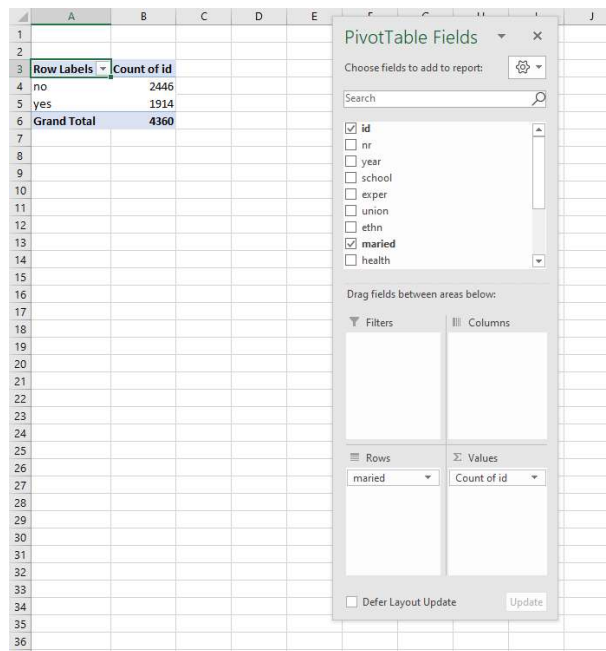


## EXPLORATORY DATA ANALYSIS IN EXCEL– DEMO NOTES

### Frequencies

Create a PivotTable from the source data.

1. Make a frequency table by selecting categories of interest in the Rows/Columns field, then place a Count of the ID field in the Values section.
  - a. To convert a field from a Sum to a Count, double-click on that variable header, and select Count in the “Summarize value field by” menu.



Row Labels	Count of id
no	2446
yes	1914
<b>Grand Total</b>	<b>4360</b>

### Downloading the Analysis ToolPak

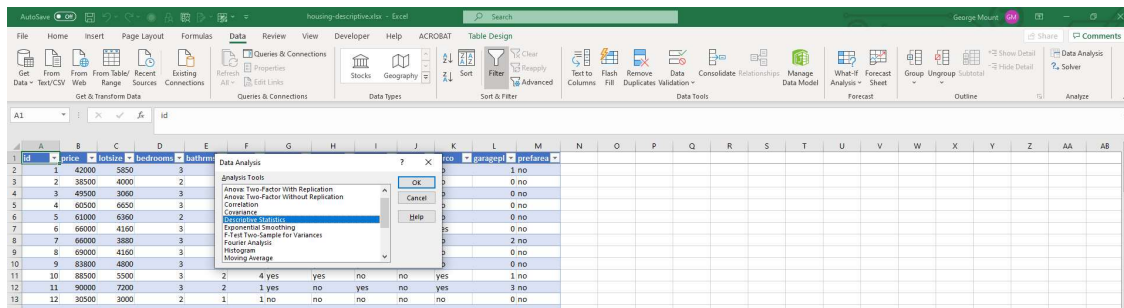
[See instructions from Microsoft here](#). Note the process is different for Windows and Mac.

### Descriptive Statistics

1. Go to the Data tab on the home ribbon.
2. Select Data Analysis from the Analyze group (far right of the menu).



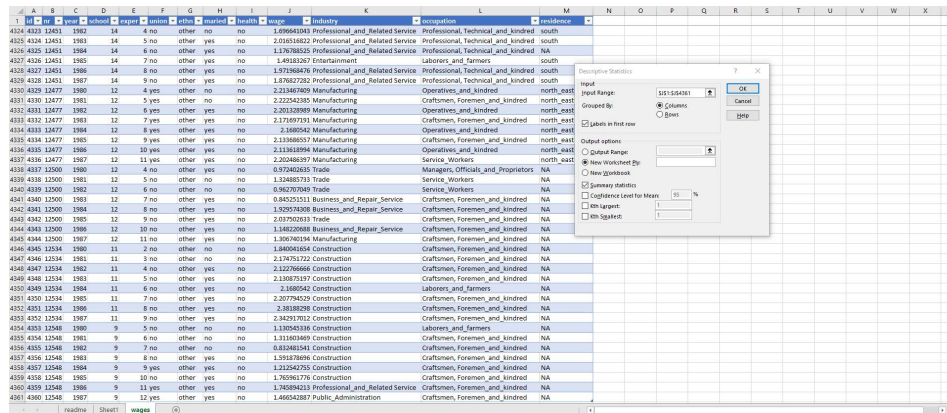
### 3. Select Descriptive Statistics from the menu.



