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:]\\$]",""))

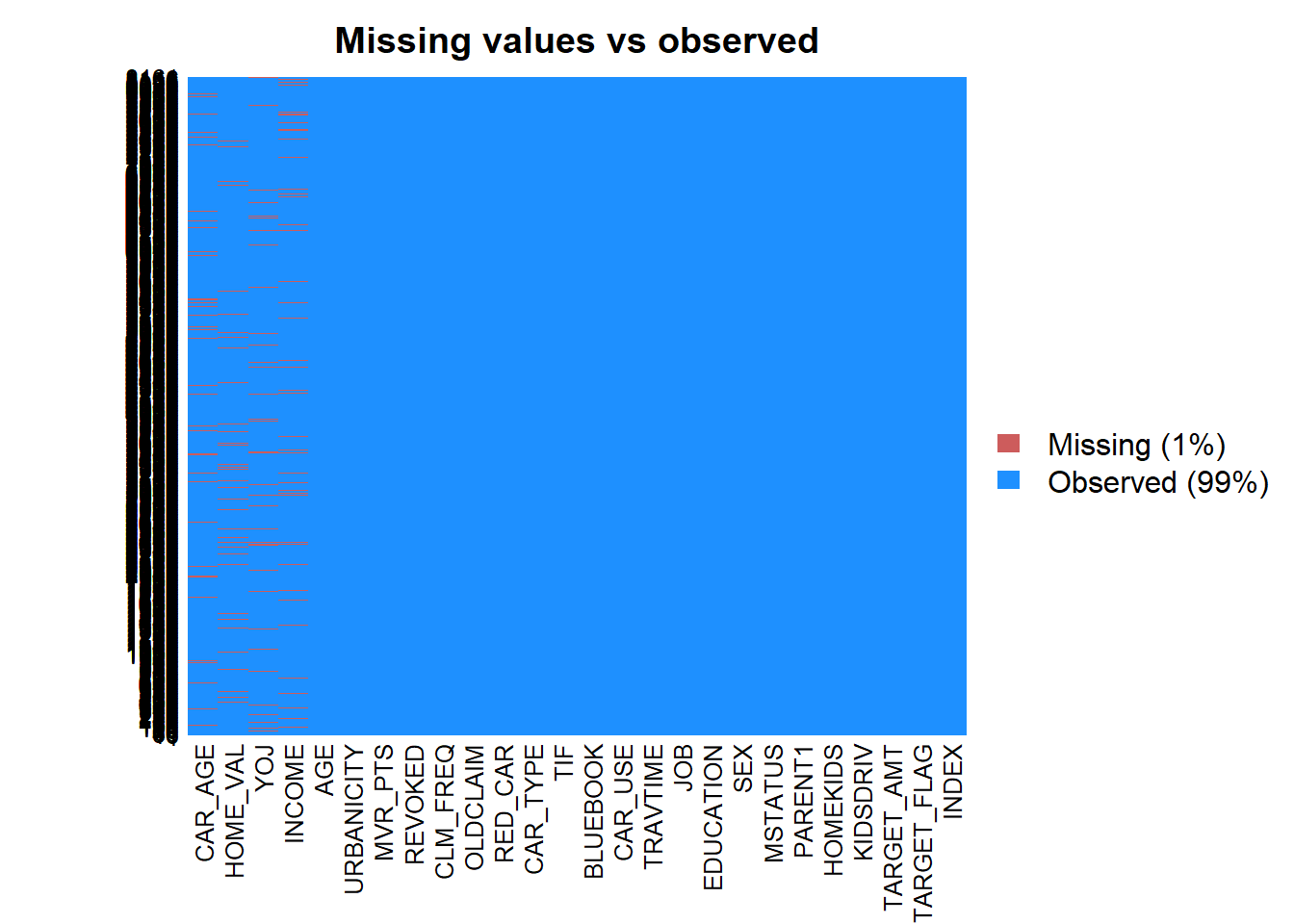
ins\_train$HOME\_VAL <- as.numeric(str\_replace\_all(ins\_train$HOME\_VAL, "[[:punct:]\\$]",""))

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

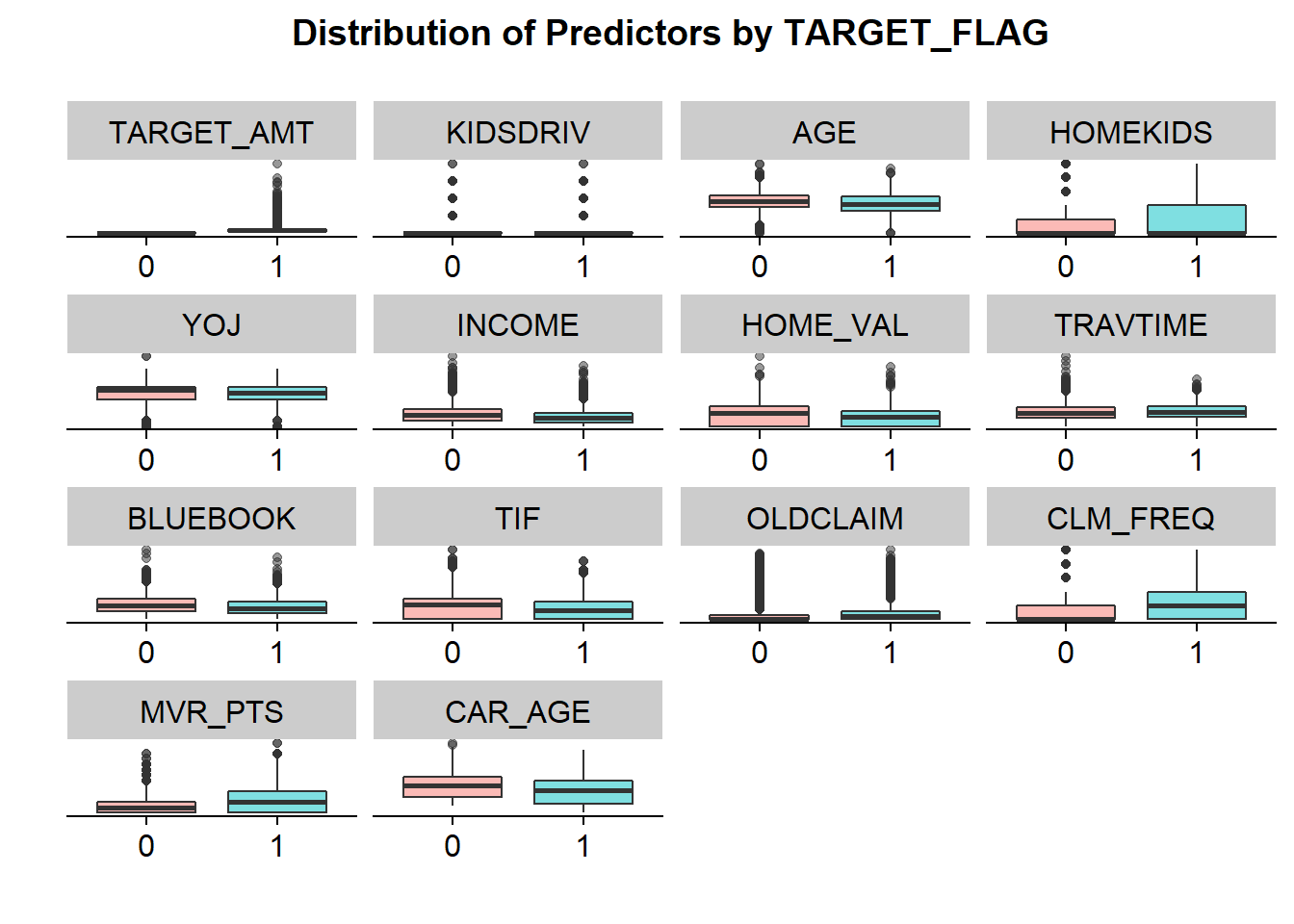
Now we will see the missing values in the dataset. For this i have used Amelia package. We can see there are missing values for CAR\_AGE, HOME\_VAL, YOJ and INCOME. There needs to be taken care while we do data preparation.



Now lets do some plots to understand the data:

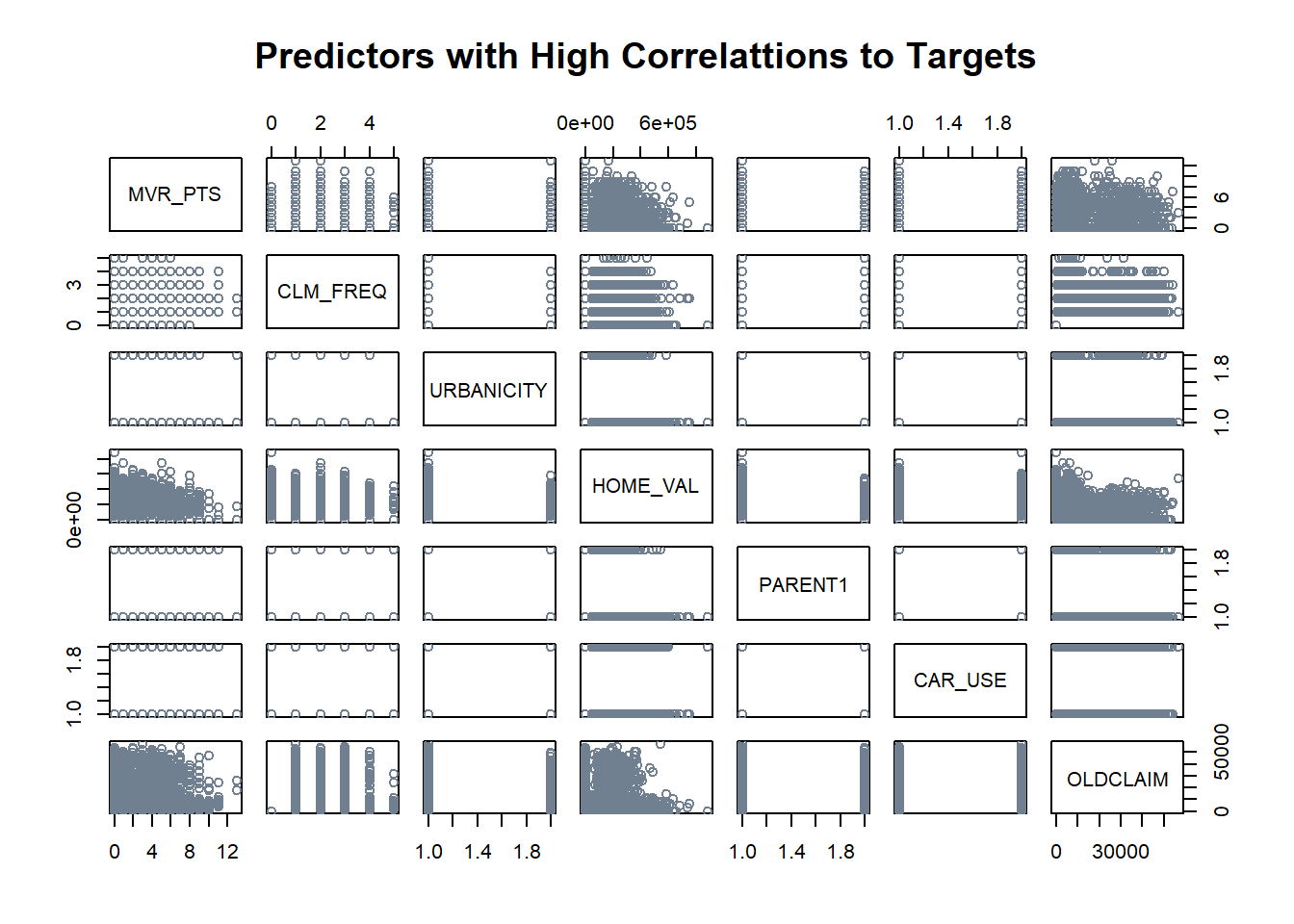
AGE - Age of Driver. Very young people tend to be risky. Maybe very old people also. We note six missing values that we’ll need to address later. The distribution of AGE is almost perfectly normal. When we break out the data by TARGET\_FLAG values, the distributions of age by TARGET\_FLAG are still roughly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

