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

Variable Name	Description	Theoretical Impact
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

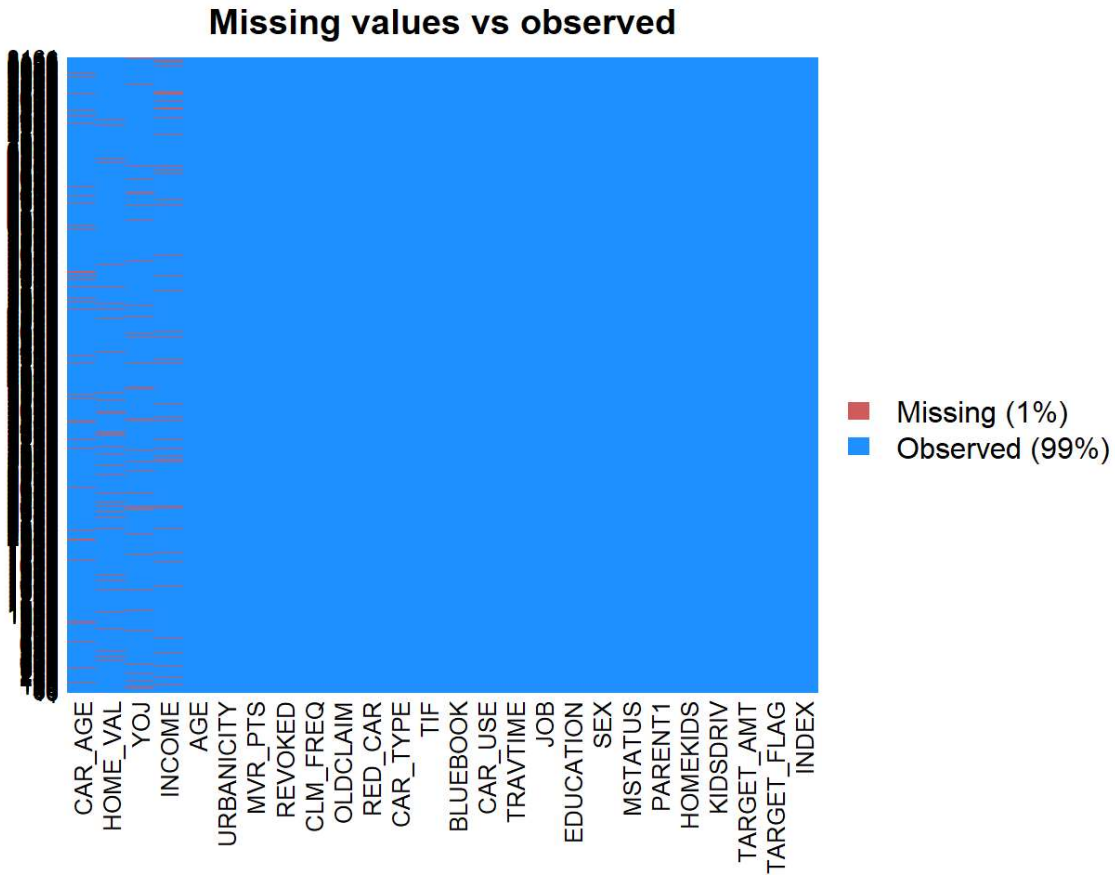
Variable Name	Description	Theoretical Impact
BLUEBOOK	Value of Vehicle	Unknown impact on collision prob, but impacts crash payout
TIF	Time in Force	Long-term customers are usually safer
CAR_TYPE	Type of Car	Unknown impact on collision prob, but impacts crash payout
RED_CAR	A Red Car	Urban legend: red 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

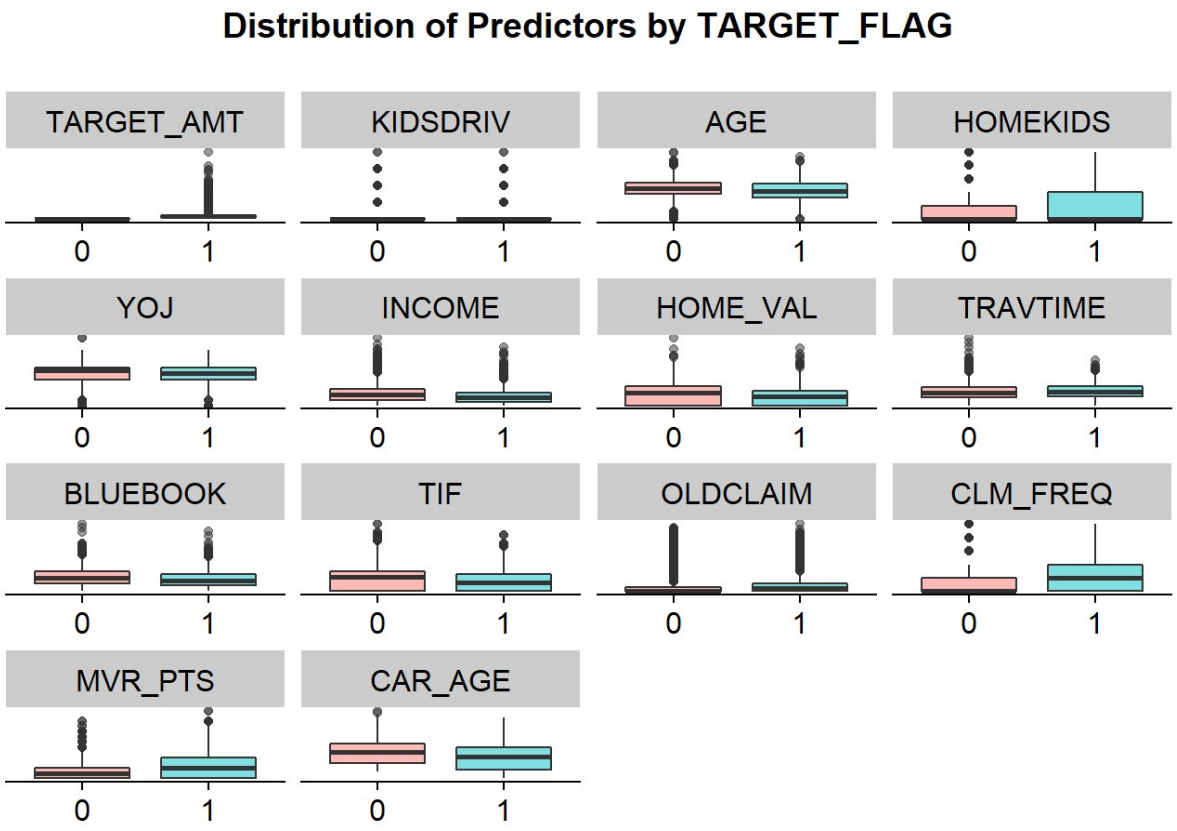
Now we will see the missing values in the dataset. For this i have used Amelia package. We can see there are missing values for CAR_AGE, HOME_VAL, YOJ and INCOME. There needs to be taken care while we do data preparation.



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.

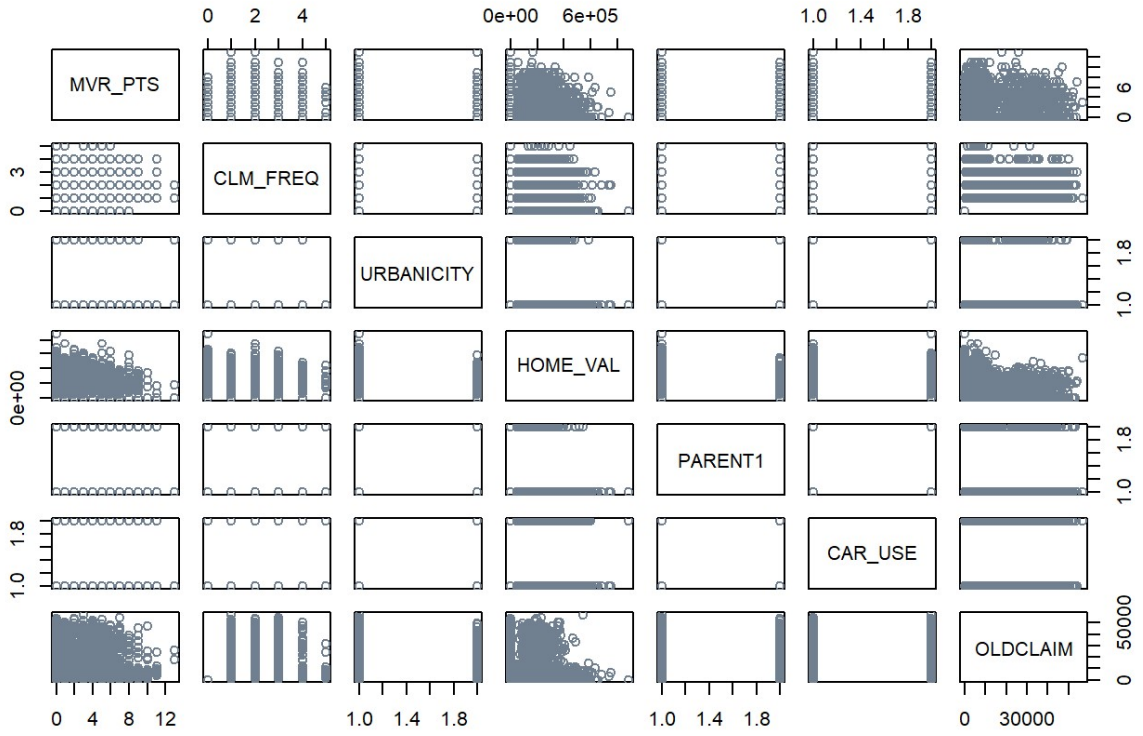
Boxplots are generated for non-binary variables split by TARGET_FLAG:



Correlation

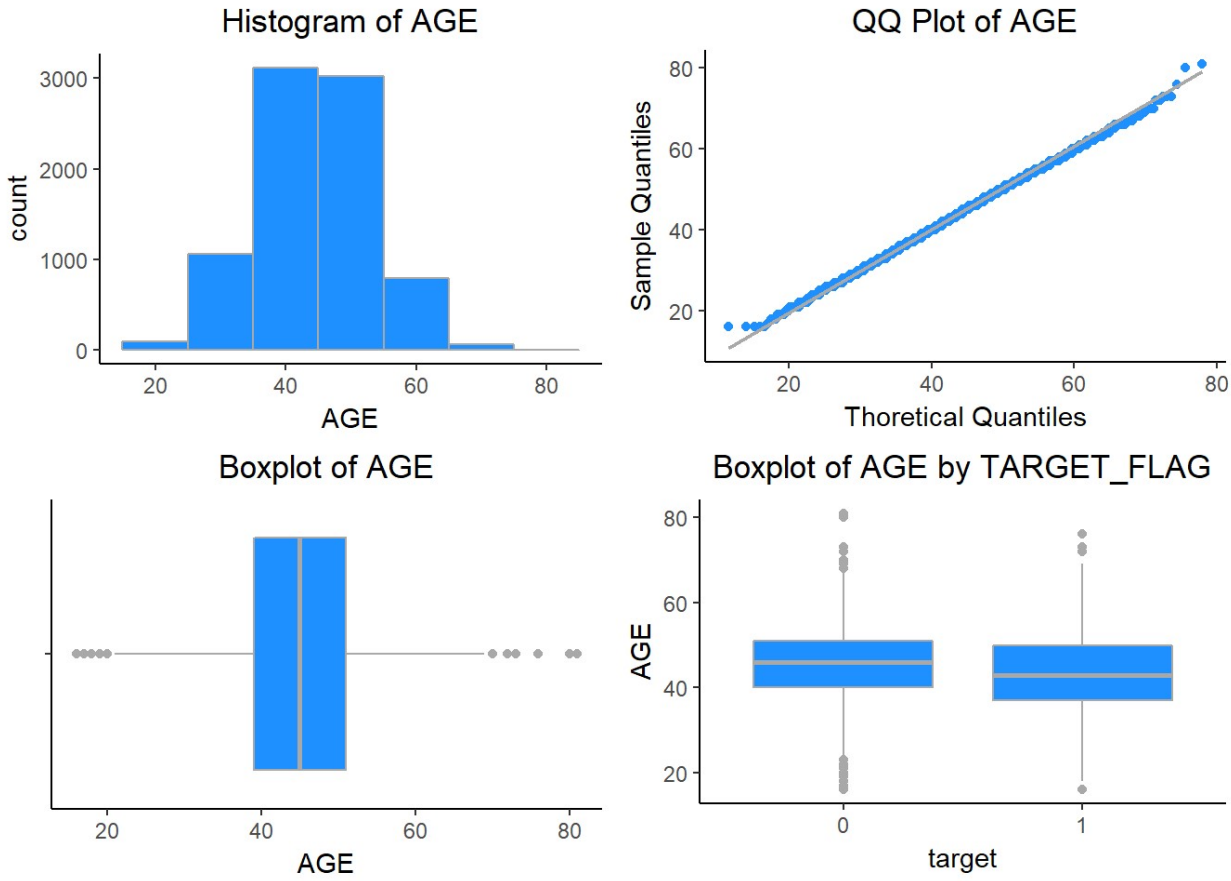
The correlation plot below shows how variables in the dataset are related to each other. Looking at the plot, we can see that certain variables are more related than others.

