Summative

100383119

Table of contents

1 Executive Summary

The main focus of this study is to identify the variables that will influence the frequency of auto insurance renewals and pricing elasticity. This academic paper offers clients a wide range of theoretical background knowledge and strategies. The dataset provided by the clients in this study is analysed for a total of 13 variables. In the statistical model, the insurance renewal rate is used as a response variable, while the other 13 variables are regarded as potential factors that may affect the renewal rate. To investigate pricing elasticity, the price and discount rate will be highlighted. Some elements of the dataset must be adjusted based on the data provided by the client, such as dealing with missing numbers or reducing biases caused by unrealistic values. As a result, data sampling and data transformation are crucial in this investigation. In this study, supervised machine learning algorithms are used to achieve accurate and representative analytical results. The logistic regression model and random forest classification model are mostly used to analyse elements that may affect the renewal rate, and statistical models are also utilised to make predictions. To ensure the rigour of the conclusions, the receiver operating characteristic (ROC) and Area Under Curve (AUC) are used to validate the models' viability. The confusion matrix will calculate the accuracy of the models in order to provide clients with an objective analysis.

2 Introduction

Some researchers have demonstrated tremendous interest in the relationship between insurance renewal and consumers, as well as the impact of price on insurance. On the other hand, due to the increasing number of insurance companies and the convenience afforded to consumers by the Internet, competition within the insurance market has increased. Using statistical models and machine learning technologies, this project intends to provide customers with pricing strategies for vehicle insurance. Customers can better understand the influence of existing factors on the renewal rate by using logistic regression and random forest models. The cost flexible concept enables customers to be more price competitive. The questions to be answered with the analysis are:

- Which factors have the greatest effect on renewal rate?

- How does price relate to renewal rates?

- What are the key factors that affect customers' response to price increase (also seen as price elasticity)? In this case price elasticity should be defined as the impact that changes in price have on a customer's likelihood to renew

- What advice would you give to this company on how they might think about pricing these customers?

3 Data Understanding

Marital Status: This variable represents the marital status of the insurance holder, indicating whether they are single, married, divorced, etc. Age: This variable represents the age of the insurance holder, indicating their age in years. Gender: This variable represents the gender of the insurance holder, indicating whether they are male or female. Car Value: This variable represents the value of the insured car, indicating the estimated monetary value of the car. Years of No Claims Bonus: This variable represents the number of consecutive years the insurance holder has not made any insurance claims, which can result in a bonus or discount on their premium. Annual Mileage: This variable represents the estimated number of miles the customer drives their car in a year. Payment Method: This variable represents the method used by the policyholder to make payments for their insurance, such as monthly installments or annual payments. Acquisition Channel: This variable represents the channel through which the policyholder acquired their insurance, such as online, through an agent, or by phone. Years of Tenure with Current Provider: This variable represents the number of years the policyholder has been insured with their current insurance provider. Price: This variable represents the price of the insurance premium for the policyholder. Actual Change in Price vs last Year: This variable represents the actual change in the insurance price compared to the previous year. Percent Change in Price vs last Year: This variable represents the percentage change in the insurance price compared to the previous year. Grouped Change in Price: This variable represents the grouped change in the insurance price, which could be categorized into different price change groups (e.g., increase, decrease, or no change). Renewed: This variable represents whether the policyholder renewed their insurance policy or not. It is a binary variable, where 1 indicates renewal and 0 indicates non-renewal.

4 Data Preparation

4.1 Data Staging

After loading the Excel file using read_xlsx(), the clean_names() function from the janitor package is applied to clean the column names of the insurance_data dataframe. The clean_names() function converts the column names to lowercase, removes special characters, and replaces spaces with underscores.

```r
insurance_data <-read_xlsx("data/insurance_data_2023.xlsx") %>%
  janitor::clean_names()
```

The filter() function is used to remove rows where the "price" column has missing values (NA).

```
insurance_data <- insurance_data %>%
  filter(!is.na(price))
```

4.2 Stage the factor value

In this stage the variables are categorized or classified in different levels of a categorical variable (factor) into specific groups or stages.

The "renewed" column of the insurance_data dataframe will be converted to a factor variable.This conversion allows R to treat the column as a categorical variable and apply statistical analysis or modeling techniques.

```
insurance_data$renewed <- factor(insurance_data$renewed,
                      levels =c ("0","1"),
                      labels =c("No","Yes"))
```

```
insurance_data$marital_status <- factor(insurance_data$marital_status)
```

A new column called "new_marital_status" is created using the case_when() function. This function allows for conditional transformations based on the values of the "marital_status" column. If the "marital_status" is equal to "M", the corresponding value in the "new_marital_status" column will be set to "Married". For all other cases, the value will be set to "Not Married".

```
insurance_data <- insurance_data %>%
mutate(new_marital_status = case_when(
marital_status =="M" ~ "Maried",
TRUE ~ "Not Maried")) %>%

mutate(new_marital_status =
         factor(new_marital_status,
 levels = c("Not Maried","Maried"),
 labels = c("Not Maried","Maried")))
```

```
insurance_data$payment_method <- factor(insurance_data$payment_method)
```

```
insurance_data$acquisition_channel<-factor(insurance_data$acquisition_channel)
```

```r
insurance_data <- insurance_data %>%
  filter(gender!="C")
```

```r
insurance_data$gender <- factor(insurance_data$gender,
                         levels= c("M","F"), labels= c("Male", "Female"))
```

4.3 Descriptive Statistics

Descriptive statistics is used to get a clear overview of the data, allowing for a better understanding of its properties and patterns. Here we will be anlysing the insurance dataset using descriptive method to obtain the required output.

```r
summary(insurance_data$price)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  96.01  264.29  357.35  422.57  501.69 4449.88 
```

The minimum value represents the lowest observed price, the maximum value represents the highest observed price, and the quartiles provide information about the spread of the data within the column. The mean represents the average price value in the dataset

```r
table1::table1(~price+car_value+
                 years_of_no_claims_bonus+
                 annual_mileage+gender+
                 new_marital_status+
                 age+payment_method+
                 acquisition_channel+
                 years_of_tenure_with_current_provider+
                 actual_change_in_price_vs_last_year+
                 percent_change_in_price_vs_last_year|
                 renewed,data =
                 insurance_data)
```

```
Get nicer `table1` LaTeX output by simply installing the `kableExtra` package
```

