

Modelling Car Insurance Claims

Jude Nogotey

2024-03-21



Data Description:

The dataset selected for this project is the Car Insurance Claim dataset sourced from Kaggle (<https://www.kaggle.com/code/kerneler/starter-car-insurance-claim-data-62f4f91c-d/input>). This dataset comprises 10,302 observations of 27 variables, each record representing a set of attributes of an insurance company's individual customer related to their socio-demographic profile and the insured vehicle. The binary response variable CLAIM_FLAG indicates whether the customer's car was involved in a crash (1) or not (0), while the continuous response variable CLM_AMT defines the cost related to the car crash if it occurred. An alternative way to define the binary response CLAIM_FLAG is whether the CLM_AMT(claim amount) given to the car owner by the insurance company was greater than \$0 (1) or not (0)

```
#Loading Libraries  
library(tidyverse)  
library(tidymodels)
```

```
library(dplyr)
library(skimr)
library(GGally)
library(DMwR2)
library(caret)
library(pROC)

#Checking for missing values among the variables
skim(car_claim)
```

Data summary

```
Name          car_claim
Number of rows 10302
Number of columns 27
```

Column type frequency:

```
character      16
numeric        11
```

```
Group variables      None
```

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
BIRTH	0	1	7	7	0	6560	0
INCOME	0	1	0	8	570	8152	0
PARENT1	0	1	2	3	0	2	0
HOME_VAL	0	1	0	8	575	6335	0
MSTATUS	0	1	3	4	0	2	0
GENDER	0	1	1	3	0	2	0
EDUCATION	0	1	3	13	0	5	0
OCCUPATION	0	1	0	13	665	9	0
CAR_USE	0	1	7	10	0	2	0
BLUEBOOK	0	1	6	7	0	2985	0
CAR_TYPE	0	1	3	11	0	6	0
RED_CAR	0	1	2	3	0	2	0
OLDCLAIM	0	1	2	7	0	3545	0
REVOKED	0	1	2	3	0	2	0
CLM_AMT	0	1	2	8	0	2346	0
URBANICITY	0	1	19	21	0	2	0

Variable type: numeric

skim_v variable	n_mi missing	complete rate	mean	sd	p0	p25	p50	p75	p100	hist
ID	0	1.00	495663 109.08	286467 479.03	63 17 5	2442 8685 6	4970 0429 3	7394 5506 9	9999 2636 8	
KIDSD RIV	0	1.00	0.17	0.51	0	0	0	0	4	
AGE	7	1.00	44.84	8.61	16	39	45	51	81	
HOME KIDS	0	1.00	0.72	1.12	0	0	0	1	5	
YOJ	548	0.95	10.47	4.11	0	9	11	13	23	
TRAVT IME	0	1.00	33.42	15.87	5	22	33	44	142	
TIF	0	1.00	5.33	4.11	1	1	4	7	25	
CLM_F REQ	0	1.00	0.80	1.15	0	0	0	2	5	
MVR_P TS	0	1.00	1.71	2.16	0	0	1	3	13	
CAR_A GE	639	0.94	8.30	5.71	-3	1	8	12	28	
CLAIM _FLAG	0	1.00	0.27	0.44	0	0	0	1	1	

1. Probability as a foundation of statistical modeling

Probability theory serves as the cornerstone of statistical modeling, providing a systematic framework for quantifying uncertainty and making probabilistic statements about random events. In the context of logistic regression, probability theory underpins the model's fundamental concepts. The logistic regression model aims to estimate the probability of a binary outcome, such as the occurrence of a car insurance claim, based on predictor variables. The odds ratio, denoted as $p_i / (1 - p_i)$, where p represents the probability of success (e.g., filing a claim), forms the basis of logistic regression. The logit function, $\log(p_i / (1 - p_i))$, transforms the odds ratio into a linear function of predictor variables. This

transformation allows logistic regression to model the log odds of the outcome as a linear combination of predictor variables.

In logistic regression, maximum likelihood estimation (MLE) is employed to estimate the model parameters. The likelihood function represents the probability of observing the given data under the assumed logistic regression model. The goal of MLE is to find the parameter values that maximize this likelihood function, effectively maximizing the probability of observing the observed data given the model. Mathematically, this involves taking the derivative of the likelihood function with respect to the parameters and setting it equal to zero to find the maximum likelihood estimates. Thus, probability theory, coupled with maximum likelihood estimation, forms the theoretical foundation of logistic regression modeling, allowing us to infer relationships between predictor variables and the probability of the binary outcome.

```
# Checking for datatypes of variables
```

```
glimpse(car_claim)
```

```
## Rows: 10,302
```

```
## Columns: 27
```

```
## $ ID          <int> 63581743, 132761049, 921317019, 727598473, 450221861, 74314...
```

```
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
```

```
## $ BIRTH       <chr> "16MAR39", "21JAN56", "18NOV51", "05MAR64", "05JUN48", "17M...
```

```
## $ AGE         <int> 60, 43, 48, 35, 51, 50, 34, 54, 40, 44, 37, 34, 50, 53, 43, ...
```

```
## $ HOMEKIDS    <int> 0, 0, 0, 1, 0, 0, 1, 0, 1, 2, 2, 0, 0, 0, 0, 0, 0, 2, 0, 3, ...
```

```
## $ YOJ         <int> 11, 11, 11, 10, 14, NA, 12, NA, 11, 12, NA, 10, 7, 14, 5, 1...
```

```
## $ INCOME      <chr> "$67,349", "$91,449", "$52,881", "$16,039", "", "$114,986", ...
```

```
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "Yes", "Ye...
```

```
## $ HOME_VAL    <chr> "$0", "$257,252", "$0", "$124,191", "$306,251", "$243,925", ...
```

```
## $ MSTATUS     <chr> "z_No", "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes", ...
```

```
## $ GENDER      <chr> "M", "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z_F", ...
```

```
## $ EDUCATION   <chr> "PhD", "z_High School", "Bachelors", "z_High School", "<Hig...
```

```
## $ OCCUPATION  <chr> "Professional", "z_Blue Collar", "Manager", "Clerical", "z_...
```

```
## $ TRAVTIME    <int> 14, 22, 26, 5, 32, 36, 46, 33, 21, 30, 44, 34, 48, 15, 36, ...
```

```
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Private", "...
```

```
## $ BLUEBOOK    <chr> "$14,230", "$14,940", "$21,970", "$4,010", "$15,440",
```

```

"$18,...
## $ TIF          <int> 11, 1, 1, 4, 7, 1, 1, 1, 6, 10, 1, 1, 7, 1, 7, 7, 6, 6,
1, ...
## $ CAR_TYPE     <chr> "Minivan", "Minivan", "Van", "z_SUV", "Minivan",
"z_SUV", "...
## $ RED_CAR      <chr> "yes", "yes", "yes", "no", "yes", "no", "no", "no",
"no", "...
## $ OLDCLAIM     <chr> "$4,461", "$0", "$0", "$38,690", "$0", "$19,217", "$0",
"$0...
## $ CLM_FREQ     <int> 2, 0, 0, 2, 0, 2, 0, 0, 1, 0, 1, 0, 0, 0, 0, 2, 0, 0,
0, 0,...
## $ REVOKED      <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No",
"No"...
## $ MVR_PTS      <int> 3, 0, 2, 3, 0, 3, 0, 0, 2, 0, 10, 0, 1, 0, 0, 3, 3, 0,
3, 0...
## $ CLM_AMT      <chr> "$0", "$0", "$0", "$0", "$0", "$0", "$2,946", "$0",
"$6,477...
## $ CAR_AGE      <int> 18, 1, 10, 10, 6, 17, 7, 1, 1, 10, 7, 1, 17, 11, 1, 9,
10, ...
## $ CLAIM_FLAG   <int> 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0,
1, 0,...
## $ URBANICITY   <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Highly
Urban...

```

2. How Logistic regression model was chosen for this insurance claim data

Binary data for a generalized linear model can come in one of two forms. One form will have each observation (row of data) as a 1 or 0 (success or failure). The other form will be having the number of successes out of the total. The choice of logistic regression as the appropriate GLM is informed by the nature of the response variable and the desired outcome prediction task. Logistic regression allows us to model the probability of a binary event, making it well-suited for predicting whether a car insurance claim will be filed based on relevant predictor variables.

The variable of interest thus in car insurance dataset is 'CLAIM_FLG' which is whether an insured filed a claim or not. This variable has two levels: 1 for claim filed and 0 for no claim. Hence it will be appropriate to utilize logistic regression whether an insured will file a claim or not.

Generally, we will let the probability of a "success" ($Y_i = 1$) be p_i and the probability of a "failure" ($Y_i = 0$) be $1 - p_i$. Therefore, the odds of a "success" are:

$$\frac{Pr(Y_i = 1)}{Pr(Y_i = 0)} = \frac{p_i}{1 - p_i}$$

We use the *logit function* (or *log odds*) to model binary outcome variables:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X$$

For positive values of β_1 , the S-shaped logistic curve will be monotonic increasing while negative values of β_1 will be monotonic decreasing. The absolute value of β_1 , will impact the steepness of the curve. In this logistic regression we will be considering several predictors which will be in the form:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

where p represents the probability of filing a claim, X_1, X_2, \dots, X_n are the predictor variables, and $\beta_1, \beta_2, \dots, \beta_n$ are the co-efficients to be estimated. By determining and applying the logistic regression model, we can effectively model and analyze the relationship between predictor variables and the likelihood of the binary outcome occurring thus an insured filing a claim or not.

```
# Converting some variables into their proper data type
car_claim$CLM_AMT <- as.numeric(gsub("[^0-9.]", "", car_claim$CLM_AMT))
#Removes leading dollar sign before converting to numeric to avoid NAs
car_claim$OLDCLAIM <- as.numeric(gsub("[^0-9.]", "", car_claim$OLDCLAIM))
car_claim$BLUEBOOK <- as.numeric(gsub("[^0-9.]", "", car_claim$BLUEBOOK))
car_claim$HOME_VAL <- as.numeric(gsub("[^0-9.]", "", car_claim$HOME_VAL))
car_claim$INCOME <- as.numeric(gsub("[^0-9.]", "", car_claim$INCOME))
car_claim$CLAIM_FLAG <- as.factor(car_claim$CLAIM_FLAG)

skim(car_claim)
```

