Home Work Assignment - 04

Critical Thinking Group 5

Contents

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 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 to build many binary logistic regression models (to predict if a person will crash the car) and also some linear regression models (to predict the amount of money it will take to fix the car after crashing). 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 following the below best practice steps and guidelines:
1 -Data Exploration
2 -Data Preparation
3 -Build Models
4 -Select Models

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 do this once all the data transformations are complete and we are ready to build the models.

While building and selecting models, We will deal with the problem in 2 parts:
- Part 1 - Here we build and select Binary Logistic Regression models using the training data set.
- Part 2 - Here we build and select Linear Regression models using only the "Crashed" data from the training data set.

1 Data Exploration Analysis

In section we will explore and gain some insights into the dataset by pursuing the below high level steps and inquiries:
-Variable identification
-Variable Relationships
-Data summary analysis
-Outliers and Missing Values Identification

1.1 Variable identification

First let's display and examine the data dictionary or the data columns as shown in table 1

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

VARIABLE_NAME	DEFINITION	THEORETICAL_EFFECT
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 probab
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.

```r
str(insure_train_full)
```

```
## '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 28
##  $ 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 553 802 746 2672 701 135 85
##  $ 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 ...
```

```r
levels(insure_train_full$MSTATUS)
```

```
## [1] "Yes"  "z_No"
```

```r
levels(insure_train_full$SEX)
```

```
## [1] "M"   "z_F"
```

```r
levels(insure_train_full$EDUCATION)
```

```
## [1] "<High School"  "Bachelors"     "Masters"       "PhD"
## [5] "z_High School"
```

```r
levels(insure_train_full$JOB)
```

```
## [1] ""               "Clerical"      "Doctor"        "Home Maker"
## [5] "Lawyer"         "Manager"       "Professional"  "Student"
## [9] "z_Blue Collar"
```

```r
levels(insure_train_full$CAR_TYPE)
```

```
## [1] "Minivan"     "Panel Truck" "Pickup"      "Sports Car"  "Van"
## [6] "z_SUV"
```

```r
levels(insure_train_full$URBANICITY)
```

```
## [1] "Highly Urban/ Urban"   "z_Highly Rural/ Rural"
```

```r
levels(insure_train_full$REVOKED)
```

```
## [1] "No"  "Yes"
```

From the output above we can make the following observations:

- some numeric variables like INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM have been converted to Factor variables. This needs to be set right.
- Some of the variables like MSTATUS, SEX, EDUCATION, JOB, CAR_TYPE, URBANICITY have some of the values encoded with "z_". Not that this will impact the analysis, but it will look a bit odd. So we will be fixing this.

- EDUCATION has 2 “High School” values - one starting with “<” and another starting with “z_”. It is assumed that both these values are to be converted to “HIGH School”.
- JOB has a “” value. This needs to be replaced with NA.
- We will also create dummy variables for all the factors.
- Please note that we will not be using INDEX variable as it serves as just an identifier for each row. And has no relationships to other variables.

Making the above fixes to the data, we now have a “clean” dataset which can be explored further.

```
#- some numeric variables like INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM have been converted to Factor varia

insure_train_full$INCOME <- as.numeric(insure_train_full$INCOME)
insure_train_full$HOME_VAL <- as.numeric(insure_train_full$HOME_VAL)
insure_train_full$BLUEBOOK <- as.numeric(insure_train_full$BLUEBOOK)
insure_train_full$OLDCLAIM <- as.numeric(insure_train_full$OLDCLAIM)

#- Some of the variables like MSTATUS, SEX, EDUCATION, JOB, CAR_TYPE, URBANICITY have some of the value

#- EDUCATION has 2 "High School" values - one starting with "<" and another starting with "z_". It is a

#- JOB has a "" value. This needs to be replaced with NA.

insure_train_full$MSTATUS <- as.factor(str_replace_all(insure_train_full$MSTATUS, "z_", ""))
insure_train_full$SEX <- as.factor(str_replace_all(insure_train_full$SEX, "z_", ""))
insure_train_full$EDUCATION <- as.factor(str_replace_all(insure_train_full$EDUCATION, "z_", ""))
insure_train_full$EDUCATION <- as.factor(str_replace_all(insure_train_full$EDUCATION, "<", ""))
insure_train_full$CAR_TYPE <- as.factor(str_replace_all(insure_train_full$CAR_TYPE, "z_", ""))
insure_train_full$URBANICITY <- as.factor(str_replace_all(insure_train_full$URBANICITY, "z_", ""))

insure_train_full$JOB[insure_train_full$JOB==""] <- NA
insure_train_full$JOB <- as.factor(str_replace_all(insure_train_full$JOB, "z_", ""))

#- We will also create dummy variables for all the factors and drop the original variables.

dummy_vars<-as.data.frame(sapply(dummy(insure_train_full), FUN = as.numeric))
dummy_vars <- dummy_vars-1

# EDU_ = factor(insure_train_full$EDUCATION)
# dummies = model.matrix(~EDU_)

insure_train_full <- cbind(select(insure_train_full, -PARENT1, -MSTATUS, -SEX, -EDUCATION, -JOB, -CAR_US

# insure_train_full <- cbind(insure_train_full, dummy_vars)

# - Please note that we will not be using INDEX variable as it serves as just an identifier for each ro

insure_train_full <- select(insure_train_full, -INDEX)
```

1.2 Variable Relationships

Since we have 2 models to build, we have 2 sets of assumptions to be checked:

- Logistic Regression for TARGET_FLAG:
  - The dependent variable need not to be normally distributed
  - Errors need to be independent but not normally distributed.
  - We will be using GLM and GLM does not assume a linear relationship between dependent and independent variables. However, it assumes a linear relationship between link function and independent variables in logit model.
  - Also does not use OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE)

NEED TO ADD SOME POINTS

- Linear Regression for TARGET_AMT:
  - The dependent variable is normally distributed
  - Errors are independent and normally distributed.

In next step below relationship between the target variable and dependent variables is shown in three charts.

1.3 Data Summary Analysis

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 in table 2.

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

First lets see the correlation for TARGET_FLAG:

Table 2: Correlation between TARGET_FLAG and predictor variables

	Correlation_TARGET_FLAG
TARGET_FLAG	1.0000000
TARGET_AMT	0.5342461
URBANICITY_Highly.Urban..Urban	0.2242509
MVR_PTS	0.2191971
CLM_FREQ	0.2161961
OLDCLAIM	0.1902875
PARENT1_Yes	0.1576222
REVOKED_Yes	0.1519391
CAR_USE_Commercial	0.1426737
EDUCATION_High.School	0.1380116
MSTATUS_No	0.1351248
HOMEKIDS	0.1156210
JOB_Blue.Collar	0.1057869
KIDSDRIV	0.1036683
JOB_Student	0.0795874

	Correlation_TARGET_FLAG
CAR_TYPE_Sports.Car	0.0572528
CAR_TYPE_Pickup	0.0566433
BLUEBOOK	0.0504453
TRAVTIME	0.0483683
CAR_TYPE_SUV	0.0450322
JOB_Clerical	0.0280954
SEX_F	0.0210786
JOB_Home.Maker	0.0114210
RED_CAR_no	0.0069473
CAR_TYPE_Van	0.0030204
CAR_TYPE_Panel.Truck	-0.0003424
RED_CAR_yes	-0.0069473
SEX_M	-0.0210786
INCOME	-0.0338365
JOB_Professional	-0.0404212
EDUCATION_Bachelors	-0.0426526
JOB_Doctor	-0.0605425
JOB_Lawyer	-0.0643342
EDUCATION_PhD	-0.0654121
YOJ	-0.0705118
EDUCATION_Masters	-0.0762960
TIF	-0.0823700
CAR_AGE	-0.1006506
AGE	-0.1032167
JOB_Manager	-0.1097548
MSTATUS_Yes	-0.1351248
CAR_TYPE_Minivan	-0.1369991
CAR_USE_Private	-0.1426737
HOME_VAL	-0.1485715
REVOKED_No	-0.1519391
PARENT1_No	-0.1576222
URBANICITY_Highly.Rural..Rural	-0.2242509

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, 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 and URBANICITY_Highly.Urban..Urban have a positive correlation.
Similarly, URBANICITY_Highly.Rural..Rural, PARENT1_No, REVOKED_No, HOME_VAL, CAR_USE_Private, 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 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 or professional. Again binning this variable will strengthen the correlation.

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

Next lets look at the correlation for TARGET_AMT.
Prior to this, we need to filter for only those records where there has been a crash. The amount incurred is relevant only when there is a crash. We then look at the correlations.

We now use the TARGET_AMT to check the correlations.

