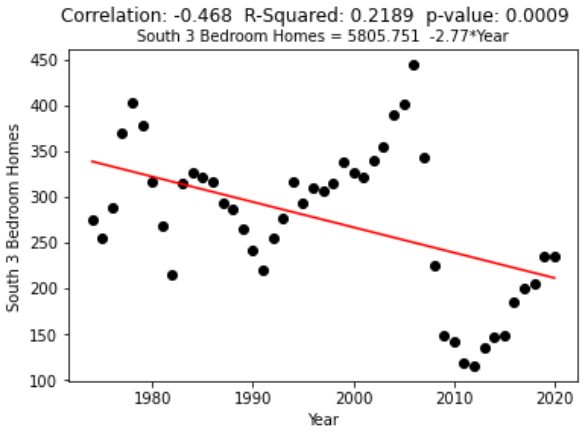
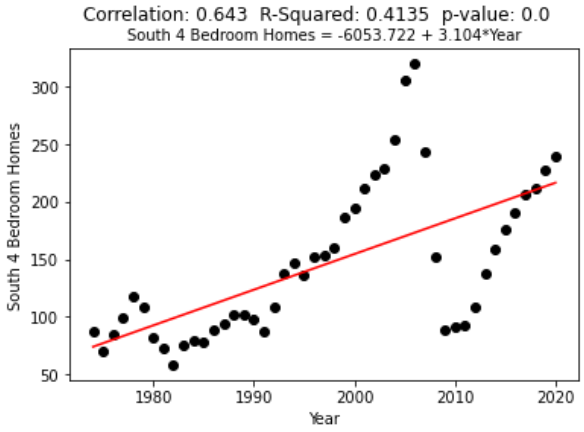
INFO 3100 FINAL PROJECT
Deja Brule

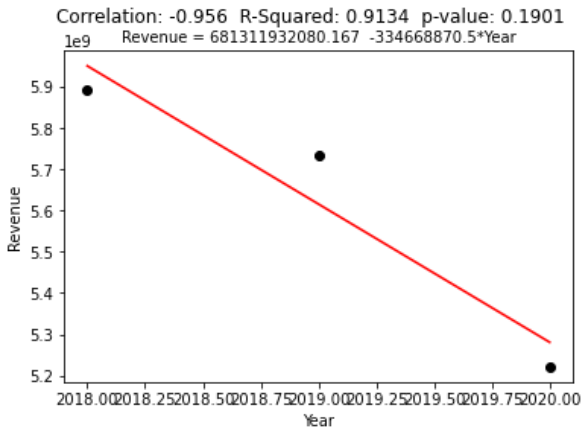
1. Use two data sets of your choice. Run at least 2 regressions on each data set. At least 1 of your regressions must have a negative correlation. At least one of your regressions must have a positive correlation. You must build graphs for each.

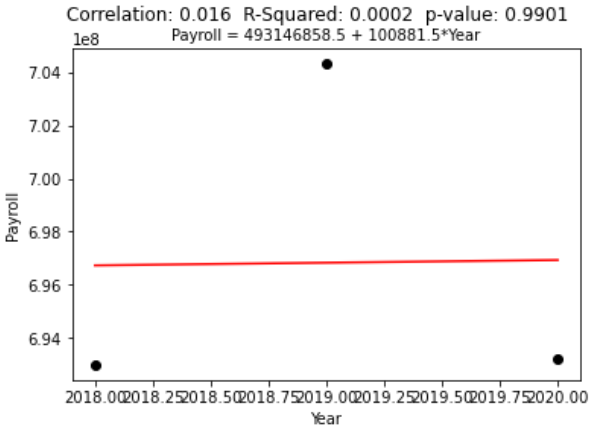
The first worksheet I used for this problem shows data on how many 3- and 4-bedroom homes were built in the South in America each year from 1973 to 2020. The second set of data I used shows the total revenue and annual payroll (both in thousands) of manufacturers in the United States in 2018, 2019, and 2020. Both of these datasets were found on the US Census website.

I did two regressions on each data set. Each graph has "Year" on the x-axis. The regressions tested to see if there was a relationship between the year and the number of 3-bedroom homes, 4-bedroom homes, revenue, or payroll, respectively.





