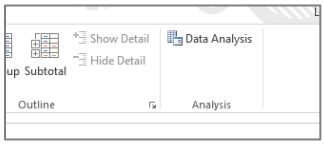
Basic Statistical Analysis in Excel

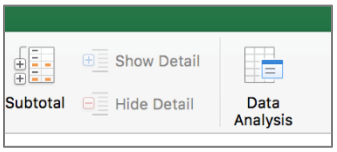
NICAR 2022 Atlanta | Norm Lewis, University of Florida | nplewis@ufl.edu

PART 1: ENSURE ANALYSIS TOOLPAK IS ENABLED ON YOUR COMPUTER



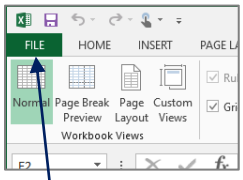
Windows

Microsoft considers Analysis TookPak an "add-in" feature. It comes with Excel (for Windows and for the latest Mac version) but you must enable it first. Check to see if it is loaded by clicking on the Data tab on the ribbon. If yours does not look like one of these examples here, follow steps below.

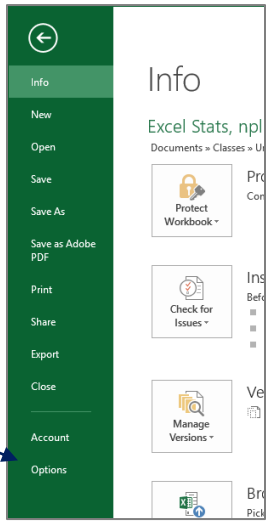


Apple

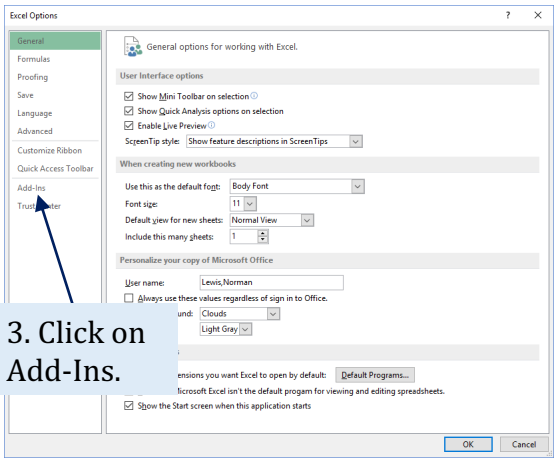
For Windows: Enabling Analysis ToolPak



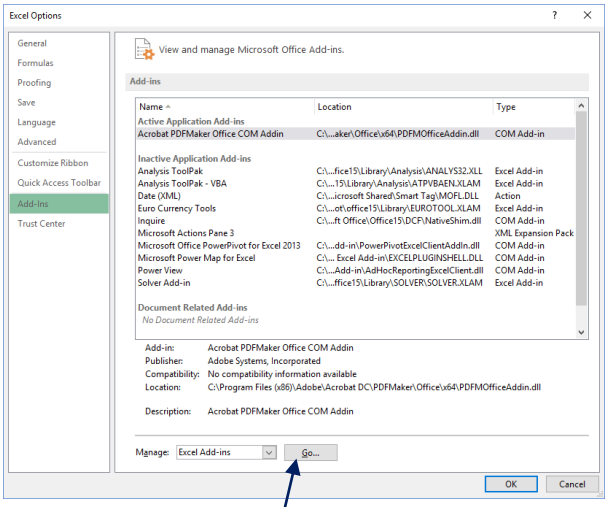
1. Click on the File tab on the ribbon.



2. Click on Options.

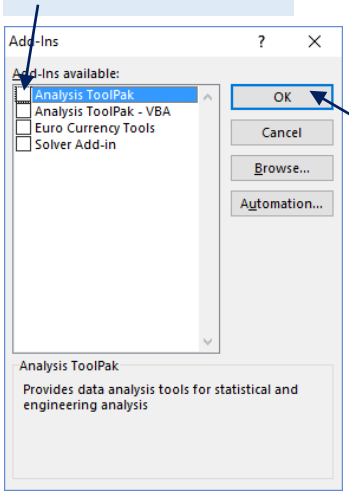


3. Click on Add-Ins.



4. Click Go ...

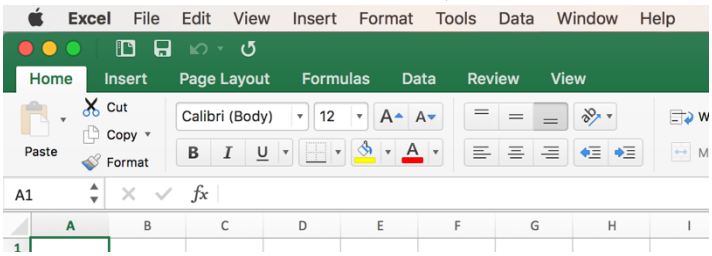
5. Click box for Analysis ToolPak.

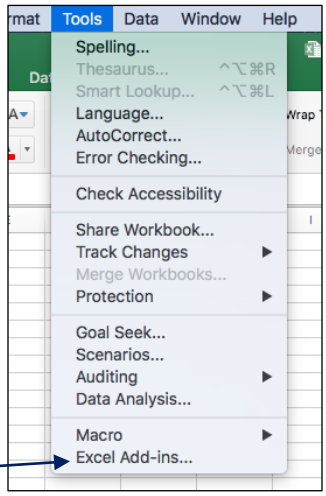


6. Click OK.

For Macintosh: Installing Analysis ToolPak
Only available in Office 365 subscription version.

