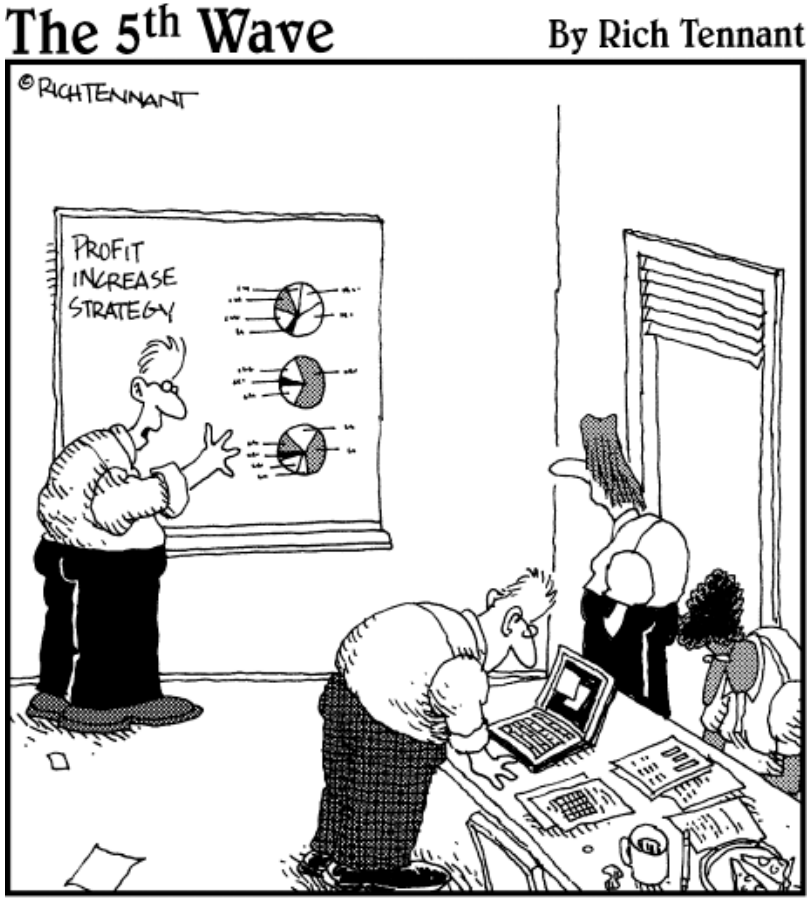
Data and Statistics in Excel

PerceptiveAnalytics

"After analyzing all your data, I think we can safely say that none of it is useful."

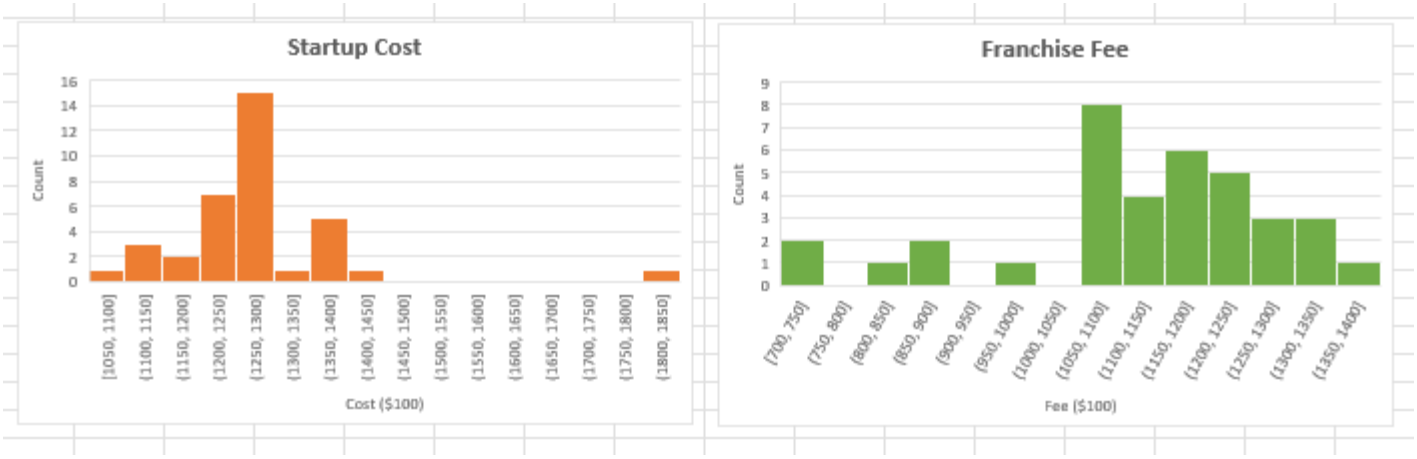


A simple way to access data (that we don't create ourselves) is to open an Excel file (.xlsx).

