pages	Date Added	Remove	Process
-----------	------	-------	------------	--------	---------

---

If you have several PDFs with the same layout, you can select the appropriate regions once, then save the selections as a Tabula Template from the Select Tables page. If someone has shared a template with you, you can upload it to Tabula at the [My Templates page](#).

Once that is done, click on Import, as seen on the picture below:



Tabula will take a few seconds to load the PDF. When the PDF is loaded, the screen will change to show a preview of the PDF with the Tabula toolbar at the top.



#### *The Tabula toolbar*

The main buttons that we are interested in are:

- Autodetect Tables, which can do as its names indicates.
- Preview & Export Extracted Data which we'll use once we're happy of our selection.

The goal of the next steps is to indicate to Tabula where the data tables are to prepare their extraction. We have two easy ways of doing that:

- Trust Tabula's table detection algorithm and use the 'Autodetect Tables' button (you can always modify the selection afterwards).
- Scroll to each data table manually.

Let's try to click on 'Autodetect tables' and see what we get.

We can see that the algorithm is not perfect. It highlighted the table of content (page 3) thinking that it was a table. But no problem! We can easily delete a specific selection by using the x button on the top right, as circled below:

rapport-itie-ci-2018-version-finale-30-12... Templates Clear All Selections Autodetect Tables Preview & Export Extracted Data

3.

Initiative pour la Transparence dans les Industries Extractives de la Côte d'Ivoire  
Rapport ITIE 2018

**Table des matières**

1. RESUME EXECUTIF .....	10
1.1. Introduction .....	10
1.2. Chiffres clés du Rapport ITIE 2018 .....	11
1.3. Etendue du rapport .....	15
1.4. Exhaustivité et fiabilité des données .....	16
1.5. Résultats des travaux de conciliation .....	18
1.6. Recommandations .....	23
2. APERÇU SUR L'ITIE COTE D'IVOIRE .....	25

Inversely, Tabula missed a table on page 11. No problem! We can add it manually to the selection by clicking on the top right corner of the table and dragging the cursor until the red highlight fully covers the table.

**Tableau 1 : Total des revenus du secteur extractif en 2018**

Payments	2018 (milliard FCFA)	%
<b>Contribution au budget national</b>	<b>206,88</b>	<b>68%</b>
Impôts et taxes payés à la DGI	74,07	24%
Produits de la vente des parts de production de l'Etat dans les champs gaziers à CIE (revenus recouverts par compensation avec les factures d'achats d'électricité)	55,41	18%
Produits de la vente des parts de production de l'Etat dans les champs pétroliers reversés à la DGI	41,92	14%
Dividendes payées à la DGTCP	16,90	6%
Droits de douanes et pénalités payés à la DGD	13,48	4%
Droits et redevances payés à la DGMG	5,10	2%
<b>Paielements collectés par PETROCI</b>	<b>90,62</b>	<b>30%</b>
Ventes des parts de PETROCI de Gaz	57,70	19%
Ventes des parts de PETROCI de pétrole brut	32,92	11%
<b>Paielements à la DGH (Formation et Equipement)</b>	<b>2,67</b>	<b>1%</b>
<b>Paielements sociaux des sociétés incluses dans le périmètre de conciliation</b>	<b>2,86</b>	<b>1%</b>
<b>Paielements collectés par PETROCI CI-11 (*)</b>	<b>-</b>	<b>0%</b>
<b>Total</b>	<b>303,03</b>	<b>100%</b>

(\*) Les revenus de la vente de la part de PETROCI CI-11 n'ont pas été rapportés par PETROCI CI-11

Using the two steps explained above, you're now able to remove all the incorrect selections and add all the missing ones. Luckily, most tables were taken into account properly! The next step is to click on the toolbar button 'Preview & Export Extracted Data.'

Tabula is now showing how the data will look like when exported. If it does not look completely right, this is normal. Unless the data tables are very clean and well structured, it is hard for Tabula to identify things like merged cells or subheadings. But we can be sure that the numbers themselves are correct, which is not the case when trying to copy the tables manually.

**Is the extracted data incorrect?**

You can revise your selected cells or try an alternate extraction method.

**Revise Selected Cells**

Data has been extracted from the cells you selected in the previous step. You can revise your selection(s) to add or remove cells.

← Revise selection(s)

**Choose Alternate Extraction Method**

The current preview uses the **Stream** extraction method. If the data is not mapped to the correct cells, try the **Lattice** method instead.

Stream Lattice

Stream looks for *whitespace* between columns, while Lattice looks for *boundary lines* between columns.

rapport-itie-ci-2018-version-finale-30-12...

Export Format: CSV

Export

Copy to Clipboard

### Preview of Extracted Tabular Data

Payments	2018
	(milliard FCFA)
Contribution au budget national	206,88
Impôts et taxes payés à la DGI	74,07
Produits de la vente des parts de production de l'Etat dans les champs	
gaziers à CIE (revenus recouvrés par compensation avec les factures d'achats	55,41
d'électricité)	
Produits de la vente des parts de production de l'Etat dans les champs	
	41,92
pétroliers reversées à la DGI	
Dividendes payées à la DGTCP	16,90

There are two main options we care about here:

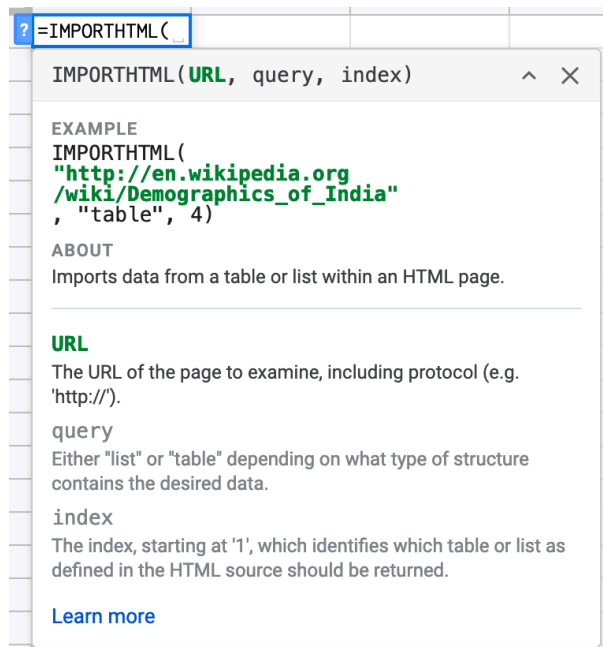
- Choose alternative extraction method: under this sidebar heading are two buttons: stream and lattice. They represent two different algorithms for identifying the boundaries of the data table. By default, Tabula uses stream, but if the preview doesn't look right, try to switch to lattice to compare. No guarantee that it will be better, though!
- Export format: this dropdown allows us to choose CSV, JSON or zip of CSVs. This last option is interesting when you do not want to merge all the tables you have selected into one big CSV/ spreadsheet. Instead, Tabula will produce one CSV per selected table.

Finalizing the data extraction is as simple as clicking on 'Export'. Voila! You have now data in a proper format that you can open in any spreadsheet software.



## Section 2: An introduction to web scraping

Data on the web is structured using HTML. Specifically, the most two common forms that you'll encounter are HTML tables and lists. As it happens, Google Sheets has a special function called 'IMPORTHTML'.



This function automates the process of extracting data from a HTML table or list and importing it into your spreadsheet.

### Walkthrough: Scraping data with IMPORTHTML

- Open a new Google Sheet.
- Choose any web page of your choice, as long as it contains a table.
  - The example will use the Freedom of Press ranking page: [https://rsf.org/en/ranking\\_table](https://rsf.org/en/ranking_table)
- Write the IMPORTHTML function in any cell of your spreadsheet and insert the two arguments it requires to function: the URL and the word 'table' or 'list.'
  - For our example, we will be using 'table' as we're not extracting a list.
  - Read the instructions well: many people miss the fact that both the URL and 'table' should be in-between double quotes.

```
=IMPORTHTML("https://rsf.org/en/ranking_table","table")
```

- After a few seconds of loading, the data should be displayed in your spreadsheet!

	A	B	C	D	E	F	G	H	
1		Ranking	Countries & regions	Abuse score	Underlying situation	Global score	Diff. score 2020	Diff. Position 2020	
2		1	Norway	0	6.72	6.72	-1.12	0	
3		2	Finland	0	6.99	6.99	-0.94	0	
4		3	Sweden	0	7.24	7.24	-2.01	1	
5		4	Denmark	0	8.57	8.57	0.44	-1	
6		5	Costa Rica	10.99	8.21	8.76	-1.77	2	
7		6	Netherlands	13.86	8.74	9.67	-0.29	-1	
8		7	Jamaica	0	9.96	9.96	-0.55	-1	
9		8	New Zealand	0	10.04	10.04	-0.65	1	
10		9	Portugal	0	10.11	10.11	-1.72	1	
11		10	Switzerland	6.93	10.55	10.55	-0.07	-2	
12		11	Belgium	6.93	11.69	11.69	-0.88	1	

There are a few things to note:

- When the formula tries to display the table, it will first check that it is not writing on top of existing data. So, make sure to clean up the cells around the formula!
- Although you can reference the scraped table in your formulas, you can't edit it directly. To do so you need to:
  - Select the full table
  - Copy it using CTRL+C (Windows) or CMD+C (Mac)
  - Create a new tab
  - Go to Edit → Paste special → Paste values only
  - You now have the original table in one tab and an editable version in another! It's always a good idea to avoid modifying the original data.

There are many other ways to scrape web data. Some require programming skills, but many do not — for a fee. Among the free ones, the web scraper Chrome/Chromium extension is a very powerful tool for which tutorials can be found at <https://webscraper.io>

## Section 3: Collecting data

Decades ago, the use of paper and pen was a painstaking and very expensive way to collect data. Most of us have experienced paper forms getting wet or damaged, or receiving paper forms that were barely answered. But as the age of smartphones and tablets arrived, mobile-based data collection technologies have also gained a huge following. The use of mobile data collection tools has also improved the conduct of surveys and assessments. Some of the advantages of using mobile data collection are:

- Most people are using smartphones for SMS messaging and have access to mobile data connection. According to Wikipedia, the Philippines is currently 12th in the world with the greatest number of cellphones, with more than 100 million cellphones, which is more than the U.S. population!
- In the absence of laptops and desktop computers, smartphones are cheap and easy to use.
- Mobile-based data collection, compared with the use of paper forms, lessens the risk of losing the data when paper forms are damaged or lost.

One of the tools that was frequently used by information managers is the Open Data Kit (ODK <https://opendatakit.org/>). This was first introduced in the Philippines during the Typhoon Pablo response in

2011 for project monitoring. In the next emergency responses, ODK has been used to conduct one-off surveys and rapid needs assessments during and directly following disasters. While there is a huge variety of online and offline data collection tools, ODK has gained a lot of users because it is free, open source, easy to use, and can be used both offline and online. Since ODK is a free and open-source set of tools which help organizations author, field, and manage mobile data collection solutions, ODK in itself has evolved into several platforms and formats such as:

- Kobo Toolbox <http://www.kobotoolbox.org/>
- GeoODK <http://geoodk.com/>
- Enketo <https://enketo.org/>

Each one seeking to customize the use of ODK according to their own needs.

ODK provides an out-of-the-box solution for users to:

- Build a data collection form or survey (XLSForm is recommended for larger forms).
- Collect the data on a mobile device and send it to a server.
- Aggregate the collected data on a server and extract it in useful formats.

This will be an introduction to using Kobo Toolbox as one of the many platforms in which ODK forms are built, collected, and aggregated for better data collection and management. According to Kobo Toolbox, acknowledging that many agencies are already using ODK, a de facto open-source standard for mobile data collection, Kobo Toolbox is fully compatible and interchangeable with ODK but delivers more functionality such as an easy-to-use form builder, question libraries and integrated data management. It also integrates other open-source ODK-based developments, such as Formhub and Enketo. Kobo Toolbox can be used online and offline. You can share the data and monitor submissions together with other users and it offers **unlimited** use for humanitarian actors.

## The data collection form

Building the data collection form or survey can be done using Microsoft Excel and XLSForm. XLSForm (<http://xlsform.org/>) is an application developed by Nafundi (<https://nafundi.com/>) used to create and validate forms for ODK (and its now open-source!). This module will only focus on creating the form using Excel, but similar steps can be replicated using another spreadsheet software. You are also able to download the full spreadsheet here: <http://bit.ly/KOBOtutorial>

### Walkthrough: Creating your data collection form

To start off, create your file with the following sheets named accordingly.

**Survey** – this sheet will include all survey questions, type of questions, label, and other instructions which ODK will interpret once the form has been uploaded.

**Settings** – this sheet will determine how your sheet will be viewed on your mobile device.

**Choices** – this sheet will include all the choices that will be used for single or multiple-choice questions from your survey sheet.

12				
13				
14				
15				

+ 
☰ 
survey 
choices 
settings

ODK lists a few other sheets that you can use depending on the type of the form that you are trying to create. For purposes of this module, we can use the three sheets listed above.

On your Survey sheet, type in the following column headers:

	A	B	C	D	E	F	G	H	I	J
1	type	name	label	hint	calculation	appearance	relevant	constraint	constraint_message	required
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

+ 
☰ 
survey 
choices 
settings

**Type (required)** – ODK recognizes a set of question types, like single select (Yes or No questions), multi-select, text, numbers, and even photos and geographical locations. You can also use grouped or repeated questions. For Kobo Collect, there are basic types of questions which you can use for the survey. For a more technical ODK form, ODK lists samples of data entry widgets (<https://opendatakit.org/help/form-design/examples/>) you can use.

**Name (required)** – this column is the headers for the responses. The name should be related to your question. Names should be unique and must not have spaces. Your names must only have letters, numbers and/or underscore e.g., pop1, no people or city.

**Label (required)** – this is basically how your survey questions will appear on your mobile device. You can type this in whatever format you choose (e.g., What is your name, age, date today, etc.). Try to copy the screenshot on your own sheet.

	A	B	C	D
1	type	name	label	hint
2	begin_group	group_profile	Sample ODK Survey	
3	date	date1	Date today	
4	date	birthdate	Birth date	
5	text	name	What is your name?	
6	end_group	profile_end		
7	begin_group	group_others	Other stuff about you	
8	geopoint	geopoint	Where are you right now?	
9	integer	siblings	How many siblings do you have?	optional
10	integer	days_of_work	How many days in a week do you work?	
11	calculate	workdays	How many days in a week do you work?	
12	note	no_work	You only have \${workdays} days of rest.	
13	select_one education	education	Highest educational attainment	
14	select_one yes_no	philippines	Are you from the Philippines?	
15	select_multiple country	countries	If not, which country are you from?	

**Hint** – as the name implies, this is where you can give your respondents a hint on how you want your question answered.

**Calculation** – for questions requiring numerical answers, ODK gives you a chance to do calculations. \${name\_of\_calculated\_field} is the expression used to perform a calculation. Under this column, you can request ODK to perform simple calculations for you. As an example, in the screenshot below, the name of your calculated field for the question “how many days in a week do you work?” is days\_of\_work. 7-\${days\_of\_work} is your calculation. This means that if your responder answers 6, Kobo Toolbox will perform the calculation 7.

9	integer	siblings	How many siblings do you have?	optional	
10	integer	days_of_work	How many days in a week do you work?		
11	calculate	workdays	How many days in a week do you work?		7 - \${days_of_work}
12	note	no_work	You only have \${workdays} days of rest.		

**Appearance** – this column determines how your question or groups of question will appear on your mobile device. Questions can be grouped together using the code field-list per group of questions.

**Constraint** – this column is used to set restrictions for numerical questions. If you would like to have a minimum or maximum value of numerical answers, you can put in a constraint (i.e., .>= 0 and .<=1000) which means you can only answer values between 0 to 1000. If you go back to calculation example, you can put a constraint that answers can only be between 0-7 to make a logical calculation. Thus, your constraint can be typed as (.>=0 and .<=7).

**Constraint message** – if you have a constraint, this column allows you to show an error message if the responder fails to follow the constraint.

**Required** – just put in yes across the question which you want to require the responder to answer.

Under the choices sheet, you need to put the following column headers:

	A	B	C
1	list_name	name	label
2	yes_no	yes	Yes
3	yes_no	no	No
4	education	basic_ed	Basic Education
5	education	highschool	Highschool Education
6	education	college	College Education
7	education	post_graduate	Post-Graduate Education
8	country	asia	Somewhere in Asia
9	country	europa	Somewhere in Europe
10	country	pacific	Somewhere in Pacific
11	country	africa	Somewhere in Africa
12	country	north_america	Somewhere in North America
13	country	south_america	Somewhere in South America
14			
15			
16			

**List\_name** – this one serves the same purpose as your type column in the survey sheet. This is your main reference in single-select or multiple-select questions in your survey sheet.

**Name** – this is similar to your name column from the survey sheet. This must be unique and is related to your choices labels. You can use letters, number and/or underscore for the names.

**Label** – this shows how your choices will appear on your device.

In the settings sheet, you can have the following headers:

	A	B	C
1	form_title	form_id	default_language
2	Sample ODK Form	sample_odk	english
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			

**Form title** – this is how the name of your form will appear in your form list.

**Form id** – ODK creates an ID for each of the forms that you create. Make sure that your form\_id is unique.

**default\_language** – this determines what the default language will be for the form. If you have translated your survey sheet into another language, you can create another default language column with your second language preference. Your user will then be prompted to choose which language they want to use in answering the form.

Once you are done with your survey form, make sure you save it in this format: name of file.xls. You can include underscores (\_) if you want. To access the file, we created during this walkthrough, download it from here: <http://bit.ly/KOBOtutorial>

## Uploading and testing for forms

We will only use Kobo Toolbox online in setting up your ODK forms. Thus, it is very important that you have already created your Kobo Toolbox account. If you have not done so, create one here:

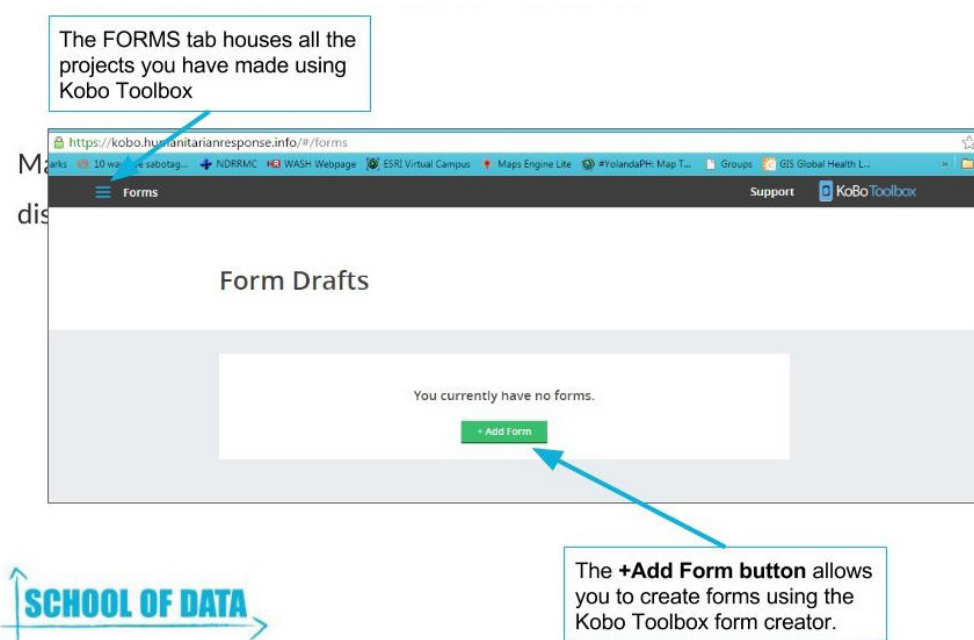
<https://kobo.humanitarianresponse.info/>

People prefer to use the Kobo Collect app on their phones because of mobility and because the forms can be used offline.

It is ideal to use the Kobo Collect app using your phone or tablet if:

- You need to do more than 50 household or field level surveys.
- You have limited access to the internet, in which case you can save your forms on your phone and submit them when you already have internet access.
- You have few enumerators and little equipment to use, thus, enumerators can just use their own phones or tablets to do data collection.

## The interface:



The Kobo Toolbox website will initially show you the form drafts that you have created using the online Kobo Toolbox form creator. However, for this module, we will only focus on uploading the Excel forms we have previously created.

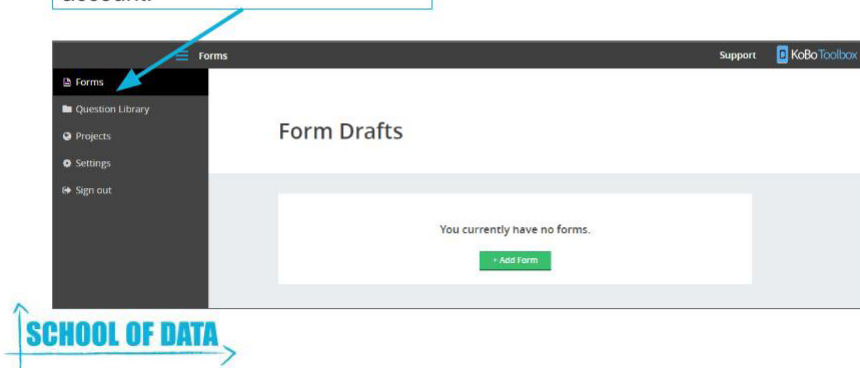
Once you click the forms tab, you will see all the forms you have uploaded.

**The forms tab** will show you:

1. **Question Library** – this is where you can save the usual questions that you are using for the forms you have made using the online form creator.
2. **Projects** – these are the projects or forms you have created both using the online form creator and the uploaded Excel forms.
3. **Settings** – this shows your user profile if you want to include metadata on your account and if you would like to require authentication to view and edit forms.

## The FORMS tab

By clicking the Forms tab, you will see your Question Library, Projects, Settings and this is where you can sign out of your account.



