**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: Shahina Athar**

**Batch Id: 10122020**

**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)**

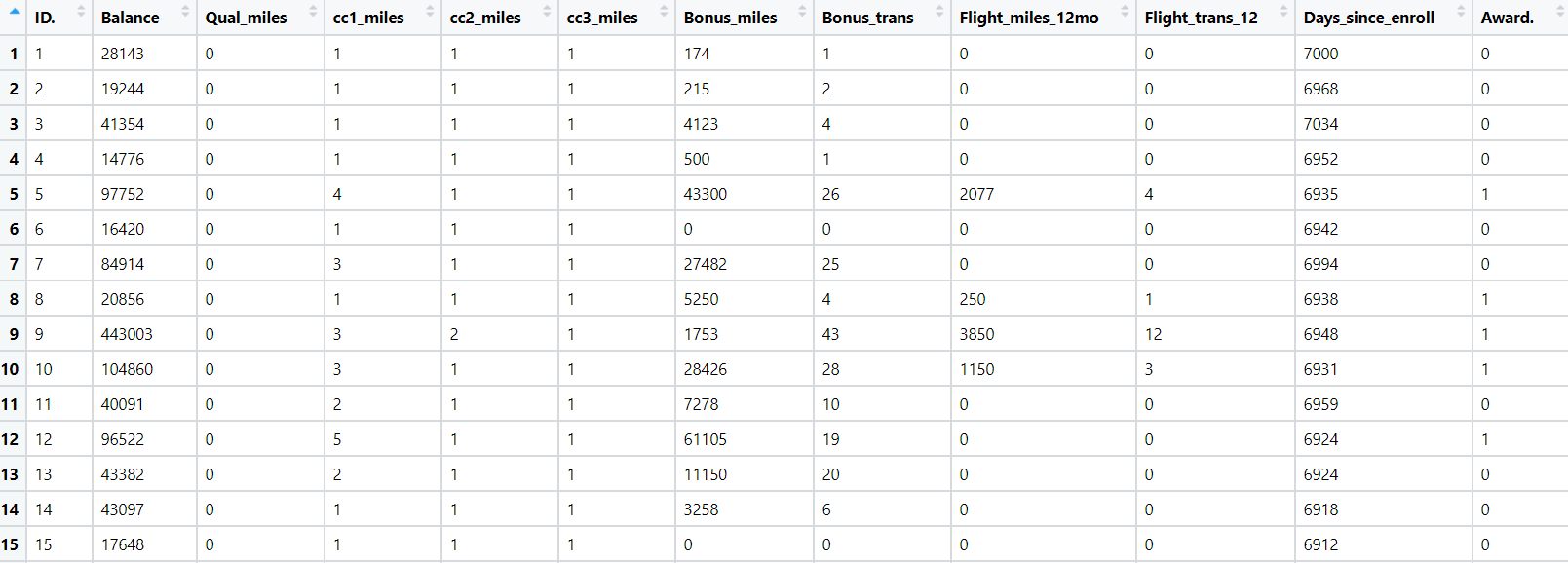
**6. Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

**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.



**Answer: 1. Business Problem**

**Objective:** We are trying to learn more about EastWest Airlines’ customers

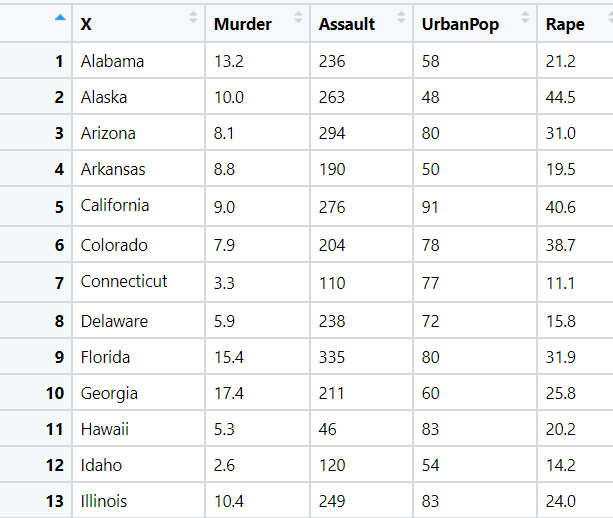
based on their flying patterns, earning and use of frequent flyer

