APPLIED DATA SCIENCE ASSIGNMENT

CLUSTERING AND FITTING

NAME: MUHAMMAD MOIZ BUTT

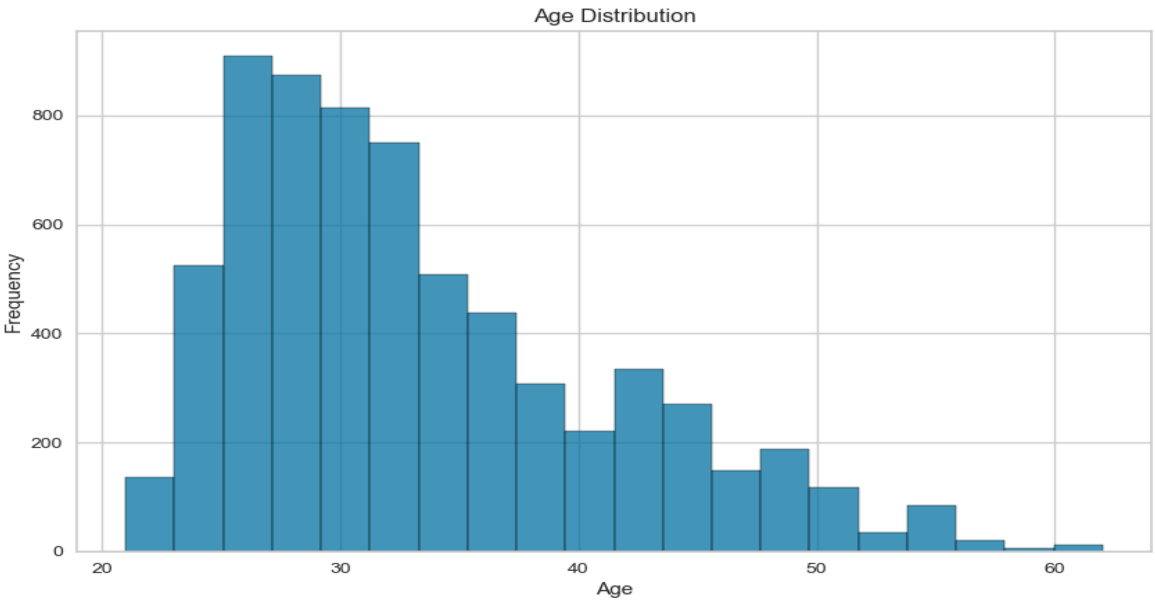
STUDENT ID NUMBER: 23038311

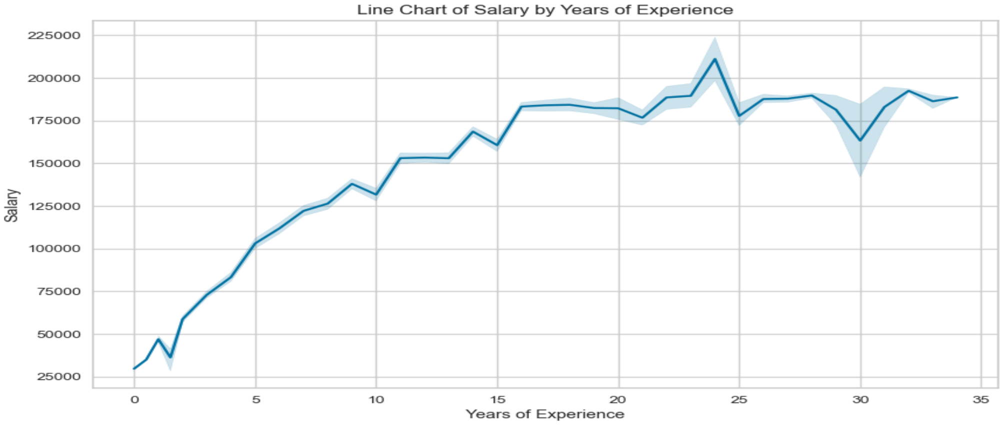
GIT LINK: https://github.com/moizbut/Assignment-.git

