# DATA621 HW4

Misha Kollontai

11/6/2020

## 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 also 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). Below is a short description of the variables of interest in the data set:

## Data Exploration

Let's calculate summary statistics and generate a box plot for further review. The income, home valuem bluebook and old claim data was converted to numeric data in order to make it easier to work with.

```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS            YOJ           INCOME
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Min.   :     0
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097
##  Median :45.00   Median :0.0000   Median :11.0   Median : 54028
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5   Mean   : 61898
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   Max.   :367030
##  NA's   :6                        NA's   :454    NA's   :445
##  PARENT1       HOME_VAL      MSTATUS      SEX              EDUCATION
##  No :7084   Min.   :     0   Yes :4894   M  :3786   <High School :1203
##  Yes:1077   1st Qu.:     0   z_No:3267   z_F:4375   Bachelors    :2242
##             Median :161160                          Masters      :1658
##             Mean   :154867                          PhD          : 728
##             3rd Qu.:238724                          z_High School:2330
```

```
##             Max.   :885282
##             NA's   :464
##            JOB            TRAVTIME            CAR_USE        BLUEBOOK
##  z_Blue Collar:1825   Min.   :  5.00   Commercial:3029   Min.   : 1500
##  Clerical     :1271   1st Qu.: 22.00   Private   :5132   1st Qu.: 9280
##  Professional :1117   Median : 33.00                     Median :14440
##  Manager      : 988   Mean   : 33.49                     Mean   :15710
##  Lawyer       : 835   3rd Qu.: 44.00                     3rd Qu.:20850
##  Student      : 712   Max.   :142.00                     Max.   :69740
##  (Other)      :1413
##       TIF                CAR_TYPE    RED_CAR       OLDCLAIM
##  Min.   : 1.000   Minivan    :2145   no :5783   Min.   :    0
##  1st Qu.: 1.000   Panel Truck: 676   yes:2378   1st Qu.:    0
##  Median : 4.000   Pickup     :1389              Median :    0
##  Mean   : 5.351   Sports Car : 907              Mean   : 4037
##  3rd Qu.: 7.000   Van        : 750              3rd Qu.: 4636
##  Max.   :25.000   z_SUV      :2294              Max.   :57037
##
##     CLM_FREQ      REVOKED       MVR_PTS          CAR_AGE
##  Min.   :0.0000   No :7161   Min.   : 0.000   Min.   :-3.000
##  1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.0000              Median : 1.000   Median : 8.000
##  Mean   :0.7986              Mean   : 1.696   Mean   : 8.328
##  3rd Qu.:2.0000              3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :5.0000              Max.   :13.000   Max.   :28.000
##                                               NA's   :510
##                  URBANICITY
##  Highly Urban/ Urban  :6492
##  z_Highly Rural/ Rural:1669
##
##
##
##
##
```

There is no missing (NA) data, though there are some zero-values in the dataset. In order to see what effect each of our variables may have on our predictive model, let's take a look and see how the variables relate to the probability of getting into an accident.

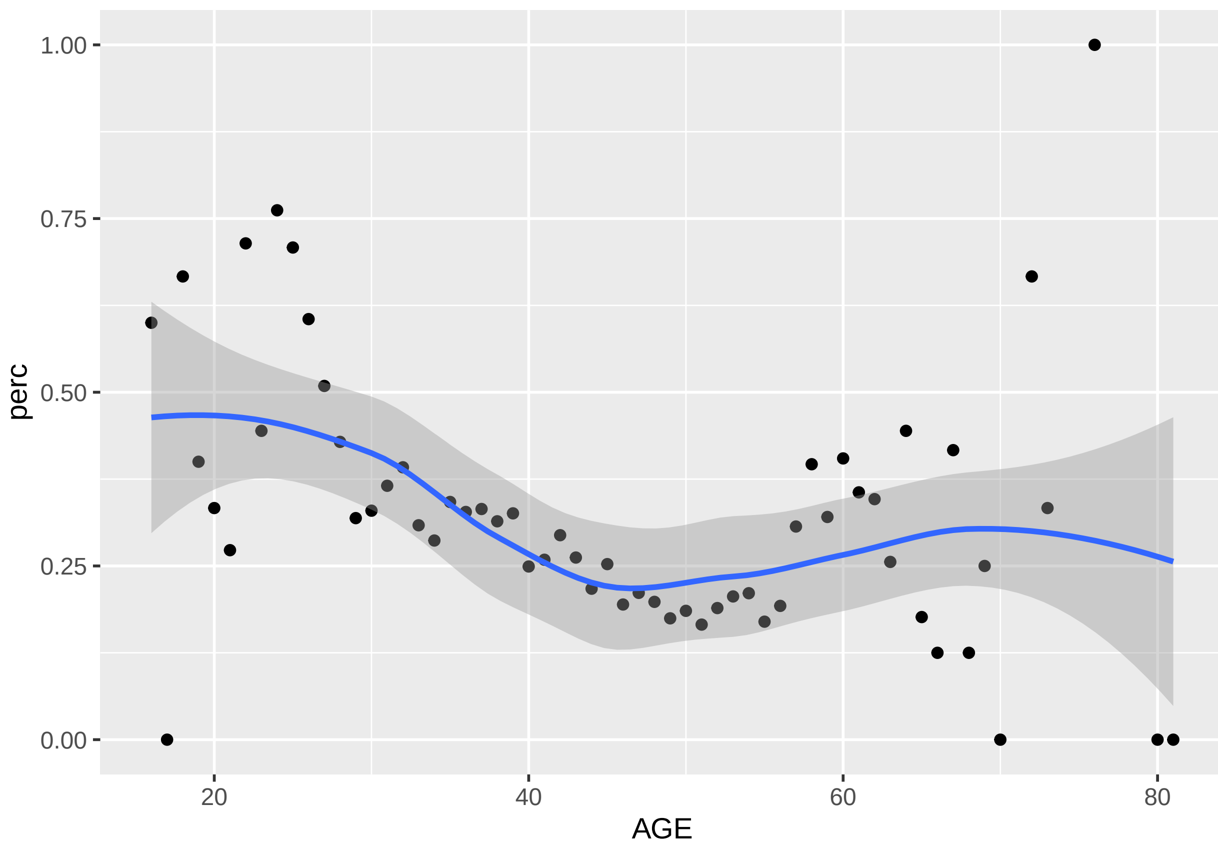
```
##
##    0    1
## 6008 2153
```

We can see that a vast majority of our data is for vehicles that did not get into an accident.

### Age

Conventional wisdom indcates that younger people tend to drive more recklessly, let's see how much our data agrees with this sentiment.

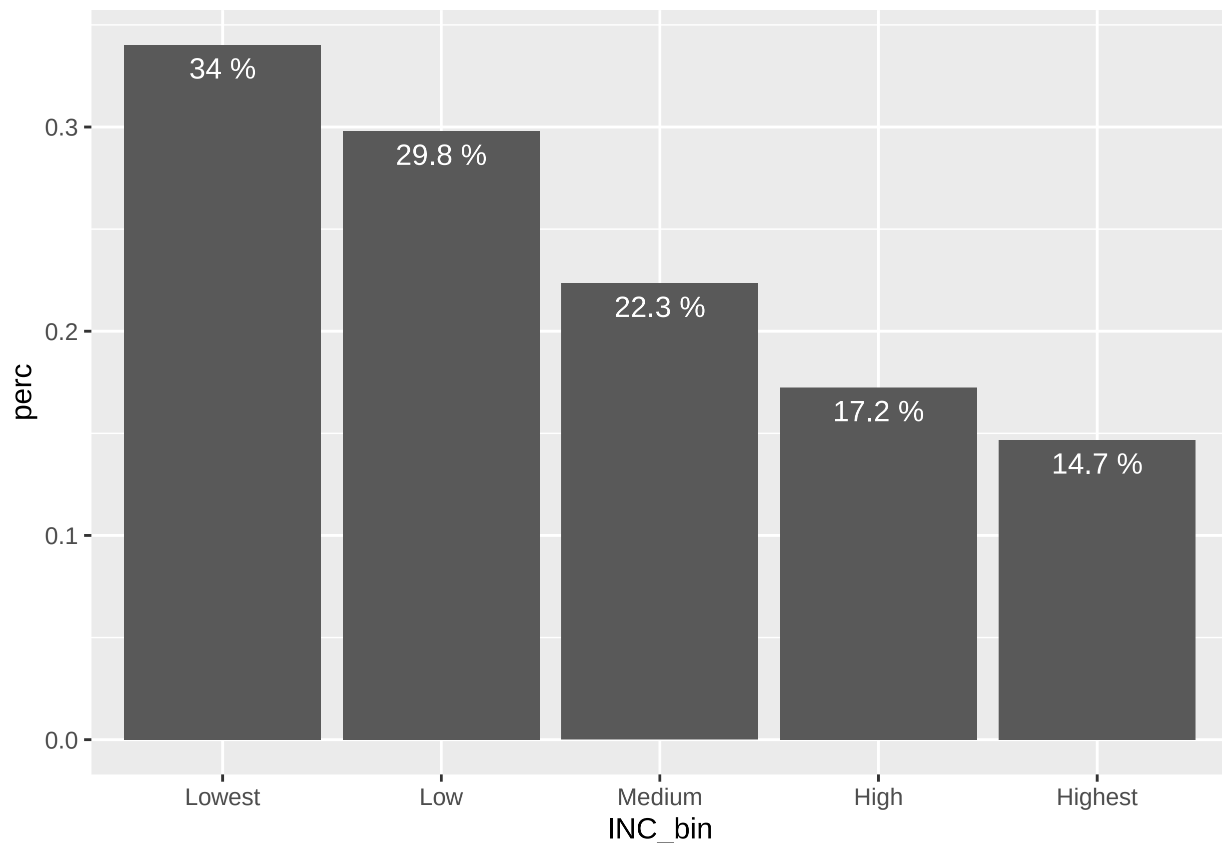
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




While the scatter in the data is significant, the trendline suggests that there are in fact two peaks, with both younger and older drivers getting into more accidents. Drivers of ages between 40 and 55 seem to be the safest, so the relationship between age and likelihood to get into an accident will not be linear.

