Home Work Assignment - 04

Critical Thinking Group 5

Contents

newpage

1 Overview

The data set contains approximately 8161 records. Each record represents a customer profile at an auto insurance company. Each record has two response variables.

The first response variable, TARGET_FLAG, is a 1 or a 0. A "1" means that the person was in a car crash. A zero means that the person was not in a car crash.

The second response variable is TARGET_AMT. This is the amount spent on repairs if there was a crash. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

We will be exploring, analyzing, and modeling the training data. Since there are 2 different predictions we have to work with, we will deal with each prediction independently. The following are the 2 predictions we will be modeling for:

1. TARGET_FLAG - This dependent variable tells whether there was a crash or not. This is a binary variable and as such we will be using a Logistic Regression Model to predict this.
2. TARGET_AMT - This dependent variable gives the amount / cost of repairs if there was a crash. This is a continuous variable and we will ve using a Linear Regression Model to predict this.

Each of the above models will be built and evaluated separately. In the first section of this document we will deal with the Logistic Model for TARGET_FLAG and in the second section we will deal with Linear Model for the TARGET_AMT

Out of the many models for each task, we will go ahead and shortlist one model that works the best. We will then use these models (one for each task) on the test / evaluation data.

To attain our objective, we will be follow the below steps for each modeling exercise:

1 -Data Exploration 2 -Data Preparation 3 -Build Models 4 -Select Models

Model Selection Strategy: As a strategy, we will split the train dataset into 2 parts - TRAIN and VALID. In the VALID dataset, we will hold out some values to validate how well the model is trained using the TRAIN dataset. We will then use the Model that performs the best on the EVALUATION data to give the required output. We will split the TRAIN / VALID data after the **Data Exploration / Preparation** before the **Build Models**.

Please Note:

- There are some common clean-up and transformations that we will carry out initially that will serve all the models.
- While working on the Linear Models for the TARGET_AMT, we will be using only a subset of the data where the TARGET_FLAG = 1. This will give us all the records where there was a crash and subsequently a repair amount.
- While Predicting the TARGET_AMT with the given Evaluation dataset, We will take the output of the TARGET_FLAG predictions on the Evaluation dataset and use only those rows that were classified as a "Crash" and use it as the input to the TARGET_AMT prediction. So this is a two step prediction, one for the TARGET_FLAG and using the output to predict TARGET_AMT.

newpage

2 Data Exploration and Cleanup / Common Transformations

In this section we go ahead and perform some common cleanup and create additional variables that will be used for modeling both the logistic as well as the linear regressions. We will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:

- Variable Identification / Relationships
- Data Clean-up
- Common Transformations
- Create Missing Flags / Impute Missing Values

2.1 Variable Identification

First let's display and examine the data dictionary or the data columns as shown in below table:

Table 1: Variable Description

VARIABLE_NAME	DEFINITION	THEORETICAL_EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old peopl
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably e
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably e
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably e
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase pr
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are li
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend
HOMEKIDS	# Children at 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
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
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probal
SEX	Gender	Urban legend says that women have less crashes then men
TIF	Time in Force	People who have been customers for a long time are usual
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

We notice that there are 2 dependent variables - TARGET_FLAG and TARGET_AMT. Apart from these 2 dependent variables, we have 23 independent or predictor variables.

2.2 Data Cleanup

[1] "Yes" "z_No" [1] "M" "z_F" [1] "

Now that we are done with the common clean-up and transformations, we can proceed to each specific model as below.

newpage

3 Logistic Regression for TARGET_FLAG

In this section we will use Logistic regression to model the TARGET_FLAG. We will first start with the Data Exploration.

3.1 Data Summary and Correlation Analysis

3.1.1 Data Summary

In this section, we will create summary data to better understand the relationship each of the variables have with our dependent variables using correlation, central tendency, and dispersion as shown below:

3.1.2 Correlations

Now we will produce the correlation table between the independent variables and the dependent variable - TARGET_FLAG

Table 2: Correlation between TARGET_FLAG and predictor variables

	Correlation_TARGET_FLAG
TARGET_FLAG	1.0000000
TARGET_AMT	0.5343138
MVR_PTS	0.2192671
CLM_FREQ	0.2159652
PARENT1_Yes	0.1576594
REVOKED_Yes	0.1517045
CAR_USE_Commercial	0.1427163
EDUCATION_High.School	0.1382094
OLDCLAIM	0.1378435
HOMEKIDS	0.1161499
KIDSDRIV	0.1040583
JOB_Blue.Collar	0.1018097
JOB_Student	0.0770293
CAR_TYPE_Sports.Car	0.0572627
CAR_TYPE_Pickup	0.0563353
TRAVTIME	0.0480461
CAR_TYPE_SUV	0.0450376
JOB_Clerical	0.0275791
JOB_Home.Maker	0.0112577
CAR_AGE_MISS	0.0085607
YOJ_MISS	0.0039126
CAR_TYPE_Van	0.0030163
CAR_TYPE_Panel.Truck	-0.0003471
HOME_VAL_MISS	-0.0016978
JOB_Unknown	-0.0031380
RED_CAR_yes	-0.0069595
INCOME_MISS	-0.0090653
SEX_M	-0.0206620
JOB_Professional	-0.0391996
EDUCATION_Bachelors	-0.0431408
JOB_Doctor	-0.0580794
JOB_Lawyer	-0.0617528
EDUCATION_PhD	-0.0652170
YOJ	-0.0684748
EDUCATION_Masters	-0.0761613
TIF	-0.0821748
CAR_AGE	-0.0974530

	Correlation_TARGET_FLAG
AGE	-0.1032152
BLUEBOOK	-0.1035337
JOB_Manager	-0.1052506
MSTATUS_Yes	-0.1347552
CAR_TYPE_Minivan	-0.1367604
INCOME	-0.1377852
HOME_VAL	-0.1785848
URBANICITY_Rural	-0.2241940

The above table suggests that none of the variables seem to have a very strong correlation with TARGET_FLAG. However, CAR_TYPE_Van, RED_CAR_no, JOB_Home.Maker, SEX_F, JOB_Clerical, CAR_TYPE_SUV, TRAVTIME, CAR_TYPE_Pickup, CAR_TYPE_Sports.Car, JOB_Student, JOB_Blue.Collar, KIDSDRIV, HOMEKIDS, MSTATUS_No, OLDCLAIM, EDUCATION_High.School, CAR_USE_Commercial, REVOKED_Yes, PARENT1_Yes, CLM_FREQ, MVR_PTS, URBANICITY_Highly.Urban..Urban have a positive correlation.

Similarly, URBANICITY_Highly.Rural..Rural, HOME_VAL, PARENT1_No, REVOKED_No, CAR_USE_Private, INCOME, CAR_TYPE_Minivan, MSTATUS_Yes, JOB_Manager, BLUEBOOK, AGE, CAR_AGE, TIF, EDUCATION_Masters, YOJ, EDUCATION_PhD, JOB_Lawyer, JOB_Doctor, EDUCATION_Bachelors, JOB_Professional, SEX_M, RED_CAR_yes, CAR_TYPE_Panel.Truck have a negative correlation.

Lets now see how values in some of the variable affects the correlation:

CAR_TYPE - If you drive Minivans and Panel Trucks you have lesser chance of being in a crash as against Pickups, Sports, SUVs and Vans. Since the distiction is clear, we believe that binning this variable accordingly will help strengthen the correlation.

EDUCATION - If you have only a high school education then you are more likely to crash than if you have a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation.

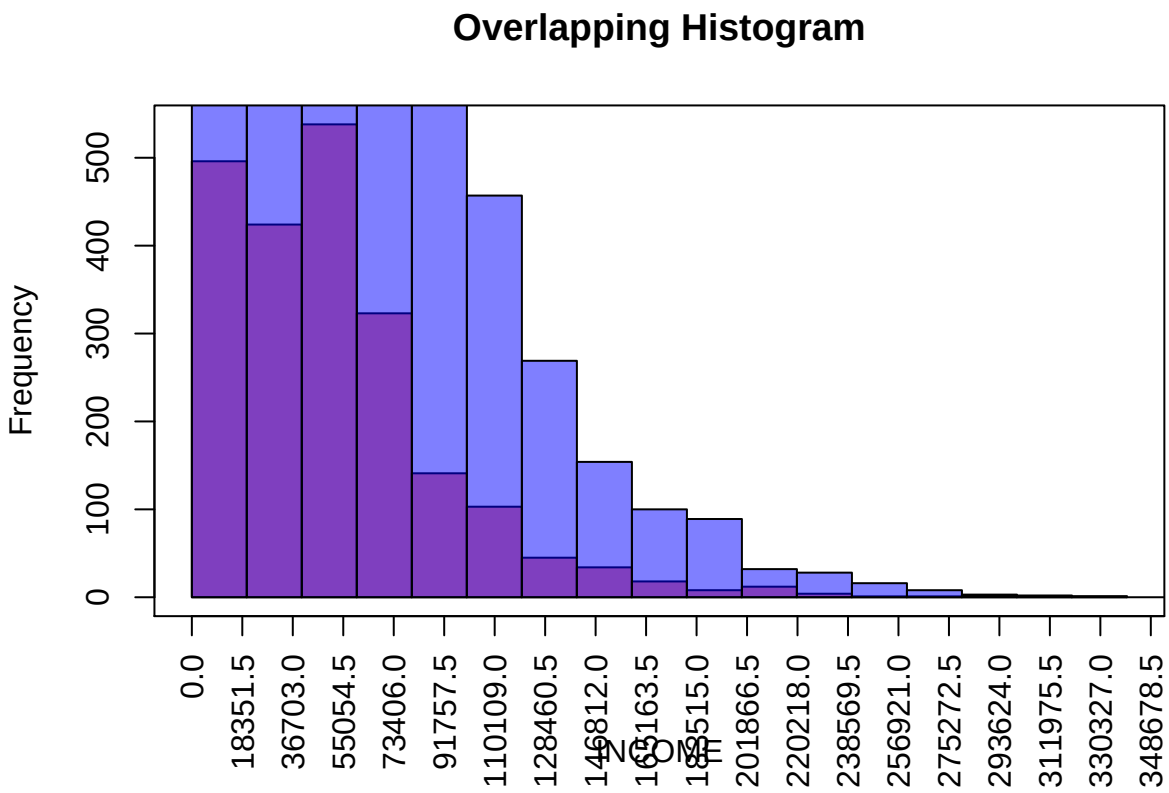
JOB - If you are a Student, Homemaker, or in a Blue Collar or Clerical job, you are more likely to be in a crash against Doctor, Lawyer, Manager, professional or Unknown job. Again binning this variable will strengthen the correlation.

3.1.3 Binning of Variables

Lets have a look at the following numeric variables to see how they are distributed vis-a-vis TARGET_FLAG: INCOME, YOJ, HOME_VAL, OLDCLAIM, CLM_FREQ, MVR_PTS, CAR_AGE, AGE, BLUEBOOK, TIF, TRAVTIME. The goal here is to see if we can bin these variables into zero and non-zero bin values and check the correlations. While doing that we will also see how the variables are distributed vis-a-vis TARGET_FLAG.

```
show_hist <- function(var) {

    col_x <- which(colnames(insure_train_full)==var)
    h0 <- select(insure_train_full[insure_train_full$TARGET_FLAG==1,], col_x)
    h1 <- select(insure_train_full[insure_train_full$TARGET_FLAG==0,], col_x)

    min_x <- min(select(insure_train_full, col_x), na.rm = TRUE)
    max_x <- max(select(insure_train_full, col_x), na.rm = TRUE)
    by_x <- (max_x - min_x) / 20
```

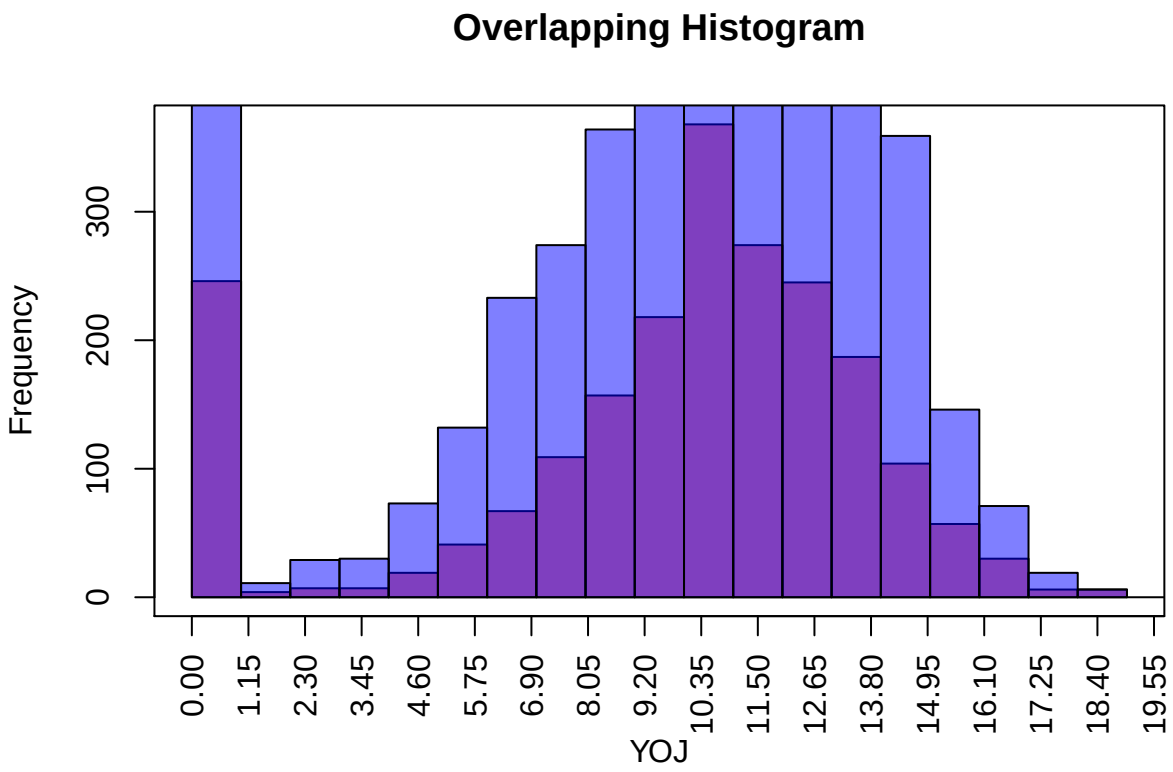
```
    hist(h0[,1], breaks = 20, col=rgb(1,0,0,0.5), main="Overlapping Histogram", xlab = var, xaxt = "n")
    axis(1, at = seq(min_x, max_x, by = by_x), las=2)

    hist(h1[,1], breaks = 20, col=rgb(0,0,1,0.5), add=T) #
#    axis(1, at = seq(min_x, max_x, by = by_x), las=2)

    box()
}

check_bins <- function(var, thresholds) {

    col_x <- which(colnames(insure_train_full)==var)
    old_x <- select(insure_train_full, col_x)
    cor_old <- cor(old_x, insure_train_full$TARGET_FLAG,use = 'na.or.complete')
    ds <- data.frame("Item" = "Original", "Correlation"= round(cor_old, 5))


    for(i in 1:length(thresholds)) {
        New_x <- ifelse(select(insure_train_full, col_x)<=thresholds[i],0,1)
        cor_new <- cor(New_x, insure_train_full$TARGET_FLAG,use = 'na.or.complete')
        ds_1 <- data.frame("Item" = as.character(thresholds[i]), "Correlation"= round(cor_new, 5))
        ds <- rbind(ds, ds_1)
    }
    return (ds)
}

show_hist("INCOME")
```




```
check_bins("INCOME", c(0, 20000, 90000, 130000))
```

```
##             Item Correlation
## INCOME  Original    -0.13779
## INCOME1        0    -0.09245
## INCOME2    20000    -0.09167
## INCOME3    90000    -0.12081
## INCOME4   130000    -0.07364
```

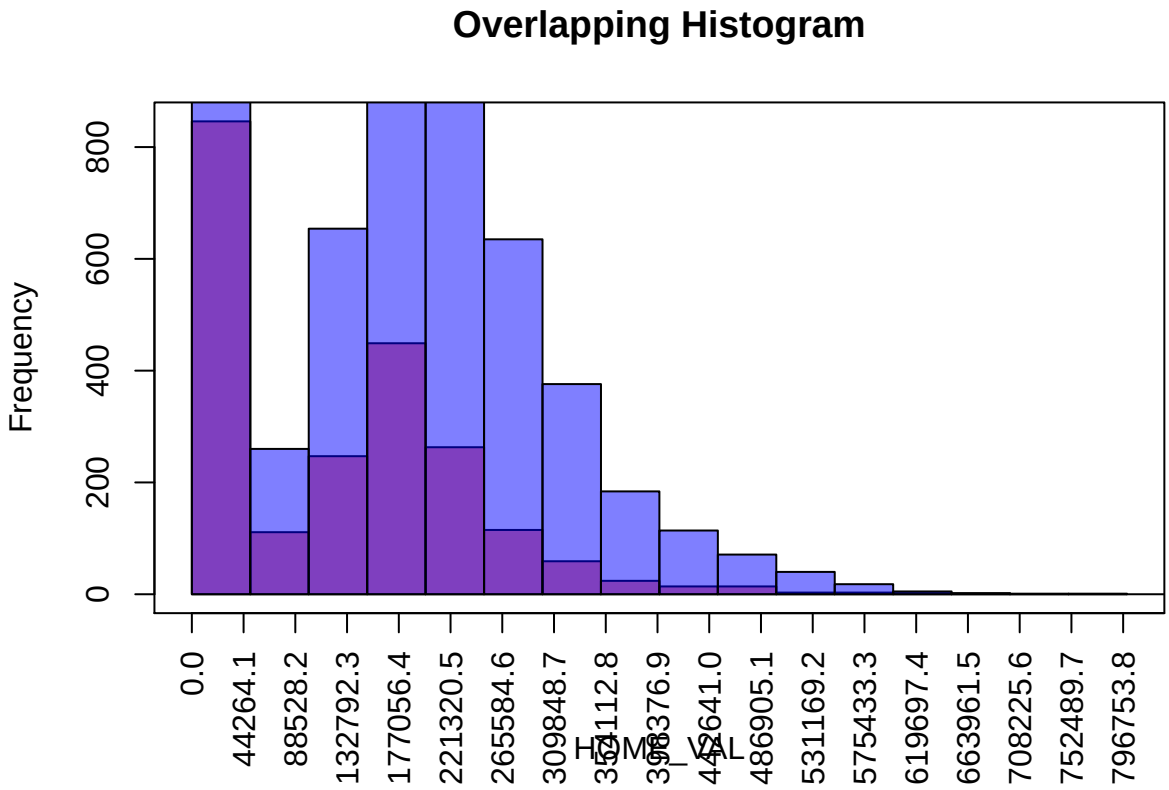
```
show_hist("YOJ")
```




```
check_bins("YOJ", c(0, 4, 8, 15))
```

```
##          Item Correlation
## YOJ  Original    -0.06847
## YOJ1        0    -0.08483
## YOJ2        4    -0.07292
## YOJ3        8    -0.04154
## YOJ4       15     0.01179
```

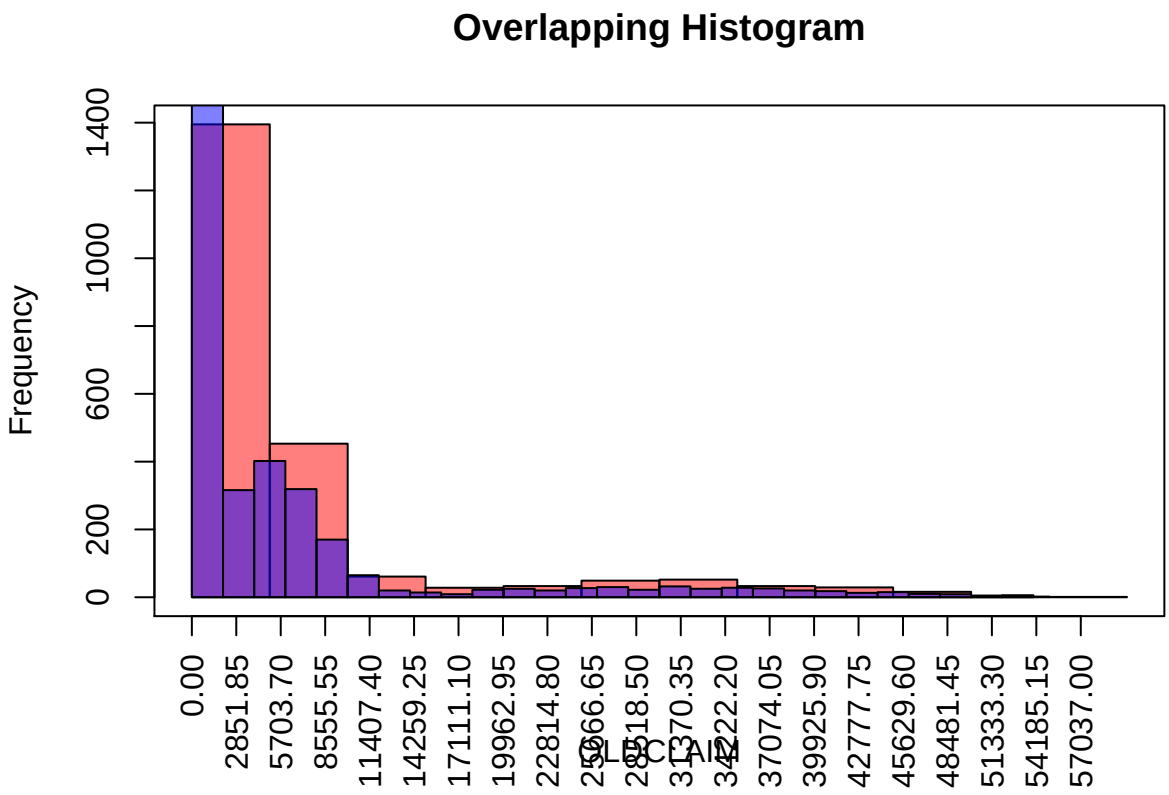
```
show_hist("HOME_VAL")
```




```
check_bins("HOME_VAL", c(0, 20000, 90000, 130000))
```

```
##               Item Correlation
## HOME_VAL  Original    -0.17858
## HOME_VAL1        0    -0.14898
## HOME_VAL2    20000    -0.14898
## HOME_VAL3    90000    -0.15142
## HOME_VAL4   130000    -0.15006
```

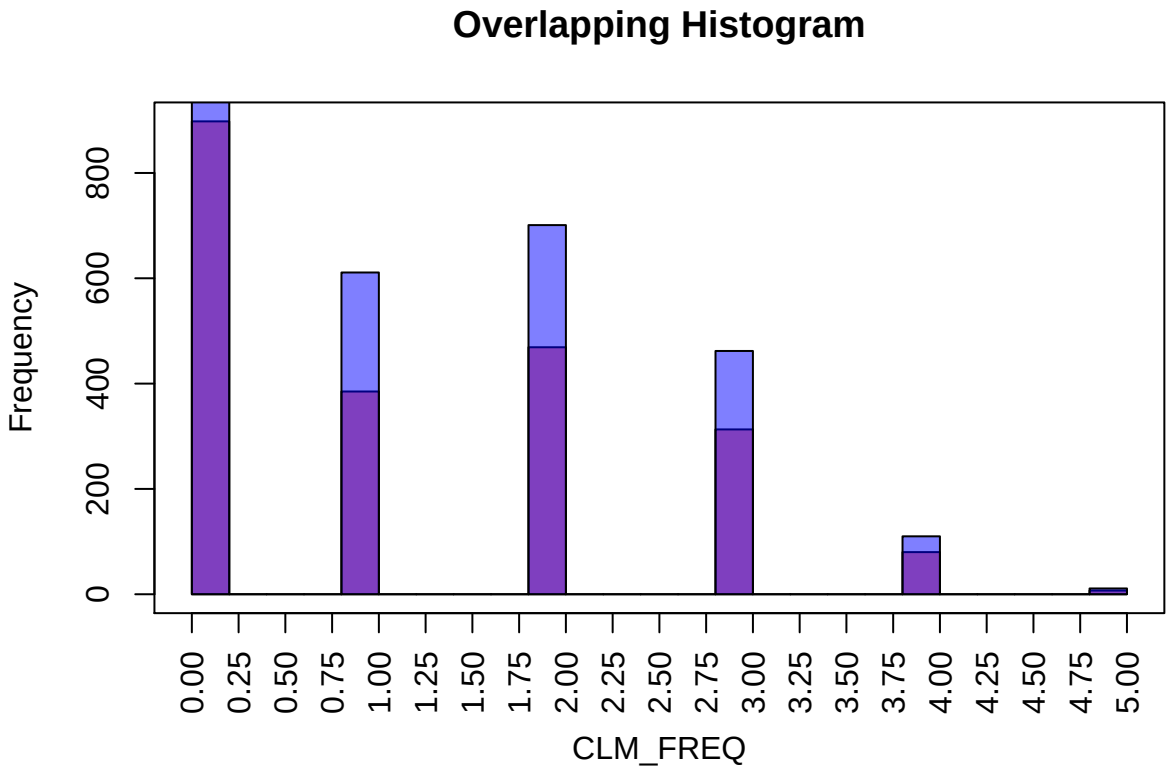
```
show_hist("OLDCLAIM")
```




```
check_bins("OLDCLAIM", c(0, 5000, 10000, 15000, 20000, 40000))
```

```
##                Item Correlation
## OLDCLAIM  Original     0.13784
## OLDCLAIM1        0     0.24183
## OLDCLAIM2     5000     0.16565
## OLDCLAIM3    10000     0.09750
## OLDCLAIM4    15000     0.08610
## OLDCLAIM5    20000     0.07898
## OLDCLAIM6    40000     0.03475
```

```
show_hist("CLM_FREQ")
```