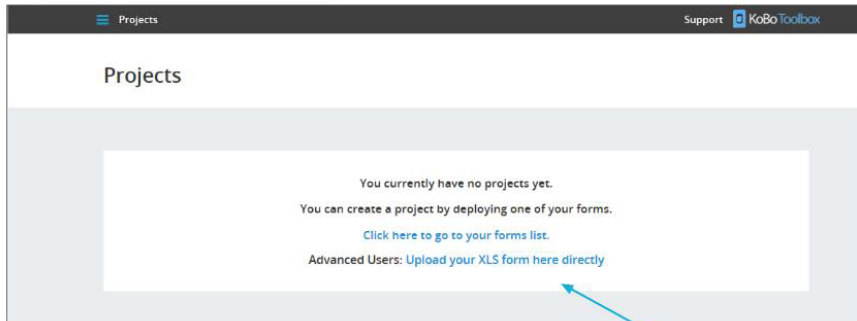
### Walkthrough: Uploading your forms

- Click PROJECTS. Here you will see all your uploaded and created projects.
- Click UPLOAD YOUR XLS FORM HERE DIRECTLY. You will be prompted to select the Excel file that you want to use.
- Once you have selected the file, it will automatically upload the form that you want to use.



# Uploading your form

To upload your Excel form, click Projects.



Once you arrive on this page, click **UPLOAD YOUR XLS FORM HERE DIRECTLY**.

If there are errors on your Excel form, you will receive an error message telling you what you need to revise on your Excel form. Otherwise, you will get the message Enter Web Form or Preview Web Form.

# Uploading your form

Once you have uploaded your form and there are no errors, you will receive this message.

Successfully published sample\_odk. [Enter Web Form](#) or [Preview Web Form](#) ✕

Projects	Active	Shared By	Date Created	Last Modified	Submissions
Sample ODK Form Sample ODK Form	✓		Aug. 26, 2015		0

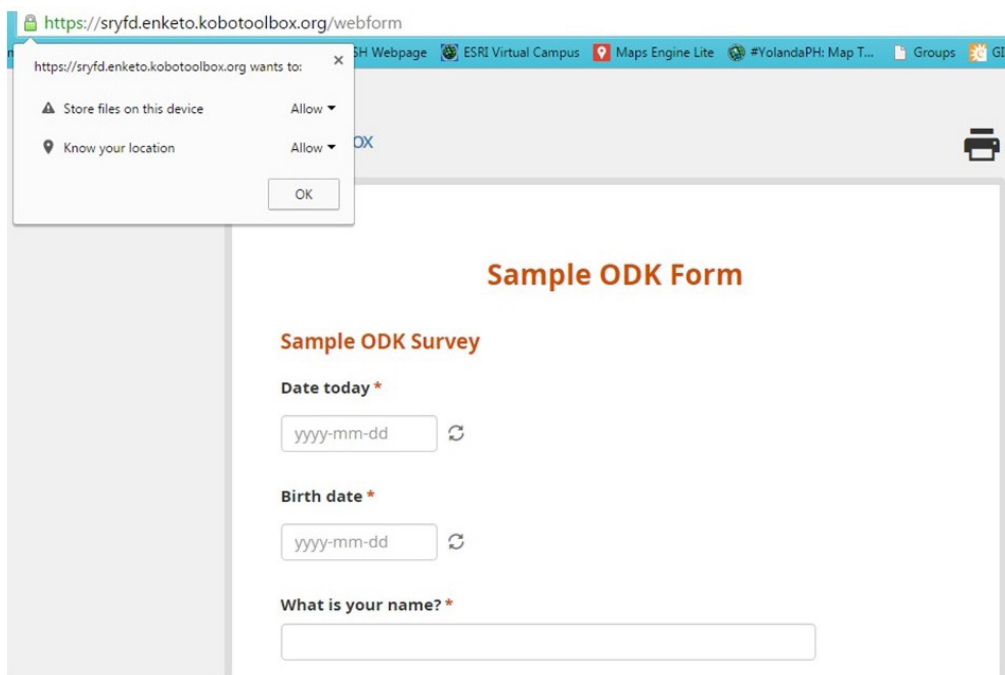
Advanced Users: [Upload your XLS form here directly](#)



However, if there are errors on your form, you will receive an error message telling you what the errors are.

**Testing the form:** once you have clicked Enter Web Form or Preview Web Form, you will see the online version of the form. Sometimes, the web form does not look exactly like the mobile version of the form.

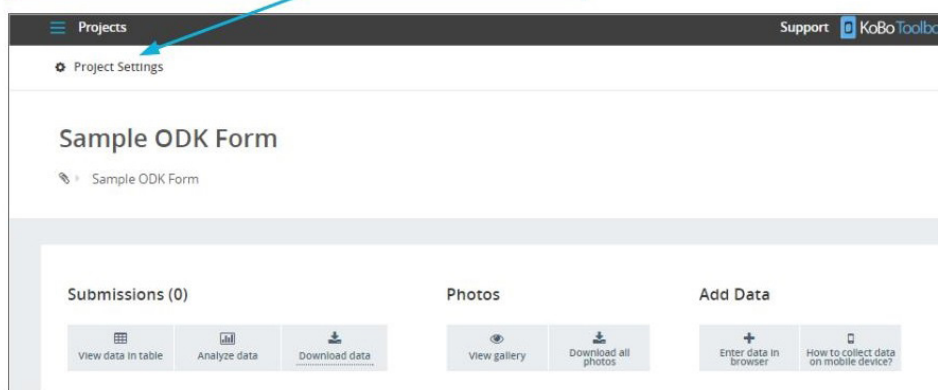
The web form will prompt you if you want to store files on your laptop and share your location. When you get errors on the form that you have selected, the form will not be uploaded.



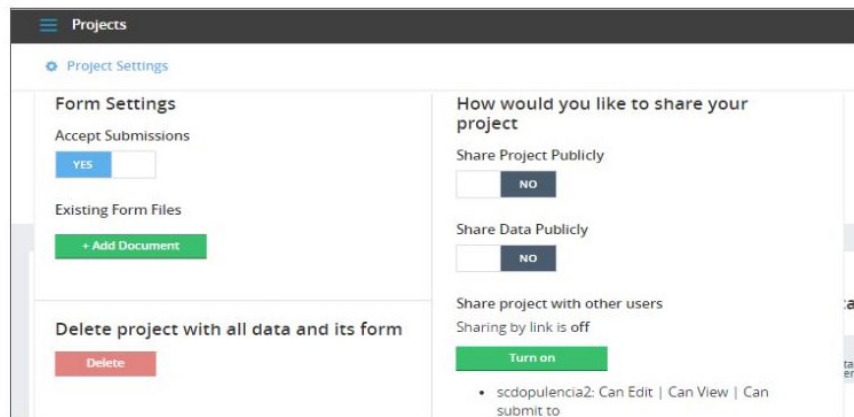
**Deleting the form:** sometimes, you will only find errors on your form when you have tested it. When you see these errors, you cannot directly change the form in the web version. You need to delete the project and re-upload the corrected Excel file.

To delete your form, click the project from the project list. When you click the project settings, you will see the option to DELETE the project.

Click PROJECT SETTINGS and you will be given the option to delete your project.



Once you have deleted your project, you can no longer retrieve it. You would need to upload it again once you've made revisions.



The screenshot shows the 'Project Settings' page in Kobo Toolbox. The page is divided into two main sections. The left section, titled 'Form Settings', includes a toggle for 'Accept Submissions' set to 'YES', a section for 'Existing Form Files' with a '+ Add Document' button, and a 'Delete project with all data and its form' section with a red 'Delete' button. The right section, titled 'How would you like to share your project', includes toggles for 'Share Project Publicly' and 'Share Data Publicly', both set to 'NO', and a 'Share project with other users' section where 'Sharing by link is off' and a 'Turn on' button is visible. Below this, a list of users is shown, including 'scdopulencia2' with permissions 'Can Edit | Can View | Can submit to'.

Be very careful in deleting your forms. When data has already been submitted through the form, it will also be deleted once you delete the project.

**Replacing the form:** to replace the form, you need to delete the form and follow the steps in Uploading the Form. The new form will appear on your projects page.

### Reminders and Tips

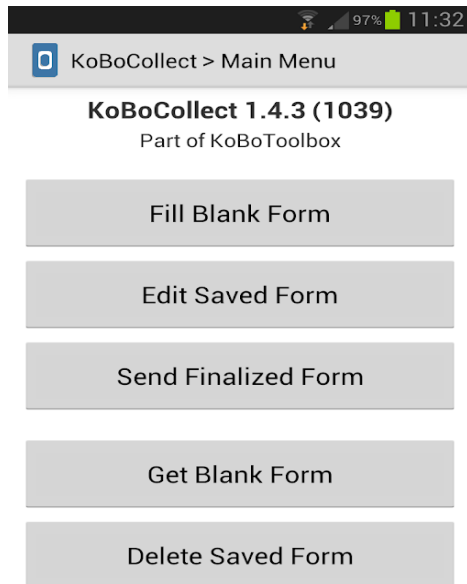
- Always save your Excel form in a different file name on your computer.
- Once you have uploaded a form, you cannot upload the same file using the same form\_id.
- Keep several versions of your Excel forms on your computer. You never know when a calculation or group questions may become useful.
- Kobo Toolbox will not allow you to revise uploaded Excel forms directly on the site. You need to delete the project, revise it, and upload it again.

## Using Kobo on an Android device

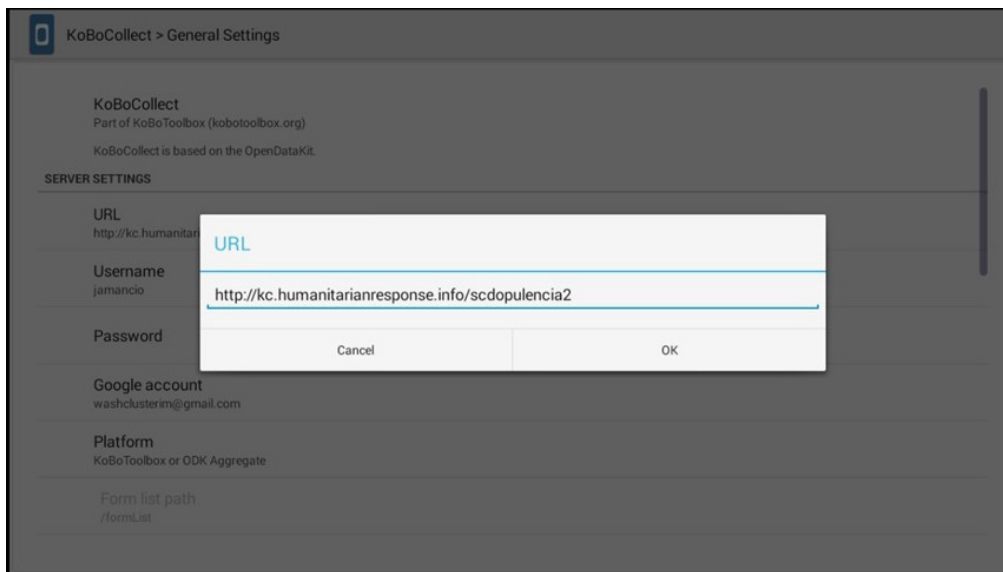
Kobo Collect can only be used on phones or tablets using the Android Operating System. It is not available in iPhones and Blackberry phones.

### Walkthrough: Installing and setting up Kobo on an Android device

- Using your phone, go to your applications and open the Play Store. Search for KOBO COLLECT.
- Install Kobo Collect on your phone.
- Once installed, open Kobo Collect and look for the General Settings (you need to press the three dots on the side of your screen or press the left button on the bottom of your phone).



- Once the app is installed and opened, you will be required to type a URL. Get your Kobo Toolbox account username and type <http://kc.humanitarianresponse.info/yourusername>.
- Type your username.
- Under Google account, type your Gmail account if you have one.
- Go back to the main page (where you will see the form options).



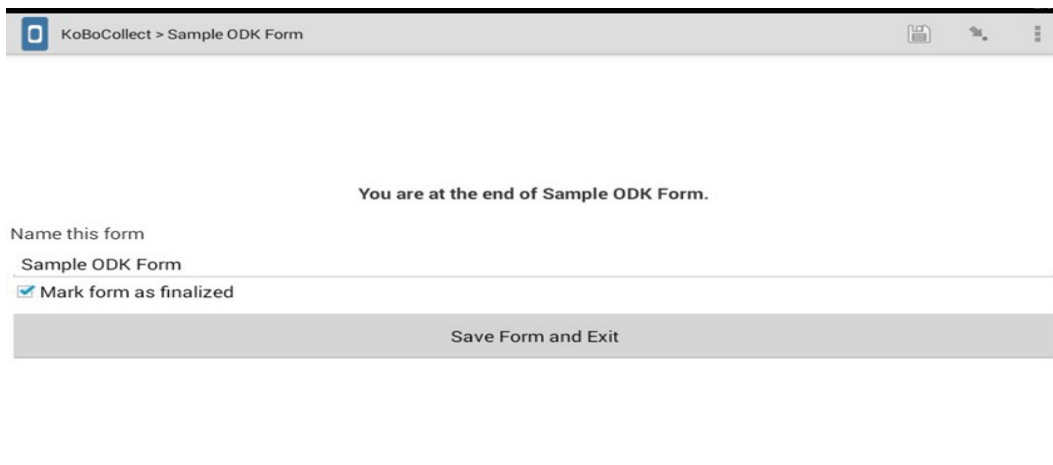
Now that we have set up the app properly, we're able to start testing the form function.

### Walkthrough: Testing your form

1. Press GET BLANK FORM. Press the check mark beside the form that you want to use, and press GET SELECTED.
2. To start collecting data, press FILL BLANK FORM and choose the form that you want to use.
3. Once you are done collecting the information, you will reach the last page, which says YOU ARE AT THE END THE FORM.
4. Change the NAME of the form if needed.
5. If you are not yet sure of your response, uncheck MARK FORM AS FINALIZED.

- Press SAVE FORM AND EXIT.
- For saved forms but not yet submitted, look at EDIT SAVED FORM and retrieve the form. Click GO TO START, review your responses until you reach the end. Repeat step 11 and make sure that MARK FORM AS FINALIZED has been checked.

- You will return to the main page. You will see your completed forms under SEND FINALIZED FORM. Once you have Internet access, check all your finalized forms and press SEND.



The screenshot shows the KoBoCollect interface for a 'Sample ODK Form'. At the top, a header bar contains the KoBoCollect logo and the text 'KoBoCollect > Sample ODK Form'. Below the header, a message states 'You are at the end of Sample ODK Form.' Underneath this message, there is a section titled 'Name this form' with a text input field containing 'Sample ODK Form'. Below the input field, there is a checkbox labeled 'Mark form as finalized' which is checked. At the bottom of the form, there is a large grey button labeled 'Save Form and Exit'.

**Tip:** If you do not have a smartphone but would like to use Kobo Collect in the comfort of your laptop or desktop, try to install Bluestacks (<http://www.bluestacks.com/>), an Android emulator for laptops and desktops (and Macs, too)!

# MODULE 4 – VERIFYING DATA

---

## Introduction

Verifying is a step that is often skipped in the process of working with data. Especially if the data was difficult to get, we are all the more eager to start on the analysis. Not including at least some basic verification will mean that, on many projects, you will end up wasting time on bad data or worse – you could publish incorrect results.

In this module we will talk about common misconceptions and pitfalls when people start analyzing and visualizing. Only if you know the common errors can you avoid making them in your own work and falling for them when they are mistakenly cited in the work of others.

## Description

The module will cover the following concepts:

- Data quality
- Data integrity
- Descriptive statistics
- Statistical fallacies

## Skills

As part of this module, you will learn the following:

- Using a pivot table
- Calculating average, median, minimum, maximum, and standard deviation
- Using conditional formatting
- The four methods for data verification
- The three aspects of quality data

## Prerequisites

- Module dataset: Global Corruption Barometer sample data: <http://bit.ly/GCBAAsiaSample>
- Basic knowledge of operating a computer
- Basic understanding of summary statistics including mean, median, mode, range, and standard deviation
- A fair understanding of all previous modules
- A working computer
- A modern internet browser
- Internet connection
- Spreadsheets (Google Sheets)



## Section 1: What is good data?

Verifying data means to check it against the criteria that define 'good' data. So, what is good data? Depending on what you're trying to do, your criteria might be more or less strict. But across all use cases, three criteria can help identify what is or what isn't good data: trustworthiness, completeness, and quality.

**Trustworthiness** refers to how well the data represent reality. Can we trust it to represent what it says it does? Or is it an unreliable account of the reality it tries to encode? Data generated by a low quality or faulty sensor may not be trustworthy. A dataset collected by a professional survey enumerator may be more trustworthy than one collected by an amateur. Whether it is sensors or humans behind the data, assessing the trustworthiness of the data requires an understanding of the methodology and choices behind the creation of the dataset.

**Completeness** refers to how well the data covers the reality it tries to represent. A dataset may be incomplete because of inconsistent data collection practices (for example many medical acts are not recorded properly due to the medical staff being too busy or focused on something else) or due to key elements being absent (a social survey dataset not recording gender will potentially deprive the data analysts of a real understanding of the social dynamics that the survey was trying to capture). Once again, understanding the methodology behind the data collection as well as the topic the dataset covers is essential to assess if the data is complete (enough).

**Quality** refers to how well the dataset is structured and documented. On one hand, a well structured dataset follows tidy data principles, as outlined in Section 3 of the first module:

When properly structuring tabular data, we want to distinguish:

- The types or categories of data points, with **one type of data point per column**. Each type of information is described across multiple observations.
- The individual observations, with **one observation per row**. Each observation is made of one or more types of information.

Of course, other principles apply for non-tabular data. But unlike other formats, like JSON, which is mostly manipulated by people who know what they're doing, tabular data is created and shared by novices and experts alike.

On the other hand, quality can also refer to the documentation of the data. Is there a data dictionary? Any metadata? Any document explaining the collection methodology? The more documented the data, the higher its quality. A properly documented dataset allows you to check for its completeness and trustworthiness, establishing the links between the three criteria that make for good data.

## Metadata, Data Dictionary, Data Inventory

A **data inventory** (or registry) is a document listing all the datasets owned by an organization. By extension, an open data inventory lists all the datasets made publicly available by that organization, such as this document by the Canadian government: <https://open.canada.ca/data/en/dataset/4ed351cf-95d8-4c10-97ac-6b3511f359b7>

A **data dictionary** or codebook is a document describing the meaning of all columns and values in a dataset. This is especially relevant for datasets that use abbreviated column headers and nonstandard values. Sometimes the dictionary may skip obvious elements, such as the data column, although it may include it, in order to describe the expected formatting of the values in the date column (e.g., DD-MM-YY).

The **metadata** is most simply described as ‘data about the data.’ It can include all relevant contextual information about the dataset, from author, to date of creation, to the expected format of the various values (text, numbers, etc.). This also includes the data dictionary, which is a subset of the metadata. The metadata can be stored alongside the data (for example, in a different tab of a spreadsheet) or shared alongside the dataset (ideally as a .json file, but most often as a .doc or .pdf file).

## Section 2: Four verification methods

The various steps that you may take to verify your data can be grouped into 4 types:

**Asking the source:** those who produced or published the data are most likely the best experts on it. Why not ask them directly? Too often new data practitioners see themselves engaged in a one-on-one battle with the data. This is not only inefficient, but it also blinds you to your potential misinterpretations.

**Asking experts:** data practitioners from the civic sector often get to work on datasets that pertain to many different topics. But on each of those topics there is probably an expert you can reach and who can contextualize the dataset for you or comment on your findings. This is an essential step to take for any serious data project.

Experts can also provide guidance when dealing with thorny data problems like: should you use the unemployment dataset published every three months by the Ministry of Work? Or the one published every six months by the National Statistics Office, which is also calculated differently? Employment experts will help guide your choice based on the objectives of your project.

**Statistical checks:** when exploring a large new dataset, diving in right away will probably result in confusion. Instead, you’ll want to make sure you understand what each column header means, what type of value to expect in the data, and if it fits what you have in mind.

A common approach is to create a statistical summary of your data – calculating the mean, median, maximum (max) and minimum (min) values and standard deviation for key columns should give you a good idea of what the data looks like. Are the min and max values within the expected range? Is the data skewed in a certain way or does it follow a normal distribution (or bell curve, with values clustered around the mean)? Spreadsheet applications do not generate a statistical summary automatically (though some other software does), which means that you’ll have to do it manually.

## The Statistical Summary

The statistical summary aims to give a quick and simple description of the data. It generally makes use of the following functions, which can be applied to all the columns in your data that are numerical:

The **mean (also called average)** is calculated by adding together all the values in the dataset before dividing the result by the number of values (or observations). The mean is influenced by extreme values: the mean of [0,1,5,2] is 2, but the mean of [0,1,5,2,300] is 61,6. When dealing with datasets where some outlier values are present, using the median is a better way to get a sense of the "middle point" of the dataset.

The **median** is the middle value, which is identified when you arrange your values in order. Said differently, half of the values will be inferior or equal to the median, while the other half will be superior or equal to it. This means that the median is not affected by outliers, as it relies on the number of observations. Calculating the difference between the mean and the median is a quick way to identify if your data includes extreme values or skews one way or another.

The **maximum** and **minimum** are, respectively, the highest and the lowest values among your observations.

The **standard deviation** gives you an idea of the variance in your data. It allows you to answer the question: are all the values in my dataset like each other, or do they vary a lot? The calculation of the standard deviation is more complex mathematically, but a dedicated formula exists in spreadsheet software. A common use of standard deviation is calculating outliers: in many datasets, if a value is superior to the mean + 2 times the standard deviation, then it is a possible outlier (extreme) value.

**Common sense check:** this is probably the most important of all. It represents the ability to identify weird patterns in the data, or a number which shouldn't be this high, or a value which shouldn't have been part of this dataset. This relies on a general sense for reading data, and most importantly, background knowledge about the topic the data is based on. Which is why data skills are not sufficient to work with any kind of data: the data needs to make sense to you, so that you won't miss important insights.

### Walkthrough: Creating a statistical summary of a dataset

We're going to work with a sample of the Global Corruption Barometer survey, which is administered every year by Transparency International: <http://bit.ly/GCBAAsiaSample>

- Open the spreadsheet linked and make a copy of it, in order to edit it, by going to File → Make a copy
- In your new spreadsheet, create a new tab by clicking on the + icon next to the 'sample dataset' tab. A new tab called 'sheet 2' (or something equivalent in your language) should appear.



- Switch to 'Sheet 2' by clicking on it.
- Because we want to apply our summary statistics to all columns, we're going to need to retrieve all the columns from the main sheet. To do so we're going to start by clicking on the cell A1 and write '='

to tell Google Sheets that we want to write a formula.

- Conveniently, the formula will follow us even if we switch to the first tab. So, let's do that: while the '=' is still active, click on the 'sample dataset' tab in order to switch to it.