**Income**

The data description suggests that “rich people tend to get into fewer crashes”. What does our data show? In order to yield a clearer visualization, we will bin the income data before plotting. To do so we will use the ‘clusters’ method from the *bin* function of the OneR package. Furthermore, there is income data missing from some entries; for the purpose of this first glance, we will simpy ignore these datapoints.

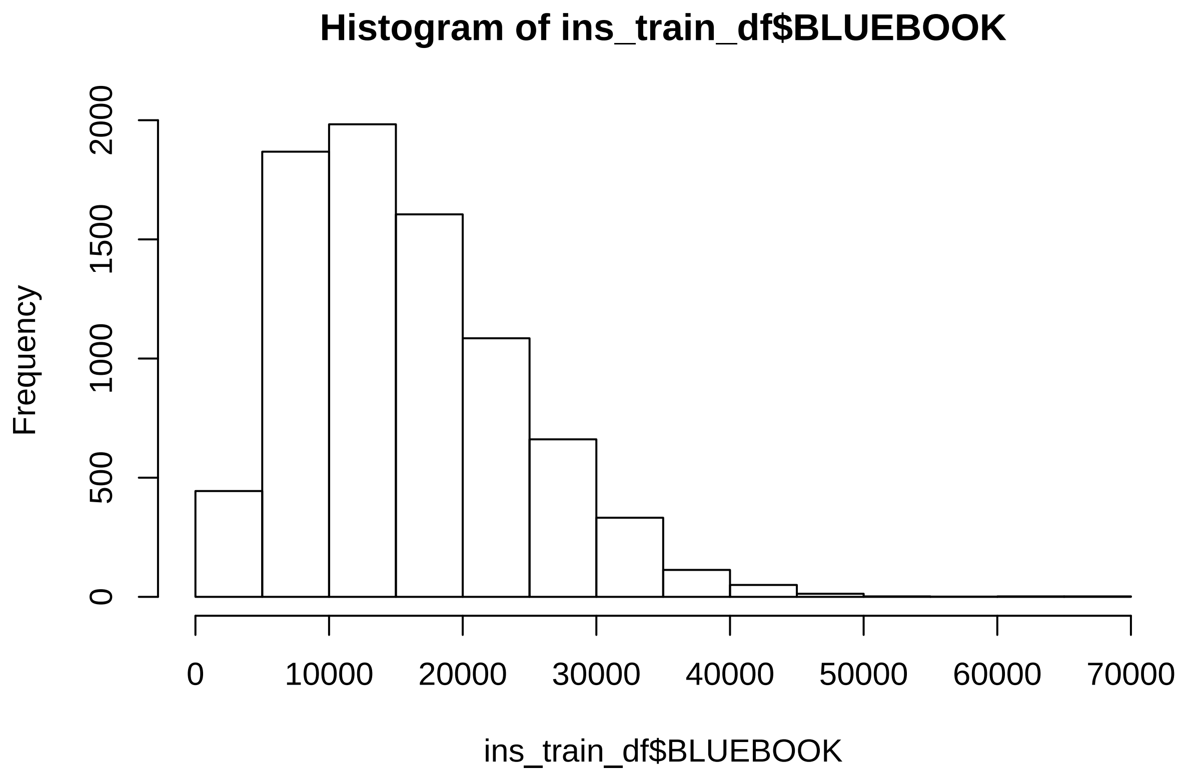


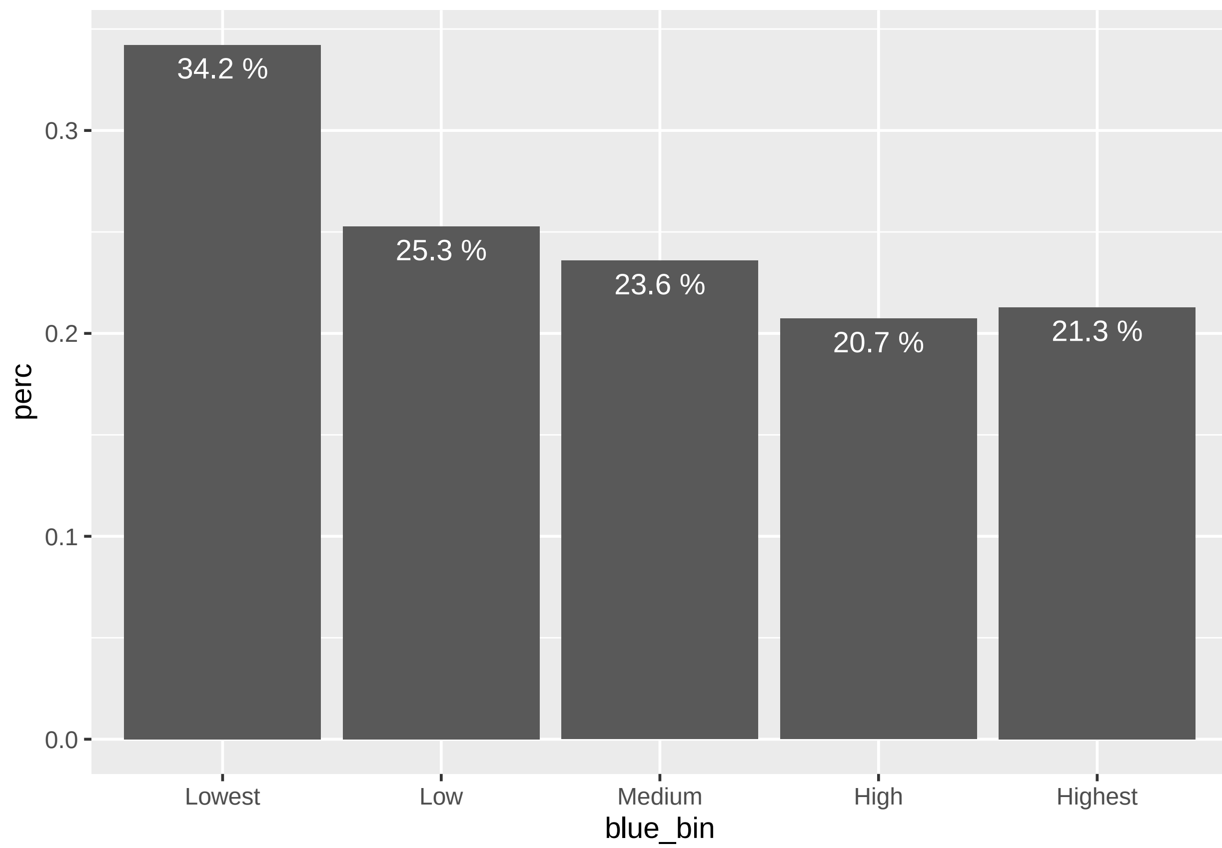

Here we can see a clear trend of higher earners getting into fewer car accidents. The highest earners are less than half as likely to get into an accident than the lowest earners.

