COURSE PROJECT: CUSTOMER BEHAVIOR ANALYSIS

AISHAT YUSUF ABDULGAFAR

TABLE OF CONTENTS

Contents

THE DATA

Dataset Description

The dataset is IBM customer data that leverages Watson Analytics, which may be used for descriptive or predictive analysis of customer behavior in order to retain customers. The data set contains 9,134 customer records with 24 variables. The dataset was gotten from Kaggle and the link to the dataset is below

IBM Watson Marketing Customer Value Data | Kaggle

Loading Libraries

Several packages were installed and loaded to begin the project using the code below. This script was updated with packages as needed as the project proceeded.

```{r}
install.packages("dplyr")
install.packages("ggplot2")
install.packages("reshape2")
library(dplyr)
library(ggplot2)
library(reshape2)
```

Loading the Dataset

The dataset was loaded from an Excel CSV file using the code below and the head () function was used to display the first 5 rows of the dataset as shown below.

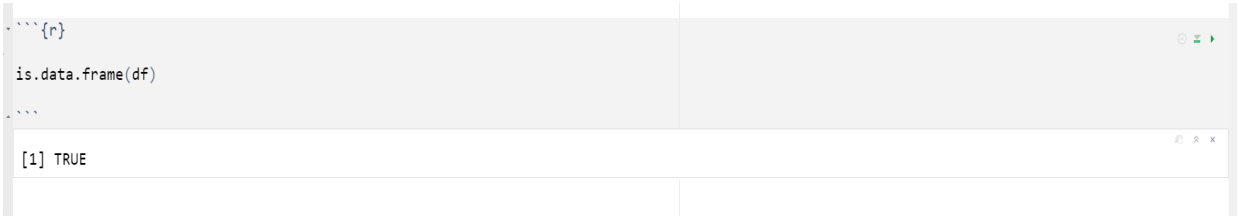
```{r}
df <- read.csv("C:/Users/Aishat/Downloads/WA_Fn-UseC_-Marketing-Customer-Value-Analysis.csv/WA_Fn-UseC_-Marketing-Customer-Value-Analysis.csv")
head(df)
```

Description: df [6 x 24]

	Customer <chr>	State <chr>	Customer.Lifetime.Value <dbl>	Response <chr>	Coverage <chr>	Education <chr>	Effective.To.Date <chr>	EmploymentStatus <chr>	Gender <chr>
1	BU79786	Washington	2763.519	No	Basic	Bachelor	2/24/11	Employed	F
2	QZ44356	Arizona	6979.536	No	Extended	Bachelor	1/31/11	Unemployed	F
3	AI49188	Nevada	12887.432	No	Premium	Bachelor	2/19/11	Employed	F
4	WW63253	California	7645.862	No	Basic	Bachelor	1/20/11	Unemployed	M
5	HB64268	Washington	2813.693	No	Basic	Bachelor	2/3/11	Employed	M
6	OC83172	Oregon	8256.298	Yes	Basic	Bachelor	1/25/11	Employed	F

6 rows | 1-10 of 24 columns

“is.data.frame” command was used to ensure that the dataset was properly loaded into R as a data frame.



```{r}
is.data.frame(df)
```

```
[1] TRUE
```

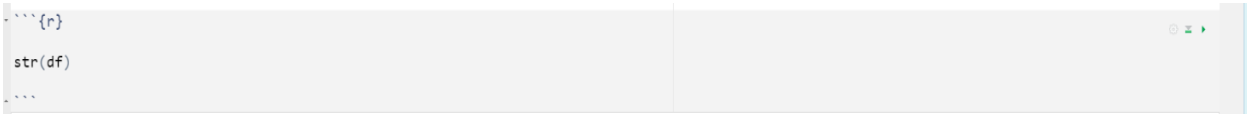
The dim () function describes the number of rows and columns present in the dataset. Here the dataset contains 9134 observations with 24 variables.

```{r}
dim(df)
```

```
[1] 9134   24
```

Original Data structures

The str () function was used to get the mode of the variables in the dataset. The above response tells that the original dataset has two modes, which include, num, int, and chr mode. With this, we will do data cleaning to our customer dataset.



```{r}
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" ...
```

colnames() function was used as seen below to display the list of the columns in the dataset.

```{r}
colnames(df)
```

```
 [1] "Customer"                  "State"                     "Customer.Lifetime.Value"
 [4] "Response"                  "Coverage"                  "Education"
 [7] "Effective.To.Date"         "EmploymentStatus"          "Gender"
[10] "Income"                    "Location.Code"             "Marital.Status"
[13] "Monthly.Premium.Auto"      "Months.Since.Last.Claim"   "Months.Since.Policy.Inception"
[16] "Number.of.Open.Complaints" "Number.of.Policies"        "Policy.Type"
[19] "Policy"                    "Renew.Offer.Type"          "Sales.Channel"
[22] "Total.Claim.Amount"        "Vehicle.Class"             "Vehicle.Size"
```

Data Cleaning

To begin, I examined the dataset to see if all of the columns had the same number of values. Specifically, whether there is no missing data. The code below was used to determine this.

```{r}
# Data Cleaning
colSums(sapply(df, is.na))
```

```
                  Customer                      State      Customer.Lifetime.Value
                         0                          0                            0
                  Response                   Coverage                    Education
                         0                          0                            0
         Effective.To.Date           EmploymentStatus                       Gender
                         0                          0                            0
                    Income              Location.Code                Marital.Status
                         0                          0                            0
      Monthly.Premium.Auto     Months.Since.Last.Claim Months.Since.Policy.Inception
                         0                          0                            0
Number.of.Open.Complaints          Number.of.Policies                   Policy.Type
                         0                          0                            0
                    Policy            Renew.Offer.Type                Sales.Channel
                         0                          0                            0
        Total.Claim.Amount               Vehicle.Class                 Vehicle.Size
                         0                          0                            0
```

```{r}
any(is.na(df))
```

```
[1] FALSE
```

According to the output above, there are no null values, so no further action is required to replace missing or null values. Using other functions as seen above.

The "State," "Coverage," and "Education" variables were saved as characters to facilitate simple modification during the study. Consequently, it has been handled as a string value. To convert the variable from a qualitative to a quantitative one, we will utilize the as.factor () function.

```{r}
df$State <- as.factor(df$State)
is.factor(df$State)
df$Coverage <- as.factor(df$Coverage)
is.factor(df$Coverage)
df$Education<- as.factor(df$Education)
is.factor(df$Education)
```

```
[1] TRUE
[1] TRUE
[1] TRUE
```

Cleaned Data Set

After the dataset has been cleaned, the str() function was then used again to provide the updated data structure of the dataset. And the head() function was used to display the first few rows of the dataset.

```{r}
str(df)
```

```
'data.frame':   9134 obs. of  24 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                     : chr  "No" "No" "No" "No" ...
 $ 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             : 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" ...
```

```{r}
head(df)
```

Description: df [6 x 24]

	Customer <chr>	State <fctr>	Customer.Lifetime.Value <dbl>	Response <chr>	Coverage <fctr>	Education <fctr>	Effective.To.Date <chr>	EmploymentStatus <chr>	Gender <chr>
1	BU79786	Washington	2763.519	No	Basic	Bachelor	2/24/11	Employed	F
2	QZ44356	Arizona	6979.536	No	Extended	Bachelor	1/31/11	Unemployed	F
3	AI49188	Nevada	12887.432	No	Premium	Bachelor	2/19/11	Employed	F
4	WW63253	California	7645.862	No	Basic	Bachelor	1/20/11	Unemployed	M
5	HB64268	Washington	2813.693	No	Basic	Bachelor	2/3/11	Employed	M
6	OC83172	Oregon	8256.298	Yes	Basic	Bachelor	1/25/11	Employed	F

6 rows | 1-10 of 24 columns

Description of the variables of the datasets

Below is a table describing each variable in the Crime-2001 to present dataset. These descriptions were gotten from the Chicago Data Portal where the original dataset was taken from as mentioned above.

Column Name	Mode	Description
Customer	Chr	Unique identifier for the Customer.
State	Chr	The state location of the customer.
Customer.Lifetime.Value	Num	The total worth of a customer.
Response	Chr	The customer's response to the offer.
Coverage	Chr	The class of the policy.
Education	Chr	Education qualification of the customer.
Effective. To.Date	Chr	The start date of the customer purchase.
EmploymentStatus	Chr	The employment status of the customer is employed/unemployed.
Gender	Chr	The gender details of the customer.
Income	Int	The per annum income of the customer.
Location.Code	Chr	The location level of the customer - Urban/Rural/Suburban.
Marital.Status	Chr	Marital status of the customer.
Monthly.Premium.Auto	Int	Auto loan monthly premium paid by the customer.
Months.Since.Last.Claim	Int	The number of months where the customer took the gap.
Months.Since.Policy.Inception	Int	The number of months since the policy was taken.
Number.Of.Open.Complaints	Int	Complaints were raised by the customer.

Number.Of.Policies	Int	Several policies issued to a customer.
Policy.Type	Chr	The category of the policy.
Policy	Chr	The category of the policy – L2/L3.
Renew.Offer.Type	Chr	Which offer type is used to renew the policy.
Sales.Channel	Chr	The channel by which the customer took the policy.
Total.Claim.Amount	Num	The claimed amount by the customer.
Vehicle.Class	Chr	The vehicle class details.
Vehicle.Size	Chr	The size of the vehicle.

Expectations

This analysis is anticipated to start by investigating the client data. I anticipate that my research will show patterns, trends, or connections between factors that might improve consumer behavior and retention. Throughout my analysis of the information, I want to get insight into consumer behavior that will help stakeholders make better data-driven decisions about customer retention and, most significantly, raise their organization's profits.

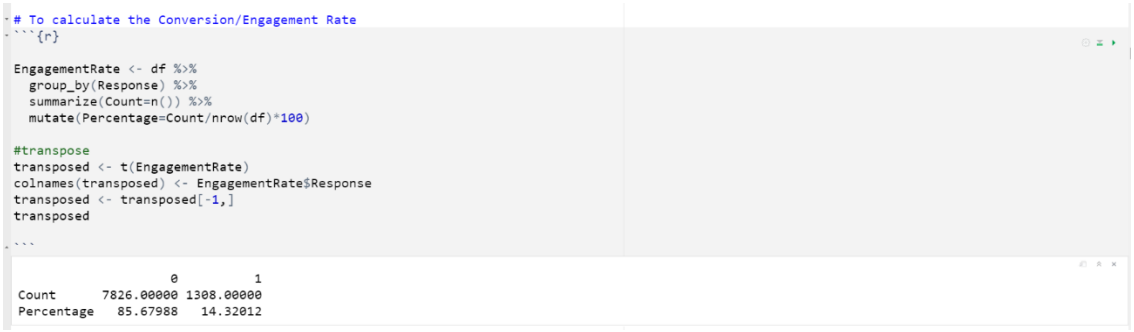
Based on my domain expertise, I anticipate that there should be a correlation between a customer's total amount of claims, which is a dependent variable, and their income, customer lifetime value, monthly auto premium, as well as the class of insurance (coverage) that they choose.

