# Analyzing Automotive Insurance: Predictive Modeling of Claim Amounts

Regression Analysis Approach for Risk Calculation


| Name | Email | Contributions |
|---|---|---|
| Jaiden Atterbury | jatter41@uw.edu | RQ3, Conclusion, Methodologies |
| Tanner Huck | thuck@uw.edu | RQ1, Introduction, Dataset Description |
| Andy Wen | xw99@uw.edu | RQ2, References, Appendix, Proofreader |

Dr. Emanuela Furfaro

STAT 423 Applied Regression and Analysis of Varience

University of Washington, Seattle


Date: March 13, 2024

# 1 Introduction

As an inspiring actuary, an important topic is calculating the risk of potential clients in order to determine insurance premium rates. This is usually accomplished through the extensive use of statistical techniques and predictive modeling. Understanding the intricate dynamics that influence a clients total claim amounts for automotive accidents is paramount for insurance companies to understand in order to tailor services effectively and make informed decisions about how much to charge for their coverage. The goal of this project is to develop robust predictive models that accurately forecast claim amounts and shed light on potential recurring themes within car insurance data. Through thorough data exploration and analysis, we aim to identify key predictors influencing total claim amounts and construct an optimal model that is capable of accurate predictions while ensuring the reliability and validity of our findings.

# 2 Dataset Description

The data set used in our analysis is comprised of different customer data collected from vehicle insurance policies. This data was found on Kaggle (3) and the original purpose of the data was for clustering insurance customers according to their driving behavior (2). This was done in order to sell more insurance policies and target advertising. Each of the 9134 entries corresponds to a unique customer and their respective insurance policy details. These details include demographics such as location code, education level, gender, marital status, income, and other factors that may have an impact on the customers insurance rates. We also have numerous categorical variables like coverage type, policy type, vehicle class, and vehicle size, which provides further insight into the characteristics of each customer and their vehicle. Overall, the data set contains a comprehensive list of features for each customer and provides the necessary tools for predictive modeling aimed at understanding the response variable total claim amount, which represents the total cost of repairs that the insurance company will have to pay in order to fix their clients' vehicle after an accident.

# 3 Research Questions

We will now explain what questions we hope to answer by the end of the project. In particular, each paragraph corresponds to its own research question. It is important to note that there are many questions in each paragraph, these correspond to sub-questions that we intend to answer in each overarching research question.

For our first research question, we want to know, in terms of predicting the total claim amount of clients, which predictors are the most important? Which variables can we assume to be good predictors of the total claim amount? What is the relationship between these predictor variables and the response variable, total claim amount? Through exploratory data analysis, we want to find which variables should and should not be included in a predictive model, as well as background information on which transformations and remedial strategies we may have to employ in our final model.

Our second research question focuses on finding the best model in terms of predictive accuracy. In particular we hope to answer things such as, how can we construct the most optimal linear regression model to predict the total claim amount? Are the identified predictor variables from our

exploratory analysis effective and form a robust linear model? From our created models, which model is the best in terms of predictive accuracy and why? Using an initial full model derived from our exploratory analysis, we will construct "candidate" models and then select a final model for further analysis, based on accuracy criterion.

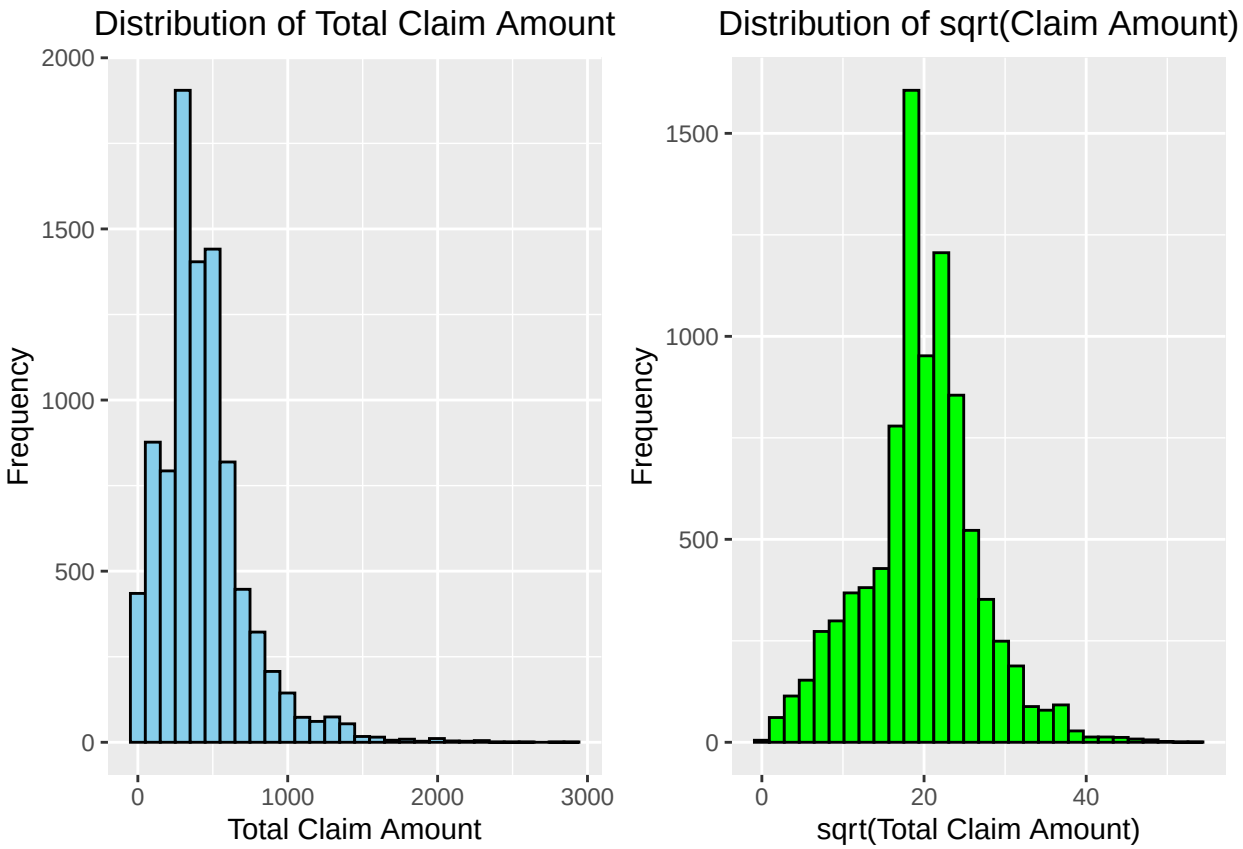
The last research question focuses on checking the validity of our final model. In particular, we hope to answer questions such as, which of these selected variables are actually significant predictors in practice? Are any of the linear regression assumptions violated in this final model? Are there any other problems with this model? If there are, why do they persist? Doing a basic residual analysis, as well as checking other important metrics, we will dive into the validity of our model which will let us know if our results are trustworthy.