Let's see if the Bluebook value of the car a person drives shows a similar correlation:

### Bluebook

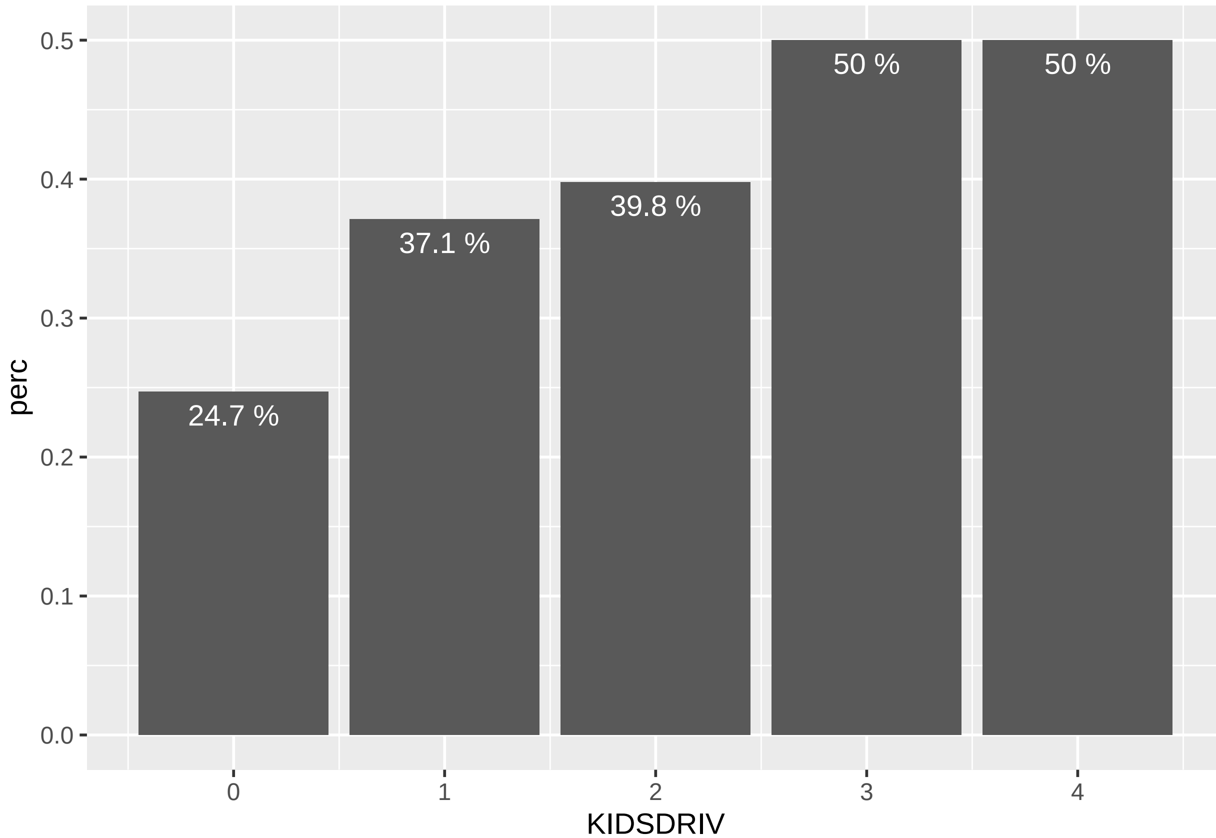
Does the 'value' of the car a person drives impact their likelihood to get into an accident? Let's repeat the analysis performed above for the income, binning hte bluebook values and ignoring NA values (for now).


Histogram of ins_train_df$BLUEBOOK
Frequency
0
500
1000
1500
2000
0
10000
20000
30000
40000
50000
60000
70000
ins_train_df$BLUEBOOK



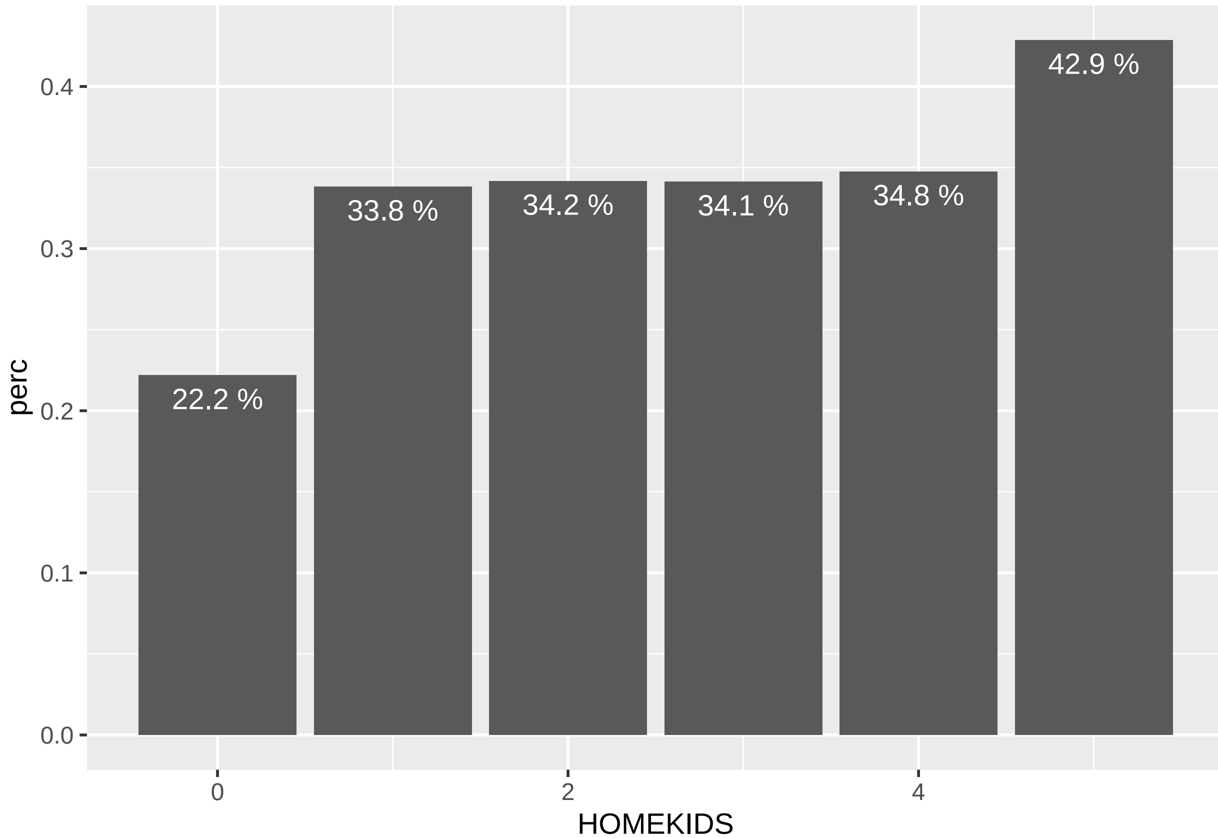

In the Bluebook value it seems that the relationship is not as negative as it was for income - the lowest value cars are most likely to be in accidents, with the other 4 bins showing fairly similar likelihoods.

**Kids in the Household Who Drive**




There seems to be a trend between the number of driving teens in a household and the likelihood of getting into an accident - this variable KIDSDRIV will likey be a strong predictor for the likelihood of an accident.

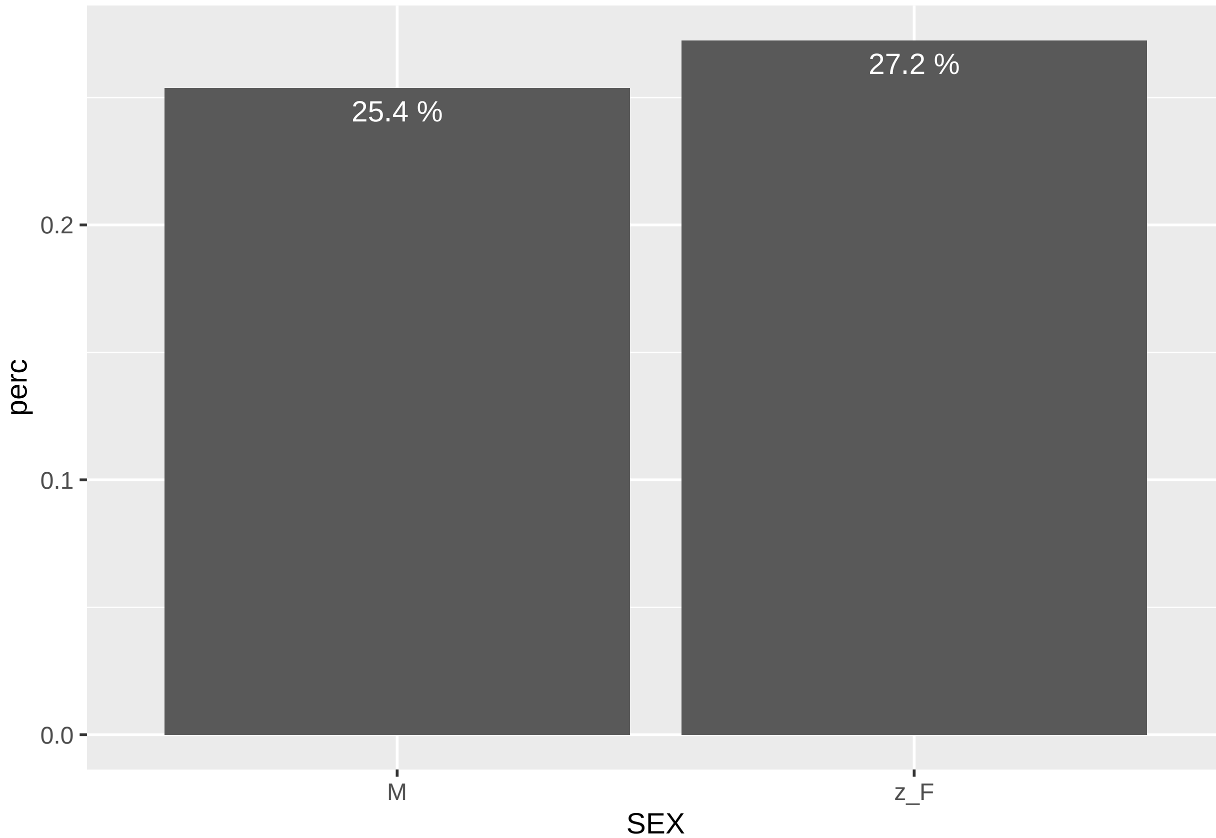
**Kids in the Household**




The overall number of kids in a household seems to follow the pattern of the data related to teenagers in the house who can drive. This could suggest that the unsafe driving practices of the eligible kids outweigh the added precautions taken by parents of many children. We will see if this makes it into our model.

