Table of Contents

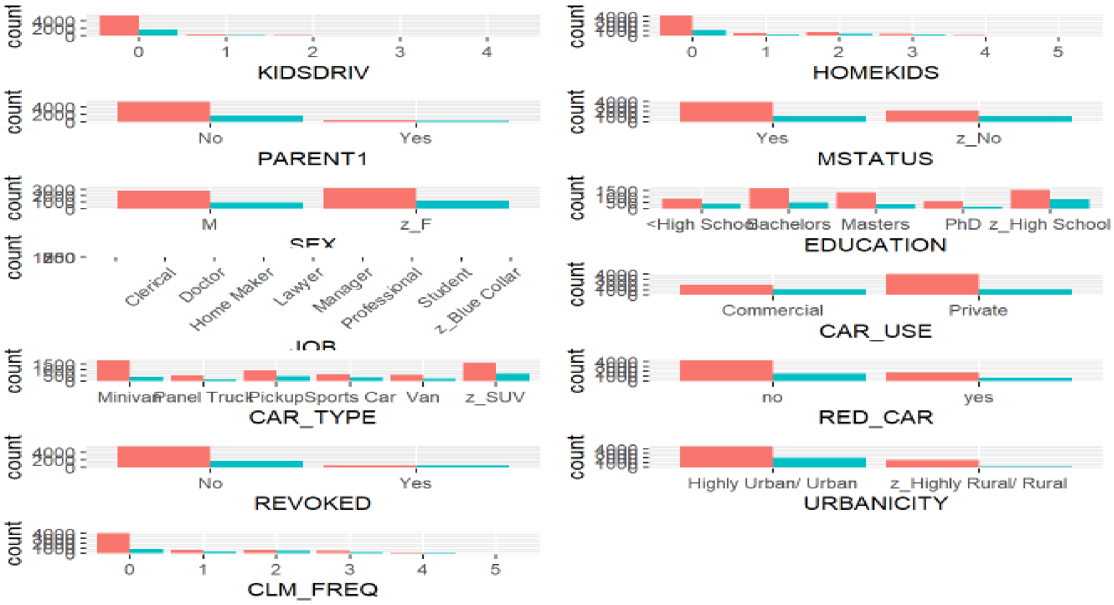
DATA EXPLORATION

The dataset has 26 variables and 8161 observations. **TARGET_FLAG** and **TARGET_AMT** are our response variables. 13 of the variables have discrete values and the rest of the variables are continuous.

```
## 'data.frame':    8161 obs. of  26 variables:
##  $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
##  $ TARGET_AMT : num  0 0 0 0 0 ...
##  $ KIDSDRIV   : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
##  $ INCOME     : chr  "$67,349" "$91,449" "$16,039" "" ...
##  $ PARENT1    : chr  "No" "No" "No" "No" ...
##  $ HOME_VAL   : chr  "$0" "$257,252" "$124,191" "$306,251" ...
##  $ MSTATUS    : chr  "z_No" "z_No" "Yes" "Yes" ...
##  $ SEX        : chr  "M" "M" "z_F" "M" ...
##  $ EDUCATION  : chr  "PhD" "z_High School" "z_High School" "<High School" ...
##  $ JOB        : chr  "Professional" "z_Blue Collar" "Clerical" "z_Blue Collar" ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : chr  "Private" "Commercial" "Private" "Private" ...
##  $ BLUEBOOK   : chr  "$14,230" "$14,940" "$4,010" "$15,440" ...
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : chr  "Minivan" "Minivan" "z_SUV" "Minivan" ...
##  $ RED_CAR    : chr  "yes" "yes" "no" "yes" ...
##  $ OLDCLAIM   : chr  "$4,461" "$0" "$38,690" "$0" ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : chr  "No" "No" "No" "No" ...
##  $ MVR_PTS    : int  3 0 3 0 3 0 0 10 0 1 ...
##  $ CAR_AGE    : int  18 1 10 6 17 7 1 7 1 17 ...
##  $ URBANICITY : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban"
...
```

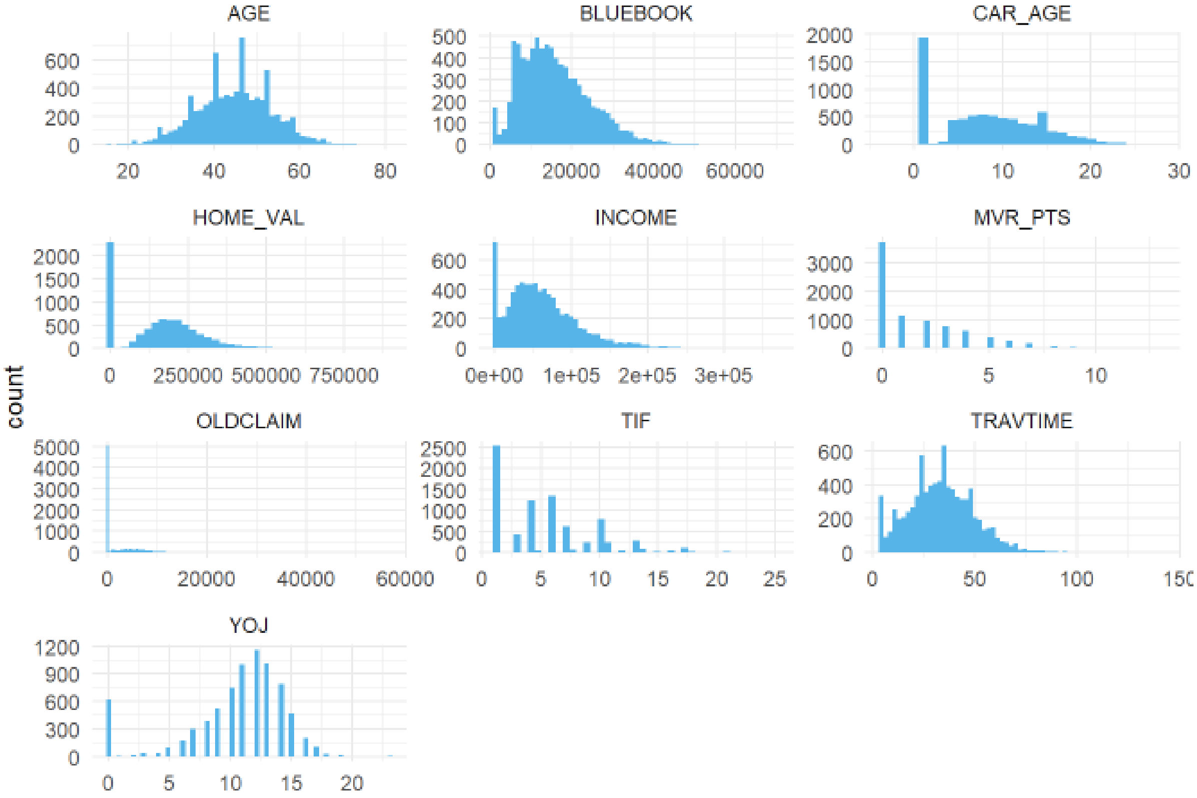
Bar Chart

Variables **HOMEKIDS**, **PARENT1, KIDSDRIV** show that having no kids results in more car crash. Sex, marital status or the type of car don't have significant effect on car crash. Blue collar employees, or SUV owners get into more car crash. If the individual's license is revoked or they are driving on an urban area, they have higher chance of car crash.



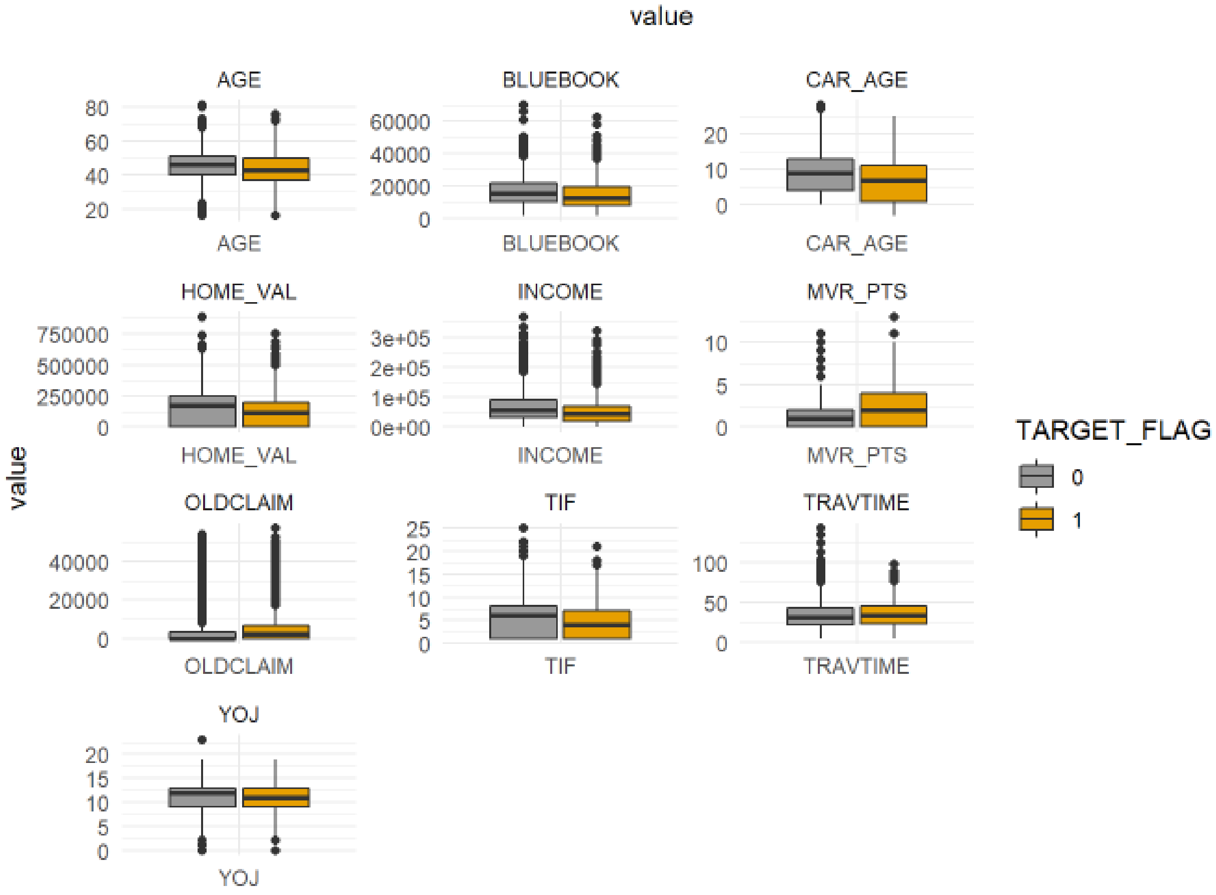
Histogram plot

`AGE` is the only variable that is normally distributed the rest are skewed.



