Eric_Hirsch_621_Assignment_4

Predicting Insurance Claims

Eric Hirsch

4/7/2022

Contents

We examine records of car insurance customers to build two predictive models: one for whether the customer would have crashed, and second, the $ amount paid for the crash.

The main issue in the dataset are outliers. Without transformation, the distribution of residuals is not normal, and there are too many influential points to create reliable models.

1. Data Exploration

A. Summary Statistics We first examine the data. The dataset consists of 8161 observations and 26 variables (including two target variables, TARGET_FLAG and TARGET_AMT). 14 of the predictor variables are numeric. Approximately 27% of customers had an accident - the rest did not. TARGET_AMT appears to be highly skewed. There is a large degree of missing values.

```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ           INCOME
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Min.   :     0
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097
##  Median :45.00   Median :0.0000   Median :11.0   Median : 54028
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5   Mean   : 61898
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   Max.   :367030
##  NA's   :6                        NA's   :454    NA's   :445
##    PARENT1             HOME_VAL        MSTATUS             SEX
##  Length:8161        Min.   :     0   Length:8161        Length:8161
##  Class :character   1st Qu.:     0   Class :character   Class :character
##  Mode  :character   Median :161160   Mode  :character   Mode  :character
##                     Mean   :154867
##                     3rd Qu.:238724
##                     Max.   :885282
##                     NA's   :464
##   EDUCATION             JOB               TRAVTIME        CAR_USE
##  Length:8161        Length:8161        Min.   :  5.00   Length:8161
##  Class :character   Class :character   1st Qu.: 22.00   Class :character
##  Mode  :character   Mode  :character   Median : 33.00   Mode  :character
##                                        Mean   : 33.49
##                                        3rd Qu.: 44.00
##                                        Max.   :142.00
##
##     BLUEBOOK          TIF            CAR_TYPE           RED_CAR
##  Min.   : 1500   Min.   : 1.000   Length:8161        Length:8161
##  1st Qu.: 9280   1st Qu.: 1.000   Class :character   Class :character
##  Median :14440   Median : 4.000   Mode  :character   Mode  :character
##  Mean   :15710   Mean   : 5.351
##  3rd Qu.:20850   3rd Qu.: 7.000
##  Max.   :69740   Max.   :25.000
##
##     OLDCLAIM        CLM_FREQ        REVOKED             MVR_PTS
##  Min.   :    0   Min.   :0.0000   Length:8161        Min.   : 0.000
##  1st Qu.:    0   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Median :    0   Median :0.0000   Mode  :character   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   Length:8161
##  1st Qu.: 1.000   Class :character
##  Median : 8.000   Mode  :character
##  Mean   : 8.328
##  3rd Qu.:12.000
```

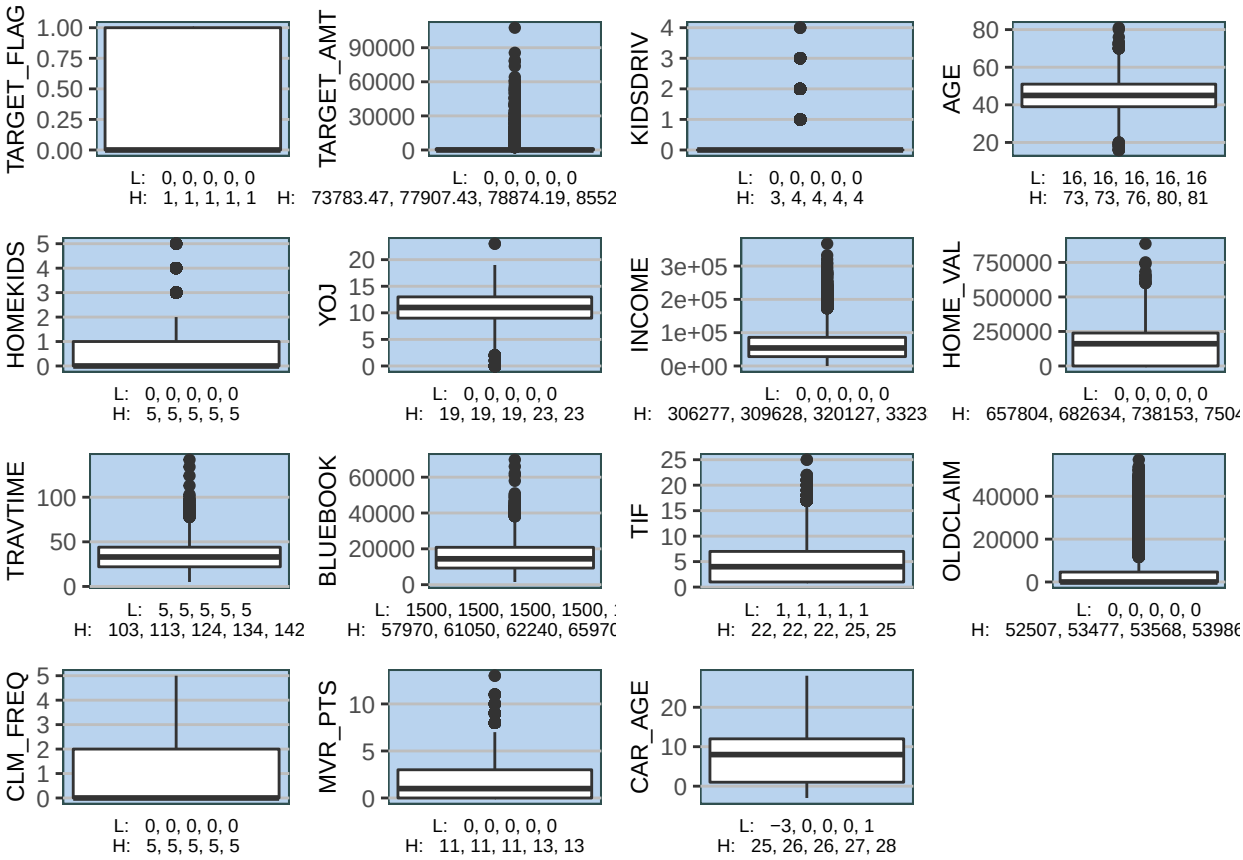
```
##  Max.   :28.000
##  NA's   :510
```

```
## 'data.frame':    8161 obs. of  26 variables:
##  $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
##  $ TARGET_AMT : num  0 0 0 0 0 ...
##  $ KIDSDRIV   : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
##  $ INCOME     : num  67349 91449 16039 NA 114986 ...
##  $ PARENT1    : chr  "No" "No" "No" "No" ...
##  $ HOME_VAL   : num  0 257252 124191 306251 243925 ...
##  $ MSTATUS    : chr  "z_No" "z_No" "Yes" "Yes" ...
##  $ SEX        : chr  "M" "M" "z_F" "M" ...
##  $ EDUCATION  : chr  "PhD" "z_High School" "z_High School" "<High School" ...
##  $ JOB        : chr  "Professional" "z_Blue Collar" "Clerical" "z_Blue Collar" ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : chr  "Private" "Commercial" "Private" "Private" ...
##  $ BLUEBOOK   : num  14230 14940 4010 15440 18000 ...
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : chr  "Minivan" "Minivan" "z_SUV" "Minivan" ...
##  $ RED_CAR    : chr  "yes" "yes" "no" "yes" ...
##  $ OLDCLAIM   : num  4461 0 38690 0 19217 ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : chr  "No" "No" "No" "No" ...
##  $ MVR_PTS    : int  3 0 3 0 3 0 0 10 0 1 ...
##  $ CAR_AGE    : int  18 1 10 6 17 7 1 7 1 17 ...
##  $ URBANICITY : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban,
```