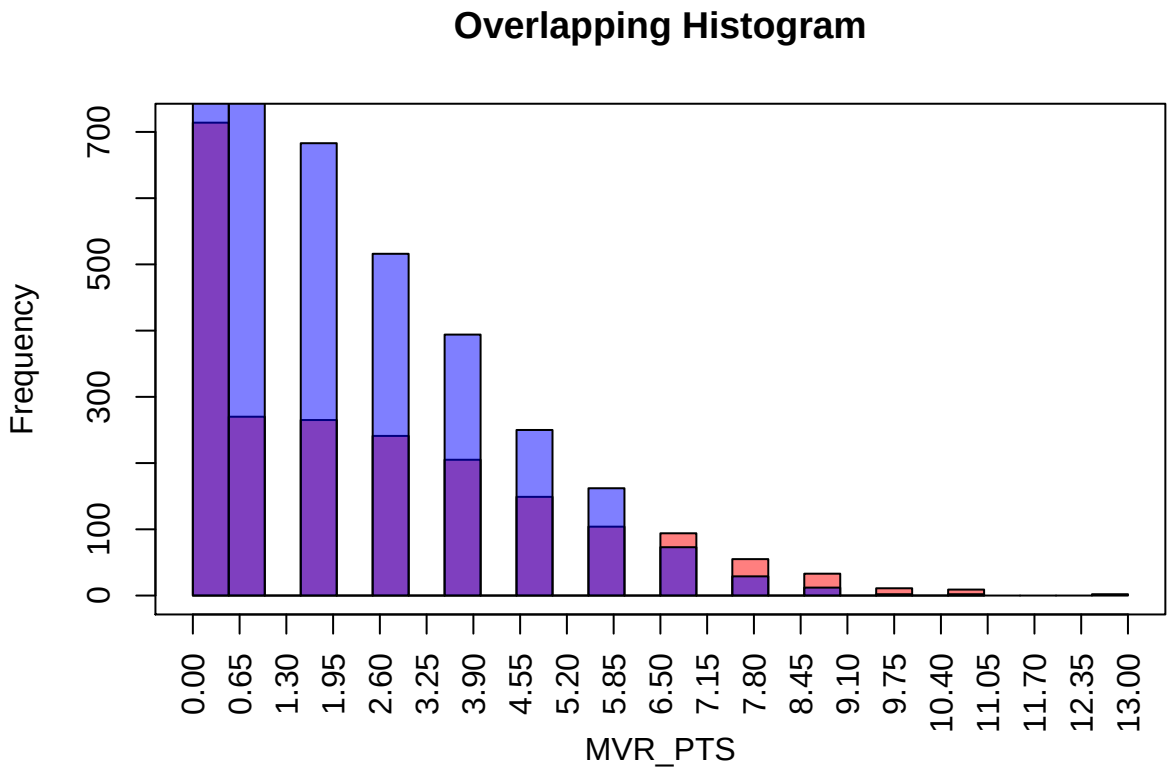

```
check_bins("CLM_FREQ", c(0, 1, 2, 3, 4))
```

```
##               Item Correlation
## CLM_FREQ  Original     0.21597
## CLM_FREQ1        0     0.24183
## CLM_FREQ2        1     0.18996
## CLM_FREQ3        2     0.12019
## CLM_FREQ4        3     0.05669
## CLM_FREQ5        4     0.01335
```

```
table(insure_train_full$MVR_PTS)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   13
## 3710 1156  948  757  599  399  266  167   84   45   13   11    2
```

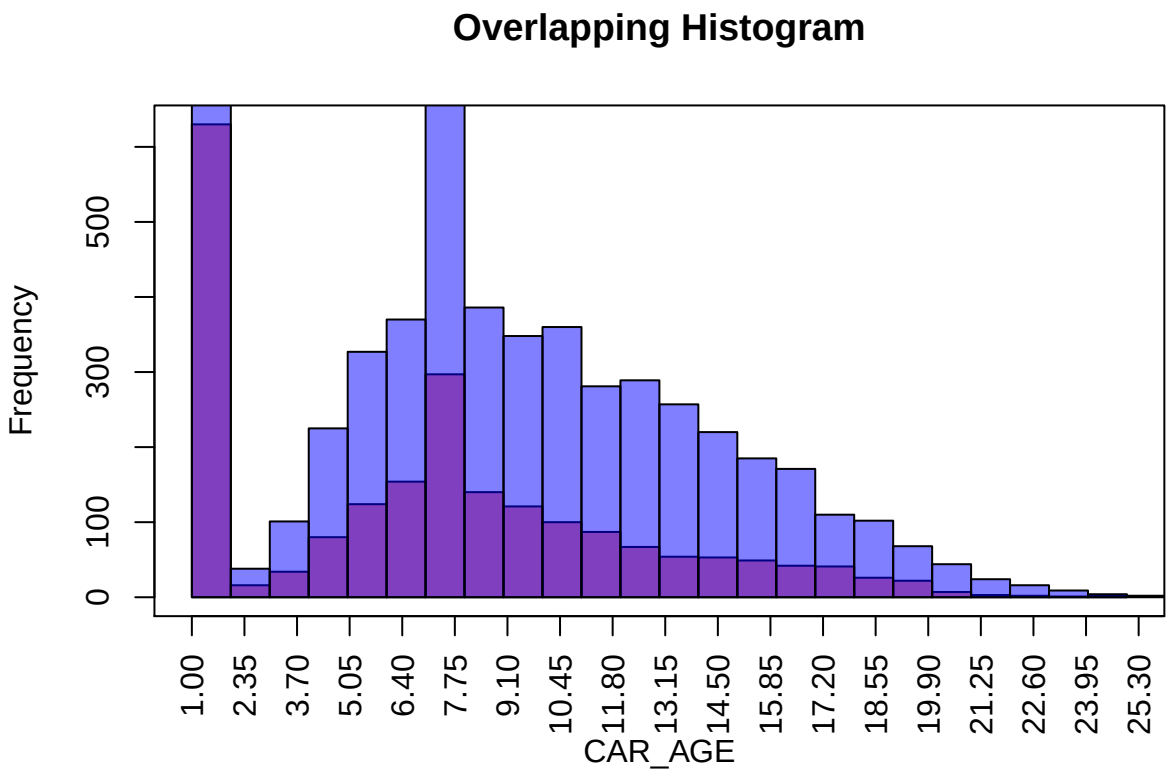
```
show_hist("MVR_PTS")
```



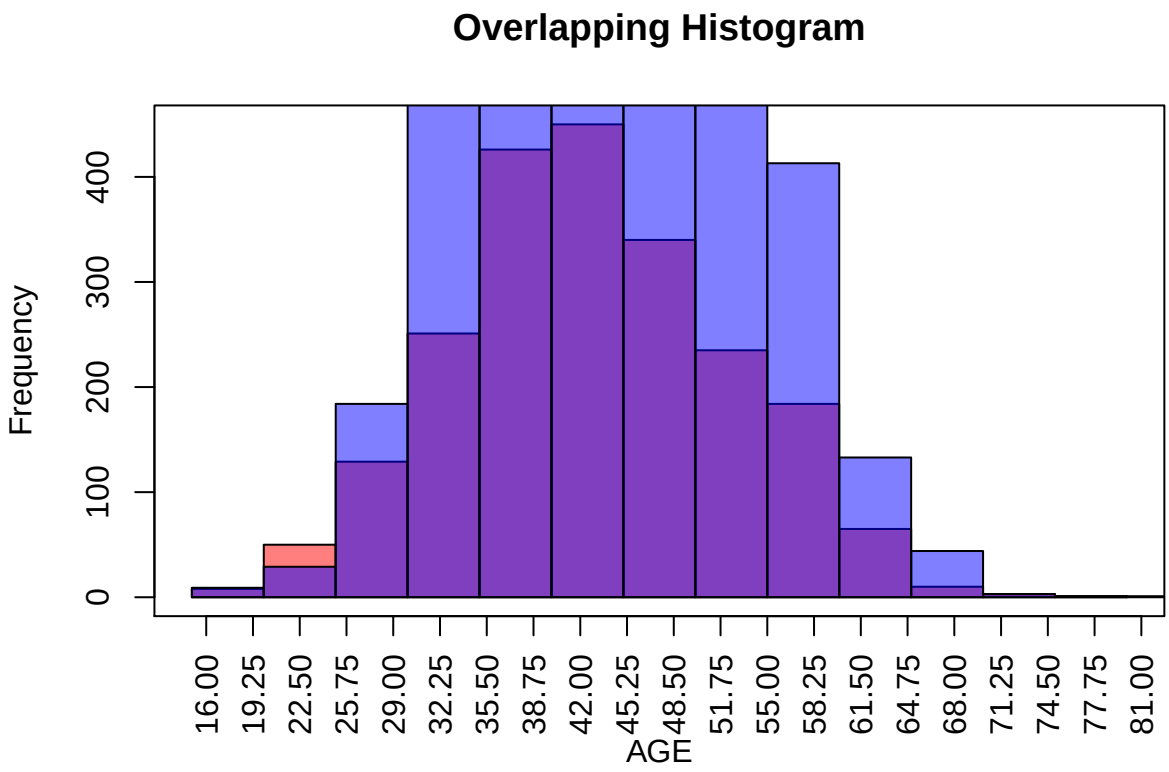

```
check_bins("MVR_PTS", c(0:12))
```

```
##                Item Correlation
## MVR_PTS    Original     0.21927
## MVR_PTS1          0     0.14792
## MVR_PTS2          1     0.16997
## MVR_PTS3          2     0.17513
## MVR_PTS4          3     0.17121
## MVR_PTS5          4     0.16770
## MVR_PTS6          5     0.16443
## MVR_PTS7          6     0.17007
## MVR_PTS8          7     0.14080
## MVR_PTS9          8     0.10861
## MVR_PTS10         9     0.07472
## MVR_PTS11        10     0.05279
## MVR_PTS12        11     0.02616
## MVR_PTS13        12     0.02616
```

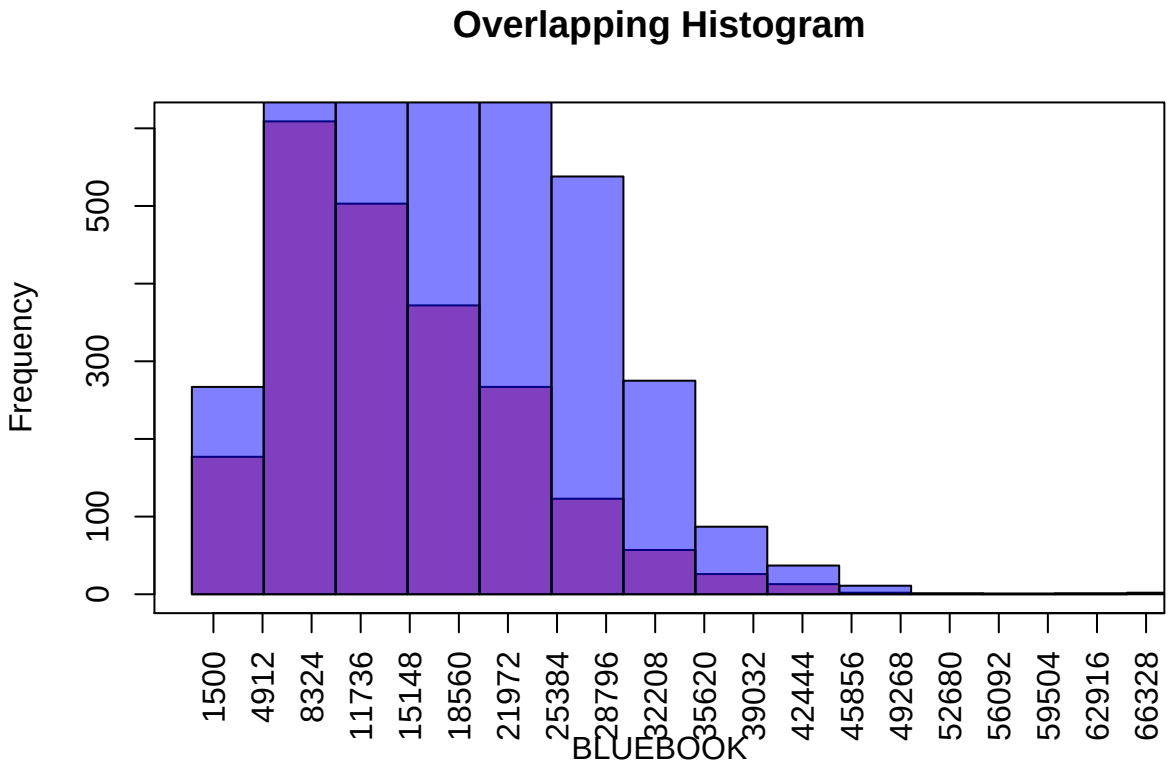
```
#table(insure_train_full$CAR_AGE)
show_hist("CAR_AGE")
```




```
#check_bins("CAR_AGE", c(1:27))

#table(insure_train_full$AGE)
show_hist("AGE")
```



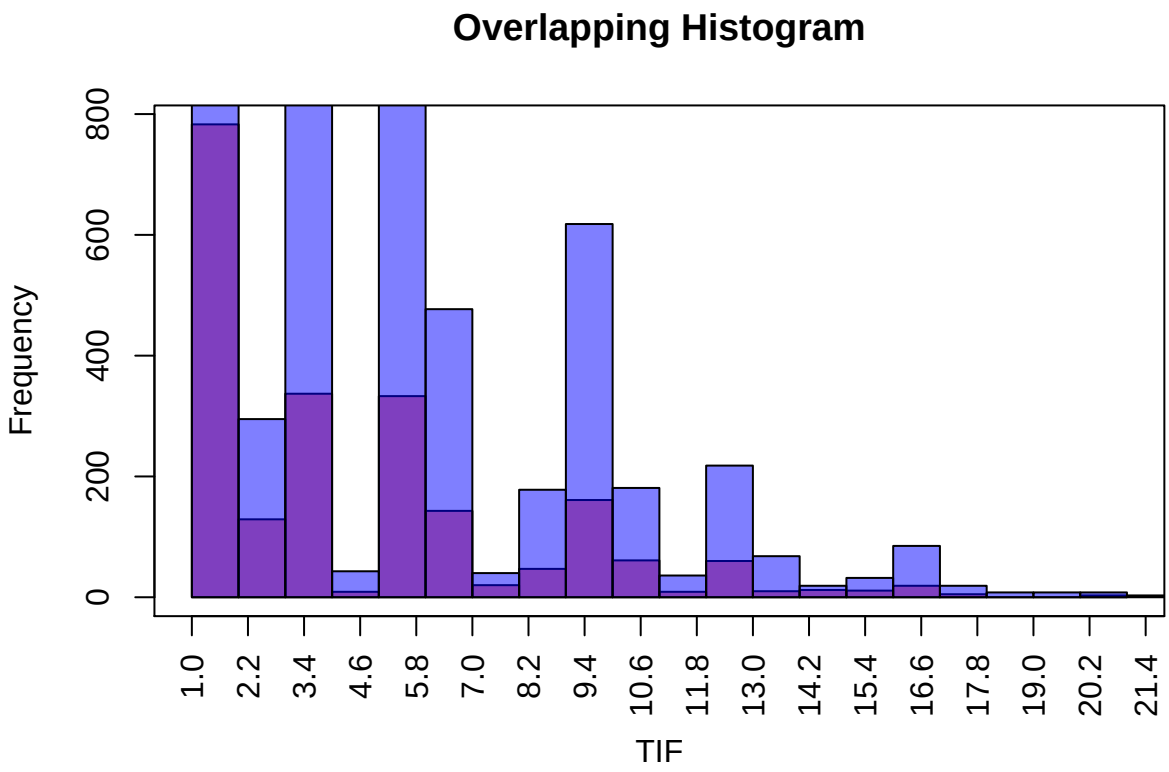

```
#check_bins("AGE", c(16:80))

#table(insure_train_full$BLUEBOOK)
show_hist("BLUEBOOK")
```




```
#check_bins("BLUEBOOK", c(11000, 41000, 41050, 57500, 58000))

table(insure_train_full$TIF)
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 2532    6  424 1241   52 1341  620   60  225  779  242   45  278   78   31
##   16   17   18   19   20   21   22   25
##   43  104   24    8    8   11    3    2
```

```
show_hist("TIF")
```



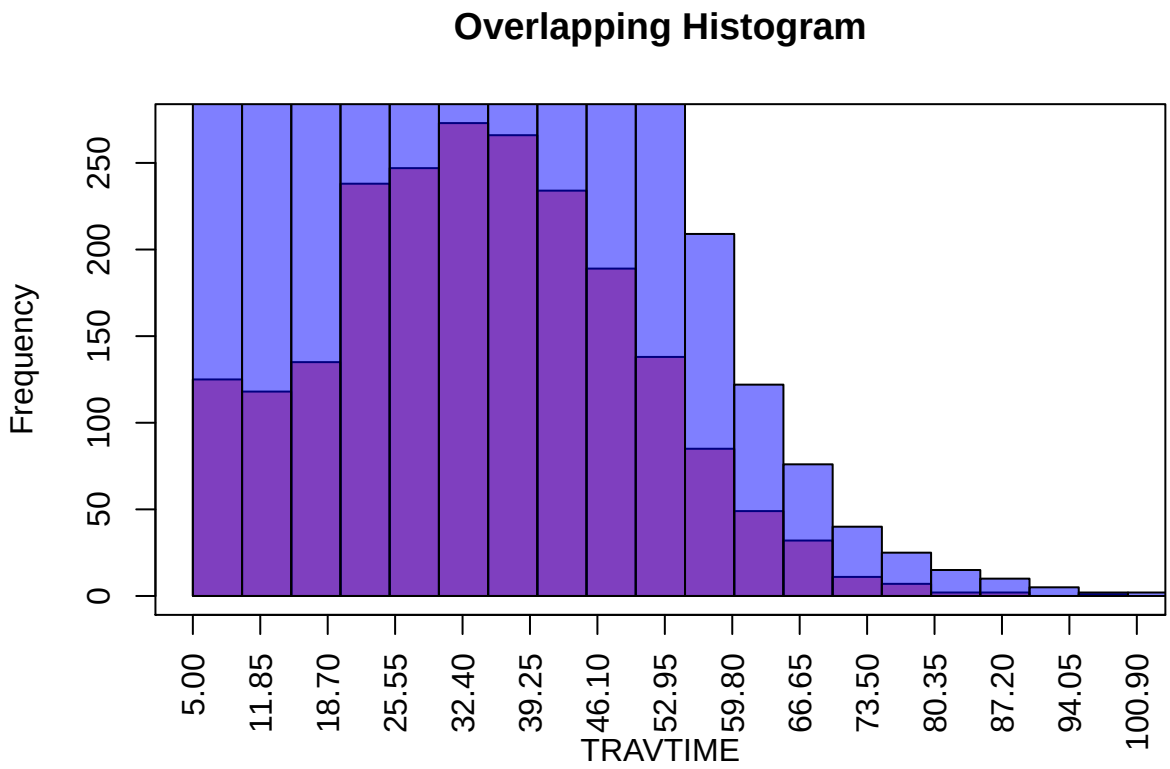

```
check_bins("TIF", c(1, 4, 6, 10, 24))
```

```
##          Item Correlation
## TIF  Original    -0.08217
## TIF1        1    -0.06734
## TIF2        4    -0.07801
## TIF3        6    -0.06872
## TIF4       10    -0.03715
## TIF5       24    -0.00937
```

```
table(insure_train_full$TRAVTIME)
```

```
##
##   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22
## 334  49  43  54  70  87  70  97  97 102  92 117 117 121 135 131 169 164
##  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40
## 157 204 212 181 173 188 207 219 190 214 206 201 219 211 202 186 175 196
##  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
## 175 154 153 162 167 145 125 128 125 106 102 110  82  62  72  73  71  56
##  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76
##  50  44  46  37  25  37  26  30  19  20  21  18  15   8   8  10  10   4
##  77  78  79  80  81  82  83  84  85  86  87  88  90  91  92  93  95  97
##   8   8   5   7   5   4   3   3   2   2   4   3   3   1   1   1   2   2
##  98 101 103 113 124 134 142
##   1   1   1   1   1   1   1
```

```
show_hist("TRAVTIME")
```




```
check_bins("TRAVTIME", c(21, 59, 120))
```

```
##                Item Correlation
## TRAVTIME   Original     0.04805
## TRAVTIME1        21     0.05432
## TRAVTIME2        59     0.00188
## TRAVTIME3       120    -0.01148
```

From the outputs above, we can come to the following conclusions:

- INCOME - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this at zero value.
- YOJ - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.
- HOME_VAL - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.
- OLDCLAIM- There is a huge difference in the coorrelation when we transform this vatiable. Binning this variable seems like a good idea.
- CLM_FREQ - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

- MVR_PTS - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.
- CAR_AGE - There are quite a few records with a 1 year car age. We will use this bound to generate a binned variable as well as retain the original varible as is.
- AGE - There is no specific pattern that emerges. We will retain this variable as is.

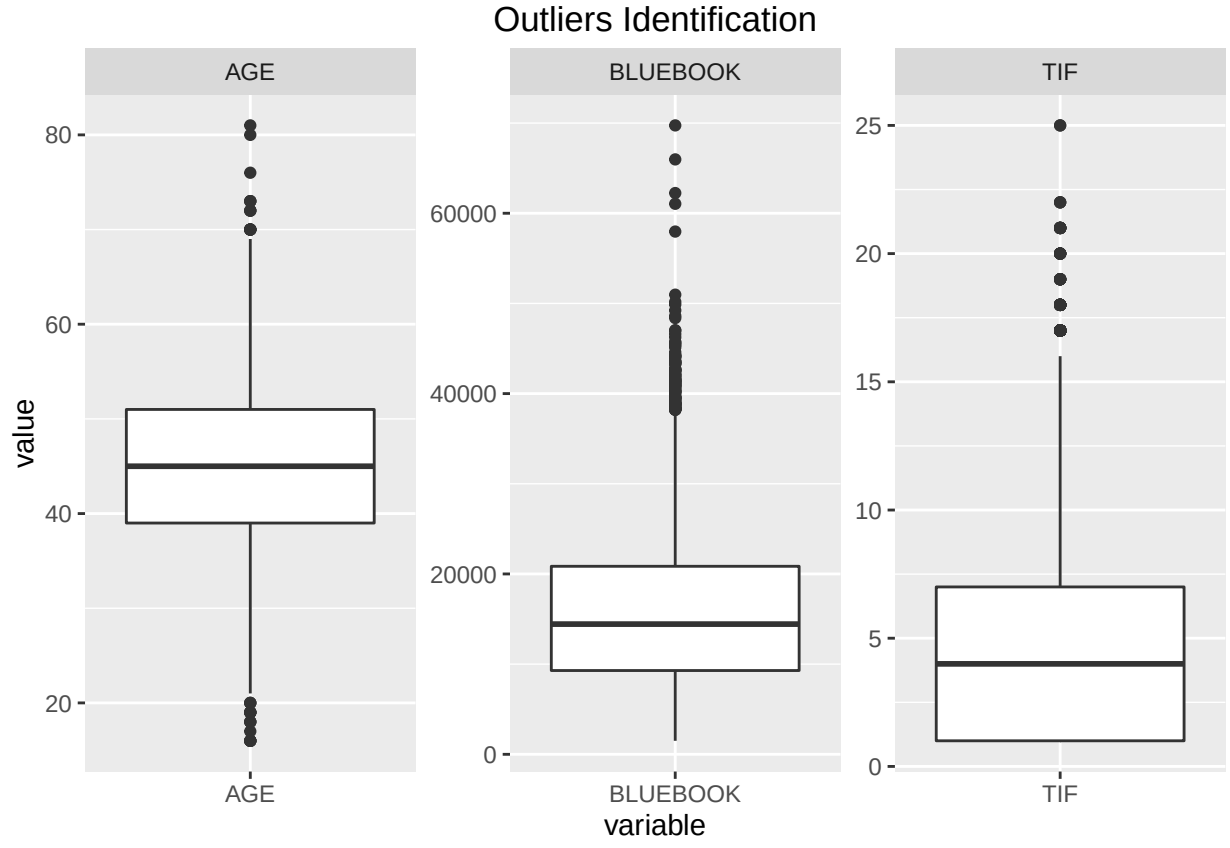
-BLUEBOOK - There is no specific pattern that emerges. We will retain the variable as is.