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



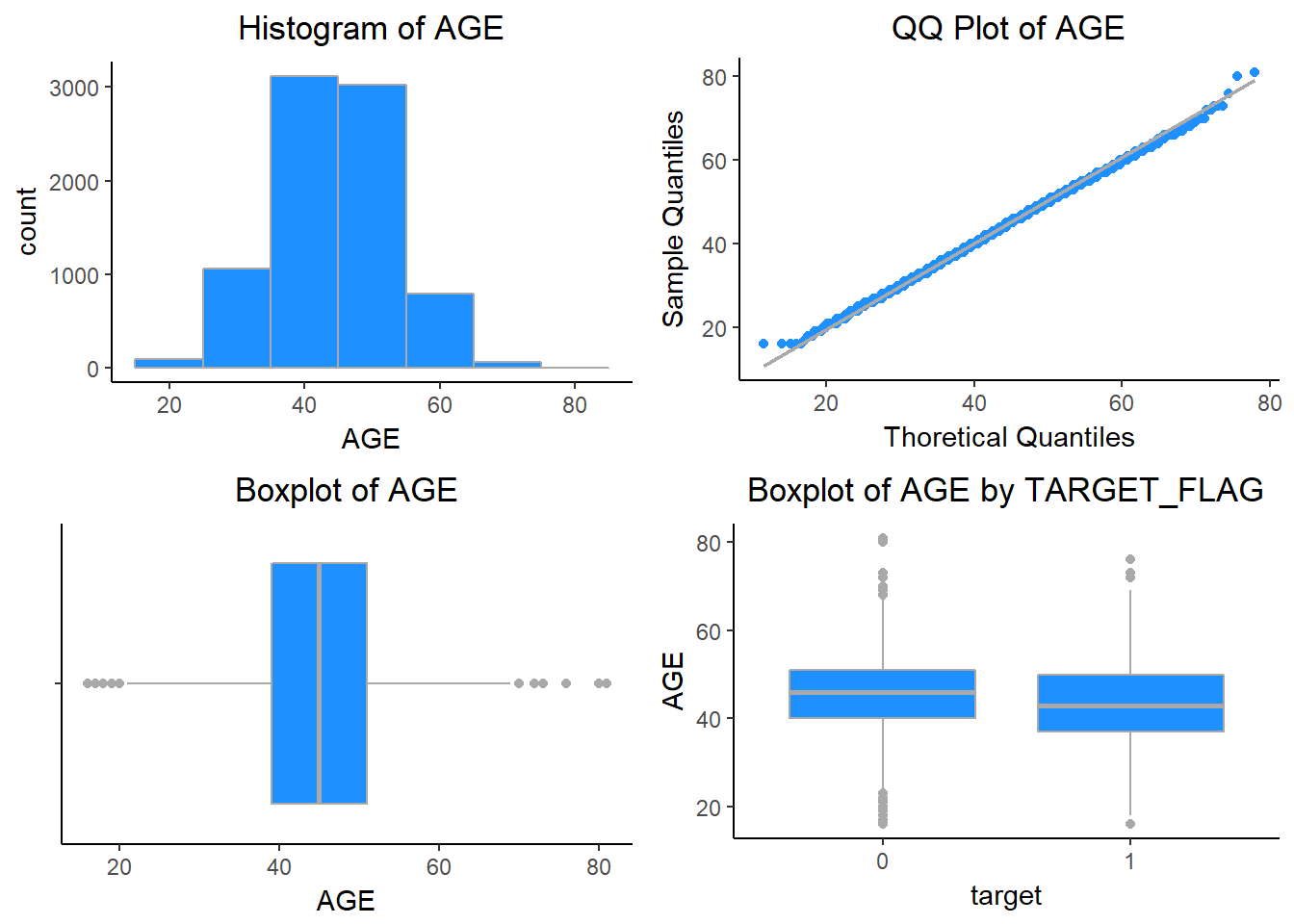
### Correlation

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.

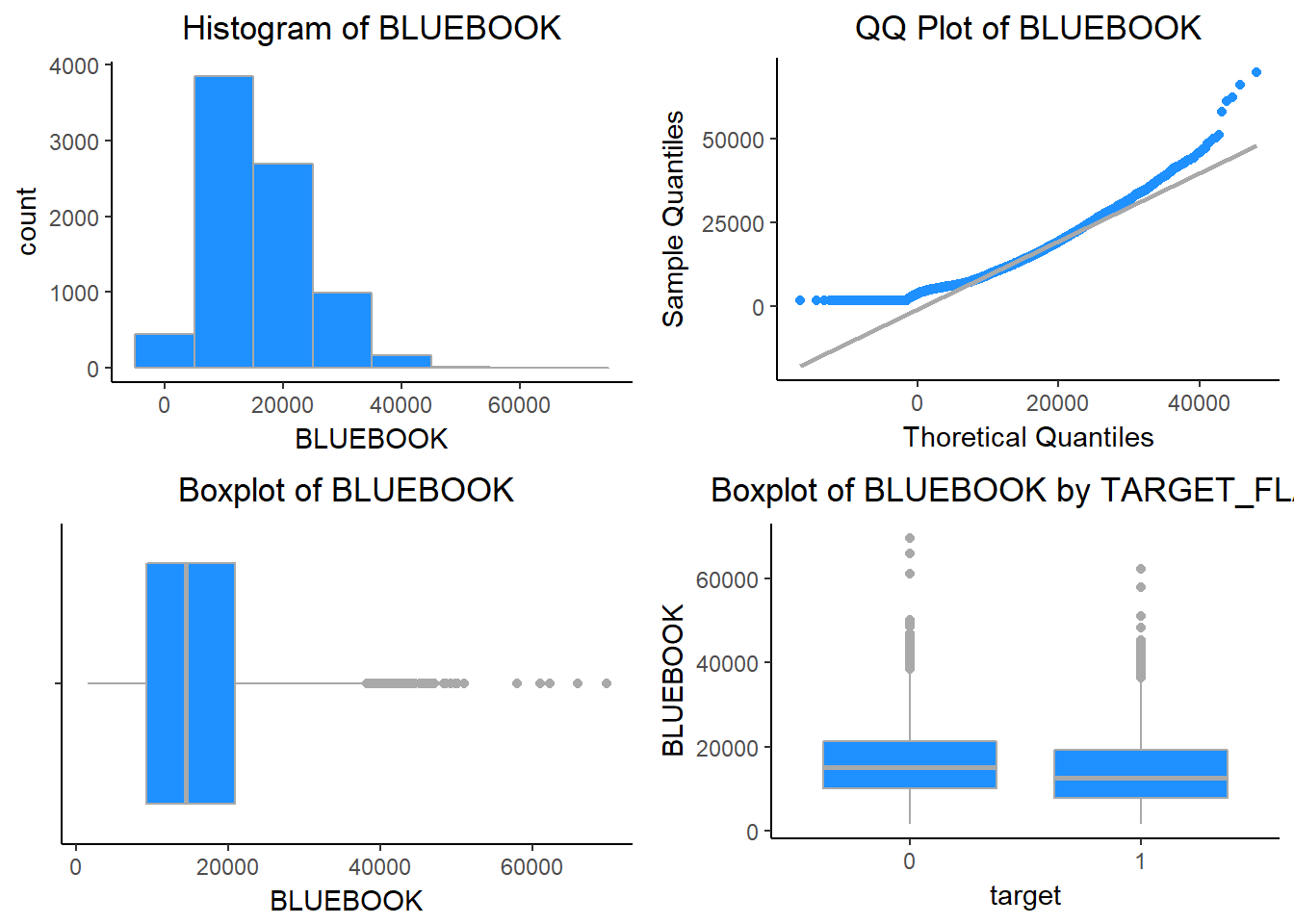


Now let’s do some plots to understand the data:

