Homework 4

Hector Santana, Zachary Safir, Mario Pena

Contents

In this assignment we will explore, analyze, and build a multiple linear regression and binary logistic model based on auto insurance data. The models will predict the probability that a person will crash their car and then the subsequent insurance cost for the accident.

We are provided with information on a little over 8,000 customers at an auto insurance company. Each record has two response variables. TARGET_FLAG has a response of 1 if the customer was involved in a crash, or 0 if the customer was not involved in a crash.TARGET_AMT has a response of 0 if the customer did not crash their car, or a value greater than 0 otherwise. Additionally, there are 23 predictor variables in the data that could be of use for the model.

Let us take a look at a snippet of the data set:

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

Before we begin the data exploration process we will clean our data a bit in order to run summary statistics and plots accurately and effectively. We remove the dollar signs in the data, remove the extra z_ character found on some variables, and convert everything to the correct data format. Some variables, such as number of kids, we decided to make factors as they are not continuous variables.

1 DATA EXPLORATION

1.1 Numeric Variable Exploratipon

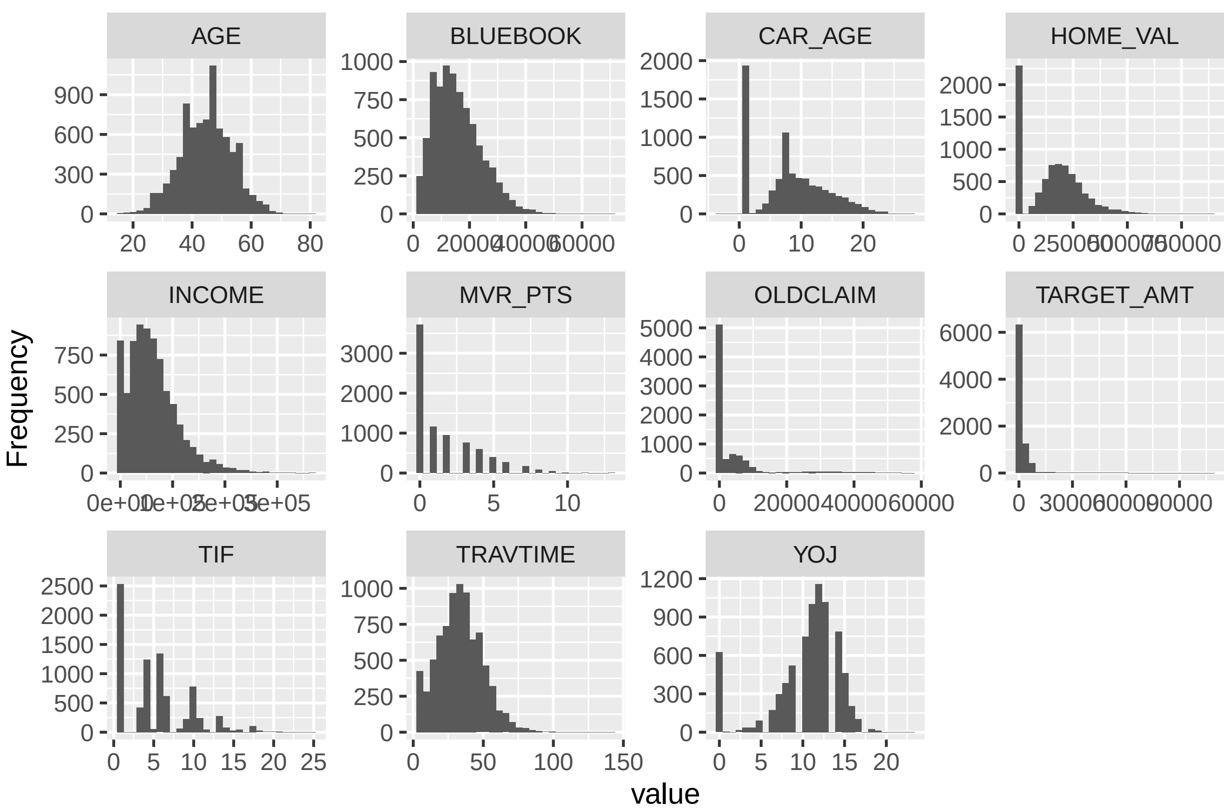
Below we have created a table with the summary statistics for our numeric predictor variables. We will explore the categorical variables later.

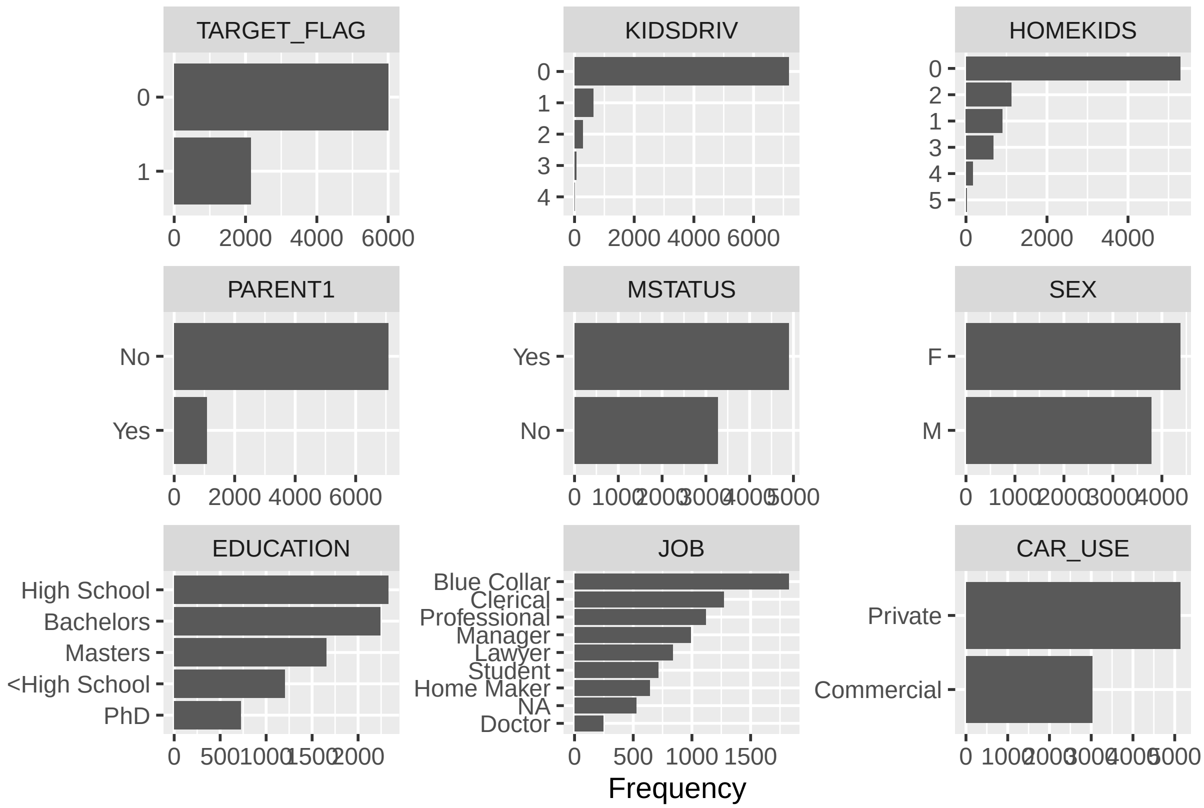
According to our summary statistics, we note that quite a few of our numeric variables appear skewed.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
TARGET_AMT	1	8161	1.504325e+03	4.704027e+03	0	5.937121e+02	0.0000	0	107586.1	107586.1	8.7063034	112.2884386	52.0712628
AGE	2	8155	4.479031e+01	8.627589e+00	45	4.483065e+01	8.8956	16	81.0	65.0	-0.0289889	-0.0617020	0.0955383
YOJ	3	7707	1.049929e+01	4.092474e+00	11	1.107119e+01	2.9652	0	23.0	23.0	-1.2029676	1.1773410	0.0466169
INCOME	4	7716	6.189809e+04	4.757268e+04	54028	5.684098e+04	41792.2701	0	367030.0	367030.0	1.1863166	2.1290163	541.5786485
HOME_VAL	5	7697	1.548673e+05	1.291238e+05	161160	1.440321e+05	147867.1110	0	885282.0	885282.0	0.4885950	-0.0160838	1471.7887185
TRAVTIME	6	8161	3.348572e+01	1.590833e+01	33	3.299541e+01	16.3086	5	142.0	137.0	0.4468174	0.6643331	0.1760974
BLUEBOOK	7	8161	1.570990e+04	8.419734e+03	14440	1.503689e+04	8450.8200	1500	69740.0	68240.0	0.7942141	0.7913559	93.2023121
TIF	8	8161	5.351305e+00	4.146635e+00	4	4.840251e+00	4.4478	1	25.0	24.0	0.8908120	0.4224940	0.0459012
OLDCLAIM	9	8161	4.037076e+03	8.777139e+03	0	1.719291e+03	0.0000	0	57037.0	57037.0	3.1190400	9.8606583	97.1586099
MVR_PTS	10	8161	1.695503e+00	2.147112e+00	1	1.313831e+00	1.4826	0	13.0	13.0	1.3478403	1.3754900	0.0237675
CAR_AGE	11	7651	8.328323e+00	5.700742e+00	8	7.963241e+00	7.4130	-3	28.0	31.0	0.2819531	-0.7489756	0.0651737

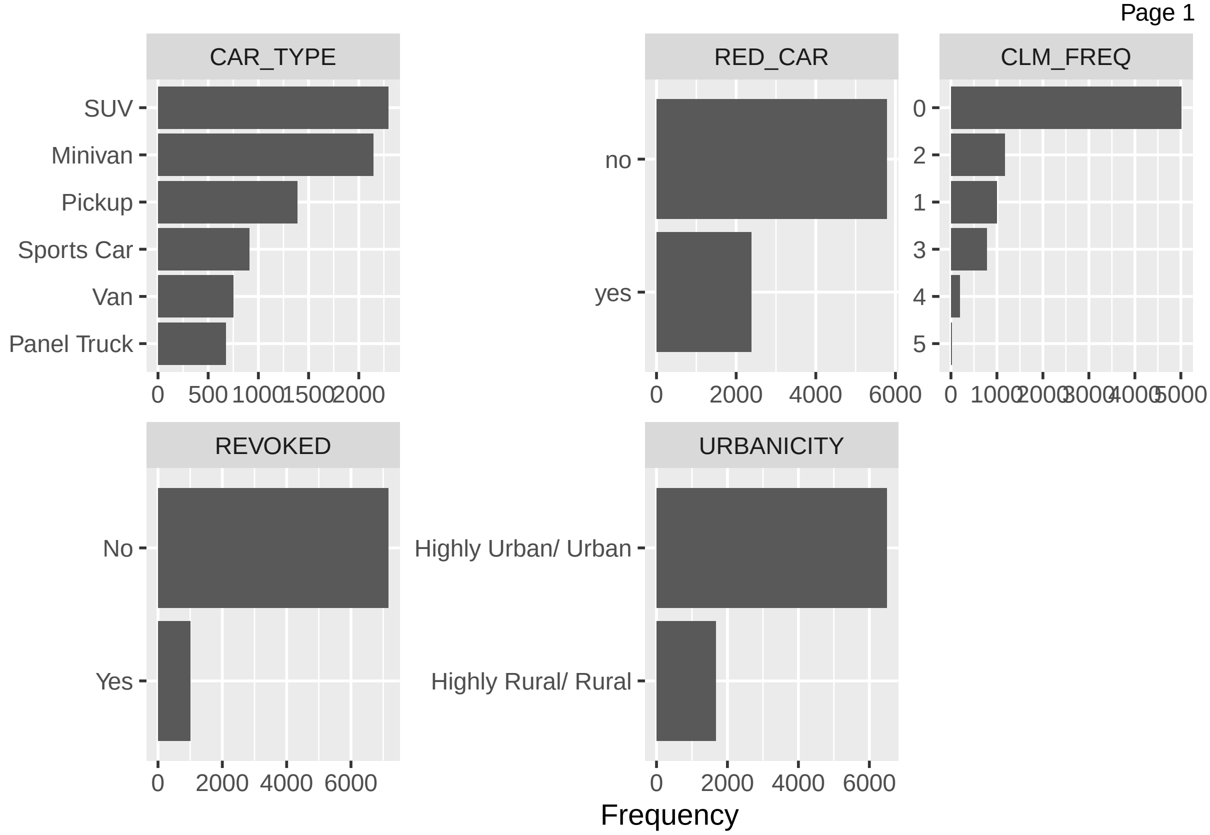
The insight gained from the statistical analysis permitted us to make note of further data of interest that needed to be analyzed in depth prior to the creation of our models. To confirm these irregularities we constructed visual representations consisting of density plots, histograms, and box plots.

We can observe from the histograms below that our second response variable "TARGET_AMT" exhibits extreme right skewness.

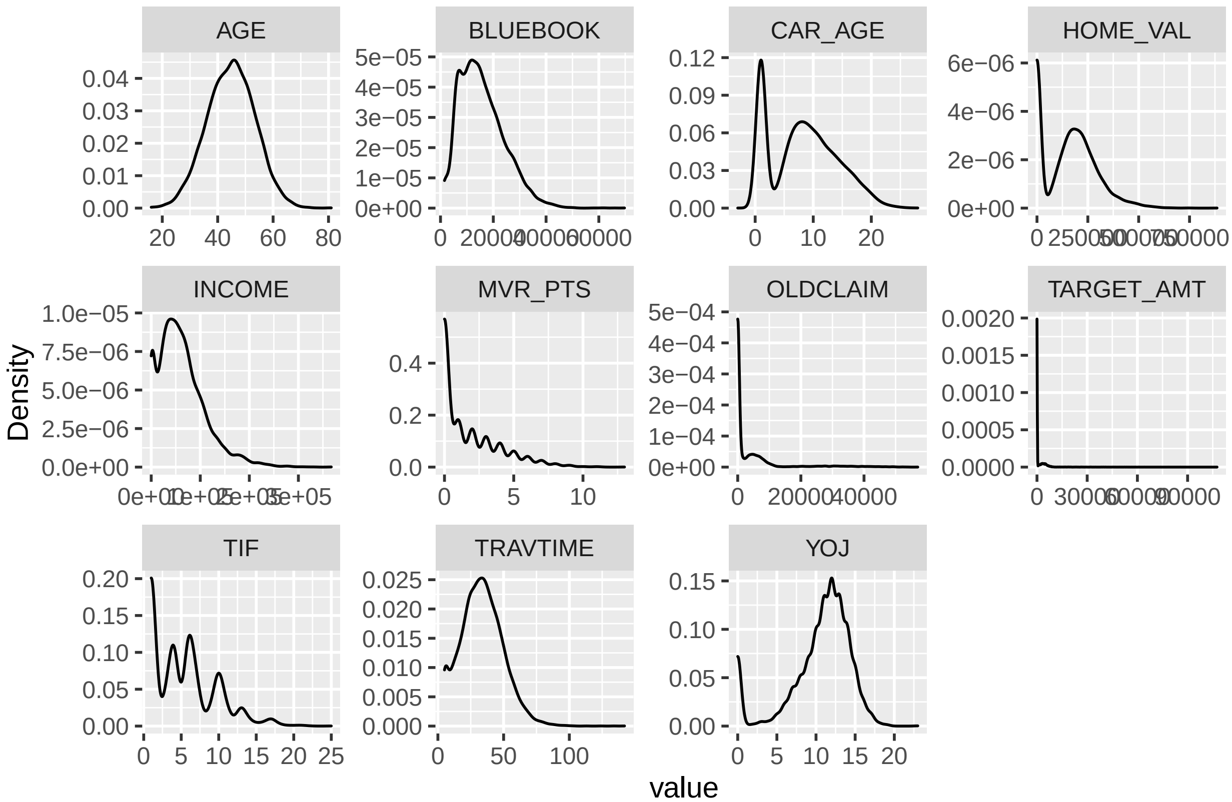


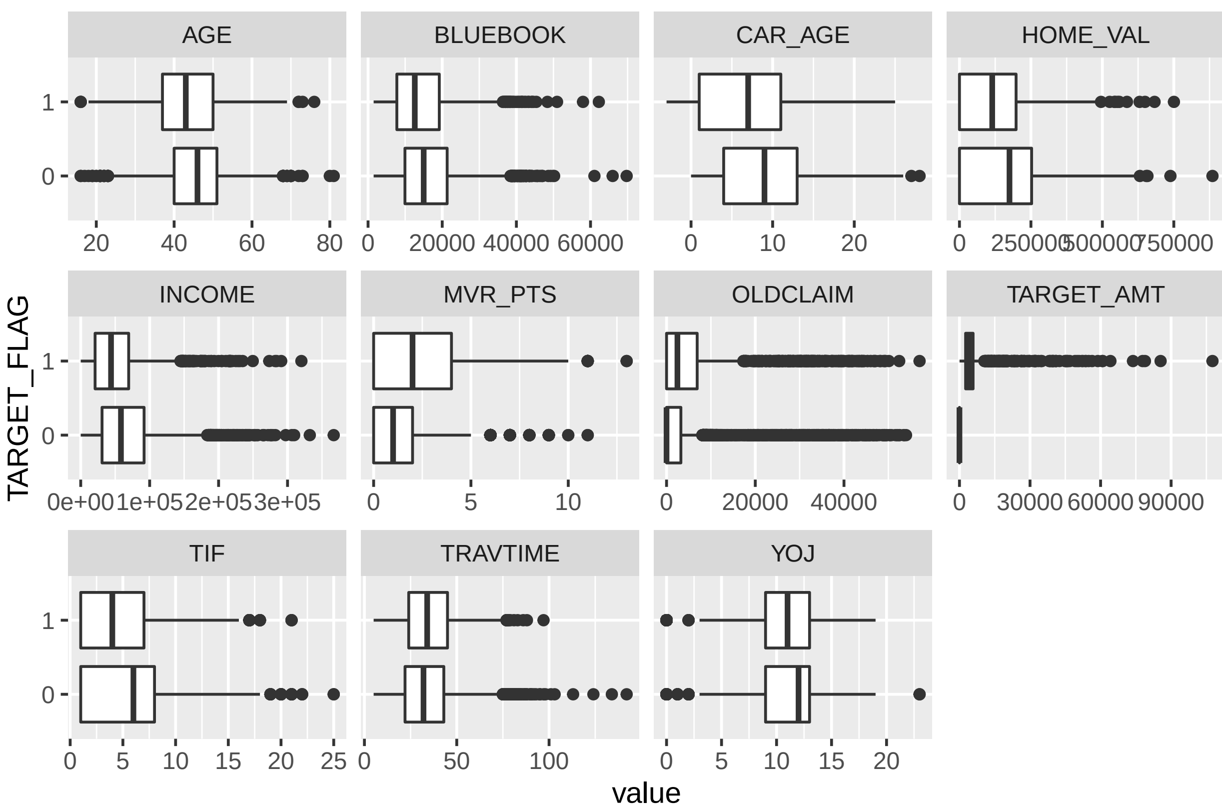

TARGET_FLAG
KIDSDRIV
HOMEKIDS
PARENT1
MSTATUS
SEX
EDUCATION
High School
Bachelors
Masters
<High School
PhD
JOB
Blue Collar
Clerical
Professional
Manager
Lawyer
Student
Home Maker
NA
Doctor
CAR_USE
Private
Commercial
Frequency

Page 1


CAR_TYPE
SUV
Minivan
Pickup
Sports Car
Van
Panel Truck
RED_CAR
no
yes
CLM_FREQ
REVOKED
No
Yes
URBANICITY
Highly Urban/ Urban
Highly Rural/ Rural
Frequency

Page 2





We can observe above that perhaps the only distribution that seems close to normal is that of the variable “AGE”, as it is also evident in the box plot against the response variable “TARGET_FLAG”. The other variable that could potentially be close to normality is “YOJ”, but it seems to have a bi-modal distribution because of the large number of people with years on the job around 0 to 1.