rewards, and use of the airline credit card. The primary objective is to identify customer segments via K-Means clustering and design targeted marketing campaigns for each segment.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| ID# | Unique ID | Quantitative,Discrete, Ordinal | Irrelevant |
| Balance | Number of miles eligible for award travel | Quantitative,Discrete,  Count | Relevant |
| Qual\_miles | Number of miles counted as qualifying for Topflight status | Quantitative,Discrete,  Count | Relevant |
| cc1\_miles | Has member earned miles with airline freq. flyer credit card in the past 12 months (1=Yes/0=No)? | Quantitative,Discrete, Nominal | Relevant |
| cc2\_miles | Has member earned miles with Rewards credit card in the past 12 months (1=Yes/0=No)? | Quantitative,Discrete, Nominal | Relevant |
| cc3\_miles | Has member earned miles with Small Business credit card in the past 12 months (1=Yes/0=No)? | Quantitative,Discrete, Nominal | Relevant |
| Bonus\_miles | Number of miles earned from non-flight bonus transactions in the past 12 months | Quantitative,Discrete,  Count | Relevant |
| Bonus\_trans | Number of non-flight bonus transactions in the past 12 months | Quantitative,Discrete,  Count | Relevant |
| Flight\_miles\_12mo | Number of flight miles in the past 12 months | Quantitative,Discrete,  Count | Relevant |
| Flight\_trans\_12 | Number of flight transactions in the past 12 months | Quantitative,Discrete,  Count | Relevant |
| Days\_since\_enroll | Number of Days since Enrollement | Quantitative,Discrete,  Count | Relevant |
| Award? | Award:  Yes or No /0 0r 1 | Quantitative,Discrete,  Binary | Relevant |

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



**Ans:**

1. **Business Problem:**

**Objective:** We are trying to learn more about Crime rate based on the states of U.S, Basically we are trying to find out the crime rate using K-Means clustering. The primary objective is to identify Crime rates segments via K-Means clustering and design targeted marketing campaigns for each segment.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| Sates | State name of U.S. | Qualitative, Discrete, Nominal | Relevant |
| Murder | |  |  | | --- | --- | | numeric | Murder arrests | | Quantitative, Continuous, Ratio | Relevant |
| Assault | |  |  | | --- | --- | | numeric | Assault arrests | | Quantitative, Discrete, Count | Relevant |
| UrbanPOp | numeric Percent urban population | Quantitative, Discrete, Count | Relevant |
| Rape | |  |  | | --- | --- | | numeric | Rape arrests | | Quantitative, Continuous, Ratio | Relevant |

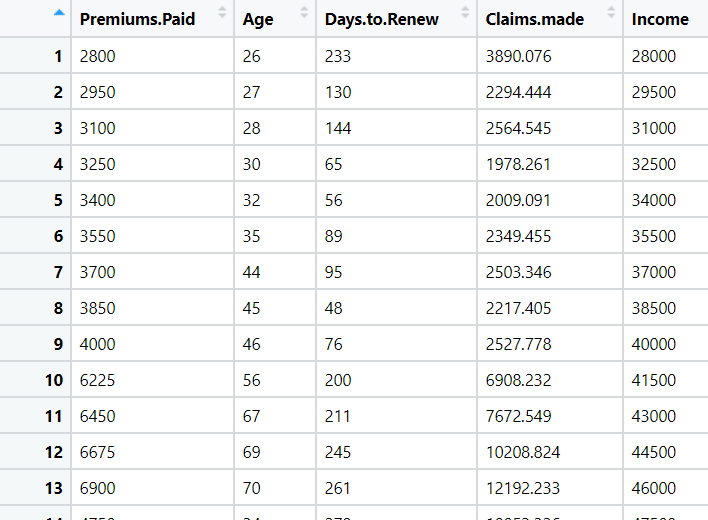
1. Analyze the information given in the following ‘Insurance Policy dataset’ to create clusters of persons falling in the same type. Refer to Insurance Dataset.csv

**1. Business Problem:**

**Objective:** We are trying to learn more about Person based on the Age, Days to Renew, Claims made, Income, Premiums Paid. Basically we are trying to find out the Person Pattern using K-Means clustering. The primary objective is to identify Person Pattern segments via K-Means clustering and design targeted marketing campaigns for each segment.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| Premiums Paid | Preminums Paid Amount | Quantitative, Continuous, Ratio | Relevant |
| Age | Age of the Person | Quantitative, Continuous, Ratio | Relevant |
| Days to Renew | Number of days to Renew | Quantitative, Continuous, Ratio | Relevant |
| Claims made | The number Claims made | Quantitative, Continuous, Ratio | Relevant |
| Income | Income of the Person | Quantitative, Continuous, Ratio | Relevant |