# 4 Methodologies

As mentioned above, we intend to analyze automotive insurance data through a structured methodology that begins with exploratory data analysis using scatter plots, histograms, and other visualization tools. This approach will help us identify which predictors are the most relevant for our multiple linear regression model. We will then split the data in half, saving one half for training, and one half for testing. In the training phase, using the full model containing the "important" variables found in the exploratory analysis phase, we will use backward selection to select predictors based off of ten different selection criterion. Using these different models we will perform K-fold cross validation to select the most accurate model on new data. Lastly, in the "testing" phase, we will use the untouched testing data set to perform a residual analysis, and check other important metrics like multicollinearity, leverage, etc. in order to ensure the validity of our model.

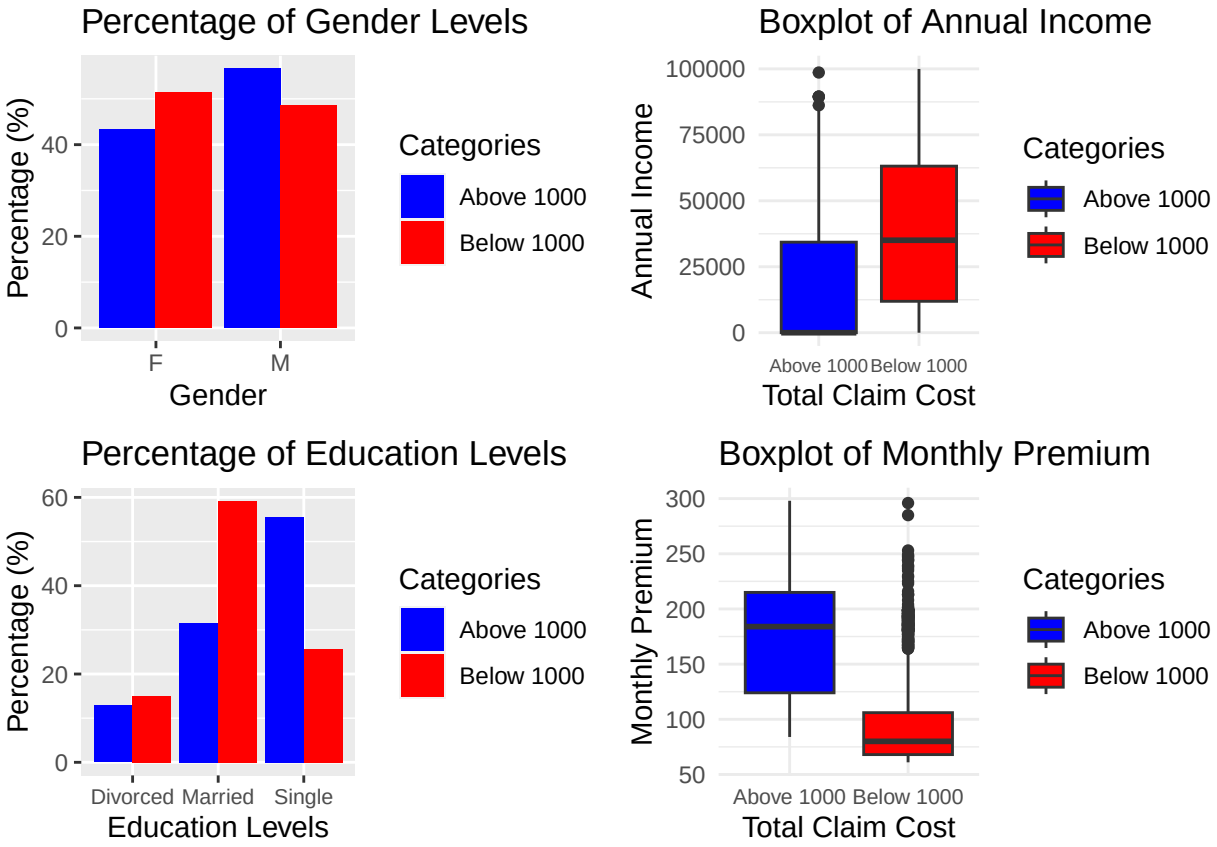
# 5 Research Question 1: What Predictors are Important? (Data Exploration)

From prior knowledge about car insurance, we can anticipate which variables will perform better than others when it comes to predicting the total claim amount of a client. For instance, looking at the Vehicle Class variable, we expect that luxury cars will have higher claim amounts, since luxury cars are more expensive to purchase and repair. Additionally, if an individual has a higher education, they might make better decisions that make them less likely to be involved in an accident that they are at fault for. With that being said, we will explore these anticipated variables to see if they have a noticeable linear relationship with total claim amount. We will also check variables that we don't anticipate to be related to claim amount to ensure we don't miss an important predictor. Understanding the relationship of these variables and total claim amount will provide the framework necessary to build our initial predictive models.