Table 3: Correlation between TARGET_AMT and predictor variables

	Correlation_TARGET_AMT
TARGET_AMT	1.0000000
CAR_TYPE_Panel.Truck	0.0683507
SEX_M	0.0515580
CAR_TYPE_Van	0.0500059
CAR_USE_Commercial	0.0498471
JOB_Professional	0.0464096
MVR_PTS	0.0398112
REVOKED_No	0.0369658
MSTATUS_No	0.0349259
YOJ	0.0342552
EDUCATION_PhD	0.0295373
AGE	0.0278814
RED_CAR_yes	0.0273333
JOB_Blue.Collar	0.0241018
PARENT1_Yes	0.0239623
INCOME	0.0230938
EDUCATION_Masters	0.0144327
JOB_Lawyer	0.0140092
EDUCATION_Bachelors	0.0132068
HOME_VAL	0.0062782
TRAVTIME	0.0051768
URBANICITY_Highly.Urban..Urban	0.0048280
CLM_FREQ	0.0019608
HOMEKIDS	0.0004689
KIDSDRIV	0.0000184
URBANICITY_Highly.Rural..Rural	-0.0048280
CAR_TYPE_Minivan	-0.0057109
TIF	-0.0060110
BLUEBOOK	-0.0073838
OLDCLAIM	-0.0095204
JOB_Clerical	-0.0109708
JOB_Doctor	-0.0115909
CAR_AGE	-0.0130178
CAR_TYPE_Sports.Car	-0.0151607
CAR_TYPE_Pickup	-0.0178884

	Correlation_TARGET_AMT
PARENT1_No	-0.0239623
JOB_Manager	-0.0241998
RED_CAR_no	-0.0273333
JOB_Home.Maker	-0.0277648
JOB_Student	-0.0309653
MSTATUS_Yes	-0.0349259
EDUCATION_High.School	-0.0356610
REVOKED_Yes	-0.0369658
CAR_TYPE_SUV	-0.0403806
CAR_USE_Private	-0.0498471
SEX_F	-0.0515580

The above table suggests that none of the variables seem to have a very strong correlation with TARGET_AMT. KIDSDRIV, HOMEKIDS, CLM_FREQ, URBANICITY_Highly.Urban..Urban, TRAVTIME, HOME_VAL, EDUCATION_Bachelors, JOB_Lawyer, EDUCATION_Masters, INCOME, PARENT1_Yes, JOB_Blue.Collar, RED_CAR_yes, AGE, EDUCATION_PhD, YOJ, MSTATUS_No, REVOKED_No, MVR_PTS, JOB_Professional, CAR_USE_Commercial, CAR_TYPE_Van, SEX_M, CAR_TYPE_Panel.Truck have a positive correlation.
Similarly, SEX_F, CAR_USE_Private, CAR_TYPE_SUV, REVOKED_Yes, EDUCATION_High.School, MSTATUS_Yes, JOB_Student, JOB_Home.Maker, RED_CAR_no, JOB_Manager, PARENT1_No, CAR_TYPE_Pickup, CAR_TYPE_Sports.Car, CAR_AGE, JOB_Doctor, JOB_Clerical, OLDCLAIM, BLUEBOOK, TIF, CAR_TYPE_Minivan, URBANICITY_Highly.Rural..Rural 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 or in a Blue Collar job, 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.

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

1.4 Missing Values and Outliers Identification

1.4.1 Missing Values

Based on the missing data from the below table, we can see that there are a few missing values for AGE, CAR_AGE, YOJ and JOB variables.

Table 4: Missing Values

	missings
TARGET_FLAG	0

	missings
TARGET_AMT	0
KIDSDRIV	0
AGE	6
HOMEKIDS	0
YOJ	454
INCOME	0
HOME_VAL	0
TRAVTIME	0
BLUEBOOK	0
TIF	0
OLDCLAIM	0
CLM_FREQ	0
MVR_PTS	0
CAR_AGE	510
PARENT1_No	0
PARENT1_Yes	0
MSTATUS_No	0
MSTATUS_Yes	0
SEX_F	0
SEX_M	0
EDUCATION_Bachelors	0
EDUCATION_High.School	0
EDUCATION_Masters	0
EDUCATION_PhD	0
JOB_Blue.Collar	526
JOB_Clerical	526
JOB_Doctor	526
JOB_Home.Maker	526
JOB_Lawyer	526
JOB_Manager	526
JOB_Professional	526
JOB_Student	526
CAR_USE_Commercial	0
CAR_USE_Private	0
CAR_TYPE_Minivan	0
CAR_TYPE_Panel.Truck	0
CAR_TYPE_Pickup	0
CAR_TYPE_Sports.Car	0
CAR_TYPE_SUV	0
CAR_TYPE_Van	0
RED_CAR_no	0
RED_CAR_yes	0
REVOKED_No	0
REVOKED_Yes	0
URBANICITY_Highly.Rural..Rural	0
URBANICITY_Highly.Urban..Urban	0

We can try and impute values to AGE, YOJ, CAR_AGE. However, we will not be able to impute values for JOB since this is a categorical variable. Though there are a few methods to do this imputation, it may not be worth it.

Lets see the impact if we have to exclude these missing records.

```
sum(!complete.cases(insure_train_full))
```

```
## [1] 1407
```

```
# Percentage of records

sum(!complete.cases(insure_train_full))/nrow(insure_train_full) *100
```

```
## [1] 17.24053
```

We see from the above that excluding the missing rows will remove about 17.24% of the records. This would not seem to have too much of an impact.

We will exclude the rows with the missing values when we do the data preparation / tranformations for the TARGET_FLAG dataset.

Similarly, based on the below analysis, we see that we are losing about 17.6% of the data for the TARGET_AMT dataset.

Table 5: Missing Values

	missings
TARGET_AMT	0
KIDSDRIV	0
AGE	5
HOMEKIDS	0
YOJ	123
INCOME	0
HOME_VAL	0
TRAVTIME	0
BLUEBOOK	0
TIF	0
OLDCLAIM	0
CLM_FREQ	0
MVR_PTS	0
CAR_AGE	142
PARENT1_No	0
PARENT1_Yes	0
MSTATUS_No	0
MSTATUS_Yes	0
SEX_F	0
SEX_M	0
EDUCATION_Bachelors	0
EDUCATION_High.School	0
EDUCATION_Masters	0
EDUCATION_PhD	0
JOB_Blue.Collar	136
JOB_Clerical	136
JOB_Doctor	136
JOB_Home.Maker	136

	missings
JOB_Lawyer	136
JOB_Manager	136
JOB_Professional	136
JOB_Student	136
CAR_USE_Commercial	0
CAR_USE_Private	0
CAR_TYPE_Minivan	0
CAR_TYPE_Panel.Truck	0
CAR_TYPE_Pickup	0
CAR_TYPE_Sports.Car	0
CAR_TYPE_SUV	0
CAR_TYPE_Van	0
RED_CAR_no	0
RED_CAR_yes	0
REVOKED_No	0
REVOKED_Yes	0
URBANICITY_Highly.Rural..Rural	0
URBANICITY_Highly.Urban..Urban	0

```
## [1] 379
```

```
## [1] 17.60334
```

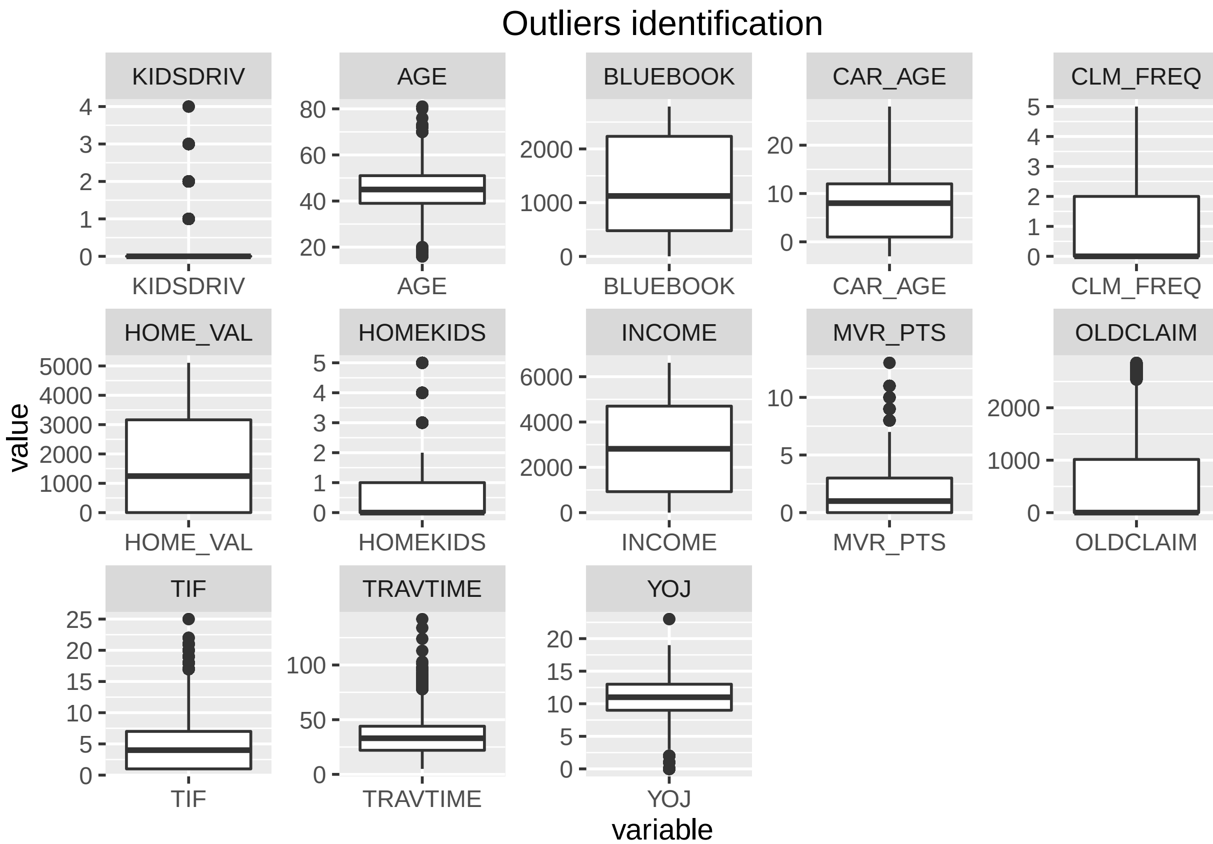
We will exclude the rows with the missing values when we do the data preparation / tranformations for the TARGET_AMT dataset.

