# R Notebook

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

# This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

title: “R Notebook” output: html_document: df_print: paged html_notebook: default pdf_document: default —

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

# Customer Lifetime Value Prediction

## Problem Statement:

To predict the Customer Lifetime Value for an insurance company offering vehicle insurance.

## Description:

Customer Lifetime Value is a commonly used metric by companies and financial institutions to assign a numeric value to their customers and thereby inform their strategy of increasing the companies profits.

It is defined as the total monetary value that a customer holds to a bank or any financial entity over the entire course of their relationship.

### The formula used to calculate CLV is = (Annual revenue per customer x Customer relationship in years) – Customer acquisition cost

The difficulty arises when some segments of customers invest a lot of money in a company over a short period of time while others might invest small sums over a longer period of time. Now, if a company were to focus only on the short-term high paying customers, they will miss out on the gradual but constant revenue invested by the latter kind of customer. Both of these kinds of customers might be of high value to the company and hence there is a need to account for these two kinds of customers as well as other factors.

In the case of insurance, customers fall into several categories. Companies design different policies as not all categories of customers will want the same policy. Some customers might go for a greater coverage, while some might go for less. This does not mean that the customers with lesser coverage are less valuable to the company, as we must take into account the cost of acquiring these customers as well.

The insurance company must therefore study their existing customers considering all these factors to find out which category of customers to target.

The dataset contains historical data of the customers already acquired by the company and the CLV for each of these customers has been computed. We must use this previously computed CLV along with the independent variables to predict the category of customers who will be profitable to the company.

To account for all these factors this metric of customer-oriented evaluation is widely used.

## Aim:

To establish the relationship between the explanatory variables and the target variable and thereby to propose a model that can predict the target variable.

In this case, the objective is to study how the outcome variable (CLV) is related to the independent variables and the subsequent model thus proposed should help the company to make an informed decision with regard to the kind of customer to target.

It is a regression task to predict how much a given customer will be valuable to an insurance company.

# Exploratory Data Analysis

Following are the packages used.

```
library(tidyverse)
library(car)
library(zoo)
library(lmtest)
library(dplyr)
library(stringr)
library(caret)
library(ggplot2)
library(timeDate)
library(plotly)
library(readxl)
library(gganimate)
library(corrplot)
library(Hmisc)
library(vtree)
library(DataExplorer)
library(caTools)
library(nortest)
library(modelr)
```

# Reading the dataset

```
Marketing_Customer_Value_Analysis_2 <- read_excel('C:/Users/HP/Downloads/Marketing-Customer-Value-Analysis 2.xlsx')
```

# Setting seed for reproducibility and overview of dataset

```
set.seed(223)
head(Marketing_Customer_Value_Analysis_2)
```

| Customer | State | Customer Lifetime Value | Response | Coverage | Education | Effective To Date |
|---|---|---:|---|---|---|---|
| <chr> | <chr> | <dbl> | <chr> | <chr> | <chr> | <chr> |
| BU79786 | Washington | 2763.519 | No | Basic | Bachelor | 2/24/11 |
| QZ44356 | Arizona | 6979.536 | No | Extended | Bachelor | 1/31/11 |
| AI49188 | Nevada | 12887.432 | No | Premium | Bachelor | 2/19/11 |

| Customer | State | Customer Lifetime Value | Response | Coverage | Education | Effective To Date |
|---|---|---:|---|---|---|---|
| <chr> | <chr> | <dbl> | <chr> | <chr> | <chr> | <chr> |
| WW63253 | California | 7645.862 | No | Basic | Bachelor | 1/20/11 |
| HB64268 | Washington | 2813.693 | No | Basic | Bachelor | 2/3/11 |
| OC83172 | Oregon | 8256.298 | Yes | Basic | Bachelor | 1/25/11 |

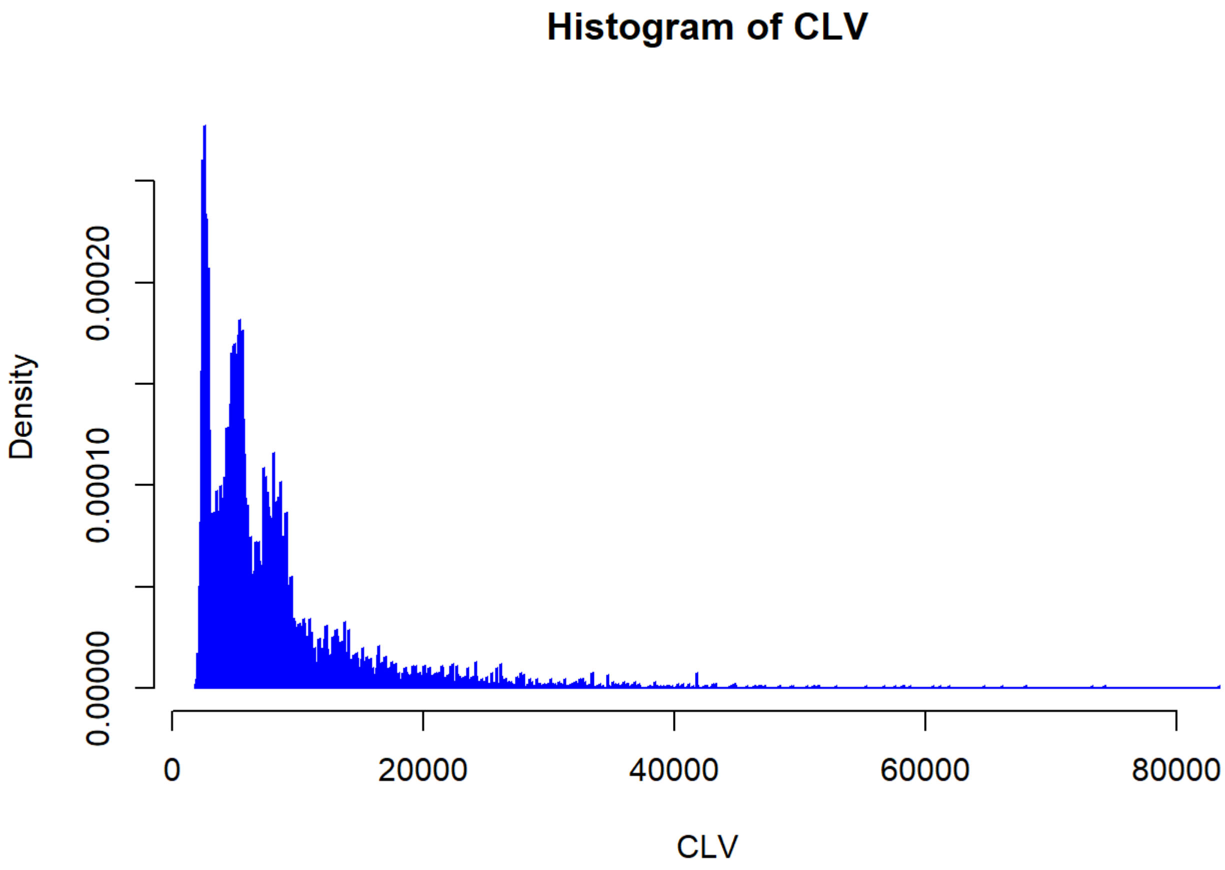
6 rows | 1-7 of 24 columns

# Histogram of the target variable (CLV)

This shows us the distribution of the target variable, where y-axis contains the probability density of the target variable.

This tells us the monetary value that the customers represent to the company.

```
Insurance_Dataset <- data.frame(Marketing_Customer_Value_Analysis_2)
hist(Insurance_Dataset$Customer.Lifetime.Value,
     breaks = 800,
     freq = FALSE,
     main = "Histogram of CLV", xlab = "CLV", border = "Blue")
```



This plot indicates that the distribution is heavily positively skewed, meaning that an overwhelming majority of the customers hold lower customer lifetime value to the company.A very small number of customers are in the higher bracket of lifetime value.

The “ideal” customers to the company are small in number and if the company is to turn a profit they must also focus on catering to the customers with lower CLV as they are more in number.

Description of dataset

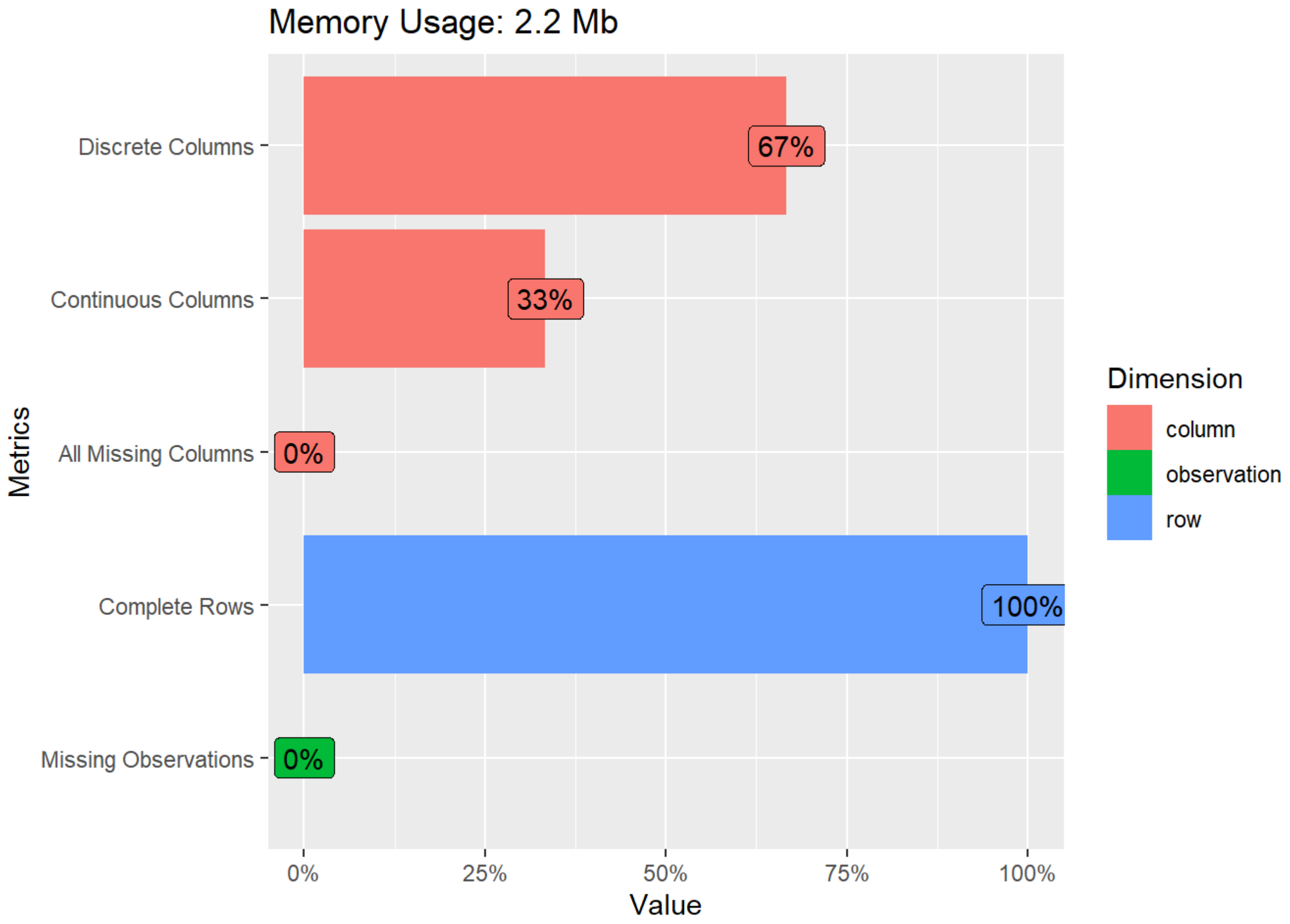
```
Insurance_Dataset %>% introduce()
```

| ro… | columns | discrete_columns | continuous_columns | all_missing_columns | total_missing_values ▶ |
|---|---|---|---|---|---|
| <int> | <int> | <int> | <int> | <int> | <int> |

| ro... | columns | discrete_columns | continuous_columns | all_missing_columns | total_missing_values |
|---|---|---|---|---|---|
| <int> | <int> | <int> | <int> | <int> | <int> |
| 9134 | 24 | 16 | 8 | 0 | 0 |

1 row | 1-6 of 9 columns

```
Insurance_Dataset %>% plot_intro()
```

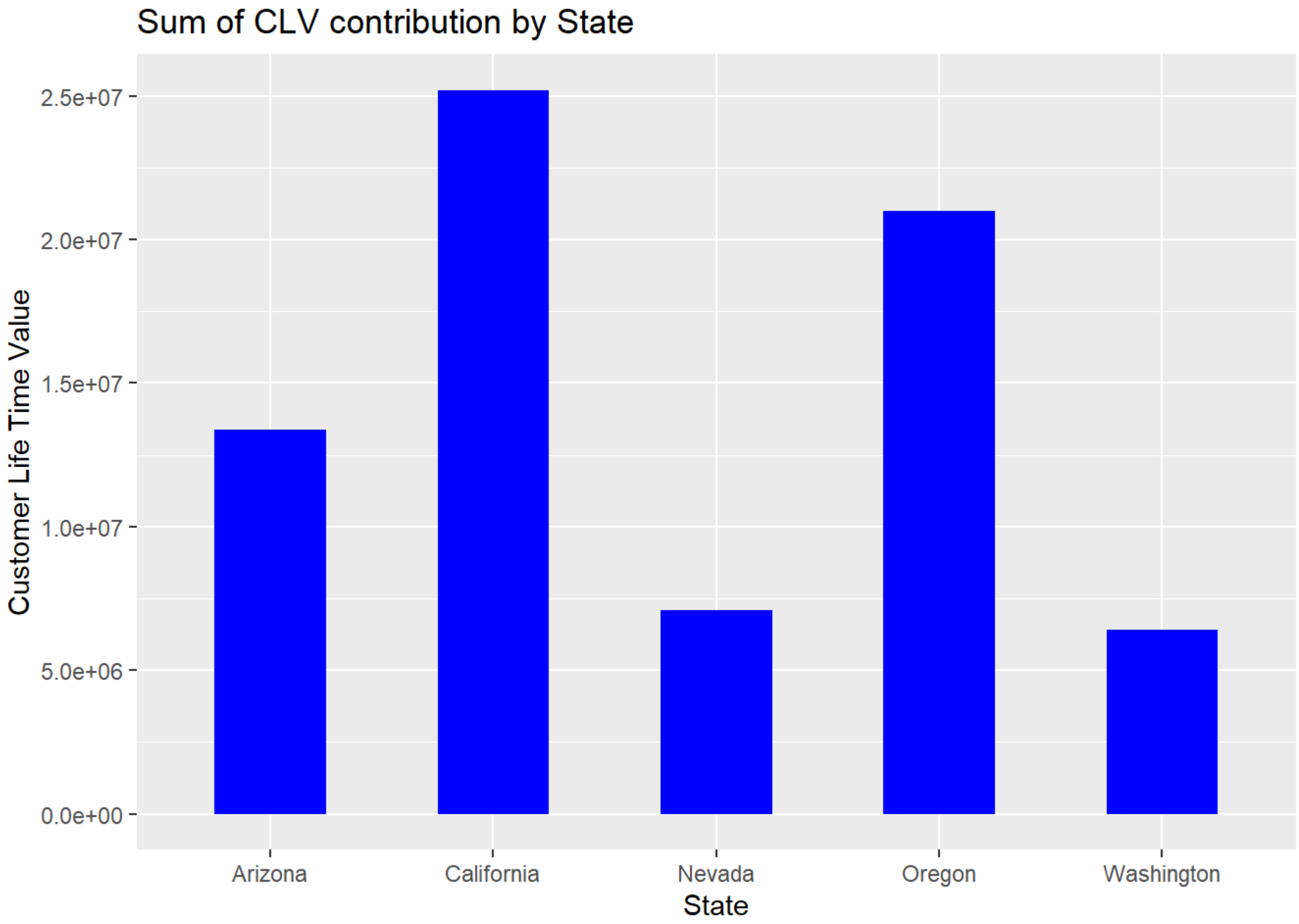


There are no null values in this dataset.

# CATEGORICAL VARIABLES VISUALIZATION

## To visualize the effect of state on CLV

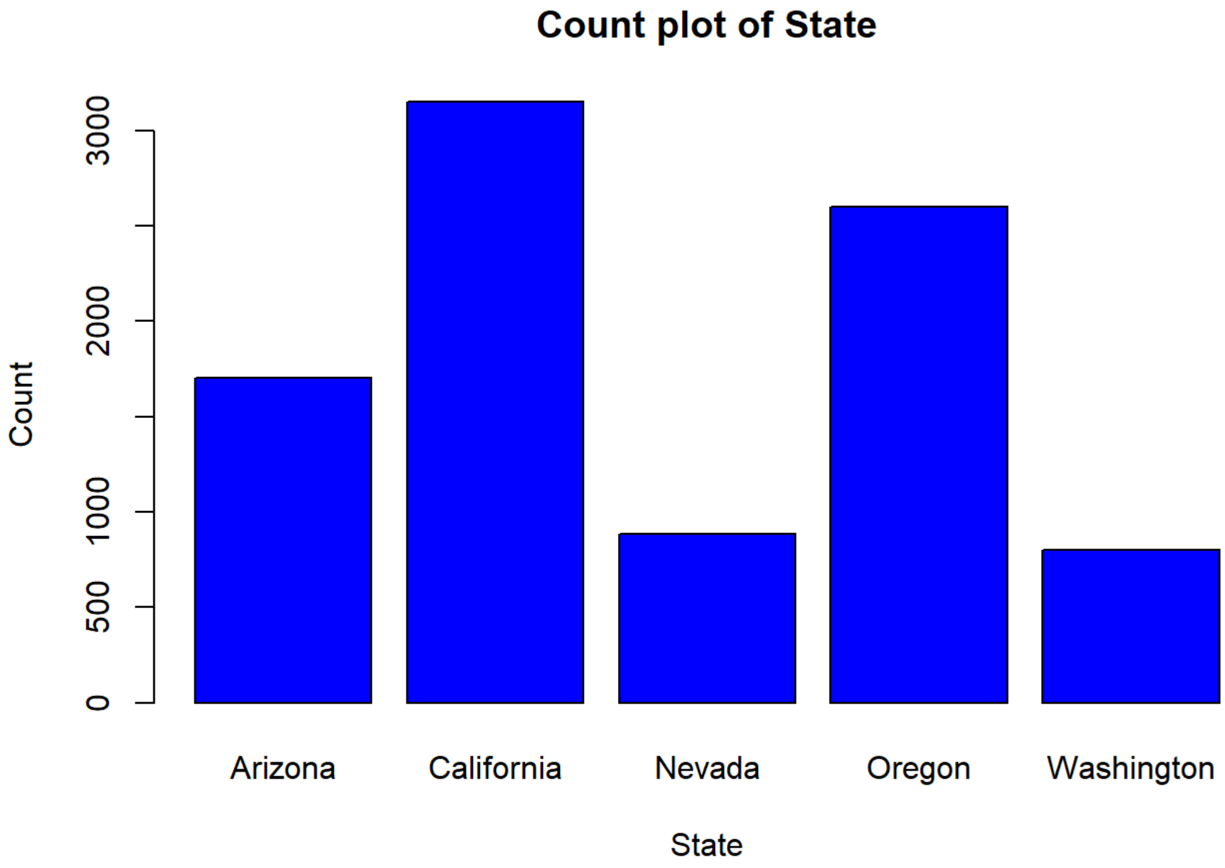
```
ggplot(Insurance_Dataset,aes (x=State ,
              y=Customer.Lifetime.Value)) + geom_bar(stat="summary",fun="sum", width=0.5, fill = "Blue")+
  labs(x="State",y = "Customer Life Time Value", fill="State") +
  ggtitle("Sum of CLV contribution by State ")
```



In this case, we’re looking at how much effect a customer’s state has on CLV. In other words, we are trying to find if a customer from a particular state is more valuable to the company than other states.

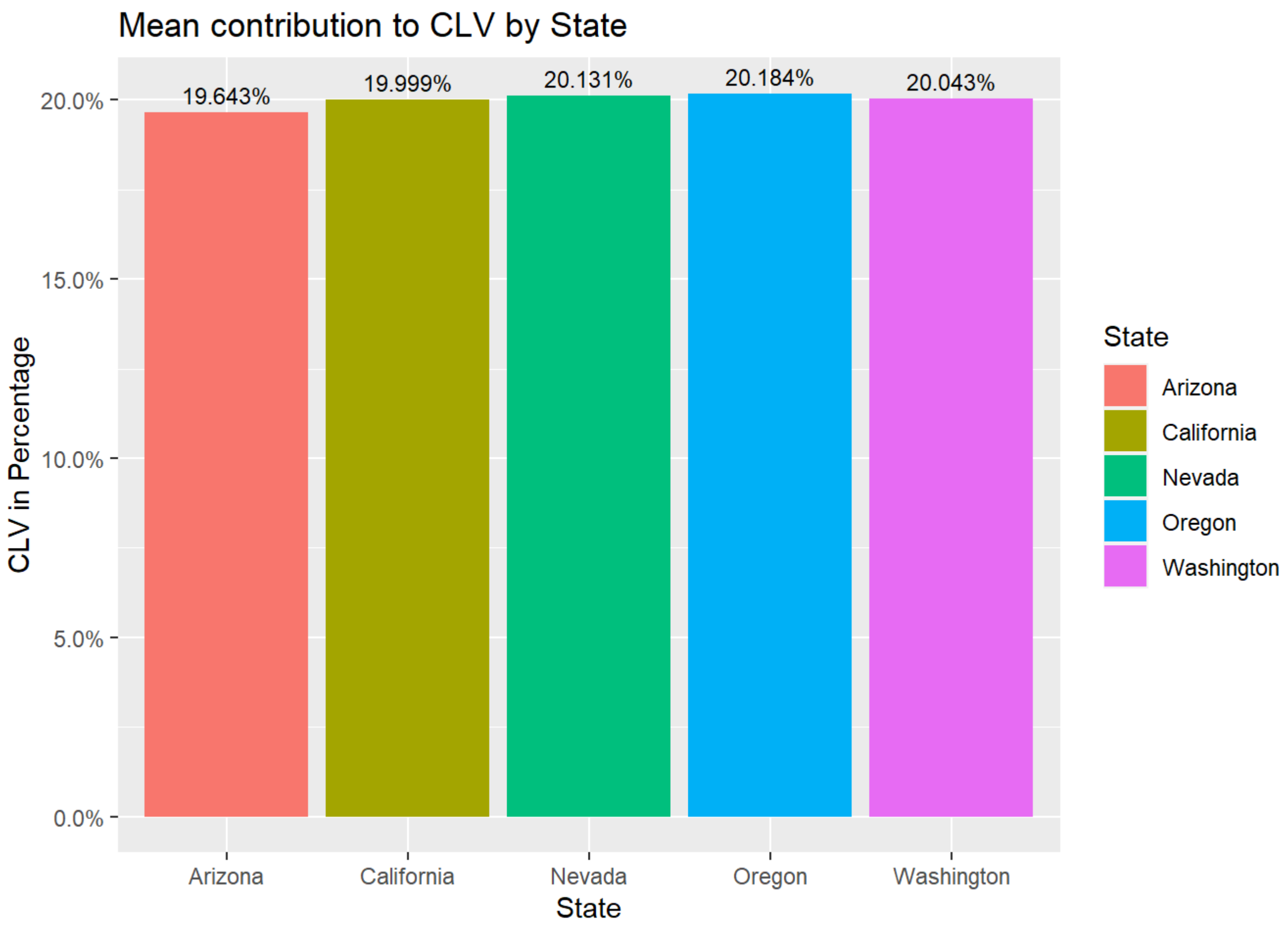
From the above chart it would appear that the company should focus their efforts on states like California or Oregon, since the sum of CLV from these states are higher. As we can see in the chart below the population of these states is a factor for the high CLV obtained.

```
count_state <- table(Insurance_Dataset$State)
barplot(count_state,
        main = "Count plot of State",col = "Blue",
        xlab = "State", ylab = "Count")
```



Let us explore this by considering the mean of the CLV by state in the subsequent chart. This measure will account for the larger populations of states like California and Oregon.

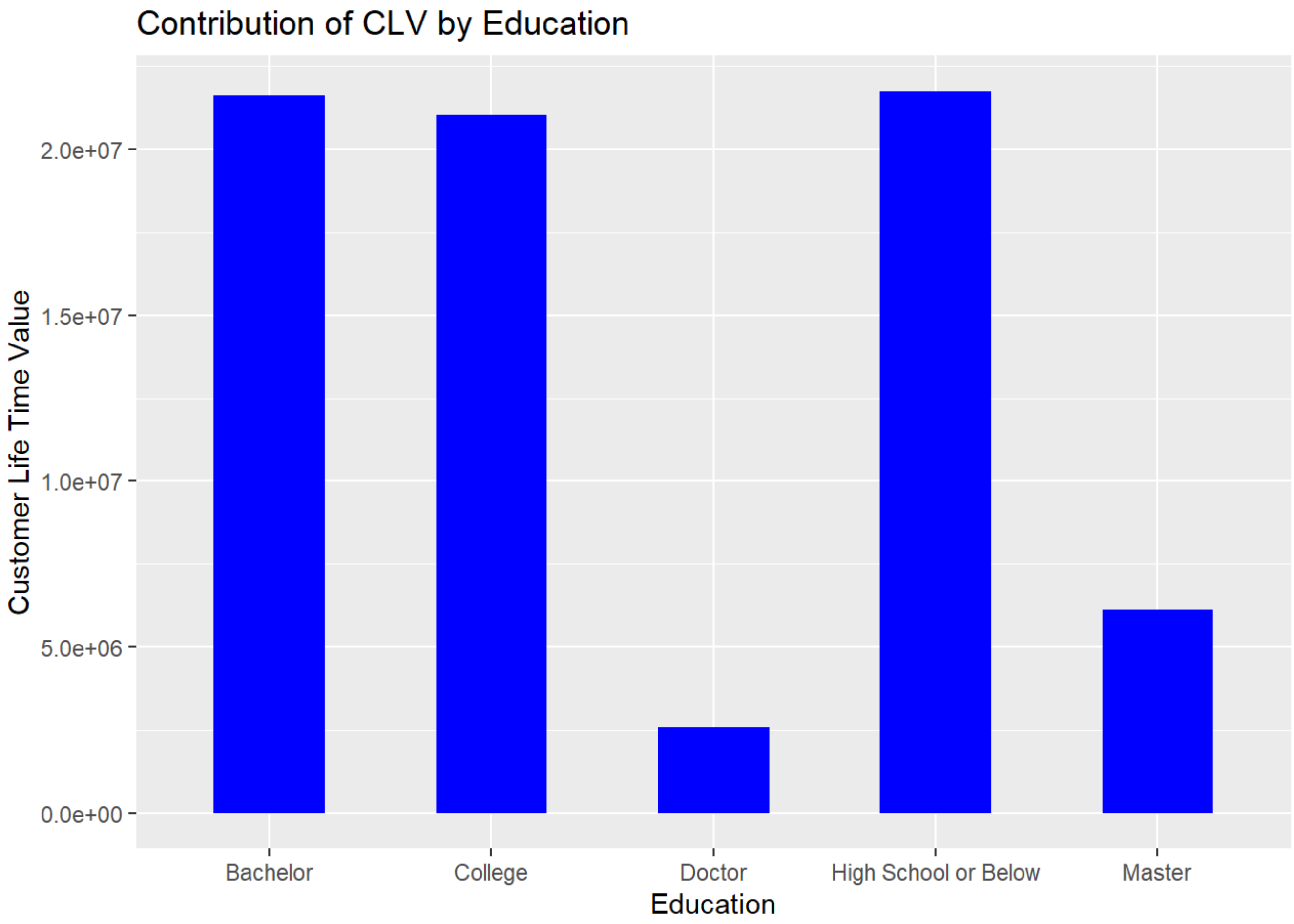
```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(State = Insurance_Dataset$State),
                     FUN = mean)
ggplot(data = aggData, aes(x = State, y = prop.table(stat(aggData$x)), fill = State, label = scales::percent(prop.table(stat
(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'State', y = 'CLV in Percentage', fill = 'State') +
  ggtitle("Mean contribution to CLV by State")
```



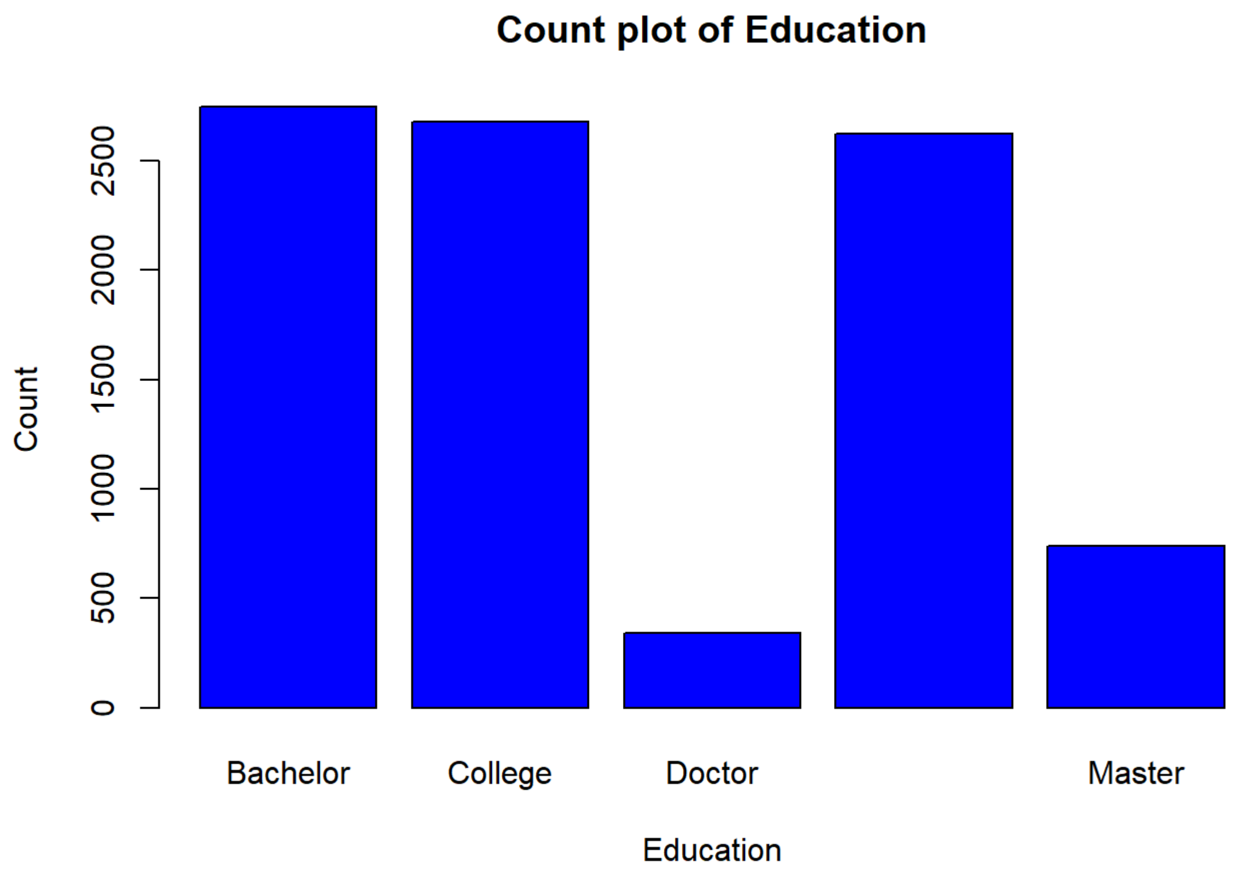
When the mean of the CLV is computed we can see that no particular state is any more economically valuable than the other, as the customers from each state on an average contribute equally to the target variable(CLV). This tells us that state is a weak indicator variable for the CLV.