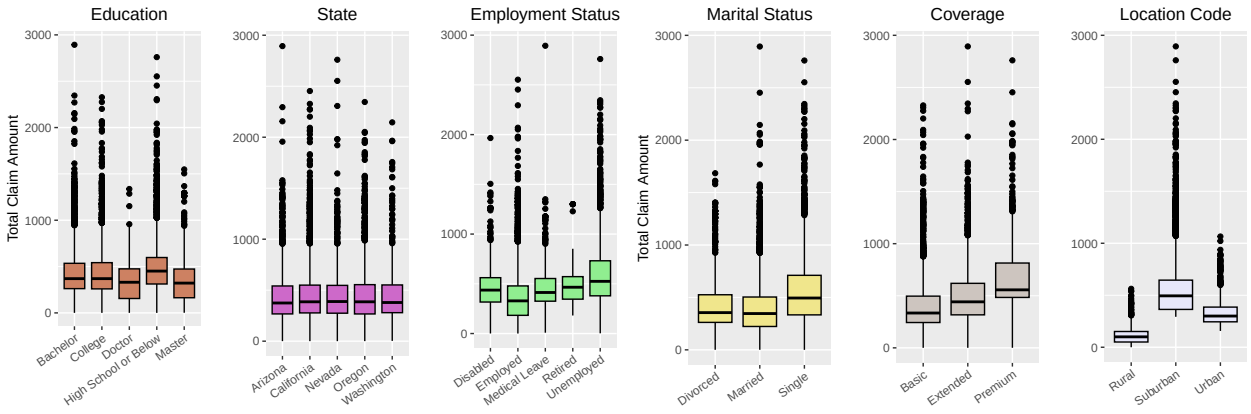


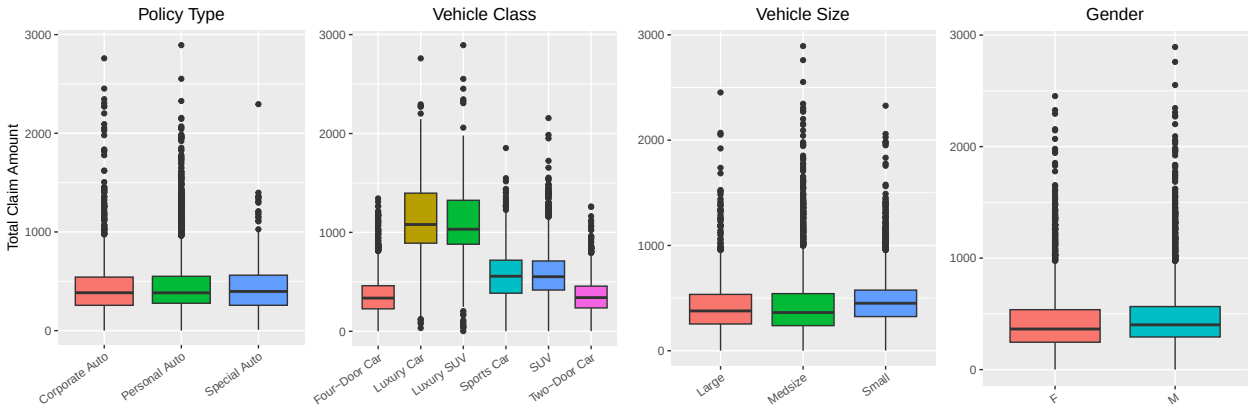

As requested in the project proposal feedback, we will explore the distribution of our response variable, total claim amount. In the above histogram (on the left), we can see the relationship of total claim amount to its frequency. The distribution seems to be right skewed with the majority of total claim amounts falling around 500 U.S. dollars with less claims falling in the higher monetary range. This skewness implies that most claims are relatively low-cost and the occurrence of more expensive claims are rare. Through the eyes of an insurance company, we can expect to allocate most resources towards less expensive claims and this insight may inform our pricing strategies to cater towards this group. Furthermore, due to this non-normality, we tried many different transformations, such as the log, cube root, and square root transformations. The one that gave us the best results was the square root transformation, and as can be seen in the above histogram (on the right) the new distribution is only slightly right skewed, and could be classified as approximately normal. Thus we will use the square root of total claim amounts in our initial model.

Furthermore, when receiving additional project feedback, we were tasked with analyzing the group of individuals who had total claim amounts above 1000 U.S. dollars (the reason for the right skewness). We were given this task in order to see if we were dealing with non-homogeneous groups. To analyze any possible difference we split the data based on total claim amount, and compared a few important variables through the use of percent bar charts and box plots.




As can be seen from the above figure, when it comes to the percentage makeup of gender, there appears to be a difference between individuals with claim amounts above and below 1000 U.S. dollars. However, this difference is something one would expect to occur with random sampling. This same conclusion can be made from the box plots of annual income split by the total claim amount categories. On the other hand, looking at the percentage makeup of the education levels variable, we see that there seems to be a noticeable difference between individuals with claim amounts above and below 1000, namely, single individuals are more prevalent in the above 1000 group. Similarly, those who got in more costly accidents also had a higher monthly premium (which is expected if the actuarial calculation was correct). Overall, we don't have enough evidence to treat these groups separately in our linear regression model without further analysis.

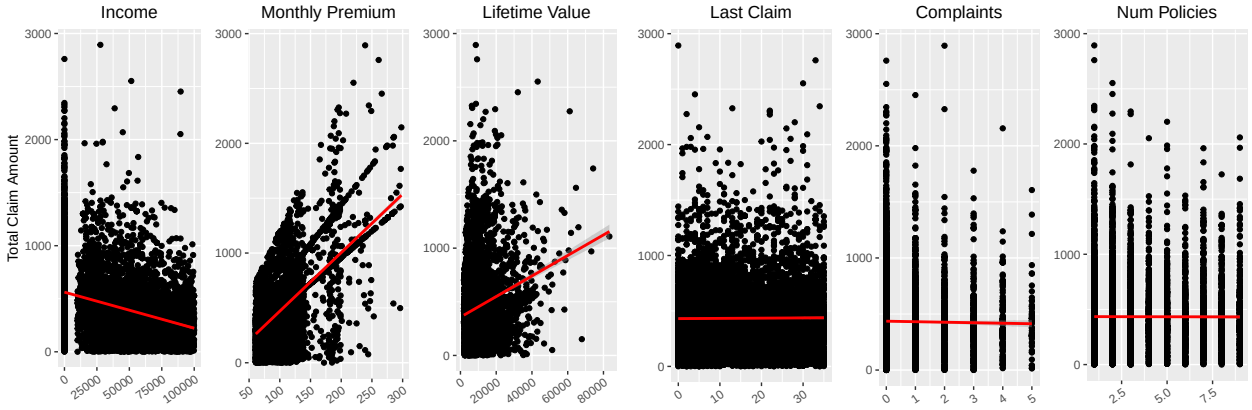





The above series of box plots explores the relationship between the total claim amount and various categorical variables that we expect to have a relationship with total claim amount. These plots reveal which variables can help explain total claim amount and which variables should be included as predictors in our predictive models. With these box plots, we are interested in seeing if there is a variation of total claim amount depending on different levels of our variables. If we can see a difference between a variables levels and total claim amount, then that variable may be an important predictor that we would want to include in a predictive model. However, similarly to the previous histogram and scatter plot, if prior knowledge leads us to believe that a predictor should be included in our initial model, we will include it anyway.

