

# Homework 4

## Insurance Linear and Binary Regression

### Group 2

4/21/2021

## Contents

Assignment Overview . . . . .	1
Deliverables . . . . .	2
Task 1: Data Exploration . . . . .	2
Data Transformation . . . . .	5
Summary Statistics . . . . .	5
Missing Values . . . . .	7
Impute Missing Values . . . . .	8
Distributions . . . . .	9
Correlations . . . . .	11
Variable Plots . . . . .	13
Task 2: Data Preparation . . . . .	13
Task 3: Build Models . . . . .	14
Binary Logistic Regression . . . . .	14
Multiple Linear Regression . . . . .	21
Task 4: Select Models . . . . .	25
Error Calculations . . . . .	25
Model 1 Confusion Matrix . . . . .	25
Model 2 Confusion Matrix . . . . .	25
Model 3 Confusion Matrix . . . . .	26
Model Comparison . . . . .	27
Model of Choice . . . . .	28
Appendix . . . . .	29

**Group 2 members:** *Diego Correa, Jagdish Chhabria, Orli Khaimova, Richard Zheng, Stephen Haslett.*

## Assignment 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
KIDSDRV	# 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?
REVOKE	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 than 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
URBANITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

Figure 1: variable information

## 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.

## Task 1: Data Exploration

Describe the size and the variables in the insurance training data set.

```
## '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 ...
## $ KIDSDRV    : 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 283 ...
## $ 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 ...
```

There are 8,161 customers or observations in the insurance training data set with 26 different variables.

```

##   INDEX TARGET_FLAG TARGET_AMT KIDSDRV 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

```

## **Data Transformation**

As we can see from the above table, The training data set contains characters that will hinder our calculations so we need to remove or transform these. We will remove dollar signs from the `INCOME`, `HOME_VAL`, `BLUEBOOK`, and `OLDCLAIM` columns, and transform spaces to underscores in the `EDUCATION`, `JOB`, `CAR_TYPE`, `URBANICITY` columns.

Also the target variable `TARGET_FLAG` is actually a boolean variable. So it is better to convert it to a factor from its numeric data type.

## **Summary Statistics**

Roughly only 25% of the customers were involved in a car crash where the cost of the crash can be up to \$107,586. Most of the drivers are middle aged and are not single parents. A little over half of the customers are married and it is about equally split between females and males. Customers drive on average half an hour to work, whereas around a third of the customers use the cars for commercial use. Most of the customers also live in urban areas and have not had their license revoked in the last 7 years.

```

## TARGET_FLAG      TARGET_AMT          KIDSDRV          AGE           HOMEKIDS
## 0:6008      Min.    : 0     Min.   :0.0000     Min.   :16.00     Min.   :0.0000
## 1:2153      1st Qu.: 0     1st Qu.:0.0000     1st Qu.:39.00     1st Qu.:0.0000
##                   Median : 0     Median :0.0000     Median :45.00     Median :0.0000
##                   Mean   : 1504   Mean   :0.1711     Mean   :44.79     Mean   :0.7212
##                   3rd Qu.: 1036   3rd Qu.:0.0000     3rd Qu.:51.00     3rd Qu.:1.0000
##                   Max.   :107586   Max.   :4.0000     Max.   :81.00     Max.   :5.0000
##                                         NA's   :6
## YOJ            INCOME           PARENT1         HOME_VAL        MSTATUS
## Min.    : 0.0   Min.    : 0     No :7084     Min.   : 0     No :3267
## 1st Qu.: 9.0   1st Qu.: 28097 Yes:1077     1st Qu.: 0     Yes:4894
## Median :11.0   Median : 54028                    Median :161160
## Mean   :10.5   Mean   : 61898                    Mean   :154867
## 3rd Qu.:13.0   3rd Qu.: 85986                    3rd Qu.:238724
## Max.   :23.0   Max.   :367030                    Max.   :885282
## NA's   :454     NA's   :445                     NA's   :464
## SEX            EDUCATION        JOB             TRAVTIME
## F:4375      <High School:1203 Blue Collar :1825     Min.   : 5.00
## M:3786      Bachelors   :2242 Clerical   :1271     1st Qu.: 22.00
##                   High School :2330 Professional:1117     Median : 33.00
##                   Masters     :1658 Manager    : 988     Mean   : 33.49
##                   PhD         : 728 Lawyer     : 835     3rd Qu.: 44.00
##                   Student     : 712       : 712     Max.   :142.00
##                   (Other)     :1413
## CAR_USE        BLUEBOOK        TIF           CAR_TYPE
## Commercial:3029 Min.    :1500     Min.   : 1.000     Minivan   :2145
## Private   :5132  1st Qu.: 9280     1st Qu.: 1.000     Panel Truck: 676
##                   Median :14440     Median : 4.000     Pickup    :1389
##                   Mean   :15710     Mean   : 5.351     Sports Car : 907
##                   3rd Qu.:20850     3rd Qu.: 7.000     SUV       :2294
##                   Max.   :69740     Max.   :25.000     Van       : 750
##
## RED_CAR        OLDCLAIM        CLM_FREQ        REVOKED        MVR PTS
## no :5783      Min.    : 0     Min.   :0.0000     No :7161     Min.   : 0.000
## yes:2378     1st Qu.: 0     1st Qu.:0.0000     Yes:1000    1st Qu.: 0.000
##                   Median : 0     Median :0.0000                    Median : 1.000
##                   Mean   : 4037   Mean   :0.7986                    Mean   : 1.696
##                   3rd Qu.: 4636   3rd Qu.:2.0000                    3rd Qu.: 3.000
##                   Max.   :57037   Max.   :5.0000                    Max.   :13.000
##
## CAR_AGE          URBANICITY
## Min.   :-3.000  Highly Rural/ Rural:1669
## 1st Qu.: 1.000  Highly Urban/ Urban:6492
## Median : 8.000
## Mean   : 8.328
## 3rd Qu.:12.000
## Max.   :28.000
## NA's   :510

```

## Missing Values

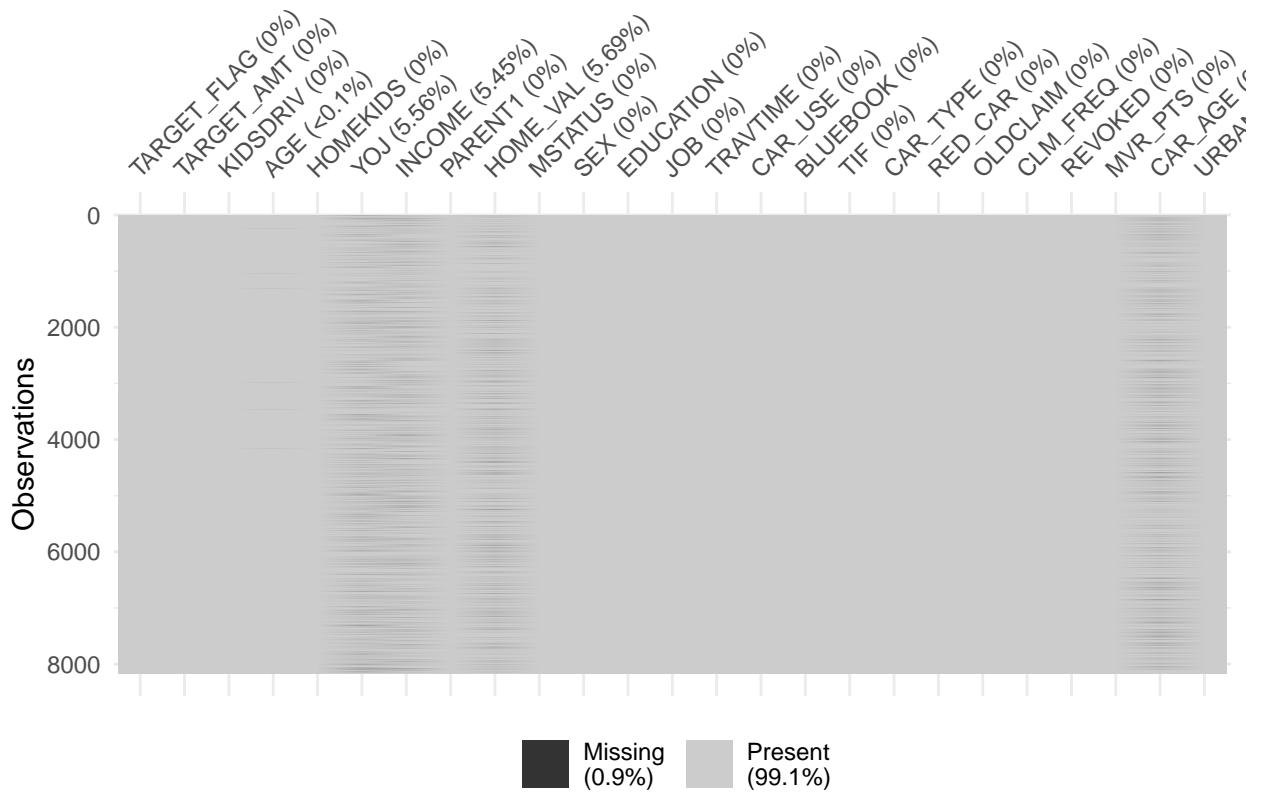
```
sapply(training_set, function(x) sum(is.na(x))) %>% sort(decreasing = TRUE) %>% kable() %>% kable_stylin
```

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

As we can see from the above table of missing values, three variables contain missing values. The “CAR\_AGE” variable has the largest amount of missing variables (510), followed by “HOME\_VALUE” (464), YOJ (454), “INCOME”(445) and finally “AGE” (6).