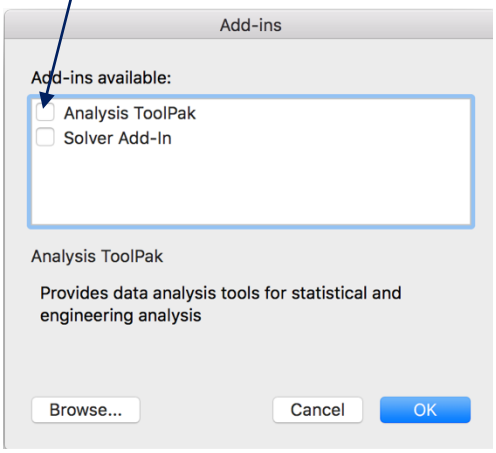
1. Click on the Tools menu above the ribbon.

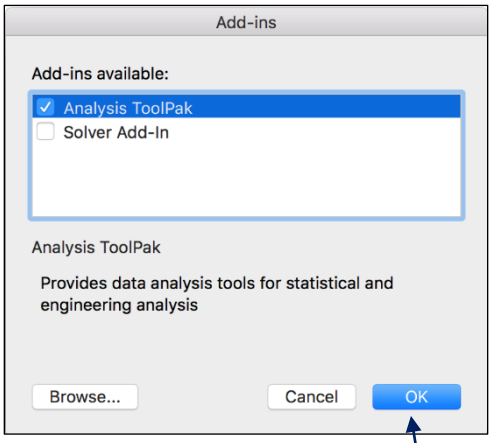




2. Select Excel Add-Ins…

3. Click on Analysis ToolPak.



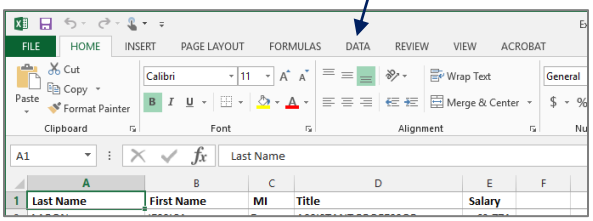


4. Click OK.

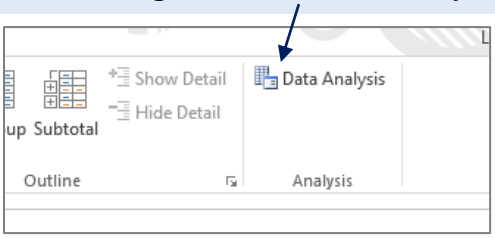
PART 2: DESCRIPTIVE STATISTICS

Choose Faculty worksheet.

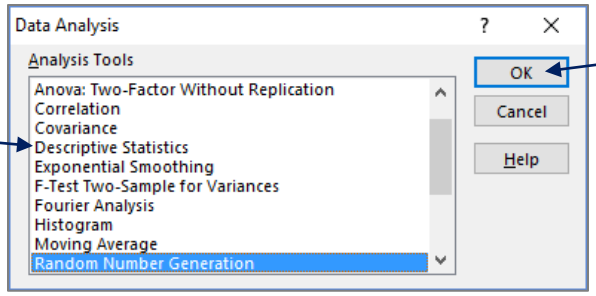
1. Click on Data tab on the ribbon.



2. On the right, click on Data Analysis.

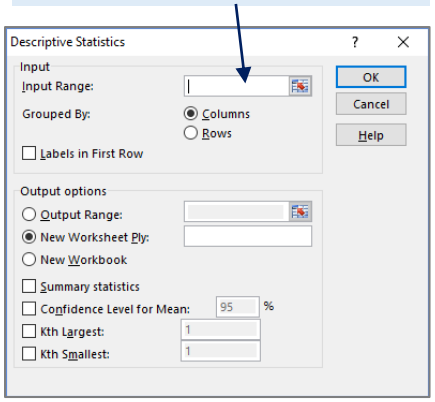


3. Click on Descriptive Statistics.

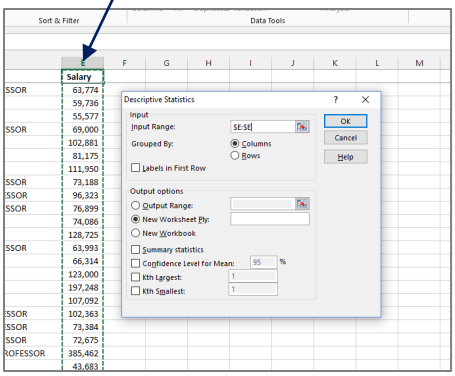


4. Click OK.

5. Click in Input Range box.

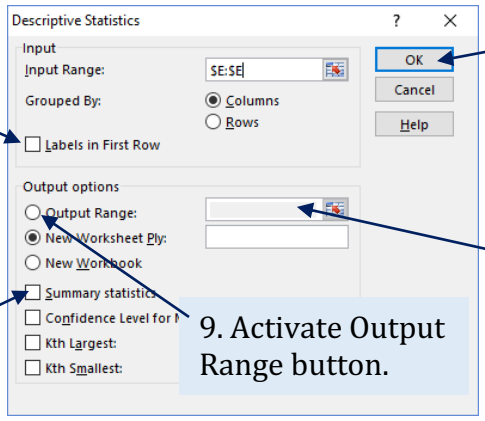


6. Click on the Salary column heading.



7. Click in the box for Labels in First Row.

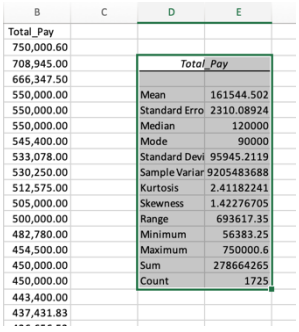
8. Click in the box beside Summary statistics.



