

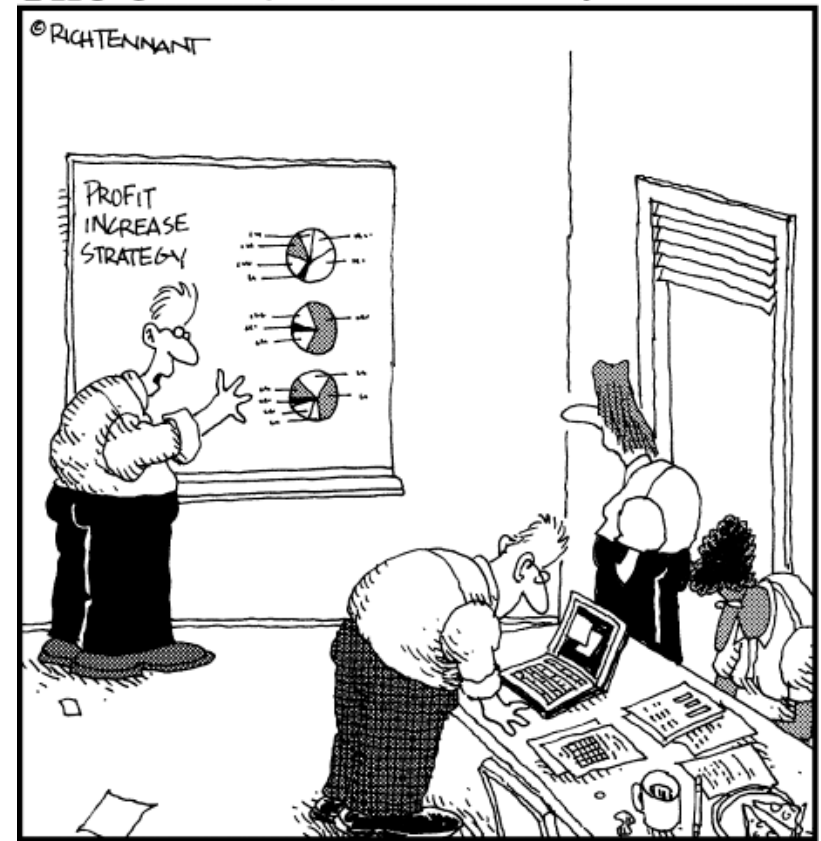
Data and Statistics in Excel



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

The 5th Wave

By Rich Tennant



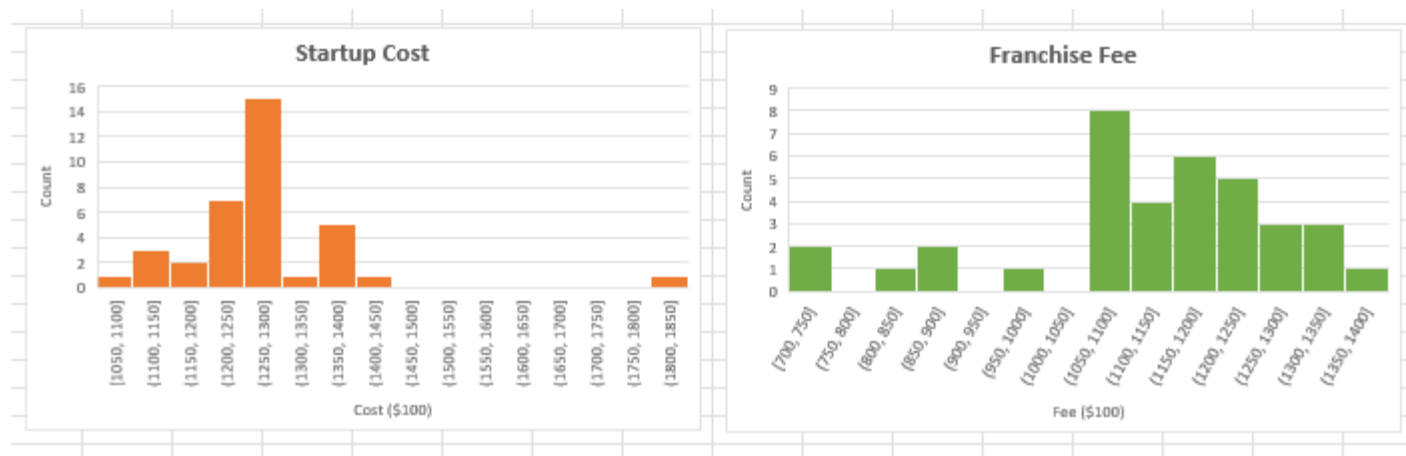
"Look-what if we just increase the size of the charts?"

