CUSTOMER LIFETIME VALUE

Code ▾

PROBLEM STATEMENT

Customer Lifetime Value (CLV) is a measure of a customer's total worth to a business over the entire period of the customer-business relationship. The main objective of this project is to predict the lifetime valuation of a customer to facilitate target marketing. Not all customers are equal. Indeed, someone who purchases an inexpensive policy is going to be less valuable to your business than someone who purchases an expensive one, and your longtime customers will bring in more money than those who buy a one-year policy and do not renew. Here we need to predict the customer lifetime value for each customer to make sure how much benefit each customer can repay to the company in exchange for the benefits he/she receives. CLV is an important figure to know as it helps a company to make decisions about how much money to invest in acquiring new customers and retaining existing ones.

With the given information regarding the customers, we can predict which bidding strategies will yield the highest lifetime revenues for the least amount of money through data analysis and exploration and predict the CLV of a given customer.

Using Watson Analytics data, we can predict customer behavior to retain customers. We can analyze all relevant customer data and develop focused customer retention programs.The question that we are trying to solve is to discover what affects customer engagement and to provide actionable recommendations for the business*Business strategy should be to* Acquire more customers + to retain more customers = To increase customer profitability.

In this project, We have tried to find the effect of different variables on the target variable through visualizations, statistical tests and statistical models and compared the results.

Loading Libraries

Required packages like ggplot2,dplyr,plotly,lubridate,modelr,Metrics and few more were loaded.

Loading data

Data was loaded in csv format and we tried to get a basic idea of data.

This data has 9134 Observations of 24 different variables.

Here, the dependent Variable is Customer Lifetime Value.

Continuous Independent Variables include CustomerLifetimeValue, Income,MonthlyPremiumAuto, MonthsSinceLastClaim, MonthsSincePolicyInception, NumberofOpenComplaints, NumberofPolicies and TotalClaimAmount

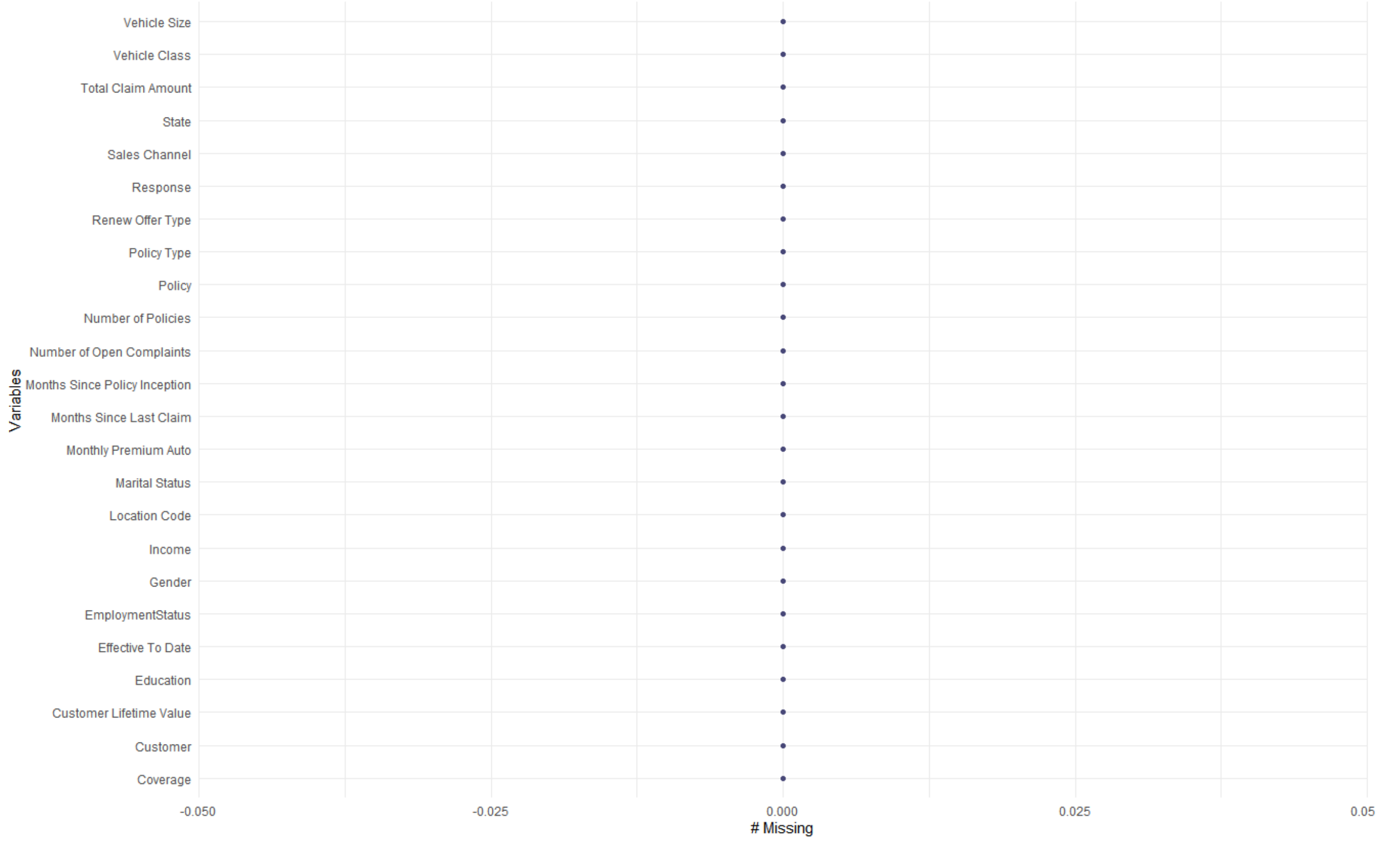
Discrete Independent Variables includes NumberofOpenComplaints, NumberofPolicies.

Categorical Independent Variables include State, Response, Coverage, Education, EmploymentStatus, Gender, LocationCode, MaritalStatus, NumberofOpenComplaints, NumberofPolicies,PolicyType, Policy, RenewOfferType, SalesChannel, VehicleClass and VehicleSize

Other functions like describe,dim,sapply were also used for better understanding of data.

Pre-processing data

Missing value analysis



We have no missing or duplicate values in data so we need not do missing value treatment.

Outlier Analysis

Hide

```
profiling_num(data)
```

variable <chr>	mean <dbl>	std_dev <dbl>	variation_coef <dbl>	p_0 <db
Customer Lifetime Value	8004.940475	6.870968e+03	0.8583409	2230.4337
Income	37657.380009	3.037990e+04	0.8067450	0.0000
Monthly Premium Auto	93.219291	3.440797e+01	0.3691078	61.0000
Months Since Last Claim	15.097000	1.007326e+01	0.6672356	0.0000
Months Since Policy Inception	48.064594	2.790599e+01	0.5805935	1.0000
Number of Open Complaints	0.384388	9.103835e-01	2.3683974	0.0000
Number of Policies	2.966170	2.390182e+00	0.8058141	1.0000
Total Claim Amount	434.088794	2.905001e+02	0.6692181	10.4028

8 rows | 1-6 of 16 columns

Skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails.These two statistics gave us insights into the shape of distribution. High kurtosis in a data set is an indicator that data has heavy outliers. Low kurtosis in a data set is an indicator that data has lack of outliers. So, high value of kurtosis displays presence of outliers for Customer Lifetime Value,Total Claim Amount,Number of Open Complaints, Monthly Premium Auto columns, where kurtosis value is more than 3.

Percentiles are observed to see change in continuous variables. Here, we observe that there is no sudden jump in the values, which means there are anomaly. And outliers of CLV actually represent important customers,and doing outlier treatment directly would mean we are losing good clients for the company.

Data Transformation

Hide

```
data$`Effective To Date`<-mdy(data$`Effective To Date`)

# Converting character variables into Factor variables

data$State <- as.factor(data$State)
data$Response <- as.factor(data$Response)
data$Coverage <- as.factor(data$Coverage)
data$Education <- as.factor(data$Education)
data$EmploymentStatus <- as.factor(data$EmploymentStatus)
data$Gender <- as.factor(data$Gender)
data$`Location Code`  <- as.factor(data$`Location Code` )
data$`Marital Status`  <- as.factor(data$`Marital Status`)
data$`Policy Type` <- as.factor(data$`Policy Type`)
data$`Renew Offer Type`<- as.factor(data$`Renew Offer Type`)
data$Policy  <- as.factor(data$Policy)
data$`Sales Channel`  <- as.factor(data$`Sales Channel`)
data$`Vehicle Class`  <- as.factor(data$`Vehicle Class`)
data$`Vehicle Size`  <- as.factor(data$`Vehicle Size`)


# Converting two numerical variables as factors
data$`Number of Open Complaints` <- as.factor(data$`Number of Open Complaints`)
data$`Number of Policies`<- as.factor(data$`Number of Policies`)
```

We converted all the categorical variables and 2 discrete numeric columns(Number of policies and number of open complaints) into factors to give each category a level. Since the format of date column was not uniform we standardized it.

Exploratory Data Analysis

To identify quantity and percentage of zeros

Hide

```
status(data)
```

	variable <chr>	q_zeros <int>	p_zeros <dbl>
Customer	Customer	0	0.000000000
State	State	0	0.000000000
Customer Lifetime Value	Customer Lifetime Value	0	0.000000000
Response	Response	0	0.000000000

	variable <chr>	q_zeros <int>	p_zeros <dbl>
Coverage	Coverage	0	0.000000000
Education	Education	0	0.000000000
Effective To Date	Effective To Date	0	0.000000000
EmploymentStatus	EmploymentStatus	0	0.000000000
Gender	Gender	0	0.000000000
Income	Income	2317	0.253667616

1-10 of 24 rows | 1-7 of 9 columns Previous **1** 2 3 Next

From here, we can observe that there are lot of people whose income is 0 since they are unemployed.

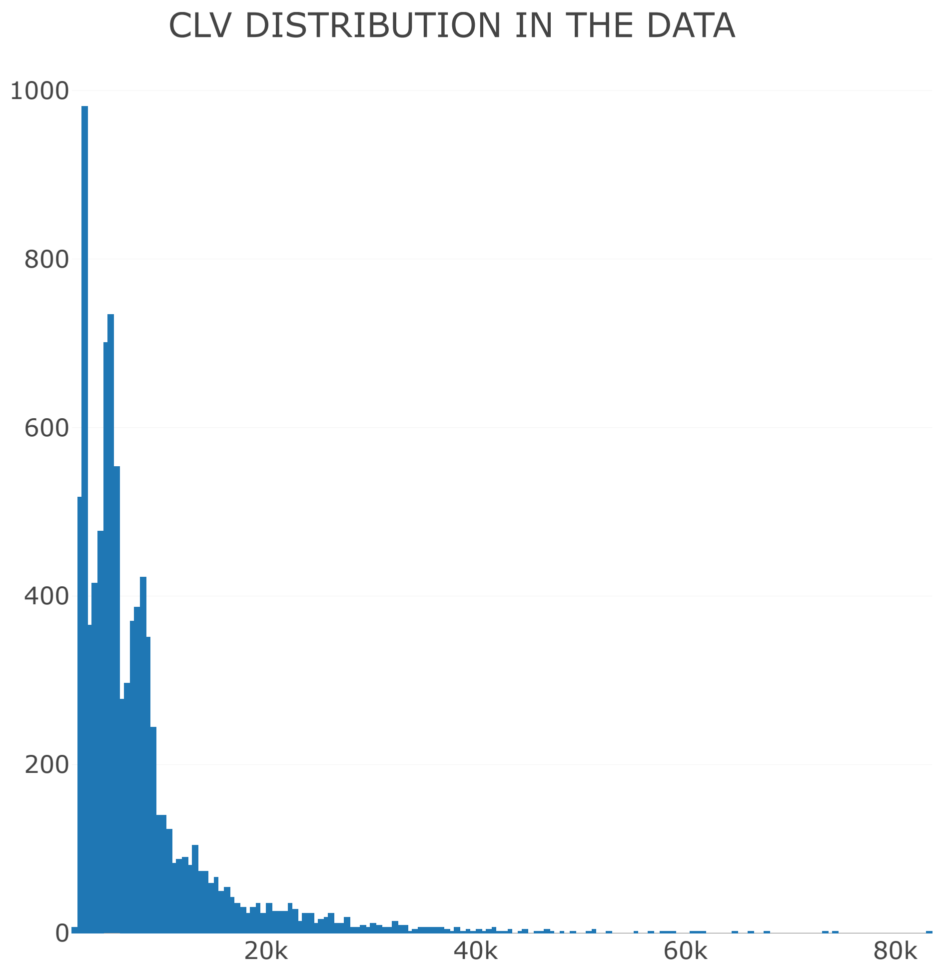
Our objective here is to visualize the given data and look for variables that can be important for modelling.

Customer Lifetime Value

This is our target variable.

Hide

```
fig_CLV <- plot_ly(x =data$`Customer Lifetime Value`, type = "histogram")%>% layout(title ="
 CLV DISTRIBUTION IN THE DATA")
fig_CLV
```



Hide

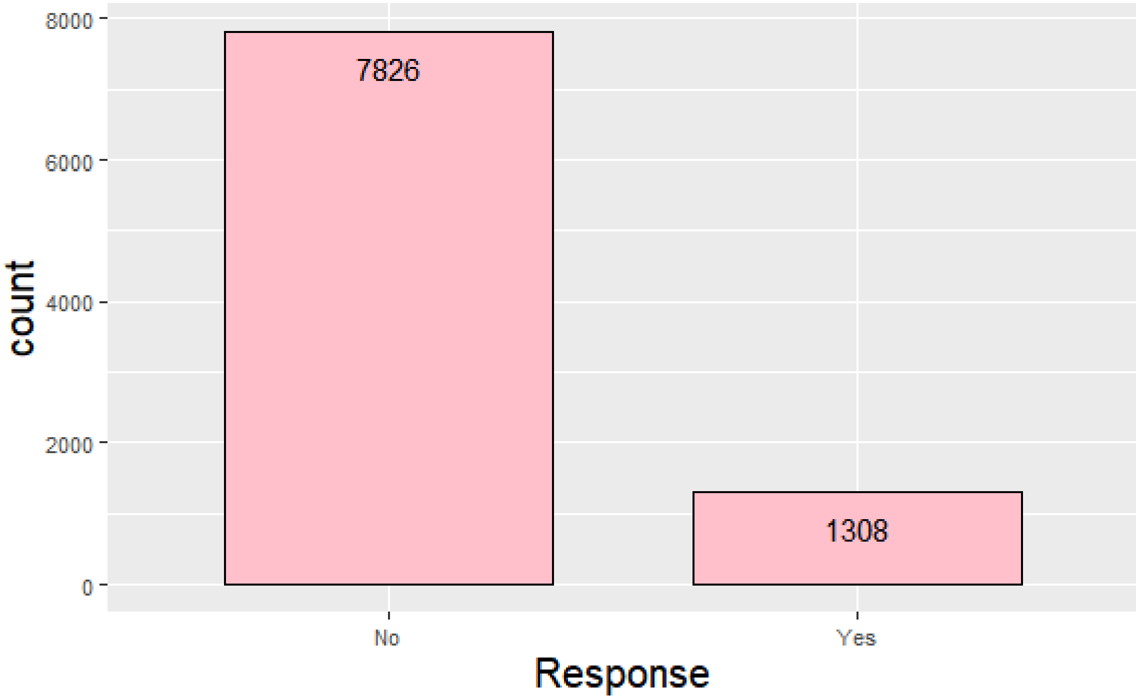
```
NA
NA
```

Customer lifetime value is positively skewed.

