Excel manual for STA 225, Fall 2025

One Quantitative Response Variable • One quantitative explanatory variable (Simple Linear Regression) ○ Descriptive statistics and graphical summaries • Scatterplot (including detection of non-linearity and outliers) • SLR model, correlation, r-square ○ Inferential statistics • P-value for testing slope, CI for slope • checking conditions	30-37	2.4, 14.1, 14.5, 14.6, 14.7,14.8, 14.9
One Quantitative Response Variable • Two or more quantitative explanatory variables (Multiple Linear Regression) ○ Descriptive statistics and graphical summaries • Scatterplots	38-39	2.4
One Quantitative Response Variable • Two or more quantitative explanatory variables (Multiple Linear Regression) ○ Descriptive statistics and graphical summaries • Correlation matrix (including collinearity)	40-41	15.5
One Quantitative Response Variable • Two or more quantitative explanatory variables (Multiple Linear Regression) ○ Descriptive statistics and graphical summaries • MLR model, R-square, Adj R-square ○ Inferential statistics • P-value for overall F test, p-value for individual slopes • Checking conditions	42-44	15.1, 15.3, 15.5, 15.8
One categorical response variable • No explanatory variables ○ Descriptive statistics and graphical summaries • Proportion • Bar chart and pie chart	45-50	2.1
One categorical response variable • No explanatory variables ○ Inferential statistics • Confidence interval and margin of error	51	8.4

General comments:

STA 225 will be taught using the Excel “add-in” called the ***Data Analysis Toolpak***. This was specifically requested by Seidman College of Business (SCB) faculty, so you should use the Data Analysis Toolpak.

Here is how to verify within Excel if the Excel Data Analysis Toolpak is activated, and how to activate if not already activated:

https://www.goskills.com/Excel/Resources/Excel-data-analysis-toolpak

This manual follows the order of topics that was used in development of STA 225 in cooperation with SCB faculty. However, individual STA faculty can determine the order of topics to use in their STA 225 sections.

This manual was developed using the version of Excel that is on the GVSU network in Fall 2025.

We assume that the reader understands that “response variable” is equivalent to “dependent variable” or “output variable”, and “explanatory variable” is equivalent to "independent variable” or “predictor variable” or “input variable”.

We acknowledge that there is always more than one way to create statistical output with Excel, so we show what we think is the simplest approach here that would be sufficient at the STA 225 level, while utilizing the Data Analysis Toolpak when that contains the simplest implementation for a given task. We generally shy away from using Excel functions, but keep in mind that your instructor may have another way that they want you to handle these issues.

Important note on data structure: Data that include a categorical variable can be organized in two different ways, and it is important to note which you have before you start analyzing it, because the data used with our textbook sometimes will be organized one way and sometimes the other, and this happens in the real world every day too. The basic difference is whether or not all the variable values and categories in each individual row in the spreadsheet correspond to one individual subject (also called observational unit). If each row corresponds to only one individual subject, the data are often called “tidy”.

For example, *the following two spreadsheets represent the same data*, but they are structured differently. For each university represented in the data, there is a quantitative variable called percent that is the percentage of students (out of 100%) from out of state, and also a categorical variable called type that is if the university is public or private. The data on the left are “untidy” because each row has percentages from two different universities that have no connection to each other (so the data are not paired), and another indicator is that there are more values in the public column than in the private column. Put another way, there is no variable observed from each university called “public” or “private”. However, the data on the right are “tidy” because both variables (percent and type) in each row come from the same university, so in other words, “percent” and “type” are variables observed from each university. Note that not all values from the actual data set were included in the following images.

	A	B
1	**Private**	**Public**
2	52.8	20.3
3	43.2	22.0
4	45.0	28.2
5	33.3	15.6
6	44.0	24.1
7	30.6	28.5
8	45.8	22.8
9	37.8	25.8
10		18.5
11		25.6
12		14.4

Untidy data

	A	B
1	**percent**	**type**
2	52.8	private
3	43.2	private
4	45.0	private
5	33.3	private
6	44.0	private
7	30.6	private
8	45.8	private
9	37.8	private
10	20.3	public
11	22.0	public
12	28.2	public
13	15.6	public
14	24.1	public
15	28.5	public
16	22.8	public
17	25.8	public
18	18.5	public
19	25.6	public
20	14.4	public

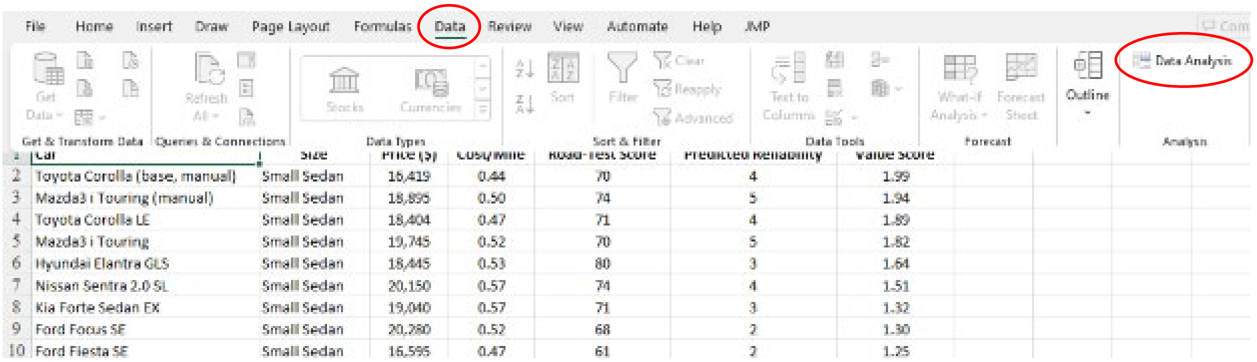
Tidy Data

As you can probably guess from the name "tidy", it is generally considered best practice to organize data in tidy form, but we can handle either tidy or untidy data with Excel, but you have to first recognize which you have, because how you use Excel depends to some extent on if the data are tidy or not. The good news is that it is easy to convert from one format to another just by doing a little copying and pasting in Excel, but that isn't necessary if you're willing to have some attention to detail. The point is that it is very important that you are aware of this issue going forward, regardless of if you choose to use tidy or untidy data in your analyses. You should follow the directions of your instructor on this issue.

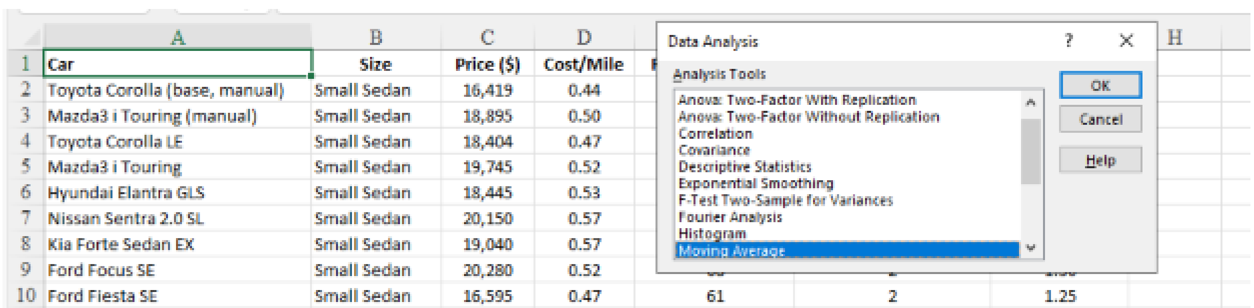
Some of the following analyses were implemented using the *carvalues* data set that comes with the Camm et al textbook (Ch 15) and others come from the *collegecosts* data set (Ch 10) partially shown above, or the *golfscores* data set (Ch 10) for a paired data example. Here are the first 10 rows of the *carvalues* data set to help understand the variables (this data set is in tidy format):

	A	B	C	D	E	F	G
1	**Car**	**Size**	**Price ($)**	**Cost/Mile**	**Road-Test Score**	**Predicted Reliability**	**Value Score**
2	Toyota Corolla (base, manual)	Small Sedan	16,419	0.44	70	4	1.99
3	Mazda3 i Touring (manual)	Small Sedan	18,895	0.50	74	5	1.94
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25

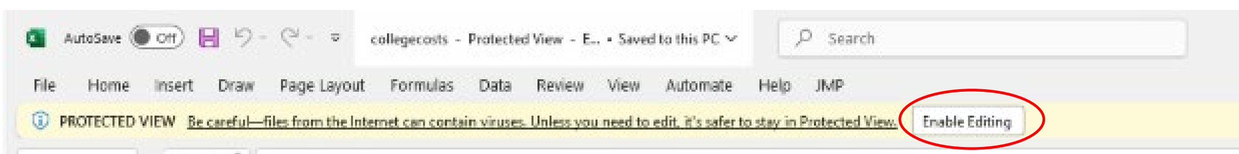
If the Data Analysis Toolpak add-in is activated, you can access the features by going to the Data tab at the top middle, and then click on “Data Analysis” on the far right at the top.



This will show you a list of features you can choose from, and we will refer to these in the rest of the document.



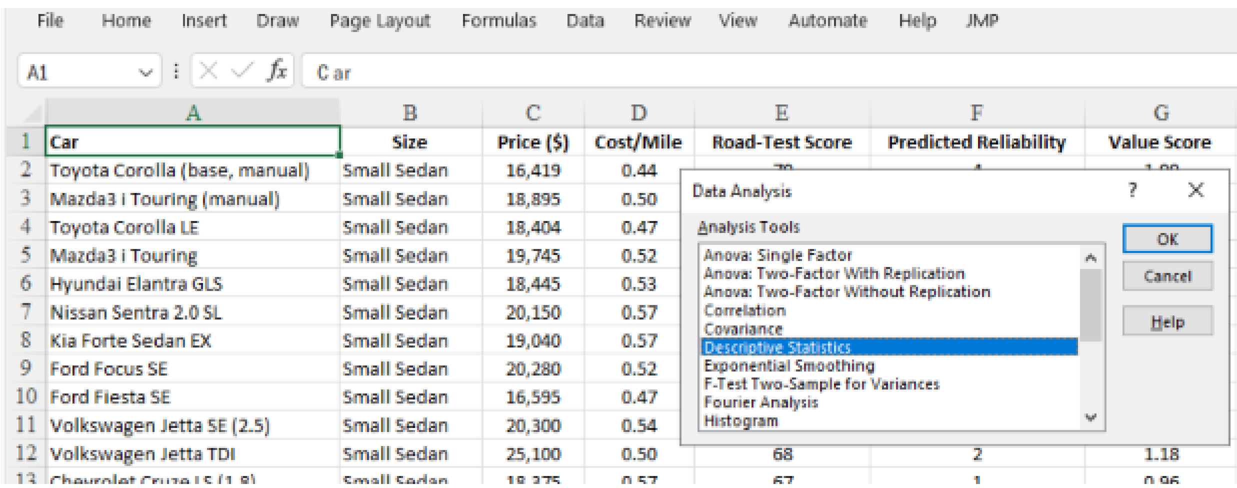
Important note: If you cannot get any of the features of this add-in to work, make sure you have clicked on “Enable Editing” if that is an option.



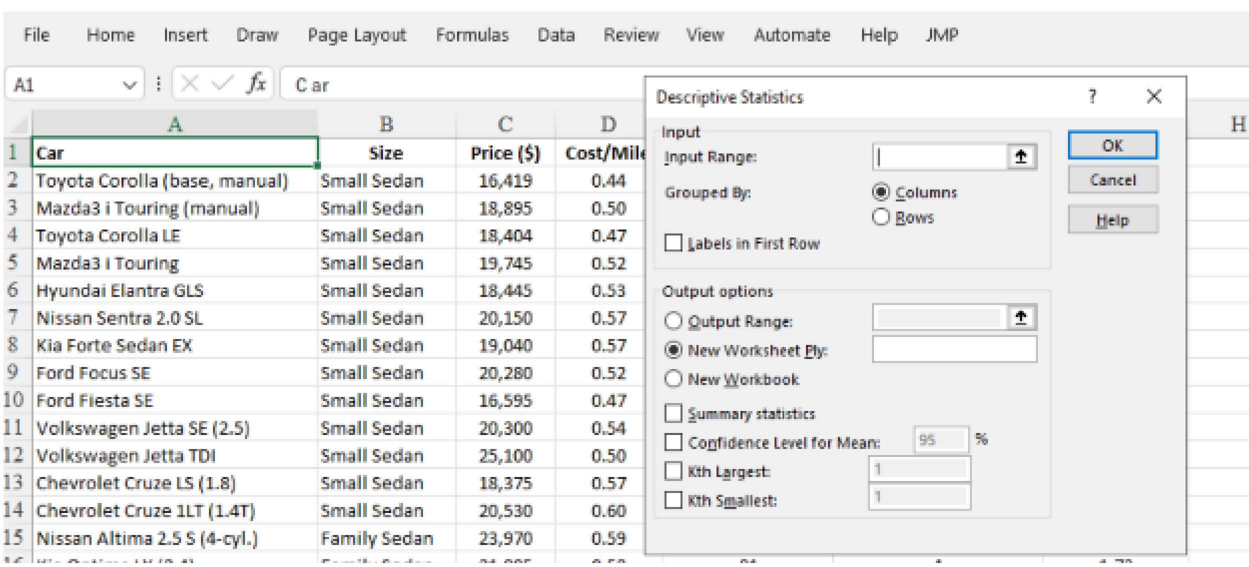
One quantitative response variable

- **No explanatory variables**
  - **Descriptive statistics**
    - **Mean, median, standard deviation, IQR, range, percentiles**
    - **5 number summary (min, Q1, median, Q3, max)**

To get Excel to calculate most of the basic descriptive statistics for one quantitative variable, we use the Data Analysis Toolpak add-in, and use the Price variable in the carvalues data set. As described above, click on the Data tab, and then the Data Analysis icon on the far right at the top. Select Descriptive Statistics option as shown below, and click OK

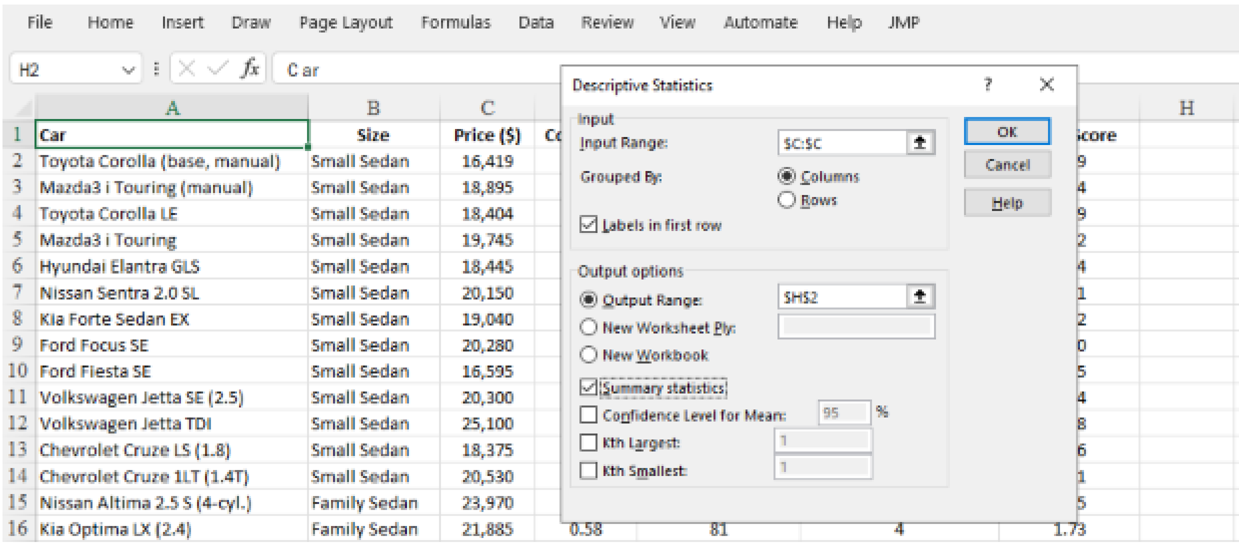


Which should show this dialog

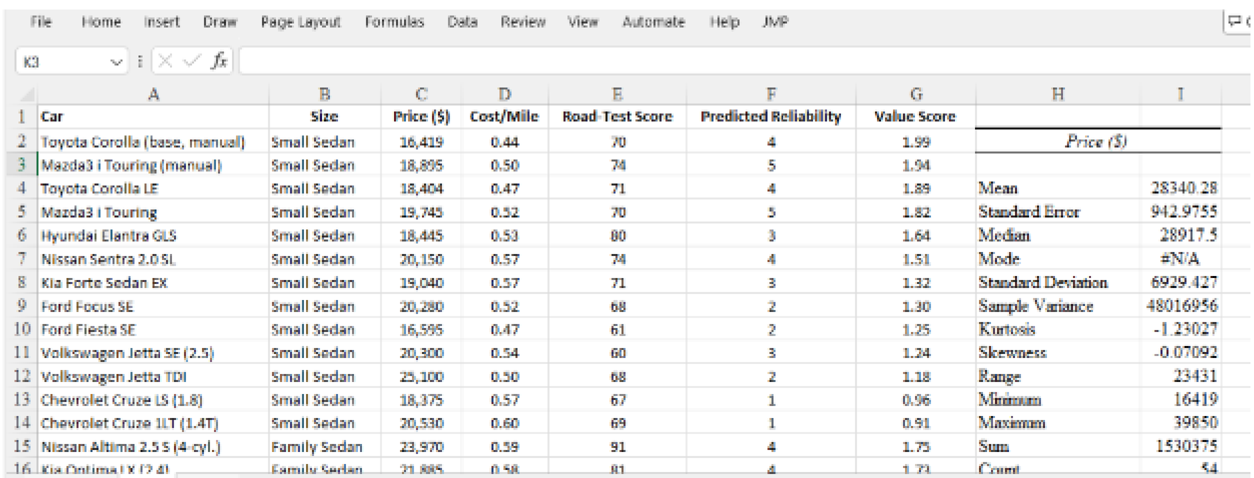


Single click in the Input Range box. Because it is appropriate to include every value of Price in this scenario, we can highlight the entire column C, but in general be careful to only highlight values that are relevant. Check the box for Labels in first row, because the variable name Price is in the first row.

To paste the output into the same spreadsheet, click on Output Range, single click in the box next to Output Range, and then click on a cell where you want the output. Check the box for Summary Statistics.



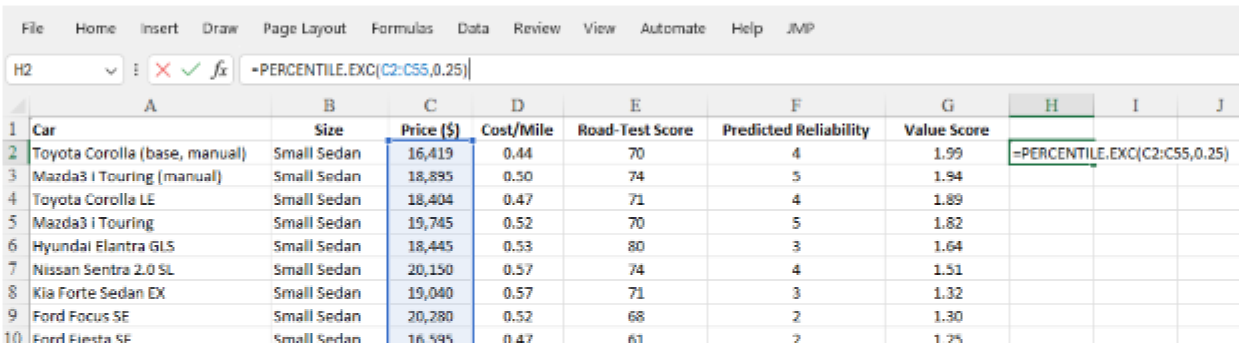
You should now see many of the common descriptive statistics to the right of the data, though we will have to do something else to get Q1, Q3 and IQR.



Excel uses a slightly different method to calculate percentiles. You should read the percentiles portion of Section 3.1 in the textbook to see how this is done. The short answer is that linear interpolation is used, so be aware that other ways you may have seen to hand calculate percentiles will likely not match what Excel is doing. The different methods generally produce similar answers, and because Excel is the official software for STA 225, we will follow the method used by Excel.

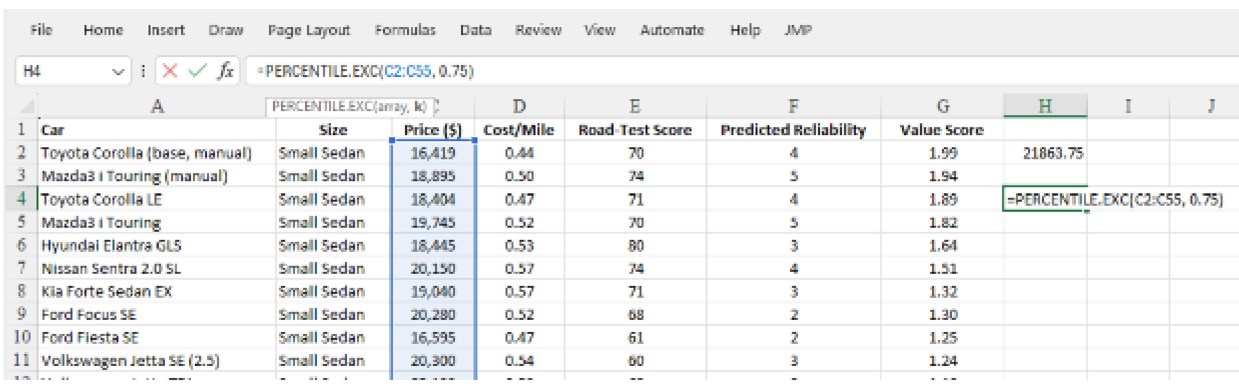
In the Data Analysis Toolpak, there is a Rank and Percentile feature, but this will not give us Q1 and Q3 directly. This feature will give a percent for each value of a given variable, which is not what we want. So this is one of the few times that we will revert to using an Excel function, because it is the easiest way to get Q1 and Q3 with Excel. Keep in mind that this will likely not match what you get in other software or the graphing calculator, and may not match methods of hand calculation you have seen prior to this class. As stated above, we choose to follow how Excel does things. The function we use, PERCENTILE.EXE, matches what is done in the textbook. You can use it to get any percentile, but we focus on Q1 and Q3 here.

Pick a cell where you would like to have the value of Q1 placed, and then start typing
=percentile.exe
You'll then need to highlight the values you are focused on, here all of the values for price (without the first row where the variable name is located). For Q1, we tell Excel we want the 0.25 percentile.



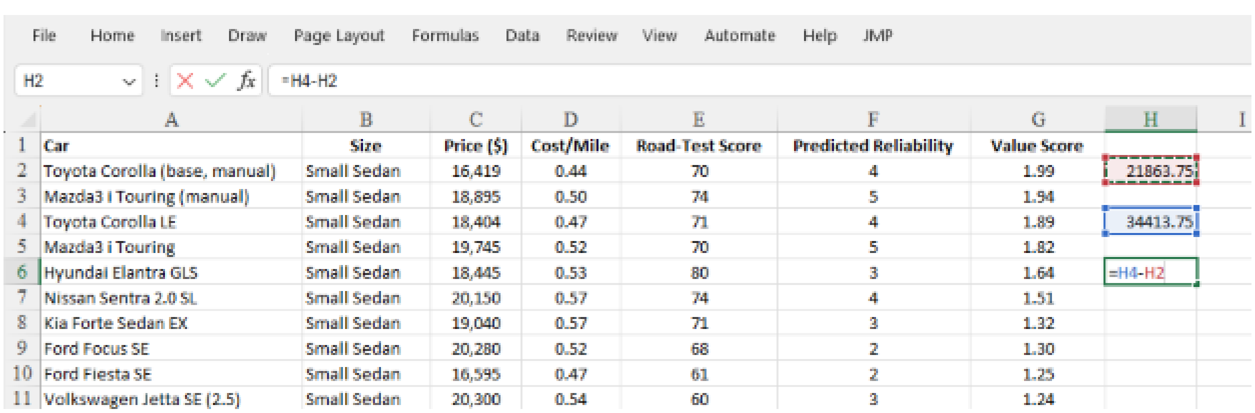
	A	B	C	D	E	F	G	H
1	Car	Size	Price ($)	Cost/Mile	Road-Test Score	Predicted Reliability	Value Score	
2	Toyota Corolla (base, manual)	Small Sedan	16,419	0.44	70	4	1.99	=PERCENTILE.EXC(C2:C55,0.25)
3	Mazda3 i Touring (manual)	Small Sedan	18,895	0.50	74	5	1.94	
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89	
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82	
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64	
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51	
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32	
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30	
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25	

If you hit enter, then you should see the value 21863.75 as shown below, and then Q3 (75th percentile) can be calculated the same way.



