621 Assignment4

Raghu

Nov 12, 2018

Contents

1 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:

2 Loading Data Set and Cleaning:

Lets see the data structure.

```
## '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 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 28
##  $ 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 85
##  $ 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 ...
```

Lets have a glimpse on the cleaned data set after removing commas, dollar signs and Z_.

```
##   TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ PARENT1 TRAVTIME
## 1           0          0        0  60        0  11      No       14
## 2           0          0        0  43        0  11      No       22
## 3           0          0        0  35        1  10      No        5
## 4           0          0        0  51        0  14      No       32
## 5           0          0        0  50        0  NA      No       36
## 6           1       2946        0  34        1  12     Yes       46
```

```
##      CAR_USE TIF RED_CAR CLM_FREQ REVOKED MVR_PTS CAR_AGE INCOME HOME_VAL
## 1    Private  11     yes        2      No       3      18  67349        0
## 2 Commercial   1     yes        0      No       0       1  91449   257252
## 3    Private   4      no        2      No       3      10  16039   124191
## 4    Private   7     yes        0      No       0       6     NA   306251
## 5    Private   1      no        2     Yes       3      17 114986   243925
## 6 Commercial   1      no        0      No       0       7 125301        0
##   BLUEBOOK OLDCLAIM MSTATUS SEX    EDUCATION          JOB   CAR_TYPE
## 1    14230     4461      No   M          PhD Professional    Minivan
## 2    14940        0      No   M  High School  Blue Collar    Minivan
## 3     4010    38690     Yes   F  High School     Clerical        SUV
## 4    15440        0     Yes   M <High School  Blue Collar    Minivan
## 5    18000    19217     Yes   F          PhD       Doctor        SUV
## 6    17430        0      No   F    Bachelors  Blue Collar Sports Car
##            URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
## 3 Highly Urban/ Urban
## 4 Highly Urban/ Urban
## 5 Highly Urban/ Urban
## 6 Highly Urban/ Urban
```

3 Data Exploration

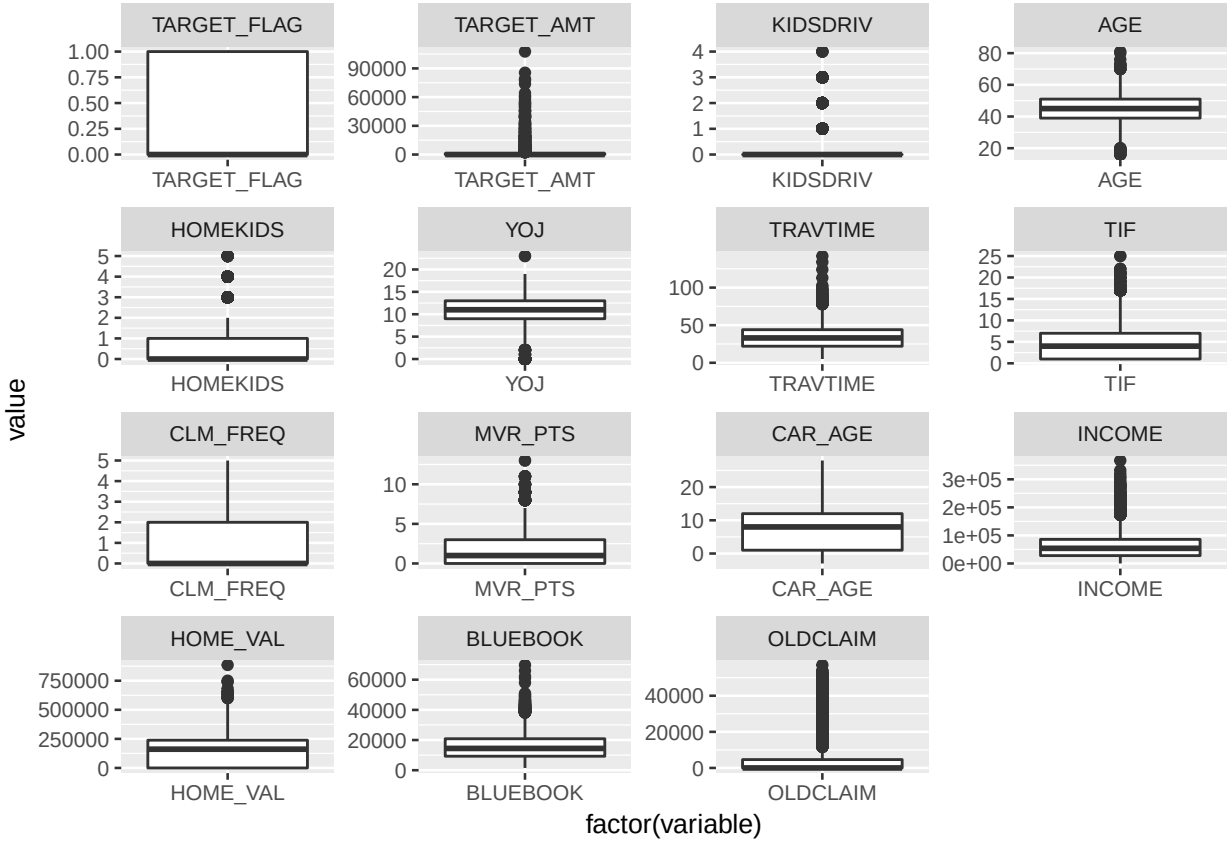
Based on the below summary statistics, we can see that there are NA's for AGE, YOJ, CAR_AGE, INCOME and HOME_VAL. There are categorical variables like PARENT1, CAR_USE, RED_CAR, REVOKED, MSTATUS, SEX, EDUCATION, JOB, CAR_TYPE and URBANICITY. Histograms shows some of the variables are skewed. Lets see if there are any correlations.

```
##   TARGET_FLAG       TARGET_AMT        KIDSDRIV           AGE
##  Min.   :0.0000   Min.   :     0   Min.   :0.0000   Min.   :16.00
##  1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000   1st Qu.:39.00
##  Median :0.0000   Median :     0   Median :0.0000   Median :45.00
##  Mean   :0.2638   Mean   :  1504   Mean   :0.1711   Mean   :44.79
##  3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000   3rd Qu.:51.00
##  Max.   :1.0000   Max.   :107586   Max.   :4.0000   Max.   :81.00
##                                                     NA's   :6
##     HOMEKIDS           YOJ        PARENT1        TRAVTIME
##  Min.   :0.0000   Min.   : 0.0   No :7084   Min.   :  5.00
##  1st Qu.:0.0000   1st Qu.: 9.0   Yes:1077   1st Qu.: 22.00
##  Median :0.0000   Median :11.0              Median : 33.00
##  Mean   :0.7212   Mean   :10.5              Mean   : 33.49
##  3rd Qu.:1.0000   3rd Qu.:13.0              3rd Qu.: 44.00
##  Max.   :5.0000   Max.   :23.0              Max.   :142.00
##                   NA's   :454
##        CAR_USE          TIF          RED_CAR       CLM_FREQ       REVOKED
##  Commercial:3029   Min.   : 1.000   no :5783   Min.   :0.0000   No :7161
##  Private   :5132   1st Qu.: 1.000   yes:2378   1st Qu.:0.0000   Yes:1000
##                    Median : 4.000              Median :0.0000
##                    Mean   : 5.351              Mean   :0.7986
##                    3rd Qu.: 7.000              3rd Qu.:2.0000
##                    Max.   :25.000              Max.   :5.0000
##
##     MVR_PTS          CAR_AGE           INCOME          HOME_VAL
##  Min.   : 0.000   Min.   :-3.000   Min.   :     0   Min.   :     0
##  1st Qu.: 0.000   1st Qu.: 1.000   1st Qu.: 28097   1st Qu.:     0
##  Median : 1.000   Median : 8.000   Median : 54028   Median :161160
##  Mean   : 1.696   Mean   : 8.328   Mean   : 61898   Mean   :154867
##  3rd Qu.: 3.000   3rd Qu.:12.000   3rd Qu.: 85986   3rd Qu.:238724
##  Max.   :13.000   Max.   :28.000   Max.   :367030   Max.   :885282
##                   NA's   :510      NA's   :445      NA's   :464
##     BLUEBOOK        OLDCLAIM     MSTATUS    SEX             EDUCATION
##  Min.   : 1500   Min.   :    0   No :3267   F:4375   <High School:1203
##  1st Qu.: 9280   1st Qu.:    0   Yes:4894   M:3786   Bachelors   :2242
##  Median :14440   Median :    0                       High School :2330
##  Mean   :15710   Mean   : 4037                       Masters     :1658
##  3rd Qu.:20850   3rd Qu.: 4636                       PhD         : 728
##  Max.   :69740   Max.   :57037
##
##            JOB              CAR_TYPE                  URBANICITY
##  Blue Collar :1825   Minivan    :2145   Highly Rural/ Rural:1669
##  Clerical    :1271   Panel Truck: 676   Highly Urban/ Urban:6492
##  Professional:1117   Pickup     :1389
##  Manager     : 988   Sports Car : 907
##  Lawyer      : 835   SUV        :2294
##  Student     : 712   Van        : 750
```

```
##  (Other)     :1413
```

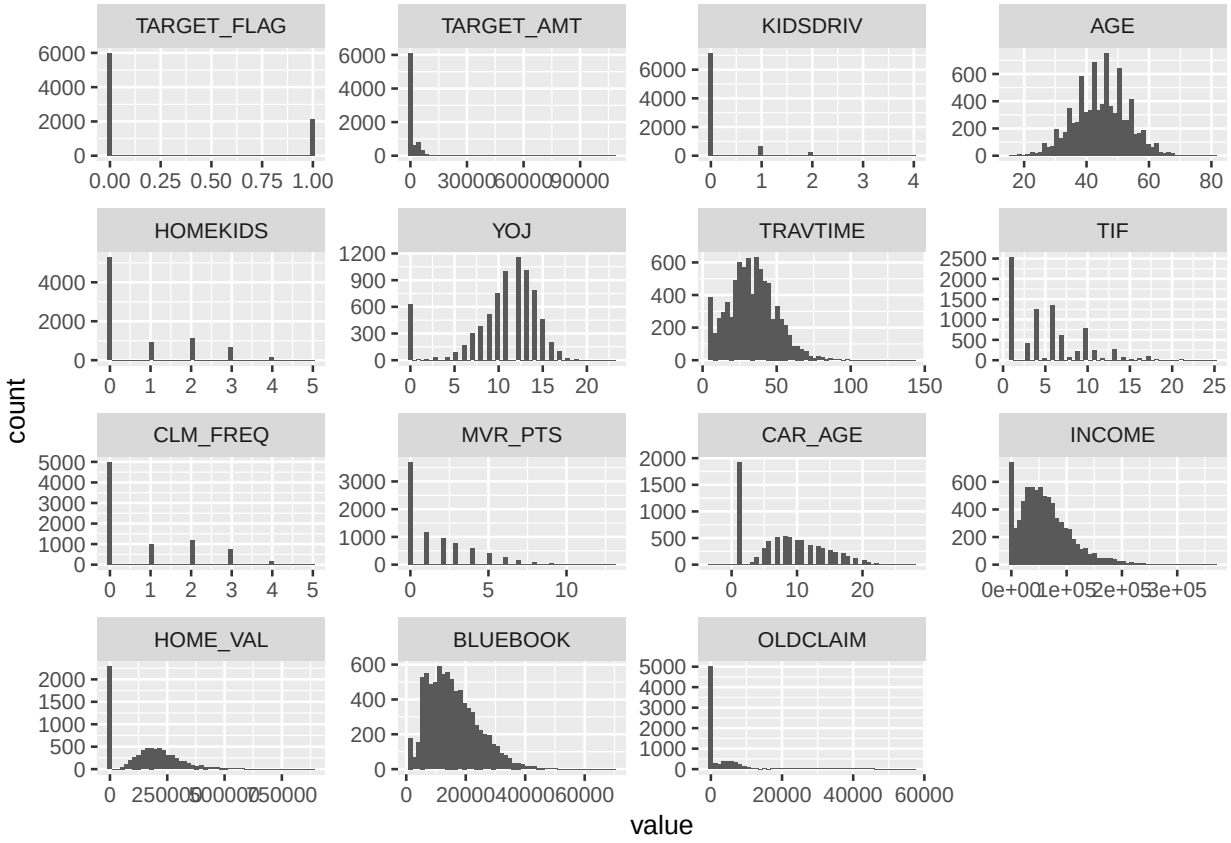
Boxplot

```
## Using PARENT1, CAR_USE, RED_CAR, REVOKED, MSTATUS, SEX, EDUCATION, JOB, CAR_TYPE, URBANICITY as id va
## Warning: Removed 1879 rows containing non-finite values (stat_boxplot).
```



Histogram

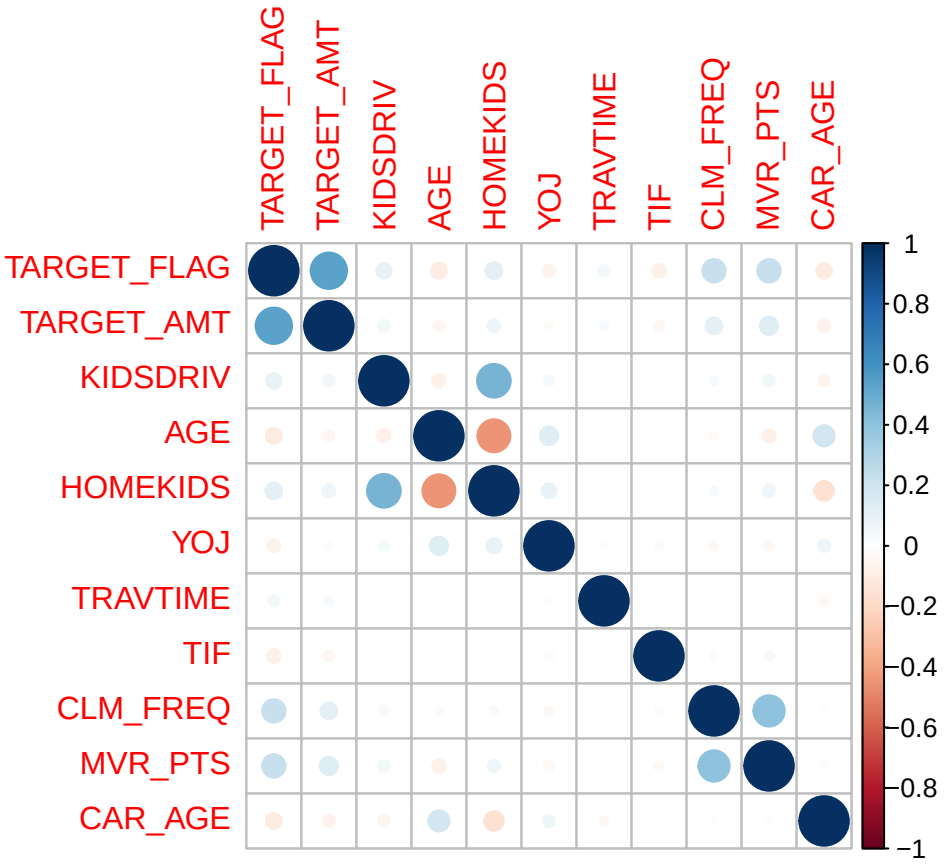
```
## Using PARENT1, CAR_USE, RED_CAR, REVOKED, MSTATUS, SEX, EDUCATION, JOB, CAR_TYPE, URBANICITY as id va
## Warning: Removed 1879 rows containing non-finite values (stat_bin).
```



Correlations:

```
##             TARGET_FLAG  TARGET_AMT     KIDSDRIV          AGE     HOMEKIDS
## TARGET_FLAG  1.00000000  0.53992929  0.097295404 -0.103649398  0.115481537
## TARGET_AMT   0.53992929  1.00000000  0.056857228 -0.046745810  0.068857861
## KIDSDRIV     0.09729540  0.05685723  1.000000000 -0.072361631  0.463046635
## AGE         -0.10364940 -0.04674581 -0.072361631  1.000000000 -0.442383841
## HOMEKIDS     0.11548154  0.06885786  0.463046635 -0.442383841  1.000000000
## YOJ         -0.06658661 -0.02144631  0.048112090  0.139566052  0.090416449
## TRAVTIME     0.05149130  0.03270817  0.008979590  0.004555303 -0.007787772
## TIF         -0.07718644 -0.04525925 -0.003423442  0.002871951  0.004673246
## CLM_FREQ     0.22338169  0.11515694  0.035087170 -0.026312189  0.030695809
## MVR_PTS      0.22526236  0.13770829  0.055019621 -0.073523273  0.062776101
## CAR_AGE     -0.10435770 -0.06283345 -0.055877063  0.182184524 -0.156534495
##                     YOJ     TRAVTIME          TIF     CLM_FREQ
## TARGET_FLAG -0.06658661  0.051491295 -0.077186438  0.223381685
## TARGET_AMT  -0.02144631  0.032708168 -0.045259254  0.115156936
## KIDSDRIV     0.04811209  0.008979590 -0.003423442  0.035087170
## AGE          0.13956605  0.004555303  0.002871951 -0.026312189
## HOMEKIDS     0.09041645 -0.007787772  0.004673246  0.030695809
## YOJ          1.00000000 -0.015762889  0.029302946 -0.030658029
## TRAVTIME    -0.01576289  1.000000000 -0.009343232  0.009306981
```

```
## TIF          0.02930295 -0.009343232  1.000000000 -0.024972898
## CLM_FREQ    -0.03065803  0.009306981 -0.024972898  1.000000000
## MVR_PTS     -0.03917262  0.009937566 -0.037174513  0.400121265
## CAR_AGE      0.06122969 -0.037055196  0.009125709 -0.011538390
##                  MVR_PTS      CAR_AGE
## TARGET_FLAG  0.225262361 -0.104357704
## TARGET_AMT   0.137708292 -0.062833451
## KIDSDRIV     0.055019621 -0.055877063
## AGE         -0.073523273  0.182184524
## HOMEKIDS     0.062776101 -0.156534495
## YOJ         -0.039172617  0.061229694
## TRAVTIME     0.009937566 -0.037055196
## TIF         -0.037174513  0.009125709
## CLM_FREQ     0.400121265 -0.011538390
## MVR_PTS      1.000000000 -0.019363647
## CAR_AGE     -0.019363647  1.000000000
```



There appears to be correlation between AGE, HOMEKIDS and accidents.

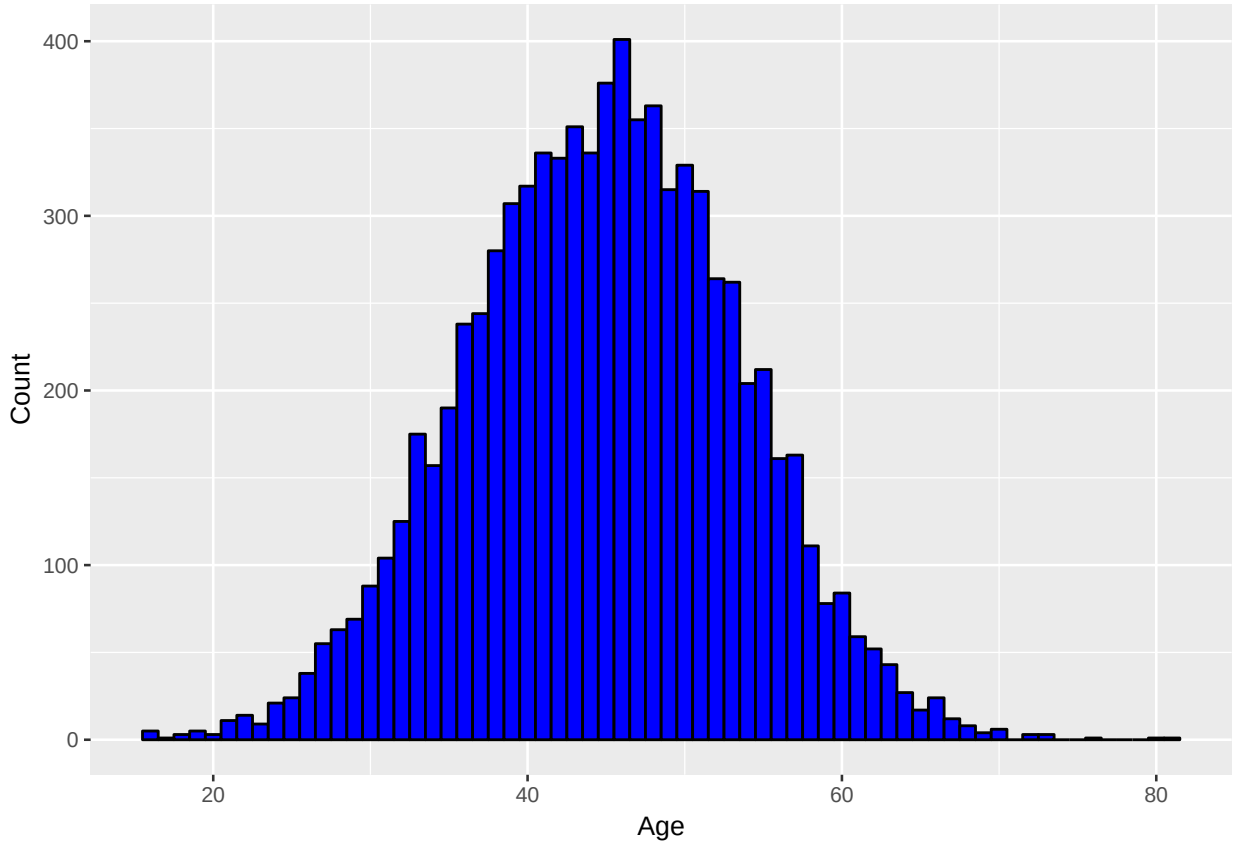
4 Factors affecting Insurance claims:

Lets see the factors affecting the claims.

