Statistics Exercise 1. Introduction to Excel

1 Introduction

MS Excel is widely used in the sciences to organize and analyze data. It includes functions for most basic data operations in a simple-to-overview interface and it is available in most professional organisations.

In this exercise we will look at

- filling cells and moving data,
- referencing cells,
- data organisation,
- the basic graphs (point, bar and line),
- applying functions.

2 Three crosses: Filling cells and moving data

The main parts of the interface is the *ribbon* at the top where various functionalities can be selected with the mouse, a *formula bar* (for a new workbook this will simply be an empty space right below the ribbon) where data or formulae can be written, and the spreadsheet area where data and output is placed.

1. Let's start by filling a column with some numbers. Write *ID* in the top left cell `A1`. Then fill the column with values from 1 to 10.

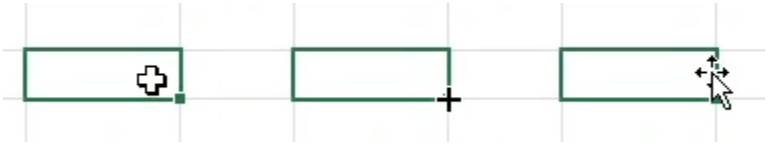


Figure 1: Three types of pointers.

If we want to mark some data (for example to copy it or to apply a function) we can click in one cell and drag to another. The marked area will have a green border. The operation of clicking and dragging has a few different uses in Excel. If the marker is a bold white, clicking and dragging will mark an area. If the marker is placed on the green border of a marked area the marker changes to a cross with arrows at the end; this marker is used to move cells. Finally, if the marker is placed on the bottom right corner of a marked area the marker changes to a thin black cross; this marker is used to auto-fill an area.

2. Write *ID2* in `B1`. Then write 1 in `B2` and 2 in `B3`. Mark the cells `B2:B3` using the bold white cross. Then place the pointer on the bottom right corner of the marked area, turning the pointer into the thin black cross. Click and drag down to `B11`. Alternatively, mark `B2:B3`, then double-click the green square at the bottom right of the marked area.

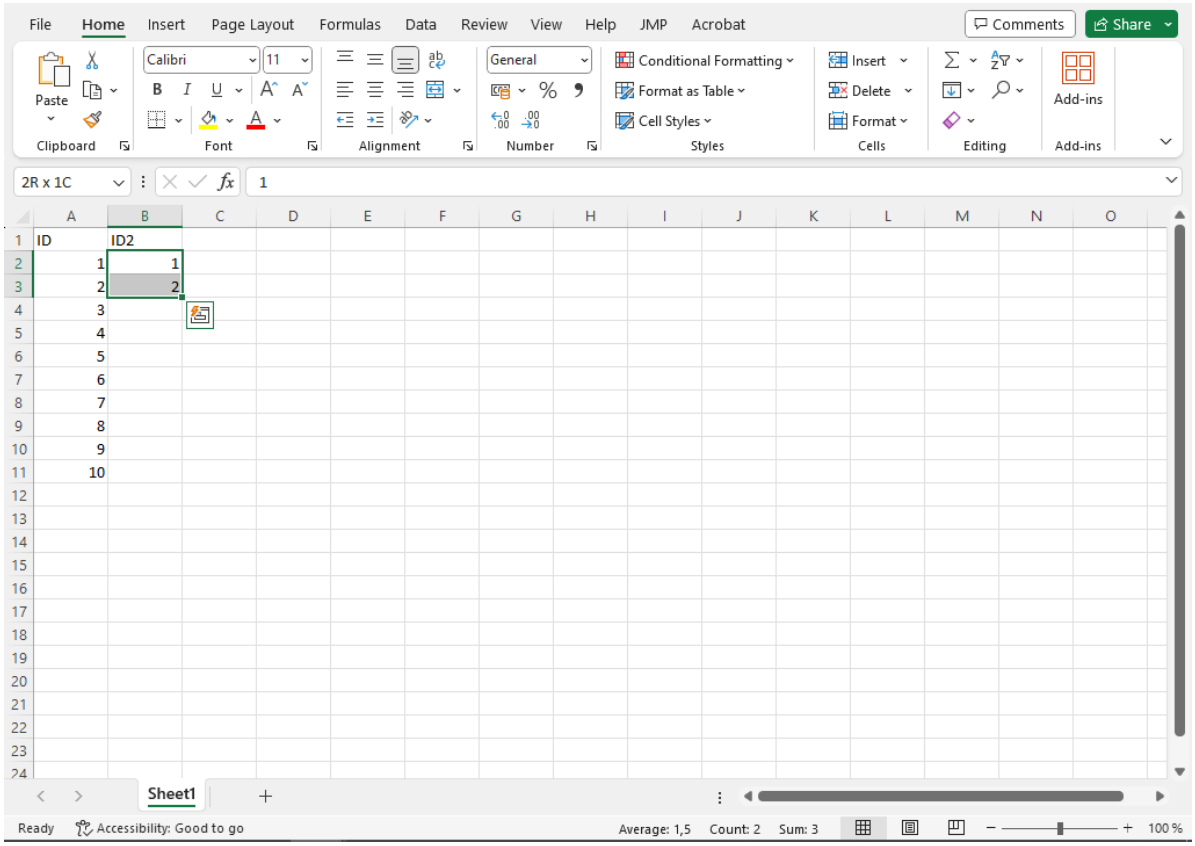


Figure 2: Auto-filling a column with inceasing numbers.

3 Mathematical operations and filling cells

One of the common data operations is to transform a column using some mathematical operation. Examples include changing the unit from hectograms to grams and calculating tree volume based on diameter and height. In Excel, operations like these can be done by doing the calculation in one cell and then filling that operation to all rows in the data.

3. We want to calculate a new column given by the ID value plus 10, divided by 2. Write a column name y in `C1`. Then go to `C2` and write `=(B2+10)/2`. Here, the equal sign is used to start a formula, and the calculation is given as the value in `B2` plus 10 and then divided by 2. The use of parenthesis follows standard mathematical rules so the operation within parenthesis is carried out first.
4. Next, we want that calculation for every row. This can be done by filling down. A few different ways are equivalent:

a. Click the calculated cell `C2` and double-click the green square at the bottom right of the marked cell.
b. Click the calculated cell `C2`, click the green square at the bottom right of the marked cell, and drag down to row 11.
c. Mark cells `C2` to `C11` and click *ctrl + D*.