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.

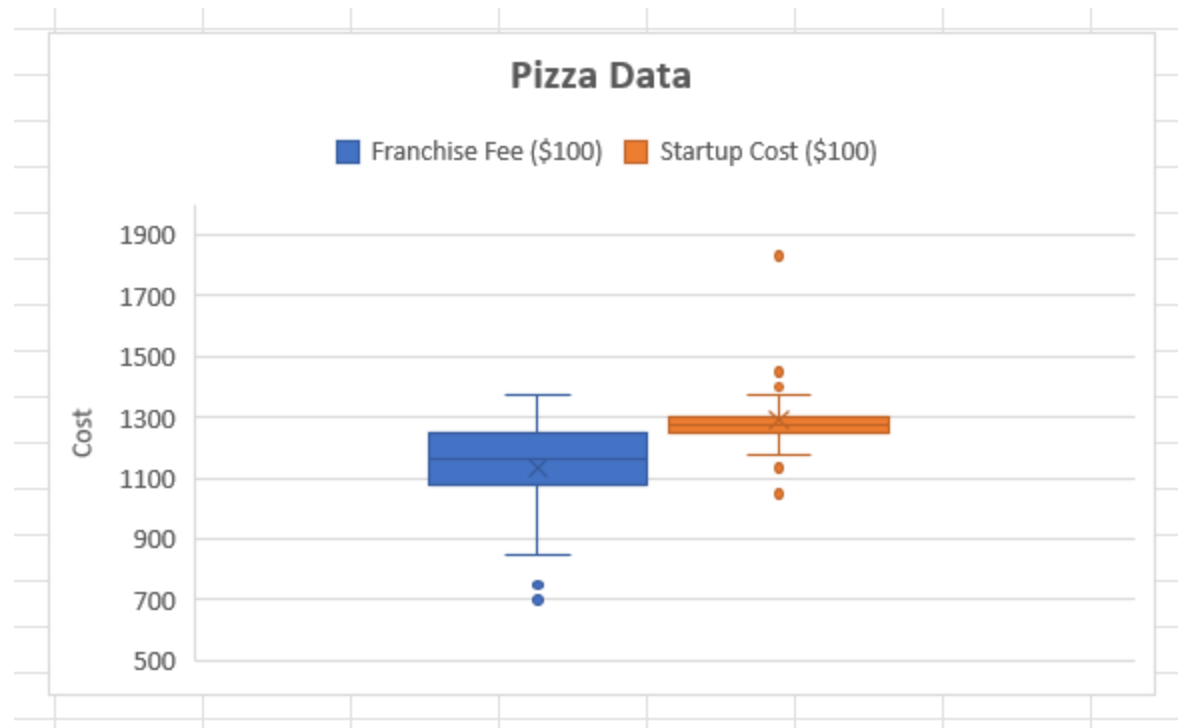


What information does a histogram show?

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:



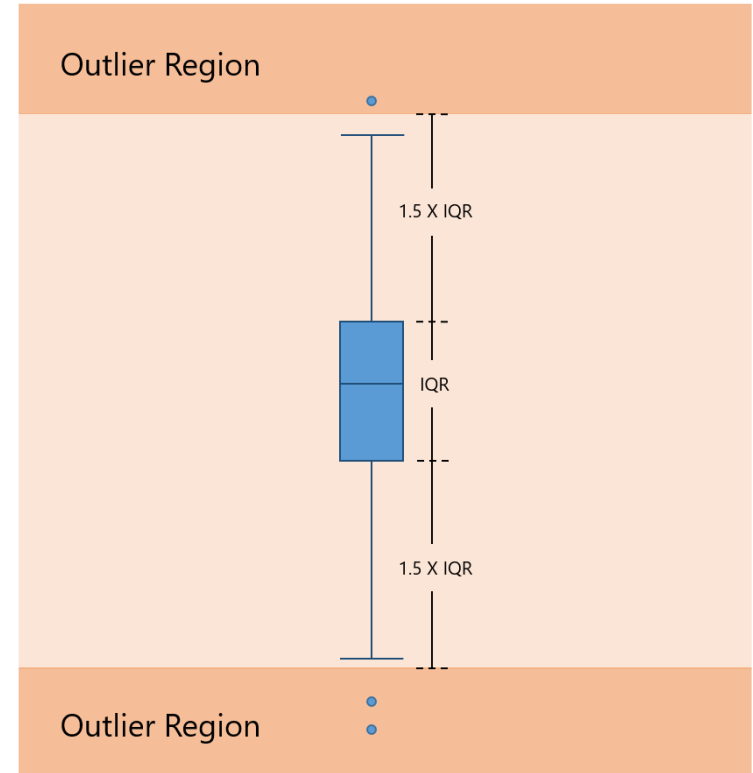
A boxplot is another useful way to visualize a distribution of data.