Predictors with High Correlattions to Targets

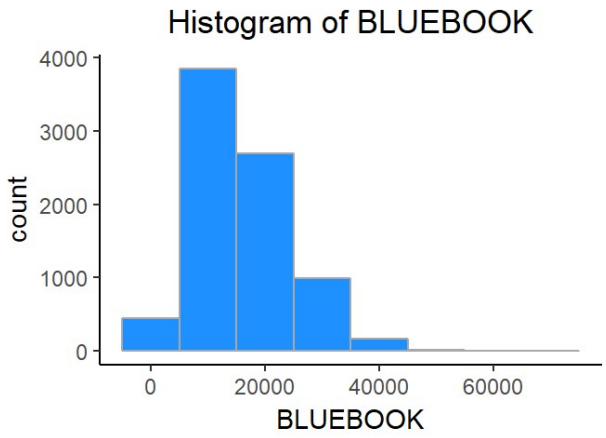


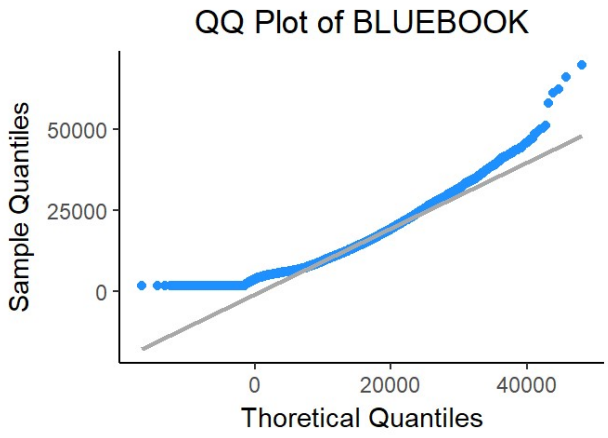
Now let's do some plots to understand the data:

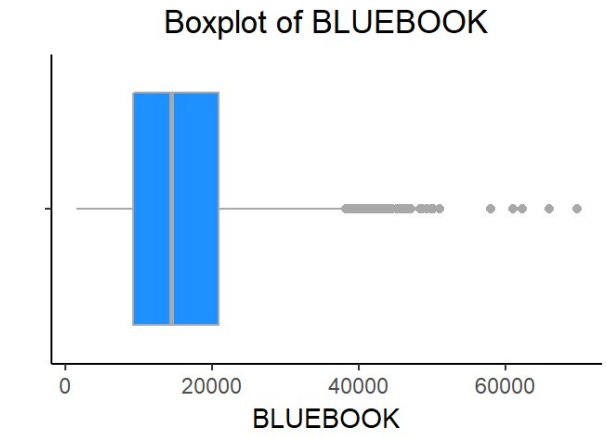
AGE - Age of Driver. Very young people tend to be risky. Maybe very old people also. We note six missing values that we'll need to address later. The distribution of AGE is almost perfectly normal. When we break out the data by TARGET_FLAG values, the distributions of age by TARGET_FLAG are still roughly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

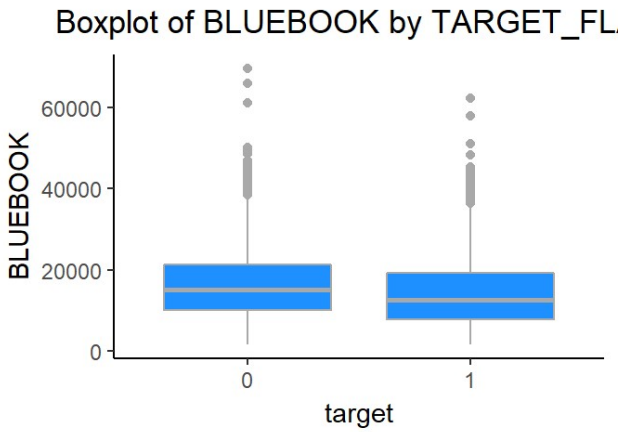


BLUEBOOK - Value of Vehicle. Unknown effect on probability of collision, but probably effect the payout if there is a crash. Individuals involved in crashes have a higher proportion of low BLUEBOOK values.

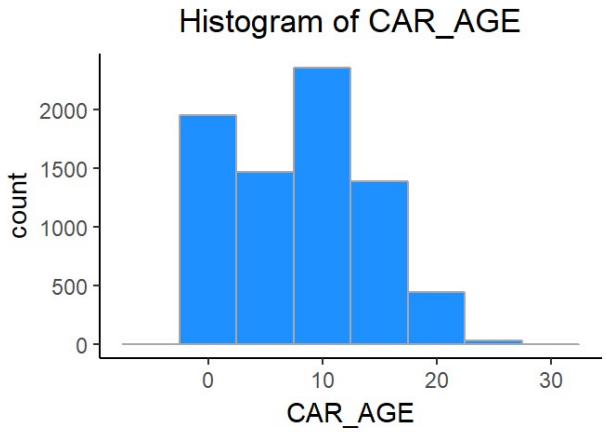


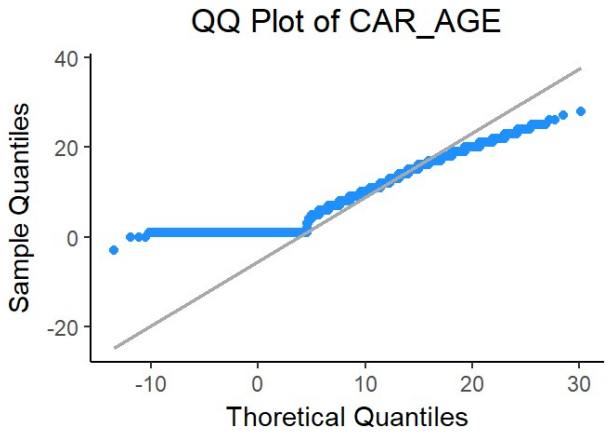


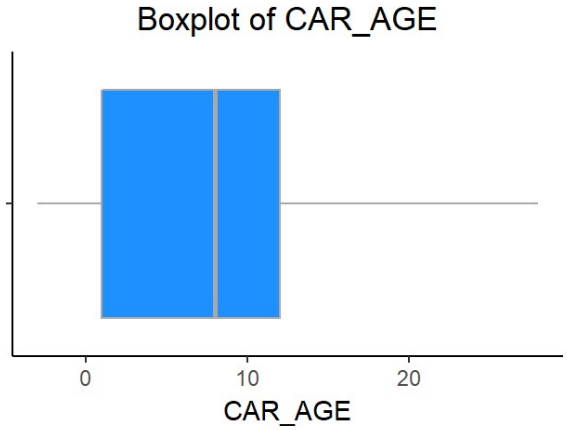


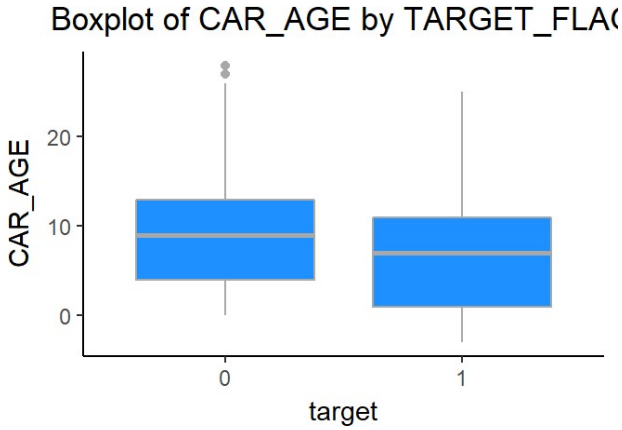


CAR_AGE - Vehicle Age. We could see there is one negative value for CAR_AGE. We have to treat this value in our data preparation step.

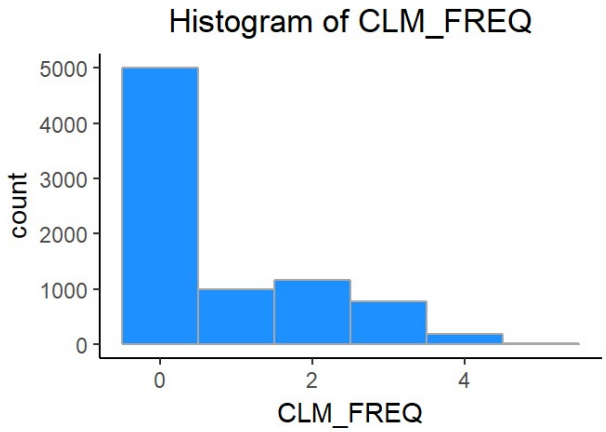


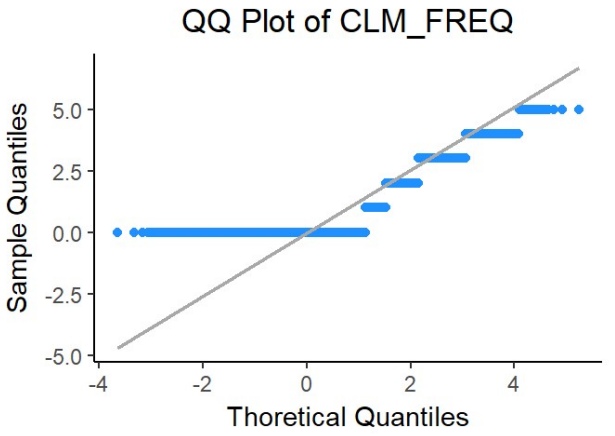


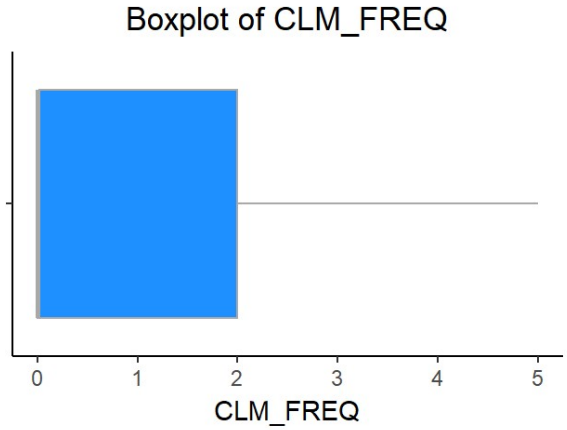


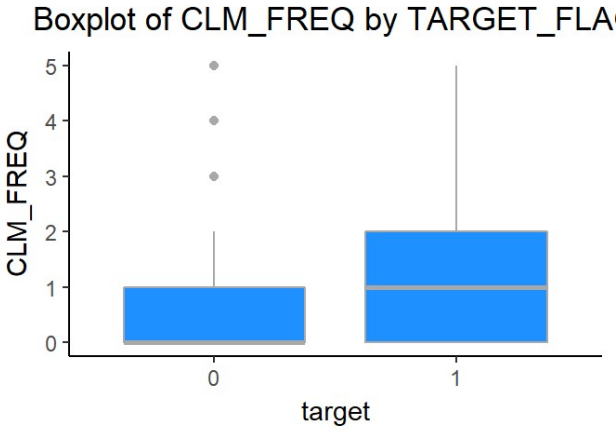


CLM_FREQ - # Claims (Past 5 Years). The more claims you filed in the past, the more you are likely to file in the future. We can see that this variable is also skewed.

