Data Analysis: A Practical Introduction for Absolute Beginners

Module 2, Lab 2: Data Walkthrough

Learning Objectives

- Analyze a real-world data set.
- Find the average of one particular variable in a data set.
- Create a scatterplot in Excel showing the relationship between two variables.
- Compute the correlation coefficient in Excel between two variables.

Data Set

Mod2Labs.csv

What You'll Need

To complete the lab, you will need the online version of Microsoft Excel.

Overview

In this lab, we'll sort variables, average variables, and explore the relationship between two different variables using our car data from the lesson. We'll also play around a bit with visualizing our data in a scatterplot.

Exercise 1: Sorting and Averaging by Engine Size

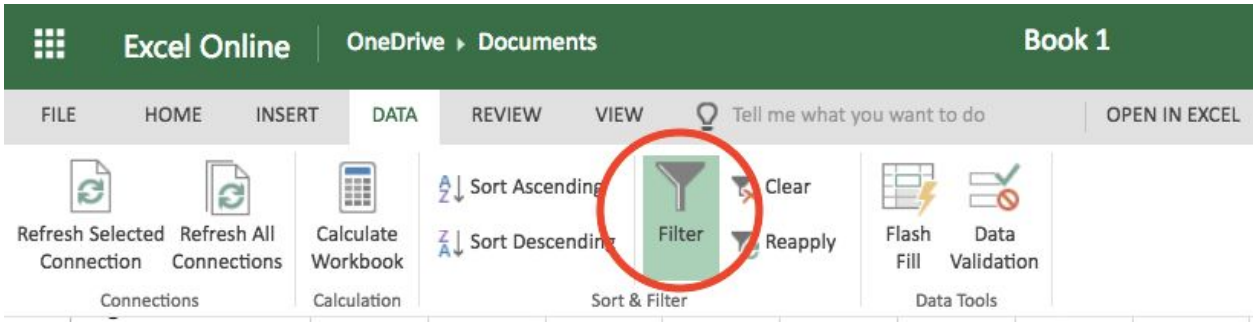
1. Open the data set in Excel. You should see several different variable columns for a group of car models. Here's a snapshot of the data:

	A	B	C	D	E	F	G	H	I	J	K	L
1	model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
2	Knight X	21	6	160	110	3.9	2.62	16.46	0	1	4	4
3	Knight X Wagon	21	6	160	110	3.9	2.875	17.02	0	1	4	4
4	Hercules 100	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
5	Wasp 4WD	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
6	Wasp Supersport	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
7	El Pasion	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
8	Road Devil	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
9	Anansi 100	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
10	Anansi 200	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
11	Anansi 200x	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4

   Once again, here's what each variable/column represents:

model = name of car model
mpg = gas mileage, in miles per (US) gallon
cyl = number of cylinders
disp = displacement, in cubic inches
hp = gross horsepower
drat = rear axle ratio
wt = weight, in thousands of pounds (1000 lb)
qsec = 1/4 mile time
vs = engine (0 = V-shaped, 1 = straight)
am = transmission (0 = automatic, 1 = manual)
gear = number of forward gears
carb = number of carburetors

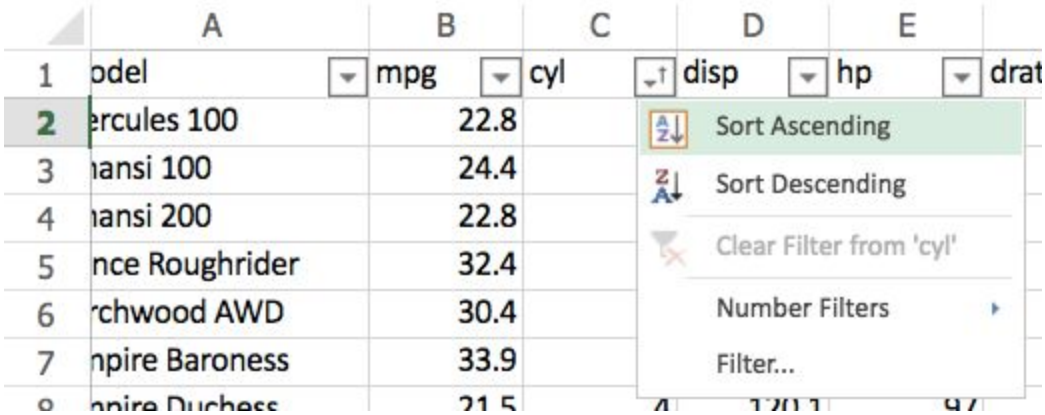
2. With the data set open, click anywhere in the spreadsheet, go to the Data tab in the ribbon, and click Filter (in the Sort & Filter tab).



