HW 4: Predictive Modeling for Car Accidents and Claim Amount

By: Christina Valore and Henry Vasquez

RMD: https://github.com/ChristinaValore/Business-Analytics-and-Data-Mining-621/blob/master/Homework4/Hw4.Rmd

Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, TARGET_FLAG, is a 1 or a 0. A “1” means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided).

Data Exploration

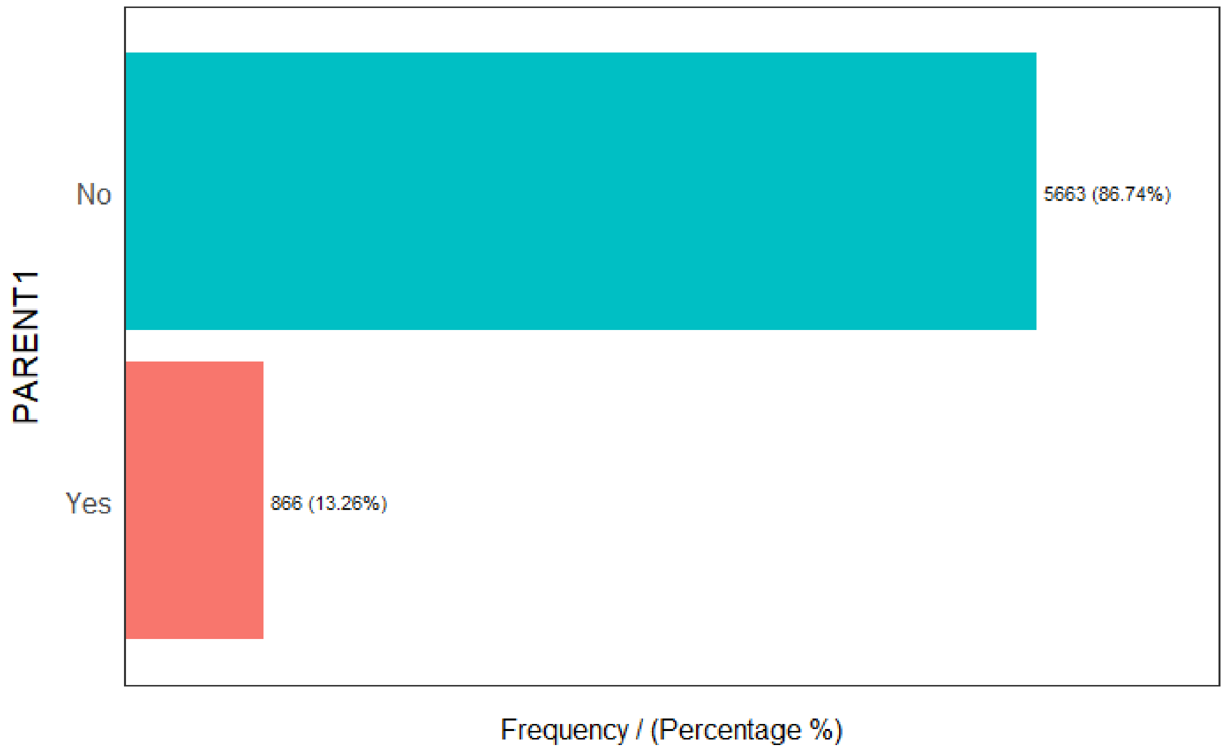
We start by importing the data into GitHub, removing the index and looking at the structure of the data to ensure all variables are the proper type. From the data, we’ll need to remove the dollar signs and commas from all values that have numbers. We do this as we want to convert those variables to numeric.

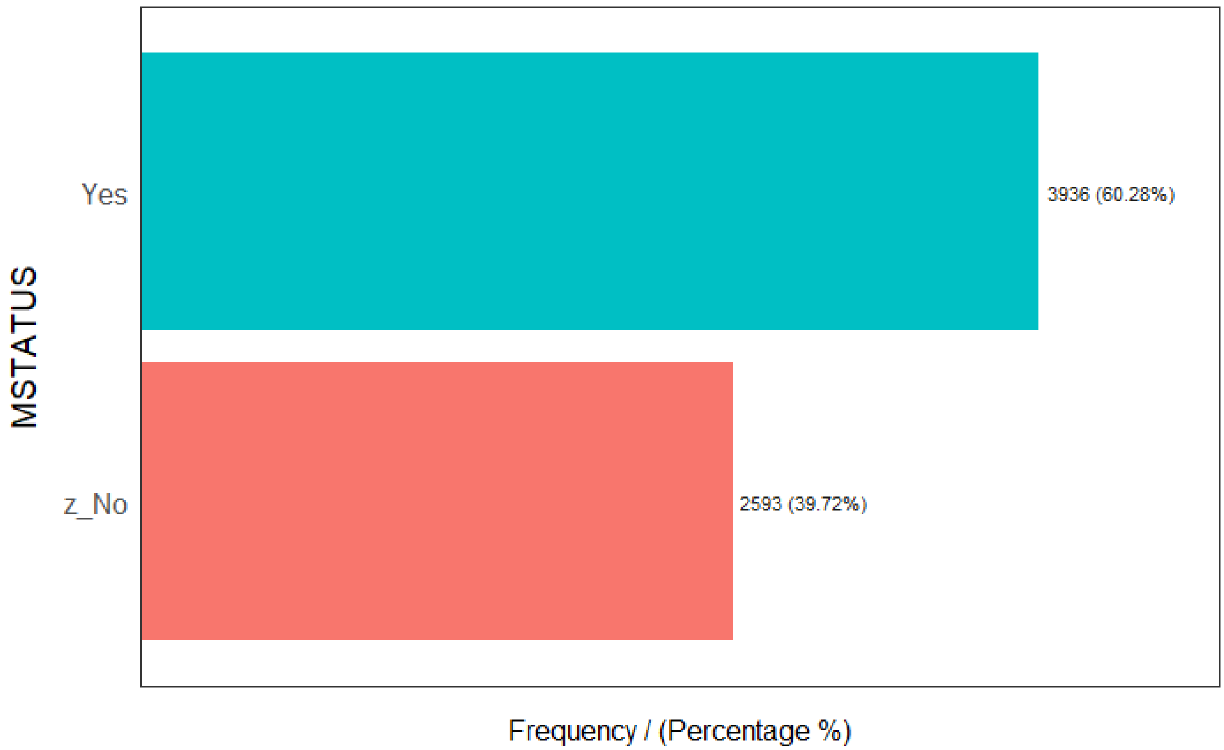
```
  TARGET_FLAG     TARGET_AMT       KIDSDRIV           AGE           HOMEKIDS            YOJ            INCOME       PARENT1      HOME_VAL     MSTATUS     SEX           EDUCATION              JOB
 Min.   :0.000   Min.   :    0   Min.   :0.0000   Min.   :16.00   Min.   :0.0000   Min.   : 0.00   Min.   :     0   No :5663   Min.   :     0   Yes :3936   M  :3033   <High School : 971   z_Blue Collar:1476
 1st Qu.:0.000   1st Qu.:    0   1st Qu.:0.0000   1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.00   1st Qu.: 27646   Yes: 866   1st Qu.:     0   z_No:2593   z_F:3496   Bachelors    :1798   Clerical     : 997
 Median :0.000   Median :    0   Median :0.0000   Median :45.00   Median :0.0000   Median :11.00   Median : 54005              Median :160945                          Masters      :1324   Professional : 901
 Mean   :0.265   Mean   : 1491   Mean   :0.1731   Mean   :44.85   Mean   :0.7265   Mean   :10.49   Mean   : 61552              Mean   :154188                          PhD          : 577   Manager      : 783
 3rd Qu.:1.000   3rd Qu.: 1102   3rd Qu.:0.0000   3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.00   3rd Qu.: 85697              3rd Qu.:238750                          z_High School:1859   Lawyer       : 665
 Max.   :1.000   Max.   :85524   Max.   :4.0000   Max.   :76.00   Max.   :5.0000   Max.   :19.00   Max.   :367030              Max.   :885282                                               Student      : 573
                                                  NA's   :6                        NA's   :370     NA's   :350                 NA's   :358                                                   (Other)      :1134
    TRAVTIME            CAR_USE         BLUEBOOK          TIF                 CAR_TYPE     RED_CAR       OLDCLAIM        CLM_FREQ       REVOKED       MVR_PTS          CAR_AGE                        URBANICITY
 Min.   :  5.00   Commercial:2440   Min.   : 1500   Min.   : 1.000   Minivan    :1706   no :4623   Min.   :    0   Min.   :0.0000   No :5742   Min.   : 0.000   Min.   : 0.000   Highly Urban/ Urban  :5169
 1st Qu.: 23.00   Private   :4089   1st Qu.: 9260   1st Qu.: 1.000   Panel Truck: 550   yes:1906   1st Qu.:    0   1st Qu.:0.0000   Yes: 787   1st Qu.: 0.000   1st Qu.: 1.000   z_Highly Rural/ Rural:1360
 Median : 33.00                     Median :14440   Median : 4.000   Pickup     :1083              Median :    0   Median :0.0000              Median : 1.000   Median : 8.000
 Mean   : 33.58                     Mean   :15684   Mean   : 5.357   Sports Car : 732              Mean   : 3982   Mean   :0.7961              Mean   : 1.695   Mean   : 8.255
 3rd Qu.: 44.00                     3rd Qu.:20800   3rd Qu.: 7.000   Van        : 612              3rd Qu.: 4633   3rd Qu.:2.0000              3rd Qu.: 3.000   3rd Qu.:12.000
 Max.   :142.00                     Max.   :65970   Max.   :25.000   z_SUV      :1846              Max.   :57037   Max.   :5.0000              Max.   :13.000   Max.   :28.000
                                                                                                                                                                NA's   :415
```