9. Activate Output Range button.

10. In the Output Range box, type G3 or click on the cell where you want the stats inserted.

11. Click OK.



Total_Pay
750,000.60
708,945.00
666,347.50
550,000.00
550,000.00
550,000.00
545,400.00
533,078.00
530,250.00
512,575.00
505,000.00
500,000.00
482,780.00
454,500.00
450,000.00
450,000.00
443,400.00
437,431.83

Total_Pay	
Mean	161544.502
Standard Erro	2310.08924
Median	120000
Mode	90000
Standard Devi	95945.2119
Sample Variar	9205483688
Kurtosis	2.41182241
Skewness	1.42276705
Range	693617.35
Minimum	56383.25
Maximum	750000.6
Sum	278664265
Count	1725

12. Adjust the columns for readability and to line up the decimal points.

Total_Pay	
Mean	161,545
Standard Error	2,310
Median	120,000
Mode	90,000
Standard Deviation	95,945
Sample Variance	9,205,483,688
Kurtosis	2
Skewness	1
Range	693,617
Minimum	56,383
Maximum	750,001
Sum	278,664,265
Count	1,725

Some key numbers from this output, a 2021 list of assistant professors (untenured, entry level) at the University of Florida.

Use **mean** for numbers that vary little: sports stats, commute times, water levels, etc.

All three averages are provided.

Use **median** for numbers with potential outliers: salaries, income, wealth, home values, etc.

±1 SD around mean includes 66% of salaries. Here: $65,600 to $257,490.

The **mode** is the most frequently occurring number; it is rarely used in journalism.

The min and max are the range.

Total_Pay	
Mean	161,545
Standard Error	2,310
Median	120,000
Mode	90,000
Standard Deviation	95,945
Sample Variance	9,205,483,688
Kurtosis	2
Skewness	1
Range	693,617
Minimum	56,383
Maximum	750,001
Sum	278,664,265
Count	1,725

For this data set, the best average is the median. Because most people think "average" can only be a mean, phrase it this way:

> The median, or midpoint, was $120,00.

PART 3: BRIEF STATISTICS PRIMER

1. Statistics help us sift meaningful patterns from random chance. Random fluctuations in data are normal. Statisticians call this *noise*. For example, we want to know if data showing racial disparities are indicative of a meaningful pattern or just noise.

2. Statistics is the science of probability. Probability is not certainty. All a statistic can tell you is whether a relationship could have happened by chance.

3. Statistics are not voodoo. They are based on empirical testing and numerical laws such as the central limit theorem. It is, after all, a *science*.

4. On the other hand, statistics are not magic. They require interpretation to avoid false interpretations or ascribe undue importance.

5. To avoid common interpretation mistakes, remember these three principles:
    a. **Correlation does not imply causation.** People who live in larger cities are more likely to consider the arts important. Is that because:
        i. Larger cities with more facilities stimulate more interest in the arts, or
        ii. People with greater arts interest tend to locate in larger cities?
    b. **Consider a plausible alternative explanation**. An increase in antidepressants among older adults means lower suicide rates. However, that may be partly due to improvements in drugs that have reduced death risk from overdoses.
    c. **Beware the law of small numbers**. Rural U.S. states have both the highest and lowest rates of cancers. This is not a meaningful pattern but the result of the greater variance inherent when population numbers are relatively small.

6. Because human behavior is so complex, social science sets the level of probability (p) to separate patterns from noise at less than 5%, or $p < .05$. Colloquially, $p < .05$ means there was a less than 5% probability a relationship was due to chance.

7. Although $p < .05$ is considered statistically *significant*, a more precise term would be statistically *noticeable* or statistically *detectable* (Jordan Ellenberg, p. 121). Whether that relationship is *meaningful* depends on human judgment – yes, you!

8. If you want to read more about statistics, I recommend:
    a. "Naked Statistics: Stripping the Dread from the Data" by Charles Wheelan
    b. "Statistics for People Who (Think They) Hate Statistics" by Neil J. Salkind
    c. "How Not to Be Wrong" by Jordan Ellenberg
    d. "Statistics Unplugged" by Sally Caldwell

PART 4: CORRELATION

Correlation 1: Crime

Question: How closely is crime correlated with population?

Correlation measures if two variables are related.

Two types of correlation:

- Positive: Either:
  - Both rise together, like more height and more weight.
  - Both fall together, like less physical activity and less life expectancy.
- Negative: One increases while other falls, like more beers & less GPA.

