Insurance Project Report

By Mimi Trinh

INTRODUCTION

The insurance project has two main goals 1) predict the probability of a customer crashing their car 2) if somebody crashes their car, predict the cost the insurance company has to pay. The project uses the logit_insurance.csv file as train dataset to build two predictive models (logistic regression model to address the first question and ordinary least square (OLS) regression to address the second question) and logit_insurance_test.csv file as test dataset to apply the predictive models to new data.

The train dataset has 8161 observations with 26 variables. Each record represents a customer at an auto insurance company. Each record has two target variables: TARGET_FLAG and TARGET_AMT. First, TARGET_FLAG indicates whether a customer has a crash with 1 means that the person was in a car crash, and 0 means that the person was not in a car crash. Second, TARGET_AMT shows what it costs the insurance company if the customer has a crash. This value is 0 if the person didn't crash their car. But if they did, this number will be greater than 0 in the record.

RESULTS

Section 1: Data Exploration

```
> str(data)
'data.frame':   8161 obs. of  26 variables:
 $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
 $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
 $ 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     : Factor w/ 6613 levels "","$0","$1,007",..: 5033 6292 1250 1 509 746
1488 315 4765 282 ...
 $ PARENT1    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
 $ HOME_VAL   : Factor w/ 5107 levels "","$0","$100,093",..: 2 3259 348 3917 3034 2
1 4167 2 2 ...
 $ 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   : Factor w/ 2789 levels "$1,500","$1,520",..: 434 503 2212 553 802 74
6 2672 701 135 852 ...
 $ 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 "no","yes": 2 2 1 2 1 1 1 2 1 1 ...
 $ OLDCLAIM   : Factor w/ 2857 levels "$0","$1,000",..: 1449 1 1311 1 432 1 1 510 1
1 ...
 $ 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 "Highly Urban/ Urban",..: 1 1 1 1 1 1 1 1 1 2 ..
```

```
> summary(data) #need to fix the following variables to the correct format in R
     INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
 Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
 1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
 Median : 5133   Median :0.0000   Median :     0   Median :0.0000
 Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
 3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
 Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000

      AGE           HOMEKIDS           YOJ             INCOME     PARENT1
 Min.   :16.00   Min.   :0.0000   Min.   : 0.0   $0      : 615   No :7084
 1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0           : 445   Yes:1077
 Median :45.00   Median :0.0000   Median :11.0   $26,840 :   4
 Mean   :44.79   Mean   :0.7212   Mean   :10.5   $48,509 :   4
 3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   $61,790 :   4
 Max.   :81.00   Max.   :5.0000   Max.   :23.0   $107,375:   3
 NA's   :6                        NA's   :454    (Other) :7086
    HOME_VAL    MSTATUS      SEX             EDUCATION
 $0      :2294   Yes :4894   M  :3786   <High School :1203
         : 464   z_No:3267   z_F:4375   Bachelors    :2242
 $111,129:   3                          Masters      :1658
 $115,249:   3                          PhD          : 728
 $123,109:   3                          z_High School:2330
 $153,061:   3
 (Other) :5391

          JOB          TRAVTIME          CAR_USE        BLUEBOOK
 z_Blue Collar:1825   Min.   : 5.00   Commercial:3029   $1,500 : 157
 Clerical     :1271   1st Qu.: 22.00  Private   :5132   $6,000 :  34
 Professional :1117   Median : 33.00                    $5,800 :  33
 Manager      : 988   Mean   : 33.49                    $6,200 :  33
 Lawyer       : 835   3rd Qu.: 44.00                    $6,400 :  31
 Student      : 712   Max.   :142.00                    $5,900 :  30
 (Other)      :1413                                     (Other):7843
      TIF              CAR_TYPE    RED_CAR       OLDCLAIM       CLM_FREQ
 Min.   : 1.000   Minivan    :2145   no :5783   $0      :5009   Min.   :0.0000
 1st Qu.: 1.000   Panel Truck: 676   yes:2378   $1,310 :   4   1st Qu.:0.0000
 Median : 4.000   Pickup     :1389              $1,391 :   4   Median :0.0000
 Mean   : 5.351   Sports Car : 907              $4,263 :   4   Mean   :0.7986
 3rd Qu.: 7.000   Van        : 750              $1,105 :   3   3rd Qu.:2.0000
 Max.   :25.000   z_SUV      :2294              $1,332 :   3   Max.   :5.0000
                                                (Other):3134
 REVOKED       MVR_PTS          CAR_AGE                  URBANICITY
 No :7161   Min.   : 0.000   Min.   :-3.000   Highly Urban/ Urban  :6492
 Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000   z_Highly Rural/ Rural:1669
            Median : 1.000   Median : 8.000
            Mean   : 1.696   Mean   : 8.328
            3rd Qu.: 3.000   3rd Qu.:12.000
            Max.   :13.000   Max.   :28.000
                             NA's   :510
```

The summary output above shows that there are 8161 observations and 26 variables along with the format, mean, median, and quantile of each variable in the train dataset.

1. INDEX: identification variable, not useful in predictive models
2. TARGET_FLAG: was car in a crash? This is the response variable in the first question using logistic regression
3. TARGET_AMT: if car was in a cash, what was the cost? This is the response variable in the second question using OLS regression
4. KIDSDRIV: number of driving children
5. AGE: age of driver
6. HOMEKIDS: number of children at home
7. YOJ: years on job
8. INCOME: income
9. PARENT1: single parent
10. HOME_VAL: home value
11. MSTATUS: marital status
12. SEX: gender
13. EDUCATION:
14. JOB: job category
15. TRAVTIME: distance to work
16. CAR_USE: vehicle use
17. BLUEBOOK: value of vehicle
18. TIF: time in force
19. CAR_TYPE: type of car
20. RED_CAR: a red car
21. OLDCLAIM: total claims in the past 5 years
22. CLM_FREQ: number of claims in the past 5 years
23. REVOKED: license revoked in the past 7 years
24. MVR_PTS: motor vehicle record points
25. CAR_AGE: vehicle age
26. URBANICITY: home/work area

From the summary output above, the following variables have missing value issues that need to be fixed in section 2 of the report.

- AGE
- YOJ
- INCOME
- HOME_VAL
- JOB
- CAR_AGE

Also, the following variables have inappropriate format that needs to be fixed prior to the data analysis process.

- TARGET_FLAG: convert two-level numeric to two-level nominal variable

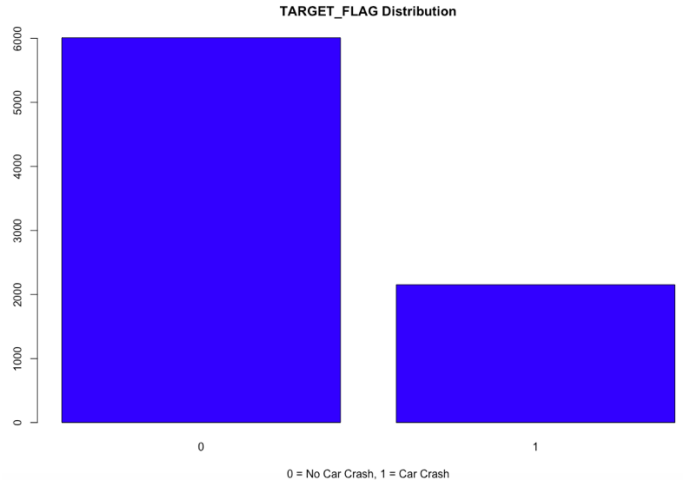
- KIDSDRIV: convert five-level numeric to two-level nominal variable
- HOMEKIDS: convert six-level numeric to two-level nominal variable
- INCOME: convert nominal to numeric variable
- HOME_VAL: convert nominal to numeric variable
- BLUEBOOK: convert nominal to numeric variable
- OLDCLAIM: convert nominal to numeric variable
- CAR_AGE: can't have negative car age (probably due to data collection error), so change this value to NA missing value

After the conversion of the variables above, we have the following output.

```
> str(data) #good format now
'data.frame':   8161 obs. of  26 variables:
 $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
 $ 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   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
 $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
 $ HOMEKIDS   : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 1 ...
 $ 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 "no","yes": 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 "Highly Urban/ Urban",..: 1 1 1 1 1 1 1 1 1 2 ..
```

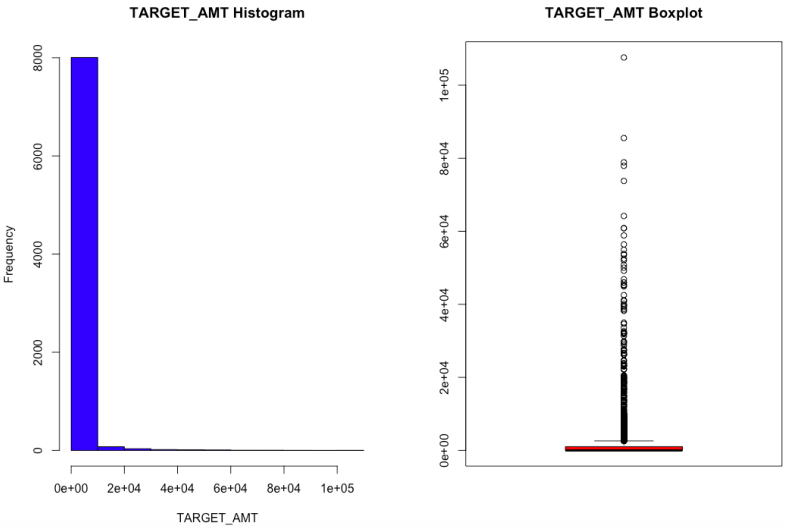
```
> summary(data)
     INDEX       TARGET_FLAG    TARGET_AMT        KIDSDRIV          AGE          HOMEKID          CAR_USE        BLUEBOOK          TIF                CAR_TYPE
S
 Min.   :    1   0:6008      Min.   :     0   0:7180   Min.   :16.00   0:5289   Commercial:3029   Min.   : 1500   Min.   : 1.000   Minivan    :2145
 1st Qu.: 2559   1:2153      1st Qu.:     0   1: 981   1st Qu.:39.00   1:2872   Private   :5132   1st Qu.: 9280   1st Qu.: 1.000   Panel Truck: 676
 Median : 5133               Median :     0            Median :45.00                             Median :14440   Median : 4.000   Pickup     :1389
 Mean   : 5152               Mean   :  1504            Mean   :44.79                             Mean   :15710   Mean   : 5.351   Sports Car : 907
 3rd Qu.: 7745               3rd Qu.:  1036            3rd Qu.:51.00                             3rd Qu.:20850   3rd Qu.: 7.000   Van        : 750
 Max.   :10302               Max.   :107586            Max.   :81.00                             Max.   :69740   Max.   :25.000   z_SUV      :2294
                                                       NA's   :6

     YOJ           INCOME       PARENT1       HOME_VAL      MSTATUS       RED_CAR       OLDCLAIM        CLM_FREQ       REVOKED       MVR_PTS
 Min.   : 0.0   Min.   :     0   No :7084   Min.   :     0   Yes :4894   no :5783   Min.   :    0   Min.   :0.0000   No :7161   Min.   : 0.000
 1st Qu.: 9.0   1st Qu.: 28097   Yes:1077   1st Qu.:     0   z_No:3267   yes:2378   1st Qu.:    0   1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000
 Median :11.0   Median : 54028              Median :161160                          Median :    0   Median :0.0000              Median : 1.000
 Mean   :10.5   Mean   : 61898              Mean   :154867                          Mean   : 4037   Mean   :0.7986              Mean   : 1.696
 3rd Qu.:13.0   3rd Qu.: 85986              3rd Qu.:238724                          3rd Qu.: 4636   3rd Qu.:2.0000              3rd Qu.: 3.000
 Max.   :23.0   Max.   :367030              Max.   :885282                          Max.   :57037   Max.   :5.0000              Max.   :13.000
 NA's   :454    NA's   :445                 NA's   :464

  SEX            EDUCATION             JOB          TRAVTIME         CAR_AGE                    URBANICITY
 M  :3786   <High School :1203   z_Blue Collar:1825   Min.   :  5.00   Min.   : 0.00   Highly Urban/ Urban  :6492
 z_F:4375   Bachelors    :2242   Clerical     :1271   1st Qu.: 22.00   1st Qu.: 1.00   z_Highly Rural/ Rural:1669
            Masters      :1658   Professional :1117   Median : 33.00   Median : 8.00
            PhD          : 728   Manager      : 988   Mean   : 33.49   Mean   : 8.33
            z_High School:2330   Lawyer       : 835   3rd Qu.: 44.00   3rd Qu.:12.00
                                 Student      : 712   Max.   :142.00   Max.   :28.00
                                 (Other)      :1413                    NA's   :511
```

Based on the summary output above, we have the correct format for all of the variables in the dataset, so we can continue with the data exploration process in this section.




```
> summary(data$TARGET_FLAG)
   0    1
6008 2153
```

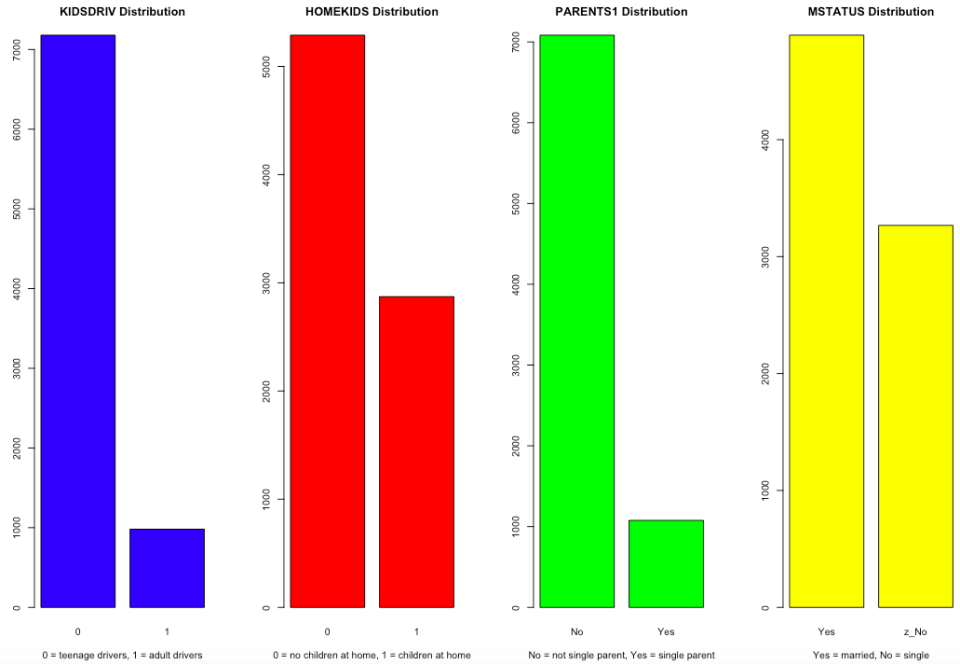
Regarding the first response variable TARGET_FLAG, 74% of customers don't have car crashes as shown in the output above, which makes sense since the majority of people shouldn't have fatal crashes. Since 26% of customers have car crash, when we build the logistic regression model, the test dataset should also yield similar result of 26% average.




```
summary(data$TARGET_AMT)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0       0       0    1504    1036  107586
```

Regarding the second response variable TARGET_AMT, it's significantly skewed toward the right, as shown by the right tail in both the histogram and boxplot above. As revealed by the

TARGET_FLAG variable above, majority of records don't have car crash. Therefore, majority of the observations in TARGET_AMT should be 0, which explains why this variable is skewed toward the right. Since the average of cost in the data is $1504, when we build the OLS regression model, the test dataset should yield similar result of 1504 average.

