Data 621 Homework 4

Critical Thinking Group 3: Vyannna Hill, Jose Rodriguez, and Christian Uriostegui

2023-12-03

Data Exploration

This assignment is a exploration of the insurance data set. The tasks are finding the optimal models for predicting if the insured driver will be involved in a car crash and estimating the value of the insurance payout.

Looking into the insurance data set Reviewing the data set below, there are 8,161 insured drivers apart of the data set. There are a few categorical variables: Parent1, Mstat, Sex, education, Job, cartype, caruse, redcar, revoked, and urbancity.

For the modeling, we will need these categorical variables to be numeric to run in our model. Let's review which variables can be updated into binary and which will need new multiple columns for dummy variables. It is noted that education, Job, and Car Type have multiple options so its dummy variables will be k-1.

In addition, there will need to be some transformations of a few non categorical variables. The values income, home val, blue book, and old claim will need to be re-define as numeric values for the regression.

```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ          INCOME
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Length:8161
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   Class :character
##  Median :45.00   Median :0.0000   Median :11.0   Mode  :character
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0
##  NA's   :6                        NA's   :454
##    PARENT1            HOME_VAL           MSTATUS              SEX
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   EDUCATION             JOB               TRAVTIME        CAR_USE
```

```
##  Length:8161        Length:8161        Min.   :  5.00   Length:8161
##  Class :character   Class :character   1st Qu.: 22.00   Class :character
##  Mode  :character   Mode  :character   Median : 33.00   Mode  :character
##                                        Mean   : 33.49
##                                        3rd Qu.: 44.00
##                                        Max.   :142.00
##
##    BLUEBOOK             TIF           CAR_TYPE           RED_CAR
##  Length:8161        Min.   : 1.000   Length:8161        Length:8161
##  Class :character   1st Qu.: 1.000   Class :character   Class :character
##  Mode  :character   Median : 4.000   Mode  :character   Mode  :character
##                     Mean   : 5.351
##                     3rd Qu.: 7.000
##                     Max.   :25.000
##
##    OLDCLAIM           CLM_FREQ        REVOKED             MVR_PTS
##  Length:8161        Min.   :0.0000   Length:8161        Min.   : 0.000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Mode  :character   Median :0.0000   Mode  :character   Median : 1.000
##                     Mean   :0.7986                      Mean   : 1.696
##                     3rd Qu.:2.0000                      3rd Qu.: 3.000
##                     Max.   :5.0000                      Max.   :13.000
##
##     CAR_AGE        URBANICITY
##  Min.   :-3.000   Length:8161
##  1st Qu.: 1.000   Class :character
##  Median : 8.000   Mode  :character
##  Mean   : 8.328
##  3rd Qu.:12.000
##  Max.   :28.000
##  NA's   :510

## $PARENT1
## [1] "No"  "Yes"
##
## $MSTATUS
## [1] "z_No" "Yes"
##
## $SEX
## [1] "M"   "z_F"
##
## $EDUCATION
## [1] "PhD"           "z_High School" "<High School"  "Bachelors"
## [5] "Masters"
##
## $JOB
## [1] "Professional"  "z_Blue Collar" "Clerical"      "Doctor"
## [5] "Lawyer"        "Manager"       ""              "Home Maker"
## [9] "Student"
##
## $CAR_USE
## [1] "Private"    "Commercial"
##
## $CAR_TYPE
```

```
## [1] "Minivan"     "z_SUV"       "Sports Car"  "Van"         "Panel Truck"
## [6] "Pickup"
##
## $RED_CAR
## [1] "yes" "no"
##
## $REVOKED
## [1] "No"  "Yes"
##
## $URBANICITY
## [1] "Highly Urban/ Urban"   "z_Highly Rural/ Rural"
```

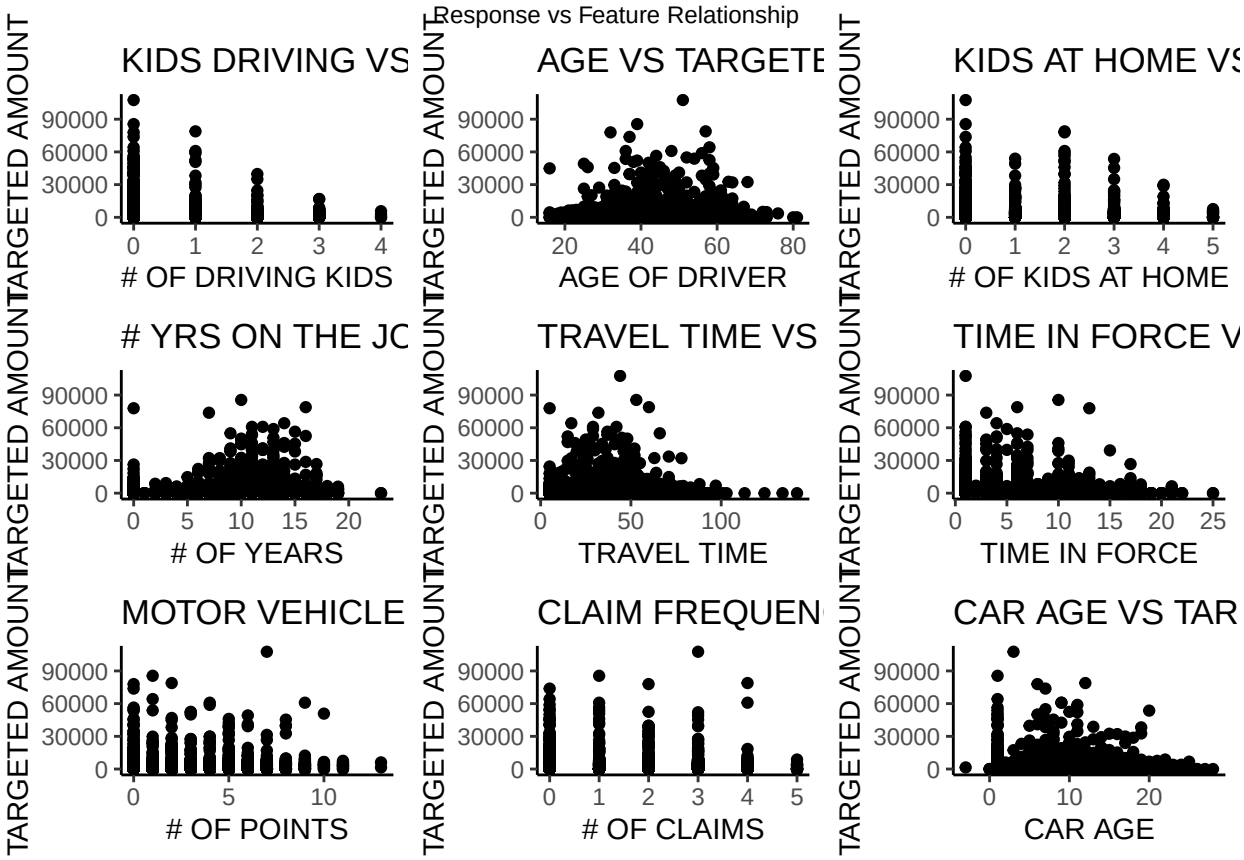
Checking for NAs and non-normal data Besides the needed transformations above, let's see if there are any missing values with the current data set. It is found that YOJ (Years on the Job), car age, and age have some missing values. The team will have to use imputation for those missing values.

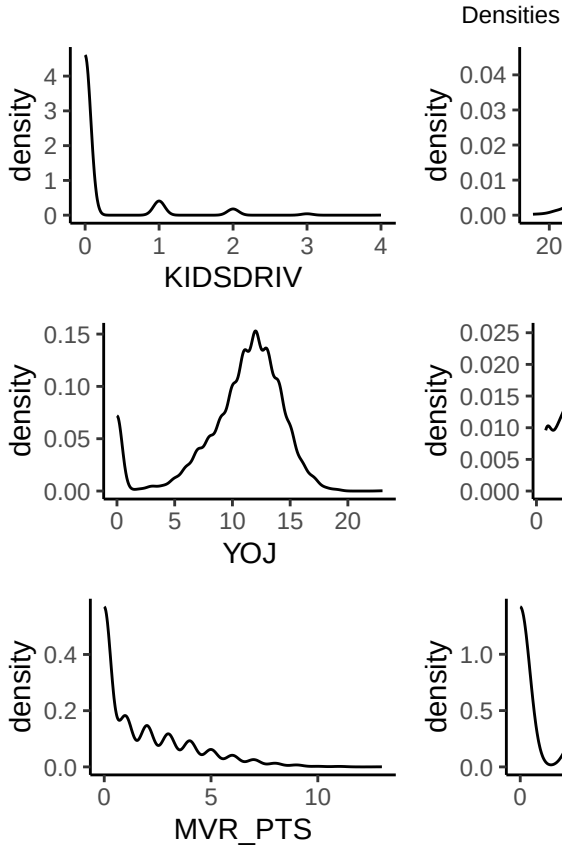
```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0           0           0           0           6           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##         454           0           0           0           0           0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##           0           0           0           0           0           0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##           0           0           0           0           0           0
##     CAR_AGE  URBANICITY
##         510           0
```

Next, the examination of the predictor variables for both linear and logistic regressions. For the linear regression, the predictor variables must pass with a linear relationship with the response variable (the targeted amount) in order to create the model.

Looking at the visual representation of their relationships below, there is a lot of non normality in the predictor values. The predictor values kids Driving, kids at home, time in force, motor vehicle points, and claim frequency resemble a linear relationship.

It can be interpreted that the claim amount lessens for the variables mentioned above. One call out is there are a ton of outliers in the variables (i.e claim frequency and mvp) that will need to be checked if they are influential points. The amount of outliers seen across the features are a concern, as these points may influence the prediction model. These predictors will need a transformation to correct its non-normality or filtered out the data set

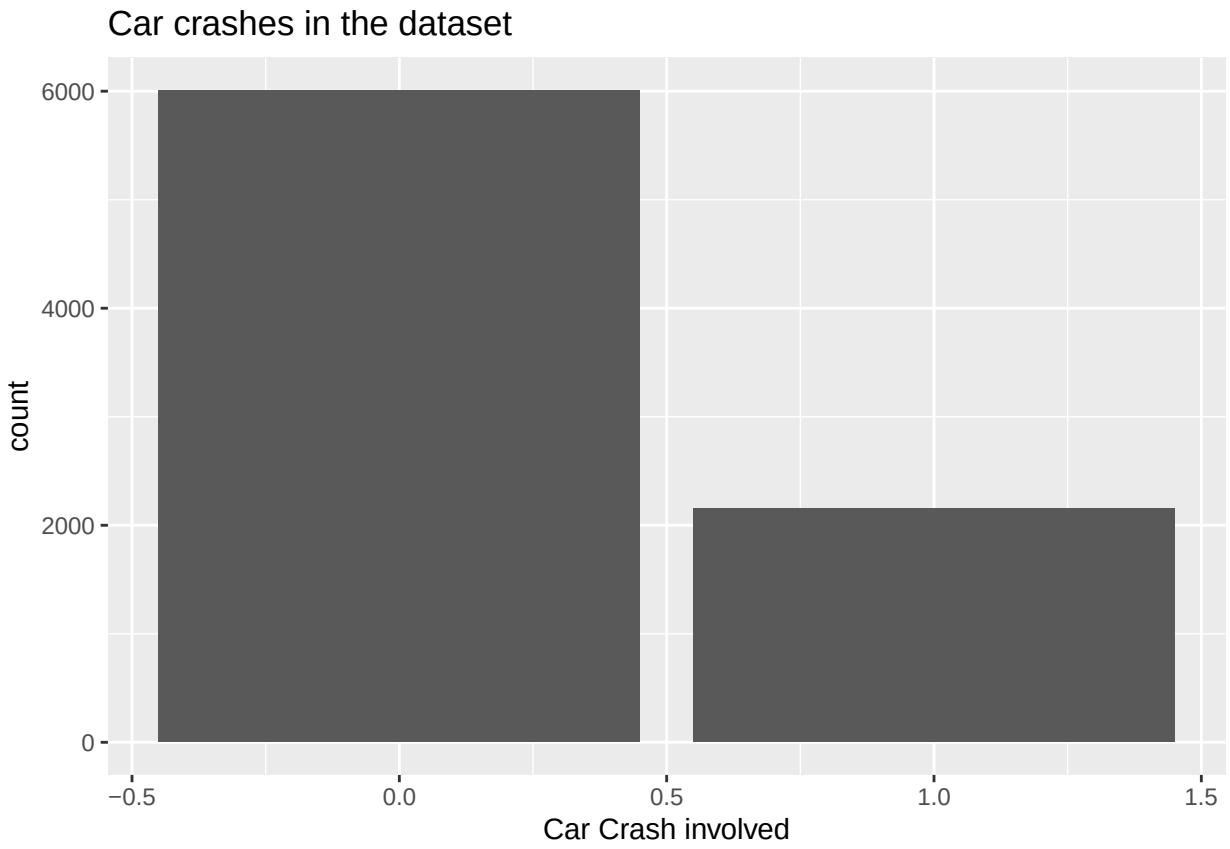

Response vs Feature Relationship
KIDS DRIVING VS
AGE VS TARGETE
KIDS AT HOME VS
YRS ON THE JC
TRAVEL TIME VS
TIME IN FORCE V
MOTOR VEHICLE
CLAIM FREQUEN
CAR AGE VS TAR
TARGETED AMOUNT
OF DRIVING KIDS
AGE OF DRIVER
OF KIDS AT HOME
OF YEARS
TRAVEL TIME
TIME IN FORCE
OF POINTS
OF CLAIMS
CAR AGE



It is apparent in the density distributions plotted below the non-normalness.

Looking if the data set is balanced Before the data preparation, the team wants to look at the distribution of cases between car crashes and non crash cases. The team noticed that the distribution of cases are heavily dominated by non car crashes. This imbalance can affect how the model predicts cases where there is a car crash case. It might be the best action to down sample the training set to help with the distribution.

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



Data Preparation

There is a list of tasks in order to begin the modeling process. The team will need to address the missing data, the categorical variables, checking column value types, and the transformations towards near normal.

Redefine Cost Variables The team noticed the cost variables income, home val, blue book, and old claim pulled as categorical. Let's transform them back into numeric values like targeted amount with the transformation below. The team did noticed after the transformation, some values of income and home value were missing. This can lead towards our second task!

Filling in the Missing In the data exploration, the team noticed some missing values that will need to be filled. To avoid any bias in the imputed data, lets use MICE to impute the data. This imputation fills the missing data with the predicted value .

Now, the data set is filled with all numeric values. Let's move onto the transformation of the categorical variables!

Transforms towards dummy variables To use the categorical variables, there will need to a transformation into dummy variables. All dummy variables will take the form K-1, which K is the number of unique values in the variable. For Parents, martial status, sex, car use, red car, revoked, and urban city it's assigning a binary true/false.

- Dictionary

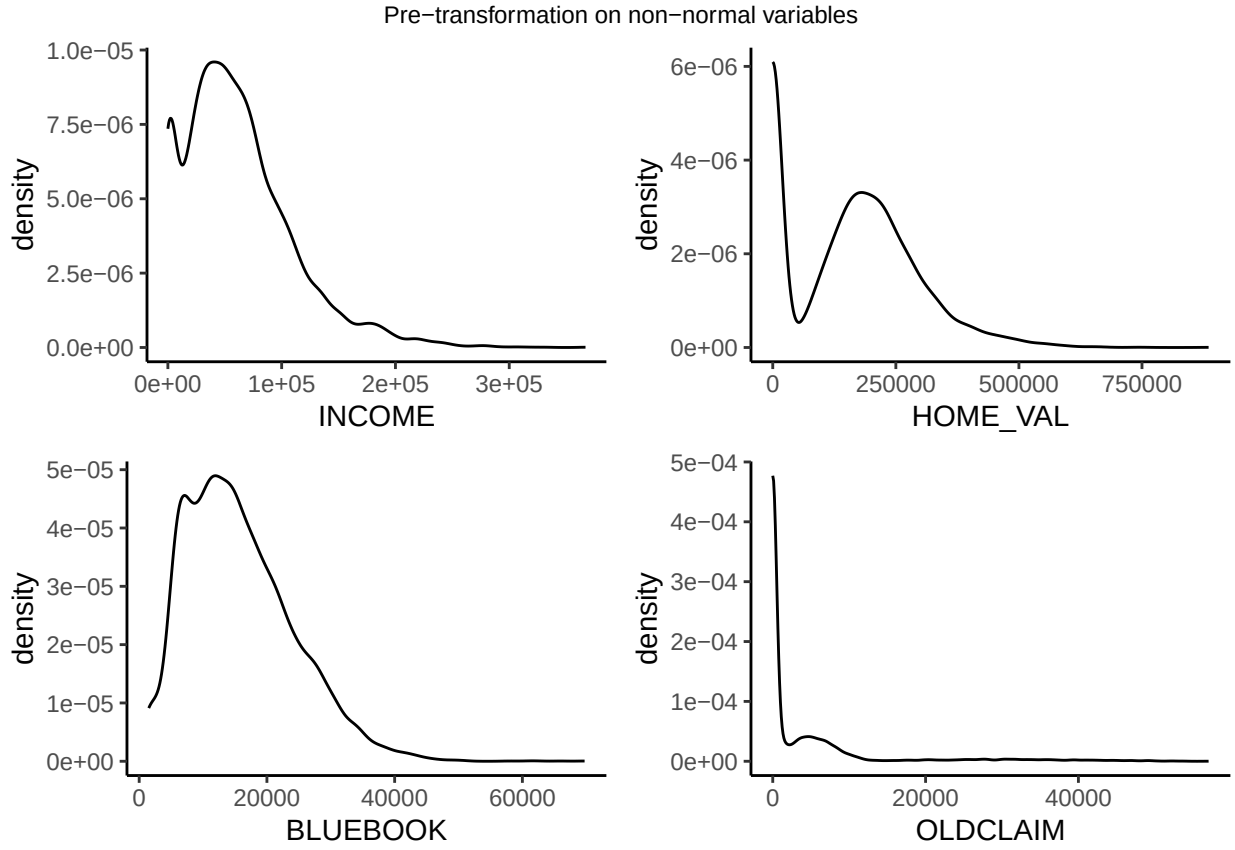
- False==0
- True==1

For job, education and car type the dummy variables will need multiple columns. For example, car type will need the structure of four new columns for the five values found. The value that is not given a column for car type is Panel trunk; there was not a reason for this value selection.

Now, all the categorical variables are converted to numeric for the model!

Converting variables towards normal distribution The next task is transforming the non-normal data seen in the numeric predictor values. The team wants to ensure all data is close to normal before applying to the regression model. First, Let's check the distribution of the new spend metrics below and if they will need to be included in the transformation.

The plots suggest these predictors will need its values transform in order to use in linear regression.

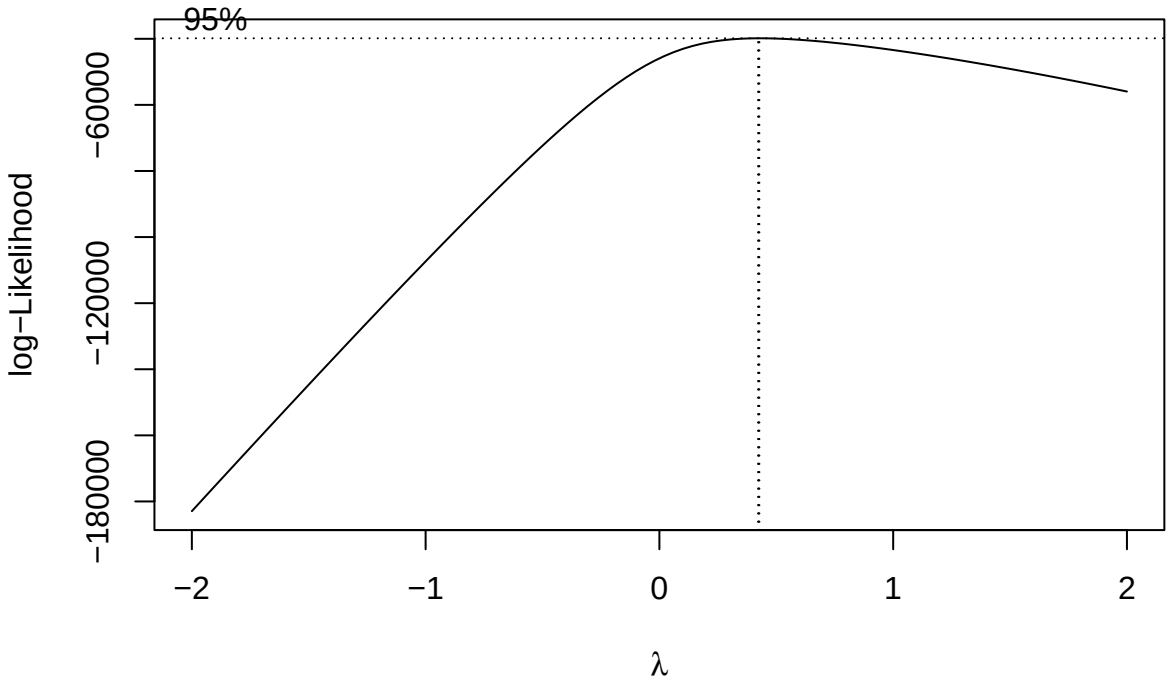


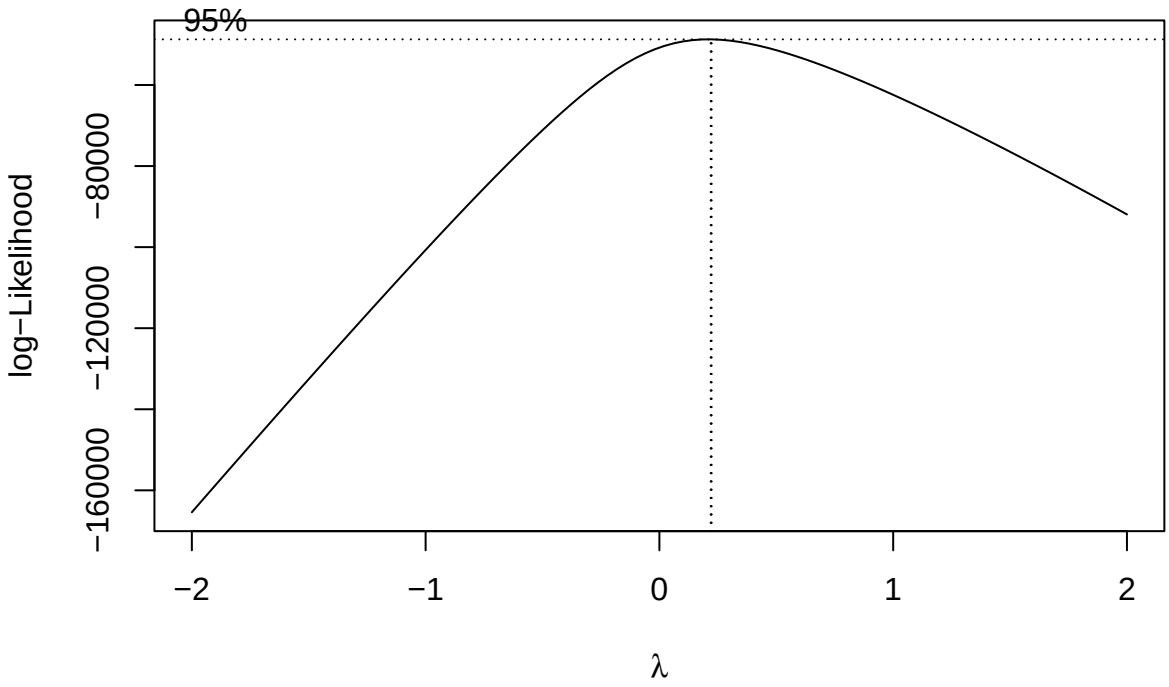
The team will use BoxCox transformations below for the following variables: INCOME, HOME_VAL, BLUEBOOK, OLDCLAIM, AGE, YOJ, MVR_PTS, CLM_FREQ, and CAR_AGE.

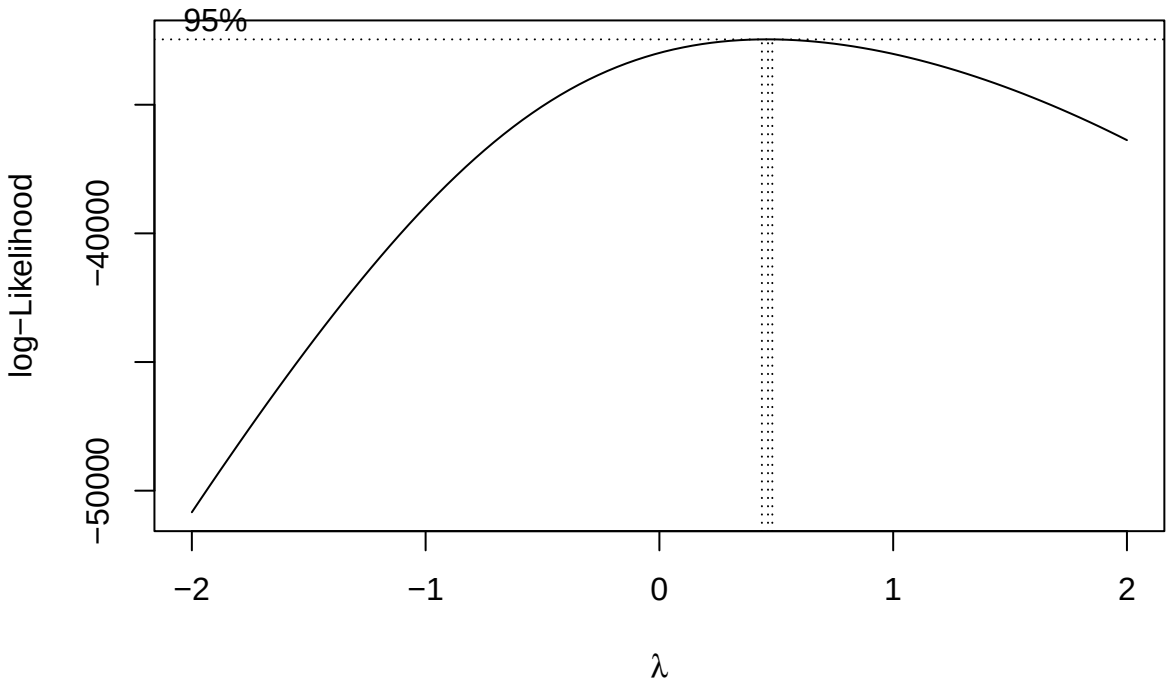
For reference, the transformations listed below will be used for the lambda value provided.