Response

Hide

```
ggplot(data,aes(Response))+geom_bar(fill="pink",col="black",width=0.7,position=position_dodge
(0.9))+
    geom_text(stat="count",aes(label = after_stat(count)),vjust=2)+
    theme(
        text=element_text(size=10),
        axis.title.x = element_text(color="black", size=15),
        axis.title.y = element_text(color="black", size=15)
    )
```

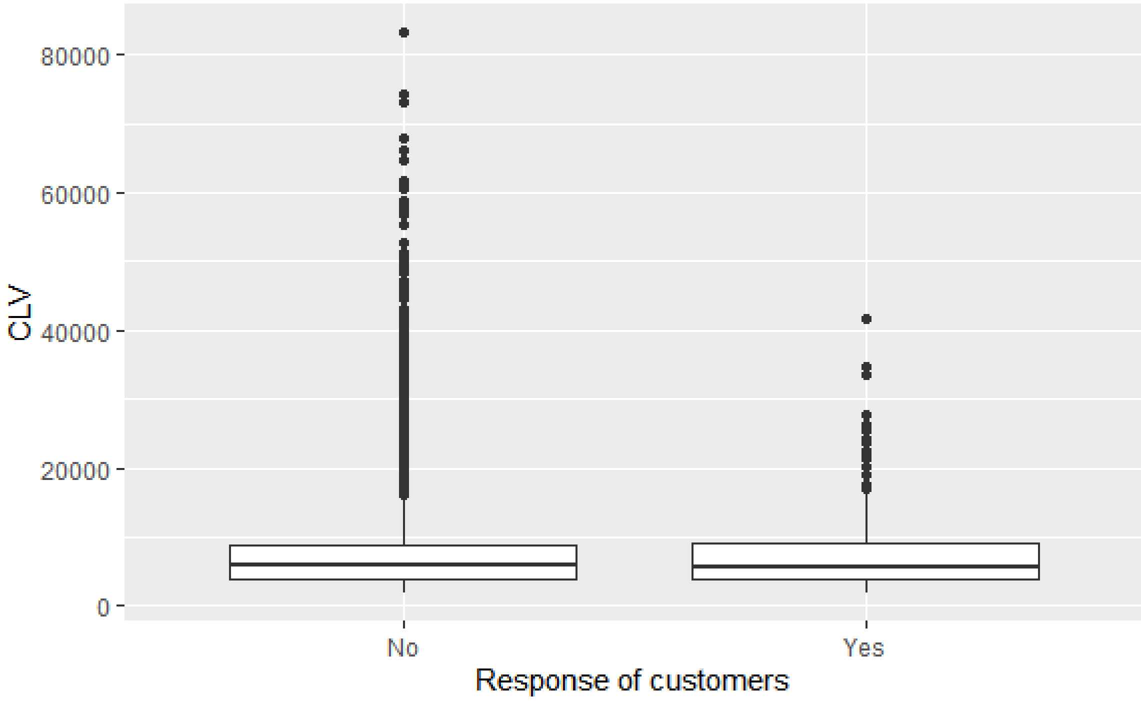


Hide

```
NA
NA
```

Hide

```
ggplot(data, aes(x =Response, y = `Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Response of customers") + ylab("CLV")
```

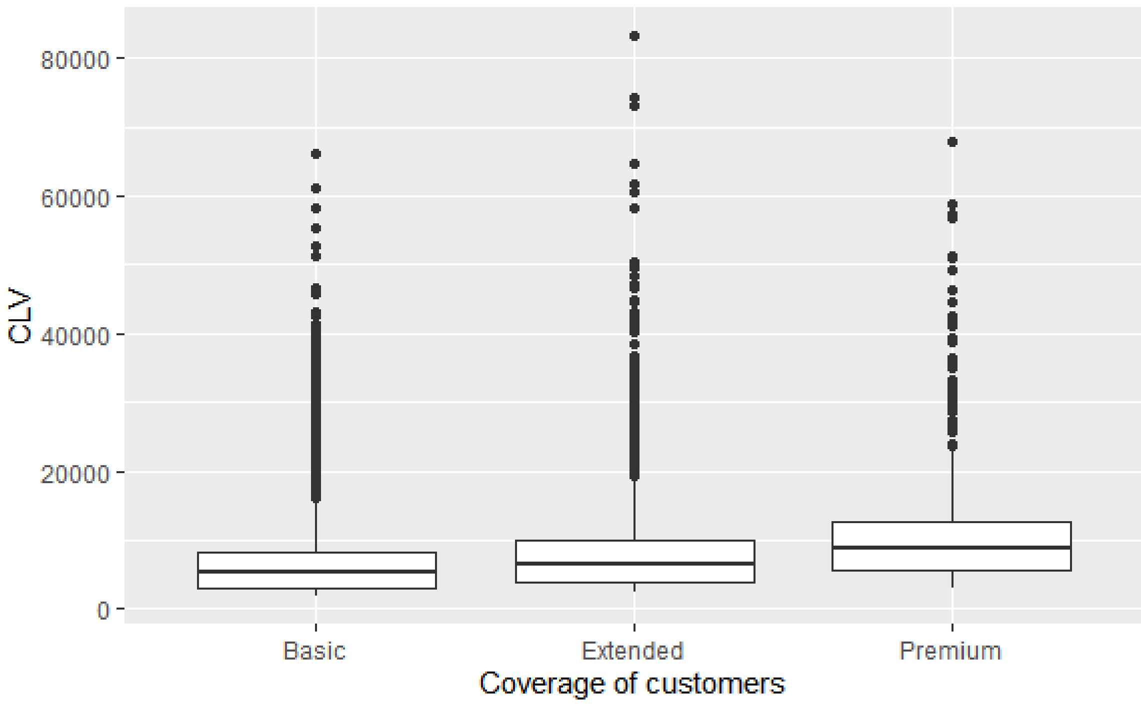


Here, we can observe that very few people have reapplied for policy. So, if our focus is to retain customers we should target customers who said NO as those customers who have said no are of high customer lifetime value.

Coverage

Hide

```
ggplot(data, aes(x =Coverage, y = `Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Coverage of customers") + ylab("CLV")
```



Hide

```
aggregate( `Customer Lifetime Value` ~ Coverage, data, mean)
```

Coverage <fctr>	Customer Lifetime Value <dbl>
Basic	7190.706
Extended	8789.678
Premium	10895.603

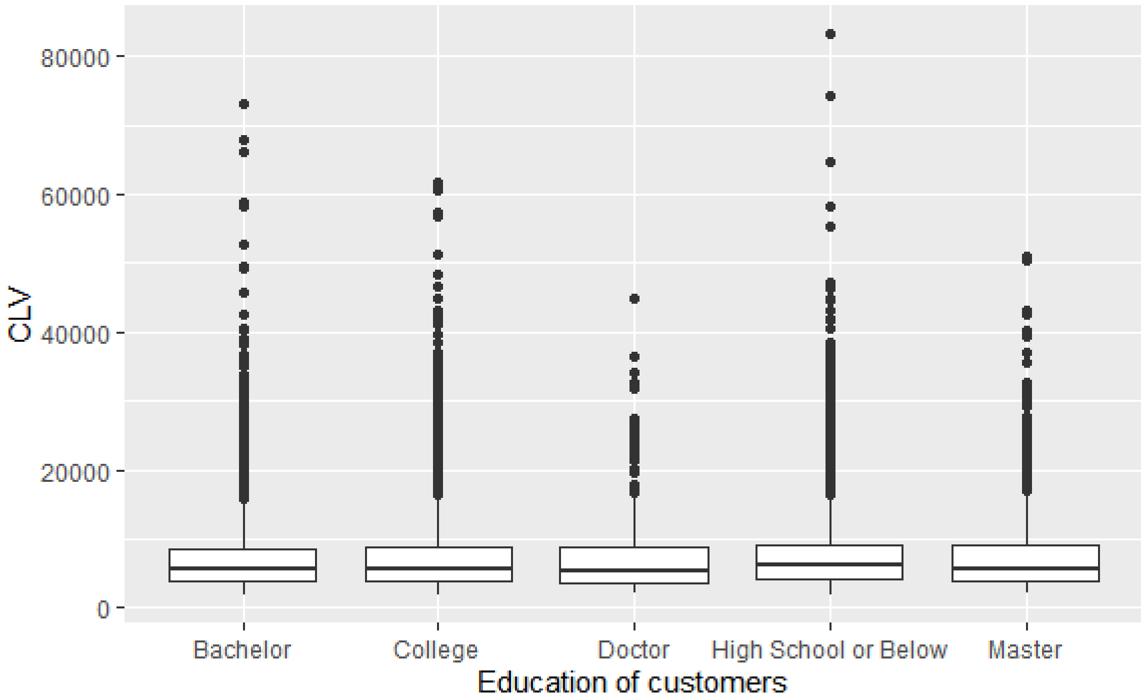
3 rows

This can be an important variable, as different groups show major differences in their value, and which kind of coverage they are choosing may help in deciding their customer life time value.

Education

Hide

```
ggplot(data, aes(x =Education, y = `Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Education of customers") + ylab("CLV")
```

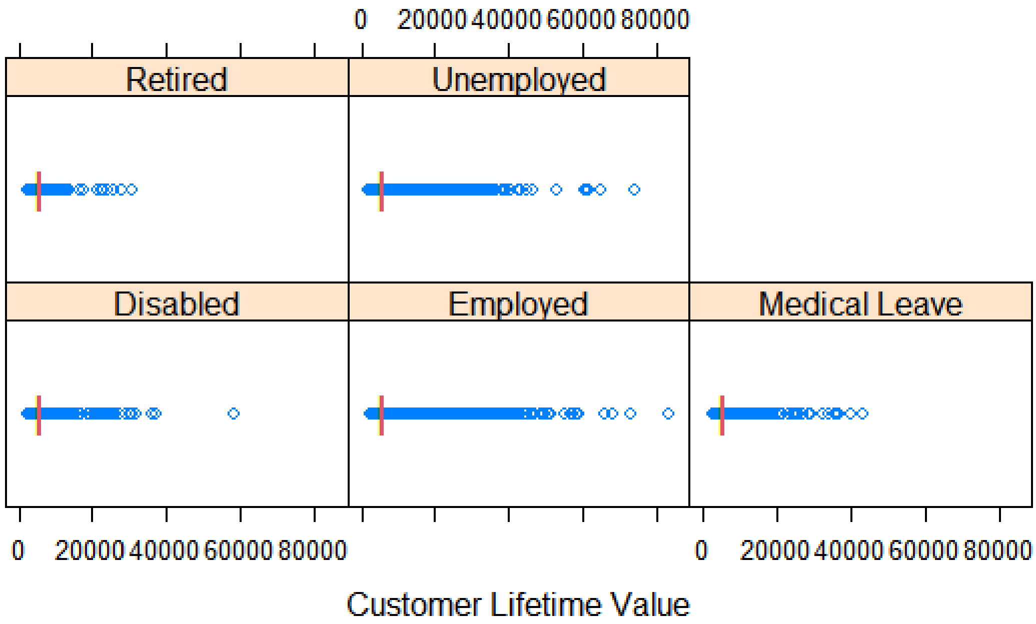


Doctor in general have low customer life time value, while Masters and High School or Below have higher customer life time value. But it’s still tough to observe any major changes across different groups. So, it may be possible that this variable is not important.

Employment Status

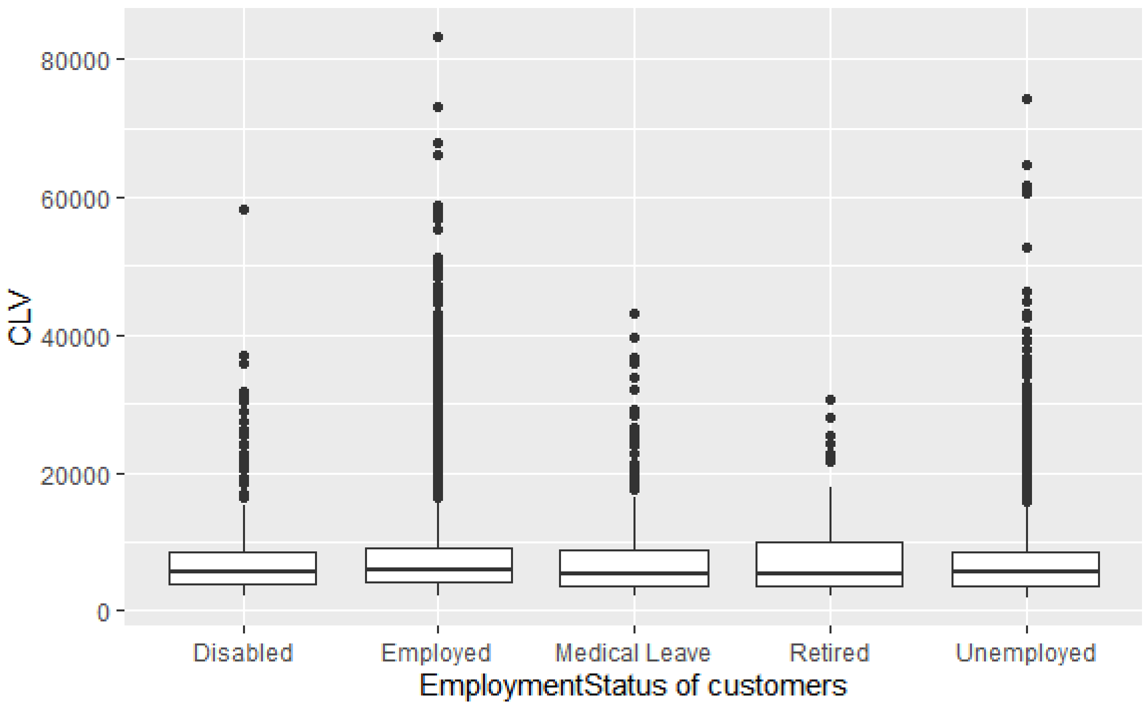
Hide

```
stripplot(~`Customer Lifetime Value`|EmploymentStatus,data,
  panel=function(x,y,...) {
     m=median(x)
     panel.stripplot(x,y,...)
     panel.stripplot(m,y,pch="|",cex=2,col=2)
  }
)
```



Hide

```
ggplot(data, aes(x =EmploymentStatus, y = `Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("EmploymentStatus of customers") + ylab("CLV")
```



Customer with high customer lifetime values lies mainly in Employed and unemployed categories, while it appears that there isn't any major difference across categories. Also, we can say that people who are Retired, on Medical leave, Disabled their customer lifetime value is less. Employed people's customer lifetime value definitely turns out to be highest among all. There is one data point in disabled which can be treated as outlier as it's value is really high, which doesn't make it a general case. So, we remove that data point.

Hide

```
subset(data,EmploymentStatus=='Disabled'& "Customer Lifetime Value">40000)
```

Custo... <chr>	State <fctr>	Customer Lifetime Value <dbl>	Respo... <fctr>	Covera... <fctr>	Education <fctr>
SV62436	Washington	3041.792	No	Extended	Bachelor
HM55802	California	2392.108	No	Basic	Bachelor
HO30839	Washington	5346.917	No	Extended	Master
HG65722	Oregon	12819.103	No	Premium	Doctor
FR46645	California	4293.997	No	Premium	Bachelor
ML29312	Oregon	4499.493	No	Extended	High School or Below
UB61619	Oregon	4059.567	No	Premium	Master
RZ33670	California	11727.776	No	Premium	College
SQ19467	Oregon	6554.216	No	Extended	College
PD27940	Arizona	4885.163	No	Extended	High School or Below

1-10 of 405 rows | 1-6 of 24 columns Previous **1** 2 3 4 5 6 ... 41 Next

Hide

```
NA
```

Hide

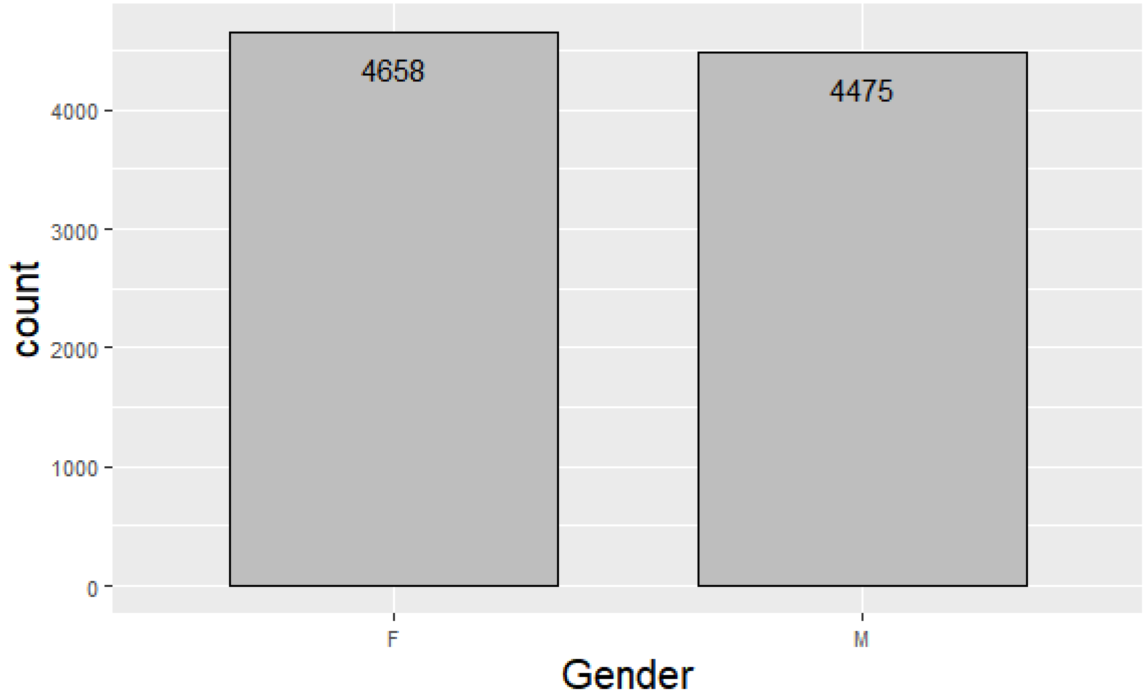
```
data=data[data$Customer !='XF89906',]
```

This row is dropped.

