Critical Thinking Group 4: DATA621 Homework 4

Table of Contents

[TEAM Members: 2](#_Toc38761704)

[**Overview** 2](#_Toc38761705)

[**Deliverables** 3](#_Toc38761706)

[**Data Exploration** 3](#_Toc38761707)

[**Data Preparation** 18](#_Toc38761708)

[**Build Models** 18](#_Toc38761709)

[**Select Models** 23](#_Toc38761710)

[**Make Predictions** 28](#_Toc38761711)

[**Appendix** 28](#_Toc38761712)

## TEAM Members:

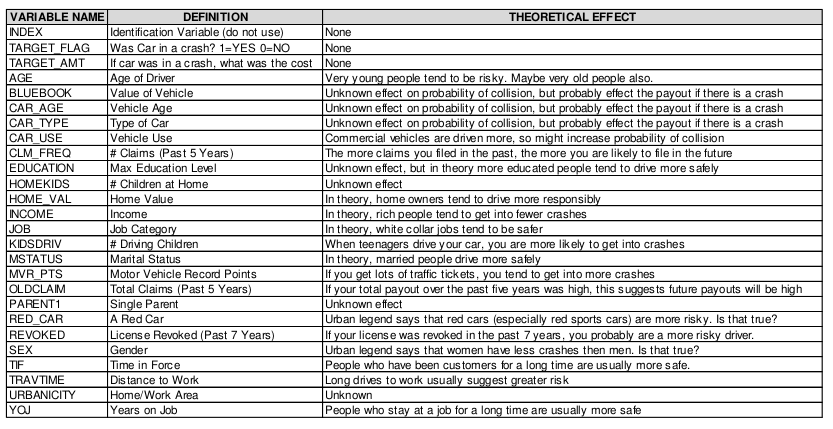
Rajwant Mishra  
Priya Shaji  
Debabrata Kabiraj  
Isabel Ramesar  
Sin Ying Wong  
Fan Xu

-------------------------------------------------

**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:

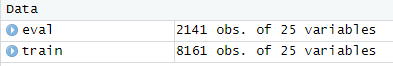


**Deliverables**

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.



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.

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

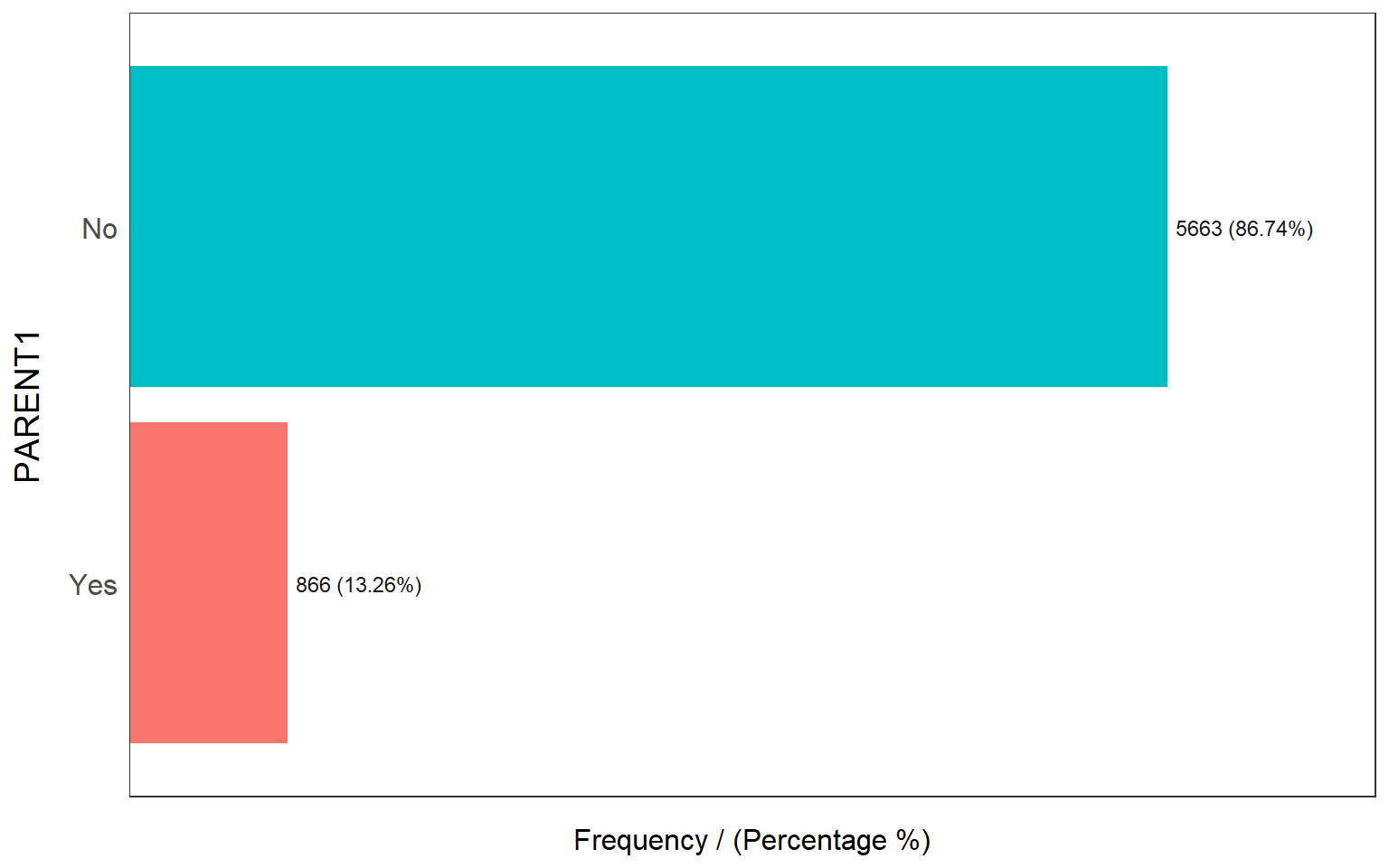
## 25 URBANICITY 0 0.00 0 0.00 0 0 factor 2