4. Select your Input Range. This will be Column J, year. If your selection includes a header row, make sure to check on the “Labels in First Row” option.

5. By default, the output will be placed in a new worksheet. If you want it elsewhere, click inside “Output Range.” Make sure to double-click inside the dialog box before selecting a new range, otherwise the input range will be re-written.

6. Check on “Summary Statistics.”

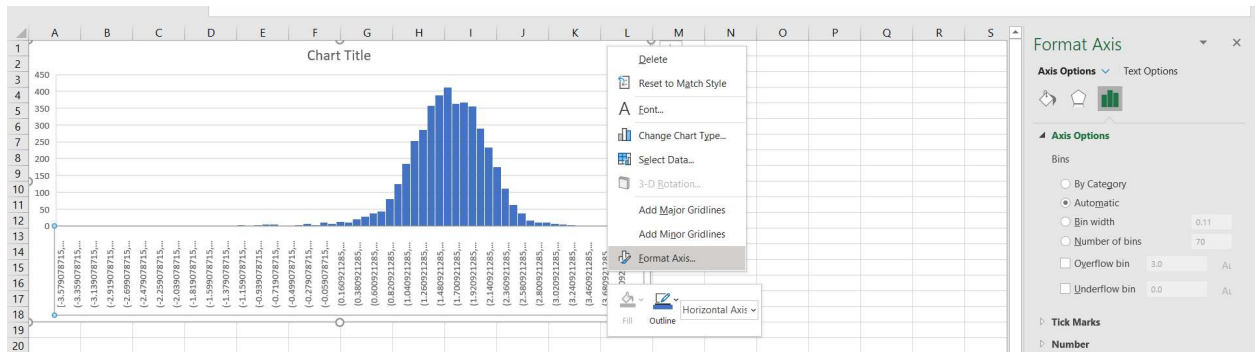


## Histograms

1. Select your input range and go to Insert -> Charts. Histogram should be your third option. Select that. You can cut and paste the resulting histogram elsewhere in the workbook.

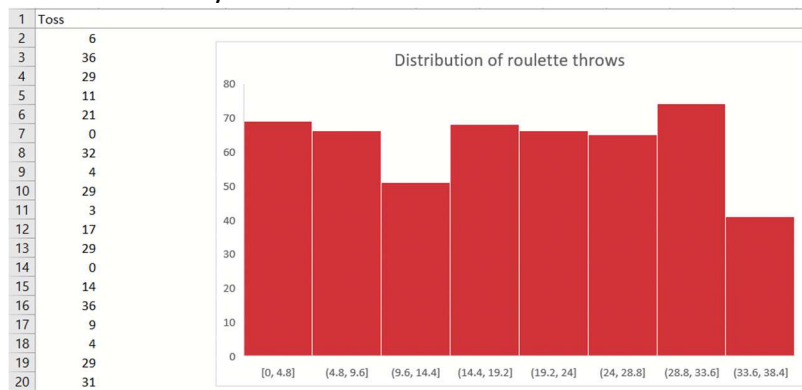
2. To change the number of bins in the histogram, right-click on the X-axis and select Format Axis. You can then customize the X-axis on the side menu. *Note: these features are not available on Excel for Mac.*