	A	B	C	D	E	F	G	H
1	Car	Size	Price ($)	Cost/Mile	Road-Test Score	Predicted Reliability	Value Score	
2	Toyota Corolla (base, manual)	Small Sedan	16,419	0.44	70	4	1.99	21863.75
3	Mazda3 i Touring (manual)	Small Sedan	18,895	0.50	74	5	1.94	
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89	=PERCENTILE.EXC(C2:C55, 0.75)
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82	
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64	
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51	
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32	
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30	
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25	
11	Volkswagen Jetta SE (2.5)	Small Sedan	20,300	0.54	60	3	1.24	

Once you have Q1 and Q3, you can manually calculate IQR = Q3-Q1, or you can tell Excel to do it

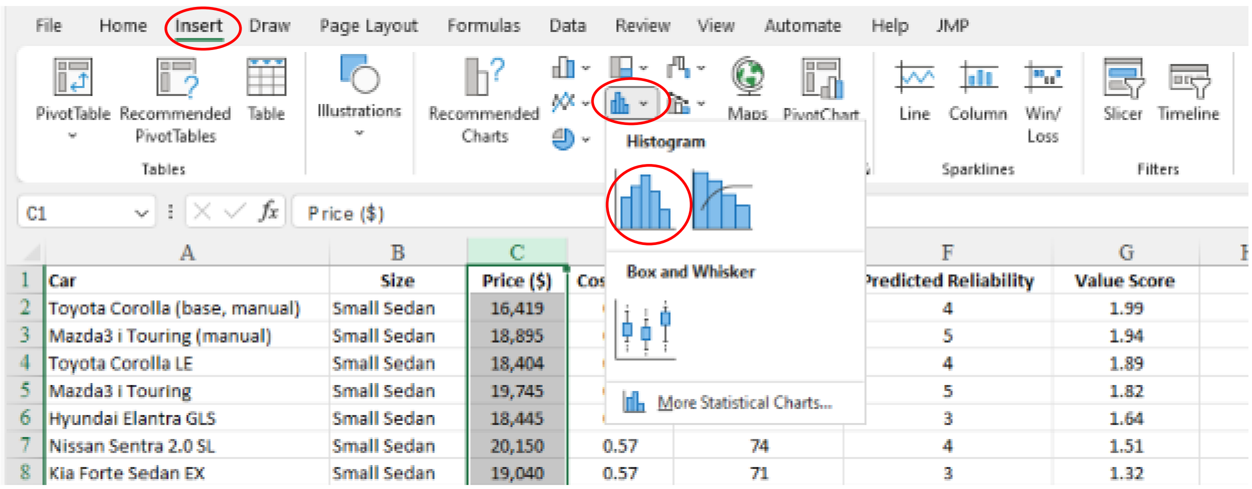


	A	B	C	D	E	F	G	H
1	Car	Size	Price ($)	Cost/Mile	Road-Test Score	Predicted Reliability	Value Score	
2	Toyota Corolla (base, manual)	Small Sedan	16,419	0.44	70	4	1.99	21863.75
3	Mazda3 i Touring (manual)	Small Sedan	18,895	0.50	74	5	1.94	
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89	34413.75
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82	
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64	=H4-H2
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51	
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32	
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30	
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25	
11	Volkswagen Jetta SE (2.5)	Small Sedan	20,300	0.54	60	3	1.24	

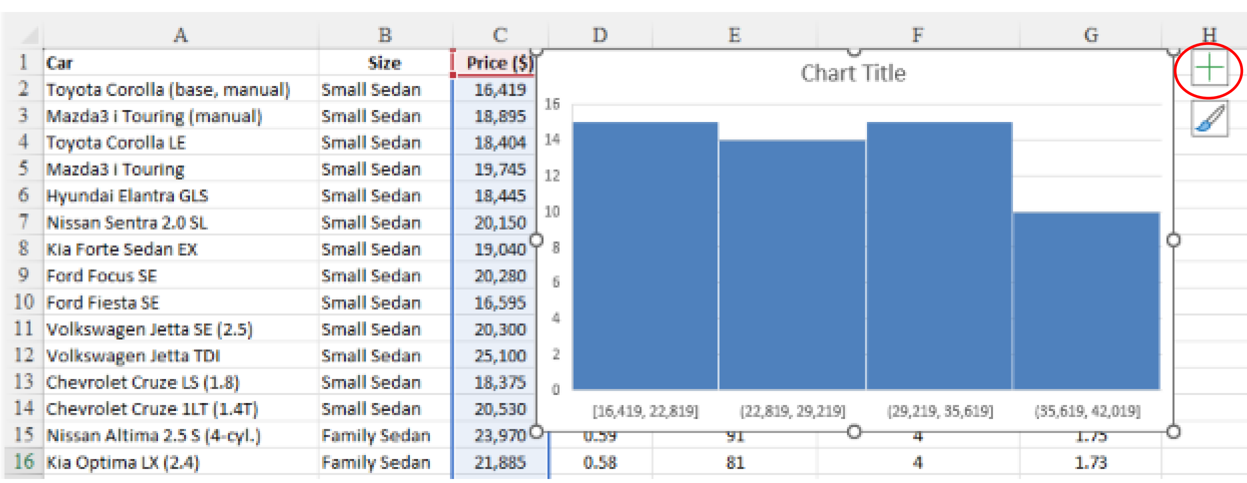
One quantitative response variable

- No explanatory variables
  - **Graphical Summaries**
    - **Boxplots and histograms**

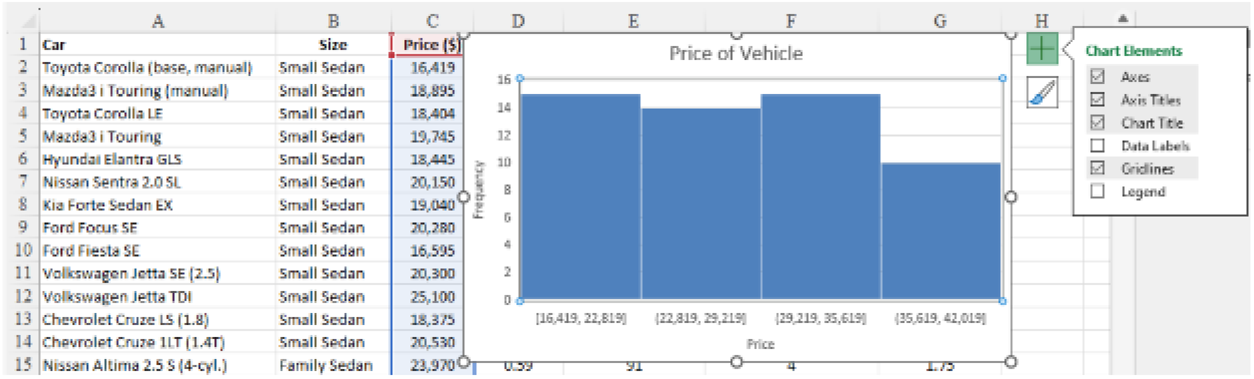
To make a histogram of one quantitative variable (here, we use the variable price from the carvalues data), first <u>highlight the Price column</u> and then click on the Insert tab, and then click on the Histogram button, and then click on the icon for the traditional histogram. Note that we are not yet interested in subsetting the data by a categorical variable like Size, so we can select the entire Price column. However, in the future, we will need to select only those cells that are relevant. The point is that you will not always be able to select an entire column.



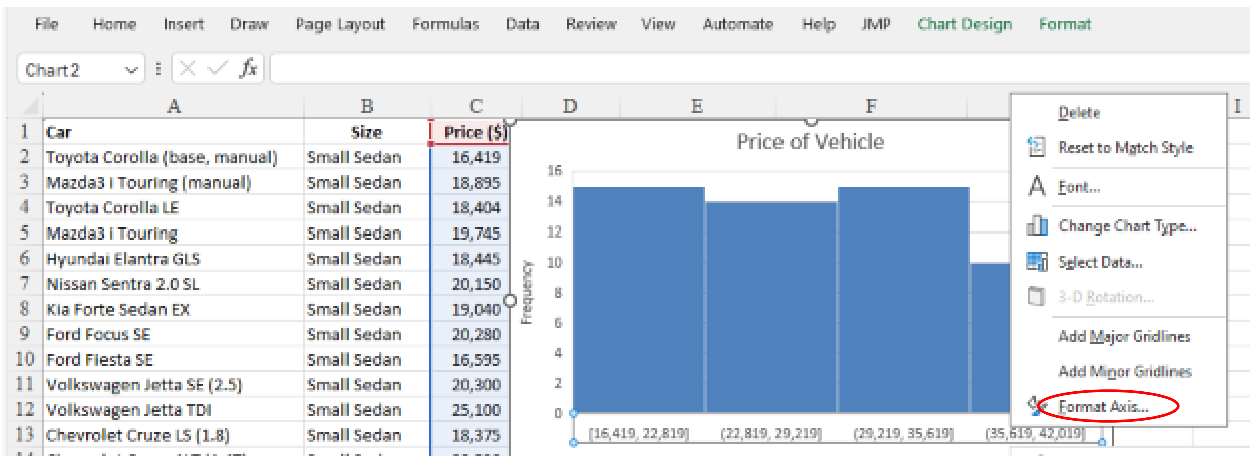
This should insert a histogram as follows, though we may wish to alter it. For example, , we may wish to change the title or axis labels. To do so, click on the + icon at the top right



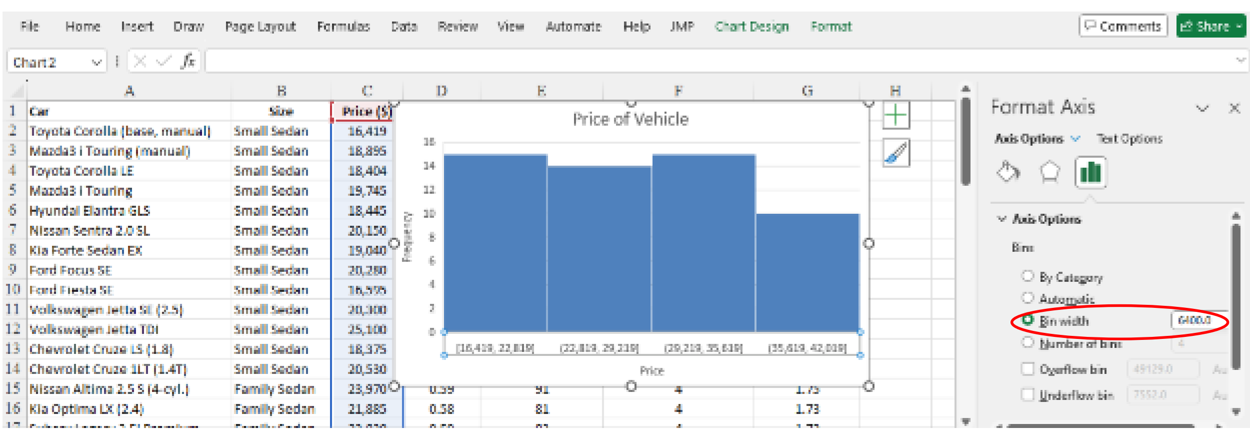
And then click on various parts of the graph to make changes like this (you can literally type the Title and axis labels inside the graph by clicking on those aspects inside the graph).



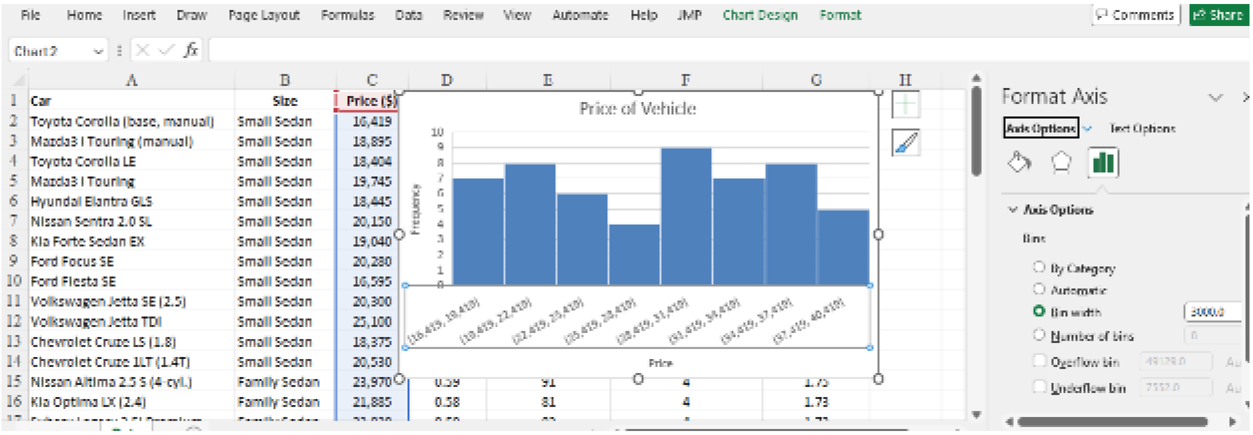
To change the width of the bins, you can right click on one of the intervals that are labeling the horizontal axis, and then select Format Axis



which will allow you to select Bin width and change the Bin width to whatever makes the most sense. This will change the number of bars in the histogram.

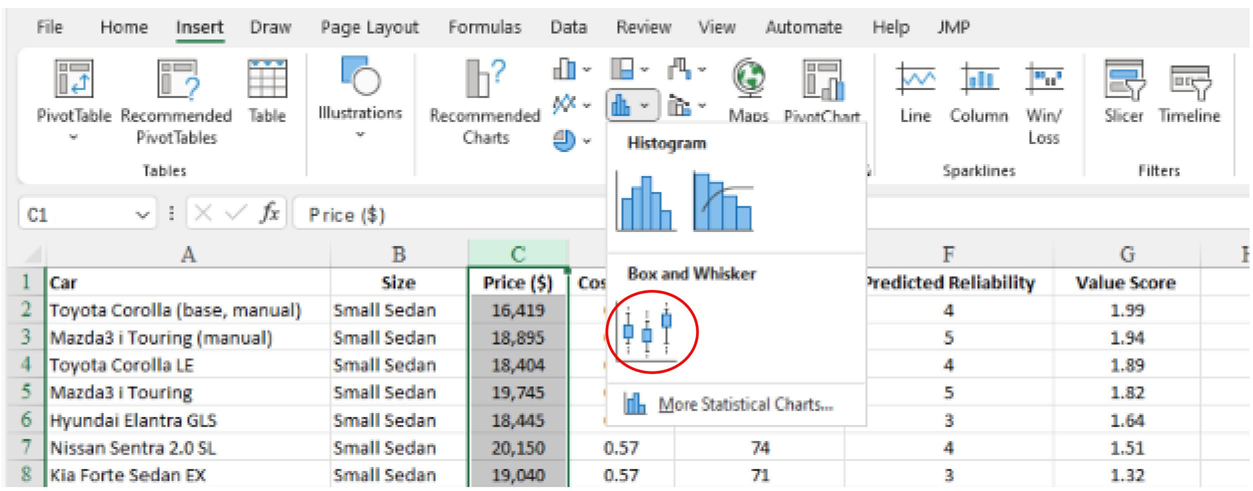


I changed the Bin width to 3000 and hit enter, which gave this histogram. You can also choose to alter the Number of bins of you prefer.

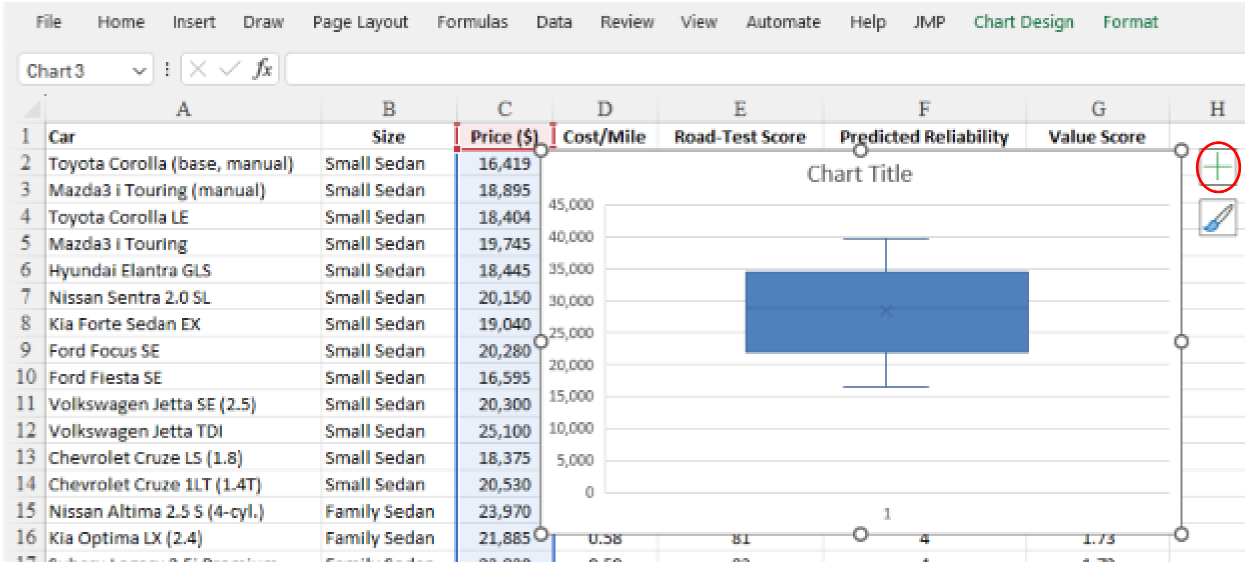


If these are all the changes your instructor would like you to make, you can copy this histogram and paste it into another document for submission.

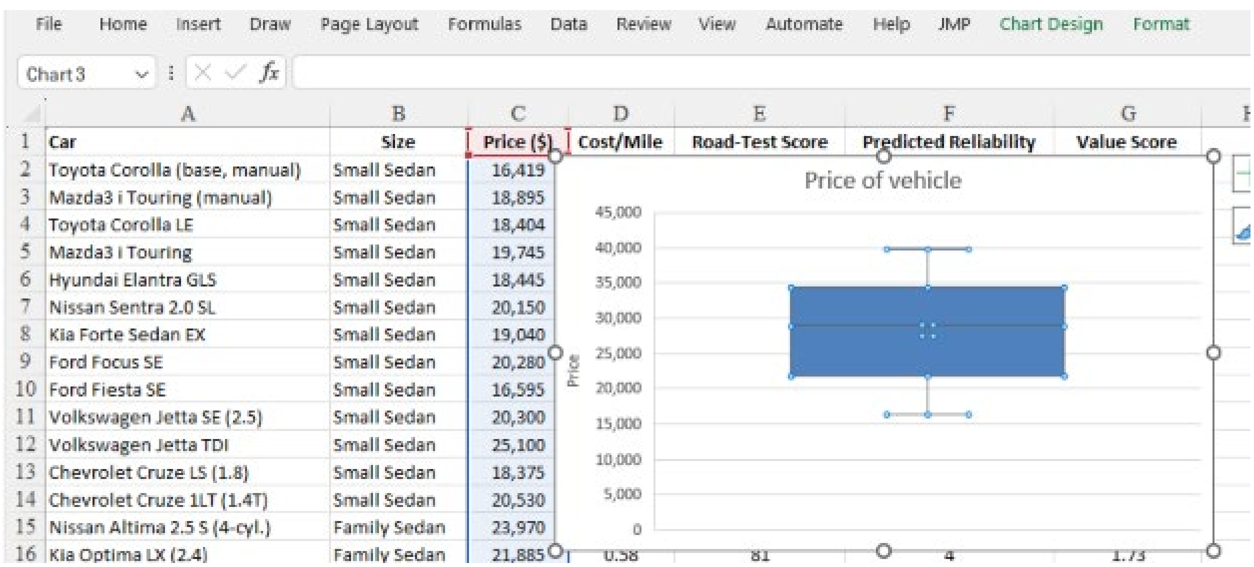
To make a boxplot of one quantitative variable, you can use the same steps as for the Histogram, but instead choose Box and Whisker. Boxplots produced this way will show outliers when they are present. Don't forget to first highlight the data that you are interested in (which may not be an entire column in general, but it is in this example):



Which should produce the basic boxplot as follows. You can make updates to the Title and axis labels as described for the Histogram above.



Here I changed the Title, axis label on the vertical axis, and deleted the horizontal axis label and the value 1 on the horizontal axis, because we only have one group in this example. We'll handle side-by-side boxplots in the coming pages.

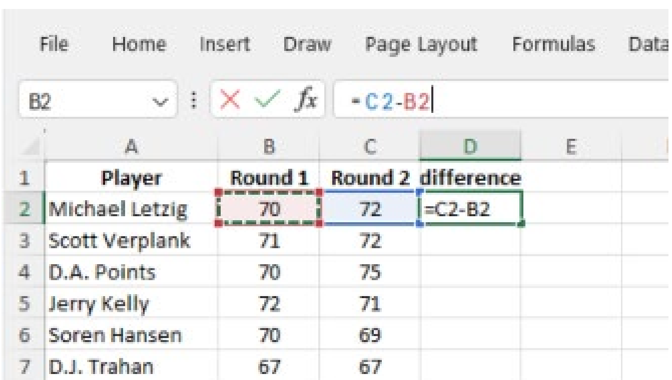


One quantitative response variable

- **No explanatory variables**
  - **Inferential statistics**
    - **Paired data**
      - **Hypothesis testing (One sample t test)**
      - **Confidence intervals (One sample t interval)**

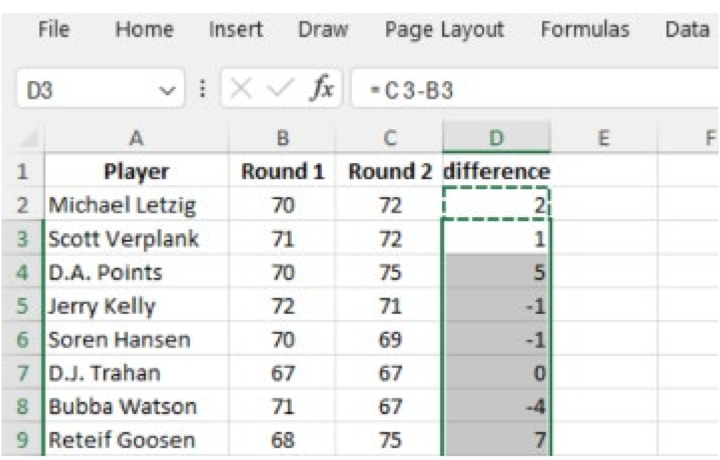
It is important to note that the feature of the Data Analysis Toolpak that can be used for the calculations for paired data t test and confidence interval do not check the assumption of the paired t test or confidence interval. In order to do that, we first must create the paired differences by creating a new variable, and check to see if the paired differences are reasonably normally distributed (unless the sample size is sufficiently large) by making appropriate graphs like histograms and boxplots.

Here we use the *golfscores* data set, because these are truly paired, meaning the two scores in each row come from the same golfer on the same course, so it makes sense to subtract the scores, to get a difference for each golfer. In fact, we'll need to do that to check to see if the differences are reasonably normal. First create a new column with the differences by doing the following. You'll need to type in a very simple formula and then hit enter. Note that I did Round 2 minus Round 1, because Round 2 came later in time.



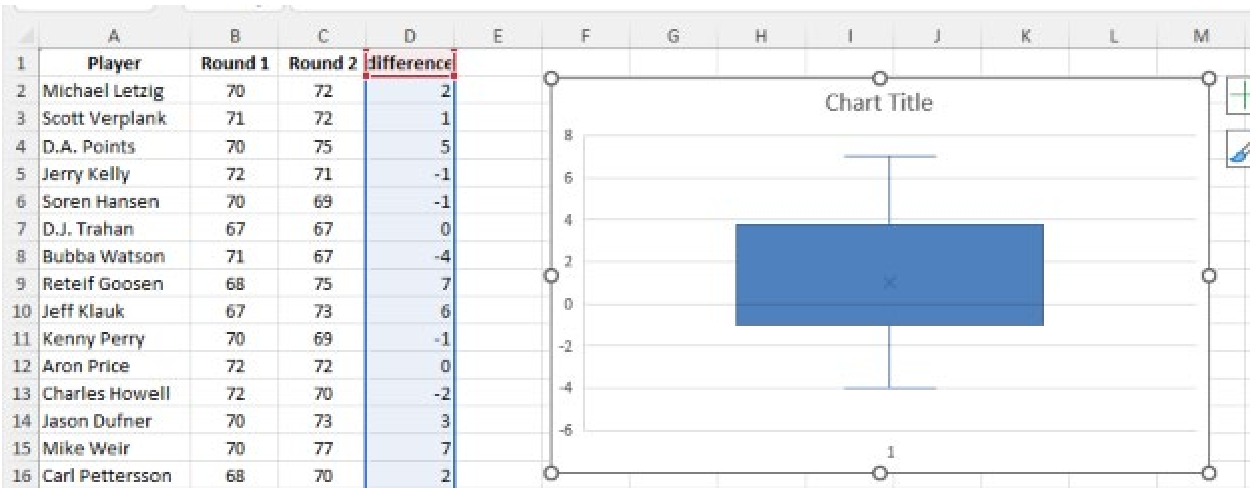
	A	B	C	D	E
1	Player	Round 1	Round 2	difference	
2	Michael Letzig	70	72	=C2-B2	
3	Scott Verplank	71	72		
4	D.A. Points	70	75		
5	Jerry Kelly	72	71		
6	Soren Hansen	70	69		
7	D.J. Trahan	67	67		

Then you can copy the formula from cell D2 to the rest of the cells below D2.



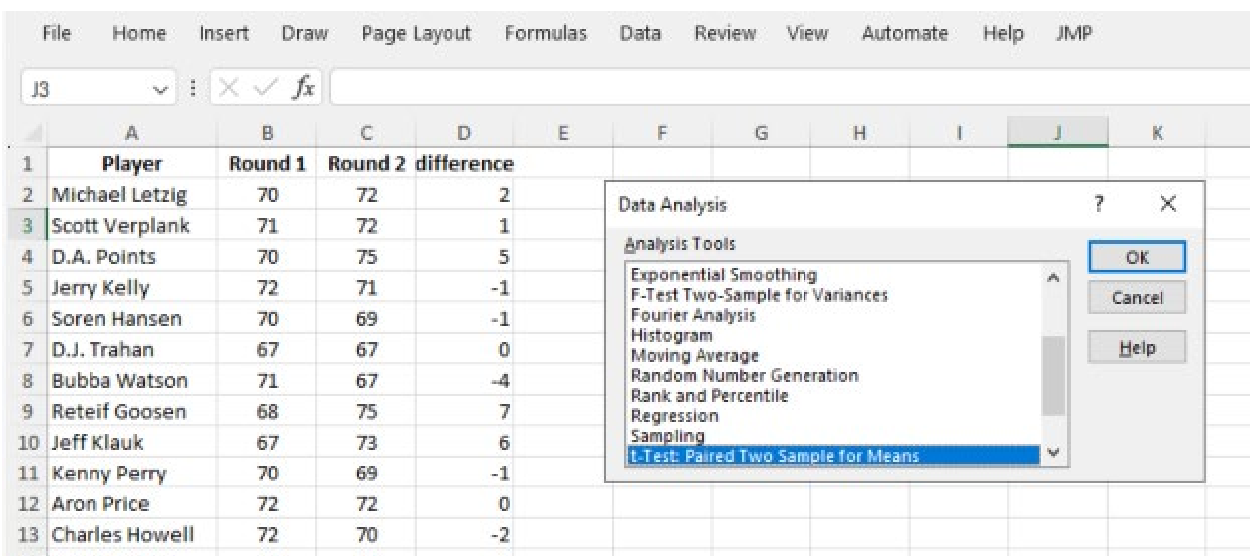
	A	B	C	D	E	F
1	Player	Round 1	Round 2	difference		
2	Michael Letzig	70	72	2		
3	Scott Verplank	71	72	1		
4	D.A. Points	70	75	5		
5	Jerry Kelly	72	71	-1		
6	Soren Hansen	70	69	-1		
7	D.J. Trahan	67	67	0		
8	Bubba Watson	71	67	-4		
9	Reteif Goosen	68	75	7		

It is a good idea to manually check a few values in the new column to make sure Excel did what you wanted it to do. If this checks out, then you can make a boxplot or histogram to check normality as described above, including altering title and axis labels (unless the sample size is sufficiently large).