Below is a visual representation of the missing data.

```
vis_miss(training_set)
```



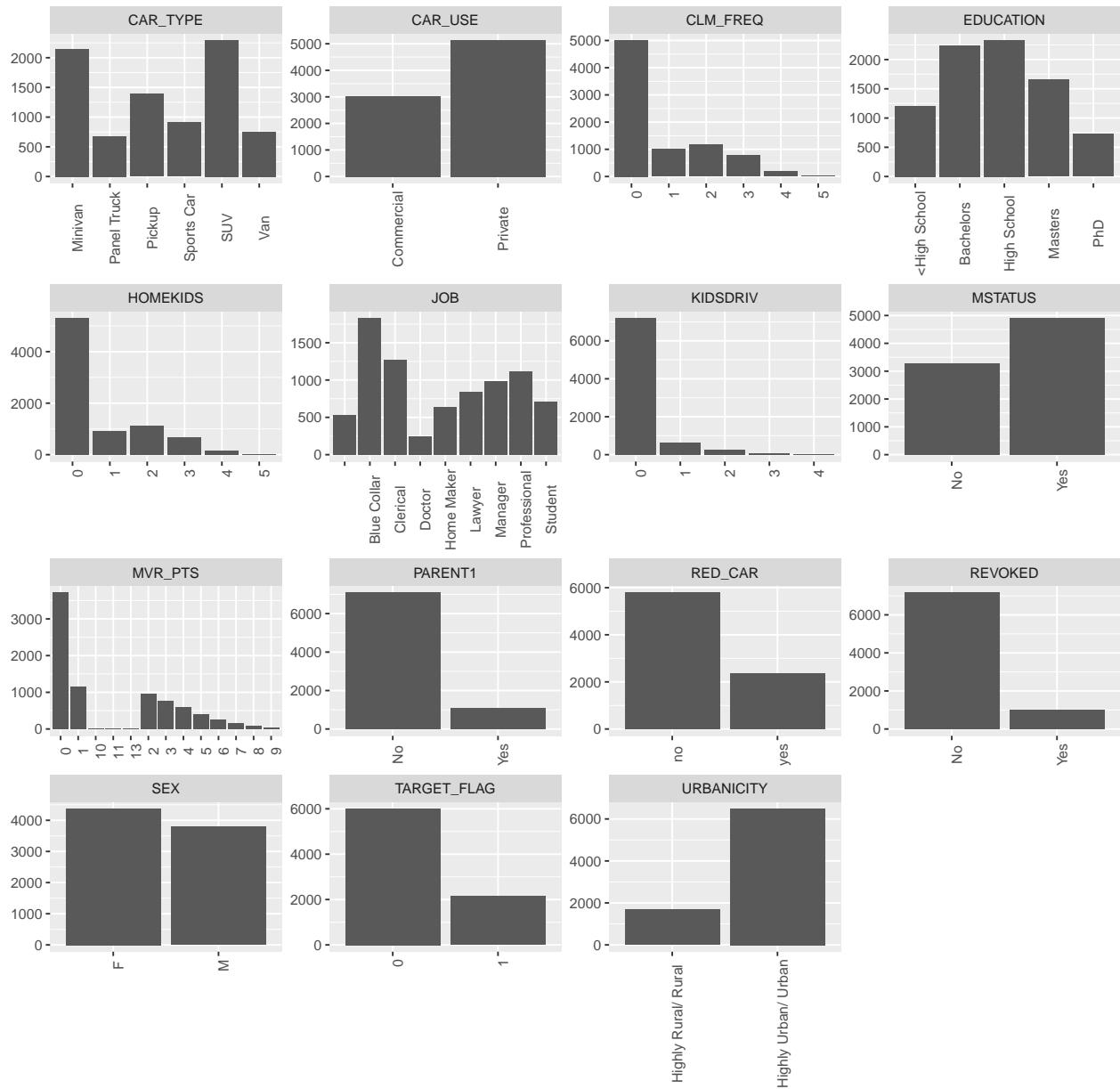
### Impute Missing Values

Data is imputed using the medians of the missing variables.

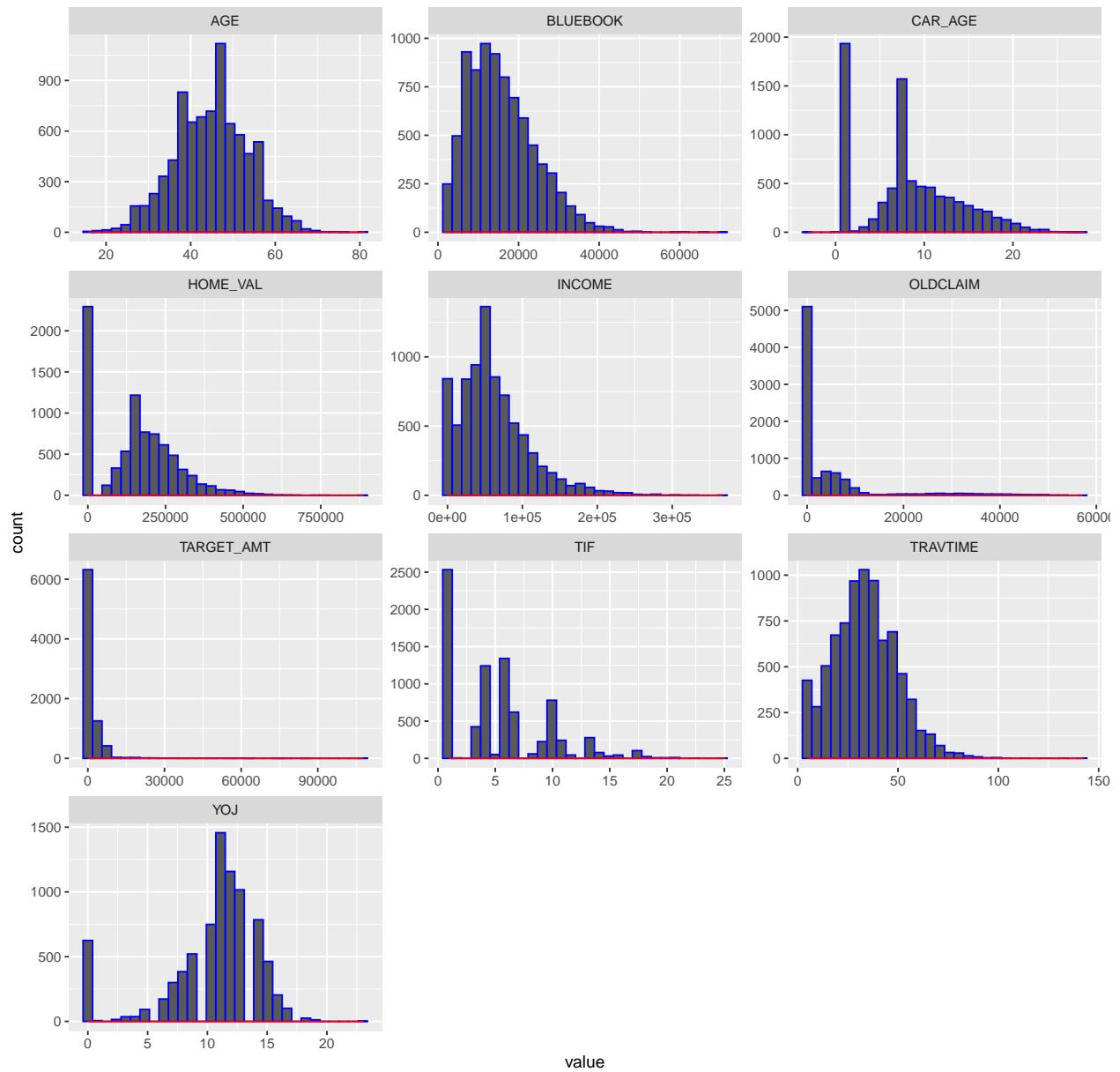
## Distributions

Having established that there are no missing values in the dataset, we will now take a look at the distribution profiles for each of the predictor variables. This will help us to decide which variables we should include in our final models.

**Distributions of Categorical Variables**



Distributions of Continuous Predictor Variables



Looking at the above distribution plots, we observe that there are a lot of skewed variables.