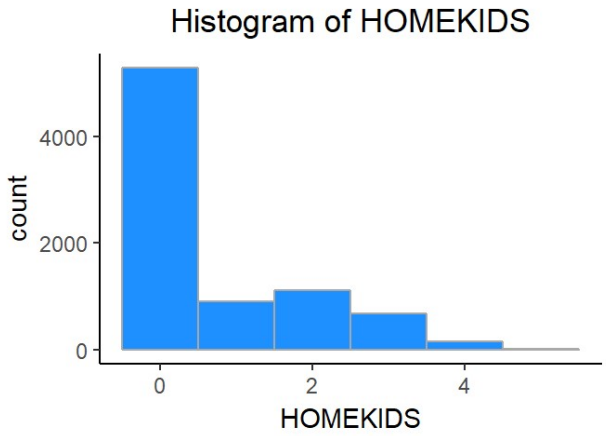


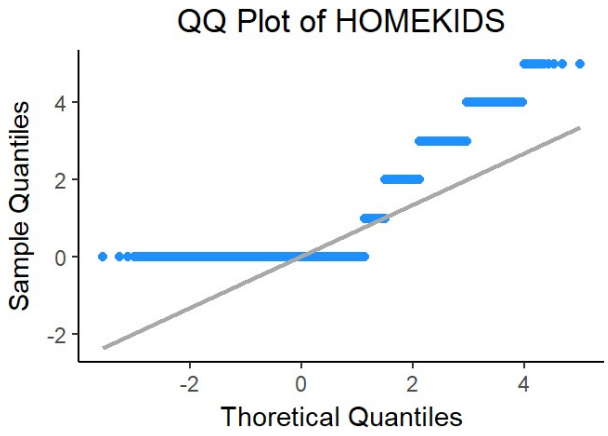


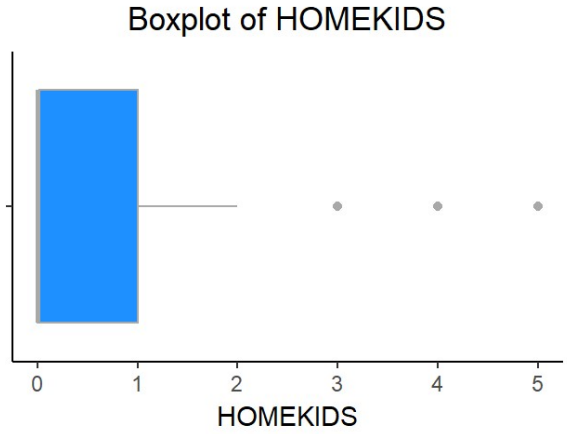


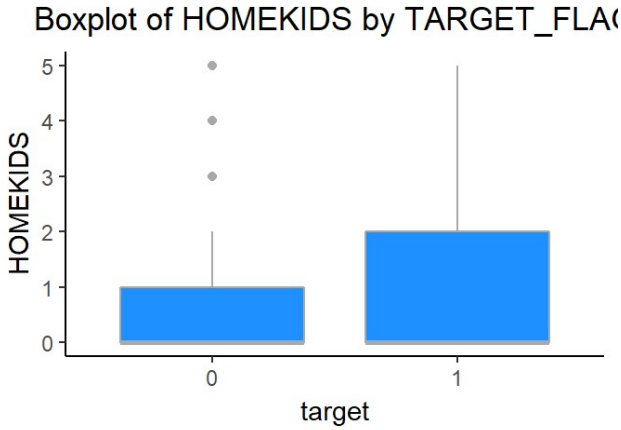


HOMEKIDS - # Children at Home. HOMEKIDS does not seem to impact the TARGET_FLAG. The distribution of this discrete variable is right skewed.

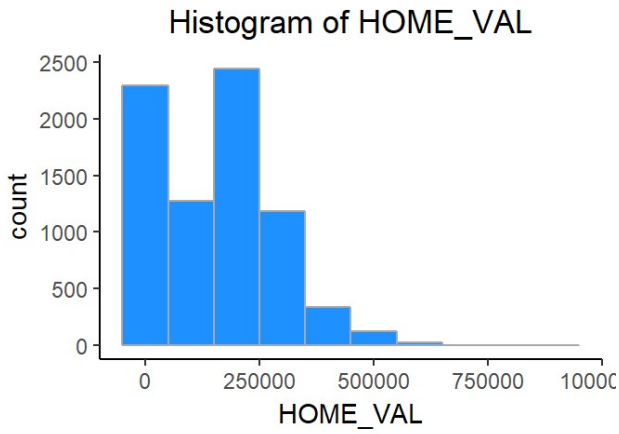


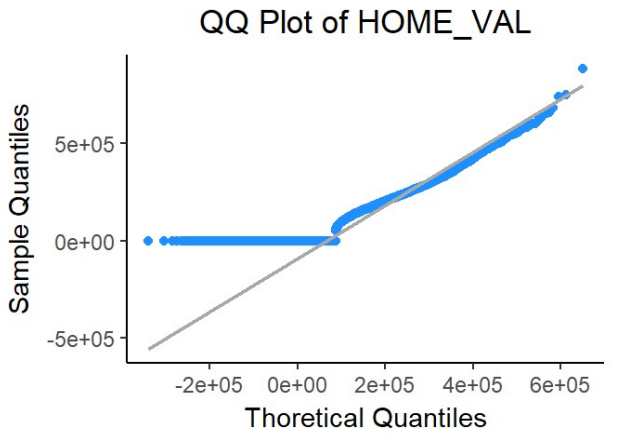


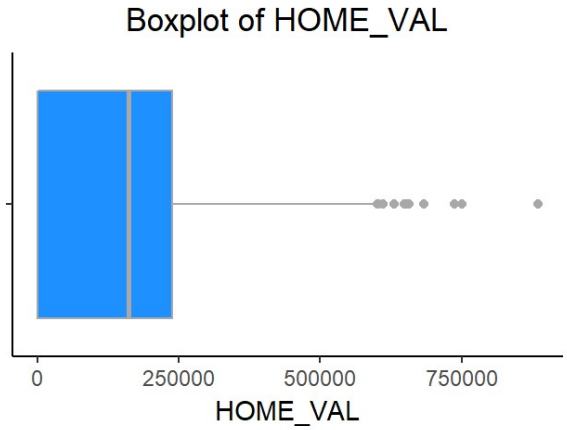


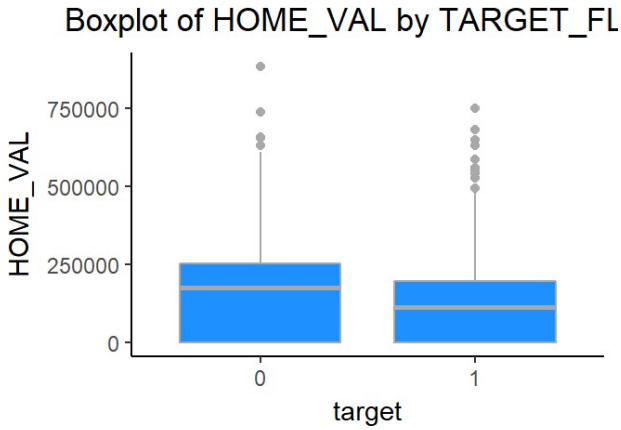


HOME_VAL - Home Value. Home owners tend to drive more responsibly. The distribution of HOME_VAL is right skewed and also we can there are some missing values.

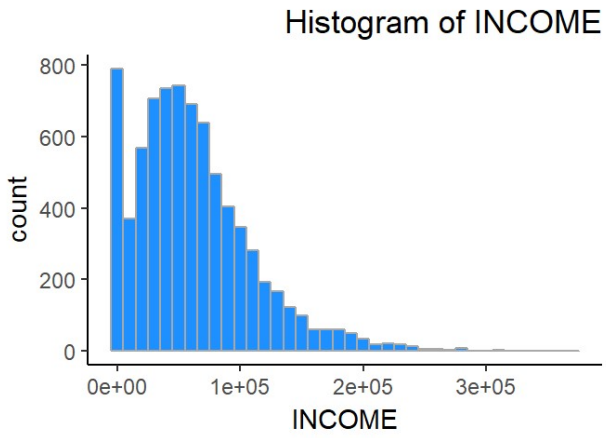


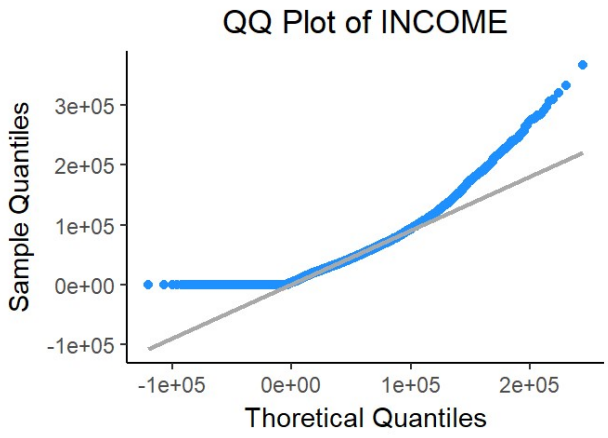


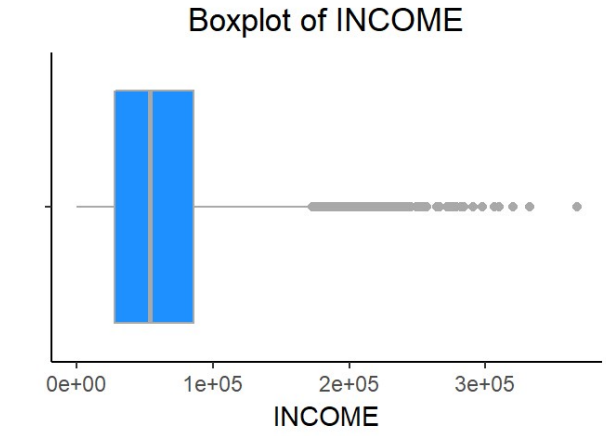


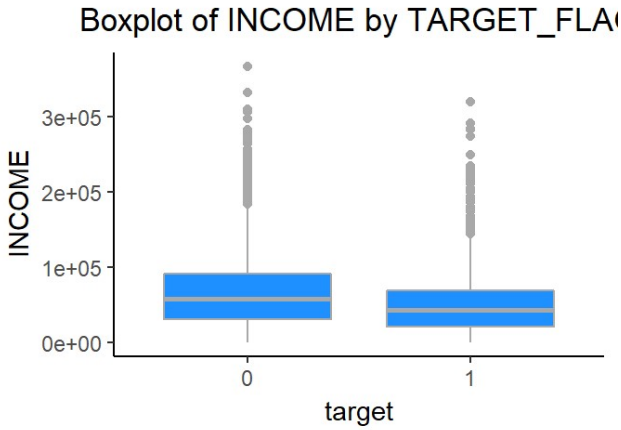


INCOME - Income of the person. Rich people tend to get into fewer crashes. The distribution of INCOME is right skewed, with a significant number of observations indicating $0 in income. There are some missing values in this as well.