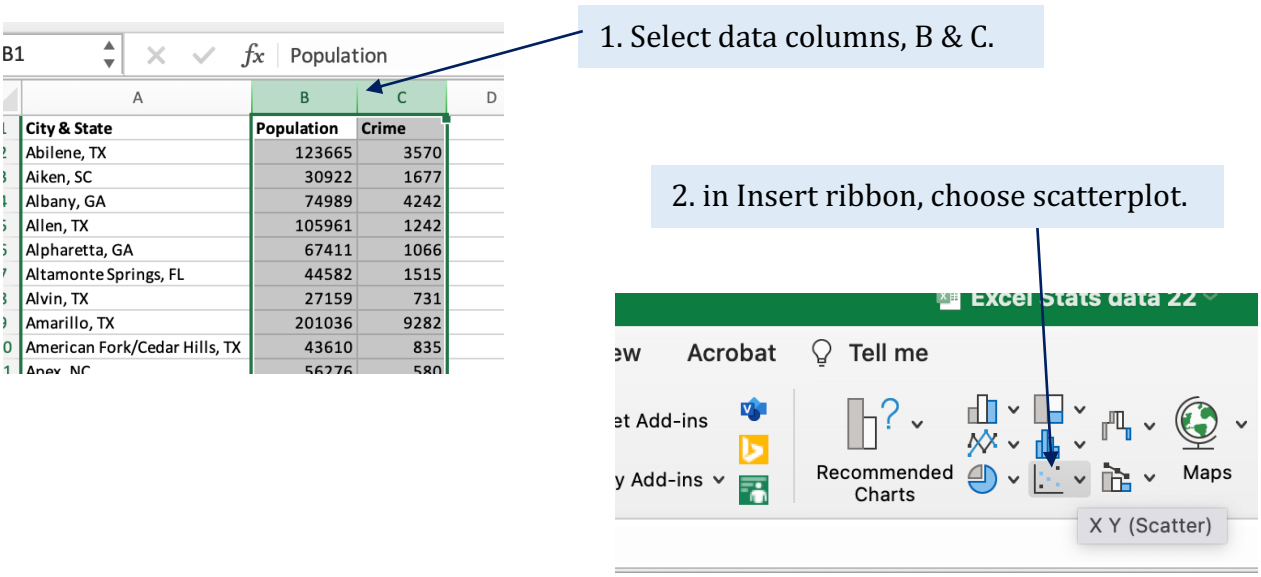
Correlation may, or may not, be causation.

Correlation measures

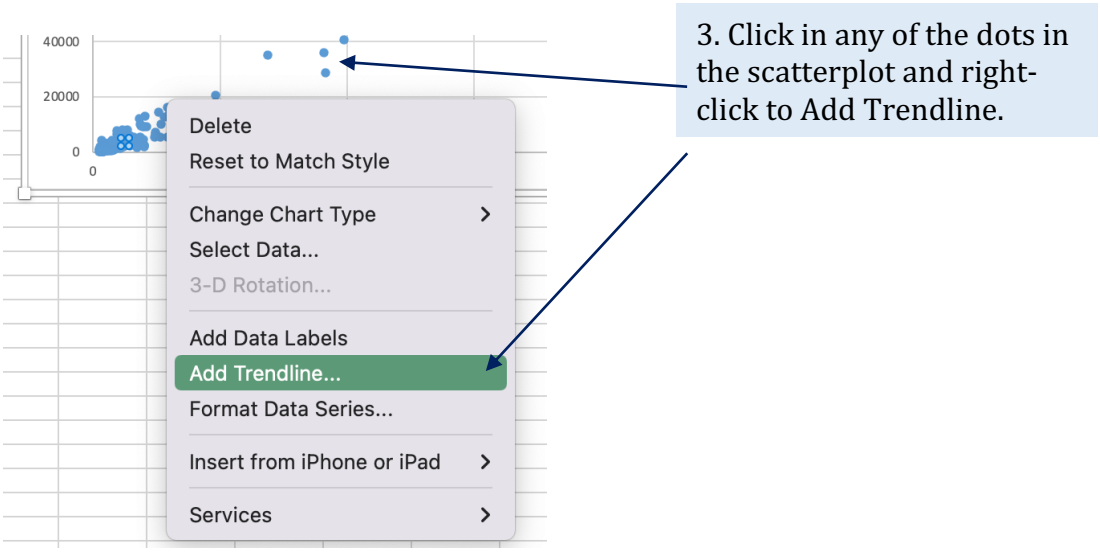
- Correlations range from 0.00 (no relationship) to 1.00 (perfect relationship).
- The correlation coefficient measures the strength of the relationship.
- A general rule of thumb (Cohen, 1988) for the correlation coefficient is:
  - 0.10 to 0.28 Small
  - 0.30 to 0.49 Moderate
  - 0.50 to 0.99 Large