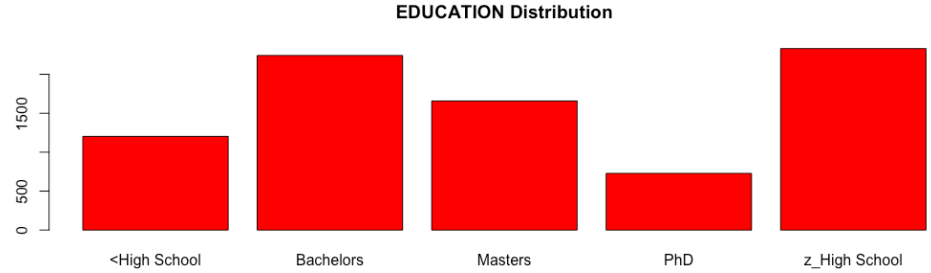



```
> summary(data$KIDSDRIV)
   0    1
7180  981
> summary(data$HOMEKIDS)
   0    1
5289 2872
> summary(data$PARENT1)
  No  Yes
7084 1077
> summary(data$MSTATUS)
 Yes z_No
4894 3267
```

From the output above, we can draw the following conclusions regarding each variable.

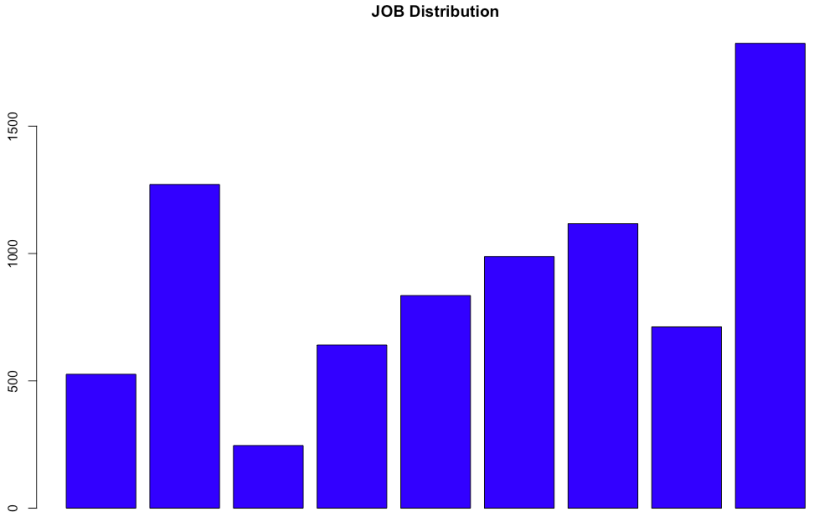
- KIDSDRIV: only 12% customers are teenage drivers
- HOMEKIDS: only 25% customers have children at home
- PARENT1: only 13% customers are single parents
- MSTATUS: 40% customers are single

The breakdown above make sense in real life, so there's no issue with data collection here. Although KIDSDRIV and PARENT1 have a significantly uneven proportion with the majority of data falling into one particular category, it's still good practice to include them in the predictive models. After running the regression analysis, we can determine to keep or remove them from the model depending on their usefulness or contribution to the models.



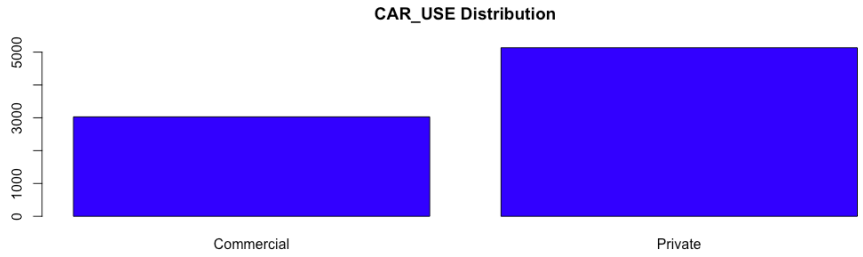

```
> summary(data$SEX)
   M  z_F
3786 4375
> summary(data$EDUCATION)
 <High School     Bachelors       Masters           PhD z_High School
         1203          2242          1658           728          2330
```

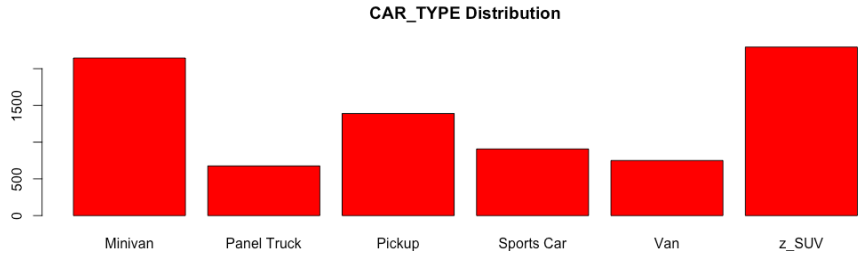
Regarding SEX, 46% drivers are male. Regarding EDUCATION, the data is very well spread out among its five categories with high school and bachelors representing the most drivers at 27% each. Since both variables have equal proportion, there's no issue to include SEX and EDUCATION in the predictive models.




```
summary(data$JOB)
                Clerical           Doctor       Home Maker           Lawyer
             526             1271              246              641              835
         Manager     Professional          Student       z_Blue Collar
             988             1117              712             1825
```

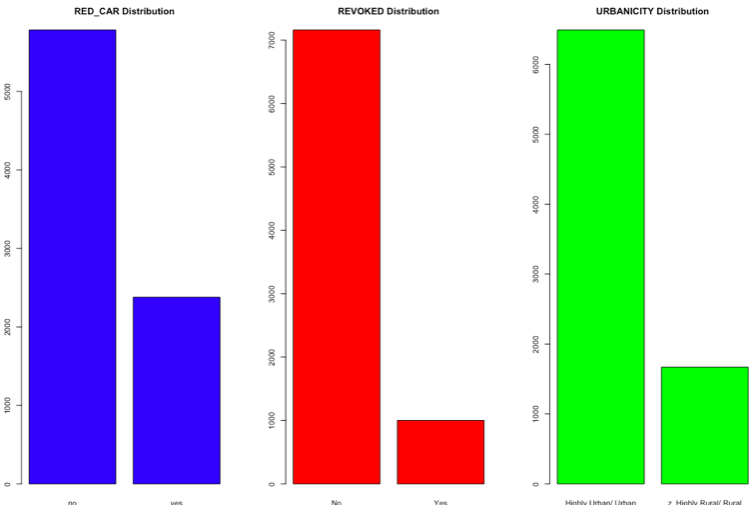
The JOB variable has equal proportion with blue collar accounting for the most drivers at 22%, follow by clerical at 16%, and professional at 14%. Also, as revealed by the output above, this variable has 526 missing values, which confirm the conclusion we draw earlier that JOB has missing value issue. This problem will be addressed in section 2 of the report prior to any data analysis steps.






```
> summary(data$CAR_USE)
Commercial    Private
      3029       5132
> summary(data$CAR_TYPE)
    Minivan Panel Truck      Pickup  Sports Car         Van       z_SUV
       2145         676        1389         907         750        2294
```

Regarding CAR_USE variable, 63% are used for private purpose. Regarding CAR_TYPE, the data is spread out evenly with SUV accounting for the most cars at 28%, follow by minivan at 26%. Since both variables are equally distributed, CAR_USE and CAR_TYPE should be included in the predictive models.



```
> summary(data$RED_CAR)
  no  yes
5783 2378
> summary(data$REVOKED)
  No  Yes
7161 1000
> summary(data$URBANICITY)
  Highly Urban/ Urban z_Highly Rural/ Rural
                6492                  1669
```

From the output above, we can draw the following conclusions.

- RED_CAR: red cars only account for 29% of vehicles in the dataset
- REVOKED: only 12% of customers have their license revoked in the past 7 years
- URBANICITY: only 20% of drivers live/work in the rural area

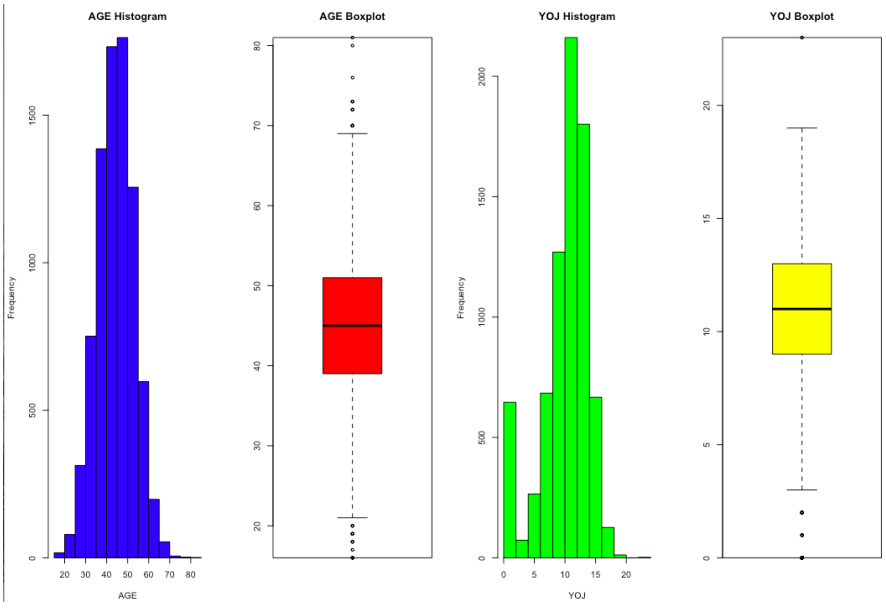
Since all variables have equal proportion, we can include RED_CAR, REVOKED, and URBANICITY in the predictive models.

```
> #x-variables numeric
> skewness(data$AGE,na.exclude(data$AGE)) #no outlier issue
[1] -0.02899428
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(data$YOJ,na.exclude(data$YOJ)) #small outlier issue
[1] -1.203202
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(data$INCOME,na.exclude(data$INCOME)) #small outlier issue
[1] 1.186547
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(data$HOME_VAL,na.exclude(data$HOME_VAL))
[1] NA
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(data$TRAVTIME) #no outlier issue
[1] 0.4468995
> skewness(data$BLUEBOOK) #no outlier issue
[1] 0.7943601
> skewness(data$TIF) #no outlier issue
[1] 0.8909758
> skewness(data$OLDCLAIM) #big outlier issue
[1] 3.119613
> skewness(data$CLM_FREQ) #small outlier issue
[1] 1.209021
> skewness(data$MVR_PTS) #small outlier issue
[1] 1.348088
> skewness(data$CAR_AGE,na.exclude(data$CAR_AGE)) #no outlier issue
[1] 0.2824556
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
```

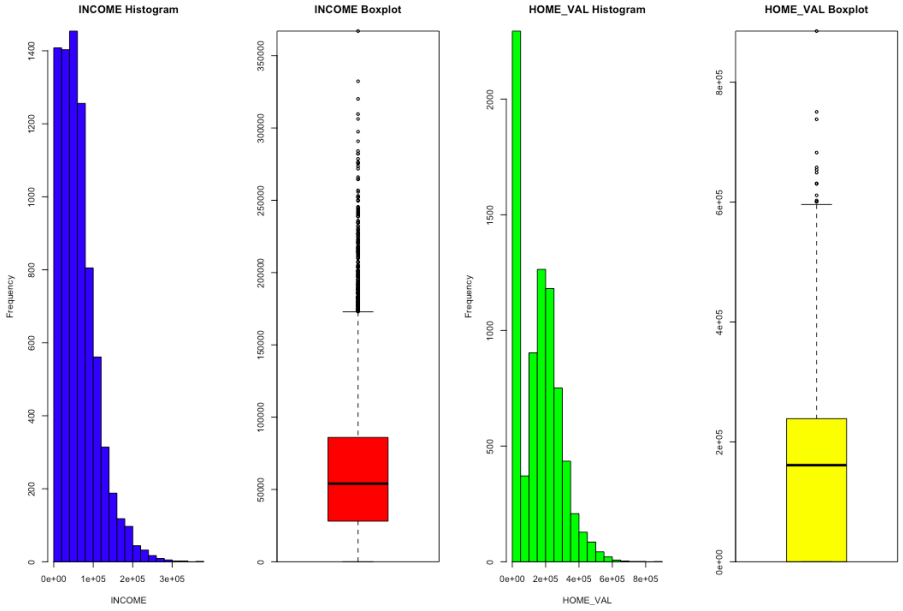
After exploring categorical variables, we will now assess numeric variables in the dataset. In order to check for outliers, we use the skewness number. If a variable has skewness of 0, there's no outlier issue since the variable follows a normal distribution. If the skewness number is above 1, the variable is skewed toward the right. If it's lower than -1, the variable is skewed toward the left. The higher the absolute value of the skewness number, the more outliers exist in the data. Using the output above, the variables with missing values yield a warning message as we calculate skewness number excluding records with missing values.