Table 1: Correlation of numeric predictors with the Target Amount

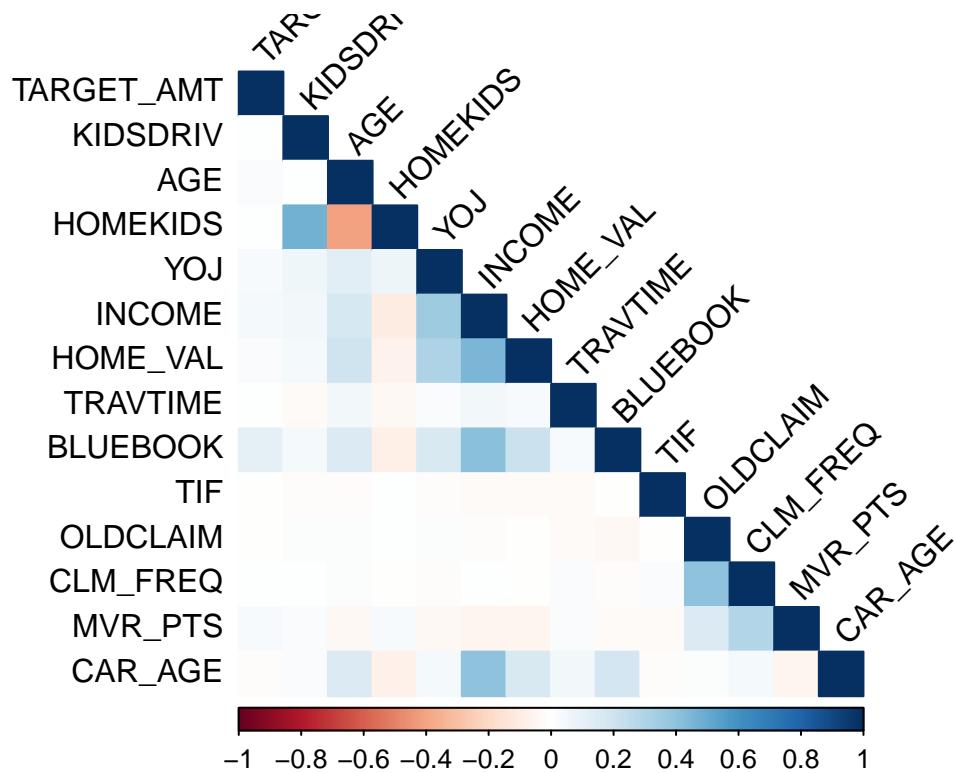
	.
TARGET_AMT	1.0000000
BLUEBOOK	0.1180862
INCOME	0.0440837
MVR_PTS	0.0398112
YOJ	0.0331754
HOME_VAL	0.0297703
AGE	0.0277963
CAR_AGE	-0.0131175
TIF	-0.0060110
OLDCLAIM	-0.0056563
TRAVTIME	0.0051768
CLM_FREQ	0.0019608
HOMEKIDS	0.0004689
KIDSDRIV	0.0000184

### Correlations

In order to run correlations, we first filter the training dataset for just the records where there was an actual accident and a non-zero insurance claim. This is because the existence of a claim is conditional on the occurrence of an accident. If we take the entire dataset for analyzing the correlation between the claim amount and the numeric variable, we would get skewed results.

After we filter the records to those with TARGET\_FLAG =1, then we further filter the columns to retain only the numeric predictor variables. In this case, that would result in the TARGET\_AMT as the response variable, and the following numeric predictor variables: AGE,BLUEBOOK,CAR\_AGE,CLM\_FREQ,HOMEKIDS,HOME\_VAL,INCOME,KIDSDRIV,MVR\_PTS,TIF,TRAVTIME, OLD CLAIM, YOJ.

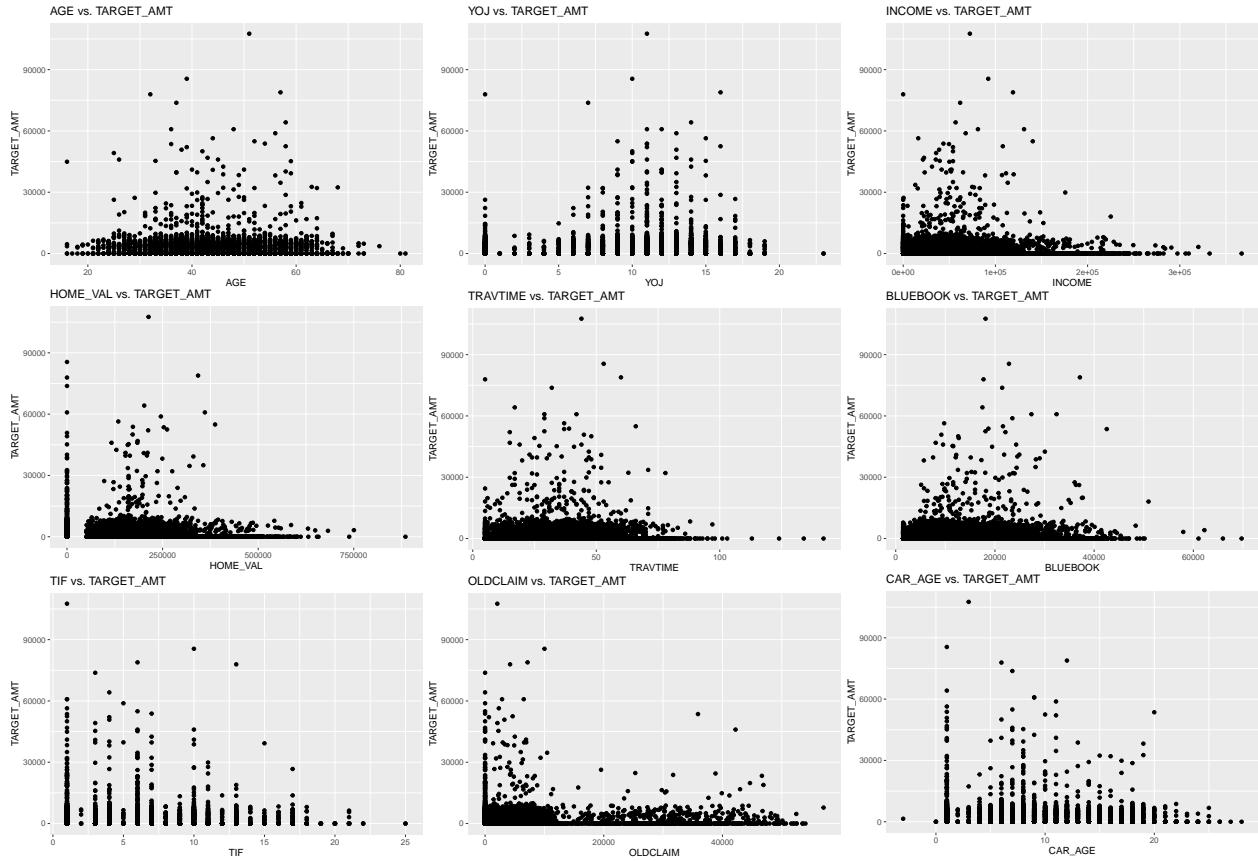
## Correlation Matrix of Training Set Predictor Variables



According to the correlation table and plot above, there is relatively higher correlation between the **TARGET\_AMT** and the **BLUEBOOK** value. We would have expected to see some correlation between the **TARGET\_AMT** and the **AGE** of the vehicle, but that doesn't seem to be the case here. In any case, the age of the car would factor into the **BLUEBOOK** value, so including both variables in a linear model with **TARGET\_AMT** is likely to introduce multi-collinearity.

## Variable Plots

Scatter plots of each variable versus the target variable.



## Task 2: Data Preparation

Describe how you have transformed the data by changing the original variables or creating new variables.

Variable transformations (such as log, square root, quadratic, inverse, etc) will be applied during model building. Also, variables that were created or refactored will also be included in the model building process.

## Task 3: Build Models

Using the training data, build at least three different binary logistic regression models, using different variables (or the same variables with different transformations).

