Report

Team1

31/07/2021

Content

Customer Lifetime Value Analysis

CLV is a measurement of how valuable a customer is to your company with an unlimited time span as opposed to just the first purchase. This metric helps you understand a reasonable cost per acquisition.

CLV is the total worth to a business of a customer over the whole period of their relationship. It’s an important metric as it costs less to keep existing customers than it does to acquire new ones, so increasing the value of your existing customers is a great way to drive growth.

CLV helps the organization in :

- management of customer relationship as an asset
- monitoring the impact of management strategies and marketing investments on the value of customer assets
- encourages marketers to focus on the long-term value of customers
- optimal allocation of limited resources for ongoing marketing activities in order to achieve a maximum return
- a good basis for selecting customers and for decision making regarding customer specific communication strategies

Problem Statement and Dataset

The task is to understand the problem of Customer Lifetime Value and provide a data-driven solution through data methodologies.To understand customer behaviour and to find out the profitable customers.To predict the lifetime valuation of a customer to facilitate target marketing

The dataset delineates the demographics and policy specifics purchased by the customers from a bank/firm. Customer Lifetime Valuation (Customer.Lifetime.Value) is the target variable that is predicted.A preview of the dataset is :

```
##   Customer      State Customer.Lifetime.Value Response Coverage Education
## 1  BU79786 Washington                2763.519       No    Basic  Bachelor
## 2  QZ44356    Arizona                6979.536       No Extended  Bachelor
## 3  AI49188     Nevada               12887.432       No  Premium  Bachelor
## 4  WW63253 California                7645.862       No    Basic  Bachelor
##   Effective.To.Date EmploymentStatus Gender Income Location.Code Marital.Status
## 1           2/24/11         Employed      F  56274      Suburban        Married
## 2           1/31/11       Unemployed      F      0      Suburban         Single
## 3           2/19/11         Employed      F  48767      Suburban        Married
## 4           1/20/11       Unemployed      M      0      Suburban        Married
##   Monthly.Premium.Auto Months.Since.Last.Claim Months.Since.Policy.Inception
## 1                   69                      32                             5
## 2                   94                      13                            42
## 3                  108                      18                            38
## 4                  106                      18                            65
##   Number.of.Open.Complaints Number.of.Policies    Policy.Type       Policy
## 1                         0                  1 Corporate Auto Corporate L3
## 2                         0                  8  Personal Auto  Personal L3
## 3                         0                  2  Personal Auto  Personal L3
## 4                         0                  7 Corporate Auto Corporate L2
##   Renew.Offer.Type Sales.Channel Total.Claim.Amount Vehicle.Class Vehicle.Size
## 1           Offer1         Agent           384.8111  Two-Door Car      Medsize
## 2           Offer3         Agent          1131.4649 Four-Door Car      Medsize
## 3           Offer1         Agent           566.4722  Two-Door Car      Medsize
## 4           Offer1   Call Center           529.8813           SUV      Medsize
```

The variables available in the dataset are :

```
##  [1] "Customer"                      "State"                        
##  [3] "Customer.Lifetime.Value"       "Response"                     
##  [5] "Coverage"                      "Education"                    
##  [7] "Effective.To.Date"             "EmploymentStatus"             
##  [9] "Gender"                        "Income"                       
## [11] "Location.Code"                 "Marital.Status"               
## [13] "Monthly.Premium.Auto"          "Months.Since.Last.Claim"      
## [15] "Months.Since.Policy.Inception" "Number.of.Open.Complaints"    
## [17] "Number.of.Policies"            "Policy.Type"                  
## [19] "Policy"                        "Renew.Offer.Type"             
## [21] "Sales.Channel"                 "Total.Claim.Amount"           
## [23] "Vehicle.Class"                 "Vehicle.Size"
```

The dimension of the dataset is (9134, 24) .

There are 15 categorical variables out of which one is the customer id, 8 numerical variables (including target variable) and one date. The structure of these variables can be inferred here :

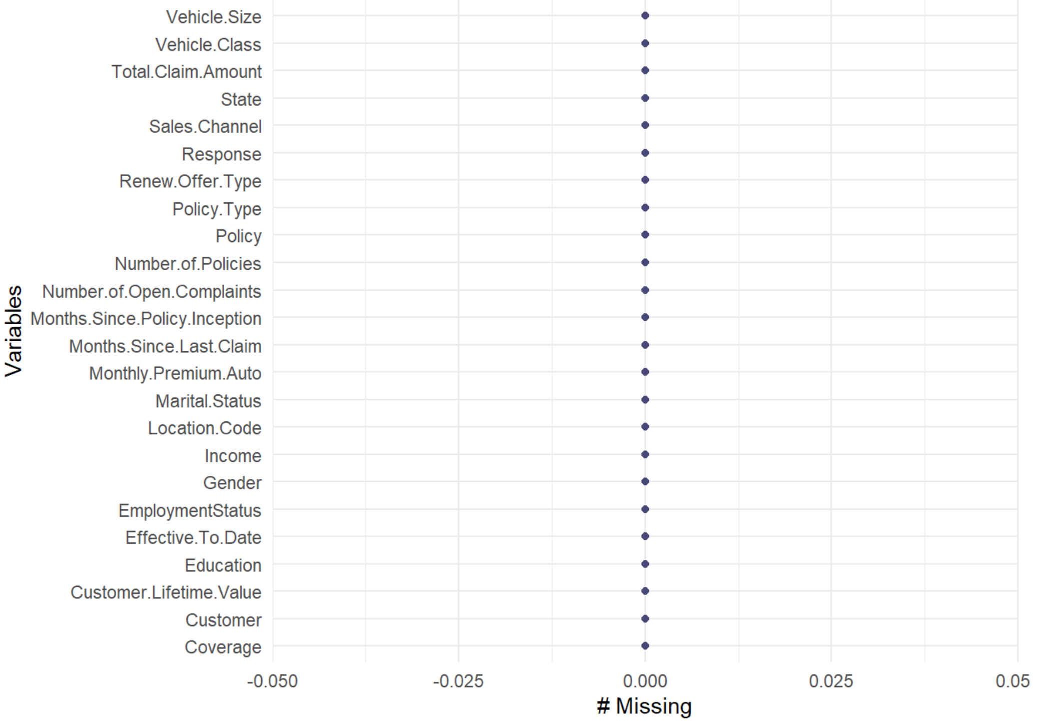
```
## 'data.frame':    9134 obs. of  24 variables:
##  $ Customer                     : chr  "BU79786" "QZ44356" "AI49188" "WW63253" ...
##  $ State                        : chr  "Washington" "Arizona" "Nevada" "California" ...
##  $ Customer.Lifetime.Value      : num  2764 6980 12887 7646 2814 ...
##  $ Response                     : chr  "No" "No" "No" "No" ...
##  $ Coverage                     : chr  "Basic" "Extended" "Premium" "Basic" ...
##  $ Education                    : chr  "Bachelor" "Bachelor" "Bachelor" "Bachelor" ...
##  $ Effective.To.Date            : chr  "2/24/11" "1/31/11" "2/19/11" "1/20/11" ...
##  $ EmploymentStatus             : chr  "Employed" "Unemployed" "Employed" "Unemployed" ...
##  $ Gender                       : chr  "F" "F" "F" "M" ...
##  $ Income                       : int  56274 0 48767 0 43836 62902 55350 0 14072 28812 ...
##  $ Location.Code                : chr  "Suburban" "Suburban" "Suburban" "Suburban" ...
##  $ Marital.Status               : chr  "Married" "Single" "Married" "Married" ...
##  $ Monthly.Premium.Auto         : int  69 94 108 106 73 69 67 101 71 93 ...
##  $ Months.Since.Last.Claim      : int  32 13 18 18 12 14 0 0 13 17 ...
##  $ Months.Since.Policy.Inception: int  5 42 38 65 44 94 13 68 3 7 ...
##  $ Number.of.Open.Complaints    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Number.of.Policies           : int  1 8 2 7 1 2 9 4 2 8 ...
##  $ Policy.Type                  : chr  "Corporate Auto" "Personal Auto" "Personal Auto" "Corporate Auto" ...
##  $ Policy                       : chr  "Corporate L3" "Personal L3" "Personal L3" "Corporate L2" ...
##  $ Renew.Offer.Type             : chr  "Offer1" "Offer3" "Offer1" "Offer1" ...
##  $ Sales.Channel                : chr  "Agent" "Agent" "Agent" "Call Center" ...
##  $ Total.Claim.Amount           : num  385 1131 566 530 138 ...
##  $ Vehicle.Class                : chr  "Two-Door Car" "Four-Door Car" "Two-Door Car" "SUV" ...
##  $ Vehicle.Size                 : chr  "Medsize" "Medsize" "Medsize" "Medsize" ...
```

The summary statistics of the numerical variables are :

```
##                               vars    n         mean           sd     median
## Customer*                        1 9134  4567.500000 2.636903e+03  4567.5000
## State*                           2 9134     2.741734 1.287771e+00     2.0000
## Customer.Lifetime.Value          3 9134  8004.940475 6.870968e+03  5780.1822
## Response*                        4 9134     1.143201 3.502971e-01     1.0000
## Coverage*                        5 9134     1.480622 6.558173e-01     1.0000
## Education*                       6 9134     2.554084 1.381978e+00     2.0000
## Effective.To.Date*               7 9134    30.057915 1.696353e+01    30.0000
## EmploymentStatus*                8 9134     2.825706 1.347793e+00     2.0000
## Gender*                          9 9134     1.490037 4.999281e-01     1.0000
## Income                          10 9134 37657.380009 3.037990e+04 33889.5000
## Location.Code*                  11 9134     1.979089 6.057325e-01     2.0000
## Marital.Status*                 12 9134     2.120210 6.368385e-01     2.0000
## Monthly.Premium.Auto            13 9134    93.219291 3.440797e+01    83.0000
## Months.Since.Last.Claim         14 9134    15.097000 1.007326e+01    14.0000
## Months.Since.Policy.Inception   15 9134    48.064594 2.790599e+01    48.0000
## Number.of.Open.Complaints       16 9134     0.384388 9.103835e-01     0.0000
## Number.of.Policies              17 9134     2.966170 2.390182e+00     2.0000
## Policy.Type*                    18 9134     1.825925 4.759888e-01     2.0000
## Policy*                         19 9134     4.797788 1.605978e+00     5.0000
## Renew.Offer.Type*               20 9134     1.970221 1.007576e+00     2.0000
## Sales.Channel*                  21 9134     2.102693 1.069452e+00     2.0000
## Total.Claim.Amount              22 9134   434.088794 2.905001e+02   383.9454
## Vehicle.Class*                  23 9134     3.036019 2.176820e+00     1.0000
## Vehicle.Size*                   24 9134     2.089556 5.373128e-01     2.0000
##                                    trimmed        mad         min      max
## Customer*                     4.567500e+03  3385.5171    1.000000  9134.00
## State*                        2.692939e+00     1.4826    1.000000     5.00
## Customer.Lifetime.Value       6.631050e+03  3658.8991 1898.007675 83325.38
## Response*                     1.054050e+00     0.0000    1.000000     2.00
## Coverage*                     1.363027e+00     0.0000    1.000000     3.00
## Education*                    2.466201e+00     1.4826    1.000000     5.00
## Effective.To.Date*            3.005761e+01    22.2390    1.000000    59.00
## EmploymentStatus*             2.712644e+00     0.0000    1.000000     5.00
## Gender*                       1.487548e+00     0.0000    1.000000     2.00
## Income                        3.572792e+04 42522.4506    0.000000 99981.00
## Location.Code*                1.973864e+00     0.0000    1.000000     3.00
## Marital.Status*               2.150246e+00     0.0000    1.000000     3.00
## Monthly.Premium.Auto          8.734031e+01    26.6868   61.000000   298.00
## Months.Since.Last.Claim       1.467761e+01    11.8608    0.000000    35.00
## Months.Since.Policy.Inception 4.786932e+01    35.5824    0.000000    99.00
## Number.of.Open.Complaints     1.325944e-01     0.0000    0.000000     5.00
## Number.of.Policies            2.541461e+00     1.4826    1.000000     9.00
## Policy.Type*                  1.855638e+00     0.0000    1.000000     3.00
## Policy*                       4.931582e+00     1.4826    1.000000     9.00
## Renew.Offer.Type*             1.837849e+00     1.4826    1.000000     4.00
## Sales.Channel*                2.003421e+00     1.4826    1.000000     4.00
## Total.Claim.Amount            4.026767e+02   213.5753    0.099007  2893.24
## Vehicle.Class*                2.920088e+00     0.0000    1.000000     6.00
## Vehicle.Size*                 2.111932e+00     0.0000    1.000000     3.00
##                                   range         skew   kurtosis           se
## Customer*                      9133.000  0.000000000 -1.2003941 2.759076e+01
## State*                            4.000  0.210353443 -1.2522205 1.347436e-02
## Customer.Lifetime.Value       81427.374  3.031284401 13.8116290 7.189313e+01
## Response*                         1.000  2.036897794  2.1491880 3.665271e-03
## Coverage*                         2.000  1.030971447 -0.1071182 6.862026e-03
## Education*                        4.000  0.333038262 -1.3708988 1.446008e-02
## Effective.To.Date*               58.000 -0.007240518 -1.1809683 1.774948e-01
## EmploymentStatus*                 4.000  0.831874159 -1.0269468 1.410239e-02
## Gender*                           1.000  0.039852474 -1.9986306 5.230908e-03
## Income                        99981.000  0.286793057 -1.0948011 3.178747e+02
## Location.Code*                    2.000  0.009507714 -0.2774384 6.337973e-03
## Marital.Status*                   2.000 -0.107517379 -0.5805584 6.663445e-03
## Monthly.Premium.Auto            237.000  2.122849037  6.1875462 3.600216e-01
## Months.Since.Last.Claim          35.000  0.278494819 -1.0741586 1.053997e-01
## Months.Since.Policy.Inception    99.000  0.040151771 -1.1334913 2.919893e-01
## Number.of.Open.Complaints         5.000  2.782348975  7.7420578 9.525634e-03
## Number.of.Policies                8.000  1.252921117  0.3615648 2.500924e-02
## Policy.Type*                      2.000 -0.468229157  0.4983517 4.980423e-03
## Policy*                           8.000 -0.427498589  0.1984827 1.680386e-02
## Renew.Offer.Type*                 3.000  0.717161897 -0.6293888 1.054259e-02
## Sales.Channel*                    3.000  0.506607401 -1.0369589 1.119002e-02
## Total.Claim.Amount             2893.141  1.714402582  5.9735064 3.039595e+00
## Vehicle.Class*                    5.000  0.254371988 -1.7698517 2.277677e-02
## Vehicle.Size*                     2.000  0.072717859  0.3436722 5.622076e-03
```

Data Preprocessing and Cleaning.

There are no missing values in our dataset. This can be inferred from the following graph :

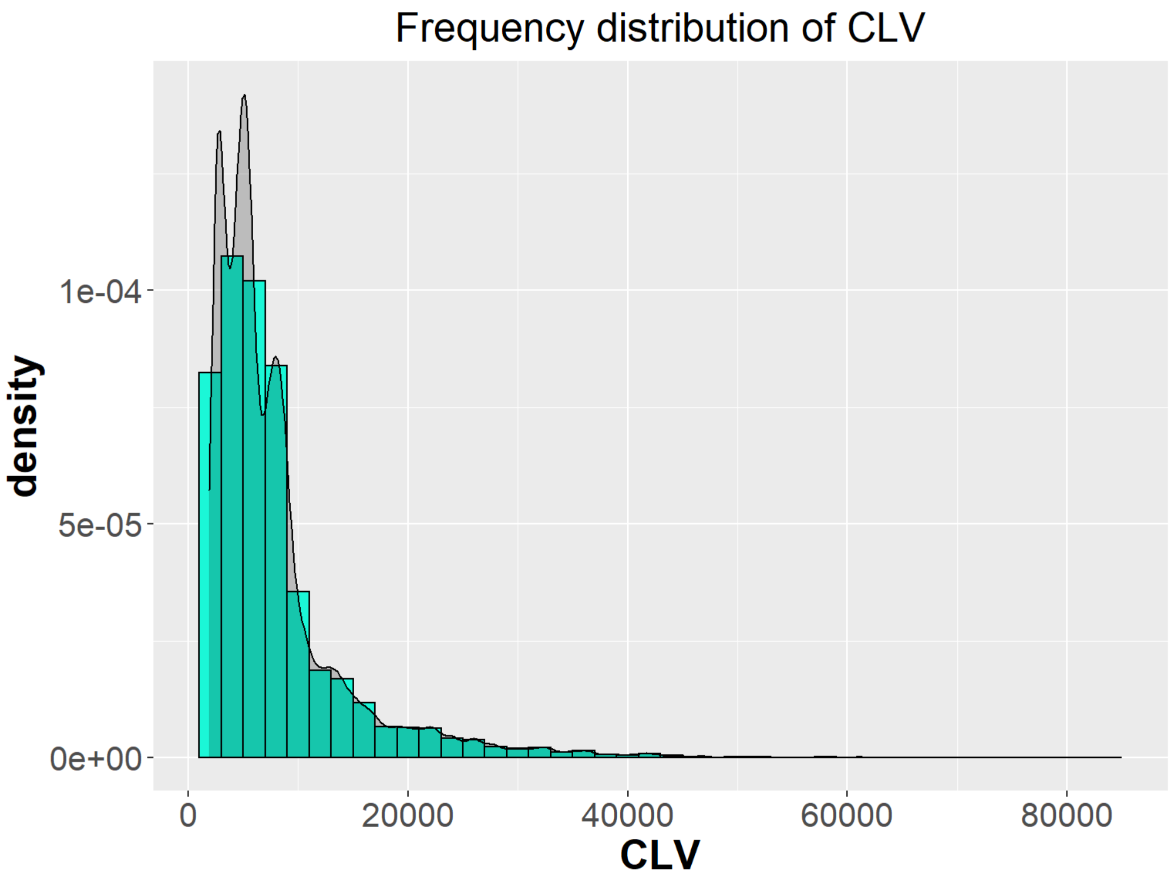


The *Effective.To.Date* column contains two different date formats : “1/20/11” and “02-03-2011”. We have cleaned this and converted them into a unique format which is : “2011-02-21”.