Gender

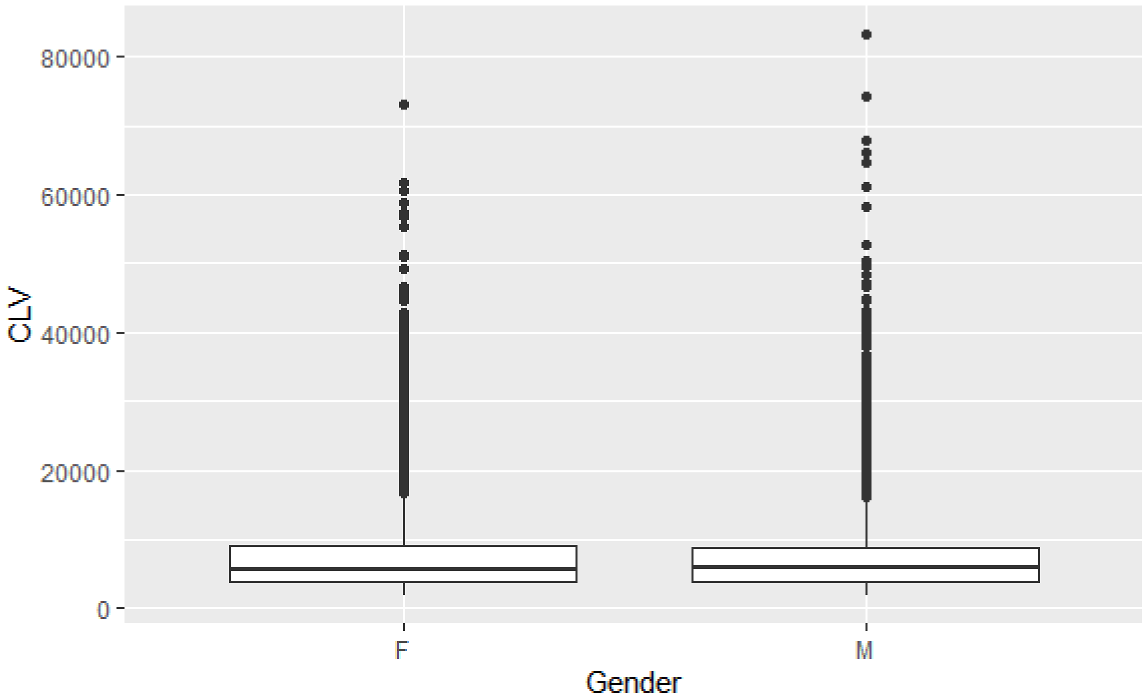
Hide

```
ggplot(data,aes(Gender))+geom_bar(fill="grey",col="black",width=0.7,position=position_dodge(
0.9))+
    geom_text(stat="count",aes(label = after_stat(count)),vjust=2)+
    theme(
        text=element_text(size=10),
        axis.title.x = element_text(color="black", size=15),
        axis.title.y = element_text(color="black", size=15)
    )
```



Hide

```
ggplot(data, aes(x =Gender, y = `Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Gender") + ylab("CLV")
```

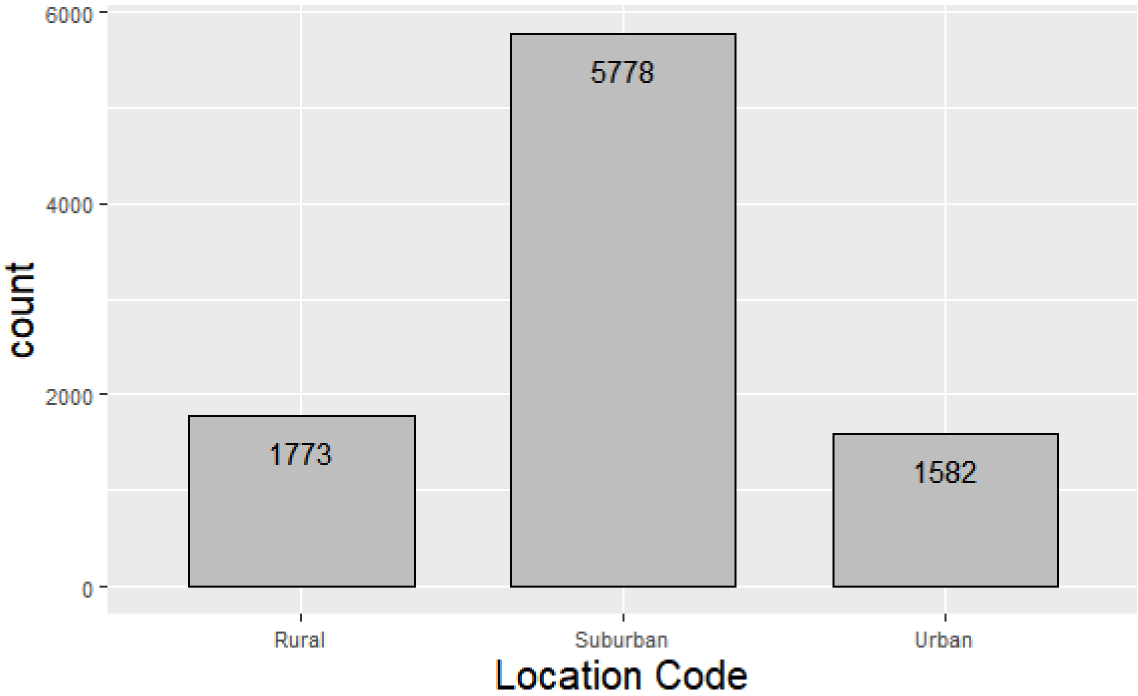


It appears that gender has no effect on customer lifetime value.

Location Code

Hide

```
ggplot(data,aes(`Location Code`))+geom_bar(fill="grey",col="black",width=0.7,position=positio
n_dodge(0.9))+
    geom_text(stat="count",aes(label = after_stat(count)),vjust=2)+
    theme(
        text=element_text(size=10),
        axis.title.x = element_text(color="black", size=15),
        axis.title.y = element_text(color="black", size=15)
    )
```



Hide

```
aggregate( `Customer Lifetime Value` ~ `Location Code`, data, mean)
```

Location Code <fctr>	Customer Lifetime Value <dbl>
Rural	7953.699
Suburban	7995.769
Urban	8064.133

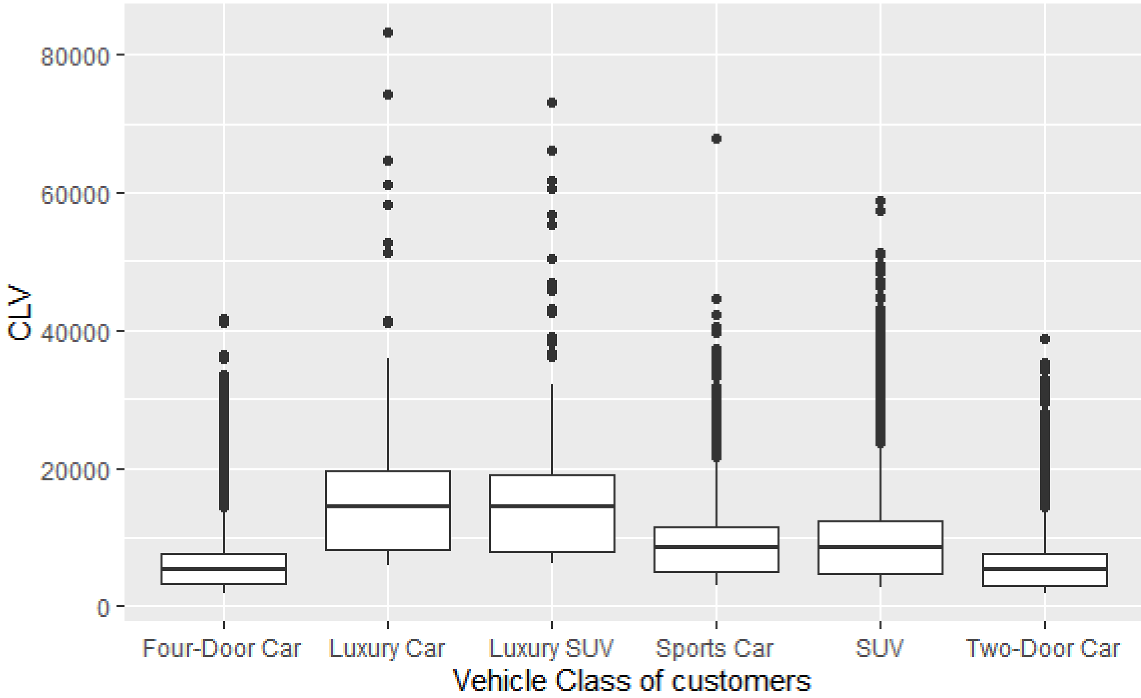
3 rows

We should focus more on suburbans. as they are our major customers We can’t say any major difference between these groups.We get same inference using boxplots as well.

Vehicle Class

Hide

```
ggplot(data, aes(x =`Vehicle Class`, y = `Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Vehicle Class of customers") + ylab("CLV")
```



Hide

```
aggregate( `Customer Lifetime Value` ~ `Vehicle Class`, data, mean)
```

Vehicle Class <fctr>	Customer Lifetime Value <dbl>
Four-Door Car	6631.727
Luxury Car	17053.348
Luxury SUV	16898.496
Sports Car	10750.989
SUV	10443.512
Two-Door Car	6671.031

6 rows

Hide

```
head(data)
```

Custo... <chr>	State <fctr>	Customer Lifetime Value <dbl>	Respo... <fctr>	Covera... <fctr>	Education <fctr>	Effectiv
BU79786	Washington	2763.519	No	Basic	Bachelor	2
QZ44356	Arizona	6979.536	No	Extended	Bachelor	2
AI49188	Nevada	12887.432	No	Premium	Bachelor	2
WW63253	California	7645.862	No	Basic	Bachelor	2
HB64268	Washington	2813.693	No	Basic	Bachelor	2
OC83172	Oregon	8256.298	Yes	Basic	Bachelor	2

6 rows | 1-7 of 24 columns

Luxury Car,Luxury SUV have high customer lifetime value, while Four-Door Car and Two-Door Car have less customer lifetime values. So, this variable may be important for prediction as it varies across different categories.

Sales Channel

Hide

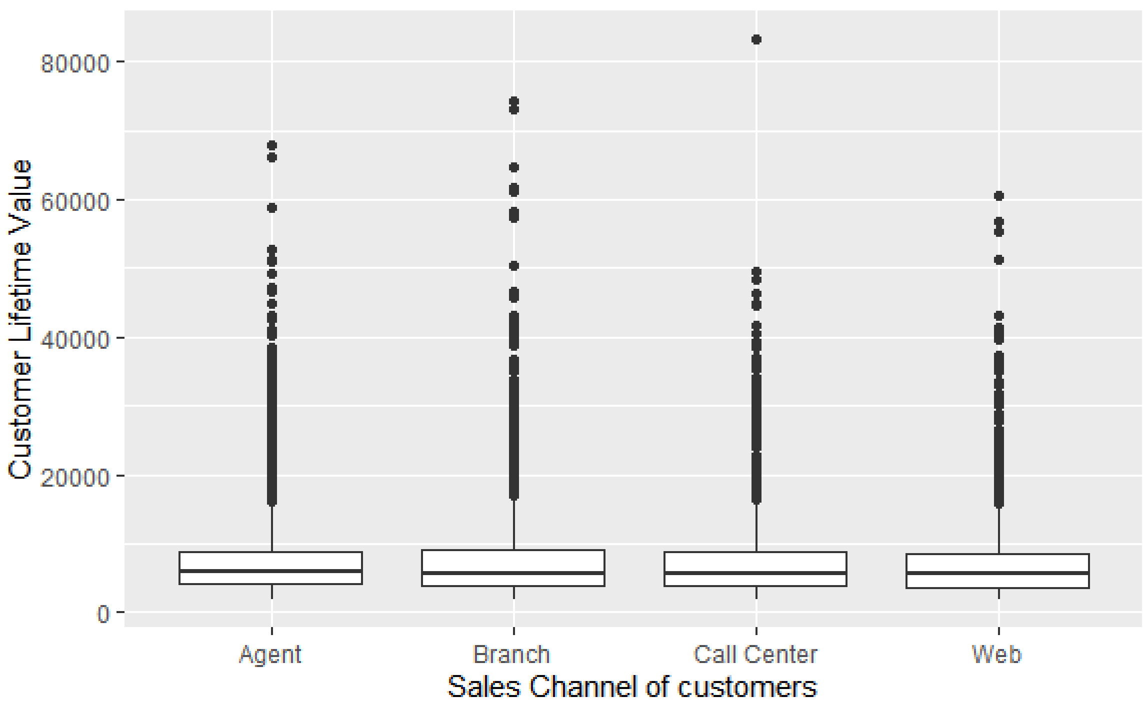
```
head(data)
```

Custo… <chr>	State <fctr>	Customer Lifetime Value <dbl>	Respo… <fctr>	Covera… <fctr>	Education <fctr>	Effectiv
BU79786	Washington	2763.519	No	Basic	Bachelor	2
QZ44356	Arizona	6979.536	No	Extended	Bachelor	2
AI49188	Nevada	12887.432	No	Premium	Bachelor	2
WW63253	California	7645.862	No	Basic	Bachelor	2
HB64268	Washington	2813.693	No	Basic	Bachelor	2
OC83172	Oregon	8256.298	Yes	Basic	Bachelor	2

6 rows | 1-7 of 24 columns

Hide

```
ggplot(data, aes(x =`Sales Channel`, y = `Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Sales Channel of customers")
```



Not much difference is observed across different categories, but we can see presence of an outlier. So, we will treat it.