1.4.2 Outliers identification

In this section univariate analysis is being carried out and boxplots diagrams are being used to determine the outliers in variables and decide on whether to act on the outliers.

We will do the outliers only on the numeric variables.

Below are the plots:

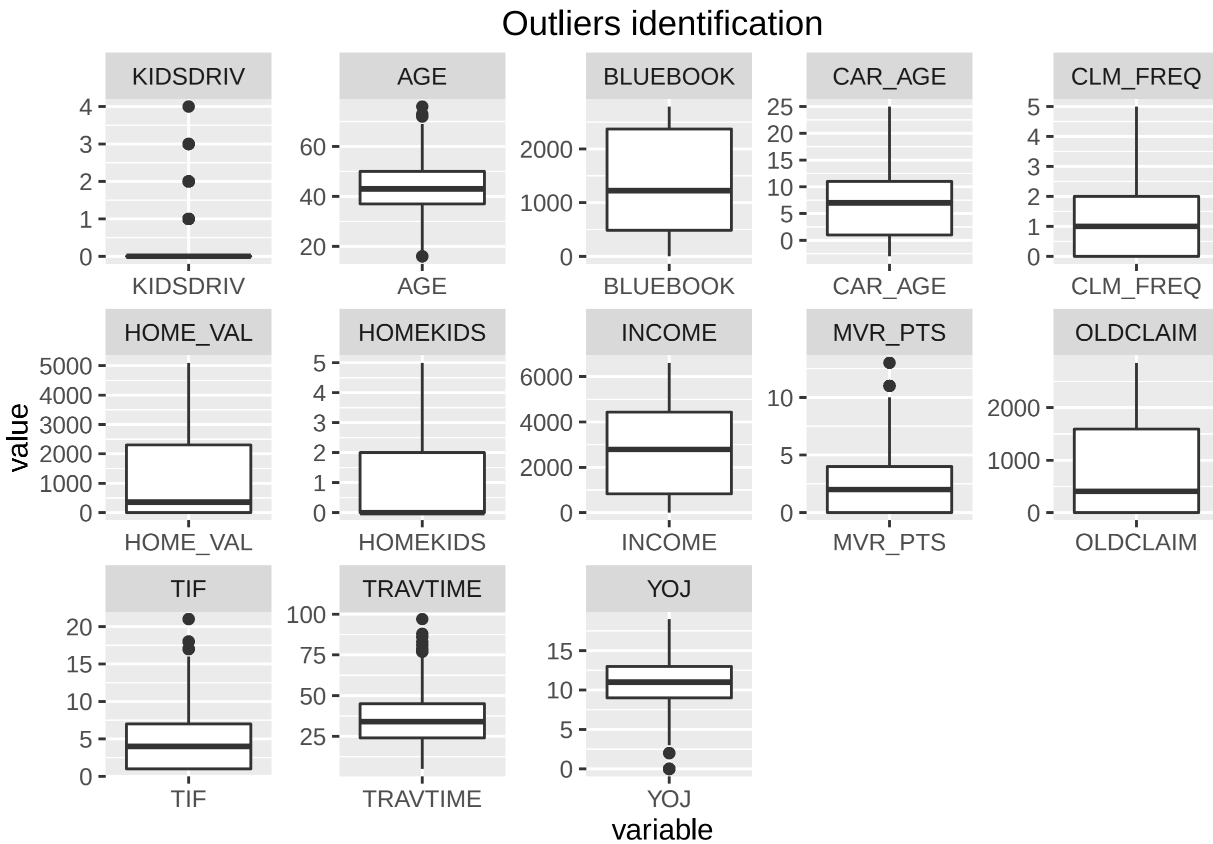


From the “Outliers identification” plot above, we see that we have few outliers that we need to treat. We see that: KIDSDRIV, AGE, HOMEKIDS, MVR_PTS, OLDCLAIM, TIF, TRAVTIME, YOJ need to be treated when we do the data preparation for modeling the TARGET_FLAG.

We carry out the same exercise for TARGET_AMT as well:

We will do the outliers only on the numeric variables.

Below are the plots:



From the "Outliers identification" plot above, we see that we have few outliers that we need to treat.
We see that: KIDSDRIV, AGE, MVR_PTS, TIF, TRAVTIME, YOJ need to be treated when we do the data preparation for modeling the TARGET_AMT.

1.5 Analysis 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:

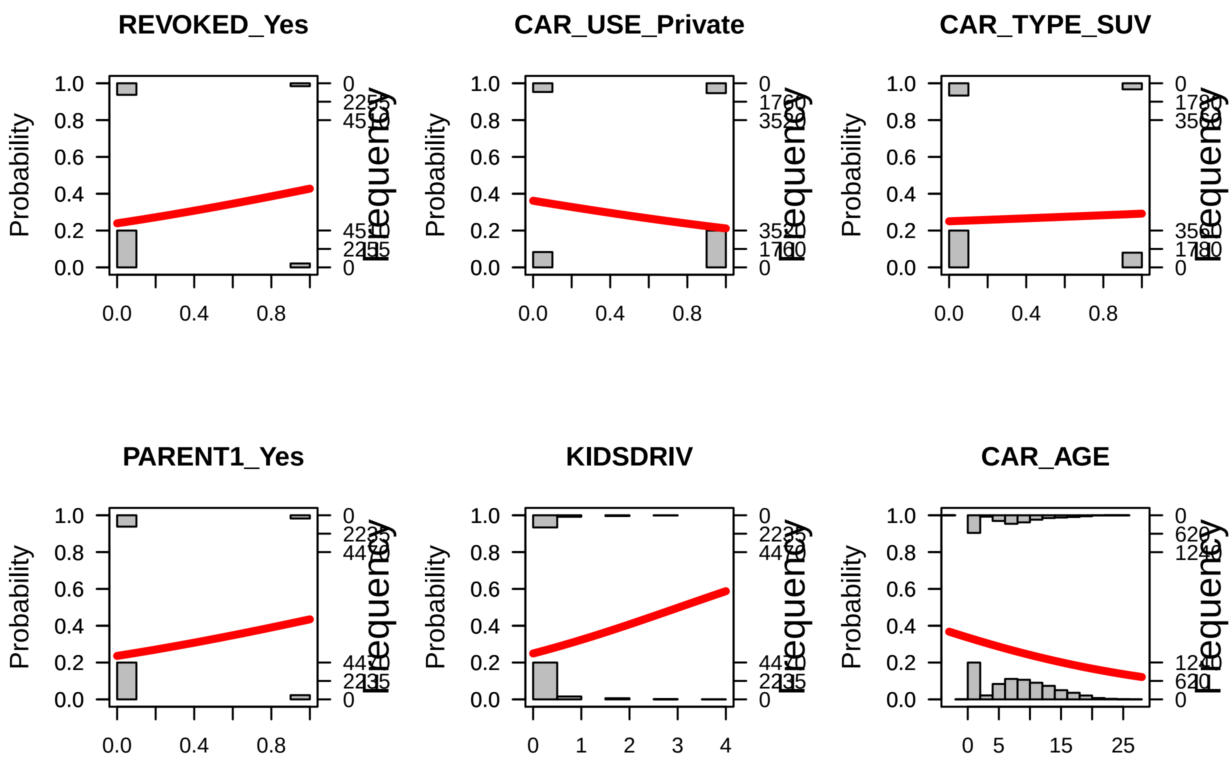
$g(E(y)) = a + Bx_1 + B_2x_2 + B_3x_3 + ...$

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.

