Big Data, Big Questions

Utrecht University

April 29, 2020

Table of Contents

First steps in Excel

- Many of you have used Excel before
- It is not a particularly user-friendly program
- Excel makes use of a **standard repertoire** of functions, and a **more advanced** set of operations called macros.
- We only concern ourselves with the **standard repertoire**

First steps in Excel [2]

- ▶ You can also use **Google Sheets**, an online version of Excel.
- ▶ If you have a Google Account, it is accessible under Google Drive.
- ▶ I will be using this version mostly throughout the tutorial series.
- ▶ It is more user-friendly than the normal Excel because it shows the syntax of the function you are using in a pop-up menu.
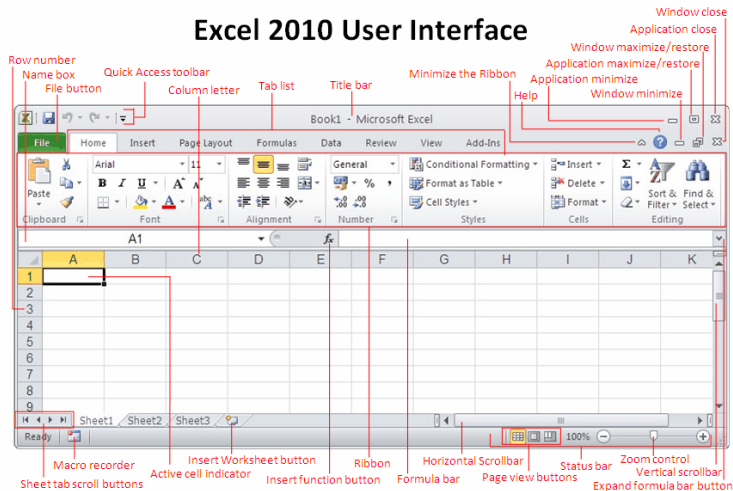
User interface

Figure 1: Source: Excelhowto.com

Three very important habits

- Don't use sheets.
  Save every separate piece of data in a separate file
- Put ONLY data in the cells.
  Notes etc. should be relegated to other files
- Save your files in .csv format, and not in .xlsx

Simple calculations in Excel..

Example

Try to type the following in cell A1:

```
= A2 + A3
```

And try to enter numbers in A2 and A3. Obviously, we can change the formula to do something more complicated. Let's try!

Simple calculations in Excel.. [2]

Example

Try to type the following in cell B1:

$$= \text{COUNTA(A:A)}$$

and insert a few random numbers in cells A1 to about A10 (decide the amount of cells for yourselves.)

- ▶ What do you think `COUNTA` does?
- ▶ What do you think its *argument* `A:A` means?
- ▶ So how do you select an entire column in Excel?

Assignment!

Try to write a function in cell B1 that determines the **mean** of an unknown
amount of observations in column A
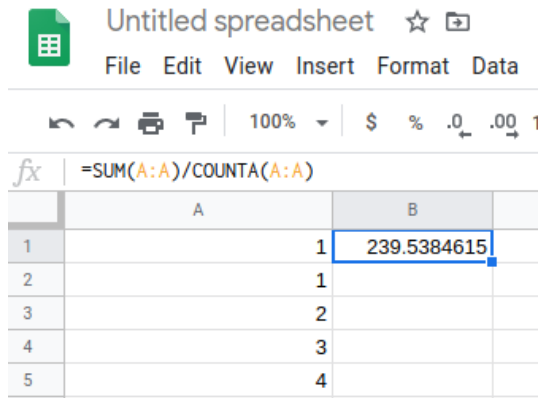Note to self: Solution on next slide

Solution

Figure 2: Source: My own

Data structures

- Usually, people talk about *variables* and *observations* when talking about data.
- What is a variable?
- What is wrong with the following *dataset*?

	A	B
1	steden & bevolking	amsterdam
2		
3		1.6
4		
5		rotterdam
6		
7		0.76
8		
9		utrecht
10		0.3
11		
12		eindhoven
13		
14		0.25

Figure 3: Source: My own

Clean data

Requirements

In general, most statistical analyses programs and programming languages require your data to be in a particular format before they can process your data to make graphs/tables/conduct analysis.

Clean data, Tidy data..

The data should have *observations* in rows, and *variables* in columns. Data in that format is called *clean data* or *tidy data*, whereas data that is not in that format is called *raw data* or *untidy data*.

Transposing data

Cleaning data

Cleaning data is one of the hardest things to do, at least, depending on the format your raw data is in. We will start with doing one of the easier things, which is called *transposing*. This will be useful when your data has observations in XX and variables in YY.

Question!

What is XX and YY? Pick between rows and columns.

Transposing data

Cleaning data

Cleaning data is one of the hardest things to do, at least, depending on the format your raw data is in. We will start with doing one of the easier things, which is called *transposing*. This will be useful when your data has observations in **columns** and variables in **rows**.

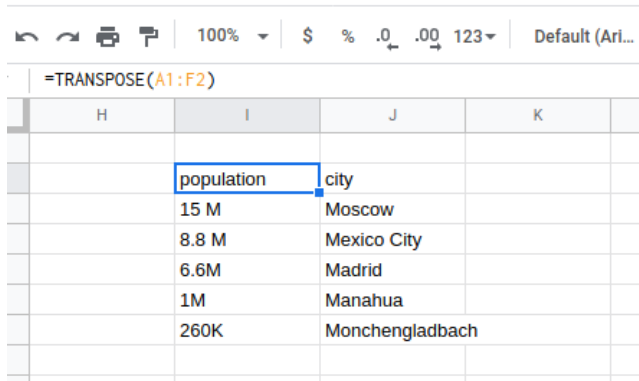## Question!

What is XX and YY? Pick between rows and columns.

Assignment!

Try to transpose the following dataset:
Link to Google Sheet

Copy this to your own spreadsheet program/Google sheets session

Solution

Figure 4: Source: My own

Data types in Excel

There are various data types in Excel. It is useful to know which is which, because:

- ▶ Default layout can differ, which can cause confusion
- ▶ You know what kind of output/input you must deal with
- ▶ When entering data, it is important to make choices about the **measurement level** of the data.

Logical

The most simple variable

The most simple variable in Excel is a logical. It returns either TRUE or FALSE.

