

**CSE303: Statistics for Data Science**

**[Summer 2023]**

**Report on Assignment 01**

**Data Visualization**

**Submitted by**

**Student ID: 2020-3-60-012**

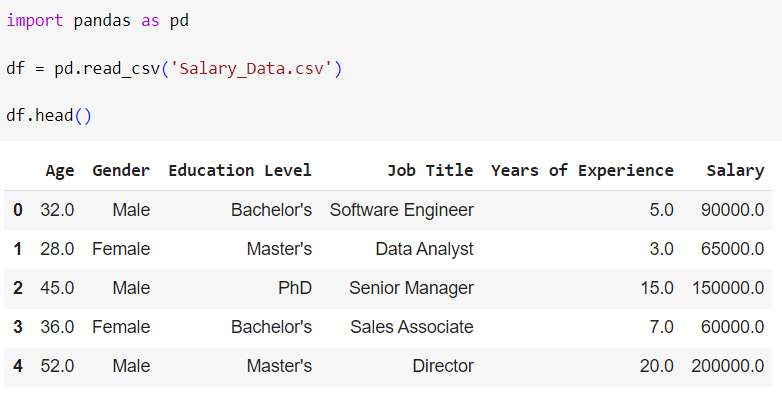
**Student Name: Sadia Islam Prova**

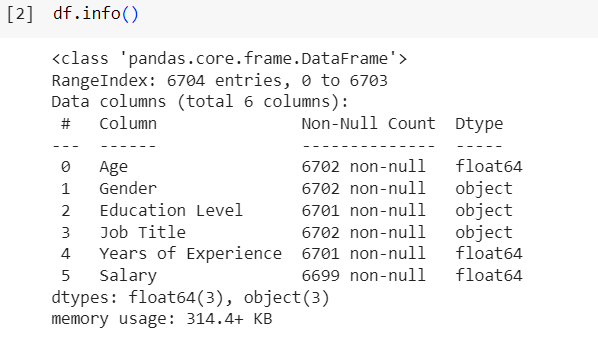
# Introduction:

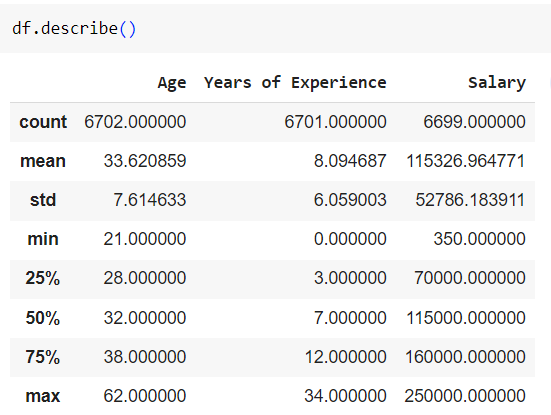
The assignment is on data visualization on a dataset of salary. From the salary information, we aim to explore and uncover trends, patterns, and valuable insights by visualizing compensation data in a meaningful and interactive manner. By leveraging various visualization techniques and interactive tools, we can grasp a comprehensive picture of salary distributions, identify potential gender and discern how different factors influence salary. We will conduct an exploratory data analysis to gain preliminary insights into salary distributions, identify outliers, and discover notable trends. Using popular data visualization libraries such as Matplotlib, Seaborn, or Plotly. We will create an array of interactive visualizations and explain them.