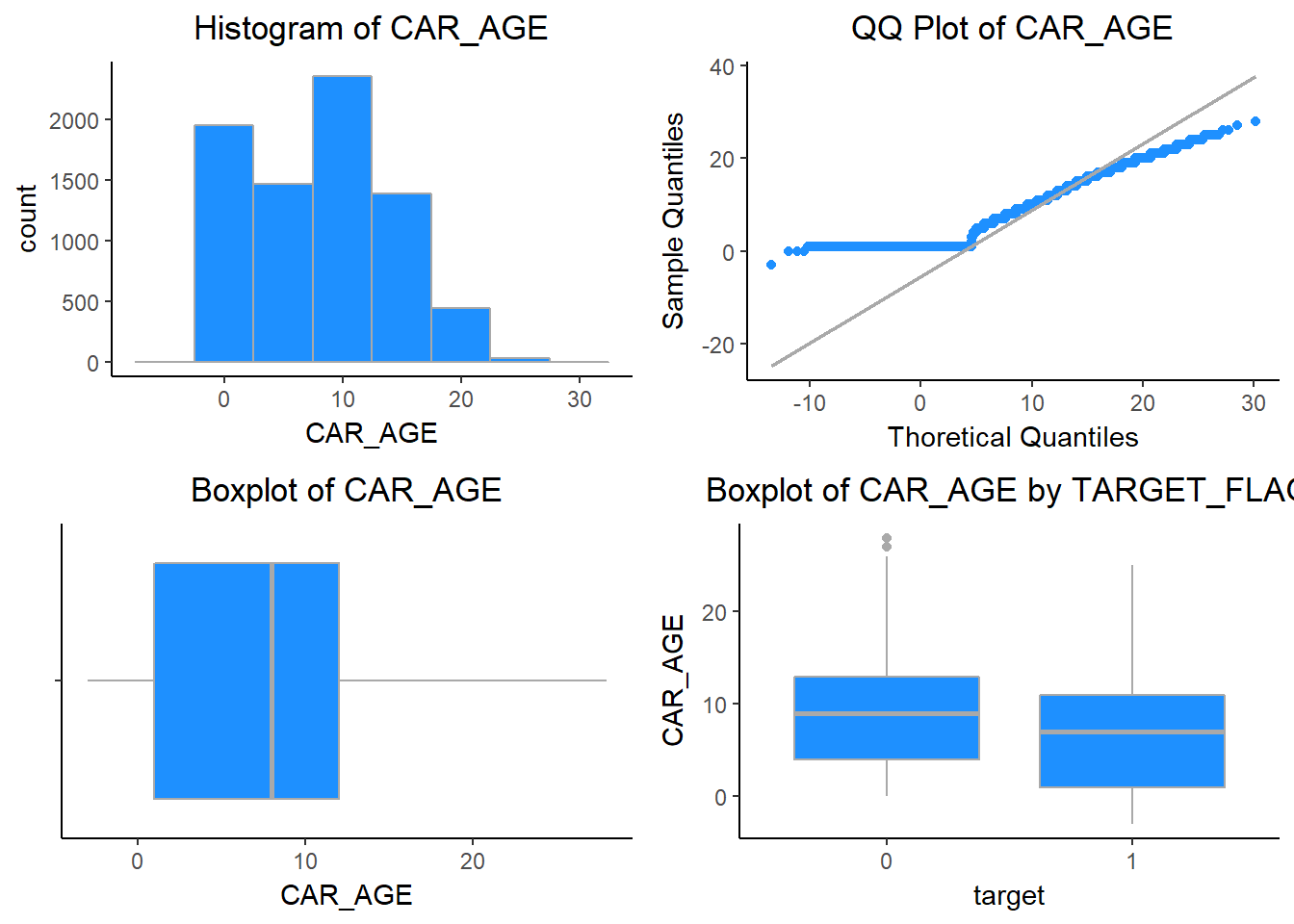
AGE - Age of Driver. Very young people tend to be risky. Maybe very old people also. We note six missing values that we’ll need to address later. The distribution of AGE is almost perfectly normal. When we break out the data by TARGET\_FLAG values, the distributions of age by TARGET\_FLAG are still roughly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



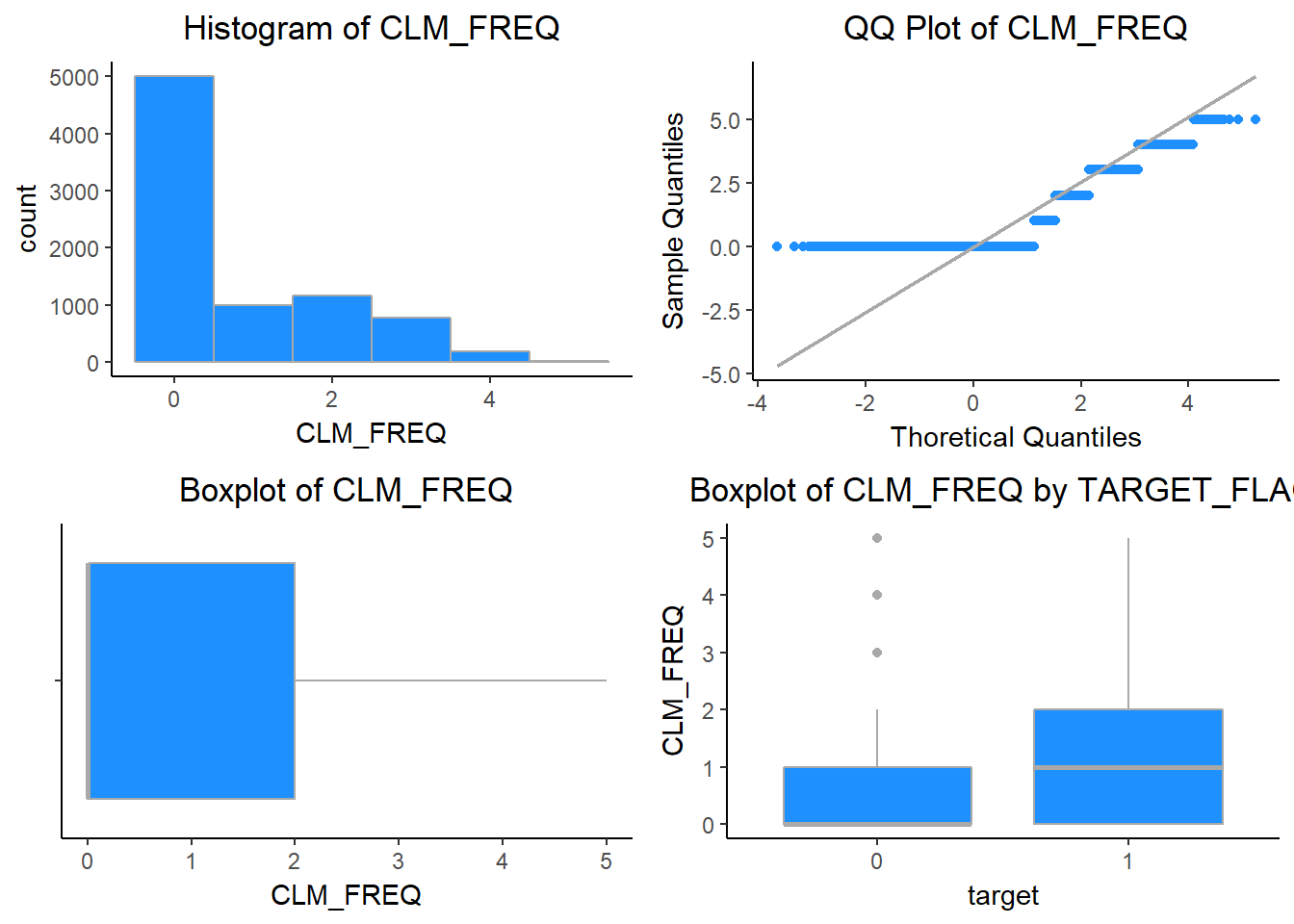
BLUEBOOK - Value of Vehicle. Unknown effect on probability of collision, but probably effect the payout if there is a crash. Individuals involved in crashes have a higher proportion of low BLUEBOOK values.



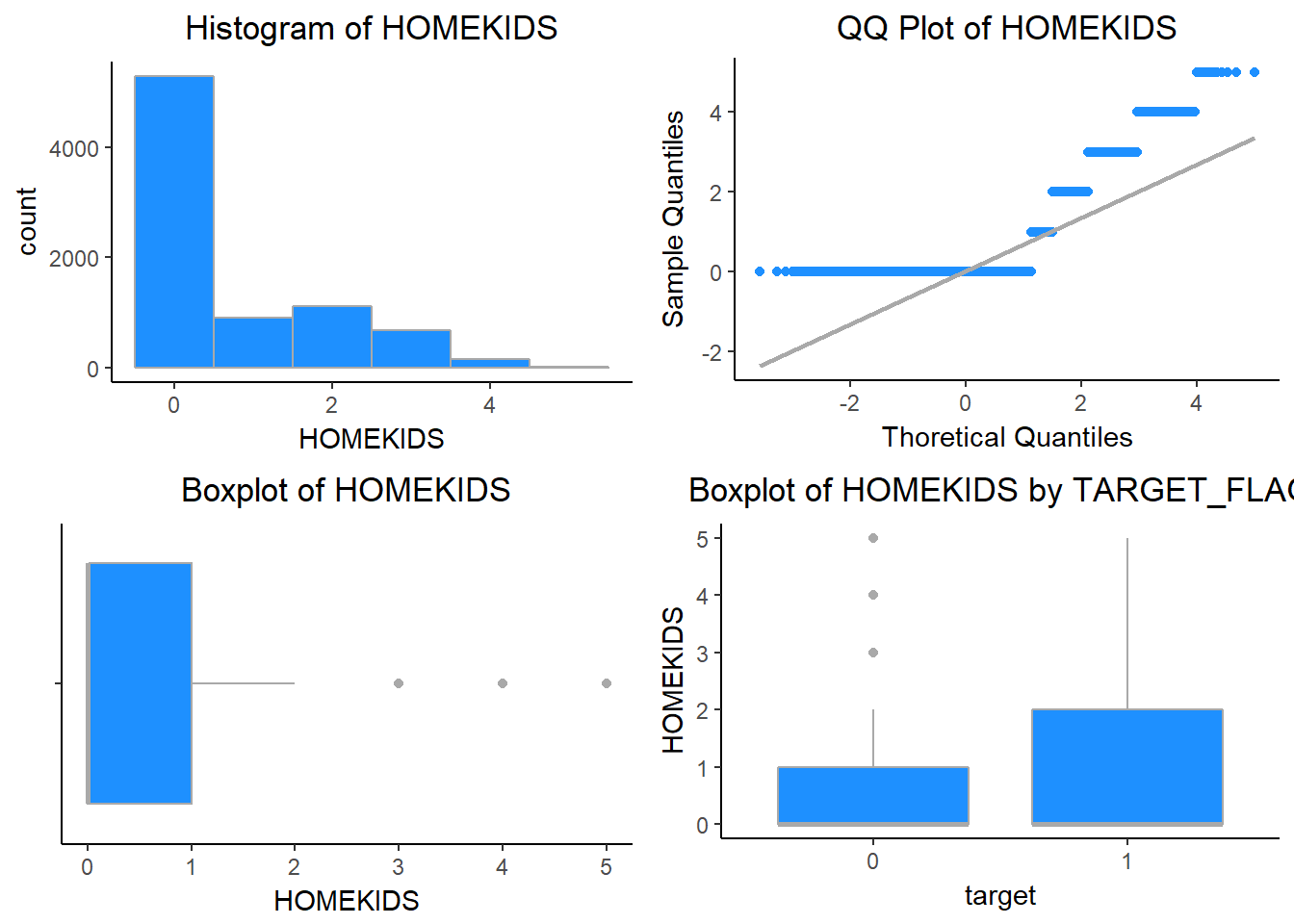
CAR\_AGE - Vehicle Age. We could see there is one negative value for CAR\_AGE. We have to treat this value in our data preparation step.