From using the skewness number as shown above, YOJ has a small outlier issue skewing toward the left with skewness of -1.203202. INCOME, CLM_FREQ, MVR_PTS have small outlier issues skewing toward the right with skewness of 1.186547, 1.209021, 1.348088 whereas OLDCLAIM has a severe outlier issue skewing toward the right with skewness of 3.119613. However, using skewness number alone is not a good way to detect outliers. We need to examine the histograms and boxplots of each variable to determine which predictors have outlier issues and thus need to be fixed before running data analysis. Using this approach of histograms and boxplots, with details in the following section, the variables below are determined to have outlier issues and thus need to be addressed in section 2 of the report before putting into the model.

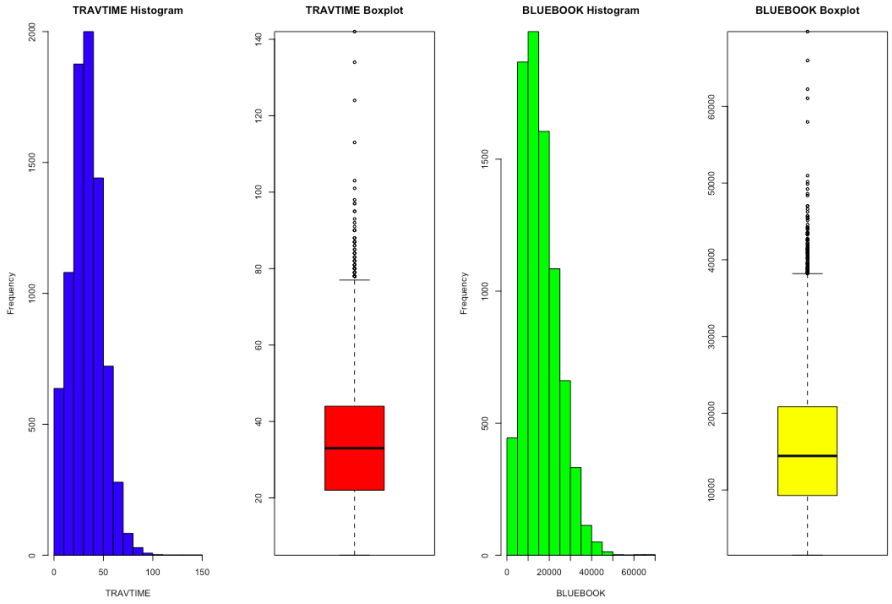
- Small outlier issues
  - YOJ
  - INCOME
  - HOME_VAL
  - TRAVTIME
  - BLUEBOOK
  - TIF
  - CLM_FREQ
  - MVR_PTS
- Big outlier issue
  - OLDCLAIM



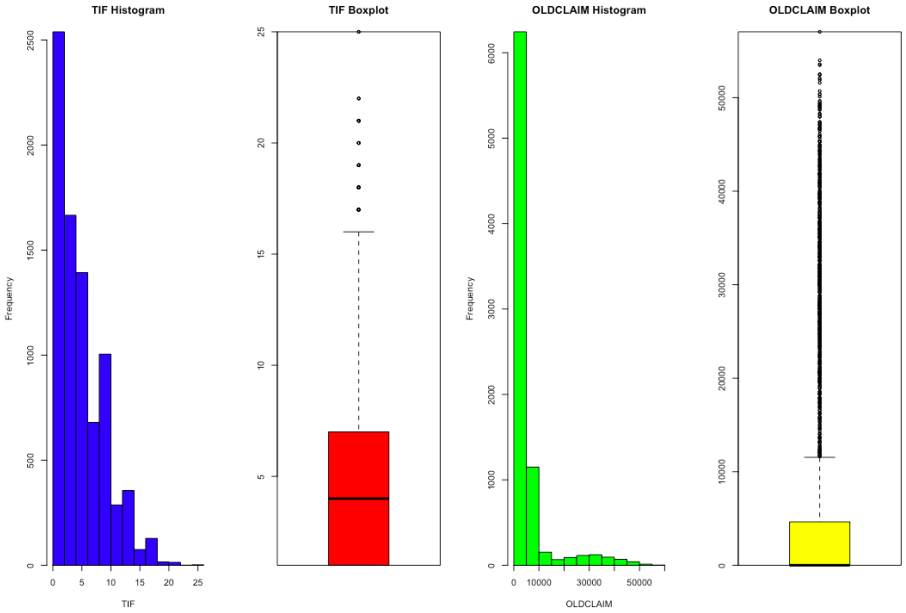
The charts show though AGE has outliers, since there are outliers on both sides, the variable still follows a normal distribution and thus doesn't need to be fixed before running data analysis. On the other hand, YOJ has some outliers on the left with a lot of records with 0 value. Therefore, this variable has a small outlier issue that needs to be fixed.



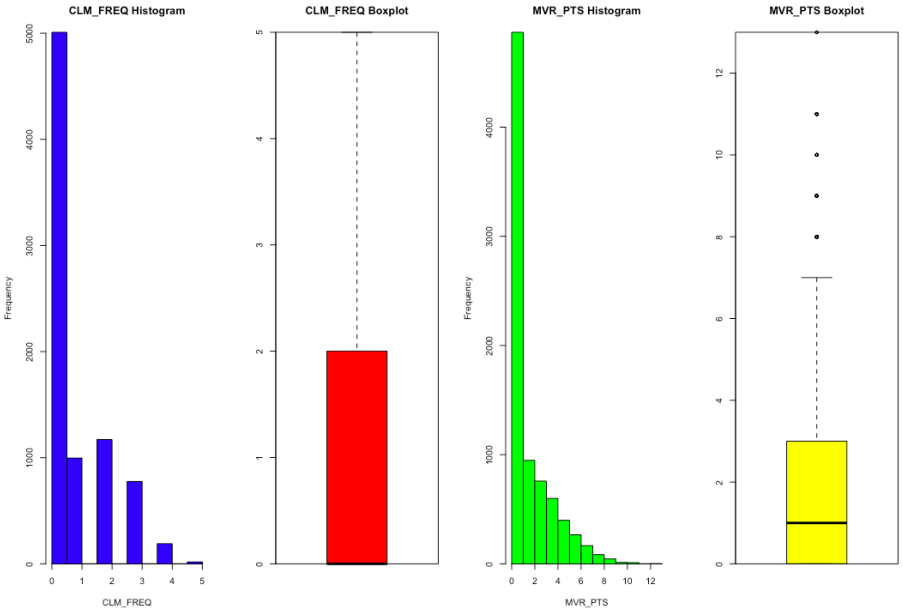
The histograms and boxplots above indicate that both INCOME and HOME_VAL have outlier issues skewing toward the right. However, INCOME has more severe issue than HOME_VAL with more outliers. Thus, both predictors need to be fixed before putting them in the models.



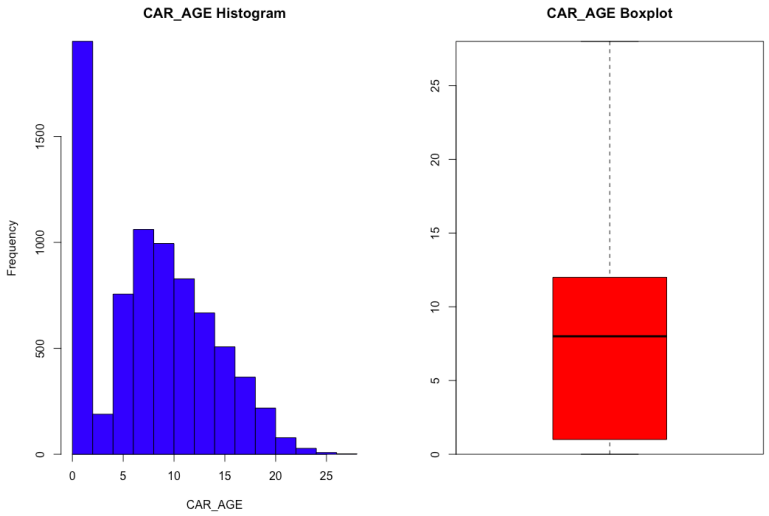
Similarly, as shown by the charts above, TRAVTIME and BLUEBOOK also have outlier issues skewing toward the right. Both need to be addressed before putting them in the models.



The histograms and boxplots show that TIF has a small outlier issue whereas OLDCLAIM has a severe outlier issue. Both are skewed toward the right and need to be fixed in section 2.

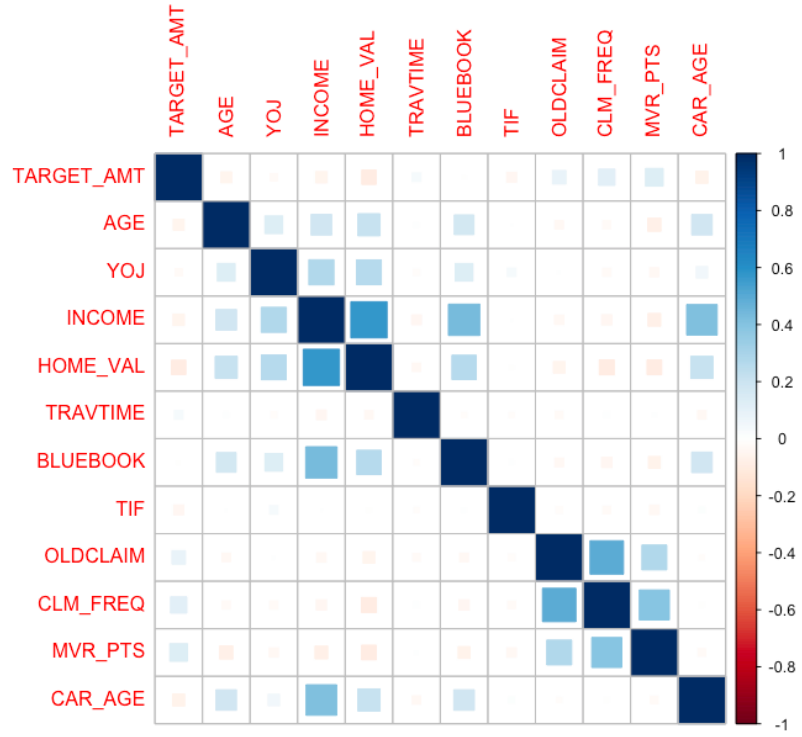


The plots above don't show any outlier in CLM_FREQ although its skewness is higher than 1, so for safety reason, we still consider this variable as one with small outlier issue. In contrast, MVR_PTS plots show that the variable has small outlier issue skewing toward the right. Both need to be fixed before putting them in the models.



Both the skewness number and plots above show that CAR_AGE doesn't have outlier issue. However, there are many vehicles with value of 0, which indicate that they are brand new cars. This is an interesting finding that can be applied to further sections of the report.

From examining the skewness numbers, histograms, and boxplots of each predictors, AGE and CAR_AGE are the only variables with no outlier issue. OLDCLAIM has a severe outlier issue whereas the remaining numeric variables have small outlier issues.



The correlation plot shows the correlation between the predictors and the second response variable TARGET_AMT along with the correlation among the predictors. The darker the color,

the stronger the relationship. Blue represents positive correlation whereas red represents negative correlation. From the plot, there's no predictor that has a strong direct relationship with TARGET_AMT. However, INCOME has a strong relationship with HOME_VAL, BLUEBOOK, CAR_AGE, so this variable may cause multicollinearity issues later on, which is something we should be cautious about. This makes sense because people with higher income are more likely to have more expensive houses and cars. However, it's interesting that there's a positive correlation between INCOME and CAR_AGE, perhaps higher income individuals buy cars with higher quality and thus last longer. Also, CLM_FREQ and OLDCLAIM have a positive relationship as well, which makes sense that people with more claims tend to have higher claim value. People with no claim with have claim value of 0.

Section 2: Data Preparation

Subsection 2.1: Missing Value Issues

As mentioned earlier, the following variables have missing values that need to be addressed.

- AGE
- YOJ
- INCOME
- HOME_VAL
- JOB
- CAR_AGE

Best practice to handle missing values is to create two additional variables. First, a flag variable (with "M" at the beginning of the variable name) is created with 1 indicating missing values and 0 indicating known values. Second, an imputed variable (with "IMP" at the beginning of the variable name) is created to replace missing values with the mean or median and keep the known values the same. The M and IMP variables will be used in the predictive models instead of the original predictors. Below is how we handle IMP variables for the predictors with missing values.

