Auto-insurance Clustering

Business Objective:- Clustering auto-insurance policies and customers together so that better business decision can be taken

Constraints:-

Features:-

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Type | Relevance |
| Customer | Customer id | Categorical,Ordinal | Customer id |
| State | State name | Categorical,Nominal | Name of state |
| Customer Lifetime Value | Customer value per lifetime | Quantitative,Nominal | Value of customer throughout life |
| Response | approval | Categorical,Ordinal | Whether insurance was approved or not |
| Coverage | type | Categorical,Ordinal | Type of isnurance plan |
| Education | Education level | Categorical,Ordinal | Education level of cliamant |
| Effective to date | Effective to date | Date/Time series | Last date upto which policy is valid |
| Employment Status | Emloyment status | Categorical,Ordinal | Whether claimant is employed or not |
| Gender | gender | Categorical,Nominal | Gender of claimant |
| Income | income | Quantitative,Nominal | Yearly income of claimant |
| Location Code | Location of claimant | Categorical,Ordinal | Location code of claimant |
| Marital Status | Marital status | Categorical,Nominal | Whether claimant is married ornot |
| Monthly Premium Auto | Monthly premium auto | Quantitative,Nominal | Monthly premium of insurance |
| Months Since Last Claim | Months since last claim | Quantitative,Nominal | Months since last claim |
| Months Since Policy Inception | Months since policy inception | Quantitative,Nominal | Months passed since inception of policy |
| Number of Open Complaints | Number of open complaints | Quantitative,Nominal | Number of open complaints |
| Number of Policies | Number of policies | Quantitative,Nominal | Number of policies |
| Policy Type | Type of policy | Categorical,Ordinal | Type of insurance policy |
| policy | policy | Cateogorical,Ordinal | Type of policy |
| Renew Offer Type | Renew offer type | Categorical,Ordinal | Type of renew offer for policy |
| Sales Channel | Sales Channel | Categorical,Ordinal | Way sale of policy was happened |
| Total Claim Amount | Claim amount | Quantitative | Total claim amount by policy |
| Vehicle Class | Class of vehicle | Categorical | Type of vehicle |
| Vehicle Size | Size of vehicle | Categorical | Size of vehicle |

Data preprocessing

Mean of dataset:-

Customer Lifetime Value 8004.940475

Income 37657.380009

Monthly Premium Auto 93.219291

Months Since Last Claim 15.097000

Months Since Policy Inception 48.064594

Number of Open Complaints 0.384388

Number of Policies 2.966170

Total Claim Amount 434.088794

Median values in dataset:-

Customer Lifetime Value 5780.182197

Income 33889.500000

Monthly Premium Auto 83.000000

Months Since Last Claim 14.000000

Months Since Policy Inception 48.000000

Number of Open Complaints 0.000000

Number of Policies 2.000000

Total Claim Amount 383.945434

Variance in dataset columns:-

Customer Lifetime Value 4.721020e+07

Income 9.229386e+08

Monthly Premium Auto 1.183908e+03

Months Since Last Claim 1.014705e+02

Months Since Policy Inception 7.787443e+02

Number of Open Complaints 8.287982e-01

Number of Policies 5.712969e+00

Total Claim Amount 8.439030e+04

Skewness in data distribution

Customer Lifetime Value 3.032280

Income 0.286887

Monthly Premium Auto 2.123546

Months Since Last Claim 0.278586

Months Since Policy Inception 0.040165

Number of Open Complaints 2.783263

Number of Policies 1.253333

Total Claim Amount 1.714966

Kurtosis of data distribution

Customer Lifetime Value 13.823533

Income -1.094326

Monthly Premium Auto 6.193605

Months Since Last Claim -1.073668

Months Since Policy Inception -1.133046

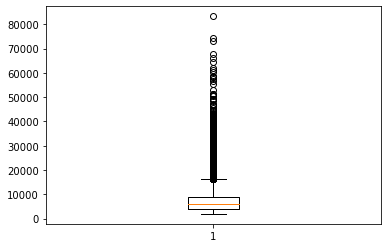
Number of Open Complaints 7.749308

Number of Policies 0.363157

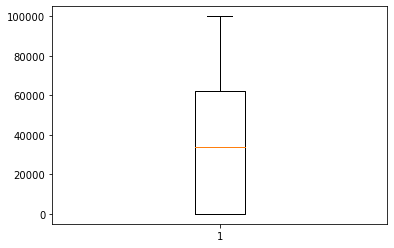
Total Claim Amount 5.979401

Boxplots to detect outliers

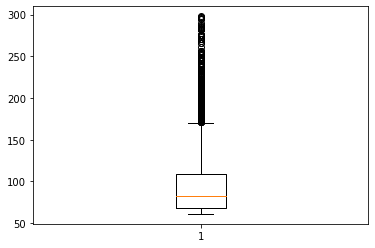
'Customer Lifetime Value'