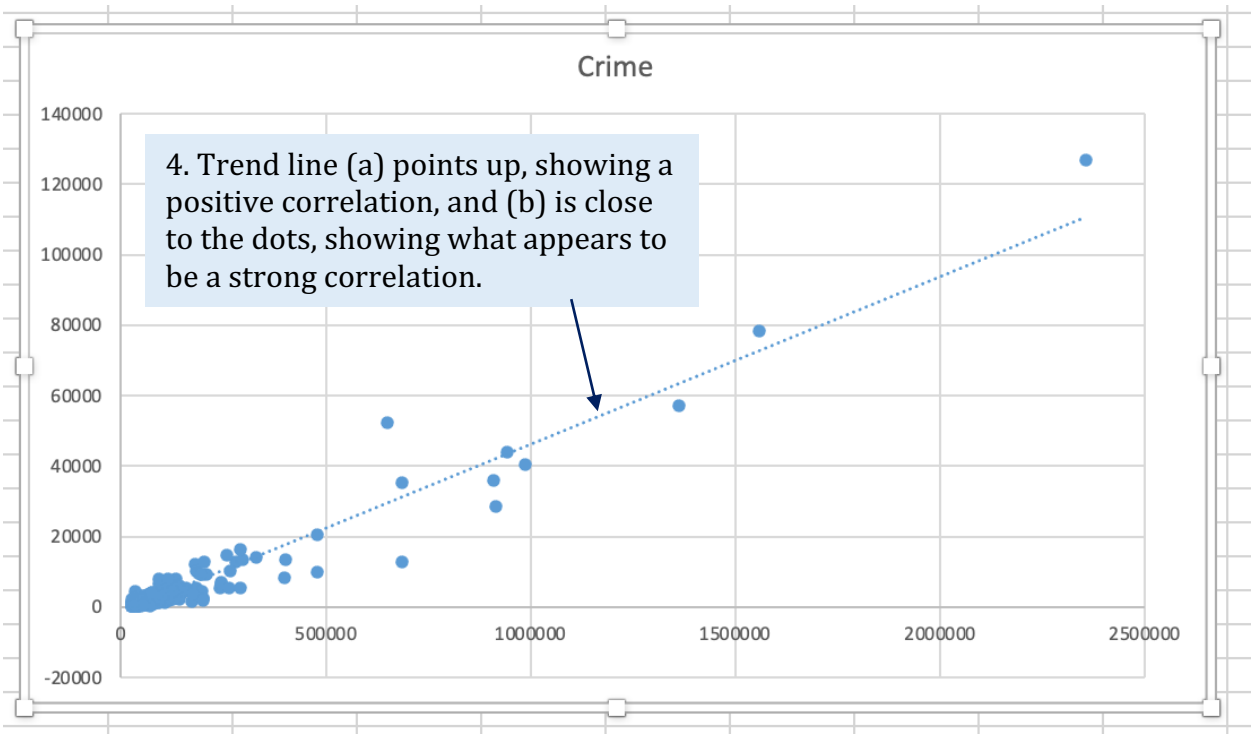
Open Crime data. These are 2019 data for cities of at least 25,000 in Georgia and nearby states, excluding Alabama (whose data are unreliable, FBI says).

First, examine a scatterplot.

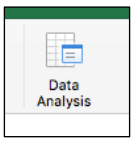


City & State	Population	Crime
Abilene, TX	123665	3570
Aiken, SC	30922	1677
Albany, GA	74989	4242
Allen, TX	105961	1242
Alpharetta, GA	67411	1066
Altamonte Springs, FL	44582	1515
Alvin, TX	27159	731
Amarillo, TX	201036	9282
American Fork/Cedar Hills, TX	43610	835