- AGE: replace missing values with the mean of 44.79031 since there's no outlier issue
- YOJ: replace missing values with the median of 11 since there's outlier issue
- INCOME: replace missing values with the median of 54,028 since there's outlier issue
- HOME_VAL: replace missing values with the median of 161,160 since there's outlier issue
- JOB: since this is a categorical variable, we create a separate category called "Unknown" to indicate records with missing values. An IMP_JOB variable is created, but no M_JOB variable is created. There's no need to flag the missing values since they're already represented as "Unknown" category in the new variable IMP_JOB.
- CAR_AGE: replace missing values with the mean of 8.329804 since there's no outlier issue

Putting all of the new variables together and removing the some of the original variables with missing values, we have a new data frame below to continue the project. The output below shows that all of the variables are in the appropriate format, and there's no more missing value in the dataset.

```
> str(newdata)
'data.frame':   8161 obs. of  31 variables:
 $ INDEX       : int  1 2 4 5 6 7 8 11 12 13 ...
 $ 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    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
 $ HOMEKIDS    : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 1 ...
 $ PARENT1     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
 $ 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 ...
 $ 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 "no","yes": 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 ...
 $ URBANICITY  : Factor w/ 2 levels "Highly Urban/ Urban",..: 1 1 1 1 1 1 1 1
1 2 ...
 $ IMP_AGE     : num  60 43 35 51 50 34 54 37 34 50 ...
 $ M_AGE       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ IMP_YOJ     : int  11 11 10 14 11 12 11 11 10 7 ...
 $ M_YOJ       : num  0 0 0 0 1 0 1 1 0 0 ...
 $ IMP_INCOME  : num  67349 91449 16039 54028 114986 ...
 $ M_INCOME    : num  0 0 0 1 0 0 0 0 0 0 ...
 $ IMP_HOME_VAL: num  0 257252 124191 306251 243925 ...
 $ M_HOME_VAL  : num  0 0 0 0 0 0 1 0 0 0 ...
 $ IMP_JOB     : Factor w/ 9 levels "Clerical","Doctor",..: 6 9 1 9 2 9 9 9 1
6 ...
 $ IMP_CAR_AGE : num  18 1 10 6 17 7 1 7 1 17 ...
 $ M_CAR_AGE   : num  0 0 0 0 0 0 0 0 0 0 ...
```

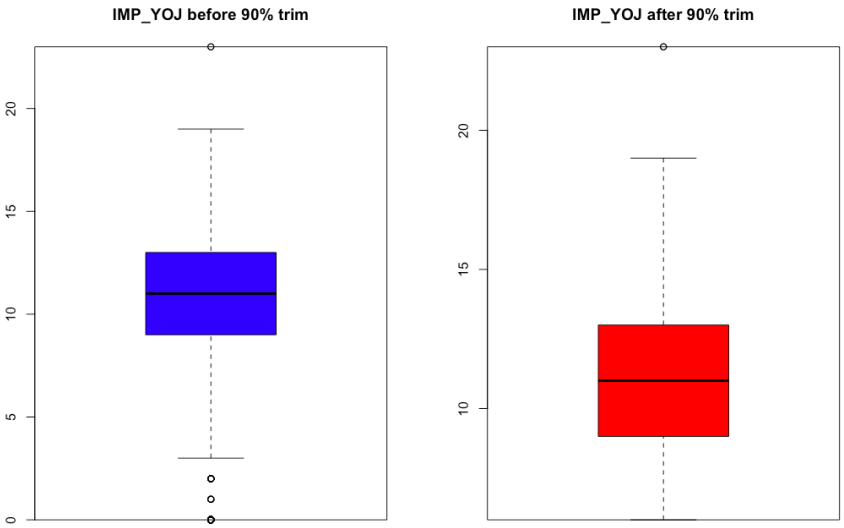
```
> summary(newdata)
     INDEX       TARGET_FLAG   TARGET_AMT     KIDSDRIV  HOMEKIDS  PARENT1                 URBANICITY      IMP_AGE          M_AGE
 Min.   :    1   0:6008      Min.   :     0   0:7180   0:5289   No :7084   Highly Urban/ Urban  :6492   Min.   :16.00   Min.   :0.0000000
 1st Qu.: 2559   1:2153      1st Qu.:     0   1: 981   1:2872   Yes:1077   z_Highly Rural/ Rural:1669   1st Qu.:39.00   1st Qu.:0.0000000
 Median : 5133               Median :     0                                                            Median :45.00   Median :0.0000000
 Mean   : 5152               Mean   :  1504                                                            Mean   :44.79   Mean   :0.0007352
 3rd Qu.: 7745               3rd Qu.:  1036                                                            3rd Qu.:51.00   3rd Qu.:0.0000000
 Max.   :10302               Max.   :107586                                                            Max.   :81.00   Max.   :1.0000000

 MSTATUS      SEX              EDUCATION       TRAVTIME          IMP_YOJ         M_YOJ          IMP_INCOME        M_INCOME
 Yes :4894   M  :3786   <High School :1203   Min.   :  5.00   Min.   : 0.00   Min.   :0.00000   Min.   :     0   Min.   :0.00000
 z_No:3267   z_F:4375   Bachelors    :2242   1st Qu.: 22.00   1st Qu.: 9.00   1st Qu.:0.00000   1st Qu.: 29707   1st Qu.:0.00000
                        Masters      :1658   Median : 33.00   Median :11.00   Median :0.00000   Median : 54028   Median :0.00000
                        PhD          : 728   Mean   : 33.49   Mean   :10.53   Mean   :0.05563   Mean   : 61469   Mean   :0.05453
                        z_High School:2330   3rd Qu.: 44.00   3rd Qu.:13.00   3rd Qu.:0.00000   3rd Qu.: 83304   3rd Qu.:0.00000
                                             Max.   :142.00   Max.   :23.00   Max.   :1.00000   Max.   :367030   Max.   :1.00000

      CAR_USE        BLUEBOOK          TIF              CAR_TYPE      IMP_HOME_VAL      M_HOME_VAL            IMP_JOB      IMP_CAR_AGE
 Commercial:3029   Min.   : 1500   Min.   : 1.000   Minivan    :2145   Min.   :     0   Min.   :0.00000   z_Blue Collar:1825   Min.   : 0.00
 Private   :5132   1st Qu.: 9280   1st Qu.: 1.000   Panel Truck: 676   1st Qu.:     0   1st Qu.:0.00000   Clerical     :1271   1st Qu.: 4.00
                   Median :14440   Median : 4.000   Pickup     :1389   Median :161160   Median :0.00000   Professional :1117   Median : 8.33
                   Mean   :15710   Mean   : 5.351   Sports Car : 907   Mean   :155225   Mean   :0.05686   Manager      : 988   Mean   : 8.33
                   3rd Qu.:20850   3rd Qu.: 7.000   Van        : 750   3rd Qu.:233352   3rd Qu.:0.00000   Lawyer       : 835   3rd Qu.:12.00
                   Max.   :69740   Max.   :25.000   z_SUV      :2294   Max.   :885282   Max.   :1.00000   Student      : 712   Max.   :28.00
                                                                                                        (Other)      :1413

 RED_CAR       OLDCLAIM        CLM_FREQ      REVOKED       MVR_PTS          M_CAR_AGE
 no :5783   Min.   :    0   Min.   :0.0000   No :7161   Min.   : 0.000   Min.   :0.00000
 yes:2378   1st Qu.:    0   1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000   1st Qu.:0.00000
            Median :    0   Median :0.0000              Median : 1.000   Median :0.00000
            Mean   : 4037   Mean   :0.7986              Mean   : 1.696   Mean   :0.06261
            3rd Qu.: 4636   3rd Qu.:2.0000              3rd Qu.: 3.000   3rd Qu.:0.00000
            Max.   :57037   Max.   :5.0000              Max.   :13.000   Max.   :1.00000
```

Subsection 2.2: Outlier Issues

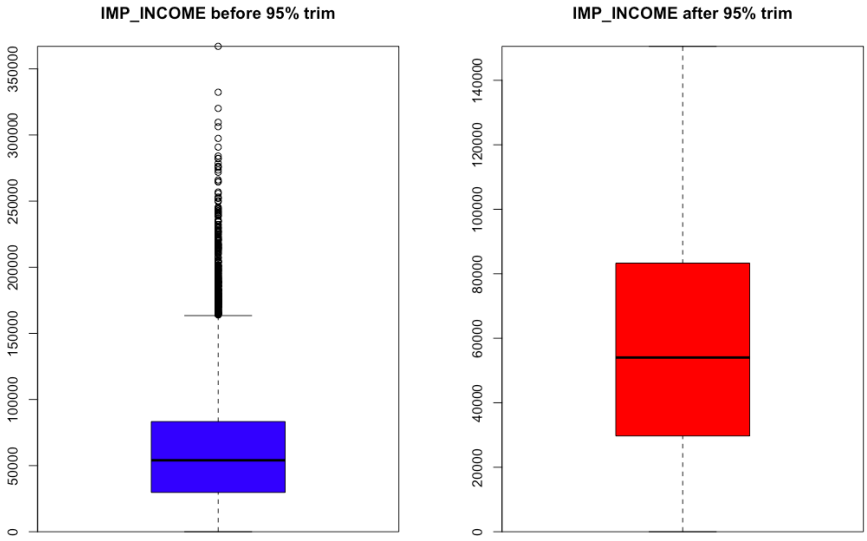
As mentioned earlier, OLDCLAIM has a severe outlier issue whereas the following variables have small outlier issues.

- YOJ
- INCOME
- HOME_VAL
- TRAVTIME
- BLUEBOOK
- TIF
- CLM_FREQ
- MVR_PTS

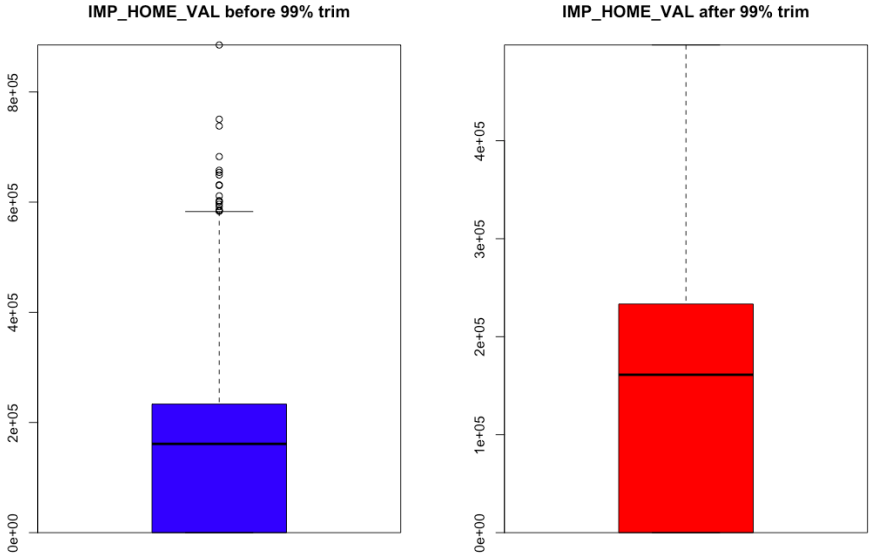
In addition, since outlier issues are caused by many records with 0 in values, new variables with "Z" at the beginning of the variable name are created with 0 means the record has 0 in value, and 1 means the record has value different from 0. These Z variables will be added to the models since they may produce impact on the target variables.



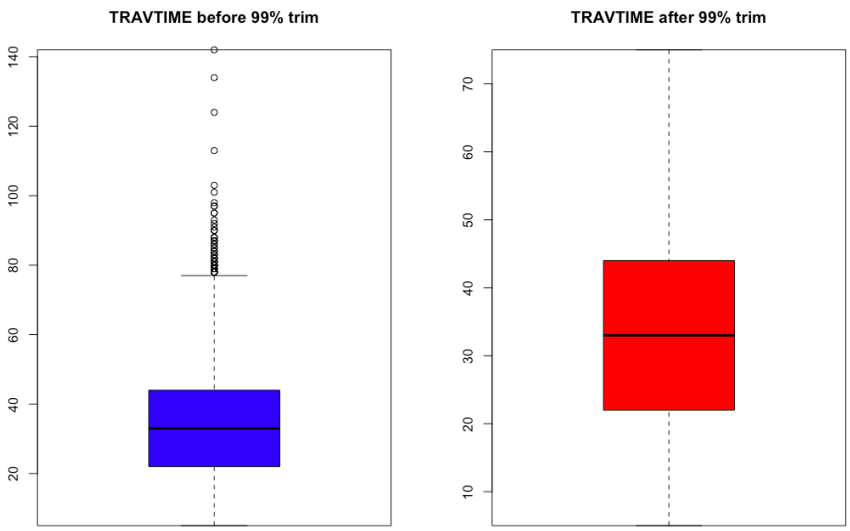
Using 90% trim, we have successfully addressed the outlier issue in YOJ as shown by the boxplots above. We tried 95% trim before, and it didn't completely address the issue, so we took a step further to use 90% trim.



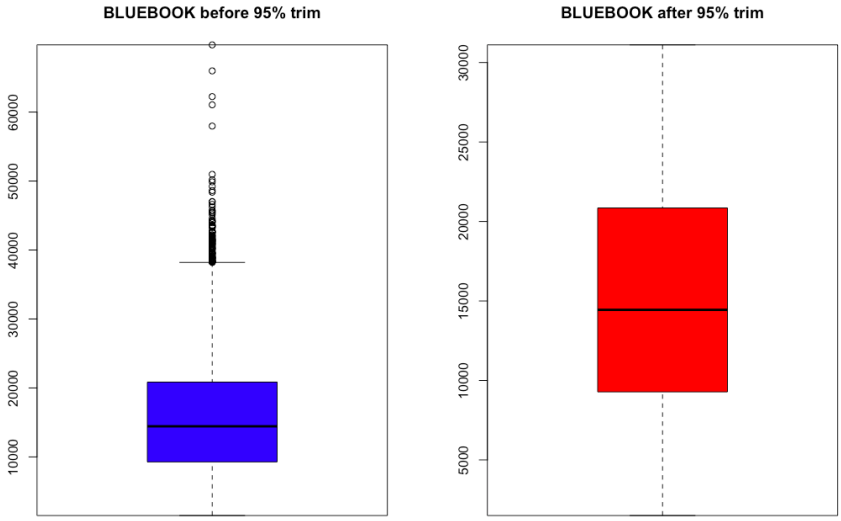
The outlier issue in INCOME is not as severe, so a 95% trim successfully addressed this issue. We tried 99% trim, and it didn't fix the issue, so we tried 95% trim, and it fixed the problem.



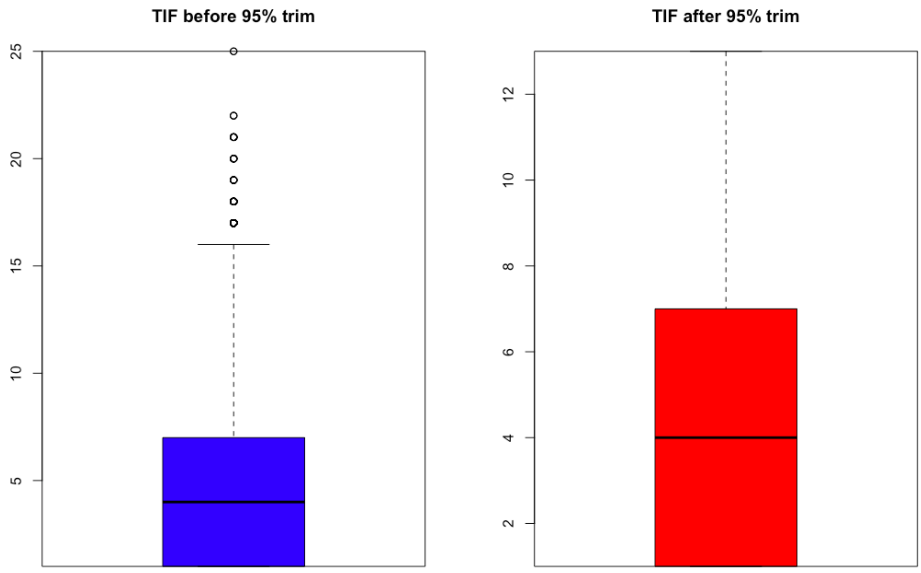
The outlier issue in HOME_VAL is small, so a 99% trim fixed the problem.