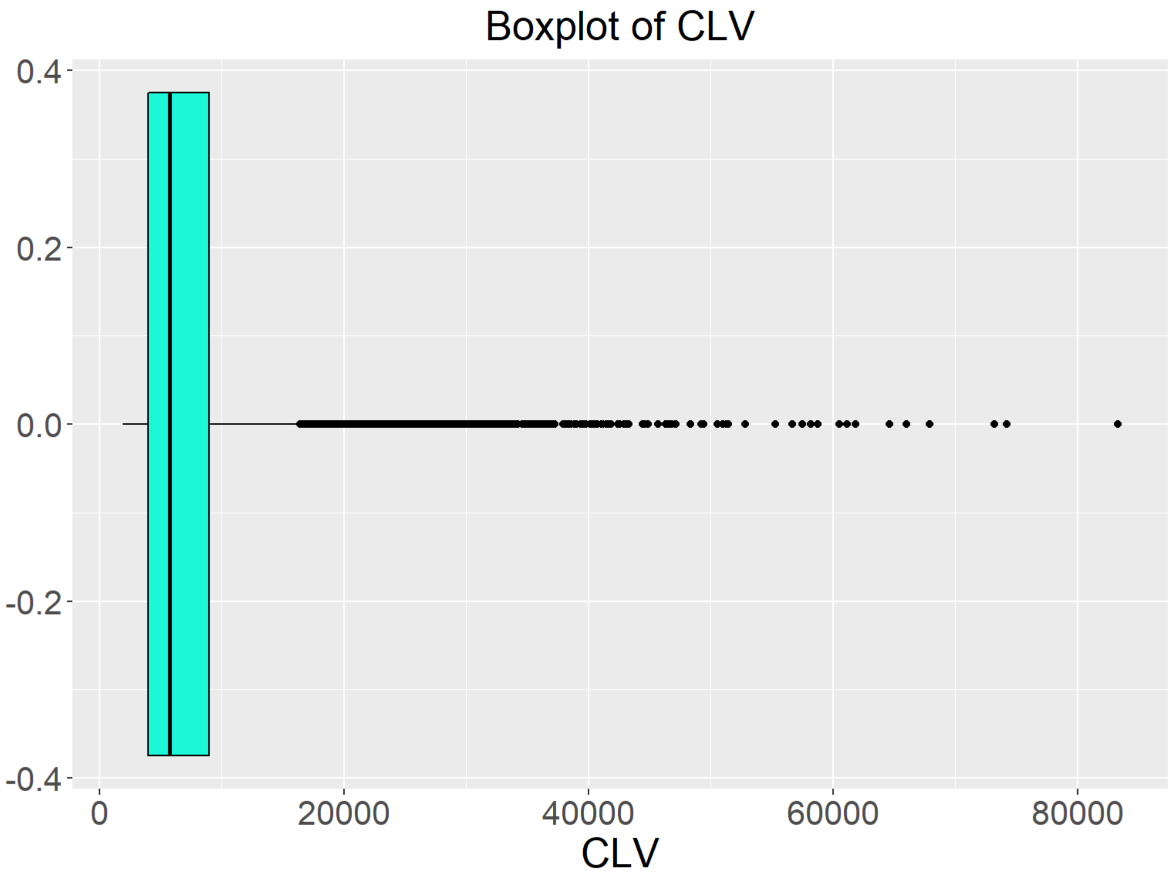
Exploratory Data Analysis.

Questions asked before approaching EDA:

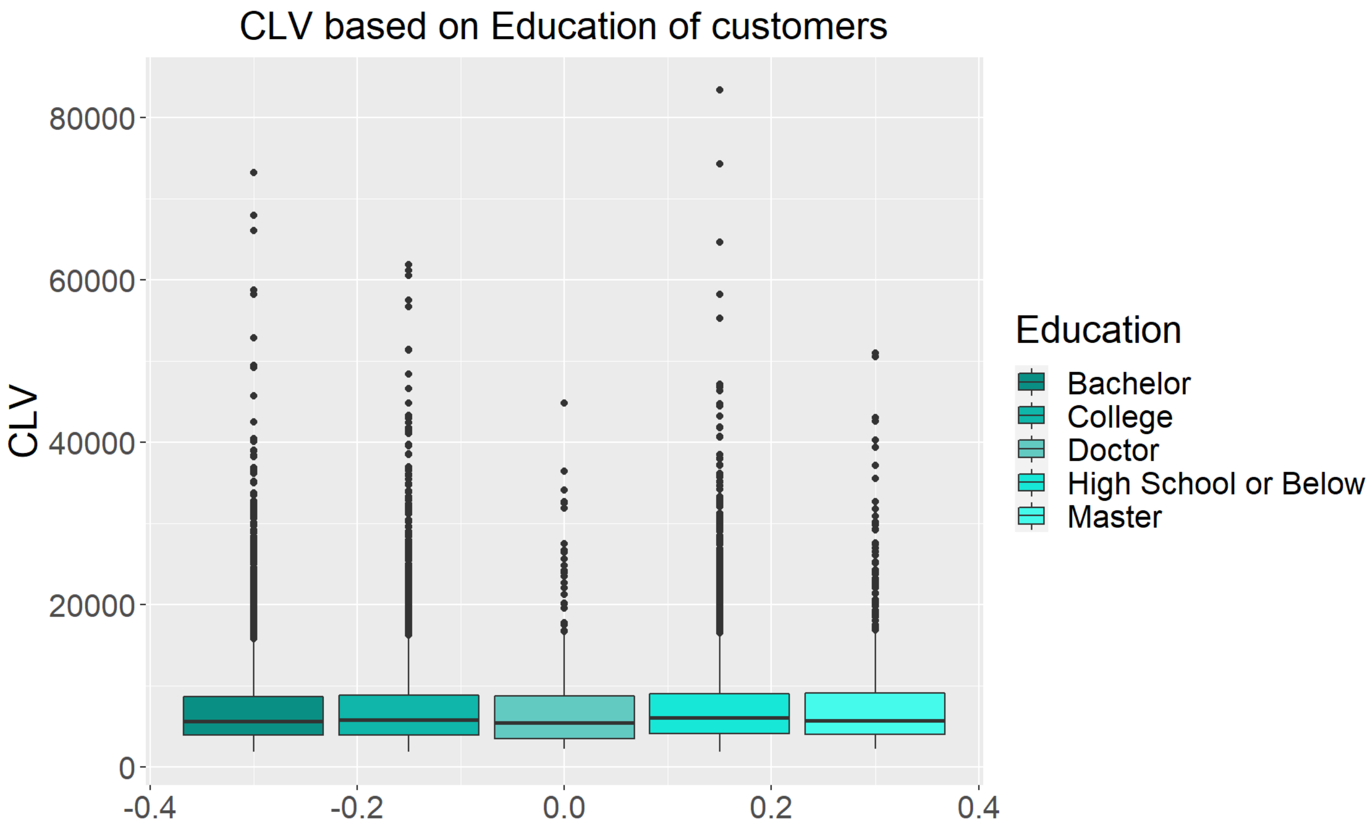
1. What is the distribution of the target variable?
2. Does the target variable contain outliers?
3. How is the target variable associated with categorical features(State,Response,Coverage,Education,EmploymentStatus,Gender,LocationCode,Saleschannel,Marital Status,Policy Type, Policy,Renew Offer Type,Sales Channel,Vehicle Class,Vehicle Size) of the dataset?
4. How is the target variable associated with Numerical features(Income,Monthly Premium Auto,Months Since Last Claim,Months Since Policy Inception,Number of Open Complaints,Number of Policies,Total Claim Amount) of the dataset?
5. Are the numerical variables correlated? If yes, to what extent?
6. What are the most important features of the dataset and how do they affect the target variable?
7. Is gender an essential feature for increase in CLV?
8. Is marital status an essential feature for increase in CLV?
9. How are customers with different policy types affecting the target variable?
10. What sales channels have a greater impact on CLV?
11. What are the types of customers the company must target based on the EDA?



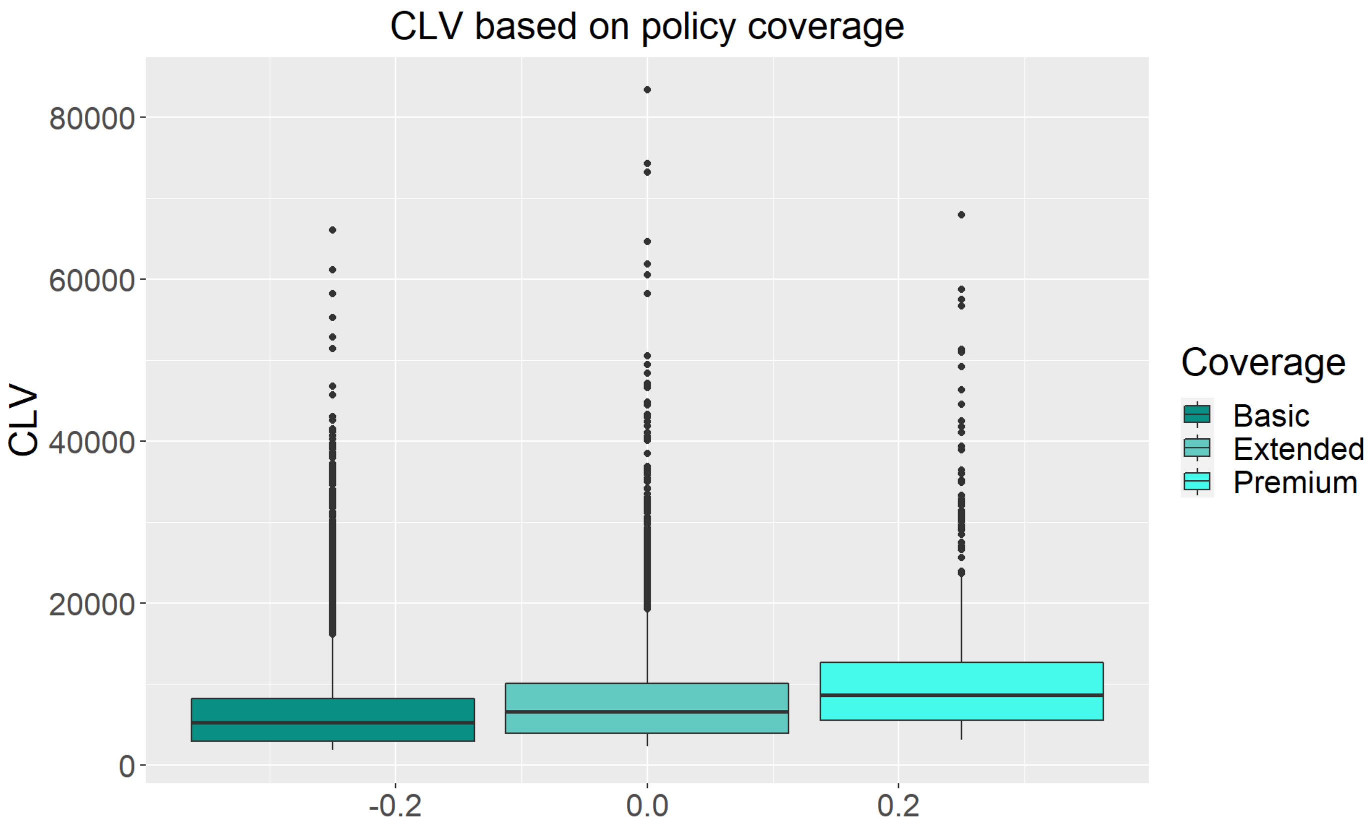
The above graph shows that the target variable, Customer Lifetime value has a positively skewed distribution.

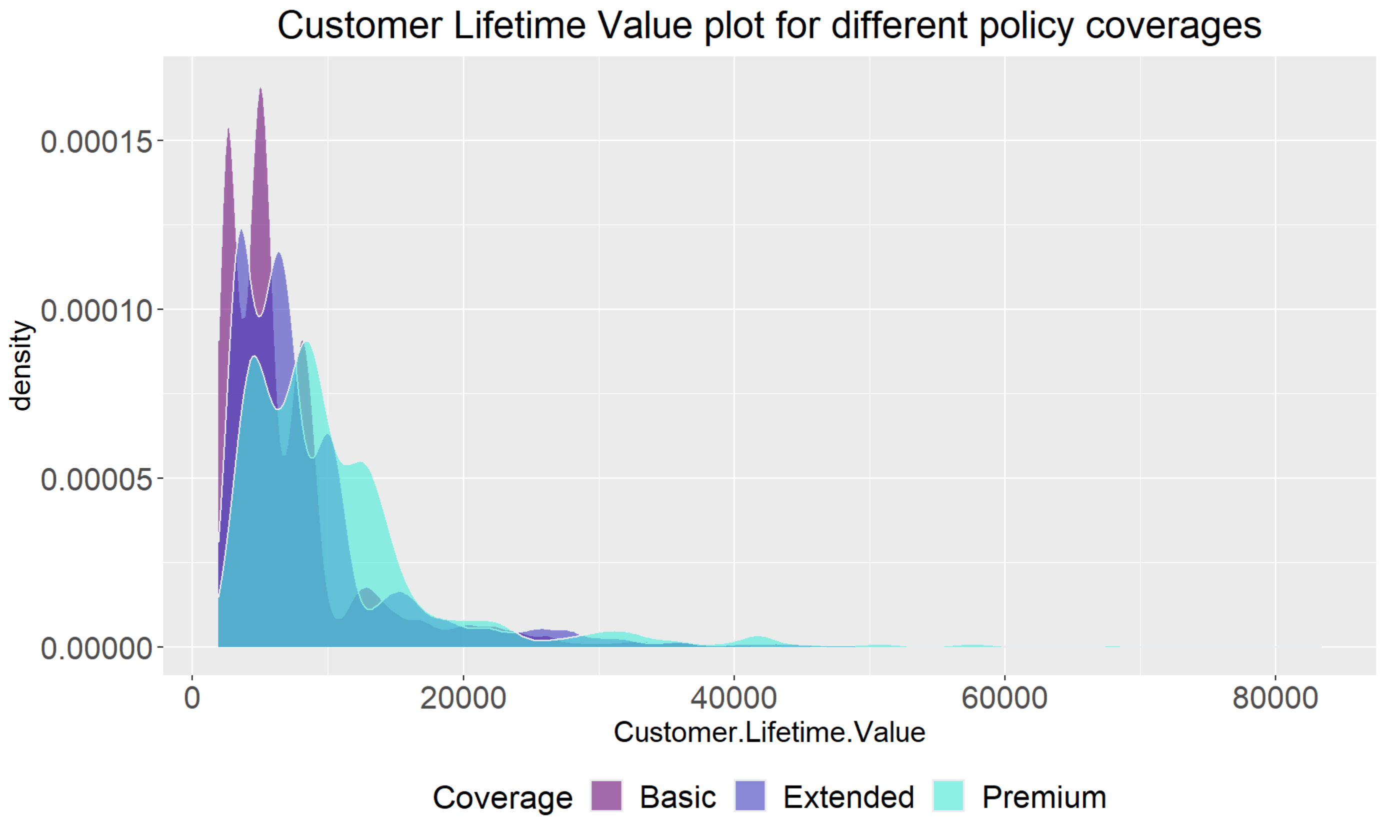


The boxplot of the target variable shows there are quite a few extreme values to the right which may also be observed in real life cases. This can be adjusted through some transformations, like log transformations.