- TIF - Looking at the plots, values and the correlations with TARGET_FLAG, we can conclude that this is not a good variable for binning. We will retain this variable as is.
- TRAVTIME - from the plot, we can see that there is a clear pattern around the value - 20. We will go ahead and create a binned variable for this.

We will carry out the above transformations in the Data Preparation phase.

3.1.4 Outliers identification

In this sub-section, we will look at the boxplots and determine the outliers in variables and decide on whether to act on the outliers. We will do the outliers only on some of the currency and few other variables. Below are the plots:



From the "Outliers identification" plot above, we see that we have few outliers that we need to treat. We will treat the outliers in this variable when we do the data preparation for modeling the TARGET_FLAG.

3.1.5 Analysis of the link function

In this section, we will investigate how our initial data aligns with a typical logistic model plot.

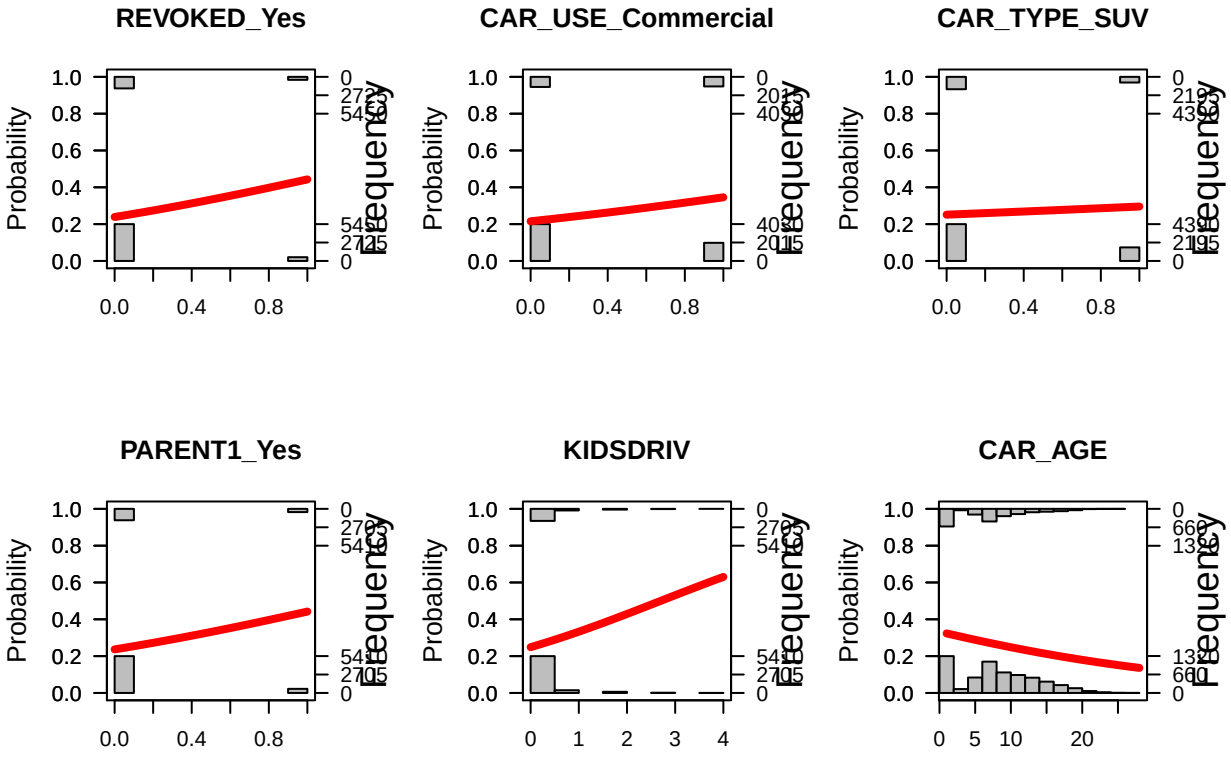
Recall the Logistic Regression is part of a larger class of algorithms known as Generalized Linear Model (glm). The fundamental equation of generalized linear model is:

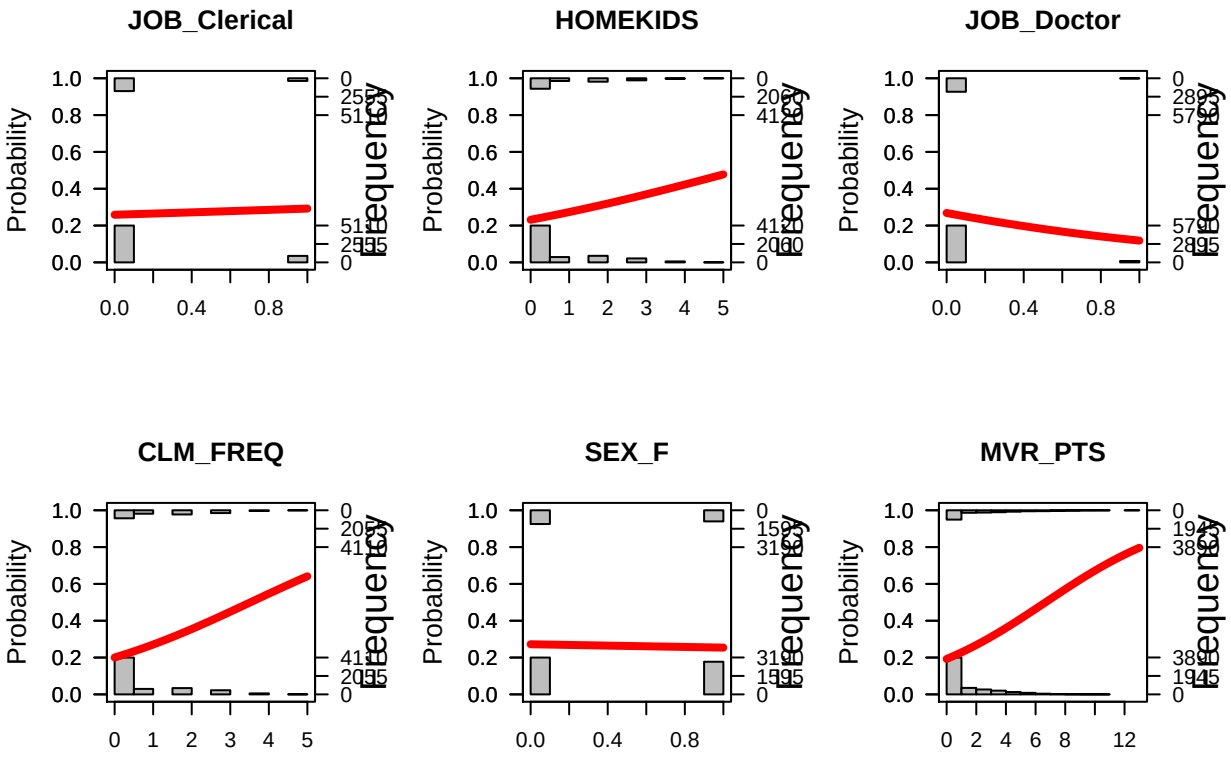
$g(E(y)) = a + Bx_1 + B_2x_2 + B_3x_3 + ...$

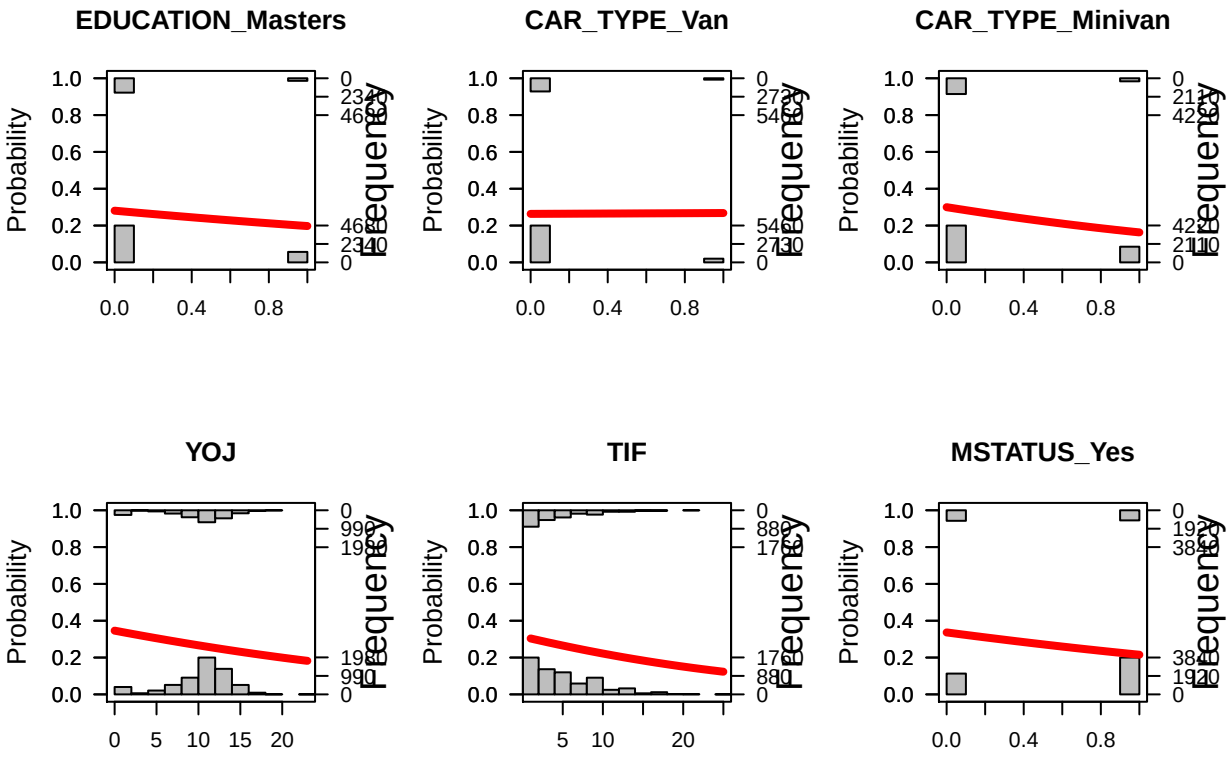
where, g() is the link function, E(y) is the expectation of target variable and $B_0 + B_1x_1 + B_2x_2 + B_3x_3$ is the linear predictor (B_0, B_1, B_2, B_3 to be predicted). The role of link function is to 'link' the expectation of y to linear predictor.

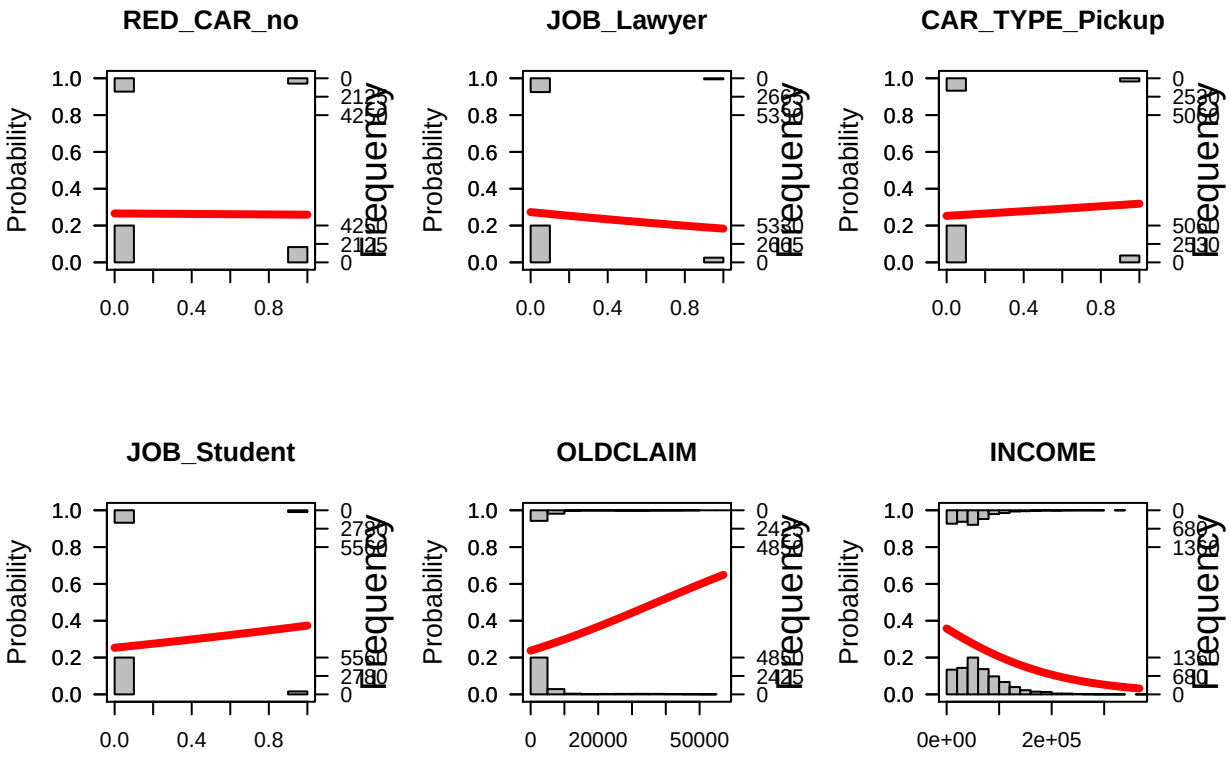
In logistic regression, we are only concerned about the probability of outcome dependent variable (success or failure). As described above, g() is the link function. This function is established using two things: Probability of Success (p) and Probability of Failure (1-p). p should meet following criteria: It must always be positive (since p >= 0) It must always be less than equals to 1 (since p <= 1).

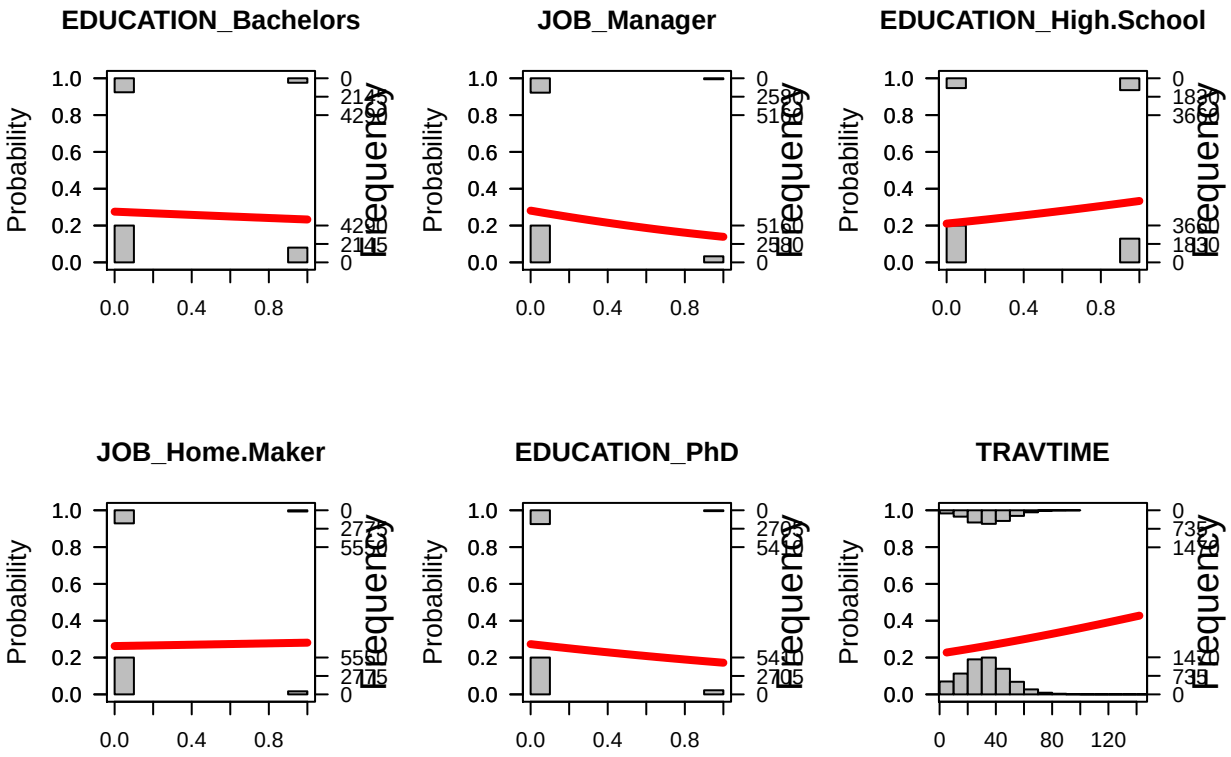
Now let's investigate how our initial data model aligns with the above criteria. In other words, we will plot regression model plots for each variable and compare it to a typical logistic model plot:

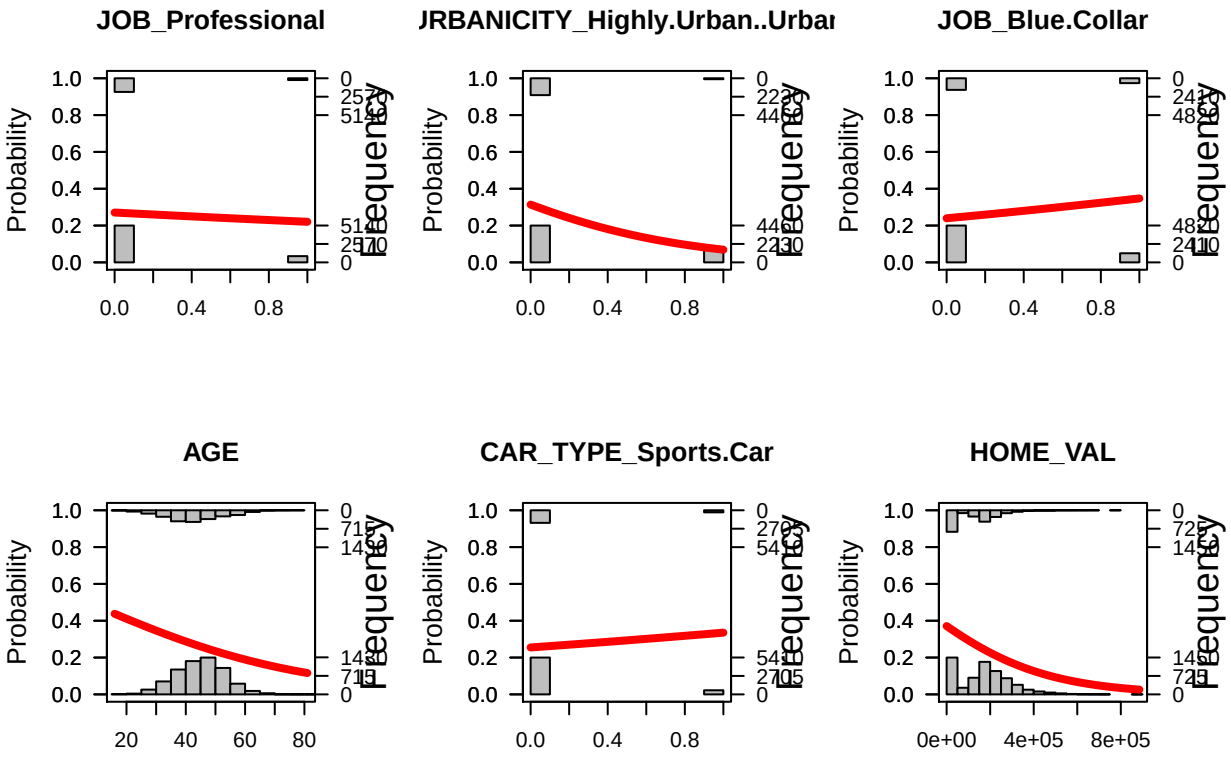


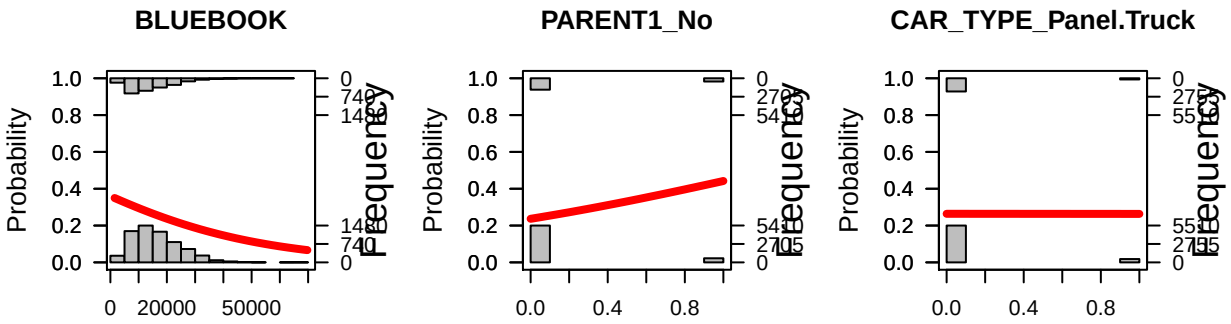

JOB_Clerical
HOMEKIDS
JOB_Doctor
CLM_FREQ
SEX_F
MVR_PTS
Probability
Frequency


EDUCATION_Masters
CAR_TYPE_Van
CAR_TYPE_Minivan
YOJ
TIF
MSTATUS_Yes
Probability
Frequency


RED_CAR_no
JOB_Lawyer
CAR_TYPE_Pickup
JOB_Student
OLDCLAIM
INCOME
Probability
Frequency
0.0
0.4
0.8
0
20000
50000
0e+00
2e+05


EDUCATION_Bachelors
JOB_Manager
EDUCATION_High.School
JOB_Home.Maker
EDUCATION_PhD
TRAVTIME
Probability
Frequency


JOB_Professional
JRBANICITY_Highly.Urban..Urbar
JOB_Blue.Collar
AGE
CAR_TYPE_Sports.Car
HOME_VAL
Probability
Frequency



3.1.5.1 Interpretation You can see that the probability of crashing increases as we get closer to the "1" classification for the CAR_TYPE_Van, RED_CAR_no, JOB_Home.Maker, SEX_F, JOB_Clerical, CAR_TYPE_SUV, TRAVTIME, BLUEBOOK, CAR_TYPE_Pickup, CAR_TYPE_Sports.Car, JOB_Student, KIDSDRIV, JOB_Blue.Collar, HOMEKIDS, MSTATUS_No, EDUCATION_High.School, CAR_USE_Commercial, REVOKED_Yes, PARENT1_Yes, OLDCLAIM, CLM_FREQ, MVR_PTS, URBANICITY_Highly.Urban..Urban variables.

You can see that the probability of crashing decreases as we get closer to the "1" classification for the HOME_VAL, CAR_TYPE_Minivan, MSTATUS_Yes, JOB_Manager, AGE, CAR_AGE, TIF, EDUCATION_Masters, YOJ, EDUCATION_PhD, JOB_Lawyer, JOB_Doctor, EDUCATION_Bachelors, JOB_Professional, INCOME, SEX_M, RED_CAR_yes, CAR_TYPE_Panel.Truck variables.

3.2 Data Preparation

Now that we have completed the data exploration / analysis, we will be transforming the data for use in analysis and modeling.