Data summary

Name	car_claim
Number of rows	10302
Number of columns	27

Column type frequency:

character	11
factor	1
numeric	15

Group variables	None
-----------------	------

Variable type: character










skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
BIRTH	0	1	7	7	0	6560	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
PARENT1	0	1	2	3	0	2	0
MSTATUS	0	1	3	4	0	2	0
GENDER	0	1	1	3	0	2	0
EDUCATION	0	1	3	13	0	5	0
OCCUPATION	0	1	0	13	665	9	0
CAR_USE	0	1	7	10	0	2	0
CAR_TYPE	0	1	3	11	0	6	0
RED_CAR	0	1	2	3	0	2	0
REVOKED	0	1	2	3	0	2	0
URBANICITY	0	1	19	21	0	2	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
CLAIM_FLAG	0	1	FALSE	2	0: 7556, 1: 2746

Variable type: numeric

skim_v variable	n_mi ssing	comple te_rate	mean	sd	p0	p25	p50	p75	p100	hist
ID	0	1.00	495663 109.08	286467 479.03	63 17 5	2442 8685 6	4970 0429 3	739455 069.00	9999 2636 8	
KIDSD RIV	0	1.00	0.17	0.51	0	0	0	0.00	4	
AGE	7	1.00	44.84	8.61	16	39	45	51.00	81	
HOME KIDS	0	1.00	0.72	1.12	0	0	0	1.00	5	
YOJ	548	0.95	10.47	4.11	0	9	11	13.00	23	
INCOM E	570	0.94	61572. 07	47457. 20	0	2758 4	5352 9	86166. 00	3670 30	
HOME_ VAL	575	0.94	154523 .02	129188 .44	0	0	1606 61	238256 .00	8852 82	
TRAVT IME	0	1.00	33.42	15.87	5	22	33	44.00	142	
BLUEB	0	1.00	15659.	8428.7	15	9200	1440	20890.	6974	

skim_v variable	n_mi ssing	comple te_rate	mean	sd	p0	p25	p50	p75	p100	hist
OOK			92	7	00		0	00	0	--
TIF	0	1.00	5.33	4.11	1	1	4	7.00	25	█--
										--
OLDCL	0	1.00	4033.9	8733.1	0	0	0	4647.5	5703	█--
AIM			8	4				0	7	--
										--
CLM_F	0	1.00	0.80	1.15	0	0	0	2.00	5	█--
REQ										--
MVR_P	0	1.00	1.71	2.16	0	0	1	3.00	13	█--
TS										--
CLM_A	0	1.00	1511.2	4725.2	0	0	0	1144.7	1232	█--
MT			7	5				5	47	--
										--
CAR_A	639	0.94	8.30	5.71	-3	1	8	12.00	28	███
GE										--

Based on Michigan Auto Law: <https://www.michiganautolaw.com/blog/2023/09/19/22-factors-that-affect-car-insurance-rates/> auto insurance companies are looked at across the board, these are the top 22 rating factors that affect car insurance rates that insurers rely on: Age, Gender, Location where a person lives, Marital status, Credit score, Profession, Driving history, Years of driving experience, Annual mileage, Use of vehicle, The actual vehicle, Safety rating of your vehicle, Size of your vehicle, Age of your vehicle, Likelihood of theft of your vehicle, Vehicle ownership status, Previous insurance coverage, Claims record, Insurance coverage levels, Insurance deductible levels, Discount options, and Who is your insurance company?

The insurance data used in this project does not contain all these variables above hence we will consider variables synonymous to the factors list above. These variables are: AGE, GENDER, URBANICITY, MSTATUS, OCCUPATION, CAR_USE, CAR_AGE, CLM_FREQ, CAR_TYPE, INCOME Removing redundant variables

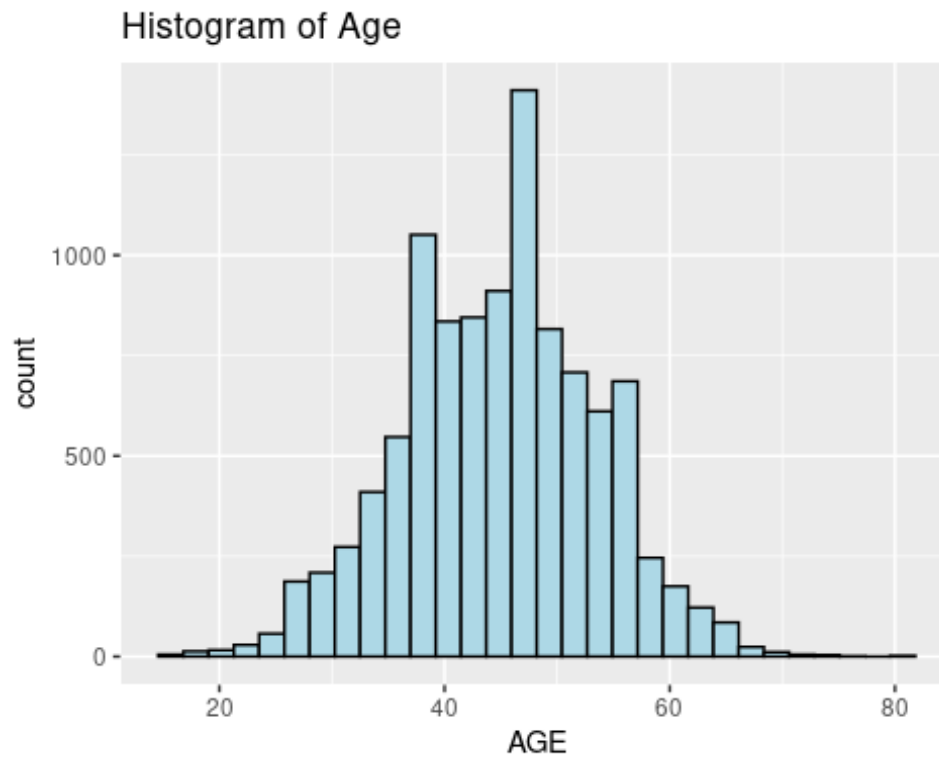
Selecting variables of interest based on information above for exploratory data analysis

```
car_claim1<-car_claim %>% select(c('AGE', 'GENDER', 'URBANICITY', 'MSTATUS',  
'OCCUPATION', 'CAR_USE', 'CAR_AGE', 'CLM_FREQ', 'CAR_TYPE',  
'INCOME', 'CLAIM_FLAG'))
```

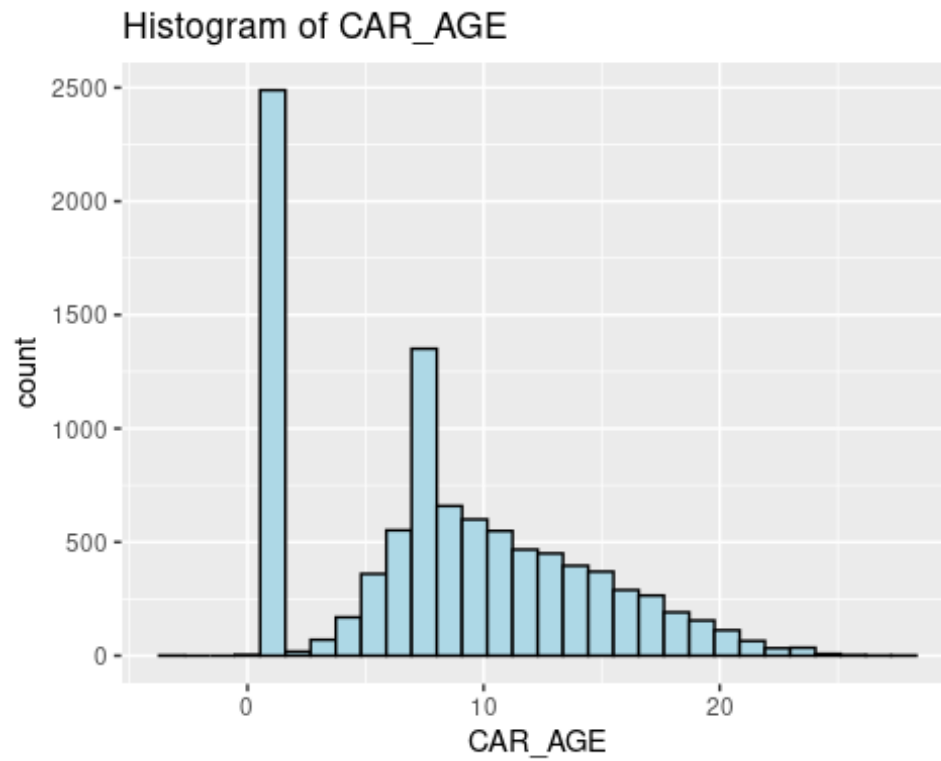
Exploratory Data Analysis

Examining and visualizing data to understand its structure, identify patterns and spot anomalies of variables of interest in the dataset.