	No	Yes	Overall
	(N=7575)	(N=12422)	(N=19997)

	No	Yes	Overall
price			
Mean (SD)	473 (318)	392 (216)	423 (262)
Median [Min, Max]	391 [103, 4450]	341 [96.0, 3470]	357 [96.0, 4450]
car_value			
Mean (SD)	3860 (4090)	3580 (3910)	3690 (3980)
Median [Min, Max]	2500 [0, 60000]	2200 [1.00, 60000]	2500 [0, 60000]
years_of_no_claims_bonus			
Mean (SD)	5.64 (2.96)	5.83 (2.85)	5.76 (2.89)
Median [Min, Max]	6.00 [0, 9.00]	6.00 [0, 9.00]	6.00 [0, 9.00]
annual_mileage			
Mean (SD)	6700 (3700)	6420 (3510)	6530 (3580)
Median [Min, Max]	6000 [1.00, 60000]	5000 [2.00, 70000]	5200 [1.00, 70000]
gender			
Male	4186 (55.3%)	6757 (54.4%)	10943 (54.7%)
Female	3389 (44.7%)	5665 (45.6%)	9054 (45.3%)
new_marital_status			
Not Maried	3522 (46.5%)	5702 (45.9%)	9224 (46.1%)
Maried	4053 (53.5%)	6720 (54.1%)	10773 (53.9%)
age			
Mean (SD)	44.2 (13.2)	45.0 (12.3)	44.7 (12.7)
Median [Min, Max]	43.0 [17.0, 89.0]	44.0 [17.0, 89.0]	44.0 [17.0, 89.0]
payment_method			
Annual	2840 (37.5%)	2646 (21.3%)	5486 (27.4%)
Monthly	4735 (62.5%)	9776 (78.7%)	14511 (72.6%)
acquisition_channel			
Aggreg	5 (0.1%)	7 (0.1%)	12 (0.1%)
Direct	1481 (19.6%)	2456 (19.8%)	3937 (19.7%)
Inbound	6086 (80.3%)	9959 (80.2%)	16045 (80.2%)
Outbound	3 (0.0%)	0 (0%)	3 (0.0%)
years_of_tenure_with_current_provider			
Mean (SD)	2.40 (0.845)	2.53 (0.854)	2.48 (0.853)
Median [Min, Max]	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]
actual_change_in_price_vs_last_year			
Mean (SD)	42.1 (574)	1.15 (266)	16.7 (411)
Median [Min, Max]	33.1 [-20600, 37000]	8.58 [-7680, 16000]	15.6 [-20600, 37000]
percent_change_in_price_vs_last_year			
Mean (SD)	0.230 (5.11)	0.0561 (0.804)	0.122 (3.21)
Median [Min, Max]	0.109 [-9.11, 441]	0.0283 [-38.7, 55.9]	0.0525 [-38.7, 441]

4.4 Data Visualisation

Data visualisation is used to represent datasets graphically to understand the data and make decisions based on them.

This is used to create a bar plot of the "years_of_tenure_with_current_provider" variable in the insurance_data dataframe.It shows majority of the people have tenure of 2 years followed by 3,4 and 1.

```
insurance_data %>%
ggplot(aes(years_of_tenure_with_current_provider))+geom_bar()
```

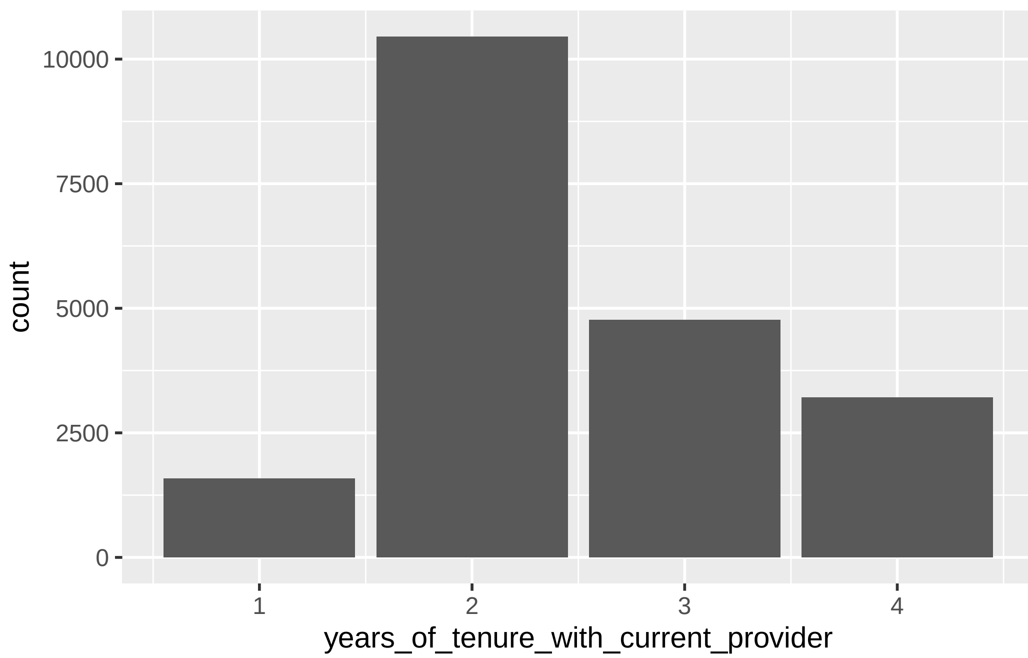
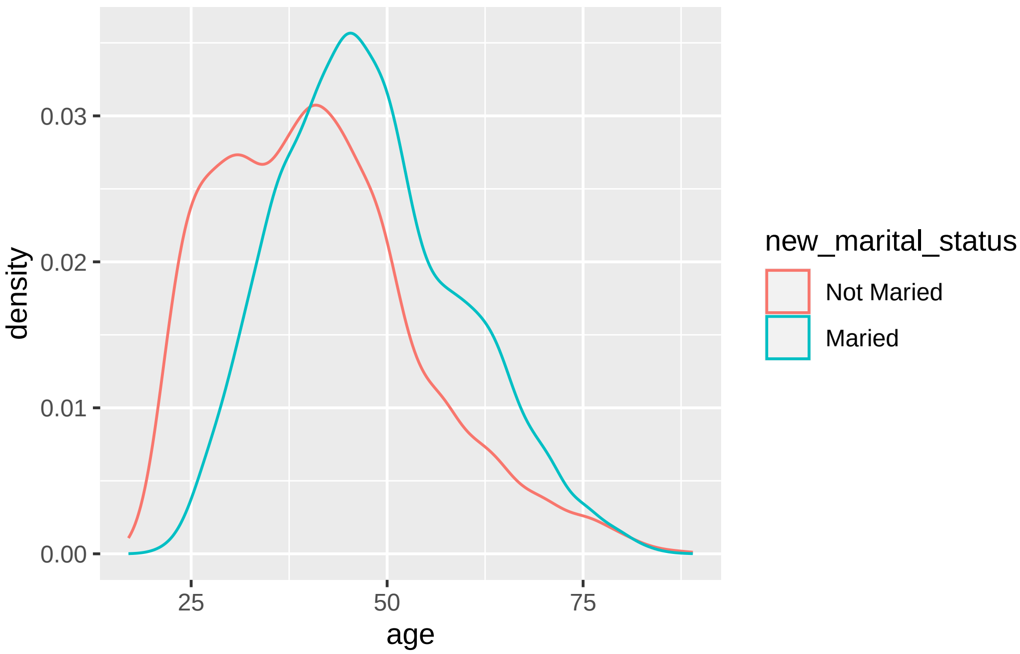