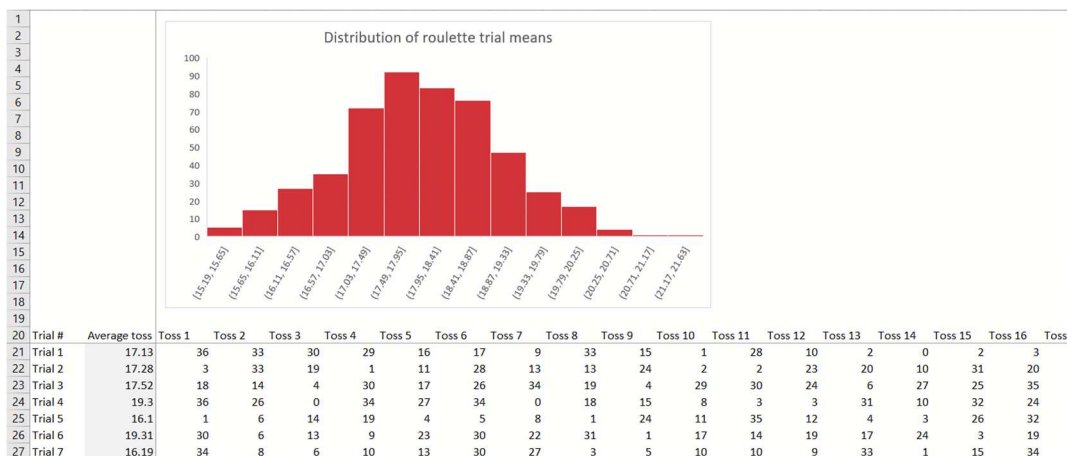


## Central limit theorem

1. Simulate 500 rounds of a roulette spin using `RANDBETWEEN(0,36)`
2. Plot the resulting frequency distribution as a histogram.
3. Use F9 while in your workbook to refresh it.



4. This is a *uniform* distribution.
5. Now simulate a roulette spin 100 times and take the average spin. Do this 500 times and plot the resulting distribution of *sample* means.



6. This time we get a normal distribution, due to the central limit theorem.

### **Law of large numbers: large-numbers.xlsx**

1. Simulate a roulette toss 500 times in Column B: `RANDBETWEEN(0,36)`
2. Take a running total in Column C: `SUM($B$2:B2)`
3. Take a running total in Column D: `C2/A2`
4. Plot Column D as a line chart. Press F9 to recalculate.
  - a. The line converges to the expected mean due to the law of large numbers.

## **FOUNDATIONS OF INFERENTIAL STATISTICS – DEMO NOTES**

### **Descriptive statistics for two categories**

1. Create a PivotTable from the raw data.
2. Create two PivotTables, each displaying the variable you want to measure in the Values, the ID variable in the Rows, and the category in the Filter:



The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable has 'airco' as the Row Label and 'Sum of price' as the Values. The PivotTable Fields task pane is open on the right, showing 'id' and 'price' selected for the Values field, and 'airco' selected for the Row Labels field. The task pane also shows options for 'Filters' and 'Columns'.

3. Remove the Grand Totals for the PivotTables by clicking inside the PivotTable, selecting Design from the home ribbon, then in the Layout group, under Grand Totals, select Off for Rows and Columns.

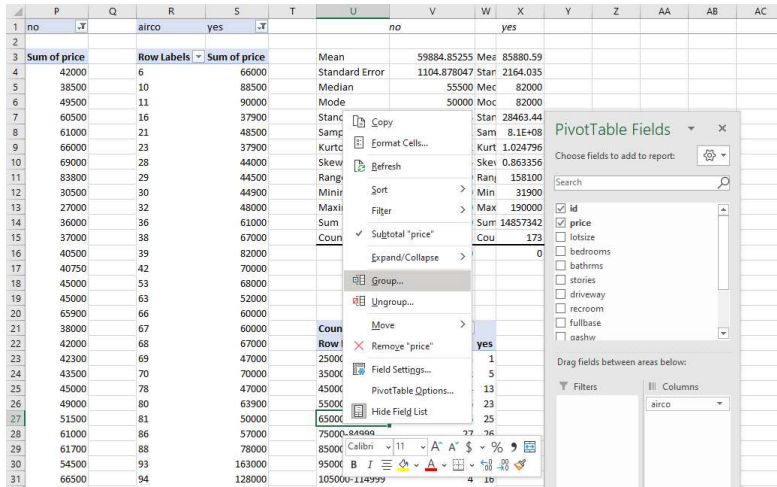
The screenshot shows the Excel Design ribbon for a PivotTable. The 'Grand Totals' dropdown menu is open, showing options for 'Off for Rows and Columns', 'On for Rows and Columns', 'On for Rows Only', and 'On for Columns Only'. The 'Off for Rows and Columns' option is selected.

