Data 621: Assignment 4

*Car Insurance Data*

*Ritesh Lohiya*

*July 8, 2018*

## *Overview*

**In this homework assignment, you will explore, analyze and model a data set containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, TARGET\_FLAG, is a 1 or a 0. A “1” means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is TARGET\_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.**

**Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:**

**Data Exploration:**

The training data set includes 8,161 observations, with 26 variables: 23 predictors, two response variables, and one record identifier. Below is a brief description of the included variables:

| **Variable Name** | **Description** | **Theoretical Impact** |
| --- | --- | --- |
| INDEX | Identification Variable (do not use) | None |
| TARGET\_FLAG | In a crash? 1=YES 0=NO | None |
| TARGET\_AMT | Cost of Crash, if applicable | None |
| KIDSDRIV | # Driving Children | When teenagers drive your car, increased crash risk |
| AGE | Age of Driver | Young and old drivers might be riskier |
| HOMEKIDS | # Children at Home | Unknown effect |
| YOJ | Years on Job | Long-term employees are usually safer |
| INCOME | Income | In theory, rich have fewer crashes |
| PARENT1 | Single Parent | Unknown impact |
| HOME\_VAL | Home Value | In theory, home owners may drive more responsibly |
| MSTATUS | Marital Status | In theory, married individuals are less risky |
| SEX | Gender | Urban legend: females are safer drivers |
| EDUCATION | Max Education Level | Unknown, but in theory more educated people tend to drive more safely |
| JOB | Job Category | In theory, white collar workers are less risky |
| TRAVTIME | Commute Distance | Long drives to work usually suggest greater risk |
| CAR\_USE | Vehicle Use | Commercial fleet driven more, may impact collision prob |
| BLUEBOOK | Value of Vehicle | Unknown impact on collision prob, but impacts crash payout |
| TIF | Time in Force | Long-term customers are usually safer |
| CAR\_TYPE | Type of Car | Unknown impact on collision prob, but impacts crash payout |
| RED\_CAR | A Red Car | Urban legend: red cars are riskier, particularly sports cars |
| OLDCLAIM | # Claims (Past 5 Years) | If total payout high, future payouts might be high |
| CLM\_FREQ | Total Claims (Past 5 Years) | Claim count should be positively correlated with future claims |
| REVOKED | License Revoked (Past 7 Years) | If your license was revoked, you probably are a riskier driver |
| MVR\_PTS | Motor Vehicle Report Points | Traffic ticket counts have postive correlation with crashes |
| CAR\_AGE | Vehicle Age | Unknown impact on collision prob, but impacts crash payout |
| URBANICITY | Home/Work Area | Unknown impact |

INCOME, HOME\_VAL, BLUEBOOK, and OLDCLAIM are represented as strings. So we will be extracting the numeric values for these.

ins\_train$INCOME <- as.numeric(str\_replace\_all(ins\_train$INCOME, "[[:punct:]\\$]",""))

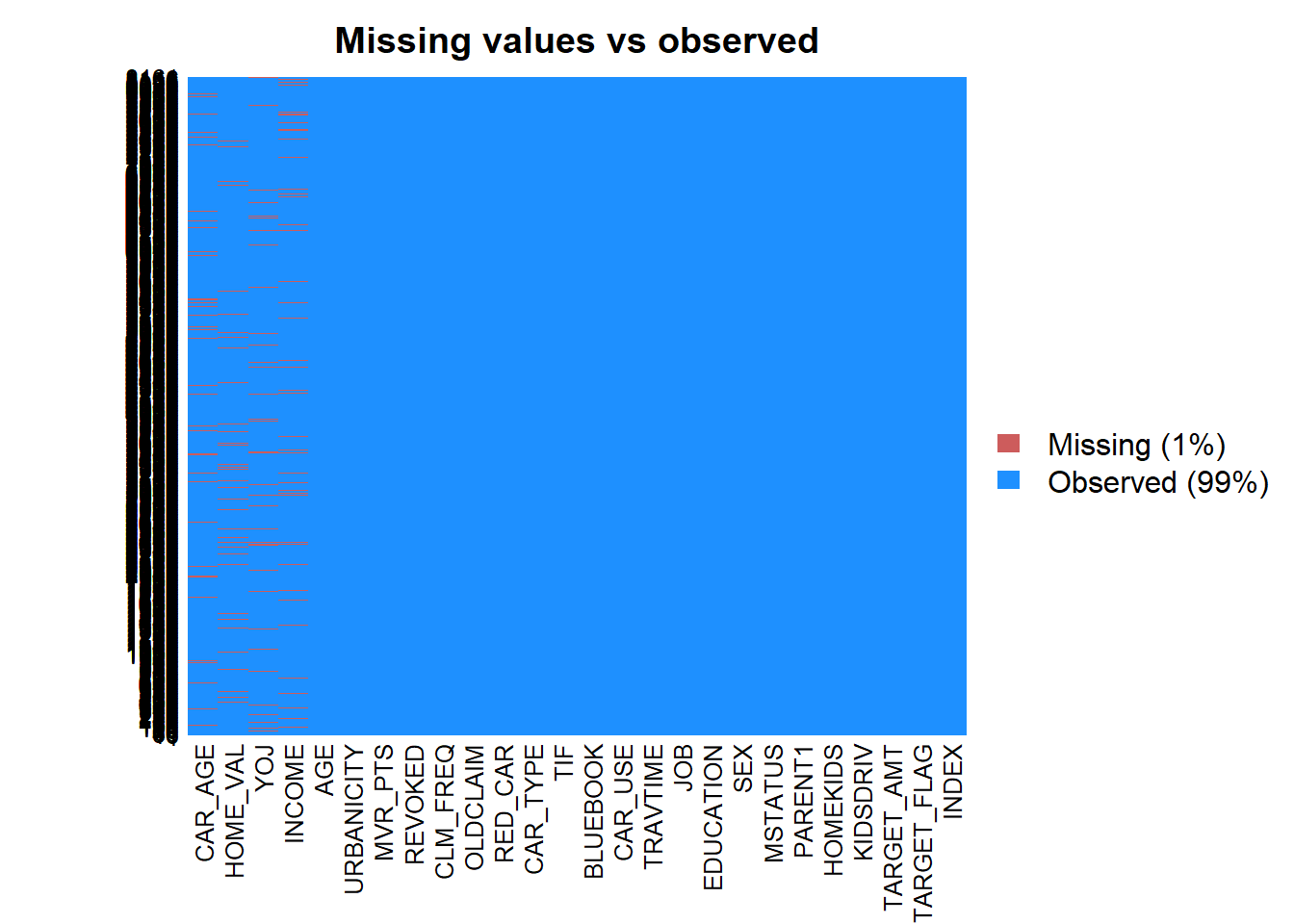
ins\_train$HOME\_VAL <- as.numeric(str\_replace\_all(ins\_train$HOME\_VAL, "[[:punct:]\\$]",""))

ins\_train$BLUEBOOK <- as.numeric(str\_replace\_all(ins\_train$BLUEBOOK, "[[:punct:]\\$]",""))

ins\_train$OLDCLAIM <- as.numeric(str\_replace\_all(ins\_train$OLDCLAIM, "[[:punct:]\\$]",""))

### Missing Values

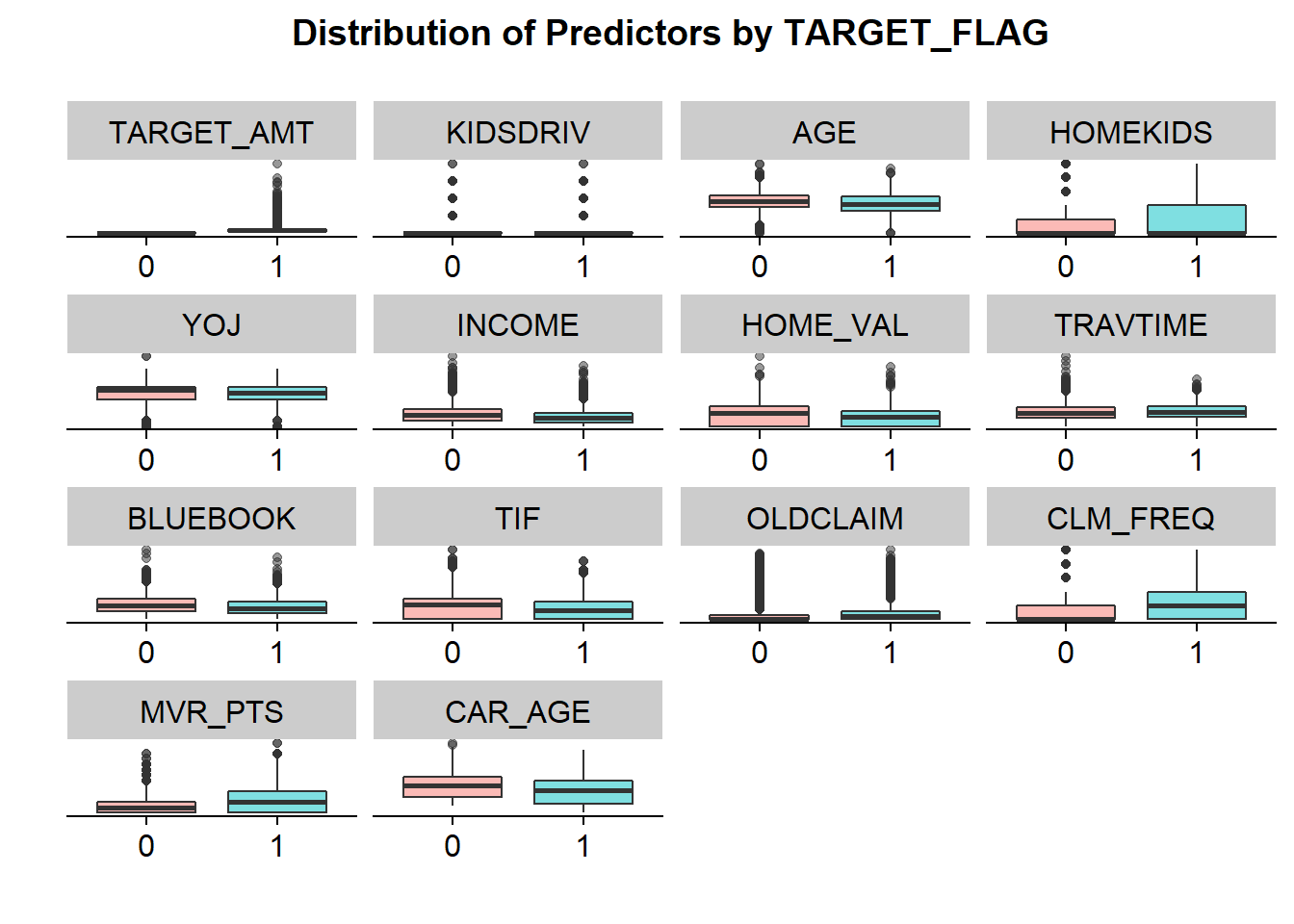
Now we will see the missing values in the dataset. For this i have used Amelia package. We can see there are missing values for CAR\_AGE, HOME\_VAL, YOJ and INCOME. There needs to be taken care while we do data preparation.



Now lets do some plots to understand the data:

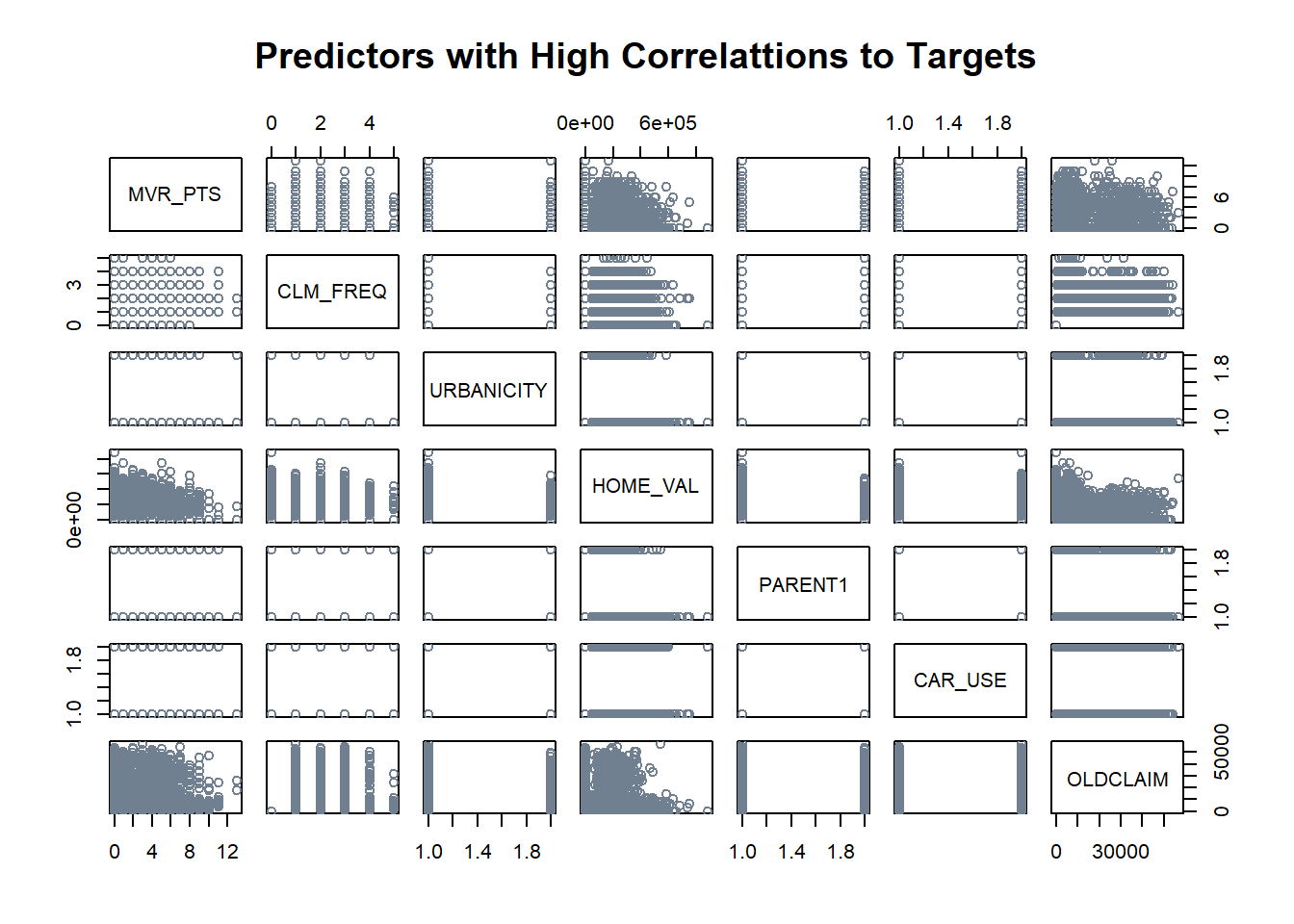
AGE - Age of Driver. Very young people tend to be risky. Maybe very old people also. We note six missing values that we’ll need to address later. The distribution of AGE is almost perfectly normal. When we break out the data by TARGET\_FLAG values, the distributions of age by TARGET\_FLAG are still roughly normal.