	Sheet2!A1	B	C	D	E
1	<b>COUNTRY</b>	<b>GENDER</b>	<b>AGE</b>	<b>AREA</b>	<b>Q30: the govern run by a few big looking out for ti</b>

- The '=' sign should have followed you, as shown in the image above (if not, just restart from the cell A1 of Sheet 2).
- Now we'll select the whole header line by clicking on the '1' of the first row. Your formula should display.

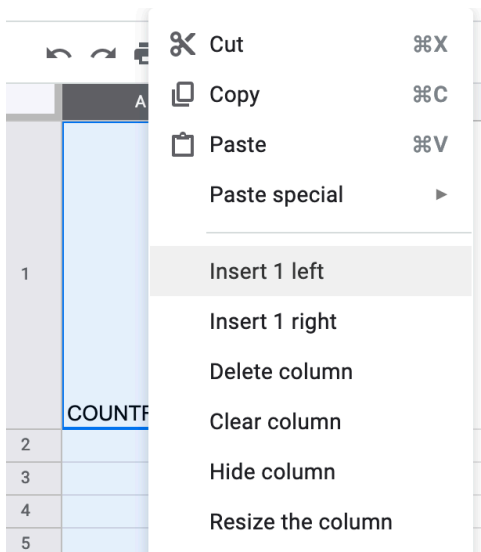
```
= 'Sample dataset' !1:1
```

(Here, 1:1 signifies that the whole row 1 is selected. The column equivalent would be A:A)

- You can now validate the formula with 'Enter.' You should be back to sheet 2, with the first line showing only the first header, 'COUNTRY'
- This is normal: by default, the reference function ('=') can only return one cell, even if a range of cells is selected. So, while the formula still shows that the whole header line was selected, the result contains only the first cell.
- The solution to this problem is to force the reference function to output a range. To do that we will modify the formula and surround it with curly brackets as seen below:

```
= { 'Sample dataset' !1:1 }
```

- Now we get the full header row! The text is not very readable, though, so let's make it so it wraps within the cell instead of being cut if it's too long. To do that, I will select the whole sheet by clicking on the gray cell at the top left corner of the sheet, between the 'A' and the '1.'
- Now go to the menu to select Format → Text wrapping → Wrap in order to have the text fully visible.
- With the header row in place, we can add the various calculations that we want to apply under each individual header. To help ourselves, we're going to add another column to the left of our header row, which will list the different types of stats we want to calculate.
- To add a column at the beginning of the sheet, right click on the 'A' of the first column and select 'Insert 1 left' as seen in the picture below.



- Our new column is now the new 'A' column. We're going to use it to indicate which line will contain which calculation. We will write:
  - MEAN in A2
  - MEDIAN in A3
  - MAX in A4
  - MIN in A5
  - MODE in A6 (mode gives us the value that appears the most frequently)
  - COUNTA in A7 (which counts of all values, whether text or numeric)
- We could add more, but that will be enough for our purpose. Now we just need to write the corresponding functions (**AVERAGE()**, **MEDIAN()** etc.) in column B and apply it to the corresponding column of the 'Sample dataset' tab.
- For column B, the corresponding column (the one with the same header) in the 'Sample dataset' tab is column A. Our first formula, in B2, is consequently:

```
=AVERAGE('Sample dataset'!A2:A)
```

Where:

- **AVERAGE()** is the formula that will give me the mean of the column I'm targeting.
- **'Sample dataset'!** Is the notation used to indicate which sheet the selected range is from when we're selecting cells in another sheet.
- **A2:A** is our selected range, which starts from the second row (A2) until the end of column A (:A).

		COUNTRY	GENDER	AGE
2	MEAN	<b>? =AVERAGE('Sample dataset'!A2:A)</b>		
3	MEDIAN			
4	MAX			
5	MIN			
6	MODE			

- After entering the formula, we get a surprising result: the error code '#DIV/0!'

COUNTRY	GENDER	AGE	ARE
#DIV/0!	<div> <b>Error</b>            Evaluation of function AVERAGE caused a divide by zero error.         </div>		

- Error results are important because they tell us what went wrong. Here the issue is the **AVERAGE()** function ended up dividing by 0, which can't be done. This is explained by the fact that the **COUNTRY** column does not have any numerical value! Functions which expect numerical values to produce a result (such as **AVERAGE()**, but also **MEDIAN()** or **MAX()**) treat text values as 0 if they encounter them in the range they are applied to.
- The error is consequently expected and should not appear when applied to columns with numerical values. Nonetheless it can be unpleasant to have an error constantly displayed on the spreadsheet, even if it is expected. To solve this, we will modify our formula and wrap our **AVERAGE()** function with another function called **IFERROR()**:

```
=IFERROR(AVERAGE('Sample dataset'!A2:A))
```

(Wrapping IFERROR around our formula with nothing else is the equivalent of saying if the formula produces an error, leave the cell blank).

- We can now apply the formula to all cells of the line, which will allow us to calculate the mean of all columns of our dataset, provided that the column contains numerical value. To quickly apply the formula to the line, put your cursor over the bottom right corner of the cell with the formula, click and drag it until the end of your range (column L).

	COUNTRY	GENDER	AGE
MEAN			
MEDIAN			
MAX			

- Given that only the column 'AGE' shows a value, we can deduce that it is the only column with numerical values.
- Now what is left is to repeat the previous steps for **MEDIAN**, **MAX**, **MIN**, **MODE**, **COUNTA**

```
=IFERROR(MEDIAN('Sample dataset'!A2:A))
```

```
=IFERROR(MAX('Sample dataset'!A2:A))
```

```
=IFERROR(MIN('Sample dataset'!A2:A))
```

```
=IFERROR(MODE('Sample dataset'!A2:A))
```

```
=IFERROR(COUNTA('Sample dataset'!A2:A))
```

- After dragging those cells until the column L, we are left with a simple summary table:

	COUNTRY	GENDER	AGE	AREA	Q30: the governments is run by a few big interests looking out for themselves	Q8A: Have you been in contact with schools in the past 12 months?	Q8B: how often, if ever, did you have to pay a bribe, give a gift, or do a favour for a teacher or school official in order to get the services you needed from the schools	Q9A: Have you been in contact with a clinic or private hospital in the path 12 months?	Q9B: how often, if ever, did you have to pay a bribe, give a gift, or do a favour for a health worker or clinic or hospital staff in order to get the medical care you needed	Q10A: have you tried to get an identity document like a birth certificate, driver's license, passport or voter's card, or a permit, from government in the past 12 months?	Q10B: how often, if ever, did you have to pay a bribe, give a gift, or do a favour for a government official in order to get the document you needed?
MEAN			40.68691824								
MEDIAN			39								
MAX	0	0	93	0	0	0	0	0	0	0	0
MIN	0	0	18	0	0	0	0	0	0	0	0
MODE			55								
COUNTA	15900	15900	15900	15900	15900	15900	15900	15900	15900	15900	15900

The zeros appear because **MIN** and **MAX** count text values as 0 instead of throwing an error, so **IFERROR()** does not apply

- Voila! From this table we can verify a number of things, such as:
  - There are no missing values anywhere in the dataset (the **COUNT** is the same for all columns).
  - **AGE** is the only column with numeric values.
  - The average age of respondents is 40.7 years.
  - The youngest respondent was 18 years old, which validates the fact that the questionnaire was only administered to people aged 18 or older.
  - The oldest respondent was 93 years old.
  - The age curve is not heavily skewed in any direction, as the mean age (40.7) is very close to the median age (39).

## Section 3: Common data issues

As you navigate the world of data, you will encounter several problems that are common to many datasets. Knowledge is half the battle, so we're providing you with an illustrative, but not exhaustive, list of those problems below:

- Missing values: is it a blank cell? A '0'? An 'NA'? Or something else?
- Missing values are replaced with a deceiving value such as 0 or a fallback date (1900, 1904, 1969, 1970 are common)
- Rows or values are duplicated
- Date formats are inconsistent
- Units are not specified
- Column names are ambiguous
- Provenance is not documented
- Suspicious values are present
- Data is too coarse (the measurement was not precise enough)
- Totals differ from published aggregates
- Spreadsheet has 65536 rows (maximum value in old Excel versions) or 255 columns (max value in old Numbers versions)

Keep an eye out for them!

This concludes the module focused on data verification. It is important to remember that not everything can be caught in this step. You will have to stay alert during the cleaning and analysis phase in case something strange appears in the data. It is very common for data projects to circle between verification and cleaning a few times until every suspicion has been rooted out.



# MODULE 5 – CLEANING DATA

## Introduction

There is a common saying among people who work regularly with data that “80 percent of the time spent working with data is spent cleaning it.” And indeed, it is often an arduous task that is not only time consuming but can also be error prone. The CLEAN step of the Data Pipeline works as a group with the VERIFY and ANALYZE steps. While simple datasets will have you go through VERIFY, CLEAN and ANALYZE linearly, more complex ones will see you go back and forth between the three. Specifically, you may find yourself

- Cleaning the dataset to be able to verify its content.
- Going back to the verify step after finding something strange during the analysis.
- Doing some basic analysis steps as part of the verification process.



Data cleaning can be broken down into three different activities that are done one after the other:

- **Data tidying**, which is the process of cleaning the structure of the data, without touching its content.
- **Data editing**, which is the process of modifying the data content to prepare it for your analysis.
- **Data consolidation**, which is the process of adding complementary data to your main dataset(s).

## Description

The module will cover the following concepts:

- Data cleaning
- Data tidying
- Data editing
- Data consolidation

## Skills

As part of this module, you will learn the following:

- How to clean a dataset to prepare it for analysis
- Common data cleaning issues
- Common spreadsheet formulas for data cleaning

## Prerequisites

- Module dataset: <http://bit.ly/DataCleaningDatasetDLC>
- A fair understanding of all previous modules
- Spreadsheets (Google Sheets)
- A working computer

- A modern internet browser
- Internet connection
- Basic knowledge of operating a computer

## Section 1: Data tidying

The goal of data cleaning is to create a consistent, human-understandable, and machine-readable dataset and prepare it for the specific analysis that you want to run on it. In the process of cleaning your data, you will encounter two main types of problems:

**Formatting problems:** these are problems related to how a dataset is written or stored. This includes having multi-line or merged headers, having a single piece of required information stored in multiple columns of a spreadsheet, or having multiple data points or types of data within single cells.

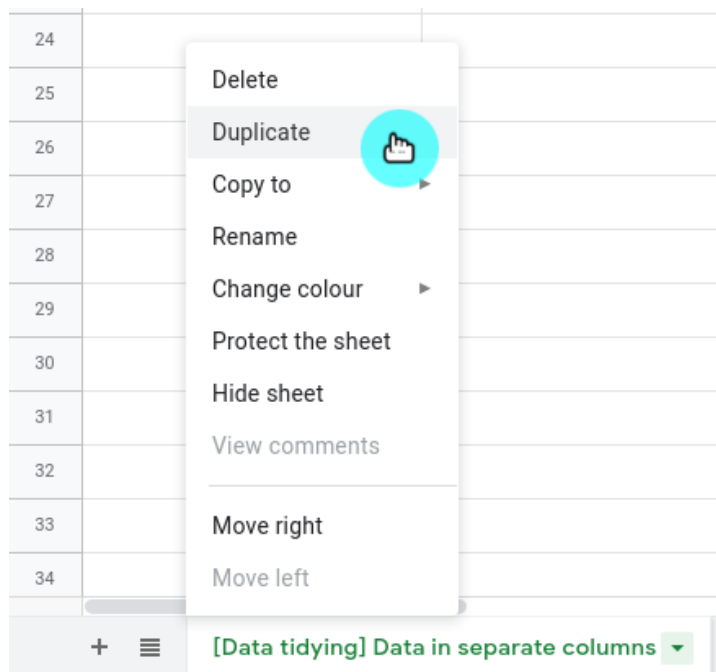
**Content problems:** these are problems related to what is written or stored in the dataset. These include mistakes in the case used, spelling, or the actual values of the data.

The data tidying step involves resolving the formatting problems before we proceed with cleaning the content problems in the data editing step.

### Walkthrough: Cleaning formatting problems

We're going to go resolve several formatting problems in the dataset here: <http://bit.ly/DataCleaningDatasetDLC>

- Open the spreadsheet linked and make a copy of it to be able to edit it by going to File → Make a copy
- For each of the examples, create a duplicate of the tab you will be cleaning by clicking on the ▼ symbol on the right of the tab → Duplicate




- This will create a duplicate tab to the right of the current tab. Rename the tab as you see fit.

### What to do when the data you need is stored in separate columns.

- Sometimes, the information we need is stored in separate columns. One example is when we need a list of full names but in the dataset that we are given a column for the last name, first name, and middle initial.

	A	B	C
1	<b>Last Name</b>	<b>First Name</b>	<b>Middle Initial</b>
2	Smith	Alex	A.
3	Jones	Samuel	S.
4	Jackson	Jason	C.
5	Doe	John	J.
6	James	Peter	S.

- Add a new column named 'Full Name'.

D1		Full Name		
	A	B	C	D
1	<b>Last Name</b>	<b>First Name</b>	<b>Middle Initial</b>	<b>Full Name</b>
2	Smith	Alex	A.	
3	Jones	Samuel	S.	
4	Jackson	Jason	C.	
5	Doe	John	J.	
6	James	Peter	S.	

- Next, we'll use the CONCATENATE function to join multiple strings together. Note that Google Sheets also has the CONCAT function which joins two strings together.

### String vs. Numeric

When analyzing the content of your data, spreadsheet and database software will try to classify it into several types, in order to know which formula can be applied or not. This is why it will show an error if you try to apply an AVERAGE() to a column of names. The four most common data types are:

Numeric, which refers to numbers that you can use in a calculation. The weight of a package would be a numeric data type, but its order number would not.

Boolean, which refers to data points expressed in 0 and 1. 0 is generally interpreted as 'no' and 1 as 'yes.'

Date, which covers data formatted as a date and allows date-specific operations such as counting the number of days between two values.

String, which covers text and everything which does not fall into the previous categories.

- Go to the D2 cell and type the following formula:

```
=CONCATENATE(B2," ",C2," ",A2)
```

Notice how we also appended spaces (" ") in between the cell values.

D2	fx	=CONCATENATE(B2, " ", C2, " ", A2)		
	A	B	C	D
1	Last Name	First Name	Middle Initial	Alex A. Smith x
2	Smith	Alex	A.	=CONCATENATE(B2, " ", C2, " ", A2)
3	Jones	Samuel	S.	
4	Jackson	Jason	C.	
5	Doe	John	J.	
6	James	Peter	S.	

D2	fx	=CONCATENATE(B2, " ", C2, " ", A2)		
	A	B	C	D
1	Last Name	First Name	Middle Initial	Full Name
2	Smith	Alex	A.	Alex A. Smith
3	Jones	Samuel	S.	
4	Jackson	Jason	C.	
5	Doe	John	J.	
6	James	Peter	S.	

- If the formula ran successfully without any errors, we can then apply it to the remaining cells by dragging like we did in the last module, or simply by copy-pasting along the range.

D6	<div><div>fx</div><div>=CONCATENATE(B6," ",C6," ",A6)</div></div>			
	A	B	C	D
1	Last Name	First Name	Middle Initial	Full Name
2	Smith	Alex	A.	Alex A. Smith
3	Jones	Samuel	S.	Samuel S. Jones
4	Jackson	Jason	C.	Jason C. Jackson
5	Doe	John	J.	John J. Doe
6	James	Peter	S.	Peter S. James

### What to do when you need multiple data that is stored in a single column.

- The inverse of the first problem is when you have multiple information stored in a single column. One example is when you need to do your analysis on a per-city or per-country level but are given a dataset with an address/area column that contains both.

	A	B
1	<b>Area</b>	<b>2021 Population</b>
2	Tokyo, Japan	37,339,804
3	Delhi, India	31,181,376
4	Shanghai, China	27,795,702
5	Sao Paulo, Brazil	22,237,472
6	Mexico City, Mexico	21,918,936
7	Dhaka, Bangladesh	21,741,090
8	Cairo, Egypt	21,322,750
9	Beijing, China	20,896,820
10	Mumbai, India	20,667,656
11	Osaka, Japan	19,110,616


- Insert two columns to the right of the Area column and name them City and Country. NOTE: If we had three values (e.g., City, State, Country) in the Area column and you wanted to extract all three, you would need to add three columns instead of two.

B1	fx	City		
	A	B	C	D
1	<b>Area</b>	<b>City</b>	<b>Country</b>	<b>2021 Population</b>
2	Tokyo, Japan			37,339,804
3	Delhi, India			31,181,376
4	Shanghai, China			27,795,702
5	Sao Paulo, Brazil			22,237,472
6	Mexico City, Mexico			21,918,936
7	Dhaka, Bangladesh			21,741,090
8	Cairo, Egypt			21,322,750
9	Beijing, China			20,896,820
10	Mumbai, India			20,667,656
11	Osaka, Japan			19,110,616

- Copy the column to be split (Area) into the city column.


B2:B21	fx	Tokyo, Japan		
	A	B	C	D
1	<b>Area</b>	<b>City</b>	<b>Country</b>	<b>2021 Population</b>
2	Tokyo, Japan	Tokyo, Japan		37,339,804
3	Delhi, India	Delhi, India		31,181,376
4	Shanghai, China	Shanghai, China		27,795,702
5	Sao Paulo, Brazil	Sao Paulo, Brazil		22,237,472
6	Mexico City, Mexico	Mexico City, Mexico		21,918,936
7	Dhaka, Bangladesh	Dhaka, Bangladesh		21,741,090
8	Cairo, Egypt	Cairo, Egypt		21,322,750
9	Beijing, China	Beijing, China		20,896,820
10	Mumbai, India	Mumbai, India		20,667,656
11	Osaka, Japan	Osaka, Japan		19,110,616

- Select the cells to be split and go to Data → Split text to columns.

 Data Cleaning Dataset ☆ 📁 ☁

File Edit View Insert Format **Data** Tools Add-ons Help [Last edit was 2 minutes ago](#)

100% ▾ £ % .0

B2:B21  Tokyo, Japan

	B		D
1	<b>City</b>	<b>Coun</b>	
2	Tokyo, Japan		<b>2021 Population</b>
3	Delhi, India		37,339,804
4	Shanghai, China		31,181,376
5	Sao Paulo, Brazil		27,795,702
6	Mexico City, Mexico		22,237,472
7	Dhaka, Bangladesh		21,918,936
8	Cairo, Egypt		21,741,090
9	Beijing, China		21,322,750
10	Mumbai, India		20,896,820
11	Osaka, Japan		20,667,656
12	Karachi, Pakistan		19,110,616
13	Chongqing, China		16,459,472
14	Istanbul, Turkey		16,382,376
15	Buenos Aires, Argentina		15,415,197
16	Kolkata, India		15,257,673
17	Kinshasa, Dr Congo		14,974,073
18	Lagos, Nigeria		14,970,460
19	Manila, Philippines		14,862,111
20	Tianjin, China		14,158,573
21	Guangzhou, China		13,794,450
22			13,635,397

Sort sheet by **column B**, A → Z

Sort sheet by **column B**, Z → A

Sort range by **column B**, A → Z

Sort range by **column B**, Z → A

Sort range

▼ Create a filter

Filter views ▶

≡ Slicer

Data validation

Pivot table


Randomise range

Named ranges

Protected sheets and ranges

Clean-up suggestions

Column stats

Split text to columns 

Remove duplicates

Trim whitespace

Group Alt+Shift+→

Ungroup Alt+Shift+←

- The columns should automatically be split using the comma as the delimiter. You can also specify the delimiter used.

B2:C21	fx	Tokyo		
	A	B	C	D
1	Area	City	Country	2021 Population
2	Tokyo, Japan	Tokyo	Japan	37,339,804
3	Delhi, India	Delhi	India	31,181,376
4	Shanghai, China	Shanghai	China	27,795,702
5	Sao Paulo, Brazil	Sao Paulo	Brazil	22,237,472
6	Mexico City, Mexico	Mexico City	Mexico	21,918,936
7	Dhaka, Bangladesh	Dhaka	Bangladesh	21,741,090
8	Cairo, Egypt	Cairo	Egypt	21,322,750
9	Beijing, China	Beijing	China	20,896,820
10	Mumbai, India	Mumbai	India	20,667,656
11	Osaka, Japan	Osaka	Japan	19,110,616
12	Karachi, Pakistan	Karachi	Pakistan	16,459,472
13	Chongqing, China	Chongqing	China	16,382,376
14	Istanbul, Turkey	Istanbul	Turkey	15,415,197
15	Buenos Aires, Argentina	Buenos Aires	Argentina	15,257,673
16	Kolkata, India	Kolkata	India	14,974,073
17	Kinshasa, Dr Congo	Kinshasa	Dr Congo	14,970,460
18	Lagos, Nigeria	Lagos	Nigeria	14,862,111
19	Manila, Philippines	Manila	Philippines	14,158,573
20	Tianjin, China	Tianjin	China	13,794,450
21	Guangzhou, China	Guangzhou	China	13,635,397
22			Separator: Detect automatically	
23				

### What to do when you have multi-line or merged headers.

- Having multiple header rows is common and makes data with multiple categories easy to understand. However, when analyzing data, spreadsheets usually prefer and expect the headers to be a single cell. Take the demographic data below for example (population in thousands):

	A	B	C	D	E	F	G	H	I
1	Location	Population							
2		0-4		5-9		10-14		15-19	
3		Male	Female	Male	Female	Male	Female	Male	Female
4	Africa	101,498	98,412	92,042	89,450	80,721	78,707	69,850	68,415
5	Asia	188,596	173,546	190,317	174,296	189,942	172,935	186,712	169,485
6	Europe	20,061	19,009	20,936	19,815	20,679	19,571	19,460	18,435
7	Latin America and the Caribbean	26,420	25,269	26,644	25,553	26,681	25,684	27,198	26,348
8	North America	11,079	10,597	11,262	10,768	11,796	11,284	11,856	11,383
9	Oceania	1,778	1,676	1,726	1,630	1,680	1,588	1,567	1,487

- We can simplify the headers so that they only utilize a single row.

