Critical Thinking Group 4: DATA621 Homework 4

Table of Contents

TEAM Members:

Rajwant Mishra
Priya Shaji
Debabrata Kabiraj
Isabel Ramesar
Sin Ying Wong
Fan Xu

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:

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
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

Deliverables

A write-up of your solutions submitted in PDF format. Assigned prediction (probabilities, classifications) for the evaluation dataset. Use 0.5 threshold.

Data Exploration

The first step we did was to import the data from GitHub, remove the index and look at the structure of the data.

Data	
eval	2141 obs. of 25 variables
train	8161 obs. of 25 variables

We removed special characters then converted variables to numbers for both the Training and Evaluation data.

```
## 'data.frame':    8161 obs. of  25 variables:
##  $ 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     : Factor w/ 6613 levels "","$0","$1,007",..: 5033 6292 1250 1 509 746 1488 315 4765
282 ...
##  $ PARENT1    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
##  $ HOME_VAL   : Factor w/ 5107 levels "","$0","$100,093",..: 2 3259 348 3917 3034 2 1 4167 2 2
...
##  $ MSTATUS    : Factor w/ 2 levels "Yes","z_No": 2 2 1 1 1 2 1 1 2 2 ...
##  $ SEX        : Factor w/ 2 levels "M","z_F": 1 1 2 1 2 2 2 1 2 1 ...
##  $ EDUCATION  : Factor w/ 5 levels "<High School",..: 4 5 5 1 4 2 1 2 2 2 ...
##  $ JOB        : Factor w/ 9 levels "","Clerical",..: 7 9 2 9 3 9 9 9 2 7 ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : Factor w/ 2 levels "Commercial","Private": 2 1 2 2 2 1 2 1 2 1 ...
##  $ BLUEBOOK   : Factor w/ 2789 levels "$1,500","$1,520",..: 434 503 2212 553 802 746 2672 701 135
852 ...
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : Factor w/ 6 levels "Minivan","Panel Truck",..: 1 1 6 1 6 4 6 5 6 5 ...
##  $ RED_CAR    : Factor w/ 2 levels "no","yes": 2 2 1 2 1 1 1 2 1 1 ...
##  $ OLDCLAIM   : Factor w/ 2857 levels "$0","$1,000",..: 1449 1 1311 1 432 1 1 510 1 1 ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 1 1 ...
##  $ 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 : Factor w/ 2 levels "Highly Urban/ Urban",..: 1 1 1 1 1 1 1 1 1 2 ...
```

We then split the training data into a train and test data set.

```{r}
```

set.seed(123)
sample <- sample.split(train,SplitRatio = 0.80)
train <- subset(train, sample == TRUE)
test <- subset(train, sample == FALSE)
```

We removed special characters then converted variables to numbers for both the Training and Evaluation data.

```
train$INCOME<-gsub("[\\$,]", "", train$INCOME)
train$HOME_VAL<-gsub("[\\$,]", "", train$HOME_VAL)
train$BLUEBOOK<-gsub("[\\$,]", "", train$BLUEBOOK)
train$OLDCLAIM<-gsub("[\\$,]", "",train$OLDCLAIM)

eval$INCOME<-gsub("[\\$,]", "", eval$INCOME)
eval$HOME_VAL<-gsub("[\\$,]", "", eval$HOME_VAL)
eval$BLUEBOOK<-gsub("[\\$,]", "", eval$BLUEBOOK)
eval$OLDCLAIM<-gsub("[\\$,]", "",eval$OLDCLAIM)

train$INCOME<-as.numeric(train$INCOME)
train$HOME_VAL<-as.numeric(train$HOME_VAL)
train$BLUEBOOK<-as.numeric(train$BLUEBOOK)
train$OLDCLAIM<-as.numeric(train$OLDCLAIM)

eval$INCOME<-as.numeric(eval$INCOME)
eval$HOME_VAL<-as.numeric(eval$HOME_VAL)
eval$BLUEBOOK<-as.numeric(eval$BLUEBOOK)
eval$OLDCLAIM<-as.numeric(eval$OLDCLAIM)
```

We then ran the summary for ‘Train’ as follows:

```
##   TARGET_FLAG      TARGET_AMT       KIDSDRIV           AGE
##  Min.   :0.000   Min.   :    0   Min.   :0.0000   Min.   :16.00
##  1st Qu.:0.000   1st Qu.:    0   1st Qu.:0.0000   1st Qu.:39.00
```
```

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

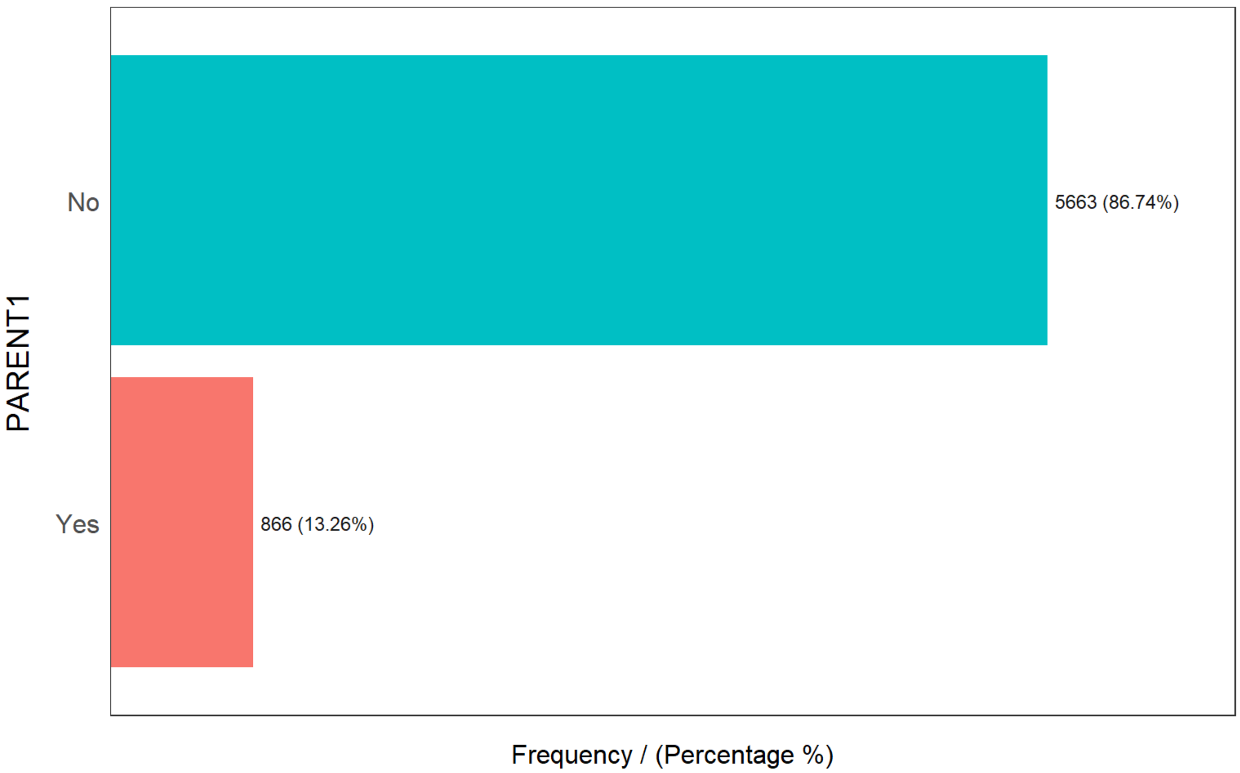
Code

Based on the data summary and bar charts below, there is not a significant amount of NA’s in most variables. There are not real issues with zeros present except variables such as KIDSDRIV, HOMEKIDS, OLDCLAIM and CLM_FREQ. The target variables have the most zeros however we will keep these while removing the rest of the variables with large percentages of zeros. Easily we can see variables with the highest factor levels are most are: drivers that are not single parents, drivers are married, female, finished high school, work blue collar jobs, use the car for leisure, cars are SVU's, not red cars, did not have their license revoked in the past 7 years and most live/work in urban area.