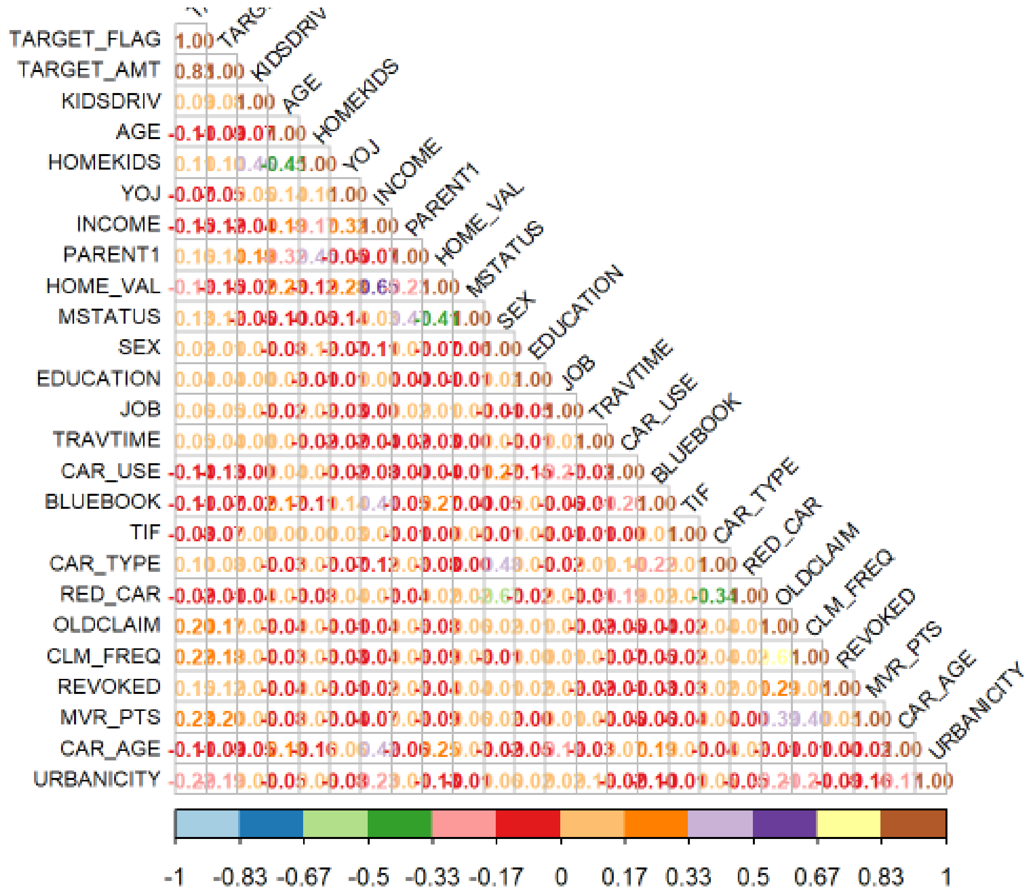
Box-plot

`BLUEBOOK`, `INCOME`, `OLDCLAIM` have high number of outliers compared to other variables. Individuals with older car, higher home value, higher income or older customer get in to less car crash. In addition, people with motor vehicle record points or high number of old claims have more car crash.



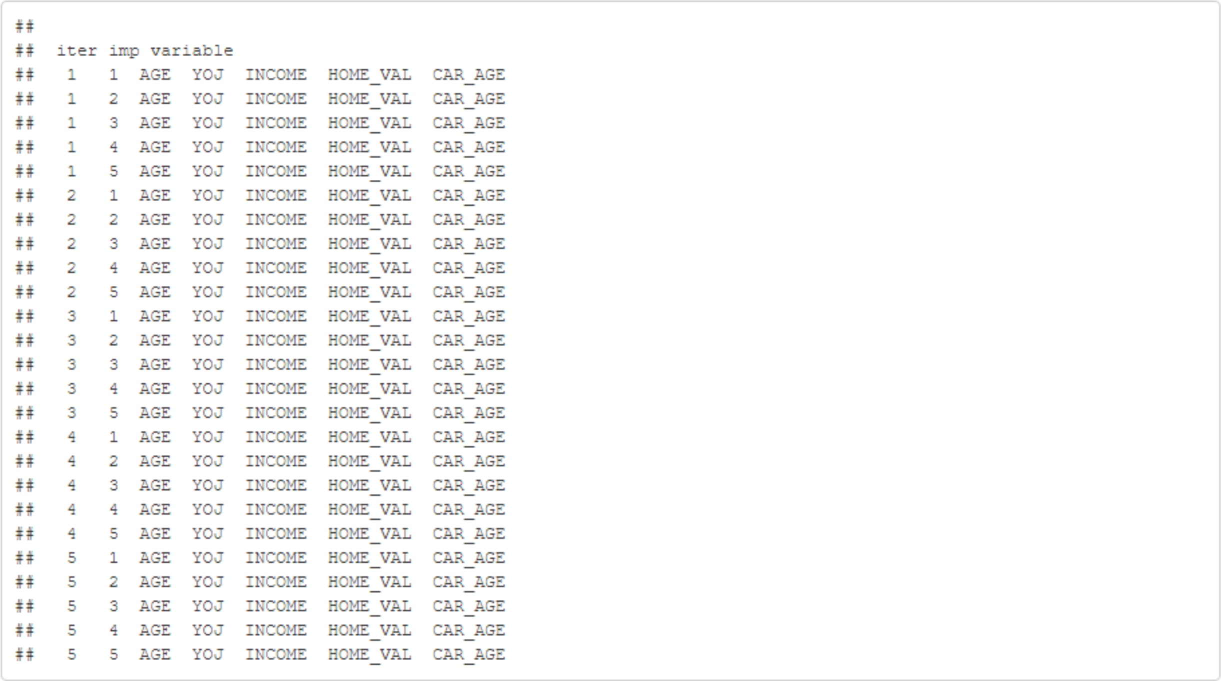
Correlations

`MVR_PTS`, `CLM_FREQ`, and `OLDCLAIM` are the most positively correlated variables with the response variables. `URBANICITY` is most negatively correlated variable. Other variables are weakly correlated.



DATA PREPARATION

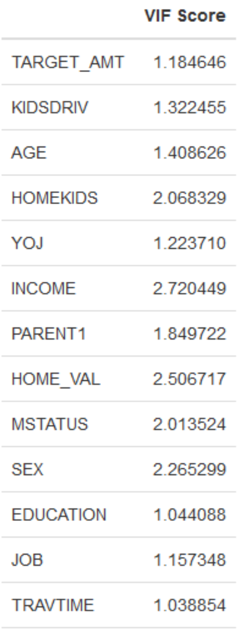
There is need to remove dollar sign and comma from **INCOME**, **HOME_VAL**, **BLUEBOOK**, **OLDCLAIM** variables and convert these variables to integer. Imputation is then performed to variables with missing data. Multiple imputations help in resolving the uncertainty for the missing data.



```
## 
##  iter imp variable
##   1   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   1   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   1   3  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   1   4  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   1   5  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   3  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   4  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   5  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   3  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   4  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   5  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   3  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   4  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   5  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   3  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   4  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   5  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
```

```
## [1] "Missing value after imputation: 0"
```

The VIF score is at a conservative level for all variables, hence no major concern on collinearity.



	VIF Score
TARGET_AMT	1.184646
KIDSDRIV	1.322455
AGE	1.408626
HOMEKIDS	2.068329
YOJ	1.223710
INCOME	2.720449
PARENT1	1.849722
HOME_VAL	2.506717
MSTATUS	2.013524
SEX	2.265299
EDUCATION	1.044088
JOB	1.157348
TRAVTIME	1.038854