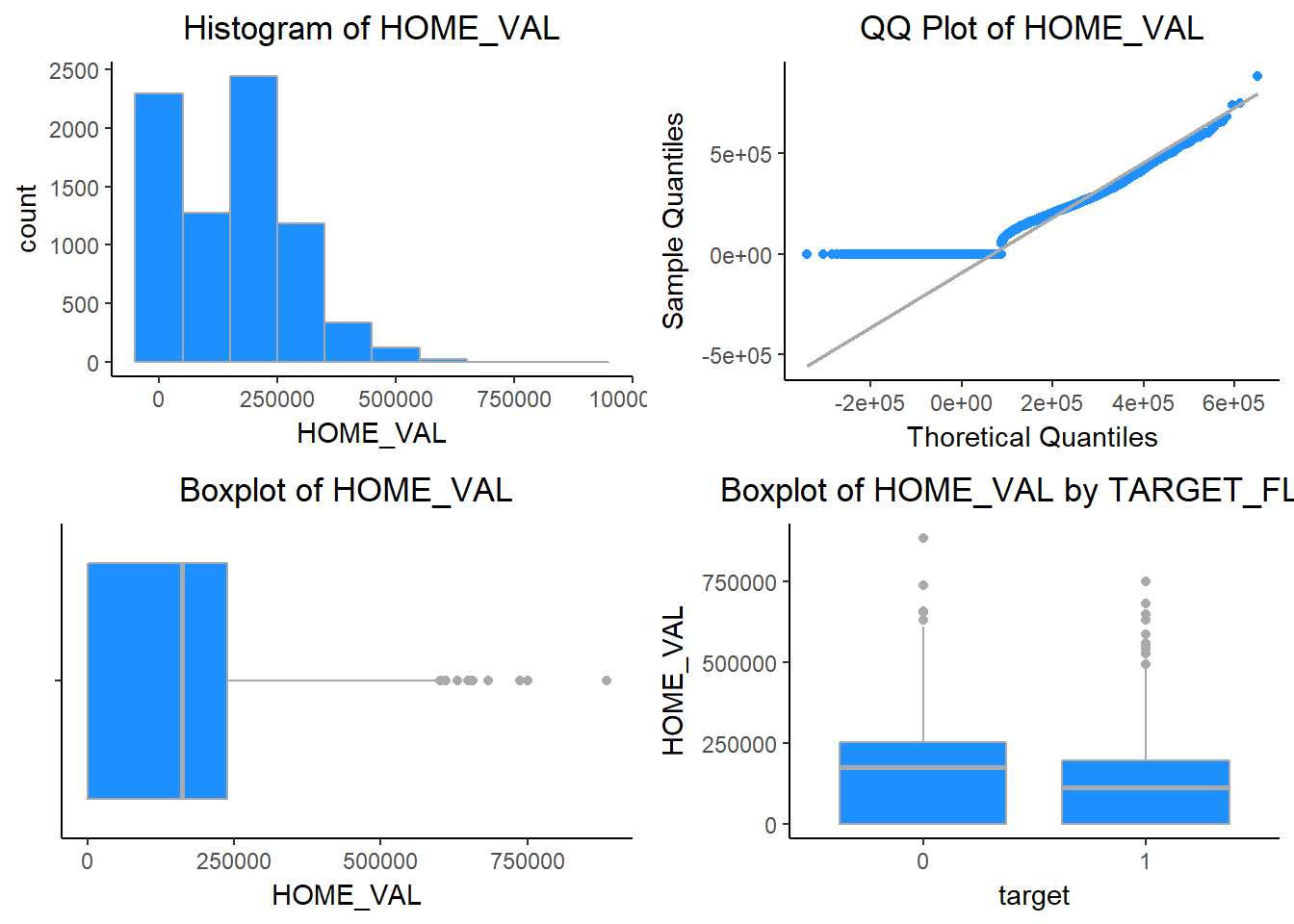
CLM\_FREQ - # Claims (Past 5 Years). The more claims you filed in the past, the more you are likely to file in the future. We can see that this variable is also skewed.



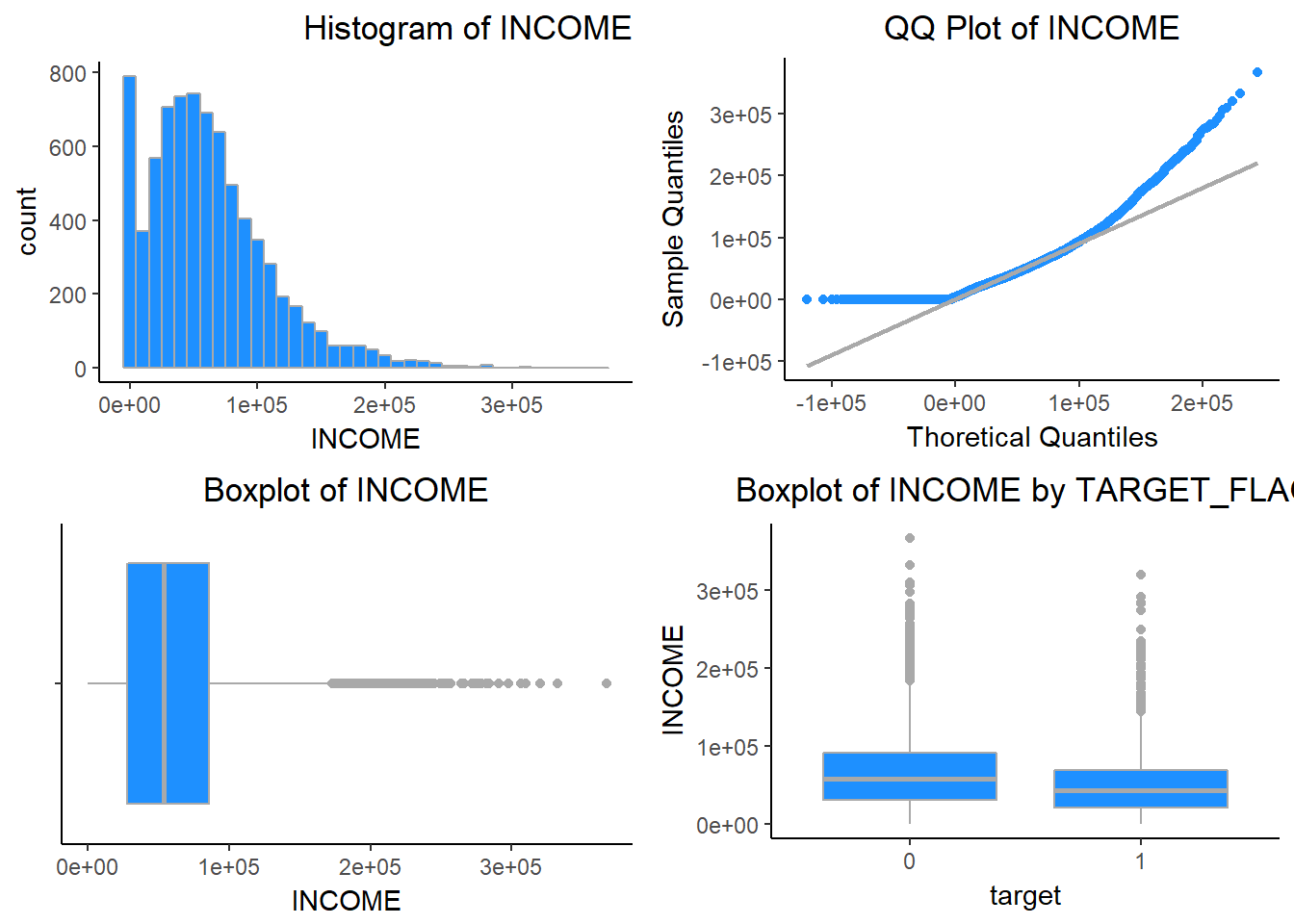
HOMEKIDS - # Children at Home. HOMEKIDS does not seem to impact the TARGET\_FLAG. The distribution of this discrete variable is right skewed.