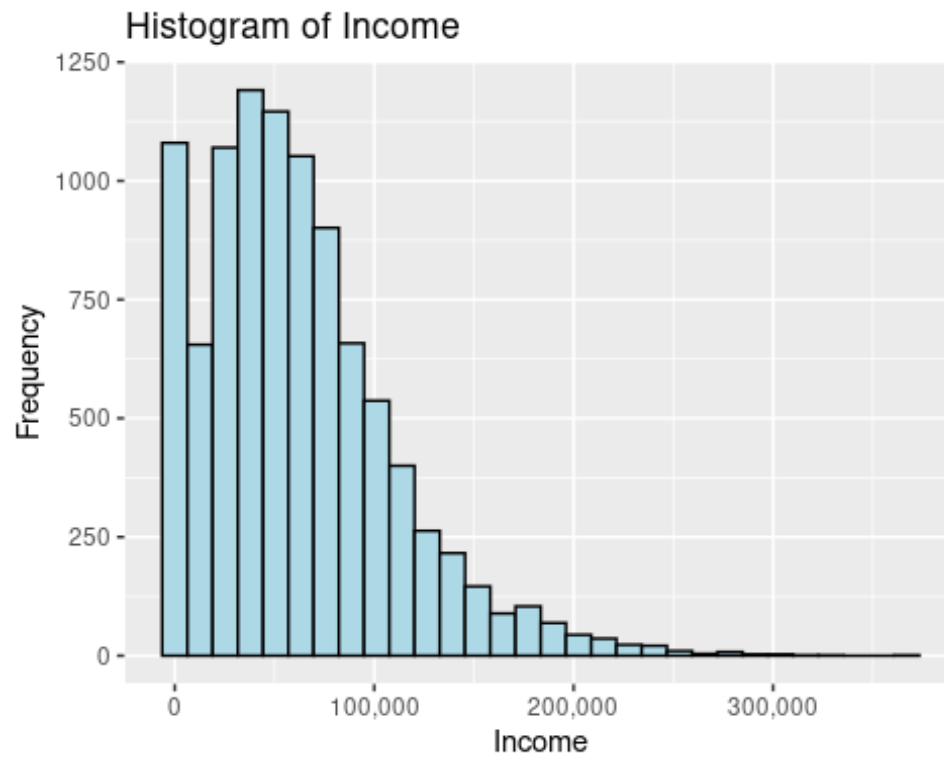

```
# Histogram of Age
car_claim1 %>%
  ggplot(aes(x=AGE))+
  geom_histogram(fill = "lightblue", color = "black") +
  ggtitle("Histogram of Age")
```



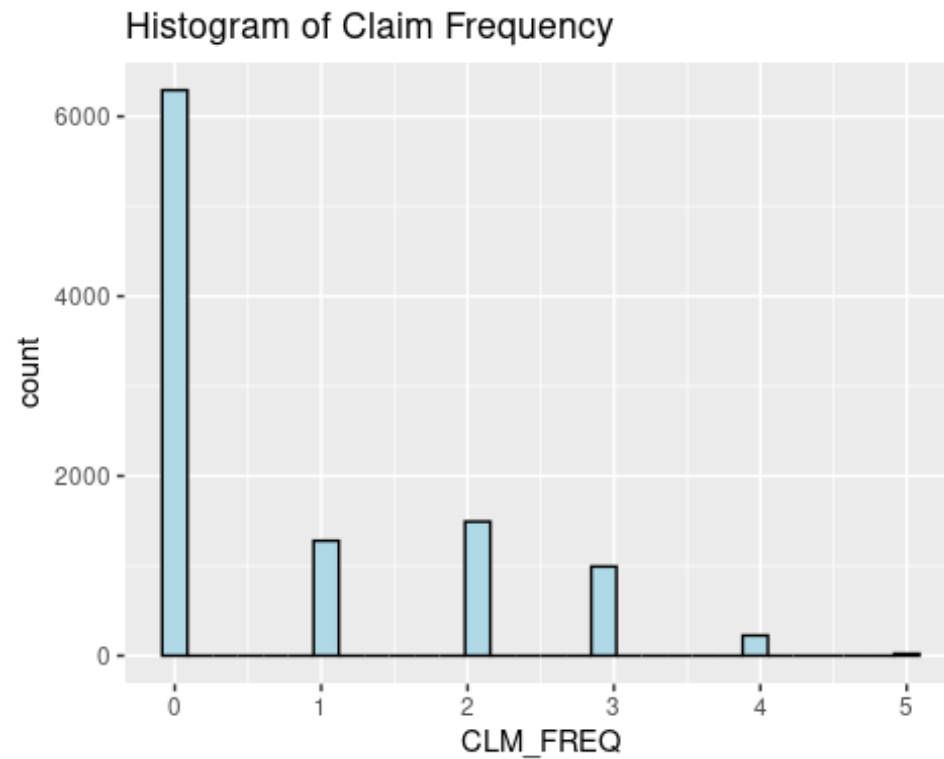
```
# Histogram of Car Age
car_claim1 %>%
  ggplot(aes(x=CAR_AGE))+
  geom_histogram(fill = "lightblue", color = "black") +
  ggtitle("Histogram of CAR_AGE")
```



```
# Histogram of Income
car_claim1 %>%
  ggplot(aes(x = INCOME)) +
  geom_histogram(fill = "lightblue", color = "black") +
  scale_x_continuous(labels = scales::comma) +
  labs(title = "Histogram of Income", x = "Income", y = "Frequency")
```



```
#histogram of Claim Frequency  
car_claim1 %>%  
  ggplot(aes(x=CLM_FREQ))+  
  geom_histogram(fill = "lightblue", color = "black") +  
  ggtitle("Histogram of Claim Frequency")
```



```
skim(car_claim1)
```

Data summary

Name car_claim1
 Number of rows 10302
 Number of columns 11

Column type frequency:

character 6
 factor 1
 numeric 4

Group variables None

Variable type: character





skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
GENDER	0	1	1	3	0	2	0
URBANICITY	0	1	19	21	0	2	0
MSTATUS	0	1	3	4	0	2	0
OCCUPATION	0	1	0	13	665	9	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
CAR_USE	0	1	7	10	0	2	0
CAR_TYPE	0	1	3	11	0	6	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
CLAIM_FLAG	0	1	FALSE	2	0: 7556, 1: 2746

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
AGE	7	1.00	44.84	8.61	16	39	45	51	81	
CAR_AGE	639	0.94	8.30	5.71	3	1	8	12	28	
CLM_FREQ	0	1.00	0.80	1.15	0	0	0	2	5	
INCOME	570	0.94	61572.07	47457.20	0	27584	53529	86166	367030	

Handling missing values using KNN imputation

K-Nearest Neighbors (KNN) imputation is a method used to fill in missing values in a dataset based on the values of its nearest neighbors. It can be applied to both numeric and categorical variables, making it quite versatile.

```
#Identifying columns with missing values
missn_cols <- colnames(car_claim1)[apply(car_claim1, 2, function(x)
any(is.na(x)))]

#Replace missing values in the numerical columns
num_cols <- sapply(car_claim1, is.numeric)
car_claim1[num_cols] <- knnImputation(car_claim1[num_cols], k =5)

#Replace missing values in the character columns with unknown
cat_col <- sapply(car_claim1, is.character)
car_claim1[cat_col] <- lapply(car_claim1[cat_col], function(x)
ifelse(is.na(x), "unknown",x))

# Distribution of Claim_flag Total
claims_total <- car_claim1 %>% mutate(CLAIM_FLAG = case_when(
  CLAIM_FLAG == 0 ~ "No",
  CLAIM_FLAG == 1 ~ "Yes"
)) %>%
```

```

count(CLAIM_FLAG) %>%
mutate(percent = round(n / sum(n) * 100, 2)) #>%
#knitr::kable()
claims_total

## CLAIM_FLAG    n percent
## 1           No 7556   73.34
## 2           Yes 2746   26.66

# Distribution of Claim_flag in Urban Areas
claims_urban <- car_claim1 %>% mutate(CLAIM_FLAG = case_when(
  CLAIM_FLAG == 0 ~ "No",
  CLAIM_FLAG == 1 ~ "Yes"
)) %>% filter(URBANICITY == 'Highly Urban/ Urban') %>%
count(CLAIM_FLAG) %>%
mutate(percent = round(n / sum(n) * 100, 2)) #>%
#knitr::kable()
claims_urban

## CLAIM_FLAG    n percent
## 1           No 5618   68.26
## 2           Yes 2612   31.74

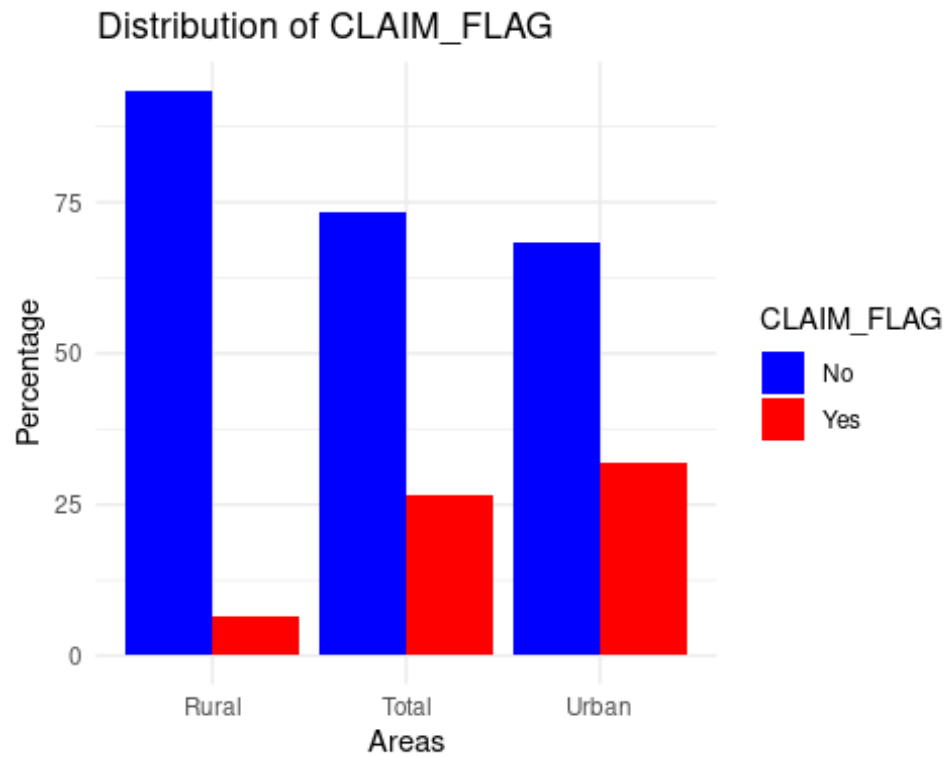
# Distribution of Claim_flag in Rural Areas
claims_rural <- car_claim1 %>% mutate(CLAIM_FLAG = case_when(
  CLAIM_FLAG == 0 ~ "No",
  CLAIM_FLAG == 1 ~ "Yes"
)) %>% filter(URBANICITY == 'z_Highly Rural/ Rural') %>%
count(CLAIM_FLAG) %>%
mutate(percent = round(n / sum(n) * 100, 2))
claims_rural

## CLAIM_FLAG    n percent
## 1           No 1938   93.53
## 2           Yes  134    6.47

# Combine the dataframes
combined_data <- bind_rows(
  mutate(claims_total, area = "Total"),
  mutate(claims_urban, area = "Urban"),
  mutate(claims_rural, area = "Rural")
)

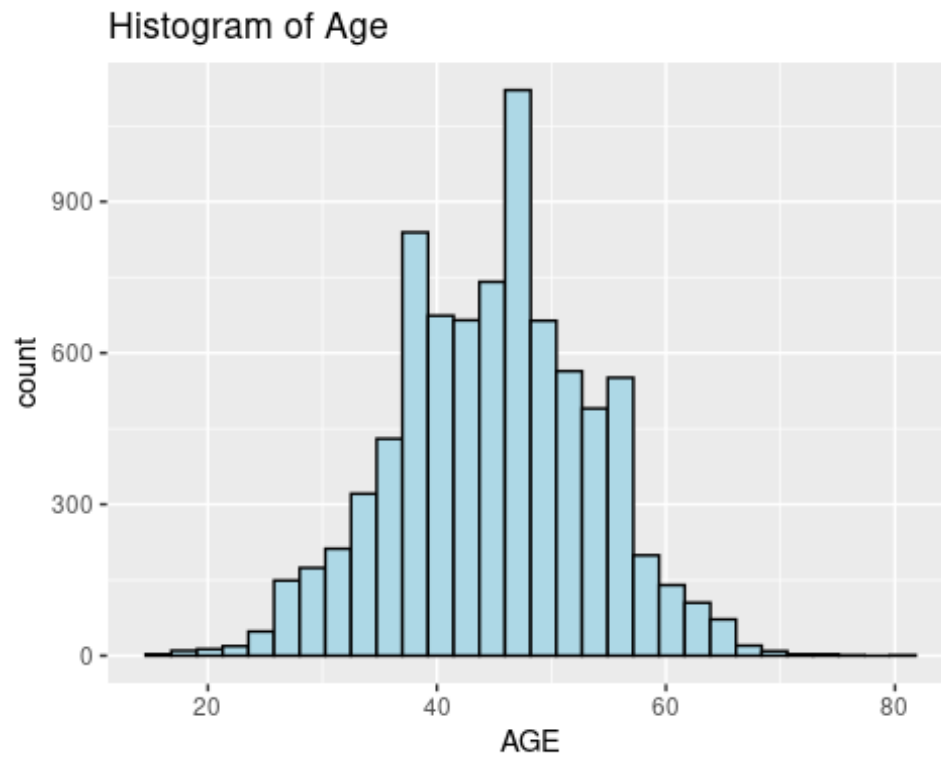
# Plotting using ggplot
ggplot(combined_data, aes(x = area, y = percent, fill = CLAIM_FLAG)) +
  geom_col(position = "dodge") +
  labs(title = "Distribution of CLAIM_FLAG",
       x = "Areas",
       y = "Percentage") +
  scale_fill_manual(values = c("No" = "blue", "Yes" = "red")) +
  theme_minimal()