B. Distributions We examine distributions of numeric variables through boxplots and histograms:

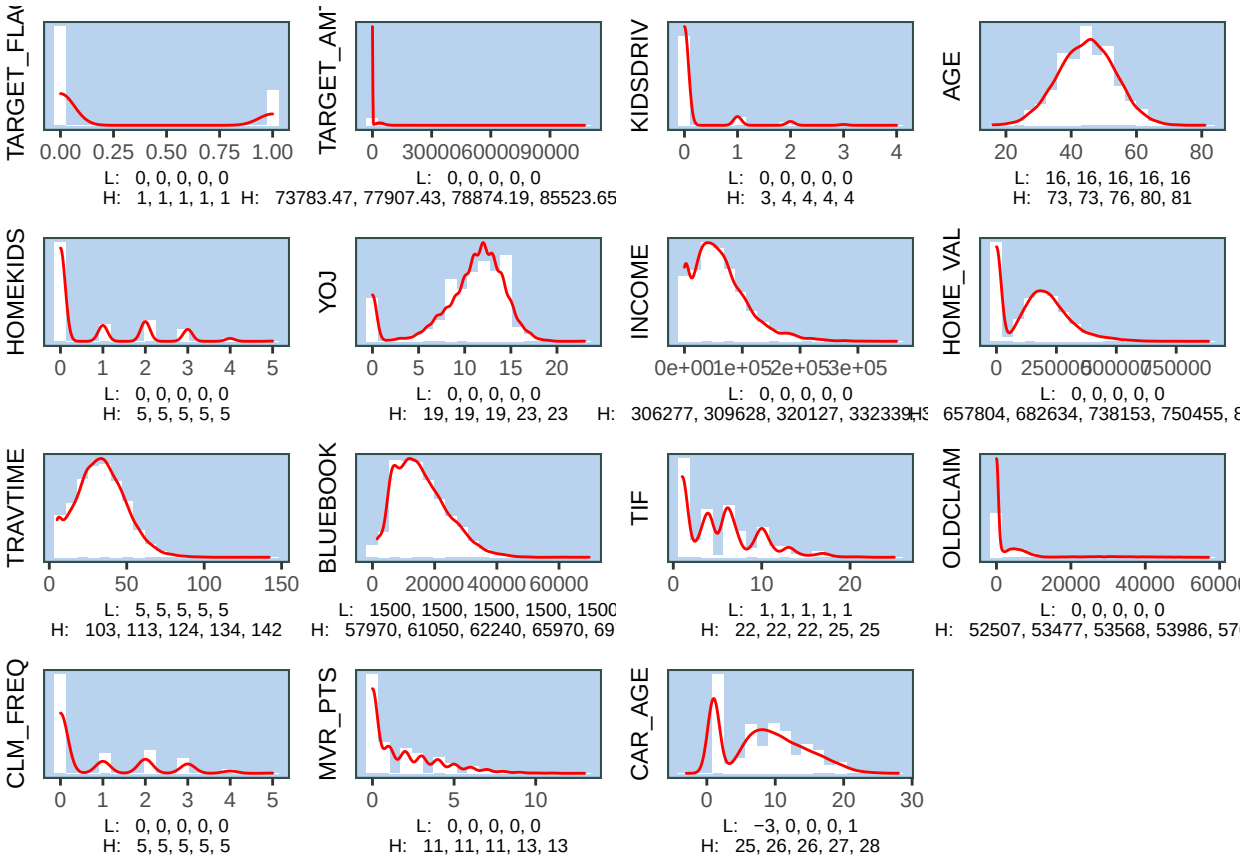
1. Boxplots

The boxplots show significant skewness and outliers.



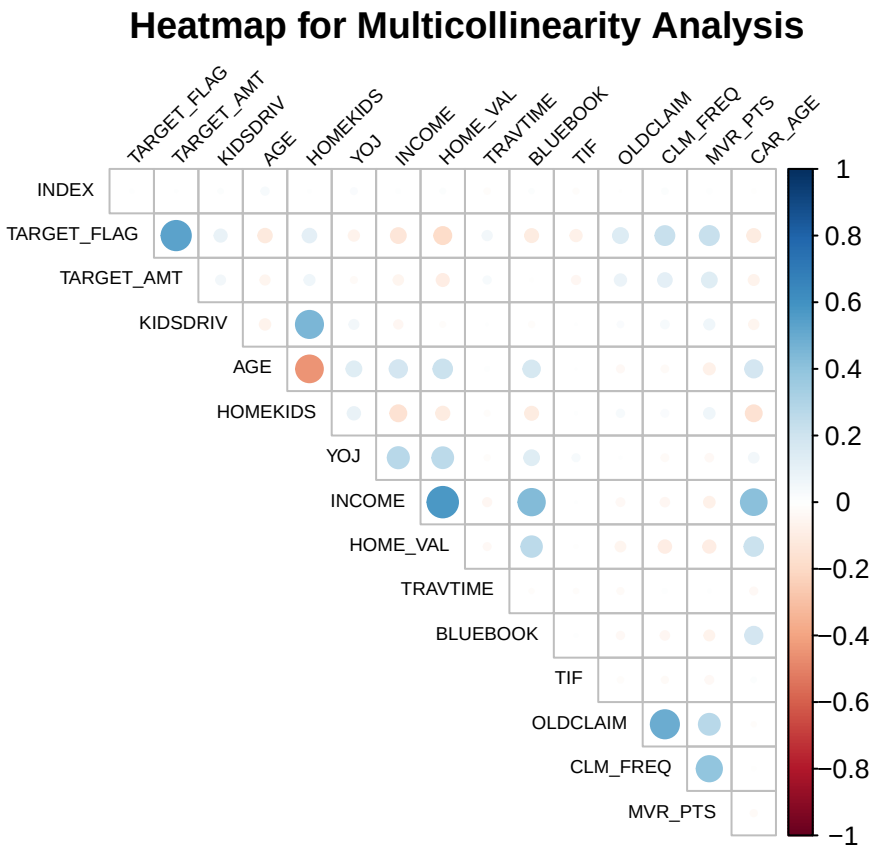
2. Histograms

We can see from the histograms a number of opportunities to perform log and other transformations.



Many of the variables are highly skewed, particularly TARGET_AMT. The level of outliers is very high.

C. Multicollinearity The chart below shows multicollinearity for numerical variables only. There are no surprises here - older people tend not to have children at home, income and home value are related, etc. Multicollinearity does not present offhand as a major issue.