**filter**(status, p\_zeros **>** 60) **%>%** .**$**variable

## [1] "TARGET\_FLAG" "TARGET\_AMT" "KIDSDRIV" "HOMEKIDS" "OLDCLAIM"

## [6] "CLM\_FREQ"

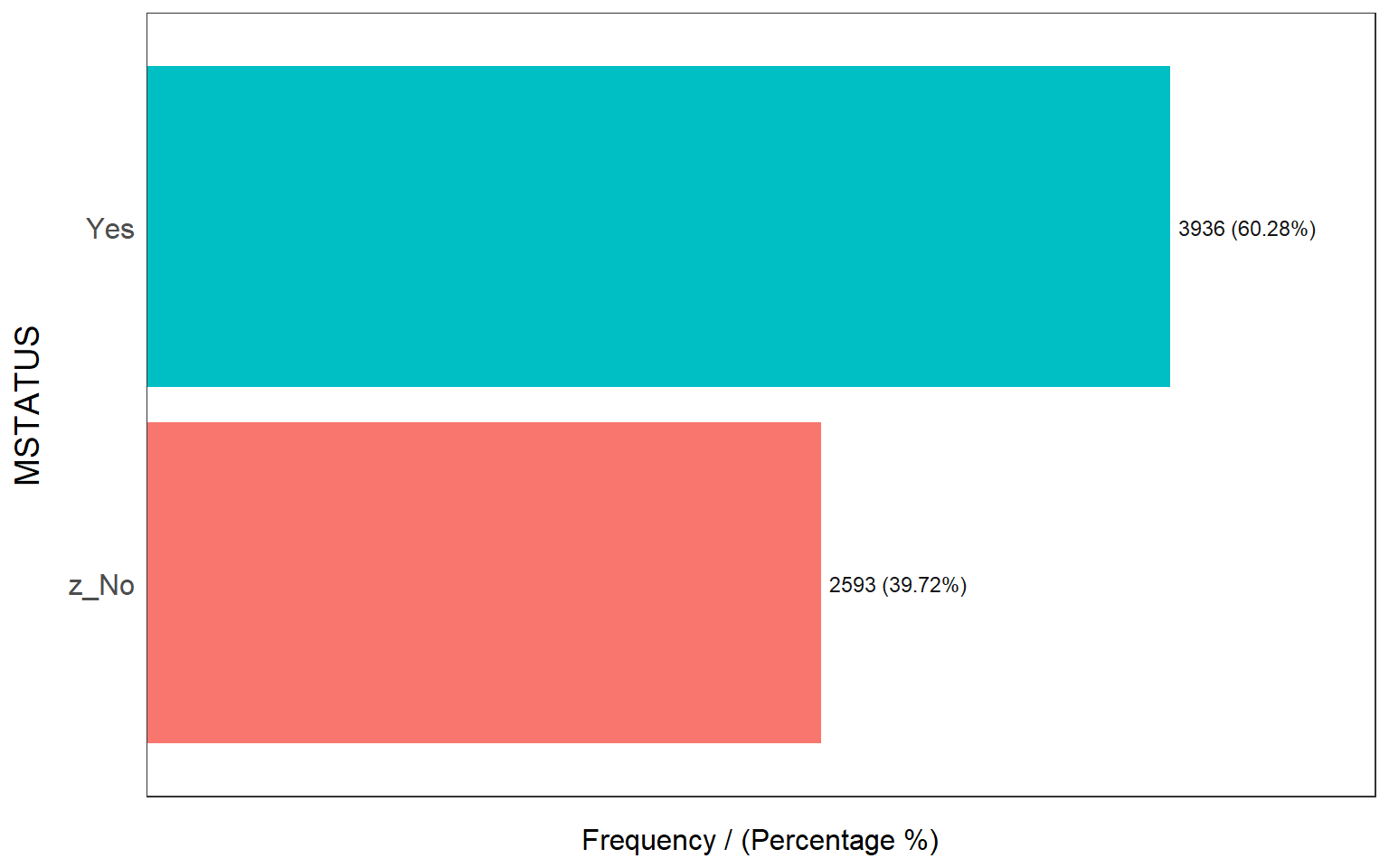
**freq**(train2)