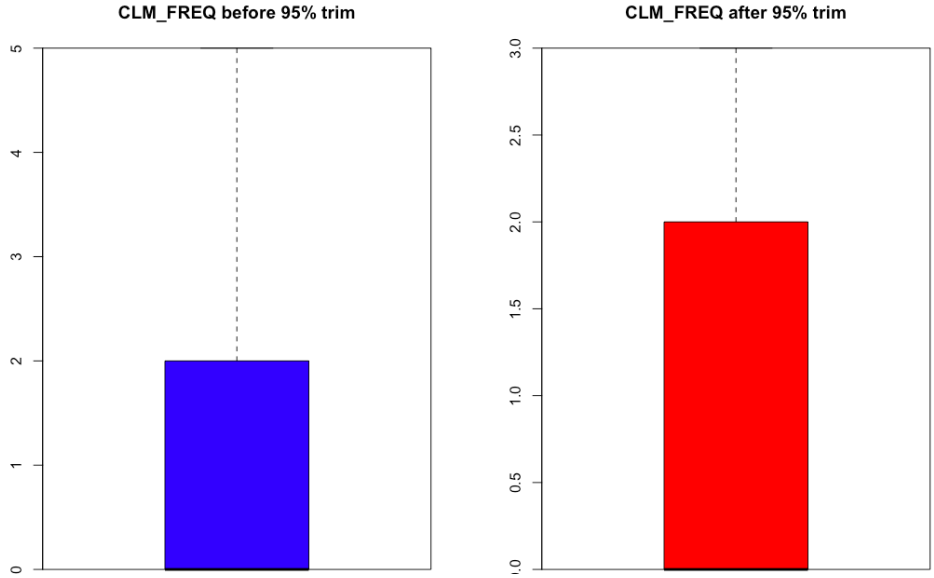
Similarly, TRAVTIME only has a small outlier issue, so a 99% trim can fix the issue.



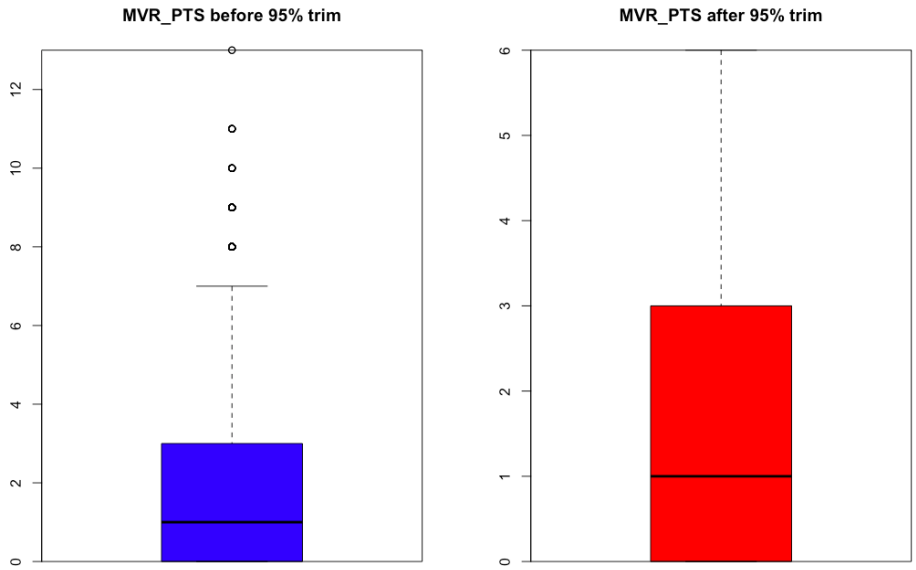
BLUEBOOK has a little more severe outlier issue, so we used 95% trim to address the problem.



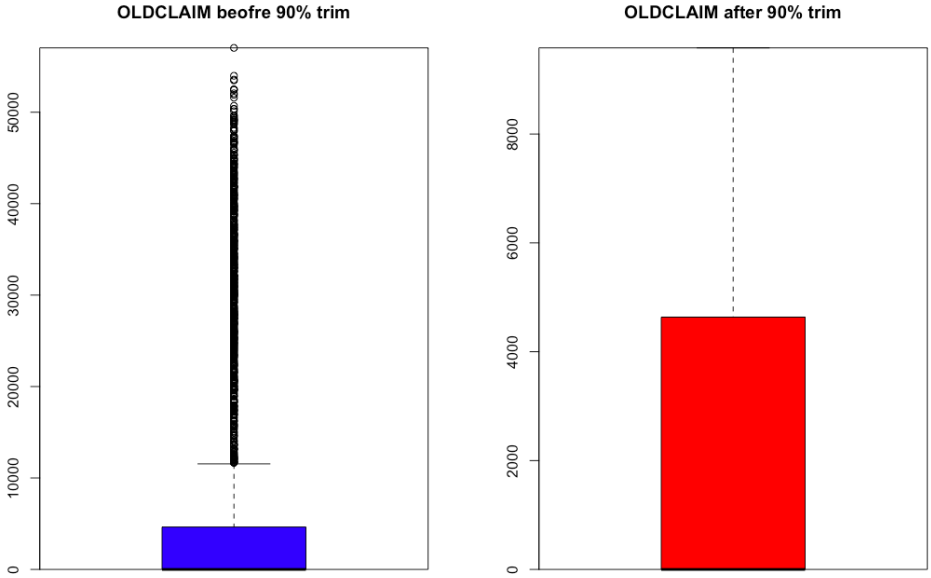
Similarly, we used 95% trim with TIF to address its moderate outlier issue.



Since there's no outlier before and after the 95% trim in CLM_FREQ and the values have a tight range from 0 to 5, we decided to keep the variable the same and use the original variable in the models.



Using a 95% trim, we successfully addressed the outlier issue in MVR_PTS.



OLDCLAIM has the most severe outlier issue among all variables with skewness number above 3, so we used a 90% trim to completely address this problem.

By adding new variables in the data frame, including M flag, IMP imputed, and Z zero variables, we have the following new data frame to start the model building process in section 3. All of the variables are in the appropriate format with no more missing value and no outlier in the dataset.

```
> str(newdata)
'data.frame':   8161 obs. of  37 variables:
 $ INDEX       : int  1 2 4 5 6 7 8 11 12 13 ...
 $ 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    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 1 ...
 $ HOMEKIDS    : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 1 ...
 $ PARENT1     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
 $ 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 ...
 $ TRAVTIME    : num  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         : num  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 "no","yes": 2 2 1 2 1 1 1 2 1 1 ...
 $ OLDCLAIM    : num  4461 0 9583 0 9583 ...
 $ 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     : num  3 0 3 0 3 0 0 6 0 1 ...
 $ URBANICITY  : Factor w/ 2 levels "Highly Urban/ Urban",..: 1 1 1 1 1 1 1 1
1 2 ...
 $ IMP_AGE     : num  60 43 35 51 50 34 54 37 34 50 ...
 $ M_AGE       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ IMP_YOJ     : num  11 11 10 14 11 12 11 11 10 7 ...
 $ M_YOJ       : num  0 0 0 0 1 0 1 1 0 0 ...
 $ IMP_INCOME  : num  67349 91449 16039 54028 114986 ...
 $ M_INCOME    : num  0 0 0 1 0 0 0 0 0 0 ...
 $ IMP_HOME_VAL: num  0 257252 124191 306251 243925 ...
 $ M_HOME_VAL  : num  0 0 0 0 0 0 1 0 0 0 ...
 $ IMP_JOB     : Factor w/ 9 levels "Clerical","Doctor",..: 6 9 1 9 2 9 9 9 1
6 ...
 $ IMP_CAR_AGE : num  18 1 10 6 17 7 1 7 1 17 ...
 $ M_CAR_AGE   : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Z_YOJ       : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ Z_INCOME    : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ Z_HOME_VAL  : Factor w/ 2 levels "0","1": 1 2 2 2 2 1 2 2 1 1 ...
 $ Z_CLM_FREQ  : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 1 2 1 1 ...
 $ Z_MVR_PTS   : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 1 2 1 2 ...
 $ Z_OLDCLAIM  : Factor w/ 2 levels "0","1": 2 1 2 1 2 1 1 2 1 1 ...
```

```
> summary(newdata)
     INDEX        TARGET_FLAG   TARGET_AMT      KIDSDRIV    HOMEKIDS    PARENT1
 Min.   :    1   0:6008      Min.   :     0   0:7180   0:5289   No :7084
 1st Qu.: 2559   1:2153      1st Qu.:     0   1: 981   1:2872   Yes:1077
 Median : 5133               Median :     0
 Mean   : 5152               Mean   :  1504
 3rd Qu.: 7745               3rd Qu.:  1036
 Max.   :10302               Max.   :107586

 MSTATUS      SEX              EDUCATION       TRAVTIME
 Yes :4894   M  :3786   <High School :1203   Min.   : 5.00
 z_No:3267   z_F:4375   Bachelors    :2242   1st Qu.:22.00
                        Masters      :1658   Median :33.00
                        PhD          : 728   Mean   :33.39
                        z_High School:2330   3rd Qu.:44.00
                                             Max.   :75.00

      CAR_USE        BLUEBOOK          TIF              CAR_TYPE
 Commercial:3029   Min.   : 1500   Min.   : 1.000   Minivan    :2145
 Private   :5132   1st Qu.: 9280   1st Qu.: 1.000   Panel Truck: 676
                   Median :14440   Median : 4.000   Pickup     :1389
                   Mean   :15459   Mean   : 5.223   Sports Car : 907
                   3rd Qu.:20850   3rd Qu.: 7.000   Van        : 750
                   Max.   :31110   Max.   :13.000   z_SUV      :2294

 RED_CAR       OLDCLAIM       CLM_FREQ      REVOKED       MVR_PTS
 no :5783   Min.   :   0   Min.   :0.0000   No :7161   Min.   :0.000
 yes:2378   1st Qu.:   0   1st Qu.:0.0000   Yes:1000   1st Qu.:0.000
            Median :   0   Median :0.0000              Median :1.000
            Mean   :2334   Mean   :0.7986              Mean   :1.623
            3rd Qu.:4636   3rd Qu.:2.0000              3rd Qu.:3.000
            Max.   :9583   Max.   :5.0000              Max.   :6.000
                  URBANICITY       IMP_AGE          M_AGE
 Highly Urban/ Urban  :6492   Min.   :16.00   Min.   :0.0000000
 z_Highly Rural/ Rural:1669   1st Qu.:39.00   1st Qu.:0.0000000
                              Median :45.00   Median :0.0000000
                              Mean   :44.79   Mean   :0.0007352
                              3rd Qu.:51.00   3rd Qu.:0.0000000
                              Max.   :81.00   Max.   :1.0000000

    IMP_YOJ          M_YOJ          IMP_INCOME        M_INCOME
 Min.   : 6.00   Min.   :0.00000   Min.   :     0   Min.   :0.00000
 1st Qu.: 9.00   1st Qu.:0.00000   1st Qu.: 29707   1st Qu.:0.00000
 Median :11.00   Median :0.00000   Median : 54028   Median :0.00000
 Mean   :11.03   Mean   :0.05563   Mean   : 59536   Mean   :0.05453
 3rd Qu.:13.00   3rd Qu.:0.00000   3rd Qu.: 83304   3rd Qu.:0.00000
 Max.   :23.00   Max.   :1.00000   Max.   :150529   Max.   :1.00000

  IMP_HOME_VAL      M_HOME_VAL              IMP_JOB      IMP_CAR_AGE
 Min.   :     0   Min.   :0.00000   z_Blue Collar:1825   Min.   : 0.00
 1st Qu.:     0   1st Qu.:0.00000   Clerical     :1271   1st Qu.: 4.00
 Median :161160   Median :0.00000   Professional :1117   Median : 8.33
 Mean   :154643   Mean   :0.05686   Manager      : 988   Mean   : 8.33
 3rd Qu.:233352   3rd Qu.:0.00000   Lawyer       : 835   3rd Qu.:12.00
 Max.   :497746   Max.   :1.00000   Student      : 712   Max.   :28.00
                                    (Other)      :1413
   M_CAR_AGE       Z_YOJ     Z_INCOME  Z_HOME_VAL Z_CLM_FREQ Z_MVR_PTS
 Min.   :0.00000   0: 625   0: 615   0:2294     0:5009     0:3712
 1st Qu.:0.00000   1:7536   1:7546   1:5867     1:3152     1:4449
 Median :0.00000
 Mean   :0.06261
 3rd Qu.:0.00000
 Max.   :1.00000

 Z_OLDCLAIM
 0:5009
 1:3152
```

Section 3: Logistic Regression Model Development

In this section, we will develop multiple logistic regression models to predict TARGET_FLAG and then use multiple metrics to choose the best model.