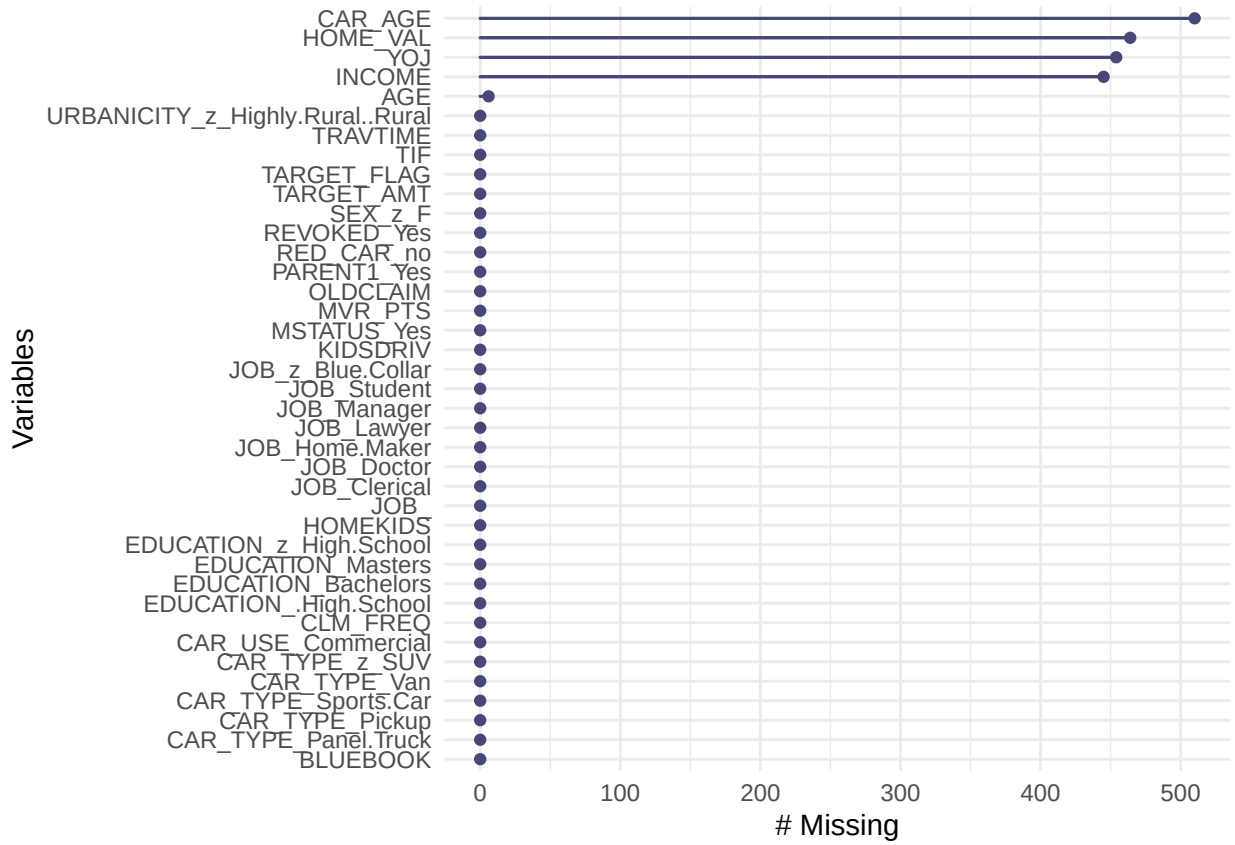
2. Data Preparation

A. Create Dummy Variables We create dummy variables from the character variables in the database.

B. Address Missing Values We consider the missing values. Over 20% of the records have missing values.

We disregard missing values in character columns because these NAs were isolated out in their own columns when we dummified the data. We convert the 0s in INCOME and HOME_VAL to NA since 0 is implausible. We create flags to track the NAs for the columns with the most significant NAs - INCOME, HOME_VAL, CAR_AGE, and YOJ. Finally we use MICE to populate the missing values.