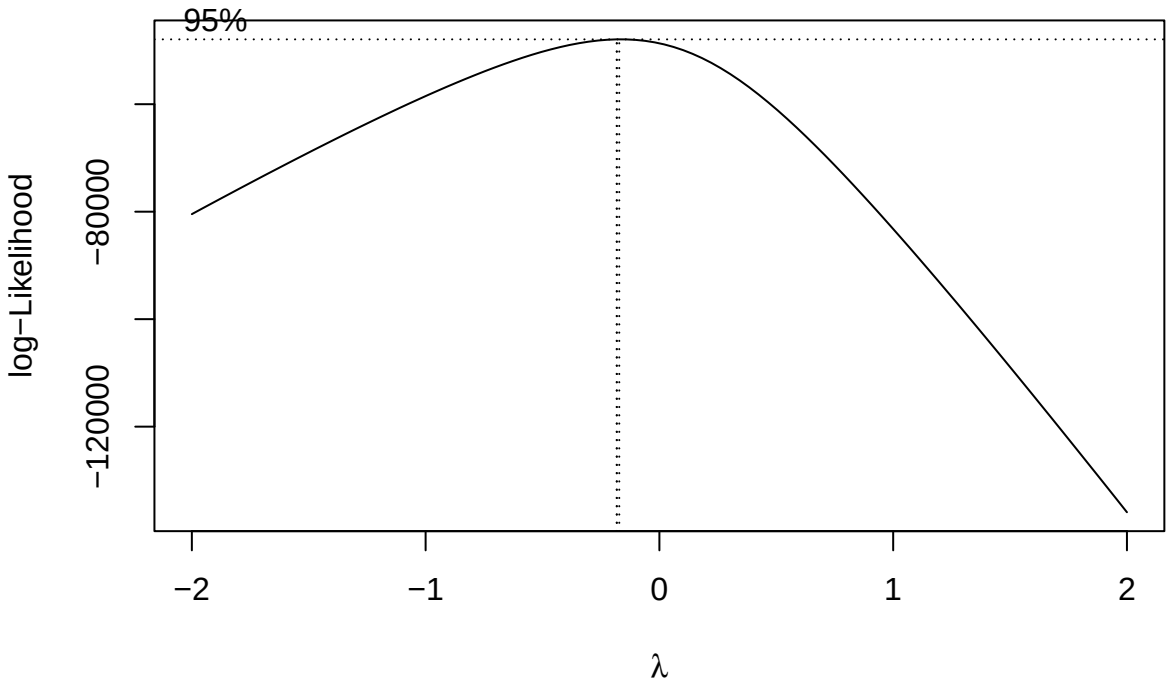
- Box Cox Transformations of λ
  - λ= -2| 1/xˆ2
  - λ=-1| 1/x
  - λ=-0.5| 1/sqrt(x) +λ=0| log(x) +λ=0.5| sqrt(x) +λ=1| x +λ=2| xˆ2

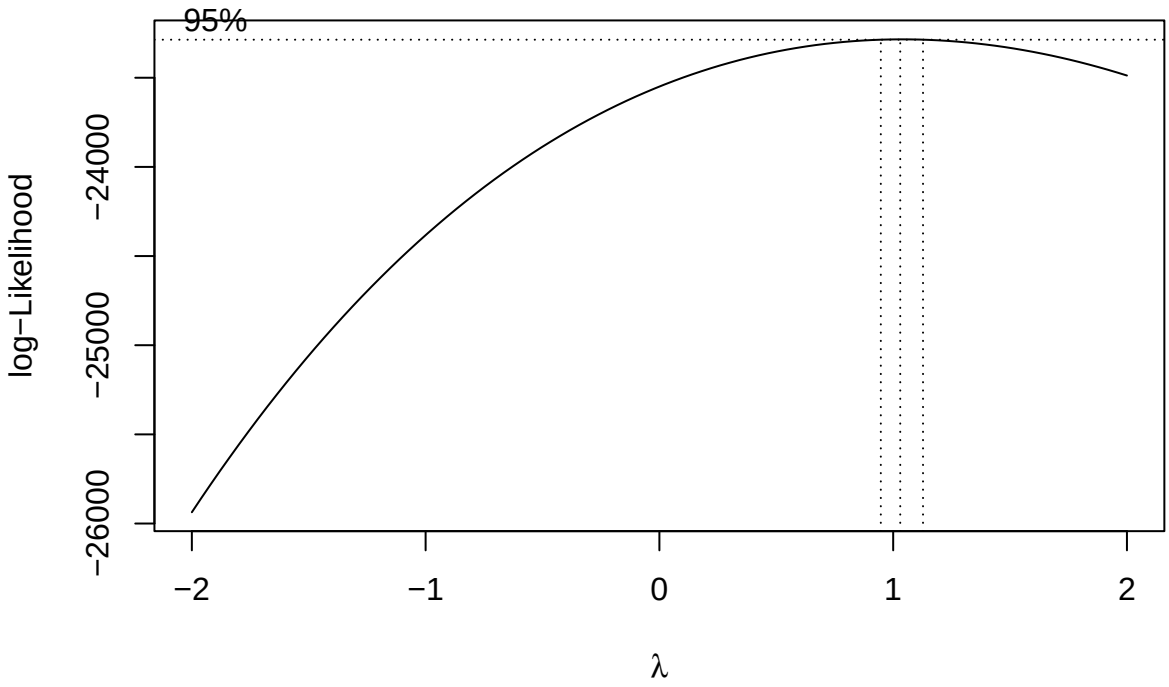
Some transformations performed needed a constant for its transformation as there are many predictors have zeros as the values provided. In order to perform the necessary transformations, a constant of 1 was applied to the needed transformation.

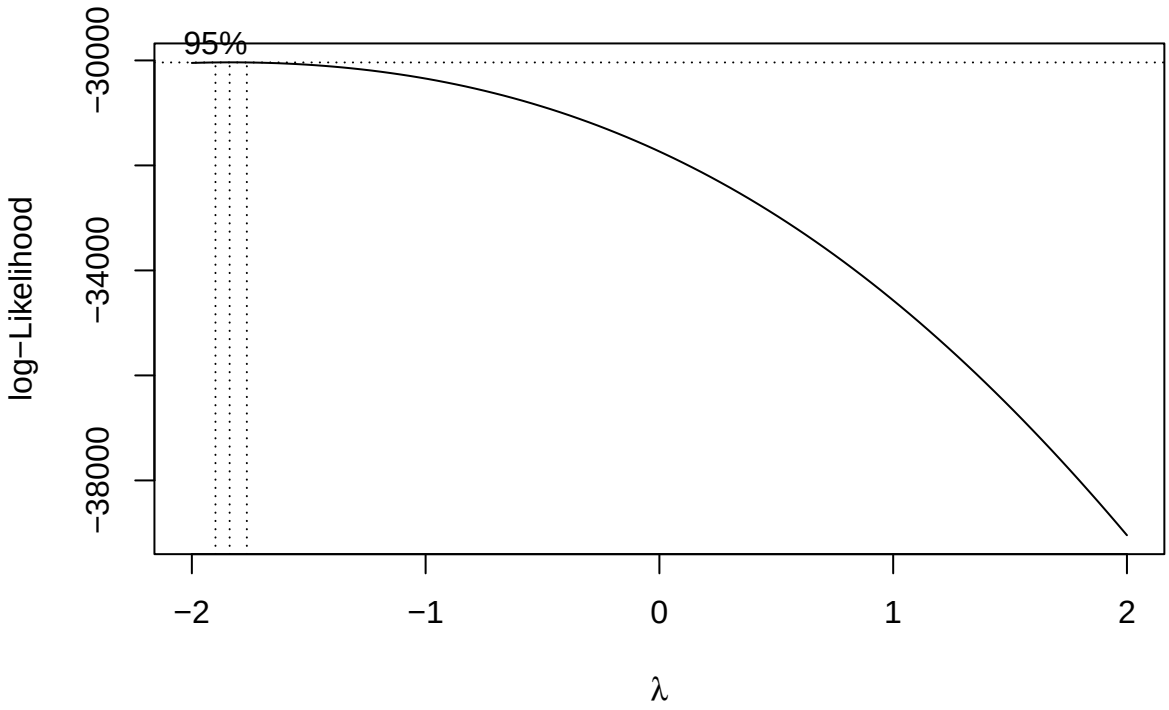


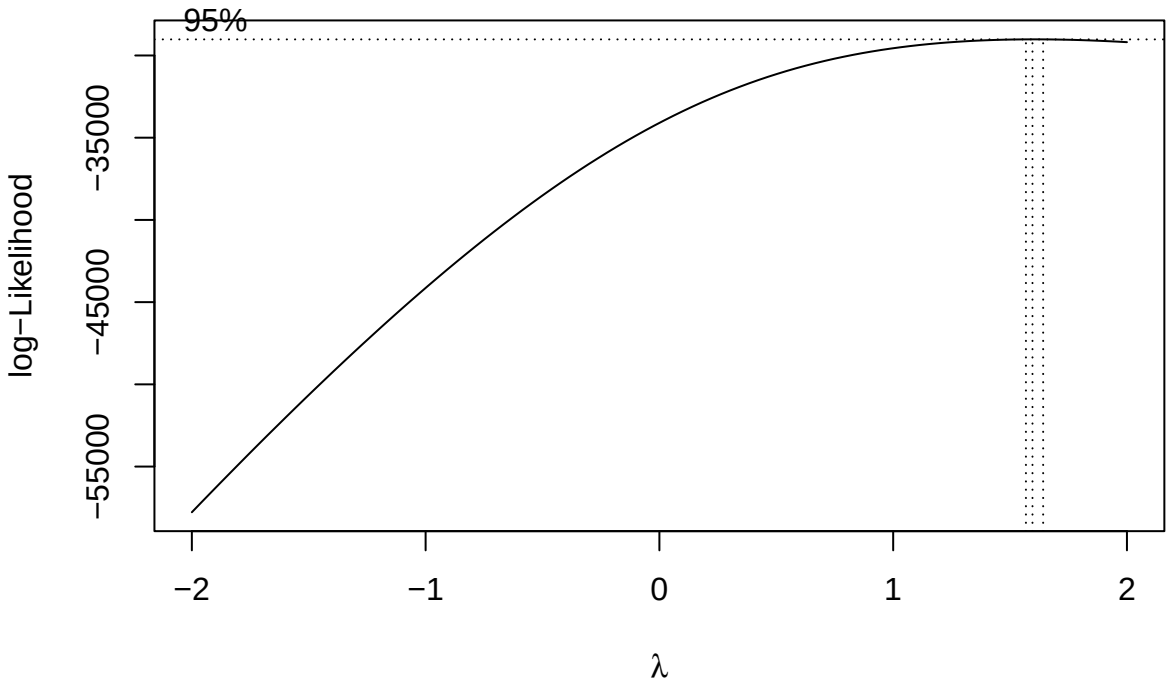

95%
log−Likelihood
−160000
−120000
−80000
−2
−1
0
1
2
λ

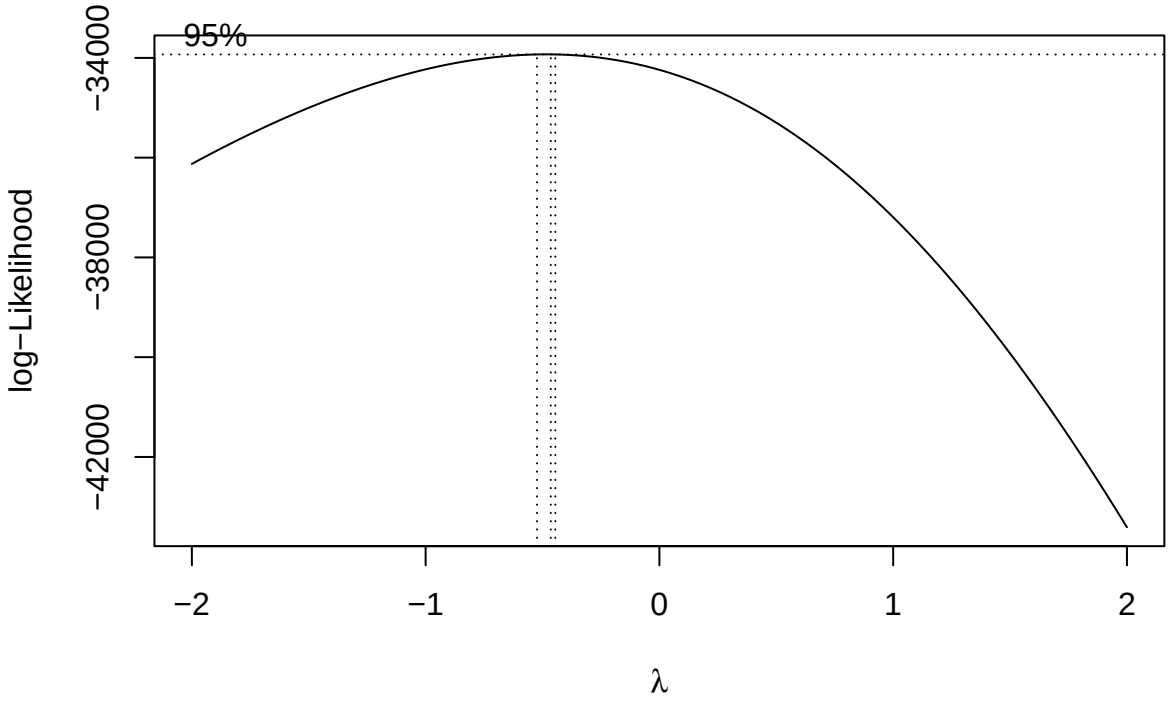

95%
log−Likelihood
−40000
−50000
−2
−1
0
1
2
λ

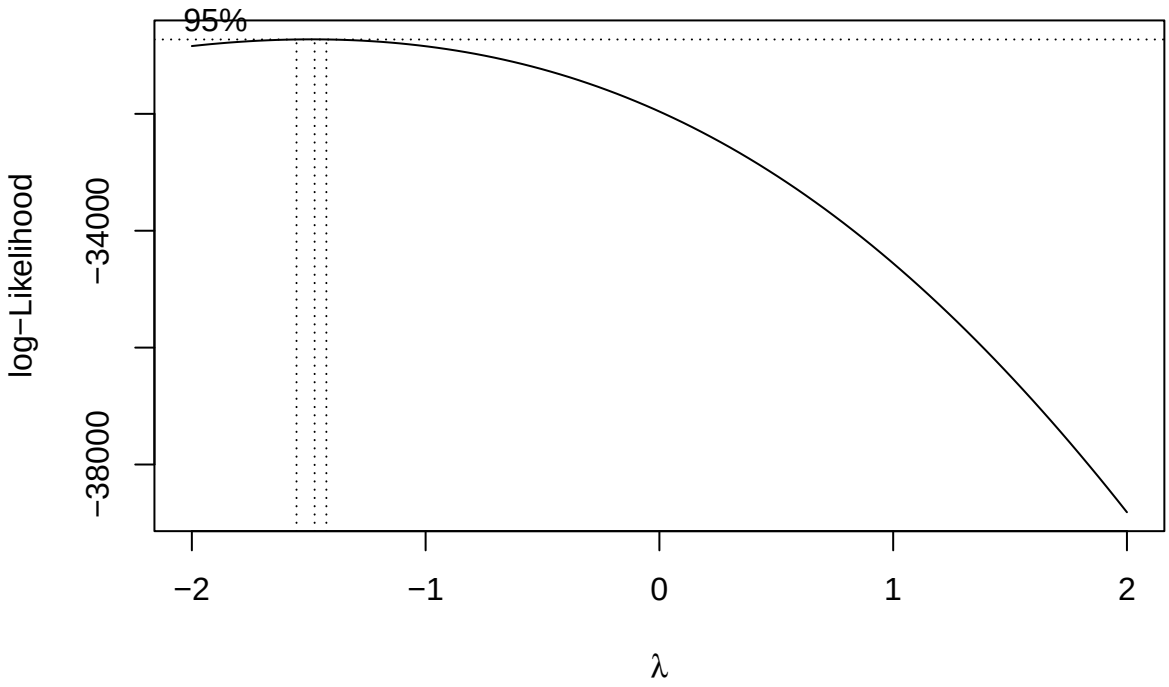

95%
log−Likelihood
−120000
−80000
−2
−1
0
1
2
λ

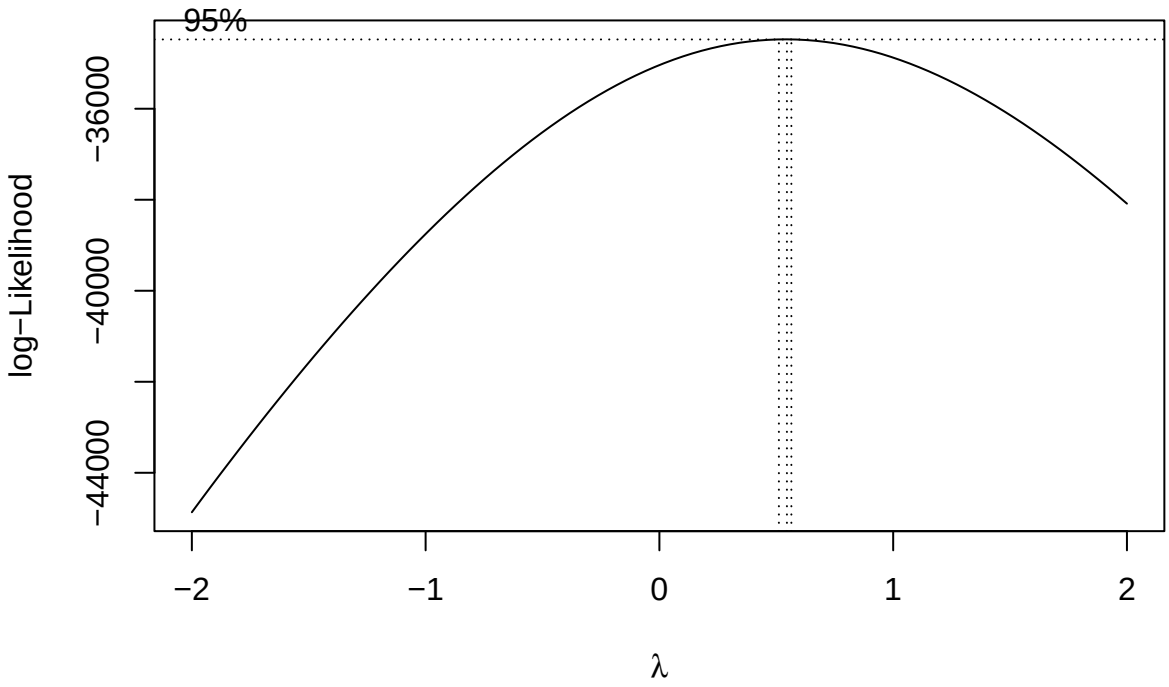

95%
log−Likelihood
−24000
−25000
−26000
−2
−1
0
1
2
λ

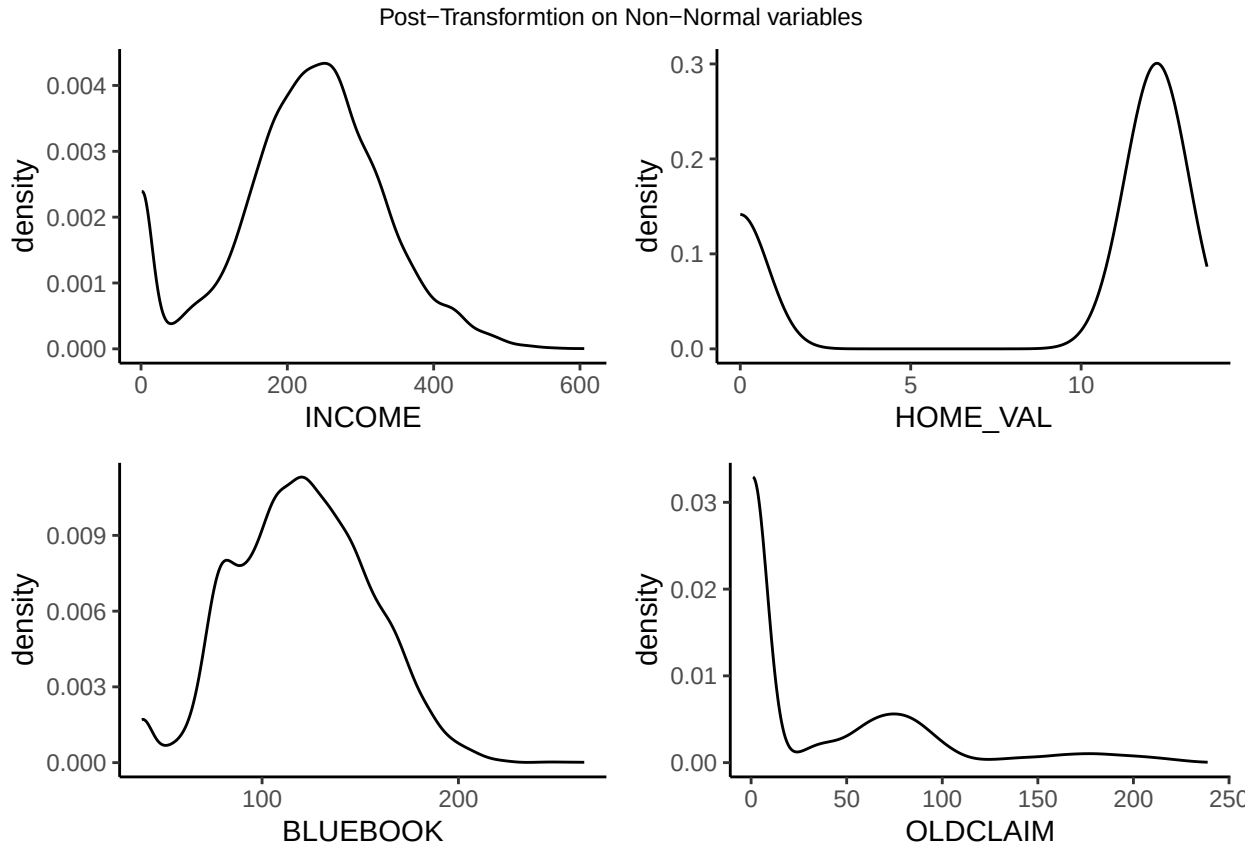

95%
log−Likelihood
−30000
−34000
−38000
−2
−1
0
1
2
λ


95%
log−Likelihood
−55000
−45000
−35000
−2
−1
0
1
2
λ


95%
−34000
−38000
−42000
log−Likelihood
−2
−1
0
1
2
λ

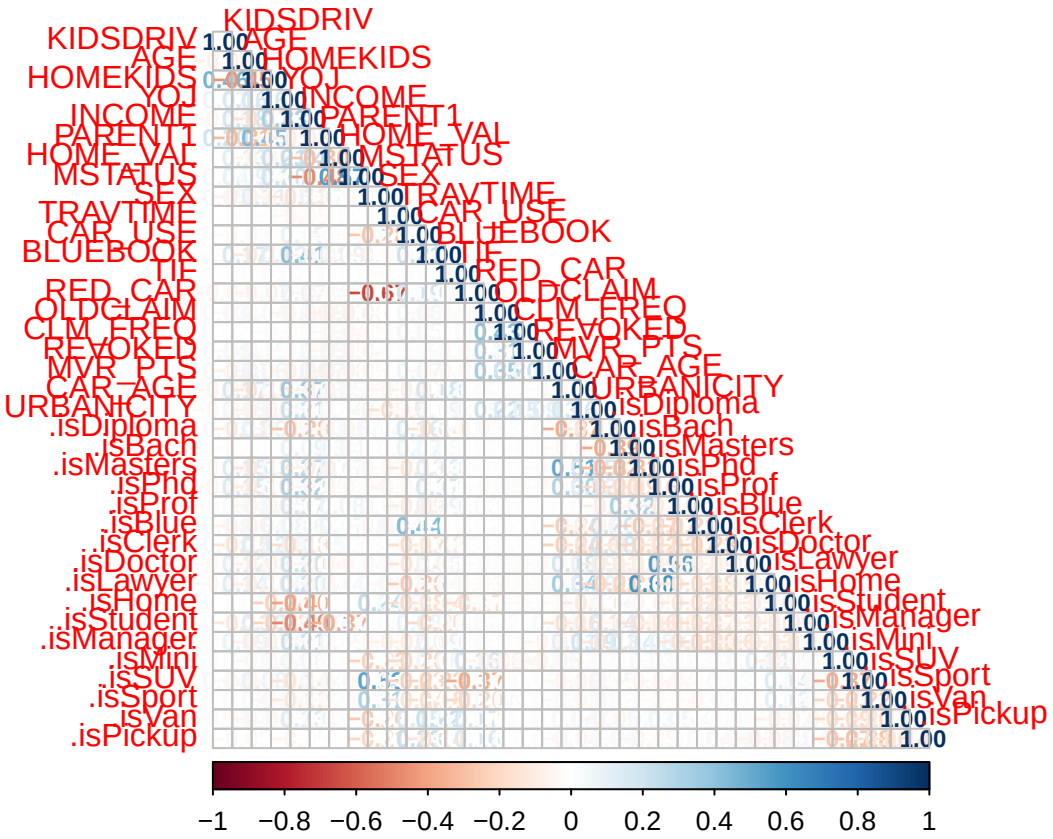

95%
log−Likelihood
−34000
−38000
−2
−1
0
1
2
λ


95%
log−Likelihood
−36000
−40000
−44000
−2
−1
0
1
2
λ



Checking for Multi-Collinearity Before any steps are taken in feature selection, the current features should be check for multi-collinearity.

