**K - Means Clustering**

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: DHEERAJ MISHRA Batch ID:**  DS\_01072020

**Topic: K Means Clustering**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

**Hints:**

**1. Business Problem**

* 1. **What is the business objective?**
  2. **Are there any constraints?**

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



**3. Data Pre-processing**

**3.1 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 K- means clustering and obtain optimum number of clusters using scree plot.**

**5.3 Validate the clusters (try with different number of clusters) – label the clusters and derive insights (compare the results from multiple approaches).**

**6. Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statements:**

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



SOLN:-

1. **Business objective :**

**Max:- Customers convenience**

**Min:- Churn rate**

**Constraints:- Demand of customers**

1. Data understanding :

|  |  |  |  |
| --- | --- | --- | --- |
| Name of Feature | Description | Types | Relevance |
| ID# | Unique ID | Discrete , nominal | Not relevant |
| Balance | Number of miles eligible for award travel | Discrete | Relevant |
| Qual\_miles | Number of miles counted as qualifying for Topflight status | Discrete | Relevant |
| cc1\_miles | Number of miles earned with freq. flyer credit card in the past 12 months | Discrete | Relevant |
| cc2\_miles | Number of miles earned with Rewards credit card in the past 12 months | Discrete | Relevant |
| cc3\_miles | Number of miles earned with Small Business credit card in the past 12 months | Discrete | Relevant |
| Bonus\_miles | Number of miles earned from non-flight bonus transactions in the past 12 months | Discrete | Relevant |
| Bonus\_trans | Number of non-flight bonus transactions in the past 12 months | Discrete | Relevant |
| Flight\_miles\_12mo | Number of flight miles in the past 12 months | Discrete | Relevant |
| Flight\_trans\_12 | Number of flight transactions in the past 12 months | Discrete | Relevant |
| Days\_since\_enroll | Number of days since enrolled in flier program | Discrete | Relevant |
| Award? | whether that person had award flight (free flight) or not | Discrete, binary count | Relevant |

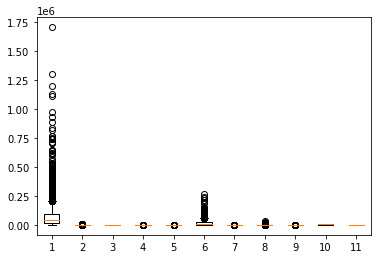
1. EDA

Summary:-

1. Calculated mean, median , sd, quartiles ,min and max values for each feature
2. No null values found in each feature
3. Duplicate row found and removed
4. All features are of type int64
5. Dataset consists of 3999 rows and 12 columns

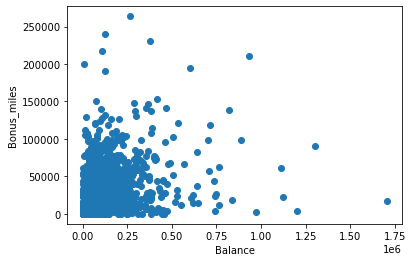
Univariate analysis :-

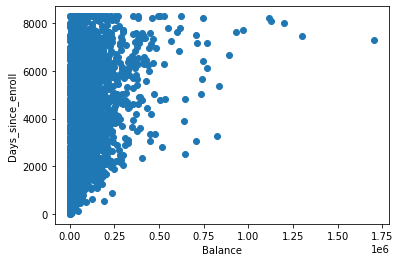
1. From boxplot most features does not have normal distribution
2. From box plot most of features are right skewed
3. From box plot outliers detected

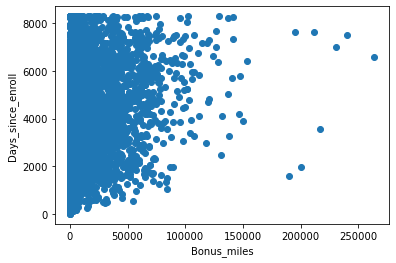


Bivariate analysis :-

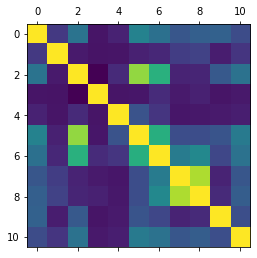
1. From scatter plots observed that there is no linear relationship between any features .
2. From scatter plot some clusters are recognized for different features