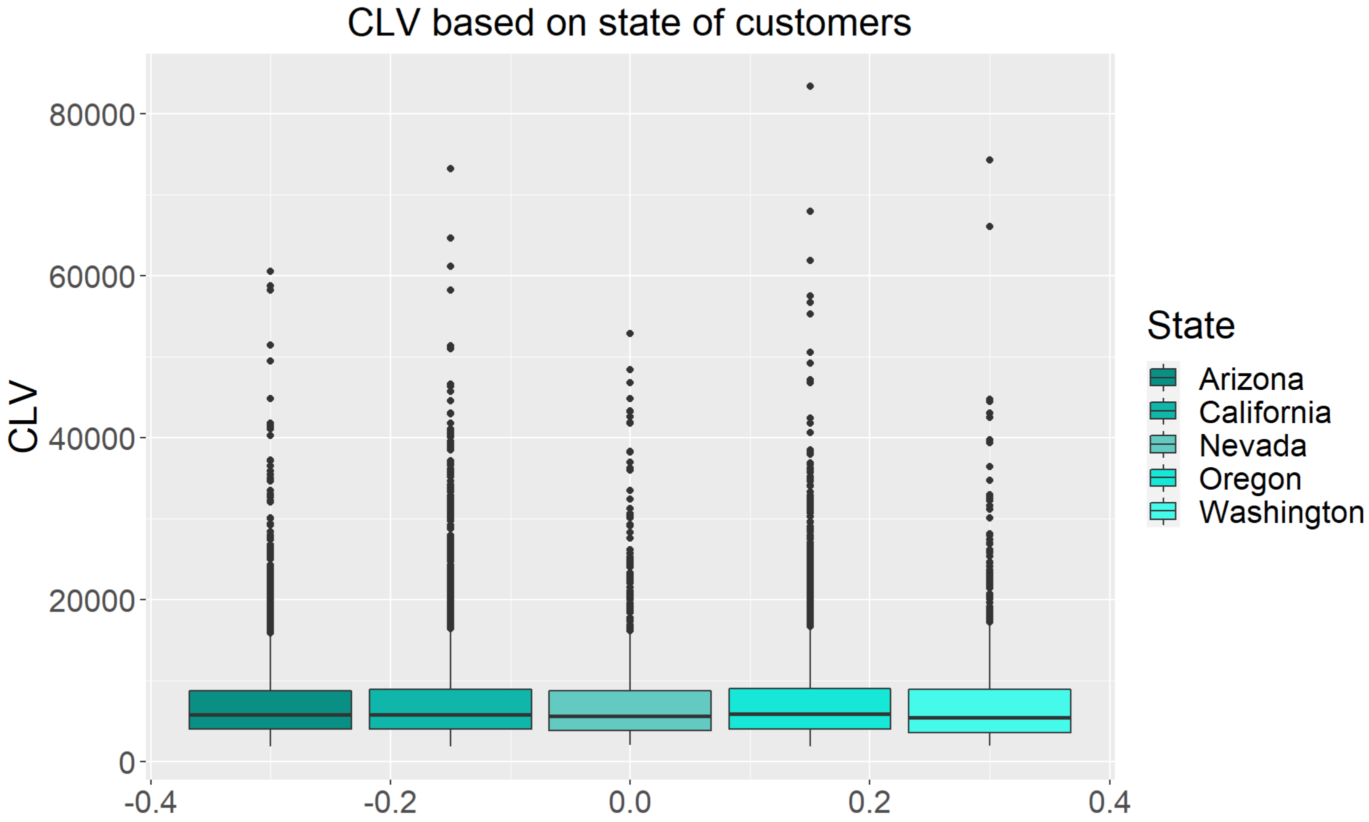


We can infer that the distribution of CLV for customers with education level below highschool and Masters are more significant than the other customers. But this plot also shows that education does not affect the target variable to a great extend.

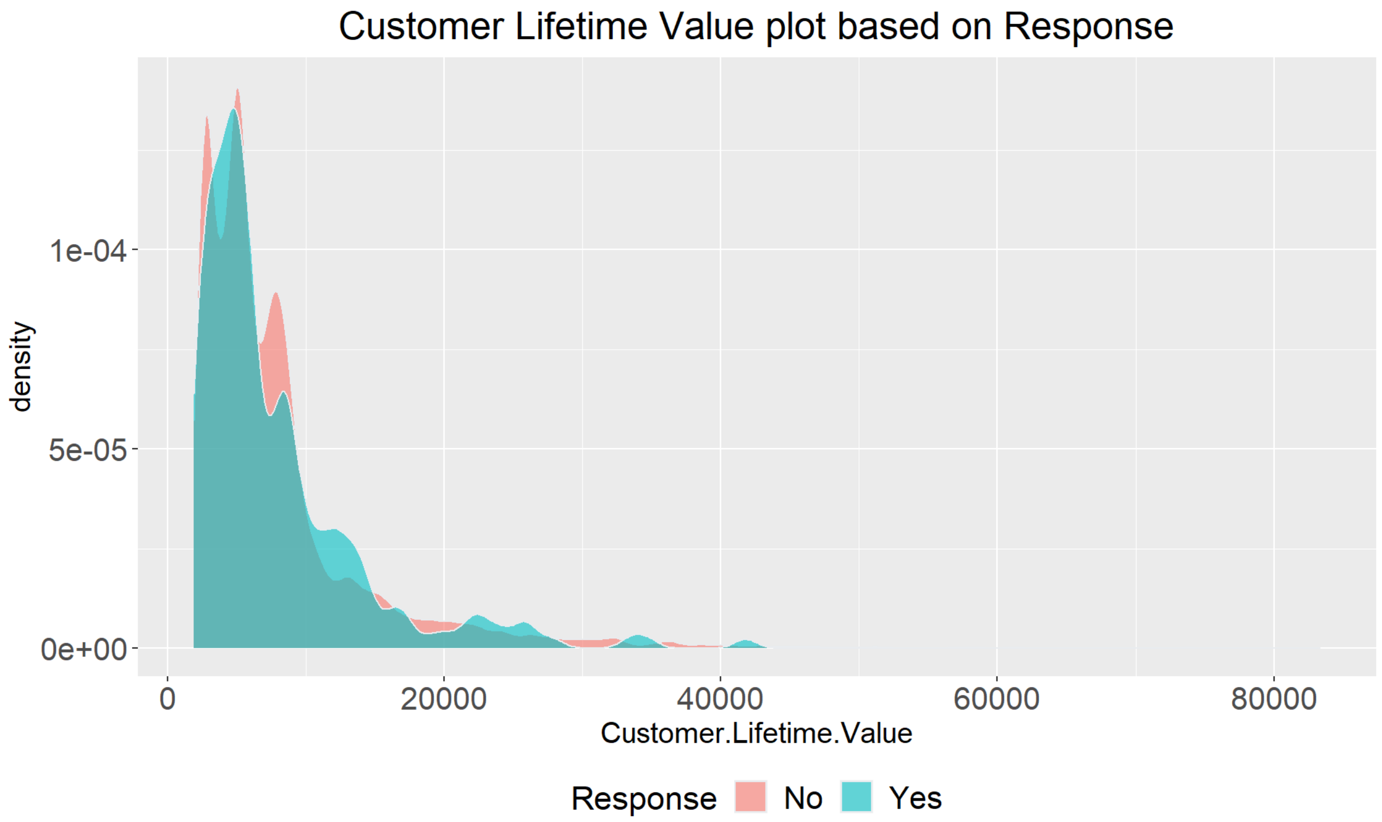




The above 2 plots suggests that the customers who opted for Premium and Extended policy coverages are more valuable to the company when compared to those customers with Basic policy coverage.

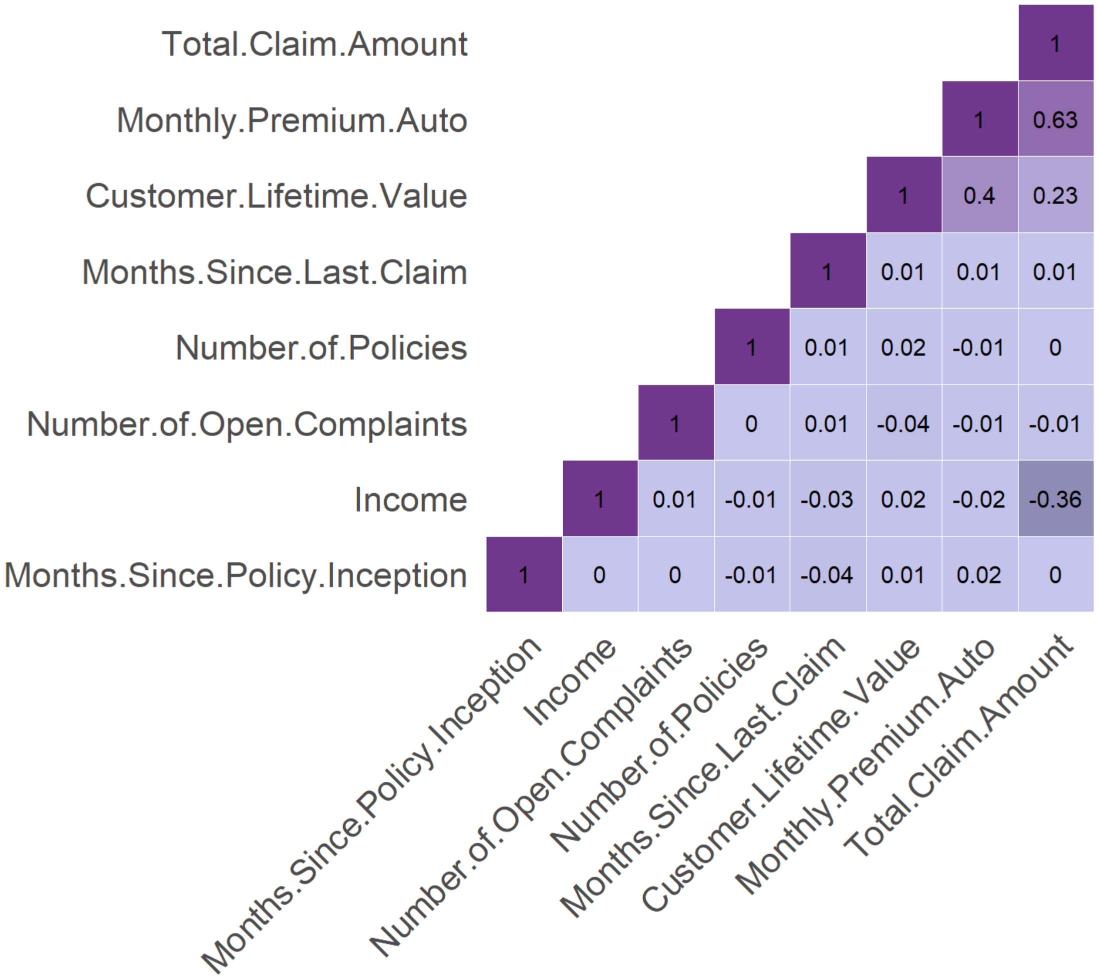


This plot shows the distribution of CLV for the customers based on the states they belong. We can see that the distributions are more or less similar. This points out that the state from which the customers come are not a valuable feature for determining their importance to the company.



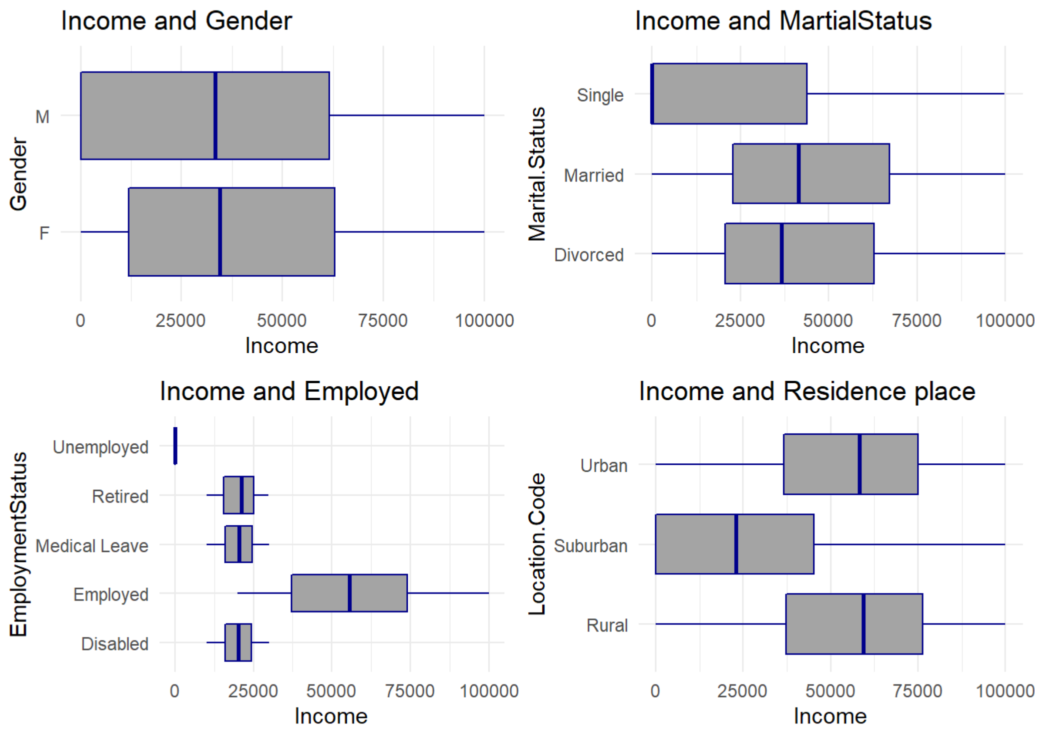
The above frequency distribution plot suggests that the target variable, CLV has more or less similar distribution irrespective of whether the response from the customers to marketting calls is positive or not.

Correlation between numerical variables.



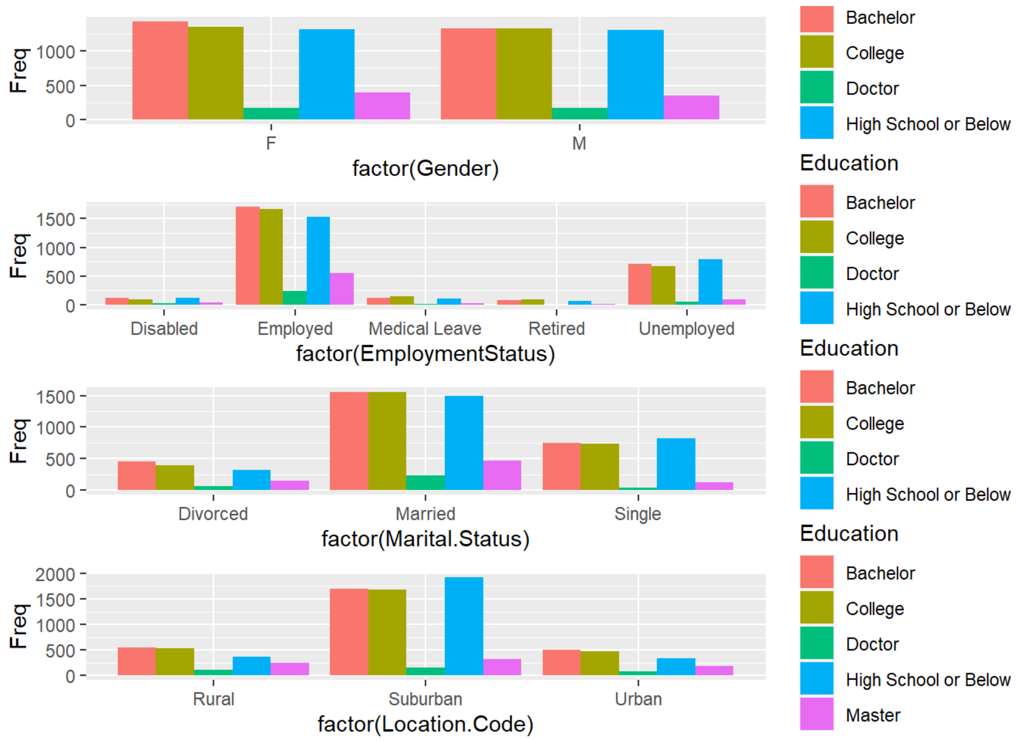
The above plot shows the correlation between numerical features of the data set. The “0” suggests that there is no correlation between Total claim amount with the variables : Income, Number of policies and Number of open complaints. We can also see that none of the pairs of variables have high correlation with the target variable and also between themselves.

Plotting of Categorical Variables To check Income of each catagories.



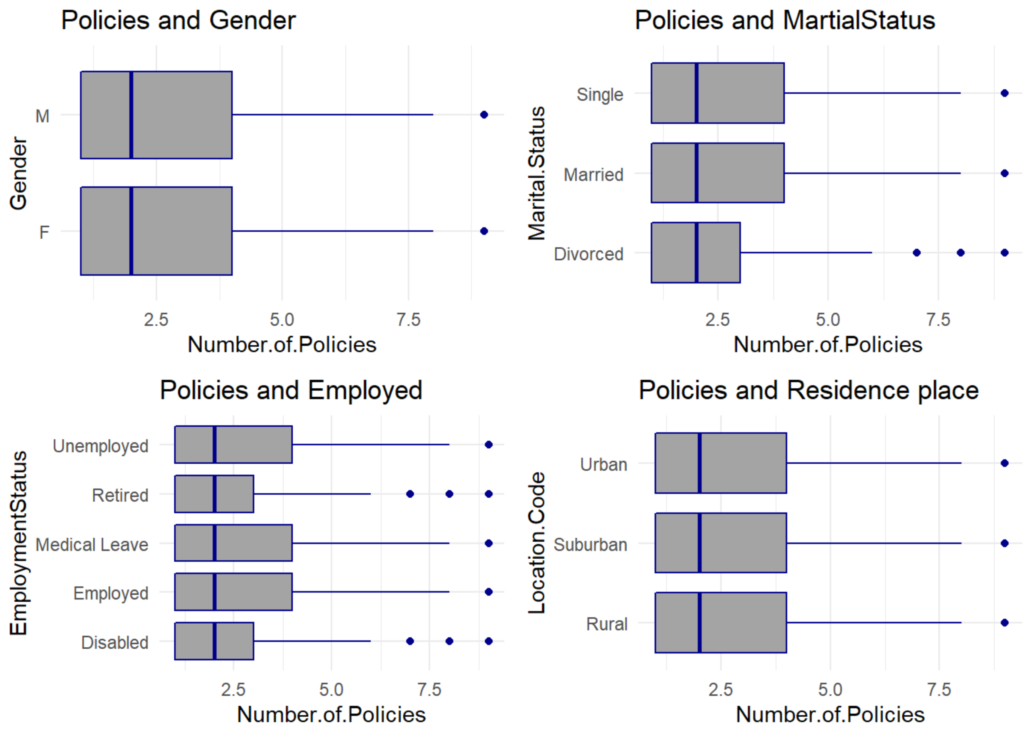
The above plots show that we do not see much difference in income for the Gender. We see that married and divorced customers have a greater income when compared to single customers. We can also see that employed customer have a higher income compared to other employees statuses. The last plot suggest that customers living in Urban and Rural residential areas have a greater income compared to suburban.

Plotting of Categorical Variables To check Education status for each categories.



