# Data 621 - Homework 4

Group 4 Layla Quinones, Ian Costello, Dmitriy Burtsev & Esteban Aramayo

11/21/2021

## 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:

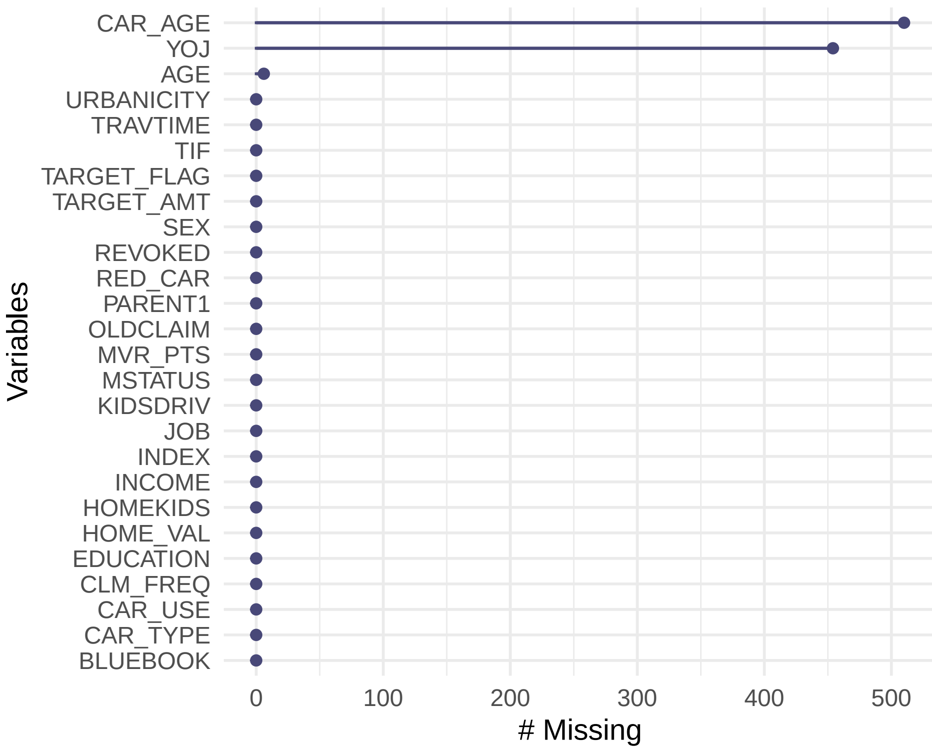
## Exploratory Data Analysis

Below is a glimpse of the Insurance Training data.

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <chr> "$67,349", "$91,449", "$16,039", "", "$114,986", "$125,301~
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", "No~
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,925", "$0"~
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes", "Yes",~
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z_F", "M"~
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High School", "~
## $ JOB         <chr> "Professional", "z_Blue Collar", "Clerical", "z_Blue Colla~
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Private", ~
## $ BLUEBOOK    <chr> "$14,230", "$14,940", "$4,010", "$15,440", "$18,000", "$17~
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV", "Sports~
## $ RED_CAR     <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes", "no", ~
## $ OLDCLAIM    <chr> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", "$0", "$~
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "N~
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Highly Urba~
```

There are 8161 observations in this data set and 26 columns. We know that `INDEX`, `TARGET_FLAG` and `TARGET_AMT` are not predictor variables. This gives us **8161 observations** with **23 predictors** that are a combination of int, double and character data types. We also see that the character variables will have to converted to factors in order for us to explore their distributions. Variables such and `INCOME`, `HOME_VAL`, `BLUEBOOK`, `OLDCLAIM` will be converted to numeric because they are numbers with values that have meaning in their hierarchy.

### Missing Values




There are missing variables in the columns `Car_AGE`, `AGE` and `YOJ`. None of these exceed the 10% missing data so we will continue with all variables for noe (not dropping any of them due to missing data)

### DATA CLEANING - CONVERTING DATA TYPES

- Let's remove the `$`, `z_`, and `,` and put in a different variable name from numeric strings.
- Let's also change all other character variables into factors.

Let's glimpse the data to confirm the data cleaning.

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <fct> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <dbl> 67349, 91449, 16039, NA, 114986, 125301, 18755, 107961, 62~
## $ PARENT1     <fct> No, No, No, No, No, Yes, No, No, No, No, No, No, No, No, N~
## $ HOME_VAL    <dbl> 0, 257252, 124191, 306251, 243925, 0, NA, 333680, 0, 0, 0,~
## $ MSTATUS     <fct> No, No, Yes, Yes, Yes, No, Yes, Yes, No, No, No, Yes, Yes,~
## $ SEX         <fct> M, M, F, M, F, F, F, M, F, M, F, F, M, M, F, F, M, F, F, F~
```

```
## $ EDUCATION   <fct> PhD, High School, High School, <High School, PhD, Bachelor~
## $ JOB         <fct> Professional, Blue Collar, Clerical, Blue Collar, Doctor, ~
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <fct> Private, Commercial, Private, Private, Private, Commercial~
## $ BLUEBOOK    <dbl> 14230, 14940, 4010, 15440, 18000, 17430, 8780, 16970, 1120~
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <fct> Minivan, Minivan, SUV, Minivan, SUV, Sports Car, SUV, Van,~
## $ RED_CAR     <fct> yes, yes, no, yes, no, no, no, yes, no, no, no, no, yes, y~
## $ OLDCLAIM    <dbl> 4461, 0, 38690, 0, 19217, 0, 0, 2374, 0, 0, 0, 0, 5028, 0,~
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <fct> No, No, No, No, Yes, No, No, Yes, No, No, No, No, Yes, No,~
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly Urban/ Ur~
```

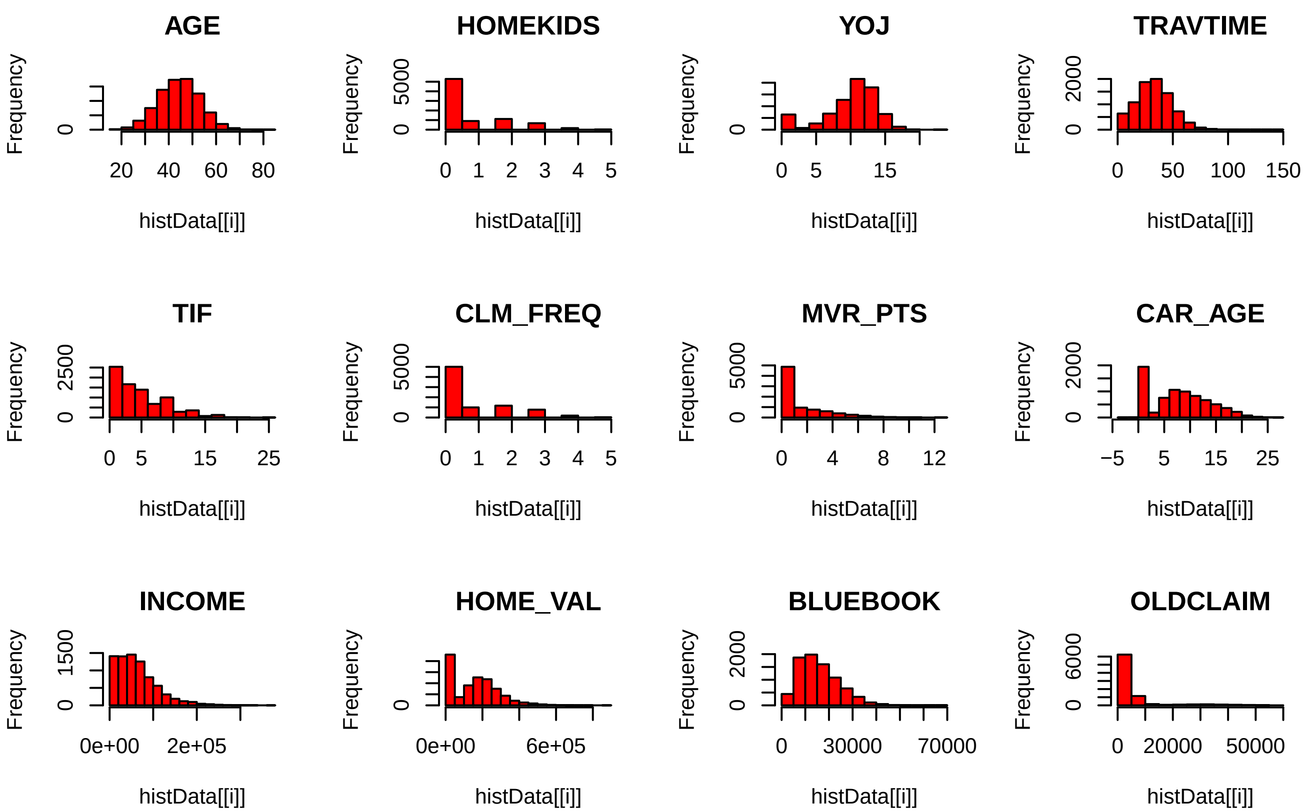
Display summary statistics again to confirm data cleaning.

```
##      INDEX        TARGET_FLAG   TARGET_AMT        KIDSDRIV           AGE
##  Min.   :    1   0:6008      Min.   :     0   Min.   :0.0000   Min.   :16.00
##  1st Qu.: 2559   1:2153      1st Qu.:     0   1st Qu.:0.0000   1st Qu.:39.00
##  Median : 5133               Median :     0   Median :0.0000   Median :45.00
##  Mean   : 5152               Mean   :  1504   Mean   :0.1711   Mean   :44.79
##  3rd Qu.: 7745               3rd Qu.:  1036   3rd Qu.:0.0000   3rd Qu.:51.00
##  Max.   :10302               Max.   :107586   Max.   :4.0000   Max.   :81.00
##                                                                NA's   :6
##     HOMEKIDS           YOJ           INCOME         PARENT1       HOME_VAL
##  Min.   :0.0000   Min.   : 0.0   Min.   :     0   No :7084   Min.   :     0
##  1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097   Yes:1077   1st Qu.:     0
##  Median :0.0000   Median :11.0   Median : 54028              Median :161160
##  Mean   :0.7212   Mean   :10.5   Mean   : 61898              Mean   :154867
##  3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986              3rd Qu.:238724
##  Max.   :5.0000   Max.   :23.0   Max.   :367030              Max.   :885282
##                   NA's   :454    NA's   :445                 NA's   :464
##  MSTATUS    SEX              EDUCATION             JOB           TRAVTIME
##  No :3267   F:4375   <High School:1203   Blue Collar :1825   Min.   :  5.00
##  Yes:4894   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   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
```