3. You should now see a little dropdown arrow show up next to each column title, like so:

	A	B	C	D	E	F	G	H	I	J	K	L
1	model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
2	Knight X	21	6	160	110	3.9	2.62	16.46	0	1	4	4
3	Knight X Wagon	21	6	160	110	3.9	2.875	17.02	0	1	4	4
4	Hercules 100	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
5	Wasp 4WD	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
6	Wasp Supersport	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2

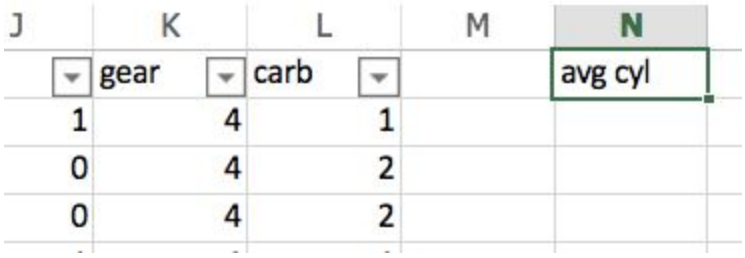
4. Click the arrow next to the number of cylinders (“cyl”), then click Sort Ascending to sort the cars by engine size.



Looks like the cars in our data have engine sizes that range from 4 to 8 cylinders.

	A	B	C	D	E	F	G	H	I	J	K	L
1	model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
2	Hercules 100	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
3	Anansi 100	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
4	Anansi 200	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
5	Lance Roughrider	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
6	Birchwood AWD	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
7	Empire Baroness	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
8	Empire Duchess	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
9	Apocalypse 100	27.3	4	79	66	4.08	1.935	18.9	1	1	4	1
10	Prince of Thieves	26	4	120.3	91	4.43	2.14	16.7	0	1	5	2

5. Now find the mean/average number of cylinders among the cars on this list. Start by creating a new column for this average, out to the right-hand side of the rest of the data.



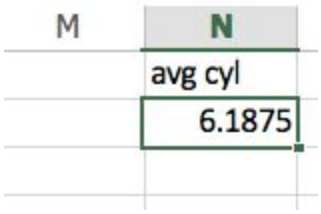
6. Use the AVERAGE function in Excel to find the mean/average number of cylinders. If you recall from our video lesson, the mean/average is just the sum of all the scores divided by the number of scores. Thankfully, Excel can crunch those numbers for us.

   (Note: It's not a requirement that you sort the data before averaging it. Sorting just makes everything easier to see.)

   The syntax for Excel's AVERAGE function is **=AVERAGE(first cell:last cell)**. Since you want the average number of cylinders, use column C as your range of data. You can either type in C2 as

the first cell and C33 as the last cell, or you can just type in =AVERAGE( ), click inside the parentheses, and highlight all the cells from C2 down to C33.

Once you hit Enter, the program will calculate the average for you from that column.



So the average car from this data set had about 6.1875 cylinders.

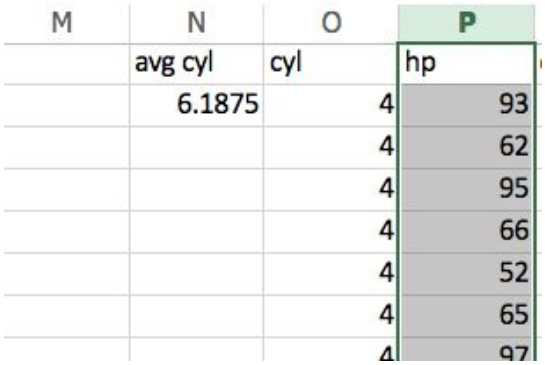
Exercise 2: Creating a Scatterplot

Next, we'll create a graph in Excel that shows the correlation between two different variables. Let's see if there's a correlation (relationship) between engine size and horsepower.

