Homework 04

Forhad Akbar

2021-04-25

Overview

In this homework assignment, you will explore, analyze and model a dataset containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first responsevariable, 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 responsevariable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

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

Data Exploration

We will explore both training and evaluation data as the first step in the data exploration.

```
## [1] "Dimension of training set:   Number of rows: 8161, Number of cols: 26"
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ   INCOME PARENT1
## 1     1           0          0        0  60        0  11  $67,349      No
## 2     2           0          0        0  43        0  11  $91,449      No
## 3     4           0          0        0  35        1  10  $16,039      No
## 4     5           0          0        0  51        0  14               No
## 5     6           0          0        0  50        0  NA $114,986      No
## 6     7           1       2946        0  34        1  12 $125,301     Yes
##   HOME_VAL MSTATUS SEX     EDUCATION           JOB TRAVTIME    CAR_USE BLUEBOOK
## 1       $0    z_No   M           PhD  Professional       14    Private  $14,230
## 2 $257,252    z_No   M z_High School z_Blue Collar       22 Commercial  $14,940
## 3 $124,191     Yes z_F z_High School      Clerical        5    Private   $4,010
## 4 $306,251     Yes   M  <High School z_Blue Collar       32    Private  $15,440
## 5 $243,925     Yes z_F           PhD        Doctor       36    Private  $18,000
## 6       $0    z_No z_F     Bachelors z_Blue Collar       46 Commercial  $17,430
##   TIF   CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1  11    Minivan     yes   $4,461        2      No       3      18
## 2   1    Minivan     yes       $0        0      No       0       1
## 3   4      z_SUV      no  $38,690        2      No       3      10
## 4   7    Minivan     yes       $0        0      No       0       6
## 5   1      z_SUV      no  $19,217        2     Yes       3      17
## 6   1 Sports Car      no       $0        0      No       0       7
##            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
```

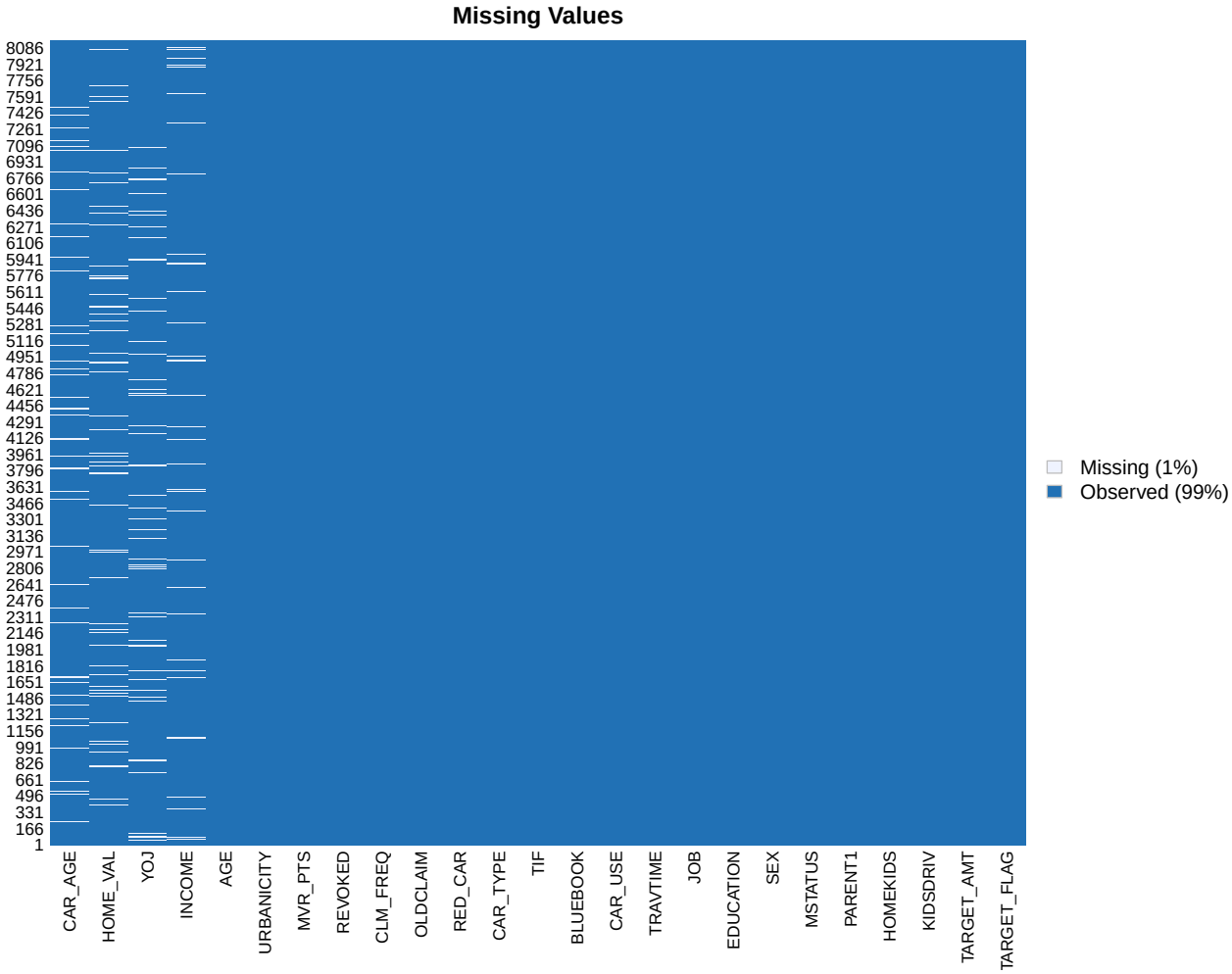
```
## [1] "Structure of training data set:"
```

```
## 'data.frame':    8161 obs. of  26 variables:
##  $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
```

```
##  $ TARGET_AMT : num  0 0 0 0 0 ...
##  $ KIDSDRIV   : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
##  $ INCOME     : chr  "$67,349" "$91,449" "$16,039" "" ...
##  $ PARENT1    : chr  "No" "No" "No" "No" ...
##  $ HOME_VAL   : chr  "$0" "$257,252" "$124,191" "$306,251" ...
##  $ MSTATUS    : chr  "z_No" "z_No" "Yes" "Yes" ...
##  $ SEX        : chr  "M" "M" "z_F" "M" ...
##  $ EDUCATION  : chr  "PhD" "z_High School" "z_High School" "<High School" ...
##  $ JOB        : chr  "Professional" "z_Blue Collar" "Clerical" "z_Blue Collar" ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : chr  "Private" "Commercial" "Private" "Private" ...
##  $ BLUEBOOK   : chr  "$14,230" "$14,940" "$4,010" "$15,440" ...
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : chr  "Minivan" "Minivan" "z_SUV" "Minivan" ...
##  $ RED_CAR    : chr  "yes" "yes" "no" "yes" ...
##  $ OLDCLAIM   : chr  "$4,461" "$0" "$38,690" "$0" ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : chr  "No" "No" "No" "No" ...
##  $ 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 : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urba
```