Hide

```
head(data)
```

Custo… <chr>	State <fctr>	Customer Lifetime Value <dbl>	Respo… <fctr>	Covera… <fctr>	Education <fctr>	Effectiv
BU79786	Washington	2763.519	No	Basic	Bachelor	2
QZ44356	Arizona	6979.536	No	Extended	Bachelor	2
AI49188	Nevada	12887.432	No	Premium	Bachelor	2
WW63253	California	7645.862	No	Basic	Bachelor	2
HB64268	Washington	2813.693	No	Basic	Bachelor	2
OC83172	Oregon	8256.298	Yes	Basic	Bachelor	2

6 rows | 1-7 of 24 columns

Hide

```
subset(data,`Sales Channel`=='Call Center' & 'Customer Lifetime Value'>50000)
```

Custo… <chr>	State <fctr>	Customer Lifetime Value <dbl>	Respo… <fctr>	Covera… <fctr>	Education <fctr>
WW63253	California	7645.862	No	Basic	Bachelor
IL66569	California	5384.432	No	Basic	College
FV94802	Nevada	2566.868	No	Basic	High School or Below
OE15005	California	3945.242	No	Basic	College
FL50705	California	8162.617	No	Premium	High School or Below
SV62436	Washington	3041.792	No	Extended	Bachelor
FS42516	Oregon	5802.066	No	Basic	College
GE62437	Arizona	12902.560	No	Premium	College
SV85652	Arizona	2454.584	No	Basic	College
PF41800	California	4715.321	No	Basic	Bachelor

1-10 of 1,765 rows | 1-6 of 24 columns Previous 1 2 3 4 5 6 ... 100 Next

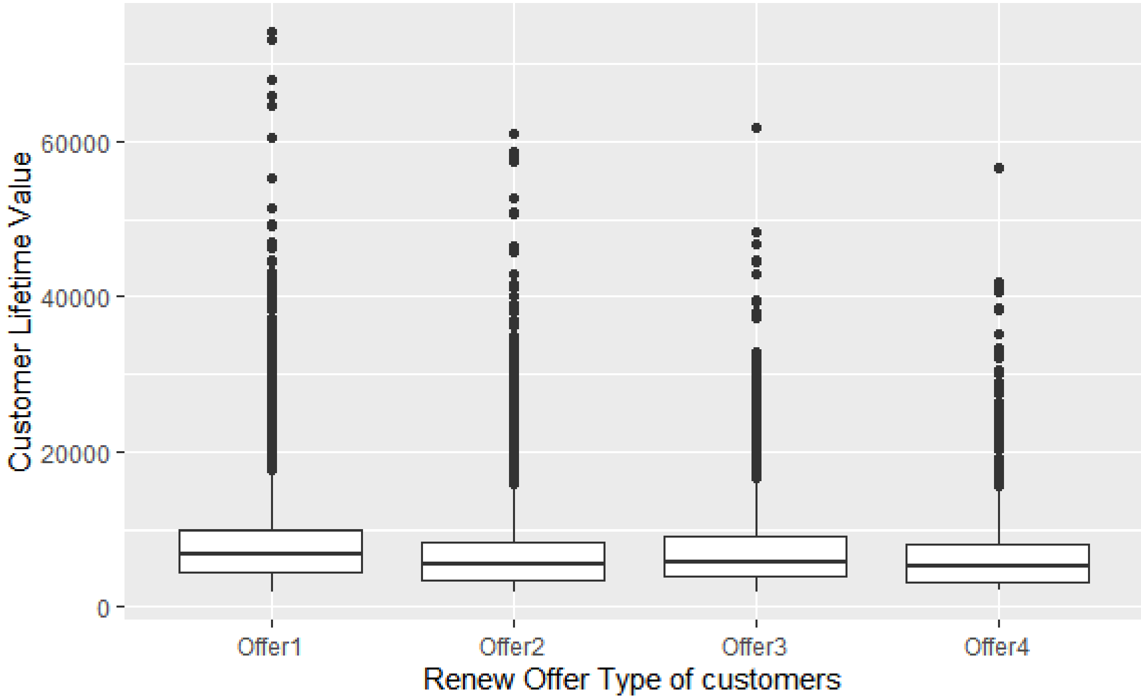
Hide

```
data=data[data$Customer !='FQ61281',]
```

Renew Offer Type

Hide

```
ggplot(data, aes(x =`Renew Offer Type`, y =`Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Renew Offer Type of customers")
```



Hide

```
aggregate( `Customer Lifetime Value` ~ `Renew Offer Type`, data, mean)
```

Renew Offer Type <fctr>	Customer Lifetime Value <dbl>
Offer1	8673.987
Offer2	7396.754
Offer3	7997.887
Offer4	7179.947

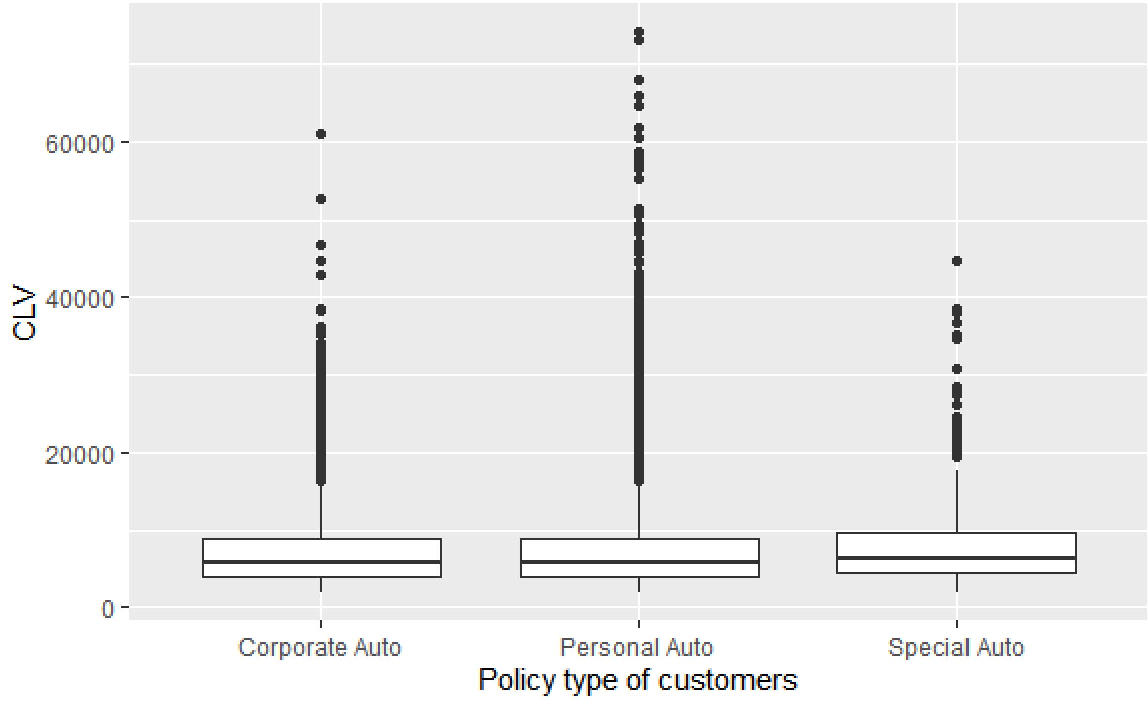
4 rows

Offer 1 and Offer 3 seems to represent little higher valued customers. But we can't say if difference across categories is significant enough.

Policy Type

Hide

```
ggplot(data, aes(x =`Policy Type`,y =`Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Policy type of customers") + ylab("CLV")
```



Hide

```
aggregate( `Customer Lifetime Value` ~ `Policy Type`, data, mean)
```

Policy Type <fctr>	Customer Lifetime Value <dbl>
Corporate Auto	7814.410
Personal Auto	8008.873
Special Auto	8594.245

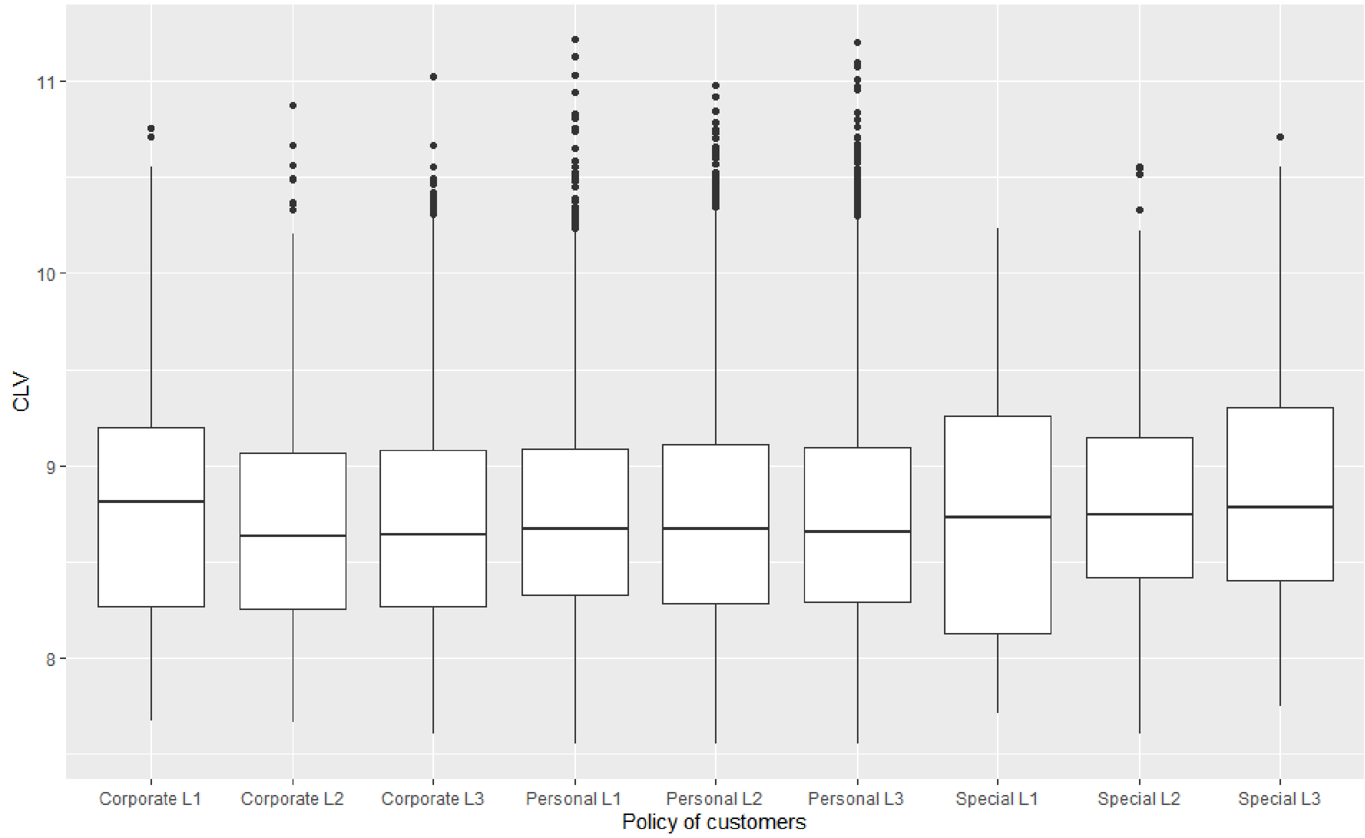
3 rows

Special Auto range of customer life time values is in general higher than others, while Personal Auto policy type is the one that is generally opted by customers.

Policy

Hide

```
ggplot(data, aes(x =Policy, y = `Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Policy of customers") + ylab("CLV")
```



Hide

```
aggregate( `Customer Lifetime Value` ~ Policy, data, mean)
```

Policy <fctr>	Customer Lifetime Value <dbl>
Corporate L1	8474.928
Corporate L2	7597.695
Corporate L3	7707.722
Personal L1	7989.762
Personal L2	8054.909
Personal L3	7987.263
Special L1	8332.763
Special L2	8326.906
Special L3	9007.092

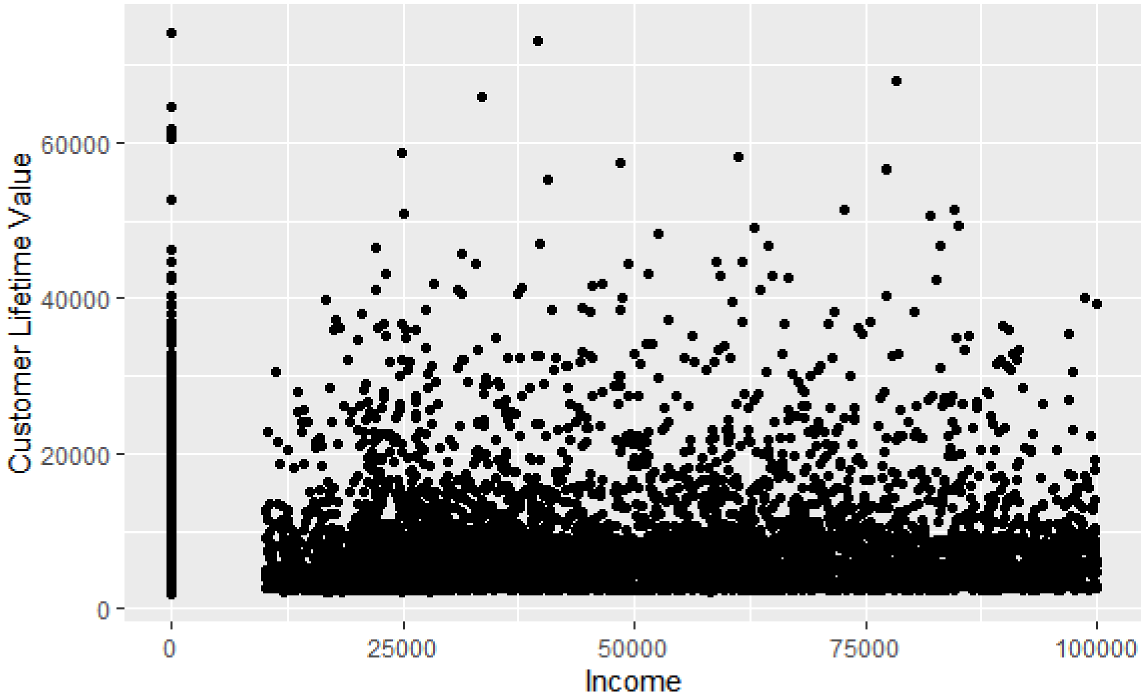
9 rows

Difference across different categories seems to be less, but may be significant across some categories. This variable seems to be highly correlated to Renew Offer Type, so we need to drop one of these variables.

Continuous variables

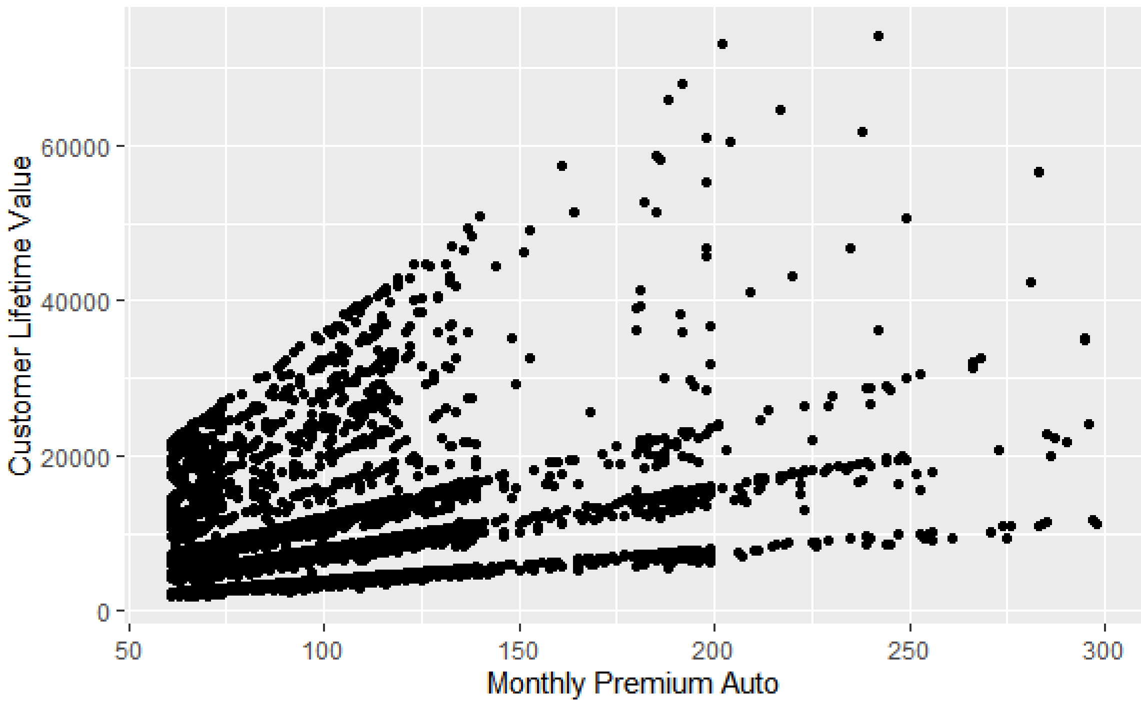
Hide

```
ggplot(data, aes(x = Income, y = `Customer Lifetime Value`)) + geom_point()
```