**Gender**

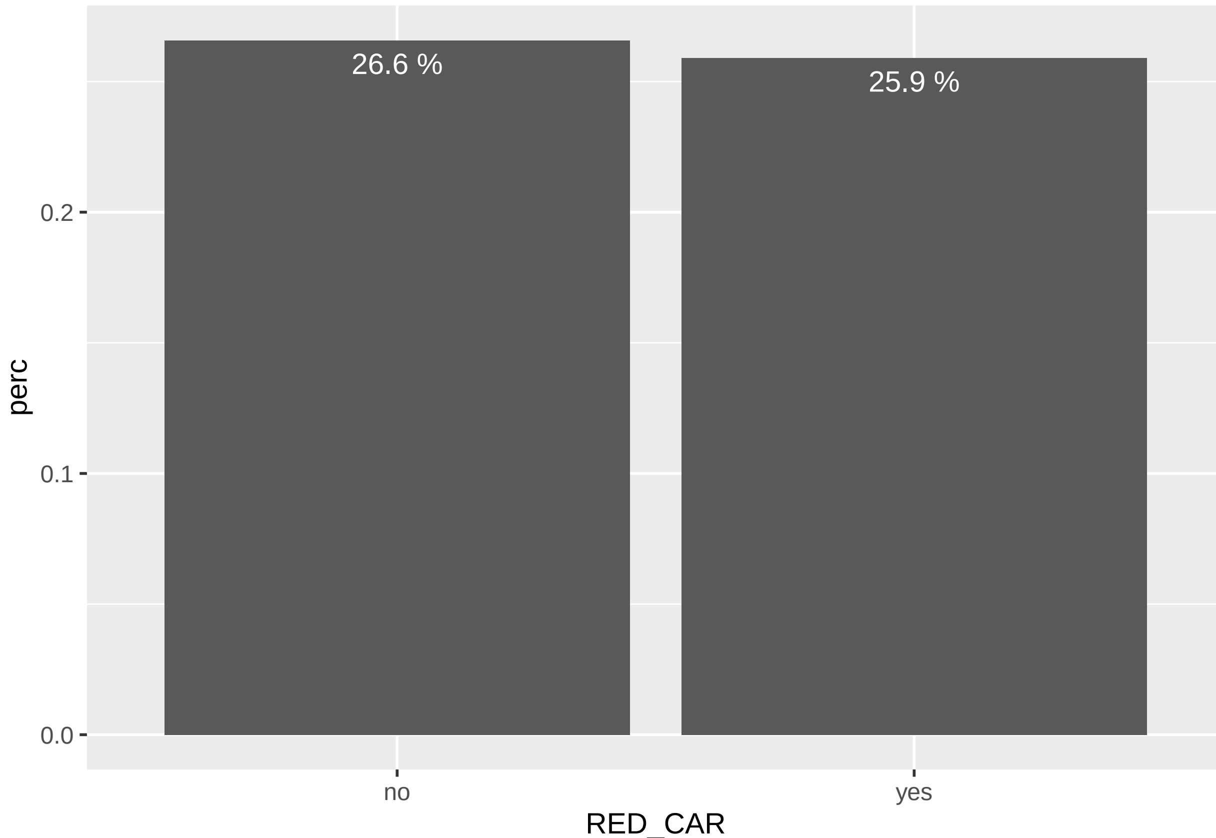
Let us look into the urban legend that women have less crashes than men:




There does appear to be a slightly larger percentage of females who get into accidents based on this data contrary to the urban legends. The difference is about 2%, however, and is unlikely to be a very strong predictor.

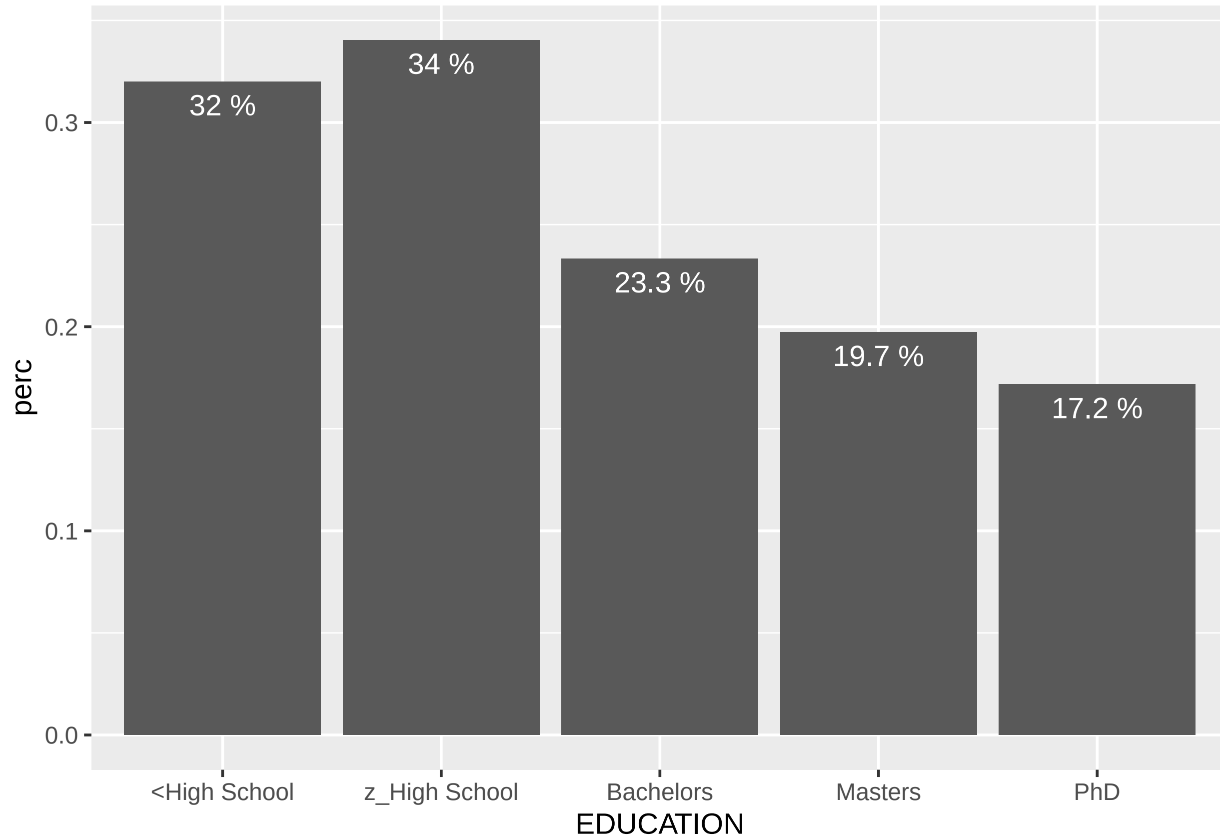
**Red Car**

Following up one urban legend with another- are red cars more likely to get into an accident than other vehicles?