```

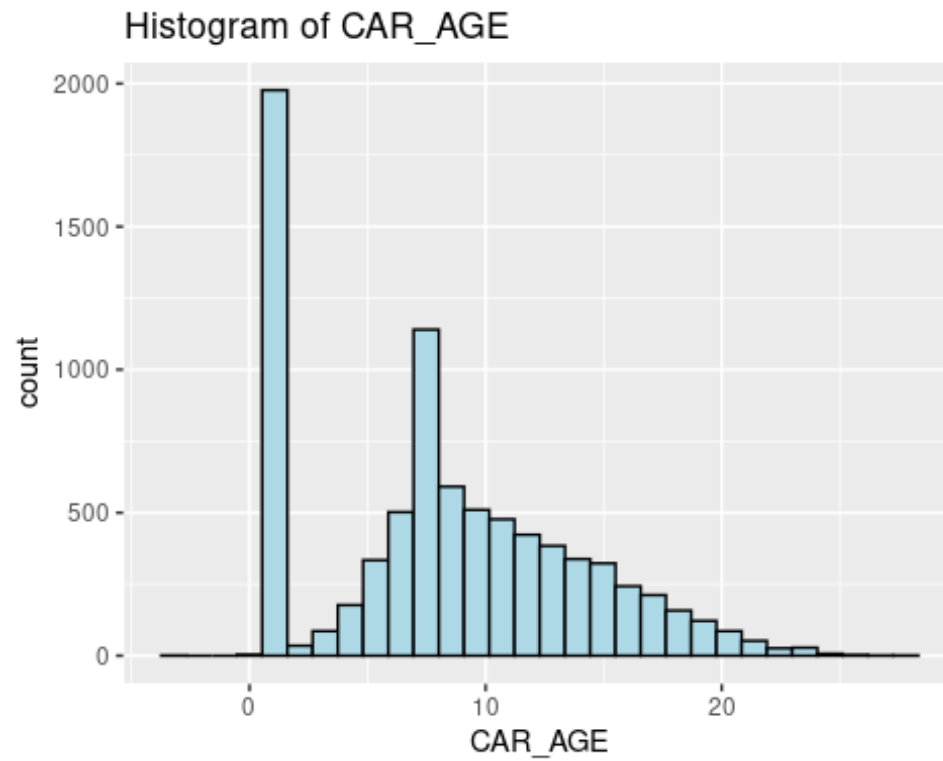


```
# Splitting data into train and test sets
set.seed(123)
claim_split <- initial_split(car_claim1, prop = 0.80)
train_claim <- training(claim_split)
test_claim <- testing(claim_split)

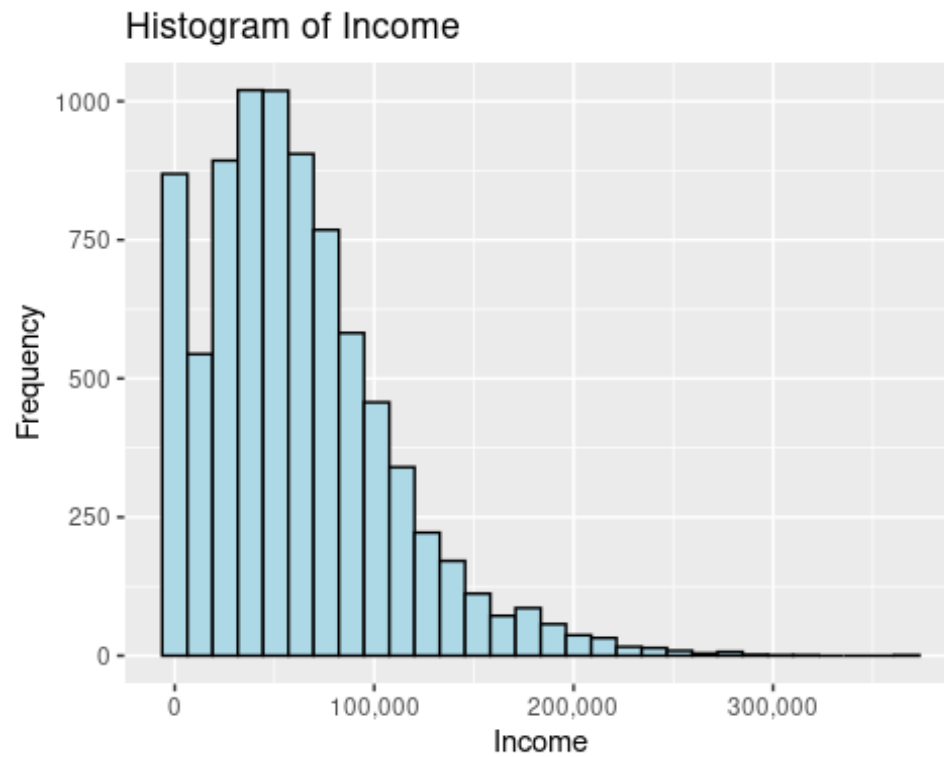
# Histogram of Age
train_claim %>%
  ggplot(aes(x=AGE))+
  geom_histogram(fill = "lightblue", color = "black") +
  ggtitle("Histogram of Age")
```



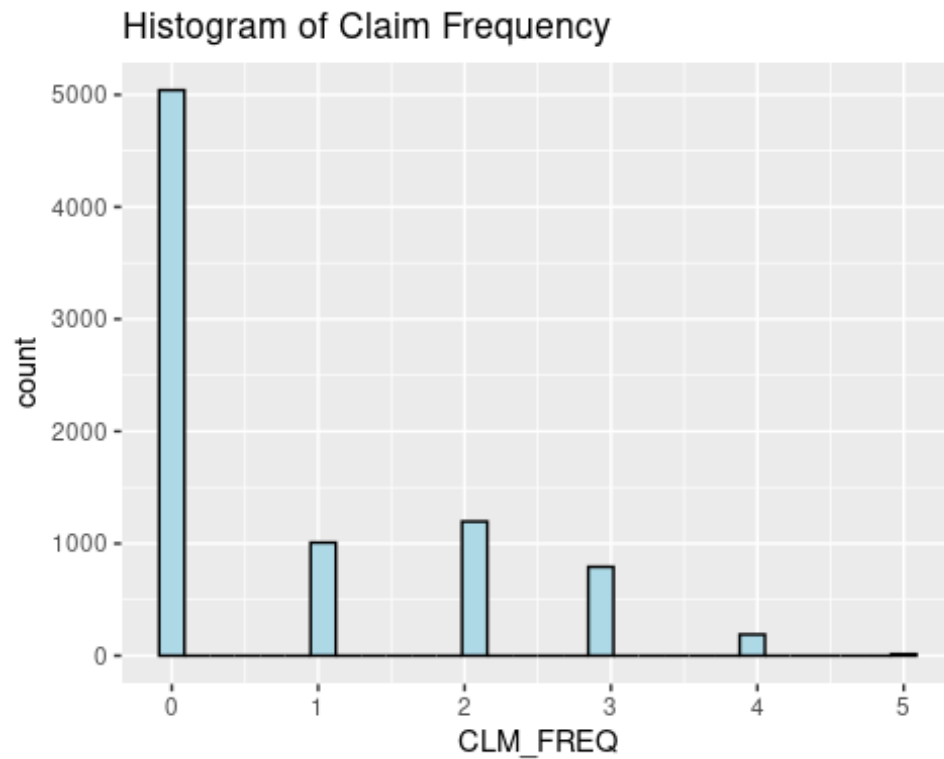
```
# Histogram of Car Age
train_claim %>%
  ggplot(aes(x=CAR_AGE))+
  geom_histogram(fill = "lightblue", color = "black") +
  ggtitle("Histogram of CAR_AGE")
```

```
# Histogram of Income
train_claim %>%
  ggplot(aes(x = INCOME)) +
  geom_histogram(fill = "lightblue", color = "black") +
  scale_x_continuous(labels = scales::comma) +
  labs(title = "Histogram of Income", x = "Income", y = "Frequency")
```