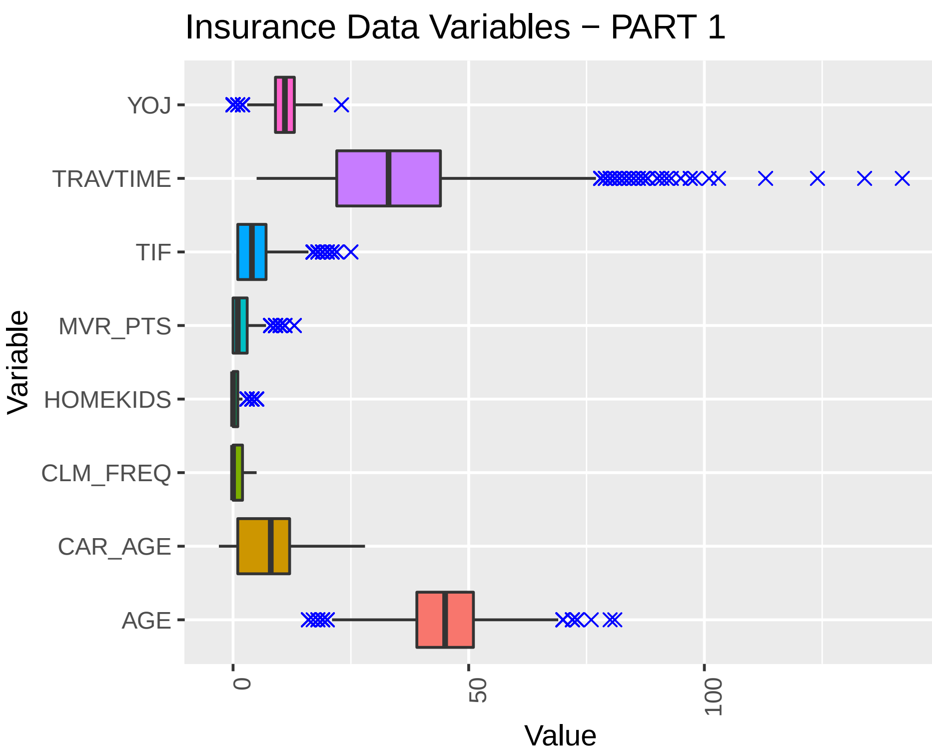
We get a better sense of the information available in each variable now with the data type changes.

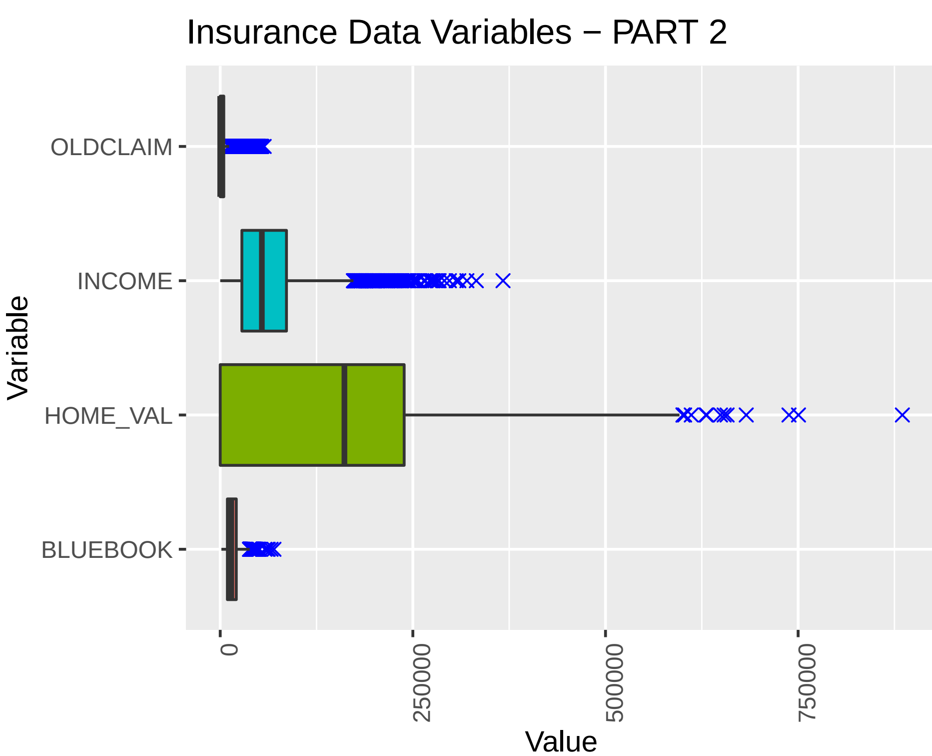
Let's plot the distribution of the numerical variables using histograms.




From the above histograms of numerical data we can see that most numerical variables have a right skew, which may indicate that a transformation will be helpful for these variables.

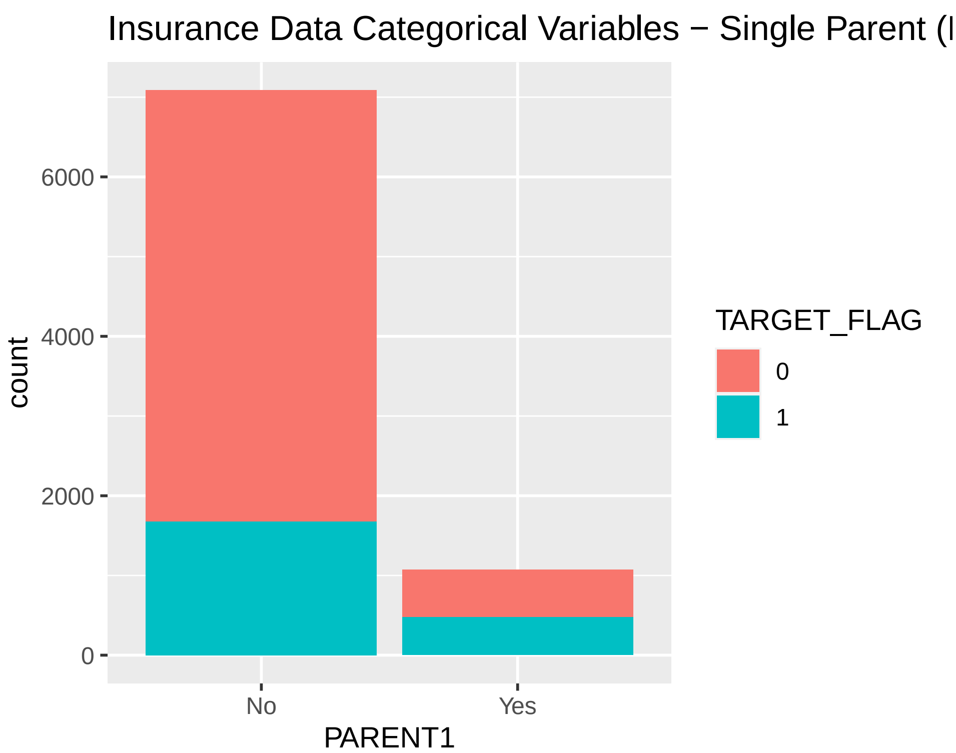
Let's identify the variables with outlier values using boxplots.

