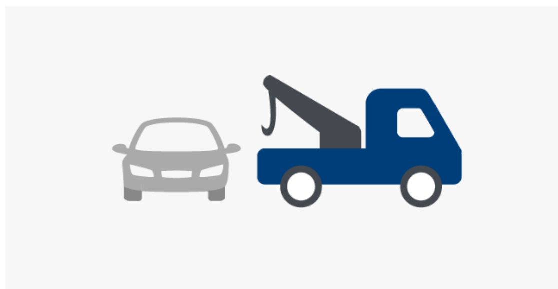
Unit 02 Assignment

Auto Insurance Logistic Regression Project



Name: Young, Brent

Predict 411 Section #: 56

Quarter: Fall 2017

Bingo Bonus:

- Used at least 1 PROBIT MODEL when building my logistic models (5 Points)
- Used MICE package for missing value imputation (20 Points)

Introduction

Context

The dataset that we will be working with is called logit_insurance (includes approximately 8000 records). Each record represents a customer at an auto insurance company. Additionally, each record has two target variables: TARGET_FLAG and TARGET_AMT. For TARGET_FLAG, a "1" means that the person was in a car crash, while a "0" means that the person was not in a car crash. For TARGET_AMT, a value of 0 means that the person did not crash their car, while a number greater than 0 means that they did crash their car.

Objectives/Purpose

The purpose of unit 2 assignment is to analyze insurance data using logistic regression to come up with a probability that a person will crash their car. Additionally, we will also come up with a simple model to predict what it will cost a customer (e.g., insurance damage) if a person does crash their car. This will be accomplished by generating logistic (or probit) regression models using different techniques (e.g., stepwise, etc.) and variables (or the same variables with different transformations). From these techniques and variables, the best model will be selected. First, an initial exploratory data analysis will be conducted using scatterplots, boxplots, summary statistics, etc. to help understand important characteristics and properties of the data that may be disguised by numerical summaries. Second, data preparation/transformations of the data will begin. This includes, but not limited to fixing missing values, conducting data transformations, and creating new variables. Third, we will begin building at least three different logistic (or probit) regression models using different variables. This will be conducted by manually selecting the variables or using variable selection techniques. We will then discuss the coefficients in the model to ensure that it makes intuitive insurance sense. Fourth, we will then decide on the "best model" using metrics such as Log Likelihood, AIC, BIC, and ROC Curve (AUC). Fourth, a Stand Alone scoring program will be conducted that will 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. The data step will include all the variable transformations such as fixing missing values and the regression formulas. Lastly, a scored data file will be produced that will contain three variables for each record: INDEX, P_TARGET_FLAG, and P_TARGET_AMT.

Section 1: Data Exploration

Figure 1: Structure and Size of the Data

```
> str(data)
'data.frame':	8161 obs. of  27 variables:
 $ INDEX       : Factor w/ 8161 levels "1","2","4","5",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ TARGET_FLAG : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 2 2 1 ...
 $ TARGET_AMT  : num  0 0 0 0 0 ...
 $ KIDSDRIV    : int  0 0 0 0 0 0 0 1 0 0 ...
 $ AGE         : int  60 43 35 51 50 34 54 37 34 50 ...
 $ HOMEKIDS    : int  0 0 1 0 0 1 0 2 0 0 ...
 $ YOJ         : int  11 11 10 14 NA 12 NA NA 10 7 ...
 $ INCOME      : num  67349 91449 16039 NA 114986 ...
 $ PARENT1     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
 $ HOME_VAL    : num  0 257252 124191 306251 243925 ...
 $ MSTATUS     : Factor w/ 2 levels "Yes","z_No": 2 2 1 1 1 2 1 1 2 2 ...
 $ SEX         : Factor w/ 2 levels "M","z_F": 1 1 2 1 2 2 2 1 2 1 ...
 $ EDUCATION   : Factor w/ 5 levels "<High School",..: 4 5 5 1 4 2 1 2 2 2 ...
 $ JOB         : Factor w/ 9 levels "","Clerical",..: 7 9 2 9 3 9 9 9 2 7 ...
 $ TRAVTIME    : int  14 22 5 32 36 46 33 44 34 48 ...
 $ CAR_USE     : Factor w/ 2 levels "Commercial","Private": 2 1 2 2 2 1 2 1 2 1 ...
 $ BLUEBOOK    : num  14230 14940 4010 15440 18000 ...
 $ TIF         : int  11 1 4 7 1 1 1 1 1 7 ...
 $ CAR_TYPE    : Factor w/ 6 levels "Minivan","Panel Truck",..: 1 1 6 1 6 4 6 5 6 5 ...
 $ RED_CAR     : Factor w/ 2 levels "0","1": 2 2 1 2 1 1 1 2 1 1 ...
 $ OLDCLAIM    : num  4461 0 38690 0 19217 ...
 $ CLM_FREQ    : int  2 0 2 0 2 0 0 1 0 0 ...
 $ REVOKED     : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 1 1 ...
 $ MVR_PTS     : int  3 0 3 0 3 0 0 10 0 1 ...
 $ CAR_AGE     : int  18 1 10 6 17 7 1 7 1 17 ...
 $ URBANICITY  : Factor w/ 2 levels "Rural","Urban": 2 2 2 2 2 2 2 2 2 1 ...
 $ DO_KIDS_DRIVE: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
```

Observations: Figure 1 shows the structure of the data, which comes out to 8161 rows and 26 variables (integers). INDEX is not considered a true variable, while TARGET_FLAG and TARGET_AMT are considered our two target variables, and the rest of the variables are considered our predictors (mixture of categorical and numerical variables).

Figure 2: Definitions of the Variables (Data Dictionary)

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None. Probability that a person will crash their car. It is a number between 0 and 1
TARGET_AMT	If car was in a crash, what was the cost	None. Insurance damage assuming that a person does crash their car. This number should be greater than 0
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	#Claims(Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	#Children @Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	#Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims(Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

Observations: Figure 2 shows the definitions of the variables that are included in the dataset and the theoretical effect in the third column.

Figure 3: Data Quality Check *(see appendix)*

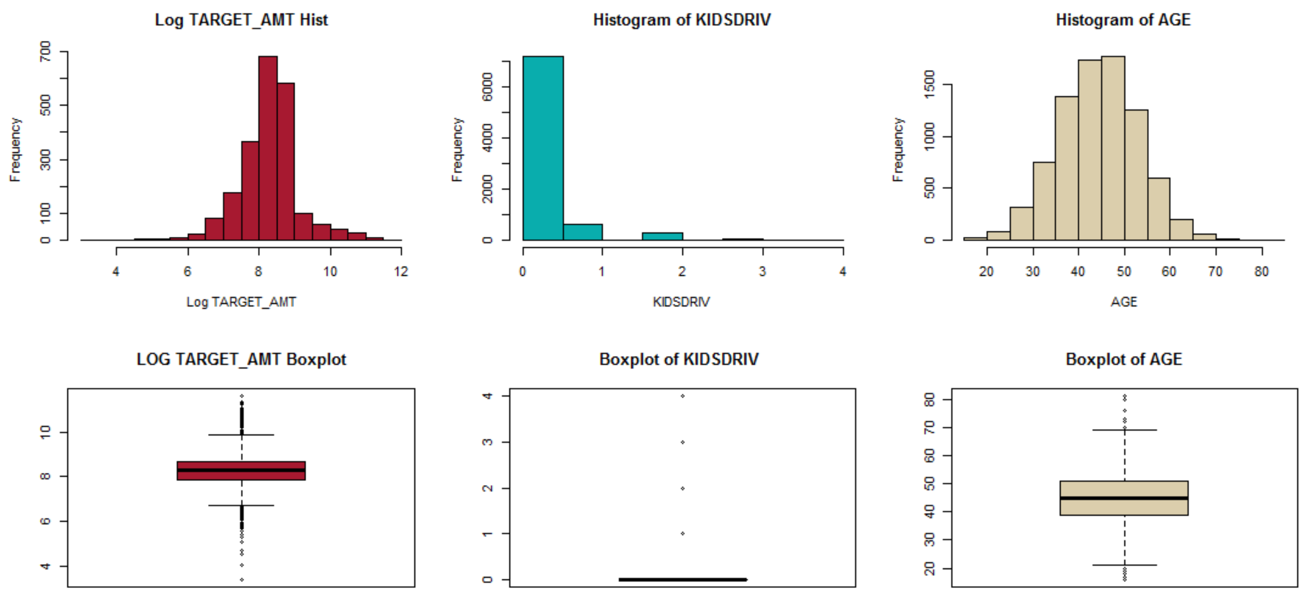
```
> summary(data)
     INDEX       TARGET_FLAG   TARGET_AMT        KIDSDRIV          AGE           HOMEKIDS          YOJ          INCOME
 1      :   1   0:6008      Min.   :     0   Min.   :0.0000   Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Min.   :     0
 2      :   1   1:2153      1st Qu.:     0   1st Qu.:0.0000   1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097
 4      :   1               Median :     0   Median :0.0000   Median :45.00   Median :0.0000   Median :11.0   Median : 54028
 5      :   1               Mean   :  1504   Mean   :0.1711   Mean   :44.79   Mean   :0.7212   Mean   :10.5   Mean   : 61898
 6      :   1               3rd Qu.:  1036   3rd Qu.:0.0000   3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986
 7      :   1               Max.   :107586   Max.   :4.0000   Max.   :81.00   Max.   :5.0000   Max.   :23.0   Max.   :367030
 (Other):8155                                                 NA's   :6                        NA's   :454    NA's   :445
 PARENT1       HOME_VAL       MSTATUS      SEX                EDUCATION              JOB           TRAVTIME           CAR_USE
 No :7084   Min.   :     0   Yes :4894   M  :3786   <High School :1203   z_Blue Collar:1825   Min.   :  5.00   Commercial:3029
 Yes:1077   1st Qu.:     0   z_No:3267   z_F:4375   Bachelors    :2242   Clerical     :1271   1st Qu.: 22.00   Private   :5132
            Median :161160                          Masters      :1658   Professional :1117   Median : 33.00
            Mean   :154867                          PhD          : 728   Manager      : 988   Mean   : 33.49
            3rd Qu.:238724                          z_High School:2330   Lawyer       : 835   3rd Qu.: 44.00
            Max.   :885282                                               Student      : 712   Max.   :142.00
            NA's   :464                                                  (Other)      :1413
    BLUEBOOK          TIF                CAR_TYPE     RED_CAR     OLDCLAIM        CLM_FREQ        REVOKED       MVR_PTS
 Min.   : 1500   Min.   : 1.000   Minivan    :2145   0:5783   Min.   :    0   Min.   :0.0000   No :7161   Min.   : 0.000
 1st Qu.: 9280   1st Qu.: 1.000   Panel Truck: 676   1:2378   1st Qu.:    0   1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000
 Median :14440   Median : 4.000   Pickup     :1389            Median :    0   Median :0.0000              Median : 1.000
 Mean   :15710   Mean   : 5.351   Sports Car : 907            Mean   : 4037   Mean   :0.7986              Mean   : 1.696
 3rd Qu.:20850   3rd Qu.: 7.000   Van        : 750            3rd Qu.: 4636   3rd Qu.:2.0000              3rd Qu.: 3.000
 Max.   :69740   Max.   :25.000   z_SUV      :2294            Max.   :57037   Max.   :5.0000              Max.   :13.000

    CAR_AGE        URBANICITY    DO_KIDS_DRIVE
 Min.   :-3.000   Rural:1669    0:7180
 1st Qu.: 1.000   Urban:6492    1: 981
 Median : 8.000
 Mean   : 8.328
 3rd Qu.:12.000
 Max.   :28.000
 NA's   :510
```

Observations: Figure 3 *(also see appendix for additional data quality checks)* shows summary statistics so that we can check for missing values, outliers, etc. The data shows that the mean target amount is 1504, while median target amount is 0 (since a value of 0 means that the person did not crash their car, while a number greater than 0 means that they did crash their car). Additionally, 74% of customers were not involved in a car crash, while 26% of customers were involved in a car crash. The data quality check also revealed that there are missing values for 5 variables: AGE, YOJ, INCOME, HOME_VAL, and CAR_AGE. Interestingly, variables such as KIDSDRIV, HOMEKIDS, YOJ, INCOME, HOME_VAL, OLDCLAIM, CLM_FREQ, and MVP_PTS have zeroes. We will have to keep this in mind as we build our models. Furthermore, the data quality check also revealed that CAR_AGE has some records with negatives. This will need to be addressed. As we go on, we will have to investigate these outliers, missing values, and decide what to do with them (e.g., conducting imputation, etc.).

Numeric Variables

Target Amount, KIDSDRIV, and Age

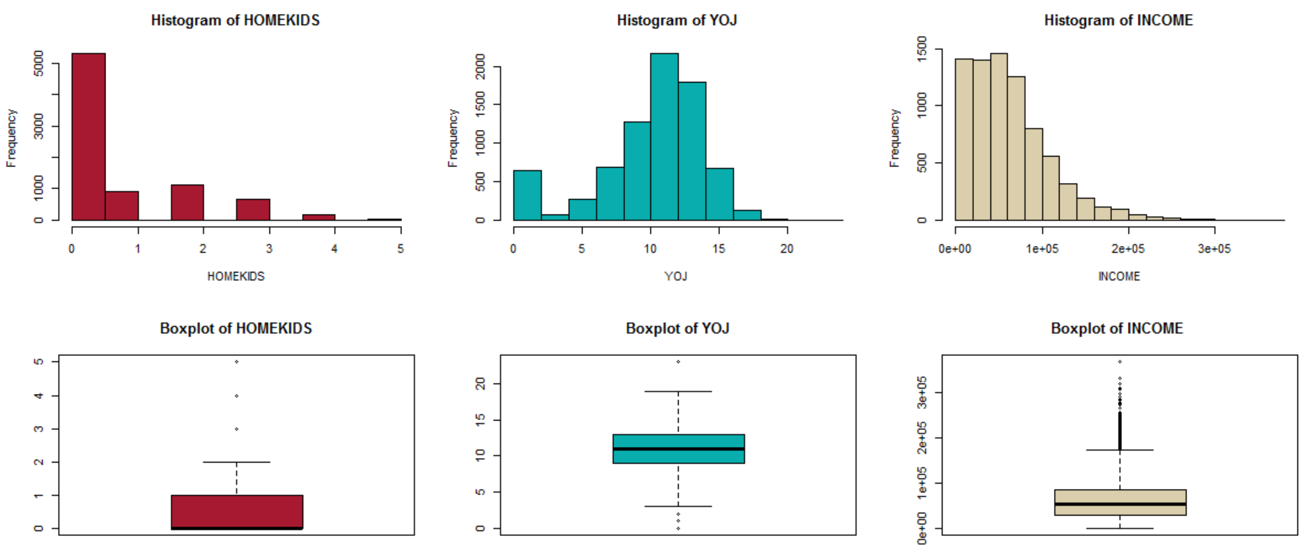
Figure 4: Histogram and Boxplot of Log TARGET AMOUNT, KIDSDRIV, and AGE



Observations: Figure 4 shows a histogram and boxplot of Log TARGET AMOUNT, KIDSDRIV, and AGE. The histogram and boxplot for Log TARGET AMOUNT shows a symmetric bell shape with noticeable outliers. The histogram and boxplot for KIDSDRIV shows a right skew with majority of the values being 0 with some outliers greater than 1. The histogram of AGE shows a symmetric bell shape with most of the values hovering around the mean of 44.79. The boxplot of AGE shows some outliers around less than age 20 and greater than age 70 and a median age of 45. Majority of the ages fall in-between 39 to 51.