	A	B	C	D	E	F
1	Location	Male (0-4 years old)	Female (0-4 years old)	Male (5-9 years old)	Female (5-9 years old)	Male (10-14 years old)
2	Africa	101,498	98,412	92,042	89,450	80,721
3	Asia	188,596	173,546	190,317	174,296	189,942
4	Europe	20,061	19,009	20,936	19,815	20,679
5	Latin America and the Caribbean	26,420	25,269	26,644	25,553	26,681
6	North America	11,079	10,597	11,262	10,768	11,796
7	Oceania	1,778	1,676	1,726	1,630	1,680



- Depending on your need, you can also replace the layout of the data.

	A	B	C	D	E	F	G	H
1	Location	Sex	0-4 years old	5-9 years old	10-14 years old	15-19 years old	20-24 years old	25-29 years old
2	Africa	Female	98,412	89,450	78,707	68,415	59,582	52,193
3	Africa	Male	101,498	92,042	80,721	69,850	60,450	52,484
4	Asia	Female	173,546	174,296	172,935	169,485	170,130	173,302
5	Asia	Male	188,596	190,317	189,942	186,712	186,539	188,681
6	Europe	Female	19,009	19,815	19,571	18,435	18,999	22,022
7	Europe	Male	20,061	20,936	20,679	19,460	19,979	23,018
8	Latin America and the Caribbean	Female	25,269	25,553	25,684	26,348	26,807	26,561
9	Latin America and the Caribbean	Male	26,420	26,644	26,681	27,198	27,230	26,751
10	North America	Female	10,597	10,768	11,284	11,383	12,095	13,039
11	North America	Male	11,079	11,262	11,796	11,856	12,535	13,538
12	Oceania	Female	1,676	1,630	1,588	1,487	1,487	1,516
13	Oceania	Male	1,778	1,726	1,680	1,567	1,555	1,587

At the end of the data tidying step, you should have a properly formatted dataset that is tailored to your needs.

## Section 2: Data editing

If the data tidying step deals with formatting problems, the data editing step is where we resolve content problems such as mistakes in spelling, values, etc.

### Walkthrough: Cleaning content problems

- Like the walkthrough for cleaning formatting problems, always duplicate the tab first before doing the actual cleaning.

### What to do when there are extra white spaces in the values.

- A common content problem that tends to be overlooked because it is hard to notice is the presence of extra white spaces in the data. This could take the form of spaces that appear before, after, or in between characters.

	A	B	C	D
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	Africa	1,182,439,000	1,340,598,000	2.51
3	aSia	4,433,475,000		0.92
4	EuropE	743,059,000	747,636,000	0.12
5	Latin America and the Caribbean	623,934,000	653,962,000	9.40
6	Northern American	357,031,000	368,870,000	0.65
7	OCeaniA	39,859,000	42,678,000	1.37
8				
9				
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))			
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>			
12				

- In the example above, how many extra spaces do you see? 1? 2? The answer is actually 3. There is an extra space before Africa, an extra space between Northern and American, and an extra space after Caribbean.
- To remove these extra spaces, we will use the TRIM function in Google Sheets.

- Add a new column to the right of the Geographic regions and use the following formula on cell B2:

=TRIM(A2)

B2	fx	=TRIM(A2)			
	A	B	C	D	E
1	<b>Geographic regions</b>	Africa x	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	Africa	=TRIM(A2)	1,182,439,000	1,340,598,000	2.51
3	aSia		4,433,475,000		0.92
4	EuropE		743,059,000	747,636,000	0.12
5	Latin America and the Caribbean		623,934,000	653,962,000	9.40
6	Northern American		357,031,000	368,870,000	0.65
7	OCeaniA		39,859,000	42,678,000	1.37
8					
9					
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))				
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>				

- If the formula ran successfully without any errors, we can then apply it to the remaining cells by dragging like we did in the last module or simply by copy-pasting along the range.

B2	fx	=TRIM(A2)			
	A	B	C	D	E
1	<b>Geographic regions</b>		<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	Africa	Africa	1,182,439,000	1,340,598,000	2.51
3	aSia		4,433,475,000		0.92
4	EuropE		743,059,000	747,636,000	0.12
5	Latin America and the Caribbean		623,934,000	653,962,000	9.40
6	Northern American		357,031,000	368,870,000	0.65
7	OCeaniA		39,859,000	42,678,000	1.37
8					
9					
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))				
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>				

B7	fx	=TRIM(A7)			
	A	B	C	D	E
1	<b>Geographic regions</b>		<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	Africa	Africa	1,182,439,000	1,340,598,000	2.51
3	aSia	aSia	4,433,475,000		0.92
4	EuropE	EuropE	743,059,000	747,636,000	0.12
5	Latin America and the Caribbean	Latin America and the Caribbean	623,934,000	653,962,000	9.40
6	Northern American	Northern American	357,031,000	368,870,000	0.65
7	OCeaniA	OCeaniA	39,859,000	42,678,000	1.37
8					
9					
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))				
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>				

- **TIP:** If you don't to keep the original column and just overwrite the wrong values, you can COPY (CTRL+C) the cleaned values (B2:B7) and perform a PASTE VALUES ONLY (CTRL+SHIFT+V) on the original values (A2:A7). Paste Values Only pastes the values shown by the cells and not the underlying formula.

- After pasting the cleaned values on the original column, you can now safely delete the extra column. Your dataset should now look like the below:

	A	B	C	D
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	Africa	1,182,439,000	1,340,598,000	2.51
3	aSia	4,433,475,000		0.92
4	EuropE	743,059,000	747,636,000	0.12
5	Latin America and the Caribbean	623,934,000	653,962,000	9.40
6	Northern American	357,031,000	368,870,000	0.65
7	OCeaniA	39,859,000	42,678,000	1.37
8				
9				
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))			
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>			

- Because there are several ways to skin an apple, you can also use the Trim white spaces function of Google Sheets directly. Highlight the cells that you want to trim the whitespaces and go to Data → Trim whitespace.

The screenshot shows the Google Sheets interface for a spreadsheet titled "Data Cleaning Dataset". The "Data" menu is open, and the "Trim whitespace" option is highlighted with a blue circle and a hand icon. The spreadsheet data is visible in the background, showing columns A, B, C, and D. Column A contains geographic regions, column B contains population data for 2015, column C contains population data for 2020, and column D contains growth rates. The "Trim whitespace" option is located at the bottom of the "Data" menu, below "Remove duplicates" and above "Group".

## What to do when the case used isn't consistent.

- We've now removed the extra white spaces in our data, but the Geographic regions column still looks wrong. As previously mentioned, the goal of cleaning is to create a consistent dataset. This usually means ensuring that the texts and strings in the data also use a consistent case. Depending on your needs, this might mean using UPPERCASE, lowercase, or Proper Case. Thankfully, switching between the three is relatively straightforward in Google Sheets using the UPPER, LOWER, and PROPER functions.
- In our case, let's say we want to use all UPPERCASE characters for our Geographic regions.

	A	B	C	D
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	Africa	1,182,439,000	1,340,598,000	2.51
3	aSia	4,433,475,000		0.92
4	EuropE	743,059,000	747,636,000	0.12
5	Latin America and the Caribbean	623,934,000	653,962,000	9.40
6	Northern American	357,031,000	368,870,000	0.65
7	OCeaniA	39,859,000	42,678,000	1.37
8				
9				
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))			
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>			

- To make all characters UPPERCASE, we will use the UPPER function in Google Sheets.
- Add a new column to the right of the Geographic regions and use the following formula on cell B2:

**=UPPER(A2)**

B2					
	A	B	C	D	E
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>	
2	Africa	=UPPER(A2)	1,182,439,000	1,340,598,000	2.51
3	aSia		4,433,475,000		0.92
4	EuropE		743,059,000	747,636,000	0.12
5	Latin America and the Caribbean		623,934,000	653,962,000	9.40
6	Northern American		357,031,000	368,870,000	0.65
7	OCeaniA		39,859,000	42,678,000	1.37
8					
9					
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))				
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>				

- If the formula ran successfully without any errors, we can then apply it to the remaining cells by dragging like we did in the last module, or simply by copy-pasting along the range.

B2		$\text{fx}$	$\text{=UPPER(A2)}$		
	A	B	C	D	E
1	<b>Geographic regions</b>		<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	Africa	AFRICA	1,182,439,000	1,340,598,000	2.51
3	aSia		4,433,475,000		0.92
4	EuropE		743,059,000	747,636,000	0.12
5	Latin America and the Caribbean		623,934,000	653,962,000	9.40
6	Northern American		357,031,000	368,870,000	0.65
7	OCeaniA		39,859,000	42,678,000	1.37
8					
9					
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))				
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>				

B7		$\text{fx}$	$\text{=UPPER(A7)}$		
	A	B	C	D	E
1	<b>Geographic regions</b>		<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	Africa	AFRICA	1,182,439,000	1,340,598,000	2.51
3	aSia	ASIA	4,433,475,000		0.92
4	EuropE	EUROPE	743,059,000	747,636,000	0.12
5	Latin America and the Caribbean	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	9.40
6	Northern American	NORTHERN AMERICAN	357,031,000	368,870,000	0.65
7	OCeaniA	OCEANIA	39,859,000	42,678,000	1.37
8					
9					
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))				
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>				

- After you paste the cleaned values in the original column, your dataset should now look like the below:

A2		$\text{fx}$	AFRICA	
	A	B	C	D
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	AFRICA	1,182,439,000	1,340,598,000	2.51
3	ASIA	4,433,475,000		0.92
4	EUROPE	743,059,000	747,636,000	0.12
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	9.40
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65
7	OCEANIA	39,859,000	42,678,000	1.37
8				
9				
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))			
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>			

### What to do when there are mistakes in spelling.

- Sometimes mistakes in spelling are present in your dataset. Finding these spelling mistakes requires knowledge of the dataset. The verify step mentioned in the previous module will also help identify and correct these spelling mistakes.
- For smaller datasets, you can simply replace the spelling mistake with the correct spelling.

- In the example below, ANERICAN should be AMERICAN, so we can simply replace that.

	A	B	C	D
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	AFRICA	1,182,439,000	1,340,598,000	2.51
3	ASIA	4,433,475,000		0.92
4	EUROPE	743,059,000	747,636,000	0.12
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	9.40
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65
7	OCEANIA	39,859,000	42,678,000	1.37
8				
9				
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))			
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>			
12				

	A	B	C	D
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	AFRICA	1,182,439,000	1,340,598,000	2.51
3	ASIA	4,433,475,000		0.92
4	EUROPE	743,059,000	747,636,000	0.12
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	9.40
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65
7	OCEANIA	39,859,000	42,678,000	1.37
8				
9				
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))			
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>			
12				

- For larger datasets, you can use the Find and replace (CTRL+H) function of Google Sheets.

A6	fx	NORTHERN AMERICAN				
	A	B	C	D	E	F
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>		
2	AFRICA	1,182,439,000	1,340,598,000	2.51		
3	ASIA	4,433,475,000		0.92		
4	EUROPE	743,059,000	747,636,000	0.12		
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	9.40		
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65		
7	OCEANIA	39,859,000	42,678,000	1.37		
8						
9						
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Po">https://population.un.org/wpp/Download/Standard/Po</a>					
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>					
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						

Find and replace

×

Find

Replace with

Search

☒ Match case
 ☐ Match entire cell contents
 ☐ Search using regular expressions [Help](#)
☐ Also search within formulae

- You can specify the following parameters of the find and replace:
  - Which sheets to perform the function
  - Matching the case of the variable
  - Matching the entire cell contents
  - Searching using regular expressions
  - Searching within formulas

### What to do when there are blank or incorrect values.


- Like spelling mistakes, cleaning blank and incorrect values requires knowledge of the dataset or taking advantage of the different verification methods discussed in the previous module.
- In our example below, there's a blank value in the 2020 column. There's also an incorrect value in the Growth rate column. Would you happen to know what that is?
- If you look at the Growth rates, there's a value that does not make sense – both in the common and mathematical sense – the 9.40 for LATIN AMERICA AND THE CARRIBEAN. If you look at the growth of the population from 2015 to 2020, it does not appear to be even close to 9.40 percent per year.

	A	B	C	D
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	AFRICA	1,182,439,000	1,340,598,000	2.51
3	ASIA	4,433,475,000		0.92
4	EUROPE	743,059,000	747,636,000	0.12
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	9.40
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65
7	OCEANIA	39,859,000	42,678,000	1.37
8				
9				
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))			
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>			
12				

- To solve these two problems, we can use the verification methods we learned in the previous module. One way is to consult the source. Fortunately, the source of the data is included in the dataset. (TIP: Metadata matters. Add metadata and data sources to your dataset).
- If you visit the data source <https://population.un.org/wpp/DataQuery/> and perform a query to recreate the data, you'll find:

population.un.org/wpp/DataQuery/

Welcome to the United Nations



**United Nations**

Department of Economic and Social Affairs  
Population Dynamics

World Population Prospects 2019

WPP Home Data Figures Documentation World Urbanization Prospects Population Division Contact Us

### Data Query

Total Population by sex (thousands)

Data Notes

Location	Sex	2015	2020
World			
Geographic regions	Both sexes combined		
Africa	Both sexes combined	1,182,439	1,340,598
Asia	Both sexes combined	4,433,475	4,641,055
Europe	Both sexes combined	743,059	747,636
Latin America and the Caribbean	Both sexes combined	623,934	653,962
Northern America	Both sexes combined	357,031	368,870
Oceania	Both sexes combined	39,859	42,678

You may sort and filter the data by clicking the arrow in the column heading. To return to the original data, just remove the sort and filter settings.


© 2019 by United Nations, made available under a Creative Commons license CC BY 3.0 IGO: <http://creativecommons.org/licenses/by/3.0/igo/>  
Citation: United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects 2019, custom data acquired via website.

<< Start Over < Back Export to Excel



population.un.org/wpp/DataQuery/

Welcome to the United Nations

 **United Nations** | Department of Economic and Social Affairs  
Population Dynamics

World Population Prospects 2019

WPP Home Data Figures Documentation World Urbanization Prospects Population Division Contact Us

### Data Query

#### Average annual rate of population change (percentage)

Data Notes

Location	2015 - 2020
World	
Geographic regions	
Africa	2.51
Asia	0.92
Europe	0.12
Latin America and the Caribbean	0.94
Northern America	0.65
Oceania	1.37

You may sort and filter the data by clicking the arrow in the column heading. To return to the original data, just remove the sort and filter settings.

© 2019 by United Nations, made available under a Creative Commons license CC BY 3.0 IGO: <http://creativecommons.org/licenses/by/3.0/igo/>  
**Citation:** United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects 2019, custom data acquired via website.

<< Start Over < Back Export to Excel

- We find that the population value for ASIA in 2020 is 4,641,055,000 (note that the values from the UN site are in thousands) and the annual growth rate for LATIN AMERICA AND THE CARRIBEAN is 0.94 (we can also compute this directly from the data if we know the formula used for computing the growth rate).

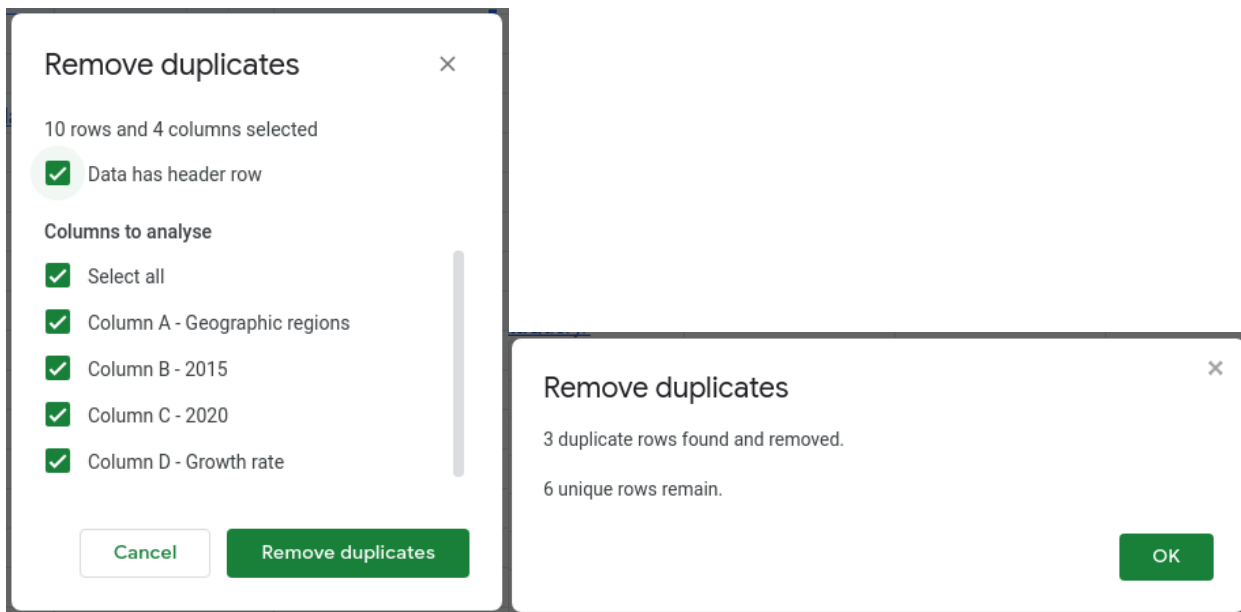
#### How to remove duplicate values, rows, columns.

- There are several methods to de-duplicate (dedup) or remove duplicates from a dataset. Google Sheets has a simple tool found under Data → Remove Duplicates.

	A	B	C	D
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	AFRICA	1,182,439,000	1,340,598,000	2.51
3	ASIA	4,433,475,000		0.92
4	EUROPE	743,059,000	747,636,000	0.12
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	0.94
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65
7	OCEANIA	39,859,000	42,678,000	1.37
8	AFRICA	1,182,439,000	1,340,598,000	2.51
9	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	0.94
10	NORTHERN AMERICAN	357,031,000	368,870,000	0.65
11				
12				
13	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))			
14	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>			
15				

- Highlight/select the cells (you can include the headers) from where you want to remove the duplicates. Then go to Data → Remove Duplicates.

The screenshot shows the Google Sheets interface for a spreadsheet titled "Data Cleaning Dataset". The "Data" menu is open, and the "Remove duplicates" option is highlighted with a hand icon. The spreadsheet data is visible in the background, showing columns A, B, C, and D. Column A contains "Geographic regions", column B contains "2015", column C contains "2020", and column D contains "Growth rate". The data rows show population figures for various regions, with some rows containing duplicate entries for the same region.



	A	B	C	D
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	AFRICA	1,182,439,000	1,340,598,000	2.51
3	ASIA	4,433,475,000		0.92
4	EUROPE	743,059,000	747,636,000	0.12
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	0.94
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65
7	OCEANIA	39,859,000	42,678,000	1.37
8				
9				
10				
11				
12				
13	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))			
14	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>			

## Section 3: Data consolidation (or “merging”)

The data consolidation stage involves adding complementary data to your cleaned dataset. In this step, it is important that both your main dataset and the complementary dataset are clean (with no formatting or content problems).

### Adding data from a complementary dataset.

- There are several ways to add/append complementary data to your main dataset.
- For small datasets, you can simply copy/paste the complementary data onto the main dataset.

	A	B	C	D
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>
2	AFRICA	1,182,439,000	1,340,598,000	2.51
3	ASIA	4,433,475,000		0.92
4	EUROPE	743,059,000	747,636,000	0.12
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	0.94
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65
7	OCEANIA	39,859,000	42,678,000	1.37
8				
9				
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))			
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>			

	A	B	C	D	E
1	<b>Geographic regions</b>	<b>2025</b>	<b>2030</b>	<b>2035</b>	<b>2040</b>
2	AFRICA	1,508,935,000	1,688,321,000	1,878,194,000	2,076,750,000
3	ASIA	4,822,629,000	4,974,092,000	5,096,362,000	5,188,949,000
4	EUROPE	745,791,000	741,303,000	735,101,000	727,811,000
5	LATIN AMERICA AND THE CARIBBEAN	681,896,000	706,254,000	726,395,000	742,348,000
6	NORTHERN AMERICAN	379,851,000	390,599,000	401,051,000	410,177,000
7	OCEANIA	45,335,000	47,919,000	50,421,000	52,814,000
8					
9					
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))				
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>				

	A	B	C	D	E	F	G	H
1	<b>Geographic regions</b>	<b>2015</b>	<b>2020</b>	<b>Growth rate</b>	<b>2025</b>	<b>2030</b>	<b>2035</b>	<b>2040</b>
2	AFRICA	1,182,439,000	1,340,598,000	2.51	1,508,935,000	1,688,321,000	1,878,194,000	2,076,750,000
3	ASIA	4,433,475,000		0.92	4,822,629,000	4,974,092,000	5,096,362,000	5,188,949,000
4	EUROPE	743,059,000	747,636,000	0.12	745,791,000	741,303,000	735,101,000	727,811,000
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	0.94	681,896,000	706,254,000	726,395,000	742,348,000
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65	379,851,000	390,599,000	401,051,000	410,177,000
7	OCEANIA	39,859,000	42,678,000	1.37	45,335,000	47,919,000	50,421,000	52,814,000
8								
9								
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))							
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>							

- Another option is to use the FILTER function of Google Sheets. The FILTER function allows you to get data from our complementary dataset based on data from our main dataset. In this case, we can get the 2025 population from our complementary dataset of each geographic region.
- Add a 2025 column to the main dataset (column E) and type the following command in E2:

```
=FILTER('[Data consolidation] Complementary dataset'!$B$2:$B$7,'[Data consolidation] Complementary dataset'!$A$2:$A$7 = A2)
```

E2	=FILTER([Data consolidation] Complementary dataset!\$B\$2:\$B\$7,[Data consolidation] Complementary dataset!\$A\$2:\$A\$7 = A2)								
	A	B	C	D	E	F	G	H	I
1	Geographic regions	2015	2020	Growth rate	1,508,935,000 ×				
2	AFRICA	1,182,439,000	1,340,598,000	2.51	=FILTER([Data consolidation] Complementary dataset!\$B\$2:\$B\$7,[Data consolidation] Complementary dataset!\$A\$2:\$A\$7 = A2)  FILTER(range, condition1, [condition2, ...])  EXAMPLE FILTER(A2:B26, A2:A26 > 5, D2:D26 < 10)  ABOUT Returns a filtered version of the source range, returning only rows or columns that meet the specified conditions.  range The data to be filtered. condition1 A column or row containing true or false values corresponding to the first column or row of 'range' or an array formula evaluating to true or false. condition2... - [optional] repeatable Additional rows or columns containing boolean values 'TRUE' or 'FALSE' indicating whether the corresponding row or column in 'range' should pass through 'FILTER'. Can also contain array formula expressions, which evaluate to such rows or columns. All conditions must be of the same type (row or column). Mixing row conditions and column conditions is not permitted.  <a href="#">Learn more</a>				
3	ASIA	4,433,475,000		0.92					
4	EUROPE	743,059,000	747,636,000	0.12					
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	0.94					
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65					
7	OCEANIA	39,859,000	42,678,000	1.37					
8									
9									
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both				range				
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>				The data to be filtered.				
12					condition1				
13					A column or row containing true or false values corresponding to the first column or row of 'range' or an array formula evaluating to true or false.				
14					condition2... - [optional] repeatable				
15					Additional rows or columns containing boolean values 'TRUE' or 'FALSE' indicating whether the corresponding row or column in 'range' should pass through 'FILTER'. Can also contain array formula expressions, which evaluate to such rows or columns. All conditions must be of the same type (row or column). Mixing row conditions and column conditions is not permitted.				
16									
17									
18									
19									

- If the formula ran successfully without any errors, we can then apply it to the remaining cells by dragging like we did in the last module, or simply by copy/pasting along the range.