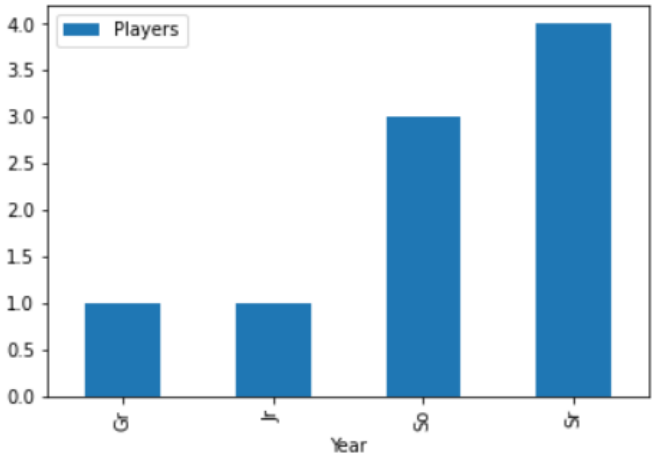




The two regressions on number of homes built proved to have no correlation. However, I failed to reject Ho for the second two regressions, which were on manufacturing data. This means that I could not prove that year does not help to predict annual revenue or payroll in the manufacturing industry.

2. Use a data set and build 3 different pivot tables that each have a least a different index row or column. (You cannot build 2 pivot tables that have the same index rows and columns but aggregate differently. You also cannot simply flip the index rows and columns.) For each pivot table, build a graph.

For my pivot tables I used data on a women's university volleyball team. Specifically, I looked at the frequencies of their year in college, US state they came from, and the position they play. I found this data on the University of Portland Athletics' website.