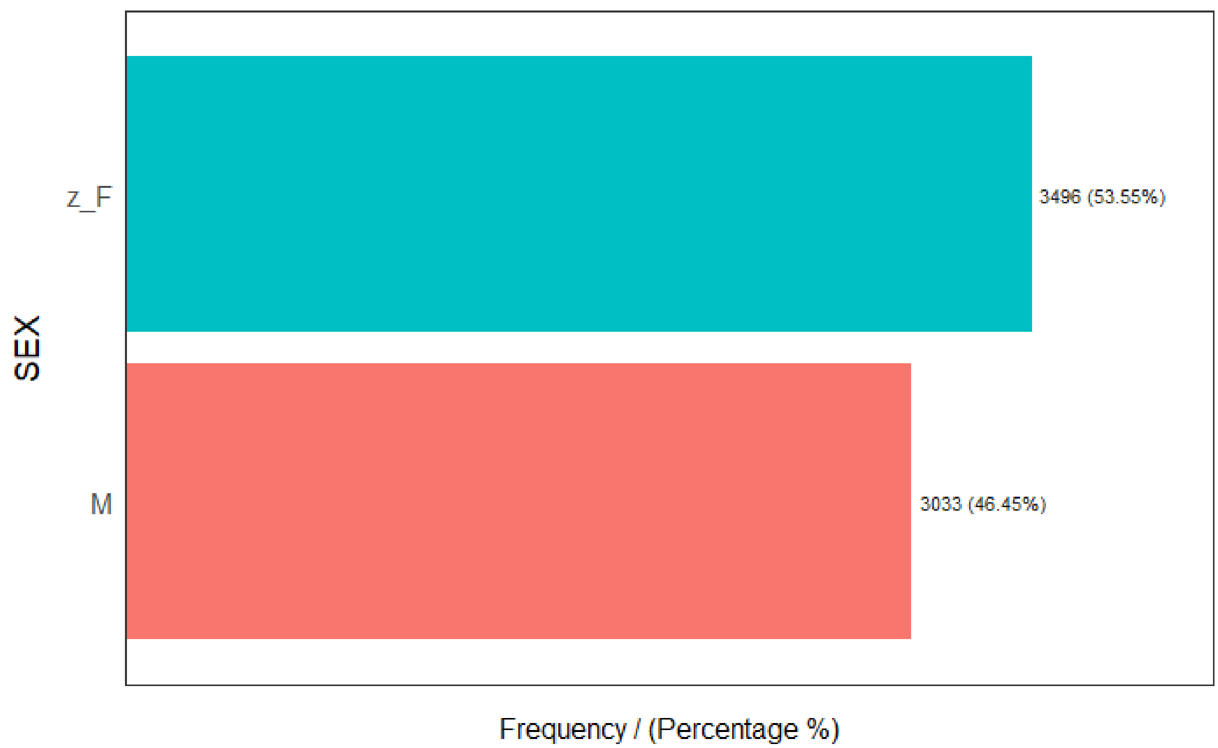
From our data summary, we notice that there is not a significant amount of NA’s in most variables. Similarly, there are not real issues with zeros present except for KIDSDRIV, HOMEKIDS, OLDCLAIM and CLM_FREQ. The target variables do have most zeros, but we will keep these while removing the rest of the variables with large percentages of zeros. Next we look at the frequency of variables that are factors or characters. Easily we can see variables with the highest factor levels such that we can say:

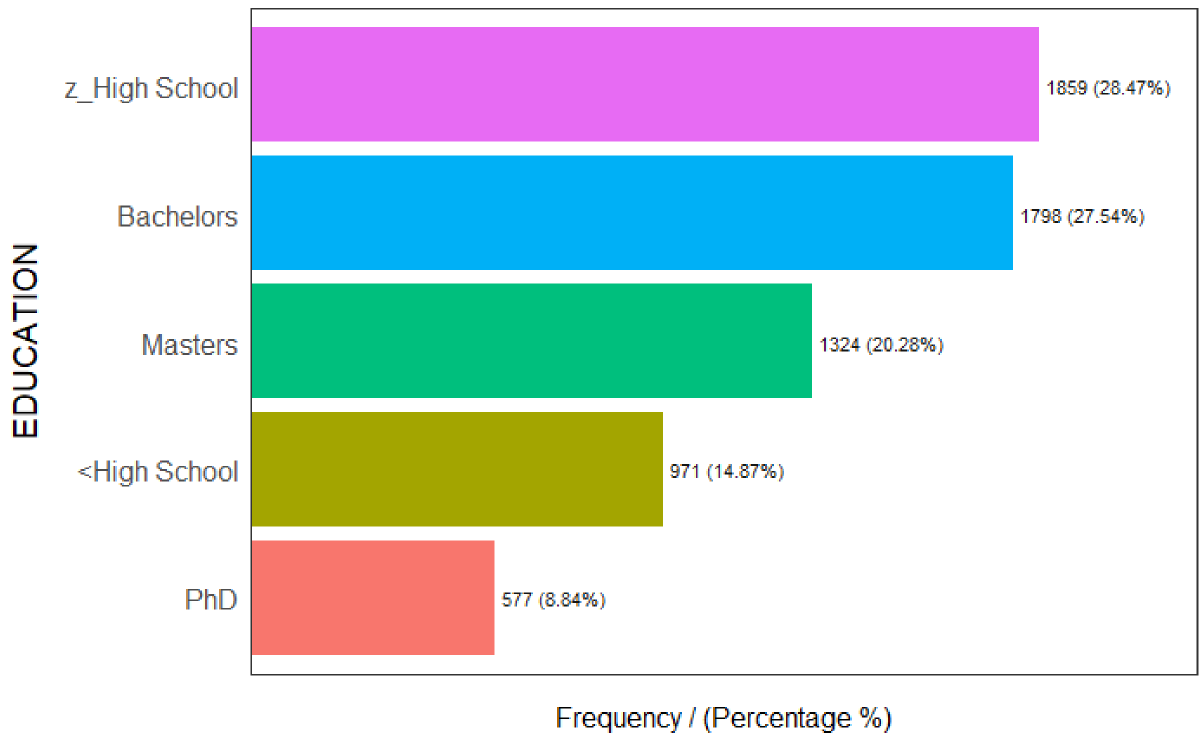
- most of the drivers are not single parents
- most of the drivers are married

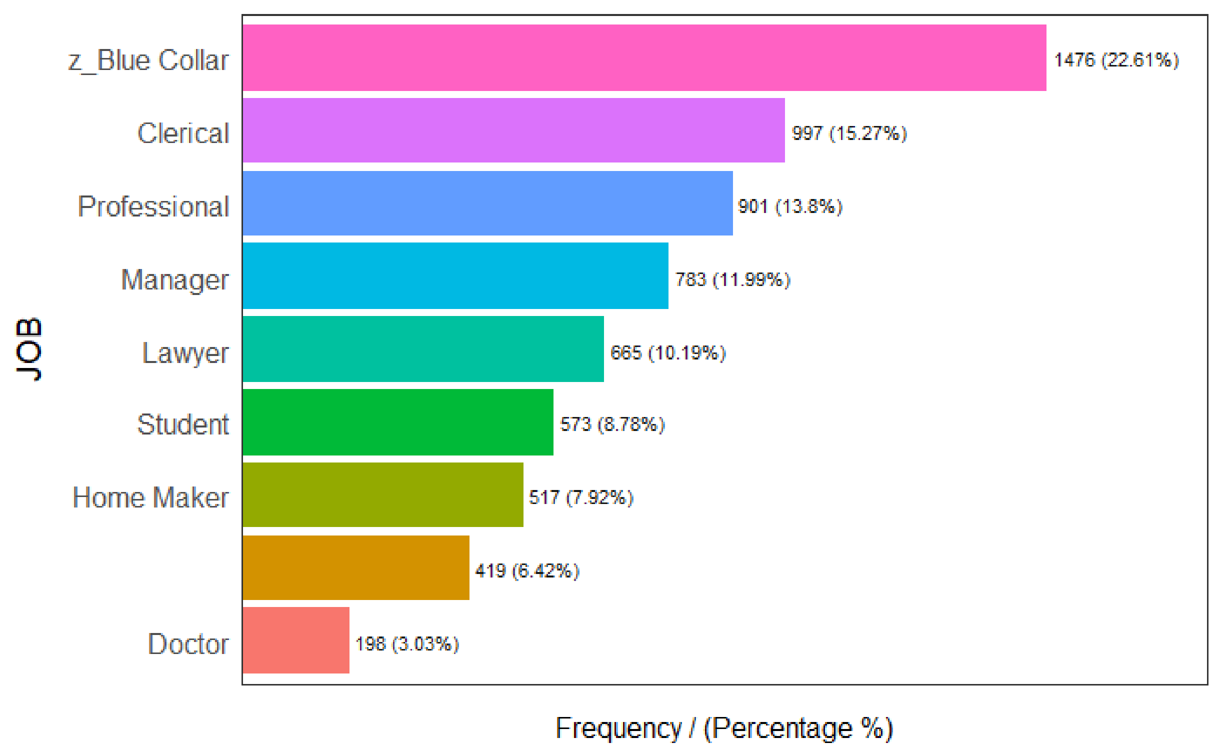
- most are female
- most have finished highschool at least
- most work blue collar jobs
- most use the car for leisure
- most of the cars are SVU's
- most are not red cars
- most did not have their license revoke in the past 7 years
- most live/work in urban area