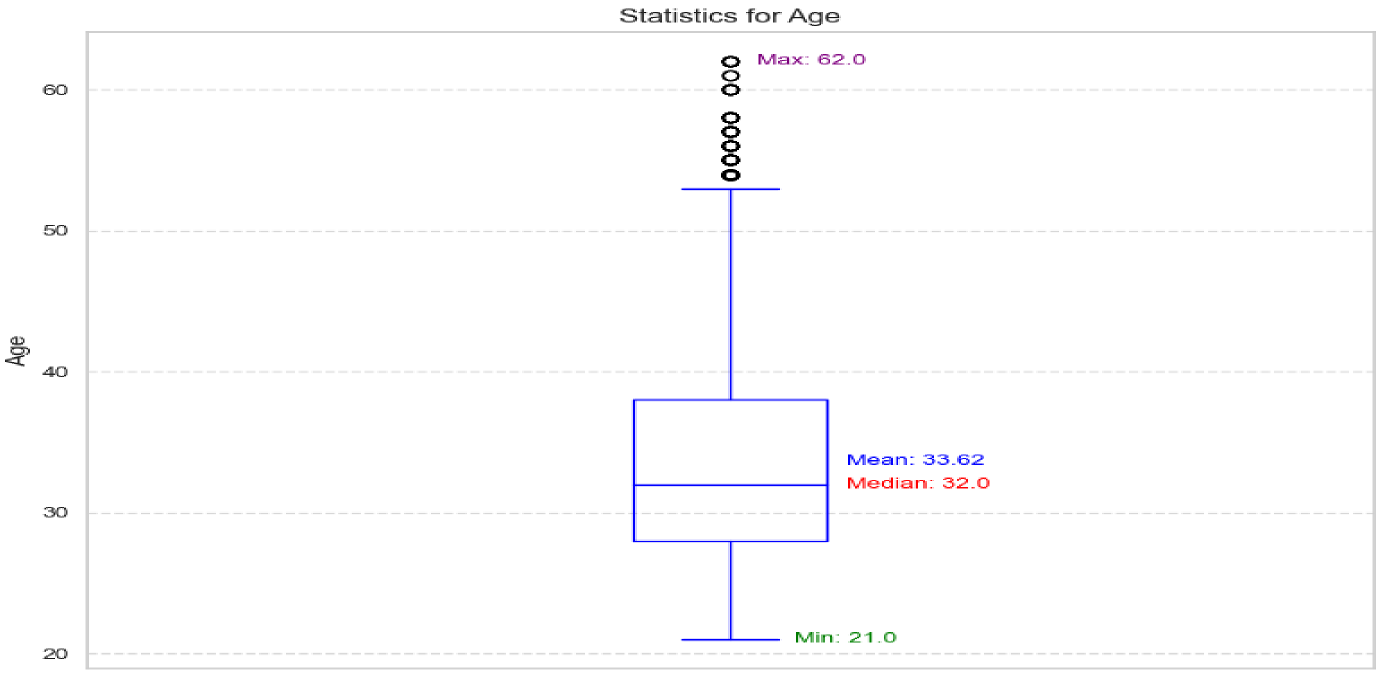
Selection of Dataset:

I have selected a dataset from Kaggle named Salary_Data. This dataset was obtained from multiple sources like surveys, online postings and other publicly available sources. It has 6705 entries and contains six variables Age, Gender, Education Level, Job Title, Years of Experience and Salary. For better understanding our data, we analyzed it and made some plots to see the relationships between different variables. First of all we did the data cleaning in order to get rid of missing columns/rows. There were two missing values and these were removed.