DATA ANALYSIS

Encoding Dataset “Response”

I started my analysis by encoding the “Response” from the customer as 0 and 1 instead of the initial “Yes” and “No” in the data frame for better analysis and manipulation. The code and output are shown below.

```
> df <- df %>%
+   mutate(Response = ifelse(Response == "No", 0, 1))
> df$Response <- as.integer(df$Response)
> head(df)
>
```

```
# ANALYSIS
# Encode conversions (default) as 0 and 1 (instead of "Yes" and "No")
```{r}

df <- df %>%
  mutate(Response = ifelse(Response == "No", 0, 1))
df$Response <- as.integer(df$Response)
head(df)
```
```

Description: df [6 x 24]

	Customer	State	Customer.Lifetime.Value	Response	Coverage	Education	Effective.To.Date	EmploymentStatus	Gender
	<chr>	<fctr>	<dbl>	<int>	<fctr>	<fctr>	<chr>	<chr>	<chr>
1	BU79786	Washington	2763.519	0	Basic	Bachelor	2/24/11	Employed	F
2	QZ44356	Arizona	6979.536	0	Extended	Bachelor	1/31/11	Unemployed	F
3	AI49188	Nevada	12887.432	0	Premium	Bachelor	2/19/11	Employed	F
4	WW63253	California	7645.862	0	Basic	Bachelor	1/20/11	Unemployed	M
5	HB64268	Washington	2813.693	0	Basic	Bachelor	2/3/11	Employed	M
6	OC83172	Oregon	8256.298	1	Basic	Bachelor	1/25/11	Employed	F

6 rows | 1-10 of 24 columns

Response rate Conversion

The code below is used to determine the response rate and the percentage ratio; non-responding customers have a higher count of 7826, which is 85.68%, than responding customers, who have a count of 1308, which is equal to 14.32%.

```
> EngagementRate <- df %>%
+   group_by(Response) %>%
+   summarize(Count=n()) %>%
+   mutate(Percentage=Count/nrow(df)*100)
>
> #transpose
> transposed <- t(EngagementRate)
```

> colnames(transposed) <- EngagementRate$Response

> transposed <- transposed[-1,]

> transposed



```
# To calculate the Conversion/Engagement Rate
```{r}

EngagementRate <- df %>%
  group_by(Response) %>%
  summarize(Count=n()) %>%
  mutate(Percentage=Count/nrow(df)*100)

#transpose
transposed <- t(EngagementRate)
colnames(transposed) <- EngagementRate$Response
transposed <- transposed[-1,]
transposed

```
```
                    0          1
Count      7826.00000 1308.00000
Percentage   85.67988   14.32012
```

I further analyzed by plotting a pie chart to look at the Response rate across the different sales channels. As you can see from the chart below, more than half of the customers that respond came from the Agents, while the non-responding customers are more evenly distributed across all sales channels.

>

> SalesChannel <- df %>%

+ group_by(Response, Sales.Channel) %>%

+ summarize(Count=n()) %>%

+ arrange(Sales.Channel)

>

> options(repr.plot.width = 20, repr.plot.height = 10)

> ggplot(SalesChannel, aes(x="", y=Count, fill=Sales.Channel)) +

+ geom_bar(stat = "identity", position = position_fill()) +

+ geom_text(aes(x=1.0, label=Count), position = position_fill(vjust=0.5)) +

+ coord_polar("y") + facet_wrap(~Response) +

+ ggtitle("Sales Channel (No, Yes)") +

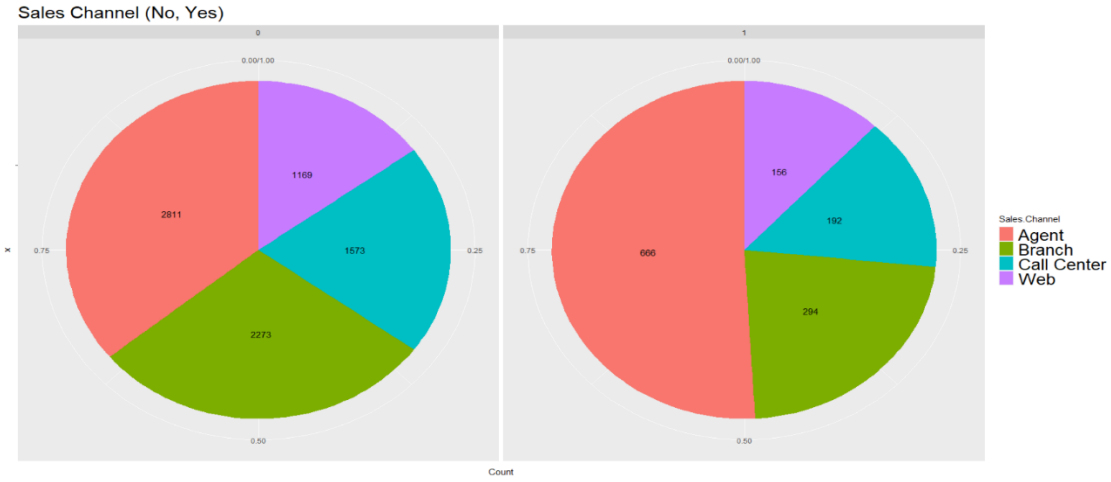
+ theme(legend.position = "right", legend.text=element_text(size=20), plot.title = element_text(size=22))

>
```

```
To Take a look at the Response rates by the different sales channels.
```{r}

SalesChannel <- df %>%
  group_by(Response, Sales.Channel) %>%
  summarize(Count=n()) %>%
  arrange(Sales.Channel)

options(repr.plot.width = 20, repr.plot.height = 10)
ggplot(SalesChannel, aes(x="", y=Count, fill=Sales.Channel)) +
  geom_bar(stat = "identity", position = position_fill()) +
  geom_text(aes(x=1.0, label=Count), position = position_fill(vjust=0.5)) +
  coord_polar("y") + facet_wrap(~Response) +
  ggtitle("Sales Channel (No, Yes)") +
  theme(legend.position = "right", legend.text=element_text(size=20), plot.title = element_text(size=22))

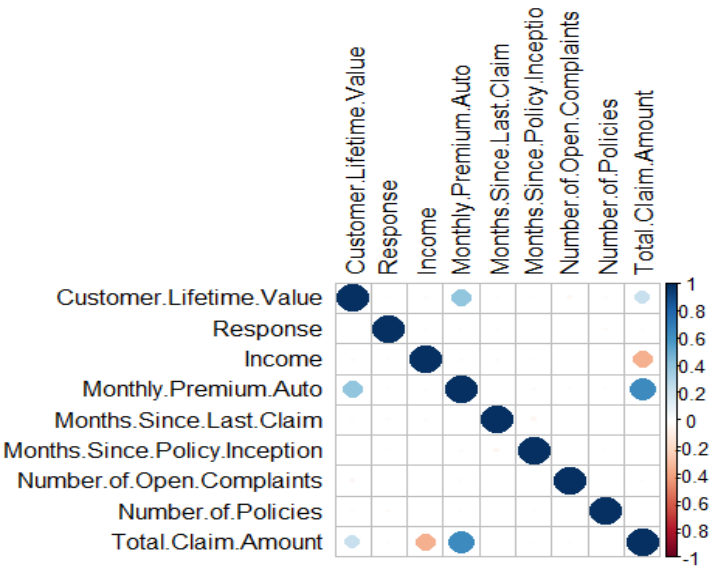
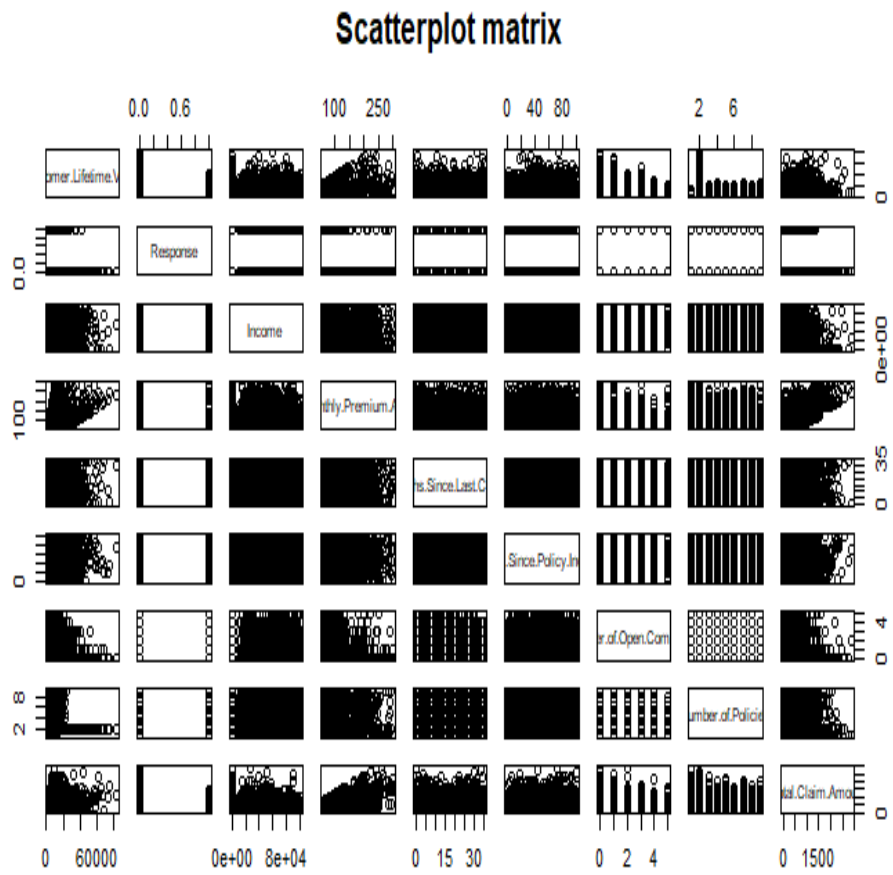
```
```



Response Rate across the "State"

Analysis to determine the response rate across the "State". From the result of the code below, we can say that the customers from Washington and Nevada response rates are less with a total count of 798 and 882 respectively than customers from other states.

```
>
> conversionsState <- df %>%
+   group_by(State) %>%
+   summarize(TotalCount=n(), NumConversions=sum(Response)) %>%
+   mutate(ConversionRate=NumConversions/TotalCount*100)
> conversionsState
>
```

```r
# Customer Conversion Rates by the States.
```{r}

conversionsState <- df %>%
  group_by(State) %>%
  summarize(TotalCount=n(), NumConversions=sum(Response)) %>%
  mutate(ConversionRate=NumConversions/TotalCount*100)
conversionsState

```
```

