**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: Nukala Ayyappa Bharthwaj**

**Batch Id: DSWDMCOS 21012022**

**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.

**Sol:**

**Business Objective:** to perform K means clustering on the east west airlines data set

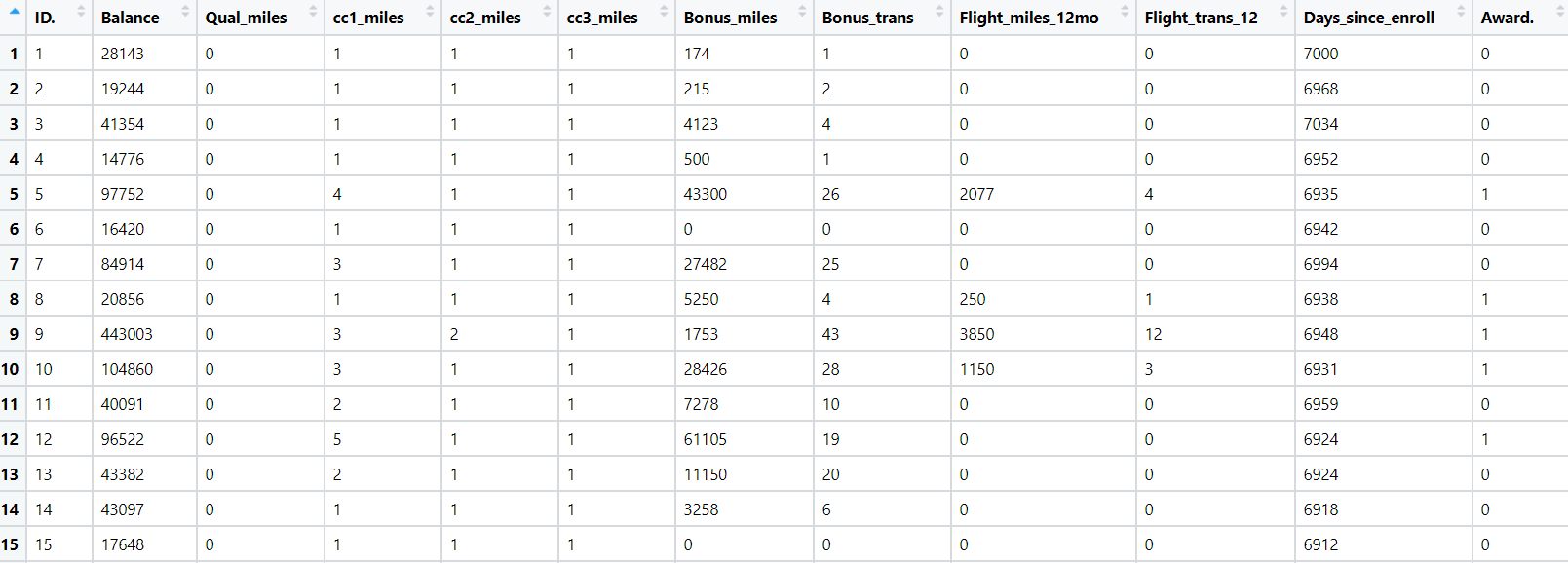
**Data Types:** the given data and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| ID | Id of the flight | Ordinal | Irrelevant since it’s a id of the flight |
| Balance | Balance amount | Ratio | Relevant |
| Qual\_miles | Number of miles travelled | Ratio | Relevant |
| cc1\_miles | CC1 miles of flight | Ratio | Relevant |
| Cc2\_miles | CC2 miles of flight | Ratio | Relevant |
| Cc3\_miles | CC3 miles of flight | Ratio | Relevant |
| Bonus\_miles | Bonus miles of the flight | Ratio | Relevant |
| Bonus\_trans | Bonus transc. Of flight | Ratio | Relevant |
| Flight\_miles\_12mo | 12 months flight miles | Ratio | Relevant |
| Flight\_trans\_12 | 12months flight transc. | Ratio | Relevant |
| Days\_since\_enroll | Number of days since flight enrolled | Ratio | Relevant |
| Award? | Whether it got award or not | Nominal | Relevant |

**Data Pre Processing:** all the variables in the given data is used for applying the clustering except the ID column, since it is not useful.

**Exploratory Data Analysis:** mean, median, variance, standard deviation, skewness, kurtosis is calculated for all the variables of the given data then normalization for all the variables of the data to apply k Means clustering.

**k-Means clustering:** after cleaning the complete data K means clustering is applied on the data and number of clusters is in front decided as 3 and based on that the complete data is clustered.