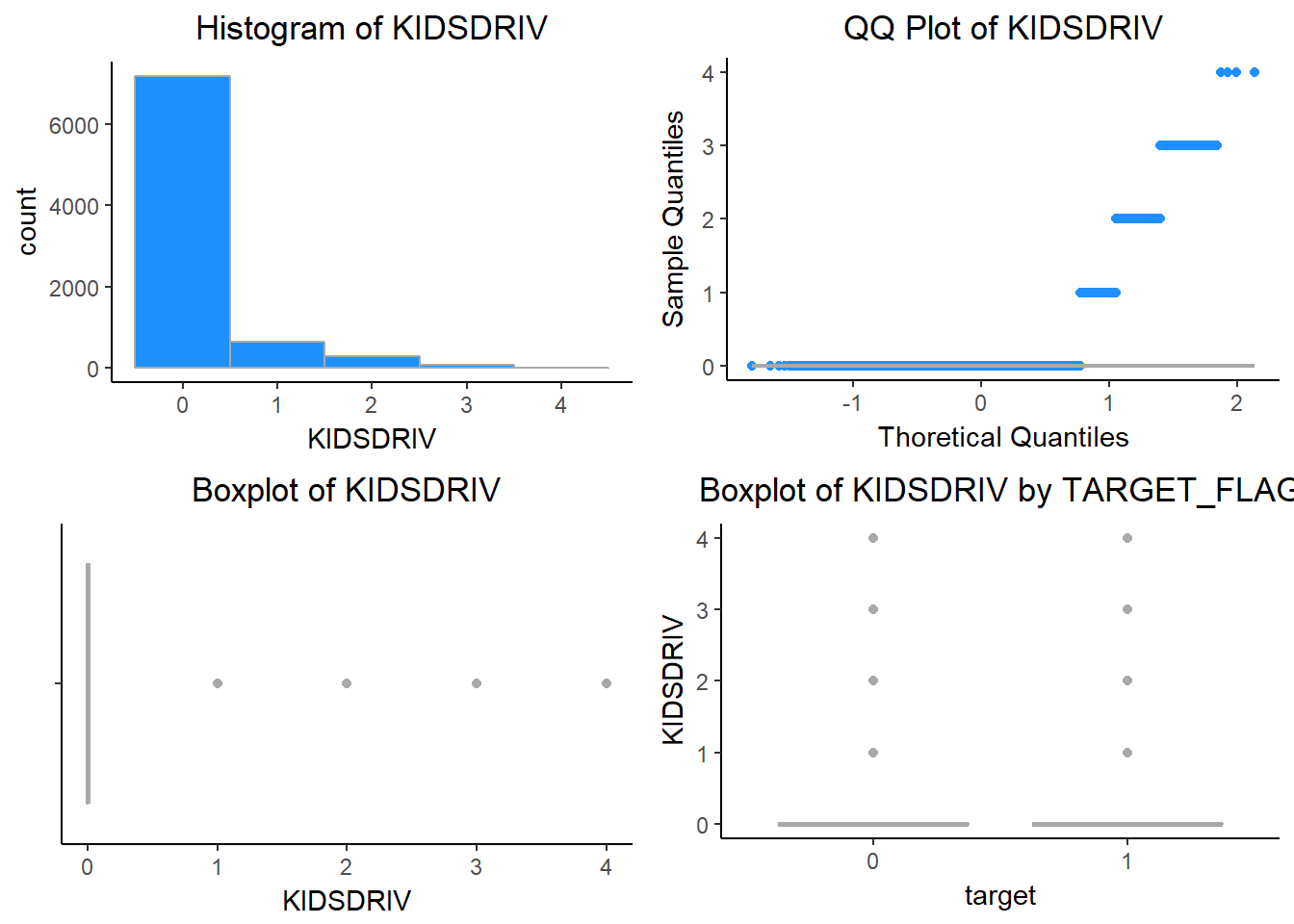
HOME\_VAL - Home Value. Home owners tend to drive more responsibly. The distribution of HOME\_VAL is right skewed and also we can there are some missing values.



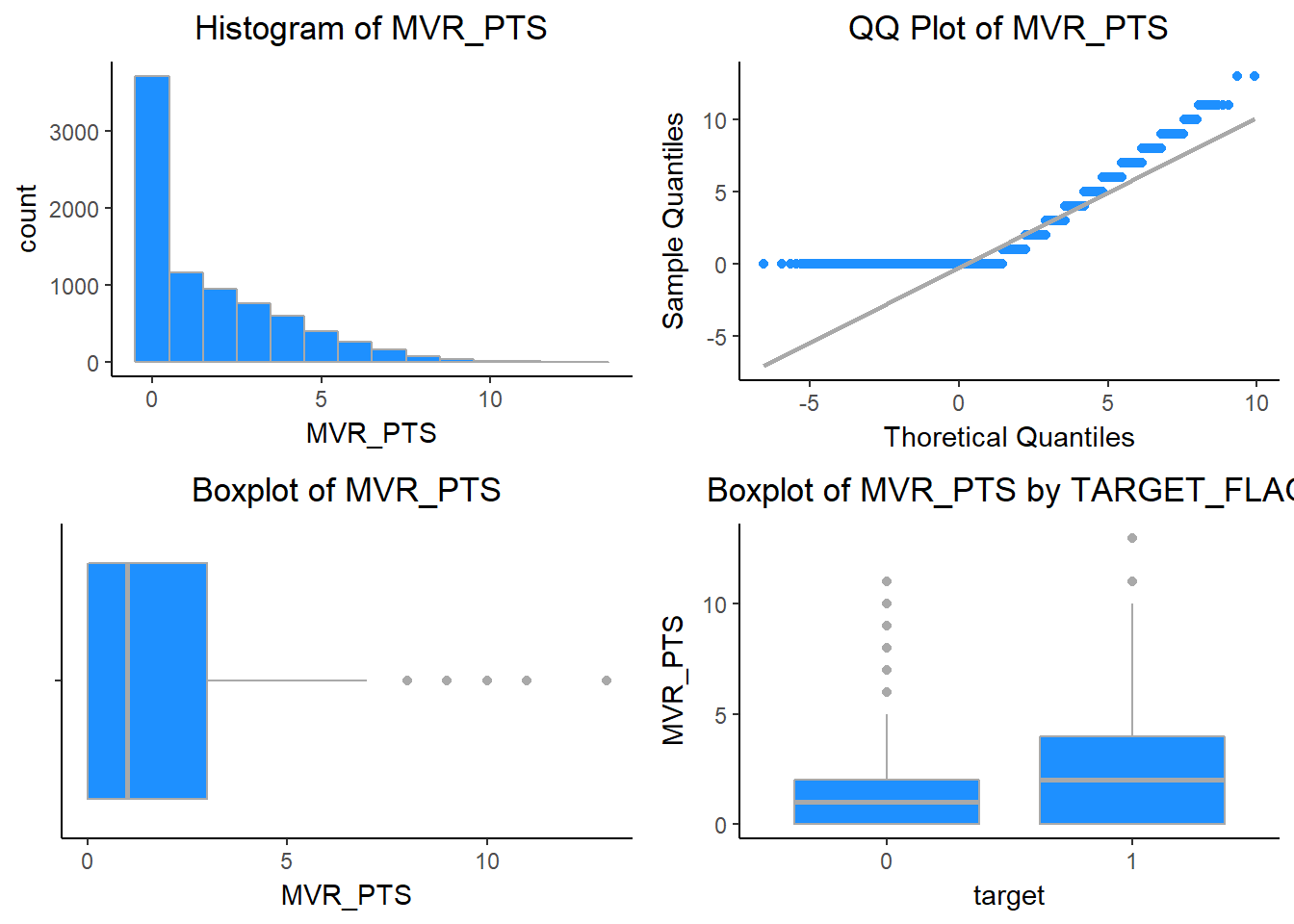
INCOME - Income of the person. Rich people tend to get into fewer crashes. The distribution of INCOME is right skewed, with a significant number of observations indicating $0 in income. There are some missing values in this as well.



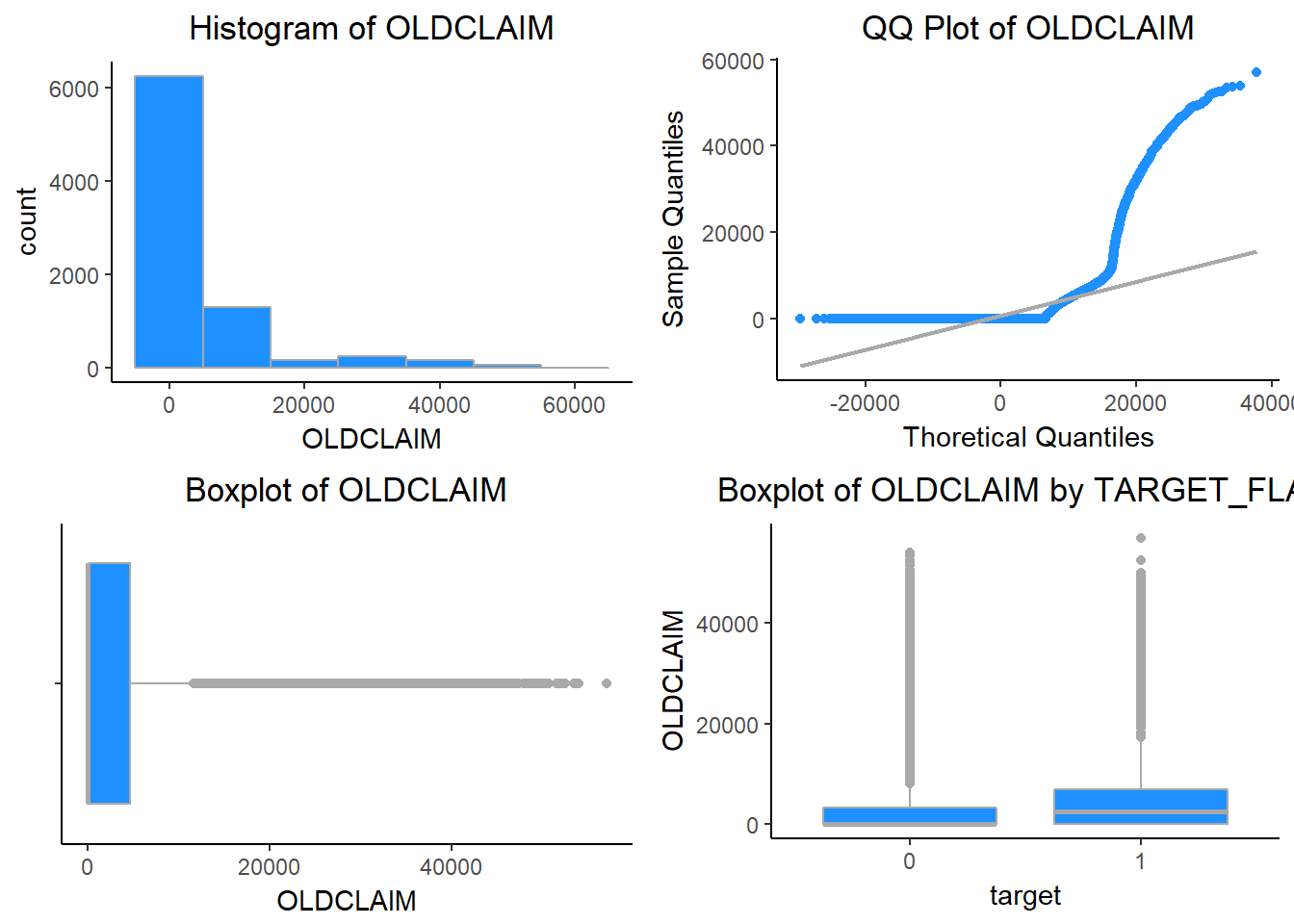
KIDSDRIV - # Driving Children. When teenagers drive your car, you are more likely to get into crashes. The discrete variable KIDSDRIV is right skewed