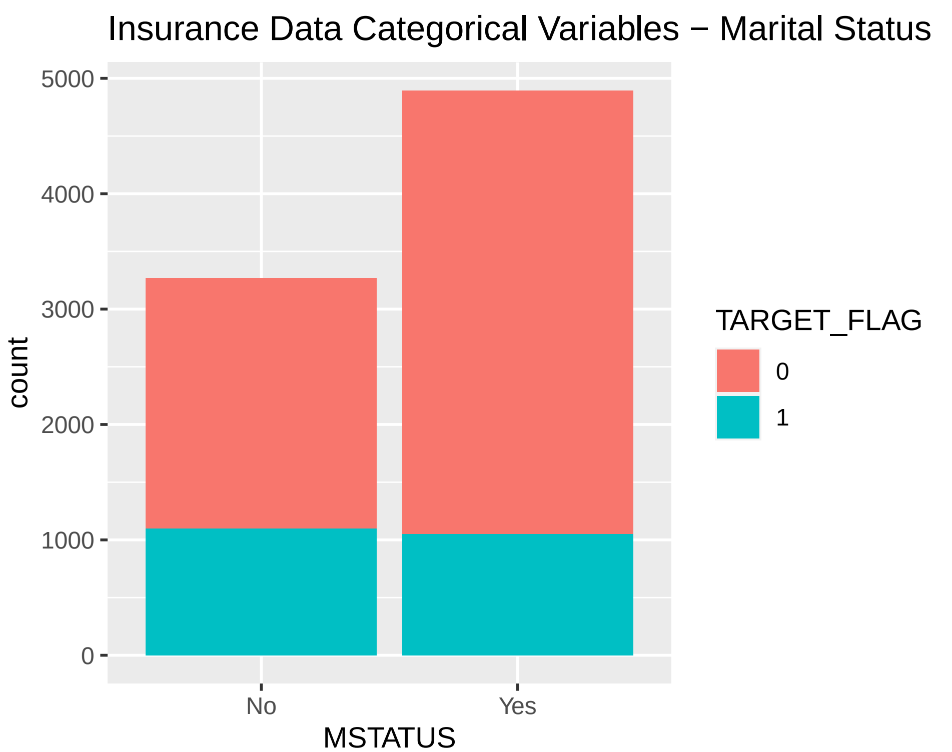


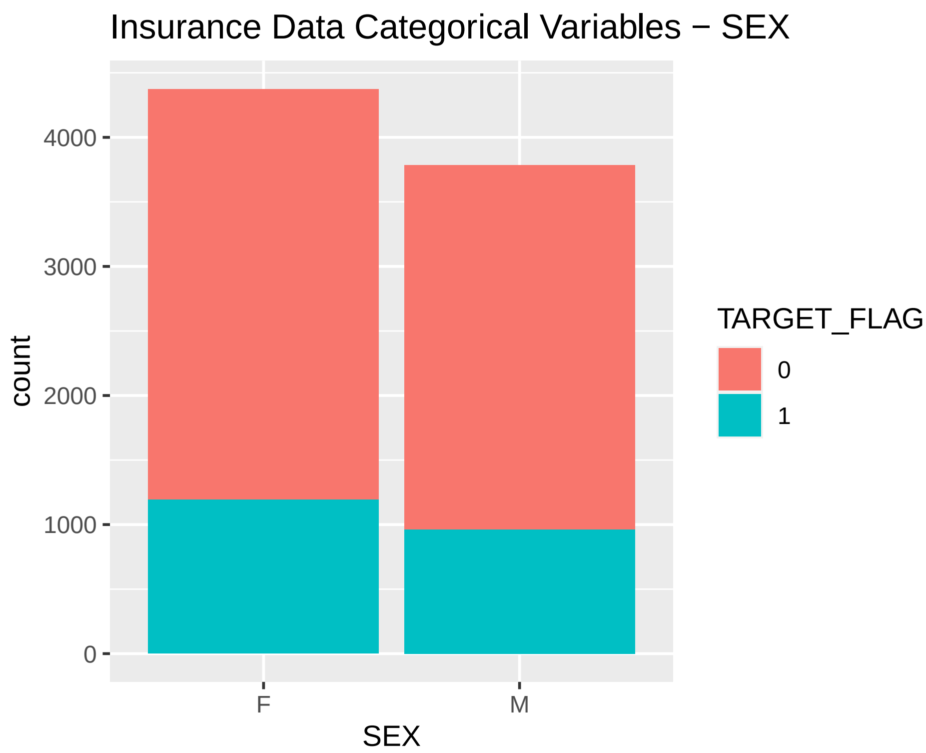



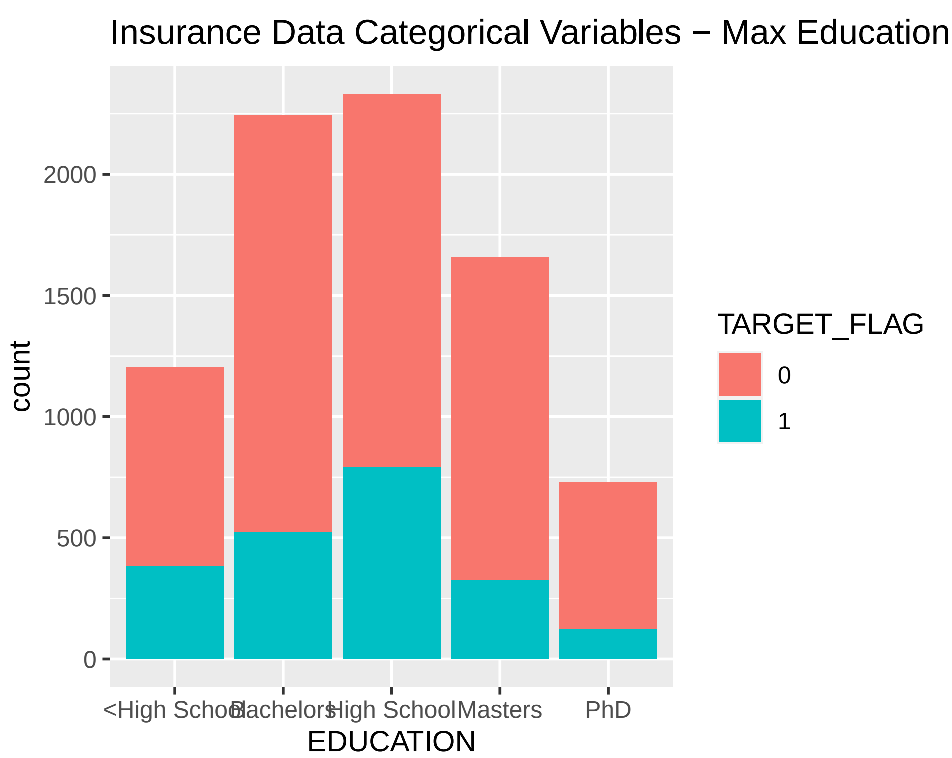

From these initial box plots we can see that there are some outliers. In particular, `TRAVTIME`, `INCOME`, and `HOME_VAL` have many outliers which are spread out more compared to the other variables.

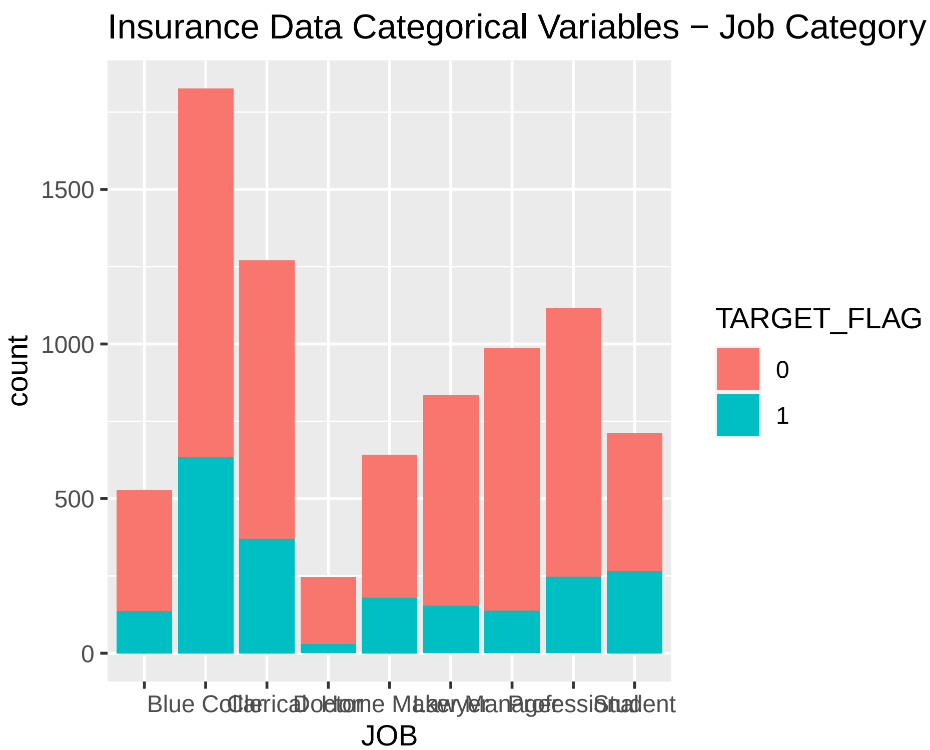
### Categorical Predictors - with target variable