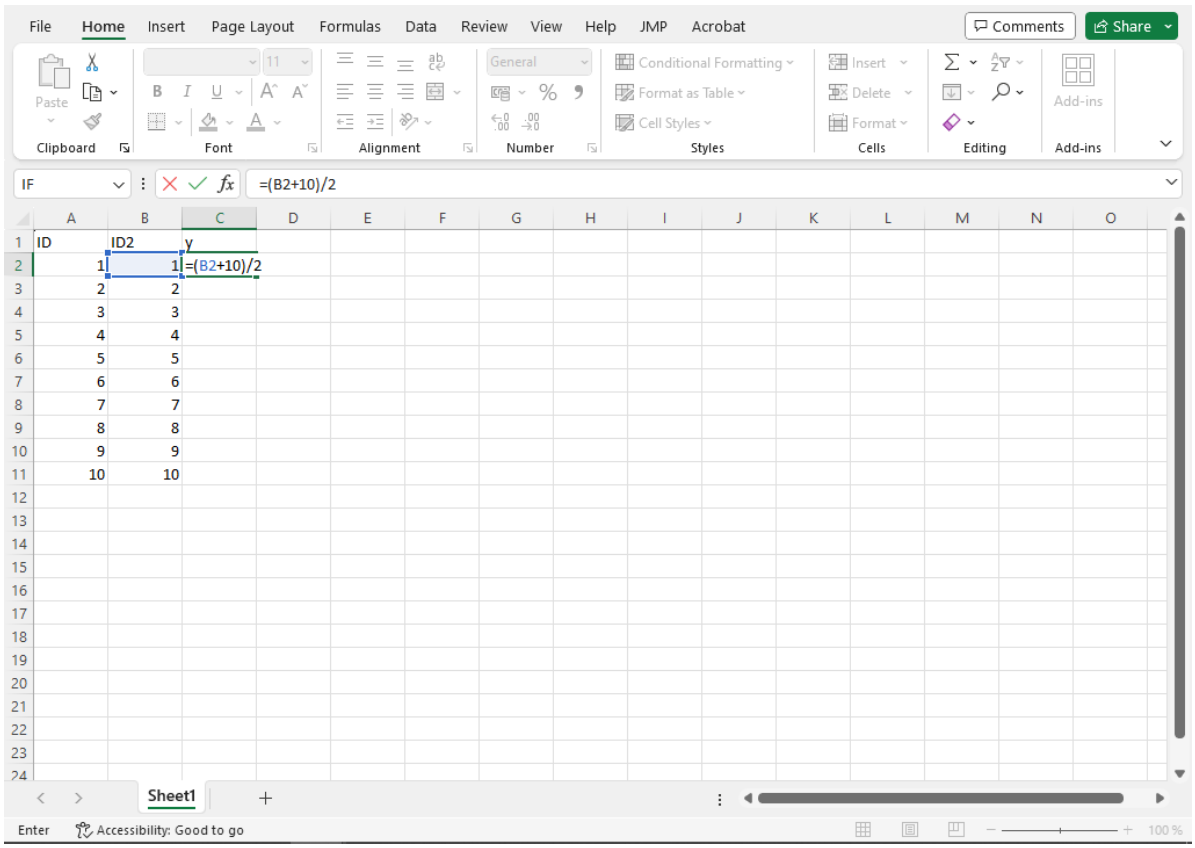


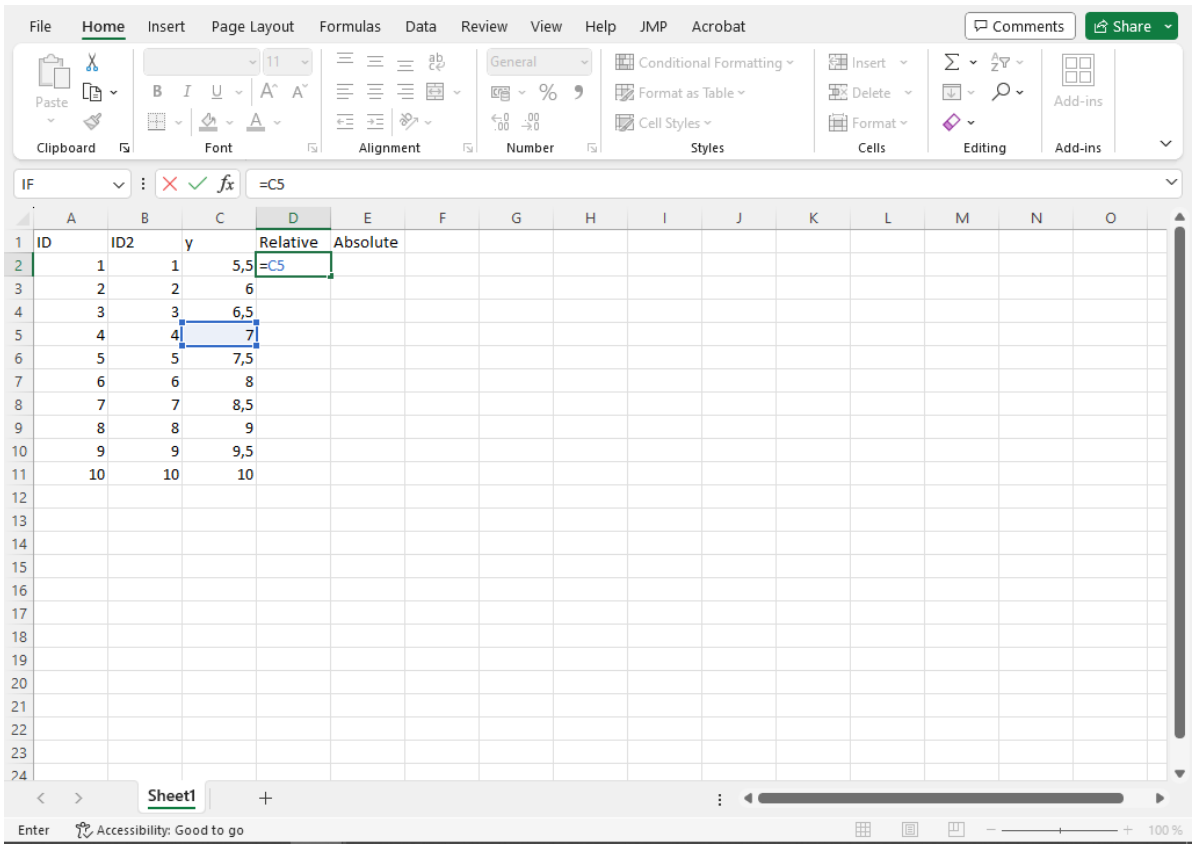
Figure 3: A mathematical formula.

4 Relative and absolute references

If we click any cell in the new column C we can see the formula in the formula bar. The calculation is different for each cell because each formula is connected to a different cell in the B column. When we filled the C column, the formula changed with each row, so that `C5` is connected to `B5`, `C6` to `B6` and so on. This is an example of a *relative* reference. The reference is relative to the current cell: a reference to a cell one step to the left will change when we fill in any direction.

If we don't want the references to change we can set an *absolute* reference. This is done with `$` such as `$C$4` to lock both column and row, alternatively `$C4` or `C$4` to only lock one of them. The F4 key is a shortcut for placing dollar signs.

5. Give the name *Relative* to column D. Write `=C5` in the cell `D2` and fill down.
6. Give the name *Absolute* to column E. Write `=$C$5` in the cell `E2` and fill down.



	A	B	C	D	E
1	ID	ID2	y	Relative	Absolute
2	1	1	5,5	=C5	
3	2	2	6		
4	3	3	6,5		
5	4	4	7		
6	5	5	7,5		
7	6	6	8		
8	7	7	8,5		
9	8	8	9		
10	9	9	9,5		
11	10	10	10		

Figure 4: A relative reference is just the cell name, for example `=C5`.

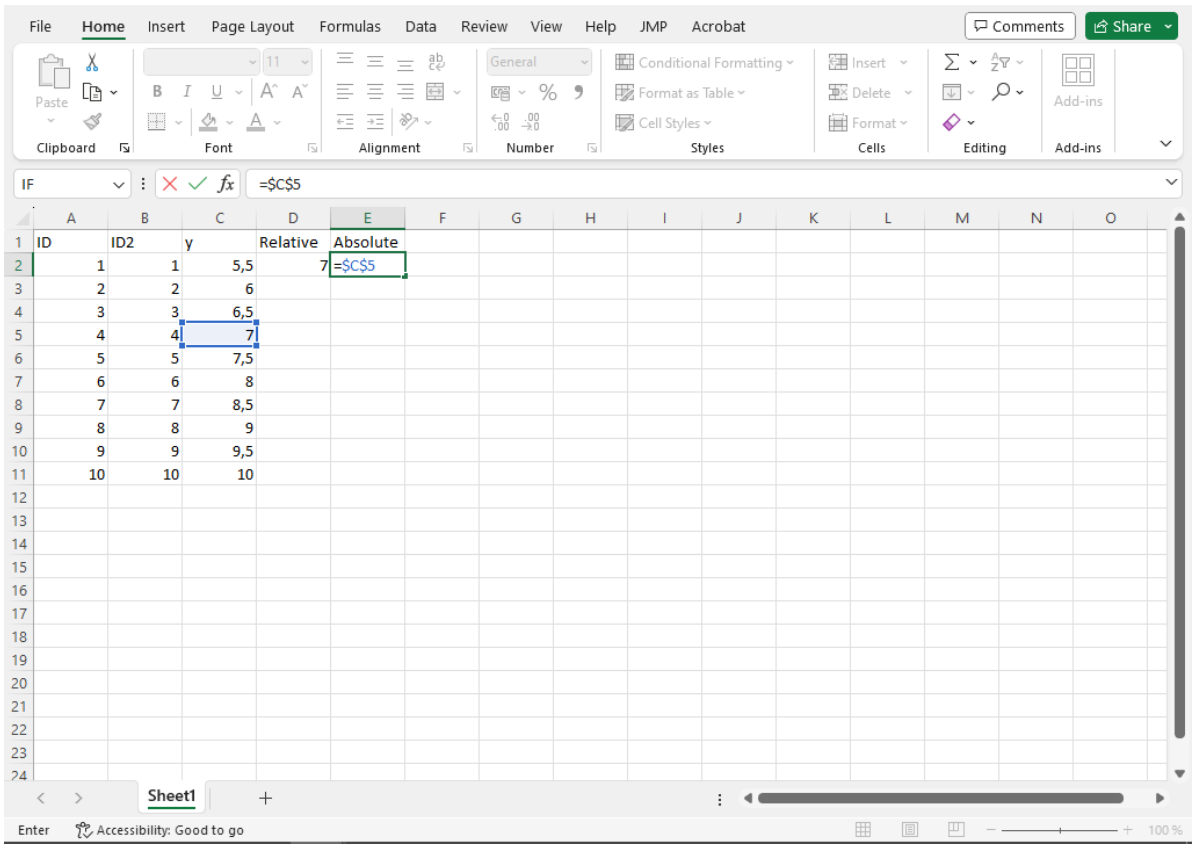


Figure 5: An absolute reference is set with a dollarsign, for example `=$C$5`.

Relative and absolute references are often mixed in a single formula. Say for example that we want to multiply some number with the values in *ID*, but we want to be flexible with what that number is. We can then use some cell to hold that number and multiply each row value with that cell.

7. Write the value 7 in an empty cell, say `I2`. Give the name *F* to column F. Write `=A2*$I$2` in `F2` and fill down. What happens when the value in `I2` is changed?