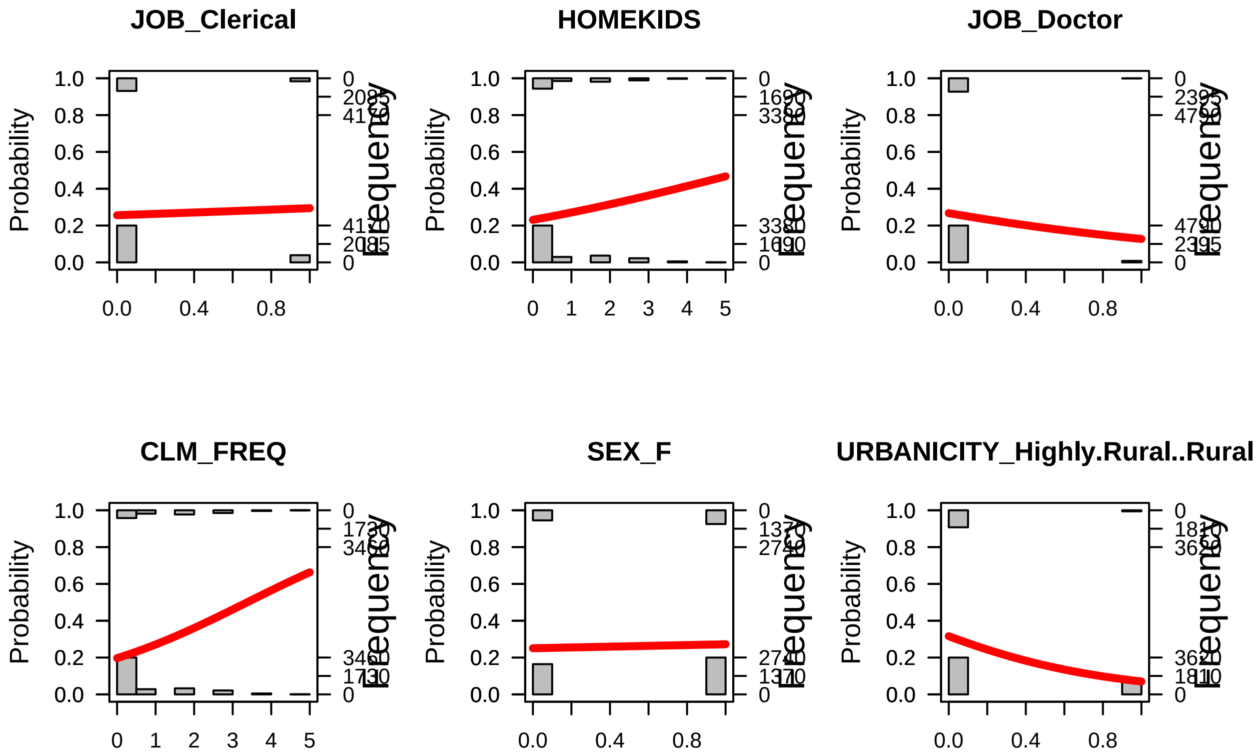
In logistic regression, we are only concerned about the probability of outcome dependent variable (success or failure). As described above, g() is the link function. This function is established using two things: Probability of Success (p) and Probability of Failure (1-p). p should meet following criteria: 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:

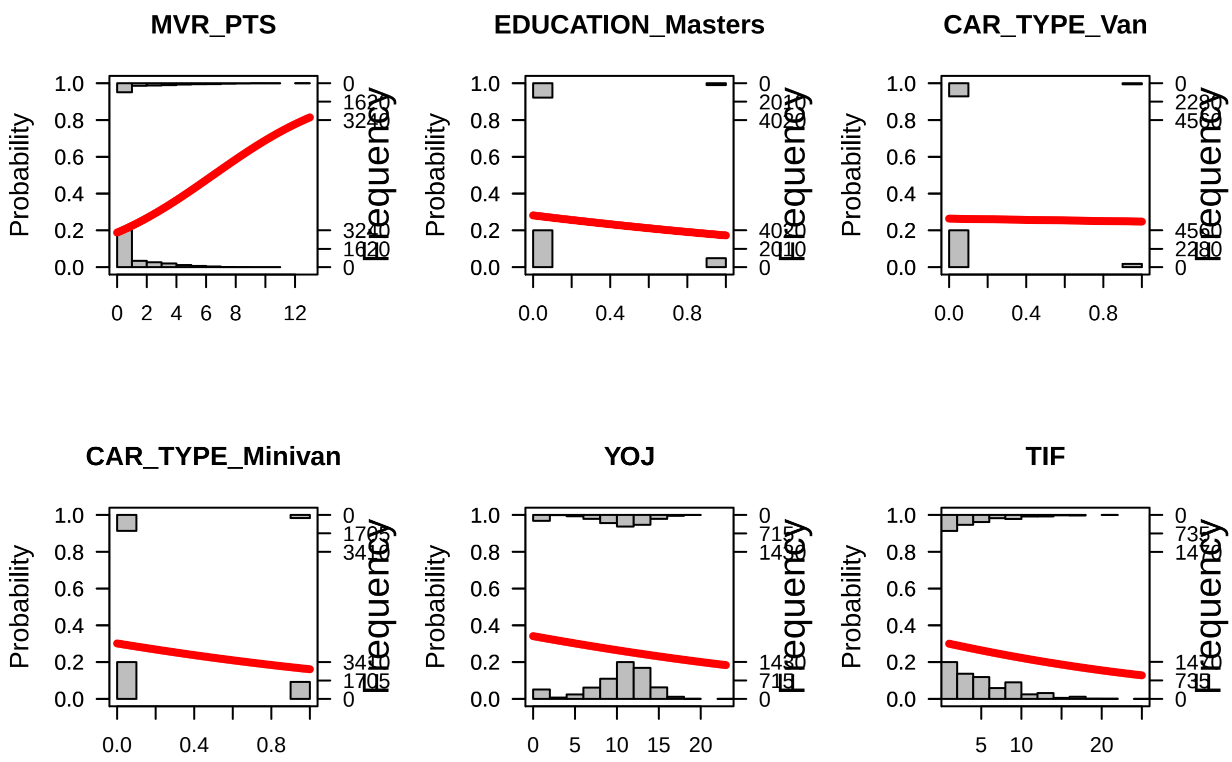
```
par(mfrow=c(2,3))

# #fun1 <- function(a, y) cor(y, a , use = 'na.or.complete')
# #Correlation_TARGET_FLAG <- sapply(x, FUN = fun1, y=insure_train_full$TARGET_FLAG)
#
# show_chart_logi.hist <- function(a, y, ...) {
# #    xlabel <- unlist(str_split(deparse(substitute(a)), pattern = "\\$"))[2]
#     xlabel <- deparse(substitute(a))
#     message(xlabel)
#     logi.hist.plot(a,y,logi.mod = 1, type="hist", boxp=FALSE,col="gray", mainlabel = xlabel)
# }

x <- insure_train_full[,-2]
x <- x[complete.cases(x),]

# sapply(x, FUN = show_chart_logi.hist, y=x$TARGET_FLAG)

logi.hist.plot(x$REVOKED_Yes,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$CAR_USE_Private,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlab
logi.hist.plot(x$CAR_TYPE_SUV,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel
logi.hist.plot(x$PARENT1_Yes,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$KIDSDRIV,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'K
logi.hist.plot(x$CAR_AGE,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'CA
```



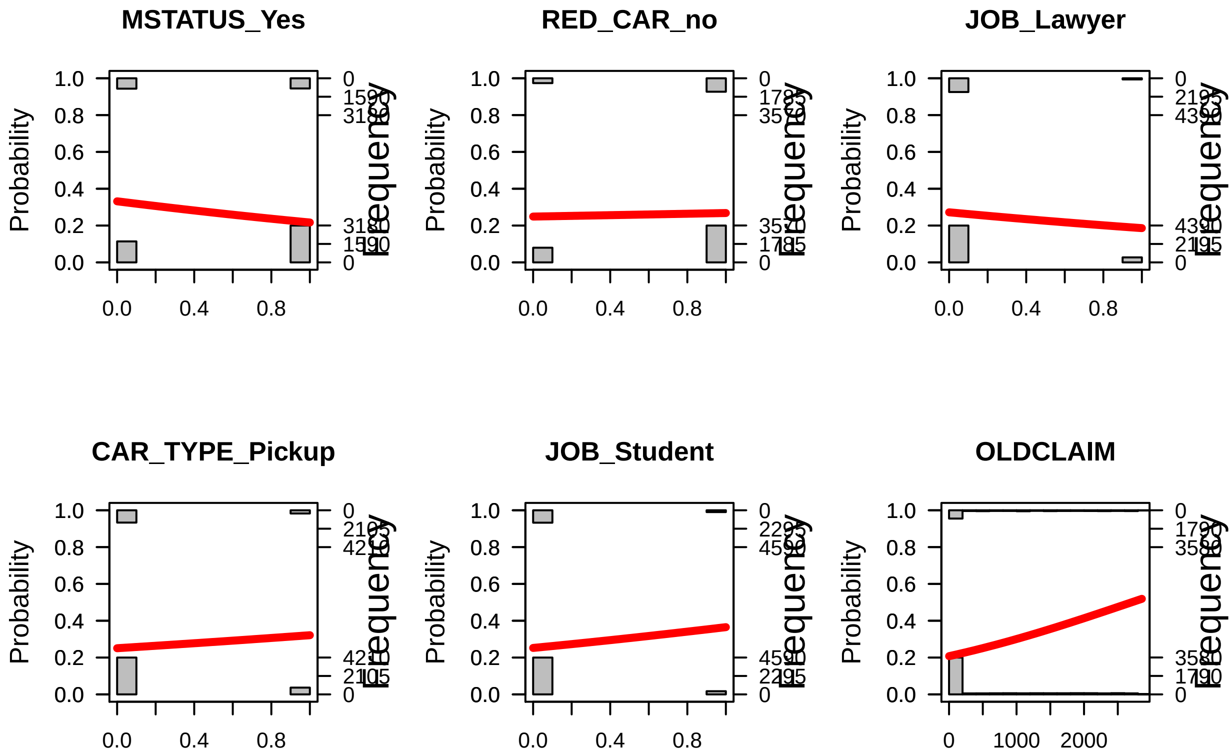

```
logi.hist.plot(x$JOB_Clerical,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel
logi.hist.plot(x$HOMEKIDS,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'I
logi.hist.plot(x$JOB_Doctor,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$CLM_FREQ,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'C
logi.hist.plot(x$SEX_F,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'SEX_
logi.hist.plot(x$URBANICITY_Highly.Rural..Rural,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col=
```