A tibble: 5 x 4

State	TotalCount	NumConversions	ConversionRate
Arizona	1703	243	14.26894
California	3150	456	14.47619
Nevada	882	124	14.05896
Oregon	2601	376	14.45598
Washington	798	109	13.65915

5 rows

Continuous Variable Plotting

> numeric1 <-select_if(df, is.numeric)

> numeric1

> cor(numeric1)

```
> numeric1
> cor(numeric1)
                          Customer.Lifetime.Value     Response        Income Monthly.Premium.Auto
Customer.Lifetime.Value                1.000000000 -0.008929582  0.0243656607          0.396261738
Response                              -0.008929582  1.000000000  0.0119322493          0.010966263
Income                                 0.024365661  0.011932249  1.0000000000         -0.016664550
Monthly.Premium.Auto                   0.396261738  0.010966263 -0.0166645503          1.000000000
Months.Since.Last.Claim                0.011516682 -0.016597147 -0.0267151791          0.005026285
Months.Since.Policy.Inception          0.009418381  0.002951561 -0.0008751978          0.020256643
Number.of.Open.Complaints             -0.036343193 -0.009881182  0.0064082726         -0.013121673
Number.of.Policies                     0.021955364 -0.020891140 -0.0086562031         -0.011233031
Total.Claim.Amount                     0.226450915  0.016877407 -0.3552543174          0.632016663
                          Months.Since.Last.Claim Months.Since.Policy.Inception Number.of.Open.Complaints
Customer.Lifetime.Value                0.011516682                  0.0094183812             -0.036343193
Response                              -0.016597147                  0.0029515608             -0.009881182
Income                                -0.026715179                 -0.0008751978              0.006408273
Monthly.Premium.Auto                   0.005026285                  0.0202566426             -0.013121673
Months.Since.Last.Claim                1.000000000                 -0.0429592057              0.005354132
Months.Since.Policy.Inception         -0.042959206                  1.0000000000             -0.001158447
Number.of.Open.Complaints              0.005354132                 -0.0011584467              1.000000000
Number.of.Policies                     0.009136079                 -0.0133328597              0.001498290
Total.Claim.Amount                     0.007562974                  0.0033349145             -0.014241441
                          Number.of.Policies Total.Claim.Amount
Customer.Lifetime.Value          0.021955364        0.226450915
Response                        -0.020891140        0.016877407
Income                          -0.008656203       -0.355254317
Monthly.Premium.Auto            -0.011233031        0.632016663
Months.Since.Last.Claim          0.009136079        0.007562974
Months.Since.Policy.Inception   -0.013332860        0.003334915
Number.of.Open.Complaints        0.001498290       -0.014241441
Number.of.Policies               1.000000000       -0.002353596
Total.Claim.Amount              -0.002353596        1.000000000
>
```

Cor plot

> corrplot(cor(numeric1),

+ type = "full",

+ diag = TRUE,

+ tl.col = "black",

```
+         bg = "white",
+         title = "",
+         col = NULL)
```



The correlation test generated a correlation matrix, which shows that there are a few positive and negative relationships between the numeric variables in the customer dataset. There is a substantial positive correlation between Total.Claim.Amount and Monthly. Premium.Auto. There is also a correlation between Monthly.Premium.Auto and Customer. Lifetime.Value. Total.Claim.Amount and Customer.Lifetime.Value also has a slight positive association. Some variables, such as Response, Monthly.Since.Last.Claim, Months.Since.Policy.Inception, Number.of.Open.Complaints, and Number.of.Policies are not correlated, which means there is no evident link between the variables. Ultimately, there is a weak negative correlation between Total.Claim.Amount and Income.

Plotting Scatterplot for the numerical value

```
> pairs(numeric1)
> plot(numeric1, main ="Scatterplot matrix")
```



Summary of Numerical Variables

> summary(numeric1)

```
# Summary of Numerical features
```{r}

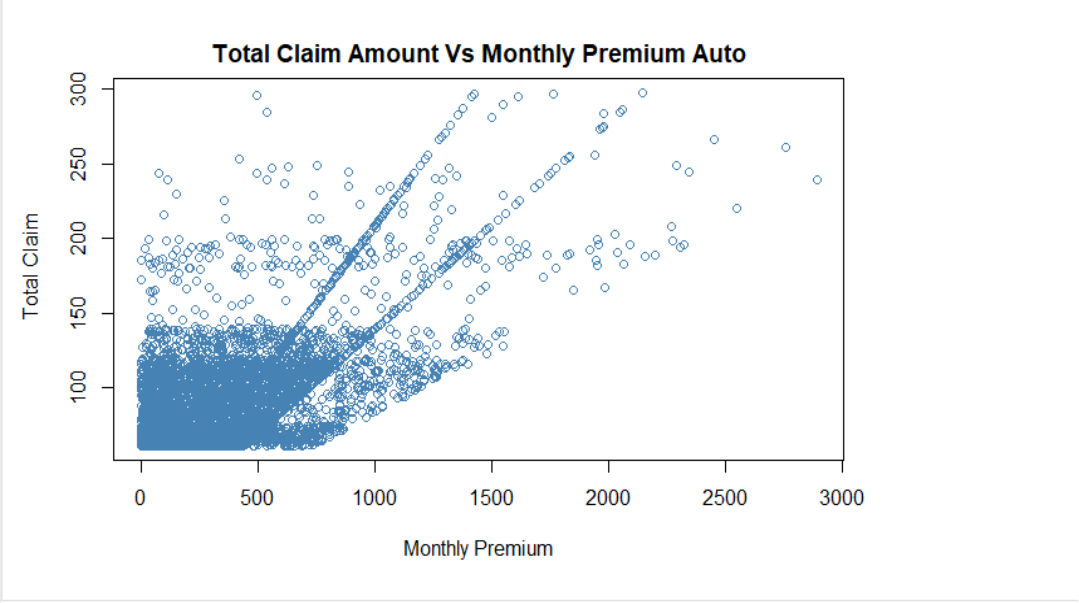
summary(numeric1)

```
```

```
 Customer.Lifetime.Value    Response          Income      Monthly.Premium.Auto Months.Since.Last.Claim
 Min.   : 1898           Min.   :0.0000   Min.   :    0   Min.   : 61.00       Min.   : 0.0
 1st Qu.: 3994           1st Qu.:0.0000   1st Qu.:    0   1st Qu.: 68.00       1st Qu.: 6.0
 Median : 5780           Median :0.0000   Median :33890   Median : 83.00       Median :14.0
 Mean   : 8005           Mean   :0.1432   Mean   :37657   Mean   : 93.22       Mean   :15.1
 3rd Qu.: 8962           3rd Qu.:0.0000   3rd Qu.:62320   3rd Qu.:109.00       3rd Qu.:23.0
 Max.   :83325           Max.   :1.0000   Max.   :99981   Max.   :298.00       Max.   :35.0
 Months.Since.Policy.Inception Number.of.Open.Complaints Number.of.Policies Total.Claim.Amount
 Min.   : 0.00                 Min.   :0.0000            Min.   :1.000      Min.   :   0.099
 1st Qu.:24.00                 1st Qu.:0.0000            1st Qu.:1.000      1st Qu.: 272.258
 Median :48.00                 Median :0.0000            Median :2.000      Median : 383.945
 Mean   :48.06                 Mean   :0.3844            Mean   :2.966      Mean   : 434.089
 3rd Qu.:71.00                 3rd Qu.:0.0000            3rd Qu.:4.000      3rd Qu.: 547.515
 Max.   :99.00                 Max.   :5.0000            Max.   :9.000      Max.   :2893.240
```

I used the pairs () function to build a jumbled figure for the correlation matrix generated earlier for the variables that were identified to be correlated and below are the graphs that were plotted.

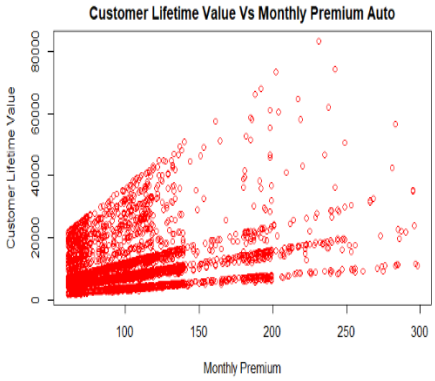
```
> plot(df$Total.Claim.Amount, df$Monthly.Premium.Auto, main ="Total Claim Amount Vs Monthly Premium Auto", xlab = "Monthly Premium", ylab = " Total Claim", col = "steelblue")
>
```



This illustrates a positive correlation between consumers' Total Claim Amount and their auto loan monthly premium paid. One may infer from the figure above that there is a positive association since the correlation matrix above shows a 0.632 correlation coefficient for both variables.

```
> plot(df$Monthly.Premium.Auto, df$Customer.Lifetime.Value, main =" Customer Lifetime Value Vs Monthly Premium Auto", xlab = "Monthly Premium", ylab = " Customer Lifetime Value ", col = "red")
>
```

```{r}
plot(df$Monthly.Premium.Auto, df$Customer.Lifetime.Value, main ="Customer Lifetime Value Vs Monthly Premium Auto", xlab = "Monthly Premium", ylab = "Customer Lifetime Value", col = "red")
```

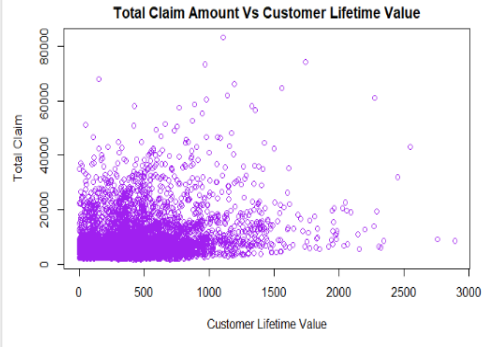


The above chart of Customer Lifetime Value versus Monthly Premium Auto shows a strong correlation between the customer's overall worth and the monthly premium for the auto loan. One might infer that the association is relatively positive from the given figure. According to the correlation matrix above, the correlation between the two variables is 0.39626.

> plot(df$Total.Claim.Amount, df$Customer.Lifetime.Value, main ="Total Claim Amount Vs Customer Lifetime Value",xlab = "Customer Lifetime Value", ylab = "Total Claim", col = "purple")

>

```{r}
plot(df$Total.Claim.Amount, df$Customer.Lifetime.Value, main ="Total Claim Amount Vs Customer Lifetime Value",xlab = "Customer Lifetime Value", ylab = "Total Claim", col = "purple")
```



The chart above of Total Claim Amount Versus Customer Lifetime Value shows a weak correlation between the total claimed amount by a customer and the customer's overall worth.

One might infer that the association is relatively positive from the given figure. According to the correlation matrix above, the correlation between the two variables is 0.226.

>

> plot(df$Total.Claim.Amount, df$Income, main ="Total Claim Amount Vs Income", xlab = "Income", ylab = "Total Claim", col = "blue")

According to the Total Claim Amount vs. Income graph above, there is a direct inverse relationship between the customer's total amount of claims and their annual income. From the presented figure, one may conclude that the correlation is negative. The correlation between the two variables is -0.35525, as shown by the correlation matrix above.