Boxplots are generated for non-binary variables split by TARGET\_FLAG:



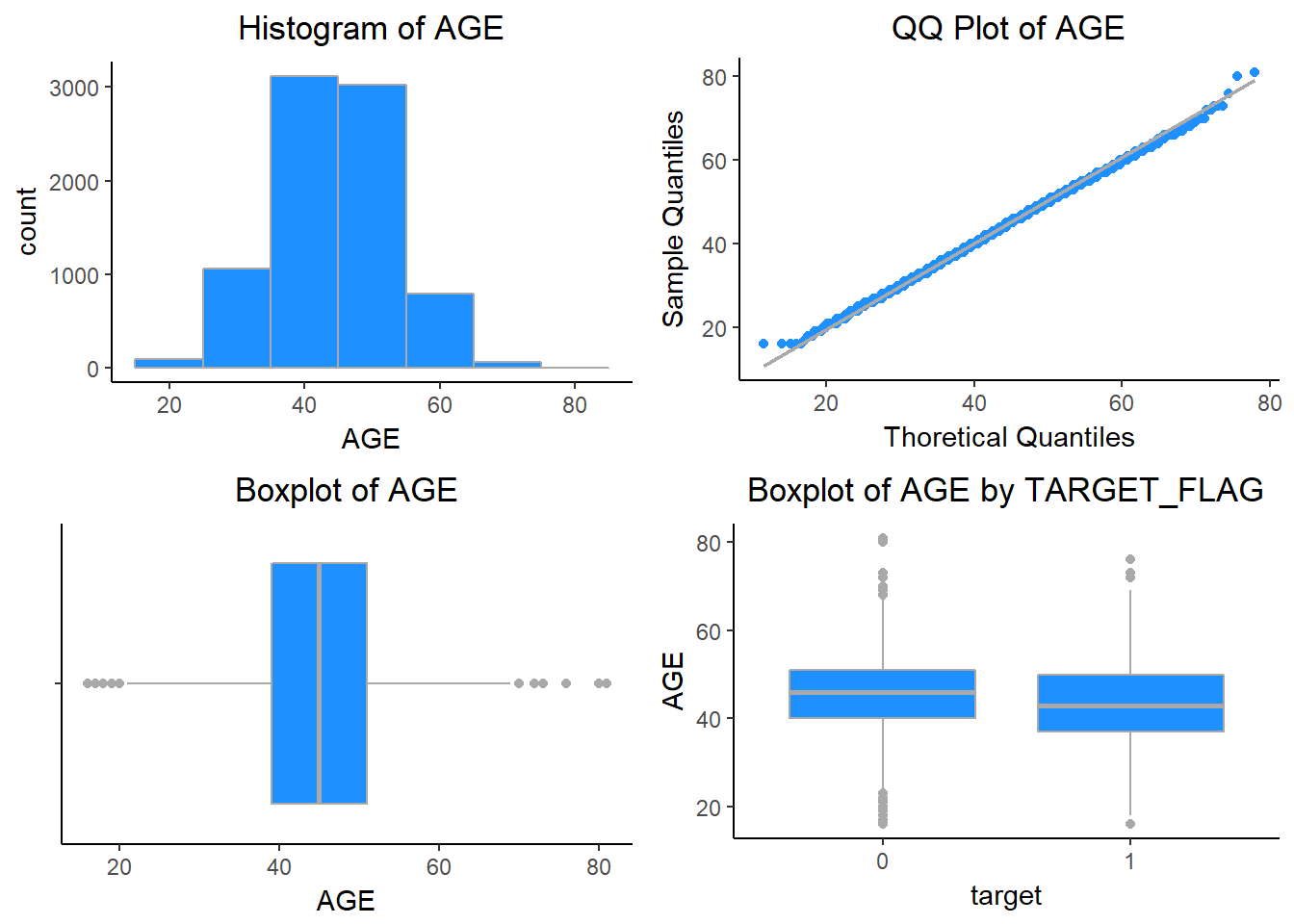
### Correlation

The correlation plot below shows how variables in the dataset are related to each other. Looking at the plot, we can see that certain variables are more related than others.

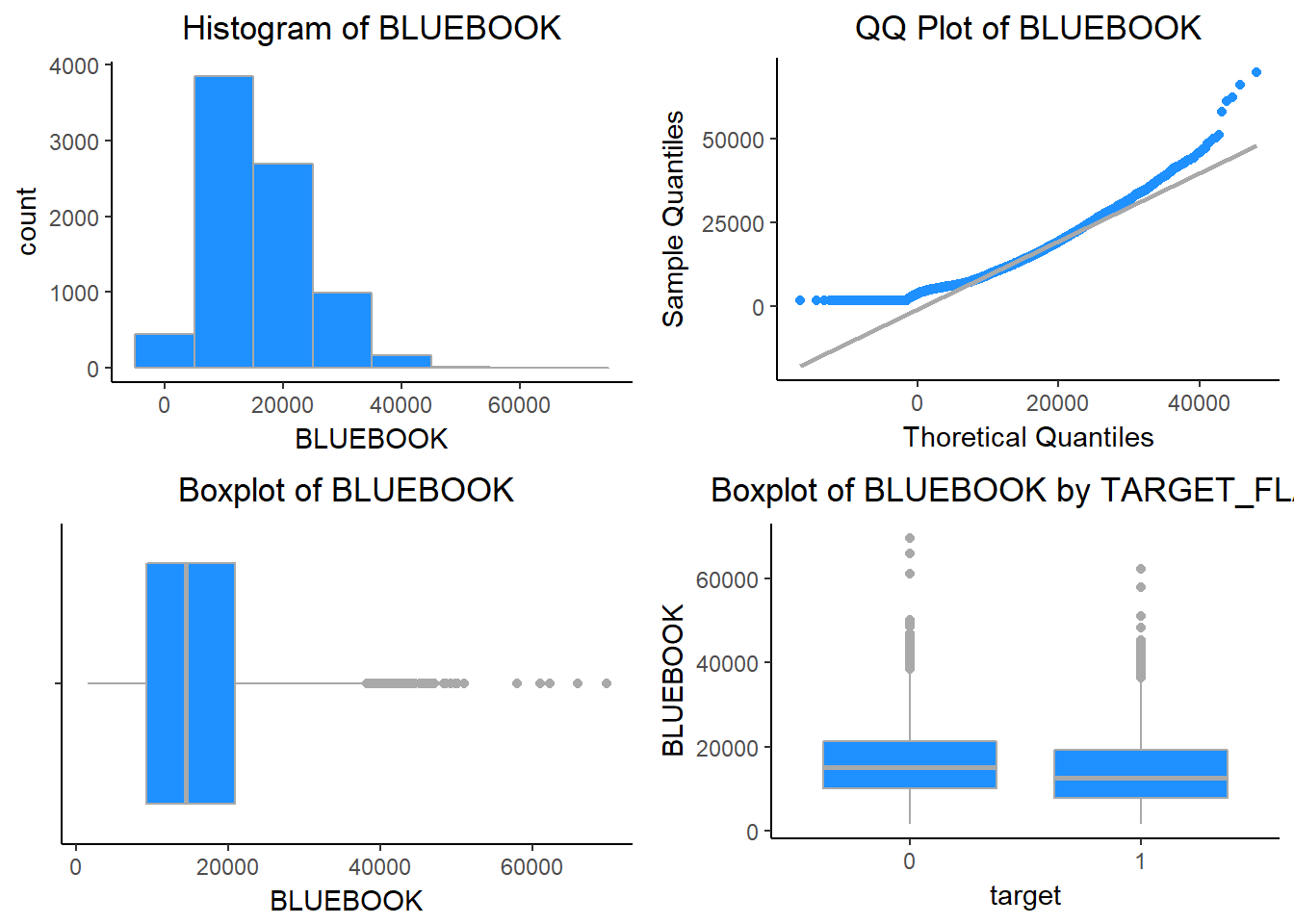


Now let’s do some plots to understand the data:

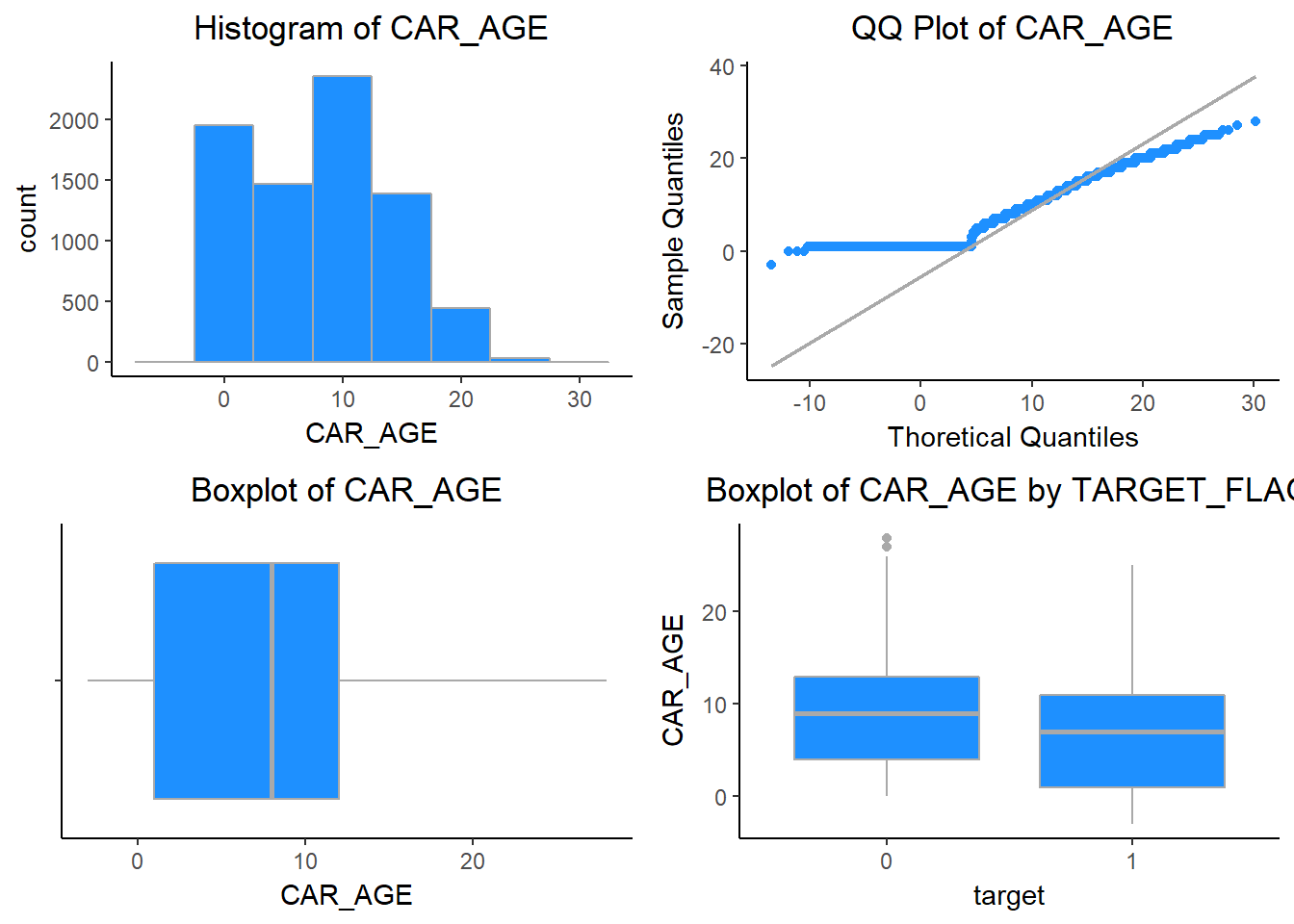
AGE - Age of Driver. Very young people tend to be risky. Maybe very old people also. We note six missing values that we’ll need to address later. The distribution of AGE is almost perfectly normal. When we break out the data by TARGET\_FLAG values, the distributions of age by TARGET\_FLAG are still roughly normal.



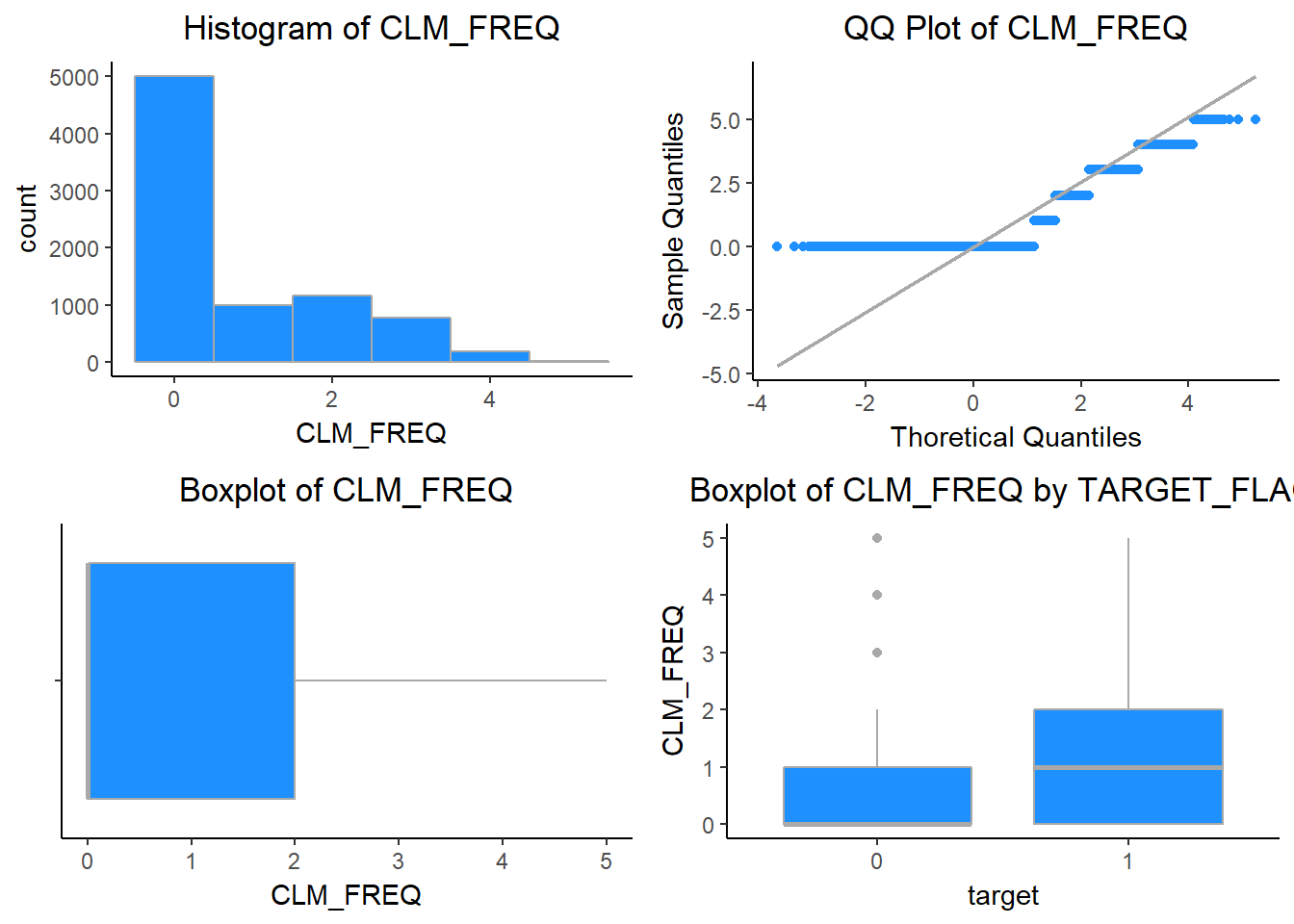
BLUEBOOK - Value of Vehicle. Unknown effect on probability of collision, but probably effect the payout if there is a crash. Individuals involved in crashes have a higher proportion of low BLUEBOOK values.