E2	=FILTER([Data consolidation] Complementary dataset!\$B\$2:\$B\$7,[Data consolidation] Complementary dataset!\$A\$2:\$A\$7 = A2)								
	A	B	C	D	E	F	G		
1	Geographic regions	2015	2020	Growth rate	2025				
2	AFRICA	1,182,439,000	1,340,598,000	2.51	1,508,935,000				
3	ASIA	4,433,475,000		0.92					
4	EUROPE	743,059,000	747,636,000	0.12					
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	0.94					
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65					
7	OCEANIA	39,859,000	42,678,000	1.37					
8									
9									
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))								
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>								

E7	=FILTER([Data consolidation] Complementary dataset!\$B\$2:\$B\$7,[Data consolidation] Complementary dataset!\$A\$2:\$A\$7 = A7)								
	A	B	C	D	E	F	G		
1	Geographic regions	2015	2020	Growth rate	2025				
2	AFRICA	1,182,439,000	1,340,598,000	2.51	1,508,935,000				
3	ASIA	4,433,475,000		0.92	4,822,629,000				
4	EUROPE	743,059,000	747,636,000	0.12	745,791,000				
5	LATIN AMERICA AND THE CARIBBEAN	623,934,000	653,962,000	0.94	681,896,000				
6	NORTHERN AMERICAN	357,031,000	368,870,000	0.65	379,851,000				
7	OCEANIA	39,859,000	42,678,000	1.37	45,335,000				
8									
9									
10	Source: <a href="https://population.un.org/wpp/Download/Standard/Population/">https://population.un.org/wpp/Download/Standard/Population/</a> (Total Population (both sexes))								
11	Source: <a href="https://population.un.org/wpp/DataQuery/">https://population.un.org/wpp/DataQuery/</a>								

## **Data cleaning best practices.**

- Always back up your data. Once you get your hands on a dataset, it's good practice to work on a duplicate or copy of the dataset so that you will always have the original or raw version of the dataset available to you.
- When using spreadsheets, make a tab for every step of the cleaning process. This way you can go back to a previous step easily if you make a mistake. Don't rely on UNDO.
- Avoid overwriting or deleting data. There is also value in unclean data. Sometimes raw, unclean datasets can be used for other purposes, such as studying the number and frequency of errors. As such, it is important to keep a version of the raw dataset.
- Documentation. Add notes and documentation on what you were doing and how you were doing it. Your future self and others will thank you for it.

# MODULE 6 – ANALYZING DATA

---

## Introduction

Here we are! Data analysis is generally what comes to mind when we think about ‘working with data.’ And yet, as you’ve seen, the actual analysis only comes at the end of a multi-step process. You will also find that many times the process of analyzing the data is trivial compared to the previous steps. That is not to say that analyzing data is always easy: depending on your goals and the data, it can also become something that only seasoned statisticians can undertake.

Luckily, most of the analysis questions that you will encounter can be tackled with simple tools. If your questions look like ‘who is most affected by X?’ or ‘how has Y evolved over the years?’ then you will most likely be able to answer them yourself. If, instead, your question is of the ‘is X influenced by Y’ kind, then you will most likely need the help of someone with statistics training to help.

We can distinguish three types of data analysis:

- **Descriptive analysis** which focuses on describing the basic features of the data, such as the mean, the median, the maximum and so on.
- **Inferential analysis** which allows you to answer more complex questions about how the phenomena described by the data behave and extrapolate those answers to similar phenomena.
- **Predictive analysis** which is a more advanced type of inferential analysis, and which aims to make predictions about future events based on past data about these events.

In this module we will explore the methodology for data analysis by coming back to the DEFINE step, which is crucial for your analysis. The ANALYZE step will be demonstrated using examples belonging to the descriptive analysis group.

## Description

The module will cover the following concepts:

- Research question
- Research hypothesis
- Data question
- Descriptive statistics
- Pivot table

## Skills

As part of this module, you will learn the following:

- How to create a data analysis plan
- How to use pivot tables for data analysis
- How to use graphs for data analysis

## Prerequisites

- Module dataset: MIT Election Lab Data <http://bit.ly/MITelectiondata>
- A fair understanding of all previous modules
- Spreadsheets (Google Sheets)
- A working computer
- A modern internet browser
- Internet connection
- Basic knowledge of operating a computer

## Section 1: From research question to hypothesis

### Producing a good research question

Any data project starts with the DEFINE step of the Data Pipeline, during which you define a research question (also called a problem statement). The goal of the research question is to give your project a focus. Without it, you only have a theme, and you take the risk of wasting a lot of time analyzing the data in ways that do not help your initial aim.

Theme	Research question
Air pollution	Are air pollution levels around primary schools in Paris higher than the city average?
Gender parity in legislative elections	Do legislative election candidate lists include 50 percent of candidates from each gender?

*Examples of themes and research questions*

What does a good research question look like? The first thing to consider is the level of precision of your question: the vaguer the question. The less useful. A precise question will be actionable, allowing you to identify the specific data that you'll need to find.

Are children affected by air pollution?	×
Are there higher levels of air pollution around primary schools in Paris?	✓

*Examples of bad and good research questions*

In the example above, the good research question tells us that we will need

- Data around air pollution in Paris
- Geographical data about primary school locations in Paris

The second thing to consider are the constraints on your project: how much time do you have? Who do you want to present the project's results to? What is realistic to do, given your level of skill/knowledge? Fully exploring a theme may require you to produce multiple research questions, each being the start of a mini project. But are all those questions relevant to your audience? Do you have the time to explore them all? Is the data easy to collect for all of them? You may not have the answers to all of these questions until you have been through the FIND and GET steps. It is indeed very common to go back and



forth between DEFINE, FIND and GET, as you refine your research question.

## Defining your hypotheses

The second phase of your analysis plan is your hypotheses. A hypothesis is a proposed explanation for some phenomenon. A good hypothesis should be simple, concise, and most importantly, testable – you should be able to prove it wrong. Hypotheses are there for two reasons:

- They further refine your project by defining the specific observations that you expect to highlight in the data.
- They structure the narrative that you are trying to evidence with data.

Let's use the air pollution example to illustrate the process: with our research question we were able to identify the two datasets we needed (air pollution levels in Paris and location of primary schools). We can make two hypotheses:

1. Primary schools close to city roads have higher levels of air pollution.
2. The levels of fine particulate matter are twice as large around primary schools right before school starts.

Those are not the only hypotheses that we could make. But they are interesting because they help refine the variables we are looking for (distance to road, levels of fine particulate matter, school start and end times...) and they are aligned with the narrative we want to build (air pollution around schools is linked to cars).

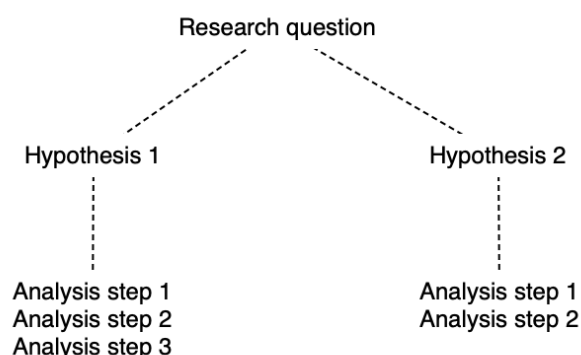
Lastly, a hypothesis is interesting only if it is grounded in something concrete – it could be an assumption shared by many people; or an observation you made in the field (such as the number of cars passing by schools in your city). Making sure that your hypothesis is grounded ensures that, regardless of your findings, the final story of your project is worth telling. Many people would be interested to hear that schools close to city roads do not have higher levels of air pollution, even if your hypothesis predicted higher levels of air pollution. For a civic organization, this is a good way to test the assumptions behind their campaigning:

- If their hypothesis is verified, then they can use the data to strengthen their campaigning.
- If their hypothesis is not verified, then it's a signal to shift their strategy.

## Section 2: Creating your analysis plan

Although research questions and hypotheses are closely linked to your final analysis, they are part of the DEFINE step and should be first produced at the start of your project. The ANALYZE step starts with the definition of your analysis plan, which is simply the series of transformations that you intend to apply to your data in order to generate the insights that you're looking for. Because the analysis plan comes after you have cleaned the data, each step can be worded in a way that includes the actual column names of your dataset.

Your goal should be that anyone should be able to reproduce your analysis step by step, given that they are provided with your cleaned data and your analysis plan.



Coming back to our previous example, the analysis plan linked to the first hypothesis (primary schools close to city roads have higher levels of air pollution) would consist of the following steps (the column names are just examples):

1. Calculate distance between PRIMARY\_SCHOOL and CITY\_ROAD
2. Create DISTANCE\_TO\_ROAD column and include results
3. For DISTANCE\_TO\_ROAD < 100M, calculate average of NOx and FPM50 levels
4. For DISTANCE\_TO\_ROAD > 100M, calculate average of NOx and FPM50 levels
5. Compare the two averages

(Here NOx means 'nitrogen oxide,' a pollutant emitted by cars, while FPM50 refers to fine particulate matter of 50 microns or less).

If you're not able to produce a full analysis plan yet, don't worry. You will naturally get better at anticipating the analysis steps that you should take to generate an insight using data. It is normal to try things directly with the data as a beginner. Just remember to add to your analysis plan once you have found the correct step. It will help you learn what the 'optimal' process should be and will also be a useful reference if you come back to the dataset in the future.

## Section 3: Common techniques for data analysis

### Using the pivot table

The basic steps of data analysis using spreadsheets are the same as the ones we used in Module 4 (Verifying Data): basic functions such as AVERAGE(), MAX(), MODE() etc. can be part of your data analysis plan and are sometimes all you need. But you will often want to slice or present your data in different ways to facilitate your analysis. For these use cases, the pivot table, a functionality found in all spreadsheet software, is the right tool.

#### Walkthrough: Using pivot tables

Let's open the dataset that we took from the MIT Election Lab (<https://electionlab.mit.edu/>): <http://bit.ly/MITelectiondata>

This is a sample of the full dataset of votes submitted for the United States' 2018 midterm elections.

Using this sample election data and the pivot table, we will apply the following analysis steps:

- Calculate total vote count per state or expressed as a percentage
- Calculate total vote count per candidate
- Calculate total vote count by mode (of voting)

Using the pivot table, we will determine 'Total vote count per state.' It will then be possible to express the vote count as a percentage of the total vote.

- Make a copy of the sheet in order to edit it: File → Make a copy
- Select the full dataset by clicking on the rectangle between the A and the 1
- Go to the Data → Pivot Table to insert your Pivot table.

state\_overall\_2018

File Edit View Insert Format Data Tools Add-ons Help Last edit was 23 minutes ago

100% 123 Arial 10

totalvotes

	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	state_gen	state_ic	office	district	stage	special	candidate	party	writein	mode	candidatvotes	totalvotes	unofficial
2	63	41	Associate Justice	statewide	gen	FALSE	Over Votes	NA	FALSE	absentee	8	1075793	FALSE
3	63	41	Associate Justice	statewide	gen	FALSE	Over Votes	NA	FALSE	election day	241	1075793	FALSE
4	63	41	Associate Justice	statewide	gen	FALSE	Over Votes	NA	FALSE	provisional	1	1075793	FALSE
5	63	41	Associate Justice	statewide	gen	FALSE	Sarah Hicks Stev	republican	FALSE	absentee	32836	1075793	FALSE
6	63	41	Associate Justice	statewide	gen	FALSE	Sarah Hicks Stev	republican	FALSE	election day	1060731	1075793	FALSE
7	63	41	Associate Justice	statewide	gen	FALSE	Sarah Hicks Stev	republican	FALSE	provisional	3568	1075793	FALSE
8	63	41	Associate Justice	statewide	gen	FALSE	Under Votes	NA	FALSE	election day	20524	1075793	FALSE
9	63	41	Associate Justice	statewide	gen	FALSE	Under Votes	NA	FALSE	election day	550653	1075793	FALSE
10	63	41	Associate Justice	statewide	gen	FALSE	Under Votes	NA	FALSE	provisional	1759	1075793	FALSE
11	63	41	Associate Justice	statewide	gen	FALSE	NA	NA	TRUE	election day	646	1075793	FALSE
12	63	41	Associate Justice	statewide	gen	FALSE	NA	NA	TRUE	election day	34460	1075793	FALSE
13	63	41	Associate Justice	statewide	gen	FALSE	NA	NA	TRUE	election day	133	1075793	FALSE
14	63	41	Associate Justice	statewide	gen	FALSE	Over Votes	NA	FALSE	absentee	8	1075503	FALSE
15	63	41	Associate Justice	statewide	gen	FALSE	Over Votes	NA	FALSE	election day	145	1075503	FALSE
16	63	41	Associate Justice	statewide	gen	FALSE	Over Votes	NA	FALSE	provisional	0	1075503	FALSE
17	63	41	Associate Justice	statewide	gen	FALSE	Tommy Bryan	republican	FALSE	absentee	32572	1075503	FALSE
18	63	41	Associate Justice	statewide	gen	FALSE	Tommy Bryan	republican	FALSE	election day	1061520	1075503	FALSE
19	63	41	Associate Justice	statewide	gen	FALSE	Tommy Bryan	republican	FALSE	provisional	2530	1075503	FALSE
20	63	41	Associate Justice	statewide	gen	FALSE	Under Votes	NA	FALSE	absentee	20801	1075503	FALSE
21	63	41	Associate Justice	statewide	gen	FALSE	Under Votes	NA	FALSE	election day	560857	1075503	FALSE
22	63	41	Associate Justice	statewide	gen	FALSE	Under Votes	NA	FALSE	provisional	1787	1075503	FALSE
23	63	41	Associate Justice	statewide	gen	FALSE	NA	NA	TRUE	absentee	620	1075503	FALSE
24	63	41	Associate Justice	statewide	gen	FALSE	NA	NA	TRUE	election day	33526	1075503	FALSE
25	63	41	Associate Justice	statewide	gen	FALSE	NA	NA	TRUE	provisional	137	1075503	FALSE
26	63	41	Associate Justice	statewide	gen	FALSE	Over Votes	NA	FALSE	absentee	6	1074954	FALSE

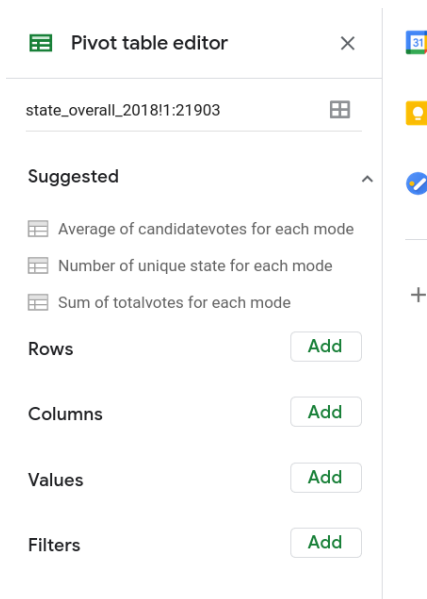
state\_overall\_2018 Pivot Table 1

Explore

## The spreadsheet

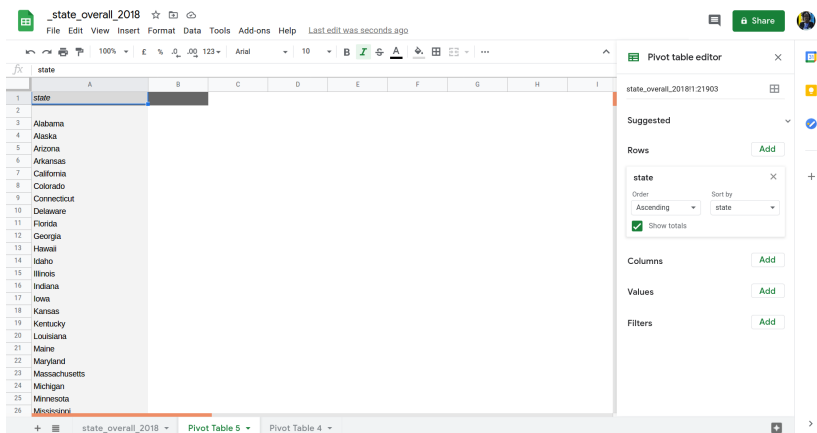
Pivot tables are all about taking big sets of original data and putting it into a report you can understand. The pivot table editor is Sheets' tool to build your pivot table out and understand your data in more detail. Let's get familiar with the pivot table editor. Google Sheets has four options on the left side to put data into a pivot table: Rows, Columns, Values, and Filter.

The pivot table editor allows you to build out a report and view the data the way you want to see it.



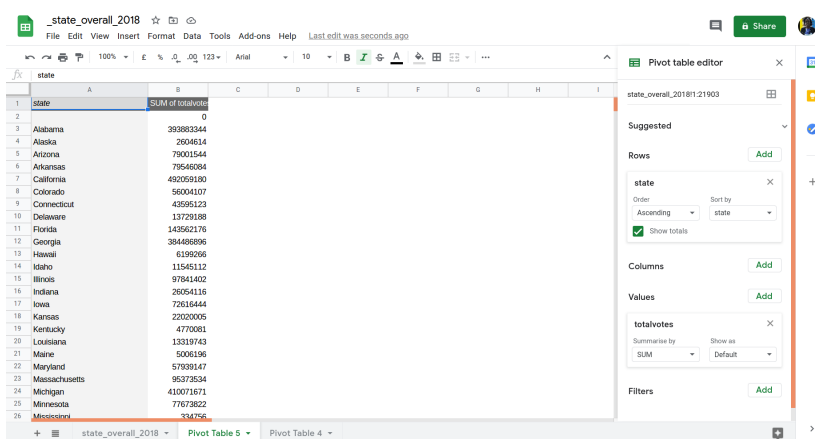
### Pivot table editor

- Click 'add field' for any of the options and you'll see the columns from your original data. Click on one of the column names to add the data in the given format. For example, if I click 'add field' next to Rows, and then click State, this is the view that you'll see:



### Add the 'State' column

- The pivot table shows each of the states on their own row. The pivot table is taking the original data and showing it in a report. We could go on to add the 'total number of votes' for each state to the pivot table editor. Click 'add field' next to Values, and then click total votes, this is the view that you'll see:



Added 'Total number of votes'

Here is what each of the four options in the report builder does when you add a field to it:

**Rows** – Add a column as a row to see each of the values on the left side of your pivot table. Rows are for the main variable you're interested in.

**Columns** – When you add a field as a column, each of the items will be shown in a column of their own. The item you add as a column should be a secondary variable that you want to cross-analyze with the main one.

Analysis step	Rows	Columns	Value
For each state, I want to see the number of candidates	State		Candidate (COUNTA)
For each state, I want to see the number of candidates per mode of vote	State	Mode	Candidate (COUNTA)
For each party, I want to see the number of votes	Party		Total votes (SUM)
For each party, I want to see the number of votes in each state	Party	State	Total votes (SUM)

**Values** – Pull numerical amounts into the values section to show. Values can be used in two ways:

- To count the number of instances in the dataset of each element of the variable set in the rows of the pivot table. For example, if we wanted to know how many times each state appeared in the dataset, we would pick the variable 'state' both in Rows and in Values.
- To calculate a value based on the variables set in Rows and Columns. This is what we did when we calculated the sum of votes for each state.

**Filter** – You can pull a field into the filter box to let you filter your data down, which is useful when you only need a smaller slice of the data.

- We'll now use a spreadsheet formula to calculate and express 'SUM of total vote' as a percentage.
- To start, move to the first row.

fx	=(B2				
	A	B	C	D	E
1	state	SUM of totalvotes			
2	Alabama	39388334	=(B2		
3	Alaska	2604614			
4	Arizona	79001544			
5	Arkansas	79546084			
6	California	492059180			
7	Colorado	56004107			
8	Connecticut	43595123			
9	Delaware	13729188			
10	Florida	143562176			
11	Georgia	384486896			
12	Hawaii	6199266			
13	Idaho	11545112			
14	Illinois	97841402			

- Now type / and select the second cell you want to add, then type ) to close the brackets. Type \* and then 100.

fx	=(B2/B52)*100			
	A	B	C	D
49	West Virginia	17448021	C2	
50	Wisconsin	5826612	=(B2/B52)*100	
51	Wyoming	6915397		
52	<b>Grand Total</b>	<b>4992392735</b>		
53				
54				
55				
56				
57				
58				

Full formula to calculate percentage

- Press Enter or tab. The formula disappears and is replaced by the value.

fx					
	A	B	C	D	E
1	state	SUM of totalvotes	percentage		
2	Alabama	393883344	7.89		
3	Alaska	2604614			
4	Arizona	79001544			
5	Arkansas	79546084			
6	California	492059180			
7	Colorado	56004107			
8	Connecticut	43595123			
9	Delaware	13729188			
10	Florida	143562176			

- Try changing the number in one of the original cells (sum of totalvotes) you should see the value in total update automatically.
- You can type each formula individually, but it is also possible to cut and paste or drag formulas across a range of cells.
- Copy the formula you have just written (using ctrl + c) and paste it into the cell below (using ctrl + v), you will get the sum of the two numbers on the row below.
- Alternatively click on the lower right corner of the cell (the blue square) and drag the formula down to the bottom of the column. The 'percentage' column will then be updated.