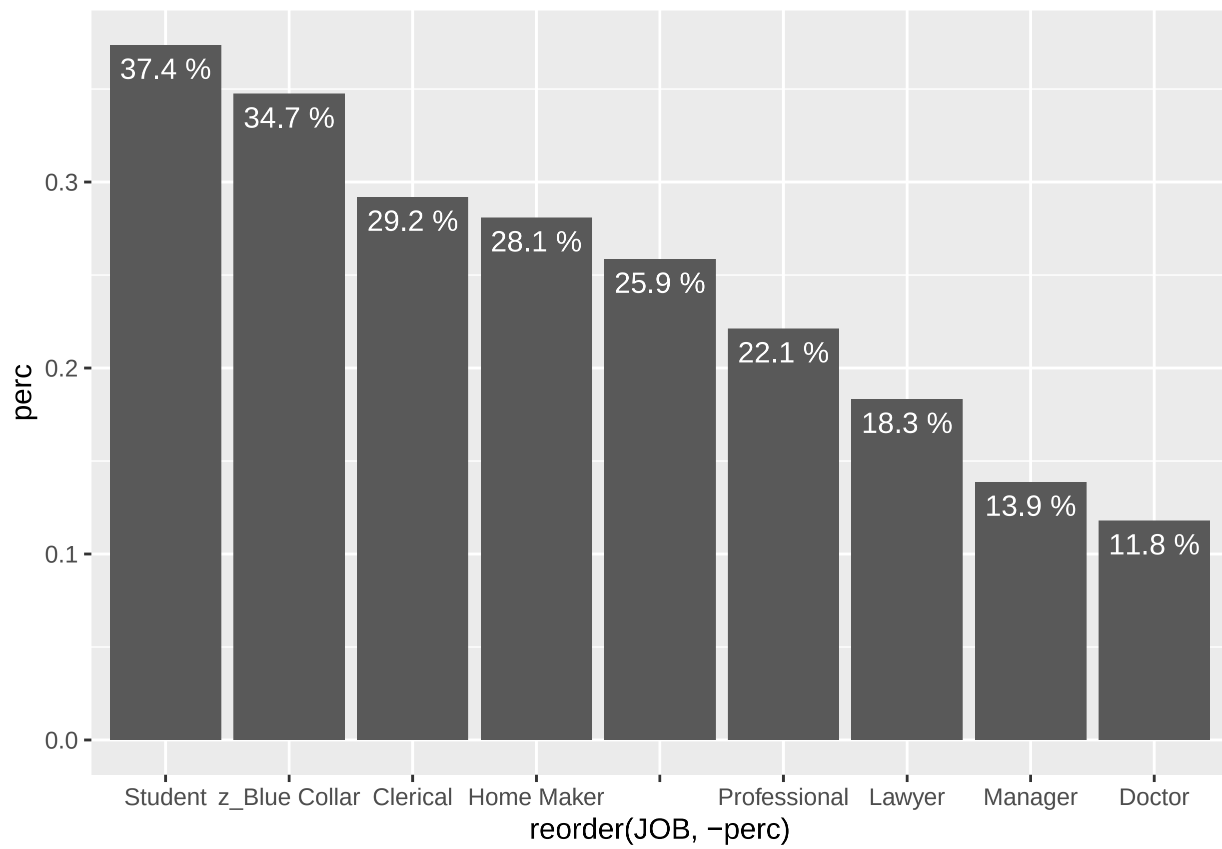

This data suggests that there is no indication of red cars getting into more accidents - a smaller proportion of red cars were involved in accidents than non-red cars were (in this dataset). Based on this data, the RED_CAR variable is unlikely to add much value to our model.

**Education**




From the image above we can clearly see that a higher percentage of people without a college education will get into an accident. Based on this information, EDUCATION will certainly be used as a predictor in our model.

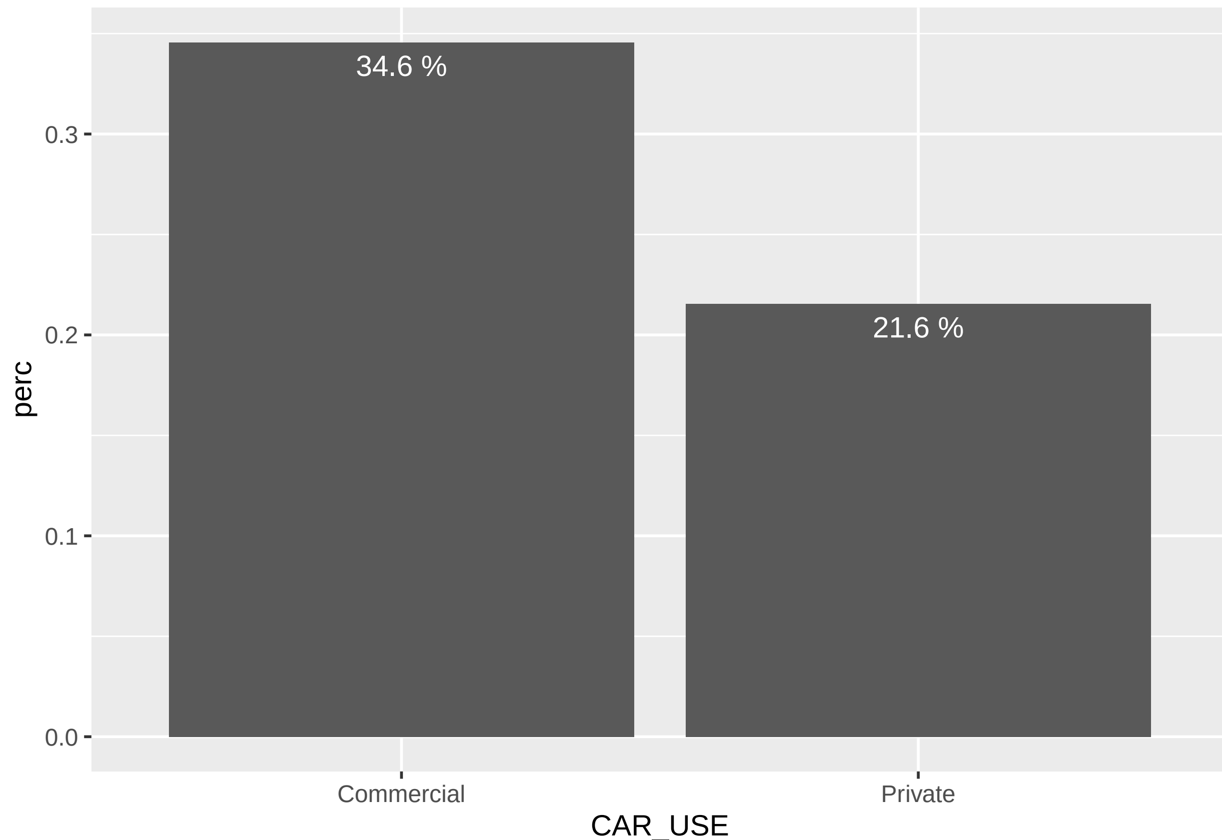
**Job**




From this breakdown based on Job we cna see that there are certain careers that correlate to a higher number of accidents. This suggests that the JOB variable wil be a valuable predictor for our model.

**Car Use**

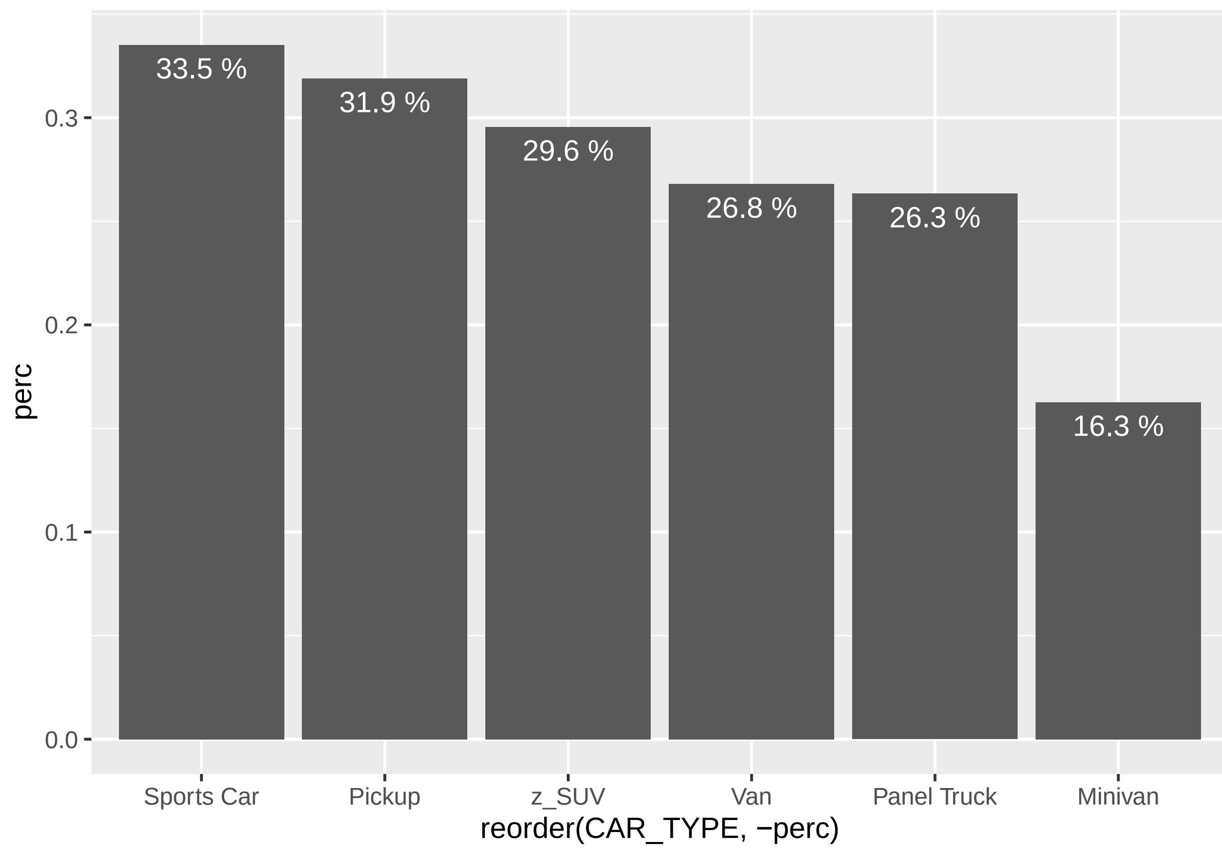
There is a suggestion that what a vehicle is used for may have an impact on accident likelihood. Commercial vehicles are driven more frequently than their private counterparts, so the vehicle is exposed to more opportunities for accidents.