1. With the car data set open in Excel Online, highlight all of column C ("cyl") and copy the data (Ctrl + C on Windows, Command + C on Mac). Paste the data into a new column, off somewhere to the right-hand side of the rest of the spreadsheet — we're pasting it into column O in this example.

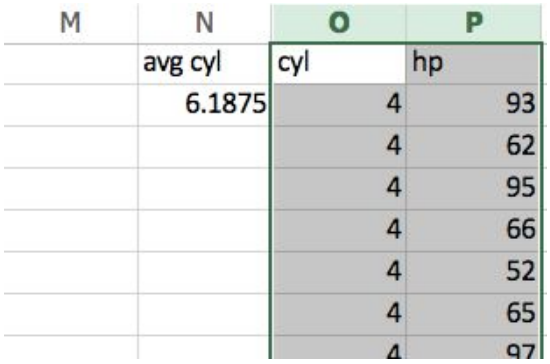
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb		avg cyl	cyl
2	Hercules 100	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1		6.1875	4
3	Anansi 100	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2			4
4	Anansi 200	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2			4
5	Lance Roughrider	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1			4
6	Birchwood AWD	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2			4
7	Empire Baroness	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1			4
8	Empire Duchess	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1			4
9	Apocalypse 100	27.3	4	79	66	4.08	1.935	18.9	1	1	4	1			4

2. Repeat Step 1 with column E ("hp"). Paste all the horsepower data into a column adjacent to the "cyl" data you just pasted. In this example, that's column P.



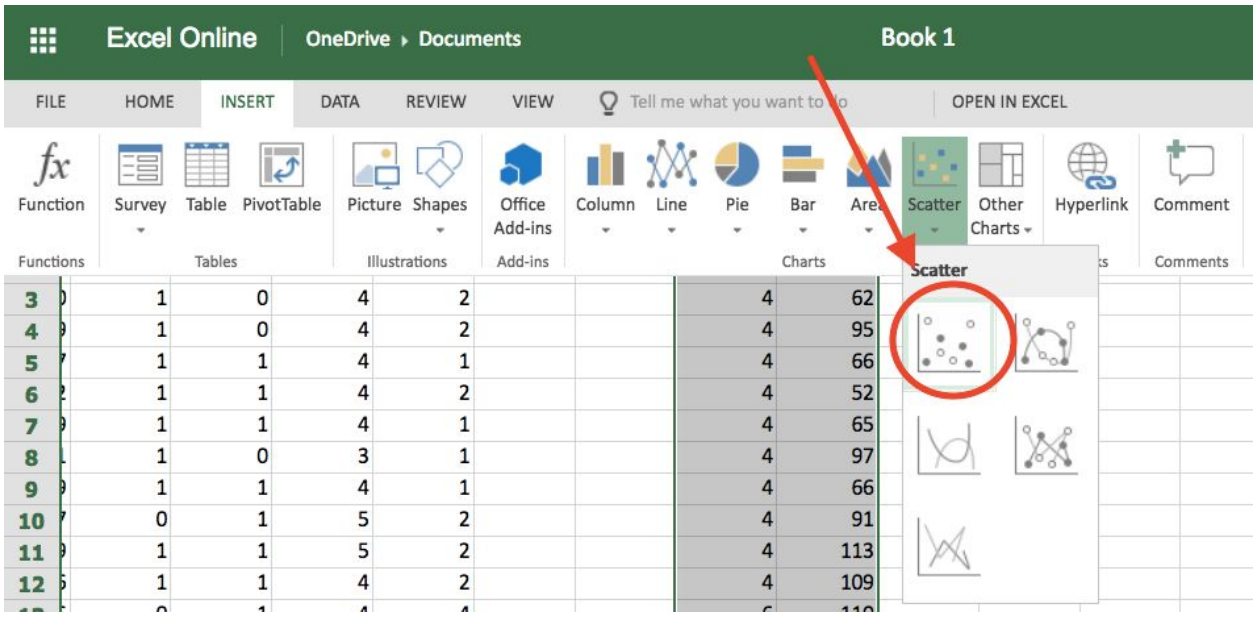
Note: Excel Online doesn't support selecting multiple non-adjacent cells or columns, which is why we're copying and pasting the two columns before creating the chart. If you have access to a desktop version of Excel, you can instead use Ctrl + click (or Command + click on a Mac) to select non-adjacent data and skip the copy/paste part.