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



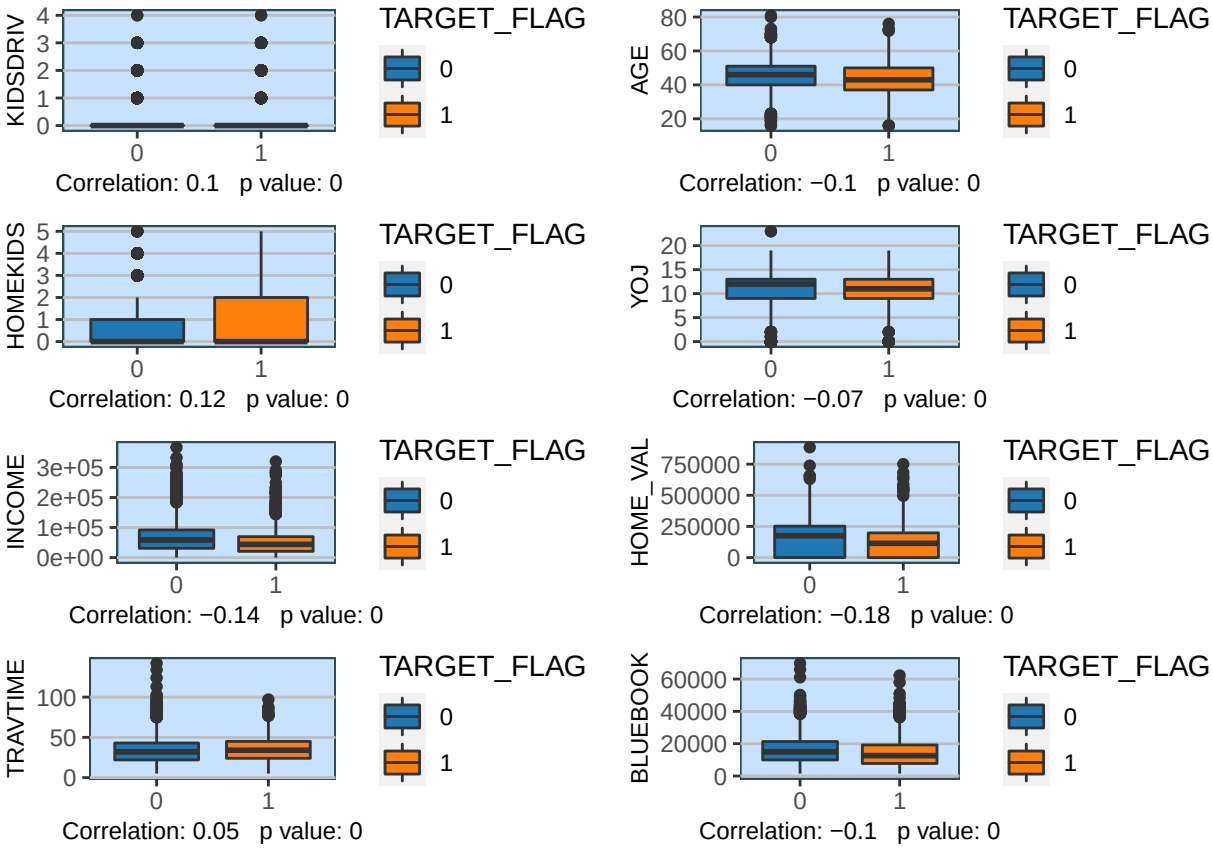
####. C. Perform Transformations

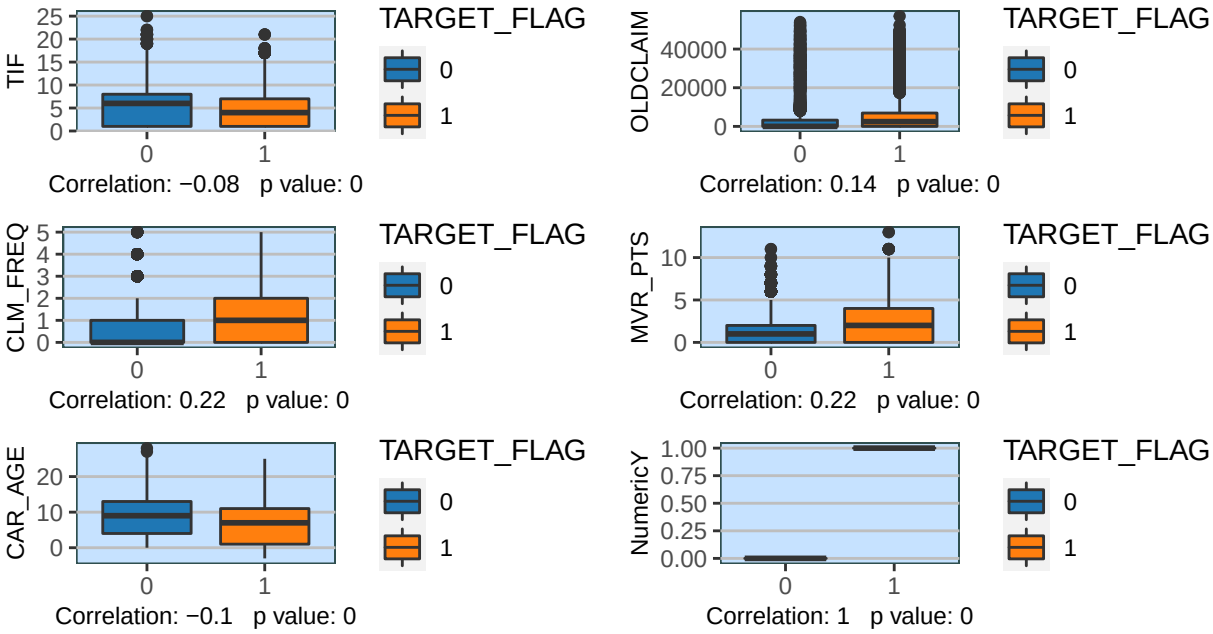
We perform log and other transformations, as well as add an interaction term, to the analysis. These transformations are based on an examination of the distributions of the indepenent variables. They include:

ageSquared
yojSquared
income_log
homeval_log
travtime_log
bluebook_log
carage_log
oldclaim_log
clm_freq_log
mvr_pts_log
tif_log
kidsdriv_log
homekids_log
inter (interaction term = KIDSDRIV*AGE

3. Predict TARGET_FLAG

A. Explore relationships We can see from the boxplots run on the original numeric variables against TARGET_FLAG that the correlations are quite low.





B. Create Models

Create Model 1 - a base model with the original numeric variables.

```
##
## Call:
## glm(formula = fla, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0236  -0.7694  -0.5736   0.9104   2.6679
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.351e-01  2.109e-01  -2.063  0.03910 *
## KIDSDRIV     2.484e-01  6.211e-02   3.999 6.35e-05 ***
## AGE         -1.042e-02  4.047e-03  -2.574  0.01006 *
## HOMEKIDS     7.194e-02  3.355e-02   2.144  0.03204 *
## YOJ         -5.932e-03  7.734e-03  -0.767  0.44309
## INCOME      -2.767e-07  9.289e-07  -0.298  0.76576
## HOME_VAL    -2.502e-06  2.932e-07  -8.532  < 2e-16 ***
## TRAVTIME     7.828e-03  1.885e-03   4.153 3.28e-05 ***
## BLUEBOOK    -1.260e-05  4.135e-06  -3.047  0.00231 **
## TIF         -4.439e-02  7.604e-03  -5.838 5.29e-09 ***
```

```
## OLDCLAIM     6.626e-06  3.501e-06   1.893  0.05842 .
## CLM_FREQ     2.683e-01  2.866e-02   9.361  < 2e-16 ***
## MVR_PTS      1.397e-01  1.410e-02   9.910  < 2e-16 ***
## CAR_AGE     -2.387e-02  5.987e-03  -3.987 6.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7445.1  on 6447  degrees of freedom
## Residual deviance: 6674.3  on 6434  degrees of freedom
##   (1713 observations deleted due to missingness)
## AIC: 6702.3
##
## Number of Fisher Scoring iterations: 4
```

```
## [[1]]
##   (Intercept)      KIDSDRIV           AGE      HOMEKIDS           YOJ
## -4.350637e-01  2.484120e-01 -1.041761e-02  7.193553e-02 -5.931651e-03
##        INCOME      HOME_VAL      TRAVTIME      BLUEBOOK           TIF
## -2.767441e-07 -2.501621e-06  7.828270e-03 -1.259763e-05 -4.438976e-02
##      OLDCLAIM      CLM_FREQ       MVR_PTS       CAR_AGE
##  6.626042e-06  2.682667e-01  1.397392e-01 -2.387229e-02
##
## [[2]]
## [1] 0
##
## [[3]]
## [1] 0
```

Most of the predictors are significant. This may in part be due to the fact that there are over 8,000 predictions. The model has an AIC of 6702. Almost 20% of the records are missing so the model is not necessarily reliable.

We run the model 100 times at a 80/20 split. The base model has an accuracy of .748, an AIC of 5266 and an AUC of .716.