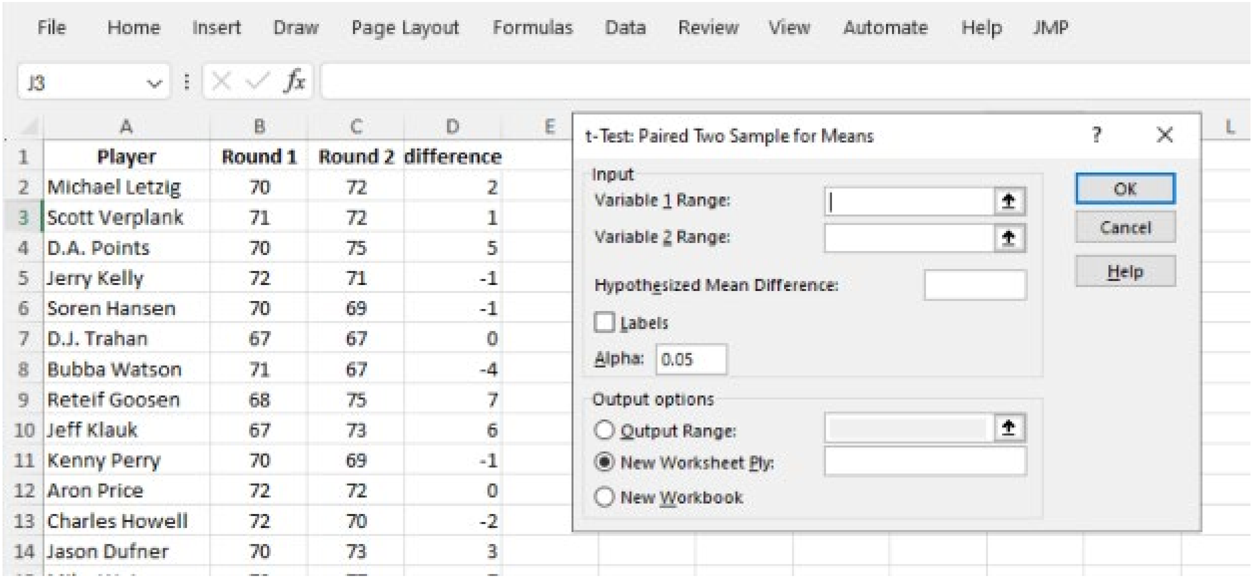
	A	B	C	D
1	Player	Round 1	Round 2	difference
2	Michael Letzig	70	72	2
3	Scott Verplank	71	72	1
4	D.A. Points	70	75	5
5	Jerry Kelly	72	71	-1
6	Soren Hansen	70	69	-1
7	D.J. Trahan	67	67	0
8	Bubba Watson	71	67	-4
9	Reteif Goosen	68	75	7
10	Jeff Klauk	67	73	6
11	Kenny Perry	70	69	-1
12	Aron Price	72	72	0
13	Charles Howell	72	70	-2
14	Jason Dufner	70	73	3
15	Mike Weir	70	77	7
16	Carl Pettersson	68	70	2

To get Excel to calculate paired t test statistic, p-value and confidence interval, you need to use the Data Analysis Toolpak add-in, but you won't use the difference column that you just created. As described above, click on the Data tab, and then the Data Analysis icon on the far right at the top. Select the t-test: Paired Two Sample for Means option as shown below, and click OK

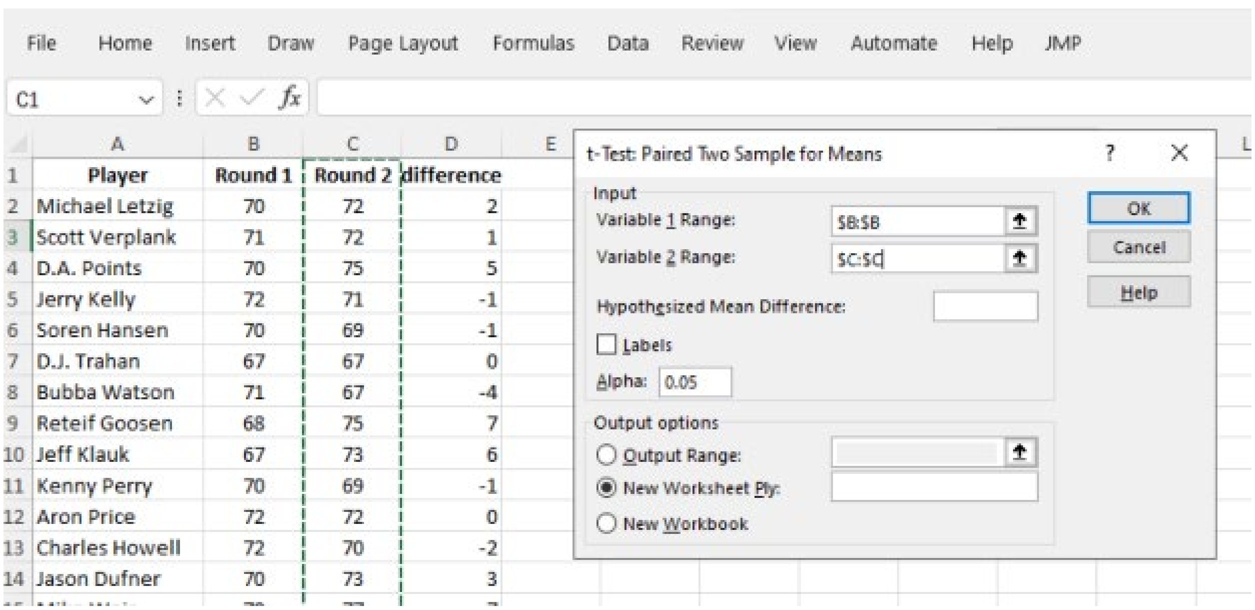


	A	B	C	D
1	Player	Round 1	Round 2	difference
2	Michael Letzig	70	72	2
3	Scott Verplank	71	72	1
4	D.A. Points	70	75	5
5	Jerry Kelly	72	71	-1
6	Soren Hansen	70	69	-1
7	D.J. Trahan	67	67	0
8	Bubba Watson	71	67	-4
9	Reteif Goosen	68	75	7
10	Jeff Klauk	67	73	6
11	Kenny Perry	70	69	-1
12	Aron Price	72	72	0
13	Charles Howell	72	70	-2

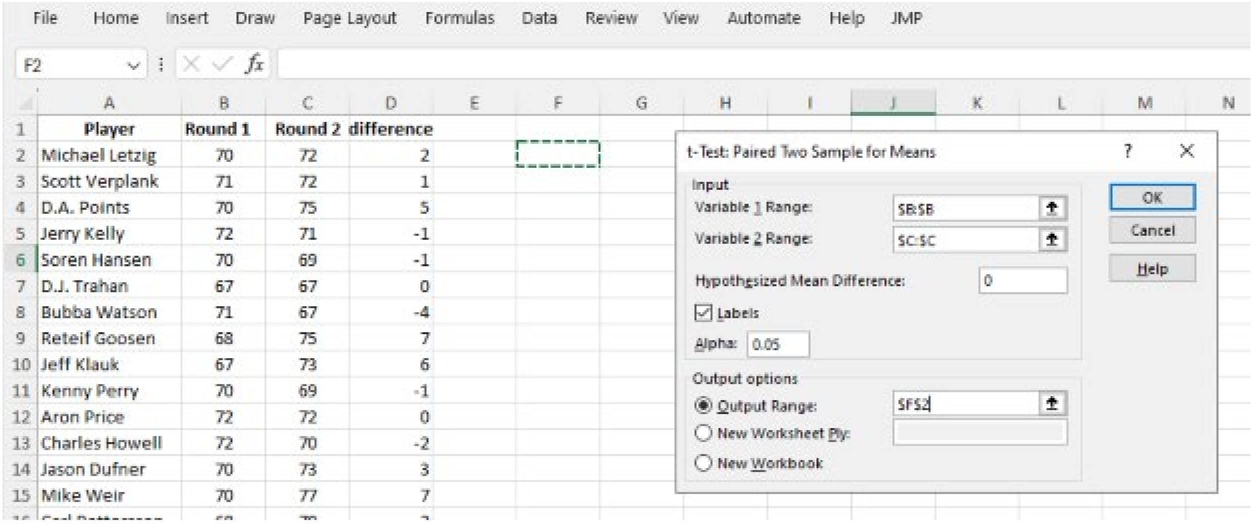
That will open this dialog



Single click in the Variable 1 Range, and then highlight the column for Round 1, and then click in Variable 2 Range and highlight the column for Round 2. We can take the whole column for each of these in this example, but in general that may not be OK. Only highlight the data you want Excel to use.

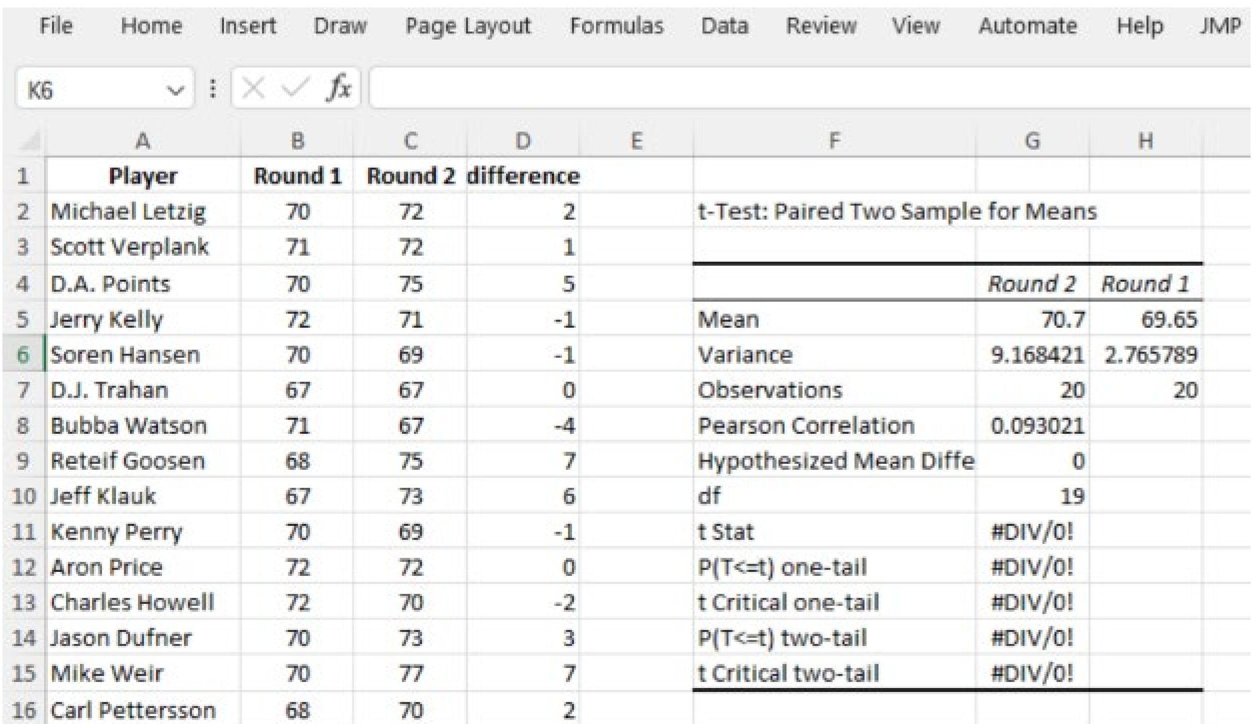


Put the null hypothesis value in the box for Hypothesized Mean Difference, which is typically 0. Check the Labels box because the first row of data has the variable names. Put in appropriate Alpha value if not 0.05. If you want the output to show up in the same spreadsheet, you can click on Output Range, click in the box to the right of Output Range, and then select a cell in the spreadsheet where you'd like the output to show up. Then click OK.



	Player	Round 1	Round 2	difference
1	Player	Round 1	Round 2	difference
2	Michael Letzig	70	72	2
3	Scott Verplank	71	72	1
4	D.A. Points	70	75	5
5	Jerry Kelly	72	71	-1
6	Soren Hansen	70	69	-1
7	D.J. Trahan	67	67	0
8	Bubba Watson	71	67	-4
9	Reteif Goosen	68	75	7
10	Jeff Klauk	67	73	6
11	Kenny Perry	70	69	-1
12	Aron Price	72	72	0
13	Charles Howell	72	70	-2
14	Jason Dufner	70	73	3
15	Mike Weir	70	77	7

This is what I get when I do this (putting Round 2 in Variable 1 Range and Round 1 in Variable 2 Range to be consistent with how the differences were formed by using Round 2 minus Round 1.

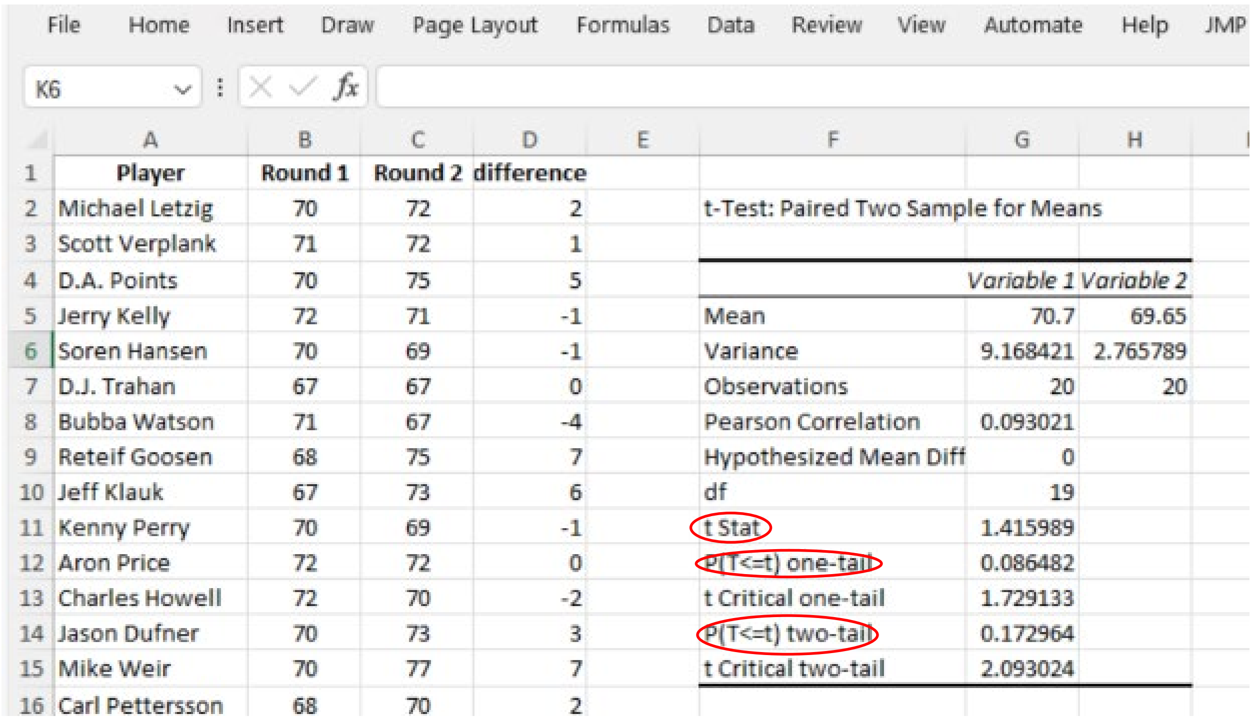


	A	B	C	D	E	F	G	H
1	Player	Round 1	Round 2	difference				
2	Michael Letzig	70	72	2		t-Test: Paired Two Sample for Means		
3	Scott Verplank	71	72	1				
4	D.A. Points	70	75	5			*Round 2*	*Round 1*
5	Jerry Kelly	72	71	-1		Mean	70.7	69.65
6	Soren Hansen	70	69	-1		Variance	9.168421	2.765789
7	D.J. Trahan	67	67	0		Observations	20	20
8	Bubba Watson	71	67	-4		Pearson Correlation	0.093021	
9	Reteif Goosen	68	75	7		Hypothesized Mean Diffe	0	
10	Jeff Klauk	67	73	6		df	19	
11	Kenny Perry	70	69	-1		t Stat	#DIV/0!	
12	Aron Price	72	72	0		P(T<=t) one-tail	#DIV/0!	
13	Charles Howell	72	70	-2		t Critical one-tail	#DIV/0!	
14	Jason Dufner	70	73	3		P(T<=t) two-tail	#DIV/0!	
15	Mike Weir	70	77	7		t Critical two-tail	#DIV/0!	
16	Carl Pettersson	68	70	2				

Note that there are some strange values in the output, specifically #DIV/0!. Excel says we asked it to divide by 0, but the real issue is the with the variable name in the first row. We can get around this (because I don't know how to fix the variable name) by just highlighting the data instead of the entire column, as you can see here, but you should uncheck Labels, because we're not including the variable name in the first row. To be consistent with how I created the difference column, I put Round 2 in Variable 1 Range and Round 1 in Variable 2 Range (you can tell by which has higher mean and the sign on the test statistic). This order will matter if you are intending to do a one tail test (to make sure you are consistent with how you set up the alternative hypothesis), but it will not matter for a two tail test.

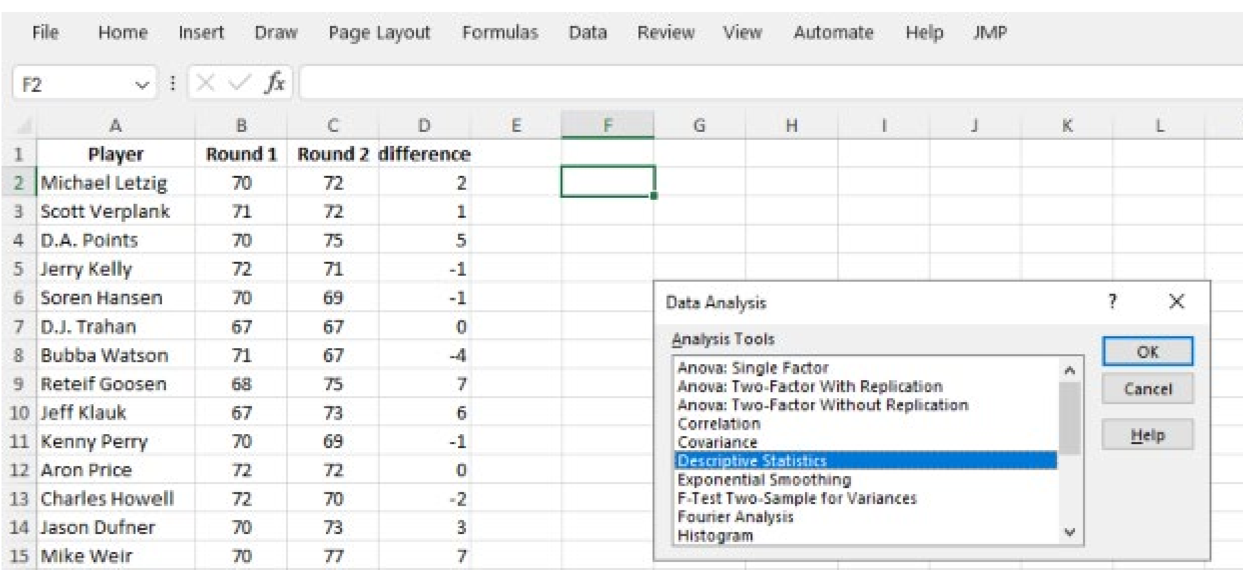
Now you can extract the paired t test statistic and p-value, being careful to note which p-value to use, one tail or two tail.

Note that it is very easy to get a paired t test wrong when you are doing a one tail test. Make sure that when you set up the hypotheses, you take into account both (1) how you subtracted to create the differences and (2) what claim you are supposed to test from the wording of the problem. This confusion can easily translate into reading Excel output as well.



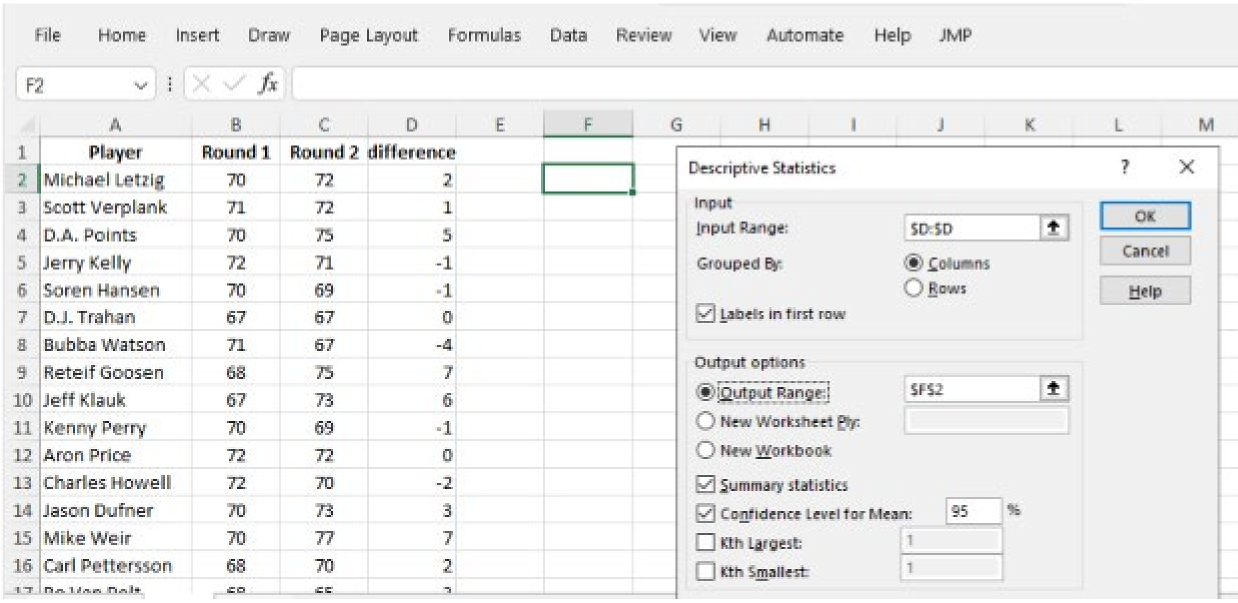
	A	B	C	D	E	F	G	H
1	Player	Round 1	Round 2	difference				
2	Michael Letzig	70	72	2		t-Test: Paired Two Sample for Means		
3	Scott Verplank	71	72	1				
4	D.A. Points	70	75	5			*Variable 1*	*Variable 2*
5	Jerry Kelly	72	71	-1		Mean	70.7	69.65
6	Soren Hansen	70	69	-1		Variance	9.168421	2.765789
7	D.J. Trahan	67	67	0		Observations	20	20
8	Bubba Watson	71	67	-4		Pearson Correlation	0.093021	
9	Reteif Goosen	68	75	7		Hypothesized Mean Diff	0	
10	Jeff Klauk	67	73	6		df	19	
11	Kenny Perry	70	69	-1		t Stat	1.415989	
12	Aron Price	72	72	0		P(T<=t) one-tail	0.086482	
13	Charles Howell	72	70	-2		t Critical one-tail	1.729133	
14	Jason Dufner	70	73	3		P(T<=t) two-tail	0.172964	
15	Mike Weir	70	77	7		t Critical two-tail	2.093024	
16	Carl Pettersson	68	70	2				

The Paired Two Sample for Means option in the Data Analysis Toolpak does not give a confidence interval for the population mean difference. However, we can use the Descriptive Statistics option shown below, applied the difference column.



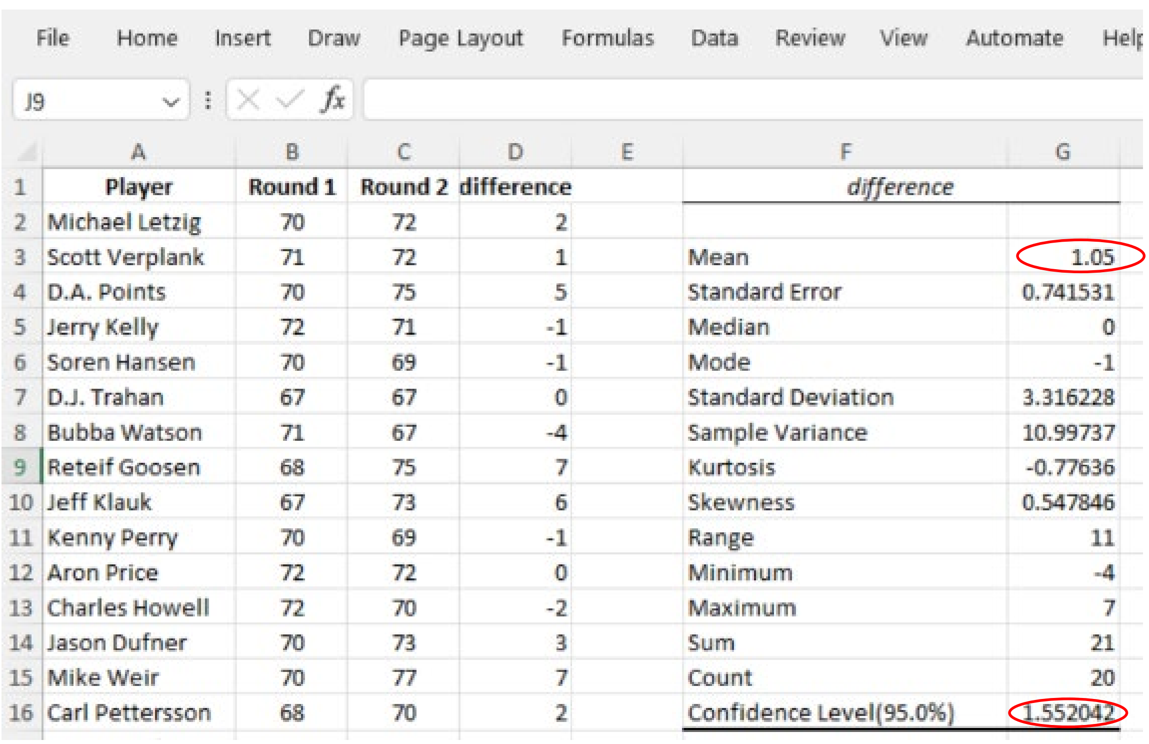
	A	B	C	D
1	Player	Round 1	Round 2	difference
2	Michael Letzig	70	72	2
3	Scott Verplank	71	72	1
4	D.A. Points	70	75	5
5	Jerry Kelly	72	71	-1
6	Soren Hansen	70	69	-1
7	D.J. Trahan	67	67	0
8	Bubba Watson	71	67	-4
9	Reteif Goosen	68	75	7
10	Jeff Klauk	67	73	6
11	Kenny Perry	70	69	-1
12	Aron Price	72	72	0
13	Charles Howell	72	70	-2
14	Jason Dufner	70	73	3
15	Mike Weir	70	77	7

To do this, you can highlight the entire column for difference (including the first row), but be sure check the box for Labels in first row. You'll also want to check the box for Confidence Level for Mean, you also need to check the box for Summary Statistics (because you'll need the value of the sample mean, as you'll see).



	Player	Round 1	Round 2	difference
2	Michael Letzig	70	72	2
3	Scott Verplank	71	72	1
4	D.A. Points	70	75	5
5	Jerry Kelly	72	71	-1
6	Soren Hansen	70	69	-1
7	D.J. Trahan	67	67	0
8	Bubba Watson	71	67	-4
9	Reteif Goosen	68	75	7
10	Jeff Klauk	67	73	6
11	Kenny Perry	70	69	-1
12	Aron Price	72	72	0
13	Charles Howell	72	70	-2
14	Jason Dufner	70	73	3
15	Mike Weir	70	77	7
16	Carl Pettersson	68	70	2

Excel gives the following output



	Player	Round 1	Round 2	difference		difference	
2	Michael Letzig	70	72	2			
3	Scott Verplank	71	72	1		Mean	1.05
4	D.A. Points	70	75	5		Standard Error	0.741531
5	Jerry Kelly	72	71	-1		Median	0
6	Soren Hansen	70	69	-1		Mode	-1
7	D.J. Trahan	67	67	0		Standard Deviation	3.316228
8	Bubba Watson	71	67	-4		Sample Variance	10.99737
9	Reteif Goosen	68	75	7		Kurtosis	-0.77636
10	Jeff Klauk	67	73	6		Skewness	0.547846
11	Kenny Perry	70	69	-1		Range	11
12	Aron Price	72	72	0		Minimum	-4
13	Charles Howell	72	70	-2		Maximum	7
14	Jason Dufner	70	73	3		Sum	21
15	Mike Weir	70	77	7		Count	20
16	Carl Pettersson	68	70	2		Confidence Level(95.0%)	1.552042

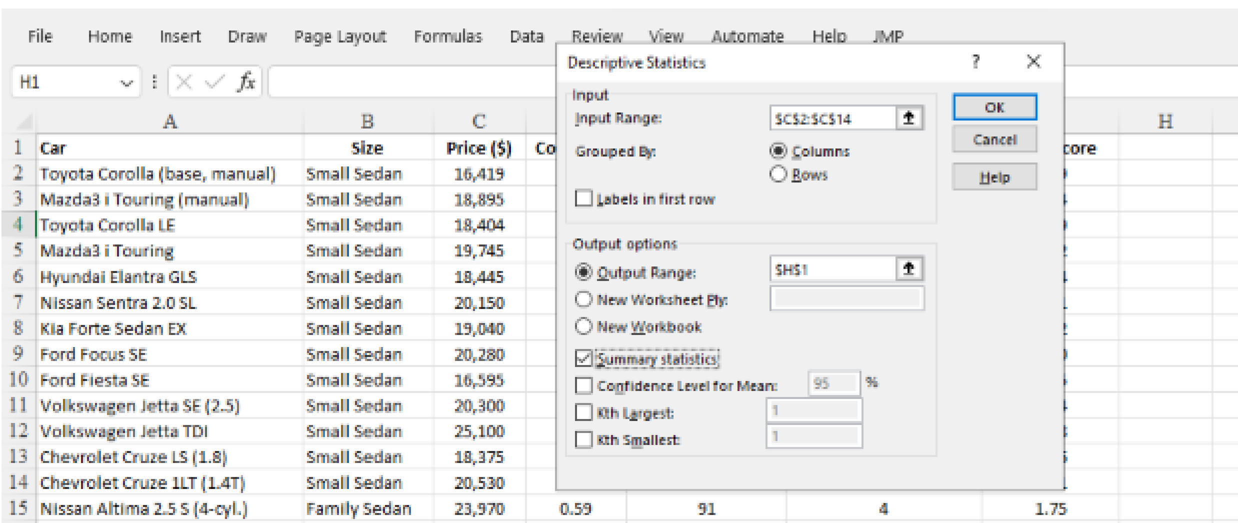
Note that Excel does not give the confidence interval directly. In the bottom row of the output, we see the value of 1.552042. This is technically the "margin of error", which by formula for one quantitative variable is $t^* s/\sqrt{n}$, or in other words, everything after the +/- in the confidence interval formula. So to get the lower and upper bounds on the confidence interval, you have to calculate mean – 1.552042 and mean + 1.552042, where mean = 1.05 from the output for this example. So the 95% confidence interval for the population mean difference is (-0.50, 2.60), rounded to two decimal places. Because we're dealing with paired data here, take care to remember which way you subtracted in creating the differences (in this example, Round 2 minus Round 1) before interpreting the confidence interval.