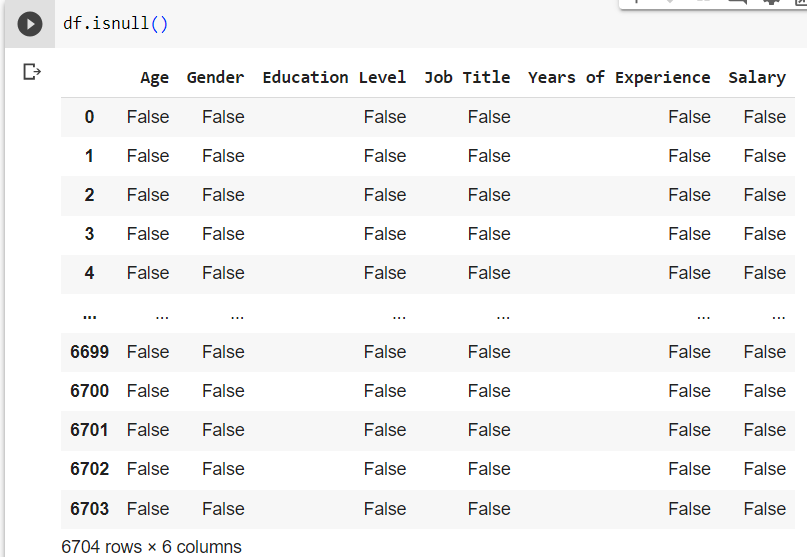
# Dataset Characteristics

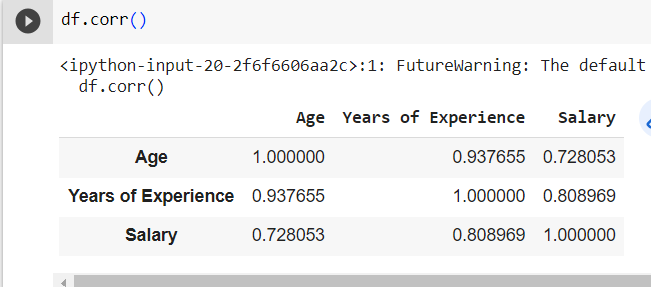
The name of the dataset is “Salary\_Data”. The data set is obtained from an online resource. A total of 6704 data points and 6 columns are in the dataset. The dataset included six variables: age, experience, job role, education level and salary. Among them are 3 numerical and 3 are categorical.