Looking at the correlation plot below, there aren't any strong relationships between the variables below to re-



move pre feature selection.

Assigning Factors For the logistic model, Factoring the binary features for the model processing.

Build Models

Logisitic Model Building First it must be identified if the person was involved in a car crash. The response variable for the binary model is 'TARGET_FLAG'. Consequently, the linear model response variable 'TARGET_AMT' will be removed to avoid over-fitting in the logistic binary models. A total of three logistic binary models will be built.

Logistic Model 1 This model includes all transformed variables along with dummy values for categorical features. It'll be used as a baseline to compare it to other model building techniques that will be used in model 2 and model 3.

Positive variables As expected, the 'URBANICITY' variable is the highest positive coefficient for predicting a car crash. Urban cities tend to have higher traffic volume as well as more real estate development with potential to lead to more car crashes. Other variables are in line with a positive expectation such as 'REVOKED', '.isSport', 'CLM_FREQ', 'MVR_PTS'.

'HOMEKIDS' was a surprise. Theoretically, one would expect parents to be more likely to think about safety under the steering wheel. On the other hand, it can also be interpreted as added stress or perhaps passenger children being a potential distraction from the road. More analysis would have to be done to make a decisive conclusion.

Negative variables Red cars having a negative coefficient came as a surprise. It is a well known myth that red cars statistically have higher crashing rates. According to this model, that myth is debunked.

Significant values Several variables are **not** statistically significant. In this model 'AGE', 'YOJ', 'SEX', 'REDCAR', 'isStudent', 'OLDCLAIM', 'CAR_AGE', profession, and education level do not add value to the model. Its possible some are affected due to multicollinearity.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial, data = train.clean.binary)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4534  -0.7145  -0.4069   0.5983   3.1208
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.754e+00  3.808e-01  -7.232 4.76e-13 ***
## KIDSDRIV     4.038e-01  6.117e-02   6.602 4.07e-11 ***
## AGE         -1.874e-03  3.648e-03  -0.514 0.607390
## HOMEKIDS     4.243e-02  3.788e-02   1.120 0.262584
## YOJ          1.982e-05  4.947e-04   0.040 0.968045
## INCOME      -2.529e-03  4.878e-04  -5.184 2.17e-07 ***
## PARENT1      3.689e-01  1.097e-01   3.364 0.000767 ***
## HOME_VAL    -2.442e-02  6.797e-03  -3.593 0.000326 ***
## MSTATUS     -5.003e-01  8.605e-02  -5.814 6.08e-09 ***
## SEX         -7.776e-02  1.106e-01  -0.703 0.482144
## TRAVTIME     1.663e-01  2.083e-02   7.983 1.43e-15 ***
## CAR_USE      7.656e-01  9.178e-02   8.342  < 2e-16 ***
## BLUEBOOK    -5.585e-03  1.214e-03  -4.602 4.18e-06 ***
## TIF         -5.553e-02  7.336e-03  -7.570 3.73e-14 ***
## RED_CAR     -1.461e-02  8.636e-02  -0.169 0.865686
## OLDCLAIM     8.063e-04  6.231e-04   1.294 0.195691
## CLM_FREQ     7.885e-03  2.156e-03   3.658 0.000255 ***
## REVOKED      6.862e-01  8.572e-02   8.005 1.19e-15 ***
## MVR_PTS      1.767e-02  1.974e-03   8.954  < 2e-16 ***
## CAR_AGE     -1.320e-04  3.800e-04  -0.347 0.728311
## URBANICITY   2.438e+00  1.123e-01  21.717  < 2e-16 ***
## .isDiploma1  5.224e-02  9.550e-02   0.547 0.584345
## .isBach1    -3.380e-01  1.133e-01  -2.984 0.002846 **
## .isMasters1 -2.386e-01  1.774e-01  -1.345 0.178476
## .isPhd1     -1.520e-01  2.098e-01  -0.724 0.468787
## .isProf1     1.979e-01  1.776e-01   1.114 0.265209
## .isBlue1     3.504e-01  1.848e-01   1.896 0.057917 .
## .isClerk1    4.196e-01  1.962e-01   2.139 0.032414 *
## .isDoctor1  -4.121e-01  2.663e-01  -1.547 0.121781
## .isLawyer1   1.331e-01  1.690e-01   0.788 0.430765
## .isHome1     1.336e-01  2.171e-01   0.616 0.538147
## .isStudent1  6.507e-04  2.246e-01   0.003 0.997688
## .isManager1 -5.374e-01  1.706e-01  -3.150 0.001635 **
## .isMini1    -5.661e-01  1.564e-01  -3.620 0.000295 ***
## .isSUV1      2.029e-01  1.914e-01   1.060 0.289100
## .isSport1    4.399e-01  2.053e-01   2.142 0.032159 *
## .isVan1      7.744e-02  1.430e-01   0.541 0.588216
## .isPickup1  -2.443e-03  1.499e-01  -0.016 0.986996
## ---
## 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: 7307.2  on 8122  degrees of freedom
## AIC: 7383.2
##
## Number of Fisher Scoring iterations: 5
```

Explore removing highly Correlated variables

'.isSUV', '.isBlue', '.isClerk', and "OLDCLAIM" were found to have VIF values above 5. In other words, there is high multicollinearity present. A model was explored to verify if removal of these variables would improve the model. Removal did not improve the AIC score, as such the original model with all variables was selected as model 1.

```
##   KIDSDRIV        AGE   HOMEKIDS        YOJ     INCOME    PARENT1   HOME_VAL
##   1.348724   1.510259   2.264947   1.412259   3.129237   1.927962   1.822392
##    MSTATUS        SEX   TRAVTIME    CAR_USE   BLUEBOOK        TIF    RED_CAR
##   2.177767   3.610999   1.033886   2.453970   2.015107   1.009512   1.833973
##   OLDCLAIM   CLM_FREQ    REVOKED    MVR_PTS    CAR_AGE URBANICITY .isDiploma
##   1.440156   1.212069   1.140321   1.122624   1.961124   1.133621   2.347329
##    .isBach .isMasters     .isPhd    .isProf    .isBlue   .isClerk  .isDoctor
##   2.931604   5.639723   3.565197   4.251966   7.784217   6.175162   1.570274
##  .isLawyer    .isHome .isStudent .isManager    .isMini     .isSUV   .isSport
##   2.769606   4.114464   5.342176   2.762135   4.630472   9.126238   5.465586
##     .isVan  .isPickup
##   2.160924   4.073448
```

Updated Model fit 1 with High VIF variables removed

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial, data = train.clean.binary[,
##     names.include])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4265  -0.7166  -0.4072   0.6024   3.1138
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.146e+00  2.911e-01  -7.371 1.69e-13 ***
## KIDSDRIV     4.040e-01  6.112e-02   6.610 3.85e-11 ***
## AGE         -1.578e-03  3.628e-03  -0.435 0.663735
## HOMEKIDS     4.262e-02  3.786e-02   1.126 0.260314
## YOJ          8.137e-05  4.939e-04   0.165 0.869140
## INCOME      -2.730e-03  4.752e-04  -5.745 9.17e-09 ***
## PARENT1      3.695e-01  1.096e-01   3.372 0.000745 ***
## HOME_VAL    -2.438e-02  6.792e-03  -3.589 0.000331 ***
## MSTATUS     -5.018e-01  8.592e-02  -5.840 5.21e-09 ***
## SEX          4.655e-03  9.177e-02   0.051 0.959546
## TRAVTIME     1.668e-01  2.082e-02   8.011 1.13e-15 ***
## CAR_USE      7.223e-01  7.943e-02   9.094  < 2e-16 ***
```

```
## BLUEBOOK    -6.438e-03  1.010e-03  -6.374 1.84e-10 ***
## TIF         -5.546e-02  7.328e-03  -7.568 3.80e-14 ***
## RED_CAR     -1.739e-02  8.620e-02  -0.202 0.840082
## CLM_FREQ     8.859e-03  1.993e-03   4.445 8.80e-06 ***
## REVOKED      7.265e-01  8.039e-02   9.038  < 2e-16 ***
## MVR_PTS      1.838e-02  1.903e-03   9.656  < 2e-16 ***
## CAR_AGE     -1.118e-04  3.800e-04  -0.294 0.768517
## URBANICITY   2.440e+00  1.116e-01  21.868  < 2e-16 ***
## .isDiploma1  5.670e-02  9.458e-02   0.600 0.548834
## .isBach1    -3.538e-01  1.117e-01  -3.167 0.001543 **
## .isMasters1 -4.335e-01  1.509e-01  -2.872 0.004075 **
## .isPhd1     -3.787e-01  1.815e-01  -2.087 0.036886 *
## .isProf1    -1.199e-01  1.026e-01  -1.168 0.242850
## .isDoctor1  -5.366e-01  2.565e-01  -2.092 0.036436 *
## .isLawyer1  -3.245e-02  1.461e-01  -0.222 0.824279
## .isHome1    -2.317e-01  1.428e-01  -1.623 0.104688
## .isStudent1 -3.888e-01  1.326e-01  -2.932 0.003363 **
## .isManager1 -8.036e-01  1.170e-01  -6.866 6.61e-12 ***
## .isMini1    -6.826e-01  8.844e-02  -7.719 1.17e-14 ***
## .isSport1    2.549e-01  9.783e-02   2.606 0.009173 **
## .isVan1     -4.610e-03  1.145e-01  -0.040 0.967897
## .isPickup1  -1.238e-01  9.017e-02  -1.373 0.169707
## ---
## 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: 7315.2  on 8126  degrees of freedom
## AIC: 7383.2
##
## Number of Fisher Scoring iterations: 5
```

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1         9415.    8159 -3658. 7383. 7621.    7315.        8126  8160
```

Logistic Model 2 This model initially includes all transformed variables along with the dummy values generated for categorical variables. The stepwise method is used, which uses a loop to remove or add variables with the best influence on the AIC score. The recursive loop terminates once all the sequential steps are executed. Ultimately, the subset with the lowest AIC value is chosen as the result. Overall, this model chose a similar subset to model fit 1 with the exception of the non-statistically significant variables.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CLM_FREQ +
##     REVOKED + MVR_PTS + URBANICITY + .isBach + .isMasters + .isPhd +
##     .isBlue + .isClerk + .isDoctor + .isManager + .isMini + .isSUV +
##     .isSport, family = "binomial", data = train.clean.binary)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.4708  -0.7177  -0.4077   0.5964   3.1228
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.7107795  0.2198664 -12.329  < 2e-16 ***
## KIDSDRIV     0.4307864  0.0550240   7.829 4.91e-15 ***
## INCOME      -0.0023609  0.0003794  -6.223 4.88e-10 ***
## PARENT1      0.4466812  0.0943375   4.735 2.19e-06 ***
## HOME_VAL    -0.0224076  0.0063980  -3.502 0.000461 ***
## MSTATUS     -0.4854138  0.0803684  -6.040 1.54e-09 ***
## TRAVTIME     0.1652188  0.0207976   7.944 1.96e-15 ***
## CAR_USE      0.7374272  0.0795061   9.275  < 2e-16 ***
## BLUEBOOK    -0.0057348  0.0009860  -5.816 6.01e-09 ***
## TIF         -0.0555266  0.0073264  -7.579 3.48e-14 ***
## CLM_FREQ     0.0089288  0.0019916   4.483 7.35e-06 ***
## REVOKED      0.7312683  0.0803119   9.105  < 2e-16 ***
## MVR_PTS      0.0185250  0.0019005   9.747  < 2e-16 ***
## URBANICITY   2.4565054  0.1116330  22.005  < 2e-16 ***
## .isBach1    -0.3680445  0.0799157  -4.605 4.12e-06 ***
## .isMasters1 -0.3355855  0.1084098  -3.096 0.001965 **
## .isPhd1     -0.2934150  0.1572210  -1.866 0.062005 .
## .isBlue1     0.2382055  0.0944791   2.521 0.011694 *
## .isClerk1    0.3072306  0.0969939   3.168 0.001537 **
## .isDoctor1  -0.5026410  0.2494463  -2.015 0.043902 *
## .isManager1 -0.6607320  0.1103351  -5.988 2.12e-09 ***
## .isMini1    -0.5870978  0.0896610  -6.548 5.83e-11 ***
## .isSUV1      0.1395102  0.0840369   1.660 0.096893 .
## .isSport1    0.3762250  0.1065152   3.532 0.000412 ***
## ---
## 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: 7315.0  on 8136  degrees of freedom
## AIC: 7363
##
## Number of Fisher Scoring iterations: 5
```

Logistic Model 3 For model 3, the top 12 predictor variables found in model 1 and model 2 were handpicked to build a subset model. The model concludes that all the selected variables are statistically significant, however, the AIC value is higher than model 1 and model 2.

```
##
## Call:
## glm(formula = train.clean.binary[, var_subset], family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4780  -0.7273  -0.4261   0.6364   3.0104
##
## Coefficients:
```

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.1245610  0.1917082 -11.082  < 2e-16 ***
## URBANICITY   2.4016592  0.1107490  21.686  < 2e-16 ***
## REVOKED      0.7388683  0.0793343   9.313  < 2e-16 ***
## CAR_USE      0.7728949  0.0615070  12.566  < 2e-16 ***
## TRAVTIME     0.1615910  0.0205084   7.879 3.29e-15 ***
## TIF         -0.0540406  0.0072226  -7.482 7.31e-14 ***
## MVR_PTS      0.0204888  0.0018731  10.939  < 2e-16 ***
## .isMini1    -0.6644157  0.0728856  -9.116  < 2e-16 ***
## BLUEBOOK    -0.0072884  0.0009436  -7.724 1.12e-14 ***
## KIDSDRIV     0.5189763  0.0527051   9.847  < 2e-16 ***
## MSTATUS     -0.7592126  0.0582794 -13.027  < 2e-16 ***
## .isManager1 -0.7990406  0.1043327  -7.659 1.88e-14 ***
## INCOME      -0.0035852  0.0003023 -11.860  < 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: 9415.3  on 8159  degrees of freedom
## Residual deviance: 7460.0  on 8147  degrees of freedom
## AIC: 7486
##
## Number of Fisher Scoring iterations: 5
```

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1         9415.    8159 -3730. 7486. 7577.    7460.        8147  8160
```

Binary Model Selection Given the large amount of observations, the AIC criterion will be used for selection. Model 2 has the best fit among all three models with an AIC score of 7351.850. Model 2 also has a better BIC score (lower score). Lastly, model 2 has a higher McFadden Pseudo R^2 score than model 3 while maintaining a score only slightly below model 1.

Logistic Model Metrics Table

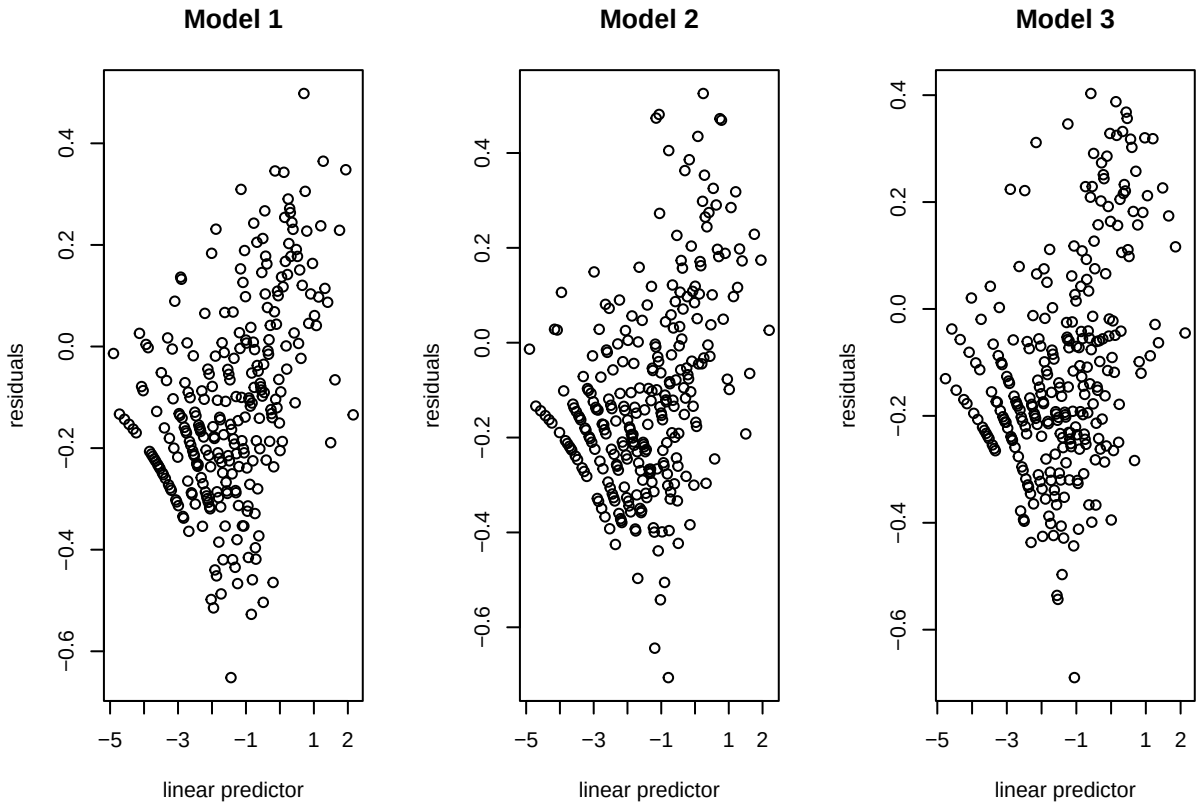
```
## fitting null model for pseudo-r2
```

```
## fitting null model for pseudo-r2
```

```
## fitting null model for pseudo-r2
```

model.build	null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs	McFaddens.R2
Bin model 1	9415.297	8159	-3653.620	7383.240	7649.506	7307.240	8122	8160	0.2238970
Bin model 2	9415.297	8159	-3657.519	7363.039	7531.207	7315.039	8136	8160	0.2230687
Bin model 3	9415.297	8159	-3730.002	7486.003	7577.094	7460.003	8147	8160	0.2076720

Residuals



Deviance (G^2)

The deviance is another measure of how well the model fits the data. Not a single model was able to reject the Null Hypothesis, therefore the deviance goodness-of-fit test confirms that all three models can be considered an adequate fit.

```
## [1] "Model 1 p-val is: 0.4979132%"
```

```
## [1] "Model 2 p-val is: 0.4979150%"
```

```
## [1] "Model 3 p-val is: 0.4979164%"
```

Homer-Lemeshow Goodness of Fit Test

The Homer-Lemeshow goodness of fit test is used to assess how a binary logistic regression model fits the observed data. Our results show that model 2 is a good fit. This is observed by the high p-value. The null hypothesis states that there is no difference between the observed and expected frequencies across the groups. In other words, the logistic regression model fits the data well.

```
library(performance)
#model1
performance_hosmer(fit1, n_bins = 272)
```

```
## # Hosmer-Lemeshow Goodness-of-Fit Test
##
##   Chi-squared: 292.791
```

```
##            df: 270
##       p-value:   0.163

## Summary: model seems to fit well.
```

```
#model2
performance_hosmer(fit2, n_bins = 272)
```

```
## # Hosmer-Lemeshow Goodness-of-Fit Test
##
##   Chi-squared: 312.897
##            df: 270
##       p-value:   0.037

## Summary: model does not fit well.
```

```
#model3
performance_hosmer(fit3, n_bins = 272)
```

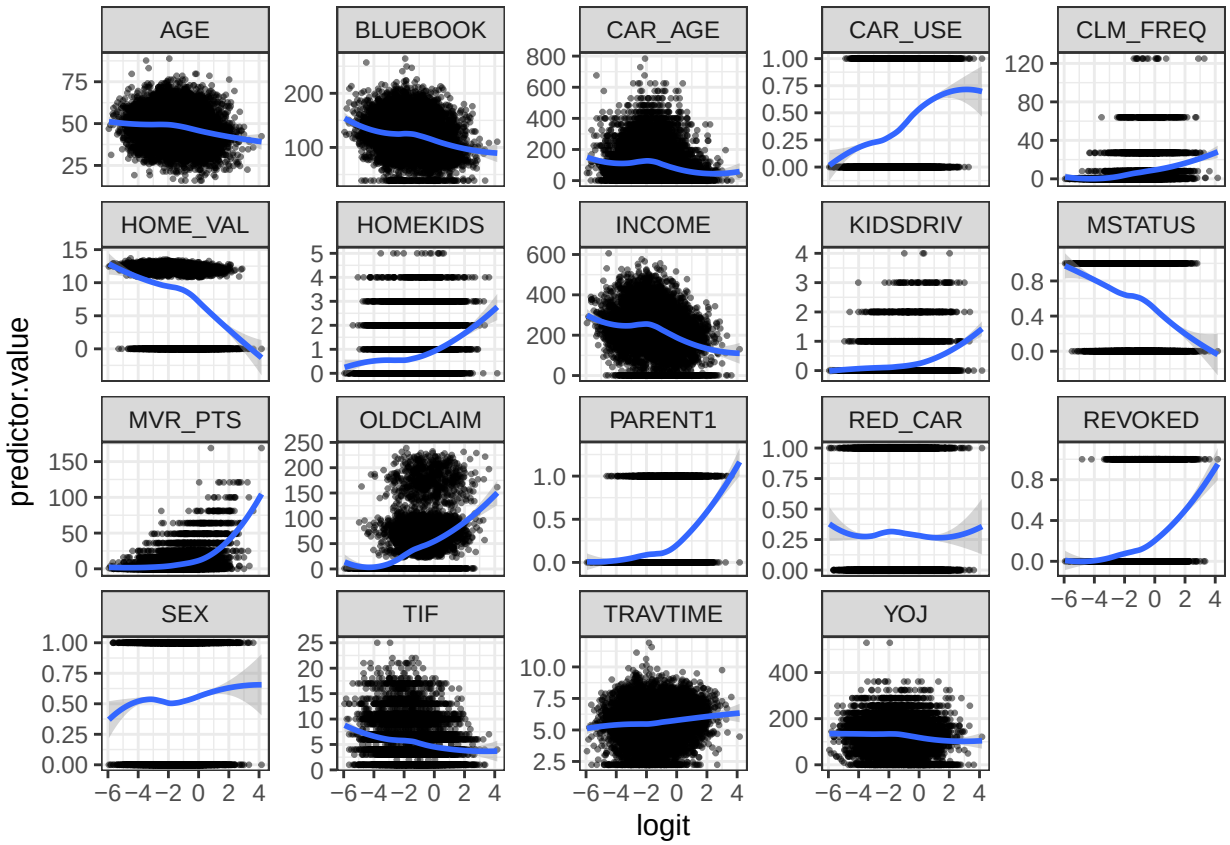
```
## # Hosmer-Lemeshow Goodness-of-Fit Test
##
##   Chi-squared: 322.051
##            df: 270
##       p-value:   0.016

## Summary: model does not fit well.
```

Binary Model Prediction This section covers using the selected model (fit2) to predict if a person was involved in an accident or not. Additional diagnostics were observed such as linear assumptions of residuals which show homoscedasticity in the numerical-type variables. A calibration plot also displays a successful agreement between predictions and observations. The model scored a 79.08% in terms of Classification Accuracy. However, specificity (True Negative Rate) are only at 42.57%, whereas Sensitivity is at 92.16% (True Positive Rate). This means the model performs well at predicting people who have been in a car crash. On the other hand, it lacks in the ability to predict people who were **not** involved in a car crash. An alternate score to consider is F1-score which takes a weighted calculation of the binary options. The F1-score is 86.64%. It should be noted the dataset is imbalanced, which may have caused the accuracy disruption of specificity.