The first plot explores the Education variable, with different levels representing different levels of education achieved by the client. We can see that clients with doctorate or masters degrees tend to have slightly less total claim amounts, which is to be expected. The next variable is State, which looks at five different states in the western United States. We can see that the total claim amounts do not vary by much in different states. Furthermore, as mentioned in the project proposal feedback, all of the variables are relevant in each state (with California being the most prevalent state). Then, looking at Employment Status, we can see if clients are employed, unemployed, retired, etc. Unemployed clients have slightly higher total claim amounts and employed clients have slightly lower total claim amounts. Similarly, for Marital Status, single clients have slightly higher total claim amounts in comparison to divorced and married persons. In the plot with Coverage, we compare the different types of insurance: basic, extended, and premium. We can see that there is a large difference between premium and basic coverage. Location code is another variable that may be considered important. We can see that suburban areas have higher total claim amounts when compared to rural areas. In the next row of graphs, we first analyze Policy Type. Regardless of the policy type, the total claim amount is relatively the same. The next graph containing Vehicle Class shows that, depending on certain vehicle classes, the total claim amount is different. Specifically, Luxury Cars and Luxury SUVs tend to have the highest claim amounts. The second to last graph covers Vehicle Size. We can see that total claim amount does not vary much depending on the size of the vehicle. Lastly, the final graph covers Gender. As can be seen, males see slightly higher total claims than females, with a few more high outliers. From these graphs we can expect Employment Status, Location Code, and Vehicle Class to be good predictors of total claim amounts. From prior knowledge, we also expect Education, Employment Status, Marital Status, Vehicle Size, and Gender to play a role in the total claim amount. Despite what the box plot says, we will not include Coverage in our initial model due to the fact that it will have high correlation with Monthly Premium (see next plot) and Vehicle Class. Lastly, we will also include State in the initial model as that will directly address instructor feedback from the project

proposal. It is important to note that all of these factor levels have a large number of observations in them, and thus we will have no issues when including them in our models.




The above scatter plots explore the relationship between total claim amount and our continuous predictors. In these scatter plots, we are searching for any clear relationships between total claim amount and each of the variables. The red lines shown above are the lines of best fit, however, none of these lines do a good job of representing the corresponding plots. The plots of months since last claim, number of open complaints, and number of policies all have a line of best fit with zero slope. This tells us that these variables would probably not be good predictors of total claim amount. The plots with positive or negative slopes for the line of best fit may make better predictors of total claim amount, since this value is changing depending on the value of the predictor. In summary, the variables income, monthly premium, and lifetime value may be significant predictors of total claim amount, and will thus be included in an initial model. However, we will also include number of policies (total number of claims) and months since last claim into our initial model, since we believe they are indicators of driver safety and could be used to predict the severity of a future accident.

Finally, as emphasized in our project proposal feedback, we want to examine the possibility of interaction effects between various predictors. Different factors may interact with one another, which in turn many influence the total claim amounts, which could be an important aspect to implement into our models. To accomplish this, we examined five different models to predict total claim amount, each including a different interaction effect. We chose these interaction effects based on which variables we thought were related to one another, as well as some of the interactions we were asked to examine in the project proposal feedback. These effects include Income and Education, Gender and Marital Status, Gender and Employment Status, Vehicle Class and Vehicle Size, and finally Income and Location Code. From running ANOVA tests and exploring model summaries, we saw that some of these effects were significant. However, only certain levels of each predictor were significant and since these effects add more complex relationships into our models, we decided to avoid adding these interactions in our initial model.

In conclusion, based on our exploratory visualizations, prior knowledge, and feedback from our project proposal, our initial model will include the following variables as predictors: Gender, Education, State, Employment Status, Marital Status, Location Code, Vehicle Class, Vehicle Size, Income, Monthly Premium, Lifetime Customer Value, Number of Policies, and Months Since Last Claim. We will also use the square root of total claim amounts to try and make our data normally distributed.

# 6 Research Question 2: Which Model is Best at Predicting Total Claim Amount?

In this research question, we split the data set in half in order to make a training and a testing set. Using the training set, as well as the initial model formulated in research question 1, we built multiple models to test on the basis of predictive accuracy using K-fold cross validation. In order to build these models, we used ten different selection criterion along with the `stepwise` function in R to iteratively choose which predictors to include in the model (4). The selection criterion that were used to build our models were: Akeike Information Criterion (AIC), corrected Akeike Information Criterion (AICc), Bayesian Information Criterion (BIC), Mallow's Cp (CP), Hannan-Quinn Information Criterion (HQ), corrected Hannan-Quinn Information Criterion (HQc), R squared (Rsq), adjusted R squared (adjRsq), Significant Levels (SL), and Schwarz Bayesian Criterion (SBC). All of these selection criterion seek to either maximize the likelihood of the data, or explain the most variability in the data, while at the same time keeping the number of predictors in mind. Descriptions of these different criterion can be found at sources (1), (5), and (6). Below is a table that summarizes the results from running backward selection using all ten of the aforementioned selection criterion.

| | Criterion | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | AIC | AICc | BIC | CP | HQ | HQc | Rsq | adjRsq | SL | SBC |
| 1 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Customer.Lifetime.Value | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Education | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| EmploymentStatus | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Gender | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Income | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Location.Code | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Marital.Status | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Monthly.Premium.Auto | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Vehicle.Class | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| Vehicle.Size | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | |
| State | | | | | ✓ | ✓ | ✓ | | | |
| Months.Since.Last.Claim | | | | | ✓ | ✓ | ✓ | | | |
| Number.of.Policies | | | | | ✓ | ✓ | ✓ | | | |

Table 1: Variables Selected by Each Criterion

As seen in the above table, the ten selection criterion chose three unique models in total. In particular, AIC, AICc, BIC, CP, adjRsq, and SL all chose a model with 10 predictors. Furthermore, HQ, HQc and Rsq selected a model with 13 predictors (the full model). Lastly, SBC selected a model with only 5 predictors. After this, K-fold cross validation was ran on all three of the models in order to determine which one performs the best on "new data." The best model in terms of accuracy will be selected as the final model. The table below summarizes the results from the K-fold cross validation step.

| | RMSE | Rsquared | MAE | RMSESD | RsquaredSD | MAESD |
|---|---|---|---|---|---|---|
| 1 | 2.802399 | 0.8317720 | 2.112475 | 0.1776773 | 0.01018507 | 0.10638653 |
| 2 | 2.799137 | 0.8321897 | 2.110090 | 0.1370693 | 0.01402996 | 0.09784324 |
| 3 | 2.813956 | 0.8303716 | 2.129564 | 0.1396840 | 0.01300625 | 0.07917600 |