Gr – Graduate Student

Jr – Junior

So – Sophomore

Sr - Senior

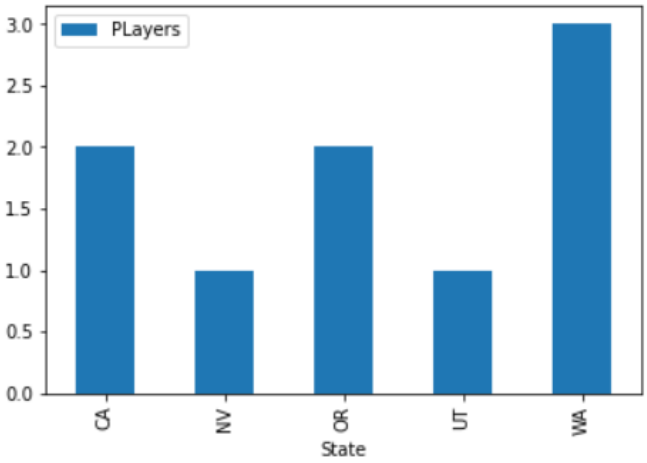
CA – California

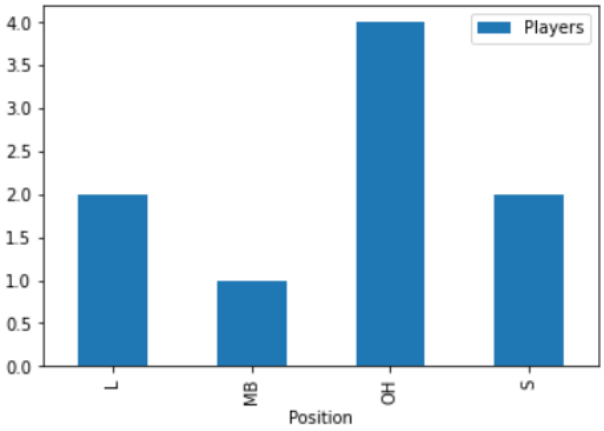
NV – Nevada

OR – Oregon

UT – Utah

WA - Washington





L – Libero

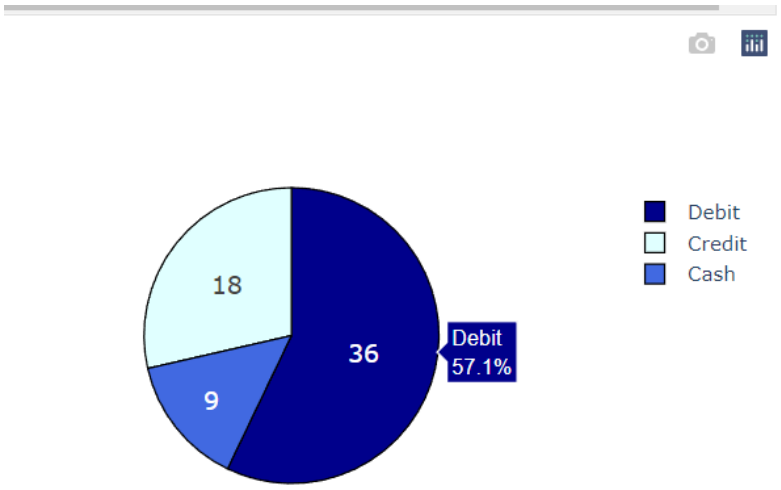
MB – Middle Blocker

OH – Outside Hitter

S - Setter

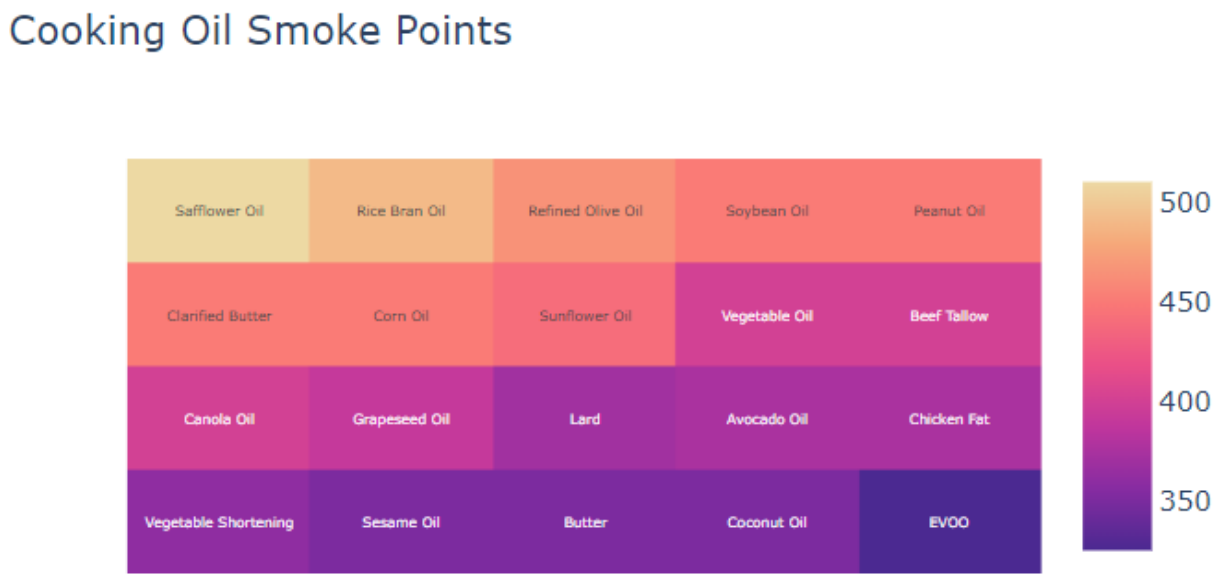
3. Explore plotly and build 2 interactive graphs. You can contrive the data for this, but it must be based on something in the real world.

   For my first graph I made up data on how many people at a coffee shop paid with debit, credit, or cash. I made this into a pie chart. Each section shows how many people used that payment method. When you hover over the section it shows the percent of customers who used the payment method. Here my mouse was over "Debit."



Next I built a choropleth map to show how each US state voted in the 2020 presidential election. Blue represents a democratic vote, red is republican, and purple means there are two districts that were split.

For my final plotly graph I made an annotated heatmap that shows the smoke points of various cooking oils. They are arranged from highest to lowest and hovering over each one tells you it's exact smoke point.



4. Use a data set and perform a lower, upper, and two-tail test. You will have to contrive the population mean. Do so in such a way that at least one of the tests results in rejecting Ho.

This dataset showed the number of full-time employees and total full-time payroll for each government department, from 2017 to 2020. I filtered the data to look specifically at the numbers for the highway department. This data was from the US Census website.