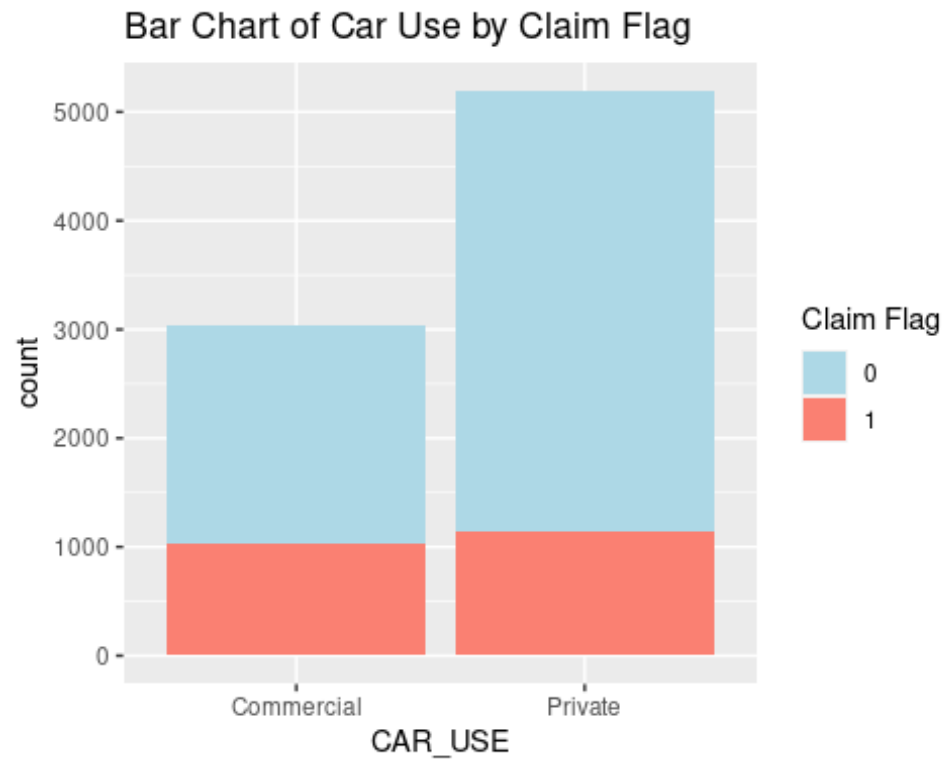


```
#histogram of Claim Frequency  
train_claim %>%  
  ggplot(aes(x=CLM_FREQ))+  
  geom_histogram(fill = "lightblue", color = "black") +  
  ggtitle("Histogram of Claim Frequency")
```



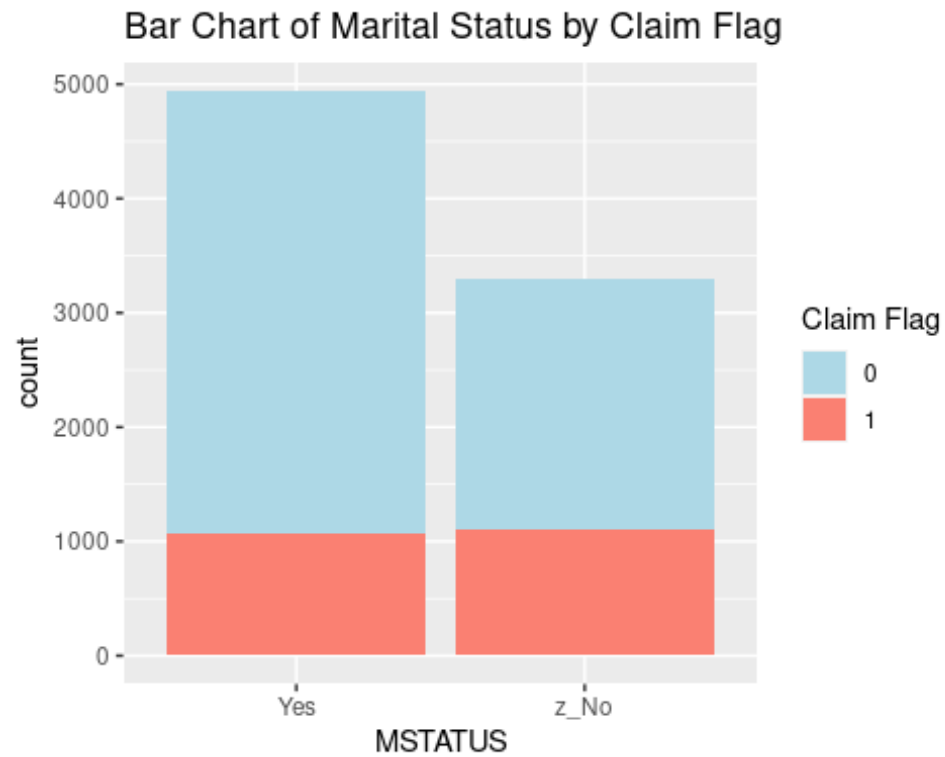
#Bar Chart of Car Use

```
train_claim %>%  
  ggplot(aes(x = CAR_USE, fill = CLAIM_FLAG)) +  
  geom_bar() +  
  ggtitle("Bar Chart of Car Use by Claim Flag") +  
  scale_fill_manual(values = c('0' = "lightblue", '1' = "salmon")) +  
  labs(fill = "Claim Flag")
```



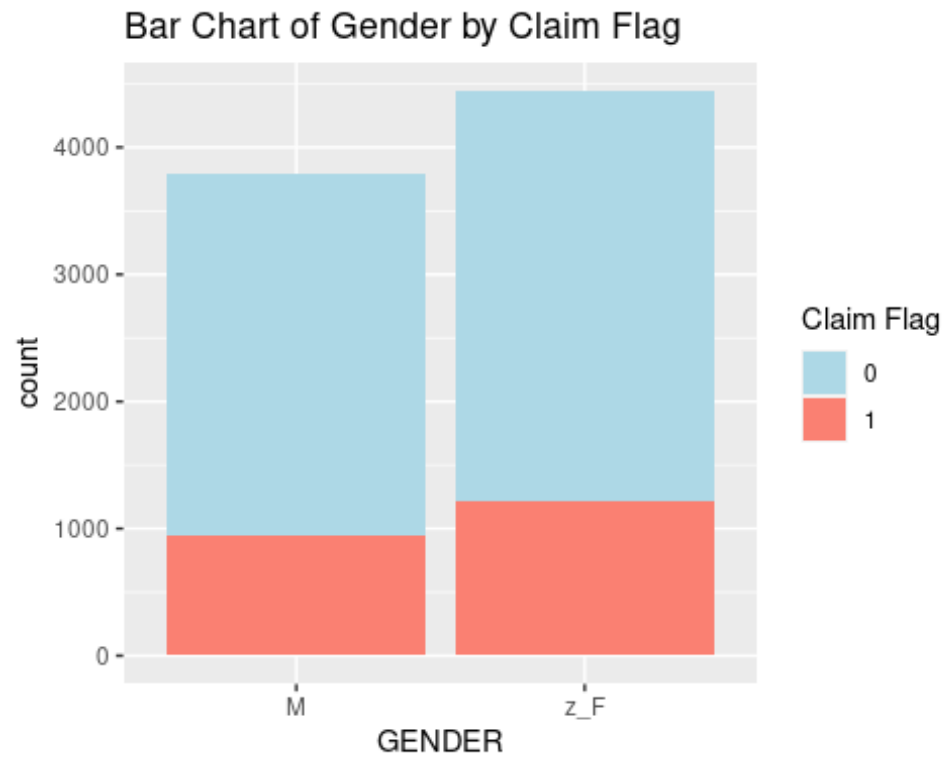
#Bar Chart of Marital Status

```
train_claim %>%  
  ggplot(aes(x = MSTATUS, fill = CLAIM_FLAG)) +  
  geom_bar() +  
  ggtitle("Bar Chart of Marital Status by Claim Flag") +  
  scale_fill_manual(values = c('0' = "lightblue", '1' = "salmon")) +  
  labs(fill = "Claim Flag")
```



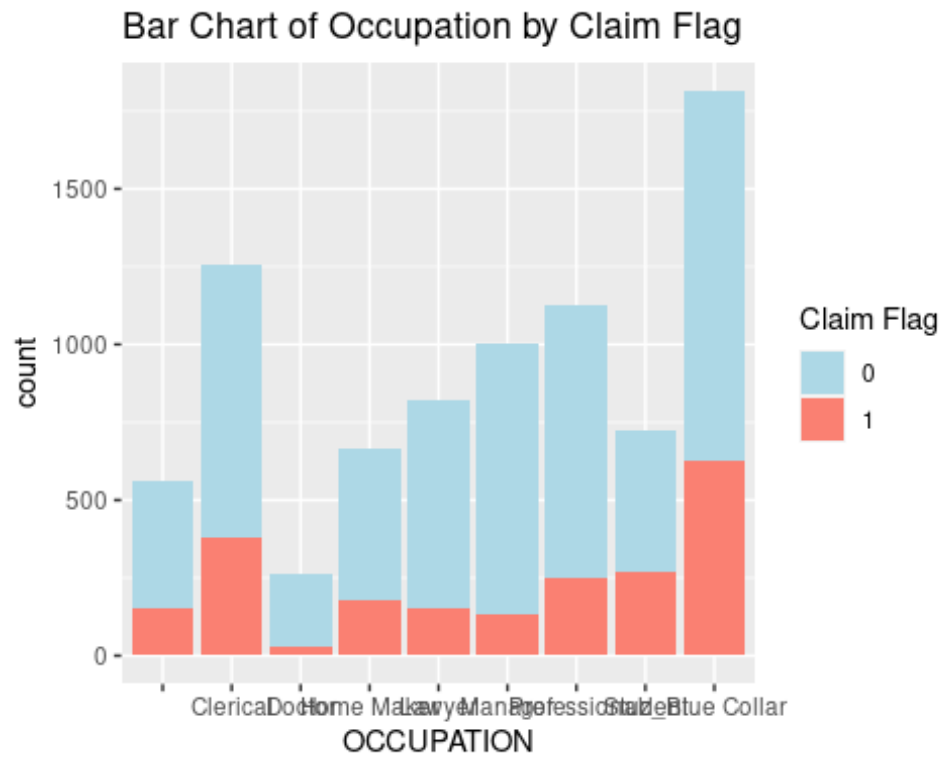
#Bar Chart of Gender

```
train_claim %>%  
  ggplot(aes(x = GENDER, fill = CLAIM_FLAG)) +  
  geom_bar() +  
  ggtitle("Bar Chart of Gender by Claim Flag") +  
  scale_fill_manual(values = c('0' = "lightblue", '1' = "salmon")) +  
  labs(fill = "Claim Flag")
```