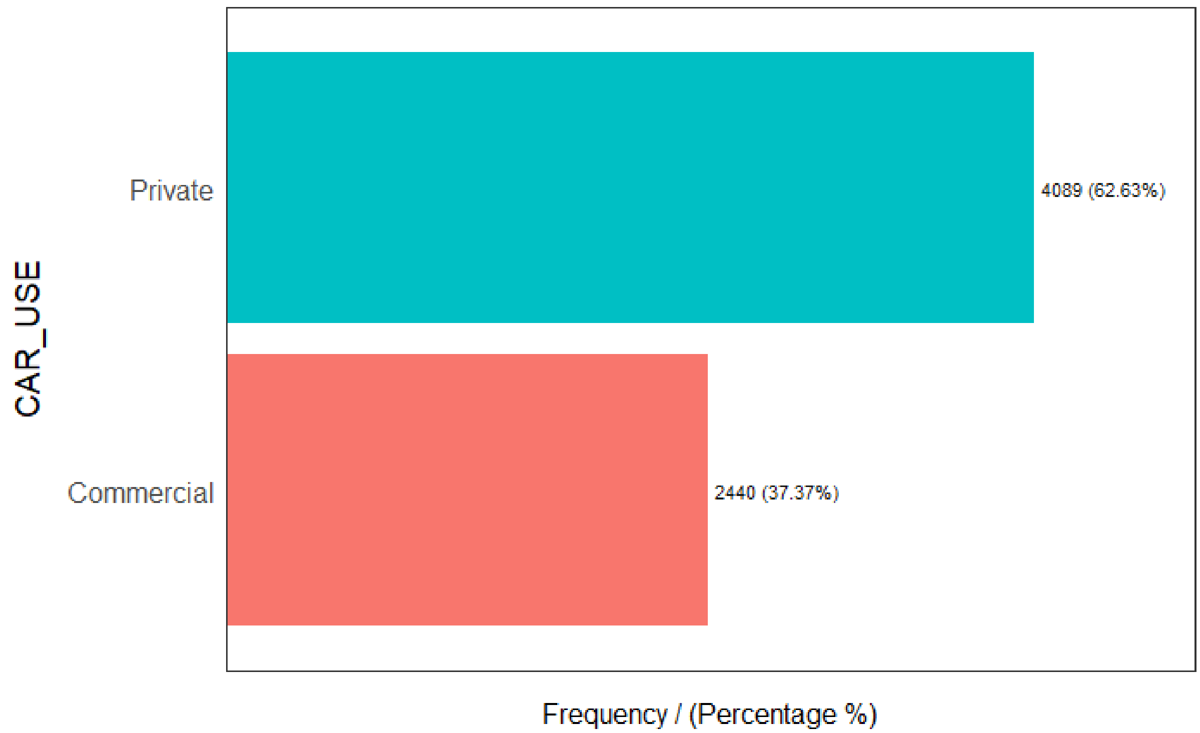


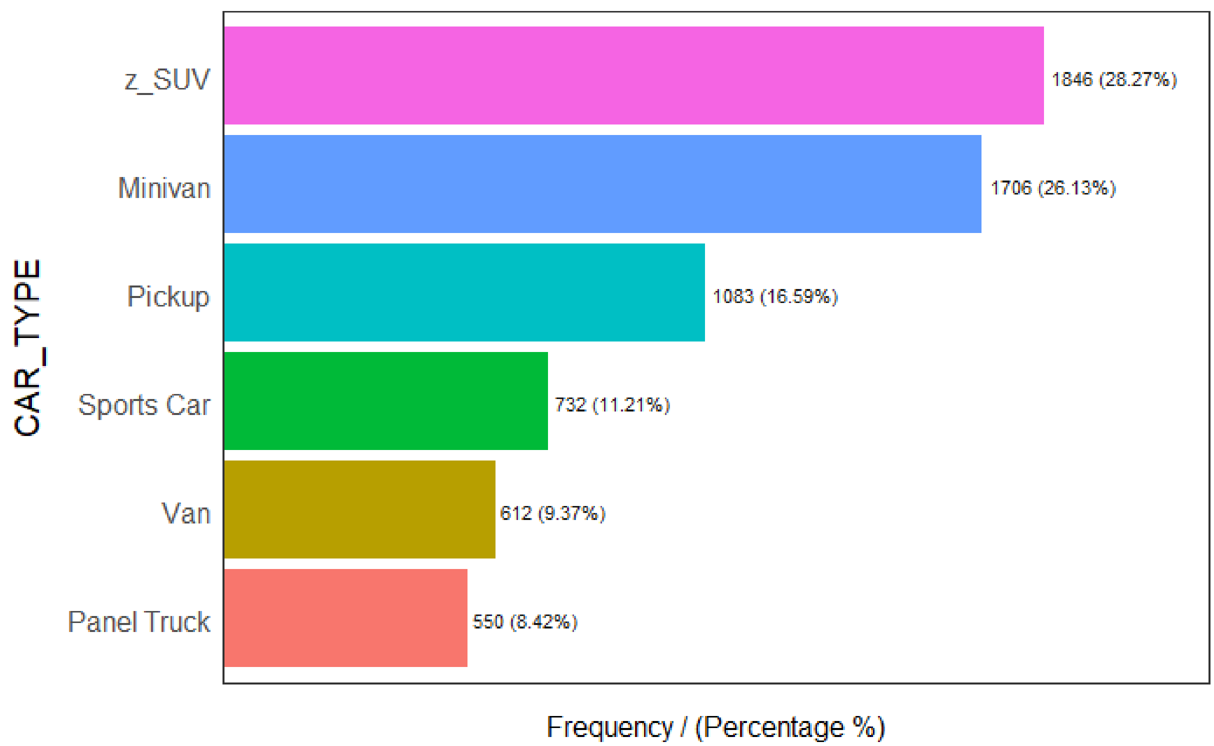

MSTATUS
Yes
z_No
3936 (60.28%)
2593 (39.72%)
Frequency / (Percentage %)

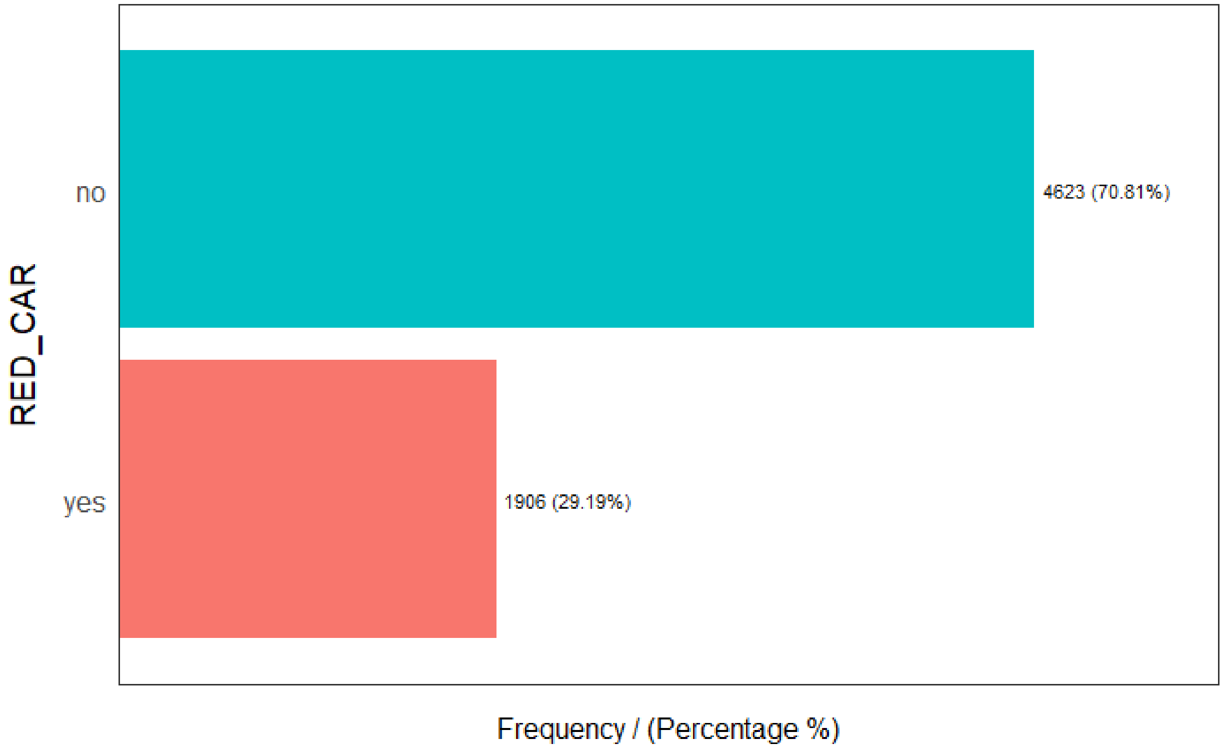

SEX
z_F
M
3496 (53.55%)
3033 (46.45%)
Frequency / (Percentage %)

