Github Repository link:

https://github.com/Gaoyuan66/Group5\_Applied\_Stats/tree/master/group\_assignment

**Executive Summary**

**Dataset Limitations:**

Although the sample size of this data set is 26765, it is important to take into account that it is not a random sample. Since it is sourced from the website askamanager.org, the respondents are already predisposed to fit some sort of group. For example, most respondents have a job already and more than half have received a college education or above. That being said, here are the key takeaways of our analysis on the askamanager salary survey.

**Key Takeaways:**

1. Gender and Salary: Overall there is a statistically significant salary gap between men and woman in the dataset. This is confirmed through multiple methods such as plotting the data, running a T.Test, calculating a confidence interval, and than testing our results through bootstrapping. As we looked deeper, however, there are some key differences in the gender salary gap in different industries. To start, the largest gap is seen in the Law industry, which has on average a $40,000 difference in salary between men and woman. In comparison, in the Leisure, Sport & Tourism, Education, and Agriculture or Forestry industries the gender gap switches, as woman have higher average salaries in the fields.
2. Changing Industry and Salary: During our data analysis we were curious to see whether changing industries lead to higher annual salaries. We created a graph to depict this and we found that those who didn't switch industry had higher salaries. We then ran a T.test that showed that this difference in average salary between those that remained in the same industry and those that switched was statically significant. In addition, when looking at industry, for both men and women Law is the industry with the highest salary levels. In addition, individuals working in law have the highest education levels. This brings us to our next point of the relationship between education and salary.
3. Education and Salary: Education level was shown to be statistically significant in explaining variation in annual salaries in our linear regression models. Those with professional degrees (MD, JD, etc) on average earned higher salaries. The gender gap is also prevalent within the respondents who have higher education levels. The pay gap between genders is highest at the high school level and gradually lowers until the professional degree level. This shows that as education levels increase, the salary gap gets lower. There is however an exception with the gender pay gap spiking with PhDs.
4. As one would expect, we found that as years of experience in the field increased, average annual salary also increased. This is because as people gain more experience, they can provide more value and garner a higher salary. This was shown in the regression models with the different levels in years of experience all being statistically significant.
5. Salary and Race: When looking into the intersection of salary and race, we picked the three races with at least 500 respondents - White, African American or Black, and Asian or Asian American. When plotting the data in a boxplot format, we could easily see that the Asian race had on average higher salaries. However, boxplots for the White and African American or Black races were very similar and we decided to run a T.test. The T.test ended up calculating a P value of .3, thus we can conclude that there isn't a statistically significant difference between the means in salary of these two races.

Regression & Results on Test data:

When working on our regression model, we tested various variables to try and determine which ones could explain salary levels with the greatest accuracy. When adding and eliminating variable, we had the following observations:

The variables that had a significant impact on salary were expirience in a specific field, industry type (data related jobs, etc.), race, and population of the state in which the respondent lives in. By taking these into account, we created a model with an adjusted r-squared of circa 42%. This means that the variables imputed in the model explains 42% the porportion of variability in annual salaries of the respondents.

1. Experience in a specific field explains more variance of salaries than the variables of age and the years of professional experience in total.
2. Data related jobs are averagely paid higher by $5046 than other jobs, scientists are paid higher by $8820.476 and data scientists are paid higher by approximately *$31300* in the survey. However, doing a database related job has an average annual salary of $12360 lower than average. Engineers on average earn $22182 more per year and managers earn $7987 more than average. Finally, doing a research job has no significant positive effects on salary.
3. Races do have influences on salaries *in the survey*. Asians (4.54% of the respondents) earn the most and white people (84.68%) earn the lowest on average. That may be caused by the bias of sampling.
4. Population of the state has positive effects on average salary. People in a state with 1 million more people averagely earn $383 more.

As we added variables to our model, the RMSE and the R-squared improved in both out train and test data. The performance of final model RMSE on train data: 28801. RMSE on test data: 28398. The fact that these numbers are not too different gives us confidence that we did not overfit our data but still space to improve. Furthermore, the R Squared on train data is 42.30% and the R Squared on the test is data 44.23%, which is also not a large difference. It makes sense that our train data R Squared is lower, as it was run on a smaller sample.