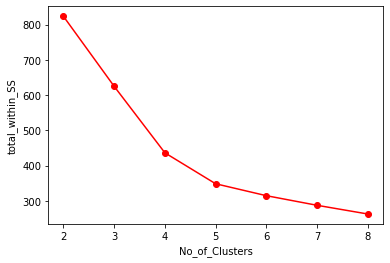




1. From mat plot weak correlations for different features



1. Data preprocessing:-
2. Id feature does not provide useful information so dropped it
3. One row contains duplicate value so it is removed by duplicates function
4. Outliers detected but they are retain for better result
5. Normalization technique used for scaling dataset for better performance
6. Model building :-
7. Model builded on scaled data set using normalization technique
8. From scree plot it is observed that we may have 4 and 5 clusters
9. For better inferences we perform two clusters 4 and 5



Inferences on cluster 3 :-

1. Not Active Customers (Cluster 2) contains customers whose flight transactions since last 12 month and no of flight miles since last 12 months is at lower side. These customers may churn unless some offers given to activate them.
2. Active Customers (Cluster 1) contains customers will qualify for top flight as their their Qual\_miles is highest. Their flight transactions since last 12 months is at higher side along with award
3. New Customers (Cluster 0) contains customers whose Days since enrollment in lowest. They are customers recently joined the program

Inferences on cluster 4 :-

1. Not Active Customers (Cluster 2) contains customers whose flight transactions since last 12 month and no of flight miles since last 12 months is at lower side. These customers may churn unless some offers given to activate them.
2. Active Customers (Cluster 1 and 3) contains customers will qualify for top flight as their their Qual\_miles is highest. Their flight transactions since last 12 months is at higher side along with award
3. New Customers (Cluster 0) contains customers whose Days since enrollment in lowest. They are customers recently joined
4. Impact :-

From above information about each clusters we can investigate on customers who may churn . Also on which customers we need to give more attention as they are high vintage customers

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



1. **Business objective :**

**Max:- Customers safety measures**

**Min:- Crime rate**

**Constraints:- Availability of data**

1. Data understanding :

|  |  |  |  |
| --- | --- | --- | --- |
| Name of Feature | Description | Types | Relevance |
| X | Name of states in United States | Char ,Nominal | Not relevant |
| Murder | No of murder cases | Continuous | Relevant |
| Assault | No of assault cases | Discrete | Relevant |
| Urban pop | Urban population in the states of united states | Discrete | Relevant |
| Rape | No of rape cases | Continuous | Relevant |

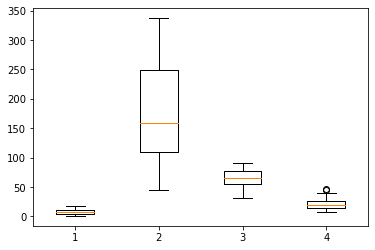
1. EDA

Summary:-

1. Calculated mean, median , sd, quartiles ,min and max values for each feature
2. No null values found in each feature
3. Duplicate records does not exists
4. Two features are of type int64 and two of type float64
5. Dataset consists of 50 rows and 5 columns

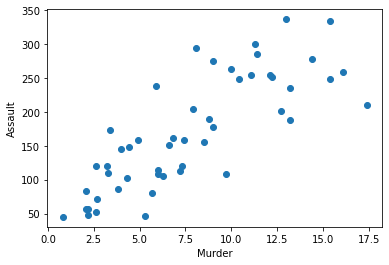
Univariate analysis :-

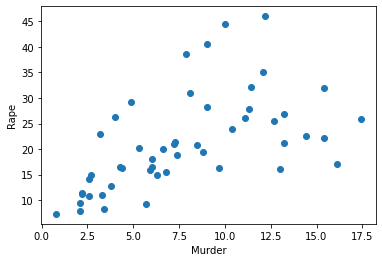
1. From boxplot most features does not have normal distribution
2. From box plot most of features are right skewed
3. From box plot outliers detected