One quantitative response variable

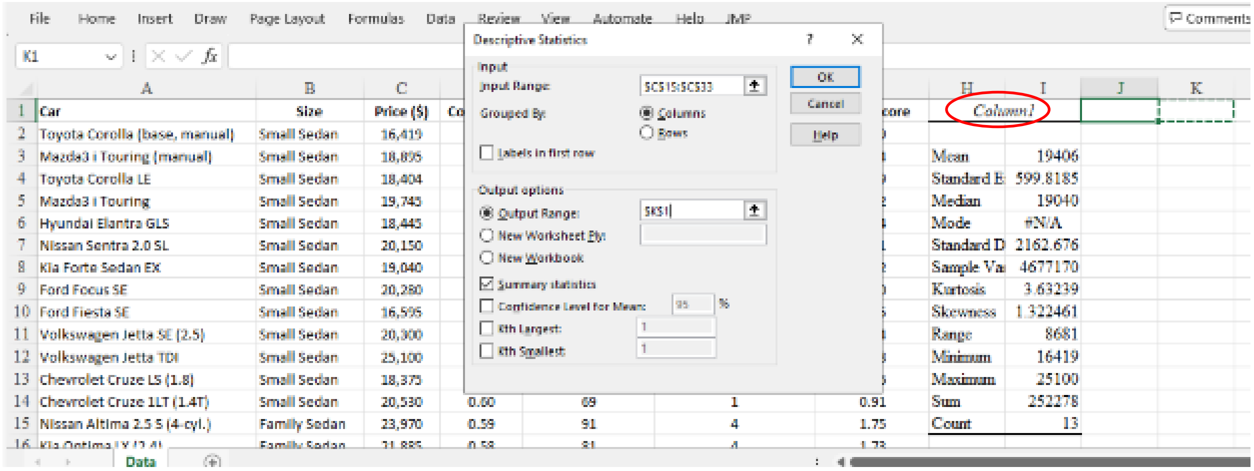
- **One categorical explanatory variable**
  - **Descriptive statistics and graphical summaries**
    - **Comparing components of a quantitative distribution (center, variability, shape, outliers) across groups**

This kind of analysis is where the issue of tidy vs untidy is relevant because there will be a categorical variable that determines which group each subject belongs to. Regardless, you'll still have to highlight the correct portion of data for the analysis, so focus on that.

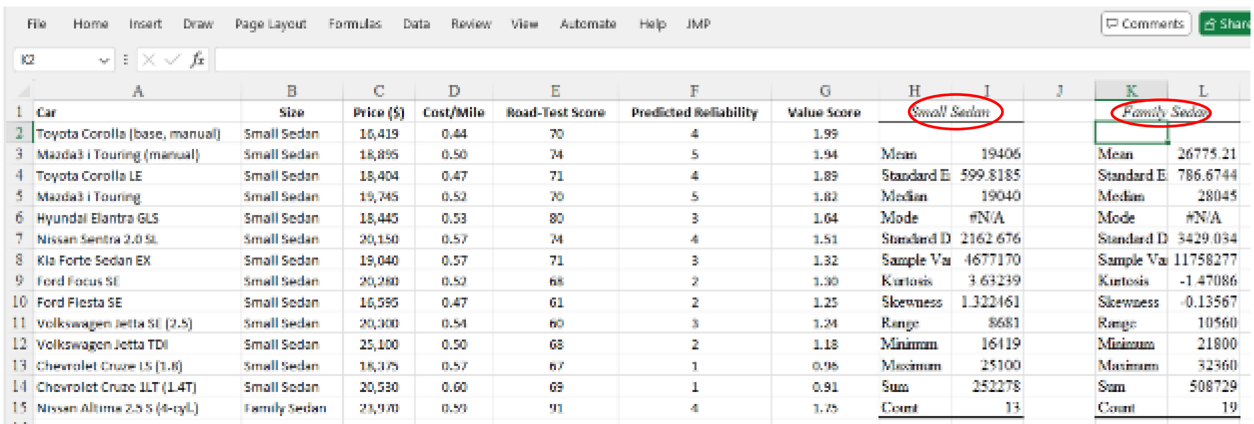
When summarizing a quantitative response variable by different groups, you just need to use the Descriptive Statistics feature of the Data Analysis Toolpak multiple times, once for each group. For example, to summarize Price for just Small Sedans, you can highlight only the values of Price that correspond to Small Sedans. You may have to first sort the data by the categorical variable (here Size). Note the range of values in Input Range are just C2 to C14.



Then you can repeat for Family Sedan. Note that Input Range now has C15 to C33, which corresponds to Family Sedan.



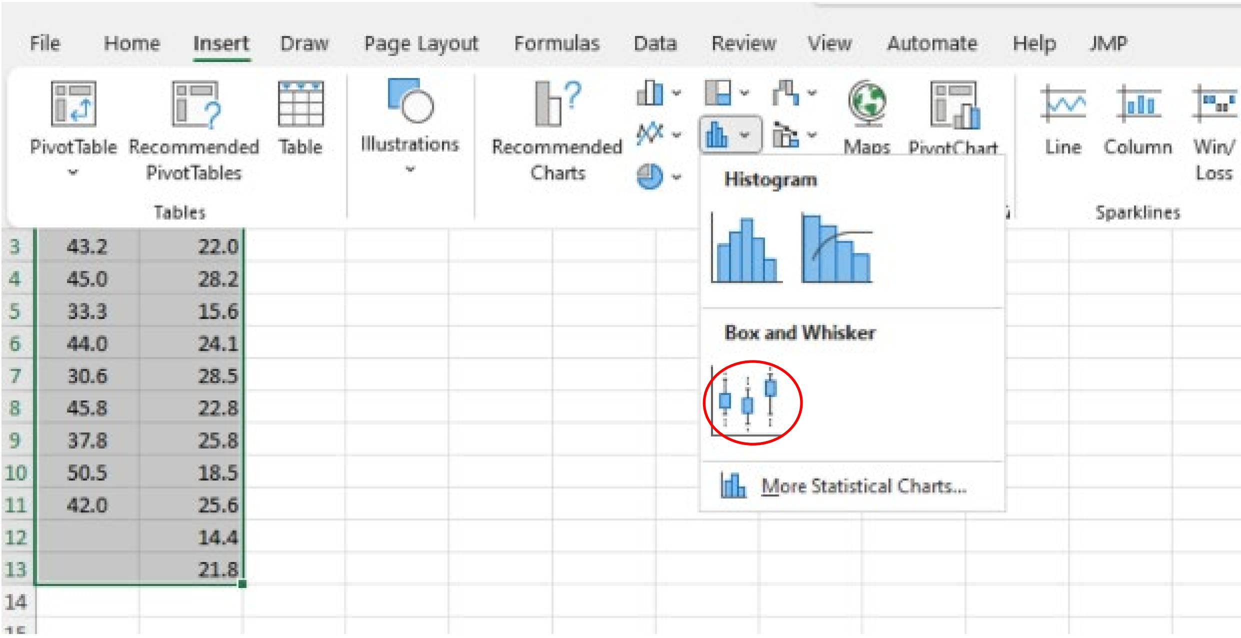
You may wish to type over "Column 1" to minimize confusion, which I replaced here with Small Sedan and Family Sedan in the first row, above the corresponding output.



Getting percentiles works the same as above, just make sure to only give Excel the range of values for a given category.

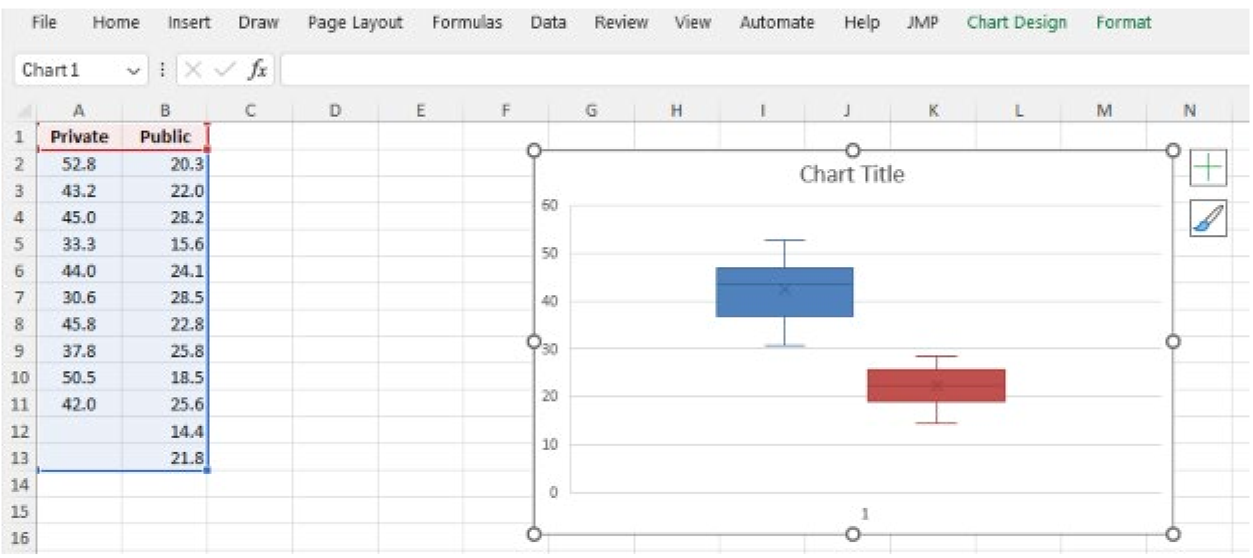
Making side-by-side boxplots is actually easier if data are in untidy form, though we show how to get side-by-side boxplots for both untidy and tidy forms. First we show untidy, so we switch to using the collegecosts data set.

First highlight the data for both Private and Public (remember, these data are untidy because the two values in each row come from unrelated universities), then go to the Insert tab, and pick boxplot as show earlier



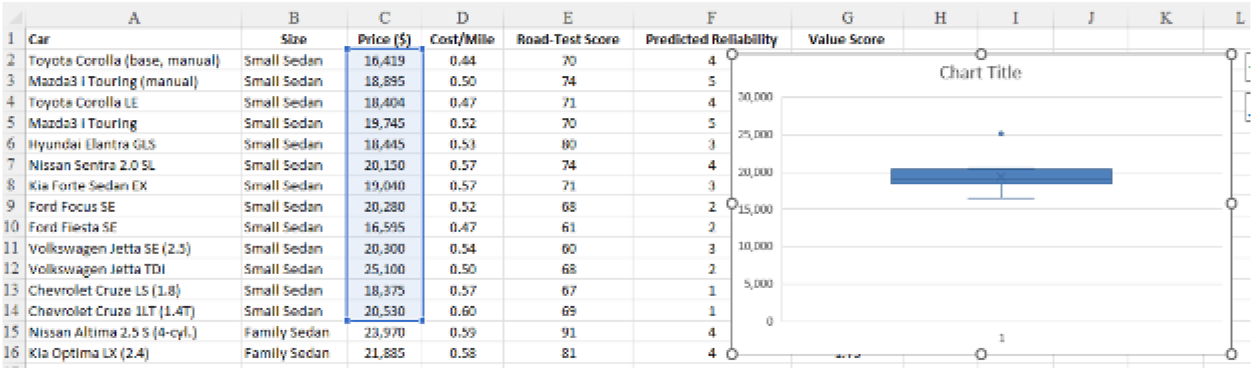
3	43.2	22.0
4	45.0	28.2
5	33.3	15.6
6	44.0	24.1
7	30.6	28.5
8	45.8	22.8
9	37.8	25.8
10	50.5	18.5
11	42.0	25.6
12		14.4
13		21.8

You should see the following output, which you can adapt as shown earlier.

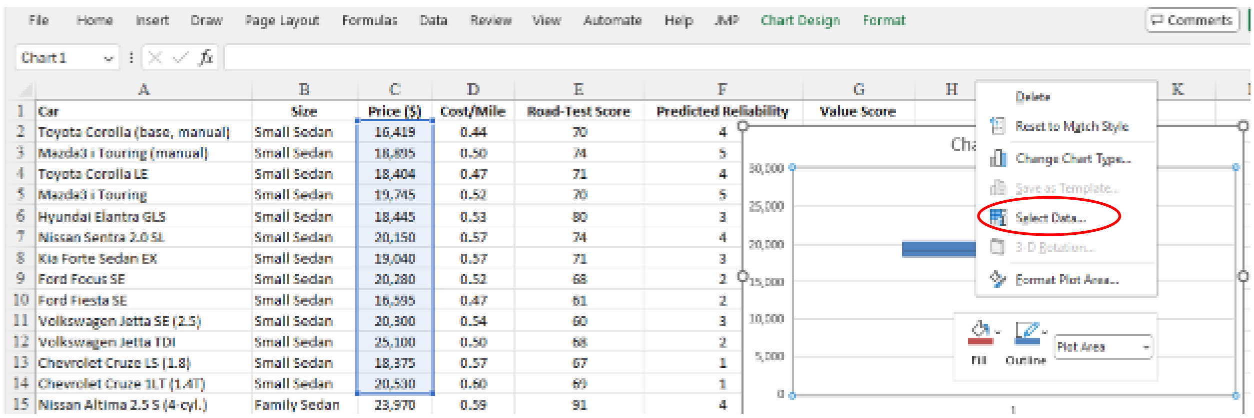


	A	B
1	Private	Public
2	52.8	20.3
3	43.2	22.0
4	45.0	28.2
5	33.3	15.6
6	44.0	24.1
7	30.6	28.5
8	45.8	22.8
9	37.8	25.8
10	50.5	18.5
11	42.0	25.6
12		14.4
13		21.8

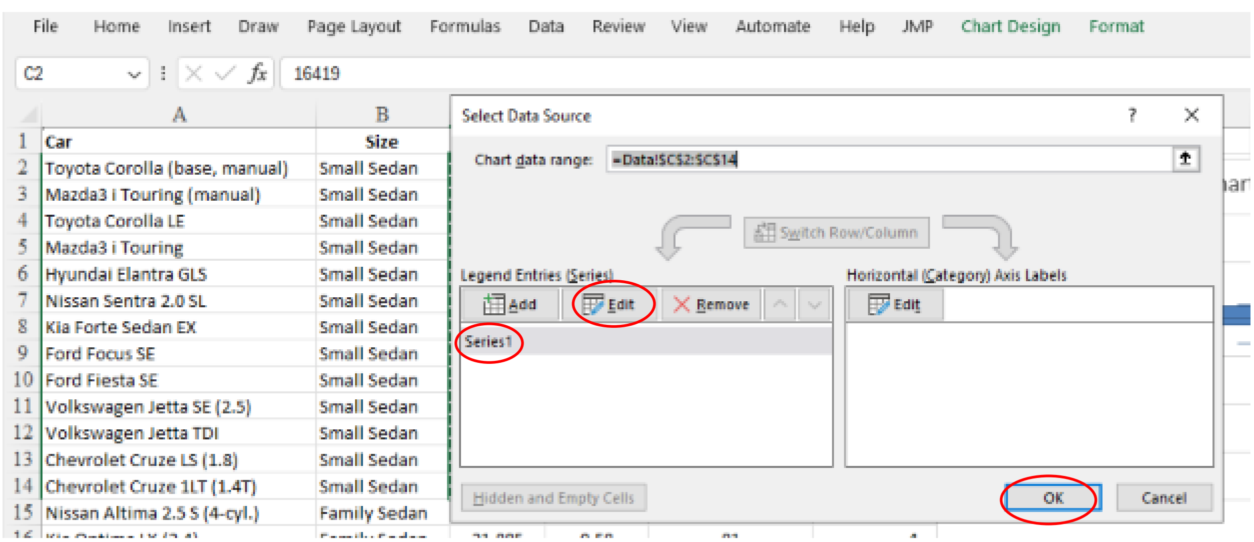
For tidy data, we use the carvalues data set (this is tidy because all values and categories in one row all come from the same car). Making side-by-side boxplots is a little more complicated than for untidy data. First make a boxplot for one group, and then you'll have to insert a second one for the other group. We'll compare prices of Small Sedans vs Family Sedans. If the data were not already sorted by the categorical variable Size, you would first want to do that.



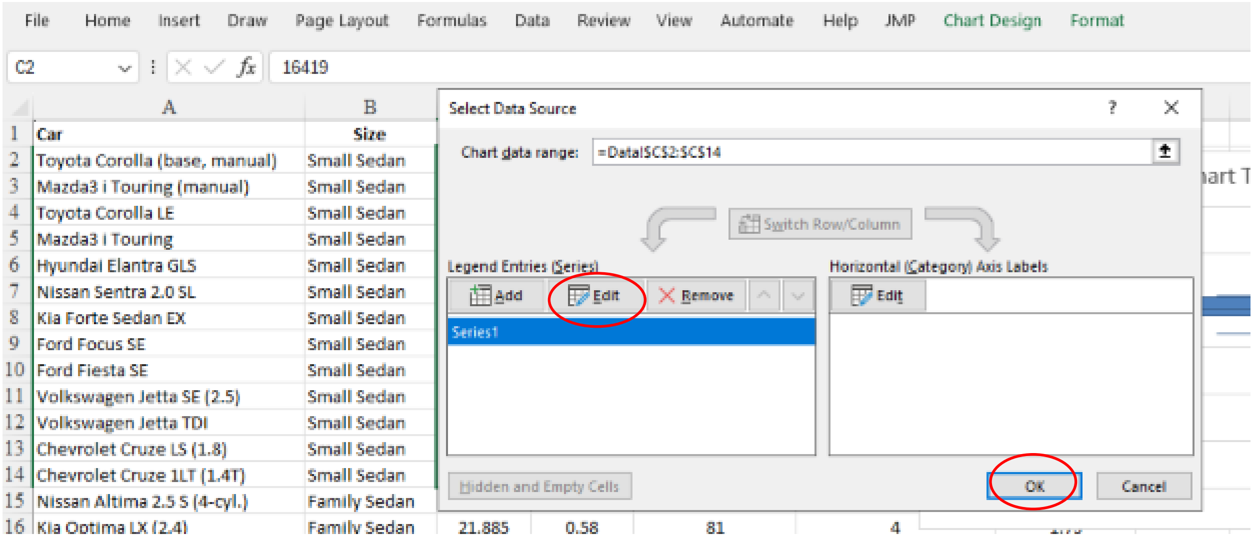
Then we want to add in a boxplot for the Family Sedans. To do that, right click in the graph so that you see the Select Data option



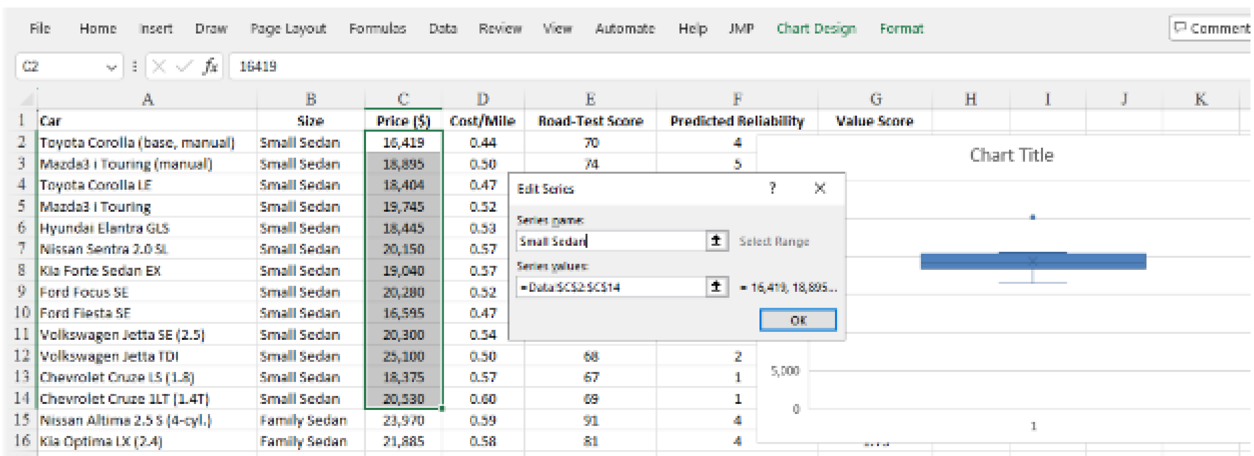
If you click on Select Data, you should see this window



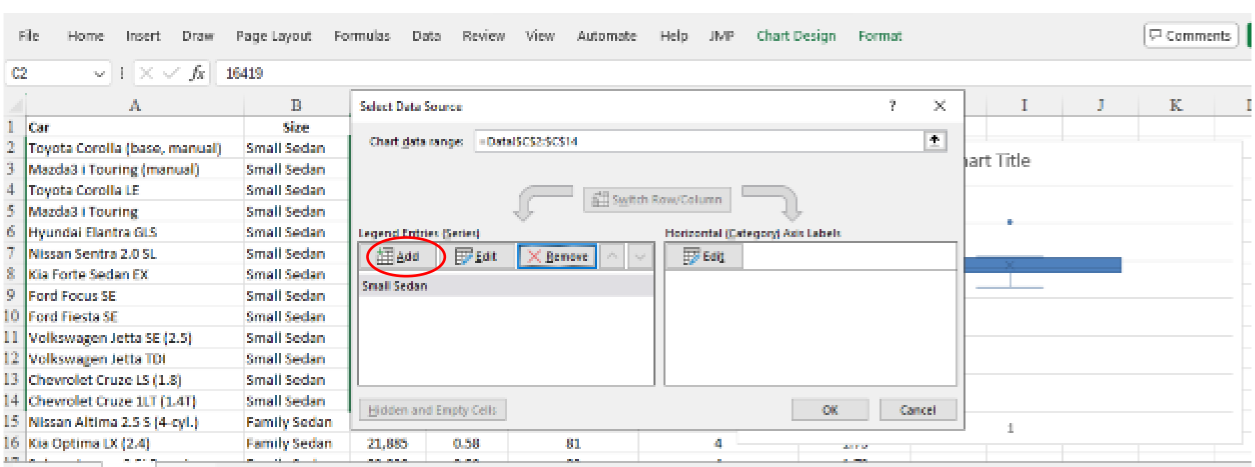
Series1 is the Small Sedan data we already plotted. We can rename that from Series1 to Small Sedan by clicking on Series1 and then Edit and then click OK as shown below



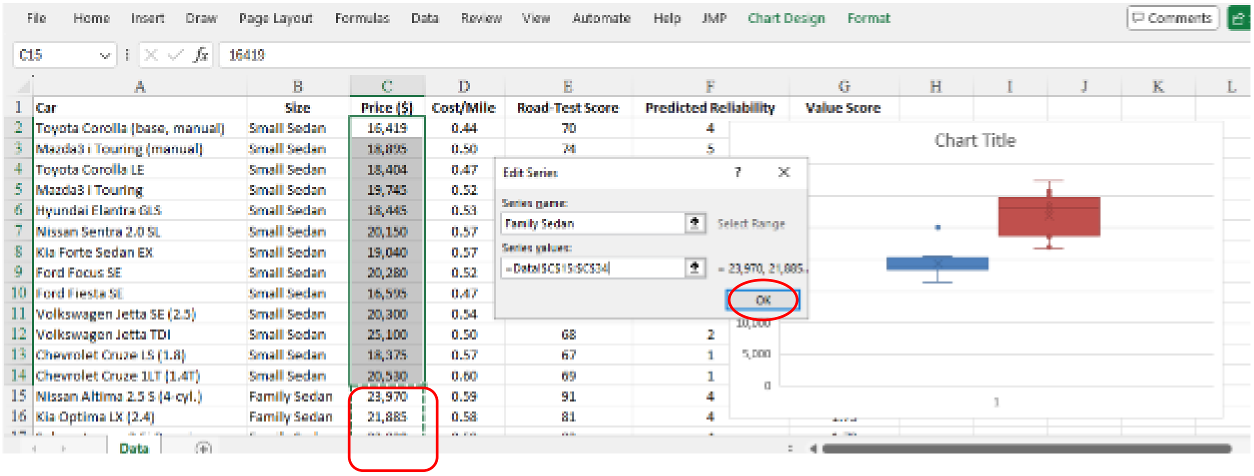
Which should show you this (after you type Small Sedan into Series name).



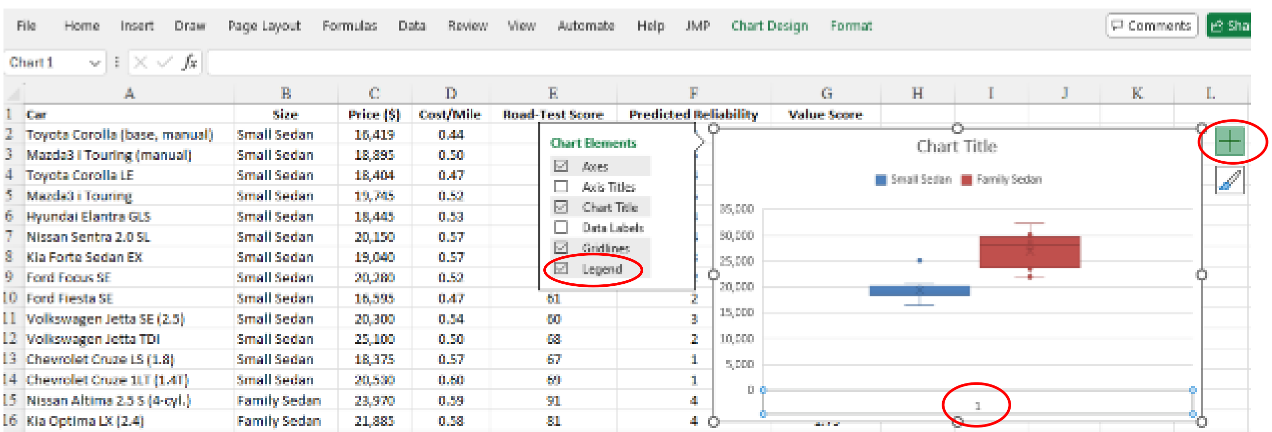
Then click OK, which should have Small Sedan in place of Series1. Now we want to add a new series for Family Sedans, so click on Add



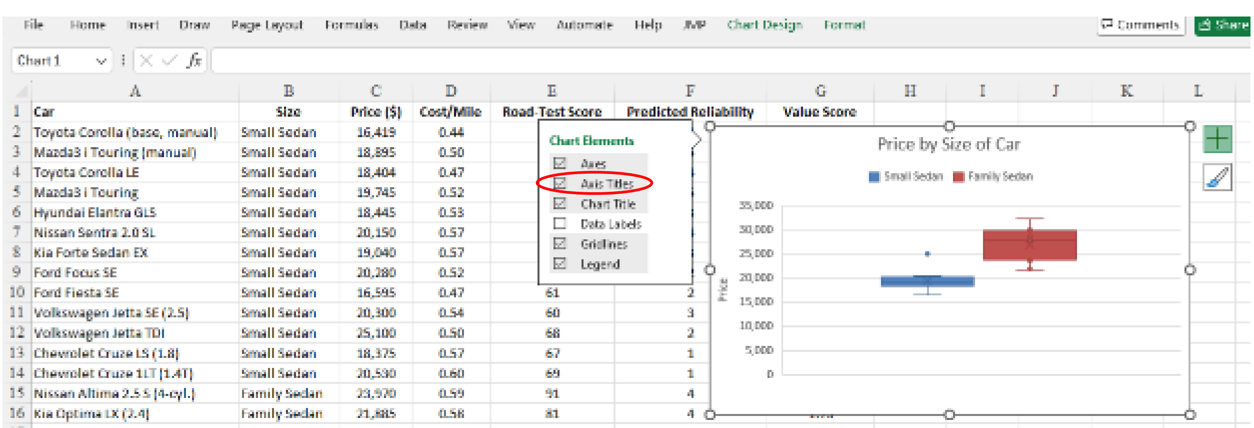
Then type Family Sedan into Series name, and then single click in Series values, and highlight the price values that correspond to the Family Sedan category (shown in red in the screen shot below), which should look like this. Then click OK, and then click OK one more time, and you have side-by-side boxplots, though we'll want to add appropriate labels, which we do next.