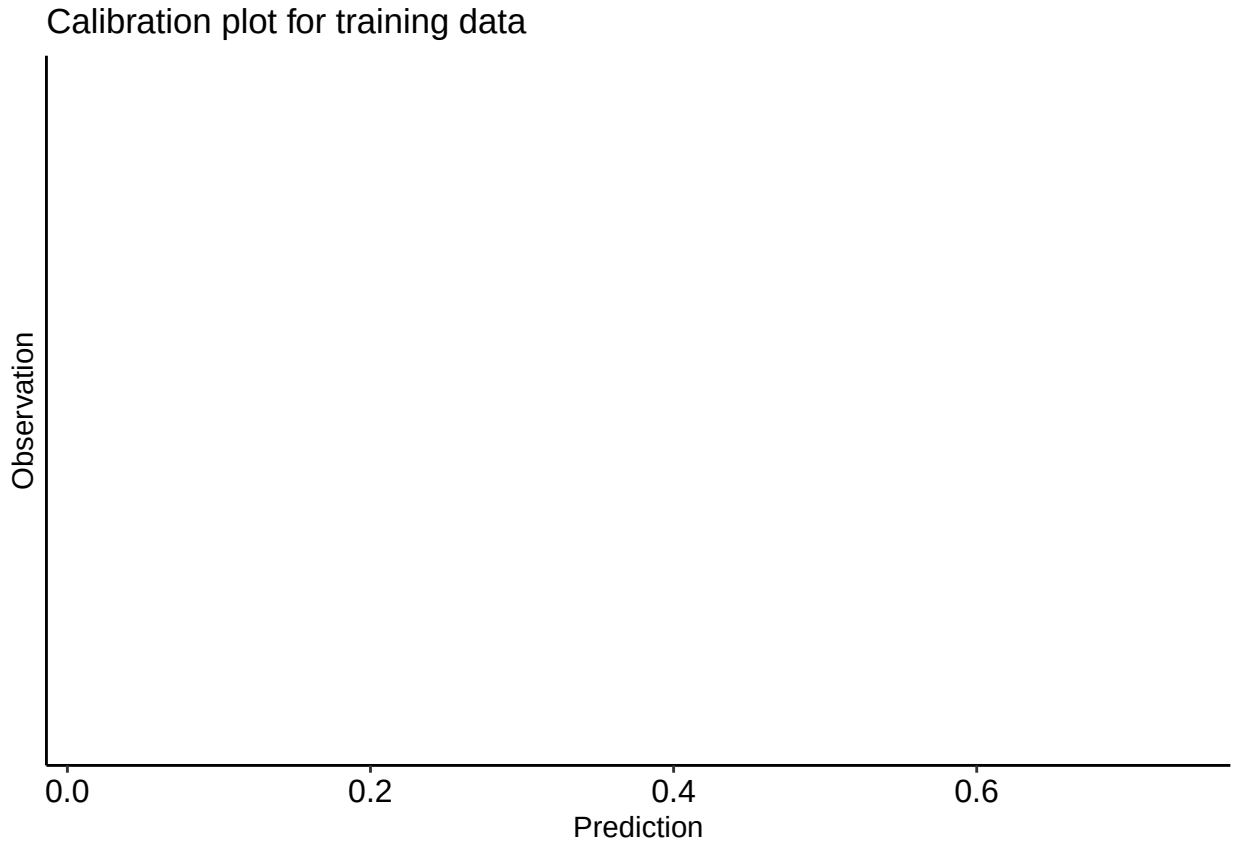
-Crosstab of traning dataset:

Var1	Freq
0	6008
1	2152

```
## `geom_smooth()` using formula = 'y ~ x'
```

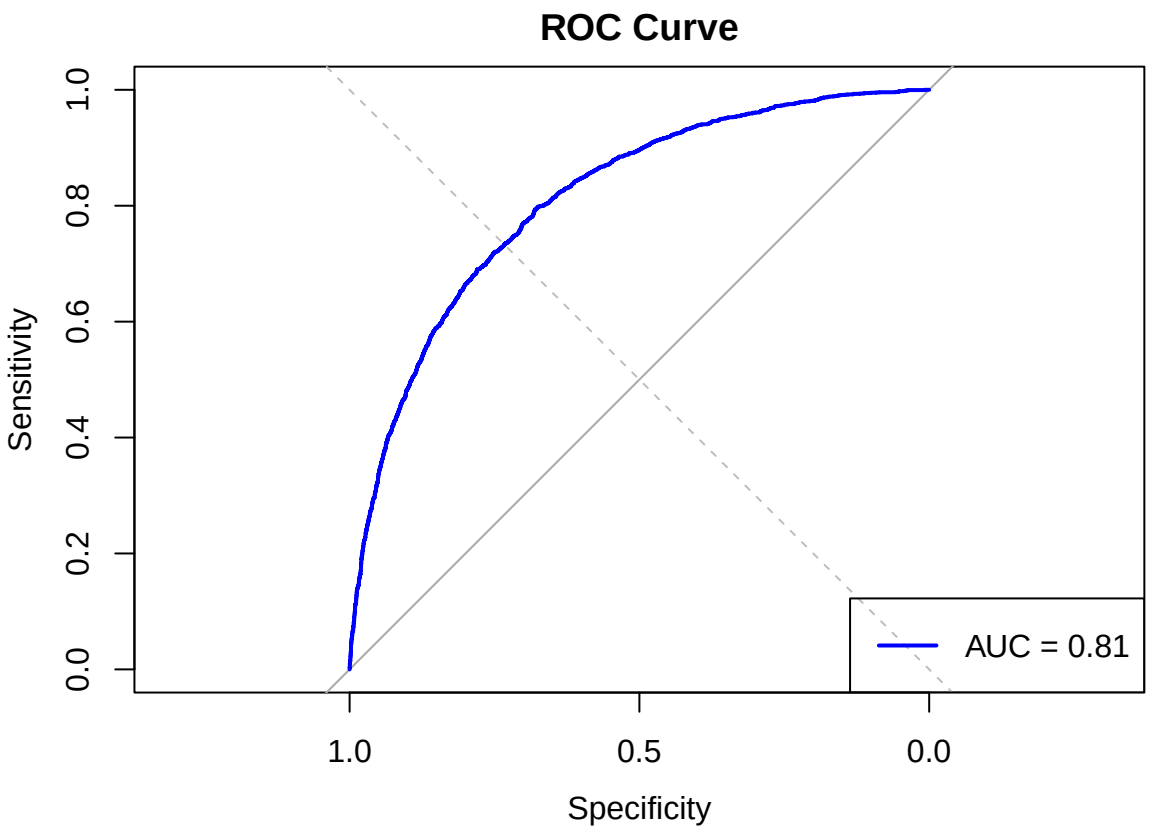



```
## Warning: There were 20 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘obsRate = mean(TARGET_FLAG/follow_up, na.rm = T)‘.
## i In group 1: ‘decile = 1‘.
## Caused by warning in ‘Ops.factor()‘:
## ! ’/’ not meaningful for factors
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 19 remaining warnings.

## $calibration_plot

## Warning: Removed 10 rows containing missing values (‘geom_point()‘).
```




```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

## [1] "Model 2 Classification Accuracy is: 79.26%"

## [1] "Model 2 Classification Error Rate is: 20.74%"

## [1] "Model 2 Precision is: 81.65%"

## [1] "Model 2 Sensitivity/Recall is: 92.66%"

## [1] "Model 2 Specificity is: 41.87%"

## [1] "Model 2 F1-score is: 86.81%"

## [1] "Model 2 AUC is: 81.21%"
```

ROC Curve



Making Predictions
The testing data set had a total of 2141 observations. Of those, 371 were classified as a car crash.

KID	BIRTH	HOME	YOJ	INCOME	PARENT	HOME	MSTATUS	SEX	EDUCATION	JOB	TRAVTIME	CAR	BLUEBOOK	TIF	CAR	OLDCLAIM	CLM	REVOKED	MVR	CLAIM	Pred	Probability
0	51.0	41	13	42229	0	96	0	0	0	5.0	0	0	14	8.2	22	81	1.0	0	0	4	1	0.1400396
1	42.4	44	16	32225	1	42	0	0	0	4.5	8	2	5	7	6	6	3	7.4	1	0	4	0.1917207
0	46.2	24	13	42208	1	53	5	0	0	5.4	1	7	7	2	6	8	0	1.0	0	0	0	0.1503186
0	36.2	64	8	51451	6	19	4	0	0	8.6	0	2	3	2	5	6	7	2	8	9	1.0	0.2026498
0	63.8	25	16	29507	3	8	0	0	0	6.7	0	8	2	0	4	1	1	7	2	9	1	0.1836501
0	49.0	6	8	8	2	1	0	5	1	6	2	4	2	9	8	2	6	4	5	7	5	0.1682588
0	64.9	5	7	4	1	9	4	7	8	4	5	1	1	5	8	2	4	0	0	0	0	0.4692444
0	58.0	7	5	4	4	1	8	2	0	2	4	4	3	9	7	3	0	9	5	1	9	0.1374513
2	37.2	7	8	5	8	6	1	3	0	1	4	6	7	4	8	9	7	2	2	3	6	0.1849192
0	53.0	6	6	6	8	4	0	9	2	3	1	0	0	0	1	0	4	6	9	0	4	0.1212699

Var1	Freq
0	1791
1	350

Linear Model Building Our goal for the linear model is to estimate the payout for those that were involved in a car crash. The variable 'TARGET_FLAG' will be removed to avoid over-fitting. Since we only want to look at those that have been in an accident, we'll also need to filter out individuals who are label as

not having been in an accident. Therefore, values under 'TARGET_AMT' that are either "0" or NA will be removed. A total of two models will be built.

#look at transformed models and see their distribution

Linear Model 1 Like the logistic model, the first model for linear model will contain all the variables. This will serve as a baseline to compare with the other models that we'll be creating.

When looking at the residuals, the range goes from a minimum of -8480 and a maximum of 99572. It also has a median of -1491. This tell us that is a wide range of variation between the observed values and the values predicted by the model.

Positive variables The '.isPhd' variable had the highest positive coefficient. This tells us that being a Phd student leads to a higher payout if there is an accident. If a sports car is involved in the accident, it will also lead to higher payouts. Other variables that lead to higher payouts also include '.Prof', '.isStudent', and 'HOMEKIDS'.

While all education variables had positive coefficients, we were surprised by the high coefficient for '.isPhd'. One might assume a higher level of education, might make a driver more conscious while on the road. One can also assume that they may drive a more expensive car, which may lead to more expensive repair/damage costs.

Negative variables

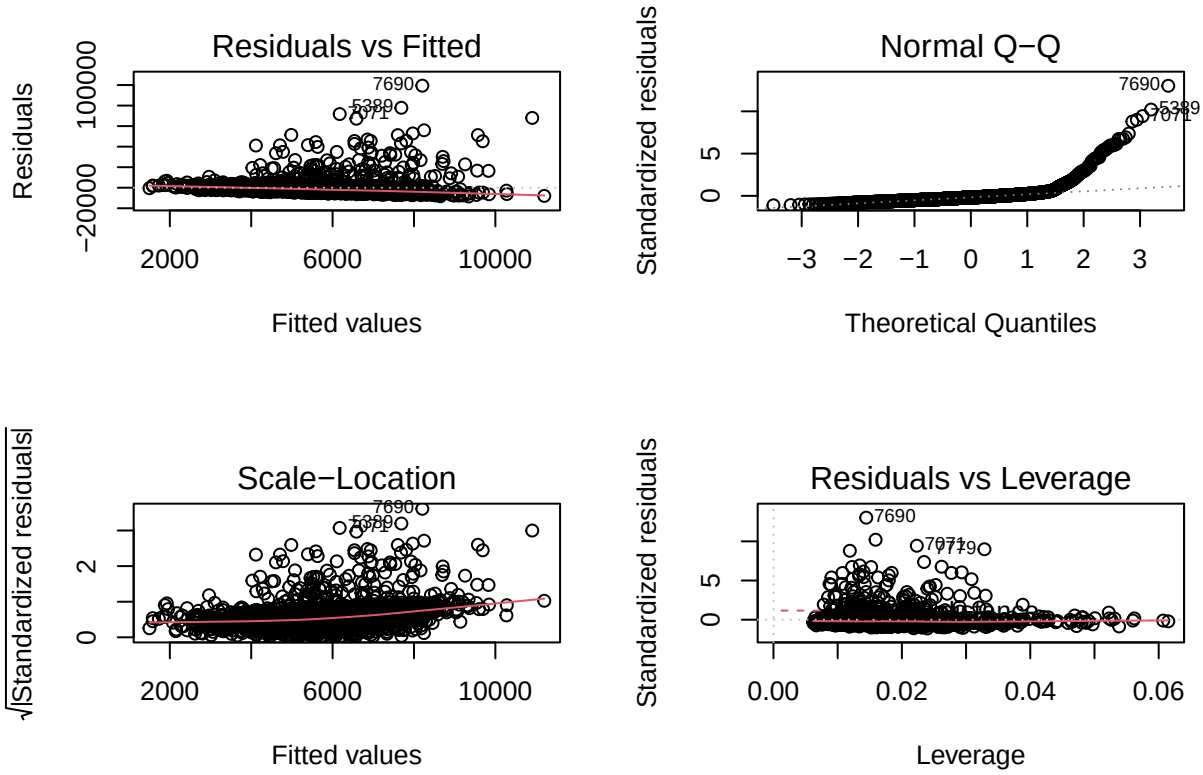
If an individual is a doctor, their payout is significantly lower, as evident by their negative coefficient score. Women also have a lower payout based on it's coefficient value.

Significant values This model only contains two significant values: "BLUEBOOK" and "SEX"

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = train.clean.linear)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8503  -3185  -1496    472  99381
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6337.1140 10823.6433  -0.585  0.55828
## KIDSDRIV     -181.5601   316.5666  -0.574  0.56635
## AGE            16.2482    19.2717   0.843  0.39926
## HOMEKIDS      206.3636   211.5307   0.976  0.32939
## YOJ             1.0952     2.8999   0.378  0.70573
## INCOME         -0.6674     2.9616  -0.225  0.82174
## PARENT1       313.2065   587.0826   0.533  0.59375
## HOME_VAL       56.6206    38.3612   1.476  0.14010
## MSTATUS      -942.2590   510.9492  -1.844  0.06530 .
## SEX         -1294.2921   646.6481  -2.002  0.04546 *
## TRAVTIME       -8.4978   121.4534  -0.070  0.94423
## CAR_USE       364.3421   521.3264   0.699  0.48471
## BLUEBOOK       29.1151     6.9101   4.213 2.62e-05 ***
## TIF           -15.7559    42.4414  -0.371  0.71050
## RED_CAR      -134.9885   496.4583  -0.272  0.78572
## OLDCLAIM        2.1743     3.5572   0.611  0.54110
## CLM_FREQ       -3.6895    11.5520  -0.319  0.74947
## REVOKED      -916.3780   461.5627  -1.985  0.04723 *
```

```
## MVR_PTS        11.6109     8.4332   1.377  0.16872
## CAR_AGE        -6.9755     2.3727  -2.940  0.00332 **
## URBANICITY     90.6472   754.1119   0.120  0.90433
## .isDiploma   -430.6307   515.3143  -0.836  0.40344
## .isBach       151.8017   625.3631   0.243  0.80823
## .isMasters   1381.6156  1086.0499   1.272  0.20346
## .isPhd       2465.4160  1288.8193   1.913  0.05589 .
## .isProf      1098.1206  1125.0530   0.976  0.32915
## .isBlue       571.8878  1139.8410   0.502  0.61591
## .isClerk      469.6616  1194.1695   0.393  0.69414
## .isDoctor   -2304.6053  1765.8633  -1.305  0.19201
## .isLawyer     354.2009  1028.0106   0.345  0.73047
## .isHome       219.5430  1292.9934   0.170  0.86519
## .isStudent    604.7964  1336.0206   0.453  0.65082
## .isManager   -786.5822  1064.6341  -0.739  0.46009
## .isMini       360.8506   920.6270   0.392  0.69513
## .isSUV       1205.0811  1107.2953   1.088  0.27658
## .isSport     1437.7792  1173.5581   1.225  0.22066
## .isVan        406.1267   831.2842   0.489  0.62521
## .isPickup     361.2560   864.2683   0.418  0.67600
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7685 on 2114 degrees of freedom
## Multiple R-squared:  0.03226,    Adjusted R-squared:  0.01532
## F-statistic: 1.905 on 37 and 2114 DF,  p-value: 0.000867
```

When looking at the residual plots, we can observe that there is linearity and homoscedasticity (constant variance of errors), however we can observe outliers, as well as a non-normal distribution of residuals as seen in the Q-Q plot.



Model 2 Model 2 will be created using the stepwise regression technique, specifically taking a "backward" approach. It essentially starts with a full model and then iteratively removes variables that are least significant until it arrives at an optimal model.

Compared to Model 1, this has more significant values at 3: MSTATUS, SEX, and BLUEBOOK

The range of the residuals is very similar. It has a minimum of -8364, a max of 100361 and Median of -1505. There is no significant improvement observed in the spread of the residuals.

The R squared of 0.02676 is lower than the R squared of Model 1(0.03079). The p value of 4.873e-08 is an improvement over Model 1 (0.001986)

```
lm2 <- stats::step(lm1,direction="backward")
```

```
## Start:  AIC=38545.59
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + RED_CAR + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY + .isDiploma + .isBach + .isMasters +
##     .isPhd + .isProf + .isBlue + .isClerk + .isDoctor + .isLawyer +
##     .isHome + .isStudent + .isManager + .isMini + .isSUV + .isSport +
##     .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - TRAVTIME    1     289114 1.2485e+11 38544
## - URBANICITY  1     853320 1.2485e+11 38544
```

```
## - .isHome     1    1702633 1.2485e+11 38544
## - INCOME      1    2998709 1.2485e+11 38544
## - .isBach     1    3479873 1.2485e+11 38544
## - RED_CAR     1    4366188 1.2485e+11 38544
## - CLM_FREQ    1    6024125 1.2485e+11 38544
## - .isLawyer   1    7010979 1.2485e+11 38544
## - TIF         1    8139234 1.2486e+11 38544
## - YOJ         1    8422761 1.2486e+11 38544
## - .isMini     1    9073231 1.2486e+11 38544
## - .isClerk    1    9135089 1.2486e+11 38544
## - .isPickup   1   10318287 1.2486e+11 38544
## - .isStudent  1   12102263 1.2486e+11 38544
## - .isVan      1   14096085 1.2486e+11 38544
## - .isBlue     1   14866463 1.2486e+11 38544
## - PARENT1     1   16808838 1.2486e+11 38544
## - KIDSDRIV    1   19426083 1.2487e+11 38544
## - OLDCLAIM    1   22065412 1.2487e+11 38544
## - CAR_USE     1   28845212 1.2488e+11 38544
## - .isManager  1   32237549 1.2488e+11 38544
## - .isDiploma  1   41242041 1.2489e+11 38544
## - AGE         1   41980232 1.2489e+11 38544
## - HOMEKIDS    1   56207476 1.2490e+11 38545
## - .isProf     1   56263752 1.2490e+11 38545
## - .isSUV      1   69948785 1.2492e+11 38545
## - .isSport    1   88644090 1.2494e+11 38545
## - .isMasters  1   95576118 1.2494e+11 38545
## - .isDoctor   1  100589699 1.2495e+11 38545
## - MVR_PTS     1  111949354 1.2496e+11 38546
## <none>                     1.2485e+11 38546
## - HOME_VAL    1  128658888 1.2498e+11 38546
## - MSTATUS     1  200844398 1.2505e+11 38547
## - .isPhd      1  216108038 1.2506e+11 38547
## - REVOKED     1  232788972 1.2508e+11 38548
## - SEX         1  236593672 1.2508e+11 38548
## - CAR_AGE     1  510434557 1.2536e+11 38552
## - BLUEBOOK    1 1048432047 1.2590e+11 38562
##
## Step:  AIC=38543.6
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + RED_CAR +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY +
##     .isDiploma + .isBach + .isMasters + .isPhd + .isProf + .isBlue +
##     .isClerk + .isDoctor + .isLawyer + .isHome + .isStudent +
##     .isManager + .isMini + .isSUV + .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - URBANICITY  1     929888 1.2485e+11 38542
## - .isHome     1    1646170 1.2485e+11 38542
## - INCOME      1    3061965 1.2485e+11 38542
## - .isBach     1    3452092 1.2485e+11 38542
## - RED_CAR     1    4401841 1.2485e+11 38542
## - CLM_FREQ    1    6122124 1.2485e+11 38542
## - .isLawyer   1    6968788 1.2485e+11 38542
## - TIF         1    8082922 1.2486e+11 38542
```

```
## - YOJ         1    8443774 1.2486e+11 38542
## - .isClerk    1    9058903 1.2486e+11 38542
## - .isMini     1    9079075 1.2486e+11 38542
## - .isPickup   1   10364902 1.2486e+11 38542
## - .isStudent  1   12060772 1.2486e+11 38542
## - .isVan      1   14187020 1.2486e+11 38542
## - .isBlue     1   14741687 1.2486e+11 38542
## - PARENT1     1   16939682 1.2486e+11 38542
## - KIDSDRIV    1   19342053 1.2487e+11 38542
## - OLDCLAIM    1   22141019 1.2487e+11 38542
## - CAR_USE     1   29160799 1.2488e+11 38542
## - .isManager  1   32341324 1.2488e+11 38542
## - .isDiploma  1   41200220 1.2489e+11 38542
## - AGE         1   41886366 1.2489e+11 38542
## - .isProf     1   56073178 1.2490e+11 38543
## - HOMEKIDS    1   56176901 1.2490e+11 38543
## - .isSUV      1   69861688 1.2492e+11 38543
## - .isSport    1   88727891 1.2494e+11 38543
## - .isMasters  1   95370843 1.2494e+11 38543
## - .isDoctor   1  100622169 1.2495e+11 38543
## - MVR_PTS     1  111744273 1.2496e+11 38544
## <none>                     1.2485e+11 38544
## - HOME_VAL    1  129044971 1.2498e+11 38544
## - MSTATUS     1  201305079 1.2505e+11 38545
## - .isPhd      1  215858300 1.2506e+11 38545
## - REVOKED     1  232755017 1.2508e+11 38546
## - SEX         1  236326093 1.2508e+11 38546
## - CAR_AGE     1  510350403 1.2536e+11 38550
## - BLUEBOOK    1 1048268120 1.2590e+11 38560
##
## Step:  AIC=38541.61
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + RED_CAR +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + .isDiploma +
##     .isBach + .isMasters + .isPhd + .isProf + .isBlue + .isClerk +
##     .isDoctor + .isLawyer + .isHome + .isStudent + .isManager +
##     .isMini + .isSUV + .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - .isHome     1    1622437 1.2485e+11 38540
## - INCOME      1    2982333 1.2485e+11 38540
## - .isBach     1    3418797 1.2485e+11 38540
## - RED_CAR     1    4388476 1.2485e+11 38540
## - CLM_FREQ    1    5998402 1.2485e+11 38540
## - .isLawyer   1    6870091 1.2486e+11 38540
## - TIF         1    7962981 1.2486e+11 38540
## - YOJ         1    8443800 1.2486e+11 38540
## - .isClerk    1    8811863 1.2486e+11 38540
## - .isMini     1    8951564 1.2486e+11 38540
## - .isPickup   1   10274801 1.2486e+11 38540
## - .isStudent  1   11694868 1.2486e+11 38540
## - .isVan      1   14107594 1.2486e+11 38540
## - .isBlue     1   14614051 1.2486e+11 38540
## - PARENT1     1   16908810 1.2487e+11 38540
```

```
## - KIDSDRIV    1   19574431 1.2487e+11 38540
## - OLDCLAIM    1   22187677 1.2487e+11 38540
## - CAR_USE     1   29188824 1.2488e+11 38540
## - .isManager  1   32227420 1.2488e+11 38540
## - .isDiploma  1   41355339 1.2489e+11 38540
## - AGE         1   41483512 1.2489e+11 38540
## - .isProf     1   55686160 1.2490e+11 38541
## - HOMEKIDS    1   55924775 1.2490e+11 38541
## - .isSUV      1   69597597 1.2492e+11 38541
## - .isSport    1   88395070 1.2494e+11 38541
## - .isMasters  1   95194655 1.2494e+11 38541
## - .isDoctor   1  100456544 1.2495e+11 38541
## - MVR_PTS     1  112438549 1.2496e+11 38542
## <none>                     1.2485e+11 38542
## - HOME_VAL    1  128402103 1.2498e+11 38542
## - MSTATUS     1  200375302 1.2505e+11 38543
## - .isPhd      1  215500860 1.2506e+11 38543
## - REVOKED     1  232191811 1.2508e+11 38544
## - SEX         1  235760555 1.2508e+11 38544
## - CAR_AGE     1  510025646 1.2536e+11 38548
## - BLUEBOOK    1 1049204261 1.2590e+11 38558
##
## Step:  AIC=38539.64
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + RED_CAR +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + .isDiploma +
##     .isBach + .isMasters + .isPhd + .isProf + .isBlue + .isClerk +
##     .isDoctor + .isLawyer + .isStudent + .isManager + .isMini +
##     .isSUV + .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - .isBach     1    3618238 1.2485e+11 38538
## - RED_CAR     1    4595783 1.2485e+11 38538
## - .isLawyer   1    5249367 1.2486e+11 38538
## - CLM_FREQ    1    5858135 1.2486e+11 38538
## - INCOME      1    5957189 1.2486e+11 38538
## - YOJ         1    7897762 1.2486e+11 38538
## - TIF         1    8130959 1.2486e+11 38538
## - .isMini     1    9271092 1.2486e+11 38538
## - .isClerk    1    9528619 1.2486e+11 38538
## - .isPickup   1   10833630 1.2486e+11 38538
## - .isVan      1   14400257 1.2486e+11 38538
## - .isStudent  1   15457255 1.2487e+11 38538
## - PARENT1     1   16579752 1.2487e+11 38538
## - .isBlue     1   17293224 1.2487e+11 38538
## - KIDSDRIV    1   19539036 1.2487e+11 38538
## - OLDCLAIM    1   21926441 1.2487e+11 38538
## - CAR_USE     1   27968306 1.2488e+11 38538
## - .isDiploma  1   40651927 1.2489e+11 38538
## - AGE         1   42325558 1.2489e+11 38538
## - HOMEKIDS    1   56547862 1.2491e+11 38539
## - .isManager  1   62802592 1.2491e+11 38539
## - .isSUV      1   70796648 1.2492e+11 38539
## - .isProf     1   82089257 1.2493e+11 38539
```