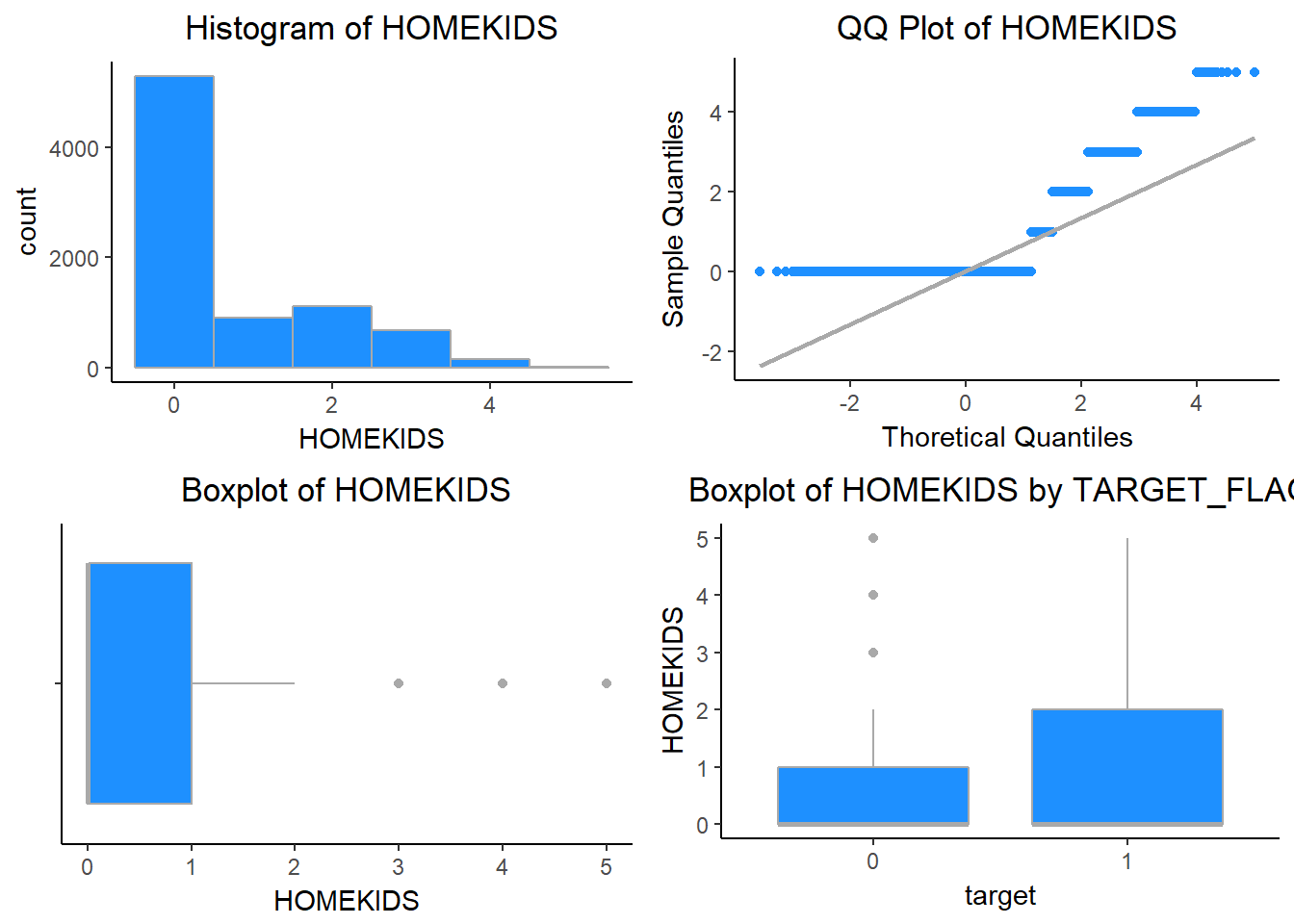
CAR\_AGE - Vehicle Age. We could see there is one negative value for CAR\_AGE. We have to treat this value in our data preparation step.



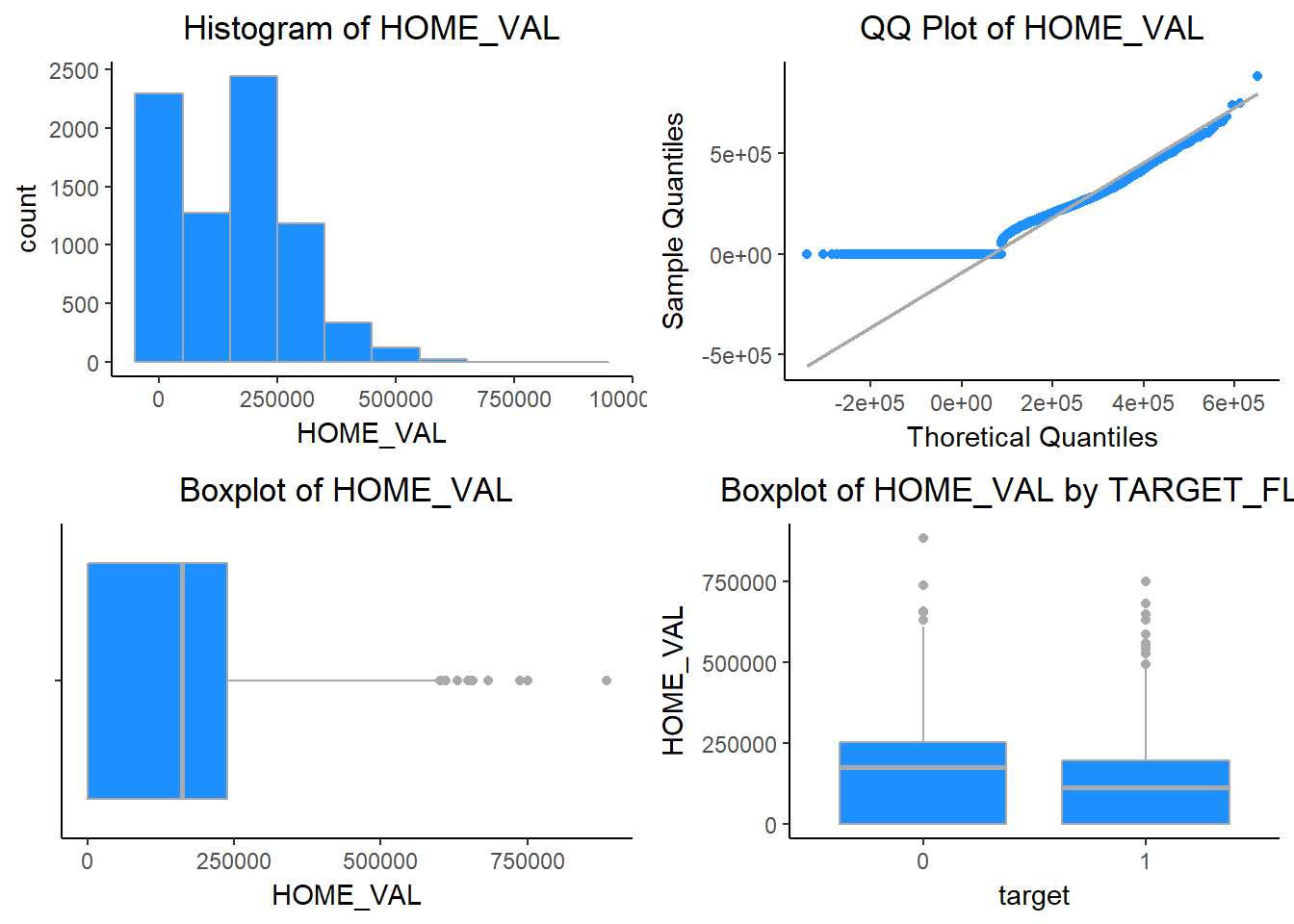
CLM\_FREQ - # Claims (Past 5 Years). The more claims you filed in the past, the more you are likely to file in the future. We can see that this variable is also skewed.



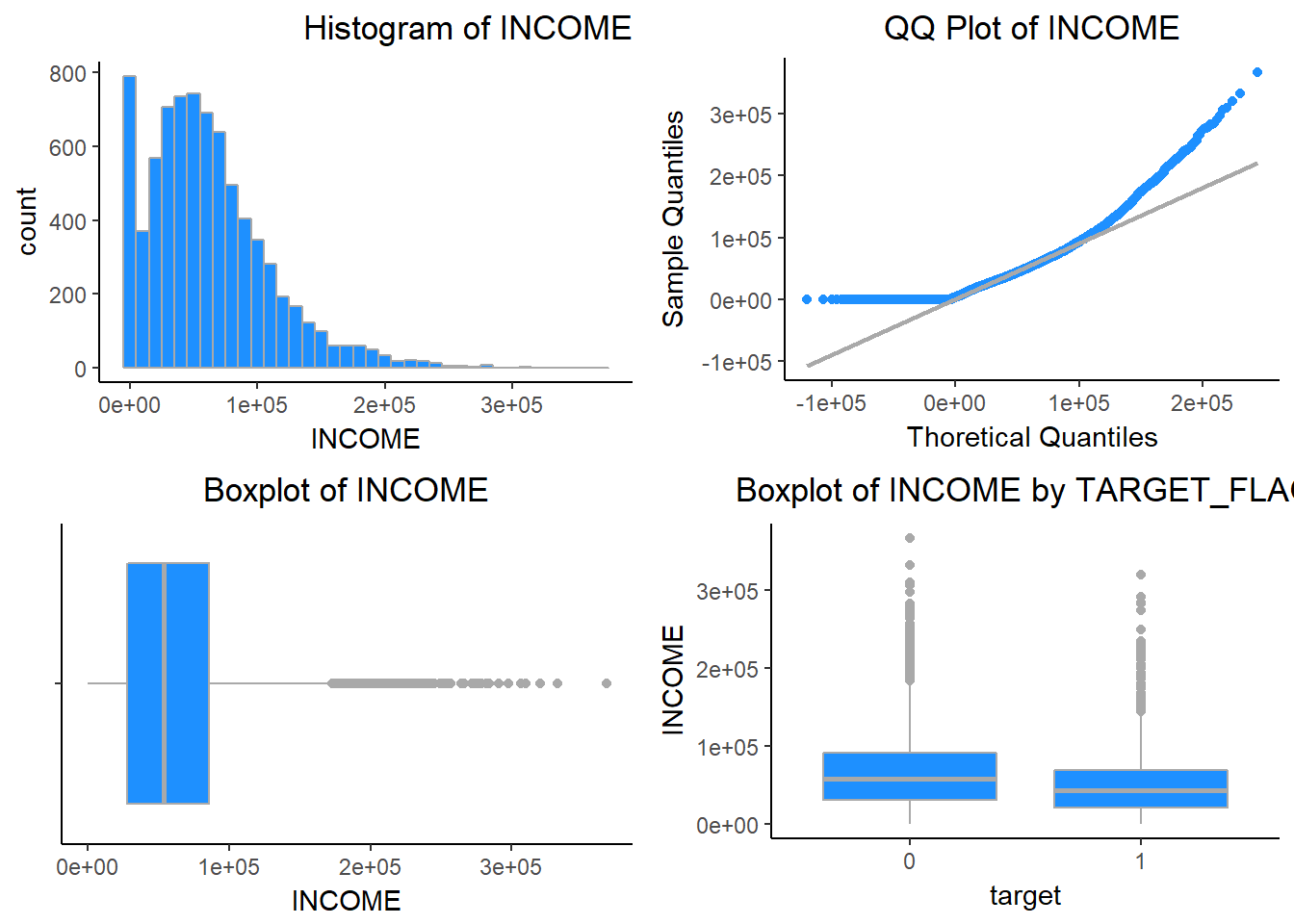
HOMEKIDS - # Children at Home. HOMEKIDS does not seem to impact the TARGET\_FLAG. The distribution of this discrete variable is right skewed.



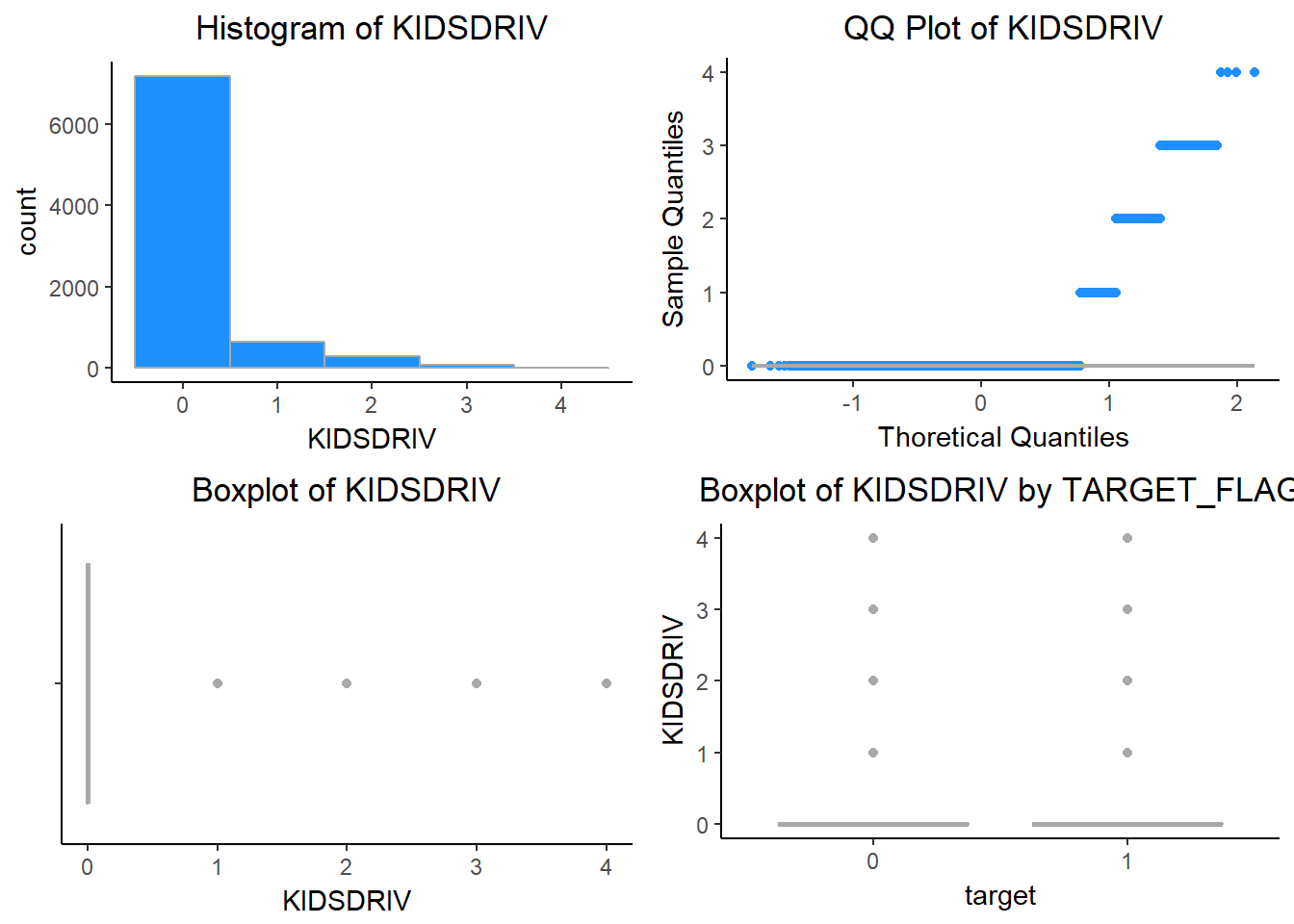
HOME\_VAL - Home Value. Home owners tend to drive more responsibly. The distribution of HOME\_VAL is right skewed and also we can there are some missing values.