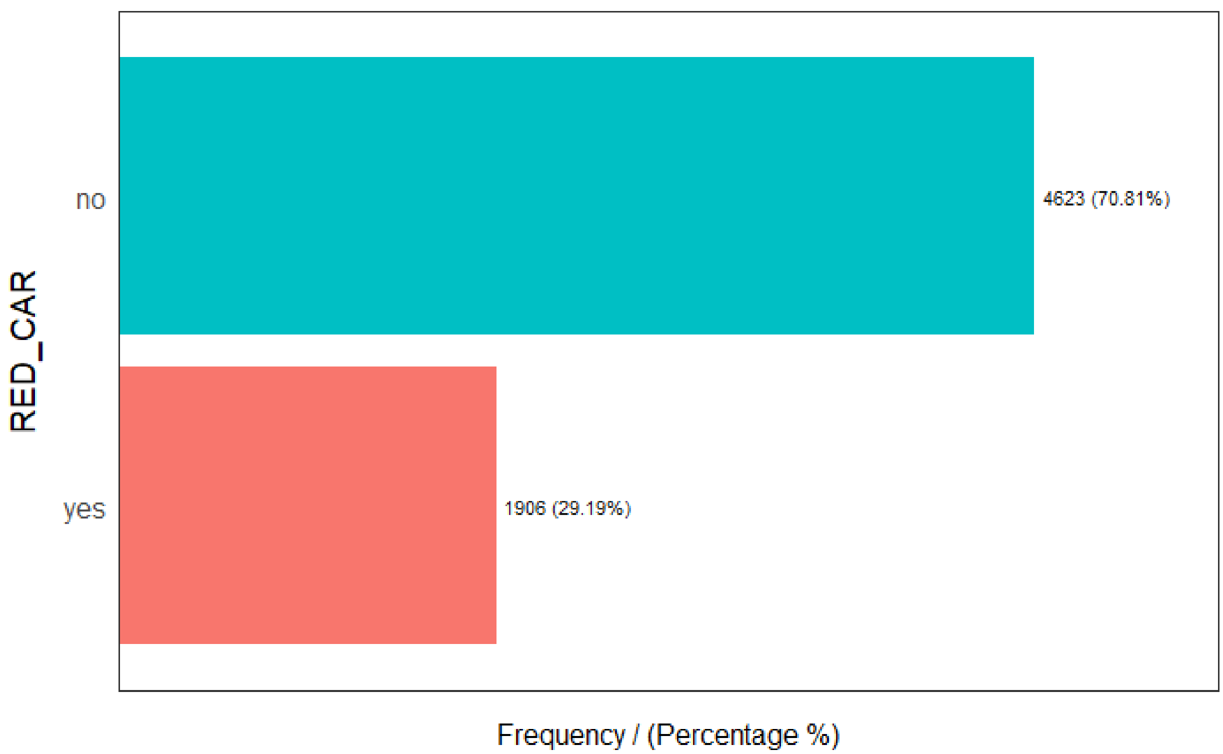

EDUCATION
z_High School
1859 (28.47%)
Bachelors
1798 (27.54%)
Masters
1324 (20.28%)
<High School
971 (14.87%)
PhD
577 (8.84%)
Frequency / (Percentage %)

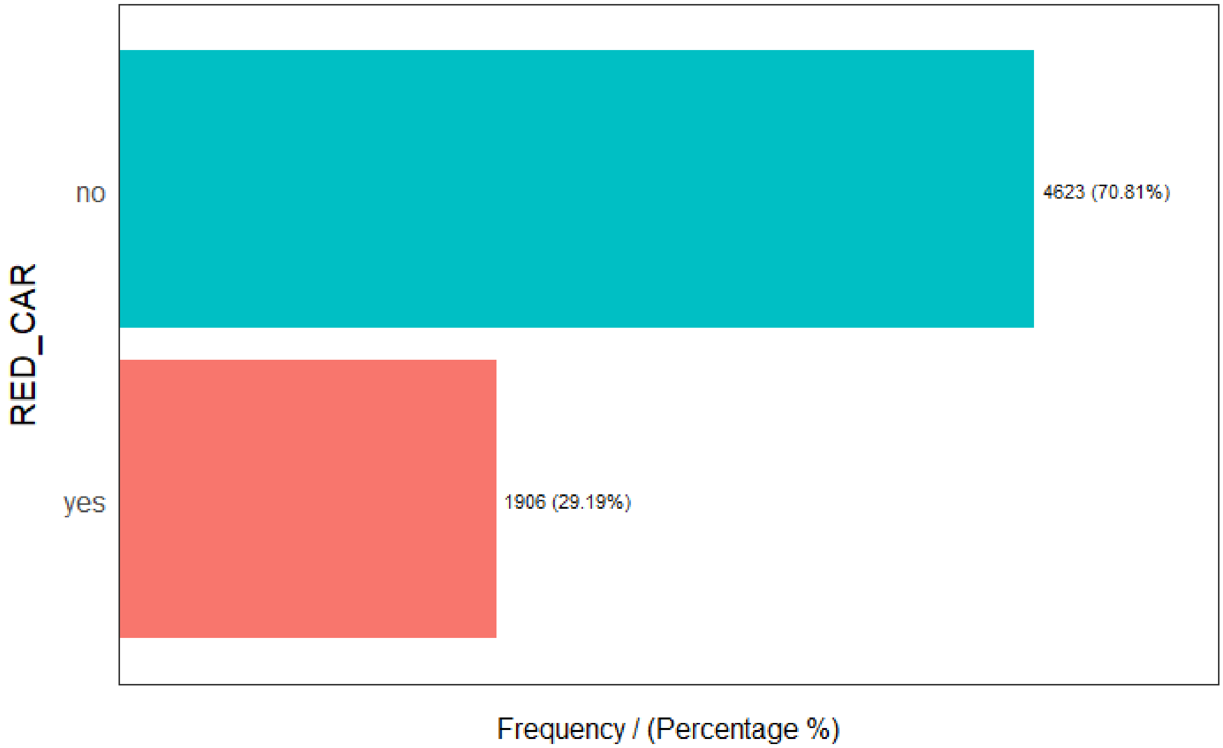

JOB
z_Blue Collar
1476 (22.61%)
Clerical
997 (15.27%)
Professional
901 (13.8%)
Manager
783 (11.99%)
Lawyer
665 (10.19%)
Student
573 (8.78%)
Home Maker
517 (7.92%)
419 (6.42%)
Doctor
198 (3.03%)
Frequency / (Percentage %)

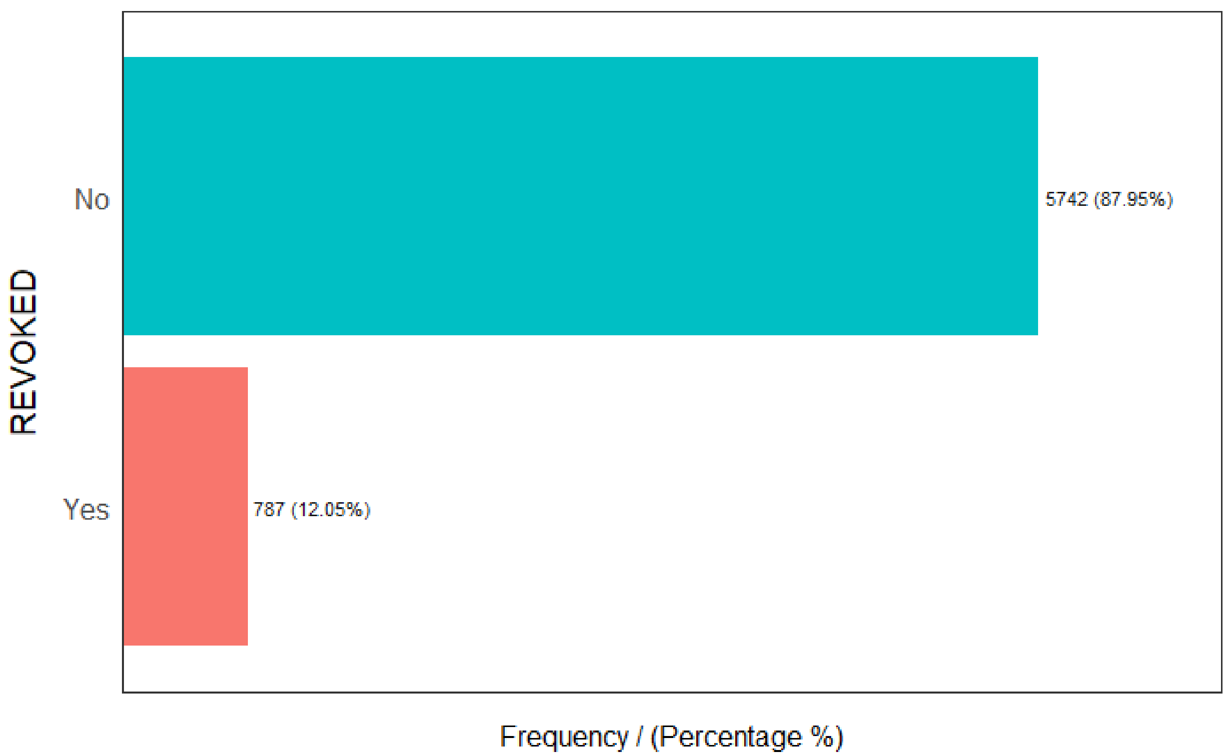

CAR_USE
Private
Commercial
4089 (62.63%)
2440 (37.37%)
Frequency / (Percentage %)