Figure 1: bar plot

The mean and median of the "age" column are computed using the mean() and median() functions, respectively. Then, the ggplot() function is used to create a density plot of age, with the "new_marital_status" variable mapped to the color aesthetic. The geom_density() function adds the density plot layer. This shows that the maximum insurance holders are married population.

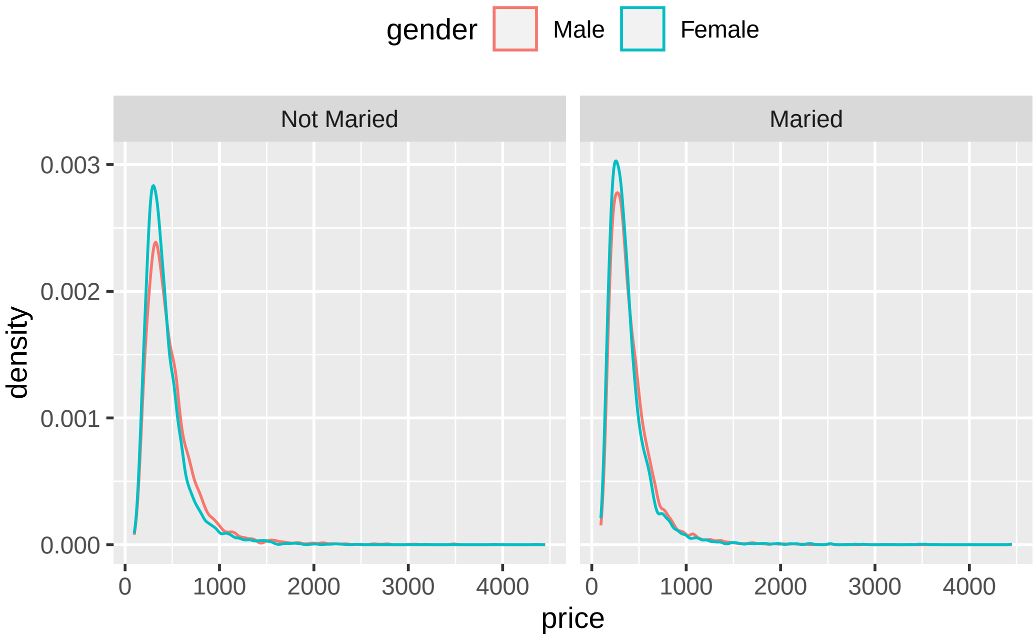
```
# |fig-cap: " age plot"
mean <- mean(insurance_data$age)
```

```r
median <- median(insurance_data$age)
insurance_data%>%
ggplot(aes(x=age, color=new_marital_status))+geom_density()
```



This provides a representation of how prices vary across gender and marital status groups in the dataset.Irrespective of the marital status, female's have highest price and unmarried men have the least.

```r
insurance_data %>%
ggplot(aes(x=price, color=gender))+geom_density()+
facet_wrap(~new_marital_status)+theme(legend.position="top")
```



The x-axis represents the “price” variable, and the y-axis represents the “renewed” variable. Each point in the plot corresponds to a data point in the “insurance_data” dataset.

```
insurance_data %>%
ggplot(aes(x=price, y=renewed)) +geom_point()
```