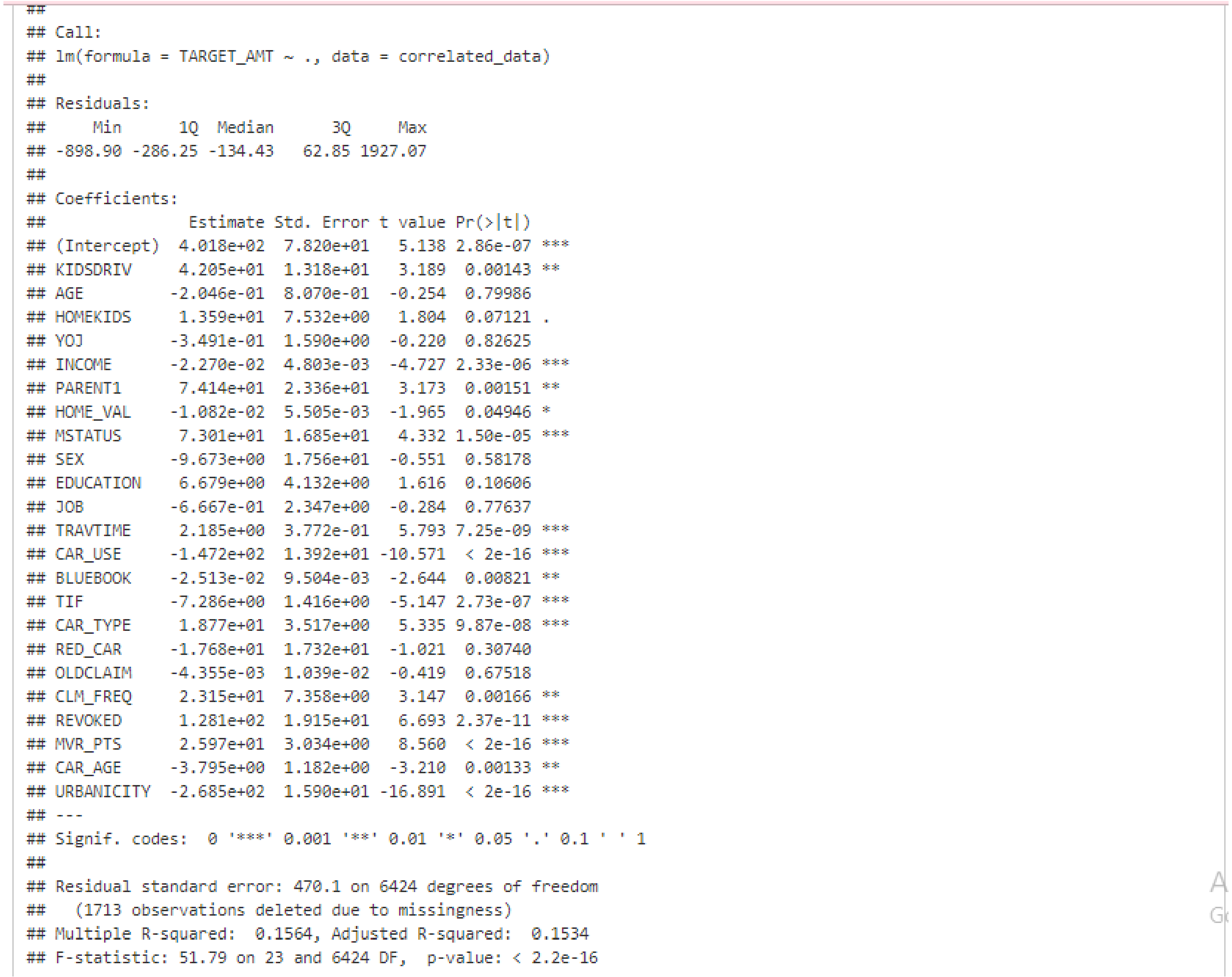
TRAVTIME	1.038854
CAR_USE	1.353302
BLUEBOOK	1.375440
TIF	1.009161
CAR_TYPE	1.409798
RED_CAR	1.809696
OLDCLAIM	2.201664
CLM_FREQ	2.131016
REVOKED	1.148628
MVR_PTS	1.249189
CAR_AGE	1.311790
URBANICITY	1.243781

BUILD MODELS

Multiple Linear Regression

Model 1

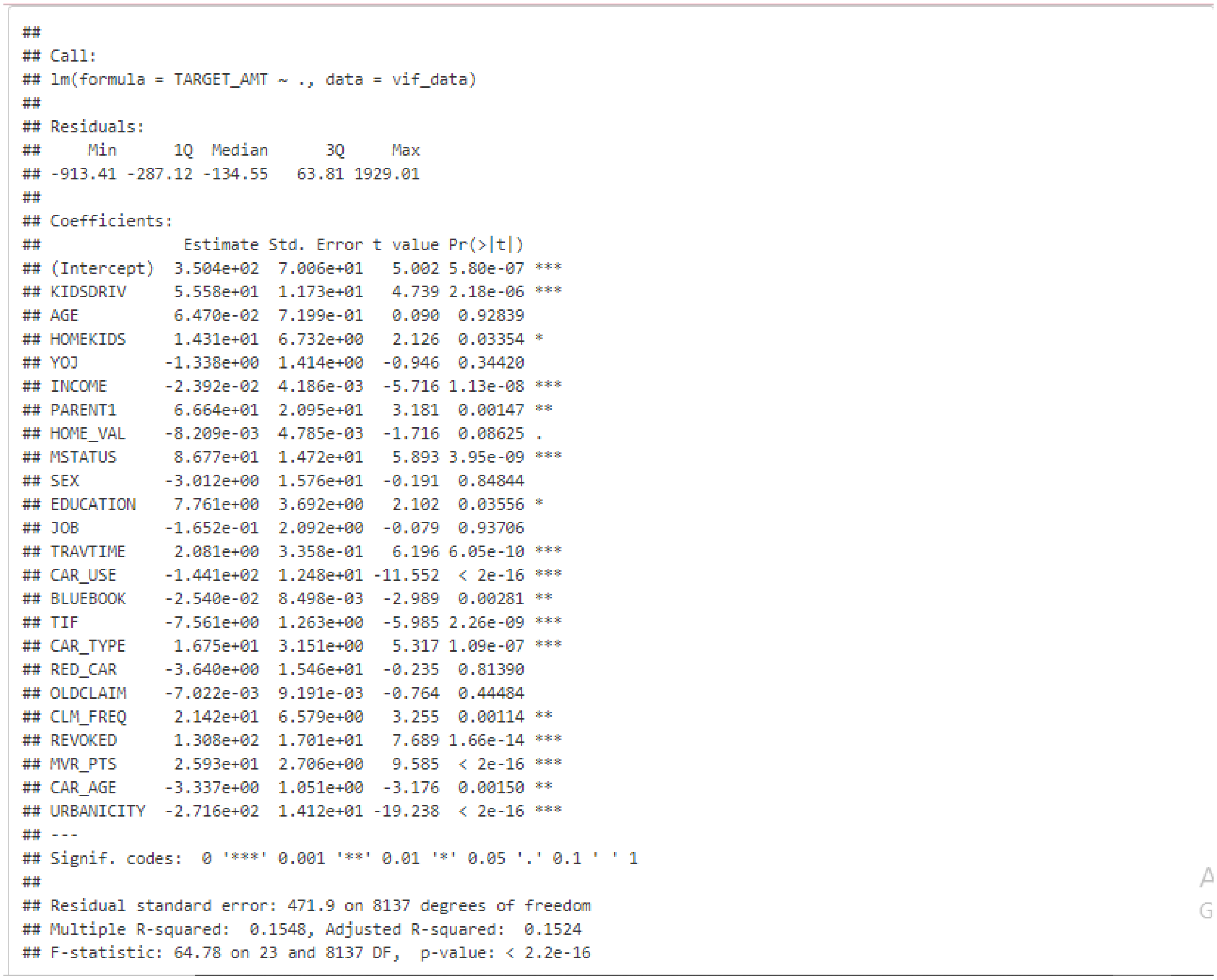
In the linear regression model below, the min-max and 1Q-3Q have different magnitudes and the median is not close to zero. R-squared is 0.1564, implying that this model explains 15.64% of the data’s variation. Hence, in overall, this is a bad model.



```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = correlated_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -898.90 -286.25 -134.43   62.85 1927.07
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.018e+02  7.820e+01   5.138 2.86e-07 ***
## KIDSDRIV     4.205e+01  1.318e+01   3.189  0.00143 **
## AGE         -2.046e-01  8.070e-01  -0.254  0.79986
## HOMEKIDS     1.359e+01  7.532e+00   1.804  0.07121 .
## YOJ         -3.491e-01  1.590e+00  -0.220  0.82625
## INCOME      -2.270e-02  4.803e-03  -4.727 2.33e-06 ***
## PARENT1      7.414e+01  2.336e+01   3.173  0.00151 **
## HOME_VAL    -1.082e-02  5.505e-03  -1.965  0.04946 *
## MSTATUS      7.301e+01  1.685e+01   4.332 1.50e-05 ***
## SEX         -9.673e+00  1.756e+01  -0.551  0.58178
## EDUCATION    6.679e+00  4.132e+00   1.616  0.10606
## JOB         -6.667e-01  2.347e+00  -0.284  0.77637
## TRAVTIME     2.185e+00  3.772e-01   5.793 7.25e-09 ***
## CAR_USE     -1.472e+02  1.392e+01 -10.571  < 2e-16 ***
## BLUEBOOK    -2.513e-02  9.504e-03  -2.644  0.00821 **
## TIF         -7.286e+00  1.416e+00  -5.147 2.73e-07 ***
## CAR_TYPE     1.877e+01  3.517e+00   5.335 9.87e-08 ***
## RED_CAR     -1.768e+01  1.732e+01  -1.021  0.30740
## OLDCLAIM    -4.355e-03  1.039e-02  -0.419  0.67518
## CLM_FREQ     2.315e+01  7.358e+00   3.147  0.00166 **
## REVOKED      1.281e+02  1.915e+01   6.693 2.37e-11 ***
## MVR_PTS      2.597e+01  3.034e+00   8.560  < 2e-16 ***
## CAR_AGE     -3.795e+00  1.182e+00  -3.210  0.00133 **
## URBANICITY  -2.685e+02  1.590e+01 -16.891  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 470.1 on 6424 degrees of freedom
##   (1713 observations deleted due to missingness)
## Multiple R-squared:  0.1564, Adjusted R-squared:  0.1534
## F-statistic: 51.79 on 23 and 6424 DF,  p-value: < 2.2e-16
```