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

**Sol:**

**Business Objective:** to perform K means clustering on the crime data set

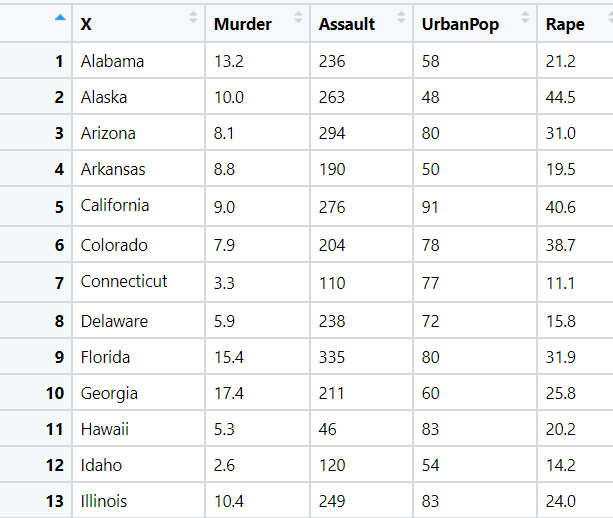
**Data Types:** the given data and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| X | Name of the murderer | Nominal | Irrelevant since name of the person |
| Murder | Murder rate | Ratio | Relevant |
| Assult | Assault rate | Ratio | Relevant |
| UrbanPop | Urban pop rate | Ratio | Relevant |
| Rape | Rape rate | Ratio | Relevant |

**Data Pre Processing:** all the variables in the given data is used for applying the clustering except the X column, since it is not useful.

**Exploratory Data Analysis:** mean, median, variance, standard deviation, skewness, kurtosis is calculated for all the variables of the given data then normalization for all the variables of the data to apply k Means clustering.

**k-Means clustering:** after cleaning the complete data K means clustering is applied on the data and number of clusters is in front decided as 4 and based on that the complete data is clustered.



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

**Sol:**

**Business Objective:** to perform K means clustering on the insurance data set

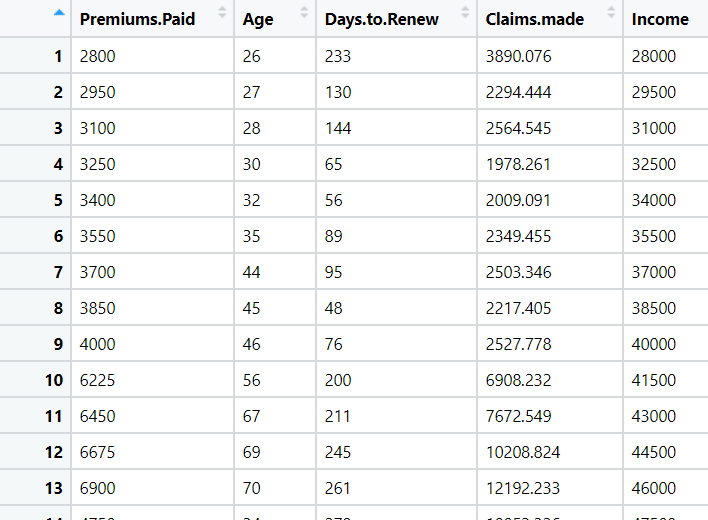
**Data Types:** the given data and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| Premiums Paid | Premiums paid by customer | Ratio | Relevant |
| Age | Age of customer | Ratio | Relevant |
| Days to Renew | Number of days to renew | Ratio | Relevant |
| Claims made | Claim amount made | Ratio | Relevant |
| Income | Income of customer | Ratio | Relevant |

**Data Pre Processing:** all the variables in the given data is used for applying the clustering.

**Exploratory Data Analysis:** mean, median, variance, standard deviation, skewness, kurtosis is calculated for all the variables of the given data then normalization for all the variables of the data to apply k Means clustering.

**k-Means clustering:** after cleaning the complete data K means clustering is applied on the data and number of clusters is in front decided as 3 and based on that the complete data is clustered.

****

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.