```
## - .isSport    1   89508251 1.2494e+11 38539
## - .isMasters  1   97007668 1.2495e+11 38539
## - MVR_PTS     1  111636121 1.2496e+11 38540
## - .isDoctor   1  114655236 1.2496e+11 38540
## <none>                     1.2485e+11 38540
## - HOME_VAL    1  128879814 1.2498e+11 38540
## - MSTATUS     1  200595286 1.2505e+11 38541
## - .isPhd      1  218404710 1.2507e+11 38541
## - REVOKED     1  232881345 1.2508e+11 38542
## - SEX         1  234284181 1.2508e+11 38542
## - CAR_AGE     1  508829323 1.2536e+11 38546
## - BLUEBOOK    1 1051920007 1.2590e+11 38556
##
## Step:  AIC=38537.7
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + RED_CAR +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + .isDiploma +
##     .isMasters + .isPhd + .isProf + .isBlue + .isClerk + .isDoctor +
##     .isLawyer + .isStudent + .isManager + .isMini + .isSUV +
##     .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - INCOME      1    4449778 1.2486e+11 38536
## - RED_CAR     1    4495931 1.2486e+11 38536
## - CLM_FREQ    1    5646982 1.2486e+11 38536
## - .isLawyer   1    5752440 1.2486e+11 38536
## - YOJ         1    7181245 1.2486e+11 38536
## - .isClerk    1    7670095 1.2486e+11 38536
## - TIF         1    8175452 1.2486e+11 38536
## - .isMini     1   10717877 1.2486e+11 38536
## - .isPickup   1   11332937 1.2487e+11 38536
## - .isStudent  1   13964079 1.2487e+11 38536
## - .isBlue     1   14781441 1.2487e+11 38536
## - .isVan      1   15598062 1.2487e+11 38536
## - PARENT1     1   16681451 1.2487e+11 38536
## - KIDSDRIV    1   19840911 1.2487e+11 38536
## - OLDCLAIM    1   21734872 1.2488e+11 38536
## - CAR_USE     1   34715060 1.2489e+11 38536
## - AGE         1   42057477 1.2490e+11 38536
## - HOMEKIDS    1   56764985 1.2491e+11 38537
## - .isManager  1   61403508 1.2492e+11 38537
## - .isSUV      1   75539360 1.2493e+11 38537
## - .isProf     1   85487902 1.2494e+11 38537
## - .isSport    1   95212684 1.2495e+11 38537
## - .isDiploma  1   98391341 1.2495e+11 38537
## - MVR_PTS     1  112020297 1.2497e+11 38538
## - .isDoctor   1  113733841 1.2497e+11 38538
## - .isMasters  1  114912535 1.2497e+11 38538
## <none>                     1.2485e+11 38538
## - HOME_VAL    1  128771511 1.2498e+11 38538
## - MSTATUS     1  201359242 1.2506e+11 38539
## - REVOKED     1  231477165 1.2509e+11 38540
## - SEX         1  236157015 1.2509e+11 38540
## - .isPhd      1  254005096 1.2511e+11 38540
```

```
## - CAR_AGE     1  520911189 1.2537e+11 38545
## - BLUEBOOK    1 1054866800 1.2591e+11 38554
##
## Step:  AIC=38535.78
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + RED_CAR + OLDCLAIM +
##     CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters +
##     .isPhd + .isProf + .isBlue + .isClerk + .isDoctor + .isLawyer +
##     .isStudent + .isManager + .isMini + .isSUV + .isSport + .isVan +
##     .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - .isLawyer   1    4219348 1.2486e+11 38534
## - RED_CAR     1    4634540 1.2486e+11 38534
## - YOJ         1    4644631 1.2486e+11 38534
## - .isClerk    1    5346748 1.2486e+11 38534
## - CLM_FREQ    1    5422272 1.2486e+11 38534
## - TIF         1    7963925 1.2487e+11 38534
## - .isBlue     1   10963948 1.2487e+11 38534
## - .isMini     1   11725356 1.2487e+11 38534
## - .isPickup   1   12546448 1.2487e+11 38534
## - .isStudent  1   15874499 1.2487e+11 38534
## - .isVan      1   16019633 1.2487e+11 38534
## - PARENT1     1   16799043 1.2488e+11 38534
## - KIDSDRIV    1   20795192 1.2488e+11 38534
## - OLDCLAIM    1   21480400 1.2488e+11 38534
## - CAR_USE     1   32352018 1.2489e+11 38534
## - AGE         1   43745367 1.2490e+11 38535
## - HOMEKIDS    1   59572066 1.2492e+11 38535
## - .isSUV      1   77218702 1.2494e+11 38535
## - .isManager  1   80712085 1.2494e+11 38535
## - .isProf     1   83637586 1.2494e+11 38535
## - .isDiploma  1   97241695 1.2496e+11 38535
## - .isSport    1   97898059 1.2496e+11 38535
## - .isMasters  1  110838148 1.2497e+11 38536
## - MVR_PTS     1  114955108 1.2497e+11 38536
## <none>                     1.2486e+11 38536
## - .isDoctor   1  121263212 1.2498e+11 38536
## - HOME_VAL    1  127189727 1.2499e+11 38536
## - MSTATUS     1  197786997 1.2506e+11 38537
## - REVOKED     1  229476095 1.2509e+11 38538
## - SEX         1  235160773 1.2509e+11 38538
## - .isPhd      1  259659712 1.2512e+11 38538
## - CAR_AGE     1  541904209 1.2540e+11 38543
## - BLUEBOOK    1 1059461632 1.2592e+11 38552
##
## Step:  AIC=38533.85
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + RED_CAR + OLDCLAIM +
##     CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters +
##     .isPhd + .isProf + .isBlue + .isClerk + .isDoctor + .isStudent +
##     .isManager + .isMini + .isSUV + .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
```

```
## - .isClerk    1    4000099 1.2487e+11 38532
## - RED_CAR     1    4756139 1.2487e+11 38532
## - YOJ         1    5777690 1.2487e+11 38532
## - CLM_FREQ    1    5984425 1.2487e+11 38532
## - TIF         1    8197512 1.2487e+11 38532
## - .isBlue     1    9846288 1.2487e+11 38532
## - .isMini     1   13965437 1.2488e+11 38532
## - .isStudent  1   14833324 1.2488e+11 38532
## - .isPickup   1   14874114 1.2488e+11 38532
## - .isVan      1   17260889 1.2488e+11 38532
## - PARENT1     1   17300687 1.2488e+11 38532
## - KIDSDRIV    1   21051933 1.2488e+11 38532
## - OLDCLAIM    1   21329704 1.2488e+11 38532
## - CAR_USE     1   29034265 1.2489e+11 38532
## - AGE         1   44273900 1.2491e+11 38533
## - HOMEKIDS    1   58704788 1.2492e+11 38533
## - .isProf     1   79603278 1.2494e+11 38533
## - .isSUV      1   82311407 1.2494e+11 38533
## - .isDiploma  1   96388662 1.2496e+11 38534
## - .isManager  1  101355757 1.2496e+11 38534
## - .isSport    1  103931763 1.2497e+11 38534
## <none>                     1.2486e+11 38534
## - MVR_PTS     1  116264667 1.2498e+11 38534
## - HOME_VAL    1  127298500 1.2499e+11 38534
## - .isDoctor   1  132800126 1.2500e+11 38534
## - .isMasters  1  145129606 1.2501e+11 38534
## - MSTATUS     1  199076634 1.2506e+11 38535
## - REVOKED     1  228336733 1.2509e+11 38536
## - SEX         1  237671644 1.2510e+11 38536
## - .isPhd      1  269556109 1.2513e+11 38536
## - CAR_AGE     1  538058861 1.2540e+11 38541
## - BLUEBOOK    1 1079648667 1.2594e+11 38550
##
## Step:  AIC=38531.92
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + RED_CAR + OLDCLAIM +
##     CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters +
##     .isPhd + .isProf + .isBlue + .isDoctor + .isStudent + .isManager +
##     .isMini + .isSUV + .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - RED_CAR     1    5017257 1.2487e+11 38530
## - CLM_FREQ    1    5692206 1.2487e+11 38530
## - .isBlue     1    5893993 1.2487e+11 38530
## - TIF         1    7829000 1.2487e+11 38530
## - YOJ         1    8819114 1.2488e+11 38530
## - .isStudent  1   10866054 1.2488e+11 38530
## - .isMini     1   14788945 1.2488e+11 38530
## - .isPickup   1   15733504 1.2488e+11 38530
## - PARENT1     1   17120340 1.2488e+11 38530
## - .isVan      1   17394906 1.2488e+11 38530
## - KIDSDRIV    1   20638006 1.2489e+11 38530
## - OLDCLAIM    1   21111740 1.2489e+11 38530
## - CAR_USE     1   29002607 1.2490e+11 38530
```

```
## - AGE         1   41264251 1.2491e+11 38531
## - HOMEKIDS    1   56902129 1.2492e+11 38531
## - .isSUV      1   83960273 1.2495e+11 38531
## - .isProf     1   92182903 1.2496e+11 38532
## - .isDiploma  1   95206286 1.2496e+11 38532
## - .isSport    1  105950794 1.2497e+11 38532
## <none>                     1.2487e+11 38532
## - MVR_PTS     1  117439041 1.2498e+11 38532
## - HOME_VAL    1  127147657 1.2499e+11 38532
## - .isDoctor   1  135298787 1.2500e+11 38532
## - .isManager  1  135882847 1.2500e+11 38532
## - .isMasters  1  162047656 1.2503e+11 38533
## - MSTATUS     1  200333823 1.2507e+11 38533
## - REVOKED     1  227430248 1.2509e+11 38534
## - SEX         1  247072426 1.2511e+11 38534
## - .isPhd      1  283913584 1.2515e+11 38535
## - CAR_AGE     1  544963996 1.2541e+11 38539
## - BLUEBOOK    1 1085800885 1.2595e+11 38549
##
## Step:  AIC=38530.01
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters + .isPhd +
##     .isProf + .isBlue + .isDoctor + .isStudent + .isManager +
##     .isMini + .isSUV + .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - .isBlue     1    5725660 1.2488e+11 38528
## - CLM_FREQ    1    5981260 1.2488e+11 38528
## - TIF         1    7693059 1.2488e+11 38528
## - YOJ         1    9036372 1.2488e+11 38528
## - .isStudent  1   10420822 1.2488e+11 38528
## - .isMini     1   15500290 1.2489e+11 38528
## - .isPickup   1   16357369 1.2489e+11 38528
## - PARENT1     1   17206048 1.2489e+11 38528
## - .isVan      1   17884606 1.2489e+11 38528
## - KIDSDRIV    1   20255940 1.2489e+11 38528
## - OLDCLAIM    1   21060974 1.2489e+11 38528
## - CAR_USE     1   29171025 1.2490e+11 38529
## - AGE         1   42216667 1.2491e+11 38529
## - HOMEKIDS    1   56860848 1.2493e+11 38529
## - .isSUV      1   86386650 1.2496e+11 38529
## - .isProf     1   92023254 1.2496e+11 38530
## - .isDiploma  1   94542021 1.2497e+11 38530
## - .isSport    1  107894392 1.2498e+11 38530
## <none>                     1.2487e+11 38530
## - MVR_PTS     1  116246847 1.2499e+11 38530
## - HOME_VAL    1  127373843 1.2500e+11 38530
## - .isDoctor   1  135523046 1.2501e+11 38530
## - .isManager  1  137145055 1.2501e+11 38530
## - .isMasters  1  162217897 1.2503e+11 38531
## - MSTATUS     1  200571308 1.2507e+11 38531
## - REVOKED     1  226635495 1.2510e+11 38532
## - SEX         1  268120151 1.2514e+11 38533
```

```
## - .isPhd      1  285349122 1.2516e+11 38533
## - CAR_AGE     1  549657795 1.2542e+11 38537
## - BLUEBOOK    1 1089661393 1.2596e+11 38547
##
## Step:  AIC=38528.11
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters + .isPhd +
##     .isProf + .isDoctor + .isStudent + .isManager + .isMini +
##     .isSUV + .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - CLM_FREQ    1    5651821 1.2488e+11 38526
## - .isStudent  1    6130536 1.2488e+11 38526
## - TIF         1    8006636 1.2489e+11 38526
## - YOJ         1   10948152 1.2489e+11 38526
## - PARENT1     1   17067729 1.2489e+11 38526
## - .isPickup   1   18586914 1.2490e+11 38526
## - KIDSDRIV    1   19450360 1.2490e+11 38526
## - .isVan      1   19936699 1.2490e+11 38526
## - OLDCLAIM    1   20113019 1.2490e+11 38526
## - .isMini     1   20326539 1.2490e+11 38526
## - AGE         1   42261202 1.2492e+11 38527
## - HOMEKIDS    1   55134907 1.2493e+11 38527
## - CAR_USE     1   57210830 1.2493e+11 38527
## - .isProf     1   87879168 1.2497e+11 38528
## - .isDiploma  1   96130793 1.2497e+11 38528
## - .isSUV      1   97357713 1.2497e+11 38528
## - MVR_PTS     1  115566961 1.2499e+11 38528
## <none>                     1.2488e+11 38528
## - .isSport    1  120651152 1.2500e+11 38528
## - HOME_VAL    1  128133774 1.2501e+11 38528
## - .isDoctor   1  134530112 1.2501e+11 38528
## - .isManager  1  153089088 1.2503e+11 38529
## - .isMasters  1  158773253 1.2504e+11 38529
## - MSTATUS     1  201471325 1.2508e+11 38530
## - REVOKED     1  225637894 1.2510e+11 38530
## - SEX         1  275190921 1.2515e+11 38531
## - .isPhd      1  282494913 1.2516e+11 38531
## - CAR_AGE     1  549617165 1.2543e+11 38536
## - BLUEBOOK    1 1096572360 1.2597e+11 38545
##
## Step:  AIC=38526.21
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + OLDCLAIM + REVOKED +
##     MVR_PTS + CAR_AGE + .isDiploma + .isMasters + .isPhd + .isProf +
##     .isDoctor + .isStudent + .isManager + .isMini + .isSUV +
##     .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - .isStudent  1    6395565 1.2489e+11 38524
## - TIF         1    8243890 1.2489e+11 38524
## - YOJ         1   11569508 1.2489e+11 38524
## - OLDCLAIM    1   15108118 1.2490e+11 38524
```

```
## - PARENT1     1   16781188 1.2490e+11 38524
## - .isPickup   1   18715055 1.2490e+11 38525
## - KIDSDRIV    1   19617229 1.2490e+11 38525
## - .isVan      1   20581831 1.2490e+11 38525
## - .isMini     1   20585603 1.2490e+11 38525
## - AGE         1   41326915 1.2492e+11 38525
## - HOMEKIDS    1   54788361 1.2494e+11 38525
## - CAR_USE     1   56546368 1.2494e+11 38525
## - .isProf     1   87394792 1.2497e+11 38526
## - .isDiploma  1   96107005 1.2498e+11 38526
## - .isSUV      1   97841620 1.2498e+11 38526
## - MVR_PTS     1  113914052 1.2500e+11 38526
## <none>                     1.2488e+11 38526
## - .isSport    1  120942076 1.2500e+11 38526
## - HOME_VAL    1  128261875 1.2501e+11 38526
## - .isDoctor   1  141169255 1.2502e+11 38527
## - .isManager  1  152379603 1.2504e+11 38527
## - .isMasters  1  159190365 1.2504e+11 38527
## - MSTATUS     1  201836456 1.2508e+11 38528
## - REVOKED     1  219994579 1.2510e+11 38528
## - SEX         1  275195244 1.2516e+11 38529
## - .isPhd      1  281590091 1.2516e+11 38529
## - CAR_AGE     1  552405248 1.2544e+11 38534
## - BLUEBOOK    1 1098830921 1.2598e+11 38543
##
## Step:  AIC=38524.32
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + OLDCLAIM + REVOKED +
##     MVR_PTS + CAR_AGE + .isDiploma + .isMasters + .isPhd + .isProf +
##     .isDoctor + .isManager + .isMini + .isSUV + .isSport + .isVan +
##     .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - YOJ         1    7933612 1.2490e+11 38522
## - TIF         1    8638071 1.2490e+11 38522
## - OLDCLAIM    1   14959473 1.2490e+11 38523
## - PARENT1     1   16565732 1.2491e+11 38523
## - .isPickup   1   18587572 1.2491e+11 38523
## - .isVan      1   20051950 1.2491e+11 38523
## - KIDSDRIV    1   20104993 1.2491e+11 38523
## - .isMini     1   20722554 1.2491e+11 38523
## - AGE         1   43634187 1.2493e+11 38523
## - HOMEKIDS    1   59605548 1.2495e+11 38523
## - CAR_USE     1   60386022 1.2495e+11 38523
## - .isProf     1   83793388 1.2497e+11 38524
## - .isDiploma  1   95694740 1.2499e+11 38524
## - .isSUV      1   97467812 1.2499e+11 38524
## - MVR_PTS     1  112826459 1.2500e+11 38524
## <none>                     1.2489e+11 38524
## - .isSport    1  120612518 1.2501e+11 38524
## - HOME_VAL    1  126659105 1.2502e+11 38524
## - .isDoctor   1  140176733 1.2503e+11 38525
## - .isManager  1  154583994 1.2504e+11 38525
## - .isMasters  1  155706911 1.2505e+11 38525
```

```
## - MSTATUS     1  196979113 1.2509e+11 38526
## - REVOKED     1  220009727 1.2511e+11 38526
## - SEX         1  277235269 1.2517e+11 38527
## - .isPhd      1  278313289 1.2517e+11 38527
## - CAR_AGE     1  554893827 1.2544e+11 38532
## - BLUEBOOK    1 1092815013 1.2598e+11 38541
##
## Step:  AIC=38522.45
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + CAR_USE + BLUEBOOK + TIF + OLDCLAIM + REVOKED +
##     MVR_PTS + CAR_AGE + .isDiploma + .isMasters + .isPhd + .isProf +
##     .isDoctor + .isManager + .isMini + .isSUV + .isSport + .isVan +
##     .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - TIF         1    8740777 1.2491e+11 38521
## - OLDCLAIM    1   15311860 1.2491e+11 38521
## - PARENT1     1   16461861 1.2491e+11 38521
## - .isPickup   1   20420684 1.2492e+11 38521
## - .isVan      1   20977659 1.2492e+11 38521
## - KIDSDRIV    1   21463120 1.2492e+11 38521
## - .isMini     1   22839712 1.2492e+11 38521
## - AGE         1   53033104 1.2495e+11 38521
## - CAR_USE     1   61050879 1.2496e+11 38522
## - HOMEKIDS    1   73081591 1.2497e+11 38522
## - .isProf     1   86354448 1.2498e+11 38522
## - .isDiploma  1   96014172 1.2499e+11 38522
## - .isSUV      1  102092520 1.2500e+11 38522
## - MVR_PTS     1  112206293 1.2501e+11 38522
## <none>                     1.2490e+11 38522
## - .isSport    1  123034139 1.2502e+11 38523
## - .isDoctor   1  137912277 1.2504e+11 38523
## - HOME_VAL    1  142827578 1.2504e+11 38523
## - .isManager  1  150473290 1.2505e+11 38523
## - .isMasters  1  159815369 1.2506e+11 38523
## - MSTATUS     1  194655589 1.2509e+11 38524
## - REVOKED     1  217059246 1.2511e+11 38524
## - .isPhd      1  278433899 1.2518e+11 38525
## - SEX         1  283563386 1.2518e+11 38525
## - CAR_AGE     1  559664254 1.2546e+11 38530
## - BLUEBOOK    1 1132208420 1.2603e+11 38540
##
## Step:  AIC=38520.6
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + CAR_USE + BLUEBOOK + OLDCLAIM + REVOKED +
##     MVR_PTS + CAR_AGE + .isDiploma + .isMasters + .isPhd + .isProf +
##     .isDoctor + .isManager + .isMini + .isSUV + .isSport + .isVan +
##     .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - OLDCLAIM    1   15246337 1.2492e+11 38519
## - PARENT1     1   16949094 1.2492e+11 38519
## - .isPickup   1   20130785 1.2493e+11 38519
## - .isVan      1   20751795 1.2493e+11 38519
```

```
## - KIDSDRIV    1   21350656 1.2493e+11 38519
## - .isMini     1   22465041 1.2493e+11 38519
## - AGE         1   53250502 1.2496e+11 38520
## - CAR_USE     1   58506421 1.2496e+11 38520
## - HOMEKIDS    1   72389904 1.2498e+11 38520
## - .isProf     1   85872280 1.2499e+11 38520
## - .isDiploma  1   96592977 1.2500e+11 38520
## - .isSUV      1  101086733 1.2501e+11 38520
## - MVR_PTS     1  113525926 1.2502e+11 38521
## <none>                     1.2491e+11 38521
## - .isSport    1  121963417 1.2503e+11 38521
## - .isDoctor   1  136125691 1.2504e+11 38521
## - HOME_VAL    1  141976858 1.2505e+11 38521
## - .isManager  1  149703095 1.2506e+11 38521
## - .isMasters  1  159785961 1.2507e+11 38521
## - MSTATUS     1  190940054 1.2510e+11 38522
## - REVOKED     1  215585495 1.2512e+11 38522
## - .isPhd      1  278586029 1.2518e+11 38523
## - SEX         1  284638565 1.2519e+11 38524
## - CAR_AGE     1  560064164 1.2547e+11 38528
## - BLUEBOOK    1 1133029219 1.2604e+11 38538
##
## Step:  AIC=38518.87
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + CAR_USE + BLUEBOOK + REVOKED + MVR_PTS +
##     CAR_AGE + .isDiploma + .isMasters + .isPhd + .isProf + .isDoctor +
##     .isManager + .isMini + .isSUV + .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - PARENT1     1   16675012 1.2494e+11 38517
## - .isPickup   1   20305802 1.2494e+11 38517
## - .isVan      1   20875498 1.2494e+11 38517
## - KIDSDRIV    1   21669890 1.2494e+11 38517
## - .isMini     1   22591726 1.2494e+11 38517
## - AGE         1   54201269 1.2498e+11 38518
## - CAR_USE     1   59155325 1.2498e+11 38518
## - HOMEKIDS    1   71743563 1.2499e+11 38518
## - .isProf     1   85011715 1.2501e+11 38518
## - .isDiploma  1   96909705 1.2502e+11 38519
## - .isSUV      1  100529774 1.2502e+11 38519
## <none>                     1.2492e+11 38519
## - .isSport    1  124063234 1.2505e+11 38519
## - .isDoctor   1  134967047 1.2506e+11 38519
## - HOME_VAL    1  141350709 1.2506e+11 38519
## - MVR_PTS     1  145091934 1.2507e+11 38519
## - .isManager  1  150246642 1.2507e+11 38519
## - .isMasters  1  158087826 1.2508e+11 38520
## - MSTATUS     1  189821651 1.2511e+11 38520
## - REVOKED     1  205747201 1.2513e+11 38520
## - .isPhd      1  274832417 1.2520e+11 38522
## - SEX         1  282495395 1.2520e+11 38522
## - CAR_AGE     1  552434314 1.2547e+11 38526
## - BLUEBOOK    1 1127662255 1.2605e+11 38536
##
```

```
## Step:  AIC=38517.15
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + HOME_VAL + MSTATUS +
##     SEX + CAR_USE + BLUEBOOK + REVOKED + MVR_PTS + CAR_AGE +
##     .isDiploma + .isMasters + .isPhd + .isProf + .isDoctor +
##     .isManager + .isMini + .isSUV + .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - KIDSDRIV    1   20412715 1.2496e+11 38516
## - .isPickup   1   21024069 1.2496e+11 38516
## - .isVan      1   21205920 1.2496e+11 38516
## - .isMini     1   23642868 1.2496e+11 38516
## - AGE         1   48505590 1.2499e+11 38516
## - CAR_USE     1   57387168 1.2500e+11 38516
## - .isProf     1   88447211 1.2503e+11 38517
## - .isDiploma  1   95226997 1.2503e+11 38517
## - .isSUV      1  102769414 1.2504e+11 38517
## <none>                     1.2494e+11 38517
## - .isSport    1  127239390 1.2507e+11 38517
## - HOMEKIDS    1  133244162 1.2507e+11 38517
## - .isDoctor   1  138104594 1.2508e+11 38518
## - HOME_VAL    1  140161633 1.2508e+11 38518
## - MVR_PTS     1  147788468 1.2509e+11 38518
## - .isManager  1  147821111 1.2509e+11 38518
## - .isMasters  1  158273860 1.2510e+11 38518
## - REVOKED     1  208198276 1.2515e+11 38519
## - .isPhd      1  275635409 1.2521e+11 38520
## - SEX         1  281684291 1.2522e+11 38520
## - MSTATUS     1  357876620 1.2530e+11 38521
## - CAR_AGE     1  552169617 1.2549e+11 38525
## - BLUEBOOK    1 1134811128 1.2607e+11 38535
##
## Step:  AIC=38515.5
## TARGET_AMT ~ AGE + HOMEKIDS + HOME_VAL + MSTATUS + SEX + CAR_USE +
##     BLUEBOOK + REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters +
##     .isPhd + .isProf + .isDoctor + .isManager + .isMini + .isSUV +
##     .isSport + .isVan + .isPickup
##
##              Df  Sum of Sq        RSS   AIC
## - .isPickup   1   20828343 1.2498e+11 38514
## - .isVan      1   22465466 1.2498e+11 38514
## - .isMini     1   23966362 1.2498e+11 38514
## - AGE         1   37422489 1.2500e+11 38514
## - CAR_USE     1   59686042 1.2502e+11 38515
## - .isProf     1   84537155 1.2504e+11 38515
## - .isDiploma  1   92790863 1.2505e+11 38515
## - .isSUV      1  104466810 1.2506e+11 38515
## - HOMEKIDS    1  115943130 1.2507e+11 38516
## <none>                     1.2496e+11 38516
## - .isSport    1  129176225 1.2509e+11 38516
## - .isDoctor   1  137166457 1.2510e+11 38516
## - HOME_VAL    1  140813625 1.2510e+11 38516
## - MVR_PTS     1  147802603 1.2511e+11 38516
## - .isMasters  1  153521552 1.2511e+11 38516
## - .isManager  1  161381981 1.2512e+11 38516
```