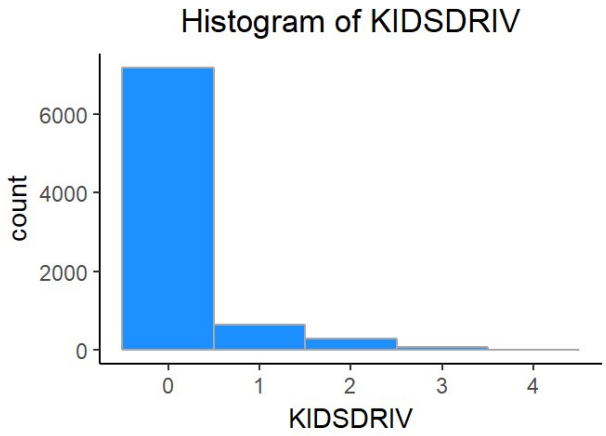


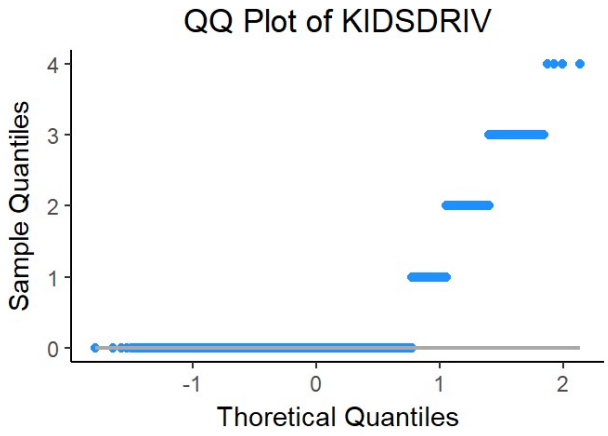


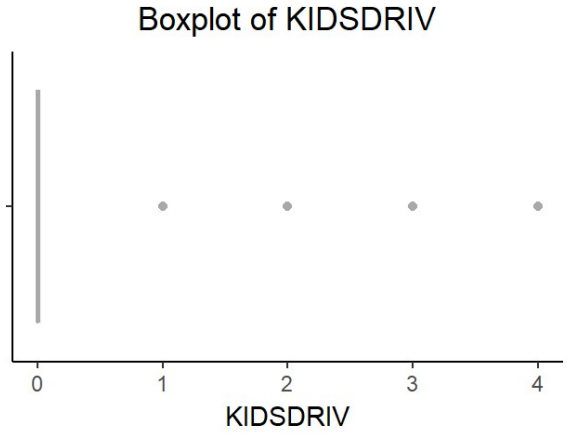


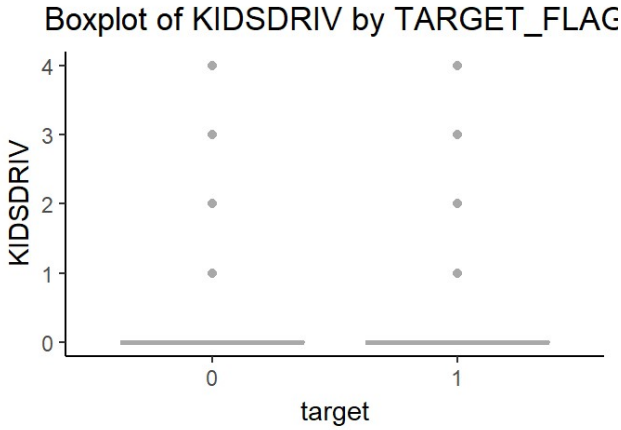


KIDSDRIV - # Driving Children. When teenagers drive your car, you are more likely to get into crashes. The discrete variable KIDSDRIV is right skewed

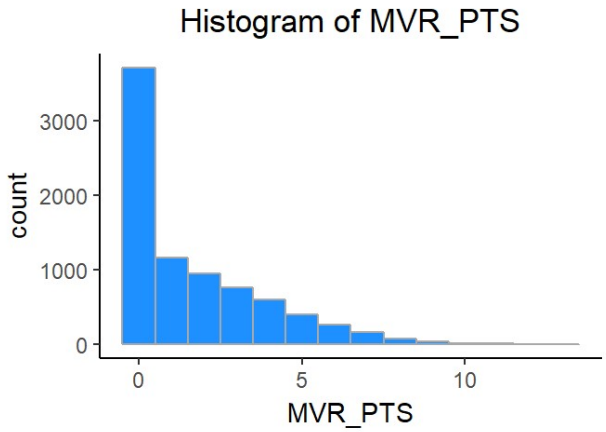


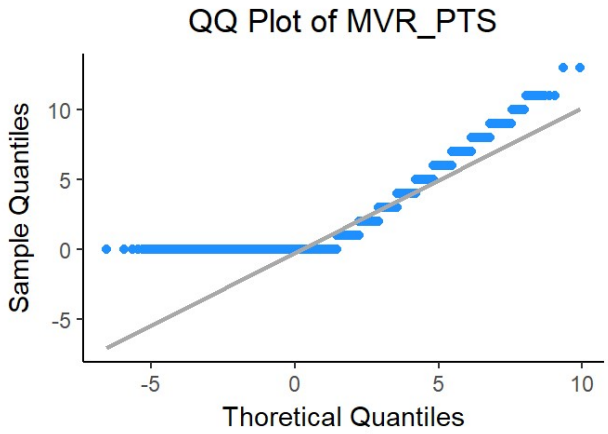


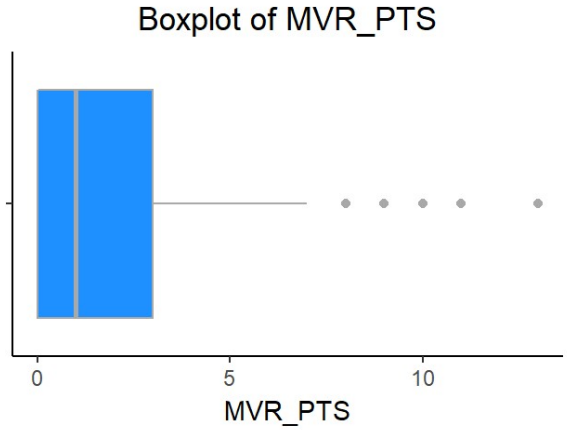


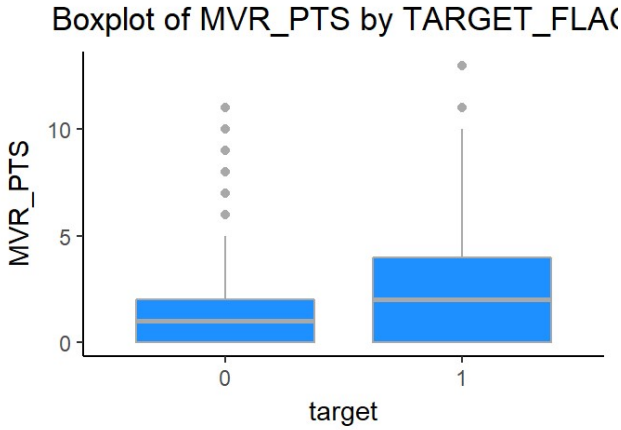


MVR_PTS - Motor Vehicle Record Points. If you get lots of traffic tickets, you tend to get into more crashes. MVR_PTS is positively skewed.

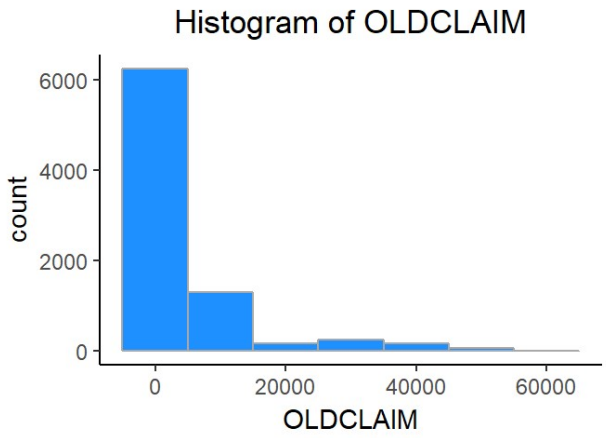


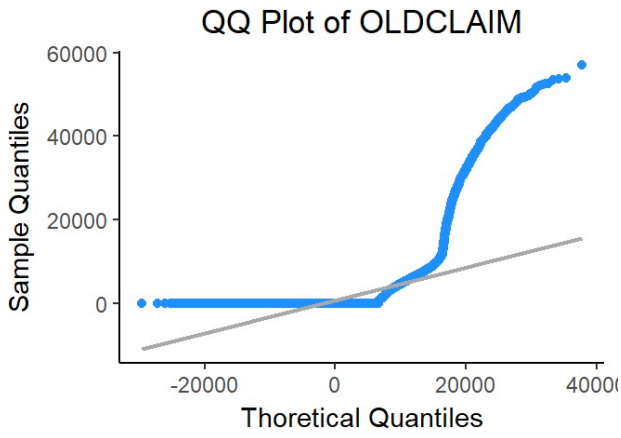


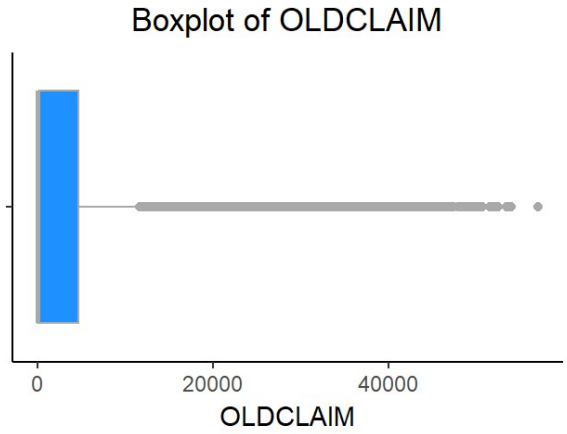


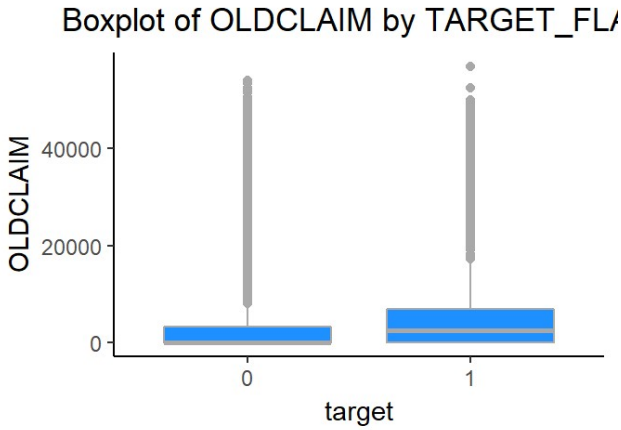


OLDCLAIM - Total Claims (Past 5 Years). If your total payout over the past five years was high, this suggests future payouts will be high. The distribution of OLDCLAIM is extremely right skewed.

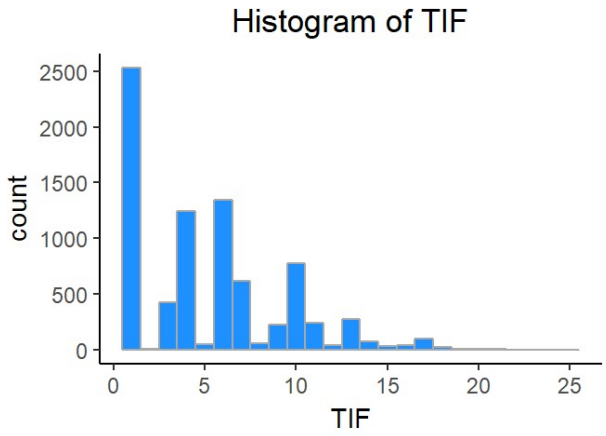


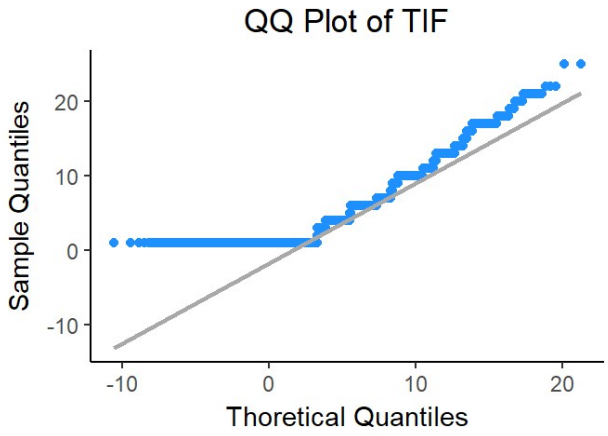


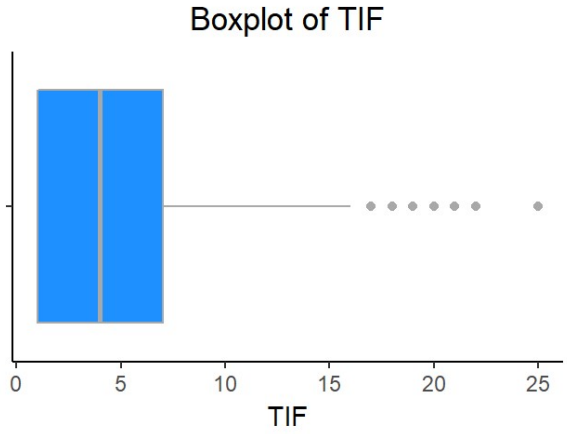


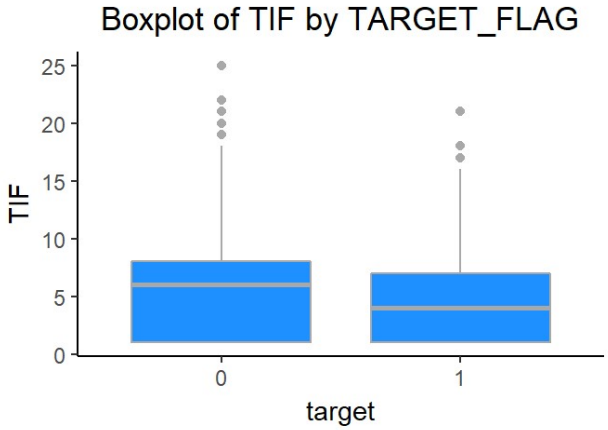


TIF - Time in Force. People who have been customers for a long time are usually more safe. The distribution is somewhat positively skewed.