Model 2

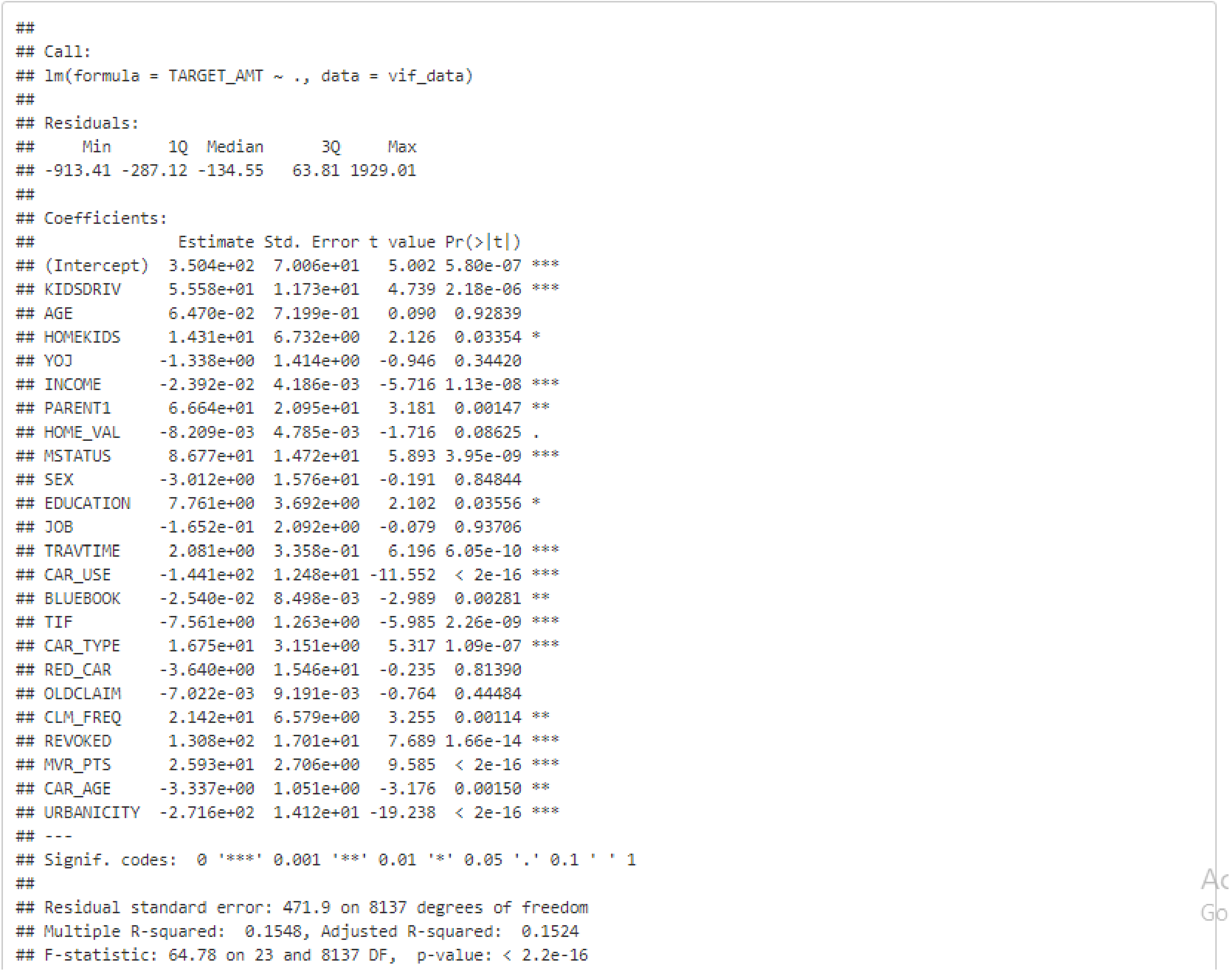
Still with this model, the min-max and 1Q-3Q have different magnitudes and the median is not close to zero. R-squared is 0.15.48, meaning the model explains 15.48% of the data’s variation. This is also a bad model.



```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = vif_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -913.41 -287.12 -134.55   63.81 1929.01
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.504e+02  7.006e+01   5.002 5.80e-07 ***
## KIDSDRIV     5.558e+01  1.173e+01   4.739 2.18e-06 ***
## AGE          6.470e-02  7.199e-01   0.090  0.92839
## HOMEKIDS     1.431e+01  6.732e+00   2.126  0.03354 *
## YOJ         -1.338e+00  1.414e+00  -0.946  0.34420
## INCOME      -2.392e-02  4.186e-03  -5.716 1.13e-08 ***
## PARENT1      6.664e+01  2.095e+01   3.181  0.00147 **
## HOME_VAL    -8.209e-03  4.785e-03  -1.716  0.08625 .
## MSTATUS      8.677e+01  1.472e+01   5.893 3.95e-09 ***
## SEX         -3.012e+00  1.576e+01  -0.191  0.84844
## EDUCATION    7.761e+00  3.692e+00   2.102  0.03556 *
## JOB         -1.652e-01  2.092e+00  -0.079  0.93706
## TRAVTIME     2.081e+00  3.358e-01   6.196 6.05e-10 ***
## CAR_USE     -1.441e+02  1.248e+01 -11.552  < 2e-16 ***
## BLUEBOOK    -2.540e-02  8.498e-03  -2.989  0.00281 **
## TIF         -7.561e+00  1.263e+00  -5.985 2.26e-09 ***
## CAR_TYPE     1.675e+01  3.151e+00   5.317 1.09e-07 ***
## RED_CAR     -3.640e+00  1.546e+01  -0.235  0.81390
## OLDCLAIM    -7.022e-03  9.191e-03  -0.764  0.44484
## CLM_FREQ     2.142e+01  6.579e+00   3.255  0.00114 **
## REVOKED      1.308e+02  1.701e+01   7.689 1.66e-14 ***
## MVR_PTS      2.593e+01  2.706e+00   9.585  < 2e-16 ***
## CAR_AGE     -3.337e+00  1.051e+00  -3.176  0.00150 **
## URBANICITY  -2.716e+02  1.412e+01 -19.238  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 471.9 on 8137 degrees of freedom
## Multiple R-squared:  0.1548, Adjusted R-squared:  0.1524
## F-statistic: 64.78 on 23 and 8137 DF,  p-value: < 2.2e-16
```

Model 3

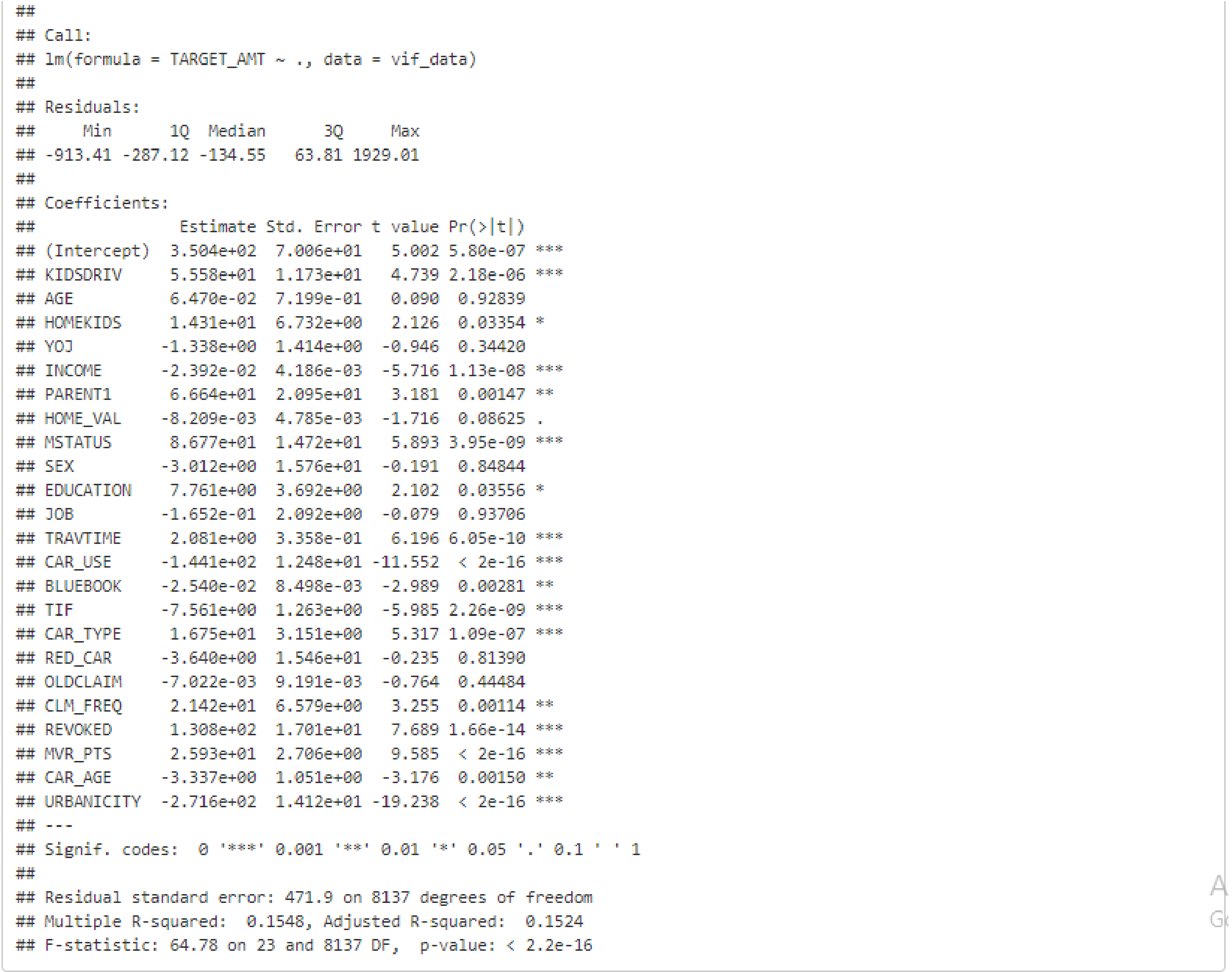
The min-max and 1Q-3Q have different magnitudes and the median is not close to zero. The p-value shows that the probability of this variables to be irrelevant is very low. R-squared is 0.1546, which means this model explains 15.46% of the data’s variation. However, there is improved p-value for several variables. It is a better model than the two previous models.



```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = vif_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -913.41 -287.12 -134.55   63.81 1929.01
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.504e+02  7.006e+01   5.002 5.80e-07 ***
## KIDSDRIV     5.558e+01  1.173e+01   4.739 2.18e-06 ***
## AGE          6.470e-02  7.199e-01   0.090  0.92839
## HOMEKIDS     1.431e+01  6.732e+00   2.126  0.03354 *
## YOJ         -1.338e+00  1.414e+00  -0.946  0.34420
## INCOME      -2.392e-02  4.186e-03  -5.716 1.13e-08 ***
## PARENT1      6.664e+01  2.095e+01   3.181  0.00147 **
## HOME_VAL    -8.209e-03  4.785e-03  -1.716  0.08625 .
## MSTATUS      8.677e+01  1.472e+01   5.893 3.95e-09 ***
## SEX         -3.012e+00  1.576e+01  -0.191  0.84844
## EDUCATION    7.761e+00  3.692e+00   2.102  0.03556 *
## JOB         -1.652e-01  2.092e+00  -0.079  0.93706
## TRAVTIME     2.081e+00  3.358e-01   6.196 6.05e-10 ***
## CAR_USE     -1.441e+02  1.248e+01 -11.552  < 2e-16 ***
## BLUEBOOK    -2.540e-02  8.498e-03  -2.989  0.00281 **
## TIF         -7.561e+00  1.263e+00  -5.985 2.26e-09 ***
## CAR_TYPE     1.675e+01  3.151e+00   5.317 1.09e-07 ***
## RED_CAR     -3.640e+00  1.546e+01  -0.235  0.81390
## OLDCLAIM    -7.022e-03  9.191e-03  -0.764  0.44484
## CLM_FREQ     2.142e+01  6.579e+00   3.255  0.00114 **
## REVOKED      1.308e+02  1.701e+01   7.689 1.66e-14 ***
## MVR_PTS      2.593e+01  2.706e+00   9.585  < 2e-16 ***
## CAR_AGE     -3.337e+00  1.051e+00  -3.176  0.00150 **
## URBANICITY  -2.716e+02  1.412e+01 -19.238  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 471.9 on 8137 degrees of freedom
## Multiple R-squared:  0.1548, Adjusted R-squared:  0.1524
## F-statistic: 64.78 on 23 and 8137 DF,  p-value: < 2.2e-16
```