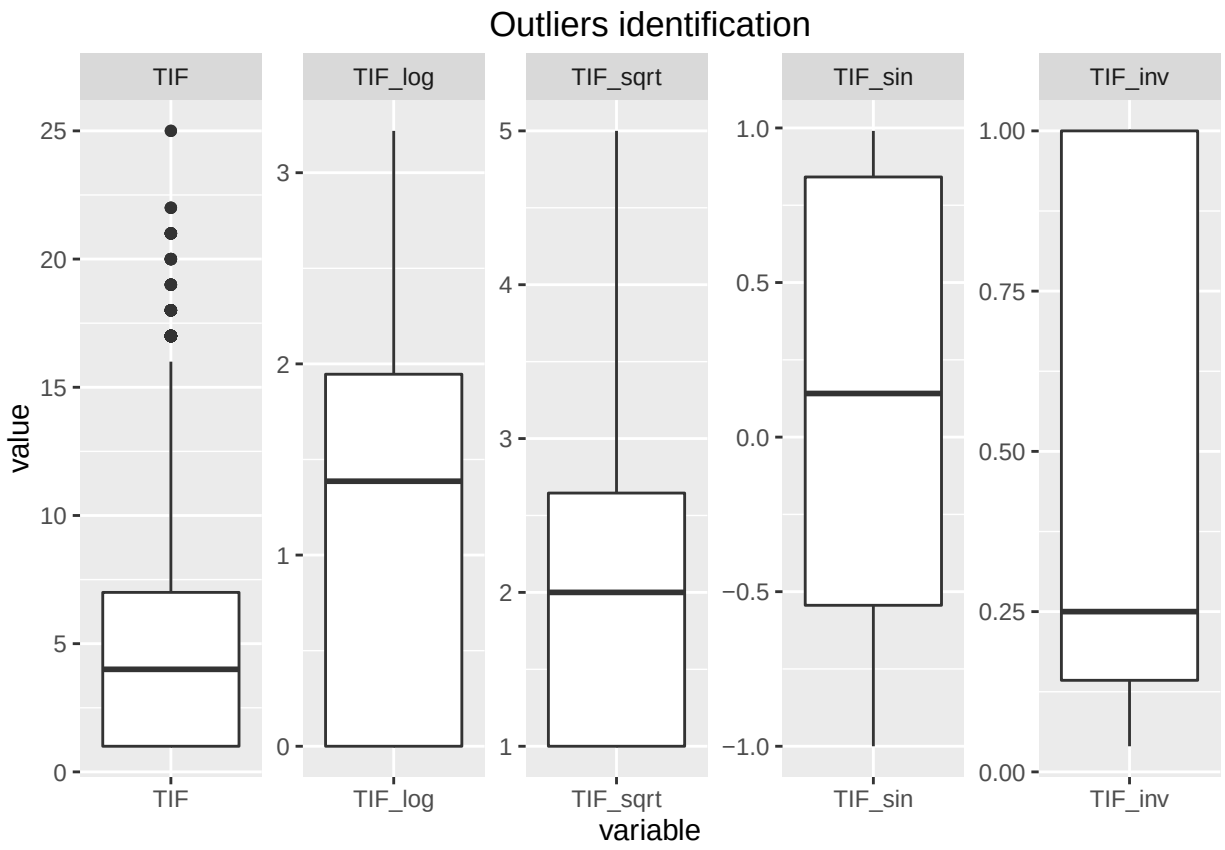
We will be following the below steps as guidelines: - Outliers treatment - Adding New Variables

3.2.1 Outliers treatment

In this sub-section, we will check different transformations for AGE, BLUEBOOK and TIF to create the appropriate outlier-handled / transformed variables.

Transformations for TIF

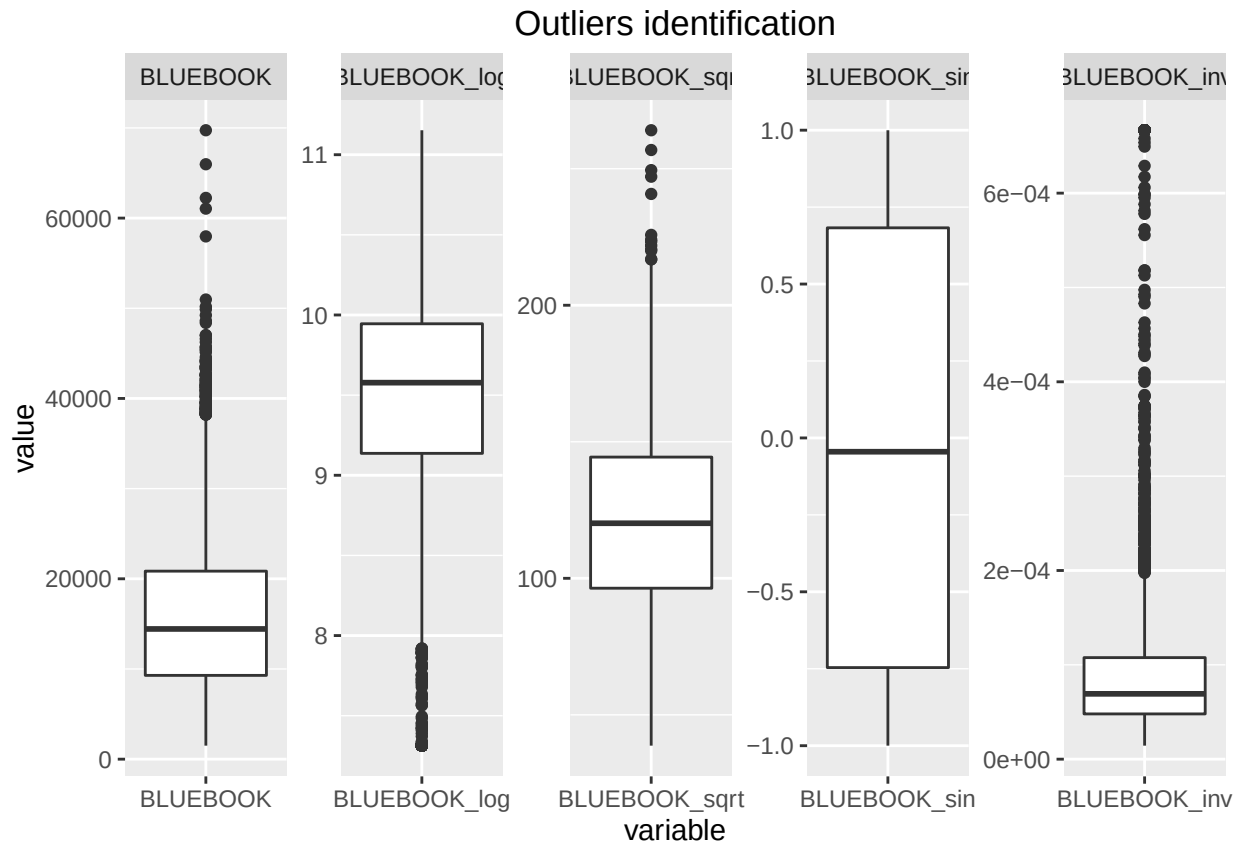
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    4.00    5.35    7.00   25.00
```



From the above charts we can see that a log, sqrt, sin or an inverse transformation works well for TIF. However, the sin transformation seems to be better distributed. Hence, We will create this variable.

Transformations for BLUEBOOK

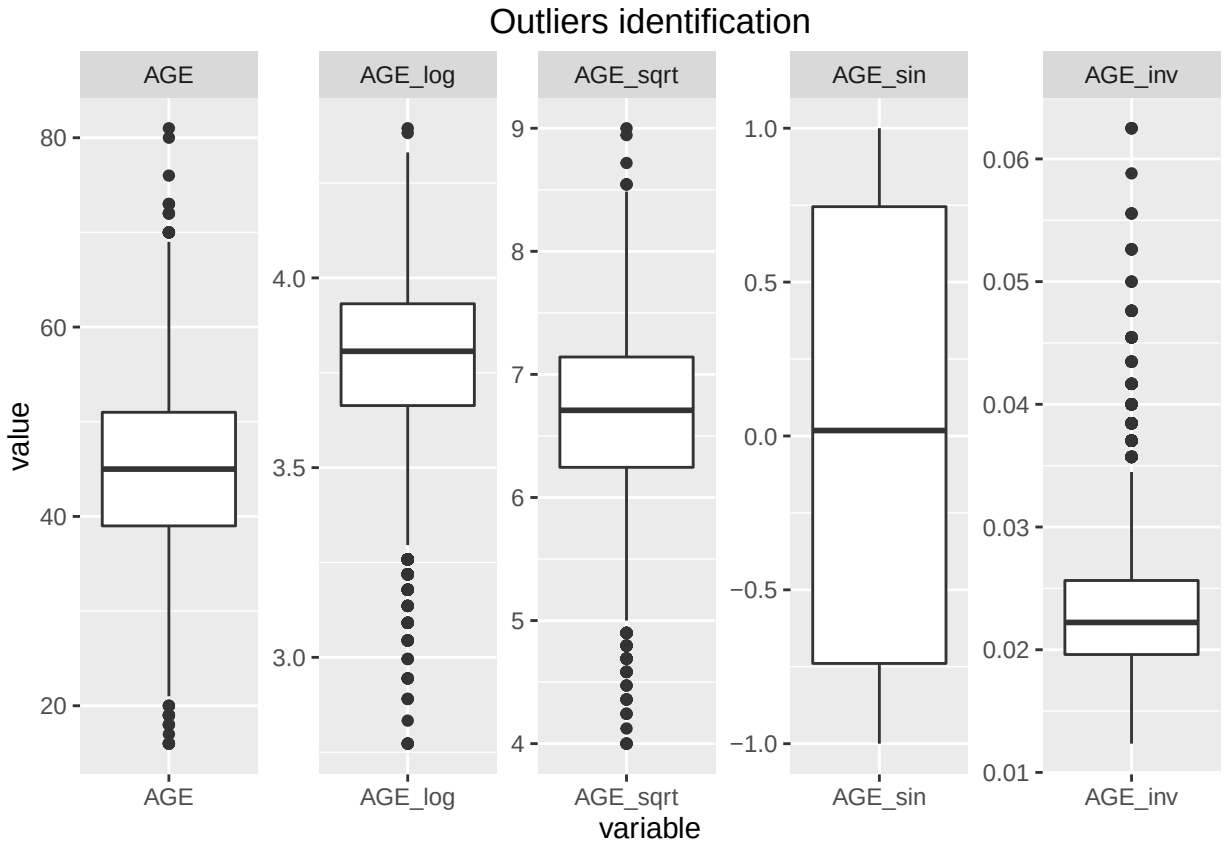
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1500    9290   14440   15710   20850   69740
```



From the above charts we can see that a sin transformation works well. Hence, We will create this variable.

Transformations for AGE

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   39.00   45.00   44.79   51.00   81.00
```



From the above charts we can see that a sin works well for AGE.Hence, We will create this variable.

3.2.2 Adding New Variables

In this section, we generate some additional variables that we feel will help the correlations. The following were some of the observations we made during the data exploration phase for TARGET_FLAG

CAR_TYPE - If you drive Minivans and Panel Trucks you have lesser chance of being in a crash as against Pickups, Sports, SUVs and Vans. Since the distiction is clear, we believe that binning this variable accordingly will help strengthen the correlation. Accordingly, we will bin this variable as below:

CAR_TYPE_FLAG_BIN :

- 1 : if CAR_TYPE is Minivans or Panel Trucks
- 0 : if CAR_TYPE is Pickups, Sports, SUVs or Vans

EDUCATION - If you have only a high school education then you are more likely to crash than if you have a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation:

EDUCATION_FLAG_BIN :

- 0 : if EDUCATION is High School
- 1 : if EDUCATION is Bachelors, Masters or Phd

JOB - If you are a Student, Homemaker, or in a Blue Collar or Clerical job, you are more likely to be in a crash against Doctor, Lawyer, Manager or professional. Again binning this variable will strengthen the correlation:

JOB_TYPE_FLAG_BIN :

- 1 : if JOB_TYPE is Student, Homemaker, or in a Blue Collar or Clerical
- 0 : if JOB_TYPE is Doctor, Lawyer, Manager, professional, Unknown
- INCOME - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this at zero value.

INCOME_FLAG_BIN :

- 1 : if INCOME <= 0
- 0 : if INCOME > 0
- YOJ - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

YOJ_FLAG_BIN :

- 1 : if YOJ <= 0
- 0 : if YOJ > 0
- HOME_VAL - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

HOME_VAL_FLAG_BIN :

- 1 : if HOME_VAL <= 0
- 0 : if HOME_VAL > 0
- OLDCLAIM- There is a huge difference in the coorrelation when we transform this vatiable. Binning this variable seems like a good idea.

OLDCLAIM_FLAG_BIN :

- 1 : if OLDCLAIM <= 0
- 0 : if OLDCLAIM > 0
- CLM_FREQ - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

CLM_FREQ_FLAG_BIN :

- 1 : if CLM_FREQ <= 0
- 0 : if CLM_FREQ > 0
- MVR_PTS - Binning this variable seems to make a difference in the correlation. We will go ahead and create a binned variable for this.

MVR_PTS_FLAG_BIN :

- 1 : if MVR_PTS <= 0
- 0 : if MVR_PTS > 0
- CAR_AGE - There are quite a few records with a 1 year car age. We will use this bound to generate a binned variable as well as retain the original varible as is.

CAR_AGE_FLAG_BIN :

- 1 : if CAR_AGE <= 1
- 0 : if CAR_AGE > 0
- AGE - There is no specific pattern that emerges. We will retain this variable as is.
- BLUEBOOK - There is no specific pattern that emerges. We will retain this variable as is.
- TRAVTIME - from the plot, we can see that there is a clear pattern around the value - 20. We will go ahead and create a binned variable for this.

TRAVTIME_FLAG_BIN :

- 1 : if TRAVTIME <= 20
- 0 : if TRAVTIME > 0

3.2.3 Additional Binned Variables

After having prepared the data, we will go ahead and drop some of the variables.

```
#write.csv(insure_train_full, file = "D:/CUNY/Courses/Business Analytics and Data Mining/Assignments/da

#DS_TARGET_FLAG <- insure_train_full
DS_TARGET_FLAG <- select(insure_train_full, -TARGET_AMT, -JOB_Blue.Collar, -JOB_Clerical, -JOB_Doctor, -

# New Additional Variables.
#-AGE, -AGE_IMPUTE, -BLUEBOOK,-CAR_AGE_IMPUTE, -CAR_AGE_MISS, -CLM_FREQ, -HOME_VAL, -HOME_VAL_IMPUTE, -

str(DS_TARGET_FLAG)
```

```
## 'data.frame':    8157 obs. of  39 variables:
##  $ TARGET_FLAG       : int  0 0 0 0 0 1 0 1 1 0 ...
##  $ KIDSDRIV          : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE               : num  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS          : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ               : num  11 11 10 14 10.5 ...
##  $ INCOME            : num  67349 91449 16039 54046 114986 ...
##  $ HOME_VAL          : num  0 257252 124191 306251 243925 ...
##  $ TRAVTIME          : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ BLUEBOOK          : num  14230 14940 4010 15440 18000 ...
##  $ TIF               : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ OLDCLAIM          : num  4461 0 38690 0 19217 ...
##  $ CLM_FREQ          : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ 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 ...
```

```
##  $ CAR_USE_Commercial: num  0 1 0 0 0 1 0 1 0 1 ...
##  $ MSTATUS_Yes       : num  0 0 1 1 1 0 1 1 0 0 ...
##  $ PARENT1_Yes       : num  0 0 0 0 0 1 0 0 0 0 ...
##  $ RED_CAR_yes       : num  1 1 0 1 0 0 0 1 0 0 ...
##  $ REVOKED_Yes       : num  0 0 0 0 1 0 0 1 0 0 ...
##  $ SEX_M             : num  1 1 0 1 0 0 0 1 0 1 ...
##  $ URBANICITY_Rural  : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ YOJ_MISS          : num  0 0 0 0 1 0 1 1 0 0 ...
##  $ INCOME_MISS       : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ HOME_VAL_MISS     : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ CAR_AGE_MISS      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TIF_sin           : num  -1 0.841 -0.757 0.657 0.841 ...
##  $ BLUEBOOK_sin      : num  -0.988 -0.988 0.971 0.8 -0.97 ...
##  $ AGE_sin           : num  -0.305 -0.832 -0.428 0.67 -0.262 ...
##  $ CAR_TYPE_FLAG_BIN : num  1 1 0 1 0 0 0 0 0 0 ...
##  $ EDUCATION_FLAG_BIN: num  1 0 0 0 1 1 0 1 1 1 ...
##  $ JOB_TYPE_FLAG_BIN : num  0 1 1 1 0 1 1 1 1 0 ...
##  $ INCOME_FLAG_BIN   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ YOJ_FLAG_BIN      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HOME_VAL_FLAG_BIN : num  1 0 0 0 0 1 0 0 1 1 ...
##  $ OLDCLAIM_FLAG_BIN : num  0 1 0 1 0 1 1 0 1 1 ...
##  $ CLM_FREQ_FLAG_BIN : num  0 1 0 1 0 1 1 0 1 1 ...
##  $ MVR_PTS_FLAG_BIN  : num  0 1 0 1 0 1 1 0 1 0 ...
##  $ CAR_AGE_FLAG_BIN  : num  0 1 0 0 0 0 1 0 1 0 ...
##  $ TRAVTIME_FLAG_BIN : num  1 0 1 0 0 0 0 0 0 0 ...
```

3.3 Build Models

In this section, we will create 3 models. Aside from using original and transformed data, we will also using different methods and functions such as Linear Discriminant Analysis, step function, and logit function to enhance our models. newline newline Below is our model definition: -Model 1- This model will be created using all the variables in train data set with logit function GLM. -Model 2: This model step function will be used to enhance the model 1. -Model 3- This model will be created using calssification and regression tree.

3.3.1 Prepare TRAIN and VALID datasets

However, prior to that, we hold out a subset of data as a validation dataset to check model performance. This will be useful when we select a model.

```
smp_size <- floor(0.80 * nrow(DS_TARGET_FLAG))

## set the seed to make your partition reproductible
set.seed(123)

train_index <- sample(seq_len(nrow(DS_TARGET_FLAG)), size = smp_size)

DS_TARGET_FLAG_TRAIN<- DS_TARGET_FLAG[train_index, ]
DS_TARGET_FLAG_VALID <- DS_TARGET_FLAG[-train_index, ]
```

3.3.2 Model 1 and enahncement of Model 1 with step function (Model 2)

In this model, we will be using all the given variables in train data set. We will create model using logit function. We will then step thru the model to remove unnecessary variables and generate the refined model. We will highlight the summary of the refined model.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = na.omit(DS_TARGET_FLAG_TRAIN))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4903  -0.7193  -0.4098   0.6561   3.1494
##
## Coefficients: (1 not defined because of singularities)
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.988e-01  3.403e-01  -1.172 0.241181
## KIDSDRIV            3.979e-01  6.763e-02   5.883 4.03e-09 ***
## AGE                -4.759e-03  4.458e-03  -1.067 0.285793
## HOMEKIDS            1.818e-02  4.169e-02   0.436 0.662766
## YOJ                 1.598e-02  1.378e-02   1.159 0.246305
## INCOME             -3.058e-06  1.260e-06  -2.427 0.015231 *
## HOME_VAL           -8.986e-07  6.553e-07  -1.371 0.170277
## TRAVTIME            1.334e-02  2.786e-03   4.786 1.70e-06 ***
## BLUEBOOK           -1.680e-05  4.762e-06  -3.527 0.000420 ***
## TIF                -5.153e-02  9.256e-03  -5.567 2.59e-08 ***
## OLDCLAIM           -2.124e-05  4.735e-06  -4.487 7.23e-06 ***
## CLM_FREQ            7.001e-02  4.953e-02   1.413 0.157567
## MVR_PTS             1.042e-01  2.119e-02   4.918 8.73e-07 ***
## CAR_AGE             5.177e-03  1.077e-02   0.481 0.630837
## CAR_USE_Commercial  7.512e-01  7.643e-02   9.828  < 2e-16 ***
## MSTATUS_Yes        -5.337e-01  9.590e-02  -5.565 2.62e-08 ***
## PARENT1_Yes         3.787e-01  1.219e-01   3.106 0.001898 **
## RED_CAR_yes        -1.602e-02  9.656e-02  -0.166 0.868237
## REVOKED_Yes         1.052e+00  1.032e-01  10.194  < 2e-16 ***
## SEX_M              -7.591e-03  9.676e-02  -0.078 0.937467
## URBANICITY_Rural   -2.313e+00  1.254e-01 -18.453  < 2e-16 ***
## YOJ_MISS           -9.088e-02  1.503e-01  -0.605 0.545393
## INCOME_MISS        -8.443e-02  1.491e-01  -0.566 0.571275
## HOME_VAL_MISS      -1.280e-02  1.414e-01  -0.091 0.927878
## CAR_AGE_MISS        2.667e-01  1.351e-01   1.975 0.048291 *
## TIF_sin             2.893e-02  5.475e-02   0.528 0.597165
## BLUEBOOK_sin       -2.722e-02  4.562e-02  -0.597 0.550670
## AGE_sin             1.864e-02  4.599e-02   0.405 0.685238
## CAR_TYPE_FLAG_BIN  -5.584e-01  8.259e-02  -6.760 1.38e-11 ***
## EDUCATION_FLAG_BIN -3.764e-01  9.592e-02  -3.923 8.73e-05 ***
## JOB_TYPE_FLAG_BIN   3.225e-01  9.760e-02   3.304 0.000953 ***
## INCOME_FLAG_BIN     4.796e-01  3.508e-01   1.367 0.171569
## YOJ_FLAG_BIN        8.043e-02  3.797e-01   0.212 0.832256
## HOME_VAL_FLAG_BIN   1.107e-02  1.552e-01   0.071 0.943148
## OLDCLAIM_FLAG_BIN  -4.899e-01  1.371e-01  -3.572 0.000354 ***
## CLM_FREQ_FLAG_BIN          NA         NA      NA       NA
## MVR_PTS_FLAG_BIN    2.785e-02  9.490e-02   0.293 0.769170
## CAR_AGE_FLAG_BIN    8.810e-02  1.170e-01   0.753 0.451457
```

```
## TRAVTIME_FLAG_BIN  -9.989e-02  1.104e-01  -0.905 0.365589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7551.7  on 6524  degrees of freedom
## Residual deviance: 5886.1  on 6487  degrees of freedom
## AIC: 5962.1
##
## Number of Fisher Scoring iterations: 5
```

Interpretation for the TF_Model1 and TF_Model1_ref *newline*

TF_Model1:

From model 1 summary we can find following important points-

(i) Variable URBANICITY_Rural has most significant association with lowest p value. negative value of log odd function indicates that chances of accidents ae higher in Urbancity areas compare to rural area.

(ii) For MSTATUS_Yes variable log odd is negative which indicates married people tend to drive slowly and have less number of accidents.

(iii) Sex variable has no significant association which means driving patters does not depend on men and women.

(iv) variable REVOKED_Yes has strong association which indicates if person's license has been revoked in last 7 years then chane of end up in accidents are much higher with log odds value of 0.809090.

(v) If person has a claim in last 5 years then chances of more claims are higher. Variable OLDCLAIM_FLAG_BI indicates that with negative log odds value(1 is here no claim -0.559409).

(vi) AIC value of the model is AIC: 6078.7 and number of iteration was 5.

TF_Model1_ref:

Model 1 has AIC value 6078.7 and enhaned model TF_Model1_ref has AIC value 6064.9 slightly better compared to model1. We will look into more details on model1_ref below-

(i) Based on the outcome from model_ref, it can be seen that following variables KIDSDRIV,PARENT1_Yes, MSTATUS_Yes,CAR_USE_Commercial, REVOKED_Yes, TIF_sin , CAR_TYPE_FLAG_BIN , EDUCATION_FLAG_BIN, JOB_TYPE_FLAG_BIN, INCOME_FLAG_BIN , HOME_VAL_FLAG_BIN,OLDCLAI , URBANICITY_Rural are only statistically significant. Most of the variables are having similar association as above model 1.

(ii)As for the statistically significant variables, URBANICITY_Rural has the lowest p-value suggesting a strong association of the URBANICITY_Rural to the target variable.Implication is also same negative value indicate lower chances of accidents in rural areas.

(iii) One interesting outcome is when childrens are dirivng your car then more chances of accidents with log odd value of 0.41327 for variable KIDSDRIV.