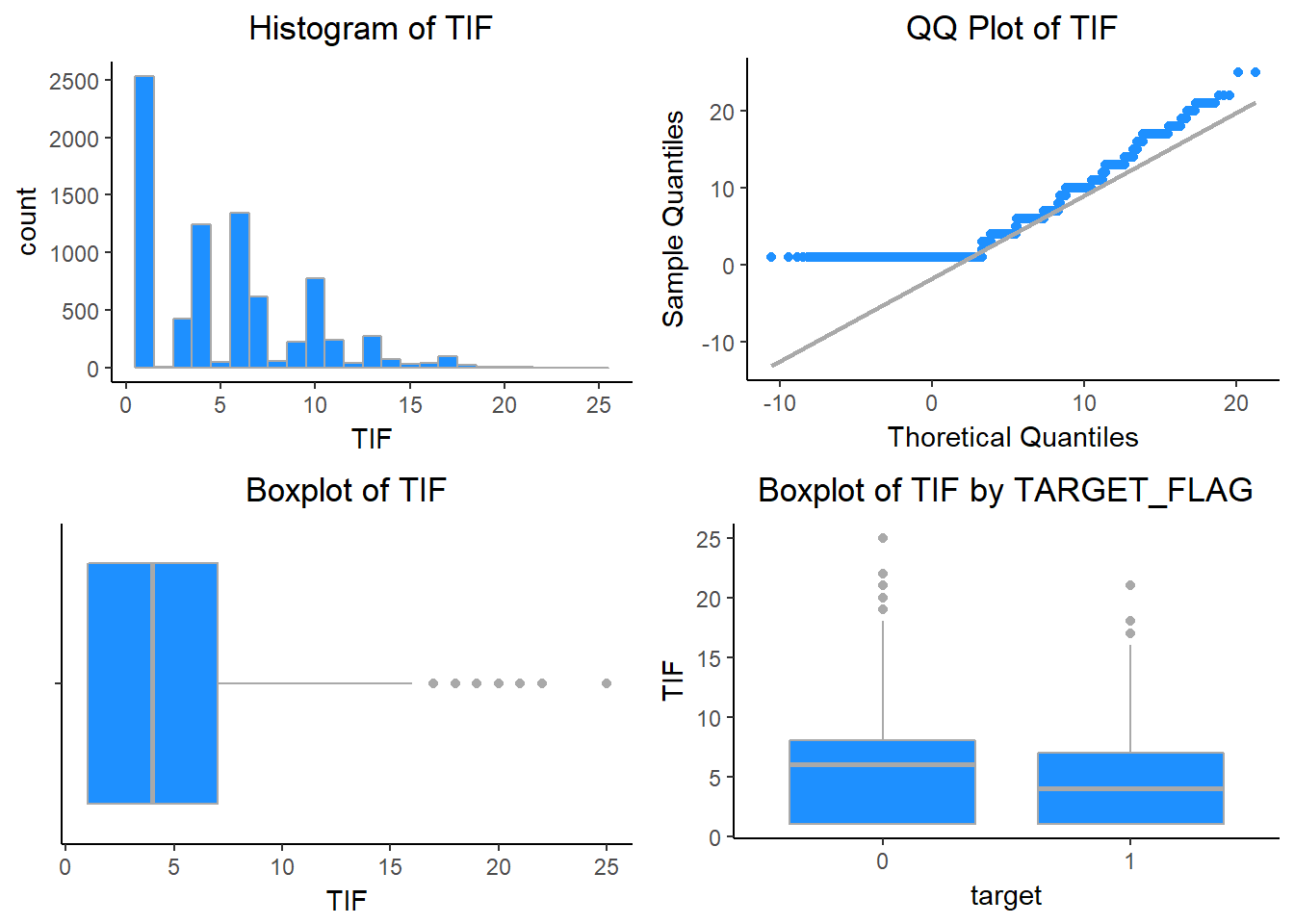
MVR\_PTS - Motor Vehicle Record Points. If you get lots of traffic tickets, you tend to get into more crashes. MVR\_PTS is positively skewed.



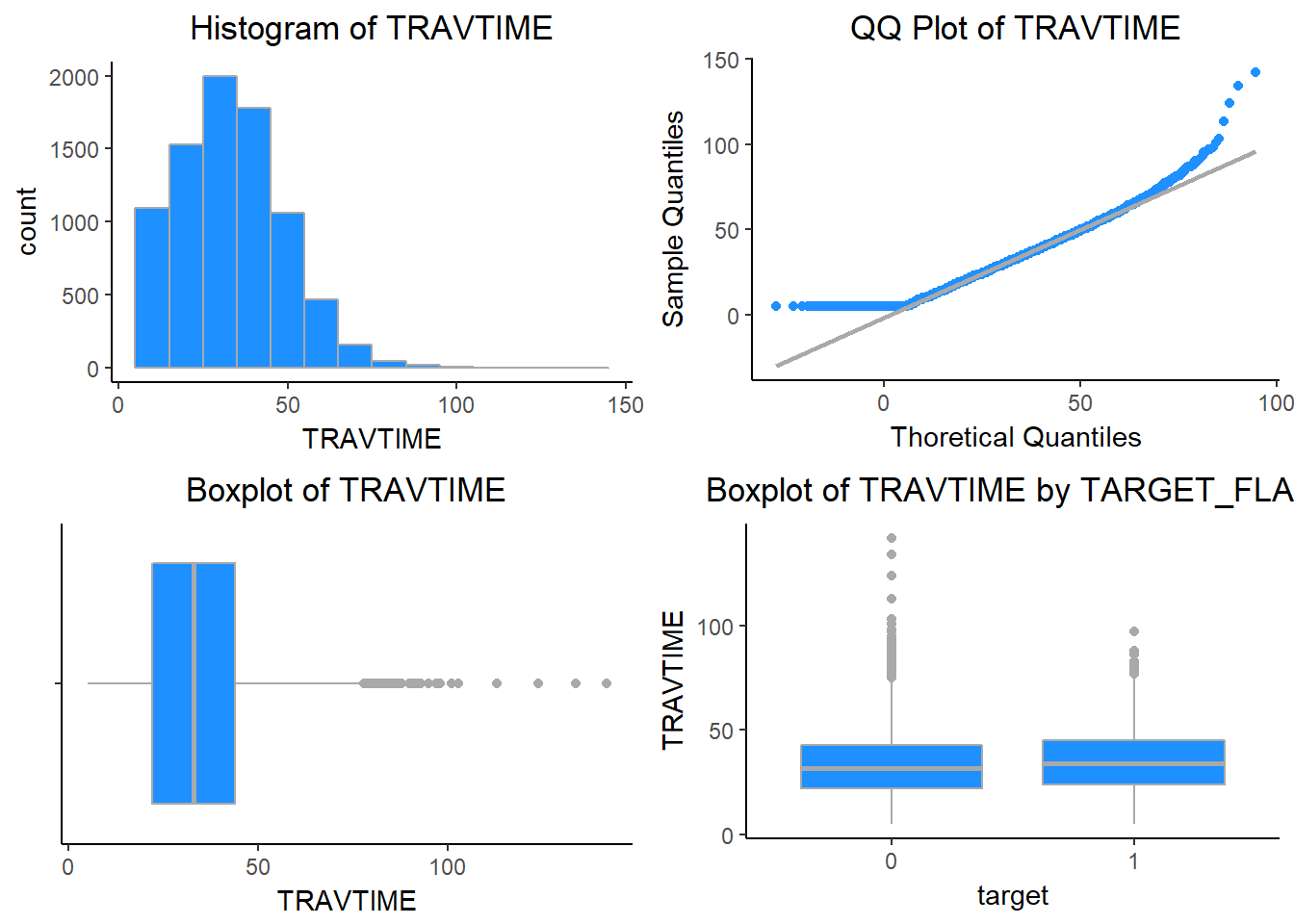
OLDCLAIM - Total Claims (Past 5 Years). If your total payout over the past five years was high, this suggests future payouts will be high. The distribution of OLDCLAIM is extremely right skewed.



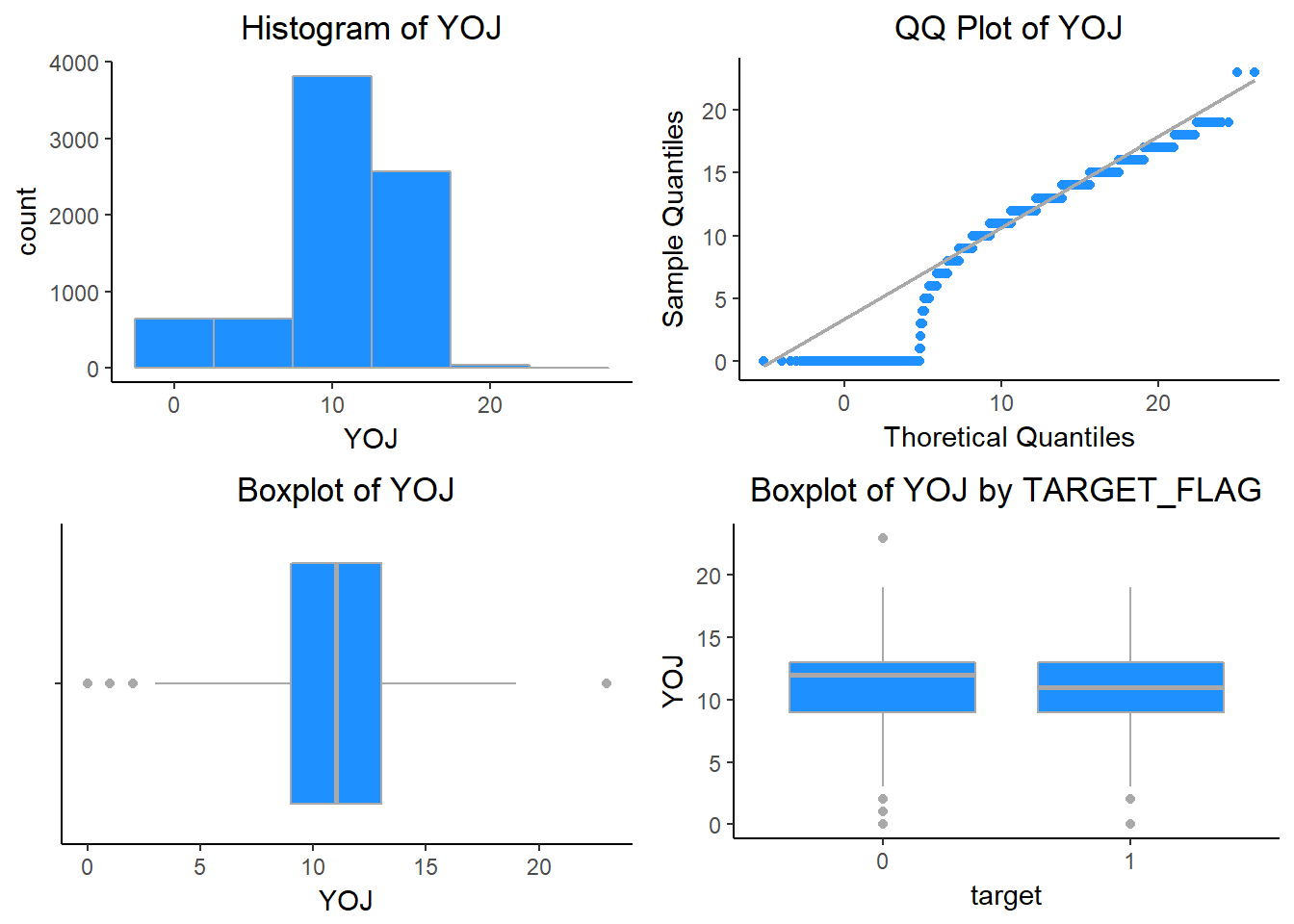
TIF - Time in Force. People who have been customers for a long time are usually more safe. The distribution is somewhat positively skewed.