Summary of The Variables' Statistics

To describe the statistical distributions and properties of the variables in the dataset. The summary () function was used to obtain a high-level overview of the dataset's distribution.

> summary(df)

```
  Customer               State      Customer.Lifetime.Value    Response          Coverage
Length:9134        Arizona   :1703   Min.   : 1898           Min.   :0.0000   Basic   :5568
Class :character   California:3150   1st Qu.: 3994           1st Qu.:0.0000   Extended:2742
Mode  :character   Nevada    : 882   Median : 5780           Median :0.0000   Premium : 824
                   Oregon    :2601   Mean   : 8005           Mean   :0.1432
                   Washington: 798   3rd Qu.: 8962           3rd Qu.:0.0000
                                     Max.   :83325           Max.   :1.0000
               Education    Effective.To.Date  EmploymentStatus      Gender              Income
Bachelor            :2748   Length:9134        Length:9134        Length:9134        Min.   :    0
College             :2681   Class :character   Class :character   Class :character   1st Qu.:    0
Doctor              : 342   Mode  :character   Mode  :character   Mode  :character   Median :33890
High School or Below:2622                                                            Mean   :37657
Master              : 741                                                            3rd Qu.:62320
                                                                                     Max.   :99981

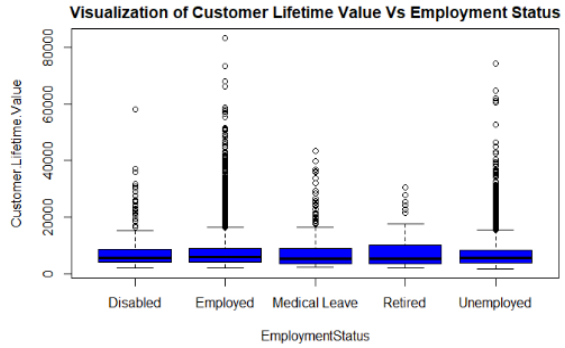
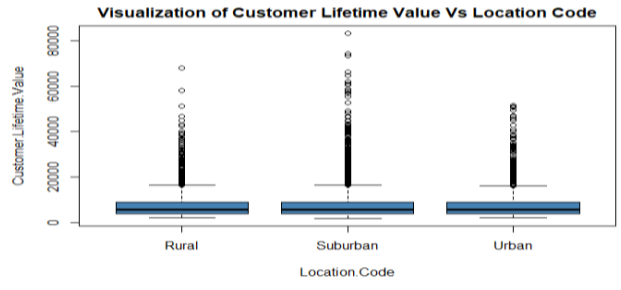
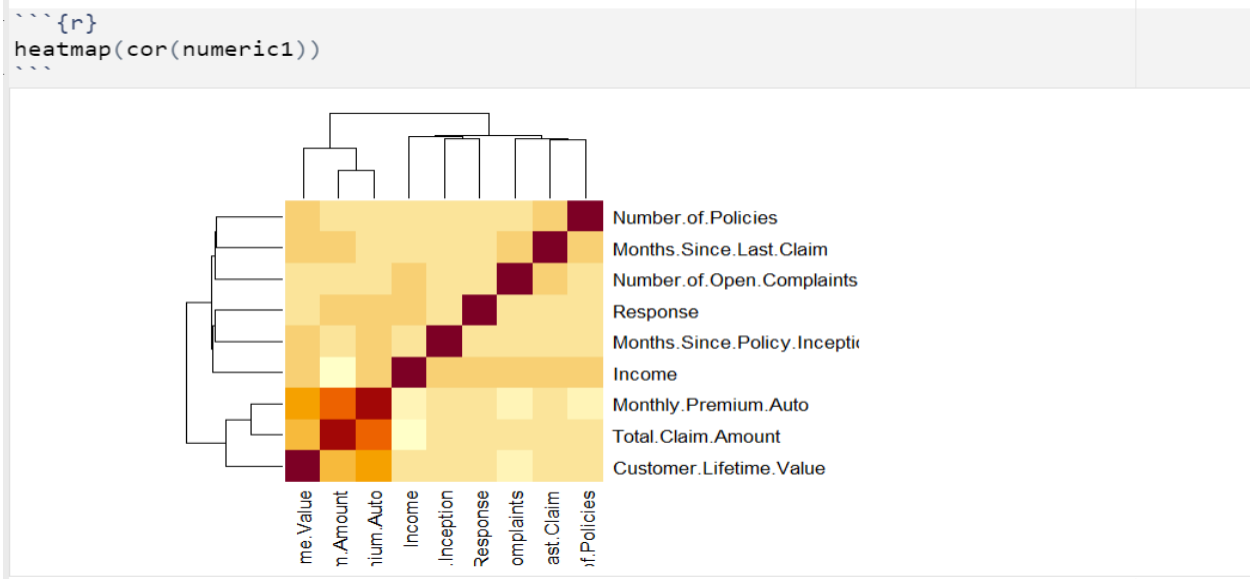
 Location.Code      Marital.Status     Monthly.Premium.Auto Months.Since.Last.Claim
 Length:9134        Length:9134        Min.   : 61.00       Min.   : 0.0
 Class :character   Class :character   1st Qu.: 68.00       1st Qu.: 6.0
 Mode  :character   Mode  :character   Median : 83.00       Median :14.0
                                       Mean   : 93.22       Mean   :15.1
                                       3rd Qu.:109.00       3rd Qu.:23.0
                                       Max.   :298.00       Max.   :35.0
 Months.Since.Policy.Inception Number.of.Open.Complaints Number.of.Policies Policy.Type
 Min.   : 0.00                 Min.   :0.0000            Min.   :1.000      Length:9134
 1st Qu.:24.00                 1st Qu.:0.0000            1st Qu.:1.000      Class :character
 Median :48.00                 Median :0.0000            Median :2.000      Mode  :character
 Mean   :48.06                 Mean   :0.3844            Mean   :2.966
 3rd Qu.:71.00                 3rd Qu.:0.0000            3rd Qu.:4.000
 Max.   :99.00                 Max.   :5.0000            Max.   :9.000
```

```
 Months.Since.Policy.Inception Number.of.Open.Complaints Number.of.Policies Policy.Type
 Min.   : 0.00                 Min.   :0.0000            Min.   :1.000      Length:9134
 1st Qu.:24.00                 1st Qu.:0.0000            1st Qu.:1.000      Class :character
 Median :48.00                 Median :0.0000            Median :2.000      Mode  :character
 Mean   :48.06                 Mean   :0.3844            Mean   :2.966
 3rd Qu.:71.00                 3rd Qu.:0.0000            3rd Qu.:4.000
 Max.   :99.00                 Max.   :5.0000            Max.   :9.000
    Policy          Renew.Offer.Type   Sales.Channel      Total.Claim.Amount Vehicle.Class
 Length:9134        Length:9134        Length:9134        Min.   :   0.099   Length:9134
 Class :character   Class :character   Class :character   1st Qu.: 272.258   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median : 383.945   Mode  :character
                                                          Mean   : 434.089
                                                          3rd Qu.: 547.515
                                                          Max.   :2893.240

 Vehicle.Size
 Length:9134
 Class :character
 Mode  :character
```

The boxplot below illustrates how the customer's lifetime value is impacted by the employee status. Customers who were employed tended to be more valued overall. Furthermore, retired customers had a higher third quartile and less suspected outliers than others.

```
>
> boxplot(Customer.Lifetime.Value ~ EmploymentStatus, df,
+         main = "Visualization of Customer Lifetime Value Vs Employment Status",
+         xlab = "EmploymentStatus",
+         ylab = "Customer.Lifetime.Value",
+         col = "blue")
>
```

```
# Effect of Employment Status on Customer Time Value
```{r}
boxplot(Customer.Lifetime.Value ~ EmploymentStatus, df,
        main = "Visualization of Customer Lifetime Value Vs Employment Status",
        xlab = "EmploymentStatus",
        ylab = "Customer.Lifetime.Value",
        col = "blue")
```
```



The boxplot below illustrates how the customer's lifetime value is impacted by the location code. Customers who live in the suburbs tend to have more customers with higher value, which is more outliers than customers in other locations.

```
> boxplot(Customer.Lifetime.Value ~ Location.Code, df,
+         main = "Visualization of Customer Lifetime Value Vs Location Code",
+         xlab = "Location.Code",
+         ylab = "Customer.Lifetime.Value",
+         col = "steelblue")
>
```

```
# Effect of Location on Customer Life Time Value
```{r}
boxplot(Customer.Lifetime.Value ~ Location.Code, df,
        main = "Visualization of Customer Lifetime Value Vs Location Code",
        xlab = "Location.Code",
        ylab = "Customer.Lifetime.Value",
        col = "steelblue")
```
```



The heatmap below depicts the correlation matrix.

```
>
```

> heatmap(cor(numeric1))



Analysis of the Continuous Variables Using Categorical Variables.

Comparing the numerical variables to the categorical variables. I decided to evaluate the three categorical factors included in the dataset with the response variable Total Claim Amount. The categorical variables that are contained in the data set are State, Coverage, and Education. Plots were made between each variable (used as a predictor variable) and the total claim amount that is the response variable.

The State Variable:

I had anticipated that in some states, the overall claim amount would be more than the national average. I employed simple linear regression and hypothesis testing to examine this proposition.

H0:β1=0

H0:β1≠0

> lg1 <- lm(df$Total.Claim.Amount ~ df$State)

> summary(lg1)

```
# ANALYSIS OF CONTINOUS VARIABLES WITH CATEGORICAL VARIABLES

```{r}
lg1 <- lm(df$Total.Claim.Amount ~ df$State)
summary(lg1)
```


Call:
lm(formula = df$Total.Claim.Amount ~ df$State)

Residuals:
    Min   1Q Median   3Q   Max
-438.65 -161.73 -51.37 114.19 2467.88

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)   425.361   7.040 60.420   <2e-16 ***
df$StateCalifornia   12.458   8.738   1.426   0.154
df$StateNevada   13.389   12.052   1.111   0.267
df$StateOregon   7.651   9.056   0.845   0.398
df$StateWashington   10.982   12.463   0.881   0.378

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

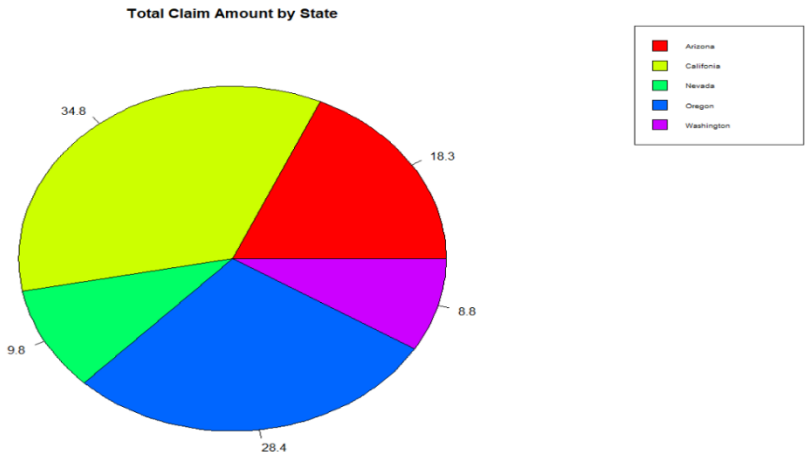
Residual standard error: 290.5 on 9129 degrees of freedom
Multiple R-squared: 0.0002592, Adjusted R-squared: -0.0001788
F-statistic: 0.5917 on 4 and 9129 DF, p-value: 0.6686
```

The linear regression output shows that the t-value is low and the p-value is 0.6686, which is more than 0.05, indicating that there is no statistical significance and that the null hypothesis is true with Total.Claim.Amount = State. The equation outcome totals 470.13 Units.