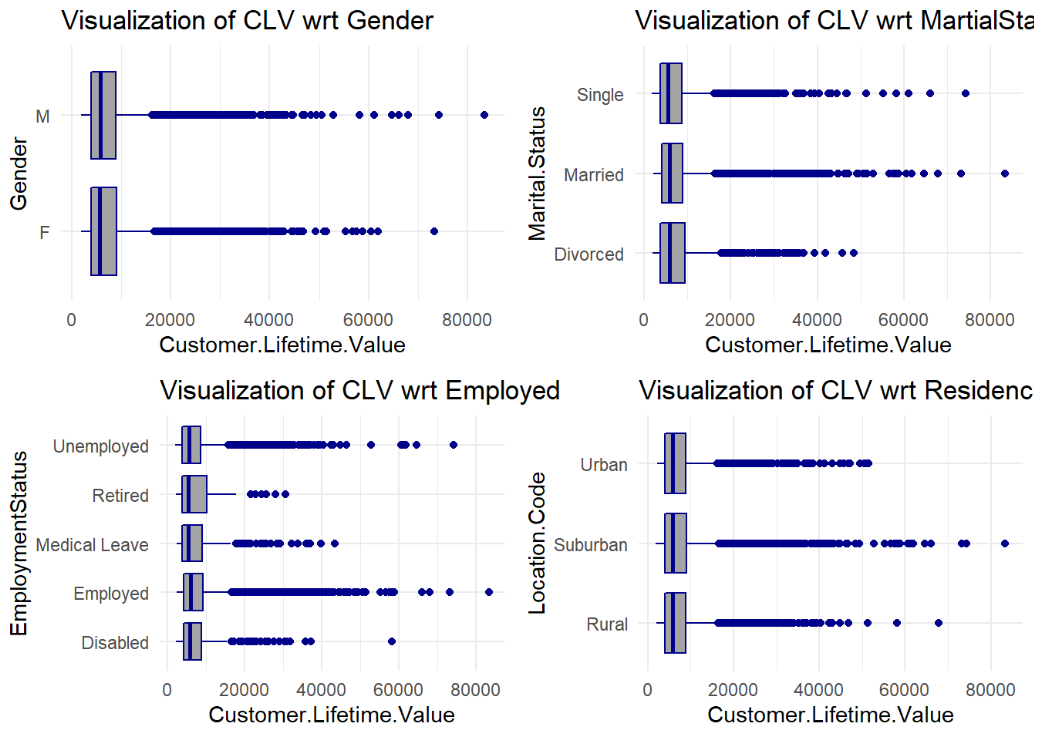
The above plots shows the basic educations of customers based on their Gender(male,female), Employment status ,Marital status and location code.

No of policies by each categories.



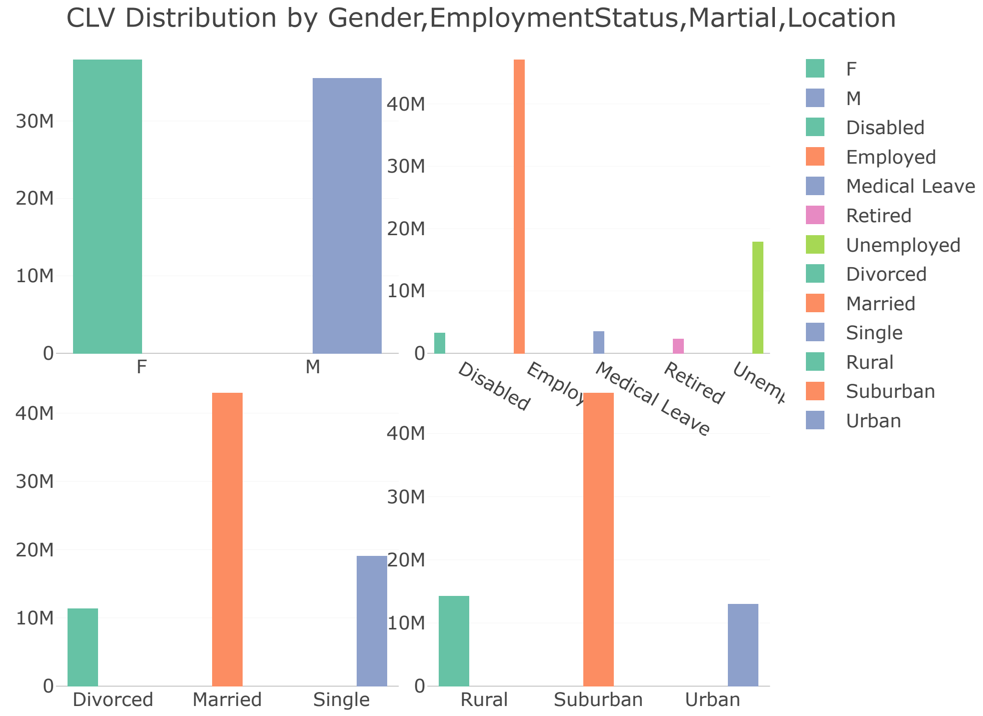
The above 4 plots suggest that there is not much difference when it comes to Number of policies for features like Gender,Marital status , Employement status and Location code.

Visualization of CLV by each categories.



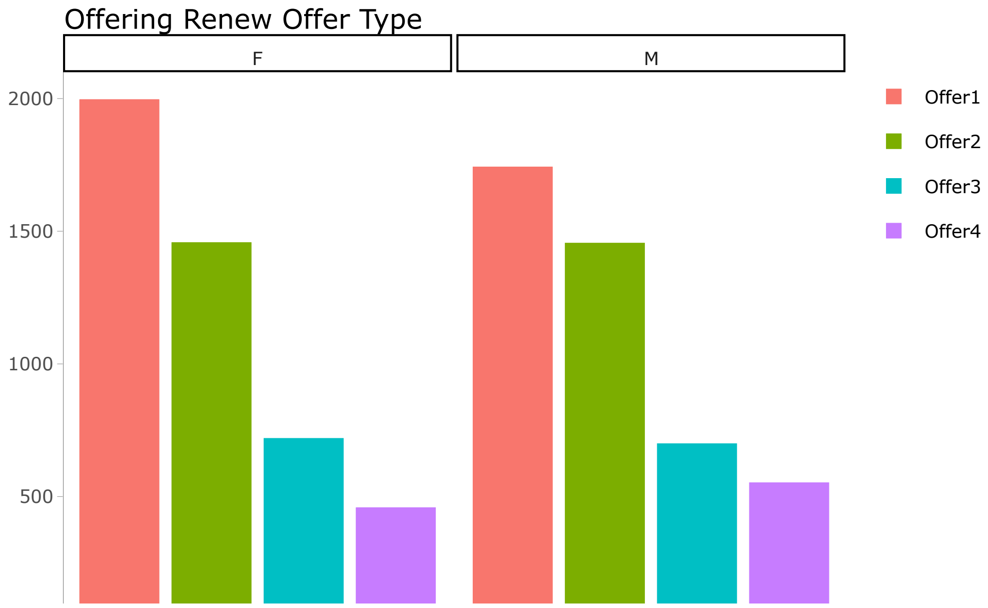
The above 4 plots suggest that there is not much difference when it comes to CLV for features like Gender,Marital status , Employment status and Location code.

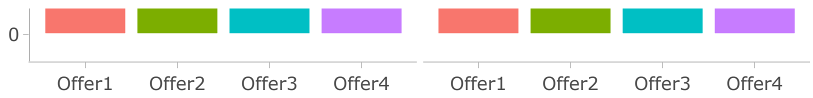
CLV Distribution by Gender,EmploymentStatus,Martial,Location



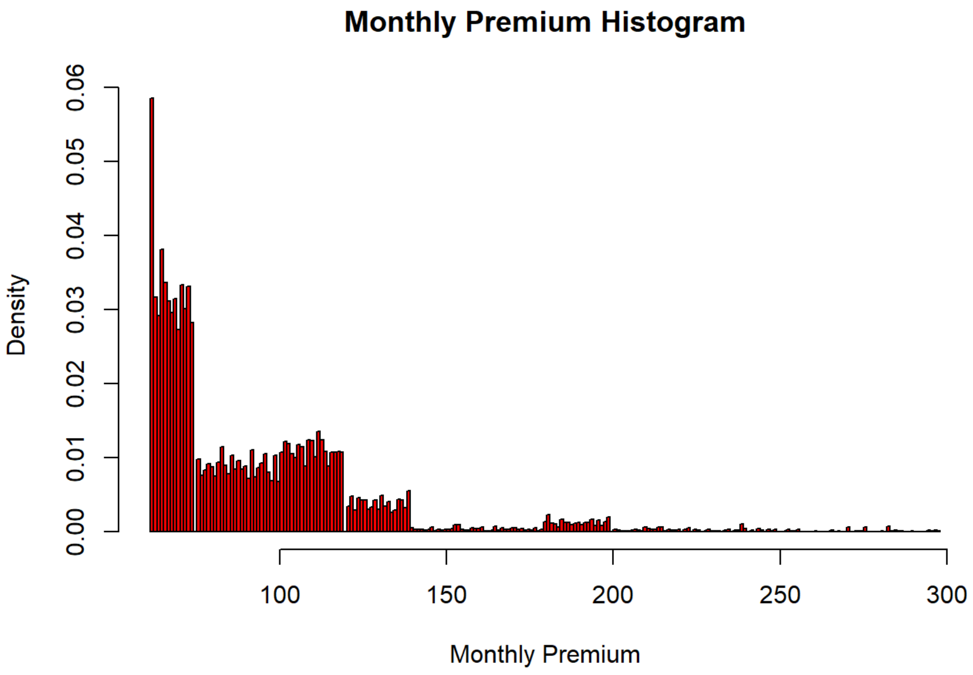
This plot shows the comparison of CLV in terms of Gender,Employment status,Marital Status and Residential Area.

Offer’s on Renew Vehicle insurance based on Gender.

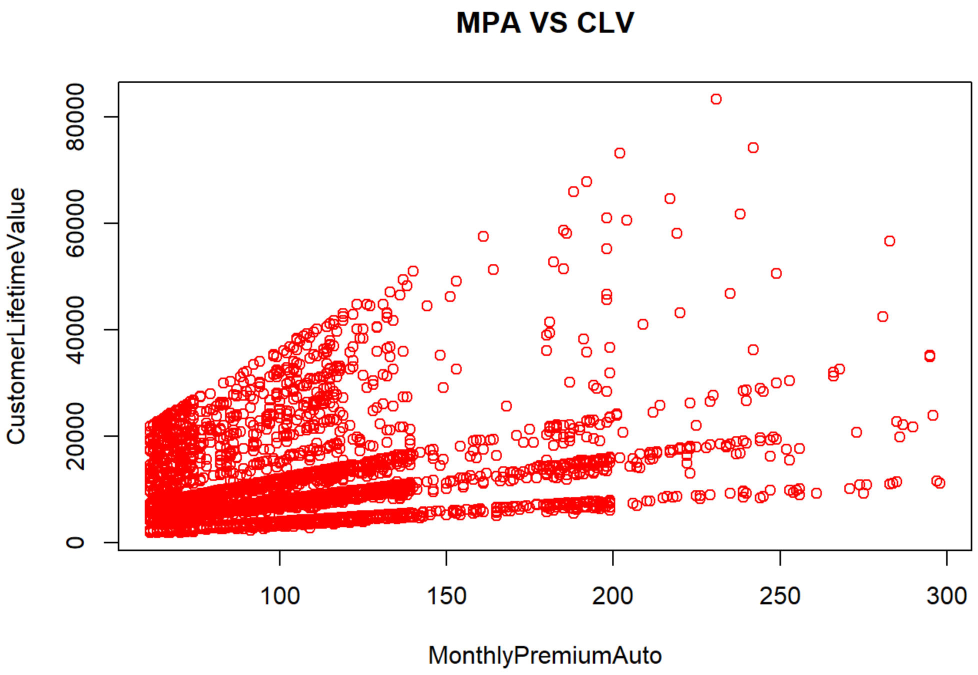




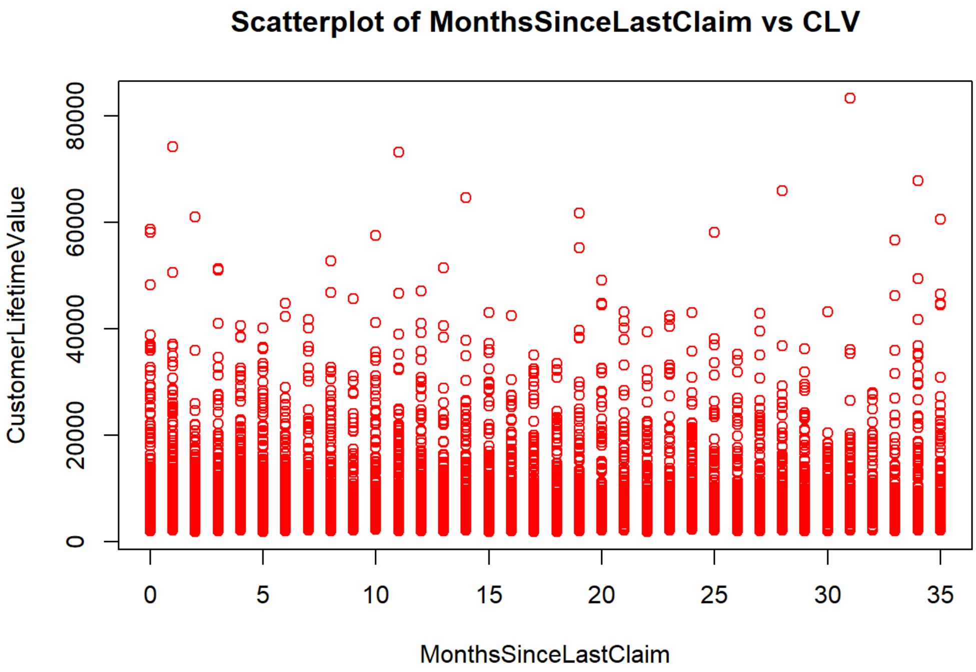
Based on bar plot above it can be new strategy for the company, How to offer the Renew Vehicle insurance based on Gender. Nevertheless, more or less nothing different between Female and Male when they decide to renew their insurance. The difference would be in Offer type 4, it shows that Male is more interested to renew their vehicle insurance in new offer type 4 than female.



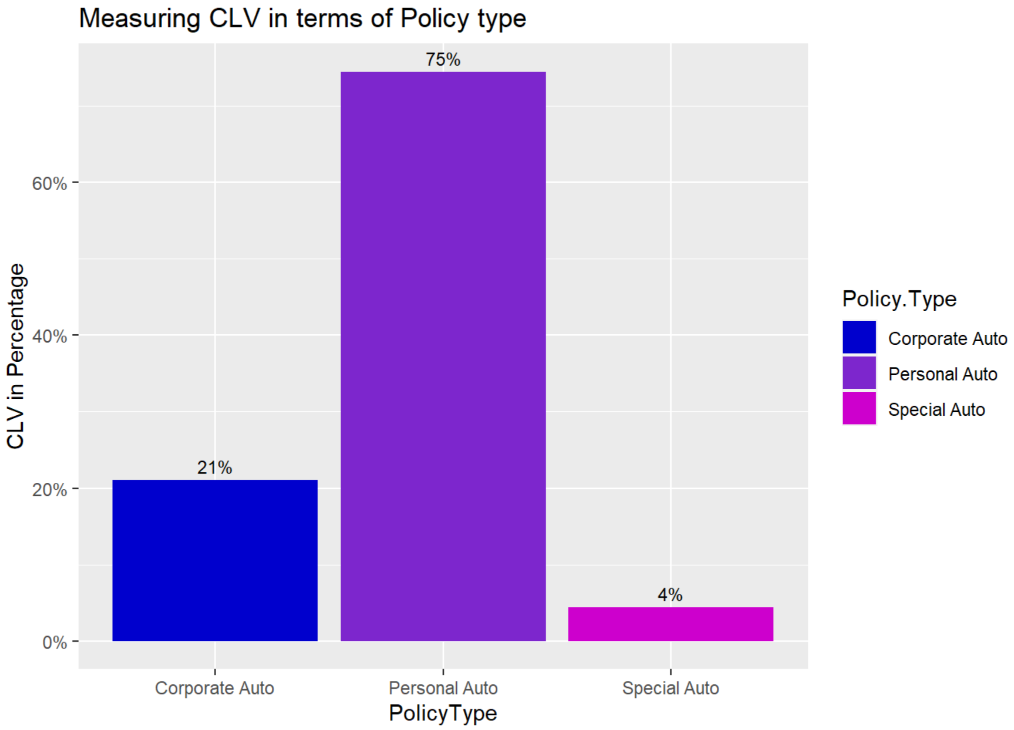
Mean of MPA(Monthly Premium Auto) is Mean of MPA is 93.21929 and the Median is 84.00 The Variance in MPA is 1183.908 and the Standard Deviation is 34.40797 There is a Positive Correlation of 39.62 % of MPA with CLV Kurtosis is 6.187546. Since kurtosis > 3, means distribution has thicker tails than normal MPA is positive Skewed



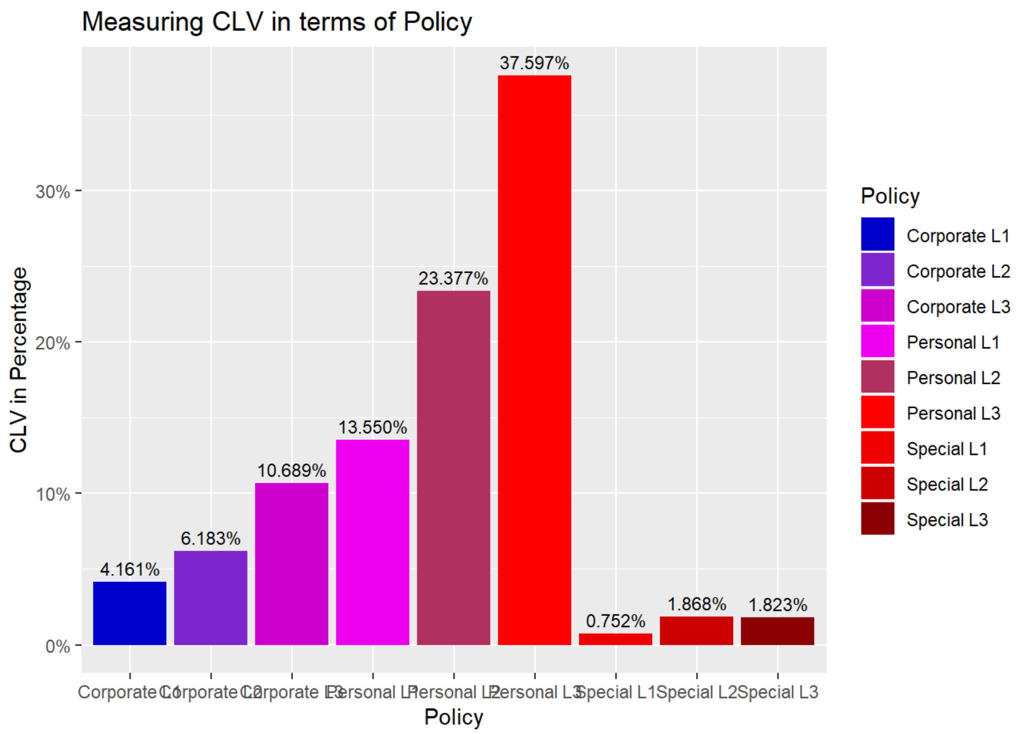
The scatter plot shows that With increase of MPA , CLV also increases.