```
## - REVOKED     1  214996443 1.2517e+11 38517
## - .isPhd      1  274031410 1.2523e+11 38518
## - SEX         1  283914601 1.2524e+11 38518
## - MSTATUS     1  356534288 1.2531e+11 38520
## - CAR_AGE     1  547164714 1.2551e+11 38523
## - BLUEBOOK    1 1125049265 1.2608e+11 38533
##
## Step:  AIC=38513.86
## TARGET_AMT ~ AGE + HOMEKIDS + HOME_VAL + MSTATUS + SEX + CAR_USE +
##     BLUEBOOK + REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters +
##     .isPhd + .isProf + .isDoctor + .isManager + .isMini + .isSUV +
##     .isSport + .isVan
##
##              Df  Sum of Sq        RSS   AIC
## - .isMini     1    4682123 1.2498e+11 38512
## - .isVan      1    5800389 1.2498e+11 38512
## - AGE         1   42184106 1.2502e+11 38513
## - CAR_USE     1   53087353 1.2503e+11 38513
## - .isProf     1   79396442 1.2506e+11 38513
## - .isDiploma  1   90059917 1.2507e+11 38513
## - HOMEKIDS    1  114535932 1.2509e+11 38514
## - .isSUV      1  115940263 1.2510e+11 38514
## <none>                     1.2498e+11 38514
## - .isDoctor   1  130803512 1.2511e+11 38514
## - HOME_VAL    1  141314059 1.2512e+11 38514
## - .isSport    1  144975958 1.2512e+11 38514
## - .isMasters  1  146965562 1.2513e+11 38514
## - MVR_PTS     1  149278761 1.2513e+11 38514
## - .isManager  1  159119310 1.2514e+11 38515
## - REVOKED     1  210651651 1.2519e+11 38515
## - .isPhd      1  263699928 1.2524e+11 38516
## - SEX         1  263768766 1.2524e+11 38516
## - MSTATUS     1  352733305 1.2533e+11 38518
## - CAR_AGE     1  543770577 1.2552e+11 38521
## - BLUEBOOK    1 1344417470 1.2632e+11 38535
##
## Step:  AIC=38511.94
## TARGET_AMT ~ AGE + HOMEKIDS + HOME_VAL + MSTATUS + SEX + CAR_USE +
##     BLUEBOOK + REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters +
##     .isPhd + .isProf + .isDoctor + .isManager + .isSUV + .isSport +
##     .isVan
##
##              Df  Sum of Sq        RSS   AIC
## - .isVan      1    3682919 1.2499e+11 38510
## - AGE         1   42958475 1.2503e+11 38511
## - CAR_USE     1   48596380 1.2503e+11 38511
## - .isProf     1   77767784 1.2506e+11 38511
## - .isDiploma  1   87958076 1.2507e+11 38511
## - HOMEKIDS    1  114793384 1.2510e+11 38512
## <none>                     1.2498e+11 38512
## - .isSUV      1  121625677 1.2511e+11 38512
## - .isDoctor   1  129664949 1.2511e+11 38512
## - HOME_VAL    1  140690866 1.2512e+11 38512
## - .isMasters  1  144262073 1.2513e+11 38512
```

```
## - MVR_PTS     1  147911235 1.2513e+11 38512
## - .isSport    1  150231754 1.2513e+11 38513
## - .isManager  1  160909820 1.2514e+11 38513
## - REVOKED     1  212063099 1.2520e+11 38514
## - SEX         1  259309027 1.2524e+11 38514
## - .isPhd      1  260116147 1.2524e+11 38514
## - MSTATUS     1  354569912 1.2534e+11 38516
## - CAR_AGE     1  543024992 1.2553e+11 38519
## - BLUEBOOK    1 1340169326 1.2632e+11 38533
##
## Step:  AIC=38510.01
## TARGET_AMT ~ AGE + HOMEKIDS + HOME_VAL + MSTATUS + SEX + CAR_USE +
##     BLUEBOOK + REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters +
##     .isPhd + .isProf + .isDoctor + .isManager + .isSUV + .isSport
##
##              Df  Sum of Sq        RSS   AIC
## - AGE         1   42861848 1.2503e+11 38509
## - CAR_USE     1   50100742 1.2504e+11 38509
## - .isProf     1   78759027 1.2507e+11 38509
## - .isDiploma  1   89516003 1.2508e+11 38510
## - HOMEKIDS    1  114189099 1.2510e+11 38510
## <none>                     1.2499e+11 38510
## - .isSUV      1  120910523 1.2511e+11 38510
## - .isDoctor   1  129212548 1.2512e+11 38510
## - HOME_VAL    1  140437644 1.2513e+11 38510
## - .isMasters  1  146771387 1.2513e+11 38511
## - MVR_PTS     1  147594378 1.2514e+11 38511
## - .isSport    1  149694685 1.2514e+11 38511
## - .isManager  1  160082826 1.2515e+11 38511
## - REVOKED     1  212328849 1.2520e+11 38512
## - .isPhd      1  260878221 1.2525e+11 38512
## - SEX         1  275718712 1.2526e+11 38513
## - MSTATUS     1  354664100 1.2534e+11 38514
## - CAR_AGE     1  546042997 1.2553e+11 38517
## - BLUEBOOK    1 1416509506 1.2640e+11 38532
##
## Step:  AIC=38508.75
## TARGET_AMT ~ HOMEKIDS + HOME_VAL + MSTATUS + SEX + CAR_USE +
##     BLUEBOOK + REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters +
##     .isPhd + .isProf + .isDoctor + .isManager + .isSUV + .isSport
##
##              Df  Sum of Sq        RSS   AIC
## - CAR_USE     1   46543065 1.2508e+11 38508
## - HOMEKIDS    1   77230965 1.2511e+11 38508
## - .isProf     1   82782262 1.2511e+11 38508
## - .isDiploma  1   90339774 1.2512e+11 38508
## <none>                     1.2503e+11 38509
## - .isDoctor   1  122729504 1.2515e+11 38509
## - .isSUV      1  144030980 1.2517e+11 38509
## - MVR_PTS     1  145441844 1.2518e+11 38509
## - HOME_VAL    1  147273454 1.2518e+11 38509
## - .isManager  1  160950252 1.2519e+11 38510
## - .isMasters  1  163385887 1.2519e+11 38510
## - .isSport    1  171012573 1.2520e+11 38510
```

```
## - REVOKED     1  205611077 1.2524e+11 38510
## - .isPhd      1  277257456 1.2531e+11 38512
## - SEX         1  300909482 1.2533e+11 38512
## - MSTATUS     1  331900010 1.2536e+11 38512
## - CAR_AGE     1  540276798 1.2557e+11 38516
## - BLUEBOOK    1 1522934229 1.2655e+11 38533
##
## Step:  AIC=38507.55
## TARGET_AMT ~ HOMEKIDS + HOME_VAL + MSTATUS + SEX + BLUEBOOK +
##     REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters + .isPhd +
##     .isProf + .isDoctor + .isManager + .isSUV + .isSport
##
##              Df  Sum of Sq        RSS   AIC
## - .isProf     1   68257060 1.2515e+11 38507
## - .isDiploma  1   69115147 1.2515e+11 38507
## - HOMEKIDS    1   70835225 1.2515e+11 38507
## <none>                     1.2508e+11 38508
## - .isSUV      1  124064830 1.2520e+11 38508
## - .isDoctor   1  140812516 1.2522e+11 38508
## - .isMasters  1  143069211 1.2522e+11 38508
## - HOME_VAL    1  150021728 1.2523e+11 38508
## - .isSport    1  150087078 1.2523e+11 38508
## - MVR_PTS     1  162522083 1.2524e+11 38508
## - .isManager  1  177827595 1.2525e+11 38509
## - REVOKED     1  195253109 1.2527e+11 38509
## - .isPhd      1  270439752 1.2535e+11 38510
## - MSTATUS     1  332164772 1.2541e+11 38511
## - SEX         1  338880409 1.2542e+11 38511
## - CAR_AGE     1  519048698 1.2560e+11 38514
## - BLUEBOOK    1 1725701399 1.2680e+11 38535
##
## Step:  AIC=38506.72
## TARGET_AMT ~ HOMEKIDS + HOME_VAL + MSTATUS + SEX + BLUEBOOK +
##     REVOKED + MVR_PTS + CAR_AGE + .isDiploma + .isMasters + .isPhd +
##     .isDoctor + .isManager + .isSUV + .isSport
##
##              Df  Sum of Sq        RSS   AIC
## - HOMEKIDS    1   62993301 1.2521e+11 38506
## - .isDiploma  1   89918300 1.2524e+11 38506
## <none>                     1.2515e+11 38507
## - .isMasters  1  117303338 1.2526e+11 38507
## - .isSUV      1  123989593 1.2527e+11 38507
## - .isDoctor   1  143812447 1.2529e+11 38507
## - .isSport    1  151492343 1.2530e+11 38507
## - MVR_PTS     1  164436496 1.2531e+11 38508
## - HOME_VAL    1  164991246 1.2531e+11 38508
## - REVOKED     1  194851199 1.2534e+11 38508
## - .isManager  1  202759879 1.2535e+11 38508
## - .isPhd      1  240465205 1.2539e+11 38509
## - MSTATUS     1  340822976 1.2549e+11 38511
## - SEX         1  346821778 1.2549e+11 38511
## - CAR_AGE     1  490278849 1.2564e+11 38513
## - BLUEBOOK    1 1834895816 1.2698e+11 38536
##
```

```
## Step:  AIC=38505.8
## TARGET_AMT ~ HOME_VAL + MSTATUS + SEX + BLUEBOOK + REVOKED +
##     MVR_PTS + CAR_AGE + .isDiploma + .isMasters + .isPhd + .isDoctor +
##     .isManager + .isSUV + .isSport
##
##              Df  Sum of Sq        RSS   AIC
## - .isDiploma  1   89021431 1.2530e+11 38505
## - .isMasters  1  109285673 1.2532e+11 38506
## <none>                     1.2521e+11 38506
## - .isSUV      1  118137722 1.2533e+11 38506
## - .isSport    1  143748899 1.2535e+11 38506
## - .isDoctor   1  146768183 1.2535e+11 38506
## - HOME_VAL    1  152554584 1.2536e+11 38506
## - MVR_PTS     1  170411386 1.2538e+11 38507
## - REVOKED     1  184682259 1.2539e+11 38507
## - .isManager  1  197734903 1.2541e+11 38507
## - .isPhd      1  232990627 1.2544e+11 38508
## - MSTATUS     1  317234423 1.2553e+11 38509
## - SEX         1  322744870 1.2553e+11 38509
## - CAR_AGE     1  491492820 1.2570e+11 38512
## - BLUEBOOK    1 1799825787 1.2701e+11 38535
##
## Step:  AIC=38505.33
## TARGET_AMT ~ HOME_VAL + MSTATUS + SEX + BLUEBOOK + REVOKED +
##     MVR_PTS + CAR_AGE + .isMasters + .isPhd + .isDoctor + .isManager +
##     .isSUV + .isSport
##
##              Df  Sum of Sq        RSS   AIC
## - .isSUV      1  112975154 1.2541e+11 38505
## <none>                     1.2530e+11 38505
## - .isSport    1  142932820 1.2544e+11 38506
## - .isDoctor   1  145186851 1.2544e+11 38506
## - .isMasters  1  155041290 1.2545e+11 38506
## - HOME_VAL    1  160163477 1.2546e+11 38506
## - MVR_PTS     1  176192754 1.2547e+11 38506
## - .isManager  1  184608295 1.2548e+11 38507
## - REVOKED     1  188391841 1.2549e+11 38507
## - .isPhd      1  272375469 1.2557e+11 38508
## - SEX         1  327450510 1.2562e+11 38509
## - MSTATUS     1  346274969 1.2564e+11 38509
## - CAR_AGE     1  441039554 1.2574e+11 38511
## - BLUEBOOK    1 1816637510 1.2711e+11 38534
##
## Step:  AIC=38505.27
## TARGET_AMT ~ HOME_VAL + MSTATUS + SEX + BLUEBOOK + REVOKED +
##     MVR_PTS + CAR_AGE + .isMasters + .isPhd + .isDoctor + .isManager +
##     .isSport
##
##              Df  Sum of Sq        RSS   AIC
## - .isSport    1   45033014 1.2546e+11 38504
## <none>                     1.2541e+11 38505
## - .isDoctor   1  133483678 1.2554e+11 38506
## - .isMasters  1  159156837 1.2557e+11 38506
## - HOME_VAL    1  168292261 1.2558e+11 38506
```

```
## - MVR_PTS     1  172497244 1.2558e+11 38506
## - .isManager  1  182182783 1.2559e+11 38506
## - REVOKED     1  186850124 1.2560e+11 38506
## - SEX         1  230044707 1.2564e+11 38507
## - .isPhd      1  265633156 1.2568e+11 38508
## - MSTATUS     1  334390615 1.2574e+11 38509
## - CAR_AGE     1  442031700 1.2585e+11 38511
## - BLUEBOOK    1 1718681889 1.2713e+11 38533
##
## Step:  AIC=38504.04
## TARGET_AMT ~ HOME_VAL + MSTATUS + SEX + BLUEBOOK + REVOKED +
##     MVR_PTS + CAR_AGE + .isMasters + .isPhd + .isDoctor + .isManager
##
##              Df  Sum of Sq        RSS   AIC
## <none>                     1.2546e+11 38504
## - .isDoctor   1  131007336 1.2559e+11 38504
## - HOME_VAL    1  164856391 1.2562e+11 38505
## - .isMasters  1  166799995 1.2562e+11 38505
## - MVR_PTS     1  175394197 1.2563e+11 38505
## - .isManager  1  177007130 1.2563e+11 38505
## - REVOKED     1  185278451 1.2564e+11 38505
## - SEX         1  187899477 1.2564e+11 38505
## - .isPhd      1  271956843 1.2573e+11 38507
## - MSTATUS     1  330356699 1.2579e+11 38508
## - CAR_AGE     1  446707971 1.2590e+11 38510
## - BLUEBOOK    1 1675254417 1.2713e+11 38531
```

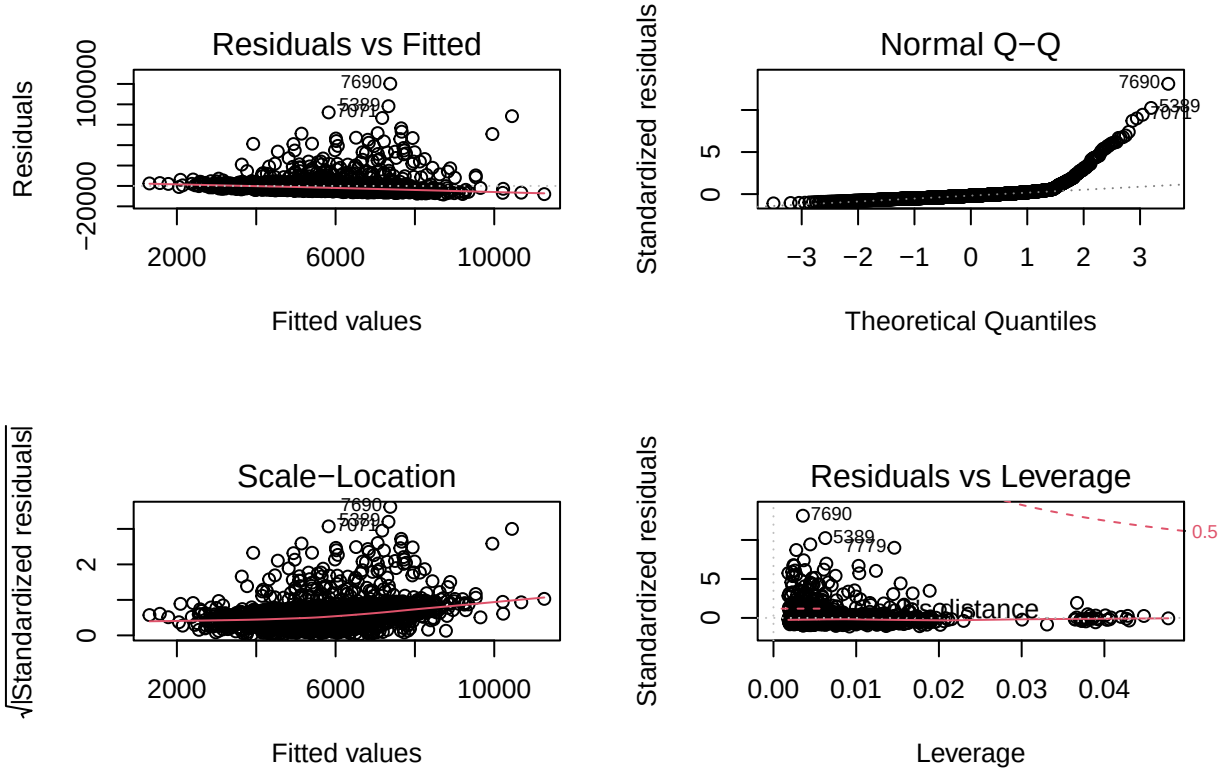
```
lm2<-lm(TARGET_AMT ~ HOME_VAL + MSTATUS + SEX + BLUEBOOK + REVOKED +MVR_PTS + CAR_AGE + .isPhd +.isDoct
summary(lm2)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ HOME_VAL + MSTATUS + SEX + BLUEBOOK +
##     REVOKED + MVR_PTS + CAR_AGE + .isPhd + .isDoctor + .isManager +
##     .isSUV + .isSport, data = train.clean.linear)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8402  -3188  -1498    426 100208
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3207.395   2161.033   1.484   0.1379
## HOME_VAL       58.317     34.163   1.707   0.0880 .
## MSTATUS      -990.868    405.289  -2.445   0.0146 *
## SEX         -1266.373    522.977  -2.421   0.0155 *
## BLUEBOOK       30.487      5.282   5.772 8.97e-09 ***
## REVOKED      -719.871    410.086  -1.755   0.0793 .
## MVR_PTS        13.976      8.037   1.739   0.0822 .
## CAR_AGE        -3.912      1.772  -2.208   0.0273 *
## .isPhd       1434.811    844.641   1.699   0.0895 .
## .isDoctor   -2510.120   1634.136  -1.536   0.1247
## .isManager  -1114.227    689.027  -1.617   0.1060
## .isSUV        801.247    567.073   1.413   0.1578
```

```
## .isSport     1082.353    663.181   1.632   0.1028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7658 on 2139 degrees of freedom
## Multiple R-squared:  0.02757,    Adjusted R-squared:  0.02212
## F-statistic: 5.054 on 12 and 2139 DF,  p-value: 2.311e-08
```

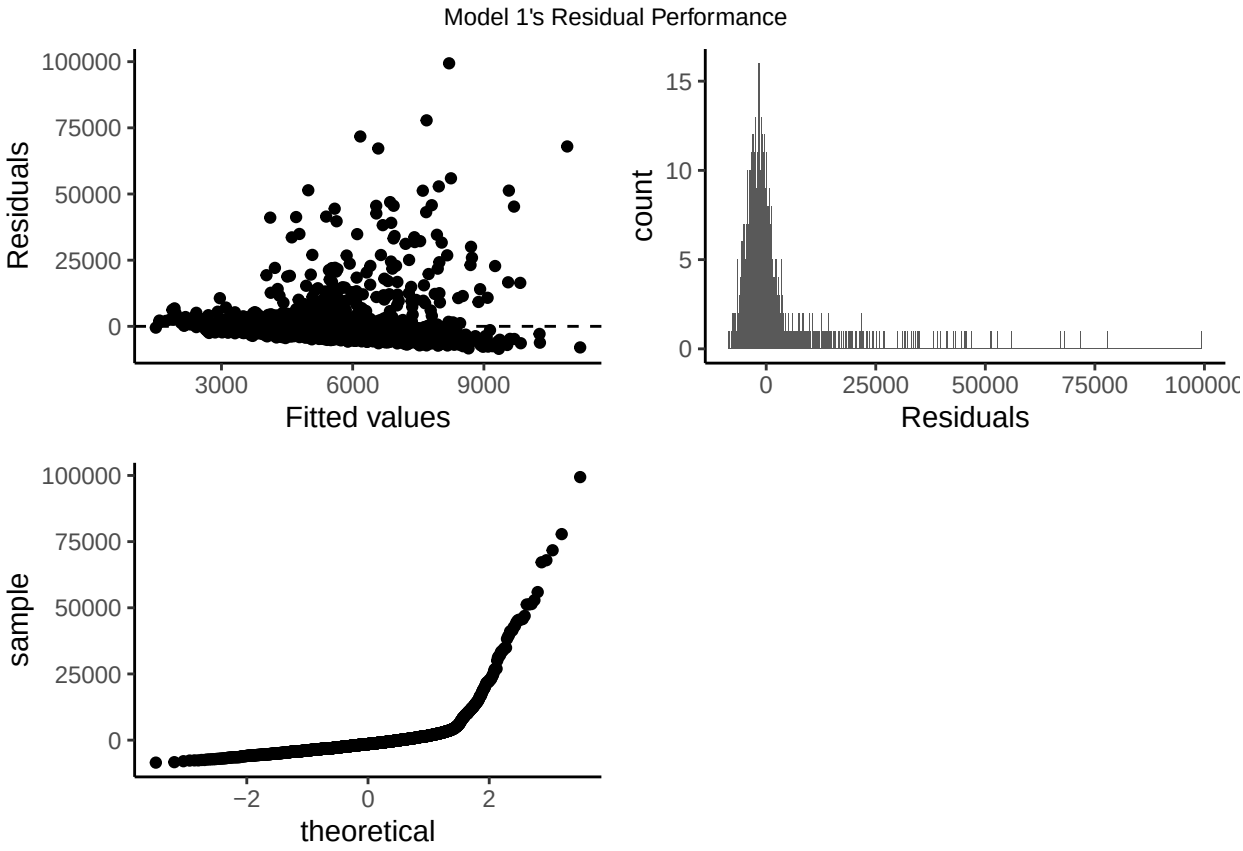
Like model 2, we can observe linearity and homoscedasticity, however, we can observe that the end of the tail veers further from the line in the Q-Q plot. There appears to have more outliers and even less normal distribution compared to model 1!

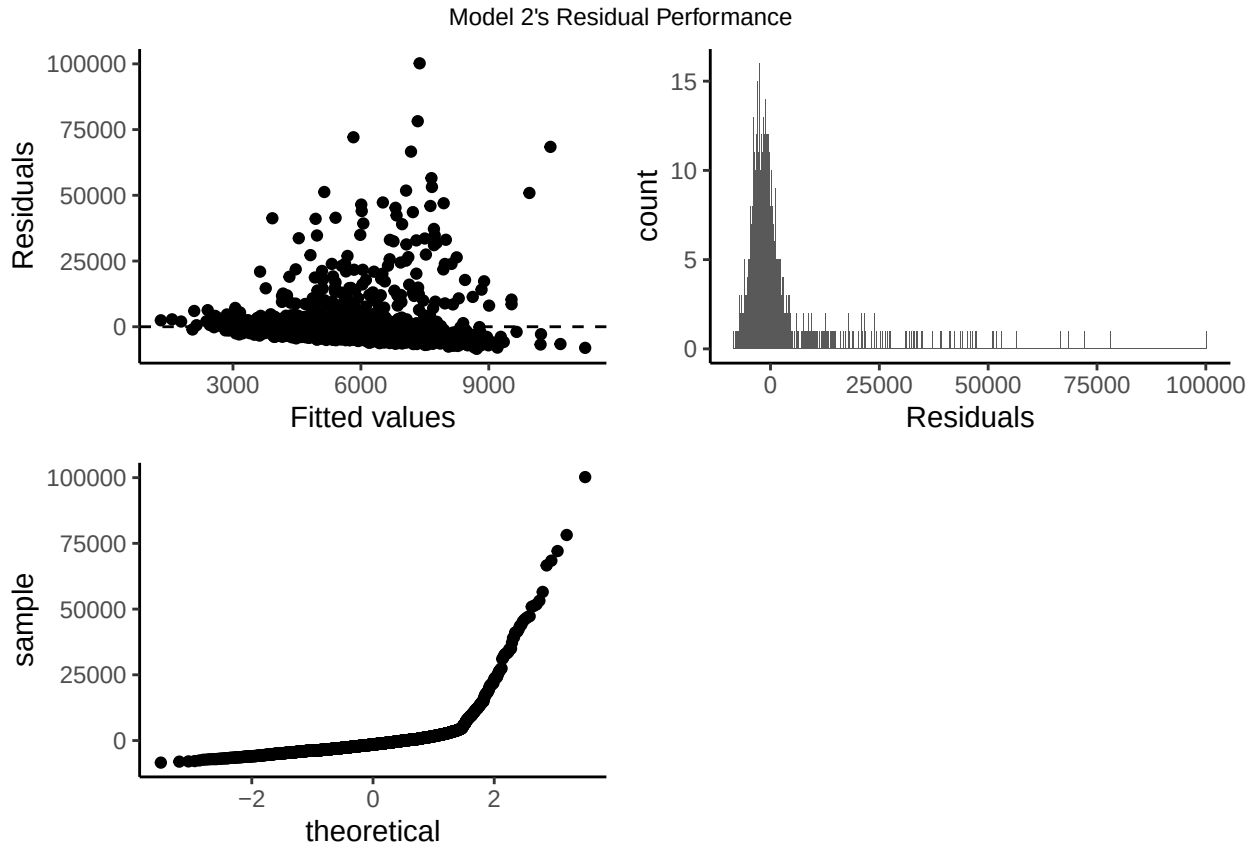
```
par(mfrow=c(2, 2))
plot(lm2)
```



Linear Model Selection Looking at the residuals plots of the linear models below.

the variance on the residuals vs fitted plot for Model 1's appears to have equal variance compared to model'2 plot. There is a slight slant pattern in the variance of model one; which indicates some error with variance.However, model 2's clustering on the plot indicates some heteroscedasticity with its residuals.


Model 1's Residual Performance
Residuals
Fitted values
count
Residuals
sample
theoretical



In the residual density plots, Model 2's density distribution is heavily right skewed while Model one is multi-modal.One possibility can the the features in model 2 bring out more outlier cases in the regression model compared to model 1. This could be seen the Q-Q ploy, where model 2's points begin to stray away.

In comparing the R-squared between the models, Model 1 (0.03079163) explains a slightly greater amount of the variance that occurs in the data compared to Model 2 (0.02676294). Despite Model 1 having a better score, both models overall have a low R-Squared value.

Model 2 has a slightly lower Stigma (Standard Deviation of Residuals) at 7661.511, which tells us that it has a slightly better prediction accuracy compared to Model 1 (7690.712)

Model 2 also has a smaller p value at 4.87e-08 compared to Model 1 (0.001986), which mean it has a stronger statistical significance.

Model 2 also has lower AIC (44616.90) and BIC (44696.33) values compared to the Model 1 AIC (44657.97) and BIC (44879.26)

