**K - Means Clustering**

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Shreyas A B**

**Batch Id: 05012021-10AM**

**Topic: K Means Clustering**

**1. Business Problem**

* 1. **Objective**
  2. **Constraints (if any)**

**2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**Using R and Python codes perform:**

**3. Data Pre-processing**

**2.1 Data Cleaning, Feature Engineering, etc.**

**4. Exploratory Data Analysis (EDA):**

**4.1. Summary**

**4.2. Univariate analysis**

**4.3. Bivariate analysis**

**5. Model Building**

**5.1 Build the model on the scaled data (try multiple options)**

**5.2 Perform the K- means clustering, visualize the clusters using scree plot**

**5.3 Validate the clusters (try with different no. of clusters) – label the clusters and derive insights (compare the results from multiple approaches)**

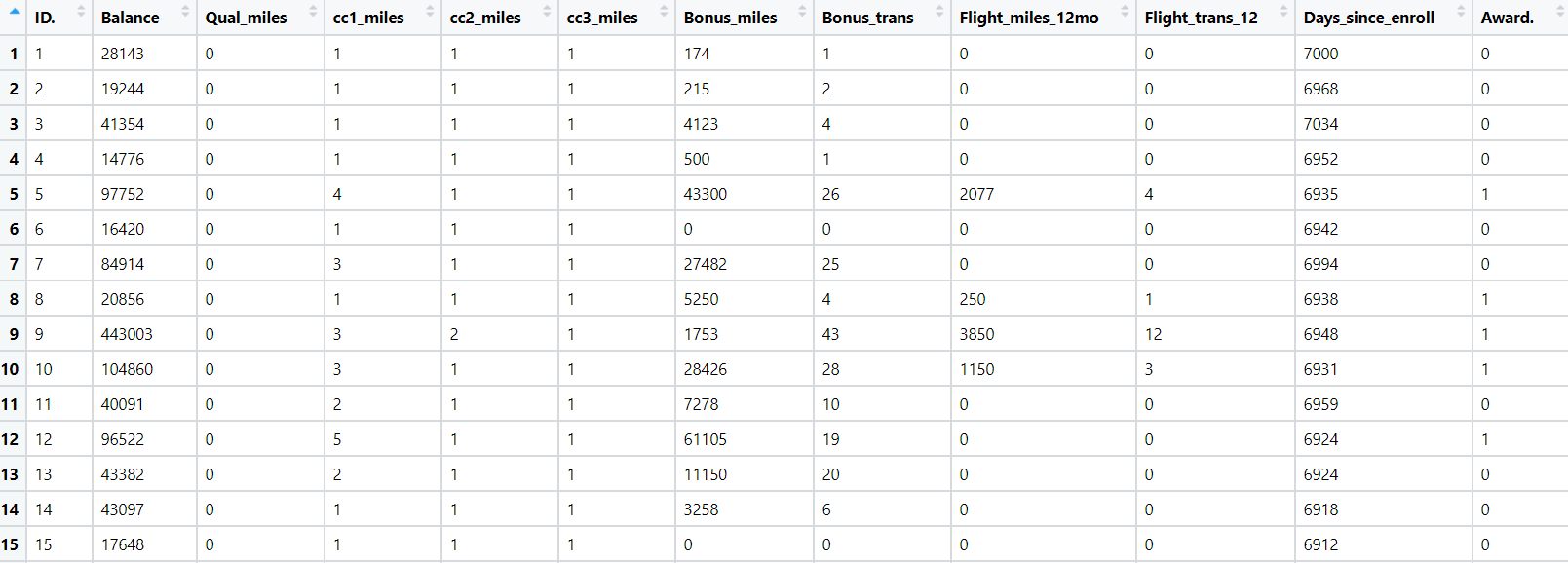
**6. Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

**Note:**

The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the modules (elaborating on steps mentioned above)

1. Perform clustering (K means clustering) for the airlines data to obtain optimum number of clusters. Draw the inferences from the clusters obtained. Refer to EastWestAirlines.xlsx dataset.



**Business Problem**

**Objective maximize the profit**

Python code steps :\_

Ans = Data Frame name is air. It has 3999 rows and 12 columns.

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important ,except ‘ID#”.