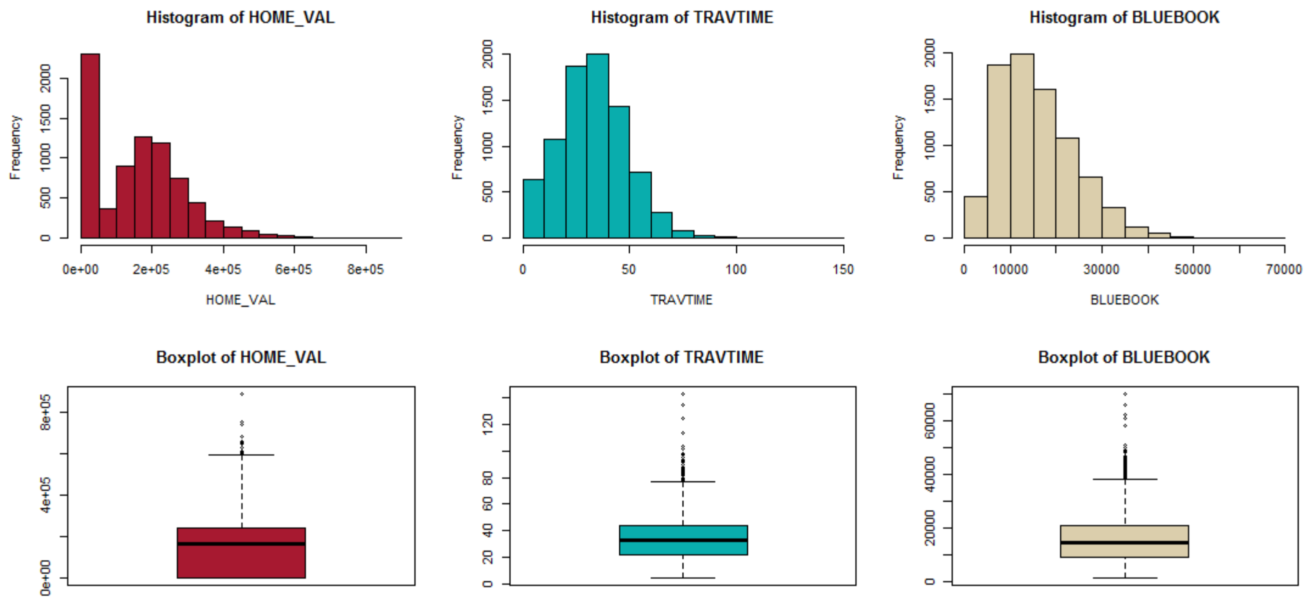
HOMEKIDS, YOJ, and INCOME

Figure 5: Histogram and Boxplot of HOMEKIDS, YOJ, and INCOME



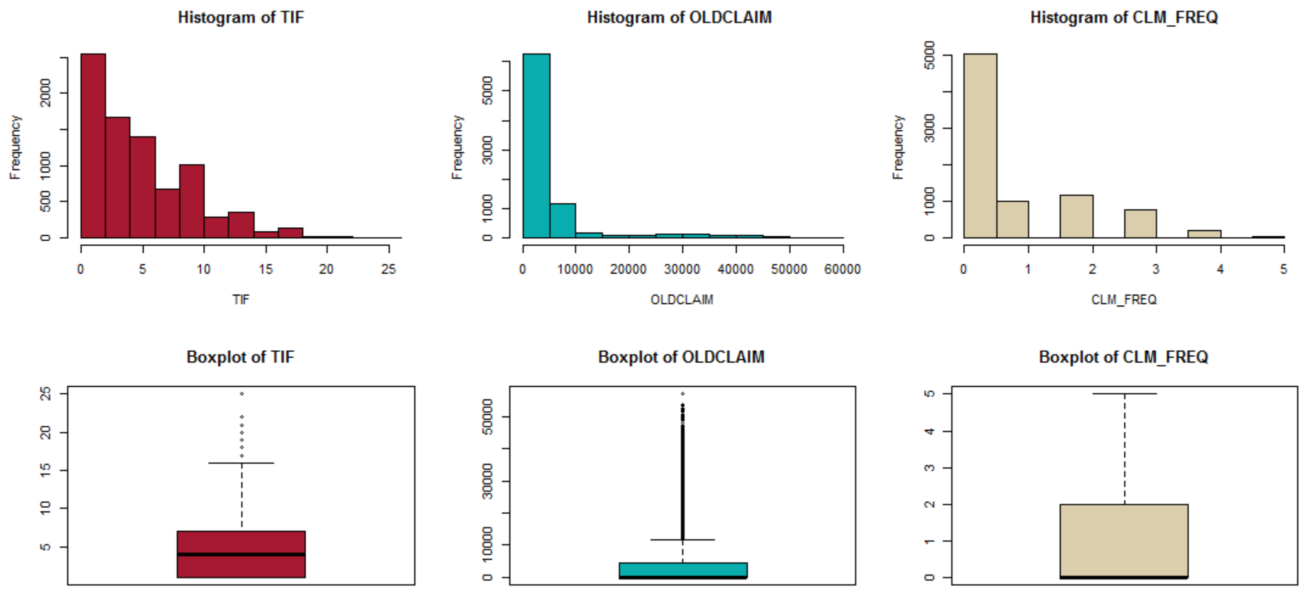
Observations: Figure 5 shows a histogram and boxplot of HOMEKIDS, YOJ, and INCOME. The histogram and boxplot for HOMEKIDS shows a right skew with majority of the values being 0 with some outliers greater than 3. The histogram of YOJ shows a slight left skew, with some 0 values, and majority of the values hovering around the mean of 10.5. The boxplot of YOJ shows some outliers around less than 3 YOJ and greater than 20 YOJ and a median YOJ of 11. Majority of YOJ fall in-between 9 to 13. The histogram of INCOME shows a right skew, with some 0 values, and majority of the values hovering around the mean of 61898. The boxplot of INCOME shows a lot of outliers greater than 175,000 and a median INCOME of 54028. Majority of INCOME fall in-between 28097 to 85986.

Figure 6: Histogram and Boxplot of HOME_VAL, TRAVELTIME, and BLUEBOOK



Observations: Figure 6 shows a histogram and boxplot of HOME_VAL, TRAVELTIME, and BLUEBOOK. The histogram of HOME_VAL shows a right skew, with a lot of 0 values, and a mean of 154867. The boxplot of HOME_VAL shows some outliers around greater than 600,000 and a median HOME_VAL of 161160. It's also important to note that the HOME_VAL variable is missing the second most data out of all the other predictor variables. The histogram of TRAVTIME shows a right skew and majority of the values hovering around the mean of 33.49. The boxplot of TRAVTIME shows outliers around greater than 80 and a median TRAVTIME of 33. Majority of TRAVTIME fall in-between 22 to 44. The histogram of BLUEBOOK shows a right skew and majority of the values hovering around the mean of 15710. The boxplot of BLUEBOOK shows outliers around greater than 40000 and a median BLUEBOOK of 14440. Majority of BLUEBOOK fall in-between 9280 to 20850.

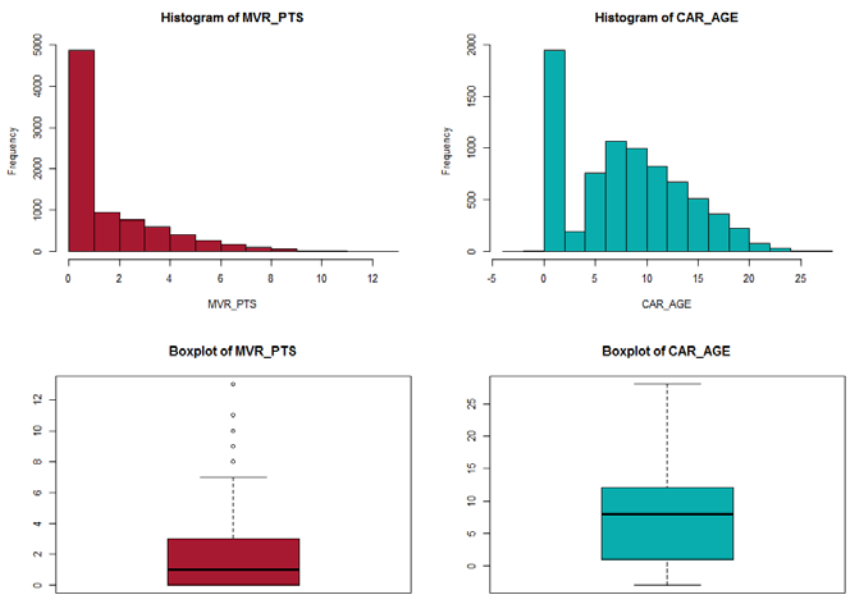
Figure 7: Histogram and Boxplot of TIF, OLDCLAIM, and CLM_FREQ



Observations: Figure 7 shows a histogram and boxplot of TIF, OLDCLAIM, and CLM_FREQ. The histogram of TIF shows a right skew, with a lot of 1 values, and a mean of 5. The boxplot of TIF shows some outliers around greater than 17 and a median TIF of 4. Majority of TIF fall in-between 1 to 7. The histogram of OLDCLAIM shows a right skew, a lot of 0 values, and a mean of 4037. The boxplot of OLDCLAIM shows extreme outliers around greater than 10000 and a median OLDCLAIM of 0. The histogram of CLM_FREQ shows a right skew and a lot of 0 values as well with a mean of 0.7986. The boxplot of CLM_FREQ shows a median CLM_FREQ of 0.

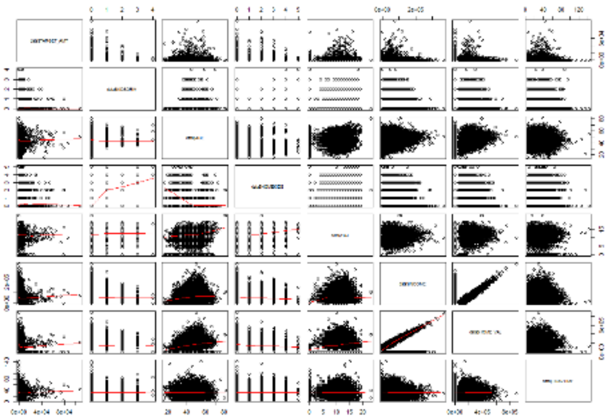
MVP_PTS and CAR_AGE

Figure 8: Histogram and Boxplot of MVP_PTS and CAR_AGE

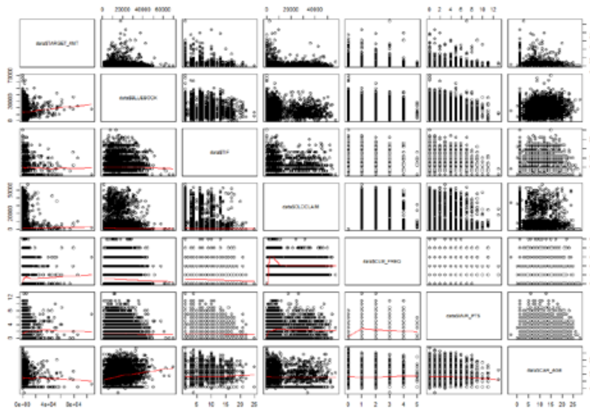


Observations: Figure 8 shows a histogram and boxplot of MVP_PTS and CAR_AGE. The histogram of MVP_PTS shows a right skew, a lot of 0 values, some outliers around greater than 8, and a mean of 1.696. The box plot of MVP_PTS also shows that the median number of MVP_PTS is 1 and shows the outliers that are seen in the histogram. The histogram of CAR_AGE shows a right skew with majority of the values hovering around 1. There is also some values that have negative age, which will need to be addressed prior to building the model. The box plot of CAR_AGE also shows that the median number of CAR_AGE is 8. Majority of CAR_AGE fall in-between 1 to 12. It’s also important to note that the CAR_AGE variable is missing the most data out of all the other predictor variables.

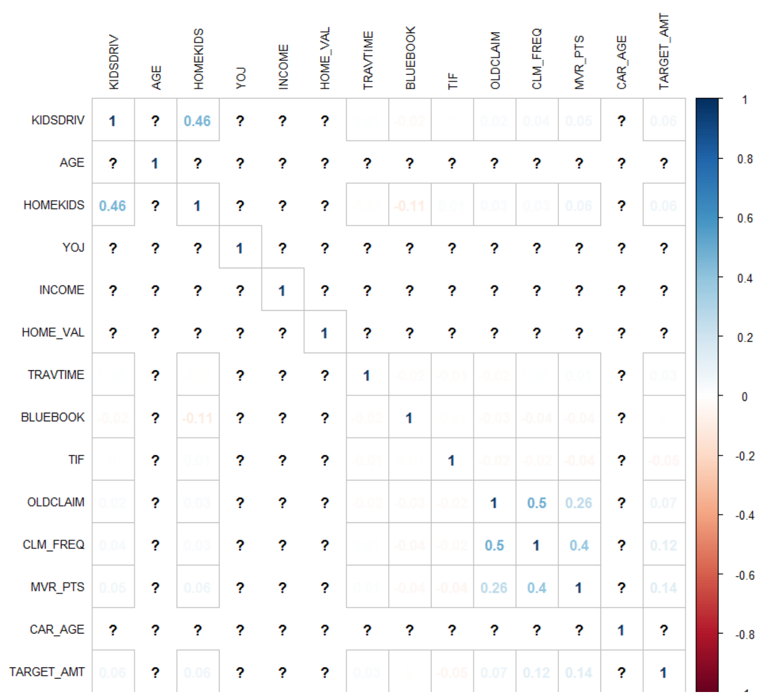
Figure 9: Scatterplot Matrices and Correlation Matrix



Scatterplot Matrix of TARGET_AMT, KIDSDRIVE, AGE, HOMEKIDS, YOJ, INCOME, HOME_VAL, and TRAVELTIME



Scatterplot Matrix of BLUEBOOK, TIF, OLDCLAIM, CLM_FREQ, MVP_PTS, and CAR_AGE

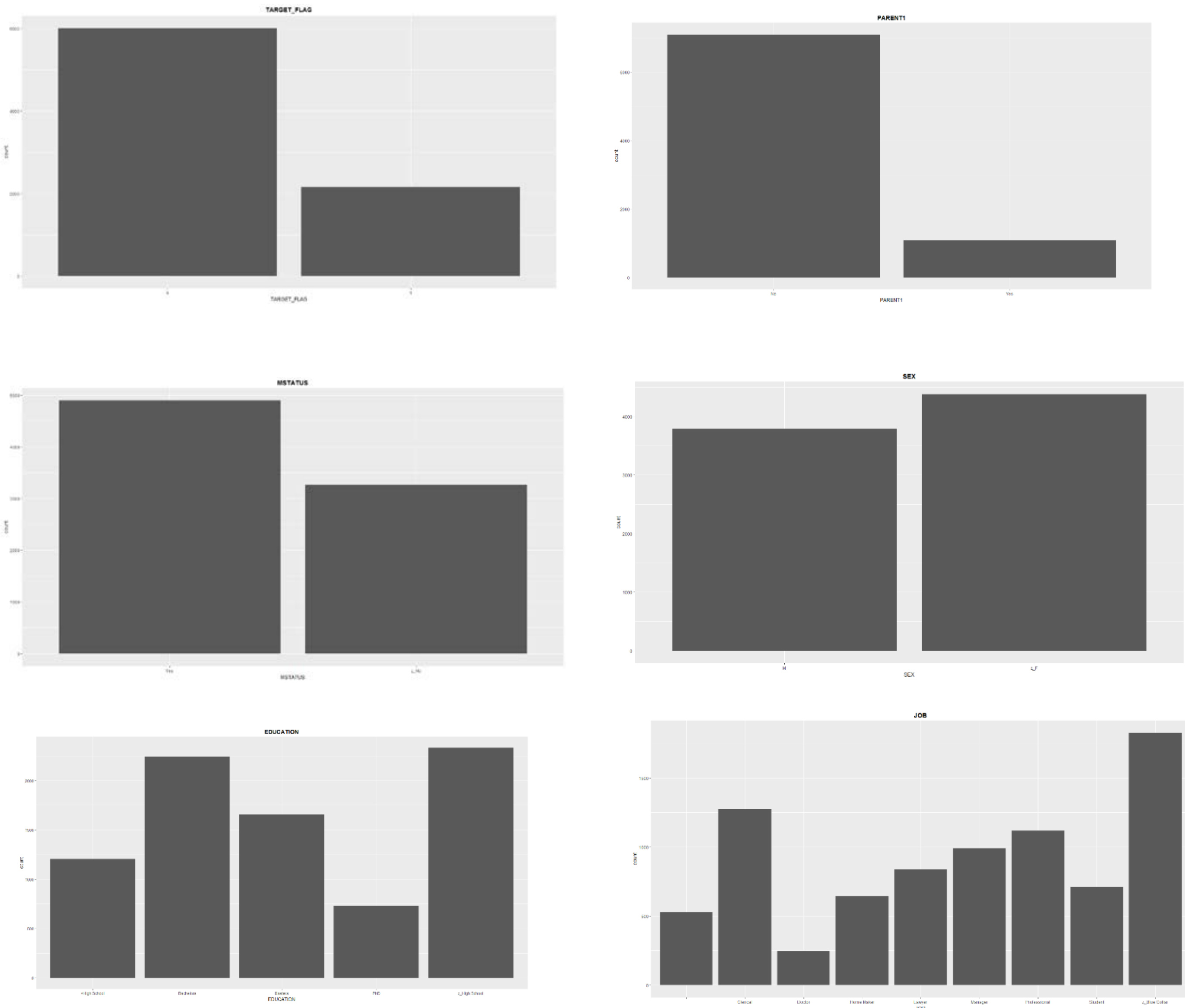


Correlation Matrix of all Numerical Variables

Observations: Figure 9 shows scatterplot matrices and a correlation matrix of all the numeric variables that were included in the dataset (excluding INDEX). This gives us an idea of the most promising predictor variables based on the predictors that are most correlated with TARGET_AMT. This also allows us to see which variables may be correlated with each other so that we can gleam interesting insights. Also, note that the correlation matrix is incomplete due to the missing values for the following 5 variables: AGE, YOJ, INCOME, HOME_VAL, and CAR_AGE. The scatterplot matrix for these variables also show N/A or no correlations. The scatterplot matrix and correlation matrix shows that MVP_PTS had the strongest positive correlation with TARGET_AMT, which makes intuitive sense since if you get lots of traffic tickets, you tend to get into more crashes and therefore have to pay a lot in insurance damage. Furthermore, the scatterplot matrix and correlation matrix also revealed strong positive correlations between KIDSDRIVE vs. HOMEKIDS (more kids at home, means more kids that drive), OLDCLAIM vs. CLM_FREQ (more claims you filed in the past, means the more you paid in the past), and CLM_FREQ vs. MVR_PTS (the more traffic tickets you receive, the more likely someone gets into crashes and therefore submits more claims). After the missing values are addressed, another correlation matrix will be created to see if we can uncover additional insights prior to model building.