Create Model 2 - a model with missing values addressed and all of the transformed and added variables included.

```
##
## Call:
## glm(formula = fla, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7164  -0.6978  -0.3849   0.5989   3.0770
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    9.370e+00  3.577e+00   2.620 0.008796 **
## KIDSDRIV                      -1.031e+00  5.121e-01  -2.013 0.044160 *
## AGE                           -1.962e-01  2.646e-02  -7.416 1.21e-13 ***
```

```
## HOMEKIDS                           -2.136e-01  1.537e-01  -1.390 0.164430
## YOJ                                -7.017e-02  3.271e-02  -2.145 0.031931 *
## INCOME                             -1.059e-05  3.245e-06  -3.263 0.001101 **
## HOME_VAL                            4.376e-06  2.282e-06   1.917 0.055225 .
## TRAVTIME                            4.900e-03  5.456e-03   0.898 0.369170
## BLUEBOOK                            4.828e-06  1.092e-05   0.442 0.658465
## TIF                                -7.933e-03  2.344e-02  -0.338 0.735042
## OLDCLAIM                           -2.782e-05  5.608e-06  -4.961 7.00e-07 ***
## CLM_FREQ                            1.874e-01  2.635e-01   0.711 0.476965
## MVR_PTS                             1.133e-01  4.321e-02   2.622 0.008736 **
## CAR_AGE                             8.331e-03  1.965e-02   0.424 0.671562
## PARENT1_Yes                         2.036e-01  1.217e-01   1.673 0.094244 .
## MSTATUS_Yes                        -6.146e-01  8.769e-02  -7.009 2.40e-12 ***
## SEX_z_F                            -4.600e-02  1.135e-01  -0.405 0.685361
## EDUCATION_.High.School              1.776e-01  2.212e-01   0.803 0.421930
## EDUCATION_Bachelors                -1.565e-01  1.875e-01  -0.835 0.403714
## EDUCATION_Masters                  -5.231e-02  1.559e-01  -0.336 0.737231
## EDUCATION_z_High.School             2.258e-01  2.033e-01   1.111 0.266763
## JOB_                               -1.751e-01  1.797e-01  -0.974 0.330022
## JOB_Clerical                        2.385e-01  1.274e-01   1.872 0.061239 .
## JOB_Doctor                         -5.960e-01  2.775e-01  -2.148 0.031718 *
## JOB_Home.Maker                     -1.446e-01  1.759e-01  -0.822 0.411063
## JOB_Lawyer                         -5.107e-02  1.694e-01  -0.301 0.763081
## JOB_Manager                        -7.251e-01  1.333e-01  -5.439 5.36e-08 ***
## JOB_Student                        -2.660e-01  1.775e-01  -1.498 0.134063
## JOB_z_Blue.Collar                   2.029e-01  1.212e-01   1.674 0.094202 .
## CAR_USE_Commercial                  7.684e-01  9.299e-02   8.263  < 2e-16 ***
## CAR_TYPE_Panel.Truck                5.206e-01  1.672e-01   3.113 0.001852 **
## CAR_TYPE_Pickup                     5.876e-01  1.019e-01   5.765 8.17e-09 ***
## CAR_TYPE_Sports.Car                 8.974e-01  1.325e-01   6.772 1.27e-11 ***
## CAR_TYPE_Van                        6.546e-01  1.281e-01   5.110 3.22e-07 ***
## CAR_TYPE_z_SUV                      7.259e-01  1.130e-01   6.422 1.35e-10 ***
## RED_CAR_no                          3.971e-02  8.747e-02   0.454 0.649852
## REVOKED_Yes                         9.685e-01  9.406e-02  10.297  < 2e-16 ***
## URBANICITY_z_Highly.Rural..Rural -2.369e+00  1.143e-01 -20.735  < 2e-16 ***
## YOJ_NA                             -5.498e-02  1.282e-01  -0.429 0.668095
## INCOME_NA                          -1.351e-01  1.274e-01  -1.061 0.288885
## CAR_AGE_NA                         -1.530e-01  1.193e-01  -1.283 0.199554
## HOME_VAL_NA                        -2.590e-01  7.785e-02  -3.327 0.000878 ***
## ageSquared                          2.107e-03  2.908e-04   7.245 4.33e-13 ***
## yojSquared                          3.284e-03  1.750e-03   1.876 0.060595 .
## income_log                         -5.300e-02  6.199e-02  -0.855 0.392558
## homeval_log                        -2.485e-01  3.413e-01  -0.728 0.466634
## travtime_log                        2.947e-01  1.542e-01   1.911 0.055958 .
## bluebook_log                       -3.583e-01  1.246e-01  -2.877 0.004015 **
## carage_log                         -8.735e-02  1.221e-01  -0.716 0.474288
## oldclaim_log                        9.244e-02  4.254e-02   2.173 0.029783 *
## clm_freq_log                       -4.504e-01  7.973e-01  -0.565 0.572122
## mvr_pts_log                        -7.221e-02  1.293e-01  -0.558 0.576596
## tif_log                            -2.875e-01  1.344e-01  -2.138 0.032482 *
## kidsdriv_log                        1.452e+00  6.156e-01   2.359 0.018334 *
## homekids_log                        4.649e-01  3.446e-01   1.349 0.177373
## inter                               1.652e-02  8.421e-03   1.962 0.049747 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9415.3  on 8159  degrees of freedom
## Residual deviance: 7156.9  on 8104  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 7268.9
##
## Number of Fisher Scoring iterations: 5
```

```
## [[1]]
##                       (Intercept)                          KIDSDRIV
##                      9.370156e+00                     -1.030672e+00
##                               AGE                          HOMEKIDS
##                     -1.962298e-01                     -2.136347e-01
##                               YOJ                            INCOME
##                     -7.016521e-02                     -1.059150e-05
##                          HOME_VAL                          TRAVTIME
##                      4.375603e-06                      4.899993e-03
##                          BLUEBOOK                               TIF
##                      4.827554e-06                     -7.932672e-03
##                          OLDCLAIM                          CLM_FREQ
##                     -2.782041e-05                      1.873694e-01
##                           MVR_PTS                           CAR_AGE
##                      1.132953e-01                      8.331462e-03
##                       PARENT1_Yes                       MSTATUS_Yes
##                      2.035901e-01                     -6.146059e-01
##                           SEX_z_F             EDUCATION_.High.School
##                     -4.600387e-02                      1.776204e-01
##                EDUCATION_Bachelors                  EDUCATION_Masters
##                     -1.565376e-01                     -5.231392e-02
##            EDUCATION_z_High.School                               JOB_
##                      2.257944e-01                     -1.750627e-01
##                       JOB_Clerical                        JOB_Doctor
##                      2.385014e-01                     -5.960141e-01
##                     JOB_Home.Maker                        JOB_Lawyer
##                     -1.446106e-01                     -5.107468e-02
##                        JOB_Manager                       JOB_Student
##                     -7.251163e-01                     -2.659940e-01
##                  JOB_z_Blue.Collar                  CAR_USE_Commercial
##                      2.028571e-01                      7.683794e-01
##              CAR_TYPE_Panel.Truck                    CAR_TYPE_Pickup
##                      5.205814e-01                      5.875796e-01
##               CAR_TYPE_Sports.Car                       CAR_TYPE_Van
##                      8.974411e-01                      6.545880e-01
##                     CAR_TYPE_z_SUV                        RED_CAR_no
##                      7.259328e-01                      3.970691e-02
##                       REVOKED_Yes URBANICITY_z_Highly.Rural..Rural
##                      9.684759e-01                     -2.369379e+00
##                            YOJ_NA                            INCOME_NA
##                     -5.497836e-02                     -1.350652e-01
##                        CAR_AGE_NA                        HOME_VAL_NA
##                     -1.529934e-01                     -2.590234e-01
```

```
##                      ageSquared                      yojSquared
##                    2.106869e-03                    3.283852e-03
##                      income_log                      homeval_log
##                    -5.299784e-02                    -2.484832e-01
##                     travtime_log                     bluebook_log
##                     2.946639e-01                    -3.583411e-01
##                      carage_log                     oldclaim_log
##                    -8.734651e-02                     9.244048e-02
##                     clm_freq_log                      mvr_pts_log
##                    -4.504258e-01                    -7.221226e-02
##                          tif_log                     kidsdriv_log
##                    -2.874765e-01                     1.451993e+00
##                     homekids_log                            inter
##                     4.648545e-01                     1.652350e-02
##
## [[2]]
## [1] 0
##
## [[3]]
## [1] 0
```

In the "kitchen sink" model many of the predictors are not significant. This model risks overprediction. AIC has increased to 7269.