### Binary Logistic Regression

#### Model 1

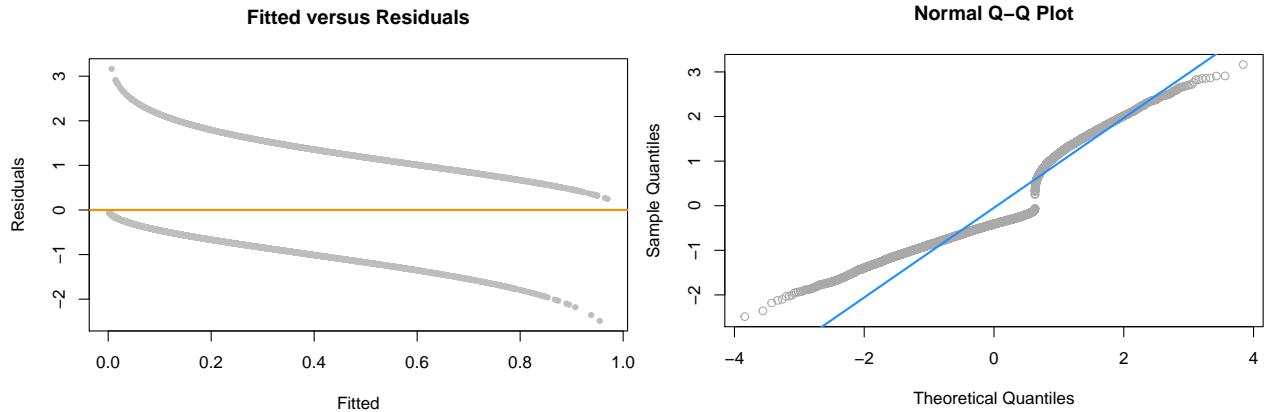
- EDUCATION was grouped as “College” and “No College”
- AGE was grouped as “Under 25” and “25 and Over”
- JOB was grouped as “Careered” and “Not Careered”
- MINIVAN grouped the CAR\_TYPE as “yes” or “no” if it was a minivan
- RED\_CAR, CAR\_AGE, SEX, YOJ , HOMEKIDS were removed due to high p value

```
##  
## Call:  
## glm(formula = TARGET_FLAG ~ KIDSDRIV + AGE + INCOME + PARENT1 +  
##       HOME_VAL + MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE +  
##       BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ + REVOKED + MVR PTS +  
##       URBANICITY + MINIVAN, family = "binomial", data = .)  
##  
## Deviance Residuals:  
##      Min        1Q     Median        3Q       Max  
## -2.4872  -0.7243  -0.4116   0.6333   3.1652  
##  
## Coefficients:  
##                                     Estimate Std. Error z value  
## (Intercept)                   -2.5384959056 0.1776677325 -14.288  
## KIDSDRIVyes                  0.6968909256 0.0875701855  7.958  
## AGEUnder 25                  0.7441640781 0.2774639894  2.682  
## INCOME                         -0.0000038457 0.0000009847 -3.905  
## PARENT1Yes                    0.4133399254 0.0946701431  4.366  
## HOME_VAL                       -0.0000011871 0.0000003309 -3.588  
## MSTATUSYes                     -0.4897519101 0.0789660463 -6.202  
## EDUCATIONNo College            0.5091517578 0.0684335720  7.440  
## JOBNot Careered               0.1932319833 0.0794187358  2.433  
## TRAVTIME                        0.0150436632 0.0018676816  8.055  
## CAR_USEPrivate                -0.7798798062 0.0637471094 -12.234  
## BLUEBOOK                        -0.0000258798 0.0000039580 -6.539  
## TIF                            -0.0552053607 0.0072966926 -7.566  
## OLDCLAIM                        -0.0000140356 0.0000038716 -3.625  
## CLM_FREQ                         0.1976855328 0.0282840192  6.989  
## REVOKEDYes                     0.9049111771 0.0904723959 10.002  
## MVR PTS                          0.1182584323 0.0135563565  8.723  
## URBANICITYHighly Urban/ Urban  2.3406979757 0.1128027392 20.750  
## MINIVANyes                     -0.6739533263 0.0738433531 -9.127  
##                                     Pr(>|z|)  
## (Intercept) < 0.0000000000000002 ***  
## KIDSDRIVyes 0.00000000000001747 ***  
## AGEUnder 25          0.007318 **  
## INCOME             0.000094095841950242 ***  
## PARENT1Yes        0.000012648083830509 ***  
## HOME_VAL           0.000334 ***  
## MSTATUSYes         0.000000000557299288 ***  
## EDUCATIONNo College 0.000000000000100619 ***
```

```

## JOBNot Careered          0.014971 *
## TRAVTIME                 0.000000000000000797 ***
## CAR_USEPrivate            < 0.0000000000000002 ***
## BLUEBOOK                  0.00000000062104046 ***
## TIF                       0.000000000000038547 ***
## OLDCLAIM                  0.000289 ***
## CLM_FREQ                  0.000000000002762585 ***
## REVOKEDYes                < 0.0000000000000002 ***
## MVR_PTS                   < 0.0000000000000002 ***
## URBANICITYHighly Urban/ Urban < 0.0000000000000002 ***
## MINIVANYes                < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7366.3  on 8142  degrees of freedom
## AIC: 7404.3
##
## Number of Fisher Scoring iterations: 5

```



## Model 2

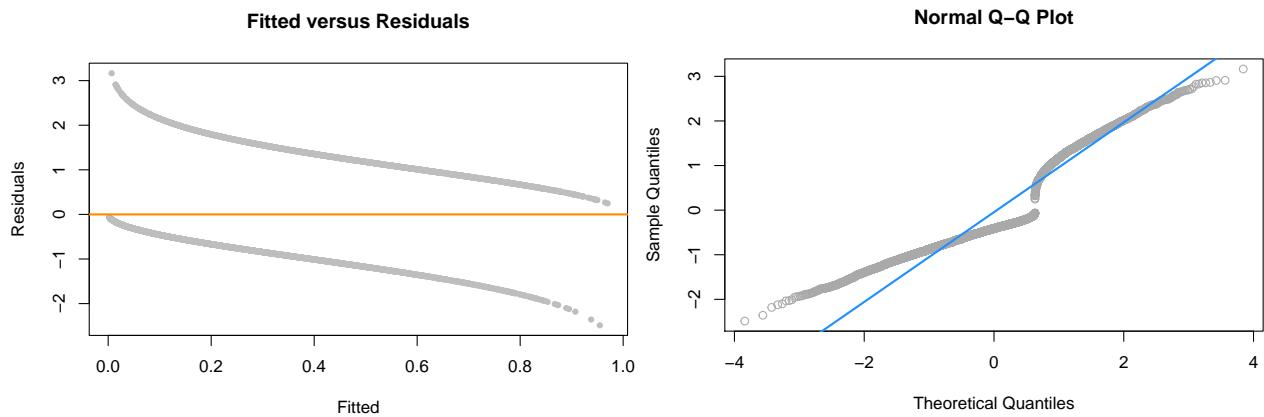
```

## 
## Call:
## glm(formula = TARGET_FLAG ~ INCOME + PARENT1 + HOME_VAL + MSTATUS +
##       EDUCATION + TRAVTIME + CAR_USE + TIF + CAR_TYPE + REVOKED +
##       URBANICITY, family = binomial(link = "logit"), data = training_set)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -2.2151   -0.7437   -0.4504    0.7317    3.0341
##
## Coefficients:
##                               Estimate Std. Error z value
## (Intercept)                -2.5097700872 0.1754970690 -14.301
## INCOME                     -0.0000057209 0.0000009443  -6.058
## PARENT1Yes                  0.6633559231 0.0888177757  7.469
## HOME_VAL                   -0.0000015424 0.0000003245  -4.754
## MSTATUSYes                 -0.3742337414 0.0763556681  -4.901
## EDUCATIONBachelors          -0.5921668518 0.0956906561  -6.188
## EDUCATIONHigh School         -0.0809393935 0.0896475530  -0.903
## EDUCATIONMasters             -0.6094601110 0.1071458499  -5.688
## EDUCATIONPhD                 0.6227665192 0.1443151490  -4.315
## TRAVTIME                      0.0156280286 0.0018302862  8.539
## CAR_USEPrivate                0.8987050507 0.0716420077 -12.544
## TIF                           -0.0566834590 0.0071612998  -7.915
## CAR_TYPEPanel Truck            0.2714278527 0.1280630609  2.119
## CAR_TYPEPickup                 0.5634830682 0.0953288916  5.911
## CAR_TYPESports Car              1.0726682146 0.1025601762 10.459
## CAR_TYPESUV                    0.8268973605 0.0823284992 10.044
## CAR_TYPEVan                     0.5223542406 0.1154076129  4.526
## REVOKEDYes                     0.7609781315 0.0782622562  9.723
## URBANICITYHighly Urban/ Urban  2.4740662702 0.1088123580 22.737
## 
## Pr(>|z|)
## (Intercept) < 0.0000000000000002 ***
## INCOME        0.00000000137465496 ***
## PARENT1Yes    0.00000000000008097 ***
## HOME_VAL      0.00000199866822317 ***
## MSTATUSYes    0.00000095257208772 ***
## EDUCATIONBachelors 0.0000000060798914 ***
## EDUCATIONHigh School           0.367
## EDUCATIONMasters               0.00000001284339097 ***
## EDUCATIONPhD                  0.00001593695169238 ***
## TRAVTIME                      < 0.0000000000000002 ***
## CAR_USEPrivate                 < 0.0000000000000002 ***
## TIF                           0.0000000000000247 ***
## CAR_TYPEPanel Truck             0.034 *
## CAR_TYPEPickup                 0.00000000340167224 ***
## CAR_TYPESports Car              < 0.0000000000000002 ***
## CAR_TYPESUV                    < 0.0000000000000002 ***
## CAR_TYPEVan                     0.00000600628479796 ***
## REVOKEDYes                     < 0.0000000000000002 ***
## URBANICITYHighly Urban/ Urban < 0.0000000000000002 ***
## ---
```

```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 9418  on 8160  degrees of freedom
## Residual deviance: 7635  on 8142  degrees of freedom
## AIC: 7673
##
## Number of Fisher Scoring iterations: 5