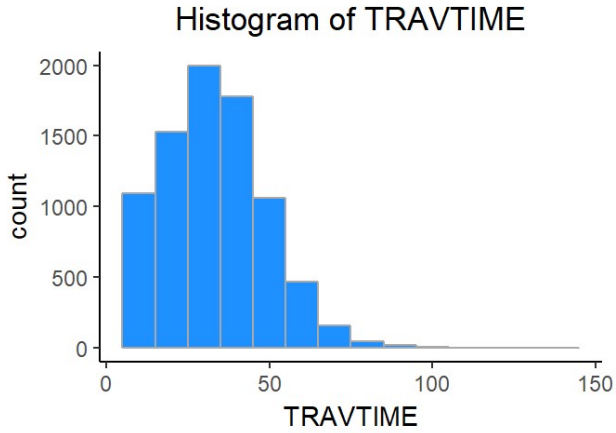


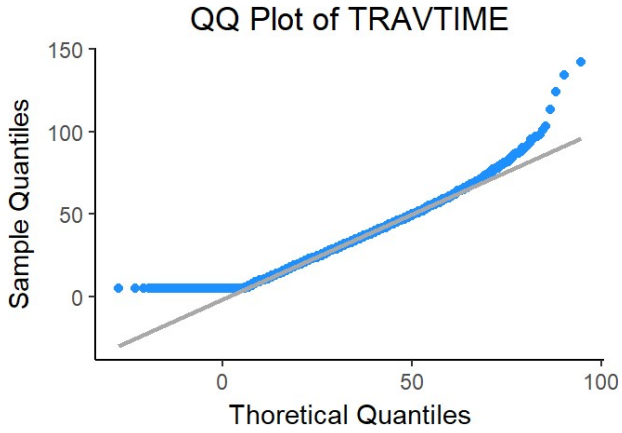


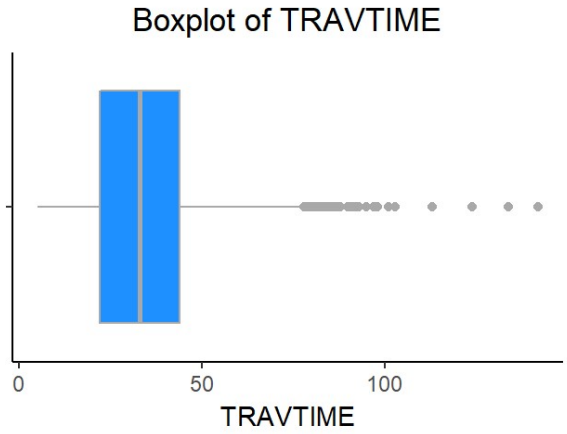


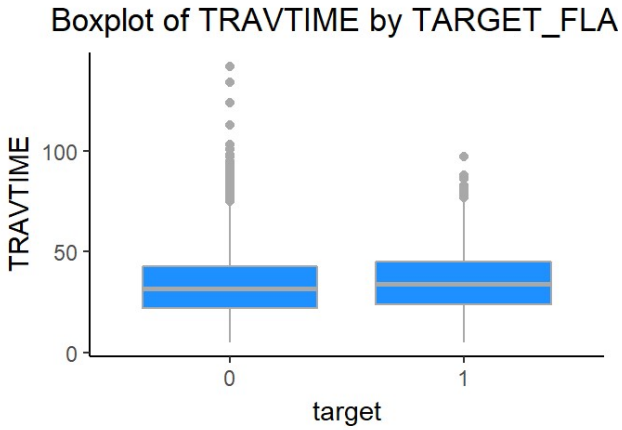


TRAVTIME - Distance to Work. Long drives to work usually suggest greater risk. The distribution has a slight positive skew. The subset of insureds with no accidents have a higher proportion of individuals with short commute times.

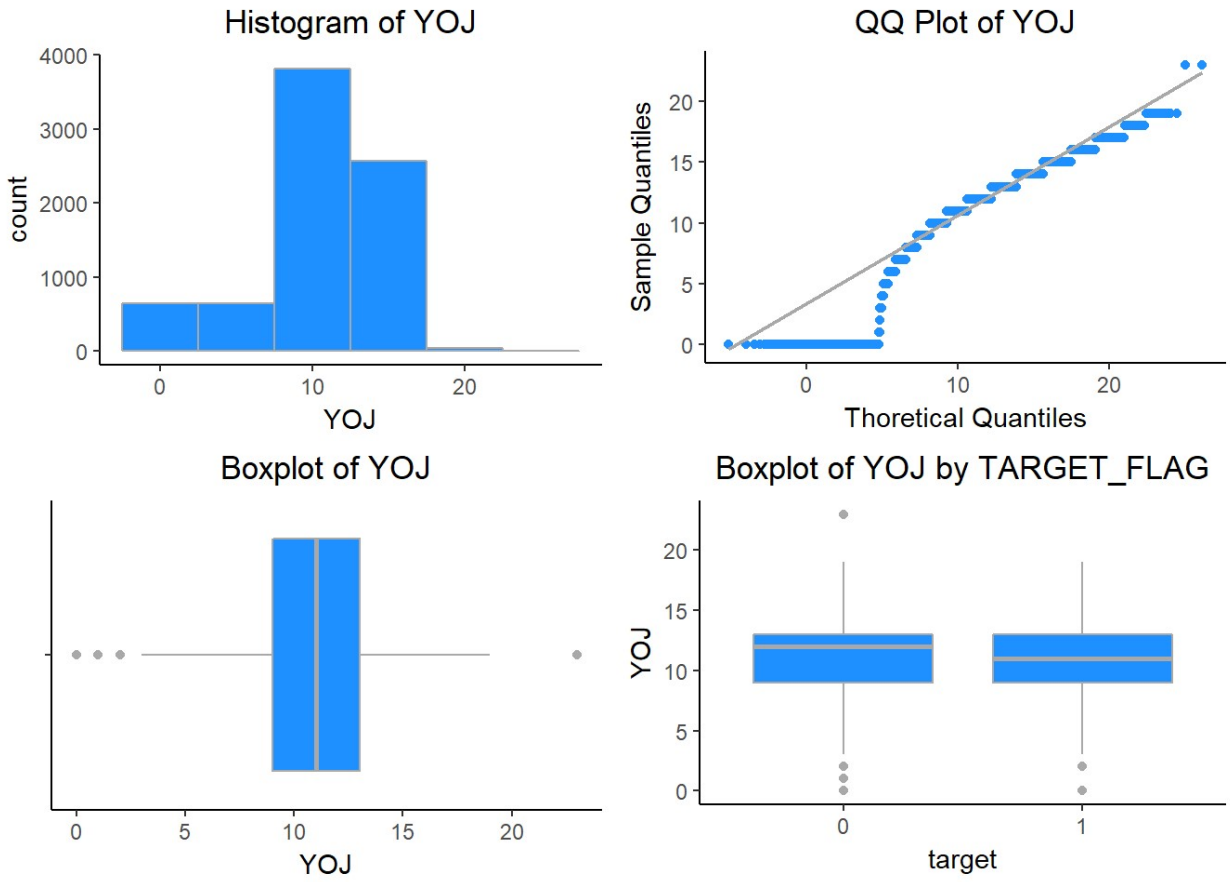








YOJ - Years on Job. People who stay at a job for a long time are usually more safe. The variable would be approximately normally distributed if it weren't for the high percentage of individuals with less than one year on the job.



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
```

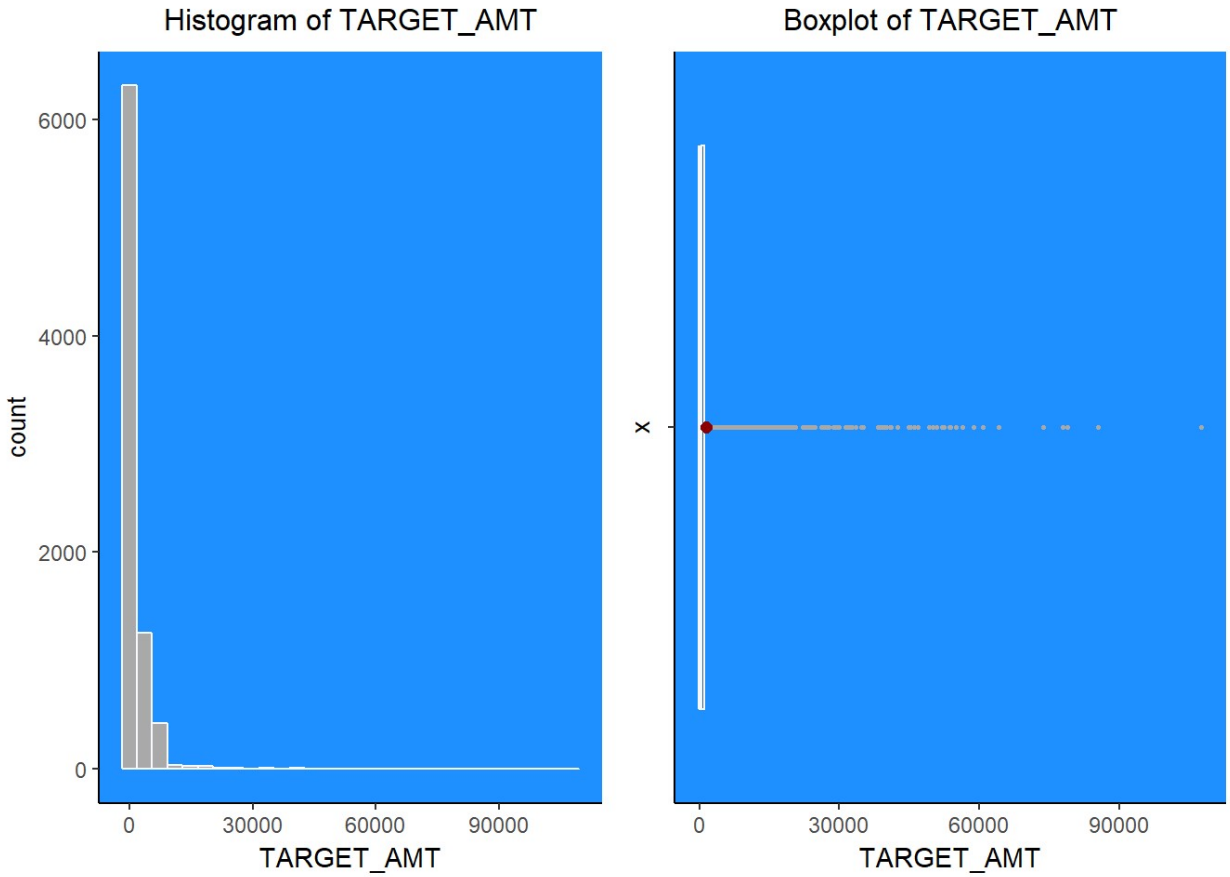
TARGET Variables

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

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.