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

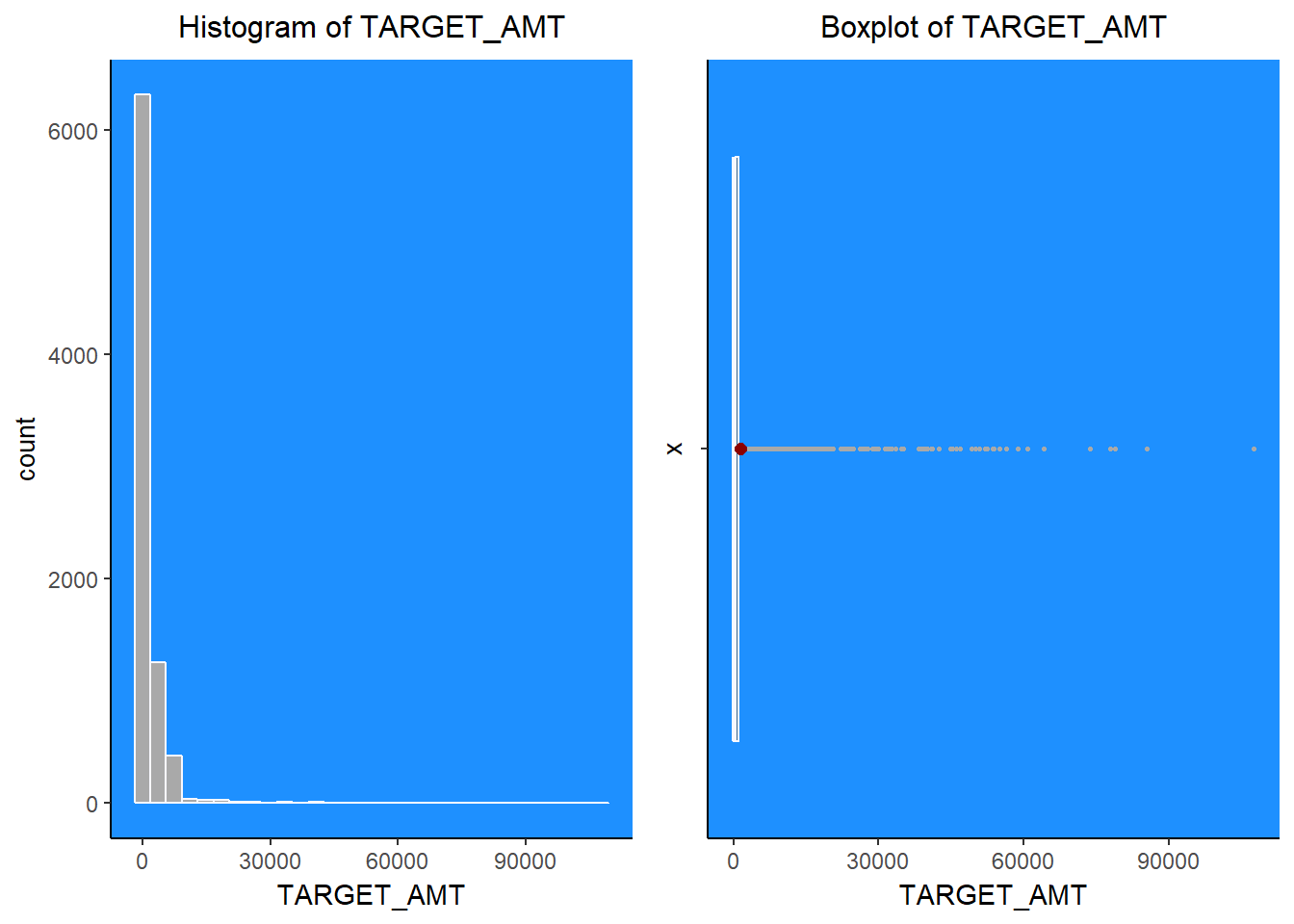
## count 6008.0 2153.0 8161

## 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

##

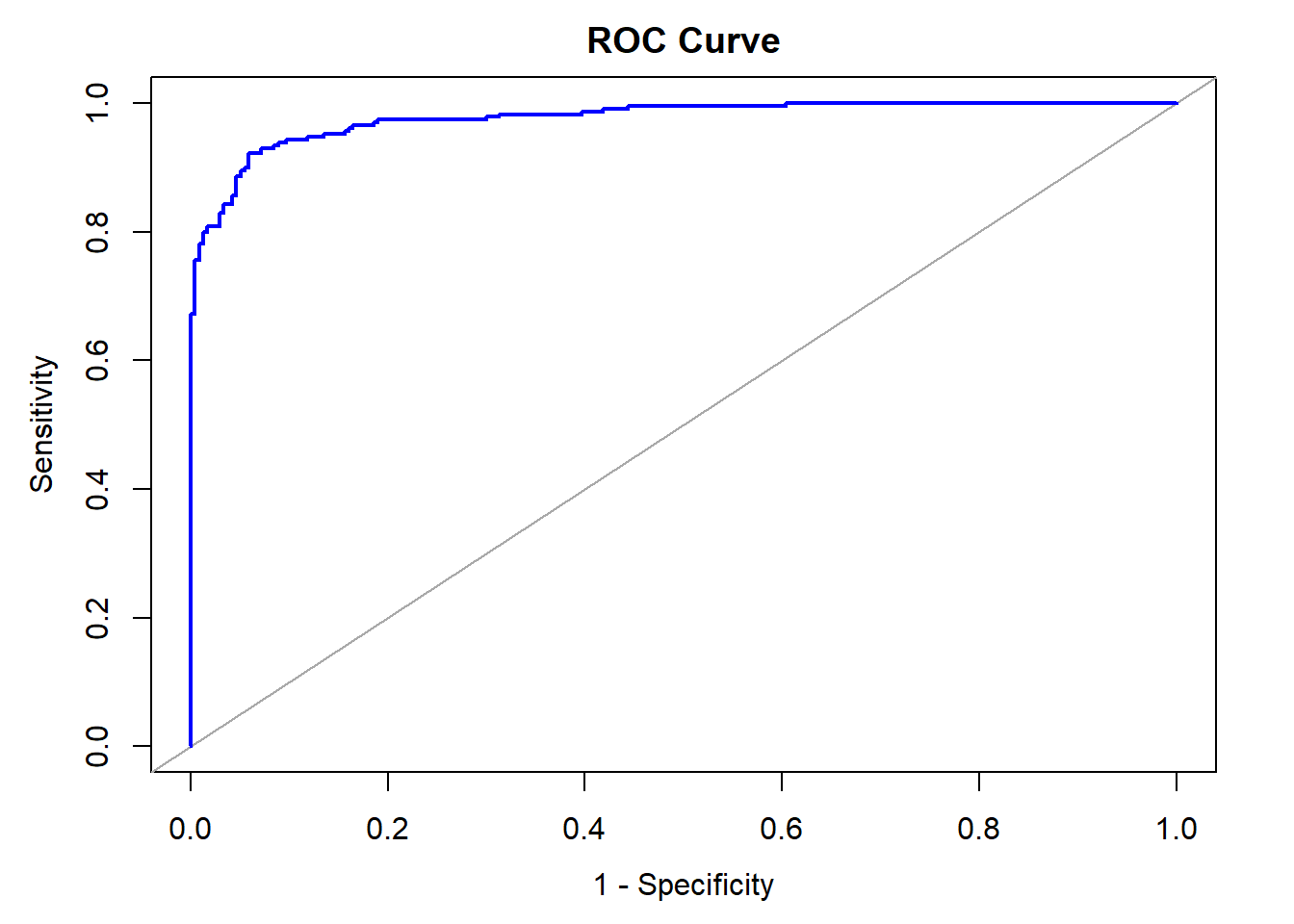
## Number of Fisher Scoring iterations: 9

# 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 Assignment3"

author: "Ritesh Lohiya"

date: "June 30, 2018"

output: html\_document

---

#Overview

objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. Below is a short description of the variables in the dataset.

zn: proportion of residential land zoned for large lots (over 25000 square feet)

indus: proportion of non-retail business acres per suburb

chas: a dummy var. for whether the suburb borders the Charles River (1) or not (0)

nox: nitrogen oxides concentration (parts per 10 million)

rm: average number of rooms per dwelling

age: proportion of owner-occupied units built prior to 1940

dis: weighted mean of distances to five Boston employment centers

rad: index of accessibility to radial highways

tax: full-value property-tax rate per $10,000

ptratio: pupil-teacher ratio by town

black: 1000 $(B\_k - 0.63)^2$ where Bk is the proportion of blacks by town

lstat: lower status of the population (percent)

medv: median value of owner-occupied homes in $1000s

target: whether the crime rate is above the median crime rate (1) or not (0) (response variable)

```{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)

```

#DATA EXPLORATION:

```{r}

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

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

summary(crime\_train)

```

Visual Exploration:

Now we will see the missing values in the dataset. For this i have used Amelia package

```{r}

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

```

There are no missing values in the dataset.