Download **4 – pizza dataset.xlsx** from D2L. Here, X represents the annual franchise fee in 100 dollar units and Y represents the startup cost.

Task:

- Change the column labels to be more descriptive, then create a **histogram** of the startup cost column of data. Label the axes and adjust the bin size to "50."
- Then, repeat for the "Franchise Fee" column of data.

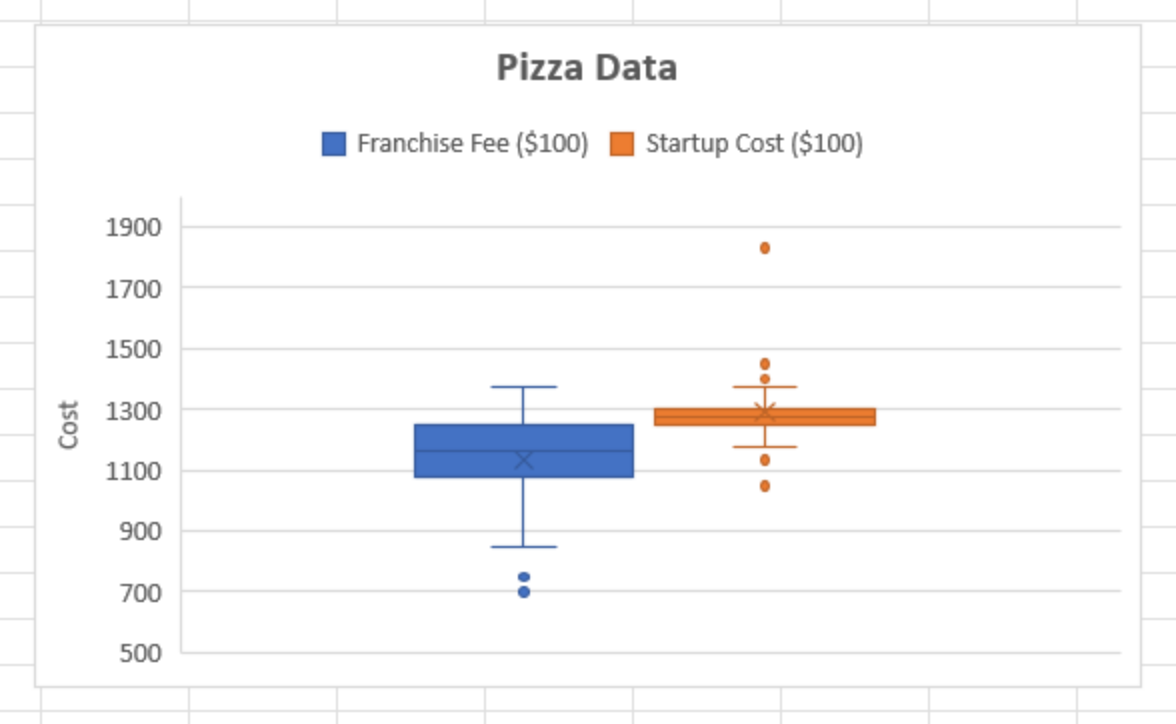


What information does a histogram show?

A boxplot is another useful way to visualize a distribution of data.

Task:

- Highlight all of your data and insert a box plot. Adjust the axis bound and format until you see something similar to the image below:



A boxplot is another useful way to visualize a distribution of data.

Features:
The data is split in to four **quartiles**:

- The line in the middle is the **median** (the middle value in a sorted list)
- The colored box extends to the **first quartile** (middle of the bottom half) and the **third quartile** (middle of the top)
- The thin lines ("whiskers") extend to the **local minimum** and **local maximum**
- Dots are outliers (more than 1.5x the inter-quartile range, i.e. the height of the shaded box)

Finally, the "X" marks the **mean** (average).