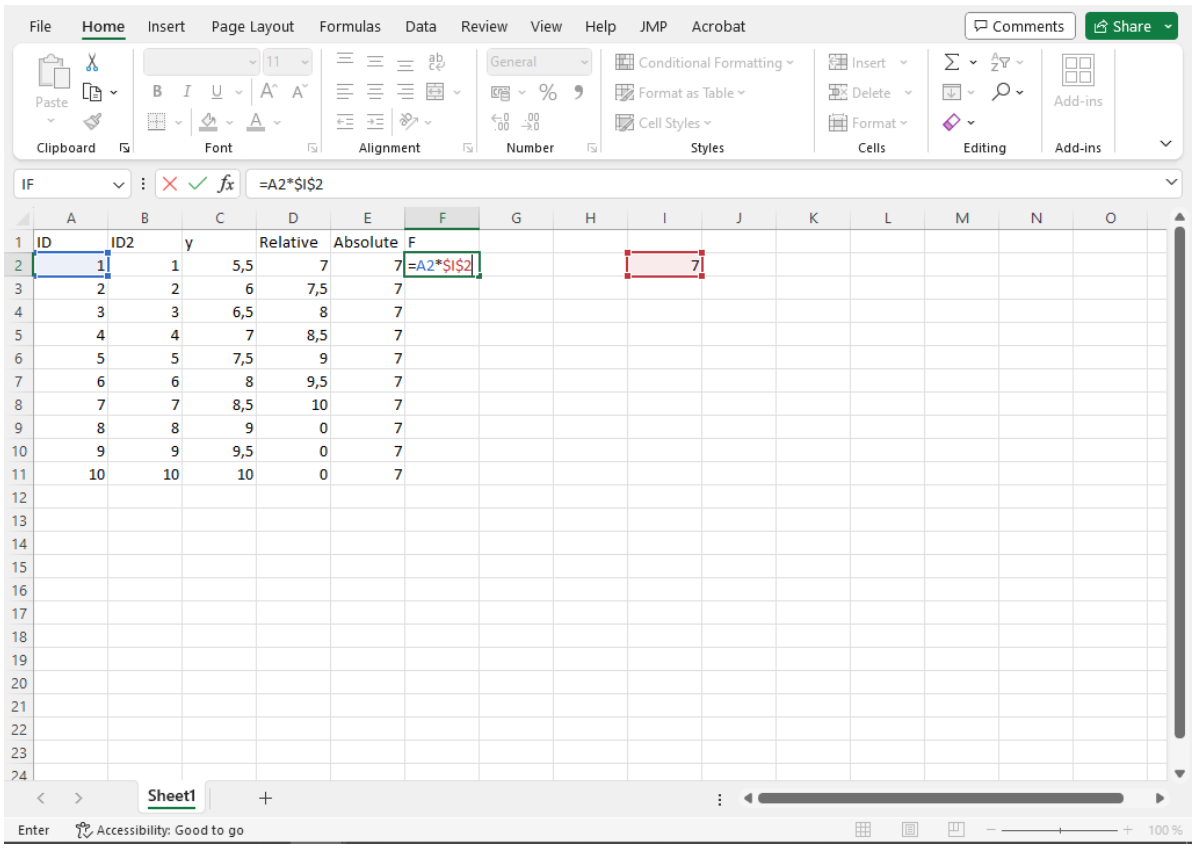


Figure 6: A mix of relative and absolute references can be useful to add interactivity in a document.

4.1 References when moving cells

As we saw when discussing the types of pointer a marked area can be used with the arrow cross by placing the cursor on the border of the marked area and clicking-and-dragging. Another alternative is to cut with *ctrl* + *X* and paste with *ctrl* + *V*. In modern Excel versions the references are fixed when moving a cell.

8. The cell `C2` references to `B2`. Move `C2` to some empty place and note the formula of the cell. Then move it back to `C2` again.
9. The cell `B2` is referenced by `C2`. Move `B2` to some empty place and note the formula in `C2`. Then move the cell back to `B2`.

5 Ready-made functions

A function is some operation that takes an input and produces an output. We saw an example of a function when calculating column *C*. Excel includes a large number of functions for different operations. You can see some of these by going to *Formula* in the ribbon and selecting a topic. Some key functions from a statistical point-of-view is to take a sum or a mean, possibly by some grouping.

10. The sum function is called `SUM` and takes a range of values as input, giving the sum of those values as output. In cell `I3`, calculate the sum of column `C`.
11. The mean function is called `AVERAGE` and takes a range of values as input, giving the mean of those values as output. In cell `I4`, calculate the mean of column `C`.
12. We almost always want to calculate sums or means for each of multiple groups. One way to do this is by `SUMIF` and `AVERAGEIF`. Start by giving column *G* a name *Grouping* and setting the first five values to *a* and the remaining to *b*. Mark an empty cell, say `I5` and write the formula `=SUMIF(G2:G11;"a";F2:F11)`. The `SUMIF` function takes three inputs: the grouping area, the condition to check, and the sum area. We can read this as *in the range G2 to G11, where the value is equal to* a*, sum the values in F2 to F11.*