## To visualize the effect of Education on CLV

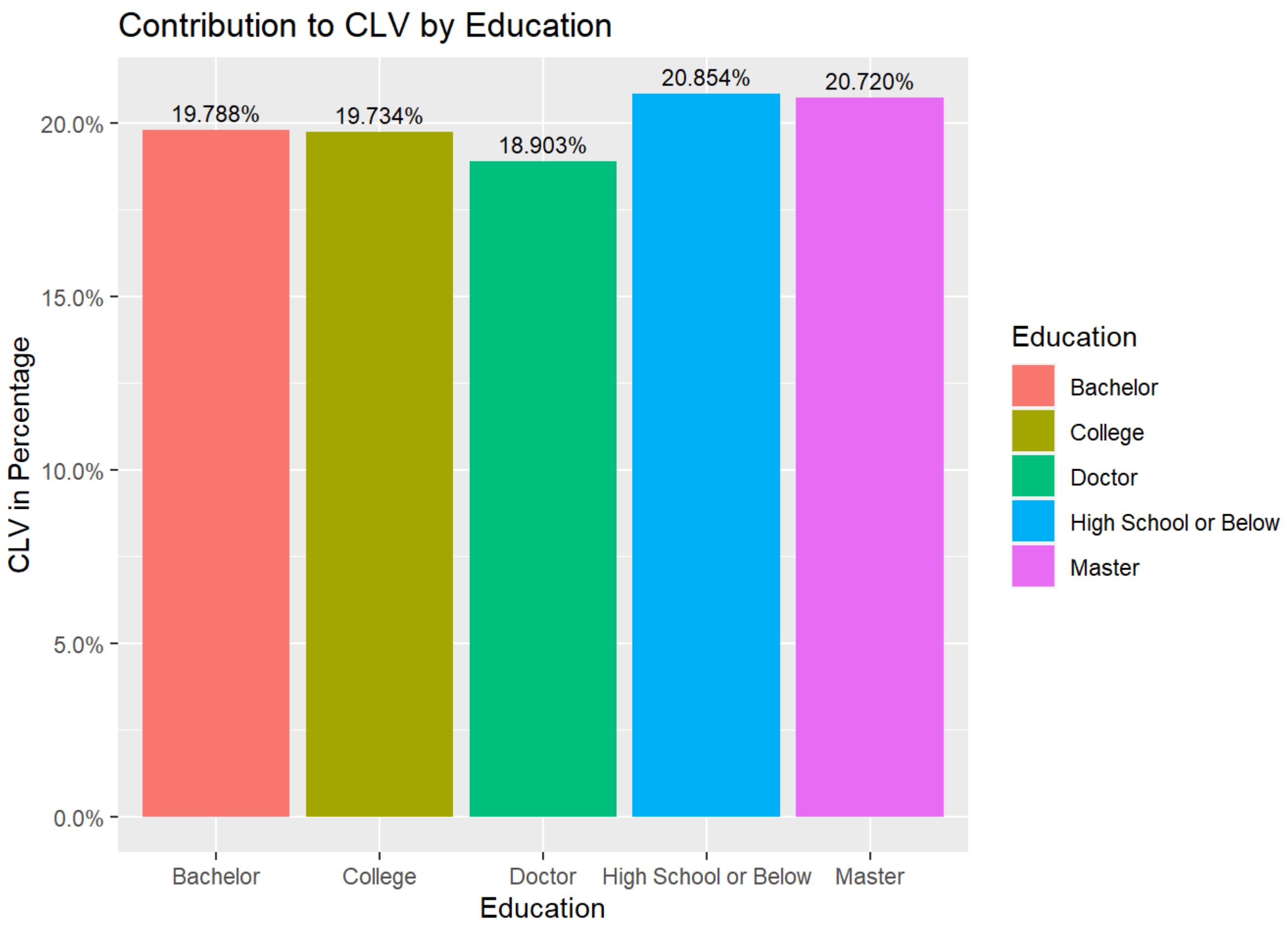
```
ggplot(Insurance_Dataset,aes (x=Education ,
              y=Customer.Lifetime.Value)) + geom_bar(stat="summary",fun="sum", width=0.5, fill = "Blue")+
  labs(x="Education",y = "Customer Life Time Value", fill="Education") +
  ggtitle("Contribution of CLV by Education")
```



```
count_education <- table(Insurance_Dataset$Education)
barplot(count_education,
        main = "Count plot of Education",col = "Blue",
        xlab = "Education", ylab = "Count")
```



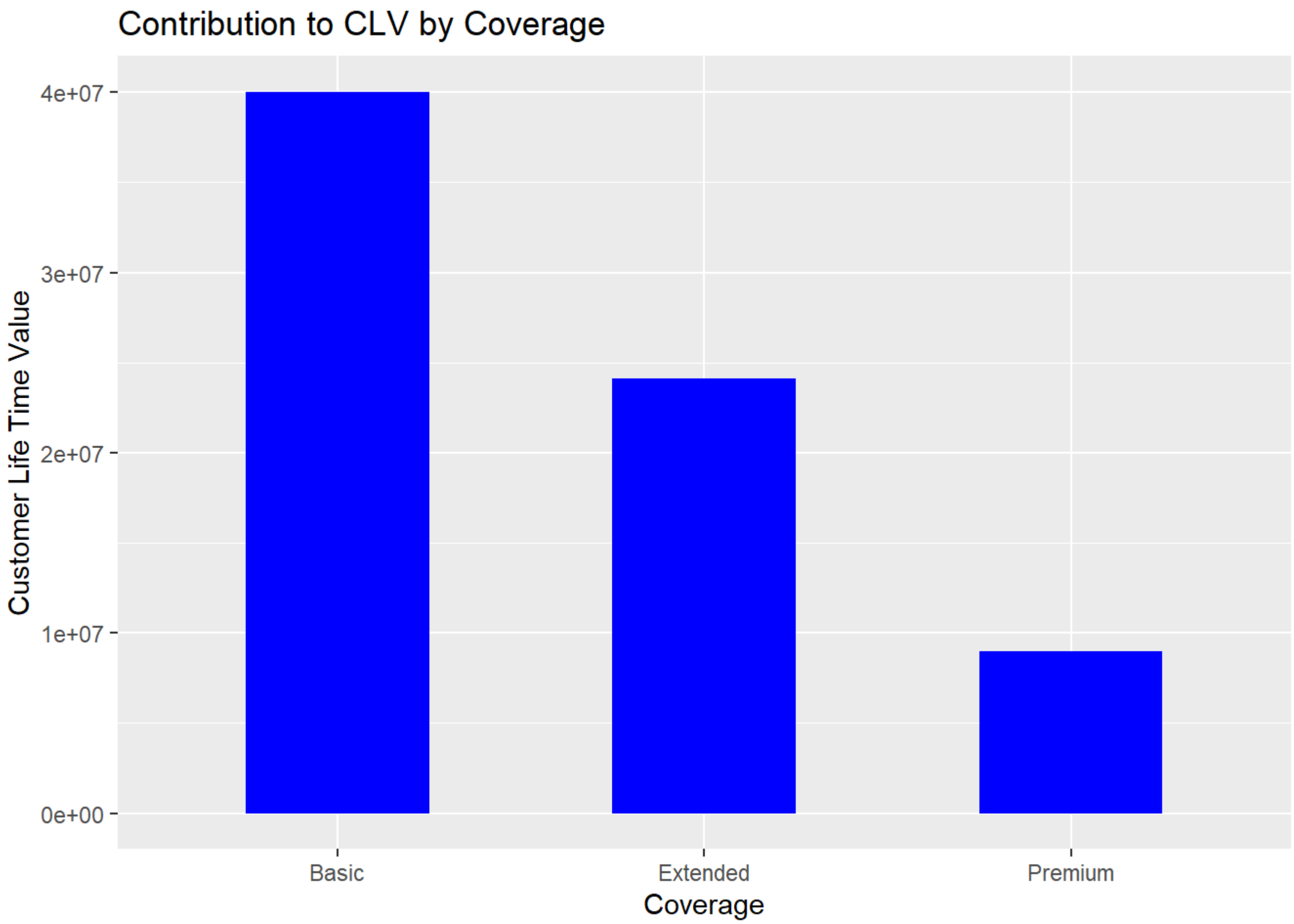
```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(Education = Insurance_Dataset$Education),
                     FUN = mean)
ggplot(data = aggData, aes(x = Education, y = prop.table(stat(aggData$x)), fill = Education, label = scales::percent(prop.ta
ble(stat(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'Education', y = 'CLV in Percentage', fill = 'Education') +
  ggtitle("Contribution to CLV by Education")
```



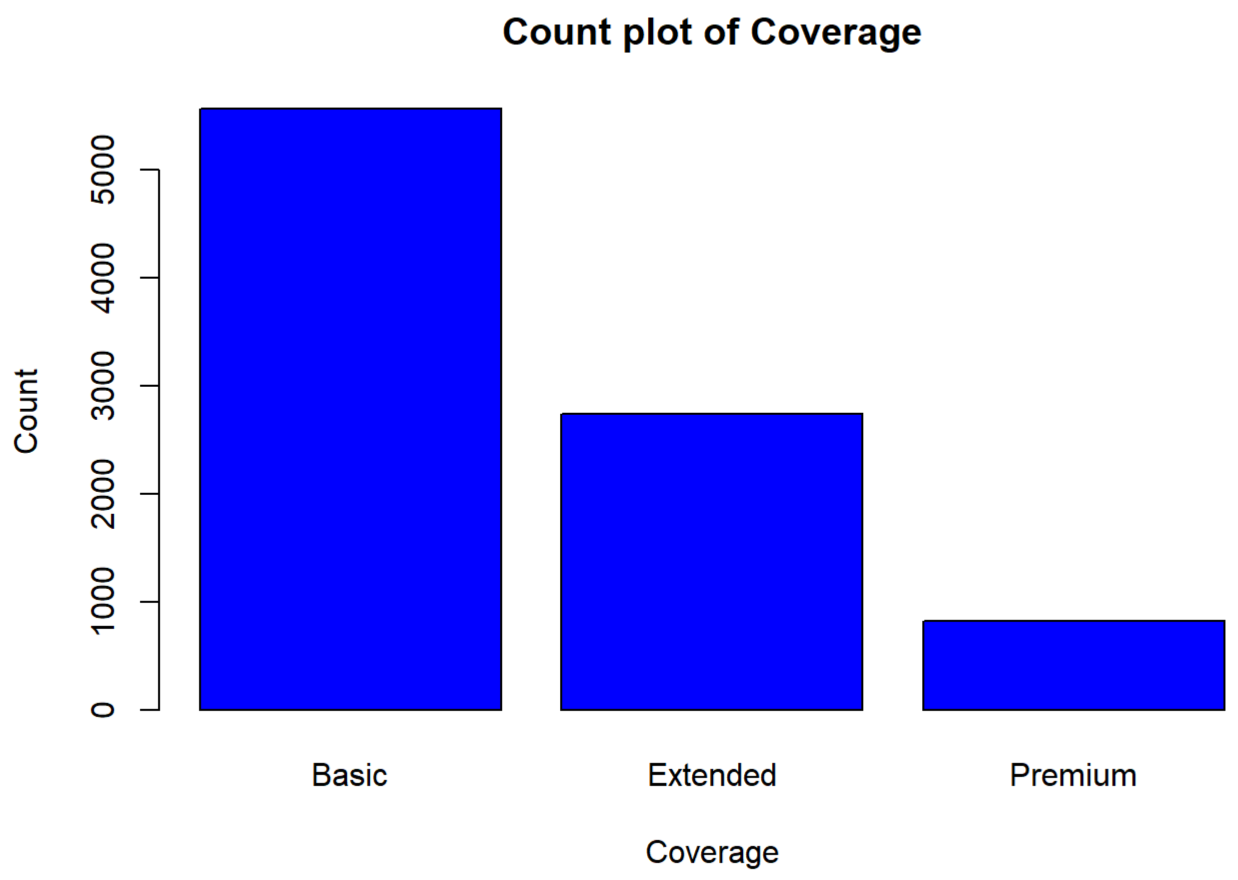
In the first plot it appears as though the contribution of customers having doctors and Master’s qualification is much lesser compared to the customers with other qualifications, which is counter-intuitive. But if we factor in the count of customers from different levels of education, we notice that since there are more customers having Bachelor’s, College and High level qualification, the contribution from these categories is more, as shown in the second chart. This point is further supported by the next chart where the average contribution of each class of qualification is almost the same, which indicates that the value of insurance policies purchased by the customers having doctors and Master’s qualification is much higher than the customers having other qualifications.

## To visualize the effect of Coverage on CLV

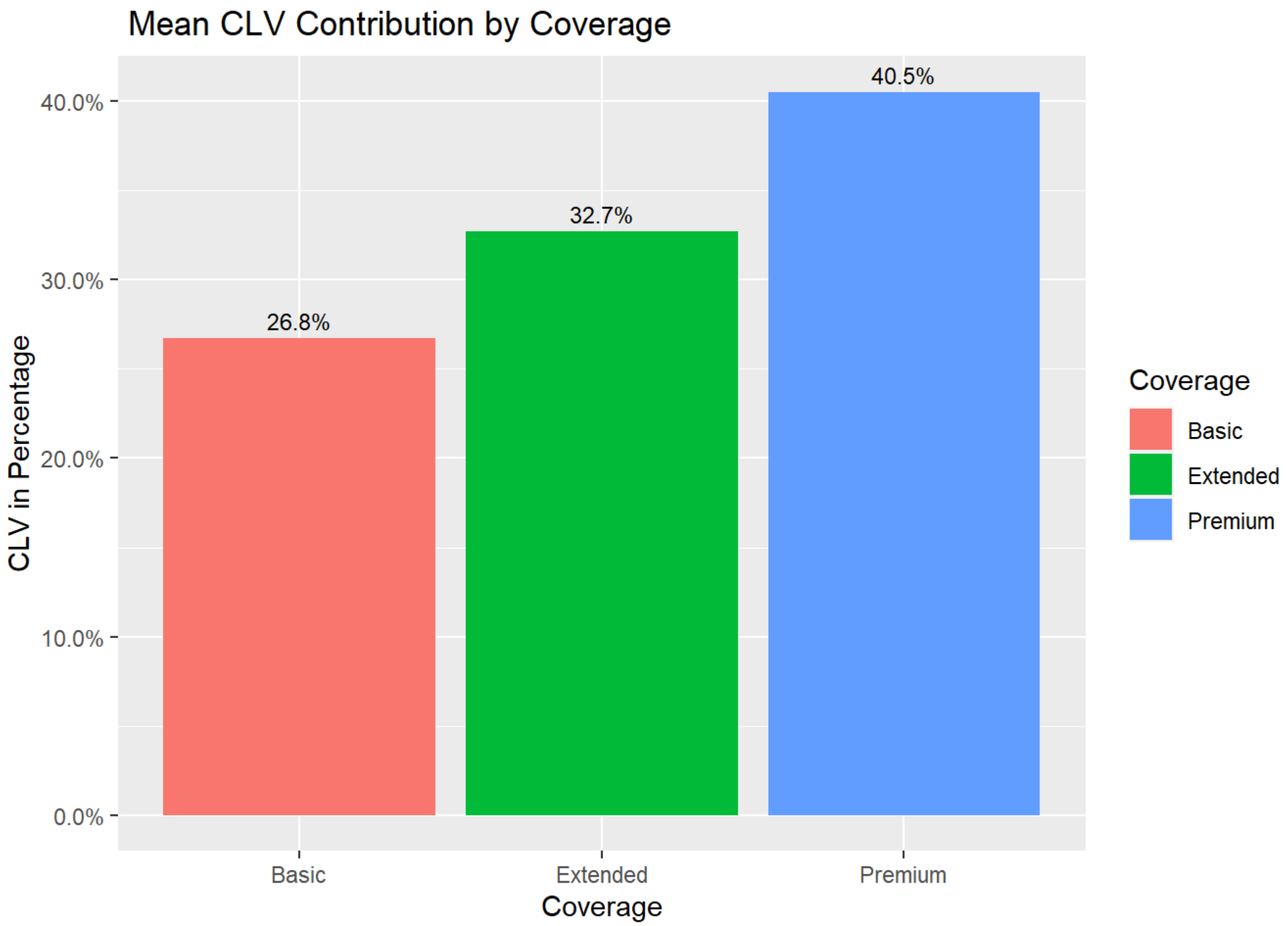
```
ggplot(Insurance_Dataset,aes (x=Coverage ,
              y=Customer.Lifetime.Value)) + geom_bar(stat="summary",fun="sum", width=0.5, fill = "Blue")+
  labs(x="Coverage",y = "Customer Life Time Value", fill="Coverage") +
  ggtitle("Contribution to CLV by Coverage")
```


Contribution to CLV by Coverage

```
count_coverage <- table(Insurance_Dataset$Coverage)
barplot(count_coverage,
        main = "Count plot of Coverage",col = "Blue",
        xlab = "Coverage", ylab = "Count")
```



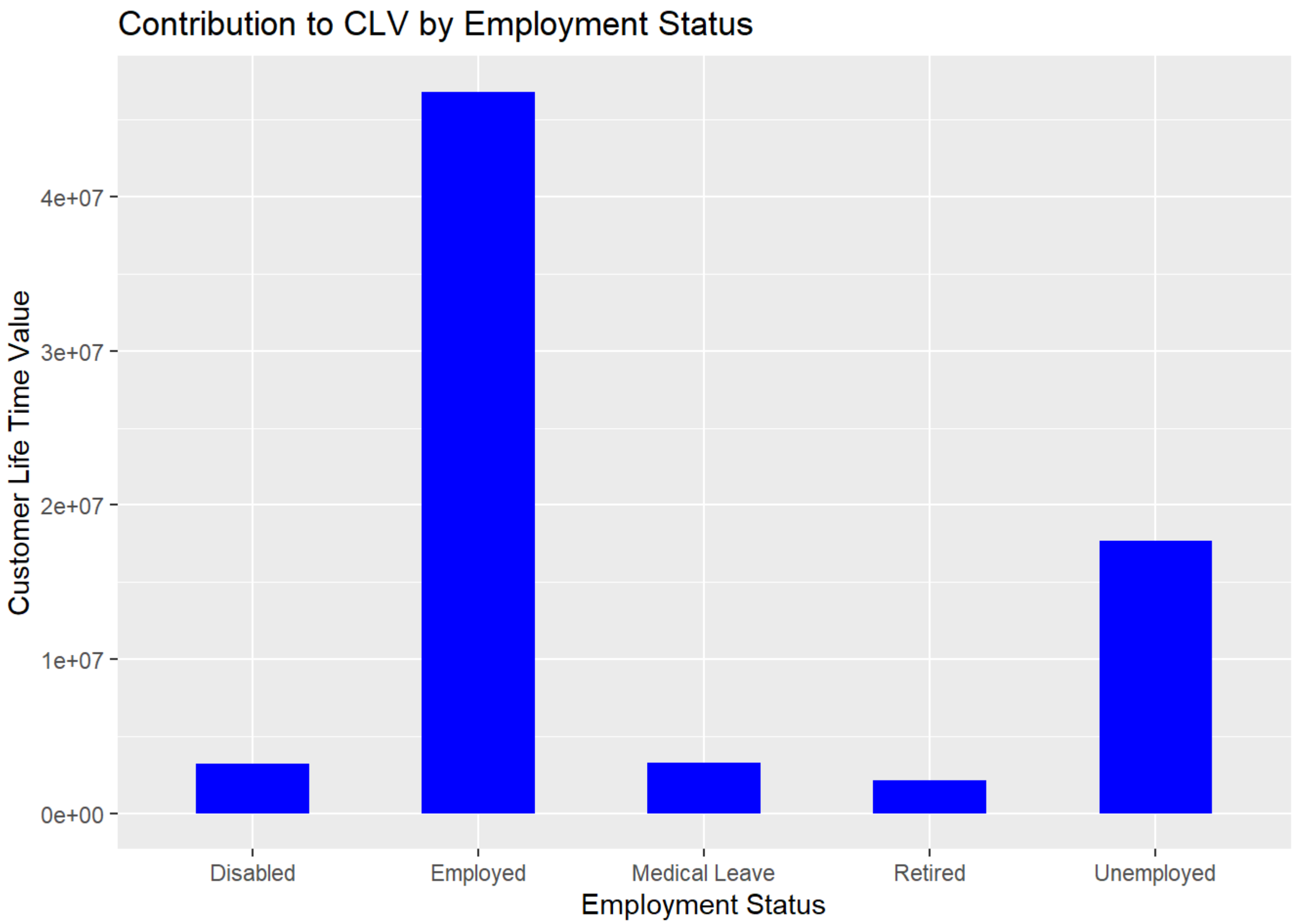
```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(Coverage = Insurance_Dataset$Coverage),
                     FUN = mean)
ggplot(data = aggData, aes(x = Coverage, y = prop.table(stat(aggData$x)), fill = Coverage, label = scales::percent(prop.table(stat(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'Coverage', y = 'CLV in Percentage', fill = 'Coverage') +
  ggtitle(" Mean CLV Contribution by Coverage")
```



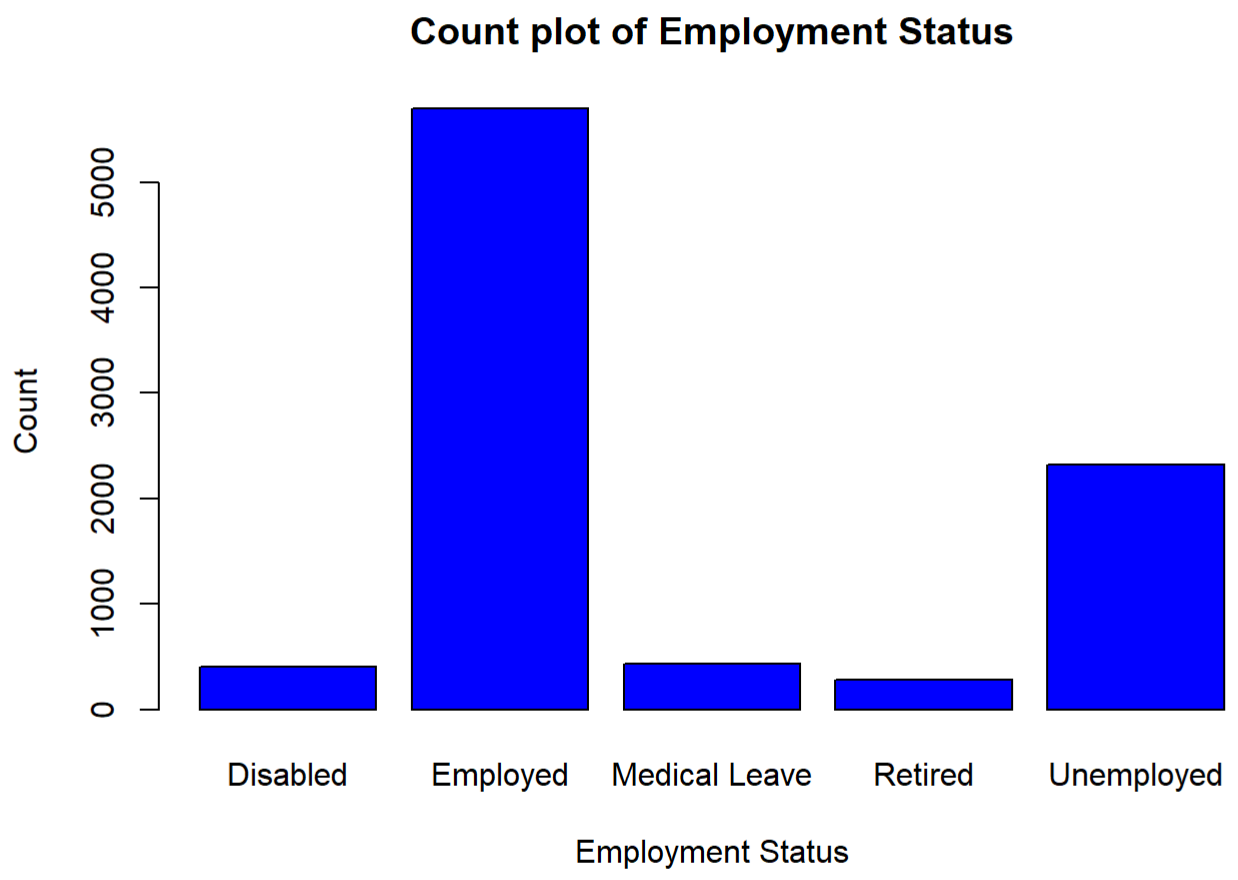
It would be apparent from the first chart that the Basic coverage plan has the most contribution to CLV because there are more takers for the basic coverage plan, as proven by the second chart. In the third chart, however, we can see that even though premium coverage plans accounted for the least volume of CLV, on an average a customer having the premium coverage has a greater contribution to CLV.

## To visualize the effect of Employment Status on CLV

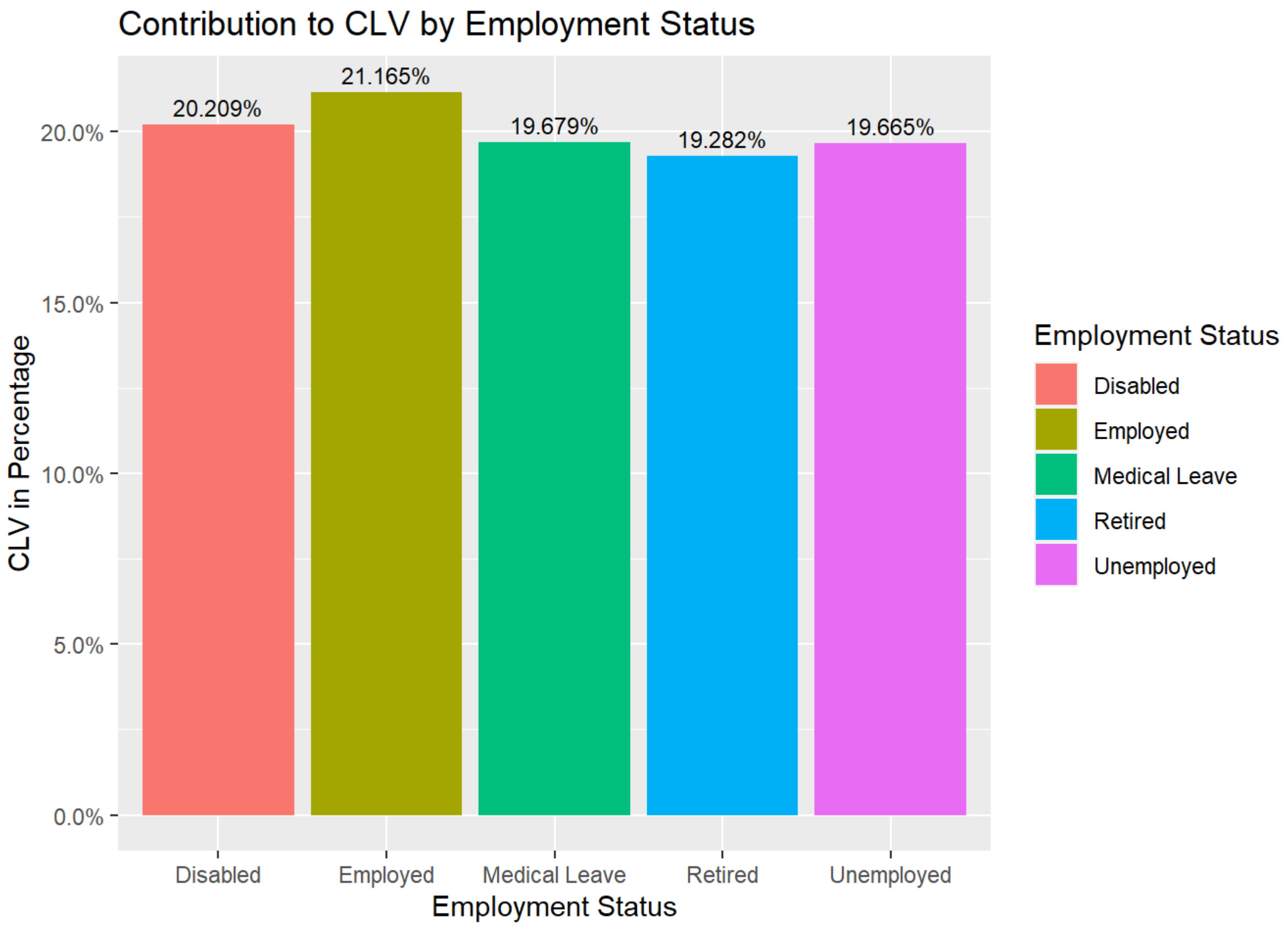
```
ggplot(Insurance_Dataset,aes (x=EmploymentStatus ,
              y=Customer.Lifetime.Value)) + geom_bar(stat="summary",fun="sum", width=0.5, fill = "Blue")+
  labs(x="Employment Status",y = "Customer Life Time Value", fill="Employment Status") +
  ggtitle("Contribution to CLV by Employment Status")
```



```
count_employmentstatus <- table(Insurance_Dataset$EmploymentStatus)
barplot(count_employmentstatus,
        main = "Count plot of Employment Status",col = "Blue",
        xlab = "Employment Status", ylab = "Count")
```



```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(EmploymentStatus = Insurance_Dataset$EmploymentStatus),
                     FUN = mean)
ggplot(data = aggData, aes(x = EmploymentStatus, y = prop.table(stat(aggData$x)), fill = EmploymentStatus, label = scales::p
ercent(prop.table(stat(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'Employment Status', y = 'CLV in Percentage', fill = 'Employment Status') +
  ggtitle("Contribution to CLV by Employment Status")
```

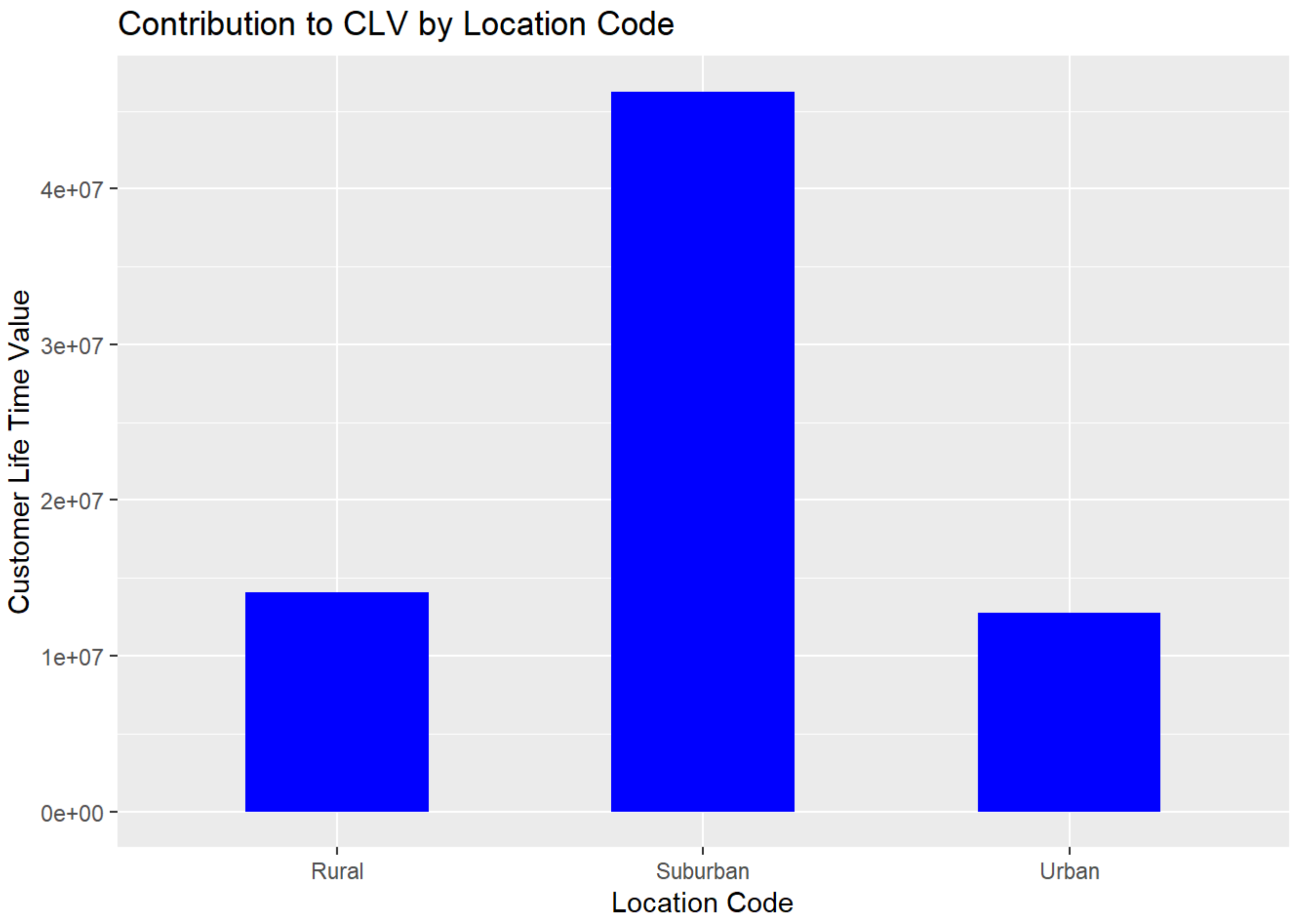


In the first chart, it is evident that the customers who are employed are of greater value to the company than the other categories. The inference drawn from this is straightforward, i.e employed customers are more likely to be able to afford the premiums and therefore contribute a major chunk to the CLV.