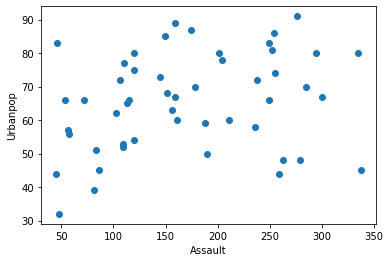


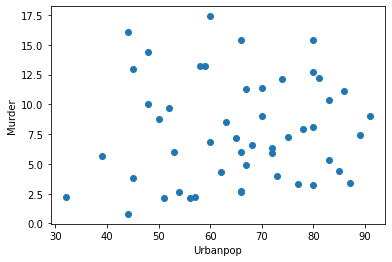
Bivariate analysis :-

1. From scatter plots observed that there is no linear relationship between any features .
2. From scatter plot some clusters are recognized for different features

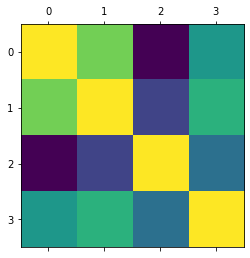




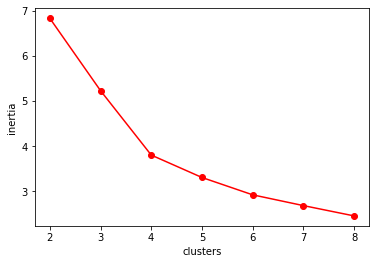




1. From mat plot weak correlations for different features



1. Data preprocessing:-
2. X feature of states name for United States does not provide useful information so dropped it
3. Outliers detected in two features and replaced by winsorizer technique with capping method as iqr and fold value 1.5m.
4. Murder and Rape features are continuous which does not make sence so converted to discrete values by rounding function
5. Murder and Rape features are float64 so converted to int64
6. Normalization technique used for scaling dataset for better performance
7. Model building :-
8. Model builded on scaled data set using normalization technique
9. From scree plot it is observed that we may have 4 and 5 clusters
10. For accuracy we perform two clusters 4 and 5



Inferences on cluster 4 :-

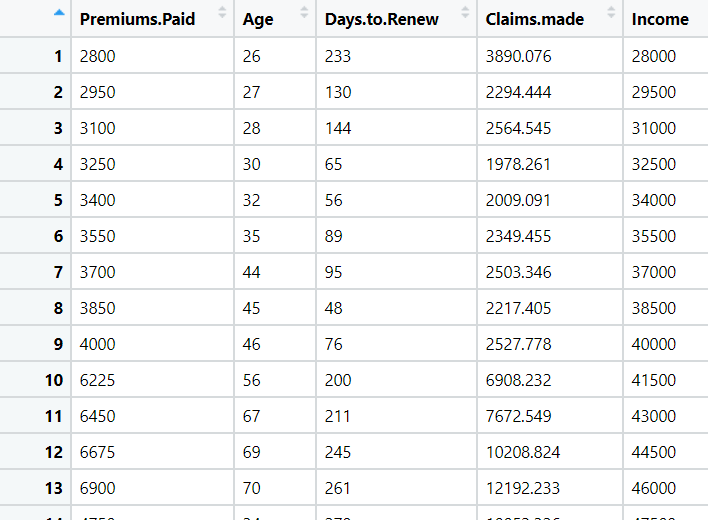
1. Cluster 0 contains states with highest Murder cases
2. Cluster 1 contains states with lowest Urban population with lowest Murder , Assault and Rape cases . Safe state for migrants
3. Cluster 2 contains states with lowest Murder ,Assault and Rape cases with respect to Urban population compared to other states.
4. Cluster 3 contains states with moderate Urban population with highest Assault cases and highest Rape cases

Inferences on cluster 5:-

1. Cluster 0 contains states with highest Assault cases and highest Rape cases.
2. Cluster 1 contains customers with low Murder and Rape cases safe for migrants
3. Cluster 2 contains states with highest Murder cases
4. Cluster 3 contains states with lowest Urban population with lowest Murder , Assault and Rape cases
5. Cluster 4 contains states with lowest Murder ,Assault and Rape cases with respect to Urban population compared to other states
6. Impact :-

From above information about each clusters we predict that in which states which crime rate is more . According to given information some actions to be taken to minimize crime cases . Also migrants can plan which state is safe for them

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 objective :**

**Max:- Customer convenience**

**Min:- Churn rate**

**Constraints:- Availability of data**

1. Data understanding :

|  |  |  |  |
| --- | --- | --- | --- |
| Name of Feature | Description | Types | Relevance |
| Premiums Paid | Premium amount paid  By customer | Discrete | Relevant |
| Age | Age of customer | Discrete | Relevant |
| Days to Renew | No of days left for  Renew the policy | Discrete | Relevant |
| Claims made | Claim amount taken | Continuous | Relevant |
| Income | Income of customer | Discrete | Relevant |