I made a pie chart graph to show the link between the Total Claim Amount and the State. It can be seen that Washington and Nevada have the lowest total amount claimed by customers with a ratio of 8.8 and 9.8. The state with the largest ratio is California.

```
> df1<-aggregate(x = df$Total.Claim.Amount,
+                by = list(df$State),
+                FUN = sum)
> df1
> piepercent3<- round(100 * df1$x / sum(df1$x), 1)
> pie(df1$x, labels =piepercent3,
+     main = "Total Claim Amount by State",col = rainbow(length(df1$x)))
> legend("topright", c("Arizona", "Califonia", "Nevada", "Oregon", "Washington"),
+        cex = 0.65, fill = rainbow(length(df1$x)))
>
```

```r
df1<-aggregate(x = df$Total.Claim.Amount,
                    by = list(df$State),
                    FUN = sum)
df1
piepercent3<- round(100 * df1$x / sum(df1$x), 1)
pie(df1$x, labels =piepercent3,
    main = "Total Claim Amount by State",col = rainbow(length(df1$x)))
legend("topright", c("Arizona", "Califonia", "Nevada","Oregon","Washington"),
       cex = 0.65, fill = rainbow(length(df1$x)))


```



The Education Variable:

I had anticipated that clients with much more education would typically have higher average total claim amounts. To test my theory, I used simple linear regression and hypothesis testing.

H0:β1=0

H0:β1≠0

>

> lg2 <- lm(df$Total.Claim.Amount ~ df$Education)

> summary(lg2)

```r
lg2 <- lm(df$Total.Claim.Amount ~ df$Education)
summary(lg2)
```

```
Call:
lm(formula = df$Total.Claim.Amount ~ df$Education)

Residuals:
    Min      1Q  Median      3Q     Max
-487.09 -175.19  -43.99  113.79 2465.25

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      427.994      5.485  78.037  < 2e-16 ***
df$EducationCollege               -4.181      7.805  -0.536    0.592
df$EducationDoctor               -89.807     16.486  -5.448 5.24e-08 ***
df$EducationHigh School or Below  59.196      7.849   7.542 5.07e-14 ***
df$EducationMaster               -77.757     11.901  -6.534 6.76e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 287.5 on 9129 degrees of freedom
Multiple R-squared:  0.02093,   Adjusted R-squared:  0.0205
F-statistic:  48.8 on 4 and 9129 DF,  p-value: < 2.2e-16
```

From the output of the linear regression, it can be deduced that the t-value is low, and the p-value is less than 0.05 associated with the Total.Claim.Amount=Education, hence, we can reject the null hypothesis and conclude that there is a significant relationship between the total amount claimed by a customer and their highest level of education. The intercept β0 can be interpreted as the average total amount by customers with different levels of education and β0 + β1 can be interpreted as the average total claim amount by customers with different levels of education. Based on the simple linear regression above, the total amount claimed across educational levels is an average of 315.445 units.

I created a pie chart to show the distribution of customers' total claim amount across different levels of education variables. From the visualization below, you can easily deduce that customers that are doctors have the lowest amount of claims of 2.9 and high school customers have the highest amount of total claims of 32.2.

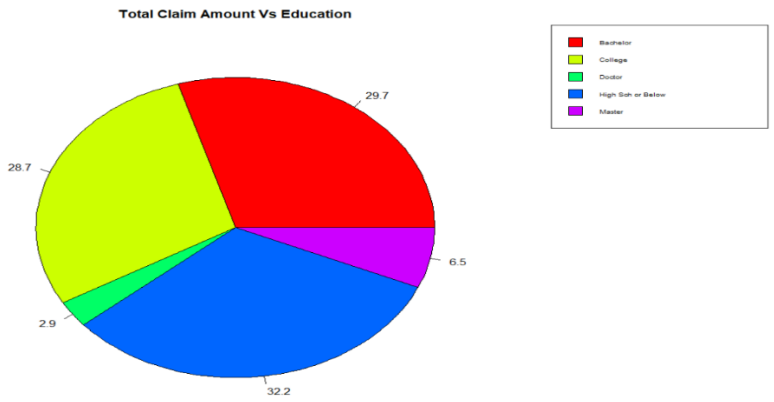
```
> df2 <-aggregate(x = df$Total.Claim.Amount,
+                 by = list(df$Education),
+                 FUN = sum)
> df2
> piepercent4<- round(100 * df2$x / sum(df2$x), 1)
> pie(df2$x, labels =piepercent4,
```

+ main = "Total Claim Amount Vs Education",col = rainbow(length(df2$x)))

> legend("topright", c("Bachelor", "College", "Doctor", "High Sch or Below", "Master"),

+ cex = 0.65, fill = rainbow(length(df2$x)))

```r
df2 <-aggregate(x = df$Total.Claim.Amount,
                by = list(df$Education),
                FUN = sum)
df2
piepercent4<- round(100 * df2$x / sum(df2$x), 1)
pie(df2$x, labels =piepercent4,
    main = "Total Claim Amount Vs Education",col = rainbow(length(df2$x)))
legend("topright", c("Bachelor", "College", "Doctor","High Sch or Below","Master"),
       cex = 0.65, fill = rainbow(length(df2$x)))
```



The Coverage Variable:

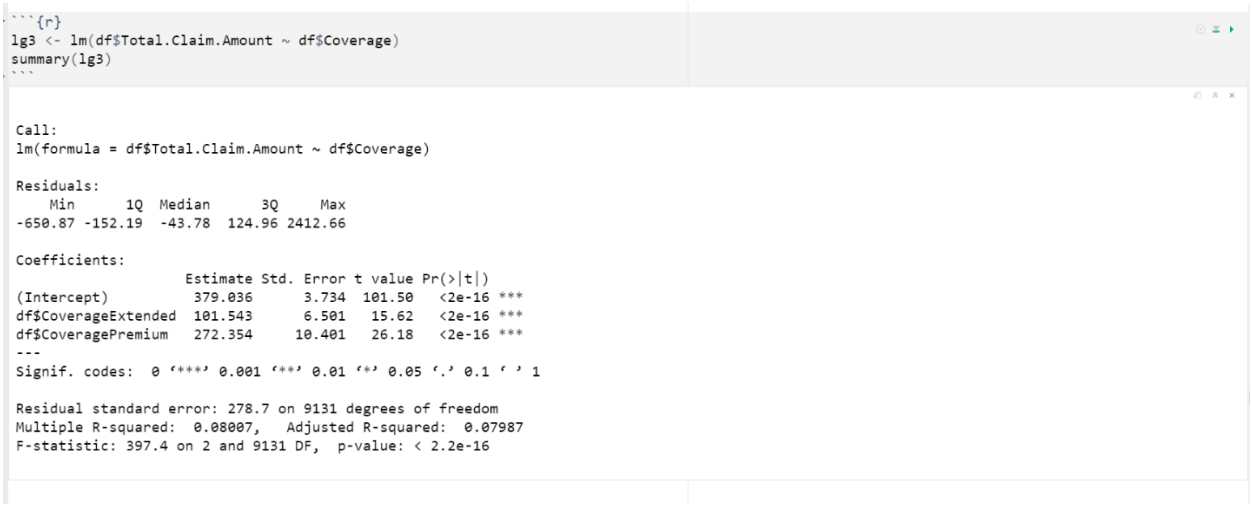
I predict that the total claim amount will increase based on the type of coverage the customer subscribes to. To test this hypothesis, I used a simple linear regression and hypothesis testing.

H0:β1=0

H0:β1≠0

> lg3 <- lm(df$Total.Claim.Amount ~ df$Coverage)

> summary(lg3)



```r
lg3 <- lm(df$Total.Claim.Amount ~ df$Coverage)
summary(lg3)
```

```
Call:
lm(formula = df$Total.Claim.Amount ~ df$Coverage)

Residuals:
    Min      1Q  Median      3Q     Max
-650.87 -152.19  -43.78  124.96 2412.66

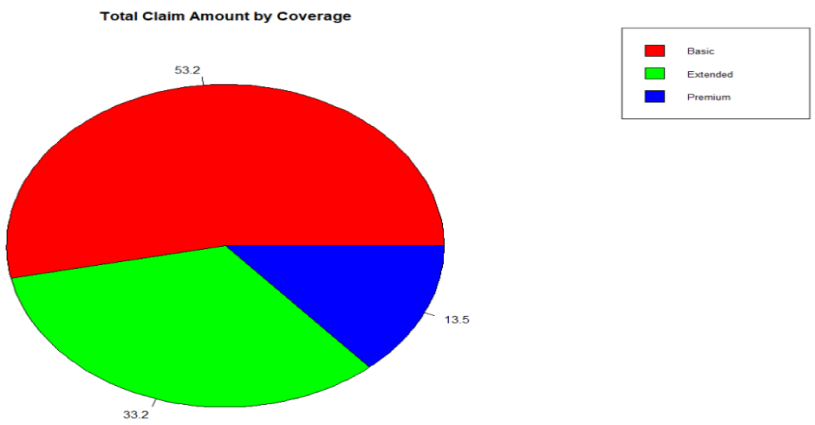
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          379.036      3.734  101.50   <2e-16 ***
df$CoverageExtended  101.543      6.501   15.62   <2e-16 ***
df$CoveragePremium   272.354     10.401   26.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 278.7 on 9131 degrees of freedom
Multiple R-squared:  0.08007,   Adjusted R-squared:  0.07987
F-statistic: 397.4 on 2 and 9131 DF,  p-value: < 2.2e-16
```

From the output of the linear regression, it can be deduced that the t-value is low, and the p-value is less than 0.05 associated with the Total.Claim.Amount=Coverage, hence, we can reject the null hypothesis and conclude that there is a significant relationship between the total amount claimed by a customer and the type of coverage that they use. The intercept β0 can be interpreted as the average total amount by customers with different levels of education and β0 + β1 can be interpreted as the average total claim amount by customers with different levels of coverage that they use. Based on the simple linear regression above, the total amount claimed across educational levels is an average of 752.933 units.