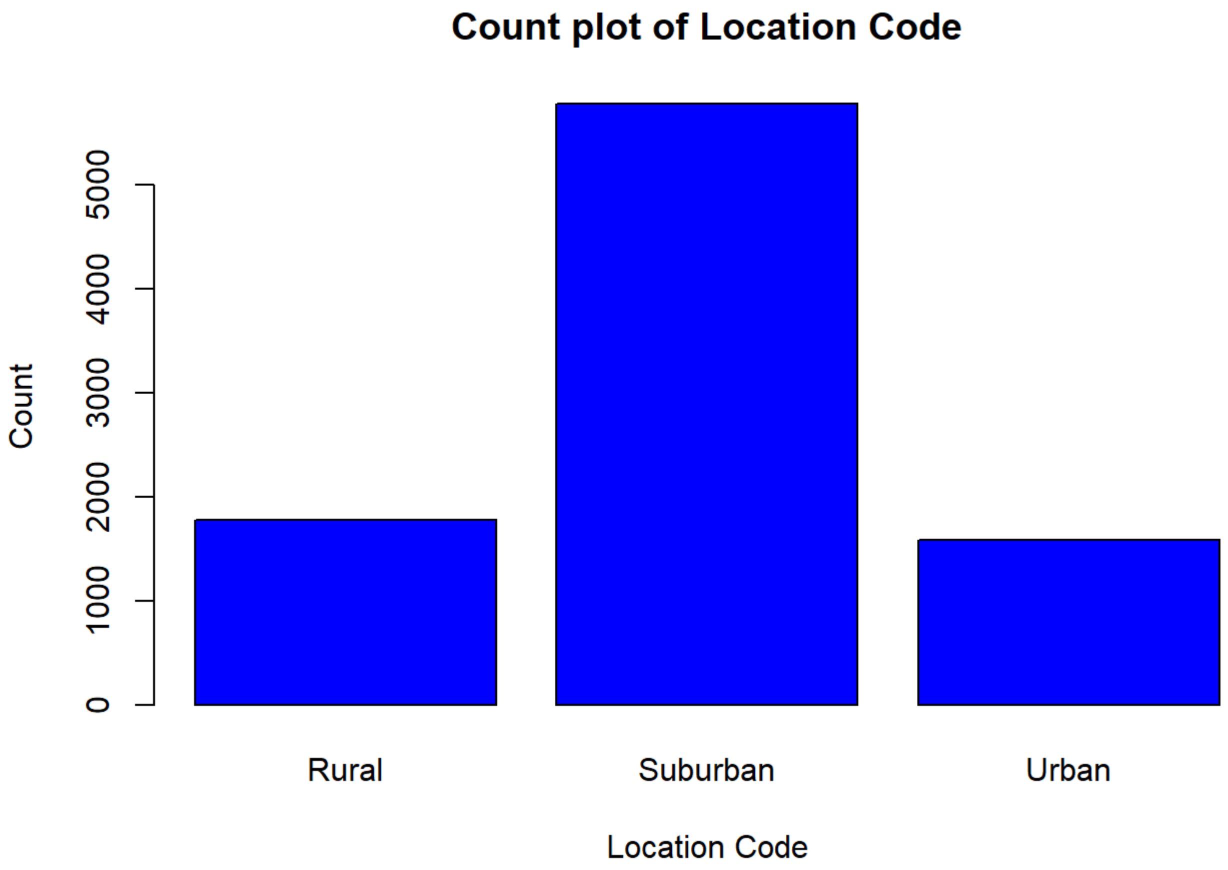
But in the second chart, when we account for the contribution on an average by the employment status, we notice that all are equally contributing to CLV. The reason why the the employed status has such a high contribution to the CLV is because the number of customers who are employed is high.

## To visualize the effect of Location Code on CLV.

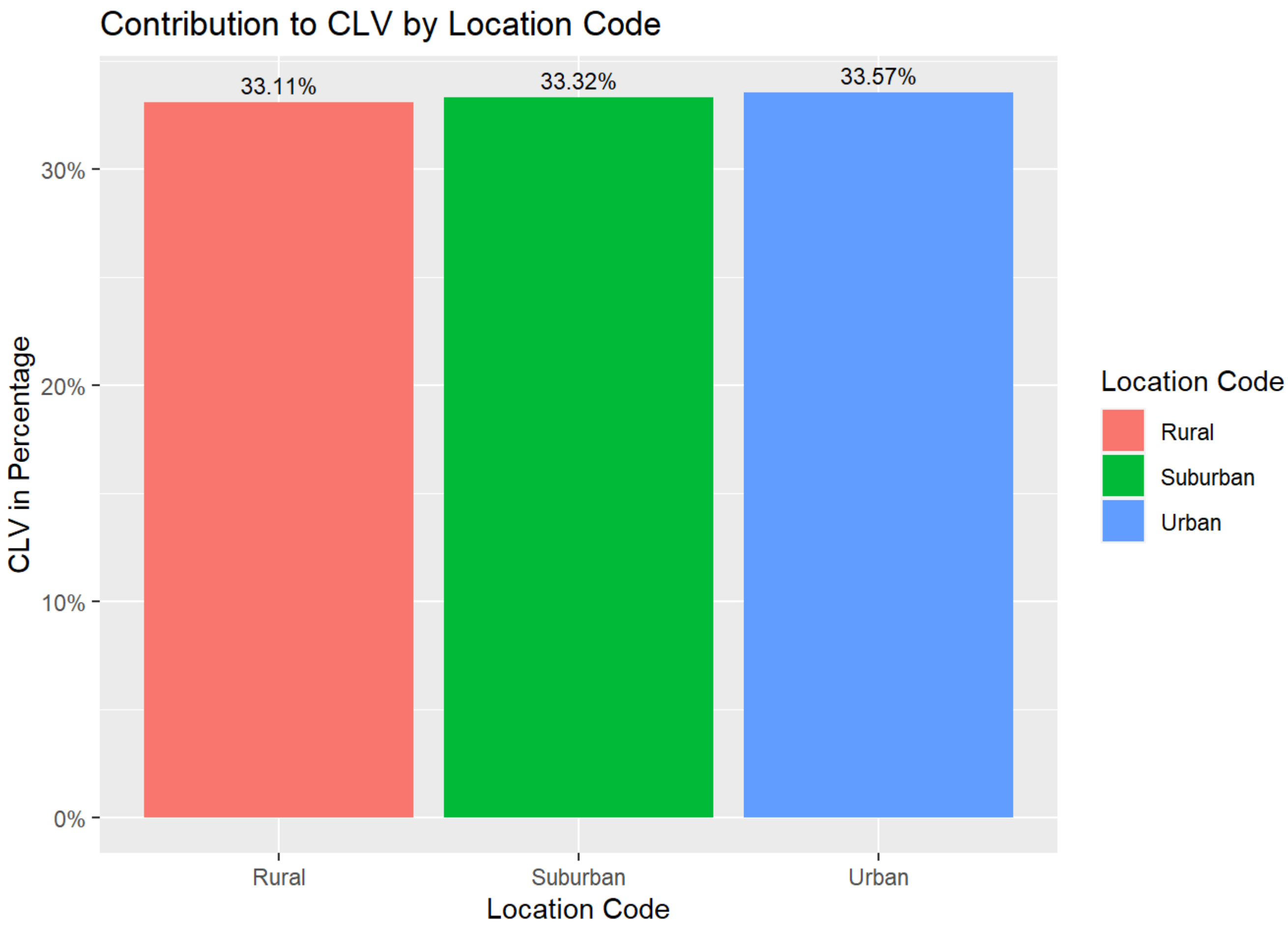
```
ggplot(Insurance_Dataset,aes (x=Location.Code ,
              y=Customer.Lifetime.Value)) + geom_bar(stat="summary",fun="sum", width=0.5, fill = "Blue")+
  labs(x="Location Code",y = "Customer Life Time Value", fill="Location Code") +
  ggtitle("Contribution to CLV by Location Code")
```



```
count_locationcode <- table(Insurance_Dataset$Location.Code)
barplot(count_locationcode,
        main = "Count plot of Location Code",col = "Blue",
        xlab = "Location Code", ylab = "Count")
```



```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(Location.Code = Insurance_Dataset$Location.Code),
                     FUN = mean)
ggplot(data = aggData, aes(x = Location.Code, y = prop.table(stat(aggData$x)), fill = Location.Code, label = scales::percent
(prop.table(stat(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'Location Code', y = 'CLV in Percentage', fill = 'Location Code') +
  ggtitle("Contribution to CLV by Location Code")
```

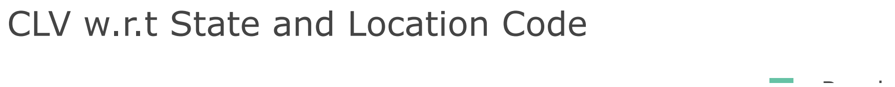


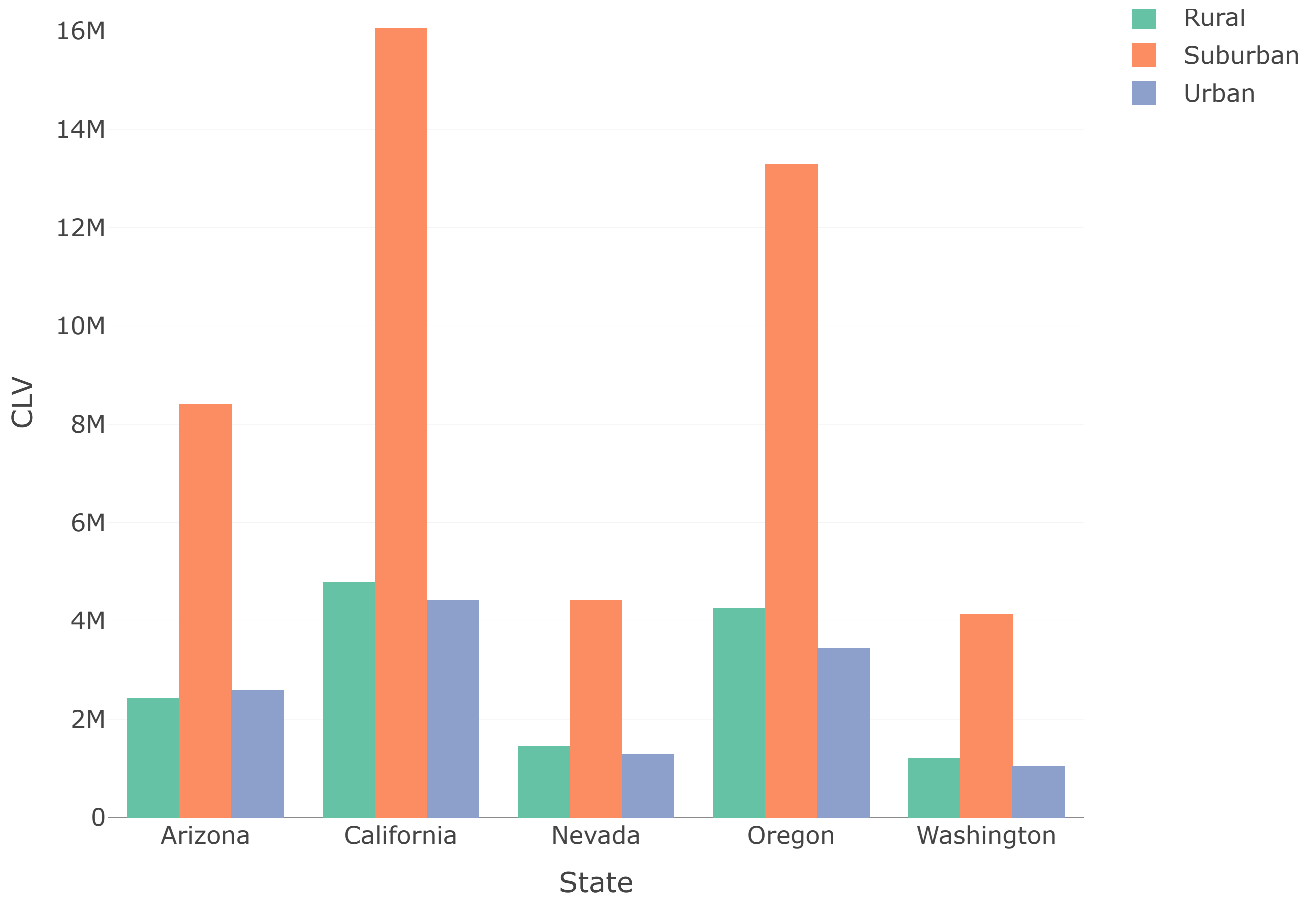
In the first chart it appears as though customers from the suburban location are a better contributor to CLV than the other areas. From the second chart it is clear that it is because of the higher number of subscribers from suburban areas.

But from the third, we see that all of the location codes on an average contribute equally to the CLV and therefore Location Code is a weak predictor of the CLV on its own.

## Effect on CLV by State and Location Code

```
p1<-plot_ly(Insurance_Dataset, x =~State, y =~Insurance_Dataset$`Customer.Lifetime.Value`,type='bar',color=~Insurance_Datase
t$`Location.Code`)
layout(p1, title ='CLV w.r.t State and Location Code', yaxis = list(title = 'CLV '))
```

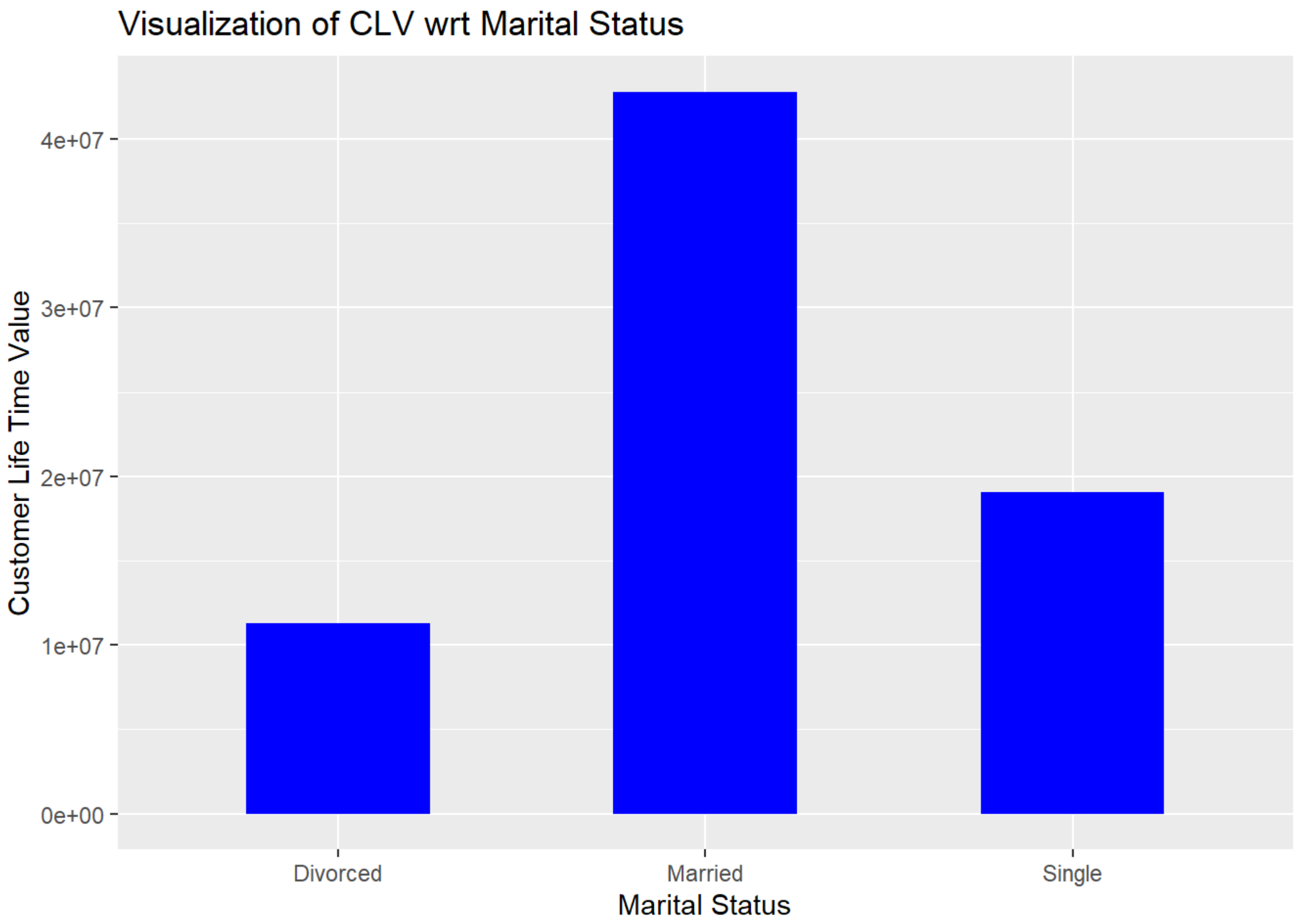




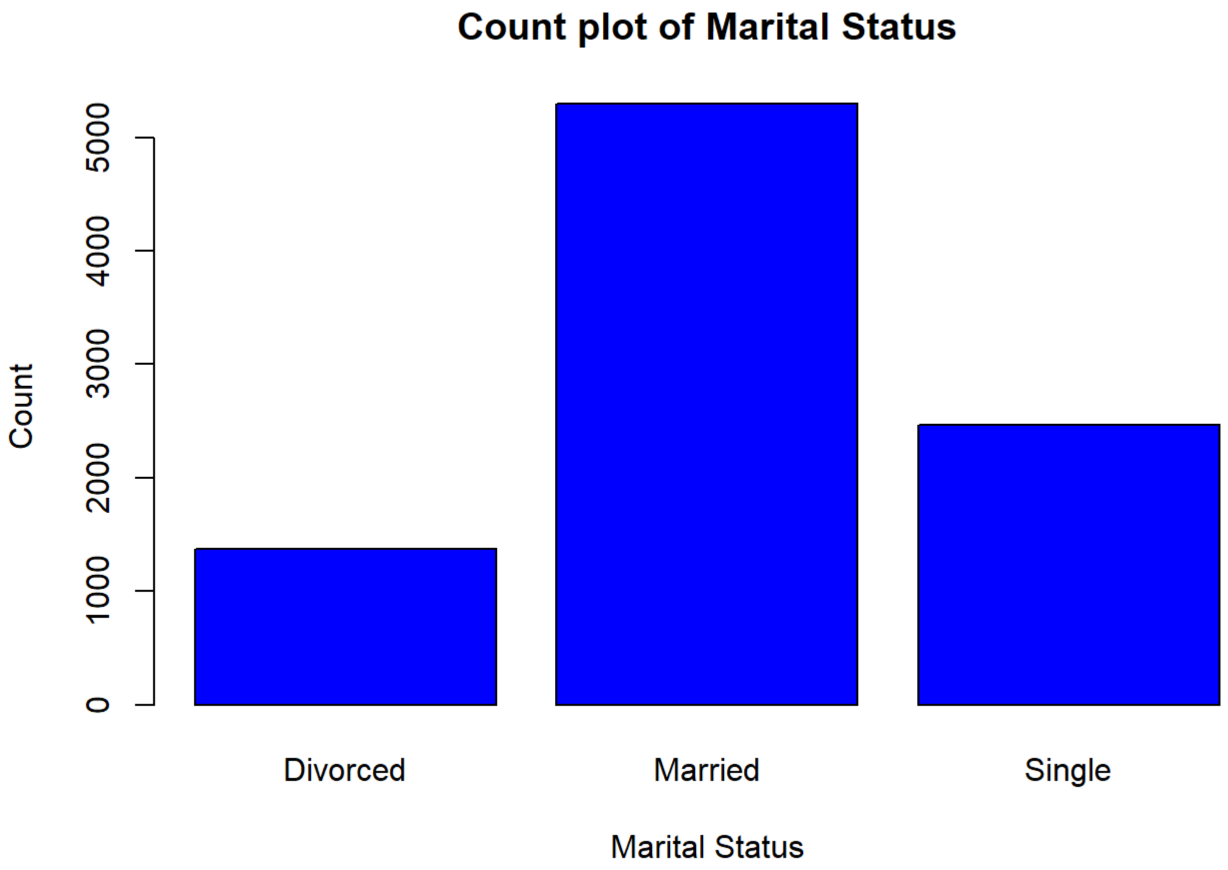
California and Oregon outperform the other states in every location code with regard to CLV.

## To visualize the effect of Marital Status on CLV

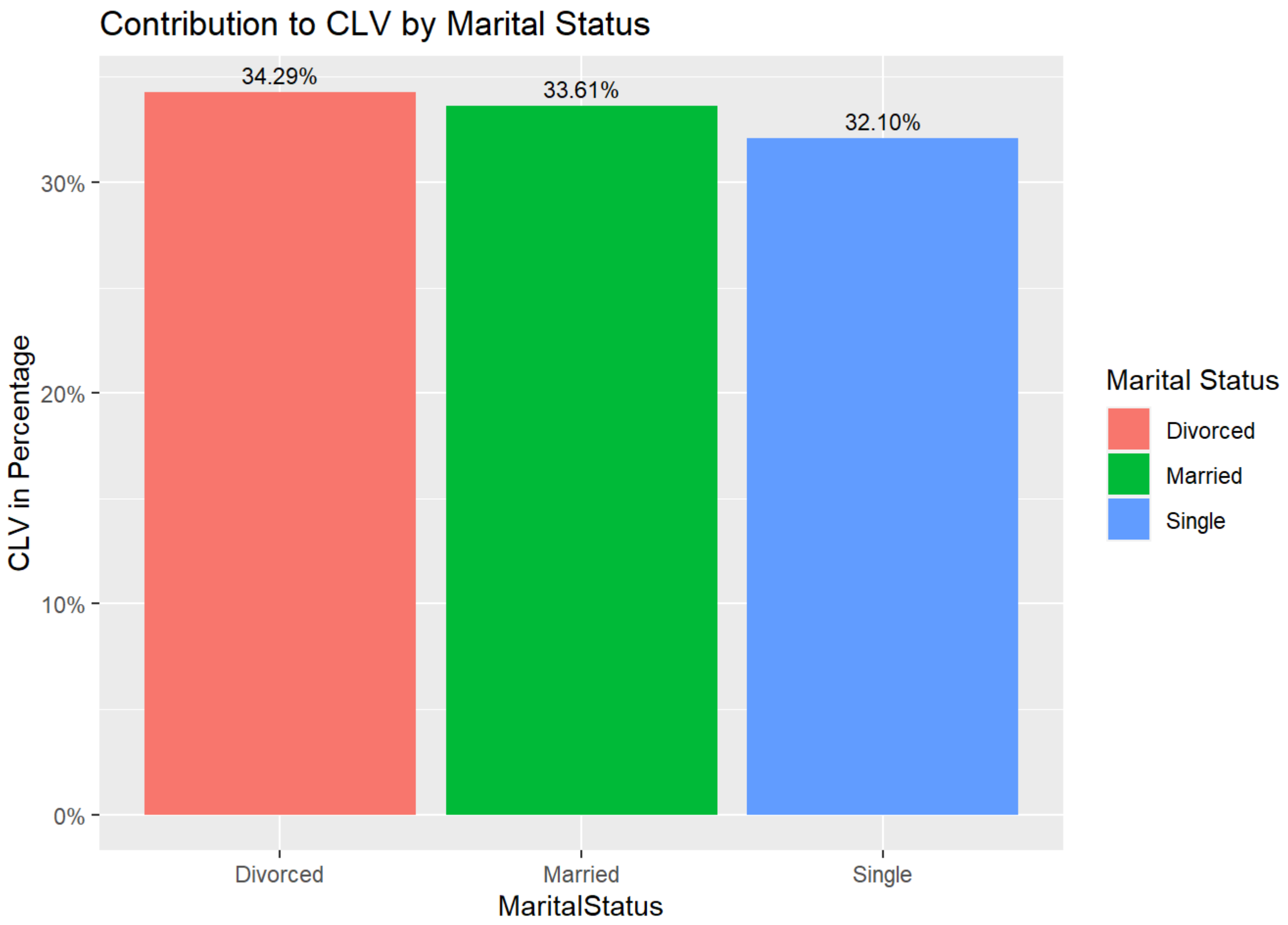
```
ggplot(Insurance_Dataset,aes (x=Insurance_Dataset$"Marital.Status", y=Insurance_Dataset$"Customer.Lifetime.Value")) + geom_b
ar(stat="summary",fun="sum", width=0.5, fill = "Blue") +
  labs(x="Marital Status",y = "Customer Life Time Value", fill="Marital Status") +
  ggtitle("Visualization of CLV wrt Marital Status")
```



```
count_maritalstatus <- table(Insurance_Dataset$Marital.Status)
barplot(count_maritalstatus,
        main = "Count plot of Marital Status",col = "Blue",
        xlab = "Marital Status", ylab = "Count")
```



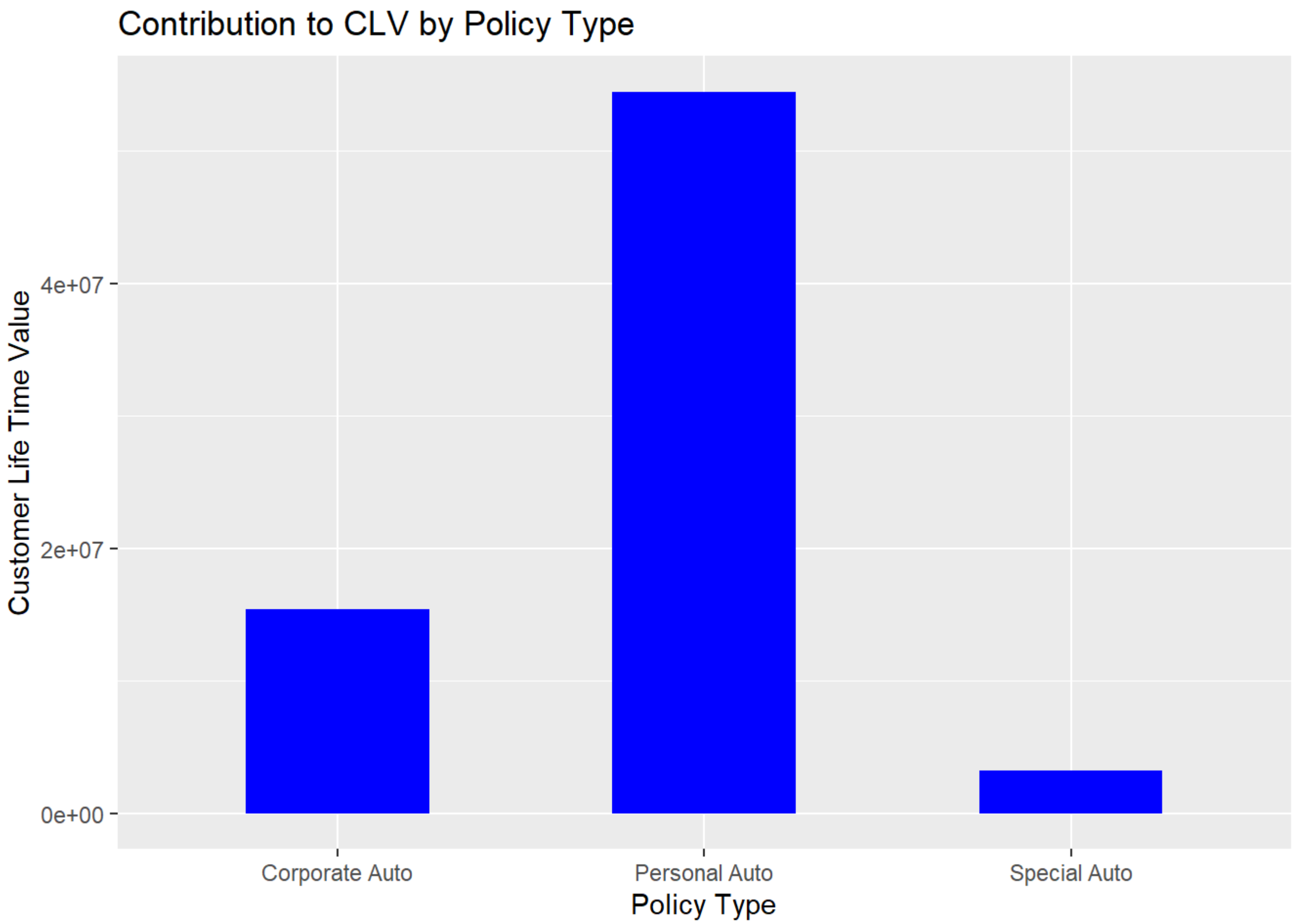
```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(Marital.Status = Insurance_Dataset$Marital.Status),
                     FUN = mean)
ggplot(data = aggData, aes(x = Marital.Status, y = prop.table(stat(aggData$x)), fill = Marital.Status, label = scales::perce
nt(prop.table(stat(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'MaritalStatus', y = 'CLV in Percentage', fill = 'Marital Status') +
  ggtitle("Contribution to CLV by Marital Status")
```



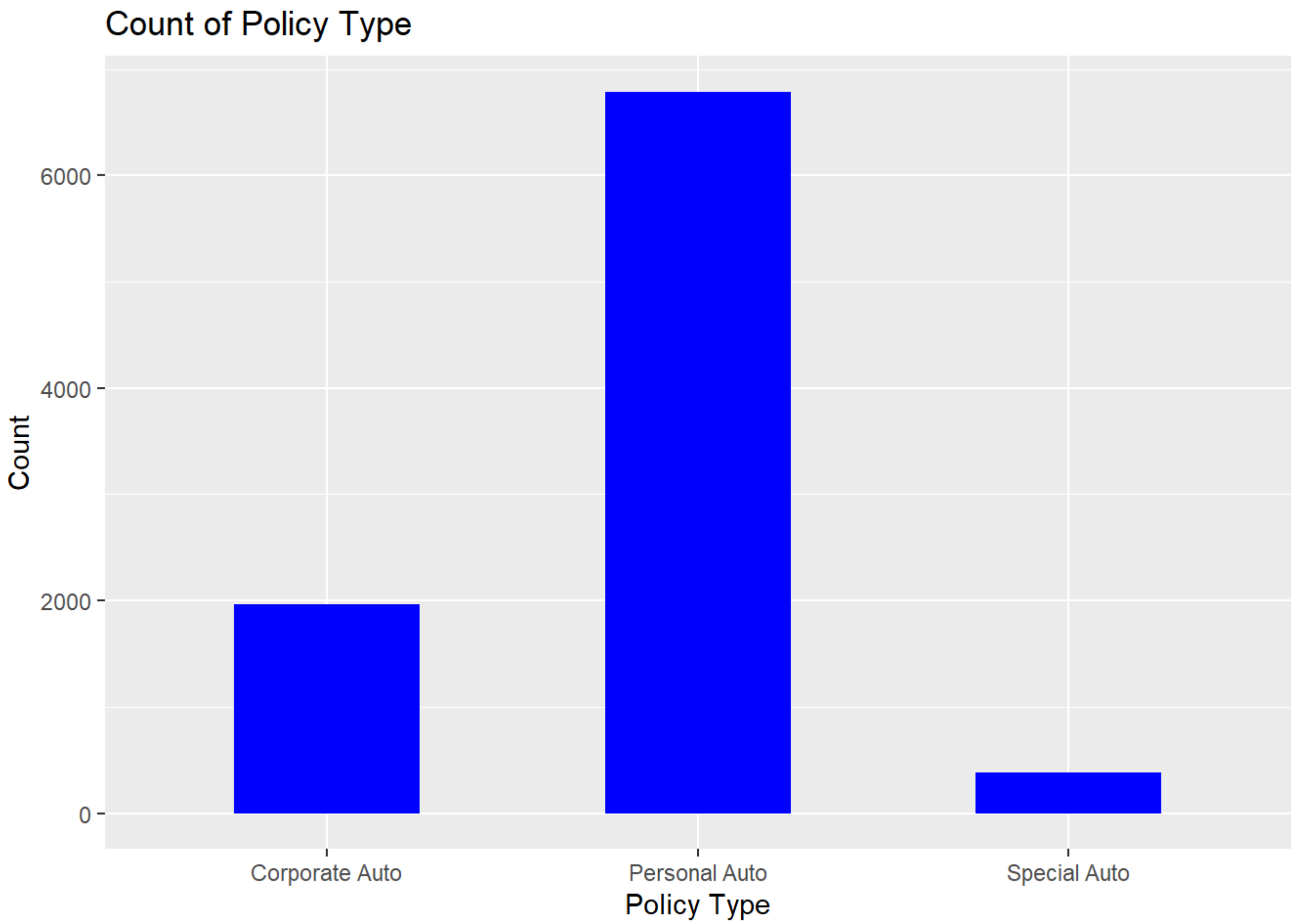
We might erroneously conclude from the first and the second chart that most married customers have high CLV but the third chart shows us that on an average there is no difference between the contributions of each sub-category to the CLV.

