Assignment – 4

**Hareesh Muppu**

**700732514**

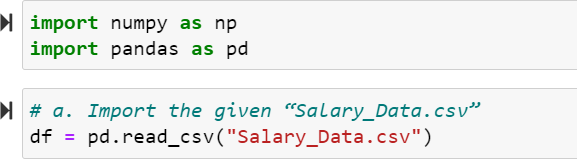
**Github link:** [**https://github.com/hareesh200/Demo**](https://github.com/hareesh200/Demo)

**Video link:** [**https://www.youtube.com/channel/UCIGk6FKMXmJt5ofU68tcNPQ**](https://www.youtube.com/channel/UCIGk6FKMXmJt5ofU68tcNPQ)

# Apply Linear Regression to the provided dataset using underlying steps.

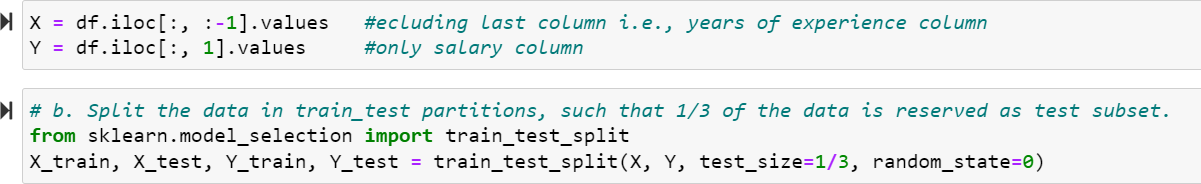
* 1. **Import the given “Salary\_Data.csv”**

Using the pandal library, I imported Salary\_Data csv file.



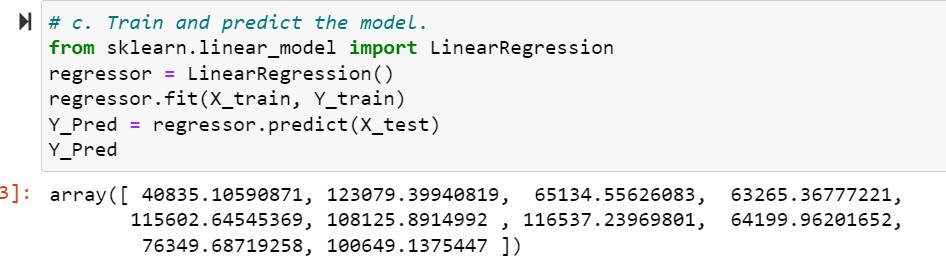
# Split the data in train\_test partitions, such that 1/3 of the data is reserved as test subset.

Using train\_test\_split method from sklearn library, we divided the data into training and testing data. By setting test\_size = 1/3 we can divide the data where the testing part contains 1/3rd of the data.



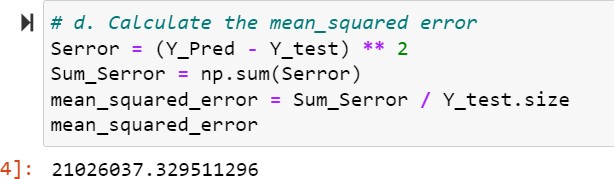
# Train and predict the model.

The relationship between the data points is used to predict the outcomes and that relationship can be represented through a line in linear regression algorithm. We trained linear regression algorithm on our training data using LinearRegression method of sklearn library and predicted the values.



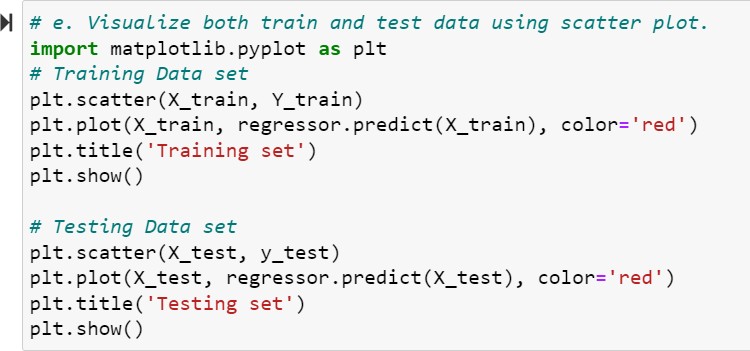
# Calculate the mean\_squared error

Mean squared error is the average squared error i.e., the average squared difference between the predicted values and the testing values.