The data does seem to support the hypothesis, with commercial vehicles ~13% mre likely to be in an accident.

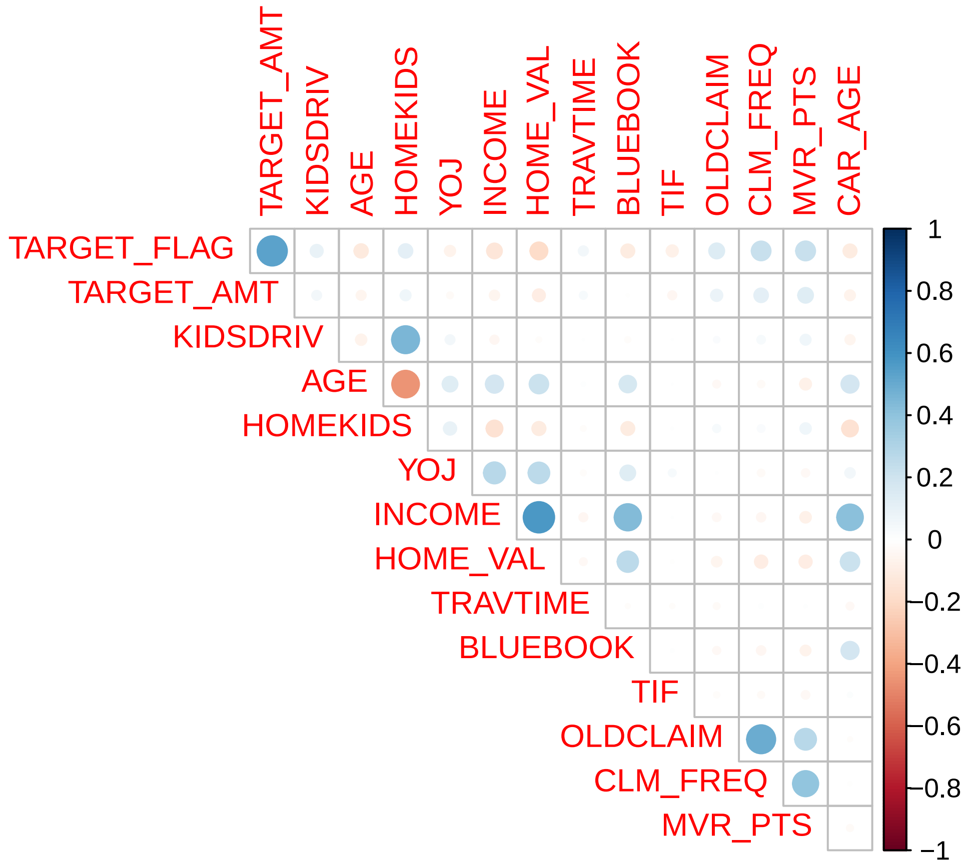
#### Car Type

What about car type - are there certain types of cars that seem to get into more accidents than others?




Based on this visualization it seems that Sports Cars are most likely to get nti ab accident, with Minivans seemingly the safest. This just about follows what we would expect as sports cars have a reputation for reckless driving, while minivans are more often owned by safety-conscious families.

Let’s check the correlation plot generated from our dataset.




With respect to the Target Flag, few variables show strong correlations in one direction or another, with Home_Val, CLM_Freq and MVR_PTS standing out somewhat.

## Data Preparation

### Imputation

What columns are missing data?

```
##        INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##            0           0           0           0           6           0
##          YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##          454         445           0         464           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     INC_bin    blue_bin
##          510           0           0           0
```

We will replace the missing $Age$, $Income$, $YearOnJob$, $HomeValue$ and $CarAge$ values with the median values for each category.

```
##        INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##            0           0           0           0           0           0
```

```
##          YOJ       INCOME      PARENT1     HOME_VAL      MSTATUS          SEX
##            0            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      INC_bin     blue_bin
##            0            0            0            0
```

#### Transforming Data

We created two new variables above, binning the Income column as well as the Bluebook columns above in order to better visualize the distribution of the data.

## Build Models

To start, let's create some binary logistic regression models that will predict whether or not someone will get into an accident. We can then use this prediction to estimate the cost associated with said accident.

#### Binary Logistic Regressions

#### Model 1 - First Binary Logistic Regression

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial, data = flag_train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5787  -0.7110  -0.3978   0.6283   3.1527
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -7.756e-01  3.331e-01  -2.328 0.019891 *
## KIDSDRIV                        3.916e-01  6.137e-02   6.381 1.76e-10 ***
## AGE                            -9.612e-04  4.033e-03  -0.238 0.811637
## HOMEKIDS                        4.814e-02  3.721e-02   1.294 0.195705
## YOJ                            -9.641e-03  8.676e-03  -1.111 0.266520
## INCOME                         -5.273e-06  2.688e-06  -1.962 0.049794 *
## PARENT1Yes                      3.837e-01  1.098e-01   3.493 0.000477 ***
## HOME_VAL                       -1.291e-06  3.409e-07  -3.787 0.000152 ***
## MSTATUSz_No                     4.985e-01  8.363e-02   5.961 2.51e-09 ***
## SEXz_F                         -9.127e-02  1.129e-01  -0.809 0.418672
## EDUCATIONBachelors             -3.528e-01  1.187e-01  -2.972 0.002962 **
## EDUCATIONMasters               -2.530e-01  1.806e-01  -1.401 0.161177
## EDUCATIONPhD                   -1.821e-01  2.149e-01  -0.847 0.396902
## EDUCATIONz_High School          2.932e-02  9.756e-02   0.301 0.763791
## JOBClerical                     4.154e-01  1.974e-01   2.105 0.035299 *
## JOBDoctor                      -3.988e-01  2.683e-01  -1.486 0.137260
## JOBHome Maker                   1.961e-01  2.180e-01   0.900 0.368371
## JOBLawyer                       1.326e-01  1.701e-01   0.780 0.435422
## JOBManager                     -5.224e-01  1.721e-01  -3.036 0.002398 **
```

```
## JOBProfessional                  1.994e-01  1.791e-01   1.113 0.265711    
## JOBStudent                       1.717e-01  2.229e-01   0.770 0.441066    
## JOBz_Blue Collar                 3.428e-01  1.862e-01   1.841 0.065683 .  
## TRAVTIME                         1.461e-02  1.886e-03   7.747 9.44e-15 ***
## CAR_USEPrivate                  -7.629e-01  9.188e-02  -8.303  < 2e-16 ***
## BLUEBOOK                        -3.271e-05  1.362e-05  -2.401 0.016329 *  
## TIF                             -5.527e-02  7.351e-03  -7.518 5.56e-14 ***
## CAR_TYPEPanel Truck              4.993e-01  1.737e-01   2.874 0.004056 ** 
## CAR_TYPEPickup                   5.318e-01  1.023e-01   5.199 2.01e-07 ***
## CAR_TYPESports Car               1.007e+00  1.304e-01   7.726 1.11e-14 ***
## CAR_TYPEVan                      6.144e-01  1.372e-01   4.479 7.51e-06 ***
## CAR_TYPEz_SUV                    7.649e-01  1.118e-01   6.840 7.90e-12 ***
## RED_CARyes                      -9.145e-03  8.646e-02  -0.106 0.915760    
## OLDCLAIM                        -1.396e-05  3.914e-06  -3.566 0.000363 ***
## CLM_FREQ                         1.965e-01  2.859e-02   6.873 6.27e-12 ***
## REVOKEDYes                       8.891e-01  9.141e-02   9.726  < 2e-16 ***
## MVR_PTS                          1.136e-01  1.365e-02   8.319  < 2e-16 ***
## CAR_AGE                         -6.989e-04  7.543e-03  -0.093 0.926184    
## URBANICITYz_Highly Rural/ Rural -2.396e+00  1.131e-01 -21.188  < 2e-16 ***
## INC_binLow                       7.962e-03  1.235e-01   0.064 0.948607    
## INC_binMedium                    2.455e-02  1.924e-01   0.128 0.898495    
## INC_binHigh                      7.769e-02  3.002e-01   0.259 0.795766    
## INC_binHighest                   4.546e-01  5.036e-01   0.903 0.366692    
## INC_binNA                       -8.068e-03  1.788e-01  -0.045 0.964002    
## blue_binLow                      2.494e-03  1.138e-01   0.022 0.982522    
## blue_binMedium                   8.909e-02  1.949e-01   0.457 0.647532    
## blue_binHigh                     1.538e-01  2.906e-01   0.529 0.596634    
## blue_binHighest                  6.725e-01  4.325e-01   1.555 0.119989    
## ---
## 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: 7287.1  on 8114  degrees of freedom
## AIC: 7381.1
## 
## Number of Fisher Scoring iterations: 5
```

Looking over some of the coefficients, we see a negative relationship with the bluebook value, Time in Force, Old Claims, while the relationships with A Revoked License history, Motor Vehicle Record Points and Travel Time is positive - this aligns with what we would exect to see.

For our second model, let's reduce the number of less significant variables and trim the model somewhat by stepwise removing variables that have insignificant p-values.

### Model 2 - Trimmed Binary Logistic Regression