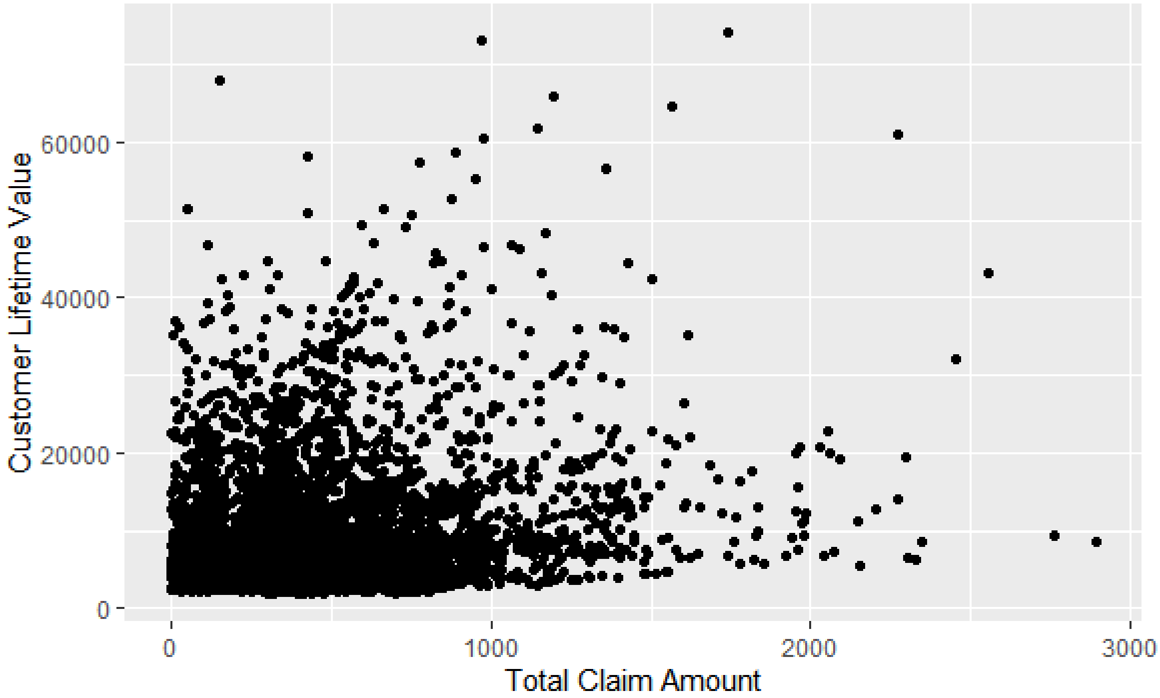
Hide

```
ggplot(data, aes(x = `Monthly Premium Auto`, y = `Customer Lifetime Value`)) + geom_point()
```



Hide

```
ggplot(data, aes(x =`Total Claim Amount`, y = `Customer Lifetime Value`)) + geom_point()
```

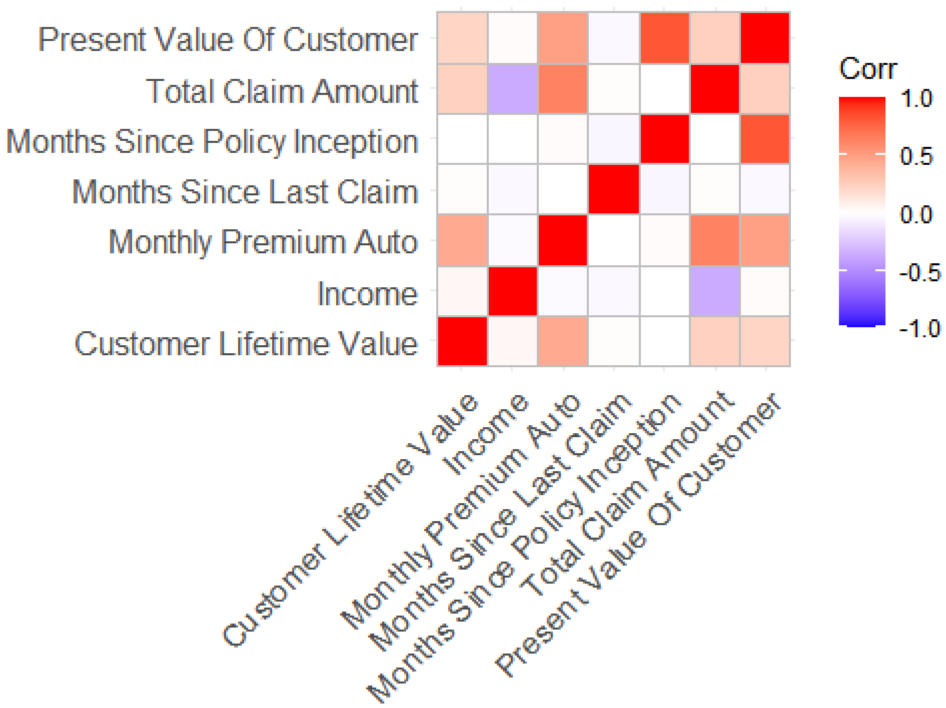


Months Since Policy Inception, Months Since Last Claim,Income doesn't show much correlation with Customer Lifetime value. While Total Claim Amount, Monthly Premium Auto appears to have some correlation to dependent variable Customer Lifetime value.

Correlation Matrix

Hide

```
library(ggcorrplot)
library("dplyr")
data_num=select_if(data, is.numeric)
corr <- round(cor(data_num), 3)
ggcorrplot(corr)
```



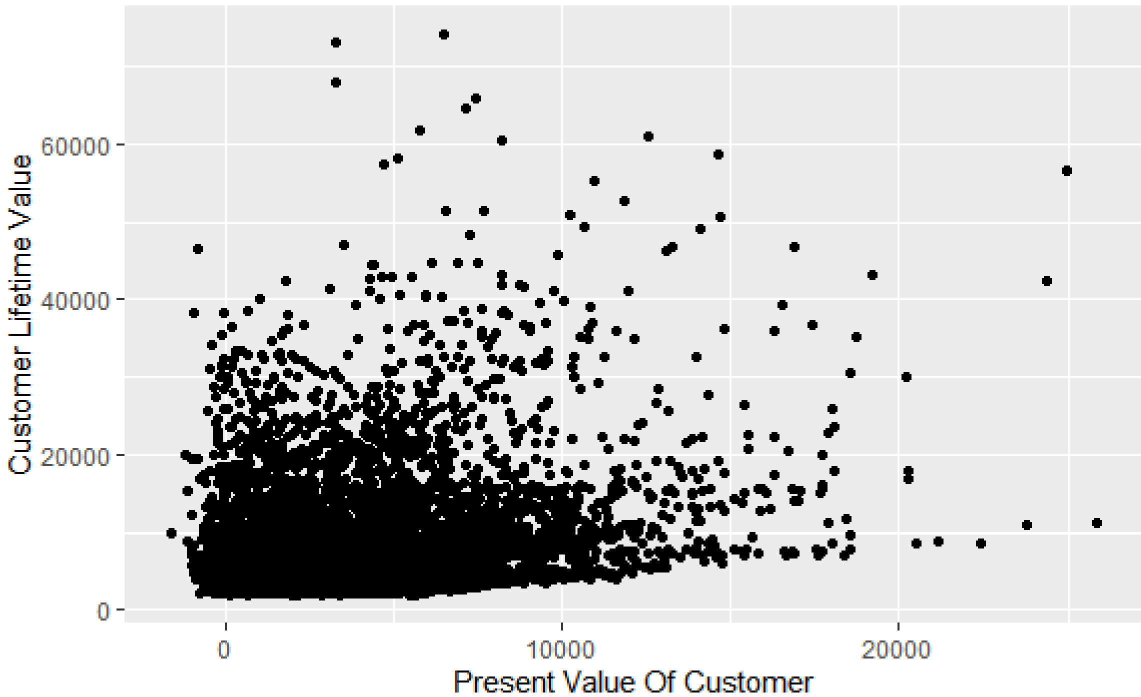
From correlation matrix we can say that Income, Total claim amount , Monthly Premium Auto are correlated to each other. So, we will drop 2 variables out of these 3 to prevent multicollinearity.

Also, we can see that dependent variable Customer lifetime value has high correlation to Monthly Premium auto, Total claim amount. So, these can be considered as important variables but due to presence of multicollinearity among these two we will most probably use just one variable.

Deriving New Column

Hide

```
data$"Present Value Of Customer"= (data$"Monthly Premium Auto" * data$"Months Since Policy In
ception") - data$"Total Claim Amount"
ggplot(data, aes(x =`Present Value Of Customer`, y = `Customer Lifetime Value`)) + geom_point
()
```



There is high correlation between Customer Lifetime value, and the derived variable Present value of customer. We will try to use this variable for modelling, and check if it helps to improve result.

Binning Columns

Hide

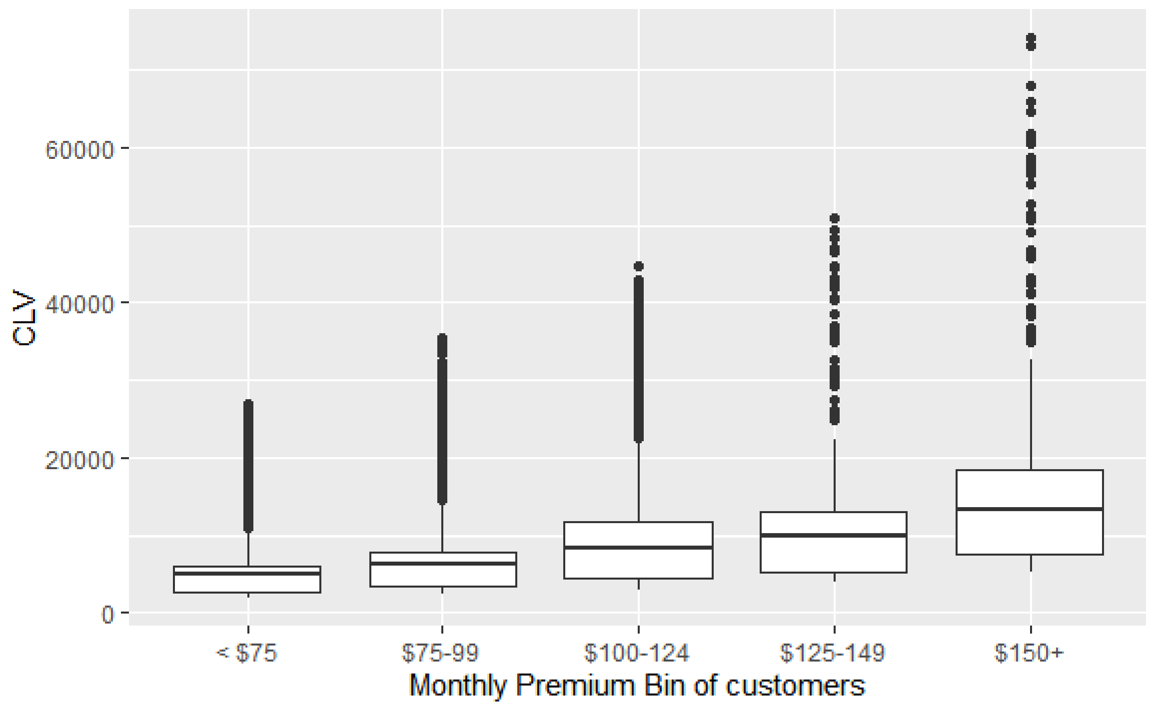
```
# Add Income bins as new column
data$"Income Bin" <- cut(data$"Income",
                         breaks = c(-1,14999,29999,44999,59999,74999,Inf),
                         labels = c("< $15000", "$15000-29999","$30000-44999",
                                    "$45000-59999", "$60000-74999", "$75000+"))

# Add Monthly Premium bins as new column
data$"Monthly Premium Bin" <- cut(data$"Monthly Premium Auto",
                            breaks = c(0,74,99,124,149,Inf),
                            labels = c("< $75", "$75-99","$100-124","$125-149","$150+"))

# Add Total Claim bins as new column
data$"Total Claim Bin" <- cut(data$"Total Claim Amount",
                        breaks = c(0,249,499,749,999,Inf),
                        labels = c("< $250", "$250-499","$500-749","$750-999","$1000+"))
```

Hide

```
ggplot(data, aes(x =`Monthly Premium Bin`, y = `Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Monthly Premium Bin of customers") + ylab("CLV")
```



Hide

```
aggregate( `Customer Lifetime Value` ~ `Monthly Premium Bin`, data, mean)
```

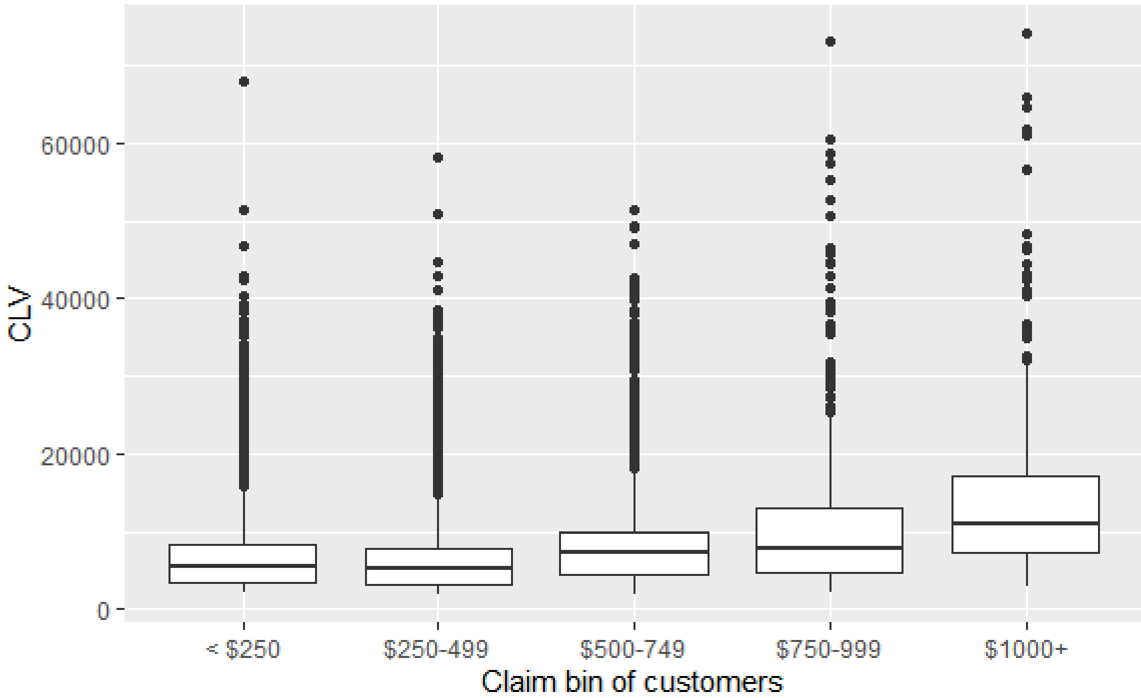
Monthly Premium Bin <fctr>	Customer Lifetime Value <dbl>
< $75	5833.855
$75-99	7444.583
$100-124	9871.735

Monthly Premium Bin <fctr>	Customer Lifetime Value <dbl>
$125-149	10985.471
$150+	15724.567

5 rows

Hide

```
ggplot(data, aes(x =`Total Claim Bin`, y = `Customer Lifetime Value`)) +
  geom_boxplot() +
  xlab("Claim bin of customers") + ylab("CLV")
```



Hide

```
aggregate( `Customer Lifetime Value` ~ `Total Claim Bin`, data, mean)
```

Total Claim Bin <fctr>	Customer Lifetime Value <dbl>
< $250	7358.637
$250-499	6925.566
$500-749	8849.857
$750-999	10596.800
$1000+	13904.052

5 rows

Creating bins for income, doesn’t show any changes across bins. But for Monthly Premium Bin and claim bin we can observe changes across categories.

Is there any preferable policy for very high valued customers?

Hide

```
subset(data,`Customer Lifetime Value`>50000)
```

Custo... <chr>	State <fctr>	Customer Lifetime Value <dbl>	Respo... <fctr>	Covera... <fctr>	Education <fctr>
OM82309	California	58166.55	No	Basic	Bachelor
YC54142	Washington	74228.52	No	Extended	High School or Below
KI58952	California	51337.91	No	Premium	College
EN65835	Arizona	58753.88	No	Premium	Bachelor
CL79250	Nevada	52811.49	No	Basic	Bachelor
AZ84403	Oregon	61850.19	No	Extended	College
JT47995	Arizona	60556.19	No	Extended	College
DU50092	Oregon	56675.94	No	Premium	College
SK66747	Washington	66025.75	No	Basic	Bachelor
BP23267	California	73225.96	No	Extended	Bachelor

1-10 of 18 rows | 1-6 of 28 columns Previous **1** 2 Next

Personal Auto policy type is the preferable policy for very high lifetime value customers.

Are agents more effective with regards to purchase of policy plans?

Hide

```
df_1=subset(data, (Response=="Yes") )
aggregate(Customer ~ `Sales Channel`,df_1, FUN = length)
```

Sales Channel <fctr>	Customer <int>
Agent	666
Branch	294
Call Center	192
Web	156

4 rows

Hide

```
aggregate(Customer ~ `Sales Channel`,data, FUN = length)
```

Sales Channel <fctr>	Customer <int>
Agent	3476
Branch	2567
Call Center	1764
Web	1325

4 rows

0.19%,0.11%,0.11%,0.12% are the respective percentages of people who responded with yes from different sales channel w.r.t total customers each channel brought.This shows that people are more likely to respond to agents rather than other sales channel mode.

Can we use knowledge of gender as leverage in any case?