```



### Model 3

```

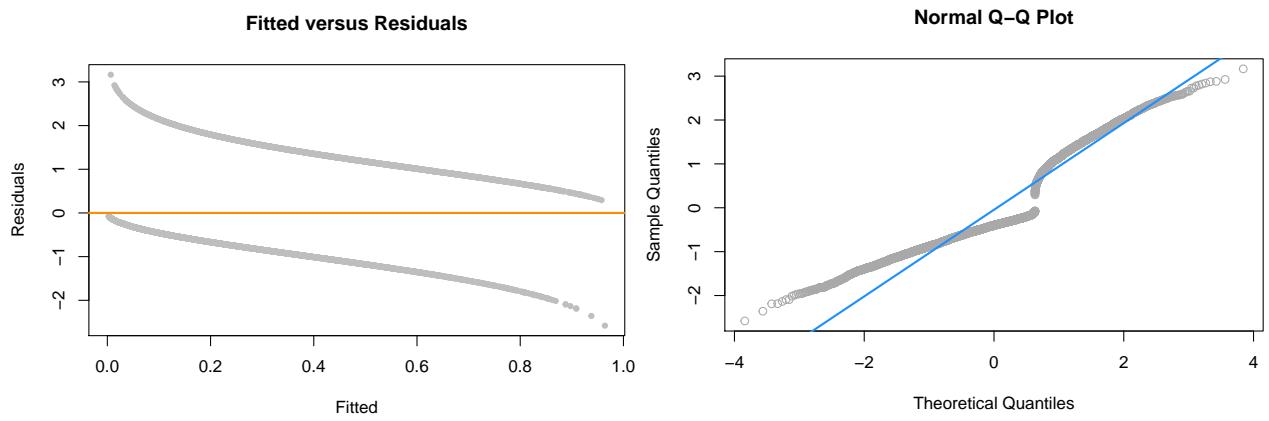
## 
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT - INCOME, family = "binomial",
##      data = training_set)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.5795 -0.7115 -0.4005  0.6204  3.1652
##
## Coefficients:
##                               Estimate Std. Error z value
## (Intercept)             -3.0679518334 0.3373699802 -9.094
## KIDSDRV               0.3830469969 0.0611286396  6.266
## AGE                  -0.0004733507 0.0040185220 -0.118
## HOMEKIDS              0.0484593218 0.0370831235  1.307
## YOJ                  -0.0132936551 0.0085550583 -1.554
## PARENT1Yes            0.3980213607 0.1094482892  3.637
## HOME_VAL              -0.0000016541 0.0000003205 -5.162
## MSTATUSYes            -0.4382163188 0.0819322446 -5.349
## SEXM                 0.0869811203 0.1119515249  0.777
## EDUCATIONBachelors   -0.4336590913 0.1143016429 -3.794
## EDUCATIONHigh School  -0.0036236357 0.0947363450 -0.038
## EDUCATIONMasters      -0.3600356708 0.1773934832 -2.030
## EDUCATIONPhD          -0.3334875706 0.2079105373 -1.604
## JOBBlue Collar       0.3637218921 0.1848138578  1.968
## JOBClerical           0.5159560435 0.1938419886  2.662
## JOBDoctor             -0.4214004991 0.2662130405 -1.583
## JOBHome Maker         0.4146033637 0.2023150912  2.049
## JOBLawyer             0.1442181947 0.1688899097  0.854
## JOBManager            -0.5273877157 0.1708828315 -3.086
## JOBProfessional       0.2034688816 0.1779215166  1.144
## JOBStudent            0.3607450708 0.2095168920  1.722
## TRAVTIME              0.0144384980 0.0018825719  7.670
## CAR_USEPrivate        -0.7555539047 0.0916823913 -8.241
## BLUEBOOK              -0.0000231194 0.0000052013 -4.445
## TIF                   -0.0550026436 0.0073328336 -7.501
## CAR_TYPEPanel Truck   0.5567027529 0.1615875110  3.445
## CAR_TYPEPickup        0.5575279306 0.1006672494  5.538
## CAR_TYPESports Car   1.0271619017 0.1298242768  7.912
## CAR_TYPESUV           0.7707255522 0.1112524500  6.928
## CAR_TYPEVan           0.6103712250 0.1263398360  4.831
## RED_CARyes            -0.0086544858 0.0863095906 -0.100
## OLDCLAIM              -0.0000137280 0.0000039072 -3.514
## CLM_FREQ              0.1966864045 0.0285294183  6.894
## REVOKEDYes            0.8851089489 0.0912284023  9.702
## MVR PTS               0.1144788910 0.0135971123  8.419
## CAR_AGE               -0.0020306511 0.0075320244 -0.270
## URBANICITYHighly Urban/ Urban  2.3853025598 0.1127354597 21.158
## 
## (Intercept)             Pr(>|z|)
## < 0.0000000000000002 ***
## KIDSDRV                  0.0000000036985964 ***
## AGE                      0.906232

```

```

## HOMEKIDS          0.191289
## YOJ              0.120210
## PARENT1Yes       0.000276 ***
## HOME_VAL         0.00000024477789715 ***
## MSTATUSYes       0.00000008867596153 ***
## SEXM             0.437186
## EDUCATIONBachelors 0.000148 ***
## EDUCATIONHigh School 0.969489
## EDUCATIONMasters 0.042398 *
## EDUCATIONPhD    0.108715
## JOBBLue Collar   0.049063 *
## JOBClerical      0.007774 **
## JOBDoctor        0.113434
## JOBHome Maker    0.040433 *
## JOBLawyer         0.393150
## JOBManager       0.002027 **
## JOBProfessional  0.252795
## JOBStudent       0.085107 .
## TRAVTIME         0.00000000000001726 ***
## CAR_USEPrivate   < 0.000000000000002 ***
## BLUEBOOK         0.00000879282137125 ***
## TIF              0.00000000000006339 ***
## CAR_TYPEPanel Truck 0.000571 ***
## CAR_TYPEPickup   0.00000003053784090 ***
## CAR_TYPESports Car 0.0000000000000253 ***
## CAR_TYPESUV     0.00000000000427684 ***
## CAR_TYPEVan     0.00000135722253494 ***
## RED_CARyes      0.920128
## OLDCLAIM        0.000442 ***
## CLM_FREQ         0.0000000000541836 ***
## REVOKEDYes      < 0.000000000000002 ***
## MVR PTS         < 0.000000000000002 ***
## CAR AGE          0.787466
## URBANICITYHighly Urban/ Urban < 0.000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7307.8  on 8124  degrees of freedom
## AIC: 7381.8
##
## Number of Fisher Scoring iterations: 5

```

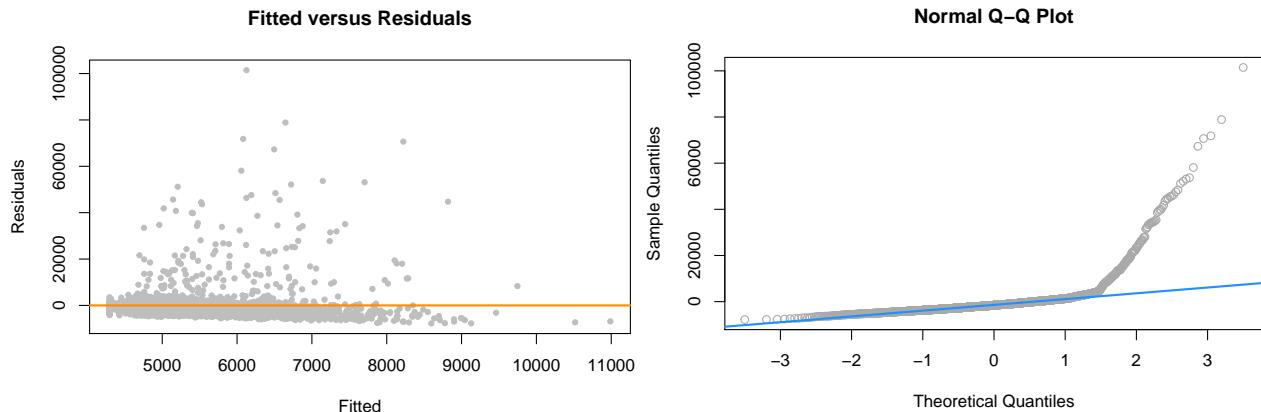


## Multiple Linear Regression

### Model 4

Here we run a linear regression (for the records that represent an actual accident) i.e. TARGET\_FLAG=1, between the TARGET\_AMT and the BLUEBOOK variable.

```
## 
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK, data = training_set_numeric)
## 
## Residuals:
##     Min      1Q Median      3Q     Max 
## -7757   -3083  -1541    295 101459 
## 
## Coefficients:
##             Estimate Std. Error t value     Pr(>|t|)    
## (Intercept) 4131.65436  329.48405 12.540 < 0.0000000000000002 *** 
## BLUEBOOK      0.11017   0.01997   5.515     0.00000039 ***  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 7691 on 2151 degrees of freedom 
## Multiple R-squared:  0.01394,   Adjusted R-squared:  0.01349 
## F-statistic: 30.42 on 1 and 2151 DF,  p-value: 0.000000039
```