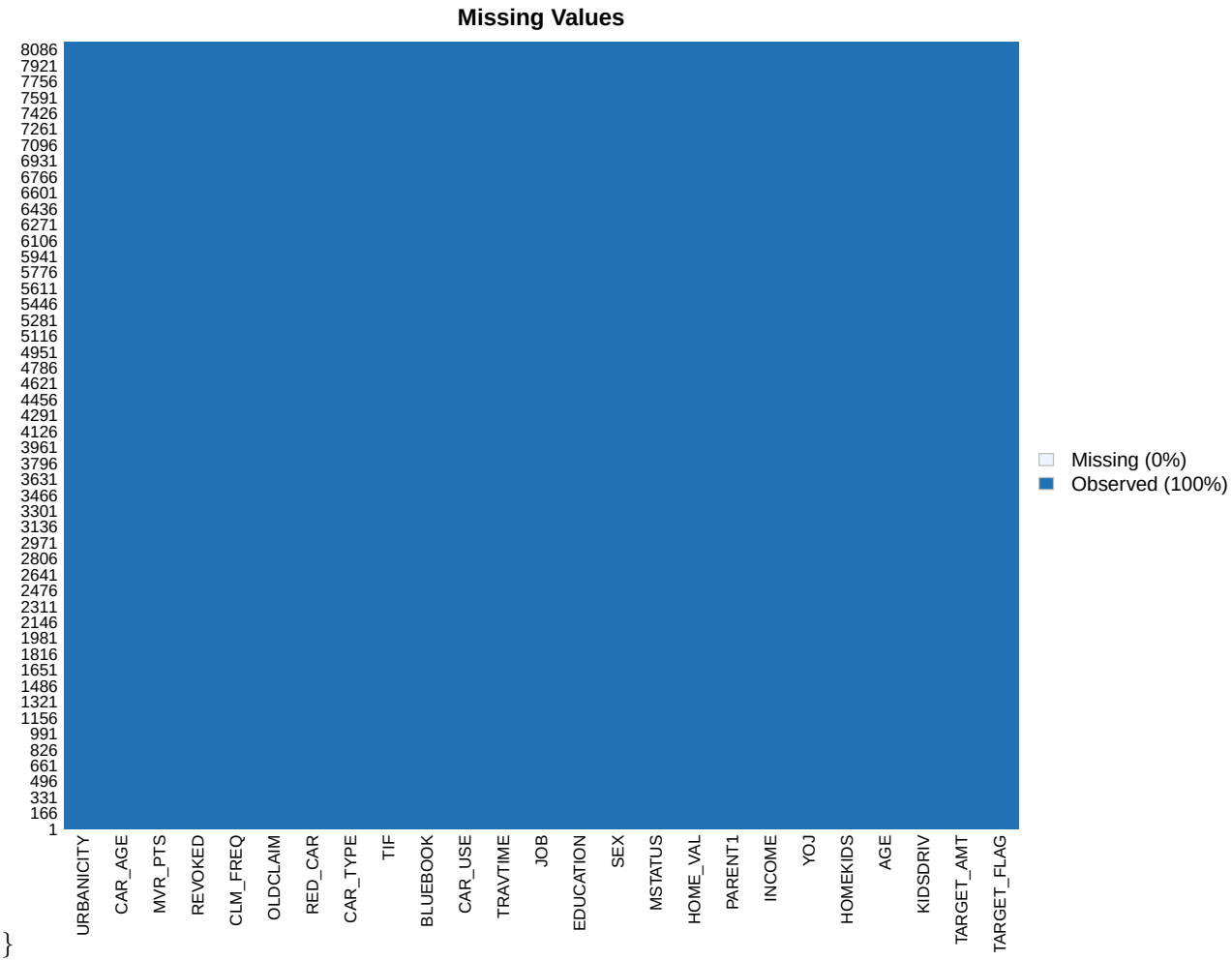
Checking for NA.

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



There are NA in our data, We will impute NA using mice().

Rechecking for NA after imputation.



}

We observe that NA were removed.

Here, We'll further explore the data. Firsr, We'll look at min, 1st quartile, median, mean, 2nd quartile, max etc.