## To visualize the effect of Policy Type on CLV

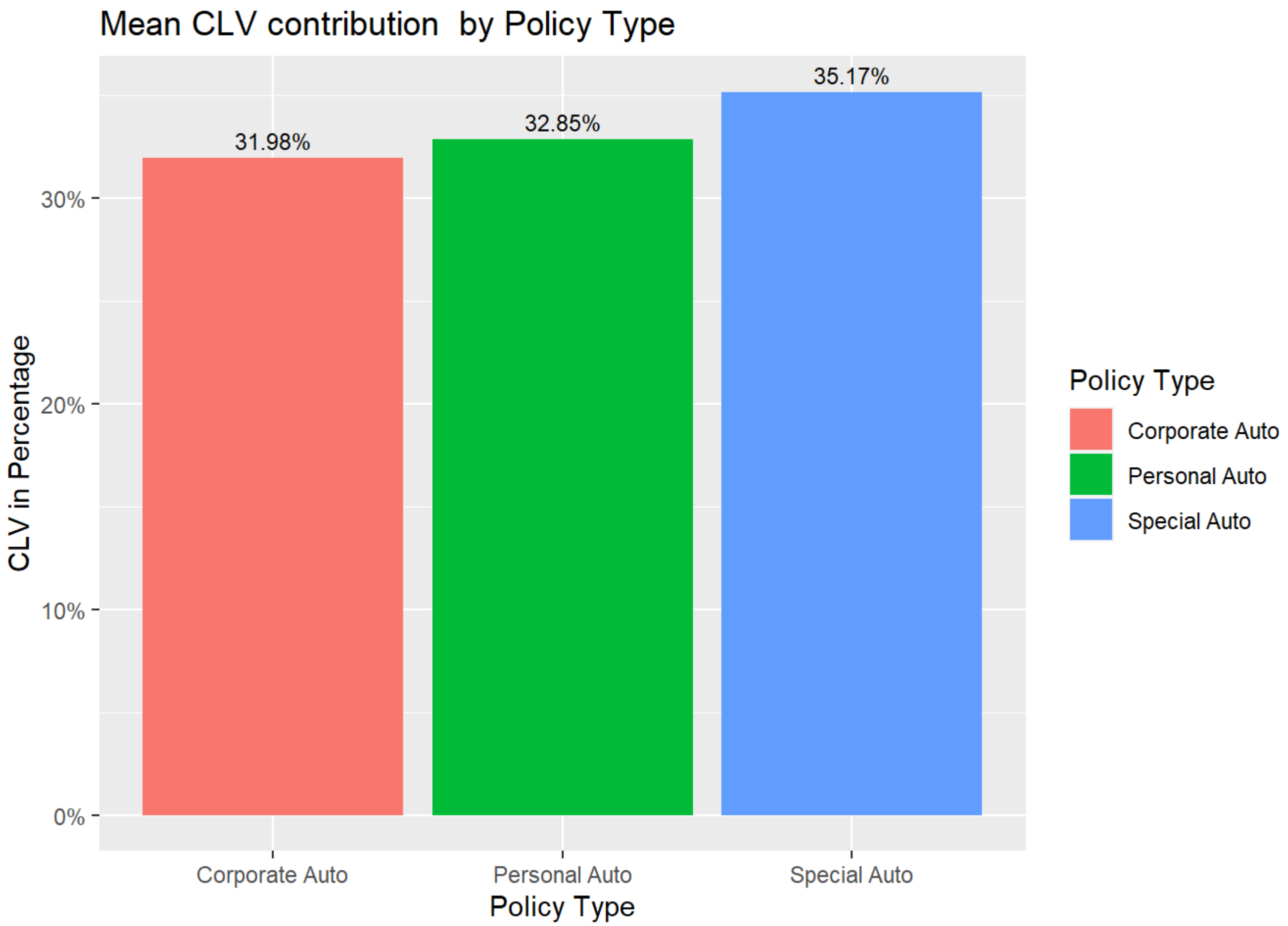
```
ggplot(Insurance_Dataset,aes (x=Policy.Type ,
              y=Customer.Lifetime.Value)) + geom_bar(stat="summary",fun="sum", width=0.5, fill = "Blue")+
  labs(x="Policy Type",y = "Customer Life Time Value", fill="Policy Type") +
  ggtitle("Contribution to CLV by Policy Type")
```



```
ggplot(Insurance_Dataset,aes (x=Policy.Type)) +
         geom_bar(stat="count", width=0.5, fill = "Blue") +
  labs(x="Policy Type",y = "Count", fill="Policy Type") +
  ggtitle("Count of Policy Type")
```



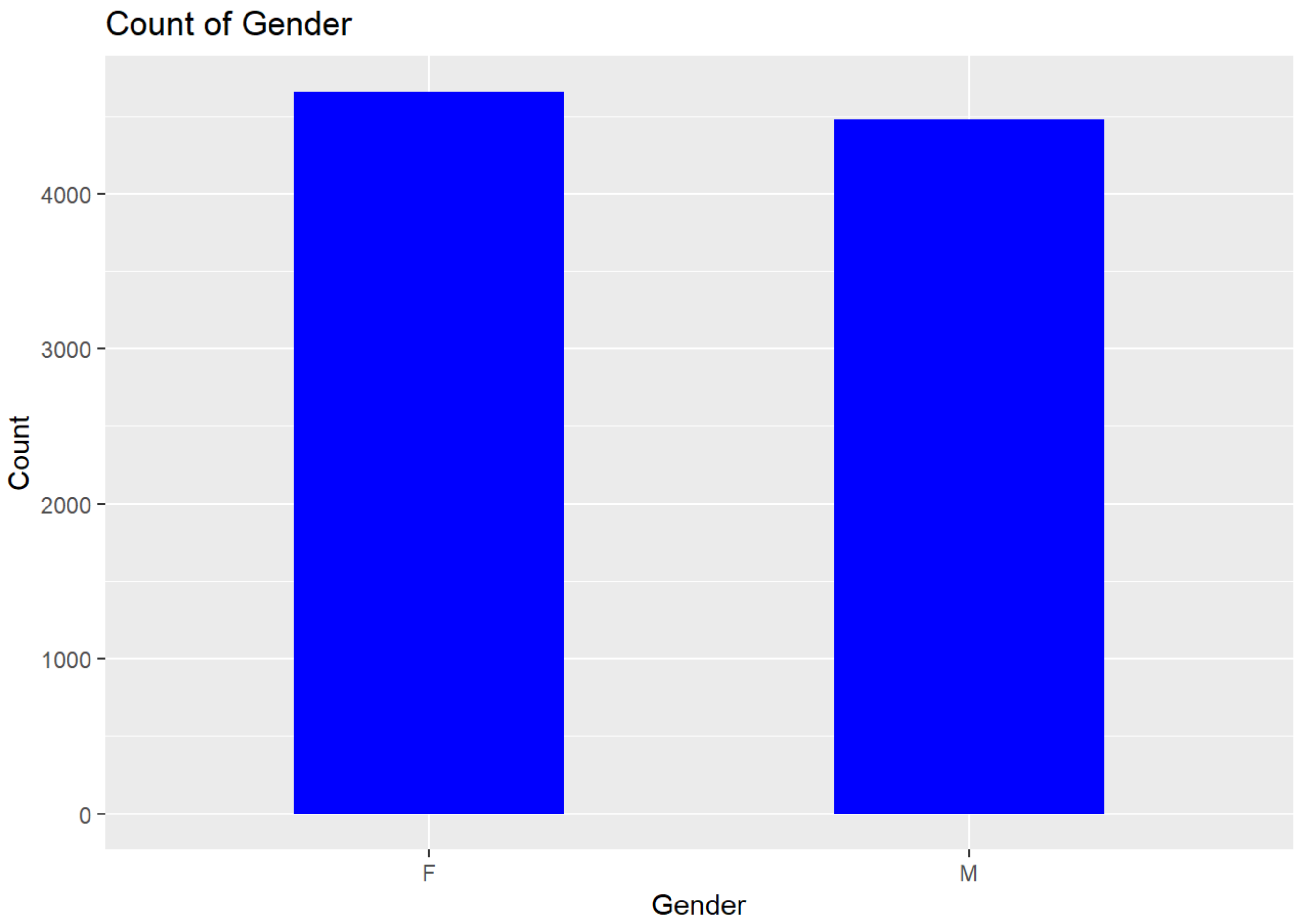
```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(Policy.Type = Insurance_Dataset$Policy.Type),
                     FUN = mean)
ggplot(data = aggData, aes(x = Policy.Type, y = prop.table(stat(aggData$x)), fill = Policy.Type, label = scales::percent(pro
p.table(stat(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'Policy Type', y = 'CLV in Percentage', fill = 'Policy Type') +
  ggtitle("Mean CLV contribution  by Policy Type")
```



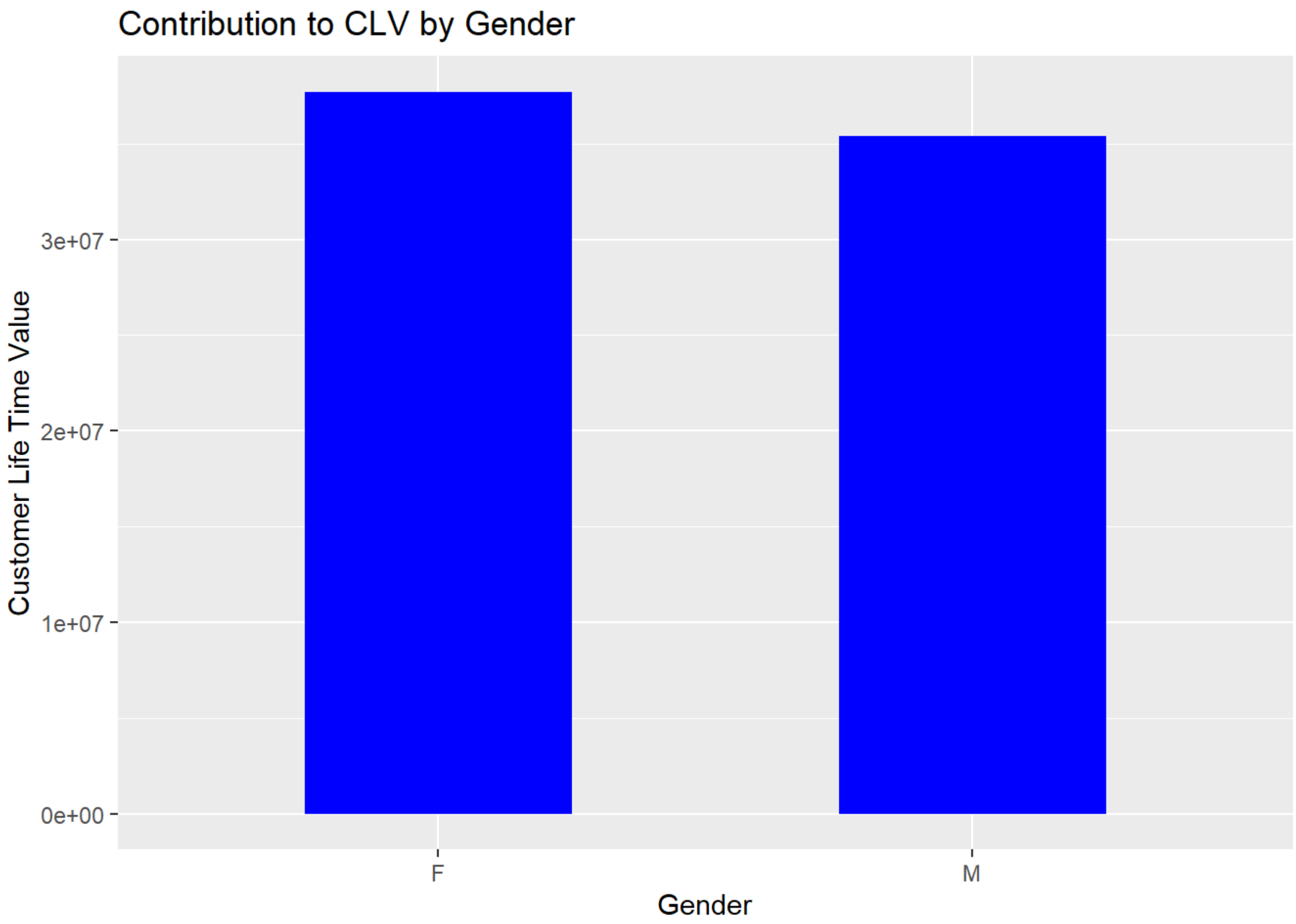
Similar results are obtained as before. Initially it may appear that the personal auto policy might be a majority contributor but further analysis shows that it seems more likely that customers who have purchased the Special Auto have a greater CLV.

## To visualize the effect of gender on CLV

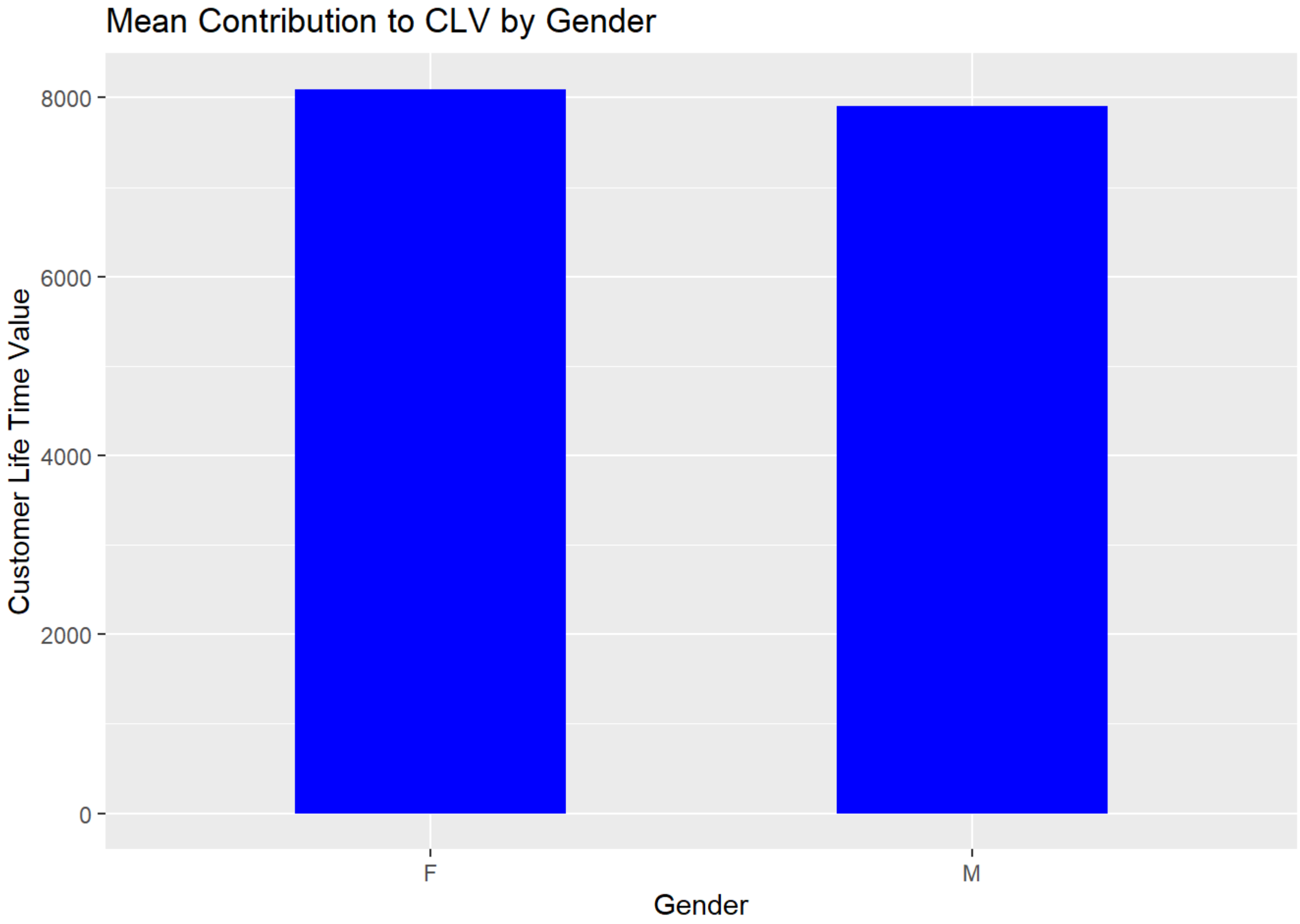
```
ggplot(Insurance_Dataset,aes (x=Gender)) +
         geom_bar(stat="count", width=0.5, fill = "Blue") +
  labs(x="Gender",y = "Count") +
  ggtitle("Count of Gender")
```



```
ggplot(Insurance_Dataset,aes (x=Gender ,
              y=Customer.Lifetime.Value)) + geom_bar(stat="summary",fun="sum", width=0.5, fill = "Blue")+
  labs(x="Gender",y = "Customer Life Time Value", fill="Gender") +
  ggtitle("Contribution to CLV by Gender")
```



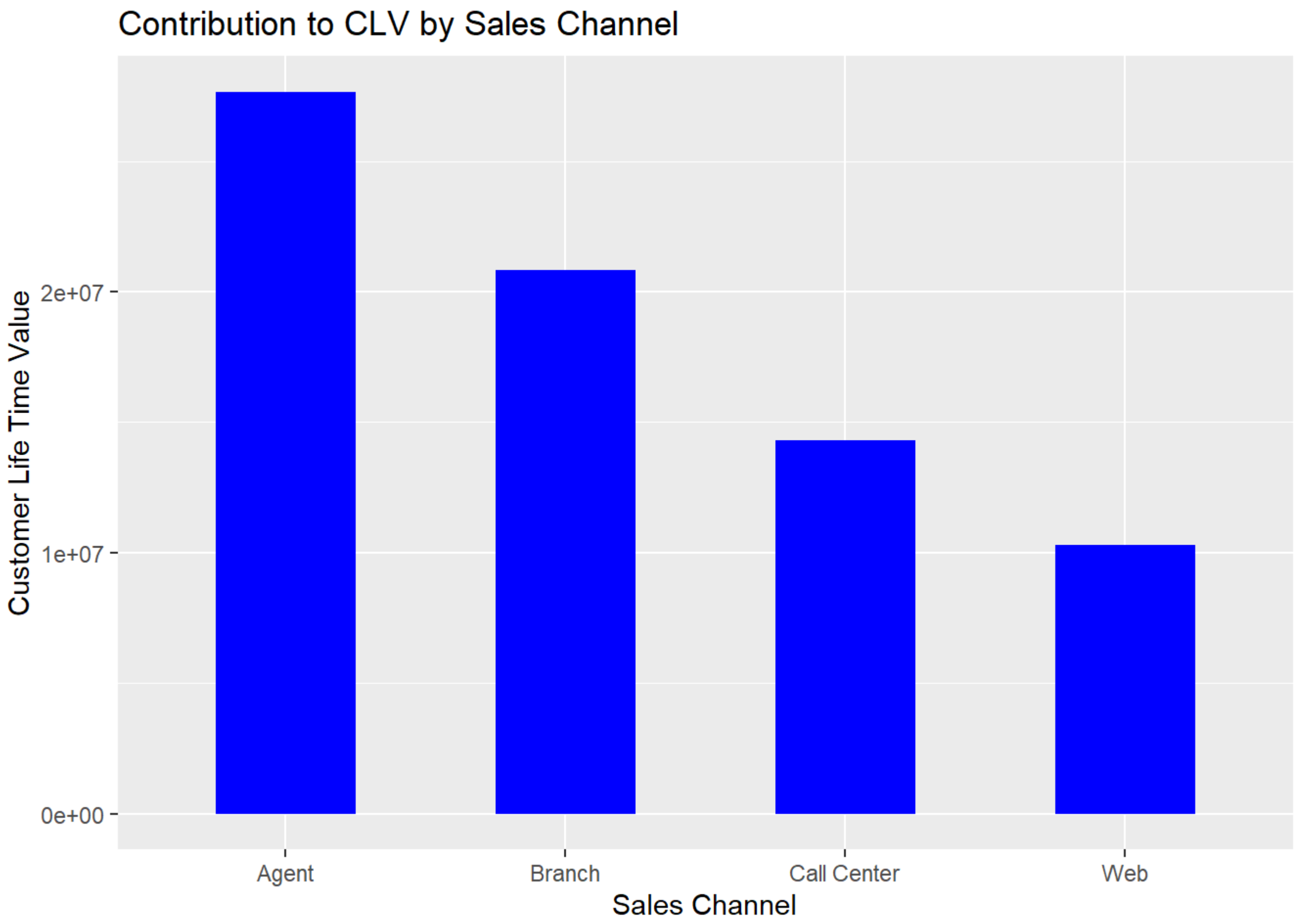
```
ggplot(Insurance_Dataset,aes (x=Gender ,
              y=Customer.Lifetime.Value)) + geom_bar(stat="summary",fun="mean", width=0.5, fill = "Blue")+
  labs(x="Gender",y = "Customer Life Time Value", fill="Gender") +
  ggtitle("Mean Contribution to CLV by Gender")
```



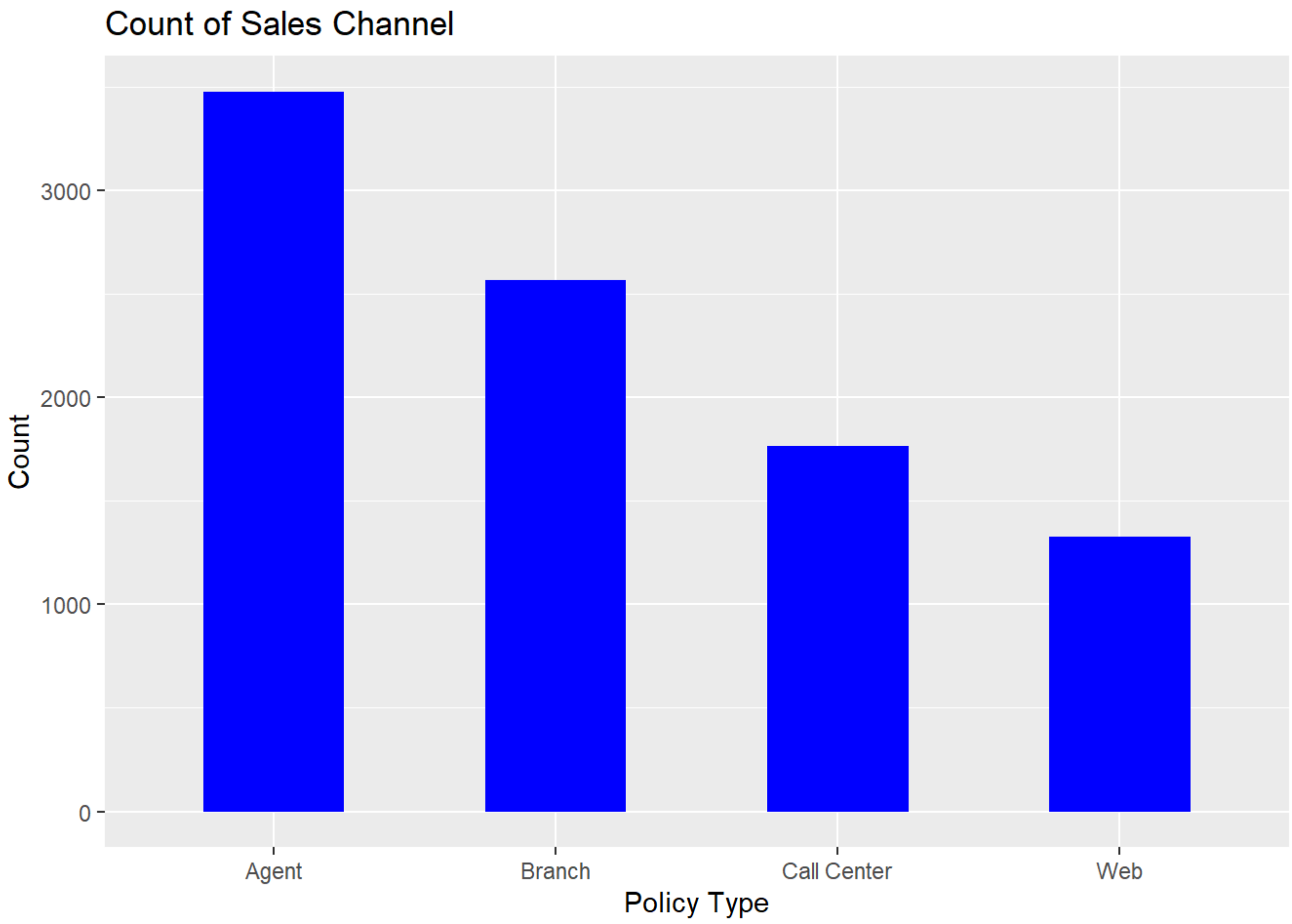
Females are on an average slightly better contributors to CLV than men as there are more female subscribers.