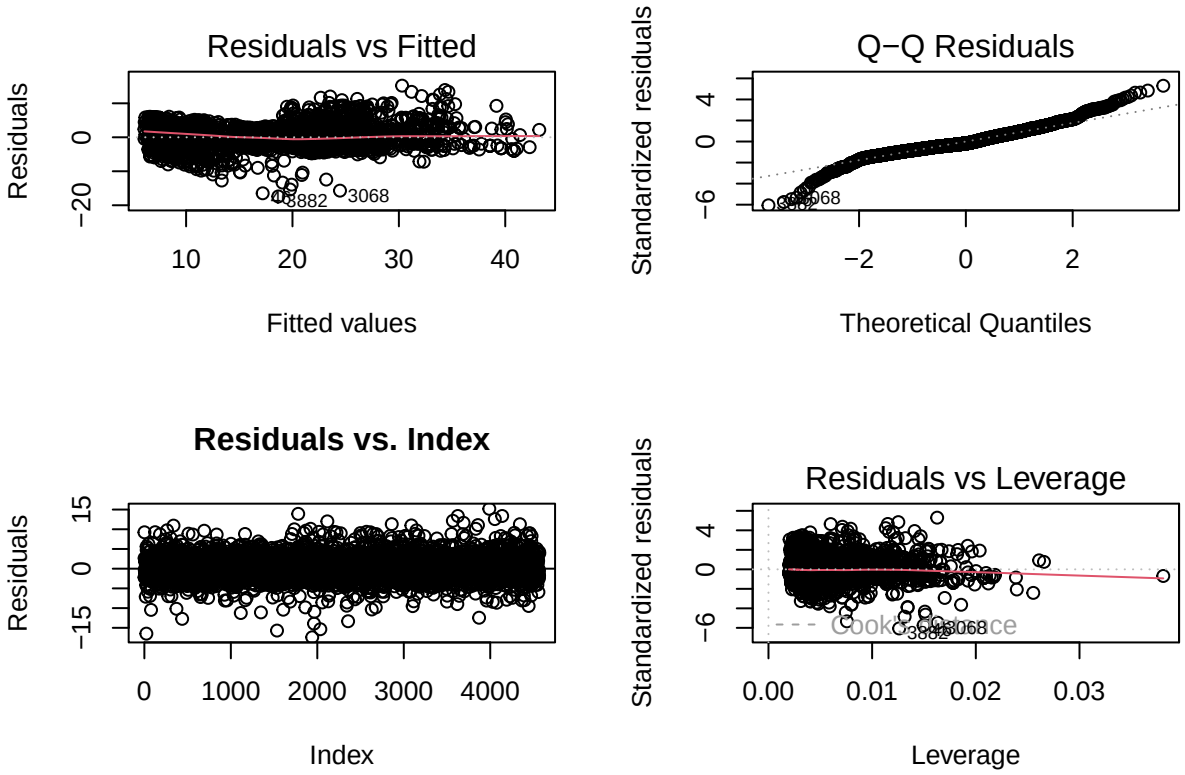
Table 2: Model Performance Metrics

The three measures used to quantify the predictive accuracy of our three models were Root Mean Squared Error (RMSE), R Squared (Rsquared), and Mean Average Error (MAE). The estimated standard deviations of these estimates were also reported. Although the results were very similar, model 2, the one with 10 predictors, had the lowest RMSE, MAE, and the highest Rsquared, while also having the most precise estimates. Ironically, the model selected by the Rsq selection criterion didn't have the highest R squared value in the K-fold cross validation step. This emphasizes one of the weaknesses of step wise selection. With that being said, model 2, which contains the variables: Customer Lifetime Value, Education, Employment Status, Gender, Income, Marital Status, Monthly Premium, Vehicle Class, and Vehicle Size, will be our final model. This model will be analyzed further in the next research question.

# 7 Research Question 3: Does the Selected Model Satisfy Regression Assumptions and Other Criteria?

In this research question, using the testing set and the final model selected in research question 2, we fit the final model to obtain parameter estimates of all of our predictors. From there, we tested all of our parameter estimates using t-tests, and ran a residual analysis to check the model assumptions. In the residual analysis we used diagnostic plots, as well as a few numerical tests to ensure model validity. We also checked our model for any outliers, leverage points, as well as the presence of any multicollinearity.

After fitting the final model using the testing data set, we used t-tests to check the significance of our predictors/parameter estimates. Out of our 10 predictors, the ones that were completely significant at the 5% level were: Gender, Income, Location, Monthly Premium Auto, and for completion, the intercept. The predictors that had one or more significant levels at the 5% level were: Education, Employment Status, Marital Status, and Vehicle Class. The predictors that were insignificant at any reasonable significance level were Customer Lifetime Value and Vehicle Size. It is important to note that many of the categorical variables may have had more significant levels had we changed the reference level. Furthermore, this model had a residual standard error of 2.89 and an adjusted R-squared value of 0.8312, which are both good indicators of the model fit and predictive accuracy. After this model was fit, we ran a residual analysis using four diagnostic plots, as well as two numeric tests; the Breusch-Pagan Test, and the Shapiro-Wilk Test. These plots are shown below.




The Residual vs. Fitted plot showed us that the errors have mean zero assumption is most likely satisfied. Furthermore, this plot also gave us evidence that the assumption of equal variance of the errors, while most likely violated, isn't violated by that much. The p-value of Breusch-Pagan Test was $1.7702147 \times 10^{-105}$, which gave us significant evidence that this equal variance assumption was violated, as expected. The last thing that we can conclude from this plot is that we have a few very large residuals. However, it is important to note that in terms of predicting total claim amounts, which is a very volatile variable with a large spread, a raw residual of around 20 isn't that concerning. Moving onto the QQ plot of the residuals, we can see that, although the transformed response (total claim amount) is more normal than it previously was, due to the deviations at the tails we can see that the normality of the errors assumption is most likely violated. The p-value of the Shapiro-Wilk Test was $1.4349714 \times 10^{-32}$, which gave us significant evidence that the normality assumption was violated, as expected. This non-normality of the response was not a big deal, since our testing set had 4567 observations, which meant that any test we ran had the Central Limit Theorem applied to it. Lastly, due to the lack of pattern in the Residuals vs. Index plot, we safely concluded that the independence of the errors assumption was satisfied in this model.