3. Select your two new adjacent columns of data (columns O and P).



M	N	O	P
	avg cyl	cyl	hp
	6.1875	4	93
		4	62
		4	95
		4	66
		4	52
		4	65
		4	97

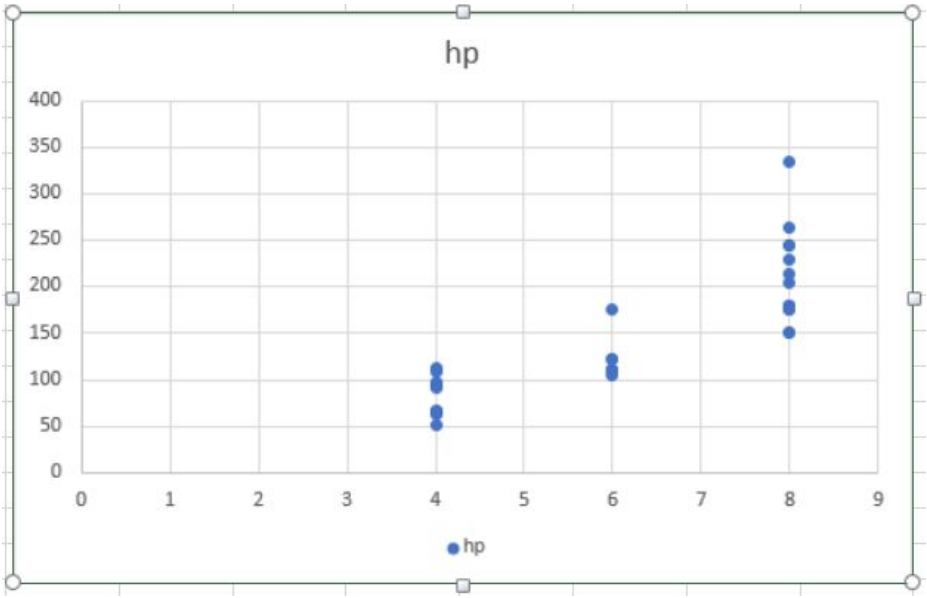
With both columns of data selected, click the Insert tab in the ribbon, then click Scatter > Scatter only with Markers.



3	0	1	0	4	2		4	62
4	9	1	0	4	2		4	95
5	7	1	1	4	1		4	66
6	2	1	1	4	2		4	52
7	9	1	1	4	1		4	65
8	1	1	0	3	1		4	97
9	9	1	1	4	1		4	66
10	7	0	1	5	2		4	91
11	9	1	1	5	2		4	113
12	5	1	1	4	2		4	109