We were also given some theoretical effects (claims) about some of the variables in the data in regards to how they influence the response variable "TARGET_FLAG" and the probability of collision.

1.2 Analyzing Catogorical Variables

So far looking at the box plots, we can see that some of the theoretical effects tend to be more true than others, however, we are unable to see the effects of our discrete variables against the response variable "TARGET_FLAG". Below we have constructed some tables to get a sense of whether the claims about these variables tend to be true or not.

PARENT1 - Single Parent. Claim: This has an unknown effect

At a glance, we can see that those customers who are single parents have a very high proportion for being in a car crash. However, it is hard to tell if there is a correlation given the majority of the data are from customers that are "Not" single parents.

```
       TARGET_FLAG
PARENT1    0    1  Sum
    No  5407 1677 7084
    Yes  601  476 1077
    Sum 6008 2153 8161
```

```
       TARGET_FLAG
PARENT1    0    1
    No  0.76 0.24
    Yes 0.56 0.44
```

MSTATUS - Marital Status. Claim: In theory, married people drive more safely.

There seems to be a balanced split between married and not married customers in our data. We can also observe that those who were involved in a car crash are evenly split between the married and not married customers. However, the proportion of those who did not crash their car tends to be higher in the married category.

```
       TARGET_FLAG
MSTATUS    0    1  Sum
    No  2167 1100 3267
    Yes 3841 1053 4894
    Sum 6008 2153 8161
```

```
       TARGET_FLAG
MSTATUS    0    1
    No  0.66 0.34
    Yes 0.78 0.22
```

SEX - Gender. Claim: Urban legend says that women have less crashes than men. Is that true?.

There seems to be a balanced split between male and female customers in our data. Below we can also observe that the data is evenly split between males and females in regards to crashing or not crashing their cars, suggesting the claim may be flawed.

```
     TARGET_FLAG
SEX      0    1  Sum
  F   3183 1192 4375
  M   2825  961 3786
  Sum 6008 2153 8161
```

	TARGET_FLAG	
SEX	0	1
F	0.73	0.27
M	0.75	0.25

EDUCATION - Max Education Level. Claim: Unknown effect, but in theory more educated people tend to drive more safely.

Given that most of the data come from those customers with high school, bachelors and masters education, the proportions also seem to correspond among those who crashed and didn't crash their car. However, there seems to be a pattern for higher proportions of car crashes within the categories with lower education.

	TARGET_FLAG		
EDUCATION	0	1	Sum
<High School	818	385	1203
Bachelors	1719	523	2242
High School	1537	793	2330
Masters	1331	327	1658
PhD	603	125	728
Sum	6008	2153	8161

	TARGET_FLAG	
EDUCATION	0	1
<High School	0.68	0.32
Bachelors	0.77	0.23
High School	0.66	0.34
Masters	0.80	0.20
PhD	0.83	0.17

JOB - Job Category. Claim: In theory, white collar jobs tend to be safer.

We can see in the table below that blue collar jobs, students and home makers have the highest proportion of customers who have crashed their cars within their category, thus the claim may have some truth to it.

	TARGET_FLAG		
JOB	0	1	Sum
Blue Collar	1191	634	1825
Clerical	900	371	1271
Doctor	217	29	246
Home Maker	461	180	641
Lawyer	682	153	835
Manager	851	137	988
Professional	870	247	1117
Student	446	266	712
Sum	5618	2017	7635

	TARGET_FLAG	
JOB	0	1
Blue Collar	0.65	0.35
Clerical	0.71	0.29
Doctor	0.88	0.12
Home Maker	0.72	0.28
Lawyer	0.82	0.18
Manager	0.86	0.14
Professional	0.78	0.22
Student	0.63	0.37
Sum	0.74	0.26

CAR_USE - Vehicle Use. Claim: Commercial vehicles are driven more, so might increase probability of collision.

About 63% of the car usage is private, but we can see that those customers who have crashed their car has a higher percentage in the category for commercial usage, suggesting that the claim is true about the increased probability of collision for commercial vehicles.

```
            TARGET_FLAG
CAR_USE         0    1  Sum
  Commercial 1982 1047 3029
  Private    4026 1106 5132
  Sum        6008 2153 8161
```

```
            TARGET_FLAG
CAR_USE         0    1
  Commercial 0.65 0.35
  Private    0.78 0.22
```

CAR_TYPE - Type of Car. Claim: Unknown effect on probability of collision, but probably affect the payout if there is a crash.

We can see that even though sports cars is about 11% of the data we have, they have the highest proportion of car crashes within their category. We can also see that SUVs and Pickups are among the categories with the highest proportions of car crashes, while Minivans have the lowest proportion of car crashes in its category.

```
             TARGET_FLAG
CAR_TYPE         0    1  Sum
  Minivan     1796  349 2145
  Panel Truck  498  178  676
  Pickup       946  443 1389
  Sports Car   603  304  907
  SUV         1616  678 2294
  Van          549  201  750
  Sum         6008 2153 8161
```

```
             TARGET_FLAG
CAR_TYPE         0    1
  Minivan     0.84 0.16
  Panel Truck 0.74 0.26
  Pickup      0.68 0.32
  Sports Car  0.66 0.34
  SUV         0.70 0.30
  Van         0.73 0.27
```

RED_CAR - A Red Car. Claim: Urban legend says that red cars (especially red sports cars) are more risky. Is that true?.

We can observe below that roughly 25% of cars in each category were involved in a car crash, and this may disprove the claim that red cars are more risky.

```
       TARGET_FLAG
RED_CAR    0    1  Sum
    no  4246 1537 5783
    yes 1762  616 2378
    Sum 6008 2153 8161
```

```
       TARGET_FLAG
RED_CAR    0    1
    no  0.73 0.27
```

```
yes 0.74 0.26
```

REVOKED - License Revoked (Past 7 Years). Claim: If your license was revoked in the past 7 years, you probably are a more risky driver.

Although only 12% of drivers in the training data have a former license suspension on record, their proportion of being involved in a car crash is twice as high as those who didn't, suggesting the claim may be true.

```
       TARGET_FLAG
REVOKED    0    1  Sum
    No  5451 1710 7161
    Yes  557  443 1000
    Sum 6008 2153 8161
```

```
       TARGET_FLAG
REVOKED    0    1
    No  0.76 0.24
    Yes 0.56 0.44
```

URBANICITY - Home/Work Area. Claim: Unknown

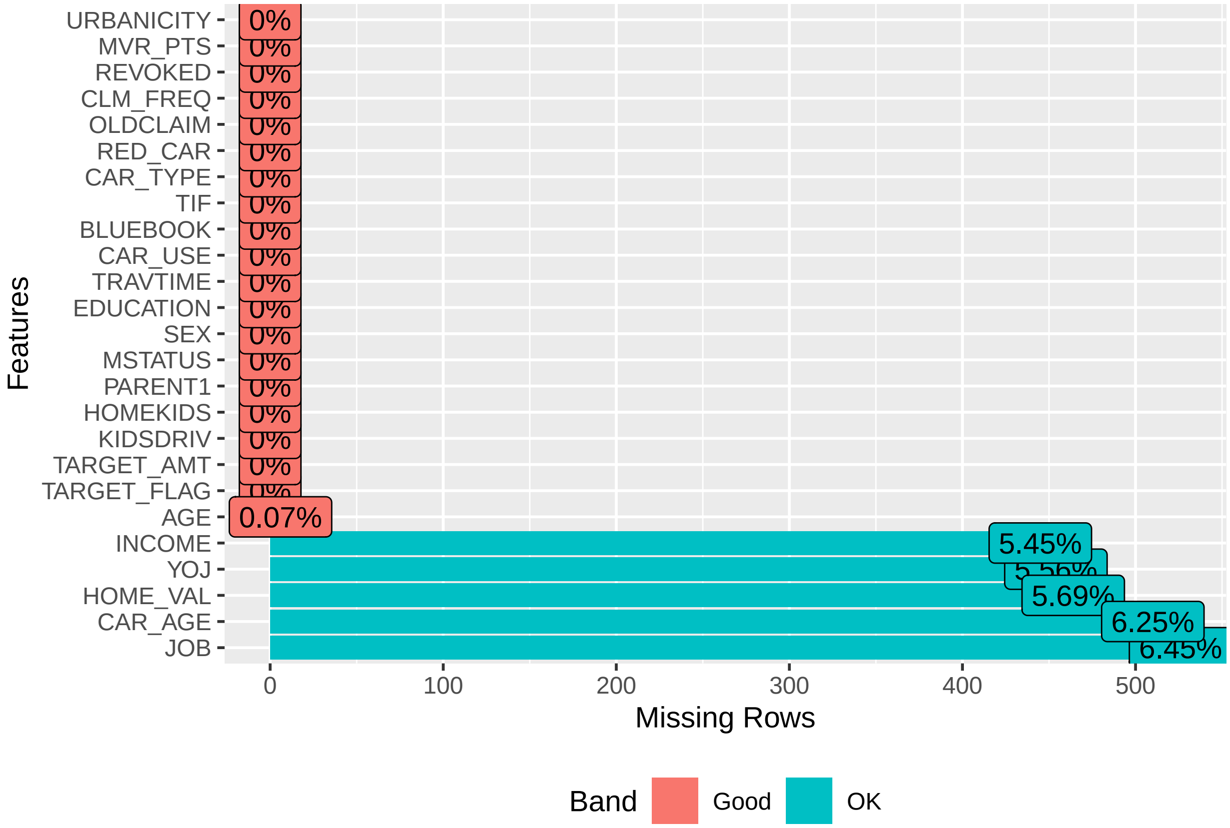
can see that the category highly urban has a higher proportion of car crashes, but this may be due to the fact that we have a lot more data from this category, roughly 80% comes from it.

```
                     TARGET_FLAG
URBANICITY               0    1  Sum
  Highly Rural/ Rural 1554  115 1669
  Highly Urban/ Urban 4454 2038 6492
  Sum                 6008 2153 8161
```

```
                     TARGET_FLAG
URBANICITY               0    1
  Highly Rural/ Rural 0.93 0.07
  Highly Urban/ Urban 0.69 0.31
```

1.3 Missing Values

Shown in our graph below, there are a few columns (variables) that have missing values. These include "AGE", "INCOME", "YOJ", "HOME_VAL", and "CAR_AGE".



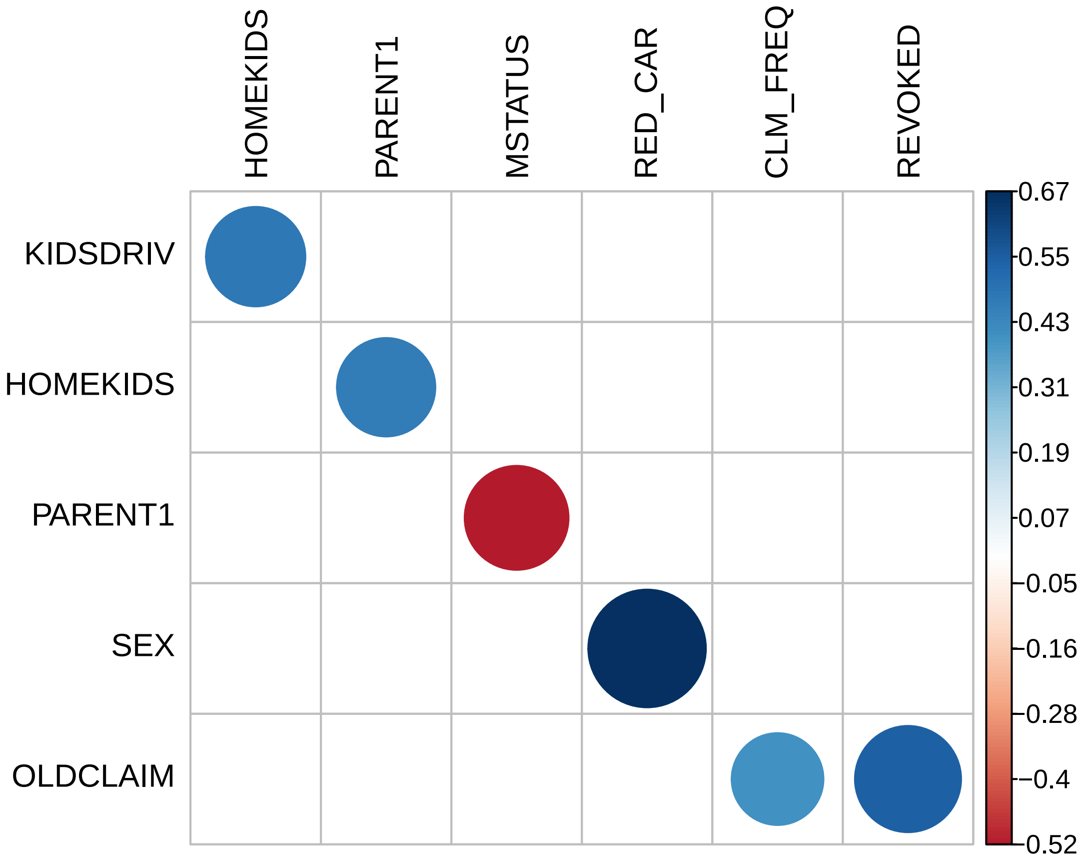
1.4 Correlation Exploration

To gain an understanding of relevant correlations we construct a function that filters our variables for correlations of interest. The result of that function can be seen below. Even though our data has a lot of variables with multiple levels, it seems that there aren't many strong correlations. Setting the minimum to .4, and we only have 7 of the many possible combination of variable correlations.

```
            Var1       Var2       Freq
461          SEX    RED_CAR  0.6666207
26   TARGET_FLAG TARGET_AMT  0.5342461
520     OLDCLAIM   CLM_FREQ  0.4951308
233      PARENT1    MSTATUS -0.4772281
103     KIDSDRIV   HOMEKIDS  0.4640152
180     HOMEKIDS    PARENT1  0.4492740
545     OLDCLAIM    REVOKED  0.4181035
```

We then apply a filter to analyze correlations where collisions occurred.

```
        Var1     Var2       Freq
461      SEX  RED_CAR  0.6678536
545 OLDCLAIM  REVOKED  0.5438075
233  PARENT1  MSTATUS -0.5212605
103 KIDSDRIV HOMEKIDS  0.4769803
180 HOMEKIDS  PARENT1  0.4673692
520 OLDCLAIM CLM_FREQ  0.4067540
```


HOMEKIDS
PARENT1
MSTATUS
RED_CAR
CLM_FREQ
REVOKED
KIDSDRIV
HOMEKIDS
PARENT1
SEX
OLDCLAIM
0.67
0.55
0.43
0.31
0.19
0.07
−0.05
−0.16
−0.28
−0.4
−0.52

We then analyze the variance inflation factors for two saturated models, one linear and on logistic. We do not find any concerning multicollinearity issues.

Table 1: Linear Model VIF Scores

	GVIF	Df	GVIF^(1/(2*Df))
KIDSDRIV	1.831636	3	1.106131
AGE	1.701204	1	1.304302
HOMEKIDS	4.310012	5	1.157305
YOJ	1.805558	1	1.343711
INCOME	3.287841	1	1.813241
PARENT1	2.931843	1	1.712262
HOME_VAL	2.198125	1	1.482607
MSTATUS	2.657837	1	1.630287
SEX	3.791626	1	1.947210
EDUCATION	11.601746	4	1.358518
JOB	29.433973	7	1.273262
TRAVTIME	1.046735	1	1.023101
CAR_USE	2.470080	1	1.571649
BLUEBOOK	2.160366	1	1.469818
TIF	1.031218	1	1.015489
CAR_TYPE	6.609987	5	1.207870
RED_CAR	1.808810	1	1.344920
OLDCLAIM	2.301839	1	1.517181
CLM_FREQ	2.010322	5	1.072325
REVOKED	1.716373	1	1.310104
MVR_PTS	1.229429	1	1.108796
CAR_AGE	2.082372	1	1.443043
URBANICITY	1.058545	1	1.028856

Table 2: Logistic Model VIF Scores

	GVIF	Df	GVIF^(1/(2*Df))
KIDSDRIV	1.611903	4	1.061494
AGE	1.608348	1	1.268206
HOMEKIDS	3.504831	5	1.133618
YOJ	1.512829	1	1.229971
INCOME	2.816302	1	1.678184
PARENT1	2.403493	1	1.550320
HOME_VAL	2.073012	1	1.439796
MSTATUS	2.362380	1	1.537004
SEX	3.554179	1	1.885253
EDUCATION	11.012664	4	1.349698
JOB	22.268379	7	1.248141
TRAVTIME	1.042313	1	1.020937
CAR_USE	2.351415	1	1.533432
BLUEBOOK	1.952979	1	1.397490
TIF	1.014362	1	1.007156
CAR_TYPE	5.433789	5	1.184432
RED_CAR	1.842568	1	1.357412
OLDCLAIM	1.920782	1	1.385923
CLM_FREQ	1.925563	5	1.067716
REVOKED	1.377369	1	1.173614
MVR_PTS	1.262024	1	1.123398
CAR_AGE	2.073252	1	1.439879
URBANICITY	1.151610	1	1.073131

2 DATA PREPARATION

We seem to have an error in one of the values for "CAR_AGE", which is -3. As we know this must be a mistake, we will turn it into a missing value.

Since the following variables with missing data ("INCOME", "YOJ", "HOME_VAL", and "CAR_AGE") are showing skewness in their distribution, we have decided to use the median as the replacement of the missing values. This will allow us to avoid any bias introduced to the mean due to the skewness itself.