4. Run the descriptives for each category using the Analysis ToolPak. See demo notes from Section 1 for a refresher.

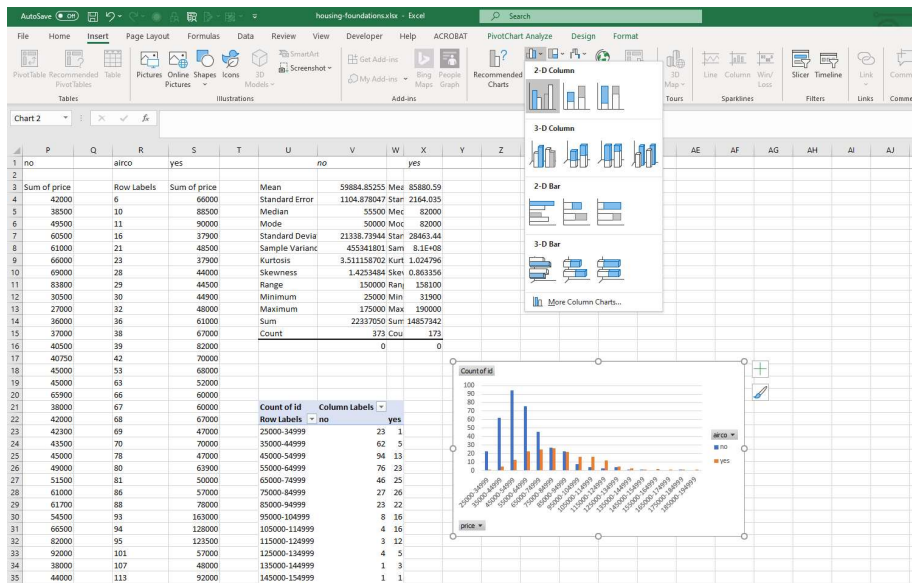
**Plot a histogram for two categories**



1. Create another PivotTable.
2. Place the two categories along the Columns, the continuous variable of interest down the Rows, and the Count of the ID variable in the Values.
3. Right-click the Rows area and select Group.



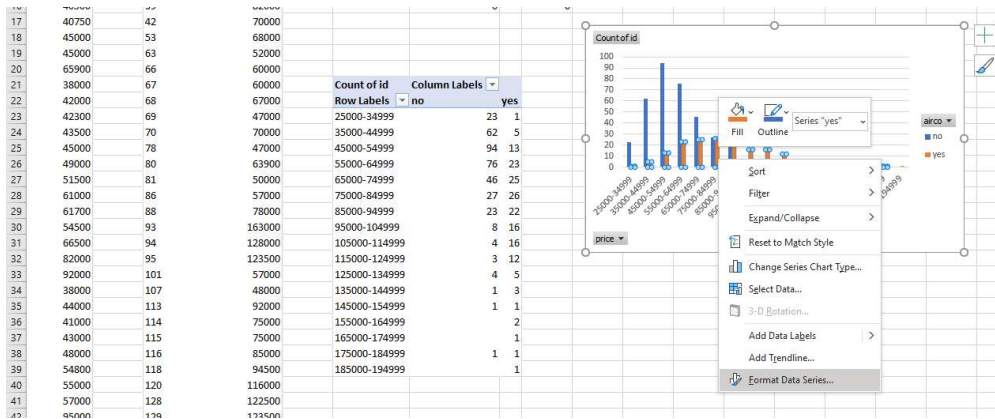
4. Go to Insert on the home ribbon and select a 2-D Column chart.