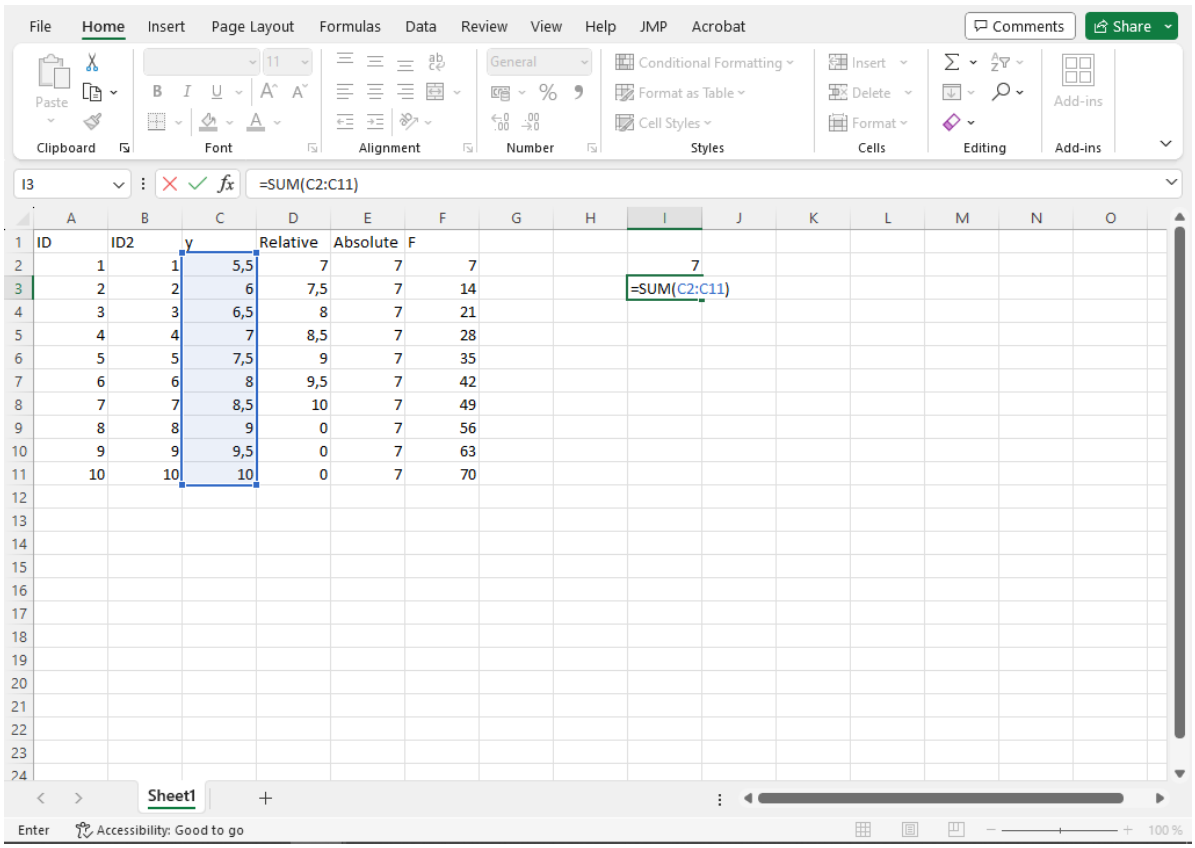


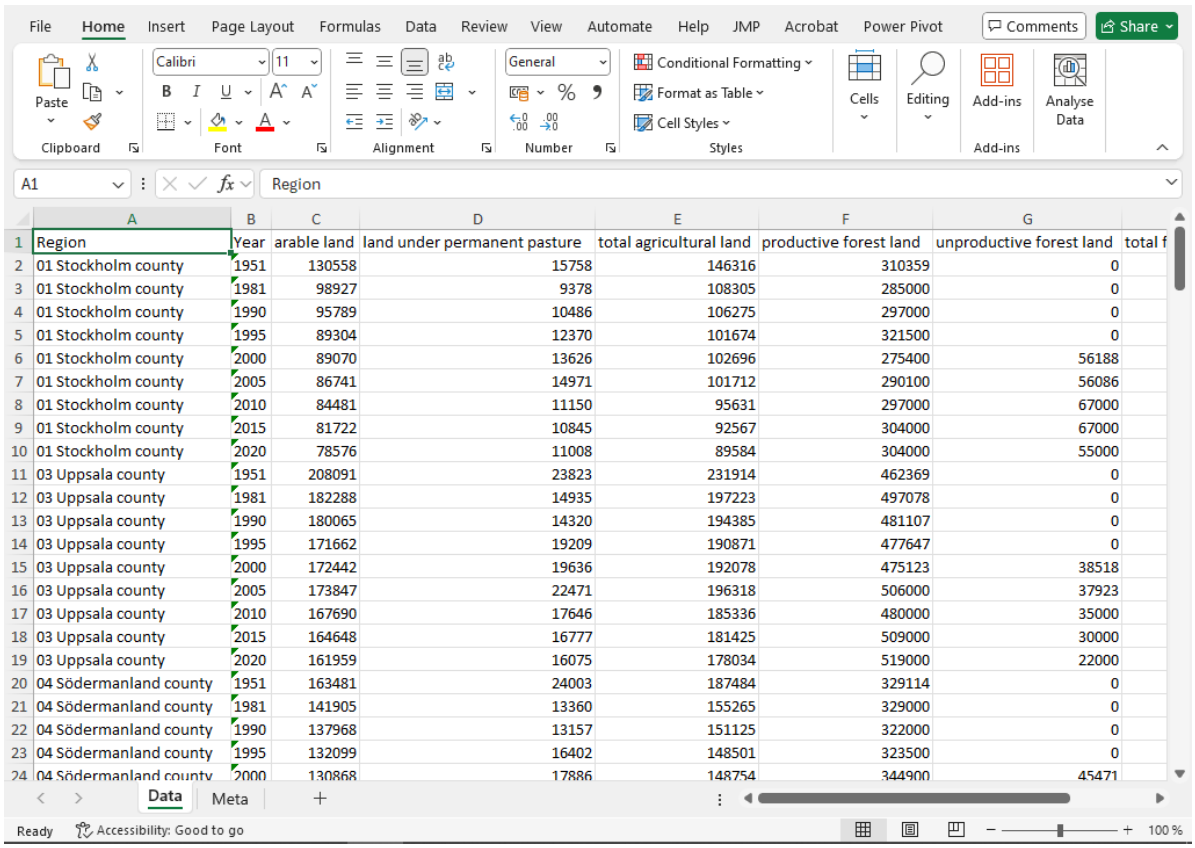
Figure 7: The sum function takes a range (multiple cells) as input and gives the sum of the valeus as output.

The `AVERAGEIF` function is one way to calculate mean values by different groups. We will later see another, more flexible, way using pivot tables.

6 A look at some actual data

Let us now turn to some real-world data to explore two common data operations: filtering (where specific observations are chosen from the full set) and sorting (where the data is ordered according to some criteria). The data covers land use in Swedish regions over time. It can be found on the course Canvas page in the file *data-exercise-1*. The source is *Statistics Sweden*.

13. Find the file on Canvas. Download it to a suitable folder. Open the file in Excel.



Region	Year	arable land	land under permanent pasture	total agricultural land	productive forest land	unproductive forest land	total f
01 Stockholm county	1951	130558	15758	146316	310359	0	
01 Stockholm county	1981	98927	9378	108305	285000	0	
01 Stockholm county	1990	95789	10486	106275	297000	0	
01 Stockholm county	1995	89304	12370	101674	321500	0	
01 Stockholm county	2000	89070	13626	102696	275400	56188	
01 Stockholm county	2005	86741	14971	101712	290100	56086	
01 Stockholm county	2010	84481	11150	95631	297000	67000	
01 Stockholm county	2015	81722	10845	92567	304000	67000	
01 Stockholm county	2020	78576	11008	89584	304000	55000	
03 Uppsala county	1951	208091	23823	231914	462369	0	
03 Uppsala county	1981	182288	14935	197223	497078	0	
03 Uppsala county	1990	180065	14320	194385	481107	0	
03 Uppsala county	1995	171662	19209	190871	477647	0	
03 Uppsala county	2000	172442	19636	192078	475123	38518	
03 Uppsala county	2005	173847	22471	196318	506000	37923	
03 Uppsala county	2010	167690	17646	185336	480000	35000	
03 Uppsala county	2015	164648	16777	181425	509000	30000	
03 Uppsala county	2020	161959	16075	178034	519000	22000	
04 Södermanland county	1951	163481	24003	187484	329114	0	
04 Södermanland county	1981	141905	13360	155265	329000	0	
04 Södermanland county	1990	137968	13157	151125	322000	0	
04 Södermanland county	1995	132099	16402	148501	323500	0	
04 Södermanland county	2000	130868	17886	148754	344900	45471	

Figure 8: Some real-world data. Each row is an observed unit and each column is a property of that unit.

6.1 Data structure

The land use data is a table where each row is given by a region (column *Region*) and a year (column *Year*). There are then numerical variables for land type. This type of data can be structured in a few different ways. The most basic structure would be to have columns for area, year and land-use with a single numerical variable. An example of this is given in the sheet *Long format*. Another format could be to let region and land-use define each row and have one column for each year.

When talking about different data structures we use the terms *longer* and *wider*. A longer dataset is one where there are more rows and a wider dataset is one where there are more columns. The operation of going from one structure to another is called *pivoting*.

7 Filter data

It is very common that we are interested in a subset of the rows in our dataset – we want to *filter* the data for certain rows. One simple way to filter in Excel is to add *filter buttons* to our dataset. This can be done by selecting the data, clicking the *Sort and Filter* button in the *Start* ribbon and selecting *Filter*. If done correctly small arrows are added to each column name. Clicking these arrows will give some options for filtering.

14. Use the filter for region and year to select the row for *12 Skåne county* in *1981*.
15. For numerical variables the filter menu has some specific options. Use these to filter rows since *1990*, i.e 1990 and all later observations.

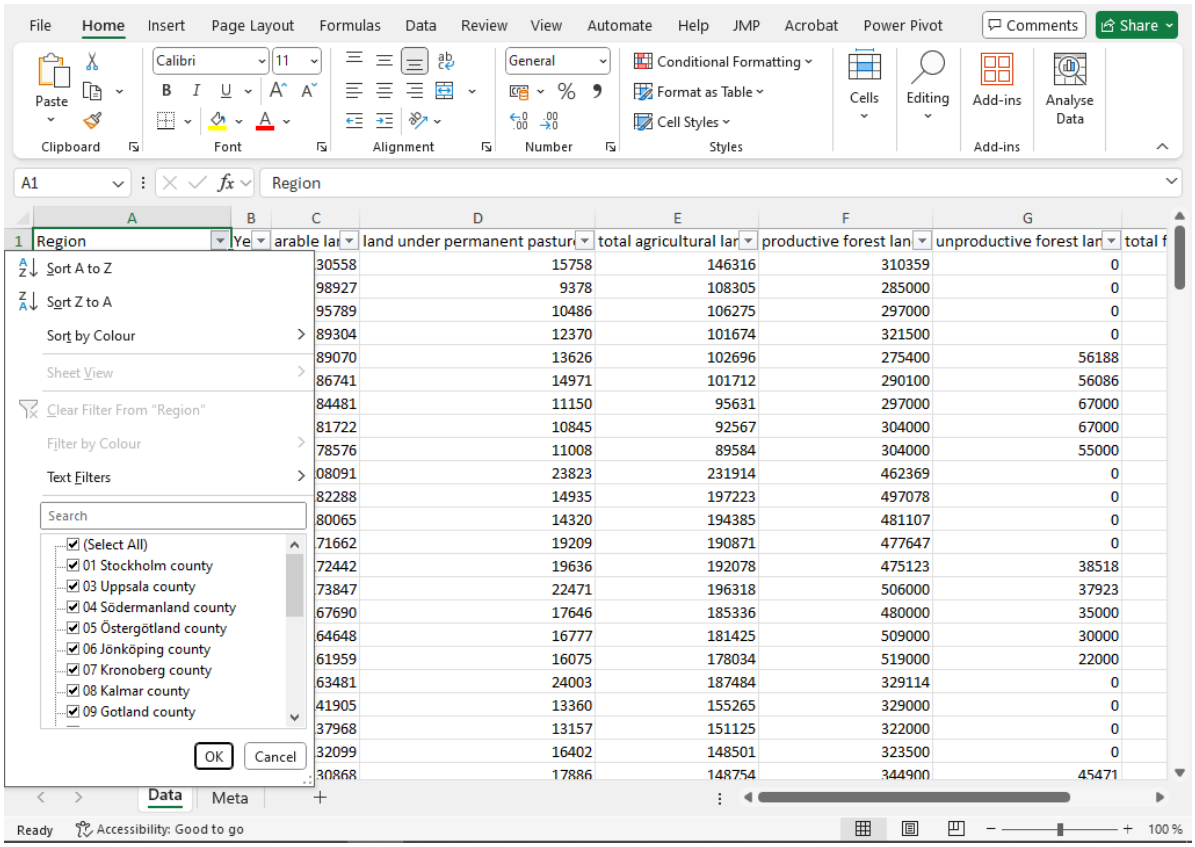


Figure 9: Setting a filter on the data using the filter button.

To remove a filter one can click the filter button and clear the current filter. This can also be done under *Sort and Filter* in the ribbon.

8 Sorting

The buttons at the top of each column can also be used for sorting the data according to some variable.

16. Remove any filter. Sort the data by *productive forest land* in decreasing order.

Since sorting changes the order of the data it is good to have a column of row IDs so that one can easily go back to the original order by sorting by ID. As a general rule the interpretation of a dataset should not rely on the order, that is it should be possible to randomly reorder the rows without losing any information.