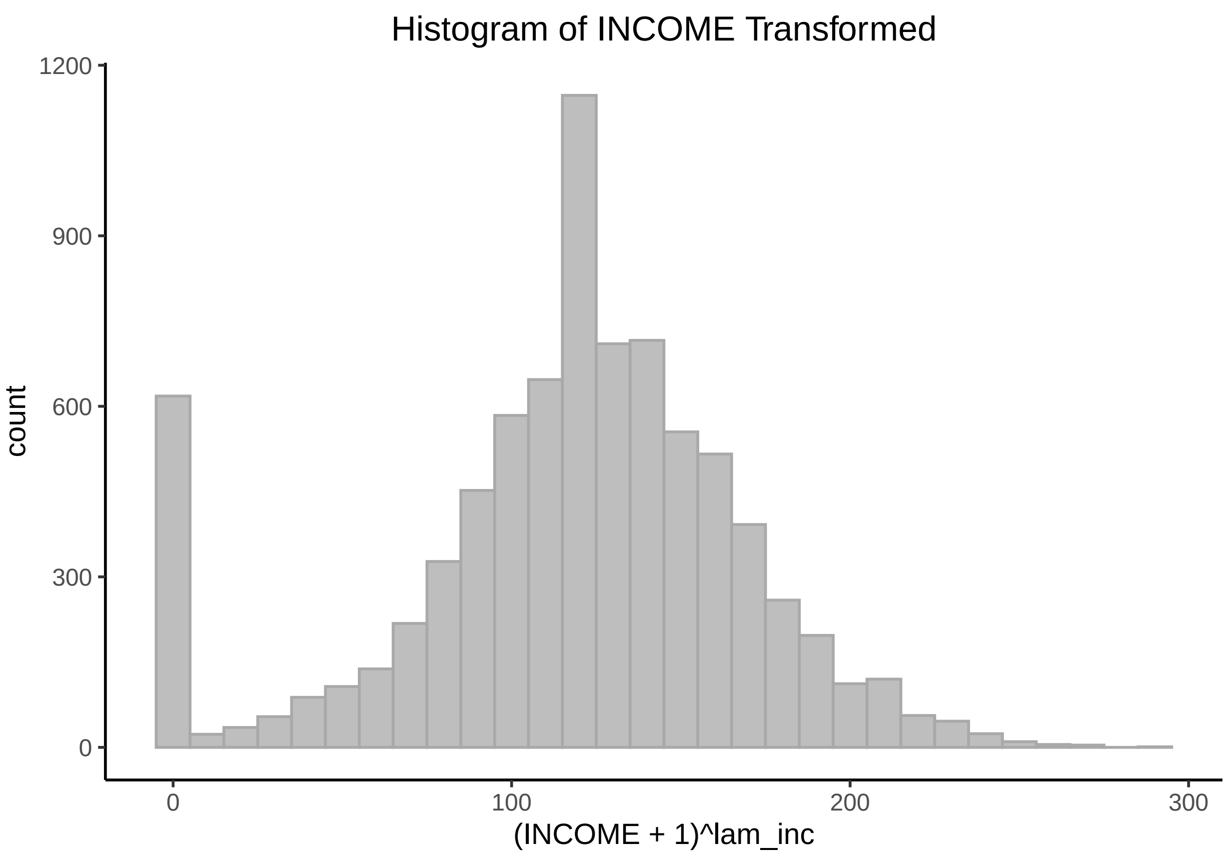
We then constructed our training sets for modeling. The linear model will only be trained on the data where an accident has occurred. We chose this approach as we do not want a model that will predict that it is possible to have no insurance cost after an accident. THe logistic model will contain all the variable, with the exception of the TARGET_AMT.

2.1 Transforming Predictors

We will next take a look at some of the variables and see what transformations may be used.

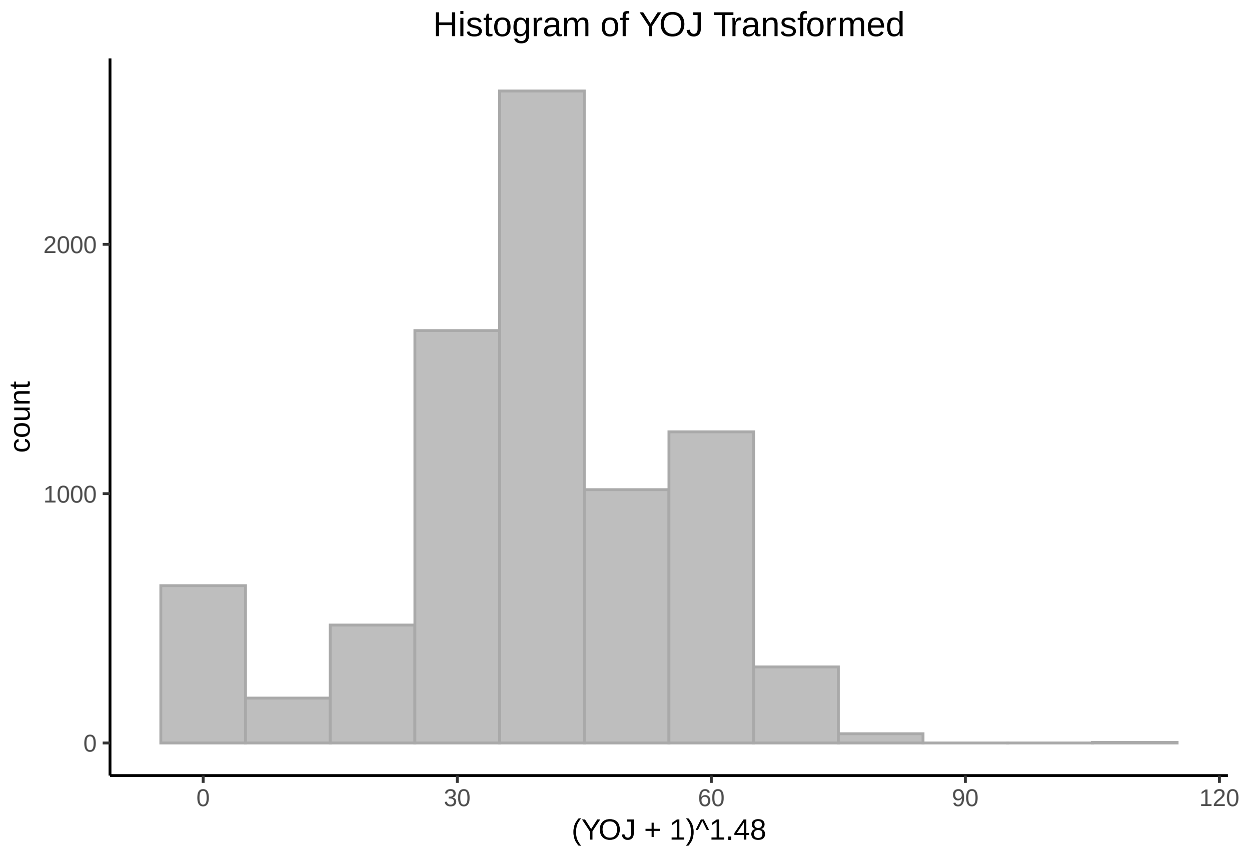
INCOME

Income is a right skewed variable with a significant number zeroes. We will apply the square root transformation suggested by the box-cox function to the original variable to reduce the overall skewness.



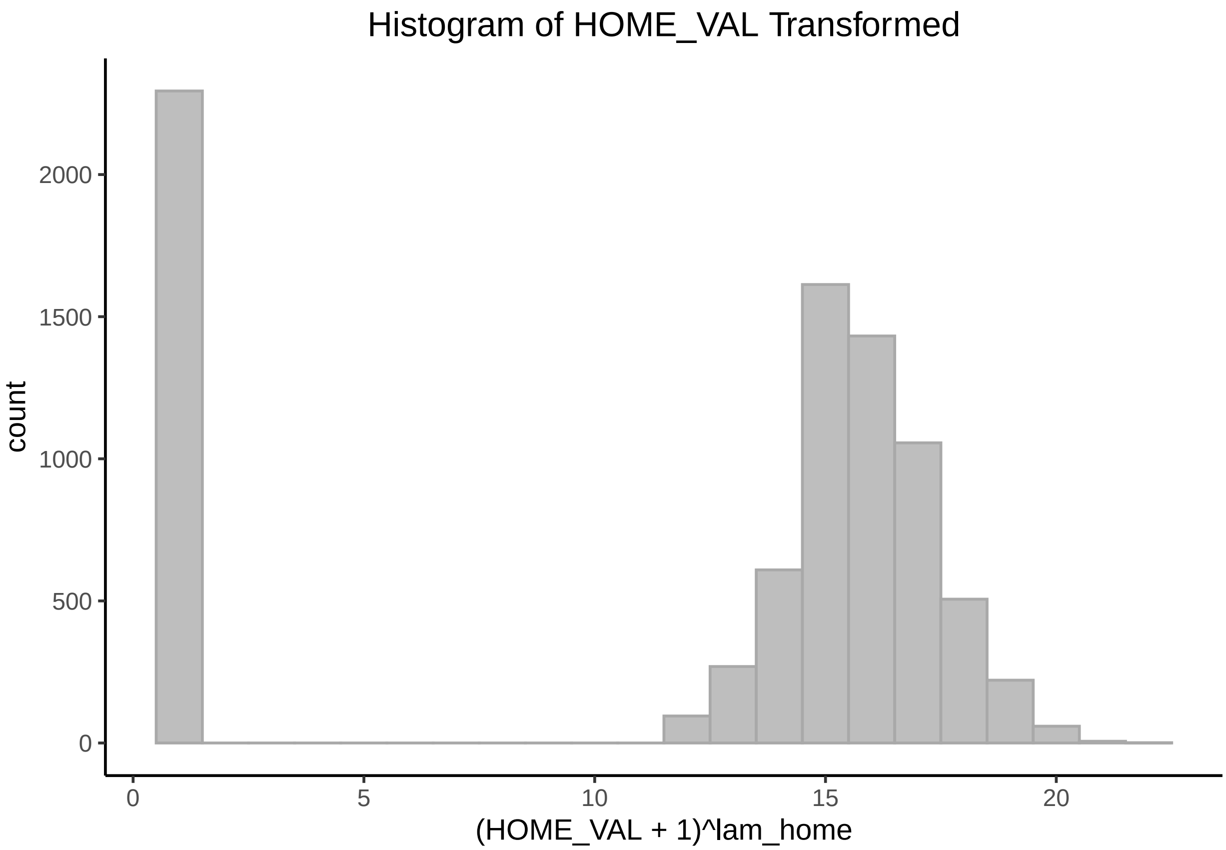
YOJ

Years on the job seems to have a bimodal distribution with a large number of customers with 0-1 years. We have applied the suggested transformation to the variable to bring it closer to normality.



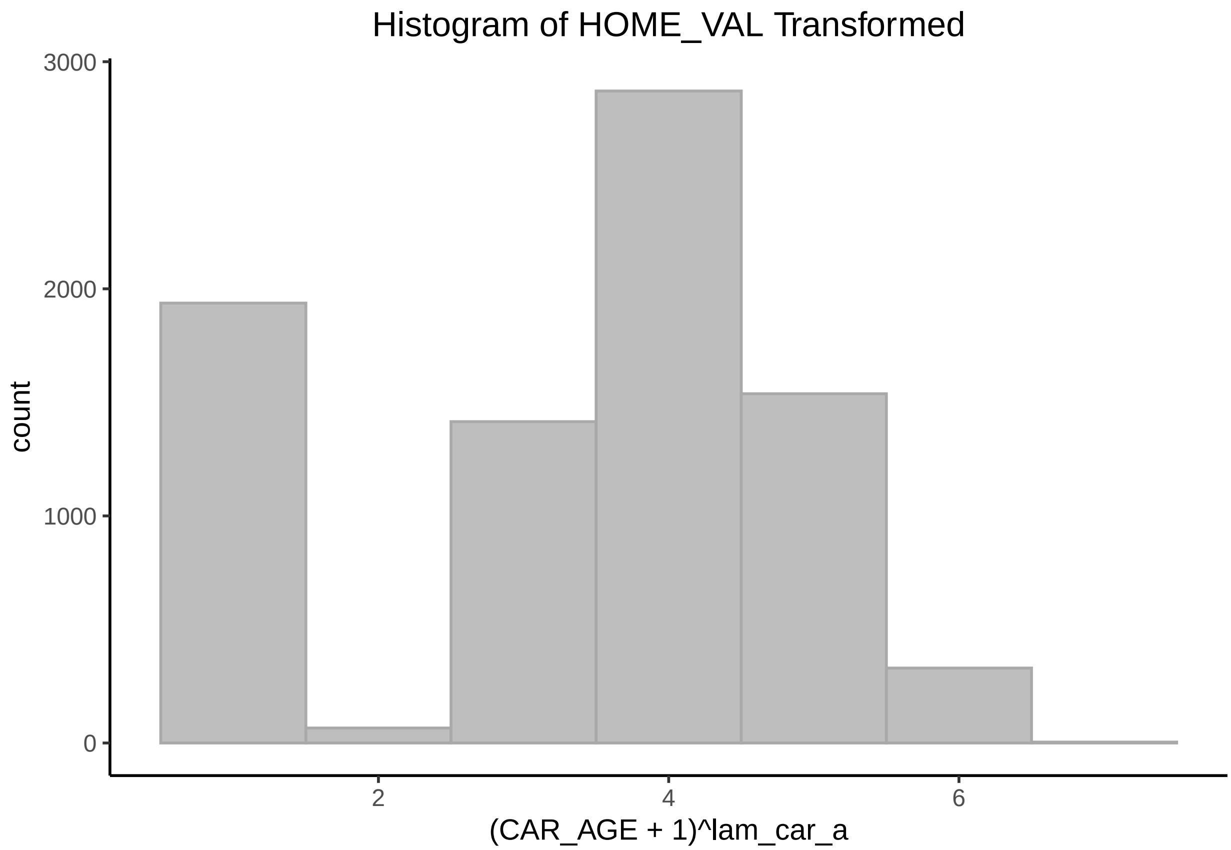
HOME_VAL

Home values are also moderately right skewed with a significant number of zeroes. We have applied the suggested transformation to this variable to reduce the overall skewness but as you can see below, it does not help much because of the significant number of 0 values in our data.



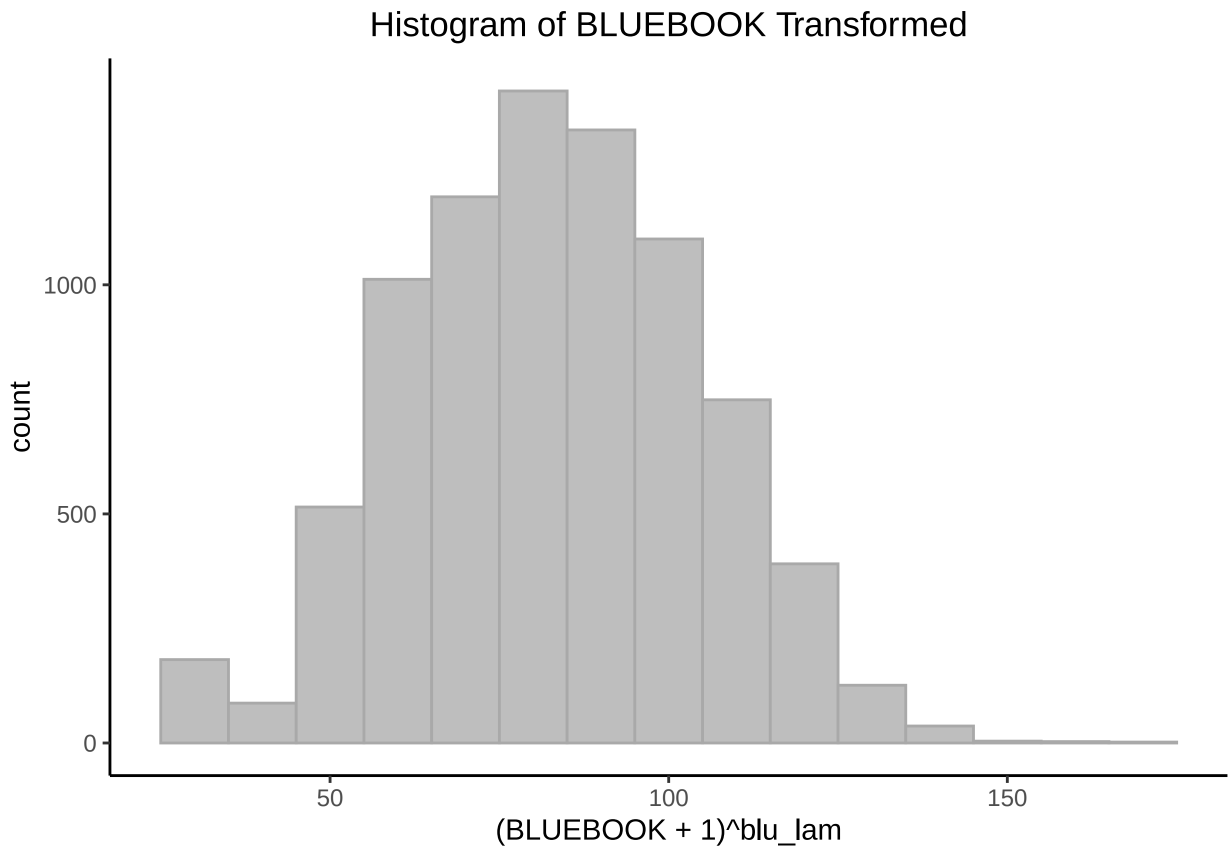
CAR_AGE

The age of the cars follow a bimodal distribution because of the significant number of cars that are close to 0 or 1 year of age. We have applied the suggested transformation, but again as we can see below, it has not helped much.



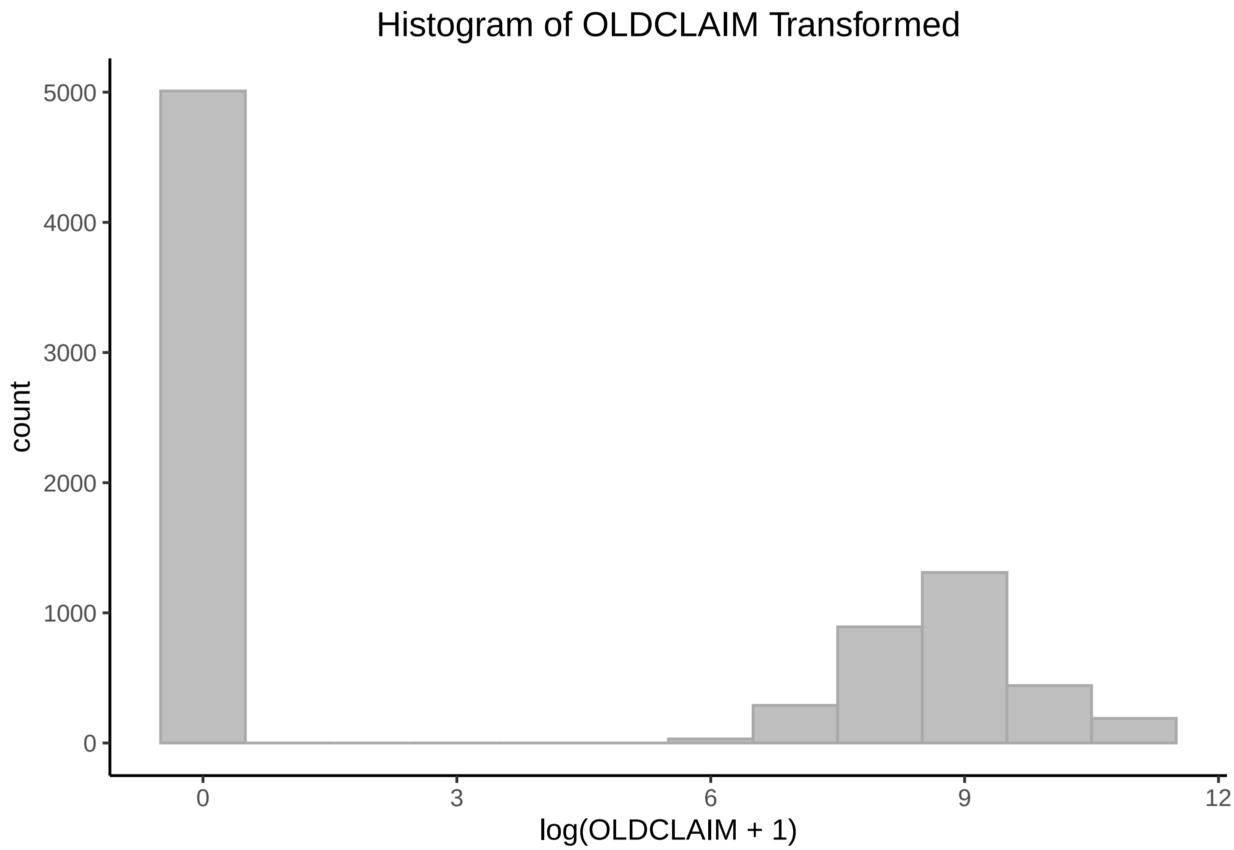
BLUEBOOK

The blue book variable is moderately right skewed. We'll apply the suggested transformation by the box-cox function.