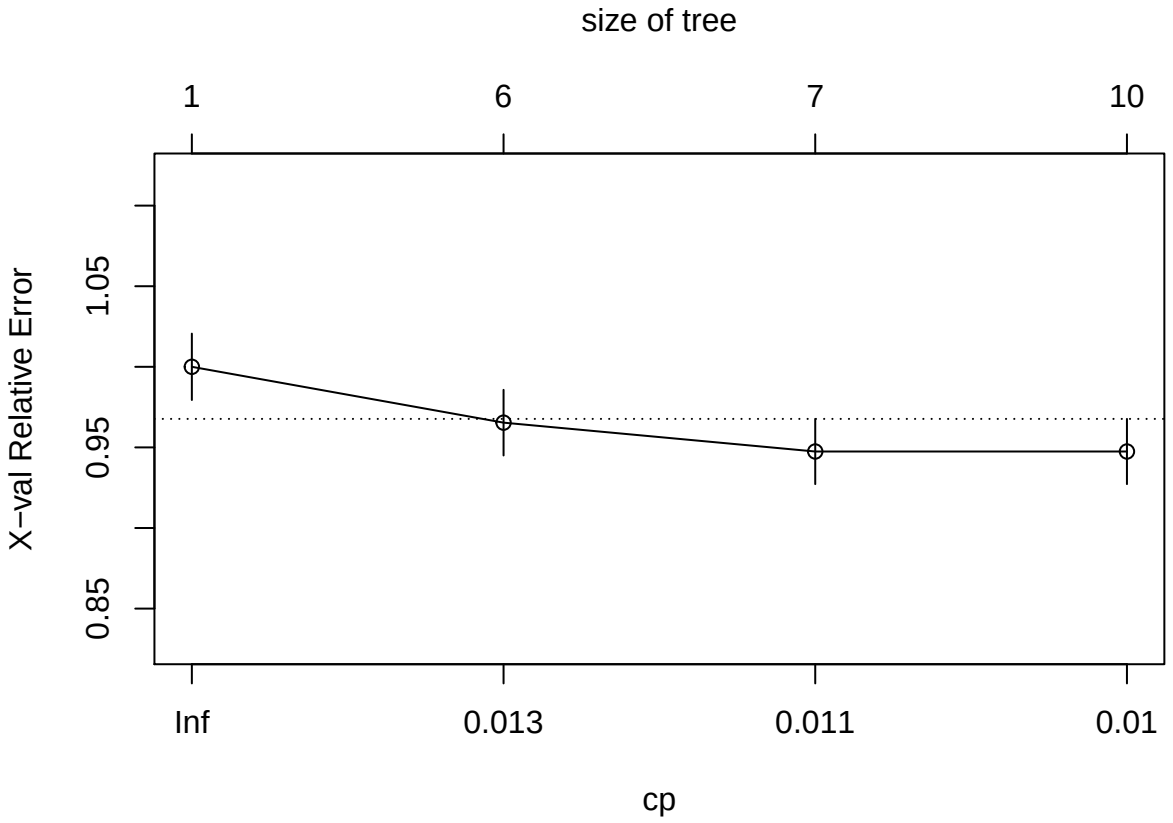
(iv) For variable CAR_TYPE_FLAG_BIN there is high negative corelation is there with log odds value of -0.65867 that means Mnivan and Panel truck has higher chance of getting into an accident.

(v) Variable EDUCATION_FLAG_BIN has negative log odds value of -0.46755 indicaating that people with higher education above high school has less chance of an accident comared to the other group.

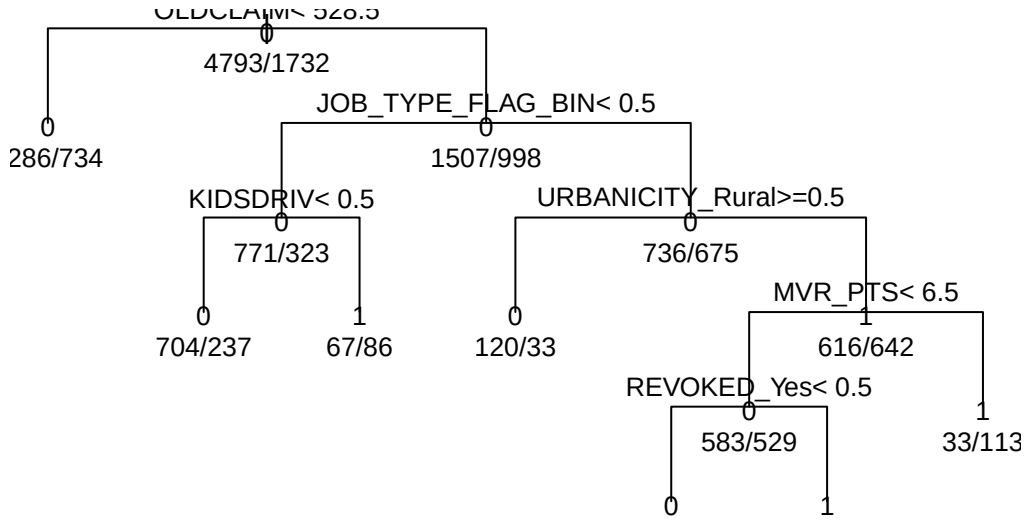
(vi) No. of iterations are 5 before lowest value of AIC was derived for this model.

3.3.3 Model 3

In this model, we will be using original variables; however we use the CART (Classification and Regression Trees) algorithm to train the model. We will then Prune the tree and have a look at the summary of this pruned model.




```
##
## Classification tree:
## rpart(formula = TARGET_FLAG ~ ., data = DS_TARGET_FLAG_TRAIN,
##     method = "class")
##
## Variables actually used in tree construction:
## [1] JOB_TYPE_FLAG_BIN KIDSDRIV          MVR_PTS           OLDCLAIM
## [5] REVOKED_Yes       URBANICITY_Rural
##
## Root node error: 1732/6525 = 0.26544
##
## n= 6525
##
##         CP nsplit rel error  xerror     xstd
## 1 0.016051      0   1.00000 1.00000 0.020594
## 2 0.010970      5   0.91975 0.96536 0.020360
## 3 0.010778      6   0.90878 0.94746 0.020235
```



Interpretation for Model 3 *newline*

Following analysis can be drawn from this model: (i)The following variables have been used for classification - OLD_CLAIM, JOB_TYPE_FLAG_BIN, URBANICITY_Rural, KIDSDRIV, MVR_PTS, REVOKED_Yes.

(ii)lowest Cp value and Xerror occured on split 7.

(iii) OLDCLAIM_FLAG_BIN is the first variable used to split the classification based on its value 0 and 1. When there is claim (1 in abobe variable) branch is further split to other branches by variable JOB_TYPE_FLAG_BIN (based on value 0 and 1). Based on value of JOB_TYPE_FLAG_BIN (0 and 1) there is two different routes in calssification. one Split (774/323) is based on variable KIDSDRIV and the other one (738/675) is based on URBANICITY_Rural variable. Using the above variable total 7 splits have been performed for classification.

3.4 Model Evaluation Using VALID Data

Lets go ahead and apply the above models to the VALID dataset that we had held out. Below is the table of predictions for each of the models:

3.4.1 Evaluation of Model 1

Table 3: Model 1 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
1	0.7947304	0.2052696	0.447619	0.6460481	0.8269948	0.3926341	0.8079817

Model 1 has good accuracy value close to 78.3%. sensitivity value is lower than the specificity value.

3.4.2 Evaluation of Model 2

Table 4: Model 2 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
2	0.7830882	0.2169118	0.2690476	0.70625	0.7914402	0.2537439	0.8079817

Model 2 has good accuracy value close to 77.3% and very close to model1. sensitivity value is lower than the specificity value.

3.4.3 Evaluation of Model 3

Table 5: Model 3 evaluation KPIs

	Accuracy	Error_Rate	Precision	sensitivity	specificity	F1_Score	AUC
1	0.7947304	0.2052696	0.4476190	0.6460481	0.8269948	0.3926341	0.8079817
2	0.7830882	0.2169118	0.2690476	0.7062500	0.7914402	0.2537439	0.8079817
3	0.7561275	0.2438725	0.1666667	0.5932203	0.7688243	0.1451789	0.6733272

This model has accuracy value of 75.4%. AUC for this model is 67.4 % and less compared to the other two models.

3.5 Final Logistic Model Selection Summary

Following is the comparison of various metrics for above 3 models

Table 6: Model Performance Metrics Comparison

Model_No	Accuracy	Error_Rate	AUC	Precision	sensitivity	specificity	F1_Score
1	0.7947304	0.2052696	0.8079817	0.4476190	0.6460481	0.8269948	0.3926341
2	0.7830882	0.2169118	0.8079817	0.2690476	0.7062500	0.7914402	0.2537439
3	0.7561275	0.2438725	0.6733272	0.1666667	0.5932203	0.7688243	0.1451789

From the comparison table, we see that Model 1 is quite superior from the accuracy and AUC perspective.

The AUC provides the best score on probability of correctly identifying the patterns at various cut off values. The Accuracy, on the other hand, is calculated as specific cut off value. For this assignment we will go with cut off value of 0.5 and choose the Model 1 based on Accuracy value for further prediction on evaluation data set.

3.5.1 Detailed Inference for Final Model

The following analysis will be carried out on the final model:

(i) Relevant variables in the model
(ii) Estimate confidence interval for coefficient
(iii) odds ratios and 95% CI
(iv) AUC curve
(v) Distribution of prediction

3.5.2 Most important variables in the model

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = na.omit(DS_TARGET_FLAG_TRAIN))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4903  -0.7193  -0.4098   0.6561   3.1494
##
## Coefficients: (1 not defined because of singularities)
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.988e-01  3.403e-01  -1.172 0.241181
## KIDSDRIV            3.979e-01  6.763e-02   5.883 4.03e-09 ***
## AGE                -4.759e-03  4.458e-03  -1.067 0.285793
## HOMEKIDS            1.818e-02  4.169e-02   0.436 0.662766
## YOJ                 1.598e-02  1.378e-02   1.159 0.246305
## INCOME             -3.058e-06  1.260e-06  -2.427 0.015231 *
## HOME_VAL           -8.986e-07  6.553e-07  -1.371 0.170277
## TRAVTIME            1.334e-02  2.786e-03   4.786 1.70e-06 ***
## BLUEBOOK           -1.680e-05  4.762e-06  -3.527 0.000420 ***
## TIF                -5.153e-02  9.256e-03  -5.567 2.59e-08 ***
## OLDCLAIM           -2.124e-05  4.735e-06  -4.487 7.23e-06 ***
## CLM_FREQ            7.001e-02  4.953e-02   1.413 0.157567
## MVR_PTS             1.042e-01  2.119e-02   4.918 8.73e-07 ***
## CAR_AGE             5.177e-03  1.077e-02   0.481 0.630837
## CAR_USE_Commercial  7.512e-01  7.643e-02   9.828  < 2e-16 ***
## MSTATUS_Yes        -5.337e-01  9.590e-02  -5.565 2.62e-08 ***
## PARENT1_Yes         3.787e-01  1.219e-01   3.106 0.001898 **
## RED_CAR_yes        -1.602e-02  9.656e-02  -0.166 0.868237
## REVOKED_Yes         1.052e+00  1.032e-01  10.194  < 2e-16 ***
## SEX_M              -7.591e-03  9.676e-02  -0.078 0.937467
## URBANICITY_Rural   -2.313e+00  1.254e-01 -18.453  < 2e-16 ***
## YOJ_MISS           -9.088e-02  1.503e-01  -0.605 0.545393
## INCOME_MISS        -8.443e-02  1.491e-01  -0.566 0.571275
## HOME_VAL_MISS      -1.280e-02  1.414e-01  -0.091 0.927878
## CAR_AGE_MISS        2.667e-01  1.351e-01   1.975 0.048291 *
## TIF_sin             2.893e-02  5.475e-02   0.528 0.597165
## BLUEBOOK_sin       -2.722e-02  4.562e-02  -0.597 0.550670
## AGE_sin             1.864e-02  4.599e-02   0.405 0.685238
## CAR_TYPE_FLAG_BIN  -5.584e-01  8.259e-02  -6.760 1.38e-11 ***
## EDUCATION_FLAG_BIN -3.764e-01  9.592e-02  -3.923 8.73e-05 ***
## JOB_TYPE_FLAG_BIN   3.225e-01  9.760e-02   3.304 0.000953 ***
```

```
## INCOME_FLAG_BIN     4.796e-01  3.508e-01   1.367 0.171569
## YOJ_FLAG_BIN        8.043e-02  3.797e-01   0.212 0.832256
## HOME_VAL_FLAG_BIN   1.107e-02  1.552e-01   0.071 0.943148
## OLDCLAIM_FLAG_BIN  -4.899e-01  1.371e-01  -3.572 0.000354 ***
## CLM_FREQ_FLAG_BIN          NA         NA      NA       NA
## MVR_PTS_FLAG_BIN    2.785e-02  9.490e-02   0.293 0.769170
## CAR_AGE_FLAG_BIN    8.810e-02  1.170e-01   0.753 0.451457
## TRAVTIME_FLAG_BIN  -9.989e-02  1.104e-01  -0.905 0.365589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7551.7  on 6524  degrees of freedom
## Residual deviance: 5886.1  on 6487  degrees of freedom
## AIC: 5962.1
##
## Number of Fisher Scoring iterations: 5
```

Following are the most relevant variables for the model: CAR_USE_Commercial, REVOKED_Yes, URBANICITY_Rural, CAR_TYPE_FLAG_BIN,TRAVTIME, OLDCLAIM_FLAG_BIN, MVR_PTS, TIF_SIN, KIDSDRIV, PARENT1_Yes, EDUCATION_FLAG_BIN, MSTATUS_Yes, BLUEBOOK, OLDCLAIM, JOB_TYPE_FLAG_BIN, HOME_VAL, INCOME_FLAG_BIN.

we can write the equation of the Model 1 as:

$log(y) = -0.4015 + 0.3431 * KIDSDRIV - 0.000001027 * HOME_VAL + 0.01557 * TRAVTIME - 0.00001858 * BLUEBOOK - 0.05103 * TIF - 0.00001873 * OLDCLAIM + 0.1043 * MVR_PTS + 0.7632 * CAR_USE_Commercial - 0.4066 * MSTATUS_Yes + 0.5246 * PARENT1_Yes + 1.025 * REVOKED_Yes - 2.221 * URBANICITY_Rural - 0.5662 * CAR_TYPE_FLAG_BIN - 0.3996 * EDUCATION_FLAG_BIN + 0.3584 * JOB_TYPE_FLAG_BIN + 0.311 * INCOME_FLAG_BIN - 0.627 * OLDCLAIM_FLAG_BIN$

3.5.3 Analysis of odds ratios of variables 95% CI

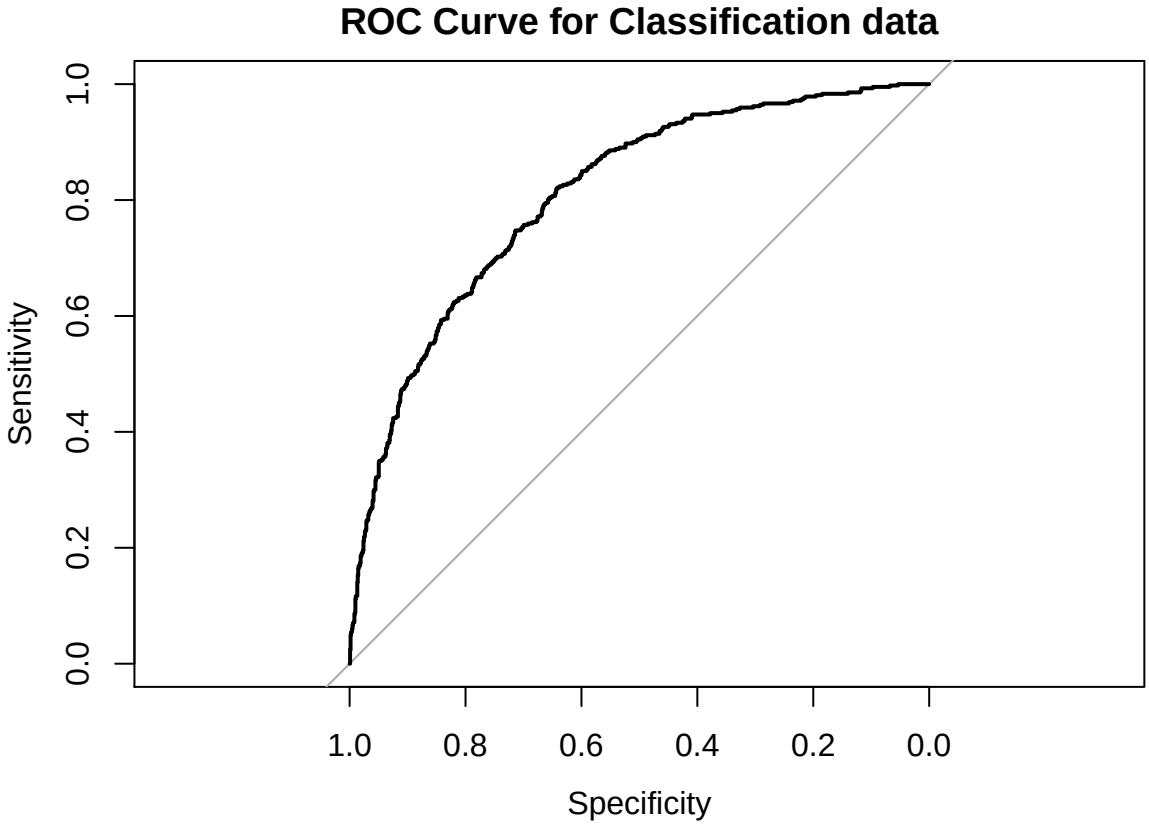
```
##                          OR      2.5 %    97.5 %
## (Intercept)        0.67113058 0.34449416 1.3074714
## KIDSDRIV           1.48862259 1.30382382 1.6996140
## AGE                0.99525237 0.98659334 1.0039874
## HOMEKIDS           1.01834688 0.93844701 1.1050495
## YOJ                1.01610813 0.98902600 1.0439318
## INCOME             0.99999694 0.99999447 0.9999994
## HOME_VAL           0.99999910 0.99999782 1.0000004
## TRAVTIME           1.01342437 1.00790509 1.0189739
## BLUEBOOK           0.99998320 0.99997387 0.9999925
## TIF                0.94977489 0.93269993 0.9671624
## OLDCLAIM           0.99997876 0.99996948 0.9999880
## CLM_FREQ           1.07251612 0.97328473 1.1818647
## MVR_PTS            1.10986959 1.06470918 1.1569455
## CAR_AGE            1.00519015 0.98418928 1.0266391
## CAR_USE_Commercial 2.11945792 1.82459644 2.4619701
## MSTATUS_Yes        0.58643036 0.48593986 0.7077019
## PARENT1_Yes        1.46035120 1.14992968 1.8545705
## RED_CAR_yes        0.98410869 0.81442639 1.1891436
## REVOKED_Yes        2.86451319 2.33977796 3.5069293
```

```
## SEX_M              0.99243744 0.82099263 1.1996844
## URBANICITY_Rural   0.09894387 0.07738989 0.1265009
## YOJ_MISS           0.91313153 0.68015614 1.2259085
## INCOME_MISS        0.91903450 0.68611221 1.2310296
## HOME_VAL_MISS      0.98728640 0.74837344 1.3024707
## CAR_AGE_MISS       1.30565907 1.00200553 1.7013335
## TIF_sin            1.02935564 0.92462226 1.1459523
## BLUEBOOK_sin       0.97314416 0.88991133 1.0641617
## AGE_sin            1.01881446 0.93100276 1.1149085
## CAR_TYPE_FLAG_BIN  0.57214143 0.48662962 0.6726796
## EDUCATION_FLAG_BIN 0.68635498 0.56871853 0.8283239
## JOB_TYPE_FLAG_BIN  1.38053840 1.14018363 1.6715608
## INCOME_FLAG_BIN    1.61542333 0.81225230 3.2127856
## YOJ_FLAG_BIN       1.08375386 0.51487406 2.2811839
## HOME_VAL_FLAG_BIN  1.01113208 0.74588194 1.3707103
## OLDCLAIM_FLAG_BIN  0.61268865 0.46828247 0.8016259
## CLM_FREQ_FLAG_BIN          NA         NA        NA
## MVR_PTS_FLAG_BIN   1.02824027 0.85372369 1.2384312
## CAR_AGE_FLAG_BIN   1.09209197 0.86830795 1.3735506
## TRAVTIME_FLAG_BIN  0.90493543 0.72885274 1.1235577
```

The following points can be made for the important variables in the model:

In keeping all other variables same, the odds of an accident increases as follow: 1.8449962 for per unit change in CAR_USE_Commercial, 2.2458626 per unit change in REVOKED_Yes, 0.1104633 for per unit change in URBANICITY_Rural, etc. Any value which is less than 1, it means that there is less chance of an event with the per unit increase of the variable.

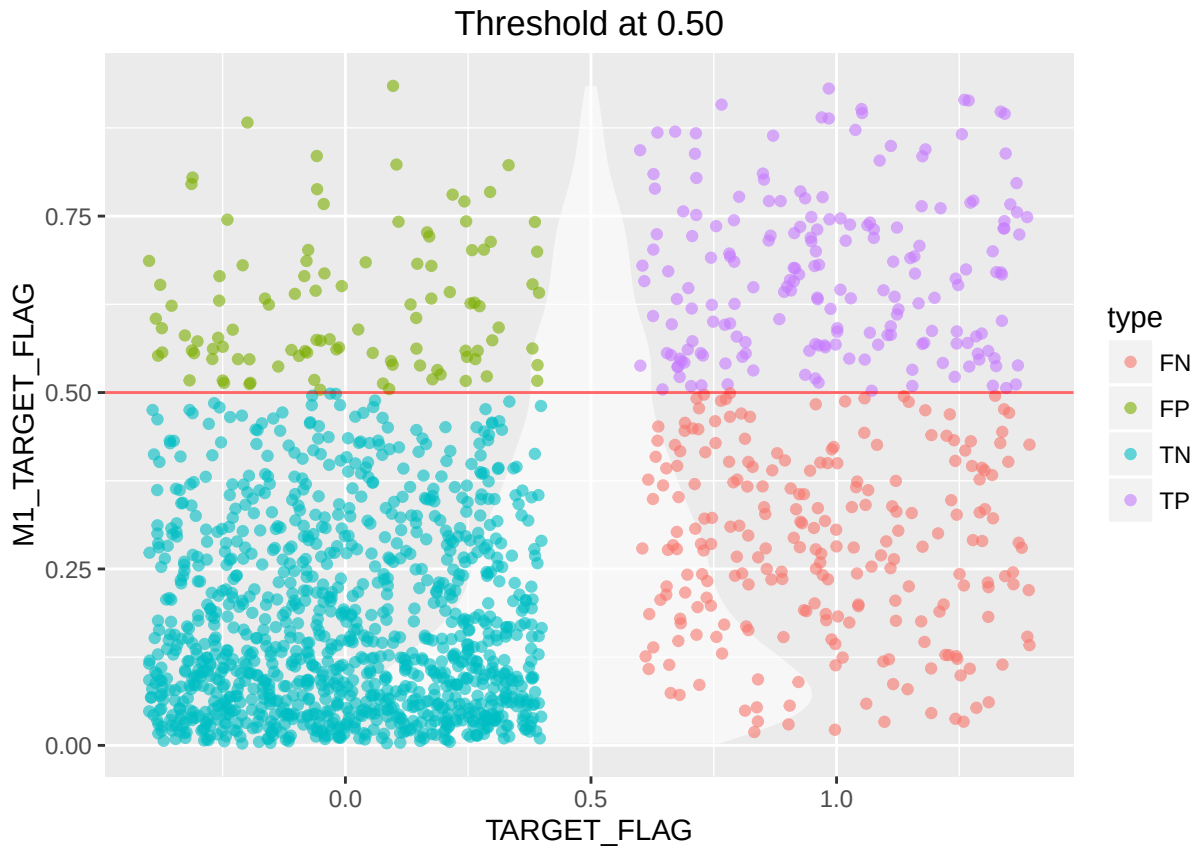
3.5.4 ROC curve for the selected model




```
##
## Call:
## roc.formula(formula = DS_TARGET_FLAG_VALID$TARGET_FLAG ~ DS_TARGET_FLAG_VALID$M1_TARGET_FLAG,     da
##
## Data: DS_TARGET_FLAG_VALID$M1_TARGET_FLAG in 1212 controls (DS_TARGET_FLAG_VALID$TARGET_FLAG 0) < 42
## Area under the curve: 0.808
```

3.5.5 Distribution of the Predictions

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```



Considering the target has value 1 (accident occurs) and 0 when no accident, then the above plot illustrates the tradeoff of choosing a reasonable threshold. In other words, if the threshold is increased, the number of false positive (FP) results is lowered; while the number of false negative (FN) results increases.

newpage

4 Linear Regression for TARGET_AMT

In this section we will use Linear regression to model the TARGET_AMT. We will first start with the Data Exploration. We will be using only those records where the TARGET_FLAG is 1. This indicates that the vehicle crashed. In such a scenario, we will be modeling the cost of repair using Linear Regression. First, lets create the required data set for the "Crashed" data from the existing "clean" full data and look at the structure of the resulting dataset. We will remove from the new "crashed" dataset all those variables that were created specifically for predicting TARGET_FLAG. We will be creating these variables separately for predicting TARGET_AMT.