Now if you click on the graph, and click on the + in the upper right, you'll be able to check the box for Legend



Which will label which boxplot is which. You will probably want to adapt the Chart title and add axis labels, but that works the same as shown above. Click on Axis Titles to add those, and just click on Chart Title at the top to alter that. You can delete the number 1 below the chart just by clicking on it and then hitting delete.

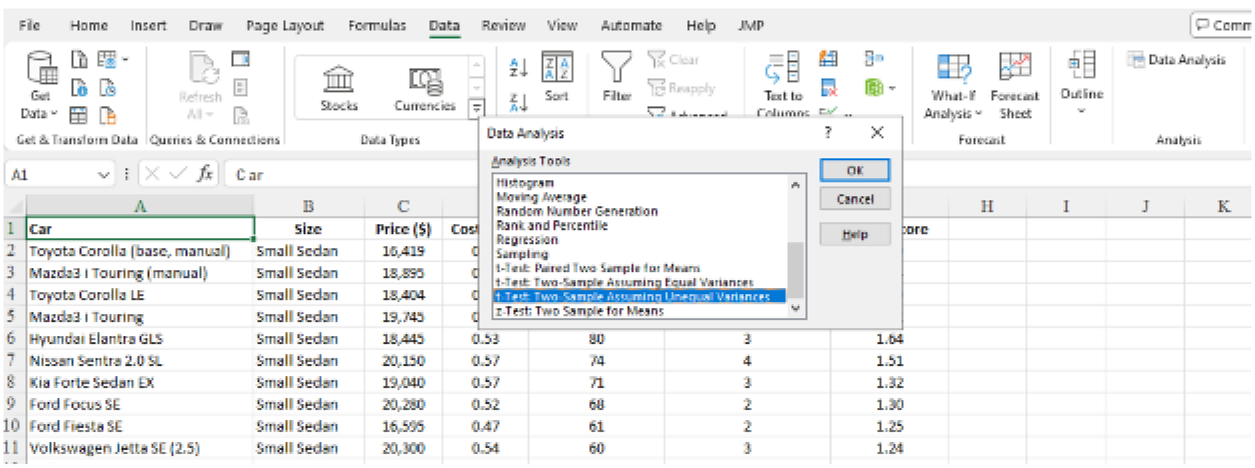


One quantitative response variable

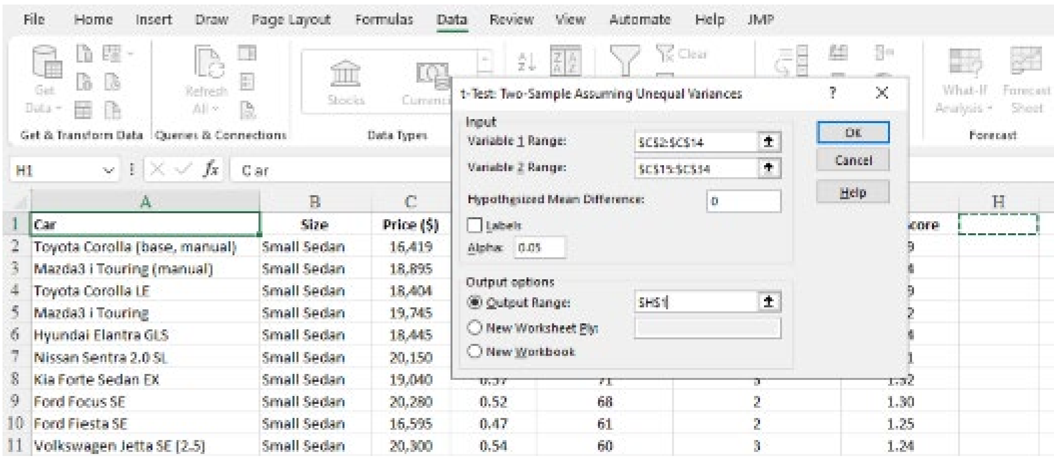
- **One categorical explanatory variable**
    - **Inferential statistics**
        - **Hypothesis testing (Two independent sample t test)**
        - **Confidence intervals (Two independent sample t interval for difference in means)**

In order to implement the two independent sample t test, you should first check conditions of the test. That may (depending on sample size) involve checking for normality of each group, which can be done using histograms or boxplots as show above. To get the test statistic and p-value from Excel, we can use the Data Analysis Toolpak as shown next. We'll use the carvalues data set which is in tidy form, but you can easily do this for untidy data as well, just by highlighting the appropriate values (so having tidy data for this purpose is not a big deal).

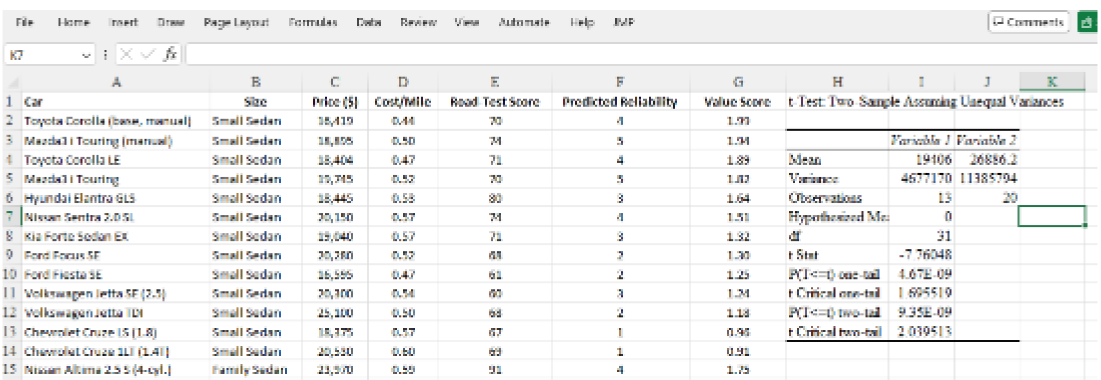
Suppose we want to test for the equality of population means of Price for small sedans vs family sedans (from the Size variable). We will always focus on using the unpooled t test (and confidence interval) in STA 225, so you should find the option show below that says "Assuming Unequal Variances".



Then you need to click in the box next to Variable 1 Range, and then highlight the Price values (without the word Price in the C1 cell) for small sedan. Then click in the box next to Variable 2 Range and highlight the Price values for family sedan. Because the null hypothesis for this test will always be that the population means are equal in STA 225, we will always put 0 in Hypothesized Mean Difference. Do NOT click on the box for Labels (because we did not highlight the C1 cell with the variable name Price in it). Then you can click on Output Range and tell Excel where you want to see the output, here I chose cell H1.



When you click on OK, you should get the following output where you told Excel to put the output. The value “t Stat” of -7.76048 is the test statistic, and then you have to choose the p-value that is consistent with your alternative hypothesis. Excel calls the p-value “P(T<=t)”, and then you decide if you need one tail or two tail. You should ignore the “t Critical” values for the purposes of the hypothesis test. Note that the degrees of freedom have been rounded to the nearest integer (the actual formula does not produce an integer).

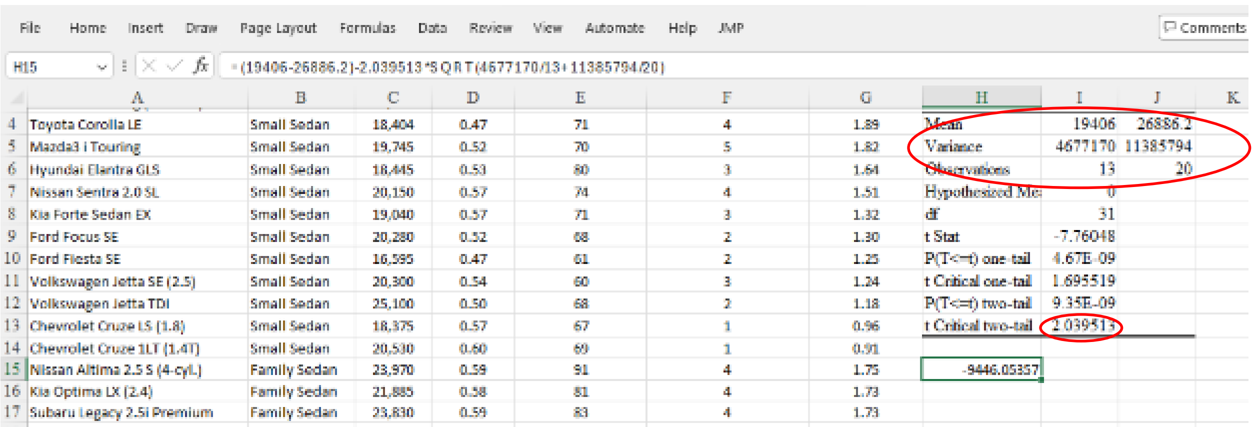


Unfortunately, the two independent sample t confidence interval is not built into Excel, even in the Data Analysis Toolpak. That means we need to work a little harder using the formula, but we can minimize our work by harnessing the output from the two independent sample t test. Recall from the textbook that the formula for the two independent sample t confidence interval is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\left\{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right\}}$$

If we want to apply this to the carvalues data set, where we want to compare population mean prices of small sedans to family sedans (as we did above with the corresponding hypothesis test), we have everything we need from the output from the hypothesis test. Specifically, we have the values of the sample means ($\bar{x}_1$ and $\bar{x}_2$) and sample variances (s_1^2 and s_2^2), and also the sample sizes (n_1 and n_2). Very important: the test statistic you see above is not the same as a critical value. In other words, t and t^* are not the same thing. The critical value t^* is also in the output called "t Critical two-tail", which is different than the test statistic which you already know is called "t Stat".

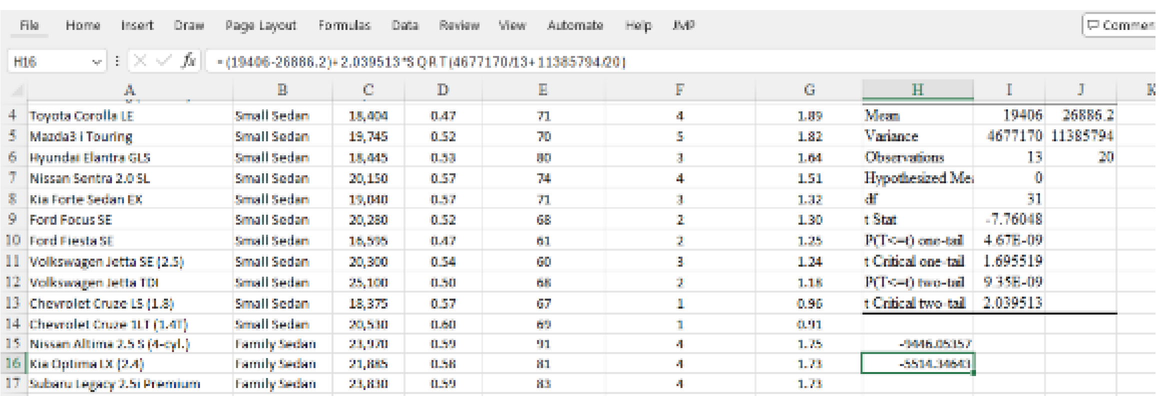
So we can put all of this together into one cell, if we are careful about order of operations, and remember that standard deviation is the square root of variance, so be careful to distinguish between sample variance s_1^2 and sample standard deviation s_1 (without the square). Excel gives us variance, which is squared standard deviation, so we don't have to square when using the formula.



H15: =(19406-26886.2)-2.039513*SQRT(4677170/13+11385794/20)

	A	B	C	D	E	F	G	H	I	J
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89	Mean	19406	26886.2
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82	Variance	4677170	11385794
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64	Observations	13	20
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51	Hypothesized Me:	0	
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32	df	31	
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30	t Stat	-7.76048	
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25	P(T<=t) one-tail	4.67E-09	
11	Volkswagen Jetta SE (2.5)	Small Sedan	20,300	0.54	60	3	1.24	t Critical one-tail	1.695519	
12	Volkswagen Jetta TDI	Small Sedan	25,100	0.50	68	2	1.18	P(T<=t) two-tail	9.35E-09	
13	Chevrolet Cruze LS (1.8)	Small Sedan	18,375	0.57	67	1	0.96	t Critical two-tail	2.039513	
14	Chevrolet Cruze 1LT (1.4T)	Small Sedan	20,530	0.60	69	1	0.91			
15	Nissan Altima 2.5 S (4-cyl.)	Family Sedan	23,970	0.59	91	4	1.75	-9446.05357		
16	Kia Optima LX (2.4)	Family Sedan	21,885	0.58	81	4	1.73			
17	Subaru Legacy 2.5i Premium	Family Sedan	23,830	0.59	83	4	1.73			

Note that in cell H15, I am only calculating the left side of the interval. We'll have to repeat in cell H16 for the right side of the interval, switching – to + in the formula: $(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\left\{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right\}}$.



H16: =(19406-26886.2)+2.039513*SQRT(4677170/13+11385794/20)

	A	B	C	D	E	F	G	H	I	J
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89	Mean	19406	26886.2
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82	Variance	4677170	11385794
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64	Observations	13	20
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51	Hypothesized Me:	0	
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32	df	31	
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30	t Stat	-7.76048	
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25	P(T<=t) one-tail	4.67E-09	
11	Volkswagen Jetta SE (2.5)	Small Sedan	20,300	0.54	60	3	1.24	t Critical one-tail	1.695519	
12	Volkswagen Jetta TDI	Small Sedan	25,100	0.50	68	2	1.18	P(T<=t) two-tail	9.35E-09	
13	Chevrolet Cruze LS (1.8)	Small Sedan	18,375	0.57	67	1	0.96	t Critical two-tail	2.039513	
14	Chevrolet Cruze 1LT (1.4T)	Small Sedan	20,530	0.60	69	1	0.91			
15	Nissan Altima 2.5 S (4-cyl.)	Family Sedan	23,970	0.59	91	4	1.75	-9446.05357		
16	Kia Optima LX (2.4)	Family Sedan	21,885	0.58	81	4	1.73	-5514.34643		
17	Subaru Legacy 2.5i Premium	Family Sedan	23,830	0.59	83	4	1.73			
18	Ford Fusion Hybrid	Family Sedan	[illegible]	[illegible]	84	5	[illegible]			

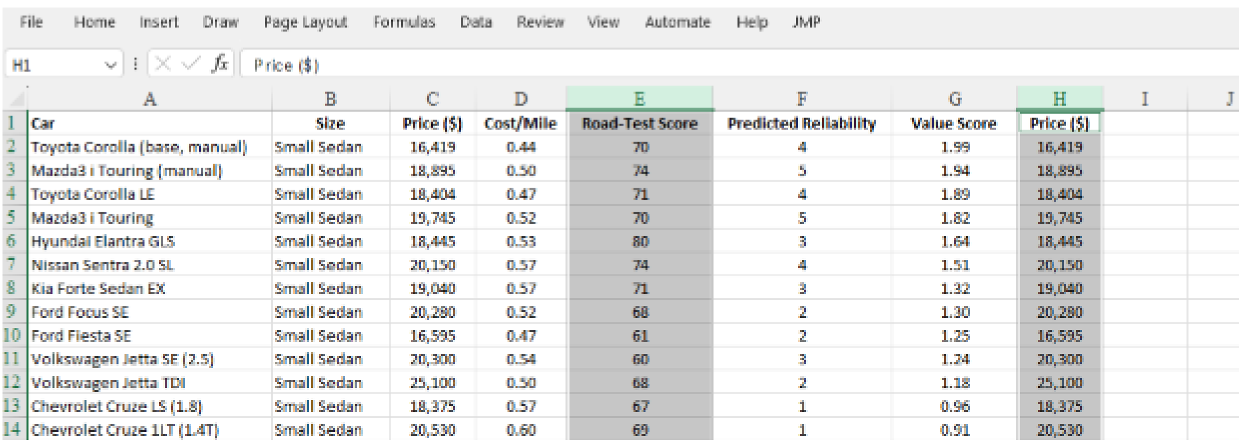
When you interpret this confidence interval make sure to remember which group you called group 1 and which group you called group 2, because the confidence interval is estimating $\mu_1 - \mu_2$. In this example, small sedan is group 1 and family sedan is group 2, so the confidence interval bounds being negative means that the population mean price for small sedans is less than the population mean price for family sedans.

One Quantitative Response Variable

- One quantitative explanatory variable (Simple Linear Regression)
  - Descriptive statistics and graphical summaries
    - Scatterplot
    - SLR model, correlation, r-square

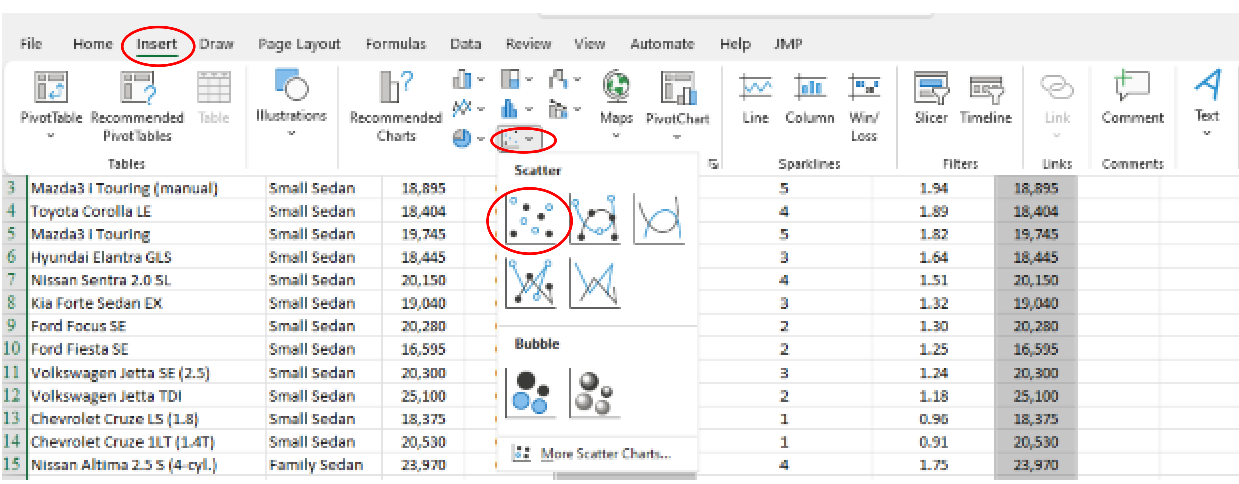
Here we'll use the carvalues data set, specifically the Price variable as the response variable and Road-Test Score as the explanatory variable. To use the following method to create a scatterplot, the response variable needs to be to the right of the explanatory variable. Because I want Price to be the response variable (and show up on the vertical axis or y-axis), I copied it to the far right. There are other ways to get around this problem, but all are more complicated and no more effective.

Then highlight the column for Road-Test Score, then hold down the ctrl button, and highlight the Price column on the far right side that you pasted on the far right side.

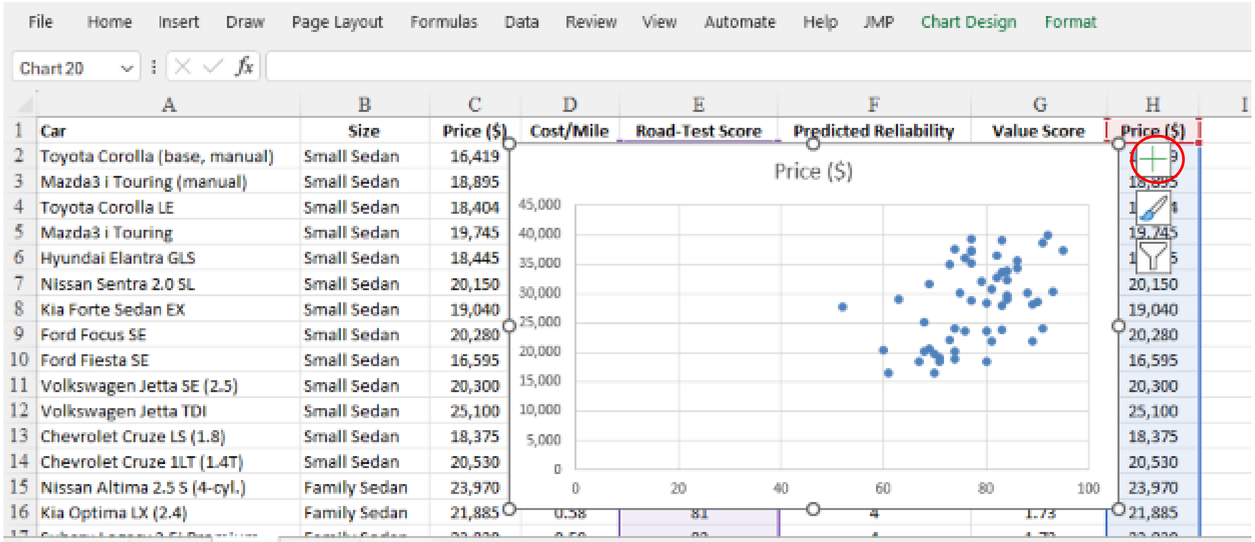


	Car	Size	Price ($)	Cost/Mile	Road-Test Score	Predicted Reliability	Value Score	Price ($)
2	Toyota Corolla (base, manual)	Small Sedan	16,419	0.44	70	4	1.99	16,419
3	Mazda3 i Touring (manual)	Small Sedan	18,895	0.50	74	5	1.94	18,895
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89	18,404
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82	19,745
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64	18,445
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51	20,150
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32	19,040
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30	20,280
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25	16,595
11	Volkswagen Jetta SE (2.5)	Small Sedan	20,300	0.54	60	3	1.24	20,300
12	Volkswagen Jetta TDI	Small Sedan	25,100	0.50	68	2	1.18	25,100
13	Chevrolet Cruze LS (1.8)	Small Sedan	18,375	0.57	67	1	0.96	18,375
14	Chevrolet Cruze 1LT (1.4T)	Small Sedan	20,530	0.60	69	1	0.91	20,530

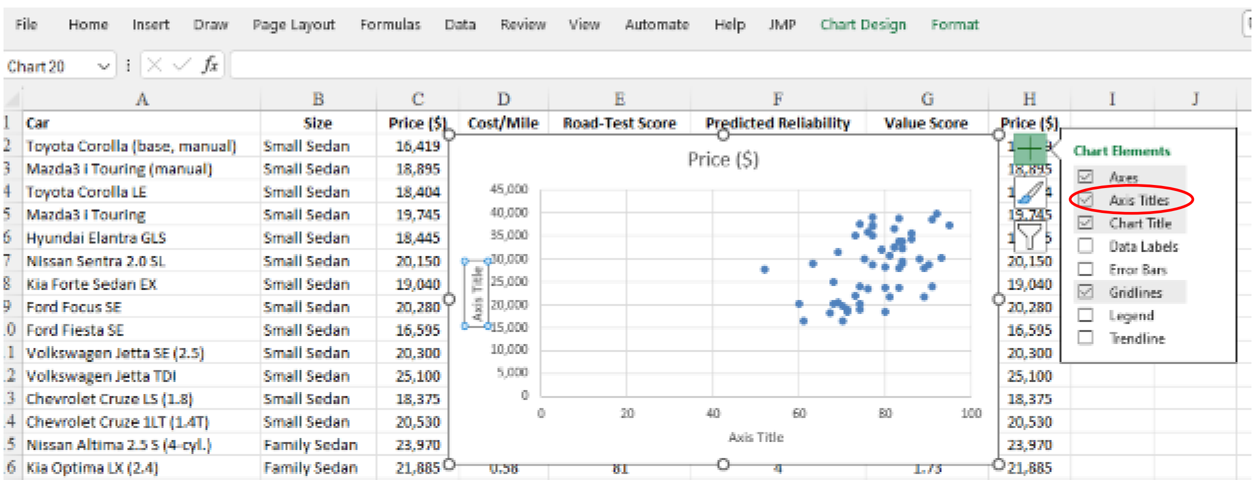
Click on the Insert tab near the top left, select the menu for scatterplots, and from that list, the scatterplot as shown below.



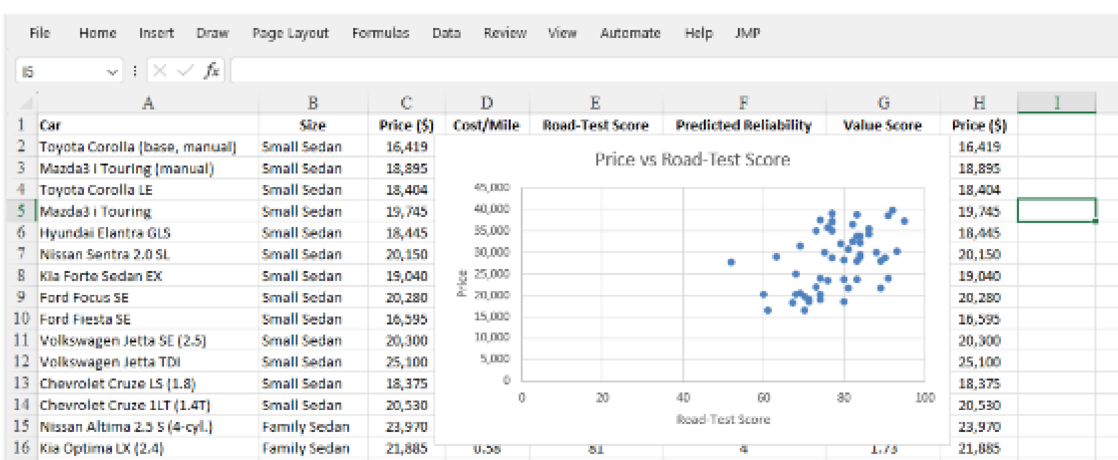
That should create the following scatterplot. Note that Price is on the vertical axis (y-axis) as it should be.



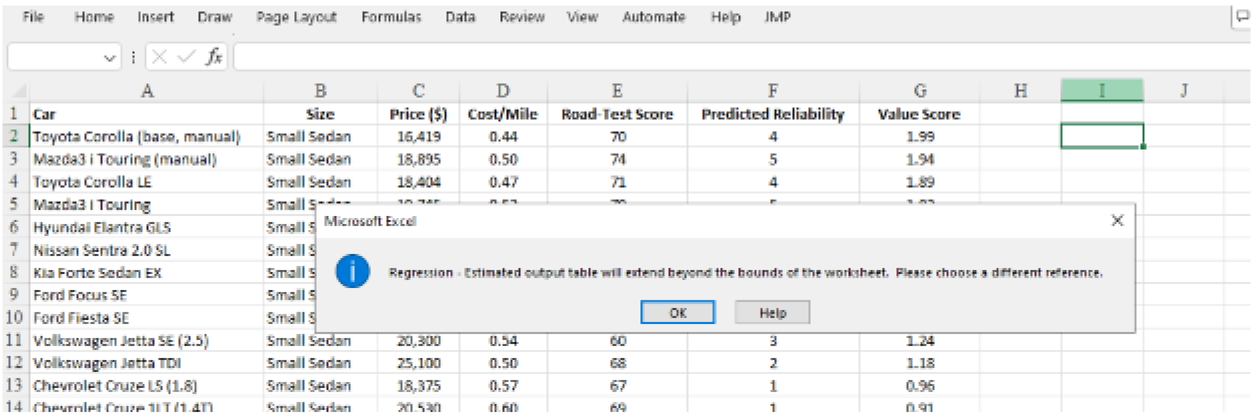
You will want to label the axes, so you can click on the + sign in the upper right (circled above). Then you check the box for Axis Titles as circled below.



