DATA 621 – Business Analytics and Data Mining

Homework 4

Ramnivas Singh

2022-05-01

1.0 Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, TARGET_FLAG, is a 1 or a 0. A "1" means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

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 people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children 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 into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

1.1 Deliverables

- A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.
- Assigned predictions (probabilities, classifications, cost) for the evaluation data set. Use 0.5 threshold.
- Include your R statistical programming code in an Appendix.

Solution Steps & Approach

- Data Exploration : The auto insurance training dataset has 26 variables and 8161 observations. Of the variables, 24 of them are predictors for two responses.
- Data Preparation : To prepare the data, we checked for any NA's or missing values. There were none.
- Build Models : We built a model using all predictors as numerics.
- Select Models :Select a suitable model
- Appendix

Import Libraries and Data

```
# Loading the data
train_df = read.csv("https://raw.githubusercontent.com/rnivas2028/MSDS/Data621/HW4/insuran
         ce_training_data.csv")
test_df = read.csv("https://raw.githubusercontent.com/rnivas2028/MSDS/Data621/HW4/insuranc
         e-evaluation-data.csv")
head(train_df)
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ   INCOME PARENT1
## 1     1           0          0        0  60        0  11  $67,349      No
## 2     2           0          0        0  43        0  11  $91,449      No
## 3     4           0          0        0  35        1  10  $16,039      No
## 4     5           0          0        0  51        0  14                No
## 5     6           0          0        0  50        0  NA $114,986      No
## 6     7           1       2946        0  34        1  12 $125,301     Yes
##   HOME_VAL MSTATUS SEX     EDUCATION          JOB TRAVTIME    CAR_USE BLUEBOOK
## 1       $0    z_No   M           PhD  Professional       14    Private  $14,230
## 2 $257,252    z_No   M z_High School z_Blue Collar       22 Commercial  $14,940
## 3 $124,191     Yes z_F z_High School      Clerical        5    Private   $4,010
## 4 $306,251     Yes   M  <High School z_Blue Collar       32    Private  $15,440
## 5 $243,925     Yes z_F           PhD        Doctor       36    Private  $18,000
## 6       $0    z_No z_F     Bachelors z_Blue Collar       46 Commercial  $17,430
##   TIF   CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1  11    Minivan     yes   $4,461        2      No       3      18
## 2   1    Minivan     yes       $0        0      No       0       1
## 3   4      z_SUV      no  $38,690        2      No       3      10
## 4   7    Minivan     yes       $0        0      No       0       6
## 5   1      z_SUV      no  $19,217        2     Yes       3      17
## 6   1 Sports Car      no       $0        0      No       0       7
##            URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
## 3 Highly Urban/ Urban
## 4 Highly Urban/ Urban
## 5 Highly Urban/ Urban
## 6 Highly Urban/ Urban
```

2.0 Data Exploration & Preparation

The auto insurance training dataset has 26 variables and 8161 observations. Of the variables, 24 of them are predictors for two responses: TARGET_FLAG and TARGET_AMT is numerical.

To explore the training data, used: * Summary function to see means, medians, and quartiles of predictors * Str function to see the data type of each predictor * Explored TARGET_FLAG in relation to some other variables such as AGE and CAR_AGE * Looked at distribution of some numerical variables such as AGE and MVR_PTS

From the summary function, the TARGET_FLAG is binary and 26% of the 8161 records were accidents. See a summary of each column in the train_df set

```
# view a summary of all columns
summary(train_df)
```

```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ          INCOME
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Length:8161
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   Class :character
##  Median :45.00   Median :0.0000   Median :11.0   Mode  :character
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0
##  NA's   :6                        NA's   :454
##    PARENT1            HOME_VAL           MSTATUS              SEX
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   EDUCATION             JOB               TRAVTIME        CAR_USE
##  Length:8161        Length:8161        Min.   :  5.00   Length:8161
##  Class :character   Class :character   1st Qu.: 22.00   Class :character
##  Mode  :character   Mode  :character   Median : 33.00   Mode  :character
##                                        Mean   : 33.49
##                                        3rd Qu.: 44.00
##                                        Max.   :142.00
##
##    BLUEBOOK              TIF           CAR_TYPE           RED_CAR
##  Length:8161        Min.   : 1.000   Length:8161        Length:8161
##  Class :character   1st Qu.: 1.000   Class :character   Class :character
##  Mode  :character   Median : 4.000   Mode  :character   Mode  :character
##                     Mean   : 5.351
##                     3rd Qu.: 7.000
##                     Max.   :25.000
##
##    OLDCLAIM           CLM_FREQ        REVOKED             MVR_PTS
##  Length:8161        Min.   :0.0000   Length:8161        Min.   : 0.000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Mode  :character   Median :0.0000   Mode  :character   Median : 1.000
##                     Mean   :0.7986                      Mean   : 1.696
##                     3rd Qu.:2.0000                      3rd Qu.: 3.000
##                     Max.   :5.0000                      Max.   :13.000
##
##     CAR_AGE        URBANICITY
##  Min.   :-3.000   Length:8161
##  1st Qu.: 1.000   Class :character
##  Median : 8.000   Mode  :character
```

```
##  Mean   : 8.328
##  3rd Qu.:12.000
##  Max.   :28.000
##  NA's   :510
```