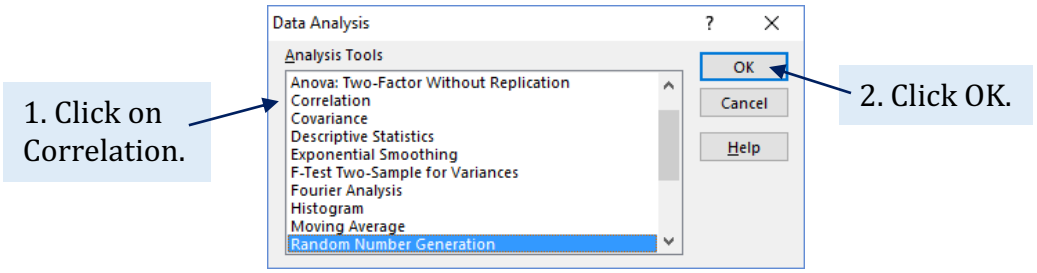


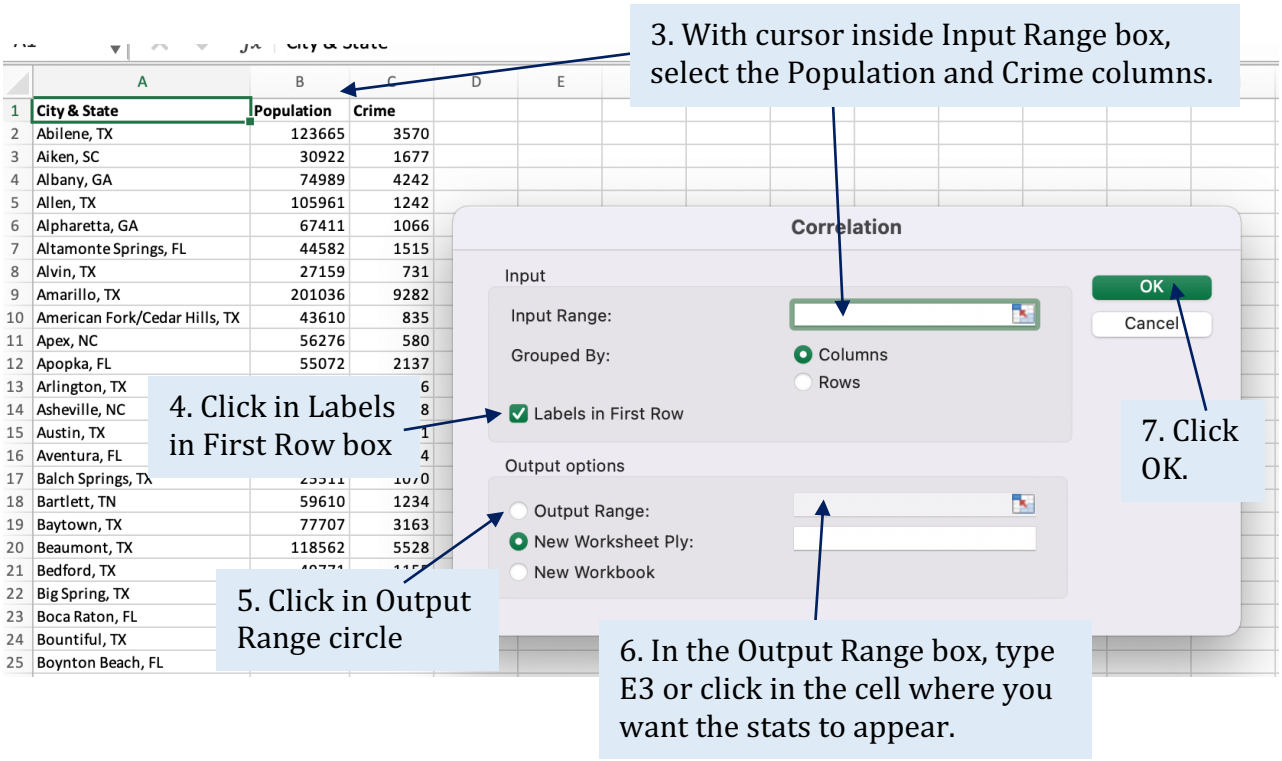


Next, measure strength of correlation.



Open the Analysis ToolPak by clicking on the Data Analysis button at the far right of the Data ribbon.





The statistic looks like this.

	Population	*Crime*
Population	1	
Crime	0.96381	1

The correlation: 0.96

Interpretation
Population and crime are almost perfectly correlated: More people = more crime.

The strong correlation does not equal uniformity. Some cities are above the line (higher crime rates) and some are below.

Also, do not treat this strong correlation as even remotely normal. In my research, 0.40 is about as strong as it gets.

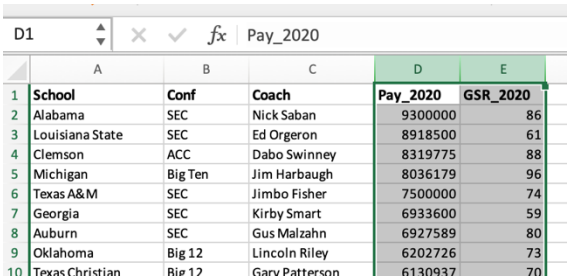
Correlation 2: Coaches

Question: Are college football coaching salaries and player graduation rates correlated?

Click on the Coaches sheet. These are college football coaches. Sources:

- Pay_2020: Total salaries paid, USA Today database.
- GSR_2020: Percent of 2014 football players who graduated within 6 years, NCAA.

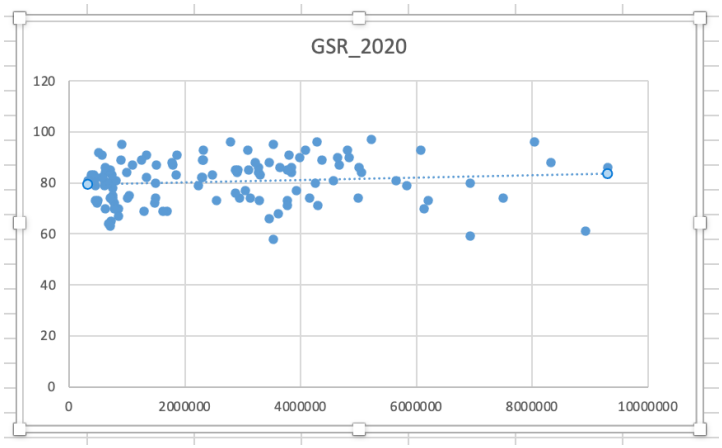
First, scatterplot.



School	Conf	Coach	Pay_2020	GSR_2020
Alabama	SEC	Nick Saban	9300000	86
Louisiana State	SEC	Ed Orgeron	8918500	61
Clemson	ACC	Dabo Swinney	8319775	88
Michigan	Big Ten	Jim Harbaugh	8036179	96
Texas A&M	SEC	Jimbo Fisher	7500000	74
Georgia	SEC	Kirby Smart	6933600	59
Auburn	SEC	Gus Malzahn	6927589	80
Oklahoma	Big 12	Lincoln Riley	6202726	73
Texas Christian	Big 12	Gary Patterson	6130937	70

Select the Pay_2020 and GSR_2020 columns.

In Insert ribbon, select the Scatterplot, as we did for the Crime sheet.

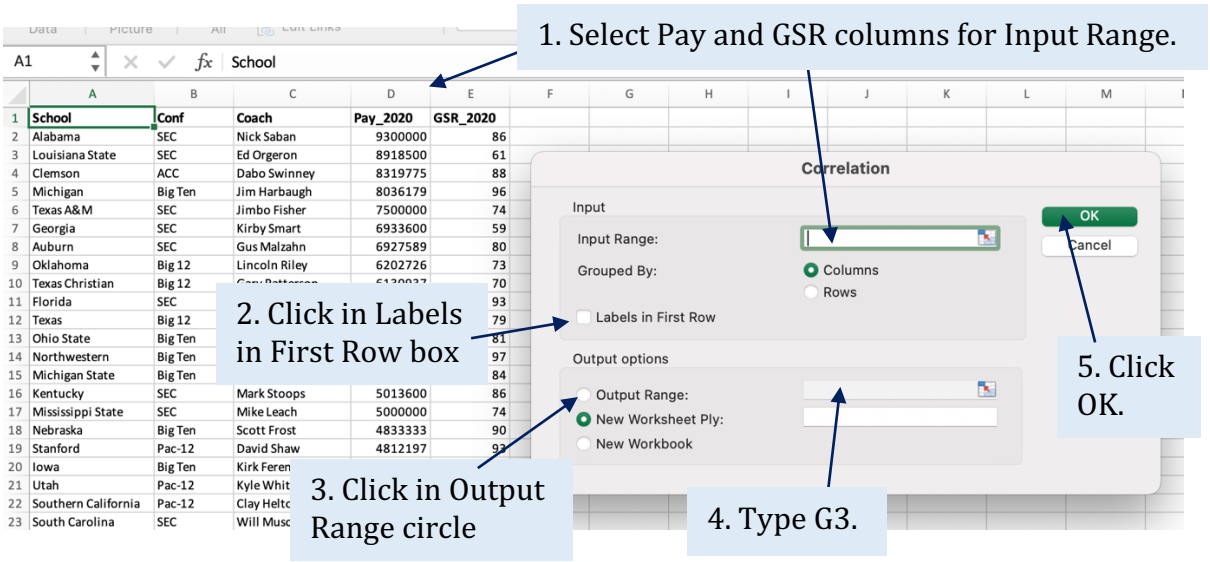


Interpretation:

First, GSR rates vary little. Most are between 70% and 85%.

Second, trend line is flat. This means that higher salaries make little difference in GSR rates.

Second, calculate correlation



	Pay_2020	*GSR_2020*
Pay_2020	1	
GSR_2020	0.11073104	1

The correlation: 0.11

Interpretation
A weak correlation with a flat line = a non-existent correlation.

Salaries paid college football coaches and the graduation rates of their players are not correlated.

Two possible explanations for the weak correlation.

1. A third and more dominant variable may be at work: money.

   Schools that pay big salaries to coaches also can afford larger academic counseling staffs who keep players on track to graduate. The presence of a third, more influential variable, is reason to consider a plausible alternative explanation.

2. The result may reflect a data definition that produces relative conformity.

   The NCAA created the GSR, or Graduation Success Rate, to be more generous than the federal data standard. The GSR allows universities to ignore athletes who transferred or who turned pro before graduation. Counting only players who stayed until the end of their athletic careers compresses potential variance.

One final correlation …

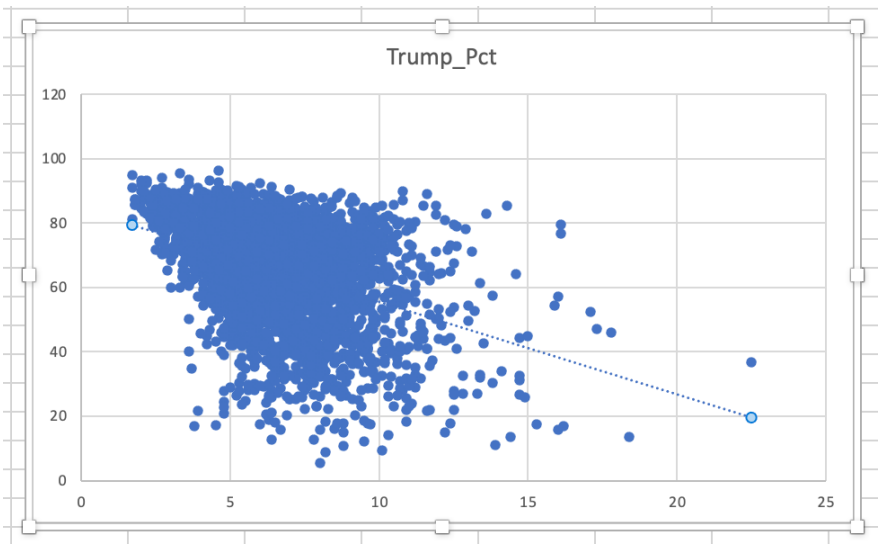
Correlation 3: Election

Question: Was the share of votes Trump received and unemployment correlated?

Sources:

- Jobless_Pct: Bureau of Labor Statistics county averages for 2020
- Trump_Pct: MIT 2020 elections project, supplemented with county elections data

Open Elections sheet. Select the two data columns and create a scatterplot with trend line.



Not what was hypothesized.

The trend line is down. Thus, this is a negative correlation. As unemployment (x-axis) increases, Trump's vote share (y-axis) decreases.

Next, use the Analysis ToolPak to calculate the correlation coefficient.

	Jobless_Pct	*Trump_Pct*
Jobless_Pct	1	
Trump_Pct	-0.3975828	1

The correlation: -0.40

Interpretation
Negative sign = negative correlation
Coefficient: -0.40 = moderate strength

Yes, Trump support and unemployment are moderately correlated. However, an assertion that greater unemployment is associated with lower voter support lacks a rationale.

Instead, this statistic is an example of how statistics must be read carefully. Despite the moderate strength of the relationship, unemployment was not a primary factor in vote share. Something else was.

PART 5: T-TEST

T-Test 1: Broadway

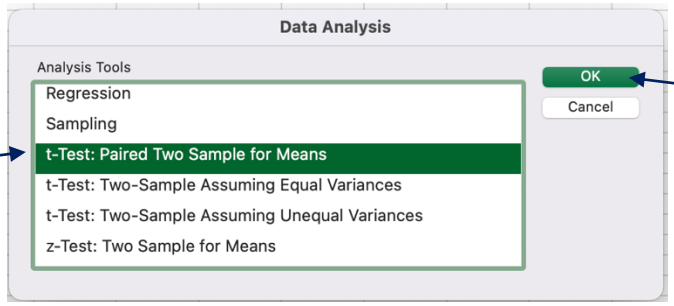
Question: Were "Hamilton" tickets substantially more expensive than for other big musicals?

The data are from the Internet Broadway Database. Data for "Hamilton" are compared to averages for four other long-running musicals: "Aladdin," "Lion King," "Book of Mormon," and "Wicked." Weekly means for those four are combined in the Big4Avg.