I created a pie chart to show the distribution of customers' total claim amount across the coverage variable. From the visualization below, you can easily deduce that customers that used the basic coverage have the highest total claim amount of 53.2, and customers that subscribe to the premium coverage have the least total claim amount.

```
>
> df3 <- aggregate(x = df$Total.Claim.Amount, by = list(df$Coverage), FUN = sum)
> piepercent1 <- round(100 * df3$x / sum(df3$x), 1)
> pie(df3$x, labels =piepercent1,
+     main = "Total Claim Amount by Coverage",col = rainbow(length(df3$x)))
> legend("topright", c("Basic", "Extended", "Premium"),
+       cex = 0.8, fill = rainbow(length(df3$x)))
```



Contingency tables for Categorical Variables

I looked at the relationships between two different categorical variables and conducted my analysis; the process is illustrated below using a contingency table.

> cont.table1 <- table(df$Coverage, df$Education)

> cont.table1

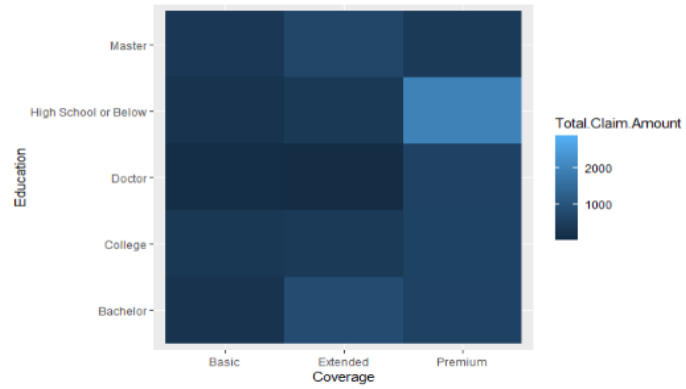
```
# Contigency tables for categorical Variables

```{r}
cont.table1 <- table(df$Coverage, df$Education)
cont.table1
```
```

```
          Bachelor College Doctor High School or Below Master
Basic         1729    1628    200                 1561    450
Extended       769     827    122                  800    224
Premium        250     226     20                  261     67
```

> ggplot(df, aes(x=Coverage, y=Education, fill=Total.Claim.Amount)) + geom_tile()

```{r}
ggplot(df, aes(x=Coverage, y=Education, fill=Total.Claim.Amount)) + geom_tile()
```



> cs.1 <- chisq.test(cont.table1)

> cs.1

```{r}
cs.1 <- chisq.test(cont.table1)
cs.1
```

```
        Pearson's Chi-squared test

data:  cont.table1
X-squared = 18.651, df = 8, p-value = 0.01684
```

Multiple Linear Regression

To choose the optimal model for my dataset, I'll employ a mixed selection method.

Model I

H0: $\beta 1 = \beta 2 = \beta 3 = 0$

H1: $\beta 1 \neq \beta 2 \neq \beta 3 = 0$

$Y = \beta 0 + \beta 1 X1 + \beta 2 Coverage + \beta 3 Education + €$

> model1 <- lm(Total.Claim.Amount ~ Monthly.Premium.Auto + Coverage + Education, data = df)

> anova(model1); summary(model1)

```
Analysis of Variance Table

Response: Total.Claim.Amount
                      Df    Sum Sq   Mean Sq   F value  Pr(>F)
Monthly.Premium.Auto    1 307866945 307866945 6259.0364 < 2e-16 ***
Coverage                2    321274    160637    3.2658 0.03821 *
Education               4  13662437   3415609   69.4405 < 2e-16 ***
Residuals            9126 448885985     49188
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = Total.Claim.Amount ~ Monthly.Premium.Auto + Coverage +
    Education, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1155.45  -102.10     8.36   104.59  1687.80

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -64.36019    7.73054  -8.325  < 2e-16 ***
Monthly.Premium.Auto           5.36780    0.07605  70.586  < 2e-16 ***
CoverageExtended             -13.10758    5.42713  -2.415   0.0157 *
CoveragePremium               -5.23467    9.14903  -0.572   0.5672
EducationCollege              -4.75222    6.02228  -0.789   0.4301
EducationDoctor              -79.40332   12.72508  -6.240 4.57e-10 ***
EducationHigh School or Below  48.06965    6.05746   7.936 2.34e-15 ***
EducationMaster              -85.76493    9.18149  -9.341  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 221.8 on 9126 degrees of freedom
Multiple R-squared:  0.4176,    Adjusted R-squared:  0.4171
F-statistic: 934.8 on 7 and 9126 DF,  p-value: < 2.2e-16
```

Interpretation

According to the value of the F-statistics and the p-value, there is a significant relationship between the response variable (Total amount claimed by customers) and the monthly premium auto, coverage, and education. This relationship is evident in the output produced by the multiple linear regression, which is based on the Analysis of the Variance table as shown above.

The adjusted R-squared value for this model is 0.4171, with a variance of the response variable that is below 45%. Given the high corrected R-squared value of the statistics, we consequently reject the null hypothesis.

I'll carry out my study further by including more variables in the multiple regression model so that I can compare them and choose the model that best fits the dataset.

Model II

H0: $\beta 1 = \beta 2 = \beta 3 = \beta 4 = \beta 5 = 0$

H1: $\beta 1 \neq \beta 2 \neq \beta 3 \neq \beta 4 \neq \beta 5 \neq 0$

```
> model2 <- lm(Total.Claim.Amount ~ Income + Monthly.Premium.Auto + Education +
Coverage, data = df)
> anova(model2); summary(model2)
```

```
Analysis of Variance Table

Response: Total.Claim.Amount
                     Df    Sum Sq   Mean Sq  F value    Pr(>F)
Income                1  97271303  97271303 2453.146 < 2.2e-16 ***
Monthly.Premium.Auto  1 302210252 302210252 7621.630 < 2.2e-16 ***
Education             4   8854241   2213560   55.825 < 2.2e-16 ***
Coverage              2    579472    289736    7.307 0.0006747 ***
Residuals          9125 361821371     39652
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = Total.Claim.Amount ~ Income + Monthly.Premium.Auto +
    Education + Coverage, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1151.45   -99.70    -1.67   103.45  1666.95

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.159e+01  7.443e+00   8.275  < 2e-16 ***
Income                      -3.226e-03  6.885e-05 -46.859  < 2e-16 ***
Monthly.Premium.Auto         5.322e+00  6.829e-02  77.937  < 2e-16 ***
EducationCollege            -4.803e+00  5.407e+00  -0.888 0.374390
EducationDoctor             -6.308e+01  1.143e+01  -5.519 3.50e-08 ***
EducationHigh School or Below  4.231e+01  5.440e+00   7.777 8.23e-15 ***
EducationMaster             -6.190e+01  8.259e+00  -7.494 7.29e-14 ***
CoverageExtended            -1.792e+01  4.874e+00  -3.676 0.000238 ***
CoveragePremium             -1.012e+00  8.215e+00  -0.123 0.901934
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 199.1 on 9125 degrees of freedom
Multiple R-squared:  0.5306,    Adjusted R-squared:  0.5301
F-statistic:  1289 on 8 and 9125 DF,  p-value: < 2.2e-16
```

Interpretation

According to the value of the F-statistics and the p-value, there is a significant relationship between the Total amount claimed by customers and their income, monthly premium auto, education, and coverage in the output produced by the multiple linear regression, as shown in the analysis of variance table above.

The total claimed amount by customers and the predictors variable employed in this model have a positive association, as seen by the modified R-squared value for this model, which comes out to 0.5301 with a variance of 53% of the response variable. As a result, given the high rate of adjusted R-squared value, we reject the null hypothesis.

Adding more variables to the multiple regression model will allow me to compare them and choose the model that best fits the dataset as I continue my investigation.

Model III

H0: $\beta 1 = \beta 2 = \beta 3 = \beta 4 = 0$

H1: $\beta 1 \neq \beta 2 \neq \beta 3 \neq \beta 4 \neq 0$

```
> model3 <- lm(Total.Claim.Amount ~ Customer.Lifetime.Value + Education + Coverage, data = df)
> anova(model3); summary(model3)
```

```
Analysis of Variance Table

Response: Total.Claim.Amount
                       Df    Sum Sq  Mean Sq F value    Pr(>F)
Customer.Lifetime.Value    1  39523388 39523388 539.043 < 2.2e-16 ***
Education                  4  15258291  3814573  52.025 < 2.2e-16 ***
Coverage                   2  46824069 23412034 319.307 < 2.2e-16 ***
Residuals               9126 669130892    73321
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = Total.Claim.Amount ~ Customer.Lifetime.Value + Education +
    Coverage, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-932.55 -147.87  -31.77  117.06 2417.25

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.204e+02  6.335e+00  50.577  < 2e-16 ***
Customer.Lifetime.Value      7.701e-03  4.185e-04  18.400  < 2e-16 ***
EducationCollege            -4.968e+00  7.353e+00  -0.676    0.499
EducationDoctor             -8.615e+01  1.554e+01  -5.545 3.02e-08 ***
EducationHigh School or Below  5.160e+01  7.396e+00   6.977 3.22e-12 ***
EducationMaster             -8.249e+01  1.121e+01  -7.358 2.02e-13 ***
CoverageExtended             8.950e+01  6.356e+00  14.080  < 2e-16 ***
CoveragePremium              2.409e+02  1.023e+01  23.552  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 270.8 on 9126 degrees of freedom
Multiple R-squared:  0.1318,    Adjusted R-squared:  0.1312
F-statistic:   198 on 7 and 9126 DF,  p-value: < 2.2e-16
```

Interpretation

The multiple linear regression output based on model III shows that all the variables are statistically significant. Based on the p-values associated with the t-values of the variables, Customer.Lifetime.Value, Education, and Coverage are all significant according to the analysis of variance in the response variable of Total.Claim.Amount. The p-value associated is low, but the adjusted R-squared value is 0.1312 indicating 13% of the change in the response variable due to changes in the predictors. This is a pretty low adjusted R-squared and therefore we can accept the null hypothesis because of the low adjusted R-square value.