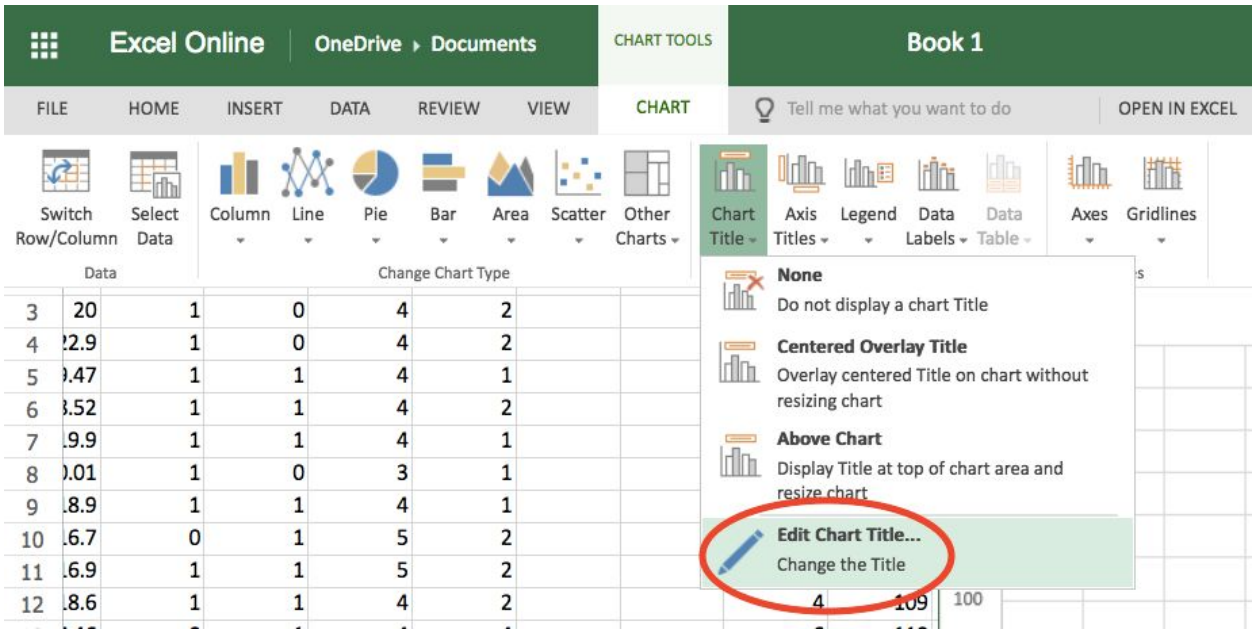
(Again, if you're using a desktop version of Excel instead of Excel Online, you can just select the two non-adjacent "cyl" and "hp" columns directly from the original data instead of copy/pasting.)

4. This should create a simple scatterplot showing the "cyl" (cylinder) data on the *x*-axis and the "hp" (horsepower) data on the *y*-axis, like so:

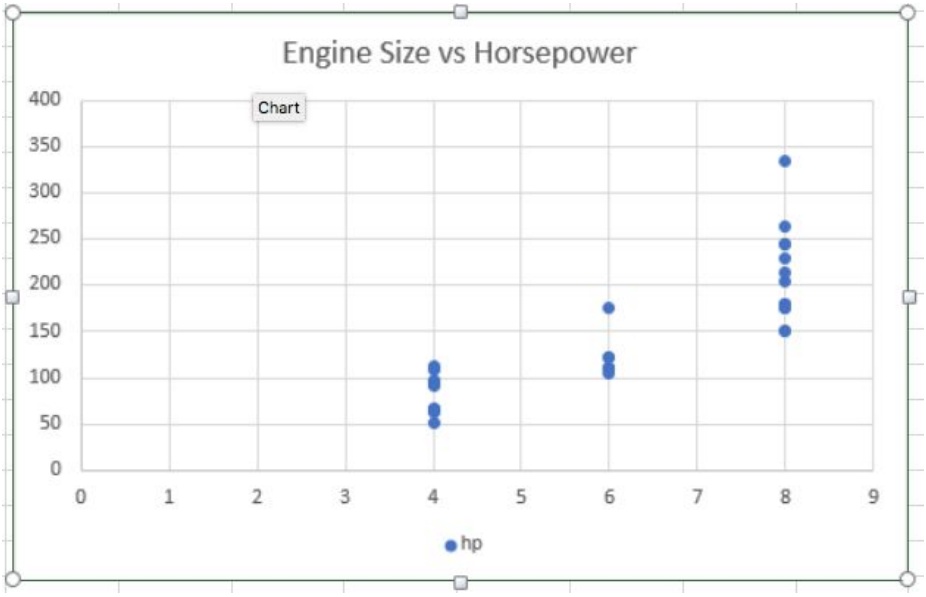


You'll need to add some titles and legends to make the chart easier to read, though.