## PARENT1 frequency percentage cumulative\_perc

## 1 No 5663 86.74 86.74

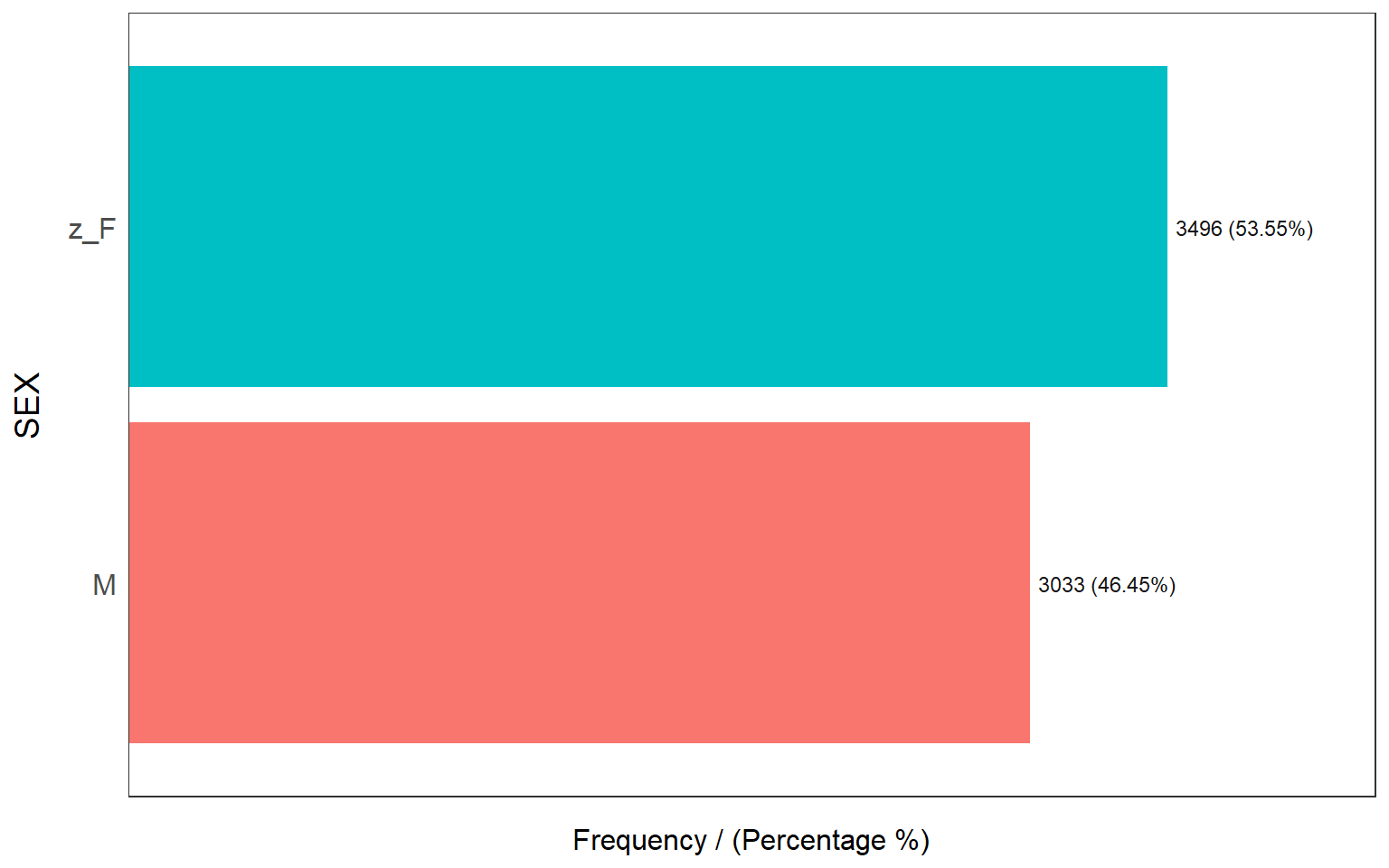
## 2 Yes 866 13.26 100.00



## MSTATUS frequency percentage cumulative\_perc

## 1 Yes 3936 60.28 60.28

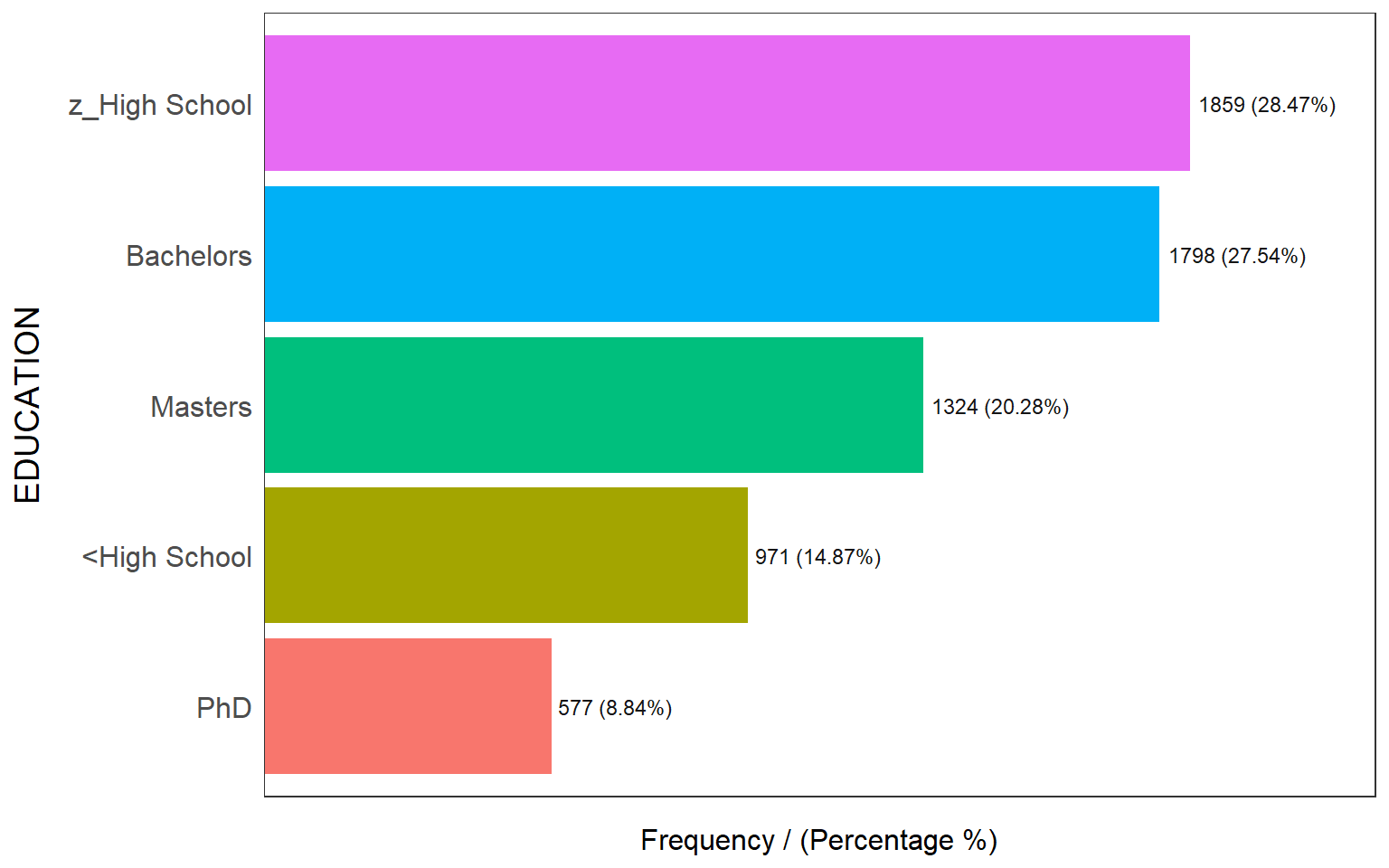
## 2 z\_No 2593 39.72 100.00



## 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

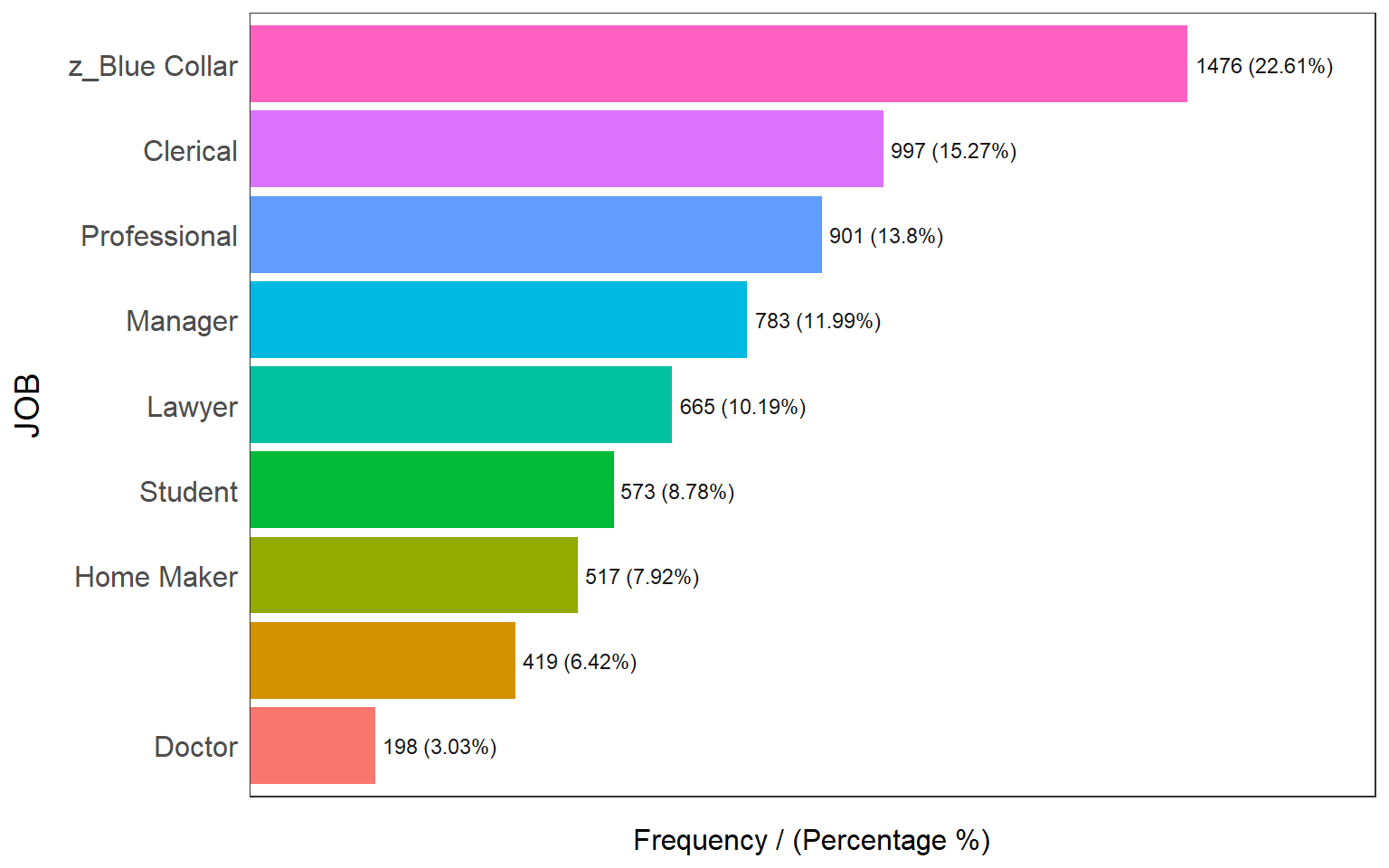