Model 4

The min-max and 1Q-3Q have quite similar magnitudes and the median is close to zero. The p-value below shows that the probability of this variables to be irrelevant is very low. R-squared is 0.2146, implying that this model explains 21.46% of the data's variation. This s a good model.

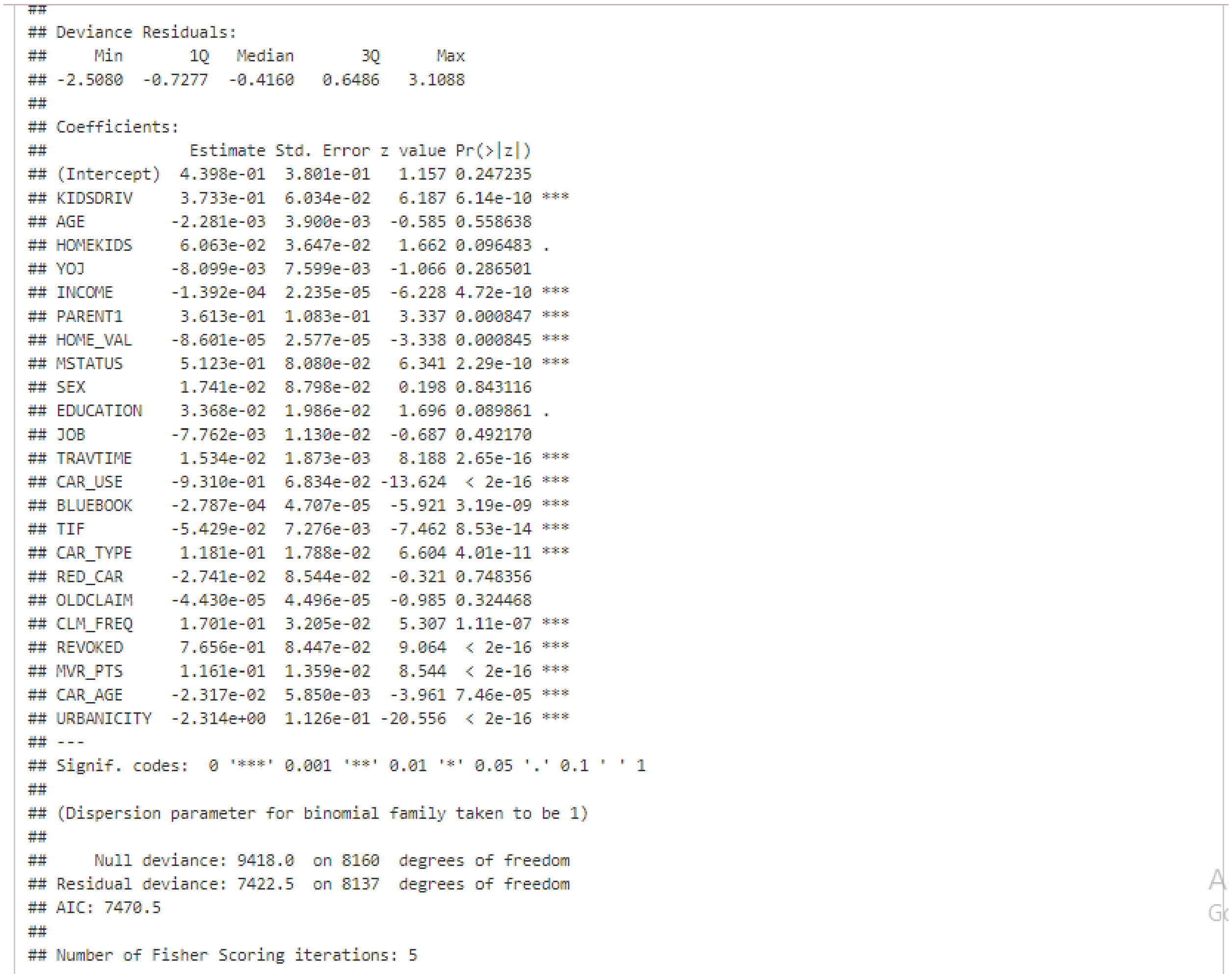


```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = vif_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -913.41 -287.12 -134.55   63.81 1929.01
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.504e+02  7.006e+01   5.002 5.80e-07 ***
## KIDSDRIV     5.558e+01  1.173e+01   4.739 2.18e-06 ***
## AGE          6.470e-02  7.199e-01   0.090  0.92839
## HOMEKIDS     1.431e+01  6.732e+00   2.126  0.03354 *
## YOJ         -1.338e+00  1.414e+00  -0.946  0.34420
## INCOME      -2.392e-02  4.186e-03  -5.716 1.13e-08 ***
## PARENT1      6.664e+01  2.095e+01   3.181  0.00147 **
## HOME_VAL    -8.209e-03  4.785e-03  -1.716  0.08625 .
## MSTATUS      8.677e+01  1.472e+01   5.893 3.95e-09 ***
## SEX         -3.012e+00  1.576e+01  -0.191  0.84844
## EDUCATION    7.761e+00  3.692e+00   2.102  0.03556 *
## JOB         -1.652e-01  2.092e+00  -0.079  0.93706
## TRAVTIME     2.081e+00  3.358e-01   6.196 6.05e-10 ***
## CAR_USE     -1.441e+02  1.248e+01 -11.552  < 2e-16 ***
## BLUEBOOK    -2.540e-02  8.498e-03  -2.989  0.00281 **
## TIF         -7.561e+00  1.263e+00  -5.985 2.26e-09 ***
## CAR_TYPE     1.675e+01  3.151e+00   5.317 1.09e-07 ***
## RED_CAR     -3.640e+00  1.546e+01  -0.235  0.81390
## OLDCLAIM    -7.022e-03  9.191e-03  -0.764  0.44484
## CLM_FREQ     2.142e+01  6.579e+00   3.255  0.00114 **
## REVOKED      1.308e+02  1.701e+01   7.689 1.66e-14 ***
## MVR_PTS      2.593e+01  2.706e+00   9.585  < 2e-16 ***
## CAR_AGE     -3.337e+00  1.051e+00  -3.176  0.00150 **
## URBANICITY  -2.716e+02  1.412e+01 -19.238  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 471.9 on 8137 degrees of freedom
## Multiple R-squared:  0.1548, Adjusted R-squared:  0.1524
## F-statistic: 64.78 on 23 and 8137 DF,  p-value: < 2.2e-16
```