```
##      model.build  r.squared adj.r.squared    sigma statistic      p.value df
## 1 Linear model 1 0.03226026    0.01532253 7684.883  1.904639 8.669687e-04 37
## 2 Linear model 2 0.02757193    0.02211651 7658.326  5.054045 2.310983e-08 12
##      logLik      AIC      BIC     deviance df.residual nobs
## 1 -22288.35 44654.71 44876.00 124847409816        2114 2152
## 2 -22293.55 44615.11 44694.54 125452248852        2139 2152
```

Despite a higher deviance and lower R-Squared value, Model 2 performs better in the other categories (Sigma, p value, AIC, BIC). The model is also less complex, making it more advantageous.

x
5970.758
6033.818
4386.837
5047.963
5792.728
7000.700
5668.590
7677.330
7089.696
7397.912

Reviewing predictions

x
5970.758
6033.818
4386.837
5047.963
5792.728
7000.700
5668.590
7677.330
7089.696
7397.912

Appendix

```
#importing data sets
library(tidymodels)
library(rpart)
library(tidyverse)
library(ggpubr)
library(mice)
library(corrplot)
library(MASS)
library(ISLR)
library(leaps)
library(bestglm)
library(pscl)
library(car)
library(lmtest)
library(performance)
library(PredictABEL)
library(predtools)
library(caret)
library(pROC)

training_data<-read.csv("https://raw.githubusercontent.com/Vy4thewin/criticalthinking3/main/insurance_t
```

```
testing_data<-read.csv("https://raw.githubusercontent.com/Vy4thewin/criticalthinking3/main/insurance-eva
#summary of the data set
summary(training_data)

#Seeing the unique categorical values
col<-training_data%>%
  dplyr::select(PARENT1,MSTATUS,SEX,EDUCATION,JOB,CAR_USE,CAR_TYPE,RED_CAR,REVOKED,URBANICITY )
lapply(col, unique)
#checking for Na values
colSums(is.na(training_data))
#Checking for linear assumption below
g1<-ggplot(training_data,aes(x=KIDSDRIV,y=TARGET_AMT))+geom_point()+labs(title = "KIDS DRIVING VS TARGE
g2<-ggplot(training_data,aes(x=AGE,y=TARGET_AMT))+geom_point()+labs(title = "AGE VS TARGETED AMOUNT",x=
g3<-ggplot(training_data,aes(x=HOMEKIDS,y=TARGET_AMT))+geom_point()+labs(title = "KIDS AT HOME VS TARGE
g4<-ggplot(training_data,aes(x=YOJ,y=TARGET_AMT))+geom_point()+labs(title = "# YRS ON THE JOB VS TARGETE
g5<-ggplot(training_data,aes(x=TRAVTIME,y=TARGET_AMT))+geom_point()+labs(title = "TRAVEL TIME VS TARGETE
g6<-ggplot(training_data,aes(x=TIF,y=TARGET_AMT))+geom_point()+labs(title = "TIME IN FORCE VS TARGETED A
g7<-ggplot(training_data,aes(x=MVR_PTS,y=TARGET_AMT))+geom_point()+labs(title = "MOTOR VEHICLE POINTS VS
g8<-ggplot(training_data,aes(x=CLM_FREQ,y=TARGET_AMT))+geom_point()+labs(title = "CLAIM FREQUENCY VS TA
g9<-ggplot(training_data,aes(x=CAR_AGE,y=TARGET_AMT))+geom_point()+labs(title = "CAR AGE VS TARGETED AM
plt<-ggarrange(g1,g2,g3,g4,g5,g6,g7,g8,g9,ncol =3 ,nrow =3)
annotate_figure(plt,top = text_grob("Response vs Feature Relationship",size=9))

#gathering a view on the current numeric variables available
#Definitely will need to revisit spent cols after transformations
g1<-ggplot(data=training_data,aes(x=KIDSDRIV))+geom_density()+theme_classic()
g2<-ggplot(data=training_data,aes(x=AGE))+geom_density()+theme_classic()
g3<-ggplot(data=training_data,aes(x=HOMEKIDS))+geom_density()+theme_classic()
g4<-ggplot(data=training_data,aes(x=YOJ))+geom_density()+theme_classic()
g5<-ggplot(data=training_data,aes(x=TRAVTIME))+geom_density()+theme_classic()
g6<-ggplot(data=training_data,aes(x=TIF))+geom_density()+theme_classic()
g7<-ggplot(data=training_data,aes(x=MVR_PTS))+geom_density()+theme_classic()
g8<-ggplot(data=training_data,aes(x=CLM_FREQ))+geom_density()+theme_classic()
g9<-ggplot(data=training_data,aes(x=CAR_AGE))+geom_density()+theme_classic()
plt<-ggarrange(g1,g2,g3,g4,g5,g6,g7,g8,g9,ncol =3 ,nrow =3)
annotate_figure(plt,top = text_grob("Densities of selected features",size=9))
#checking the distribution of the response
training_data%>%ggplot(aes(fill=TARGET_FLAG))+geom_bar(aes(x=TARGET_FLAG))+labs(title="Car crashes in t

#Removing index
train.clean<-training_data%>%dplyr::select(-(INDEX))

#converting the spend columns back to numeric
train.clean<-train.clean%>%mutate_at(c("INCOME","HOME_VAL","BLUEBOOK","OLDCLAIM"),~parse_number(.))
#Using MICE to impute missing values, using pmm to avoid neg impute values
train.clean<-complete(mice(train.clean,method = "pmm",seed = 333))

#Doubling checking no NAs arise from imputation
colSums(is.na(train.clean))
#Mutating the columns with two value into a binary dummy version below
#It's under the assumption, false==0 and true==1
train.clean<-train.clean%>%mutate(PARENT1=if_else(PARENT1=="No",0,1))
train.clean<-train.clean%>%mutate(MSTATUS=if_else(MSTATUS=="z_No",0,1))
```

```r
train.clean<-train.clean%>%mutate(SEX=if_else(SEX=="M",0,1))
train.clean<-train.clean%>%mutate(CAR_USE=if_else(CAR_USE=="Private",0,1))
train.clean<-train.clean%>%mutate(RED_CAR=if_else(RED_CAR=="no",0,1))
train.clean<-train.clean%>%mutate(REVOKED=if_else(REVOKED=="No",0,1))
train.clean<-train.clean%>%mutate(URBANICITY=if_else(URBANICITY=="z_Highly Rural/ Rural",0,1))

# Following the K-1 format, each variable lowest choice does not receive a column
#i.e highest education, the no high school diploma does not get a column
train.clean<-train.clean%>%mutate(.isDiploma=if_else(EDUCATION=="z_High School",1,0),
                                  .isBach=if_else(EDUCATION=="Bachelors",1,0),
                                  .isMasters=if_else(EDUCATION=="Masters",1,0),
                                  .isPhd=if_else(EDUCATION=="PhD",1,0)
                                  )
#assuming unemployed lowest level to not deal with a NA
train.clean<-train.clean%>%mutate(.isProf=if_else(JOB=="Professional",1,0),
                                  .isBlue=if_else(JOB=="z_Blue Collar",1,0),
                                  .isClerk=if_else(JOB=="Clerical",1,0),
                                  .isDoctor=if_else(JOB=="Doctor",1,0),
                                  .isLawyer=if_else(JOB=="Lawyer",1,0),
                                  .isHome=if_else(JOB=="Home Maker",1,0),
                                  .isStudent=if_else(JOB=="Student",1,0),
                                  .isManager=if_else(JOB=="Manager",1,0)
                                  )

#assume panel truck is lowest
train.clean<-train.clean%>%mutate(.isMini=if_else(CAR_TYPE=="Minivan",1,0),
                                  .isSUV=if_else(CAR_TYPE=="z_SUV",1,0),
                                  .isSport=if_else(CAR_TYPE=="Sports Car",1,0),
                                  .isVan=if_else(CAR_TYPE=="Van",1,0),
                                  .isPickup=if_else(CAR_TYPE=="Pickup",1,0)
                                  )

#removing categorical columns after dummies are set
train.clean<-train.clean%>%dplyr::select(-c(EDUCATION,CAR_TYPE,JOB))


#Checking distribution of the variables below before boxcox
g1<-ggplot(data=train.clean,aes(x=INCOME))+geom_density()+theme_classic()
g2<-ggplot(data=train.clean,aes(x=HOME_VAL))+geom_density()+theme_classic()
g3<-ggplot(data=train.clean,aes(x=BLUEBOOK))+geom_density()+theme_classic()
g4<-ggplot(data=train.clean,aes(x=OLDCLAIM))+geom_density()+theme_classic()
plt<-ggarrange(g1,g2,g3,g4,nrow = 2,ncol = 2)
annotate_figure(plt,top = text_grob("Pre-transformation on non-normal variables",size=9))

#importing mass in this section to avoid errors with dplyr
library(MASS)

#removing random car age of -3
train.clean<-train.clean%>%filter(!CAR_AGE==-3)

# Using BoxCox to select best fit transformation based on the lambda value of each predictor
#Performing box cox on the predictors and retrieving their lambdas
#adding a constant 1 as some observations are zero (i.e income)
```

```
lamb.INCOME<-boxcox((train.clean$INCOME+1)~1)
lamb.HOME_VAL<-boxcox((train.clean$HOME_VAL+1)~1)
lamb.BLUEBOOK<-boxcox((train.clean$ BLUEBOOK+1)~1)
lamb.OLDCLAIM<-boxcox((train.clean$OLDCLAIM+1)~1)
lamb.AGE<-boxcox((train.clean$AGE+1)~1)
lamb.HOMEKIDS<-boxcox((train.clean$HOMEKIDS+1)~1)
lamb.YOJ<-boxcox((train.clean$YOJ+1)~1)
lamb.MVR_PTS<-boxcox((train.clean$MVR_PTS+1)~1)
lamb.CLM_FREQ<-boxcox((train.clean$CLM_FREQ+1)~1)
lamb.CAR_AGE<-boxcox((train.clean$CAR_AGE+1)~1)

#retrieving the exact lambda for transformation
lamb.INCOME<-lamb.INCOME$x[which.max(lamb.INCOME$y)]#.042
lamb.HOME_VAL<-lamb.HOME_VAL$x[which.max(lamb.HOME_VAL$y)]#.22
lamb.BLUEBOOK<-lamb.BLUEBOOK$x[which.max(lamb.BLUEBOOK$y)]#.46
lamb.OLDCLAIM<-lamb.OLDCLAIM$x[which.max(lamb.OLDCLAIM$y)]#-.018
lamb.AGE<-lamb.AGE$x[which.max(lamb.AGE$y)]#1.03
lamb.HOMEKIDS<-lamb.HOMEKIDS$x[which.max(lamb.HOMEKIDS$y)]#-1.83
lamb.YOJ<-lamb.YOJ$x[which.max(lamb.YOJ$y)]#1.59
lamb.MVR_PTS<-lamb.MVR_PTS$x[which.max(lamb.MVR_PTS$y)]#-0.46
lamb.CLM_FREQ<-lamb.CLM_FREQ$x[which.max(lamb.CLM_FREQ$y)]#-1.47
lamb.CAR_AGE<-lamb.CAR_AGE$x[which.max(lamb.CAR_AGE$y)]#1.03

#Performing the aligned transformation. For spend, added a constant 1 to prevent transformations toward
train.clean<-train.clean%>%mutate(INCOME=sqrt(INCOME+1))#sqrt
train.clean<-train.clean%>%mutate(HOME_VAL=log(HOME_VAL+1))
train.clean<-train.clean%>%mutate(BLUEBOOK=sqrt(BLUEBOOK))
train.clean<-train.clean%>%mutate(OLDCLAIM=sqrt(OLDCLAIM+1))
train.clean<-train.clean%>%mutate(AGE=(AGE**lamb.AGE-1)/lamb.AGE)
train.clean<-train.clean%>%mutate(TRAVTIME=sqrt(TRAVTIME))
train.clean<-train.clean%>%mutate(YOJ=YOJ**2)
train.clean<-train.clean%>%mutate(MVR_PTS=MVR_PTS**2)
train.clean<-train.clean%>%mutate(CLM_FREQ=CLM_FREQ**3)
train.clean<-train.clean%>%mutate(CAR_AGE=CAR_AGE**2)

#Checking distribution of the variables below after boxcox
g1<-ggplot(data=train.clean,aes(x=INCOME))+geom_density()+theme_classic()
g2<-ggplot(data=train.clean,aes(x=HOME_VAL))+geom_density()+theme_classic()
g3<-ggplot(data=train.clean,aes(x=BLUEBOOK))+geom_density()+theme_classic()
g4<-ggplot(data=train.clean,aes(x=OLDCLAIM))+geom_density()+theme_classic()
plt<-ggarrange(g1,g2,g3,g4,nrow = 2,ncol = 2)
annotate_figure(plt,top = text_grob("Post-Transformtion on Non-Normal variables",size=9))

#checking for highly correlated variables
corrplot(cor(train.clean[,3:39]),method = "number",type="lower", tl.srt = .71,number.cex=0.75)

#assigning dummy variables as factors
train.clean<-train.clean%>%mutate_at(c(".isDiploma",".isBach",".isMasters",".isPhd",".isProf",".isBlue"

#Remove TARGET_AMT
train.clean.binary <- train.clean%>%dplyr::select(-c(TARGET_AMT))
#Forward selection
fit1 <- glm(TARGET_FLAG ~ ., data=train.clean.binary, family=binomial)
```

```r
summary(fit1)
# #Sort pvals in model 1
# idx <- order(coef(summary(fit1))[,4])  # sort out the p-values
# out <- coef(summary(fit1))[idx,]       # reorder coef, SE, etc. by increasing p
# (as.data.frame(out))
#Check VIF
vif_values <- vif(fit1)
print(vif_values)
#Explore removing highly Correlated variables
variables_to_exclude <- c(".isSUV", ".isBlue", ".isClerk", "OLDCLAIM")
names.include <- names(train.clean.binary)[!(names(train.clean.binary) %in% variables_to_exclude)]

updated_fit1 <- glm(TARGET_FLAG ~ ., train.clean.binary[, names.include], family=binomial)

summary(updated_fit1)
glance(updated_fit1)
#Stepwise Selection
fit2 <- glm(TARGET_FLAG ~ ., data = train.clean.binary, family="binomial") %>%
  stepAIC(direction = "both", trace=FALSE)

summary(fit2)
var_subset <- c("TARGET_FLAG", "URBANICITY", "REVOKED", "CAR_USE", "TRAVTIME", "TIF", "MVR_PTS", ".isMi

#Custom selection
fit3 <- glm(train.clean.binary[, var_subset], family=binomial)

summary(fit3)
glance(fit3)
#Calc McFaddens pseudo r^2 for each binary model
pseudo_r2.m1 <- pR2(fit1, method = "mcfadden")
pseudo_r2.m2 <- pR2(fit2, method = "mcfadden")
pseudo_r2.m3 <- pR2(fit3, method = "mcfadden")
mcfads_vals <- c(pseudo_r2.m1[4], pseudo_r2.m2[4], pseudo_r2.m3[4])

model_res <- bind_rows(glance(fit1), glance(fit2), glance(fit3))
model_names <- c("Bin model 1","Bin model 2","Bin model 3")
model_res <- cbind(model.build = model_names, model_res)
model_res <- cbind(model_res,McFaddens.R2 = mcfads_vals)

knitr::kable(model_res, "pipe")
par(mfrow = c(1, 3))

#model 1
resid.df1 <- mutate(train.clean.binary, residuals=residuals(fit1), linpred=predict(fit1))
gdf1 <- group_by(resid.df1, cut(linpred, breaks=unique(quantile(linpred,(1:272)/273))))
diagdf1 <- summarise(gdf1, residuals=mean(residuals), linpred=mean(linpred))
plot(residuals ~ linpred, diagdf1, xlab="linear predictor", main="Model 1")

#model 2
resid.df2 <- mutate(train.clean.binary, residuals=residuals(fit2), linpred=predict(fit2))
gdf2 <- group_by(resid.df2, cut(linpred, breaks=unique(quantile(linpred,(1:272)/273))))
diagdf2 <- summarise(gdf2, residuals=mean(residuals), linpred=mean(linpred))
plot(residuals ~ linpred, diagdf2, xlab="linear predictor", main="Model 2")
```

```r
#model 3
resid.df3 <- mutate(train.clean.binary, residuals=residuals(fit3), linpred=predict(fit3))
gdf3 <- group_by(resid.df3, cut(linpred, breaks=unique(quantile(linpred,(1:272)/273))))
diagdf3 <- summarise(gdf3, residuals=mean(residuals), linpred=mean(linpred))
plot(residuals ~ linpred, diagdf3, xlab="linear predictor", main="Model 3")
#the p-value for the test of the hypothesis that at least one of the predictors is related to the respo
#Values are large for all models, we cannot directly conclude a relationship.
sprintf("Model 1 p-val is: %.7f%%",(1-pchisq(model_res[["df.residual"]][1],fit1$df.residual)))  #model1
sprintf("Model 2 p-val is: %.7f%%",(1-pchisq(model_res[["df.residual"]][2],fit2$df.residual)))  #model2
sprintf("Model 3 p-val is: %.7f%%",(1-pchisq(model_res[["df.residual"]][3],fit3$df.residual)))  #model3
library(performance)
#model1
performance_hosmer(fit1, n_bins = 272)

#model2
performance_hosmer(fit2, n_bins = 272)

#model3
performance_hosmer(fit3, n_bins = 272)

knitr::kable(table(train.clean.binary$TARGET_FLAG), "pipe")
# Predict the probability (p) of crime
probabilities <- predict(fit2, type = "response")
predicted.classes <- ifelse(probabilities < 0.5, 0, 1)

# Select only numeric predictors
num_predictors <- train.clean.binary[,2:20]

predictors <- colnames(num_predictors)

# Bind the logit and tidying the data for plot
num_predictors <- num_predictors %>%
  mutate(logit = log(probabilities/(1-probabilities))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)

#Create Scatter plots
ggplot(num_predictors, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
library(predtools)

train.clean.binary$pred <- predict.glm(fit2, type = 'response')
calibration_plot(data = train.clean.binary, obs = "TARGET_FLAG", pred = "pred", title = "Calibration pl
train.clean.binary <- mutate(train.clean.binary, predout=ifelse(pred < 0.5, 0, 1))

#Create confusion matrix
cm <- confusionMatrix(as.factor(train.clean.binary$predout), as.factor(train.clean.binary$TARGET_FLAG))

#Calculate AUC
auc_res <- auc(train.clean.binary$TARGET_FLAG, train.clean.binary$pred)
```

```r
sprintf("Model 2 Classification Accuracy is: %.2f%%",(cm$overall[1])*100)
sprintf("Model 2 Classification Error Rate is: %.2f%%",(1-cm$overall[1])*100)
sprintf("Model 2 Precision is: %.2f%%",(cm$byClass['Pos Pred Value']*100))
sprintf("Model 2 Sensitivity/Recall is: %.2f%%",(cm$byClass['Sensitivity']*100))
sprintf("Model 2 Specificity is: %.2f%%",(cm$byClass['Specificity']*100))
sprintf("Model 2 F1-score is: %.2f%%",(cm$byClass['F1']*100))
sprintf("Model 2 AUC is: %.2f%%",(auc_res[1]*100))
# par(pty="s")
# roc_score= roc(train.clean.binary$TARGET_FLAG, fit2$fitted.values) #AUC score
# plot(roc_score ,main ="ROC curve -- Logistic Regression ", legacy.axes=TRUE)

true_labels <- train.clean.binary$TARGET_FLAG
roc_curve <- roc(true_labels, fit2$fitted.values)
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)
abline(a = 0, b = 1, col = "gray", lty = 2)
legend("bottomright", legend = paste("AUC =", round(auc(roc_curve), 2)), col = "blue", lwd = 2)
#Removing index and TARGET_AMT
testing.set <-testing_data %>%
  dplyr::select(-c(INDEX, TARGET_AMT, TARGET_FLAG))

#converting the spend columns back to numeric
testing.clean <- testing.set %>%
  mutate_at(c("INCOME","HOME_VAL","BLUEBOOK","OLDCLAIM"),~parse_number(.))

#Using MICE to impute missing values, using pmm to avoid neg impute values
testing.clean <- complete(mice(testing.clean,method = "pmm",seed = 333))


#Mutating the columns with two value into a binary dummy version below
#It's under the assumption, false==0 and true==1
testing.clean<-testing.clean%>%mutate(PARENT1=if_else(PARENT1=="No",0,1))
testing.clean<-testing.clean%>%mutate(MSTATUS=if_else(MSTATUS=="z_No",0,1))
testing.clean<-testing.clean%>%mutate(SEX=if_else(SEX=="M",0,1))
testing.clean<-testing.clean%>%mutate(CAR_USE=if_else(CAR_USE=="Private",0,1))
testing.clean<-testing.clean%>%mutate(RED_CAR=if_else(RED_CAR=="no",0,1))
testing.clean<-testing.clean%>%mutate(REVOKED=if_else(REVOKED=="No",0,1))
testing.clean<-testing.clean%>%mutate(URBANICITY=if_else(URBANICITY=="z_Highly Rural/ Rural",0,1))

# Following the K-1 format, each variable lowest choice does not receive a column
#i.e highest education, the no high school diploma does not get a column
testing.clean<-testing.clean%>%mutate(.isDiploma=if_else(EDUCATION=="z_High School",1,0),
                                      .isBach=if_else(EDUCATION=="Bachelors",1,0),
                                      .isMasters=if_else(EDUCATION=="Masters",1,0),
                                      .isPhd=if_else(EDUCATION=="PhD",1,0)
                                      )
#assuming unemployed lowest level to not deal with a NA
testing.clean<-testing.clean%>%mutate(.isProf=if_else(JOB=="Professional",1,0),
                                      .isBlue=if_else(JOB=="z_Blue Collar",1,0),
                                      .isClerk=if_else(JOB=="Clerical",1,0),
                                      .isDoctor=if_else(JOB=="Doctor",1,0),
                                      .isLawyer=if_else(JOB=="Lawyer",1,0),
                                      .isHome=if_else(JOB=="Home Maker",1,0),
                                      .isStudent=if_else(JOB=="Student",1,0),
```

```
                                    .isManager=if_else(JOB=="Manager",1,0)
                                    )

#assume panel truck is lowest
testing.clean<-testing.clean%>%mutate(.isMini=if_else(CAR_TYPE=="Minivan",1,0),
                                    .isSUV=if_else(CAR_TYPE=="z_SUV",1,0),
                                    .isSport=if_else(CAR_TYPE=="Sports Car",1,0),
                                    .isVan=if_else(CAR_TYPE=="Van",1,0),
                                    .isPickup=if_else(CAR_TYPE=="Pickup",1,0)
                                    )

#removing categorical columns after dummies are set
testing.clean<-testing.clean%>%dplyr::select(-c(EDUCATION,CAR_TYPE,JOB))

#Applying boxcox transformations
testing.clean<-testing.clean%>%mutate(INCOME=sqrt(INCOME+1))
testing.clean<-testing.clean%>%mutate(HOME_VAL=log(HOME_VAL+1))
testing.clean<-testing.clean%>%mutate(BLUEBOOK=sqrt(BLUEBOOK))
testing.clean<-testing.clean%>%mutate(OLDCLAIM=sqrt(OLDCLAIM+1))
testing.clean<-testing.clean%>%mutate(AGE=(AGE**lamb.AGE-1)/lamb.AGE)
testing.clean<-testing.clean%>%mutate(TRAVTIME=sqrt(TRAVTIME))
testing.clean<-testing.clean%>%mutate(YOJ=YOJ**2)
testing.clean<-testing.clean%>%mutate(MVR_PTS=MVR_PTS**2)
testing.clean<-testing.clean%>%mutate(CLM_FREQ=CLM_FREQ**3)
testing.clean<-testing.clean%>%mutate(CAR_AGE=CAR_AGE**2)

#assigning dummy variables as factors
testing.clean<-testing.clean%>%mutate_at(c(".isDiploma",".isBach",".isMasters",".isPhd",".isProf",".isB

testing.clean$pred_prob <- predict(fit2, testing.clean, type="response")
testing.clean <- mutate(testing.clean, predout=ifelse(pred_prob < 0.5, 0, 1))

knitr::kable(head(testing.clean,10) , "pipe")

knitr::kable(table(testing.clean$predout), "pipe")

#Remove TARGET_FLAG
train.clean.linear <- subset(train.clean, select = -c(TARGET_FLAG))
#Remove 0 and Nulls
train.clean.linear <- train.clean.linear[!(train.clean.linear$TARGET_AMT %in% c('0', 'NA')), ]

train.clean.linear<-train.clean.linear%>%mutate_at(c(".isDiploma",".isBach",".isMasters",".isPhd",".isP
#first model with all predictors
lm1 <- lm(TARGET_AMT~., data = train.clean.linear)
summary(lm1)
par(mfrow=c(2, 2))
plot(lm1)
lm2 <- stats::step(lm1,direction="backward")
lm2<-lm(TARGET_AMT ~ HOME_VAL + MSTATUS + SEX + BLUEBOOK + REVOKED +MVR_PTS + CAR_AGE + .isPhd +.isDoct
summary(lm2)
par(mfrow=c(2, 2))
plot(lm2)
#Plotting residual vs fitted values for visual representation of model performance
```