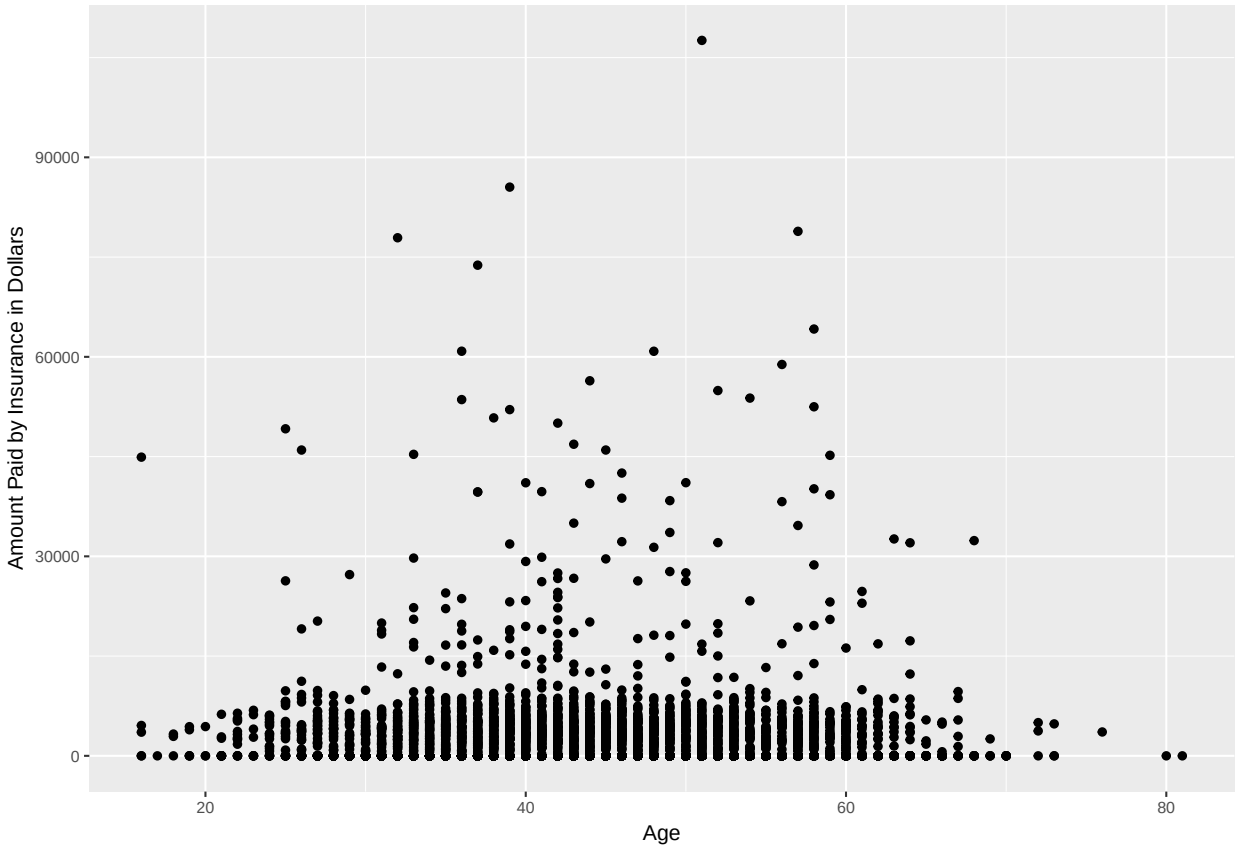
4.1 1. Age and Gender:

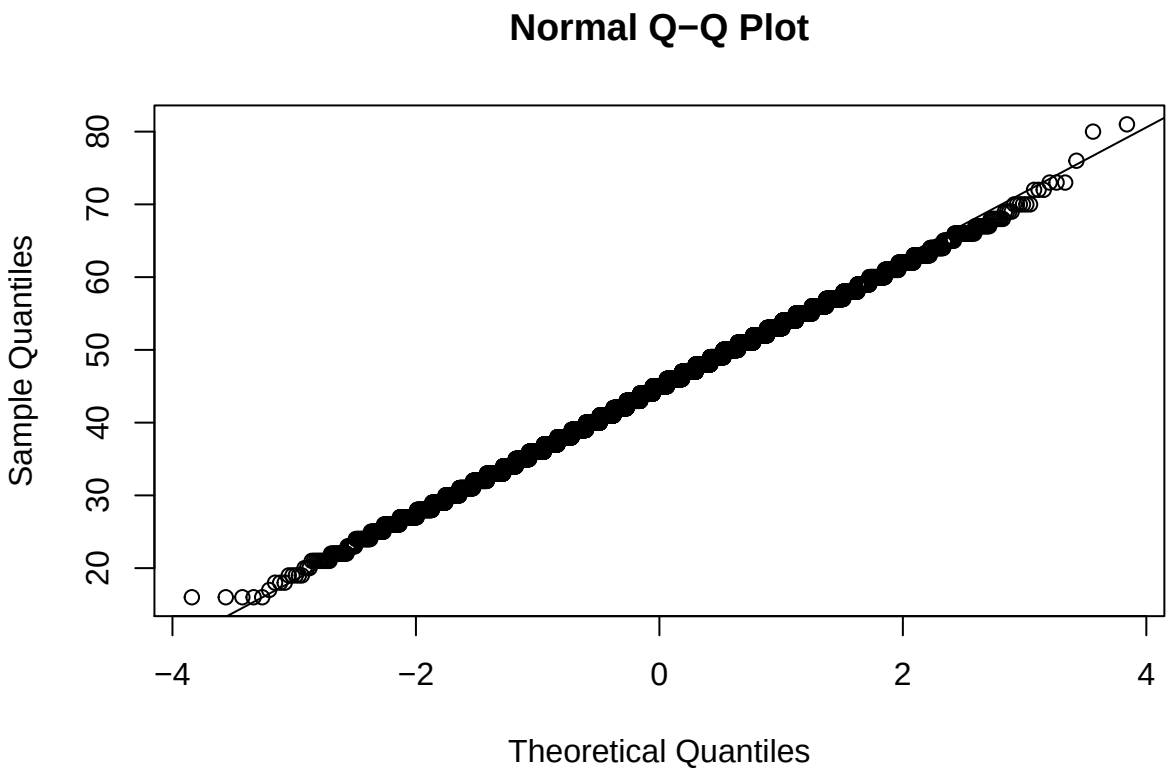
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   16.00   39.00   45.00   44.79   51.00   81.00       6
```

```
## Warning: Removed 6 rows containing non-finite values (stat_bin).
```




```
## Warning: Removed 6 rows containing missing values (geom_point).
```


Amount Paid by Insurance in Dollars
90000
60000
30000
0
20
40
60
80
Age




```
##         TARGET_FLAG
## AGE         0    1
##   Young   222  187
##   Middle 5603 1882
##   Old     182   79
```

```
## [1] "Percentage of Young Drivers: 30 and below in a Crash? 0.457"
```

```
## [1] "Percentage of Middle Age Drivers: 30 to 60 in a Crash? 0.251"
```

```
## [1] "Percentage of Older Drivers: 60 and older in a Crash? 0.303"
```

```
## [1] "Correlation between Age and car crash? -0.103"
```

```
## [1] "Correlation between Age and Amount Paid in car crash? -0.042"
```

Analysis: It appears that young and old drivers are more involved in crash compared to the percentage of middleaged drivers. But the number of middle aged drivers are more and the incidents is high in the middle aged drivers group. There does not appear to be any correlation between age and amount paid.

Let's take a look at the linear regression model using only age as the predictor variable and TARGET_AMT as the response variable.

```
##
## Call:
## lm(formula = TARGET_AMT ~ AGE, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2158  -1566  -1407   -516 106225
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2522.162    275.257   9.163  < 2e-16 ***
## AGE          -22.757      6.035  -3.771 0.000164 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4701 on 8153 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.001741,   Adjusted R-squared:  0.001619
## F-statistic: 14.22 on 1 and 8153 DF,  p-value: 0.0001637
```

Age seems to be statistically significant but Rsquare is not good.

Let's take a look at the logistic regression for Age vs. TARGET_FLAG:

```
##
## Call:
## glm(formula = TARGET_FLAG ~ AGE, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0723  -0.8035  -0.7395   1.4260   2.0183
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.186140   0.132052   1.410    0.159
## AGE         -0.027408   0.002956  -9.272   <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: 9404.0  on 8154  degrees of freedom
## Residual deviance: 9316.8  on 8153  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 9320.8
##
## Number of Fisher Scoring iterations: 4
```

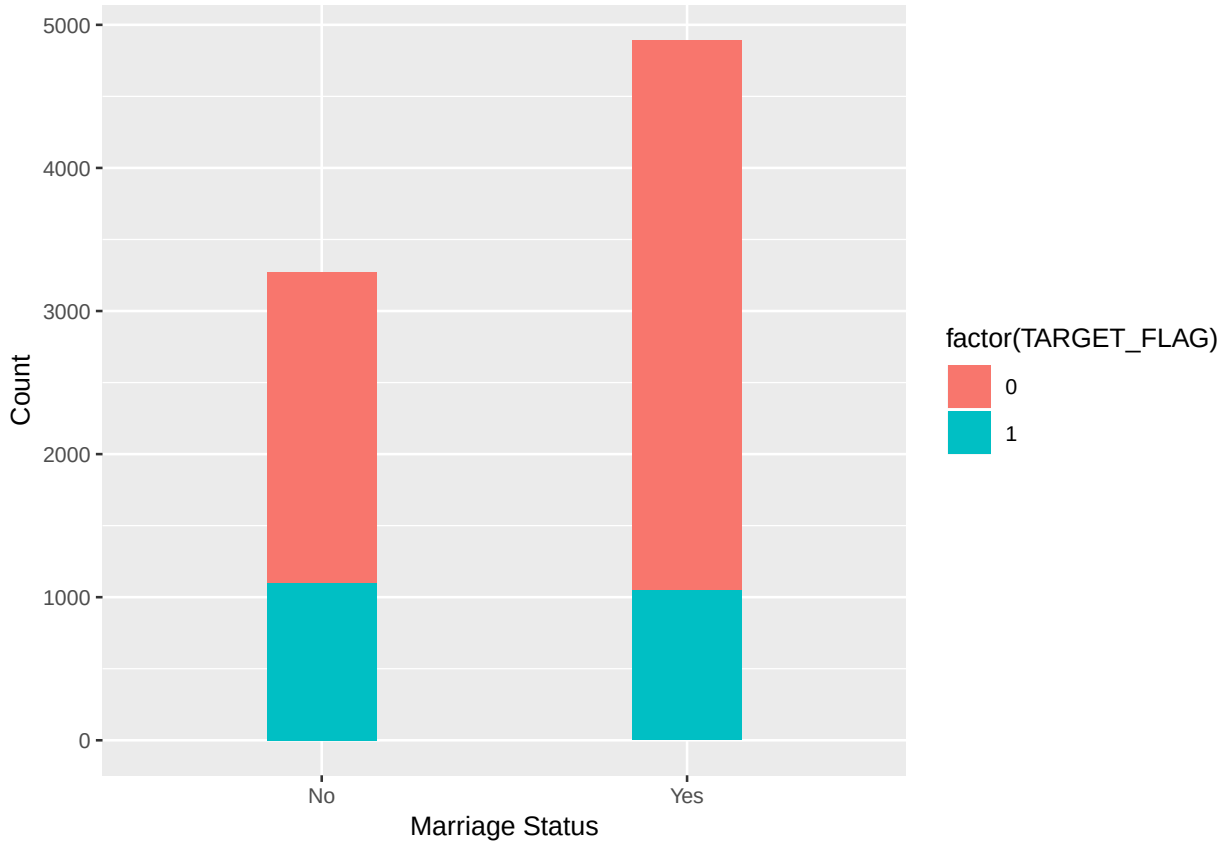
Age seems to be statistically significant.

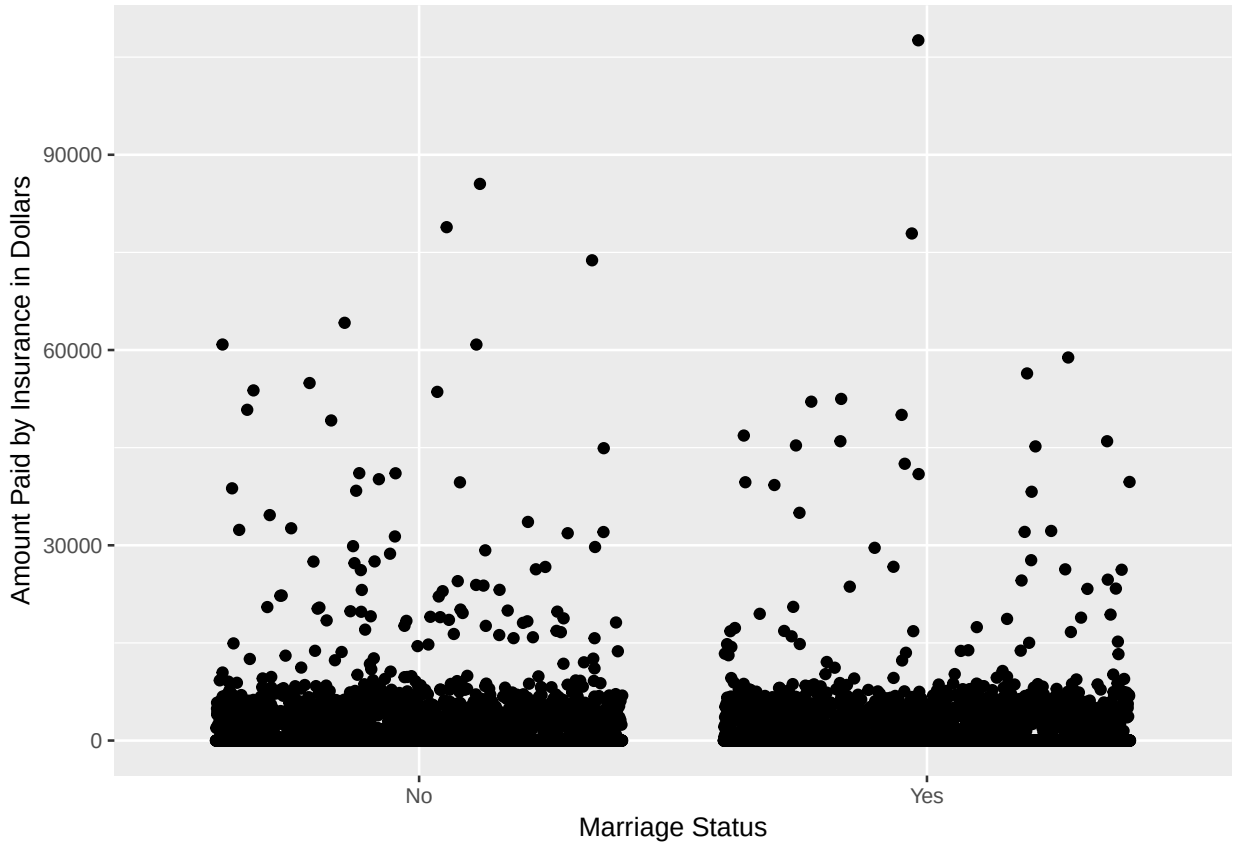
Age appears to be normally distributed.

4.2 2. Marital Status:

Percentage of married people involved in crash is less compared to unmarried. Looks like married people tend to drive very carefully.

```
##   No  Yes
## 3267 4894
```






```
##        TARGET_FLAG
## MSTATUS    0    1
##     No  2167 1100
##     Yes 3841 1053
```

```
## [1] "Percentage of Unmarried People Involved in Car Crashes: 0.337"
```

```
## [1] "Percentage of Married People Involved in Car Crashes: 0.215"
```

4.3 3. Place of living(Urban vs Rural):

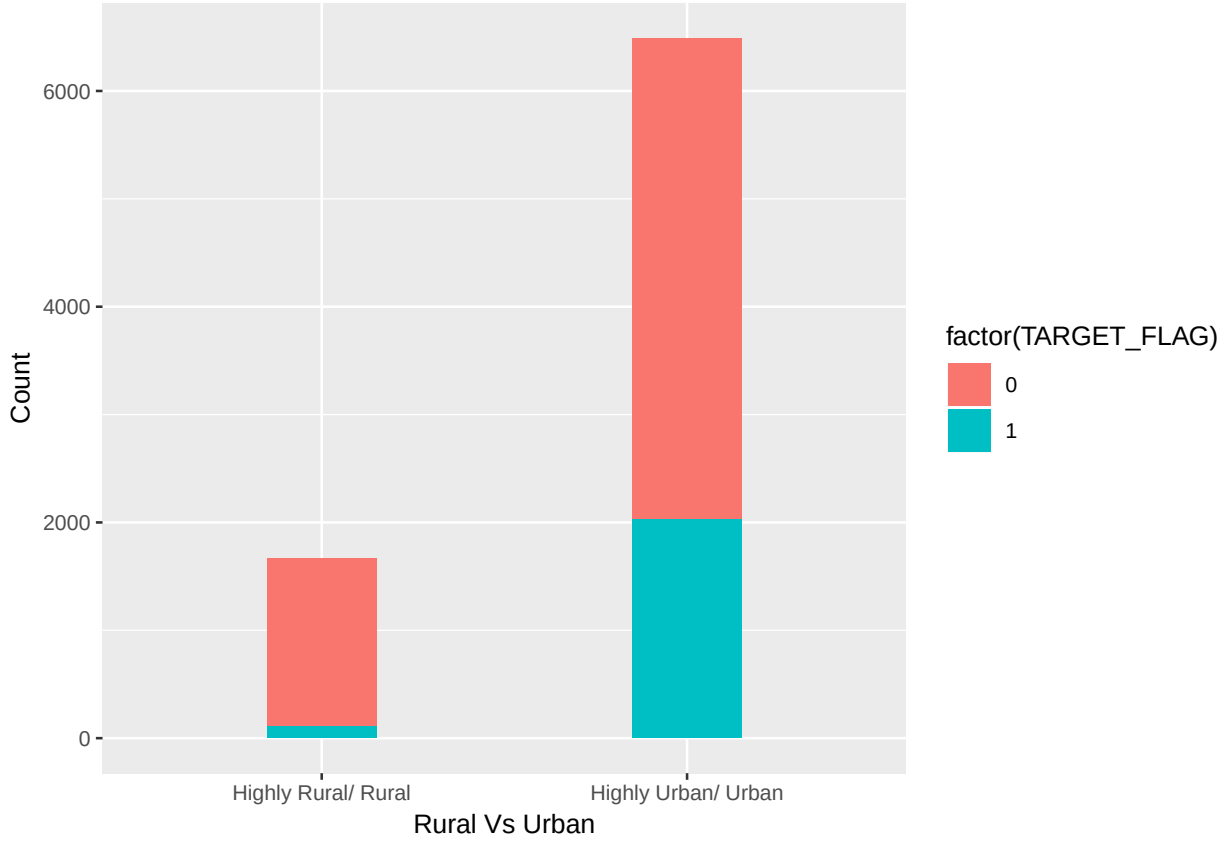
Its very common that most accidents occur in Urban areas where its highly crowded. Most densely populated neighborhoods are at high risk for accidents and also the insurance rates are high. Also, Urban areas with unemployment rates have lot of uninsured drivers.

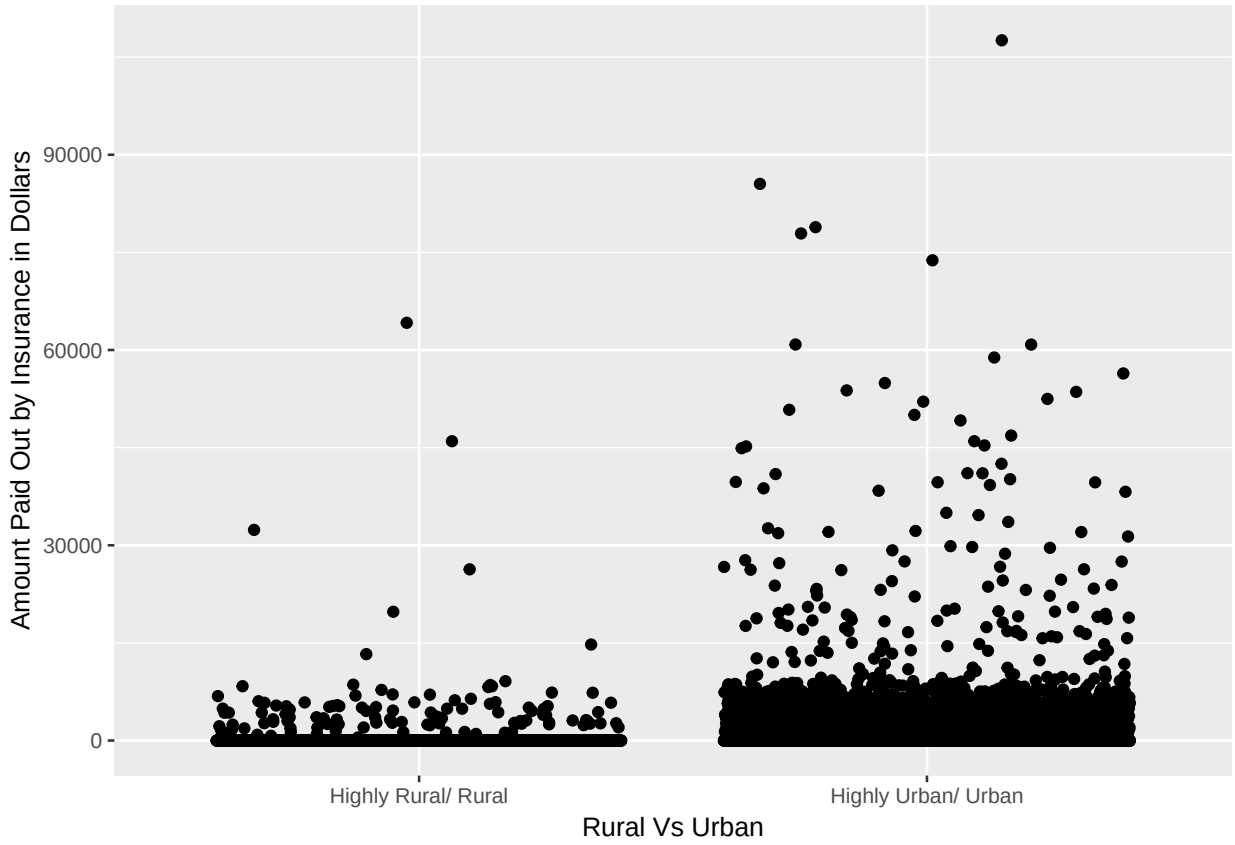
```
## Highly Rural/ Rural Highly Urban/ Urban
##                1669                6492
```

```
##                       TARGET_FLAG
## URBANICITY                  0    1
##   Highly Rural/ Rural 1554  115
##   Highly Urban/ Urban 4454 2038
```

```
## [1] "Percentage of People Living in Rural Areas Involved in Car Crashes: 0.069"
```

```
## [1] "Percentage of People Living in Urban Areas Involved in Car Crashes: 0.314"
```




Amount Paid Out by Insurance in Dollars
90000
60000
30000
0
Highly Rural/ Rural
Highly Urban/ Urban
Rural Vs Urban

4.4 4. Profession:

Insurance industry considers the profession to calculate the risk of accident makers. For example, taxi drivers or truck drivers are on the road constantly where as some professionals does not spend much time on the road or they are very careful.

It appears that jobs such as doctor, manager, lawyer, and professional seem to have less percent crashed. Let's categorize the data into two buckets, Professional and NonProfessional.