My upper tail test was asking if the highway department hires more employees than the other government departments.

```
Government Employment - Upper tail test
Population Mean:  470000
Sample Size:  4
Sample Mean:  471608.5
Sample Standard Deviation:  2906.9762640929835
Level of Confidence:  0.95
Alpha:  0.05

t Statistic:  1.1066481827651757
Raw P Vale:  0.3492102094865546
Adjusted P value:  0.1746051047432773
Fail to Reject Ho: Cannot say that Highways hires more employees.
```

My lower tail test was asking if the highway department has a lower payroll than the other government departments.

```
Government payroll - lower tail test
Population Mean:  6000000000
Sample Size:  4
Sample Mean:  2369363576.0
Sample standard deviation:  90661209.88043483
Level of Confidence:  0.95
Alpha:  0.050000000000000044

t statistic:  -80.09238854826954
raw p value:  4.289960917029715e-06
Adjusted p value:  2.1449804585148574e-06
Reject Ho: The highway department has a lower payroll than other departments.
```

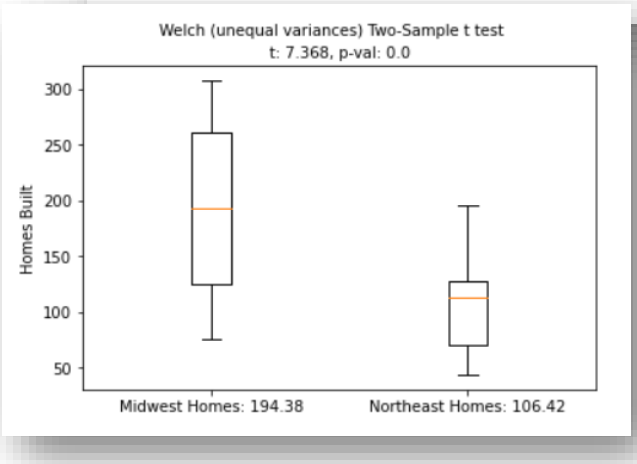
My two-tail test was seeing if the highway department hires about the same number of full-time employees as the other government departments.

```
Government Employment - Two Tail Test
Population Mean:  600000
Sample Size:  4
Sample Mean:  471608.5
Sample standard deviation:  2906.9762640929835
Level of Confidence:  0.95
Alpha:  0.050000000000000044

t statistic:  -88.33336658843336
pvalue:  3.198129622627241e-06
Reject Ho: Employment numbers among departments are not equal
```

5. Use one or more data sets and perform 2 two-sample hypothesis tests. One of the tests must result in rejecting Ho. Build a box plot for each test.

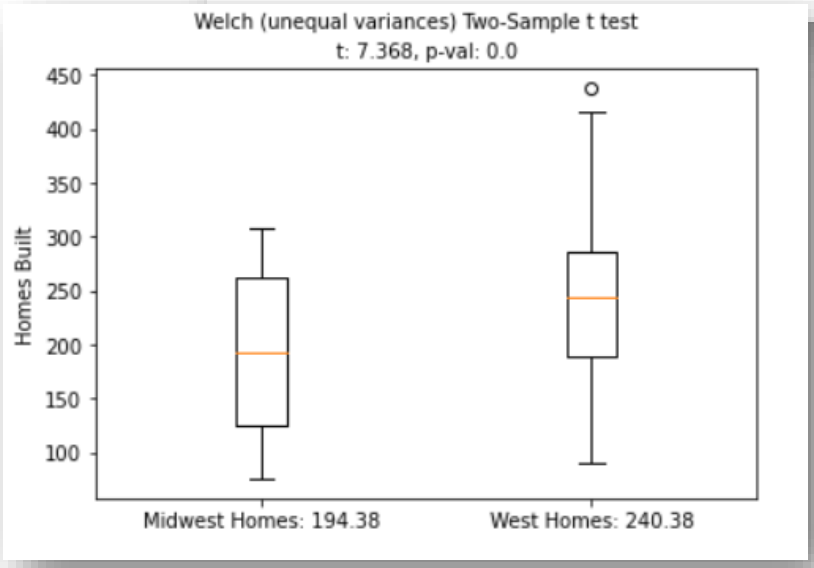
   This data from the Census shows the total number of homes built in each region of the US from 1973-2020. My first hypothesis test was to see if there was a difference between the average number of homes built in the Midwest and the Northeast. I performed a two-sample hypothesis test and built a box plot for the data.

```
Ho: The means for Homes Built in the Midwest and Northeast are the same.
Ha: The means for Homes Built in the Midwest and Northeast are not the same.


This is a Welch (unequal variances) Two-Sample t test of equal means
The t test statistic is 7.368 and the p-value is 0.0

p-value = 0.0
Because p-value is less than alpha,
Conclusion: Reject Ho: The means are not equal
```