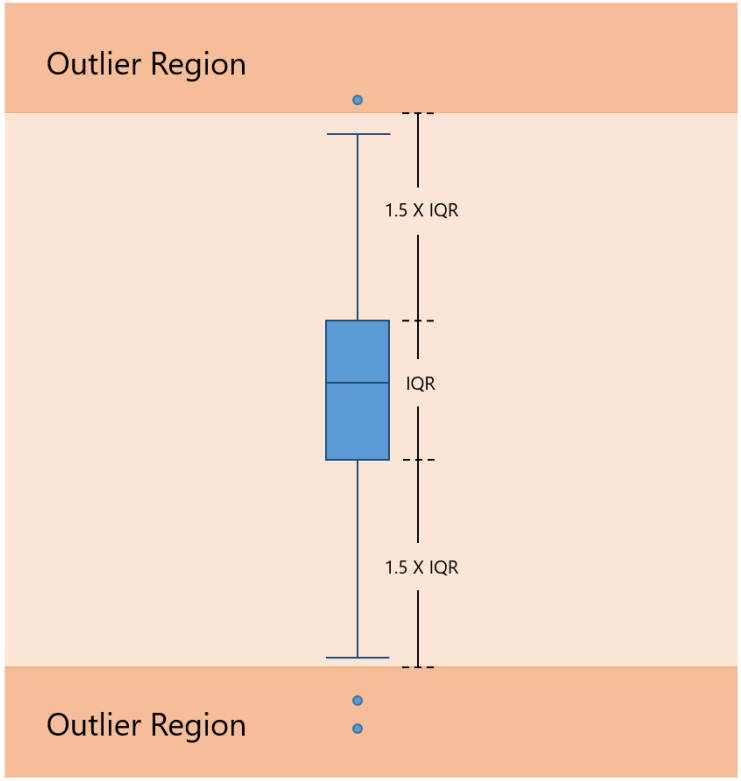
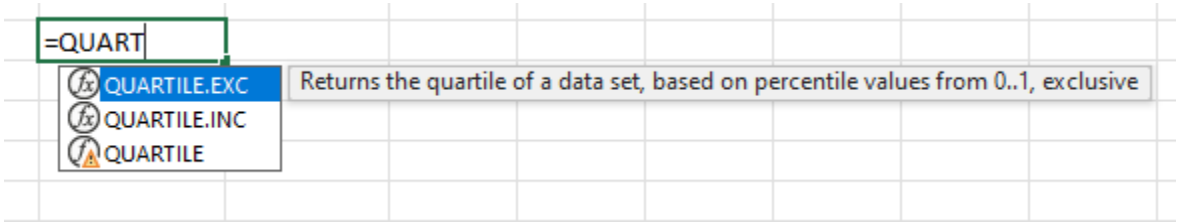


Figure from Microsoft

We can also calculate the numerical values directly with a formula.

Near your box and whisker plot, choose a cell and begin typing **=QUARTILE**. You should see an autocompletion box pop up with some suggestions:



What is the difference between QUARTILE.EXC and QUARTILE.INC? Take a moment and look it up!

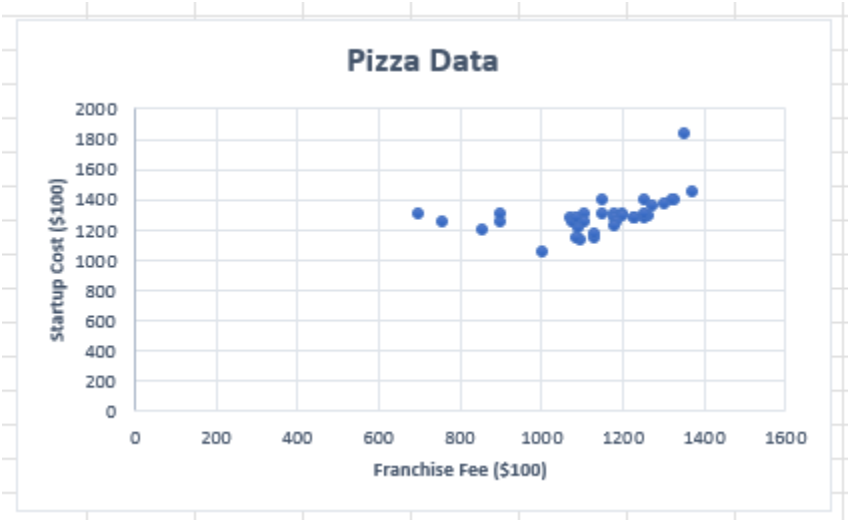
Task:

- Compute the 1st, 2nd, and 3rd quartile (use the method that Excel chose by default on your plot) for both distributions.
- Compute the mean, minimum, and maximum of both distributions.