Once we had checked all of the assumptions of our regression model, we tested other indicators of model fit such as the presence of multicollinearity, outliers, and leverage points, which we were requested to analyze in the project proposal feedback. Based on the Residuals vs. Leverage plot, there appears to be a few data points that could be deemed as outliers or leverage points. In terms of the outlier heuristic of having a standardized residual greater than 3, there were 66 outliers. Similarly, in terms of the leverage heuristic of having a leverage value greater than 2 times the average leverage, there were 324 leverage points. Lastly, in terms of the Cook's distance heuristic, calculated as $D_i \geq F_{0.5,p+1,n-p-1}$, which in our case was 0.9725067, none of our individuals in the

test set have a large Cook's distance. In the context of our problem, these higher residual values don't immediately indicate the presence of an outlier. However, we decided to check these leverage points, to further understand why they're considered leverage points in the first place.

Based on plots similar to the ones used in the exploratory analysis (omitted due to space limitations), we observed that there were virtually no differences between the makeup of the leverage point group and the rest of the data. Thus, we concluded that to be a leverage point, an individual had to have a combination of extreme values across different factors. These plots gave us evidence that the individuals considered as leverage points don't differ from the other individuals in terms of general characteristics and makeup. Lastly, we checked for multicollinearity through the use of the Generalized Variance Inflation Factor (GVIF), which is an extension of the Variance Inflation Factor (VIF) that accounts for categorical variables. The results of using the `vif` function in R are shown below.

| Predictor | GVIF | Df | $GVIF^{\frac{1}{2\times Df}}$ |
|---|---|---|---|
| Customer.Lifetime.Value | 1.202357 | 1 | 1.096520 |
| Education | 1.068521 | 4 | 1.008319 |
| EmploymentStatus | 3.412050 | 4 | 1.165808 |
| Gender | 1.017946 | 1 | 1.008933 |
| Income | 3.086289 | 1 | 1.756784 |
| Location.Code | 1.441631 | 2 | 1.095755 |
| Marital.Status | 1.256761 | 2 | 1.058798 |
| Monthly.Premium.Auto | 4.788263 | 1 | 2.188210 |
| Vehicle.Class | 4.714441 | 5 | 1.167732 |
| Vehicle.Size | 1.056882 | 2 | 1.013927 |

Table 3: Variance Inflation Factors (VIF) for Predictors

Since none of our predictors had a GVIF value of above 5, we concluded that there wasn't any severe multicollinearity present in our final model, which is mainly due to our choice of leaving out Coverage Type from our initial model. Overall, our model fits the data surprisingly well, with only a few assumptions marginally violated, as well as passing other heuristics such as lack of multicollinearity and serious outliers.

# 8 Conclusion

Overall, our final model chosen through step wise selection and K-fold cross validation, predicts new data with surprising accuracy (given the range of the response). Furthermore, this model minimally violated a few assumptions and heuristics, but for the most part it was free from any serious outliers and multicollinearity.

Some further analysis that could improve the performance of this model include: further analysis of outliers and leverage points to understand what makes them leverage points, as well as further analysis of individuals who have total claim amounts over 1000 U.S. Dollars to see if these individuals are significantly different from the rest of the people in the data set. If it is the case that these individuals can be labeled as significantly different, this model can be improved through the use of a classifier that first classifies an individual as having a claim amount above or below 1000 U.S. dollars, and then running a separate regression model dependent on this result.

# 9 References


(1) Pannier, S. (2021). Combining data sources: A path to improved ... - refubium. CombiningDataSources: A Path to Improved Understanding and Prediction. https://refubium.fu-berlin.de/bitstream/handle/fub188/31720/Diss_Pannier.pdf?sequence=3&isAllowed=y

(2) Madhushreesannigrahi. (2022a, July 24). Jenks natural breaks and K-means clustering. Kaggle. https://www.kaggle.com/code/madhushreesannigrahi/jenks-natural-breaks-and-k-means-clustering

(3) Sarkar, R. (2019, May 31). Insurance. Kaggle. https://www.kaggle.com/datasets/ranja7/vehicle-insurance-customer-data

(4) Li, J., Lu, X., Cheng, K., & Liu, W. (2022). StepReg: Stepwise regression analysis. https://cran.r-project.org/web/packages/StepReg/StepReg.pdf

(5) Mohamad. (2016). Appendix E: Hannan-Quinn Information Criterion (HQC) – help center. Hannan-Quinn Information Criterion (HQC). https://support.numxl.com/hc/en-us/articles/215531183-Appendix-E-Hannan-Quinn-Information-Criterion-HQC-

(6) Bayesian information criterion. Bayesian Information Criterion - an overview | ScienceDirect Topics. (2005). https://www.sciencedirect.com/topics/social-sciences/bayesian-information-criterion