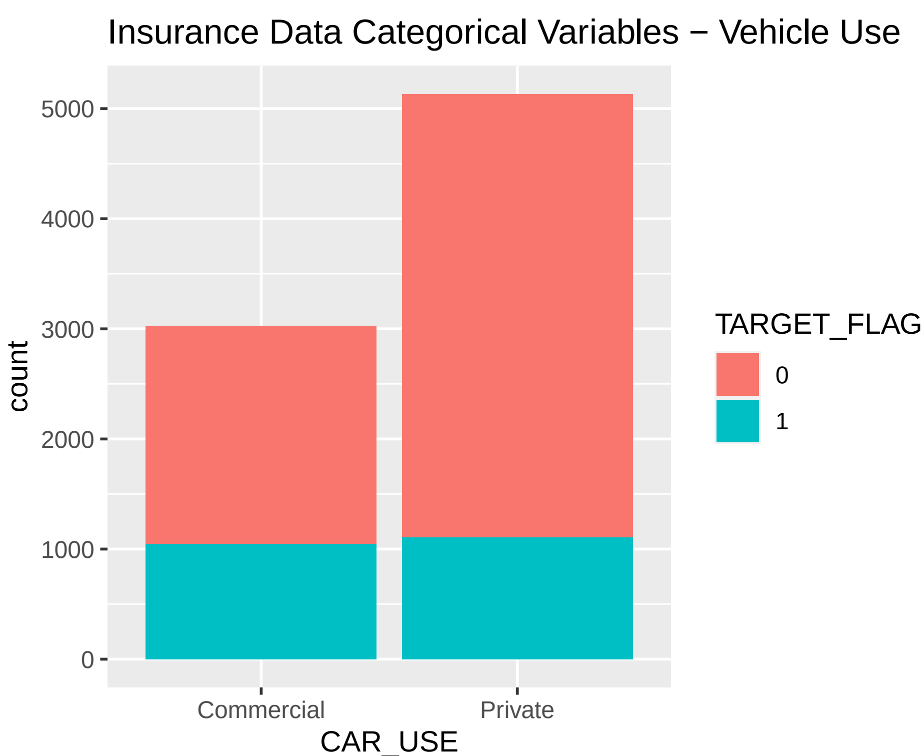


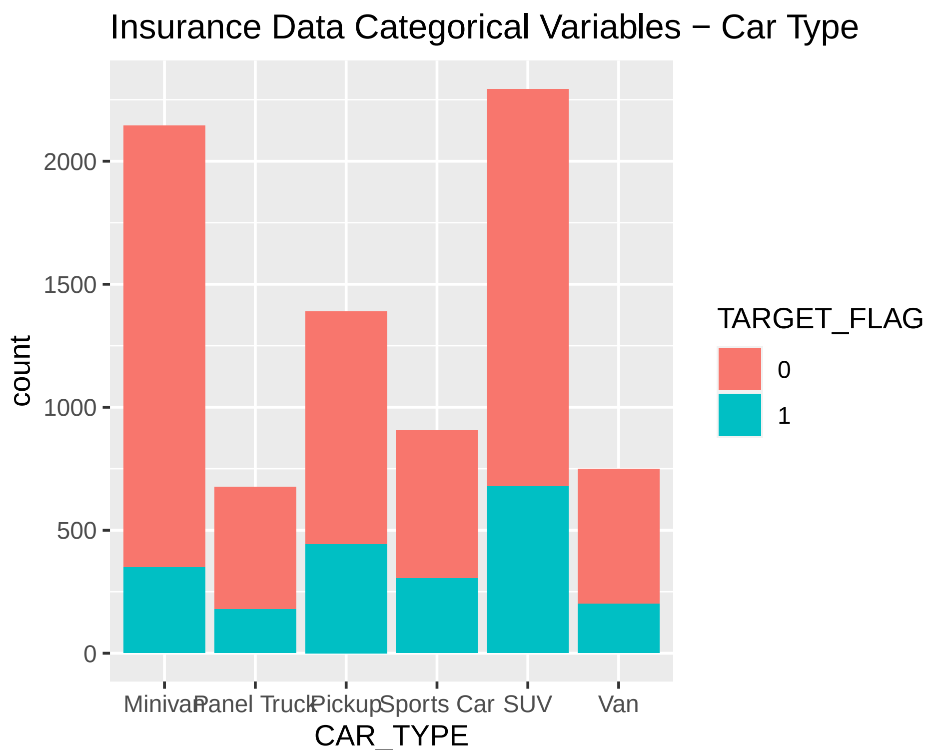



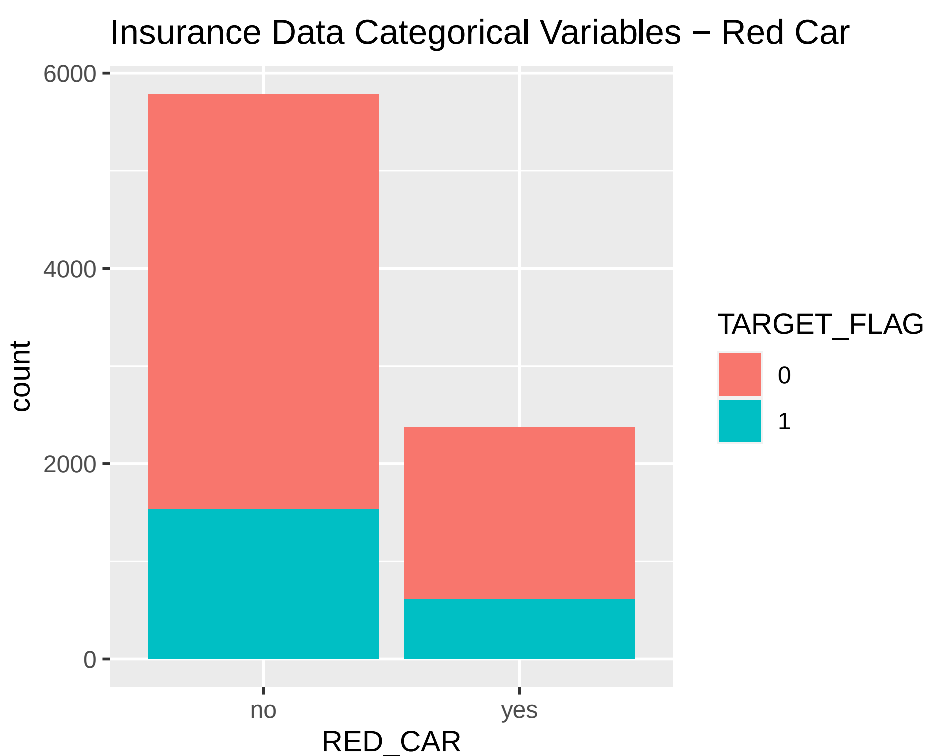

Insurance Data Categorical Variables − SEX
count
4000
3000
2000
1000
0
F
M
SEX
TARGET_FLAG
0
1

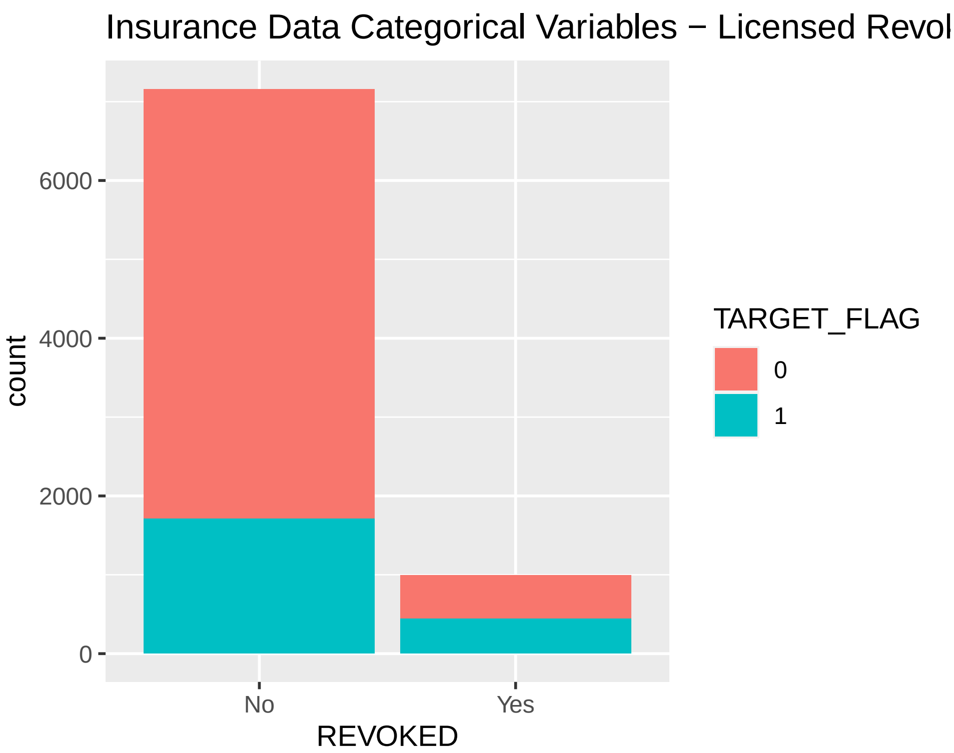

Insurance Data Categorical Variables − Max Education
count
2000
1500
1000
500
0
<High School
Bachelors
High School
Masters
PhD
EDUCATION
TARGET_FLAG
0
1