```
## [1] ""             "Blue Collar"  "Clerical"     "Doctor"
## [5] "Home Maker"   "Lawyer"       "Manager"      "Professional"
## [9] "Student"
```

	0	1	Percent_Crashed
	390	136	0.2585551
Blue Collar	1191	634	0.3473973
Clerical	900	371	0.2918961
Doctor	217	29	0.1178862
Home Maker	461	180	0.2808112
Lawyer	682	153	0.1832335
Manager	851	137	0.1386640
Professional	870	247	0.2211280
Student	446	266	0.3735955

```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

```
## Warning in `==.default`(prof_df$JOB, c("Doctor", "Manager", "Lawyer",
## "Professional")): longer object length is not a multiple of shorter object
## length
```

```
##                   TARGET_FLAG
## JOB_CAT               0    1
##   NonProfessional 5354 2030
##   Professional     654  123
```

```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

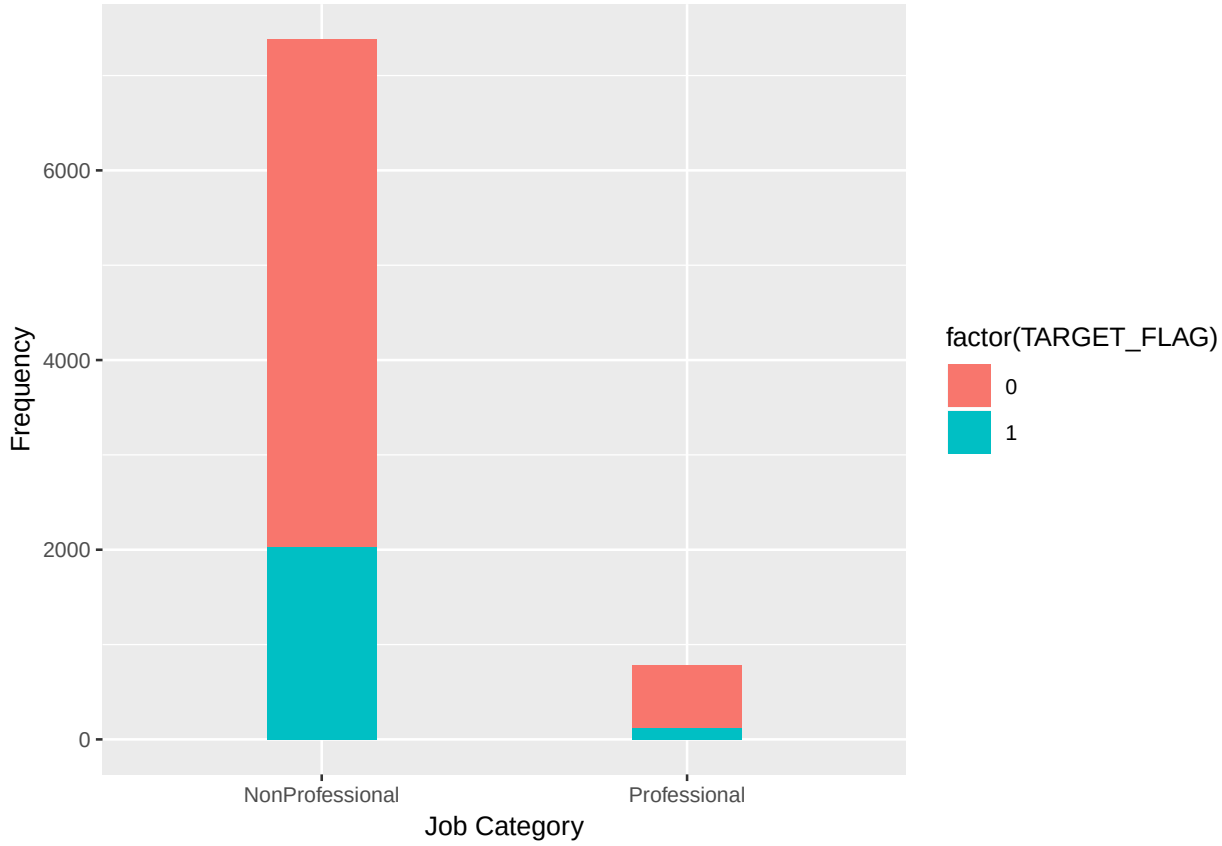
```
## Warning in `==.default`(JOB, c("Doctor", "Manager", "Lawyer",
## "Professional")): longer object length is not a multiple of shorter object
## length
```

```
##   TARGET_FLAG       TARGET_AMT
##  Min.   :0.0000   Min.   :    0.0
##  1st Qu.:0.0000   1st Qu.:    0.0
##  Median :0.0000   Median :    0.0
##  Mean   :0.1583   Mean   :  668.3
##  3rd Qu.:0.0000   3rd Qu.:    0.0
##  Max.   :1.0000   Max.   :32363.2
```

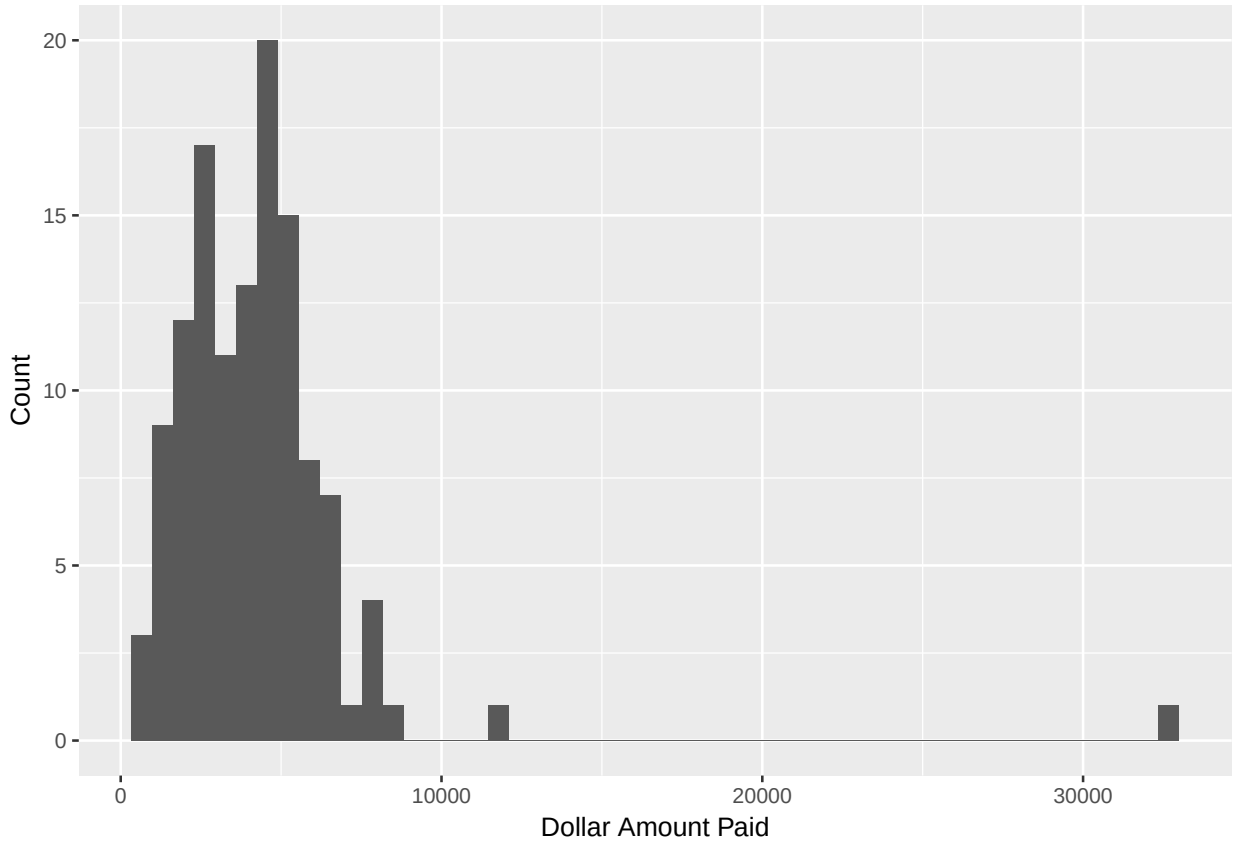
Lets get rid of the zeros from TARGET_AMT as that will heavily skew the TARGET_AMT data.

```
##    TARGET_AMT
##  Min.   :  343.9
##  1st Qu.: 2506.5
##  Median : 3993.0
##  Mean   : 4221.7
##  3rd Qu.: 5060.5
```

```
##  Max.   :32363.2
## [1] "There are a total of 777 professionals in this dataset."
## [1] "How many were involved in a car accident? 123"
## [1] "Percent Crashed: 0.158"
```




```
## No id variables; using all as measure variables
```




```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

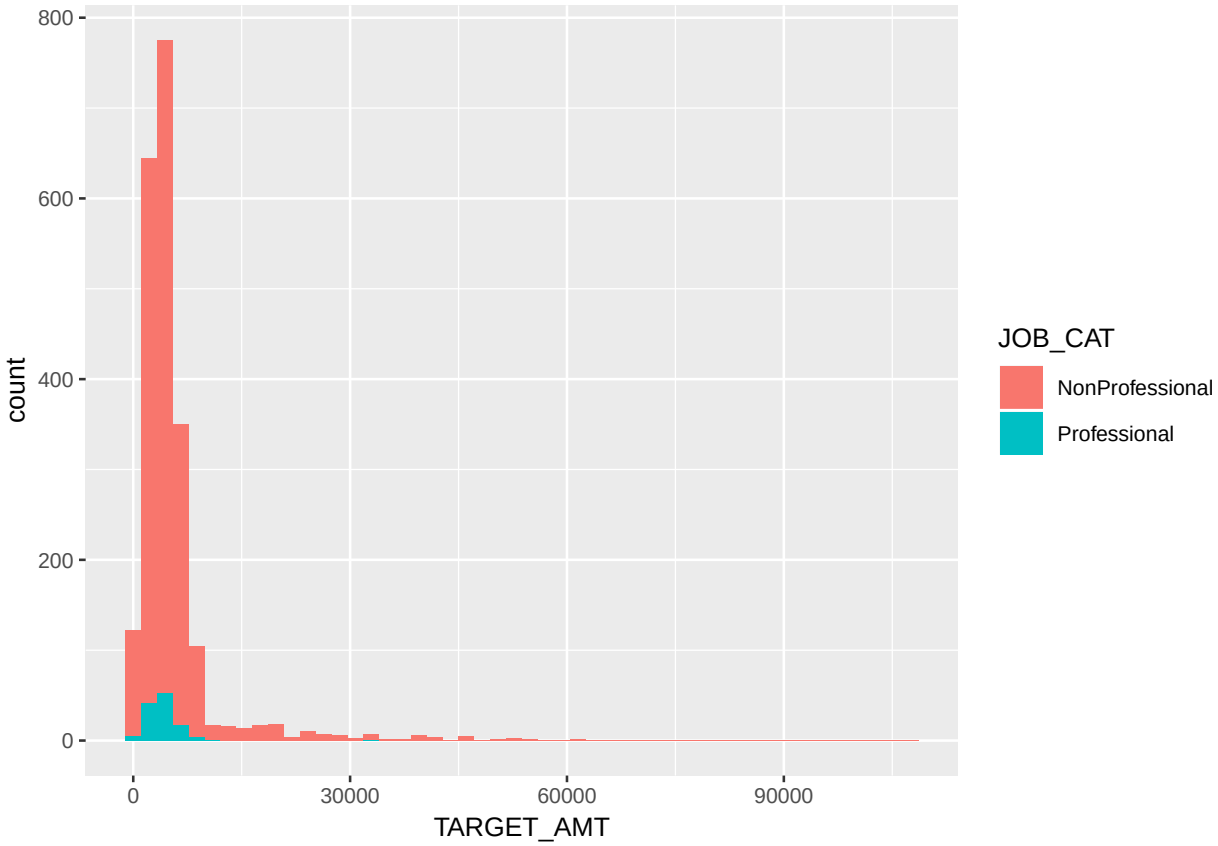
```
## Warning in `!=.default`(JOB, c("Doctor", "Manager", "Lawyer",
## "Professional")): longer object length is not a multiple of shorter object
## length
```

```
##    TARGET_AMT
##  Min.   :    30.28
##  1st Qu.:  2635.50
##  Median :  4114.57
##  Mean   :  5791.88
##  3rd Qu.:  5811.75
##  Max.   :107586.14
```

```
## [1] "There are a total of 7384 nonprofessionals in this dataset."
```

```
## [1] "How many were involved in a car accident? 2030"
```

```
## [1] "Percent Crashed: 0.275"
```


800
600
400
200
0
count
0
30000
60000
90000
TARGET_AMT
JOB_CAT
NonProfessional
Professional

4.5 5. Vehicle Size:

Larger cars have high insurance compared to smaller cars. On the other hand, Larger cars are generally safer than smaller cars in an accident. Cars with larger engines relative to body size tend to have higher rates - for instance, insurance for a sports car with a V8 engine costs much more than a small car with a V4 engine.

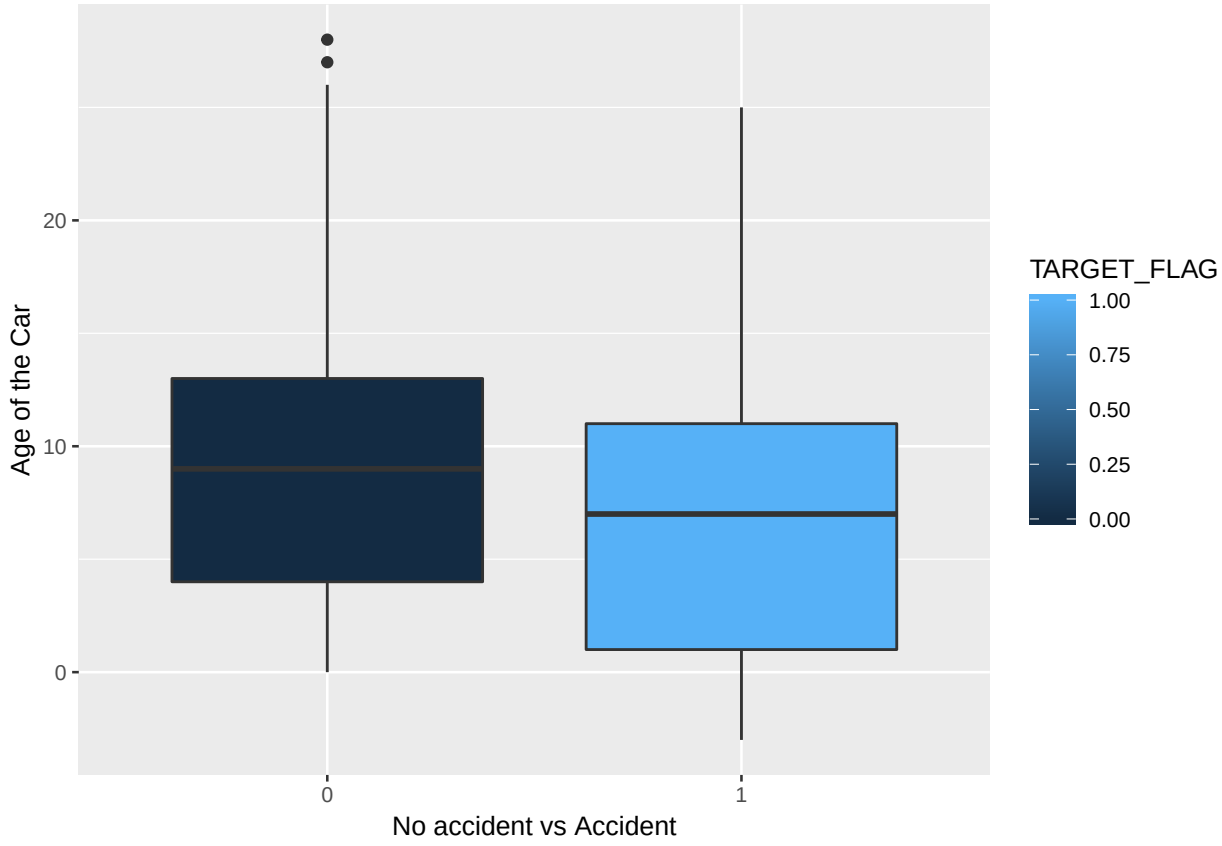
```
## [1] "Minivan"     "Panel Truck" "Pickup"      "Sports Car"  "SUV"
## [6] "Van"
```

```
##              TARGET_FLAG
## CAR_TYPE           0    1
##   Minivan       1796  349
##   Panel Truck    498  178
##   Pickup         946  443
##   Sports Car     603  304
##   SUV           1616  678
##   Van            549  201
```