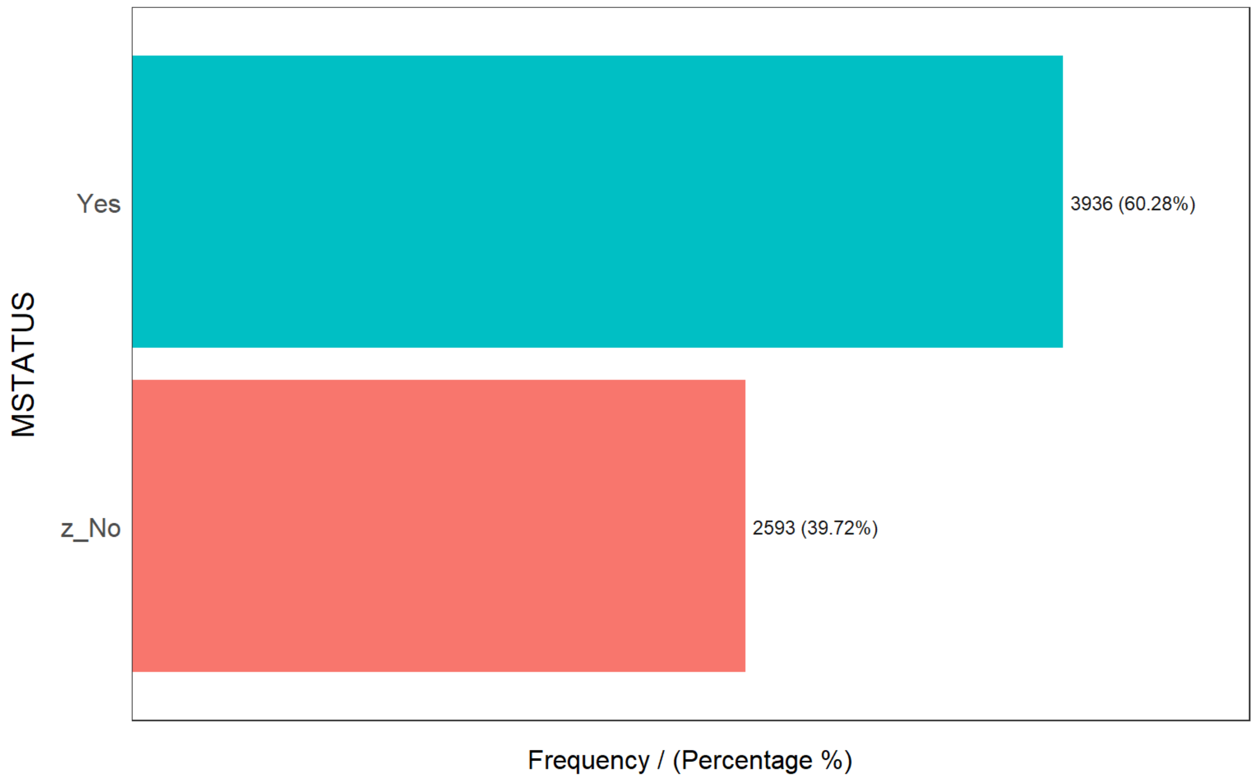
```r
status <- df_status(train, print_results = TRUE)
##        variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1  TARGET_FLAG    4799   73.50    0 0.00     0     0 integer      2
## 2   TARGET_AMT    4799   73.50    0 0.00     0     0 numeric   1595
## 3     KIDSDRIV    5735   87.84    0 0.00     0     0 integer      5
## 4          AGE       0    0.00    6 0.09     0     0 integer     57
## 5     HOMEKIDS    4219   64.62    0 0.00     0     0 integer      6
## 6          YOJ     512    7.84  370 5.67     0     0 integer     20
## 7       INCOME     507    7.77  350 5.36     0     0 numeric   5347
## 8      PARENT1       0    0.00    0 0.00     0     0  factor      2
## 9     HOME_VAL    1852   28.37  358 5.48     0     0 numeric   4121
## 10     MSTATUS       0    0.00    0 0.00     0     0  factor      2
## 11         SEX       0    0.00    0 0.00     0     0  factor      2
## 12   EDUCATION       0    0.00    0 0.00     0     0  factor      5
## 13         JOB       0    0.00    0 0.00     0     0  factor      9
## 14    TRAVTIME       0    0.00    0 0.00     0     0 integer     95
## 15     CAR_USE       0    0.00    0 0.00     0     0  factor      2
## 16    BLUEBOOK       0    0.00    0 0.00     0     0 numeric   2572
## 17         TIF       0    0.00    0 0.00     0     0 integer     23
## 18    CAR_TYPE       0    0.00    0 0.00     0     0  factor      6
## 19     RED_CAR       0    0.00    0 0.00     0     0  factor      2
## 20    OLDCLAIM    4006   61.36    0 0.00     0     0 numeric   2336
## 21    CLM_FREQ    4006   61.36    0 0.00     0     0 integer      6
## 22     REVOKED       0    0.00    0 0.00     0     0  factor      2
## 23     MVR_PTS    2967   45.44    0 0.00     0     0 integer     13
## 24     CAR_AGE       2    0.03  415 6.36     0     0 integer     28
## 25  URBANICITY       0    0.00    0 0.00     0     0  factor      2
```

```r
filter(status, p_zeros > 60)  %>% .$variable
```

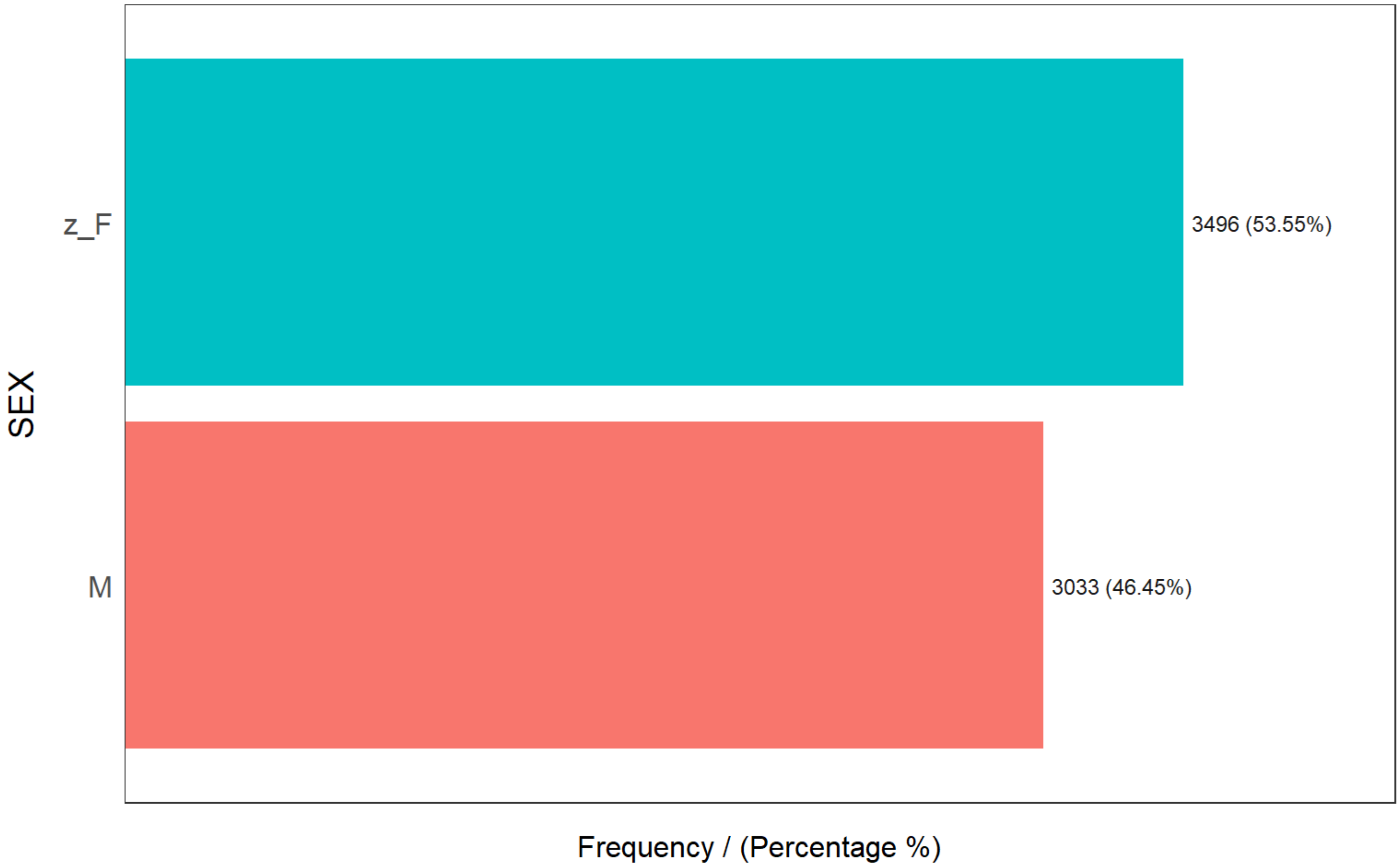
```r
## [1] "TARGET_FLAG" "TARGET_AMT"  "KIDSDRIV"    "HOMEKIDS"    "OLDCLAIM"
## [6] "CLM_FREQ"
freq(train2)
```