**Sol:**

**Business Objective:** to perform K means clustering on the tele customer data set

**Data Types:** the given data and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| Customer ID | Id of the customer | Nominal | Irrelevant |
| Count | Count number | Nominal | Relevant |
| Quarter | Type of quarter | Nominal | Relevant |
| Referred a Friend | Whether yes or no | Nominal | Relevant |
| Number of Referrals | Number of referals | Internal | Relevant |
| Tenure in Months | Tenure of the package | Ratio | Relevant |
| Offer | Type of offer given | Nominal | Relevant |
| Phone Service | Whether yes or no | Nominal | Relevant |
| Avg Monthly Long Distance Charges | Long distance charges for the udser | Ratio | Relevant |
| Multiple Lines | Whether yes or no | Nominal | Relevant |
| Internet Service | Whether yes or no | Nominal | Relevant |
| Internet Type | Type of internet | Nominal | Relevant |
| Avg Monthly GB Download | It is total GB downloaded in month | Ratio | Relevant |
| Online Security | Whether yes or no | Nominal | Relevant |
| Online Backup | Whether yes or no | Nominal | Relevant |
| Device Protection Plan | Whether yes or no | Nominal | Relevant |
| Premium Tech Support | Whether yes or no | Nominal | Relevant |
| Streaming TV | Whether yes or no | Nominal | Relevant |
| Streaming Movies | Whether yes or no | Nominal | Relevant |
| Streaming Music | Whether yes or no | Nominal | Relevant |
| Unlimited Data | Whether yes or no | Nominal | Relevant |
| Paperless Billing | Whether yes or no | Nominal | Relevant |
| Payment Method | Type of payment | Nominal | Relevant |
| Monthly Charge | Monthly charge to user | Ratio | Relevant |
| Total Charges | Total charge to user | Ratio | Relevant |
| Total Refunds | Total refunds to user | Ratio | Relevant |
| Total Extra Data Charges | Total extra data charge to user | Ratio | Relevant |
| Total Long Distance Charges | Total long distance chrge to user | Ratio | Relevant |
| Total Revenue | Total revenue to user | Ratio | Relevant |

**Data Pre Processing:** all the variables in the given data is used for applying the clustering except the customer id and the columns which are non numeric are converted into numeric data using dummy variables.

**Exploratory Data Analysis:** mean, median, variance, standard deviation, skewness, kurtosis is calculated for all the variables of the given data then normalization for all the variables of the data to apply k Means clustering.

**k-Means clustering:** after cleaning the complete data K means clustering is applied on the data and number of clusters is in front decided as 4 and based on that the complete data is clustered.

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.

**Sol:**

**Business Objective:** to perform K means clustering on the auto insurance data set

**Data Types:** the given data and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| Customer | Id of the customer | Nominal | Irrelevant |
| State | State of the customer | Nominal | Relevant |
| Customer Lifetime Value | Life time value of the customer | Ratio | Relevant |
| Response | Whether yes or no | Nominal | Relevant |
| Coverage | Type of coverage | Nominal | Relevant |
| Education | Education of cust. | Nominal | Relevant |
| Effective To Date | Insurance effective date | Nominal | Relevant |
| EmploymentStatus | Employment status of cust | Nominal | Relevant |
| Gender | Gender of cust. | Nominal | Relevant |
| Income | Income of the cust. | Ratio | Relevant |
| Location Code | Location code of the cust. | Nominal | Relevant |
| Marital Status | Marital status of cust. | Nominal | Relevant |
| Monthly Premium Auto | Monthly premium auto amount | Ratio | Relevant |
| Months Since Last Claim | No of months since last claim | Ratio | Relevant |
| Months Since Policy Inception | No of months since policy inception | Ratio | Relevant |
| Number of Open Complaints | No of open complaints on cust. | Ratio | Relevant |
| Number of Policies | No of policies of the cust. | Ratio | Relevant |
| Policy Type | Policy type of cust. | Nominal | Relevant |
| Policy | Policy amount of cust. | Ratio | Relevant |
| Renew Offer Type | Renew offer type of cust. | Nominal | Relevant |
| Sales Channel | Type of sales channel | Nominal | Relevant |
| Total Claim Amount | Total claim amount of the cust. | Ratio | Relevant |
| Vehicle Class | Vehicle class of cust. | Nominal | Relevant |
| Vehicle Size | Vehicle size of cust. | Nominal | Relevant |

**Data Pre Processing:** all the variables in the given data is used for applying the clustering except the customer column, since it is not useful. The data which is in the form of categorical type is converted into numeric type so that it can be used for the clustering.

**Exploratory Data Analysis:** mean, median, variance, standard deviation, skewness, kurtosis is calculated for all the variables of the given data then normalization for all the variables of the data to apply k Means clustering.

**k-Means clustering:** after cleaning the complete data K means clustering is applied on the data and number of clusters is in front decided as 5 and based on that the complete data is clustered.