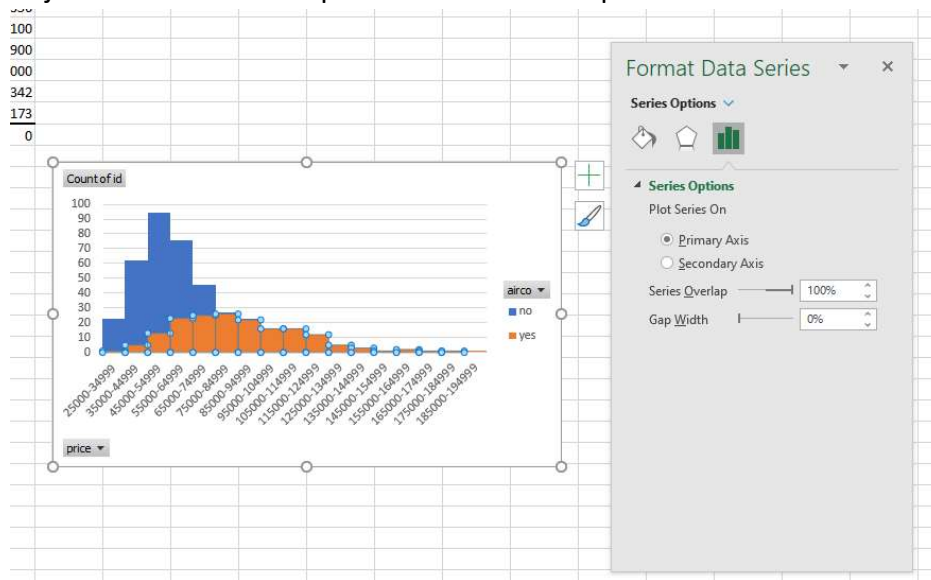
5. Right-click on any of the bars in the resulting bar chart. Select Format Data Series.







6. Adjust the Series Overlap to 100% and the Gap Width to 0%.



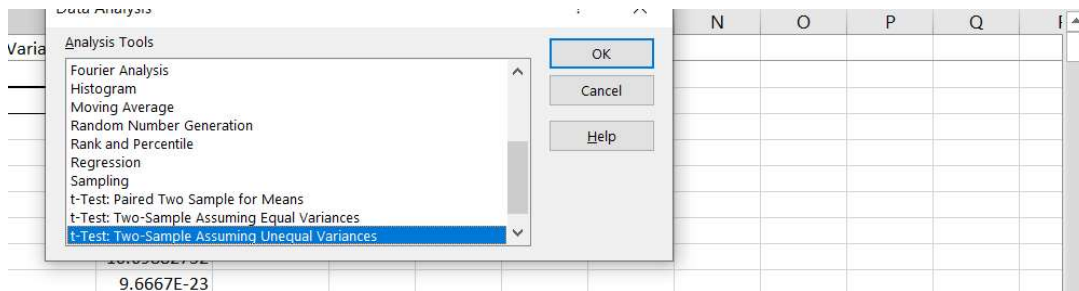
## T-TESTS FOR BUSINESS IMPACT – DEMO NOTES