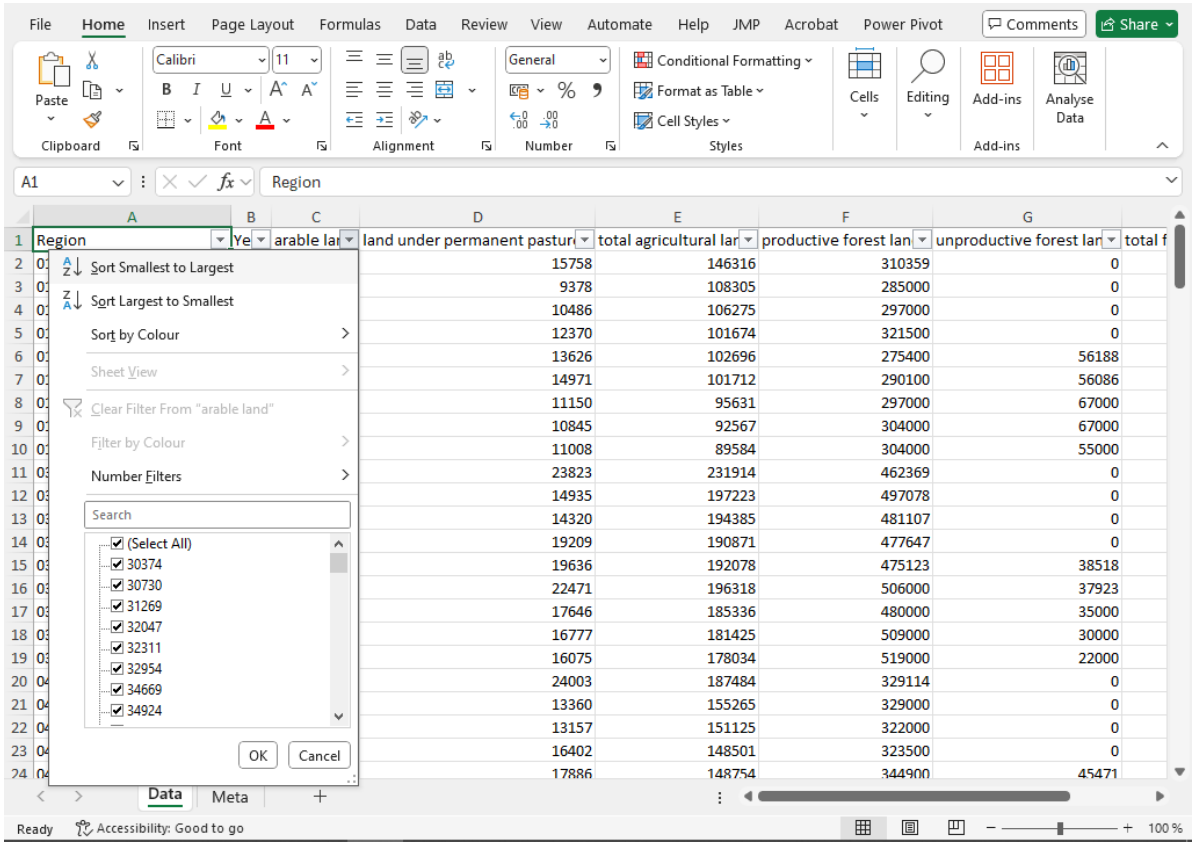


Figure 10: Sorting can be done using the button at the top of a column.

9 Graphs

Excel includes a large number of functions for graphs. The behaviour of Excels graph function depend on the structure of the data, and the amount of work needed to create an understandable,

neat-looking graph can vary quite a bit depending on data format and graph type. Nonetheless, Excel is a flexible tool, both in the types of graphs that are available and in the detailed appearance of the graph. Here we look at three fundamental graph types: a scatter plot, a bar chart, and a pie chart.

9.1 Scatter plot

A scatter plot illustrates two numeric variables, one on the x-axis and one on the y-axis. Each observation is a point in the graph located at the (x,y)-coordinates given by the two variables. Common variations are to color points according to a group variable or to connect the size of the point to a third variable (a bubble graph).

18. In the first sheet of the land use data, filter the data for the year 2020. Then mark the columns for total agricultural land and total forest land. This can be done by clicking and dragging while holding the *Ctrl* key. Then go to the *Insert* ribbon and find the mini-button for the scatter plot (bottom middle of the graph icons). Click it to see some suggested graph. Select the basic one, with only points.
19. The default graph is missing axis labels. These can be added in a few different ways, for example by clicking in a white space of the graph and adding a graph element with the plus buttom which should appear to the top-right of the graph. Use this to add labels to the graph.
20. The appearance of a graph can be changed to fairly great detail. Try double-clicking a point to reach an editor. The menu at the top of the editor window (the text typically start with *Alternatives for*) can be used to select a specific part of the graph to change. Try to find a way to change the color of the points.

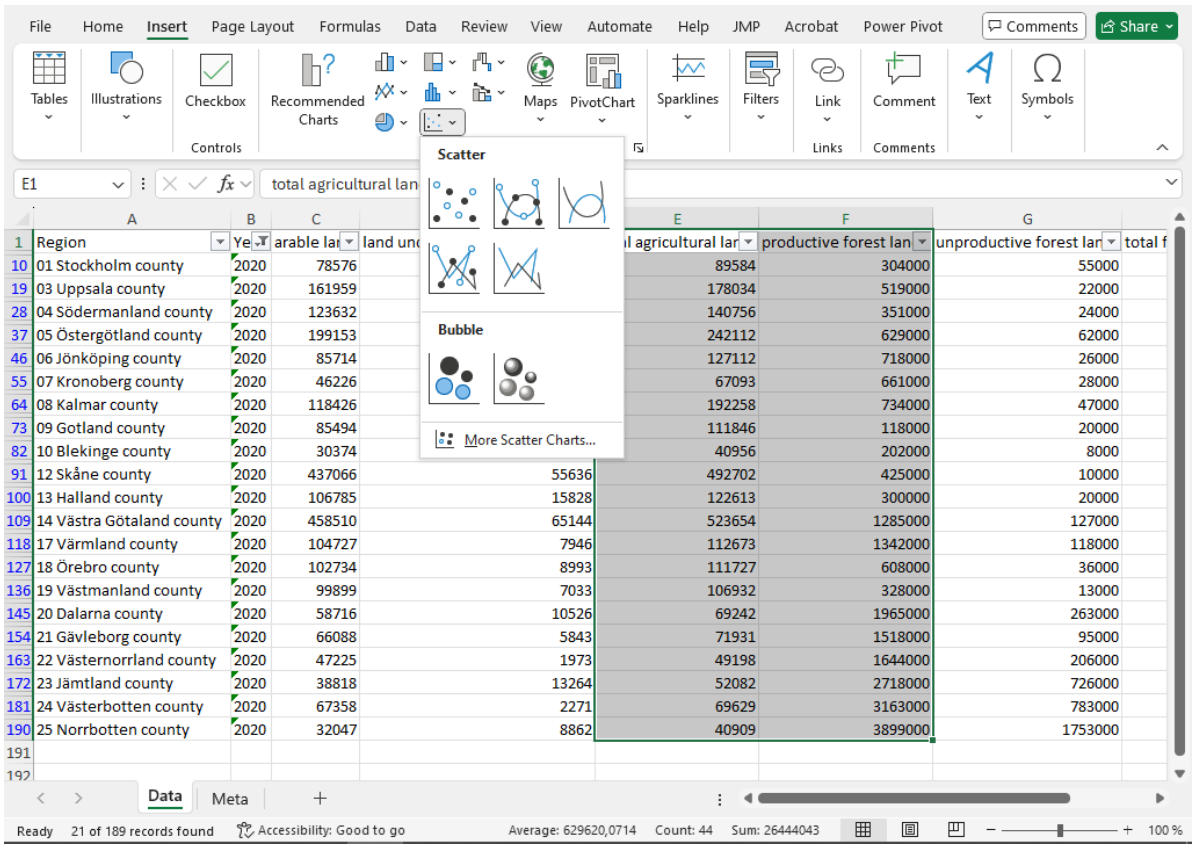
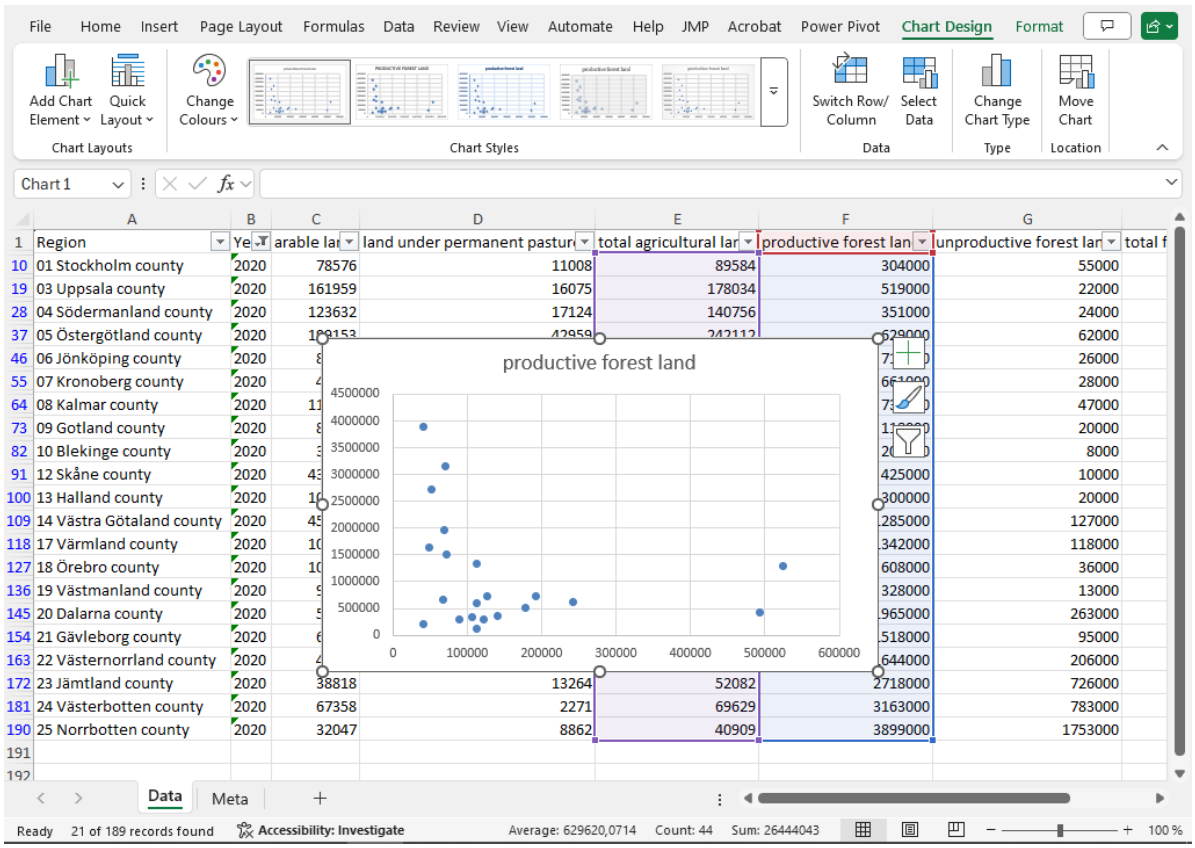