```
train_amount <- ins_train[,-c(1)] #Training dataset with response of claim am
ount

amount_full_model1 <- lm(TARGET_AMT ~., data = train_amount)

summary(amount_full_model1)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = train_amount)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5894  -1698   -764    359 103840
##
## Coefficients: (2 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -8.847e+02  7.305e+02  -1.211 0.225919
## KIDSDRIV       3.159e+02  1.132e+02   2.790 0.005282 **
## AGE            4.905e+00  7.091e+00   0.692 0.489098
## HOMEKIDS       7.569e+01  6.576e+01   1.151 0.249771
## YOJ           -2.644e+00  1.707e+01  -0.155 0.876919
```

```
## INCOME        -4.386e-03  4.002e-03  -1.096 0.273171
## PARENT1        5.697e+02  2.024e+02   2.815 0.004885 **
## HOME_VAL      -1.486e-04  1.094e-03  -0.136 0.891922
## MSTATUS       -5.541e+02  1.498e+02  -3.698 0.000218 ***
## SEX            3.346e+02  1.625e+02   2.059 0.039497 *
## TRAVTIME       1.196e+01  3.223e+00   3.713 0.000207 ***
## CAR_USE        8.086e+02  1.630e+02   4.961 7.14e-07 ***
## BLUEBOOK       3.550e-03  3.328e-02   0.107 0.915038
## TIF           -4.800e+01  1.218e+01  -3.940 8.23e-05 ***
## RED_CAR               NA         NA      NA       NA
## OLDCLAIM      -1.847e-02  8.994e-03  -2.053 0.040078 *
## CLM_FREQ       3.730e+01  8.734e+01   0.427 0.669341
## REVOKED        5.906e+02  1.755e+02   3.366 0.000766 ***
## MVR_PTS        1.648e+02  2.672e+01   6.170 7.18e-10 ***
## CAR_AGE       -2.682e+01  1.280e+01  -2.096 0.036147 *
## URBANICITY     1.618e+03  1.405e+02  11.514  < 2e-16 ***
## HSDropout      1.071e+02  1.725e+02   0.621 0.534564
## Bachelors     -1.978e+02  1.570e+02  -1.260 0.207677
## Masters       -8.223e+01  2.302e+02  -0.357 0.720910
## PhD            1.485e+02  2.933e+02   0.506 0.612678
## Panel_Truck    1.720e+02  2.766e+02   0.622 0.533880
## Pickup         3.426e+02  1.696e+02   2.020 0.043410 *
## Sports_Car     1.016e+03  2.181e+02   4.659 3.22e-06 ***
## Van            4.622e+02  2.115e+02   2.186 0.028875 *
## SUV            7.388e+02  1.795e+02   4.116 3.89e-05 ***
## Professional   7.615e+01  1.961e+02   0.388 0.697813
## Blue_Collar           NA         NA      NA       NA
## Clerical       7.619e+01  1.899e+02   0.401 0.688277
## Doctor        -6.970e+02  3.888e+02  -1.793 0.073074 .
## Lawyer        -9.410e+00  2.569e+02  -0.037 0.970778
## Manager       -8.037e+02  2.031e+02  -3.957 7.64e-05 ***
## Home_Maker    -6.325e+01  2.908e+02  -0.218 0.827819
## Student       -2.296e+02  2.839e+02  -0.809 0.418678
## INCOME_MOD    -7.343e-01  4.403e+00  -0.167 0.867547
```

```
## HOME_VAL_MOD  -2.905e+01  7.127e+01  -0.408 0.683578
## BLUEBOOK_MOD   4.230e+00  1.238e+01   0.342 0.732642
## OLD_CLAIM_MOD  4.485e+01  2.920e+01   1.536 0.124545
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4545 on 8121 degrees of freedom
## Multiple R-squared:  0.07105,    Adjusted R-squared:  0.06659
## F-statistic: 15.93 on 39 and 8121 DF,  p-value: < 2.2e-16
```

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.

```
amount_reduced_model2 <- update(amount_full_model1, .~.-HSDropout-Home_Maker-
Bachelors-Masters-PhD-Panel_Truck-Blue_Collar-Professional-Student-HOMEKIDS-C
AR_AGE-YOJ-Lawyer-SEX-AGE-Doctor-Clerical-INCOME-HOME_VAL-BLUEBOOK-RED_CAR--C
LM_FREQ-INCOME_MOD-HOME_VAL_MOD-BLUEBOOK_MOD-OLD_CLAIM_MOD-OLDCLAIM)
summary(amount_reduced_model2)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + PARENT1 + MSTATUS + TRAVTIME +
##     CAR_USE + TIF + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY +
##     Pickup + Sports_Car + Van + SUV + Manager, data = train_amount)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5685  -1698   -800    304 103866
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -669.790    217.622  -3.078 0.002093 **
## KIDSDRIV     393.362    102.187   3.849 0.000119 ***
## PARENT1      730.172    175.662   4.157 3.26e-05 ***
## MSTATUS     -534.231    118.994  -4.490 7.23e-06 ***
## TRAVTIME      12.187      3.225   3.779 0.000158 ***
```

```
## CAR_USE      899.712    112.672   7.985 1.59e-15 ***
## TIF          -46.869     12.184  -3.847 0.000121 ***
## CLM_FREQ     127.463     48.762   2.614 0.008965 **
## REVOKED      473.061    155.100   3.050 0.002296 **
## MVR_PTS      178.285     25.813   6.907 5.33e-12 ***
## URBANICITY  1467.231    134.707  10.892  < 2e-16 ***
## Pickup       317.736    151.471   2.098 0.035965 *
## Sports_Car   790.710    176.754   4.473 7.80e-06 ***
## Van          437.232    188.775   2.316 0.020574 *
## SUV          514.705    131.086   3.926 8.69e-05 ***
## Manager     -961.429    159.097  -6.043 1.58e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4556 on 8145 degrees of freedom
## Multiple R-squared:  0.06366,    Adjusted R-squared:  0.06194
## F-statistic: 36.92 on 15 and 8145 DF,  p-value: < 2.2e-16
```

Interpretation of the Model1:

The Residual standard error is 4545

Multiple R-squared: 0.07105

Adjusted R-squared: 0.06659

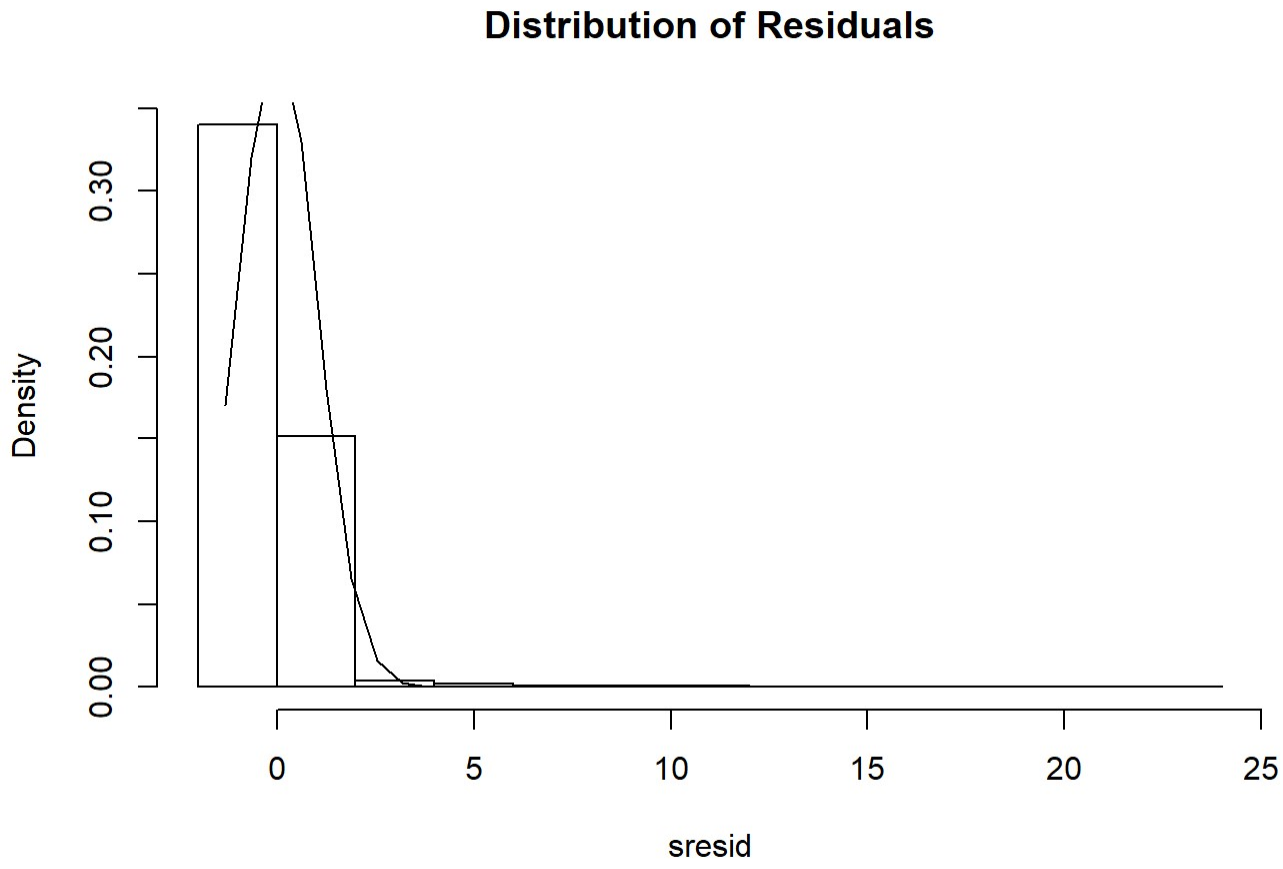
F-statistic: 15.93 on 39 and 8121 DF

p-value: < 2.2e-16

Analysis of plot on residuals to verify normal distribution of residuals

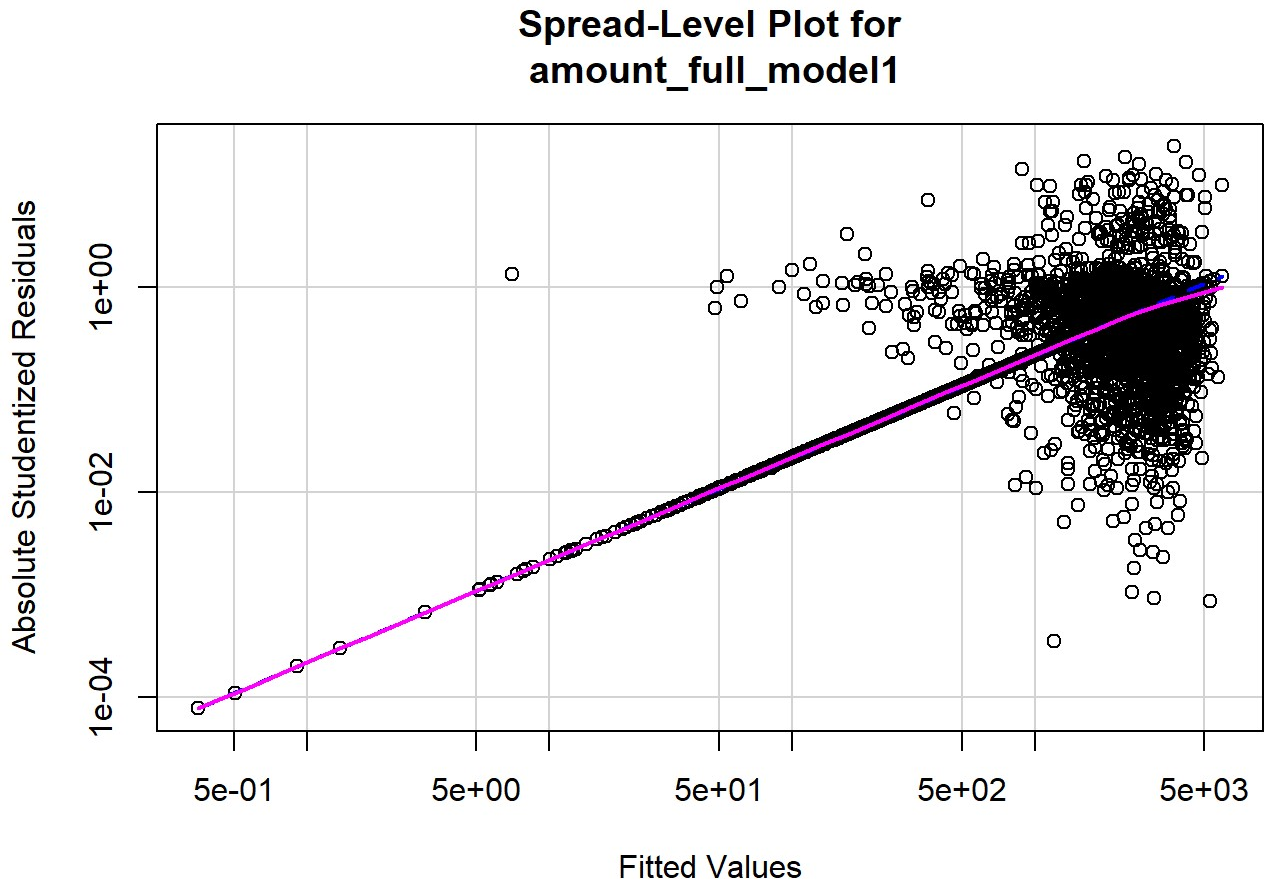
```
sresid <- studres(amount_full_model1)
hist(sresid, freq=FALSE,
     main="Distribution of Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
```