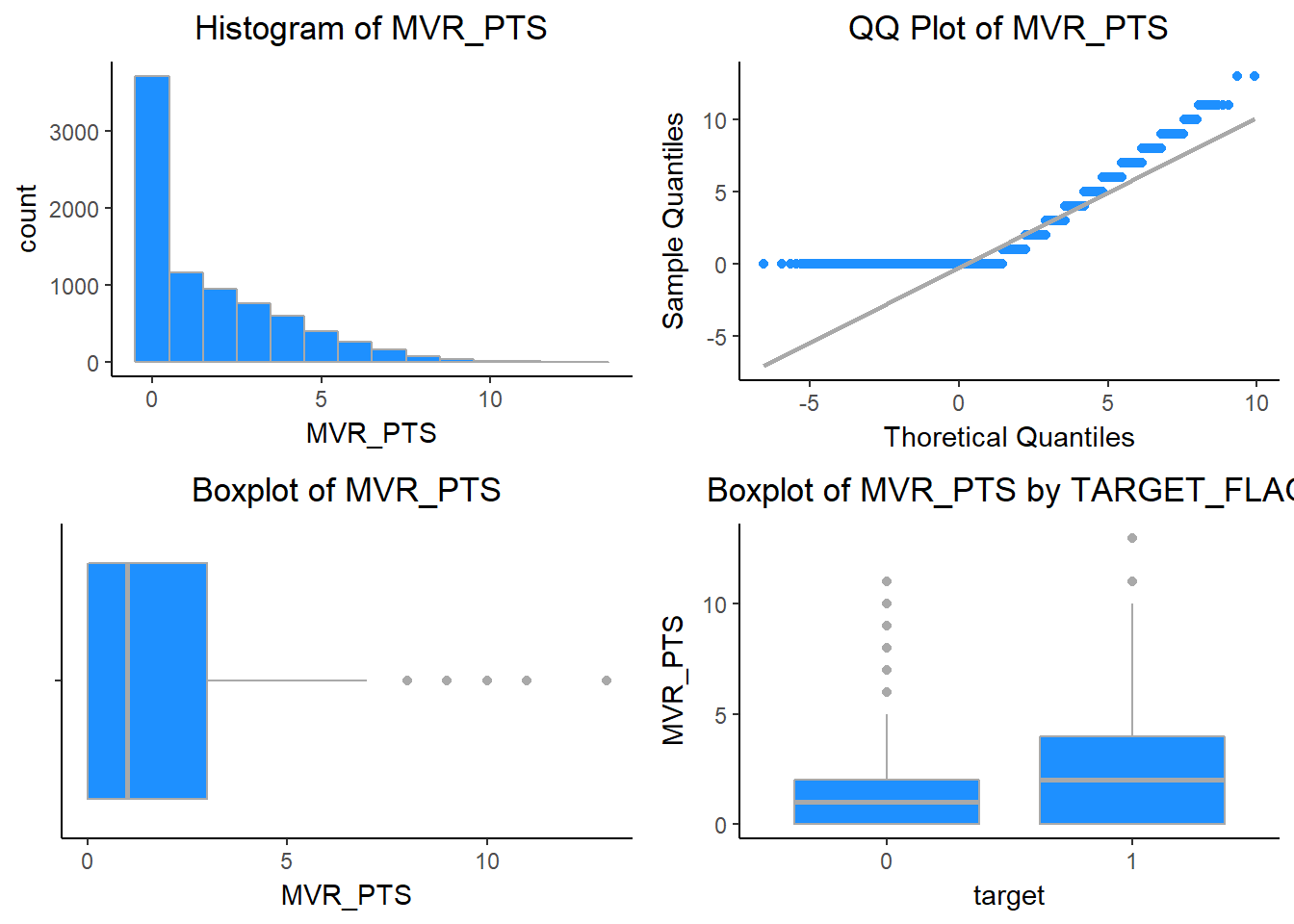
INCOME - Income of the person. Rich people tend to get into fewer crashes. The distribution of INCOME is right skewed, with a significant number of observations indicating $0 in income. There are some missing values in this as well.



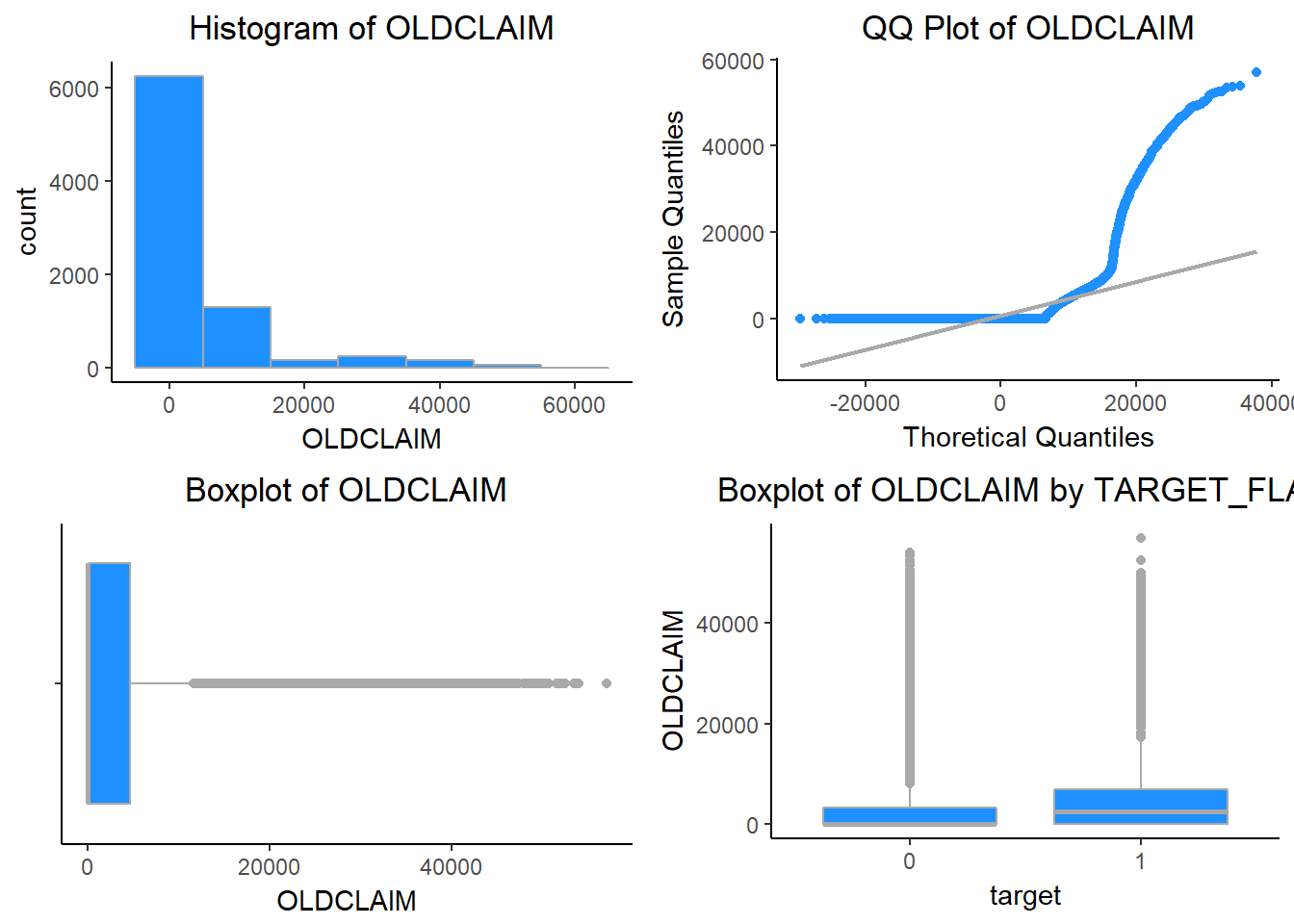
KIDSDRIV - # Driving Children. When teenagers drive your car, you are more likely to get into crashes. The discrete variable KIDSDRIV is right skewed



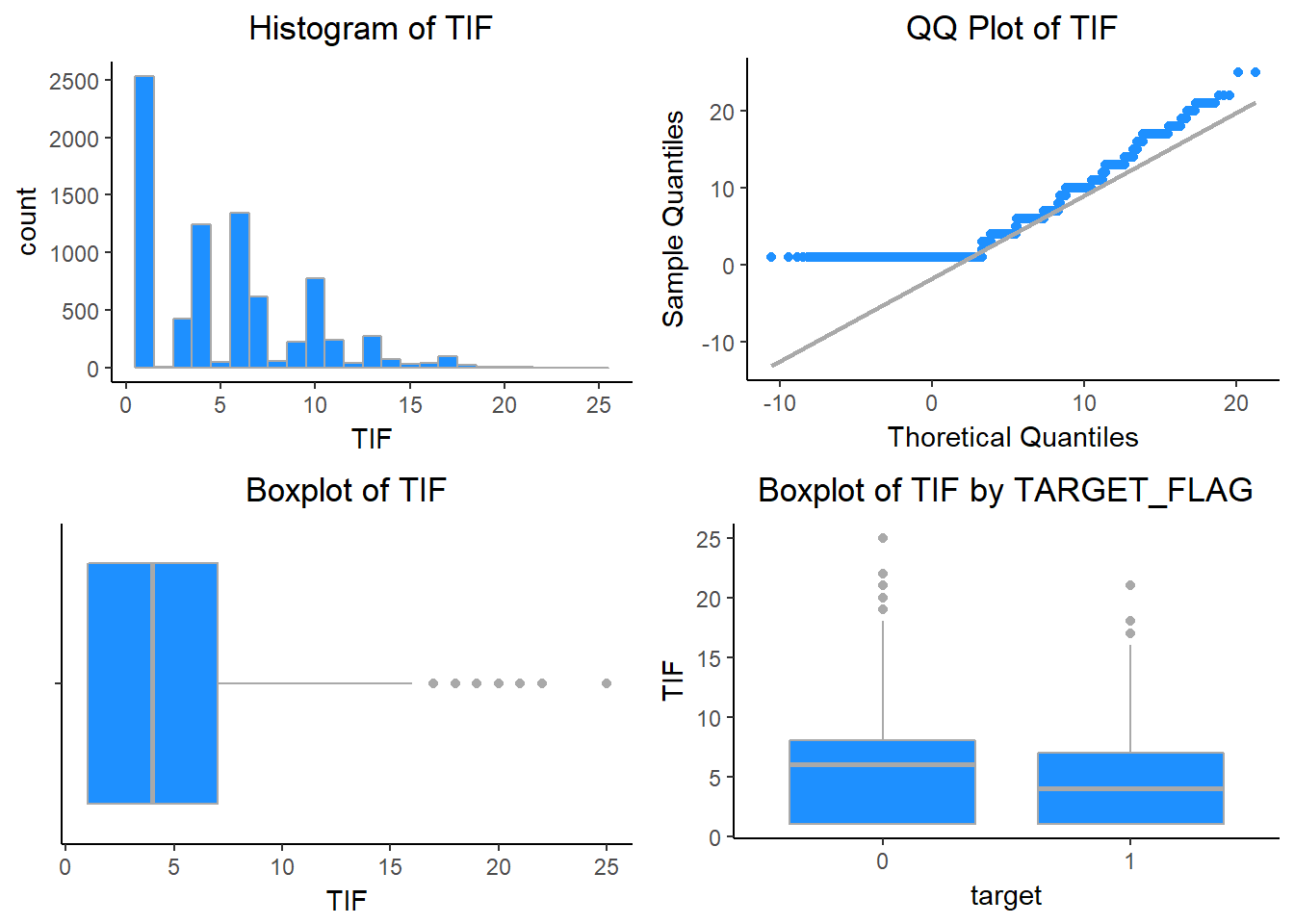
MVR\_PTS - Motor Vehicle Record Points. If you get lots of traffic tickets, you tend to get into more crashes. MVR\_PTS is positively skewed.



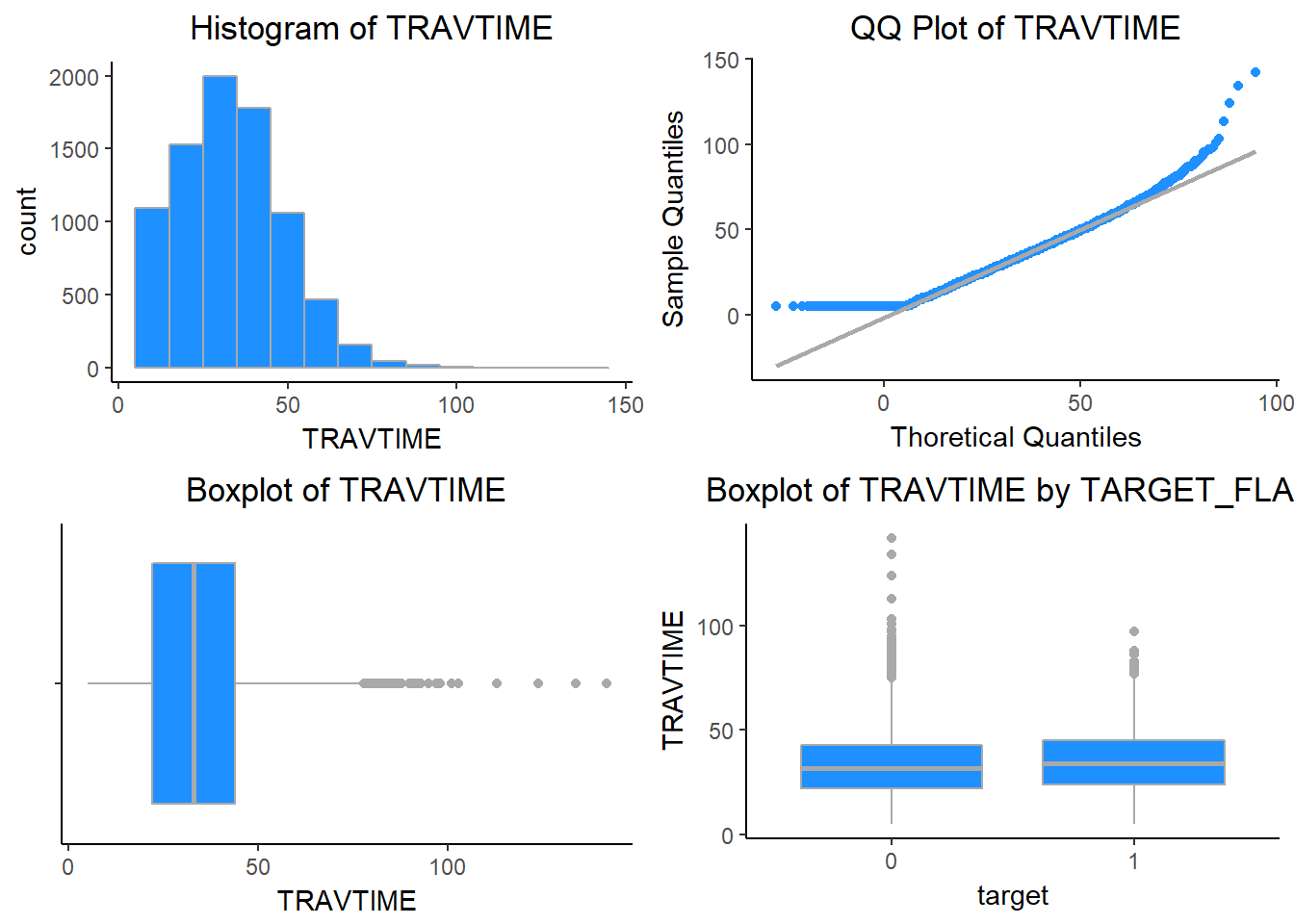
OLDCLAIM - Total Claims (Past 5 Years). If your total payout over the past five years was high, this suggests future payouts will be high. The distribution of OLDCLAIM is extremely right skewed.