Quartile	Franchise Fee	Startup Cost
1	1080	1250
2	1162.5	1277.5
3	1250	1300
Mean:	1134.777778	1291.05556
Min:	700	1050
Max:	1375	1830

So far, we've analyzed both columns **independently**. Can we assess the extent to which they are related?

Probably the simplest approach here is with a **scatter plot** and **correlation analysis**. Highlight the entire set of data and insert the plot:



To calculate the correlation, try **=CORREL(<array 1>,<array 2>)**.
(You substitute in the appropriate arrays!)

The correlation ranges from -1 (as one increases, the other decreases) to 0 (no relationship) to +1 (as one increases, so does the other).

Your turn!

Check out **4 – Fires and Thefts.xlsx** on D2L. X is the fires per 1000 housing units, and Y is the number of thefts per 1000 population.

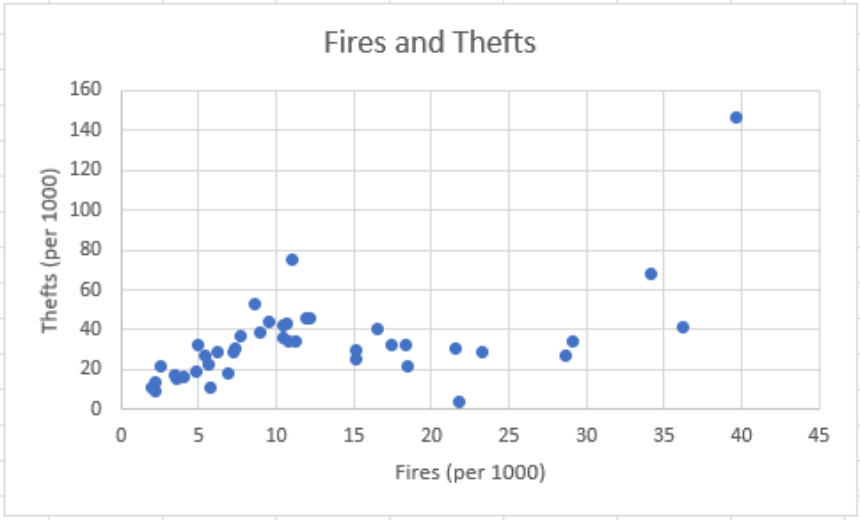
Task:

- Update the column headers.
- Create a box and whisker plot of both data sets.
- Determine the interquartile range for each set of data.
- Make a scatter plot with fires on the x axis and thefts on the y axis.
- Calculate the correlation coefficient between the data sets. Interpret.

Label your graphs and keep everything tidy in your workbook!