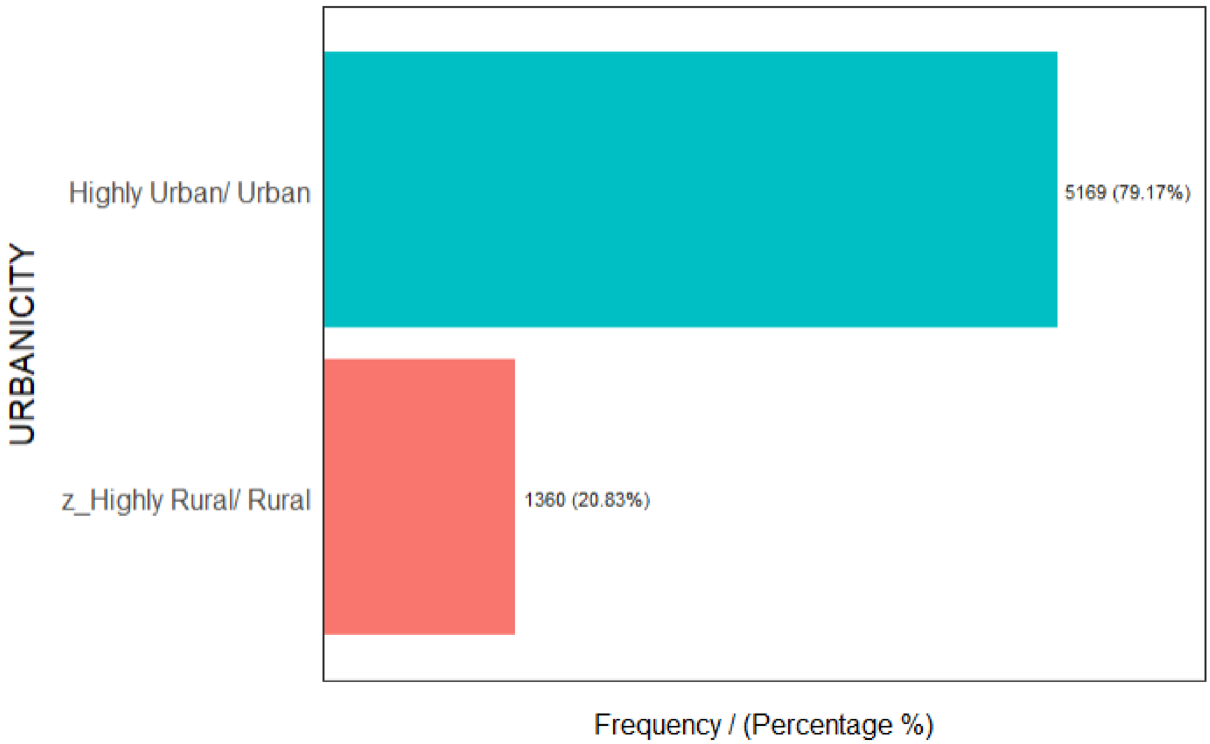

CAR_TYPE
z_SUV
Minivan
Pickup
Sports Car
Van
Panel Truck
1846 (28.27%)
1706 (26.13%)
1083 (16.59%)
732 (11.21%)
612 (9.37%)
550 (8.42%)
Frequency / (Percentage %)


RED_CAR
no
yes
4623 (70.81%)
1906 (29.19%)
Frequency / (Percentage %)

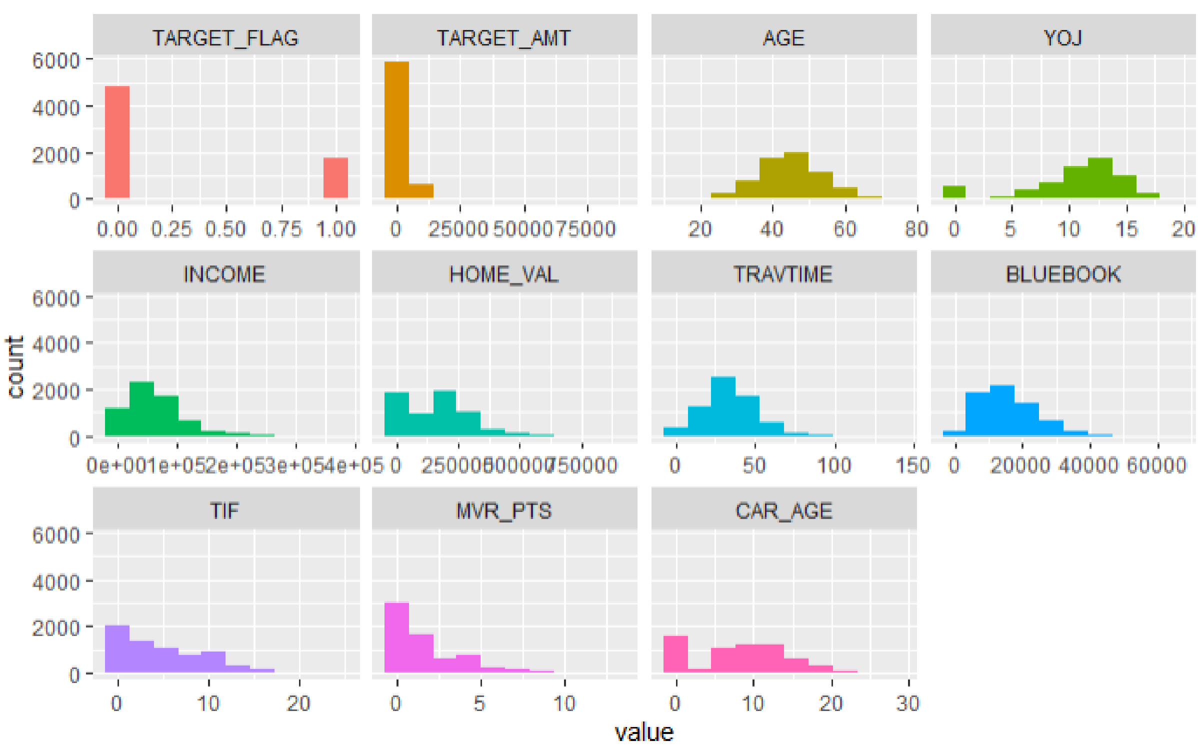

RED_CAR
no
yes
4623 (70.81%)
1906 (29.19%)
Frequency / (Percentage %)


RED_CAR
no
yes
4623 (70.81%)
1906 (29.19%)
Frequency / (Percentage %)


REVOKED
No
Yes
5742 (87.95%)
787 (12.05%)
Frequency / (Percentage %)



Looking at the distributions of the remaining variables, we can see that income,YOJ, TIV, MVR_PTS are all skewed right.



We can also see variables with skewness and high kurtosis (indicating outliers) here. As seen before visually, we can verify here that YOJ and income are highly skewed and have high kurtosis. Also, bluebook, tif and mvr_pts are also similar.

Data Preparation

Most of the data preparation was already done above, which included transforming variables that contained special characters like dollar signs or dropping variables with too many zeros. The remaining preparation includes imputing missing NA values with the median using the Hmisc package. We'll apply this to age, yoj, income and car age. One other thing we'll do is introduce a new variable, PTS_AGE. This variable is equal to MVR_PTS/AGE which says that if the ratio is higher then one is a driver with more points.

Build Models

Predicting Car Crash

Model 1

All predictors and their corresponding coefficients fall in line with their theoretical effect, except for sex. The theoretical effect suggest females are more at risk, but the model has a negative coefficient suggesting the opposite. However, sex is not statistically significant therefore we will not continue with the variable going forward. The variable YOJ whose coefficient is in line with the theoretical effect turned out to be statistically insignificant as well. Single parents were suggested more likely to be involved in an accident according to the model while Urban City Rural suggests less of a risk. The red car theory also suggests less risk but is insignificant based on its p-value. We'll go ahead and remove contradicting and insignificant variables in model 2. Our created variable PTS_AGE also tends to be significant with a corresponding coefficient as well.

```
Call:
glm(formula = TARGET_FLAG ~ YOJ + INCOME + PARENT1 + HOME_VAL +
    MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + TIF +
    CAR_TYPE + RED_CAR + REVOKED + URBANICITY + PTS_AGE, family = "binomial",
    data = train2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1603  -0.7234  -0.4181   0.6649   3.0602

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.154e+00  3.097e-01  -3.727 0.000194 ***
YOJ                             -6.191e-03  9.490e-03  -0.652 0.514169
INCOME                          -2.730e-06  1.238e-06  -2.204 0.027514 *
PARENT1Yes                       5.639e-01  1.047e-01   5.383 7.32e-08 ***
HOME_VAL                        -1.351e-06  3.913e-07  -3.454 0.000553 ***
MSTATUSz_No                      3.830e-01  9.266e-02   4.133 3.58e-05 ***
SEXz_F                          -2.449e-01  1.175e-01  -2.085 0.037062 *
EDUCATIONBachelors              -3.601e-01  1.244e-01  -2.896 0.003784 **
EDUCATIONMasters                -3.924e-01  1.868e-01  -2.101 0.035649 *
EDUCATIONPhD                    -1.700e-01  2.270e-01  -0.749 0.453831
EDUCATIONz_High School           7.008e-02  1.083e-01   0.647 0.517416
JOBClerical                      4.164e-01  2.240e-01   1.859 0.063050 .
JOBDoctor                       -6.475e-01  3.043e-01  -2.128 0.033362 *
JOBHome Maker                    2.450e-01  2.379e-01   1.030 0.303225
JOBLawyer                        9.244e-02  1.911e-01   0.484 0.628575
JOBManager                      -6.692e-01  1.978e-01  -3.383 0.000717 ***
JOBProfessional                  8.490e-02  2.034e-01   0.417 0.676417
JOBStudent                       3.574e-01  2.444e-01   1.462 0.143642
JOBz_Blue Collar                 2.867e-01  2.122e-01   1.351 0.176615
TRAVTIME                         1.593e-02  2.122e-03   7.509 5.94e-14 ***
CAR_USEPrivate                  -6.998e-01  1.050e-01  -6.665 2.64e-11 ***
TIF                             -5.058e-02  8.294e-03  -6.099 1.07e-09 ***
CAR_TYPEPanel Truck              3.056e-01  1.613e-01   1.895 0.058144 .
CAR_TYPEPickup                   5.584e-01  1.151e-01   4.853 1.22e-06 ***
CAR_TYPESports Car               1.199e+00  1.374e-01   8.724  < 2e-16 ***
CAR_TYPEVan                      4.925e-01  1.393e-01   3.536 0.000407 ***
CAR_TYPEz_SUV                    9.610e-01  1.162e-01   8.272  < 2e-16 ***
RED_CARyes                      -5.146e-02  9.856e-02  -0.522 0.601606
REVOKEDYes                       7.648e-01  9.198e-02   8.315  < 2e-16 ***
URBANICITYz_Highly Rural/ Rural -2.436e+00  1.255e-01 -19.415  < 2e-16 ***
PTS_AGE                          5.356e+00  5.792e-01   9.247  < 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: 7129.6  on 6170  degrees of freedom
Residual deviance: 5609.8  on 6140  degrees of freedom
  (358 observations deleted due to missingness)
AIC: 5671.8

Number of Fisher Scoring iterations: 5
```