## Using data visualizations for data analysis

While pivot tables are useful to slice and reshape your data in many ways, they still present a series of numbers, which can be unhelpful when trying to pick out trends or identify outliers. This is where data visualization can be useful – a simple line graph can be more useful to identify a trend over time than looking at a row of numbers.

Combining pivot table and data visualization is also possible: the pivot table is particularly suited for a type of visualization called a heatmap. As explained by the Dataviz catalogue: “Heatmaps are good for showing variance across multiple variables, revealing any patterns, displaying whether any variables are similar to each other, and for detecting if any correlations exist in between them.”

(<https://datavizcatalogue.com/methods/heatmap.html>)

### Walkthrough: Create a heatmap with your dataset


We're going to compare the share of the votes for each party in the Midwest (Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin). We must reword our objective to clarify the pivot table steps:

“For each state of the Midwest, I want to see the share of the counted votes for each party.”

- Using the existing election data spreadsheet, create a new pivot table after selecting the whole dataset.
- In Rows, add the variable 'state.'
- In Filter, add the variable 'state' then use the dropdown to select only the Midwest states.
- In Columns, add the variable 'party.'
- In Values, add the variable 'total vote.'

We now have a share of the vote by party for the Midwest states. But we'd like to see better which party got the biggest share of each state. This is where using a heatmap could be convenient. To create a heatmap we will use the conditional formatting functionality and more specifically its color scale option.

- The pivot table currently shows the absolute number of votes for each state/party combination. This makes it difficult to compare across states as they do not have the same population size. To have comparable numbers, we will use percentages. To do so, go to the total votes variable in Values and change the option 'show as' to display '% of row.' We want it across rows because we want to see which party got the best score in each state. Had we used '% of column' the percentage would tell us in which state each party got the most votes.


**Pivot table editor**
×

**Rows**
Add

**state**
×

Order

Ascending

Sort by

state

☒ Show totals

**Columns**
Add

**party**
×

Order

Ascending

Sort by

party

☒ Show totals

**Values**
Add

**totalvotes**
×

Summarise by

SUM

Show as

% of row

**Filters**
Add

**state**
×

Status

Showing 12 items

- Now select your whole table at the exception of the 'grand total' line and column (you can also remove by checking the 'show totals' boxes in Rows and Columns).
- Go to Format → Conditional formatting. A sidebar will appear on the right.
- Select Color Scale and change the color, from white (lower values) to red (higher values).



Conditional format rules

Single colour
Colour scale

Apply to range
B3:V14

Format rules

Preview
Default

Minpoint
Min value

Midpoint
None

Maxpoint
Max value

Cancel Done

- Click on 'Done'.
- And voila! We can now easily identify outliers in the data such as the share of the vote that the Democratic party got in South Dakota.

	A	B	C	D	E	F	G	H	I	J
1	SUM of total vote party									
2	state	clear water	conservative	constitution	democrat	democratic-farm	democratic-npl	downstate	unitex	grassroots-legali green
3	Illinois		4.65%		29.54%			0.08%		
4	Indiana				36.36%					
5	Iowa	3.66%			26.01%					
6	Kansas				27.45%					
7	Michigan				19.04%					6.31%
8	Minnesota					16.43%			6.58%	
9	Missouri			10.02%	23.78%					10.60%
10	Nebraska				32.70%					
11	North Dakota						26.64%			
12	Ohio				24.29%					4.01%
13	South Dakota				46.85%					
14	Wisconsin			9.04%	24.01%					
15	Grand Total	0.28%	0.49%	0.82%	21.19%	1.37%	0.16%	0.01%	0.55%	3.55%
16										

**Tip 1:** The table shows votes across all offices. We could have added a filter based on the 'office' variable to show the votes for a specific office.

**Tip 2:** Pivot tables are also a great tool to dive into a specific subsection of the data: Double-clicking on the cell of the pivot table opens a new sheet with the slice of data corresponding to that cell.

	A	B	C	D	E	F	G	H	I
1	year	state	state_po	state_fips	state_cen	state_ic	office	district	stage
2	2018	South Dakota	SD	46	45	37	Attorney General	statewide	gen
3	2018	South Dakota	SD	46	45	37	Commissioner of State Lands	statewide	gen
4	2018	South Dakota	SD	46	45	37	Governor and Lieutenant Governor	statewide	gen
5	2018	South Dakota	SD	46	45	37	Public Utilities Commission	statewide	gen
6	2018	South Dakota	SD	46	45	37	Secretary of State	statewide	gen
7	2018	South Dakota	SD	46	45	37	State Auditor	statewide	gen
8	2018	South Dakota	SD	46	45	37	State Representative	District 1	gen
9	2018	South Dakota	SD	46	45	37	State Representative	District 1	gen
10	2018	South Dakota	SD	46	45	37	State Representative	District 10	gen
11	2018	South Dakota	SD	46	45	37	State Representative	District 10	gen
12	2018	South Dakota	SD	46	45	37	State Representative	District 11	gen
13	2018	South Dakota	SD	46	45	37	State Representative	District 11	gen
14	2018	South Dakota	SD	46	45	37	State Representative	District 12	gen
15	2018	South Dakota	SD	46	45	37	State Representative	District 12	gen
16	2018	South Dakota	SD	46	45	37	State Representative	District 13	gen
17	2018	South Dakota	SD	46	45	37	State Representative	District 14	gen
18	2018	South Dakota	SD	46	45	37	State Representative	District 14	gen
19	2018	South Dakota	SD	46	45	37	State Representative	District 15	gen
20	2018	South Dakota	SD	46	45	37	State Representative	District 15	gen
21	2018	South Dakota	SD	46	45	37	State Representative	District 16	gen
22	2018	South Dakota	SD	46	45	37	State Representative	District 16	gen
23	2018	South Dakota	SD	46	45	37	State Representative	District 17	gen
24	2018	South Dakota	SD	46	45	37	State Representative	District 17	gen
25	2018	South Dakota	SD	46	45	37	State Representative	District 18	gen
26	2018	South Dakota	SD	46	45	37	State Representative	District 18	gen
27	2018	South Dakota	SD	46	45	37	State Representative	District 19	gen
28	2018	South Dakota	SD	46	45	37	State Representative	District 2	gen
29	2018	South Dakota	SD	46	45	37	State Representative	District 2	gen
30	2018	South Dakota	SD	46	45	37	State Representative	District 20	gen
31	2018	South Dakota	SD	46	45	37	State Representative	District 20	gen

New sheet created to show details

DELETE

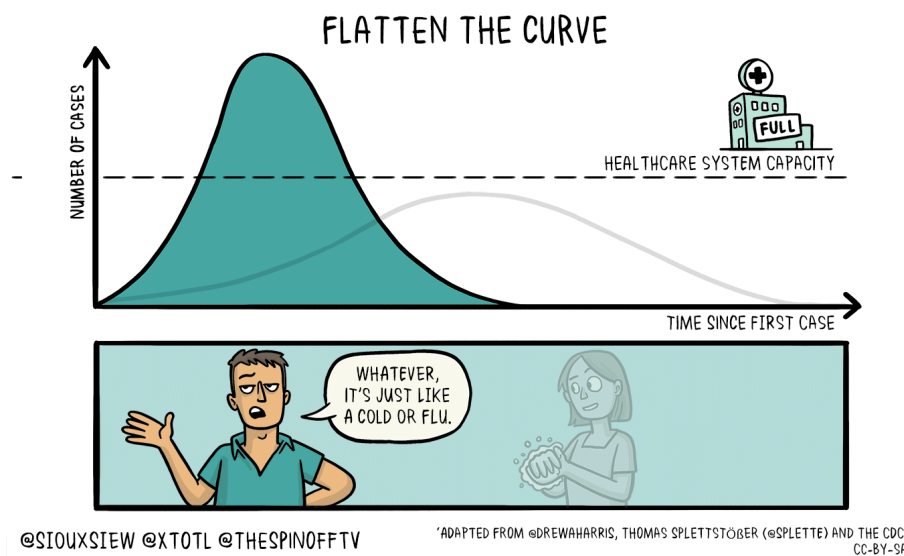


# MODULE 7 – PRESENTING DATA

## Introduction

The first question that you must ask yourself at this stage of the Data Pipeline is not ‘how do I visualize the data?’ but rather, ‘what do I want my audience to get out of my project?’ Said differently, the key parameters of data presentation are your audience and your message.

All too often, people default to the idea that a data analysis should be presented with a graph of some sort, which is not always the case. After all, how would you present your results on the radio? To people with a visual impairment? To remote communities with no data visualization literacy? In all those cases, a graph wouldn’t work. On the other hand, the message itself is also key: there are many ways to present your results, and sometimes an engaging presentation provides more value than a strict adherence to the data.



All of this to say that the use of the title ‘Presenting Data’ instead of ‘Visualizing Data’ is deliberate. There’s also one more reason that you may have noticed-- data visualizations are used throughout the Data Pipeline! They are frequently used for data verification and analysis, which make them not exclusive to the presentation step.

With that said, there will often be cases where you need to use an actual data visualization. For those cases you’ll need to understand what the various types of data visualizations are best at, and how to pick the right one for your project.

## Description

The module will cover the following concepts:

- Presenting vs visualizing
- Data visualization tasks
- Continuous variables
- Categorical variables

## Skills

As part of this module, you will learn the following:

- Choosing the right presentation format for your analysis
- Choosing the right data visualization format for your analysis
- Labelling your charts correctly
- Creating graphs with Google Sheets

## Prerequisites

- Module dataset: <http://bit.ly/GCBAAsiaSample>
- A fair understanding of all previous modules
- Spreadsheets (Google Sheets)
- A working computer
- A modern internet browser
- Internet connection
- Basic knowledge of operating a computer

## Section 1: Choosing your emphasis

Over the past 15 years, data journalism has grown immensely and both in reach and depth. While just a few journalists were doing data work 10 years ago, now teams hundreds strong collaborate on far reaching projects. And as more and more data-driven articles are published, more and more of them did away with visualizations as a way to communicate findings. And it makes sense – after all, data is a means to an end, much like a witness interview is.

Beyond data journalism, a civic organization working with one dataset may choose to present it differently depending on its audience:

- If within a report, then it may choose to include a table and a simple graph.
- If to a specialist audience, the results could be shared through a flashy interactive data visualization.
- If to a general audience, then an infographic should suffice.

## Data Visualization vs. Infographics

The two are used interchangeably, but they refer to two different things, as hinted at by their names.

**Data visualizations** are tightly linked to their underlying data, which is both an opportunity (e.g., for interactive visualizations which allow people to explore the data in a visual way, and which react to changes in the underlying data) and a constraint (there are a limited number of viable data visualization formats that translate the data faithfully while still being easy to understand for a human reader).

**Infographics** are, as the name describes, information graphics. They are used to present fact-based stories in a visual way. They frequently include data visualizations but are not constrained by any underlying data. This allows for more creativity in terms of presentation but limits their use to general communication, unlike data visualizations which can also be used for specialized and technical uses.

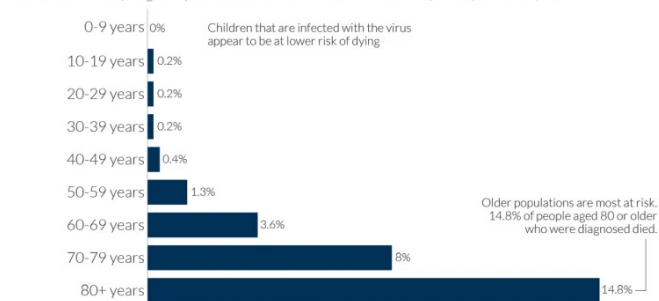
There are three key elements that you can choose to emphasize when presenting data: the data, the design, the message.

**Emphasizing the data** is done through simple visualizations which allow the audience to quickly get a sense of key figures. With web visualizations, interactivity allows even more precise exploration of the underlying data, and sometimes personalization.

### Coronavirus: early-stage case fatality rates by age-group in China

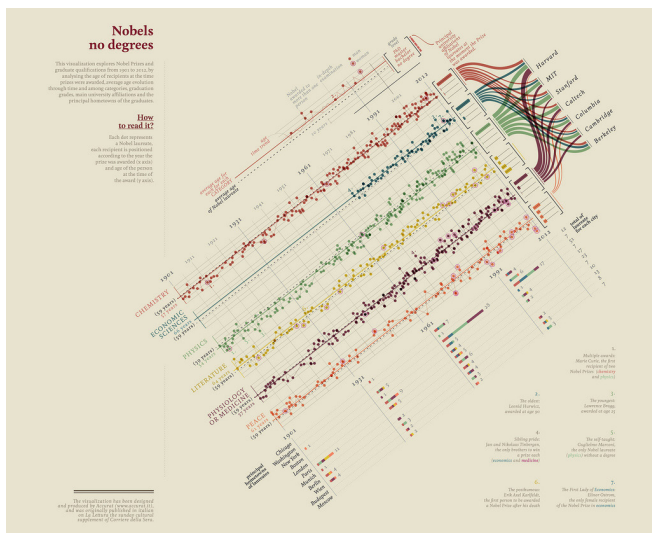
Our World in Data

Case fatality rate (CFR) is calculated by dividing the total number of deaths from a disease by the number of confirmed cases. Data is based on early-stage analysis of the COVID-19 outbreak in China in the period up to February 11, 2020.



Data source: Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. Vital surveillance: the epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020. China CDC Weekly. OurWorldinData.org - Research and data to make progress against the world's largest problems. Licensed under CC-BY by the authors.

**Emphasizing the design** requires design chops to successfully deliver but can produce very engaging visualizations that make up for their lack of familiarity with the attractiveness of their presentation. The goal here is to make the content memorable.



*Data visualization by La Repubblica*

**Emphasizing the message** generally means putting storytelling at the forefront. That can be done with interactive formats that weave data visualizations inside a larger story, or with simple infographics.



Visual scrollytelling by the New York Times

## Section 2: Choosing your visualization

So, you've decided to develop a data visualization. How do you choose which format best conveys your message? You can let the software choose for you, when the option is possible (such as in Google Sheets) but although the software can guess what you're trying to do based on the features of the data, it does not know your audience nor the context of your project. This can lead to visualizations that are technically correct, but which will confuse your readers.

There are two main things to consider in choosing the data visualization format: the type of data and the visualization task.

### Categorical and continuous data

At the basic level you can distinguish categorical and continuous data (or variable).

**Categorical data** is a type of qualitative data that is split across multiple groups that you can list: gender, colors, age groups, job category, etc. Certain charts, such as bar charts or pie charts, are perfectly suited for categorical data. Specifically, they associate one categorical variable (e.g., job category) with at least one continuous variable (e.g., average income, or percentage of all jobs in the case of a pie chart). It is not possible to associate only categorical variables in a chart; at least one continuous variable is needed.

**Continuous data** is data which is countable, and which can take any value on a scale weight, income, speed, etc. A few charts, such as the scatterplot (<https://datavizcatalogue.com/methods/scatterplot.html>) or the histogram (<https://datavizcatalogue.com/methods/histogram.html>), take continuous data on both axis. But most charts require a mix of continuous and categorical data. It is important to note that continuous data can be displayed as categorical data – this is often the case for time, which can be presented either as a continuous timeline or distinct timeframes (weeks, months, years, etc.). Similarly, age and income are often grouped into categories for analysis. But although the groupings are less natural than for time, the same can be done with any other continuous variable.

### Visualization tasks

Different visualizations represent data differently. Specifically, they highlight a certain relationship between the values of the dataset. This action is called the visualization task (or function). The most common relationships include comparison (e.g., a bar chart), part-to-a-whole (e.g., a pie chart), change over time (e.g., a line chart), or localization (maps).

There is no canonical list of tasks that represent all possible data visualizations, but the tasks of the common charts are well understood. A single chart can be linked to multiple visualization tasks, and multiple types of data may be displayed in the same visualization. The Dataviz catalogue (<https://datavizcatalogue.com>), one of the references for data visualization information, lists the following tasks:

- Comparisons
- Proportions
- Relationships
- Hierarchy
- Location

- Part-to-a-whole
- Distribution
- Flow
- Patterns
- Range
- Data over time

Multiple websites feature chart choosers that can help you identify which charts fits your objective best:

- Depict Data Studio <https://depictdatastudio.com/charts/>
- Juice Analytics: <http://labs.juiceanalytics.com/chartchooser/index.html>
- Dataviz Catalogue: <https://datavizcatalogue.com/search.html>

Just like we did for the pivot table, identifying a visualization task is a matter of properly wording our goal. An easy way to do it is to start with 'I want to show...'

I want to show...	Task
The number of adults and children who exercise regularly	Comparison
The proportion of adults and children who exercise regularly	Comparison, part-to-a-whole
The evolution over time of the number of adults and children exercising regularly	Comparison, pattern, data over time
The share of adults and children exercising regularly across men and women	Part-to-a-whole, comparison, hierarchy

### Walkthrough: From dataset to data visualization

We will use the Global Corruption Barometer sample data: <http://bit.ly/GCBAAsiaSample> to analyze the variation in opinion of people of different ages about corruption.

- Select the whole dataset and create a pivot table.
- Let's now formulate our analysis goal properly. For each unique age value, I want the share of people who answered Agree or Strongly Agree to the statement 'Q30: the government is run by a few big interests looking out for themselves.' Our main variable is 'age,' and our value will be a count of the column 'Q30,' filtered to keep only the Agree and Strongly Agree values.
- Let's do as we planned. In the pivot table sidebar, add 'age' to Rows, 'Q30' to Value and Q30 to Filter. In the Filter dropdown, unselect everything but 'Agree' and 'Strongly Agree.'
- You should obtain something like this:



	A	B
1	AGE	COUNTA of Q30
2	18	113
3	19	125
4	20	183
5	21	141
6	22	142
7	23	160
8	24	228
9	25	164
10	26	108
11	27	120
12	28	132
13	29	101
14	30	243

- We want to compare the respondents' opinions across all ages, but not all ages are represented equally. We consequently need to look at the percentage that each of those lines represent. For example, we want to clarify how many respondents were 18 years old in order to identify the percentage of the total that the 113 who responded 'Agree' or 'Strongly Agree' represent.
- To do so, we will create a third column called '% age group' in column C and include the formula that will help us calculate the percentage:

```
=B2/ COUNTIF( 'Sample dataset' !C:C,A2)
```

Where:

- B2 is the number of respondents who answered 'Agree' or 'Strongly Agree'
- 'Sample dataset'!C:C, is the 'age' column of the dataset
- A2 is age whose population we're calculating
- COUNTIF( ) is a formula that counts all the lines of the column 'Age' where the value is equal to A2
- By dragging down the formula, the percentage is calculated, and we obtain the following:

	A	B	C
1	AGE	COUNTA of Q30: t	% age group
2	18	113	0.4346153846
3	19	125	0.4222972973
4	20	183	0.3995633188
5	21	141	0.4433962264
6	22	142	0.3523573201
7	23	160	0.387409201
8	24	228	0.3870967742
9	25	164	0.3877068558
10	26	108	0.3686006826
11	27	120	0.3715170279
12	28	132	0.3739376771
13	29	101	0.3976377953
14	30	243	0.3790951638
15	31	78	0.3611111111
16	32	156	0.3768115942

- To have the values displayed as percentages, we just need to select the column and go to Format → Number → Percent.

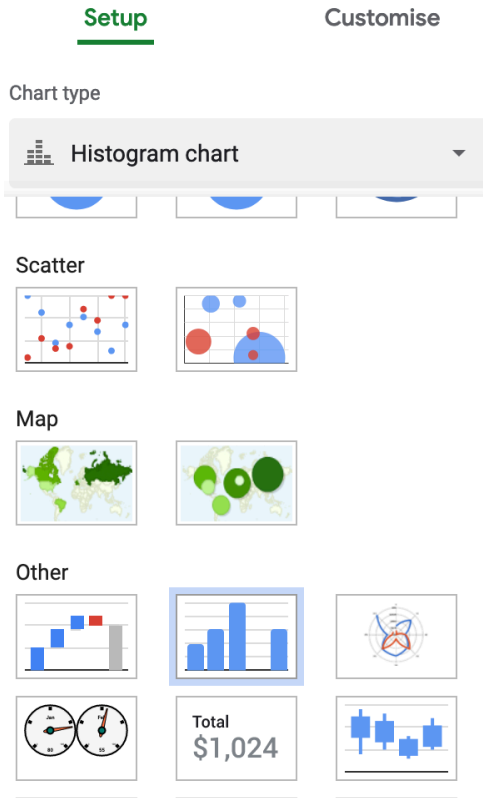
- Our dataset is now ready to be visualized, which leads us to our main goal: identifying our visualization task, which is 'I want to show how opinion about corruption varies across ages.' Age is continuous data and a relationship between age and opinion implies a pattern (there is no time variable, so it is not data over time). If we look at the choice of graphs under the 'Pattern' category in the Dataviz catalogue, we have many options:



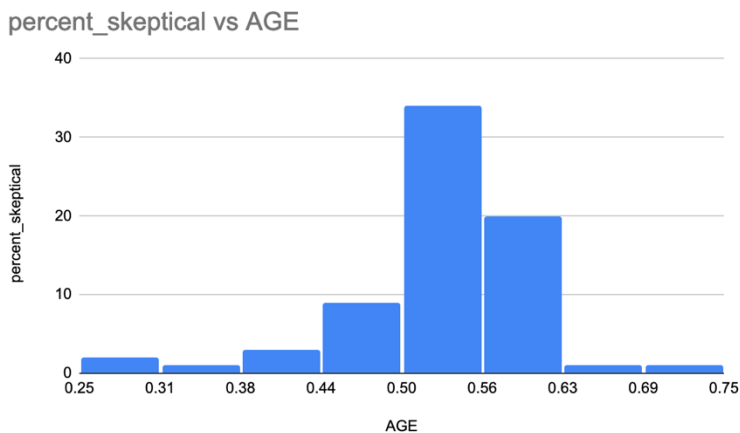
*A screenshot from [datavizcatalogue.com](https://datavizcatalogue.com)*

- To restrict further our choices, we can take into account that:
  - Our chart only has two variables: age and opinion about corruption. This eliminates some options such as the bubble chart, the Box & Whisker Plot, the multi-set bar chart, the population pyramid...
  - We have two continuous variables (age and percentage of respondents) and one categorical variable (opinion on corruption), which eliminates charts that require at least one categorical variable (see continuous vs categorical section below). We are left with charts with two continuous variables, such as the line chart, the histogram, or the density plot. And that works well for us because these charts are also great at showing patterns, which is what we want.
- It is then a matter of choosing among the leftover charts with the message you want to send as a guide. The line chart is often used for time series and may give the idea that opinion is fluctuating over time, instead of by age. Histograms and density plots do not have that issue, and histograms in particular have an added benefit: they allow us to split our age variable into groups called 'buckets' (or 'bins'). We have the choice of setting the bucket size based on age groups (e.g., comparing people aged 20-29 to those aged 30-39) or number of responses (each bar of the histogram could correspond to 10,000 responses).
- To finish our visualization, we simply need to select both the columns A and C:

- First click on the A to select the full column A.
- Then maintain CTRL (CMD on Mac) while clicking on the C to select the full column C while maintaining the selection of column A.
- Go to Insert → Chart to finalize the data visualization. Here you need to change the proposed visualization to histogram (under 'Other').