#Bar Chart of Occupation

```
train_claim %>%  
  ggplot(aes(x = OCCUPATION, fill = CLAIM_FLAG)) +  
  geom_bar() +  
  ggtitle("Bar Chart of Occupation by Claim Flag") +  
  scale_fill_manual(values = c('0' = "lightblue", '1' = "salmon")) +  
  labs(fill = "Claim Flag")
```



#Bar Chart of Urban and Rural City

`train_claim %>%`

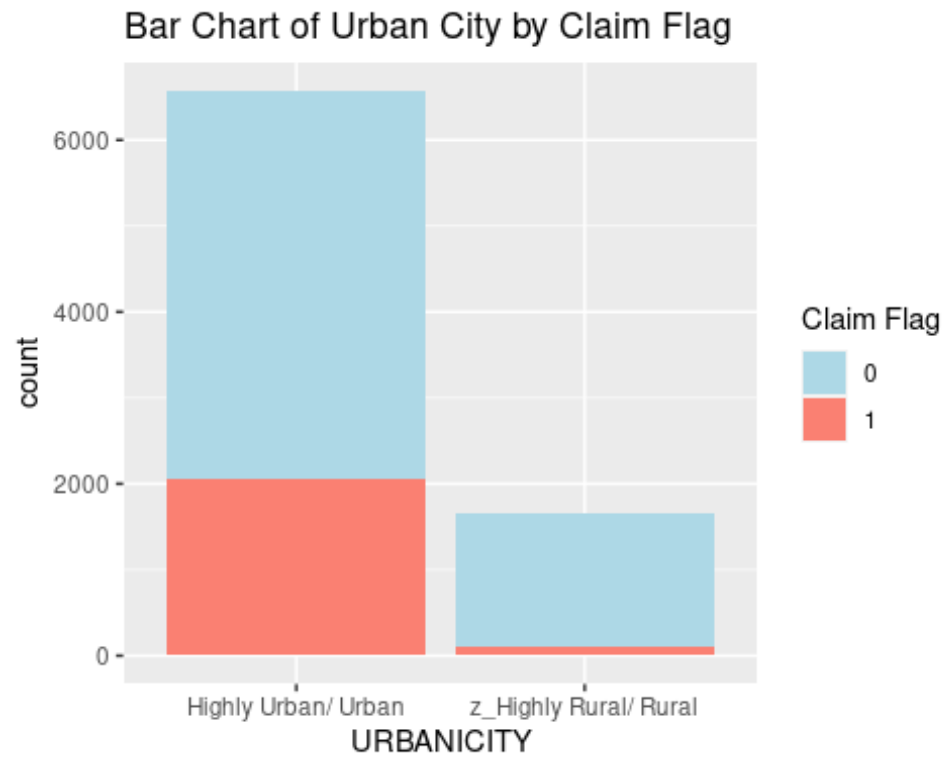
`ggplot(aes(x = URBANICITY, fill = CLAIM_FLAG)) +`

`geom_bar() +`

`ggtitle("Bar Chart of Urban City by Claim Flag") +`

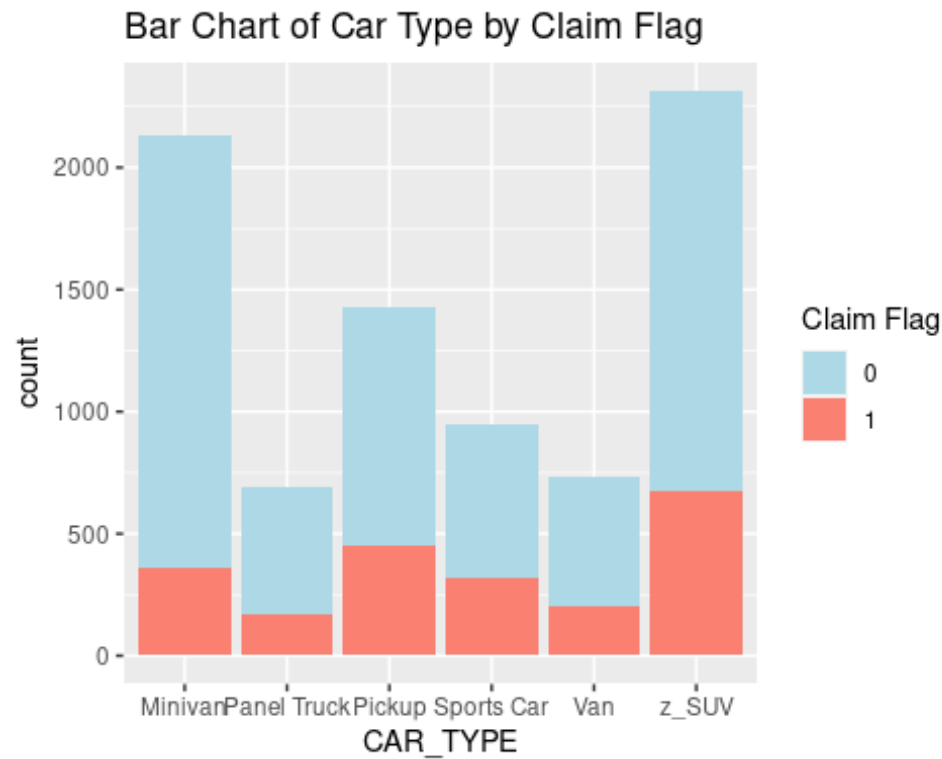
`scale_fill_manual(values = c('0' = "lightblue", '1' = "salmon")) +`

`labs(fill = "Claim Flag")`

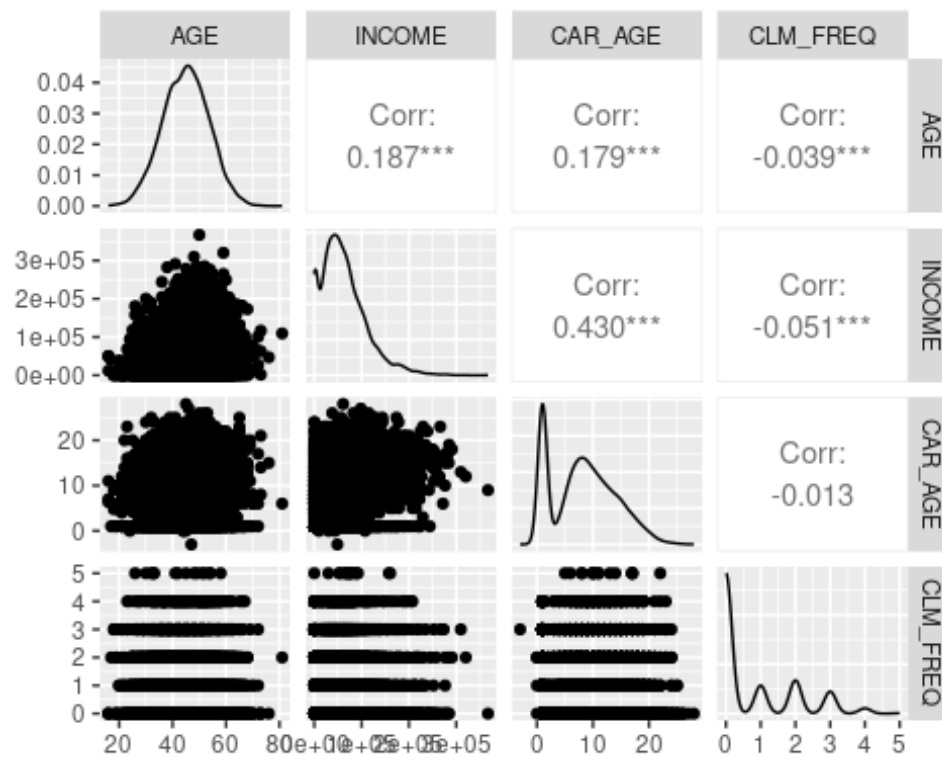


#Bar Chart of Car Type

```
train_claim %>%  
  ggplot(aes(x = CAR_TYPE, fill = CLAIM_FLAG)) +  
  geom_bar() +  
  ggtitle("Bar Chart of Car Type by Claim Flag") +  
  scale_fill_manual(values = c('0' = "lightblue", '1' = "salmon")) +  
  labs(fill = "Claim Flag")
```

```
train_claim %>%
  select(AGE, INCOME, CAR_AGE, CLM_FREQ) %>%
  ggpairs()
```



From the correlation values above, there is no multicollinearity exhibited by the numeric variables of interest.

3. Model selection for a set of candidate models and communicating the results of statistical models to a general audience (Course objectives 3 and 4 merged)

Model selection is a critical step in statistical modeling, aimed at identifying the most appropriate model from a set of candidate models. This process is essential for ensuring the chosen model adequately captures the underlying relationships in the data while avoiding overfitting. In the context of logistic regression modeling for car insurance claims, we employ model selection techniques to compare different logistic regression models with varying sets of predictor variables.

Moreover, effectively communicating the results of statistical models to a general audience is paramount for ensuring comprehension and facilitating informed decision-making. To achieve this, it is essential to present the findings in a clear, concise, and understandable manner, using a combination of text, tables, and figures. Visualizations such as ROC curves, confusion matrices, and predicted vs. observed plots provide intuitive representations of model performance and are accessible to a wide audience.