Which statistic to use?

- Comparing two means, so t-Test
- Data are "paired" by week, so a paired, two-sample test

1. In Data Analysis, select t-Test: Paired Two Sample for Means

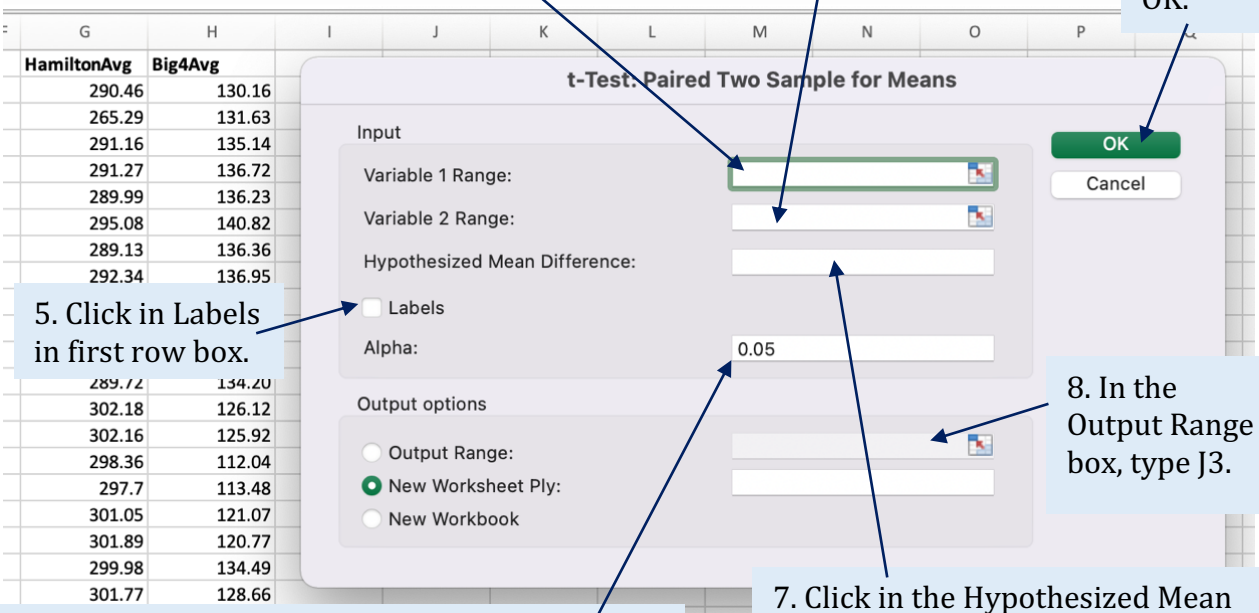


2. Click OK.

3. With cursor inside Variable 1 Range box, click on Hamilton Avg (does not matter which one)

4. With cursor inside Variable 2 Range box, choose Big4Avg

9. Click OK.



5. Click in Labels in first row box.

8. In the Output Range box, type J3.

7. Click in the Hypothesized Mean Difference box and type 0 (zero).

6. Leave Alpha at the default, 0.05, the standard level for determining statistical significance.

Mean ticket price for "Hamilton": $294.64

Mean ticket price for other 4 musicals: $130.88

t-Test: Paired Two Sample for Means		
	HamiltonAvg	*Big4Avg*
Mean	294.6375	130.879214
Variance	366.606803	245.474958
Observations	52	52
Pearson Correlation	0.64756416	
Hypothesized Mean Difference	0	
df	51	
t Stat	78.9785296	
P(T<=t) one-tail	2.6705E-55	
t Critical one-tail	1.67528495	
P(T<=t) two-tail	5.341E-55	
t Critical two-tail	2.00758377	

The t-statistic. We could look up the significance level in a stats book … or let Excel do it for us.

We hypothesized a direction ("Hamilton" more costly), so use the one-tailed result.

The p-value is < .05. Thus, reject the no-difference hypothesis and conclude difference is statistically significant.

Well, d'oh! We did not need a fancy statistics test to discover that "Hamilton" tickets cost more – in fact, twice what the other four big musicals did.

But We Are Not Throwing Away Our Shot, so do another test …

T-Test 2: Firefighters

Question: Is pay for firefighters in Bloomington, Ind., different from the national average?

Sources:

- Salaries for front-line firefighters for 2017, via Data.gov.
- National firefighter pay mean, Bureau of Labor Statistics for 2017.

This hypothesis has no direction (could be lower or higher), so use the two-tailed test.

Repeat the same steps above, with a paired-sample t-test.

Bloomington mean: $58,283 National mean: $51,930

t-Test: Paired Two Sample for Means		
	Salary	*National*
Mean	58282.8214	51930
Variance	9358522.57	0
Observations	49	49
Pearson Correlation	#DIV/0!	
Hypothesized Mean Difference	0	
df	48	
t Stat	14.5365398	
P(T<=t) one-tail	1.6543E-19	
t Critical one-tail	1.6772242	
P(T<=t) two-tail	3.3086E-19	
t Critical two-tail	2.01063476	

No direction hypothesized, so use the one-tailed result.

The p-value is < .05. Thus, reject the no-difference hypothesis and conclude difference is statistically significant.

Difference in firefighter pay, 11 percent, is not small, but not large, either.

Therefore, the finding of statistical significance offers assurance the difference is beyond what would be expected by chance. You are free to write a story:

- Bloomington firefighters are paid above the national average.

Please note The statistic does not give license to use "significantly" in that sentence. All the t-Test can tell us is that the difference was unlikely to occur by chance.