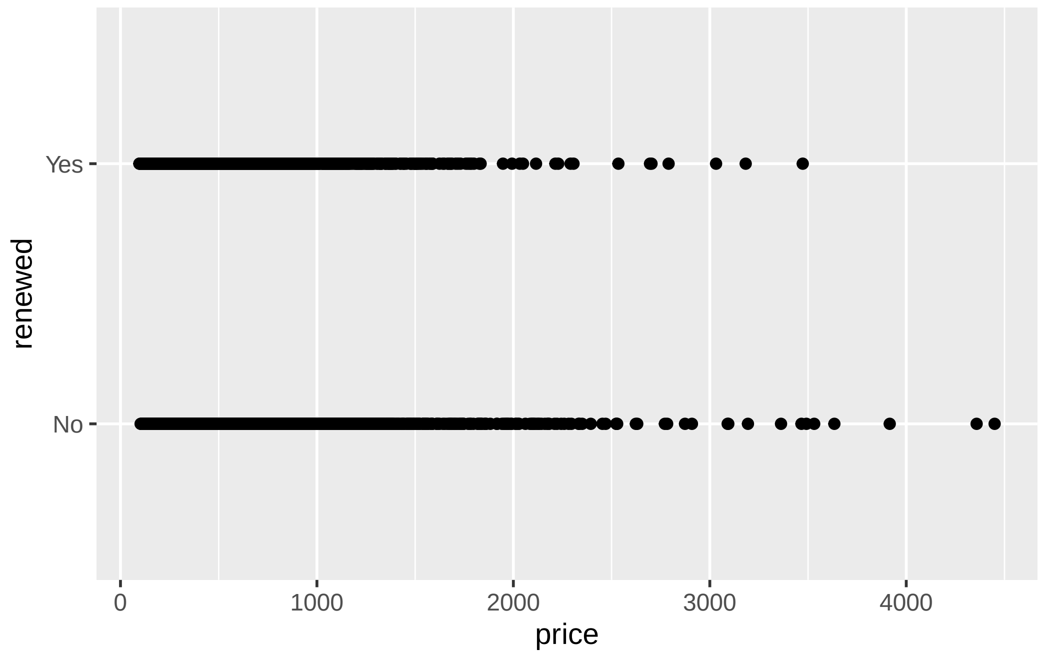
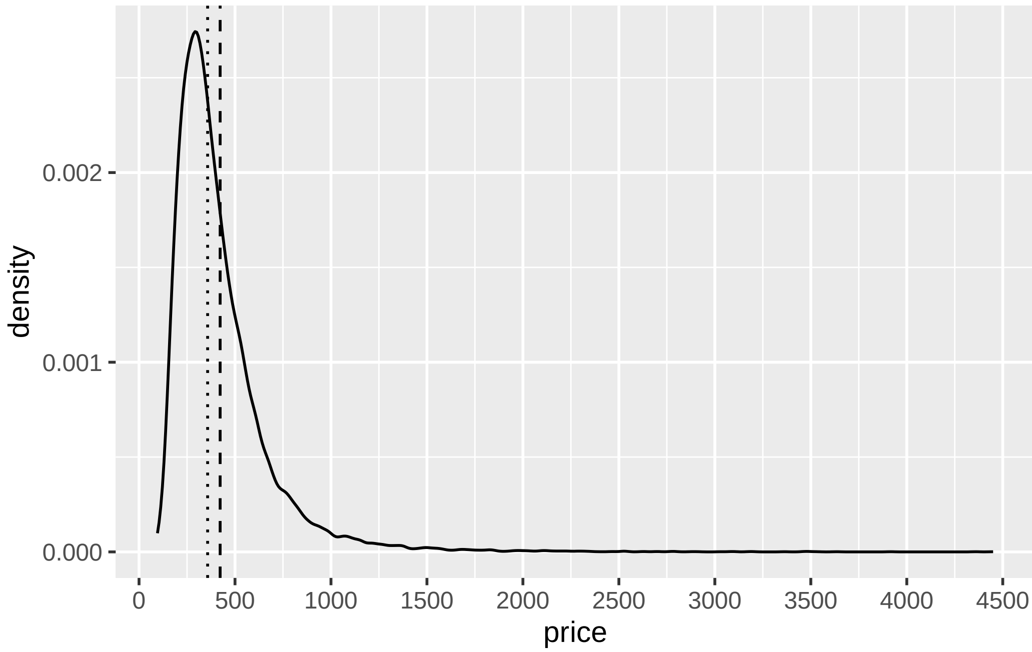


Figure 2: prise vs renewed plot

This calculates the average and median prices from the "insurance_data" dataset and adds vertical lines representing these values on top of the density plot.

```
average_price <- mean(insurance_data$price)
median_price <- median(insurance_data$price)
insurance_data %>%
  ggplot(aes(x=price)) + geom_density() + geom_vline(xintercept = average_price,
                                                     linetype = "dashed") +
  geom_vline(xintercept = median_price, linetype = "dotted") + scale_x_continuous(n.breaks
```

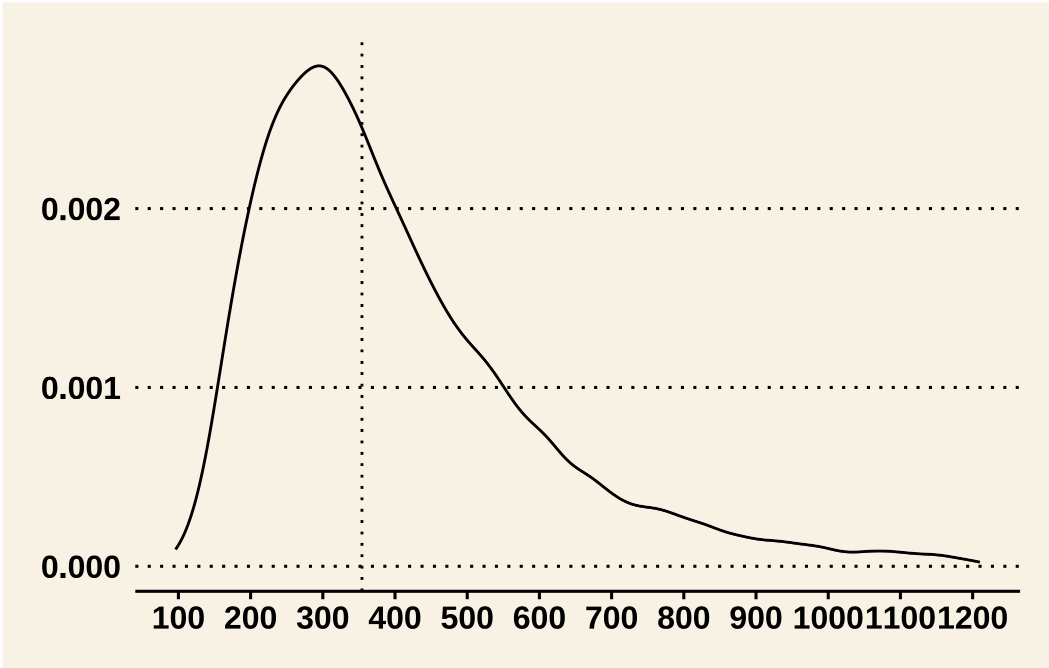


This chunk calculates the cutoff for outliers based on the mean and standard deviation of the "price" variable in the dataset. It then filters the dataset to include only the rows where the price is less than or equal to the outliers_cutoff value.

```
outliers_cutoff <- mean(insurance_data$price)+3*sd(insurance_data$price)
insurance_data <- insurance_data %>% filter(price<=outliers_cutoff)
```