Example

```
A1 = 3
A2 = 3
A3 = A1=A2
```

What will A3 return?

Treating logicals

Suppose you would want to write a logical condition like this: You have data on various countries' of the world's GDP, but you want to add only the GDP of countries in Europe. You construct a logical vector like this:

Figure 1: Source: My own

How would you construct a function that does this?

Answer

Answer:

```
=SUMIF(B:B, TRUE, C:C)
```

Number

Numbers are easy, but..

Numbers are the most natural and easy-to-understand class of cells in Excel. Often, you might want to enter numeric variables, and Excel will recognize them as such. Sometimes, however, something which you might not want to be treated as a number will end up as a number.

Hint

Sometimes, when a number is very large (or has many decimals), you will see the string ####### pop-up your screen. You should widen the column a bit if you want to see the entire number.

Number

Another hint
Alternatively, you might want to use the the tools in the toolbar (find out where!) to adjust the number of decimals displayed.

Different representations
You might also want to set your cells to be a certain category, such as number, percentage, currency, fraction, or scientific. You can even change it to date or text in the menu adjacent to the decimal settings.

Measurement levels
One stores **interval** and **ratio**-level variables in numbers.

Text

Text
Text is also a very basic class. Text can often be used to store **nominal** and **categorical** variables.

Question
Enter five numbers in cells `A1:A5` in Excel. Now, set the cells to text. Attempt some calculation with these five cells. What happens? What happens to the formula you've used to do the calculation?

Date

- ▶ You have to be very careful with dates.
- ▶ Oftentimes, you don't have to store years as a date.
- ▶ Better store it as text instead.
- ▶ Only if you have data with ymd or times, you are advised to use this class.
- ▶ If you convert dates to another type of variable, information can be lost!

Formulas and values

Another typology

Try to enter $15 * 10^6$ (15 million) in cell A1, and 1 in A2. Now try to add those. What happens?

Values (numbers) are generally raw numbers. For Excel, $15 * 10^6$ is not a number, whereas $= 15 * 10^6$ is.

Formulas are instructions for Excel to perform calculations.

Formatting the cells

Try to do the following:

1. Enter `1` in A1
2. Enter `=A1+1` in A2
3. Now drag down A2 below to (about) cell 10.
4. Copy Column A to Column B.

What do you see?

Figure 2: Source: My own

You can see that cells B2:B10 have been filled with =B2+1, =B3+1, etc.

Suppose you want to select only the values

Click on the small symbol Paste Options -> Values only

Warning! The formulas will be replaced (and lost if you don't save them)
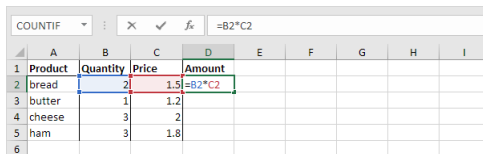
Relative References

Relative references

What we have just used are *relative cell references*. Relative references are used by default in Excel.

Copying cells

If you copy a cell using relative references, the function will be defined on the relative basis of those cells (so the object of the function will be determined by its position *relative* to the original cell.
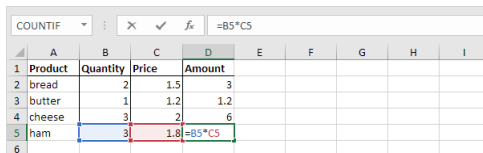
Relative References

A small example:

Figure 3: Source: excel-easy.com

Figure 4: Source: excel-easy.com

Absolute References

You can also use absolute references. These are used by putting a dollar-sign (\$ in front of either a column number or a row number, or both:

```
=$H3    #H is absolute, 3 is relative
=H$3    #H is relative, 3 is absolute
=$H$3   #Both H and 3 are absolute
```

Some Examples

Figure 5: Source: excel-easy.com

Figure 6: Source: excel-easy.com

Excel's Syntax

Now, you know how to navigate and aim your functions on a particular sheet.
You already know the following:

1. To select many cells in the same column, you use `A1:A2`
2. To select many cells in the same row, you use `A1:E1`
3. To select an array (a subtable), you use `A1:E5`
4. You separate arguments for various functions using `;`
5. The difference between absolute and relative references

Excel's Syntax

Some other important things which you need to know:

- ▶ Refer to another sheet using: `Sheet2!A1:A5`
- ▶ You can use logical operators: `>`, `<`, `!=` etc.
- ▶ You can refer to another file using:
  `=[OtherExcelFile.xlsx]Sheet1!$A$1:$A$10`

Excel in other languages

If you use other language-versions of Excel:

- ▶ Functions have a different name! `MEAN`, `GEMIDDELDE`, etc.
- ▶ This can be really annoying if you are used to one version.
- ▶ More importantly, some versions of Excel use comma's (`,`) to separate function arguments, others use semicons (`;`)
- ▶ Pay attention when copying functions from Google

Usage of filter

New dataset
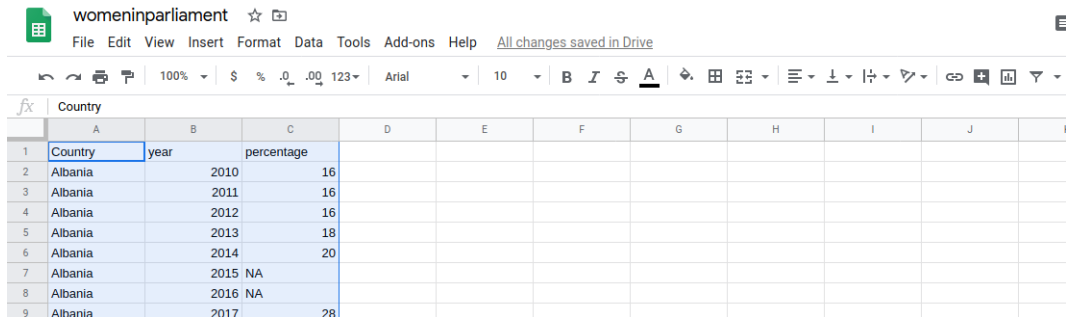
Try to load this dataset about the proportion of women in parliament in several European countries: **click**

Try to copy and paste the data to your own Excel session (or your Google Sheets session).

Usage of filter

We can try to look at the observations from one particular country or year using the filter function.