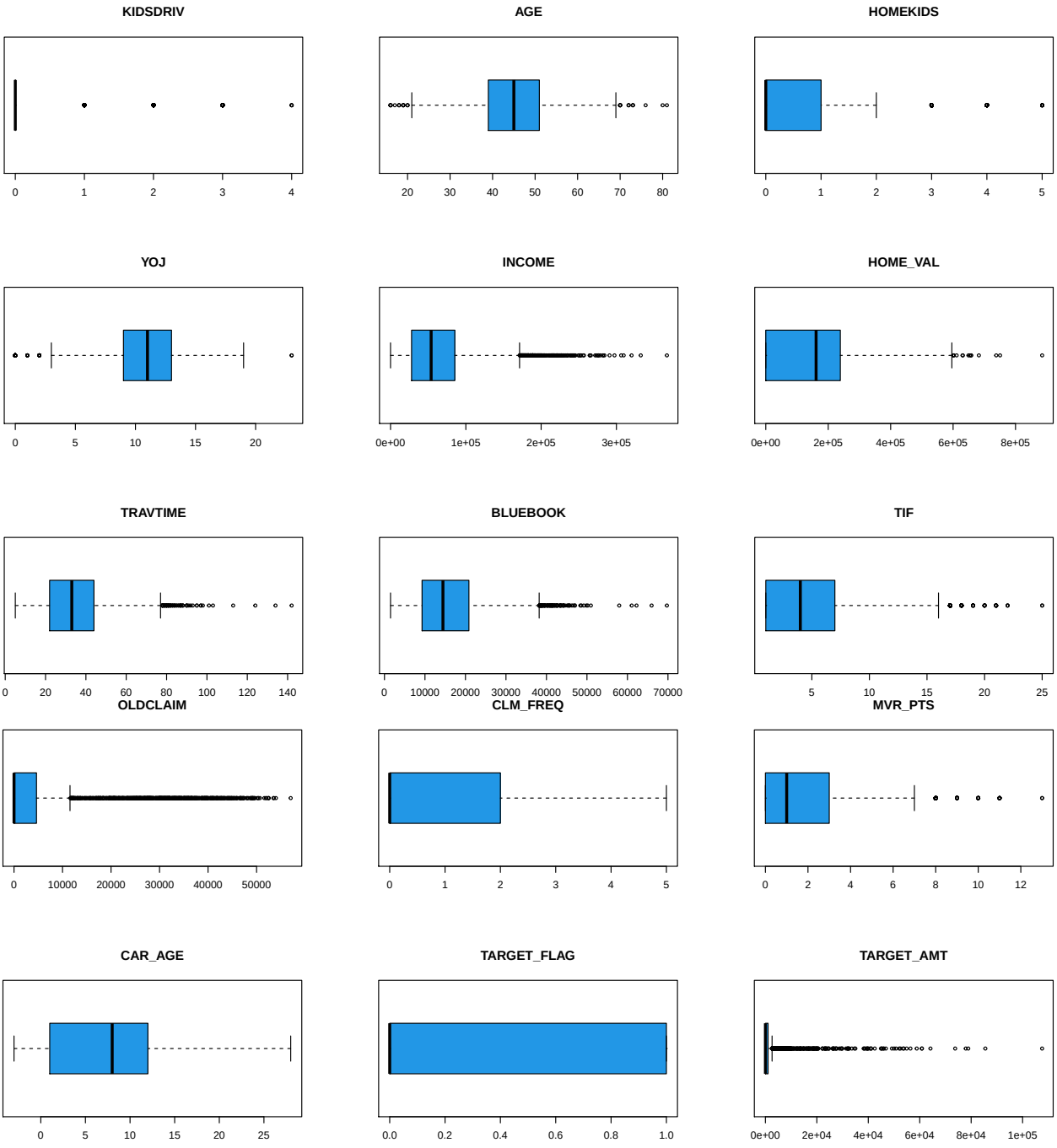
	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME
	Min. :0.0000	Min. : 0	Min. :0.0000	Min. :16.00	Min. :0.0000	Min. : 0.00	Min. : 0
	1st Qu.:0.0000	1st Qu.: 0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000	1st Qu.: 9.00	1st Qu.: 2
	Median :0.0000	Median : 0	Median :0.0000	Median :45.00	Median :0.0000	Median :11.00	Median :
	Mean :0.2638	Mean : 1504	Mean :0.1711	Mean :44.78	Mean :0.7212	Mean :10.52	Mean : 6
	3rd Qu.:1.0000	3rd Qu.: 1036	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:13.00	3rd Qu.:
	Max. :1.0000	Max. :107586	Max. :4.0000	Max. :81.00	Max. :5.0000	Max. :23.00	Max. :36

Data reordering

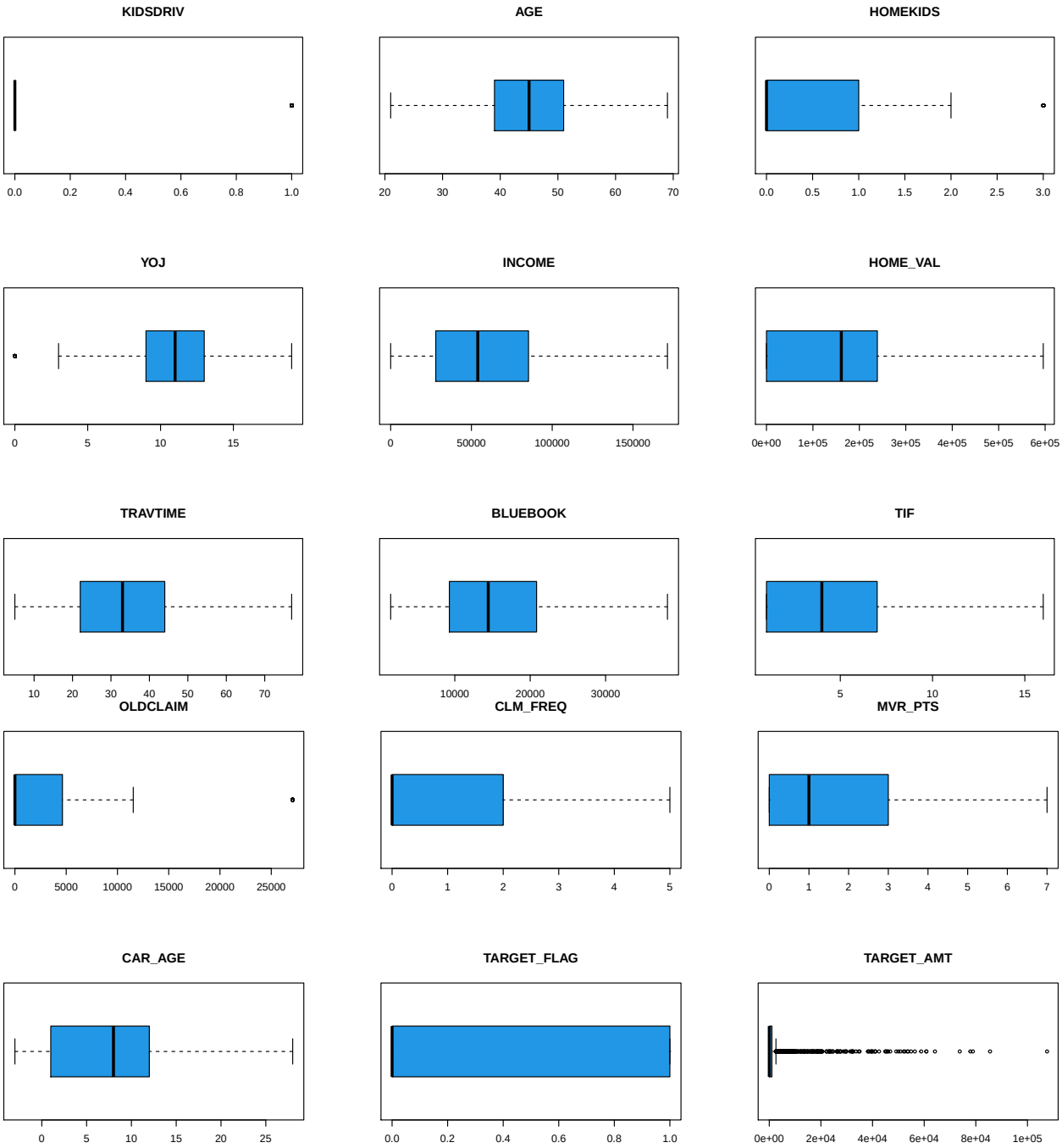
For downstream analysis, we'll reorder the columns into categorical, numeric and target.