```r
g1<-ggplot(data = lm1 , aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")+theme_classic()
g2<-ggplot(data = lm1 , aes(x = .resid)) +
    geom_histogram(binwidth = 25) +
    xlab("Residuals")+theme_classic()
g3<-ggplot(data = lm1 , aes(sample = .resid)) +
  stat_qq()+theme_classic()
plt<-ggarrange(g1,g2,g3)
annotate_figure(plt,top = text_grob("Model 1's Residual Performance",size=9))

g1<-ggplot(data = lm2 , aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")+theme_classic()
g2<-ggplot(data = lm2 , aes(x = .resid)) +
    geom_histogram(binwidth = 25) +
    xlab("Residuals")+theme_classic()
g3<-ggplot(data = lm2 , aes(sample = .resid)) +
  stat_qq()+theme_classic()
plt<-ggarrange(g1,g2,g3)
annotate_figure(plt,top = text_grob("Model 2's Residual Performance",size=9))

linear_model_res <- bind_rows(glance(lm1), glance(lm2))
model_names1 <- c("Linear model 1","Linear model 2")
linear_model_res <- cbind(model.build = model_names1, linear_model_res)
linear_model_res
#updating testing set values for linear prediction
testing.clean<-testing.clean%>%mutate_at(c(".isDiploma",".isBach",".isMasters",".isPhd",".isProf",".isB

#predicting the target amount w/ model2
pred<-predict(lm2,newdata =testing.clean)

knitr::kable(head(pred,10))

#predicting the target amount w/ model2
pred<-predict(lm2,newdata =testing.clean)

knitr::kable(head(pred,10))
#importing data sets
library(tidymodels)
library(rpart)
library(tidyverse)
library(ggpubr)
library(mice)
library(corrplot)
library(MASS)
library(ISLR)
library(leaps)
library(bestglm)
```

```
library(pscl)
library(car)
library(lmtest)
library(performance)
library(PredictABEL)
library(predtools)
library(caret)
library(pROC)

training_data<-read.csv("https://raw.githubusercontent.com/Vy4thewin/criticalthinking3/main/insurance_t
testing_data<-read.csv("https://raw.githubusercontent.com/Vy4thewin/criticalthinking3/main/insurance-eva
#summary of the data set
summary(training_data)

#Seeing the unique categorical values
col<-training_data%>%
  dplyr::select(PARENT1,MSTATUS,SEX,EDUCATION,JOB,CAR_USE,CAR_TYPE,RED_CAR,REVOKED,URBANICITY )
lapply(col, unique)

#checking for Na values
colSums(is.na(training_data))
#Checking for linear assumption below
g1<-ggplot(training_data,aes(x=KIDSDRIV,y=TARGET_AMT))+geom_point()+labs(title = "KIDS DRIVING VS TARGE
g2<-ggplot(training_data,aes(x=AGE,y=TARGET_AMT))+geom_point()+labs(title = "AGE VS TARGETED AMOUNT",x=
g3<-ggplot(training_data,aes(x=HOMEKIDS,y=TARGET_AMT))+geom_point()+labs(title = "KIDS AT HOME VS TARGE
g4<-ggplot(training_data,aes(x=YOJ,y=TARGET_AMT))+geom_point()+labs(title = "# YRS ON THE JOB VS TARGET
g5<-ggplot(training_data,aes(x=TRAVTIME,y=TARGET_AMT))+geom_point()+labs(title = "TRAVEL TIME VS TARGET
g6<-ggplot(training_data,aes(x=TIF,y=TARGET_AMT))+geom_point()+labs(title = "TIME IN FORCE VS TARGETED
g7<-ggplot(training_data,aes(x=MVR_PTS,y=TARGET_AMT))+geom_point()+labs(title = "MOTOR VEHICLE POINTS V
g8<-ggplot(training_data,aes(x=CLM_FREQ,y=TARGET_AMT))+geom_point()+labs(title = "CLAIM FREQUENCY VS TA
g9<-ggplot(training_data,aes(x=CAR_AGE,y=TARGET_AMT))+geom_point()+labs(title = "CAR AGE VS TARGETED AM
ggarrange(g1,g2,g3,g4,g5,g6,g7,g8,g9,ncol =3 ,nrow =3)

#gathering a view on the current numeric variables available
#Definitely will need to revisit spent cols after transformations
g1<-ggplot(data=training_data,aes(x=KIDSDRIV))+geom_density()+theme_classic()
g2<-ggplot(data=training_data,aes(x=AGE))+geom_density()+theme_classic()
g3<-ggplot(data=training_data,aes(x=HOMEKIDS))+geom_density()+theme_classic()
g4<-ggplot(data=training_data,aes(x=YOJ))+geom_density()+theme_classic()
g5<-ggplot(data=training_data,aes(x=TRAVTIME))+geom_density()+theme_classic()
g6<-ggplot(data=training_data,aes(x=TIF))+geom_density()+theme_classic()
g7<-ggplot(data=training_data,aes(x=MVR_PTS))+geom_density()+theme_classic()
g8<-ggplot(data=training_data,aes(x=CLM_FREQ))+geom_density()+theme_classic()
g9<-ggplot(data=training_data,aes(x=CAR_AGE))+geom_density()+theme_classic()
ggarrange(g1,g2,g3,g4,g5,g6,g7,g8,g9,ncol =3 ,nrow =3)

#checking the distribution of the response
training_data%>%ggplot(aes(fill=TARGET_FLAG))+geom_bar(aes(x=TARGET_FLAG))+labs(title="Car crashes in t

#Removing index
train.clean<-training_data%>%dplyr::select(-(INDEX))

#converting the spend columns back to numeric
```

```r
train.clean<-train.clean%>%mutate_at(c("INCOME","HOME_VAL","BLUEBOOK","OLDCLAIM"),~parse_number(.))

#Using MICE to impute missing values, using pmm to avoid neg impute values
train.clean<-complete(mice(train.clean,method = "pmm",seed = 333))

#Doubling checking no NAs arise from imputation
colSums(is.na(train.clean))


#Mutating the columns with two value into a binary dummy version below
#It's under the assumption, false==0 and true==1
train.clean<-train.clean%>%mutate(PARENT1=if_else(PARENT1=="No",0,1))
train.clean<-train.clean%>%mutate(MSTATUS=if_else(MSTATUS=="z_No",0,1))
train.clean<-train.clean%>%mutate(SEX=if_else(SEX=="M",0,1))
train.clean<-train.clean%>%mutate(CAR_USE=if_else(CAR_USE=="Private",0,1))
train.clean<-train.clean%>%mutate(RED_CAR=if_else(RED_CAR=="no",0,1))
train.clean<-train.clean%>%mutate(REVOKED=if_else(REVOKED=="No",0,1))
train.clean<-train.clean%>%mutate(URBANICITY=if_else(URBANICITY=="z_Highly Rural/ Rural",0,1))

# Following the K-1 format, each variable lowest choice does not receive a column
#i.e highest education, the no high school diploma does not get a column
train.clean<-train.clean%>%mutate(.isDiploma=if_else(EDUCATION=="z_High School",1,0),
                                  .isBach=if_else(EDUCATION=="Bachelors",1,0),
                                  .isMasters=if_else(EDUCATION=="Masters",1,0),
                                  .isPhd=if_else(EDUCATION=="PhD",1,0)
                                  )
#assuming unemployed lowest level to not deal with a NA
train.clean<-train.clean%>%mutate(.isProf=if_else(JOB=="Professional",1,0),
                                  .isBlue=if_else(JOB=="z_Blue Collar",1,0),
                                  .isClerk=if_else(JOB=="Clerical",1,0),
                                  .isDoctor=if_else(JOB=="Doctor",1,0),
                                  .isLawyer=if_else(JOB=="Lawyer",1,0),
                                  .isHome=if_else(JOB=="Home Maker",1,0),
                                  .isStudent=if_else(JOB=="Student",1,0),
                                  .isManager=if_else(JOB=="Manager",1,0)
                                  )

#assume panel truck is lowest
train.clean<-train.clean%>%mutate(.isMini=if_else(CAR_TYPE=="Minivan",1,0),
                                  .isSUV=if_else(CAR_TYPE=="z_SUV",1,0),
                                  .isSport=if_else(CAR_TYPE=="Sports Car",1,0),
                                  .isVan=if_else(CAR_TYPE=="Van",1,0),
                                  .isPickup=if_else(CAR_TYPE=="Pickup",1,0)
                                  )

#removing categorical columns after dummies are set
train.clean<-train.clean%>%dplyr::select(-c(EDUCATION,CAR_TYPE,JOB))

#Checking distribution of the variables below before boxcox
g1<-ggplot(data=train.clean,aes(x=INCOME))+geom_density()+theme_classic()
g2<-ggplot(data=train.clean,aes(x=HOME_VAL))+geom_density()+theme_classic()
g3<-ggplot(data=train.clean,aes(x=BLUEBOOK))+geom_density()+theme_classic()
g4<-ggplot(data=train.clean,aes(x=OLDCLAIM))+geom_density()+theme_classic()
```

```
ggarrange(g1,g2,g3,g4,nrow = 2,ncol = 2)

#importing mass in this section to avoid errors with dplyr
library(MASS)

#removing random car age of -3
train.clean<-train.clean%>%filter(!CAR_AGE==-3)

# Using BoxCox to select best fit transformation based on the lambda value of each predictor
#Performing box cox on the predictors and retrieving their lambdas
#adding a constant 1 as some observations are zero (i.e income)
lamb.INCOME<-boxcox((train.clean$INCOME+1)~1)
lamb.HOME_VAL<-boxcox((train.clean$HOME_VAL+1)~1)
lamb.BLUEBOOK<-boxcox((train.clean$ BLUEBOOK+1)~1)
lamb.OLDCLAIM<-boxcox((train.clean$OLDCLAIM+1)~1)
lamb.AGE<-boxcox((train.clean$AGE+1)~1)
lamb.HOMEKIDS<-boxcox((train.clean$HOMEKIDS+1)~1)
lamb.YOJ<-boxcox((train.clean$YOJ+1)~1)
lamb.MVR_PTS<-boxcox((train.clean$MVR_PTS+1)~1)
lamb.CLM_FREQ<-boxcox((train.clean$CLM_FREQ+1)~1)
lamb.CAR_AGE<-boxcox((train.clean$CAR_AGE+1)~1)

#retrieving the exact lambda for transformation
lamb.INCOME<-lamb.INCOME$x[which.max(lamb.INCOME$y)]#.042
lamb.HOME_VAL<-lamb.HOME_VAL$x[which.max(lamb.HOME_VAL$y)]#.22
lamb.BLUEBOOK<-lamb.BLUEBOOK$x[which.max(lamb.BLUEBOOK$y)]#.46
lamb.OLDCLAIM<-lamb.OLDCLAIM$x[which.max(lamb.OLDCLAIM$y)]#-.018
lamb.AGE<-lamb.AGE$x[which.max(lamb.AGE$y)]#1.03
lamb.HOMEKIDS<-lamb.HOMEKIDS$x[which.max(lamb.HOMEKIDS$y)]#-1.83
lamb.YOJ<-lamb.YOJ$x[which.max(lamb.YOJ$y)]#1.59
lamb.MVR_PTS<-lamb.MVR_PTS$x[which.max(lamb.MVR_PTS$y)]#-0.46
lamb.CLM_FREQ<-lamb.CLM_FREQ$x[which.max(lamb.CLM_FREQ$y)]#-1.47
lamb.CAR_AGE<-lamb.CAR_AGE$x[which.max(lamb.CAR_AGE$y)]#1.03

#Performing the aligned transformation. For spend, added a constant 1 to prevent transformations toward
train.clean<-train.clean%>%mutate(INCOME=sqrt(INCOME+1))#sqrt
train.clean<-train.clean%>%mutate(HOME_VAL=log(HOME_VAL+1))
train.clean<-train.clean%>%mutate(BLUEBOOK=sqrt(BLUEBOOK))
train.clean<-train.clean%>%mutate(OLDCLAIM=sqrt(OLDCLAIM+1))
train.clean<-train.clean%>%mutate(AGE=(AGE**lamb.AGE-1)/lamb.AGE)
train.clean<-train.clean%>%mutate(TRAVTIME=sqrt(TRAVTIME))
train.clean<-train.clean%>%mutate(YOJ=YOJ**2)
train.clean<-train.clean%>%mutate(MVR_PTS=MVR_PTS**2)
train.clean<-train.clean%>%mutate(CLM_FREQ=CLM_FREQ**3)
train.clean<-train.clean%>%mutate(CAR_AGE=CAR_AGE**2)


#Checking distribution of the variables below after boxcox
g1<-ggplot(data=train.clean,aes(x=INCOME))+geom_density()+theme_classic()
g2<-ggplot(data=train.clean,aes(x=HOME_VAL))+geom_density()+theme_classic()
g3<-ggplot(data=train.clean,aes(x=BLUEBOOK))+geom_density()+theme_classic()
g4<-ggplot(data=train.clean,aes(x=OLDCLAIM))+geom_density()+theme_classic()
ggarrange(g1,g2,g3,g4,nrow = 2,ncol = 2)
```

```r
#checking for highly correlated variables
corrplot(cor(train.clean[,3:39]),method = "number",type="lower", tl.srt = .71,number.cex=0.75)


#assigning dummy variables as factors
train.clean<-train.clean%>%mutate_at(c(".isDiploma",".isBach",".isMasters",".isPhd",".isProf",".isBlue"

#Remove TARGET_AMT
train.clean.binary <- subset(train.clean, select = -c(TARGET_AMT)) #|>
  # rename(y = TARGET_FLAG) |>
  # relocate(y, .after = last_col())

#Forward selection
fit1 <- glm(TARGET_FLAG ~ ., data=train.clean.binary, family=binomial)

#Check VIF
vif_values <- vif(fit1)
print(vif_values)

#Explore removing highly Correlated variables
variables_to_exclude <- c(".isSUV", ".isBlue", ".isClerk", "OLDCLAIM")
names.include <- names(train.clean.binary)[!(names(train.clean.binary) %in% variables_to_exclude)]

updated_fit1 <- glm(TARGET_FLAG ~ ., train.clean.binary[, names.include], family=binomial)

summary(updated_fit1)
glance(updated_fit1)

#Stepwise Selection
fit2 <- glm(TARGET_FLAG ~ ., data = train.clean.binary, family="binomial") %>%
  stepAIC(direction = "both", trace=FALSE)

summary(fit2)

var_subset <- c("TARGET_FLAG", "URBANICITY", "REVOKED", "CAR_USE", "TRAVTIME", "TIF", "MVR_PTS", ".isMi

#Custom selection
fit3 <- glm(train.clean.binary[, var_subset], family=binomial)

summary(fit3)
glance(fit3)

#Calc McFaddens pseudo r^2 for each binary model
pseudo_r2.m1 <- pR2(fit1, method = "mcfadden")
pseudo_r2.m2 <- pR2(fit2, method = "mcfadden")
pseudo_r2.m3 <- pR2(fit3, method = "mcfadden")
mcfads_vals <- c(pseudo_r2.m1[4], pseudo_r2.m2[4], pseudo_r2.m3[4])

model_res <- bind_rows(glance(fit1), glance(fit2), glance(fit3))
model_names <- c("Bin model 1","Bin model 2","Bin model 3")
model_res <- cbind(model.build = model_names, model_res)
model_res <- cbind(model_res,McFaddens.R2 = mcfads_vals)
```

```r
knitr::kable(model_res, "pipe")

par(mfrow = c(1, 3))

#model 1
resid.df1 <- mutate(train.clean.binary, residuals=residuals(fit1), linpred=predict(fit1))
gdf1 <- group_by(resid.df1, cut(linpred, breaks=unique(quantile(linpred,(1:272)/273))))
diagdf1 <- summarise(gdf1, residuals=mean(residuals), linpred=mean(linpred))
plot(residuals ~ linpred, diagdf1, xlab="linear predictor", main="Model 1")

#model 2
resid.df2 <- mutate(train.clean.binary, residuals=residuals(fit2), linpred=predict(fit2))
gdf2 <- group_by(resid.df2, cut(linpred, breaks=unique(quantile(linpred,(1:272)/273))))
diagdf2 <- summarise(gdf2, residuals=mean(residuals), linpred=mean(linpred))
plot(residuals ~ linpred, diagdf2, xlab="linear predictor", main="Model 2")

#model 3
resid.df3 <- mutate(train.clean.binary, residuals=residuals(fit3), linpred=predict(fit3))
gdf3 <- group_by(resid.df3, cut(linpred, breaks=unique(quantile(linpred,(1:272)/273))))
diagdf3 <- summarise(gdf3, residuals=mean(residuals), linpred=mean(linpred))
plot(residuals ~ linpred, diagdf3, xlab="linear predictor", main="Model 3")

#the p-value for the test of the hypothesis that at least one of the predictors is related to the respo
#Values are large for all models, we cannot directly conclude a relationship.
sprintf("Model 1 p-val is: %.7f%%",(1-pchisq(model_res[["df.residual"]][1],fit1$df.residual)))  #model1
sprintf("Model 2 p-val is: %.7f%%",(1-pchisq(model_res[["df.residual"]][2],fit2$df.residual)))  #model2
sprintf("Model 3 p-val is: %.7f%%",(1-pchisq(model_res[["df.residual"]][3],fit3$df.residual)))  #model3

library(performance)
#model1
performance_hosmer(fit1, n_bins = 272)

#model2
performance_hosmer(fit2, n_bins = 272)

#model3
performance_hosmer(fit3, n_bins = 272)

knitr::kable(table(train.clean.binary$TARGET_FLAG), "pipe")

# Predict the probability (p) of crime
probabilities <- predict(fit2, type = "response")
predicted.classes <- ifelse(probabilities < 0.5, 0, 1)

# Select only numeric predictors
num_predictors <- train.clean.binary[,2:20]

predictors <- colnames(num_predictors)

# Bind the logit and tidying the data for plot
num_predictors <- num_predictors %>%
  mutate(logit = log(probabilities/(1-probabilities))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)
```

```
#Create Scatter plots
ggplot(num_predictors, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")

library(predtools)

train.clean.binary$pred <- predict.glm(fit2, type = 'response')
calibration_plot(data = train.clean.binary, obs = "TARGET_FLAG", pred = "pred", title = "Calibration pl

train.clean.binary <- mutate(train.clean.binary, predout=ifelse(pred < 0.5, 0, 1))

#Create confusion matrix
cm <- confusionMatrix(as.factor(train.clean.binary$predout), as.factor(train.clean.binary$TARGET_FLAG))

#Calculate AUC
auc_res <- auc(train.clean.binary$TARGET_FLAG, train.clean.binary$pred)

sprintf("Model 2 Classification Accuracy is: %.2f%%",(cm$overall[1])*100)
sprintf("Model 2 Classification Error Rate is: %.2f%%",(1-cm$overall[1])*100)
sprintf("Model 2 Precision is: %.2f%%",(cm$byClass['Pos Pred Value']*100))
sprintf("Model 2 Sensitivity/Recall is: %.2f%%",(cm$byClass['Sensitivity']*100))
sprintf("Model 2 Specificity is: %.2f%%",(cm$byClass['Specificity']*100))
sprintf("Model 2 F1-score is: %.2f%%",(cm$byClass['F1']*100))
sprintf("Model 2 AUC is: %.2f%%",(auc_res[1]*100))


true_labels <- train.clean.binary$TARGET_FLAG
roc_curve <- roc(true_labels, fit2$fitted.values)
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)
abline(a = 0, b = 1, col = "gray", lty = 2)
legend("bottomright", legend = paste("AUC =", round(auc(roc_curve), 2)), col = "blue", lwd = 2)

#Removing index and TARGET_AMT
testing.set <-testing_data %>%
  dplyr::select(-c(INDEX, TARGET_AMT, TARGET_FLAG))

#converting the spend columns back to numeric
testing.clean <- testing.set %>%
  mutate_at(c("INCOME","HOME_VAL","BLUEBOOK","OLDCLAIM"),~parse_number(.))

#Using MICE to impute missing values, using pmm to avoid neg impute values
testing.clean <- complete(mice(testing.clean,method = "pmm",seed = 333))


#Mutating the columns with two value into a binary dummy version below
#It's under the assumption, false==0 and true==1
testing.clean<-testing.clean%>%mutate(PARENT1=if_else(PARENT1=="No",0,1))
testing.clean<-testing.clean%>%mutate(MSTATUS=if_else(MSTATUS=="z_No",0,1))
testing.clean<-testing.clean%>%mutate(SEX=if_else(SEX=="M",0,1))
testing.clean<-testing.clean%>%mutate(CAR_USE=if_else(CAR_USE=="Private",0,1))
```

```
testing.clean<-testing.clean%>%mutate(RED_CAR=if_else(RED_CAR=="no",0,1))
testing.clean<-testing.clean%>%mutate(REVOKED=if_else(REVOKED=="No",0,1))
testing.clean<-testing.clean%>%mutate(URBANICITY=if_else(URBANICITY=="z_Highly Rural/ Rural",0,1))

# Following the K-1 format, each variable lowest choice does not receive a column
#i.e highest education, the no high school diploma does not get a column
testing.clean<-testing.clean%>%mutate(.isDiploma=if_else(EDUCATION=="z_High School",1,0),
                                      .isBach=if_else(EDUCATION=="Bachelors",1,0),
                                      .isMasters=if_else(EDUCATION=="Masters",1,0),
                                      .isPhd=if_else(EDUCATION=="PhD",1,0)
                                      )
#assuming unemployed lowest level to not deal with a NA
testing.clean<-testing.clean%>%mutate(.isProf=if_else(JOB=="Professional",1,0),
                                      .isBlue=if_else(JOB=="z_Blue Collar",1,0),
                                      .isClerk=if_else(JOB=="Clerical",1,0),
                                      .isDoctor=if_else(JOB=="Doctor",1,0),
                                      .isLawyer=if_else(JOB=="Lawyer",1,0),
                                      .isHome=if_else(JOB=="Home Maker",1,0),
                                      .isStudent=if_else(JOB=="Student",1,0),
                                      .isManager=if_else(JOB=="Manager",1,0)
                                      )

#assume panel truck is lowest
testing.clean<-testing.clean%>%mutate(.isMini=if_else(CAR_TYPE=="Minivan",1,0),
                                      .isSUV=if_else(CAR_TYPE=="z_SUV",1,0),
                                      .isSport=if_else(CAR_TYPE=="Sports Car",1,0),
                                      .isVan=if_else(CAR_TYPE=="Van",1,0),
                                      .isPickup=if_else(CAR_TYPE=="Pickup",1,0)
                                      )

#removing categorical columns after dummies are set
testing.clean<-testing.clean%>%dplyr::select(-c(EDUCATION,CAR_TYPE,JOB))

#Applying boxcox transformations
testing.clean<-testing.clean%>%mutate(INCOME=sqrt(INCOME+1))
testing.clean<-testing.clean%>%mutate(HOME_VAL=log(HOME_VAL+1))
testing.clean<-testing.clean%>%mutate(BLUEBOOK=sqrt(BLUEBOOK))
testing.clean<-testing.clean%>%mutate(OLDCLAIM=sqrt(OLDCLAIM+1))
testing.clean<-testing.clean%>%mutate(AGE=(AGE**lamb.AGE-1)/lamb.AGE)
testing.clean<-testing.clean%>%mutate(TRAVTIME=sqrt(TRAVTIME))
testing.clean<-testing.clean%>%mutate(YOJ=YOJ**2)
testing.clean<-testing.clean%>%mutate(MVR_PTS=MVR_PTS**2)
testing.clean<-testing.clean%>%mutate(CLM_FREQ=CLM_FREQ**3)
testing.clean<-testing.clean%>%mutate(CAR_AGE=CAR_AGE**2)

#assigning dummy variables as factors
testing.clean<-testing.clean%>%mutate_at(c(".isDiploma",".isBach",".isMasters",".isPhd",".isProf",".isB

testing.clean$pred_prob <- predict(fit2, testing.clean, type="response")
testing.clean <- mutate(testing.clean, predout=ifelse(pred_prob < 0.5, 0, 1))

knitr::kable(head(testing.clean,10) , "pipe")
```

```r
knitr::kable(table(testing.clean$predout), "pipe")

#Remove TARGET_FLAG
train.clean.linear <- subset(train.clean, select = -c(TARGET_FLAG))

#Remove 0 and Nulls
train.clean.linear <- train.clean.linear[!(train.clean.linear$TARGET_AMT %in% c('0', 'NA')), ]

train.clean.linear<-train.clean.linear%>%mutate_at(c(".isDiploma",".isBach",".isMasters",".isPhd",".isP

#first model with all predictors
lm1 <- lm(TARGET_AMT~., data = train.clean.linear)
summary(lm1)

par(mfrow=c(2, 2))
plot(lm1)

lm2 <- stats::step(lm1,direction="backward")
lm2<-lm(TARGET_AMT ~ HOME_VAL + MSTATUS + SEX + BLUEBOOK + REVOKED +MVR_PTS + CAR_AGE + .isPhd +.isDocto
summary(lm2)

par(mfrow=c(2, 2))
plot(lm2)
linear_model_res <- bind_rows(glance(lm1), glance(lm2))
model_names1 <- c("Linear model 1","Linear model 2")
linear_model_res <- cbind(model.build = model_names1, linear_model_res)
linear_model_res

#Plotting residual vs fitted values for visual representation of model peformance
g1<-ggplot(data = lm1 , aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")+theme_classic()
g2<-ggplot(data = lm1 , aes(x = .resid)) +
    geom_histogram(binwidth = 25) +
    xlab("Residuals")+theme_classic()
g3<-ggplot(data = lm1 , aes(sample = .resid)) +
  stat_qq()+theme_classic()
plt<-ggarrange(g1,g2,g3)
annotate_figure(plt,top = text_grob("Model 1's Residual Performance",size=9))

g1<-ggplot(data = lm2 , aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")+theme_classic()
g2<-ggplot(data = lm2 , aes(x = .resid)) +
    geom_histogram(binwidth = 25) +
    xlab("Residuals")+theme_classic()
g3<-ggplot(data = lm2 , aes(sample = .resid)) +
  stat_qq()+theme_classic()
plt<-ggarrange(g1,g2,g3)
```

```r
annotate_figure(plt,top = text_grob("Model 2's Residual Performance",size=9))
```