Lets now dig into our response variables.

1. Response Variable zn - proportion of residential land zoned for large lots (over 25000 square feet). We can see there are more zeros values for zn and also has positive skewness. Also there appears to be relationship between crime rates and zn.

```{r}

with(crime\_train, c(summary(zn), SD=sd(zn), Skew=skewness(zn), Kurt=kurtosis(zn)))

hist <- ggplot(crime\_train, aes(zn)) + geom\_histogram(fill = 'dodgerblue', binwidth = 20, color = 'darkgray' ) +

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), zn)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

2. Response Variable: indus - proportion of non-retail business acres per suburb. The histogram below indicates a bi-modal quality to the variable’s distribution, with many values clustering in two ranges.

```{r}

with(crime\_train, c(summary(indus), SD=sd(indus), Skew=skewness(indus), Kurt=kurtosis(indus)))

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

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), indus)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

3. Response Variable: chas - a dummy var. for whether the suburb borders the Charles River (1) or not (0). This variable tells us if the neighborhood borders the Charles River (1) or not (0). Close to 7% of the neighborhood borders the Charles River. Of the areas bordering the Charles river 21 are in high crime areas.

```{r}

addmargins(table(crime\_train$chas))

addmargins(table(crime\_train$chas, crime\_train$target))

```

```{r}

ggplot(crime\_train, aes(x=target, y=chas)) + geom\_jitter(color='seagreen4') + theme\_classic() +

labs(title ='Jittered Scatter Plot of chas vs.target') + theme(plot.title = element\_text(hjust = 0.5))

```

4. Response Variable: nox - nitrogen oxides concentration (parts per 10 million). The variable nox represents the concentration of nitrogen oxide in each Boston area. There is also positive skewness. We also see moderately higher nox variance in high crime areas.

```{r}

with(crime\_train, c(summary(nox), SD=sd(nox), Skew=skewness(nox), Kurt=kurtosis(nox)))

hist <- ggplot(crime\_train, aes(nox)) + geom\_histogram(fill = 'dodgerblue', binwidth = .05, color = 'darkgray' ) +

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), nox)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

5. Response Variable: rm - average number of rooms per dwelling. The predictor rm is count measure describing the average number of rooms per dwelling. The distridution has heavy tail and has bell curve.

```{r}

with(crime\_train, c(summary(rm), SD=sd(rm), Skew=skewness(rm), Kurt=kurtosis(rm)))

hist <- ggplot(crime\_train, aes(rm)) + geom\_histogram(fill = 'dodgerblue', binwidth = 0.5, color = 'darkgray' ) +

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), rm)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

6. Response Variable: age - proportion of owner-occupied units built prior to 1940. The variable age indicates the proportion of owner occupied units built prior to 1940. This variable has high left skewness. Also there is significantly higher mean percentage of older homes in high crime areas.

```{r}

with(crime\_train, c(summary(age), SD=sd(age), Skew=skewness(age), Kurt=kurtosis(age)))

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

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

qq\_plot <- ggplot(crime\_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(crime\_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(crime\_train, aes(x=factor(target), age)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

7. Response Variable: dis - weighted mean of distances to five Boston employment centers. The predictor dist describes the average distance to Boston employment centers. The variable is moderately right skewed. Also we can see that low crime areas are associated with higher average distances to employment centers.

```{r}

with(crime\_train, c(summary(dis), SD=sd(dis), Skew=skewness(dis), Kurt=kurtosis(dis)))

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

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), dis)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

8. Response Variable: rad - index of accessibility to radial highways. The rad variable is an integer-valued index measure indicating an area’s accessibility to radial highways. In the boxplots below, there appears to be a significant positive association between high crime rates and rad value.

```{r}

with(crime\_train, c(summary(rad), SD=sd(rad), Skew=skewness(rad), Kurt=kurtosis(rad)))

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

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), rad)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

9. Response Variable: tax - full-value property-tax rate per $10,000. The tax variable refers to the the tax rate per $10k of property value. High crime areas also appear to have a strong, positive association with the tax value. This variable is densely distributed around two of the following approximate values: 300 and 700.

```{r}

with(crime\_train, c(summary(tax), SD=sd(tax), Skew=skewness(tax), Kurt=kurtosis(tax)))

hist <- ggplot(crime\_train, aes(tax)) + geom\_histogram(fill = 'dodgerblue', binwidth = 20, color = 'darkgray' ) +

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), tax)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

10. Response Variable: ptratio - ptratio: pupil-teacher ratio by town. The predictor ptratio indicates the average school, pupil-to-student ratio, and has a left skewed distribution. We can see a positive relationship between ptratio and high crime.

```{r}

with(crime\_train, c(summary(ptratio), SD=sd(ptratio), Skew=skewness(ptratio), Kurt=kurtosis(ptratio)))

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

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), ptratio)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

11. Response Variable: black - 1000 (Bk−0.63)2(Bk−0.63)2 where Bk is the proportion of blacks by town. This variable is heavily left skewed.

```{r}

with(crime\_train, c(summary(black), SD=sd(black), Skew=skewness(black), Kurt=kurtosis(black)))

hist <- ggplot(crime\_train, aes(black)) + geom\_histogram(fill = 'dodgerblue', binwidth = 40, color = 'darkgray' ) +

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), black)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

12. Response Variable: lstat - lower status of the population (percent). The variable lstat indicates the proportion of the population deemed to be of lower status. lstat is right skewed. High crime areas tend to have be associated with larger lstat values.

```{r}

with(crime\_train, c(summary(lstat), SD=sd(lstat), Skew=skewness(lstat), Kurt=kurtosis(lstat)))

hist <- ggplot(crime\_train, aes(lstat)) + geom\_histogram(fill = 'dodgerblue', binwidth = 2, color = 'darkgray' ) +

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), lstat)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

13. Response Variable: medv - median value of owner-occupied homes in $1000s. The median value of residential homes in a given area. The variable is slightly right skewed, and high values of medv appear to be associated with lower crime rates.

```{r}

with(crime\_train, c(summary(medv), SD=sd(medv), Skew=skewness(medv), Kurt=kurtosis(medv)))

hist <- ggplot(crime\_train, aes(medv)) + geom\_histogram(fill = 'dodgerblue', binwidth = 2, color = 'darkgray' ) +

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

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

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

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

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

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

box\_target <- ggplot(crime\_train, aes(x=factor(target), medv)) + geom\_boxplot(fill='dodgerblue', color='darkgrey') +

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

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

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

```

Finding correlations: 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.

```{r}

names(crime\_train)

cor(drop\_na(crime\_train))

```

```{r}

pairs.panels(crime\_train[1:14])

```

#DATA PREPARATION:

a. Missing Values - there are no missing values, so we will not do any missing value treatment.

b. outliers: I think we dont have any outliers that we should be removing at this stage.

c. Transformation -

age and lstat are both skewed, so lets see boxcox transformation suggestions.

```{r}

boxcoxfit(crime\_train$age)

boxcoxfit(crime\_train$lstat)

```

so for age the boxcox fit suggested power transformation of 1.3 and for lstat boxcox fit suggested power transformation of 0.23. Lets apply the same.

```{r}

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

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

```

The predictor dis, rm and medv has a moderate positive skew. Let's transform using the box-cox transformation

```{r}

boxcoxfit(crime\_train$dis)

boxcoxfit(crime\_train$rm)

boxcoxfit(crime\_train$medv)

```

so for medv and rm the boxcox fit suggested power transformation of .23. Lets apply the same

```{r}

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

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

```

The lamda for the boxcoxfit for is dis is alose to 0, so we can apply log transformation.

```{r}

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

```

Lets plot to see the status of the variables after transformation:

```{r}

hist\_medv <- ggplot(crime\_train, aes(medv\_mod)) + geom\_histogram(fill = 'dodgerblue', binwidth = 0.2, color = 'darkgray' ) +

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

hist\_lstat <- ggplot(crime\_train, aes(lstat\_mod)) + geom\_histogram(fill = 'dodgerblue', binwidth = 0.2, color = 'darkgray' ) +

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

hist\_dis <- ggplot(crime\_train, aes(dis\_mod)) + geom\_histogram(fill = 'dodgerblue', binwidth = 0.2, color = 'darkgray' ) +

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

hist\_rm <- ggplot(crime\_train, aes(rm\_mod)) + geom\_histogram(fill = 'dodgerblue', binwidth = 0.025, color = 'darkgray' ) +

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

hist\_age <- ggplot(crime\_train, aes(age\_mod)) + geom\_histogram(fill = 'dodgerblue', binwidth = 50, color = 'darkgray' ) +

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

grid.arrange(hist\_medv, hist\_lstat, hist\_dis, hist\_rm , hist\_age, ncol=2)

```

We can see that the skewness of the transformed variables improved.

#BUILD MODELS:

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.

```{r}

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

summary(model1)

```

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.

```{r}

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

summary(model2)

```

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 transfornation 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.

```{r}

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)

```

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.

```{r}

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

summary(model4)

```

#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

```{r}

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

roc\_model3 <- roc(crime\_train$target, crime\_train$predict, plot=T, asp=NA,

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

roc\_model3["auc"]

```

Now lets do the confusion matrix:

```{r}

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)

```

```{r}

Accuracy <- function(data) {

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

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(crime\_train\_cm$predict\_target,crime\_train\_cm$target)

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(crime\_train\_cm$predict\_target,crime\_train\_cm$target)

TP=tb[2,2]

FP=tb[1,2]

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

}

Precision(data)

```

```{r}

Sensitivity <- function(data) {

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

TP=tb[2,2]

FN=tb[2,1]

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

}

Sensitivity(data)

```

```{r}

Specificity <- function(data) {

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

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(crime\_train\_cm$predict\_target,crime\_train\_cm$target)

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}

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)

```

The Predicted Evaluation data is present at https://github.com/Riteshlohiya/Data621-Week3-Assignment3/blob/master/Evaluation\_Data.csv