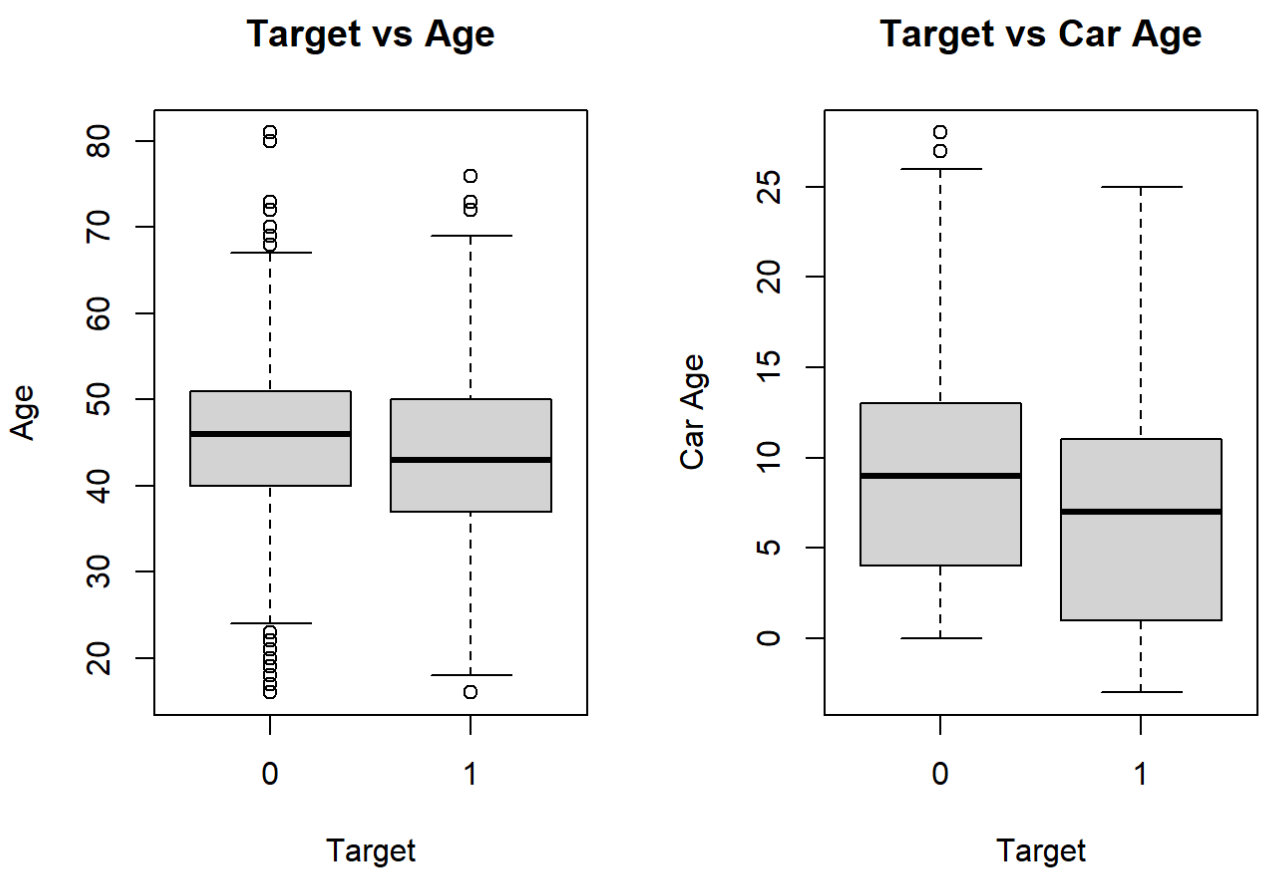
Look at the data type of each variable

```
# data type of predictors
str(train_df)
```

```
## '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     : chr  "$67,349" "$91,449" "$16,039" "" ...
##  $ PARENT1    : chr  "No" "No" "No" "No" ...
##  $ HOME_VAL   : chr  "$0" "$257,252" "$124,191" "$306,251" ...
##  $ MSTATUS    : chr  "z_No" "z_No" "Yes" "Yes" ...
##  $ SEX        : chr  "M" "M" "z_F" "M" ...
##  $ EDUCATION  : chr  "PhD" "z_High School" "z_High School" "<High School" ...
##  $ JOB        : chr  "Professional" "z_Blue Collar" "Clerical" "z_Blue Collar" ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : chr  "Private" "Commercial" "Private" "Private" ...
##  $ BLUEBOOK   : chr  "$14,230" "$14,940" "$4,010" "$15,440" ...
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : chr  "Minivan" "Minivan" "z_SUV" "Minivan" ...
##  $ RED_CAR    : chr  "yes" "yes" "no" "yes" ...
##  $ OLDCLAIM   : chr  "$4,461" "$0" "$38,690" "$0" ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : chr  "No" "No" "No" "No" ...
##  $ 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 : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban"
"Highly Urban/ Urban" ...
```