Binary Logistic Regression

Model 5

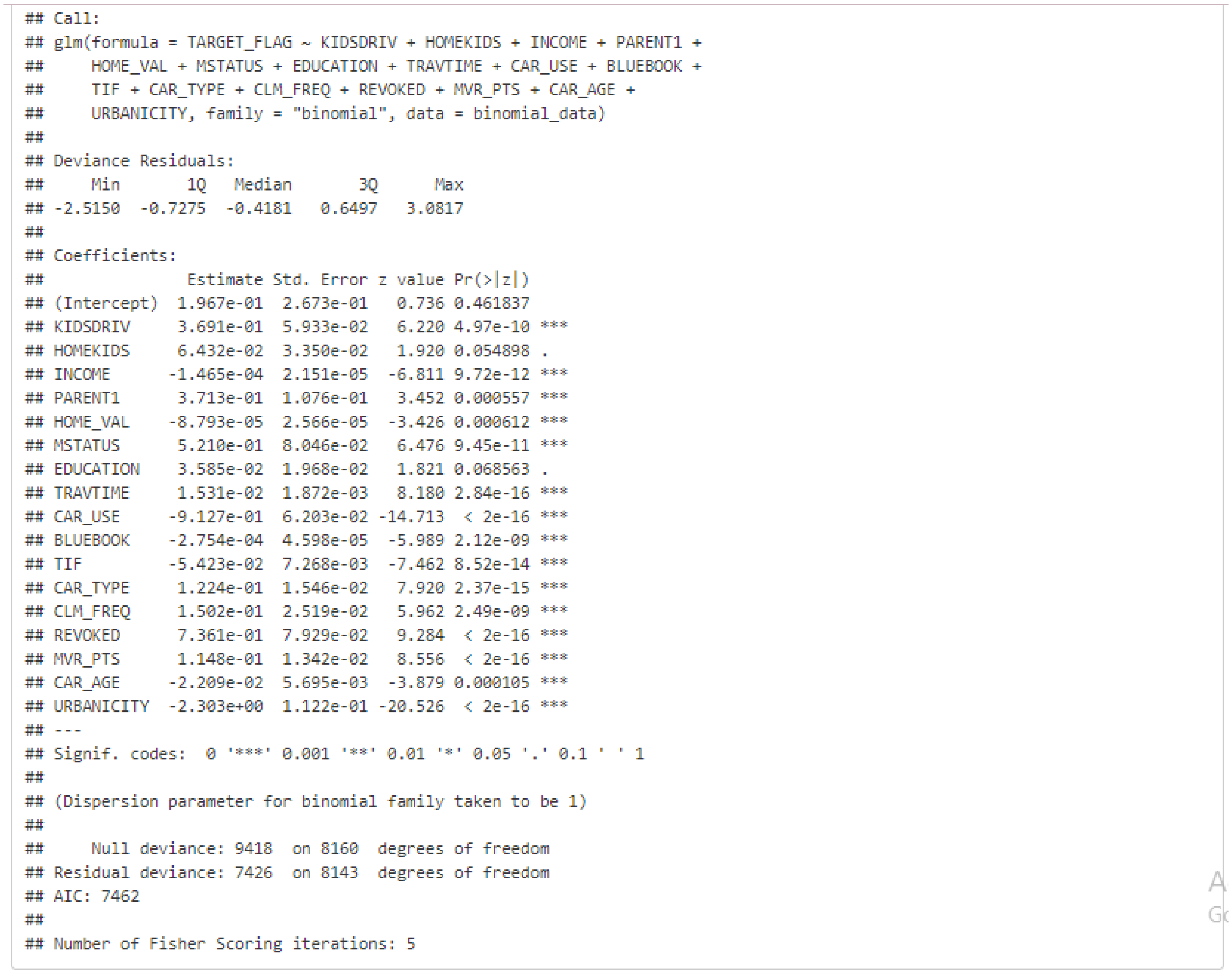
The min-max and 1Q-3Q magnitudes is quite close and the median is close to zero. The P-value shows many variables are significant.



```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5080  -0.7277  -0.4160   0.6486   3.1088
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.398e-01  3.801e-01   1.157 0.247235
## KIDSDRIV     3.733e-01  6.034e-02   6.187 6.14e-10 ***
## AGE         -2.281e-03  3.900e-03  -0.585 0.558638
## HOMEKIDS     6.063e-02  3.647e-02   1.662 0.096483 .
## YOJ         -8.099e-03  7.599e-03  -1.066 0.286501
## INCOME      -1.392e-04  2.235e-05  -6.228 4.72e-10 ***
## PARENT1      3.613e-01  1.083e-01   3.337 0.000847 ***
## HOME_VAL    -8.601e-05  2.577e-05  -3.338 0.000845 ***
## MSTATUS      5.123e-01  8.080e-02   6.341 2.29e-10 ***
## SEX          1.741e-02  8.798e-02   0.198 0.843116
## EDUCATION    3.368e-02  1.986e-02   1.696 0.089861 .
## JOB         -7.762e-03  1.130e-02  -0.687 0.492170
## TRAVTIME     1.534e-02  1.873e-03   8.188 2.65e-16 ***
## CAR_USE     -9.310e-01  6.834e-02 -13.624  < 2e-16 ***
## BLUEBOOK    -2.787e-04  4.707e-05  -5.921 3.19e-09 ***
## TIF         -5.429e-02  7.276e-03  -7.462 8.53e-14 ***
## CAR_TYPE     1.181e-01  1.788e-02   6.604 4.01e-11 ***
## RED_CAR     -2.741e-02  8.544e-02  -0.321 0.748356
## OLDCLAIM    -4.430e-05  4.496e-05  -0.985 0.324468
## CLM_FREQ     1.701e-01  3.205e-02   5.307 1.11e-07 ***
## REVOKED      7.656e-01  8.447e-02   9.064  < 2e-16 ***
## MVR_PTS      1.161e-01  1.359e-02   8.544  < 2e-16 ***
## CAR_AGE     -2.317e-02  5.850e-03  -3.961 7.46e-05 ***
## URBANICITY  -2.314e+00  1.126e-01 -20.556  < 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: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7422.5  on 8137  degrees of freedom
## AIC: 7470.5
##
## Number of Fisher Scoring iterations: 5
```

Model 6

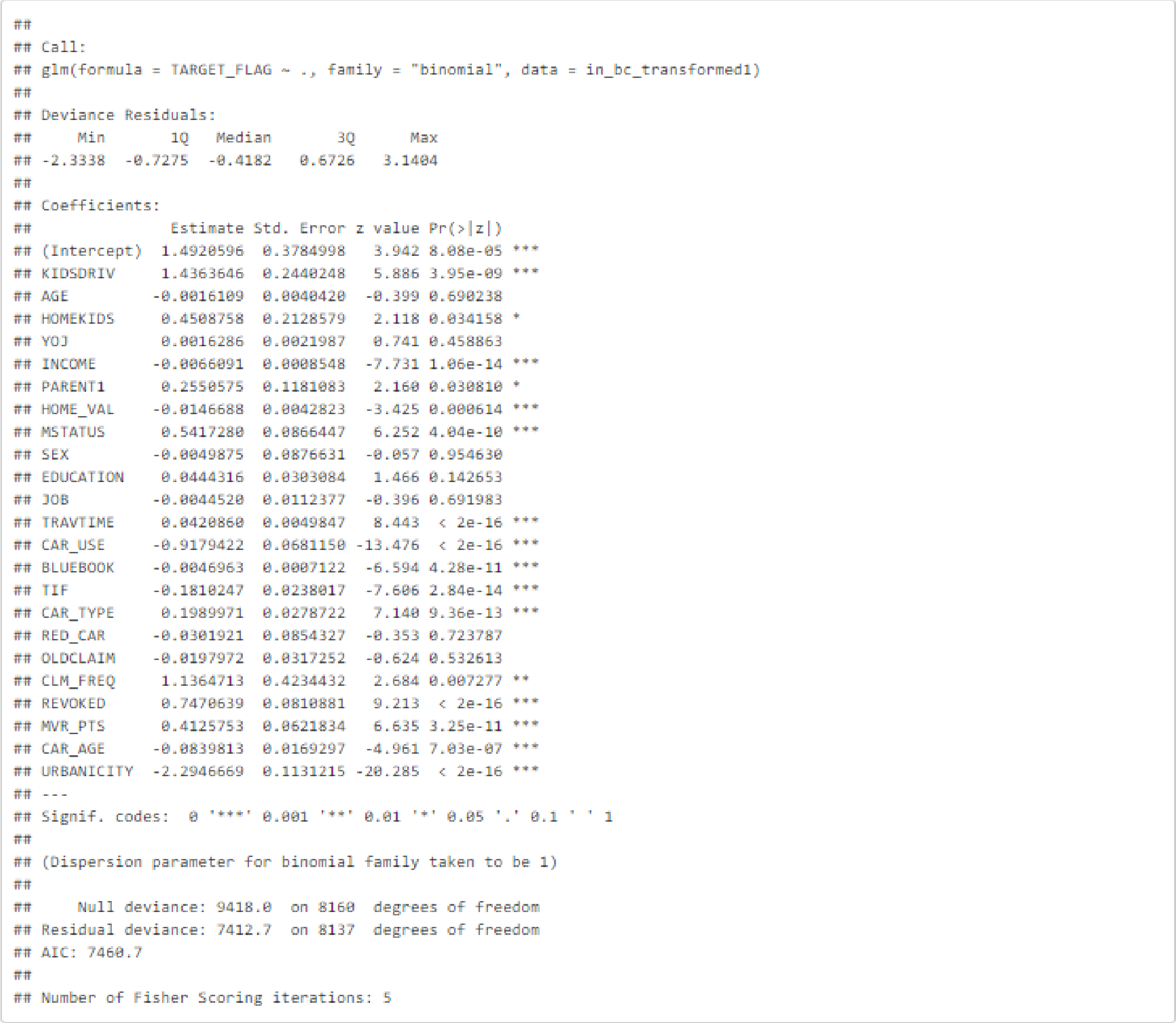
This model made use of forward and backward step-wise variables selection algorithm. The min-max and 1Q-3Q magnitudes is quite close and the median is close to zero. This model's variables selection is better with better p-value. However, AIC score has not improved from the previous model.



```
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 + 
##     HOME_VAL + MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK + 
##     TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + 
##     URBANICITY, family = "binomial", data = binomial_data)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.5150  -0.7275  -0.4181   0.6497   3.0817  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  1.967e-01  2.673e-01   0.736 0.461837    
## KIDSDRIV     3.691e-01  5.933e-02   6.220 4.97e-10 ***
## HOMEKIDS     6.432e-02  3.350e-02   1.920 0.054898 .  
## INCOME      -1.465e-04  2.151e-05  -6.811 9.72e-12 ***
## PARENT1      3.713e-01  1.076e-01   3.452 0.000557 ***
## HOME_VAL    -8.793e-05  2.566e-05  -3.426 0.000612 ***
## MSTATUS      5.210e-01  8.046e-02   6.476 9.45e-11 ***
## EDUCATION    3.585e-02  1.968e-02   1.821 0.068563 .  
## TRAVTIME     1.531e-02  1.872e-03   8.180 2.84e-16 ***
## CAR_USE     -9.127e-01  6.203e-02 -14.713  < 2e-16 ***
## BLUEBOOK    -2.754e-04  4.598e-05  -5.989 2.12e-09 ***
## TIF         -5.423e-02  7.268e-03  -7.462 8.52e-14 ***
## CAR_TYPE     1.224e-01  1.546e-02   7.920 2.37e-15 ***
## CLM_FREQ     1.502e-01  2.519e-02   5.962 2.49e-09 ***
## REVOKED      7.361e-01  7.929e-02   9.284  < 2e-16 ***
## MVR_PTS      1.148e-01  1.342e-02   8.556  < 2e-16 ***
## CAR_AGE     -2.209e-02  5.695e-03  -3.879 0.000105 ***
## URBANICITY  -2.303e+00  1.122e-01 -20.526  < 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: 9418  on 8160  degrees of freedom
## Residual deviance: 7426  on 8143  degrees of freedom
## AIC: 7462
## 
## Number of Fisher Scoring iterations: 5
```

Model 7

The min-max and 1Q-3Q magnitudes is quite close and the median is close to zero. Many variables have significant p-value. This model has the best AIC score among the three models.