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)

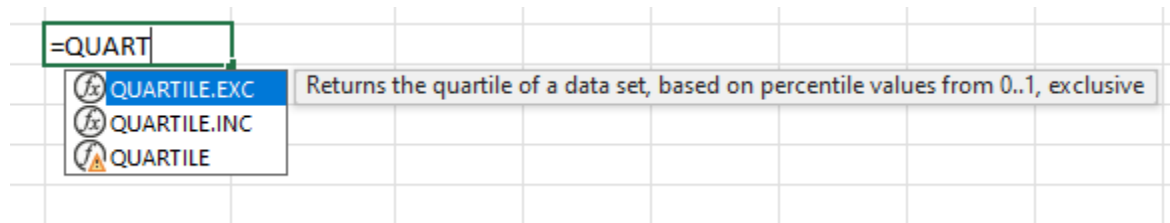
Finally, the “X” marks the **mean** (average).



[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!

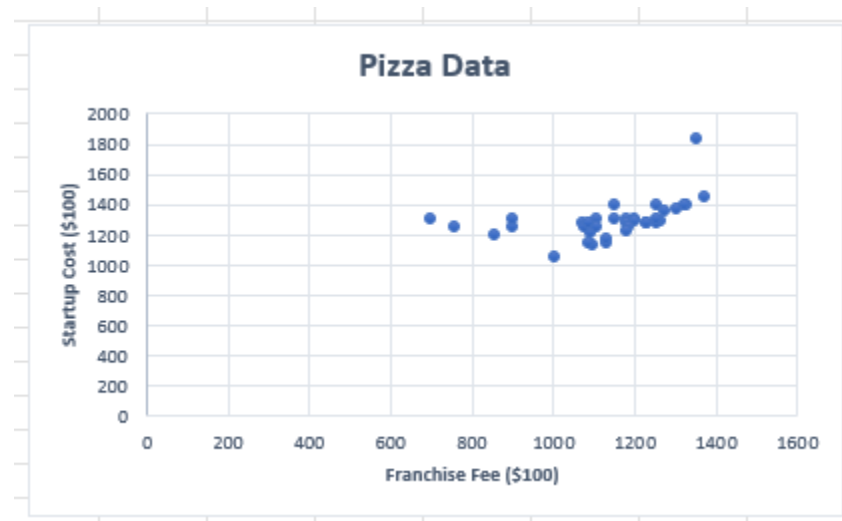
Task:

- 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:

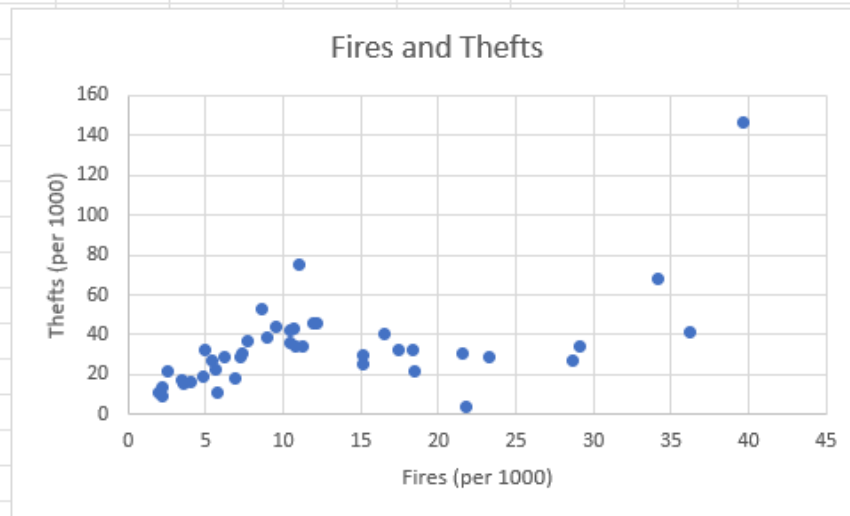
- Update the column headers.
- Create a box and whisker plot of both data sets.
- Determine the interquartile range for each set of data.
- Make a scatter plot with fires on the x axis and thefts on the y axis.
- Calculate the correlation coefficient between the data sets.
Interpret.