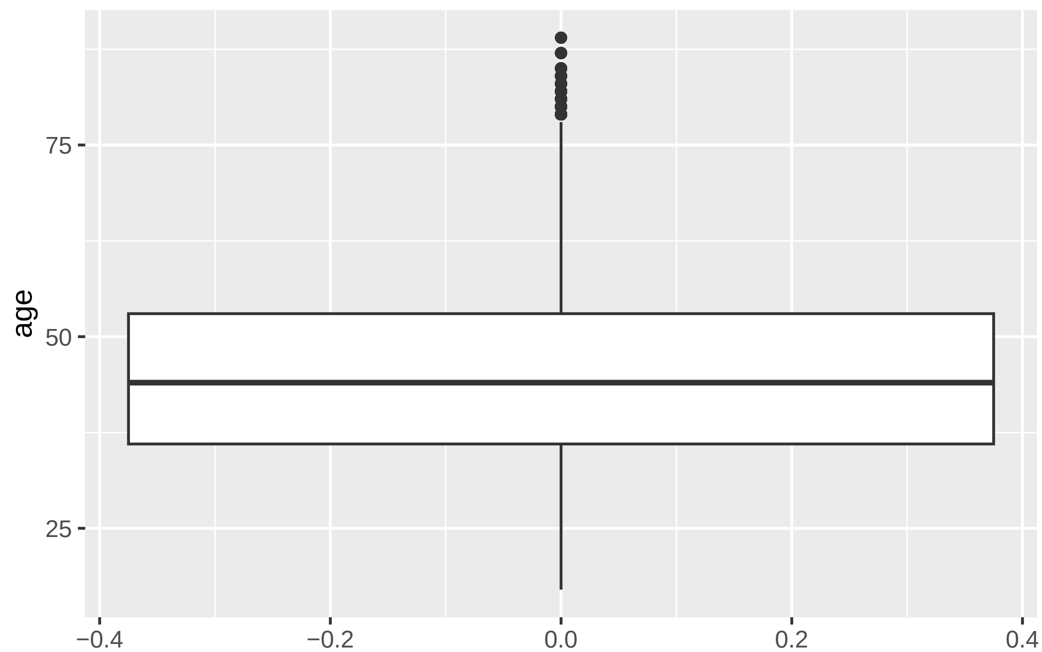
This chunk create a density plot of the "price" variable in dataset, along with a vertical line indicating the median price.

```
avg_price <- mean(insurance_data$price)
median_price <- median(insurance_data$price)
insurance_data %>%
ggplot(aes(x=price))+geom_density()+geom_vline(xintercept=median_price, linetype="dotted")
```



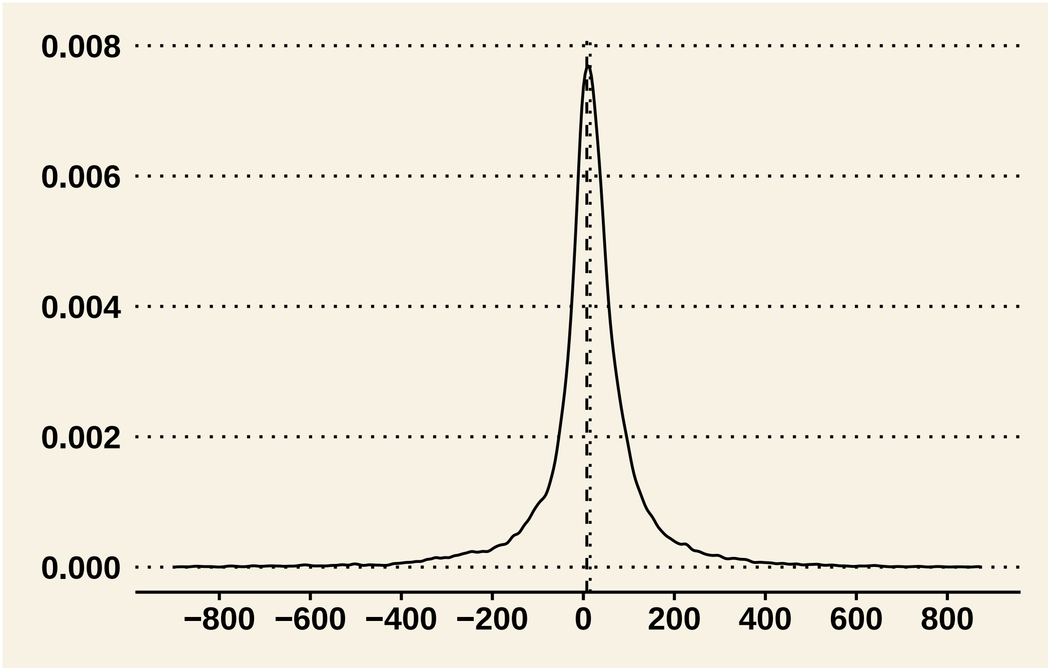
The plot, specifies the "age" variable on the y-axis and using the geom_boxplot() function to create the boxplot.

```
#|fig-cap: Price boxplot
insurance_data %>%
  ggplot(aes(y=age))+geom_boxplot()
```



This code calculates the mean and median of the "actual_change_in_price_vs_last_year" variable. It then filters the "insurance_data" dataset to exclude outliers beyond 3 standard deviations from the mean. Finally, it creates a density plot with the filtered data and adds vertical lines for the mean and median using the geom_vline() function and scale_x_continuous(n.breaks = 10) sets the number of breaks on the x-axis to 10.

```
mean_change <- mean(insurance_data$
              actual_change_in_price_vs_last_year)
median_change <- median(insurance_data$
              actual_change_in_price_vs_last_year)
outliers_change <- mean_change+3*sd(insurance_data$
                                    actual_change_in_price_vs_last_year)
insurance_data %>%
  filter(actual_change_in_price_vs_last_year<
           outliers_change &
           actual_change_in_price_vs_last_year>-
           outliers_change) %>%
  ggplot(aes(x=actual_change_in_price_vs_last_year))+
  geom_density()+
  geom_vline(xintercept = mean_change, linetype="dashed")+geom_vline(
    xintercept = median_change,linetype="dotted")+
  ggthemes::theme_wsj()+scale_x_continuous(n.breaks=10)
```



5 Correlation

Correlation refers to the statistical relationship between two or more variables. It measures the degree to which changes in one variable correspond to changes in another variable. Correlation gives insight about the strength and direction of the relationship between variables, providing insights into how they are related.

