**Data Science Programming 1 Project Report**  
Pace University, December 23, 2021

Analysis on Data Science and STEM Salaries

**V Sai Pranav, Master’s in Data Science**

**Instructor: Professor Yegin Genc  
Pace University**

**Introduction:**

This project is mainly concentrated on the analysis of the salary based on the levels, titles, compensation, location, gender, etc. in different companies. Analysis was needed here to extract the information on which gender is more experienced. Also, a comparison between the earnings based on gender was successfully fulfilled.

**Data:**

The dataset used in this project is a salary dataset of Data Science and STEM companies and is a public dataset from Kaggle.

**Columns**:

There were 29 variables in the dataset of which

‘timestamp’, ‘company’, ‘level’, ‘yearsofexperience’, ‘yearsatcompany’, ‘bonus’, ‘gender’

were used to get the desired results.

**Rows**:

The number of rows in the dataset is ’62,642’.

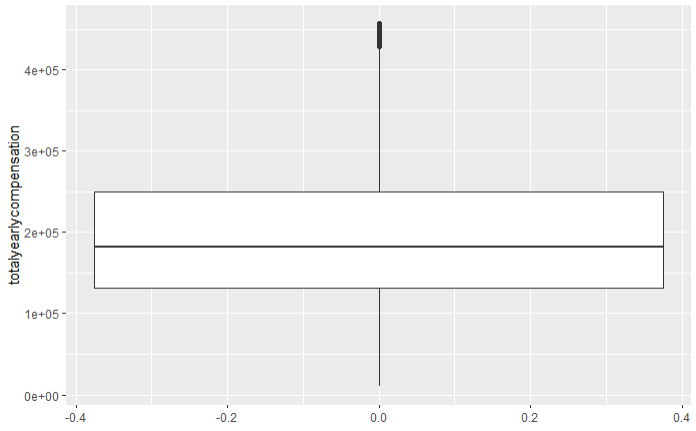
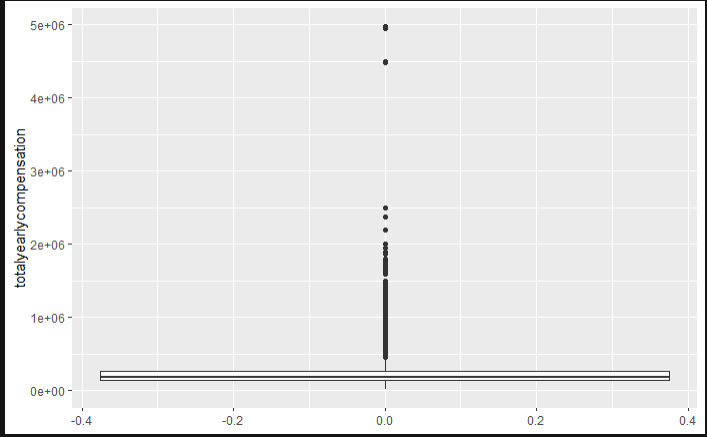
**Data Cleaning:**

**Outliers Detection & Removal:**

Generally, Outliers are the data points that are outside the limits of the existing actual data points. With the outliers and without the outliers the accuracy of the model differs. Better performance is achieved after the outlier Removal.

To remove the outliers, the quantile technique was used where we set the maximum threshold and minimum threshold. Data points outside these ranges can be identified as outliers. Below are the maximum and minimum values in the dataset.

The maximum Threshold was set to 0.75 which is 75% and the minimum threshold was set to 0.25 which is 25%. There were 3133 outliers in the dataset.



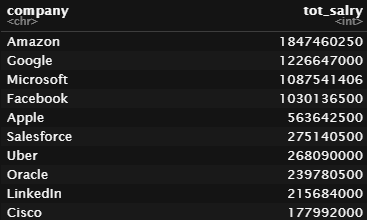
As you can see in the above image, the plot on the left is showing us the outliers which are the dark spots above the boxplot. The plot on the right side shows us that outliers in the data have been removed.

**Nulls Detection and Removal:**

Dataset also had Null values in the data which were affecting the outcome of the graphs. ‘108708’ was the count of the null values which was severely influencing the distribution graphs. na.omit() was the function used to remove all the null values in the whole dataset. 21,575 rows was the count before null removal, 20,750 rows was the count after null removal which says that there were around 850 rows with Nulls.

**Descriptive Statistics:**

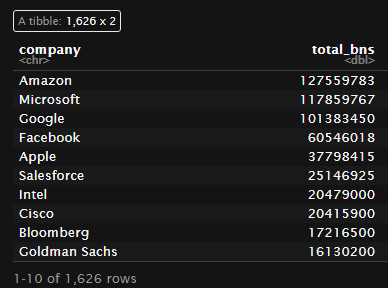
In the first example companies that offer more compensation were extracted by sorting and summing up the 'totalyearlycompensation' column in descending order. Another contradicting scenario, to concentrate on here is that as we do not know the number of employees in the companies that can be the issue to concentrate on. Like if Amazon has a large number of employees and Cisco has a smaller number of employees compared to Amazon that can affect the outcome too.

****

In this example we are, extracting the information on in which location the salary being compensated is maximum.