You can click on the label for each axis to type the variable name for each. You can also click on the main title and change it if you wish, which I did in the image below.

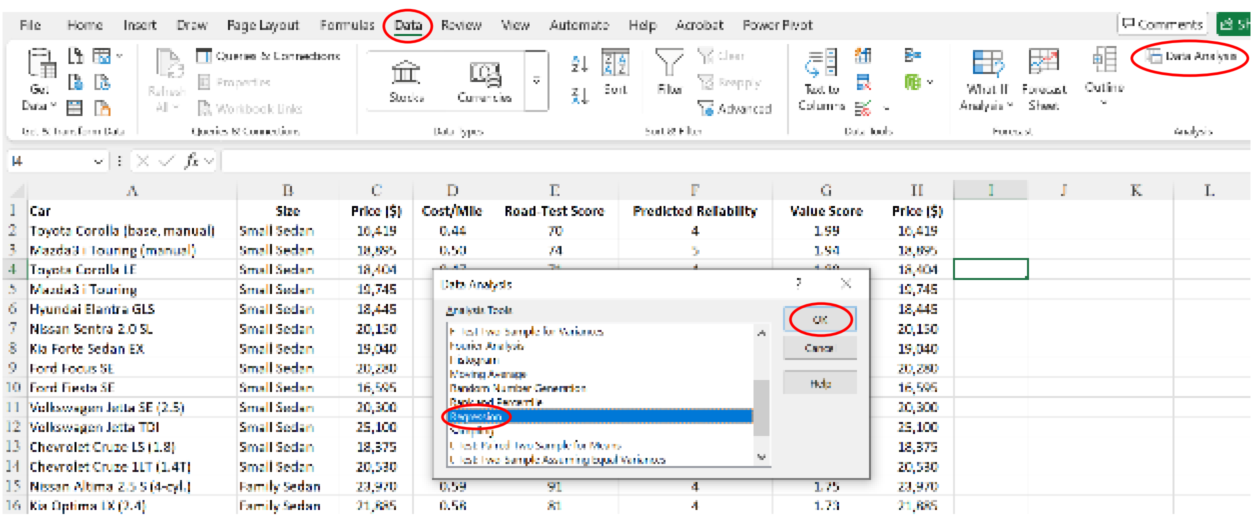


Very important: For producing statistical analyses (not graphs) for regression, when you highlight the data for each variable, do NOT highlight the entire column, because this will cause an error, where Excel will show the following error message.

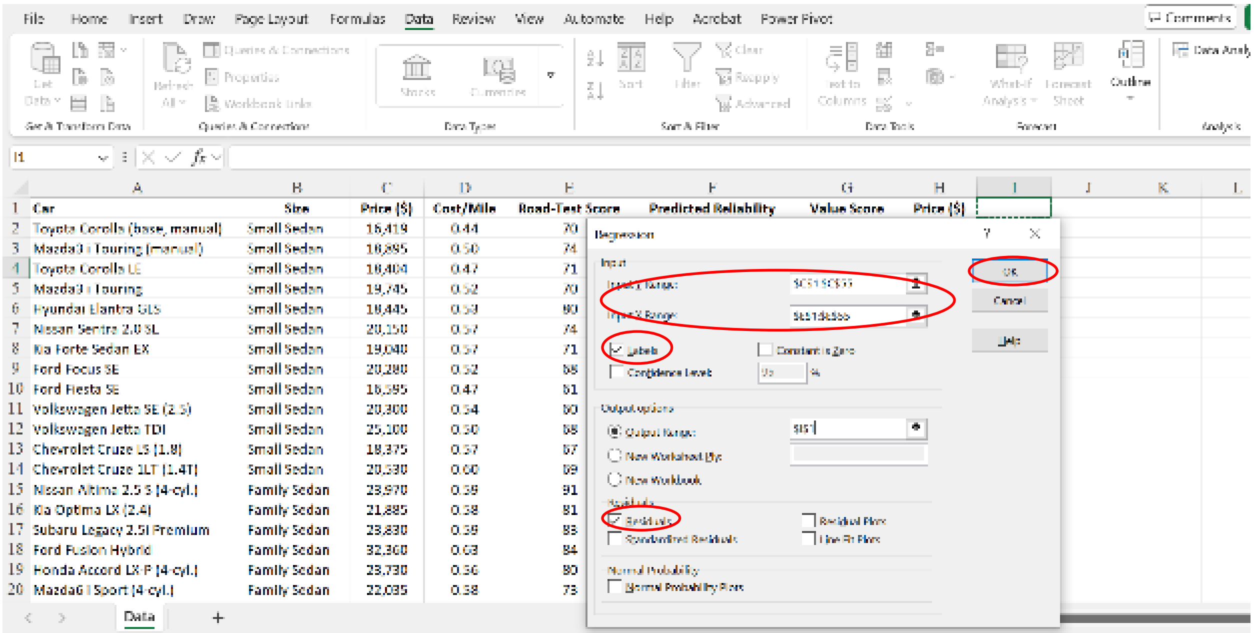


You should only highlight the cells with numbers in them (rather than also including the empty cells below the cells with numbers in them), along with the variable name in the first row. Here we want to use Price as the response variable and Road-Test Score as the explanatory variable, just as we did above for producing the scatterplot.

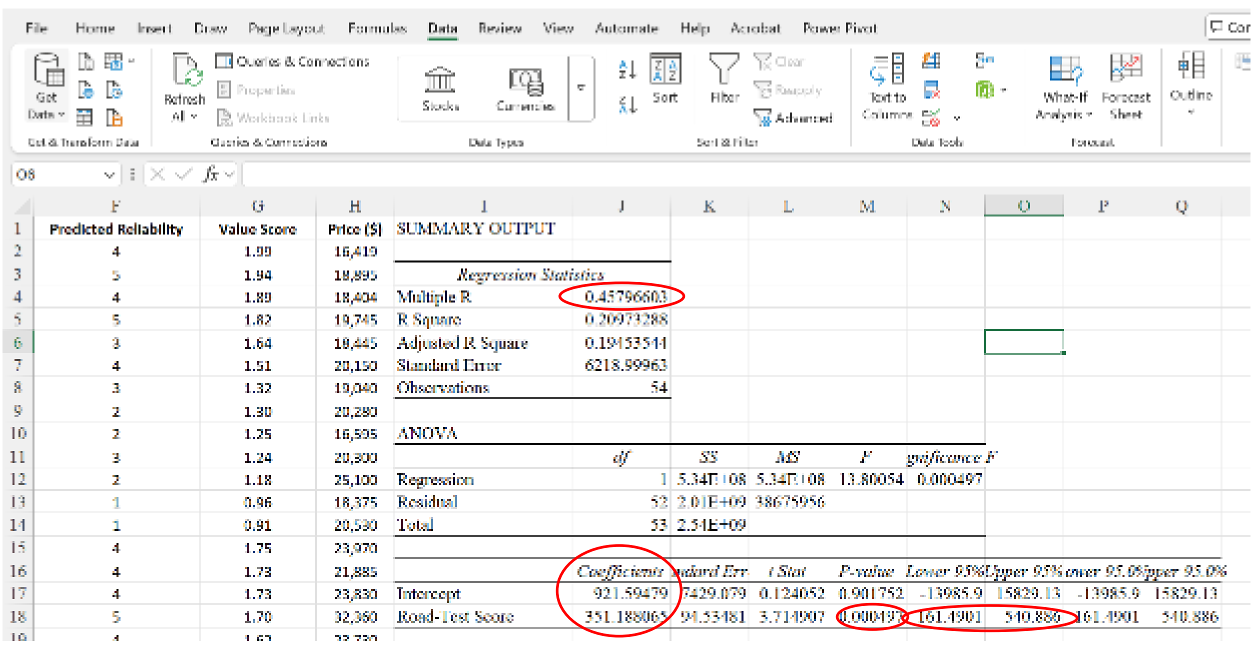
Click on the Data tab at the top, and then Data Analysis on far right, and select Regression, and then click OK as shown below:

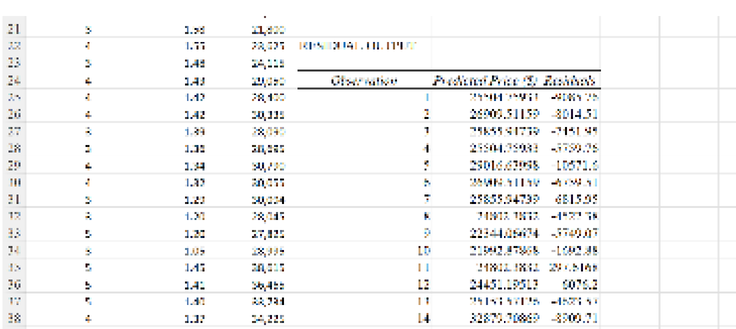


This will create the following dialog, where you can tell Excel that Price is the Y (response) variable in Input Y Range, and that Road-Test Score is the X (explanatory) variable in Input X Range. Note that I have included the first row (but no empty cells below the cells with numbers in them), which is the variable names, so I should click on Labels. Then click on Residuals. Your instructor may want you to select more options, but this is all we need for further analysis as shown in this document, so click OK. Note that I told Excel to put the output in cell I1 (by clicking in Output Range and then clicking on cell I1), but you can put it wherever you like.



This should produce output that looks like this (you'll need to scroll to right and/or down to see all of it, and perhaps adjust column widths to see all labels below).





The (linear) correlation coefficient is equivalent to what Excel calls "Multiple R" (circled above) because we only have one explanatory variable (simple linear regression). The equation of the least squares line can be written using the Coefficients values circled above. The p-value to test the null hypothesis that the population slope can be found under P-value as circled above. The 95% confidence interval for the population slope is also circled above. For both the p-value and confidence interval, make sure you use the second row, which correspond to the explanatory variable, here Road-Test, as opposed to the first row which are for the y-intercept, which we generally will not be interested in.

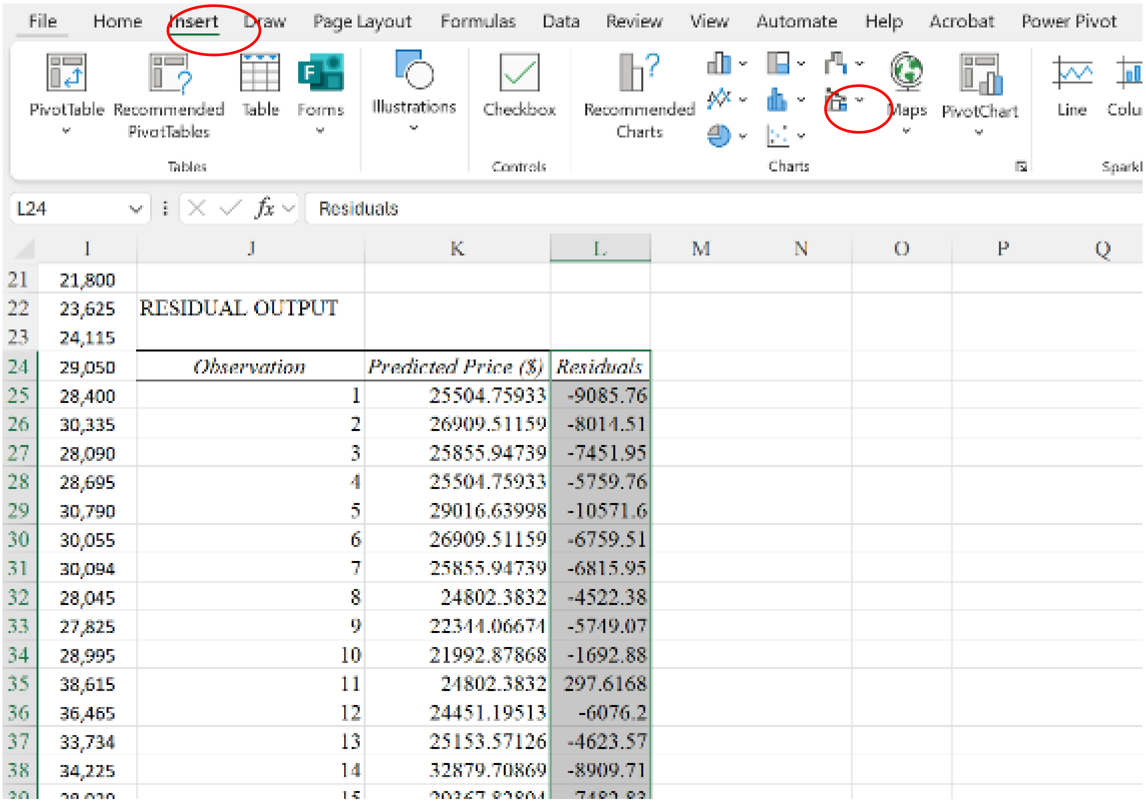
- Residuals, including graph of residual vs predicted values

In order for the p-value and confidence interval to be valid, we need to check assumptions, just as for any hypothesis test or confidence interval. First, we need that the scatterplot of the response variable and explanatory variable are linear and don't have any extreme outliers, which can be done with a scatterplot as produced above.

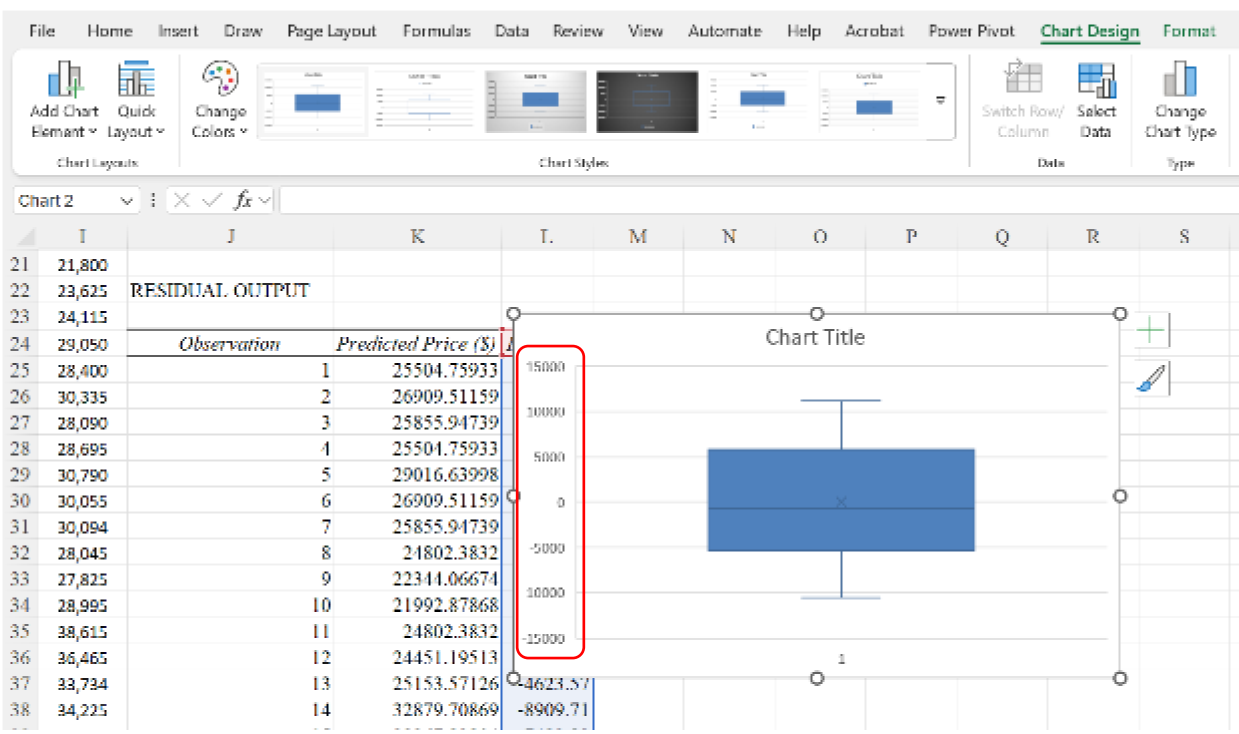
We also need that the residuals are normally distributed, and that the scatterplot of residuals vs predicted values shows constant variance (so does not have a funnel shape). We have everything we need to create these plots because we told Excel to create these, which are circled below. We can also detect outliers in either of the following plots, and look for non-linearity in the scatterplot of residuals vs predicted.

21	3	1.58	21,800			
22	4	1.55	23,625	RESIDUAL OUTPUT		
23	3	1.48	24,115			
24	4	1.43	29,050	*Observation*	*Predicted Price ($)*	*Residuals*
25	4	1.42	28,400	1	25504.75933	-9085.76
26	4	1.42	30,335	2	26909.51159	-8014.51
27	3	1.39	28,090	3	25855.94739	-7451.95
28	3	1.36	28,695	4	25504.75933	-5759.76
29	4	1.34	30,790	5	29016.63998	-10571.6
30	4	1.32	30,055	6	26909.51159	-6759.51
31	3	1.29	30,094	7	25855.94739	-6815.95
32	3	1.20	28,045	8	24802.3832	-4522.38
33	5	1.20	27,825	9	22344.06674	-5749.07
34	3	1.05	28,995	10	21992.87868	-1692.88
35	5	1.45	38,615	11	24802.3832	297.6168
36	5	1.41	35,465	12	24451.19513	-6076.2
37	5	1.40	33,734	13	25153.57126	-4623.57
38	4	1.37	34,225	14	32879.70869	-8909.71

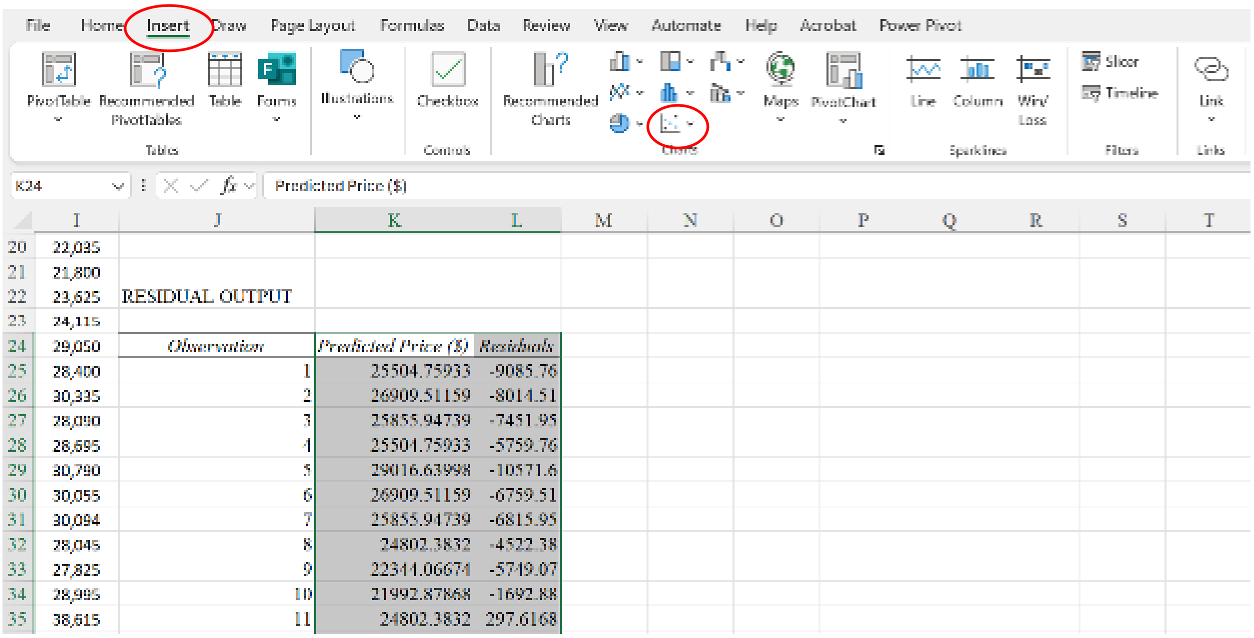
You can make whatever kind of plot your instructor prefers for checking normality, and we showed earlier how to make histograms and boxplots. You just have to apply them to the residuals. I'll use a boxplot here. Highlight the residuals and then click on the Insert tab, and click on the icon that looks like a histogram as shown here (and as shown earlier), and then select Box and Whisker.



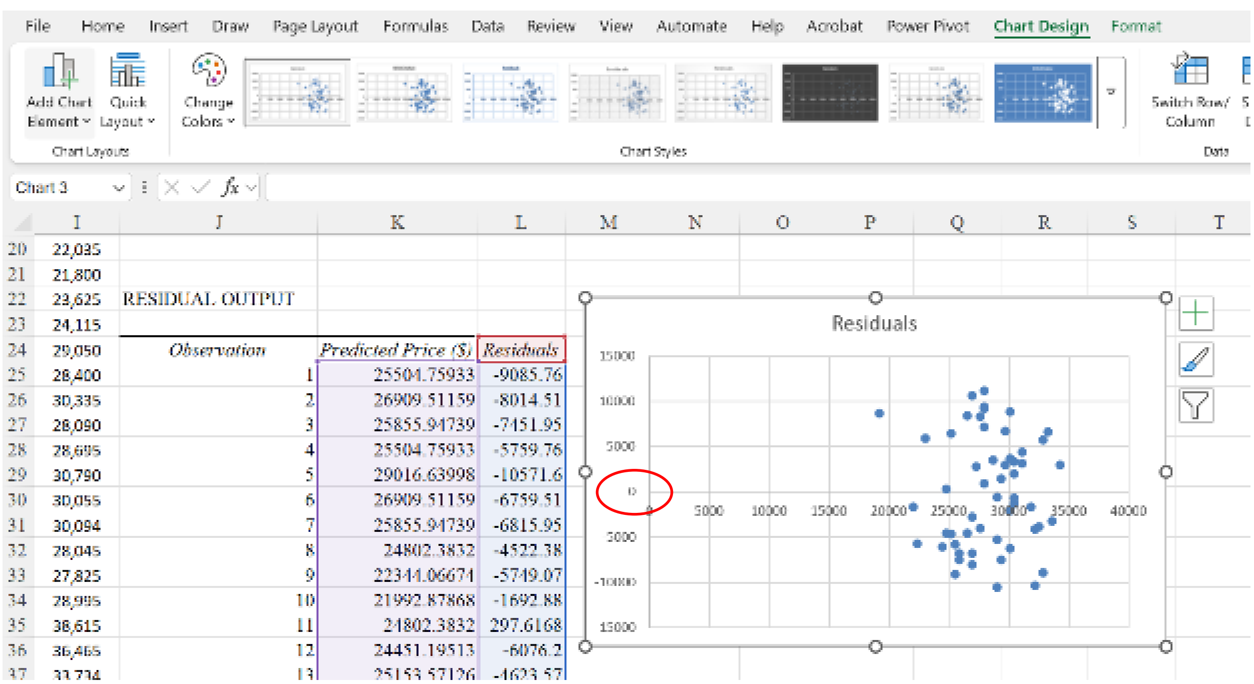
This should produce the following boxplot of the residuals. We can make the labels more appropriate, but regardless, the boxplot looks pretty symmetric and does not show any outliers. Note the scale on the vertical axis has 0 in the middle, which is how you can verify you are actually plotting residuals. In other words, residuals can be both positive and negative as circled below.



To make a scatterplot of residuals vs predicted values, you can do exactly what we did earlier for scatterplots, but first highlight the predicted values (here, Predicted Price) and the residuals, and then click on Insert tab, and then the icon that looks like a scatterplot shown here:



Which should produce this plot. We can clean up the labels as before, but the point is that this plot does not show any funnel shape. It also does not show any extreme outliers, and does not have any general pattern at all (like a linear pattern), which is what we hope to see.



Because residuals can be negative, note that the vertical axis again has 0 in the middle (as circled above), which means the residuals are on the vertical axis as we would like.

One Quantitative Response Variable

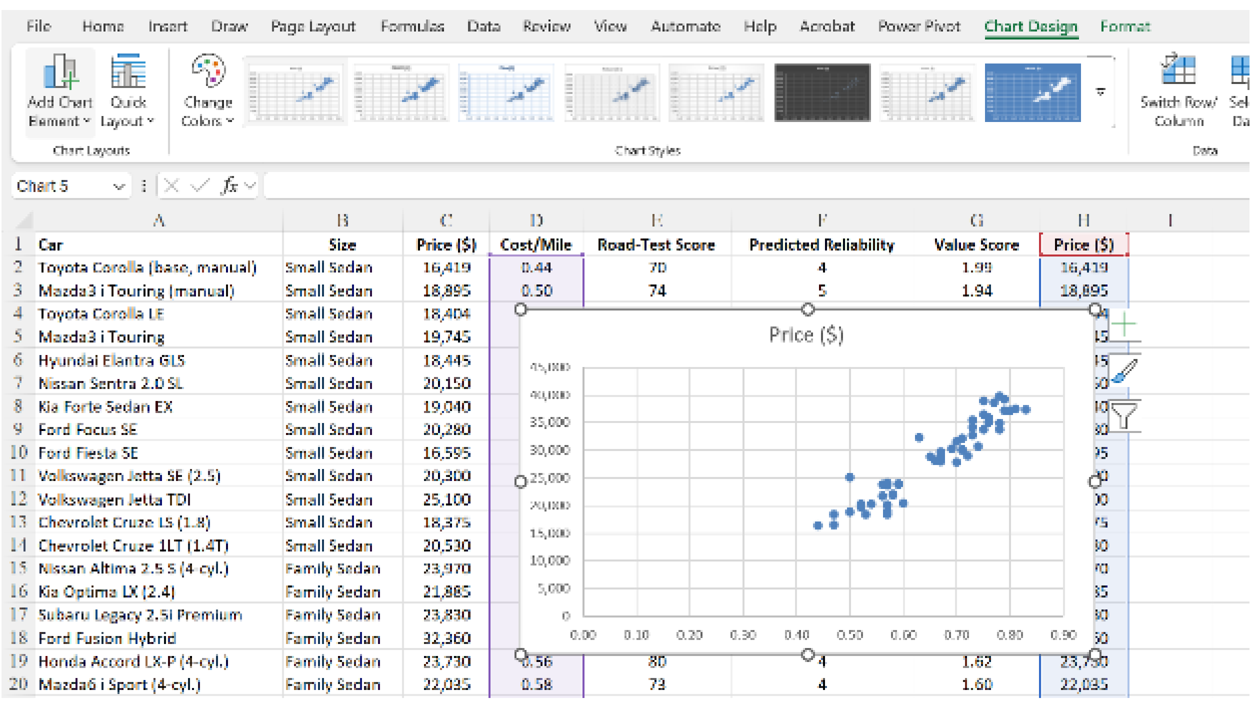
- **Two or more explanatory variables (Multiple Linear Regression)**
  - **Descriptive statistics and graphical summaries**
    - **Scatterplots of response variable vs each quantitative explanatory variable (including detection of non-linearity and outliers)**

Note that in STA 225, we do not cover categorical explanatory variables (because SCB faculty preferred that we skipped that material), so our focus here is limited to quantitative explanatory variables.

Very important: In order to use the Regression feature of the Data Analysis Toolpak, the columns containing the explanatory variables must be next to each other in the spreadsheet. The column for the response variable can be anywhere.

In general, when performing multiple linear regression, the first step is to make scatterplots of all quantitative variables in pairs, regardless of which variable is the response variable and which variables are the explanatory variables. This is often called a scatter matrix. However, neither Excel nor the Data Analysis Toolpak has a way to do this. This just means you should make all of these scatterplots one by one, to look for linearity, strong relationships and outliers. Implementing these in Excel is no different than show above, you just need to make more of them. Remember that the response variable needs to be to the right of the explanatory variable in the spreadsheet.

For example, using the carvalues data set with Price as the response variable, I made one scatterplot of Price vs Cost/Mile as show below, by highlighting Cost/Mile, hold down the ctrl button, and then highlight Price, then make the scatterplot by clicking on the scatterplot icon as before.