OLDCLAIM

Old claim is has an extremely right skewed distribution. We'll apply a log transformation to reduce the overall skewness.



We then construct a training data set inclusive of our desired transformations. We used the 'dplyr' libraries mutate and across functionality to quickly and efficiently create a variety of transformations.

We can now compare the skewness of the various "TARGET_AMT" variables. It appears that the log and Standard Box-Cox transformation have the least skewing.

Table 3: Skewness Values for Target Amt Variable

Transformation	Skewness
Original	5.6346576
Square Root	2.7630888
Log	-0.0118216
Box Cox Standard	0.0017396
Alternate Box Cox	-0.9597139

3 Model Building

3.1 Building Logistic Models

We begin the model building process by creating a partition of our data to train our models. Doing so allows us to test the accuracy and performance of our constructed models.

For the first logistic model, we construct it using the untransformed data.

```
Call:
glm(formula = TARGET_FLAG ~ ., family = binomial, data = glm_additional_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5456  -0.7151  -0.3849   0.6209   3.1662

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.539e+00  3.218e-01  -7.890 3.02e-15 ***
KIDSDRIV1                      6.915e-01  1.281e-01   5.400 6.66e-08 ***
KIDSDRIV2                      7.456e-01  1.810e-01   4.121 3.78e-05 ***
KIDSDRIV3                      1.184e+00  3.591e-01   3.296 0.000982 ***
KIDSDRIV4                      1.421e+01  2.294e+02   0.062 0.950591
AGE                           -2.194e-03  4.696e-03  -0.467 0.640412
HOMEKIDS1                      2.790e-01  1.332e-01   2.094 0.036287 *
HOMEKIDS2                      1.498e-01  1.330e-01   1.127 0.259905
HOMEKIDS3                      1.762e-01  1.534e-01   1.148 0.250827
HOMEKIDS4                      1.944e-01  2.384e-01   0.815 0.414864
HOMEKIDS5                     -5.629e-02  7.607e-01  -0.074 0.941012
YOJ                           -3.929e-03  9.633e-03  -0.408 0.683364
INCOME                        -3.657e-06  1.223e-06  -2.989 0.002797 **
PARENT1Yes                     2.099e-01  1.361e-01   1.542 0.122986
HOME_VAL                      -1.174e-06  3.844e-07  -3.054 0.002260 **
MSTATUSYes                    -5.921e-01  9.731e-02  -6.085 1.16e-09 ***
SEXM                           7.166e-02  1.248e-01   0.574 0.565802
EDUCATIONBachelors            -2.802e-01  1.297e-01  -2.160 0.030772 *
EDUCATIONHigh School           6.227e-02  1.069e-01   0.582 0.560359
EDUCATIONMasters              -2.391e-01  2.021e-01  -1.183 0.236740
EDUCATIONPhD                  -2.258e-01  2.474e-01  -0.913 0.361299
JOBClerical                    4.973e-02  1.202e-01   0.414 0.679195
JOBDoctor                     -7.479e-01  3.390e-01  -2.206 0.027372 *
JOBHome Maker                 -1.136e-01  1.717e-01  -0.662 0.508285
JOBLawyer                     -1.618e-01  2.117e-01  -0.764 0.444682
JOBManager                    -8.847e-01  1.570e-01  -5.636 1.74e-08 ***
JOBProfessional               -1.616e-01  1.331e-01  -1.214 0.224736
JOBStudent                    -1.496e-01  1.453e-01  -1.029 0.303383
JOBUnknown                    -2.490e-01  2.099e-01  -1.186 0.235634
TRAVTIME                       1.499e-02  2.115e-03   7.087 1.37e-12 ***
CAR_USEPrivate                -7.570e-01  1.039e-01  -7.286 3.19e-13 ***
BLUEBOOK                      -2.696e-05  5.939e-06  -4.539 5.65e-06 ***
TIF                           -5.472e-02  8.309e-03  -6.586 4.52e-11 ***
CAR_TYPEPanel Truck            5.768e-01  1.813e-01   3.182 0.001464 **
CAR_TYPEPickup                 5.113e-01  1.130e-01   4.527 5.98e-06 ***
CAR_TYPESports Car             9.765e-01  1.449e-01   6.740 1.58e-11 ***
CAR_TYPESUV                    6.959e-01  1.238e-01   5.619 1.92e-08 ***
```

```
CAR_TYPEVan                     5.135e-01  1.454e-01   3.531 0.000414 ***
RED_CARyes                     -1.967e-02  9.717e-02  -0.202 0.839592    
OLDCLAIM                       -1.943e-05  4.687e-06  -4.146 3.38e-05 ***
CLM_FREQ1                       5.715e-01  1.112e-01   5.137 2.79e-07 ***
CLM_FREQ2                       5.878e-01  1.057e-01   5.561 2.67e-08 ***
CLM_FREQ3                       5.848e-01  1.193e-01   4.901 9.56e-07 ***
CLM_FREQ4                       7.676e-01  1.947e-01   3.942 8.07e-05 ***
CLM_FREQ5                       1.063e+00  5.518e-01   1.926 0.054073 .  
REVOKEDYes                      1.017e+00  1.040e-01   9.779  < 2e-16 ***
MVR_PTS                         1.014e-01  1.571e-02   6.455 1.08e-10 ***
CAR_AGE                        -1.728e-03  8.490e-03  -0.204 0.838716    
URBANICITYHighly Urban/ Urban  2.347e+00  1.264e-01  18.573  < 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: 7536.3  on 6529  degrees of freedom
Residual deviance: 5777.0  on 6481  degrees of freedom
AIC: 5875

Number of Fisher Scoring iterations: 11
```

We then use the "stepAIC" function from the MASS library to find the best model for our data.

```
Call:
glm(formula = TARGET_FLAG ~ KIDSDRIV + INCOME + PARENT1 + HOME_VAL +
    MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
    TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
    URBANICITY, family = binomial, data = glm_additional_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5423  -0.7164  -0.3878   0.6215   3.1415

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.614e+00  2.235e-01 -11.694  < 2e-16 ***
KIDSDRIV1                      7.702e-01  1.178e-01   6.539 6.21e-11 ***
KIDSDRIV2                      7.960e-01  1.677e-01   4.746 2.07e-06 ***
KIDSDRIV3                      1.247e+00  3.463e-01   3.602 0.000316 ***
KIDSDRIV4                      1.395e+01  2.294e+02   0.061 0.951507
INCOME                        -3.610e-06  1.216e-06  -2.968 0.002997 **
PARENT1Yes                     3.951e-01  1.069e-01   3.695 0.000220 ***
HOME_VAL                      -1.232e-06  3.832e-07  -3.216 0.001301 **
MSTATUSYes                    -5.212e-01  8.950e-02  -5.823 5.77e-09 ***
EDUCATIONBachelors            -2.923e-01  1.221e-01  -2.393 0.016691 *
EDUCATIONHigh School           6.305e-02  1.065e-01   0.592 0.553753
EDUCATIONMasters              -2.683e-01  1.820e-01  -1.474 0.140422
EDUCATIONPhD                  -2.647e-01  2.309e-01  -1.146 0.251722
JOBClerical                    5.825e-02  1.199e-01   0.486 0.627194
JOBDoctor                     -7.590e-01  3.384e-01  -2.243 0.024918 *
JOBHome Maker                 -1.153e-01  1.623e-01  -0.710 0.477681
JOBLawyer                     -1.742e-01  2.111e-01  -0.825 0.409286
JOBManager                    -8.995e-01  1.565e-01  -5.749 8.98e-09 ***
JOBProfessional               -1.664e-01  1.328e-01  -1.254 0.209993
JOBStudent                    -1.132e-01  1.386e-01  -0.817 0.413995
JOBUnknown                    -2.441e-01  2.096e-01  -1.164 0.244275
TRAVTIME                       1.481e-02  2.112e-03   7.014 2.32e-12 ***
CAR_USEPrivate                -7.514e-01  1.036e-01  -7.250 4.17e-13 ***
BLUEBOOK                      -2.902e-05  5.321e-06  -5.454 4.92e-08 ***
TIF                           -5.441e-02  8.286e-03  -6.567 5.15e-11 ***
CAR_TYPEPanel Truck            6.234e-01  1.686e-01   3.697 0.000218 ***
CAR_TYPEPickup                 5.099e-01  1.128e-01   4.522 6.12e-06 ***
CAR_TYPESports Car             9.375e-01  1.209e-01   7.752 9.06e-15 ***
CAR_TYPESUV                    6.619e-01  9.656e-02   6.855 7.12e-12 ***
CAR_TYPEVan                    5.337e-01  1.404e-01   3.801 0.000144 ***
OLDCLAIM                      -1.932e-05  4.681e-06  -4.128 3.67e-05 ***
CLM_FREQ1                      5.707e-01  1.110e-01   5.140 2.75e-07 ***
CLM_FREQ2                      5.861e-01  1.055e-01   5.557 2.75e-08 ***
CLM_FREQ3                      5.803e-01  1.192e-01   4.868 1.13e-06 ***
CLM_FREQ4                      7.539e-01  1.943e-01   3.880 0.000104 ***
CLM_FREQ5                      1.076e+00  5.511e-01   1.952 0.050979 .
REVOKEDYes                     1.018e+00  1.038e-01   9.805  < 2e-16 ***
MVR_PTS                        1.025e-01  1.566e-02   6.544 5.99e-11 ***
URBANICITYHighly Urban/ Urban  2.351e+00  1.265e-01  18.593  < 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: 7536.3  on 6529  degrees of freedom
Residual deviance: 5783.8  on 6491  degrees of freedom
AIC: 5861.8

Number of Fisher Scoring iterations: 11
```

For our second model, we construct it using the transformed variables we created.

```
Call:
glm(formula = TARGET_FLAG ~ ., family = binomial(link = "logit"),
    data = glm_additional_train_tran)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7462  -0.6985  -0.3759   0.5564   2.9832

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -6.442e+04  5.713e+04  -1.128  0.25947
KIDSDRIV1                  6.844e-01  1.359e-01   5.035 4.78e-07 ***
KIDSDRIV2                  9.799e-01  1.908e-01   5.136 2.80e-07 ***
KIDSDRIV3                  1.122e+00  3.439e-01   3.263  0.00110 **
KIDSDRIV4                  2.820e+00  1.481e+00   1.904  0.05688 .
AGE                        1.575e+05  8.164e+04   1.929  0.05374 .
HOMEKIDS1                  2.349e-02  1.418e-01   0.166  0.86842
HOMEKIDS2                 -6.693e-02  1.396e-01  -0.480  0.63153
HOMEKIDS3                 -1.133e-01  1.634e-01  -0.693  0.48833
HOMEKIDS4                 -2.533e-01  2.572e-01  -0.985  0.32467
HOMEKIDS5                  3.938e-01  7.680e-01   0.513  0.60809
YOJ                        1.203e+00  3.375e+00   0.356  0.72147
INCOME                    -8.159e-06  2.667e-05  -0.306  0.75962
PARENT1Yes                 2.404e-01  1.392e-01   1.728  0.08400 .
HOME_VAL                   4.497e-04  2.295e-04   1.960  0.05003 .
MSTATUSYes                -5.996e-01  1.035e-01  -5.793 6.91e-09 ***
SEXM                       2.703e-02  1.293e-01   0.209  0.83444
EDUCATIONBachelors        -3.761e-01  1.398e-01  -2.691  0.00712 **
EDUCATIONHigh School       1.195e-03  1.125e-01   0.011  0.99152
EDUCATIONMasters          -2.566e-01  2.100e-01  -1.222  0.22184
EDUCATIONPhD              -1.919e-01  2.467e-01  -0.778  0.43673
JOBClerical                1.031e-02  1.254e-01   0.082  0.93450
JOBDoctor                 -9.549e-01  3.390e-01  -2.817  0.00484 **
JOBHome Maker             -1.810e-01  2.054e-01  -0.881  0.37810
JOBLawyer                 -2.880e-01  2.162e-01  -1.332  0.18280
JOBManager                -8.823e-01  1.594e-01  -5.533 3.14e-08 ***
JOBProfessional           -1.455e-01  1.361e-01  -1.069  0.28510
JOBStudent                -2.565e-01  1.928e-01  -1.331  0.18335
JOBUnknown                -4.194e-01  2.118e-01  -1.980  0.04775 *
TRAVTIME                  -8.698e+00  4.686e+01  -0.186  0.85277
CAR_USEPrivate            -7.313e-01  1.047e-01  -6.983 2.88e-12 ***
BLUEBOOK                   7.037e-03  3.848e-03   1.829  0.06745 .
TIF                       -8.165e+00  9.622e+00  -0.849  0.39613
CAR_TYPEPanel Truck        6.331e-01  1.977e-01   3.203  0.00136 **
CAR_TYPEPickup             6.103e-01  1.206e-01   5.061 4.17e-07 ***
CAR_TYPESports Car         9.605e-01  1.527e-01   6.291 3.16e-10 ***
CAR_TYPESUV                7.892e-01  1.298e-01   6.080 1.20e-09 ***
CAR_TYPEVan                7.992e-01  1.455e-01   5.495 3.91e-08 ***
RED_CARyes                -2.720e-03  9.907e-02  -0.027  0.97810
OLDCLAIM                   8.876e-04  7.020e-04   1.265  0.20605
CLM_FREQ1                 -2.019e+04  1.675e+04  -1.205  0.22804
CLM_FREQ2                 -2.019e+04  1.675e+04  -1.205  0.22804
```

```
CLM_FREQ3                       -2.019e+04  1.675e+04  -1.205  0.22804
CLM_FREQ4                       -2.019e+04  1.675e+04  -1.205  0.22805
CLM_FREQ5                       -2.019e+04  1.675e+04  -1.205  0.22806
REVOKEDYes                       1.002e+00  1.088e-01   9.207  < 2e-16 ***
MVR_PTS                          5.833e+00  5.315e+00   1.097  0.27249
CAR_AGE                          5.893e+00  1.096e+01   0.538  0.59072
URBANICITYHighly Urban/ Urban    2.400e+00  1.292e-01  18.574  < 2e-16 ***
AGE_lam_one                     -3.808e+04  1.971e+04  -1.932  0.05339 .
AGE_lam_two                     -1.209e+05  6.275e+04  -1.928  0.05391 .
AGE_sqrt                         1.333e+04  7.054e+03   1.890  0.05877 .
AGE_log                         -4.472e+03  2.459e+03  -1.818  0.06905 .
YOJ_lam_one                     -2.664e-01  9.308e-01  -0.286  0.77468
YOJ_lam_two                      1.160e-01  3.574e-01   0.325  0.74547
YOJ_sqrt                        -4.924e+00  8.285e+00  -0.594  0.55231
YOJ_log                          4.647e+00  1.241e+01   0.375  0.70803
INCOME_lam_one                  -6.700e-01  1.302e+00  -0.515  0.60673
INCOME_lam_two                   6.837e-01  1.101e+00   0.621  0.53467
INCOME_sqrt                      2.242e-01  4.702e-01   0.477  0.63348
INCOME_log                      -1.518e+00  2.101e+00  -0.722  0.47007
HOME_VAL_lam_one                -9.276e+01  5.322e+01  -1.743  0.08135 .
HOME_VAL_lam_two                -3.144e+00  1.720e+00  -1.828  0.06761 .
HOME_VAL_sqrt                    1.213e+01  6.684e+00   1.814  0.06964 .
HOME_VAL_log                     3.595e+01  2.116e+01   1.699  0.08936 .
TRAVTIME_lam_one                 3.094e+02  2.372e+03   0.130  0.89622
TRAVTIME_lam_two                 1.261e+03  1.320e+04   0.095  0.92393
TRAVTIME_sqrt                   -2.840e+03  2.777e+04  -0.102  0.91854
TRAVTIME_log                     1.938e+00  9.147e+02   0.002  0.99831
BLUEBOOK_lam_one                 2.056e+02  1.116e+02   1.842  0.06551 .
BLUEBOOK_lam_two                -2.085e+02  1.131e+02  -1.843  0.06529 .
BLUEBOOK_sqrt                   -1.139e+02  6.184e+01  -1.841  0.06559 .
BLUEBOOK_log                     3.577e+02  1.944e+02   1.840  0.06575 .
TIF_lam_one                     -5.032e+03  6.434e+03  -0.782  0.43414
TIF_lam_two                     -6.982e+01  9.626e+01  -0.725  0.46824
TIF_sqrt                         2.336e+02  2.858e+02   0.817  0.41371
TIF_log                          5.563e+02  7.346e+02   0.757  0.44887
OLDCLAIM_lam_one                -2.887e+03  2.462e+03  -1.173  0.24097
OLDCLAIM_lam_two                 7.171e+02  5.943e+02   1.207  0.22756
OLDCLAIM_sqrt                   -1.131e+00  8.927e-01  -1.266  0.20534
OLDCLAIM_log                    -1.056e+03  8.790e+02  -1.201  0.22980
MVR_PTS_lam_one                 -9.353e+00  3.007e+01  -0.311  0.75575
MVR_PTS_lam_two                  3.457e+00  4.146e+00   0.834  0.40436
MVR_PTS_sqrt                    -5.716e+01  6.384e+01  -0.895  0.37060
MVR_PTS_log                      3.507e+01  5.276e+01   0.665  0.50622
CAR_AGE_lam_one                  1.909e+02  3.438e+02   0.555  0.57879
CAR_AGE_lam_two                 -2.204e+03  4.350e+03  -0.507  0.61233
CAR_AGE_sqrt                     4.280e+03  8.512e+03   0.503  0.61511
CAR_AGE_log                     -4.496e+01  1.228e+02  -0.366  0.71420
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7536.3  on 6529  degrees of freedom
Residual deviance: 5659.5  on 6441  degrees of freedom
```

```
AIC: 5837.5

Number of Fisher Scoring iterations: 9
```