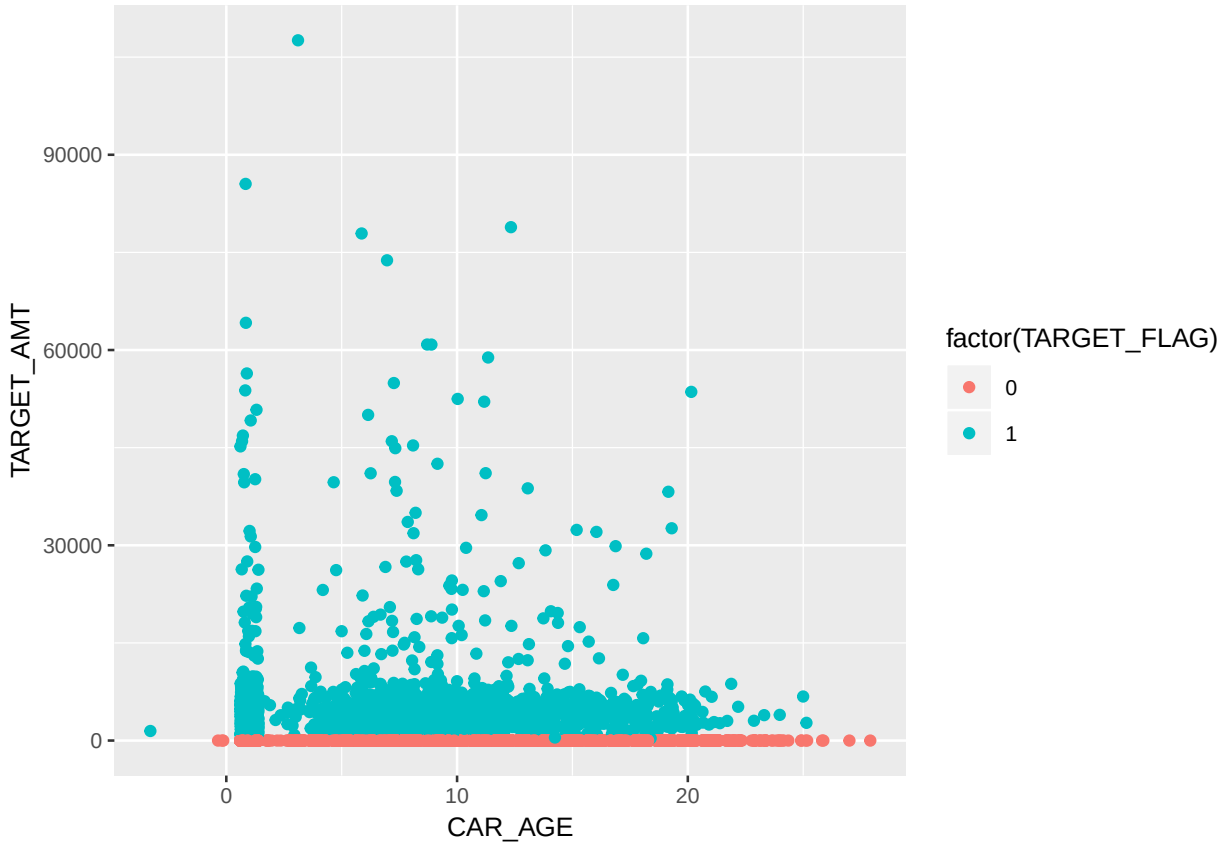
4.6 6. Age of the vehicle:

Newer vehicle usually tend to have higher premiums compared to older vehicle. As the vehicle cost depreciates over the years, premiums also drop minimally. cost to replace a newer vehicle involved in accident is expensive compared to older vehicle.

```
## Warning: Removed 510 rows containing non-finite values (stat_boxplot).
```

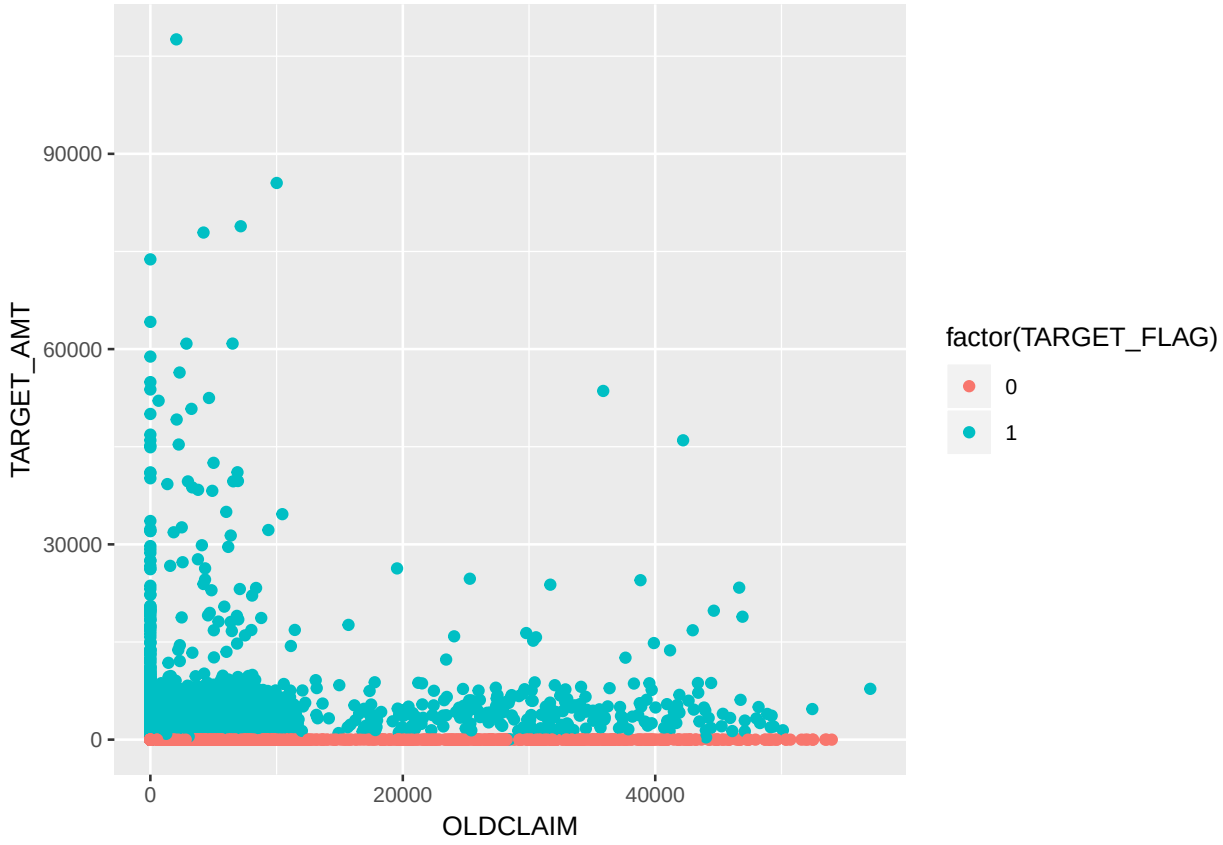



```
## Warning: Removed 510 rows containing missing values (geom_point).
```


90000
60000
30000
0
TARGET_AMT
0
10
20
CAR_AGE
factor(TARGET_FLAG)
0
1

4.7 7. Driving History:

Prior accidents will lead to higher premiums. In this data set, we have CLM_FREQ and OLDCLAIM fields which indicates how many times accidents happened and how much was claimed. The higher the frequency, the more likely the person will meet with accident in future and submit claim.




```
##               OLDCLAIM  CLM_FREQ TARGET_AMT
## OLDCLAIM   1.00000000 0.4951308 0.07095329
## CLM_FREQ   0.49513081 1.0000000 0.11641916
## TARGET_AMT 0.07095329 0.1164192 1.00000000
```

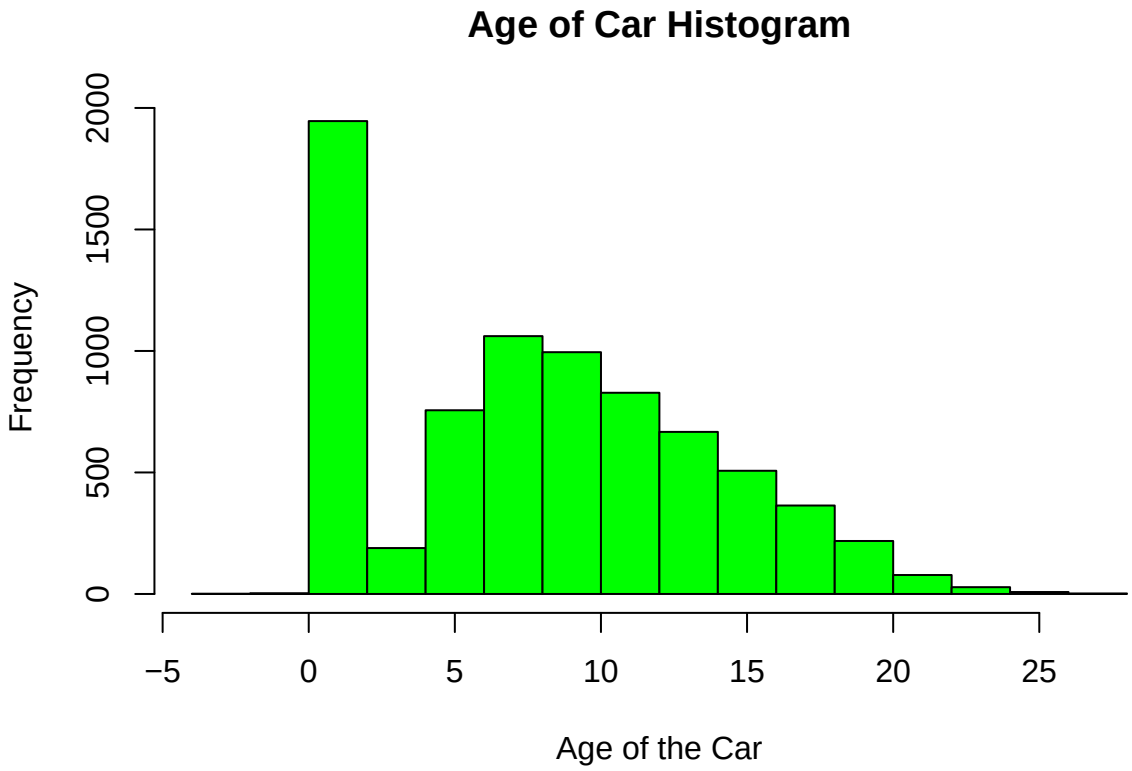
5 Data Preparation and Model Building

Earlier we noticed that there were significant missing data. Lets see how many are missing overall.

```
## [1] 1879
```

```
##     CAR_AGE    HOME_VAL         YOJ      INCOME         AGE  URBANICITY
##         510         464         454         445           6           0
##    CAR_TYPE         JOB   EDUCATION         SEX     MSTATUS    OLDCLAIM
##           0           0           0           0           0           0
##    BLUEBOOK     MVR_PTS     REVOKED    CLM_FREQ     RED_CAR         TIF
##           0           0           0           0           0           0
##     CAR_USE    TRAVTIME     PARENT1    HOMEKIDS    KIDSDRIV  TARGET_AMT
##           0           0           0           0           0           0
## TARGET_FLAG
##           0
```

Let's see if we can impute these values. Let's start with CAR_AGE. We will take only complete cases and create a histogram to see how it is distributed.



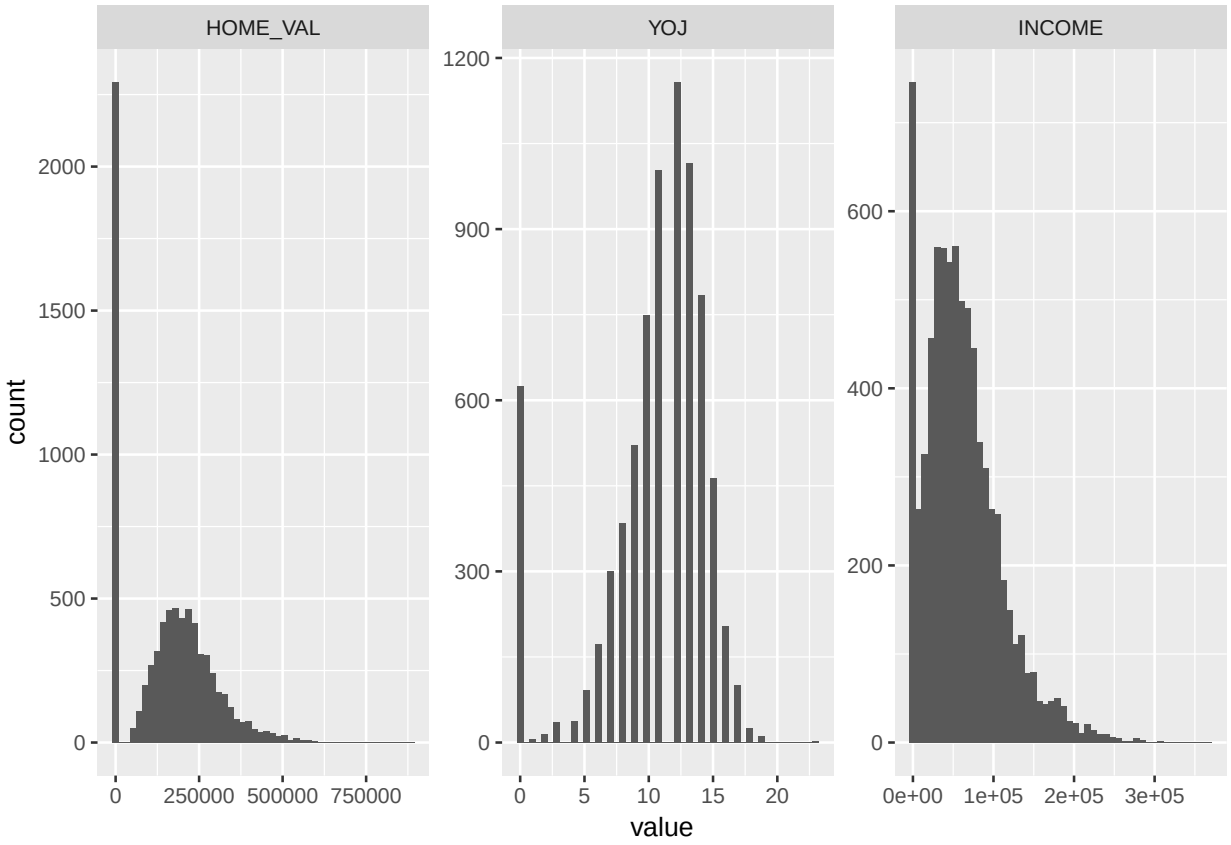

```
## [1] 1
```

It is right skewed, with one bin appearing to be the most frequent. It may be best to impute the missing data with the mode value, as opposed to the mean or median value.

Now, lets Impute other predictor Variables HOME_VAL, YOJ and INCOME.

```
## No id variables; using all as measure variables
```

```
## Warning: Removed 1363 rows containing non-finite values (stat_bin).
```



Looking at the histograms, the best imputation for HOME_VAL, YOJ, and INCOME would be the mode value, median value, and median value respectively. Let's go ahead and impute these values into the dataset.

There are 6 missing points for AGE which will be substitued with median as well.

Let's start transforming some of our predictor variables using box cox as not all the data points are normally distributed. By performing a BoxCox transformation, It might normalize some of these predictors, which could potentially yield a better model. Box cox transformation can be done only to positive and not-zero values. Since CAR_AGE, YOJ and INCOME have zeros, I will substite that with 1.

```
## $TIF
## Box-Cox Transformation
##
## 8160 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   4.000   5.351   7.000  25.000
##
## Largest/Smallest: 25
## Sample Skewness: 0.891
##
## Estimated Lambda: 0.2
##
##
## $BLUEBOOK
## Box-Cox Transformation
##
```

```
## 8160 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1500    9280   14440   15710   20850   69740
##
## Largest/Smallest: 46.5
## Sample Skewness: 0.794
##
## Estimated Lambda: 0.5
##
##
## $TRAVTIME
## Box-Cox Transformation
##
## 8160 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   22.00   33.00   33.48   44.00  142.00
##
## Largest/Smallest: 28.4
## Sample Skewness: 0.447
##
## Estimated Lambda: 0.7
##
##
## $AGE
## Box-Cox Transformation
##
## 8160 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   39.00   45.00   44.79   51.00   81.00
##
## Largest/Smallest: 5.06
## Sample Skewness: -0.029
##
## Estimated Lambda: 1
## With fudge factor, no transformation is applied
##
##
## $CAR_AGE
## Box-Cox Transformation
##
## 8160 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   9.000   8.872  13.000  29.000
##
## Largest/Smallest: 29
## Sample Skewness: 0.363
```

```
##
## Estimated Lambda: 0.4
##
##
## $YOJ
## Box-Cox Transformation
##
## 8160 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   10.00   12.00   11.53   14.00   24.00
##
## Largest/Smallest: 24
## Sample Skewness: -1.26
##
## Estimated Lambda: 1.6
##
##
## $INCOME
## Box-Cox Transformation
##
## 8160 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1   29710   54030   61470   83310  367000
##
## Largest/Smallest: 367000
## Sample Skewness: 1.24
##
## Estimated Lambda: 0.4
```

Box cox has suggested transformations and this will be considered for training the model.

Let's now proceed with building models for both logistic regression to determine whether or not a crash will occur and a linear regression to determine how much money is paid out if a crash did indeed occur. We will build the models and evaluate them in terms of their efficacy.

5.1 Model1: Binary Logistic Regression Model

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = "binomial", data = full_bin_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5827  -0.7114  -0.3982   0.6266   3.1578
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.901e+00  3.405e-01  -8.519  < 2e-16 ***
## KIDSDRIV                    3.861e-01  6.119e-02   6.311 2.77e-10 ***
## AGE                        -1.231e-03  4.016e-03  -0.307 0.759182
## HOMEKIDS                    5.084e-02  3.711e-02   1.370 0.170667
## YOJ                        -1.116e-02  8.583e-03  -1.301 0.193401
## PARENT1Yes                  3.794e-01  1.096e-01   3.463 0.000534 ***
## TRAVTIME                    1.462e-02  1.883e-03   7.762 8.37e-15 ***
## CAR_USEPrivate             -7.566e-01  9.170e-02  -8.251  < 2e-16 ***
## TIF                        -5.538e-02  7.343e-03  -7.543 4.61e-14 ***
## RED_CARyes                 -7.382e-03  8.633e-02  -0.086 0.931852
## CLM_FREQ                    1.967e-01  2.854e-02   6.894 5.41e-12 ***
## REVOKEDYes                  8.884e-01  9.128e-02   9.733  < 2e-16 ***
## MVR_PTS                     1.130e-01  1.361e-02   8.305  < 2e-16 ***
## CAR_AGE                    -3.945e-03  6.867e-03  -0.575 0.565619
## INCOME                     -3.624e-06  1.075e-06  -3.372 0.000747 ***
## HOME_VAL                   -1.084e-06  3.171e-07  -3.419 0.000627 ***
## BLUEBOOK                   -2.072e-05  5.260e-06  -3.939 8.20e-05 ***
## OLDCLAIM                   -1.389e-05  3.910e-06  -3.553 0.000380 ***
## MSTATUSYes                 -5.222e-01  8.176e-02  -6.388 1.68e-10 ***
## SEXM                        8.015e-02  1.120e-01   0.715 0.474379
## EDUCATIONBachelors         -3.694e-01  1.141e-01  -3.238 0.001205 **
## EDUCATIONHigh School        1.925e-02  9.493e-02   0.203 0.839295
## EDUCATIONMasters           -2.662e-01  1.756e-01  -1.516 0.129557
## EDUCATIONPhD               -1.419e-01  2.113e-01  -0.672 0.501883
## JOBBlue Collar              3.082e-01  1.855e-01   1.662 0.096601 .
## JOBClerical                 4.115e-01  1.966e-01   2.093 0.036345 *
## JOBDoctor                  -4.445e-01  2.668e-01  -1.666 0.095662 .
## JOBHome Maker               2.368e-01  2.100e-01   1.127 0.259589
## JOBLawyer                   1.093e-01  1.694e-01   0.645 0.518722
## JOBManager                 -5.566e-01  1.714e-01  -3.247 0.001167 **
## JOBProfessional             1.581e-01  1.784e-01   0.887 0.375226
## JOBStudent                  2.307e-01  2.143e-01   1.076 0.281781
## CAR_TYPEPanel Truck         5.538e-01  1.617e-01   3.425 0.000615 ***
## CAR_TYPEPickup              5.511e-01  1.007e-01   5.473 4.42e-08 ***
## CAR_TYPESports Car          1.023e+00  1.299e-01   7.875 3.40e-15 ***
## CAR_TYPESUV                 7.651e-01  1.113e-01   6.877 6.12e-12 ***
## CAR_TYPEVan                 6.129e-01  1.264e-01   4.849 1.24e-06 ***
```

```
## URBANICITYHighly Urban/ Urban  2.392e+00  1.129e-01  21.193  < 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: 7300.4  on 8123  degrees of freedom
## AIC: 7376.4
##
## Number of Fisher Scoring iterations: 5
```

5.2 Model2: Backwards stepwise approach Logistic Regression Model

In this model, I will take the full model and attempt to shrink the amount of predictor variables needed by using a backwards stepwise approach.