Dataframe has no missing values . ‘ID#’ Index irrelevant index does not provide any useful information.so we create a new data frame without “ID#” columns named as air\_1 . We also check for unique values for each columns named as “col\_uni” . . now we do exploratory data analysis for whole data frame and saved the details as “EDA”.

For graphical representation we have done scatter and histogram using seaborn libarary .

for continuous data we have done min max scaling and for discrete data we performed onehot encoding and saved as air\_norm .

after that we go for scree plot(elbow curve ) where k range from 3 to 9 on normalized data (air\_norm).and look for clusters ,so we can check for cluster number .

After we have seen elbow curve we go for 4 cluster number and check the details of each cluster .

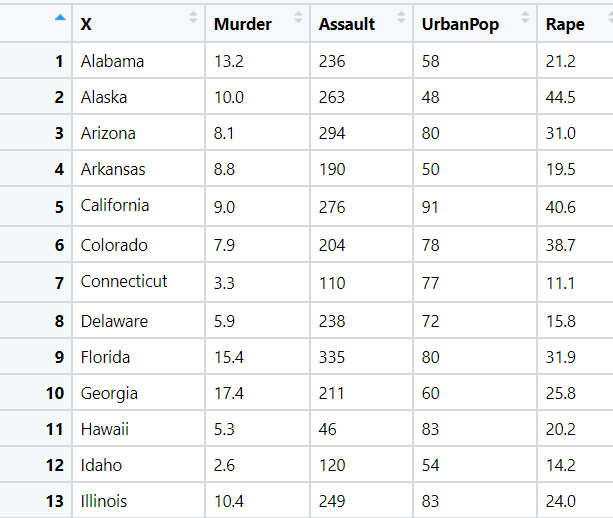
we go with 4 cluster and save the details as clust4\_details.

After looking of all cluster details we go with 4 cluster and save the data as ‘air\_final.’

And cluster details as ‘fclust\_details’

Used library:- Pandas , matplotlib, Seaborn, sklearn , scipy

1. Perform clustering for the crime data and identify the number of clusters formed and draw inferences. Refer to crime\_data.csv dataset.



**Business Problem**

**Objective minimize crime**

Python code step :-

Ans = Data Frame name is ‘crime\_df’. It has 50 entries and 5 columns.

Dataframe has no missing values . We also check for unique values for each columns named as “col\_uni” .

now we do exploratory data analysis for whole data frame and saved the details as “EDA”.

For graphical representation we have done scatter and histogram using seaborn libarary .

for continuous data we have done min max scaling and for 1st column has no name so we named as “state” and this is nominal data with n unique value .(where n is number of rows .)

after that we go for scree plot(elbow curve ) where k range from 2 to 6 on normalized data (crime\_norm).and look for clusters ,so we can check for cluster number .

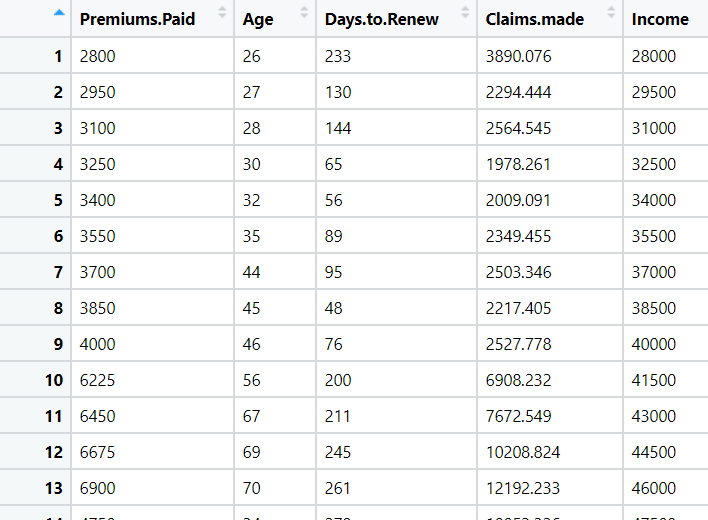
After we have seen scree plot we go for 4 cluster number and check the details of each cluster .

we go with 4 cluster and save the details as clust4\_details.

After looking of all cluster details we go with 4 cluster and save the data as ‘crime\_final.’