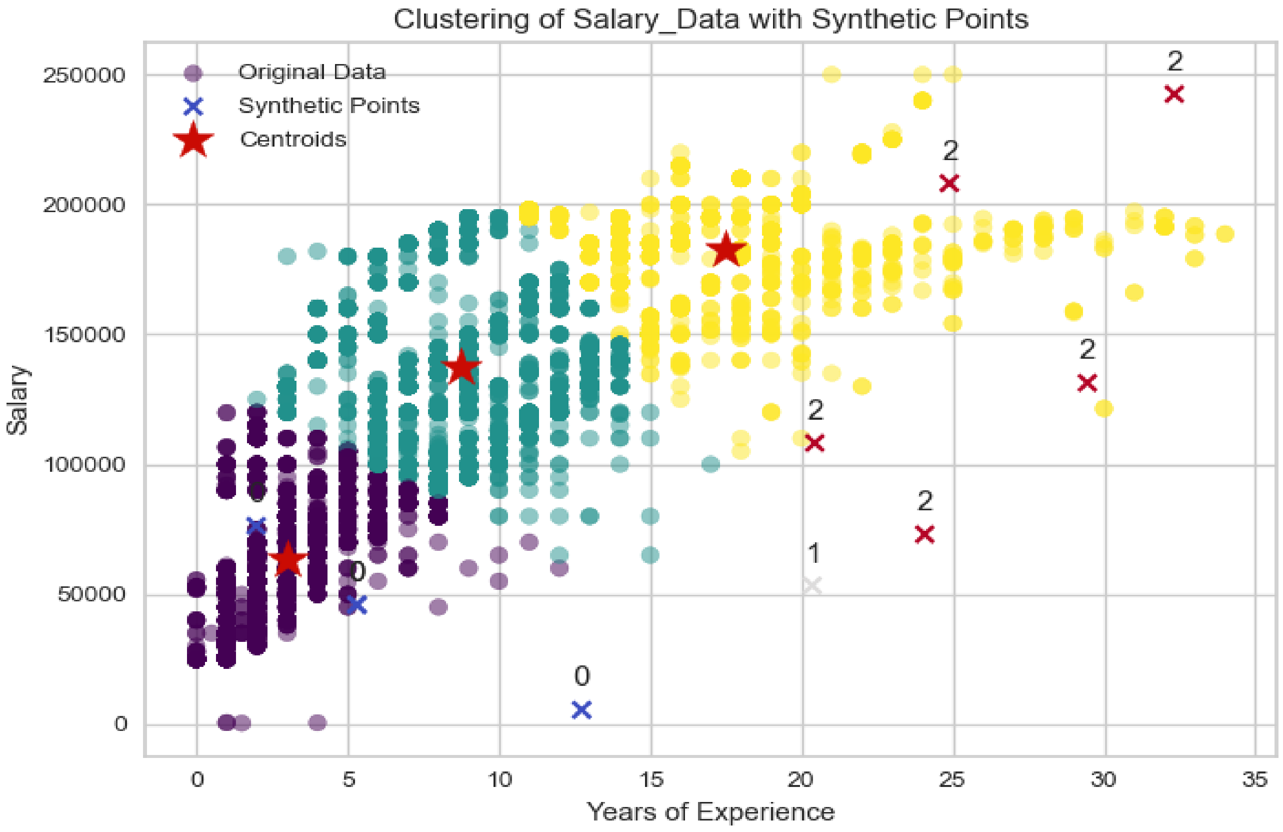
According to the requirements a box plot, histogram and line chart have been produced. This tells us that the average age of the people in dataset is 33.63 years with 62 years as the highest and 21 years as the lowest. As we plotted the line chart, it gave an insight to test a relationship between two variables, **Years of Experience** and **Salary** and we further performed our analysis on these two variables.