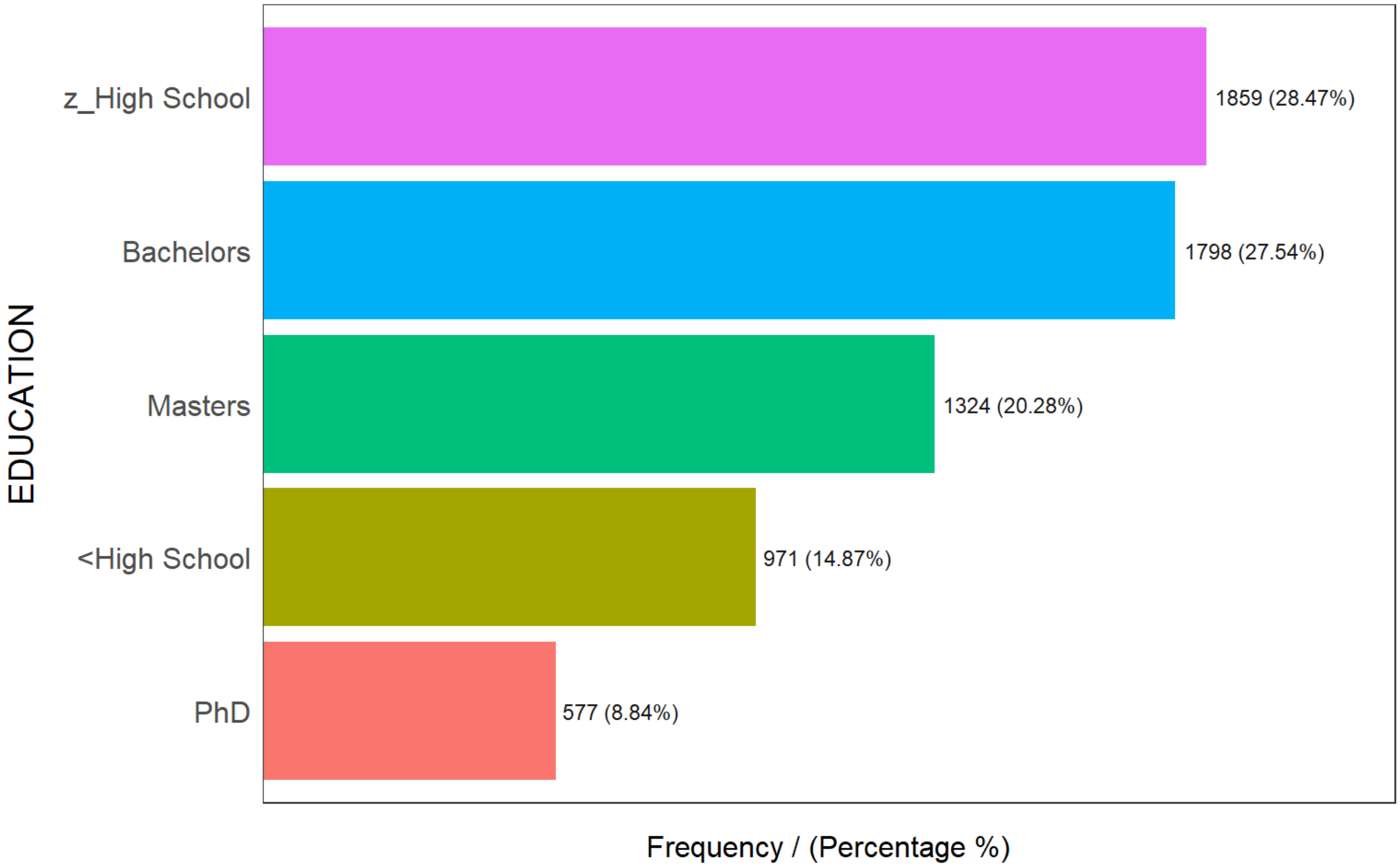

```
##   PARENT1 frequency percentage cumulative_perc
## 1      No      5663      86.74           86.74
## 2     Yes       866      13.26          100.00
```



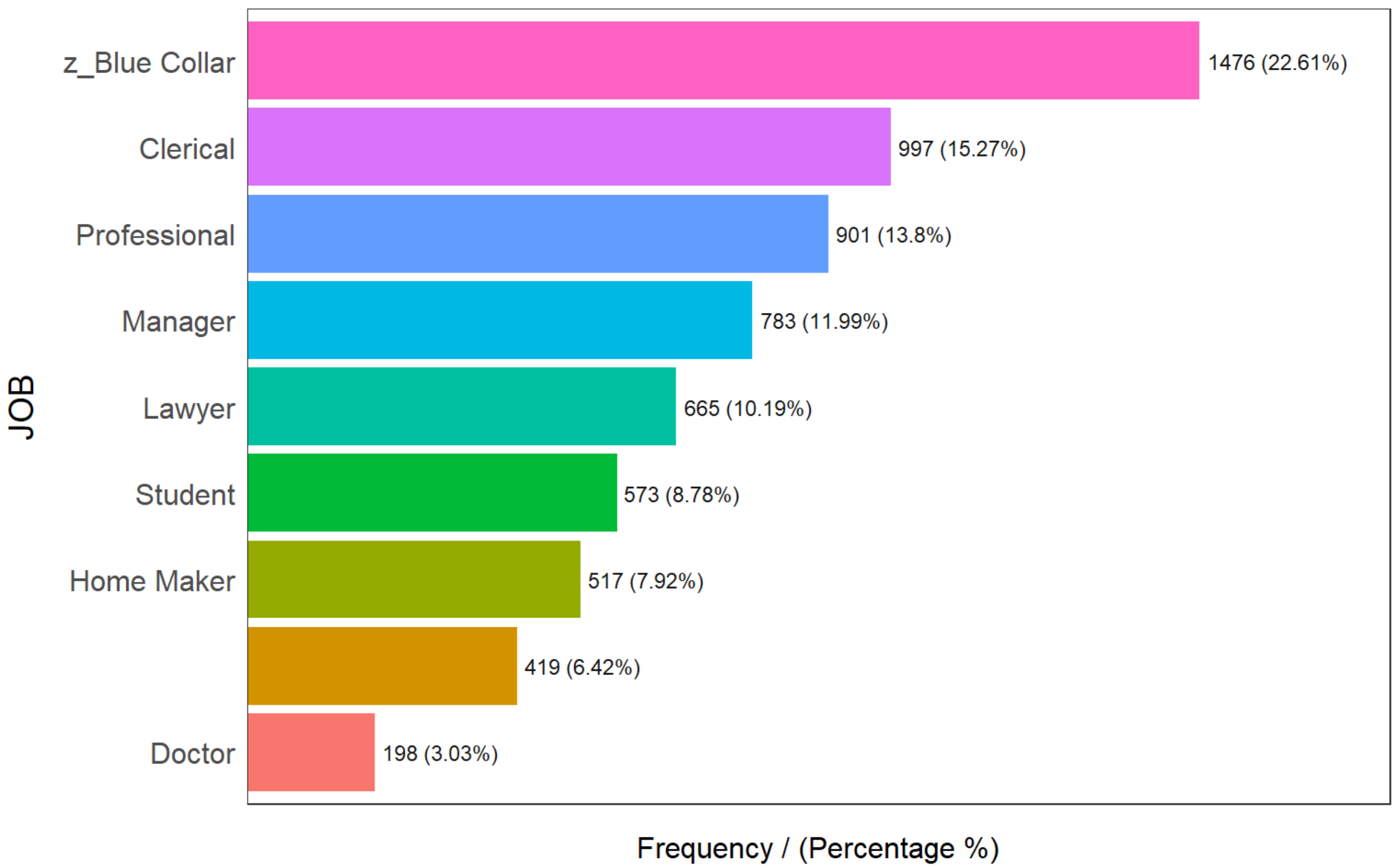

```
##   MSTATUS frequency percentage cumulative_perc
## 1     Yes      3936      60.28           60.28
## 2    z_No      2593      39.72          100.00
```