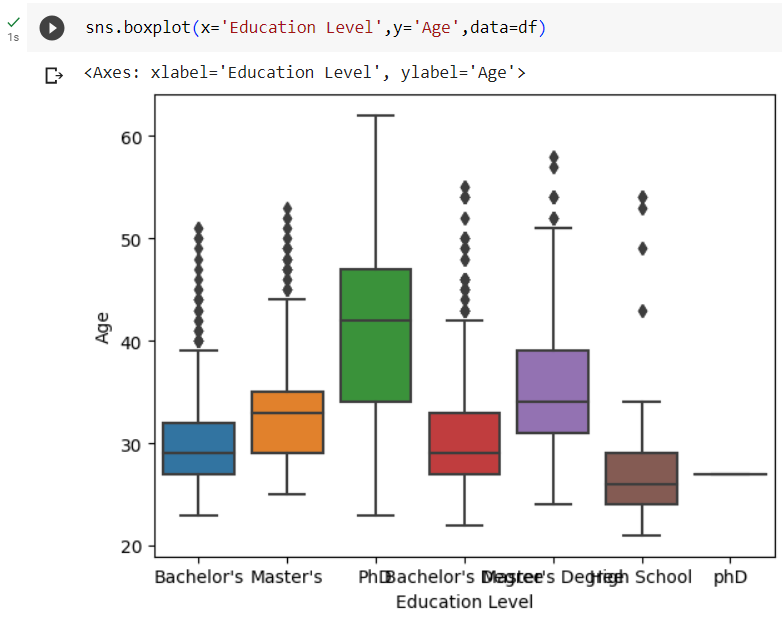


# Exploratory Data Analysis:



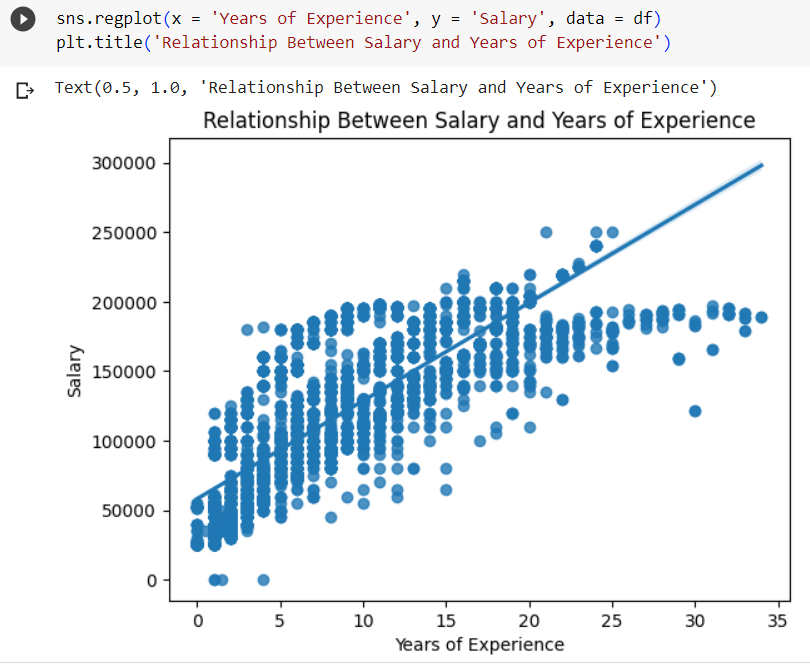
**Figure: 01**

This is boxplot represents the relation between salary and gender. From the above figure, we can say that the salary of male participants is a little bit higher than that of female.



**Figure: 02**

This is boxplot represents the relation between age and education level. From the above figure, we can say that the sage of PhD participants is a little bit higher than that of other degree. we can see that there are a lot of outliers in the model



**Figure: 03**

From the above figure, we can say that there is a positive correlation between years of experience and salary.

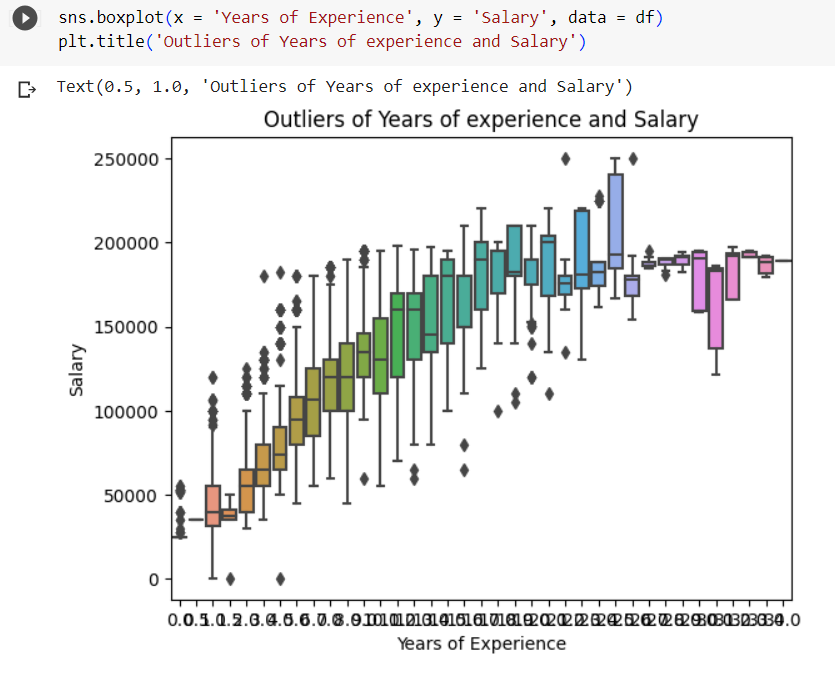
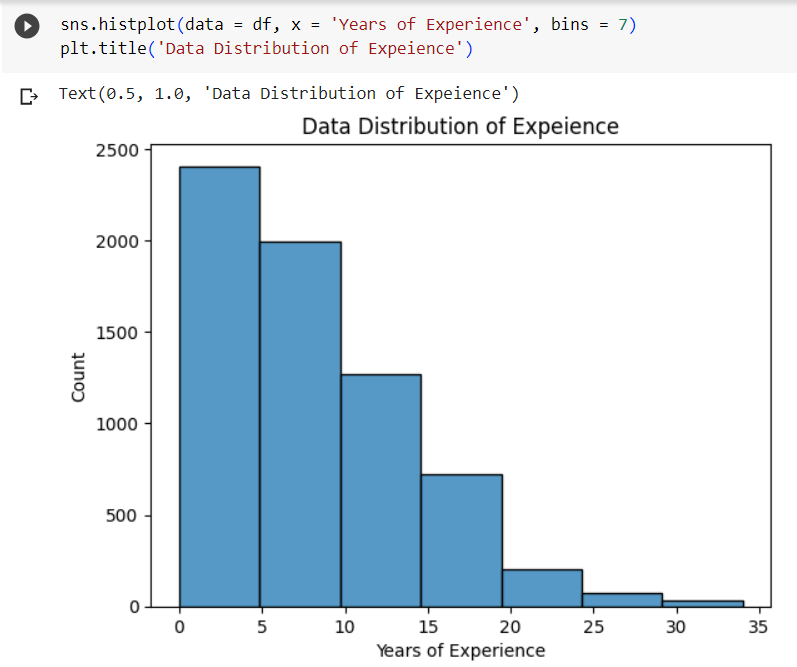