Hide

```
subset(data, (Response=="Yes")& (`Customer Lifetime Value`>25000))
```

Custo... <chr>	State <fctr>	Customer Lifetime Value <dbl>	Respo... <fctr>	Covera... <fctr>	Education <fctr>
PY51963	California	33473.35	Yes	Basic	Bachelor
BL90769	California	33473.35	Yes	Basic	Bachelor
HB67642	Arizona	34611.38	Yes	Basic	High School or Below
NV61299	Arizona	27789.69	Yes	Extended	Bachelor
UH35128	Oregon	25807.06	Yes	Extended	College
VL84149	Oregon	25807.06	Yes	Extended	College
TI61458	California	27789.69	Yes	Extended	Bachelor
WS53288	Oregon	25464.82	Yes	Extended	College
NQ67659	Nevada	33473.35	Yes	Basic	Bachelor
QL45827	Washington	25807.06	Yes	Extended	College

1-10 of 42 rows | 1-6 of 28 columns Previous **1** 2 3 4 5 Next

Hide

```
NA
```

After observing graphs it appeared that gender has no role to play, but we observe that all high valued customers who gave yes as response for policy renewal are females. So, we should focus more on high valued female customers who gave no as response, as they have more chance to say yes. People prefer to take 2 policies in general, so we can use it to our leverage.

Based on EDA, which variables are not important for prediction?

Customer,State, Marital Status, Education,Gender,Location code,Sales Channel,Renew offer type,Policy,Months Since Policy Inception, Months Since Last Claim,Income don't come across as important variables from our data exploration.Monthly Premium Auto is collinear with Total Claim Amount, so we need to drop one of these variables.

Based on EDA, which variables are important for prediction?

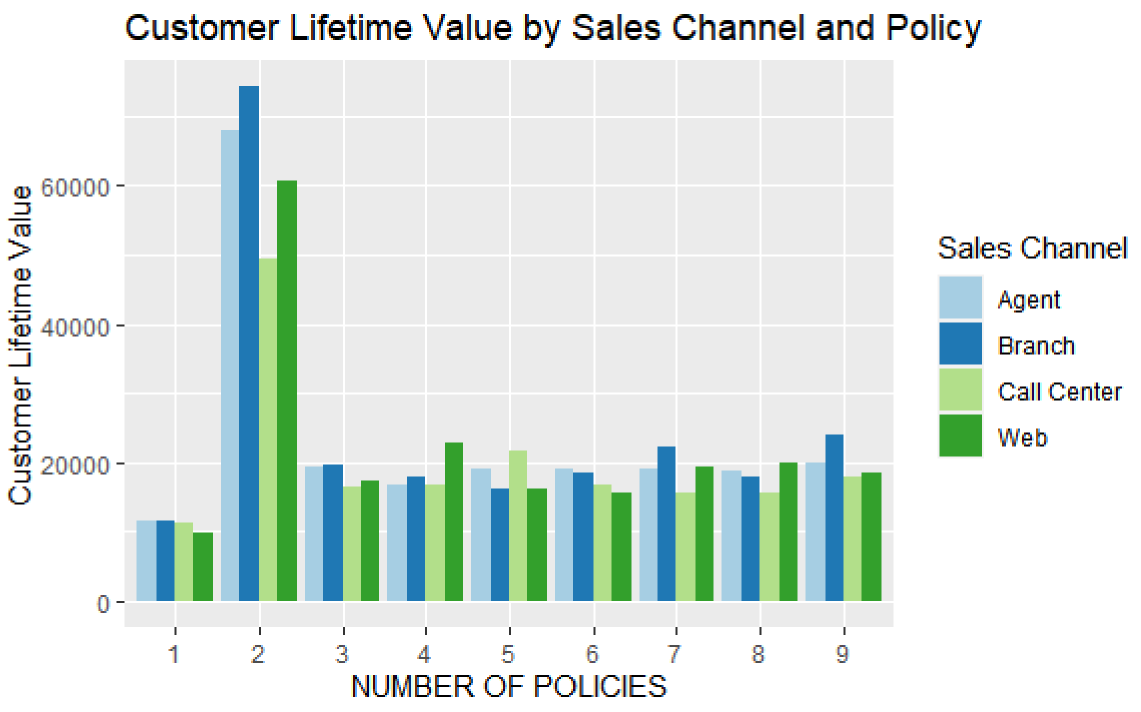
Response,Coverage,EmploymentStatus,Vehicle class,Policy type,Total Claim Amount appear to be important variables.

Multivariate Analysis

Hide

```
#install.packages("RColorBrewer")
library(RColorBrewer)

CLV_Type <- ggplot(data, aes(x=`Number of Policies`, y=`Customer Lifetime Value`, fill = `Sal
es Channel` ))+
        geom_col(position="dodge") + xlab("NUMBER OF POLICIES") + ylab("Customer Lifetime Val
ue") +
        ggtitle("Customer Lifetime Value by Sales Channel and Policy") +
         scale_fill_brewer(palette = "Paired")
CLV_Type
```



Hide

```
NA
NA
```

The Average number of policies that the Company issues comes to be around 2-3 in the the given time frame and it is noticeable that Call Centers fetch the most valuable customer.It is also to be noted that Customers having higher number of policies directly approach the Branch.Here the Policies to be issued vs the Lifetime Value of a Customer could be a trade off, since our focus is Lifetime Value we should be focusing on attracting more customers through Call centers.

Hide

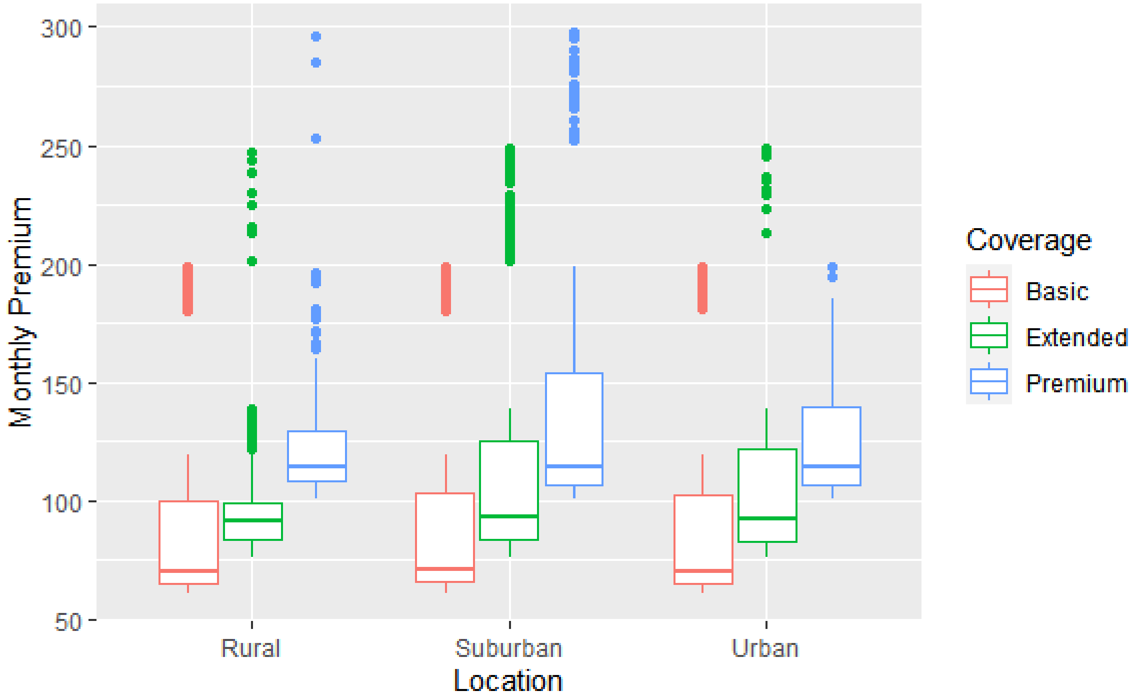
```
Claim_Type <- ggplot(data, aes(x=`Number of Policies`, y=data$`Total Claim Amount`, fill =`Po
licy Type` ))+
        geom_col(position="dodge") + xlab("NUMBER OF POLICIES") + ylab("Total Claim Amt") +
        ggtitle("Total Claim by no.of policy and Policy type") +
         scale_fill_brewer(palette = "Paired")
Claim_Type
```

It is evident that our major share of Customers procure Personal insurance and since the Total Claim Value determines the Lifetime Value of a customer the No.of Policy sold per category plays a crucial role to generate value to the company.

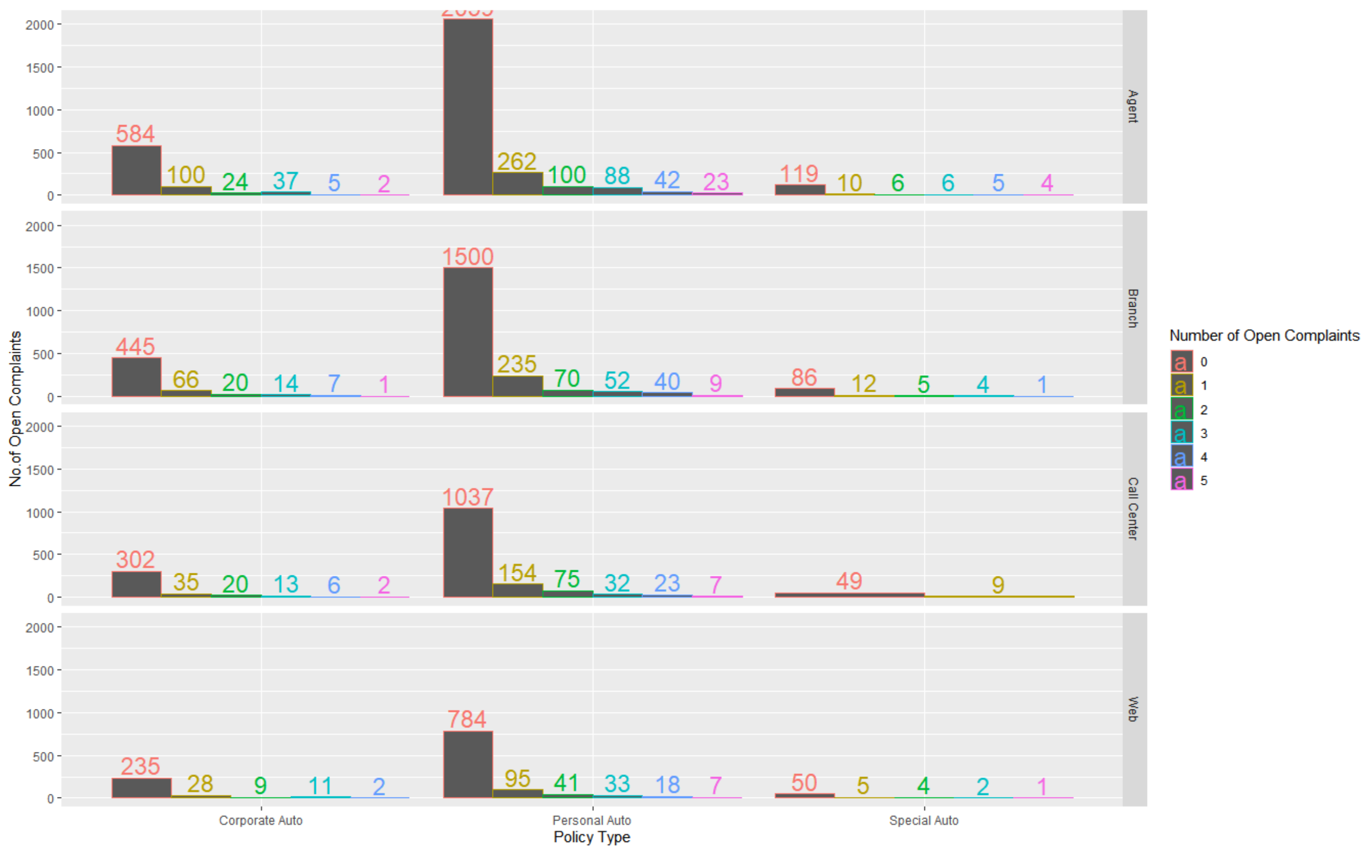
Hide

```
loc<-count(data,`Location Code`,Coverage)
#View(loc)


ggplot(data = data, aes(x = `Location Code` , y =`Monthly Premium Auto` , color = Coverage))
 +
  geom_boxplot() +
  xlab("Location") + ylab("Monthly Premium")
```



This to strategize the Plan to be promoted in different Locations, since we know most of our Customers come from Sub-Urban locations, where the Premium coverage numbers are high it is also seen that Premium coverage is preferred by most irrespective of the location.Hence prioritizing Premium Customers can benefit the business.



No.of Complaints can reflect upon relationship of the Customer with the Company given that the services are not managed well by the company.Most of our policies are for Personal purpose and are marketed by Call-Center.Personal policies being our focal point we see that those policies distributed by Agents have major number of Complaints which indicates poor-job by the Agents and delayed response by the company to the Target customers.

Modelling

Linear Regression

Linear Regression is the oldest, simple and widely used supervised machine learning algorithm for predictive analysis. It is a method to predict a target variable(Y) by fitting the best linear relationship between the dependent(Y) and independent variable(X). It helps determine:

> If a independent variable does a good job in predicting the dependent variable.

> Which independent variable plays a significant role in predicting the dependent variable.

In our analysis, we identified the target variable to be predicted as Customer Lifetime Value and all the others as dependent variable to run a multiple regression model.

Factorisation

Factor in R is a variable used to categorize and store the data, having a limited number of different values. It stores the data as a vector of integer values.

We converted all the categorical variables and 2 discrete numeric columns(Number of policies and number of open complaints) into factors to give each category a level to help the regression analysis.

Train-Test Split

The train-test split is a technique for evaluating the performance of the model

> Train Dataset: Used to fit the machine learning model.

> Test Dataset: Used to evaluate the fit machine learning model.

The objective is to estimate the performance of the model on new data: data not used to train the model.

Here, the data is randomly split in the ratio of 7:3 where training data constitutes 70% of the data and testing data is 30% of the complete dataset.

Hide

```
#LR TILL TRAIN TEST SPLIT
data$`Customer Lifetime Value`=log(data$`Customer Lifetime Value`)
set.seed(123)
sample <- sample(c(TRUE, FALSE), nrow(data), replace = T, prob = c(0.7,0.3))
train <- data[sample, ]
test <- data[!sample, ]
```

Approach

1. We initially built a base model with all the variables and got R^2 value of approximately 0.64. When we constructed a graph of Residual v/s Fitted the graph was funnel shaped indicating heteroscedasticity which is against the linear regression assumptions. This was due to the skewness in the target variable. Hence we log transformed the target variable.
2. Using the log transformed target column when we built a model, the score jumped up to 0.896. But the number of dependent variables used were 22. [However, the customer id column was initially dropped as it is of no importance] In order to optimize the usage of 22 dependent variables, we did various trial and error methods to get maximum information from minimum columns.
3. Statistical tests like correlation, Kruskal Wallis tests were also performed to see the dependence of each independent variable for the prediction of CLV.
4. However, we finally came up with the efficient model using anova for feature importance.