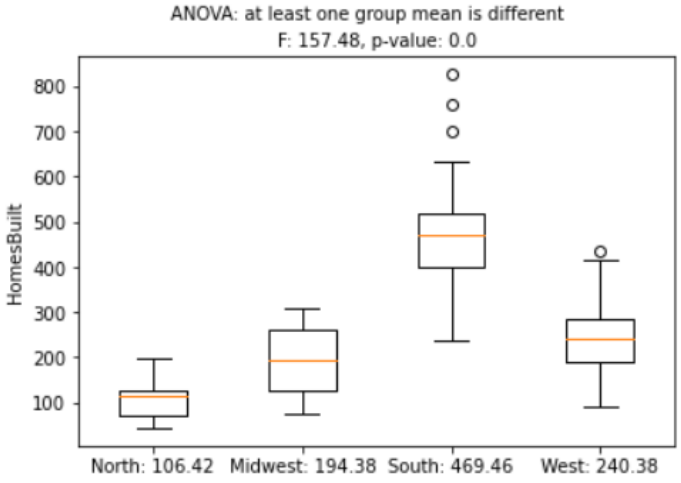
For my next hypothesis test I looked at the number of homes built in the Midwest and the West, since those means were more similar. Again, I performed a hypothesis test and made a box plot. They were much closer, but still not similar enough to be equal.

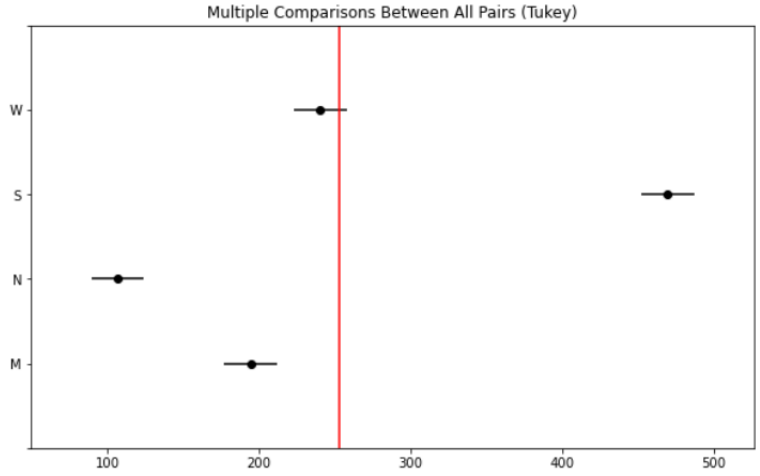
```
Ho: The means for Homes Built in the Midwest and West are the same.
Ha: The means for Homes Built in the Midwest and the West are not the same.

This is a Welch (unequal variances) Two-Sample t test of equal means
The t test statistic is -2.917 and the p-value is 0.0044
p-value = 0.0044
Because p-value is less than alpha,
Conclusion: Reject Ho: The means are not equal
```



6. Use a data set and perform an ANOVA. The data must be broken into at least 4 groups, with each containing at least 25 observations. Build a box plot and a vlines graph.

   For my ANOVA testing I went back to my large dataset on homes that were built between 1973 and 2020. I compared the total number of homes built each year in the North, Midwest, South, and West.

