**Data Mining Report(Clustering)**

The given dataset looks like the following:

Data columns (total 15 columns):

Age 4579 non-null int64

WorkClass 4579 non-null object

Fnlwght 4579 non-null int64

Education 4579 non-null object

EducationNumber 4579 non-null int64

MaritalStatus 4579 non-null object

Occupation 4579 non-null object

Relationship 4579 non-null object

Race 4579 non-null object

Sex 4579 non-null object

CapitalGain 4579 non-null int64

CapitalLoss 4579 non-null int64

HoursPerWeek 4579 non-null int64

NativeCountry 4579 non-null object

Class 4579 non-null object

dtypes: int64(6), object(9)

The **Class** attribute is the label.

All the null values and the duplicates in the dataset are removed. The unknown values in the dataset are marked as ‘?’ . All the rows which contains ‘?’ in any of the column are removed.

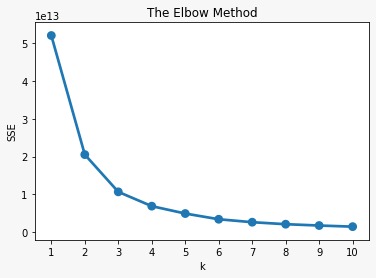
The shape of the dataset is now: (4579,15)

The dataset contains categorical features. we need to encode the categorical features to numeric quantities. Hence, one-hot encoding is performed on the dataset and the shape of the dataframe is now(4579,102) .

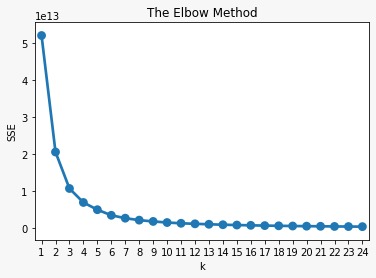
1-a)

To find the right number of clusters we use elbow method which lets us identify the appropriate k after plotting a graph.

When I initially choose k-value between 1 and 10 , the best k using elbow method is suggested as 3.

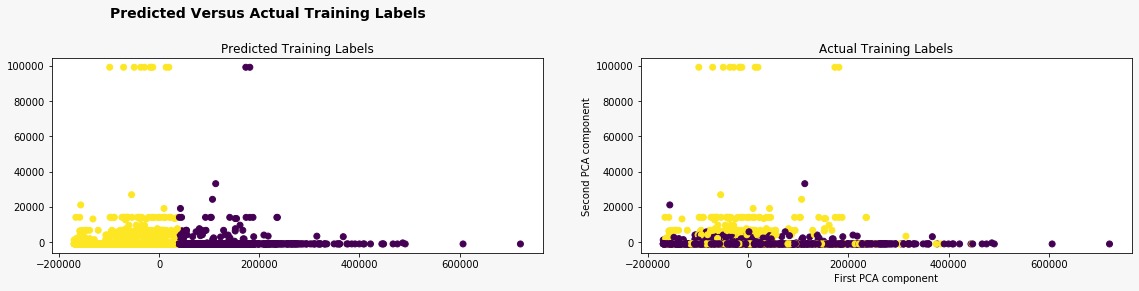


When I choose k-value between 1 and 10 , the best k using elbow method is suggested as 3 as well but this time it is very clear as 3. Finally, when you put the value as the number of rows SSE becomes zero as each row is grouped into a different cluster and the classification error is 0.



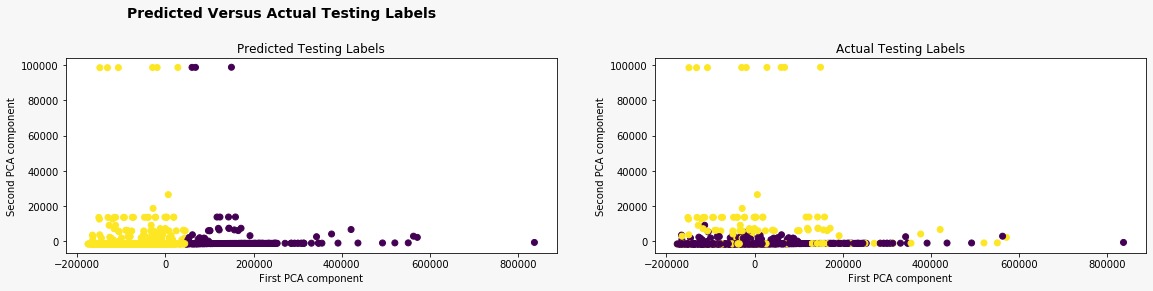
**1-b-2**

The features are reduced to two in order to plot a scatter plot using Principal Component Analysis. When we plot the graph of first component vs second component it looks like the following. We observe that it did not perform well and also the confusion matrix confirms with many number of missclassifications.