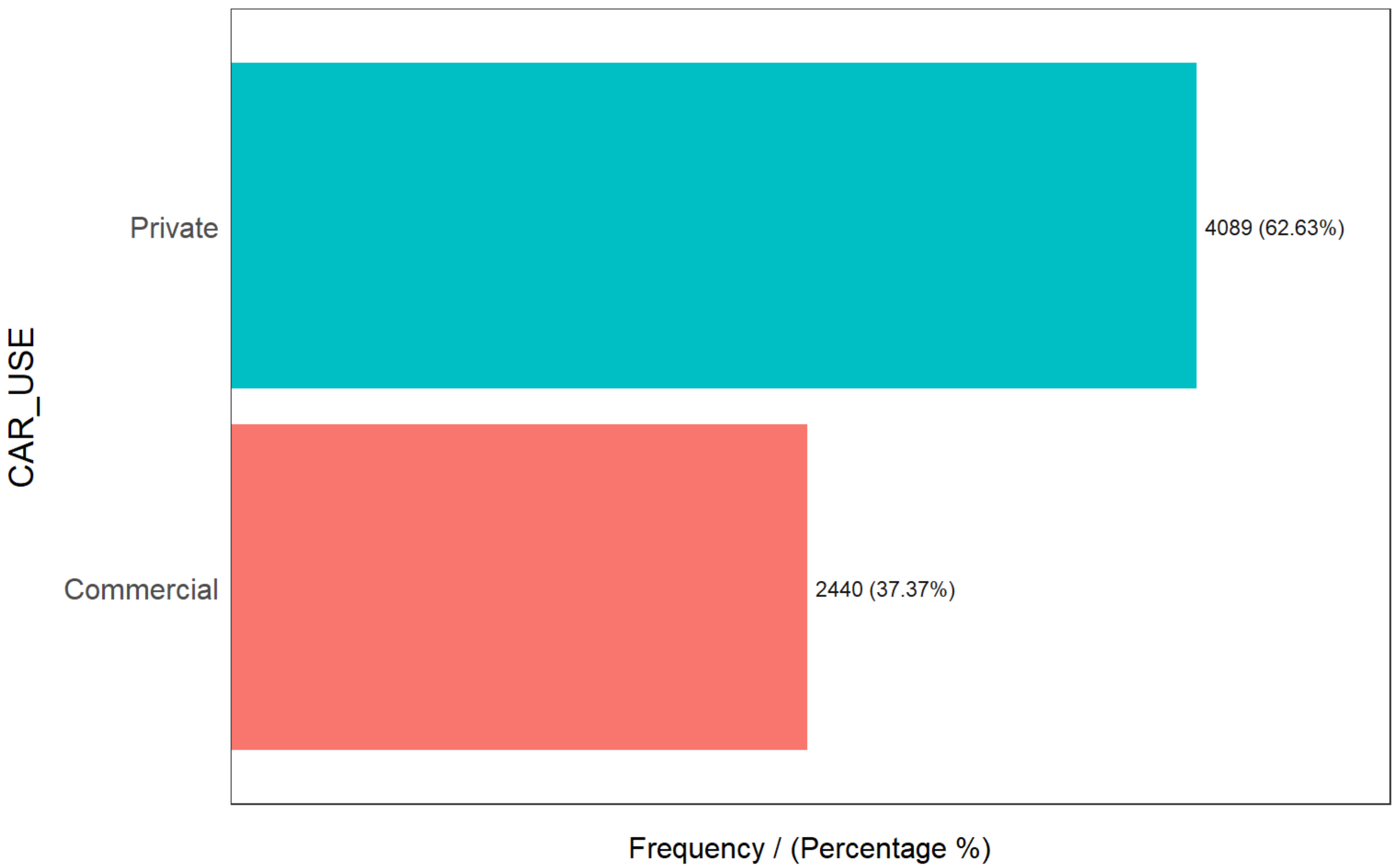

```
##    SEX frequency percentage cumulative_perc
## 1 z_F       3496      53.55           53.55
## 2   M       3033      46.45          100.00
```



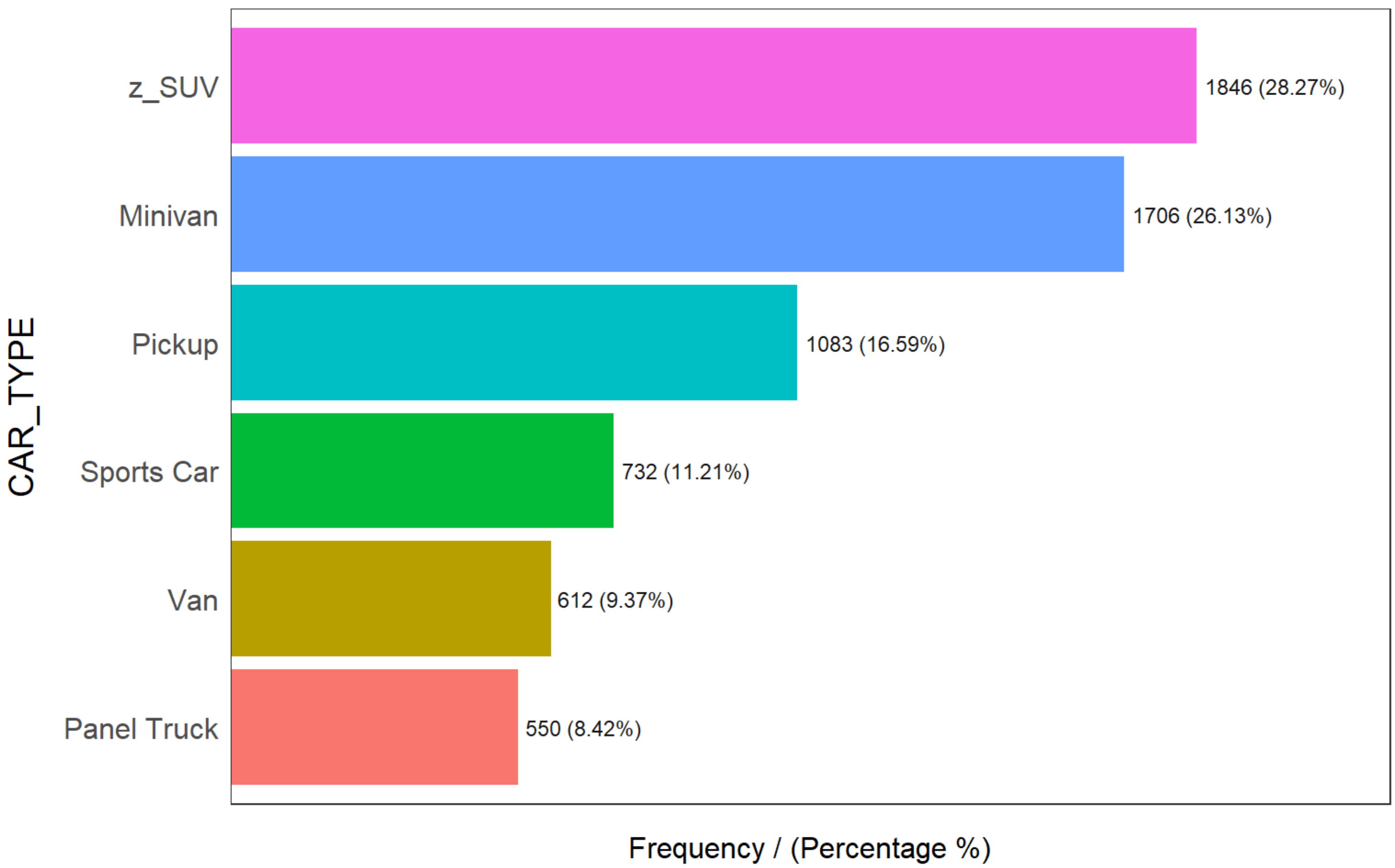

```
##        EDUCATION frequency percentage cumulative_perc
## 1 z_High School      1859      28.47           28.47
## 2     Bachelors      1798      27.54           56.01
## 3       Masters      1324      20.28           76.29
## 4  <High School       971      14.87           91.16
## 5           PhD       577       8.84          100.00
```