## Model 5

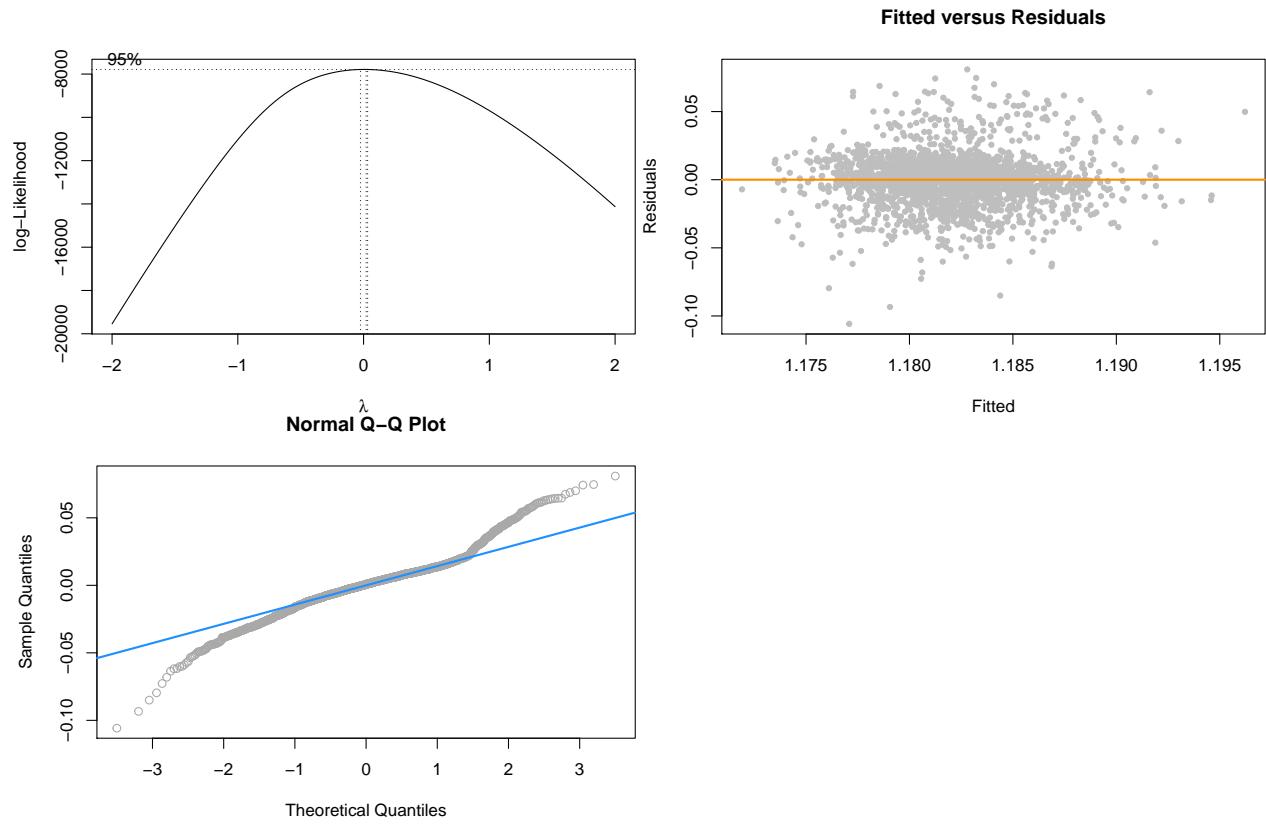
Using the Boxcox method for linear model.

```
##
## Call:
## lm(formula = TARGET_AMT^(1) ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.105762 -0.009634  0.000749  0.009599  0.080927 
##
## Coefficients:
##                               Estimate Std. Error t value
## (Intercept)           1.174097356347 0.004994461397 235.080
## KIDSDRV             -0.000766473804 0.000795200280 -0.964
## AGE                  0.000045046895 0.000053353774  0.844
## HOMEKIDS            0.0005589595294 0.000520290239  1.074
## YOJ                 -0.000022475456 0.000123521645 -0.182
## INCOME              -0.000000031826 0.000000016935 -1.879
## PARENT1Yes          0.000574179574 0.001475101290  0.389
## HOME_VAL            0.000000003635 0.000000005075  0.716
## MSTATUSYes          -0.002348666282 0.001239526656 -1.895
## SEXM                0.002268833851 0.001648695394  1.376
## EDUCATIONBachelors -0.000694956993 0.001611004017 -0.431
## EDUCATIONHigh School 0.000181505648 0.001292348608  0.140
## EDUCATIONMasters    0.003601788764 0.002720875911  1.324
## EDUCATIONPhD        0.005937436887 0.003293860792  1.803
## JOBBlue Collar     0.001502882033 0.002878623843  0.522
## JOBClerical         0.001311249797 0.003021539147  0.434
## JOBDoctor          -0.001035979944 0.004426297830 -0.234
## JOBHome Maker      -0.001188889198 0.003180809525 -0.374
## JOBLawyer           -0.000275474113 0.002585417744 -0.107
## JOBManager          0.000404067259 0.002677330932  0.151
## JOBProfessional    0.002547884962 0.002836000195  0.898
## JOBStudent          0.000493947066 0.003229747475  0.153
## TRAVTIME            -0.000006113871 0.000027832363 -0.220
## CAR_USEPrivate     -0.000342473962 0.001310185426 -0.261
## BLUEBOOK            0.000000287558 0.000000076694  3.749
## TIF                 -0.000046975803 0.000106792688 -0.440
## CAR_TYPEPanel Truck 0.000009851578 0.002412559681  0.004
## CAR_TYPEPickup     0.000676285873 0.001499046060  0.451
## CAR_TYPESports Car 0.001400828129 0.001884367093  0.743
## CAR_TYPESUV        0.002215842438 0.001674775528  1.323
## CAR_TYPEVan        -0.000284207621 0.001935958933 -0.147
## RED_CARyes         0.000495668458 0.001247007441  0.397
## OLDCLAIM            0.000000105747 0.000000056854  1.860
## CLM_FREQ            -0.000860607923 0.000396908405 -2.168
## REVOKEDYes         -0.002298635135 0.001297695301 -1.771
## MVR PTS             0.000355027712 0.000172119610  2.063
## CAR_AGE             -0.000058375161 0.000110405223 -0.529
## URBANICITYHighly Urban/ Urban 0.001352819972 0.001899489885  0.712
## 
## (Intercept)           Pr(>|t|) < 0.0000000000000002 ***
## KIDSDRV                  0.335219
```

```

## AGE                      0.398594
## HOMEKIDS                 0.282801
## YOJ                       0.855635
## INCOME                     0.060343 .
## PARENT1Yes                  0.697132
## HOME_VAL                    0.473873
## MSTATUSYes                   0.058254 .
## SEXM                        0.168924
## EDUCATIONBachelors          0.666235
## EDUCATIONHigh School          0.888321
## EDUCATIONMasters              0.185726
## EDUCATIONPhD                  0.071597 .
## JOBBlue Collar                0.601667
## JOBClerical                     0.664356
## JOBDoctor                      0.814968
## JOBHome Maker                  0.708613
## JOBLawyer                      0.915157
## JOBManager                     0.880052
## JOBProfessional                  0.369070
## JOBStudent                      0.878463
## TRAVTIME                      0.826151
## CAR_USEPrivate                  0.793815
## BLUEBOOK                      0.000182 ***
## TIF                           0.660070
## CAR_TYPEPanel Truck             0.996742
## CAR_TYPEPickup                  0.651932
## CAR_TYPESports Car              0.457325
## CAR_TYPESUV                      0.185956
## CAR_TYPEVan                      0.883300
## RED_CARyes                      0.691049
## OLDCLAIM                      0.063029 .
## CLM_FREQ                         0.030249 *
## REVOKEDYes                      0.076651 .
## MVR PTS                         0.039265 *
## CAR AGE                          0.597044
## URBANICITYHighly Urban/ Urban      0.476418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01932 on 2115 degrees of freedom
## Multiple R-squared:  0.02649,   Adjusted R-squared:  0.009457
## F-statistic: 1.555 on 37 and 2115 DF,  p-value: 0.01792

```



## Task 4: Select Models

Decide on the criteria for selecting the best binary logistic regression model.

### Error Calculations

#### Model 1 Confusion Matrix

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   0     1
##           0 5570 1277
##           1  438  876
##
##                 Accuracy : 0.7899
##                 95% CI : (0.7809, 0.7986)
## No Information Rate : 0.7362
## P-Value [Acc > NIR] : < 0.00000000000000022
##
##                 Kappa : 0.3817
##
## Mcnemar's Test P-Value : < 0.00000000000000022
##
##                 Sensitivity : 0.9271
##                 Specificity : 0.4069
## Pos Pred Value : 0.8135
## Neg Pred Value : 0.6667
## Prevalence : 0.7362
## Detection Rate : 0.6825
## Detection Prevalence : 0.8390
## Balanced Accuracy : 0.6670
##
## 'Positive' Class : 0
##
```

#### Model 2 Confusion Matrix

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   0     1
##           0 5582 1382
##           1  426  771
##
##                 Accuracy : 0.7785
##                 95% CI : (0.7693, 0.7874)
## No Information Rate : 0.7362
## P-Value [Acc > NIR] : < 0.00000000000000022
##
##                 Kappa : 0.3349
##
## Mcnemar's Test P-Value : < 0.00000000000000022
##
##                 Sensitivity : 0.9291
##                 Specificity : 0.3581
```

```

##           Pos Pred Value : 0.8016
##           Neg Pred Value : 0.6441
##           Prevalence : 0.7362
##           Detection Rate : 0.6840
##   Detection Prevalence : 0.8533
##           Balanced Accuracy : 0.6436
##
##           'Positive' Class : 0
##

```

### Model 3 Confusion Matrix

```

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0     1
##           0 5561 1230
##           1  447  923
##
##           Accuracy : 0.7945
##           95% CI : (0.7856, 0.8032)
##   No Information Rate : 0.7362
##   P-Value [Acc > NIR] : < 0.00000000000000022
##
##           Kappa : 0.4011
##
##   Mcnemar's Test P-Value : < 0.00000000000000022
##
##           Sensitivity : 0.9256
##           Specificity : 0.4287
##           Pos Pred Value : 0.8189
##           Neg Pred Value : 0.6737
##           Prevalence : 0.7362
##           Detection Rate : 0.6814
##   Detection Prevalence : 0.8321
##           Balanced Accuracy : 0.6772
##
##           'Positive' Class : 0
##

```

## Model Comparison

```
##                                     Model 1 Model 2 Model 3
## accuracy                           0.7899  0.7785  0.7945
## classification error rate          0.2101  0.2215  0.2055
## precision                          0.6667  0.6441  0.6737
## sensitivity                         0.9271  0.9291  0.9256
## specificity                         0.4069  0.3581  0.4287
## F1 score                            0.5053  0.4603  0.5240
## AUC                                 0.6670  0.6436  0.6772
```