Boxplots

First look at the boxplots.



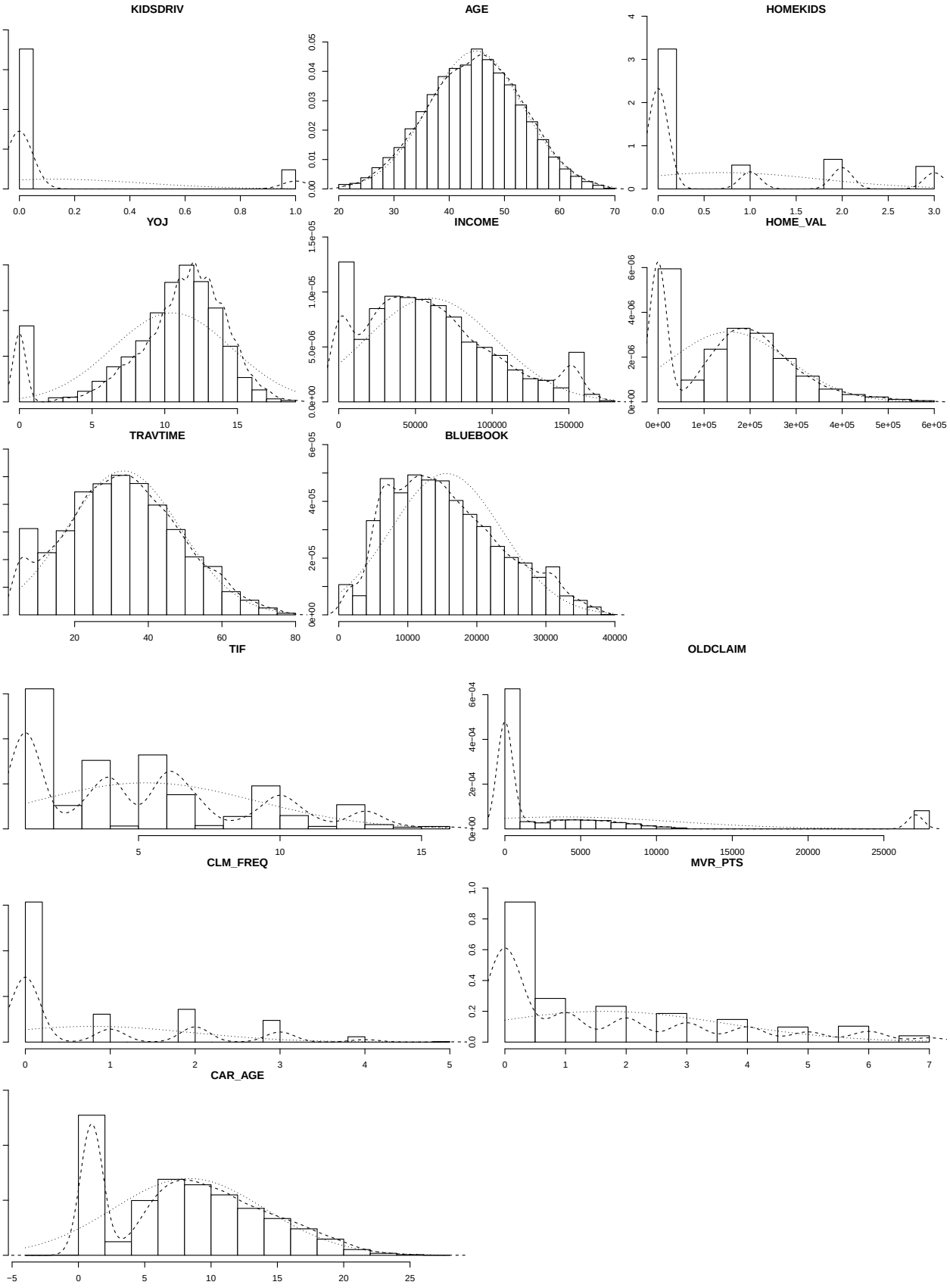
The boxplots show that some of the variables have outliers in them. So, we'll cap them.



The fields AGE, HOMEKIDS, INCOME, HOME_VAL, TRVTIME, BLUEBOOK, TIF, CLM_FREQ, MVR_PTS, CAR_AGE have higher variance. We will ignore the boxplots for TARGET_FLAG and TARGET_AMT here.

Histograms

We will see how the data is distributed (numeric fields) using histrogram.


KIDSDRIV
AGE
HOMEKIDS
YOJ
INCOME
HOME_VAL
TRAVTIME
BLUEBOOK
TIF
OLDCLAIM
CLM_FREQ
MVR_PTS
CAR_AGE

We can see that AGE, YOJ, INCOME, HOME_VAL, TRAVTIME, BLUEBOOK and CAR_AGE are approximately normally distributed. HOME_VALUE, CAR_AGE and CLM_FREQ are quite dispersed.

Categorical variables

Now, let's explore the Categorical variables.

```
## PARENT1:

##
##   No  Yes
## 7084 1077

## MSTATUS:

##
##   No  Yes
## 3267 4894

## SEX:

##
##    F    M
## 4375 3786

## EDUCATION:

##
## <High School    Bachelors  High School      Masters          PhD
##         1203         2242         2330         1658          728

## JOB:

##
##               Blue Collar     Clerical       Doctor   Home Maker       Lawyer
##          526         1825         1271          246          641          835
##      Manager Professional      Student
##          988         1117          712

## CAR_USE:

##
## Commercial    Private
##       3029       5132

## CAR_TYPE:

##
##     Minivan Panel Truck      Pickup  Sports Car         SUV         Van
##        2145         676        1389         907        2294         750

## RED_CAR:

##
##   no  yes
## 5783 2378

## REVOKED:

##
##   No  Yes
## 7161 1000
```