Figure 11: A scatter plot can be produced by marking the relevant data and selecting a scatter plot in the Insert ribbon.



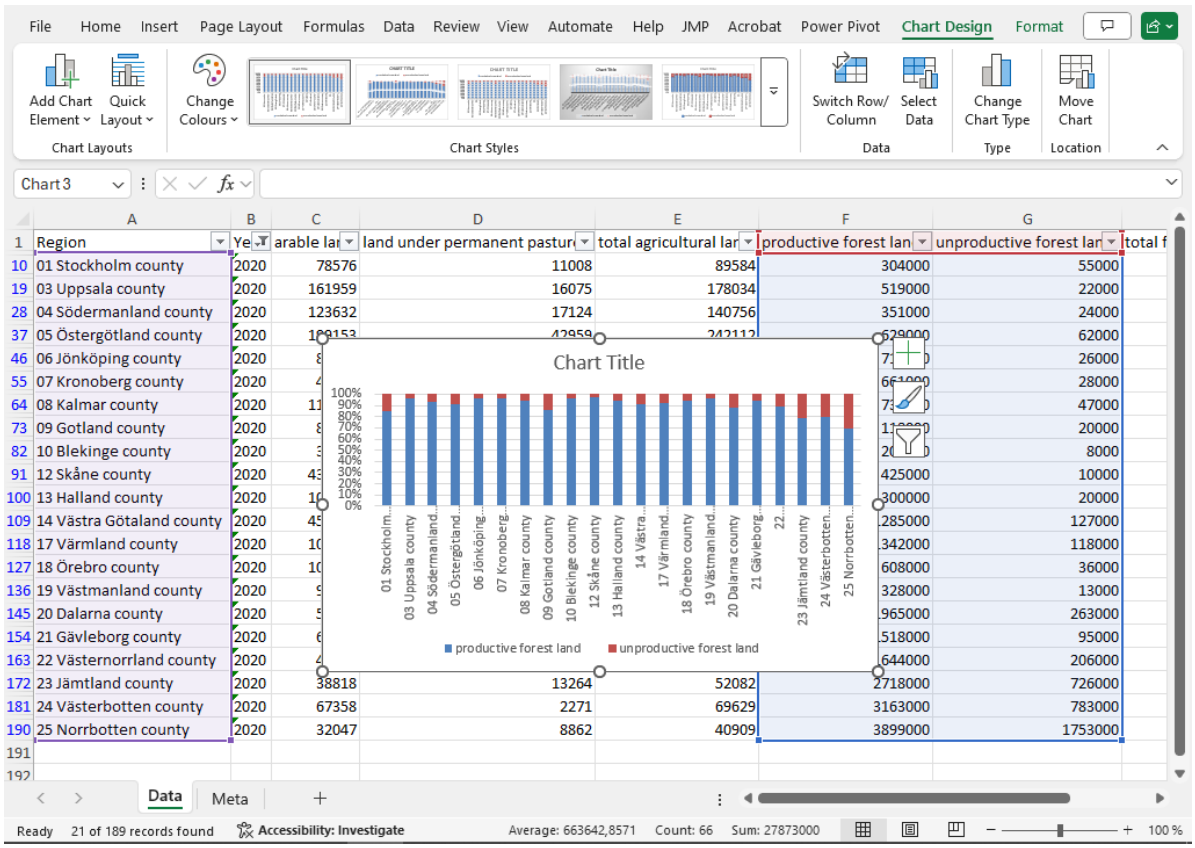
	Region	Ye	arable lar	land under permanent pastur	total agricultural lar	productive forest lan	unproductive forest lar	total f
10	01 Stockholm county	2020	78576	11008	89584	304000	55000	
19	03 Uppsala county	2020	161959	16075	178034	519000	22000	
28	04 Södermanland county	2020	123632	17124	140756	351000	24000	
37	05 Östergötland county	2020					62000	
46	06 Jönköping county	2020					26000	
55	07 Kronoberg county	2020					28000	
64	08 Kalmar county	2020					47000	
73	09 Gotland county	2020					20000	
82	10 Blekinge county	2020					8000	
91	12 Skåne county	2020				425000	10000	
100	13 Halland county	2020				300000	20000	
109	14 Västra Götaland county	2020				285000	127000	
118	17 Värmland county	2020				342000	118000	
127	18 Örebro county	2020				608000	36000	
136	19 Västmanland county	2020				328000	13000	
145	20 Dalarna county	2020				965000	263000	
154	21 Gävleborg county	2020				518000	95000	
163	22 Västernorrland county	2020				644000	206000	
172	23 Jämtland county	2020	38818	13264	52082	2718000	726000	
181	24 Västerbotten county	2020	67358	2271	69629	3163000	783000	
190	25 Norrbotten county	2020	32047	8862	40909	3899000	1753000	
191								
192								

Figure 12: The produced graph can be alterd in the formatting window to the right.

9.2 Bar chart

A bar chart illustrates a numerical value as the height of a bar. Commonly multiple bars are shown in one graph, with each bar connected to a group in a categorical variable. The bar can be divided into different sections by a filled-in color.

21. We want to create a bar chart which for a given year shows each land use for each region. Filter for a single year, say 2020. Then mark the columns for region and the numerical columns *arable land*, *unproductive forest land*, and *productive forest land*. This can be done by clicking in the column header (where there is a letter indicating the column) and then clicking the other three columns while holding the *ctrl* key (on a windows system). Then we go to the *Insert* ribbon and select the bar chart icon among the different graph options. There are a few different options depending on whether one want grouped bars (the sections standing next to each other) or stacked bars (the sections placed on top of each other). Pick whichever you prefer.