## To visualize the effect of Sales Channel on CLV.

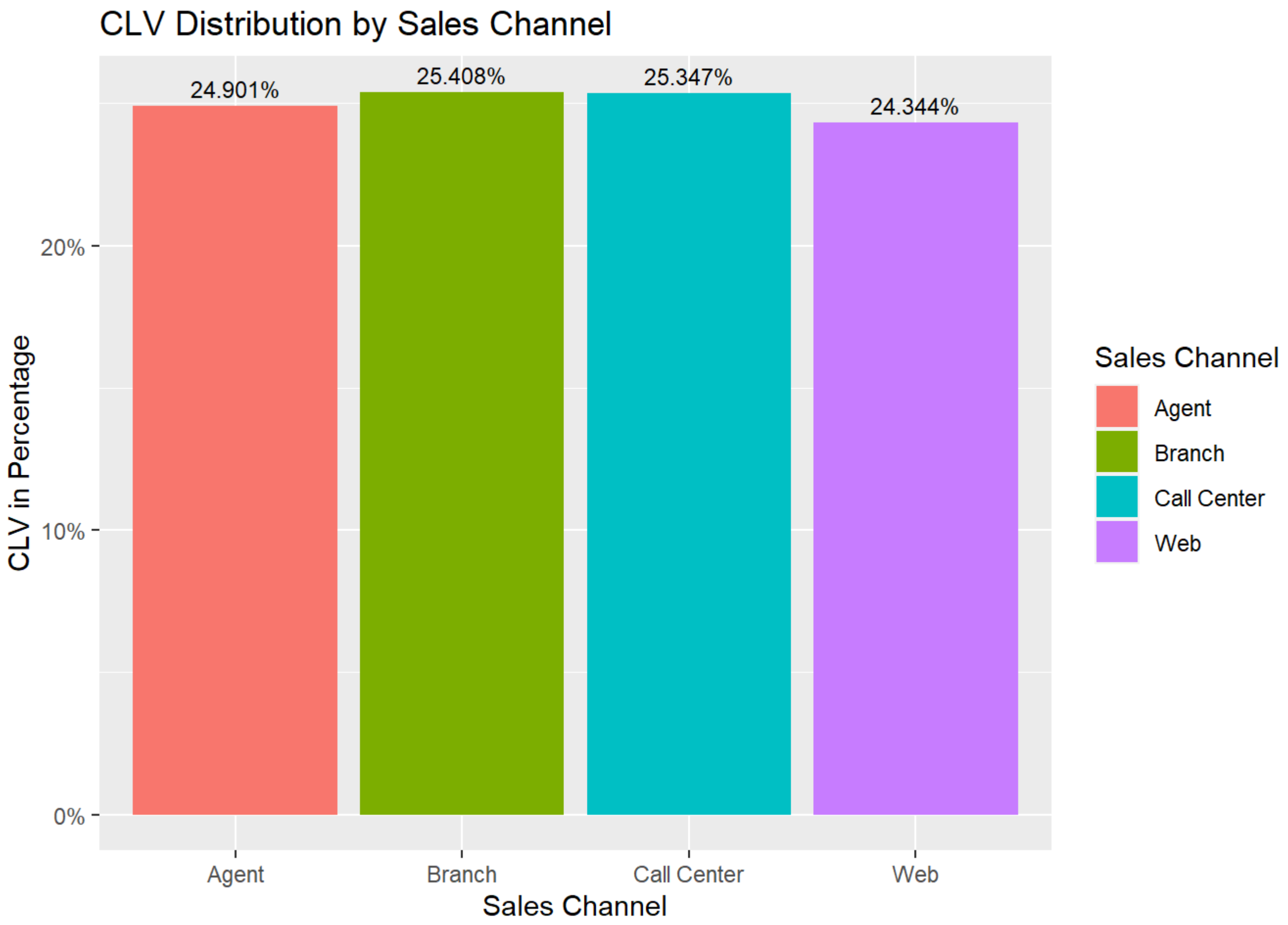
```
ggplot(Insurance_Dataset,aes (x=Sales.Channel ,
              y=Customer.Lifetime.Value)) + geom_bar(stat="summary",fun="sum", width=0.5, fill = "Blue")+
  labs(x="Sales Channel",y = "Customer Life Time Value", fill="Sales Channel") +
  ggtitle("Contribution to CLV by Sales Channel")
```



```
ggplot(Insurance_Dataset,aes (x=Sales.Channel)) +
         geom_bar(stat="count", width=0.5, fill = "Blue") +
  labs(x="Policy Type",y = "Count") +
  ggtitle("Count of Sales Channel")
```



```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(Sales.Channel = Insurance_Dataset$Sales.Channel),
                     FUN = mean)
ggplot(data = aggData, aes(x = Sales.Channel, y = prop.table(stat(aggData$x)), fill = Sales.Channel, label = scales::percent
(prop.table(stat(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'Sales Channel', y = 'CLV in Percentage', fill = 'Sales Channel') +
  ggtitle("CLV Distribution by Sales Channel")
```

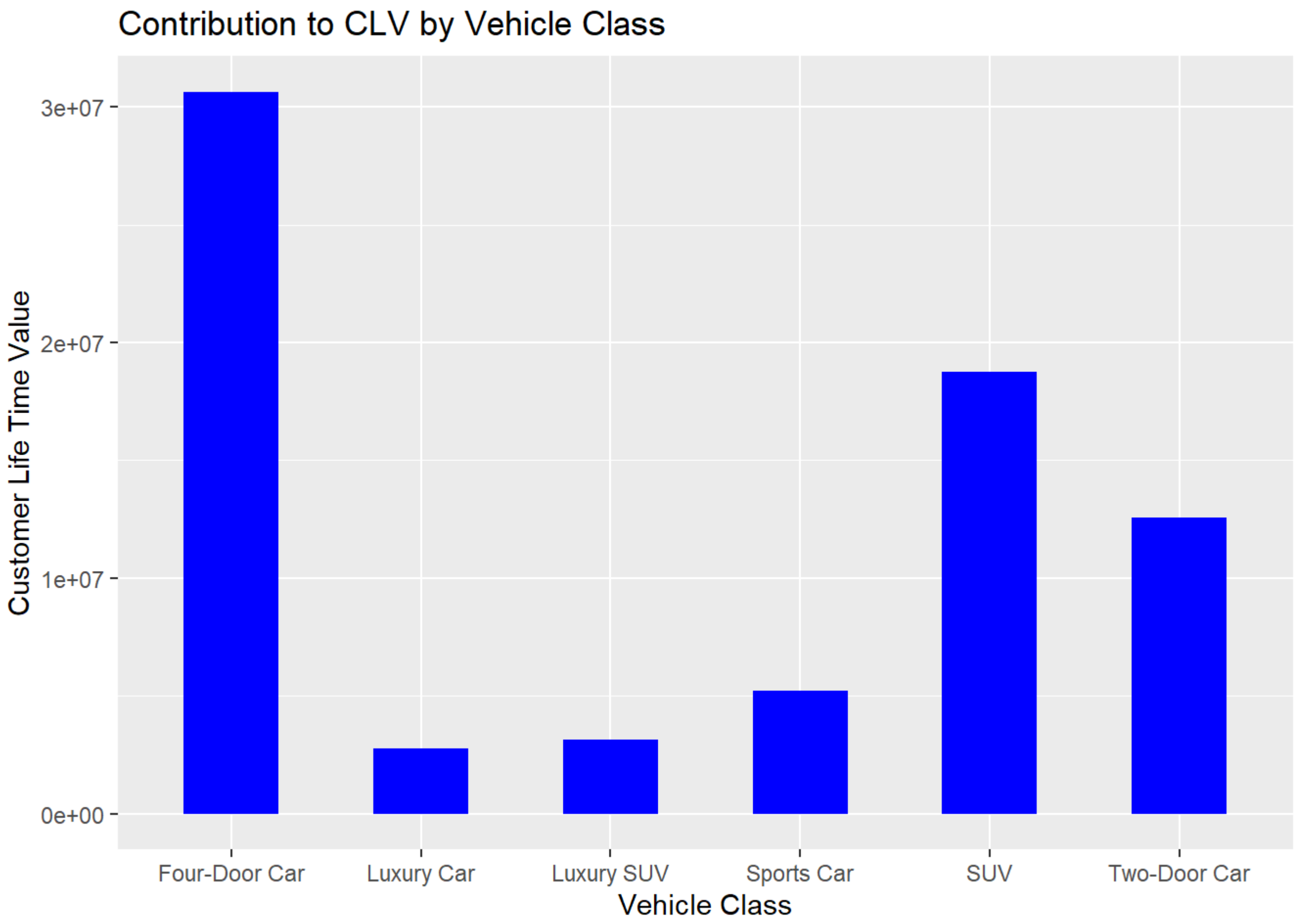


The customers procured through agents are contributing to higher CLV.

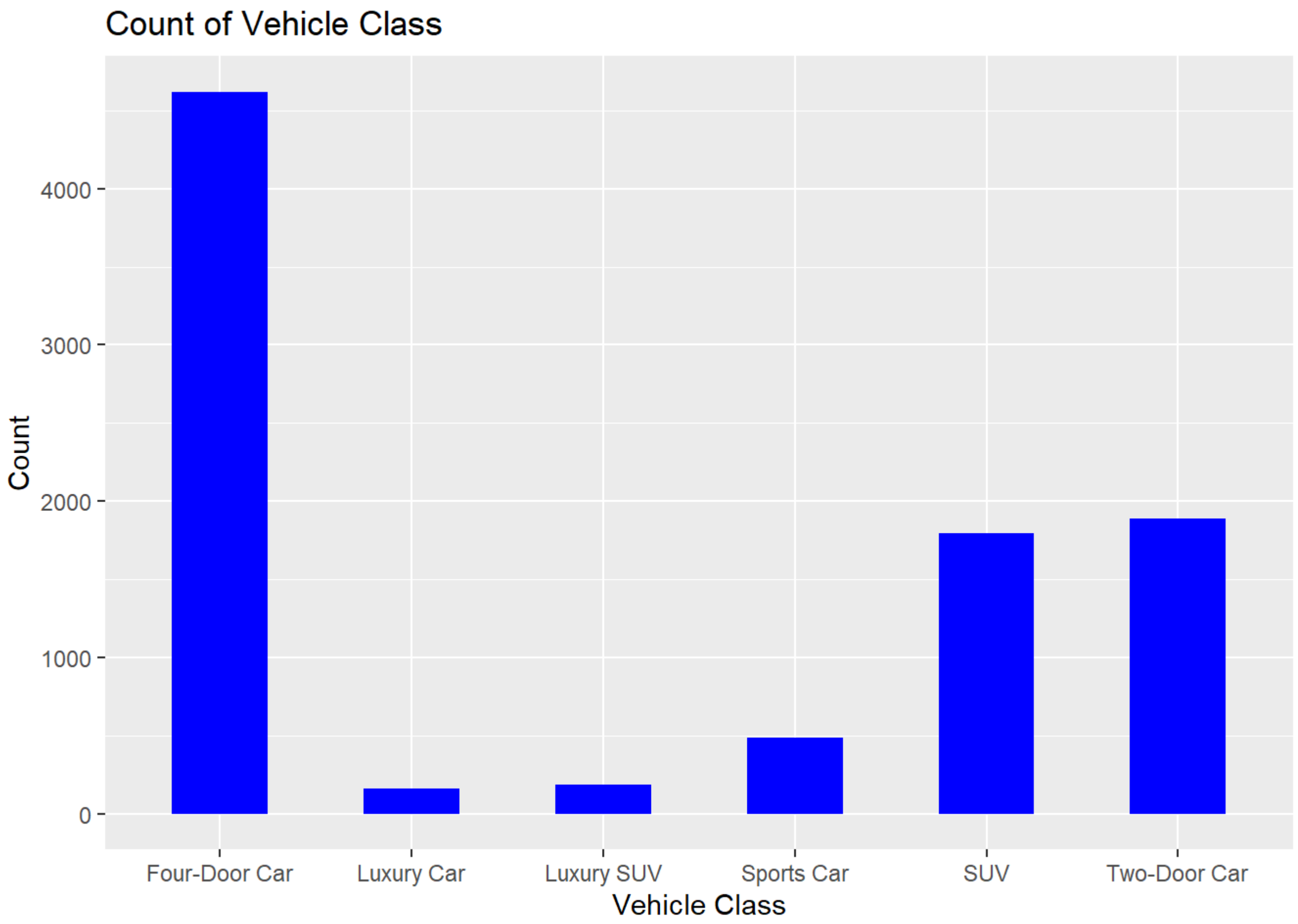
From the third chart, it is evident that it is hard to predict CLV from Sales Channel as all the sub-categories are equal contributors on an average to CLV. Therefore the insurance company needs to promote the channel which costs the least to sustain operations.

## To visualize the effect of Vehicle Class on CLV

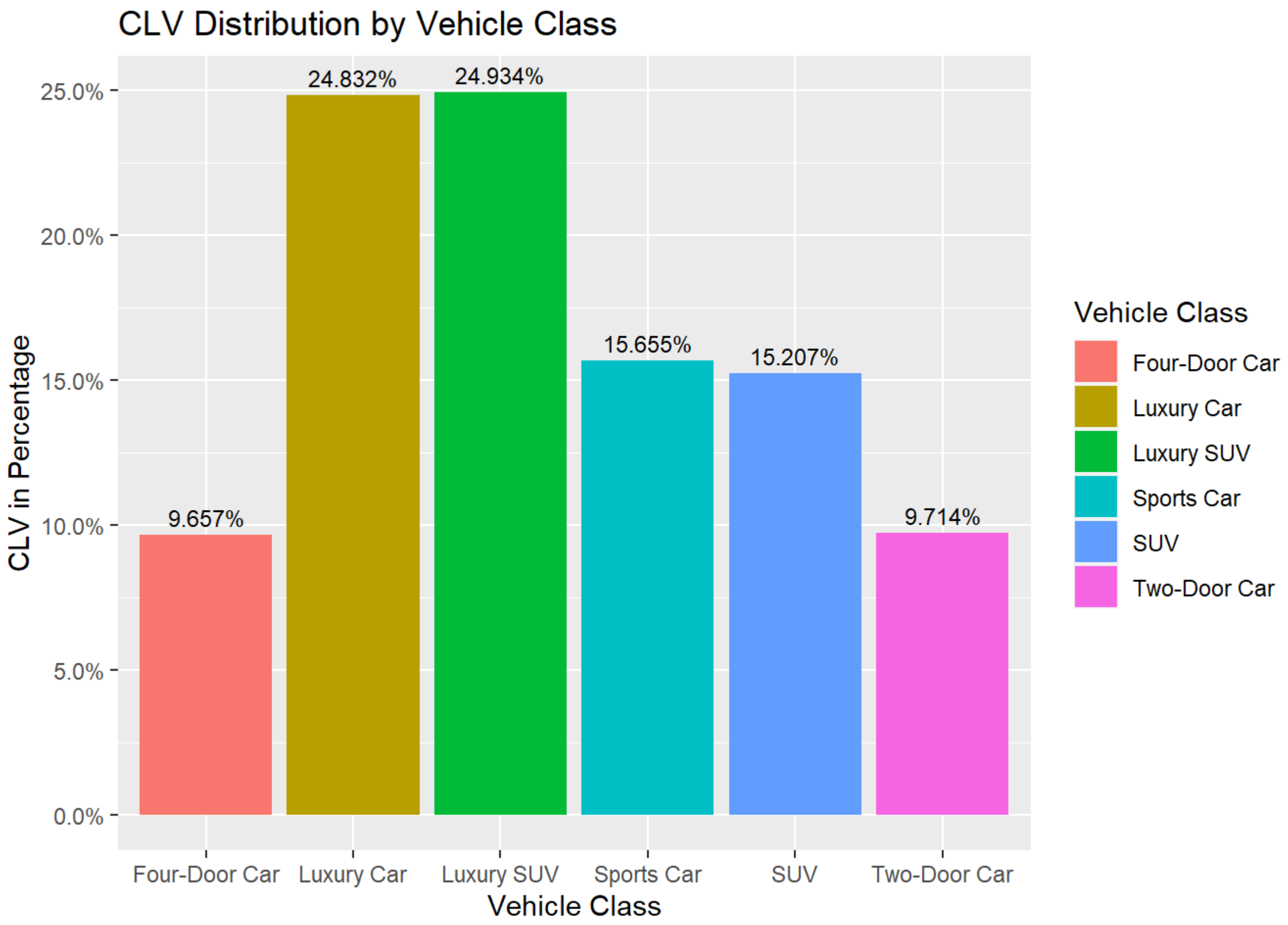
```
ggplot(Insurance_Dataset,aes (x=Vehicle.Class ,
              y=Customer.Lifetime.Value)) + geom_bar(stat="summary",fun="sum", width=0.5, fill = "Blue")+
  labs(x="Vehicle Class",y = "Customer Life Time Value", fill="Vehicle Class") +
  ggtitle("Contribution to CLV by Vehicle Class")
```



```
ggplot(Insurance_Dataset,aes (x=Vehicle.Class)) +
         geom_bar(stat="count", width=0.5, fill = "Blue") +
  labs(x="Vehicle Class",y = "Count") +
  ggtitle("Count of Vehicle Class")
```



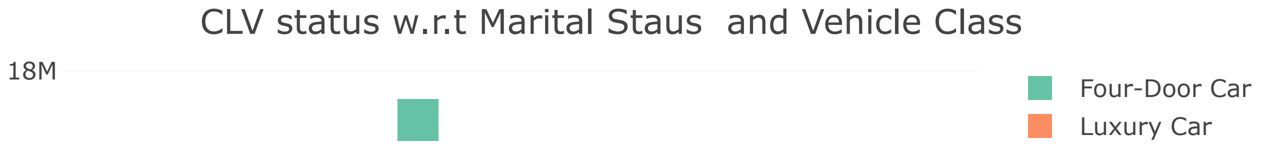
```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(Vehicle.Class = Insurance_Dataset$Vehicle.Class),
                     FUN = mean)
ggplot(data = aggData, aes(x = Vehicle.Class, y = prop.table(stat(aggData$x)), fill = Vehicle.Class, label = scales::percent
(prop.table(stat(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'Vehicle Class', y = 'CLV in Percentage', fill = 'Vehicle Class') +
  ggtitle("CLV Distribution by Vehicle Class")
```

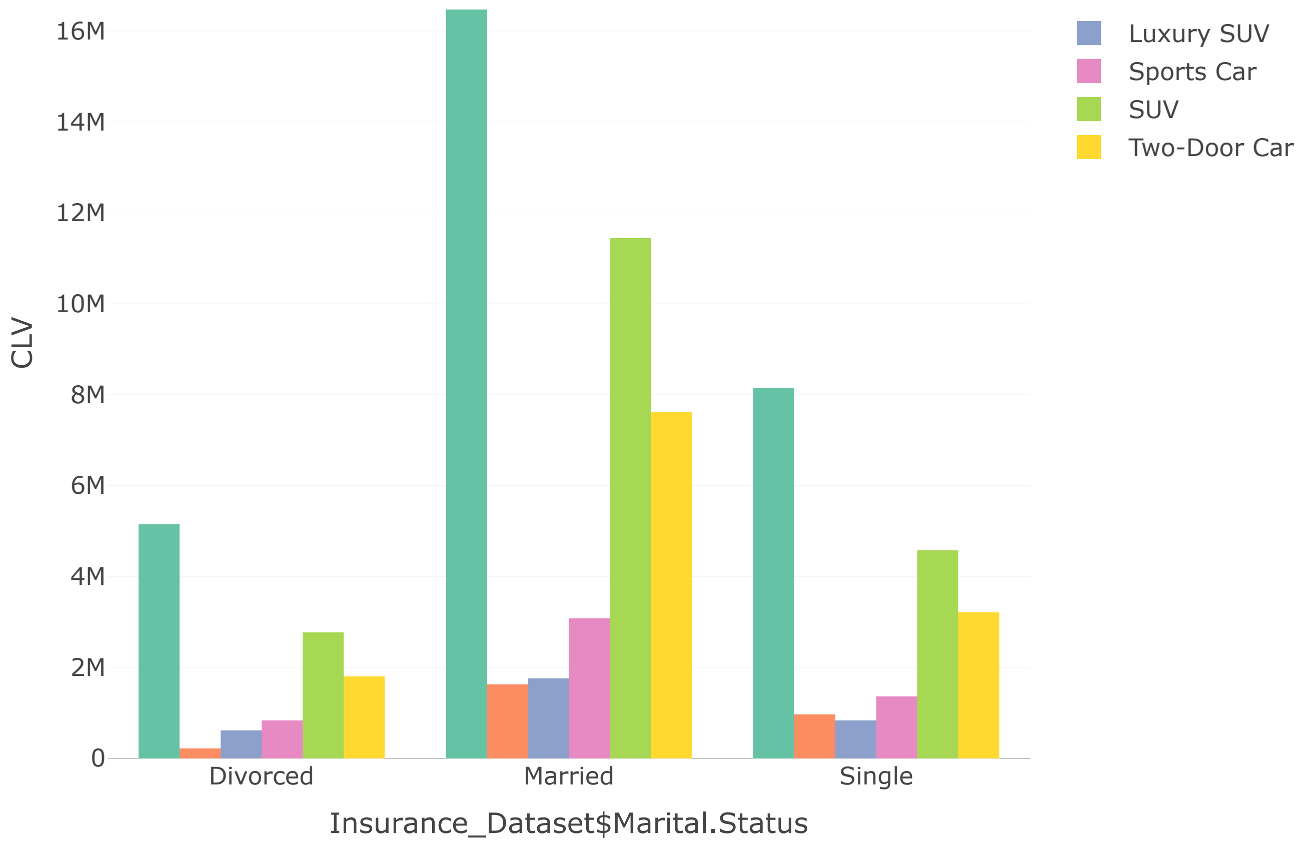


These two charts show us that although the customers owning Luxury, and Luxury SUV are a small fraction, on an average they contribute to almost 50% of CLV. Therefore we can make a conclusion that if a customer owns a Luxury car or Luxury SUV car, there is a high likelihood that she/he she will have high CLV.

## Effect on CLV by Marital Status and Vehicle Class

```
p1<-plot_ly(Insurance_Dataset, x =~Insurance_Dataset$`Marital.Status`, y =~Insurance_Dataset$`Customer.Lifetime.Value`,type=
'bar',color=~Insurance_Dataset$`Vehicle.Class`)
layout(p1, title ='CLV status w.r.t Marital Staus  and Vehicle Class', yaxis = list(title = 'CLV '))
```

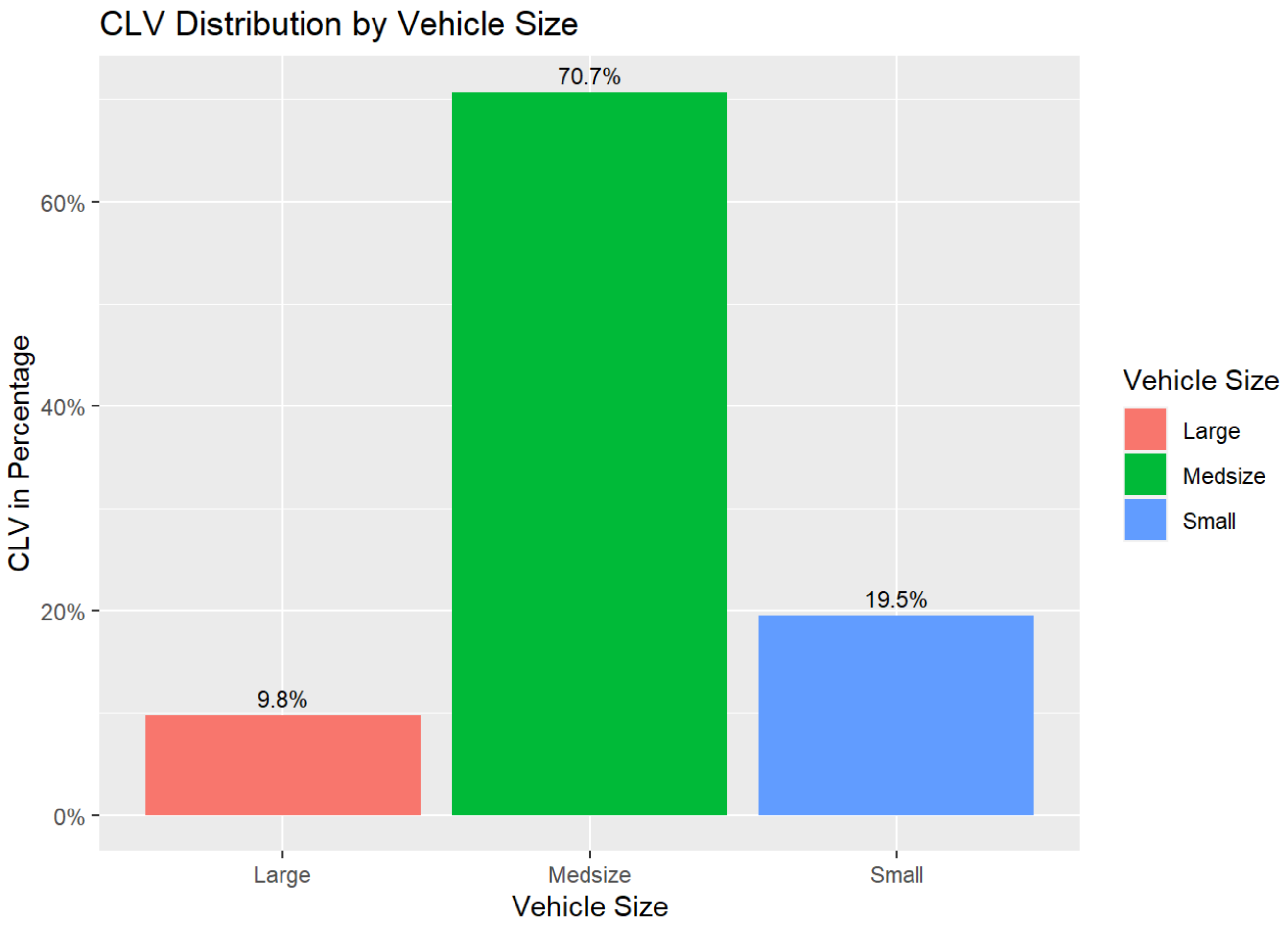




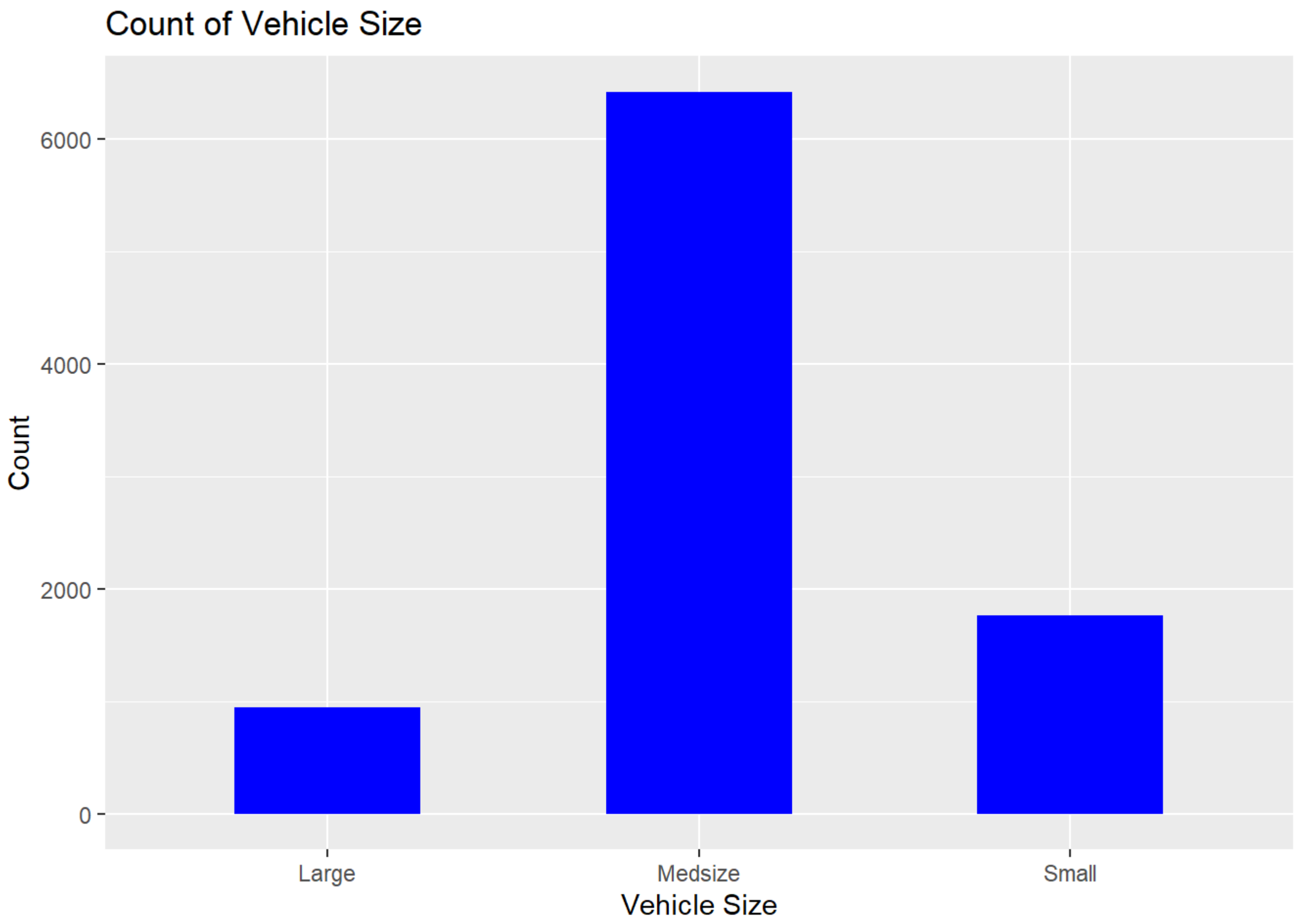
When we consider Marital Status and Vehicle Class we notice that across all the marital statuses the customers owning Four-Door cars and SUVs are better contributors to CLV.

## To visualize the effect of Vehicle Size on CLV.

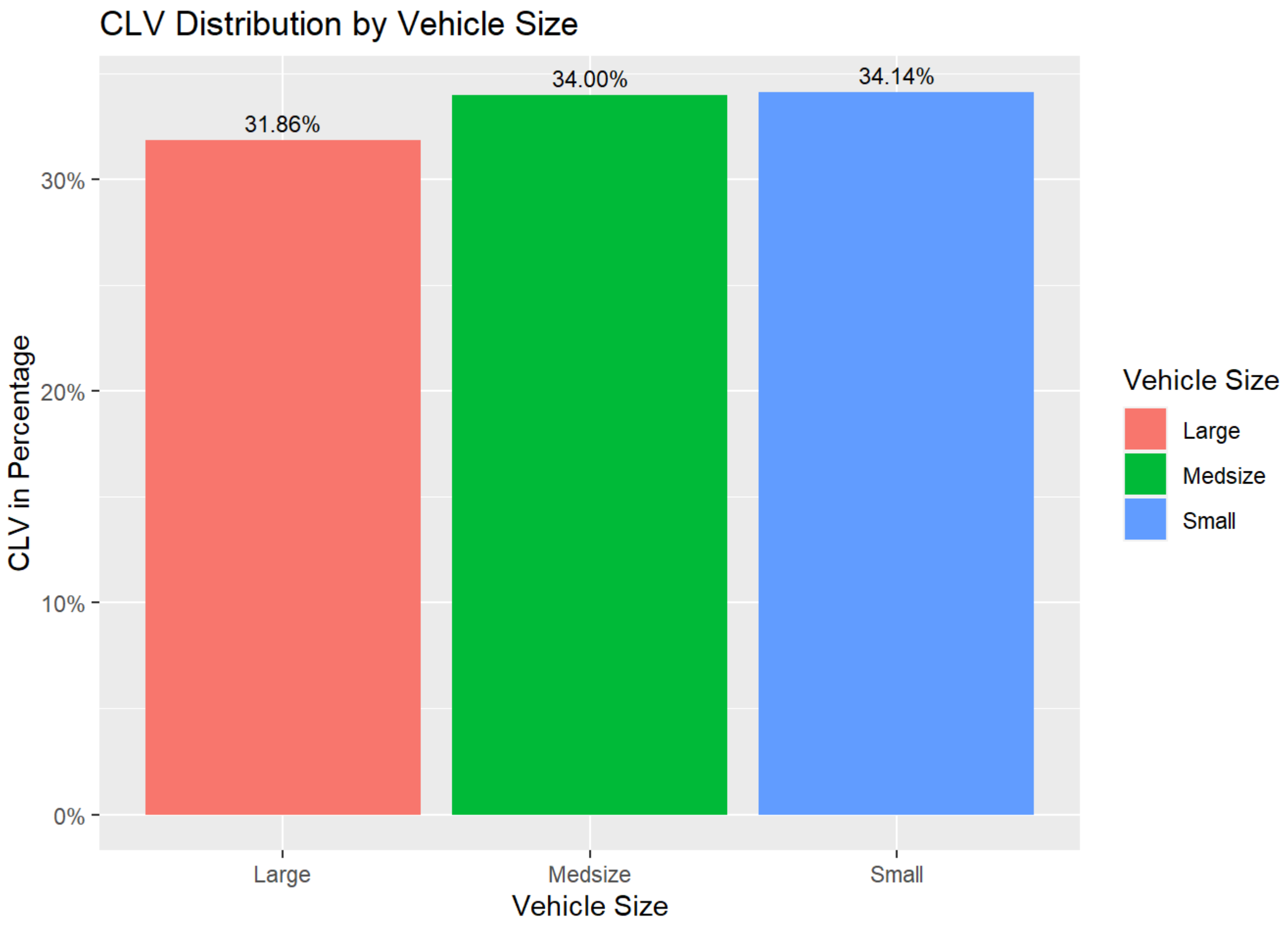
```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(Vehicle.Size = Insurance_Dataset$Vehicle.Size),
                     FUN = sum)
ggplot(data = aggData, aes(x = Vehicle.Size, y = prop.table(stat(aggData$x)), fill = Vehicle.Size, label = scales::percent(prop.table(stat(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'Vehicle Size', y = 'CLV in Percentage', fill = 'Vehicle Size') +
  ggtitle("CLV Distribution by Vehicle Size")
```



```
ggplot(Insurance_Dataset,aes (x=Vehicle.Size)) +
         geom_bar(stat="count", width=0.5, fill = "Blue") +
  labs(x="Vehicle Size",y = "Count") +
  ggtitle("Count of Vehicle Size")
```



```
aggData <- aggregate(x = Insurance_Dataset$Customer.Lifetime.Value,
                     by=list(Vehicle.Size = Insurance_Dataset$Vehicle.Size),
                     FUN = mean)
ggplot(data = aggData, aes(x = Vehicle.Size, y = prop.table(stat(aggData$x)), fill = Vehicle.Size, label = scales::percent(p
rop.table(stat(aggData$x))))) +
  geom_bar(stat="identity", position = "dodge") +
  geom_text(stat = 'identity', position = position_dodge(.9),  vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(x = 'Vehicle Size', y = 'CLV in Percentage', fill = 'Vehicle Size') +
  ggtitle("CLV Distribution by Vehicle Size")
```



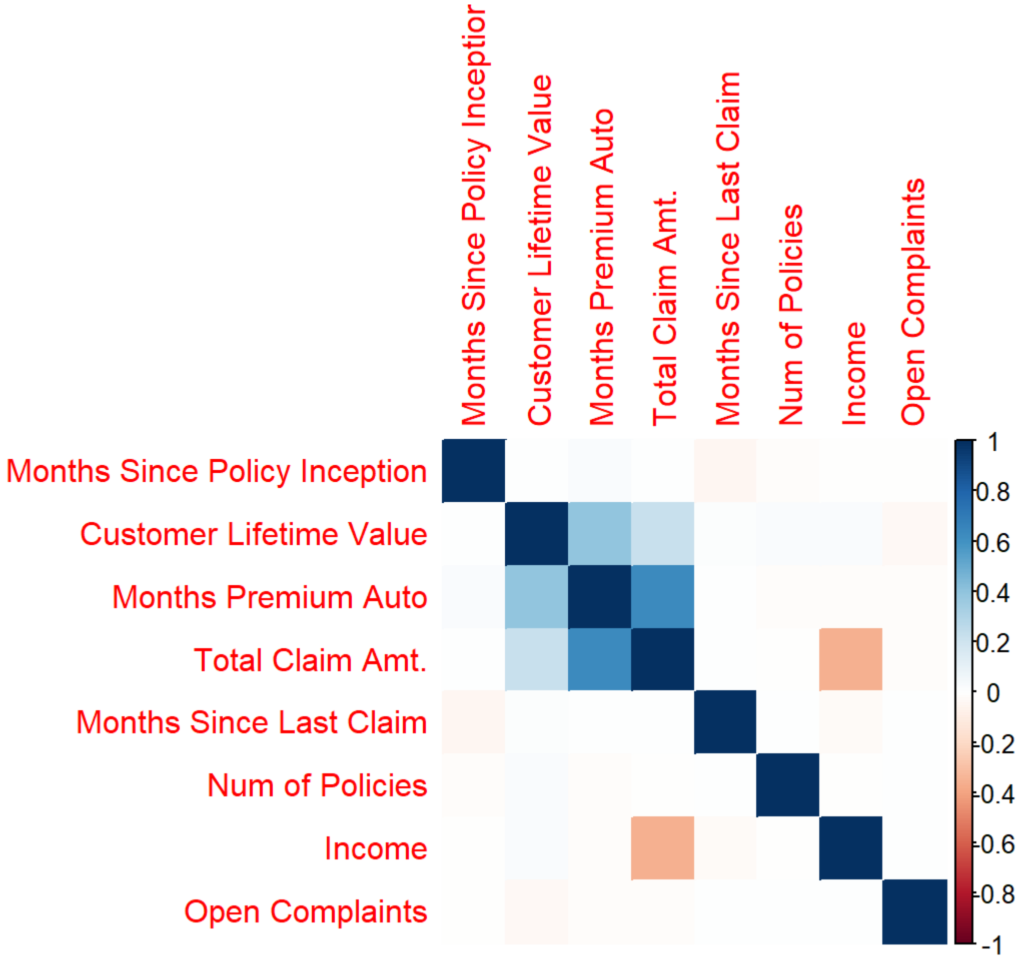
As we can see, the variable vehicle size is a weak predictor because all the sub-categories contribute equally to the CLV on an average.