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

There are 526 rows empty in JOB column. So, we'll impute them with "Unknown".

```
## JOB:
```

```
##
##  Blue Collar     Clerical       Doctor   Home Maker       Lawyer      Manager
##         1825         1271          246          641          835          988
## Professional      Student      Unknown
##         1117          712          526
```

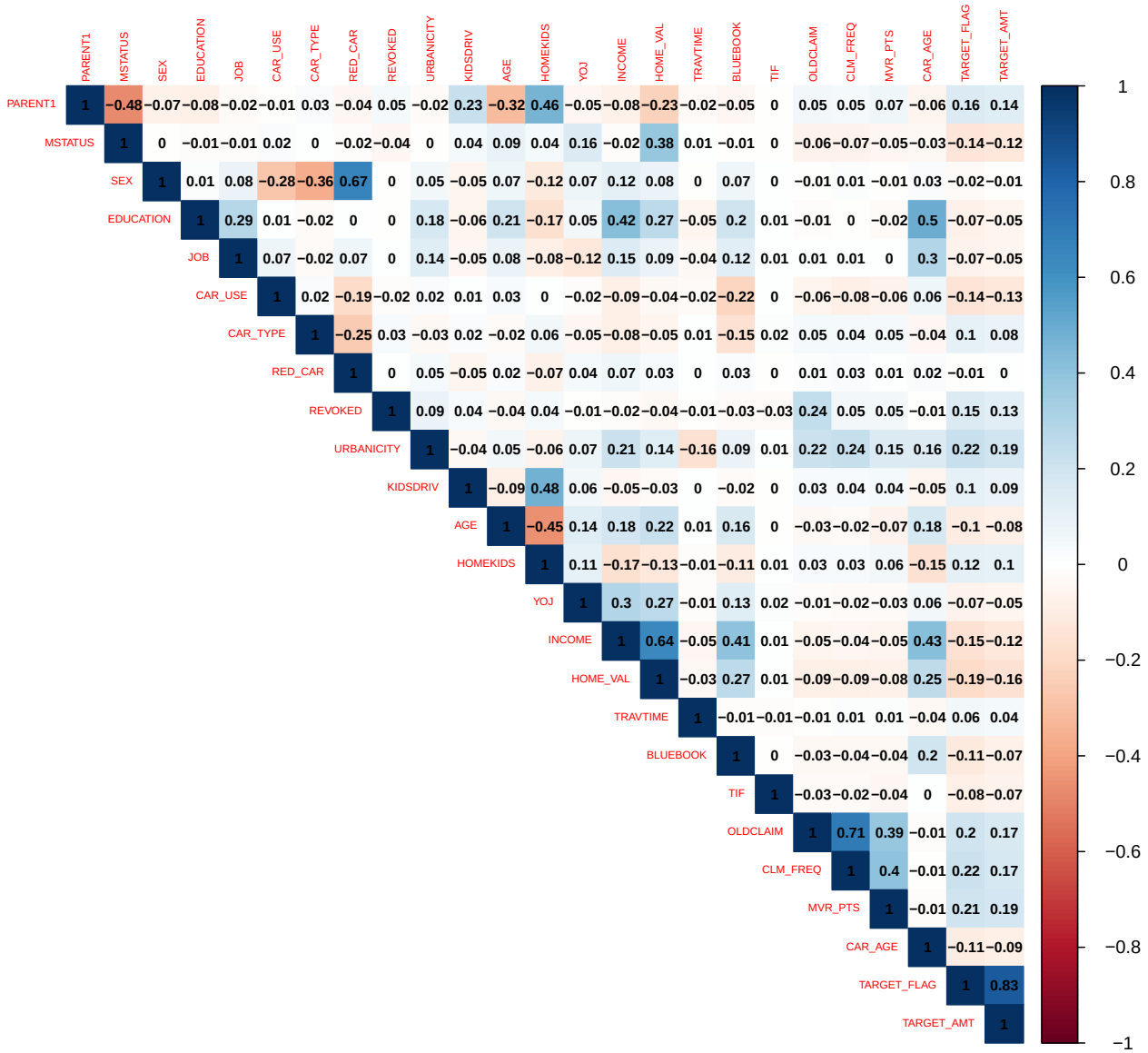
Correlations

At this point the data is prepared. So, we'll explore the top correlated variables.

There are 25 variables, among which 15 are numeric and 10 are non-categorical. In order to find the top correlated variables, we'll give numerical values to the correlated variables.

	Top Correlated Variables	
	TARGET_FLAG	TARGET_AMT
TARGET_FLAG	1.0000000	1.0000000
TARGET_AMT	0.8334240	0.8334240
URBANICITY	0.2242509	0.1904945
CLM_FREQ	0.2161961	0.1869848
MVR_PTS	0.2075451	0.1741927
OLDCLAIM	0.2004106	0.1652881
PARENT1	0.1576222	0.1359305
REVOKED	0.1519391	0.1263285
HOMEKIDS	0.1175903	0.1014967
KIDSDRIV	0.1036217	0.0863553
CAR_TYPE	0.1035765	0.0827170
TRAVTIME	0.0550685	0.0440349
RED_CAR	-0.0069473	0.0005877
SEX	-0.0210786	-0.0088270
YOJ	-0.0661180	-0.0509930
JOB	-0.0669944	-0.0512803
EDUCATION	-0.0734429	-0.0519743
TIF	-0.0818308	-0.0690666
AGE	-0.1044668	-0.0712728
CAR_AGE	-0.1092440	-0.0817563
BLUEBOOK	-0.1100800	-0.0851285
MSTATUS	-0.1351248	-0.1214174
CAR_USE	-0.1426737	-0.1214701
INCOME	-0.1504646	-0.1287263
HOME_VAL	-0.1874506	-0.1554778

Now, we'll look at the correlation matrix of the variables.