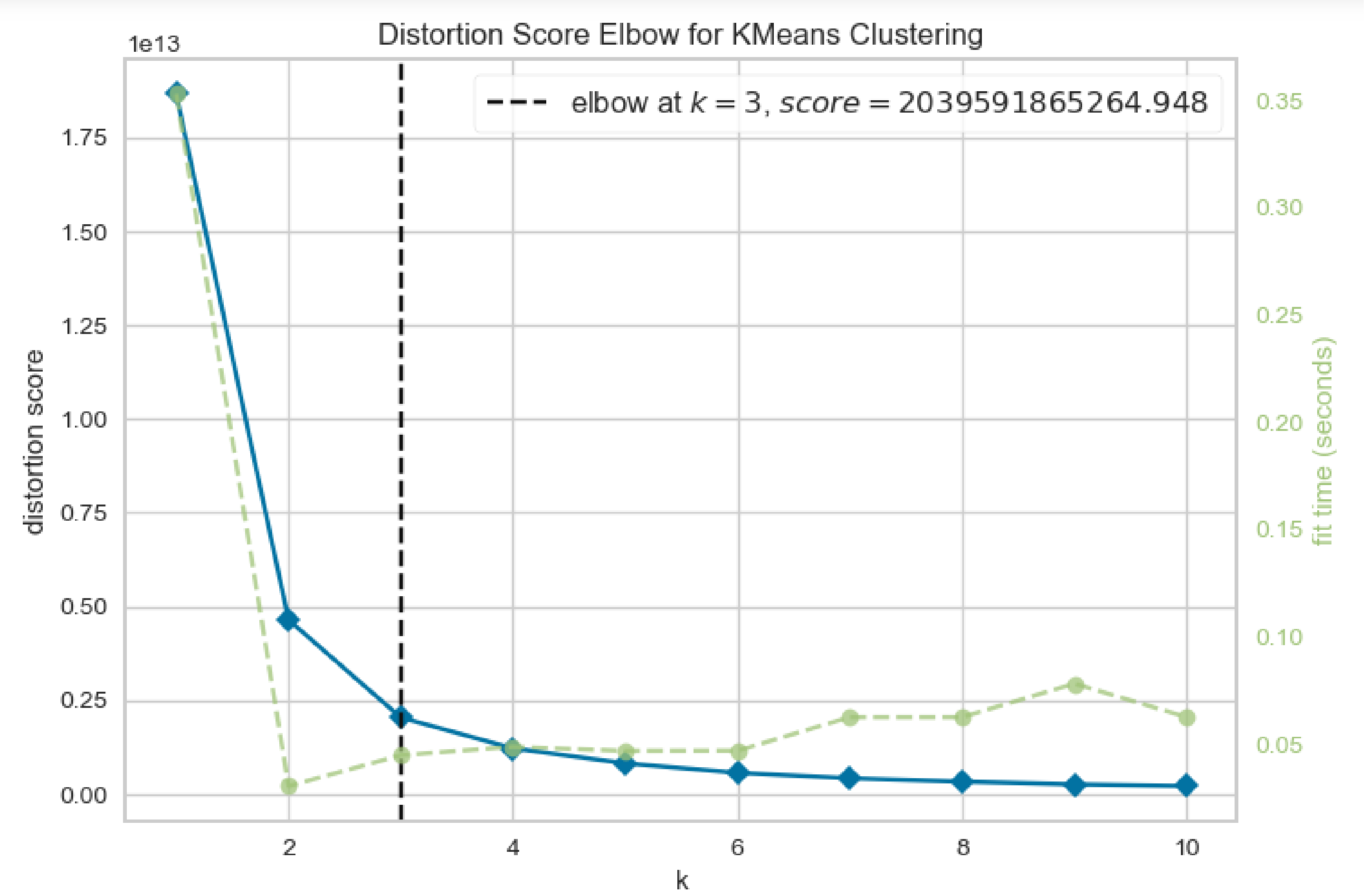




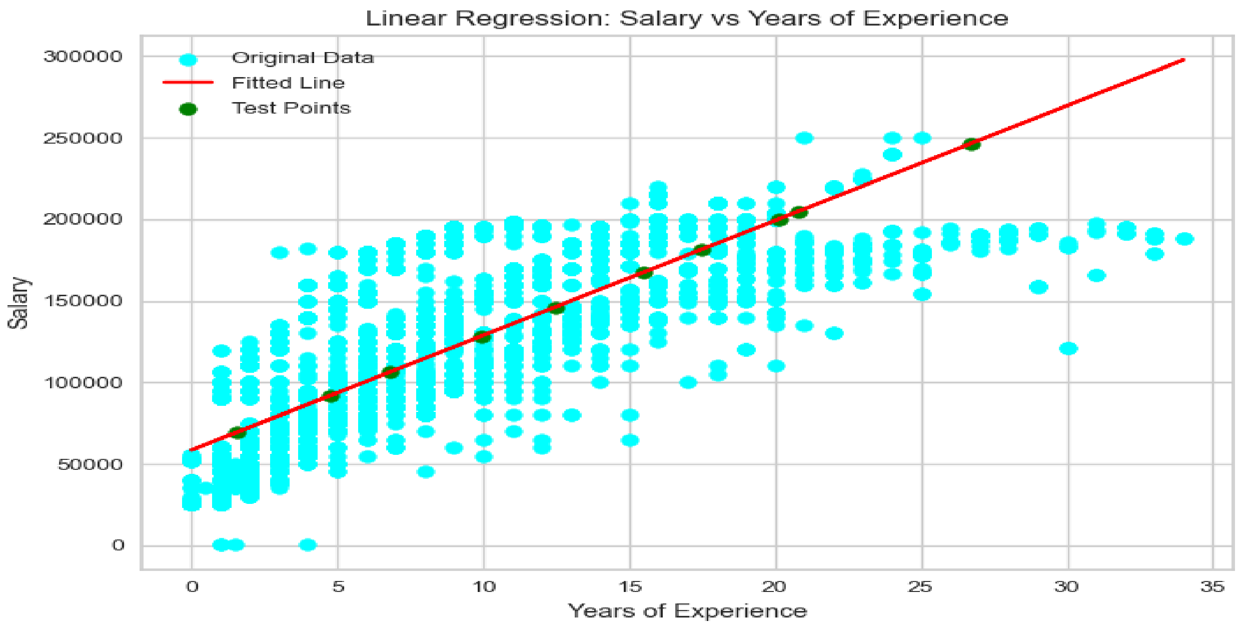


We took these two variables and performed clustering on them. Initially with number of clusters as 5, but then after using the elbow method it suggested the value **k=3** for the appropriate number of clusters. We calculated the **Silhouette Score** for this clustering to be **0.49**, which means they are well separated overall. The score of 0.49 suggests clustering algorithm has performed well in separating the data into distinct clusters but there is still room for improvement. We have also used synthetic points to test the clustering prediction as shown in the figure.





For the fitting we used Linear Regression to predict the relationship between these two variables. We took 10 random points and tested the fitting prediction by this, the test points are shown as green for testing this model's predictive performance. We also calculated the **R squared score** for this which came out to be **0.654**, which means **65.4%** of the variability in the dependent variable values can be accounted for by the independent variable in the model, and the rest could be attributed to other factors not included in the model or random noise.


Linear Regression: Salary vs Years of Experience
Original Data
Fitted Line
Test Points
Salary
Years of Experience