- By default, Google Sheets suggests a bucket size based on the age variable, with each bucket corresponding to six years. A histogram gives us an idea of where our values are concentrated: in our case, the age groups with the highest percentage of positive responses are 50-56, 56-63 and 44-50.



## Section 3: The tools for data presentation

There are many, many tools available for presenting your data. Listing all their functionalities would be an immense task whose result would be quickly obsolete as these tools are being updated frequently.

But we can distinguish three types of tools used for data presentation

**Data wrangling applications:** many applications that are designed for data wrangling (i.e., cleaning, tidying, merging, and reshaping), from spreadsheet software, to business intelligence software (Tableau (<https://public.tableau.com/en-us/s/>), Google Data Studio (<https://datastudio.google.com/u/0/>), etc.), to data analysis software (R, SPSS, Workbench, etc.) also include some data visualization functionalities.

**Data visualization applications:** these include applications like Datawrapper, Flourish, Raw Graphs, Khartis and many others which are focused on producing powerful and aesthetically pleasing visualizations. Data visualizations libraries (D3.js, Plot.ly...), which can only be used through programming, can also be included here.

**Information graphics applications:** these may include some data visualization functionalities but are mainly used for building visual narratives around your data. This includes Piktochart, Canvas, and Illustrator, which is also widely used for putting the finishing touches on data visualizations produced through other software.

**Storytelling applications:** this category includes Timeline.js, Storymap, but also specific functionalities from other tools such as the dashboard functionality of Tableau or the multi-step visualization functionality of Flourish. At the other extreme in terms of complexity, some programming libraries allow for 'scrollytelling' types of presentation, where the visualization can be animated as the reader scrolls down the page.

# MODULE 8 – USING GEOGRAPHICAL DATA

---

## Introduction

The story of Geographical Information Systems (GIS) begins in the world of maps. A map is a simplified and conventional visual representation of real things from the real world.

Just as the name suggests, a GIS is a computer system which allows you to work with geographical information – just like you would use a word processor to type a document or a spreadsheet to do numerical calculations and make graphs to visualize your data.

A GIS is a computer-based system that provides the following four sets of capabilities to handle georeferenced data:

- Data capture and preparation
- Data management, including storage and maintenance
- Data manipulation and analysis
- Data presentation

GIS allows you to build powerful maps and applications using geographic data. A good map, and therefore a good GIS, is a collection of thematic layers that lets you identify relationships and trends in your data, perform spatial analysis, and inform decision making.

You'll notice that a significant component of a GIS is geographic data.

## Data and common data sources

There are two types of data that a GIS can read:

**Vector data:** a representation of the world using points, lines, and polygons (closed shapes). Vector information is stored in x, y coordinates. These data have discrete boundaries like country borders, land parcels and roads.

**Raster data:** raster data is any pixelated (or gridded) data where each pixel is associated with a specific geographical location. The value of a pixel can be continuous (e.g., elevation) or categorical (e.g., land use). If this sounds familiar, it is because this data structure is very common, it's how we represent any digital image.

The thematic layers in your GIS can come from many sources. They can be base layers, such as street maps and imagery from services like Google Maps (<https://www.google.com/maps>) and OpenStreetMap (<https://www.openstreetmap.org/>). They can be foundational layers stored in a geodatabase, such as a parcel or administrative boundary layers complete with associated rules and behavior. Often, state agencies make these data available. And they can be layers you've pulled into the map from external systems, such as open data portals using commonly used services like the Web Map Service (WMS) or Tile Map Service (TMS).

## **Introduction to QGIS**

QGIS is a free and open-source cross-platform desktop geographic information system application that supports viewing, editing, and analysis of geospatial data.

### **Description**

The module will cover the following concepts:

- Geodata
- Layers
- Geographical features
- Coordinates
- Projections
- Openstreetmap

### **Skills**

As part of this module, you will learn the following:

- Reading and understanding geodata formats
- Creating an offline map
- Creating an online map
- Styling a map for publication
- Embed an interactive map in a website
- Export data from OpenStreetMap

### **Prerequisites**

- Module dataset: <http://bit.ly/totalvotesUSA>
- A fair understanding of all previous modules
- QGIS software
- Google Sheets
- A working computer
- A modern internet browser
- Internet connection
- Basic knowledge of operating a computer

## Section 1: The basics of GIS

Let's start with an example: an election happened. You'd now like to know how the total votes were distributed across the U.S. states. Below is an example tutorial using Datawrapper to show how you might go about it.

Geographic and location information have become ubiquitous in the 21st century, at all levels of granularity. We have satellites, aircraft, and even commercial drones that can capture large quantities of raw data over large areas over long periods of time. There are also sensors that collect and gather location information, such as the GPS and other applications on our phones and gadgets.

With all of this data available to us, knowing how to properly manage, analyze, and present it is imperative. This is where Geographic Information Systems, or more commonly known as GIS, come in.

GIS is a tool/framework/system that allows us to work with geographic (or spatial) information and its corresponding attribute (non-spatial) information. A GIS should be able to:

- **Collect, store, edit, manipulate**, or generally, **manage** spatial data.
- **Analyze** the data using its spatial component (where the data is) or attribute (the value of the data).
- **Present** the data and analysis as meaningful information through maps, charts, plots, and other visualizations.

*GIS is not just mapping.* Although map-making is one of the most important and commonly known aspects of GIS, it is just one of its many parts and features. We can view the modern GIS framework as being composed of:

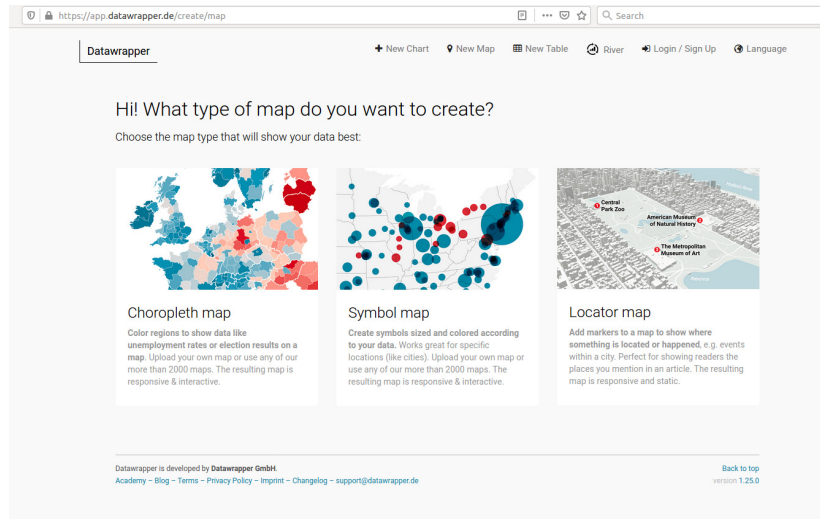
- **Data** – these refer to the pieces of information used by GIS which may or may not have location information (e.g., names of stores, locations of stores).
- **Technology** – these include both the hardware and software components of a GIS – the machines, instruments, and computers with the software applications used to do geospatial work.
- **Methods & Analysis** – these refer to the science and mathematics in GIS which allows it to perform both spatial and non-spatial (attribute) queries and computations.
- **Visualization** – although similar to methods & analysis, visualization can be considered as a separate component, as these refer specifically to the means by which a GIS represents data graphically. These include not just maps, although that's the most common case, but also tables, charts, graphs, etc.
- **People** – these are the people and/or groups of people who are involved in the use, development, teaching, and maintenance of GIS.

All parts of the modern GIS framework – from the data to the people – are integral in ensuring its success.

For a more comprehensive review of GIS key concepts, see: <http://bit.ly/GISbasics>

## Section 2: Creating an online map

- Datawrapper lets you create maps and graphs using your own data and embed them in your site or share on social media. In this tutorial, we'll focus on the mapping functionality which takes geographic data and analyzes it using inbuilt routines. The output is a simple visual in the form of a map that helps you understand your data better.



### Datawrapper

- We'll use this CSV file (<http://bit.ly/totalvotesUSA>) (Total votes by each U.S. state) to demonstrate how Datawrapper may be used to create shareable maps. Two files were joined within QGIS to arrive at this dataset: the results from the pivot table and a shapefile from the U.S. Census Bureau with state boundary data. You'll notice that the file contains 'sum of total votes' and 'percentage' of total votes for each state.

E	F	G	H	I	J	K
STUSPS	NAME	LSAD	ALAND	AWATER	_state_ove	percentage_vote
MD	Maryland		0 25151100280	6979966958	57939147	1.161
IA	Iowa		0 144661267977	1084180812	72616444	1.455
DE	Delaware		0 5045925646	1399985648	13729188	0.275
OH	Ohio		0 105828882568	10268850702	116369036	2.331
PA	Pennsylvania		0 115884442321	3394589990	32895529	0.659
NE	Nebraska		0 198956658395	1371829134	14286784	0.286
WA	Washington		0 172112588220	12559278850	23099156	0.463
PR	Puerto Rico		0 8868896030	4922382562		
AL	Alabama		0 131174048583	4593327154	393883344	7.89
AR	Arkansas		0 134768872727	2962859592	79546084	1.593
NM	New Mexico		0 314196306401	728776523	12739040	0.255
TX	Texas		0 676653171537	19006305260	306948546	6.148
CA	California		0 403503931312	20463871877	492059180	9.856

### CSV file

- Once we have explored the data, we are now ready to start using Datawrapper. The first step is to choose the map type that will show your data best. Select the choropleth map.
- Select the geographic region of interest. In this case it'll be U.S.A. states.



*Geographic region*

- Next, Click the Add your data tab. Click the Import button to start importing your data.

*Start import*

- Match the columns of the dataset, specify the corresponding column with ISO codes, in this case, the corresponding column is the GEOID column.

## Match your columns

Please select which column in your dataset contains 'ISO-Codes'.

MATCH  
AS  
ISO-  
CODES

MATCH  
AS  
ISO-  
CODES

MATCH  
AS  
ISO-  
CODES

MATCH  
AS  
ISO-  
CODES

MATCH  
AS  
ISO-  
CODES

MATCH  
AS  
ISO-  
CODES

MATCH  
AS  
ISO-  
CODES

MATCH  
AS  
ISO-  
CODES

MATCH  
AS  
ISO-  
CODES

MATCH  
AS  
ISO-  
CODES

MATCH  
AS  
ISO-  
CODES

STATE	STATE	AFFGE	GEOID	STUSF	NAME	LSAD	ALAN	AWAT	_state	percer
24	01714	04000	24	MD	Maryla	00	25151	69799	57939	1.161
19	01779	04000	19	IA	Iowa	00	14466	10841	72616	1.455
10	01779	04000	10	DE	Delawa	00	50459	13999	13729	0.275
39	01085	04000	39	OH	Ohio	00	10582	10268	11636	2.331
42	01779	04000	42	PA	Penns	00	11588	33945	32895	0.659
31	01779	04000	31	NE	Nebras	00	19895	13718	14286	0.286
53	01779	04000	53	WA	Washii	00	17211	12559	23099	0.463
72	01779	04000	72	PR	Puerto Rico	00	88688	49223		
01	01779	04000	01	AL	Alabar	00	13117	45933	39388	7.890
05	00068	04000	05	AR	Arkans	00	13476	29628	79546	1.593
35	00897	04000	35	NM	New	00	31419	72877	12739	0.255

☒ First row as caption

NEXT

### Match the columns

- The completed import should look like the screenshot below. At this point, the data can be visualized. Remember to specify the column with Geocoded data. This could be a name of a place e.g., state names.

Datawrapper

[+ New Chart](#)
[New Map](#)
[New Table](#)
[River](#)
[My Charts](#)

This map is in My Charts

1 Select your map

2 Add your data

3 Visualize

4 Publish & Embed

Now it's time to add data to your map.

Fill the table below with the values you want to visualize. You can add additional columns to the table by right-clicking. You can also import your dataset automatically by clicking on the import button below the table.

Geo-Code: Names 🔍

STUSPS	NAME	STATEFP	STATENS	AFFGEOID	GEO
NE	Nebraska	31	01779792	0400000US31	31
ND	North Dakota	38	01779797	0400000US38	38
NC	North Carolina	37	01027616	0400000US37	37
MT	Montana	30	00767982	0400000US30	30
MS	Mississippi	28	01779790	0400000US28	28
MO	Missouri	29	01779791	0400000US29	29
MN	Minnesota	27	00662849	0400000US27	27
MI	Michigan	26	01779789	0400000US26	26
ME	Maine	23	01779787	0400000US23	23
MD	Maryland	24	01714934	0400000US24	24
MA	Massachusetts	25	00606926	0400000US25	25
LA	Louisiana	22	01629543	0400000US22	22
KY	Kentucky	21	01779786	0400000US21	21
KS	Kansas	20	00481813	0400000US20	20
IN	Indiana	18	00448508	0400000US18	18
IL	Illinois	17	01779784	0400000US17	17

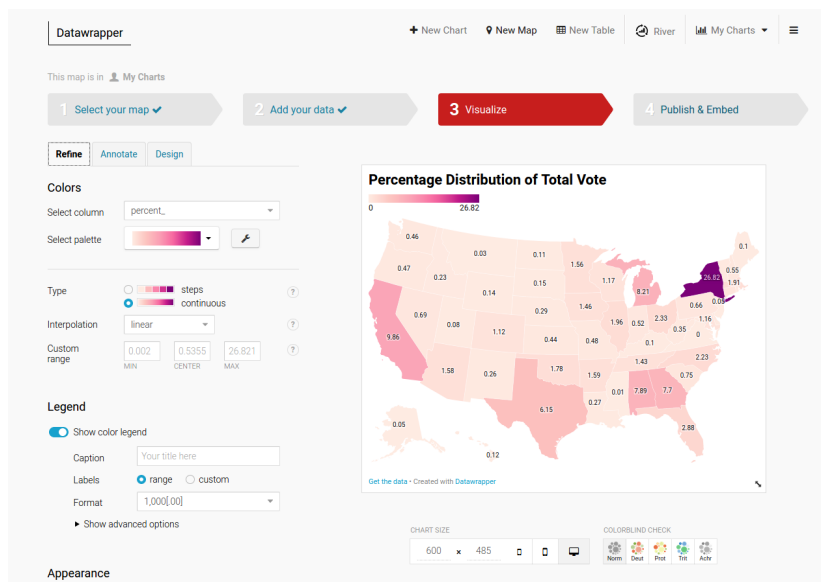
Import your dataset

Looking for the other data table?

Proceed

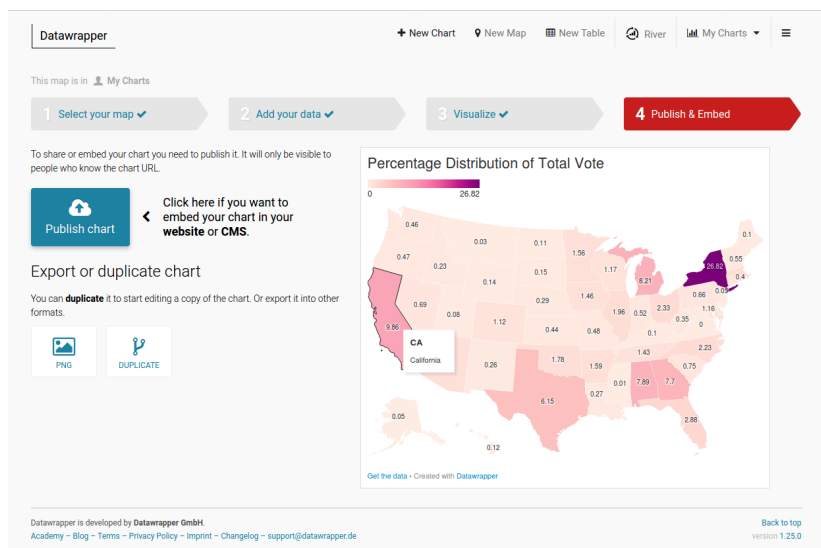
### Specify Geocode

- Visualize; Select column to visualize – in this case it's 'percent\_'. Under the Annotate tab, specify the Map Label.



*Select column, specify map label*

- To embed or share the map in a website, click Publish Chart. You will receive a shareable link and embedded code that allows you to add the map to your website.



*Embed or share map*

- By applying some analysis techniques using formulas and pivot tables, you're able to summarize and make sense of spreadsheet data. Additionally, you just made an election map from a spreadsheet. With this map, the vote distribution across states is visualized. It's also possible to use a desktop GIS application like QGIS

# MODULE 9 – GOING FURTHER WITH SPREADSHEETS

---

## Introduction

Spreadsheets are used extensively across fields all around the world, from financiers tracking stock movements, to governments monitoring the progress of contagious diseases, to individual companies tracking their Twitter presence. The power and flexibility of a spreadsheet is unparalleled when you consider the accessibility of its interface and functionalities.

In this module we will explore two functionalities of Google Sheets: the creation of dashboards and the use of custom functions.

## Description

The module will cover the following concepts:

- Dashboards
- Conditional formatting
- Data validation
- Google app scripts
- Custom functions
- Pivot table

## Skills

As part of this module, you will learn the following:

- How to create interactive dashboards
- How to make use of custom functions for specialized calculations

## Prerequisites

- Module dataset: <http://bit.ly/GCBAAsiaSample>
- A fair understanding of all previous modules
- A Google Sheet
- A working computer
- A modern internet browser
- Internet connection
- Basic knowledge of operating a computer

## Section 1: Creating a dashboard in Google Sheets

Once again, we will make use of our sample of the Global Corruption Barometer dataset: <http://bit.ly/GCBAAsiaSample>

A dashboard is a format of data presentation which aims to:

- Surface multiple insights on the same interface
- Make it easy to monitor trends
- Be easily shareable and printable

You will create dashboards if you plan to collect the same data over time and if you want to monitor something with the help of that data. Dashboards are meant to be dynamic (i.e., they receive new data regularly). If not, a dashboard is just an infographic. We will see spreadsheet-based dashboards can not only be dynamic but also interactive.

We will practice by putting together a dashboard made of three sections:

- A title section
- An interactive section
- A pivot table section

### Title section

The title section of our dashboard includes the title ("Survey Dashboard") and the number of days elapsed since the administration of the survey. This number is automatically updated.

	A	B	C	
1	Survey Dashboard			
2				
3	489	days since the survey		
4				
5				
6				
7				

### Walkthrough: Creating the title section of the dashboard

- After copying the module spreadsheet, create a new sheet.
- To make our dashboard more visually appealing we will start by changing some visual parameters of the sheet: Go to View → Gridlines to remove the grid of the spreadsheet.
- Now select the first 30 rows of the spreadsheet: click on the '1' to select the full row 1 then maintain shift while clicking on row 30 to select all rows between 1 and 30.
- Right click on the row number of any of the selected rows r to display the pop-up menu.
- From there, click on 'Resize rows 1-30' and change the value to 30 to give more breathing room to content.
- To center the text in the cells, select all the cells of the sheet and go to Format → Align → Middle

- We can write the title: in cell A1, write 'Survey Dashboard' then change the text formatting to make it stand out (in the picture the title uses a size 24 font, is in bold and uses the color dark red 1).
- Let's create the day counter: in cell A3, we will write:

```
=DATEDIF("2020/01/01",TODAY(),"D")
```

Where:

- **DATEDIF()** is the function calculating the difference between the two dates used as parameters.
- **TODAY()** provides the date of the day automatically (it refreshes whenever the spreadsheet is loaded).
- "D" is the parameter indicating that we want the time elapsed between TODAY() and 2020/01/01 to be in days.
- Because we're still in the title section, we can make the result stand out a bit more by increasing the font size and applying bold to the text (in the picture the text is size 18).
- Finally, we add the text in B3 to complete the title section: "days since the survey."

## The interactive section

This section presents some key figures at the country level. It relies on data validation in order to allow the spreadsheet user to pick the country that they want to explore as well as conditional formatting to highlight certain indicators.

D	E	F	G
<b>Country focus</b>			Philippines ▼
	Average age		35.71
	Share of rural respondents		48.00%
	Bribery in education		5.90%
	Bribery in health services		4.10%

### Walkthrough: Creating the interactive section of the dashboard

- In E1 write the section's title: "Country focus" and change its formatting to increase its visibility (the picture's text uses bold and a font size 18)
- Click on G1 then go to Data → Data validation.

**Data validation** ×

Cell range:

Criteria:

Tip: Use absolute references (e.g. =\$A\$1:\$B\$1) to lock rows & columns.