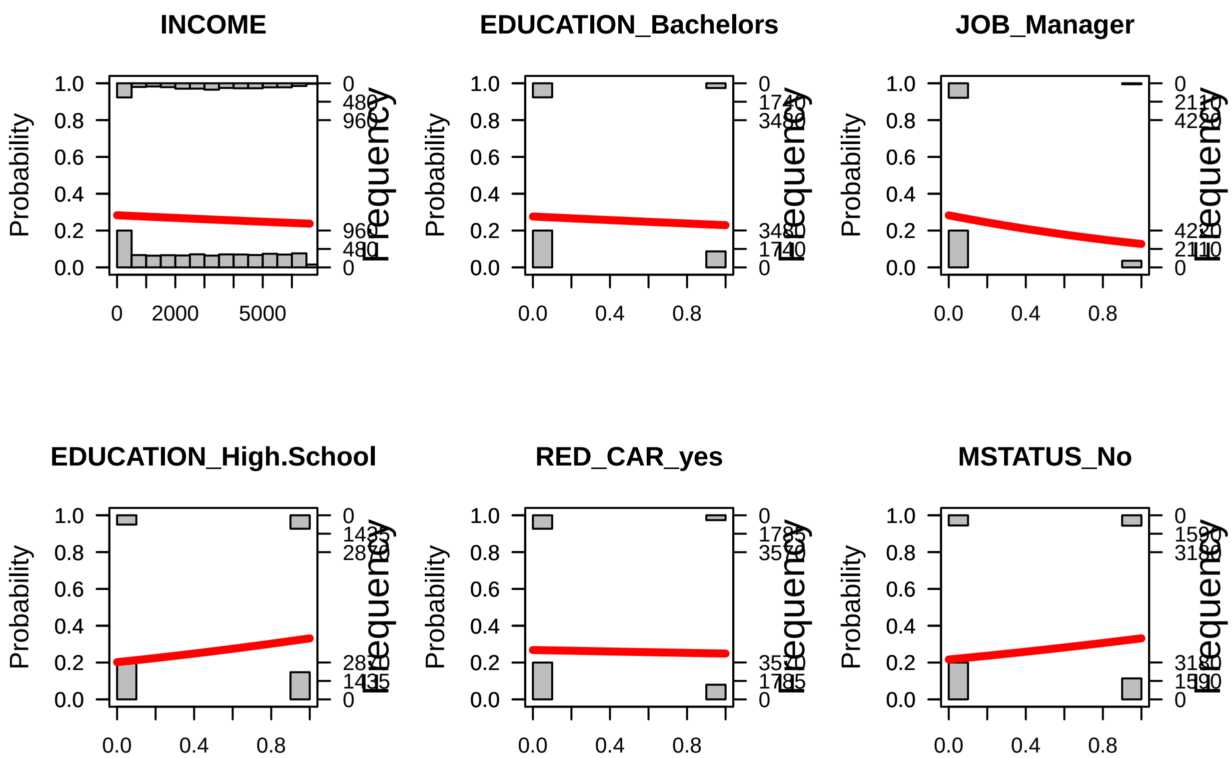

```
logi.hist.plot(x$MVR_PTS,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'M
logi.hist.plot(x$EDUCATION_Masters,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', main
logi.hist.plot(x$CAR_TYPE_Van,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel
logi.hist.plot(x$CAR_TYPE_Minivan,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainl
logi.hist.plot(x$YOJ,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'YOJ')
logi.hist.plot(x$TIF,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'TIF')
```



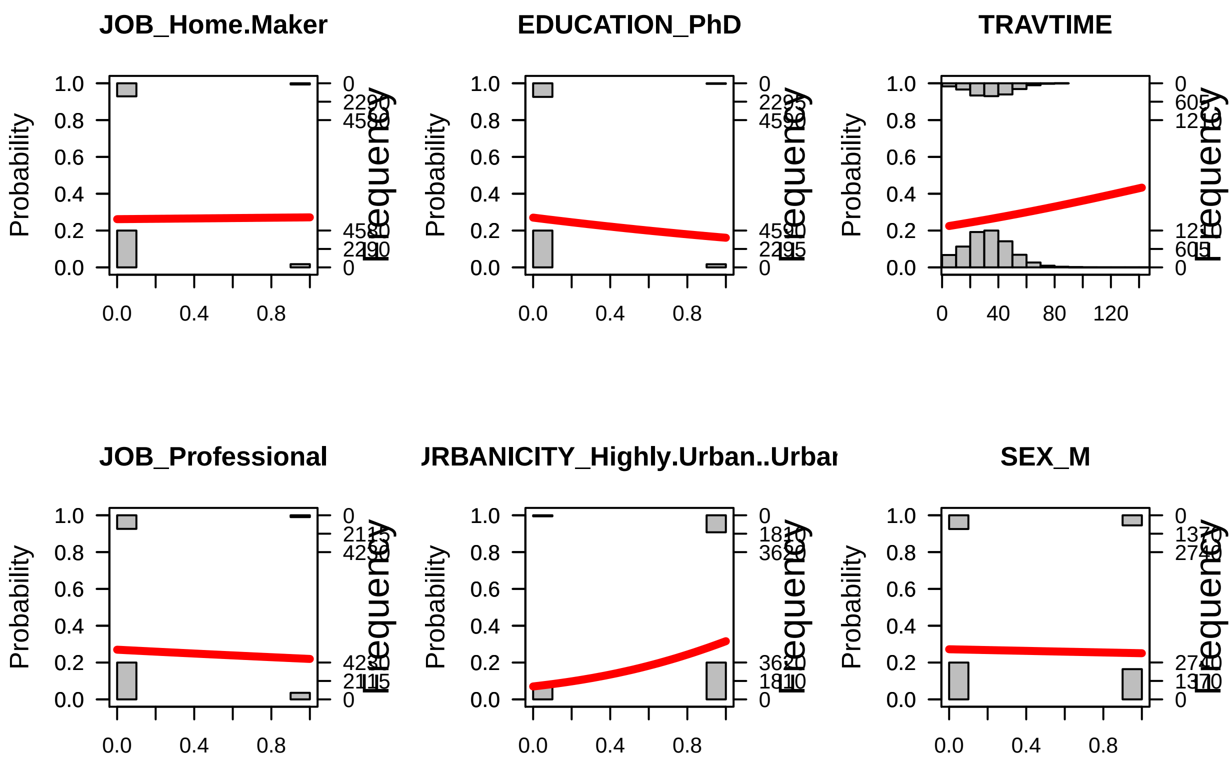

```
logi.hist.plot(x$MSTATUS_Yes,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$RED_CAR_no,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$JOB_Lawyer,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$CAR_TYPE_Pickup,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlab
logi.hist.plot(x$JOB_Student,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$OLDCLAIM,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = '(
```



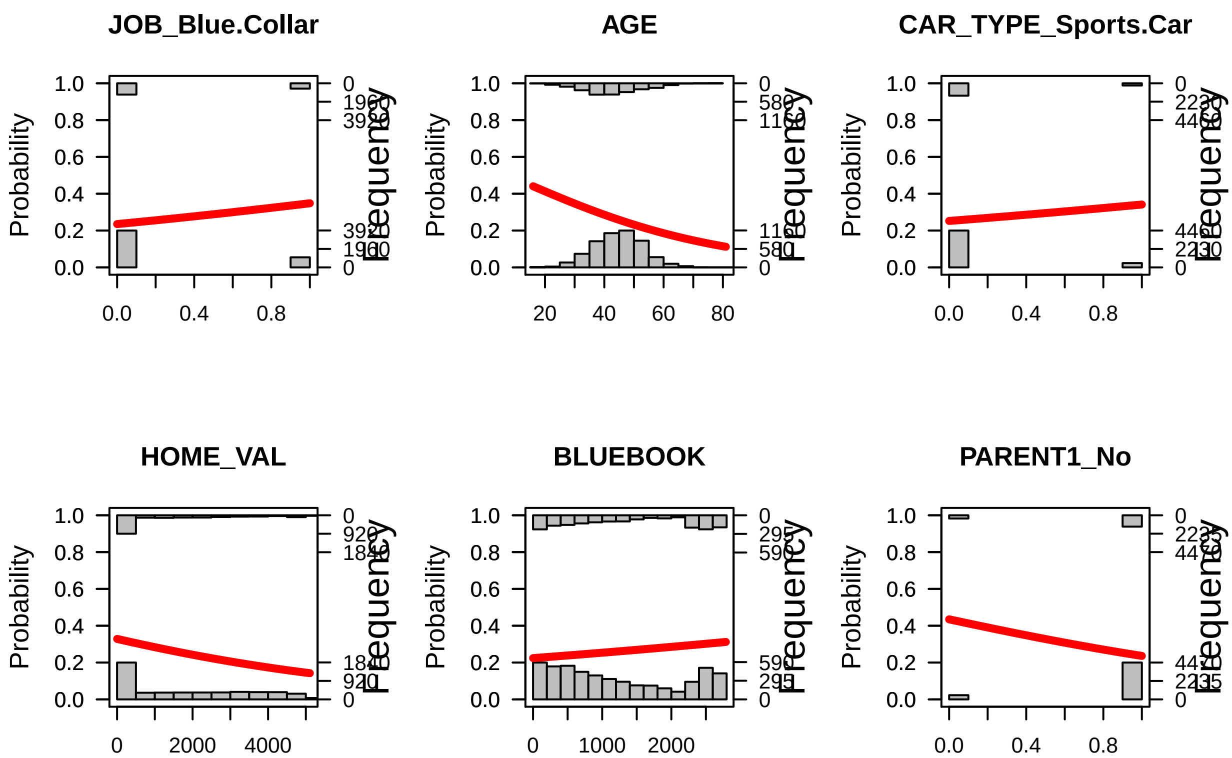

```
logi.hist.plot(x$INCOME,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'INC
logi.hist.plot(x$EDUCATION_Bachelors,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mai
logi.hist.plot(x$JOB_Manager,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$EDUCATION_High.School,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', m
logi.hist.plot(x$RED_CAR_yes,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$MSTATUS_No,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
```