```
## 'data.frame':    2152 obs. of  45 variables:
##  $ TARGET_AMT           : num  2946 4021 2501 6077 1267 ...
##  $ KIDSDRIV             : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ AGE                  : num  34 37 34 53 53 45 28 43 32 40 ...
##  $ HOMEKIDS             : int  1 2 0 0 0 0 1 0 1 0 ...
##  $ YOJ                  : num  12 10.5 10 14 11 ...
##  $ INCOME               : num  125301 107961 62978 77100 130795 ...
##  $ HOME_VAL             : num  0 333680 0 0 0 ...
##  $ TRAVTIME             : int  46 44 34 15 64 48 29 52 26 20 ...
```

```
##  $ BLUEBOOK             : num  17430 16970 11200 18300 28340 ...
##  $ TIF                  : int  1 1 1 1 6 1 6 1 1 4 ...
##  $ OLDCLAIM             : num  0 2374 0 0 0 ...
##  $ CLM_FREQ             : int  0 1 0 0 0 0 2 0 0 1 ...
##  $ MVR_PTS              : int  0 10 0 0 3 3 0 3 0 13 ...
##  $ CAR_AGE              : int  7 7 1 11 10 5 1 1 1 6 ...
##  $ CAR_USE_Commercial   : num  1 1 0 0 1 0 1 1 0 1 ...
##  $ MSTATUS_Yes          : num  0 1 0 0 0 1 1 1 1 0 ...
##  $ PARENT1_Yes          : num  1 0 0 0 0 0 0 0 0 0 ...
##  $ RED_CAR_yes          : num  0 1 0 0 1 0 0 1 0 1 ...
##  $ REVOKED_Yes          : num  0 1 0 0 0 0 0 0 0 1 ...
##  $ SEX_M                : num  0 1 0 0 1 0 0 1 0 1 ...
##  $ URBANICITY_Rural     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ EDUCATION_Bachelors  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ EDUCATION_High.School: num  0 0 0 0 0 1 1 1 1 1 ...
##  $ EDUCATION_Masters    : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ EDUCATION_PhD        : num  0 0 0 0 1 0 0 0 0 0 ...
##  $ JOB_Blue.Collar      : num  1 1 0 0 0 0 1 1 0 1 ...
##  $ JOB_Clerical         : num  0 0 1 0 0 0 0 0 1 0 ...
##  $ JOB_Doctor           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JOB_Home.Maker       : num  0 0 0 0 0 1 0 0 0 0 ...
##  $ JOB_Lawyer           : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ JOB_Manager          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JOB_Professional     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JOB_Student          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ JOB_Unknown          : num  0 0 0 0 1 0 0 0 0 0 ...
##  $ CAR_TYPE_Minivan     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CAR_TYPE_Panel.Truck : num  0 0 0 0 1 0 0 1 0 0 ...
##  $ CAR_TYPE_Pickup      : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ CAR_TYPE_Sports.Car  : num  1 0 0 1 0 0 0 0 0 0 ...
##  $ CAR_TYPE_SUV         : num  0 0 1 0 0 1 1 0 1 0 ...
##  $ CAR_TYPE_Van         : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ YOJ_MISS             : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ INCOME_MISS          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HOME_VAL_MISS        : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ CAR_AGE_MISS         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ TRAVTIME_FLAG_BIN    : num  0 0 0 1 0 0 0 0 0 1 ...
```

We notice that the dependent variable here is TARGET_AMT. Apart from the dependent variables, we have 49 independent or predictor variables.

Also, since we created this dataset from the "Clean" full dataset, we already have taken care of the missing values.

However, we may need to look into the outliers and correlations again since we have a new target variable to correlate against.

4.1 Data Summary and Correlation Analysis

4.1.1 Data Summary

In this section, we will create summary data to better understand the relationship each of the variables have with our dependent variables using correlation, central tendency, and dispersion as shown below:

4.1.2 Correlations

Now we will produce the correlation table between the independent variables and the dependent variable - TARGET_AMT

Table 7: Correlation between TARGET_AMT and predictor variables

	Correlation_TARGET_AMT
TARGET_AMT	1.0000000
BLUEBOOK	0.1181297
CAR_TYPE_Panel.Truck	0.0682806
SEX_M	0.0513430
CAR_TYPE_Van	0.0499290
CAR_USE_Commercial	0.0496142
INCOME	0.0440737
JOB_Professional	0.0406747
JOB_Unknown	0.0402613
MVR_PTS	0.0396710
YOJ	0.0328277
HOME_VAL	0.0299847
EDUCATION_PhD	0.0294767
AGE	0.0279078
RED_CAR_yes	0.0271768
PARENT1_Yes	0.0238302
YOJ_MISS	0.0187194
HOME_VAL_MISS	0.0163608
JOB_Blue.Collar	0.0155259
EDUCATION_Masters	0.0143267
EDUCATION_Bachelors	0.0136662
JOB_Lawyer	0.0102382
TRAVTIME	0.0053657
CLM_FREQ	0.0023251
HOMEKIDS	0.0002698
KIDSDRIV	-0.0000869
TRAVTIME_FLAG_BIN	-0.0001208
INCOME_MISS	-0.0017793
URBANICITY_Rural	-0.0048888
OLDCLAIM	-0.0049723
CAR_TYPE_Minivan	-0.0058234
TIF	-0.0060620
CAR_AGE_MISS	-0.0114633
JOB_Doctor	-0.0122018
CAR_AGE	-0.0136248
JOB_Clerical	-0.0151891
CAR_TYPE_Sports.Car	-0.0152654
CAR_TYPE_Pickup	-0.0174060
JOB_Manager	-0.0256129
JOB_Home.Maker	-0.0293974
JOB_Student	-0.0331511
MSTATUS_Yes	-0.0351848
EDUCATION_High.School	-0.0359529
REVOKED_Yes	-0.0365018
CAR_TYPE_SUV	-0.0405600

The above table suggests that none of the variables seem to have a very strong correlation with TARGET_AMT.

However, BLUEBOOK, CAR_TYPE_Panel.Truck, SEX_M, CAR_TYPE_Van, CAR_USE_Commercial, INCOME, INCOME_IMPUTE, JOB_Professional, JOB_Unknown, MVR_PTS, YOJ, YOJ_IMPUTE, HOME_VAL_IMPUTE, EDUCATION_PhD, HOME_VAL, AGE, AGE_IMPUTE, RED_CAR_yes, PARENT1_Yes, YOJ_MISS, HOME_VAL_MISS, JOB_Blue.Collar, EDUCATION_Masters, EDUCATION_Bachelors, JOB_Lawyer, TRAVTIME, CLM_FREQ, HOMEKIDS have a positive correlation.

Similarly, KIDSDRIV, TRAVTIME_FLAG_BIN, INCOME_MISS, URBANICITY_Rural, OLDCLAIM, CAR_TYPE_Minivan, TIF, CAR_AGE_MISS, JOB_Doctor, CAR_AGE, CAR_AGE_IMPUTE, JOB_Clerical, CAR_TYPE_Sports.Car, CAR_TYPE_Pickup, JOB_Manager, JOB_Home.Maker, JOB_Student, MSTATUS_Yes, EDUCATION_High.School, REVOKED_Yes, CAR_TYPE_SUV have a negative correlation.

Lets now see how values in some of the variable affects the correlation:

CAR_TYPE - If you drive Vans or Panel Trucks your cost of repair seems to increase as against Minivan, Pickup, Sports.Car, SUV. Since the distiction is clear, we believe that binning this variable accordingly will help strengthen the correlation.

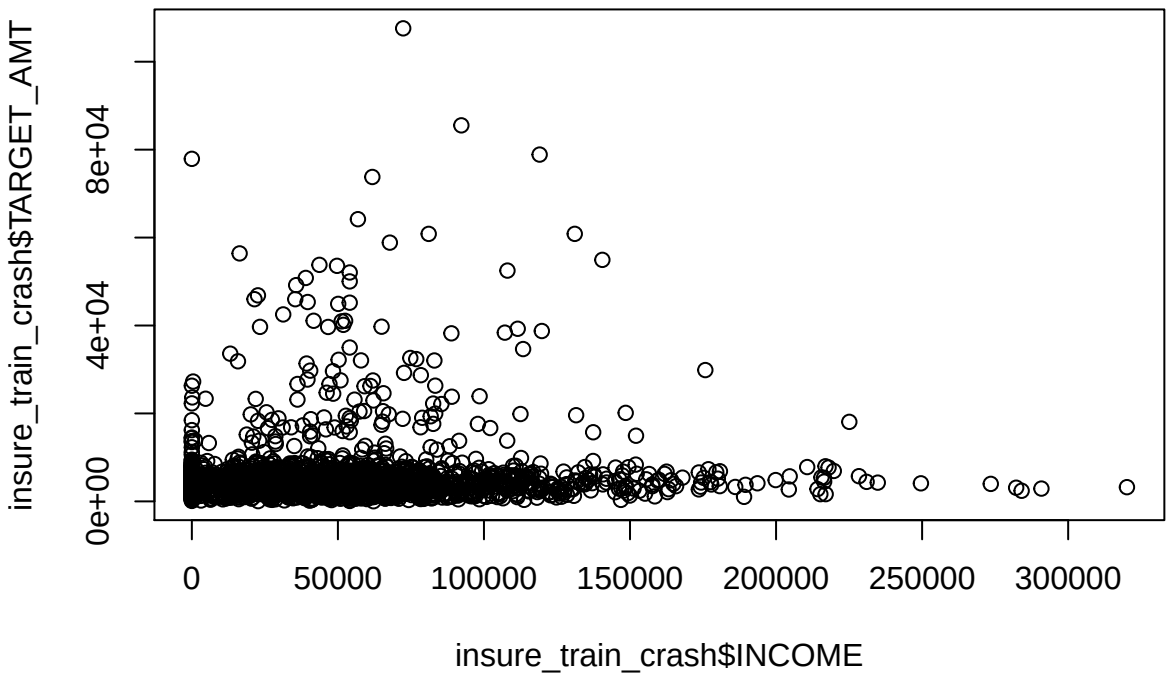
EDUCATION - If you have only a high school education then your cost of repair is less compared to a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation.

JOB - If you are a Lawyer, Professional, in a Blue Collar job or the job is unknown, you spend more on repairs as compared to a Doctor, Manager, Home Maker, Student, or Clerical job. Again binning this variable will strengthen the correlation.

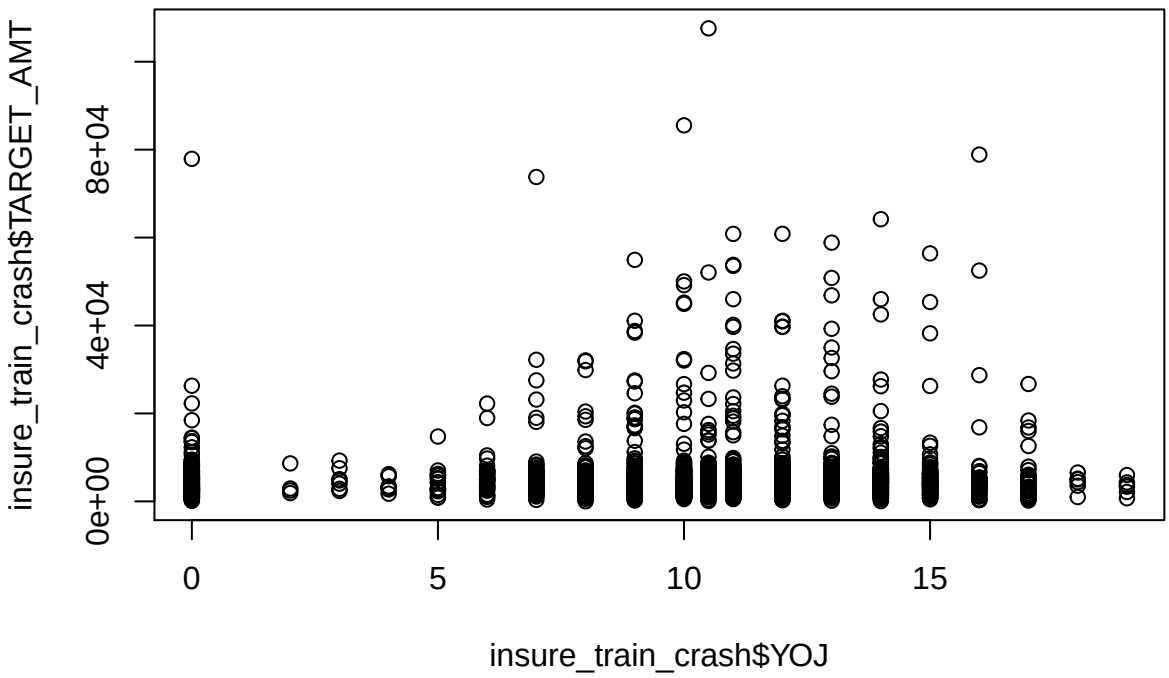
4.1.3 Binning of Variables

Lets have a look at the following numeric variables that have 0 as one of their values: INCOME, YOJ, HOME_VAL, OLDCLAIM, CLM_FREQ, MVR_PTS, CAR_AGE, AGE, BLUEBOOK, TIF, TRAVTIME. The goal here is to see if we can bin these variables into zero and non-zero bin values and check the correlations. While doing that we will also see how the variables are distributed vis-a-vis TARGET_AMT.

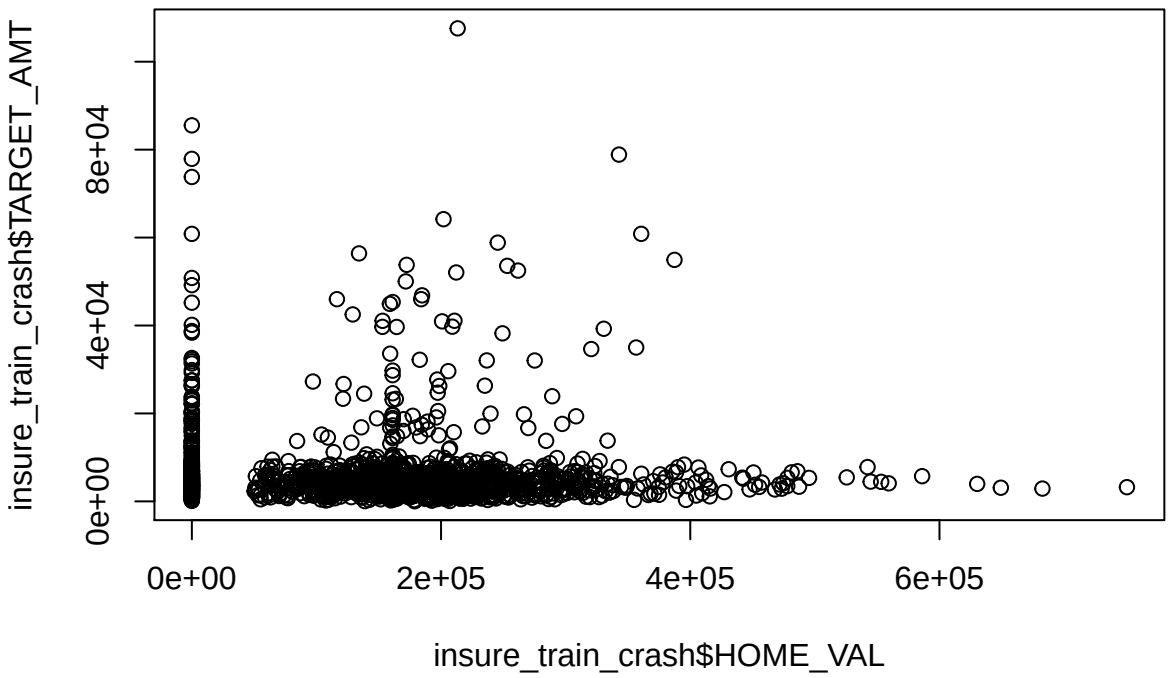
```
check_bins <- function(var, thresholds) {
    col_x <- which(colnames(insure_train_crash)==var)
    old_x <- select(insure_train_crash, col_x)
    cor_old <- cor(old_x, insure_train_crash$TARGET_AMT,use = 'na.or.complete')
    ds <- data.frame("Item" = "Original", "Correlation"= round(cor_old, 5))

    old_tresh <- 0
    for(i in 1:length(thresholds)) {
        New_x <- ifelse((select(insure_train_crash, col_x) >= old_tresh & select(insure_train_crash, co

        cor_new <- cor(New_x, insure_train_crash$TARGET_AMT,use = 'na.or.complete')

        ds_1 <- data.frame("Item" = as.character(thresholds[i]), "Correlation"= round(cor_new, 5))

        ds <- rbind(ds, ds_1)
        old_tresh <- thresholds[i]
    }

    return (ds)
}

plot(insure_train_crash$INCOME, insure_train_crash$TARGET_AMT)
```



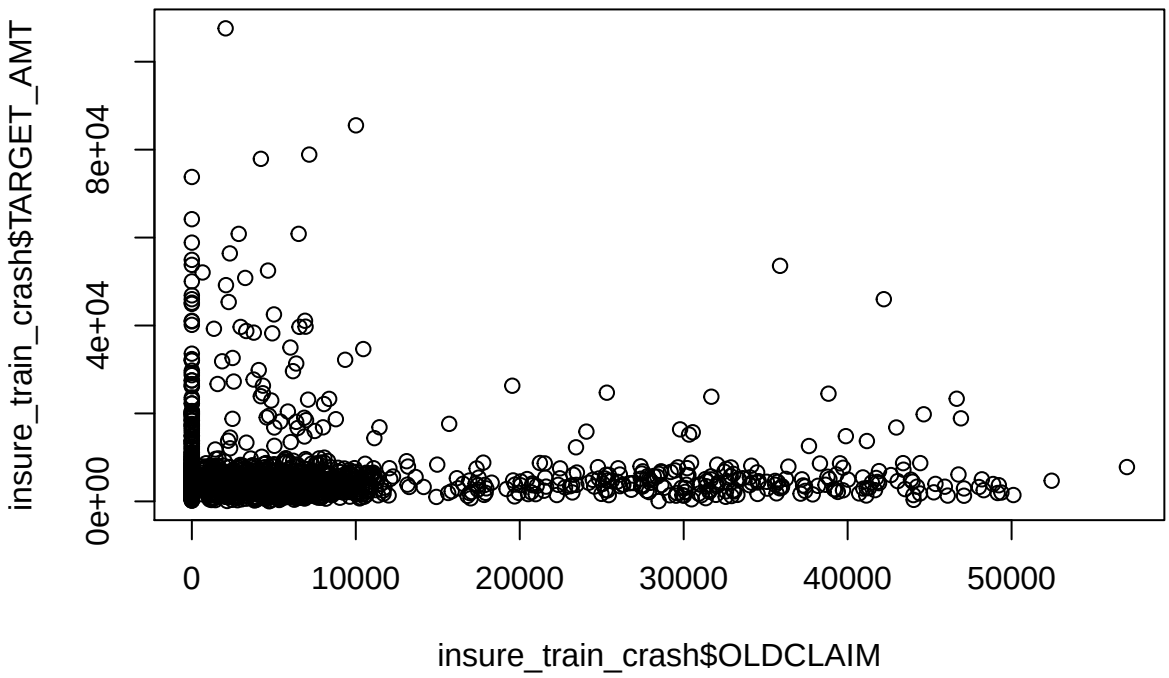

```
#check_bins("INCOME", c(0, 50000, 125000, 200000))

plot(insure_train_crash$YOJ, insure_train_crash$TARGET_AMT)
```



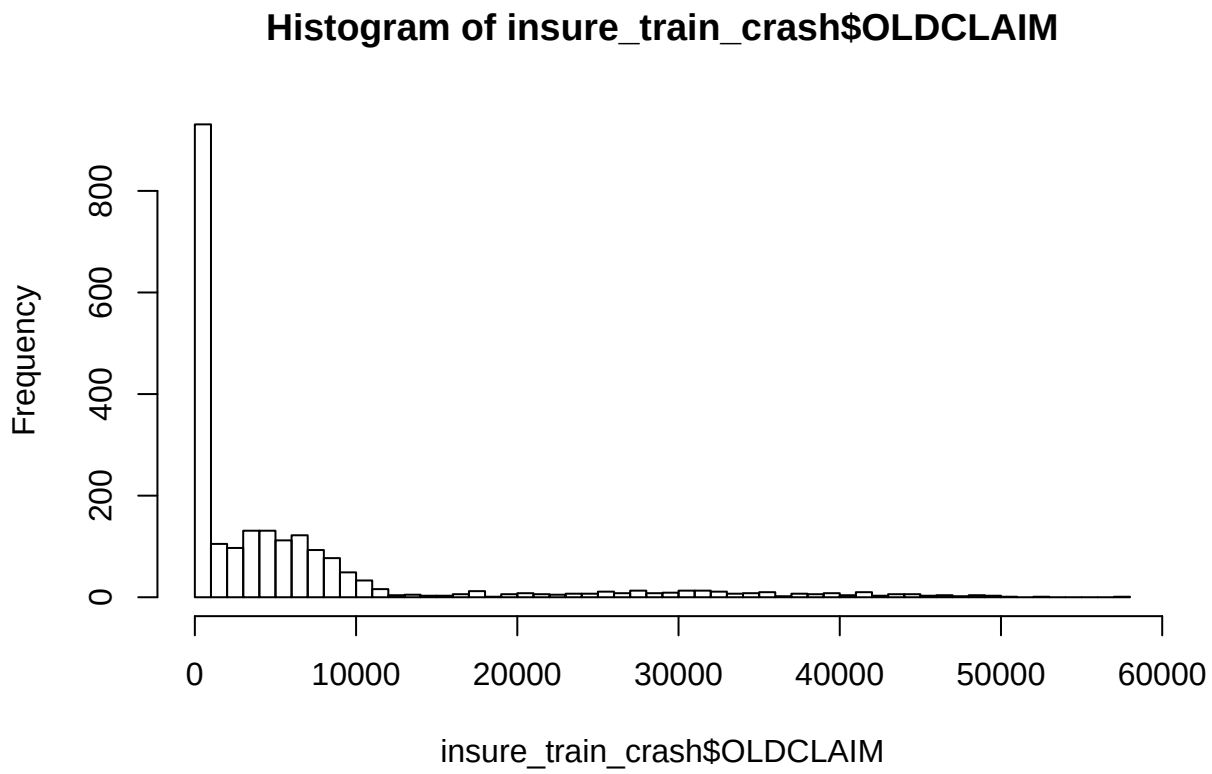

```
#check_bins("YOJ", c(0:19))

plot(insure_train_crash$HOME_VAL, insure_train_crash$TARGET_AMT)
```



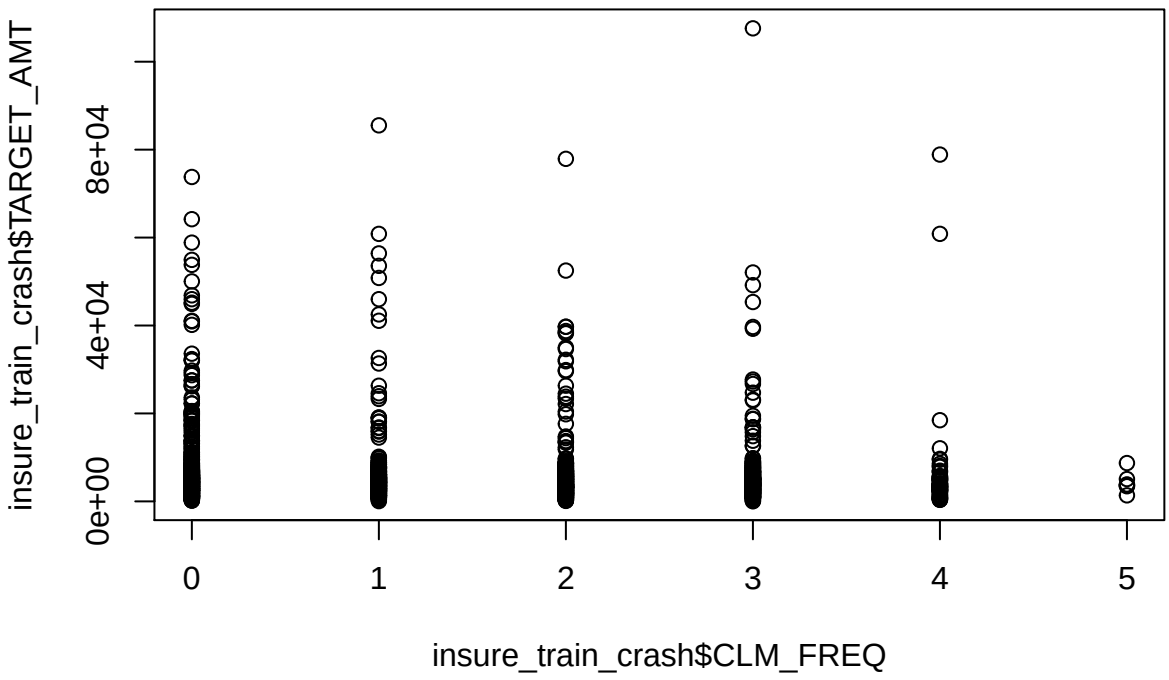