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

The image shows a screenshot of the Microsoft Excel interface. The 'Formulas' ribbon is active, and the 'More Functions' dropdown menu is open, showing the 'Statistical' category. The 'CORREL' function is selected in the list. A tooltip for the 'CORREL' function is visible, showing the formula 'CORREL(array1,array2)' and its description: 'Returns the correlation coefficient between two data sets.' A 'Tell me more' link is also present. In the foreground, the 'Function Arguments' dialog box for the 'CORREL' function is open, showing the 'Array1' and 'Array2' arguments. A 'Help on this function' link is circled in the dialog box. Orange arrows indicate the path from the 'Formulas' ribbon to 'More Functions', then to 'Statistical', then to 'CORREL', and finally to the 'Help on this function' link in the dialog box.

Function Arguments

CORREL

Array1 = array

Array2 = array

Returns the correlation coefficient between two data sets.

Array1 is a cell range of values. The values should be numbers, names, arrays, or references that contain numbers.

Formula result =

[Help on this function](#) OK Cancel

CORREL(array1,array2)
Returns the correlation coefficient between two data sets.
[Tell me more](#)

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

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

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

The screenshot shows the Microsoft Excel interface with the 'Data' ribbon selected. The 'Text to Columns' button in the 'Data Tools' group is highlighted with an orange circle. Below it, the 'Convert Text to Columns Wizard - Step 1 of 3' dialog box is open. The wizard indicates that the data is delimited and offers two options: 'Delimited' (selected) and 'Fixed width'. The 'Delimited' option is further highlighted with an orange circle. The preview of the selected data shows the first five rows of the CSV file, with semicolons separating the fields. The 'Next >' button at the bottom right of the wizard is also highlighted with an orange circle.

Country	Area(sq km)	Birth rate(births/1000 population)	Current account balance	Death rate(deaths/1000 population)	Debt - external	Electricity																	
Afghanistan	647500	47.02	20.75	8000000000	6522000000	540000000	215000000																
Akrotiri	123																						
Albania	28748	15.08	-5040000000	5.12	14100000000	6760000000	5680000000	552400000	174														
Algeria	2381740	17.13	11900000000	4.60	21900000000	23610000000	25760000000	321600															
American Samoa	199	23.13	3.33	120900000	130000000	30000000	500000000	8000	185														
Andorra	458	9.00	6.07	580000000	1900000000	26800	2.00	269	1077000000	4.05	4.30	4											
Angola	1246700	44.64	-378800000	25.90	10450000000	1587000000	1707000000	1276000000															
Anguilla	102	14.26	5.43	8800000	42600000	2600000	112000000	7500	2.80	105	8090000												
Antarctica	14000000											0											
Antigua and Barbuda	443	17.26	5.44	231000000	103000000	110800000	689000000	750000															
Argentina	2766890	16.90	5473000000	7.56	157700000000	81650000000	81390000000	3378															
Armenia	9800	11.76	-240400000	8.16	905000000	5797000000	6492000000	850000000	136														
Aruba	193	11.26	6.57	285000000	751200000	807700000	128000000	1940000000	28000	1													
Ashmore and Cartier Islands	5																						
Australia	7686850	12.26	-38300000000	7.44	308700000000	195600000000	210300000000	8															
Austria	83870	8.81	-3283000000	9.70	15500000000	55090000000	58490000000	1027000000															
Azerbaijan	86600	20.40	-2899000000	9.86	1832000000	17370000000	17550000000	316800															
Bahamas	13940	17.87	8.97	308500000	1596000000	1716000000	636000000	52950000															
Bahrain	65	18.10	5861000000	4.08	62150000000	63790000000	68600000000	8205000000	130100000000	19200	5.60	0.20	200	600	3459	5870000000	2.00	17.27	2.10	1334	195700	12.80	3700

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!