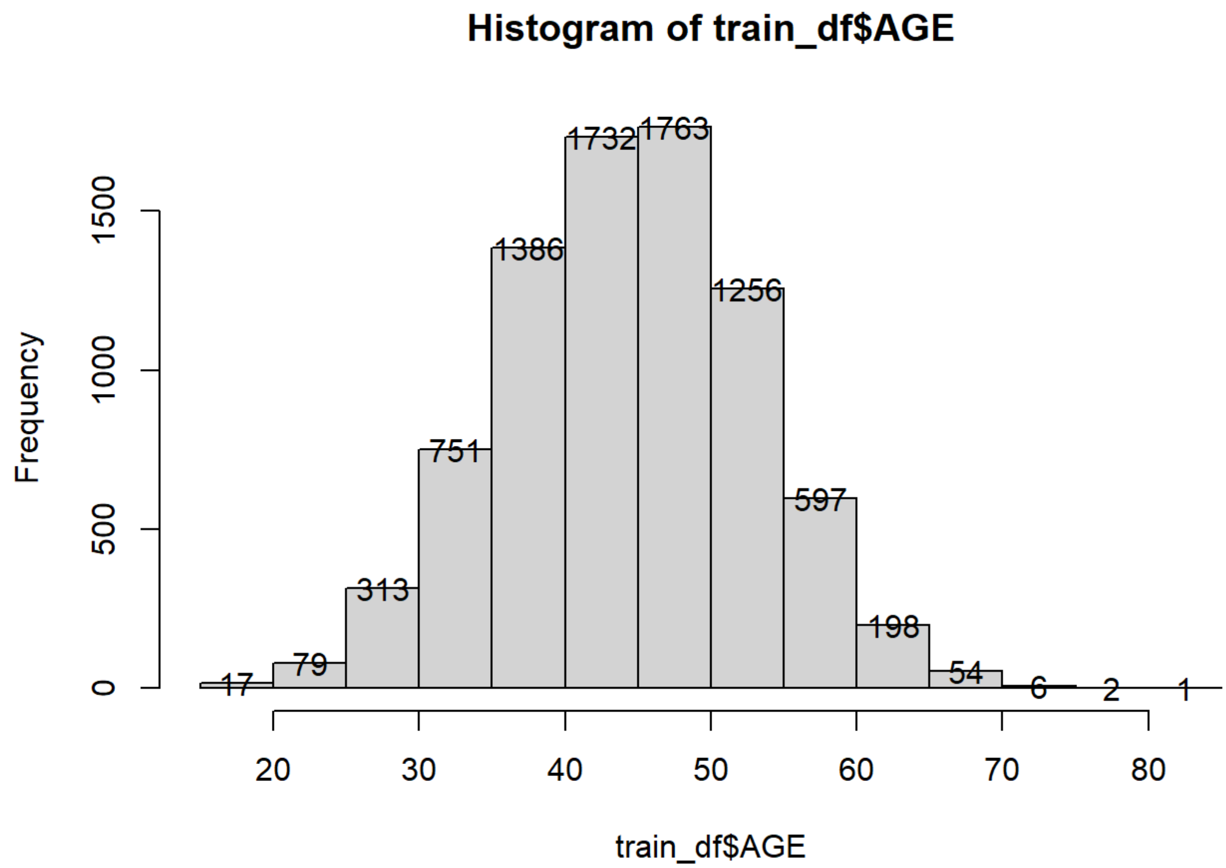
Look at the relationship between TARGET_FLAG and some of the numerical variables.

```r
par(mfrow=c(1,2))
# plot response variable "target" against predictor variable "age" and "car_age"
boxplot(AGE ~ TARGET_FLAG, train_df,
        main="Target vs Age",
        xlab="Target",
        ylab="Age")
boxplot(CAR_AGE ~ TARGET_FLAG, train_df,
        main="Target vs Car Age",
        xlab="Target",
        ylab="Car Age")
```