Move your mouse cursor to the 'Country' cell, and select the following button:

Figure 1: Source: My own

Filtering..

Press clear (meaning no countries are selected), select a country of your liking, and have a look!
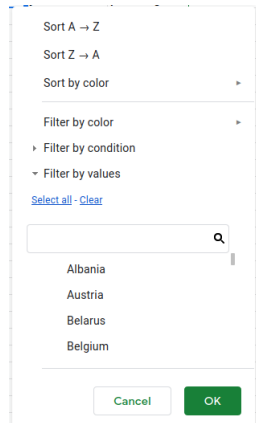
Figure 2: Source: My own

Calculations by group

- Let us now proceed to do some more complicated calculations.
- For example, suppose we want to know the mean women representation in every year in the dataset.
- How would we go about this?

Calculations by group [2]

Let us first reset the filter! Simply click the button again, and the filter will be reset.

Now, let us find what years are present in the dataset, so we can make a small table of all years:

```
=MIN(B:B)
=MAX(B:B)
```

These turn out to be 2010 to 2017. A very easy way to make a mini-table containing this information is to write 2010 and 2011, select both cells, and drag them down.
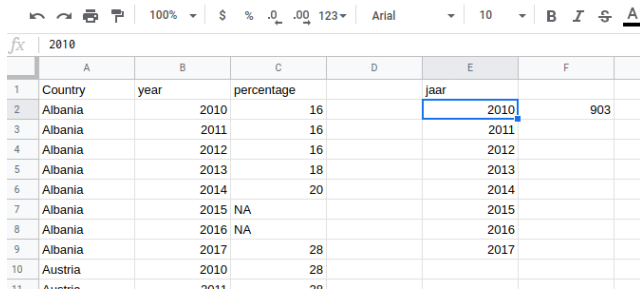
Calculations by group [3]

We will use the `SUMIF` and `COUNTIF` functions to perform calculations per group.

- ▶ `SUMIF` has the following syntax: =`SUMIF`(Cells on which to apply a criterion, the value of the criterion, Cells on which to apply the sum)
- ▶ In our case, this looks like this: =`SUMIF(B:B,2010,C:C)`

Calculations by group [4]

If we make put the years in cells starting at E2, like this:

Figure 3: Source: My own

We can also "parametrize" the year, and use the 'drag down' feature to perform these calculations automatically!

Parametrization

If we use the following code:

```
=SUMIF(B:B,E2,C:C)
```

we make sure that we change the criteria from 2010 to 2011, 2012, etc. as we use the drag down function.

E2 automatically changes to E3, E4, etc., and then takes the values of 2011, 2012, etc. This is smart use of *relative* cell reference.

Examples of absolute and relative cell references

Example

Let's use some example data to demonstrate the power of absolute and relative cell references in the context of a large dataset.

Dataset on Wars

Here is a dataset of (virtually) all major wars in history and estimation of their casualties, and some other info.

Freeze panes

- First, in a large dataset such as this one, we might want to browse the data, and not lose sight of the column that contains the variable names.
- Otherwise, we might be looking at a number, and forgetting what the variable in question is.
- At the top of your sheet, go to View -> Freeze -> 1 Row (Google Sheets)
- View -> Freeze Top Row (Excel)

Indicator variable

Death count larger than X

Suppose we want to make an **indicator variable**, indicating whether a particular war has had an estimated death count > a certain number, which we want to specify later.

- ▶ What do we do?
- ▶ First, put a number, say, 100.000, in cell I2
- ▶ Now, attempt to enter the following function in H2, and drag it down:

        =IF(G2>I2;1;0)

- ▶ Why doesn't it work?

Combining absolute and relative references

We must combine absolute and relative cell references in order to make it work:

```
=IF(G2>$I$2;1;0)
```

Filter

Use the filter function to investigate which wars have had more than 100,000 casualties.

Play with the I2 cell and try to find out which wars have been least and most deadly of all.

Copy this dataset to your Excel session

Let's take the **Dataset about historical wars.**

Let's try to separate the two opposing sides of the wars, and put them into separate columns, and find out which polity was involved in war most often of all.

Editing strings

- ▶ MS Excel has the feature of extracting parts of strings (text) using the `=LEFT,` `=RIGHT` and `=MID` commands.
- ▶ The syntax of those commands is as follows:

  ```
  =LEFT(STRING,FIRST X CHAR)
  =RIGHT(STRING, LAST X CHAR)
  =MID(STRING, START, HOWMANY)
  ```

- ▶ We will cleverly combine them with `FIND` to extract relevant parts of the strings.

Separating the belligerents

- ▶ After inspection of the data, you have probably noticed that belligerents are separated by `vs..`
- ▶ Let us make use of that to find the occurrence of `vs` in column D.
- ▶ `=FIND("vs", D2 ,1)`
- ▶ Always put strings between quotation marks

Assignment

Try to extract the names of the two belligerents for yourself using the `LEFT` and `RIGHT` functions

Hint: the `LEN` function captures the full length of a string

Solution

```
=LEFT(D2, J2-1)

=RIGHT(D2,LEN(D2)-J2-3)
```

▶ If you were to start from LEN(D2-J2+1), you would extract everything starting from vs.
▶ The string is split up into two parts, one up until (J2-1) and one from LEN(D2-J2+1)
▶ But, you do not want to take vs. along in the 2nd string.
▶ Therefore, you extract 4 characters (including the space!) less

Find and replace

Find and Replace

Alternatively, you can use the find & replace function to replace parts of a string which you want to get rid of.

For example, if you were only able to split the string in half, you can select the entire column, and replace vs. by nothing.
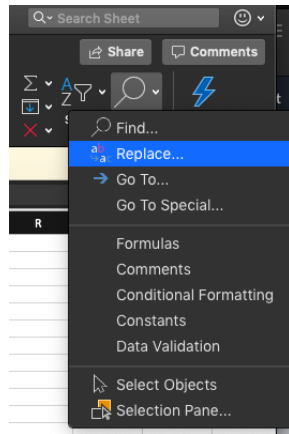
Figure 1: Source: My own

Vlookup

- In this tutorial, we will merge two datasets on the basis of a common identifier.
- We will learn what an identifier is, and how we can use it to match certain observations to certain others, even if they are not in the exact same order.
- We will use a dataset about Olympic Games medals to learn how to merge a dataset with obtained medals in the Summer Olympics and in the Winter Olympics for each country.