Model 2

In this model, all coefficients fall in line with their theoretical effects. Only concern would be that most job categories are not statistically significant. For the next model, well go ahead and remove these.

```
Call:
glm(formula = TARGET_FLAG ~ INCOME + PARENT1 + HOME_VAL + MSTATUS +
    EDUCATION + JOB + TRAVTIME + CAR_USE + TIF + CAR_TYPE + REVOKED +
    URBANICITY + PTS_AGE, family = "binomial", data = train2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1772  -0.7240  -0.4179   0.6595   3.0730

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -1.320e+00  2.822e-01  -4.678 2.90e-06 ***
INCOME                         -3.088e-06  1.227e-06  -2.516 0.011870 *
PARENT1Yes                      5.466e-01  1.042e-01   5.244 1.57e-07 ***
HOME_VAL                       -1.326e-06  3.907e-07  -3.394 0.000690 ***
MSTATUSz_No                     3.960e-01  9.169e-02   4.319 1.57e-05 ***
EDUCATIONBachelors             -3.576e-01  1.243e-01  -2.877 0.004009 **
EDUCATIONMasters               -3.857e-01  1.864e-01  -2.069 0.038557 *
EDUCATIONPhD                   -1.727e-01  2.266e-01  -0.762 0.445899
EDUCATIONz_High School          6.977e-02  1.082e-01   0.645 0.519032
JOBClerical                     4.136e-01  2.239e-01   1.847 0.064725 .
JOBDoctor                      -6.248e-01  3.040e-01  -2.055 0.039876 *
JOBHome Maker                   2.384e-01  2.316e-01   1.030 0.303221
JOBLawyer                       9.764e-02  1.911e-01   0.511 0.609456
JOBManager                     -6.671e-01  1.978e-01  -3.373 0.000745 ***
JOBProfessional                 8.711e-02  2.033e-01   0.428 0.668306
JOBStudent                      3.876e-01  2.404e-01   1.612 0.106889
JOBz_Blue Collar                2.947e-01  2.120e-01   1.390 0.164465
TRAVTIME                        1.594e-02  2.119e-03   7.523 5.33e-14 ***
CAR_USEPrivate                 -6.950e-01  1.048e-01  -6.631 3.33e-11 ***
TIF                            -5.045e-02  8.291e-03  -6.084 1.17e-09 ***
CAR_TYPEPanel Truck             3.760e-01  1.580e-01   2.379 0.017347 *
CAR_TYPEPickup                  5.690e-01  1.148e-01   4.957 7.17e-07 ***
CAR_TYPESports Car              1.066e+00  1.206e-01   8.841  < 2e-16 ***
CAR_TYPEVan                     5.440e-01  1.372e-01   3.964 7.36e-05 ***
CAR_TYPEz_SUV                   8.279e-01  9.618e-02   8.608  < 2e-16 ***
REVOKEDYes                      7.673e-01  9.190e-02   8.349  < 2e-16 ***
URBANICITYz_Highly Rural/ Rural -2.436e+00  1.255e-01 -19.418  < 2e-16 ***
PTS_AGE                         5.345e+00  5.781e-01   9.245  < 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: 7129.6  on 6170  degrees of freedom
Residual deviance: 5614.9  on 6143  degrees of freedom
  (358 observations deleted due to missingness)
AIC: 5670.9

Number of Fisher Scoring iterations: 5
```

Model 3

The model has most of the variables with significant p-values except for 2 categories of education (high school) and car type (truck). All coefficients of the variables also fall in line with theoretical effects.

```
Call:
glm(formula = TARGET_FLAG ~ INCOME + PARENT1 + HOME_VAL + MSTATUS +
    EDUCATION + TRAVTIME + CAR_USE + TIF + CAR_TYPE + REVOKED +
    URBANICITY + PTS_AGE, family = "binomial", data = train2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1696  -0.7337  -0.4349   0.6606   3.0671

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -8.296e-01  1.684e-01  -4.928 8.31e-07 ***
INCOME                      -4.457e-06  1.120e-06  -3.981 6.87e-05 ***
PARENT1Yes                   5.555e-01  1.031e-01   5.385 7.22e-08 ***
HOME_VAL                    -1.425e-06  3.774e-07  -3.775 0.000160 ***
MSTATUSz_No                  3.718e-01  9.037e-02   4.115 3.88e-05 ***
EDUCATIONBachelors          -5.966e-01  1.115e-01  -5.352 8.68e-08 ***
EDUCATIONMasters            -6.731e-01  1.251e-01  -5.380 7.44e-08 ***
EDUCATIONPhD                -6.456e-01  1.665e-01  -3.877 0.000106 ***
EDUCATIONz_High School      -4.559e-02  1.044e-01  -0.437 0.662453
TRAVTIME                     1.646e-02  2.102e-03   7.827 4.99e-15 ***
CAR_USEPrivate              -8.303e-01  8.391e-02  -9.895  < 2e-16 ***
TIF                         -4.973e-02  8.240e-03  -6.035 1.59e-09 ***
CAR_TYPEPanel Truck          2.685e-01  1.481e-01   1.813 0.069811 .
CAR_TYPEPickup               5.028e-01  1.118e-01   4.496 6.93e-06 ***
CAR_TYPESports Car           1.044e+00  1.186e-01   8.808  < 2e-16 ***
CAR_TYPEVan                  4.819e-01  1.342e-01   3.590 0.000330 ***
CAR_TYPEz_SUV                8.294e-01  9.490e-02   8.739  < 2e-16 ***
REVOKEDYes                   7.795e-01  9.108e-02   8.559  < 2e-16 ***
URBANICITYz_Highly Rural/ Rural -2.360e+00  1.250e-01 -18.875  < 2e-16 ***
PTS_AGE                      5.541e+00  5.745e-01   9.645  < 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: 7129.6  on 6170  degrees of freedom
Residual deviance: 5677.2  on 6151  degrees of freedom
  (358 observations deleted due to missingness)
AIC: 5717.2

Number of Fisher Scoring iterations: 5
```