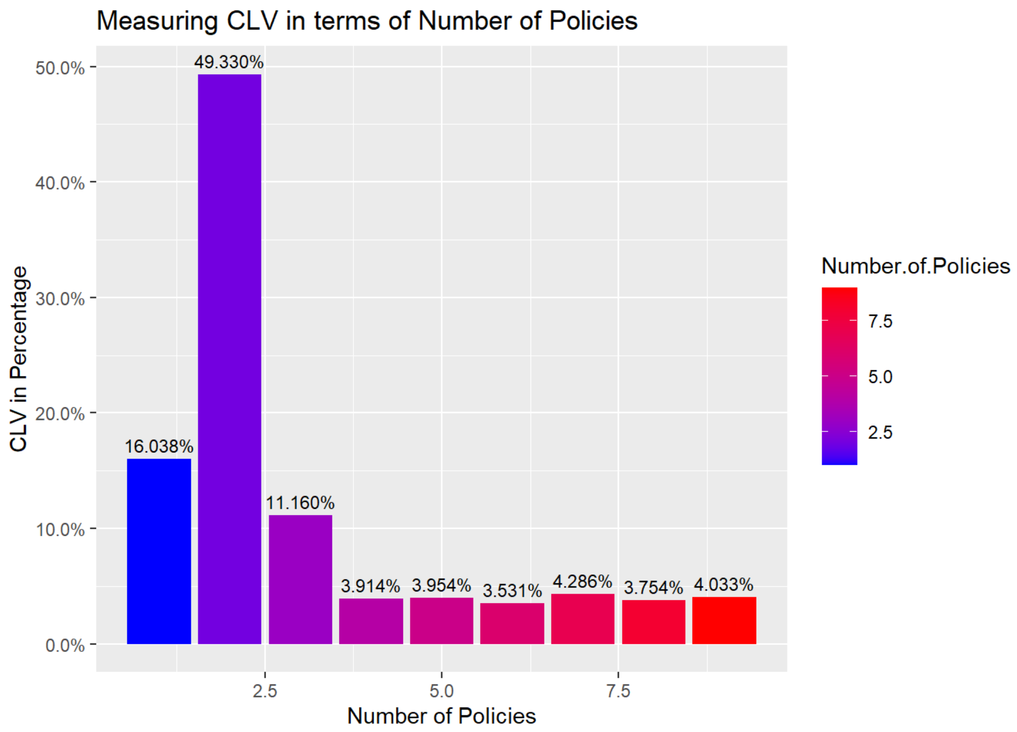
The positive correlation values close to zero show that that there is no strong relationship MonthsSinceLastClaim with CLV.

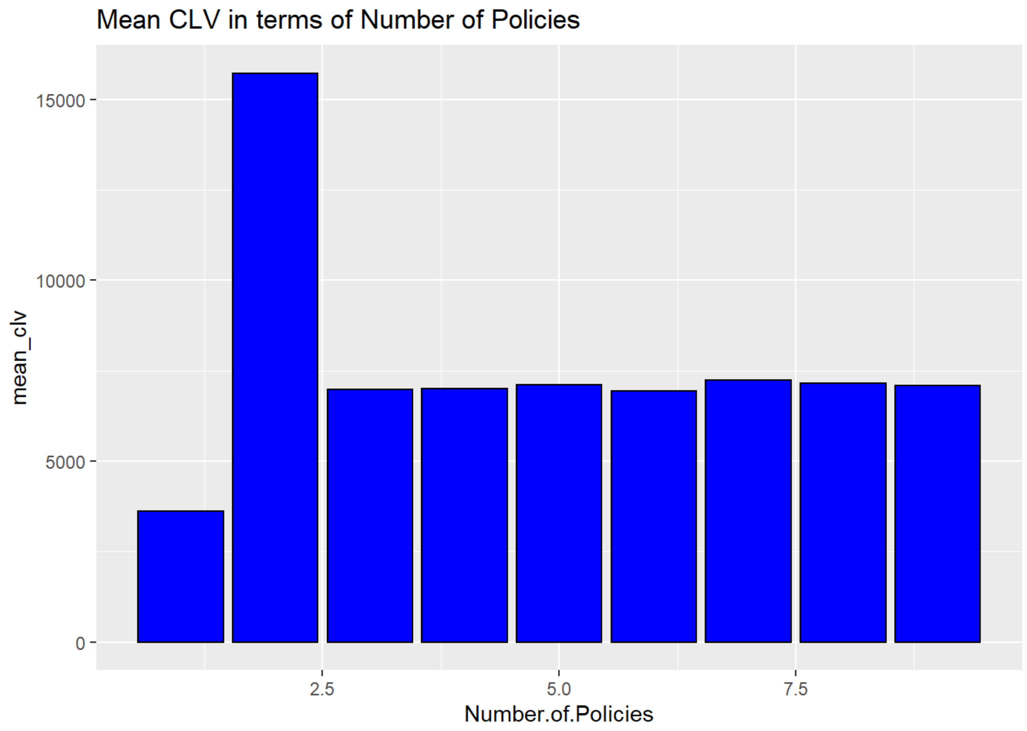


Customers who have the personal policy type seem to be more valuable to the company when compared to Corporate and special type.

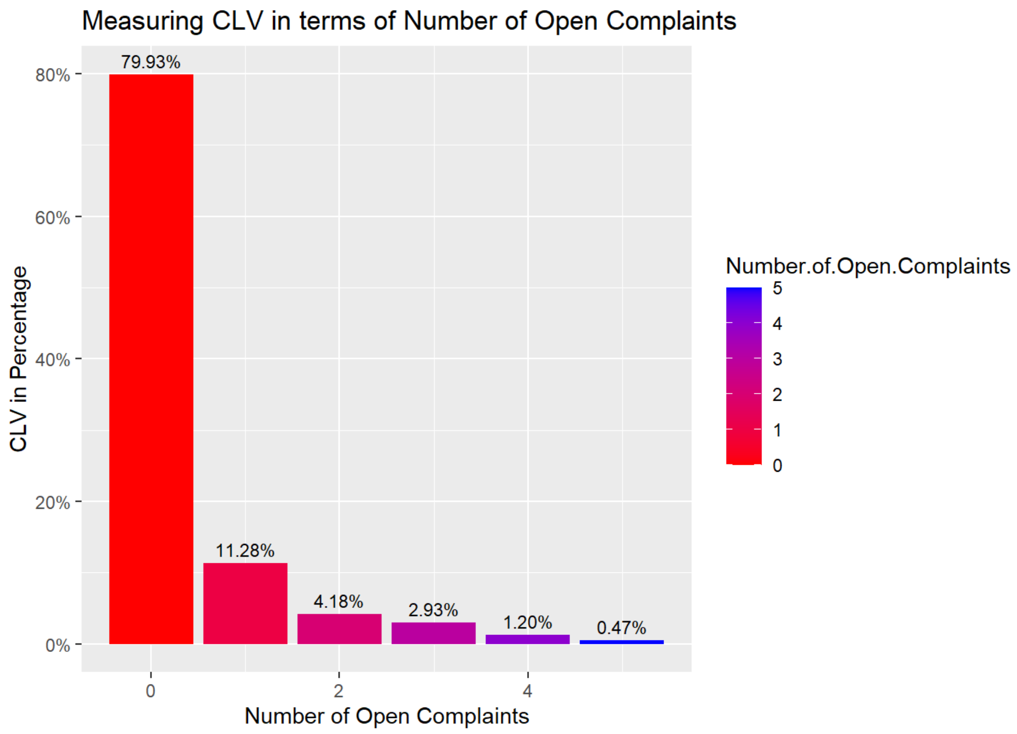


We had already established that customers with personal policy type were more valuable. Digging deeper, we see that personal L3 customers are highly valuable to the company.

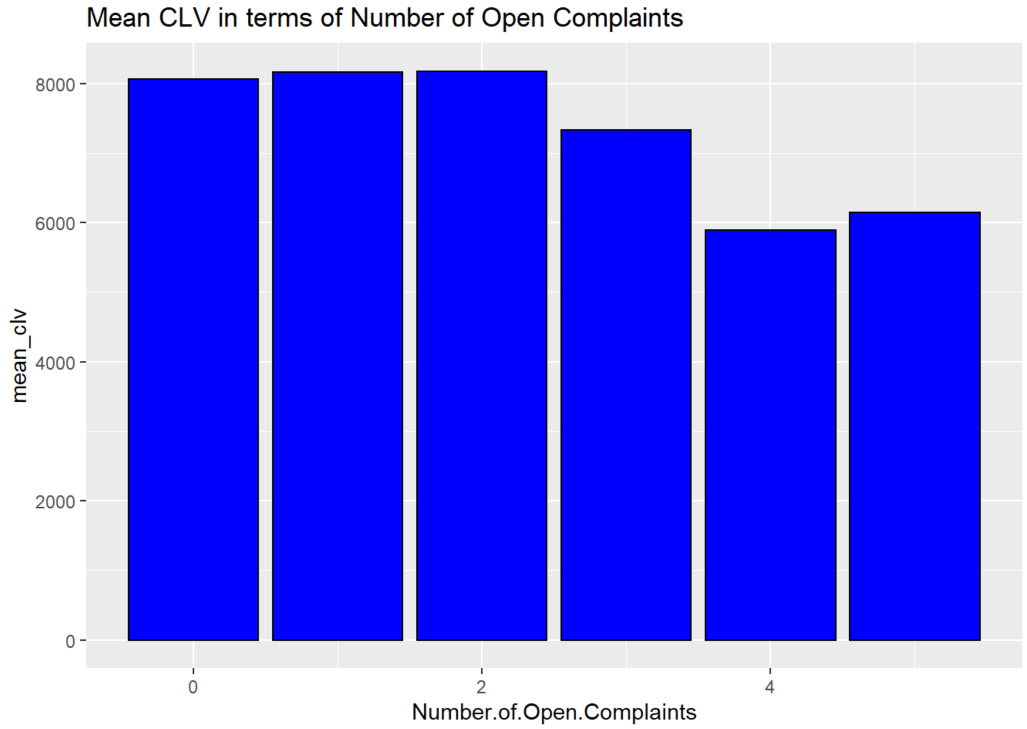




According to above two plot, customers with 2 Number of policies have higher CLV than that of others. It also suggests that the policy they opted for the first time was satisfactory hence they applied for the second one too. Customers who have more than 3 policies show a similar pattern in their CLV values. This may suggest that those customers having more than three policies may not significant in our model. This is visible in the second plot. While modelling a possible approach can be binning the Number.of.Policies variable. For eg : those with policies more than 3 can be combined into a single group.

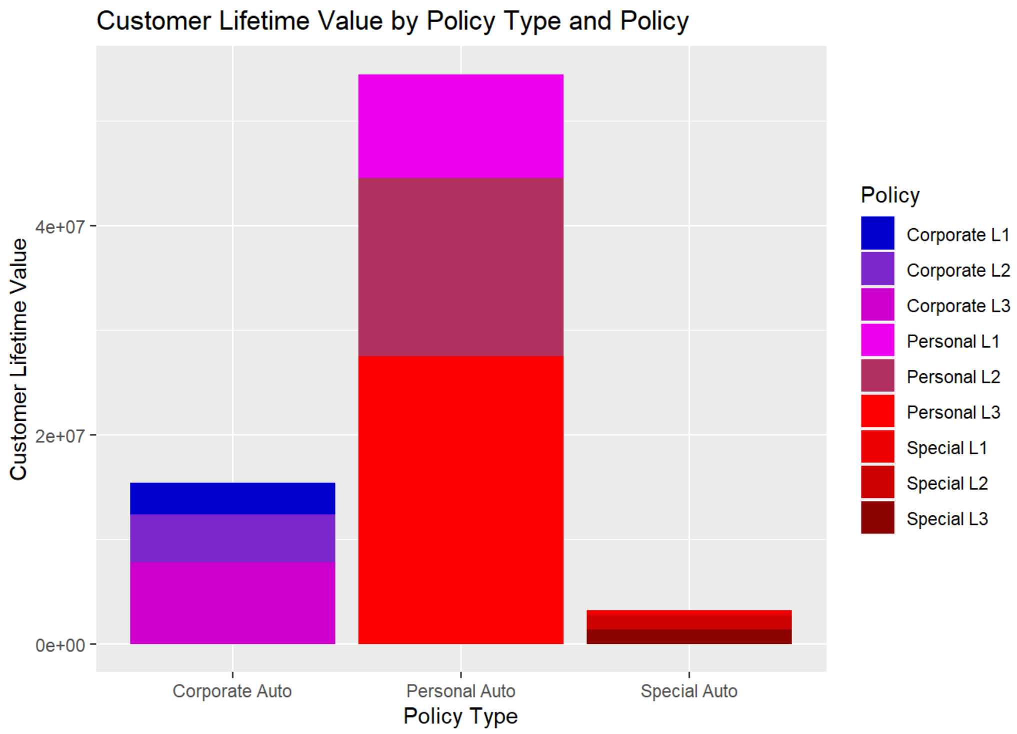


Measuring CLV in terms of Number of Open Complaints

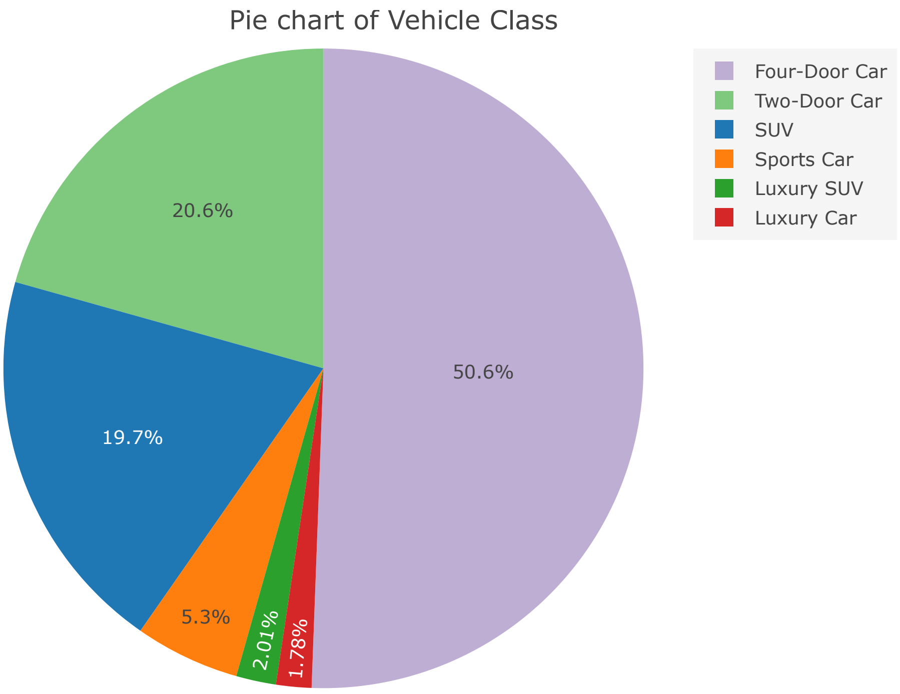


Mean CLV in terms of Number of Open Complaints

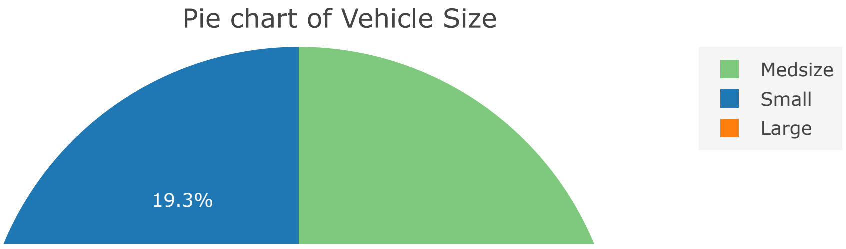
Most of the customers have no complaints with the service provided. The mean value of CLV for customers with number of complaints below 3 shows similar pattern and hence those customers with more number of complaints who also have relatively less mean CLV should be taken care for better profit for the organization. A similar approach of binning can be applied here also as that suggested for the variable Number.of.Policies.