## Model of Choice

- Picked Models 2 and 4 because they had the best results

```

evaluation_set = evaluation_set %>%
  mutate(AGE = ifelse(is.na(AGE), median(training_set$AGE, na.rm = TRUE), AGE),
        YOJ = ifelse(is.na(YOJ), median(training_set$YOJ, na.rm = TRUE), YOJ),
        INCOME = ifelse(is.na(INCOME), median(training_set$INCOME, na.rm = TRUE), INCOME),
        HOME_VAL = ifelse(is.na(HOME_VAL), median(training_set$HOME_VAL, na.rm = TRUE), HOME_VAL),
        CAR_AGE = ifelse(is.na(CAR_AGE), median(training_set$CAR_AGE, na.rm = TRUE), CAR_AGE))
evaluation_set$TARGET_FLAG = to_binary(predict(model2, evaluation_set, type = "response"), 0.5)
evaluation_set$TARGET_AMT = predict(model4, evaluation_set)
head(evaluation_set)

##   TARGET_FLAG TARGET_AMT KIDSDRV AGE HOMEKIDS YOJ INCOME PARENT1 HOME_VAL
## 1          0  6552.017     0  48      0  11  52881    No      0
## 2          0   6217.110     1  40      1  11  50815   Yes      0
## 3          0   4781.638     0  44      2  12  43486   Yes      0
## 4          0   5148.493     0  35      2  11  21204   Yes      0
## 5          0   5830.425     0  59      0  12  87460    No      0
## 6          0   6958.532     0  46      0  14  54028    No  207519
##   MSTATUS SEX EDUCATION           JOB TRAVTIME   CAR_USE BLUEBOOK TIF
## 1     No   M   Bachelors   Manager      26 Private  21970    1
## 2     No   M High School Manager      21 Private  18930    6
## 3     No   F High School Blue Collar      30 Commercial  5900  10
## 4     No   M High School Clerical      74 Private  9230    6
## 5     No   M High School Manager      45 Private  15420    1
## 6   Yes   M   Bachelors Professional      7 Commercial  25660    1
##   CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR PTS CAR_AGE
## 1       Van yes      0      0    No     2    10
## 2   Minivan no     3295      1   No     2     1
## 3      SUV no      0      0   No     0    10
## 4   Pickup no      0      0 Yes     0     4
## 5   Minivan yes    44857      2   No     4     1
## 6 Panel Truck no    2119      1   No     2    12
##   URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
## 3 Highly Rural/ Rural
## 4 Highly Rural/ Rural
## 5 Highly Urban/ Urban
## 6 Highly Urban/ Urban

```

## Appendix

```
# =====
# Load Required Libraries
# =====

knitr::opts_chunk$set(echo = TRUE, warning = FALSE, include = TRUE)

# Load required libraries.
library(tidyverse)
library(caret)
library(pROC)
library(grid)
library(Amelia)
library(ggplot2)
library(kableExtra)
library(corrplot)
library(reshape2)
library(e1071)
library(visdat)
library(mice)
library(MASS)

# =====
# Load The Datasets and Look at the Structure of the Data
# =====

# Pull in the provided crime training and evaluation datasets.
training_set <- read.csv('https://raw.githubusercontent.com/Jagdish16/CUNY_DATA_621/main/homework_4/inst'
evaluation_set <- read.csv('https://raw.githubusercontent.com/Jagdish16/CUNY_DATA_621/main/homework_4/inst

# List the structure of the training dataset.
str(training_set)

# =====
# Summarize the Training Data
# =====

# Summarize the training dataset.
summary(training_set)

# =====
# Cleaning Data
# =====

training_set <- training_set %>%
  # remove the index variable
  dplyr::select(-INDEX) %>%
  # remove characters such as $ | , | _Z
  mutate_all(~ str_remove_all(., "\$\|,\|z_")) %>%
  # convert spaces to underscores
  # mutate_all(~ str_replace_all(., " ", "_")) %>%
```

```
# convert types automatically from chr type
type.convert(.) %>%
# convert TARGET_FLAG to factor
mutate(TARGET_FLAG = as.factor(TARGET_FLAG))

evaluation_set <- evaluation_set %>%
# remove the index variable
dplyr::select(-INDEX) %>%
# remove characters such as $ | , | _Z
mutate_all(~ str_remove_all(., "\$\|,\|z_")) %>%
# convert spaces to underscores
# mutate_all(~ str_replace_all(., " ", "_")) %>%
# convert types automatically from chr type
type.convert(.) %>%
# convert TARGET_FLAG to factor
mutate(TARGET_FLAG = as.factor(TARGET_FLAG))
```

```
# =====
# Check for Missing Values
# =====

# Check for missing values using the Amelia package's missmap() function.
sapply(training_set, function(x) sum(is.na(x))) %>% sort(decreasing = TRUE) %>% kable() %>% kable_stylin
```

As we can see from the above table of missing values, three variables contain missing values. The "CAR\_

Below is a visual representation of the missing data.

```
vis_miss(training_set)
```

```
# =====
# Imputing Missing Values
# =====

training_set <- training_set %>%
  mutate(AGE = ifelse(is.na(AGE), median(training_set$AGE, na.rm =TRUE), AGE),
         YOJ = ifelse(is.na(YOJ), median(training_set$YOJ, na.rm =TRUE), YOJ),
         INCOME = ifelse(is.na(INCOME), median(training_set$INCOME, na.rm =TRUE), INCOME),
         HOME_VAL = ifelse(is.na(HOME_VAL), median(training_set$HOME_VAL, na.rm =TRUE), HOME_VAL),
         CAR_AGE = ifelse(is.na(CAR_AGE), median(training_set$CAR_AGE, na.rm =TRUE), CAR_AGE))
```

```

# =====
# Distribution Plots
# =====

# Using the Dplyr package, massage the data by removing the target value prior
# to plotting a histogram for each predictor variable.
training_set %>%
  mutate(CLM_FREQ = as.factor(CLM_FREQ),
         HOMEKIDS = as.factor(HOMEKIDS),
         KIDSDRIV = as.factor(KIDSDRIV),
         MVR PTS = as.factor(MVR PTS)) %>%
  select_if(is.factor) %>%
  gather(variable, value) %>%
  ggplot(., aes(x = value)) +
  geom_histogram(bins = 25, stat = 'count') +
  labs(title = 'Distributions of Categorical Variables') +
  facet_wrap(~variable, scales = "free", ncol = 4) +
  labs(x = element_blank(), y = element_blank()) +
  theme(axis.text.x = element_text(angle = 90))

training_set %>%
  mutate(CLM_FREQ = as.factor(CLM_FREQ),
         HOMEKIDS = as.factor(HOMEKIDS),
         KIDSDRIV = as.factor(KIDSDRIV),
         MVR PTS = as.factor(MVR PTS)) %>%
  select_if(negate(is.factor)) %>%
  gather(variable, value) %>%
  ggplot(., aes(x = value)) +
  geom_histogram(bins = 30, color = 'blue') +
  labs(title = 'Distributions of Continuous Predictor Variables') +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_density(aes(x = value), color = 'red') +
  facet_wrap(. ~variable, scales = 'free', ncol = 3)

```

```
# =====
# Data Correlation Table and Matrix Plot
# =====

training_set_claims = training_set%>%filter(TARGET_FLAG==1)
training_set_numeric<-select_if(training_set_claims,is.numeric)
correlation_table <- cor(training_set_numeric, method = 'pearson', use = 'complete.obs')[,1]
correlation_table %>%
  as.data.frame() %>%
  arrange(desc(abs(.))) %>%
kable(caption = 'Correlation of numeric predictors with the Target Amount') %>%
kable_styling(bootstrap_options = c("striped", "hover"))
```

```

# =====
# Scatter Plots of Each Variable Versus the Target Variable
# =====

# Scatter plots for each of the variables against the target.
col_size = dim(training_set)[2]
cols = training_set %>%
  mutate(CLM_FREQ = as.factor(CLM_FREQ),
         HOMEKIDS = as.factor(HOMEKIDS),
         KIDSDRIV = as.factor(KIDSDRIV),
         MVR PTS = as.factor(MVR PTS)) %>%
  select_if(negate(is.factor)) %>%
  names(.)
for (col in cols[2:10]) {
  plot = training_set %>%
    ggplot(aes_string(x = col, y = 'TARGET_AMT')) +
    geom_point(stat = 'identity') +
    labs(title = paste(col,'vs.', 'TARGET_AMT'))
  print(plot)
}

# =====
# Model One
# =====

- `EDUCATION` was grouped as "College" and "No College"
- `AGE` was grouped as "Under 25" and "25 and Over"
- `JOB` was grouped as "Careered" and "Not Careered"
- `MINIVAN` grouped the `CAR_TYPE` as "yes" or "no" if it was a minivan
- `RED_CAR`, `CAR_AGE`, `SEX`, `YOJ` , `HOMEKIDS` were removed due to high p value

options(scipen = 999)
model1 <- training_set %>%
  mutate(EDUCATION = ifelse(EDUCATION=="Bachelors" | EDUCATION=="Masters" | EDUCATION=="PhD",
                           "College", "No College"),
         AGE = ifelse(AGE < 25, "Under 25", "Over 25"),
         JOB = ifelse(JOB=="Student" | JOB == "Home Maker" | JOB == "Clerical", "Not Careered", "Careered"),
         MINIVAN = as.factor(ifelse(CAR_TYPE == "Minivan", "yes", "no")),
         KIDSDRIV = as.factor(ifelse(KIDSDRIV == 0, "no", "yes")),
         ) %>%

  glm(TARGET_FLAG ~ KIDSDRIV + AGE + INCOME + PARENT1 + HOME_VAL + MSTATUS +
       EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + OLDCALL +
       CLM_FREQ + REVOKED + MVR PTS + URBANICITY + MINIVAN, family = "binomial", data = .)

summary(model1)

plot(fitted(model1), resid(model1), col = "grey", pch = 20,
      xlab = "Fitted", ylab = "Residuals", main = "Fitted versus Residuals")
abline(h = 0, col = "darkorange", lwd = 2)
qqnorm(resid(model1), main = "Normal Q-Q Plot", col = "darkgrey")
qqline(resid(model1), col = "dodgerblue", lwd = 2)