PARENT1 MSTATUS SEX EDUCATION JOB CAR_USE CAR_TYPE RED_CAR REVOKED URBANICITY KIDSDRIV AGE HOMEKIDS YOJ INCOME HOME_VAL TRAVTIME BLUEBOOK TIF OLDCLAIM CLM_FREQ MVR_PTS CAR_AGE TARGET_FLAG TARGET_AMT
1 −0.48 −0.07 −0.08 −0.02 −0.01 0.03 −0.04 0.05 −0.02 0.23 −0.32 0.46 −0.05 −0.08 −0.23 −0.02 −0.05 0 0.05 0.05 0.07 −0.06 0.16 0.14
1 0 −0.01 −0.01 0.02 0 −0.02 −0.04 0 0.04 0.09 0.04 0.16 −0.02 0.38 0.01 −0.01 0 −0.06 −0.07 −0.05 −0.03 −0.14 −0.12
1 0.01 0.08 −0.28 −0.36 0.67 0 0.05 −0.05 0.07 −0.12 0.07 0.12 0.08 0 0.07 0 −0.01 0.01 −0.01 0.03 −0.02 −0.01
1 0.29 0.01 −0.02 0 0 0.18 −0.06 0.21 −0.17 0.05 0.42 0.27 −0.05 0.2 0.01 −0.01 0 −0.02 0.5 −0.07 −0.05
1 0.07 −0.02 0.07 0 0.14 −0.05 0.08 −0.08 −0.12 0.15 0.09 −0.04 0.12 0.01 0.01 0.01 0 0.3 −0.07 −0.05
1 0.02 −0.19 −0.02 0.02 0.01 0.03 0 −0.02 −0.09 −0.04 −0.02 −0.22 0 −0.06 −0.08 −0.06 0.06 −0.14 −0.13
1 −0.25 0.03 −0.03 0.02 −0.02 0.06 −0.05 −0.08 −0.05 0.01 −0.15 0.02 0.05 0.04 0.05 −0.04 0.1 0.08
1 0 0.05 −0.05 0.02 −0.07 0.04 0.07 0.03 0 0.03 0 0.01 0.03 0.01 0.02 −0.01 0
1 0.09 0.04 −0.04 0.04 −0.01 −0.02 −0.04 −0.01 −0.03 −0.03 0.24 0.05 0.05 −0.01 0.15 0.13
1 −0.04 0.05 −0.06 0.07 0.21 0.14 −0.16 0.09 0.01 0.22 0.24 0.15 0.16 0.22 0.19
1 −0.09 0.48 0.06 −0.05 −0.03 0 −0.02 0 0.03 0.04 0.04 −0.05 0.1 0.09
1 −0.45 0.14 0.18 0.22 0.01 0.16 0 −0.03 −0.02 −0.07 0.18 −0.1 −0.08
1 0.11 −0.17 −0.13 −0.01 −0.11 0.01 0.03 0.03 0.06 −0.15 0.12 0.1
1 0.3 0.27 −0.01 0.13 0.02 −0.01 −0.02 −0.03 0.06 −0.07 −0.05
1 0.64 −0.05 0.41 0.01 −0.05 −0.04 −0.05 0.43 −0.15 −0.12
1 −0.03 0.27 0.01 −0.09 −0.09 −0.08 0.25 −0.19 −0.16
1 −0.01 −0.01 −0.01 0.01 0.01 −0.04 0.06 0.04
1 0 −0.03 −0.04 −0.04 0.2 −0.11 −0.07
1 −0.03 −0.02 −0.04 0 −0.08 −0.07
1 0.71 0.39 −0.01 0.2 0.17
1 0.4 −0.01 0.22 0.17
1 −0.01 0.21 0.19
1 −0.11 −0.09
1 0.83
1
1 0.8 0.6 0.4 0.2 0 −0.2 −0.4 −0.6 −0.8 −1

At this point exploration, preparation and pair-wise correlations of **insurance_training_data.csv** are done. So, I'll begin the same exercise for **insurance-evaluation-data.csv**.

Data Exploration of insurance-evaluation-data.csv.

Initially, We'll do a cursory exploration of the data. After that, we'll iteratively prepare and explore the data, wherever required.

```
## [1] "Dimension of training set:   Number of rows: 2141, Number of cols: 26"
```

```
## [1] "Head of training data set:"
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ  INCOME PARENT1
## 1     3          NA         NA        0  48        0  11 $52,881      No
## 2     9          NA         NA        1  40        1  11 $50,815     Yes
## 3    10          NA         NA        0  44        2  12 $43,486     Yes
## 4    18          NA         NA        0  35        2  NA $21,204     Yes
## 5    21          NA         NA        0  59        0  12 $87,460      No
## 6    30          NA         NA        0  46        0  14              No
##   HOME_VAL MSTATUS SEX     EDUCATION           JOB TRAVTIME    CAR_USE BLUEBOOK
## 1       $0    z_No   M     Bachelors       Manager       26    Private  $21,970
## 2       $0    z_No   M z_High School       Manager       21    Private  $18,930
## 3       $0    z_No z_F z_High School z_Blue Collar       30 Commercial   $5,900
## 4       $0    z_No   M z_High School      Clerical       74    Private   $9,230
## 5       $0    z_No   M z_High School       Manager       45    Private  $15,420
## 6 $207,519     Yes   M     Bachelors  Professional        7 Commercial  $25,660
##   TIF    CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS CAR_AGE
## 1   1         Van     yes       $0        0      No       2      10
## 2   6     Minivan      no   $3,295        1      No       2       1
## 3  10       z_SUV      no       $0        0      No       0      10
## 4   6      Pickup      no       $0        0     Yes       0       4
## 5   1     Minivan     yes  $44,857        2      No       4       1
## 6   1 Panel Truck      no   $2,119        1      No       2      12
##              URBANICITY
## 1   Highly Urban/ Urban
## 2   Highly Urban/ Urban
## 3 z_Highly Rural/ Rural
## 4 z_Highly Rural/ Rural
## 5   Highly Urban/ Urban
## 6   Highly Urban/ Urban
```

```
## [1] "Structure of training data set:"
```

```
## 'data.frame':    2141 obs. of  26 variables:
##  $ INDEX      : int  3 9 10 18 21 30 31 37 39 47 ...
##  $ TARGET_FLAG: logi  NA NA NA NA NA NA ...
##  $ TARGET_AMT : logi  NA NA NA NA NA NA ...
##  $ KIDSDRIV   : int  0 1 0 0 0 0 0 0 2 0 ...
##  $ AGE        : int  48 40 44 35 59 46 60 54 36 50 ...
##  $ HOMEKIDS   : int  0 1 2 2 0 0 0 0 2 0 ...
##  $ YOJ        : int  11 11 12 NA 12 14 12 12 12 8 ...
##  $ INCOME     : chr  "$52,881" "$50,815" "$43,486" "$21,204" ...
##  $ PARENT1    : chr  "No" "Yes" "Yes" "Yes" ...
##  $ HOME_VAL   : chr  "$0" "$0" "$0" "$0" ...
##  $ MSTATUS    : chr  "z_No" "z_No" "z_No" "z_No" ...
```

```
##  $ SEX        : chr  "M" "M" "z_F" "M" ...
##  $ EDUCATION  : chr  "Bachelors" "z_High School" "z_High School" "z_High School" ...
##  $ JOB        : chr  "Manager" "Manager" "z_Blue Collar" "Clerical" ...
##  $ TRAVTIME   : int  26 21 30 74 45 7 16 27 5 22 ...
##  $ CAR_USE    : chr  "Private" "Private" "Commercial" "Private" ...
##  $ BLUEBOOK   : chr  "$21,970" "$18,930" "$5,900" "$9,230" ...
##  $ TIF        : int  1 6 10 6 1 1 1 4 4 4 ...
##  $ CAR_TYPE   : chr  "Van" "Minivan" "z_SUV" "Pickup" ...
##  $ RED_CAR    : chr  "yes" "no" "no" "no" ...
##  $ OLDCLAIM   : chr  "$0" "$3,295" "$0" "$0" ...
##  $ CLM_FREQ   : int  0 1 0 0 2 1 0 0 0 0 ...
##  $ REVOKED    : chr  "No" "No" "No" "Yes" ...
##  $ MVR_PTS    : int  2 2 0 0 4 2 0 5 0 3 ...
##  $ CAR_AGE    : int  10 1 10 4 1 12 1 NA 9 1 ...
##  $ URBANICITY : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "z_Highly Rural/ Rural" "z_Highly Rural/ R
```