And cluster details as ‘fclust\_details’

1. Analyze the information given in the following ‘Insurance Policy dataset’ to create clusters of persons falling in the same type. Refer to Insurance Dataset.csv

****

**Business Problem**

**Objective maximize profit**

Python code steps :\_

Ans = Data Frame name is insurance. It has 100 rows and 5 columns.

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

Dataframe has no missing values .

We also check for unique values for each columns named as “col\_uni” . . now we do exploratory data analysis for whole data frame and saved the details as “EDA”.

For graphical representation we have done scatter and histogram using seaborn libarary .

for continuous data we have done min max scaling and saved as insurance\_norm .

after that we go for scree plot(elbow curve ) where k range from 2 to 6 on normalized data (insurance\_norm).and look for clusters ,so we can check for cluster number .

After we have seen elbow curve we go for 4 cluster number and check the details of each cluster .

1st we go with 4 cluster and save the details as clust4\_details.

After looking of all cluster details we go with 4 cluster and save the data as ‘insurance\_final.’

And cluster details as ‘fclust\_details’

Used library:- Pandas , matplotlib, Seaborn, sklearn , scipy

1. Perform clustering analysis on the telecom dataset. The data is a mixture of both categorical and numerical data. It consists the number of customers who churn. Derive insights and get possible information on factors that may affect the churn decision. Refer to Telco\_customer\_churn.xlsx dataset.

Hint:

* Perform EDA and remove unwanted columns.
* Use Gower dissimilarity matrix and In R use daisy() function.



**Business Problem**

**Objective minimize the churn rate**

Ans = Data Frame name is tel. It has 7043 rows and 30 columns.

Then we create a data frame that’s contain details of each column’s data types ,and save the details named as data\_details .all of them are important ,except "Customer ID","Count",'Quarter'.

Dataframe has no missing values . ‘"Customer ID","Count",'Quarter' columns irrelevant columns does not provide any useful information.so we create a new data frame without ‘"Customer ID","Count",'Quarter' columns named as tel\_1 . We also check for unique values for each columns named as “col\_uni” . . now we do exploratory data analysis for whole data frame and saved the details as “EDA”.

For graphical representation we have done scatter and histogram using seaborn libarary and also boxplot for continuous columns .

for continuous data we have done min max scaling and for discrete data we performed onehot encoding and saved as tel\_norm .

after that we go for scree plot on normalized data (tel\_norm).and look for clusters ,so we can check for cluster number .

After we have seen elbow curve. we go for 8 cluster number and check the details of each cluster .

we go with 8 cluster and save the details as clust8\_details.

After looking of all cluster details we go with 8 cluster and save the data as ‘tel\_final.’

And cluster details as ‘fclust\_details’

Used library:- Pandas , matplotlib, Seaborn, sklearn , scipy

1. Perform clustering on mixed data convert the categorical variables to numeric by using dummies or Label Encoding and perform normalization techniques. The data set consists details of customers related to auto insurance. Refer to Autoinsurance.csv dataset.



**Business Problem**

**Objective maximize profit**

Data Frame name is insu. It has 9134 rows and 24 columns.

Then we create a data frame that’s contain details of each column’s data types ,and save the details named as data\_details .all of them are important ,except "Customer”.

Dataframe has no missing values . ‘"Customer column irrelevant columns does not provide any useful information.so we create a new data frame without ‘"Customer “ columns named as insu\_1 . We also check for unique values for each columns named as “col\_uni” . .

now we do exploratory data analysis for whole data frame and saved the details as “EDA”.

For graphical representation we have done scatter and histogram using seaborn libarary and also boxplot for continuous columns .

for continuous data we have done min max scaling and for discrete data we performed onehot encoding and saved as insu\_norm .

after that we go for scree plot on normalized data (insu\_norm).and look for clusters ,so we can check for cluster number .

After we have seen elbow curve we go for 7 cluster number and check the details of each cluster .

we go with 7 cluster and save the details as clust7\_details.

After looking of all cluster details we go with 7 cluster and save the data as ‘insu\_final.’

And cluster details as ‘fclust\_details’

Used library:- Pandas , matplotlib, Seaborn, sklearn , scipy