Subsection 3.1: Model #1 – Full Model

By putting all of the variables in the data frame into the logistic model, we have the following results.

```
> summary(model1)

Call:
glm(formula = newdata$TARGET_FLAG ~ ., family = binomial(), data = newdata)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-1.901e-03  -2.000e-08  -2.000e-08   2.000e-08   2.924e-03

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.930e+01  3.550e+03  -0.005    0.996
INDEX                          1.251e-03  1.118e-01   0.011    0.991
TARGET_AMT                     2.827e-01  2.997e+00   0.094    0.925
KIDSDRIV1                      2.330e+01  2.599e+03   0.009    0.993
HOMEKIDS1                     -2.485e+01  2.530e+03  -0.010    0.992
PARENT1Yes                    -3.631e+00  2.035e+03  -0.002    0.999
MSTATUSz_No                    1.944e+00  6.922e+02   0.003    0.998
SEXz_F                        -3.762e+00  1.745e+03  -0.002    0.998
EDUCATIONBachelors            -2.632e+01  2.074e+03  -0.013    0.990
EDUCATIONMasters              -2.364e+01  5.870e+03  -0.004    0.997
EDUCATIONPhD                  -1.979e+01  5.732e+03  -0.003    0.997
EDUCATIONz_High School        -4.023e-01  7.615e+02  -0.001    1.000
TRAVTIME                       3.949e-02  1.552e+01   0.003    0.998
CAR_USEPrivate                -1.432e+01  1.203e+03  -0.012    0.991
BLUEBOOK                      -1.978e-03  8.645e-02  -0.023    0.982
TIF                           -8.600e-01  9.888e+01  -0.009    0.993
CAR_TYPEPanel Truck           -6.347e+01  1.284e+05   0.000    1.000
CAR_TYPEPickup                -2.778e-01  6.480e+02   0.000    1.000
CAR_TYPESports Car             7.976e+00  1.566e+03   0.005    0.996
CAR_TYPEVan                    1.109e+00  2.945e+03   0.000    1.000
CAR_TYPEz_SUV                 -1.862e+01  2.156e+03  -0.009    0.993
RED_CARyes                    -2.069e+00  8.264e+02  -0.003    0.998
OLDCLAIM                       4.944e-05  8.070e-02   0.001    1.000
CLM_FREQ                      -2.607e-01  4.444e+02  -0.001    1.000
REVOKEDYes                     7.483e+00  7.175e+02   0.010    0.992
MVR_PTS                        6.127e+00  4.479e+02   0.014    0.989
URBANICITYz_Highly Rural/ Rural -6.289e+00  9.617e+02  -0.007    0.995
```

```
IMP_AGE                      -2.733e-01  4.774e+01  -0.006    0.995
M_AGE                        -6.144e+02  1.173e+05  -0.005    0.996
IMP_YOJ                       1.541e-01  2.016e+02   0.001    0.999
M_YOJ                         9.639e+00  2.589e+03   0.004    0.997
IMP_INCOME                    2.975e-05  2.968e-02   0.001    0.999
M_INCOME                     -8.518e+00  2.237e+03  -0.004    0.997
IMP_HOME_VAL                  4.744e-05  1.442e-02   0.003    0.997
M_HOME_VAL                   -1.072e+01  1.606e+03  -0.007    0.995
IMP_JOBDoctor                -1.965e+02  3.975e+05   0.000    1.000
IMP_JOBHome Maker            -6.140e+00  3.439e+03  -0.002    0.999
IMP_JOBLawyer                -8.799e+00  2.406e+04   0.000    1.000
IMP_JOBManager               -2.382e+00  5.560e+03   0.000    1.000
IMP_JOBProfessional          -8.711e+00  1.651e+04  -0.001    1.000
IMP_JOBStudent                6.450e+00  1.994e+03   0.003    0.997
IMP_JOBUnknown               -1.493e+01  7.503e+04   0.000    1.000
IMP_JOBz_Blue Collar          1.323e-01  7.315e+02   0.000    1.000
IMP_CAR_AGE                   1.259e+00  8.819e+01   0.014    0.989
M_CAR_AGE                     7.859e+00  6.658e+02   0.012    0.991
Z_YOJ1                        9.449e-01  6.036e+03   0.000    1.000
Z_INCOME1                    -5.387e-01  5.886e+03   0.000    1.000
Z_HOME_VAL1                   3.176e+00  1.881e+03   0.002    0.999
Z_CLM_FREQ1                   1.545e+01  1.798e+03   0.009    0.993
Z_MVR_PTS1                   -3.585e+01  2.439e+03  -0.015    0.988
Z_OLDCLAIM1                          NA         NA      NA       NA

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9.418e+03  on 8160  degrees of freedom
Residual deviance: 6.147e-05  on 8111  degrees of freedom
AIC: 100

Number of Fisher Scoring iterations: 25

> str(newdata$predict1)
 num [1:8161] 2.22e-16 2.22e-16 2.22e-16 2.22e-16 2.22e-16 ...
> summary(newdata$predict1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.2638  1.0000  1.0000
```

When we apply model #1 to the train dataset, we have a mean predicted value of 0.2638, which is very close to the mean in the actual data, so that's good.

If a predictor has p-value less than 0.10, we conclude that variable has a statistically significant impact on the response variable at 90% confidence level. If a predictor has p-value less than 0.05, we conclude that variable has a statistically significant impact on the response variable at 95% confidence level. Using this p-value metric, none of the predictor in model #1 has a significant impact on TARGET_FLAG. This may be a result of overfitting since we run a full model here with all of the variables in the dataset in the model. Therefore, since we don't have good results here from model #1, we will need to break out this full model into smaller models in subsection 3.2 and 3.3 and then put it all together in section 3.4 of this section.

Subsection 3.2: Model #2 – Original Predictors

Using only original predictors in model #2, we have the following results.

```
Call:
glm(formula = newdata$TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + PARENT1 +
    MSTATUS + SEX + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK +
    TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ + REVOKED +
    MVR_PTS + URBANICITY, family = binomial(), data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3216  -0.7264  -0.4223   0.6633   3.1607

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -9.403e-01  1.711e-01  -5.496 3.88e-08 ***
KIDSDRIV1                       5.489e-01  9.464e-02   5.800 6.62e-09 ***
HOMEKIDS1                       2.877e-01  8.752e-02   3.287 0.001011 **
PARENT1Yes                      2.132e-01  1.190e-01   1.791 0.073259 .
MSTATUSz_No                     7.065e-01  7.457e-02   9.474  < 2e-16 ***
SEXz_F                         -3.285e-03  1.088e-01  -0.030 0.975924
EDUCATIONBachelors             -7.691e-01  9.393e-02  -8.188 2.65e-16 ***
EDUCATIONMasters               -8.658e-01  1.016e-01  -8.524  < 2e-16 ***
EDUCATIONPhD                   -1.068e+00  1.320e-01  -8.093 5.81e-16 ***
EDUCATIONz_High School         -1.291e-01  9.092e-02  -1.420 0.155592
TRAVTIME                        1.581e-02  1.900e-03   8.323  < 2e-16 ***
CAR_USEPrivate                 -8.301e-01  7.261e-02 -11.431  < 2e-16 ***
BLUEBOOK                       -3.583e-05  5.299e-06  -6.762 1.37e-11 ***
TIF                            -5.985e-02  7.734e-03  -7.738 1.01e-14 ***
CAR_TYPEPanel Truck             5.574e-01  1.496e-01   3.725 0.000195 ***
CAR_TYPEPickup                  5.026e-01  9.759e-02   5.150 2.60e-07 ***
CAR_TYPESports Car              9.681e-01  1.274e-01   7.598 3.00e-14 ***
CAR_TYPEVan                     5.918e-01  1.229e-01   4.816 1.46e-06 ***
CAR_TYPEz_SUV                   7.151e-01  1.094e-01   6.538 6.23e-11 ***
RED_CARyes                     -6.457e-03  8.532e-02  -0.076 0.939677
OLDCLAIM                       -6.303e-08  1.159e-05  -0.005 0.995662
CLM_FREQ                        1.618e-01  3.383e-02   4.783 1.73e-06 ***
REVOKEDYes                      7.540e-01  8.270e-02   9.116  < 2e-16 ***
MVR_PTS                         1.172e-01  1.512e-02   7.755 8.83e-15 ***
URBANICITYz_Highly Rural/ Rural -2.271e+00  1.112e-01 -20.414  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9418.0  on 8160  degrees of freedom
Residual deviance: 7434.5  on 8136  degrees of freedom
AIC: 7484.5

Number of Fisher Scoring iterations: 5
```

Using p-value metric, the following variables are not statistically significant at 90% level.

- SEX
- RED_CAR
- OLDCLAIM

Removing these predictors from the model, we have the following results.

```
> summary(model2)

Call:
glm(formula = newdata$TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + PARENT1 +
    MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK + TIF +
    CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY, family = binomial(),
    data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3226  -0.7265  -0.4222   0.6635   3.1599

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -9.444e-01  1.623e-01  -5.819 5.91e-09 ***
KIDSDRIV1                      5.491e-01  9.457e-02   5.806 6.39e-09 ***
HOMEKIDS1                      2.876e-01  8.711e-02   3.301 0.000963 ***
PARENT1Yes                     2.133e-01  1.190e-01   1.793 0.073009 .
MSTATUSz_No                    7.064e-01  7.454e-02   9.477  < 2e-16 ***
EDUCATIONBachelors            -7.692e-01  9.390e-02  -8.192 2.56e-16 ***
EDUCATIONMasters              -8.657e-01  1.016e-01  -8.525  < 2e-16 ***
EDUCATIONPhD                  -1.068e+00  1.319e-01  -8.098 5.60e-16 ***
EDUCATIONz_High School        -1.291e-01  9.092e-02  -1.420 0.155629
TRAVTIME                       1.581e-02  1.899e-03   8.325  < 2e-16 ***
CAR_USEPrivate                -8.301e-01  7.261e-02 -11.433  < 2e-16 ***
BLUEBOOK                      -3.581e-05  4.827e-06  -7.419 1.18e-13 ***
TIF                           -5.985e-02  7.734e-03  -7.739 1.00e-14 ***
CAR_TYPEPanel Truck            5.570e-01  1.405e-01   3.964 7.38e-05 ***
CAR_TYPEPickup                 5.026e-01  9.756e-02   5.152 2.58e-07 ***
CAR_TYPESports Car             9.687e-01  1.056e-01   9.173  < 2e-16 ***
CAR_TYPEVan                    5.917e-01  1.190e-01   4.972 6.62e-07 ***
CAR_TYPEz_SUV                  7.158e-01  8.465e-02   8.456  < 2e-16 ***
CLM_FREQ                       1.616e-01  2.532e-02   6.384 1.72e-10 ***
REVOKEDYes                     7.538e-01  7.931e-02   9.505  < 2e-16 ***
MVR_PTS                        1.172e-01  1.490e-02   7.866 3.66e-15 ***
URBANICITYz_Highly Rural/ Rural -2.271e+00  1.111e-01 -20.432  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9418.0  on 8160  degrees of freedom
Residual deviance: 7434.5  on 8139  degrees of freedom
AIC: 7478.5

Number of Fisher Scoring iterations: 5

> newdata$predict2=predict(model2,type='response')
> str(newdata$predict2)
 num [1:8161] 0.0826 0.3511 0.3707 0.0962 0.303 ...
> summary(newdata$predict2)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002363 0.086829 0.207831 0.263816 0.396171 0.949374
```

The predicted mean of model #2 is 0.263816, which is very close to the actual mean, so that's good. Using p-value, all of the predictors are significant at 90% confidence level, so we have a good model #2 here.

Subsection 3.3.: Model #3 – New Variables

Using the new created variables in model #3, we have the following results.

```
Call:
glm(formula = newdata$TARGET_FLAG ~ IMP_AGE + M_AGE + IMP_YOJ +
    M_YOJ + IMP_INCOME + M_INCOME + IMP_HOME_VAL + M_HOME_VAL +
    IMP_JOB + IMP_CAR_AGE + M_CAR_AGE + Z_YOJ + Z_INCOME + Z_HOME_VAL +
    Z_CLM_FREQ + Z_MVR_PTS + Z_OLDCLAIM, family = binomial(),
    data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1583  -0.7737  -0.5697   0.9685   2.6541

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          3.740e-02  2.118e-01   0.177 0.859877
IMP_AGE             -1.288e-02  3.234e-03  -3.983 6.81e-05 ***
M_AGE                2.241e+00  1.125e+00   1.993 0.046269 *
IMP_YOJ              1.588e-02  1.134e-02   1.400 0.161364
M_YOJ                7.138e-02  1.195e-01   0.597 0.550256
IMP_INCOME          -4.976e-06  1.269e-06  -3.922 8.78e-05 ***
M_INCOME            -4.731e-02  1.225e-01  -0.386 0.699297
IMP_HOME_VAL        -6.922e-07  5.875e-07  -1.178 0.238740
M_HOME_VAL           9.592e-02  1.198e-01   0.801 0.423408
IMP_JOBDoctor       -6.302e-01  2.324e-01  -2.712 0.006683 **
IMP_JOBHome Maker   -3.104e-01  1.326e-01  -2.341 0.019242 *
IMP_JOBLawyer       -1.743e-01  1.320e-01  -1.320 0.186829
IMP_JOBManager      -5.999e-01  1.258e-01  -4.767 1.87e-06 ***
IMP_JOBProfessional -7.875e-02  1.084e-01  -0.727 0.467452
IMP_JOBStudent      -3.492e-01  1.285e-01  -2.717 0.006594 **
IMP_JOBUnknown       2.587e-01  1.467e-01   1.764 0.077718 .
IMP_JOBz_Blue Collar 4.256e-01  8.626e-02   4.934 8.04e-07 ***
IMP_CAR_AGE         -8.489e-03  6.100e-03  -1.392 0.164013
M_CAR_AGE            1.318e-01  1.089e-01   1.211 0.226008
Z_YOJ1              -8.873e-02  2.879e-01  -0.308 0.757931
Z_INCOME1           -4.490e-01  2.839e-01  -1.581 0.113806
Z_HOME_VAL1         -5.231e-01  1.351e-01  -3.873 0.000108 ***
Z_CLM_FREQ1          9.392e-01  5.607e-02  16.750  < 2e-16 ***
Z_MVR_PTS1           3.868e-01  5.765e-02   6.709 1.96e-11 ***
Z_OLDCLAIM1                 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: 8438.9  on 8137  degrees of freedom
AIC: 8486.9

Number of Fisher Scoring iterations: 4
```