# EDA OF NUMERIC DEPENDENT VARIABLES VS CLV.

## To visualize the correlation between the variables

A correlation heat map is plotted for all the numeric variables. This also checks for multi-collinearity between variables.

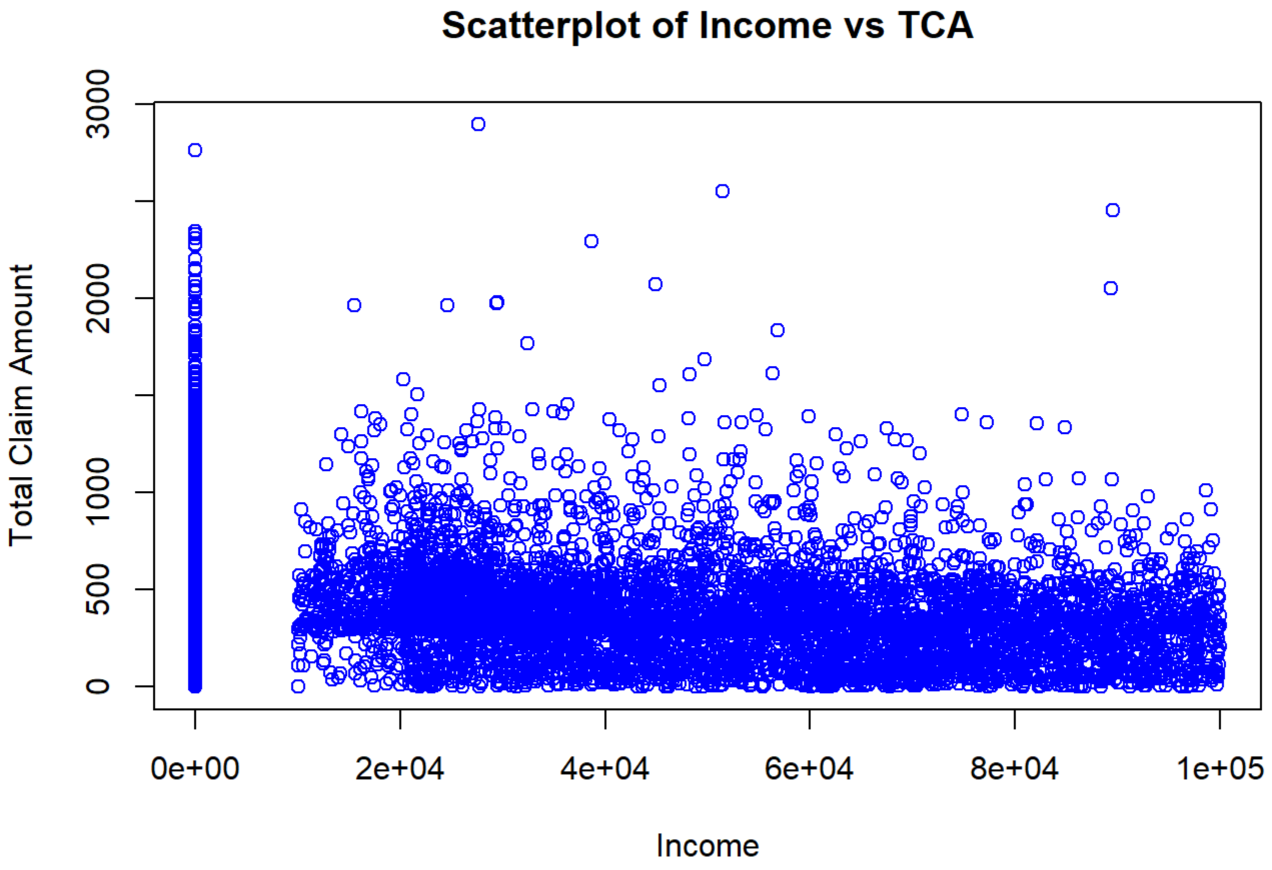
```
autoCorr <- Insurance_Dataset[,c(3,10,13:17,22)]
colnames(autoCorr) <- c("Customer Lifetime Value", "Income", "Months Premium Auto", "Months Since Last Claim", "Months Since
Policy Inception",
                        "Open Complaints", "Num of Policies", "Total Claim Amt.")
autoCorr <- cor(autoCorr)
# Plot the correlation table
corrplot(autoCorr, method = "color", order = "hclust")
```



As is evident from the correlation plot, Monthly Premium Auto and Total Claim Amount are moderately correlated while the other variables are weakly correlated. Other than Monthly Premium Auto and Total Claim Amount negligible multicollinearity is seen between the remaining independent variables.

### To explore the effect of Income and Total Claim Amount

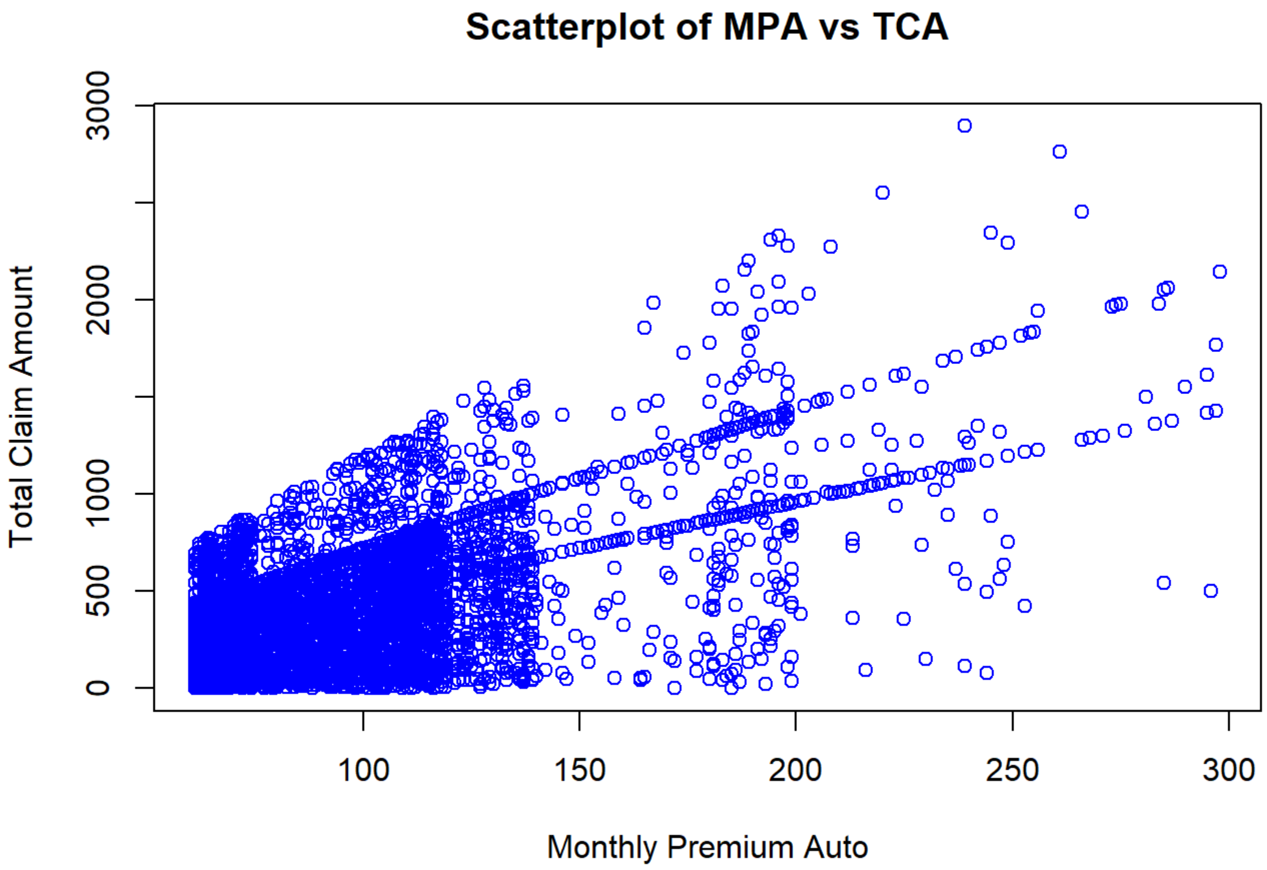
```
plot(x=Insurance_Dataset$"Income", y=Insurance_Dataset$"Total.Claim.Amount", col="Blue", cex=1, xlab="Income",
     ylab="Total Claim Amount",main="Scatterplot of Income vs TCA")
```



We see in this chart that there is no linear positive or negative relationship between variables. This means that they are independent of each other

## To visualize the effect of Monthly Premium Auto and Total Claim Amount

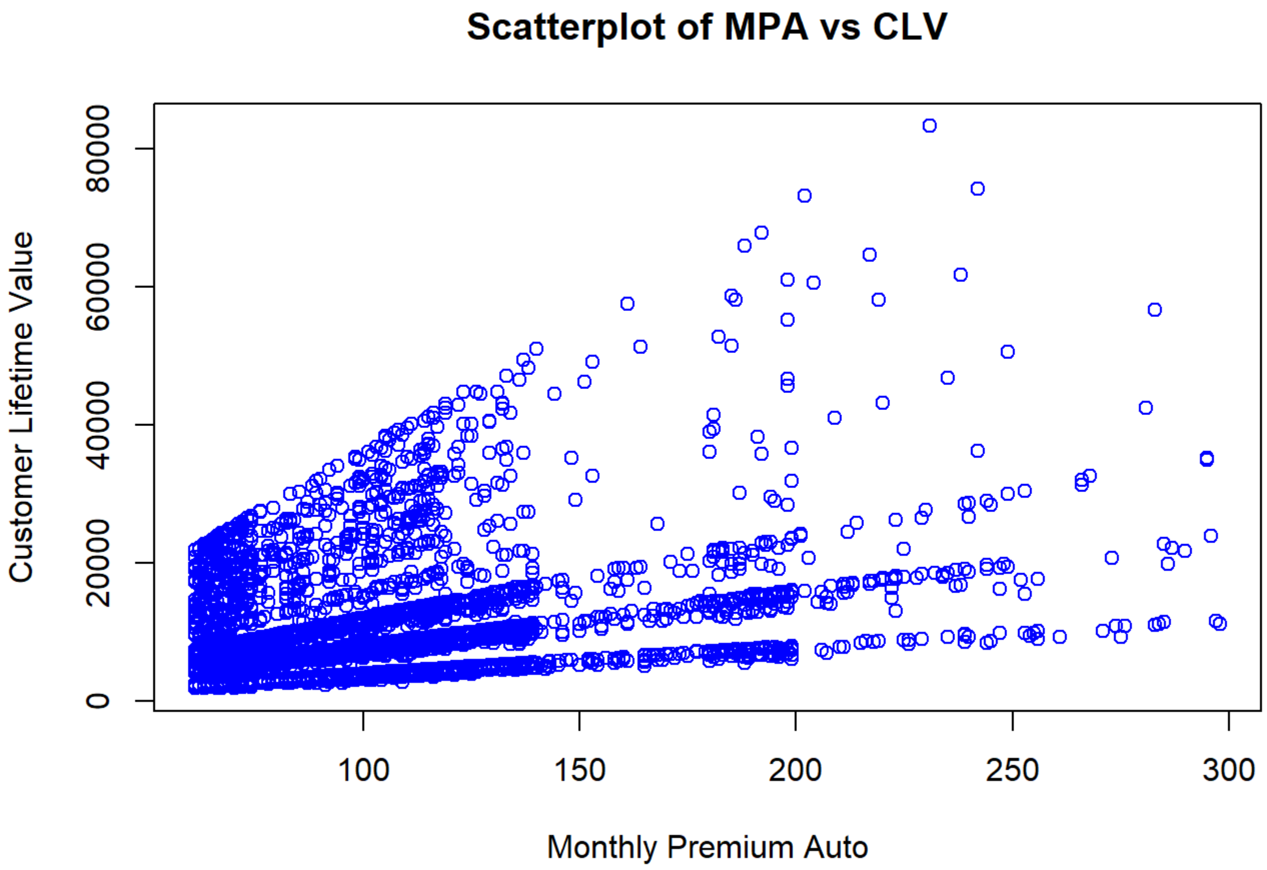
```
plot(x=Insurance_Dataset$"Monthly.Premium.Auto", y=Insurance_Dataset$"Total.Claim.Amount", col="Blue", cex=1, xlab="Monthly
 Premium Auto",
     ylab="Total Claim Amount",main="Scatterplot of MPA vs TCA")
```



Here we see the relationship of MPA and TCA, we notice that a few clusters have a positive linear relationship, as evidenced by the upward slopes in the chart.

## To visualize the effect of Monthly Premium Auto and CLV

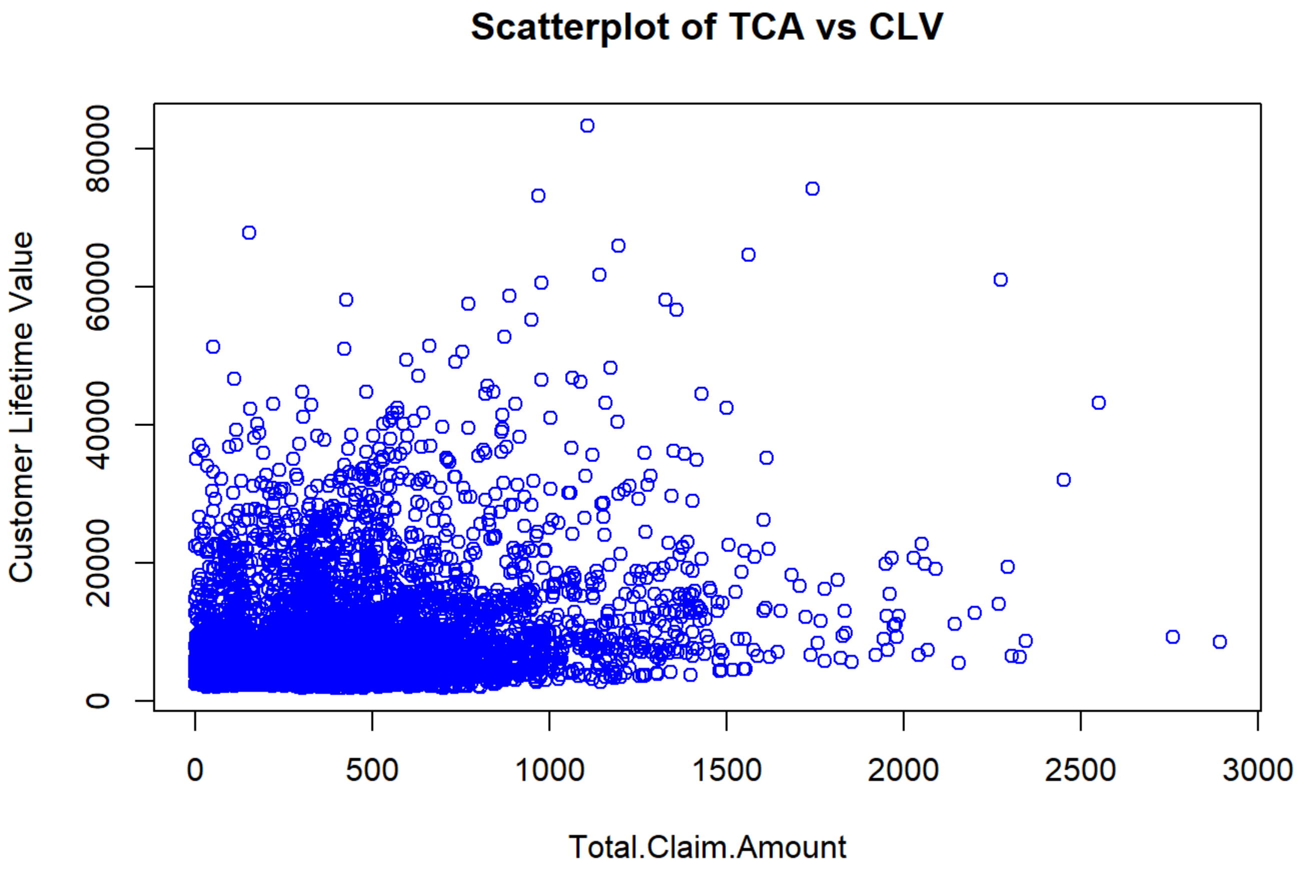
```
plot(x=Insurance_Dataset$"Monthly.Premium.Auto", y=Insurance_Dataset$"Customer.Lifetime.Value", col="Blue", cex=1, xlab="Monthly Premium Auto",
     ylab="Customer Lifetime Value",main="Scatterplot of MPA vs CLV")
```



From the scatterplot it is evident that higher the MPA, higher is the CLV.

## To visualize the effect of Total Claim Amount on CLV.

```
plot(x=Insurance_Dataset$"Total.Claim.Amount", y=Insurance_Dataset$"Customer.Lifetime.Value", col="Blue", cex=1, xlab="Tota
l.Claim.Amount",
     ylab="Customer Lifetime Value", main="Scatterplot of TCA vs CLV")
```



There is no evidence that there is any linear relationship between Total Claim Amount and CLV as the scatterplot is inconclusive. There is no clear slope either downward or upward in the chart indicating that these two variables are independent of each other.

# Feature Engineering, Feature Selection and Model Building

## Introduction

Having done the EDA, Feature Engineering, Feature Selection and Model Building was carried out.

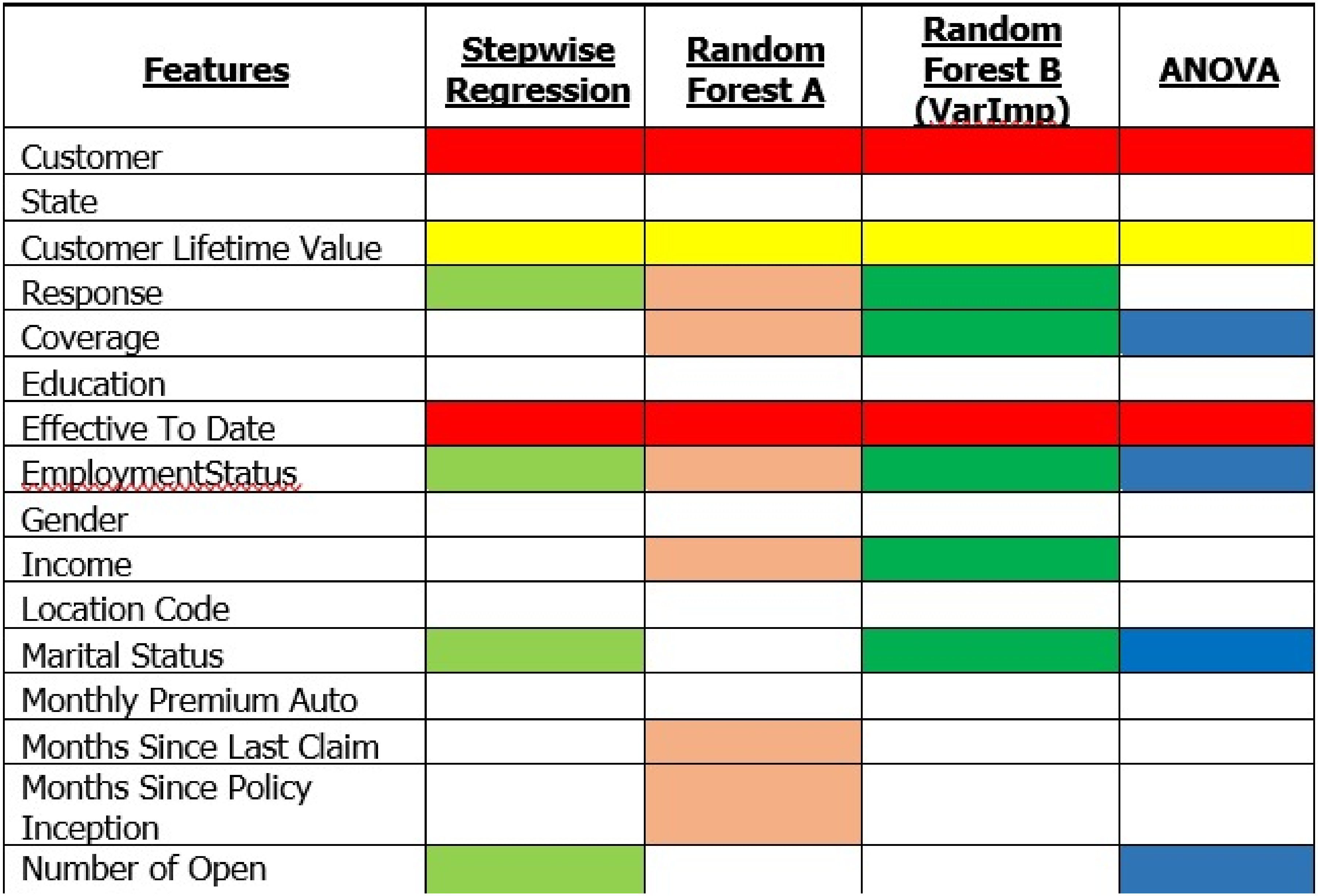
## Feature Engineering:

1. Sqrt and Log transformations have been used in trying out various models.

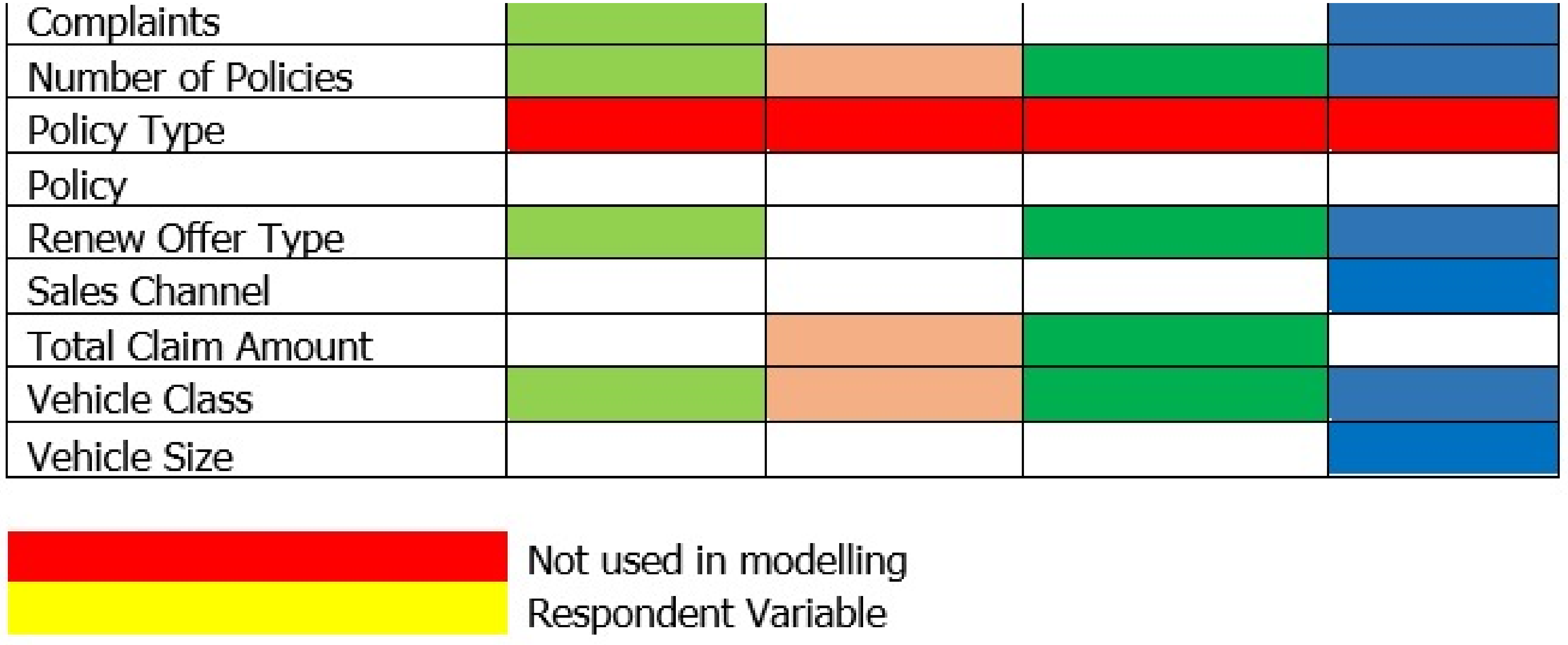
2. Variations of relationship between “Monthly Premium Auto” and “Number of Policies” were tried out but their effect was redundant.