This calculates the correlation between the sum of price and actual_change_in_price_vs_last_year variables and the sum of car_value and annual_mileage and with age variables. A correlation value of 0.0817959 indicates a positive correlation between the variables.

```
cor(insurance_data$age+insurance_data$price+
       insurance_data$
       actual_change_in_price_vs_last_year,
     insurance_data$car_value+
       insurance_data$annual_mileage)
```

```
[1] 0.0817959
```

This perform a correlation test between the sum of price and actual_change_in_price_vs_last_year variables and the sum of car_value and annual_mileage variables.There is a statistically significant positive correlation between the combined price and actual_change_in_price_vs_last_year variables and the combined car_value and annual_mileage variables in the insurance data. This means that as the combined price and actual_change_in_price_vs_last_year increase, the combined car_value and annual_mileage also tend to increase.

```
cor.test(insurance_data$price+
           insurance_data$actual_change_in_price_vs_last_year,
         insurance_data$car_value+
           insurance_data$annual_mileage)
```

```
    Pearson's product-moment correlation

data:  insurance_data$price + insurance_data$actual_change_in_price_vs_last_year and insuranc
t = 11.211, df = 19649, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06581645 0.09360218
sample estimates:
      cor
0.0797248
```

There is a statistically significant positive correlation between the price and car value variables in the insurance data set.

```
cor.test(insurance_data$price,
         insurance_data$car_value)
```

```
    Pearson's product-moment correlation

data:  insurance_data$price and insurance_data$car_value
t = 21.768, df = 19649, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1397696 0.1670747
sample estimates:
      cor
0.1534514
```

5.1 T-test

This test proves that there is significant difference between the “yes” or “no” group.

```r
t.test(insurance_data$price+insurance_data$
         actual_change_in_price_vs_last_year+
         insurance_data$annual_mileage+
         insurance_data$car_value+
         insurance_data$age ~
         insurance_data$renewed)
```

```
    Welch Two Sample t-test

data:  insurance_data$price + insurance_data$actual_change_in_price_vs_last_year + insurance_
t = 7.3781, df = 14821, p-value = 1.691e-13
alternative hypothesis: true difference in means between group No and group Yes is not equal
95 percent confidence interval:
 472.1523 813.7866
sample estimates:
 mean in group No mean in group Yes
         11033.04          10390.07
```

There is a significant difference in means between the "No" and "Yes" groups in terms of car value.

```r
t.test(insurance_data$car_value ~
         insurance_data$renewed)
```

```
    Welch Two Sample t-test

data:  insurance_data$car_value by insurance_data$renewed
t = 4.8534, df = 14960, p-value = 1.226e-06
alternative hypothesis: true difference in means between group No and group Yes is not equal
95 percent confidence interval:
 168.009 395.656
sample estimates:
 mean in group No mean in group Yes
         3829.507          3547.675
```

There is a significant difference in means between the "No" and "Yes" groups in terms of price.

```r
t.test(insurance_data$price ~
         insurance_data$renewed)
```

```
	Welch Two Sample t-test

data:  insurance_data$price by insurance_data$renewed
t = 18.355, df = 13645, p-value < 2.2e-16
alternative hypothesis: true difference in means between group No and group Yes is not equal
95 percent confidence interval:
 47.95155 59.41773
sample estimates:
 mean in group No mean in group Yes
         434.1793          380.4946
```

There is a significant difference in means between the “No” and “Yes” groups in terms of age.

```r
t.test(insurance_data$age~
         insurance_data$renewed)
```

```
	Welch Two Sample t-test

data:  insurance_data$age by insurance_data$renewed
t = -3.2833, df = 14651, p-value = 0.001028
alternative hypothesis: true difference in means between group No and group Yes is not equal
95 percent confidence interval:
 -0.9933104 -0.2506608
sample estimates:
 mean in group No mean in group Yes
         44.48326          45.10524
```

5.2 ANOVA

Here we are using ANOVA test to analyze the effect of “gender” on the combination of “age” and “price”. This modifies formula specifies that “age” and “price” are the dependent variables, “gender” is the independent variable.

The p-value associated with this variable is less than 0.001, which proves the evidence to reject the null hypothesis. Therefore, there is a significant difference in the means across the different levels of the “gender” variable.

```r
test_aov <-aov(insurance_data$age+insurance_data$price ~ insurance_data$gender)
summary(test_aov)
```

```
                        Df    Sum Sq Mean Sq F value Pr(>F)    
insurance_data$gender     1   3690127 3690127   104.4 <2e-16 ***
Residuals             19649 694836523   35362                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant differences in the means of the dependent variable across the levels of each independent variables insurance_data*renewed*, *insurance*$_d$*ata*gender and insurance_data$new_marital_status.