****

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. **Business Problem:**

**Objective:** Customer churn, also known as customer attrition, occurs when customers stop doing business with a company or stop using a company’s services. By being aware of and monitoring churn rate, companies are equipped to determine their customer retention success rates and identify strategies for improvement. We will use a machine learning model to understand the precise customer behaviors and attributes.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| **Customer ID** | Unique ID | Quantitative,Discrete, Ordinal | Irrelevant |
| **Count** | Number of Count | Quantitative,Discrete,  Count | Irrelevant |
| **Quarter** | Q3 Quarter | Qualitative,Discrete,  Nominal | Irrelevant |
| **Referred a Friend** | Is customer Referred a Friend  (yes or no) | Quanlitative,Discrete, Nominal | Relevant |
| **Number of Referrals** | It shows number of Referrals | Quantitative,Discrete, Count | Relevant |
| **Tenure in Months** | Number of months the customer has stayed with the company | Quantitative,Discrete, Count | Relevant |
| **Offer** | Which offers customer has | Qualitative,Discrete,  Ordinal | Relevant |
| **Phone Service** | Whether the customer has Phone Service or not (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Avg Monthly Long Distance Charges** | How much Avg Monthly Long Distance Charges of the customers | Quantitative,Continuous,  Ratio | Relevant |
| **Multiple Lines** | Whether the customer has Multiple Lines or not (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Internet Service** | Whether the customer has Internet Service or not  (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Internet Type** | Customer’s internet service provider (DSL, Fiber optic, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Avg Monthly GB Download** | How much Avg Monthly GB Download by customer | Quantitative,Discrete,  Count | Relevant |
| **Online Security** | Whether the customer has Online Security or not  (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Online Backup** | Whether the customer has Online Backup or not  (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Device Protection Plan** | Whether the customer has Device Protection Plan or not  (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Premium Tech Support** | Whether the customer has Preminum Tech Support or not  (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Streaming TV** | Whether the customer has  Streaming TV or not  (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Streaming Movies** | Whether the customer has Streaming Movies or not  (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Streaming Music** | Whether the customer has Streaming Music or not  (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Unlimited Data** | Whether the customer has Unlimited Data or not  (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Contract** | How many months, years of contract of the customers | Qualitative,Discrete,  Ordinal | Relevant |
| **Paperless Billing** | Whether the customer has Paperless Billing or not  (Yes, No) | Qualitative,Discrete,  Nominal | Relevant |
| **Payment Method** | Payment method of a customer | Qualitative,Discrete,  Nominal | Relevant |
| **Monthly Charge** | How much Monthly Charge of the customers | Quantitative,Continuous,  Ratio | Relevant |
| **Total Charges** | How much Total Charge of the customers | Quantitative,Continuous,  Ratio | Relevant |
| **Total Refunds** | How much Total Refund of the customers | Quantitative,Continuous,  Ratio | Relevant |
| **Total Extra Data Charges** | How much Total Extra Data Charges of the customers | Quantitative,Continuous,  Ratio | Relevant |
| **Total Long Distance Charges** | How much Total Long Distance Charges of the customers | Quantitative,Continuous,  Ratio | Relevant |
| **Total Revenue** | How much Total Revenue of the customers | Quantitative,Continuous,  Ratio | Relevant |

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.
2. **Business Problem:**

**Objective:** We are trying to learn more about Auto insurance pattern on the basis of various features, Basically we are trying to find out the auto insurance pattern using K-Means clustering. The primary objective is to identify Auto insurance pattern segments via K-Means clustering and design targeted marketing campaigns for each segment.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| **Customer** | Unique Customer Id | Qualitative, Discrete, Nominal | Irrelevant |
| **State** | Names of States | Qualitative, Discrete, Nominal | Relevant |
| **Customer Lifetime Value** | Value of Customer Lifetime | Quantitative, Continuous,  Ratio | Relevant |
| **Response** | Response  (Yes/No) | Qualitative, Discrete, Nominal | Relevant |
| **Coverage** | Coverage  (Basic, Extended,Premium) | Qualitative, Discrete, Ordinal | Relevant |
| **Education** | Education  (Status) | Qualitative, Discrete, Ordinal | Relevant |
| **Effective To Date** | Dates | Qualitative, Discrete, Nominal | Irrelevant |
| **EmploymentStatus** | Shows Employment Status | Qualitative, Discrete, Nominal | Relevant |
| **Gender** | Shows Gender status | Qualitative, Discrete, Nominal | Relevant |
| **Income** | It shows How much income of the customers | Quantitative, Continuous,  Ratio | Relevant |
| **Location Code** | It shows code location of the area | Qualitative, Discrete, Nominal | Relevant |
| **Marital Status** | It shows Marital Status | Qualitative, Discrete, Nominal | Relevant |
| **Monthly Premium Auto** | Number of Monthly Premium Auto | Quantitative, Discrete, Count | Relevant |
| **Months Since Last Claim** | Number of Months Since Last Claim | Quantitative, Discrete, Count | Relevant |
| **Months Since Policy Inception** | Number of Months Since Policy Inception | Quantitative, Discrete, Count | Relevant |
| **Number of Open Complaints** | Number of open complains | Quantitative, Discrete, Count | Relevant |
| **Number of Policies** | Number of Policies | Quantitative, Discrete, Count | Relevant |
| **Policy Type** | It shows Policy Type | Qualitative, Discrete,  Nominal | Relevant |
| **Policy** | Policy  (Corporate L3,  Personal L3….) | Qualitative, Discrete,  Nominal | Relevant |
| **Renew Offer Type** | Renew Offer Types | Qualitative, Discrete,  Ordinal | Relevant |
| **Sales Channel** | Various Sales Channel | Qualitative, Discrete,  Nominal | Relevant |
| **Total Claim Amount** | Total Claim Amount | Quantitative, Continuous,  Ratio | Relevant |
| **Vehicle Class** | Vehicle Class (types) | Qualitative, Discrete,  Nominal | Relevant |
| **Vehicle Size** | Vehicle Size | Qualitative, Discrete,  Ordinal | Relevant |