Insurance Data Categorical Variables − Job Category
count
1500
1000
500
0
Blue Collar
Clerical
Doctor
Home Maker
Lawyer
Manager
Professional
Student
JOB
TARGET_FLAG
0
1

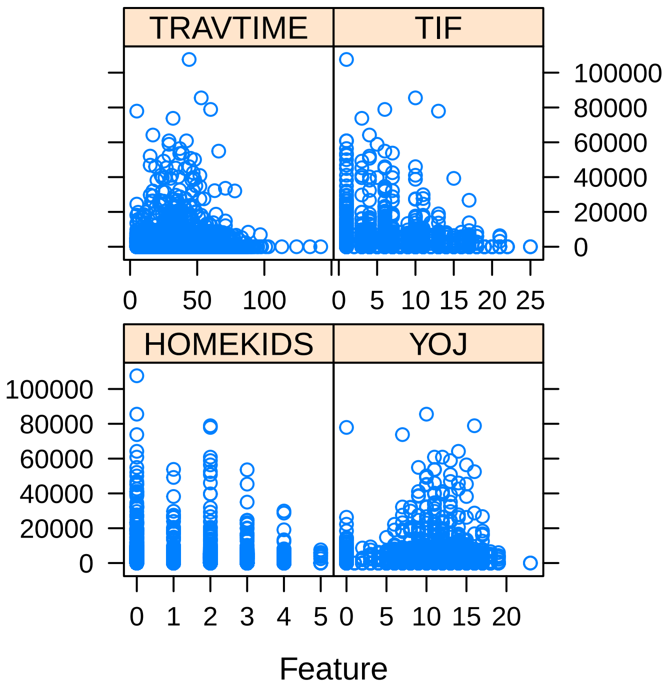

Insurance Data Categorical Variables − Vehicle Use
count
5000
4000
3000
2000
1000
0
Commercial
Private
CAR_USE
TARGET_FLAG
0
1

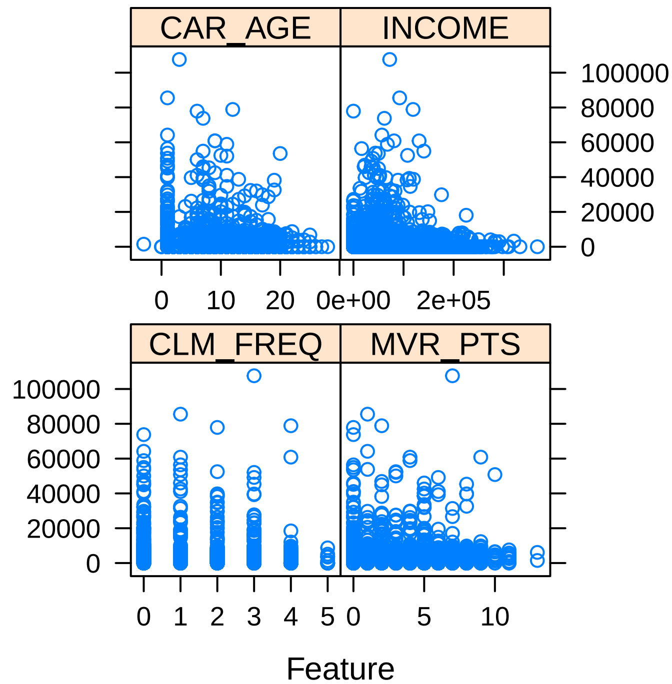

Insurance Data Categorical Variables − Car Type
count
2000
1500
1000
500
0
Minivan
Panel Truck
Pickup
Sports Car
SUV
Van
CAR_TYPE
TARGET_FLAG
0
1

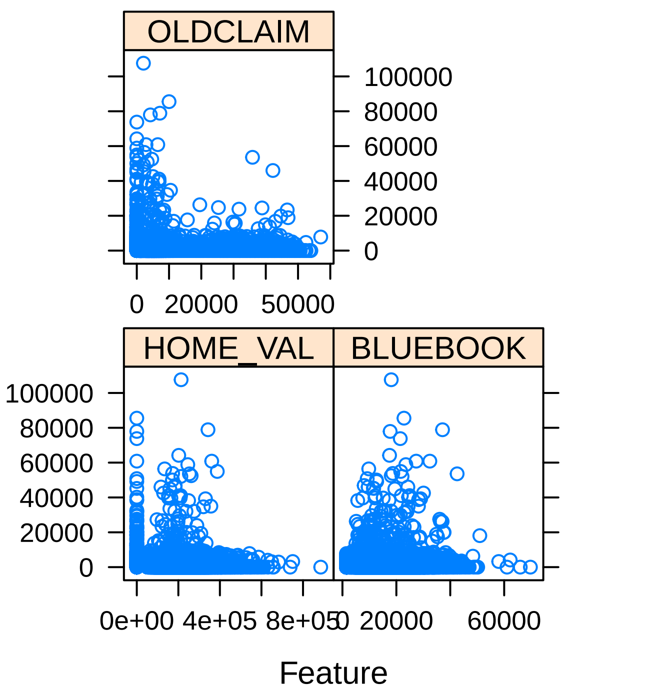

Insurance Data Categorical Variables − Red Car
count
6000
4000
2000
0
no
yes
RED_CAR
TARGET_FLAG
0
1


Insurance Data Categorical Variables − Licensed Revol
count
6000
4000
2000
0
No
Yes
REVOKED
TARGET_FLAG
0
1

### Numeric Data - Relationship to Target

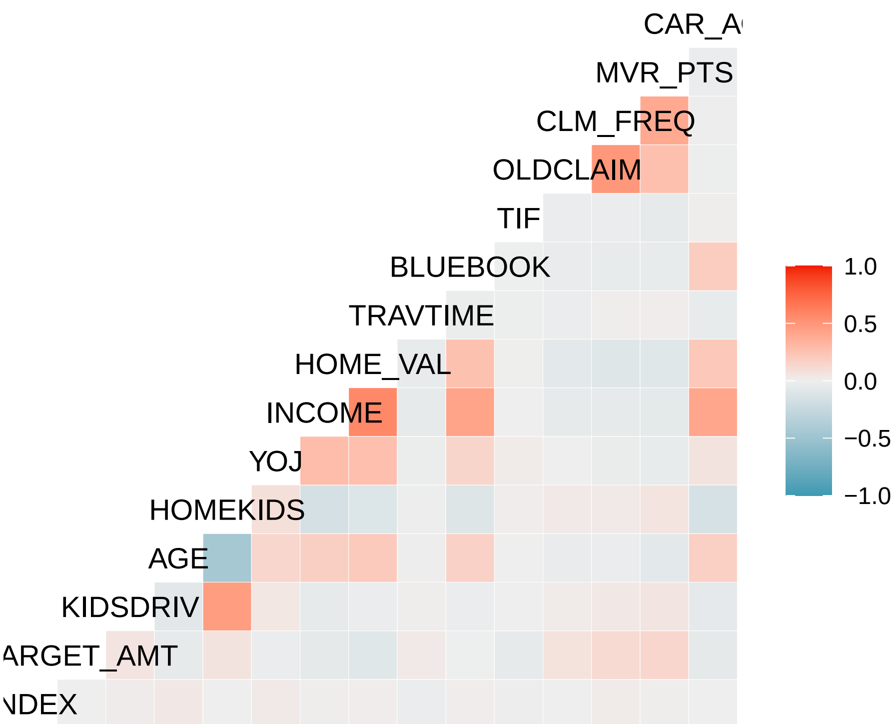



### Correlation

Let's use a heat map to see the level of correlation of the numeric predictor variables.




Let's check if there are any highly correlated variables (correlation higher than 0.75) and drop them if necessary.

```
## All correlations <= 0.75
```

```
## character(0)
```

# Data Preparation

## Data Cleaning

- Missing values are handled by imputing them as follows:
  - Use the mean to impute missing values for `Age` and `YOJ`.
  - Use the `median` to impute missing values for `HOME_VAL`, `INCOME`, and `CAR_AGE`.
- Outlier values non-factor variables are being normalized.