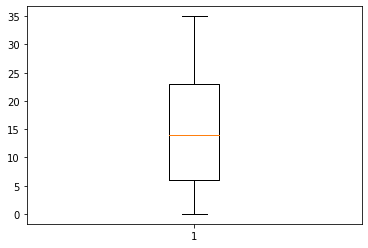
Income



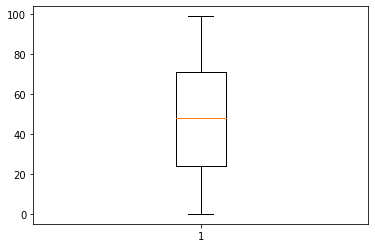
'Monthly Premium Auto'



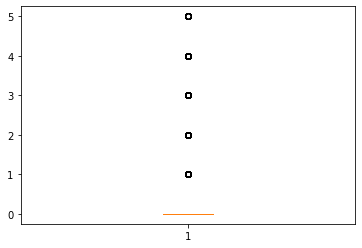
'Months Since Last Claim'



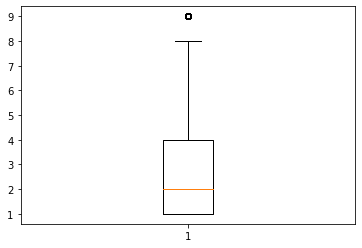
Months Since Policy Inception



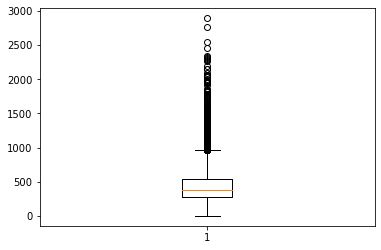
'Number of Open Complaints'



'Number of Policies'



Total Claim amount



Outliers treated through winsorization

Quantitative features were normalized

Dataset post normalization was verified

Customer Lifetime Value ... Months Since Policy Inception

count 9134.000000 ... 9134.000000

mean 0.364266 ... 0.485501

std 0.291685 ... 0.281879

min 0.000000 ... 0.000000

25% 0.144409 ... 0.242424

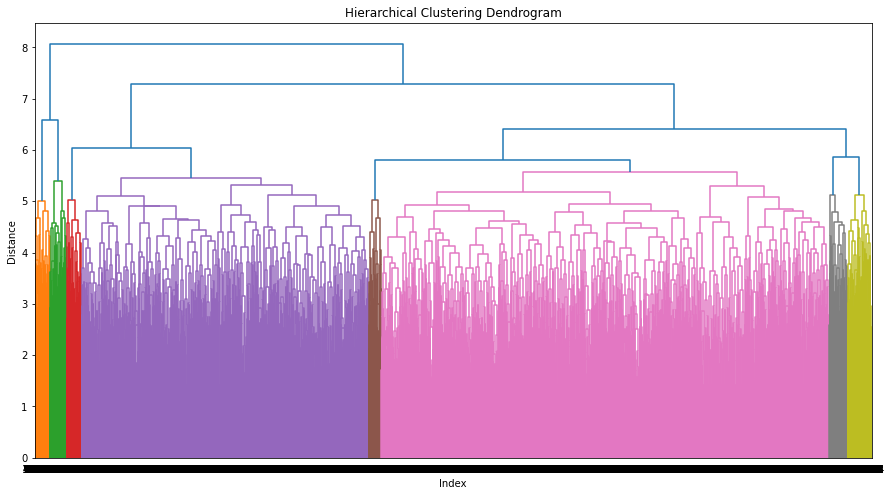
50% 0.267440 ... 0.484848

75% 0.486645 ... 0.717172

max 1.000000 ... 1.000000

Nominal Categorical variables were encoded with one hot encoding while Ordinal Categorical variables were encoded with label encoding variables.

Dendrogram was drawn to get the idea of how many clusters should be made



After this agglomeration was performed and clusters were made

Final Clustered dataset description was as follows

Customer Lifetime Value Monthly Premium Auto ... Education Vehicle Size

clust ...

0 0.359834 0.273349 ... 0.530783 1.097975

1 0.343145 0.240188 ... 3.207650 1.032787

2 0.378877 0.290589 ... 3.175215 1.089200

3 0.381030 0.276767 ... 1.791045 0.940299

4 0.306285 0.259128 ... 0.528517 1.068441

5 0.373505 0.270009 ... 2.987421 1.220126

6 0.417443 0.309436 ... 0.431472 1.010152

7 0.250877 0.247542 ... 0.629412 1.035294

Clustered dataset was saved as a new csv file