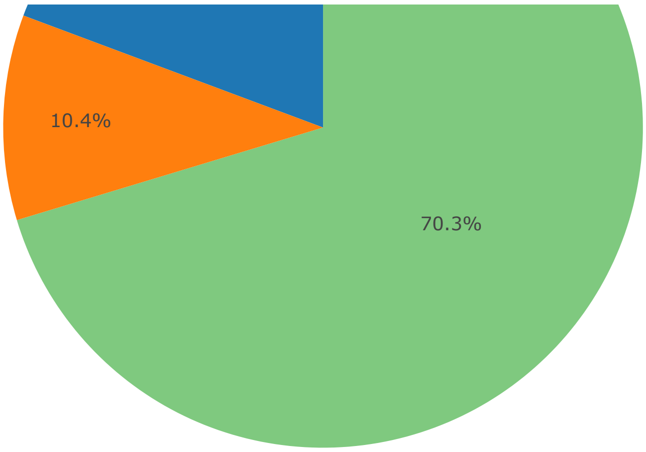


This bar plot Combines policy type and policy for Visualizing purposes. The personal policy type L3 has higher density among others.

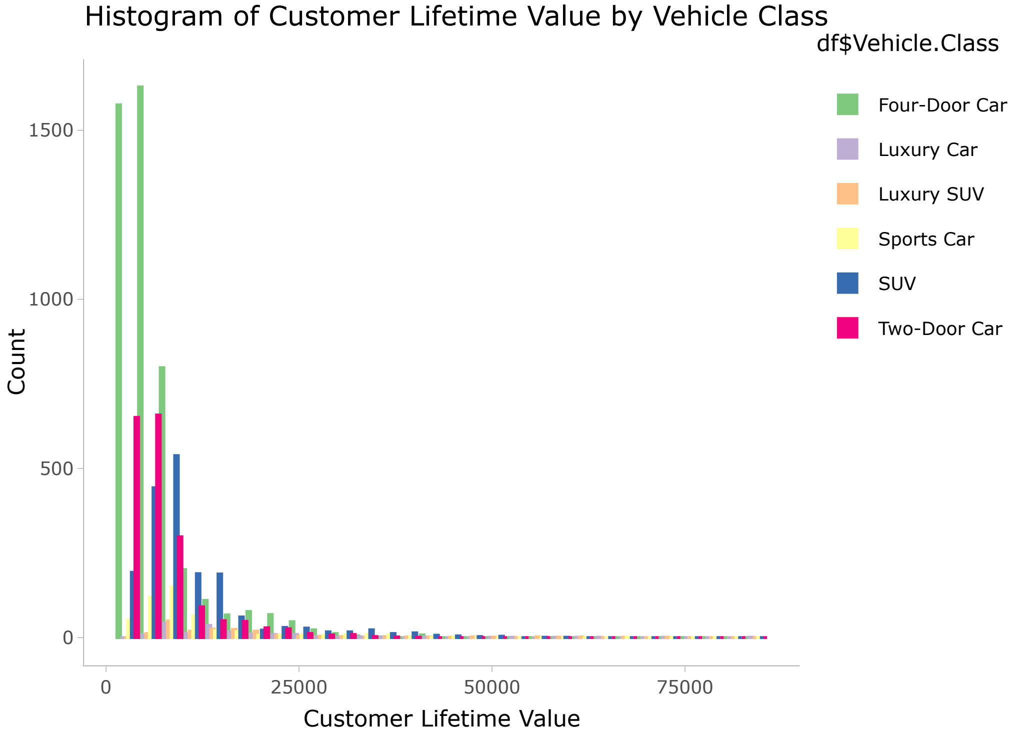


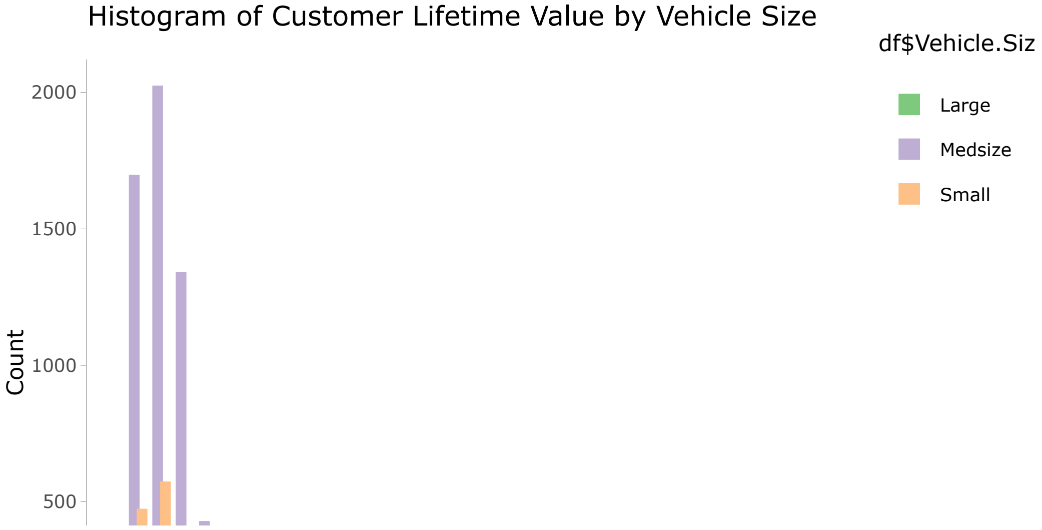
This pie chart shows the proportion of Vehicle Class. We see that this data set has 50% of customers with “four door cars”. Which may also suggest that there is a higher number of middle class customers.

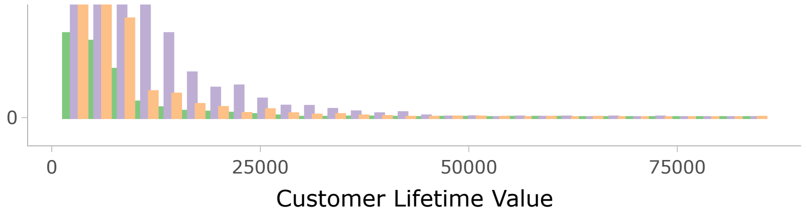




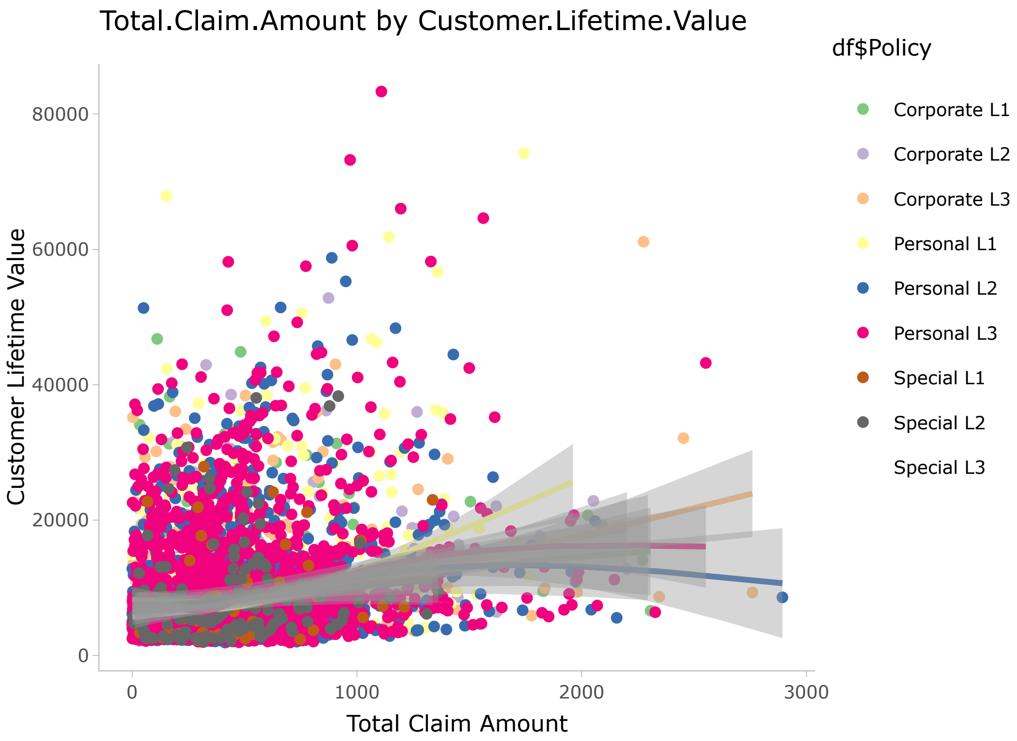
This pie chart shows the proportion of Vehicle size. We see that this data set has 70.3% of customers with “Medium Size”, which could support our previous claim.



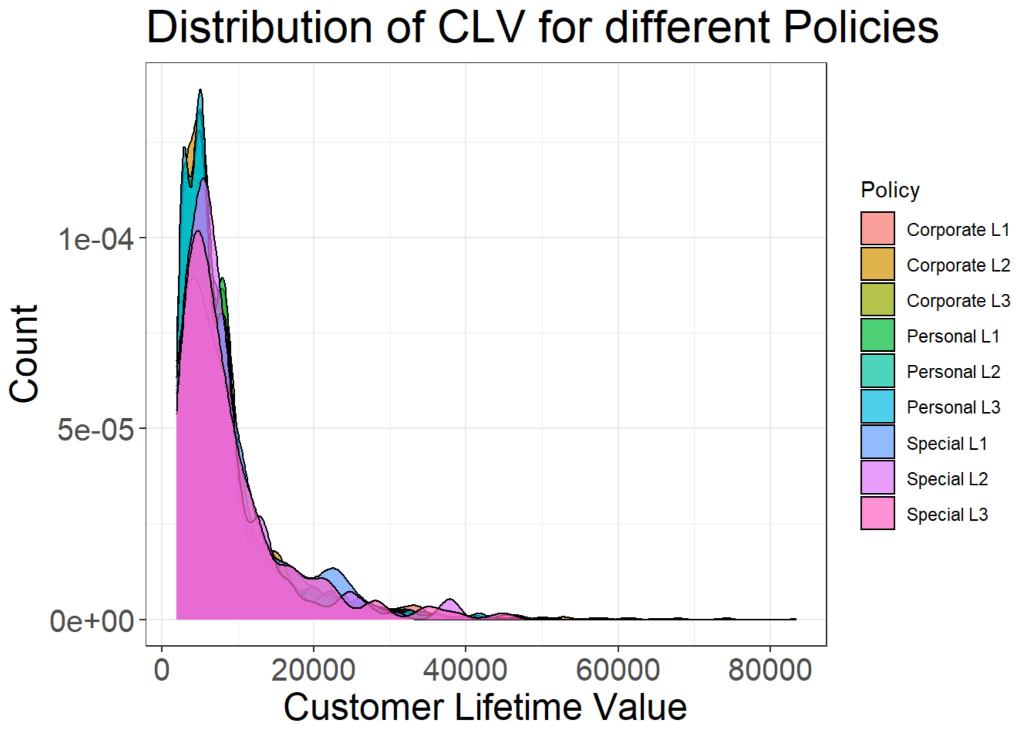


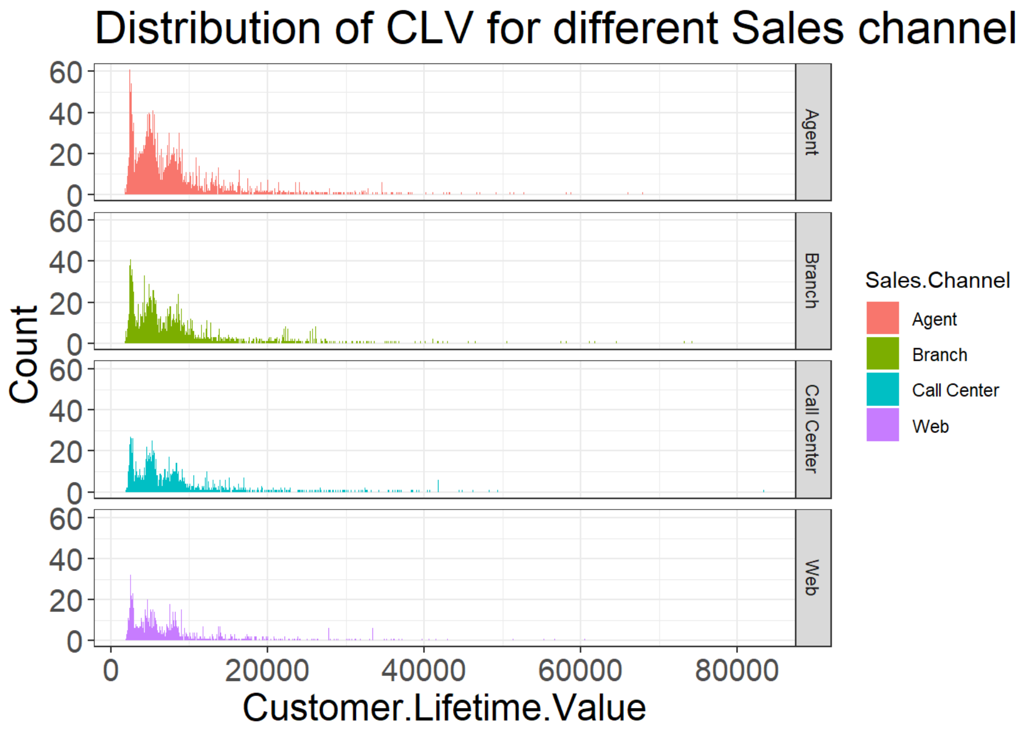


The above two plots show the distribution of CLV in terms of Vehicle size and Vehicle class.The customers who own luxury cars have relatively higher CLV when compared to others, which is obvious. But we have more number of regular middle-class customers.



This plot shows that CLV increases with the increase in total claim amount for different policies. But there is not prefect linear relationship.





This plot shows that the company must invest in agent and branches as their primary source of sales than Call center and web. The marketing through call centers and Web should be improved which can lead to more profit to the company.

Base model with all the variables.

Train Result

```
## [1] "MAE : 3871.59106307986"
## [1] "RMSE : 6241.96232879225"
## [1] "MSE : 38962093.7140616"
## [1] "R2 : 0.172241959232048"
## [1] "Adjusted R2 : 0.169741866302766"
```

Test Result

```
## [1] "MAE : 3922.20512370246"
## [1] "RMSE : 6410.31417826123"
## [1] "MSE : 41092127.8640169"
## [1] "R2 : 0.139383348571966"
## [1] "Adjusted R2 : 0.128888023554551"
```

Kruskal - Wallis test on categorical variables.

he Kruskal–Wallis test is a nonparametric method and hence does not assume a normal distribution of the residuals, unlike the analogous one-way analysis of variance.The test can be implemented in R using the kruskal.test(x,g) function. The x parameter is a continuous (interval/ratio) variable. The g parameter is the categorical variable representing different groups to which the continuous values belong.The test does not identify where this stochastic dominance occurs or for how many pairs of groups stochastic dominance obtains. Even if we use ANOVA we get similar results.

The null hypothesis with a Kruskal-Wallace test is that all the different groups represented by the samples are very similar based on the median value.

Here let us see how the test works with categorical variables with two examples.