Tutorial: How to merge datasets

Tutorial
We will use our skills, in particular, doing calculations over arrays, and our knowledge of the `VLOOKUP` function to merge two datasets on the basis of an **identifier** variable or **key**.

What is an identifier?
An identifier is a variable, present in both datasets, on the basis of which you can match an observation in the first dataset with an observation in the second dataset.

Example of an identifier

Have a look at these two (mini)datasets:

Short	ID	Gold	Silver	Bronze	Total
AFG	0	0	0	0	0
ALG	3	0	0	0	0
ARG	19	0	0	0	0
ARM	7	0	0	0	0
ANZ	0	0	0	0	0
AUS	19	5	5	5	15
AUT	23	64	81	87	232
AZE	6	0	0	0	0
BAH	0	0	0	0	0
BRN	0	0	0	0	0
BAR	0	0	0	0	0
BLR	7	8	5	5	18
BEL	21	1	2	3	6
BER	8	0	0	0	0
BOH	0	0	0	0	0

Figure 2: Olypmic Medals (Summer)

Short	ID	Gold	Silver	Bronze	Total
AFG	14	0	0	2	2
ALG	13	5	4	8	17
ARG	24	21	25	28	74
ARM	6	2	6	6	14
ANZ	2	3	4	5	12
AUS	26	147	163	187	497
AUT	27	18	33	36	87
AZE	6	7	11	24	42
BAH	16	6	2	6	14
BRN	9	2	1	0	3
BAR	12	0	0	1	1
BLR	6	12	27	39	78
BEL	26	40	53	55	148
BER	18	0	0	1	1
BOH	3	0	1	3	4

Figure 3: Olympic Medals (Winter)

Question

What would be the identifier in both datasets? Why?

What would you do to merge one particular column of the second dataset with the first dataset?

Answer

Vlookup

You would use the `VLOOKUP` function. Recall that `VLOOKUP` has several arguments: the variable you want to look up, the table in which you want to look, the column number of the table in which you want to look, and an (irrelevant) literal match option.

Examplary Answer

Suppose the identifiers would be in column A, the data in column B and C, and you want to add the second column from the other dataset based on the other identifier in column D. You will specify in cell D2:

```
= VLOOKUP(A2, Otherdata!(A:B), 2, FALSE)
```

Practice

Let's now try this in practice. Through **this link**, you will see the spreadsheets shown before. Your task is to merge the two datasets to Sheet 1.

I will give you 10 minutes try-out time, and then go through the steps required to do this in the next slides.

Solution

- ▶ Let us first rename the variables Gold, Silver and Bronze to Summer_Gold, Summer_Silver and Summer_Bronze.
- ▶ Let us create three new columns, Winter_Gold, Winter_Silver and Winter_Bronze.
- ▶ In cell F2:

  ```
  = VLOOKUP(A:A, Winter!A:E, 3, FALSE)
  ```
- ▶ Now, let's 'drag down' cell F2 along the entire column F.

Solution [2]

Now, we have extracted all Gold medals in the Winter Olympics for each country.

▶ We also want to extract the silver and bronze medal count. Let's drag cell F2 to the right, to G2. What happens/

▶ We obtain:

```
= VLOOKUP(B:B, Winter!B:F, 3, FALSE)
```

▶ This is wrong! We still want to still look up `A:A`, and we want to look it up in `Winter!A:E`.

Solution [3]

▶ As we know, this can be solved using 'absolute cell references': enter in F2:

```
=VLOOKUP($A:$A, Winter!$A:$E, 3, FALSE)
```

▶ In this case, if we drag to the right, we will obtain in cell G2:

```
=VLOOKUP($A:$A, WINTER!$A:$E, 3, FALSE)
```

▶ We therefore have to manually change 3 to 4 to extract the silver medals.

▶ Afterwards, we change it to 5 to extract also the bronze medals.

▶ Dragging down the cell in G2 and H2 will accomplish our task and merge the datasets.

Introduction

- In this lecture, we will proceed to do one of the more fun things in Excel, creating graphs!
- Creating graphs can be really nice, but it is also dangerous.
- Oftentimes, people create meaningless or bad graphs, containing either the wrong information, too little information, or too much information.
- To strike the right balance, use your common sense. Can I explain the content of this graph within 1 minute to my parents? Is this graph intelligible?

Bad graphs

► Some examples of bad graphs..
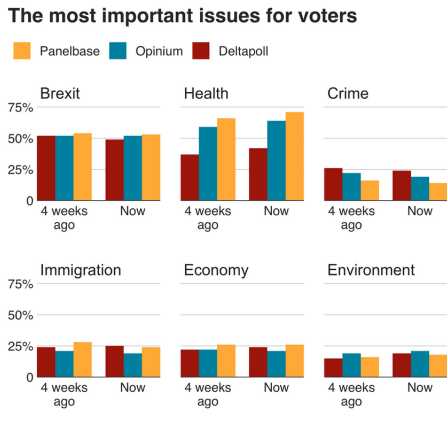

Figure 1: Source: My own

Bad graphs [2]

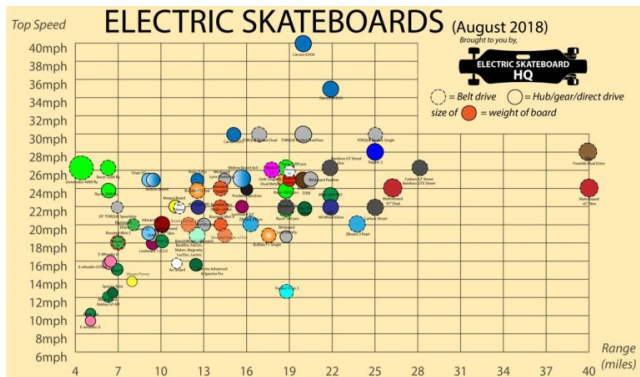Figure 2: Source: twitter.com/BBCBadGraphs

Bad graphs [3]

Figure 3: Source: Here

Constructing your own graphs

We will concentrate on creating our own (good) graphs.

Let us start out with some basic graphs, which are also the most used: tracking the development of a variable (Y) over time (X), and tracking the development of a couple of variables ($Y_1, \ldots, Y_n$ over time).

