DATA WORK:

I decided to work with data on engineers (specifically combining mechanical and civil engineers, codes 1460 and 1360 respectively). I created the dataframe, new variables, and inspected the data.

DISTRIBUTION OF WAGES FOR ENGINEERS (IRRESPECTIVE OF SEX)

Firstly, I wanted to see the distribution of wages for these engineers using bar charts(not focusing on the sexes). I found that the average wages per hour were about $35.

WAGES FOR ENGINEERS (FOCUSING ON SEX)- Regressions and Scatter Plots

-I then changed the data to find which of the engineers were female. I found that about 12.5% of the engineers were female- after completing the assignment, I feel that the little number of women in these jobs may have impacted the findings..

-I then looked at the gender gap between male and female wages for the occupations. I took a level-level regression and discovered it didn't fit the data too well (R-squared was 0.002 or the regression line fit 0.2% of the data). From the slope I saw that female mechanical/civil engineers earn 2% less, on average, than male engineers. We can be 95% confident that the average difference between hourly earnings of female mechanical/civil engineers versus male was -5% to 1%. The CI includes 0 which means we can't say with 95% confidence that their average earnings (between males and females) are the not same. The t-stat is 1.2 which is less than 1.96 so we can't reject the H0 hypothesis. The p-value is also 0.2 which is much larger than 0.05. The coefficient can't be considered statistically significant at 5%.

-I then took a log-level regression and got a really weird reading/summary with mostly 'nan' values. I found that the data looked somewhat odd and tended not to fit the regression line at all for the log-level regression.

-After the TA session, I went back and did a Robust SE regression (for heteroskedasticity) and found similar findings to my first level-level regression: we can't reject the null hypothesis (more in depth analysis seen on Jupyter Notebook).

WAGES v. HIGHEST LEVEL OF EDUCATION/GENDER- Regression and Scatter Plot

-I categorized the grade92 data into dummy variables of ed_low (1st-middle school), ed_mid (some/all highschool), and ed_high (college and beyond). I performed regressions to compare the three levels of education. I found that, as expected, those with 1st-middle school education earned the least, those with college degrees and beyond earned the most, and those with some/all highschool fell in between the other two categories. Interestingly enough, I also found that with education, the coefficient on 'female' was a larger negative than without. They tended to earn a greater amount less than males when education was included in comparison to when it wasn't. (There is a more in-depth analysis on these regressions found in the Jupyter Notebook as well).

Lastly, I created a scatterplot to model the gender gap based on education level/gender. I had to first create a new column in the DF, 'education_level' with the values low, medium, and high to be able to represent them as categorical variables on the visualization. I then modeled it and found that education level has some say in how wages look, however, gender did not seem to play a large role.