```
## Start:  AIC=7376.36
## TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + TRAVTIME +
##     CAR_USE + TIF + RED_CAR + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + INCOME + HOME_VAL + BLUEBOOK + OLDCLAIM + MSTATUS +
##     SEX + EDUCATION + JOB + CAR_TYPE + URBANICITY
##
##              Df Deviance    AIC
## - RED_CAR     1   7300.4 7374.4
## - AGE         1   7300.5 7374.5
## - CAR_AGE     1   7300.7 7374.7
## - SEX         1   7300.9 7374.9
## - YOJ         1   7302.1 7376.1
## - HOMEKIDS    1   7302.2 7376.2
## <none>            7300.4 7376.4
## - INCOME      1   7311.9 7385.9
## - HOME_VAL    1   7312.1 7386.1
## - PARENT1     1   7312.4 7386.4
## - OLDCLAIM    1   7313.2 7387.2
## - EDUCATION   4   7321.6 7389.6
## - BLUEBOOK    1   7316.1 7390.1
## - KIDSDRIV    1   7340.2 7414.2
## - MSTATUS     1   7340.9 7414.9
## - CLM_FREQ    1   7347.3 7421.3
## - JOB         8   7362.5 7422.5
## - TIF         1   7359.4 7433.4
## - TRAVTIME    1   7360.8 7434.8
## - CAR_USE     1   7369.5 7443.5
## - MVR_PTS     1   7369.9 7443.9
## - CAR_TYPE    5   7391.5 7457.5
## - REVOKED     1   7393.5 7467.5
## - URBANICITY  1   7950.0 8024.0
##
## Step:  AIC=7374.37
## TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + TRAVTIME +
##     CAR_USE + TIF + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE +
##     INCOME + HOME_VAL + BLUEBOOK + OLDCLAIM + MSTATUS + SEX +
##     EDUCATION + JOB + CAR_TYPE + URBANICITY
##
##              Df Deviance    AIC
## - AGE         1   7300.5 7372.5
## - CAR_AGE     1   7300.7 7372.7
## - SEX         1   7300.9 7372.9
## - YOJ         1   7302.1 7374.1
## - HOMEKIDS    1   7302.2 7374.2
## <none>            7300.4 7374.4
## - INCOME      1   7311.9 7383.9
## - HOME_VAL    1   7312.1 7384.1
## - PARENT1     1   7312.4 7384.4
## - OLDCLAIM    1   7313.3 7385.3
```

```
## - EDUCATION   4   7321.6 7387.6
## - BLUEBOOK    1   7316.1 7388.1
## - KIDSDRIV    1   7340.3 7412.3
## - MSTATUS     1   7340.9 7412.9
## - CLM_FREQ    1   7347.3 7419.3
## - JOB         8   7362.5 7420.5
## - TIF         1   7359.4 7431.4
## - TRAVTIME    1   7360.8 7432.8
## - CAR_USE     1   7369.5 7441.5
## - MVR_PTS     1   7369.9 7441.9
## - CAR_TYPE    5   7391.6 7455.6
## - REVOKED     1   7393.5 7465.5
## - URBANICITY  1   7950.0 8022.0
##
## Step:  AIC=7372.46
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + PARENT1 + TRAVTIME +
##     CAR_USE + TIF + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE +
##     INCOME + HOME_VAL + BLUEBOOK + OLDCLAIM + MSTATUS + SEX +
##     EDUCATION + JOB + CAR_TYPE + URBANICITY
##
##              Df Deviance    AIC
## - CAR_AGE     1   7300.8 7370.8
## - SEX         1   7301.0 7371.0
## - YOJ         1   7302.3 7372.3
## <none>            7300.5 7372.5
## - HOMEKIDS    1   7303.0 7373.0
## - INCOME      1   7311.9 7381.9
## - HOME_VAL    1   7312.3 7382.3
## - PARENT1     1   7312.8 7382.8
## - OLDCLAIM    1   7313.3 7383.3
## - EDUCATION   4   7321.7 7385.7
## - BLUEBOOK    1   7316.8 7386.8
## - KIDSDRIV    1   7341.0 7411.0
## - MSTATUS     1   7341.0 7411.0
## - CLM_FREQ    1   7347.3 7417.3
## - JOB         8   7363.1 7419.1
## - TIF         1   7359.4 7429.4
## - TRAVTIME    1   7360.8 7430.8
## - CAR_USE     1   7369.5 7439.5
## - MVR_PTS     1   7370.2 7440.2
## - CAR_TYPE    5   7391.7 7453.7
## - REVOKED     1   7393.7 7463.7
## - URBANICITY  1   7950.9 8020.9
##
## Step:  AIC=7370.8
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + PARENT1 + TRAVTIME +
##     CAR_USE + TIF + CLM_FREQ + REVOKED + MVR_PTS + INCOME + HOME_VAL +
##     BLUEBOOK + OLDCLAIM + MSTATUS + SEX + EDUCATION + JOB + CAR_TYPE +
##     URBANICITY
##
##              Df Deviance    AIC
## - SEX         1   7301.3 7369.3
## - YOJ         1   7302.7 7370.7
## <none>            7300.8 7370.8
```

```
## - HOMEKIDS    1   7303.3 7371.3
## - INCOME      1   7312.4 7380.4
## - HOME_VAL    1   7312.5 7380.5
## - PARENT1     1   7313.1 7381.1
## - OLDCLAIM    1   7313.7 7381.7
## - BLUEBOOK    1   7317.1 7385.1
## - EDUCATION   4   7326.6 7388.6
## - MSTATUS     1   7341.4 7409.4
## - KIDSDRIV    1   7341.4 7409.4
## - CLM_FREQ    1   7347.5 7415.5
## - JOB         8   7363.4 7417.4
## - TIF         1   7359.9 7427.9
## - TRAVTIME    1   7361.1 7429.1
## - CAR_USE     1   7369.7 7437.7
## - MVR_PTS     1   7370.5 7438.5
## - CAR_TYPE    5   7392.3 7452.3
## - REVOKED     1   7394.0 7462.0
## - URBANICITY  1   7951.3 8019.3
##
## Step:  AIC=7369.34
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + PARENT1 + TRAVTIME +
##     CAR_USE + TIF + CLM_FREQ + REVOKED + MVR_PTS + INCOME + HOME_VAL +
##     BLUEBOOK + OLDCLAIM + MSTATUS + EDUCATION + JOB + CAR_TYPE +
##     URBANICITY
##
##              Df Deviance    AIC
## - YOJ         1   7303.2 7369.2
## <none>            7301.3 7369.3
## - HOMEKIDS    1   7303.7 7369.7
## - HOME_VAL    1   7313.0 7379.0
## - INCOME      1   7313.1 7379.1
## - PARENT1     1   7313.6 7379.6
## - OLDCLAIM    1   7314.2 7380.2
## - EDUCATION   4   7327.0 7387.0
## - BLUEBOOK    1   7324.6 7390.6
## - KIDSDRIV    1   7341.9 7407.9
## - MSTATUS     1   7341.9 7407.9
## - CLM_FREQ    1   7348.2 7414.2
## - JOB         8   7363.6 7415.6
## - TIF         1   7360.4 7426.4
## - TRAVTIME    1   7361.7 7427.7
## - CAR_USE     1   7370.2 7436.2
## - MVR_PTS     1   7371.0 7437.0
## - REVOKED     1   7394.8 7460.8
## - CAR_TYPE    5   7409.3 7467.3
## - URBANICITY  1   7952.0 8018.0
##
## Step:  AIC=7369.16
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + PARENT1 + TRAVTIME + CAR_USE +
##     TIF + CLM_FREQ + REVOKED + MVR_PTS + INCOME + HOME_VAL +
##     BLUEBOOK + OLDCLAIM + MSTATUS + EDUCATION + JOB + CAR_TYPE +
##     URBANICITY
##
##              Df Deviance    AIC
```

```
## - HOMEKIDS    1   7305.0 7369.0
## <none>            7303.2 7369.2
## - HOME_VAL    1   7315.1 7379.1
## - INCOME      1   7315.7 7379.7
## - PARENT1     1   7315.8 7379.8
## - OLDCLAIM    1   7316.3 7380.3
## - EDUCATION   4   7328.6 7386.6
## - BLUEBOOK    1   7327.0 7391.0
## - KIDSDRIV    1   7344.3 7408.3
## - MSTATUS     1   7345.7 7409.7
## - CLM_FREQ    1   7350.1 7414.1
## - JOB         8   7365.5 7415.5
## - TIF         1   7362.5 7426.5
## - TRAVTIME    1   7363.3 7427.3
## - CAR_USE     1   7372.7 7436.7
## - MVR_PTS     1   7373.8 7437.8
## - REVOKED     1   7396.9 7460.9
## - CAR_TYPE    5   7411.5 7467.5
## - URBANICITY  1   7953.0 8017.0
##
## Step:  AIC=7369.01
## TARGET_FLAG ~ KIDSDRIV + PARENT1 + TRAVTIME + CAR_USE + TIF +
##     CLM_FREQ + REVOKED + MVR_PTS + INCOME + HOME_VAL + BLUEBOOK +
##     OLDCLAIM + MSTATUS + EDUCATION + JOB + CAR_TYPE + URBANICITY
##
##              Df Deviance    AIC
## <none>            7305.0 7369.0
## - INCOME      1   7317.2 7379.2
## - HOME_VAL    1   7317.3 7379.3
## - OLDCLAIM    1   7318.2 7380.2
## - EDUCATION   4   7330.8 7386.8
## - PARENT1     1   7329.0 7391.0
## - BLUEBOOK    1   7329.2 7391.2
## - MSTATUS     1   7346.0 7408.0
## - CLM_FREQ    1   7352.1 7414.1
## - JOB         8   7368.6 7416.6
## - KIDSDRIV    1   7362.3 7424.3
## - TIF         1   7364.0 7426.0
## - TRAVTIME    1   7364.7 7426.7
## - CAR_USE     1   7374.3 7436.3
## - MVR_PTS     1   7376.1 7438.1
## - REVOKED     1   7399.6 7461.6
## - CAR_TYPE    5   7413.8 7467.8
## - URBANICITY  1   7955.0 8017.0
##
## Call:  glm(formula = TARGET_FLAG ~ KIDSDRIV + PARENT1 + TRAVTIME + CAR_USE +
##     TIF + CLM_FREQ + REVOKED + MVR_PTS + INCOME + HOME_VAL +
##     BLUEBOOK + OLDCLAIM + MSTATUS + EDUCATION + JOB + CAR_TYPE +
##     URBANICITY, family = "binomial", data = full_bin_df)
##
## Coefficients:
##                   (Intercept)                       KIDSDRIV
##                    -2.985e+00                      4.187e-01
```

```
##                   PARENT1Yes                       TRAVTIME
##                    4.614e-01                      1.451e-02
##                CAR_USEPrivate                            TIF
##                    -7.566e-01                     -5.532e-02
##                     CLM_FREQ                     REVOKEDYes
##                    1.969e-01                      8.941e-01
##                      MVR_PTS                         INCOME
##                    1.140e-01                     -3.710e-06
##                     HOME_VAL                       BLUEBOOK
##                    -1.110e-06                     -2.298e-05
##                     OLDCLAIM                     MSTATUSYes
##                    -1.405e-05                     -5.003e-01
##            EDUCATIONBachelors           EDUCATIONHigh School
##                    -3.911e-01                      1.383e-02
##              EDUCATIONMasters                    EDUCATIONPhD
##                    -3.121e-01                     -1.882e-01
##                JOBBlue Collar                    JOBClerical
##                     3.060e-01                      4.135e-01
##                     JOBDoctor                  JOBHome Maker
##                    -4.482e-01                      2.784e-01
##                     JOBLawyer                     JOBManager
##                     9.984e-02                     -5.659e-01
##                JOBProfessional                     JOBStudent
##                     1.495e-01                      2.898e-01
##           CAR_TYPEPanel Truck                 CAR_TYPEPickup
##                     6.031e-01                      5.477e-01
##            CAR_TYPESports Car                    CAR_TYPESUV
##                     9.711e-01                      7.128e-01
##                   CAR_TYPEVan  URBANICITYHighly Urban/ Urban
##                     6.420e-01                      2.392e+00
##
## Degrees of Freedom: 8160 Total (i.e. Null);  8129 Residual
## Null Deviance:       9418
## Residual Deviance: 7305  AIC: 7369
```

The stepwise model has produced a model that is not better in regards to fit as evidenced by the residual deviance and AIC, but it does contain less coefficients, which may be ultimately a better choice for us.

Variable Importance: To assess the relative importance of individual predictors in the model, we can also look at the absolute value of the t-statistic for each model parameter.

	Overall
URBANICITYHighly Urban/ Urban	21.195625
REVOKEDYes	9.806389
CAR_TYPESports Car	9.041959
MVR_PTS	8.394753
CAR_TYPESUV	8.293251
CAR_USEPrivate	8.261928

it is interesting to note that these are the top six variables that is considered likely the most important. This may not ultimately impact our final model for logistic regression, but it is also interesting to look out for this type of information.

5.3 Model3: Logistic Regression Model using Transformed data

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + I(YOJ^1.6) +
##     PARENT1 + I(TRAVTIME^0.7) + CAR_USE + I(TIF^0.2) + RED_CAR +
##     CLM_FREQ + REVOKED + MVR_PTS + I(CAR_AGE^0.4) + I(INCOME^0.4) +
##     HOME_VAL + I(BLUEBOOK^0.5) + OLDCLAIM + MSTATUS + SEX + EDUCATION +
##     JOB + CAR_TYPE + URBANICITY, family = "binomial", data = full_bin_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5595  -0.7117  -0.3984   0.6201   3.1439
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.455e+00  4.078e-01  -3.567 0.000360 ***
## KIDSDRIV                       3.930e-01  6.144e-02   6.397 1.59e-10 ***
## AGE                           -2.059e-03  4.067e-03  -0.506 0.612705
## HOMEKIDS                       3.538e-02  3.772e-02   0.938 0.348204
## I(YOJ^1.6)                     8.961e-04  1.609e-03   0.557 0.577457
## PARENT1Yes                     3.850e-01  1.098e-01   3.506 0.000455 ***
## I(TRAVTIME^0.7)                6.032e-02  7.615e-03   7.921 2.35e-15 ***
## CAR_USEPrivate                -7.575e-01  9.192e-02  -8.241  < 2e-16 ***
## I(TIF^0.2)                    -9.424e-01  1.205e-01  -7.818 5.37e-15 ***
## RED_CARyes                    -1.201e-02  8.643e-02  -0.139 0.889493
## CLM_FREQ                       1.969e-01  2.858e-02   6.888 5.66e-12 ***
## REVOKEDYes                     8.885e-01  9.142e-02   9.719  < 2e-16 ***
## MVR_PTS                        1.131e-01  1.365e-02   8.290  < 2e-16 ***
## I(CAR_AGE^0.4)                -4.034e-02  4.667e-02  -0.864 0.387420
## I(INCOME^0.4)                 -8.746e-03  1.764e-03  -4.957 7.14e-07 ***
## HOME_VAL                      -1.029e-06  3.122e-07  -3.295 0.000983 ***
## I(BLUEBOOK^0.5)               -5.465e-03  1.214e-03  -4.502 6.74e-06 ***
## OLDCLAIM                      -1.411e-05  3.918e-06  -3.602 0.000316 ***
## MSTATUSYes                    -5.462e-01  8.204e-02  -6.658 2.77e-11 ***
## SEXM                           7.075e-02  1.109e-01   0.638 0.523523
## EDUCATIONBachelors            -3.071e-01  1.152e-01  -2.665 0.007699 **
## EDUCATIONHigh School           5.745e-02  9.555e-02   0.601 0.547717
## EDUCATIONMasters              -2.106e-01  1.720e-01  -1.224 0.220925
## EDUCATIONPhD                  -1.183e-01  2.058e-01  -0.575 0.565250
## JOBBlue Collar                 3.203e-01  1.854e-01   1.728 0.084031 .
## JOBClerical                    3.817e-01  1.963e-01   1.944 0.051870 .
## JOBDoctor                     -4.317e-01  2.671e-01  -1.617 0.105960
## JOBHome Maker                  1.847e-02  2.205e-01   0.084 0.933219
## JOBLawyer                      1.204e-01  1.694e-01   0.711 0.477124
## JOBManager                    -5.499e-01  1.713e-01  -3.209 0.001331 **
## JOBProfessional                1.600e-01  1.784e-01   0.897 0.369802
## JOBStudent                     1.529e-02  2.235e-01   0.068 0.945455
## CAR_TYPEPanel Truck            5.539e-01  1.567e-01   3.534 0.000410 ***
## CAR_TYPEPickup                 5.473e-01  1.009e-01   5.425 5.79e-08 ***
## CAR_TYPESports Car             1.009e+00  1.299e-01   7.766 8.07e-15 ***
## CAR_TYPESUV                    7.630e-01  1.103e-01   6.919 4.55e-12 ***
## CAR_TYPEVan                    6.446e-01  1.264e-01   5.098 3.43e-07 ***
## URBANICITYHighly Urban/ Urban  2.399e+00  1.130e-01  21.229  < 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: 9415.3  on 8159  degrees of freedom
## Residual deviance: 7274.2  on 8122  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 7350.2
##
## Number of Fisher Scoring iterations: 5
```

AIC has not improved in this model.

	Overall
URBANICITYHighly Urban/ Urban	21.229229
REVOKEDYes	9.719202
MVR_PTS	8.289974
CAR_USEPrivate	8.240706
I(TRAVTIME^0.7)	7.921222
I(TIF^0.2)	7.817932

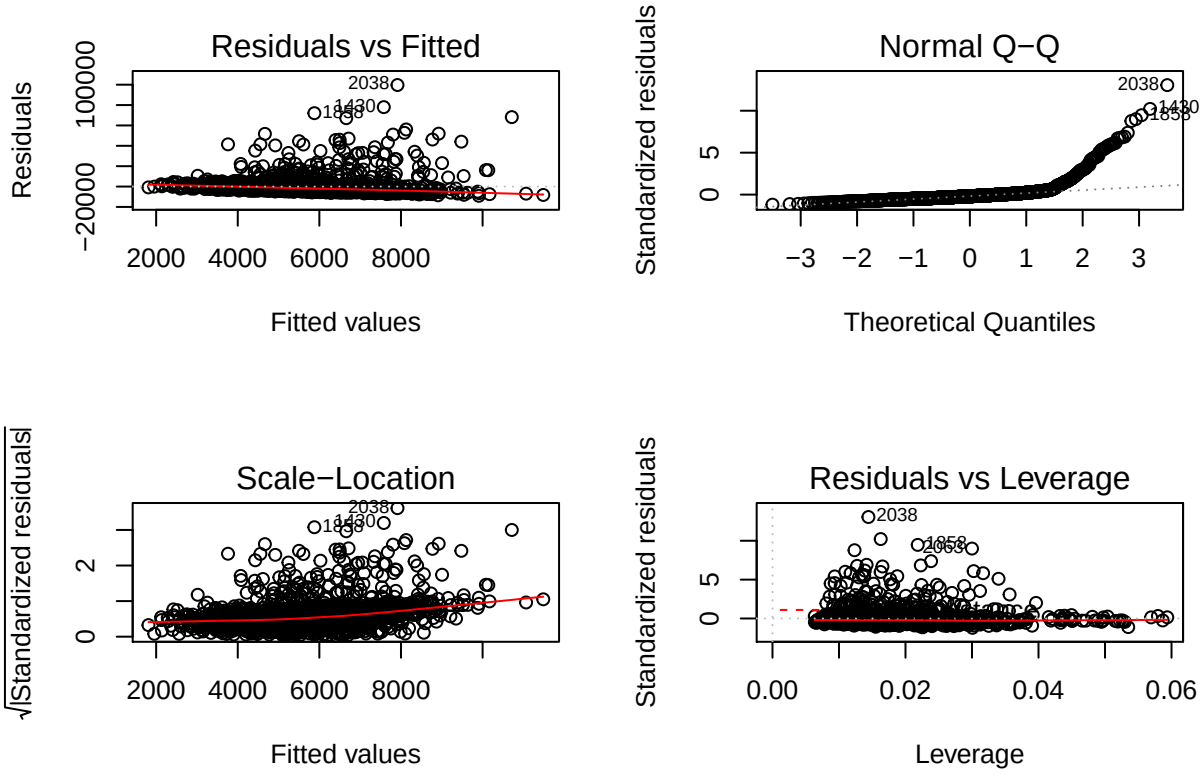
In both model 3 and model 2, URBANCITY seemed to have the highest variable importance.

5.4 Model4: Linear Regression Model with untransformed variables.

Lets develop a linear regression model for the amount paid if the person did get into an accident. Given that we are only looking at people who have ultimately crashed, we need to filter out the results for the zero dollar payout (as these are the people who have not crashed). We are interested in creating a model for the amount paid out in the event of a crash.