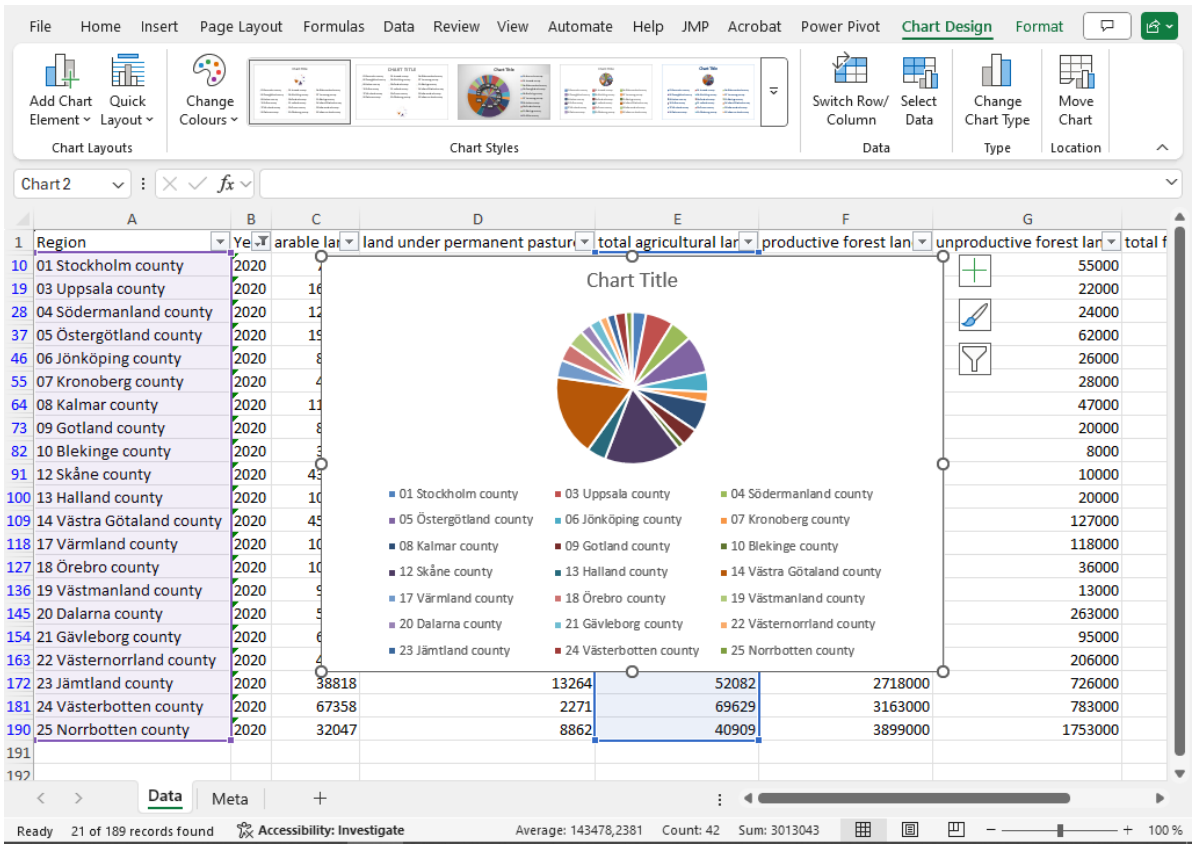
	Region	Year	arable land	land under permanent pasture	total agricultural land	productive forest land	unproductive forest land	total f
10	01 Stockholm county	2020	78576	11008	89584	304000	55000	
19	03 Uppsala county	2020	161959	16075	178034	519000	22000	
28	04 Södermanland county	2020	123632	17124	140756	351000	24000	
37	05 Östergötland county	2020					62000	
46	06 Jönköping county	2020					26000	
55	07 Kronoberg county	2020					28000	
64	08 Kalmar county	2020					47000	
73	09 Gotland county	2020					20000	
82	10 Blekinge county	2020					8000	
91	12 Skåne county	2020				425000	10000	
100	13 Halland county	2020				300000	20000	
109	14 Västra Götaland county	2020				285000	127000	
118	17 Värmland county	2020				342000	118000	
127	18 Örebro county	2020				608000	36000	
136	19 Västmanland county	2020				328000	13000	
145	20 Dalarna county	2020				965000	263000	
154	21 Gävleborg county	2020				518000	95000	
163	22 Västernorrland county	2020				644000	206000	
172	23 Jämtland county	2020	38818	13264	52082	2718000	726000	
181	24 Västerbotten county	2020	67358	2271	69629	3163000	783000	
190	25 Norrbotten county	2020	32047	8862	40909	3899000	1753000	

Figure 13: An example of the bar chart.

9.3 Pie chart

A pie chart illustrates the parts of a total as sectors of a circle (the *pie*). It shows the relative sizes of different groups. The pie chart has met some criticism for being difficult to read and it is generally recommended to not use a pie chart when the number of groups is large or when the differences between groups is small. In those cases, a bar chart with grouped bars may be clearer.

22. A pie chart can be made with just a single column of numerical data, but typically we also want a categorical variable to create labels for each sector. In the first sheet of the excel file, filter for a particular year, then mark the column with region and the numerical values for total forest land. Then select the pie chart icon in the *Insert* ribbon. Pick the top-left standard pie chart. The resulting pie chart should have a legend showing which sector belongs to what region.



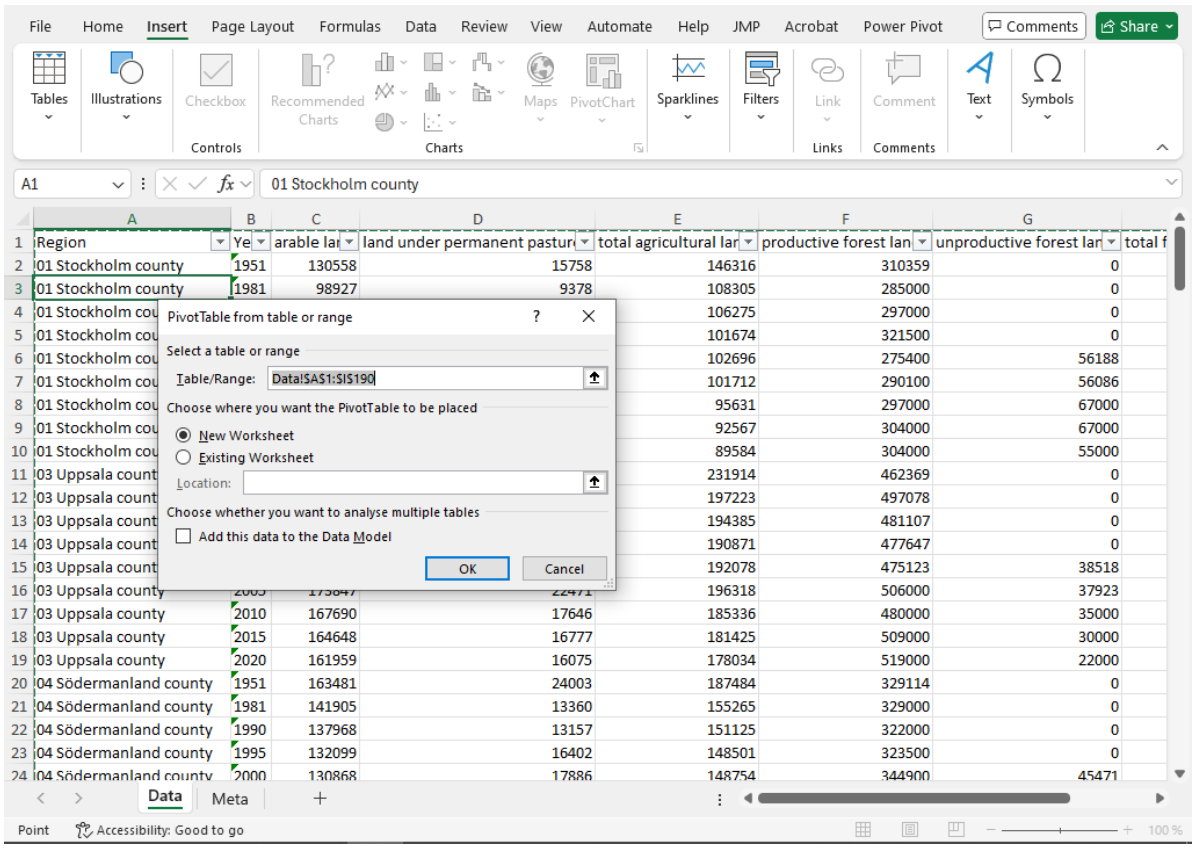
	Region	Ye	arable la	land under permanent pastur	total agricultural lar	productive forest lan	unproductive forest lar	total f
10	01 Stockholm county	2020					55000	
19	03 Uppsala county	2020	16				22000	
28	04 Södermanland county	2020	12				24000	
37	05 Östergötland county	2020	19				62000	
46	06 Jönköping county	2020	8				26000	
55	07 Kronoberg county	2020	4				28000	
64	08 Kalmar county	2020	11				47000	
73	09 Gotland county	2020	8				20000	
82	10 Blekinge county	2020	3				8000	
91	12 Skåne county	2020	43				10000	
100	13 Halland county	2020	10				20000	
109	14 Västra Götaland county	2020	45				127000	
118	17 Värmland county	2020	10				118000	
127	18 Örebro county	2020	10				36000	
136	19 Västmanland county	2020	9				13000	
145	20 Dalarna county	2020	5				263000	
154	21 Gävleborg county	2020	6				95000	
163	22 Västernorrland county	2020	4				206000	
172	23 Jämtland county	2020	38818	13264	52082	2718000	726000	
181	24 Västerbotten county	2020	67358	2271	69629	3163000	783000	
190	25 Norrbotten county	2020	32047	8862	40909	3899000	1753000	
191								
192								

Figure 14: An example of a pie chart.

10 Pivot tables

We earlier saw how specific functions like `SUM` and `AVERAGE` can be used to summarise an entire group of values into a single number (to *aggregate*). Excel has a built in tool, pivot tables, which makes this operation a bit simpler. The steps to constructing a pivot table is to start with some data, create an empty pivot table, for example using the *Insert* ribbon, and then connecting the rows and columns of the pivot table to the columns of the original data.

23. Go to the first sheet in the land use data. Click the *Insert* ribbon and the PivotTable button. This should open a dialog window showing the range of the data (it should automatically identify the full data as the ingoing data) and an option of where to place the pivot table. Select *New Worksheet* and click *Ok*.