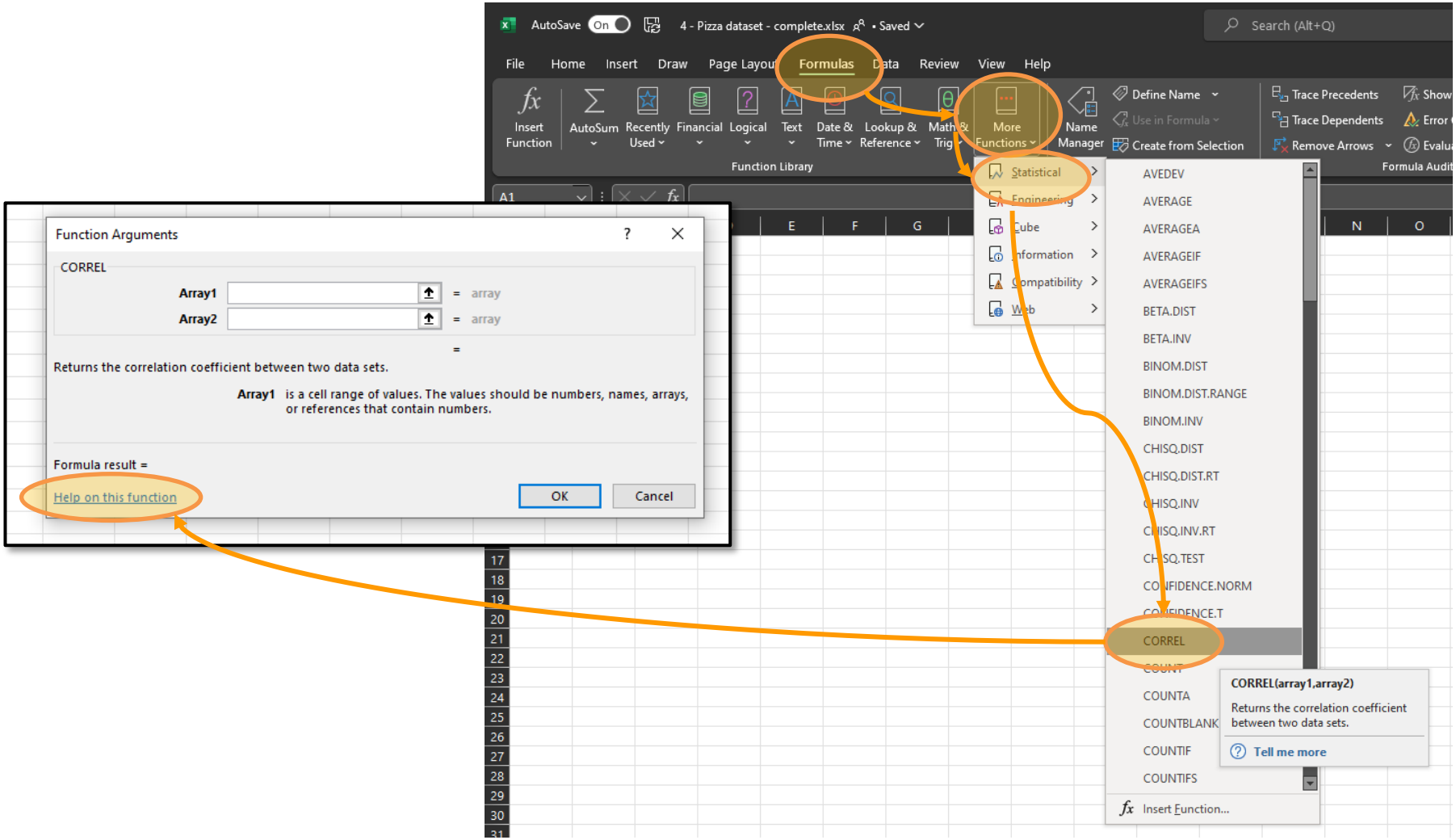
Your turn!

Quartile	Fires	Thefts
1	5.55	21.25
2	10.5	31
3	17.65	40.25
IQR:	12.1	19
Correl:	0.551121	
	(middle positive correlation)	



To see how the correlation is calculated, try looking up the official documentation.

Formulas Ribbon -> More Functions -> Statistical -> CORREL -> Help on this function



Let's calculate the fires & thefts correlation "by hand."

$$Correl(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Task:

- Calculate the average of each data set.
- Insert a new column; call it "x – x avg", and calculate the difference.
- Repeat the two steps above for a column "y – y avg".
- Make *another* new column that multiplies the "x – x avg" and "y – y avg" columns. At the bottom of the column, determine the sum. This is the numerator of the correlation function.
- Continue in a similar fashion to calculate the denominator, then determine the overall correlation. Compare to your earlier result.
- Finally, insert a linear trendline on your scatter plot and show the equation and R^2 value. How does it compare to the correlation?