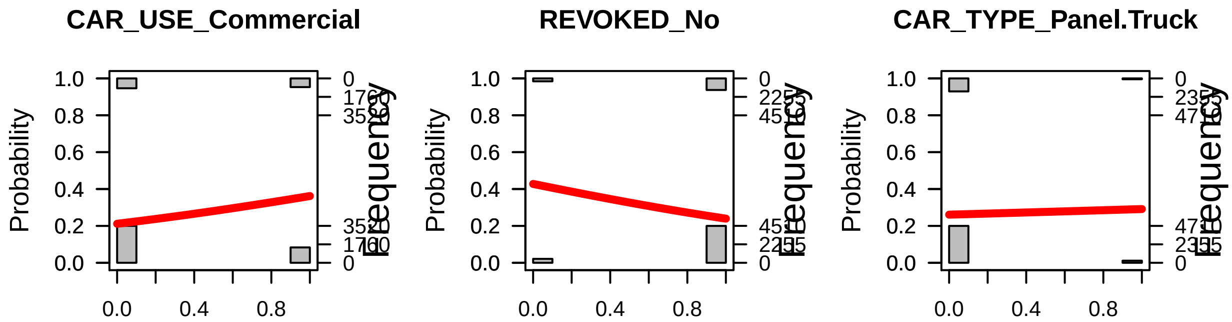

```
logi.hist.plot(x$JOB_Home.Maker,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabe
logi.hist.plot(x$EDUCATION_PhD,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel
logi.hist.plot(x$TRAVTIME,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = ''
logi.hist.plot(x$JOB_Professional,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainla
logi.hist.plot(x$URBANICITY_Highly.Urban..Urban,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col=
logi.hist.plot(x$SEX_M,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'SEX_
```




```
logi.hist.plot(x$JOB_Blue.Collar,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlal
logi.hist.plot(x$AGE,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'AGE')
logi.hist.plot(x$CAR_TYPE_Sports.Car,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', ma
logi.hist.plot(x$HOME_VAL,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'I
logi.hist.plot(x$BLUEBOOK,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel = 'I
logi.hist.plot(x$PARENT1_No,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
```




```
logi.hist.plot(x$CAR_USE_Commercial,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mai
logi.hist.plot(x$REVOKED_No,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', mainlabel =
logi.hist.plot(x$CAR_TYPE_Panel.Truck,x$TARGET_FLAG,logi.mod = 1, type='hist', boxp=FALSE,col='gray', ma
```



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 URBANICITY_Highly.Rural..Rural, PARENT1_No, REVOKED_No, HOME_VAL, CAR_USE_Private, 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.

2. Data Preparation

Now that we have completed the data exploration / analysis, we will be cleaning and consolidating data into two datasets for use in analysis and modeling.

One dataset will be used for building and selecting models for TARGET_FLAG and the other dataset with only the "crash" records will be used for building and selecting models for TARGET_AMT.

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

2.1 Outliers treatment

As the outliers have no impact on the outcome of TARGET_FLAG, we will not be carrying out any transformations for this dataset to handle the outlier.

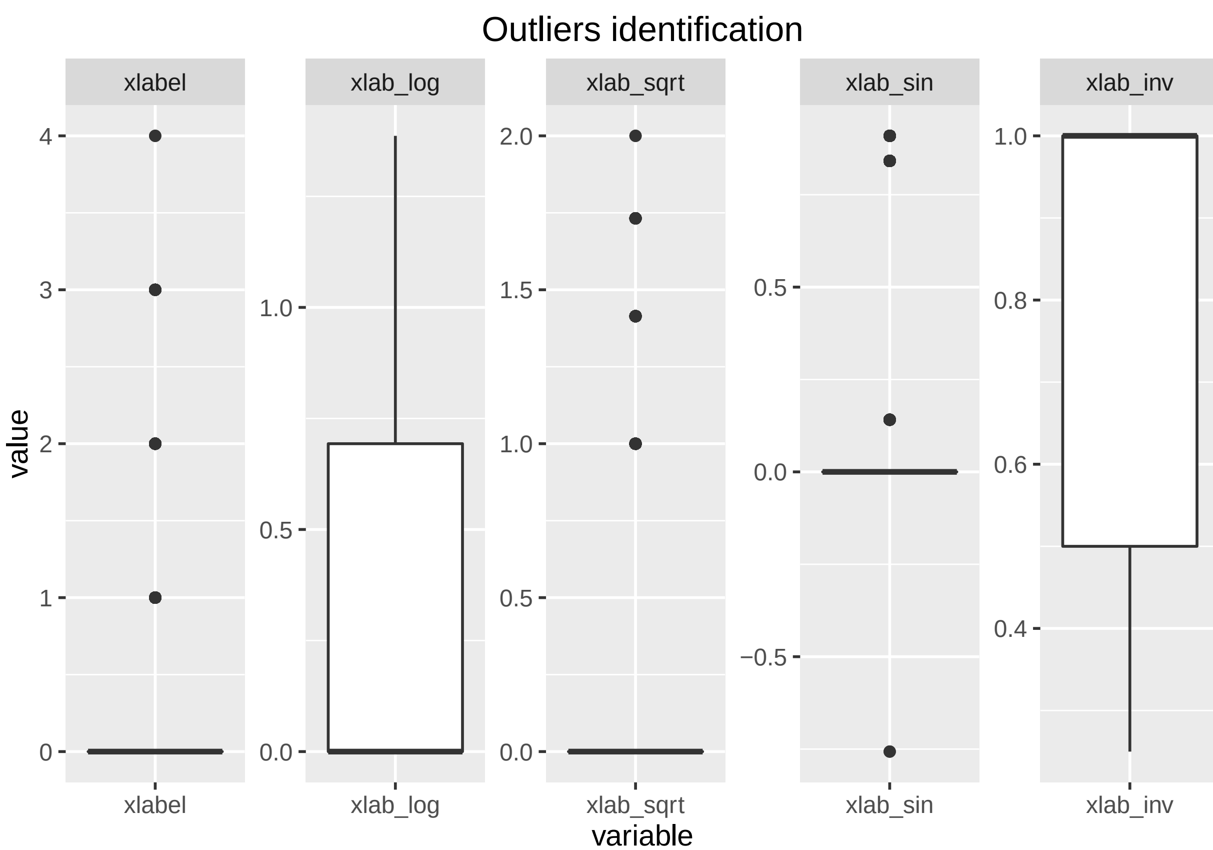
However, the "crashed" dataset where we predict the "TARGET_AMT" needs to have its outliers handled.

In the sections below, we will check different transformations for each of the variables - KIDSDRIV, AGE, HOMEKIDS, MVR_PTS, OLDCLAIM, TIF, TRAVTIME and YOJ - and create the appropriate outlier-handled / transformed variables.

** Transformations for KIDSDRIV**

```r
show_charts(insure_train_crash$KIDSDRIV)
```

```
## Warning: Removed 3546 rows containing non-finite values (stat_boxplot).
```