You should make a scatterplot of Road-Test vs Price, Predicted Reliability vs Price, and Value Score vs Price, to verify linearity and look for outliers. It is also a good idea to make scatterplots of just the explanatory variables, so Cost/Mile vs Road-Test Score, Cost/Mile vs Predicted Reliability, etc., and look for strong relationships between the explanatory variables (to identify redundancy also known as collinearity), in addition to linearity and outliers.

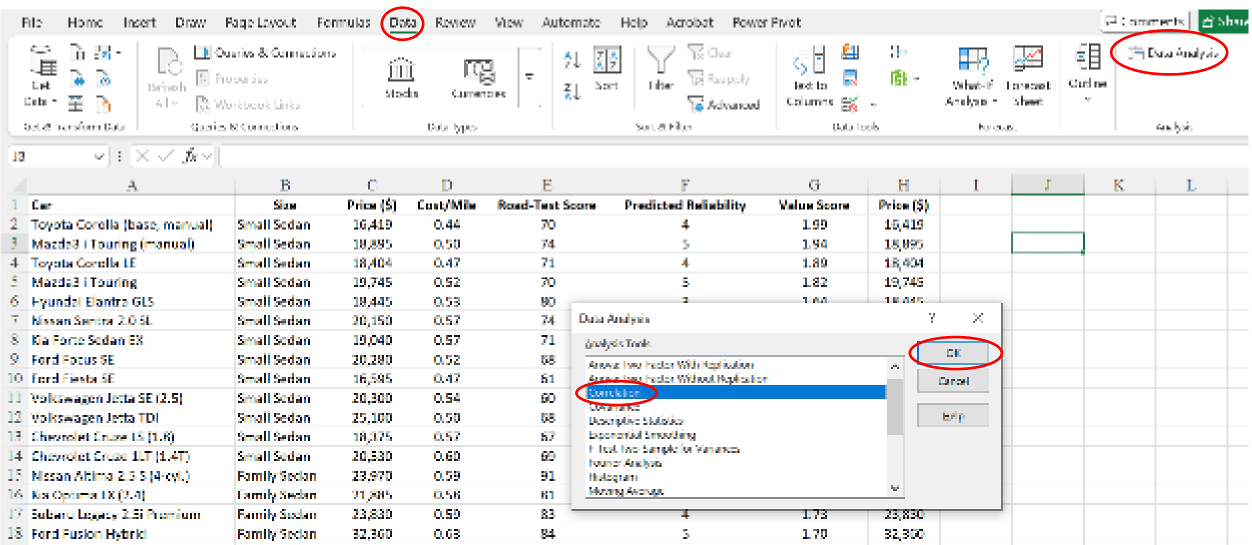
We are going to calculate correlations amongst all these pairs of variables next, and for correlations to be valid, we have to first verify linearity and no extreme outliers.

One Quantitative Response Variable

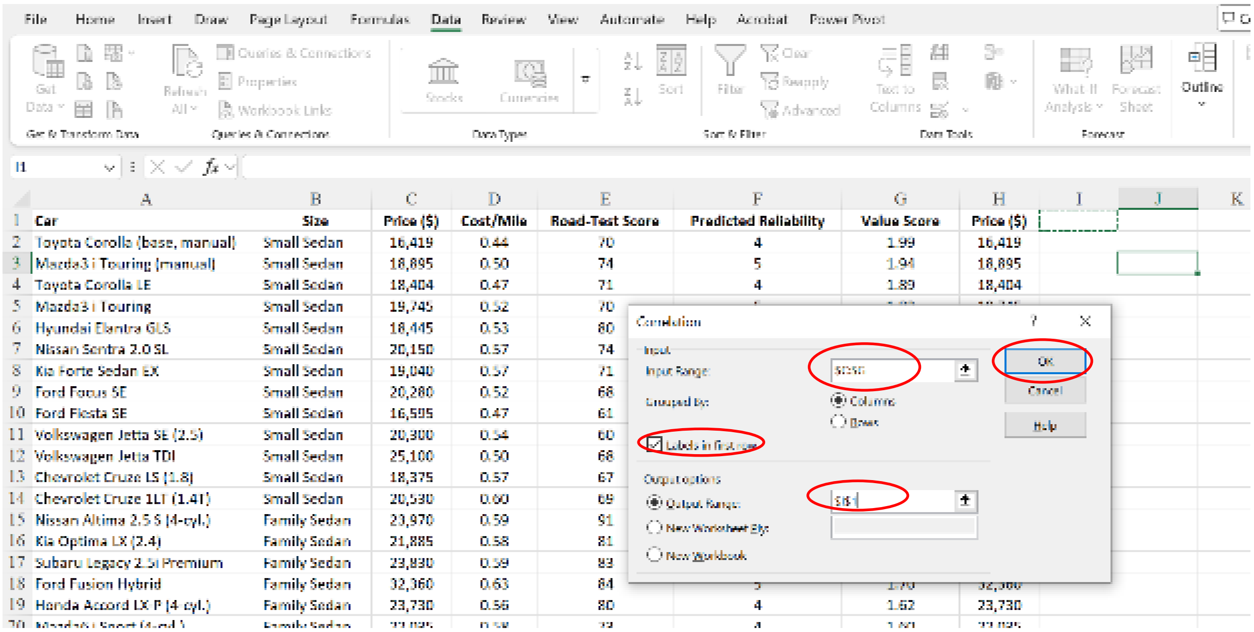
- **Two or more quantitative explanatory variables (Multiple Linear Regression)**
    - **Descriptive statistics and graphical summaries**
        - **Correlation matrix (including collinearity)**

Even though Excel will not make a scattermatrix for us, it will make a correlation matrix. This gives us a lot of information all at once.

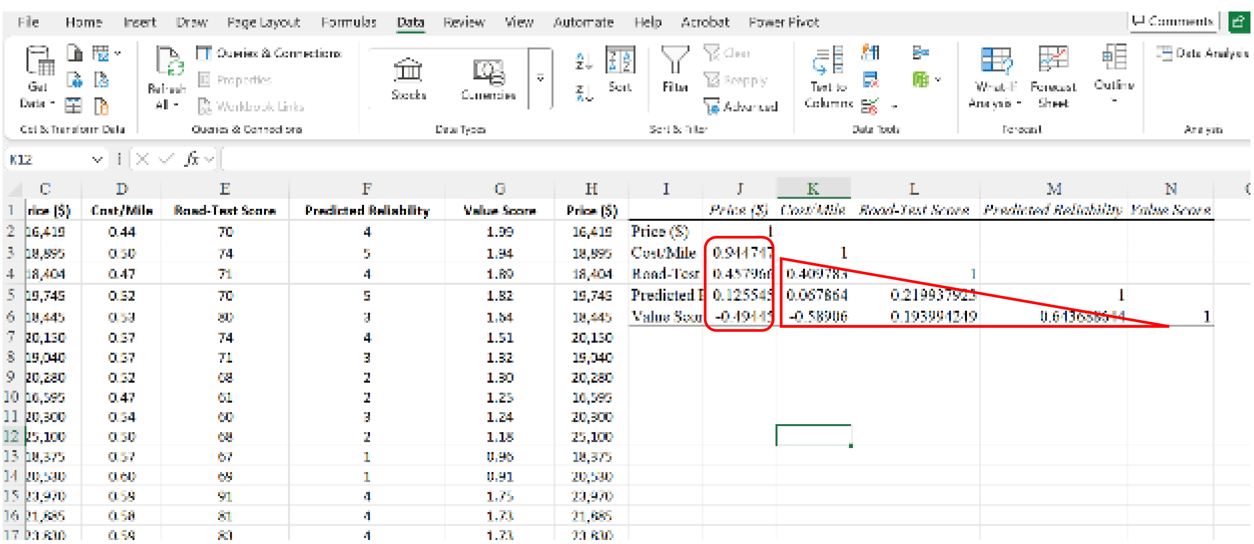
Go to Data tab, click on Data Analysis on far right, and select Correlation, and then click OK, as circled below.



Click in Input Range, then highlight all relevant quantitative variables (both response variable and all explanatory variables). Because I highlighted the variable names in row 1, I clicked on Labels in first row. Then you can tell Excel where to put the output. I chose cell I1. Then click OK.



After adjusting column widths in the resulting table so I could see the variable names, I get the following table.



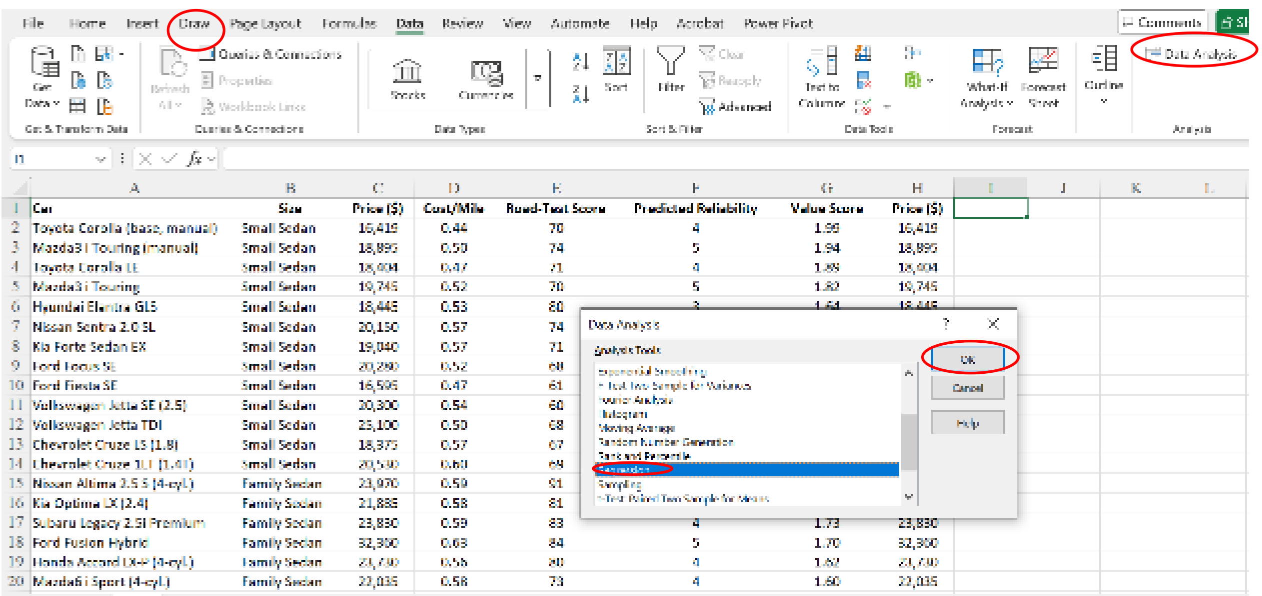
The values (in the table we just produced) under Price circled above are the correlations of each explanatory variable with Price. The other values (separately circled above) are the correlations between the explanatory variables, and ideally, we'd like all of these to be near 0. For example, Value Score and Predicted Reliability have a moderate correlation of 0.64, which indicates we may not need both of them in a multiple regression model. We'll ignore this going forward for the purposes of this document, but that can be unwise in practice, because this is an indication of collinearity.

One Quantitative Response Variable

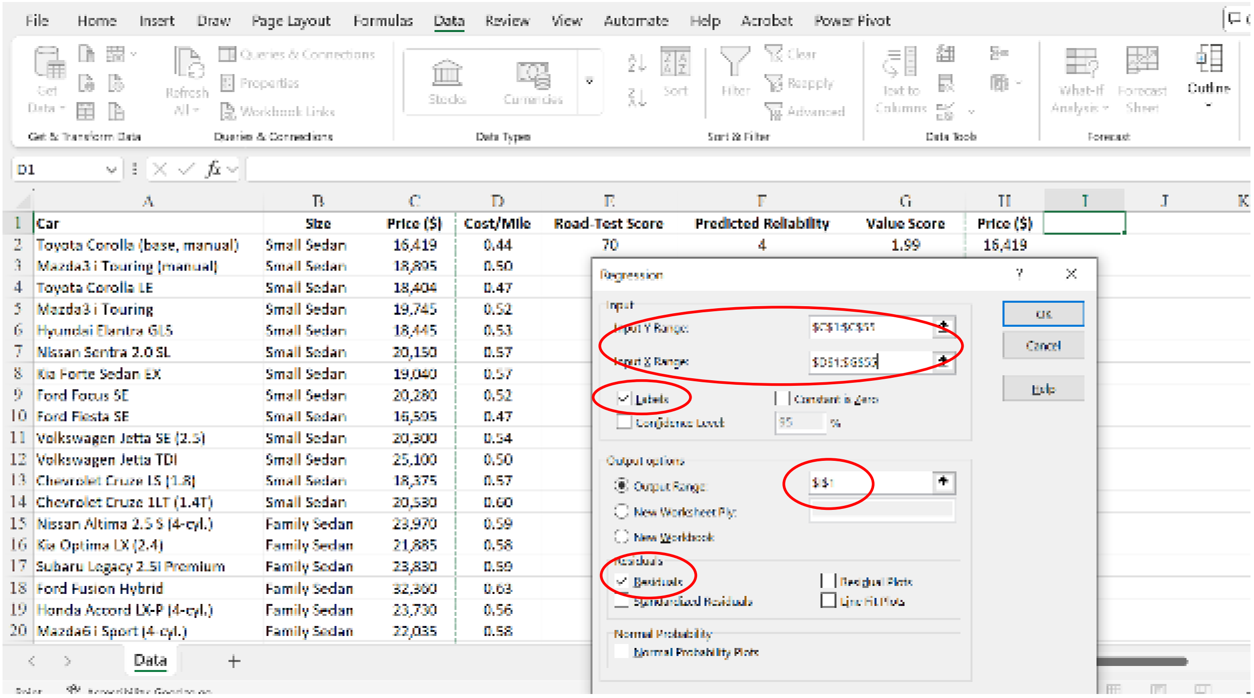
- **Two or more quantitative explanatory variables (Multiple Linear Regression)**
    - **Descriptive statistics and graphical summaries**
        - **MLR model and prediction, especially as it relates to forecasting**
    - **Inferential statistics**
        - **P-value for overall F test, p-value for individual slopes**
        - **Checking conditions**

To get multiple linear regression output, you can follow the same steps as we did earlier for simple linear regression. Remember that you should only highlight cells with numbers in them, as opposed to entire columns.

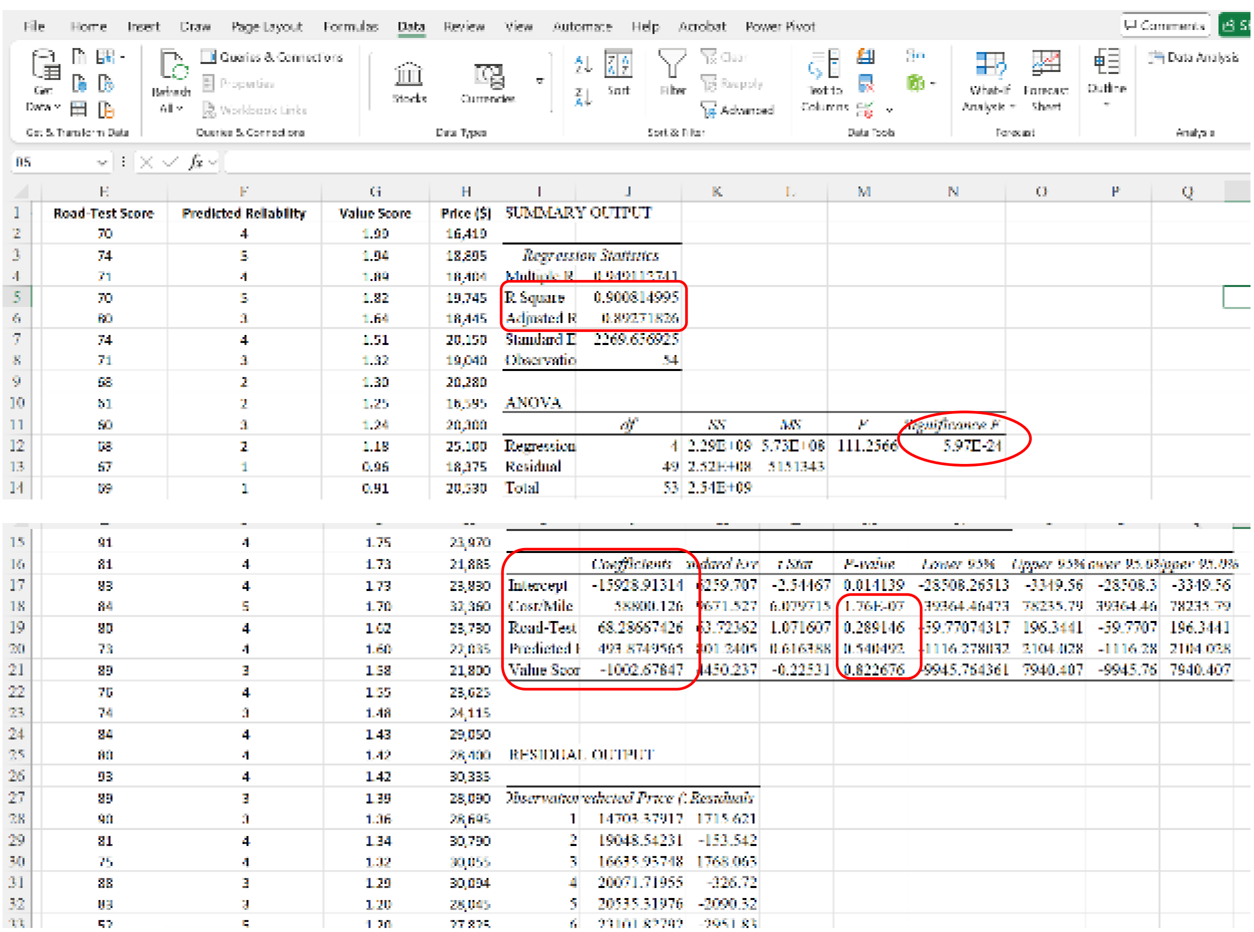
Very important: As noted above, all explanatory variables must be next to each other in the spreadsheet. If they are not, then you need to do some copying and pasting to make this true before asking Excel to create output for multiple linear regression.



Click in Input Y Range, and highlight cells under Price (not entire column), including the first row with the name Price in it. Then click in Input X Range and highlight the cells in columns for Cost/Mile, Road-Test Score, Predicted Reliability and Value Score, making sure not to highlight the entire columns as before. Click on Labels and Residuals. I asked Excel to put the output in cell I1 again.

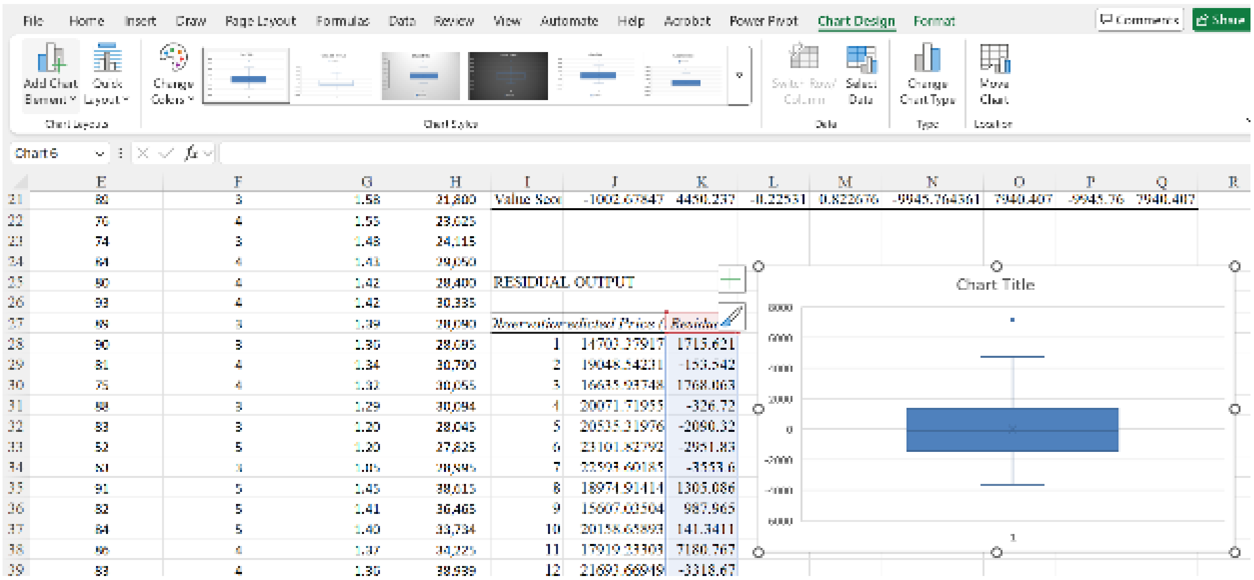


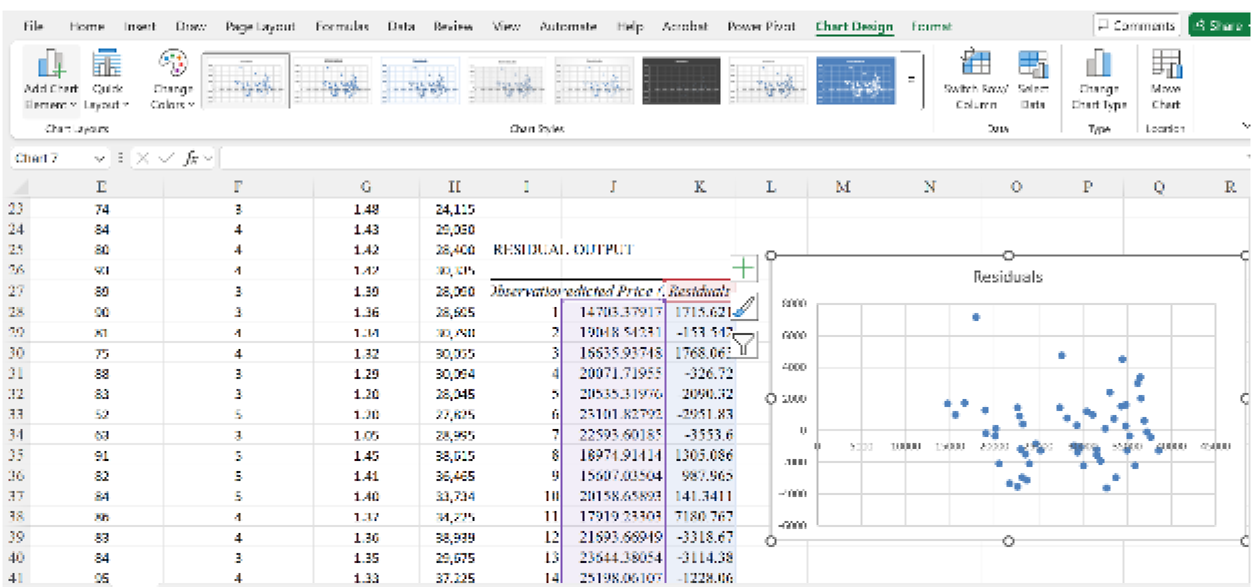
This should produce the following output, after clicking OK, and after adjusting the width of a couple columns so I could more easily read the labels.



R-Square and Adjusted R-Square are circled above. In the ANOVA table, Significance F is circled, which is the p-value simultaneously testing if all population slopes are 0. The next table of output, under Coefficients, gives the sample y-intercept and sample slopes, which are circled. In the same table under P-value are the p-values for testing if each individual population slope is equal to 0 (these are two-tailed p-values). Note we are not interested in the p-value for the y-intercept.

We still need to check for normality of residuals (boxplot of residuals) and constant variance in the residuals (scatterplot of residuals vs predicted values. This is done the same way that we showed for simple linear regression, but we show it again here.





One categorical response variable

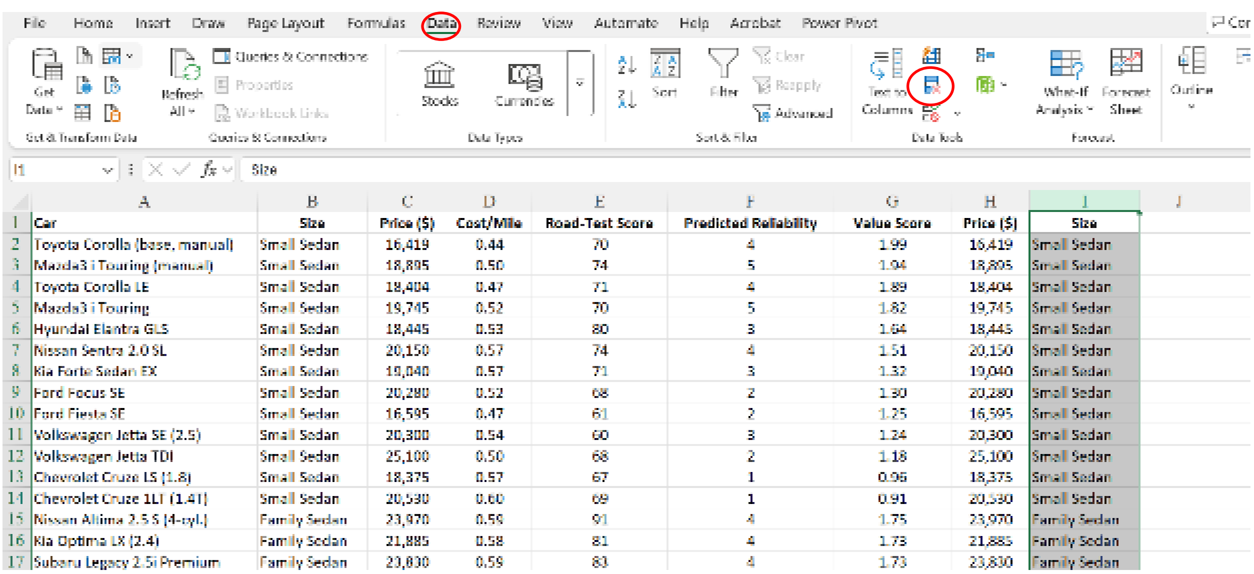
- **No explanatory variables**
  - **Descriptive statistics and graphical summaries**
    - **Proportion**
    - **Bar chart and pie chart**

Here we use the carvalues data set, and focus on the Size variable, and supposed we want to focus on percent of cars that are small sedans. We would technically need that our data set came from a random sample, but we'll not worry about that for this example (though you should in general).

First, we need to get Excel to tell us how many times each category appears. You'd think this would be pretty easy in Excel, but it is not, or it is not as easy as it should be, meaning that there is nothing in the Data Analysis Toolpak to do this.

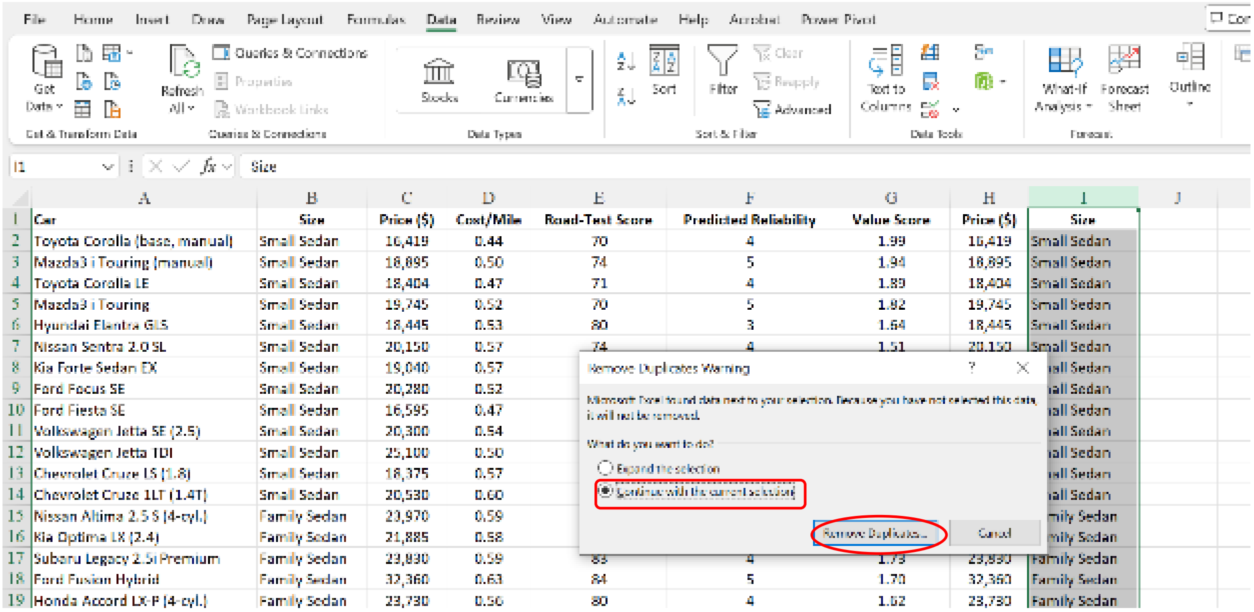
The first step is to copy the categorical variable Size to the far right, because we are going to tell Excel to eliminate duplicate values, and we don't want to do that in the original column. This will allow us to tell Excel to count how many times each unique category appears in our data.