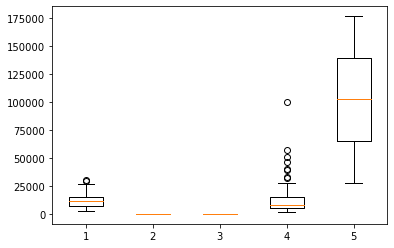
1. EDA

Summary:-

1. Calculated mean, median , sd, quartiles ,min and max values for each feature
2. No null values found in each feature
3. Duplicate records does not exists
4. Features are of type int64 and type float64
5. Dataset consists of 100 rows and 5 columns

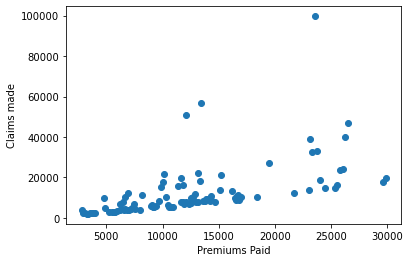
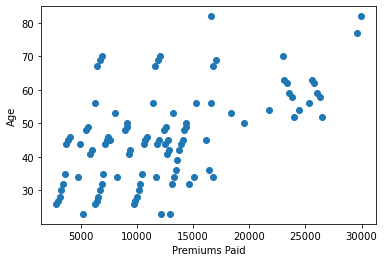
Univariate analysis :-

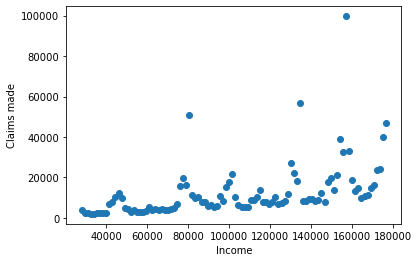
1. From boxplot most features does not have normal distribution
2. From box plot most of features are right skewed
3. From box plot outliers detected



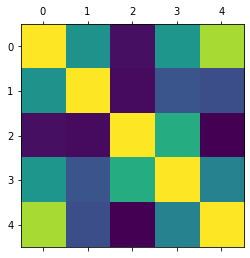
Bivariate analysis :-

1. From scatter plots observed that there is no linear relationship between any features .
2. From scatter plot some clusters are recognized for different features

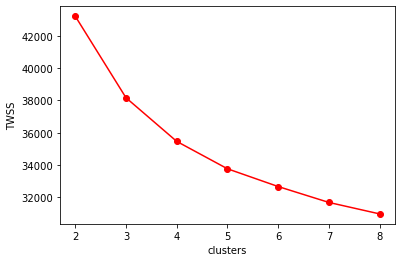




1. Plotting for correlation that is weak correlation



1. Data preprocessing:
2. Outliers detected and retained for better result
3. There is no duplicates record
4. Normalization technique used for scaling dataset for better result
5. Model building :-
6. Model builded on scaled data set using normalization technique
7. From scree plot it is observed that we may have 3 and 4 clusters
8. For accuracy we perform two clusters 3 and 4



Inferences on cluster 3 :-

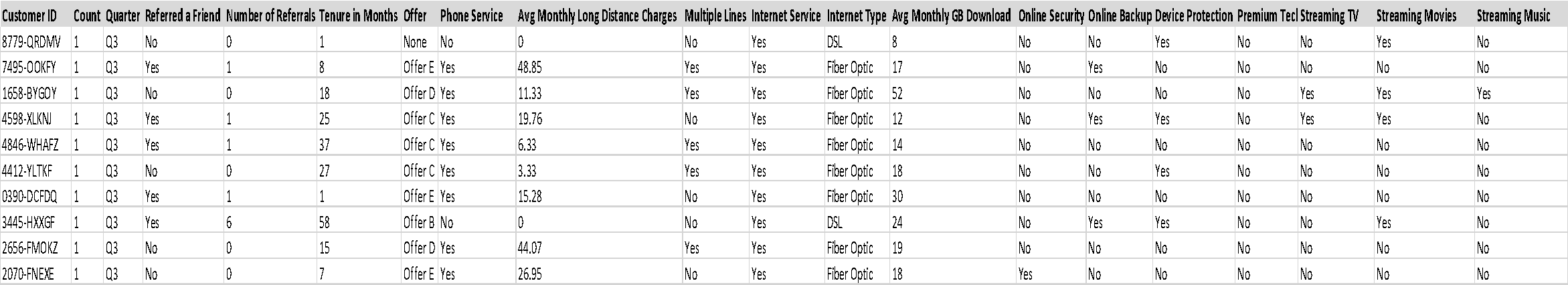
1. High premium (cluster 0) contains customer generated high revenue . They are aged customers with high income . There renewal will due in some time .
2. Moderate premium (cluster 1) contains customer whose renewal date is at higher side .
3. Low premium (cluster 2) contains customer whose renewal date is lowest . Some offer should be provided unless they churn