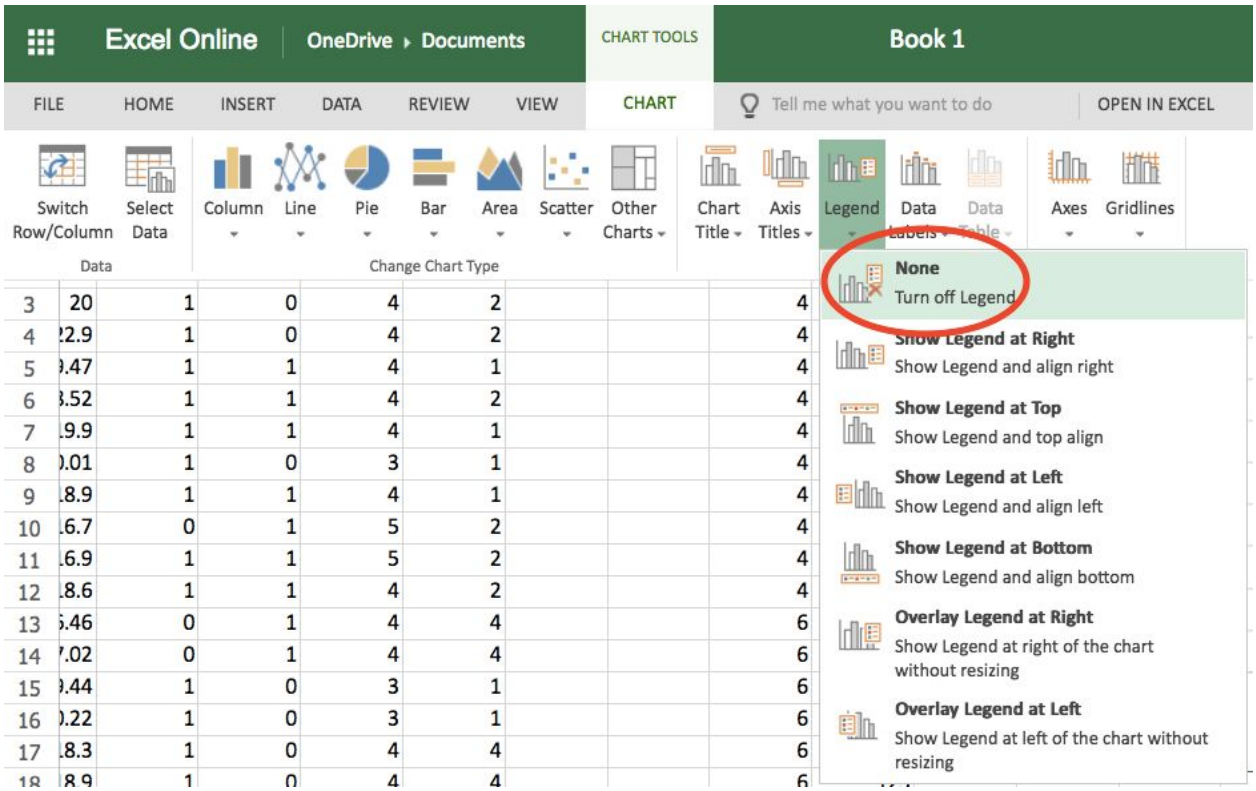
5. To add a better chart title, click on the scatterplot itself, then click Chart Tools in the ribbon and go to Chart Title > Edit Chart Title.



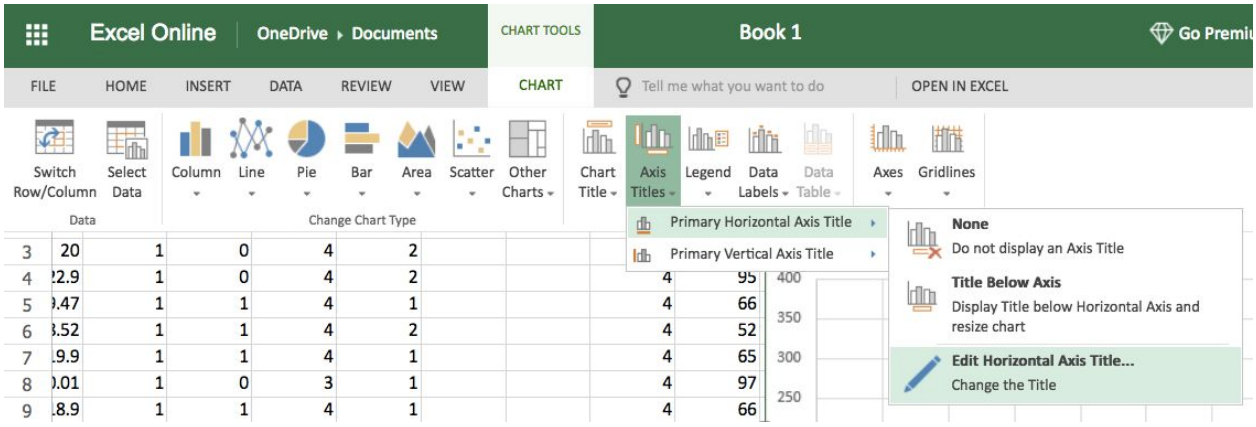
Type in "Engine Size vs Horsepower" under Title Text, then click OK. Your scatterplot now has a shiny new title.