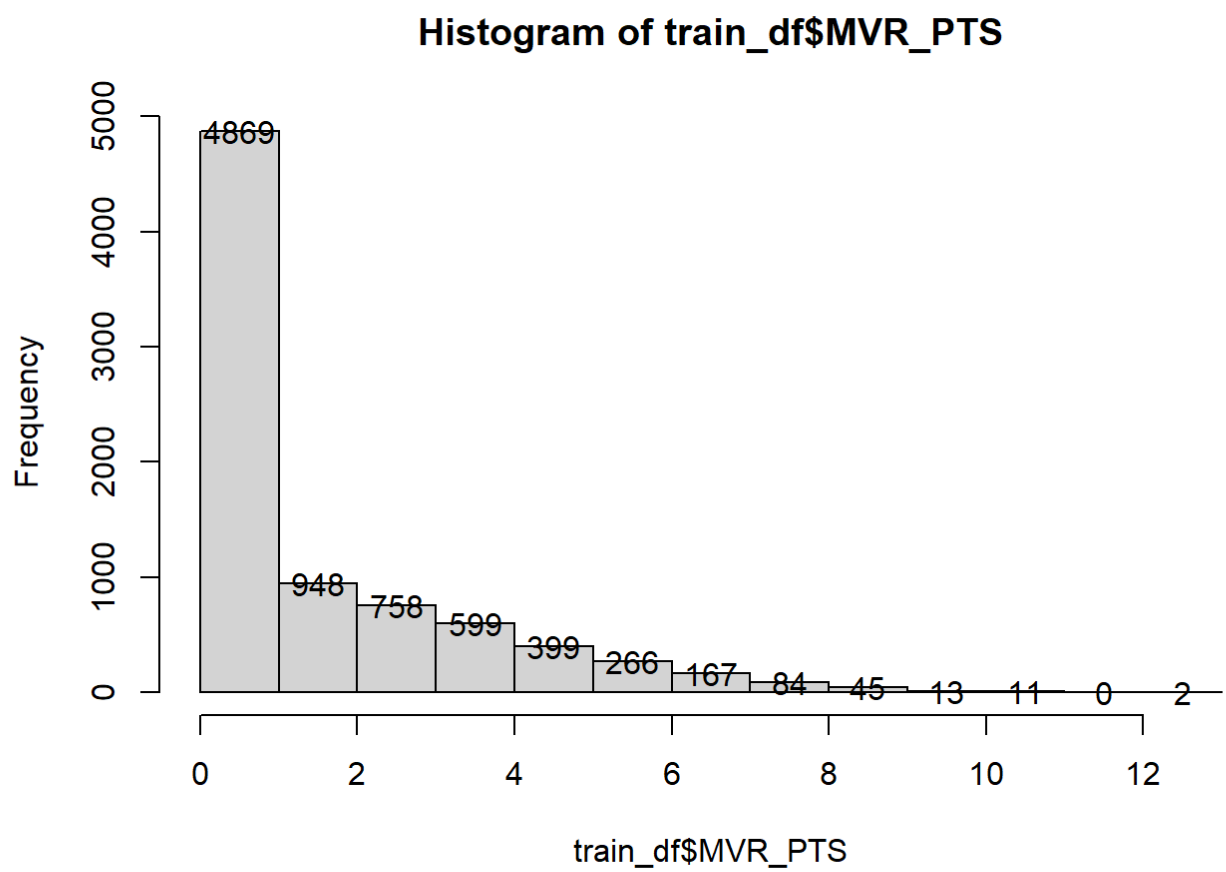


Look at the distribution of some numerical variables.

```r
h <- hist(train_df$AGE)
text(h$mids,h$counts,labels=h$counts)
```




```
h <- hist(train_df$MVR_PTS)
text(h$mids,h$counts,labels=h$counts)
```



This data was prepared to build both a binary logistic model and a multiple linear regression model. The binary logisitc model was used to predict the TARGET_FLAG response variable and the multiple linear regression model was used to predict the TARGET_AMT variable. Thus, there was a different training dataset prepared for each model.

In both training datasets, all 948 records with at least one missing value were removed.

Then, in the multiple linear regression training dataset all records with TARGET_AMT = 0 were removed.

Check for NA's

```
has_NA = names(which(sapply(train_df, anyNA)))
has_NA
```

```
## [1] "AGE"     "YOJ"     "CAR_AGE"
```

Remove rows with NA's train_df will be used for binary logistic regression model

```
train_df <- train_df[complete.cases(train_df), ]
```

The training dataset for the binary logistic regression model was labelled train_df. The training dataset for the multiple linear regression model was titled train_amt_df.

Create train_amt_df dataframe for multiple linear regression model

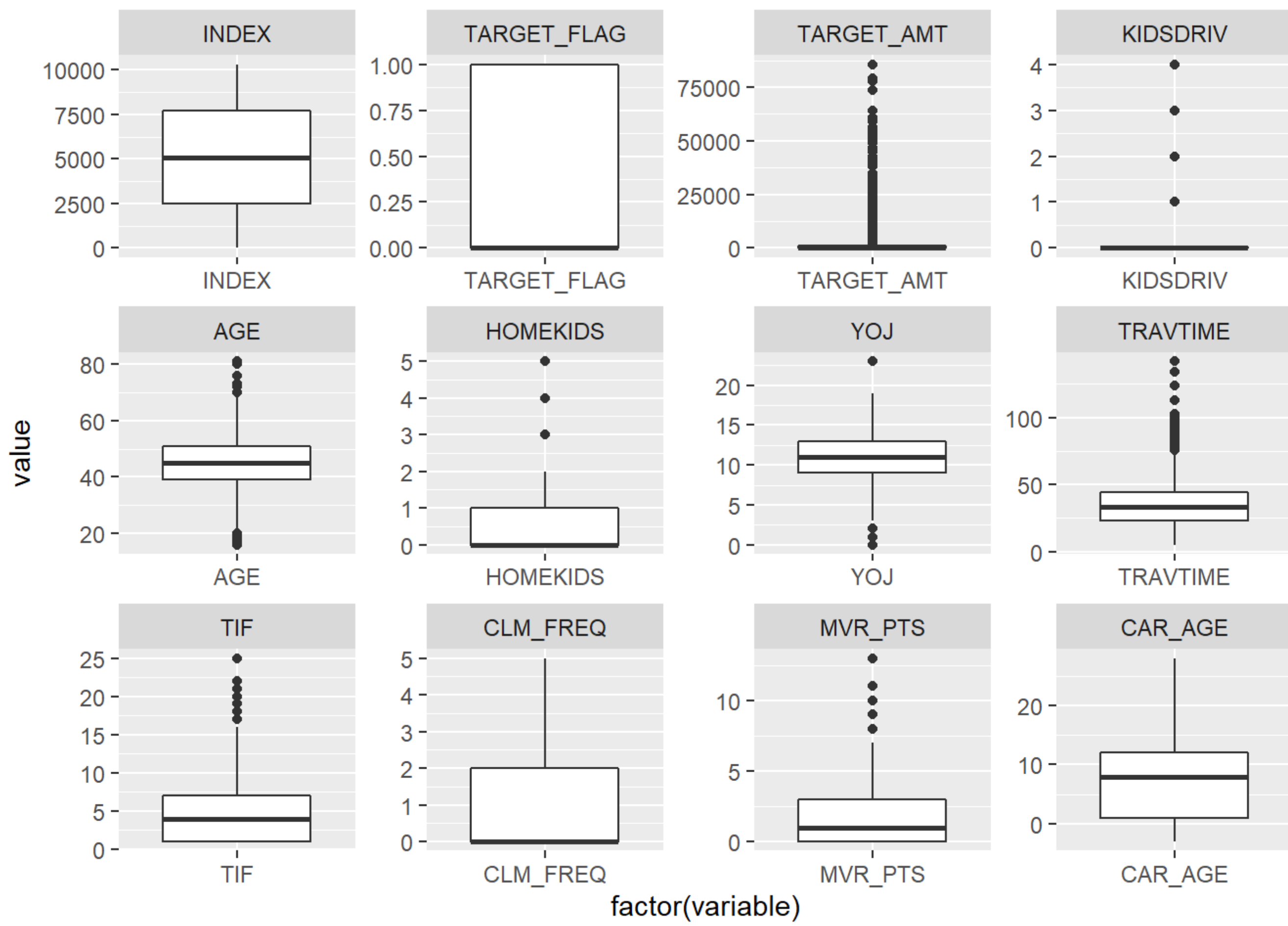
```r
train_amt_df <- subset(train_df, TARGET_AMT > 0)
summary(train_amt_df$TARGET_FLAG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       1       1       1       1       1       1
```