## Feature Selection

Feature selection using Stepwise Regression, Random Forest and ANOVA were carried out. These features were used in the various models that were tried out, with and without transformation. The details are in the report below.



| **Features** | **Stepwise Regression** | **Random Forest A** | **Random Forest B (VarImp)** | **ANOVA** |
|---|---|---|---|---|
| Customer | 🟥 | 🟥 | 🟥 | 🟥 |
| State | | | | |
| Customer Lifetime Value | 🟨 | 🟨 | 🟨 | 🟨 |
| Response | 🟩 | 🟧 | 🟩 | |
| Coverage | | 🟧 | 🟩 | 🟦 |
| Education | | | | |
| Effective To Date | 🟥 | 🟥 | 🟥 | 🟥 |
| EmploymentStatus | 🟩 | 🟧 | 🟩 | 🟦 |
| Gender | | | | |
| Income | | 🟧 | 🟩 | |
| Location Code | | | | |
| Marital Status | 🟩 | | 🟩 | 🟦 |
| Monthly Premium Auto | | | | |
| Months Since Last Claim | | 🟧 | | |
| Months Since Policy Inception | | 🟧 | | |
| Number of Open | 🟩 | | | 🟦 |



| | | | | |
|---|---|---|---|---|
| Complaints | 🟩 | | | 🟦 |
| Number of Policies | 🟩 | 🟧 | 🟩 | 🟦 |
| Policy Type | 🟥 | 🟥 | 🟥 | 🟥 |
| Policy | | | | |
| Renew Offer Type | 🟩 | | 🟩 | 🟦 |
| Sales Channel | | | | 🟦 |
| Total Claim Amount | | 🟧 | 🟩 | |
| Vehicle Class | 🟩 | 🟧 | 🟩 | 🟦 |
| Vehicle Size | | | | 🟦 |

🟥 Not used in modelling

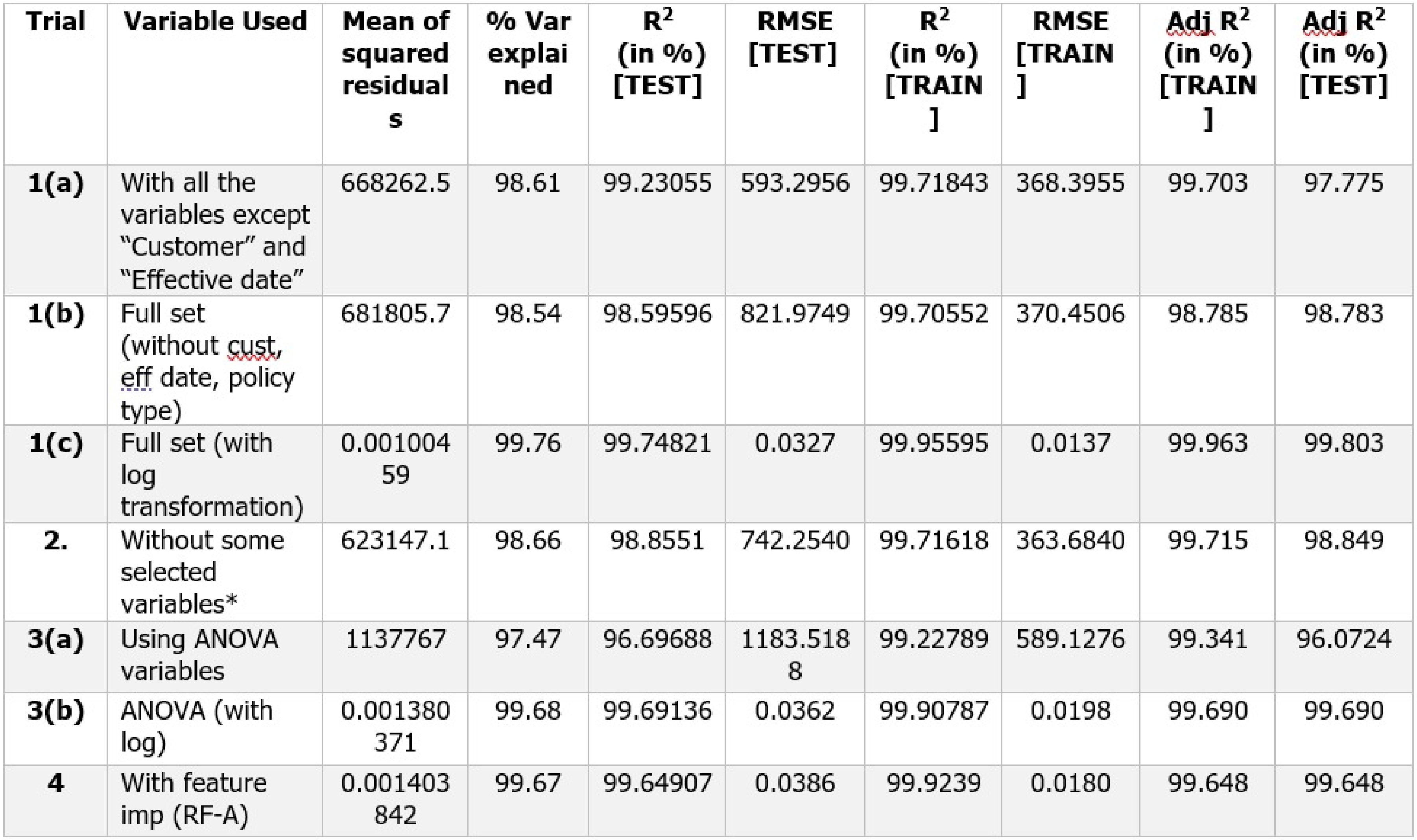
🟨 Respondent Variable

Feature Selection

# Model Building

## 1. Random Forest.

One of the models that was chosen was Random Forest. Since the data had a lot of outliers, random forest was selected as it is resilient to outliers. Random Forest algorithm itself is not robust to outliers but the base learner on which it is built - the decision tree, is.The R2 value (in percent) and Adjusted R2 values (in percent) of the RF model for all the various trials with different variables was 96% or higher, as shown in the table below:-



| Trial | Variable Used | Mean of squared residuals | % Var explained | $R^2$ (in %) [TEST] | RMSE [TEST] | $R^2$ (in %) [TRAIN] | RMSE [TRAIN] | Adj $R^2$ (in %) [TRAIN] | Adj $R^2$ (in %) [TEST] |
|---|---|---|---|---|---|---|---|---|---|
| **1(a)** | With all the variables except "Customer" and "Effective date" | 668262.5 | 98.61 | 99.23055 | 593.2956 | 99.71843 | 368.3955 | 99.703 | 97.775 |
| **1(b)** | Full set (without cust, eff date, policy type) | 681805.7 | 98.54 | 98.59596 | 821.9749 | 99.70552 | 370.4506 | 98.785 | 98.783 |
| **1(c)** | Full set (with log transformation) | 0.00100459 | 99.76 | 99.74821 | 0.0327 | 99.95595 | 0.0137 | 99.963 | 99.803 |
| **2.** | Without some selected variables* | 623147.1 | 98.66 | 98.8551 | 742.2540 | 99.71618 | 363.6840 | 99.715 | 98.849 |
| **3(a)** | Using ANOVA variables | 1137767 | 97.47 | 96.69688 | 1183.5188 | 99.22789 | 589.1276 | 99.341 | 96.0724 |
| **3(b)** | ANOVA (with log) | 0.001380371 | 99.68 | 99.69136 | 0.0362 | 99.90787 | 0.0198 | 99.690 | 99.690 |
| **4** | With feature imp (RF-A) | 0.001403842 | 99.67 | 99.64907 | 0.0386 | 99.9239 | 0.0180 | 99.648 | 99.648 |

Summary

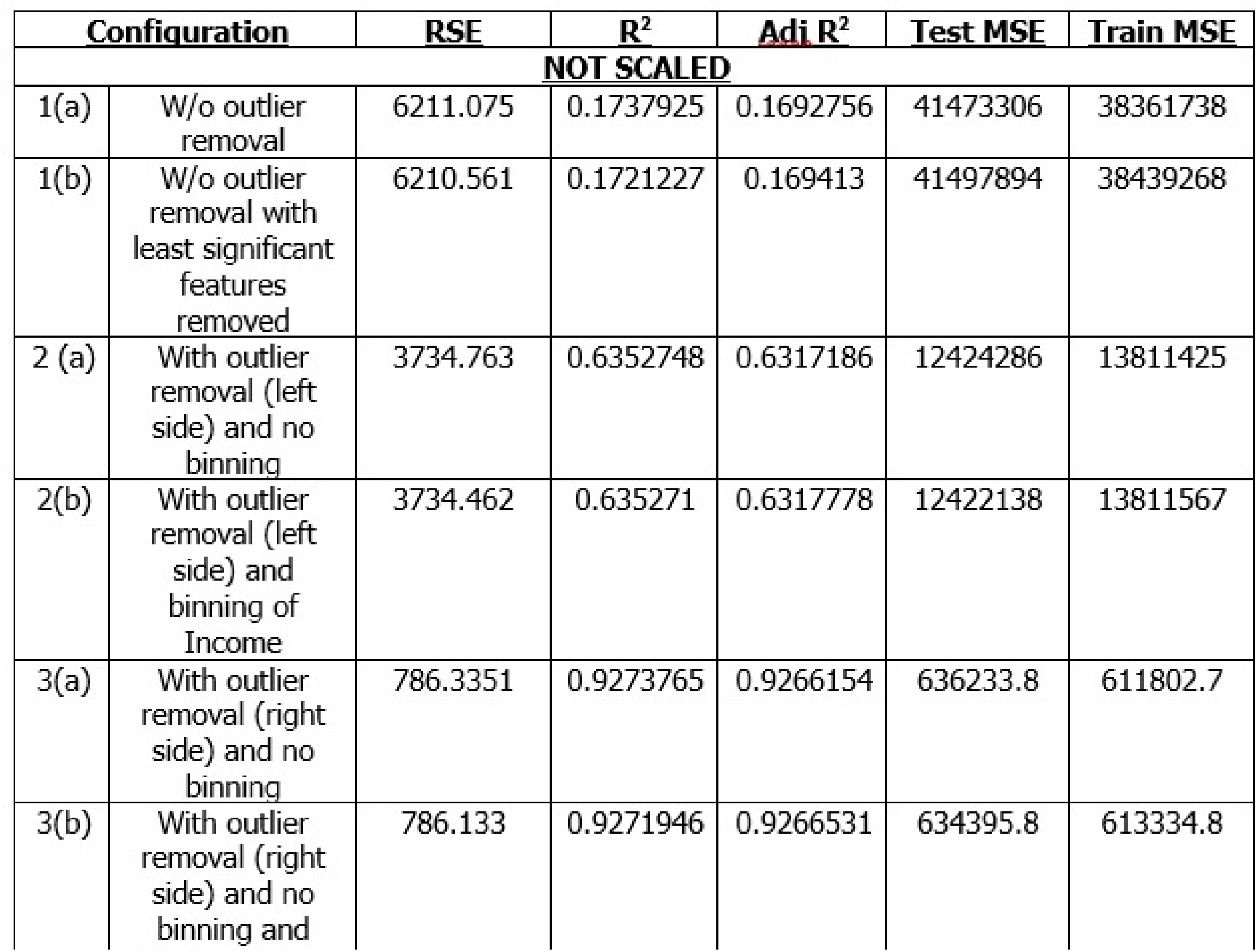
Based on research carried out on the internet for this model, Random Forest is biased towards specific factors (like categorical variables with different levels) because it provided exceptionally high results for each experiment, which does not seem realistic. As a result, it was decided to not go ahead with Random Forest.

- In trial 2, some of the variables were not included which are Customer, State, Response, Effective to Date, Income, Policy Type and Income_Bin. The selection of these variables was through trial and error and were selected as these were giving better results.

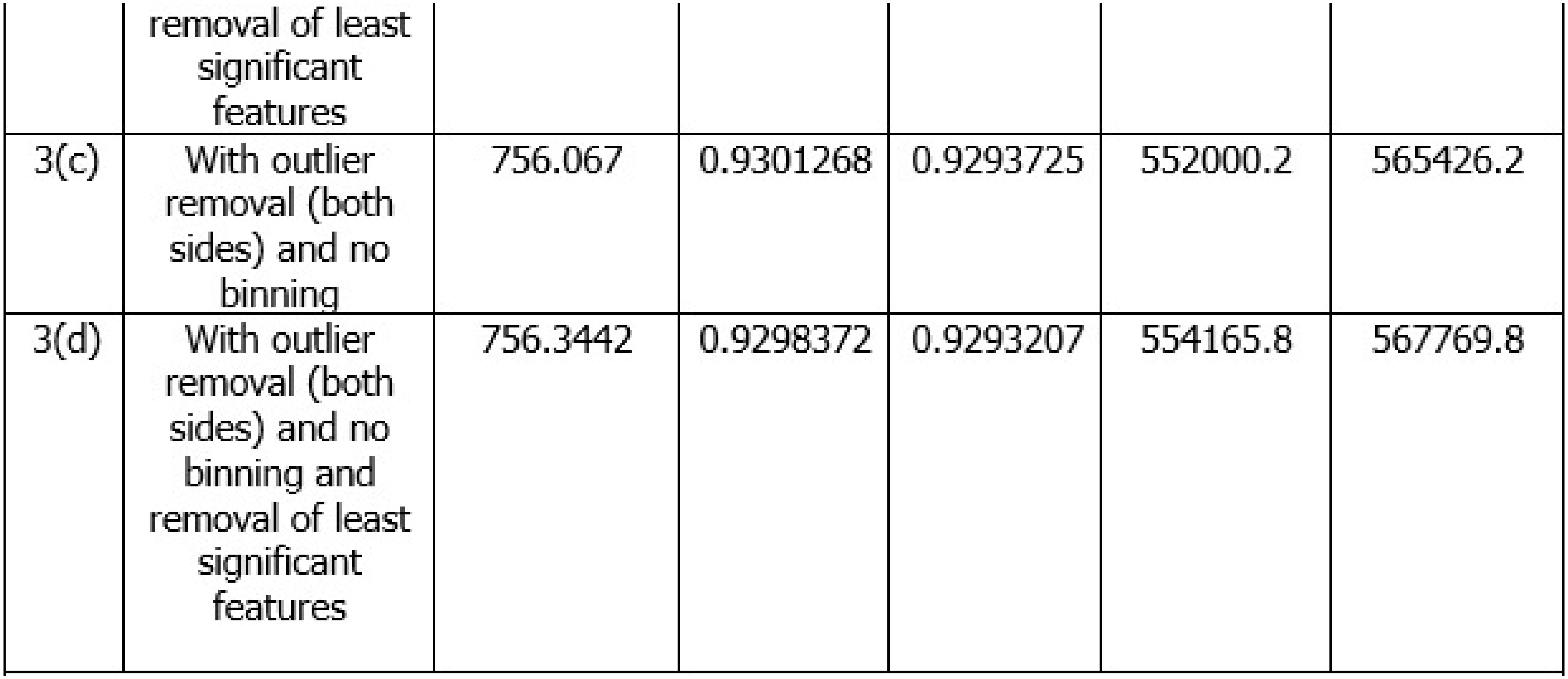
## 2. Linear Regression.

The second model was Linear Regression. The summary of the various models tried out is as given below :-

SUMMARY : REGRESSION

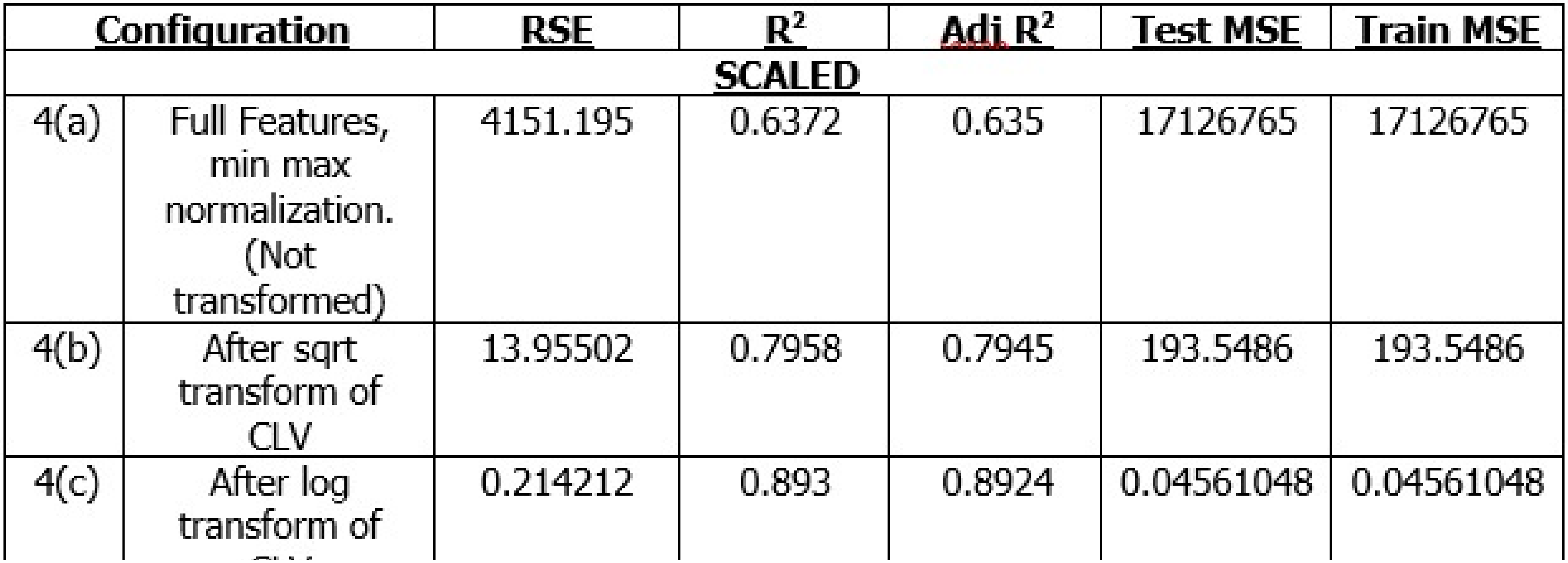


| Configuration | | RSE | $R^2$ | Adj $R^2$ | Test MSE | Train MSE |
|---|---|---|---|---|---|---|
| NOT SCALED | | | | | | |
| 1(a) | W/o outlier removal | 6211.075 | 0.1737925 | 0.1692756 | 41473306 | 38361738 |
| 1(b) | W/o outlier removal with least significant features removed | 6210.561 | 0.1721227 | 0.169413 | 41497894 | 38439268 |
| 2 (a) | With outlier removal (left side) and no binning | 3734.763 | 0.6352748 | 0.6317186 | 12424286 | 13811425 |
| 2(b) | With outlier removal (left side) and binning of Income | 3734.462 | 0.635271 | 0.6317778 | 12422138 | 13811567 |
| 3(a) | With outlier removal (right side) and no binning | 786.3351 | 0.9273765 | 0.9266154 | 636233.8 | 611802.7 |
| 3(b) | With outlier removal (right side) and no binning and | 786.133 | 0.9271946 | 0.9266531 | 634395.8 | 613334.8 |

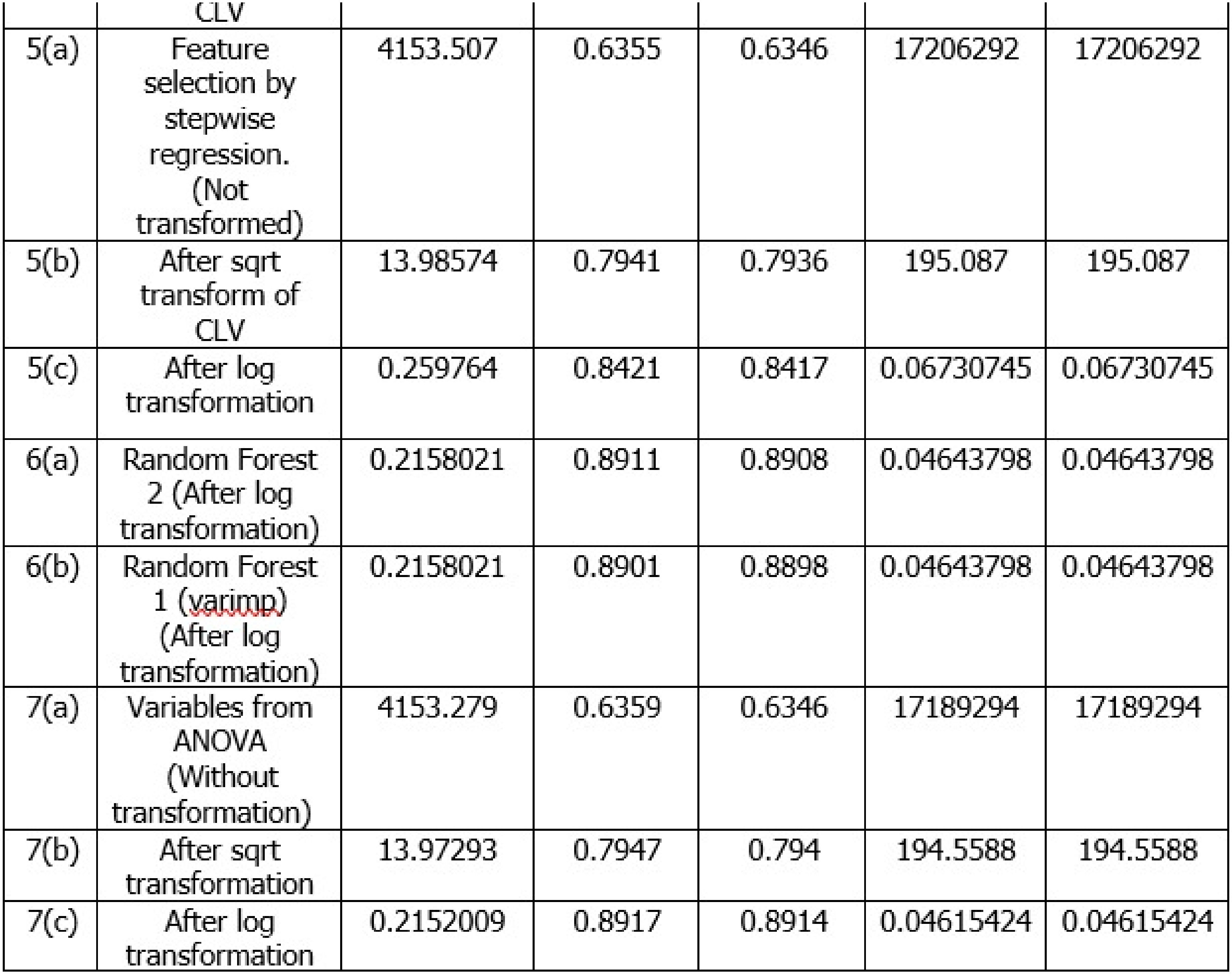


| | | | | | | |
|---|---|---|---|---|---|---|
| | removal of least significant features | | | | | |
| 3(c) | With outlier removal (both sides) and no binning | 756.067 | 0.9301268 | 0.9293725 | 552000.2 | 565426.2 |
| 3(d) | With outlier removal (both sides) and no binning and removal of least significant features | 756.3442 | 0.9298372 | 0.9293207 | 554165.8 | 567769.8 |

Not Scaled



| **Configuration** | | **RSE** | **$R^2$** | **Adj $R^2$** | **Test MSE** | **Train MSE** |
|---|---|---|---|---|---|---|
| **SCALED** | | | | | | |
| 4(a) | Full Features, min max normalization. (Not transformed) | 4151.195 | 0.6372 | 0.635 | 17126765 | 17126765 |
| 4(b) | After sqrt transform of CLV | 13.95502 | 0.7958 | 0.7945 | 193.5486 | 193.5486 |
| 4(c) | After log transform of | 0.214212 | 0.893 | 0.8924 | 0.04561048 | 0.04561048 |



| | | | | | | |
|---|---|---|---|---|---|---|
| | CLV | | | | | |
| 5(a) | Feature selection by stepwise regression. (Not transformed) | 4153.507 | 0.6355 | 0.6346 | 17206292 | 17206292 |
| 5(b) | After sqrt transform of CLV | 13.98574 | 0.7941 | 0.7936 | 195.087 | 195.087 |
| 5(c) | After log transformation | 0.259764 | 0.8421 | 0.8417 | 0.06730745 | 0.06730745 |
| 6(a) | Random Forest 2 (After log transformation) | 0.2158021 | 0.8911 | 0.8908 | 0.04643798 | 0.04643798 |
| 6(b) | Random Forest 1 (varimp) (After log transformation) | 0.2158021 | 0.8901 | 0.8898 | 0.04643798 | 0.04643798 |
| 7(a) | Variables from ANOVA (Without transformation) | 4153.279 | 0.6359 | 0.6346 | 17189294 | 17189294 |
| 7(b) | After sqrt transformation | 13.97293 | 0.7947 | 0.794 | 194.5588 | 194.5588 |
| 7(c) | After log transformation | 0.2152009 | 0.8917 | 0.8914 | 0.04615424 | 0.04615424 |

Scaled

# CONCLUSIONS

Various models were tried out and their performance measures tabulated as given above. The models were modelled without following feature selection methods, initially, and, later, using feature selection methods. A summary of the features selected in various models is as shown in a subsequent section.

## Without Feature selection and any feature engineering :-

1. If outliers are not removed, then the performance is extremely poor, even after removing the least significant features.
2. If outliers are removed from the indicator variables, then the performance of Adj R2improves to approx. 63%. There is hardly any effect in the Adj R2 value when insignificant variables are removed and even after binning is effected. Removal of outliers results in removal of approx. 8% of the records being removed.
3. When outliers are removed from the respondent variable, clv, the performance dramatically improves to 93%. There is no change in this figure even after binning. Removal of outliers results in removal of approx. 12% of the records being removed.
4. When outliers are removed from the respondent variable, clv, there are outliers still available in Total Claim Amount and Monthly Premium Auto. Removal of outliers from these features results in removal of a total of 16.88% of records. The performance, however, varies only from the third decimal point onwards wrt the case in point 3 above, and hence there is marginal improvement in performance with all outliers removed.
5. However, steps 3 and 4 were only to check the effect of removing outliers from the respondent variable. This is not being followed.

## With Feature Selection and/or sqrt/log transformation:

1. The Adj R2 value remains around 63% irrespective of whether outliers are removed or not and whether binning is done or not.Therefore, we retain all outliers as indicators of variation in the data and do not carry out binning. Instead, we carry out scaling of the data.
2. When we apply sqrt transformation to the respondent variable the Adj R2 value goes up to 79 – 80% approx.
3. When we apply log transformation to the respondent variable, we obtain Adj R2 values in the region of 89-90%.
4. The best values are obtained in the model where all the features are taken and the log transformation is applied to the respondent variable.
5. “Monthly Premium Auto” was omitted from modelling due to its correlation with “Total Claim Amount”.

## Pertinent Take-aways:

1. Binning has negligible effect on performance.
2. Transformation of Respondent variable (sqrt/Log) has a significant improvement in performance vis-à-vis the non transformed variants.
3. Removal of outliers improved performance, but also caused significant loss of data. Hence, outliers were retained.

4. Removal of least significant features hardly caused an improvement in performance. Hence, feature selection techniques were employed.
5. Converting “Number of Open Complaints” and “Number of Policies” to factors improved accuracy of the models.

# Best Model

The best model was one which used all the features and had the log transformation applied to the respondent variable.

## Code for the model with the best performance

```
df<- read.csv("C:\\Users\\HP\\Downloads\\Marketing-Customer-Value-Analysis.csv")
str(df)
```

```
## '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" ...
```

```
glimpse(df)
```

```
## Rows: 9,134
## Columns: 24
## $ Customer                      <chr> "BU79786", "QZ44356", "AI49188", "WW6325~
## $ State                         <chr> "Washington", "Arizona", "Nevada", "Cali~
## $ Customer.Lifetime.Value       <dbl> 2763.519, 6979.536, 12887.432, 7645.862,~
## $ Response                      <chr> "No", "No", "No", "No", "No", "Yes", "Ye~
## $ Coverage                      <chr> "Basic", "Extended", "Premium", "Basic",~
## $ Education                     <chr> "Bachelor", "Bachelor", "Bachelor", "Bac~
## $ Effective.To.Date             <chr> "2/24/11", "1/31/11", "2/19/11", "1/20/1~
## $ EmploymentStatus              <chr> "Employed", "Unemployed", "Employed", "U~
## $ Gender                        <chr> "F", "F", "F", "M", "M", "F", "F", "M", ~
## $ Income                        <int> 56274, 0, 48767, 0, 43836, 62902, 55350,~
## $ Location.Code                 <chr> "Suburban", "Suburban", "Suburban", "Sub~
## $ Marital.Status                <chr> "Married", "Single", "Married", "Married~
## $ Monthly.Premium.Auto          <int> 69, 94, 108, 106, 73, 69, 67, 101, 71, 9~
## $ Months.Since.Last.Claim       <int> 32, 13, 18, 18, 12, 14, 0, 0, 13, 17, 23~
## $ Months.Since.Policy.Inception <int> 5, 42, 38, 65, 44, 94, 13, 68, 3, 7, 5, ~
## $ Number.of.Open.Complaints     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1~
## $ Number.of.Policies            <int> 1, 8, 2, 7, 1, 2, 9, 4, 2, 8, 3, 3, 3, 8~
## $ Policy.Type                   <chr> "Corporate Auto", "Personal Auto", "Pers~
## $ Policy                        <chr> "Corporate L3", "Personal L3", "Personal~
## $ Renew.Offer.Type              <chr> "Offer1", "Offer3", "Offer1", "Offer1", ~
## $ Sales.Channel                 <chr> "Agent", "Agent", "Agent", "Call Center"~
## $ Total.Claim.Amount            <dbl> 384.81115, 1131.46493, 566.47225, 529.88~
## $ Vehicle.Class                 <chr> "Two-Door Car", "Four-Door Car", "Two-Do~
## $ Vehicle.Size                  <chr> "Medsize", "Medsize", "Medsize", "Medsiz~
```

```
#-----------------------------------Min Max normalization all numeric variables (Scaling)----------------------------#

#Income
df$Income<- (df$Income-min(df$Income))/(max(df$Income)-min(df$Income))
```

```
#Months.Since.Last.Claim
df$Months.Since.Last.Claim<- (df$Months.Since.Last.Claim-min(df$Months.Since.Last.Claim))/(max(df$Months.Since.Last.Claim)-m
in(df$Months.Since.Last.Claim))
```

```
#Months.Since.Policy.Inception
df$Months.Since.Policy.Inception<- (df$Months.Since.Policy.Inception-min(df$Months.Since.Policy.Inception))/(max(df$Months.Since.Policy.Inception)-min(df$Months.Since.Policy.Inception))
```

```
#Total.Claim.Amount
df$Total.Claim.Amount<- (df$Total.Claim.Amount-min(df$Total.Claim.Amount))/(max(df$Total.Claim.Amount)-min(df$Total.Claim.Amount))
```

```
#converting categorical feauters to factors.
df$State <- as.factor(df$State)
df$Response <- as.factor(df$Response)
df$Coverage <- as.factor(df$Coverage)
df$Education <- as.factor(df$Education)
df$EmploymentStatus <- as.factor(df$EmploymentStatus)
df$Gender <- as.factor(df$Gender)
df$Location.Code  <- as.factor(df$Location.Code)
df$Marital.Status  <- as.factor(df$Marital.Status)
df$Policy.Type  <- as.factor(df$Policy.Type)
df$Renew.Offer.Type  <- as.factor(df$Renew.Offer.Type)
df$Policy  <- as.factor(df$Policy)
df$Sales.Channel  <- as.factor(df$Sales.Channel)
df$Vehicle.Class  <- as.factor(df$Vehicle.Class)
df$Vehicle.Size  <- as.factor(df$Vehicle.Size)
```

```
#Converting no. of open complaints and policies also to factor.
df$Number.of.Open_Complaints <- as.factor(df$Number.of.Open.Complaints)
df$Number.of.Policies <- as.factor(df$Number.of.Policies)
str(df)
```

```
## 'data.frame':    9134 obs. of  25 variables:
##  $ Customer                     : chr  "BU79786" "QZ44356" "AI49188" "WW63253" ...
##  $ State                        : Factor w/ 5 levels "Arizona","California",..: 5 1 3 2 5 4 4 1 4 4 ...
##  $ Customer.Lifetime.Value      : num  2764 6980 12887 7646 2814 ...
##  $ Response                     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
##  $ Coverage                     : Factor w/ 3 levels "Basic","Extended",..: 1 2 3 1 1 1 1 3 1 2 ...
##  $ Education                    : Factor w/ 5 levels "Bachelor","College",..: 1 1 1 1 1 1 2 5 1 2 ...
##  $ Effective.To.Date            : chr  "2/24/11" "1/31/11" "2/19/11" "1/20/11" ...
##  $ EmploymentStatus             : Factor w/ 5 levels "Disabled","Employed",..: 2 5 2 5 2 2 2 5 3 2 ...
##  $ Gender                       : Factor w/ 2 levels "F","M": 1 1 1 2 2 1 1 2 2 1 ...
##  $ Income                       : num  0.563 0 0.488 0 0.438 ...
##  $ Location.Code                : Factor w/ 3 levels "Rural","Suburban",..: 2 2 2 2 1 1 2 3 2 3 ...
##  $ Marital.Status               : Factor w/ 3 levels "Divorced","Married",..: 2 3 2 2 3 2 2 3 1 2 ...
##  $ Monthly.Premium.Auto         : int  69 94 108 106 73 69 67 101 71 93 ...
##  $ Months.Since.Last.Claim      : num  0.914 0.371 0.514 0.514 0.343 ...
##  $ Months.Since.Policy.Inception: num  0.0505 0.4242 0.3838 0.6566 0.4444 ...
##  $ Number.of.Open.Complaints    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Number.of.Policies           : Factor w/ 9 levels "1","2","3","4",..: 1 8 2 7 1 2 9 4 2 8 ...
##  $ Policy.Type                  : Factor w/ 3 levels "Corporate Auto",..: 1 2 2 1 2 2 1 1 1 3 ...
##  $ Policy                       : Factor w/ 9 levels "Corporate L1",..: 3 6 6 2 4 6 3 3 3 8 ...
##  $ Renew.Offer.Type             : Factor w/ 4 levels "Offer1","Offer2",..: 1 3 1 1 1 2 1 1 1 2 ...
##  $ Sales.Channel                : Factor w/ 4 levels "Agent","Branch",..: 1 1 1 3 1 4 1 1 1 2 ...
##  $ Total.Claim.Amount           : num  0.133 0.3911 0.1958 0.1831 0.0477 ...
##  $ Vehicle.Class                : Factor w/ 6 levels "Four-Door Car",..: 6 1 6 5 1 6 1 1 1 1 ...
##  $ Vehicle.Size                 : Factor w/ 3 levels "Large","Medsize",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Number.of.Open_Complaints    : Factor w/ 6 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
#-----------------------------------Log transformation only on CLV---------------------------#


df$Customer.Lifetime.Value=log(df$Customer.Lifetime.Value)
```

```
#-----------------------------------Splitting the data to test and train.---------------------------#


split <- sample.split(df, SplitRatio = 0.7)
split
```

```
##  [1] FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
## [13]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE
## [25]  TRUE
```

```
train <- subset(df, split="true")
test <-subset(df, split="false")
train
```

| | Customer | State | Customer.Lifetime.Value | Response | Coverage | Education | ▸ |
|---|---|---|---|---|---|---|---|
| | <chr> | <fct> | <dbl> | <fct> | <fct> | <fct> | |
| 1 | BU79786 | Washington | 7.924260 | No | Basic | Bachelor | |
| 2 | QZ44356 | Arizona | 8.850738 | No | Extended | Bachelor | |
| 3 | AI49188 | Nevada | 9.464008 | No | Premium | Bachelor | |
| 4 | WW63253 | California | 8.941920 | No | Basic | Bachelor | |
| 5 | HB64268 | Washington | 7.942253 | No | Basic | Bachelor | |
| 6 | OC83172 | Oregon | 9.018732 | Yes | Basic | Bachelor | |
| 7 | XZ87318 | Oregon | 8.590611 | Yes | Basic | College | |
| 8 | CF85061 | Arizona | 8.884070 | No | Premium | Master | |
| 9 | DY87989 | Oregon | 10.091108 | Yes | Basic | Bachelor | |
| 10 | BQ94931 | Oregon | 8.907636 | No | Extended | College | |