## 1 z\_High School 1859 28.47 28.47

## 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

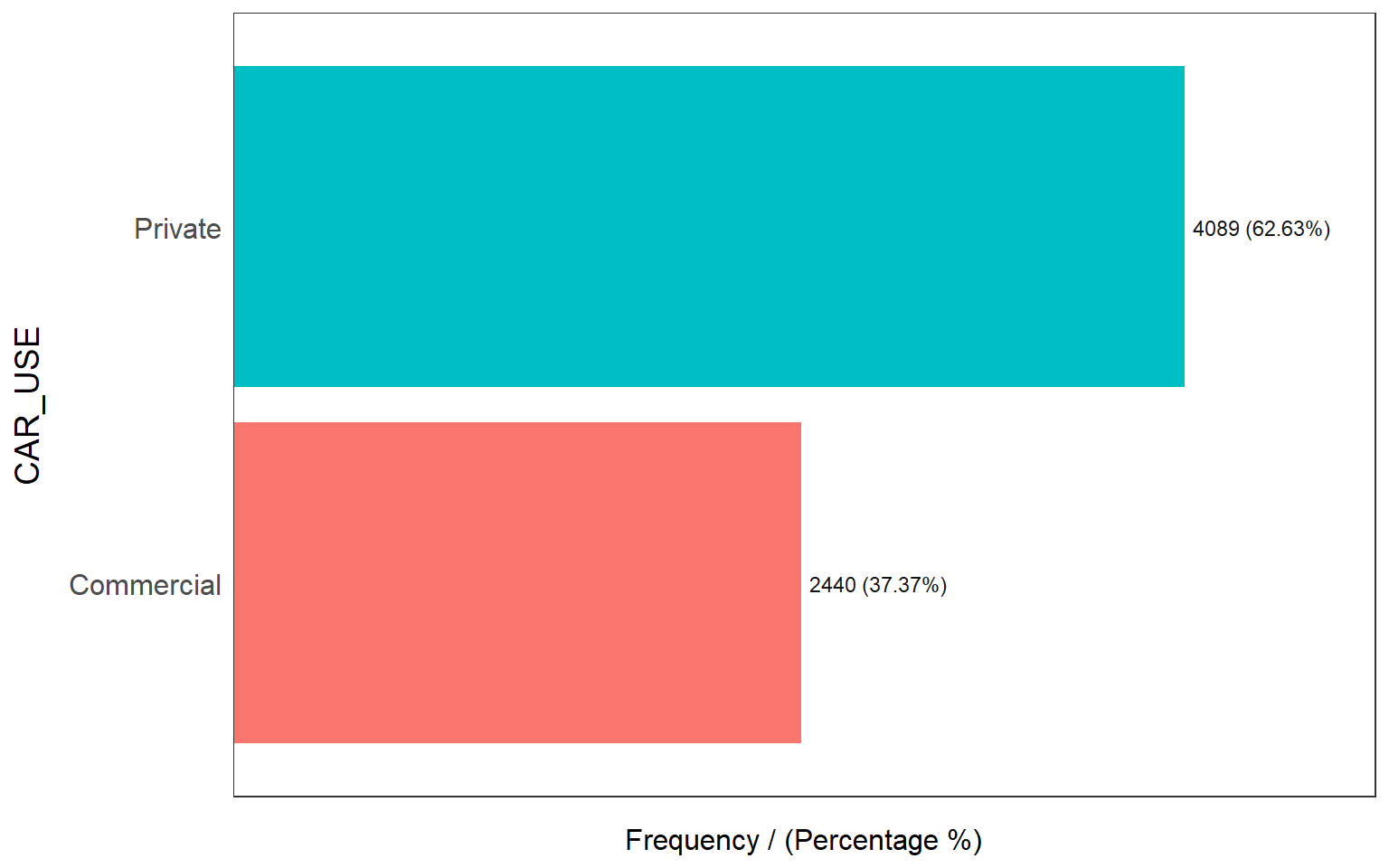
## 5 Lawyer 665 10.19 73.86

## 6 Student 573 8.78 82.64

## 7 Home Maker 517 7.92 90.56

## 8 419 6.42 96.98

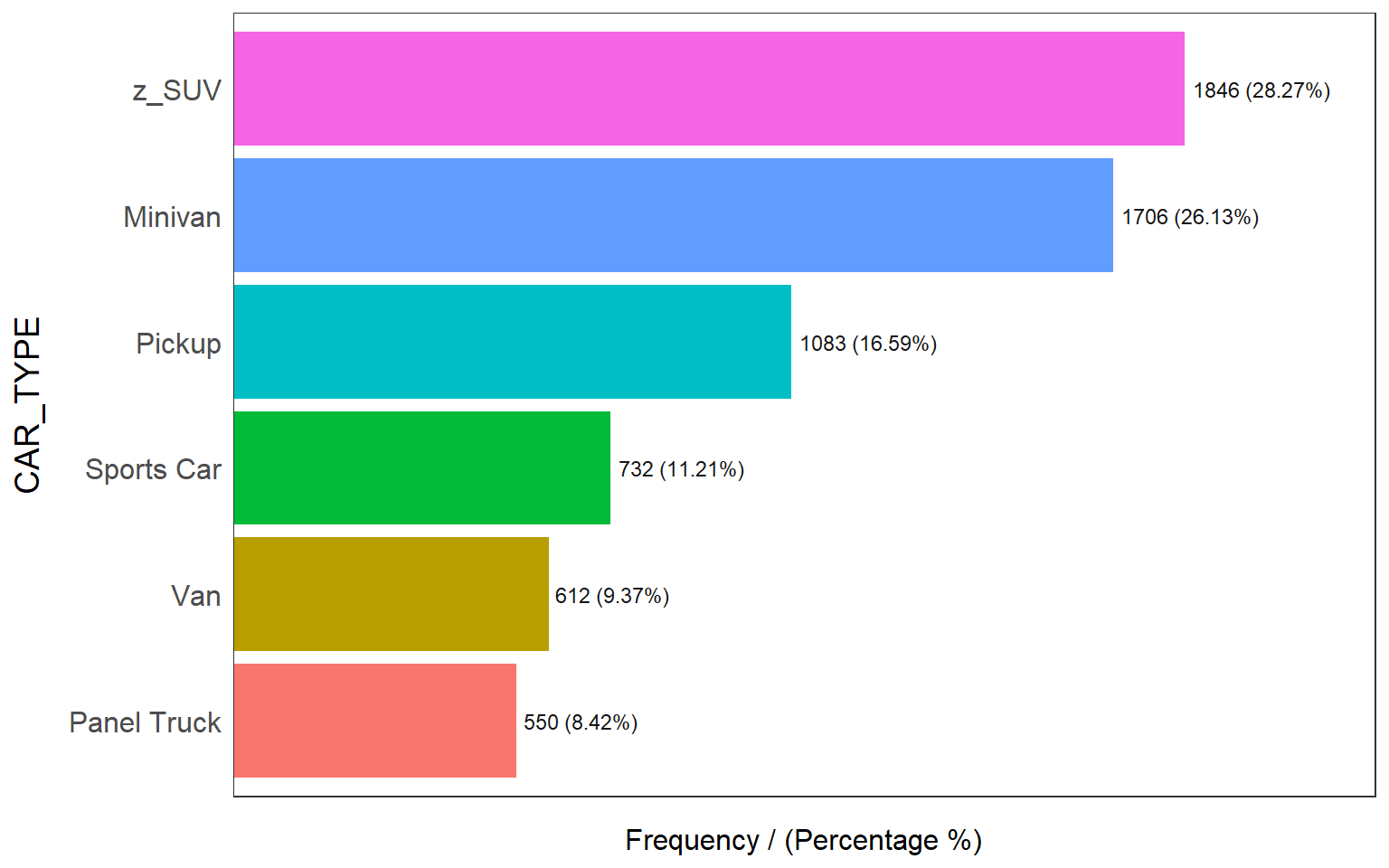
## 9 Doctor 198 3.03 100.00



## 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

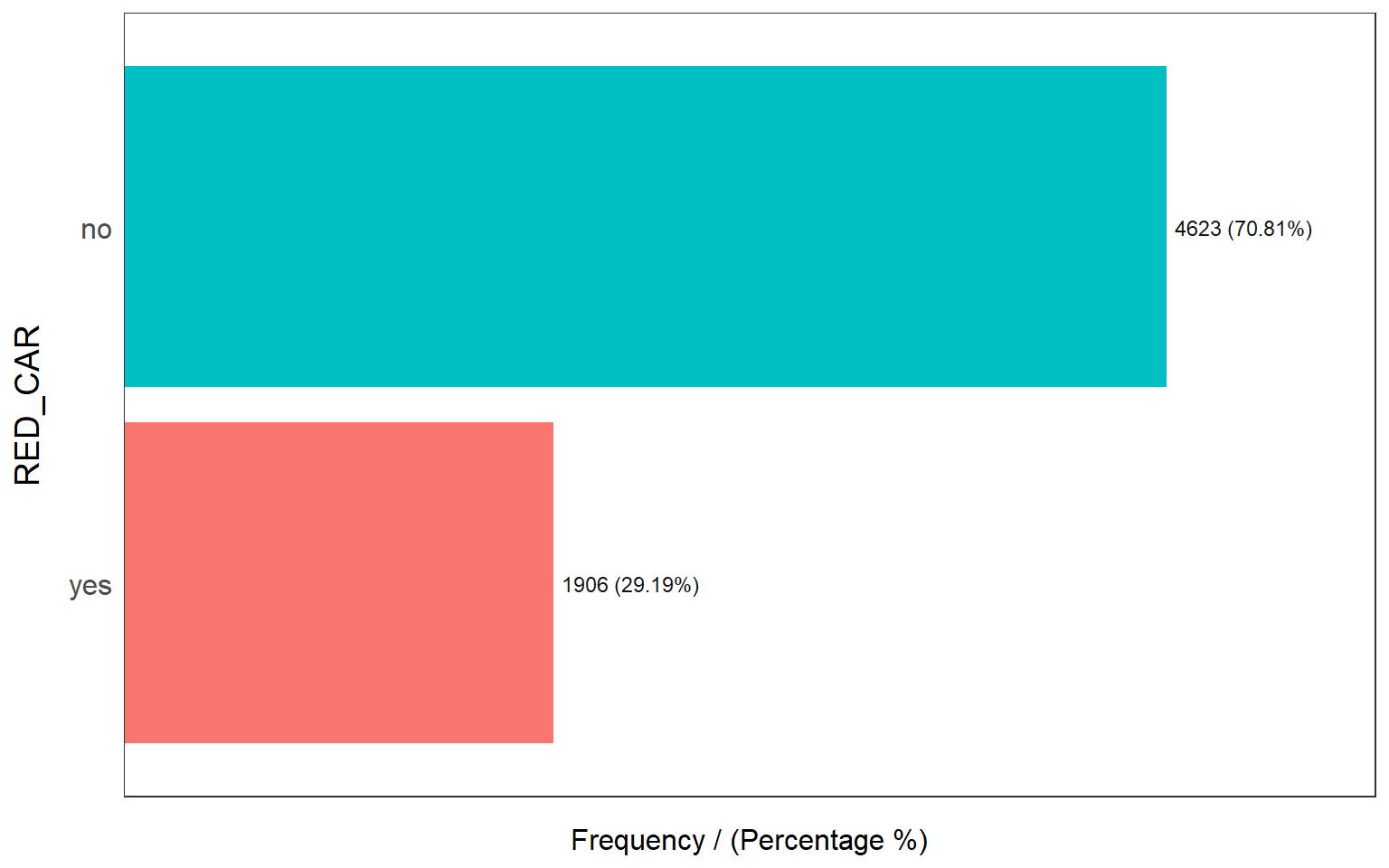