Alternatively, a line graph could track the same variable over time for a number of different categories. We will see an example of this soon.

Line graphs

Line graphs

There are a multitude of other graphs you can make using Excel, but of the most often-used is a line graph.

Line graphs can also be abused by adding to many lines to the graph.

A good general principle is to make sure each individual line is tractable.

If you have too many categories, you should find another way of communicating your data.

Creating a line graph

Line graphs are very easy to make.

1. You select two columns out of a data series.
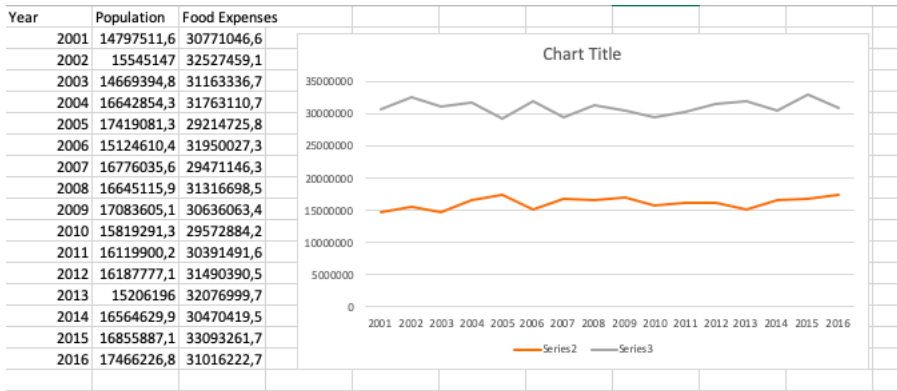2. Insert -> Line Symbol -> 2D Line

Short Example I

Figure 4: Source: My own

Short Example II

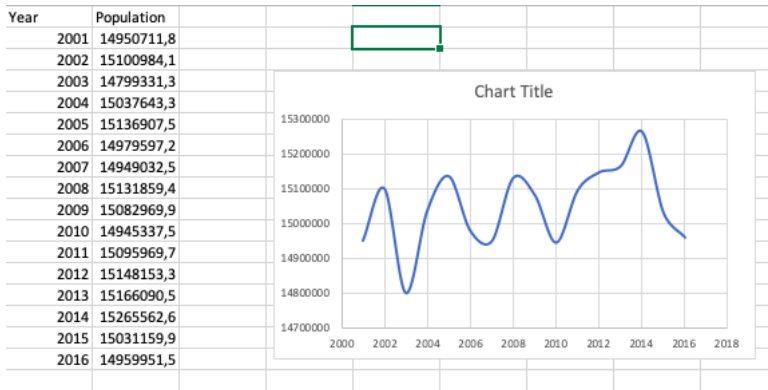You can create X-Y line graphs through X Y Scatter.

Figure 5: Caption

Scatterplots

Scatterplots are very similar to line graphs, expect that the observations might not be connected to each other through time. A short example:
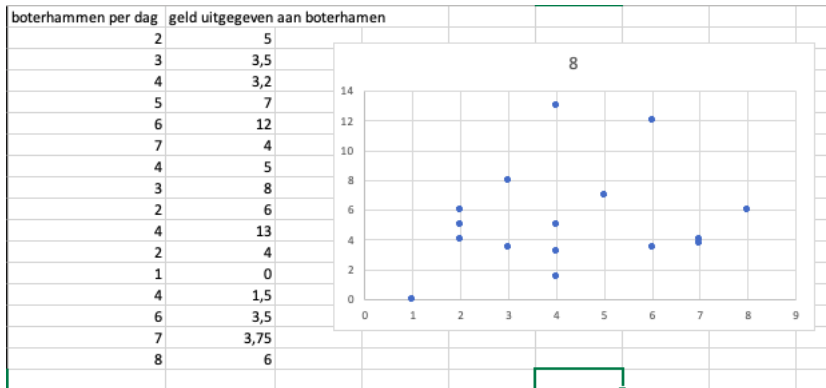

Figure 6: Source: My own

Manually adding the data

Problems..
Sometimes, Excel understands which graph you want to make automatically. Other times, it doesn't. Then, you have to add the data manually to the graph.

How to do this?
You create the graph as before, right-click -> Select Data..

Manually adding the data

Figure 7: Source: My own

A short demonstration

1. Let's open the Coronavirus database **over here**
2. Filter the data on Belgium and Netherlands
3. Put the two next to each other in a New sheet (CTRL-X and CTRL-V)
4. Let's focus only on the Confirmed cases..

Selecting The Data

- ▶ Let's attempt to draw a graph using the tools at our disposal.
- ▶ Select the 'Date' and 'Confirmed' columns for both countries
- ▶ You can name the columns 'Belgium' and 'Netherlands'.
- ▶ Now go to insert, charts, 2D-line, and select the first option.
- ▶ You will end up with the following:

A bad graph

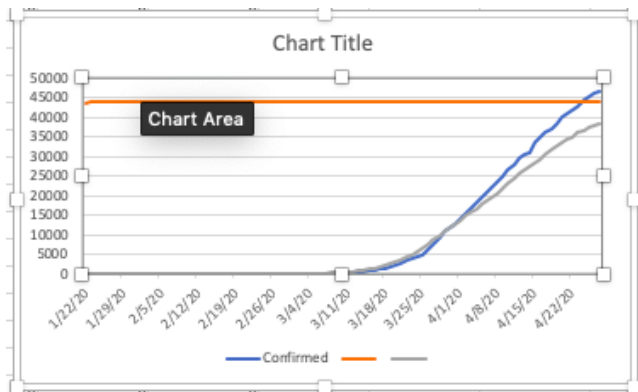Not quite what we wanted. How to proceed?

Figure 8: Source: My own

Selecting the data

Right click the graph > Select data:

1. You will see a couple of legend entries.
2. The first one is called Belgium. If you look at the Y-values and X-values, you will see that they are correctly specified.
3. The second entry, Netherlands if you gave it a name, also has a correct correspondence between X- and Y-values.
4. The third entry, however, doesn't. Delete it by clicking on the minus-sign.