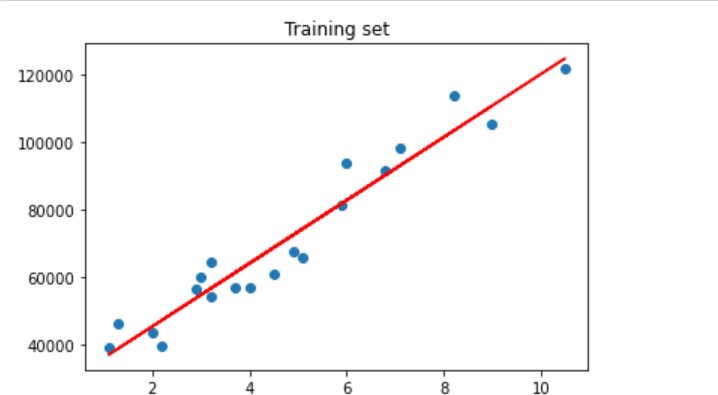


# Visualize both train and test data using scatter plot.

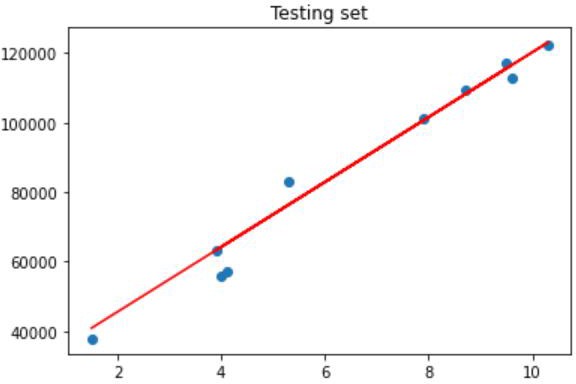
The scatter method of matplotlib library is used to represent the training and testing data visually through a scatter plot and plot method to draw a line between the data points.



**Testing data:**

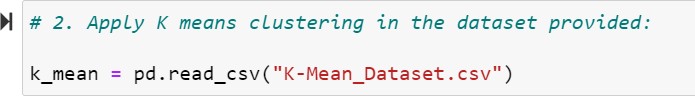


**Training data:**



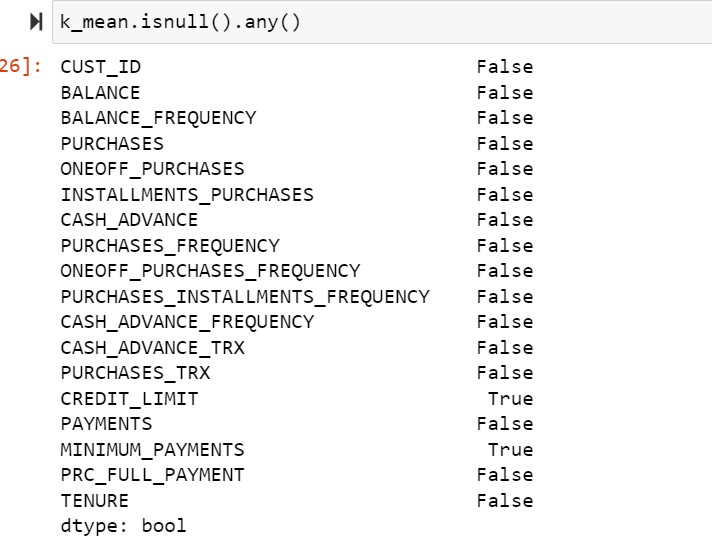
1. **Apply K means clustering in the dataset provided:**

Using the pandal library, I imported K-Mean\_Dataset csv file.

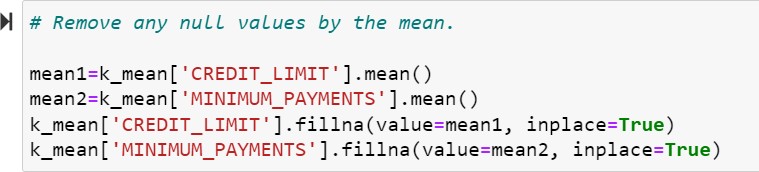


# Remove any null values by the means.

Using isnull method we can check that the dataset contains any null values.

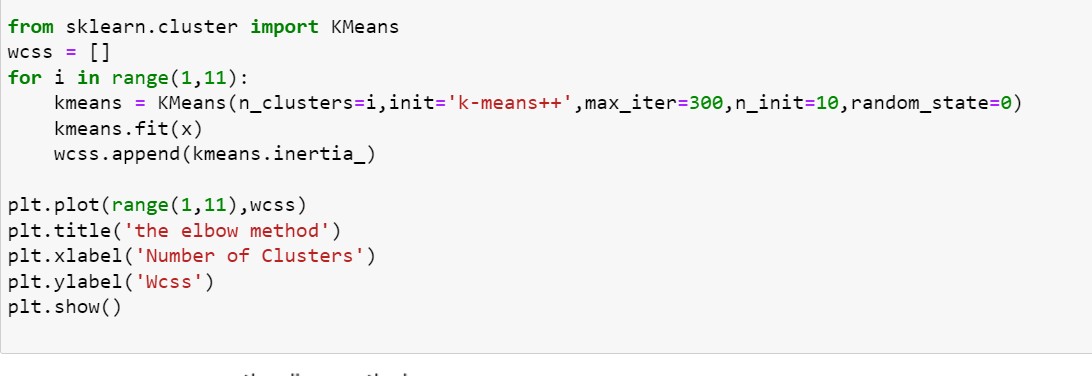


If any null values are present, we need to replace those values by means of individual columns using fillna method.