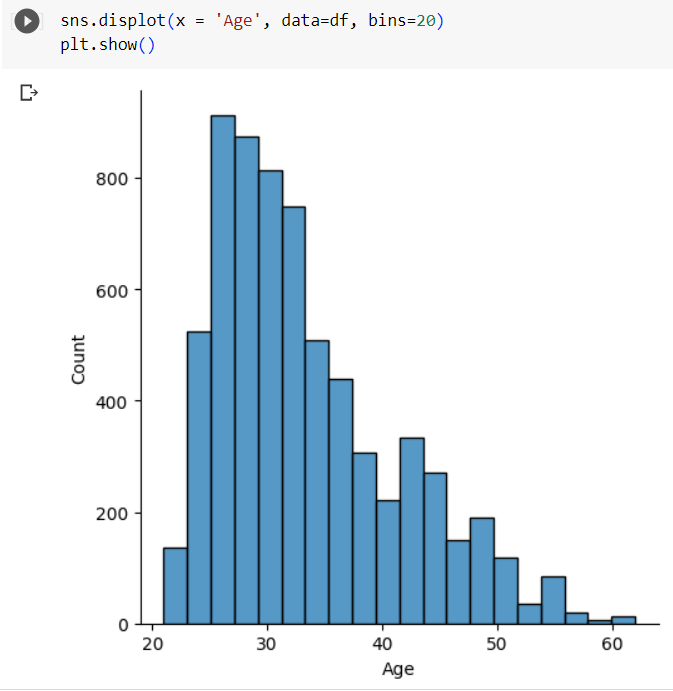
Figure: 04

From the above figure, we can find that the independent and dependent variable have a positive relationship. From the Box plot we can see that there are a lot of outliers in the model



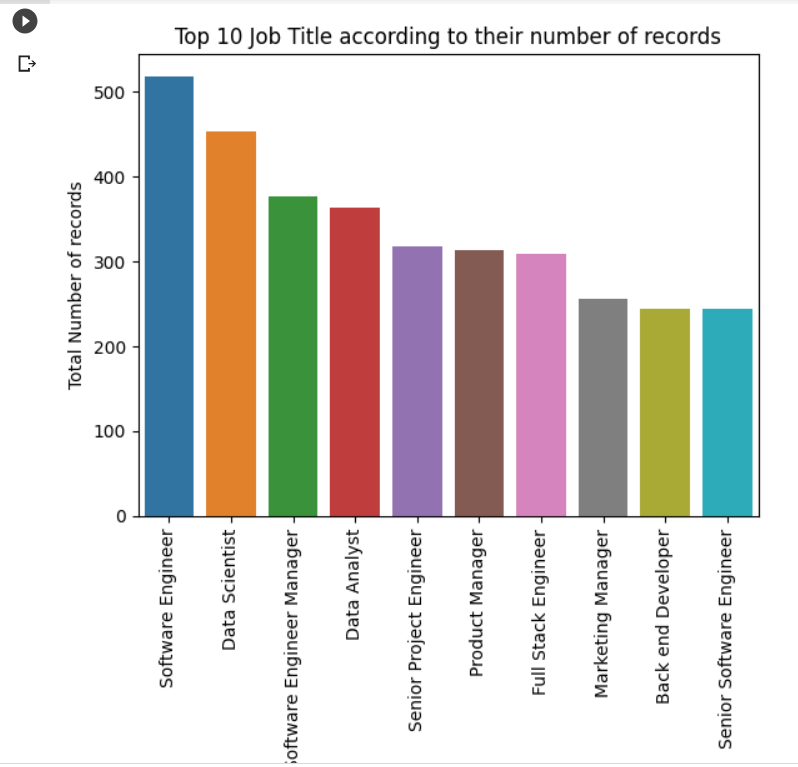
**Figure: 05**

The above figure is for analysing experience. We can see that 0-5 range of experinence has the highest count.