Model IV

> model4 <- lm(Total.Claim.Amount ~ Monthly.Premium.Auto + Income + Customer.Lifetime.Value + Education + Coverage, data = df)

> anova(model4); summary(model4)

```
Analysis of Variance Table

Response: Total.Claim.Amount
                          Df    Sum Sq   Mean Sq   F value    Pr(>F)
Monthly.Premium.Auto       1 307866945 307866945 7767.7024 < 2.2e-16 ***
Income                     1  91614611  91614611 2311.5019 < 2.2e-16 ***
Customer.Lifetime.Value    1    162266    162266    4.0941 0.0430627 *
Education                  4   8888320   2222080   56.0647 < 2.2e-16 ***
Coverage                   2    581734    290867    7.3388 0.0006537 ***
Residuals               9124 361622764     39634
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = Total.Claim.Amount ~ Monthly.Premium.Auto + Income +
    Customer.Lifetime.Value + Education + Coverage, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1152.09  -100.10    -0.09   102.82  1658.80

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  6.175e+01  7.442e+00   8.297  < 2e-16 ***
Monthly.Premium.Auto         5.382e+00  7.331e-02  73.413  < 2e-16 ***
Income                      -3.221e-03  6.888e-05 -46.765  < 2e-16 ***
Customer.Lifetime.Value     -7.399e-04  3.305e-04  -2.239 0.025211 *
EducationCollege            -4.829e+00  5.406e+00  -0.893 0.371759
EducationDoctor             -6.327e+01  1.143e+01  -5.536 3.18e-08 ***
EducationHigh School or Below  4.251e+01  5.440e+00   7.815 6.13e-15 ***
EducationMaster             -6.175e+01  8.258e+00  -7.478 8.23e-14 ***
CoverageExtended            -1.801e+01  4.873e+00  -3.695 0.000221 ***
CoveragePremium             -1.345e+00  8.214e+00  -0.164 0.869991
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 199.1 on 9124 degrees of freedom
Multiple R-squared:  0.5308,    Adjusted R-squared:  0.5303
F-statistic:  1147 on 9 and 9124 DF,  p-value: < 2.2e-16
```

Interpretation

The output generated from the multiple linear regression above shows that based on the p-values associated with the t-values of the variables, Monthly.Premium.Auto, Income, Customer.Lifetime.Value, Coverage are all significant according to the analysis of the variance table in response variable of Total.Claim.Amount. their associated p-values are small enough to

reject the null hypothesis and determine that it is unlikely to see the such association between the predictor and response by chance. The adjusted R-squared value is 0.5303 indicating roughly 53% of the change in the response variable is due to changes in the predictors.

Model V

```
> model5 <- lm(Total.Claim.Amount ~ Income + Monthly.Premium.Auto +
Customer.Lifetime.Value + Response + Education + Coverage, data = df)
> anova(model5); summary(model5)
```

```
Analysis of Variance Table

Response: Total.Claim.Amount
                         Df    Sum Sq   Mean Sq   F value    Pr(>F)
Income                    1  97271303  97271303 2455.4759 < 2.2e-16 ***
Monthly.Premium.Auto      1 302210252 302210252 7628.8685 < 2.2e-16 ***
Customer.Lifetime.Value   1    162266    162266    4.0962 0.0430096 *
Response                  1    149162    149162    3.7654 0.0523547 .
Education                 4   8966565   2241641   56.5871 < 2.2e-16 ***
Coverage                  2    578261    289131    7.2987 0.0006804 ***
Residuals              9123 361398830     39614
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
lm(formula = Total.Claim.Amount ~ Income + Monthly.Premium.Auto +
    Customer.Lifetime.Value + Response + Education + Coverage,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1149.87  -101.47    -1.21   103.48  1661.11

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   6.003e+01  7.475e+00   8.030 1.10e-15 ***
Income                       -3.223e-03  6.886e-05 -46.799  < 2e-16 ***
Monthly.Premium.Auto          5.379e+00  7.330e-02  73.381  < 2e-16 ***
Customer.Lifetime.Value      -7.286e-04  3.305e-04  -2.204 0.027517 *
Response                      1.415e+01  5.950e+00   2.378 0.017447 *
EducationCollege             -5.036e+00  5.405e+00  -0.932 0.351484
EducationDoctor              -6.380e+01  1.143e+01  -5.583 2.43e-08 ***
EducationHigh School or Below  4.260e+01  5.438e+00   7.834 5.25e-15 ***
EducationMaster              -6.209e+01  8.257e+00  -7.519 6.02e-14 ***
CoverageExtended             -1.794e+01  4.872e+00  -3.683 0.000232 ***
CoveragePremium              -1.281e+00  8.212e+00  -0.156 0.876066
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 199 on 9123 degrees of freedom
Multiple R-squared:  0.5311,    Adjusted R-squared:  0.5306
F-statistic:  1033 on 10 and 9123 DF,  p-value: < 2.2e-16
```

Interpretation

The output generated by Model V shows almost all the predictor variables are statistically significant except for Response. If we look at the F-statistics and its p-value, it is lower than 0.05, therefore, the null hypothesis that all the estimated betas are equal to zero can be rejected. It can be concluded that all predictor variables except for Response have some relationship with the response variable (Total.Claim.Amount).

The adjusted R-squared value for this model v is 0.5306 indicating roughly 53% of the change in the response variable is due to changes in the predictors. When comparing with model IV. This model is better because the adjusted R-squared is slightly higher than model IV and model II which is why I decided to go with this model.

> confint(model2)

```{r}
confint(model2)
```

	2.5 %	97.5 %
(Intercept)	47.004514141	76.185026568
Income	-0.003361361	-0.003091424
Monthly.Premium.Auto	5.188116312	5.455825994
EducationCollege	-15.402358314	5.795868738
EducationDoctor	-85.490044264	-40.677389878
EducationHigh School or Below	31.643625223	52.971124296
EducationMaster	-78.087617877	-45.707462167
CoverageExtended	-27.471790393	-8.364245175
CoveragePremium	-17.115371776	15.090843606

Other Multiple Regression

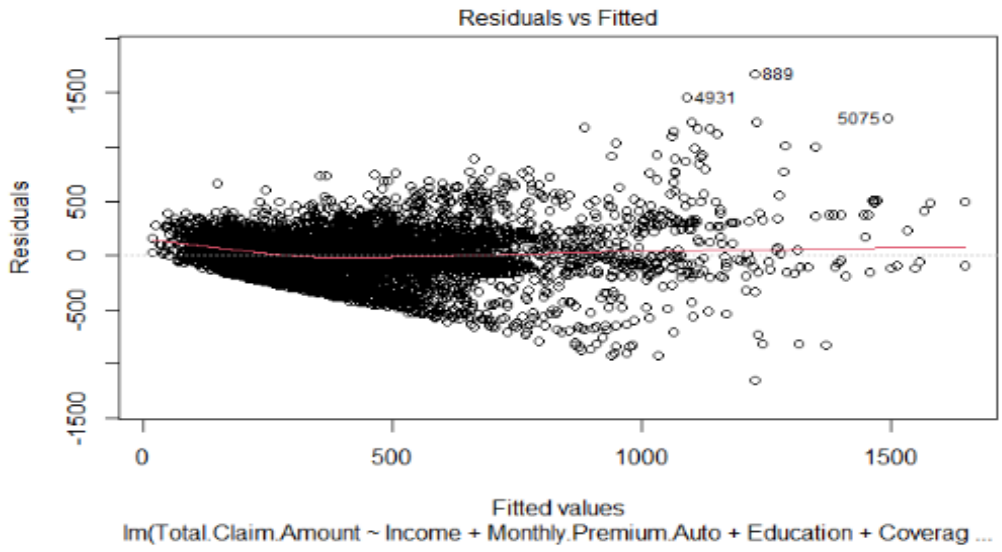
Based on the results of the various multiple regressions that were carried out, I carried out further multiple regressions by switching between different predictor variables and compared my output with models I through V. Model V is the most effective model for this dataset, in my opinion. The model analysis does not finish here. By examining the issues with model V, I intend to proceed with the data analysis.

Potential Problems of the Model

After deciding to use model V, I further examine the model to determine whether it has any regression issues. The processes used to determine that model is a good one to employ are as follows.

A) Non-linearity of the model

I plotted the residual plot to observe the model's nonlinearity. If there is a pattern in the residuals of the plot versus the expected values, it indicates that our linear model has a problem. However, I can find no trend in the comparison of residuals against anticipated values in my figure below. There are both positive and negative residuals around the regression line, indicating that there does not appear to be a distinct trend that would suggest concerns with data non-linearity. As a result, the linear model utilized to work with the data looks to be effective. In the graph, there are three outliers where customers received the highest total claim amount. 889, 4931, and 50750, to be exact.



B) Correlation of Error Term

We may deduce that the dataset's error term is uncorrelated since it was not collected at the same intervals. As a result, for mode 1, there is no discernable pattern of the error term.

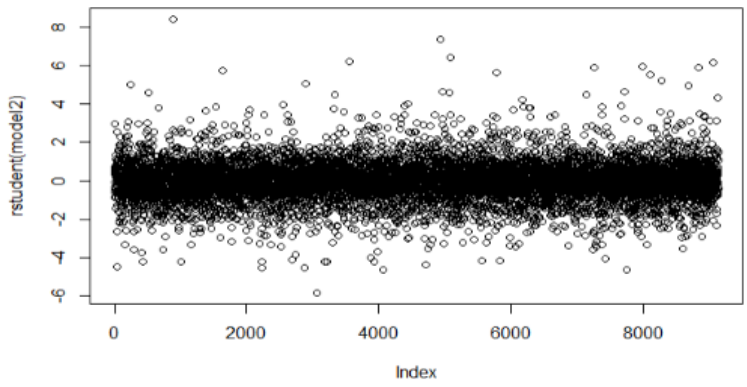
C) Non-constant Variance of Error

We examine the residual plot for the presence of a funnel-shaped pattern or signs of heteroscedasticity to assess the non-constant variance in mistakes of our model. However, based on the residual plot of the model I have, there is no indication of heteroscedasticity. It would not be necessary to change the data in order to produce a different residual plot.

D) Outliers

I continued my analysis by looking for outliers in the model. To see if there were any outliers in the model, I displayed the studentized residuals of multiple regression model V. The graph below displays residuals that are more than + or -6 standard deviations. This demonstrates the occurrence of outliers. However, I do not believe the outliers are caused by a mistake in the dataset. I am leaving the dataset without looking into the causes of the outliers.