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = amt_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9069  -3182  -1483    478  99672
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.571e+03  1.987e+03   1.294   0.1959
## KIDSDRIV                    -1.810e+02  3.170e+02  -0.571   0.5680
## AGE                          1.930e+01  2.123e+01   0.909   0.3635
## HOMEKIDS                     2.176e+02  2.074e+02   1.049   0.2942
## YOJ                          1.947e+01  4.921e+01   0.396   0.6924
## PARENT1Yes                   2.746e+02  5.876e+02   0.467   0.6403
## TRAVTIME                     7.601e-01  1.108e+01   0.069   0.9453
## CAR_USEPrivate              -4.674e+02  5.219e+02  -0.896   0.3706
## TIF                         -1.575e+01  4.253e+01  -0.370   0.7112
## RED_CARyes                  -1.868e+02  4.967e+02  -0.376   0.7068
## CLM_FREQ                    -1.163e+02  1.581e+02  -0.736   0.4619
## REVOKEDYes                  -1.137e+03  5.167e+02  -2.201   0.0279 *
## MVR_PTS                      1.137e+02  6.858e+01   1.658   0.0975 .
## CAR_AGE                     -8.022e+01  4.085e+01  -1.964   0.0497 *
## INCOME                      -8.433e-03  6.728e-03  -1.253   0.2102
## HOME_VAL                     1.601e-03  1.918e-03   0.835   0.4037
## BLUEBOOK                     1.248e-01  3.054e-02   4.086 4.55e-05 ***
## OLDCLAIM                     2.519e-02  2.265e-02   1.112   0.2661
## MSTATUSYes                  -7.316e+02  4.882e+02  -1.498   0.1342
## SEXM                         1.412e+03  6.566e+02   2.150   0.0317 *
## EDUCATIONBachelors           1.559e+02  6.337e+02   0.246   0.8057
## EDUCATIONHigh School        -4.282e+02  5.140e+02  -0.833   0.4049
## EDUCATIONMasters             1.042e+03  1.073e+03   0.971   0.3318
## EDUCATIONPhD                 2.233e+03  1.301e+03   1.716   0.0863 .
## JOBBlue Collar               5.384e+02  1.146e+03   0.470   0.6387
## JOBClerical                  3.475e+02  1.204e+03   0.289   0.7728
## JOBDoctor                   -2.070e+03  1.763e+03  -1.174   0.2405
## JOBHome Maker               -2.549e+01  1.267e+03  -0.020   0.9839
## JOBLawyer                    3.221e+02  1.030e+03   0.313   0.7546
## JOBManager                  -7.829e+02  1.066e+03  -0.734   0.4628
## JOBProfessional              1.086e+03  1.129e+03   0.962   0.3363
## JOBStudent                   8.088e+01  1.286e+03   0.063   0.9499
## CAR_TYPEPanel Truck         -6.420e+02  9.609e+02  -0.668   0.5041
## CAR_TYPEPickup              -7.582e+01  5.970e+02  -0.127   0.8990
## CAR_TYPESports Car           1.070e+03  7.505e+02   1.425   0.1542
## CAR_TYPESUV                  9.086e+02  6.670e+02   1.362   0.1733
## CAR_TYPEVan                  4.957e+01  7.714e+02   0.064   0.9488
```

```
## URBANICITYHighly Urban/ Urban  9.313e+01  7.565e+02   0.123   0.9020
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7693 on 2115 degrees of freedom
## Multiple R-squared:  0.02986,    Adjusted R-squared:  0.01289
## F-statistic: 1.759 on 37 and 2115 DF,  p-value: 0.003273
```



VIF

```
##                   KIDSDRIV                        AGE
##                   1.436027                   1.496272
##                   HOMEKIDS                        YOJ
##                   2.234314                   1.694815
##                  PARENT1Yes                   TRAVTIME
##                   2.162926                   1.031280
##                CAR_USEPrivate                      TIF
##                   2.474947                   1.017530
##                   RED_CARyes                   CLM_FREQ
##                   1.833052                   1.416592
##                   REVOKEDYes                    MVR_PTS
##                   1.587162                   1.137529
##                    CAR_AGE                     INCOME
##                   1.888068                   2.859254
##                   HOME_VAL                    BLUEBOOK
##                   1.855777                   2.336869
##                   OLDCLAIM                 MSTATUSYes
```

```
##                       1.891359                       2.166634
##                           SEXM             EDUCATIONBachelors
##                       3.876110                       2.686288
##           EDUCATIONHigh School               EDUCATIONMasters
##                       2.236481                       5.394925
##                    EDUCATIONPhD                 JOBBlue Collar
##                       3.367247                       9.932732
##                     JOBClerical                      JOBDoctor
##                       7.516017                       1.502302
##                   JOBHome Maker                      JOBLawyer
##                       4.472591                       2.548098
##                     JOBManager                 JOBProfessional
##                       2.463689                       4.712299
##                     JOBStudent             CAR_TYPEPanel Truck
##                       6.516638                       2.547406
##                  CAR_TYPEPickup              CAR_TYPESports Car
##                       2.118701                       2.484660
##                     CAR_TYPESUV                    CAR_TYPEVan
##                       3.491773                       1.832417
## URBANICITYHighly Urban/ Urban
##                       1.052710
```

The adjusted R-squared value is quite low here: 0.01289. The model does not appear to explain the variation in the response variable quite well.

There appears to be significant outliers in this dataset, and given the skew in this data (as demonstrated in the Q-Q plot), this may be affecting the data quite adversely. Would the response variable benefit from a transformation? Let's try performing a Box Cox transformation to the response variable and see what the results are.

value of VIF is high which suggests to check for multicollinearity.

5.5 Model5: Linear Regression Model with transformation

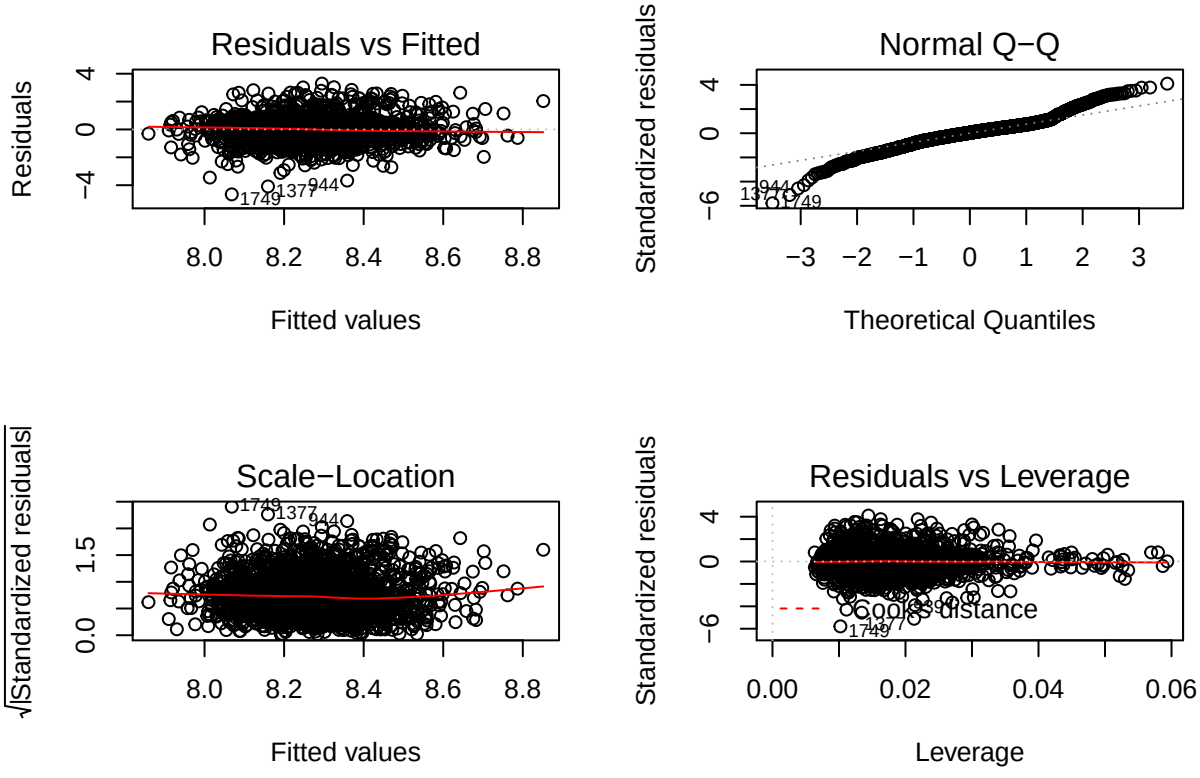
Let's try Box Cox transformation to the response variable and see what the results are.

```
## Box-Cox Transformation
##
## 2153 data points used to estimate Lambda
##
## Input data summary:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     30.28   2610.00   4104.00   5702.00   5787.00 107600.00
##
## Largest/Smallest: 3550
## Sample Skewness: 5.63
##
## Estimated Lambda: 0
## With fudge factor, Lambda = 0 will be used for transformations
```

With lambda = 0, let's create a logarithmic transformation of the response variable.

```
##
## Call:
## lm(formula = l_TARGET_AMT ~ ., data = amt_df_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6581 -0.4056  0.0305  0.4078  3.2910
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   7.936e+00  2.089e-01  37.989  < 2e-16 ***
## KIDSDRIV                     -3.251e-02  3.333e-02  -0.975 0.329508
## AGE                           1.935e-03  2.233e-03   0.867 0.386172
## HOMEKIDS                      2.272e-02  2.180e-02   1.042 0.297539
## YOJ                          -8.763e-04  5.174e-03  -0.169 0.865528
## PARENT1Yes                    2.454e-02  6.178e-02   0.397 0.691261
## TRAVTIME                     -2.795e-04  1.165e-03  -0.240 0.810509
## CAR_USEPrivate               -1.420e-02  5.487e-02  -0.259 0.795832
## TIF                          -1.961e-03  4.472e-03  -0.438 0.661158
## RED_CARyes                    2.146e-02  5.222e-02   0.411 0.681119
## CLM_FREQ                     -3.659e-02  1.662e-02  -2.201 0.027827 *
## REVOKEDYes                   -9.675e-02  5.433e-02  -1.781 0.075071 .
## MVR_PTS                       1.477e-02  7.211e-03   2.049 0.040586 *
## CAR_AGE                      -1.240e-03  4.295e-03  -0.289 0.772920
## INCOME                       -1.205e-06  7.074e-07  -1.704 0.088567 .
## HOME_VAL                      2.079e-08  2.016e-07   0.103 0.917876
## BLUEBOOK                      1.200e-05  3.212e-06   3.735 0.000193 ***
## OLDCLAIM                      4.489e-06  2.381e-06   1.885 0.059540 .
## MSTATUSYes                   -8.309e-02  5.133e-02  -1.619 0.105683
## SEXM                          9.462e-02  6.904e-02   1.370 0.170695
## EDUCATIONBachelors           -3.515e-02  6.663e-02  -0.528 0.597834
## EDUCATIONHigh School          7.069e-03  5.405e-02   0.131 0.895957
## EDUCATIONMasters              1.406e-01  1.128e-01   1.246 0.212834
## EDUCATIONPhD                  2.378e-01  1.368e-01   1.738 0.082299 .
## JOBBlue Collar                6.276e-02  1.205e-01   0.521 0.602631
```

```
## JOBClerical                      5.525e-02  1.266e-01   0.437 0.662440
## JOBDoctor                       -3.585e-02  1.854e-01  -0.193 0.846668
## JOBHome Maker                   -5.047e-02  1.332e-01  -0.379 0.704811
## JOBLawyer                       -1.214e-02  1.083e-01  -0.112 0.910737
## JOBManager                       1.987e-02  1.121e-01   0.177 0.859314
## JOBProfessional                  1.083e-01  1.187e-01   0.912 0.361655
## JOBStudent                       1.316e-02  1.352e-01   0.097 0.922510
## CAR_TYPEPanel Truck              3.761e-03  1.010e-01   0.037 0.970313
## CAR_TYPEPickup                   2.818e-02  6.277e-02   0.449 0.653494
## CAR_TYPESports Car               5.762e-02  7.891e-02   0.730 0.465372
## CAR_TYPESUV                      9.353e-02  7.013e-02   1.334 0.182478
## CAR_TYPEVan                     -1.152e-02  8.111e-02  -0.142 0.887095
## URBANICITYHighly Urban/ Urban    5.586e-02  7.955e-02   0.702 0.482577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8089 on 2115 degrees of freedom
## Multiple R-squared:  0.02605,    Adjusted R-squared:  0.009017
## F-statistic: 1.529 on 37 and 2115 DF,  p-value: 0.02192
```



VIF

```
##                   KIDSDRIV                        AGE
##                   1.436027                   1.496272
##                   HOMEKIDS                        YOJ
##                   2.234314                   1.694815
##                  PARENT1Yes                   TRAVTIME
```

```
##                     2.162926                     1.031280
##                CAR_USEPrivate                          TIF
##                     2.474947                     1.017530
##                    RED_CARyes                     CLM_FREQ
##                     1.833052                     1.416592
##                    REVOKEDYes                      MVR_PTS
##                     1.587162                     1.137529
##                      CAR_AGE                       INCOME
##                     1.888068                     2.859254
##                     HOME_VAL                     BLUEBOOK
##                     1.855777                     2.336869
##                     OLDCLAIM                   MSTATUSYes
##                     1.891359                     2.166634
##                         SEXM             EDUCATIONBachelors
##                     3.876110                     2.686288
##           EDUCATIONHigh School               EDUCATIONMasters
##                     2.236481                     5.394925
##                  EDUCATIONPhD                 JOBBlue Collar
##                     3.367247                     9.932732
##                   JOBClerical                    JOBDoctor
##                     7.516017                     1.502302
##                  JOBHome Maker                    JOBLawyer
##                     4.472591                     2.548098
##                    JOBManager                JOBProfessional
##                     2.463689                     4.712299
##                    JOBStudent            CAR_TYPEPanel Truck
##                     6.516638                     2.547406
##                CAR_TYPEPickup             CAR_TYPESports Car
##                     2.118701                     2.484660
##                   CAR_TYPESUV                  CAR_TYPEVan
##                     3.491773                     1.832417
## URBANICITYHighly Urban/ Urban
##                     1.052710
```

The adjusted R-squared is even worse.The model is certainly better, and the distribution does indeed appear to be improved. However, with a poor adjusted R-squared value, it may benefit us to also look at the transformed predictive values as well.

value of VIF is high which suggests to check for multicollinearity.

Check for Multicollinearity

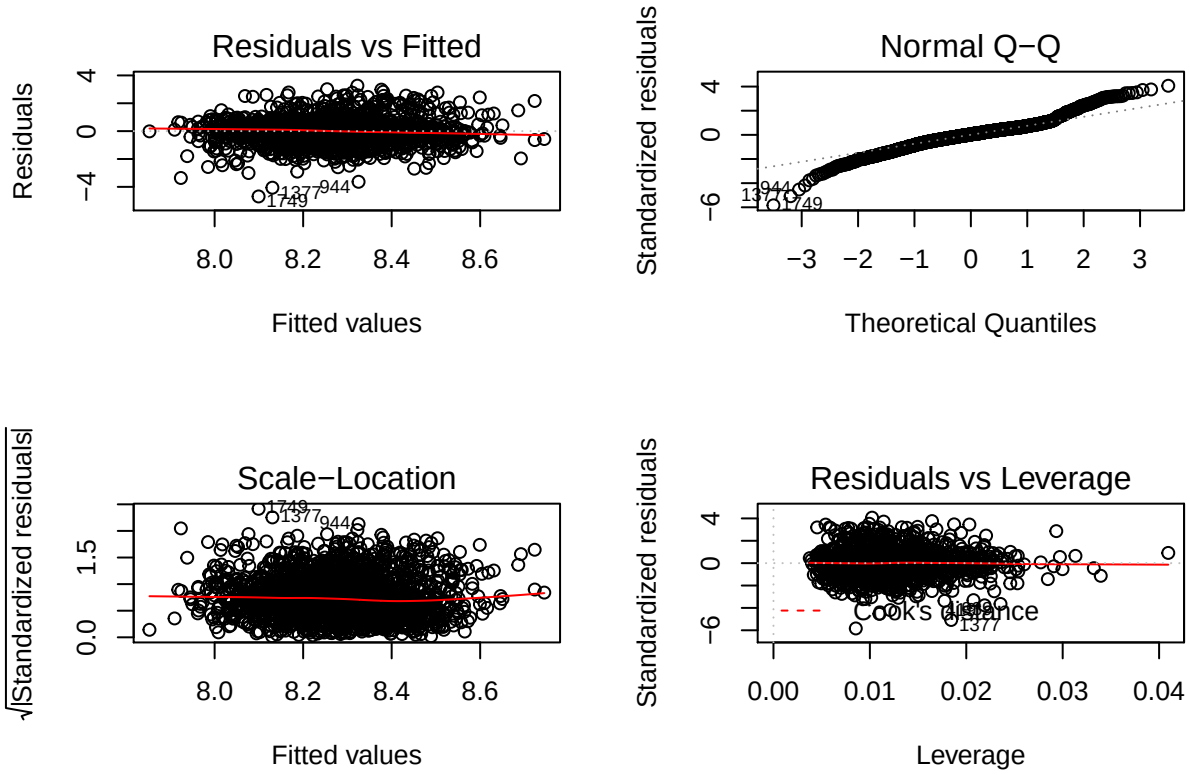
```
##           KIDSDRIV   AGE HOMEKIDS   YOJ TRAVTIME   TIF CLM_FREQ MVR_PTS
## KIDSDRIV      1.00  0.00     0.48  0.07    -0.02 -0.01     0.01    0.02
## AGE           0.00  1.00    -0.41  0.12     0.06 -0.02     0.02   -0.03
## HOMEKIDS      0.48 -0.41     1.00  0.07    -0.04  0.00    -0.01    0.04
## YOJ           0.07  0.12     0.07  1.00     0.03 -0.02    -0.02   -0.04
## TRAVTIME     -0.02  0.06    -0.04  0.03     1.00 -0.02     0.03    0.02
## TIF          -0.01 -0.02     0.00 -0.02    -0.02  1.00     0.02   -0.03
## CLM_FREQ      0.01  0.02    -0.01 -0.02     0.03  0.02     1.00    0.30
## MVR_PTS       0.02 -0.03     0.04 -0.04     0.02 -0.03     0.30    1.00
## CAR_AGE       0.01  0.15    -0.07  0.05     0.06 -0.01     0.05   -0.04
## INCOME        0.05  0.17    -0.10  0.35     0.05 -0.03     0.00   -0.05
## HOME_VAL      0.04  0.19    -0.06  0.30     0.04 -0.02    -0.01   -0.07
## BLUEBOOK      0.04  0.15    -0.08  0.16     0.03 -0.01    -0.02   -0.03
## OLDCLAIM      0.01  0.02     0.01  0.02    -0.03 -0.01     0.41    0.15
```