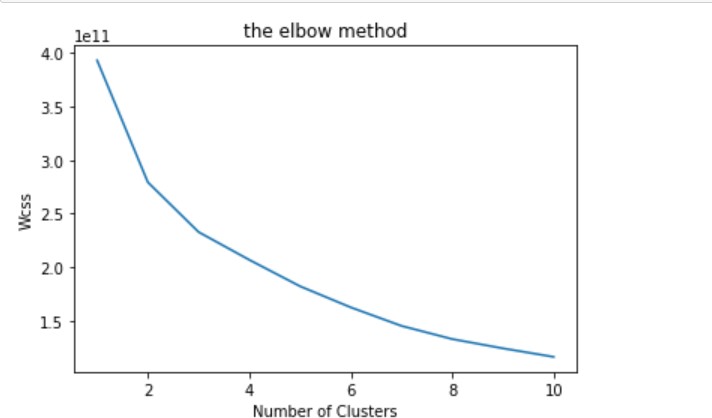


# Use the elbow method to find a good number of clusters with the K-Means algorithm

To find a good number of clusters required to fit our data together into clusters we have the elbow method. This elbow method results in optimal value of K. In K-Means algorithms, where K is the number of clusters. We can find the k value from the graph where the elbow point that the inertia values start decreases linearly.

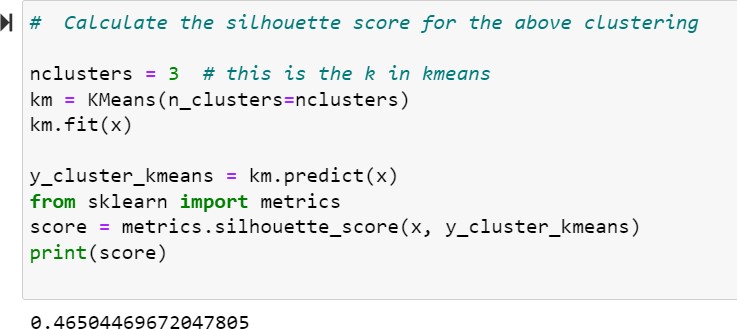


From the graph below, we can define the number of clusters required to train KMeans algorithm on data set. From the graph, we can see at elbow point 3, wcss values start decreases linearly. So, we need 3 clusters to fit out data into clusters.



# Calculate the silhouette score for the above clustering

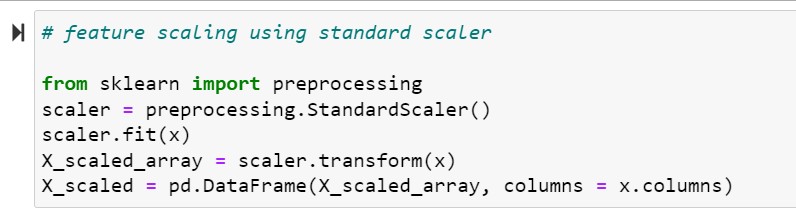
With k = 3 we can train KMeans algorithm on K-Mean\_Dataset and predit the values using KMeans method of sklearn library. Using the predicted values, we can find the silhouette score. Silhouette score is used to calculate how good the clustering technique is. Silhouette score is the difference between the point and the nearest cluster that the point is not part of the cluster.



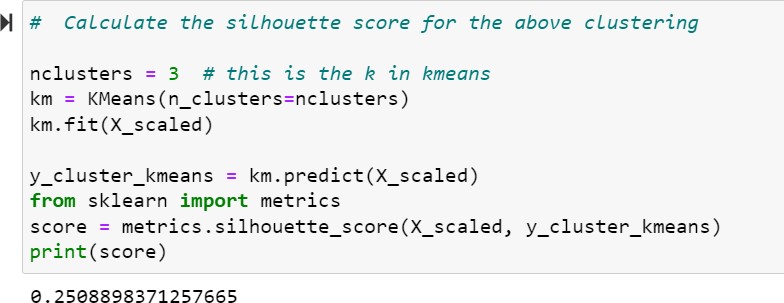
# Try feature scaling and then apply K-Means on the scaled features. Did that improve the Silhouette score? If yes, can you justify why

We can perform feature scaling using many ways i.e., min-max scaler, standard scaler, and robust scaler. Here, we are using standard scaler to perform feature scaling. Sklearn library contains a preprocessing module to preprocess our raw data. Using preprocessing module, we will perform feature scaling. Feature scaling is used to normalize the range of all features.

Initially some features contain a range in thousands or even hundreds. Among varies type of ranges, the features which have high range have some superiority over others.



Here we are training KMeans algorithm on the preprocessed data set and found the silhouette score for this model.



Even after preprocessing raw data, it does not improve the silhouette score. The silhouette score for this preprocessed data is less than the silhouette score for raw data.