We then use "stepAIC" again to find the best model, shown below.

```
Call:
glm(formula = TARGET_FLAG ~ KIDSDRIV + AGE + YOJ + INCOME + PARENT1 +
    HOME_VAL + MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE +
    BLUEBOOK + CAR_TYPE + REVOKED + URBANICITY + AGE_lam_one +
    AGE_lam_two + AGE_sqrt + YOJ_lam_one + YOJ_lam_two + YOJ_sqrt +
    INCOME_lam_two + INCOME_log + HOME_VAL_lam_one + HOME_VAL_lam_two +
    HOME_VAL_sqrt + TRAVTIME_lam_one + TRAVTIME_sqrt + BLUEBOOK_lam_one +
    BLUEBOOK_lam_two + BLUEBOOK_sqrt + BLUEBOOK_log + TIF_lam_one +
    OLDCLAIM_sqrt + OLDCLAIM_log + MVR_PTS_lam_one + MVR_PTS_lam_two +
    MVR_PTS_sqrt + MVR_PTS_log, family = binomial(link = "logit"),
    data = glm_additional_train_tran)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6654  -0.6982  -0.3823   0.5746   3.0286

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -8.752e+03  3.491e+03  -2.507 0.012174 *
KIDSDRIV1                      8.551e-01  1.221e-01   7.005 2.48e-12 ***
KIDSDRIV2                      1.108e+00  1.690e-01   6.554 5.62e-11 ***
KIDSDRIV3                      1.075e+00  3.574e-01   3.008 0.002626 **
KIDSDRIV4                      1.388e+00  1.204e+00   1.153 0.248985
AGE                            1.474e+04  6.389e+03   2.306 0.021084 *
YOJ                            2.506e+00  1.544e+00   1.622 0.104703
INCOME                        -9.166e-06  3.916e-06  -2.340 0.019262 *
PARENT1Yes                     1.719e-01  1.163e-01   1.478 0.139283
HOME_VAL                       7.157e-05  3.966e-05   1.805 0.071130 .
MSTATUSYes                    -5.897e-01  9.792e-02  -6.023 1.72e-09 ***
EDUCATIONBachelors            -3.320e-01  1.273e-01  -2.609 0.009089 **
EDUCATIONHigh School           3.785e-02  1.097e-01   0.345 0.730122
EDUCATIONMasters              -3.445e-01  1.876e-01  -1.837 0.066270 .
EDUCATIONPhD                  -3.228e-01  2.308e-01  -1.399 0.161919
JOBClerical                    2.392e-02  1.230e-01   0.194 0.845802
JOBDoctor                     -8.747e-01  3.318e-01  -2.636 0.008387 **
JOBHome Maker                 -4.731e-01  2.061e-01  -2.296 0.021692 *
JOBLawyer                     -8.861e-02  2.135e-01  -0.415 0.678154
JOBManager                    -9.280e-01  1.586e-01  -5.850 4.93e-09 ***
JOBProfessional               -1.834e-01  1.360e-01  -1.348 0.177552
JOBStudent                    -4.729e-01  1.867e-01  -2.532 0.011337 *
JOBUnknown                    -3.156e-01  2.127e-01  -1.484 0.137798
TRAVTIME                      -2.179e-01  1.251e-01  -1.742 0.081544 .
CAR_USEPrivate                -7.065e-01  1.042e-01  -6.782 1.18e-11 ***
BLUEBOOK                       7.059e-03  3.945e-03   1.789 0.073562 .
CAR_TYPEPanel Truck            6.607e-01  1.791e-01   3.689 0.000225 ***
CAR_TYPEPickup                 6.431e-01  1.190e-01   5.405 6.46e-08 ***
CAR_TYPESports Car             8.785e-01  1.249e-01   7.036 1.99e-12 ***
CAR_TYPESUV                    7.858e-01  9.913e-02   7.927 2.24e-15 ***
CAR_TYPEVan                    7.468e-01  1.375e-01   5.432 5.56e-08 ***
REVOKEDYes                     8.630e-01  1.062e-01   8.123 4.55e-16 ***
URBANICITYHighly Urban/ Urban  2.385e+00  1.282e-01  18.602  < 2e-16 ***
AGE_lam_one                   -3.671e+03  1.595e+03  -2.301 0.021400 *
```

```
AGE_lam_two                -1.115e+04  4.829e+03  -2.309 0.020937 *
AGE_sqrt                    7.811e+02  3.335e+02   2.342 0.019173 *
YOJ_lam_one                -5.759e-01  3.846e-01  -1.498 0.134259
YOJ_lam_two                 2.201e-01  1.550e-01   1.420 0.155617
YOJ_sqrt                   -2.340e+00  1.359e+00  -1.722 0.085025 .
INCOME_lam_two              2.652e-02  1.879e-02   1.412 0.158030
INCOME_log                 -1.891e-01  1.073e-01  -1.762 0.078144 .
HOME_VAL_lam_one           -2.439e+00  1.715e+00  -1.422 0.155136
HOME_VAL_lam_two           -2.418e-01  1.535e-01  -1.575 0.115321
HOME_VAL_sqrt               8.353e-01  5.391e-01   1.549 0.121297
TRAVTIME_lam_one            2.537e+00  1.403e+00   1.808 0.070593 .
TRAVTIME_sqrt              -3.600e+00  2.047e+00  -1.758 0.078674 .
BLUEBOOK_lam_one            2.112e+02  1.139e+02   1.854 0.063770 .
BLUEBOOK_lam_two           -2.171e+02  1.153e+02  -1.884 0.059564 .
BLUEBOOK_sqrt              -1.167e+02  6.312e+01  -1.849 0.064442 .
BLUEBOOK_log                3.768e+02  1.978e+02   1.904 0.056865 .
TIF_lam_one                -2.273e+00  3.366e-01  -6.753 1.45e-11 ***
OLDCLAIM_sqrt              -6.900e-03  1.476e-03  -4.674 2.95e-06 ***
OLDCLAIM_log                1.190e-01  1.835e-02   6.485 8.89e-11 ***
MVR_PTS_lam_one            -6.241e+01  1.789e+01  -3.489 0.000485 ***
MVR_PTS_lam_two            -1.775e+00  4.877e-01  -3.639 0.000273 ***
MVR_PTS_sqrt                1.974e+01  5.091e+00   3.877 0.000106 ***
MVR_PTS_log                -3.587e+01  9.697e+00  -3.699 0.000217 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7536.3  on 6529  degrees of freedom
Residual deviance: 5706.0  on 6473  degrees of freedom
AIC: 5820

Number of Fisher Scoring iterations: 5
```

3.2 Building Linear Models

Now that the binary logistic regression model is constructed we can proceed to our linear models. We do the same process as before, partitioning our data and then beginning with a simple linear model with no transformations.

```
Call:
lm(formula = TARGET_AMT ~ ., data = lm_additional_train)

Residuals:
   Min     1Q Median     3Q    Max
 -9355  -3098  -1351    531  77783

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.626e+03  1.730e+03   2.096   0.0362 *
KIDSDRIV1                      1.427e+02  6.666e+02   0.214   0.8305
KIDSDRIV2                     -2.934e+02  9.104e+02  -0.322   0.7473
KIDSDRIV3                     -1.064e+03  1.528e+03  -0.696   0.4864
KIDSDRIV4                     -4.490e+03  8.293e+03  -0.541   0.5883
AGE                            1.453e+01  2.322e+01   0.626   0.5317
HOMEKIDS1                      2.351e+02  7.281e+02   0.323   0.7468
HOMEKIDS2                      1.098e+03  7.184e+02   1.528   0.1266
HOMEKIDS3                      6.399e+02  8.200e+02   0.780   0.4353
HOMEKIDS4                      4.063e+02  1.245e+03   0.326   0.7441
HOMEKIDS5                      1.661e+03  3.749e+03   0.443   0.6578
YOJ                            4.337e+00  5.210e+01   0.083   0.9337
INCOME                        -9.518e-03  7.128e-03  -1.335   0.1820
PARENT1Yes                     1.252e+02  7.142e+02   0.175   0.8608
HOME_VAL                       1.870e-03  2.157e-03   0.867   0.3860
MSTATUSYes                    -9.174e+02  5.535e+02  -1.657   0.0977 .
SEXM                           1.691e+03  7.062e+02   2.394   0.0168 *
EDUCATIONBachelors            -3.534e+02  6.837e+02  -0.517   0.6053
EDUCATIONHigh School          -6.709e+02  5.459e+02  -1.229   0.2192
EDUCATIONMasters               7.187e+02  1.185e+03   0.607   0.5442
EDUCATIONPhD                   1.689e+03  1.395e+03   1.210   0.2263
JOBClerical                   -6.187e+02  6.241e+02  -0.991   0.3217
JOBDoctor                     -2.012e+03  1.947e+03  -1.033   0.3016
JOBHome Maker                 -5.567e+02  9.210e+02  -0.604   0.5456
JOBLawyer                      1.523e+02  1.279e+03   0.119   0.9052
JOBManager                    -7.509e+02  1.021e+03  -0.736   0.4620
JOBProfessional                3.876e+02  7.268e+02   0.533   0.5939
JOBStudent                    -3.617e+02  7.468e+02  -0.484   0.6282
JOBUnknown                     6.725e+02  1.247e+03   0.539   0.5896
TRAVTIME                       1.841e+00  1.183e+01   0.156   0.8764
CAR_USEPrivate                -1.868e+02  5.558e+02  -0.336   0.7369
BLUEBOOK                       1.303e-01  3.215e-02   4.054 5.27e-05 ***
TIF                            3.778e+01  4.497e+01   0.840   0.4010
CAR_TYPEPanel Truck           -9.169e+02  1.019e+03  -0.900   0.3684
CAR_TYPEPickup                -5.699e+01  6.302e+02  -0.090   0.9280
CAR_TYPESports Car             9.760e+02  7.982e+02   1.223   0.2215
CAR_TYPESUV                    9.225e+02  7.080e+02   1.303   0.1927
CAR_TYPEVan                   -1.515e+03  8.077e+02  -1.876   0.0608 .
RED_CARyes                    -6.250e+02  5.380e+02  -1.162   0.2455
OLDCLAIM                       2.907e-02  2.575e-02   1.129   0.2591
```

```
CLM_FREQ1                      1.142e+02  5.944e+02   0.192   0.8477
CLM_FREQ2                     -2.010e+02  5.577e+02  -0.360   0.7186
CLM_FREQ3                     -3.393e+02  6.257e+02  -0.542   0.5877
CLM_FREQ4                     -1.956e+02  1.031e+03  -0.190   0.8495
CLM_FREQ5                     -1.714e+02  3.338e+03  -0.051   0.9591
REVOKEDYes                    -1.068e+03  5.598e+02  -1.907   0.0566 .
MVR_PTS                        5.679e+01  7.523e+01   0.755   0.4504
CAR_AGE                       -7.469e+01  4.724e+01  -1.581   0.1141
URBANICITYHighly Urban/ Urban -2.792e+02  7.829e+02  -0.357   0.7214
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7280 on 1676 degrees of freedom
Multiple R-squared:  0.03388,   Adjusted R-squared:  0.006208
F-statistic: 1.224 on 48 and 1676 DF,  p-value: 0.1415
```

We then use the "stepAIC" function again to find the best model.

```
Call:
lm(formula = TARGET_AMT ~ PARENT1 + BLUEBOOK, data = lm_additional_train)

Residuals:
   Min     1Q Median     3Q    Max
 -8021  -3021  -1493    434  79234

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.977e+03  3.646e+02  10.909  < 2e-16 ***
PARENT1Yes  7.786e+02  4.190e+02   1.858   0.0633 .
BLUEBOOK    1.013e-01  2.112e-02   4.796 1.76e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7253 on 1722 degrees of freedom
Multiple R-squared:  0.01473,   Adjusted R-squared:  0.01359
F-statistic: 12.87 on 2 and 1722 DF,  p-value: 2.818e-06
```

In the second model, we use the transformed “TARGET_AMT” variable “TARGET_AMT_lam_one” and the transformed predictors to see if we can find a better fitting model.

```
Call:
lm(formula = TARGET_AMT_lam_one ~ . - TARGET_AMT - TARGET_AMT_lam_two -
    TARGET_AMT_sqrt - TARGET_AMT_log, data = lm_additional_train_trans_lam)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0107583 -0.0009220  0.0000684  0.0009867  0.0071792

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -2.311e+02  2.691e+02  -0.859   0.3905
KIDSDRIV1                      6.022e-05  1.843e-04   0.327   0.7439
KIDSDRIV2                     -7.969e-05  2.471e-04  -0.323   0.7471
KIDSDRIV3                     -6.919e-04  4.072e-04  -1.699   0.0894 .
KIDSDRIV4                     -2.148e-03  2.177e-03  -0.986   0.3241
AGE                            2.656e-01  1.299e+00   0.204   0.8381
HOMEKIDS1                      9.500e-05  1.989e-04   0.478   0.6330
HOMEKIDS2                      1.828e-04  1.964e-04   0.931   0.3521
HOMEKIDS3                      3.094e-04  2.301e-04   1.344   0.1790
HOMEKIDS4                      2.587e-04  3.414e-04   0.758   0.4486
HOMEKIDS5                      9.044e-04  1.001e-03   0.903   0.3666
YOJ                           -8.295e-03  8.192e-03  -1.013   0.3114
INCOME                         1.010e-08  3.519e-08   0.287   0.7741
PARENT1Yes                     4.316e-05  1.892e-04   0.228   0.8195
HOME_VAL                      -2.814e-07  2.500e-07  -1.126   0.2604
MSTATUSYes                    -2.865e-04  1.497e-04  -1.914   0.0558 .
SEXM                           1.180e-04  1.895e-04   0.623   0.5335
EDUCATIONBachelors            -1.400e-04  1.912e-04  -0.732   0.4644
EDUCATIONHigh School          -3.113e-05  1.487e-04  -0.209   0.8342
EDUCATIONMasters               4.117e-04  3.257e-04   1.264   0.2064
EDUCATIONPhD                   5.928e-04  3.746e-04   1.583   0.1137
JOBClerical                   -3.820e-09  1.708e-04   0.000   1.0000
JOBDoctor                      1.014e-04  5.194e-04   0.195   0.8453
JOBHome Maker                  1.188e-04  2.925e-04   0.406   0.6846
JOBLawyer                      7.033e-05  3.401e-04   0.207   0.8362
JOBManager                     1.629e-04  2.706e-04   0.602   0.5472
JOBProfessional                1.860e-04  1.915e-04   0.971   0.3315
JOBStudent                     4.036e-04  2.614e-04   1.544   0.1228
JOBUnknown                     9.464e-05  3.314e-04   0.286   0.7752
TRAVTIME                       2.203e-02  6.221e-02   0.354   0.7233
CAR_USEPrivate                -3.199e-05  1.473e-04  -0.217   0.8280
BLUEBOOK                      -1.812e-06  4.016e-06  -0.451   0.6518
TIF                            1.272e-02  1.155e-02   1.101   0.2709
CAR_TYPEPanel Truck            8.484e-05  2.877e-04   0.295   0.7681
CAR_TYPEPickup                -8.985e-06  1.716e-04  -0.052   0.9583
CAR_TYPESports Car             1.990e-05  2.136e-04   0.093   0.9258
CAR_TYPESUV                    5.898e-05  1.903e-04   0.310   0.7566
CAR_TYPEVan                   -2.122e-04  2.140e-04  -0.992   0.3215
RED_CARyes                    -1.325e-05  1.423e-04  -0.093   0.9258
OLDCLAIM                      -1.423e-07  3.075e-07  -0.463   0.6436
CLM_FREQ1                      1.409e+02  1.237e+02   1.138   0.2552
```

```
CLM_FREQ2                        1.409e+02  1.237e+02   1.138   0.2552
CLM_FREQ3                        1.409e+02  1.237e+02   1.138   0.2552
CLM_FREQ4                        1.409e+02  1.237e+02   1.138   0.2552
CLM_FREQ5                        1.409e+02  1.237e+02   1.138   0.2552
REVOKEDYes                      -2.083e-04  1.549e-04  -1.345   0.1788
MVR_PTS                         -7.532e-03  5.450e-03  -1.382   0.1672
CAR_AGE                          5.063e-03  1.659e-02   0.305   0.7603
URBANICITYHighly Urban/ Urban    1.086e-04  2.078e-04   0.523   0.6012
AGE_lam_one                     -3.601e+00  1.821e+01  -0.198   0.8432
AGE_lam_two                      1.362e+01  7.102e+01   0.192   0.8480
AGE_sqrt                        -2.545e+01  1.346e+02  -0.189   0.8501
AGE_log                          2.023e+00  1.188e+01   0.170   0.8648
YOJ_lam_one                      2.809e-03  2.771e-03   1.014   0.3108
YOJ_lam_two                     -5.004e-04  4.891e-04  -1.023   0.3064
YOJ_sqrt                        -1.740e-02  1.912e-02  -0.910   0.3629
YOJ_log                          2.755e-02  2.871e-02   0.960   0.3373
INCOME_lam_one                  -6.976e-04  3.585e-03  -0.195   0.8457
INCOME_lam_two                   7.391e-04  2.853e-03   0.259   0.7956
INCOME_sqrt                      6.021e-05  4.892e-04   0.123   0.9020
INCOME_log                      -1.481e-03  4.740e-03  -0.312   0.7547
HOME_VAL_lam_one                 1.169e+00  8.829e-01   1.324   0.1857
HOME_VAL_lam_two                -9.567e-03  7.577e-03  -1.263   0.2069
HOME_VAL_sqrt                    4.207e-03  3.396e-03   1.239   0.2156
HOME_VAL_log                    -1.512e-01  1.127e-01  -1.341   0.1800
TRAVTIME_lam_one                -3.313e-01  1.084e+00  -0.306   0.7600
TRAVTIME_lam_two                 4.474e-01  1.607e+00   0.278   0.7807
TRAVTIME_sqrt                   -4.723e-01  2.008e+00  -0.235   0.8141
TRAVTIME_log                    -1.843e-02  1.235e-01  -0.149   0.8814
BLUEBOOK_lam_one                -6.909e-02  1.266e-01  -0.546   0.5853
BLUEBOOK_lam_two                 5.038e-01  8.625e-01   0.584   0.5592
BLUEBOOK_sqrt                    9.402e-03  1.784e-02   0.527   0.5983
BLUEBOOK_log                    -6.221e-01  1.053e+00  -0.591   0.5546
TIF_lam_one                      4.260e+01  3.822e+01   1.115   0.2652
TIF_lam_two                      1.107e-01  9.950e-02   1.112   0.2662
TIF_sqrt                        -3.140e-01  2.834e-01  -1.108   0.2680
TIF_log                          2.362e+00  2.120e+00   1.114   0.2654
OLDCLAIM_lam_one                 1.004e+02  8.819e+01   1.139   0.2550
OLDCLAIM_lam_two                -8.425e+00  7.402e+00  -1.138   0.2552
OLDCLAIM_sqrt                   -4.295e-04  4.315e-04  -0.995   0.3197
OLDCLAIM_log                     1.729e+00  1.523e+00   1.135   0.2564
MVR_PTS_lam_one                 -2.047e+01  1.345e+01  -1.522   0.1283
MVR_PTS_lam_two                 -1.223e-02  8.420e-03  -1.453   0.1464
MVR_PTS_sqrt                     1.145e-01  7.881e-02   1.453   0.1463
MVR_PTS_log                     -5.960e-01  3.912e-01  -1.524   0.1278
CAR_AGE_lam_one                  1.571e-01  5.470e-01   0.287   0.7740
CAR_AGE_lam_two                  6.119e-01  2.033e+00   0.301   0.7635
CAR_AGE_sqrt                    -1.254e+00  4.180e+00  -0.300   0.7641
CAR_AGE_log                     -6.810e-02  2.308e-01  -0.295   0.7680
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001895 on 1636 degrees of freedom
Multiple R-squared:  0.05451,   Adjusted R-squared:  0.003658
F-statistic: 1.072 on 88 and 1636 DF,  p-value: 0.3087
```