Boxplots

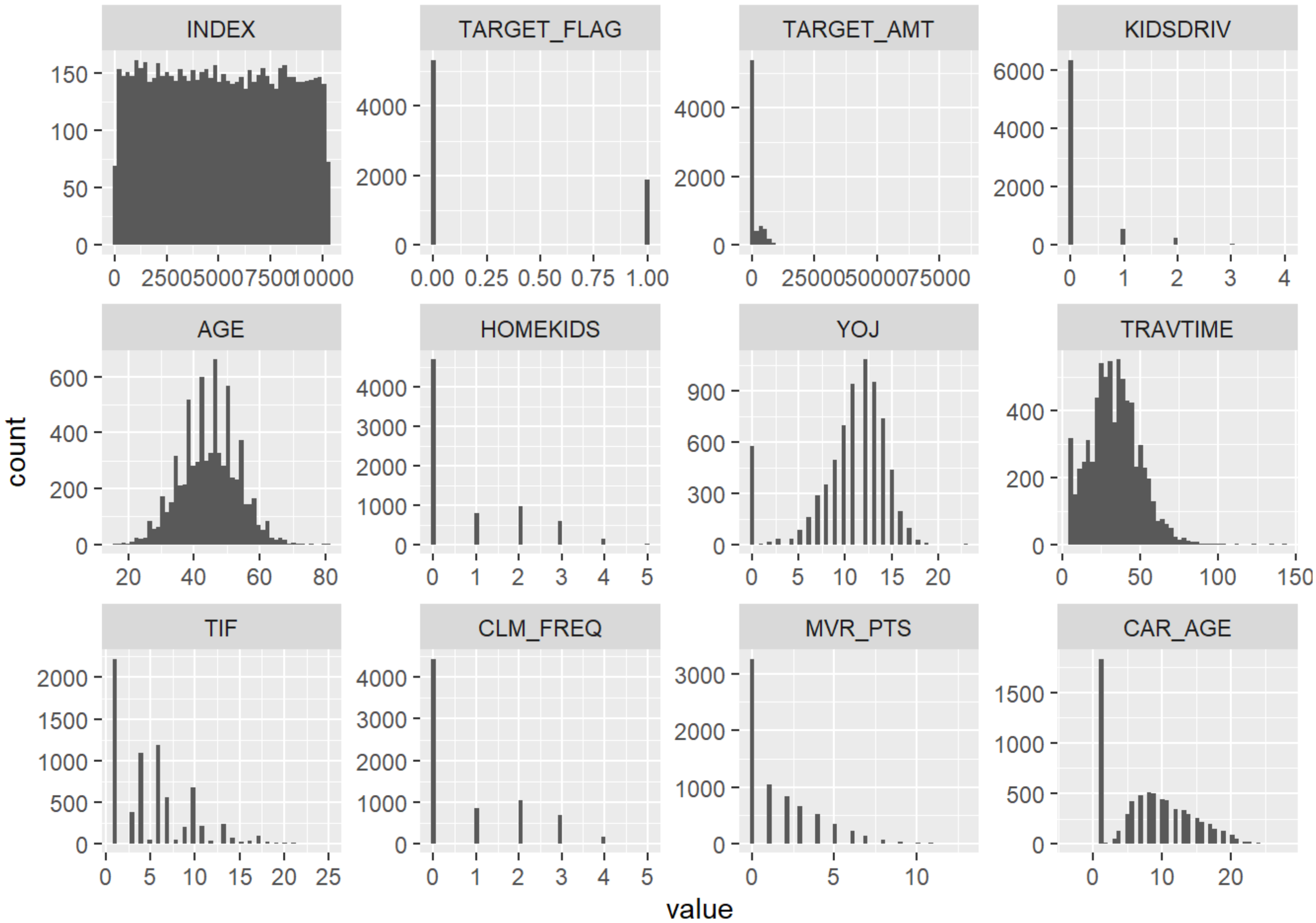
The below boxplots show all of the variables listed in the dataset. This visualization will assist in showing how the data is spread for each variable.

```r
ggplot(melt(train_df), aes(x=factor(variable), y=value)) +
  facet_wrap(~variable, scale="free") +
  geom_boxplot()
```