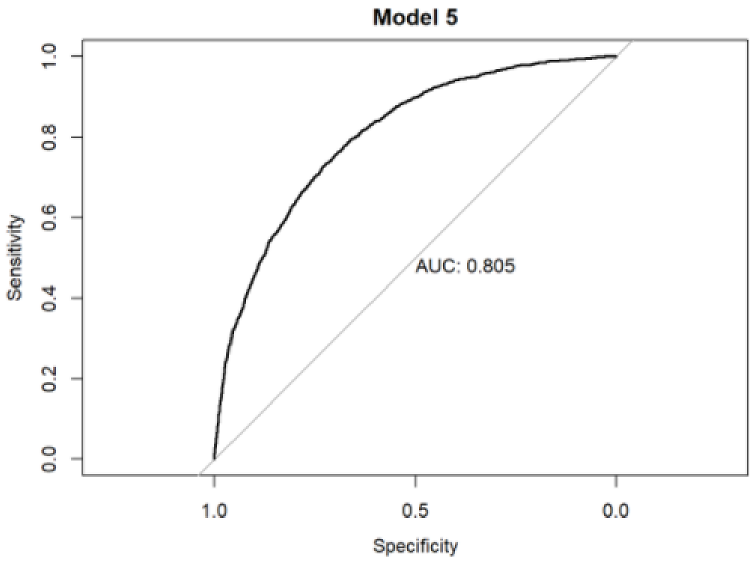


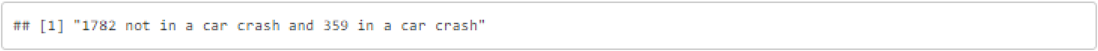
```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = in_bc_transformed1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3338  -0.7275  -0.4182   0.6726   3.1404
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.4920596  0.3784998   3.942 8.08e-05 ***
## KIDSDRIV     1.4363646  0.2440248   5.886 3.95e-09 ***
## AGE         -0.0016109  0.0040420  -0.399 0.690238
## HOMEKIDS     0.4508758  0.2128579   2.118 0.034158 *
## YOJ          0.0016286  0.0021987   0.741 0.458863
## INCOME      -0.0066091  0.0008548  -7.731 1.06e-14 ***
## PARENT1      0.2550575  0.1181083   2.160 0.030810 *
## HOME_VAL    -0.0146688  0.0042823  -3.425 0.000614 ***
## MSTATUS      0.5417280  0.0866447   6.252 4.04e-10 ***
## SEX         -0.0049875  0.0876631  -0.057 0.954630
## EDUCATION    0.0444316  0.0303084   1.466 0.142653
## JOB         -0.0044520  0.0112377  -0.396 0.691983
## TRAVTIME     0.0420860  0.0049847   8.443  < 2e-16 ***
## CAR_USE     -0.9179422  0.0681150 -13.476  < 2e-16 ***
## BLUEBOOK    -0.0046963  0.0007122  -6.594 4.28e-11 ***
## TIF         -0.1810247  0.0238017  -7.606 2.84e-14 ***
## CAR_TYPE     0.1989971  0.0278722   7.140 9.36e-13 ***
## RED_CAR     -0.0301921  0.0854327  -0.353 0.723787
## OLDCLAIM    -0.0197972  0.0317252  -0.624 0.532613
## CLM_FREQ     1.1364713  0.4234432   2.684 0.007277 **
## REVOKED      0.7470639  0.0810881   9.213  < 2e-16 ***
## MVR_PTS      0.4125753  0.0621834   6.635 3.25e-11 ***
## CAR_AGE     -0.0839813  0.0169297  -4.961 7.03e-07 ***
## URBANICITY  -2.2946669  0.1131215 -20.285  < 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: 9418.0  on 8160  degrees of freedom
## Residual deviance: 7412.7  on 8137  degrees of freedom
## AIC: 7460.7
##
## Number of Fisher Scoring iterations: 5
```

SELECT MODELS

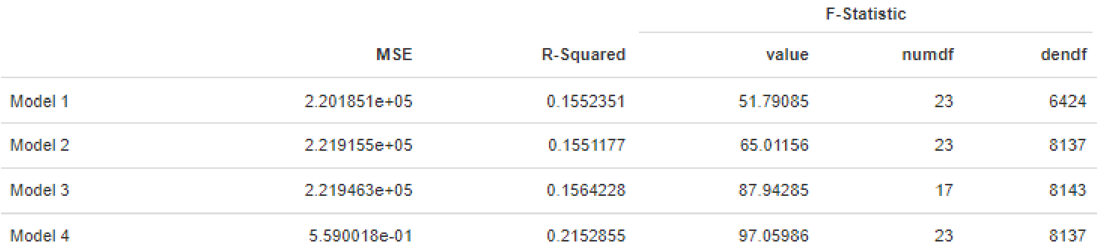
Multiple Linear Regression Metrics

From the residual plot below, the variance of residuals is not uniform. In addition, from the Q-Q plot, the residuals are not normally distributed. Hence, the model is not a good model.