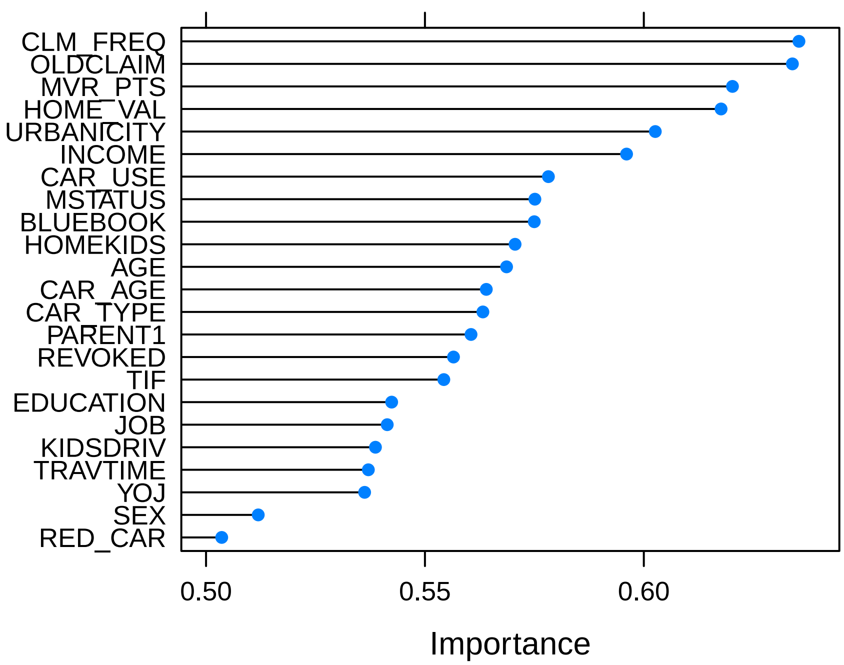
## Variable Importance

To determine the variable importance the following steps were taken:

- A training data frame `prepTrainA` was prepared for the `TARGET_FLAG` response variable and its associated predictor variables.
- A training data frame `prepTrainB` was prepared for the `TARGET_AMT` response variable and its associated predictor variables.
- Using the `prepTrainA` data frame, a classification model `modelA` was trained using the `Learning Vector Quantization (lvq)` method. From it, the variable importance was summarized and plotted.

```
## ROC curve variable importance
##
##   only 20 most important variables shown (out of 23)
##
```

```
##           Importance
## CLM_FREQ       0.6354
## OLDCLAIM       0.6339
## MVR_PTS        0.6202
## HOME_VAL       0.6176
## URBANICITY     0.6026
## INCOME         0.5961
## CAR_USE        0.5782
## MSTATUS        0.5751
## BLUEBOOK       0.5750
## HOMEKIDS       0.5706
## AGE            0.5686
## CAR_AGE        0.5640
## CAR_TYPE       0.5632
## PARENT1        0.5605
## REVOKED        0.5565
## TIF            0.5543
## EDUCATION      0.5424
## JOB            0.5414
## KIDSDRIV       0.5387
## TRAVTIME       0.5371
```

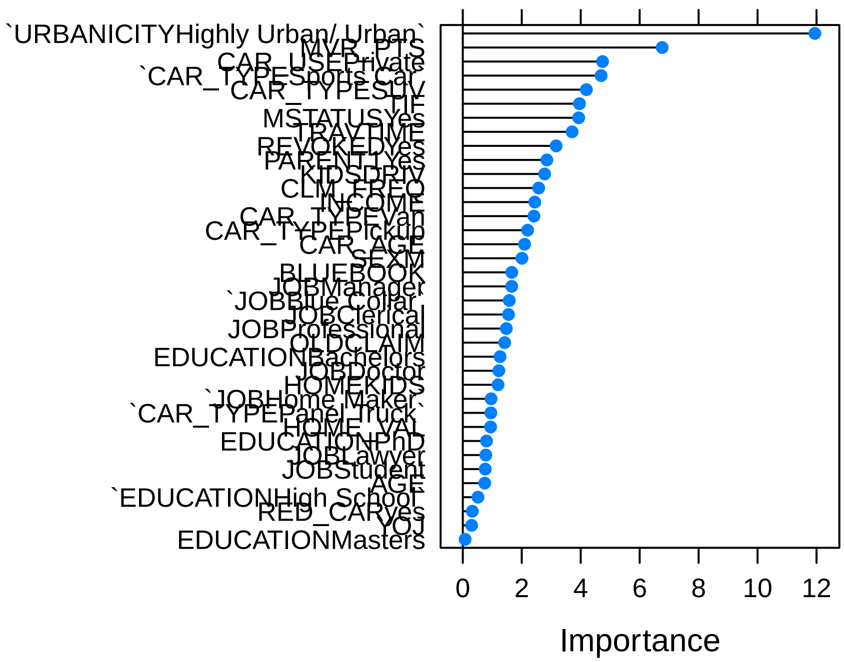



According to the plots above, we can predict which variables would contribute best to the categorical predictions for `TARGET_FLAG`. We can use this to inform our data transformations.

- Using the `prepTrainB` data frame, a classification/regression model `modelB` was trained using the `Generalized Linear Model (glm)` method. From it, the variable importance was summarized and plotted.

```
## glm variable importance
##
##   only 20 most important variables shown (out of 37)
##
##                                Overall
## `URBANICITYHighly Urban/ Urban`  11.944
## MVR_PTS                           6.764
```

```
## CAR_USEPrivate                    4.741
## `CAR_TYPESports Car`              4.692
## CAR_TYPESUV                       4.193
## TIF                               3.958
## MSTATUSYes                        3.932
## TRAVTIME                          3.708
## REVOKEDYes                        3.166
## PARENT1Yes                        2.852
## KIDSDRIV                          2.776
## CLM_FREQ                          2.574
## INCOME                            2.441
## CAR_TYPEVan                       2.413
## CAR_TYPEPickup                    2.200
## CAR_AGE                           2.096
## SEXM                              2.007
## BLUEBOOK                          1.663
## JOBManager                        1.660
## `JOBBlue Collar`                  1.578
```




According to the plots above, we can predict which variables would contribute best to the numerical predictions for `TARGET_AMT`. We can use this to inform our data transformations.

### Train Test Split

We partition the training data in two data sets. One to be used for training purposes and one for validation/testing purposes.

## Models

Using the training data set, build at least two different multiple linear regression models and three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? For example, if a person has a lot of traffic tickets, you would reasonably expect that person to have more car crashes. If the coefficient is negative (suggesting that the person is a safer driver), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

## Binary Logistic Regression

### Binary Logistic Regression Model 1

We begin with a `baseline` model that includes all the predictor variables and the response variable `TARGET_FLAG`.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6207  -0.7138  -0.3982   0.6320   3.1760
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.794e+00  3.811e-01  -7.331 2.29e-13 ***
## KIDSDRIV                       3.954e-01  6.933e-02   5.703 1.18e-08 ***
## AGE                           -3.360e-03  4.509e-03  -0.745 0.456212
## HOMEKIDS                       2.628e-02  4.177e-02   0.629 0.529287
## YOJ                           -1.639e-02  9.646e-03  -1.699 0.089301 .
## INCOME                        -2.356e-06  1.194e-06  -1.972 0.048596 *
## PARENT1Yes                     4.746e-01  1.226e-01   3.871 0.000108 ***
## HOME_VAL                      -1.381e-06  3.795e-07  -3.640 0.000273 ***
## MSTATUSYes                    -4.922e-01  9.386e-02  -5.244 1.57e-07 ***
## SEXM                           6.883e-02  1.256e-01   0.548 0.583642
## EDUCATIONBachelors            -4.420e-01  1.295e-01  -3.413 0.000643 ***
## EDUCATIONHigh School          -5.567e-02  1.070e-01  -0.520 0.602836
## EDUCATIONMasters              -3.802e-01  2.010e-01  -1.891 0.058579 .
## EDUCATIONPhD                  -2.484e-01  2.370e-01  -1.048 0.294649
## JOBBlue Collar                 3.697e-01  2.081e-01   1.777 0.075644 .
## JOBClerical                    4.590e-01  2.202e-01   2.085 0.037058 *
## JOBDoctor                     -2.672e-01  2.901e-01  -0.921 0.357022
## JOBHome Maker                  3.097e-01  2.358e-01   1.314 0.188979
## JOBLawyer                      1.798e-01  1.916e-01   0.938 0.348195
## JOBManager                    -4.673e-01  1.928e-01  -2.424 0.015348 *
## JOBProfessional                2.623e-01  2.002e-01   1.310 0.190294
## JOBStudent                     2.746e-01  2.409e-01   1.140 0.254280
## TRAVTIME                       1.493e-02  2.105e-03   7.091 1.33e-12 ***
## CAR_USEPrivate                -7.869e-01  1.025e-01  -7.680 1.59e-14 ***
## BLUEBOOK                      -2.070e-05  5.921e-06  -3.496 0.000473 ***
## TIF                           -5.618e-02  8.141e-03  -6.901 5.17e-12 ***
## CAR_TYPEPanel Truck            5.310e-01  1.829e-01   2.903 0.003694 **
## CAR_TYPEPickup                 5.420e-01  1.125e-01   4.818 1.45e-06 ***
## CAR_TYPESports Car             1.067e+00  1.446e-01   7.377 1.62e-13 ***
## CAR_TYPESUV                    7.894e-01  1.239e-01   6.369 1.91e-10 ***
## CAR_TYPEVan                    7.015e-01  1.403e-01   5.002 5.68e-07 ***
## RED_CARyes                    -1.634e-02  9.674e-02  -0.169 0.865834
## OLDCLAIM                      -1.115e-05  4.394e-06  -2.537 0.011172 *
## CLM_FREQ                       1.718e-01  3.196e-02   5.377 7.55e-08 ***
## REVOKEDYes                     7.916e-01  1.026e-01   7.715 1.21e-14 ***
## MVR_PTS                        1.124e-01  1.523e-02   7.381 1.57e-13 ***
## CAR_AGE                       -3.696e-03  8.409e-03  -0.440 0.660251
```