1. Create a new worksheet, including the PivotTables with the records of the two categories you want to compare.

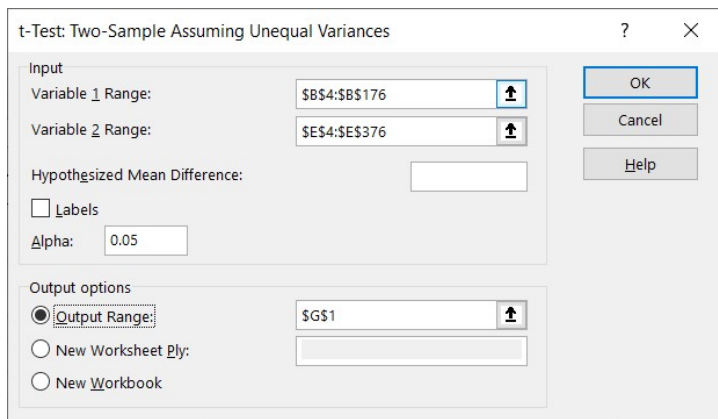
	A	B	C	D	E
1	airco	yes		airco	no
2					
3	Row Labels	Sum of price		Row Labels	Sum of price
4	6	66000		1	42000
5	10	88500		2	38500
6	11	90000		3	49500
7	16	37900		4	60500
8	21	48500		5	61000
9	23	37900		7	66000
10	28	44000		8	69000
11	29	44500		9	83800
12	30	44900		12	30500
13	32	48000		13	27000



- On the ribbon, go to Data -> Data Analysis -> t-Test: Two-Sample Assuming Unequal Variances



- Select your variable ranges, and set the output range to somewhere on the same worksheet.



- This gives you the p-value. Return to slides for explanation of confidence interval.
- To calculate the confidence interval, follow with the formulas used below.





	F	G	H	I	J	K
1		Ho: $\mu_1 - \mu_2 = 0$				
2		Ha: $\mu_1 - \mu_2 \neq 0$				
3		t-Test: Two-Sample Assuming Unequal Variances				
4						
5			yes	no		
6		Mean	85880.5896	59884.85255		
7		Variance	810167352.2	455341801		
8		Observations	173	373		
9		Hypothesized Mean Difference	0			
10		df	265			
11		t Stat	10.69882732			
12		P(T<=t) one-tail	9.6667E-23			
13		t Critical one-tail	1.650623976			
14		P(T<=t) two-tail	1.93334E-22			
15		t Critical two-tail	1.968956281			
16						
17		total sample size	546	=SUM(H8:I8)		
18		mean difference	25995.73705	=H6-I6		
19		standard error of difference	2429.774429	=SQRT((H7/H8)+(I7/I8))		
20						
21		Margin of error	4784.119625	=H19*H15		
22		Lower limit	21211.61742	=H18-H21		
23		Upper limit	30779.85667	=H18+H21		
24						

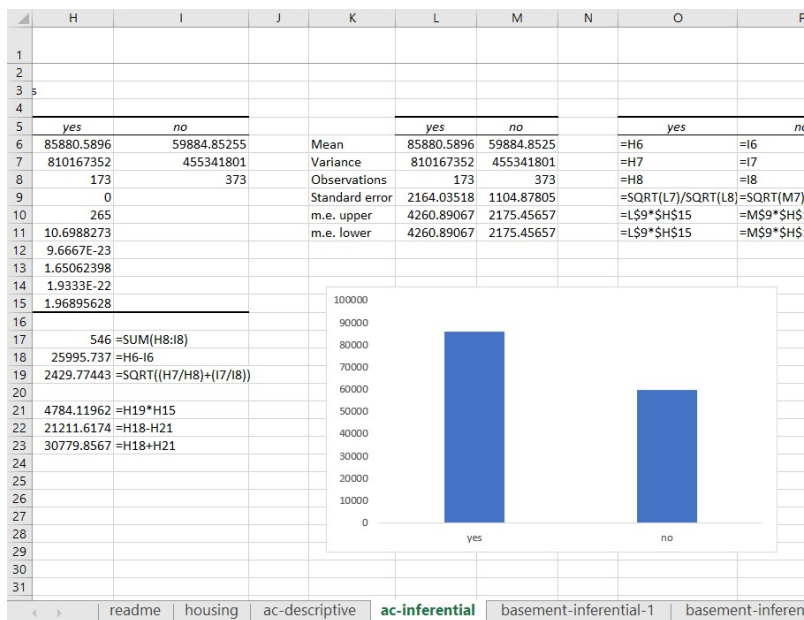
Return to slides for explanation of visualizing t-test results