Using p-value, many predictors in model #3 are not significant. So removing these variables that are not significant at 90% level from the model, we have the following result.

```
Call:
glm(formula = newdata$TARGET_FLAG ~ IMP_AGE + M_AGE + IMP_INCOME +
    IMP_JOB + Z_HOME_VAL + Z_CLM_FREQ + Z_MVR_PTS + Z_OLDCLAIM,
    family = binomial(), data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0268  -0.7769  -0.5725   0.9873   2.6000

Coefficients: (1 not defined because of singularities)
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.777e-01  1.595e-01  -1.741   0.0816 .
IMP_AGE             -1.289e-02  3.185e-03  -4.047 5.19e-05 ***
M_AGE                2.275e+00  1.119e+00   2.033   0.0421 *
IMP_INCOME          -6.793e-06  1.005e-06  -6.756 1.42e-11 ***
IMP_JOBDoctor       -6.162e-01  2.272e-01  -2.712   0.0067 **
IMP_JOBHome Maker   -1.236e-01  1.150e-01  -1.075   0.2824
IMP_JOBLawyer       -2.040e-01  1.233e-01  -1.655   0.0980 .
IMP_JOBManager      -6.052e-01  1.232e-01  -4.913 8.95e-07 ***
IMP_JOBProfessional -7.780e-02  1.066e-01  -0.730   0.4653
IMP_JOBStudent      -1.917e-01  1.125e-01  -1.704   0.0885 .
IMP_JOBUnknown       2.442e-01  1.396e-01   1.749   0.0803 .
IMP_JOBz_Blue Collar 4.372e-01  8.610e-02   5.078 3.82e-07 ***
Z_HOME_VAL1         -6.428e-01  6.078e-02 -10.576  < 2e-16 ***
Z_CLM_FREQ1          9.448e-01  5.590e-02  16.901  < 2e-16 ***
Z_MVR_PTS1           3.860e-01  5.749e-02   6.713 1.91e-11 ***
Z_OLDCLAIM1                 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: 8459.8  on 8146  degrees of freedom
AIC: 8489.8

Number of Fisher Scoring iterations: 4
```

The output above has NA for Z_OLDCLAIM because it has a big multicollinearity issue with other predictors. So we remove this variable from the model #3 and get the following result.

```
> summary(model3)

Call:
glm(formula = newdata$TARGET_FLAG ~ IMP_AGE + M_AGE + IMP_INCOME +
    IMP_JOB + Z_HOME_VAL + Z_CLM_FREQ + Z_MVR_PTS, family = binomial(),
    data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0268  -0.7769  -0.5725   0.9873   2.6000

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.777e-01  1.595e-01  -1.741   0.0816 .
IMP_AGE             -1.289e-02  3.185e-03  -4.047 5.19e-05 ***
M_AGE                2.275e+00  1.119e+00   2.033   0.0421 *
IMP_INCOME          -6.793e-06  1.005e-06  -6.756 1.42e-11 ***
IMP_JOBDoctor       -6.162e-01  2.272e-01  -2.712   0.0067 **
IMP_JOBHome Maker   -1.236e-01  1.150e-01  -1.075   0.2824
IMP_JOBLawyer       -2.040e-01  1.233e-01  -1.655   0.0980 .
IMP_JOBManager      -6.052e-01  1.232e-01  -4.913 8.95e-07 ***
IMP_JOBProfessional -7.780e-02  1.066e-01  -0.730   0.4653
IMP_JOBStudent      -1.917e-01  1.125e-01  -1.704   0.0885 .
IMP_JOBUnknown       2.442e-01  1.396e-01   1.749   0.0803 .
IMP_JOBz_Blue Collar 4.372e-01  8.610e-02   5.078 3.82e-07 ***
Z_HOME_VAL1         -6.428e-01  6.078e-02 -10.576  < 2e-16 ***
Z_CLM_FREQ1          9.448e-01  5.590e-02  16.901  < 2e-16 ***
Z_MVR_PTS1           3.860e-01  5.749e-02   6.713 1.91e-11 ***
---
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: 8459.8  on 8146  degrees of freedom
AIC: 8489.8

Number of Fisher Scoring iterations: 4

> summary(newdata$predict3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03405 0.14730 0.23022 0.26382 0.36287 0.93884
```

The predicted value of model #3 has a mean of 0.26382, which is close to the actual mean value of TARGET_FLAG, so that's good. After removing several predictors, all of the variables in the model now are significant at 90% confidence level.

Subsection 3.4: Model #4 – Combined Model

By combining the significant variables in models #2 and #3 together in model #4, we have the following result.

```
Call:
glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + PARENT1 + MSTATUS +
    EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE +
    CLM_FREQ + REVOKED + MVR_PTS + URBANICITY + IMP_AGE + M_AGE +
    IMP_INCOME + IMP_JOB + Z_HOME_VAL + Z_CLM_FREQ + Z_MVR_PTS,
    family = binomial(), data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3392  -0.7192  -0.3989   0.6385   3.1568

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -5.391e-01  2.783e-01  -1.937 0.052749 .
KIDSDRIV1                      5.836e-01  9.811e-02   5.949 2.70e-09 ***
HOMEKIDS1                      2.363e-01  9.793e-02   2.413 0.015838 *
PARENT1Yes                     2.305e-01  1.206e-01   1.911 0.056001 .
MSTATUSz_No                    5.278e-01  8.887e-02   5.939 2.87e-09 ***
EDUCATIONBachelors            -3.618e-01  1.096e-01  -3.300 0.000965 ***
EDUCATIONMasters              -2.461e-01  1.625e-01  -1.515 0.129878
EDUCATIONPhD                  -1.768e-01  1.990e-01  -0.889 0.374235
EDUCATIONz_High School         2.603e-02  9.490e-02   0.274 0.783863
TRAVTIME                       1.553e-02  1.921e-03   8.087 6.11e-16 ***
CAR_USEPrivate                -7.754e-01  9.190e-02  -8.437  < 2e-16 ***
BLUEBOOK                      -2.760e-05  5.035e-06  -5.481 4.23e-08 ***
TIF                           -6.047e-02  7.837e-03  -7.716 1.20e-14 ***
CAR_TYPEPanel Truck            6.374e-01  1.502e-01   4.244 2.20e-05 ***
CAR_TYPEPickup                 5.517e-01  1.008e-01   5.471 4.48e-08 ***
CAR_TYPESports Car             9.495e-01  1.081e-01   8.786  < 2e-16 ***
CAR_TYPEVan                    6.603e-01  1.225e-01   5.391 7.02e-08 ***
CAR_TYPEz_SUV                  6.982e-01  8.628e-02   8.093 5.83e-16 ***
CLM_FREQ                       5.473e-02  4.445e-02   1.231 0.218236
REVOKEDYes                     7.323e-01  8.051e-02   9.095  < 2e-16 ***
MVR_PTS                        8.542e-02  2.264e-02   3.773 0.000161 ***
URBANICITYz_Highly Rural/ Rural -2.361e+00  1.127e-01 -20.947  < 2e-16 ***
IMP_AGE                        3.105e-04  4.058e-03   0.077 0.939007
M_AGE                          2.342e+00  1.291e+00   1.813 0.069763 .
IMP_INCOME                    -6.292e-06  1.154e-06  -5.455 4.91e-08 ***
IMP_JOBDoctor                 -7.271e-01  2.853e-01  -2.549 0.010809 *
IMP_JOBHome Maker             -1.449e-01  1.373e-01  -1.055 0.291279
IMP_JOBLawyer                 -2.678e-01  1.859e-01  -1.440 0.149736
IMP_JOBManager                -9.182e-01  1.445e-01  -6.356 2.07e-10 ***
IMP_JOBProfessional           -1.952e-01  1.250e-01  -1.561 0.118489
IMP_JOBStudent                -2.846e-01  1.330e-01  -2.139 0.032405 *
IMP_JOBUnknown                -3.905e-01  1.964e-01  -1.988 0.046777 *
IMP_JOBz_Blue Collar          -5.876e-02  1.072e-01  -0.548 0.583485
Z_HOME_VAL1                   -3.522e-01  8.363e-02  -4.211 2.54e-05 ***
Z_CLM_FREQ1                    3.213e-01  1.134e-01   2.835 0.004588 **
Z_MVR_PTS1                     5.577e-02  9.088e-02   0.614 0.539430
---
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: 7288.5  on 8125  degrees of freedom
AIC: 7360.5

Number of Fisher Scoring iterations: 5
```

Some of the predictors are no longer significant, probably due to multicollinearity. By removing CLM_FREQ, IMP_AGE, and Z_MVR_PTS from the model, we have the following result.

```
> summary(model4)

Call:
glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + PARENT1 + MSTATUS +
    EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE +
    REVOKED + MVR_PTS + URBANICITY + M_AGE + IMP_INCOME + IMP_JOB +
    Z_HOME_VAL + Z_CLM_FREQ, family = binomial(), data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3482  -0.7197  -0.3989   0.6386   3.1532

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -5.119e-01  2.112e-01  -2.424  0.01533 *
KIDSDRIV1                      5.875e-01  9.589e-02   6.127 8.96e-10 ***
HOMEKIDS1                      2.310e-01  8.855e-02   2.608  0.00909 **
PARENT1Yes                     2.304e-01  1.206e-01   1.911  0.05603 .
MSTATUSz_No                    5.288e-01  8.850e-02   5.976 2.29e-09 ***
EDUCATIONBachelors            -3.597e-01  1.096e-01  -3.282  0.00103 **
EDUCATIONMasters              -2.411e-01  1.624e-01  -1.484  0.13768
EDUCATIONPhD                  -1.699e-01  1.987e-01  -0.855  0.39241
EDUCATIONz_High School         2.783e-02  9.488e-02   0.293  0.76930
TRAVTIME                       1.556e-02  1.920e-03   8.107 5.20e-16 ***
CAR_USEPrivate                -7.756e-01  9.189e-02  -8.441  < 2e-16 ***
BLUEBOOK                      -2.774e-05  5.012e-06  -5.535 3.11e-08 ***
TIF                           -6.047e-02  7.835e-03  -7.717 1.19e-14 ***
CAR_TYPEPanel Truck            6.398e-01  1.501e-01   4.263 2.02e-05 ***
CAR_TYPEPickup                 5.534e-01  1.008e-01   5.491 3.99e-08 ***
CAR_TYPESports Car             9.512e-01  1.078e-01   8.823  < 2e-16 ***
CAR_TYPEVan                    6.587e-01  1.225e-01   5.379 7.50e-08 ***
CAR_TYPEz_SUV                  6.974e-01  8.621e-02   8.090 5.98e-16 ***
REVOKEDYes                     7.308e-01  8.047e-02   9.081  < 2e-16 ***
MVR_PTS                        9.505e-02  1.560e-02   6.094 1.10e-09 ***
URBANICITYz_Highly Rural/ Rural -2.362e+00  1.127e-01 -20.960  < 2e-16 ***
M_AGE                          2.380e+00  1.312e+00   1.813  0.06982 .
IMP_INCOME                    -6.273e-06  1.153e-06  -5.439 5.36e-08 ***
IMP_JOBDoctor                 -7.217e-01  2.845e-01  -2.537  0.01119 *
IMP_JOBHome Maker             -1.439e-01  1.372e-01  -1.049  0.29409
IMP_JOBLawyer                 -2.713e-01  1.855e-01  -1.463  0.14355
IMP_JOBManager                -9.170e-01  1.441e-01  -6.364 1.97e-10 ***
IMP_JOBProfessional           -1.939e-01  1.248e-01  -1.554  0.12028
IMP_JOBStudent                -2.839e-01  1.330e-01  -2.134  0.03286 *
IMP_JOBUnknown                -3.916e-01  1.962e-01  -1.995  0.04600 *
IMP_JOBz_Blue Collar          -5.656e-02  1.071e-01  -0.528  0.59735
Z_HOME_VAL1                   -3.514e-01  8.356e-02  -4.205 2.61e-05 ***
Z_CLM_FREQ1                    4.328e-01  6.543e-02   6.615 3.73e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9418.0  on 8160  degrees of freedom
Residual deviance: 7290.5  on 8128  degrees of freedom
AIC: 7356.5

Number of Fisher Scoring iterations: 5
> summary(newdata$predict4)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002313 0.077466 0.200982 0.263816 0.401619 0.971379
```

The predicted mean of model #4 is 0.263816, which is very close to the actual mean of TARGET_FLAG, so that's good. All of the predictors are significant at 90% level, so we have a good model here. Unlike OLS regression, coefficients in logistic regression have technical meanings that are difficult to interpret and use. Thus, instead of using the magnitude (or absolute value) of the coefficients, we will use the direction (positive vs. negative sign) of the coefficients to interpret the model. Below are the coefficient interpretations from the model.

- KIDSDRIV: moving from 0 (no kid driving) to 1 (teenage driver), the probability of getting a car crash increases. This makes sense since teenage drivers are more reckless.
- HOMEKIDS: moving from 0 (no kid at home) to 1 (have kids at home), the probability of car crash increases. This is an unknown effect.
- PARENT1: single parents have higher car crash probability, which makes sense since married people typically drive more safely.
- MSTATUS: singles have higher car crash probability than married, which makes sense
- EDUCATION: by default, R uses below high school as the base level for this variable. The only significant difference is between below high school and bachelors to predict car crash probability. None of the other levels make a difference, as indicated by high p-values in the regression result.
- TRAVTIME: the longer the distance to work, the higher the probability of car crash, which makes sense since there's more exposure.
- CAR_USE: commercial car usage has higher probability of crash than private usage, which makes sense since commercial vehicles are driven more so it has more exposure on the road.
- BLUEBOOK: the higher the value, the lower the probability of car crash. This is an unknown effect.
- TIF: people who have been customers for longer time have lower probability of crash, which makes sense since they're probably safer drivers.
- CAR_TYPE: by default, R uses minivan as the base level for this variable. Comparing to the other types, minivan has the lowest probability of crashes.
- REVOKED: drivers with license revoked in the past 7 years have higher probability of crashes.
- MVR_PTS: drivers with more traffic drivers have higher probability of crashes.
- URBANICITY: city drivers have higher probability of crashes, which makes sense since there's more traffic and cars in the urban settings.
- M_AGE: people with missing age values are more likely to have crashes. This is an interesting finding since we didn't expect this at the beginning of the project. We may have to speak to industry expert to understand better about the data collection process.
- IMP_INCOME: the higher the income, the less likely the crashes occur, which makes sense.
- IMP_JOB: by default, R uses clerical as the base level for this variable. Clerical has the highest probability of crashes comparing to other groups.
- Z_HOME_VAL: people with 0 in home value have higher probability of crashes. This makes sense since these people are probably renters and not home owners. In theory, home owners tend to drive more safely.