Inferences on cluster 4 :-

1. High premium (cluster 3) contains customer generated high revenue . They are aged customers with high income . There renewal will due in some time .
2. Moderate premium (cluster 2) contains customer whose renewal date is at higher side .
3. Low premium (cluster 0 and 1) contains customer whose renewal date is lowest . Some offer should be provided unless they churn
4. Impact :-

From above information about each clusters we predict that customer whose renewal date is lower side . Some immediate steps to be taken so that they may not churn

1. Perform clustering analysis on the telecom dataset. The data is a mixture of both categorical and numerical data. It consists of 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.



1. **Business objective :**

**Min:- Churn rate**

**Max:- Customers convenience**

**Constraints:- Size of market**

1. Data understanding

|  |  |  |  |
| --- | --- | --- | --- |
| Name of Feature | Description | Types | Relevance |
| Customer ID | A unique ID that identifies each customer. | Discrete , Nominal | Not relevant |
| Count | A value used in reporting or dashboarding to sum up the number of customers in a filtered set. | Discrete , Nominal | Not relevant |
| Quarter | The fiscal quarter that the data has been derived | Char , Nominal | Not relevant |
| Referred a Friend | Indicates if the customer has ever referred a friend or family member to this company: | Char , Nominal | Not relevant |
| Number of Referrals | Indicates the number of referrals to date that the customer has made. | Discrete | Relevant |
| Tenure in Months | Indicates the total amount of months that the customer has been with the company by end of quarter | Discrete | Relevant |
| Offer | Identifies the last marketing offer that the customer accepted, if applicable | Char | Relevant |
| Phone Service | Indicates if the customer subscribes to home phone service with the company | Char | Relevant |
| Avg Monthly Long Distance Charges | Indicates the customer’s average long distance charges, calculated to the end of the quarter | Continious | Relevant |
| Multiple Lines | Indicates if the customer subscribes to multiple telephone lines with the company | Char | Relevant |
| Internet Service | Indicates if the customer subscribes to Internet service with the company | Char , Nominal | Not relevant |
| Internet Type | Indicates which type of internet used by customer | Charge | Relevant |
| Avg Monthly GB Download | Indicates the customer’s average download volume in gigabytes | Discrete | Relevant |
| Online Security | Indicates if the customer subscribes to an additional online security service provided by the company | Char | Relevant |
| Online Backup | Indicates if the customer subscribes to an additional online backup service provided by the company | Char | Relevant |
| Device Protection Plan | Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company | Char | Relevant |
| Premium Tech Support | Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times | Char | Relevant |
| Streaming TV | Indicates if the customer uses their Internet service to stream television programing from a third party provide | Char | Relevant |
| Streaming Movies | Indicates if the customer uses their Internet service to stream movies from a third party provider | Char | Relevant |
| Streaming Music | Indicates if the customer uses their Internet service to stream music from a third party provider: | Char | Relevant |
| Unlimited Data | Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: | Char | Relevant |
| Contract | Indicates the customer’s current contract type | Char | Relevant |
| Paperless Billing | Indicates if the customer has chosen paperless billing | Char | Relevant |
| Payment Method | Indicates how the customer pays their bill | Char , Nominal | Not relevant |
| Monthly Charge | Indicates the customer’s current total monthly charge for all their services from the company | Continious | Relevant |
| Total Charges | Indicates the customer’s total charges, calculated to the end of the quarter specified above. | Continious | Relevant |
| Total Refunds | **:** Indicates the customer’s total refunds, calculated to the end of the quarter specified above | Continious | Relevant |
| Total Extra Data Charges | Indicates the customer’s total charges for extra data downloads above those specified in their plan | Discrete | Relevant |
| Total Long Distance Charges | Indicates the customer’s total charges for long distance above those specified in their plan | Continious | Relevant |
| Total Revenue | Indicates revenue generated by each customer for the company | Continious | Relevant |