****

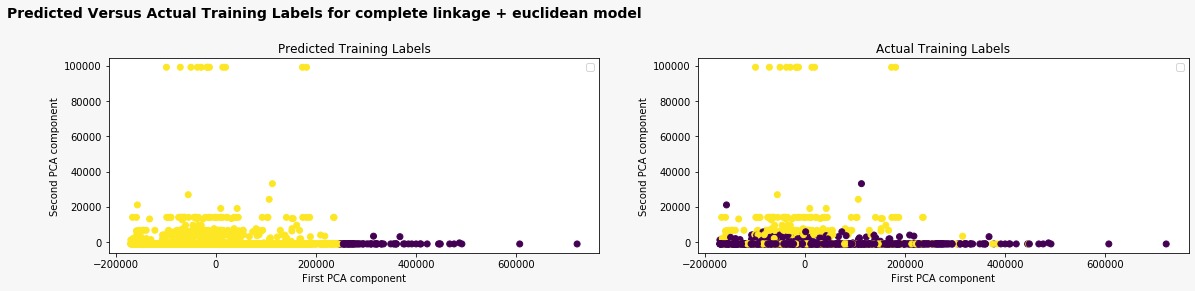
**1-b-3**

This model performed well when compared to the above model which is for test data. And the confusion matrix also confirms the same . Almost 78% of the classes are correctly classified.

****

**2 a)** When I calculated F-1 score for all combinations of affinity and linkage for hierarchial Agglomerative clustering, affinity=eucledian, linkage = complete has the highest value which is 0.4. Although the accuracy is higher for affinity=manhattan and linkage=’average’.

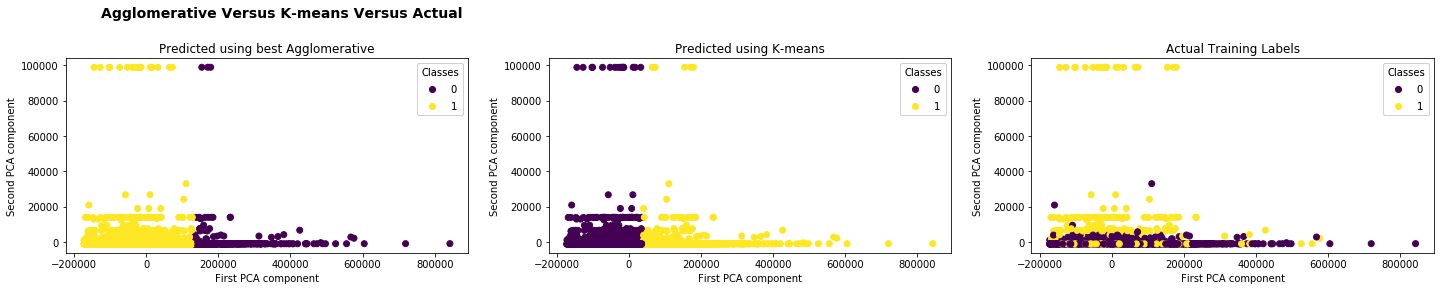
**2-b**

****

By looking at the graph we can say that our model did not perform well. By looking at the confusion matrix we observed that one class label is highly misclassified . And hence the graph is dominant with only one class label.

This approach is not good.

**3-a**

****

As you can see not much has changed in the performance when the entire dataset is used. However, from the graph Agglomerative looks better when compared to K-means. Lets check this by classification matrix in the below section.

**Task3b:**

K means model and Agglomerative clustering model are built for the given data set. After predicting the labels for the training data, we have calculated Accuracy, Precision, F1-Score, Recall for both the models.

The Accuracy for K-means clustering is greater than that of best Agglomerative method. Also it has greater values of Recall and F1-score . But it has slightly less precision score than Agglomerative. Hence, we can conclude that K-Means **is the best clustering model for our data set.**