```
lines(xfit, yfit)
```



Check for Homoscedasticity:

```
ncvTest(amount_full_model1)
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3209.905    Df = 1     p = 0
spreadLevelPlot(amount_full_model1)
## Warning in spreadLevelPlot.lm(amount_full_model1):
## 1011 negative fitted values removed
```




```
##
## Suggested power transformation:  0.000210097
```

Interpretation of the Model2:

The Residual standard error is 4556

Multiple R-squared: 0.06366

Adjusted R-squared: 0.06194

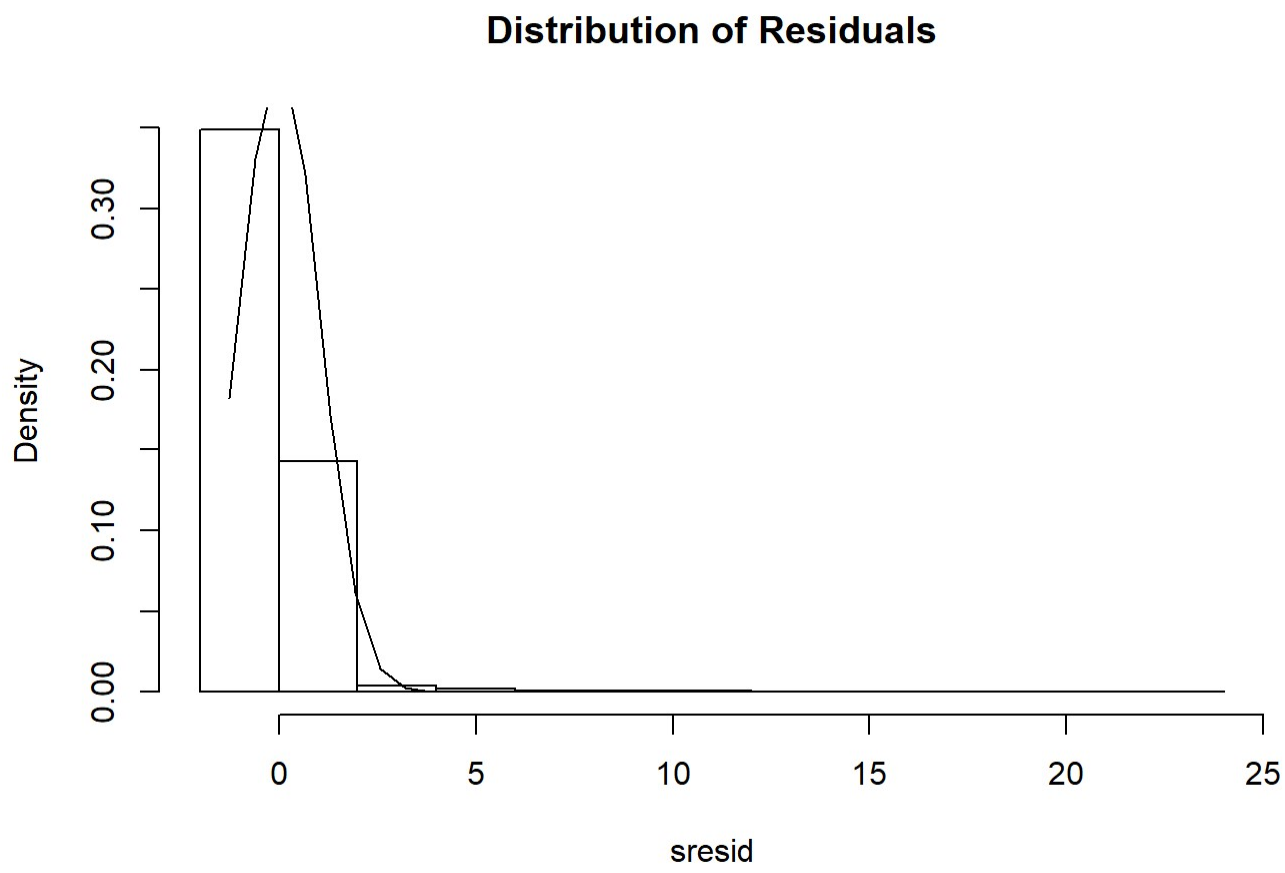
F-statistic: 36.92 on 15 and 8145 DF

p-value: < 2.2e-16

Analysis of plot on residuals to verify normal distribution of residuals

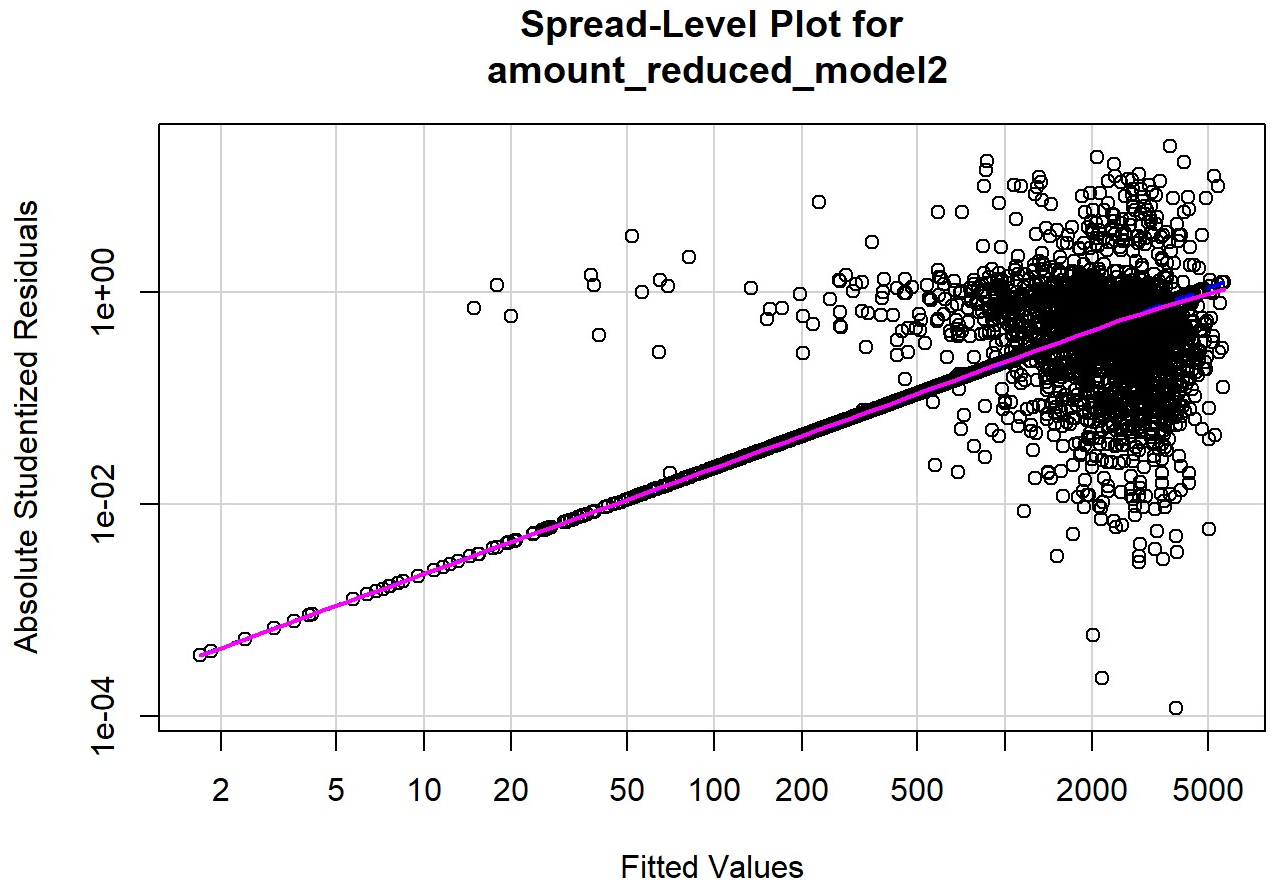
```
sresid <- studres(amount_reduced_model2)
hist(sresid, freq=FALSE,
     main="Distribution of Residuals")
```

```
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



Check for Homoscedasticity:

```
ncvTest(amount_reduced_model2)
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2762.247    Df = 1     p = 0
spreadLevelPlot(amount_reduced_model2)
## Warning in spreadLevelPlot.lm(amount_reduced_model2):
## 858 negative fitted values removed
```




```
##
## Suggested power transformation:  7.842944e-05
```

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.

```
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)
```

```
##
## Call:
```

```
## glm(formula = TARGET_FLAG ~ . - INCOME_MOD - HOME_VAL_MOD - BLUEBOOK_MOD -
##     OLD_CLAIM_MOD, family = binomial(link = "logit"), data = train_flag)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5940  -0.7152  -0.3989   0.6316   3.1401
##
## Coefficients: (2 not defined because of singularities)
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.360e+00  2.819e-01 -11.920  < 2e-16 ***
## KIDSDRIV      3.890e-01  6.119e-02   6.357 2.06e-10 ***
## AGE          -1.005e-03  4.018e-03  -0.250 0.802461
## HOMEKIDS      4.965e-02  3.712e-02   1.337 0.181061
## YOJ          -1.100e-02  8.585e-03  -1.281 0.200271
## INCOME       -3.587e-06  1.078e-06  -3.326 0.000880 ***
## PARENT1       3.813e-01  1.095e-01   3.481 0.000499 ***
## HOME_VAL     -1.304e-06  3.418e-07  -3.815 0.000136 ***
## MSTATUS      -4.937e-01  8.354e-02  -5.910 3.41e-09 ***
## SEX           7.723e-02  1.002e-01   0.771 0.440722
## TRAVTIME      1.464e-02  1.882e-03   7.778 7.36e-15 ***
## CAR_USE       7.769e-01  9.101e-02   8.537  < 2e-16 ***
## BLUEBOOK     -2.074e-05  5.258e-06  -3.945 7.99e-05 ***
## TIF          -5.556e-02  7.344e-03  -7.566 3.85e-14 ***
## RED_CAR              NA         NA      NA       NA
## OLDCLAIM     -1.397e-05  3.910e-06  -3.574 0.000352 ***
## CLM_FREQ      1.958e-01  2.853e-02   6.864 6.68e-12 ***
## REVOKED       8.876e-01  9.132e-02   9.719  < 2e-16 ***
## MVR_PTS       1.130e-01  1.360e-02   8.308  < 2e-16 ***
## CAR_AGE      -5.727e-04  7.548e-03  -0.076 0.939521
## URBANICITY    2.386e+00  1.129e-01  21.137  < 2e-16 ***
## HSDropout    -6.142e-03  9.479e-02  -0.065 0.948336
## Bachelors    -4.135e-01  8.908e-02  -4.642 3.46e-06 ***
## Masters      -4.523e-01  1.340e-01  -3.375 0.000738 ***
## PhD          -3.481e-01  1.715e-01  -2.030 0.042385 *
```

```
## Panel_Truck   5.037e-01  1.582e-01   3.184 0.001454 **
## Pickup        5.326e-01  9.999e-02   5.327 1.00e-07 ***
## Sports_Car    1.024e+00  1.299e-01   7.888 3.08e-15 ***
## Van           5.906e-01  1.254e-01   4.709 2.49e-06 ***
## SUV           7.670e-01  1.112e-01   6.897 5.30e-12 ***
## Professional -7.245e-02  1.109e-01  -0.654 0.513400
## Blue_Collar          NA         NA      NA       NA
## Clerical      1.324e-01  1.051e-01   1.260 0.207523
## Doctor       -5.477e-01  2.599e-01  -2.108 0.035069 *
## Lawyer       -1.933e-02  1.516e-01  -0.127 0.898557
## Manager      -7.582e-01  1.229e-01  -6.168 6.91e-10 ***
## Home_Maker   -1.721e-02  1.483e-01  -0.116 0.907601
## Student      -7.218e-02  1.282e-01  -0.563 0.573549
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7300.5  on 8125  degrees of freedom
## AIC: 7372.5
##
## Number of Fisher Scoring iterations: 5
```