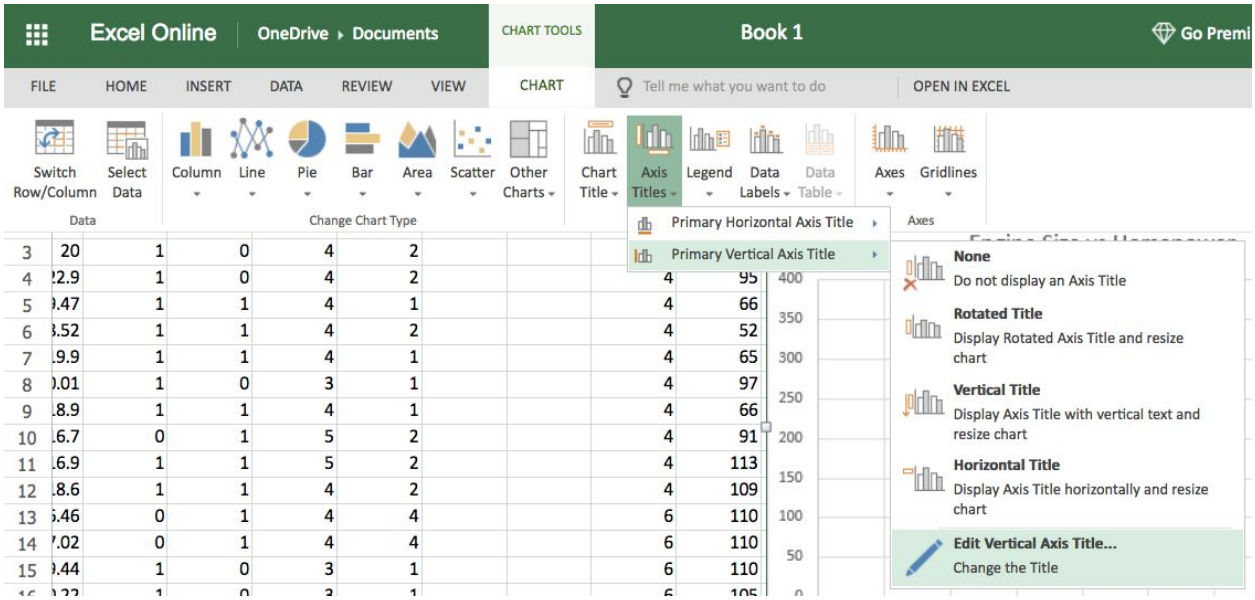
6. Click into Chart Tools in the ribbon again, and click Legend > None to get rid of that little blue dot at the bottom that says “hp.”