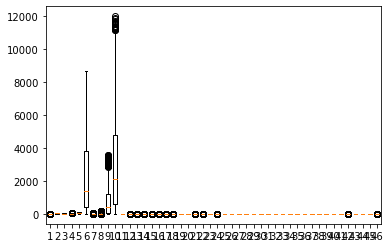
1. EDA

Summary:-

1. Calculated mean, median , sd, quartiles ,min and max values for each feature
2. No null values found in each feature
3. Duplicate records does not exists
4. Features are mixture of int64 ,float64 and char data type
5. Data set contains 7043 rows and 30 columns

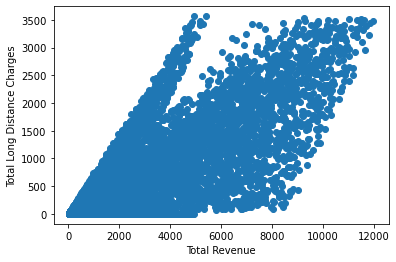
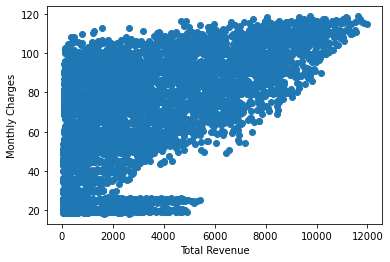
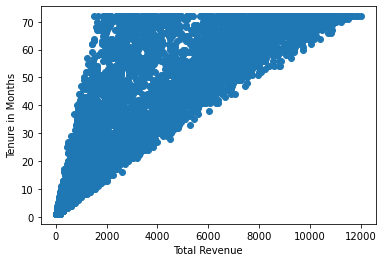
Univariate analysis :-

1. From boxplot most features does not have normal distribution
2. From box plot most of features are right skewed
3. From box plot outliers detected but retained

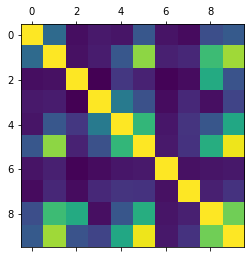


Bivariate analysis :-

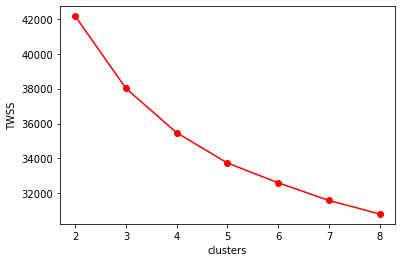
1. From scatter plots observed that there is no linear relations of features .
2. From scatter plot some clusters are recognized for different features

  cc

1. Plotting mat show for weak correlation



1. Data preprocessing:-
2. Customer ID, Count, Quarter, Referred a Friend, Internet Service, Paperless Billing feature of states name for United States does not provide useful information so dropped it
3. Outliers detected but not removed they were retained for better result
4. Some features contains char so they were converted to discrete using dummy variables
5. Normalization technique used for scaling dataset for better performance
6. Model building :-
7. Model builded on scaled data set using normalization technique
8. From scree plot it is observed that we may have 3 and 4 clusters
9. For accuracy we perform two clusters 3 and 4



Inferences on cluster 3 :-

1. High Tenure (Cluster 2) contains customers with highest tenure in months, highest monthly GB download, highest monthly charges, highest total charges, highest total long distance charges, highest total revenue generators, uses more offer B , having highest internet type optical fibre, highest security, highest protection plan, highest online backup, highest streaming TV ,Movies, Music, highest unlimited data plans, highest contract for month to month, highest payment method Bank withdrawal .High vintage customers
2. Moderate Tenure (Cluster 1) contains customers with lowest tenure in months, highest long distance charges, uses more offer E, highest phone service, highest contract month to month. They are not using internet services most
3. Low Tenure (Cluster 0) contains customers with lowest tenure in months, highest long distance charges, uses more offer E, having highest internet type none, highest contract month to month.