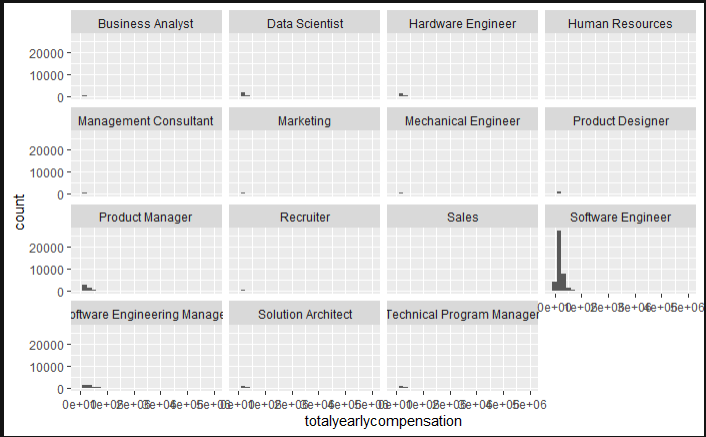
****

The below example gives us the companies that are offering more bonuses to the employees.

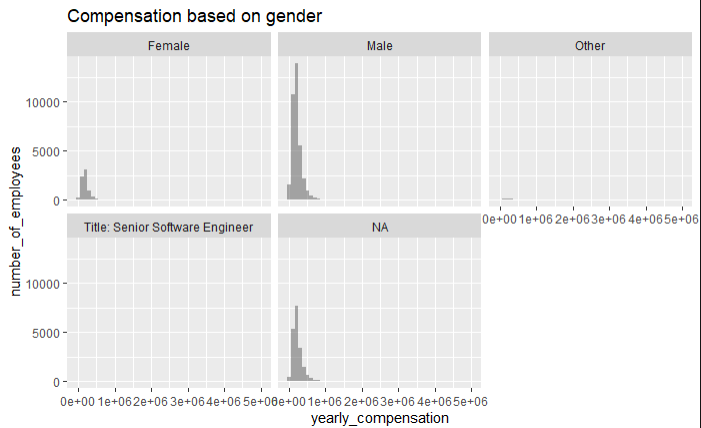
****

**Exploratory Analysis:**

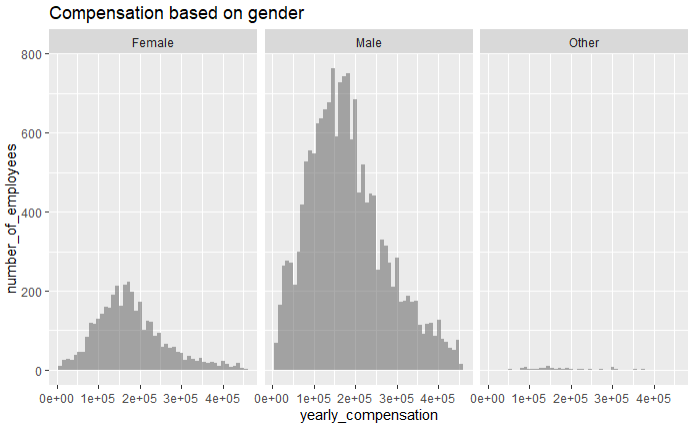
First is the distribution graph, looking at which we can say that most of the data in the ‘totalyearlycompensation’ column are part of Software Engineer.



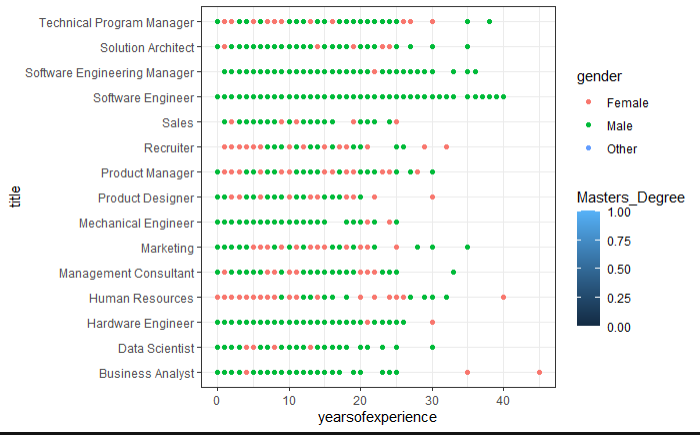
Second, comes the yearly compensation based on Gender, as you can see below graph that we have NA’s as well as the unwanted data.



Here is the distribution graph after the removal of NA’s and the unwanted data. Here we can see the distribution is clearer and we can understand that yearly compensation of male gender is more when compared to female and the other category.



This graph tells us which gender is most experienced based on their work experience. Looking at the Human Resources title we can say that females are more experienced for the specific role. Next comes the Data Scientist & Hardware Engineer where we can see that males are the ones who are more experienced. Now if we look at the Marketing role both men and women hold an equal amount of work experience.



**Results & Conclusions:**

In this project, the main concentration was to do a comparison between the different genders available in the dataset. The results were achieved successfully as expected. In addition, another observation was that if there are Nulls and outliers the results, we get will not be accurate there will be deviations, but if we concentrate on removing both Nulls and Outliers from the dataset, we will get better results, as it can be seen in the above graphs. In last, I would like to add that there are contradicting scenarios which I would like to concentrate in my future work on this dataset.

**References:**

1. <https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries> (Dataset)
2. <https://datascienceplus.com/missing-values-in-r/>
3. <https://okanbulut.github.io/bigdata/visualizing-big-data.html>