## 2 Minivan 1706 26.13 54.40

## 3 Pickup 1083 16.59 70.99

## 4 Sports Car 732 11.21 82.20

## 5 Van 612 9.37 91.57

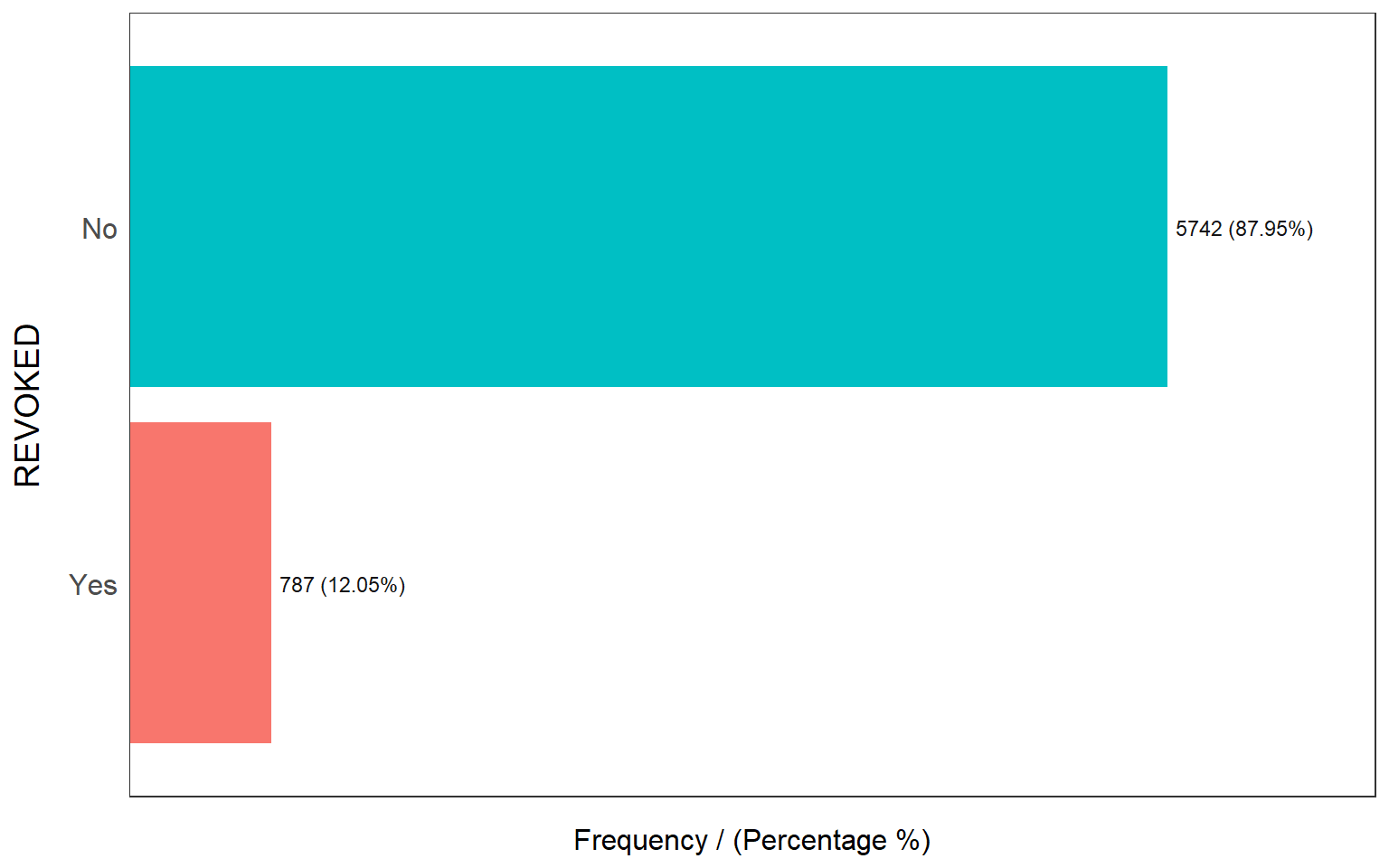
## 6 Panel Truck 550 8.42 100.00



## RED\_CAR frequency percentage cumulative\_perc

## 1 no 4623 70.81 70.81

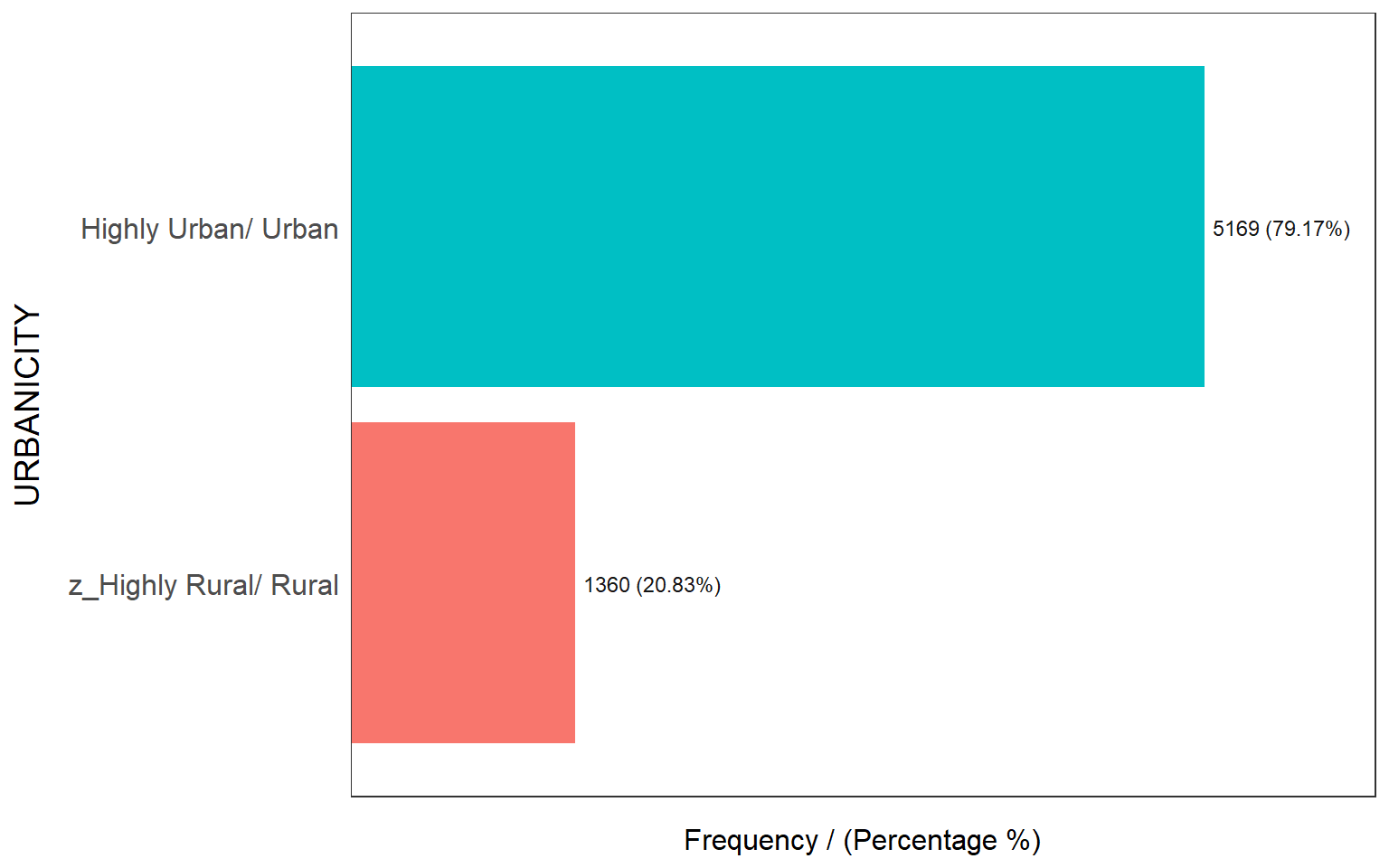
## 2 yes 1906 29.19 100.00



## REVOKED frequency percentage cumulative\_perc

## 1 No 5742 87.95 87.95

## 2 Yes 787 12.05 100.00



## URBANICITY frequency percentage cumulative\_perc

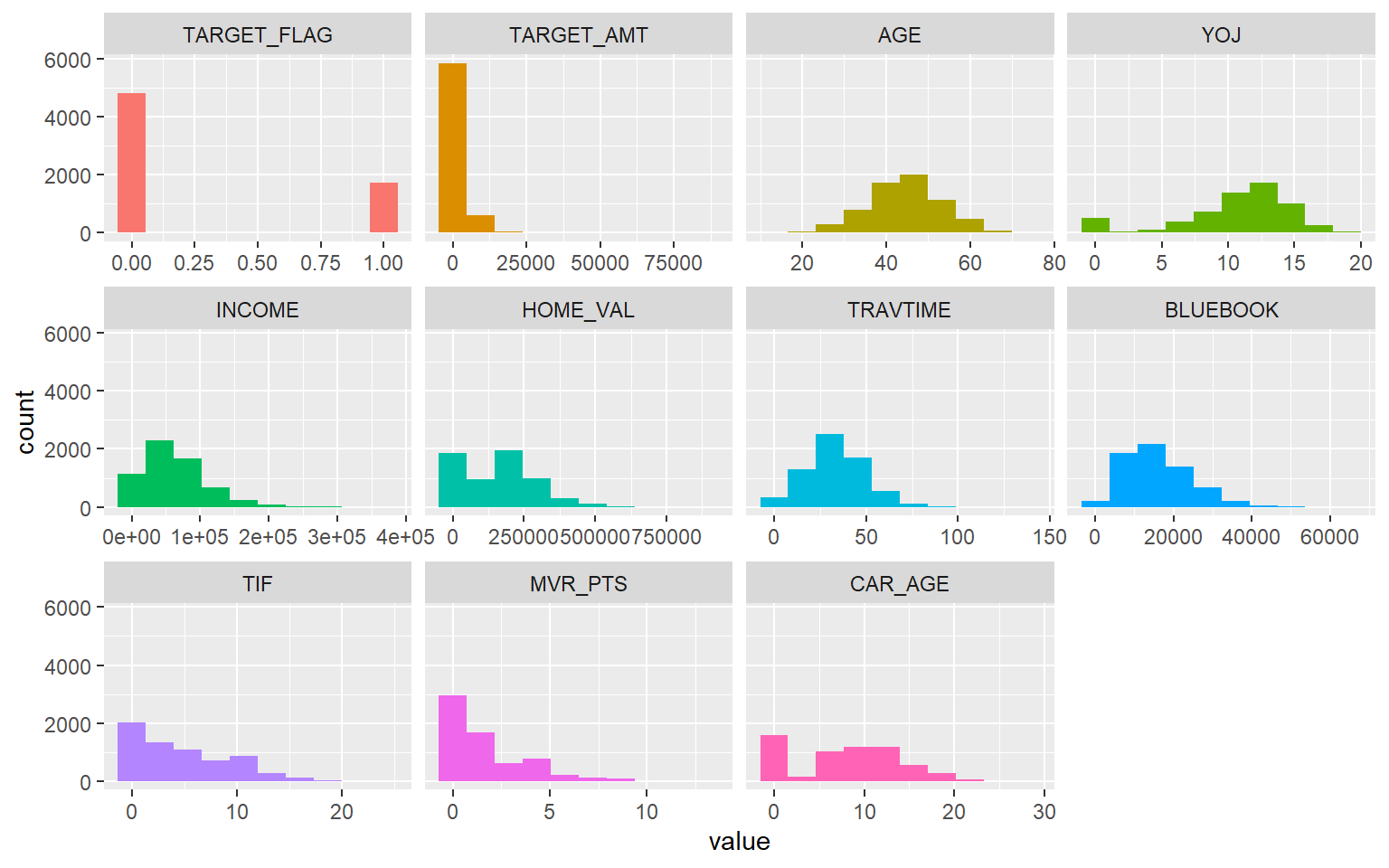