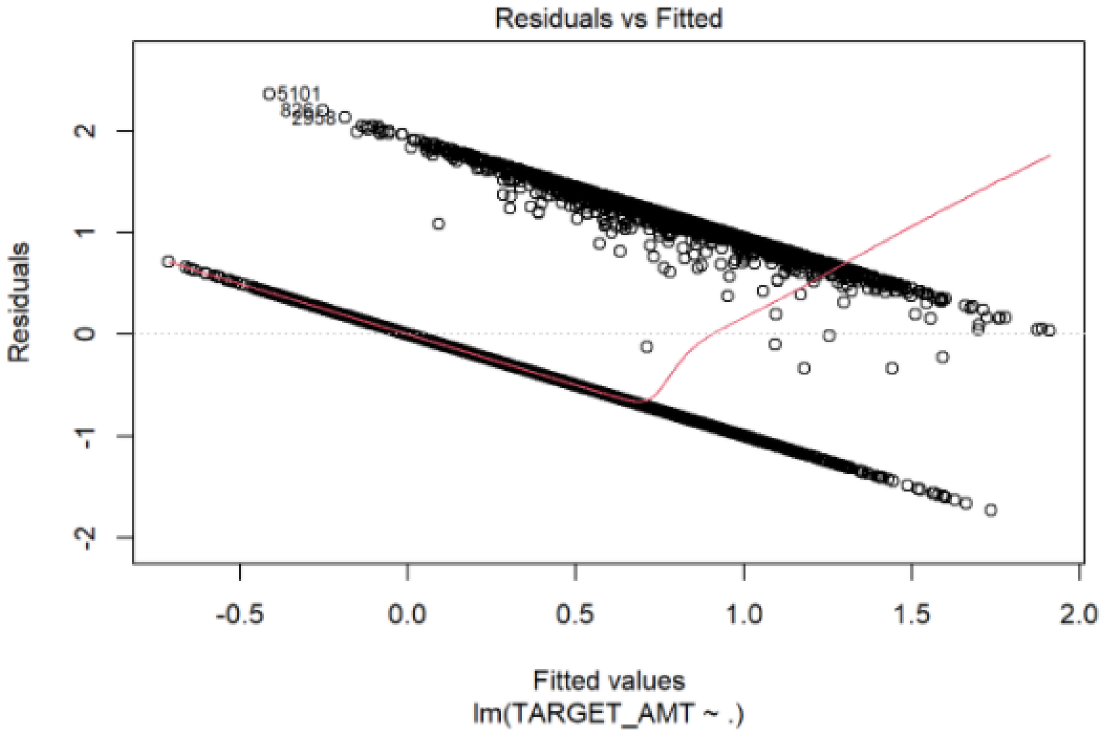


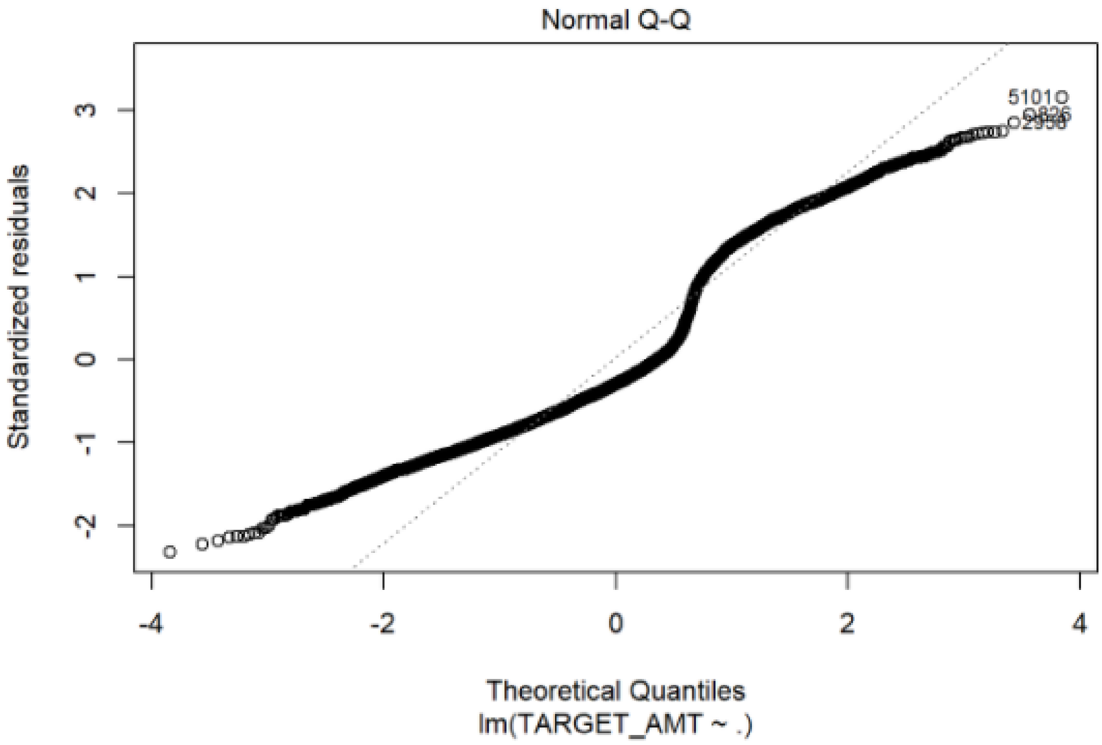


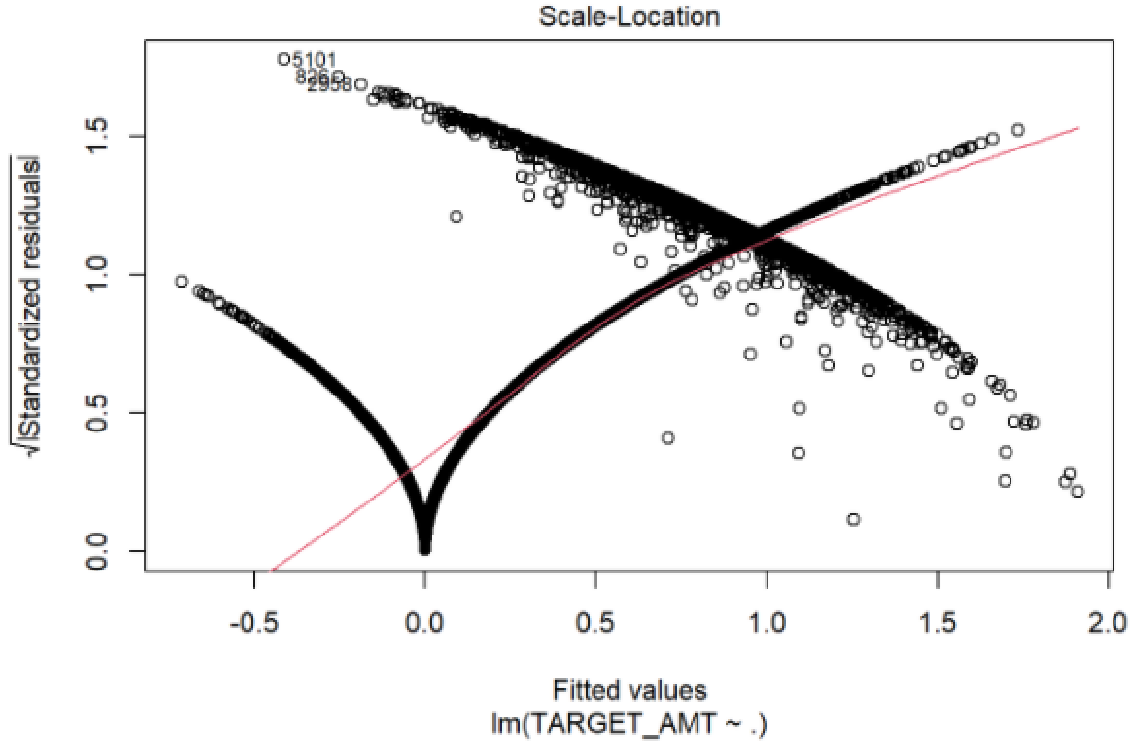
```
## [1] "1782 not in a car crash and 359 in a car crash"
```

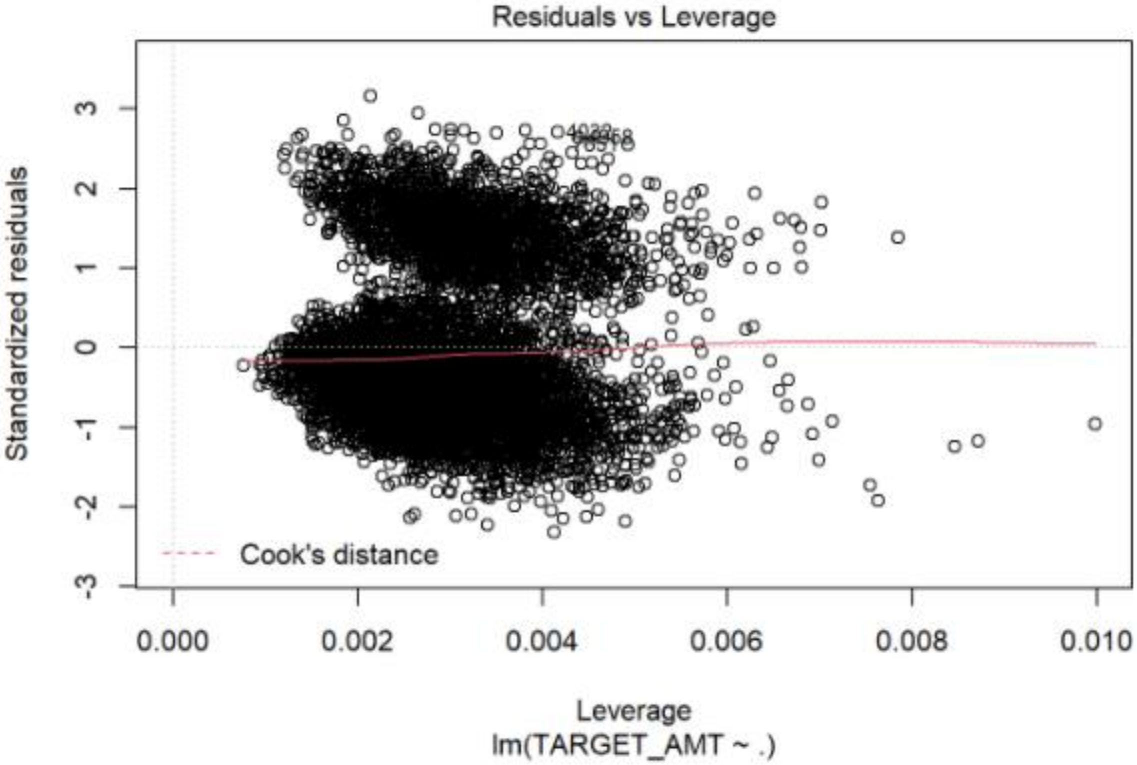


			F-Statistic		
	MSE	R-Squared	value	numdf	dendf
Model 1	2.201851e+05	0.1552351	51.79085	23	6424
Model 2	2.219155e+05	0.1551177	65.01156	23	8137
Model 3	2.219463e+05	0.1564228	87.94285	17	8143
Model 4	5.590018e-01	0.2152855	97.05986	23	8137


Residuals vs Fitted
Residuals
Fitted values
lm(TARGET_AMT ~ .)
5101
8265
2958
2
1
0
-1
-2
-0.5
0.0
0.5
1.0
1.5
2.0

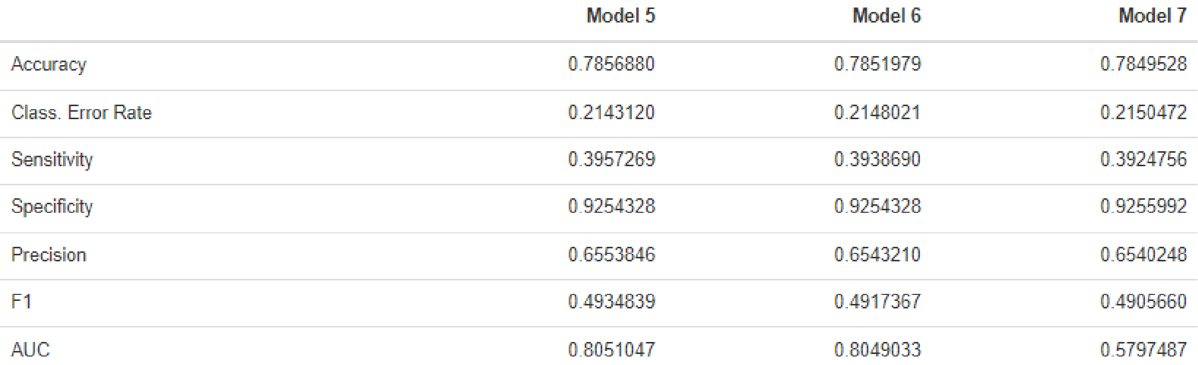

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(TARGET_AMT ~ .)
5101
8265
2958
3
2
1
0
-1
-2
-4
-2
0
2
4


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
5101
8260
2958
-0.5
0.0
0.5
1.0
1.5
2.0
Fitted values
lm(TARGET_AMT ~ .)


Residuals vs Leverage
Standardized residuals
3
2
1
0
-1
-2
-3
4023
Cook's distance
0.000
0.002
0.004
0.006
0.008
0.010
Leverage
lm(TARGET_AMT ~ .)

Binary Logistic Regression Metrics

Even though all models yield similar metrics value, model 5 has the highest AUC value. As a result, I select model 5 with imputed values for my prediction.



	Model 5	Model 6	Model 7
Accuracy	0.7856880	0.7851979	0.7849528
Class. Error Rate	0.2143120	0.2148021	0.2150472
Sensitivity	0.3957269	0.3938690	0.3924756
Specificity	0.9254328	0.9254328	0.9255992
Precision	0.6553846	0.6543210	0.6540248
F1	0.4934839	0.4917367	0.4905660
AUC	0.8051047	0.8049033	0.5797487

References

- A Modern Approach to Regression with R: Simon Sheather
- Linear Models with R: Julian Faraway.

Appendix

Description of the variables

Variable Name	Definition	Theoretical Effect
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children at Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high

Variable Name	**Definition**	**Theoretical Effect**
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

R Code

GitHub: https://github.com/nathtrish334/Data-621/blob/main/HW4/Homework04.rmd

RPubs: https://rpubs.com/trishitanath/838996