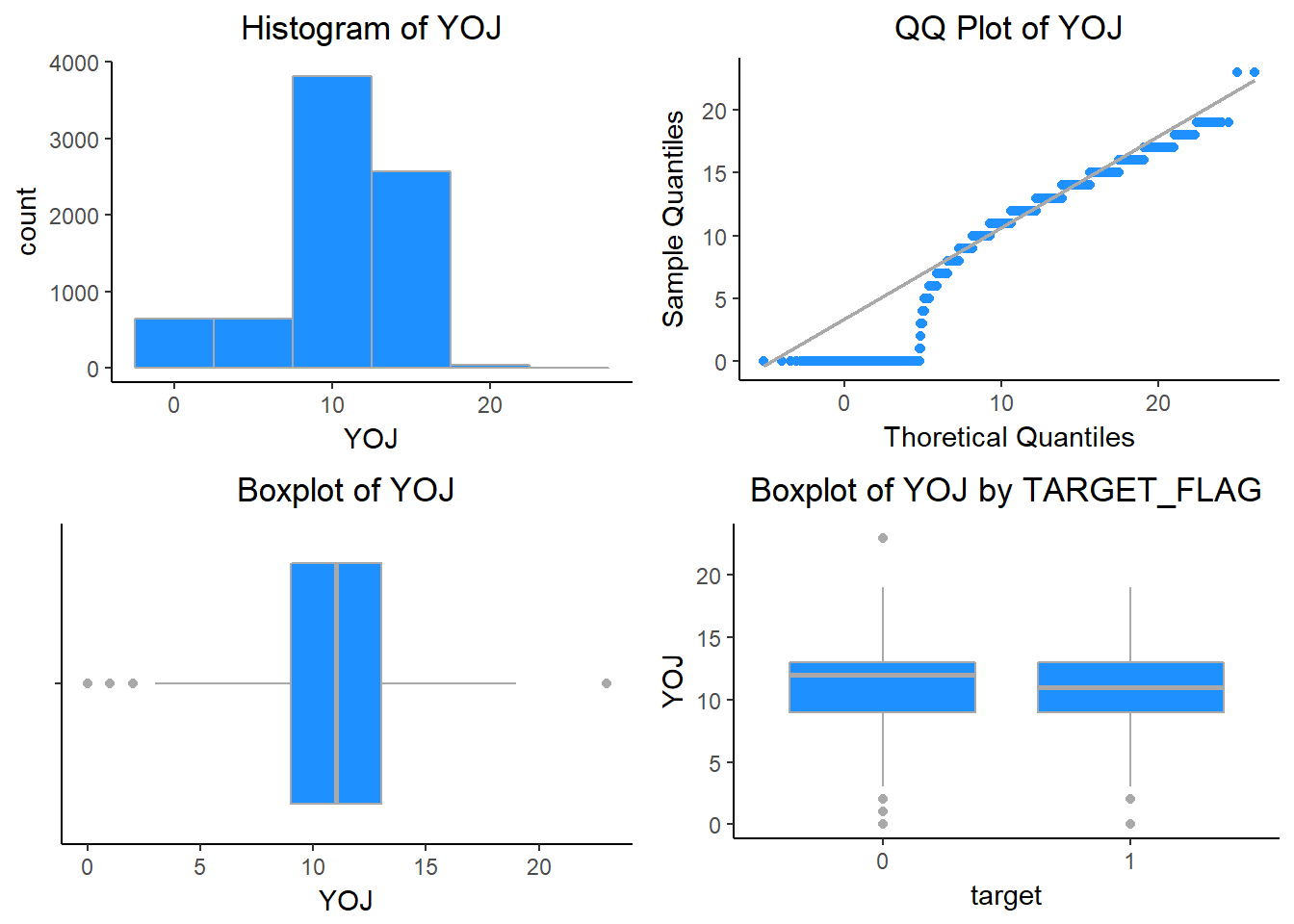
TIF - Time in Force. People who have been customers for a long time are usually more safe. The distribution is somewhat positively skewed.



TRAVTIME - Distance to Work. Long drives to work usually suggest greater risk. The distribution has a slight positive skew. The subset of insureds with no accidents have a higher proportion of individuals with short commute times.



YOJ - Years on Job. People who stay at a job for a long time are usually more safe. The variable would be approximately normally distributed if it weren’t for the high percentage of individuals with less than one year on the job.



EDUCATION - Unknown effect, but in theory more educated people tend to drive more safely.

## <High School Bachelors Masters PhD z\_High School Sum

## count 1203.0 2242.0 1658.0 728.0 2330.0 8161

## percent 14.7 27.5 20.3 8.9 28.6 100

REVOKED - License Revoked (Past 7 Years). If your license was revoked in the past 7 years, you probably are a more risky driver. Only 12% of drivers in the training data have a former license suspension on record.

## TARGET\_FLAG

## REVOKED 0 1 Sum

## No 5451 1710 7161

## Yes 557 443 1000

## Sum 6008 2153 8161

RED\_CAR - A Red Car. Urban legend says that red cars (especially red sports cars) are more risky. Is that true?. 30% of vehicles in the red category.

## TARGET\_FLAG

## RED\_CAR 0 1 Sum

## no 4246 1537 5783

## yes 1762 616 2378

## Sum 6008 2153 8161

CAR\_USE - Vehicle Use. Commercial vehicles are driven more, so might increase probability of collision. 60% car usage is private.

## TARGET\_FLAG

## CAR\_USE 0 1 Sum

## Commercial 1982 1047 3029

## Private 4026 1106 5132

## Sum 6008 2153 8161

SEX - Gender. Urban legend says that women have less crashes then men. Is that true?. The split between males and females is split almost 50/50.

## TARGET\_FLAG

## SEX 0 1 Sum

## M 2825 961 3786

## z\_F 3183 1192 4375

## Sum 6008 2153 8161

Probability test for SEX.

##

## 2-sample test for equality of proportions with continuity correction

##

## data: tbl[1:2, 1:2]

## X-squared = 3.5307, df = 1, p-value = 0.06024

## alternative hypothesis: two.sided

## 95 percent confidence interval:

## -0.0007561151 0.0380106016

## sample estimates:

## prop 1 prop 2

## 0.7461701 0.7275429

MSTATUS - Marital Status. In theory, married people drive more safely. There is a fairly balanced split (60/40) between married and single insureds.

## TARGET\_FLAG

## MSTATUS 0 1 Sum

## Yes 3841 1053 4894

## z\_No 2167 1100 3267

## Sum 6008 2153 8161

Probability test for MSTATUS.

##

## 2-sample test for equality of proportions with continuity correction

##

## data: tbl[1:2, 1:2]

## X-squared = 148.38, df = 1, p-value < 2.2e-16

## alternative hypothesis: two.sided

## 95 percent confidence interval:

## 0.1014053 0.1416726

## sample estimates:

## prop 1 prop 2

## 0.7848386 0.6632997

PARENT1 - Single Parent. The is a 20% difference in the calculated proportions. This difference is statistically significant.

## TARGET\_FLAG

## PARENT1 0 1 Sum

## No 5407 1677 7084

## Yes 601 476 1077

## Sum 6008 2153 8161

Probability test for PARENT1.

##

## 2-sample test for equality of proportions with continuity correction

##

## data: tbl[1:2, 1:2]

## X-squared = 201.7, df = 1, p-value < 2.2e-16

## alternative hypothesis: two.sided

## 95 percent confidence interval:

## 0.1734351 0.2370404

## sample estimates:

## prop 1 prop 2

## 0.7632693 0.5580316

CAR\_TYPE. Type of Car. We can see sports cars are having the highest proportion of accidents, and minivan have the lowest.

## TARGET\_FLAG

## CAR\_TYPE 0 1 Sum

## Minivan 1796 349 2145

## Panel Truck 498 178 676

## Pickup 946 443 1389

## Sports Car 603 304 907

## Van 549 201 750

## z\_SUV 1616 678 2294

## Sum 6008 2153 8161

Probability test for CAR\_TYPE.

##

## 6-sample test for equality of proportions without continuity correction

##

## data: tbl[1:6, 1:2]

## X-squared = 170.38, df = 5, p-value < 2.2e-16

## alternative hypothesis: two.sided

## sample estimates:

## prop 1 prop 2 prop 3 prop 4 prop 5 prop 6

## 0.8372960 0.7366864 0.6810655 0.6648291 0.7320000 0.7044464

TARGET Variables

TARGET\_FLAG - The response variable TARGET\_FLAG has a moderate imbalance, with three-quarters of the observations indicating no crashes.

## 0 1 Sum

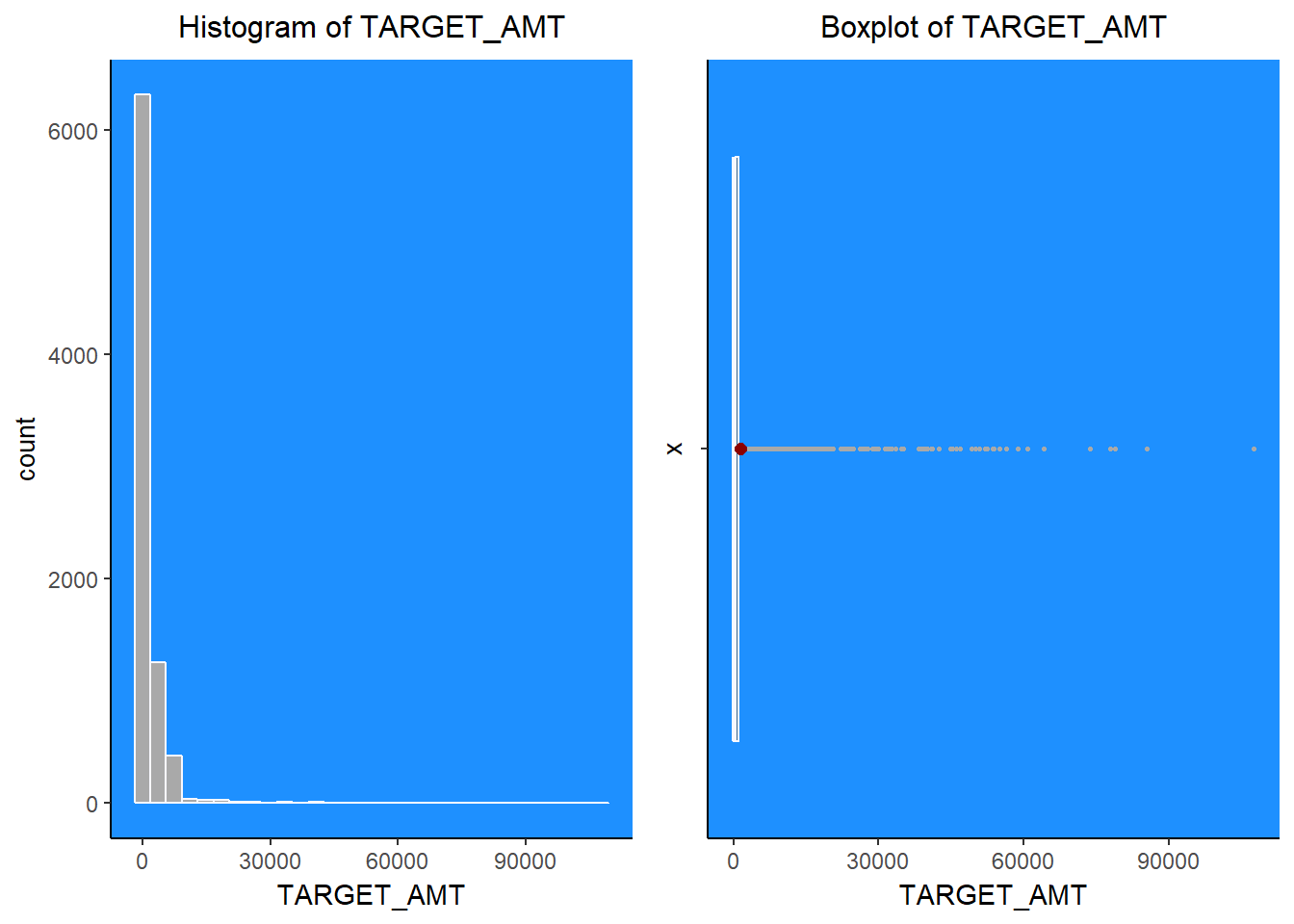
## count 6008.0 2153.0 8161

## percent 73.6 26.4 100

TARGET\_AMT - exhibits extreme, positive skewness and high kurtosis.

## Min. 1st Qu. Median Mean 3rd Qu. Max. StdD Skew Kurt

## 0.00 0.00 0.00 1504.32 1036.00 107586.14 4704.03 8.71 115.32



**Data Preparation**

Data preparation or the preprocessing is the most important part in model development. We need to remove the noise in the data so as to build a good model. We may use the transformation such as log, power transformation etc

Transformation –

*#Convert indicator variables to 0s and 1s; 1 = Yes, Male for Sex, Commercial for Car Use, Red for RED\_CAR, and Highly Urban for URBANICITY*

ins\_train$PARENT1 <- ifelse(ins\_train$PARENT1=="Yes", 1, 0)

ins\_train$MSTATUS <- ifelse(ins\_train$MSTATUS=="Yes", 1, 0)

ins\_train$SEX <- ifelse(ins\_train$SEX=="M", 1, 0)

ins\_train$CAR\_USE <- ifelse(ins\_train$CAR\_USE=="Commercial", 1, 0)

ins\_train$RED\_CAR <- ifelse(ins\_train$RED\_CAR=="Yes", 1, 0)

ins\_train$REVOKED <- ifelse(ins\_train$REVOKED=="Yes", 1, 0)

ins\_train$URBANICITY <- ifelse(ins\_train$URBANICITY == "Highly Urban/ Urban", 1, 0)

*#Convert categorical predictor values to indicator variables - EDUCATION, CAR\_TYPE, JOB*

*#EDUCATION, High school graduate is base case*

ins\_train$HSDropout <- ifelse(ins\_train$EDUCATION=="<High School", 1, 0)

ins\_train$Bachelors <- ifelse(ins\_train$EDUCATION=="Bachelors", 1, 0)

ins\_train$Masters <- ifelse(ins\_train$EDUCATION=="Masters", 1, 0)

ins\_train$PhD <- ifelse(ins\_train$EDUCATION=="PhD", 1, 0)

*#CAR\_TYPE, base case is minivan*

ins\_train$Panel\_Truck <- ifelse(ins\_train$CAR\_TYPE=="Panel Truck", 1, 0)

ins\_train$Pickup <- ifelse(ins\_train$CAR\_TYPE=="Pickup", 1, 0)

ins\_train$Sports\_Car <- ifelse(ins\_train$CAR\_TYPE=="Sports Car", 1, 0)

ins\_train$Van <- ifelse(ins\_train$CAR\_TYPE=="Van", 1, 0)

ins\_train$SUV <- ifelse(ins\_train$CAR\_TYPE=="z\_SUV", 1, 0)

*#JOB, base case is ""*

ins\_train$Professional <- ifelse(ins\_train$JOB == "Professional", 1, 0)

ins\_train$Blue\_Collar <- ifelse(ins\_train$JOB == "Professional", 1, 0)

ins\_train$Clerical <- ifelse(ins\_train$JOB == "Clerical", 1, 0)

ins\_train$Doctor <- ifelse(ins\_train$JOB == "Doctor", 1, 0)

ins\_train$Lawyer <- ifelse(ins\_train$JOB == "Lawyer", 1, 0)

ins\_train$Manager <- ifelse(ins\_train$JOB == "Manager", 1, 0)

ins\_train$Home\_Maker <- ifelse(ins\_train$JOB == "Home Maker", 1, 0)

ins\_train$Student <- ifelse(ins\_train$JOB == "Student", 1, 0)

Let’s look into the variables and see what transformation to use.

INCOME

Income is a positively skewed variable with a significant number zeroes. We will apply the square root transformation suggested by the box-cox procedure to the original variable to reduce the overall skew.

boxcoxfit(ins\_train$INCOME[ins\_train$INCOME >0])

HOME\_VAL

Home values are also moderately right skewed with a significant number of zeroes. We’ll apply a quarter root transformation to the original variable to reduce the overall skew.

ins\_train$HOME\_VAL\_MOD <- ins\_train$HOME\_VAL^0.113

BLUEBOOK

The BLUEBOOK variable has a moderate right skew. We’ll apply the square root transformation suggested by the box-cox procedure.

ins\_train$BLUEBOOK\_MOD <- ins\_train$BLUEBOOK^0.461

OLDCLAIM

OLDCLAIM is extremely right skewed. We’ll apply a log(x+1) transformation to reduce the overall skew.

ins\_train$OLD\_CLAIM\_MOD <- log(ins\_train$OLDCLAIM + 1)

**Build Models**

We will build 4 models and see which one is a good fit model.