## 1 Highly Urban/ Urban 5169 79.17 79.17

## 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"

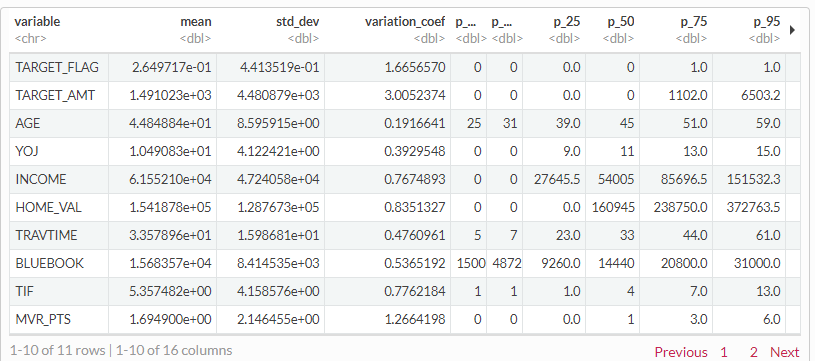
We can determine the skewness and kurtosis of the data. Looking at the distributions of the remaining variables, we can see that the following variables are all skewed right. We can also see variables with skewness and high kurtosis (indicating outliers). As seen before visually, we can verify here that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



**Data Preparati****on**

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

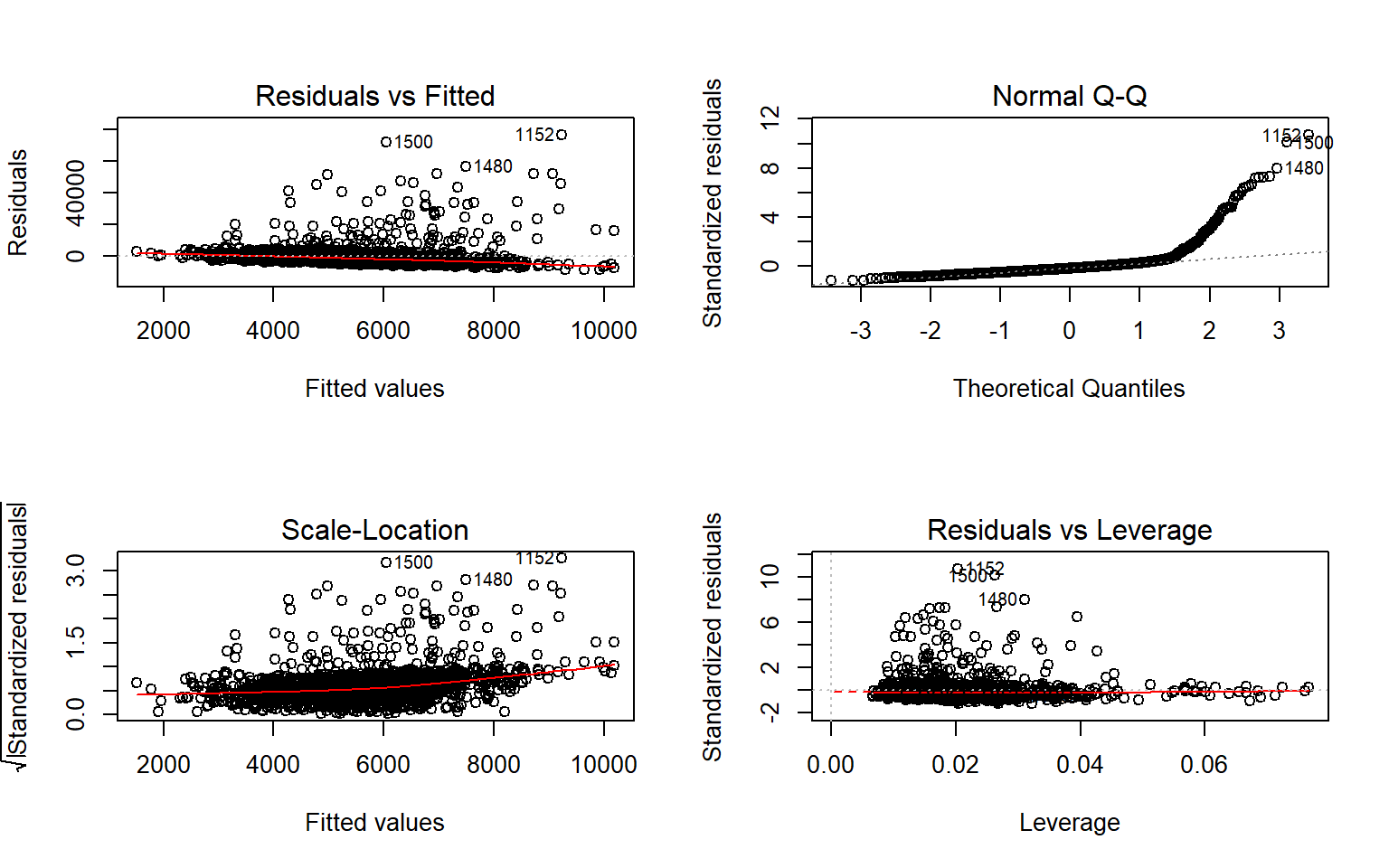
When analyzing the r-squared value for each of the linear models we notice that each performed

relatively poor. The r-squared values were 0.03284 and 0.01529 for models 1 and 2 respectively.