# 10 Appendix

```
# Load in the data (Tanner's CSV)
data <- read_csv("AutoInsurance.csv")

# Load in the full data set
data2 <- read.csv("AutoInsurance.csv", stringsAsFactors = TRUE)

# Remove the columns that we don't deem necessary (from prior knowledge and
# exploratory analysis)
data_new <- data2[,-c(1, 4, 5, 7, 15, 16, 18, 19, 20, 21)]

# Split the sample in half in order to find a training and testing set
sample <- sample.split(data_new, SplitRatio = 0.5)
train  <- subset(data_new, sample == TRUE)
test   <- subset(data_new, sample == FALSE)

# Split the data to the observations above and below 1000 dollars in claims
above_1000 <- data2[which(data2$Total.Claim.Amount >= 1000),]
below_1000 <- data2[-which(data2$Total.Claim.Amount >= 1000),]
```

```r
# Histogram of total claims
plot1 <-  ggplot(data = data, aes(x = `Total Claim Amount`)) +
  geom_histogram(fill = "skyblue", color = "black") +
  labs(title = "Distribution of Total Claim Amount",
       x = "Total Claim Amount",
       y = "Frequency") +
  theme(plot.title = element_text(hjust = 0.5))

# Histogram of the square root of total claims
plot2 <-  ggplot(data = data, aes(x = sqrt(`Total Claim Amount`))) +
  geom_histogram(fill = "green", color = "black") +
  labs(title = "Distribution of sqrt(Claim Amount)",
       x = "sqrt(Total Claim Amount)",
       y = "Frequency") +
  theme(plot.title = element_text(hjust = 0.5))

# Merge the plots together
ggarrange(plot1, plot2, ncol = 2, nrow = 1)
```

```r
# Compare aspects of the individuals with above and below 1000 claim amount
above_1000$Categories <- "Above 1000"
below_1000$Categories <- "Below 1000"

combined_data <- rbind(above_1000, below_1000) %>%
  group_by(Categories, Gender) %>%
  summarise(Count = n()) %>%
  mutate(Percentage = Count / sum(Count) * 100)

p1 <- ggplot(aes(x=Gender, y=Percentage, fill=Categories), data = combined_data) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Percentage of Gender Levels",
       x="Gender",
       y="Percentage (%)") +
  scale_fill_manual(values=c("Above 1000"="blue", "Below 1000"="red"))

p2 <- ggplot(rbind(above_1000, below_1000), aes(x=Categories, y=Income,
                                                fill=Categories)) +
  geom_boxplot() +
  labs(title="Boxplot of Annual Income",
       x="Total Claim Cost",
       y="Annual Income") +
  scale_fill_manual(values=c("Above 1000"="blue", "Below 1000"="red")) +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 7))

combined_data <- rbind(above_1000, below_1000) %>%
  group_by(Categories, Marital.Status) %>%
```

```
  summarise(Count = n()) %>%
  mutate(Percentage = Count / sum(Count) * 100)

p3 <- ggplot(aes(x=Marital.Status, y=Percentage, fill=Categories),
             data = combined_data) +
  geom_bar(stat="identity", position="dodge") +
  labs(title="Percentage of Education Levels",
       x="Education Levels",
       y="Percentage (%)") +
  scale_fill_manual(values=c("Above 1000"="blue", "Below 1000"="red")) +
  theme(axis.text.x = element_text(size = 8))

p4 <- ggplot(rbind(above_1000, below_1000), aes(x=Categories,
                                                y=Monthly.Premium.Auto,
                                                fill=Categories)) +
  geom_boxplot() +
  labs(title="Boxplot of Monthly Premium",
       x="Total Claim Cost",
       y="Monthly Premium") +
  scale_fill_manual(values=c("Above 1000"="blue", "Below 1000"="red")) +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 8))

ggarrange(p1,p2,p3,p4,ncol = 2,nrow = 2)
```

```
# Box plot with Education
plot1 <- ggplot(data = data, aes(x = Education, y = `Total Claim Amount`)) +
  geom_boxplot(fill = "lightsalmon3", color = "black") +
  labs(title = "Education",
       x = "",
       y = "Total Claim Amount") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))

# Box plot with State
plot2 <- ggplot(data = data, aes(x = State, y = `Total Claim Amount`)) +
  geom_boxplot(fill = "orchid3", color = "black") +
  labs(title = "State",
       x = "",
       y = "") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))

# Box plot plot of Total Claim Amount to Employment Status
plot3 <- ggplot(data, aes(x = EmploymentStatus, y = `Total Claim Amount`,
                          fill = EmploymentStatus)) +
  geom_boxplot(fill = "lightgreen", color = "black") +
```

```
  labs(title = "Employment Status",
       x = "",
       y = "") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))

# Box plot plot of Total Claim Amount to Marital Status
plot4 <- ggplot(data, aes(x = `Marital Status`, y = `Total Claim Amount`,
                          fill = `Marital Status`)) +
  geom_boxplot(fill = "khaki", color = "black") +
  labs(title = "Marital Status",
       x = "",
       y = "Total Claim Amount") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))

# Box plot plot of Total Claim Amount to Coverage
plot5 <- ggplot(data, aes(x = Coverage, y = `Total Claim Amount`,
                          fill = Coverage)) +
  geom_boxplot(fill = "seashell3", color = "black") +
  labs(title = "Coverage",
       x = "",
       y = "") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))

# Box plot plot of Total Claim Amount to Location Code
plot6 <- ggplot(data, aes(x = `Location Code`, y = `Total Claim Amount`,
                          fill = `Location Code`)) +
  geom_boxplot(fill = "lavender", color = "black") +
  labs(title = "Location Code",
       x = "",
       y = "") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5))

# Plot comparing Total Claim Amount to Policy Type
plot7 <- ggplot(data, aes(x = `Policy Type`, y = `Total Claim Amount`,
                          fill = `Policy Type`)) +
  geom_boxplot() +
  labs(title = "Policy Type",
       x = "",
       y = "Total Claim Amount") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  guides(fill = FALSE)
```

```r
# Plot comparing Total Claim Amount to Vehicle Class
plot8 <- ggplot(data, aes(x = `Vehicle Class`, y = `Total Claim Amount`,
                          fill = `Vehicle Class`)) +
  geom_boxplot() +
  labs(title = "Vehicle Class",
       x = "",
       y = "") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.title.y = element_blank()) +
  guides(fill = FALSE)

# Plot comparing Total Claim Amount to Vehicle Size
plot9 <- ggplot(data, aes(x = `Vehicle Size`, y = `Total Claim Amount`,
                          fill = `Vehicle Size`)) +
  geom_boxplot() +
  labs(title = "Vehicle Size",
       x = "",
       y = "") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.title.y = element_blank()) +
  guides(fill = FALSE)

# Plot comparing Total Claim Amount to Gender
plot10 <- ggplot(data, aes(x = Gender, y = `Total Claim Amount`, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Gender",
       x = "",
       y = "") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1)) +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.title.y = element_blank()) +
  guides(fill = FALSE)

# Merge the plots together
ggarrange(plot1, plot2, plot3, ncol = 3, nrow = 1)
ggarrange(plot4, plot5, plot6, ncol = 3, nrow = 1)
ggarrange(plot7, plot8, ncol=2, nrow=1)
ggarrange(plot9, plot10, ncol = 2, nrow = 1)
```

```r