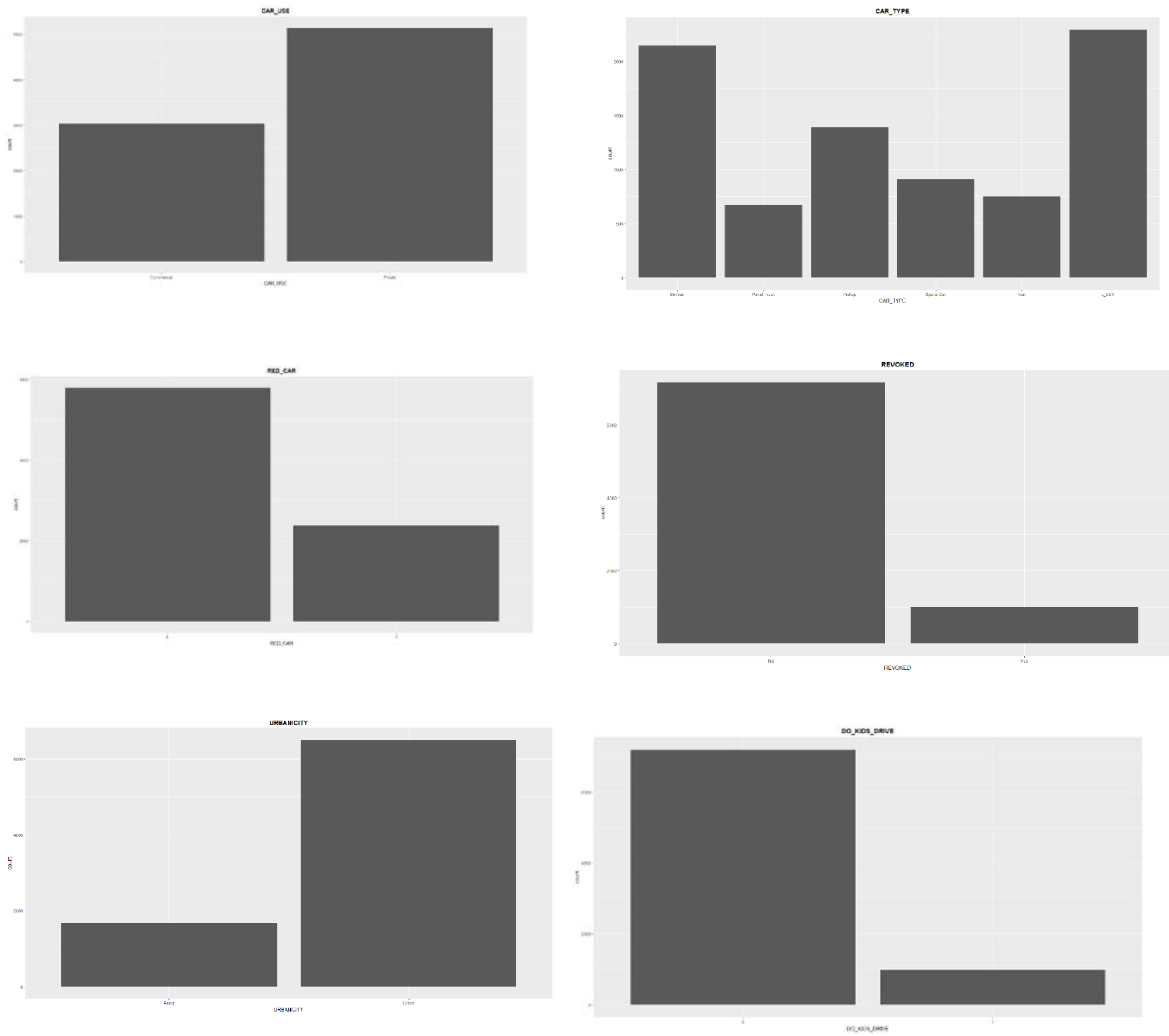
Categorical Variables

Figure 10: Bar plots of TARGET_FLAG, PARENT1, MSTATUS, SEX, EDUCATION, and JOB



Observations: Figure 10 shows bar plots of TARGET_FLAG, PARENT1, MSTATUS, SEX, EDUCATION, and JOB. The bar plots revealed that 26% of the customers in the data set were involved in a car crash, 87% are not single parents, 60% are married, 54% are female, 44% have high school level education and below, and 22% are blue collar workers.

Figure 11: Bar plots of CAR_USE, CAR_TYPE, RED_CAR, REVOKED, URBANCITY, and DO_KIDS_DRIVE



Observations: Figure 11 shows bar plots of CAR_USE, CAR_TYPE, RED_CAR, REVOKED, URBANCITY, and DO_KIDS_DRIVE. The bar plots revealed that 63% of the customers in the data set use their car for private use, majority of the customers drive minivans (26%) or SUV's (28%), 71% of customers do not drive a red car, 88% of customers did not have their license revoked in the past 7 years, 80% of customers live in an urban city, and 88% of customers do not have teenagers that drive.

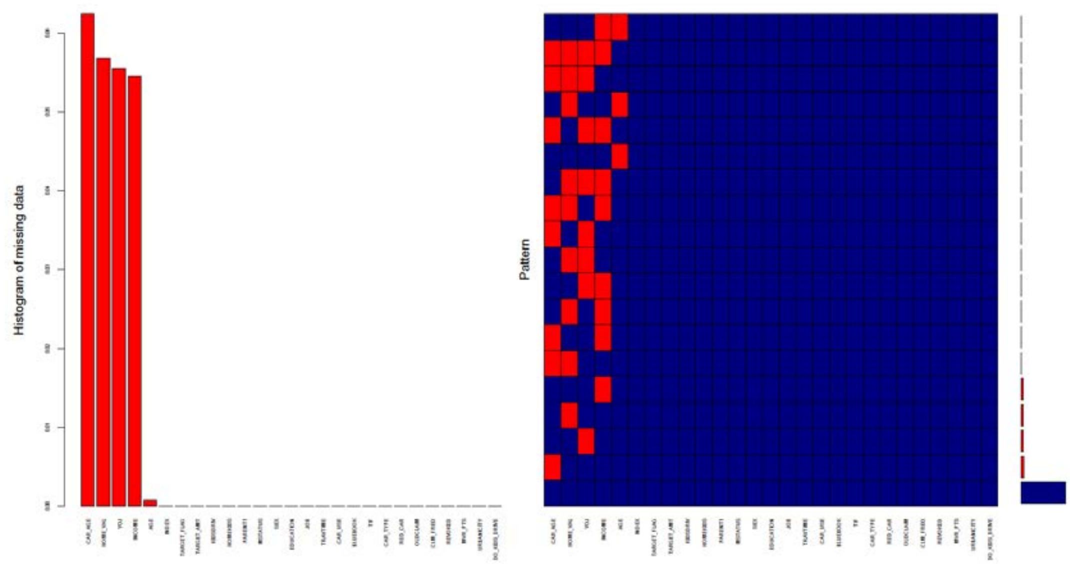
Section 2: Data Preparation

Figure 12: Missing Values for Variables

```
     INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS         YOJ
         0           0           0           0           6           0         454
    INCOME     PARENT1    HOME_VAL     MSTATUS         SEX   EDUCATION         JOB
       445           0         464           0           0           0           0
  TRAVTIME     CAR_USE    BLUEBOOK         TIF    CAR_TYPE     RED_CAR    OLDCLAIM
         0           0           0           0           0           0           0
  CLM_FREQ     REVOKED     MVR_PTS     CAR_AGE  URBANICITY DO_KIDS_DRIVE
         0           0           0         510           0           0
```

Observations: Figure 12 shows variables in the data set that have missing data. We will use the MICE package (pmm = predictive mean matching) to impute the missing data. The MICE package uses an algorithm in such a way that uses information from other variables in dataset to predict and impute the missing values. We need to address the missing values because logistic regression cannot handle missing values and must be addressed prior to utilizing this modeling technique.

Figure 13: Percentage of Missing Values for Variables




```
     INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS         YOJ
 0.0000000   0.0000000   0.0000000   0.0000000   0.0735204   0.0000000   5.5630437
    INCOME     PARENT1    HOME_VAL     MSTATUS         SEX   EDUCATION         JOB
 5.4527631   0.0000000   5.6855777   0.0000000   0.0000000   0.0000000   0.0000000
  TRAVTIME     CAR_USE    BLUEBOOK         TIF    CAR_TYPE     RED_CAR    OLDCLAIM
 0.0000000   0.0000000   0.0000000   0.0000000   0.0000000   0.0000000   0.0000000
  CLM_FREQ     REVOKED     MVR_PTS     CAR_AGE  URBANICITY DO_KIDS_DRIVE
 0.0000000   0.0000000   0.0000000   6.2492342   0.0000000   0.0000000
```

Observations: Figure 13 shows the percentage of missing variables in the data set. CAR_AGE had the most data missing, while AGE had the least.

Figure 14: Summary of Imputation using Predictive Mean Matching

```
Multiply imputed data set
Call:
mice(data = subdatnum.df, m = 5, method = "pmm", maxit = 50,
    seed = 500)
Number of multiple imputations:  5
Missing cells per column:
     INDEX  KIDSDRIV       AGE  HOMEKIDS       YOJ    INCOME  HOME_VAL  TRAVTIME
         0         0         6         0       454       445       464         0
  BLUEBOOK       TIF  OLDCLAIM  CLM_FREQ   MVR_PTS   CAR_AGE TARGET_AMT
         0         0         0         0         0       510         0
Imputation methods:
     INDEX  KIDSDRIV       AGE  HOMEKIDS       YOJ    INCOME  HOME_VAL  TRAVTIME
     "pmm"     "pmm"     "pmm"     "pmm"     "pmm"     "pmm"     "pmm"     "pmm"
  BLUEBOOK       TIF  OLDCLAIM  CLM_FREQ   MVR_PTS   CAR_AGE TARGET_AMT
     "pmm"     "pmm"     "pmm"     "pmm"     "pmm"     "pmm"     "pmm"
```

```
  INDEX  TARGET_FLAG   TARGET_AMT    KIDSDRIV         AGE     HOMEKIDS         YOJ
      0            0            0           0           0            0           0
     INCOME      PARENT1     HOME_VAL     MSTATUS         SEX    EDUCATION         JOB
          0            0            0           0           0            0           0
   TRAVTIME      CAR_USE     BLUEBOOK         TIF    CAR_TYPE      RED_CAR    OLDCLAIM
          0            0            0           0           0            0           0
   CLM_FREQ      REVOKED      MVR_PTS     CAR_AGE  URBANICITY DO_KIDS_DRIVE
          0            0            0           0           0            0
```

Observations: Figure 14 shows imputation being applied to the missing values using predictive mean matching. The result shows that all the missing values have been replaced.

Figure 15: Transformation of Variables

Discussion: I created 5 flag variables called: HAVE_HOME_KIDS (1=Yes, 0 =No, using HOMEKIDS), EMPLOYED (1=Yes, 0 =No, using YOJ), HOME_OWNER (1=Yes, 0 =No, using HOME_VAL), SUBMITTED_CLAIM (1=Yes, 0 =No), and HAVE_MVR_PTS (1=Yes, 0 =No). I created these flag variables because these variables contained “0” values in the data set.

Additionally, TRAVTIME, BLUEBOOK, MVR_PTS, TIF, and OLDCLAIM were five variables that I initially transformed using SQRT and LOG since they had the most outliers. I will experiment with these transformations later on in the model building section.

Furthermore, I also employed binning on the following variables since these variables are based on dollar amounts and there was high variability. Binning these variables allows us to analyze the data in a more simplistic way and helps to merge small, medium, and high values into a single group, etc.

- INCOME:
    - Missing vales <- "NA"
    - 0 <- "Zero"
    - < 30000 <- "Low"
    - >= 30000 & < 80000 <- "Medium"
    - >= 80000 <- "High"

- HOME_VAL:
    - Missing values <- "NA"
    - 0 <- "Zero"
    - >= 1 < 125000 <- "Low"
    - >= 125000 & < 300000 <- "Medium"
    - >= 300000 <- "High"

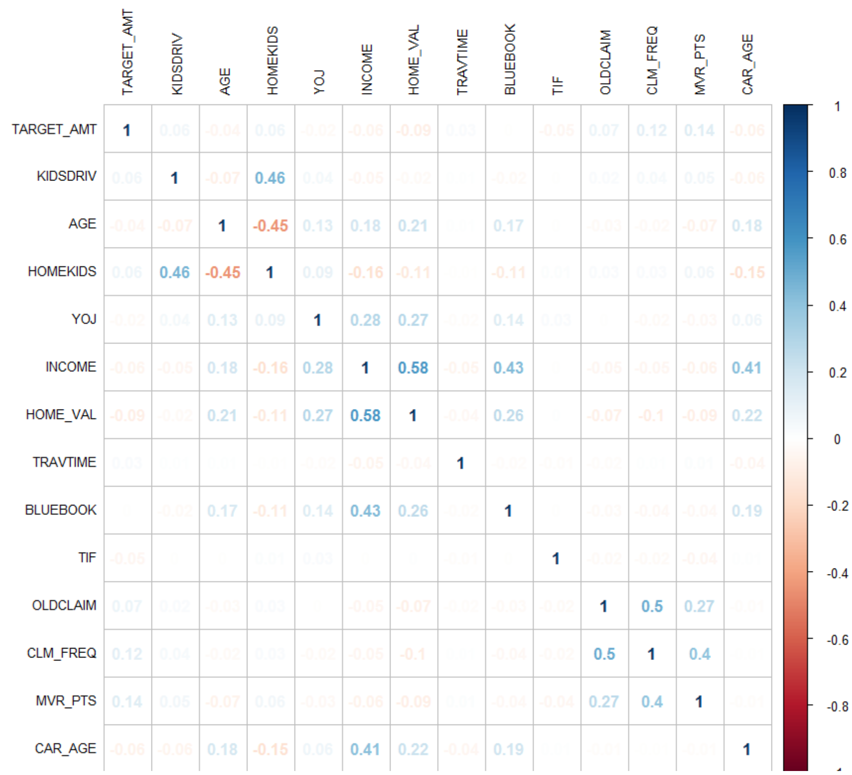
- OLDCLAIM
    - Missing values <- "NA"
    - 0 <- "Zero"
    - >= 1 & < 1000 <- "Low"
    - >= 1000 & < 4500] <- "Medium"
    - >= 4500 <- "High"

Figure 16: Handling Outliers

- CAR_AGE < 0 = 0
- YOJ >= 20 = 20
- INCOME >= 300000 = 300000
- HOME_VAL >= 650000 = 650000
- TRAVTIME >= 100) = 100
- BLUEBOOK >= 55000 = 55000
- TIF >= 17 = 17
- MVR_PTS >=8 = 8

Discussion: Figure 16 shows how I handled the outliers from the insurance data set based on the EDA in section 1 (e.g., box plots, bar graphs, and summary statistics). Addressing outliers is important because outliers can exert significant influence on model parameters. For instance, the model may be less accurate and the model may give a different interpretation or understanding that actually exists. Additionally, outliers can significantly impact a predictive model. For example, an outlier can cause a large difference in the coefficient or “beta” value in a regression model. As a result, the primary technique that I used to handle the outliers was trimming the data (e.g., when a variable exceeds a certain limit, it is simply truncated so that it cannot exceed the limit).

Figure 17: Correlation Matrix After Missing Values, Outliers, etc. Have Been Addressed



Observations: Figure 17 shows a correlation matrix of all numeric variables that were included in the dataset (excluding INDEX) after the missing values, outliers, etc. have been addressed. The correlation matrix allows us to see which variables may be correlated with each other so that we can gleam interesting insights. The correlation matrix revealed strong positive correlations that we saw earlier in our EDA such as KIDSDRIVE vs. HOMEKIDS (more kids at home, means more kids that drive), OLDCLAIM vs. CLM_FREQ (more claims you filed in the past, means the more you paid in the past), and CLM_FREQ vs. MVR_PTS (the more traffic tickets you receive, the more likely someone gets into crashes and therefore submits more claims). The correlation matrix also revealed additional strong positive correlations that we did not see earlier in our EDA such as HOME_VAL vs. INCOME (the more money you make, the more money you will have to buy a more expensive house), BLUEBOOK vs. INCOME (the more money you make, the more money you will have to buy a more expensive car and therefore the BLUEBOOK value would be higher), CAR_AGE vs. INCOME (as income increases, car age also increases) and strong negative correlations such as AGE vs. HOMEKIDS (tend to have kids when you are younger and when kids get older they tend to move out).