```
## 
## Call:
## glm(formula = TARGET_FLAG ~ . - AGE - HOMEKIDS - YOJ - INCOME - 
##     INC_bin - blue_bin - CAR_AGE - RED_CAR - SEX, family = binomial, 
##     data = flag_train_data)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5951  -0.7142  -0.4017   0.6251   3.1588
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.169e+00  2.532e-01  -4.617 3.90e-06 ***
## KIDSDRIV                        4.130e-01  5.502e-02   7.506 6.11e-14 ***
## PARENT1Yes                      4.689e-01  9.418e-02   4.979 6.39e-07 ***
## HOME_VAL                       -1.693e-06  3.194e-07  -5.302 1.15e-07 ***
## MSTATUSz_No                     4.199e-01  7.814e-02   5.373 7.74e-08 ***
## EDUCATIONBachelors             -4.467e-01  1.072e-01  -4.166 3.09e-05 ***
## EDUCATIONMasters               -3.864e-01  1.596e-01  -2.422  0.01545 *
## EDUCATIONPhD                   -3.618e-01  1.932e-01  -1.872  0.06116 .
## EDUCATIONz_High School         -8.091e-03  9.435e-02  -0.086  0.93166
## JOBClerical                     5.211e-01  1.937e-01   2.690  0.00716 **
## JOBDoctor                      -4.204e-01  2.659e-01  -1.581  0.11387
## JOBHome Maker                   4.732e-01  1.950e-01   2.427  0.01521 *
## JOBLawyer                       1.396e-01  1.685e-01   0.828  0.40761
## JOBManager                     -5.324e-01  1.707e-01  -3.120  0.00181 **
## JOBProfessional                 1.991e-01  1.778e-01   1.120  0.26282
## JOBStudent                      4.319e-01  2.051e-01   2.105  0.03527 *
## JOBz_Blue Collar                3.643e-01  1.848e-01   1.972  0.04864 *
## TRAVTIME                        1.435e-02  1.880e-03   7.631 2.32e-14 ***
## CAR_USEPrivate                 -7.576e-01  9.157e-02  -8.273  < 2e-16 ***
## BLUEBOOK                       -2.551e-05  4.647e-06  -5.490 4.03e-08 ***
## TIF                            -5.496e-02  7.329e-03  -7.500 6.40e-14 ***
## CAR_TYPEPanel Truck             6.083e-01  1.508e-01   4.034 5.48e-05 ***
## CAR_TYPEPickup                  5.540e-01  1.005e-01   5.510 3.58e-08 ***
## CAR_TYPESports Car              9.726e-01  1.074e-01   9.059  < 2e-16 ***
## CAR_TYPEVan                     6.405e-01  1.220e-01   5.248 1.54e-07 ***
## CAR_TYPEz_SUV                   7.143e-01  8.592e-02   8.314  < 2e-16 ***
## OLDCLAIM                       -1.390e-05  3.904e-06  -3.561  0.00037 ***
## CLM_FREQ                        1.972e-01  2.850e-02   6.917 4.61e-12 ***
## REVOKEDYes                      8.903e-01  9.112e-02   9.771  < 2e-16 ***
## MVR_PTS                         1.155e-01  1.357e-02   8.512  < 2e-16 ***
## URBANICITYz_Highly Rural/ Rural -2.384e+00  1.127e-01 -21.154  < 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: 7312.4  on 8130  degrees of freedom
## AIC: 7374.4
##
## Number of Fisher Scoring iterations: 5
```

The two professions that seem to stand out in this model seem to be the 'Clerical' and 'Manager' designations. Lets remove the overall Job variable and create two new ones designating whether the car owner falls into one of those categories.

#### Model 3 - Third Binary Logistic Regression

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - AGE - HOMEKIDS - YOJ - INCOME -
##     INC_bin - blue_bin - CAR_AGE - RED_CAR - SEX - JOB, family = binomial,
##     data = flag_train_data_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6134  -0.7173  -0.4043   0.6300   3.1250
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -7.239e-01  1.616e-01  -4.479 7.51e-06 ***
## KIDSDRIV                         4.168e-01  5.496e-02   7.584 3.34e-14 ***
## PARENT1Yes                       4.793e-01  9.388e-02   5.105 3.31e-07 ***
## HOME_VAL                        -1.888e-06  2.977e-07  -6.343 2.26e-10 ***
## MSTATUSz_No                      3.875e-01  7.659e-02   5.060 4.20e-07 ***
## EDUCATIONBachelors              -5.127e-01  9.876e-02  -5.191 2.09e-07 ***
## EDUCATIONMasters                -6.007e-01  1.088e-01  -5.522 3.35e-08 ***
## EDUCATIONPhD                    -7.429e-01  1.400e-01  -5.306 1.12e-07 ***
## EDUCATIONz_High School          -3.287e-02  9.183e-02  -0.358 0.720352
## TRAVTIME                         1.439e-02  1.878e-03   7.662 1.83e-14 ***
## CAR_USEPrivate                  -7.931e-01  7.627e-02 -10.398  < 2e-16 ***
## BLUEBOOK                        -2.646e-05  4.609e-06  -5.741 9.42e-09 ***
## TIF                             -5.503e-02  7.320e-03  -7.517 5.60e-14 ***
## CAR_TYPEPanel Truck              5.748e-01  1.436e-01   4.002 6.29e-05 ***
## CAR_TYPEPickup                   5.377e-01  9.876e-02   5.445 5.19e-08 ***
## CAR_TYPESports Car               9.978e-01  1.063e-01   9.388  < 2e-16 ***
## CAR_TYPEVan                      6.150e-01  1.200e-01   5.125 2.98e-07 ***
## CAR_TYPEz_SUV                    7.319e-01  8.512e-02   8.599  < 2e-16 ***
## OLDCLAIM                        -1.375e-05  3.900e-06  -3.526 0.000421 ***
## CLM_FREQ                         1.961e-01  2.846e-02   6.890 5.59e-12 ***
## REVOKEDYes                       8.866e-01  9.097e-02   9.746  < 2e-16 ***
## MVR_PTS                          1.148e-01  1.355e-02   8.471  < 2e-16 ***
## URBANICITYz_Highly Rural/ Rural -2.359e+00  1.121e-01 -21.042  < 2e-16 ***
## Manager                         -7.519e-01  1.070e-01  -7.026 2.13e-12 ***
## Clerical                         1.675e-01  8.793e-02   1.905 0.056782 .
## ---
## 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: 7325.8  on 8136  degrees of freedom
## AIC: 7375.8
##
## Number of Fisher Scoring iterations: 5
```

Let's use this model to predict the likelihood of an accident. This data could then be used as an input for the determination of how high the accident value would be.

```
##   TARGET_FLAG KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1 HOME_VAL MSTATUS
## 1  0.12157108        0  60        0  11  67349      No        0    z_No
## 2  0.27116301        0  43        0  11  91449      No   257252    z_No
```

```
## 3  0.28192415        0  35        1  10  16039      No   124191     Yes
## 4  0.08102923        0  51        0  14  54028      No   306251     Yes
## 5  0.34455331        0  50        0  11 114986      No   243925     Yes
## 6  0.68434008        0  34        1  12 125301     Yes        0    z_No
##   SEX     EDUCATION           JOB TRAVTIME    CAR_USE BLUEBOOK TIF
## 1   M           PhD  Professional       14    Private    14230  11
## 2   M z_High School z_Blue Collar       22 Commercial    14940   1
## 3 z_F z_High School      Clerical        5    Private     4010   4
## 4   M  <High School z_Blue Collar       32    Private    15440   7
## 5 z_F           PhD        Doctor       36    Private    18000   1
## 6 z_F     Bachelors z_Blue Collar       46 Commercial    17430   1
##     CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1    Minivan     yes     4461        2      No       3      18
## 2    Minivan     yes        0        0      No       0       1
## 3      z_SUV      no    38690        2      No       3      10
## 4    Minivan     yes        0        0      No       0       6
## 5      z_SUV      no    19217        2     Yes       3      17
## 6 Sports Car      no        0        0      No       0       7
##            URBANICITY INC_bin blue_bin Manager Clerical
## 1 Highly Urban/ Urban  Medium      Low       0        0
## 2 Highly Urban/ Urban  Medium      Low       0        0
## 3 Highly Urban/ Urban  Lowest   Lowest       0        1
## 4 Highly Urban/ Urban      NA      Low       0        0
## 5 Highly Urban/ Urban    High   Medium       0        0
## 6 Highly Urban/ Urban    High   Medium       0        0
```

### Linear Logistic Regressions

Using our previously calcuated prediction for the accident likelihood as one of the inputs, we can create a linear model for calculating the amount expected to be associated with an accident.