Highlight the column for Size on the far right, then click on Data tab, and then in the Data Tools section, click on Remove Duplicates as circled below.

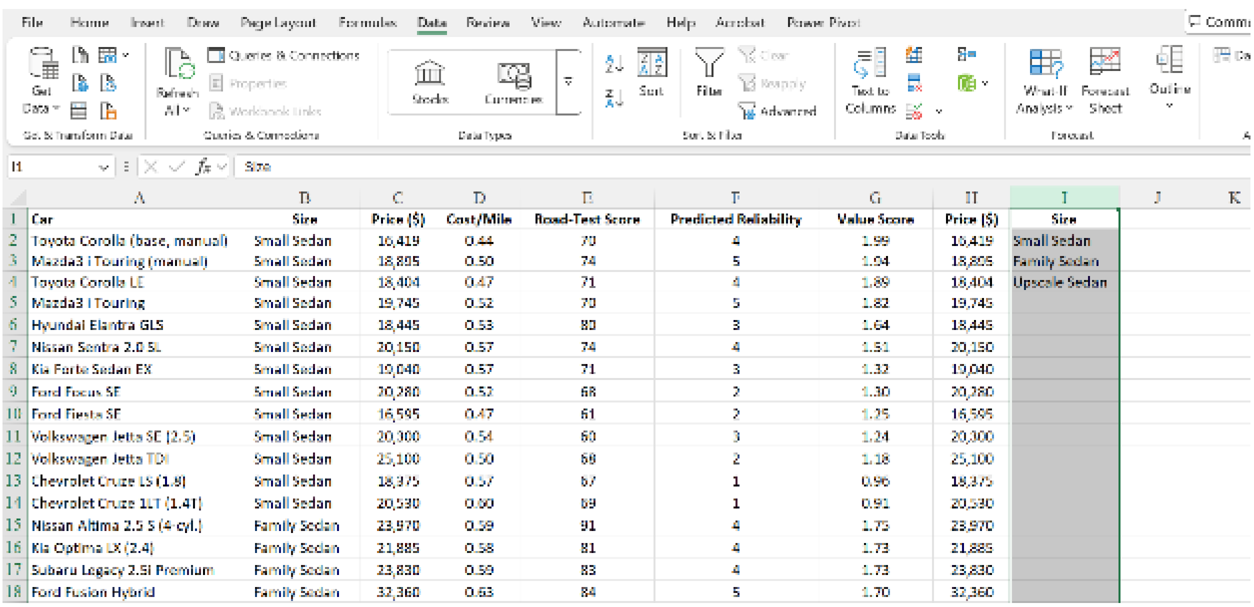


	Car	Size	Price ($)	Cost/Mile	Road-Test Score	Predicted Reliability	Value Score	Price ($)	Size
2	Toyota Corolla (base, manual)	Small Sedan	16,419	0.44	70	4	1.99	16,419	Small Sedan
3	Mazda3 i Touring (manual)	Small Sedan	18,895	0.50	74	5	1.94	18,895	Small Sedan
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89	18,404	Small Sedan
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82	19,745	Small Sedan
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64	18,445	Small Sedan
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51	20,150	Small Sedan
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32	19,040	Small Sedan
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30	20,280	Small Sedan
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25	16,595	Small Sedan
11	Volkswagen Jetta SE (2.5)	Small Sedan	20,300	0.54	60	3	1.24	20,300	Small Sedan
12	Volkswagen Jetta TDI	Small Sedan	25,100	0.50	68	2	1.18	25,100	Small Sedan
13	Chevrolet Cruze LS (1.8)	Small Sedan	18,375	0.57	67	1	0.96	18,375	Small Sedan
14	Chevrolet Cruze 1LT (1.4T)	Small Sedan	20,530	0.60	69	1	0.91	20,530	Small Sedan
15	Nissan Altima 2.5 S (4-cyl.)	Family Sedan	23,970	0.59	91	4	1.75	23,970	Family Sedan
16	Kia Optima LX (2.4)	Family Sedan	21,885	0.58	81	4	1.73	21,885	Family Sedan
17	Subaru Legacy 2.5i Premium	Family Sedan	23,830	0.59	83	4	1.73	23,830	Family Sedan

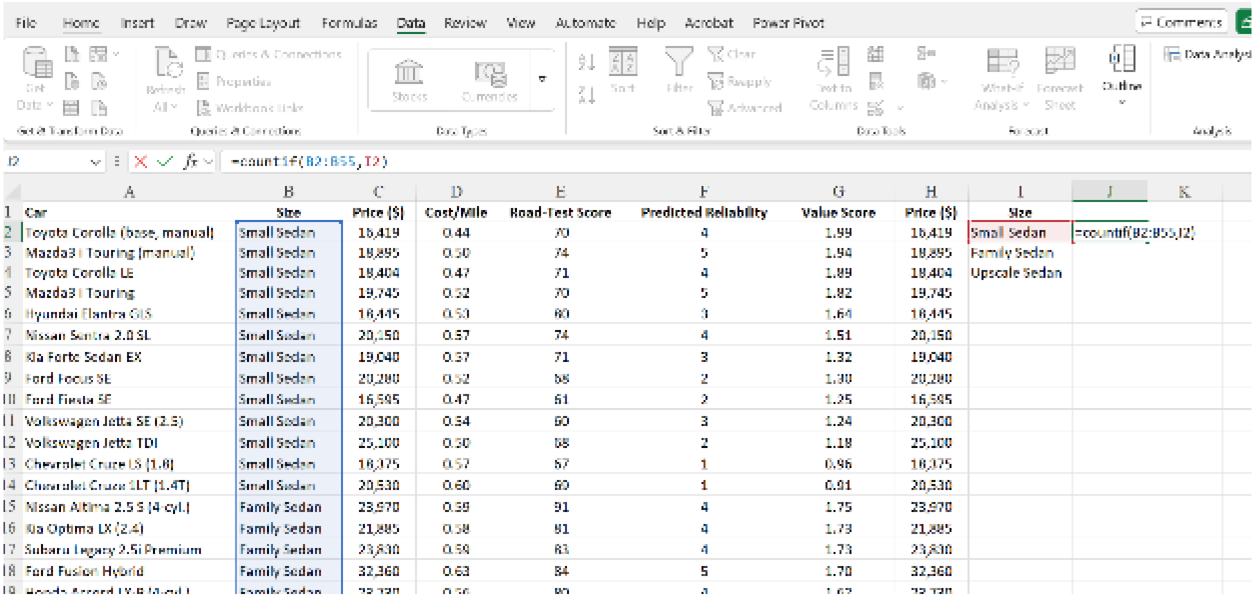
Select Continue with the current selection in the box as show below, and then Remove Duplicates, and then in the next box that pops up, click OK.



You should now see, so we can now use the COUNTIF function to get the frequencies for each of the three categories.



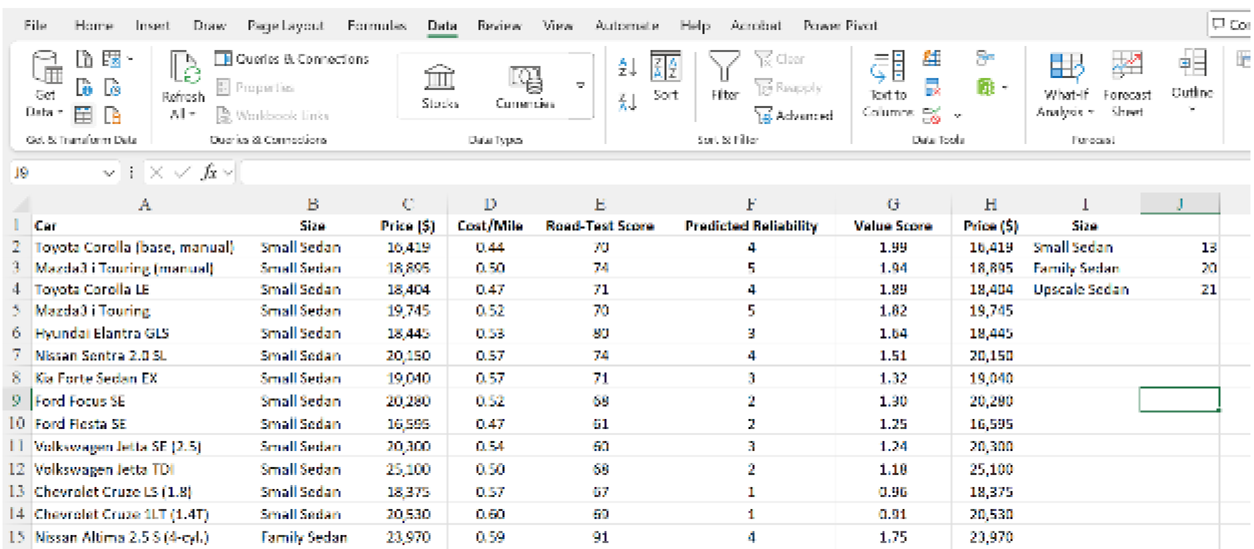
First, we can get the number of Small Sedans by highlighting the original Size column (without highlighting row 1 where the variable name Size is) by clicking in the J2 cell, and then then typing the following as shown (which tells Excel to count how many times the category in I2 cell appears), and then hitting enter



=countif(B2:B55,I2)

	A	B	C	D	E	F	G	H	I	J	K
1	Car	Size	Price ($)	Cost/Mile	Road-Test Score	Predicted Reliability	Value Score	Price ($)	Size		
2	Toyota Corolla (base, manual)	Small Sedan	16,419	0.44	70	4	1.99	16,419	Small Sedan	=countif(B2:B55,I2)	
3	Mazda3 i Touring (manual)	Small Sedan	18,895	0.50	74	5	1.94	18,895	Family Sedan		
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89	18,404	Upscale Sedan		
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82	19,745			
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64	18,445			
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51	20,150			
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32	19,040			
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30	20,280			
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25	16,595			
11	Volkswagen Jetta SE (2.5)	Small Sedan	20,300	0.54	60	3	1.24	20,300			
12	Volkswagen Jetta TDI	Small Sedan	25,100	0.50	68	2	1.18	25,100			
13	Chevrolet Cruze LS (1.8)	Small Sedan	18,375	0.57	67	1	0.96	18,375			
14	Chevrolet Cruze 1LT (1.4T)	Small Sedan	20,530	0.60	60	1	0.91	20,530			
15	Nissan Altima 2.5 S (4-cyl.)	Family Sedan	23,970	0.59	91	4	1.75	23,970			
16	Kia Optima LX (2.4)	Family Sedan	21,885	0.58	81	4	1.73	21,885			
17	Subaru Legacy 2.5i Premium	Family Sedan	23,830	0.59	83	4	1.73	23,830			
18	Ford Fusion Hybrid	Family Sedan	32,360	0.63	84	5	1.70	32,360			

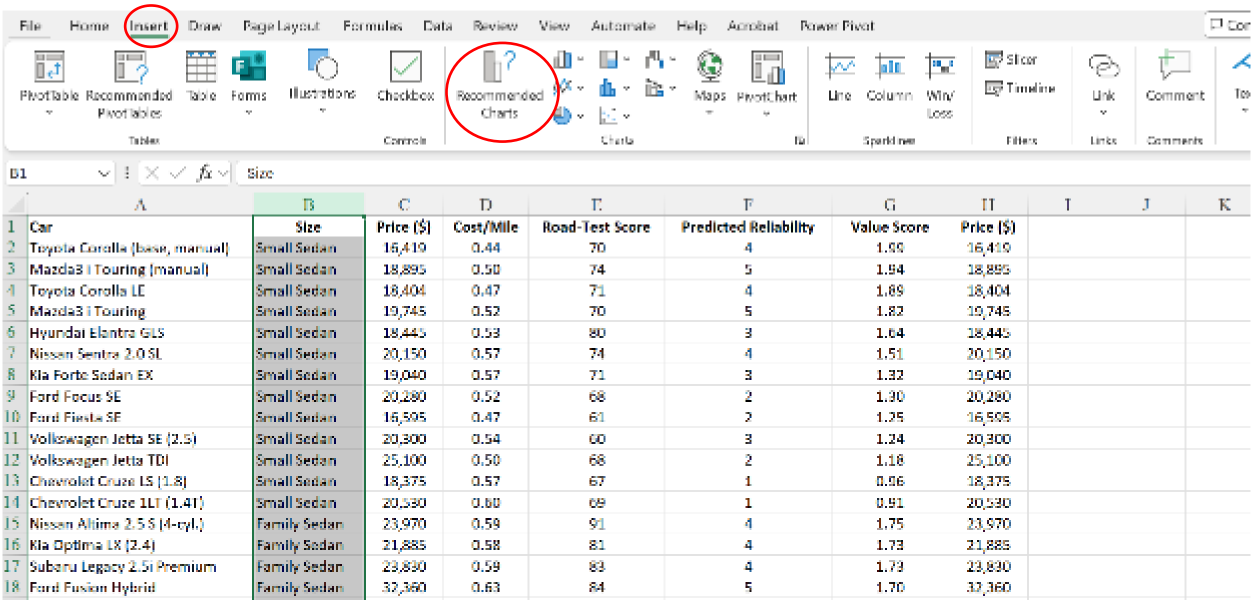
Then copy the formula in cell J2 and paste into cells J3 and J4 to get the following



	A	B	C	D	E	F	G	H	I	J
1	Car	Size	Price ($)	Cost/Mile	Road-Test Score	Predicted Reliability	Value Score	Price ($)	Size	
2	Toyota Corolla (base, manual)	Small Sedan	16,419	0.44	70	4	1.99	16,419	Small Sedan	13
3	Mazda3 i Touring (manual)	Small Sedan	18,895	0.50	74	5	1.94	18,895	Family Sedan	20
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89	18,404	Upscale Sedan	21
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82	19,745		
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64	18,445		
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51	20,150		
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32	19,040		
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30	20,280		
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25	16,595		
11	Volkswagen Jetta SE (2.5)	Small Sedan	20,300	0.54	60	3	1.24	20,300		
12	Volkswagen Jetta TDI	Small Sedan	25,100	0.50	68	2	1.18	25,100		
13	Chevrolet Cruze LS (1.8)	Small Sedan	18,375	0.57	67	1	0.96	18,375		
14	Chevrolet Cruze 1LT (1.4T)	Small Sedan	20,530	0.60	60	1	0.91	20,530		
15	Nissan Altima 2.5 S (4-cyl.)	Family Sedan	23,970	0.59	91	4	1.75	23,970		

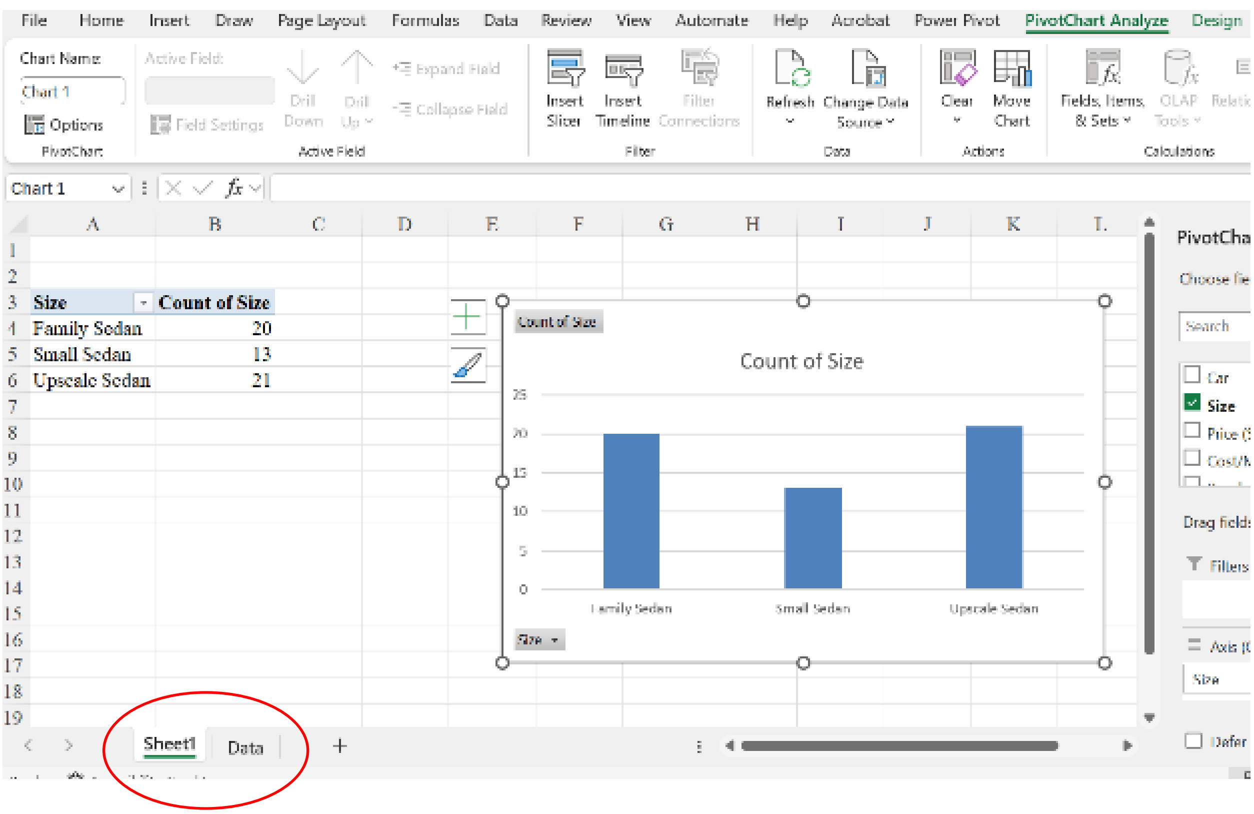
Which says that there are 13 Small Sedans, 20 Family Sedans, and 21 Upscale Sedans. If we care about proportion of Small Sedans, we can just calculate 13 out of the sample size of n = 13+20+21 = 54, meaning that the sample proportion of Small sedans is $\hat{p}$ = 13/54 = 0.25. In other words, 25% of the sample is Small Sedan.

To get a bar graph for one categorical variable, highlight the column with the categorical variable, and then click on Insert tab, then Recommended Charts



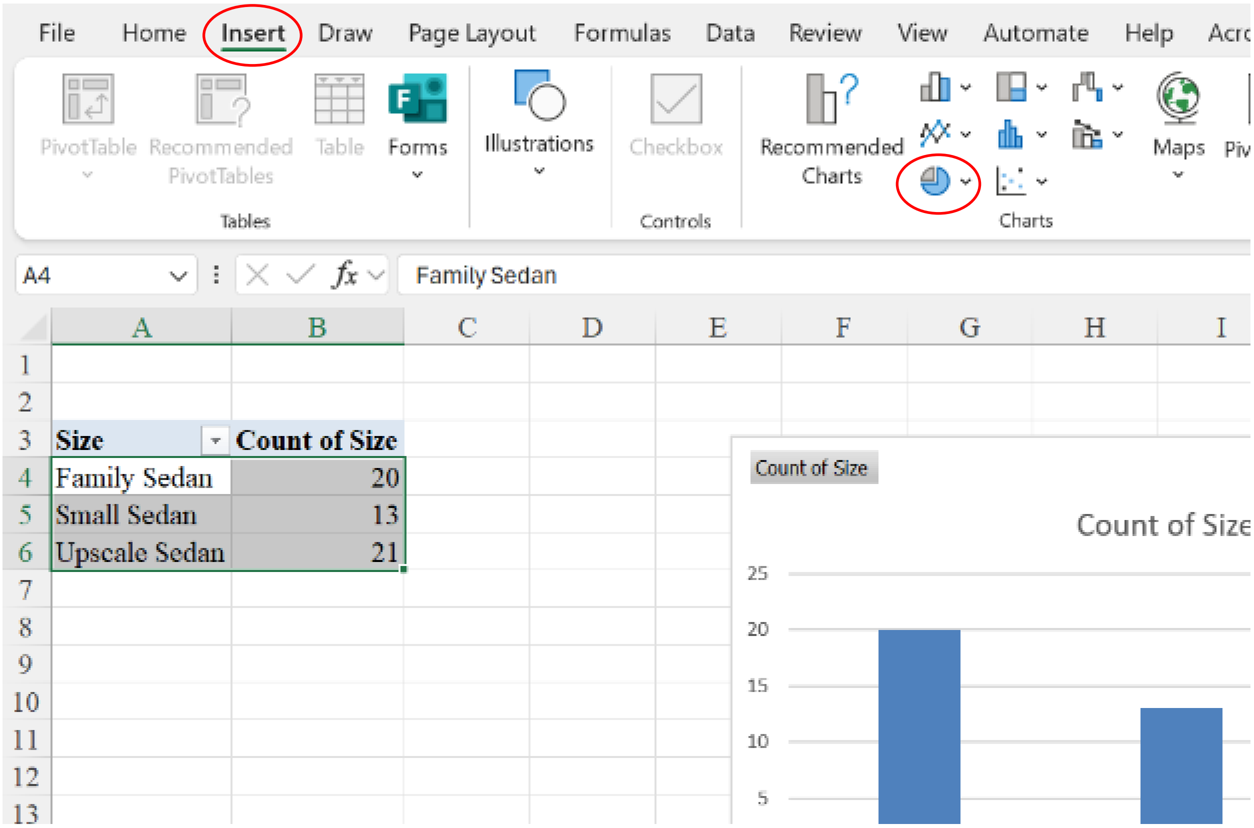
	Car	Size	Price ($)	Cost/Mile	Road-Test Score	Predicted Reliability	Value Score	Price ($)
2	Toyota Corolla (base, manual)	Small Sedan	16,419	0.44	70	4	1.99	16,419
3	Mazda3 i Touring (manual)	Small Sedan	18,895	0.50	74	5	1.94	18,895
4	Toyota Corolla LE	Small Sedan	18,404	0.47	71	4	1.89	18,404
5	Mazda3 i Touring	Small Sedan	19,745	0.52	70	5	1.82	19,745
6	Hyundai Elantra GLS	Small Sedan	18,445	0.53	80	3	1.64	18,445
7	Nissan Sentra 2.0 SL	Small Sedan	20,150	0.57	74	4	1.51	20,150
8	Kia Forte Sedan EX	Small Sedan	19,040	0.57	71	3	1.32	19,040
9	Ford Focus SE	Small Sedan	20,280	0.52	68	2	1.30	20,280
10	Ford Fiesta SE	Small Sedan	16,595	0.47	61	2	1.25	16,595
11	Volkswagen Jetta SE (2.5)	Small Sedan	20,300	0.54	60	3	1.24	20,300
12	Volkswagen Jetta TDI	Small Sedan	25,100	0.50	68	2	1.18	25,100
13	Chevrolet Cruze LS (1.8)	Small Sedan	18,375	0.57	67	1	0.96	18,375
14	Chevrolet Cruze 1LT (1.4T)	Small Sedan	20,530	0.60	69	1	0.91	20,530
15	Nissan Altima 2.5 S (4-cyl.)	Family Sedan	23,970	0.59	91	4	1.75	23,970
16	Kia Optima LX (2.4)	Family Sedan	21,885	0.58	81	4	1.73	21,885
17	Subaru Legacy 2.5i Premium	Family Sedan	23,830	0.59	83	4	1.73	23,830
18	Ford Fusion Hybrid	Family Sedan	32,360	0.63	84	5	1.70	32,360

And then click OK. This will make a bar graph in an new sheet, and provide the same counts as given above. See below for output. The graph can be altered in the usual way.

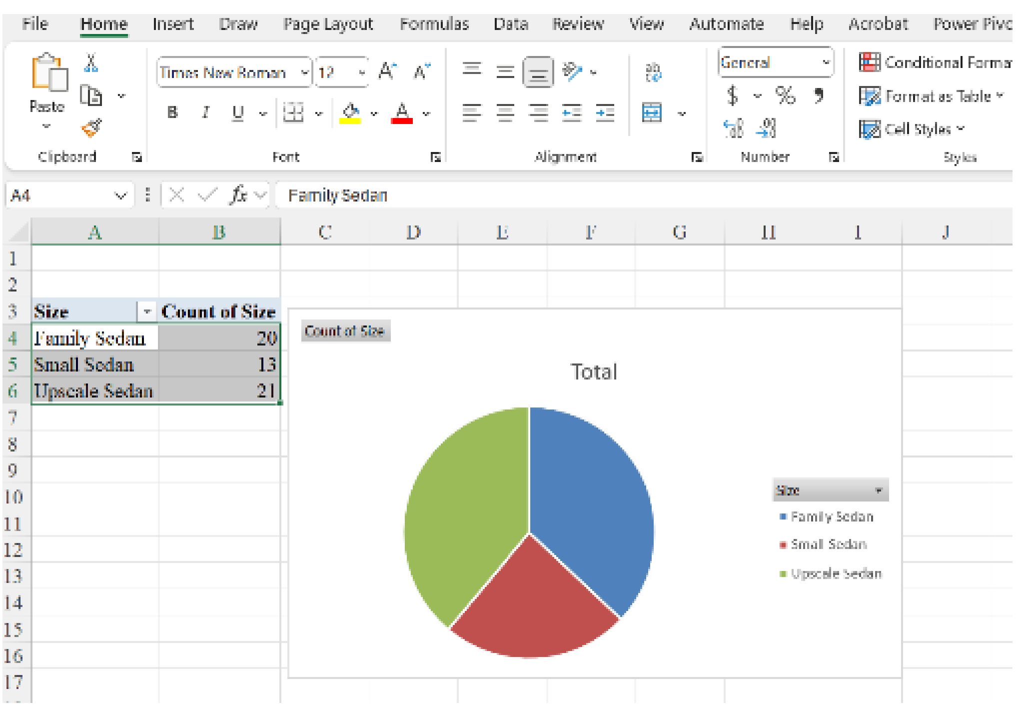


Size	Count of Size
Family Sedan	20
Small Sedan	13
Upscale Sedan	21

To get a pie chart, you have to first summarize the data into frequencies as we did above, or as done by the bar graph in the above image (Counts of 20, 13 and 21). Highlight this table of counts, then click on Insert tab, and then the icon for Pie Charts circled below, and then select the first image under 2-D Pie.



This should produce the following Pie chart, which can be altered in the usual way.



One categorical response variable

- **No explanatory variables**
  - **Inferential statistics**
    - **Confidence interval and margin of error**

Though we don't show it here, you should first check conditions of a confidence interval (CI), which are that both $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ are true, and that the sample was taken randomly from the population.

To get margin of error and confidence interval, we have to use the formulas because they are not built into Excel. Recall that for a one categorical variable problem, the confidence interval formula f or the population proportion is $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and the margin of error is just the part of this formula after the $\pm$, so ME = $z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

We know from above that $\hat{p}$ = 0.25 and *n* = 54. If we use 95% confidence, then z^* = 1.96, and then the CI or ME can be calculated with a calculator, or typed into a cell in Excel.