Section 3: Build Models

Model 1

Figure 18: Model 1 (Full Model)

```
> summary(Model1)

Call:
glm(formula = TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ +
    INCOME + PARENT1 + HOME_VAL + MSTATUS + SEX + EDUCATION +
    JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + RED_CAR +
    OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY +
    DO_KIDS_DRIVE + HAVE_HOME_KIDS + EMPLOYED + HOME_OWNER +
    SUBMITTED_CLAIM + HAVE_MVR_PTS + INCOME_bin + HOME_VAL_bin +
    OLDCLAIM_bin, family = binomial(link = "logit"), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4939  -0.7073  -0.3894   0.6247   3.1368

Coefficients: (2 not defined because of singularities)
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.974e+00  3.769e-01  -7.891 3.01e-15 ***
KIDSDRIV                2.458e-01  1.267e-01   1.939 0.052478 .
AGE                    -1.673e-03  4.255e-03  -0.393 0.694185
HOMEKIDS               -7.955e-02  5.485e-02  -1.450 0.146955
YOJ                     1.813e-02  1.220e-02   1.486 0.137270
INCOME                 -5.435e-07  1.726e-06  -0.315 0.752789
PARENT1Yes              2.183e-01  1.217e-01   1.794 0.072889 .
HOME_VAL                7.505e-07  8.985e-07   0.835 0.403575
MSTATUSz_No             5.699e-01  8.940e-02   6.375 1.83e-10 ***
SEXz_F                 -9.021e-02  1.127e-01  -0.801 0.423415
EDUCATIONBachelors     -3.572e-01  1.185e-01  -3.016 0.002565 **
EDUCATIONMasters       -2.374e-01  1.803e-01  -1.316 0.188078
EDUCATIONPhD           -2.155e-01  2.162e-01  -0.997 0.318864
EDUCATIONz_High School  1.696e-02  9.774e-02   0.174 0.862214
JOBClerical             4.215e-01  1.983e-01   2.126 0.033530 *
JOBDoctor              -3.680e-01  2.672e-01  -1.377 0.168513
JOBHome Maker           6.346e-02  2.233e-01   0.284 0.776281
JOBLawyer               1.059e-01  1.704e-01   0.621 0.534359
JOBManager             -5.480e-01  1.719e-01  -3.188 0.001431 **
JOBProfessional         1.674e-01  1.796e-01   0.932 0.351134
JOBStudent              8.797e-03  2.289e-01   0.038 0.969341
JOBz_Blue Collar        3.386e-01  1.869e-01   1.812 0.070006 .
TRAVTIME                1.477e-02  1.905e-03   7.750 9.21e-15 ***
CAR_USEPrivate         -7.539e-01  9.238e-02  -8.160 3.34e-16 ***
BLUEBOOK               -2.052e-05  5.306e-06  -3.868 0.000110 ***
TIF                    -5.507e-02  7.439e-03  -7.403 1.33e-13 ***
CAR_TYPEPanel Truck     5.539e-01  1.628e-01   3.402 0.000668 ***
CAR_TYPEPickup          5.668e-01  1.012e-01   5.598 2.16e-08 ***
CAR_TYPESports Car      1.021e+00  1.306e-01   7.819 5.32e-15 ***
CAR_TYPEVan             6.220e-01  1.273e-01   4.886 1.03e-06 ***
CAR_TYPEz_SUV           7.629e-01  1.119e-01   6.818 9.25e-12 ***
RED_CAR1               -1.946e-03  8.681e-02  -0.022 0.982112
OLDCLAIM               -2.279e-05  4.756e-06  -4.792 1.65e-06 ***
CLM_FREQ                4.583e-02  4.458e-02   1.028 0.303979
```

```
REVOKEDYes              9.750e-01  9.348e-02  10.430  < 2e-16 ***
MVR_PTS                 8.999e-02  1.996e-02   4.509 6.50e-06 ***
CAR_AGE                -3.128e-03  7.290e-03  -0.429 0.667837
URBANICITYUrban         2.373e+00  1.138e-01  20.861  < 2e-16 ***
DO_KIDS_DRIVE1          2.505e-01  2.041e-01   1.227 0.219691
HAVE_HOME_KIDS1         3.692e-01  1.449e-01   2.548 0.010831 *
EMPLOYED1              -3.673e-01  2.790e-01  -1.317 0.187968
HOME_OWNER1            -6.596e-01  3.324e-01  -1.984 0.047205 *
SUBMITTED_CLAIM1        5.806e-01  1.347e-01   4.311 1.62e-05 ***
HAVE_MVR_PTS1           2.775e-02  8.699e-02   0.319 0.749763
INCOME_binLow          -4.005e-01  2.503e-01  -1.600 0.109523
INCOME_binMedium       -4.744e-01  2.685e-01  -1.767 0.077180 .
INCOME_binHigh         -8.703e-01  3.103e-01  -2.805 0.005032 **
HOME_VAL_binLow         4.269e-01  2.597e-01   1.644 0.100146
HOME_VAL_binMedium      1.804e-01  1.819e-01   0.992 0.321267
HOME_VAL_binHigh               NA         NA      NA       NA
OLDCLAIM_binLow        -4.413e-01  2.432e-01  -1.814 0.069637 .
OLDCLAIM_binMedium     -2.530e-02  1.006e-01  -0.251 0.801536
OLDCLAIM_binHigh               NA         NA      NA       NA
---
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: 7234.6  on 8110  degrees of freedom
AIC: 7336.6

Number of Fisher Scoring iterations: 5
```

```
Analysis of Deviance Table (Type II tests)

Response: TARGET_FLAG
                Df    Chisq Pr(>Chisq)
KIDSDRIV         1   3.7604  0.0524785 .
AGE              1   0.1546  0.6941847
HOMEKIDS         1   2.1036  0.1469546
YOJ              1   2.2083  0.1372699
INCOME           1   0.0992  0.7527890
PARENT1          1   3.2167  0.0728892 .
HOME_VAL         1   0.6976  0.4035749
MSTATUS          1  40.6431  1.827e-10 ***
SEX              1   0.6408  0.4234149
EDUCATION        4  18.0548  0.0012040 **
JOB              8  60.8439  3.183e-10 ***
TRAVTIME         1  60.0572  9.214e-15 ***
CAR_USE          1  66.5930  3.338e-16 ***
BLUEBOOK         1  14.9606  0.0001098 ***
TIF              1  54.8010  1.334e-13 ***
CAR_TYPE         5  86.3437  < 2.2e-16 ***
RED_CAR          1   0.0005  0.9821117
OLDCLAIM         1  22.9607  1.653e-06 ***
CLM_FREQ         1   1.0567  0.3039787
REVOKED          1 108.7874  < 2.2e-16 ***
MVR_PTS          1  20.3336  6.505e-06 ***
CAR_AGE          1   0.1841  0.6678375
URBANICITY       1 435.1777  < 2.2e-16 ***
DO_KIDS_DRIVE    1   1.5064  0.2196912
HAVE_HOME_KIDS   1   6.4928  0.0108312 *
EMPLOYED         1   1.7335  0.1879685
HOME_OWNER       1   3.9381  0.0472052 *
SUBMITTED_CLAIM  1  18.5859  1.624e-05 ***
HAVE_MVR_PTS     1   0.1017  0.7497630
INCOME_bin       3  12.4037  0.0061208 **
HOME_VAL_bin     2   3.3037  0.1916964
OLDCLAIM_bin     2   3.3113  0.1909685
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                             0.231834
Cox and Snell (ML)                   0.234741
Nagelkerke (Cragg and Uhler)         0.342871

$Likelihood.ratio.test
 Df.diff LogLik.diff   Chisq p.value
     -50     -1091.7 2183.4       0
```

Observations: Figure 18 shows a summary of the logistic regression model TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 + HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY + DO_KIDS_DRIVE + HAVE_HOME_KIDS + EMPLOYED + HOME_OWNER + SUBMITTED_CLAIM + HAVE_MVR_PTS + INCOME_bin + HOME_VAL_bin + OLDCLAIM_bin. Model 1 (full model) includes all the variables in the insurance dataset in addition to the binned variables that were created. This model gives us a “starting point/base model” that we can work off of. The deviance of residuals, which is a measure of model fit of a generalized linear model, shows that the null deviance is 9418.0 and the residual deviance is 7234.6. Since a null deviance shows how well the response variable is predicted by the model that includes only the intercept, the results shows that there was a significant reduction in deviance, even though the deviance of the residuals are high. The results also show an AIC of 7336.6, which provides a method for assessing the quality of the model (e.g., model complexity) that we can use later on when we compare this model to other models (lower AIC the better).

The Analysis of Deviance tables shows the difference between the null deviance and the residual deviance (wider the gap, the better). The table shows that adding variables such as URBANCITY, REVOKED, and CAR_TYPE significantly reduces the residual deviance, whereas variables such as AGE, HOMEKIDS, YOJ seem to improve the model less as indicated by the low deviance and large p-values (without the variable explains more or less the same amount of variation). Ultimately, after employing stepwise regression, my hope is that the variables with low deviance and large-p-values (insignificant) will be dropped. Lastly, the results show Pseudo R-Square Metrics for McFadden: 0.231834, Cox and Snell (ML): 0.234741, and Nagelkerke (Cragg and Uhler): 0.342871, which helps estimate the coefficient of determination (larger the better). We will use these model fit metrics to compare our models later on. It’s also important to note that since this model includes all the variables, it’s a very complex model. As a result, in our next model we will use stepwise regression to help us create a more parsimonious and simple model.

For the most part, most of the coefficients in the model make sense. For example, TRAVTIME and MVP_PTS are positive, while BLUEBOOK is negative. This means for every one unit change in TRAVTIME and MVP_PTS, the log odds of getting into a car crash increases, which makes intuitive insurance sense (e.g., long drives to work and more traffic tickets, suggests greater risk and increases the likelihood of getting into crashes). On the other hand, for BLUEBOOK, for every one unit change in BLUEBOOK, the log odds getting into a car crash decreases, which also makes intuitive insurance sense (e.g., a car that is worth more, most likely means that a car is newer or has better performance and therefore decreases the likelihood of getting into a car crash). A more in-depth analysis of the coefficients will be provided in Model 2 (stepwise) and in Model 3 (stepwise with transformations).

Model 2

Figure 19: Model 2 (Stepwise Model)

```
> summary(Model2)

Call:
glm(formula = TARGET_FLAG ~ URBANICITY + JOB + MVR_PTS + MSTATUS + 
    CAR_TYPE + REVOKED + DO_KIDS_DRIVE + INCOME_bin + CAR_USE + 
    TRAVTIME + TIF + OLDCLAIM_bin + BLUEBOOK + OLDCLAIM + HAVE_HOME_KIDS + 
    EDUCATION + HOME_OWNER + PARENT1 + KIDSDRIV, family = binomial(link = "logit")
,
    data = data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.4618  -0.7098  -0.3926   0.6284   3.1450  

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -3.050e+00  3.104e-01  -9.825  < 2e-16 ***
URBANICITYUrban         2.370e+00  1.137e-01  20.855  < 2e-16 ***
JOBClerical             4.232e-01  1.967e-01   2.151 0.031464 *  
JOBDoctor              -3.531e-01  2.656e-01  -1.330 0.183632    
JOBHome Maker           1.112e-01  2.164e-01   0.514 0.607331    
JOBLawyer               1.123e-01  1.694e-01   0.663 0.507401    
JOBManager             -5.415e-01  1.711e-01  -3.164 0.001554 ** 
JOBProfessional         1.736e-01  1.786e-01   0.972 0.331204    
JOBStudent              1.781e-02  2.249e-01   0.079 0.936907    
JOBz_Blue Collar        3.375e-01  1.858e-01   1.817 0.069283 .  
MVR_PTS                 9.421e-02  1.445e-02   6.519 7.10e-11 ***
MSTATUSz_No             5.612e-01  8.798e-02   6.378 1.79e-10 ***
CAR_TYPEPanel Truck     5.959e-01  1.515e-01   3.933 8.38e-05 ***
CAR_TYPEPickup          5.609e-01  1.010e-01   5.553 2.80e-08 ***
CAR_TYPESports Car      9.512e-01  1.080e-01   8.810  < 2e-16 ***
CAR_TYPEVan             6.394e-01  1.227e-01   5.210 1.89e-07 ***
CAR_TYPEz_SUV           6.929e-01  8.653e-02   8.008 1.17e-15 ***
REVOKEDYes              9.714e-01  9.333e-02  10.408  < 2e-16 ***
DO_KIDS_DRIVE1          2.958e-01  1.990e-01   1.486 0.137198    
INCOME_binLow          -5.601e-01  1.340e-01  -4.181 2.90e-05 ***
INCOME_binMedium       -6.707e-01  1.533e-01  -4.376 1.21e-05 ***
INCOME_binHigh         -1.107e+00  1.690e-01  -6.548 5.84e-11 ***
CAR_USEPrivate         -7.568e-01  9.219e-02  -8.209 2.23e-16 ***
TRAVTIME                1.481e-02  1.903e-03   7.784 7.02e-15 ***
TIF                    -5.511e-02  7.428e-03  -7.420 1.17e-13 ***
OLDCLAIM_binLow         2.368e-01  2.372e-01   0.998 0.318085    
OLDCLAIM_binMedium      6.485e-01  9.218e-02   7.035 1.99e-12 ***
OLDCLAIM_binHigh        6.757e-01  9.672e-02   6.986 2.83e-12 ***
BLUEBOOK               -2.273e-05  4.725e-06  -4.810 1.51e-06 ***
OLDCLAIM               -2.261e-05  4.751e-06  -4.760 1.94e-06 ***
HAVE_HOME_KIDS1         2.396e-01  8.892e-02   2.694 0.007050 ** 
EDUCATIONBachelors     -3.812e-01  1.115e-01  -3.419 0.000628 ***
EDUCATIONMasters       -2.742e-01  1.636e-01  -1.676 0.093773 .  
EDUCATIONPhD           -2.605e-01  1.973e-01  -1.320 0.186743    
EDUCATIONz_High School  4.185e-03  9.706e-02   0.043 0.965611    
HOME_OWNER1            -2.950e-01  8.200e-02  -3.598 0.000321 ***
PARENT1Yes              2.182e-01  1.211e-01   1.803 0.071452 .  
```

```
KIDSDRIV                1.969e-01  1.215e-01   1.620 0.105193
---
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: 7244.6  on 8123  degrees of freedom
AIC: 7320.6

Number of Fisher Scoring iterations: 5
```

```
Analysis of Deviance Table (Type II tests)

Response: TARGET_FLAG
               Df    Chisq Pr(>Chisq)
URBANICITY      1 434.9173  < 2.2e-16 ***
JOB             8  60.5175  3.689e-10 ***
MVR_PTS         1  42.4925  7.095e-11 ***
MSTATUS         1  40.6782  1.795e-10 ***
CAR_TYPE        5  98.4544  < 2.2e-16 ***
REVOKED         1 108.3191  < 2.2e-16 ***
DO_KIDS_DRIVE   1   2.2091  0.1371978
INCOME_bin      3  50.3637  6.684e-11 ***
CAR_USE         1  67.3921  2.225e-16 ***
TRAVTIME        1  60.5929  7.018e-15 ***
TIF             1  55.0550  1.172e-13 ***
OLDCLAIM_bin    3  68.6572  8.275e-15 ***
BLUEBOOK        1  23.1322  1.512e-06 ***
OLDCLAIM        1  22.6564  1.937e-06 ***
HAVE_HOME_KIDS  1   7.2601  0.0070504 **
EDUCATION       4  22.1186  0.0001898 ***
HOME_OWNER      1  12.9451  0.0003208 ***
PARENT1         1   3.2493  0.0714522 .
KIDSDRIV        1   2.6250  0.1051933
```

```
$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                             0.230763
Cox and Snell (ML)                   0.233795
Nagelkerke (Cragg and Uhler)         0.341489

$Likelihood.ratio.test
 Df.diff LogLik.diff  Chisq p.value
     -37     -1086.7 2173.3       0
```

Observations: Figure 19 shows a summary of the logistic regression model TARGET_FLAG ~ URBANICITY + JOB + MVR_PTS + MSTATUS + CAR_TYPE + REVOKED + DO_KIDS_DRIVE + INCOME_bin + CAR_USE + TRAVTIME + TIF + OLDCLAIM_bin + BLUEBOOK + OLDCLAIM + HAVE_HOME_KIDS + EDUCATION + HOME_OWNER + PARENT1 + KIDSDRIV. These variables were chosen using stepwise regression. The deviance of residuals, which is a measure of model fit of a generalized linear model, shows that the null deviance is 9418.0 and the residual deviance is 7244.6. Since a null deviance shows how well the response variable is predicted by the model that includes only the intercept, the results shows that there was a significant reduction in deviance, even though the deviance of the residuals are high. The results also show an AIC of 7320.6. The Analysis of Deviance tables shows the difference between the null deviance and the residual deviance. In comparison to model 1 where many variables were insignificant, in Model 2, majority of the variables are significant. Additionally, the table shows that adding variables such as URBANCITY, REVOKED, and CAR_TYPE significantly reduces the residual deviance (similar to Model 1), whereas variables such as DO_KIDS_DRIVE, PARENT1, and KIDSDRIV seem to improve the model less as indicated by the low deviance and large p-values. DO_KIDS_DRIVE (duplicative with KIDSDRIVE) and PARENT1 (insignificant and low deviance) will be dropped from my next model so that we can create a more simple and parsimonious model. Lastly, the results show Pseudo R-Square Metrics for McFadden: 0.230763, Cox and Snell (ML): 0.233795, and Nagelkerke (Cragg and Uhler): 0.341489.

For the most part, the coefficients in the model make sense. For example, the positive coefficient for URBANICITYUrban suggests that if a customer lives in an urban area, a customer is more likely to get into a car crash because the area is more heavily populated than a rural area. Furthermore, variables such as MVR_PTS, MSTATUS_NO, REVOKEDYes, DO_KIDS_DRIVE1, TRAVTIME, HAVE_HOME_KIDS1, and KIDSDRIV, which have positive coefficients also make intuitive insurance sense. For instance, customers who get more traffic tickets, have longer commutes (greater risk), and have a lot of teenagers that drive their car are more likely to get into car crashes. Furthermore, customers who are single (possibly less responsible) vs. customers who are married (more responsible), customers who had their license revoked (more reckless) vs. customers who didn't (less reckless) are also more likely to get into car crashes.

On the other hand, variables such as CAR_USEPrivate, TIF, BLUEBOOK, OLDCLAIM, and HOME_OWNER1, which have negative coefficients also make intuitive insurance sense. For example, customers who use their car for non-commercial (drive less) in comparison to those who use their car for commercial use (drive more) are less likely to get into a collision. Furthermore, people who have been customers for a longer time are usually safer and are less likely to get into a car crash. Additionally, a car that is worth more, might mean that the car is newer, or has better performance and therefore decreases the likelihood of getting into a car crash. Furthermore, high total payout of past claims suggests that a customer is more likely to get into a car crash in the future. Likewise, customers who own a home vs. customers who don't are possibly more responsible drivers and therefore are also less likely to get into a car crash.