Predicting Claim Amount

Model 1

A lot of the variables are insignificant, which makes sense. Most of these variables' theoretical effects have to do with their probabilities influencing accidents and not claim amount. Now since we're looking at claim amount the significant variables make sense with minor exceptions. Marital status no, suggests higher payments claim which is not what would originally be expected. The positive coefficient of bluebook makes sense since the company measures value for vehicles and higher bluebook value suggests higher payout. Car age is also in line with theoretical effect meaning older cars depreciate in cost (in most cases). For the next model, we'll remove the insignificant predictors except for car type since it should have an effect on amount (usually).

```
Call:
lm(formula = TARGET_AMT ~ . - TARGET_FLAG, data = train2_claims)

Residuals:
   Min     1Q Median     3Q    Max 
 -8473  -3015  -1393    568  76295 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      3.085e+03  1.773e+03   1.741 0.081949 .  
YOJ                              4.300e+01  5.164e+01   0.833 0.405148    
INCOME                          -4.142e-03  7.301e-03  -0.567 0.570612    
PARENT1Yes                      -3.944e+02  5.176e+02  -0.762 0.446170    
HOME_VAL                         1.232e-03  2.192e-03   0.562 0.574090    
MSTATUSz_No                      1.161e+03  5.091e+02   2.281 0.022660 *  
SEXz_F                          -1.011e+03  7.043e+02  -1.436 0.151154    
EDUCATIONBachelors               8.035e+01  6.935e+02   0.116 0.907772    
EDUCATIONMasters                 1.442e+03  1.182e+03   1.220 0.222527    
EDUCATIONPhD                     1.492e+03  1.393e+03   1.071 0.284439    
EDUCATIONz_High School          -7.167e+02  5.571e+02  -1.287 0.198413    
JOBClerical                      6.019e+02  1.300e+03   0.463 0.643432    
JOBDoctor                       -1.132e+03  1.927e+03  -0.587 0.557010    
JOBHome Maker                    1.299e+03  1.359e+03   0.956 0.339060    
JOBLawyer                        9.975e+02  1.103e+03   0.904 0.366077    
JOBManager                      -1.581e+02  1.193e+03  -0.133 0.894599    
JOBProfessional                  2.152e+03  1.219e+03   1.766 0.077621 .  
JOBStudent                       1.523e+03  1.385e+03   1.099 0.271811    
JOBz_Blue Collar                 1.619e+03  1.241e+03   1.304 0.192348    
TRAVTIME                        -2.845e+00  1.181e+01  -0.241 0.809624    
CAR_USEPrivate                  -1.720e+02  5.619e+02  -0.306 0.759581    
BLUEBOOK                         1.186e-01  3.280e-02   3.617 0.000308 ***
TIF                              3.672e+00  4.486e+01   0.082 0.934772    
CAR_TYPEPanel Truck             -5.591e+02  1.028e+03  -0.544 0.586808    
CAR_TYPEPickup                   1.181e+01  6.455e+02   0.018 0.985405    
CAR_TYPESports Car               1.345e+03  7.953e+02   1.691 0.091001 .  
CAR_TYPEVan                     -4.801e+02  8.319e+02  -0.577 0.563937    
CAR_TYPEz_SUV                    8.016e+02  7.101e+02   1.129 0.259130    
RED_CARyes                      -1.670e+01  5.347e+02  -0.031 0.975087    
REVOKEDYes                      -9.291e+02  4.458e+02  -2.084 0.037277 *  
CAR_AGE                         -1.147e+02  4.753e+01  -2.414 0.015877 *  
URBANICITYz_Highly Rural/ Rural -5.489e+02  8.108e+02  -0.677 0.498498    
PTS_AGE                          2.351e+03  2.599e+03   0.904 0.365915    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7210 on 1599 degrees of freedom
  (98 observations deleted due to missingness)
Multiple R-squared:  0.03284,	Adjusted R-squared:  0.01349 
F-statistic: 1.697 on 32 and 1599 DF,  p-value: 0.009073
```

Model 2

The predictors' coefficients all align with theoretical values. The only issue would be car type not having a significant p-value. We'll go ahead and remove this in the final model and keep car age along with bluebook value and marital status.

```
Call:
lm(formula = TARGET_AMT ~ MSTATUS + BLUEBOOK + CAR_AGE + CAR_TYPE, 
    data = train2_claims)

Residuals:
   Min     1Q Median     3Q    Max 
 -7741  -3012  -1481    361  77866 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         4237.37167  620.99693   6.823 1.23e-11 ***
MSTATUSz_No          749.91715  348.11398   2.154 0.031361 *  
BLUEBOOK               0.09922    0.02705   3.668 0.000252 ***
CAR_AGE              -61.30905   33.13971  -1.850 0.064482 .  
CAR_TYPEPanel Truck -115.57619  841.86723  -0.137 0.890821    
CAR_TYPEPickup        45.10293  581.95360   0.078 0.938233    
CAR_TYPESports Car   466.60774  634.46715   0.735 0.462176    
CAR_TYPEVan          -40.14408  728.66929  -0.055 0.956071    
CAR_TYPEz_SUV        -60.33257  532.58654  -0.113 0.909820    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7208 on 1721 degrees of freedom
Multiple R-squared:  0.01529,	Adjusted R-squared:  0.01071 
F-statistic:  3.34 on 8 and 1721 DF,  p-value: 0.0008428
```

Model 3

In this linear model, the coefficients are in line with theoretical effects. There is no need to remove any variables.

```
Call:
lm(formula = TARGET_AMT ~ MSTATUS + BLUEBOOK + CAR_AGE, data = train2_claims)

Residuals:
   Min     1Q Median     3Q    Max 
 -7721  -3027  -1490    351  78332 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4339.86307  423.06857  10.258  < 2e-16 ***
MSTATUSz_No  754.61699  347.16539   2.174   0.0299 *  
BLUEBOOK       0.09451    0.02106   4.487 7.68e-06 ***
CAR_AGE      -60.72690   33.03295  -1.838   0.0662 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

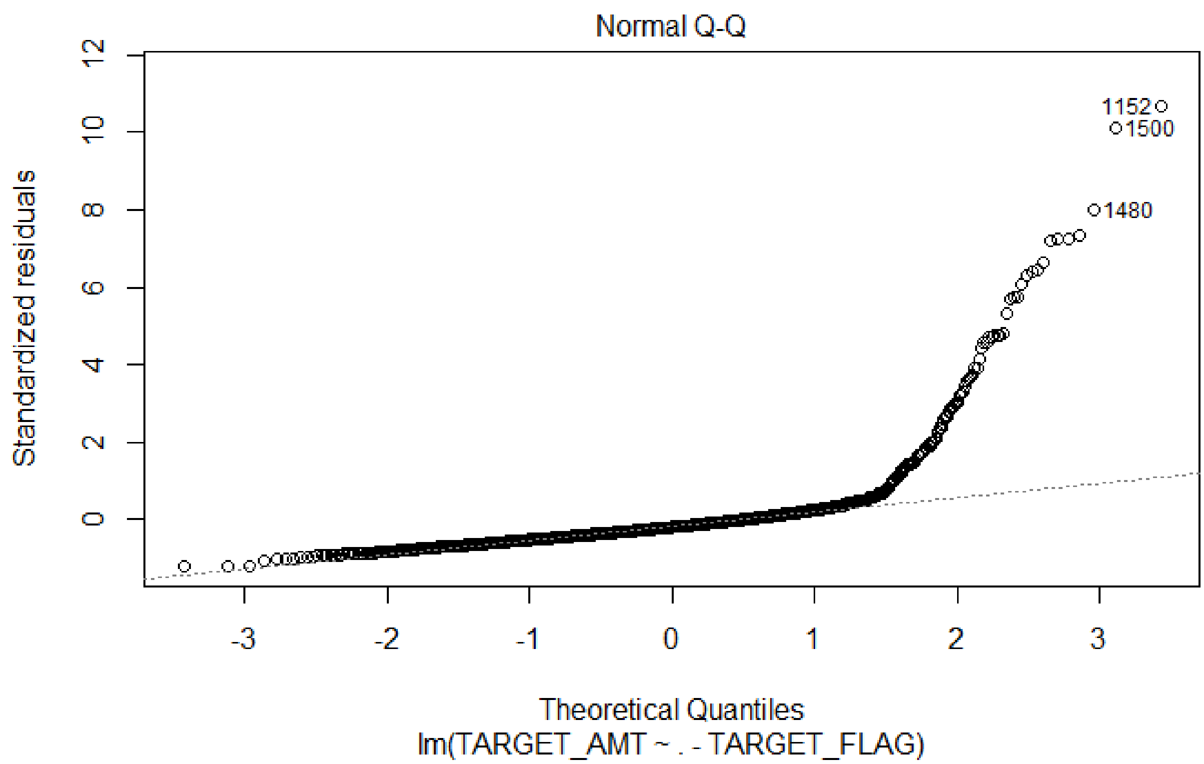
Residual standard error: 7200 on 1726 degrees of freedom
Multiple R-squared:  0.01471,	Adjusted R-squared:  0.013 
F-statistic: 8.591 on 3 and 1726 DF,  p-value: 1.163e-05
```

