**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: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Batch Id: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**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)**

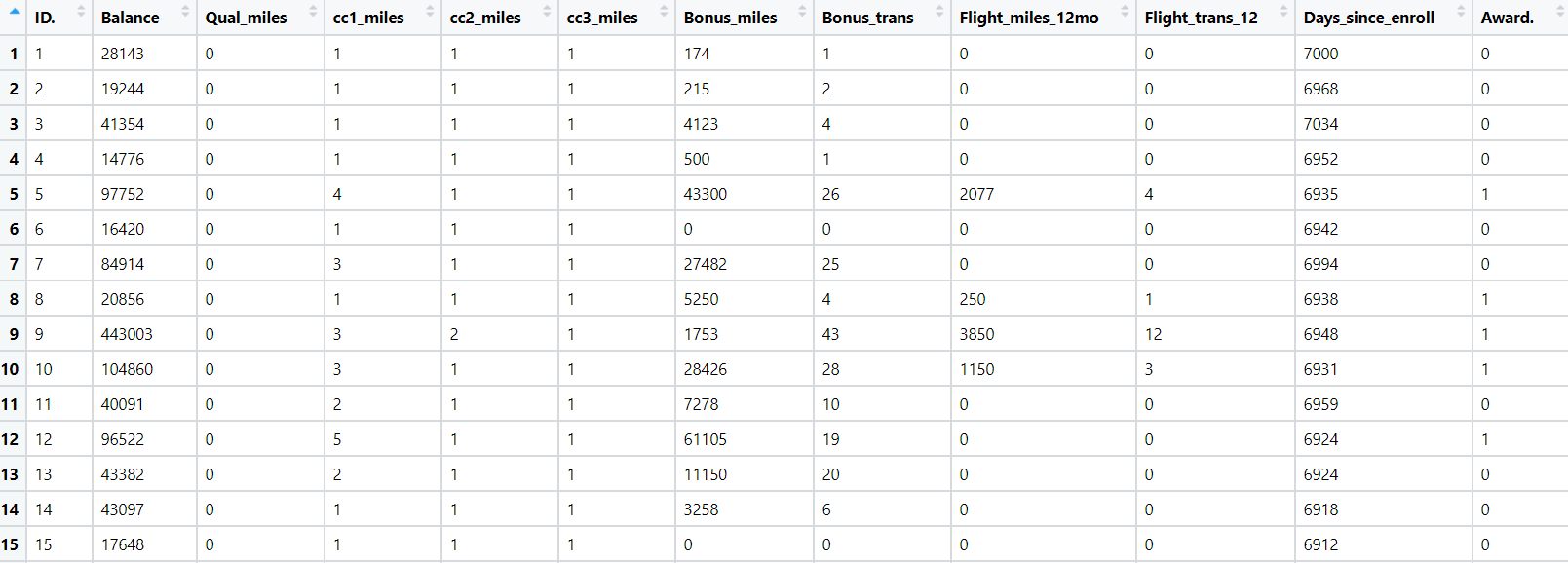
**6. Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

**Note:**

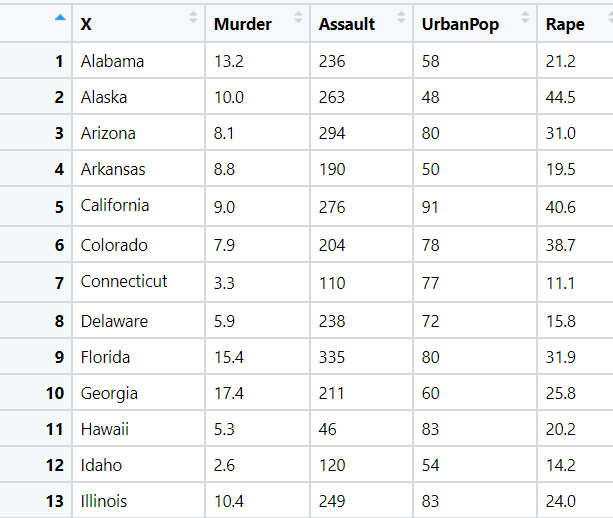
The assignment should be submitted in the following format:

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

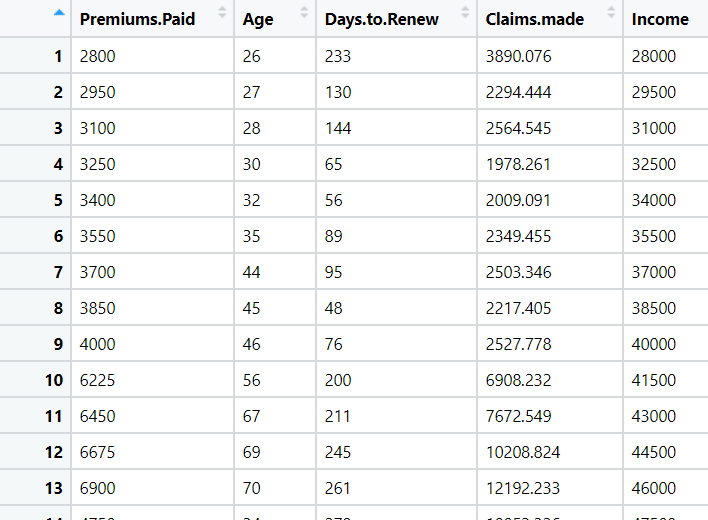
1. Perform clustering (K means clustering) for the airlines data to obtain optimum number of clusters. Draw the inferences from the clusters obtained. Refer to EastWestAirlines.xlsx dataset.



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



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

****

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.



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.