```

```

# =====
# Model Two
# =====

options(scipen = 999)
model2 = glm(TARGET_FLAG ~ INCOME +
             PARENT1 + HOME_VAL +
             MSTATUS + EDUCATION +
             TRAVTIME + CAR_USE +
             TIF + CAR_TYPE +
             REVOKED + URBANICITY,
             data = training_set, family = binomial(link = 'logit'))
summary(model2, digits = 2)

plot(fitted(model1), resid(model1), col = "grey", pch = 20,
      xlab = "Fitted", ylab = "Residuals", main = "Fitted versus Residuals")
abline(h = 0, col = "darkorange", lwd = 2)
qqnorm(resid(model1), main = "Normal Q-Q Plot", col = "darkgrey")
qqline(resid(model1), col = "dodgerblue", lwd = 2) col = "dodgerblue", lwd = 2)

```

```

# =====
# Model Three
# =====

model3 = glm(TARGET_FLAG~.-TARGET_AMT-INCOME,training_set,family ='binomial')

summary(model3)

plot(fitted(model3), resid(model3), col = "grey", pch = 20,
      xlab = "Fitted", ylab = "Residuals", main = "Fitted versus Residuals")
abline(h = 0, col = "darkorange", lwd = 2)
qqnorm(resid(model3), main = "Normal Q-Q Plot", col = "darkgrey")
qqline(resid(model3), col = "dodgerblue", lwd = 2)

# =====
# Model Four
# =====

Here we run a linear regression (for the records that represent an actual accident) i.e. TARGET_FLAG=1,
```

model4 = lm(TARGET\_AMT ~ BLUEBOOK, data = training\_set\_numeric)

summary(model4)

plot(fitted(model4), resid(model4), col = "grey", pch = 20,
 xlab = "Fitted", ylab = "Residuals", main = "Fitted versus Residuals")
abline(h = 0, col = "darkorange", lwd = 2)
qqnorm(resid(model4), main = "Normal Q-Q Plot", col = "darkgrey")
qqline(resid(model4), col = "dodgerblue", lwd = 2)

```

# =====
# Model Five
# =====

data <- training_set %>% filter(TARGET_FLAG == 1)
data <- data[-1]
b <- boxcox(TARGET_AMT ~ . , data = data)

lamda <- b$x
lik <- b$y
bc <- cbind(lamda, lik)
l <- bc[order(-lik),] [1,1]

model5 <- lm(TARGET_AMT^(1) ~ . , data = data)

summary(model5)

plot(fitted(model5), resid(model5), col = "grey", pch = 20,
      xlab = "Fitted", ylab = "Residuals", main = "Fitted versus Residuals")
abline(h = 0, col = "darkorange", lwd = 2)
qqnorm(resid(model5), main = "Normal Q-Q Plot", col = "darkgrey")
qqline(resid(model5), col = "dodgerblue", lwd = 2)

```

```

# =====
# Model Selection
# =====

# Function that creates a vector of binary values based on threshold.
to_binary = function(arr,thresh) {
  binary = c()
  for (i in arr) {
    if (i >= thresh) {
      binary = c(binary, 1)
    }
    else {
      binary = c(binary, 0)
    }
  }
  return(binary)
}

# Predictions based on a threshold of 0.5.
predictions = training_set[c('target')]
predictions$model1 = to_binary(predict(model1,type ='response'),0.5)
predictions$model2 = to_binary(predict(model2,type ='response'),0.5)
predictions$model3 = to_binary(predict(model3,type ='response'),0.5)
head(predictions)

```

```
# =====
# Error Calculations
# =====

predictions = predictions %>%
  mutate(target = as.factor(target),
         model1 = as.factor(model1),
         model2 = as.factor(model2),
         model3 = as.factor(model3)
     )

# Model 1.
confusionMatrix(predictions$model1,predictions$target)

# Model 2.
confusionMatrix(predictions$model2,predictions$target)

# Model 3.
confusionMatrix(predictions$model3,predictions$target)
```

```

# =====
# Model Comparison
# =====

accuracy <- function(df,col1,col2) {
  true = df[,col1]
  predict = df[,col2]
  # total events
  len = length(true)
  # total correct predictions
  correct = 0
  for (i in seq(len)){
    if (true[i] == predict[i]){
      correct = correct + 1
    }
  }
  # accuracy
  return (correct/len)
}

class_error_rate <- function(df,col1,col2) {
  true = df[,col1]
  predict = df[,col2]
  # total events
  len = length(true)
  # total errors
  error = 0
  for (i in seq(len)){
    if (true[i] != predict[i]){
      error = error + 1
    }
  }
  # error rate
  return (error/len)
}

precision <- function(col1, col2) {
  # Calculate the total number of true positives in the dataset.
  true_positive <- sum(col1 == 1 & col2 == 1)
  # Calculate the total number of false positives in the dataset.
  false_positive <- sum(col1 == 0 & col2 == 1)
  # Perform the precision calculation and round the result to 2 decimal places.
  prediction_precision <- true_positive / (true_positive + false_positive)
  return(prediction_precision)
}

sensitivity <- function(col1, col2) {

  true_positive <- sum(col1 == 1 & col2 == 1)
  false_negative <- sum(col1 == 1 & col2 == 0)

  sensitivity<- true_positive / (true_positive + false_negative)

  return(sensitivity)
}

specificity <- function(col1, col2) {

```

```

true_negative <- sum(col2 == 0 & col1 == 0)
false_positive <- sum(col2 == 1 & col1 == 0)

specificity <- true_negative / (true_negative + false_positive)

return(specificity)
}

f1_score <- function(col1, col2) {
  sens <- sensitivity(col1, col2)
  prec <- precision(col1, col2)
  f1 <- 2 * sens * prec / (prec+sens)
  return(f1)
}

roc_model1 <- roc(predictions$TARGET_FLAG, as.numeric(predictions$model1))
roc_model2 <- roc(predictions$TARGET_FLAG, as.numeric(predictions$model2))
roc_model3 <- roc(predictions$TARGET_FLAG, as.numeric(predictions$model3))

#accuracy
acc <- c(accuracy(predictions, 'TARGET_FLAG', 'model1'), accuracy(predictions, 'TARGET_FLAG', 'model2'),
        accuracy(predictions, 'TARGET_FLAG', 'model3'))

#classification error rate
class_error <- c(class_error_rate(predictions, 'TARGET_FLAG', 'model1'),
                  class_error_rate(predictions, 'TARGET_FLAG', 'model2'),
                  class_error_rate(predictions, 'TARGET_FLAG', 'model3'))

#precision
prec <- c(precision(predictions$TARGET_FLAG, predictions$model1),
           precision(predictions$TARGET_FLAG, predictions$model2),
           precision(predictions$TARGET_FLAG, predictions$model3))

#specificity
spec <- c(specificity(predictions$TARGET_FLAG, predictions$model1),
           specificity(predictions$TARGET_FLAG, predictions$model2),
           specificity(predictions$TARGET_FLAG, predictions$model3))

#sensitivity
sens <- c(sensitivity(predictions$TARGET_FLAG, predictions$model1),
           sensitivity(predictions$TARGET_FLAG, predictions$model2),
           sensitivity(predictions$TARGET_FLAG, predictions$model3))

#f1 score
f1 <- c(f1_score(predictions$TARGET_FLAG, predictions$model1),
         f1_score(predictions$TARGET_FLAG, predictions$model2),
         f1_score(predictions$TARGET_FLAG, predictions$model3))

#AUC
a_u_c <- c(auc(roc_model1), auc(roc_model2), auc(roc_model3))

model_comparison <- rbind(acc, class_error, prec, spec, sens, f1, a_u_c) %>%
  as.data.frame() %>%
  magrittr::set_rownames(c('accuracy', 'classification error rate', 'precision', 'sensitivity',
                           'specificity', 'F1 score', 'AUC')) %>%
  magrittr::set_colnames(c('Model 1', 'Model 2', 'Model 3')) %>%
  round(., 4)

model_comparison

```

```

# =====
# Evaluating Dataset
# =====

- Picked Models 2 and 4 because they had the best results

evaluation_set = evaluation_set%>%
  mutate(AGE = ifelse(is.na(AGE), median(training_set$AGE, na.rm =TRUE), AGE),
         YOJ = ifelse(is.na(YOJ), median(training_set$YOJ, na.rm =TRUE), YOJ),
         INCOME = ifelse(is.na(INCOME), median(training_set$INCOME, na.rm =TRUE), INCOME),
         HOME_VAL = ifelse(is.na(HOME_VAL), median(training_set$HOME_VAL, na.rm =TRUE), HOME_VAL),
         CAR_AGE = ifelse(is.na(CAR_AGE), median(training_set$CAR_AGE, na.rm =TRUE), CAR_AGE))
evaluation_set$TARGET_FLAG = to_binary(predict(model2, evaluation_set, type = "response"), 0.5)
evaluation_set$TARGET_AMT = predict(model4, evaluation_set)
head(evaluation_set)

```