There were also some coefficients that did not make intuitive sense or were "interesting" that I wanted to point out. For instance, I would think that all of the white collar jobs would have a negative coefficient, but this was not the case (e.g., JOBLawyer, JOBProfessional). This is something that could be explored further. Furthermore, it seemed like if you drove a CAR_TYPESports Car you are more likely to get into a car crash (highest positive coefficient) within the CAR_TYPE variable when compared to CAR_TYPESports Minivans. Also, those who have medium or high incomes are also less likely to get into a car crash than those who have less income (smaller negative coefficient than Low income). Furthermore, customers who have low total payouts of past claims are also less likely to get into car crashes than those who have medium and high payouts. Lastly, customers who are more educated (drive more safely) than customers who are less educated are less likely to get into a car crash.

Model 3

Figure 20: Model 3 (Reduced Model + Transformations)

```
> summary(Model3)

Call:
glm(formula = TARGET_FLAG ~ URBANICITY + JOB + MVR_PTS + MSTATUS +
    CAR_TYPE + REVOKED + INCOME_bin + CAR_USE + SQRT_TRAVTIME +
    SQRT_TIF + OLDCLAIM_bin + LOG_BLUEBOOK + OLDCLAIM + HAVE_HOME_KIDS +
    EDUCATION + HOME_OWNER + KIDSDRIV, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4345  -0.7075  -0.3900   0.6295   3.1412

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -7.165e-01  5.984e-01  -1.197 0.231147
URBANICITYUrban         2.362e+00  1.133e-01  20.849  < 2e-16 ***
JOBClerical             4.340e-01  1.969e-01   2.204 0.027521 *
JOBDoctor              -3.476e-01  2.653e-01  -1.310 0.190144
JOBHome Maker           1.093e-01  2.168e-01   0.504 0.614322
JOBLawyer               1.375e-01  1.695e-01   0.811 0.417137
JOBManager             -5.216e-01  1.711e-01  -3.048 0.002302 **
JOBProfessional         1.902e-01  1.788e-01   1.064 0.287468
JOBStudent              1.041e-02  2.251e-01   0.046 0.963093
JOBz_Blue Collar        3.548e-01  1.860e-01   1.908 0.056389 .
MVR_PTS                 9.396e-02  1.445e-02   6.504 7.82e-11 ***
MSTATUSz_No             6.504e-01  7.429e-02   8.755  < 2e-16 ***
CAR_TYPEPanel Truck     5.245e-01  1.450e-01   3.618 0.000297 ***
CAR_TYPEPickup          5.642e-01  1.008e-01   5.596 2.19e-08 ***
CAR_TYPESports Car      9.409e-01  1.081e-01   8.704  < 2e-16 ***
CAR_TYPEVan             6.560e-01  1.228e-01   5.343 9.16e-08 ***
CAR_TYPEz_SUV           7.013e-01  8.615e-02   8.140 3.95e-16 ***
REVOKEDYes              9.730e-01  9.334e-02  10.423  < 2e-16 ***
INCOME_binLow          -5.468e-01  1.343e-01  -4.073 4.64e-05 ***
INCOME_binMedium       -6.477e-01  1.537e-01  -4.215 2.50e-05 ***
INCOME_binHigh         -1.103e+00  1.690e-01  -6.524 6.85e-11 ***
CAR_USEPrivate         -7.571e-01  9.225e-02  -8.206 2.28e-16 ***
SQRT_TRAVTIME           1.664e-01  2.096e-02   7.940 2.02e-15 ***
SQRT_TIF               -2.496e-01  3.256e-02  -7.666 1.77e-14 ***
OLDCLAIM_binLow         2.543e-01  2.372e-01   1.072 0.283699
OLDCLAIM_binMedium      6.547e-01  9.221e-02   7.100 1.25e-12 ***
OLDCLAIM_binHigh        6.777e-01  9.677e-02   7.003 2.51e-12 ***
LOG_BLUEBOOK           -3.121e-01  5.537e-02  -5.636 1.74e-08 ***
OLDCLAIM               -2.263e-05  4.752e-06  -4.763 1.91e-06 ***
HAVE_HOME_KIDS1         3.609e-01  6.810e-02   5.300 1.16e-07 ***
EDUCATIONBachelors     -3.723e-01  1.116e-01  -3.336 0.000850 ***
EDUCATIONMasters       -2.614e-01  1.638e-01  -1.596 0.110584
EDUCATIONPhD           -2.578e-01  1.975e-01  -1.306 0.191678
EDUCATIONz_High School  1.103e-02  9.715e-02   0.114 0.909618
HOME_OWNER1            -2.905e-01  8.179e-02  -3.552 0.000382 ***
KIDSDRIV                3.528e-01  5.891e-02   5.988 2.12e-09 ***
```

```
---
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: 7234.3  on 8125  degrees of freedom
AIC: 7306.3

Number of Fisher Scoring iterations: 5
```

```
> Anova(Model3, type="II", test="Wald")
Analysis of Deviance Table (Type II tests)

Response: TARGET_FLAG
               Df   Chisq Pr(>Chisq)
URBANICITY      1 434.700  < 2.2e-16 ***
JOB             8  60.761  3.304e-10 ***
MVR_PTS         1  42.303  7.818e-11 ***
MSTATUS         1  76.650  < 2.2e-16 ***
CAR_TYPE        5  97.998  < 2.2e-16 ***
REVOKED         1 108.647  < 2.2e-16 ***
INCOME_bin      3  51.695  3.478e-11 ***
CAR_USE         1  67.344  2.280e-16 ***
SQRT_TRAVTIME   1  63.050  2.015e-15 ***
SQRT_TIF        1  58.775  1.768e-14 ***
OLDCLAIM_bin    3  69.438  5.631e-15 ***
LOG_BLUEBOOK    1  31.765  1.740e-08 ***
OLDCLAIM        1  22.684  1.909e-06 ***
HAVE_HOME_KIDS  1  28.090  1.158e-07 ***
EDUCATION       4  21.661  0.0002341 ***
HOME_OWNER      1  12.618  0.0003819 ***
KIDSDRIV        1  35.861  2.120e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                             0.231859
Cox and Snell (ML)                   0.234764
Nagelkerke (Cragg and Uhler)         0.342904

$Likelihood.ratio.test
 Df.diff LogLik.diff  Chisq p.value
     -35     -1091.8 2183.6       0
```

Observations: Figure 20 shows a summary of the logistic regression model TARGET_FLAG ~ URBANICITY + JOB + MVR_PTS + MSTATUS + CAR_TYPE + REVOKED + INCOME_bin + CAR_USE + SQRT_TRAVTIME + SQRT_TIF + OLDCLAIM_bin + LOG_BLUEBOOK + OLDCLAIM + HAVE_HOME_KIDS + EDUCATION + HOME_OWNER + KIDSDRIV. This model removes DO_KIDS_DRIVE1 and PARENT1 and also adds log and sqrt transformations on TRAVTIME, TIF, and BLUEBOOK to improve the model fit. The deviance of residuals, which is a measure of model fit of a generalized linear model, shows that the null deviance is 9418.0 and the residual deviance is 7234.3. Since a null deviance shows how well the response variable is predicted by the model that includes only the intercept, the results shows that there was a significant reduction in deviance, even though the deviance of the residuals are high. The results also show an AIC of 7306.3, which is lower than Models 1 and 2. The Analysis of Deviance tables shows the difference between the null deviance and the residual deviance. The results show that all the variables are significant and are similar to Models 1 & 2. Additionally, the table shows that adding variables such as URBANCITY, REVOKED, and CAR_TYPE significantly reduces the residual deviance (similar to Models 1 and 2). It’s also important to note that in the summary statistics, some variables within EDUCATION and JOB are not statistically significant, but I left them in the model because removing them degrades my fit statistics (e.g., AIC, BIC, etc.). The results show Pseudo R-Square Metrics for McFadden: 0.231859, Cox and Snell (ML): 0.234764, and Nagelkerke (Cragg and Uhler 0.342904.

For the most part, the coefficients in the model make sense and are very similar to Model 2 (stepwise). For example, variables such as URBANICITYUrban, MVR_PTS, MSTATUS_NO, REVOKEDYes, SQRT_TRAVTIME, HAVE_HOME_KIDS1, and KIDSDRIV, which have positive coefficients make intuitive insurance sense (more likely to get into a car crash). On the other hand, variables such as

CAR_USEPrivate, SQRT_TIF, LOG_BLUEBOOK, OLDCLAIM, and HOME_OWNER1, which have negative coefficients also make intuitive insurance sense (less likely to get into a car crash).

There were also some coefficients that still did not make intuitive sense or were “interesting” that I wanted to point out. For instance, JOBLawyer and JOBProfessional still showed up positive when in theory it should be negative (people with white collar jobs are safer drivers). This is something that could be explored further. Furthermore, CAR_TYPESports Car continued to have high positive coefficients within the CAR_TYPE variable when compared to CAR_TYPESports Minivans, indicating that people who drive these cars are more likely to get into a car crash. Additionally, those who have medium or high incomes are also less likely to get into a car crash than those who have less income (smaller negative coefficient than low income). Lastly, customers who are more educated (drive more safely) than customers who are less educated are less likely to get into a car crash.

Model 4

BONUS: Figure 21: Model 4 (Reduced Model + Transformations using Probit Link)

```
> summary(Model4)

Call:
glm(formula = TARGET_FLAG ~ URBANICITY + JOB + MVR_PTS + MSTATUS +
    CAR_TYPE + REVOKED + INCOME_bin + CAR_USE + SQRT_TRAVTIME +
    SQRT_TIF + OLDCLAIM_bin + LOG_BLUEBOOK + OLDCLAIM + HAVE_HOME_KIDS +
    EDUCATION + HOME_OWNER + KIDSDRIV, family = binomial(link = "probit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4558  -0.7264  -0.3904   0.6562   3.4368

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -3.919e-01  3.449e-01  -1.136 0.255850
URBANICITYUrban        1.298e+00  5.889e-02  22.035  < 2e-16 ***
JOBClerical            2.580e-01  1.134e-01   2.276 0.022865 *
JOBDoctor             -2.103e-01  1.471e-01  -1.430 0.152834
JOBHome Maker          5.104e-02  1.252e-01   0.408 0.683467
JOBLawyer              8.152e-02  9.749e-02   0.836 0.403023
JOBManager            -2.748e-01  9.713e-02  -2.829 0.004667 **
JOBProfessional        1.285e-01  1.028e-01   1.250 0.211322
JOBStudent             2.166e-02  1.299e-01   0.167 0.867610
JOBz_Blue Collar       2.225e-01  1.070e-01   2.080 0.037496 *
MVR_PTS                5.446e-02  8.498e-03   6.409 1.47e-10 ***
MSTATUSz_No            3.744e-01  4.291e-02   8.725  < 2e-16 ***
CAR_TYPEPanel Truck    2.845e-01  8.385e-02   3.393 0.000691 ***
CAR_TYPEPickup         3.107e-01  5.779e-02   5.376 7.61e-08 ***
CAR_TYPESports Car     5.344e-01  6.204e-02   8.614  < 2e-16 ***
CAR_TYPEVan            3.632e-01  7.049e-02   5.152 2.57e-07 ***
CAR_TYPEz_SUV          3.964e-01  4.896e-02   8.097 5.65e-16 ***
REVOKEDYes             5.610e-01  5.438e-02  10.316  < 2e-16 ***
INCOME_binLow         -3.186e-01  7.806e-02  -4.081 4.48e-05 ***
INCOME_binMedium      -3.893e-01  8.928e-02  -4.360 1.30e-05 ***
INCOME_binHigh        -6.446e-01  9.778e-02  -6.592 4.33e-11 ***
```

```
CAR_USEPrivate         -4.280e-01  5.355e-02  -7.992 1.33e-15 ***
SQRT_TRAVTIME           9.511e-02  1.198e-02   7.937 2.07e-15 ***
SQRT_TIF               -1.467e-01  1.872e-02  -7.834 4.72e-15 ***
OLDCLAIM_binLow         1.552e-01  1.411e-01   1.100 0.271328
OLDCLAIM_binMedium      3.900e-01  5.427e-02   7.186 6.66e-13 ***
OLDCLAIM_binHigh        3.978e-01  5.663e-02   7.025 2.14e-12 ***
LOG_BLUEBOOK           -1.769e-01  3.190e-02  -5.544 2.95e-08 ***
OLDCLAIM               -1.277e-05  2.773e-06  -4.604 4.14e-06 ***
HAVE_HOME_KIDS1         2.088e-01  3.950e-02   5.287 1.24e-07 ***
EDUCATIONBachelors     -2.115e-01  6.491e-02  -3.258 0.001120 **
EDUCATIONMasters       -1.367e-01  9.363e-02  -1.460 0.144226
EDUCATIONPhD           -1.312e-01  1.122e-01  -1.170 0.242015
EDUCATIONz_High School  1.530e-02  5.662e-02   0.270 0.786913
HOME_OWNER1            -1.646e-01  4.770e-02  -3.451 0.000558 ***
KIDSDRIV                1.999e-01  3.436e-02   5.820 5.90e-09 ***
---
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: 7244.6  on 8125  degrees of freedom
AIC: 7316.6
Number of Fisher Scoring iterations: 5
```

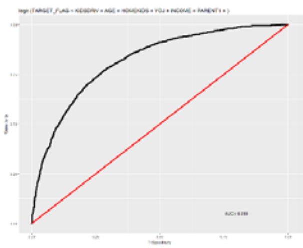
```
> Anova(Model4, type="II", test="Wald")
Analysis of Deviance Table (Type II tests)

Response: TARGET_FLAG
               Df   Chisq Pr(>Chisq)
URBANICITY      1 485.522  < 2.2e-16 ***
JOB             8  61.027  2.929e-10 ***
MVR_PTS         1  41.074  1.466e-10 ***
MSTATUS         1  76.124  < 2.2e-16 ***
CAR_TYPE        5  97.394  < 2.2e-16 ***
REVOKED         1 106.413  < 2.2e-16 ***
INCOME_bin      3  51.947  3.074e-11 ***
CAR_USE         1  63.867  1.331e-15 ***
SQRT_TRAVTIME   1  62.998  2.069e-15 ***
SQRT_TIF        1  61.375  4.718e-15 ***
OLDCLAIM_bin    3  71.251  2.304e-15 ***
LOG_BLUEBOOK    1  30.741  2.949e-08 ***
OLDCLAIM        1  21.199  4.139e-06 ***
HAVE_HOME_KIDS  1  27.957  1.240e-07 ***
EDUCATION       4  22.171  0.0001853 ***
HOME_OWNER      1  11.911  0.0005580 ***
KIDSDRIV        1  33.868  5.899e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                             0.230771
Cox and Snell (ML)                   0.233802
Nagelkerke (Cragg and Uhler)         0.341499

$Likelihood.ratio.test
 Df.diff LogLik.diff  Chisq p.value
     -35     -1086.7 2173.4       0
```

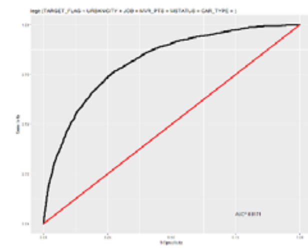
Observations: Figure 21 shows a summary of the logistic regression model using a probit link TARGET_FLAG ~ URBANICITY + JOB + MVR_PTS + MSTATUS + CAR_TYPE + REVOKED + INCOME_bin + CAR_USE + SQRT_TRAVTIME + SQRT_TIF + OLDCLAIM_bin + LOG_BLUEBOOK + OLDCLAIM + HAVE_HOME_KIDS + EDUCATION + HOME_OWNER + KIDSDRIV. This model is the same as Model 3, but instead of using a logit link, we are using a probit link instead. The deviance of residuals, which is a measure of model fit of a generalized linear model, shows that the null deviance is 9418.0 and the residual deviance is 7244.6. Since a null deviance shows how well the response variable is predicted by the model that includes only the intercept, the results shows that there was a significant reduction in deviance, even though the deviance of the residuals are high. The results also show an AIC of 7316.6. The Analysis of Deviance tables shows the difference between the null deviance and the residual deviance. The results show that all the variables are significant, similar to the other models. Additionally, the table shows that adding variables such as URBANCITY, REVOKED, and CAR_TYPE significantly reduces the residual deviance (similar to Models 1 to 3). The results show Pseudo R-Square Metrics for McFadden: 0.230771, Cox and Snell (ML): 0.233802, and Nagelkerke (Cragg and Uhler): 0.341499.

Section 4: Selection Models

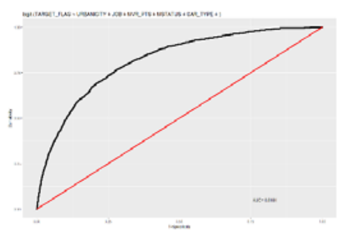
Figure 22: ROC Curves & Area Under Curve (AUC) for Models & KS Statistic



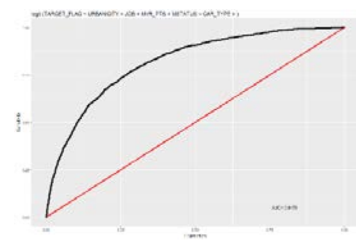
Model 1



Model 2



Model 3



Model 4

Model Name	KS Statistic
Model 1 (Full)	0.4851
Model 2 (Stepwise)	0.4876
Model 3 (Reduced + Transformations)	0.4826
Model 4 (Probit Model)	0.4826

Observations: Figure 22 shows ROC Curves along with the Area Under the Curve or "AUC" and KS Statistic for all 4 models. The ROC Curve measures how well a model can differentiate between True Positives (TP) and False Positives (FP), while Area Under the Curve or "AUC" is the percentage of the area that is under the curve (higher the better). All of the ROC curves look similar in the sense that there is a nice bow and all the lift is in the beginning. This means that the models are good at rank ordering and when the model says that there is a high chance that a value is positive, then it is in fact positive. Furthermore, in regards to AUC, Model 3 had the best AUC (0.8181), followed by Model 1, Model 4, and Model 2. This means that for Model 3, there would be an 81.81% chance that the true positive value would be greater than the true negative value because the AUC is 81.81%.