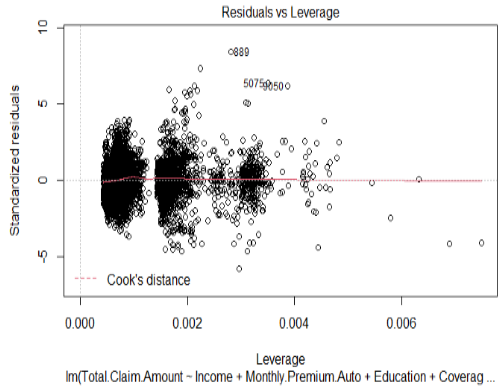
```
> plot(rstudent(model2))
>
```

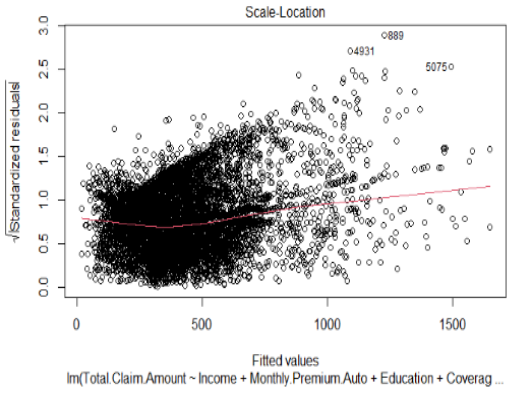


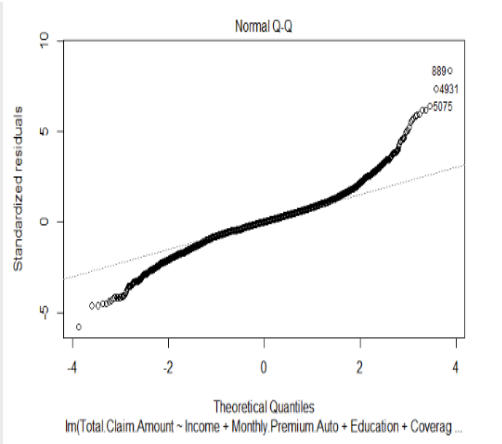
E) High Leverage Points

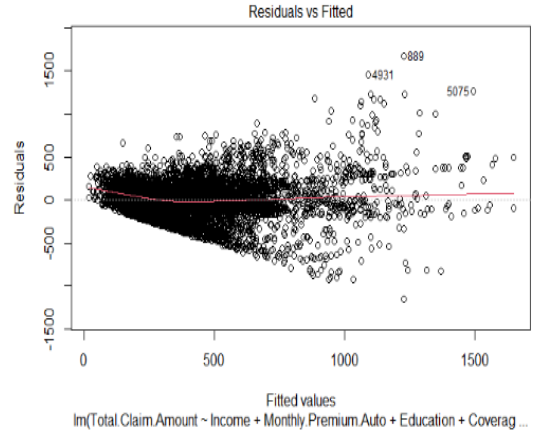
We cannot find any high leverage points in our model based on the residuals vs. leverage plot. This leads us to the conclusion that we do not require modification to any variables to suit our model. Model V, the model of use, is an excellent choice. The normal Q-Q plot demonstrates a normal distribution in the model, indicating that the model chosen is appropriate.

```
> plot(lm(Total.Claim.Amount ~ Income + Monthly.Premium.Auto + Education + Coverage,
data = df))
>
```









Clustering Techniques

For my clustering approaches, I chose Kmeans clustering to cluster instances of various dataset properties. The cluster codes and output that were used are listed below.

```{r}
set.seed(55)
cluster.df <- kmeans(df[, 22, 6], 3, nstart = 20)
cluster.df
```

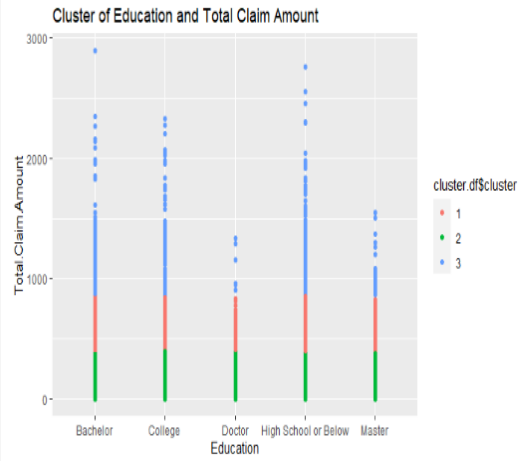
```
K-means clustering with 3 clusters of sizes 3788, 4675, 671

Cluster means:
       [,1]
1  553.7127
2  233.2363
3 1158.1576

Clustering vector:
   [1] 2 3 1 1 2 2 2 2 1 1 1 1 1 1 2 2 1 1 2 1 2 2 1 1 1 2 1 1 1 2 1 1 2 2 2 2 1 1 1 2 1 3 1 2 2 1 1 1 2 1 1 2 1 1 2 2 2 1 1 2 1 1 2 1 1 2
  [89] 2 2 1 1 1 2 2 3 1 2 2 2 2 1 1 2 1 3 2 2 1 2 2 1 1 1 2 1 1 1 1 2 2 1 2 2 3 2 1 2 1 2 2 1 2 2 1 1 3 2 1 2 2 1 1 3 1 1 2 2 2 2 1 2 2 1 2 2 2 2 1 1 2 1 1 1 2 1 1 2 2 2 1 2 2 2 1 2
 [177] 2 2 1 1 1 3 1 1 1 1 1 1 2 2 1 1 2 2 2 1 3 2 1 2 1 1 1 3 1 3 2 3 3 1 2 1 2 2 2 2 1 1 1 1 2 2 2 1 1 1 2 1 2 2 1 2 2 1 2 2 2 1 2 2 2 2 2 2 1 1 1 2 2 3 2 1 1 1 1 2 1 2 2 1 2 2 2 2 2
 [265] 1 2 2 1 2 1 2 1 1 1 1 2 1 2 1 1 2 2 2 1 2 2 2 2 1 2 2 2 1 2 1 1 2 1 2 1 2 2 1 2 1 1 2 1 1 2 1 1 1 1 2 2 2 2 2 2 3 1 1 2 2 2 2 2 2 1 2 2
 [353] 2 2 2 2 2 3 2 1 2 2 3 2 2 1 1 1 2 2 2 1 1 1 2 2 1 3 1 2 2 2 2 1 1 1 2 2 1 2 2 2 3 1 2 1 2 1 1 2 2 1 1 2 1 2 2 1 1 3 1 2 2 2 1 1 2 1 3 1 1 2 3 2 1 2 2 2 1 1 1 2 1 1 1 3 2 2 2 1
 [441] 1 1 1 2 2 2 1 1 1 2 1 1 1 2 2 1 1 3 2 2 2 1 2 2 2 2 2 1 1 2 2 2 2 1 1 1 2 1 2 2 1 2 2 2 2 2 3 1 1 2 2 2 2 1 1 2 1 3 1 1 1 1 1 3 2 1 1 2 2 1 1 3 1 2 1 3 2 2 2 3 3 1 1 3 1 2 1 2 2
 [529] 1 2 2 1 2 1 2 1 2 2 2 3 1 1 1 1 2 1 2 2 1 1 1 2 1 2 2 1 2 1 1 1 2 2 2 1 2 1 2 1 2 1 2 2 3 2 2 1 2 1 2 2 1 1 1 2 1 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 3 1 1 3 2 2 2 2 2 1
 [617] 1 2 2 1 2 2 2 1 2 1 2 1 1 2 2 2 2 3 2 2 3 2 1 1 2 2 1 2 2 3 1 2 1 3 2 2 1 2 1 1 2 2 2 2 2 2 1 2 1 2 2 1 3 2 2 2 1 2 2 2 1 2 1 2 1 1 1 2 1 2 1 2 1 1 1 2 1 1 2 2 2 2 2
 [705] 2 1 1 2 2 1 2 2 1 3 1 2 2 2 1 2 1 2 1 1 1 2 2 2 2 1 2 2 2 1 2 2 2 2 1 1 1 2 1 2 1 1 2 2 2 2 1 2 1 1 1 2 1 2 3 2 3 2 1 1 2 2 2 1 2 1 1 1 1 2 1 1 3 2 1 2 1 1 1 1 1 3 2 2 1 1 2 2 2
 [793] 1 2 2 2 1 2 1 2 2 2 2 2 1 2 2 2 1 2 3 2 2 2 2 2 2 1 3 1 1 2 1 1 1 1 1 2 1 2 1 3 2 2 3 2 1 2 1 2 2 2 3 1 1 2 2 2 1 2 2 1 2 2 1 2 1 1 1 1 2 2 1 2 2 2 1 2 1 2 2 1 2 1 2 3 3 2 2 2 2 2 1 2 2
 [881] 1 2 1 2 2 1 2 2 3 1 2 3 1 2 2 1 2 2 2 2 2 2 3 2 1 1 1 2 2 1 2 1 2 1 1 2 2 2 3 2 2 2 2 1 2 2 2 1 1 3 1 1 2 2 2 1 2 2 2 1 1 2 2 2 1 1 2 2 2 2 2 2 2 1 1 2 2 2 1
 [969] 2 2 1 1 2 2 2 1 2 3 2 1 1 1 2 2 2 1 3 2 2 1 2 1 1 1 2 2 2 1 2 2 1
 [ reached getOption("max.print") -- omitted 8134 entries ]

Within cluster sum of squares by cluster:
[1] 50662582 61225823 64255958
 (between_SS / total_SS =  77.1 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"         "ifault"
```

```r
table(cluster.df$cluster, df$Education)
```

```
  Bachelor College Doctor High School or Below Master
1     1090    1073    122                1275    228
2     1456    1428    214                1099    478
3      202     180      6                 248     35
```

```r
cluster.df$cluster <- as.factor(cluster.df$cluster)
```

```r
ggplot(df, aes(Education, Total.Claim.Amount, color = cluster.df$cluster)) + geom_point() + ggtitle("Cluster of Education and Total Claim Amount")
```



Future Works

In future work, it would be interesting to conduct predictive modeling to estimate the total amount that will be claimed by an unemployed customer, who has a high school education or other levels of education and has a low income to anticipate the total amount that will be claimed by consumers like that.

SUMMARY

I was able to create different visual representations of my dataset using a pie chart, box plot, scatterplot, heatmap, and correlation matrix. I was also able to determine the model for my dataset. My expectation that my research will show patterns and connections between the variables provided in the dataset was met because the p-value was close to 0 and the and the R-square of the statistics was also 0.5306 which is 53% of the coefficient of variance.

Based on my analysis, for the Income variable, there was a relationship between the predictor (Income) and the response variable (Total Claim Amount) because the p-value of the F-statistics was very low and close to 0, thereby we can reject the null hypothesis.

The Monthly Premium Auto variable, had a statistically significant relationship between the predictor (Monthly.Premium.Auto) and the response variable (Total.Claim.Amount) because the p-value was also very low thereby we accept the null hypothesis.

The coverage variable had a slightly low relationship between the predictor (Coverage) and the response variable (Total.Claim.Amount) because the p-value of the F-statistics was pretty high therefore we cannot reject the null hypothesis.

The Education variable, had a statistically significant relationship between the predictor (Education) and the response variable (Total.Claim.Amount) because the p-value of the F-statistics was low. Hence, we reject the null hypothesis.

The Customer Lifetime Value variable had a slightly low significant relationship between the predictor (Customer.Lifetime.Value) and the response variable (Total.Claim.Amount) because the p-value of the F-statistics was slightly on the high side. Therefore, we cannot reject the null hypothesis.

I tested five models; model V was the best and the chosen model because it had the highest adjusted R squared value to show that the changes made in the predictors were related to the change in the response variable. The addition of Coverage that had a slightly low significance had an impact on the model of choice which is model V over model II. The other models were ignored because of the value of their adjusted R squared which was smaller than model V.

Model v did not seem to have any problem attached to it except for the three outliers in the residual plot which were the customers that received the highest amount of claims. So, we leave the model as it is without transforming it.