```
##          CAR_AGE INCOME HOME_VAL BLUEBOOK OLDCLAIM
## KIDSDRIV    0.01   0.05     0.04     0.04     0.01
## AGE         0.15   0.17     0.19     0.15     0.02
## HOMEKIDS   -0.07  -0.10    -0.06    -0.08     0.01
## YOJ         0.05   0.35     0.30     0.16     0.02
## TRAVTIME    0.06   0.05     0.04     0.03    -0.03
## TIF        -0.01  -0.03    -0.02    -0.01    -0.01
## CLM_FREQ    0.05   0.00    -0.01    -0.02     0.41
## MVR_PTS    -0.04  -0.05    -0.07    -0.03     0.15
## CAR_AGE     1.00   0.40     0.16     0.18     0.02
## INCOME      0.40   1.00     0.45     0.42    -0.02
## HOME_VAL    0.16   0.45     1.00     0.21    -0.01
## BLUEBOOK    0.18   0.42     0.21     1.00    -0.03
## OLDCLAIM    0.02  -0.02    -0.01    -0.03     1.00
```

5.6 Model6: Linear Regression Model with transformed variabled and elimination by VIF

Based on VIF and the correlation matrix, we can start to eliminate some variables. In this case, let's remove HOMEKIDS, EDUCATION, and JOB. Let's create model with the elimination of these variables and with the addition of the other transformed variables.

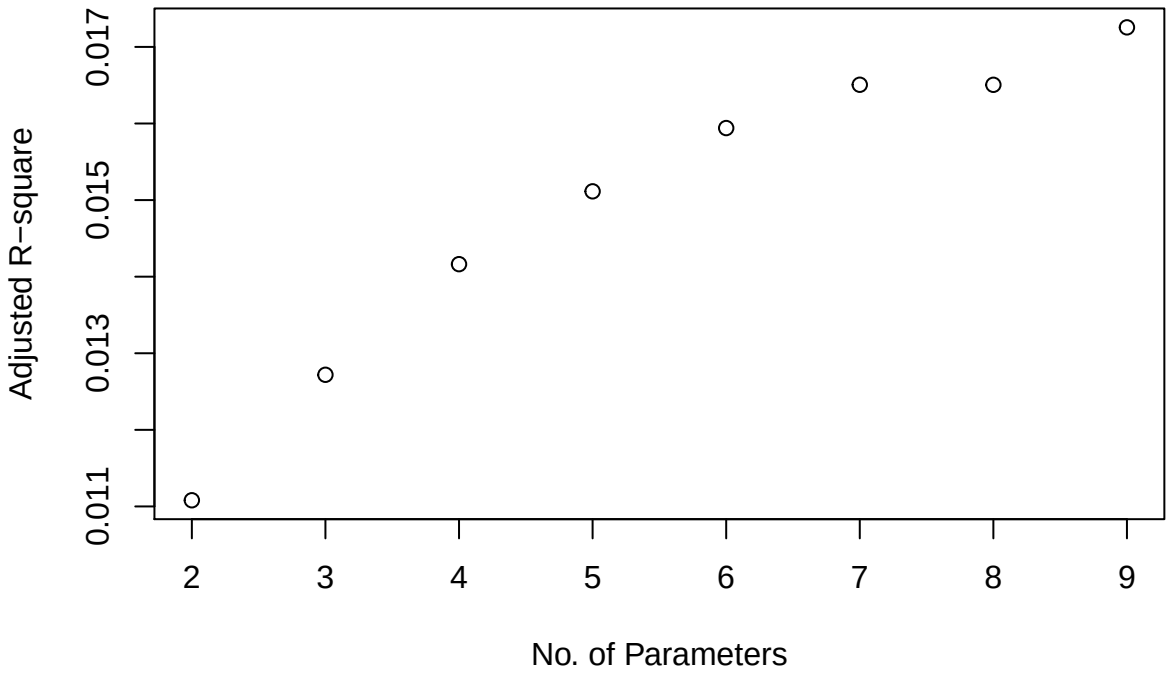
```
##
## Call:
## lm(formula = l_TARGET_AMT ~ KIDSDRIV + AGE + I(YOJ^1.6) + PARENT1 +
##     I(TRAVTIME^0.7) + CAR_USE + I(TIF^0.2) + RED_CAR + CLM_FREQ +
##     REVOKED + MVR_PTS + I(CAR_AGE^0.4) + I(INCOME^0.4) + HOME_VAL +
##     I(BLUEBOOK^0.5) + OLDCLAIM + MSTATUS + SEX + CAR_TYPE + URBANICITY,
##     data = amt_df_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6889 -0.3988  0.0301  0.4088  3.2635
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    7.823e+00  1.921e-01  40.726  < 2e-16 ***
## KIDSDRIV                      -1.966e-02  2.948e-02  -0.667   0.5049
## AGE                            1.043e-03  2.037e-03   0.512   0.6086
## I(YOJ^1.6)                    -1.551e-04  9.224e-04  -0.168   0.8665
## PARENT1Yes                     5.030e-02  5.530e-02   0.910   0.3631
## I(TRAVTIME^0.7)               -1.526e-03  4.657e-03  -0.328   0.7432
## CAR_USEPrivate                 4.707e-03  4.200e-02   0.112   0.9108
## I(TIF^0.2)                    -4.900e-02  7.261e-02  -0.675   0.4999
## RED_CARyes                     1.756e-02  5.194e-02   0.338   0.7354
## CLM_FREQ                      -3.598e-02  1.648e-02  -2.183   0.0291 *
## REVOKEDYes                    -8.875e-02  5.396e-02  -1.645   0.1002
## MVR_PTS                        1.568e-02  7.153e-03   2.192   0.0285 *
## I(CAR_AGE^0.4)                 6.700e-03  2.358e-02   0.284   0.7763
## I(INCOME^0.4)                 -1.959e-04  7.838e-04  -0.250   0.8026
## HOME_VAL                      -1.399e-08  1.935e-07  -0.072   0.9423
## I(BLUEBOOK^0.5)                3.225e-03  7.199e-04   4.480 7.85e-06 ***
## OLDCLAIM                       4.397e-06  2.367e-06   1.858   0.0634 .
## MSTATUSYes                    -5.864e-02  4.863e-02  -1.206   0.2280
## SEXM                           1.068e-01  6.714e-02   1.591   0.1119
## CAR_TYPEPanel Truck            1.423e-02  9.152e-02   0.155   0.8765
## CAR_TYPEPickup                 4.366e-02  6.112e-02   0.714   0.4750
## CAR_TYPESports Car             7.125e-02  7.806e-02   0.913   0.3615
## CAR_TYPESUV                    9.717e-02  6.883e-02   1.412   0.1582
## CAR_TYPEVan                   -2.347e-02  7.843e-02  -0.299   0.7648
## URBANICITYHighly Urban/ Urban  4.652e-02  7.866e-02   0.591   0.5543
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8069 on 2127 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.02472,    Adjusted R-squared:  0.01371
## F-statistic: 2.246 on 24 and 2127 DF,  p-value: 0.0004957
```



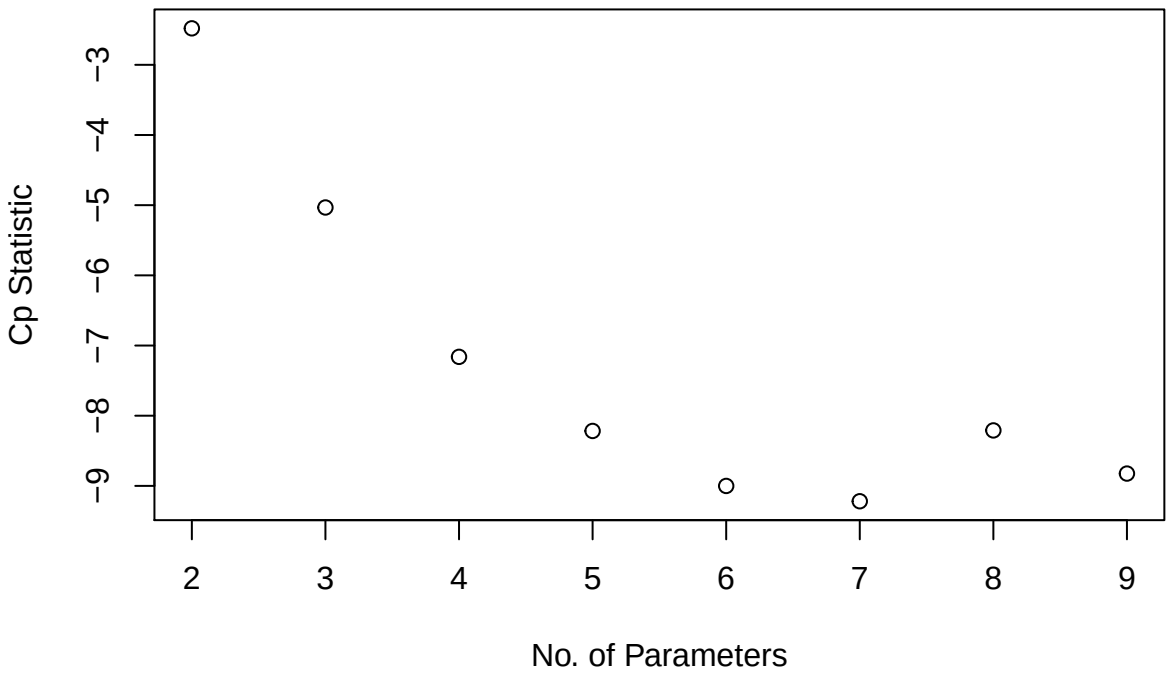
So far, this model has outperformed the other models with a modest increase in the adjusted R-square to 0.01371.

5.7 Model7: Linear Regressions with LEAPS

We will be using the leaps package in this model. For each size of model p, it finds the variables that produces the minimum RSS.




```
## [1] "How many variables that maximizes the adjusted R-squared value? 8"
```



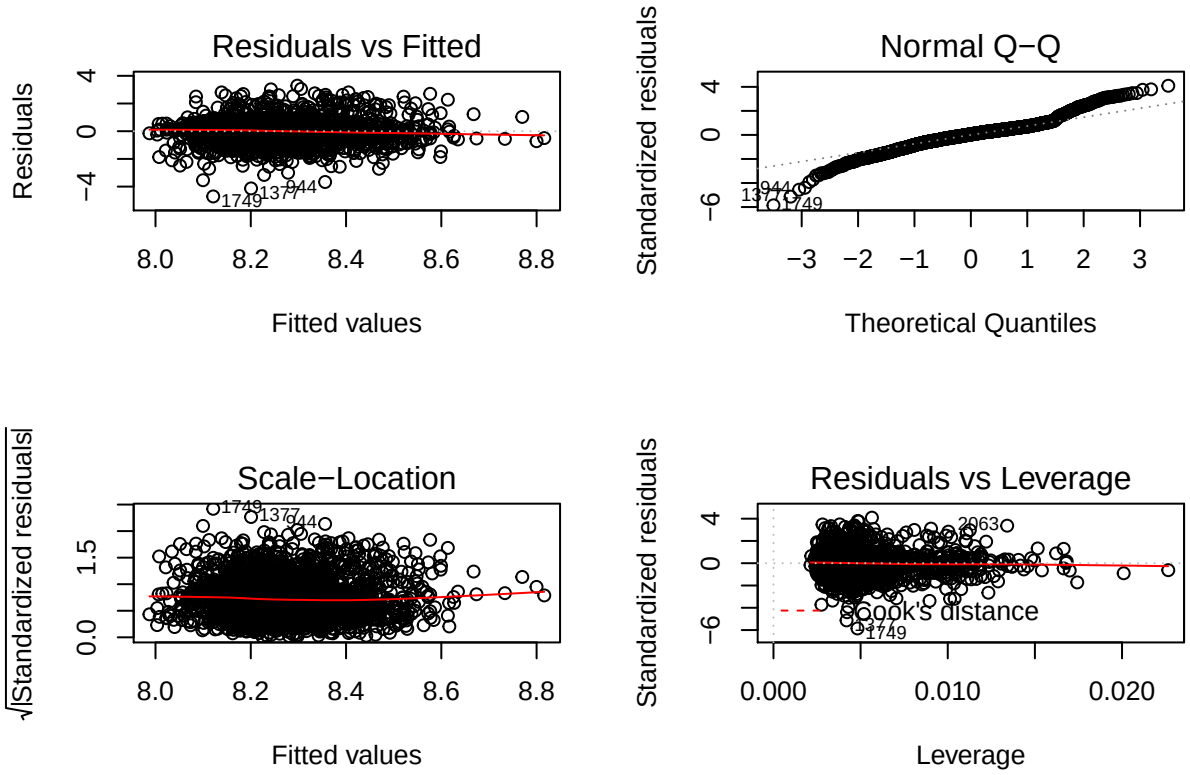
Looks like according to the model that would satisfy the Cp Statistic and increases the adjusted R-squared would be 8 variables. Which variables should be included?

```
##                   (Intercept)                     KIDSDRIV
##                          TRUE                        FALSE
##                           AGE                     HOMEKIDS
##                         FALSE                        FALSE
##                           YOJ                   PARENT1Yes
##                         FALSE                        FALSE
##                      TRAVTIME                CAR_USEPrivate
##                         FALSE                        FALSE
##                           TIF                   RED_CARyes
##                         FALSE                        FALSE
##                      CLM_FREQ                   REVOKEDYes
##                          TRUE                         TRUE
##                       MVR_PTS                      CAR_AGE
##                          TRUE                        FALSE
##                        INCOME                     HOME_VAL
##                         FALSE                        FALSE
##                     BLUEBOOK                     OLDCLAIM
##                          TRUE                         TRUE
##                    MSTATUSYes                        SEXM
##                          TRUE                         TRUE
##             EDUCATIONBachelors          EDUCATIONHigh School
##                          TRUE                        FALSE
##               EDUCATIONMasters                 EDUCATIONPhD
##                         FALSE                        FALSE
```

```
##                JOBBlue Collar                   JOBClerical
##                         FALSE                         FALSE
##                     JOBDoctor                JOBHome Maker
##                         FALSE                         FALSE
##                     JOBLawyer                   JOBManager
##                         FALSE                         FALSE
##               JOBProfessional                   JOBStudent
##                         FALSE                         FALSE
##           CAR_TYPEPanel Truck               CAR_TYPEPickup
##                         FALSE                         FALSE
##            CAR_TYPESports Car                  CAR_TYPESUV
##                         FALSE                         FALSE
##                  CAR_TYPEVan URBANICITYHighly Urban/ Urban
##                         FALSE                         FALSE
```

So this includes the Intercept, CLM_FREQ, MVR_PTS, BLUEBOOK, MSTATUS, EDUCATION, SEX, CAR_AGE. Let's create a Linear Regression Model from just using these variables and the logairthmic response variable of TARGET_AMT.

```
##
## Call:
## lm(formula = l_TARGET_AMT ~ CLM_FREQ + MVR_PTS + BLUEBOOK + MSTATUS +
##     EDUCATION + SEX + CAR_AGE, data = amt_df_trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7107 -0.4026  0.0361  0.4002  3.2884
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           8.139e+00  5.970e-02 136.330  < 2e-16 ***
## CLM_FREQ             -2.233e-02  1.464e-02  -1.525   0.1274
## MVR_PTS               1.728e-02  7.077e-03   2.442   0.0147 *
## BLUEBOOK              9.790e-06  2.192e-06   4.465 8.41e-06 ***
## MSTATUSYes           -7.726e-02  3.499e-02  -2.208   0.0273 *
## EDUCATIONBachelors   -5.064e-02  5.854e-02  -0.865   0.3871
## EDUCATIONHigh School  5.173e-03  5.035e-02   0.103   0.9182
## EDUCATIONMasters      4.416e-02  7.520e-02   0.587   0.5571
## EDUCATIONPhD          8.359e-02  9.567e-02   0.874   0.3823
## SEXM                  5.881e-02  3.516e-02   1.673   0.0945 .
## CAR_AGE              -1.116e-03  4.252e-03  -0.262   0.7930
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8064 on 2142 degrees of freedom
## Multiple R-squared:  0.01973,    Adjusted R-squared:  0.01515
## F-statistic: 4.311 on 10 and 2142 DF,  p-value: 5.49e-06
```



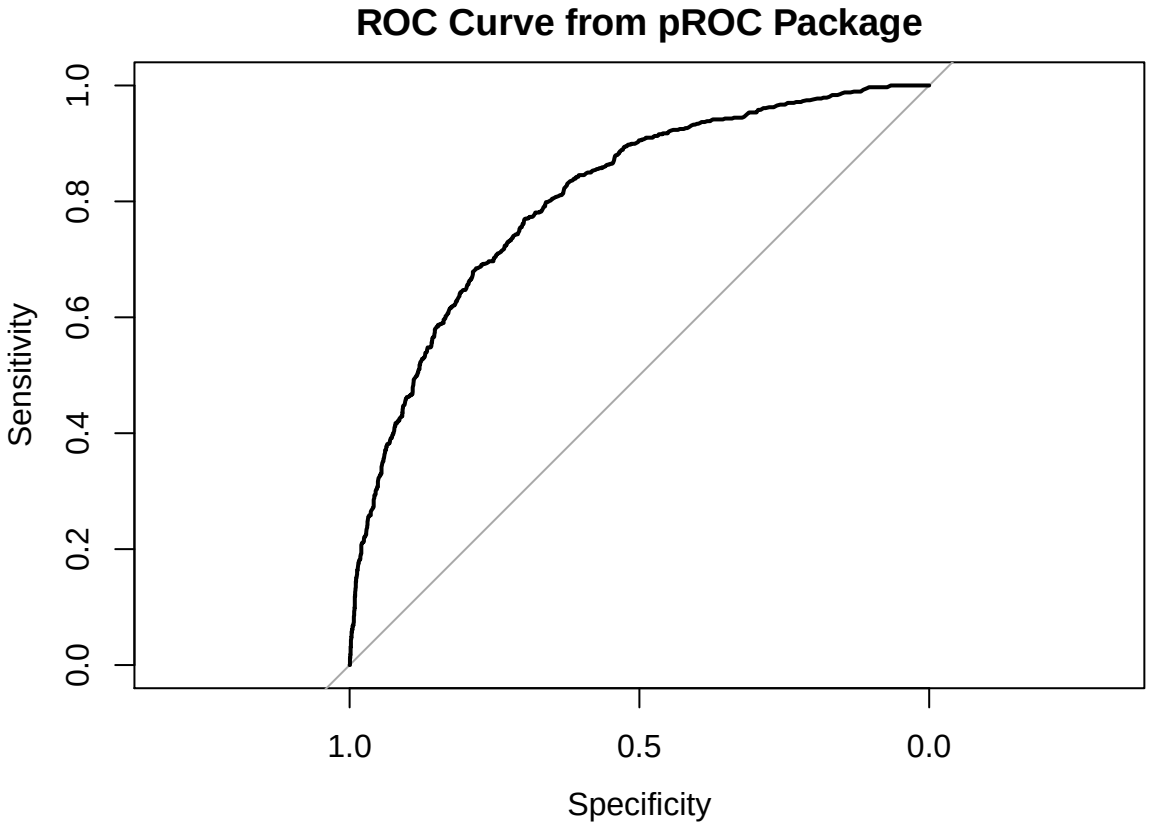
This Model appears to be further improved from previous model, with an adjusted R-square of 0.01515.

Based on the above results, let me choose Model2 and Model7 as final Models. We had already evaluated the performance strength on the linear regression models by using adjusted R-square values, so let's focus on evaluating the logistic regression model.

```
## Confusion Matrix and Statistics
##
##
##        0    1
##   0 1296  180
##   1  486  486
##
##                Accuracy : 0.7279
##                  95% CI : (0.7098, 0.7455)
##     No Information Rate : 0.7279
##     P-Value [Acc > NIR] : 0.5104
##
##                   Kappa : 0.3995
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7297
##             Specificity : 0.7273
##          Pos Pred Value : 0.5000
##          Neg Pred Value : 0.8780
##              Prevalence : 0.2721
##          Detection Rate : 0.1985
```

```
##    Detection Prevalence : 0.3971
##       Balanced Accuracy : 0.7285
##
##        'Positive' Class : 1
##
```

```
##          Sensitivity          Specificity       Pos Pred Value
##            0.7297297            0.7272727            0.5000000
##       Neg Pred Value            Precision               Recall
##            0.8780488            0.5000000            0.7297297
##                   F1           Prevalence       Detection Rate
##            0.5934066            0.2720588            0.1985294
## Detection Prevalence    Balanced Accuracy
##            0.3970588            0.7285012
```




```
## Area under the curve: 0.8065
```

The accuracy is 75%. I will use this for Prediction.

6 Prediction

```
##   TARGET_FLAG KIDSDRIV AGE HOMEKIDS YOJ PARENT1 TRAVTIME    CAR_USE TIF
## 1          NA        0  48        0  11      No       26    Private   1
## 2          NA        1  40        1  11     Yes       21    Private   6
## 3          NA        0  44        2  12     Yes       30 Commercial  10
## 4          NA        0  35        2  NA     Yes       74    Private   6
```