There are few fields, which have missing values, which we'll investigate in greater details later.

Data Preparation of insurance-evaluation-data.csv.

At this stage, We'll explore and prepare iteratively. First we'll convert the fields, which are supposed to be numeric, into proper numeric format and strings into string format. After reformatting, we'll check for NA. After that if required, we'll impute them.

After that we'll show some boxplots of the numeric fields.

Checking for NA.

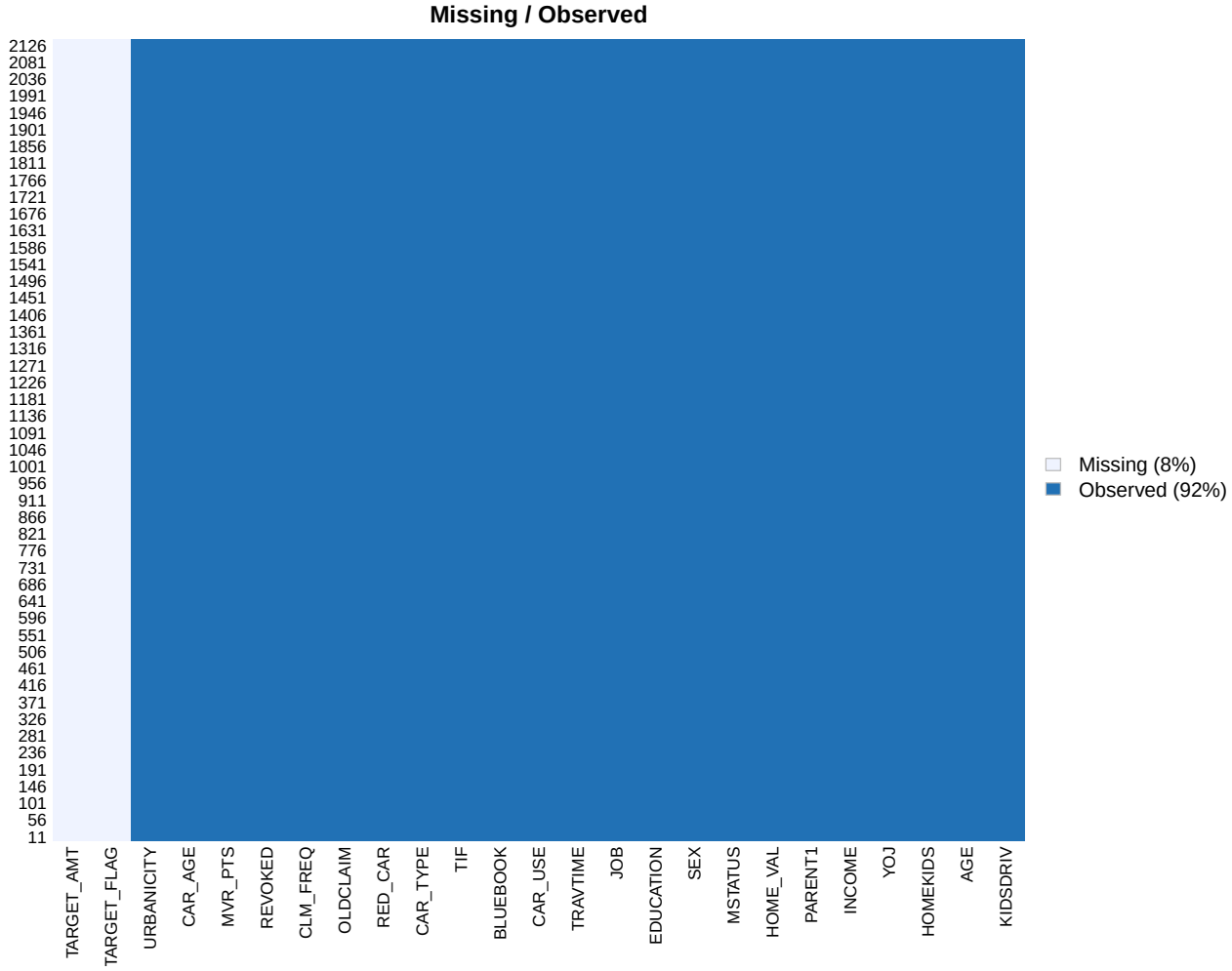
```
## [1] TRUE
```

NA does exist. So, we'll impute with mice().

Rechecking for NA after imputation.

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

We observe that NA were removed in all columns except TARGET_FLAG and TARGET_AMT, which is what we want. In the following, we'll visualize with missmap().



Both is.na() and missmap() confirm that NA were eliminated.

More Data exploration of insurance-evaluation-data.csv.

Now, we'll explore the data a little further. First, we'll take a quick look at min, 1st quartile, median, mean, 2nd quartile, max etc.