The KS statistic is a metric that measures the maximum difference the cumulative true positive rate and the cumulative false positive. In other words, it measures the difference between the proportion of those who were in a car crash and those who were not in a car crash. It is a way of determining to what extent our model discriminates two groups of drivers: those who were in a car crash and those who were not in a car crash. The KS Statistic for all 4 models are all around 0.48.

Figure 23: Model Comparison and Criteria for Selecting the "Best Model"

Model Name	AIC	Rank	BIC	Rank	Log Likelihood	Rank	ROC Curve (AUC)	Rank	Total Points & Best Model*
Model 1 (Full)	7336.56	4	7693.923	4	7234.56	2	0.8180	2	8
Model 2 (Stepwise)	7320.643	3	7586.913	3	7244.643	1	0.8174	4	9
Model 3 (Reduced + Transformations)	7306.321	1	7558.577	1	7234.321	2	0.8181	1	15*
Model 4 (Probit Model)	7316.573	2	7568.829	2	7244.573	1	0.8179	3	12

**Points: Rank 1 = 4 points, Rank 2 = 3 points, Rank 3 = 2 points, Rank 4 = 1 point*

Observations: Figure 23 shows the model comparisons so that we can compare the in-sample fit and predictive accuracy of our models so that we can select the best model. The results above show the computations for AIC, BIC, log likelihood, and AUC for each of these models. Each of these metrics represent some concept of 'fit' (e.g., rewarding for accuracy and penalizing for complexity). Additionally, each model was ranked on each metric. Points were then allotted to each model based on how they ranked on each metric.

As a result, given the criteria above. Model 3 is the best model because it ranked in the upper echelon on all the metrics and as a result received the most total points (e.g., was the most accurate and most parsimonious model).

The formula/coefficients for Model 3 to predict the probability that a person will crash their car is:

```
> coef(Model3)
        (Intercept)      URBANICITYUrban          JOBClerical            JOBDoctor
      -7.165286e-01         2.362444e+00         4.339772e-01        -3.476275e-01
     JOBHome Maker            JOBLawyer           JOBManager      JOBProfessional
       1.092646e-01         1.375239e-01        -5.216251e-01         1.902077e-01
         JOBStudent      JOBz_Blue Collar              MVR_PTS          MSTATUSz_No
       1.041397e-02         3.548308e-01         9.395602e-02         6.504291e-01
  CAR_TYPEPanel Truck      CAR_TYPEPickup   CAR_TYPESports Car          CAR_TYPEVan
       5.244682e-01         5.642352e-01         9.409179e-01         6.559717e-01
       CAR_TYPEz_SUV           REVOKEDYes        INCOME_binLow     INCOME_binMedium
       7.012982e-01         9.729508e-01        -5.468149e-01        -6.476624e-01
      INCOME_binHigh       CAR_USEPrivate         SQRT_TRAVTIME             SQRT_TIF
      -1.102723e+00        -7.570603e-01         1.664025e-01        -2.496234e-01
    OLDCLAIM_binLow    OLDCLAIM_binMedium      OLDCLAIM_binHigh          LOG_BLUEBOOK
       2.542819e-01         6.546705e-01         6.776900e-01        -3.120594e-01
           OLDCLAIM       HAVE_HOME_KIDS1   EDUCATIONBachelors     EDUCATIONMasters
      -2.263104e-05         3.609349e-01        -3.723003e-01        -2.613930e-01
       EDUCATIONPhD EDUCATIONz_High School           HOME_OWNER1             KIDSDRIV
      -2.578184e-01         1.102833e-02        -2.905255e-01         3.527899e-01
```

For the most part, the coefficients in the model make sense. For example, the positive coefficient for URBANICITYUrban suggests that if a customer lives in an urban area, a customer is more likely to get into a car crash because the area is more heavily populated than a rural area. Furthermore, variables such as MVR_PTS, MSTATUS_NO, REVOKEDYes, SQRT_TRAVTIME, HAVE_HOME_KIDS1, and KIDSDRIV, which have positive coefficients also make intuitive insurance sense. For instance, customers who get more traffic tickets, have longer commutes (greater risk), and have a lot of teenagers that drive their car are more likely to get into car crashes. Furthermore, customers who are single (possibly less responsible) vs. customers who are married (more responsible), customers who had their license revoked (more reckless) vs. customers who didn't (less reckless) are also more likely to get into car crashes.

On the other hand, variables such as CAR_USEPrivate, SQRT_TIF, LOG_BLUEBOOK, OLDCLAIM, and HOME_OWNER1, which have negative coefficients also make intuitive insurance sense. For example, customers who use their car for non-commercial (drive less) in comparison to those who use their car for commercial use (drive more) are less likely to get into a collision. Furthermore, people who have been customers for a longer time are usually safer and are less likely to get into a car crash. Additionally, a car that is worth more, might mean that the car is newer, or has better performance and therefore decreases the likelihood of getting into a car crash. Furthermore, high total payout of past claims suggests that a customer is more likely to get into a car crash in the future. Likewise, customers who own a home vs. customers who don't are possibly more responsible drivers and therefore are also less likely to get into a car crash.

There were also some coefficients that did not make intuitive sense or were "interesting" that I wanted to point out. For instance, I would think that all of the white collar jobs would have a negative coefficient, but this was not the case (e.g., JOBLawyer, JOBProfessional). This is something that could be explored further in the future (e.g., conducting interaction variables). It's also important to note that in the summary statistics, variables within EDUCATION and JOB are not statistically significant, but I left them in the model because removing them degrades my fit statistics (e.g., AIC, BIC, etc.). Furthermore, it seemed like if you drove a CAR_TYPESports Car you are more likely to get into a car crash (highest positive coefficient) within the CAR_TYPE variable when compared to CAR_TYPESports Minivans. Also, those who have medium or high incomes are also less likely to get into a car crash than those who have less income (smaller negative coefficient than Low income). Furthermore, customers who have low total payouts of past claims are also less likely to get into car crashes than those who have medium and high payouts. Lastly, customers who are more educated (drive more safely) than customers who are less educated are less likely to get into a car crash.

Section 5: Stand Alone Scoring Program

#Part 5: Stand Alone Scoring Program

```
setwd("~/R/Data")
mytest <- read.csv("logit_insurance_test.csv")
```

#Test Data

```
mytest$INDEX <- as.numeric(mytest$INDEX)
mytest$TARGET_FLAG <- as.factor(mytest$TARGET_FLAG)
mytest$SEX <- as.factor(mytest$SEX)
mytest$EDUCATION <- as.factor(mytest$EDUCATION)
mytest$PARENT1 <- as.factor(mytest$PARENT1)
mytest$INCOME <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mytest$INCOME)))

mytest$HOME_VAL <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mytest$HOME_VAL)))
mytest$MSTATUS <- as.factor(mytest$MSTATUS)
mytest$REVOKED <- as.factor(mytest$REVOKED)
mytest$RED_CAR <- as.factor(ifelse(mytest$RED_CAR=="yes", 1, 0))
mytest$URBANICITY <- ifelse(mytest$URBANICITY == "Highly Urban/ Urban", "Urban", "Rural")
mytest$URBANICITY <- as.factor(mytest$URBANICITY)
mytest$JOB <- as.factor(mytest$JOB)
mytest$CAR_USE <- as.factor(mytest$CAR_USE)
mytest$CAR_TYPE <- as.factor(mytest$CAR_TYPE)
mytest$DO_KIDS_DRIVE <- as.factor(ifelse(mytest$KIDSDRIV > 0, 1, 0 ))
mytest$OLDCLAIM <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mytest$OLDCLAIM)))
mytest$BLUEBOOK <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mytest$BLUEBOOK)))
summary(mytest)
```

 # Fix NA's for Test Data

```
library(mice)
```

#Check for missing values

```
sapply(mytest, function(x) sum(is.na(x)))
```

#Check missing data percentage

```
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(mytest,2,pMiss)

library(VIM)
aggr_plot <- aggr(mytest, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(mytest),
cex.axis=.5, gap=2, ylab=c("Histogram of missing data","Pattern"))
```

#Split datasets into numerical and categorical

#Numeric
subdatnumtest <- subset(mytest, select=c(
"INDEX",
"KIDSDRIV",
"AGE",
"HOMEKIDS",
"YOJ",
"INCOME",
"HOME_VAL",
"TRAVTIME",
"BLUEBOOK",
"TIF",
"OLDCLAIM",
"CLM_FREQ",
"MVR_PTS",
"CAR_AGE",
"TARGET_AMT"))

subdatnumtest.df <- data.frame(subdatnumtest)

#Categorical
subdatcattest <- subset(mytest, select=c(
"INDEX",
"TARGET_FLAG",
"PARENT1",
"MSTATUS",
"SEX",
"EDUCATION",
"JOB",
"CAR_USE",
"CAR_TYPE",
"RED_CAR",
"REVOKED",
"URBANICITY",
"DO_KIDS_DRIVE"))

subdatcattest.df <- data.frame(subdatcattest)

 # Fix NA's for Test Data

#Run imputation
tempDatatest <- mice(subdatnumtest.df,m=5,maxit=50,meth='pmm',seed=500)

summary(tempDatatest)

 # Inspecting the distribution of original and imputed data for the variables that contained N/A
xyplot(tempDatatest,TARGET_AMT~ CAR_AGE + HOME_VAL + YOJ + INCOME + AGE,pch=18,cex=1)

densityplot(tempDatatest)

#Check N/A values have been removed
subdatnumimptest <- complete(tempDatatest,1)
apply(subdatnumimptest,2,pMiss)
summary(subdatnumimptest)
sapply(subdatnumimptest, function(x) sum(is.na(x)))

#Merge Numeric and Categorical datasets back
test <- merge(subdatnumimptest, subdatcattest.df, by=c("INDEX"))

#Check data
str (test)
summary (test)

#Trim Test
test$CAR_AGE[test$CAR_AGE < 0] <- 0
test$YOJ [(test$YOJ >= 20)] = 20
test$INCOME [(test$INCOME >= 300000)] = 300000
test$HOME_VAL [(test$HOME_VAL >= 650000)] = 650000
test$TRAVTIME [(test$TRAVTIME >= 100)] = 100
test$BLUEBOOK [(test$BLUEBOOK >= 55000)] = 55000
test$TIF [(test$TIF >= 17)] = 17
test$MVR_PTS [(test$MVR_PTS >=8)] = 8

#Create Flag Variables
test$HAVE_HOME_KIDS <- as.factor(ifelse(test$HOMEKIDS > 0, 1, 0))
test$EMPLOYED <- as.factor(ifelse(test$YOJ > 0, 1, 0))
test$HOME_OWNER <- as.factor(ifelse(test$HOME_VAL> 0, 1, 0))
test$SUBMITTED_CLAIM <- as.factor(ifelse(test$CLM_FREQ > 0, 1, 0))
test$HAVE_MVR_PTS<- as.factor(ifelse(test$MVR_PTS> 0, 1, 0))

#Create SQRT Transformations of Some of the Variables
test$SQRT_TRAVTIME <- sqrt(test$TRAVTIME)
test$SQRT_BLUEBOOK <- sqrt(test$BLUEBOOK)
test$SQRT_TIF <- sqrt(test$TIF)
test$LOG_TRAVTIME <- log(test$TRAVTIME)
test$LOG_BLUEBOOK <- log(test$BLUEBOOK)
test$LOG_TIF <- log(test$TIF)
test$LOG_MVR_PTS <- log(test$MVR_PTS)
test$LOG_OLDCLAIM <- log(test$OLDCLAIM)

 # Bins for Test Data

#Income

```
test$INCOME_bin[is.na(test$INCOME)] <- "NA"
test$INCOME_bin[test$INCOME == 0] <- "Zero"
test$INCOME_bin[test$INCOME >= 1 & test$INCOME < 30000] <- "Low"
test$INCOME_bin[test$INCOME >= 30000 & test$INCOME < 80000] <- "Medium"
test$INCOME_bin[test$INCOME >= 80000] <- "High"
test$INCOME_bin <- factor(test$INCOME_bin)
test$INCOME_bin <- factor(test$INCOME_bin, levels=c("NA","Zero","Low","Medium","High"))
```

#HOME_VAL

```
test$HOME_VAL_bin[is.na(test$HOME_VAL)] <- "NA"
test$HOME_VAL_bin[test$HOME_VAL == 0] <- "Zero"
test$HOME_VAL_bin[test$HOME_VAL >= 1 & test$HOME_VAL < 125000] <- "Low"
test$HOME_VAL_bin[test$HOME_VAL >= 125000 & test$HOME_VAL < 300000] <- "Medium"
test$HOME_VAL_bin[test$HOME_VAL >= 300000] <- "High"
test$HOME_VAL_bin <- factor(test$HOME_VAL_bin)
test$HOME_VAL_bin <- factor(test$HOME_VAL_bin, levels=c("NA","Zero","Low","Medium","High"))
```

#OLDCLAIM

```
test$OLDCLAIM_bin[is.na(test$OLDCLAIM)] <- "NA"
test$OLDCLAIM_bin[test$OLDCLAIM == 0] <- "Zero"
test$OLDCLAIM_bin[test$OLDCLAIM >= 1 & test$OLDCLAIM < 1000] <- "Low"
test$OLDCLAIM_bin[test$OLDCLAIM >= 1000 & test$OLDCLAIM < 4500] <- "Medium"
test$OLDCLAIM_bin[test$OLDCLAIM >= 4500] <- "High"
test$OLDCLAIM_bin <- factor(test$OLDCLAIM_bin)
test$OLDCLAIM_bin <- factor(test$OLDCLAIM_bin, levels=c("NA","Zero","Low","Medium","High"))

summary(test)
```

#Stand Alone Scoring Program

```
data0<- subset(data, TARGET_FLAG == 1 )

test$P_TARGET_FLAG <- predict(Model3, newdata = test, type = "response")

targetbycar <- aggregate(data0$TARGET_AMT, list(data0$CAR_TYPE), mean)

test$P_TARGET_AMT <- ifelse(test$CAR_TYPE=="Minivan", 5601.665%*%.27,
                  ifelse(test$CAR_TYPE=="Panel Truck", 7464.703%*%.27,
                        ifelse(test$CAR_TYPE=="Pickup", 5430.106%*%.27,
                              ifelse(test$CAR_TYPE=="Sports Car", 5412.733%*%.27,
                                    ifelse(test$CAR_TYPE=="Van", 6908.553%*%0.27, 5241.104%*%.27)))))
```

 # Scored Data File

```
#subset of data set for the deliverable "Scored data file"
scores <- test[c("INDEX","P_TARGET_FLAG", "P_TARGET_AMT")]

#Note, this next function will output a csv file in your work environment called write.csv.

write.csv(scores, file = "CI_Scored.csv")
write.csv(as.data.frame(scores), file = "logit_insurance_test.csv",
        sheetName = "Scored Data File", row.names = FALSE)
```

Section 6: Scored Data File

Summary Statistics ***Probability that a person will crash their car for Quality Control Purposes***	
MEAN	0.2718
MEDIAN	0.2133
MAX	0.9623
MIN	0.0025

Summary Statistics ***Target Amount for Quality Control Purposes***	
MEAN	1540.67
MEDIAN	1466.13
MAX	2015.47
MIN	1415.10

Conclusion

In section 1, we conducted an initial exploratory data analysis using scatterplots, boxplots, summary statistics, etc. to help understand important characteristics and properties of the data that may be disguised by numerical summaries. The EDA revealed outliers and missing values for 5 variables: AGE, YOJ, INCOME, HOME_VAL, and CAR_AGE.

In section 2, we conducted data preparation/transformations of the data by fixing the missing values using predictive mean matching, conducting data transformations, binning variables, and handling outliers.

In section 3, we built 4 logistic regression models (3 logit link and 1 probit link) using different variables (or the same variables with different transformations). This was conducted using variable selection techniques. We then ran model diagnostics and discussed the coefficients in the model to ensure that it makes intuitive insurance sense.

In section 4, we selected Model 3 as our “best model” based on ‘fit’ (AIC, BIC, log likelihood, and AUC) metrics.

Lastly, in section 5, a Stand Alone scoring program was conducted that scored the new data and predicted the probably that that a person will crash their car. The summary statistics showed the following: mean (0.2718), median (0.2133), max (0.9623), and min (0.0025). The data step also included all the variable transformations such as fixing missing values and the logistic regression formula.