```r
test_aov <- aov(insurance_data$age+insurance_data$price ~insurance_data$renewed+insurance_
summary(test_aov)
```

```
                                    Df    Sum Sq  Mean Sq F value Pr(>F)    
insurance_data$renewed               1  12951671 12951671   378.9 <2e-16 ***
insurance_data$gender                1   3576614  3576614   104.6 <2e-16 ***
insurance_data$new_marital_status    1  10474077 10474077   306.4 <2e-16 ***
Residuals                        19647 671524287    34179                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6 Regression Analysis

Logistic regression is commonly used when the dependent variable is binary or categorical in nature. These regressions will be performed with the help of above performed tests.

```r
insurance_data$renewed <-ordered(insurance_data$renewed)
```

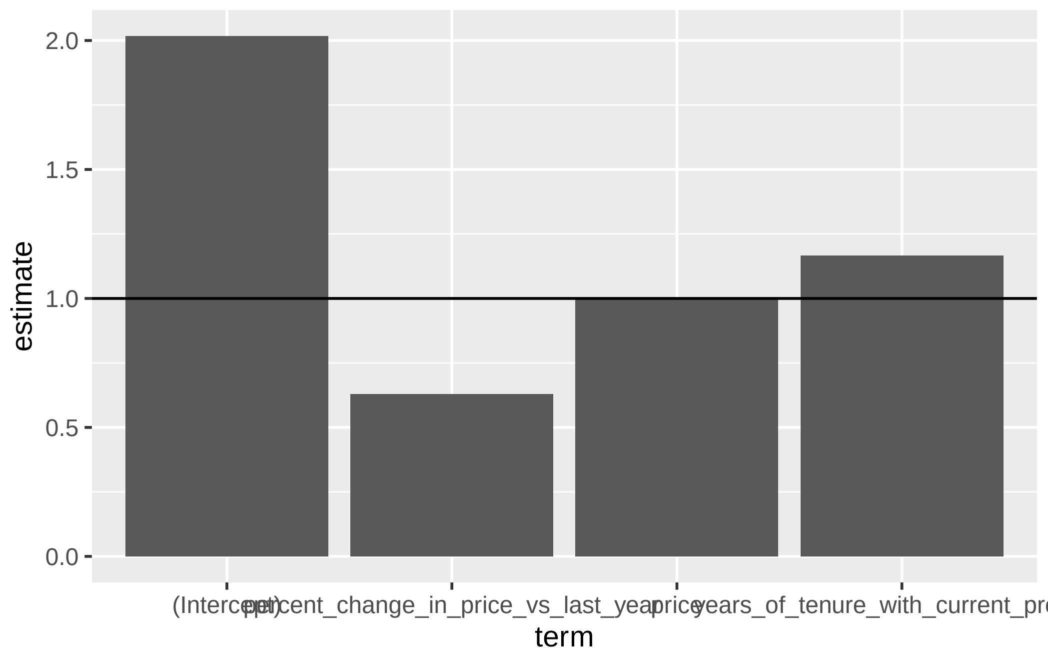
```r
regression_model <- glm(renewed~price+
                    years_of_tenure_with_current_provider+
                    percent_change_in_price_vs_last_year,
                    data = insurance_data,
                    family = "binomial")
summary(regression_model)
```

```
Call:
glm(formula = renewed ~ price + years_of_tenure_with_current_provider + 
    percent_change_in_price_vs_last_year, family = "binomial", 
    data = insurance_data)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1064  -1.3330   0.8556   0.9559   7.1628

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            7.018e-01  5.725e-02  12.259   <2e-16
price                                 -1.281e-03  7.772e-05 -16.480   <2e-16
years_of_tenure_with_current_provider  1.529e-01  1.779e-02   8.598   <2e-16
percent_change_in_price_vs_last_year  -4.650e-01  4.668e-02  -9.963   <2e-16

(Intercept)                           ***
price                                 ***
years_of_tenure_with_current_provider ***
percent_change_in_price_vs_last_year  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 25977  on 19650  degrees of freedom
Residual deviance: 25413  on 19647  degrees of freedom
AIC: 25421

Number of Fisher Scoring iterations: 6
```

This step will tidy the model results, round the estimates and p-values, select the desired columns, and create a bar plot of the exponentiated coefficients with a horizontal line at y = 1.Through this we get to know that percent_change_in_price_vs_last_year has negative relationship, where as price has no significant effect.

```
broom::tidy(regression_model,
            exponentiate = TRUE, digits=2) %>%
  mutate(estimate=round(estimate,3)) %>%
  mutate(p.value=round(p.value,3))%>%
  select(term,estimate,p.value) %>%
  ggplot(aes(x=term,y=estimate))+
  geom_bar(stat="identity")+
  geom_hline(yintercept=1)
```




```
model_pred <- predict(regression_model, type="response")
```

7 In sample analysis

This is to splits the insurance_data_sample into training and test datasets. It assigns 80% of the rows to the train_data dataframe and the remaining 20% to the test_data dataset. In this code, insurance_data_sample is a dataset containing the insurance data. split_data calculates the row index to split the data based on the 80% threshold. The first 80% of rows are extracted and assigned to train_data, while the remaining rows are assigned to test_data.

```
insurance_data_sample <-insurance_data
split_data <-round(0.8*nrow(insurance_data_sample))
train_data <-insurance_data_sample[1:split_data,]
test_data <- insurance_data_sample[(split_data+1):nrow(insurance_data_sample),]
```

This is a logistic regression model with response variable renewed which is modeled as a function of the predictor variables price, age, and gender. The model will specify with the family argument set to “binomial”, indicating that a binomial distribution with a logit link function will be used for the logistic regression.

```
model <-glm(renewed~price+age+gender,data=train_data,family="binomial")
```

The pred_model is a vector of predicted probabilities for the observations in the test_data. These probabilities represent the model's estimated likelihood of renewal based on the predictor variables price, age, and gender.

```
pred_model<- predict(model, test_data,type="response")
```

In this code, insurance_data_sample represents your training data. The formula renewed ~ price + age + gender specifies the dependent variable (renewed) and the predictor variables (price, age, and gender). method = "cv" specifies that you want to perform cross-validation, and number = 5 indicates that you want to use 5-fold cross-validation. The verboseIter = TRUE argument enables verbose output during the training process.

```
model <-train(renewed~price+age+gender,insurance_data_sample,method="glm", trControl=train
```

```
+ Fold1: parameter=none
- Fold1: parameter=none
+ Fold2: parameter=none
- Fold2: parameter=none
+ Fold3: parameter=none
- Fold3: parameter=none
+ Fold4: parameter=none
- Fold4: parameter=none
+ Fold5: parameter=none
- Fold5: parameter=none
Aggregating results
Fitting final model on full training set
```