**Model 1 - All original variables model**

 In this model we will use all the variables. This can be our base model and this model will not include any transformations. We can see which variables are significant. This will help us in looking at the P-Values and removing the non-significant variables.

model1 <- glm(target ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + black + lstat + medv , family="binomial", data=crime\_train)

summary(model1)

##

## Call:

## glm(formula = target ~ zn + indus + chas + nox + rm + age + dis +

## rad + tax + ptratio + black + lstat + medv, family = "binomial",

## data = crime\_train)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.2854 -0.1372 -0.0017 0.0020 3.4721

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -36.839521 7.028726 -5.241 1.59e-07 \*\*\*

## zn -0.061720 0.034410 -1.794 0.072868 .

## indus -0.072580 0.048546 -1.495 0.134894

## chas 1.032352 0.759627 1.359 0.174139

## nox 50.159513 8.049503 6.231 4.62e-10 \*\*\*

## rm -0.692145 0.741431 -0.934 0.350548

## age 0.034522 0.013883 2.487 0.012895 \*

## dis 0.765795 0.234407 3.267 0.001087 \*\*

## rad 0.663015 0.165135 4.015 5.94e-05 \*\*\*

## tax -0.006593 0.003064 -2.152 0.031422 \*

## ptratio 0.442217 0.132234 3.344 0.000825 \*\*\*

## black -0.013094 0.006680 -1.960 0.049974 \*

## lstat 0.047571 0.054508 0.873 0.382802

## medv 0.199734 0.071022 2.812 0.004919 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 645.88 on 465 degrees of freedom

## Residual deviance: 186.15 on 452 degrees of freedom

## AIC: 214.15

##

## Number of Fisher Scoring iterations: 9

**Model 2: - All significant original variables model.**

I came up with this models after analyzing the output of model1. I removed all the variables that are not significant after seeing their P-Value.

model2 <- glm(target ~ nox + age + dis + rad + tax + ptratio + black + medv , family="binomial", data=crime\_train)

summary(model2)

##

## Call:

## glm(formula = target ~ nox + age + dis + rad + tax + ptratio +

## black + medv, family = "binomial", data = crime\_train)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.42422 -0.19292 -0.01400 0.00279 3.06740

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -32.301655 6.382694 -5.061 4.17e-07 \*\*\*

## nox 42.160350 6.674149 6.317 2.67e-10 \*\*\*

## age 0.031017 0.010681 2.904 0.003684 \*\*

## dis 0.437803 0.172533 2.538 0.011165 \*

## rad 0.703446 0.140296 5.014 5.33e-07 \*\*\*

## tax -0.008744 0.002611 -3.348 0.000813 \*\*\*

## ptratio 0.395580 0.112482 3.517 0.000437 \*\*\*

## black -0.012490 0.006760 -1.848 0.064662 .

## medv 0.101177 0.034116 2.966 0.003020 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 645.88 on 465 degrees of freedom

## Residual deviance: 198.28 on 457 degrees of freedom

## AIC: 216.28

##

## Number of Fisher Scoring iterations: 9

**Model 3: - All variables with transformations(will keep variables that were not transformed)**

Model 3 includes original variables, plus the transformed variables from the transformations like power transformation and log transformations. This transformation should help in reducing the skewness in the data or help them to become more normalized. This will help us in looking at the P-Values and removing the non-significant variables.

model3 <- glm(target ~ zn + indus + chas + nox + rm\_mod + age\_mod + dis\_mod + rad + tax + ptratio + black + lstat\_mod + medv\_mod , family="binomial", data=crime\_train)

summary(model3)

##

## Call:

## glm(formula = target ~ zn + indus + chas + nox + rm\_mod + age\_mod +

## dis\_mod + rad + tax + ptratio + black + lstat\_mod + medv\_mod,

## family = "binomial", data = crime\_train)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.4018 -0.1416 -0.0029 0.0032 3.4233

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -42.515655 17.813038 -2.387 0.016997 \*

## zn -0.037515 0.029842 -1.257 0.208703

## indus -0.051749 0.049379 -1.048 0.294636

## chas 0.970813 0.768970 1.262 0.206774

## nox 54.149495 8.472349 6.391 1.64e-10 \*\*\*

## rm\_mod -15.802136 12.885763 -1.226 0.220076

## age\_mod 0.010277 0.003204 3.208 0.001336 \*\*

## dis\_mod 3.824093 0.986732 3.876 0.000106 \*\*\*

## rad 0.634929 0.164849 3.852 0.000117 \*\*\*

## tax -0.004892 0.003173 -1.542 0.123132

## ptratio 0.500107 0.141497 3.534 0.000409 \*\*\*

## black -0.013934 0.007189 -1.938 0.052588 .

## lstat\_mod 0.363908 1.824782 0.199 0.841930

## medv\_mod 11.900134 4.008860 2.968 0.002993 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 645.88 on 465 degrees of freedom

## Residual deviance: 182.76 on 452 degrees of freedom

## AIC: 210.76

##

## Number of Fisher Scoring iterations: 9

**Model 4: - Only the significant variables from model3 are used in this model.**

I removed all the variables that are not significant after seeing their P-Value.

model4 <- glm(target ~ nox + age\_mod + dis\_mod + rad + ptratio + medv\_mod , family="binomial", data=crime\_train)

summary(model4)

##

## Call:

## glm(formula = target ~ nox + age\_mod + dis\_mod + rad + ptratio +

## medv\_mod, family = "binomial", data = crime\_train)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -1.8866 -0.2127 -0.0217 0.0064 3.2168

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -57.955389 9.233814 -6.276 3.46e-10 \*\*\*

## nox 46.172648 7.022160 6.575 4.86e-11 \*\*\*

## age\_mod 0.009192 0.002487 3.695 0.000220 \*\*\*

## dis\_mod 3.488834 0.807859 4.319 1.57e-05 \*\*\*

## rad 0.529064 0.123587 4.281 1.86e-05 \*\*\*

## ptratio 0.398295 0.110069 3.619 0.000296 \*\*\*

## medv\_mod 7.928413 1.927376 4.114 3.90e-05 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 645.88 on 465 degrees of freedom

## Residual deviance: 203.43 on 459 degrees of freedom

## AIC: 217.43

##

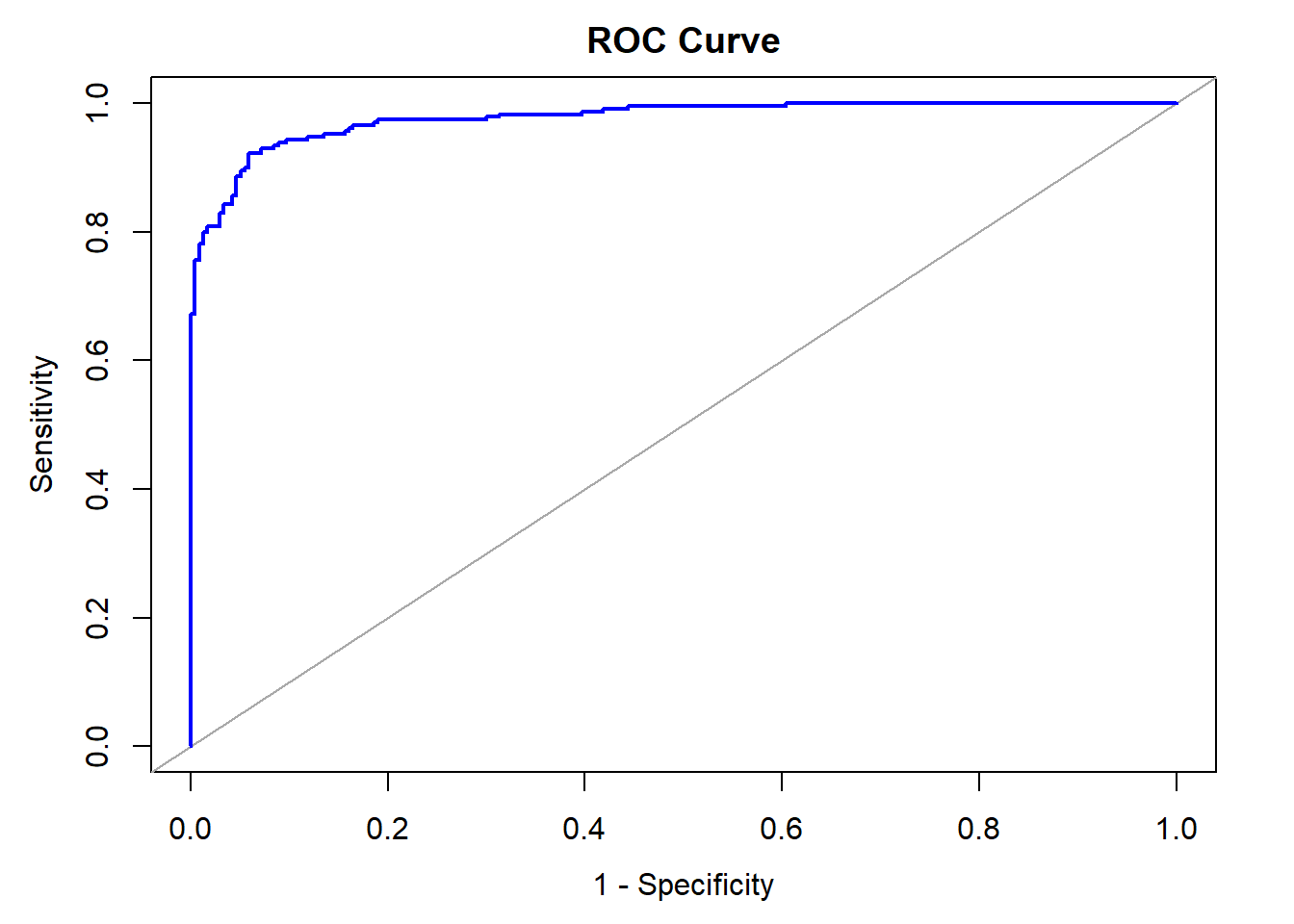
## Number of Fisher Scoring iterations: 9

# Model Selection.

I would like to select Model3. The AIC and residual deviance for this model seemed to give the best values that would be suited for the prediction. Below is the ROC curve for model3 and to me it looks good. So i would like to proceed with model3.

**Validating the model:**

I would like to validate the model using some techniques such as ROC curve, confusion Matrix as see the Accuracy, CER, Precision, Sensitivity, Specificity and F1 Score.



Area under the curve: 0.9766

Now let’s do the confusion matrix:

crime\_train$predict\_target <- ifelse(crime\_train$predict >=0.5, 1, 0)

crime\_train$predict\_target <- as.integer(crime\_train$predict\_target)

myvars <- c("target", "predict\_target")

crime\_train\_cm <- crime\_train[myvars]

cm <- table(crime\_train\_cm$predict\_target,crime\_train\_cm$target)

knitr:: kable(cm)

## PredictedValue

## ActualValue FALSE TRUE

## 0 221 18

## 1 16 211

Accuracy : 0.9270386

Classification Error Rate: 0.07296137

Precision: 0.9213974

Sensitivity: 0.9295154

Specificity: 0.9246862

F1 Score: 0.9254386

## **Testing the evaluation data with mode 3**

In this final step we will be testing the evaluation data using model3. We need to first pre-preprocess the data in the exact similar way as we did for train data. The Predicted Evaluation data is present at https://github.com/Riteshlohiya/Data621-Week3-Assignment3/blob/master/Evaluation\_Data.csv

crime\_eval <- read.csv("https://raw.githubusercontent.com/Riteshlohiya/Data621-Week3-Assignment3/master/crime-evaluation-data.csv")

crime\_eval$age\_mod <- crime\_eval$age^1.3

crime\_eval$lstat\_mod <- crime\_eval$lstat^0.23

crime\_eval$dis\_mod <- log(crime\_eval$dis)

crime\_eval$medv\_mod <- crime\_eval$medv^0.23

crime\_eval$rm\_mod <- crime\_eval$rm^0.23

crime\_eval$predict\_prob <- predict(model3, crime\_eval, type='response')

crime\_eval$predict\_target <- ifelse(crime\_eval$predict\_prob >= 0.50, 1,0)

write.csv(crime\_eval,"Evaluation\_Data.csv", row.names=FALSE)

**Appendix**

---

title: "Data621 Assignment 4"

author: "Ritesh Lohiya"

date: "July 6, 2018"

output: html\_document

---

Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, TARGET\_FLAG, is a 1 or a 0. A â1â means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is TARGET\_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero. Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

#install.packages('pander')

```{r}

library(readr)

library(kableExtra)

library(tidyverse)

library(knitr)

library(psych)

library(gridExtra)

library(usdm)

library(mice)

library(ggiraph)

library(cowplot)

library(reshape2)

library(corrgram)

library(caTools)

library(caret)

library(ROCR)

library(pROC)

library(reshape2)

library(Amelia)

library(qqplotr)

library(moments)

library(car)

library(MASS)

library(geoR)

library(pander)

```

#DATA EXPLORATION:

The dataset of interest contains information about customers of an auto insurance company. The dataset has 8161 rows (each representing a customer) and 25 variables. There are 23 predictor variables and 2 response variables: TARGET\_FLAG, a binary categorical variable representing whether each customer has been in an accident; and TARGET\_AMT, a numerical variable indicating the cost of a crash that a customer was in.

```{r}

ins\_train <- read.csv("https://raw.githubusercontent.com/Riteshlohiya/Data621-Assignment-4/master/insurance\_training\_data.csv")

summary(ins\_train)

var\_class <- data.frame(Class = rep(NA, ncol(ins\_train) - 1), Levels = rep(NA, ncol(ins\_train) - 1), stringsAsFactors = FALSE, check.names = FALSE, row.names = names(ins\_train)[-1])

for(i in 2:ncol(ins\_train)) {

var\_class[i - 1, 1] <- class(ins\_train[, i])

var\_class[i - 1, 2] <- ifelse(length(levels(ins\_train[, i])) == 0, '-', length(levels(ins\_train[, i])))

}

pander(var\_class)

```

INCOME, HOME\_VAL, BLUEBOOK, and OLDCLAIM are represented as strings. So we will be extracting the numeric values for these.

```{r}

ins\_train$INCOME <- as.numeric(str\_replace\_all(ins\_train$INCOME, "[[:punct:]\\$]",""))

ins\_train$HOME\_VAL <- as.numeric(str\_replace\_all(ins\_train$HOME\_VAL, "[[:punct:]\\$]",""))

ins\_train$BLUEBOOK <- as.numeric(str\_replace\_all(ins\_train$BLUEBOOK, "[[:punct:]\\$]",""))

ins\_train$OLDCLAIM <- as.numeric(str\_replace\_all(ins\_train$OLDCLAIM, "[[:punct:]\\$]",""))

```

Visual Exploration:

Boxplots are generated for non-binary variables split by TARGET\_FLAG:

```{r}

numeric <- ins\_train %>% dplyr::select(c(TARGET\_FLAG, TARGET\_AMT, KIDSDRIV, AGE, HOMEKIDS, YOJ, INCOME, HOME\_VAL, TRAVTIME, BLUEBOOK, TIF, OLDCLAIM, CLM\_FREQ, MVR\_PTS, CAR\_AGE))

numeric <- melt(numeric, id.vars="TARGET\_FLAG")

numeric$TARGET\_FLAG <- factor(numeric$TARGET\_FLAG)

ggplot(numeric, aes(TARGET\_FLAG, value)) + geom\_boxplot(aes(fill = TARGET\_FLAG), alpha = 0.5) + facet\_wrap(~variable, scale="free") + scale\_fill\_discrete(guide = FALSE) + scale\_y\_continuous('', labels = NULL, breaks = NULL) + scale\_x\_discrete('') + ggtitle("Distribution of Predictors by TARGET\_FLAG\n")

```

Now lets see the correlations:

```{r}

pairs(~MVR\_PTS+CLM\_FREQ+URBANICITY+HOME\_VAL+PARENT1+CAR\_USE+OLDCLAIM, data=ins\_train, main="Predictors with High Correlattions to Targets", col="slategrey")

```

Now we will see the missing values in the dataset. For this i have used Amelia package. We can see there are missing values for CAR\_AGE, HOME\_VAL, YOJ and INCOME. There needs to be taken care while we do data preparation.

```{r}

missmap(ins\_train, main = "Missing values vs observed", color='dodgerblue')

```

Now lets do some plots to understand the data:

AGE - Age of Driver. Very young people tend to be risky. Maybe very old people also. We note six missing values that we'll need to address later.

The distribution of AGE is almost perfectly normal. When we break out the data by TARGET\_FLAG values, the distributions of age by TARGET\_FLAG are still roughly normal.

```{r}

with(ins\_train, c(summary(AGE), SD=sd(AGE), Skew=skewness(AGE), Kurt=kurtosis(AGE)))

hist <- ggplot(ins\_train, aes(AGE)) + geom\_histogram(fill = 'dodgerblue', binwidth = 10, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of AGE') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=AGE)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of AGE") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", AGE)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of AGE', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), AGE)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of AGE by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

BLUEBOOK - Value of Vehicle. Unknown effect on probability of collision, but probably effect the payout if there is a crash. Individuals involved in crashes have a higher proportion of low BLUEBOOK values.