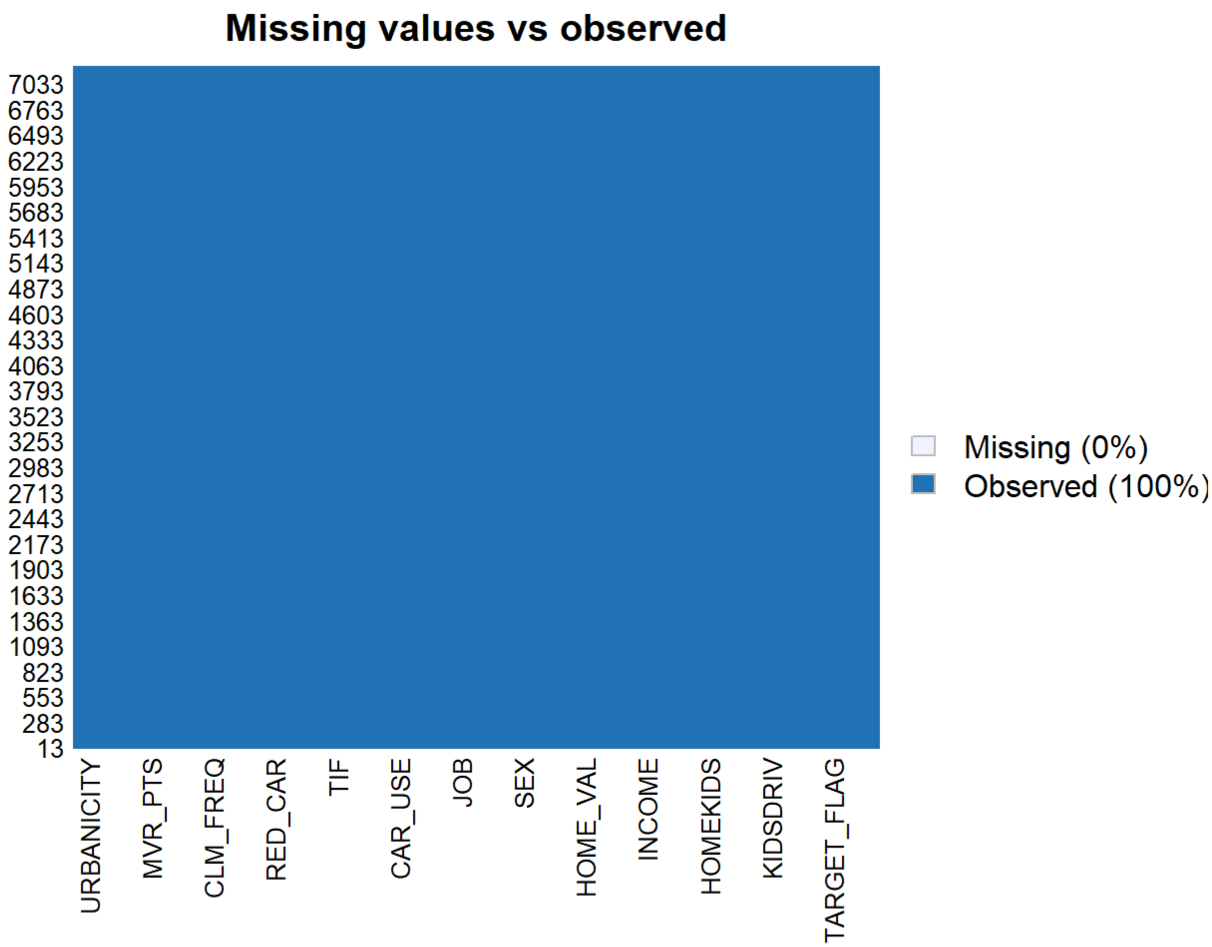


Histograms

```r
ggplot(melt(train_df), aes(x=value)) +
  facet_wrap(~variable, scale="free") +
  geom_histogram(bins=50)
```


INDEX
TARGET_FLAG
TARGET_AMT
KIDSDRIV
AGE
HOMEKIDS
YOJ
TRAVTIME
TIF
CLM_FREQ
MVR_PTS
CAR_AGE
count
value


TIF
YOJ
CAR_AGE
AGE
INDEX
CLM_FREQ
TRAVTIME
MVR_PTS
TARGET_AMT
TARGET_FLAG
KIDSDRIV
HOMEKIDS
0.03 0.01 0.00 -0.01 -0.02 -0.01 -0.04 -0.05 -0.08 -0.00 0.00
0.06 0.14 0.03 -0.03 -0.02 -0.04 -0.02 -0.07 0.05 0.09
0.18 -0.00 -0.01 -0.04 -0.02 -0.06 -0.10 -0.06 -0.16
0.04 -0.03 0.00 -0.07 -0.05 -0.10 -0.07 -0.44
0.02 -0.02 0.01 -0.00 0.00 0.01 -0.00
0.01 0.40 0.12 0.22 0.04 0.03
0.01 0.03 0.05 0.01 -0.01
0.14 0.23 0.06 0.06
0.54 0.06 0.07
0.10 0.12
0.46



3.0 Build Models

First, we built two models using most predictors as numerics. Then we used the step AIC function to find the best variables for each model.

One model was a Binary Logistic Regression model for the TARGET_FLAG response titled step_BLR. The second model was a Multiple Linear Regression for the TARGET_AMT response titled MLR_all_vars

Binary Logistic Regression

```
# preliminary exploration with one predictor
model1 <- glm(formula = TARGET_FLAG ~ AGE, family = binomial(), data = train_df)
summary(model1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE, family = binomial(), data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0712  -0.8017  -0.7376   1.4215   2.0219
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.184991   0.140255   1.319    0.187
## AGE         -0.027504   0.003141  -8.756   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8303.6  on 7212  degrees of freedom
## Residual deviance: 8225.7  on 7211  degrees of freedom
## AIC: 8229.7
##
## Number of Fisher Scoring iterations: 4
```

Binary Logistic Regression Model with more variables

```
BLR_all_vars = glm(TARGET_FLAG ~ AGE +
                  CAR_AGE +
                  MVR_PTS +
                  YOJ +
                  CLM_FREQ +
                  TIF, family = binomial(), data = train_df)
summary(BLR_all_vars)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CAR_AGE + MVR_PTS + YOJ + CLM_FREQ +
##     TIF, family = binomial(), data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8003  -0.7558  -0.6057   0.9552   2.4008
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.004828   0.162509   0.030 0.976299
## AGE         -0.019102   0.003313  -5.766 8.12e-09 ***
## CAR_AGE     -0.037685   0.005134  -7.341 2.12e-13 ***
## MVR_PTS      0.152214   0.013185  11.544  < 2e-16 ***
## YOJ         -0.023014   0.006747  -3.411 0.000648 ***
## CLM_FREQ     0.302335   0.024479  12.351  < 2e-16 ***
## TIF         -0.042139   0.007117  -5.921 3.21e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8303.6  on 7212  degrees of freedom
## Residual deviance: 7647.6  on 7206  degrees of freedom
## AIC: 7661.6
##
## Number of Fisher Scoring iterations: 4
```