State

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  df$Customer.Lifetime.Value and as.factor(df$State)
## Kruskal-Wallis chi-squared = 5.0721, df = 4, p-value = 0.28
```

```
##     df$State df$Customer.Lifetime.Value
## 1    Arizona                   7861.341
## 2 California                   8003.648
## 3     Nevada                   8056.707
## 4     Oregon                   8077.901
## 5 Washington                   8021.472
```

There is 28% chance that the means are same.Therefore, we fail to reject the null hypothesis. There is no significant difference in CLV among the customers from different States.Thus State variable can be avoided in our model. We can also infer that there is no discernible difference in the mean values of CLV for each State.

Coverage

```
## 
##  Kruskal-Wallis rank sum test
## 
## data:  df$Customer.Lifetime.Value and as.factor(df$Coverage)
## Kruskal-Wallis chi-squared = 502.5, df = 2, p-value < 2.2e-16
```

```
##   df$Coverage df$Customer.Lifetime.Value
## 1       Basic                   7190.706
## 2    Extended                   8789.678
## 3     Premium                  10895.603
```

The p value here is < 0.05.Therefore, we reject the null hypothesis. There is significant difference in CLV among the customers with different policy coverages.Thus Coverage variable may add value to our model. By observing the mean of CLV for different policy coverages, it can be inferred that Extended and Premium policy customers are an asset to the organization.

Therefore based on Kruskal Wallace test, the categorical variables that would help in predicting the CLV are:

- Vehicle.Size
- Vehicle.Class
- Renew.Offer.Type
- Marital.Status
- Coverage

- Education
- EmploymentStatus

Feature Engineering and Feature Selection.

Feature Engineering.

We have implemented different models with different transformations of the variables for predicting the Customer Lifetime Value. The transformations that have been attempted with numerical variables includes :

- Log and sqrt transformations on the continous independent variables.
- Log and sqrt transformations on the dependent variable.
- Binning of the continous variables based on their distribution.
- Binning of the variables : Number.of.Open.Complaints and Number.of.Policies as mentioned in the EDA.
- Label encoding the categorical features. R provides us with ‘superml‘ package that contains a set of functions to apply Label Encoder to our data.

Feature selection.

Relative Importance is a technique that is specific to linear regression models.

Relative importance can be used to assess which variables contributed how much in explaining the linear model’s R-squared value. So, if you sum up the produced importances, it will add up to the model’s R- squared value. In essence, it is not directly a feature selection method, because you have already provided the features that go in the model. But after building the model, the relaimpo package can provide a sense of how important each feature is in contributing to the R- squared, or in other words, in ‘explaining the Y variable’.It is implemented in the relaimpo package. Basically, you build a linear regression model and pass that as the main argument to calc.relimp(). The relaimpo has multiple options to compute the relative importance, but the recommended method is to use type=‘lmg’, and we have used the same for feature selection.

Step wise Forward and Backward Selection.

It searches for the best possible regression model by iteratively selecting and dropping variables to arrive at a model with the lowest possible AIC. It can be implemented using the step() function and you need to provide it with a lower model, which is the base model from which it won’t remove any features and an upper model, which is a full model that has all possible features you want to have.

Modelling.

Lasso Regression.

Here we have used all the features available.

Train Result

```
## [1] "MAE : 3871.59106307986"
## [1] "RMSE : 6241.96232879225"
## [1] "MSE : 38962093.7140616"
## [1] "R2 : 0.172241959232048"
## [1] "Adjusted R2 : 0.169741866302766"
```

Test Result

```
## [1] "MAE : 3922.20512370246"
## [1] "RMSE : 6410.31417826123"
## [1] "MSE : 41092127.8640169"
## [1] "R2 : 0.139383348571966"
## [1] "Adjusted R2 : 0.128888023554551"
```

Stepwise Regression.

Train Result

```
## [1] "MAE : 3871.42551414636"
## [1] "RMSE : 6245.92550222668"
## [1] "MSE : 39011585.3793656"
## [1] "R2 : 0.171190498183599"
## [1] "Adjusted R2 : 0.168687229507053"
```

Test Result

```
## [1] "MAE : 3919.26780041461"
## [1] "RMSE : 6408.03198748838"
## [1] "MSE : 41062873.9526743"
## [1] "R2 : 0.139996030477955"
## [1] "Adjusted R2 : 0.129508177191101"
```

Polynomial Regression on selected Variables.

Based on the EDA and feature selection methods the variables used in building the polynomial regression model are : Income , Monthly.Premium.Auto , poly(Number.of.Policies,5), Coverage, EmploymentStatus, Renew.Offer.Type, Vehicle.Size, Vehicle.Class, Marital.Status and Education.

Train Result

```
## [1] "MAE : 2584.50345837054"
## [1] "RMSE : 4306.71655779153"
## [1] "MSE : 18547807.5091557"
## [1] "R2 : 0.605947849800001"
## [1] "Adjusted R2 : 0.604757686798299"
```

Test Result

```
## [1] "MAE : 2662.66955678632"
## [1] "RMSE : 4423.50642411934"
## [1] "MSE : 19567409.0842251"
## [1] "R2 : 0.590188219531591"
## [1] "Adjusted R2 : 0.585190514891732"
```

XGBoost Regression.

The variables used in XGBoost model are : Vehicle.Class, Coverage, Renew.Offer.Type, EmploymentStatus, Policy.Type, Monthly.Premium.Auto, Number.of.Open.Complaints, Total.Claim.Amount, Number.of.Policies, Income and Education.

Train Result

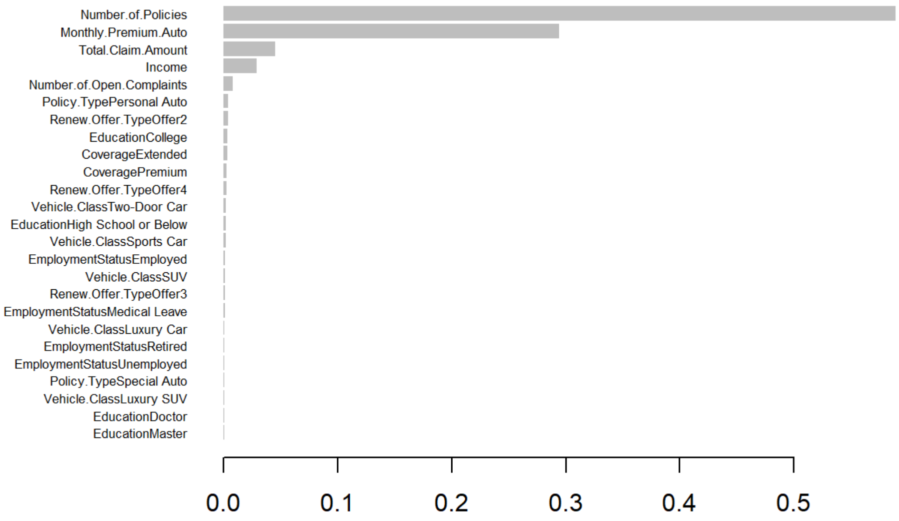
```
## [1] "MAE : 1316.14598626418"
## [1] "RMSE : 2909.08410449344"
## [1] "MSE : 8462770.3270164"
## [1] "R2 : 0.823283591115292"
## [1] "Adjusted R2 : 0.823017233140815"
```

Test Result

```
## [1] "MAE : 1633.98400708824"
## [1] "RMSE : 3640.70099583986"
## [1] "MSE : 13254703.7411094"
## [1] "R2 : 0.701888792154881"
## [1] "Adjusted R2 : 0.700079066279441"
```

Feature Importance

```
##                        Feature         Gain       Cover   Frequency
##  1:          Number.of.Policies 0.5890670138 0.117021206 0.080800772
##  2:         Monthly.Premium.Auto 0.2939819705 0.313251959 0.247708635
##  3:          Total.Claim.Amount 0.0454546654 0.218050711 0.178967680
##  4:                      Income 0.0291492083 0.204920679 0.165701881
##  5:     Number.of.Open.Complaints 0.0078809364 0.038151506 0.068982151
##  6:      Policy.TypePersonal Auto 0.0036945356 0.004585525 0.017366136
##  7:        Renew.Offer.TypeOffer2 0.0035931683 0.001976270 0.020742885
##  8:            EducationCollege 0.0035505528 0.001706600 0.010371442
##  9:            CoverageExtended 0.0032933913 0.006331472 0.024602026
## 10:             CoveragePremium 0.0025567925 0.005135678 0.011336228
## 11:        Renew.Offer.TypeOffer4 0.0022332308 0.003534000 0.011336228
## 12:     Vehicle.ClassTwo-Door Car 0.0019955431 0.000944072 0.014471780
## 13: EducationHigh School or Below 0.0019754082 0.003549416 0.006753497
## 14:       Vehicle.ClassSports Car 0.0016069991 0.007375634 0.014712976
## 15:      EmploymentStatusEmployed 0.0015136332 0.019128789 0.031114327
## 16:             Vehicle.ClassSUV 0.0014643702 0.001547376 0.014471780
## 17:        Renew.Offer.TypeOffer3 0.0011973708 0.007706278 0.012059817
## 18: EmploymentStatusMedical Leave 0.0011792107 0.001354098 0.005065123
## 19:       Vehicle.ClassLuxury Car 0.0008404111 0.001586722 0.014712976
## 20:       EmploymentStatusRetired 0.0007782971 0.009556687 0.009889050
## 21:    EmploymentStatusUnemployed 0.0007681527 0.004514426 0.016160154
## 22:       Policy.TypeSpecial Auto 0.0006328992 0.006314675 0.003859141
## 23:       Vehicle.ClassLuxury SUV 0.0005620279 0.001372735 0.006753497
## 24:              EducationDoctor 0.0005215955 0.001138501 0.004582730
## 25:              EducationMaster 0.0005086154 0.019244986 0.007477086
##                        Feature         Gain       Cover   Frequency
```



Final Model.

The final model takes in the variables :

Monthly.Premium.Auto, I(Coverage == “Premium”), Total.Claim.Amount, Income, I(Number.of.Policies == 1), I(Number.of.Policies == 2), I(EmploymentStatus == “Employed”), I(Policy.Type == “Special Auto”), I(Number.of.Open.Complaints == 3), I(Number.of.Open.Complaints == 4), I(Number.of.Open.Complaints == 5), I(Vehicle.Class== “Sports Car”).

Model Summary

```
## 
## Call:
## lm(formula = Customer.Lifetime.Value ~ Monthly.Premium.Auto + 
##     I(Coverage == "Premium") + Total.Claim.Amount + Income + 
##     I(Number.of.Policies == 1) + I(Number.of.Policies == 2) + 
##     I(EmploymentStatus == "Employed") + I(Policy.Type == "Special Auto") + 
##     I(Number.of.Open.Complaints == 3) + I(Number.of.Open.Complaints == 
##     4) + I(Number.of.Open.Complaints == 5) + I(Vehicle.Class == 
##     "Sports Car"), data = traindf3)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
##  -8034   -792     12    645  55959 
## 
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -8.808e+02  1.724e+02  -5.108 3.33e-07
## Monthly.Premium.Auto                8.362e+01  2.048e+00  40.831  < 2e-16
## I(Coverage == "Premium")TRUE       -5.902e+02  1.856e+02  -3.180  0.00148
## Total.Claim.Amount                 -3.213e-01  2.454e-01  -1.309  0.19042
## Income                              3.836e-03  2.712e-03   1.414  0.15731
## I(Number.of.Policies == 1)TRUE     -3.623e+03  1.136e+02 -31.889  < 2e-16
## I(Number.of.Policies == 2)TRUE      8.623e+03  1.253e+02  68.841  < 2e-16
## I(EmploymentStatus == "Employed")TRUE  4.383e+02  1.679e+02   2.610  0.00907
## I(Policy.Type == "Special Auto")TRUE   2.220e+02  2.493e+02   0.891  0.37321
## I(Number.of.Open.Complaints == 3)TRUE -6.136e+02  2.877e+02  -2.132  0.03301
## I(Number.of.Open.Complaints == 4)TRUE -7.730e+02  3.960e+02  -1.952  0.05098
## I(Number.of.Open.Complaints == 5)TRUE -1.187e+03  5.954e+02  -1.994  0.04621
## I(Vehicle.Class == "Sports Car")TRUE   3.929e+02  2.246e+02   1.750  0.08023
## 
## (Intercept)                        ***
## Monthly.Premium.Auto               ***
## I(Coverage == "Premium")TRUE       ** 
## Total.Claim.Amount                    
## Income                                
## I(Number.of.Policies == 1)TRUE     ***
## I(Number.of.Policies == 2)TRUE     ***
## I(EmploymentStatus == "Employed")TRUE ** 
## I(Policy.Type == "Special Auto")TRUE    
## I(Number.of.Open.Complaints == 3)TRUE * 
## I(Number.of.Open.Complaints == 4)TRUE . 
## I(Number.of.Open.Complaints == 5)TRUE * 
## I(Vehicle.Class == "Sports Car")TRUE  . 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4192 on 7297 degrees of freedom
## Multiple R-squared:  0.6338, Adjusted R-squared:  0.6332 
## F-statistic:  1052 on 12 and 7297 DF,  p-value: < 2.2e-16
```

Train Result

```
## [1] "MAE : 2077.39229479091"
## [1] "RMSE : 4187.85907289204"
## [1] "MSE : 17538163.6144041"
## [1] "R2 : 0.633774618404111"
## [1] "Adjusted R2 : 0.633222620706447"
```

Test Result

```
## [1] "MAE : 2011.60854846471"
## [1] "RMSE : 4001.44979156014"
## [1] "MSE : 16011600.4343767"
## [1] "R2 : 0.639883498095751"
## [1] "Adjusted R2 : 0.637697360391034"
```

Repeated 10 fold Cross Validation.

```
## Linear Regression
##
## 9134 samples
##    9 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 8221, 8221, 8220, 8220, 8222, 8220, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4145.278  0.6370282  2057.793
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Final Model Interpretation :-

Null Hypothesis - None of the independed variables are significant for CLV.

Alternate Hypothesis – At least some of the independent variables are significant and can effect the CLV.

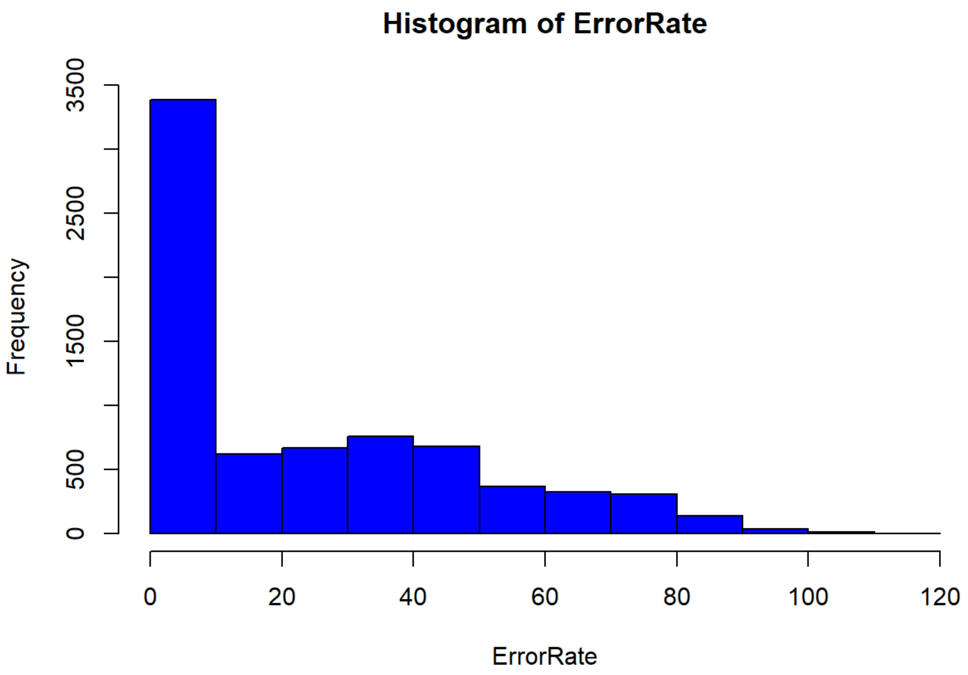
1. p-value of model is less than 0.05, so atleast one of the independent variables are significant.
2. p-value of MonthlyPremiumAuto, NumberofOpenComplaints, NumberofPolicies (Buyer’s == 2 and 1),Coverage(“Premium”), and EmploymentStatus (“People who are Employed”) are less than 0.05. So atlest one of them independed variables are significant and can effect the CLV.
3. However R squared is 63.38% which means that 93.8% of the actual variance in CLV can be explained by our multiple linear regression model.
4. Adjusted R squared is 0.6329 which is less than R squared and is the best among all the models we had. We have seen that eventhough XGBoost gives higher adjusted R square, its highly overfitting.
5. Residual standard error is 4193 which is average, so it means the actual CLV will deviate from the true regression line by approximately 4193 on an average. The smaller the standard error, the less the spread and the more likely it is that any sample mean is close to the population mean. A small standard error is thus a Good Thing.F-statistic: The lower the F-statistic, the closer to a non-significant model. So F-statistic in Final Model is high means it is significant model.
6. The results of cross validation also suggests that our final model can prove to be good with the unseen data points.

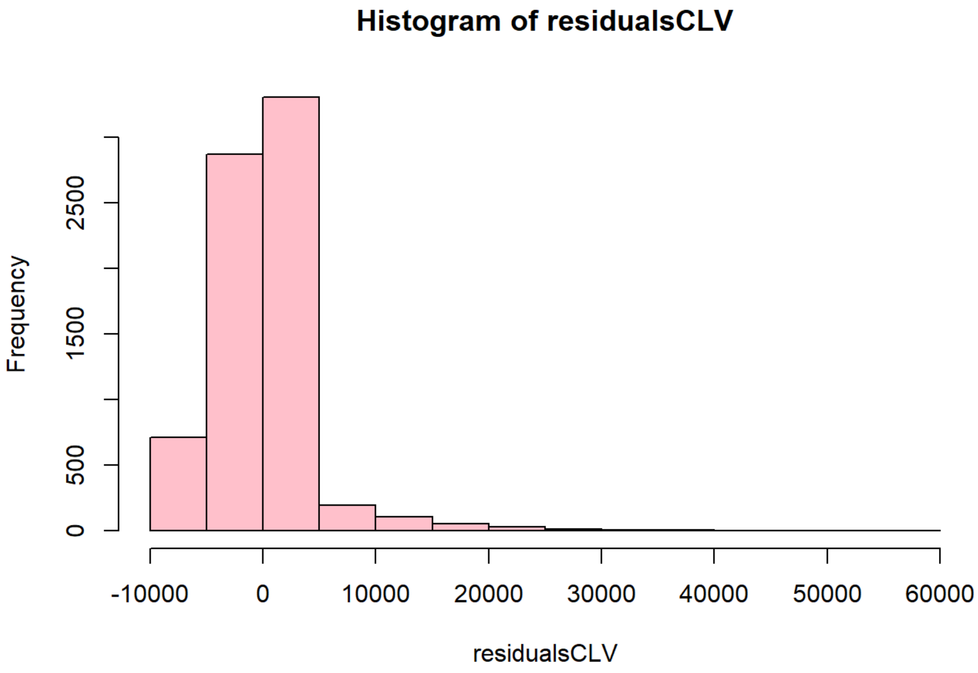
Model Valuation and Checking Assumptions of Linear Regression.

Average error rate of model is 22.9093%, which is low and we can say that model is good.

```
## [1] 23.95436
```

Normality of Residuals.





Shapiro Test.