```
#check_bins("HOME_VAL", c(seq(0, 600000, 10000)))

plot(insure_train_crash$OLDCLAIM, insure_train_crash$TARGET_AMT)
```



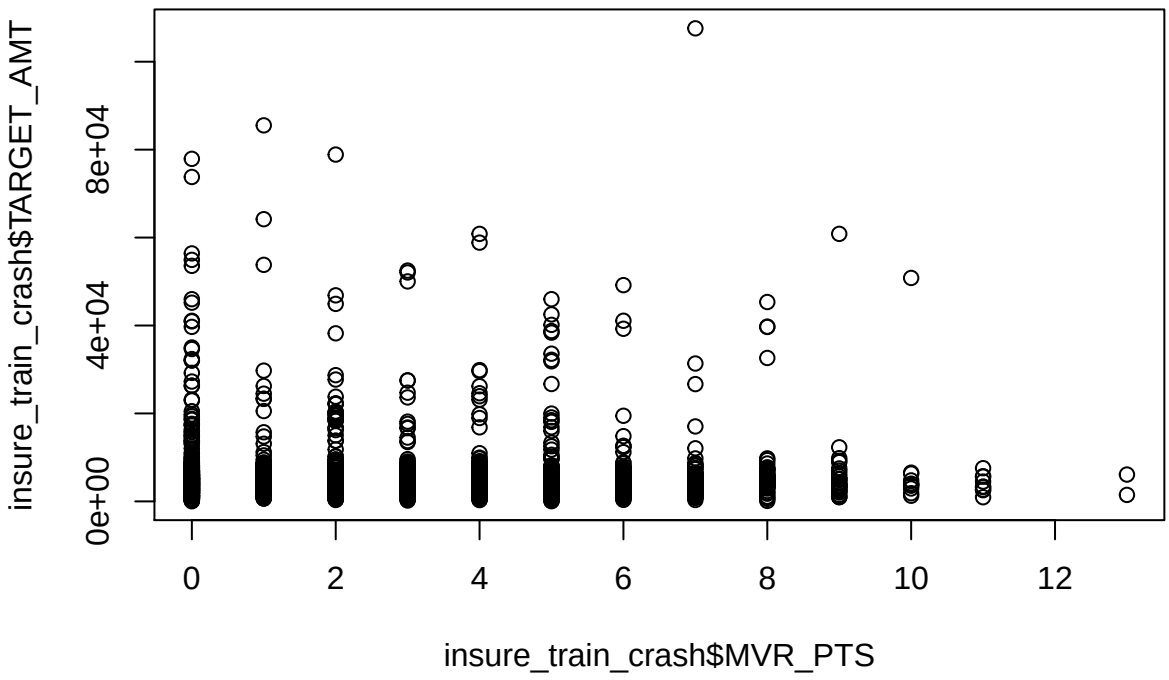

```
hist(insure_train_crash$OLDCLAIM, breaks=50)
```




```
#check_bins("OLDCLAIM", c(seq(0, 50000, 1000)))

#show_hist("CLM_FREQ")
plot(insure_train_crash$CLM_FREQ, insure_train_crash$TARGET_AMT)
```



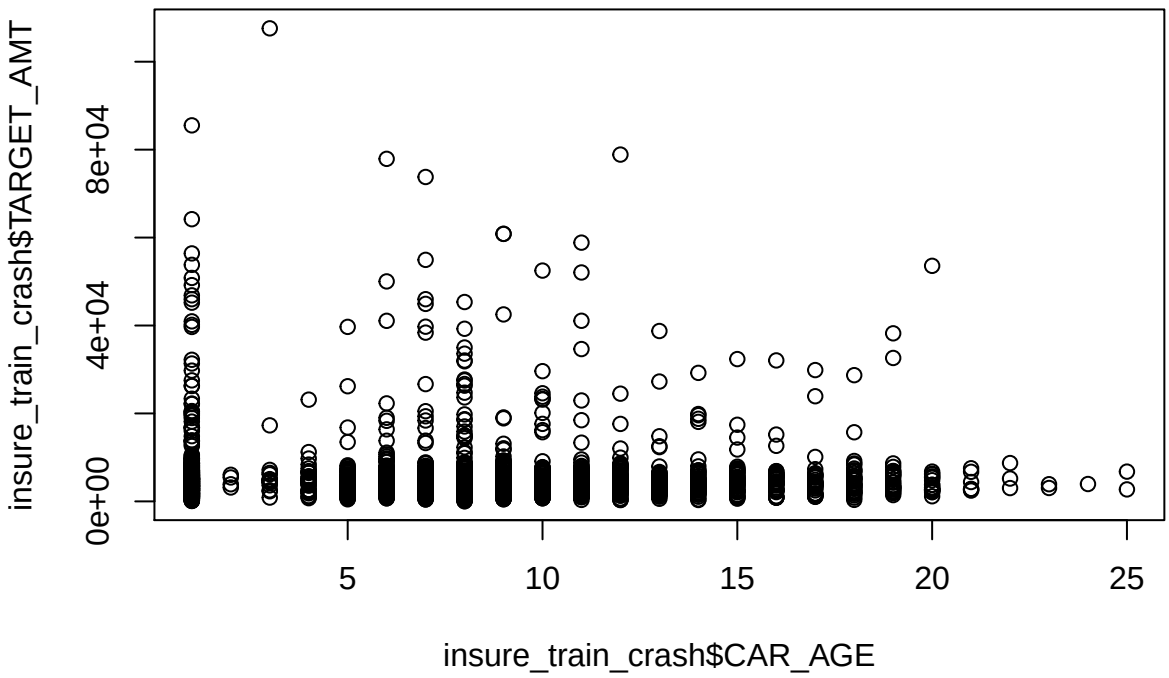

```
#check_bins("CLM_FREQ", c(0, 1, 2, 3, 4))

#table(insure_train_full$MVR_PTS)
plot(insure_train_crash$MVR_PTS, insure_train_crash$TARGET_AMT)
```



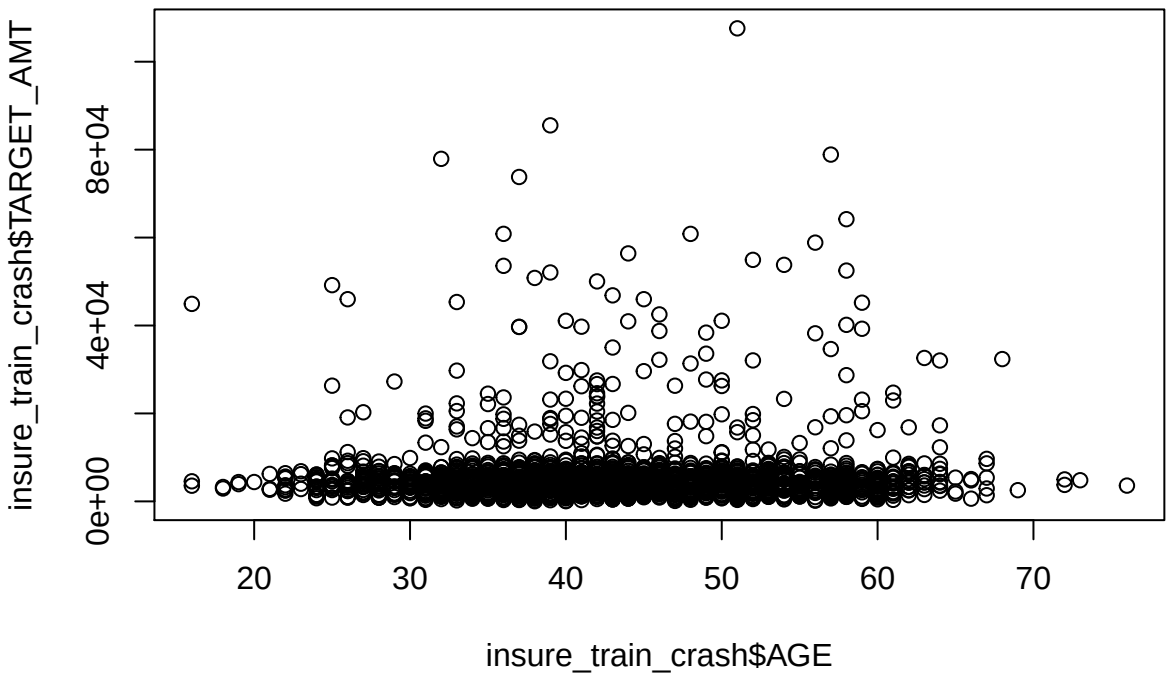

```
#check_bins("MVR_PTS", c(0:12))

table(insure_train_full$CAR_AGE)
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 1934   12   54  135  305  451  524 1047  526  469  460  368  356  311  273
##   16   17   18   19   20   21   22   23   24   25   26   27   28
##  234  213  151  128   90   51   27   18   10    6    2    1    1
```

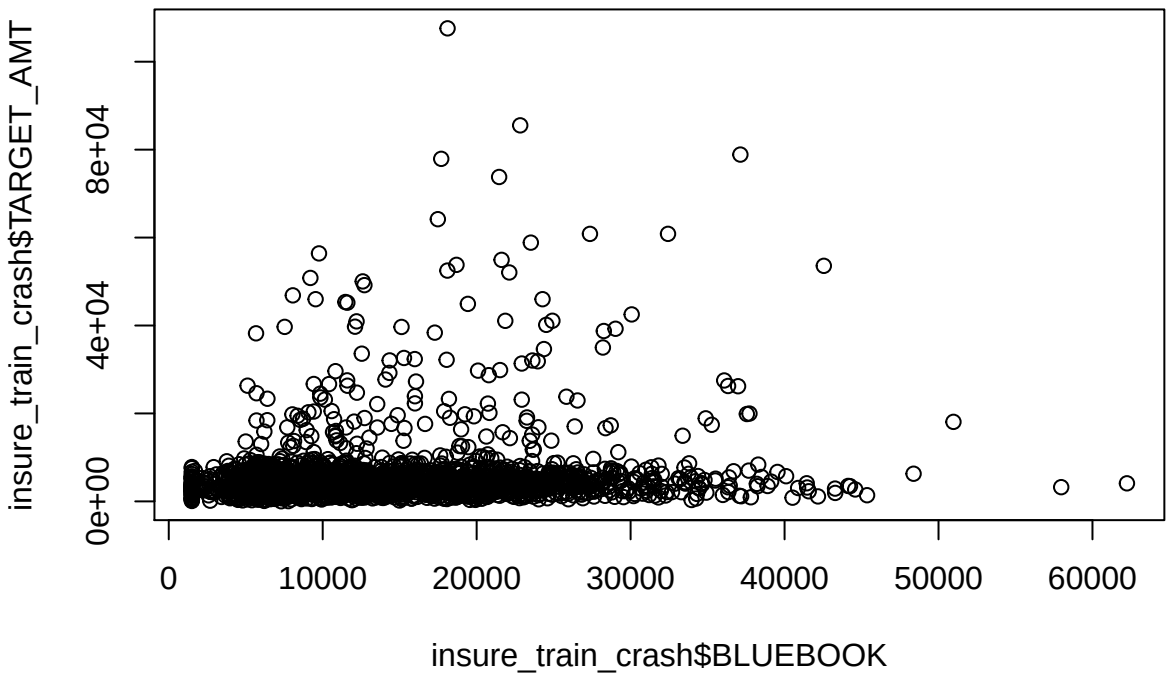
```
#show_hist("CAR_AGE")
plot(insure_train_crash$CAR_AGE, insure_train_crash$TARGET_AMT)
```



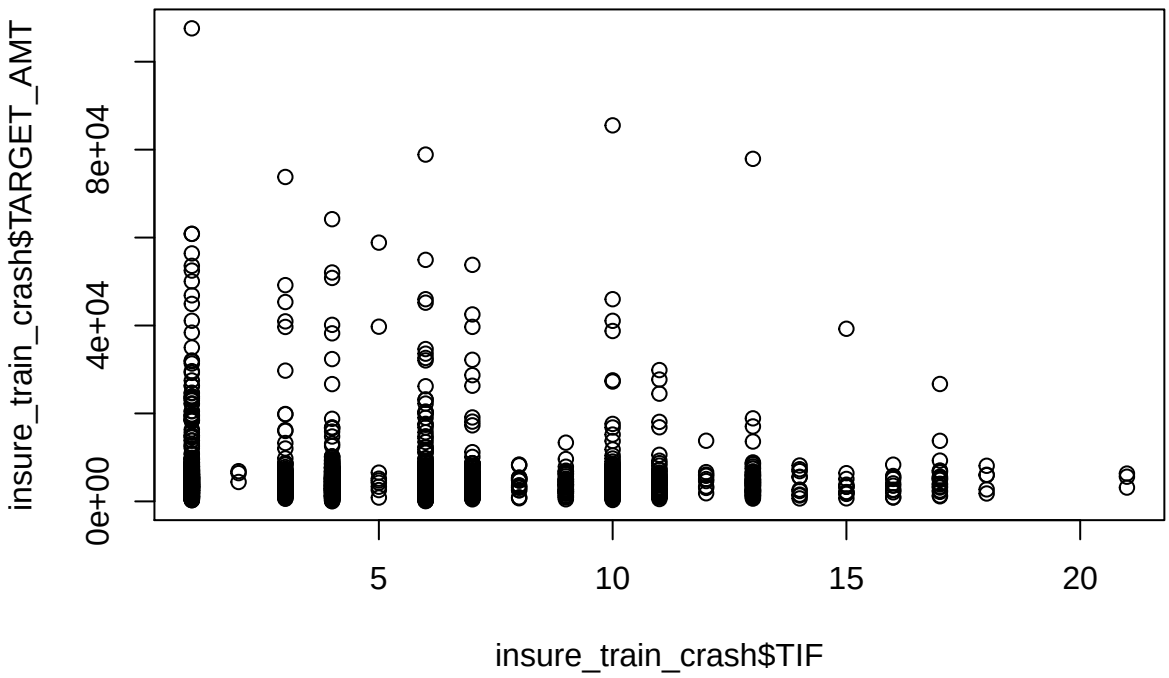

```
#check_bins("CAR_AGE", c(1:27))

#table(insure_train_full$AGE)
plot(insure_train_crash$AGE, insure_train_crash$TARGET_AMT)
```



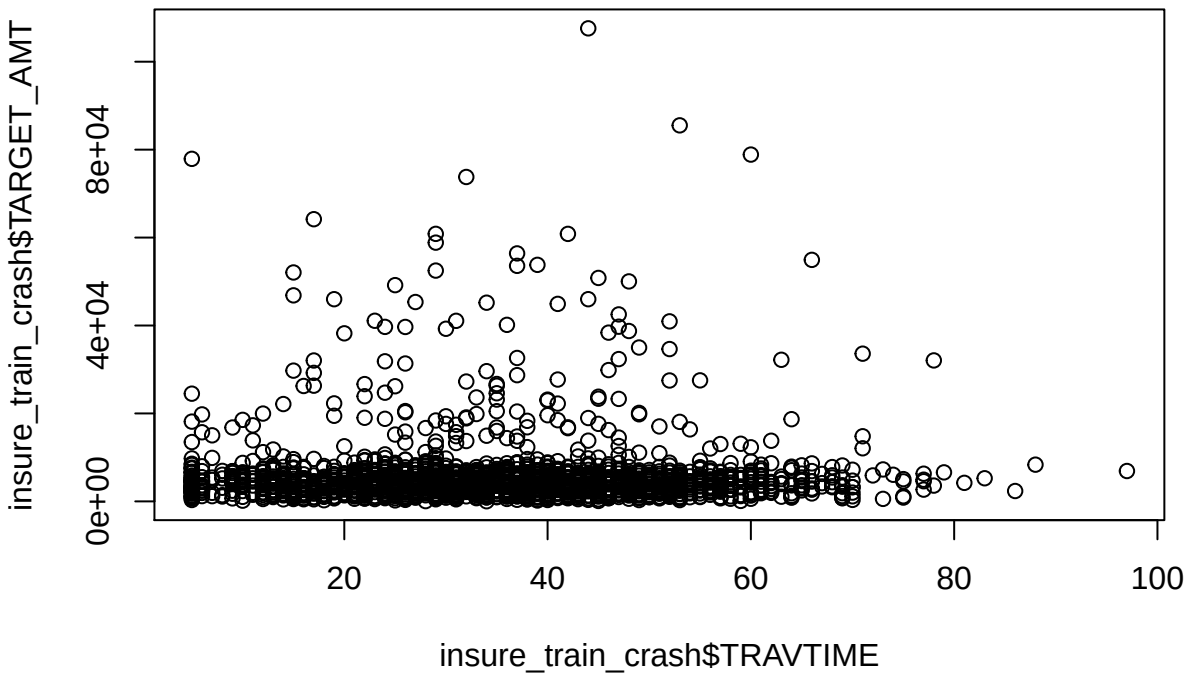

```
#show_hist("AGE")
#check_bins("AGE", c(16:80))

#table(insure_train_full$BLUEBOOK)
plot(insure_train_crash$BLUEBOOK, insure_train_crash$TARGET_AMT)
```




```
#show_hist("BLUEBOOK")
#check_bins("BLUEBOOK", c(5000, 10000, 20000, 30000, 45000, 57500, 58000))

# table(insure_train_full$TIF)
# show_hist("TIF")
plot(insure_train_crash$TIF, insure_train_crash$TARGET_AMT)
```




```
# check_bins("TIF", c(1, 4, 6, 10, 24))

#table(insure_train_full$TRAVTIME)
plot(insure_train_crash$TRAVTIME, insure_train_crash$TARGET_AMT)
```