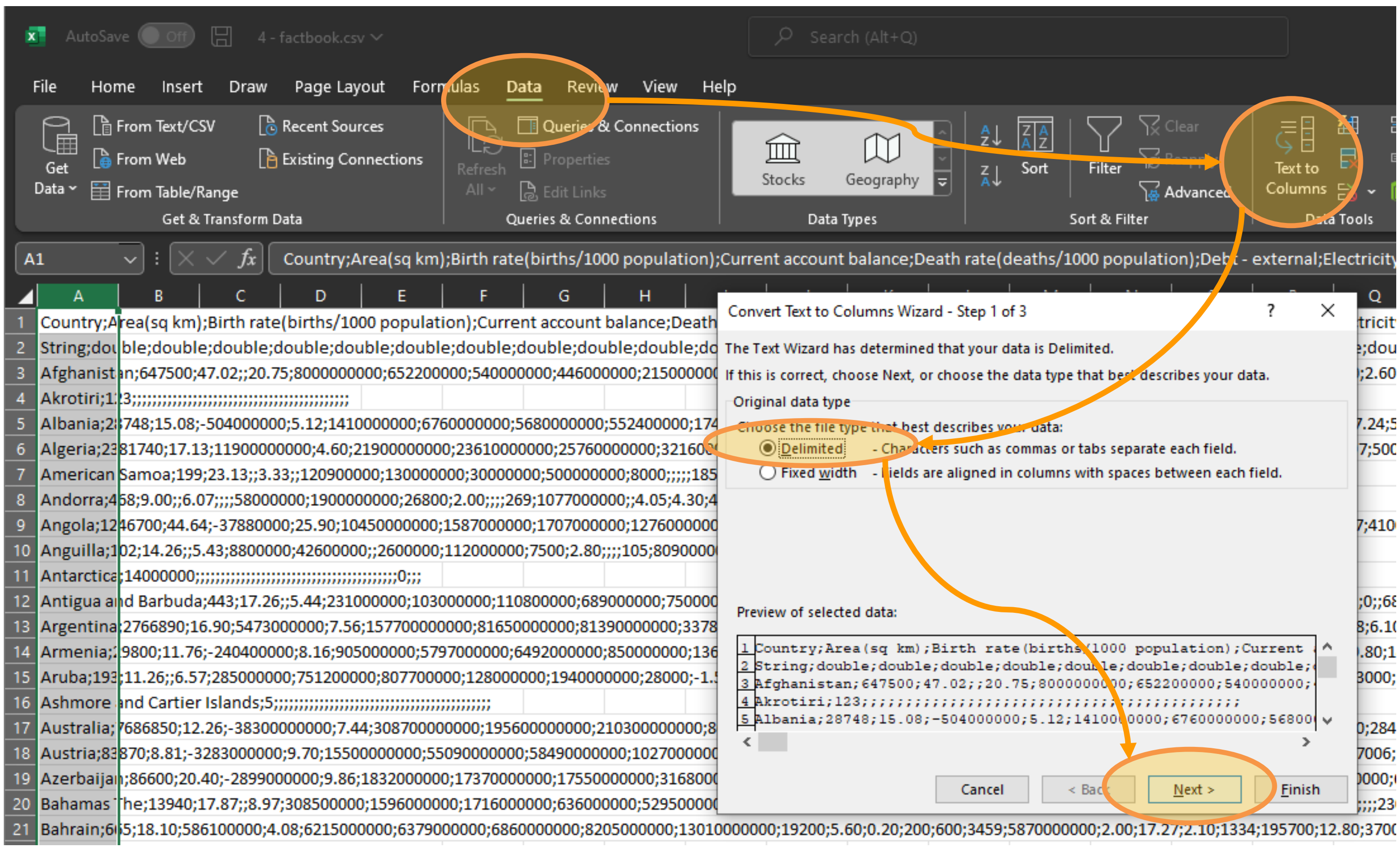
A brief aside on **types** of statistical analysis…

- **Descriptive Statistics** quantitatively describe the main features of some data (mean, median, mode, IQR…). Visualizations like a box and whisker plot give some sense of these properties.
- **Diagnostic Statistics** are used for discovery, or to determine *why* something happened (correlation is an example here, though correlation isn't the same as causation!).
- **Predictive Statistics/Analytics** attempts to forecast into the future (consider the line of best fit and **extrapolating** farther along the x axis than what the available data covers).
- **Prescriptive Analytics** is similar to predictive analytics but encompasses the situation where additional data can be coming in.
- **Exploratory Analytics** refers to the search for new patterns in data, and often encompasses quite a bit of visualization.

Sometimes we'd like to pull data in to Excel that isn't already in an .xlsx format!

Download **4 – factbook.csv** from D2L and open it in Excel. (It will look gross!). Select column A, then through the following:

Data Ribbon -> Text to Columns -> Delimited -> Semicolon -> Finish



Now that the data is cleaned up and readable, spend some time doing "exploratory data analysis."

At a minimum, do the following:

- Make 1 pie chart.
- Report 3 different correlations and visualize with a scatter plot.
- Make 1 box-and-whisker plot and identify which countries are high and low outliers.

Make sure your charts are labeled and look professional!

This data set was pulled from https://perso.telecom-paristech.fr/eagan/class/igr204/datasets