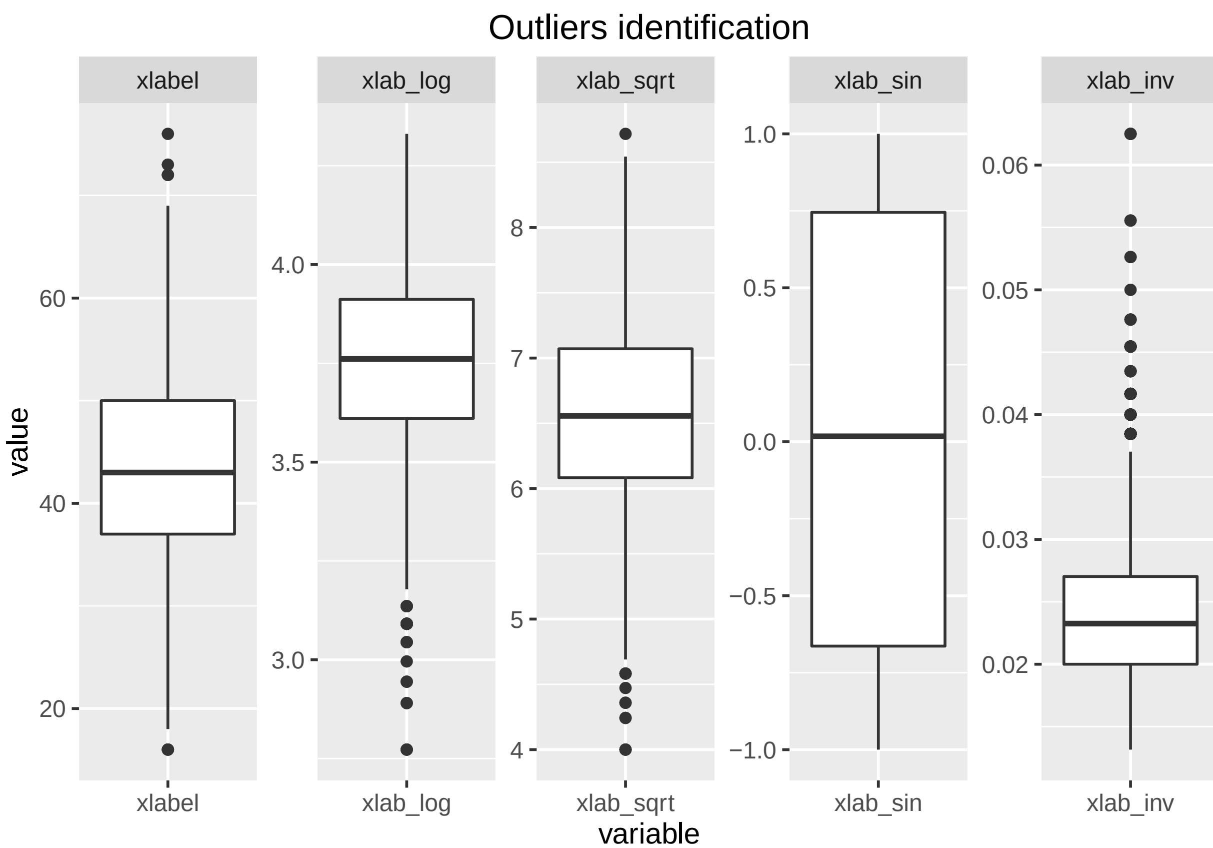
```
insure_train_crash$KIDSDRIV_log <- log(insure_train_crash$KIDSDRIV)
insure_train_crash$KIDSDRIV_inv <- 1 / insure_train_crash$KIDSDRIV
```

From the above charts we can see that a log or an inverse transformation works well for KIDSDRV. Hence, We create both these variables.

** Transformations for AGE**

```
show_charts(insure_train_crash$AGE)
```

```
## Warning: Removed 25 rows containing non-finite values (stat_boxplot).
```



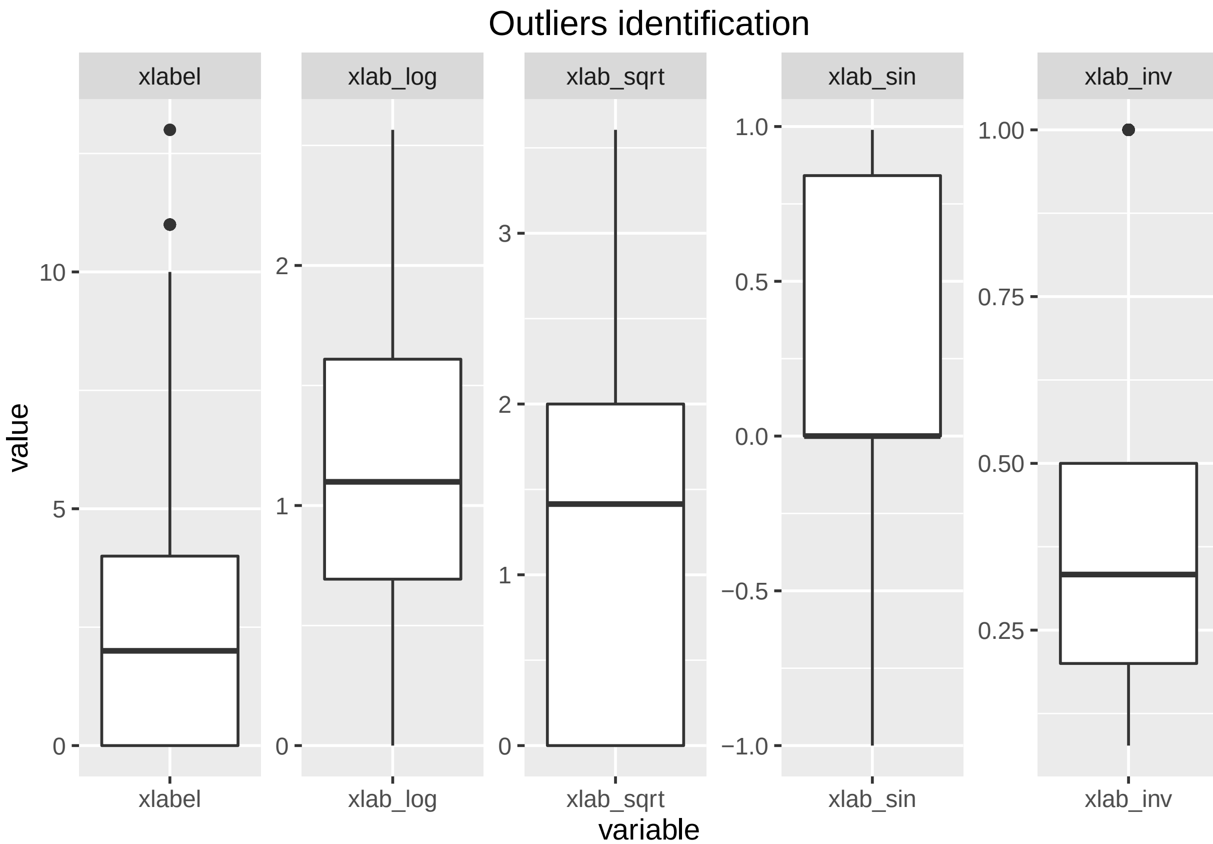

```
insure_train_crash$AGE_sin <- sin(insure_train_crash$AGE)
```

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

** Transformations for MVR_PTS**

```
show_charts(insure_train_crash$MVR_PTS)
```

```
## Warning: Removed 1428 rows containing non-finite values (stat_boxplot).
```

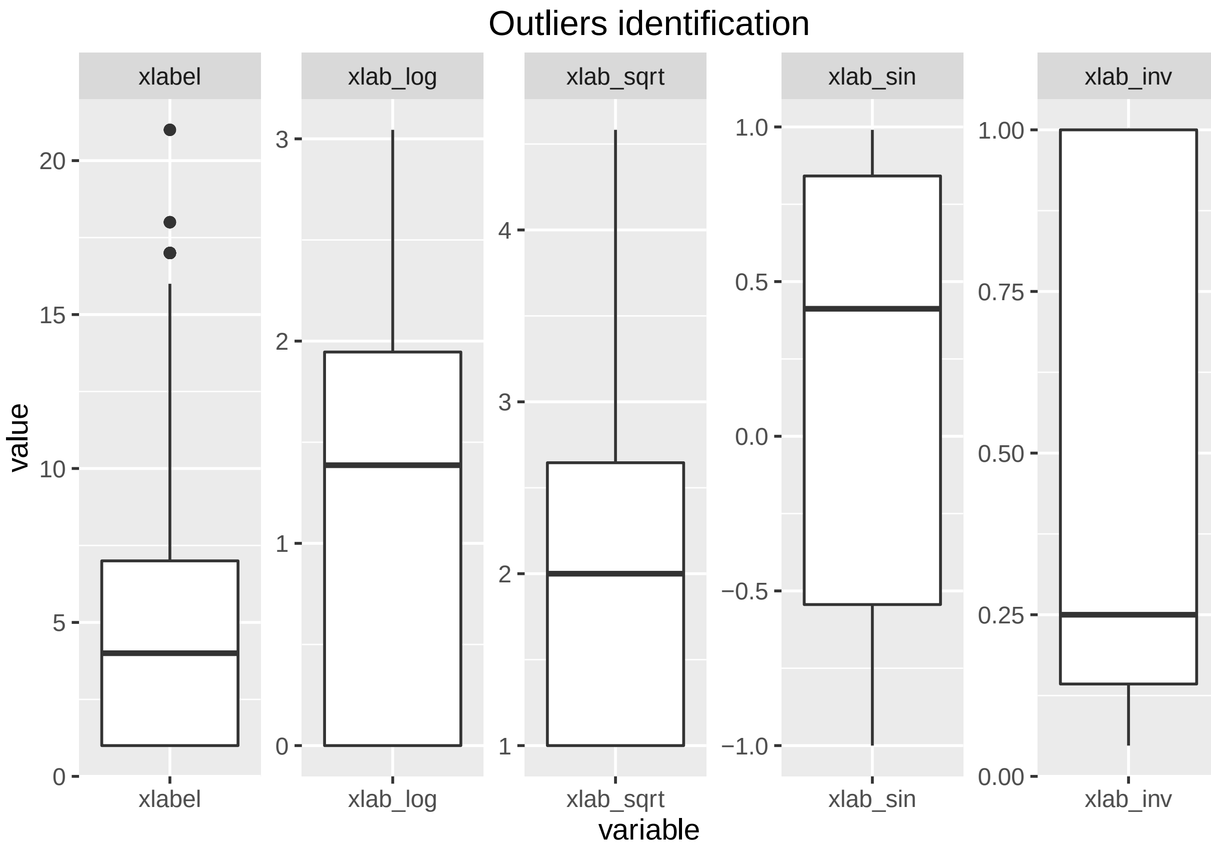



```r
insure_train_crash$MVR_PTS_log <- log(insure_train_crash$MVR_PTS)
insure_train_crash$MVR_PTS_sqrt <- sqrt(insure_train_crash$MVR_PTS)
```

From the above charts we can see that a log, sqrt or an inverse transformation works well for MVR_PTS. Hence, We will create these variables.