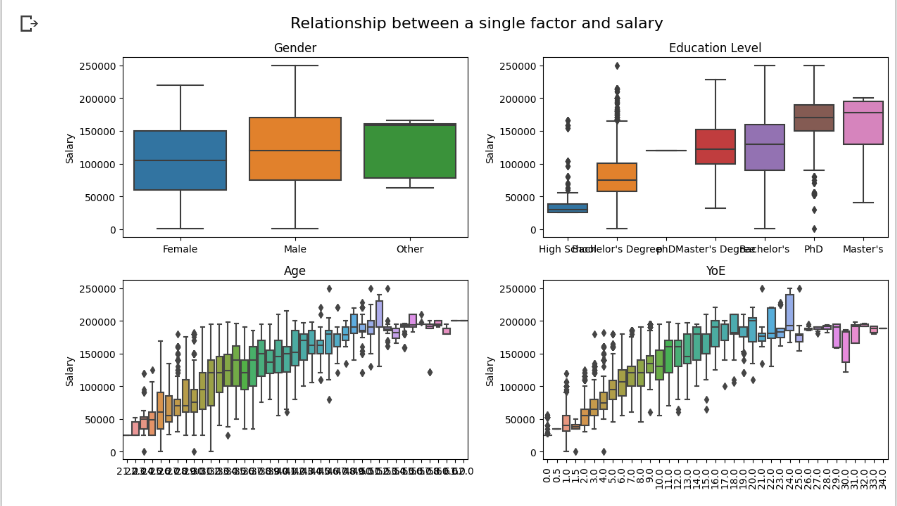
**Figure: 06**

The above figure is for analysing age. We can see that 20-30 range o age has the highest count.



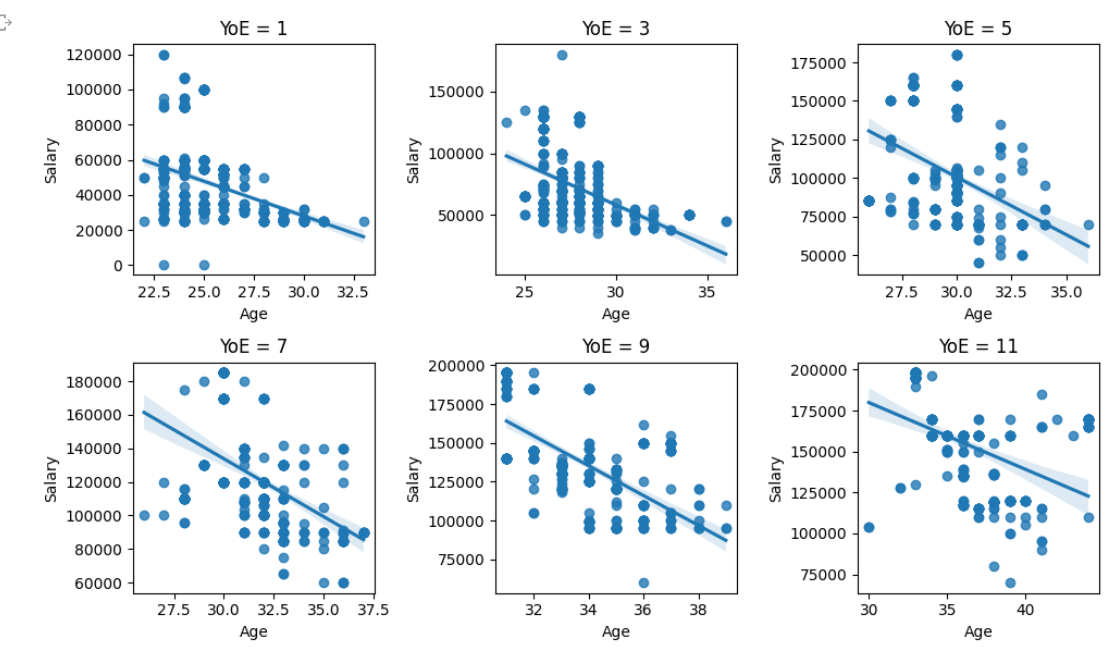
**Figure: 07**

Through the above bar chart, we have represented the relation between total number of records and job title. We can see that software Engineer has the higher number of records.