	A	B	C	D	E	F	G
1	Region	Ye	arable lar	land under permanent pastur	total agricultural lar	productive forest lan	unproductive forest lar
2	01 Stockholm county	1951	130558	15758	146316	310359	0
3	01 Stockholm county	1981	98927	9378	108305	285000	0
4	01 Stockholm cou				106275	297000	0
5	01 Stockholm cou				101674	321500	0
6	01 Stockholm cou				102696	275400	56188
7	01 Stockholm cou				101712	290100	56086
8	01 Stockholm cou				95631	297000	67000
9	01 Stockholm cou				92567	304000	67000
10	01 Stockholm cou				89584	304000	55000
11	03 Uppsala count				231914	462369	0
12	03 Uppsala count				197223	497078	0
13	03 Uppsala count				194385	481107	0
14	03 Uppsala count				190871	477647	0
15	03 Uppsala count				192078	475123	38518
16	03 Uppsala county				196318	506000	37923
17	03 Uppsala county	2010	167690	17646	185336	480000	35000
18	03 Uppsala county	2015	164648	16777	181425	509000	30000
19	03 Uppsala county	2020	161959	16075	178034	519000	22000
20	04 Södermanland county	1951	163481	24003	187484	329114	0
21	04 Södermanland county	1981	141905	13360	155265	329000	0
22	04 Södermanland county	1990	137968	13157	151125	322000	0
23	04 Södermanland county	1995	132099	16402	148501	323500	0
24	04 Södermanland county	2000	130868	17886	148754	344900	45471

Figure 15: Creating a pivot table.

Done correctly, we should now have an empty pivot table in a new sheet. To the right of the window we can see a list of the variables of the original data and four fields: *Filters*, *Columns*, *Rows* and *Values*. We fill a pivot table by dragging variables from the list into one of the four fields.

24. Drag the variable *Year* to the field *Rows* and the variable *total arable land* to the field *Values*. Interpret the resulting pivot table.

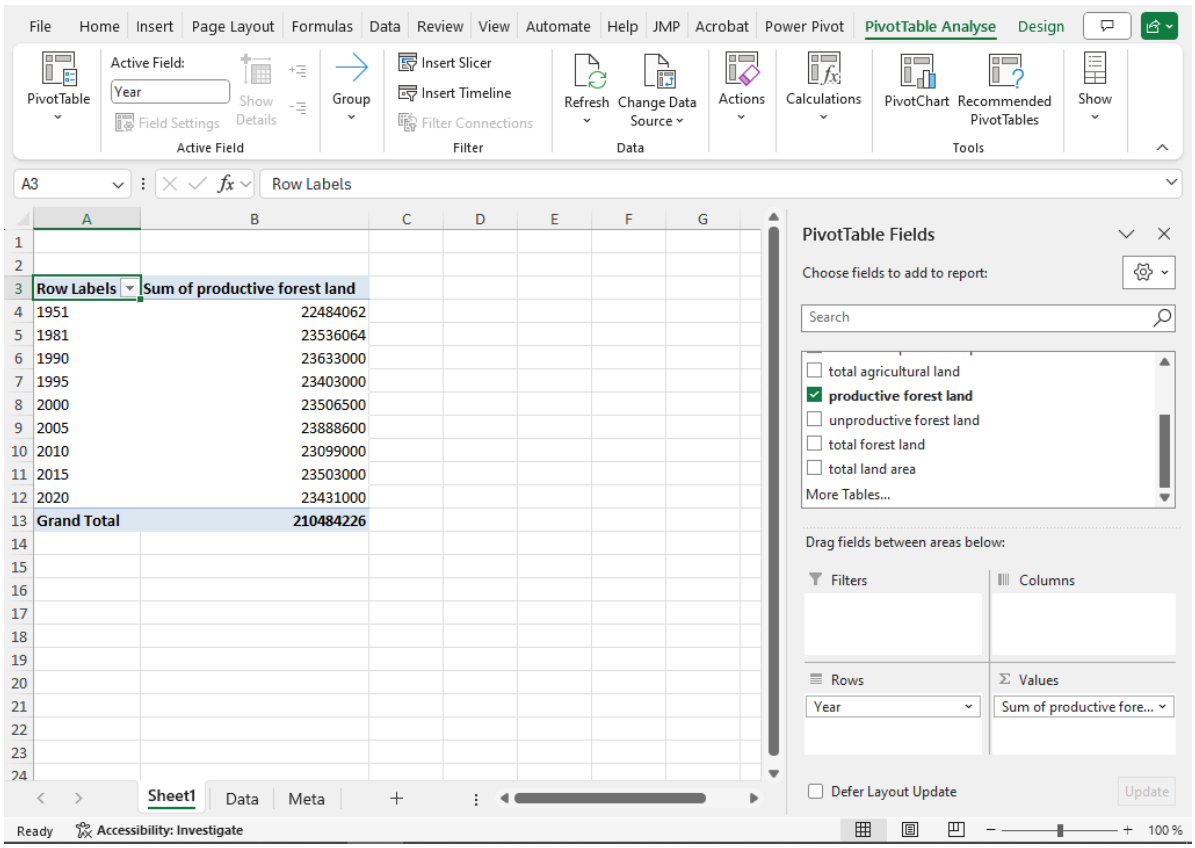


Figure 16: A pivot table with rows and values.

25. The default aggregation function is the sum. To change this we can go to the *Values* field in the window to the right and click the arrow button new to the selected variable. Under *Value Field Settings* there is a number of possible aggregation functions, including the mean and the standard deviation (*StdDev*). Change it to a mean value for now.
26. Finally a quick look at the other two field of the table. Drag the variable *Region* to the field *Row* and observe the outcome. Then drag the variable *Region* to *Filters* and observe the outcome.

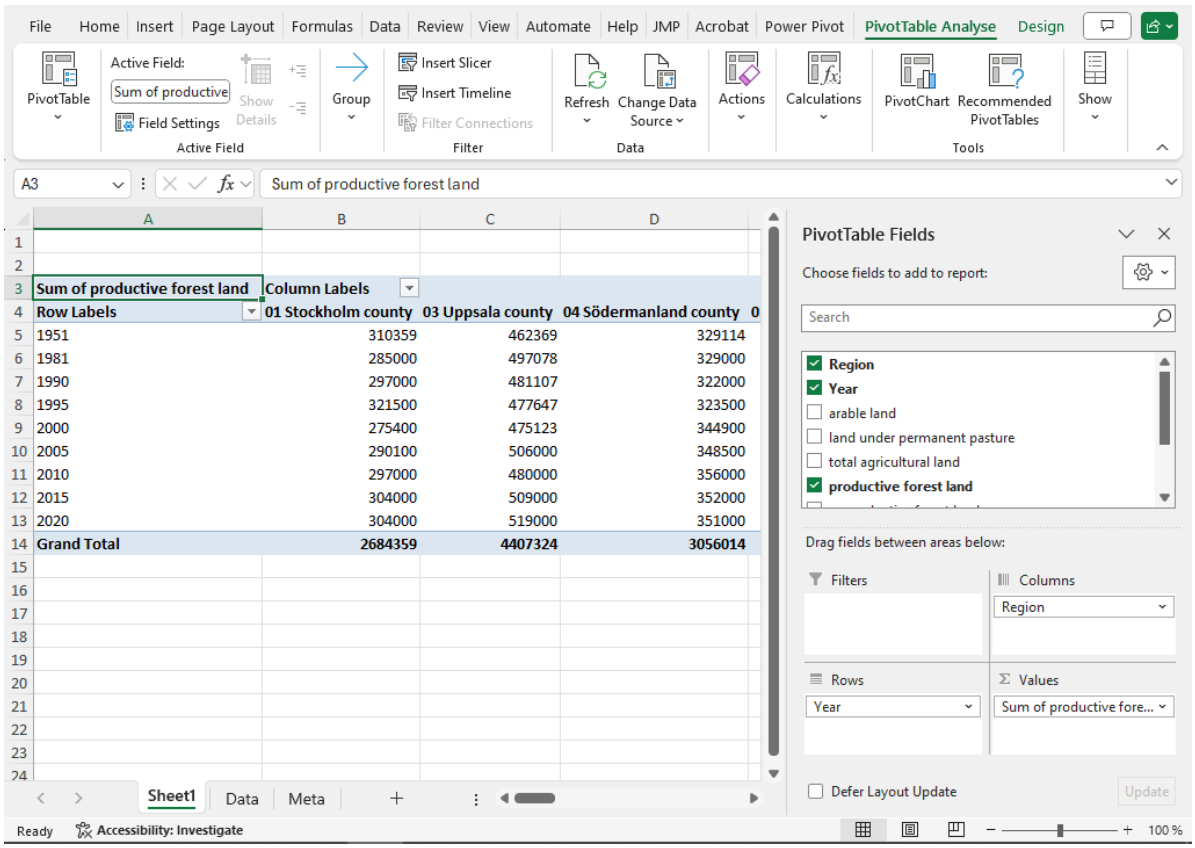


Figure 17: A pivot table with columns, rows and values.

11 Upcoming material

In the next computer exercise we will look at

- some of the standard scientific graphs (boxplots, barcharts with errorbars, and the histogram),
- statistical tests to compare groups,
- correlation and regression to examine the relation between two numeric variables.