```
## 5          NA        0  59        0  12      No       45    Private   1
## 6          NA        0  46        0  14      No        7 Commercial   1
##   RED_CAR CLM_FREQ REVOKED MVR_PTS CAR_AGE INCOME HOME_VAL BLUEBOOK
## 1     yes        0      No       2      10  52881        0    21970
## 2      no        1      No       2       1  50815        0    18930
## 3      no        0      No       0      10  43486        0     5900
## 4      no        0     Yes       0       4  21204        0     9230
## 5     yes        2      No       4       1  87460        0    15420
## 6      no        1      No       2      12     NA   207519    25660
##   OLDCLAIM MSTATUS SEX   EDUCATION          JOB    CAR_TYPE
## 1        0      No   M   Bachelors      Manager         Van
## 2     3295      No   M High School      Manager     Minivan
## 3        0      No   F High School  Blue Collar         SUV
## 4        0      No   M High School     Clerical      Pickup
## 5    44857      No   M High School      Manager     Minivan
## 6     2119     Yes   M   Bachelors Professional Panel Truck
##            URBANICITY
## 1 Highly Urban/ Urban
## 2 Highly Urban/ Urban
## 3 Highly Rural/ Rural
## 4 Highly Rural/ Rural
## 5 Highly Urban/ Urban
## 6 Highly Urban/ Urban
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	0	0	0	NA	1	1	0	0	0	1	NA	0	0
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
NA	1	0	1	1	0	1	0	1	1	1	1	1	0	0
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
0	1	0	0	0	0	0	0	0	1	1	1	0	1	0
46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
0	0	1	0	1	0	1	1	0	NA	0	1	1	0	1
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
0	0	1	0	0	1	1	1	0	0	NA	0	1	1	1
76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
0	1	1	0	0	1	1	1	0	1	1	1	1	0	NA
91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
0	0	NA	0	0	0	0	1	1	0	1	NA	1	1	NA
106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
0	0	0	1	0	1	0	0	0	NA	0	0	1	1	0
121	122	123	124	125	126	127	128	129	130	131	132	133	134	135
0	1	1	NA	0	1	1	0	0	0	0	0	0	0	0
136	137	138	139	140	141	142	143	144	145	146	147	148	149	150
NA	1	1	0	0	0	1	0	0	0	1	0	NA	1	NA
151	152	153	154	155	156	157	158	159	160	161	162	163	164	165
1	NA	NA	1	1	1	NA	0	1	1	0	0	0	0	1
166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
0	0	NA	1	0	0	1	NA	1	1	1	NA	1	1	1
181	182	183	184	185	186	187	188	189	190	191	192	193	194	195
1	0	0	1	0	1	0	0	0	0	1	1	NA	1	0
196	197	198	199	200	201	202	203	204	205	206	207	208	209	210
1	1	0	0	0	0	1	0	0	0	0	NA	0	0	0
211	212	213	214	215	216	217	218	219	220	221	222	223	224	225
0	0	1	1	0	1	1	0	0	0	0	0	1	1	0

```
##  226  227  228  229  230  231  232  233  234  235  236  237  238  239  240
##    1    1    1    1    0    0    0    1    0    0    0    0    0    0    1
##  241  242  243  244  245  246  247  248  249  250  251  252  253  254  255
##    0   NA    1    0    0    1    1    1    0    1   NA   NA    0    1    1
##  256  257  258  259  260  261  262  263  264  265  266  267  268  269  270
##    1   NA    0    1    0    0    0    0   NA    0    0    0    0    1    1
##  271  272  273  274  275  276  277  278  279  280  281  282  283  284  285
##    1    0    1   NA    0    0   NA   NA    0    0    0    0    0    1    1
##  286  287  288  289  290  291  292  293  294  295  296  297  298  299  300
##    1    0    1    1    1    1    0    0    1    0    1    0    1    0    0
##  301  302  303  304  305  306  307  308  309  310  311  312  313  314  315
##    0    0    1    1    1    1    0    1    0    0   NA   NA    0    1    0
##  316  317  318  319  320  321  322  323  324  325  326  327  328  329  330
##    0    0    0    1    0    0    1    0    0    1    1   NA   NA   NA    0
##  331  332  333  334  335  336  337  338  339  340  341  342  343  344  345
##   NA   NA    1    0    0    1   NA    1   NA    0    1    1    1    1    0
##  346  347  348  349  350  351  352  353  354  355  356  357  358  359  360
##    0    0    0    0    0    0    1   NA    1    0    1    1    0    0    0
##  361  362  363  364  365  366  367  368  369  370  371  372  373  374  375
##    1    0    0    1    0    1    0    1   NA    0    0    0    1    0    0
##  376  377  378  379  380  381  382  383  384  385  386  387  388  389  390
##    1    0   NA    0    0    0    1    1    1    0   NA   NA    0    0    1
##  391  392  393  394  395  396  397  398  399  400  401  402  403  404  405
##    0    0    0    0    0    1    0    1    1   NA    1   NA    0    0    0
##  406  407  408  409  410  411  412  413  414  415  416  417  418  419  420
##    0    0    1    0    0   NA    1    1    0    1    1    0    1    0    0
##  421  422  423  424  425  426  427  428  429  430  431  432  433  434  435
##    1    1    1    0   NA    1    0    0    1    1    1    0    0   NA    1
##  436  437  438  439  440  441  442  443  444  445  446  447  448  449  450
##    1    0    0    0    0    0    0    0    0    0    0    0    0    1    1
##  451  452  453  454  455  456  457  458  459  460  461  462  463  464  465
##    0    0    1    0   NA    1    1    1    0    1    0    0    0    1    0
##  466  467  468  469  470  471  472  473  474  475  476  477  478  479  480
##    0    1    1    0   NA   NA    1    0    0    0    0    1    1    0   NA
##  481  482  483  484  485  486  487  488  489  490  491  492  493  494  495
##   NA    0    0    1    1    1    0    1    0    1    1    0    0    1    0
##  496  497  498  499  500  501  502  503  504  505  506  507  508  509  510
##    1   NA    0   NA    1    0   NA    1   NA    1    0    1    0    0    0
##  511  512  513  514  515  516  517  518  519  520  521  522  523  524  525
##    0    0   NA    0    0    0   NA    1    1    1    0    0    0    0    0
##  526  527  528  529  530  531  532  533  534  535  536  537  538  539  540
##    0    0    0    0    0    0    0    0    1    0    0    0    0    0    0
##  541  542  543  544  545  546  547  548  549  550  551  552  553  554  555
##    0    0    1    1    0    0   NA    1    1    0    0    0    0    1    0
##  556  557  558  559  560  561  562  563  564  565  566  567  568  569  570
##    0    1    0    1    1    0    0    1    0    1    0    1    1    1    1
##  571  572  573  574  575  576  577  578  579  580  581  582  583  584  585
##    0   NA    0    0    0    0    0    0   NA    0    0    1    0    1    0
##  586  587  588  589  590  591  592  593  594  595  596  597  598  599  600
##    0    0    0    1    1    1    0    0    1    1   NA   NA    1    0    1
##  601  602  603  604  605  606  607  608  609  610  611  612  613  614  615
##    1    0    0    0    1    1    1   NA    0    0    0    1    0    0    1
##  616  617  618  619  620  621  622  623  624  625  626  627  628  629  630
##    0    0   NA   NA    1    0    0    0   NA    0    1    1    0    0    1
```

```
##  631  632  633  634  635  636  637  638  639  640  641  642  643  644  645
##    0    1   NA    0    0    0   NA    1    0    0    0    1    0    0    0
##  646  647  648  649  650  651  652  653  654  655  656  657  658  659  660
##    1    1    1    0    1   NA    0    1    0    0    0    1    0    0    0
##  661  662  663  664  665  666  667  668  669  670  671  672  673  674  675
##    1    1    1    0    0    1    1   NA    1   NA    0    1    1    0    1
##  676  677  678  679  680  681  682  683  684  685  686  687  688  689  690
##    0    0    1    0    0    1    0    1    0    1    0    0    1    0    0
##  691  692  693  694  695  696  697  698  699  700  701  702  703  704  705
##   NA    0    0    0    0    1   NA    1    1    0    0    0    1    0    0
##  706  707  708  709  710  711  712  713  714  715  716  717  718  719  720
##    0    1    1    0    0    0    0   NA    0    0    0    0    0    0   NA
##  721  722  723  724  725  726  727  728  729  730  731  732  733  734  735
##    1    0    0    0    0    0    0    0    0    0    1    1    1    0    1
##  736  737  738  739  740  741  742  743  744  745  746  747  748  749  750
##   NA    0    0    1    0    1    0    1    0    1    1   NA    0    0    1
##  751  752  753  754  755  756  757  758  759  760  761  762  763  764  765
##    0    0    1    1   NA    1    1    1    0    1    0    1    1    1    1
##  766  767  768  769  770  771  772  773  774  775  776  777  778  779  780
##    1    1    0    1    0    1    1    0   NA    0    1    0    1    0    0
##  781  782  783  784  785  786  787  788  789  790  791  792  793  794  795
##    1    1    0    0    0    1    0   NA    1    0    0    0    1    0    0
##  796  797  798  799  800  801  802  803  804  805  806  807  808  809  810
##    1    0    1    1    1   NA    0    0    1    0    0    0    0    0    1
##  811  812  813  814  815  816  817  818  819  820  821  822  823  824  825
##    1    0   NA    0    1    0    0    1    1    1    1    0   NA    1    1
##  826  827  828  829  830  831  832  833  834  835  836  837  838  839  840
##   NA   NA    0    0   NA    0    0    1    0    0    0   NA   NA    0    0
##  841  842  843  844  845  846  847  848  849  850  851  852  853  854  855
##    0    0    0    0    0   NA    1    1    1    1    1    0    0    0    1
##  856  857  858  859  860  861  862  863  864  865  866  867  868  869  870
##   NA    0    0    1    0    0    1    0    0    1    0    1    0    0   NA
##  871  872  873  874  875  876  877  878  879  880  881  882  883  884  885
##    0    1   NA    1    1    0    0    0    0    1    1    0    0    0    1
##  886  887  888  889  890  891  892  893  894  895  896  897  898  899  900
##    1    1    0    0    0    1    0    0    0    0    1    0    0   NA    1
##  901  902  903  904  905  906  907  908  909  910  911  912  913  914  915
##    0    0    1    0    0    0    1    0    1    1    1    0    0    0    0
##  916  917  918  919  920  921  922  923  924  925  926  927  928  929  930
##   NA    1    1    0    1    0    0    0    0    0    0    0    0    1    0
##  931  932  933  934  935  936  937  938  939  940  941  942  943  944  945
##    0    1   NA    1   NA    1    0   NA    1    1    1    0    1    0    1
##  946  947  948  949  950  951  952  953  954  955  956  957  958  959  960
##   NA    0    0    0    1    0    0    1    1    0    1    0    0    1    0
##  961  962  963  964  965  966  967  968  969  970  971  972  973  974  975
##    1    1    1    0    0    1    1    1    0   NA    1   NA    0    1    0
##  976  977  978  979  980  981  982  983  984  985  986  987  988  989  990
##    1    1    0    0    1    0    0    1    1    1    1   NA    0    1    1
##  991  992  993  994  995  996  997  998  999 1000 1001 1002 1003 1004 1005
##    0   NA    0    0    1    0    0    0    0   NA    1    1    1    0    1
## 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020
##    0    0   NA    1    0    1    0    0   NA    0    1    0    1    0    0
## 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035
##    0    0    1    1    1    1    1    0   NA    0    0    0    0    0    0
```

```
## 1036 1037 1038 1039 1040 1041 1042 1043 1044 1045 1046 1047 1048 1049 1050
##    0    0    0    0    0    0    1    1    0    1    0    1    1    1   NA
## 1051 1052 1053 1054 1055 1056 1057 1058 1059 1060 1061 1062 1063 1064 1065
##    1    1    1    0    1    0    1    0    1    1    0   NA    1   NA    0
## 1066 1067 1068 1069 1070 1071 1072 1073 1074 1075 1076 1077 1078 1079 1080
##    0    0    0    0    1    0   NA    1    1    0    0    0    0    1    0
## 1081 1082 1083 1084 1085 1086 1087 1088 1089 1090 1091 1092 1093 1094 1095
##    1    1    0    1    1    1    0    0    0    0    0    0    0    1   NA
## 1096 1097 1098 1099 1100 1101 1102 1103 1104 1105 1106 1107 1108 1109 1110
##    0    0    0    1    1    0    0    1    1    1    1    0    0   NA    1
## 1111 1112 1113 1114 1115 1116 1117 1118 1119 1120 1121 1122 1123 1124 1125
##    1   NA    0    0    0    0    1    1    1    0    1    1    0    0    0
## 1126 1127 1128 1129 1130 1131 1132 1133 1134 1135 1136 1137 1138 1139 1140
##    0    1    0    0    0   NA    0    1    0    1    0    0    0    0    0
## 1141 1142 1143 1144 1145 1146 1147 1148 1149 1150 1151 1152 1153 1154 1155
##    0    0    1    1    1    0    0    1    0    1    1    1    0   NA    1
## 1156 1157 1158 1159 1160 1161 1162 1163 1164 1165 1166 1167 1168 1169 1170
##    1    0    0    0    0    0    0    0    0    0    1    0    1    1    0
## 1171 1172 1173 1174 1175 1176 1177 1178 1179 1180 1181 1182 1183 1184 1185
##    1   NA    1    1    0    0   NA    0    1    1   NA    1    0    1    1
## 1186 1187 1188 1189 1190 1191 1192 1193 1194 1195 1196 1197 1198 1199 1200
##   NA    0    0   NA    0    1    0   NA    1    0    0    0    0    1    1
## 1201 1202 1203 1204 1205 1206 1207 1208 1209 1210 1211 1212 1213 1214 1215
##    0    0    0    1    0    0   NA    0    0    1    0    0    1    0    0
## 1216 1217 1218 1219 1220 1221 1222 1223 1224 1225 1226 1227 1228 1229 1230
##    0    1    1    0    0    0    0    1    0   NA    0    0    0    1    1
## 1231 1232 1233 1234 1235 1236 1237 1238 1239 1240 1241 1242 1243 1244 1245
##    0    1    1    1    0    0    0    1    0    0    1    0    0    0    0
## 1246 1247 1248 1249 1250 1251 1252 1253 1254 1255 1256 1257 1258 1259 1260
##    1    1    0    0    0    0    1    0    1    0    0    1    0    1    0
## 1261 1262 1263 1264 1265 1266 1267 1268 1269 1270 1271 1272 1273 1274 1275
##    0    1    0    1   NA    1    1    0    1    0    1    0    1    1    0
## 1276 1277 1278 1279 1280 1281 1282 1283 1284 1285 1286 1287 1288 1289 1290
##    0    1    0    0    0    1    0    0   NA    1    0    1    0    0    0
## 1291 1292 1293 1294 1295 1296 1297 1298 1299 1300 1301 1302 1303 1304 1305
##    1   NA    0    1   NA    0    1   NA    1   NA    0    0   NA    0    0
## 1306 1307 1308 1309 1310 1311 1312 1313 1314 1315 1316 1317 1318 1319 1320
##    0    1    1    0    1    1    0   NA    0    0    0    0   NA    0    1
## 1321 1322 1323 1324 1325 1326 1327 1328 1329 1330 1331 1332 1333 1334 1335
##    0   NA    1    0    0    1    0    1    0   NA    0    0    0    0    1
## 1336 1337 1338 1339 1340 1341 1342 1343 1344 1345 1346 1347 1348 1349 1350
##   NA    0    0    1    0    0    1    0    0    1    0    0    1    0   NA
## 1351 1352 1353 1354 1355 1356 1357 1358 1359 1360 1361 1362 1363 1364 1365
##    0    1    0    0    0    0    0   NA    0    0    0    0    1    0   NA
## 1366 1367 1368 1369 1370 1371 1372 1373 1374 1375 1376 1377 1378 1379 1380
##    0    1    1    1    1    0    0    0    0    0    1    1   NA    0   NA
## 1381 1382 1383 1384 1385 1386 1387 1388 1389 1390 1391 1392 1393 1394 1395
##    1    1    1    0    1    0    0   NA    0    1    1    1    0    1    0
## 1396 1397 1398 1399 1400 1401 1402 1403 1404 1405 1406 1407 1408 1409 1410
##    0    0    1    1    1    0    0    1    1   NA    1    0    0    0    1
## 1411 1412 1413 1414 1415 1416 1417 1418 1419 1420 1421 1422 1423 1424 1425
##    0    0   NA    0    0    0    0    0    1    0    0    1    1    0    1
## 1426 1427 1428 1429 1430 1431 1432 1433 1434 1435 1436 1437 1438 1439 1440
##    1    0    0    1    0    1    0    0    0    0    1    1    1    0    0
```

```
## 1441 1442 1443 1444 1445 1446 1447 1448 1449 1450 1451 1452 1453 1454 1455
##    0    1    1    0   NA    1    1    1    0    0    0    0    0    0   NA
## 1456 1457 1458 1459 1460 1461 1462 1463 1464 1465 1466 1467 1468 1469 1470
##   NA    1   NA    0    0    0    0   NA    1    0    1    0    0    0    1
## 1471 1472 1473 1474 1475 1476 1477 1478 1479 1480 1481 1482 1483 1484 1485
##    0    0    0    0    1    0    1    0    0    0    1   NA    1    0    1
## 1486 1487 1488 1489 1490 1491 1492 1493 1494 1495 1496 1497 1498 1499 1500
##    1    0    1    1    0    0    1    0    0    1    1    0    0    0    0
## 1501 1502 1503 1504 1505 1506 1507 1508 1509 1510 1511 1512 1513 1514 1515
##    0    1    1    1    0   NA    1   NA    0    0    0   NA    0    0    1
## 1516 1517 1518 1519 1520 1521 1522 1523 1524 1525 1526 1527 1528 1529 1530
##    1    0    1    0    0    1    0    0   NA   NA    1    0   NA    1    1
## 1531 1532 1533 1534 1535 1536 1537 1538 1539 1540 1541 1542 1543 1544 1545
##    1    1    0    0    1    1    0    1    1    1    0    0   NA    1   NA
## 1546 1547 1548 1549 1550 1551 1552 1553 1554 1555 1556 1557 1558 1559 1560
##    0    0    1    0   NA   NA    0    0    1   NA    0    0    0    1    0
## 1561 1562 1563 1564 1565 1566 1567 1568 1569 1570 1571 1572 1573 1574 1575
##    0    1    1    1    1   NA   NA    0    0    1    1    0    0   NA   NA
## 1576 1577 1578 1579 1580 1581 1582 1583 1584 1585 1586 1587 1588 1589 1590
##    1    1    0    0    0    1    0    0    0    0    0    0    1    0    1
## 1591 1592 1593 1594 1595 1596 1597 1598 1599 1600 1601 1602 1603 1604 1605
##    1    1   NA   NA    0    0    1   NA    0    1    1    0    1    1    0
## 1606 1607 1608 1609 1610 1611 1612 1613 1614 1615 1616 1617 1618 1619 1620
##    1    0    0    0    1    1    0    0   NA    1    1    0    1    1   NA
## 1621 1622 1623 1624 1625 1626 1627 1628 1629 1630 1631 1632 1633 1634 1635
##    1    0    1    0    0    0    1    0    0    1    0    1    0    1    0
## 1636 1637 1638 1639 1640 1641 1642 1643 1644 1645 1646 1647 1648 1649 1650
##    0    1    0    0    1    0    0   NA    0    1    0    0    0    0    1
## 1651 1652 1653 1654 1655 1656 1657 1658 1659 1660 1661 1662 1663 1664 1665
##    0    0    0    0    1    1    0    0    0    0   NA    1    1   NA    1
## 1666 1667 1668 1669 1670 1671 1672 1673 1674 1675 1676 1677 1678 1679 1680
##    0   NA    1    1   NA    1    0    1   NA    0    0    0    0    0    0
## 1681 1682 1683 1684 1685 1686 1687 1688 1689 1690 1691 1692 1693 1694 1695
##    0    1   NA    0    1    0    0    1    0    0    0    0    0    1    0
## 1696 1697 1698 1699 1700 1701 1702 1703 1704 1705 1706 1707 1708 1709 1710
##    1    0    1    1    0    1    0    0    1    0    0   NA   NA    1    0
## 1711 1712 1713 1714 1715 1716 1717 1718 1719 1720 1721 1722 1723 1724 1725
##    1    0    1    0    1    0    0    0    1   NA    1    1    0    1    1
## 1726 1727 1728 1729 1730 1731 1732 1733 1734 1735 1736 1737 1738 1739 1740
##   NA   NA    1    1    1    1    0    0    1    0    0    1   NA    0    0
## 1741 1742 1743 1744 1745 1746 1747 1748 1749 1750 1751 1752 1753 1754 1755
##    1    0    0    0    1    0    0    0    1    0    0    1    0    1    0
## 1756 1757 1758 1759 1760 1761 1762 1763 1764 1765 1766 1767 1768 1769 1770
##    0    0    1    0    0    1    0    0    0    1   NA    0    1    0    0
## 1771 1772 1773 1774 1775 1776 1777 1778 1779 1780 1781 1782 1783 1784 1785
##    0   NA    0    1    1    0   NA    1    1    1   NA    0    1    0    0
## 1786 1787 1788 1789 1790 1791 1792 1793 1794 1795 1796 1797 1798 1799 1800
##    0    0    0    1    1    0    0    0    0   NA    1    0    0    0    1
## 1801 1802 1803 1804 1805 1806 1807 1808 1809 1810 1811 1812 1813 1814 1815
##    1   NA    0    0    0   NA   NA    1    0    0    0    0    1    1    1
## 1816 1817 1818 1819 1820 1821 1822 1823 1824 1825 1826 1827 1828 1829 1830
##    0    0    0    0    0    0    0    0    1    1    1    0    1    0    0
## 1831 1832 1833 1834 1835 1836 1837 1838 1839 1840 1841 1842 1843 1844 1845
##    0    0    0    0    1    0    0    1    0    1    0    0    1   NA    0
```

```
## 1846 1847 1848 1849 1850 1851 1852 1853 1854 1855 1856 1857 1858 1859 1860
##    0    0    0   NA    0    1    0    1   NA    0    1    0    0    0    0
## 1861 1862 1863 1864 1865 1866 1867 1868 1869 1870 1871 1872 1873 1874 1875
##    1    0    0    1    1    0    0    0    1    1   NA    1   NA    0    1
## 1876 1877 1878 1879 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1890
##    0    0    1    1    1   NA   NA    0    0    1    1    0    0    1    1
## 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900 1901 1902 1903 1904 1905
##    0    1    0    0    1   NA    0    0    0    1    0    1    0    1    0
## 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 1920
##    0    0    0    1    1    1    1    0    0    1   NA    0    0   NA    0
## 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935
##    0    1    0    0    1    0   NA    0    0    0   NA    0    1    1    0
## 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945 1946 1947 1948 1949 1950
##    1    1    0    0    0    0    0   NA    0    0    0   NA   NA    1    0
## 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965
##    1    0    1    1    0   NA    0    0    1   NA    1   NA    1    1    0
## 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980
##   NA    0    1    0   NA    0    0    1    1    0    1    0    1    1   NA
## 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995
##    0    1   NA    0    0    0    0    0   NA    1    0   NA    1    1    0
## 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010
##    0    1    1    1   NA    1    1   NA   NA    1    0    1    0    1    1
## 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025
##    1    0    1    0    0    1    0    1    1    0    0    0   NA    0    1
## 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040
##    0    0    0    0    1    0    0    0    0    1    1    0    0    1    0
## 2041 2042 2043 2044 2045 2046 2047 2048 2049 2050 2051 2052 2053 2054 2055
##    0    0    0    1    0    0    0    1    1    0    1    1    1    0   NA
## 2056 2057 2058 2059 2060 2061 2062 2063 2064 2065 2066 2067 2068 2069 2070
##    0    0    0    0    0    0    1    1    0    1    0    1    1    1    0
## 2071 2072 2073 2074 2075 2076 2077 2078 2079 2080 2081 2082 2083 2084 2085
##    0    0   NA    1    0    0    1    0    0    1    0    1   NA    0    0
## 2086 2087 2088 2089 2090 2091 2092 2093 2094 2095 2096 2097 2098 2099 2100
##    0    1    1   NA    1    0    0    1    0    1    1    0    1    1    1
## 2101 2102 2103 2104 2105 2106 2107 2108 2109 2110 2111 2112 2113 2114 2115
##    1    1    1    0   NA    0    1   NA    1    0    1    0    1    0    0
## 2116 2117 2118 2119 2120 2121 2122 2123 2124 2125 2126 2127 2128 2129 2130
##   NA    1    1    1    0    0    0    1    1    1    0   NA    0    0    0
## 2131 2132 2133 2134 2135 2136 2137 2138 2139 2140 2141
##    0    0    0    0    1    0    0    0    0    0    0
```

7 Appendix

For full code visit:

https://raw.githubusercontent.com/raghu74us/DATA-621/master/Assignment4/621_Assignment4.Rmd