☒ Show drop-down list in cell

On invalid data: ☒ Show warning ☐ Reject input

Appearance: ☐ Show validation help text:

- Click on the little window to start selecting the range that you need. We will go to the 'Sample Dataset' sheet and select the whole COUNTRY column, which gives us:

```
= 'Sample dataset' !A:A
```

- The consequence of that action is that a dropdown will appear in G1 with each country being selectable. But we don't want the header row to also appear in that drop box, so we will tweak the formula to skip it.

```
= 'Sample dataset' !A2:A
```

- We now have the 'switch' that will make our dashboard interactive. To highlight it, you can change the background of the cell to a light grey.
- In the cells E3, E4, E5 and E6, write
  - Average age
  - Share of rural respondents
  - Bribery in education
  - Bribery in health services
- We will now write the formulas that will calculate those values based on the country displayed in G1.
- For the average age, the normal function would be AVERAGE(). But here we want to calculate the average of a specific country only (the one appearing on G1), which means that we need to add a condition. The appropriate formula for this scenario is AVERAGEIF(). We will then write:

```
=iferror(AVERAGEIF('Sample dataset'!A:A,G1,'Sample dataset'!C:C))
```

Where:

- 'Sample dataset'!A:A, is the COUNTRY column, used for filtering.
- G1 is the filtering condition used by **AVERAGEIF()**.
- 'Sample dataset'!C:C) is the range whose values will be averaged.
- IFERROR()** is there to clean up the error that appears when G1 is left empty.
- Now we will calculate the share of rural respondents, using a similar principle. But instead of finding the mean, we're trying to count the number of respondents for which the column AREA indicates 'rural.' The formula looks like this:

```
=iferror(COUNTIFS('Sample dataset'!D:D,"rural",'Sample dataset'!A:A,G1)/  
COUNTIFS('Sample dataset'!D2:D,"<>",'Sample dataset'!A2:A,G1))
```

Where:

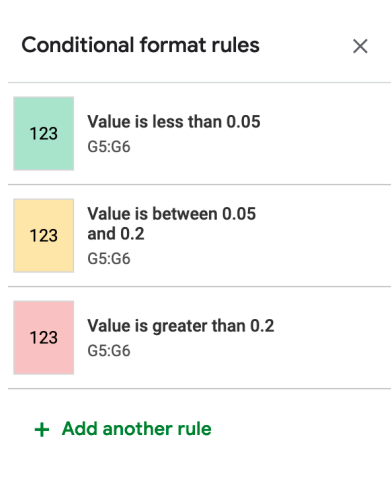
- COUNTIFS is the version of COUNTIF that accepts multiple conditions
- COUNTIFS('Sample dataset'!D:D,"rural",'Sample dataset'!A:A,G1) is a conditional count where the conditions are that the value in the column D of 'Sample dataset' should be 'rural' and the value of the column A of 'Sample Dataset' should be equal to G1.
- COUNTIFS('Sample dataset'!D2:D,"<>",'Sample dataset'!A2:A,G1) is a conditional count where the conditions are that the value in column D of 'Sample dataset' should be non-empty ("<>") and the value of the column A of 'Sample dataset' should be equal to G1.
- The process is repeated for bribery in education and health services, with one nuance: instead of filtering by a single value (e.g., rural) we must include all the positive responses in the dataset: "Once or Twice", "A few times", "Often." This gives the following formula:

```
=(COUNTIFS('Sample dataset'!G2:G,"Once or Twice",'Sample  
dataset'!A2:A,G1)+COUNTIFS('Sample dataset'!G2:G,"A few times",'Sample  
dataset'!A2:A,G1)+COUNTIFS('Sample dataset'!G2:G,"Often",'Sample  
dataset'!A2:A,G1))/COUNTIFS('Sample dataset'!G2:G,"<>",'Sample  
dataset'!A2:A,G1)
```

Where:

- COUNTIFS('Sample dataset'!G2:G,"Once or Twice",'Sample dataset'!A2:A,G1) filters the count by "Once or Twice" in the column G of the 'Sample Dataset' sheet as well as by the value of G1 in the column A of the 'Sample dataset' sheet.
- COUNTIFS('Sample dataset'!G2:G,"A few times",'Sample dataset'!A2:A,G1) is the same as above, but with a different filter.
- COUNTIFS('Sample dataset'!G2:G,"Often",'Sample dataset'!A2:A,G1) is also the same.
- Now let's add some colors: select the cells G5 and G6, which represent the share of respondents who have faced a corruption situation.
- Go to Format → Conditional Formatting and add three rules:
  - If the value is below 0.05, then change the background color of the cell to green.
  - If the value is between 0.05 and 0.2, change the background color of the cell to orange.
  - If the value is over 0.2, change the background color of the cell to red.





- You can now go to Format → Number and display the values as percentages.

Try changing the country in G1 using the dropdown to see the colors changing. The breaks we have chosen between the colors are arbitrary and were set for pedagogical purposes only.

## Pivot table section

For this section we will produce an alternative version of the pivot table we created during the 'Presenting Data' module. Instead of selecting the data and creating a pivot table in a new sheet, we will insert a pivot table directly in our dashboard.

9	<b>Age breakdown</b>						
10							
11	<i>COUNTA of Q30: t Grouped AGE</i>						
12	<b>Q30: the governm</b>	<b>18 - 32</b>	<b>33 - 47</b>	<b>48 - 62</b>	<b>63 - 77</b>	<b>78 - 92</b>	<b>93 - 107</b>
13	Agree	37.85%	35.65%	36.27%	36.60%	34.78%	66.67%
14	Disagree	23.89%	23.90%	24.15%	25.41%	26.09%	33.33%
15	Don't know	1.82%	2.06%	1.39%	1.81%	4.35%	
16	Neither agree or di	4.70%	6.79%	5.19%	5.92%	18.26%	
17	Strongly agree	16.82%	17.51%	17.77%	18.91%	9.57%	
18	Strongly disagree	14.91%	14.07%	15.23%	11.35%	6.96%	
19							

## Walkthrough: Pivot table section

- In A9 write the title of the section 'Age breakdown.' Increase the font size to make it a proper title.
- Click on A11 then go to Data → Pivot table. This time you will have to navigate to the 'Sample dataset' sheet to select the range for the pivot table. Then select 'this sheet' to have the pivot table integrated in the dashboard.
- The pivot table should be set up with:
  - Q30 as Rows
  - AGE as Columns
  - Q30 as Values, shown as % of column

- Q30 in the filter in order to filter out the blank cells
- Now we want to create age groups, instead of having each age treated separately. To do so, right click on the dark blue bar with the age and select 'set pivot group.'
- You can now decide how to group your age variable. Set the min value to 18 and the interval size to 15.
- Our last step is to reproduce the visualizations to the right of the table. This is a function called **SPARKLINE()**:

```
=SPARKLINE(B13:G13,{"charttype","column";"ymax",0.75;"color","lightblue"})
```

Where:

- **SPARKLINE()** is the function that produces a chart contained to the cell the function is in.
- B13:G13 is the range that **SPARKLINE()** is visualizing: specifically, this is the variation in the share of people who answered 'Agree' across age groups.
- {"charttype","column";"ymax",0.75;"color","lightblue"} are the parameters of the chart. By default, the chart is a line chart, but here we made it a bar chart (called column chart here).

	A	B	C	D	E	F	G	H
1	<b>Survey Dashboard</b>			<b>Country focus</b>			Philippines	
2								
3	<b>489</b> days since the survey			Average age			35.71	
4				Share of rural respondents			48.00%	New York
5				Bribery in education			5.90%	
6				Bribery in health services			4.10%	
7								
8								
9	<b>Age breakdown</b>							
10								
11	COUNTA of Q30: t Grouped AGE							
12	Q30: the governm	18 - 32	33 - 47	48 - 62	63 - 77	78 - 92	93 - 107	
13	Agree	37.85%	35.65%	36.27%	36.60%	34.78%	66.67%	
14	Disagree	23.89%	23.90%	24.15%	25.41%	26.09%	33.33%	
15	Don't know	1.82%	2.06%	1.39%	1.81%	4.35%		
16	Neither agree or di	4.70%	6.79%	5.19%	5.92%	18.26%		
17	Strongly agree	16.82%	17.51%	17.77%	18.91%	9.57%		
18	Strongly disagree	14.91%	14.07%	15.23%	11.35%	6.96%		
19								

The final dashboard

## Section 2: Using custom scripts in Google Sheets

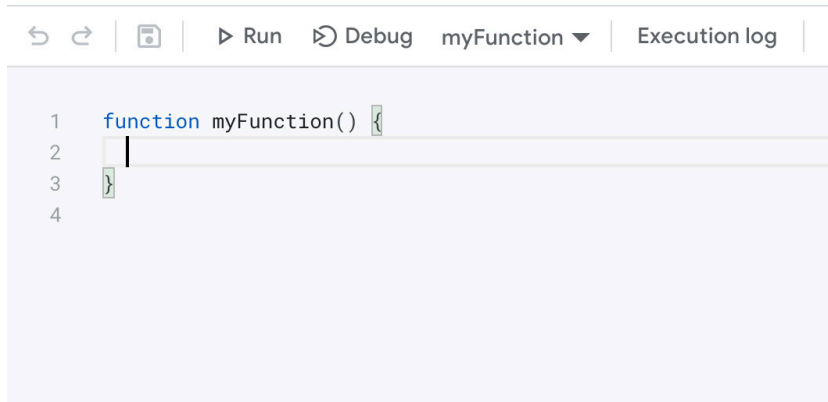
Google Sheets takes full advantage of the fact that it is an online tool to propose functionalities that rely on internet access. We saw one example during the GET phase where we used the **IMPORTHTML()** function. Another example of those internet-enabled functionalities is Google Apps Script – this is a functionality of Google Sheets allowing you to run scripts (that you have written yourself or found on the internet).

Those scripts could be used to simply implement complex functions that are not part of Google Sheets, but also to call on external APIs in order to create functions that pull data from elsewhere or do calculations that are normally impossible within a spreadsheet.

For the purpose of this module, we install a small script (found here: <http://bit.ly/scriptdistance>) that will allow us to calculate the distance between two locations.

### Walkthrough: Setting up a custom function

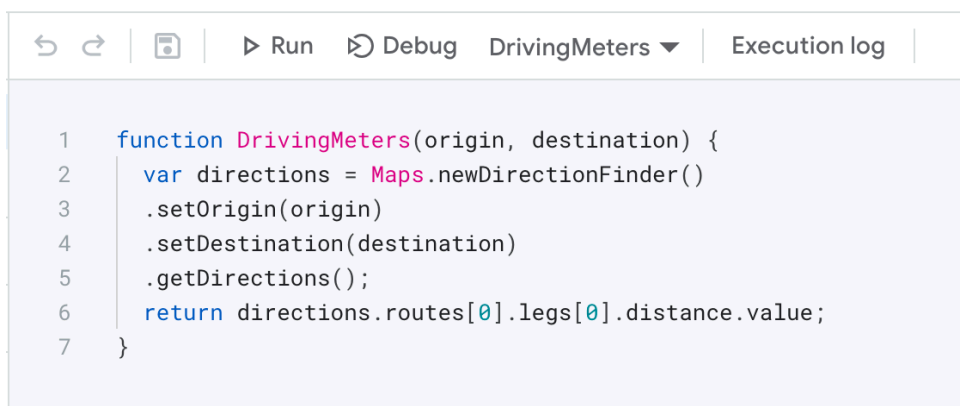
- Using any blank sheet, go to Tools → Script Editor



- You are now in the interface where you can create and manage your scripts. Copy the script below and paste it over the existing text.

```
function DrivingMeters(origin, destination) {  
  var directions = Maps.newDirectionFinder()  
  .setOrigin(origin)  
  .setDestination(destination)  
  .getDirections();  
  return directions.routes[0].legs[0].distance.value;  
}
```

- You can now save the script.



- Before the script becomes active in your spreadsheet, you need to run it once. There may be an error displayed in the log, but you can ignore it.
- What's left is simply for you to test the script by writing two locations just like you would write them in Google Maps. Then you can write your function and point it to the two locations. Note that custom functions do not have autocomplete, so you need to make sure you write them properly.

# MODULE 10 – USING DATA TO EVALUATE YOUR PROJECTS

---

## Introduction

Data plays an ever-increasing role in how civic projects are designed, planned, delivered, and evaluated. It does not magically improve the very manual acts that are key to many projects – visiting a family to check on their health, teaching a disenfranchised group about their rights, attending a municipal council to voice the concerns of the community. But data allows you to reinforce the effectiveness of the infrastructure of your projects – the activities that enable your project to optimize its reach, sustainability, efficiency, and impact.

Civic organizations and activists tend to be very focused on the people they are trying to help and the key deliverables of their projects. Working with data involves taking a step back in order to formalize your processes and generate different types of insights than the ones you get from simply delivering your projects.

There are many ways that data can play a role in your project, but a helpful way to understand the topic is to think about the three main phases of your projects:

1. **The planning phase** is when you use data to refine the design of your project (for example, with a baseline survey) and design your monitoring and evaluation strategy (e.g., with a scorecard system).
2. **The delivery phase** is when you get to make use of the systems set up during the planning phase – that may mean filling out the scorecard or, if your project includes training, analyzing the results of the pre-training survey to tweak your training agenda.
3. **The evaluation phase** which is when you may complete the final items of your scorecard or launch an evaluation survey that is adapted to the reality of the project delivery (rather than simply based on the initial planning).

For each of those phases we’re going to explore the role that data can play at the project level, but we will look at how the same phases can structure your approach to the delivery of a training, in order to give a more concrete example of how these concepts can be deployed.

## Description

The module will cover the following concepts:

- Baseline survey
- Pre-training survey
- Scorecards
- Post-training survey
- Outcome assessment
- Impact assessment

## Skills

As part of this module, you will learn the following:

- How to structure your project planning to incorporate data
- How to design pre- and post-training surveys

## Prerequisites

- Google Sheets
- A working computer
- A modern internet browser
- Internet connection
- Basic knowledge of operating a computer

## Section 1: Planning

### At the project level

Most civic work is spurred by events rather than data. People gather to fight for change because they have a shared experience of injustice or unmet needs. The normal behavior then is to address those needs directly: provide food where there is food scarcity; supply medicine where people are stricken by disease; organize communities whose rights are being ignored; protest decision-makers to highlight an unfair situation.

But taking action is not enough. To maximize the impact of the project, organizers with a strategic vision will take a step back: which village should be served by which food center? Can medicine be delivered strategically to prevent the spread of the disease, rather than simply playing catch-up? What skills or resources are disenfranchised communities missing in order to mobilize themselves? This is the strategic analysis (or assessment) process. And this is where data can play a role.

The role that data can play falls into two different categories:

**Baseline assessment:** this is the process of identifying the key indicators of success for your project and measuring their value before the start of the project. Doing a baseline assessment provides two benefits:

- It forces you to identify the key information that you need access to, or be able to measure, in order to complete your strategic analysis (should you measure the distance between food centers and famine-stricken villages, or the time for delivery instead? The number of people in the village? The time between deliveries across all villages?).
- It helps you evaluate the outcomes of your project by comparing the baseline assessment with a post- (or mid) project evaluation. Although the concept of 'evaluation' can be scary, it is important to do it in order to learn from your project in a systematic way and improve your future ones.

## Output, Outcome, Impact

Output, outcome, and impact are words that can often come up in casual discussion, and even more when discussing civic work. The meaning behind them will vary depending on who's talking, but there are generally agreed definitions when using them to evaluate a project:

**Output** refers to the deliverables produced by your project, be they activities (e.g., 30 people trained, 14 women increase knowledge) or products (e.g., 1 report, 2 action plans). Usually, outputs can be identified as soon as the activities have been conducted.

**Outcome** refers to what your project has achieved, beyond what was produced – they are often in line with your project goals (e.g., training participants assert their rights based on training materials, political participation among women has increased during townhalls or public meetings), but are not limited to them (e.g., participants trust each other more and will continue collaborating in the future). In many cases, measuring outcomes can only be done several weeks, months or years after the activity has concluded. Some time may be needed for the effects of your project to become visible.

**Impact** refers to the broader influence that your project has played on the main overall problem it was targeting (e.g., strengthening a country's democracy). Impact can only be measured over a longer period of time, and measuring it is tricky; your project is necessarily only one of the many events affecting a broader problem. Specifically in the context of "impact evaluation," impact may refer to a change in outcomes that can be attributed to your program; mathematically it is the difference in outcome for units that participated in a program relative to a "counterfactual," or what would have happened in the absence of the program.

**Analysis scaling:** it is one thing to decide on your mobilization strategy for one province. But what happens if you want to do it across dozens of provinces? You need a way to summarize and organize the information about each province to define priorities, match the right resources to each specific need and, more generally, track what is happening. Data can help with this, by giving you a structured view of disparate elements of your project or campaign.

## At the training level

Data also plays a role in the preparation for a training. Specifically, it does so in two preparatory activities:

**Pre-training assessment:** this is a questionnaire you send participants to get a sense of who they are, what they expect from the training, as well as logistical details (e.g., videoconferencing equipment if the training is remote). Typical questions cover:

- Identity (name, surname, job, etc...)
- Information about the organization they represent (if relevant)
- Detailed information about their work or any activity relevant to the training
- General questions about data literacy (e.g., do you usually work with data?)
- Their expectations (e.g., what they want out of the training, or the part of the program they would like to focus on)
- Logistical details

**Pre-training quiz:** this is a questionnaire used for assessing the level of data literacy (or any specific proficiency that you want to measure). The pre-training quiz is only necessary if your training is technical in nature and if you're unsure of the baseline capacity of the participants. It would have little value for a training on making freedom of information requests, for example: participants might not be expected to have any prior knowledge of this process.

However, pre-testing becomes valuable when you want to anticipate the possibility of having people of very different skill levels. Having to discover this during the training will leave you unprepared. The pre-training quiz's counterpart is the post-training quiz, which will help you understand how much the participants have learned.

### Data Minimization and Security

Collecting data about your participants is a balancing act: some information that may appear essential to your planning may also endanger participants in certain political contexts. Maintaining awareness of that risk is essential when designing a plan that is both successful and safe for your beneficiaries. There are two strategies to keep in mind:

**Data minimization** is the process of collecting as little data as possible on your participants/beneficiaries. It requires you to have a clear idea about project plan: what are you trying to do? Is that specific data point necessary for this? Is that level of precision necessary? Can alternative, safer, data points be used to achieve the same goal?

**Data security** is the process of ensuring that the data you collect cannot be directly accessed or reused by anyone other than the people you granted access. Data security is a field in itself, but you should research simple best practices around:

- Data anonymization, which is the process of stripping a dataset from personally identifiable information.
- Access policies, to make sure that the computer or server you store your data in is not unknowingly accessible to anyone.
- Device security, which defines how secure the devices you store your data on are. This includes making sure that the devices are up to date, that they have a passcode login, etc.

## Section 2: Project monitoring

### At the project level

The monitoring of your project is the process of tracking metrics that help you guide your actions and adjust if needed. It requires you to define target metrics that are good indicators of your project's progress. Interesting metrics can include:

**Countdown to target:** if your project has a specific target number (of signatures, events, people to train, etc.) to reach, then the percentage of progress toward that goal is a simple metric to track.

**Rate of change:** if there is not a clear target value, the percentage of progress becomes meaningless. Instead, it may be more interesting to track the rate of change, which tells you if the project is

progressing steadily or faster in certain regions than others. For example, even if you do not have a target number of social media interactions, you could measure if the number of interactions is stable week over week.

**Time for completion:** if your project includes activities which are time critical, such as the delivery of medicine or the gathering of signatures for a petition, measuring the time needed to accomplish the task allows you to see if some areas require additional resources.

**Real-time feedback:** getting regular feedback about a campaign may allow you to identify if your message is getting through or if some adjustments are needed. This can be especially useful for months-long campaigns, as the social and political environment may change between the time you designed the campaign and the time it is deployed. Because the value of this metric is the speed at which you can measure it, you should limit your polling process to a couple of questions that can be answered with yes or no, (e.g., do you agree/disagree with the following statement) making it easy for you to process the results and act on them.

## At the training level

During a training it may prove useful to get **real-time feedback**: a short anonymous survey at the end of each day, with a few questions (prepared in advance) allows you to adjust your training agenda or methods from one day to the next. Although fundamental elements of the training should not change, this kind of survey is useful to identify if some participants are not able to keep up with the pace of the training or if some concept was not understood properly.

### Example: Daily feedback survey

#### Q1: Which day(s) of the training have you completed?

Question type: Multiple choice

Response data type: Categorical

Potential answers: Day1, Day 2, Day 3

#### Q2: The content covered by the training today was useful for my work.

Question type: Multiple choice

Response data type: Ordinal

Potential answers: Strongly disagree, disagree, neither agree nor disagree, agree, strongly agree

#### Q3: Do you have any improvements to suggest for the next session?

Question type: Open response

Response data type: Qualitative (unstructured)

Potential answers: Open response



## Section 3: Project evaluation

### At the project level

At the project level, evaluation refers to the assessment of the success of the project, measured against its initial goals. Evaluation can take many forms, and bigger organizations often hire research teams to perform external evaluations of key projects. But a smaller organization has several avenues to generate data that can be used for evaluating the project:

**Post-project assessment:** using the same criteria as the baseline assessment (described in the planning phase), the project coordinator can get a before and after view of the situation and deduce the project's outcomes. A scorecard system (which is essentially a spreadsheet) can be used to compile the information.

#### Understanding Scorecards

A scorecard is essentially a spreadsheet aimed at synthesizing multiple indicators in order to create an overview score. For example, a project aimed at mobilizing citizens to increase voter turnout may be broken into a dozen smaller goals, each with their own indicators and evaluation criteria:

##### Goal 1: create more awareness around voting rights

- Indicator 1.1. number of people reached through social media campaigns
  - Evaluation criteria: success = 50,000+ people reached
- Indicator 1.2. number of workshop attendees
  - Evaluation criteria: success = 300+ workshops attendees
- Indicator 1.3. number of people mentioning having heard of the campaign during surveys
  - Evaluation criteria: success = 20+% of survey respondents

##### Goal 2: organize 50 village meetings

- Indicator 2.1. number of village meetings organized
  - Evaluation criteria: percent of progress

The indicators 1.1., 1.2., 1.3. and 2.1. would then be given a numeric value (either a progress percentage or a 0 or 1 value indicating failure or success) which would be combined in a final dashboard (see module 9 on dashboards) with summary insights (e.g., 75 percent of the objectives have been successfully reached).

The exact shape of the scorecard varies a lot between projects and will have to be adjusted for your specific needs.

**Stakeholder interviews:** although interview transcripts are initially unstructured data, through coding and other forms of qualitative analysis and transformation, these data can be converted to machine readable formats. And whether structured for analysis or unstructured, they can be an important part of the evaluation process to get a nuanced understanding of the project outcomes. Researchers commonly transform interview transcripts into structured datasets through a process called coding – they review the content with specific criteria in mind (Did they mention a specific, predetermined theme? Do they

show signs of confidence? Do common unexpected themes emerge across interview respondents? Do respondents respond positively or negatively to a theme?) and either extract the relevant text from the transcript or encode the results into specific values (e.g., 0 for no mention of them, and 1 if mention of that theme).

### **At the training level**

At the training level the evaluation can consist mainly of a post-training quiz which should be designed similarly to the pre-training survey mentioned in Section 1; even if the exercises are different, the concepts evaluated should be the same, so that analysts can determine whether participation in program activities increased knowledge or capacity.

The post-training quiz can be accompanied by a post-training survey and/or participant interviews to get more nuanced feedback.