6. To visualize t-test results, first set up the below formulas.

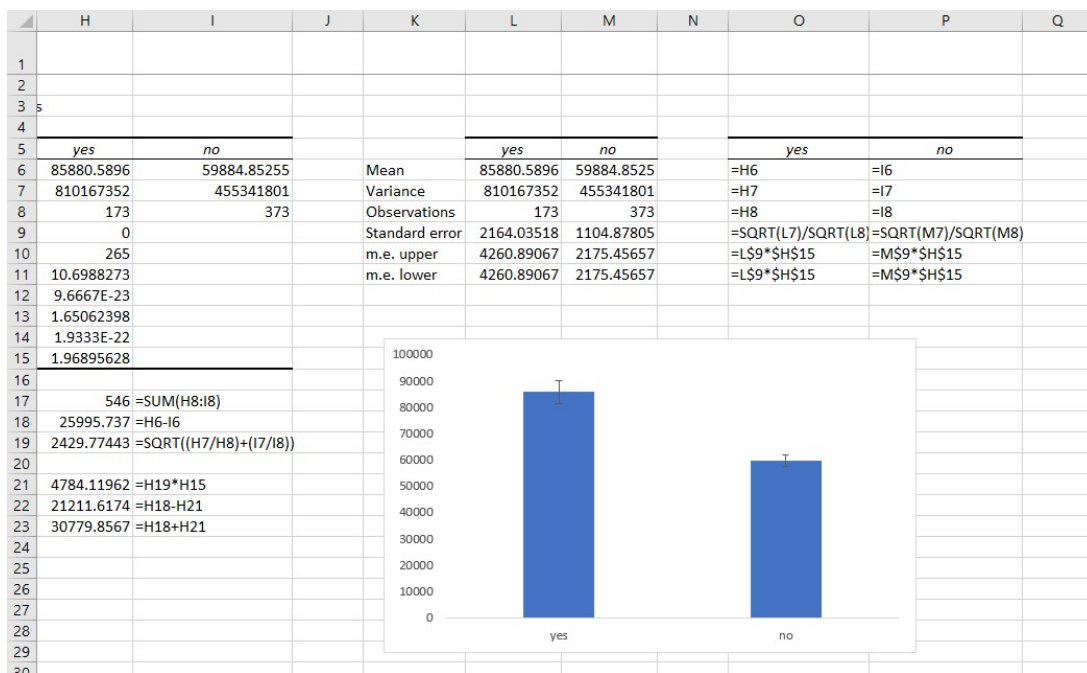
	J	K	L	M	N	O	P	Q
1								
2								
3								
4								
5			yes	no		yes	no	
6		Mean	85880.5896	59884.85255		=H6	=I6	
7		Variance	810167352.2	455341801		=H7	=I7	
8		Observations	173	373		=H8	=I8	
9		Standard error	2164.035184	1104.878047		=SQRT(L7)/SQRT(L8)	=SQRT(M7)/SQRT(M8)	
10		m.e. upper	4260.890669	2175.456571		=L\$9*\$H\$15	=M\$9*\$H\$15	
11		m.e. lower	4260.890669	2175.456571		=L\$9*\$H\$15	=M\$9*\$H\$15	
12								
13								
14								

7. Create a bar chart based on the means of each category





8. Click on the plus-sign next to the bar chart and select Error Bars, hit the right arrow next to it and select More Options.
9. Go to Custom and the bottom and set the error bars to be the margin of error values.
10. The bar chart now has error bars representing the 95% confidence interval for each sample. If the bars intersect between the two charts, there is no significant difference in means.



11. Adjust the y axis depending on the circumstances, with the knowledge that it is the best practice to start a y axis at zero.



**Demo: margin-of-error.xlsx**

Column position	Column label	Formula
C	Sample mean	=AVERAGE(\$B\$3:B4)
D	Variance	=VAR.S(\$B\$3:B4)
E	Standard Error	=SQRT(D4)/SQRT(A4)
F	Critical value	=VLOOKUP(\$A4,'critical-value'!\$A\$1:\$B\$34,2)
G	Margin of error	=E4*F4
H	Margin of error as % of mean	=G4/C4

Plot column H as a line chart.