# Scatter plot for Income
plot1 <- ggplot(data, aes(x = Income, y = `Total Claim Amount`)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Income", x = "", y = "Total Claim Amount") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1),
```

```
        plot.title = element_text(hjust = 0.5))

# Scatter plot for Monthly Premium Auto
plot2 <- ggplot(data, aes(x = `Monthly Premium Auto`, y = `Total Claim Amount`)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Monthly Premium", x = "", y = "") +
  theme(axis.title.y = element_blank(),
        axis.text.x = element_text(angle = 35, hjust = 1),
        plot.title = element_text(hjust = 0.5))

# Scatter plot for Customer Lifetime Value
plot3 <- ggplot(data, aes(x = `Customer Lifetime Value`,
                          y = `Total Claim Amount`)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Lifetime Value", x = "", y = "") +
  theme(axis.title.y = element_blank(),
        axis.text.x = element_text(angle = 35, hjust = 1),
        plot.title = element_text(hjust = 0.5))

# Scatter plot for Months Since Last Claim Amount
plot4 <- ggplot(data, aes(x = `Months Since Last Claim`,
                          y = `Total Claim Amount`)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Last Claim", x = "", y = "Total Claim Amount") +
  theme(axis.title.y = element_blank(),
        axis.text.x = element_text(angle = 35, hjust = 1),
        plot.title = element_text(hjust = 0.5))

# Scatter plot for Number of Open Complaints
plot6 <- ggplot(data, aes(x = `Number of Open Complaints`,
                          y = `Total Claim Amount`)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Complaints", x = "", y = "") +
  theme(axis.title.y = element_blank(),
        axis.text.x = element_text(angle = 35, hjust = 1),
        plot.title = element_text(hjust = 0.5))

# Scatter plot for Number of Policies
plot7 <- ggplot(data, aes(x = `Number of Policies`, y = `Total Claim Amount`)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Num Policies", x = "", y = "") +
  theme(axis.title.y = element_blank(),
```

```
        axis.text.x = element_text(angle = 35, hjust = 1),
        plot.title = element_text(hjust = 0.5))

ggarrange(plot1, plot2, plot3, ncol = 3, nrow = 1)
ggarrange(plot4, plot6, plot7, ncol = 3, nrow = 1)
```

```
# Checking the interaction of Income and Education
no_interaction <- lm(`Total Claim Amount` ~ Income + Education, data = data)
interaction <- lm(`Total Claim Amount` ~ Income * Education, data = data)

anova_result <- anova(no_interaction, interaction)
print(anova_result)

summary(interaction)

# Checking the interaction of Gender and Marital Status
no_interaction <- lm(`Total Claim Amount` ~ Gender + `Marital Status`, data = data)
interaction <- lm(`Total Claim Amount` ~ Gender * `Marital Status`, data = data)

anova_result <- anova(no_interaction, interaction)
print(anova_result)

summary(interaction)

# Checking the interaction of Gender and Employment Status
no_interaction <- lm(`Total Claim Amount` ~ Gender + `EmploymentStatus`, data = data)
interaction <- lm(`Total Claim Amount` ~ Gender * `EmploymentStatus`, data = data)

anova_result <- anova(no_interaction, interaction)
print(anova_result)

summary(interaction)

# Checking the interaction of Vehicle Class and vehicle Size
no_interaction <- lm(`Total Claim Amount` ~ `Vehicle Class` + `Vehicle Size`, data = data)
interaction <- lm(`Total Claim Amount` ~ `Vehicle Class` * `Vehicle Size`, data = data)

anova_result <- anova(no_interaction, interaction)
print(anova_result)

summary(interaction)

# Checking the interaction of Income and Location Code
no_interaction <- lm(`Total Claim Amount` ~ Income + `Location Code`, data = data)
interaction <- lm(`Total Claim Amount` ~ Income * `Location Code`, data = data)

anova_result <- anova(no_interaction, interaction)
```

```r
print(anova_result)

summary(interaction)
```

```r
# Selecting models to run K-fold cross validation on based on backward selection
# using 10 different criterion.
method_names <- c("AIC", "AICc", "BIC", "CP", "HQ", "HQc", "Rsq", "adjRsq", "SL",
                  "SBC")

models <- matrix(NA, nrow = 10, ncol = 14)

for (i in 1:length(method_names)) {
  temp <- stepwise(formula=sqrt(Total.Claim.Amount)~.,
         data=train,
         include=NULL,
         selection="backward",
         select=method_names[i])
  names(temp$`Selected Varaibles`) <- NULL
  models[i,] <- c(unlist(temp$`Selected Varaibles`[1,]),
                  rep(NA, 14-as.integer(summary(temp)[4,1])))
}

rownames(models) <- method_names
colnames(models) <- c("Variable 1", "Variable 2", "Variable 3", "Variable 4",
                      "Variable 5", "Variable 6", "Variable 7", "Variable 8",
                      "Variable 9", "Variable 10", "Variable 11", "Variable 12",
                      "Variable 13", "Variable 14")
```

```r
# Define the formulas of the selected models
formula_1 <- sqrt(Total.Claim.Amount)~.

formula_2 <- as.formula(paste("sqrt(Total.Claim.Amount)~", ".-State",
                              "-Months.Since.Last.Claim",
                              "-Number.of.Policies"))

formula_3 <- as.formula(paste("sqrt(Total.Claim.Amount)~", ".-State",
                              "-Customer.Lifetime.Value", "-Education",
                              "-Income", "-Months.Since.Last.Claim",
                              "-Number.of.Policies", "-Vehicle.Class",
                              "-Vehicle.Size"))

formulas <- c(formula_1, formula_2, formula_3)
```

```r
# Run K-fold cross-validation on the selected models
accuracy <- matrix(NA, nrow = 3, ncol = 6)

for (i in 1:length(formulas)) {
```

```
  train_control <- trainControl(method = "cv",
                                number = 10)

  model <- train(formulas[[i]], data = train,
                 trControl = train_control,
                 method = "lm")

  names(model$results) <- NULL

  accuracy[i,] <- unlist(model$results)[-c(1)]
}

colnames(accuracy) = c("RMSE", "Rsquared", "MAE", "RMSESD", "RsquaredSD", "MAESD")
```

```
# Fit a model on the selected formula using the test set
chosen_model <- lm(formula=formula_2, data=test)

# Get a summary of the fitted model
summary(chosen_model)
```

```
# Breusch-Pagan Test (constant variance assumption)
bpt_results <- lmtest::bptest(chosen_model)

# Shapiro-Wilk Test (Normality assumption)
shap_results <- shapiro.test(chosen_model$residuals)

# Find the number of points that could be considered outliers
num_out <- sum(abs(rstandard(chosen_model)) > 3)

# Find the leverage point threshold for the data set
threshold <- 2*(23+1)/nrow(test)

# Number of high leverage points in model
num_lev <- length(which(hatvalues(chosen_model) > threshold))

# Find the cooks distance threshold for the data set
f_val <- qf(0.5, 24, 4543)

# Find VIF of predictors in the model
gvif_results <- vif(chosen_model)

# Create plotting frame
par(mfrow=c(2,2))

# Diagnostic Plots
plot(chosen_model, which=1)
plot(chosen_model, which=2)
```

```r
# Residuals against index
plot(resid(chosen_model)~c(1:length(resid(chosen_model))),
     main="Residuals vs. Index",  xlab="Index", ylab="Residuals")
abline(h=0)

# Diagnostic Plots
plot(chosen_model, which=5)
```