We run the model 100 times at a 80/20 split. The kitchen sink model has an accuracy of .79, an AIC of 5723 and an AUC of .815. Despite possible overprediction, this model offers significant improvement.

Create Model 3 - Use backward elmination to choose the best model: We use backward elimination to achieve a better fit and lower AIC.

```
##
## Call:
## glm(formula = fla, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5304  -0.7393  -0.4235   0.7285   2.8625
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -9.854e-01  1.277e-01  -7.716 1.20e-14 ***
## INCOME                          -7.501e-06  7.651e-07  -9.805  < 2e-16 ***
## TRAVTIME                         1.395e-02  1.837e-03   7.594 3.11e-14 ***
## BLUEBOOK                        -1.321e-05  4.388e-06  -3.010  0.00261 **
## TIF                             -5.591e-02  7.193e-03  -7.773 7.69e-15 ***
## OLDCLAIM                        -4.926e-06  3.875e-06  -1.271  0.20362
## PARENT1_Yes                      7.579e-01  8.255e-02   9.181  < 2e-16 ***
## SEX_z_F                         -2.041e-01  8.798e-02  -2.319  0.02038 *
## JOB_Manager                     -8.476e-01  1.041e-01  -8.139 4.00e-16 ***
## CAR_USE_Commercial               9.257e-01  6.478e-02  14.291  < 2e-16 ***
## CAR_TYPE_Pickup                  3.612e-01  8.695e-02   4.155 3.26e-05 ***
## CAR_TYPE_Sports.Car              9.408e-01  1.232e-01   7.639 2.19e-14 ***
## CAR_TYPE_z_SUV                   6.999e-01  1.049e-01   6.671 2.54e-11 ***
```

```
## URBANICITY_z_Highly.Rural..Rural -2.342e+00  1.119e-01 -20.933  < 2e-16 ***
## HOME_VAL_NA                      -4.871e-01  6.124e-02  -7.954 1.80e-15 ***
## oldclaim_log                      7.761e-02  8.507e-03   9.123  < 2e-16 ***
## inter                             9.684e-03  1.208e-03   8.018 1.08e-15 ***
## ---
## 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: 7553.3  on 8144  degrees of freedom
## AIC: 7587.3
##
## Number of Fisher Scoring iterations: 5
```

```
## [[1]]
##                       (Intercept)                            INCOME
##                     -9.853727e-01                     -7.501381e-06
##                          TRAVTIME                          BLUEBOOK
##                      1.395330e-02                     -1.320759e-05
##                               TIF                           OLDCLAIM
##                     -5.590842e-02                     -4.925857e-06
##                       PARENT1_Yes                            SEX_z_F
##                      7.579452e-01                     -2.040512e-01
##                       JOB_Manager                CAR_USE_Commercial
##                     -8.475772e-01                      9.257383e-01
##                   CAR_TYPE_Pickup              CAR_TYPE_Sports.Car
##                      3.612426e-01                      9.407812e-01
##                   CAR_TYPE_z_SUV URBANICITY_z_Highly.Rural..Rural
##                      6.999295e-01                     -2.341684e+00
##                       HOME_VAL_NA                       oldclaim_log
##                     -4.870927e-01                      7.760677e-02
##                             inter
##                      9.684202e-03
##
## [[2]]
## [1] 0
##
## [[3]]
## [1] 0
```

The refined model has an AIC of 7587. This is not an improvement.

We run the model 100 times at a 80/20 split. The refined model has an accuracy of .77, an AIC of 5805 and an AUC of .79. Despite lower AIC, this model does not predict the data as well. Again, the AIC does not fall and the model does not predict as well.

4. Select model Below is a table of results:

	Base Model	Kitchen Sink Model	Refined Model
Accuracy	0.748	0.790	0.770
AIC	5266.000	5723.000	5805.000

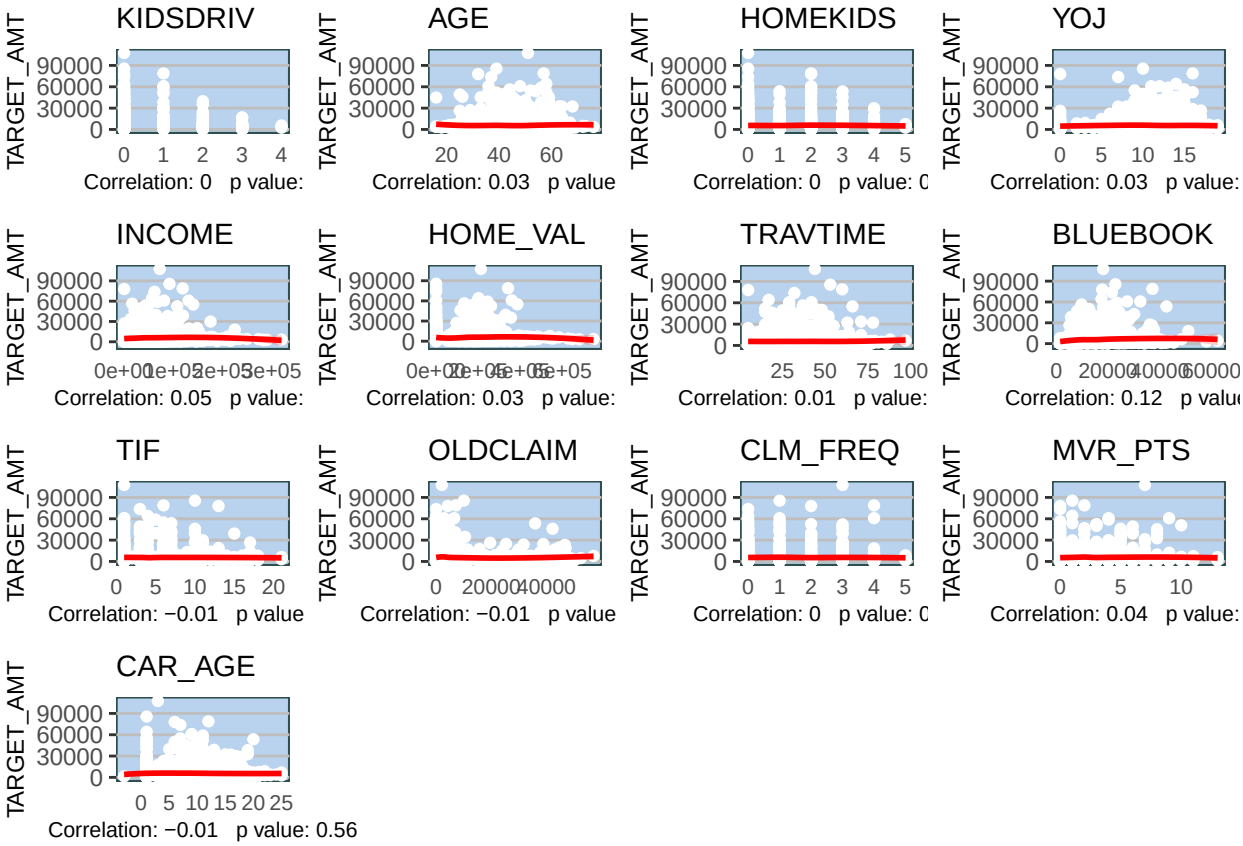
	Base Model	Kitchen Sink Model	Refined Model
AUC	0.716	0.815	0.795