Inferences on cluster 6 :-

1. High Tenure (Cluster 3 ) contains customers with highest tenure in months, highest monthly GB download, highest monthly charges, highest total charges, highest total long distance charges, highest total revenue generators, uses more offer B , highest security, highest protection plan, highest online backup, highest streaming TV ,Movies, Music, highest unlimited data plans, highest contract for month to month, highest payment method Bank withdrawal .High vintage customers
2. Moderate Tenure (Cluster 1 and 2) contains customers with lowest tenure in months, highest long distance charges, uses more offer E, highest phone service, having highest internet type optical fibre, highest contract month to month. They are also revenue generators
3. Low Tenure (Cluster 0) contains customers with lowest tenure in months, highest long distance charges, uses more offer E, highest phone service, having highest internet type none, highest contract month to month. They are not using internet services most
4. Impact :-

From above information about each clusters we predict that the which customers are revenue generators . Customer uses which

type of internet services . Which customers uses phone services. Customers payment method . According to above information we can provide services to customers so that we can reduce churn rate.

1. Perform clustering on mixed data. Convert the categorical variables to numeric by using dummies or label encoding and perform normalization techniques. The dataset has the details of customers related to their auto insurance. Refer to Autoinsurance.csv dataset.



1. **Business objective :**

**Min:- Churn rate**

**Max:- CLV value**

**Constraints:- Size of market**

1. Data understanding

|  |  |  |  |
| --- | --- | --- | --- |
| Name of Feature | Description | Types | Relevance |
| Customer | Customer identification no | Char + discrete, Nominal | Not relevant |
| State | State from which customer belongs to | Char , Nominal | Not relevant |
| Customer Lifetime Value | Total revenue generated by customer | Continuous | Relevant |
| Response | Response of customers towards insurance policy | Char | Relevant |
| Coverage | Type of coverage | Char | Relevant |
| Education | Customers qualification | Char | Relevant |
| Effective To Date | Policy start date | Time series | Relevant |
| Employment Status | Whether customer is employed or not | Char | Relevant |
| Gender | Gender of customer | Char | Relevant |
| Income | Income of customer | Discrete | Relevant |
| Location Code | Location of customers | Char | Relevant |
| Marital Status | Customer marital status | Char | Relevant |
| Monthly Premium Auto | Premium amount calculated monthly | Discrete | Relevant |
| Months Since Last Claim | Not taken claims since last months | Discrete | Relevant |
| Months Since Policy Inception | Months since policy started | Discrete | Relevant |
| Number of Open Complaints | Complaints raised by customers | Discrete | Relevant |
| Number of Policies | No of policies taken by customers | Discrete | Relevant |
| Policy Type | Type of policy | Char | Relevant |
| Policy | Policy name | Char | Relevant |
| Renew Offer Type | Renewal offer to customers | Char | Relevant |
| Sales Channel | Through which policy purchased | Char | Relevant |
| Total Claim Amount | Total claim taken by customers | Continuous | Relevant |
| Vehicle Class | Type of vehicle | Char | Relevant |
| Vehicle Size | Size of vehicle | Char | Relevant |

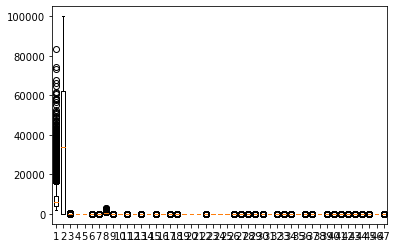
1. EDA

Summary:-

1. Calculated mean, median , sd, quartiles ,min and max values for each feature
2. No null values found in each feature
3. Duplicate records does not exists
4. Features are mixture of int64 ,float64 and char data type
5. Data set contains 9134 rows and 24 columns