```
## URBANICITYHighly Urban/ Urban  2.449e+00  1.263e-01  19.388  < 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: 7533.1  on 6527  degrees of freedom
## Residual deviance: 5827.2  on 6490  degrees of freedom
## AIC: 5903.2
##
## Number of Fisher Scoring iterations: 5
```

### Binary Logistic Regression Model 2

For our second model, we only include the top 10 most important predictor variables that we gathered from our importance trained model `modelA`.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ CLM_FREQ + OLDCLAIM + MVR_PTS + HOME_VAL +
##     URBANICITY + INCOME + CAR_USE + MSTATUS + BLUEBOOK + HOMEKIDS,
##     family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1010  -0.7694  -0.4683   0.7749   2.9756
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.521e+00  1.395e-01 -10.906  < 2e-16 ***
## CLM_FREQ                      1.275e-01  2.987e-02   4.269 1.96e-05 ***
## OLDCLAIM                      4.720e-06  3.670e-06   1.286    0.198
## MVR_PTS                       1.222e-01  1.459e-02   8.375  < 2e-16 ***
## HOME_VAL                     -1.455e-06  3.542e-07  -4.108 3.99e-05 ***
## URBANICITYHighly Urban/ Urban  2.103e+00  1.197e-01  17.567  < 2e-16 ***
## INCOME                       -6.435e-06  9.351e-07  -6.881 5.94e-12 ***
## CAR_USEPrivate               -8.736e-01  6.572e-02 -13.293  < 2e-16 ***
## MSTATUSYes                   -5.942e-01  7.582e-02  -7.838 4.58e-15 ***
## BLUEBOOK                     -2.787e-05  4.318e-06  -6.453 1.09e-10 ***
## HOMEKIDS                      2.164e-01  2.745e-02   7.883 3.20e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7533.1  on 6527  degrees of freedom
## Residual deviance: 6243.0  on 6517  degrees of freedom
## AIC: 6265
##
## Number of Fisher Scoring iterations: 5
```

### Binary Logistic Regression Model 3

For our third model, we only include the predictor variables that have `theoretical effect on probability of collition`, which was provided as part of the definition of the variables.

```
##
## Call:
```

```
## glm(formula = TARGET_FLAG ~ AGE + CAR_USE + CLM_FREQ + EDUCATION +
##     HOME_VAL + INCOME + JOB + KIDSDRIV + MSTATUS + MVR_PTS +
##     RED_CAR + REVOKED + SEX + TIF + TRAVTIME + YOJ, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0442  -0.7570  -0.5291   0.8024   2.7606
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.908e-01  2.983e-01   0.975 0.329660
## AGE                  -1.121e-02  3.763e-03  -2.979 0.002888 **
## CAR_USEPrivate       -6.565e-01  8.440e-02  -7.778 7.35e-15 ***
## CLM_FREQ              2.605e-01  2.686e-02   9.699  < 2e-16 ***
## EDUCATIONBachelors   -3.575e-01  1.114e-01  -3.209 0.001334 **
## EDUCATIONHigh School -3.664e-02  9.636e-02  -0.380 0.703748
## EDUCATIONMasters     -2.865e-01  1.709e-01  -1.676 0.093723 .
## EDUCATIONPhD         -1.718e-01  2.099e-01  -0.818 0.413228
## HOME_VAL             -1.237e-06  3.608e-07  -3.429 0.000606 ***
## INCOME               -3.644e-06  1.127e-06  -3.234 0.001223 **
## JOBBlue Collar        1.313e-01  1.958e-01   0.670 0.502629
## JOBClerical           6.651e-02  2.111e-01   0.315 0.752723
## JOBDoctor            -2.807e-01  2.851e-01  -0.985 0.324745
## JOBHome Maker        -7.372e-02  2.246e-01  -0.328 0.742751
## JOBLawyer             6.305e-02  1.847e-01   0.341 0.732825
## JOBManager           -3.990e-01  1.890e-01  -2.110 0.034819 *
## JOBProfessional       8.855e-02  1.942e-01   0.456 0.648320
## JOBStudent           -1.120e-01  2.269e-01  -0.494 0.621485
## KIDSDRIV              3.454e-01  5.581e-02   6.188 6.08e-10 ***
## MSTATUSYes           -5.268e-01  7.498e-02  -7.026 2.13e-12 ***
## MVR_PTS               1.370e-01  1.438e-02   9.527  < 2e-16 ***
## RED_CARyes           -2.078e-02  9.144e-02  -0.227 0.820268
## REVOKEDYes            7.874e-01  8.558e-02   9.201  < 2e-16 ***
## SEXM                 -2.172e-01  8.694e-02  -2.499 0.012467 *
## TIF                  -5.078e-02  7.723e-03  -6.576 4.84e-11 ***
## TRAVTIME              6.685e-03  1.908e-03   3.504 0.000458 ***
## YOJ                  -1.066e-02  8.778e-03  -1.215 0.224445
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7533.1  on 6527  degrees of freedom
## Residual deviance: 6501.1  on 6501  degrees of freedom
## AIC: 6555.1
##
## Number of Fisher Scoring iterations: 4
```

**Binary Logistic Regression Model 4**

For our third model, we only include the predictor variables that have `theoretical effect on probability of collition`, which was provided as part of the definition of the variables. Additionally, we remove the variables that were deemed as "urban legends", such as `RED_CAR` and `SEX`.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CAR_USE + CLM_FREQ + EDUCATION +
##     HOME_VAL + INCOME + JOB + KIDSDRIV + MSTATUS + MVR_PTS +
```

```
##     REVOKED + TIF + TRAVTIME + YOJ, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0349  -0.7577  -0.5320   0.8043   2.7934
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           8.655e-02  2.921e-01   0.296 0.766990
## AGE                  -1.207e-02  3.747e-03  -3.220 0.001280 **
## CAR_USEPrivate       -5.827e-01  8.125e-02  -7.171 7.42e-13 ***
## CLM_FREQ              2.595e-01  2.682e-02   9.672  < 2e-16 ***
## EDUCATIONBachelors   -3.274e-01  1.109e-01  -2.952 0.003155 **
## EDUCATIONHigh School -6.902e-03  9.588e-02  -0.072 0.942611
## EDUCATIONMasters     -2.528e-01  1.707e-01  -1.481 0.138589
## EDUCATIONPhD         -1.280e-01  2.092e-01  -0.612 0.540746
## HOME_VAL             -1.264e-06  3.602e-07  -3.510 0.000448 ***
## INCOME               -3.542e-06  1.124e-06  -3.152 0.001624 **
## JOBBlue Collar        2.100e-01  1.943e-01   1.081 0.279809
## JOBClerical           1.108e-01  2.105e-01   0.526 0.598656
## JOBDoctor            -2.654e-01  2.850e-01  -0.931 0.351633
## JOBHome Maker         5.402e-02  2.214e-01   0.244 0.807204
## JOBLawyer             9.267e-02  1.840e-01   0.504 0.614523
## JOBManager           -3.701e-01  1.886e-01  -1.962 0.049763 *
## JOBProfessional       1.159e-01  1.939e-01   0.598 0.549789
## JOBStudent           -4.309e-02  2.258e-01  -0.191 0.848620
## KIDSDRIV              3.518e-01  5.565e-02   6.322 2.59e-10 ***
## MSTATUSYes           -5.218e-01  7.485e-02  -6.971 3.15e-12 ***
## MVR_PTS               1.381e-01  1.437e-02   9.611  < 2e-16 ***
## REVOKEDYes            7.900e-01  8.551e-02   9.239  < 2e-16 ***
## TIF                  -5.038e-02  7.705e-03  -6.539 6.18e-11 ***
## TRAVTIME              6.659e-03  1.905e-03   3.495 0.000475 ***
## YOJ                  -1.024e-02  8.766e-03  -1.169 0.242569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7533.1  on 6527  degrees of freedom
## Residual deviance: 6512.9  on 6503  degrees of freedom
## AIC: 6562.9
##
## Number of Fisher Scoring iterations: 4
```

#### Binary Logistic Regression Model 5

For our third model, we only include the predictor variables that have `theoretical effect on probability of collition`, which was provided as part of the definition of the variables.

Additionally, we remove the variables that were as

- "urban legends", such as `RED_CAR` and `SEX`.
- having a theoretical "unknown effect" on probability of collision, such as `EDUCATION`.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE + CAR_USE + CLM_FREQ + HOME_VAL +
##     INCOME + JOB + KIDSDRIV + MSTATUS + MVR_PTS + REVOKED + TIF +
##     TRAVTIME + YOJ, family = binomial(link = "logit"), data = train)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0607  -0.7577  -0.5370   0.8177   2.8031
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -6.265e-02  2.593e-01  -0.242 0.809111
## AGE             -1.202e-02  3.727e-03  -3.225 0.001258 **
## CAR_USEPrivate  -5.537e-01  7.862e-02  -7.043 1.88e-12 ***
## CLM_FREQ         2.581e-01  2.677e-02   9.641  < 2e-16 ***
## HOME_VAL        -1.327e-06  3.610e-07  -3.675 0.000238 ***
## INCOME          -4.107e-06  1.080e-06  -3.804 0.000143 ***
## JOBBlue Collar   2.759e-01  1.452e-01   1.900 0.057447 .
## JOBClerical      1.945e-01  1.690e-01   1.151 0.249816
## JOBDoctor       -2.060e-01  2.650e-01  -0.777 0.436982
## JOBHome Maker    4.789e-02  2.004e-01   0.239 0.811094
## JOBLawyer        1.522e-02  1.784e-01   0.085 0.932011
## JOBManager      -4.282e-01  1.715e-01  -2.496 0.012546 *
## JOBProfessional  3.635e-02  1.598e-01   0.227 0.820111
## JOBStudent       2.181e-02  1.899e-01   0.115 0.908548
## KIDSDRIV         3.511e-01  5.551e-02   6.324 2.54e-10 ***
## MSTATUSYes      -5.051e-01  7.465e-02  -6.765 1.33e-11 ***
## MVR_PTS          1.374e-01  1.435e-02   9.579  < 2e-16 ***
## REVOKEDYes       7.957e-01  8.529e-02   9.329  < 2e-16 ***
## TIF             -5.016e-02  7.687e-03  -6.525 6.78e-11 ***
## TRAVTIME         6.508e-03  1.903e-03   3.420 0.000625 ***
## YOJ             -9.587e-03  8.753e-03  -1.095 0.273386
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7533.1  on 6527  degrees of freedom
## Residual deviance: 6528.6  on 6507  degrees of freedom
## AIC: 6570.6
##
## Number of Fisher Scoring iterations: 4
```