Fulll Code

#Part 0: Load & Prepare Data

```
library(readr)
library(dplyr)
library(zoo)
library(psych)
library(ROCR)
library(corrplot)
library(car)
library(InformationValue)
library(rJava)
library(pbkrtest)
library(car)
library(leaps)
library(MASS)
library(corrplot)
library(glm2)
library(aod)
library(mice)
library(Hmisc)
library(xlsxjars)
library(xlsx)
library(VIM)
library(pROC)
```

 # Data Import and Variable Type Changes

```
setwd("~/R/Insurance")
mydata <- read.csv("logit_insurance.csv")
```

#Training Data

```
mydata$INDEX <- as.numeric(mydata$INDEX)
mydata$TARGET_FLAG <- as.factor(mydata$TARGET_FLAG)
mydata$SEX <- as.factor(mydata$SEX)
mydata$EDUCATION <- as.factor(mydata$EDUCATION)
mydata$PARENT1 <- as.factor(mydata$PARENT1)
mydata$INCOME <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mydata$INCOME)))

mydata$HOME_VAL <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mydata$HOME_VAL)))
mydata$MSTATUS <- as.factor(mydata$MSTATUS)
mydata$REVOKED <- as.factor(mydata$REVOKED)
mydata$RED_CAR <- as.factor(ifelse(mydata$RED_CAR=="yes", 1, 0))
mydata$URBANICITY <- ifelse(mydata$URBANICITY == "Highly Urban/ Urban", "Urban", "Rural")
mydata$URBANICITY <- as.factor(mydata$URBANICITY)
mydata$JOB <- as.factor(mydata$JOB)
```

```
mydata$CAR_USE <- as.factor(mydata$CAR_USE)
mydata$CAR_TYPE <- as.factor(mydata$CAR_TYPE)
mydata$DO_KIDS_DRIVE <- as.factor(ifelse(mydata$KIDSDRIV > 0, 1, 0 ))
mydata$OLDCLAIM <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mydata$OLDCLAIM)))
mydata$BLUEBOOK <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mydata$BLUEBOOK)))
summary(mydata)
```

#Part 1: Data Exploration

#Mydata Quality Check

```
str(mydata)
summary(mydata)

library(Hmisc)
describe(mydata)
```

 # EDA for Numeric Variables

```
mydata0<- subset(mydata, TARGET_FLAG == 1 )

par(mfrow=c(3,3))
hist(log(mydata0$TARGET_AMT), col = "#A71930", xlab = "Log TARGET_AMT", main = "Log TARGET_AMT
Hist")
hist(mydata$KIDSDRIV, col = "#09ADAD", xlab = "KIDSDRIV", main = "Histogram of KIDSDRIV")
hist(mydata$AGE, col = "#DBCEAC", xlab = "AGE", main = "Histogram of AGE")
boxplot(log(mydata0$TARGET_AMT), col = "#A71930", main = "LOG TARGET_AMT Boxplot")
boxplot(mydata$KIDSDRIV, col = "#09ADAD", main = "Boxplot of KIDSDRIV ")
boxplot(mydata$AGE, col = "#DBCEAC", main = "Boxplot of AGE")
par(mfrow=c(1,1))

par(mfrow=c(3,3))
hist(mydata$HOMEKIDS, col = "#A71930", xlab = "HOMEKIDS", main = "Histogram of HOMEKIDS")
hist(mydata$YOJ, col = "#09ADAD", xlab = "YOJ ", main = "Histogram of YOJ")
hist(mydata$INCOME, col = "#DBCEAC", xlab = "INCOME", main = "Histogram of INCOME")
boxplot(mydata$HOMEKIDS, col = "#A71930", main = "Boxplot of HOMEKIDS")
boxplot(mydata$YOJ, col = "#09ADAD", main = "Boxplot of YOJ")
boxplot(mydata$INCOME, col = "#DBCEAC", main = "Boxplot of INCOME ")
par(mfrow=c(1,1))

par(mfrow=c(3,3))
hist(mydata$HOME_VAL, col = "#A71930", xlab = "HOME_VAL", main = "Histogram of HOME_VAL")
hist(mydata$TRAVTIME, col = "#09ADAD", xlab = "TRAVTIME", main = "Histogram of TRAVTIME")
hist(mydata$BLUEBOOK, col = "#DBCEAC", xlab = "BLUEBOOK", main = "Histogram of BLUEBOOK")
boxplot(mydata$HOME_VAL, col = "#A71930", main = "Boxplot of HOME_VAL")
boxplot(mydata$TRAVTIME, col = "#09ADAD", main = "Boxplot of TRAVTIME")
boxplot(mydata$ BLUEBOOK, col = "#DBCEAC", main = "Boxplot of BLUEBOOK")
par(mfrow=c(1,1))
```

```
par(mfrow=c(3,3))
hist(mydata$TIF, col = "#A71930", xlab = "TIF", main = "Histogram of TIF")
hist(mydata$OLDCLAIM, col = "#09ADAD", xlab = "OLDCLAIM", main = "Histogram of OLDCLAIM ")
hist(mydata$CLM_FREQ, col = "#DBCEAC", xlab = "CLM_FREQ", main = "Histogram of CLM_FREQ")
boxplot(mydata$TIF, col = "#A71930", main = "Boxplot of TIF")
boxplot(mydata$OLDCLAIM, col = "#09ADAD", main = "Boxplot of OLDCLAIM")
boxplot(mydata$CLM_FREQ, col = "#DBCEAC", main = "Boxplot of CLM_FREQ")
par(mfrow=c(1,1))


par(mfrow=c(2,2))
hist(mydata$MVR_PTS, col = "#A71930", xlab = "MVR_PTS", main = "Histogram of MVR_PTS")
hist(mydata$CAR_AGE, col = "#09ADAD", xlab = "CAR_AGE", main = "Histogram of CAR_AGE")
boxplot(mydata$MVR_PTS, col = "#A71930", main = "Boxplot of MVR_PTS")
boxplot(mydata$CAR_AGE, col = "#09ADAD", main = "Boxplot of CAR_AGE")
par(mfrow=c(1,1))
```

 # Scatterplot Matrix

```
panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(~ mydata$TARGET_AMT + mydata$KIDSDRIV + mydata$AGE + mydata$HOMEKIDS + mydata$YOJ +
mydata$INCOME + mydata$HOME_VAL + mydata$TRAVTIME, lower.panel = panel.smooth)
par(mfrow=c(1,1))

pairs(~ mydata$TARGET_AMT + mydata$BLUEBOOK+ mydata$TIF+ mydata$OLDCLAIM + mydata$CLM_FREQ +
mydata$MVR_PTS + mydata$CAR_AGE, lower.panel = panel.smooth)
par(mfrow=c(1,1))
```

#Correlation Matrix

```
subdatnumcor <- subset(mydata, select=c(
"KIDSDRIV",
"AGE",
"HOMEKIDS",
"YOJ",
"INCOME",
"HOME_VAL",
```

```
"TRAVTIME",
"BLUEBOOK",
"TIF",
"OLDCLAIM",
"CLM_FREQ",
"MVR_PTS",
"CAR_AGE",
"TARGET_AMT"))

require(corrplot)
mcor <- cor(subdatnumcor)
corrplot(mcor, method="number", shade.col=NA, tl.col="black",tl.cex=0.8)
```

#EDA for Categorical Variables

```
library(ggplot2)
```

#TARGET_FLAG

```
require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(TARGET_FLAG) ) +
  ggtitle("TARGET_FLAG") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))
```

 # PARENT1

```
require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(PARENT1) ) +
  ggtitle("PARENT1") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))
```

#MSTATUS

```
require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(MSTATUS) ) +
  ggtitle("MSTATUS") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))
```

 # SEX

```
require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(SEX) ) +
  ggtitle("SEX") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))
```

```
# EDUCATION
require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(EDUCATION) ) +
  ggtitle("EDUCATION") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))


#JOB

require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(JOB) ) +
  ggtitle("JOB") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))


#CAR_USE
require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(CAR_USE) ) +
  ggtitle("CAR_USE") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))

#CARTYPE
require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(CAR_TYPE) ) +
  ggtitle("CAR_TYPE") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))

#RED_CAR
require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(RED_CAR) ) +
  ggtitle("RED_CAR") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))


#REVOKED
require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(REVOKED) ) +
  ggtitle("REVOKED") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))
```

```
#URBANICITY

require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(URBANICITY) ) +
  ggtitle("URBANICITY") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))

#DO_KIDS_DRIVE
require(ggplot2)
ggplot(mydata) +
  geom_bar( aes(DO_KIDS_DRIVE) ) +
  ggtitle("DO_KIDS_DRIVE") +
  theme(plot.title=element_text(lineheight=0.8, face="bold", hjust=0.5))

########### Part 2: Data Transformation ##################

#Part 2: Data Preparation
library(mice)

#Check for missing values
sapply(mydata, function(x) sum(is.na(x)))

#Check missing data percentage
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(mydata,2,pMiss)

library(VIM)
aggr_plot <- aggr(mydata, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(mydata),
cex.axis=.5, gap=2, ylab=c("Histogram of missing data","Pattern"))

#Split datasets into numerical and categorical

#Numeric
subdatnum <- subset(mydata, select=c(
"INDEX",
"KIDSDRIV",
"AGE",
"HOMEKIDS",
"YOJ",
"INCOME",
"HOME_VAL",
"TRAVTIME",
"BLUEBOOK",
"TIF",
"OLDCLAIM",
"CLM_FREQ",
"MVR_PTS",
```

```
"CAR_AGE",
"TARGET_AMT"))

subdatnum.df <- data.frame(subdatnum)


#Categorical
subdatcat <- subset(mydata, select=c(
"INDEX",
"TARGET_FLAG",
"PARENT1",
"MSTATUS",
"SEX",
"EDUCATION",
"JOB",
"CAR_USE",
"CAR_TYPE",
"RED_CAR",
"REVOKED",
"URBANICITY",
"DO_KIDS_DRIVE"))

subdatcat.df <- data.frame(subdatcat)

# Fix NA's for Training Data

#Run imputation
tempData <- mice(subdatnum.df,m=5,maxit=50,meth='pmm',seed=500)

summary(tempData)

# Inspecting the distribution of original and imputed data for the variables that contained N/A
xyplot(tempData,TARGET_AMT~ CAR_AGE + HOME_VAL + YOJ +  INCOME + AGE,pch=18,cex=1)

densityplot(tempData)

#Check N/A values have been removed
subdatnumimp <- complete(tempData,1)
apply(subdatnumimp,2,pMiss)
summary(subdatnumimp)
sapply(subdatnumimp, function(x) sum(is.na(x)))


#Merge Numeric and Categorical datasets back
data <- merge(subdatnumimp, subdatcat.df, by=c("INDEX"))

#Check data
str (data)
```

```
summary (data)

#Trim Data
data$CAR_AGE[data$CAR_AGE < 0 ] <- 0
data$YOJ [(data$YOJ >= 20)] = 20
data$INCOME [(data$INCOME >= 300000)] = 300000
data$HOME_VAL  [(data$HOME_VAL  >= 650000)] = 650000
data$TRAVTIME [(data$TRAVTIME >= 100)] = 100
data$BLUEBOOK [(data$BLUEBOOK >= 55000)] = 55000
data$TIF [(data$TIF >= 17)] = 17
data$MVR_PTS [(data$MVR_PTS >=8)] = 8

#Create Flag Variables
data$HAVE_HOME_KIDS <- as.factor(ifelse(data$HOMEKIDS > 0, 1, 0 ))
data$EMPLOYED <- as.factor(ifelse(data$YOJ > 0, 1, 0 ))
data$HOME_OWNER <- as.factor(ifelse(data$HOME_VAL> 0, 1, 0 ))
data$SUBMITTED_CLAIM <- as.factor(ifelse(data$CLM_FREQ  > 0, 1, 0 ))
data$HAVE_MVR_PTS<- as.factor(ifelse(data$MVR_PTS> 0, 1, 0 ))

#Create SQRT Transformations of Some of the Variables
data$SQRT_TRAVTIME <- sqrt(data$TRAVTIME)
data$SQRT_BLUEBOOK <- sqrt(data$BLUEBOOK)
data$SQRT_TIF <- sqrt(data$TIF)
data$LOG_TRAVTIME <- log(data$TRAVTIME)
data$LOG_BLUEBOOK <- log(data$BLUEBOOK)
data$LOG_TIF <- log(data$TIF)
data$LOG_MVR_PTS <- log(data$MVR_PTS)
data$LOG_OLDCLAIM <- log(data$OLDCLAIM)

# Bins for Training Data

#Income
data$INCOME_bin[is.na(data$INCOME)] <- "NA"
data$INCOME_bin[data$INCOME == 0] <- "Zero"
data$INCOME_bin[data$INCOME >= 1 & data$INCOME < 30000] <- "Low"
data$INCOME_bin[data$INCOME >= 30000 & data$INCOME < 80000] <- "Medium"
data$INCOME_bin[data$INCOME >= 80000] <- "High"
data$INCOME_bin <- factor(data$INCOME_bin)
data$INCOME_bin <- factor(data$INCOME_bin, levels=c("NA","Zero","Low","Medium","High"))

#HOME_VAL
data$HOME_VAL_bin[is.na(data$HOME_VAL)] <- "NA"
data$HOME_VAL_bin[data$HOME_VAL == 0] <- "Zero"
data$HOME_VAL_bin[data$HOME_VAL >= 1 & data$HOME_VAL < 125000] <- "Low"
data$HOME_VAL_bin[data$HOME_VAL >= 125000 & data$HOME_VAL < 300000] <- "Medium"
data$HOME_VAL_bin[data$HOME_VAL >= 300000] <- "High"
data$HOME_VAL_bin <- factor(data$HOME_VAL_bin)
data$HOME_VAL_bin <- factor(data$HOME_VAL_bin, levels=c("NA","Zero","Low","Medium","High"))
```

```
#OLDCLAIM
data$OLDCLAIM_bin[is.na(data$OLDCLAIM)] <- "NA"
data$OLDCLAIM_bin[data$OLDCLAIM == 0] <- "Zero"
data$OLDCLAIM_bin[data$OLDCLAIM >= 1 & data$OLDCLAIM < 1000] <- "Low"
data$OLDCLAIM_bin[data$OLDCLAIM >= 1000 & data$OLDCLAIM < 4500] <- "Medium"
data$OLDCLAIM_bin[data$OLDCLAIM >= 4500] <- "High"
data$OLDCLAIM_bin <- factor(data$OLDCLAIM_bin)
data$OLDCLAIM_bin <- factor(data$OLDCLAIM_bin, levels=c("NA","Zero","Low","Medium","High"))

summary(data)

#Correlation Matrix
subdatnum2 <- subset(data, select = c(TARGET_AMT, KIDSDRIV, AGE, HOMEKIDS, YOJ, INCOME, HOME_VAL,
TRAVTIME, BLUEBOOK, TIF, OLDCLAIM, CLM_FREQ, MVR_PTS, CAR_AGE), na.rm = TRUE)

par(mfrow=c(1,1))
require(corrplot)
mcor <- cor(subdatnum2)
corrplot(mcor, method="number", shade.col=NA, tl.col="black",tl.cex=0.8)
par(mfrow=c(1,1))


############# Part 3: Model Development ######################

# Full Model

Model1 = glm(TARGET_FLAG ~ KIDSDRIV  + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 + HOME_VAL +
MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + RED_CAR +
OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY + DO_KIDS_DRIVE +
HAVE_HOME_KIDS + EMPLOYED + HOME_OWNER  + SUBMITTED_CLAIM + HAVE_MVR_PTS + INCOME_bin +
HOME_VAL_bin + OLDCLAIM_bin, data=data, family = binomial(link="logit"))

summary(Model1)
library(car)
Anova(Model1, type="II", test="Wald")
library(rcompanion)
nagelkerke(Model1)
data$Model1Prediction <- predict(Model1, type = "response")

# Model 2 - Stepwise

model.lower = glm(TARGET_FLAG ~ 1, data=data, family = binomial(link="logit"))

Model1 = glm(TARGET_FLAG ~ KIDSDRIV  + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 + HOME_VAL +
MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + RED_CAR +
OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY + DO_KIDS_DRIVE +
HAVE_HOME_KIDS + EMPLOYED + HOME_OWNER  + SUBMITTED_CLAIM + HAVE_MVR_PTS + INCOME_bin +
HOME_VAL_bin + OLDCLAIM_bin, data=data, family = binomial(link="logit"))
```

```
step(model.lower, scope = list(upper=Model1), direction="both", test="Chisq", data=data)

Model2 = glm(formula = TARGET_FLAG ~ URBANICITY + JOB + MVR_PTS + MSTATUS +
   CAR_TYPE + REVOKED + DO_KIDS_DRIVE + INCOME_bin + CAR_USE +
   TRAVTIME + TIF + OLDCLAIM_bin + BLUEBOOK + OLDCLAIM + HAVE_HOME_KIDS +
   EDUCATION + HOME_OWNER + PARENT1 + KIDSDRIV, family = binomial(link = "logit"), data = data)
summary(Model2)
library(car)
Anova(Model2, type="II", test="Wald")
library(rcompanion)
nagelkerke(Model2)
data$Model2Prediction <- predict(Model2, type = "response")
```

 # Model 3- Reduced Model with Transformations

```
Model3 = glm(formula = TARGET_FLAG ~ URBANICITY + JOB + MVR_PTS + MSTATUS +
   CAR_TYPE + REVOKED + INCOME_bin + CAR_USE +
   SQRT_TRAVTIME + SQRT_TIF + OLDCLAIM_bin + LOG_BLUEBOOK + OLDCLAIM + HAVE_HOME_KIDS +
   EDUCATION + HOME_OWNER + KIDSDRIV, family = binomial(link = "logit"), data = data)

summary(Model3)
library(car)
Anova(Model3, type="II", test="Wald")
library(rcompanion)
nagelkerke(Model3)
data$Model3Prediction <- predict(Model3, type = "response")
```

#Probit Link Model

```
Model4 = glm(formula = TARGET_FLAG ~ URBANICITY + JOB + MVR_PTS + MSTATUS +
   CAR_TYPE + REVOKED + INCOME_bin + CAR_USE +
   SQRT_TRAVTIME + SQRT_TIF + OLDCLAIM_bin + LOG_BLUEBOOK + OLDCLAIM + HAVE_HOME_KIDS +
   EDUCATION + HOME_OWNER + KIDSDRIV, family = binomial(link = "probit"), data = data)

summary(Model4)
library(car)
Anova(Model4, type="II", test="Wald")
library(rcompanion)
nagelkerke(Model4)
data$Model4Prediction <- predict(Model4, type = "response")
```