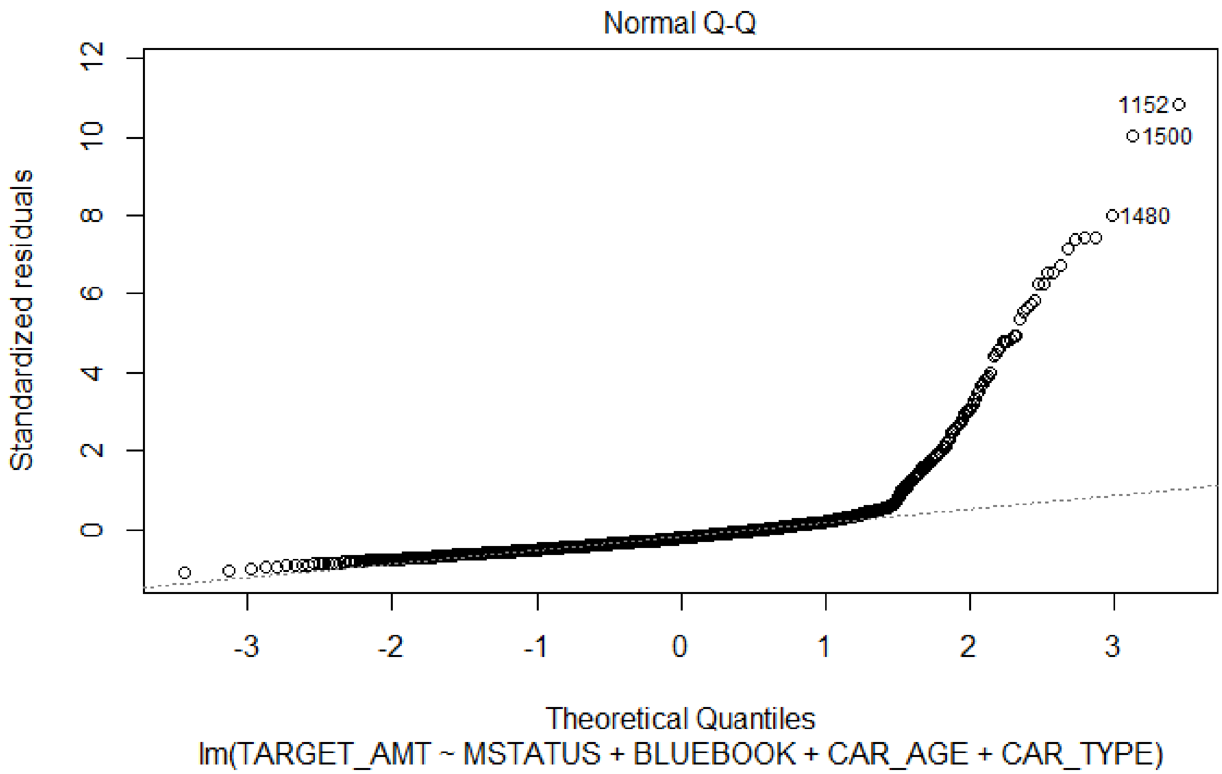
Select Models

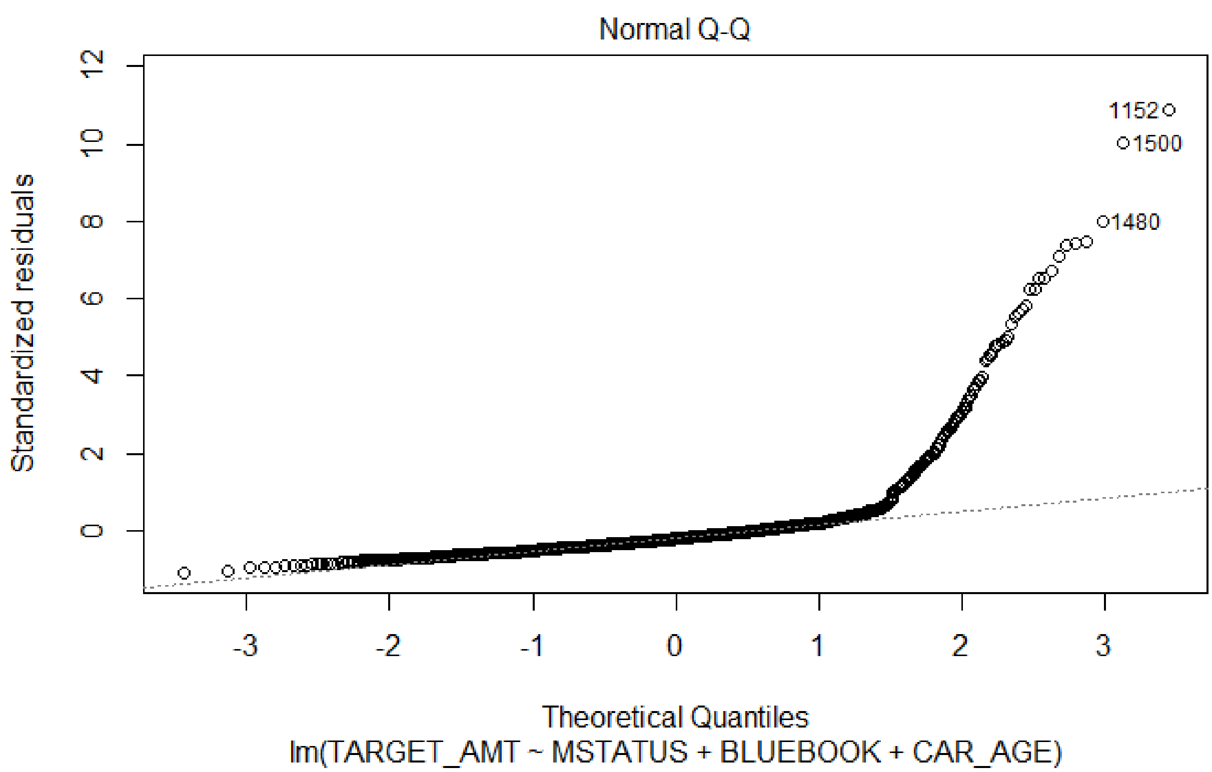
Linear Models

Looking at the r-squared value for each of the three linear models we notice that each performed relatively poor. The r-squared values were 0.03284, 0.01529 and 0.01471 for models 1, 2 and 3 respectively. The f-statistic for all models also appeared to be significant.

Looking at the plots of the models the biggest issues in each of the models is the normal qq plot. The quantile points do not appear to lie on the theoretical normal line. See below for models 1, 2 and 3 respectively:




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(TARGET_AMT ~ MSTATUS + BLUEBOOK + CAR_AGE + CAR_TYPE)
1152
1500
1480


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(TARGET_AMT ~ MSTATUS + BLUEBOOK + CAR_AGE)
1152
1500
1480

Based off the information presented, the models are ideally not what we would consider moving forward with. For the purpose of the project however, model 2 makes the most to proceed with. It has a better r-squared than model 3 and has variables that make sense regarding claim amount and not probability of crashing.

Logistic Models

To decide on selecting between the models we used ANOVA and McFaddens R^2. For ANOVA, we are looking for the widest gap between the null and residual deviance. Here is the ANOVA for the original model with all variables:

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: TARGET_FLAG

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                        6170     7129.6              
YOJ         1    29.14      6169     7100.5 6.726e-08 ***
INCOME      1    98.01      6168     7002.5 < 2.2e-16 ***
PARENT1     1   133.87      6167     6868.6 < 2.2e-16 ***
HOME_VAL    1    51.84      6166     6816.8 6.034e-13 ***
MSTATUS     1     9.15      6165     6807.6 0.0024927 ** 
SEX         1     0.08      6164     6807.6 0.7784726    
EDUCATION   4    48.59      6160     6759.0 7.120e-10 ***
JOB         8    95.42      6152     6663.6 < 2.2e-16 ***
TRAVTIME    1    11.45      6151     6652.1 0.0007136 ***
CAR_USE     1    58.43      6150     6593.7 2.104e-14 ***
TIF         1    41.36      6149     6552.3 1.267e-10 ***
CAR_TYPE    5    95.96      6144     6456.3 < 2.2e-16 ***
RED_CAR     1     0.03      6143     6456.3 0.8682045    
REVOKED     1   110.48      6142     6345.8 < 2.2e-16 ***
URBANICITY  1   647.79      6141     5698.0 < 2.2e-16 ***
PTS_AGE     1    88.25      6140     5609.8 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: TARGET_FLAG

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        6170     7129.6
INCOME      1   122.55      6169     7007.1 < 2.2e-16 ***
PARENT1     1   135.19      6168     6871.9 < 2.2e-16 ***
HOME_VAL    1    54.60      6167     6817.3 1.477e-13 ***
MSTATUS     1     9.46      6166     6807.8 0.0020975 **
EDUCATION   4    47.65      6162     6760.2 1.119e-09 ***
JOB         8    92.07      6154     6668.1 < 2.2e-16 ***
TRAVTIME    1    11.30      6153     6656.8 0.0007744 ***
CAR_USE     1    49.51      6152     6607.3 1.972e-12 ***
TIF         1    41.47      6151     6565.8 1.195e-10 ***
CAR_TYPE    5   102.97      6146     6462.9 < 2.2e-16 ***
REVOKED     1   111.18      6145     6351.7 < 2.2e-16 ***
URBANICITY  1   648.52      6144     5703.2 < 2.2e-16 ***
PTS_AGE     1    88.24      6143     5614.9 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: TARGET_FLAG

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        6170     7129.6
INCOME      1   122.55      6169     7007.1 < 2.2e-16 ***
PARENT1     1   135.19      6168     6871.9 < 2.2e-16 ***
HOME_VAL    1    54.60      6167     6817.3 1.477e-13 ***
MSTATUS     1     9.46      6166     6807.8 0.0020975 **
EDUCATION   4    47.65      6162     6760.2 1.119e-09 ***
TRAVTIME    1    14.90      6161     6745.3 0.0001133 ***
CAR_USE     1   103.78      6160     6641.5 < 2.2e-16 ***
TIF         1    41.37      6159     6600.1 1.258e-10 ***
CAR_TYPE    5   100.32      6154     6499.8 < 2.2e-16 ***
REVOKED     1   113.88      6153     6385.9 < 2.2e-16 ***
URBANICITY  1   612.61      6152     5773.3 < 2.2e-16 ***
PTS_AGE     1    96.10      6151     5677.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
 McFadden
0.2569846
 McFadden
0.2563038
 McFadden
0.2480526
```

The ANOVA for each model is in order above, as are the McFadden scores. Based on this information, even though model 2 had a slightly lower R2 than model 1, it makes the most sense as far as variable coefficients and AIC. Going forward with testing this model on the prediction set, we get an accuracy of 78%.

```
```{r logit_2 pred}
fitted.results = predict(logit_2, test, type = 'response')
fitted.results = ifelse(fitted.results > 0.5, 1, 0)

misClasificError = mean(fitted.results != test$TARGET_FLAG, na.rm = TRUE)
print(paste('Accurancy', round(1-misClasificError, 3)))
```
```

```
[1] "Accurancy 0.784"
```

Logistic Model Prediction: https://github.com/ChristinaValore/Business-Analytics-and-Data-Mining-621/blob/master/Homework4/logistic_model_eval.csv

Linear Regression Model Prediction: https://github.com/ChristinaValore/Business-Analytics-and-Data-Mining-621/blob/master/Homework4/linear_model_eval.csv