---

## Linear Regression Models

### Linear Regression Model 1

We begin with a `baseline` model that includes all the predictor variables and the response variable `TARGET_AMT`.

```
##
## Call:
## lm(formula = TARGET_AMT ~ . - TARGET_FLAG, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5952  -1694   -762    352 103691
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -6.848e+01  6.487e+02  -0.106 0.915928
## KIDSDRIV                     3.233e+02  1.284e+02   2.517 0.011852 *
```

```
## AGE                          5.760e+00  7.980e+00   0.722 0.470441
## HOMEKIDS                     5.912e+01  7.359e+01   0.803 0.421818
## YOJ                         -8.437e+00  1.709e+01  -0.494 0.621472
## INCOME                      -4.093e-03  2.015e-03  -2.031 0.042292 *
## PARENT1Yes                   7.032e+02  2.278e+02   3.086 0.002034 **
## HOME_VAL                    -6.449e-04  6.619e-04  -0.974 0.329934
## MSTATUSYes                  -5.344e+02  1.638e+02  -3.263 0.001109 **
## SEXM                         4.437e+02  2.076e+02   2.137 0.032619 *
## EDUCATIONBachelors          -4.616e+02  2.309e+02  -1.999 0.045621 *
## EDUCATIONHigh School        -1.545e+02  1.948e+02  -0.793 0.427742
## EDUCATIONMasters            -1.577e+02  3.407e+02  -0.463 0.643392
## EDUCATIONPhD                 8.225e+01  3.995e+02   0.206 0.836881
## JOBBlue Collar               2.763e+02  3.644e+02   0.758 0.448271
## JOBClerical                  2.679e+02  3.863e+02   0.694 0.487944
## JOBDoctor                   -6.130e+02  4.577e+02  -1.339 0.180560
## JOBHome Maker                2.059e+02  4.123e+02   0.499 0.617610
## JOBLawyer                    9.217e+01  3.354e+02   0.275 0.783475
## JOBManager                  -6.918e+02  3.283e+02  -2.107 0.035141 *
## JOBProfessional              3.274e+02  3.507e+02   0.933 0.350668
## JOBStudent                   9.435e+01  4.233e+02   0.223 0.823627
## TRAVTIME                     1.127e+01  3.625e+00   3.109 0.001885 **
## CAR_USEPrivate              -8.458e+02  1.851e+02  -4.570 4.97e-06 ***
## BLUEBOOK                     1.762e-02  9.787e-03   1.800 0.071865 .
## TIF                         -5.279e+01  1.362e+01  -3.875 0.000108 ***
## CAR_TYPEPanel Truck         -1.110e+02  3.170e+02  -0.350 0.726208
## CAR_TYPEPickup               3.659e+02  1.913e+02   1.912 0.055891 .
## CAR_TYPESports Car           9.907e+02  2.452e+02   4.040 5.42e-05 ***
## CAR_TYPESUV                  7.973e+02  2.021e+02   3.944 8.09e-05 ***
## CAR_TYPEVan                  5.866e+02  2.397e+02   2.447 0.014433 *
## RED_CARyes                  -1.247e+02  1.678e+02  -0.743 0.457580
## OLDCLAIM                    -1.191e-02  8.465e-03  -1.408 0.159304
## CLM_FREQ                     1.114e+02  6.196e+01   1.798 0.072230 .
## REVOKEDYes                   4.205e+02  1.961e+02   2.145 0.032007 *
## MVR_PTS                      1.754e+02  2.927e+01   5.992 2.18e-09 ***
## CAR_AGE                     -2.352e+01  1.446e+01  -1.626 0.103938
## URBANICITYHighly Urban/ Urban  1.726e+03  1.565e+02  11.029  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4589 on 6490 degrees of freedom
## Multiple R-squared:  0.07099,    Adjusted R-squared:  0.06569
## F-statistic:  13.4 on 37 and 6490 DF,  p-value: < 2.2e-16
```

**Linear Regression Model 2**

For our second model, we only include the top 10 most important predictor variables that we gathered from our importance trained model `modelB`.

```
##
## Call:
## lm(formula = TARGET_AMT ~ URBANICITY + MVR_PTS + CAR_USE + CAR_TYPE +
##     CAR_TYPE + TIF + MSTATUS + TRAVTIME + REVOKED + PARENT1,
##     data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5989  -1671   -852    249 103828
##
## Coefficients:
```

```
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    227.510    267.099   0.852 0.394365
## URBANICITYHighly Urban/ Urban 1407.279    145.912   9.645  < 2e-16 ***
## MVR_PTS                        210.704     27.108   7.773 8.86e-15 ***
## CAR_USEPrivate                -971.332    139.845  -6.946 4.13e-12 ***
## CAR_TYPEPanel Truck            -43.760    255.789  -0.171 0.864166
## CAR_TYPEPickup                 369.661    185.113   1.997 0.045872 *
## CAR_TYPESports Car             799.531    204.728   3.905 9.50e-05 ***
## CAR_TYPESUV                    615.412    155.300   3.963 7.49e-05 ***
## CAR_TYPEVan                    590.626    225.135   2.623 0.008725 **
## TIF                            -53.139     13.671  -3.887 0.000103 ***
## MSTATUSYes                    -454.592    132.811  -3.423 0.000624 ***
## TRAVTIME                        12.849      3.632   3.537 0.000407 ***
## REVOKEDYes                     384.468    176.236   2.182 0.029178 *
## PARENT1Yes                     990.605    192.623   5.143 2.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4616 on 6514 degrees of freedom
## Multiple R-squared:  0.05666,    Adjusted R-squared:  0.05478
## F-statistic:  30.1 on 13 and 6514 DF,  p-value: < 2.2e-16
```

**Linear Regression Model 3**

For our third model, we only include the predictor variables that have `theoretical probably of effecting the payout if there is a crash`, which was provided as part of the definition of the variables.

```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK + CAR_AGE + CAR_TYPE + CLM_FREQ +
##     OLDCLAIM, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3763  -1597  -1117   -297 104469
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.042e+03  1.967e+02   5.295 1.23e-07 ***
## BLUEBOOK            1.810e-03  8.597e-03   0.210 0.833307
## CAR_AGE            -4.808e+01  1.072e+01  -4.486 7.37e-06 ***
## CAR_TYPEPanel Truck  7.741e+02  2.612e+02   2.963 0.003054 **
## CAR_TYPEPickup      6.882e+02  1.822e+02   3.777 0.000160 ***
## CAR_TYPESports Car   7.034e+02  2.115e+02   3.326 0.000886 ***
## CAR_TYPESUV         5.532e+02  1.611e+02   3.435 0.000597 ***
## CAR_TYPEVan         9.643e+02  2.268e+02   4.253 2.14e-05 ***
## CLM_FREQ            4.042e+02  5.779e+01   6.995 2.93e-12 ***
## OLDCLAIM            4.720e-03  7.728e-03   0.611 0.541369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4705 on 6518 degrees of freedom
## Multiple R-squared:  0.01945,    Adjusted R-squared:  0.01809
## F-statistic: 14.36 on 9 and 6518 DF,  p-value: < 2.2e-16
```

# Model Selection

## Conclusions

## Code Appendix