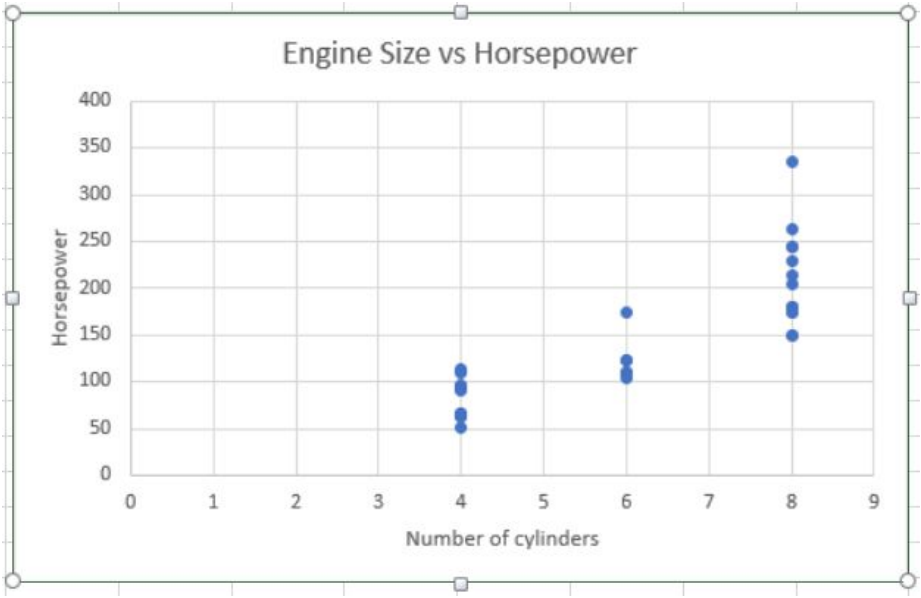
7. Finish up by adding titles to the *x*- and *y*-axis. To do this, head back to Chart Tools in the ribbon again, then click Axis Titles > Primary Horizontal Axis Title > Edit Horizontal Axis title. Type “Number of Cylinders” and hit OK.



8. Do the same thing for the vertical axis: Go to Chart Tools > Axis Titles > Primary Vertical Axis Title > Edit Vertical Axis Title. Type in “Horsepower” as the title and click OK.



9. Now your scatterplot should be nice and easy to read.



You can see from the plot that there does seem to be a correlation between an engine's number of cylinders and its horsepower. The higher dots on the right show that a higher number of cylinders usually means a higher horsepower. We call that a **positive correlation**.

Exercise 3: Correlating Two Variables

We saw in the previous exercise that the engine size and horsepower variables were correlated (as one increases, so does the other), but the questions is: *How* correlated are they? Luckily, Excel's CORREL function can give a specific value that shows how strong the relationship is between those two variables.

1. With the car data set open in Excel Online, set up a new column for the correlation between the "cyl" and "hp" variables. We'll use the blank column M.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	correlation
2	Hercules 100	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1	
3	Anansi 100	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2	
4	Anansi 200	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2	
5	Lance Roughrider	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1	
6	Birchwood AWD	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2	
7	Empire Baroness	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1	

2. Use the CORREL function to track down a value called the **correlation coefficient**. This is a number between -1 and 1 that shows how strongly two variables are correlated.

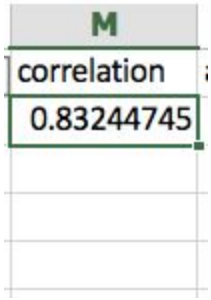
   A value of -1 means the two variables have a perfect negative correlation (e.g. as one variable increases, the other decreases at exactly the same rate). A value close to -1 (like -0.9) indicates a **strong negative correlation**.

   A value of 0 means the two variables aren't correlated at all. Negative values close to 0 (like -0.1) indicate a **weak negative correlation**. Positive values close to 0 (like 0.1) indicate a **weak positive correlation**.

A value of 1 means the two variables have a perfect positive correlation (e.g. as one variable increases, the other increases at exactly the same rate). A value close to 1 (like 0.9) indicates a **strong positive correlation**.

Here's the syntax: **=CORREL(first range of values, second range of values)**. In this case, you want the correlation coefficient between the number of cylinders ("cyl") and the horsepower ("hp"), so your first range of values is everything in column C (starting with cell C2) and your second range of values is everything in column E (starting with cell E2).

Hit Enter.



3. The correlation coefficient is about 0.832. That's getting fairly close to the max value of 1, so this value indicates a fairly strong positive correlation. That makes sense, since the scatterplot we made in Exercise 2 showed that the horsepower increased as the number of cylinders increased. In other words, the two variables are positively correlated.