** Transformations for TIF**

```r
#TIF, TRAVTIME and YOJ

show_charts(insure_train_crash$TIF)
```

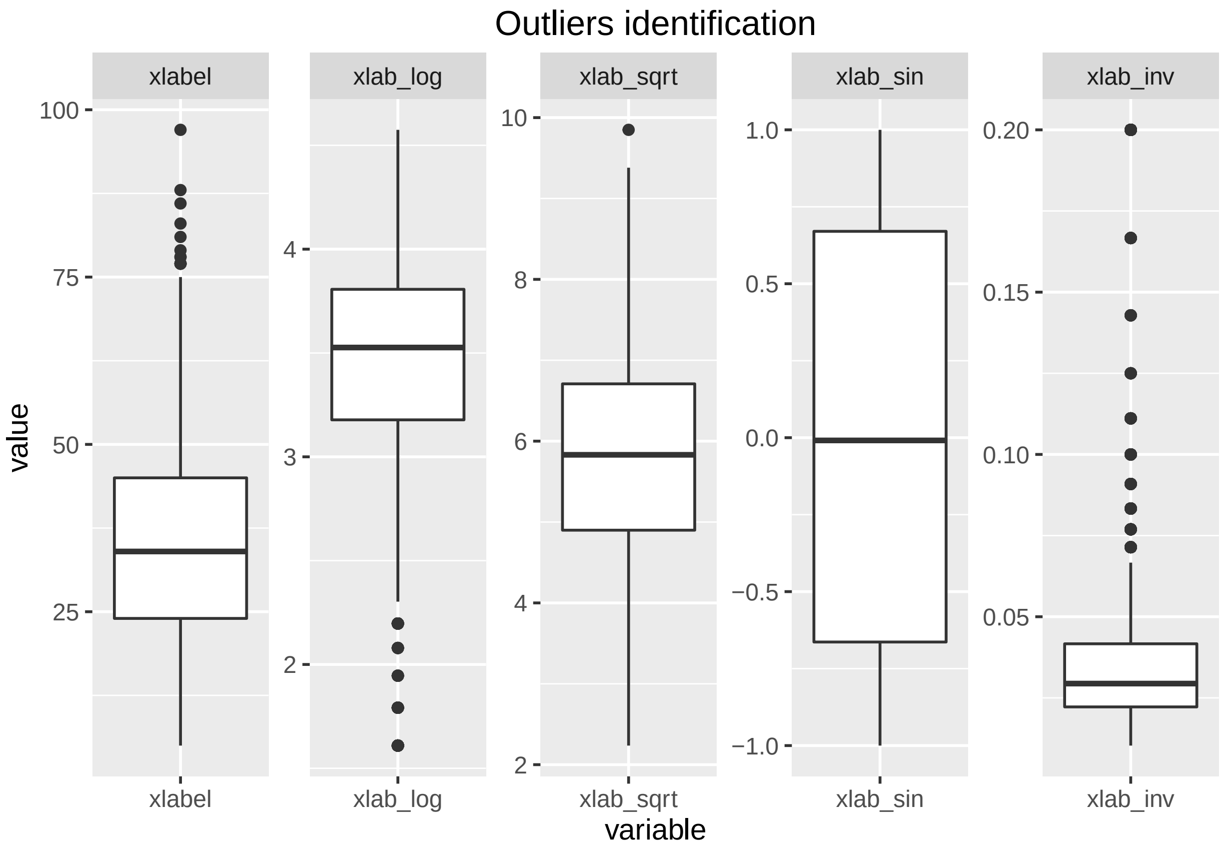



```
insure_train_crash$TIF_log <- log(insure_train_crash$TIF)
insure_train_crash$TIF_sqrt <- sqrt(insure_train_crash$TIF)
insure_train_crash$TIF_sin <- sin(insure_train_crash$TIF)
insure_train_crash$TIF_inv <- 1 / insure_train_crash$TIF
```

From the above charts we can see that a log, sqrt, sin or an inverse transformation works well for TIF. Hence, We will create these variables.

** Transformations for TRAVTIME**

```
#TRAVTIME and YOJ

show_charts(insure_train_crash$TRAVTIME)
```



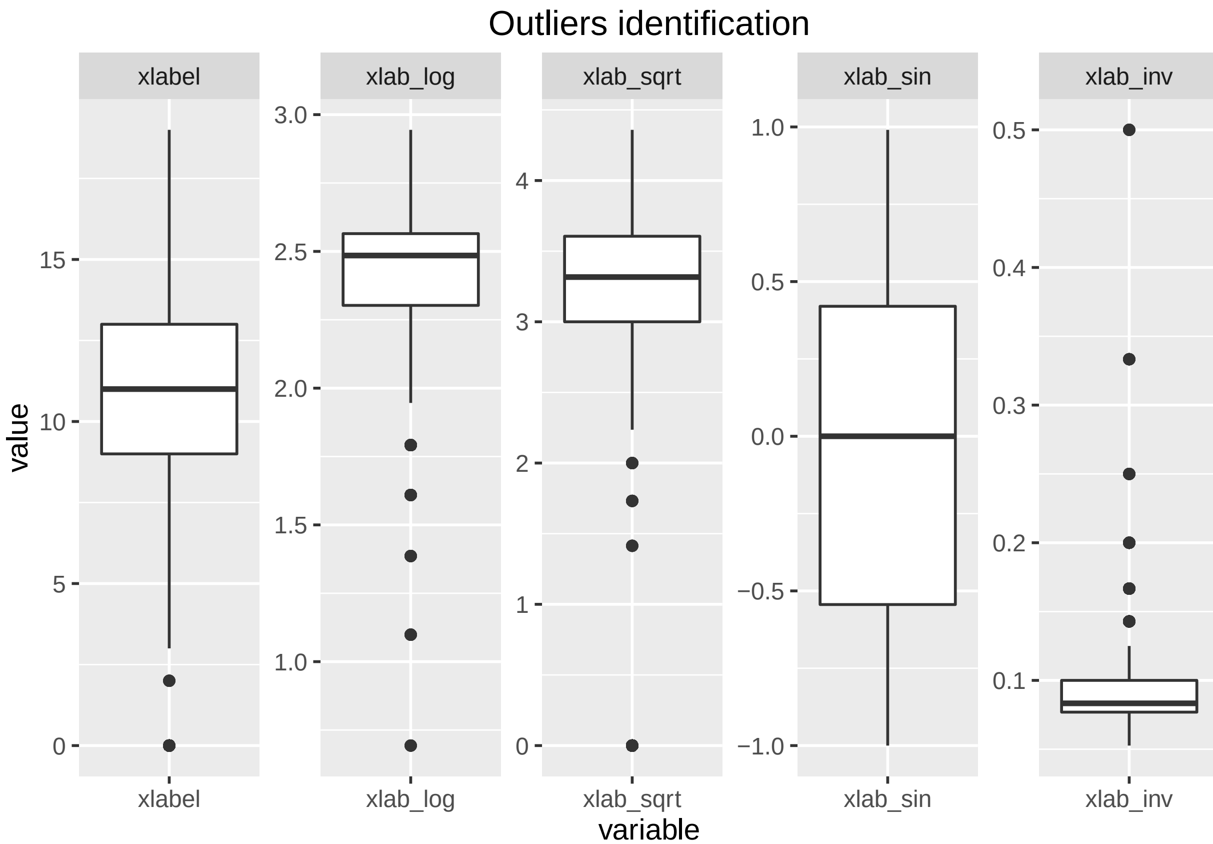

```
insure_train_crash$TRAVTIME_sin <- sin(insure_train_crash$TRAVTIME)
```

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

** Transformations for YOJ**

```
#YOJ

show_charts(insure_train_crash$YOJ)
```

```
## Warning: Removed 1107 rows containing non-finite values (stat_boxplot).
```




```
insure_train_crash$YOJ_sin <- sin(insure_train_crash$YOJ)
```

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

2.2 Missing Values treatment

As we have seen in the data exploration phase, we can do with removing the rows that contain missing values. We now do this for both the datasets:

```
insure_train_full <- insure_train_full[complete.cases(insure_train_full),]
insure_train_crash <- insure_train_crash[complete.cases(insure_train_crash),]
```

2.3 Adding New Variables

In this section, we generate some additional variables that we feel will help the correlations. As before, we do it for both the datasets.

2.3.1 New Variables for Full Dataset (TARGET_FLAG)

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 these variables 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

```
insure_train_full$CAR_TYPE_FLAG_BIN <- ifelse(insure_train_full$CAR_TYPE_Minivan | insure_train_full$CAR
```

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.

Accordingly, we will bin these variables as below:
EDUCATION_FLAG_BIN :

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

```
insure_train_full$EDUCATION_FLAG_BIN <- ifelse(insure_train_full$EDUCATION_High.School, 1, 0)
```

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.

Accordingly, we will bin these variables as below:
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 or professional

```
insure_train_full$JOB_TYPE_FLAG_BIN <- ifelse(insure_train_full$JOB_Student |  insure_train_full$JOB_Hom
```

2.3.2 New Variables for Crashed Dataset (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 or in a Blue Collar job, 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 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
```