Step through AIC scores to find best model

```
step_BLR = stepAIC(BLR_all_vars)
```

```
## Start:  AIC=7661.59
## TARGET_FLAG ~ AGE + CAR_AGE + MVR_PTS + YOJ + CLM_FREQ + TIF
##
##            Df Deviance    AIC
## <none>          7647.6 7661.6
## - YOJ       1   7659.1 7671.1
## - AGE       1   7681.1 7693.1
## - TIF       1   7683.7 7695.7
## - CAR_AGE   1   7702.5 7714.5
## - MVR_PTS   1   7781.4 7793.4
## - CLM_FREQ  1   7796.8 7808.8
```

```
summary(step_BLR)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CAR_AGE + MVR_PTS + YOJ + CLM_FREQ +
##     TIF, family = binomial(), data = train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8003  -0.7558  -0.6057   0.9552   2.4008
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.004828   0.162509   0.030 0.976299
## AGE         -0.019102   0.003313  -5.766 8.12e-09 ***
## CAR_AGE     -0.037685   0.005134  -7.341 2.12e-13 ***
## MVR_PTS      0.152214   0.013185  11.544  < 2e-16 ***
## YOJ         -0.023014   0.006747  -3.411 0.000648 ***
## CLM_FREQ     0.302335   0.024479  12.351  < 2e-16 ***
## TIF         -0.042139   0.007117  -5.921 3.21e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8303.6  on 7212  degrees of freedom
## Residual deviance: 7647.6  on 7206  degrees of freedom
## AIC: 7661.6
##
## Number of Fisher Scoring iterations: 4
```

Multiple Linear Regression

Multiple Linear Regression models with many variables

```
MLR_all_vars = lm(TARGET_AMT ~ AGE +
                  CAR_AGE +
                  MVR_PTS +
                  YOJ +
                  CLM_FREQ +
                  TIF, data = train_amt_df)
summary(MLR_all_vars)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ AGE + CAR_AGE + MVR_PTS + YOJ + CLM_FREQ +
##     TIF, data = train_amt_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6127  -3068  -1561    142  79965
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4636.72     920.08   5.039 5.11e-07 ***
## AGE            15.56      18.58   0.837    0.402
## CAR_AGE       -24.37      32.32  -0.754    0.451
## MVR_PTS       112.96      71.34   1.583    0.114
## YOJ            50.51      39.47   1.280    0.201
## CLM_FREQ     -135.92     148.13  -0.918    0.359
## TIF           -14.20      44.46  -0.319    0.749
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7618 on 1886 degrees of freedom
## Multiple R-squared:  0.003076,   Adjusted R-squared:  -9.516e-05
## F-statistic:  0.97 on 6 and 1886 DF,  p-value: 0.444
```

4.0 Select Models

We used Stepwise AIC (both backward and forward) to do model selection and ended with a Binary Logistic 7661.4

Appendix

- Diez, D.M., Barr, C.D., & Cetinkaya-Rundel, M. (2015). OpenIntro Statistics, Third Edition. Open Source. Print
- Faraway, J. J. (2015). Extending linear models with R, Second Edition. Boca Raton, FL: Chapman & Hall/CRC. Print
- https://www.sciencedirect.com/topics/computer-science/binary-logistic-regression (https://www.sciencedirect.com/topics/computer-science/binary-logistic-regression)
- https://bookdown.org/chua/ber642_advanced_regression/binary-logistic-regression.html (https://bookdown.org/chua/ber642_advanced_regression/binary-logistic-regression.html)
- http://wise.cgu.edu/wp-content/uploads/2016/07/Introduction-to-Logistic-Regression.pdf (http://wise.cgu.edu/wp-content/uploads/2016/07/Introduction-to-Logistic-Regression.pdf)

````
---
title: "DATA 621 - Business Analytics and Data Mining"
subtitle: "Homework 4"
author: "Ramnivas Singh"
date: "`r Sys.Date()`"
output:
  html_document:
    theme: default
    highlight: espresso
    toc: yes
    toc_depth: 5
    toc_float:
      collapsed: yes
  pdf_document:
    toc: yes
    toc_depth: '5'
  editor_options:
   chunk_output_type: inline
  always_allow_html: true
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
```

\clearpage

--------------------------------------------------------------------------------

# 1.0 Overview

![](overview.png)

## 1.1 Deliverables
![](deliverables.png)

\clearpage

--------------------------------------------------------------------------------

## Solution Steps & Approach

* Data Exploration : The auto insurance training dataset has 26 variables and
8161 observations. Of the variables, 24 of them are predictors for two
responses.

* Data Preparation : To prepare the data, we checked for any NA's or missing
values. There were none.

* Build Models : We built a model using all predictors as numerics.