#Part 4: Performance

```
AIC(Model1)
AIC(Model2)
AIC(Model3)
AIC(Model4)

BIC(Model1)
```

```
BIC(Model2)
BIC(Model3)
BIC(Model4)

print(-2*logLik(Model1, REML = TRUE))
print(-2*logLik(Model2, REML = TRUE))
print(-2*logLik(Model3, REML = TRUE))
print(-2*logLik(Model4, REML = TRUE))

ks_stat(actuals= data$TARGET_FLAG, predictedScores=data$Model1Prediction)
ks_stat(actuals= data$TARGET_FLAG, predictedScores=data$Model2Prediction)
ks_stat(actuals= data$TARGET_FLAG, predictedScores=data$Model3Prediction)
ks_stat(actuals= data$TARGET_FLAG, predictedScores=data$Model4Prediction)

library(Deducer)
rocplot(Model1)
rocplot(Model2)
rocplot(Model3)
rocplot(Model4)

coef(Model3)
```

#Part 5: Stand Alone Scoring Program

```
setwd("~/R/Data")
mytest <- read.csv("logit_insurance_test.csv")
```

#Test Data

```
mytest$INDEX <- as.numeric(mytest$INDEX)
mytest$TARGET_FLAG <- as.factor(mytest$TARGET_FLAG)
mytest$SEX <- as.factor(mytest$SEX)
mytest$EDUCATION <- as.factor(mytest$EDUCATION)
mytest$PARENT1 <- as.factor(mytest$PARENT1)
mytest$INCOME <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mytest$INCOME)))

mytest$HOME_VAL <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mytest$HOME_VAL)))
mytest$MSTATUS <- as.factor(mytest$MSTATUS)
mytest$REVOKED <- as.factor(mytest$REVOKED)
mytest$RED_CAR <- as.factor(ifelse(mytest$RED_CAR=="yes", 1, 0))
mytest$URBANICITY <- ifelse(mytest$URBANICITY == "Highly Urban/ Urban", "Urban", "Rural")
mytest$URBANICITY <- as.factor(mytest$URBANICITY)
mytest$JOB <- as.factor(mytest$JOB)
mytest$CAR_USE <- as.factor(mytest$CAR_USE)
mytest$CAR_TYPE <- as.factor(mytest$CAR_TYPE)
mytest$DO_KIDS_DRIVE <- as.factor(ifelse(mytest$KIDSDRIV > 0, 1, 0 ))
mytest$OLDCLAIM <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mytest$OLDCLAIM)))
mytest$BLUEBOOK <- suppressWarnings(as.numeric(gsub("[^0-9.]", "", mytest$BLUEBOOK)))
```

summary(mytest)

 # Fix NA's for Test

library(mice)

#Check for missing values
sapply(mytest, function(x) sum(is.na(x)))

#Check missing data percentage
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(mytest,2,pMiss)

library(VIM)
aggr_plot <- aggr(mytest, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(mytest), cex.axis=.5, gap=2, ylab=c("Histogram of missing data","Pattern"))

#Split datasets into numerical and categorical

#Numeric
subdatnumtest <- subset(mytest, select=c(
"INDEX",
"KIDSDRIV",
"AGE",
"HOMEKIDS",
"YOJ",
"INCOME",
"HOME_VAL",
"TRAVTIME",
"BLUEBOOK",
"TIF",
"OLDCLAIM",
"CLM_FREQ",
"MVR_PTS",
"CAR_AGE",
"TARGET_AMT"))

subdatnumtest.df <- data.frame(subdatnumtest)

#Categorical
subdatcattest <- subset(mytest, select=c(
"INDEX",
"TARGET_FLAG",
"PARENT1",
"MSTATUS",
"SEX",
"EDUCATION",

```
"JOB",
"CAR_USE",
"CAR_TYPE",
"RED_CAR",
"REVOKED",
"URBANICITY",
"DO_KIDS_DRIVE"))

subdatcattest.df <- data.frame(subdatcattest)


# Fix NA's for Test Data

#Run imputation
tempDatatest <- mice(subdatnumtest.df,m=5,maxit=50,meth='pmm',seed=500)

summary(tempDatatest)

# Inspecting the distribution of original and imputed data for the variables that contained N/A
xyplot(tempDatatest,TARGET_AMT~ CAR_AGE + HOME_VAL + YOJ +  INCOME + AGE,pch=18,cex=1)

densityplot(tempDatatest)

#Check N/A values have been removed
subdatnumimptest <- complete(tempDatatest,1)
apply(subdatnumimptest,2,pMiss)
summary(subdatnumimptest)
sapply(subdatnumimptest, function(x) sum(is.na(x)))

#Merge Numeric and Categorical datasets back
test <- merge(subdatnumimptest, subdatcattest.df, by=c("INDEX"))

#Check data
str (test)
summary (test)

#Trim Test
test$CAR_AGE[test$CAR_AGE < 0 ] <- 0
test$YOJ [(test$YOJ >= 20)] = 20
test$INCOME [(test$INCOME >= 300000)] = 300000
test$HOME_VAL  [(test$HOME_VAL  >= 650000)] = 650000
test$TRAVTIME [(test$TRAVTIME >= 100)] = 100
test$BLUEBOOK [(test$BLUEBOOK >= 55000)] = 55000
test$TIF [(test$TIF >= 17)] = 17
test$MVR_PTS [(test$MVR_PTS >=8)] = 8

#Create Flag Variables
test$HAVE_HOME_KIDS <- as.factor(ifelse(test$HOMEKIDS > 0, 1, 0 ))
```

```
test$EMPLOYED <- as.factor(ifelse(test$YOJ > 0, 1, 0 ))
test$HOME_OWNER <- as.factor(ifelse(test$HOME_VAL> 0, 1, 0 ))
test$SUBMITTED_CLAIM <- as.factor(ifelse(test$CLM_FREQ  > 0, 1, 0 ))
test$HAVE_MVR_PTS<- as.factor(ifelse(test$MVR_PTS> 0, 1, 0 ))

#Create SQRT Transformations of Some of the Variables
test$SQRT_TRAVTIME <- sqrt(test$TRAVTIME)
test$SQRT_BLUEBOOK <- sqrt(test$BLUEBOOK)
test$SQRT_TIF <- sqrt(test$TIF)
test$LOG_TRAVTIME <- log(test$TRAVTIME)
test$LOG_BLUEBOOK <- log(test$BLUEBOOK)
test$LOG_TIF <- log(test$TIF)
test$LOG_MVR_PTS <- log(test$MVR_PTS)
test$LOG_OLDCLAIM <- log(test$OLDCLAIM)

# Bins for Test Data

#Income
test$INCOME_bin[is.na(test$INCOME)] <- "NA"
test$INCOME_bin[test$INCOME == 0] <- "Zero"
test$INCOME_bin[test$INCOME >= 1 & test$INCOME < 30000] <- "Low"
test$INCOME_bin[test$INCOME >= 30000 & test$INCOME < 80000] <- "Medium"
test$INCOME_bin[test$INCOME >= 80000] <- "High"
test$INCOME_bin <- factor(test$INCOME_bin)
test$INCOME_bin <- factor(test$INCOME_bin, levels=c("NA","Zero","Low","Medium","High"))

#HOME_VAL
test$HOME_VAL_bin[is.na(test$HOME_VAL)] <- "NA"
test$HOME_VAL_bin[test$HOME_VAL == 0] <- "Zero"
test$HOME_VAL_bin[test$HOME_VAL >= 1 & test$HOME_VAL < 125000] <- "Low"
test$HOME_VAL_bin[test$HOME_VAL >= 125000 & test$HOME_VAL < 300000] <- "Medium"
test$HOME_VAL_bin[test$HOME_VAL >= 300000] <- "High"
test$HOME_VAL_bin <- factor(test$HOME_VAL_bin)
test$HOME_VAL_bin <- factor(test$HOME_VAL_bin, levels=c("NA","Zero","Low","Medium","High"))

#OLDCLAIM
test$OLDCLAIM_bin[is.na(test$OLDCLAIM)] <- "NA"
test$OLDCLAIM_bin[test$OLDCLAIM == 0] <- "Zero"
test$OLDCLAIM_bin[test$OLDCLAIM >= 1 & test$OLDCLAIM < 1000] <- "Low"
test$OLDCLAIM_bin[test$OLDCLAIM >= 1000 & test$OLDCLAIM < 4500] <- "Medium"
test$OLDCLAIM_bin[test$OLDCLAIM >= 4500] <- "High"
test$OLDCLAIM_bin <- factor(test$OLDCLAIM_bin)
test$OLDCLAIM_bin <- factor(test$OLDCLAIM_bin, levels=c("NA","Zero","Low","Medium","High"))


summary(test)
```

#Stand Alone Scoring Program

```
data0<- subset(data, TARGET_FLAG == 1 )

test$P_TARGET_FLAG <- predict(Model3, newdata = test, type = "response")

targetbycar <- aggregate(data0$TARGET_AMT, list(data0$CAR_TYPE), mean)

test$P_TARGET_AMT <- ifelse(test$CAR_TYPE=="Minivan", 5601.665%*%.27,
                ifelse(test$CAR_TYPE=="Panel Truck", 7464.703%*%.27,
                    ifelse(test$CAR_TYPE=="Pickup", 5430.106%*%.27,
                        ifelse(test$CAR_TYPE=="Sports Car", 5412.733%*%.27,
                            ifelse(test$CAR_TYPE=="Van", 6908.553%*%0.27, 5241.104%*%.27)))))
```

 # Scored Data File

```
#subset of data set for the deliverable "Scored data file"
scores <- test[c("INDEX","P_TARGET_FLAG", "P_TARGET_AMT")]

#Note, this next function will output a csv file in your work environment called write.csv.

write.csv(scores, file = "CI_Scored.csv")
write.csv(as.data.frame(scores), file = "logit_insurance_test.csv",
      sheetName = "Scored Data File", row.names = FALSE)
```

Appendix

Data Quality Check (Figure 3)

```
> describe(data)
data

 27  Variables      8161  Observations
---------------------------------------------------------------------------------------------
-------------------------------------------
INDEX
       n  missing distinct
    8161        0     8161

lowest : 1     2     4     5     6    , highest: 10297 10298 10299 10301 10302
---------------------------------------------------------------------------------------------
-------------------------------------------
TARGET_FLAG
       n  missing distinct
    8161        0        2

Value          0     1
Frequency   6008  2153
Proportion 0.736 0.264
---------------------------------------------------------------------------------------------
-------------------------------------------
TARGET_AMT
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75      .90      .95
    8161        0     1949    0.601     1504     2574        0        0        0        0
1036     4904     6452

lowest :      0.00000     30.27728     58.53106     95.56732    108.74150, highest:  73783.
46592  77907.43028  78874.19056  85523.65335 107586.13616
---------------------------------------------------------------------------------------------
-------------------------------------------
KIDSDRIV
       n  missing distinct     Info     Mean      Gmd
    8161        0        5    0.318   0.1711   0.3095

Value          0     1     2     3     4
Frequency   7180   636   279    62     4
Proportion 0.880 0.078 0.034 0.008 0.000
---------------------------------------------------------------------------------------------
-------------------------------------------
AGE
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75      .90      .95
    8155        6       60    0.999    44.79    9.747       30       34       39       45
51       56       59

lowest : 16 17 18 19 20, highest: 72 73 76 80 81
---------------------------------------------------------------------------------------------
-------------------------------------------
HOMEKIDS
       n  missing distinct     Info     Mean      Gmd
    8161        0        6    0.723   0.7212    1.056
```

```
Value          0     1     2     3     4     5
Frequency   5289   902  1118   674   164    14
Proportion 0.648 0.111 0.137 0.083 0.020 0.002
---------------------------------------------------------------------------------------------
-------------------------------------------
YOJ
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75      .90      .95
    7707      454       21    0.989     10.5     4.29        0        5        9       11
13       15       15

lowest :  0  1  2  3  4, highest: 16 17 18 19 23
---------------------------------------------------------------------------------------------
-------------------------------------------
INCOME
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75      .90      .95
    7716      445     6612    0.999    61898    51302        0     4380    28097    54028
85986   123180   152274

lowest :      0      5      7     18     70, highest: 306277 309628 320127 332339 367030
---------------------------------------------------------------------------------------------
-------------------------------------------
PARENT1
       n  missing distinct
    8161        0        2

Value         No   Yes
Frequency   7084  1077
Proportion 0.868 0.132
---------------------------------------------------------------------------------------------
-------------------------------------------
HOME_VAL
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75      .90      .95
    7697      464     5106    0.974   154867   143664        0        0        0   161160
238724   316543   374871

lowest :      0  50223  50343  50964  51038, highest: 657804 682634 738153 750455 885282
---------------------------------------------------------------------------------------------
-------------------------------------------
MSTATUS
       n  missing distinct
    8161        0        2

Value       Yes z_No
Frequency  4894 3267
Proportion  0.6  0.4
---------------------------------------------------------------------------------------------
-------------------------------------------
SEX
       n  missing distinct
    8161        0        2

Value          M   z_F
Frequency   3786  4375
```

```
Proportion 0.464 0.536
--------------------------------------------------------------------------------------------
-------------------------------------------
EDUCATION 
       n  missing distinct 
    8161        0        5 
                                                                                
Value       <High School     Bachelors       Masters           PhD z_High School
Frequency           1203          2242          1658           728          2330
Proportion         0.147         0.275         0.203         0.089         0.286
--------------------------------------------------------------------------------------------
-------------------------------------------
JOB 
       n  missing distinct 
    8161        0        9 

Value                            Clerical        Doctor    Home Maker        Lawyer       Mana
ger  Professional       Student
Frequency            526          1271           246           641           835           
988          1117           712
Proportion         0.064         0.156         0.030         0.079         0.102          0.
121         0.137         0.087
                          
Value      z_Blue Collar
Frequency           1825
Proportion         0.224
--------------------------------------------------------------------------------------------
-------------------------------------------
TRAVTIME 
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50 
.75      .90      .95 
    8161        0       97        1    33.49    17.85        7       13       22       33 
44       54       60 

lowest :   5   6   7   8   9, highest: 103 113 124 134 142
--------------------------------------------------------------------------------------------
-------------------------------------------
CAR_USE 
       n  missing distinct 
    8161        0        2 
                                
Value      Commercial    Private
Frequency        3029       5132
Proportion      0.371      0.629
--------------------------------------------------------------------------------------------
-------------------------------------------
BLUEBOOK 
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50 
.75      .90      .95 
    8161        0     2789        1    15710     9354     4900     6000     9280    14440 
20850    27460    31110 

lowest :  1500  1520  1530  1540  1590, highest: 57970 61050 62240 65970 69740
--------------------------------------------------------------------------------------------
-------------------------------------------
TIF 
```

```
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50 
.75      .90      .95 
    8161        0       23    0.961    5.351    4.512        1        1        1        4 
7       11       13 

lowest :  1  2  3  4  5, highest: 19 20 21 22 25
---------------------------------------------------------------------------------------------
-------------------------------------------
CAR_TYPE 
       n  missing distinct 
    8161        0        6 

Value          Minivan Panel Truck      Pickup  Sports Car         Van       z_SUV
Frequency         2145         676        1389         907         750        2294
Proportion       0.263       0.083       0.170       0.111       0.092       0.281
---------------------------------------------------------------------------------------------
-------------------------------------------
RED_CAR 
       n  missing distinct 
    8161        0        2 

Value          0     1
Frequency   5783  2378
Proportion 0.709 0.291
---------------------------------------------------------------------------------------------
-------------------------------------------
OLDCLAIM 
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50 
.75      .90      .95 
    8161        0     2857    0.769     4037     6563        0        0        0        0 
4636     9583    27090 

lowest :     0   502   506   518   519, highest: 52507 53477 53568 53986 57037
---------------------------------------------------------------------------------------------
-------------------------------------------
CLM_FREQ 
       n  missing distinct     Info     Mean      Gmd 
    8161        0        6    0.763   0.7986    1.129 

Value          0     1     2     3     4     5
Frequency   5009   997  1171   776   190    18
Proportion 0.614 0.122 0.143 0.095 0.023 0.002
---------------------------------------------------------------------------------------------
-------------------------------------------
REVOKED 
       n  missing distinct 
    8161        0        2 

Value         No   Yes
Frequency   7161  1000
Proportion 0.877 0.123
---------------------------------------------------------------------------------------------
-------------------------------------------
MVR_PTS 
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50 
.75      .90      .95 
```

```
    8161        0       13      0.9    1.696    2.187        0        0        0        1
3        5        6

Value          0     1     2     3     4     5     6     7     8     9    10    11    13
Frequency   3712  1157   948   758   599   399   266   167    84    45    13    11     2
Proportion 0.455 0.142 0.116 0.093 0.073 0.049 0.033 0.020 0.010 0.006 0.002 0.001 0.000
---------------------------------------------------------------------------------------------
-------------------------------------------
CAR_AGE
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
.75      .90      .95
    7651      510       30    0.982    8.328    6.459        1        1        1        8
12       16       18

lowest : -3  0  1  2  3, highest: 24 25 26 27 28
---------------------------------------------------------------------------------------------
-------------------------------------------
URBANICITY
       n  missing distinct
    8161        0        2

Value      Rural Urban
Frequency   1669  6492
Proportion 0.205 0.795
---------------------------------------------------------------------------------------------
-------------------------------------------
DO_KIDS_DRIVE
       n  missing distinct
    8161        0        2

Value         0    1
Frequency  7180  981
Proportion 0.88 0.12
---------------------------------------------------------------------------------------------
-------------------------------------------
```