To start, let's see how our model would look using all of the original available variables:

### Model 4

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = amt_train_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5858  -1696   -765    351 103803
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.041e+03  5.792e+02   1.798 0.072230 .
## KIDSDRIV                        3.153e+02  1.133e+02   2.784 0.005387 **
## AGE                             5.130e+00  7.077e+00   0.725 0.468519
## HOMEKIDS                        7.902e+01  6.544e+01   1.208 0.227260
## YOJ                            -4.420e+00  1.523e+01  -0.290 0.771678
## INCOME                         -4.325e-03  4.322e-03  -1.001 0.316936
## PARENT1Yes                      5.713e+02  2.022e+02   2.825 0.004737 **
## HOME_VAL                       -5.314e-04  5.917e-04  -0.898 0.369108
```

```
## MSTATUSz_No                        5.747e+02  1.450e+02   3.963 7.46e-05 ***
## SEXz_F                            -3.532e+02  1.854e+02  -1.905 0.056829 .
## EDUCATIONBachelors                -2.675e+02  2.107e+02  -1.270 0.204287
## EDUCATIONMasters                   1.632e+01  3.033e+02   0.054 0.957082
## EDUCATIONPhD                       2.748e+02  3.569e+02   0.770 0.441400
## EDUCATIONz_High School            -1.037e+02  1.762e+02  -0.588 0.556234
## JOBClerical                        5.329e+02  3.428e+02   1.555 0.120082
## JOBDoctor                         -4.631e+02  4.109e+02  -1.127 0.259852
## JOBHome Maker                      3.854e+02  3.780e+02   1.020 0.307913
## JOBLawyer                          2.416e+02  2.964e+02   0.815 0.414952
## JOBManager                        -4.703e+02  2.895e+02  -1.625 0.104283
## JOBProfessional                    4.634e+02  3.096e+02   1.497 0.134506
## JOBStudent                         3.246e+02  3.899e+02   0.833 0.405121
## JOBz_Blue Collar                   5.064e+02  3.227e+02   1.569 0.116571
## TRAVTIME                           1.195e+01  3.224e+00   3.707 0.000211 ***
## CAR_USEPrivate                    -7.873e+02  1.646e+02  -4.783 1.76e-06 ***
## BLUEBOOK                           1.398e-02  2.268e-02   0.617 0.537491
## TIF                               -4.808e+01  1.219e+01  -3.946 8.03e-05 ***
## CAR_TYPEPanel Truck                3.387e+02  2.935e+02   1.154 0.248475
## CAR_TYPEPickup                     3.823e+02  1.734e+02   2.205 0.027478 *
## CAR_TYPESports Car                 1.012e+03  2.190e+02   4.620 3.90e-06 ***
## CAR_TYPEVan                        4.605e+02  2.311e+02   1.993 0.046329 *
## CAR_TYPEz_SUV                      7.310e+02  1.802e+02   4.057 5.01e-05 ***
## RED_CARyes                        -4.856e+01  1.491e+02  -0.326 0.744737
## OLDCLAIM                          -1.060e-02  7.440e-03  -1.424 0.154366
## CLM_FREQ                           1.424e+02  5.508e+01   2.585 0.009743 **
## REVOKEDYes                         5.503e+02  1.736e+02   3.170 0.001532 **
## MVR_PTS                            1.749e+02  2.595e+01   6.740 1.69e-11 ***
## CAR_AGE                           -2.703e+01  1.280e+01  -2.112 0.034723 *
## URBANICITYz_Highly Rural/ Rural -1.662e+03  1.395e+02 -11.914  < 2e-16 ***
## INC_binLow                         6.644e+01  2.156e+02   0.308 0.758011
## INC_binMedium                      1.463e+01  3.234e+02   0.045 0.963920
## INC_binHigh                        4.667e+01  4.893e+02   0.095 0.924009
## INC_binHighest                    -5.418e+01  8.068e+02  -0.067 0.946460
## INC_binNA                          1.600e+01  3.034e+02   0.053 0.957935
## blue_binLow                        4.943e+00  1.990e+02   0.025 0.980183
## blue_binMedium                     5.270e+01  3.289e+02   0.160 0.872685
## blue_binHigh                      -1.750e+02  4.881e+02  -0.359 0.719936
## blue_binHighest                    2.219e+02  7.286e+02   0.305 0.760696
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4546 on 8114 degrees of freedom
## Multiple R-squared:  0.07136,    Adjusted R-squared:  0.0661
## F-statistic: 13.56 on 46 and 8114 DF,  p-value: < 2.2e-16
```

Looking over the summary of this model, we can see that many of the variables do not appear to be very significant. There appears to be value in removing some of these less significant variables and perhaps adding our prediction of rhte flag as an additional one.

#### Model 5 - More significant variables along with the Flag prediction

```
##
```

```
## Call:
## lm(formula = TARGET_AMT ~ . - AGE - HOMEKIDS - YOJ - INCOME -
##     KIDSDRIV - INC_bin - blue_bin - CAR_AGE - RED_CAR - SEX -
##     JOB - CAR_TYPE - TRAVTIME, data = amt_train_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5931  -1446   -615      2 103767
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -3.477e+02  3.651e+02  -0.952  0.34095
## PARENT1Yes                        1.553e+02  1.824e+02   0.851  0.39473
## HOME_VAL                          3.527e-04  5.240e-04   0.673  0.50094
## MSTATUSz_No                       1.415e+02  1.328e+02   1.065  0.28693
## EDUCATIONBachelors               -6.275e+01  1.799e+02  -0.349  0.72726
## EDUCATIONMasters                 -6.260e+01  1.989e+02  -0.315  0.75298
## EDUCATIONPhD                     -7.342e+01  2.461e+02  -0.298  0.76546
## EDUCATIONz_High School           -1.649e+02  1.645e+02  -1.003  0.31601
## CAR_USEPrivate                   -1.268e+02  1.344e+02  -0.944  0.34537
## BLUEBOOK                          2.546e-02  7.023e-03   3.625  0.00029 ***
## TIF                              -2.925e+00  1.289e+01  -0.227  0.82054
## OLDCLAIM                          2.969e-03  7.536e-03   0.394  0.69359
## CLM_FREQ                         -3.629e+01  5.846e+01  -0.621  0.53480
## REVOKEDYes                       -3.064e+02  1.947e+02  -1.573  0.11571
## MVR_PTS                           5.627e+01  2.871e+01   1.960  0.05003 .
## URBANICITYz_Highly Rural/ Rural  3.725e+00  2.006e+02   0.019  0.98519
## Manager                          -1.782e+02  1.740e+02  -1.024  0.30585
## Clerical                          6.645e+00  1.546e+02   0.043  0.96572
## TARGET_FLAG                       5.603e+03  5.533e+02  10.127  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4532 on 8142 degrees of freedom
## Multiple R-squared:  0.0737, Adjusted R-squared:  0.07165
## F-statistic: 35.99 on 18 and 8142 DF,  p-value: < 2.2e-16
```

We can clearly see that our FLAG prediction is by far the most significant of predictor. This will be partially because this variable has already accounted for many of the other variables in the equation. One could argue that this would be double-dipping into variables by accounting for them more than once, but the predicted FLAG variable is actually a compicated combination of many of the variables and should provide valuable new information. The Bluebook vlaue is the other strongly significant variable in this model which makes sense as the value of one of the cars involved in an accident drives the value associated with said accident. We would expect this to be a strong predictor with a positive relationship.

## Choose Model

Though the third binary model doesn't have the lowest AIC value, it's simplicity more than makes up for the slight difference there, so we will use it to predict our FLAG value in the original data. Once that is done we will go with our 5th model (2nd linear regression) to predict the amount associated with an accident. This model has a slightly higher R-squared, but also incorporates our custom FLAG prediction variable, which we believe to be a very good indicator of the amount associated wiht an accident. The relative strength of the Bluebook variable is another argument in favor of this model.

To start, we must calculate the FLAG predictions of our binary model after imputing missing data:

```
##    INDEX TARGET_FLAG TARGET_AMT
## 1      3           0   881.7777
## 2      9           0  1359.8568
## 3     10           0   709.0925
## 4     18           0  1130.0635
## 5     21           0  1025.0292
## 6     30           0  1604.6587
## 7     31           0  2320.4203
## 8     37           0  2917.0971
## 9     39           0   870.8603
## 10    47           0  1785.1647
## 11    60           0   416.4551
## 12    62           1  3314.7809
## 13    63           1  4914.9318
## 14    64           0   380.5161
## 15    68           0     0.0000
## 16    75           1  3932.1687
## 17    76           1  4095.3488
## 18    83           0  1421.0658
## 19    87           1  3016.5605
## 20    92           0  2253.1173
```

We could have customized the amount to display 0 if the flag was predicted to be 0, but since there is a significant level of uncertainty here, we wil leave the amount prediction capped negatiely at 0, but ignoring the FLAG variable prediction.