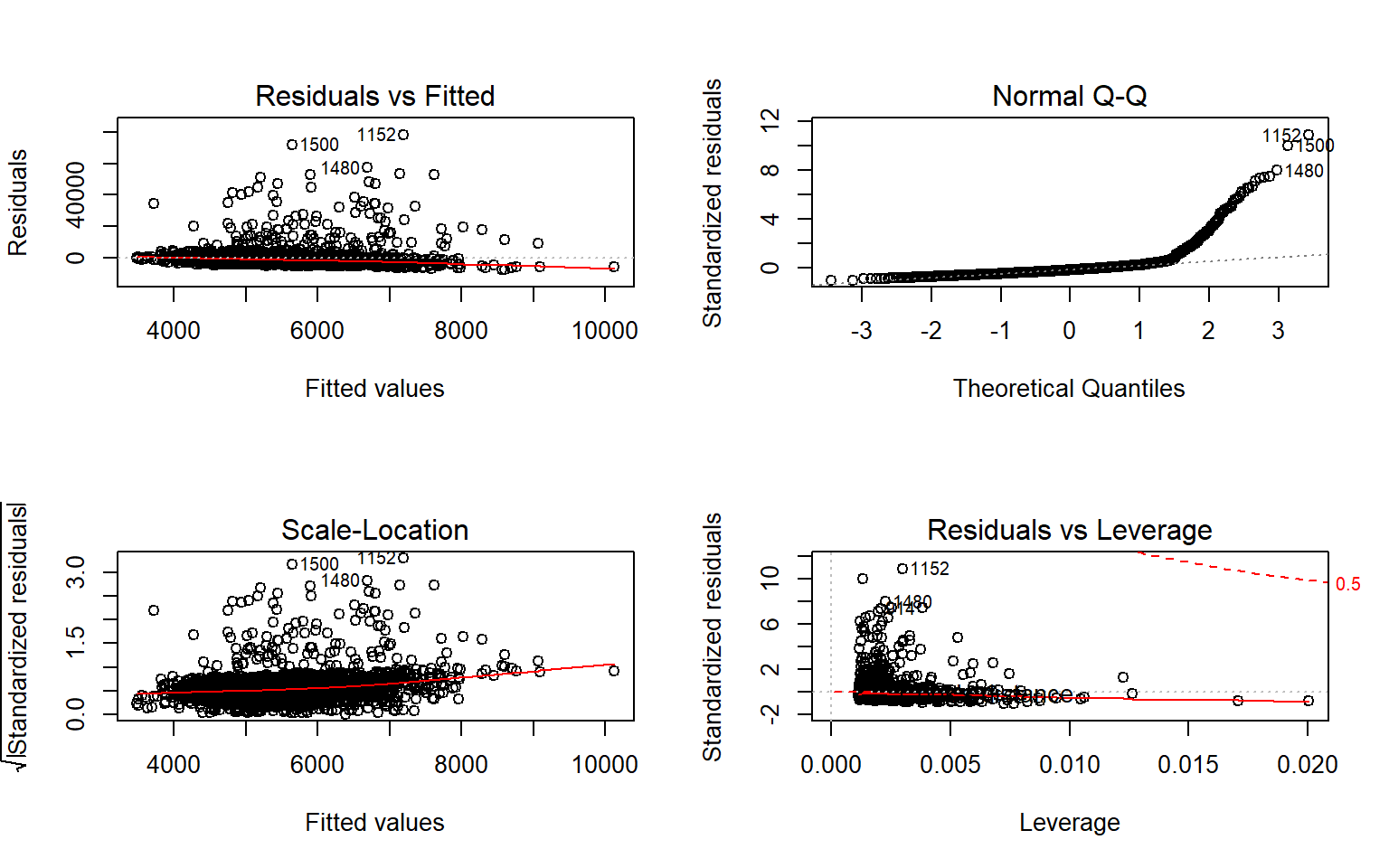
The f-statistic for all models also appeared to be significant. When viewing the plots of the models the biggest issues in each of the models is the Normal Q-Q plot. The quantile points do not appear to lie on the theoretical normal line. The models are ideally not what we would consider moving

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



Model 2



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

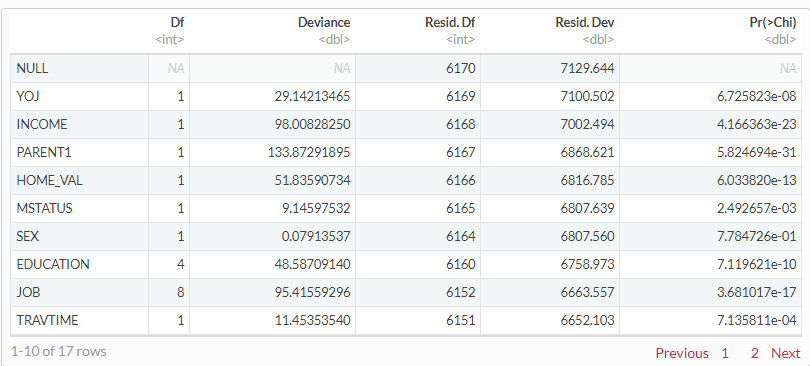
## [2,] 47757957

## [3,] 47757957

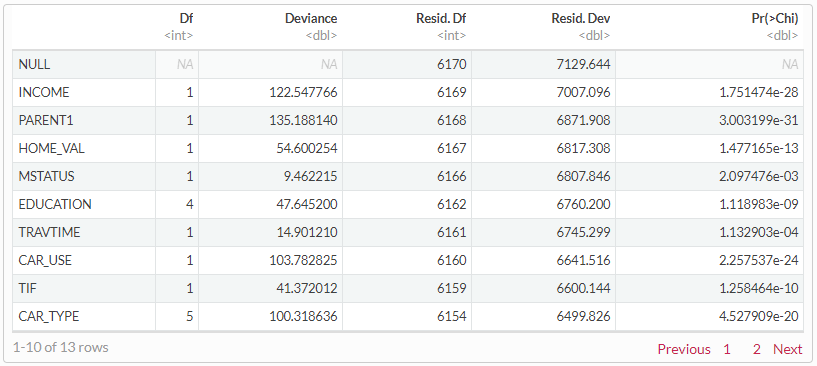
**Logit Models**

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



Model 2



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**