Selection of variables for candidate models

Two models were considered for this project. Considering domain knowledge in the insurance industry and with reference from Michigan Auto Law (<https://www.michiganautolaw.com/blog/2023/09/19/22-factors-that-affect-car-insurance-rates/>). The following variables were used to build the first model:

AGE - Age of Insured

GENDER - Gender of Insured

URBANICITY - Location of insured: whether insured lives in rural or urban area

MSTATUS - Marital status of the insured

OCCUPATION - Occupation of the insured

CAR_USE - The use of the insured's car

CAR_AGE - The age of the car

CLM_FREQ - Number of claims in the past 5 years

CAR_TYPE - Type of car

INCOME - Income of insured

The second model contains the above variables with three interactions: AGE x MSTATUS

CAR_AGE x CAR_USE

CLM_FREQ x URBANCITY

##Fitting the models

```
logistic_spec <- logistic_reg() %>%  
  set_engine("glm")  
  
logistic_spec  
  
## Logistic Regression Model Specification (classification)  
##  
## Computational engine: glm  
  
#Building model 1 with all variables selected based on domain knowledge  
claim_model_1 <- logistic_spec %>%  
  fit(CLAIM_FLAG ~ . , data = train_claim , family = "binomial")  
  
#Building model 2 with all variables selected based on domain knowledge and  
some interactons  
claim_model_2 <- logistic_spec %>%  
  fit(CLAIM_FLAG ~ . + AGE:MSTATUS + CAR_TYPE:CAR_USE + CLM_FREQ:URBANCITY ,  
data = train_claim , family = "binomial")
```

Model 1

```
tidy(claim_model_1, exponentiate = TRUE, conf.int = TRUE) %>%  
  knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.758	0.241	-1.152	0.250	0.472	1.214
AGE	0.985	0.003	-4.637	0.000	0.978	0.991
GENDERz_F	0.857	0.087	-1.783	0.075	0.722	1.015
URBANCITYz_Highly Rural/ Rural	0.104	0.110	-20.589	0.000	0.083	0.128
MSTATUSz_No	2.007	0.057	12.236	0.000	1.795	2.244
OCCUPATIONClerical	1.614	0.154	3.115	0.002	1.195	2.183
OCCUPATIONDoctor	0.647	0.234	-1.864	0.062	0.404	1.012
OCCUPATIONHome Maker	1.241	0.176	1.228	0.220	0.879	1.752
OCCUPATIONLawyer	1.072	0.155	0.446	0.656	0.791	1.452
OCCUPATIONManager	0.507	0.148	-4.594	0.000	0.379	0.677
OCCUPATIONProfessional	1.022	0.139	0.158	0.874	0.779	1.343
OCCUPATIONStudent	1.507	0.170	2.412	0.016	1.081	2.104
OCCUPATIONz_Blue Collar	1.357	0.138	2.213	0.027	1.037	1.780
CAR_USEPrivate	0.481	0.084	-8.717	0.000	0.408	0.567

term	estimate	std.error	statistic	p.value	conf.low	conf.high
CAR_AGE	0.984	0.006	-2.538	0.011	0.972	0.996
CLM_FREQ	1.276	0.023	10.641	0.000	1.220	1.335
CAR_TYPEPanel Truck	1.159	0.137	1.078	0.281	0.886	1.514
CAR_TYPEPickup	1.739	0.095	5.846	0.000	1.445	2.094
CAR_TYPESports Car	3.157	0.116	9.906	0.000	2.516	3.966
CAR_TYPEVan	1.597	0.118	3.967	0.000	1.266	2.011
CAR_TYPEz_SUV	2.334	0.099	8.558	0.000	1.924	2.837
INCOME	1.000	0.000	-5.843	0.000	1.000	1.000

From table of our claim model 1, The coefficient for the AGE predictor is 0.985. This means that for each one-unit increase in AGE, the odds of filing a claim increases multiplicatively by 0.985, assuming all other variables are held constant. The confidence interval lies between 0.978 and 0.991. Since the interval does not contain 1, it shows AGE is a significant. We can interpret the other predictors in similar fashion.

Model 2

```
tidy(claim_model_2, exponentiate = TRUE, conf.int = TRUE) %>%
  knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.487	0.294	-2.447	0.014	0.273	0.866
AGE	0.989	0.004	-2.457	0.014	0.981	0.998
GENDERz_F	0.864	0.087	-1.689	0.091	0.728	1.023
URBANICITYz_Highly Rural/ Rural	0.078	0.131	-	0.000	0.060	0.100
MSTATUSz_No	3.143	0.292	3.926	0.000	1.776	5.574
OCCUPATIONClerical	1.703	0.155	3.443	0.001	1.259	2.308
OCCUPATIONDoctor	0.691	0.234	-1.574	0.115	0.431	1.083
OCCUPATIONHome Maker	1.302	0.176	1.496	0.135	0.921	1.841
OCCUPATIONLawyer	1.131	0.156	0.790	0.430	0.833	1.537
OCCUPATIONManager	0.524	0.148	-4.367	0.000	0.391	0.700
OCCUPATIONProfessional	1.060	0.139	0.419	0.675	0.807	1.394
OCCUPATIONStudent	1.632	0.173	2.840	0.005	1.165	2.291
OCCUPATIONz_Blue Collar	1.535	0.142	3.028	0.002	1.164	2.029
CAR_USEPrivate	0.607	0.153	-3.254	0.001	0.450	0.821
CAR_AGE	0.985	0.006	-2.410	0.016	0.973	0.997
CLM_FREQ	1.244	0.023	9.304	0.000	1.188	1.302
CAR_TYPEPanel Truck	1.426	0.167	2.119	0.034	1.028	1.981

term	estimate	std.error	statistic	p.value	conf.low	conf.high
CAR_TYPEPickup	2.211	0.147	5.386	0.000	1.659	2.957
CAR_TYPESports Car	2.686	0.226	4.374	0.000	1.727	4.189
CAR_TYPEVan	2.217	0.168	4.727	0.000	1.595	3.087
CAR_TYPEz_SUV	2.458	0.170	5.297	0.000	1.765	3.434
INCOME	1.000	0.000	-5.762	0.000	1.000	1.000
AGE:MSTATUSz_No	0.990	0.006	-1.556	0.120	0.977	1.003
CAR_USEPrivate:CAR_TYPEPanel Truck	NA	NA	NA	NA	NA	NA
CAR_USEPrivate:CAR_TYPEPickup	0.652	0.200	-2.143	0.032	0.440	0.962
CAR_USEPrivate:CAR_TYPESports Car	1.174	0.246	0.651	0.515	0.724	1.899
CAR_USEPrivate:CAR_TYPEVan	0.489	0.254	-2.812	0.005	0.295	0.801
CAR_USEPrivate:CAR_TYPEz_SUV	0.919	0.186	-0.456	0.649	0.638	1.322
URBANICITYz_Highly Rural/ Rural:CLM_FREQ	1.636	0.096	5.123	0.000	1.353	1.975

From table of our claim model 2, the coefficient for the CLM_FREQ predictor is 1.244. This means that for each one-unit increase in having a claim in the last 5 years, the odds of filing a claim increases multiplicatively by 1.244, assuming all other variables are held constant. The confidence interval lies between 1.188 and 1.302. Since the interval does not contain 1, it shows CLM_FREQ is a significant predictor. We can interpret the other predictors in similar fashion.

4. Assessing statistical models performance

Evaluating model performance metrics such as accuracy, sensitivity, specificity and the area under the ROC curve (AUC-ROC) provides insight into how well the models discriminate between positive and negative outcomes. In the context of car insurance claims, where correctly identifying claimants is crucial for risk assessment and pricing, metrics like precision (the proportion of predicted claimants who actually file a claim) and recall (the proportion of actual claimants correctly identified by the model) are particularly important.

Additionally, the AUC-ROC metric summarizes the model's ability to distinguish between claimants and non-claimants across various decision thresholds, providing a comprehensive assessment of its discriminative power. Assessment of both models is conducted using the test set to prevent the likelihood of overfitting.

Assessing model 1's performance

```
predictions <- predict(claim_model_1, new_data = test_claim, type = "prob")
# 1. Confusion Matrix and Classification Metrics
# Convert predicted probabilities to class predictions
class_predictions <- ifelse(predictions$.pred_0 > 0.5, 1, 0)

# Ensure that class predictions and actual outcomes are factors with the same levels
class_predictions <- factor(class_predictions, levels =
levels(test_claim$CLAIM_FLAG))

# Calculate confusion matrix
conf_matrix <- confusionMatrix(class_predictions, test_claim$CLAIM_FLAG)

tidy(conf_matrix, exponentiate = TRUE, conf.int = TRUE) %>%
  knitr::kable(digits = 3)
```

term	class	estimate	conf.low	conf.high	p.value
accuracy	NA	0.247	0.229	0.267	1
kappa	NA	-0.142	NA	NA	NA
mcnemar	NA	NA	NA	NA	0
sensitivity	0	0.077	NA	NA	NA
specificity	0	0.690	NA	NA	NA
pos_pred_value	0	0.390	NA	NA	NA
neg_pred_value	0	0.224	NA	NA	NA
precision	0	0.390	NA	NA	NA
recall	0	0.077	NA	NA	NA
f1	0	0.128	NA	NA	NA
prevalence	0	0.721	NA	NA	NA
detection_rate	0	0.055	NA	NA	NA
detection_prevalence	0	0.142	NA	NA	NA
balanced_accuracy	0	0.383	NA	NA	NA

Interpreting the performance metrics of model 1 based on the confusion matrix:

Accuracy: The accuracy of 0.247 implies that approximately 24.7% of the model's predictions regarding whether a claim was filed are correct. However, this accuracy rate is relatively low, indicating that the model's overall performance in accurately classifying claim filings is poor.

Sensitivity: With a sensitivity of 0.077, the model correctly identifies only 7.7% of the actual instances where a claim was filed. A low sensitivity suggests that the model struggles to capture instances of filed claims effectively. This shortfall is particularly concerning for

insurers, as it means that a substantial portion of claim filings may go undetected or misclassified, potentially leading to inadequate coverage or loss for policyholders.