Final graph
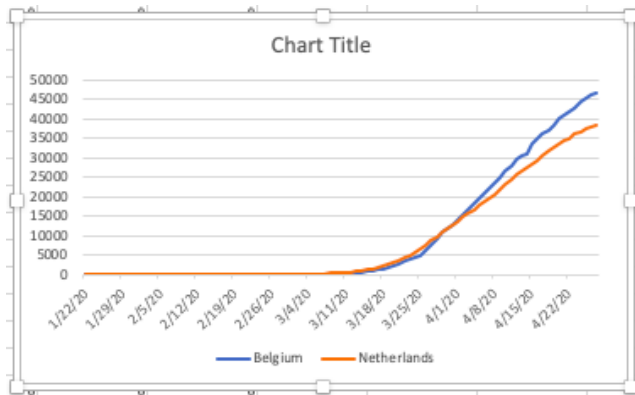
Then, the graph will look like this:

Figure 9: Source: My own

Tutorial

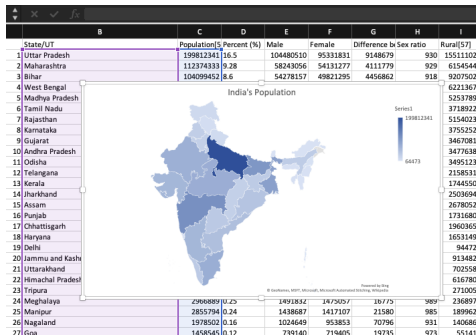In this tutorial, we are going to learn how to make maps of this kind:

Figure 1: Source: My own

We will see that it is surprisingly easy to do so! All we need is clean and tidy data!

Start with the data

You can find the dataset, which I took from **this Wikipedia page** over here:

Click!

Copy this to your own Excel session, and delete the last observation (total of India) from your file.
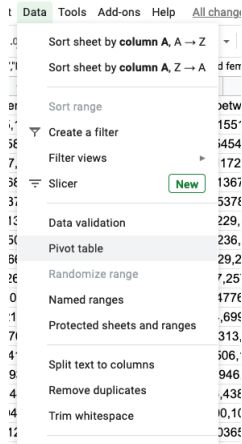
Ordering the data

Why does it look so strange?
The data seems to be in a messy format: not each observation seems to be in its own row. What is the reason?

.csv files
.csv files are sometimes recognized automatically by Excel (and Google Sheets), but sometimes they aren't. In that case:

Split to columns

Figure 2: Source: My own

- ▶ You select the entire column of cells you want to split.
- ▶ You go to the tab Data
- ▶ (Split) text to columns
- ▶ You manually select the sign that separates the cells
- ▶ in .csv, this is always a comma
- ▶ **Comma** separated values

Attempt to make the map

Let's attempt to make the map now.

1. Select the columns containing the Indian states (the geographical area) and the corresponding population
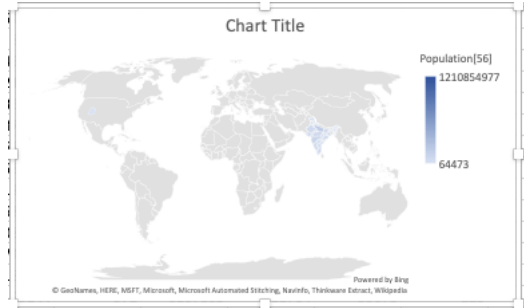2. Go to insert..
3. Select maps -> fill map

Figure 3: Source: My own

Attempt to make the map [2]

Nice, but not yet what we want.

We can see that Indian states are colored according to their population, but we might want to zoom in on India. This is not straightforward.

1. We double-click on the actual figure
2. To the right, a menu "Format Data Series" will appear
3. Go to the third tab, Series options
4. Select "Only regions with data"

Attempt to make the map [3]
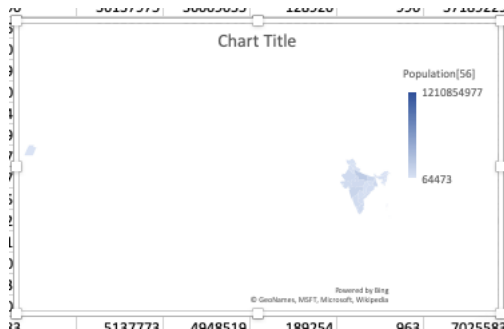
What we get is the following.

Figure 4: Source: My own

Still not exactly what we want. The problem appears to be the data!

Cleaning the Data

Excel is not recognizing the States data perfectly. Notice that some states have a suffix, UT behind them. Let's remove them quickly. Create two new columns between State and Population.

- ▶ In the first column:

        =FIND("(UT)";B2;1)

- ▶ In the second column:

        =IF(ISNUMBER(C2);LEFT(B2;C2-1);B2)

Cleaning the Data [2]

Of course, we could also remove them manually. It is useful to know how to do this with functions, because you might come across datasets where it takes a lot of time to do this manually.

- ▶ Now select the newly created column, and the column containing the population data.
- ▶ Again go to Insert -> Maps -> Fill Map
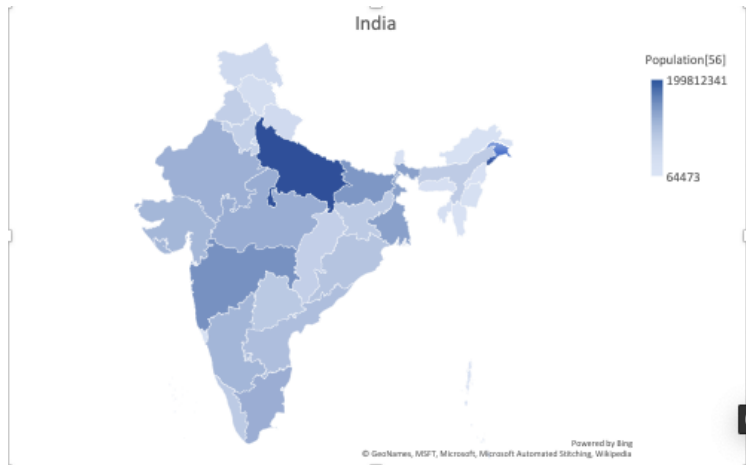- ▶ And you will (automatically) end up with..

This!

Figure 5: Source: My own

Copying tables

Broadly speaking, there are two (easy) ways to get appropriate data for your own chloropleths.