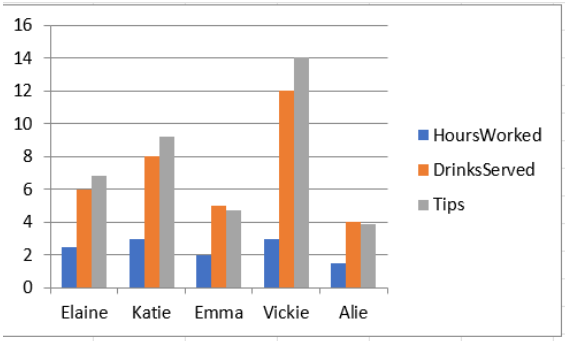


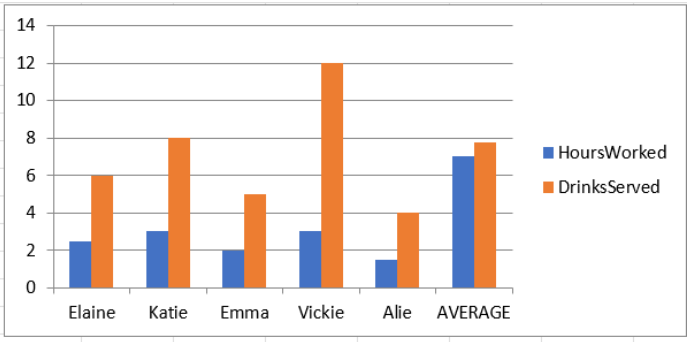


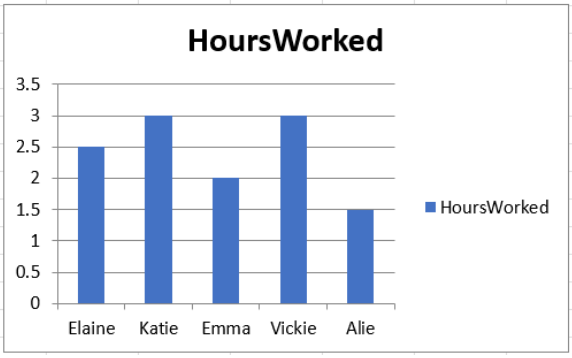
7. Use a data set (can be contrived), create an Excel workbook using XLWings. You must do the things I showed you in class (adding a dataframe, adding some summaries, and adding 3 different graphs.

   Here I made up some data about a coffee shop. I came up with hours each barista worked, the number of drinks they served, and their individual tips. Then I exported the data to excel and made the graphs from it.

	HoursWor	DrinksServ	Tips
Elaine	2.5	6	6.8
Katie	3	8	9.25
Emma	2	5	4.75
Vickie	3	12	14.1
Alie	1.5	4	3.9
AVERAGE	7	7.76	2.4



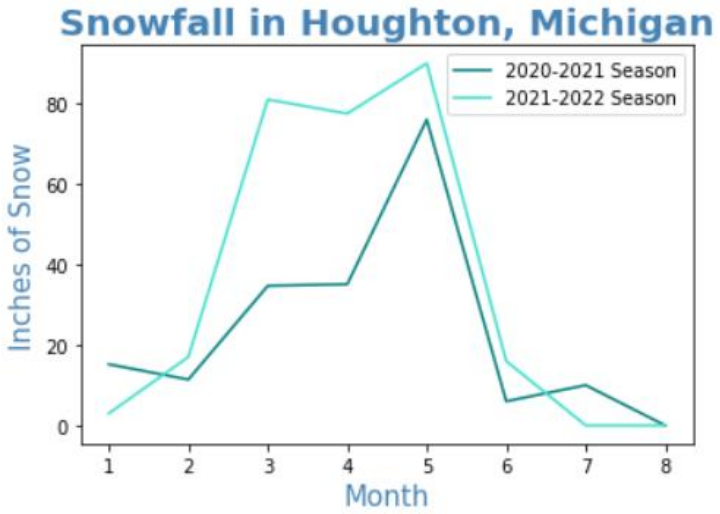




8. Use one or more data sets to build a variety of graphs (not the plotly graphs from 3). Explore graphing and add graph elements that we didn't address in class.

   For my first graph I used data I found online that had monthly snowfall in my hometown. I collected the data only from the last two winter seasons. I made a two-line plot to examine the difference by month in the last two seasons. I customized the colors, alignment, size, and weight of labels.

   (Yes, we have *minimum* 8 months of snowfall each year. It's horrible.)



   My second graph was built from data on DU's hockey team. I looked at the points they scored during each game this season, and the result (if DU won or lost). I put this on a scatter plot with game result on the x axis and points scored on the y axis to see if we had more points in winning games. I played around with different markers before settling on the thin diamond, and I made the colors of the graph maroon and gold.