1-10 of 9,134 rows | 1-7 of 26 columns Previous **1** 2 3 4 5 6 ... 914 Next

```
#Training the model with normalized data columns and log transformed clv
fit3<- lm(Customer.Lifetime.Value ~     State+Response+Coverage+
           Education+EmploymentStatus+Gender+
           Income+Location.Code+Marital.Status+
           Months.Since.Last.Claim+Months.Since.Policy.Inception+
           Number.of.Open.Complaints+Number.of.Policies+Policy+            Renew.Offer.Type+Sales.Channel+Total.Claim.Amount
+Vehicle.Class+Vehicle.Size , data=train)

summary(fit3)
```

```
## 
## Call:
## lm(formula = Customer.Lifetime.Value ~ State + Response + Coverage + 
##     Education + EmploymentStatus + Gender + Income + Location.Code + 
##     Marital.Status + Months.Since.Last.Claim + Months.Since.Policy.Inception + 
##     Number.of.Open.Complaints + Number.of.Policies + Policy + 
##     Renew.Offer.Type + Sales.Channel + Total.Claim.Amount + Vehicle.Class + 
##     Vehicle.Size, data = train)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.49334 -0.08354 -0.00878  0.05855  0.95022 
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    7.809e+00  2.100e-02 371.809  < 2e-16 ***
## StateCalifornia                2.396e-03  6.464e-03   0.371  0.71087    
## StateNevada                    1.583e-02  8.916e-03   1.776  0.07582 .  
## StateOregon                    4.360e-03  6.690e-03   0.652  0.51456    
## StateWashington                9.225e-03  9.213e-03   1.001  0.31669    
## ResponseYes                   -5.999e-04  7.085e-03  -0.085  0.93253    
## CoverageExtended               2.171e-01  5.224e-03  41.550  < 2e-16 ***
## CoveragePremium                4.225e-01  8.906e-03  47.437  < 2e-16 ***
## EducationCollege              -3.350e-03  5.843e-03  -0.573  0.56642    
## EducationDoctor                2.352e-02  1.239e-02   1.899  0.05766 .  
## EducationHigh School or Below  1.313e-02  5.920e-03   2.219  0.02654 *  
## EducationMaster                2.656e-02  8.972e-03   2.960  0.00308 ** 
## EmploymentStatusEmployed       6.501e-02  1.210e-02   5.371 8.04e-08 ***
## EmploymentStatusMedical Leave  3.594e-02  1.489e-02   2.414  0.01581 *  
## EmploymentStatusRetired       -1.364e-03  1.717e-02  -0.079  0.93666    
## EmploymentStatusUnemployed    -1.294e-02  1.221e-02  -1.060  0.28912    
## GenderM                       -2.956e-02  4.535e-03  -6.518 7.51e-11 ***
## Income                         3.550e-02  1.318e-02   2.694  0.00708 ** 
## Location.CodeSuburban         -2.811e-02  8.888e-03  -3.163  0.00157 ** 
## Location.CodeUrban            -1.260e-02  8.196e-03  -1.537  0.12429    
## Marital.StatusMarried          8.239e-03  6.650e-03   1.239  0.21536    
## Marital.StatusSingle          -3.234e-02  7.696e-03  -4.202 2.67e-05 ***
## Months.Since.Last.Claim        1.607e-02  7.829e-03   2.053  0.04011 *  
## Months.Since.Policy.Inception  3.052e-03  8.042e-03   0.380  0.70430    
## Number.of.Open.Complaints     -2.071e-02  2.474e-03  -8.372  < 2e-16 ***
## Number.of.Policies2            1.403e+00  5.898e-03 237.848  < 2e-16 ***
## Number.of.Policies3            6.955e-01  7.367e-03  94.418  < 2e-16 ***
## Number.of.Policies4            6.982e-01  1.131e-02  61.749  < 2e-16 ***
## Number.of.Policies5            7.016e-01  1.133e-02  61.898  < 2e-16 ***
## Number.of.Policies6            6.923e-01  1.180e-02  58.678  < 2e-16 ***
## Number.of.Policies7            6.942e-01  1.104e-02  62.893  < 2e-16 ***
## Number.of.Policies8            6.986e-01  1.162e-02  60.142  < 2e-16 ***
## Number.of.Policies9            7.029e-01  1.122e-02  62.649  < 2e-16 ***
## PolicyCorporate L2            -1.686e-02  1.435e-02  -1.175  0.24008    
## PolicyCorporate L3            -5.712e-03  1.318e-02  -0.433  0.66483    
## PolicyPersonal L1             -9.331e-03  1.287e-02  -0.725  0.46852    
## PolicyPersonal L2             -4.154e-03  1.225e-02  -0.339  0.73448    
## PolicyPersonal L3             -6.310e-03  1.191e-02  -0.530  0.59631    
## PolicySpecial L1              -2.324e-02  2.879e-02  -0.807  0.41955    
## PolicySpecial L2               1.791e-03  2.025e-02   0.088  0.92955    
## PolicySpecial L3               3.217e-02  2.097e-02   1.534  0.12503    
## Renew.Offer.TypeOffer2         6.280e-03  5.671e-03   1.107  0.26812    
## Renew.Offer.TypeOffer3         6.948e-03  6.841e-03   1.016  0.30982    
## Renew.Offer.TypeOffer4         2.589e-03  7.988e-03   0.324  0.74585    
## Sales.ChannelBranch            8.831e-03  5.625e-03   1.570  0.11646    
## Sales.ChannelCall Center       5.967e-03  6.358e-03   0.939  0.34801    
## Sales.ChannelWeb               8.507e-06  7.034e-03   0.001  0.99904    
## Total.Claim.Amount             2.261e-01  4.544e-02   4.977 6.59e-07 ***
## Vehicle.ClassLuxury Car        9.240e-01  2.049e-02  45.098  < 2e-16 ***
## Vehicle.ClassLuxury SUV        9.440e-01  1.940e-02  48.667  < 2e-16 ***
## Vehicle.ClassSports Car        4.632e-01  1.077e-02  43.030  < 2e-16 ***
## Vehicle.ClassSUV               4.374e-01  6.845e-03  63.895  < 2e-16 ***
## Vehicle.ClassTwo-Door Car      3.066e-03  5.876e-03   0.522  0.60182    
## Vehicle.SizeMedsize            4.447e-03  7.500e-03   0.593  0.55324    
## Vehicle.SizeSmall              4.554e-03  8.739e-03   0.521  0.60230    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2142 on 9079 degrees of freedom
## Multiple R-squared:  0.893,  Adjusted R-squared:  0.8924 
## F-statistic:  1403 on 54 and 9079 DF,  p-value: < 2.2e-16
```

```
#Finding RSE
sigma(fit3)
```

```
## [1] 0.214212
```

```
# computing test MSE
test %>%
  add_predictions(fit3) %>%
  summarise(MSE = mean((Customer.Lifetime.Value - pred)^2))
```

| **MSE** |
|---:|
| <dbl> |
| 0.04561048 |

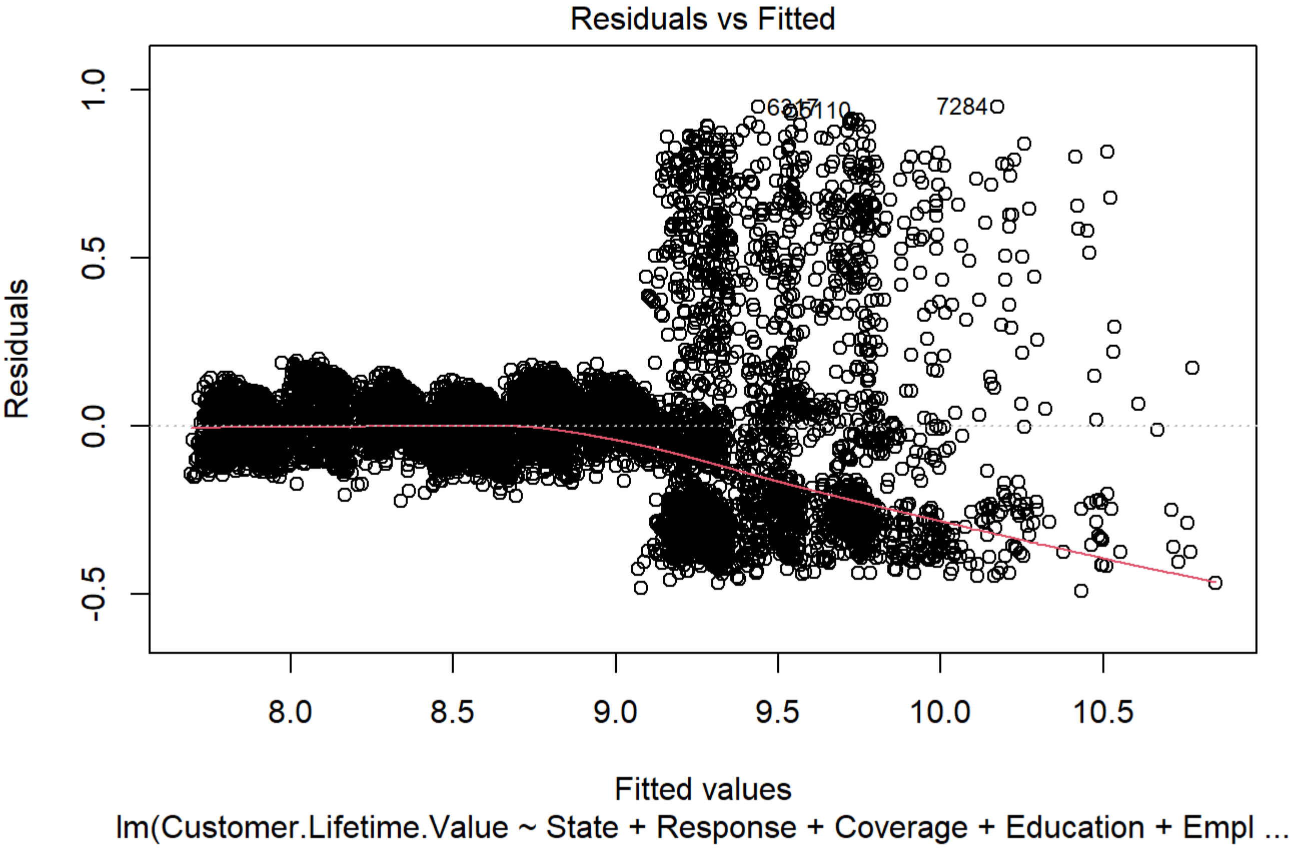
1 row

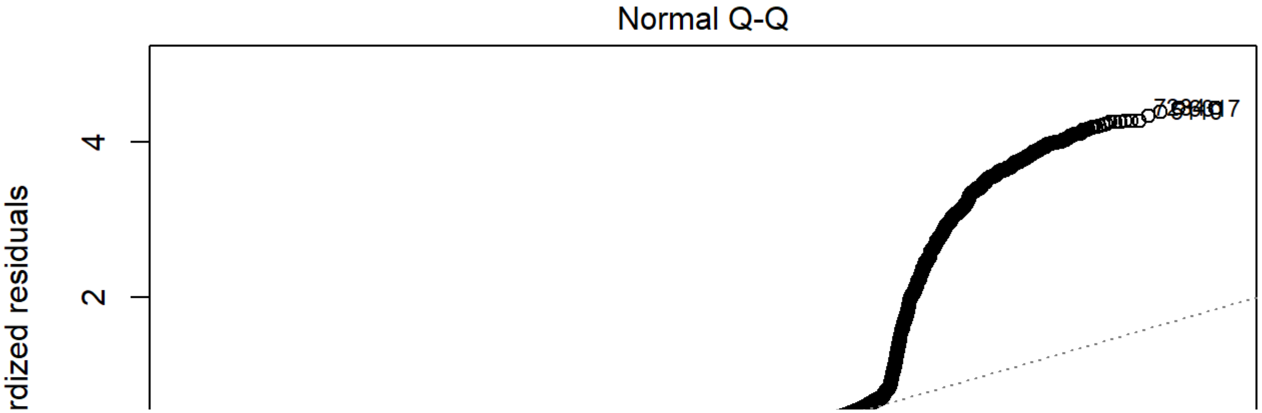
```
# computing train MSE
train %>%
  add_predictions(fit3) %>%
  summarise(MSE = mean((Customer.Lifetime.Value - pred)^2))
```

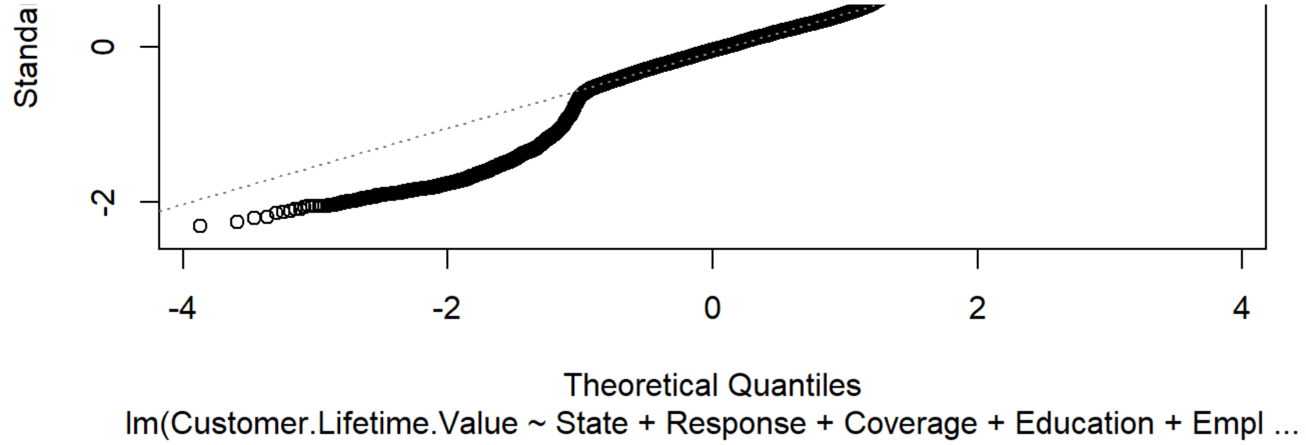
| **MSE** |
|---:|
| <dbl> |
| 0.04561048 |

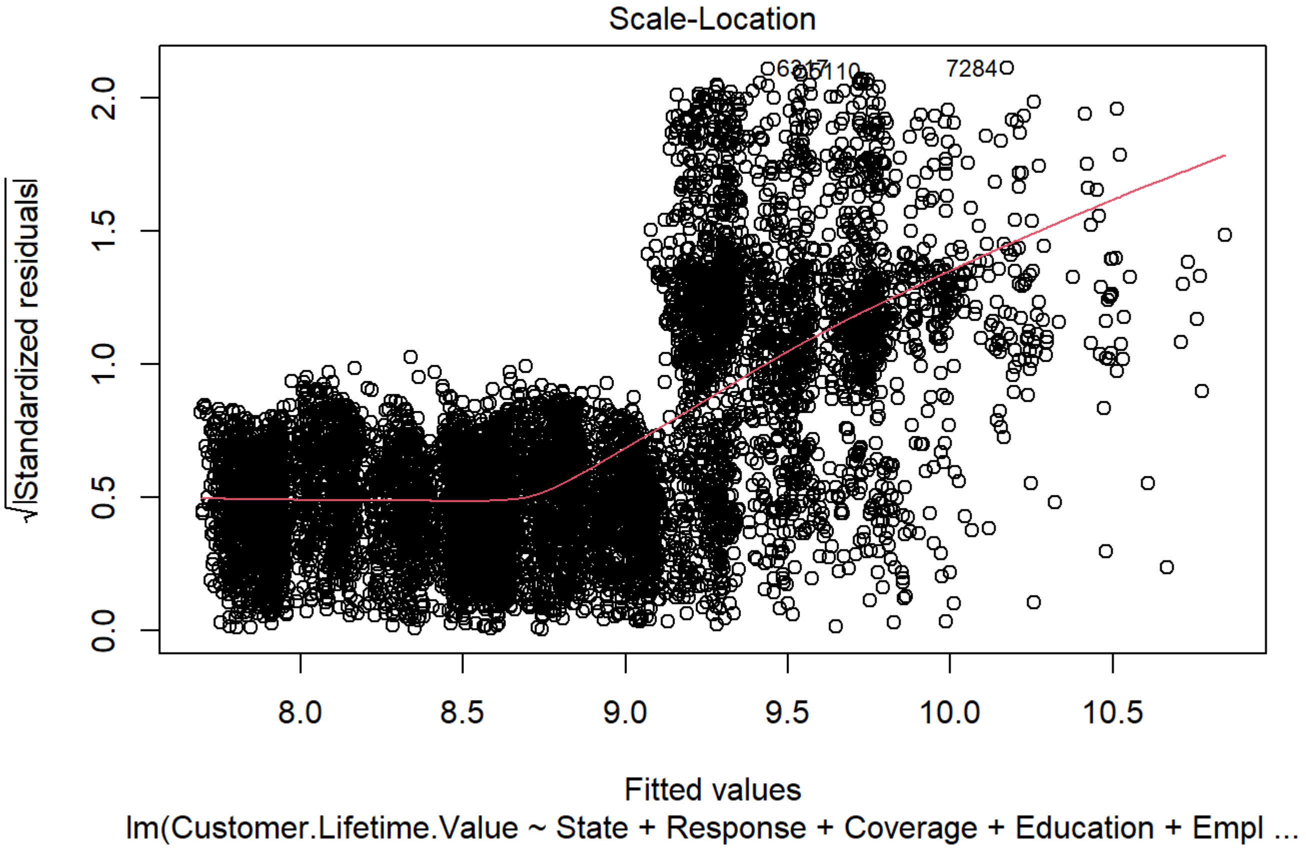
1 row

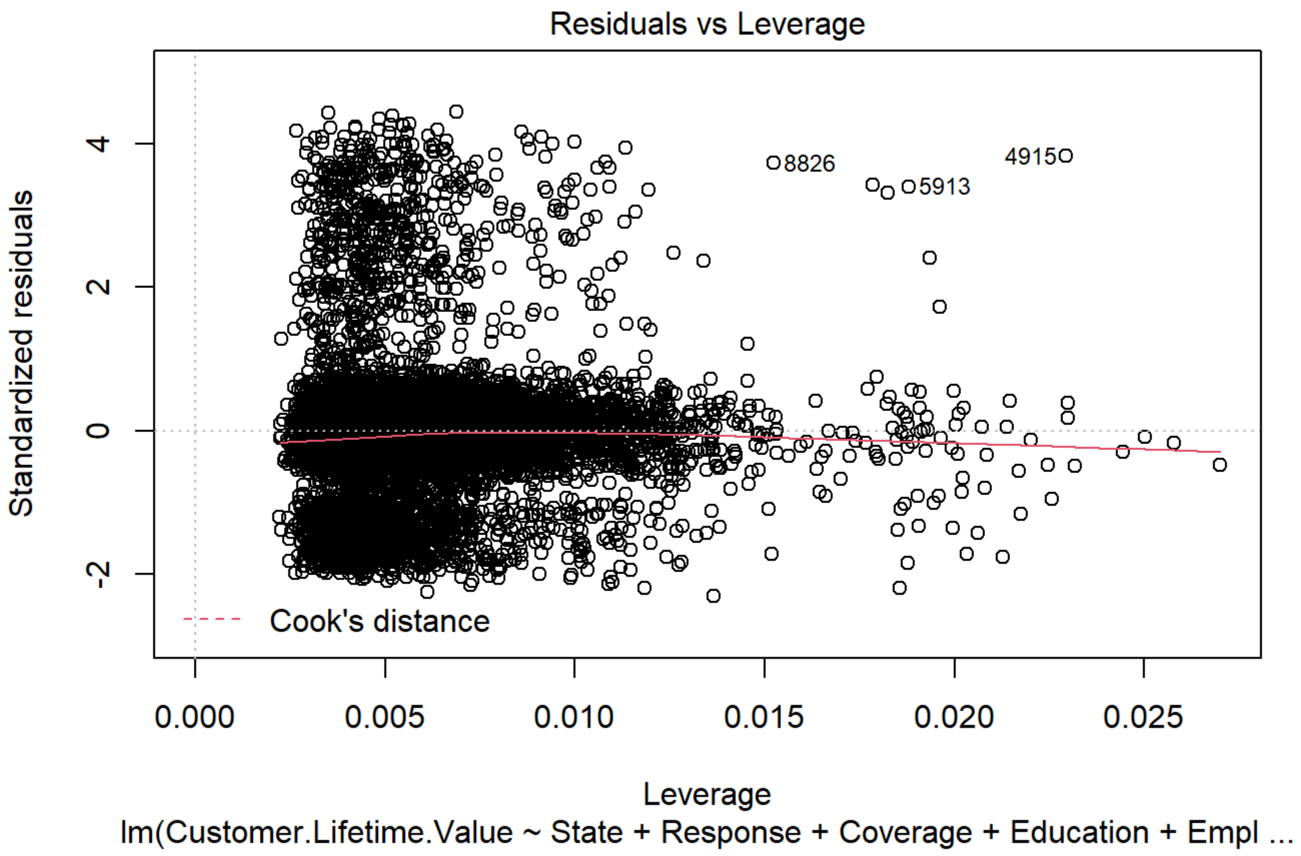
```
#Plotting the model.
plot(fit3)
```


Residuals vs Fitted
Residuals
1.0
0.5
0.0
-0.5
6317
5110
7284
8.0
8.5
9.0
9.5
10.0
10.5
Fitted values
lm(Customer.Lifetime.Value ~ State + Response + Coverage + Education + Empl ...


Normal Q-Q
rdized residuals
4
2

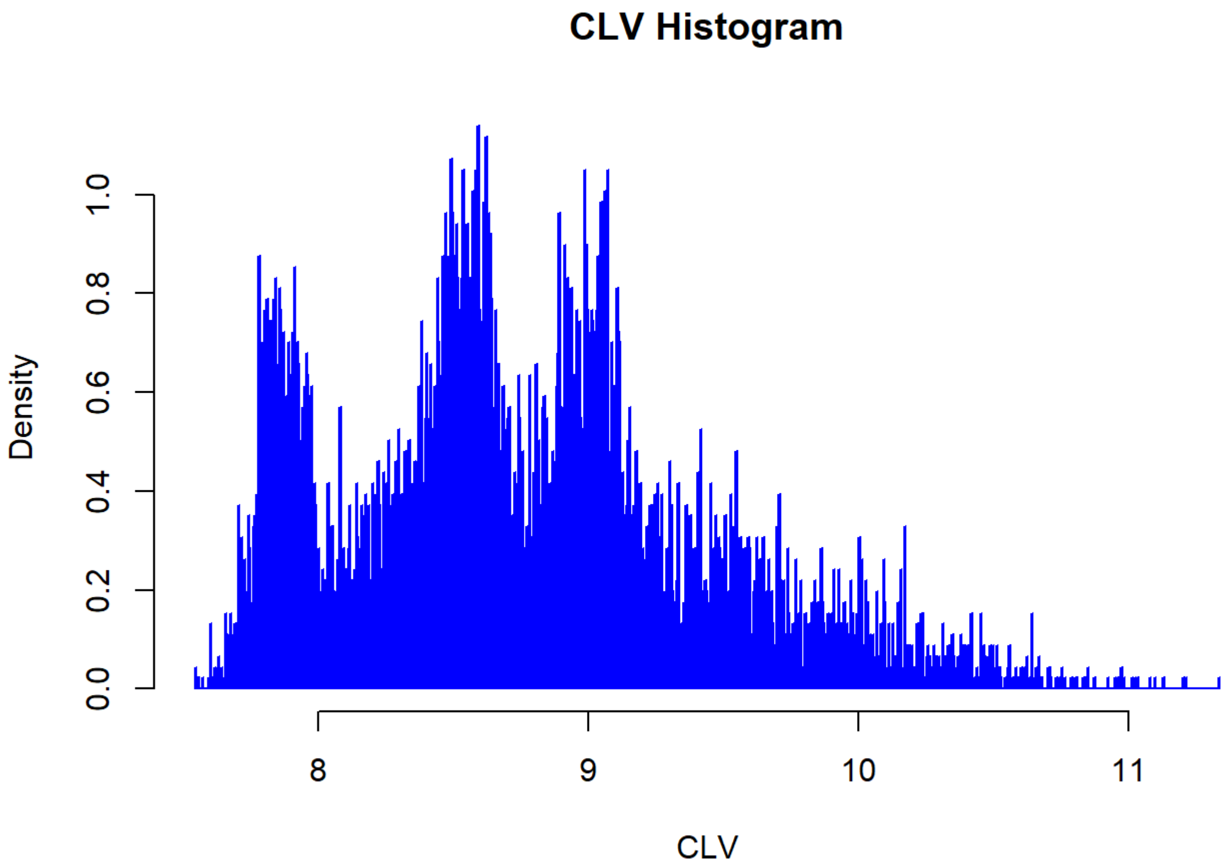

Standa
0
-2
-4
-2
0
2
4
Theoretical Quantiles
lm(Customer.Lifetime.Value ~ State + Response + Coverage + Education + Empl ...


Scale-Location
6317
6110
7284
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
8.0
8.5
9.0
9.5
10.0
10.5
Fitted values
lm(Customer.Lifetime.Value ~ State + Response + Coverage + Education + Empl ...


Residuals vs Leverage
Standardized residuals
4
2
0
-2
8826
4915
5913
Cook's distance
0.000
0.005
0.010
0.015
0.020
0.025
Leverage
lm(Customer.Lifetime.Value ~ State + Response + Coverage + Education + Empl ...

```
##################################### Checking of Assumption ############################################


# 1. In the residual vs fitted graph we cannot see any funnel shape in the residues, hence the assumption of homoskedasticity is satisfied.

# 2. The points in the center part of the graphs follow the Q-Q plot. The trailing portion deviates from the Q-Q plot by a small amount. However, the leading portion deviates significantly from the Q-Q plot indicating non adherence to normality. Therefore, Log transformation has been applied to the target variable. The graph of the Log transformed target variable is displayed below. As we can see graph resembles the normal curve.

# 3. Residuals are spread equally along the ranges of predictors, indicating homoscedasticity. We can see a  horizontal line with equally (randomly) spread points.

# 4. Even though there seems to be extreme values, the regression line is more or less straight.
```

```
# Plot of Log transformed CLV
hist(df$Customer.Lifetime.Value,
breaks = 800,
freq = FALSE,
main = "CLV Histogram", xlab = "CLV", border = "Blue")
```



```
# Residuals should be uncorrelated.There should be no Autocorrelation.
# Null H0: residuals from a linear regression are uncorrelated.
# D-W Statistic should be close to 2.


durbinWatsonTest(fit3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     0.008965194      1.982047    0.34
##  Alternative hypothesis: rho != 0
```

```
#Since, the p-value is >0.05, we fail to reject H0: (No Autocorrelation)
```

```
# Checking multicollinearity

vif(fit3)
```

```
##                                   GVIF Df GVIF^(1/(2*Df))
## State                         1.022224  4        1.002751
## Response                      1.226011  1        1.107254
## Coverage                      1.291001  2        1.065937
## Education                     1.089738  4        1.010800
## EmploymentStatus              3.918643  4        1.186156
## Gender                        1.022844  1        1.011358
## Income                        3.190906  1        1.786311
## Location.Code                 2.625380  2        1.272911
## Marital.Status                1.328824  2        1.073660
## Months.Since.Last.Claim       1.010612  1        1.005292
## Months.Since.Policy.Inception 1.022789  1        1.011330
## Number.of.Open.Complaints     1.009651  1        1.004814
## Number.of.Policies            1.085342  8        1.005132
## Policy                        1.043708  8        1.002677
## Renew.Offer.Type              1.275012  3        1.041324
## Sales.Channel                 1.066915  3        1.010854
## Total.Claim.Amount            4.143505  1        2.035560
## Vehicle.Class                 2.236223  5        1.083806
## Vehicle.Size                  1.067929  2        1.016566
```

```
# The values of VIF should be within 2. And in no case it should be greater than 10.
# Since all values are from 1 to 4, absence of multicollinearity is witnessed.
```

```
#  After checking the assumption of the linear regression model we can say that the assumptions seems to be largely satisfie
d.
```

## Additional Data that could have better predicted the outcome variable

1.In order to better analyze the CLV of a customer, the company needs to make sure the customer stays with them. This can be calculated using “Customer Retention Rate” : Customer Retention Rate is the number of customers retained by a company over a certain time period. It’s expressed as a percentage of a company’s existing customers who remain loyal within that time frame. CRR = [(E-N)/S] x 100 Where : The number of existing customers at the start of the time period (S) The number of total customers at the end of the time period (E) The number of new customers added within the time period (N)

2.If a customer has made no claims for a period of n years, company can provide perks such as “no claim bonus”, which would essentially reduce the premium amount payable at next renewal, while keeping the insurance cover at the same or higher value. We have the data regarding “Months since last claim” , however if we have an additional data regarding customers response to opting for “no claim bonus” it would help us analyze customer preference . Hence, leading to better prediction of customers with consistent CLV

3. If the cost of sustaining Sales Channel was provided it would have helped us analyze the sustainability cost of each Sales Channel, with which we could have essentially found out which Sales Channel is contributing effectively to the CLV. Based on the outcome, we can suggest courses of action to the company reduce such overheads.

# Contribution of team members

To coordinate the team activities, formal meetings were held every day at 1200 hrs and 1900 hrs, besides informal meetings on teams, whatsapp calls, phone calls and innumerable chats at all hours of the day.

Initially everybody carried out EDA. The results of EDA were discussed. Further courses of action on EDA were discussed and carried out. Once EDA was satisfactorily done, feature selection and model building was carried out. Joel, Alex and Tashi carried out feature selection using Stepwise Regression. Sarah, Aishwarya and Bart carried out Feature selection using Random Forest and ANOVA. Sarah, Aishwarya and Bart carried out model building using Random Forest implementing various models with and without transformation as brought out in the report above. Joel, Alex and Tashi constructed various models using Regression, with and without transformation as well as, with and without scaling as brought out in the report above. Retention and deletion of outliers and effect of binning was also experimented with, in the models by both sub-teams.

Finally, having compared the various results and zeroing on the most operative model was done, the report was prepared drawing from the EDA and Model Building done earlier. All members sat together and constructed the report vetting all aspects of the report collectively.

All members were involved at every stage of the process from beginning to end.

# References

https://www.geeksforgeeks.org/random-forest-approach-in-r-programming/ (https://www.geeksforgeeks.org/random-forest-approach-in-r-programming/)

https://github.com/abhiyerasi/CLV-Auto-Insurance (https://github.com/abhiyerasi/CLV-Auto-Insurance)

http://www.sthda.com/english/wiki/two-way-anova-test-in-r (http://www.sthda.com/english/wiki/two-way-anova-test-in-r) http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r (http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r)

https://www.kaggle.com/juancarlosventosa/models-to-improve-customer-retention (https://www.kaggle.com/juancarlosventosa/models-to-improve-customer-retention)

https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80 (https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80)

https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why (https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why)

https://medium.com/@TheDataGyan/day-8-data-transformation-skewness-normalization-and-much-more-4c144d370e55 (https://medium.com/@TheDataGyan/day-8-data-transformation-skewness-normalization-and-much-more-4c144d370e55)

https://anshikaaxena.medium.com/how-skewed-data-can-skrew-your-linear-regression-model-accuracy-and-transfromation-can-help-62c6d3fe4c53 (https://anshikaaxena.medium.com/how-skewed-data-can-skrew-your-linear-regression-model-accuracy-and-transfromation-can-help-62c6d3fe4c53)

http://r-statistics.co/Variable-Selection-and-Importance-With-R.html (http://r-statistics.co/Variable-Selection-and-Importance-With-R.html)

https://dataaspirant.com/feature-selection-techniques-r/ (https://dataaspirant.com/feature-selection-techniques-r/)

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.