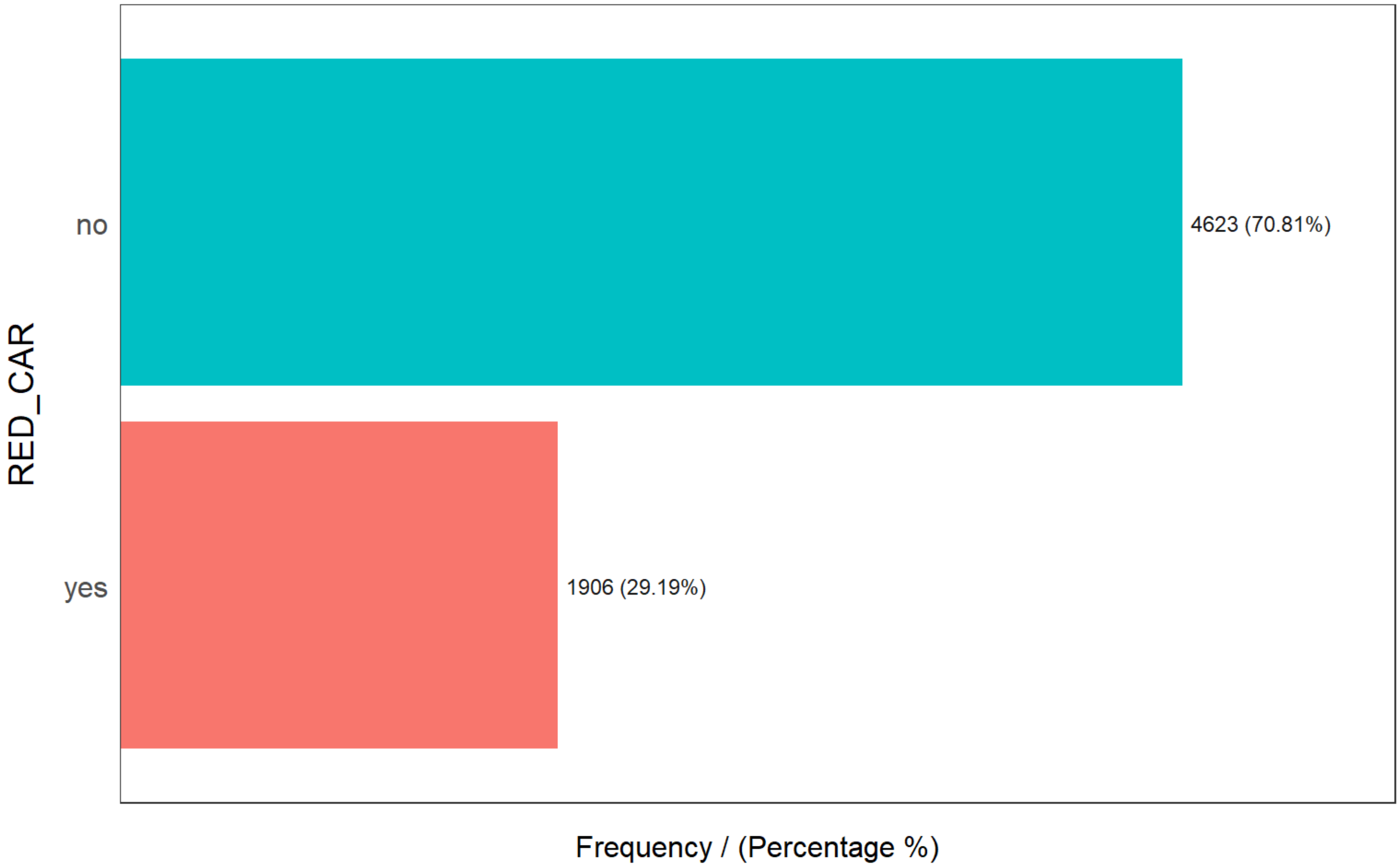

```
##             JOB frequency percentage cumulative_perc
## 1 z_Blue Collar      1476      22.61           22.61
## 2      Clerical       997      15.27           37.88
## 3  Professional       901      13.80           51.68
## 4       Manager       783      11.99           63.67
## 5        Lawyer       665      10.19           73.86
## 6       Student       573       8.78           82.64
## 7     Home Maker      517       7.92           90.56
## 8                     419       6.42           96.98
## 9        Doctor       198       3.03          100.00
```



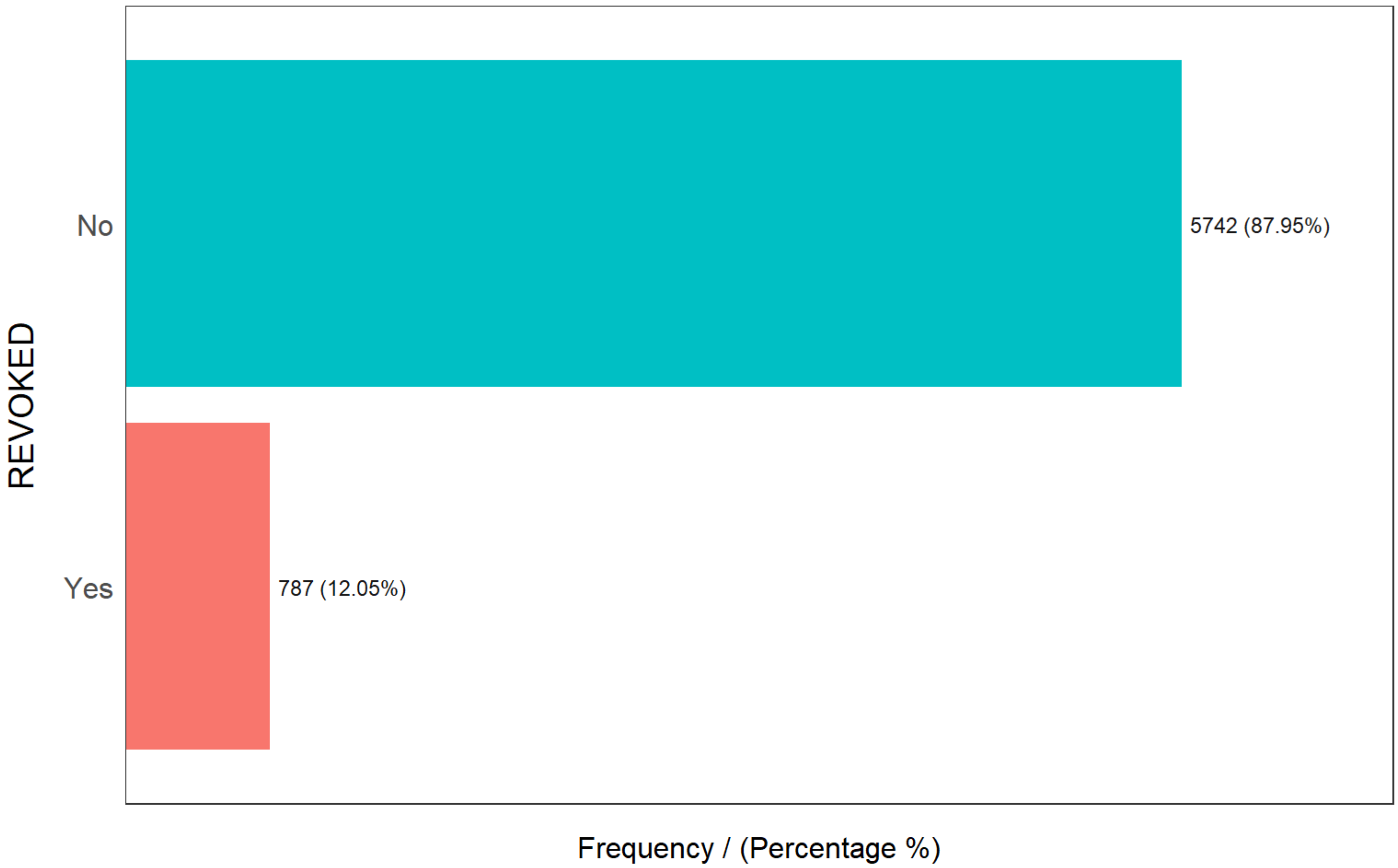

```
##      CAR_USE frequency percentage cumulative_perc
## 1    Private      4089      62.63           62.63
## 2 Commercial      2440      37.37          100.00
```



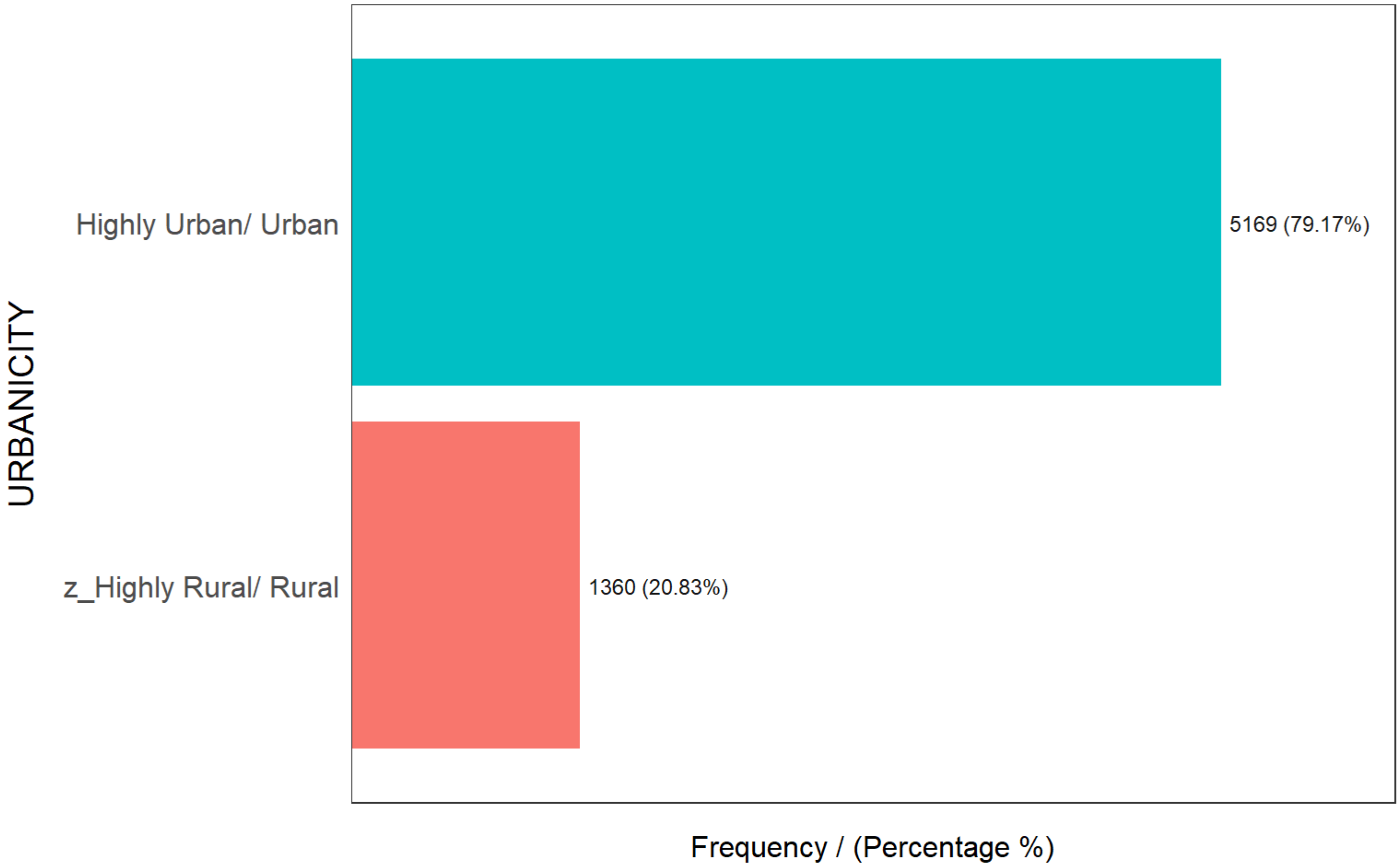

```
##      CAR_TYPE frequency percentage cumulative_perc
## 1       z_SUV      1846      28.27           28.27
## 2     Minivan      1706      26.13           54.40
## 3      Pickup      1083      16.59           70.99
## 4  Sports Car       732      11.21           82.20
## 5         Van       612       9.37           91.57
## 6 Panel Truck       550       8.42          100.00
```




```
##   RED_CAR frequency percentage cumulative_perc
## 1      no      4623      70.81           70.81
## 2     yes      1906      29.19          100.00
```




```
##   REVOKED frequency percentage cumulative_perc
## 1      No      5742      87.95           87.95
## 2     Yes       787      12.05          100.00
```