Oftentimes, you can just copy-and-paste tables from Wikipedia to Excel.
For example:

Take the table on this page and paste it to Excel

Make sure to paste and select 'Match destination formatting'

US Incarcerated Population

Then, you will end up with something like this (use find and replace to remove the comma's):
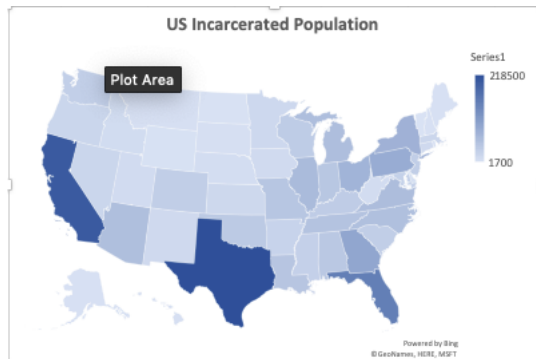

Figure 6: Source: My own

Use geographical datasets

Geographical datasets
The other way is to compile data yourselves from primary sources. In this case, make sure to be very careful when coding.

Checking
It pays to make up some random data and see if Excel 'understands' the coding of your geographical variables. Otherwise, you might want to reconsider.

The Match Function

- ▶ Today, we will have a look at the `MATCH` function in Excel.
- ▶ The `MATCH` function is not super useful in itself
- ▶ However, it is useful to use in conjunction with other functions
- ▶ We will learn how to use `MATCH` and `VLOOKUP` to efficiently merge datasets

Syntax

The `MATCH` function has several arguments:

```
=MATCH(lookup value, lookup table, match_type)
```

`match_type` will not be important to us. You should (probably) always set it to 0, or FALSE, meaning exact match.

Let us watch a short video explaining what the match function does in itself:
Here

Selecting variables

- In our case, match is most useful for looking up the position of certain variables in a different dataset
- We can find a variable with a certain name in a different sheet and its position in an array
- We can then use `VLOOKUP` to merge a dataset at once, by only having to drag one `VLOOKUP`-command throughout rows and cells

Example

A simple dataset

The best way to illustrate this is by means of an example with very simple data. We will now proceed to do so using this dataset: **Here**. You can download it and follow along step-by-step.

Relative and absolute arguments

You will also see that it's very important to make smart use of relative and absolute cell references. It probably takes some practice to fully get this.

Structure of the data

- ▶ The dataset contains data about gender equality in various countries in the world at various times.
- ▶ The first sheet contains three variables, the second sheet contains three variables
- ▶ Naturally, you want to merge them into one frame, but..
- ▶ You cannot just copy and paste. The items are not sorted in the same order, and not every observation has a 'match' in another sheet.

What do we do?

Vlookup and Match

We attempt to use `VLOOKUP` and `MATCH` to make **one** command, which we can drag to all relevant cells, and make Excel automatically find matches for all present observations in the first dataset.

Preparatory Work

Before we can do so, we have to `CONCATENATE` our id variables into one vector (a limitation of Excel). (Dutch: `TEKST.SAMENVOEGEN`)

Concatenating the observations

Figure 1: Source: My own

Assignment

Do the same thing in sheet 2

The real job

Step 1
First, copy the variable names of the variables you want to merge (i.e. Historical Gender Equality Index, Sex Ratio, Share of Women in Parliament to `H1:J1` in sheet 1.

Step 2
Now, in cell `H2`, insert: `=MATCH(H$1,sheet2.$D$1:$G$1, 0)`.
Mind the syntax (! instead of . in Excel and Google Sheets).
Drag H2 to I2 and J2. Why is the first argument of the match function relative/absolute, whereas the second argument, the array, is absolute?

The real job [2]

The trick consists of using MATCH *inside* the VLOOKUP function! This way, it'll automatically match the variable name in sheet 1 to a variable name in the selected array in the MATCH function. In cell H2:

```
=VLOOKUP(
        $D2,
        sheet2!sheet2!$D:$G,
        MATCH(
            H$1,
            sheet2!$D$1:$G$1,
            0),
        FALSE
        )
```

The real job [3]

Drag the function in H2 down, drag it to I2, and J2 and then down. We're done:

Figure 2: Source: My own

Exercise

1. **Verify that the observations have been matched correctly by comparing some random observations to the data in the second sheet. Use the filter function**
2. **This method also allows you to change the order of variables. Undo all merges, and try for yourself to change the order of the (to be merged) variables. The `MATCH` function will still look up the correct variables in the second sheet!**

Discussion

- ► We have now learned how to merge datasets in Excel
- ► Disadvantage: The procedure requires quite a lot of thinking, and careful placement of relative and absolute references
- ► BIG disadvantage: the amount of observations and the dataset you will end up with depends on the dataset you use as a reference!
- ► Other programming languages, such as R and Python, have way more hospitable ways of doing this
- ► There is even a separate programming language, SQL, designed to do these kinds of things such that the dataset you end up with will not depend on the reference dataset.

Assignment

Have a look again at the following dataset:
Link to Google Sheet

Copy this to your own spreadsheet program/Google sheets session

Transpose the dataset

If and else statements in Excel

If-else conditions

Suppose we would want to calculate the average population of those five cities. What if we try to take the sum, i.e.:

```
=SUM(I3:I7)/5
```

(or wherever we decide to place the data). That doesn't work, and the reason is that Excel does not recognize that those numbers (M and K) stand for 'million' and 'thousand' respectively.

If and else statements [2]

Function

Let us now try to write a program that converts numbers like these in 'readable' numbers. We will do this in several steps.

- ► First, we will see whether we can find an 'M' or a 'K'.
- ► Second, if we find one, we want to find which number stands before the M or K.
- ► If not, we may want to keep the number as it was..
- ► Third, we want to multiply the number with the appropriate coefficient to generate a readable number.

Organize your data..

Try to organize the data in the following way, and follow along:

population	city		find M	find K	which one should I take?	extract everything before M and K	final number
15 M	Moscow			4			
8.8 M	Mexico City						
6.6M	Madrid						
1M	Manahua						
260K	Monchengladbach						

Figure 1: Source: My own

If and else statements [3]

Step 1:

```
=FIND("M";A2) #or wherever your data is located
=FIND("M";A3)
etc.
```

..and similarly for `FIND("K")`.