Performing "stepAIC" we get this for our final linear model.

```
Call:
lm(formula = TARGET_AMT_lam_one ~ MSTATUS + EDUCATION + TRAVTIME +
    REVOKED + AGE_lam_two + AGE_sqrt + HOME_VAL_lam_one + HOME_VAL_lam_two +
    HOME_VAL_sqrt + HOME_VAL_log + TRAVTIME_lam_one + TRAVTIME_lam_two +
    TRAVTIME_sqrt + BLUEBOOK_lam_one + BLUEBOOK_lam_two + OLDCLAIM_lam_one +
    OLDCLAIM_log + MVR_PTS_sqrt, data = lm_additional_train_trans_lam)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0109451 -0.0009487  0.0001036  0.0009616  0.0072188

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.267e+00  9.993e-01   2.268   0.0235 *
MSTATUSYes           -2.360e-04  1.125e-04  -2.097   0.0361 *
EDUCATIONBachelors   -1.191e-04  1.461e-04  -0.815   0.4150
EDUCATIONHigh School -1.276e-05  1.314e-04  -0.097   0.9226
EDUCATIONMasters      2.628e-04  1.721e-04   1.527   0.1268
EDUCATIONPhD          3.739e-04  2.360e-04   1.584   0.1133
TRAVTIME              2.840e-02  1.958e-02   1.450   0.1471
REVOKEDYes           -1.968e-04  1.380e-04  -1.426   0.1539
AGE_lam_two           2.714e-03  1.628e-03   1.667   0.0957 .
AGE_sqrt             -7.491e-03  4.486e-03  -1.670   0.0951 .
HOME_VAL_lam_one      1.979e-01  1.217e-01   1.627   0.1040
HOME_VAL_lam_two     -1.133e-03  7.258e-04  -1.561   0.1187
HOME_VAL_sqrt         4.126e-04  2.695e-04   1.531   0.1259
HOME_VAL_log         -2.773e-02  1.689e-02  -1.642   0.1008
TRAVTIME_lam_one     -4.481e-01  3.082e-01  -1.454   0.1462
TRAVTIME_lam_two      6.252e-01  4.296e-01   1.455   0.1458
TRAVTIME_sqrt        -7.034e-01  4.828e-01  -1.457   0.1453
BLUEBOOK_lam_one     -6.949e-05  4.556e-05  -1.525   0.1274
BLUEBOOK_lam_two      6.034e-04  2.745e-04   2.198   0.0281 *
OLDCLAIM_lam_one      3.933e-03  2.371e-03   1.659   0.0974 .
OLDCLAIM_log         -3.730e-04  2.250e-04  -1.658   0.0976 .
MVR_PTS_sqrt          9.161e-05  4.975e-05   1.842   0.0657 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001879 on 1703 degrees of freedom
Multiple R-squared:  0.03299,   Adjusted R-squared:  0.02107
F-statistic: 2.767 on 21 and 1703 DF,  p-value: 3.078e-05
```

4 Model Selection

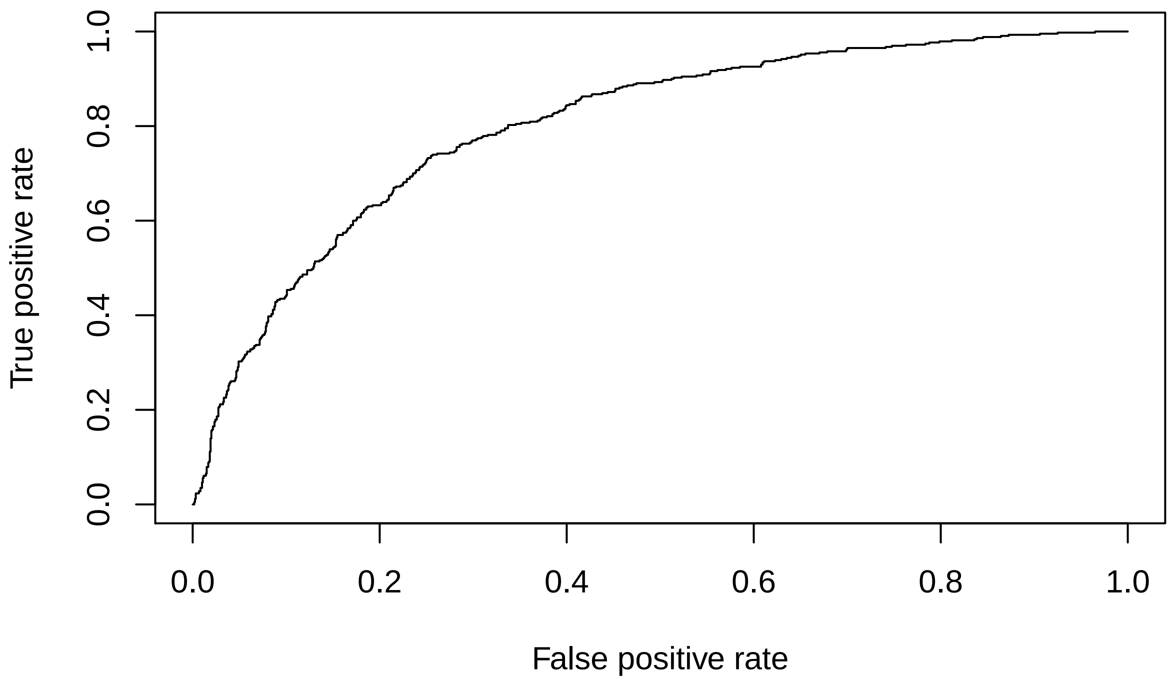
4.1 Binary Model Selecion

We can predict results to discern performance metrics.

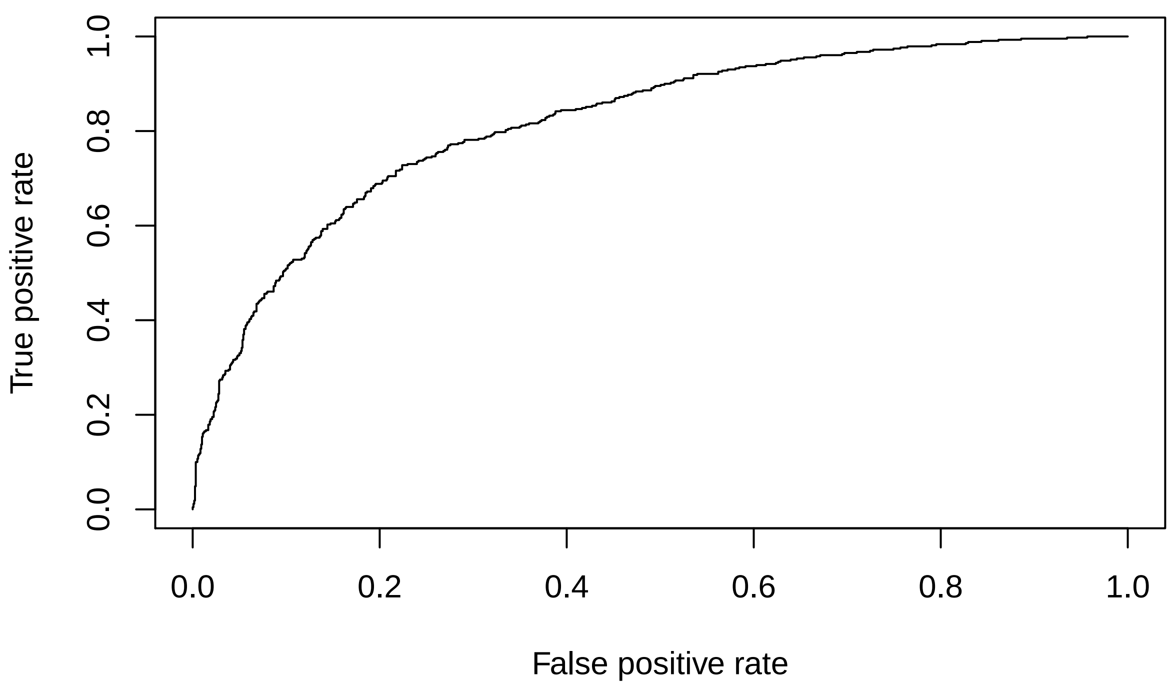
In selecting the best model, first we need to measure performance of the models prior to selection. We can do so by looking at the confusion matrix and AUC curve for our models. For the first model we have:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1102  259
         1   99  171

               Accuracy : 0.7805
                 95% CI : (0.7596, 0.8004)
    No Information Rate : 0.7364
    P-Value [Acc > NIR] : 2.138e-05

                  Kappa : 0.358

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9176
            Specificity : 0.3977
         Pos Pred Value : 0.8097
         Neg Pred Value : 0.6333
             Prevalence : 0.7364
         Detection Rate : 0.6757
   Detection Prevalence : 0.8345
      Balanced Accuracy : 0.6576

       'Positive' Class : 0


Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1109  234
         1   92  196

               Accuracy : 0.8001
                 95% CI : (0.7799, 0.8193)
    No Information Rate : 0.7364
    P-Value [Acc > NIR] : 1.089e-09

                  Kappa : 0.4242

 Mcnemar's Test P-Value : 5.752e-15

            Sensitivity : 0.9234
            Specificity : 0.4558
         Pos Pred Value : 0.8258
```

```
         Neg Pred Value : 0.6806
             Prevalence : 0.7364
         Detection Rate : 0.6800
   Detection Prevalence : 0.8234
      Balanced Accuracy : 0.6896

       'Positive' Class : 0
```



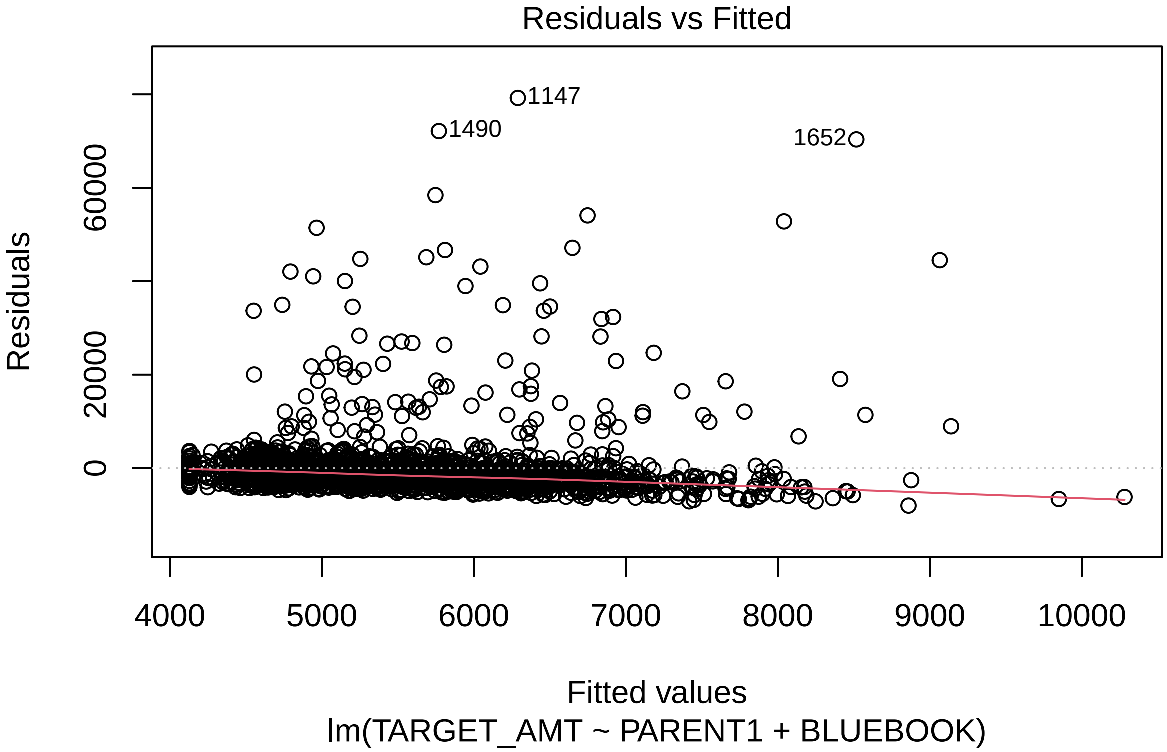

```
[1] 0.8012393
```

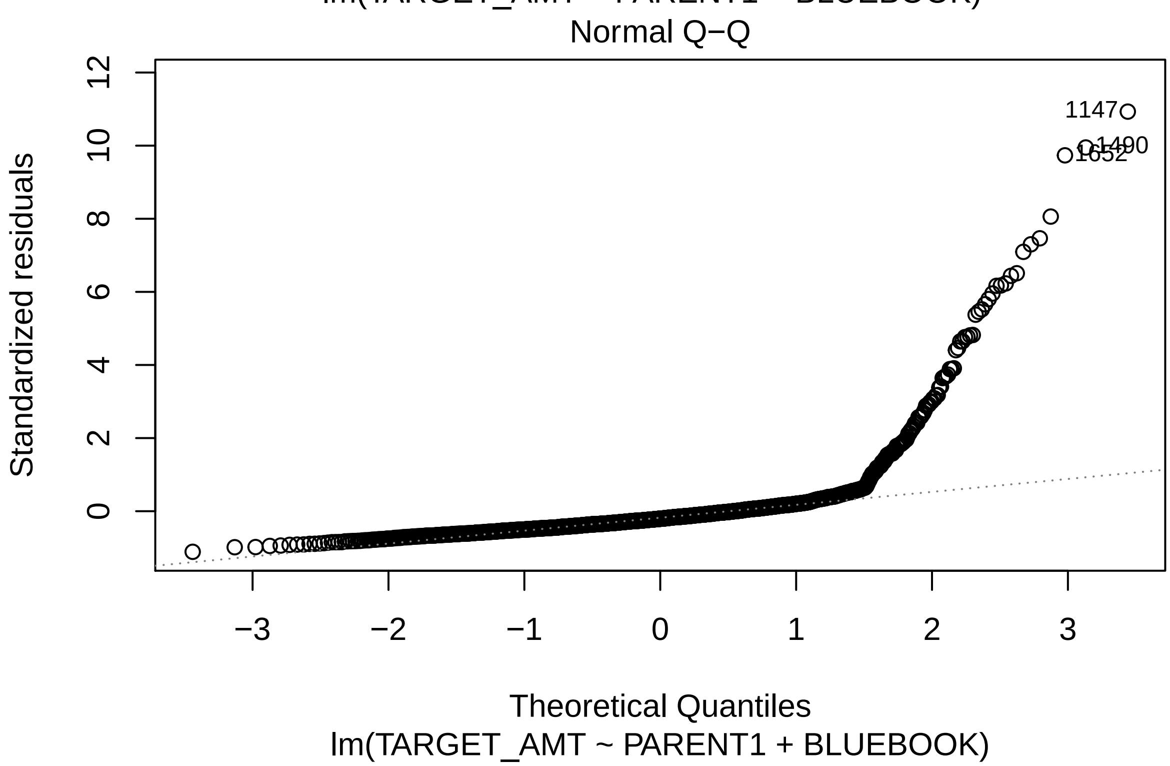


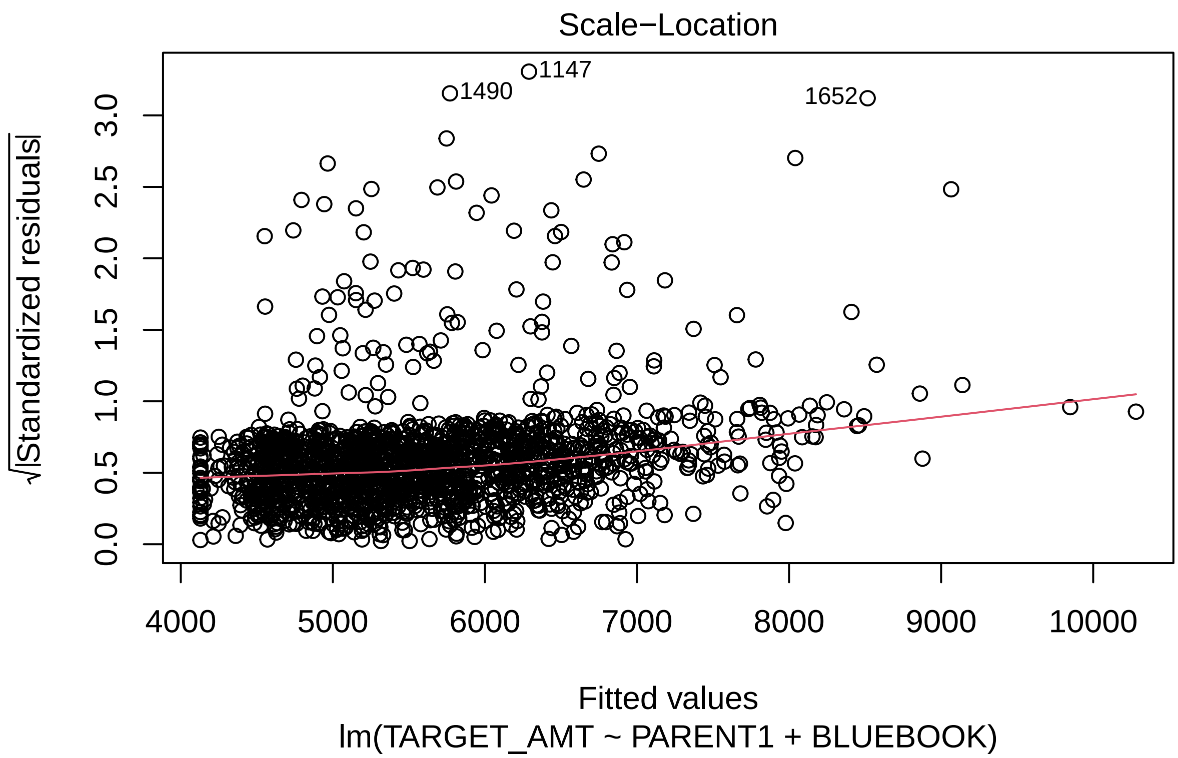