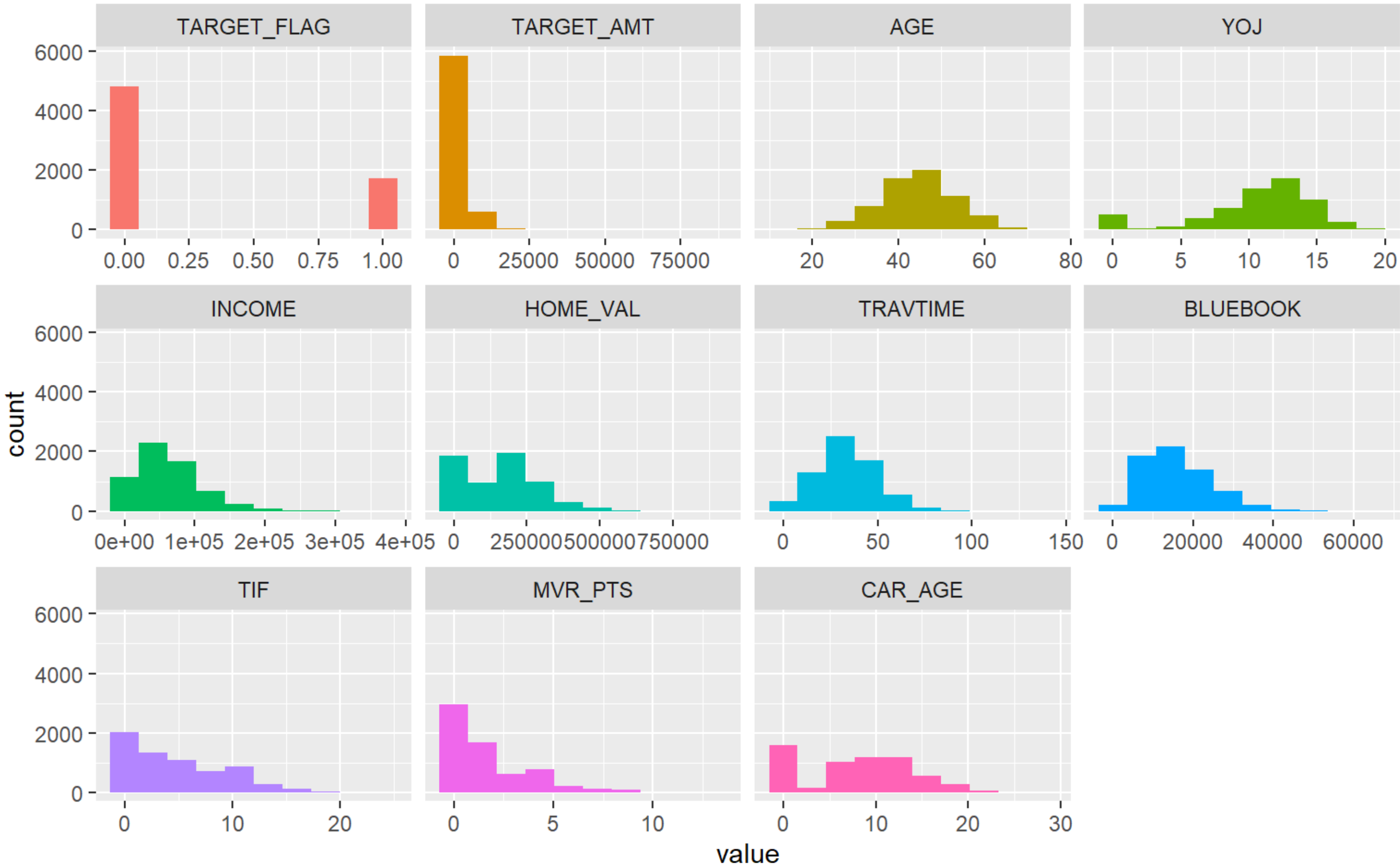



```
##            URBANICITY frequency percentage cumulative_perc
## 1   Highly Urban/ Urban      5169      79.17           79.17
## 2 z_Highly Rural/ Rural      1360      20.83          100.00


## [1] "Variables processed: PARENT1, MSTATUS, SEX, EDUCATION, JOB, CAR_USE, CAR_TYPE, RED_CAR,
REVOKED, URBANICITY"
```

We can determine the skewness and kurtosis of the data. Looking at the distributions of the remaining variables, we can see that the following variables are all skewed right. We can also see variables with skewness and high kurtosis (indicating outliers). 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.

```
plot_num(train2)
```



Below we can see the Mean, Standard deviation, Variation Coefficient and P values for each variable.

```
profiling_num(train2)
```

variable <chr>	mean <dbl>	std_dev <dbl>	variation_coef <dbl>	p_... <dbl>	p_... <dbl>	p_25 <dbl>	p_50 <dbl>	p_75 <dbl>	p_95 <dbl>
TARGET_FLAG	2.649717e-01	4.413519e-01	1.6656570	0	0	0.0	0	1.0	1.0
TARGET_AMT	1.491023e+03	4.480879e+03	3.0052374	0	0	0.0	0	1102.0	6503.2
AGE	4.484884e+01	8.595915e+00	0.1916641	25	31	39.0	45	51.0	59.0
YOJ	1.049083e+01	4.122421e+00	0.3929548	0	0	9.0	11	13.0	15.0
INCOME	6.155210e+04	4.724058e+04	0.7674893	0	0	27645.5	54005	85696.5	151532.3
HOME_VAL	1.541878e+05	1.287673e+05	0.8351327	0	0	0.0	160945	238750.0	372763.5
TRAVTIME	3.357896e+01	1.598681e+01	0.4760961	5	7	23.0	33	44.0	61.0
BLUEBOOK	1.568357e+04	8.414535e+03	0.5365192	1500	4872	9260.0	14440	20800.0	31000.0
TIF	5.357482e+00	4.158576e+00	0.7762184	1	1	1.0	4	7.0	13.0
MVR_PTS	1.694900e+00	2.146455e+00	1.2664198	0	0	0.0	1	3.0	6.0

1-10 of 11 rows | 1-10 of 16 columns   Previous 1 2 Next

Data Preparation

We prepared the data in the previous section which included transformation of variables that contained special characters and removing zeros. The remaining preparation includes imputing missing NA values. We used the Hmisc package. We applied this to AGE, YOJ, INCOME and CAR_AGE. In this section we created a new variable called PTSAGE.

```
train2$AGE<-impute(train2$AGE, median)

train2$YOJ<-impute(train2$YOJ, median)

train2$INCOME<-impute(train2$INCOME, median)

train2$CAR_AGE<-impute(train2$CAR_AGE, median)


eval$AGE<-impute(eval$AGE, median)

eval$YOJ<-impute(eval$YOJ, median)

eval$INCOME<-impute(eval$INCOME, median)

eval$CAR_AGE<-impute(eval$CAR_AGE, median)
```