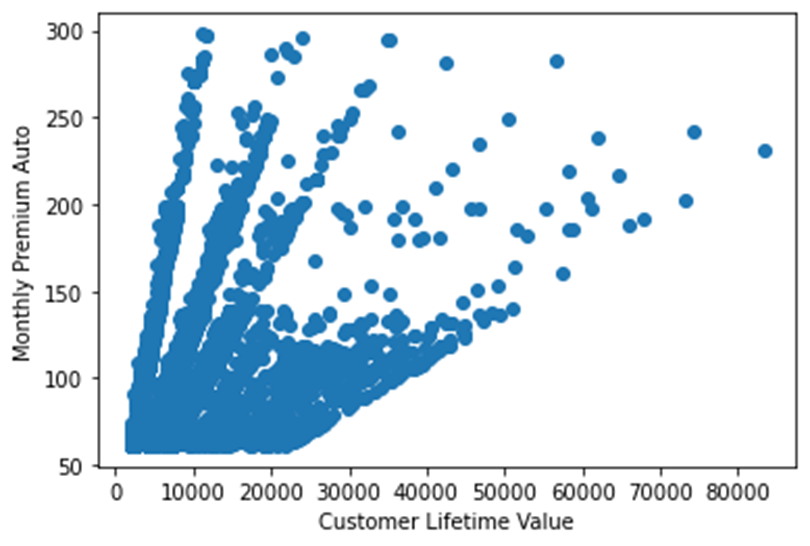
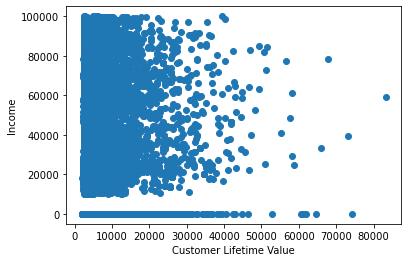
Univariate analysis :-

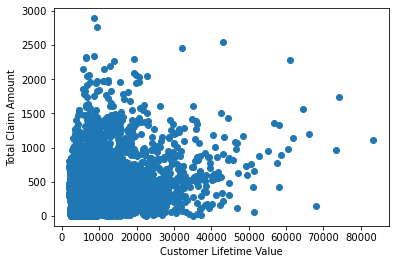
1. From boxplot most features does not have normal distribution
2. From box plot most of features are right skewed
3. From box plot outliers detected but retained for better result



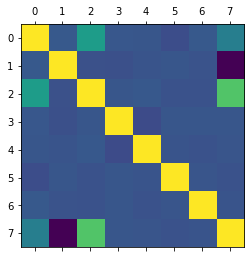
Bivariate analysis :-

1. From scatter plots observed that there is no linear relations of features .
2. From scatter plot some clusters are recognized for different features

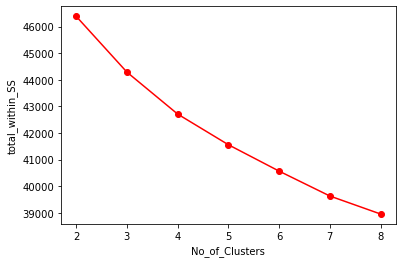
 



1. Matshow for correlation



1. Data preprocessing:-
2. Customer Customer , State , Effective To Date does not provide useful information so dropped it
3. Outliers detected but not removed they were retained for better result
4. Some features contains char so they were converted to discrete using dummy variables
5. Normalization technique used for scaling dataset for better performance
6. Model building :-
7. Model builded on scaled data set using normalization technique
8. From scree plot it is observed that we may have 3 and 4 clusters
9. For accuracy we perform two clusters 3 and 4



Inferences on cluster 3 :-

1. High CLV (Cluster 1) contains customers with highest customer lifetime value. They have high response towards the policy. They are educated and retired .Male customers are more. Sales channel is branch. Vehical class luxury. High vintage customers needs to give more offer to retain them
2. Moderate CLV (Cluster 2) contains customers with highest income. They raised more complains. Response is at lower side. They are highly educated . Class of vehicle is small. Need to get touch with them to increase the response
3. Low CLV (Cluster 0) contains customers total claim amount . Female customers are more . Vehicle size SUV. More attention required to increase CLV value

Inferences on cluster 4 :-

1. High CLV (Cluster 0 and 3) contains customers with highest customer lifetime value. They have high response towards the policy. They are educated and retired .Male customers are more. Sales channel is branch. Vehicle class luxury. High vintage customers needs to give more offer to retain them
2. Moderate CLV (Cluster 2) contains customers with highest income. They raised more complains. Response is at lower side. They are highly educated . Class of vehicle is small. Need to get touch with them to increase the response
3. Low CLV (Cluster 1) contains customers total claim amount . Female customers are more . Vehicle size SUV. More attention required to increase CLV value
4. Impact :-

From above information about each clusters we require to in increase CLV values for customers by reducing complaints and increasing response.