```
[1] 0.818229
```

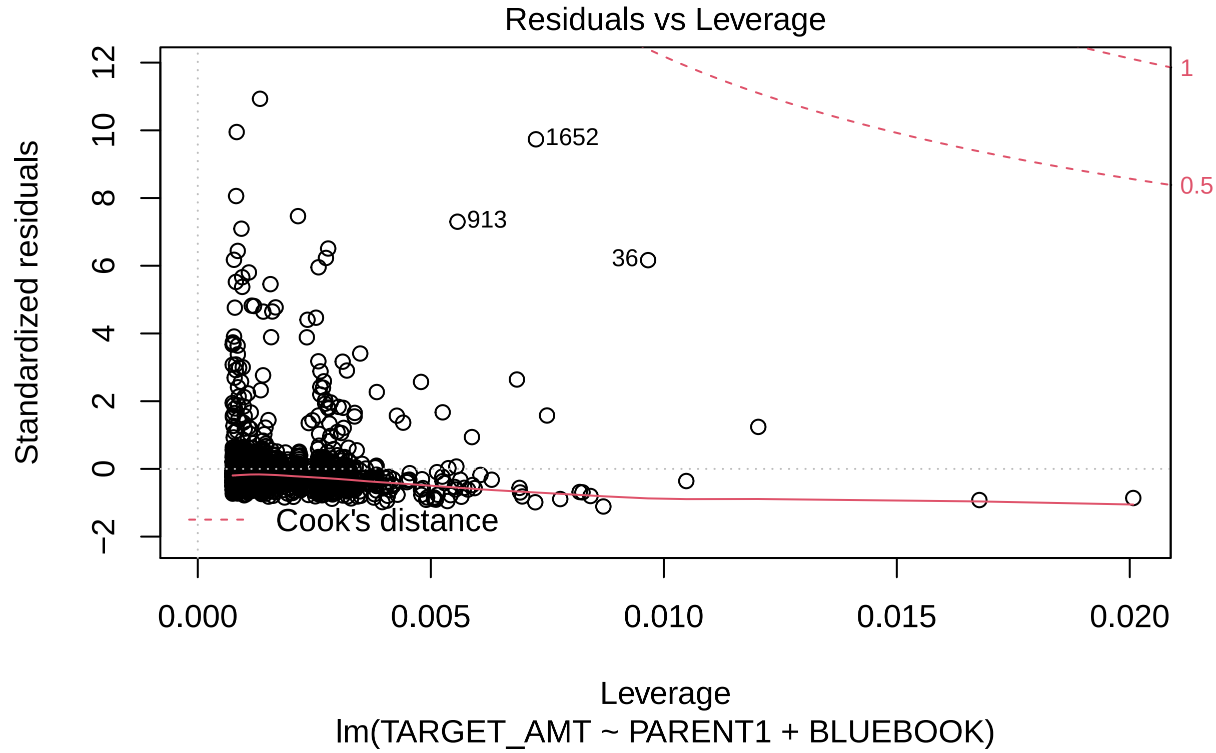
4.2 Linear Model Selection

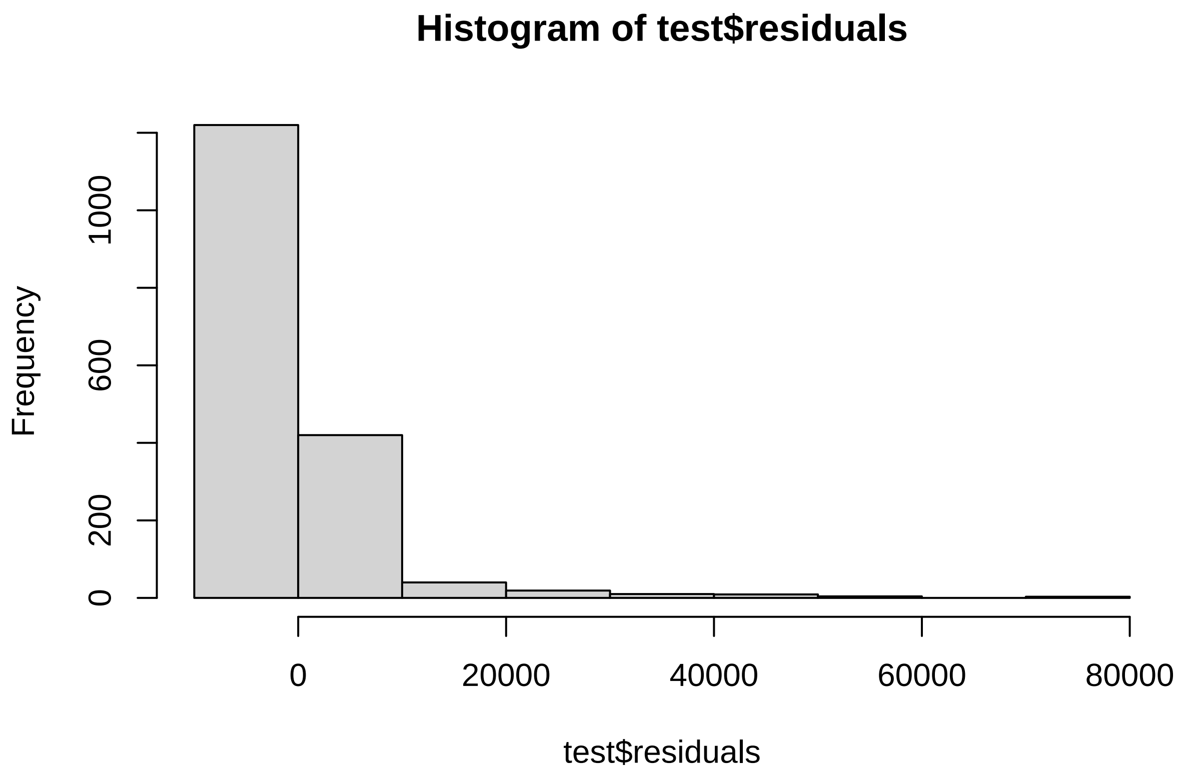
Model One diagnostics, highly skewed and horrible R-Squared value when testing on hold out data.



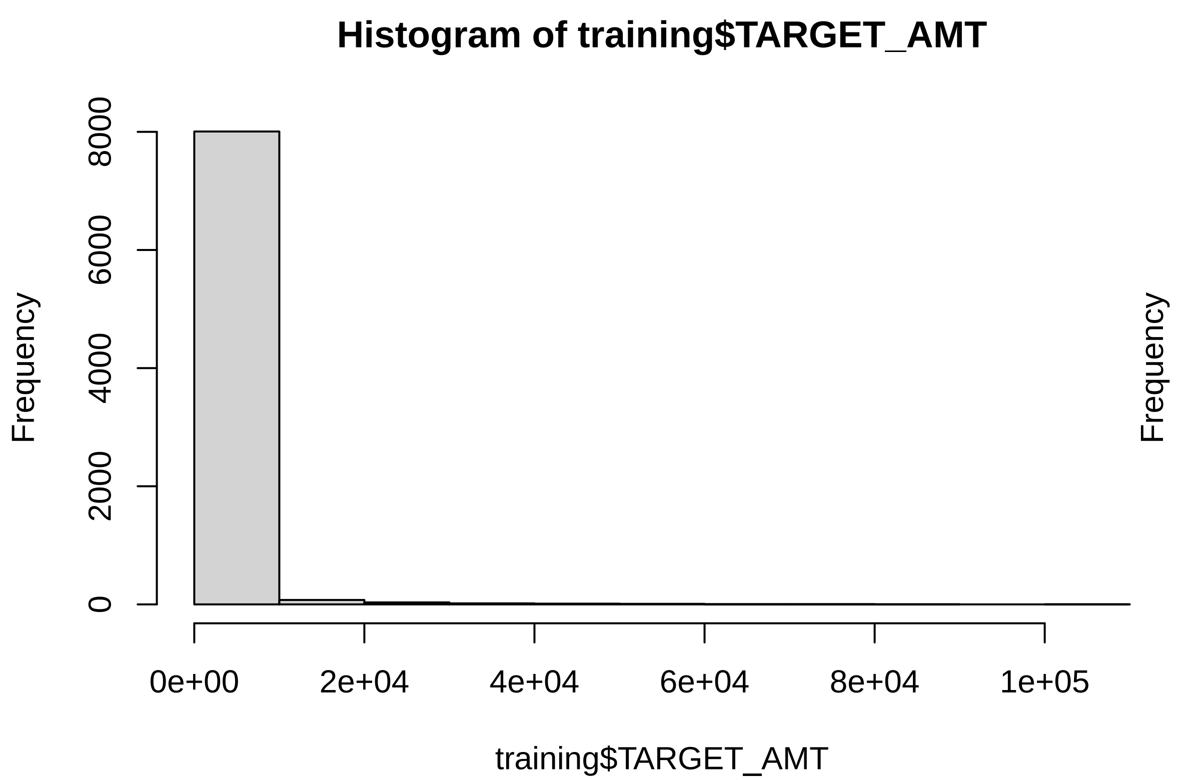


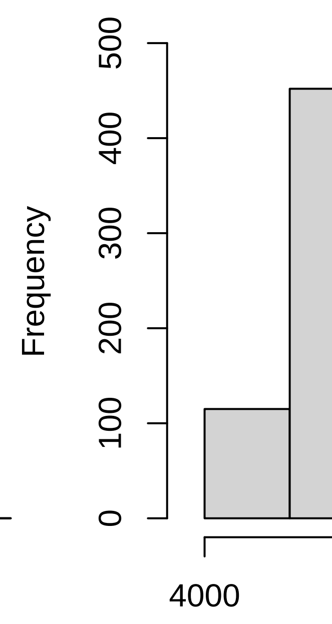

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1147
1490
1652
4000
5000
6000
7000
8000
9000
10000
Fitted values
lm(TARGET_AMT ~ PARENT1 + BLUEBOOK)


Residuals vs Leverage
Standardized residuals
−2
0
2
4
6
8
10
12
1652
913
36
1
0.5
Cook's distance
0.000
0.005
0.010
0.015
0.020
Leverage
lm(TARGET_AMT ~ PARENT1 + BLUEBOOK)




```
[1] 5.232757
```

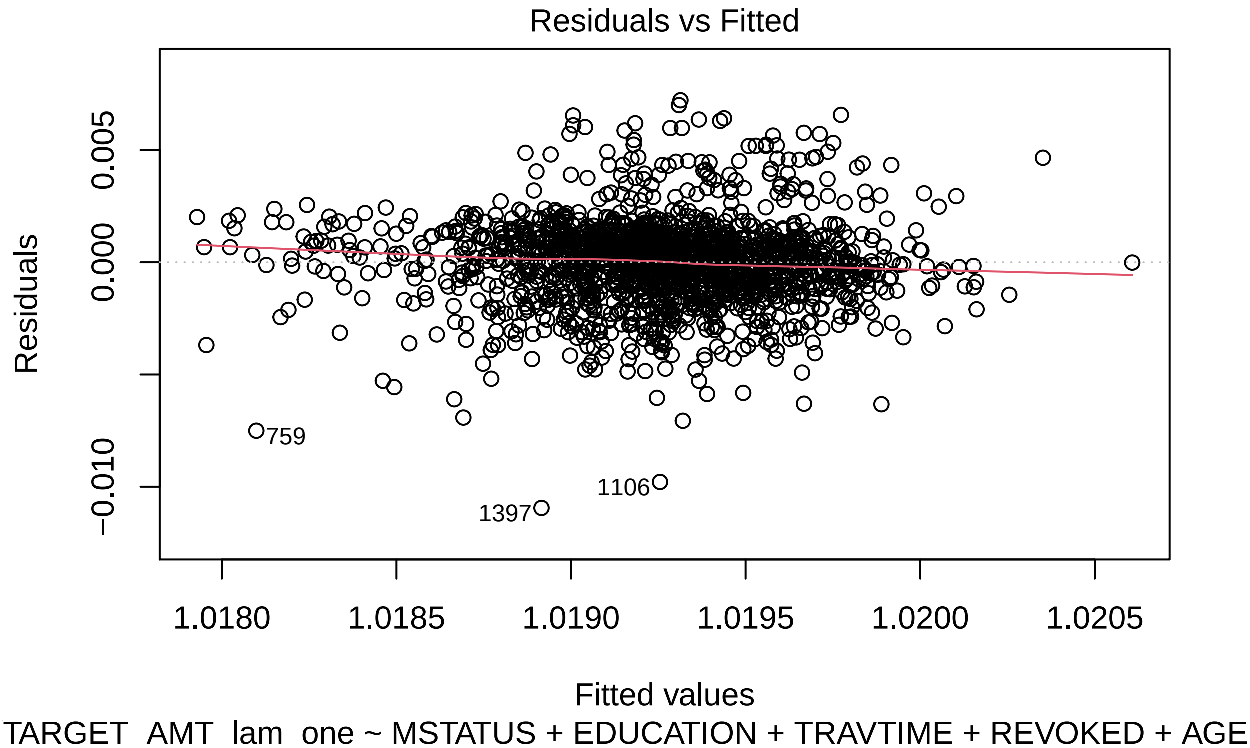


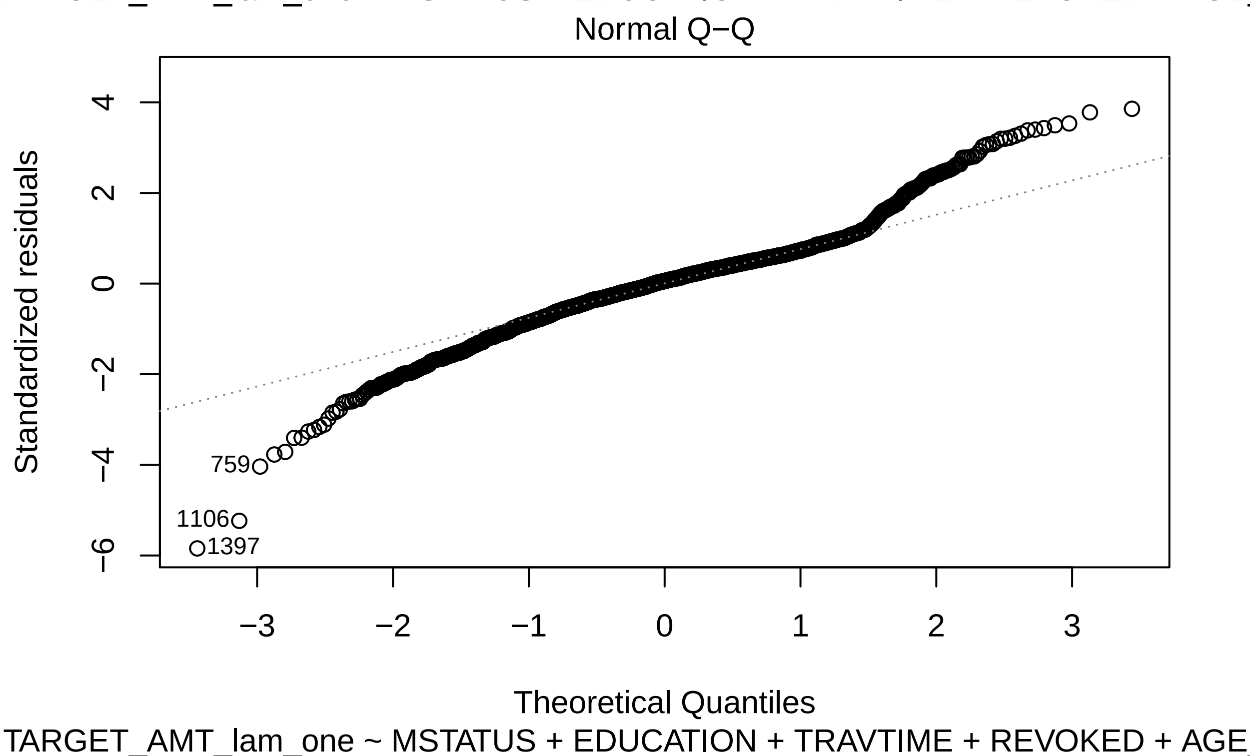


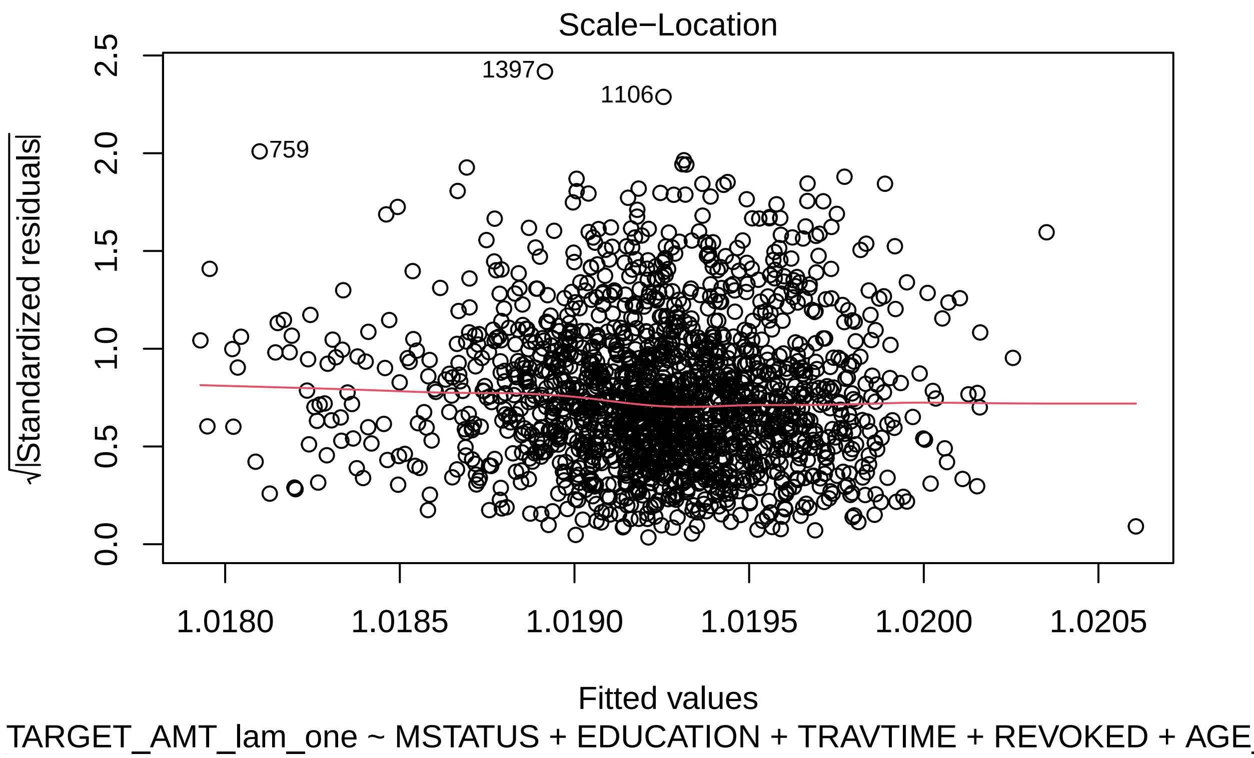

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
4129    4955    5505    5641    6227    9036

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

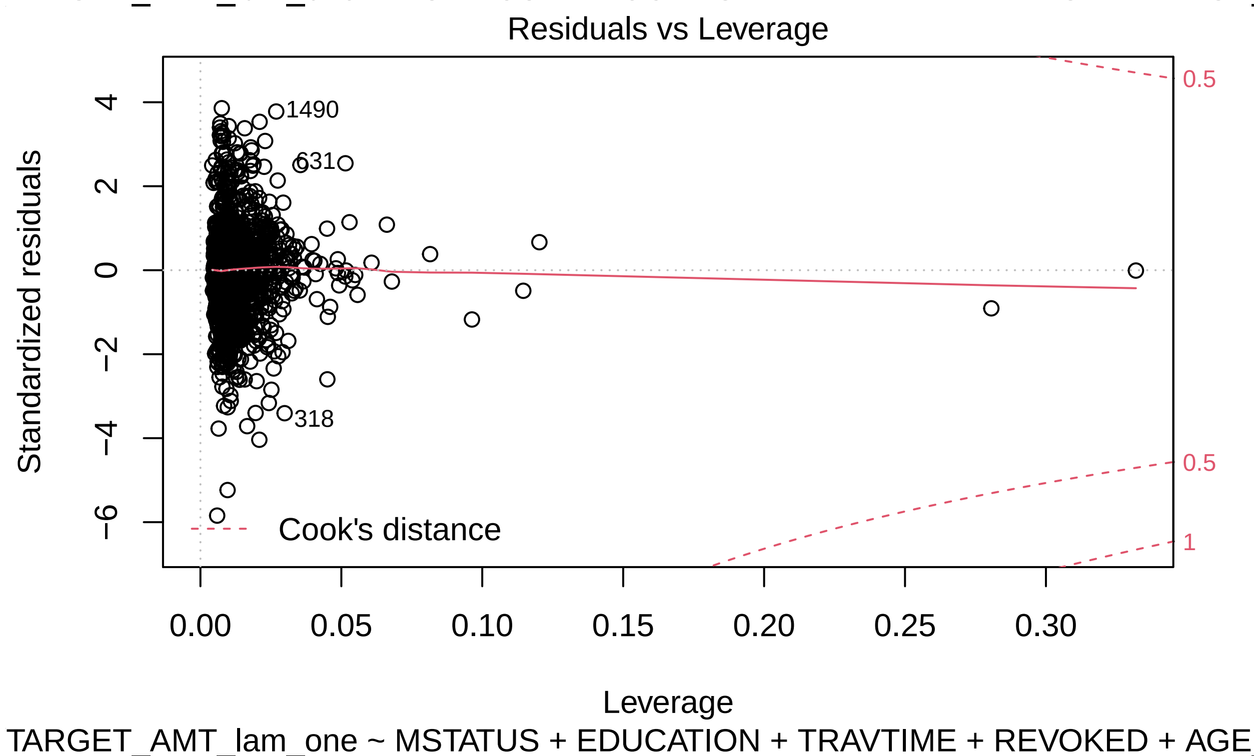
```
    30.28   2609.78   4104.00   5702.18   5787.00 107586.14