Create new variable

We created new variable which is PTSAGE = MVR_PTS/AGE. This variable is equal to MVR_PTS/AGE. This variable indicates that if the ratio is higher than one is a driver with more points.

```
train2$PTSAGE <- train2$MVR_PTS/train2$AGE
test$PTSAGE <- test$MVR_PTS/test$AGE

train2 <- dplyr::select(train2, -c(MVR_PTS,AGE))

test <- dplyr::select(test, -c(MVR_PTS,AGE))
```

Build Models

Predicting car crash

All predictors and their corresponding coefficients are within the theoretical effect, except for SEX. The theoretical effect suggest that females are more at risk, but the model has a negative coefficient suggesting the opposite. SEX and YOJ is not statistically significant therefore we will not continue with the variable. 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 removed contradicting and insignificant variables in model 2. The variable we created, PTSAGE also tended to be significant with a corresponding coefficient as well. In the model, we selected the following variables.

```
model1 = glm(TARGET_FLAG ~ YOJ + INCOME + PARENT1 + HOME_VAL + MSTATUS + SEX + EDUCATION + JOB +
TRAVTIME + CAR_USE + TIF + CAR_TYPE + RED_CAR + REVOKED + URBANICITY + PTSAGE,data = train2, family
= 'binomial')
summary(model1)
##
## Call:
## glm(formula = TARGET_FLAG ~ YOJ + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + TIF +
##     CAR_TYPE + RED_CAR + REVOKED + URBANICITY + PTSAGE, 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 ***
## PTSAGE                           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
```

However, we removed variables that deemed insufficient. In this model, all coefficients are in line with their theoretical effects. The only concern was that most job categories are not statistically significant and for the next model, well go ahead and remove these.

```
model2 = glm(TARGET_FLAG ~ INCOME + PARENT1 + HOME_VAL + MSTATUS + EDUCATION + TRAVTIME + CAR_USE +
TIF + CAR_TYPE + REVOKED + URBANICITY + PTSAGE, data = train2, family = 'binomial')
summary(model2)
##
## Call:
## glm(formula = TARGET_FLAG ~ INCOME + PARENT1 + HOME_VAL + MSTATUS +
##     EDUCATION + TRAVTIME + CAR_USE + TIF + CAR_TYPE + REVOKED +
##     URBANICITY + PTSAGE, 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 ***
```

```
## PTSAGE                           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
```

After removing the unnecessary variables, all coefficients fall in line with their theoretical effects.

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

Amount Predicted

A lot of the variables are insignificant, which makes sense. Most of these variables' theoretical effects Are in line with their probabilities influencing accidents and not claim amount. We looked At the claim amount the significant variables. Marital status 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 a higher BLUEBOOK value suggests a higher payout. CAR_AGE is also in line with theoretical effect. Older cars depreciate in cost a majority of the time. In the next model we removed the insignificant predictors except for car type.

```
train2_claims = train2 %>% filter(TARGET_FLAG == 1)
test_claims = test %>% filter(TARGET_FLAG == 1)
linearmodel1 = lm(TARGET_AMT ~ .-TARGET_FLAG, data = train2_claims)
summary(linearmodel1)
##
## 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
## PTSAGE                         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
```

A lot of the variables are insignificant so we will limit the variables in the next model to make it more significant.

The predictors' coefficients all align with theoretical values. The only issue would be car type not having a significant p-value. We removed this in the final model and keep car age along with BLUEBOOK value and Marital Status.

```
linearmodel2 = lm(TARGET_AMT ~ MSTATUS + BLUEBOOK + CAR_AGE, data = train2_claims)
summary(linearmodel2)
##
## 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
```

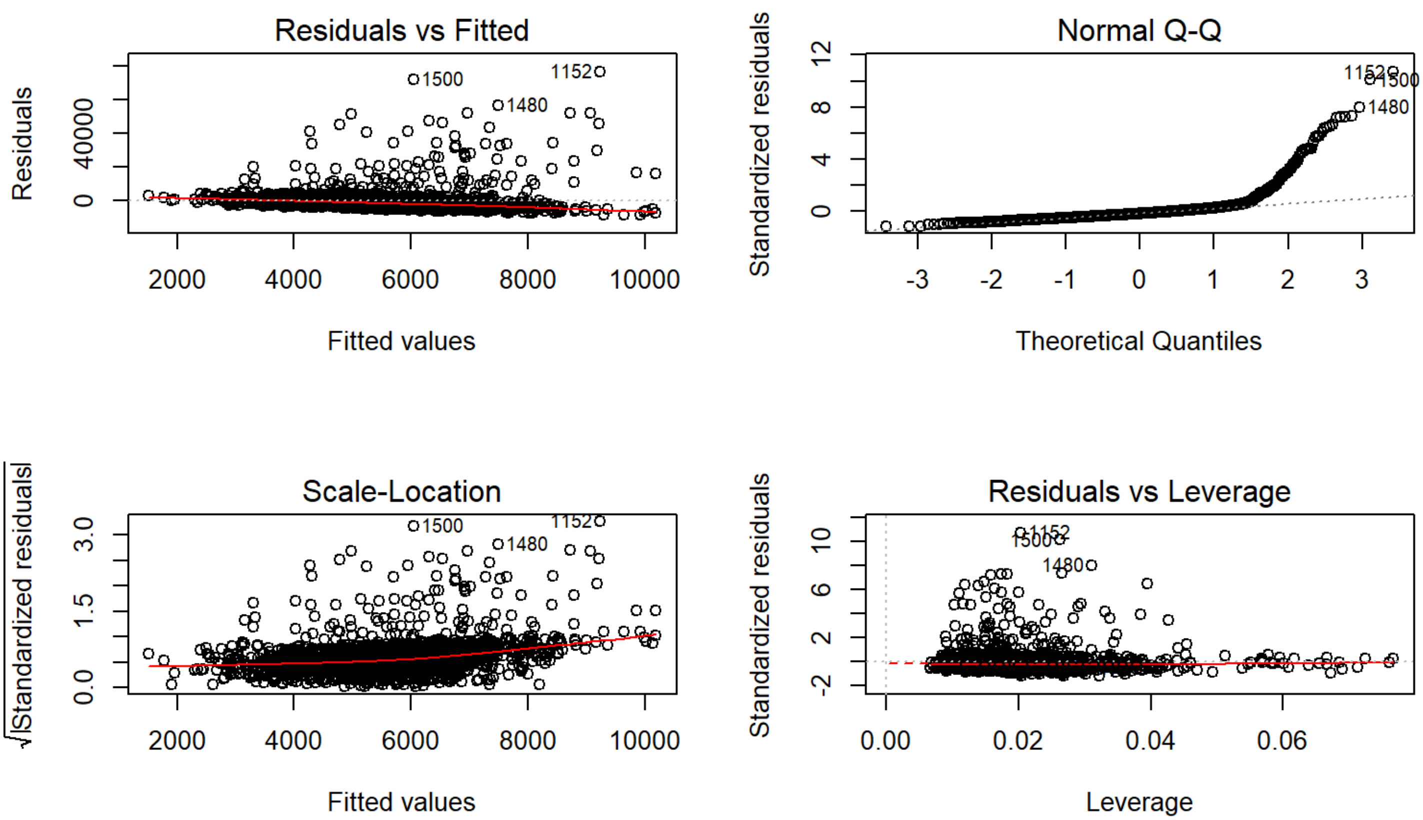
The coefficients are in line with theoretical effects in this model.