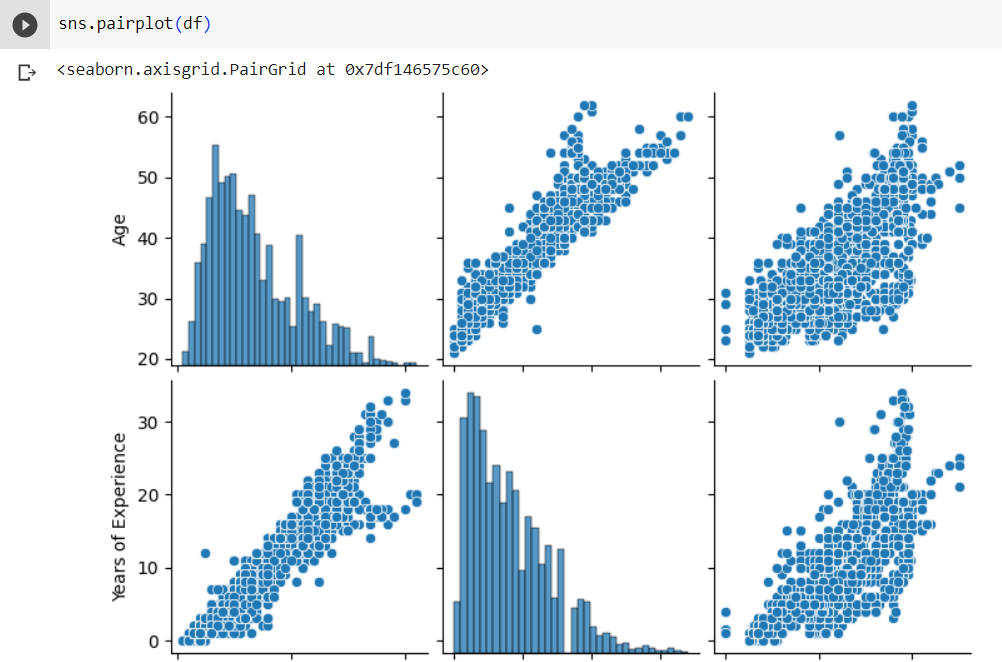
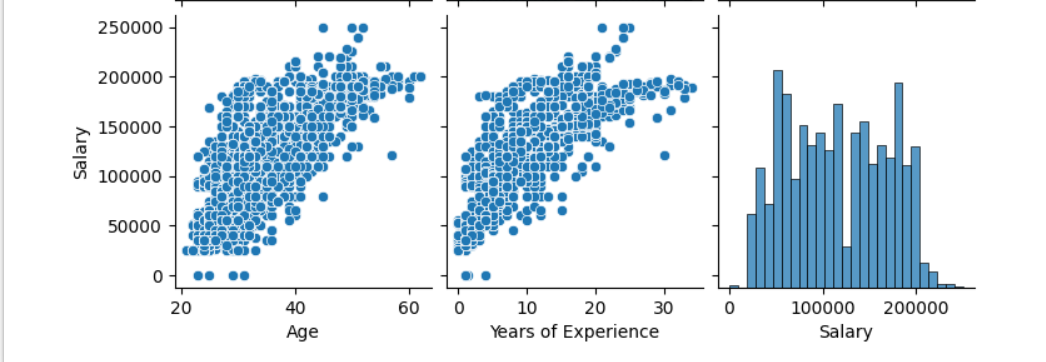
**Figure: 08**

From the above figure, we can say that the salary of male participants is a little bit higher than that of female. The salary of participants with a higher degree is higher. The salary roughly increases as the age and YoE increase.

****

**Figure: 09**

The above figure is for finding out the effect of age on salary. From the relationship between the age and the salary, it reveals that when the YoE is the same, the age and the salary are negatively correlated.

**Figure: 10**

From the above figure, we can see that there is positive correlation between YOE and Salary, Age and Salary.

# Conclusion:

**Link of Colab Notebook:**

<https://colab.research.google.com/drive/18gRru3YgKJNnLQauWs2dQDv10UKjZkYj#scrollTo=eHbgWhdbuKW7>

I have learned how to visualize and analysis data from a data set. How to learn about their characteristics. Also learnt how to visualize data using python and how data are dependent on each other.After finishing the assignment, I have learnt how to plot data and describe them. This assignment was truly helpful for learning data plotting and analysis.