        RMSE     Rsquared          MAE
9.246177e+03 1.476241e-02 3.950348e+03
```

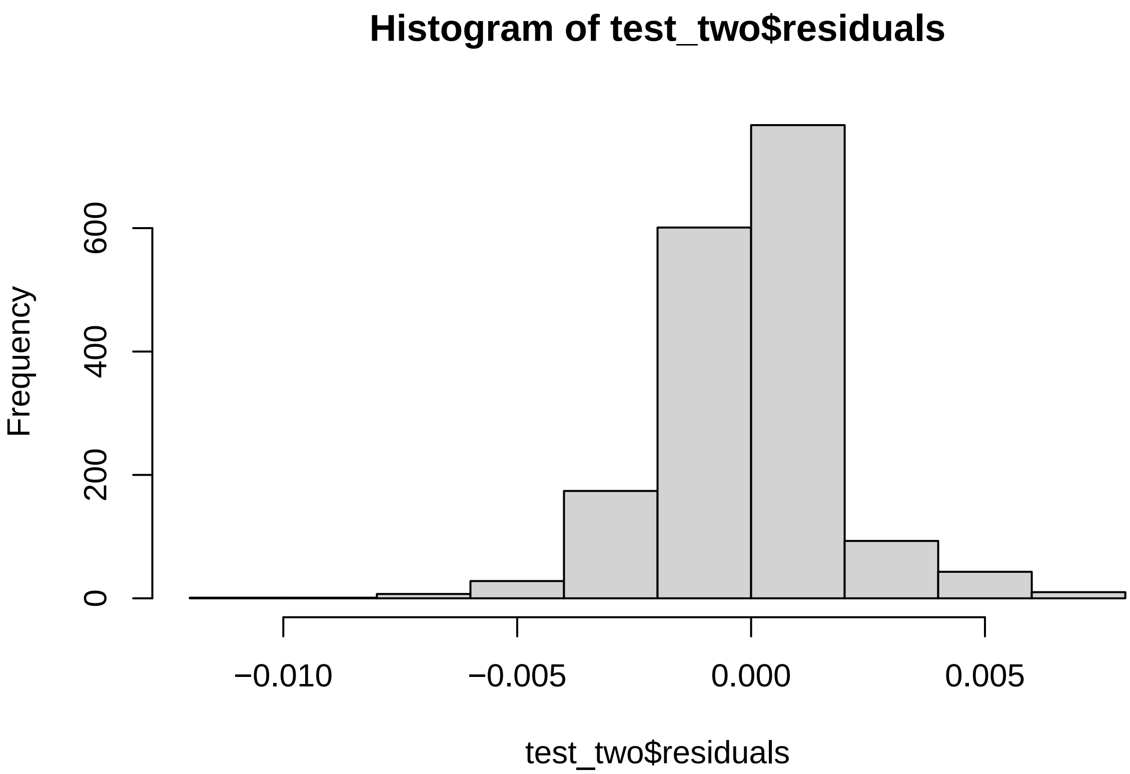
We can see belkw that model two is no longer skewed, and while the R-Squared is small, is now above one percent accuracy. Far better than the first model



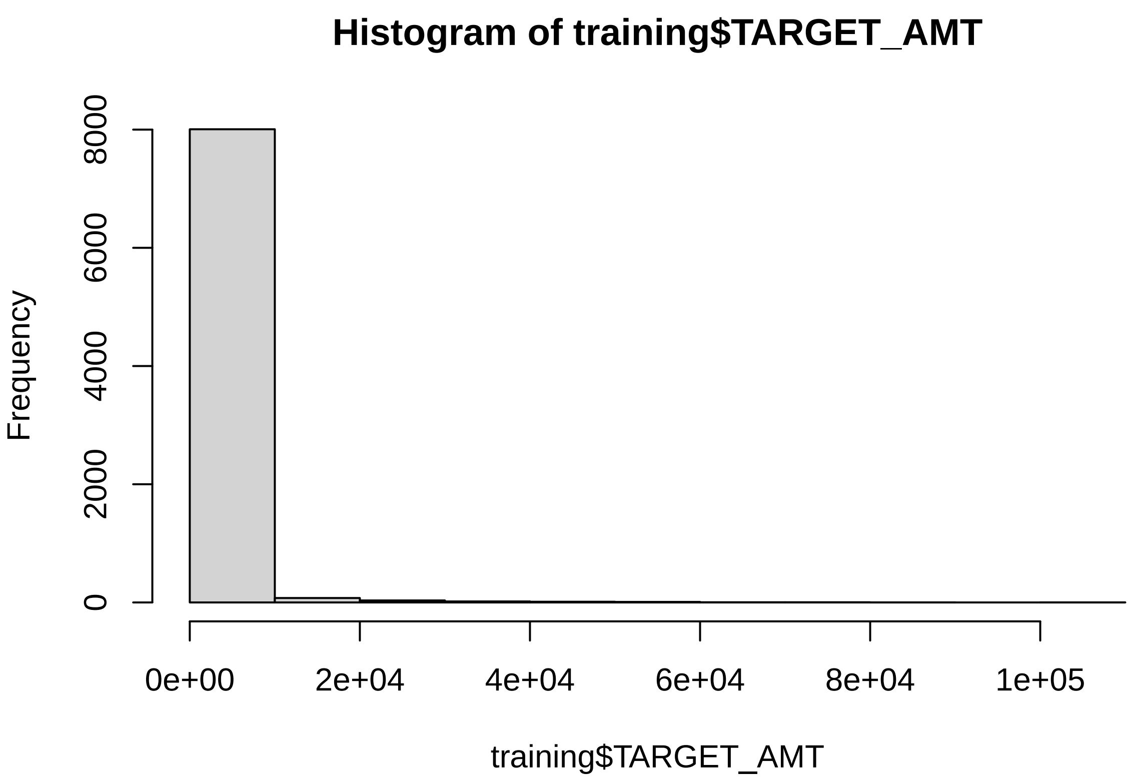


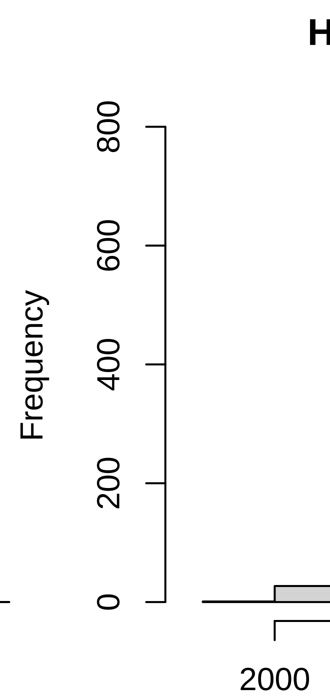

Scale−Location
√|Standardized residuals|
Fitted values
TARGET_AMT_lam_one ~ MSTATUS + EDUCATION + TRAVTIME + REVOKED + AGE
1397
1106
759


Residuals vs Leverage
Standardized residuals
Leverage
TARGET_AMT_lam_one ~ MSTATUS + EDUCATION + TRAVTIME + REVOKED + AGE
1490
631
318
Cook's distance
0.5
0.5
1




```
[1] -0.0882088
```






```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1853    3584    3924    3963    4314    7849
```

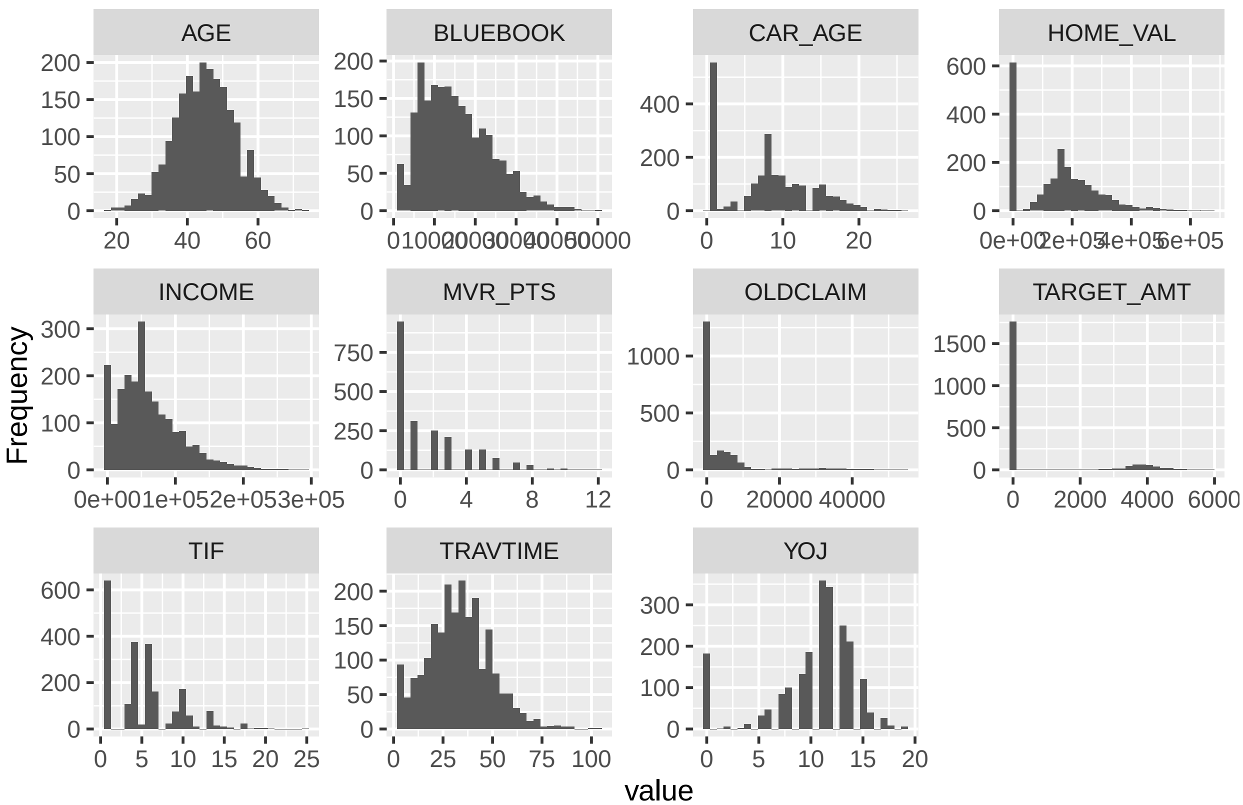
```
  Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
```

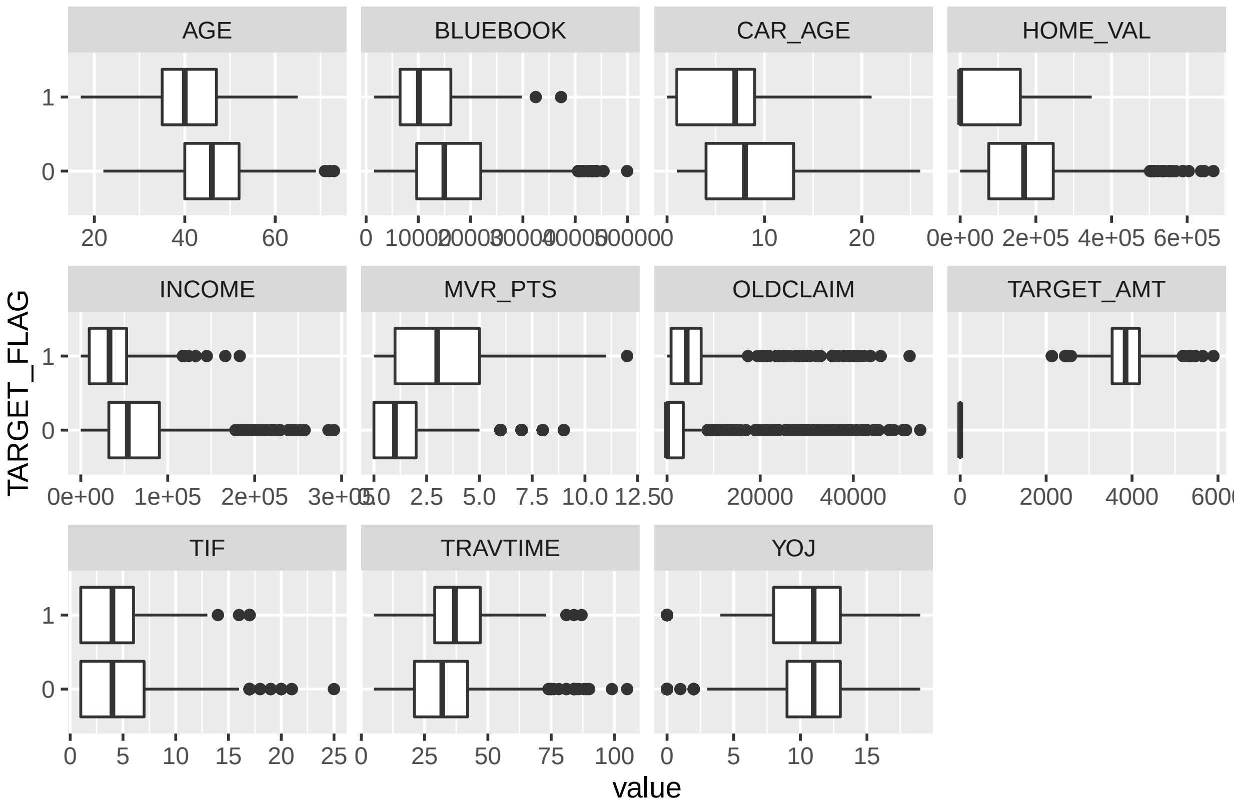
```
    30.28   2609.78   4104.00   5702.18   5787.00 107586.14

        RMSE     Rsquared          MAE
9.523520e+03 4.897731e-03 3.617693e+03
```

4.3 Making Predictions for the Evaluation Data

Using model two for both our binary logistic model and our our linear regression model, we update our evaluation data set with the final required predictions. We can now see how our evaluation predictions look below.

	vars	n	mean	sd	min	max	range	se
TARGET_AMT	1	2141	6.862709e+02	1.495171e+03	0	5893.581	5893.581	32.3133892
AGE	2	2141	4.501681e+01	8.523014e+00	17	73.000	56.000	0.1841980
YOJ	3	2141	1.040635e+01	4.079380e+00	0	19.000	19.000	0.0881629
INCOME	4	2141	5.982530e+04	4.565402e+04	0	291182.000	291182.000	986.6672438
HOME_VAL	5	2141	1.535092e+05	1.260609e+05	0	669271.000	669271.000	2724.4080232
TRAVTIME	6	2141	3.315227e+01	1.572239e+01	5	105.000	100.000	0.3397898
BLUEBOOK	7	2141	1.546943e+04	8.462367e+03	1500	49940.000	48440.000	182.8872917
TIF	8	2141	5.244745e+00	3.971026e+00	1	25.000	24.000	0.0858212
OLDCLAIM	9	2141	4.022168e+03	8.565379e+03	0	54399.000	54399.000	185.1135707
MVR_PTS	10	2141	1.765997e+00	2.203413e+00	0	12.000	12.000	0.0476198
CAR_AGE	11	2141	8.172349e+00	5.589936e+00	0	26.000	26.000	0.1208088




AGE
BLUEBOOK
CAR_AGE
HOME_VAL
INCOME
MVR_PTS
OLDCLAIM
TARGET_AMT
TIF
TRAVTIME
YOJ
TARGET_FLAG
value

5 Conclusion

The underlying nature of this data set had a few subtle complexities. In the way of modifications, there was a need to use regular expressions and re coded factors in order to make the data more interpretable to our models. In addition, there was a large focus on transforming our variables to smooth out the distributions and reduce skewness. After processing the data and transforming the necessary variables, we were able to determine that the second iteration of our models performed the best. It most accurately interpreted the data and seemed best poised to deal data abstractions.