Feature Selection

The main aim of a regression analysis is to fetch as much information as possible with the least number of variables which are significant.To achieve this we need to select the most significant independent varibles that can explain the variation of the dependent variable. Feature selection was performed using a statistical test called ANOVA. ANOVA is a statistical test for estimating how a quantitative dependent variable changes according to the levels of one or more categorical independent variables

Hide

```
#ANOVA
#install.packages('AICcmodavg')
#library(AICcmodavg)

anova <- aov(`Customer Lifetime Value` ~
                 State+Response+Coverage +
                 Education+`Effective To Date` +EmploymentStatus+Gender+
                 Income+`Location Code`+`Months Since Policy Inception`+
                `Marital Status`+ `Months Since Last Claim`+`Policy Type`+
                 Policy+`Sales Channel`+
                 `Renew Offer Type` +`Vehicle Size`+
                 `Monthly Premium Auto`+
                 `Number of Open Complaints`+
                 `Number of Policies`+
                 `Total Claim Amount`+
                 `Vehicle Class`, data = data)

summary(anova)
```

```
                                  Df Sum Sq Mean Sq   F value   Pr(>F)    
State                              4    1.3     0.3     7.404 5.99e-06 ***
Response                           1    0.2     0.2     3.610   0.0575 .  
Coverage                           2  195.2    97.6  2228.672  < 2e-16 ***
Education                          4    4.0     1.0    22.599  < 2e-16 ***
`Effective To Date`                1    0.3     0.3     5.766   0.0164 *  
EmploymentStatus                   4   13.1     3.3    74.957  < 2e-16 ***
Gender                             1    0.2     0.2     3.944   0.0471 *  
Income                             1    0.0     0.0     0.876   0.3492    
`Location Code`                    2    0.4     0.2     4.142   0.0159 *  
`Months Since Policy Inception`    1    0.0     0.0     0.003   0.9586    
`Marital Status`                   2    2.6     1.3    29.332 2.01e-13 ***
`Months Since Last Claim`          1    0.2     0.2     4.094   0.0431 *  
`Policy Type`                      2    2.1     1.0    23.654 5.67e-11 ***
Policy                             6    2.2     0.4     8.216 6.88e-09 ***
`Sales Channel`                    3    1.2     0.4     9.462 3.08e-06 ***
`Renew Offer Type`                 3   72.0    24.0   548.299  < 2e-16 ***
`Vehicle Size`                     2    3.2     1.6    36.309  < 2e-16 ***
`Monthly Premium Auto`             1  545.7   545.7 12461.628  < 2e-16 ***
`Number of Open Complaints`        5   10.2     2.0    46.387  < 2e-16 ***
`Number of Policies`               8 2619.8   327.5  7478.553  < 2e-16 ***
`Total Claim Amount`               1    0.1     0.1     1.901   0.1680    
`Vehicle Class`                    5   22.4     4.5   102.221  < 2e-16 ***
Residuals                       9073  397.3     0.0                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the results obtatined from ANOVA we selected 9 variables to be significant with the Customer Lifetime Value variable based on their p values. All the variables significant at 0.1 level of significance and below were considered for building the model. Namely, Coverage,Education, Effective To Date, EmploymentStatus, Policy,Renew Offer Type,Monthly Premium Auto,Vehicle Class,Number of Open Complaints, Number of Policies. However we considered Policy over Policy type as they both explain similar things. We also neglected Effective To Date column as the results did not affect much with it. Also, we have not used the derived column “Present Value of Customer” as only Monthly Premium Auto was considered significant whereas other variables used weren’t. The R^2 value dropped when the derived column was used.

Model Building
Using the important variables obtained from ANOVA we trained a linear regression model using lm function on the training data.

Accuracy Measures

a. R^2 value- It represents the proportion of variance explained and always takes on a value between 0 and 1. The higher the value, better the model. It is independent of the scale of Y.
b. RMSE - Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are. RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

Hide

```
result = broom::glance(fit_log)
result
```

r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>	df <dbl>	logLik <dbl>	AIC <dbl>	BIC <dbl>
0.8959918	0.8953414	0.2106082	1377.694	0	40	914.2711	-1744.542	-1460.203

1 row | 1-10 of 12 columns

Hide

```
NA
NA
NA
```

Hide

```
#install.packages("modelr")
library(modelr)
#install.packages("broom")# provides easy pipeline modeling functions
library(broom)
```

```

Attaching package: 'broom'

The following object is masked from 'package:modelr':

    bootstrap
```

Hide

```
#install.packages("Metrics")
library(Metrics)
```

```

Attaching package: 'Metrics'

The following objects are masked from 'package:modelr':

    mae, mape, mse, rmse

The following objects are masked from 'package:caret':

    precision, recall
```

Hide

```
library(dplyr)
#install.packages("pbkrtest", dependencies = TRUE)
#library(caret)
test %>%
  add_predictions(fit_log)
```

Custo… <chr>	State <fctr>	Customer Lifetime Value <dbl>	Respo… <fctr>	Covera… <fctr>	Education <fctr>

Custo… <chr>	State <fctr>	Customer Lifetime Value <dbl>	Respo… <fctr>	Covera… <fctr>	Education <fctr>
QZ44356	Arizona	8.850738	No	Extended	Bachelor
WW63253	California	8.941920	No	Basic	Bachelor
HB64268	Washington	7.942253	No	Basic	Bachelor
CF85061	Arizona	8.884070	No	Premium	Master
SX51350	California	8.463580	No	Basic	College
BW63560	Oregon	8.917731	No	Basic	Bachelor
FL50705	California	9.007320	No	Premium	High School or Below
ZK25313	Oregon	7.962782	No	Basic	High School or Below
TZ98966	Nevada	7.803921	No	Basic	Bachelor
FS42516	Oregon	8.665969	No	Basic	College

1-10 of 2,696 rows | 1-6 of 25 columns Previous **1** 2 3 4 5 6 ... 100 Next

Hide

```
  #summarise(MSE = mean((`Customer Lifetime Value`-pred)^2))

y_train = predict(fit_log, newdata = train)
R2 <- 1- (sum((train$`Customer Lifetime Value`-y_train)^2) / sum((train$`Customer Lifetime Va
lue` - mean(train$`Customer Lifetime Value`))^2))
print(R2 * 100)
```

```
[1] 89.59918
```

Hide

```
caret::RMSE(y_train,train$`Customer Lifetime Value`)
```

```
[1] 0.2099365
```

Hide

```
y_test = predict(fit_log, newdata = test)
R3 <- 1- (sum((test$`Customer Lifetime Value`-y_test)^2) / sum((test$`Customer Lifetime Value
` - mean(test$`Customer Lifetime Value`))^2))
print(R3 * 100)
```

```
[1] 89.82091
```

Hide

```
caret::RMSE(y_test,test$`Customer Lifetime Value`)
```

```
[1] 0.2096528
```

After training the model , we can get a summary of results using broom package which gives r^2, adj r^2, RSE and other accuracy measures to asses the model performace.

However, we assessed the training and testing predictions by calculting the R^ value and RMSE for training and testing data.

The results obtained are: train data : R^2=0.8959482 and RMSE=0.2099805 test data : R^2=0.8982095 and RMSE=0.2096523

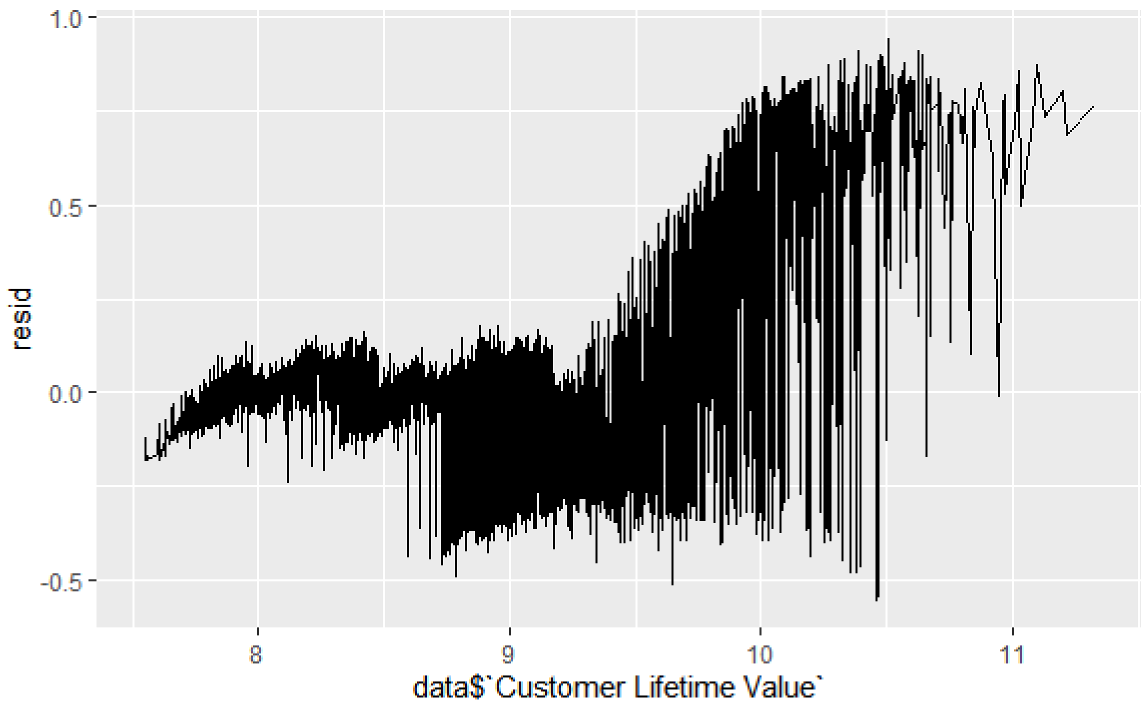
The obtained results were satisfactory with less error and hence concluded this to be the best fit model for prediction of the target variable Customer Lifetime Value from the most significant variables.

Assumptions and Residual Plots for Accuracy Measure

Correlation of Error Terms

Hide

```
data %>%
add_residuals(fit_log) %>%
   ggplot(aes(data$`Customer Lifetime Value`, resid)) +
  geom_line()
```



Correlation of error term An important assumption of the linear regression model is that the error terms are uncorrelated. This scatterplot is used to detect a particular form of non-independence of the error terms, namely serial correlation. A Residual vs. order plot helps to see if there is any correlation between the error terms that are near each other in the sequence.However, if we look at our modelâ€™s residuals we see that adjacent residuals do not tend to take on similar values, hence error terms are not correlated.

Detecting Multicollinearity using GVIF

Hide

```
car::vif(fit_log)
```

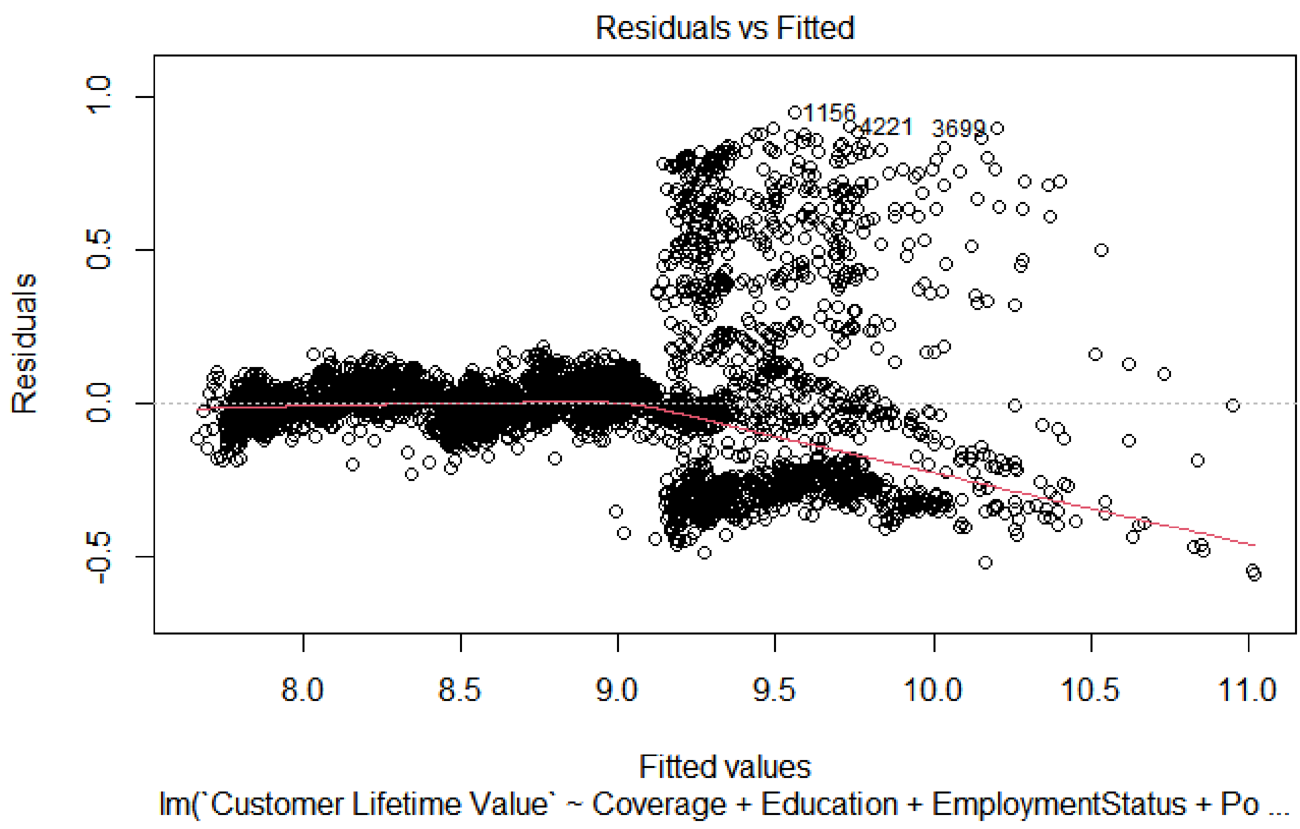
```
                                GVIF Df GVIF^(1/(2*Df))
Coverage                    6.080094  2        1.570282
Education                   1.050078  4        1.006127
EmploymentStatus            1.101621  4        1.012171
Policy                      1.038022  8        1.002335
`Renew Offer Type`          1.131443  3        1.020796
`Monthly Premium Auto`     25.656367  1        5.065211
`Number of Open Complaints` 1.049274  5        1.004821
`Number of Policies`        1.070495  8        1.004267
`Vehicle Class`            20.693570  5        1.353891
```

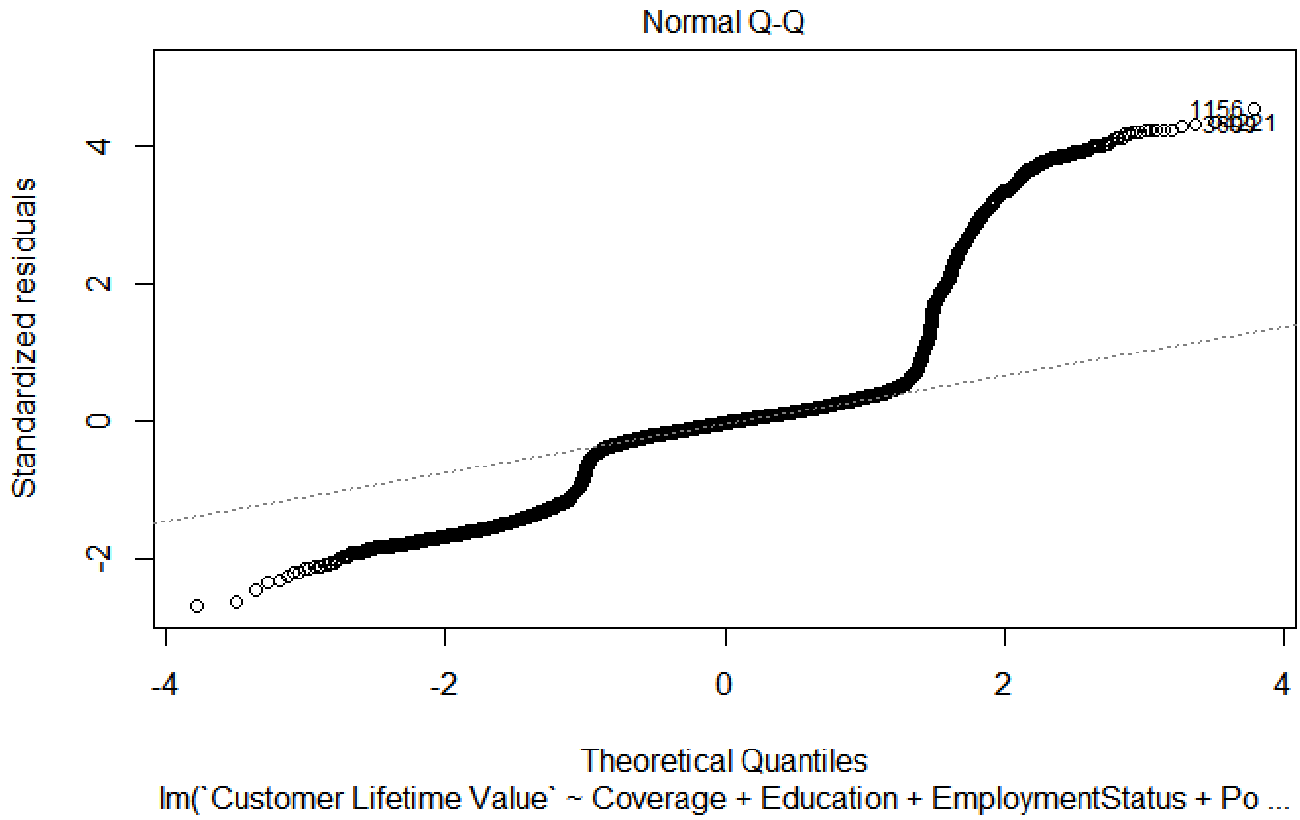
Multicollinearity can be assessed by computing the VIF(variance inflation factor) value. Any variable with a high VIF value should be removed from the model. In our model all the variables have VIF values between 1 and 5 which satisfies the condition, means no severity of multicollinearity. Here gvif is the square root of the VIF for individual predictors and thus can be used equivalently

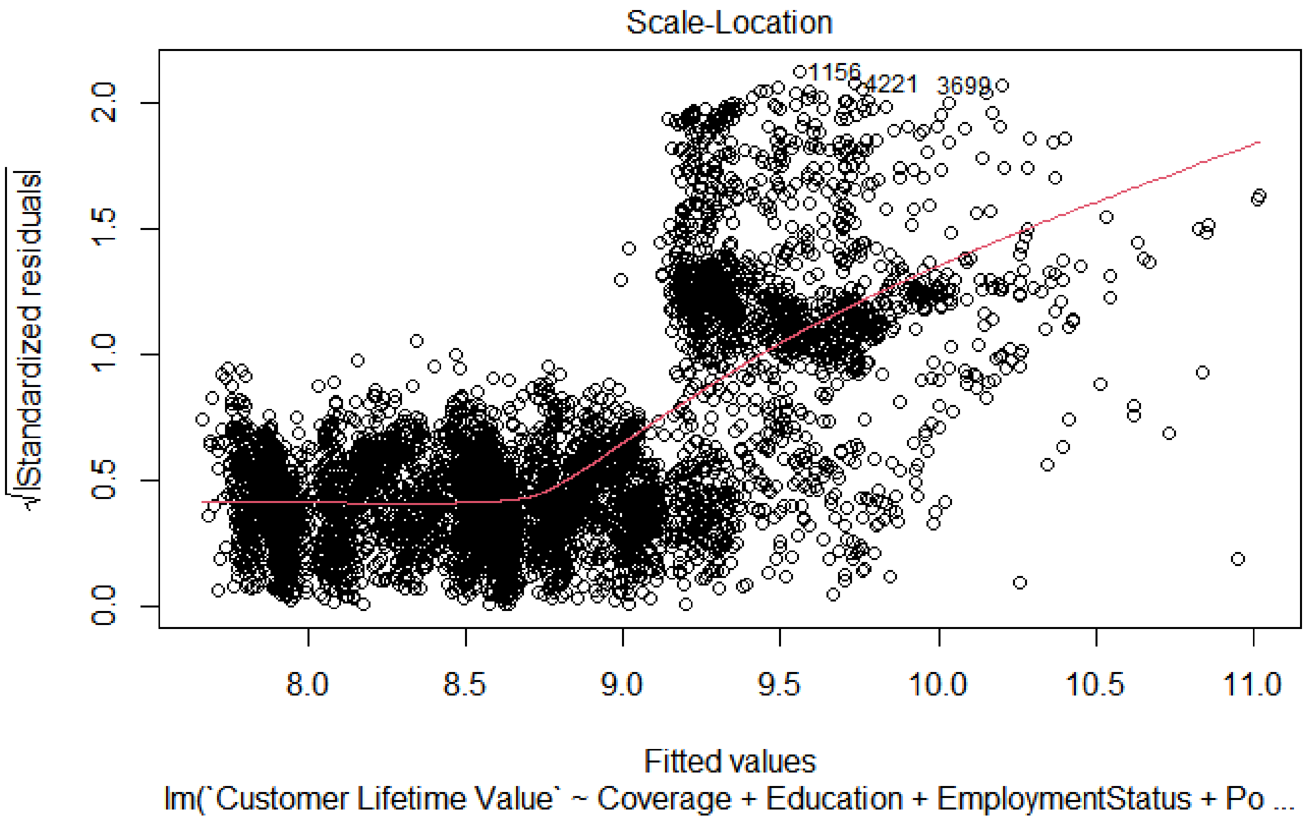
Residual Plots

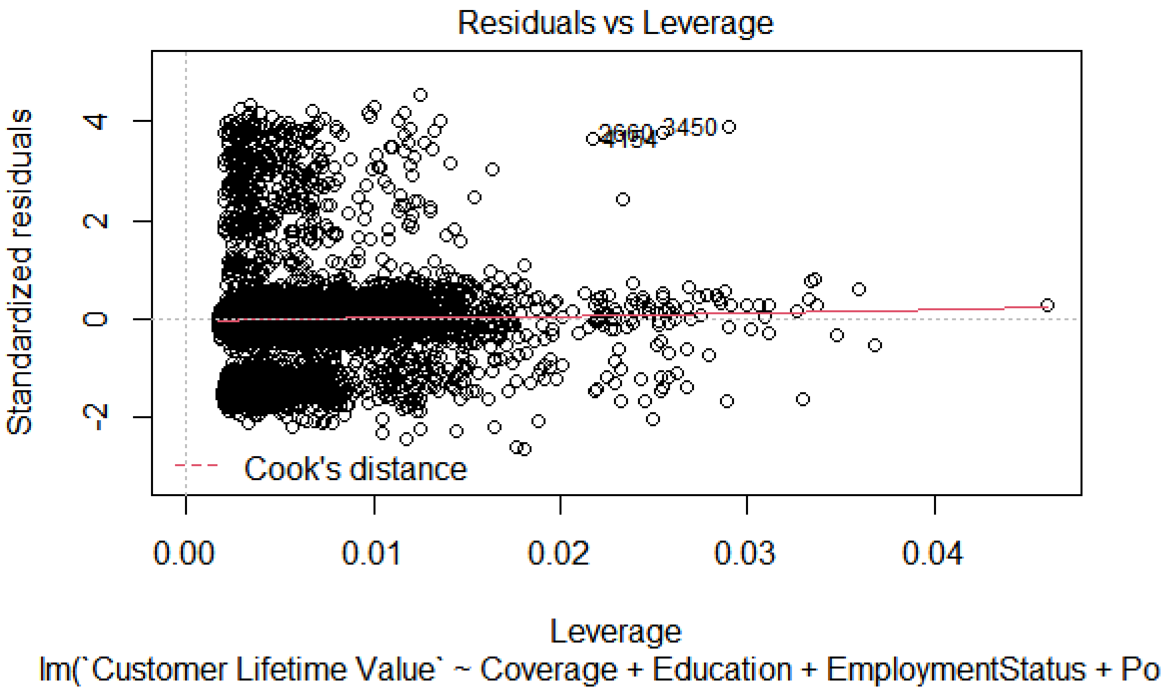
Hide