In this chunk test_data is the dataset containing the test data. You assign the predicted probabilities from the pred_model to the predicted column in cross_validation. Then, using mutate(), you create the predicted_class column by converting the predicted probabilities to classes. If the predicted probability is greater than 0.5, it is classified as "Yes"; otherwise, it is classified as "No".

```
cross_validation <- test_data
cross_validation$predicted <- pred_model

cross_validation <- cross_validation %>%
  mutate(predicted_class = ifelse(predicted>0.5,"Yes","No")) %>%
  mutate(predicted_class = factor(predicted_class,levels = c("Yes","No"), labels = c("Yes"
```

```r
table(cross_validation$renewed, cross_validation$predicted_class)
```

```
      Yes   No
 No  1605  150
 Yes 2074  101
```

The confusionMatrix() function uses the actual classes (cross_validation*renewed*)*andthepredictedclasses*(*cross*_
as input and returns the confusion matrix along with various performance metrics such as accuracy, sensitivity, specificity.

```r
caret::confusionMatrix(cross_validation$renewed,
                       cross_validation$predicted_class)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  Yes   No
       Yes 2074  101
       No  1605  150

               Accuracy : 0.5659
                 95% CI : (0.5502, 0.5815)
    No Information Rate : 0.9361
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0426

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.56374
            Specificity : 0.59761
         Pos Pred Value : 0.95356
         Neg Pred Value : 0.08547
             Prevalence : 0.93613
         Detection Rate : 0.52774
   Detection Prevalence : 0.55344
      Balanced Accuracy : 0.58067

       'Positive' Class : Yes
```

```r
model_auc<-pROC::auc(cross_validation$renewed,cross_validation$predicted)
```

```
Setting levels: control = No, case = Yes

Setting direction: controls < cases
```

```r
print(model_auc)
```

```
Area under the curve: 0.5584
```

8 Prediction Analysis

```r
model2 <- glm(renewed~price+percent_change_in_price_vs_last_year+annual_mileage+car_value+
```

```r
predicted_data <-predict(model2,test_data,type="response")
```

```r
model2 <-train(renewed~price+percent_change_in_price_vs_last_year+annual_mileage+car_value
```

```
+ Fold1: parameter=none
- Fold1: parameter=none
+ Fold2: parameter=none
- Fold2: parameter=none
+ Fold3: parameter=none
- Fold3: parameter=none
+ Fold4: parameter=none
- Fold4: parameter=none
+ Fold5: parameter=none
- Fold5: parameter=none
Aggregating results
Fitting final model on full training set
```

```r
cross_validation <- test_data
cross_validation$predicted <- predicted_data

cross_validation <- cross_validation %>%
```

```r
  mutate(predicted_class = ifelse(predicted>0.5,"Yes","No")) %>%
  mutate(predicted_class = factor(predicted_class,levels = c("Yes","No"), labels = c("Yes"
```

```r
table(cross_validation$renewed, cross_validation$predicted_class)
```

```
       Yes   No
  No  1291  464
  Yes 1985  190
```

```r
caret::confusionMatrix(cross_validation$renewed,cross_validation$predicted_class)
```

```
Confusion Matrix and Statistics

          Reference
Prediction  Yes   No
       Yes 1985  190
       No  1291  464

               Accuracy : 0.6232
                 95% CI : (0.6078, 0.6383)
    No Information Rate : 0.8336
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1884

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6059
            Specificity : 0.7095
         Pos Pred Value : 0.9126
         Neg Pred Value : 0.2644
             Prevalence : 0.8336
         Detection Rate : 0.5051
   Detection Prevalence : 0.5534
      Balanced Accuracy : 0.6577

       'Positive' Class : Yes
```

```r
model_auc <- pROC::auc(cross_validation$renewed, cross_validation$predicted)
```

```
Setting levels: control = No, case = Yes
```

```
Setting direction: controls < cases
```

```r
print(model_auc)
```

```
Area under the curve: 0.6938
```

9 Evaluation

To measure the performance of the classification problem, two most common approach is used that is ROC, AUC and Confusion matrix. This was used by spilting the data by using 80% for training and 20% for testing. Basically same data set was used to train the dataset and obtain the output from. We utilised the statistical model for predicting the likelihoods on the actual dataset. Train models were used to predict the renewal variable using the generalised linear model based on the price, percent Change in Price vs last year, age and genders.The logistic regressiom model accurately analysed that the age of customer, car value, mileage, insurance price are the main factors affecting the insurance.

The ROC (Receiver operating characrteristics) and the AUC (Area under curve) was used for true positive percentage and false positive percentage.AUC provides a cumulative measure of performance across all classification levels.

The confusion matrix from the training dataset represents the proportion of correct predictions out of the total predictions made by the model. In this case, the accuracy is calculated as 0.5659 or 56.59% , while the acuracy with the prediction model is 0.6059 or 60.59%.

AUC can be interpreted as the likelihood that the model ranks a random positive case higher than a random negative example.An ROC curve is a graph showing the performance of a classification model at all classification thresholds. It indicates the model's ability to distinguish between positive and negative classes based on anticipated probabilities. With an AUC of 0.5584, the model's ability to discriminate between positive and negative classes is relatively low. The model's predictions are only marginally better than chance. It implies that the model's ability to differentiate between the two classes is restricted. Medium level of distinction is made from the value of 0.6938.

10 Conclusion

Insurance Renewal rate can be increased by if the change of price from every year is not much increased. This can be observed from the regression model.

From the Visualisation part we understand that the price has little effect on renewal rates. As prices rise, fewer customers renew their insurance; however, we see that married females who purchase insurance at a higher rate have a higher chance of renewal than non-married males.

Percentage change has a greater impact on renewal rates than price change. We can observe from the correlation tests that price has no influence on the renewal rate. The most beneficial variables in terms of renewal rate is the percentage change in price vs last year. According to the logistic regression analysis and covariance test, it has a negative impact, which means that if the percentage change is greater, the chances of an individual renewing the insurance are lower.

Characteristics such as gender and new marital status have a moderate effect. We can also see from the logistic regression that more tenure is more likely to enhance the renewal rate.