```
#show_hist("TRAVTIME")
#check_bins("TRAVTIME", c(21, 59, 120))
```

From the outputs above, we can come to the following conclusions:

- INCOME - From the plot we can see that there is a marked difference in the chart at around 125000. We will use this value to bin this variable.
- YOJ - We can see that from 7 - 17 years, there is a visible change in the TARGET_AMT. We will use this bound to create the binned variable.
- HOME_VAL - We see from the plot 3 distinct segments - Between 0-10000, 60000-400000 and the rest. We will use these values to create 2 bins.
- OLDCLAIM- We can visualize 3 clusters in the data - 0-2000, 2000-10000, > 10000, We will use these values to create 2 bins.
- CLM_FREQ - Values less than 4 seem to have a positive correlation. We will use this value for binning.
- MVR_PTS - We can see from the plot that after 2, the TARGET_AMT starts decreasing. We will use this value for binning.
- CAR_AGE - There are quite a few records with a 1 year car age. We will use this bound to generate a binned variable as well as retain the original varible as is.
- AGE - There is no specific pattern that emerges in AGE. We will retain the variable as is.
- BLUEBOOK - There is no specific pattern that emerges. We will retain the variable as is.

- TIF - Looking at the plot we can conclude that this is not a good variable for binning. We will retain this variable as is.
- TRAVTIME - from the plot, we can see that there is a clear pattern around the value - 20. We will go ahead and create a binned variable for this.

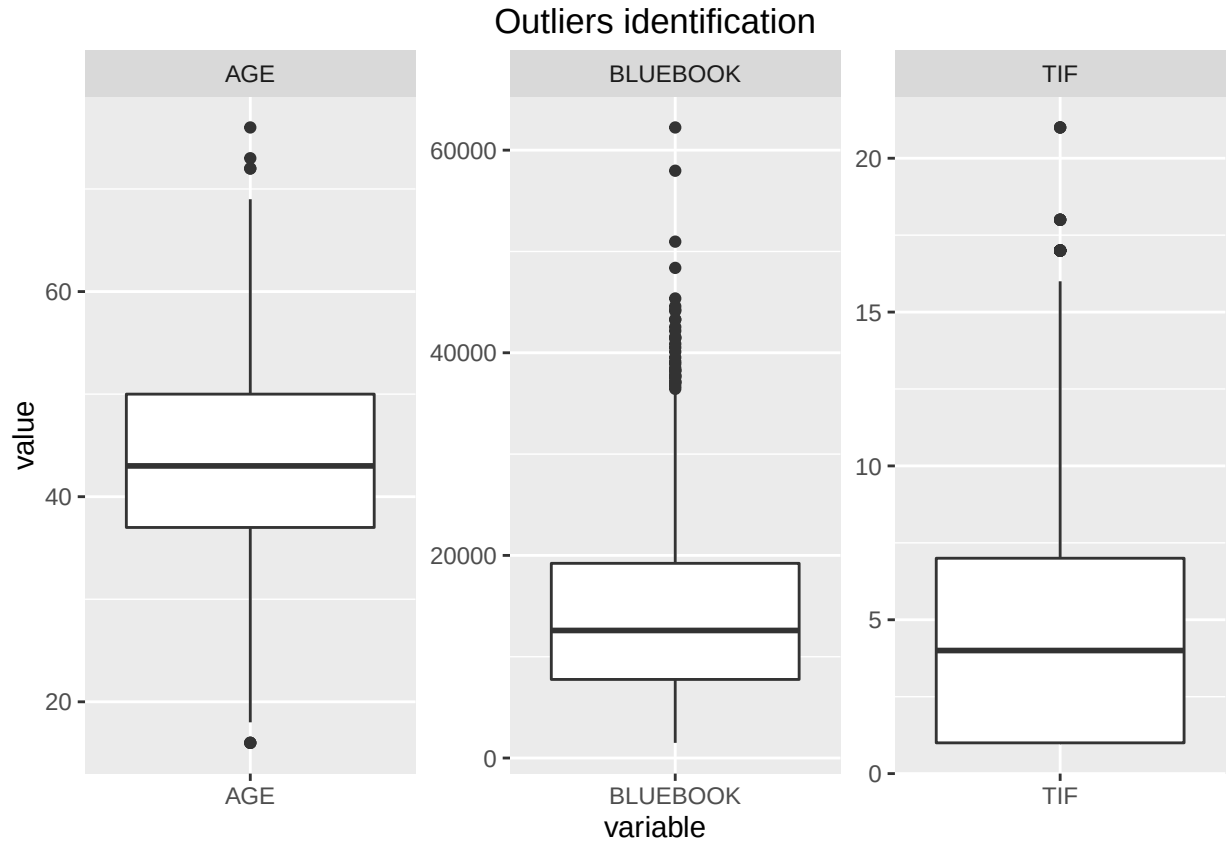
We will carry out the above transformations in the Data Preparation phase.

4.1.4 Outliers identification

In this sub-section, we will look at the boxplots and determine the outliers in variables and decide on whether to act on the outliers.

We will do the outliers only on the numeric variables: AGE, BLUEBOOK and TIF. The other variables will be binned and would not beed outlier handling.

Below are the plots:



From the "Outliers identification" plot above, we see that we have few outliers that we need to treat.

We see thatall the 3 variables need to be treated when we do the data preparation for modeling the TARGET_AMT.

4.2 Data Preparation

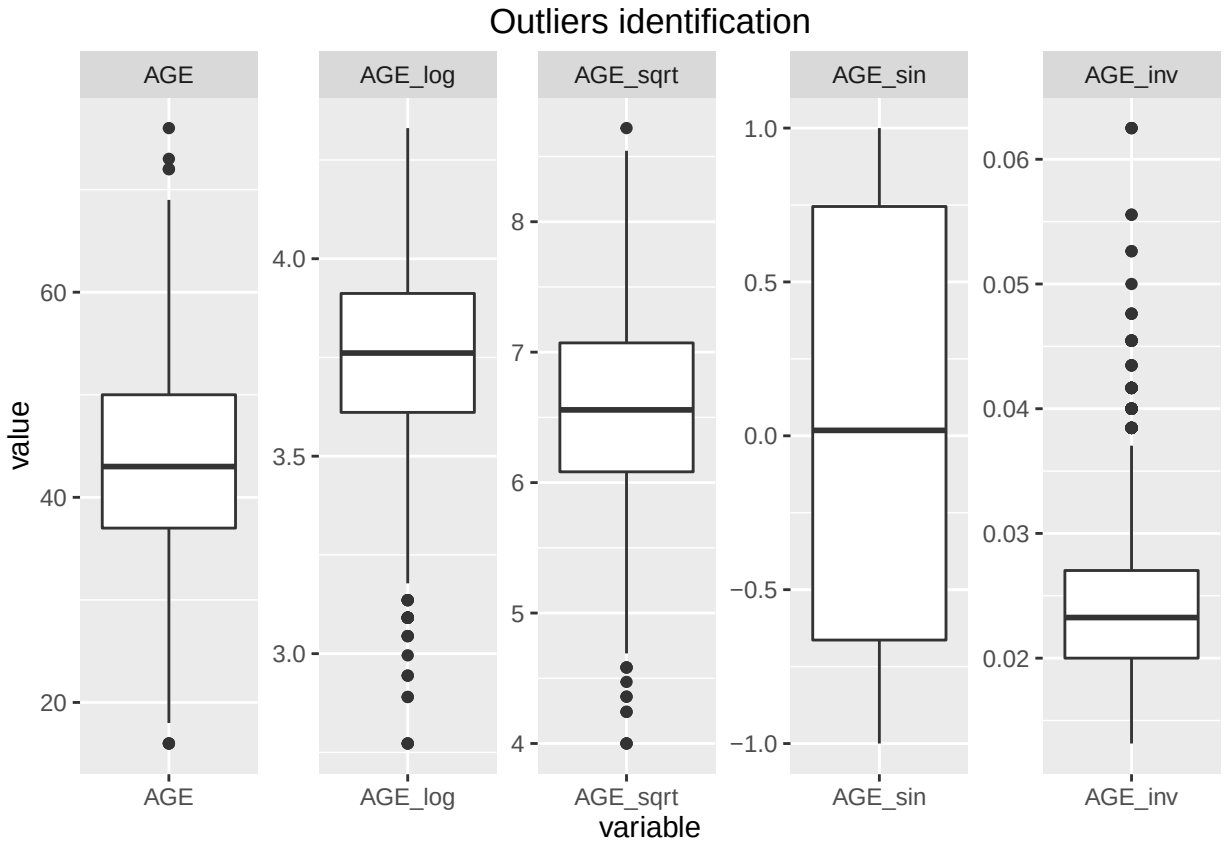
Now that we have completed the data exploration / analysis, we will be transforming the data for use in analysis and modeling.

We will be following the below steps as guidelines: - Outliers treatment - Adding New Variables

4.2.1 Outliers treatment

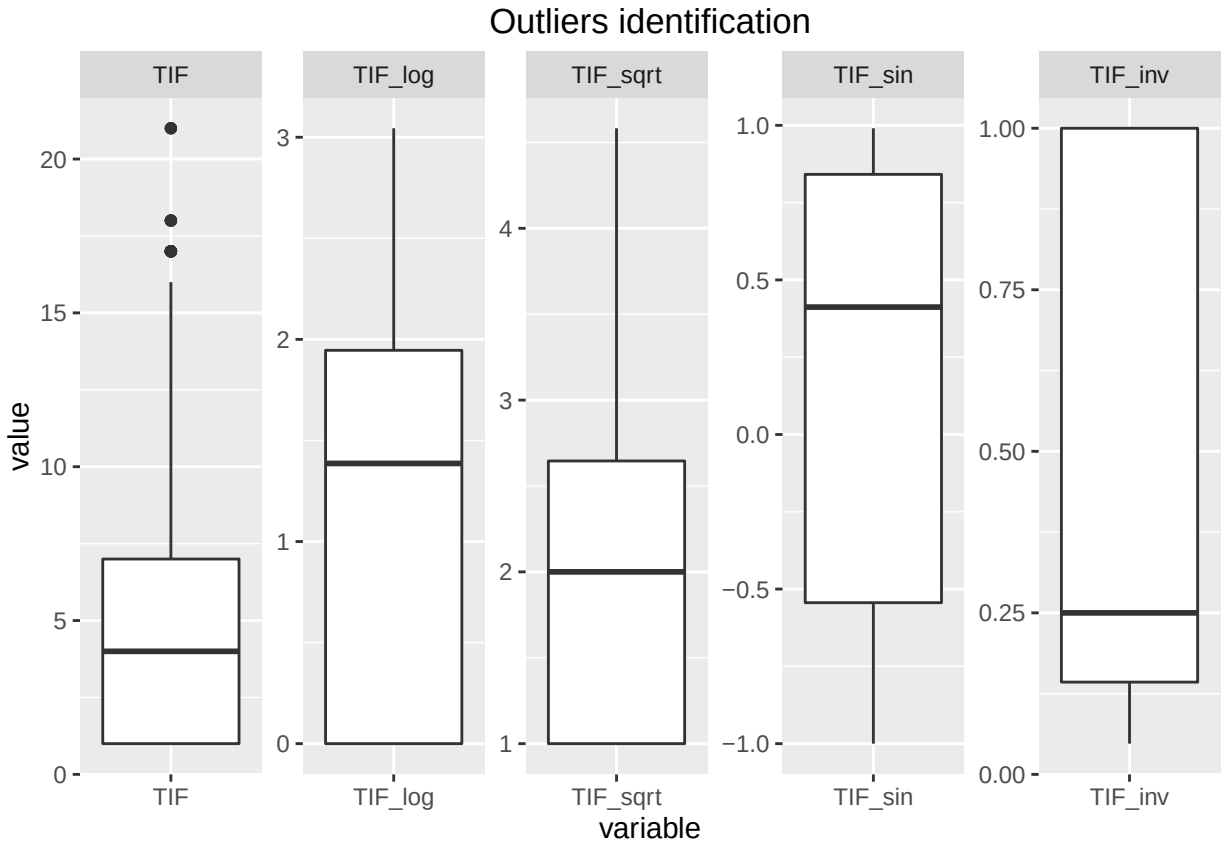
In this sub-section, we will check different transformations for each of the variables - AGE, BLUEBOOK, TIF - and create the appropriate outlier-handled / transformed variables.

Transformations for AGE



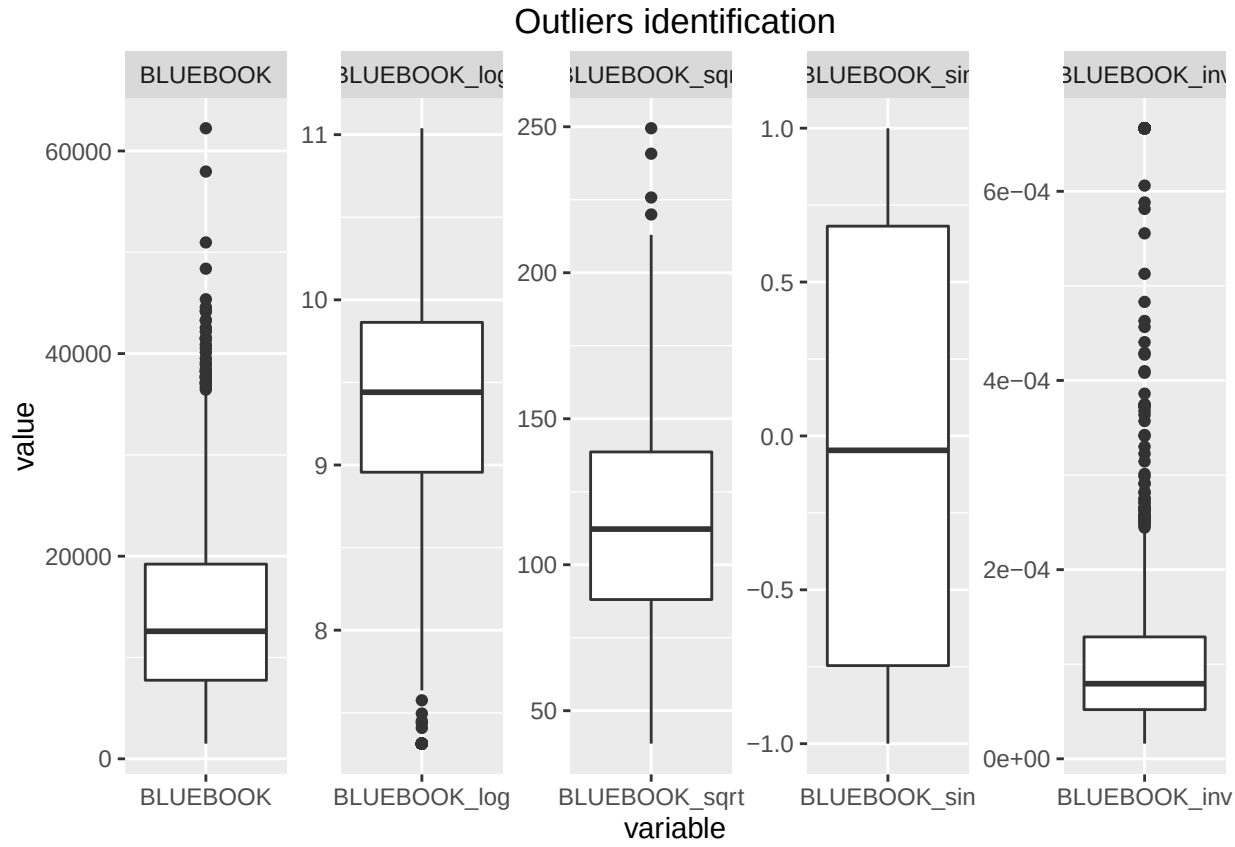
From the above charts we can see that a sin transformation works well for AGE. We will create this variable.

Transformations for TIF



From the above charts we can see that a log, sqrt, sin or an inverse transformation works well for TIF. However, a sin transformation seems to be more appropriate as it is well centered. Hence, We will create these variables.

Transformations for BLUEBOOK



From the above charts we can see that a sin transformation works well for BLUEBOOK. We will create these variables.

4.2.2 Adding New Variables

In this section, we generate some additional variables that we feel will help the correlations. The following were some of the observations we made during the data exploration phase for TARGET_AMT.

The following were some of the observations we made during the data exploration phase for TARGET_AMT

CAR_TYPE - If you drive Vans or Panel Trucks your cost of repair seems to increase as against Minivan, Pickup, Sports.Car, SUV. Since the distiction is clear, we believe that binning this variable accordingly will help strengthen the correlation.

Accordingly, we will bin these variables as below:

CAR_TYPE_AMT_BIN :

- 1 : if CAR_TYPE is Vans or Panel Trucks
- 0 : if CAR_TYPE is Pickups, Sports, SUVs or Minivans

```
insure_train_crash$CAR_TYPE_AMT_BIN <- ifelse(insure_train_crash$CAR_TYPE_Van | insure_train_crash$CAR_T
```

EDUCATION - If you have only a high school education then your cost of repair is less compared to a Bachelors, Masters or a Phd. Again binning this variable will strengthen the correlation.

Accordingly, we will bin these variables as below:

EDUCATION_AMT_BIN :

- 1 : if EDUCATION is High School
- 0 : if EDUCATION is Bachelors, Masters or Phd

```
insure_train_crash$EDUCATION_AMT_BIN <- ifelse(insure_train_crash$EDUCATION_High.School, 1, 0)
```

JOB - If you are a Lawyer, Professional, in a Blue Collar job or the job is unknown, you spend more on repairs as compared to a Doctor, Manager, Home Maker, Student, or Clerical job. Again binning this variable will strengthen the correlation.

Accordingly, we will bin these variables as below:

JOB_TYPE_AMT_BIN :

- 1 : if JOB_TYPE is Lawyer, Professional, Unknown or in a Blue Collar
- 0 : if JOB_TYPE is Doctor, Manager, Home Maker, Student, or Clerical

```
insure_train_crash$JOB_TYPE_AMT_BIN <- ifelse(insure_train_crash$JOB_Lawyer |  insure_train_crash$JOB_P
```

- INCOME - From the plot we can see that there is a marked difference in the chart at around 125000. We will use this value to bin this variable.

INCOME_AMT_BIN :

- 1 : if INCOME <= 125000
- 0 : if INCOME > 125000
- YOJ - We can see that from 7 - 17 years, there is a visible change in the TARGET_AMT. We will use this bound to create the binned variable.

YOJ_AMT_BIN :

- 1 : if YOJ >=7 and YOJ<= 17
- 0 : ELSE 0
- HOME_VAL - We see from the plot 3 distinct segments - Between 0-10000, 60000-400000 and the rest. We will use these values to create 2 bins.

HOME_VAL_AMT_0_10K_BIN :

- 1 : if HOME_VAL >=0 and HOME_VAL<= 10000
- 0 : ELSE 0

HOME_VAL_AMT_60K_400K_BIN :

- 1 : if HOME_VAL >=60000 and HOME_VAL<= 400000
- 0 : ELSE 0
- OLDCLAIM- We can visualize 3 clusters in the data - 0-2000, 2000-10000, > 10000, We will use these values to create 2 bins.

OLDCLAIM_AMT_0_2K_BIN :

- 1 : if OLDCLAIM >=0 and OLDCLAIM<= 2000
- 0 : ELSE 0

OLDCLAIM_AMT_2K_10K_BIN :

- 1 : if OLDCLAIM >=2000 and OLDCLAIM<= 10000
- 0 : ELSE 0
- CLM_FREQ - Values less than 4 seem to have a positive correlation. We will use this value for binning.

CLM_FREQ_AMT_BIN :

- 1 : if CLM_FREQ < 4
- 0 : if CLM_FREQ >= 4
- MVR_PTS - We can see from the plot that after 2, the TARGET_AMT starts decreasing. We will use this value for binning.

MVR_PTS_AMT_BIN :

- 1 : if MVR_PTS <=2
- 0 : if MVR_PTS > 0
- CAR_AGE - There are quite a few records with a 1 year car age. We will use this bound to generate a binned variable as well as retain the original varible as is.

CAR_AGE_AMT_BIN :

- 1 : if CAR_AGE <= 1
- 0 : if CAR_AGE > 0
- TRAVTIME - from the plot, we can see that there is a clear pattern around the value - 20. We will go ahead and create a binned variable for this.

TRAVTIME_AMT_BIN :

- 1 : if TRAVTIME <= 20
- 0 : if TRAVTIME > 0

4.2.3 Additional Binned Variables

Next we will create some more additional binned variables.

```
#write.csv(insure_train_crash, file = "D:/CUNY/Courses/Business Analytics and Data Mining/Assignments/d

DS_TARGET_AMT <- insure_train_crash

#DS_TARGET_AMT <- select(insure_train_crash, -AGE, -BLUEBOOK, -CAR_AGE, -CAR_TYPE_Minivan, -CAR_TYPE_Pa
```

4.3 Build Models

Now that we have the dataset in a shape that can be modeled, we will go ahead and train the model for TARGET_AMT. We will train 2 models and select the best among these 2 models. The following will be the model specifications:

- Model1 (Original Variables with cleanup) - This will use the standard lm for building the model. We will use only those variables that were part of the original set of variables.
- Model2 - (Original Variables + Additional Binned variables) - This will use the standard lm for building the model. We will use all the variables that were part of the original set of variables and the additional variables that we created as part of the binning for TARGET_AMT. We will then step thru the model to refine it further.

We will then generate inferences from these models.

4.3.1 Prepare TRAIN and VALID datasets

However, prior to that, we hold out a subset of data as a validation dataset to check model performance. This will be useful when we select a model.

```
smp_size <- floor(0.80 * nrow(DS_TARGET_AMT))

## set the seed to make your partition reproductible
set.seed(123)

train_index <- sample(seq_len(nrow(DS_TARGET_AMT)), size = smp_size)

DS_TARGET_AMT_TRAIN<- DS_TARGET_AMT[train_index, ]
DS_TARGET_AMT_VALID <- DS_TARGET_AMT[-train_index, ]
```

4.3.2 Model 1

In this model, we will be using the stadard lm modeling technique. We will use only those variables that were part of the original set of variables.

```
##
## Call:
## lm(formula = TARGET_AMT ~ . - TARGET_FLAG, data = na.omit(insure_orig))
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5824  -1702   -767    347 103577
##
## Coefficients: (3 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.310e+03  6.026e+02   2.173 0.029779 *
## KIDSDRIV               3.124e+02  1.134e+02   2.756 0.005861 **
## AGE                    5.170e+00  7.070e+00   0.731 0.464607
## HOMEKIDS               7.942e+01  6.542e+01   1.214 0.224771
## YOJ                   -4.526e+00  1.510e+01  -0.300 0.764429
## INCOME                -4.434e-03  1.803e-03  -2.458 0.013975 *
```

```
## HOME_VAL              -5.547e-04  5.912e-04  -0.938 0.348080
## TRAVTIME               1.202e+01  3.225e+00   3.727 0.000195 ***
## BLUEBOOK               1.426e-02  8.632e-03   1.653 0.098467 .
## TIF                   -4.822e+01  1.219e+01  -3.956 7.68e-05 ***
## OLDCLAIM              -1.040e-02  7.443e-03  -1.398 0.162197
## CLM_FREQ               1.409e+02  5.505e+01   2.559 0.010529 *
## MVR_PTS                1.752e+02  2.595e+01   6.751 1.57e-11 ***
## CAR_AGE               -2.772e+01  1.279e+01  -2.167 0.030230 *
## CAR_USE_Commercial     7.549e+02  1.578e+02   4.785 1.74e-06 ***
## MSTATUS_Yes           -5.745e+02  1.450e+02  -3.963 7.47e-05 ***
## PARENT1_Yes            5.717e+02  2.021e+02   2.829 0.004684 **
## RED_CAR_yes           -4.863e+01  1.492e+02  -0.326 0.744509
## REVOKED_Yes            5.489e+02  1.736e+02   3.162 0.001574 **
## SEX_M                  3.720e+02  1.840e+02   2.022 0.043259 *
## URBANICITY_Rural      -1.665e+03  1.395e+02 -11.939  < 2e-16 ***
## EDUCATION_Bachelors   -5.553e+02  2.886e+02  -1.924 0.054369 .
## EDUCATION_High.School -3.656e+02  3.222e+02  -1.135 0.256427
## EDUCATION_Masters     -2.649e+02  2.543e+02  -1.042 0.297655
## EDUCATION_PhD                 NA         NA      NA       NA
## JOB_Blue.Collar        5.244e+02  3.212e+02   1.633 0.102569
## JOB_Clerical           5.398e+02  3.415e+02   1.581 0.113997
## JOB_Doctor            -5.077e+02  4.088e+02  -1.242 0.214358
## JOB_Home.Maker         3.472e+02  3.644e+02   0.953 0.340764
## JOB_Lawyer             2.211e+02  2.950e+02   0.749 0.453583
## JOB_Manager           -4.844e+02  2.881e+02  -1.681 0.092767 .
## JOB_Professional       4.469e+02  3.084e+02   1.449 0.147298
## JOB_Student            2.947e+02  3.738e+02   0.788 0.430530
## JOB_Unknown                   NA         NA      NA       NA
## CAR_TYPE_Minivan      -5.226e+02  2.126e+02  -2.459 0.013970 *
## CAR_TYPE_Panel.Truck  -2.401e+02  2.629e+02  -0.913 0.361128
## CAR_TYPE_Pickup       -1.336e+02  2.214e+02  -0.604 0.546172
## CAR_TYPE_Sports.Car    5.007e+02  2.951e+02   1.696 0.089848 .
## CAR_TYPE_SUV           2.289e+02  2.666e+02   0.859 0.390607
## CAR_TYPE_Van                  NA         NA      NA       NA
## YOJ_MISS               2.444e+02  2.201e+02   1.110 0.266821
## INCOME_MISS           -1.487e+01  2.224e+02  -0.067 0.946704
## HOME_VAL_MISS          7.472e+01  2.178e+02   0.343 0.731519
## CAR_AGE_MISS           6.205e+00  2.085e+02   0.030 0.976257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4546 on 8116 degrees of freedom
## Multiple R-squared:  0.07115,    Adjusted R-squared:  0.06657
## F-statistic: 15.54 on 40 and 8116 DF,  p-value: < 2.2e-16
```

Interpretation of the Model

Based on the backward stepwise selection, below are the characteristics of the refined model :

- The Residual standard error is 4500
- Multiple R-squared: 0.07549
- Adjusted R-squared: 0.0703
- F-statistic: 14.53 on 36 and 6407 DF
- p-value: < 2.2e-16

Table 8: Coefficients for the refined model 1

Coefficients
1309.6007428
312.4197609
5.1702497
79.4201735
-4.5263536
-0.0044338
-0.0005547
12.0168871
0.0142638
-48.2227268
-0.0104037
140.8589206
175.1617136
-27.7182174
754.8818232
-574.4596123
571.7354754
-48.6253999
548.9412355
371.9628103
-1664.9890333
-555.2646268
-365.6399212
-264.8523580
NA
524.3573176
539.7935709
-507.6600913
347.1909450
221.0938865
-484.3864004
446.9202139
294.6757847
NA
-522.5974902
-240.0680598
-133.5933497
500.6775127
228.8762923
NA
244.4025369
-14.8696866
74.7161324
6.2052721

Based on the above coefficients, we can see that some of the coefficients are counter-intutive to the Theoretical impact.

- TEAM_BATTING_H (-0.034), TEAM_BATTING_2B (-0.049), TEAM_FIELDING_DP (-0.112), TEAM_PITCHING_SO (-0.054) have a negative coefficient even though they are theoretically supposed

to have a positive impact on wins. This means that a unit change in each of these variables will decrease the number of a wins.

- Similarly, TEAM_BATTING_SO (0.033), TEAM_PITCHING_H (0.06) have a positive coefficient even though they are theoretically supposed to have a negative impact on wins. This means that a unit change in each of these variables will increase the number of a wins.
- TEAM_BATTING_3B (0.183), TEAM_BATTING_HR (0.1), TEAM_BATTING_BB (0.118), TEAM_BASERUN_SB (0.069), TEAM_FIELDING_E (-0.119), TEAM_PITCHING_BB (-0.08) have the intended theoretical impact on wins. This means that a unit change in each of these variables will either decrease or increase the number of a wins as intended by the theoretical impact.

Since we have already seen this result in our data exploration phase, we will retain this model as is for comparision with other models.

4.2.3 Model 2

In this model, we will be using the rpart package to do a tree based regression. We will create model and carry out further pruning for the tree.

Interpretation of the Model

Based on the backward stepwise selection, below are the characteristics of the refined model :

- The Residual standard error is 10.18
- Multiple R-squared: 0.4058
- Adjusted R-squared: 0.4019
- F-statistic: 103.7 on 12 and 1822 DF
- p-value: $<$ 2.2e-16

Based on the above coefficients, we can see that some of the coefficients are counter-intutive to the Theoretical impact.

- TEAM_BATTING_H (-0.034), TEAM_BATTING_2B (-0.049), TEAM_FIELDING_DP (-0.112), TEAM_PITCHING_SO (-0.054) have a negative coefficient even though they are theoretically supposed to have a positive impact on wins. This means that a unit change in each of these variables will decrease the number of a wins.
- Similarly, TEAM_BATTING_SO (0.033), TEAM_PITCHING_H (0.06) have a positive coefficient even though they are theoretically supposed to have a negative impact on wins. This means that a unit change in each of these variables will increase the number of a wins.
- TEAM_BATTING_3B (0.183), TEAM_BATTING_HR (0.1), TEAM_BATTING_BB (0.118), TEAM_BASERUN_SB (0.069), TEAM_FIELDING_E (-0.119), TEAM_PITCHING_BB (-0.08) have the intended theoretical impact on wins. This means that a unit change in each of these variables will either decrease or increase the number of a wins as intended by the theoretical impact.

Since we have already seen this result in our data exploration phase, we will retain this model as is for comparision with other models.

4.4 Model Evaluation Using VALID Data

4.4.1 Evaluation of Model 1

4.4.2 Evaluation of Model 2

3.4.3 Final Linear Model Selection Summary

5 Prediction Using Evaluation Data

Now that we have selected the final models for both the TARGET_FLAG and the TARGET_AMT, we will go ahead and use these models to predict the results for the evaluation dataset. After transforming the data to meet the needs of the trained models, we will apply the models in 2 steps.

Step 1 - Here we use the transformed evaluation dataset to predict for the TARGET_FLAG using the requisite predictors.

Step 2 - Once we have the prediction for the TARGET_FLAG, we will filter this data for only those rows that were predicted for a CRASH. We then use this smaller dataset to precict for the TARGET_AMT.

5.1 Tranformation of Evaluation Data

First we need to transform the evaluation dataset to account for all the predictors that were used in both the models.

5.2 Model Output for Logistic Regression

We now apply the final Logistic regression model that was trained for predicting the TARGET_FLAG. Below are the predictions.

5.3 Model Output for Linear Regression

Next we filter for the "predicted" crashes.

Next, we apply the final linear model to this smaller dataset and see the results.

Table 9: Outcome on evaluation data set

INDEX	TARGET_FLAG	TARGET_FLAG_prob	TARGET_AMT
62	1	0.6180450	3040.482
186	1	0.6048017	3164.902
195	1	0.5406679	2815.494
241	1	0.6238824	2783.981
281	1	0.8111059	4246.278
308	1	0.5982238	2454.777
348	1	0.8113404	3964.364
350	1	0.6081842	3721.647
367	1	0.6748337	3424.494
376	1	0.6944851	3397.645