Despite the apparent superior predictability of the second model, we choose the third. This model has a lower AIC and is more interpretable and coherent than the second model.

4. Predict TARGET_AMT

Now we predict the target amount for those customers who have had an accident.

A.Explore Relationships We look at scatterplots of numeric variables:

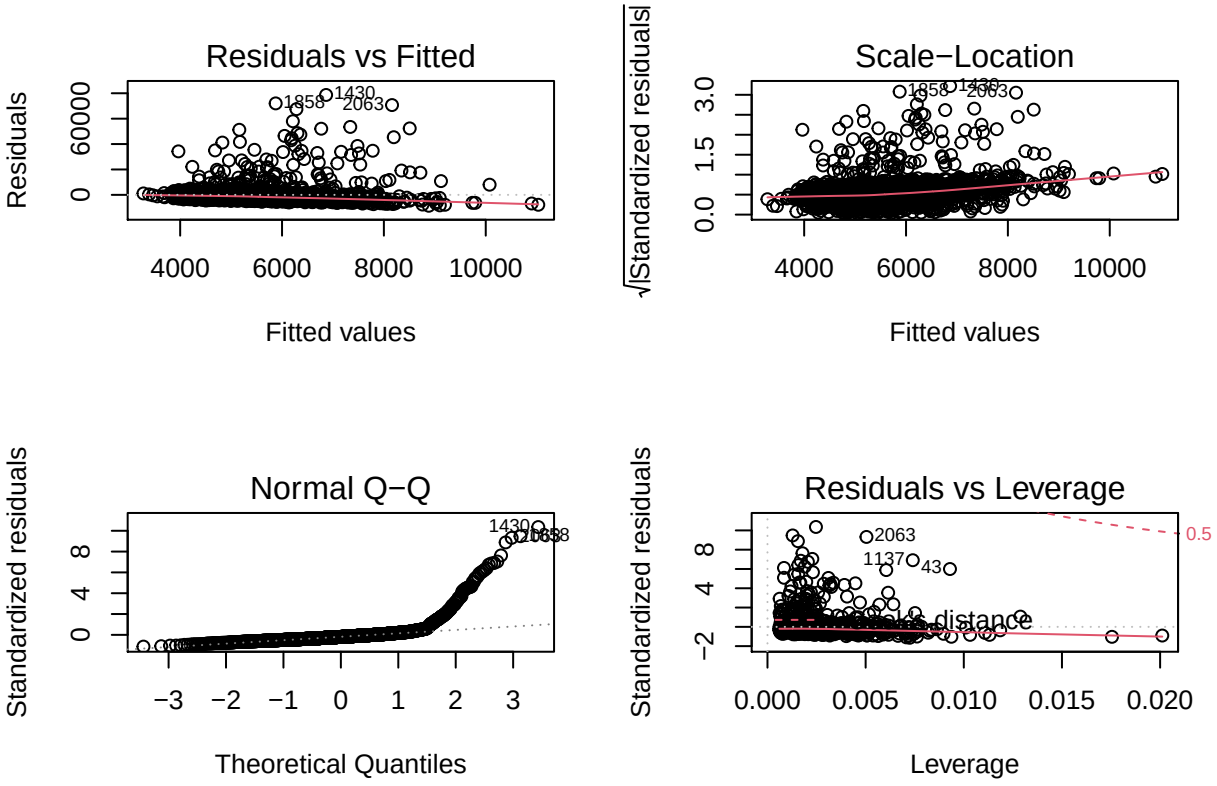


We can see that correlations are quite low.

B. Create models

Create Model 1 - the base model with the original numeric variables. We use stepAIC to choose the best model. The model retains very few predictor variables.

```
##
## Call:
## lm(formula = TARGET_AMT ~ BLUEBOOK + MVR_PTS + CAR_AGE, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8566  -3086  -1519    287  78658
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4021.48678  459.06110   8.760  < 2e-16 ***
## BLUEBOOK       0.12199    0.02243   5.439 6.15e-08 ***
## MVR_PTS      109.75419   70.96899   1.547    0.122
## CAR_AGE      -50.90781   34.04373  -1.495    0.135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7605 on 1699 degrees of freedom
## Multiple R-squared:  0.01821,    Adjusted R-squared:  0.01648
## F-statistic: 10.51 on 3 and 1699 DF,  p-value: 7.581e-07
##
## [1] "VIF Analysis"
## BLUEBOOK  MVR_PTS  CAR_AGE
## 1.037901 1.005499 1.036879
```