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

```
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)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial(link = "logit"),
##     data = train_flag)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.5796  -0.7115  -0.3909   0.6205   3.1550
##
## Coefficients: (2 not defined because of singularities)
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.911e+00  4.179e-01  -4.574 4.79e-06 ***
## KIDSDRIV       3.985e-01  6.149e-02   6.481 9.14e-11 ***
## AGE           -2.086e-03  4.043e-03  -0.516 0.605950
## HOMEKIDS       3.398e-02  3.753e-02   0.905 0.365297
## YOJ            5.434e-03  9.881e-03   0.550 0.582369
## INCOME         1.837e-06  2.374e-06   0.774 0.439039
## PARENT1        3.714e-01  1.101e-01   3.373 0.000745 ***
## HOME_VAL      -3.082e-07  6.582e-07  -0.468 0.639657
## MSTATUS       -4.781e-01  8.679e-02  -5.509 3.62e-08 ***
## SEX            1.092e-01  1.007e-01   1.084 0.278163
## TRAVTIME       1.480e-02  1.891e-03   7.825 5.06e-15 ***
## CAR_USE        7.736e-01  9.148e-02   8.456  < 2e-16 ***
## BLUEBOOK       4.075e-05  1.969e-05   2.069 0.038513 *
## TIF           -5.478e-02  7.364e-03  -7.439 1.01e-13 ***
## RED_CAR               NA         NA      NA       NA
## OLDCLAIM      -2.615e-05  4.774e-06  -5.476 4.35e-08 ***
## CLM_FREQ       4.548e-02  4.412e-02   1.031 0.302637
## REVOKED        9.527e-01  9.305e-02  10.238  < 2e-16 ***
## MVR_PTS        9.540e-02  1.410e-02   6.768 1.31e-11 ***
## CAR_AGE        2.274e-05  7.562e-03   0.003 0.997600
## URBANICITY     2.361e+00  1.141e-01  20.700  < 2e-16 ***
## HSDropout     -4.795e-02  9.597e-02  -0.500 0.617378
## Bachelors     -4.035e-01  8.963e-02  -4.501 6.75e-06 ***
## Masters       -4.650e-01  1.347e-01  -3.451 0.000558 ***
## PhD           -4.661e-01  1.747e-01  -2.668 0.007625 **
## Panel_Truck    3.581e-01  1.626e-01   2.202 0.027696 *
## Pickup         5.378e-01  1.004e-01   5.358 8.40e-08 ***
## Sports_Car     1.024e+00  1.304e-01   7.855 4.01e-15 ***
## Van            5.956e-01  1.255e-01   4.745 2.08e-06 ***
```

```
## SUV             7.944e-01  1.122e-01   7.083 1.41e-12 ***
## Professional   -8.353e-02  1.111e-01  -0.752 0.452119
## Blue_Collar            NA         NA      NA       NA
## Clerical        8.731e-02  1.061e-01   0.823 0.410447
## Doctor         -5.135e-01  2.588e-01  -1.984 0.047251 *
## Lawyer         -2.688e-02  1.522e-01  -0.177 0.859771
## Manager        -7.676e-01  1.230e-01  -6.238 4.42e-10 ***
## Home_Maker     -2.989e-01  1.704e-01  -1.754 0.079346 .
## Student        -4.639e-01  1.590e-01  -2.917 0.003531 **
## INCOME_MOD     -8.396e-03  2.592e-03  -3.240 0.001197 **
## HOME_VAL_MOD   -6.994e-02  4.124e-02  -1.696 0.089928 .
## BLUEBOOK_MOD   -2.301e-02  7.155e-03  -3.216 0.001300 **
## OLD_CLAIM_MOD   6.907e-02  1.510e-02   4.574 4.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7254.7  on 8121  degrees of freedom
## AIC: 7334.7
##
## Number of Fisher Scoring iterations: 5
```

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

```
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-R
ED_CAR-CLM_FREQ-CAR_AGE-HSDropout-Professional-Blue_Collar-Clerical-Lawyer-Ho
me_Maker-HOME_VAL_MOD-Student-Doctor, data = train_flag, family = binomial(li
nk='logit'))

summary(flag_reduced_mod)
```

```
##
## Call:
## glm(formula = 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, family = binomial(link = "logit"),
##     data = train_flag)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4861  -0.7200  -0.3998   0.6352   3.1436
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.293e+00  3.328e-01  -6.890 5.57e-12 ***
## KIDSDRIV       4.223e-01  5.501e-02   7.677 1.63e-14 ***
## PARENT1        4.533e-01  9.423e-02   4.810 1.51e-06 ***
## MSTATUS       -6.218e-01  6.876e-02  -9.042  < 2e-16 ***
## TRAVTIME       1.481e-02  1.882e-03   7.868 3.59e-15 ***
## CAR_USE        7.868e-01  7.175e-02  10.965  < 2e-16 ***
## BLUEBOOK       3.501e-05  1.892e-05   1.851  0.06420 .
## TIF           -5.372e-02  7.335e-03  -7.324 2.40e-13 ***
## OLDCLAIM      -2.723e-05  4.673e-06  -5.826 5.68e-09 ***
## REVOKED        9.706e-01  9.257e-02  10.485  < 2e-16 ***
## MVR_PTS        9.784e-02  1.402e-02   6.978 3.00e-12 ***
## URBANICITY     2.365e+00  1.141e-01  20.736  < 2e-16 ***
## Bachelors     -4.712e-01  7.483e-02  -6.297 3.03e-10 ***
## Masters       -5.234e-01  8.887e-02  -5.890 3.87e-09 ***
## PhD           -6.617e-01  1.260e-01  -5.251 1.51e-07 ***
## Panel_Truck    4.290e-01  1.460e-01   2.939  0.00329 **
## Pickup         5.274e-01  9.814e-02   5.374 7.69e-08 ***
## Sports_Car     9.240e-01  1.071e-01   8.628  < 2e-16 ***
## Van            6.242e-01  1.194e-01   5.229 1.71e-07 ***
## SUV            6.976e-01  8.554e-02   8.155 3.50e-16 ***
## Manager       -7.115e-01  1.065e-01  -6.682 2.36e-11 ***
## INCOME_MOD    -5.309e-03  7.736e-04  -6.863 6.75e-12 ***
## BLUEBOOK_MOD  -2.195e-02  7.004e-03  -3.134  0.00172 **
## OLD_CLAIM_MOD  8.201e-02  9.799e-03   8.369  < 2e-16 ***
```

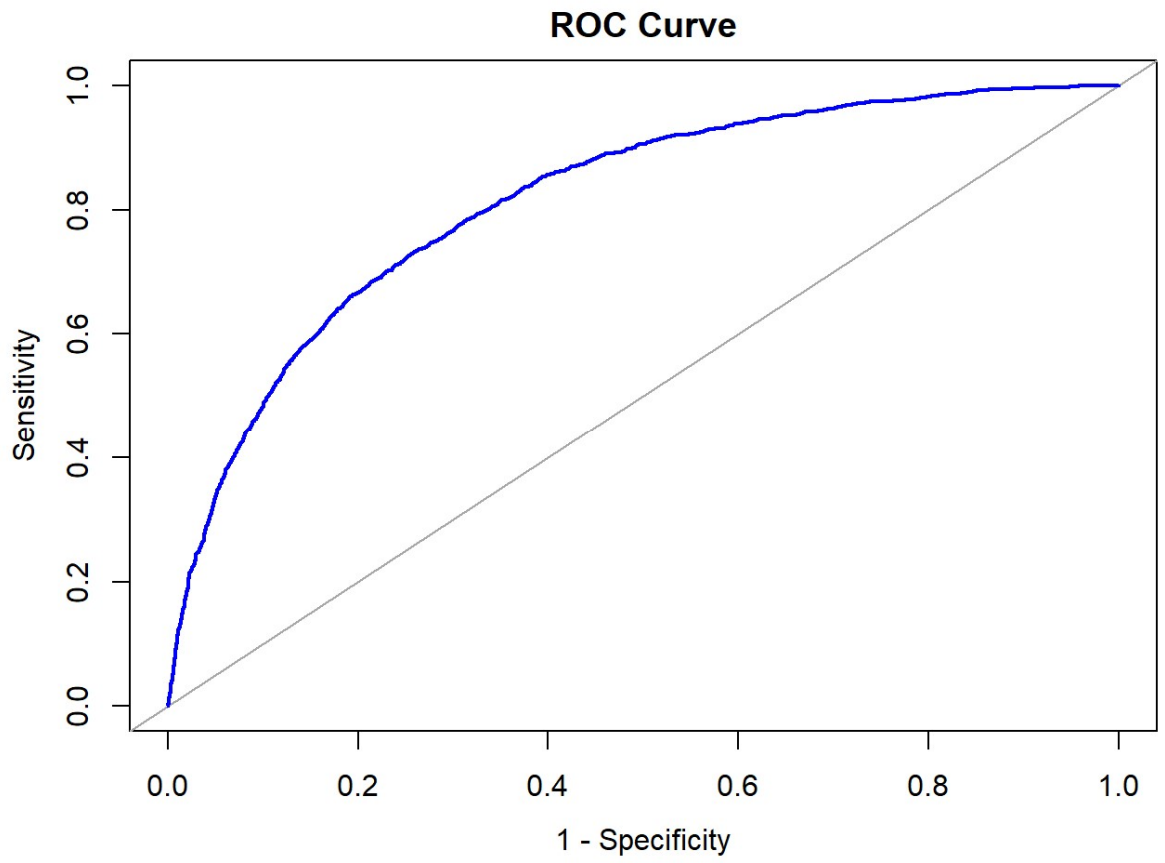
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7290.2  on 8137  degrees of freedom
## AIC: 7338.2
##
## Number of Fisher Scoring iterations: 5
```

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.

Validating the model:

ROC Curve for the Model5:



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.

Now lets do the confusion matrix:

```
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)
```

	0	1
0	5560	1254
1	448	899

```
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)
```

```
## [1] 0.7914471
```

```
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)
```

```
## [1] 0.2085529
```

```
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)
```

```
## [1] 0.4175569
```

```
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)
```

```
## [1] 0.6674091
```

```
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)
```

```
## [1] 0.8159671
```

```
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)
```

```
## [1] 0.5137143
```

Testing the evaluation data with mode 3

In the final step we will test our model by using the test data. 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 below given github locations:

For TARGET_FLAG Predictions: https://github.com/Riteshlohiya/Data621-Assignment-4/blob/master/Evaluation_Data.csv

For TARGET_GLAG = 1 and TARGET_AMOUNT: https://github.com/Riteshlohiya/Data621-Assignment-4/blob/master/Evaluation_Full_Data.csv

```
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
```

```r
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)
```

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 “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:

#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+OLD CLAIM, 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)
```
```
```