```
plot(fit_log)
```




Normal Q-Q
Standardized residuals
-2
0
2
4
-4
-2
0
2
4
Theoretical Quantiles
lm(`Customer Lifetime Value` ~ Coverage + Education + EmploymentStatus + Po ...


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1156
4221
3699
8.0
8.5
9.0
9.5
10.0
10.5
11.0
Fitted values
lm(`Customer Lifetime Value` ~ Coverage + Education + EmploymentStatus + Po ...



1. Residual Vs Fitted Values In this scatter plot, the distribution of residuals (errors) vs fitted values (predicted values) is depicted.Since the plot now does not show a funnel shape, it is an indication of constant variance, i.e.homoskedasticity. Also, since there is no recognizable pattern seen, it indicates that the assumption of linearity is fair.

2. Normal Q-Q Plot This q-q, or quantile-quantile, scatter plot helps in the validation of the normal distribution assumption in our data set. We can infer if the data has a normal distribution by looking at this graph. If this is the case, the plot will tend to be fairly straight line. In our case, there is strong deviation from the diagonal line which shows that our residuals are not normally distributed.

3. Scale-Location Plot This plot can be used to detect homoskedasticity (assumption of equal variance). It displays how the residuals are distributed across the predictor range. It’s similar to the residual vs. fitted value plot, but it actually uses standardised residual values. Here, we can see a diagonal line with somewhat equally distributed points which shows less homoskedasticity.

4. Residuals Vs Leverage Plot This is also known as Cookâ€™s Distance plot. It is a method of determining which points have more influence than others. Such influential locations have a significant impact on the regression line. In our case, Cookâ€™s distance scores are high and are clustered near the top of our leverage plot, indicating that they have a significant influence on the regression results.

Random Forest

Hide

```
#install.packages("randomForest")
library(randomForest)
```

```
Warning: package ‘randomForest’ was built under R version 4.0.5
randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.

Attaching package: ‘randomForest’

The following object is masked from ‘package:gridExtra’:

    combine

The following object is masked from ‘package:ggplot2’:

    margin

The following object is masked from ‘package:dplyr’:

    combine
```

Hide

```
#head(train)
#colnames(train)
RF_fit<- randomForest(`Customer Lifetime Value` ~
             Coverage+
             Education+
             EmploymentStatus+
             train$`Policy` +
             train$`Renew Offer Type`+
             train$`Monthly Premium Auto`+
             train$`Number of Open Complaints`+
             train$`Number of Policies`+
             train$`Vehicle Class`,
             data=train)
RF_fit
```

```


Call:
 randomForest(formula = `Customer Lifetime Value` ~ Coverage +      Education + EmploymentSta
tus + train$Policy + train$`Renew Offer Type` +      train$`Monthly Premium Auto` + train$`Nu
mber of Open Complaints` +      train$`Number of Policies` + train$`Vehicle Class`, data = tr
ain)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.04106598
                    % Var explained: 90.31
```

Hide

```
#caret::RMSE(y_test,test$`Customer Lifetime Value`)
```

1. From the summary results of the predicted values, when the Random Forest Regressor is tasked with the problem of predicting for values not previously seen, it will always predict an average of the values seen previously. Obviously the average of a sample can not fall outside the highest and lowest values in

the sample. The Random Forest Regressor is unable to discover trends that would enable it in extrapolating values that fall outside the training set. When faced with such a scenario, the regressor assumes that the prediction will fall close to the maximum value in the training set.

2. The obtained train score is 95.73442 and test score is 89.81142 for Random Forest. This show that there is a slight overfitting of the model.
3. The complexity of Random Forest model is high as it is based on bootstrap aggregation and bagging techniques.

Support Vector Machine

Hide

```
#SVM
#install.packages("e1071")
library(e1071)
```

```
Warning: package ‘e1071’ was built under R version 4.0.5

Attaching package: ‘e1071’

The following object is masked from ‘package:Hmisc’:

    impute
```

Hide

```
fit<- svm(data$`Customer Lifetime Value` ~
           data$Education+data$`Effective To Date` +data$EmploymentStatus+data$Gender+
           data$Income+data$`Location Code` +data$`Months Since Policy Inception`+
           data$`Marital Status`+ data$`Months Since Last Claim`+data$`Policy Type`+
           data$`Monthly Premium Auto` + data$Policy +data$`Sales Channel`+
           data$`Number of Open Complaints` + data$`Number of Policies` +
           data$`Renew Offer Type` + data$`Total Claim Amount`+
           data$`Vehicle Class`+data$`Vehicle Size`, data=train)
summary(fit)
```

```
Call:
svm(formula = data$`Customer Lifetime Value` ~ data$Education + data$`Effective To Date` + 
    data$EmploymentStatus + data$Gender + data$Income + data$`Location Code` + data$`Months S
ince Policy Inception` + 
    data$`Marital Status` + data$`Months Since Last Claim` + data$`Policy Type` + data$`Month
ly Premium Auto` + 
    data$Policy + data$`Sales Channel` + data$`Number of Open Complaints` + data$`Number of P
olicies` + 
    data$`Renew Offer Type` + data$`Total Claim Amount` + data$`Vehicle Class` + data$`Vehicl
e Size`, 
    data = train)


Parameters:
   SVM-Type:  eps-regression 
 SVM-Kernel:  radial 
       cost:  1 
      gamma:  0.01785714 
    epsilon:  0.1 


Number of Support Vectors:  2853
```

Hide

```
predictedY_ <- predict(fit, train)
error_2 <- train$"Customer Lifetime Value" - predictedY_
svm_error <- sqrt(mean(error_2^2))
svm_error
```

```
[1] 0.8497346
```

Hide

```
predictedY <- predict(fit, test)
error_2 <- test$`Customer Lifetime Value` - predictedY
svm_error <- sqrt(mean(error_2^2))
svm_error
```

```
[1] 0.8459781
```

On applying SVM model, we get 7900 RMSE which is much higher than the basic model of Linear regression which gave us approx 4000 RMSE . These RMSE's have been compared without log transformation, and if we want better prediction from SVM we need to optimize the model in a better way.

Results of important variables from EDA, ANOVA and Linear Regression

> 1. EDA

Response,Total Claim Amount,Coverage,EmploymentStatus,Vehicle class,Policy type, Monthly Premium Auto [Due multicollinearity b/w Total Claim Amount and Monthly Premium Auto, rejected Total Claim Amount to be an important variable.]

2. ANOVA

Coverage,Education,EmploymentStatus, Policy,Renew Offer Type,Monthly Premium Auto,Vehicle Class,Number of Open Complaints,Number of Policies

3. LINEAR REGRESSION

Education,Employment Status, Gender, Monthly premium Auto, Sales Channel, Number of open complaints, Number of policieS, Vehicle Class

Conclusion

Why CLV?

Ultimately, the company just needs to be mindful of the value that a customer provides over their lifetime relationship with it. By understanding the customers' details regarding various aspects and analyzing all key touchpoints, one can understand the key drivers of CLV. CLV is indeed a great metric that should be used to improve business strategies.

How good is the analysis?

Performing Exploratory Data Analysis, Statistical tests and using Statistical models like linear regression we analysed the effect of different variables on the variation of the target variable - Customer Lifetime Value and also concluded which are the important variables that are sufficient to give maximum information to predict the CLV. The results obtained from all the 3 methods were approximately similar and hence the analysis seems to be efficient.

How good is the final model?

Considering number of factors, we can conclude that Linear Regression model, being one of the simplest models has given the best R^2 value of 0.89 with least error rate of 0.20 compared to other models like random forest and Support Vector Machine which has its own cons over Linear regression. We have optimized the model to predict the Customer Lifetime Value from 9 independent and hence can be concluded as a pretty good model for prediction.

Business recommendations

From the model and the analysis we designed, we can suggest that:

1. The company should target mainly the customers who are employed.
2. The customers whose Education is master level should be targeted more whereas doctors do not serve to be valuable customers.
3. The number of complaints should be reduced because more number of complaints the company is more prone towards losing the customer. Compalints above 2 should focused.

4. More attention should be given to the Extended and premium customers.
5. The target audience should be female as they are easier to convince according to our analysis of response given towards the policy.
6. The company should start increasing their policy advertisement through branch and agents as the number of policy affects the CLV.

Suggestions

Suggesting new variables that can improve the analysis: 1. Debt to income ratio 2. Number of houses owned 3. Number of cars owned 4. Type of purchase- Installment or one-off 5. Price of insured commodity

Contribution:

Dona Sam (20BDA02) - EDA + Assumptions of Linear Regression

Prathibha K S (20BDA15) - EDA + Linear Regression + Random Forest + Report writing

Reba Susan Joseph (20BDA37) - EDA + Assumptions of Linear Regression

Jayasree C (20BDA53) - EDA + Compiling for the final report + Report Writing

Abhijith P K (20BDA60) - EDA + Statistical Tests

Ananya Kumari (20BDA68) - Compiling and deriving insights from EDA + SVM + Stepwise Regression + Report Writing

References:

1. https://medium.com/@aritraadhikari.b3/predicting-customer-lifetime-value-for-an-auto-insurance-company-3b24d8bf4e24 (https://medium.com/@aritraadhikari.b3/predicting-customer-lifetime-value-for-an-auto-insurance-company-3b24d8bf4e24)
2. https://mediaalpha.com/article/why-auto-insurance-advertisers-should-optimize-for-customer-lifetime-value/ (https://mediaalpha.com/article/why-auto-insurance-advertisers-should-optimize-for-customer-lifetime-value/)
3. https://github.com/abhiyerasi/CLV-Auto-Insurance/blob/master/Code/Clustering.R (https://github.com/abhiyerasi/CLV-Auto-Insurance/blob/master/Code/Clustering.R)
4. https://www.kaggle.com/dktalaicha/predict-customer-life-time-value-clv#1.-Objective-:--Predict-Customer-Life-time-Value-(CLV-)for-an-Auto-Insurance-Company (https://www.kaggle.com/dktalaicha/predict-customer-life-time-value-clv#1.-Objective-:--Predict-Customer-Life-time-Value-(CLV-)for-an-Auto-Insurance-Company).
5. https://www.kaggle.com/juancarlosventosa/models-to-improve-customer-retention (https://www.kaggle.com/juancarlosventosa/models-to-improve-customer-retention)
6. https://www.kaggle.com/prathibhaks/notebookaaa340ee38/edit (https://www.kaggle.com/prathibhaks/notebookaaa340ee38/edit)

Hide