* Select Models :Select a suitable model

*  Appendix
````

\clearpage

Import Libraries and Data

```{r echo=FALSE}
library(ggplot2)
library(dplyr)
library(corrplot)
library(MASS)
library(RCurl)
library(pROC)
library(tidyverse)
library(knitr)
library(corrgram)
library(reshape2)
library(Amelia)
```


```{r}
# Loading the data
train_df = read.csv("https://raw.githubusercontent.com/rnivas2028/MSDS/
Data621/HW4/insurance_training_data.csv")
test_df = read.csv("https://raw.githubusercontent.com/rnivas2028/MSDS/Data621/
HW4/insurance-evaluation-data.csv")
head(train_df)
```

2.0 Data Exploration & Preparation

The auto insurance training dataset has 26 variables and 8161 observations. Of the variables, 24 of them are predictors for two responses: TARGET_FLAG and TARGET_AMT is numerical.

To explore the training data, used:
 * Summary function to see means, medians, and quartiles of predictors
 * Str function to see the data type of each predictor
 * Explored TARGET_FLAG in relation to some other variables such as AGE and CAR_AGE
 * Looked at distribution of some numerical variables such as AGE and MVR_PTS

From the summary function, the TARGET_FLAG is binary and 26% of the 8161 records were accidents.
See a summary of each column in the train_df set
```{r train_dfing_data_summary}
# view a summary of all columns
summary(train_df)
```

Look at the data type of each variable
```{r}
```

````
# data type of predictors
str(train_df)
```

Look at the relationship between TARGET_FLAG and some of the numerical
variables.
```{r}
par(mfrow=c(1,2))
# plot response variable "target" against predictor variable "age" and
"car_age"
boxplot(AGE ~ TARGET_FLAG, train_df,
        main="Target vs Age",
        xlab="Target",
        ylab="Age")
boxplot(CAR_AGE ~ TARGET_FLAG, train_df,
        main="Target vs Car Age",
        xlab="Target",
        ylab="Car Age")
```

Look at the distribution of some numerical variables.
```{r}
h <- hist(train_df$AGE)
text(h$mids,h$counts,labels=h$counts)
```

```{r}
h <- hist(train_df$MVR_PTS)
text(h$mids,h$counts,labels=h$counts)
```

This data was prepared to build both a binary logistic model and a multiple
linear regression model. The binary logisitc model was used to predict the
TARGET_FLAG response variable and the multiple linear regression model was
used to predict the TARGET_AMT variable. Thus, there was a different training
dataset prepared for each model.

In both training datasets, all 948 records with at least one missing value
were removed.

Then, in the multiple linear regression training dataset all records with
TARGET_AMT = 0 were removed.

Check for NA's
```{r}
has_NA = names(which(sapply(train_df, anyNA)))
has_NA
```

Remove rows with NA's
train_df will be used for binary logistic regression model
```{r}
train_df <- train_df[complete.cases(train_df), ]
```
````
````

The training dataset for the binary logistic regression model was labelled train_df. The training dataset for the multiple linear regression model was titled train_amt_df.

Create train_amt_df dataframe for multiple linear regression model
```{r}
train_amt_df <- subset(train_df, TARGET_AMT > 0)
summary(train_amt_df$TARGET_FLAG)
```

Boxplots

The below boxplots show all of the variables listed in the dataset. This visualization will assist in showing how the data is spread for each variable.

```{r}
ggplot(melt(train_df), aes(x=factor(variable), y=value)) +
  facet_wrap(~variable, scale="free") +
  geom_boxplot()
```

Histograms

```{r}
ggplot(melt(train_df), aes(x=value)) +
  facet_wrap(~variable, scale="free") +
  geom_histogram(bins=50)
```

```{r echo=FALSE, message=FALSE, warning=FALSE}
corrgram(drop_na(train_df), order=TRUE,
         upper.panel=panel.cor, main="")
```

```{r echo=FALSE, message=FALSE, warning=FALSE}
missmap(train_df, main = "Missing values vs observed")
```

3.0 Build Models

First, we built two models using most predictors as numerics. Then we used the step AIC function to find the best variables for each model.

One model was a Binary Logistic Regression model for the TARGET_FLAG response titled step_BLR.
The second model was a Multiple Linear Regression for the TARGET_AMT response titled MLR_all_vars

````
## Binary Logistic Regression

```{r}
# preliminary exploration with one predictor
model1 <- glm(formula = TARGET_FLAG ~ AGE, family = binomial(), data =
train_df)
summary(model1)
```

Binary Logistic Regression Model with more variables
```{r}
BLR_all_vars = glm(TARGET_FLAG ~ AGE +
                  CAR_AGE +
                  MVR_PTS +
                  YOJ +
                  CLM_FREQ +
                  TIF, family = binomial(), data = train_df)
summary(BLR_all_vars)
```

Step through AIC scores to find best model
```{r}
step_BLR = stepAIC(BLR_all_vars)
summary(step_BLR)
```

## Multiple Linear Regression

Multiple Linear Regression models with many variables
```{r}
MLR_all_vars = lm(TARGET_AMT ~ AGE +
                  CAR_AGE +
                  MVR_PTS +
                  YOJ +
                  CLM_FREQ +
                  TIF, data = train_amt_df)
summary(MLR_all_vars)
```

\clearpage

--------------------------------------------------------------------------------

# 4.0 Select Models

We used Stepwise AIC (both backward and forward) to do model selection and
ended with a Binary Logistic 7661.4

--------------------------------------------------------------------------------

\clearpage

# Appendix
````

* Diez, D.M., Barr, C.D., & Cetinkaya-Rundel, M. (2015). OpenIntro Statistics, Third Edition. Open Source. Print

* Faraway, J. J. (2015). Extending linear models with R, Second Edition. Boca Raton, FL: Chapman & Hall/CRC. Print

* https://www.sciencedirect.com/topics/computer-science/binary-logistic-regression

* https://bookdown.org/chua/ber642_advanced_regression/binary-logistic-regression.html

* http://wise.cgu.edu/wp-content/uploads/2016/07/Introduction-to-Logistic-Regression.pdf