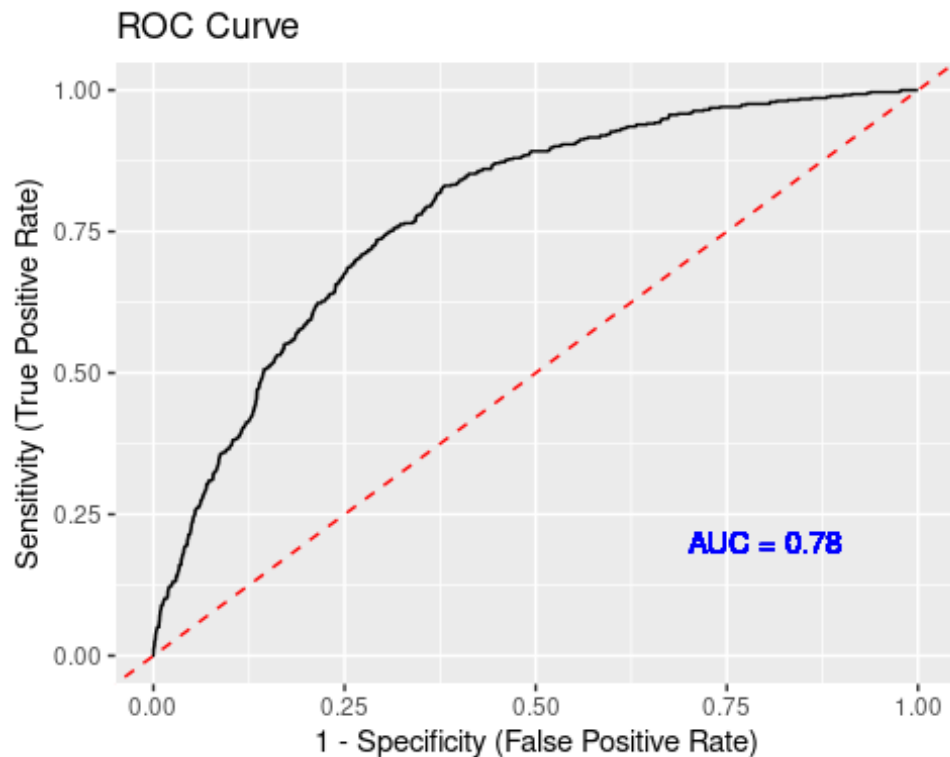
Specificity: The specificity of 0.690 indicates that the model performs better at correctly identifying instances where no claim was filed, with 69.0% of such instances being classified accurately. While high specificity is desirable to avoid incorrectly flagging non-claims, it must be balanced with sensitivity to ensure that actual claim filings are not overlooked. In the insurance context, maintaining a balance between sensitivity and specificity is crucial for accurately assessing risk, determining premiums, and providing timely support to policyholders.

```
# Receiver Operating Characteristic (ROC) Curve and AUC
roc_curve <- roc(test_claim$CLAIM_FLAG, predictions$.pred_0)

# Calculate AUC value
auc_value <- auc(roc_curve)

# Plot ROC curve using ggplot
roc_data <- data.frame(
  specificity = 1 - roc_curve$specificities,
  sensitivity = roc_curve$sensitivities
)

ggplot(roc_data, aes(x = specificity, y = sensitivity)) +
  geom_line() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  labs(x = "1 - Specificity (False Positive Rate)", y = "Sensitivity (True
Positive Rate)", title = "ROC Curve") +
  geom_text(x = 0.8, y = 0.2, label = paste("AUC =", round(auc_value, 2)),
color = "blue")
```



AUC of 0.78 signifies that the model has moderate discriminatory power and performs better than random chance in distinguishing between positive and negative instances.

Assessing model 2's performance

```
predictions <- predict(claim_model_2, new_data = test_claim, type = "prob")
# 1. Confusion Matrix and Classification Metrics
# Convert predicted probabilities to class predictions
class_predictions <- ifelse(predictions$.pred_0 > 0.5, 1, 0)

# Ensure that class predictions and actual outcomes are factors with the same levels
class_predictions <- factor(class_predictions, levels =
levels(test_claim$CLAIM_FLAG))

# Calculate confusion matrix
conf_matrix <- confusionMatrix(class_predictions, test_claim$CLAIM_FLAG)

tidy(conf_matrix, exponentiate = TRUE, conf.int = TRUE) %>%
  knitr::kable(digits = 3)
```

term	class	estimate	conf.low	conf.high	p.value
accuracy	NA	0.248	0.229	0.267	1
kappa	NA	-0.144	NA	NA	NA
mcnemar	NA	NA	NA	NA	0

term	class	estimate	conf.low	conf.high	p.value
sensitivity	0	0.079	NA	NA	NA
specificity	0	0.685	NA	NA	NA
pos_pred_value	0	0.395	NA	NA	NA
neg_pred_value	0	0.223	NA	NA	NA
precision	0	0.395	NA	NA	NA
recall	0	0.079	NA	NA	NA
f1	0	0.132	NA	NA	NA
prevalence	0	0.721	NA	NA	NA
detection_rate	0	0.057	NA	NA	NA
detection_prevalence	0	0.145	NA	NA	NA
balanced_accuracy	0	0.382	NA	NA	NA

Interpreting the performance metrics of model 2 based on the confusion matrix:

Accuracy: The accuracy of 0.248 implies that approximately 24.8% of the model's predictions regarding whether a claim was filed are correct. However, this accuracy rate is relatively low, indicating that the model's overall performance in accurately classifying claim filings is poor.

Sensitivity: With a sensitivity of 0.079, the model correctly identifies only 7.9% of the actual instances where a claim was filed. A low sensitivity suggests that the model struggles to capture instances of filed claims effectively. This deficiency is particularly concerning for insurers, as it means that a substantial portion of claim filings may go undetected or misclassified, potentially leading to inadequate coverage or loss for policyholders.

Specificity: The specificity of 0.685 indicates that the model performs better at correctly identifying instances where no claim was filed, with 68.5% of such instances being classified accurately. While high specificity is desirable to avoid incorrectly flagging non-claims, it must be balanced with sensitivity to ensure that actual claim filings are not overlooked. In the insurance context, maintaining a balance between sensitivity and specificity is crucial for accurately assessing risk, determining premiums, and providing timely support to policyholders.

These values and interpretations are similar to values obtained in the performance of model 1, which suggests that interaction variables in model 2 may be insignificant.

```
# Receiver Operating Characteristic (ROC) Curve and AUC
roc_curve <- roc(test_claim$CLAIM_FLAG, predictions$.pred_0)

# Calculate AUC value
auc_value <- auc(roc_curve)

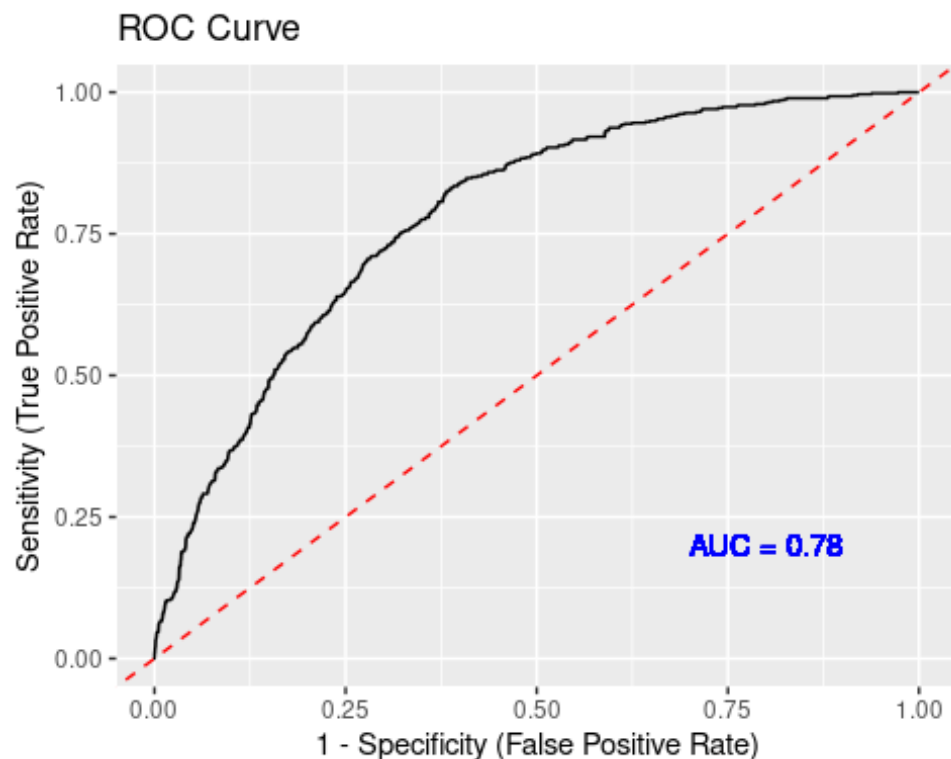
# Plot ROC curve using ggplot
roc_data <- data.frame(
```

```

    specificity = 1 - roc_curve$specificities,
    sensitivity = roc_curve$sensitivities
)

ggplot(roc_data, aes(x = specificity, y = sensitivity)) +
  geom_line() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  labs(x = "1 - Specificity (False Positive Rate)", y = "Sensitivity (True
Positive Rate)", title = "ROC Curve") +
  geom_text(x = 0.8, y = 0.2, label = paste("AUC =", round(auc_value, 2)),
color = "blue")

```



AUC of 0.78 obtained in model 2 suggestions interactions may be insignificant in the model.

Model 1 preferred

The striking similarity in the performance scores of both models suggests that the interactions included in Model 2 are likely insignificant. Therefore, Model 1, which excludes these interactions, should be preferred over Model 2. Moving forward, expanding domain knowledge in the insurance sector and exploring additional significant variables could lead to further improvements in model accuracy and robustness.

5.Conclusion

The modeling process encompasses various key components, including exploratory data analysis (EDA) for insights, model selection and training using logistic regression, and thorough evaluation of model performance using metrics such as accuracy, sensitivity, specificity, and AUC-ROC.

Furthermore, interpreting model results, such as the AUC value, is critical for measuring its predictive capability and making informed decisions. While an AUC of 0.78 suggests reasonably good discriminatory power, the model has the potential to do better with the inclusion of significant predictors.

In summary, predictive modeling for car insurance claims holds immense potential for enhancing decision-making processes and optimizing business outcomes for the insurance sector in general. By harnessing data-driven insights and employing robust modeling methodologies, insurance companies can strengthen their risk assessment capabilities and deliver excellent services to their clients.