```{r}

with(ins\_train, c(summary(BLUEBOOK), SD=sd(BLUEBOOK), Skew=skewness(BLUEBOOK), Kurt=kurtosis(BLUEBOOK)))

hist <- ggplot(ins\_train, aes(BLUEBOOK)) + geom\_histogram(fill = 'dodgerblue', binwidth = 10000, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of BLUEBOOK') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=BLUEBOOK)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of BLUEBOOK") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", BLUEBOOK)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of BLUEBOOK', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), BLUEBOOK)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of BLUEBOOK by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

CAR\_AGE - Vehicle Age. We could see there is one negative value for CAR\_AGE. We have to treat this value in our data preparation step.

```{r}

with(ins\_train, c(summary(CAR\_AGE), SD=sd(CAR\_AGE), Skew=skewness(CAR\_AGE), Kurt=kurtosis(CAR\_AGE)))

hist <- ggplot(ins\_train, aes(CAR\_AGE)) + geom\_histogram(fill = 'dodgerblue', binwidth = 5, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of CAR\_AGE') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=CAR\_AGE)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of CAR\_AGE") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", CAR\_AGE)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of CAR\_AGE', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), CAR\_AGE)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of CAR\_AGE by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

CLM\_FREQ - # Claims (Past 5 Years). The more claims you filed in the past, the more you are likely to file in the future. We can see that this variable is also skewed.

```{r}

with(ins\_train, c(summary(CLM\_FREQ), SD=sd(CLM\_FREQ), Skew=skewness(CLM\_FREQ), Kurt=kurtosis(CLM\_FREQ)))

hist <- ggplot(ins\_train, aes(CLM\_FREQ)) + geom\_histogram(fill = 'dodgerblue', binwidth = 1, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of CLM\_FREQ') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=CLM\_FREQ)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of CLM\_FREQ") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", CLM\_FREQ)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of CLM\_FREQ', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), CLM\_FREQ)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of CLM\_FREQ by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

HOMEKIDS - # Children at Home. HOMEKIDS does not seem to impact the TARGET\_FLAG. The distribution of this discrete variable is right skewed.

```{r}

with(ins\_train, c(summary(HOMEKIDS), SD=sd(HOMEKIDS), Skew=skewness(HOMEKIDS), Kurt=kurtosis(HOMEKIDS)))

hist <- ggplot(ins\_train, aes(HOMEKIDS)) + geom\_histogram(fill = 'dodgerblue', binwidth = 1, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of HOMEKIDS') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=HOMEKIDS)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of HOMEKIDS") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", HOMEKIDS)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of HOMEKIDS', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), HOMEKIDS)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of HOMEKIDS by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

HOME\_VAL - Home Value. Home owners tend to drive more responsibly. The distribution of HOME\_VAL is right skewed and also we can there are some missing values.

```{r}

with(ins\_train, c(summary(HOME\_VAL), SD=sd(HOME\_VAL), Skew=skewness(HOME\_VAL), Kurt=kurtosis(HOME\_VAL)))

hist <- ggplot(ins\_train, aes(HOME\_VAL)) + geom\_histogram(fill = 'dodgerblue', binwidth = 100000, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of HOME\_VAL') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=HOME\_VAL)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of HOME\_VAL") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", HOME\_VAL)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of HOME\_VAL', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), HOME\_VAL)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of HOME\_VAL by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

INCOME - Income of the person. Rich people tend to get into fewer crashes. The distribution of INCOME is right skewed, with a significant number of observations indicating $0 in income. There are some missing values in this aswell.

```{r}

with(ins\_train, c(summary(INCOME), SD=sd(INCOME), Skew=skewness(INCOME), Kurt=kurtosis(INCOME)))

hist <- ggplot(ins\_train, aes(INCOME)) + geom\_histogram(fill = 'dodgerblue', binwidth = 10000, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of INCOME') + theme(plot.title = element\_text(hjust = 1))

qq\_plot <- ggplot(ins\_train, aes(sample=INCOME)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of INCOME") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", INCOME)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of INCOME', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), INCOME)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of INCOME by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

KIDSDRIV - # Driving Children. When teenagers drive your car, you are more likely to get into crashes. The discrete variable KIDSDRIV is right skewed

```{r}

with(ins\_train, c(summary(KIDSDRIV), SD=sd(KIDSDRIV), Skew=skewness(KIDSDRIV), Kurt=kurtosis(KIDSDRIV)))

hist <- ggplot(ins\_train, aes(KIDSDRIV)) + geom\_histogram(fill = 'dodgerblue', binwidth = 1, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of KIDSDRIV') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=KIDSDRIV)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of KIDSDRIV") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", KIDSDRIV)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of KIDSDRIV', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), KIDSDRIV)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of KIDSDRIV by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

MVR\_PTS - Motor Vehicle Record Points. If you get lots of traffic tickets, you tend to get into more crashes. MVR\_PTS is positively skewed.

```{r}

with(ins\_train, c(summary(MVR\_PTS), SD=sd(MVR\_PTS), Skew=skewness(MVR\_PTS), Kurt=kurtosis(MVR\_PTS)))

hist <- ggplot(ins\_train, aes(MVR\_PTS)) + geom\_histogram(fill = 'dodgerblue', binwidth = 1, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of MVR\_PTS') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=MVR\_PTS)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of MVR\_PTS") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", MVR\_PTS)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of MVR\_PTS', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), MVR\_PTS)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of MVR\_PTS by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

OLDCLAIM - Total Claims (Past 5 Years). If your total payout over the past five years was high, this suggests future payouts will be high. The distribution of OLDCLAIM is extremely right skewed.

```{r}

with(ins\_train, c(summary(OLDCLAIM), SD=sd(OLDCLAIM), Skew=skewness(OLDCLAIM), Kurt=kurtosis(OLDCLAIM)))

hist <- ggplot(ins\_train, aes(OLDCLAIM)) + geom\_histogram(fill = 'dodgerblue', binwidth = 10000, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of OLDCLAIM') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=OLDCLAIM)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of OLDCLAIM") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", OLDCLAIM)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of OLDCLAIM', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), OLDCLAIM)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of OLDCLAIM by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

TIF - Time in Force. People who have been customers for a long time are usually more safe. The distribution is somewhat positively skewed.

```{r}

with(ins\_train, c(summary(TIF), SD=sd(TIF), Skew=skewness(TIF), Kurt=kurtosis(TIF)))

hist <- ggplot(ins\_train, aes(TIF)) + geom\_histogram(fill = 'dodgerblue', binwidth = 1, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of TIF') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=TIF)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of TIF") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", TIF)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of TIF', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), TIF)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of TIF by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

TRAVTIME - Distance to Work. Long drives to work usually suggest greater risk. The distribution has a slight positive skew. The subset of insureds with no accidents have a higher proportion of individuals with short commute times.

```{r}

with(ins\_train, c(summary(TRAVTIME), SD=sd(TRAVTIME), Skew=skewness(TRAVTIME), Kurt=kurtosis(TRAVTIME)))

hist <- ggplot(ins\_train, aes(TRAVTIME)) + geom\_histogram(fill = 'dodgerblue', binwidth = 10, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of TRAVTIME') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=TRAVTIME)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of TRAVTIME") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", TRAVTIME)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of TRAVTIME', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), TRAVTIME)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of TRAVTIME by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

YOJ - Years on Job. People who stay at a job for a long time are usually more safe. The variable would be approximately normally distributed if it weren't for the high percentage of individuals with less than one year on the job.

```{r}

with(ins\_train, c(summary(YOJ), SD=sd(YOJ), Skew=skewness(YOJ), Kurt=kurtosis(YOJ)))

hist <- ggplot(ins\_train, aes(YOJ)) + geom\_histogram(fill = 'dodgerblue', binwidth = 5, color = 'darkgray' ) +

theme\_classic() + labs(title = 'Histogram of YOJ') + theme(plot.title = element\_text(hjust = 0.5))

qq\_plot <- ggplot(ins\_train, aes(sample=YOJ)) + stat\_qq\_point(color='dodgerblue') + stat\_qq\_line(color='darkgray') +

labs(x="Thoretical Quantiles", y="Sample Quantiles", title = "QQ Plot of YOJ") + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

box\_plot <- ggplot(ins\_train, aes(x="", YOJ)) + geom\_boxplot(fill='dodgerblue', color='darkgray')+ theme\_classic() +

labs(title = 'Boxplot of YOJ', x="") + theme(plot.title = element\_text(hjust = 0.5)) + coord\_flip()

box\_target <- ggplot(ins\_train, aes(x=factor(TARGET\_FLAG), YOJ)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

labs(x='target', title = 'Boxplot of YOJ by TARGET\_FLAG') + theme\_classic() +

theme(plot.title = element\_text(hjust = 0.5))

grid.arrange(hist, qq\_plot, box\_plot, box\_target, ncol=2)

```

EDUCATION - Unknown effect, but in theory more educated people tend to drive more safely.

```{r}

options(width=100)

tbl <- with(ins\_train, rbind(addmargins(table(EDUCATION)),addmargins(prop.table(table(EDUCATION)))\*100))

row.names(tbl) <- c('count','percent')

round(tbl,1)

```

REVOKED - License Revoked (Past 7 Years). If your license was revoked in the past 7 years, you probably are a more risky driver. Only 12% of drivers in the training data have a former license suspension on record.

```{r}

tbl <- addmargins(table(REVOKED=ins\_train$REVOKED,TARGET\_FLAG=ins\_train$TARGET\_FLAG))

tbl

```

RED\_CAR - A Red Car. Urban legend says that red cars (especially red sports cars) are more risky. Is that true?. 30% of vehicles in the red category.

```{r}

tbl <- addmargins(table(RED\_CAR=ins\_train$RED\_CAR,TARGET\_FLAG=ins\_train$TARGET\_FLAG))

tbl

```

CAR\_USE - Vehicle Use. Commercial vehicles are driven more, so might increase probability of collision. 60% car usage is private.

```{r}

tbl <- addmargins(table(CAR\_USE=ins\_train$CAR\_USE,TARGET\_FLAG=ins\_train$TARGET\_FLAG))

tbl

```

SEX - Gender. Urban legend says that women have less crashes then men. Is that true?. The split between males and females is split almost 50/50.

```{r}

tbl <- addmargins(table(SEX=ins\_train$SEX,TARGET\_FLAG=ins\_train$TARGET\_FLAG))

tbl

round(prop.table(tbl[1:2,1:2], margin=1),2)

prop.test(tbl[1:2,1:2])

```

MSTATUS - Marital Status. In theory, married people drive more safely. There is a fairly balanced split (60/40) between married and single insureds.

```{r}

tbl <- addmargins(table(MSTATUS=ins\_train$MSTATUS,TARGET\_FLAG=ins\_train$TARGET\_FLAG))

tbl

round(prop.table(tbl[1:2,1:2], margin=1),2)

prop.test(tbl[1:2,1:2])

```

PARENT1 - Single Parent. The is a 20% difference in the calculated proportions. This difference is statistically significant:

```{r}

tbl <- addmargins(table(PARENT1=ins\_train$PARENT1,TARGET\_FLAG=ins\_train$TARGET\_FLAG))

tbl

round(prop.table(tbl[1:2,1:2], margin=1),2)

prop.test(tbl[1:2,1:2])

```

CAR\_TYPE. Type of Car. We can see sports cars are having the highest proportion of accidents, and minivan have the lowest.

```{r}

tbl <- with(ins\_train, addmargins(table(CAR\_TYPE, TARGET\_FLAG)))

tbl

pt <- round(prop.table(tbl[1:6,1:2], margin=1),2)

pt

prop.test(tbl[1:6,1:2])

```

TARGET Variables

TARGET\_FLAG - The response variable TARGET\_FLAG has a moderate imbalance, with three-quarters of the observations indicating no crashes.

```{r}

tbl <- with(ins\_train,rbind(round(addmargins(table(TARGET\_FLAG)),0),

addmargins(prop.table(table(TARGET\_FLAG)))\*100))

row.names(tbl) <- c('count','percent')

round(tbl,1)

```

TARGET\_AMT - exhibits extreme, positive skewness and high kurtosis.

```{r}

options(width=100)

round(with(ins\_train, c(summary(TARGET\_AMT), StdD=sd(TARGET\_AMT), Skew=skewness(TARGET\_AMT), Kurt=kurtosis(TARGET\_AMT))),2)

```

```{r}

h <- ggplot(ins\_train, aes(TARGET\_AMT)) +

geom\_histogram(color="ghostwhite", fill="darkgrey") +

theme\_classic()+ labs(title = 'Histogram of TARGET\_AMT') +

theme(plot.title = element\_text(hjust = 0.5),axis.title.y=element\_text(size=10)) +

theme(legend.position = c(1,1),legend.justification = c(1,1), legend.background = element\_rect(fill='dodgerblue')) +

scale\_fill\_manual("TARGET\_FLAG",values=c("dodgerblue","dodgerblue")) +

theme(plot.title = element\_text(size=12),legend.title=element\_text(size=8),

legend.text=element\_text(size=7),panel.background = element\_rect(fill = "dodgerblue"))

b <- ggplot(ins\_train, aes(x="",y=TARGET\_AMT)) +

geom\_boxplot(color="ghostwhite", fill="steelblue4",outlier.color="darkgrey", outlier.size = 0.5) +

theme\_classic()+ labs(title = 'Boxplot of TARGET\_AMT') +

theme(plot.title = element\_text(hjust = 0.5),axis.title.y=element\_text(size=10)) +

theme(legend.position = c(1,1),legend.justification = c(1,1), legend.background = element\_rect(fill='dodgerblue')) +

scale\_fill\_manual("TARGET\_FLAG",values=c("dodgerblue","dodgerblue")) +

theme(plot.title = element\_text(size=12),legend.title=element\_text(size=8),

legend.text=element\_text(size=7),panel.background = element\_rect(fill = "dodgerblue")) + coord\_flip() +

stat\_summary(fun.y=mean, colour="darkred", geom="point", shape=16, size=2)

grid.arrange(h,b, ncol=2)

```

#DATA PREPARATION:

There are 7 variables that have only 2 values, so we can make them binary.