Hint: you can 'slide down' formulas that you've entered once, so they extrapolate to other cells.

If and else statements [4]

Your table might look something like this afterwards:

population	city	find M	find K
15 M	Moscow	4	#VALUE!
8.8 M	Mexico City	5	#VALUE!
6.6M	Madrid	4	#VALUE!
1M	Manahua	2	#VALUE!
260K	Monchengladbac	#VALUE!	4

Figure 2: Source: My own

which indicates the places where the M's and K's are. We need these to find out which number stands before this, which we will do next.

If and else statements [5]

The next variable is a kind of merging variable, which looks for the place of an M in case there is an M, and looks for the place of K in case there is a K.

```
=IF(ISNUMBER(C2),
    C2,
    IF(ISNUMBER(D2),
        D2,
        "none")
    )
```

Here is the first `IF`-statement: we ask ourselves: if C2 is a number, then return C2. Otherwise, if D2 is a number, then return D2. Otherwise, return "none". (We will see why next)

If and else statements [6]

We will extract the number before "M" or "K" sign using the =LEFT function, which extracts the left part of a cell up to a certain number.

That certain number is the number we've just calculated (well, almost):

```
=IF(ISNUMBER(E2),
    LEFT(A2,E2-1),
    A2)
```

In this chunk, I am saying: if the previous column did NOT give back "none" (so if it is a number), then give me the left part of A2, up to the M or K. Otherwise, just give me back the original number.

If and else statements - Final

Finally, let us do the calculation.

```
=IF(ISNUMBER(C2),
    F2*10^6,
    IF(ISNUMBER(D2),
        F2*10^3,
        F2)
    )
```

We are saying: if you could find an M, take the extracted number, and multiply it by one million. If you could not, and if you could find a K, take the extracted number, and multiply it by one thousand! If you could find neither, then just take the number!

Comments

Conclusion
Now, we can calculate the mean population. We can also do something more: we can add new observations, and now, the data will be understandable if we use either 'K', 'M' or normal number notation!
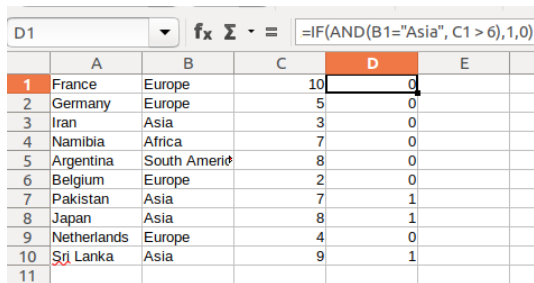
Watch out
Of course, all the auxiliary columns that we created can be removed. Be sure, however, to paste the output column as a value, not as a formula, because if the auxiliary columns disappear, the formula won't work anymore.

Multiple criteria..

More complicated functions

You can also think of situations where we have to deal with more criteria.
The `AND` and `OR` functions can help us, in combination with the `IF` function.
A small example:

Figure 3: Source: My own

Or and Not

- The `OR` and `NOT` functions work in the same way.
- OR means 1 of the criteria should be met, NOT means a criterion should not be met.
- You can use logic to come up with more complex criteria. What if you wanted at least 1 of several criteria not to be met?

Figure 4: Source: My own

Frequency tables

▶ As you have probably seen in the Excel tutorial, we can use Excel's pivot table function to generate several interesting aspects of the data, such as the frequency of a categorical variable.

▶ Afterwards, we also show how to quickly generate descriptive statistics (e.g. mean, standard deviation, etc.) from a given dataset.

Creating a frequency table

Creating frequency tables in excel is very easy. Let us take **this dataset** to demonstrate.

This dataset contains variables indicators about the development of human capital in various countries at various times.

Count observations

- In Excel, go to Insert > Pivot table
- On Google sheets, go to Data > Pivot table > create..
- Add country.name to rows, and (for example) Avg. years of education to values
- Excel: right click on the table > summarize by > count
- Sheets: in the menu on the right, click summarize by > count

Frequency table

This is what you get:

Figure 1: Source: My own

Other variables

- By using count, the frequency table automatically omits NA observations
- You can also add other variables to the table
- Excel: drag the variable to the 'value' menu
- Sheets: click on 'add', and select the variable you want to add

Descriptives

- It is also very easy to generate descriptive statistics in Excel.
- Let's use **this dataset** about public finances from the early 19th century onwards
- Suppose we are interested in the distribution of each of these variables, for example, the smallest and highest government bond yields, or the average debt-to-GDP ratio.

Descriptives [2]

1. Copy the dataset to Excel
2. Select everything.
3. Go to Data
4. Go to Data Analaysis > Descriptive Statistics
5. Select an output range
6. You can use a new sheet

Outcome

	Column 1	Column 2	Column 3	Column 4	Column 5
Mean	79.44273721	39.71556437	0.161534818	6.031391669	59.39890596
Standard Err	7.055362916	3.894485801	0.008011879	0.082628723	1.065297057
Mode	1	1	0	3.24	28.9
Median	8.664305	2.477087	0	4.95	45.96918
First Quartile	1.895748	0.97875125	0	3.79	26.366135
Third Quartile	19.58437	6.0722	0	7.1079165	74.433025
Variance	105081.6659	32017.57849	0.135505511	14.41286504	2395.684859
Standard Dev	324.1630237	178.9345648	0.368110731	3.796427932	48.94573382
Kurtosis	32.59832782	47.34943111	1.38940079	103.5500803	3.587370358
Skewness	5.670190134	6.577895133	1.840675241	6.108507912	1.79411797
Range	2875.355	1909.439	1	86.37258	277.6480121
Minimum	0	0	0	1.00325	0.002787855
Maximum	2875.355	1909.439	1	87.37583	277.6508
Sum	167703.6182	83839.55639	341	12732.26781	125391.0905
Count	2111	2111	2111	2111	2111

Figure 2: Source: My own

Finance

- It also pays to check out this function in combination with the filter.
- First filter the countries and years to your liking, and then executive descriptie statistics using the relevant subset of your data.
- You can think of criteria, and how to implement them, yourself (`IF`, `AND`, `OR`-kind of functions)

Questions

What is the minimum and maximum government bond yield in France between 1880 and 1960?
What is the minimum and maximum of Gold standard?
What is the standard deviation of the exchange rate to the US dollar of Argentina?