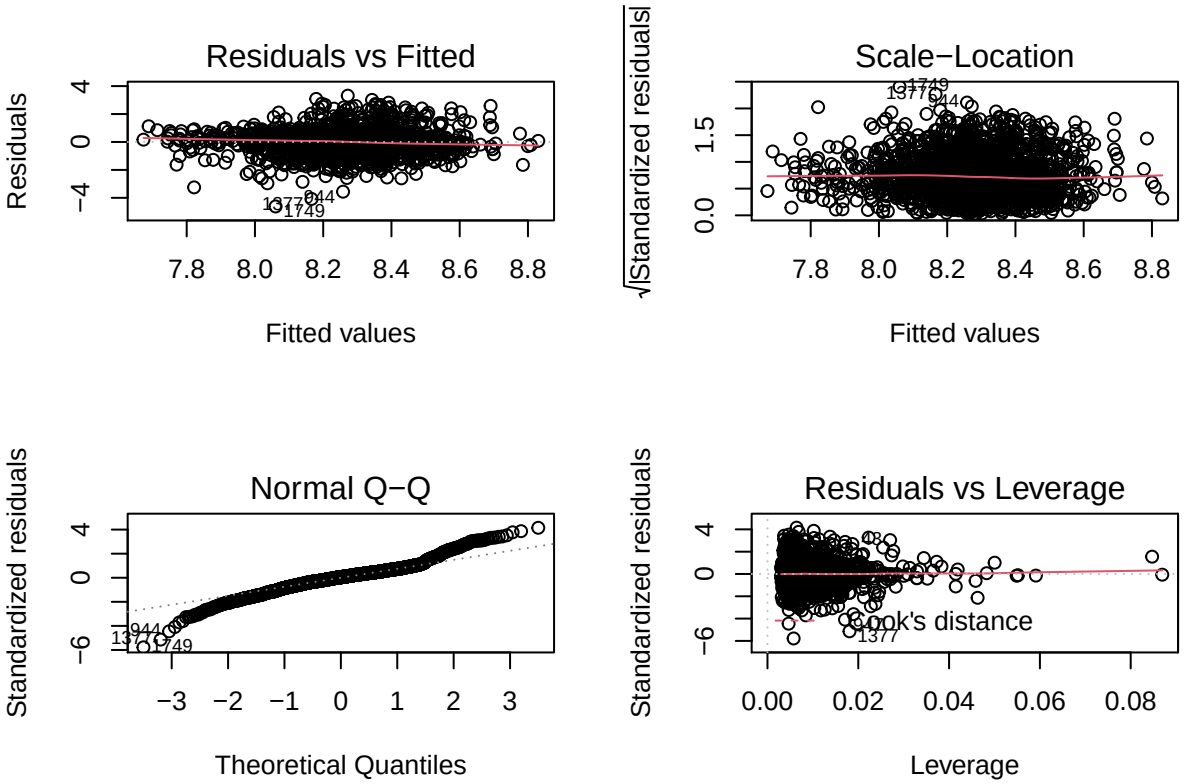

```
## NULL
```

The base model shows a number of issues with the residuals including heteroskedastcity and particularly non-normal residuals. We cannot use this model without some transformation.

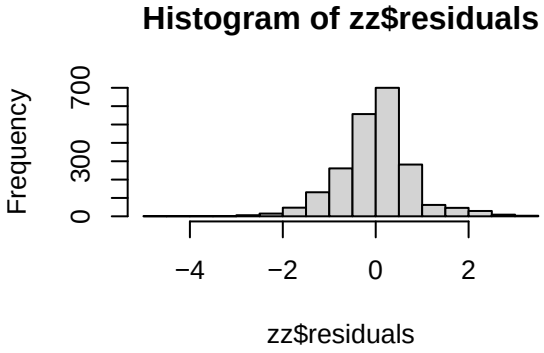
Create Model 2 - a model with missing values addressed and all of the transformed and added variables included. We also take the log of TARGET_AMT. We find the model with the highest AIC using stepAIC from the MASS package in R.

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + INCOME + BLUEBOOK + OLDCLAIM +
##     CLM_FREQ + CAR_AGE + MSTATUS_Yes + EDUCATION_.High.School +
##     EDUCATION_Bachelors + EDUCATION_Masters + EDUCATION_z_High.School +
##     CAR_TYPE_Panel.Truck + REVOKED_Yes + homeval_log + bluebook_log +
##     carage_log + mvr_pts_log + inter, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6176 -0.4021  0.0342  0.3988  3.3141
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.665e+00  9.932e-01   4.696 2.82e-06 ***
## KIDSDRIV                -3.378e-01  1.693e-01  -1.995  0.04616 *
## INCOME                  -1.943e-06  9.973e-07  -1.949  0.05148 .
## BLUEBOOK                -1.227e-05  6.129e-06  -2.002  0.04539 *
## OLDCLAIM                 4.793e-06  2.344e-06   2.045  0.04101 *
## CLM_FREQ                -3.893e-02  1.630e-02  -2.388  0.01704 *
## CAR_AGE                 -2.693e-02  1.255e-02  -2.146  0.03197 *
## MSTATUS_Yes             -8.359e-02  3.493e-02  -2.393  0.01681 *
## EDUCATION_.High.School  -2.633e-01  1.118e-01  -2.356  0.01855 *
## EDUCATION_Bachelors     -2.908e-01  9.546e-02  -3.047  0.00234 **
## EDUCATION_Masters       -1.310e-01  9.008e-02  -1.455  0.14586
## EDUCATION_z_High.School -2.579e-01  1.034e-01  -2.493  0.01273 *
## CAR_TYPE_Panel.Truck     1.635e-01  7.864e-02   2.079  0.03777 *
## REVOKED_Yes             -9.493e-02  5.343e-02  -1.777  0.07578 .
## homeval_log              1.223e-01  8.066e-02   1.516  0.12962
## bluebook_log             2.757e-01  6.974e-02   3.953 7.96e-05 ***
## carage_log               1.483e-01  7.448e-02   1.991  0.04664 *
## mvr_pts_log              5.922e-02  2.315e-02   2.559  0.01058 *
## inter                    7.489e-03  3.852e-03   1.944  0.05199 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8011 on 2133 degrees of freedom
## Multiple R-squared:  0.03496,    Adjusted R-squared:  0.02681
## F-statistic: 4.292 on 18 and 2133 DF,  p-value: 3.959e-09
##
## [1] "VIF Analysis"
##                KIDSDRIV                  INCOME                BLUEBOOK
##               37.786241                5.589714                8.678143
##                OLDCLAIM                CLM_FREQ                 CAR_AGE
##                1.862833                1.388044               15.925940
##             MSTATUS_Yes  EDUCATION_.High.School     EDUCATION_Bachelors
```

```
##             1.022736             6.152726             5.614654
##   EDUCATION_Masters EDUCATION_z_High.School  CAR_TYPE_Panel.Truck
##             3.506406             8.351117             1.573691
##          REVOKED_Yes          homeval_log          bluebook_log
##             1.562762             4.746742             7.136432
##          carage_log          mvr_pts_log                inter
##            12.705195             1.108395            37.848324
```




```
## NULL
```



The adjusted-R-Squared is very low, despite the number of significant variables and significance overall. The distribution of the residuals improves but the tails are still an issue. There still appear to be a large number of outliers.

Create Model 3 - a model using robust regression. If we remove a large number of outliers form the data our results improve dramatically. However, this must be true by definition as the variance will decrease when outliers are removed. A better way to discover the underlying pattern beneath the outliers (and to check that pattern against our second model) is robust regression.

4. Select model The following table summarizes the RSE and RSME when using robust regression on the base and refined models:

	Base Model	Refined Model
Standard - Sigma	7605.000	0.800
Standard - Adj-R^2	0.016	0.026
Robust - Sigma	2334.000	0.573

It should be noted that the refined model sigma is not comparable to the base model because we have taken the log of the dependent variable. The refined model clearly outperforms the base model. In addition, the robust model clearly outperforms the standard model. We therefore choose the robust refined model.

5. Conclusion

We examined 466 records of town statistics to create a predictive model of whether crime rates were above the median or not. We used a logistic regression to do this, testing our models on an 80/20 split 100 times and taking the average accuracy and AIC.

Several enhancements to the model increased accuracy and lowered AIC. First, some predictors were transformed with the log or square to improve fit. Second, dummy variables were introduced to capture the fact that highly industrial areas appeared to operate by a different logic than mixed use areas. interaction terms to model this phenomenon did not improve the model. The final model 93% accurate.