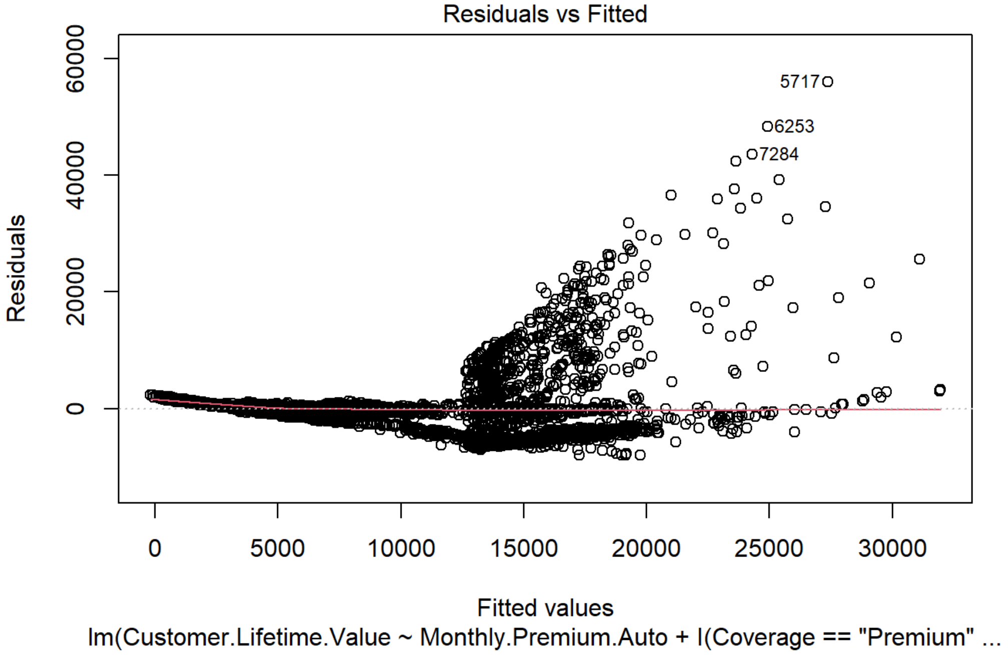
Null Hypotheses - Errors are normally distributed.

Alt Hypothese - Errors are not normally distributed.

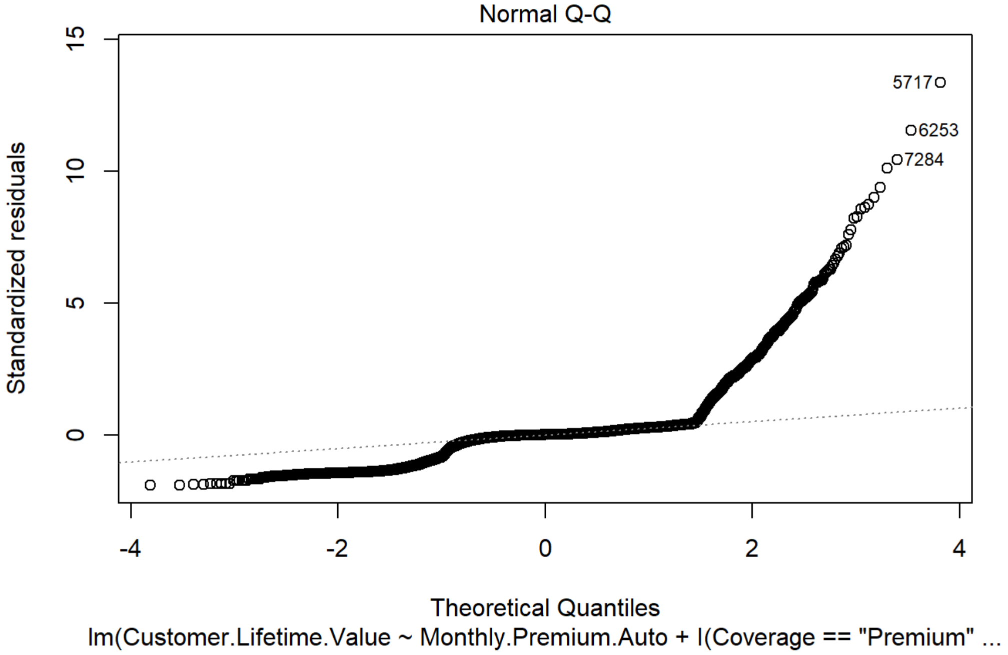
```
##
##  Shapiro-Wilk normality test
##
## data:  residualsCLV[0:5000]
## W = 0.6857, p-value < 2.2e-16
```

p-value < 0.05, Null Hypotheses get rejected, and so the errors are not normally distributed.

Residuals vs Fitted Plot



Here we can observe a recognizable pattern. This suggests either non-linearity or that other attributes have not been adequately captured.

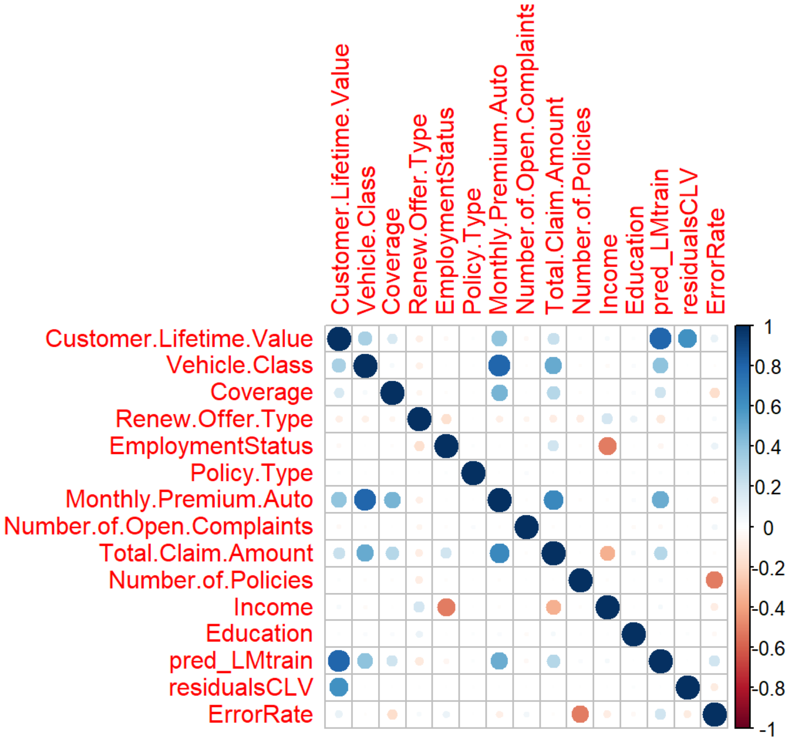


The Q-Q plot plots the distribution of our residuals against the theoretical normal distribution. There is strong snaking or deviations from the diagonal line towards the right and hence we should consider our residuals non-normally distributed.

Detecting multicollinearity.

Variance Inflation Factor.

```
##            Monthly.Premium.Auto         I(Coverage == "Premium")
##                        2.107712                         1.174011
##              Total.Claim.Amount                           Income
##                        2.155873                         2.809679
##       I(Number.of.Policies == 1)       I(Number.of.Policies == 2)
##                        1.232306                         1.231919
## I(EmploymentStatus == "Employed")  I(Policy.Type == "Special Auto")
##                        2.763738                         1.001672
## I(Number.of.Open.Complaints == 3) I(Number.of.Open.Complaints == 4)
##                        1.001090                         1.001641
## I(Number.of.Open.Complaints == 5)  I(Vehicle.Class == "Sports Car")
##                        1.001868                         1.046825
```



Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.If there is high correlation between two independed variables (high multicollinearity), then you will not be able to seperate out the impact of individual independed variable on depended variable.Instead of inspecting the correlation matrix, a better way to assess multi- collinearity is to compute the variance inflation factor (VIF). The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. Typically in practice there is a small amount of collinearity among the predictors. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

Due to multicolinearity we can’t define the complete impact of only one independed variable on the depended variable.

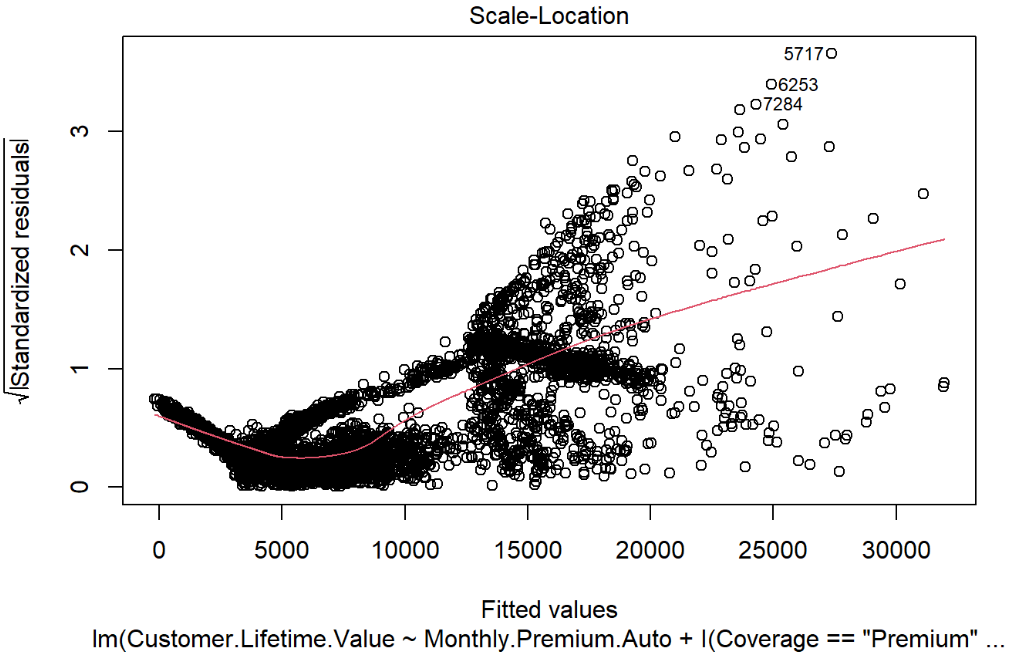
Detecting Homoscedasticity.

Breusch-Pagan test.

Null Hypothesis - Homoscedasticity is present in Residuals.

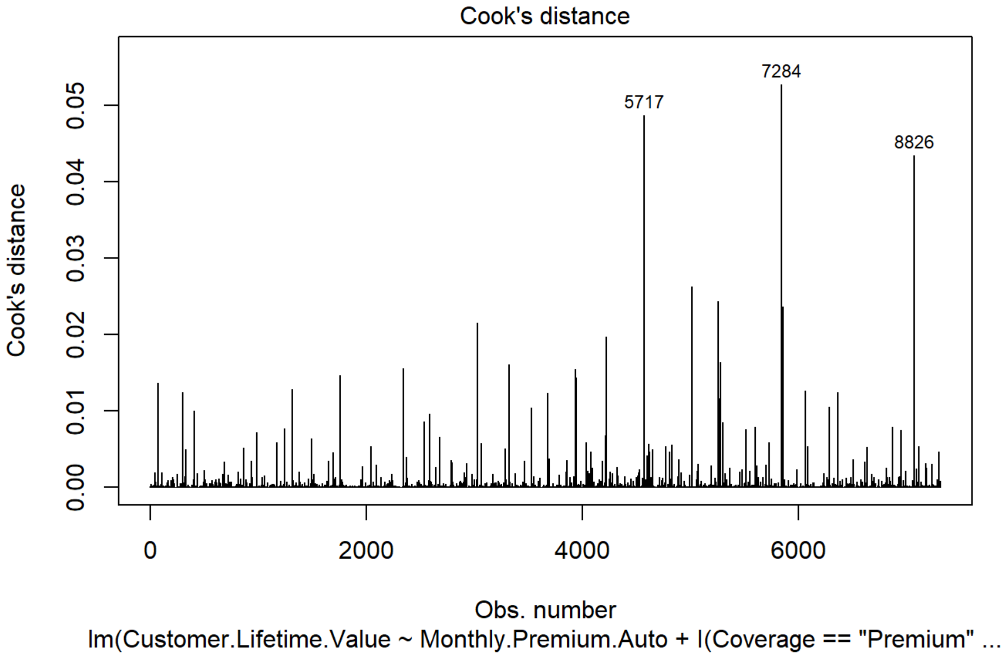
Alternate hypothesis - Heteroskedasticity is present in residuals.

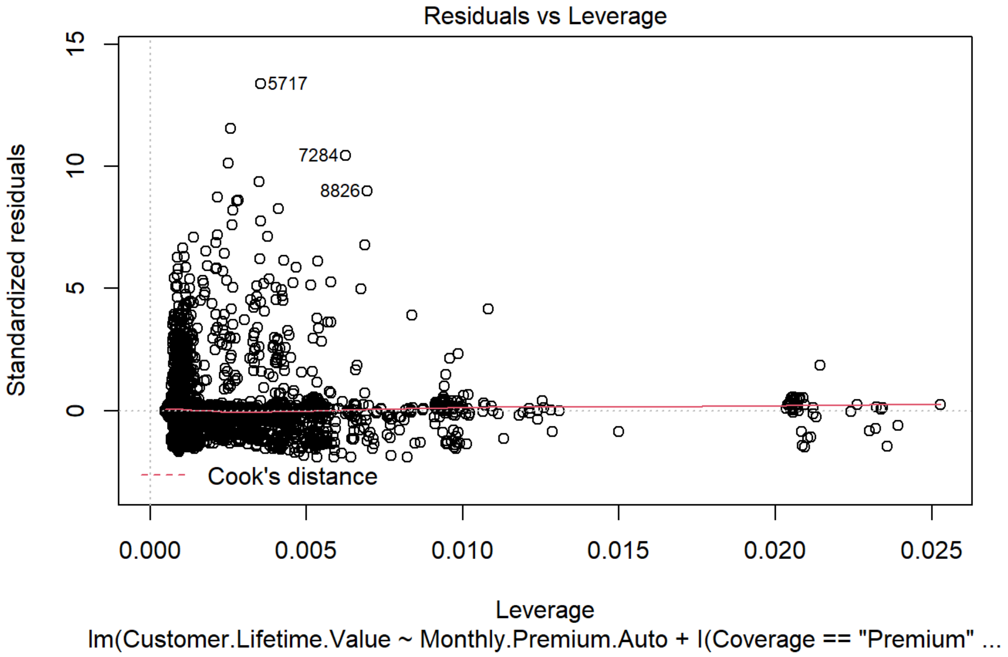
```
## 
##  studentized Breusch-Pagan test
## 
## data:  Reg_4
## BP = 940.67, df = 12, p-value < 2.2e-16
```



p-value < 0.05, so it rejects that errors are homoscedasticity. So errors terms are heteroscedastic ie, they doesn’t have constant variance which is not good for model. The Scale - Location plot also suggests that there is heteroscedasticity in our residuals as it has a funnel shaped pattern and are not randomly distributed.

Model - Plots.





These are the Cook’s Distance and residuals versus leverage plot. These plot helps us to find influential cases (i.e., subjects) if any. We can see that the data points 5717, 7284 and 8826 are influential cases against our regression line.The regression results can be altered if we exclude these cases.

Suggestions for Additional data

- The data could provide more insights if it had variables like ‘expenses caused to the company by the customer’ , ‘initial cost of acquiring a customer’ , ‘mode of contacting the customer’ ’ ‘no:of purchases made by each customer’ etc.
- The data provided would be more efficient if the time period given was extended(more than 2 months).

- Churn would be another necessary feature which would help in predicting if a customer would leave the company. If yes, what would be the offers given to retain the customer.
- Credit based Insurance score would be efficient in providing information about the customer and how much of monetary value he can provide to the company.

Business Solutions.

We would like recommend the following business solutions according to our analysis above.

- Insurance company should target Married customers that have premium policy type, as their employment will be “employed”.
- The company must give preference to “agent” based and “branch” based sales.
- The company should encourage the customers to take policy coverage with Extended and premium coverage and also try to retain such customers. As per our analysis web and call center based marketing should be strengthened.
- Higher number of open complaints can decrease the value of CLV.
- Other modes of marketing can be implemented to increase the profit of the company. Call based marketing strategies show lesser customer involvement.
- The company must also target customers who are willing to invest on policy type “personal”, more deeply the personal L3 policy customers have been more valuable to the company.
- The company must retain customers from Suburban and Urban areas. And also try to allure more customers from rural areas.
- They must also target customers who own Luxury SUVs or sports cars as these customers would be more valuable to the company. And they are also likely to take premium policies.
- Incentivizing long-term customer loyalty would be a good option to retain customers.

References.

- Xgboost (https://analyticsindiamag.com/complete-guide-to-xgboost-with-implementation-in-r/)
- Book (https://github.com/PacktPublishing/Hands-On-Data-Science-for-Marketing)
- Feature Selection (http://r-statistics.co/Variable-Selection-and-Importance-With-R.html)
- DataCamp (https://www.datacamp.com/community/tutorials/introduction-customer-segmentation-python)
- Modelling CLV (https://www.custora.com/blog/story/how-bayesian-probability-models-can-make-clv-predictions-12x-more-accurate)
- Kaggle (https://www.kaggle.com/pankajjsh06/ibm-watson-marketing-customer-value-data)
- Stepwise Modelling (http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/)
- Markdown (https://rmarkdown.rstudio.com/)
- Statistical Tests (https://michaelminn.net/tutorials/r-categorical/index.html)
- Customer Lifetime Value (https://en.wikipedia.org/wiki/Customer_lifetime_value)
- Lasso Regression (https://www.pluralsight.com/guides/linear-lasso-and-ridge-regression-with-r)

Team 1

Team Members:

1. Abraham G K (20BDA20)
2. Josmi Agnes Jose(20BDA27)
3. Sanjana Ramesh(20BDA34)
4. Nidhi Teresa George(20BDA35)
5. Rakshith Kumar K.N(20BDA47)