- Z_CLM_FREQ: people with 0 claim have lower probability of crashes, which makes sense.

From going through the direction/sign of each beta, the interpretation of each predictors makes sense regarding the probability of car crashes.

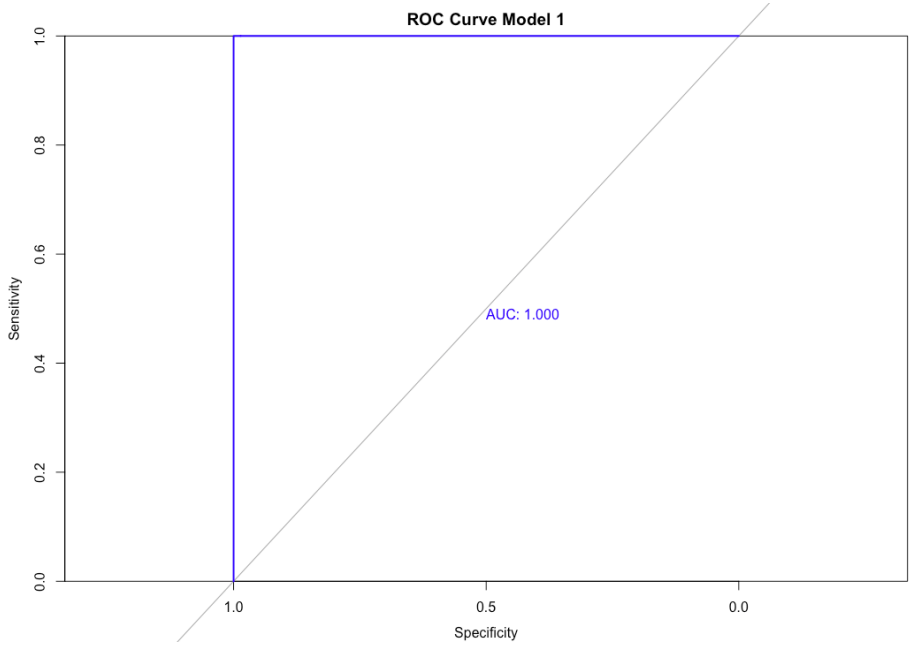
Section 4: Model Selection

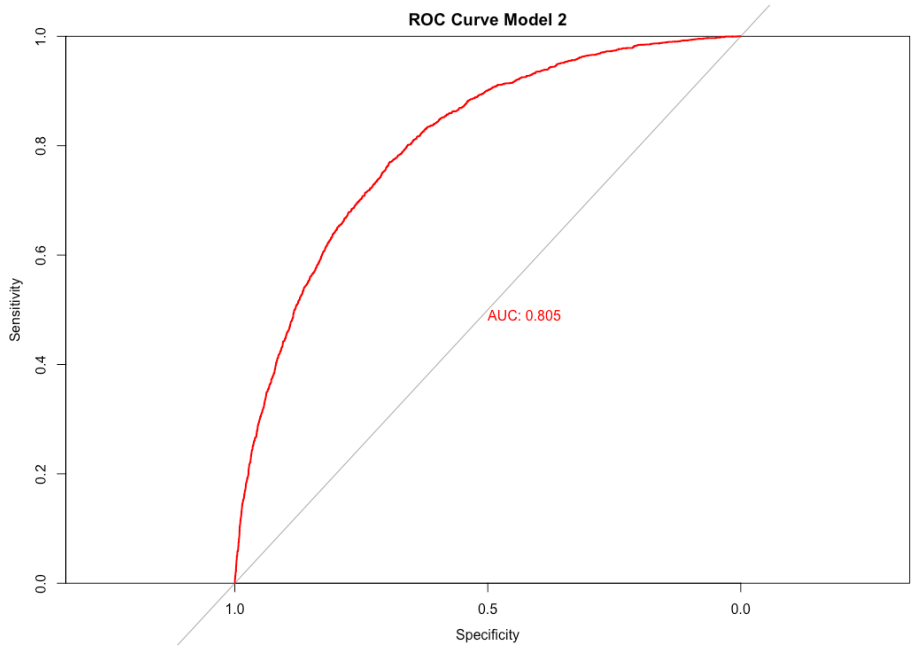
In order to compare the four models in section 3 and ultimate choose the best one, we will use the following metrics.

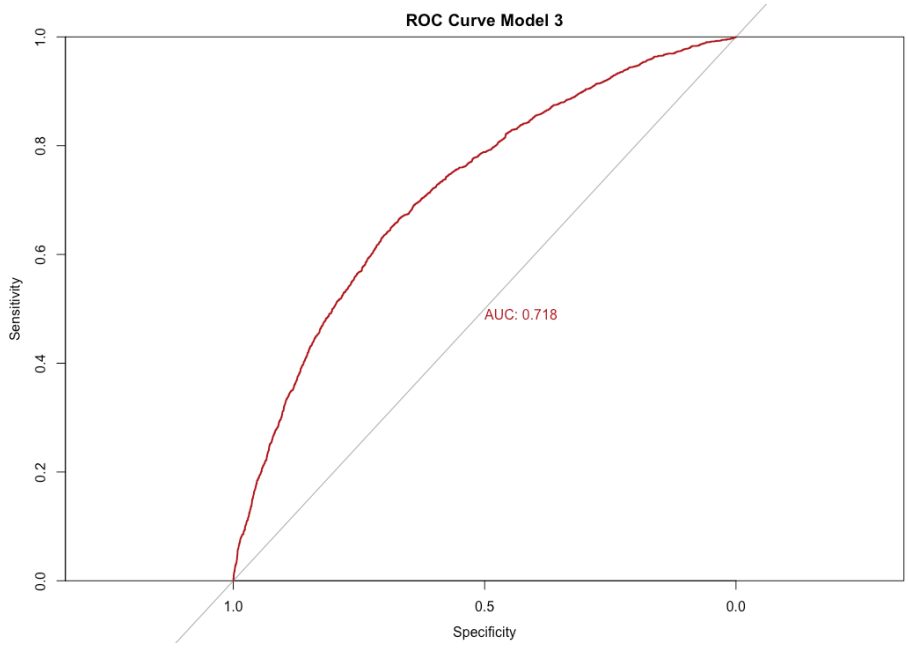
- AIC: the lower the metric, the better the model
- BIC: the lower the metric, the better the model
- Log likelihood: the higher the metric, the better the model
- KS statistic: the lower the metric, the better the model
- AUC (area under the curve) in ROC Curve: the higher the metric, the better the model with AUC = 1 representing a perfect model

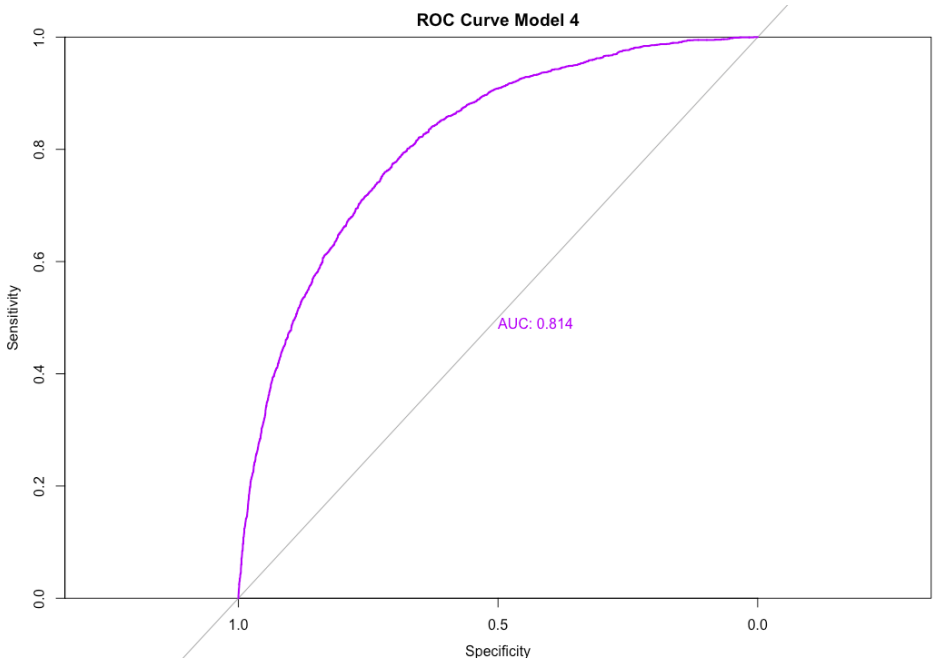
	model 1	model 2	model 3	model 4
AIC	100.00	7478.47	8489.81	7356.49
BIC	450.36	7632.63	8594.91	7587.72
LL	0.00	7434.47	8459.81	7290.49
KS	0.95	0.45	0.33	0.47

Based on the results above, using AIC and BIC, model #1 is the best model. Using log likelihood and KS statistic, model #3 is the best model.




ROC Curve Model 2
Sensitivity
Specificity
AUC: 0.805
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0


ROC Curve Model 3
Sensitivity
Specificity
AUC: 0.718
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0



Using AUC metric, model #1 is the perfect model. However, that's the problem with model #1: too perfect to be true. It has perfect score on the metric probably due to the fact that the model is overfitted. Therefore, we will remove model #1 from consideration.

If we only examine models #2, #3, and #4, using AIC and BIC, model #4 is the best. Using log likelihood and KS statistic, model #3 is the best. Using AUC metric, model #4 is the best. If we purely look at statistics using a quantitative approach, model #4 is ranked the highest in 3 out of 5 metrics, so that's the best one. If we use qualitative approach, model #4 is a combined model with the significant variables of models #2 and #3, so it's the best of the best. As a result, model #4 is the chosen model for this project to predict TARGET_FLAG using logistic regression analysis.

Section 5: OLS Regression Model Development

In this section, we will develop an OLS regression model using stepwise variable selection method to predict TARGET_AMT. After removing insignificant variables, we have the following result.

```
Call:
lm(formula = newdata$TARGET_AMT ~ newdata$TRAVTIME + newdata$BLUEBOOK +
    newdata$TIF + newdata$CLM_FREQ + newdata$MVR_PTS + newdata$IMP_AGE +
    newdata$IMP_INCOME + newdata$IMP_HOME_VAL)

Residuals:
   Min     1Q Median     3Q    Max
 -4204  -1578   -986   -154 104368

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.715e+03  3.145e+02   5.454 5.08e-08 ***
newdata$TRAVTIME      7.609e+00  3.300e+00   2.306   0.0211 *
newdata$BLUEBOOK      1.498e-02  7.237e-03   2.070   0.0385 *
newdata$TIF          -4.998e+01  1.343e+01  -3.721   0.0002 ***
newdata$CLM_FREQ      2.933e+02  4.847e+01   6.051 1.50e-09 ***
newdata$MVR_PTS       2.229e+02  2.883e+01   7.731 1.20e-14 ***
newdata$IMP_AGE      -1.202e+01  6.137e+00  -1.959   0.0502 .
newdata$IMP_INCOME   -2.385e-03  1.588e-03  -1.502   0.1331
newdata$IMP_HOME_VAL -2.271e-03  4.994e-04  -4.547 5.53e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4635 on 8152 degrees of freedom
Multiple R-squared:  0.02994,   Adjusted R-squared:  0.02899
F-statistic: 31.46 on 8 and 8152 DF,  p-value: < 2.2e-16
```

The overall model has p-value less than 0.05 alpha, so it’s statistically significant at 95% confidence level. Using the coefficients above, we have the following formula to predict TARGET_AMT.

```
formula=1.715e+03+
  newdata$TRAVTIME*7.609e+00+
  newdata$BLUEBOOK*1.498e-02+
  newdata$TIF*-4.998e+01+
  newdata$CLM_FREQ*2.933e+02+
  newdata$MVR_PTS*2.229e+02+
  newdata$IMP_AGE*-1.202e+01+
  newdata$IMP_INCOME*-2.385e-03+
  newdata$IMP_HOME_VAL*-2.271e-03
```

Applying this formula to the train dataset to forecast TARGET_AMT, we have the following result.

```
 summary(formula)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-747.2   913.0  1386.5  1504.0  2032.5  4225.1
```

The predicted mean is 1504.0, which is very close to the actual mean, so that’s good.

CONCLUSION

In conclusion, the insurance project starts with a train dataset of 8161 observations and 26 variables. Among them, there are two response variables: TARGET_FLAG (used to predict the probability of car crashes) and TARGET_AMT (used to predict the cost of the crash). There are five stages project goes through 1) data exploration to understand the data via visuals such as histograms, boxplots, and correlation plot as well as to identify variables with missing value and outlier issues 2) data preparation to address missing value and outlier issues. During this stage, three types of new variables are created: M to show the distinction between missing and known values, IMP to indicate imputed variables for missing values, and Z to show the distinction between records with 0 value and other values 3) logistic regression model development to build four models to forecast TARGET_FLAG variable 4) model selection to compare these four models using different metrics such as AIC, BIC, log likelihood, KS statistic, and AUC under ROC curve. From this comparison, model #4 is selected 5) OLS regression model development to build an OLS regression model to forecast TARGET_AMT.

The next step for this project is to build a stand alone data step that can apply the result in model #4 from section 3 logistic regression model development and OLS model in section 5 OLS regression model development to new datasets. If the predicted mean of the first model is approximately 26% and the predicted mean of the second model is approximately 1500, we conclude that both models are solid since these numbers are the average of the actual TARGET_FLAG and TARGET_AMT in the train dataset.

Beside testing the two models on new dataset using the stand alone data step, future researchers should consider adding more variables to increase its accuracy. However, if that's the case, precision may suffer, so future researchers need to consider the balance between accuracy and precision. Perhaps researchers can utilize tools such as AIC and BIC to determine the right mix of accuracy and precision in future models.