Select Models

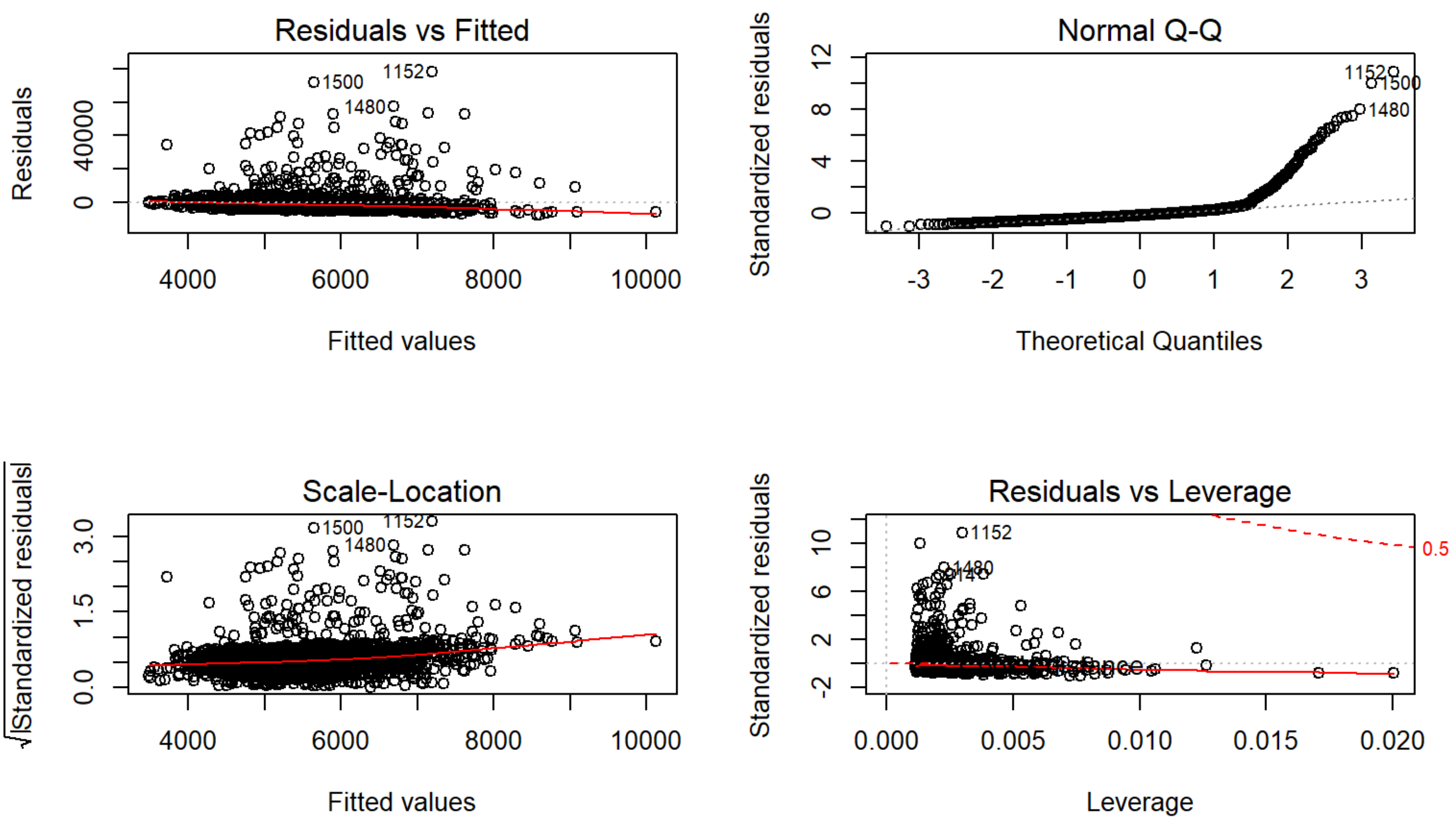
Linear Models

When analyzing the r-squared value for each of the linear models we notice that each performed relatively poor. The r-squared values were 0.03284 and 0.01529 for models 1 and 2 respectively. The f-statistic for all models also appeared to be significant. When viewing the plots of the models the biggest issues in each of the models is the Normal Q-Q plot. The quantile points do not appear to lie on the theoretical normal line. The models are ideally not what we would consider moving forward with however, we proceeded with Model 2 which has a better r-squared and has variables that make sense regarding claim amount and a probability of not crashing.

Model 1


Residuals vs Fitted
Residuals
Fitted values
1500
1152
1480
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Leverage

Model 2




```r
amt = test_claims$TARGET_AMT

summary(test_claims)
```

```
##   TARGET_FLAG   TARGET_AMT          YOJ            INCOME       PARENT1
##  Min.   :1     Min.   :  159.2   Min.   : 0.00   Min.   :     0   No :275
##  1st Qu.:1     1st Qu.: 2632.2   1st Qu.: 9.00   1st Qu.: 17853   Yes: 81
##  Median :1     Median : 4159.5   Median :11.00   Median : 41299
##  Mean   :1     Mean   : 5616.0   Mean   :10.08   Mean   : 49377
##  3rd Qu.:1     3rd Qu.: 5727.7   3rd Qu.:13.00   3rd Qu.: 70128
##  Max.   :1     Max.   :60838.1   Max.   :19.00   Max.   :320127
##                                  NA's   :24      NA's   :16
##     HOME_VAL       MSTATUS      SEX              EDUCATION             JOB
##  Min.   :     0   Yes :172   M  :162   <High School : 58   z_Blue Collar:97
##  1st Qu.:     0   z_No:184   z_F:194   Bachelors    : 84   Clerical     :66
##  Median :101563                        Masters      : 55   Student      :46
##  Mean   :108545                        PhD          : 18   Home Maker   :37
##  3rd Qu.:190761                        z_High School:141   Professional :36
##  Max.   :750455                                                         :25
##  NA's   :18                                             (Other)         :49
```

```
##     TRAVTIME           CAR_USE       BLUEBOOK          TIF
##  Min.   : 5.00   Commercial:178   Min.   : 1500   Min.   : 1.000
##  1st Qu.:24.00   Private   :178   1st Qu.: 7338   1st Qu.: 1.000
##  Median :35.00                    Median :12245   Median : 4.000
##  Mean   :35.17                    Mean   :14643   Mean   : 4.747
##  3rd Qu.:46.00                    3rd Qu.:20215   3rd Qu.: 7.000
##  Max.   :81.00                    Max.   :62240   Max.   :18.000
##
##         CAR_TYPE    RED_CAR    REVOKED      CAR_AGE
##  Minivan    : 65   no :254   No :297   Min.   : 1.000
##  Panel Truck: 35   yes:102   Yes: 59   1st Qu.: 1.000
##  Pickup     : 76                       Median : 7.000
##  Sports Car : 49                       Mean   : 7.061
##  Van        : 29                       3rd Qu.:10.750
##  z_SUV      :102                       Max.   :22.000
##                                        NA's   :30
##                    URBANICITY      PTSAGE
##  Highly Urban/ Urban  :343   Min.   :0.00000
##  z_Highly Rural/ Rural: 13   1st Qu.:0.00000
##                              Median :0.04651
##                              Mean   :0.06580
##                              3rd Qu.:0.10217
##                              Max.   :0.42308
##
```

```
as.matrix(c(mean((amt - predict.lm(linearmodel1, newdata = test_claims))^2, na.rm = TRUE), mean((amt
- predict.lm(linearmodel2, newdata = test_claims))^2, na.rm = TRUE), mean((amt -
predict.lm(linearmodel2, newdata = test_claims))^2, na.rm = TRUE)))
##          [,1]
## [1,] 45889850
## [2,] 47757957
## [3,] 47757957
```

Logit Models

To decide on which model should be selected, we used ANOVA and McFaddens R^2. When using ANOVA, we looked for the widest gap between the null and residual deviance. Below is the ANOVA for the original model with all variables:

Model 1

	Df <int>	Deviance <dbl>	Resid. Df <int>	Resid. Dev <dbl>	Pr(>Chi) <dbl>
NULL	NA	NA	6170	7129.644	NA
YOJ	1	29.14213465	6169	7100.502	6.725823e-08
INCOME	1	98.00828250	6168	7002.494	4.166363e-23
PARENT1	1	133.87291895	6167	6868.621	5.824694e-31
HOME_VAL	1	51.83590734	6166	6816.785	6.033820e-13
MSTATUS	1	9.14597532	6165	6807.639	2.492657e-03
SEX	1	0.07913537	6164	6807.560	7.784726e-01
EDUCATION	4	48.58709140	6160	6758.973	7.119621e-10
JOB	8	95.41559296	6152	6663.557	3.681017e-17
TRAVTIME	1	11.45353540	6151	6652.103	7.135811e-04

1-10 of 17 rows Previous 1 2 Next

Model 2

	Df <int>	Deviance <dbl>	Resid. Df <int>	Resid. Dev <dbl>	Pr(>Chi) <dbl>
NULL	NA	NA	6170	7129.644	NA
INCOME	1	122.547766	6169	7007.096	1.751474e-28
PARENT1	1	135.188140	6168	6871.908	3.003199e-31
HOME_VAL	1	54.600254	6167	6817.308	1.477165e-13
MSTATUS	1	9.462215	6166	6807.846	2.097476e-03
EDUCATION	4	47.645200	6162	6760.200	1.118983e-09
TRAVTIME	1	14.901210	6161	6745.299	1.132903e-04
CAR_USE	1	103.782825	6160	6641.516	2.257537e-24
TIF	1	41.372012	6159	6600.144	1.258464e-10
CAR_TYPE	5	100.318636	6154	6499.826	4.527909e-20

1-10 of 13 rows Previous 1 2 Next

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

```
fitted.results = predict(model2, 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"
```

Make Predictions

Predictions can be found in the following:

https://github.com/Rajwantmishra/DATA621_CR4/blob/master/HW4/linear_model_eval.csv

https://github.com/Rajwantmishra/DATA621_CR4/blob/master/HW4/logistic_model_eval.csv

Appendix

https://github.com/Rajwantmishra/DATA621_CR4/blob/master/HW4/Homework4_Final.Rmd

Thank you