PARENT1 - Convert yes to 1

MSTATUS - Convert yes to 1

RED\_CAR - Convert yes to 1

REVOKED - Convert yes to 1

SEX - Convert male to 1

CAR\_USE - Convert Commercial to 1

URBANICITY: Conver Highly Urban/ Urban to 1

```{r}

#Convert indicator variables to 0s and 1s; 1 = Yes, Male for Sex, Commercial for Car Use, Red for RED\_CAR, and Highly Urban for URBANICITY

ins\_train$PARENT1 <- ifelse(ins\_train$PARENT1=="Yes", 1, 0)

ins\_train$MSTATUS <- ifelse(ins\_train$MSTATUS=="Yes", 1, 0)

ins\_train$SEX <- ifelse(ins\_train$SEX=="M", 1, 0)

ins\_train$CAR\_USE <- ifelse(ins\_train$CAR\_USE=="Commercial", 1, 0)

ins\_train$RED\_CAR <- ifelse(ins\_train$RED\_CAR=="Yes", 1, 0)

ins\_train$REVOKED <- ifelse(ins\_train$REVOKED=="Yes", 1, 0)

ins\_train$URBANICITY <- ifelse(ins\_train$URBANICITY == "Highly Urban/ Urban", 1, 0)

#Convert categorical predictor values to indicator variables - EDUCATION, CAR\_TYPE, JOB

#EDUCATION, High school graduate is base case

ins\_train$HSDropout <- ifelse(ins\_train$EDUCATION=="<High School", 1, 0)

ins\_train$Bachelors <- ifelse(ins\_train$EDUCATION=="Bachelors", 1, 0)

ins\_train$Masters <- ifelse(ins\_train$EDUCATION=="Masters", 1, 0)

ins\_train$PhD <- ifelse(ins\_train$EDUCATION=="PhD", 1, 0)

#CAR\_TYPE, base case is minivan

ins\_train$Panel\_Truck <- ifelse(ins\_train$CAR\_TYPE=="Panel Truck", 1, 0)

ins\_train$Pickup <- ifelse(ins\_train$CAR\_TYPE=="Pickup", 1, 0)

ins\_train$Sports\_Car <- ifelse(ins\_train$CAR\_TYPE=="Sports Car", 1, 0)

ins\_train$Van <- ifelse(ins\_train$CAR\_TYPE=="Van", 1, 0)

ins\_train$SUV <- ifelse(ins\_train$CAR\_TYPE=="z\_SUV", 1, 0)

#JOB, base case is ""

ins\_train$Professional <- ifelse(ins\_train$JOB == "Professional", 1, 0)

ins\_train$Blue\_Collar <- ifelse(ins\_train$JOB == "Professional", 1, 0)

ins\_train$Clerical <- ifelse(ins\_train$JOB == "Clerical", 1, 0)

ins\_train$Doctor <- ifelse(ins\_train$JOB == "Doctor", 1, 0)

ins\_train$Lawyer <- ifelse(ins\_train$JOB == "Lawyer", 1, 0)

ins\_train$Manager <- ifelse(ins\_train$JOB == "Manager", 1, 0)

ins\_train$Home\_Maker <- ifelse(ins\_train$JOB == "Home Maker", 1, 0)

ins\_train$Student <- ifelse(ins\_train$JOB == "Student", 1, 0)

```

Missing/ Error Values treatment:

Due to the skewness illustrated by some of the variables with missing data, the median is used to avoid any bias introduced into the mean by the skewness of these variables' distribution.

```{r}

ins\_train$CAR\_AGE[ins\_train$CAR\_AGE == -3] <- NA

ins\_train <- ins\_train %>% dplyr::select(-c(INDEX,EDUCATION,CAR\_TYPE,JOB))

fillwithmedian <- function(x) {

median\_val = median(x, na.rm = TRUE)

x[is.na(x)] = median\_val

return(x)

}

ins\_train <- data.frame(lapply(ins\_train, fillwithmedian))

```

Lets look into the variables and see what transformation to use.

INCOME

Income is a positively skewed variable with a significant number zeroes. We will apply the square root transformation suggested by the box-cox procedure to the original variable to reduce the overall skew.

```{r}

boxcoxfit(ins\_train$INCOME[ins\_train$INCOME >0])

ins\_train$INCOME\_MOD <- ins\_train$INCOME ^0.433

```

HOME\_VAL

Home values are also moderately right skewed with a significant number of zeroes. We'll apply a quarter root transformation to the original variable to reduce the overall skew.

```{r}

boxcoxfit(ins\_train$HOME\_VAL[ins\_train$HOME\_VAL > 0])

ins\_train$HOME\_VAL\_MOD <- ins\_train$HOME\_VAL^0.113

```

BLUEBOOK

The BLUEBOOK variable has a moderate right skew. We'll apply the square root transformation suggested by the box-cox procedure.

```{r}

boxcoxfit(ins\_train$BLUEBOOK)

ins\_train$BLUEBOOK\_MOD <- ins\_train$BLUEBOOK^0.461

```

OLDCLAIM

OLDCLAIM is extremely right skewed. We'll apply a log(x+1) transformation to reduce the overall skew.

```{r}

boxcoxfit(ins\_train$OLDCLAIM[ins\_train$OLDCLAIM>0])

ins\_train$OLD\_CLAIM\_MOD <- log(ins\_train$OLDCLAIM + 1)

```

#BUILD MODELS:

1. Multiple linear regression models:

Model 1 - : In this model we will use all the variables. This can be our base model.We can see which variables are significant. This will help us in looking at the P-Values and removing the non significant variables.

```{r}

train\_amount <- ins\_train[,-c(1)] #Training dataset with response of claim amount

amount\_full\_model1 <- lm(TARGET\_AMT ~., data = train\_amount)

summary(amount\_full\_model1)

```

Model 2 - Reduced model- I came up with this models after analyzing the output of model1. I removed all the variables that are not significant after seeing their P-Value.

```{r}

amount\_reduced\_model2 <- update(amount\_full\_model1, .~.-HSDropout-Home\_Maker-Bachelors-Masters-PhD-Panel\_Truck-Blue\_Collar-Professional-Student-HOMEKIDS-CAR\_AGE-YOJ-Lawyer-SEX-AGE-Doctor-Clerical-INCOME-HOME\_VAL-BLUEBOOK-RED\_CAR--CLM\_FREQ-INCOME\_MOD-HOME\_VAL\_MOD-BLUEBOOK\_MOD-OLD\_CLAIM\_MOD-OLDCLAIM)

summary(amount\_reduced\_model2)

```

Interpretation of the Model1:

The Residual standard error is 4545

Multiple R-squared: 0.07105

Adjusted R-squared: 0.06659

F-statistic: 15.93 on 39 and 8121 DF

p-value: < 2.2e-16

Analysis of plot on residuals to verify normal distribution of residuals

```{r}

sresid <- studres(amount\_full\_model1)

hist(sresid, freq=FALSE,

main="Distribution of Residuals")

xfit<-seq(min(sresid),max(sresid),length=40)

yfit<-dnorm(xfit)

lines(xfit, yfit)

```

Check for Homoscedasticity:

```{r}

ncvTest(amount\_full\_model1)

spreadLevelPlot(amount\_full\_model1)

```

Interpretation of the Model2:

The Residual standard error is 4556

Multiple R-squared: 0.06366

Adjusted R-squared: 0.06194

F-statistic: 36.92 on 15 and 8145 DF

p-value: < 2.2e-16

Analysis of plot on residuals to verify normal distribution of residuals

```{r}

sresid <- studres(amount\_reduced\_model2)

hist(sresid, freq=FALSE,

main="Distribution of Residuals")

xfit<-seq(min(sresid),max(sresid),length=40)

yfit<-dnorm(xfit)

lines(xfit, yfit)

```

Check for Homoscedasticity:

```{r}

ncvTest(amount\_reduced\_model2)

spreadLevelPlot(amount\_reduced\_model2)

```

2. Binary Logistic Regression models:

Model 3: Base Model: All variables without transformation.

All of the variables will be tested to determine the base model they provided. This will allow us to see which variables are significant in our dataset, and allow us to make other models based on that.

```{r}

train\_flag <- ins\_train[,-c(2)] #Training dataset with response of crash

flagfull <- glm(TARGET\_FLAG ~.-INCOME\_MOD-HOME\_VAL\_MOD-BLUEBOOK\_MOD-OLD\_CLAIM\_MOD, data = train\_flag, family = binomial(link='logit'))

summary(flagfull)

```

Model 4: We will now add the transformed data to the model.

```{r}

train\_flag <- ins\_train[,-c(2)] #Training dataset with response of crash

flagfull\_mod <- glm(TARGET\_FLAG ~., data = train\_flag, family = binomial(link='logit'))

summary(flagfull\_mod)

```

Model5: We will only keep only the significant variables for our reduced model3.

```{r}

train\_flag <- ins\_train[,-c(2)] #Training dataset with response of crash

flag\_reduced\_mod <- glm(TARGET\_FLAG ~.-AGE-HOMEKIDS-YOJ-INCOME-HOME\_VAL-SEX-RED\_CAR-CLM\_FREQ-CAR\_AGE-HSDropout-Professional-Blue\_Collar-Clerical-Lawyer-Home\_Maker-HOME\_VAL\_MOD-Student-Doctor, data = train\_flag, family = binomial(link='logit'))

summary(flag\_reduced\_mod)

```

#MODEL SELECTION:

I would like to select model5 for Binary Logistic Regression models. The AIC and residual deviance for this model seemed to give the best values that would be suited for the prediction. Below is the ROC curve for model5 and to me it looks good. So i would like to proceed with model5. For Multiple linear model i wouldd like to go for model2.

```{r}

train\_flag$predict <- predict(flag\_reduced\_mod, train\_flag, type='response')

roc\_model3 <- roc(train\_flag$TARGET\_FLAG, train\_flag$predict, plot=T, asp=NA,

legacy.axes=T, main = "ROC Curve", col="blue")

roc\_model3["auc"]

```

Now lets do the confusion matrix:

```{r}

train\_flag$predict\_target <- ifelse(train\_flag$predict >=0.5, 1, 0)

train\_flag$predict\_target <- as.integer(train\_flag$predict\_target)

myvars <- c("TARGET\_FLAG", "predict\_target")

train\_flag\_cm <- train\_flag[myvars]

cm <- table(train\_flag\_cm$predict\_target,train\_flag\_cm$TARGET\_FLAG)

knitr:: kable(cm)

```

```{r}

Accuracy <- function(data) {

tb <- table(train\_flag\_cm$predict\_target,train\_flag\_cm$TARGET\_FLAG)

TN=tb[1,1]

TP=tb[2,2]

FN=tb[2,1]

FP=tb[1,2]

return((TP+TN)/(TP+FP+TN+FN))

}

Accuracy(data)

```

```{r}

CER <- function(data) {

tb <- table(train\_flag\_cm$predict\_target,train\_flag\_cm$TARGET\_FLAG)

TN=tb[1,1]

TP=tb[2,2]

FN=tb[2,1]

FP=tb[1,2]

return((FP+FN)/(TP+FP+TN+FN))

}

CER(data)

```

```{r}

Precision <- function(data) {

tb <- table(train\_flag\_cm$predict\_target,train\_flag\_cm$TARGET\_FLAG)

TP=tb[2,2]

FP=tb[1,2]

return((TP)/(TP+FP))

}

Precision(data)

```

```{r}

Sensitivity <- function(data) {

tb <- table(train\_flag\_cm$predict\_target,train\_flag\_cm$TARGET\_FLAG)

TP=tb[2,2]

FN=tb[2,1]

return((TP)/(TP+FN))

}

Sensitivity(data)

```

```{r}

Specificity <- function(data) {

tb <- table(train\_flag\_cm$predict\_target,train\_flag\_cm$TARGET\_FLAG)

TN=tb[1,1]

TP=tb[2,2]

FN=tb[2,1]

FP=tb[1,2]

return((TN)/(TN+FP))

}

Specificity(data)

```

```{r}

F1\_score <- function(data) {

tb <- table(train\_flag\_cm$predict\_target,train\_flag\_cm$TARGET\_FLAG)

TN=tb[1,1]

TP=tb[2,2]

FN=tb[2,1]

FP=tb[1,2]

Precision = (TP)/(TP+FP)

Sensitivity = (TP)/(TP+FN)

Precision =(TP)/(TP+FP)

return((2\*Precision\*Sensitivity)/(Precision+Sensitivity))

}

F1\_score(data)

```

#TEST DATA PREPARATION AND TESTING THE MODEL ON EVALUATION DATA:

In the final step we will test our model by using the test data.

```{r}

ins\_eval <- read.csv("https://raw.githubusercontent.com/Riteshlohiya/Data621-Assignment-4/master/insurance\_evaluation\_data.csv")

ins\_eval$INCOME <- as.numeric(str\_replace\_all(ins\_eval$INCOME, "[[:punct:]\\$]",""))

ins\_eval$HOME\_VAL <- as.numeric(str\_replace\_all(ins\_eval$HOME\_VAL, "[[:punct:]\\$]",""))

ins\_eval$BLUEBOOK <- as.numeric(str\_replace\_all(ins\_eval$BLUEBOOK, "[[:punct:]\\$]",""))

ins\_eval$OLDCLAIM <- as.numeric(str\_replace\_all(ins\_eval$OLDCLAIM, "[[:punct:]\\$]",""))

#Convert indicator variables to 0s and 1s; 1 = Yes, Male for Sex, Commercial for Car Use, Red for RED\_CAR, and Highly Urban for URBANICITY

ins\_eval$PARENT1 <- ifelse(ins\_eval$PARENT1=="Yes", 1, 0)

ins\_eval$MSTATUS <- ifelse(ins\_eval$MSTATUS=="Yes", 1, 0)

ins\_eval$SEX <- ifelse(ins\_eval$SEX=="M", 1, 0)

ins\_eval$CAR\_USE <- ifelse(ins\_eval$CAR\_USE=="Commercial", 1, 0)

ins\_eval$RED\_CAR <- ifelse(ins\_eval$RED\_CAR=="Yes", 1, 0)

ins\_eval$REVOKED <- ifelse(ins\_eval$REVOKED=="Yes", 1, 0)

ins\_eval$URBANICITY <- ifelse(ins\_eval$URBANICITY == "Highly Urban/ Urban", 1, 0)

#Convert categorical predictor values to indicator variables - EDUCATION, CAR\_TYPE, JOB

#EDUCATION, High school graduate is base case

ins\_eval$HSDropout <- ifelse(ins\_eval$EDUCATION=="<High School", 1, 0)

ins\_eval$Bachelors <- ifelse(ins\_eval$EDUCATION=="Bachelors", 1, 0)

ins\_eval$Masters <- ifelse(ins\_eval$EDUCATION=="Masters", 1, 0)

ins\_eval$PhD <- ifelse(ins\_eval$EDUCATION=="PhD", 1, 0)

#CAR\_TYPE, base case is minivan

ins\_eval$Panel\_Truck <- ifelse(ins\_eval$CAR\_TYPE=="Panel Truck", 1, 0)

ins\_eval$Pickup <- ifelse(ins\_eval$CAR\_TYPE=="Pickup", 1, 0)

ins\_eval$Sports\_Car <- ifelse(ins\_eval$CAR\_TYPE=="Sports Car", 1, 0)

ins\_eval$Van <- ifelse(ins\_eval$CAR\_TYPE=="Van", 1, 0)

ins\_eval$SUV <- ifelse(ins\_eval$CAR\_TYPE=="z\_SUV", 1, 0)

#JOB, base case is ""

ins\_eval$Professional <- ifelse(ins\_eval$JOB == "Professional", 1, 0)

ins\_eval$Blue\_Collar <- ifelse(ins\_eval$JOB == "Professional", 1, 0)

ins\_eval$Clerical <- ifelse(ins\_eval$JOB == "Clerical", 1, 0)

ins\_eval$Doctor <- ifelse(ins\_eval$JOB == "Doctor", 1, 0)

ins\_eval$Lawyer <- ifelse(ins\_eval$JOB == "Lawyer", 1, 0)

ins\_eval$Manager <- ifelse(ins\_eval$JOB == "Manager", 1, 0)

ins\_eval$Home\_Maker <- ifelse(ins\_eval$JOB == "Home Maker", 1, 0)

ins\_eval$Student <- ifelse(ins\_eval$JOB == "Student", 1, 0)

ins\_eval <- ins\_eval %>% dplyr::select(-c(INDEX,EDUCATION,CAR\_TYPE,JOB))

fillwithmedian <- function(x) {

median\_val = median(x, na.rm = TRUE)

x[is.na(x)] = median\_val

return(x)

}

ins\_eval <- data.frame(lapply(ins\_eval, fillwithmedian))

ins\_eval$INCOME\_MOD <- ins\_eval$INCOME ^0.433

ins\_eval$HOME\_VAL\_MOD <- ins\_eval$HOME\_VAL^0.113

ins\_eval$BLUEBOOK\_MOD <- ins\_eval$BLUEBOOK^0.461

ins\_eval$OLD\_CLAIM\_MOD <- log(ins\_eval$OLDCLAIM + 1)

ins\_eval$predict\_prob <- predict(flag\_reduced\_mod, ins\_eval, type='response')

ins\_eval$predict\_target <- ifelse(ins\_eval$predict\_prob >= 0.50, 1,0)

write.csv(ins\_eval,"Evaluation\_Data.csv", row.names=FALSE)

ins\_eval$TARGET\_AMT1 <- 0

ins\_eval1 <- filter(ins\_eval, predict\_target == 1)

ins\_eval1$predict\_target<-as.numeric(ins\_eval1$predict\_target)

ins\_eval1$TARGET\_AMT1 <- predict(amount\_reduced\_model2, newdata=ins\_eval1)

write.csv(ins\_eval1,"Evaluation\_Full\_Data.csv", row.names=FALSE)

```