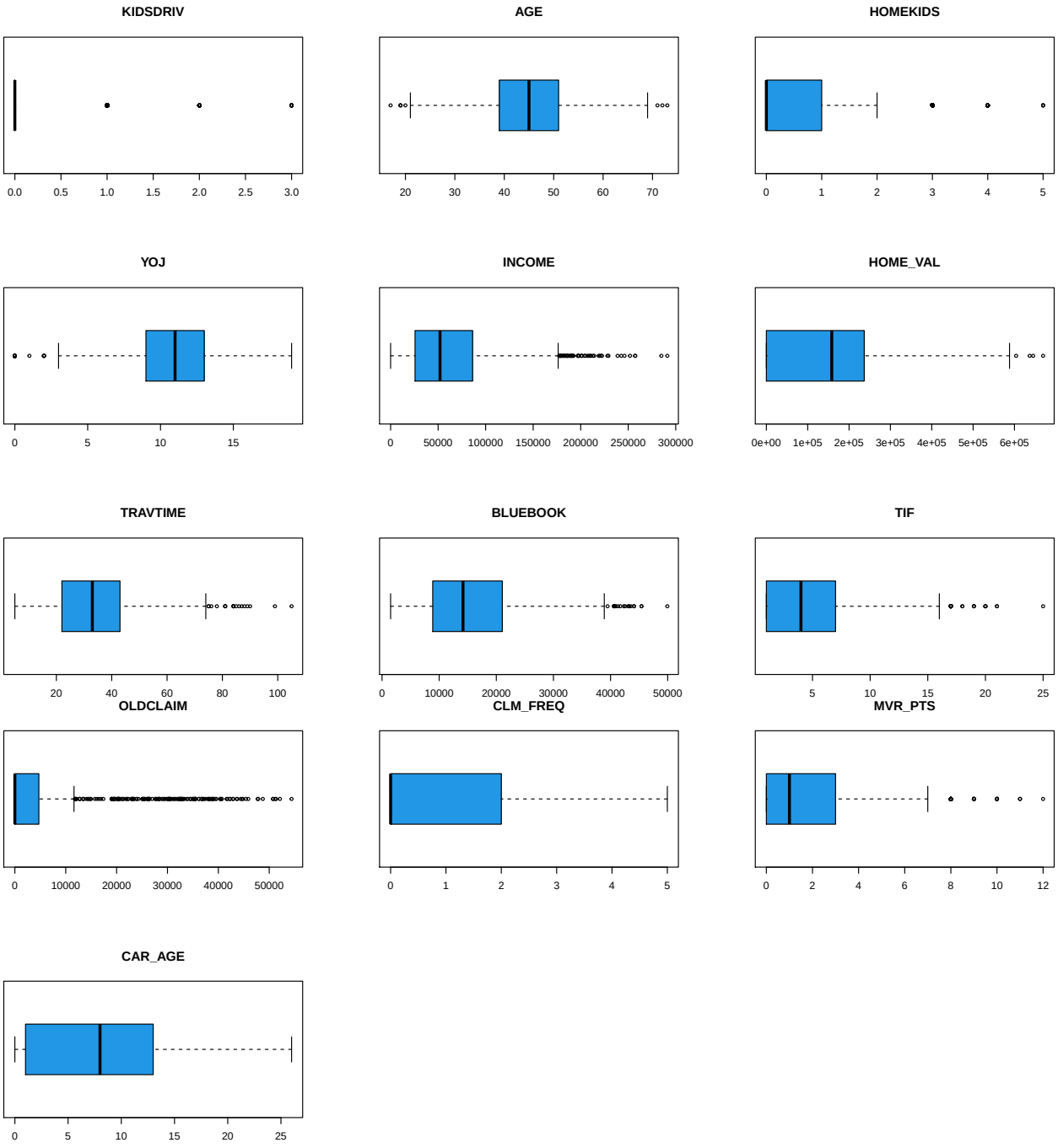
	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME
	Min. : NA	Min. : NA	Min. :0.0000	Min. :17.00	Min. :0.0000	Min. : 0.0	Min. : 0
	1st Qu.: NA	1st Qu.: NA	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000	1st Qu.: 9.0	1st Qu.: 25
	Median : NA	Median : NA	Median :0.0000	Median :45.00	Median :0.0000	Median :11.0	Median : 5
	Mean :NaN	Mean :NaN	Mean :0.1625	Mean :45.02	Mean :0.7174	Mean :10.4	Mean : 60
	3rd Qu.: NA	3rd Qu.: NA	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:13.0	3rd Qu.: 8
	Max. : NA	Max. : NA	Max. :3.0000	Max. :73.00	Max. :5.0000	Max. :19.0	Max. :291
	NA's :2141	NA's :2141	NA	NA	NA	NA	NA

Data reordering

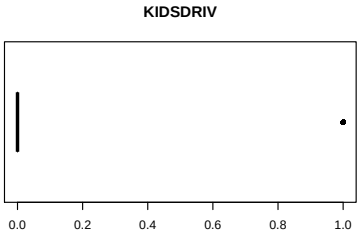
For downstream analysis, we'll reorder the columns into categorical, numeric and target.

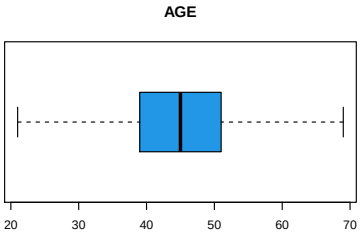
Boxplots

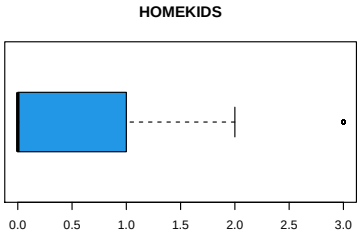
Let's take a first look at the boxplots

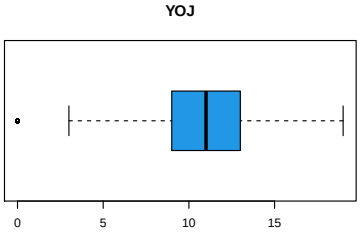


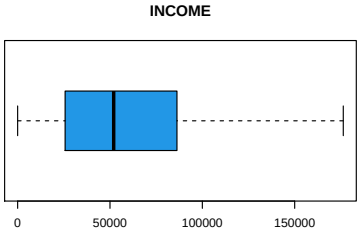
The boxplots show that some of the variables have outliers in them. So, we'll cap them.

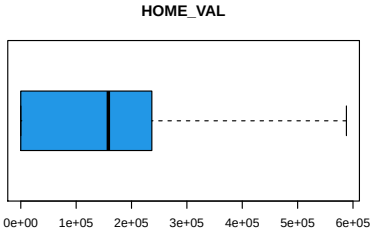


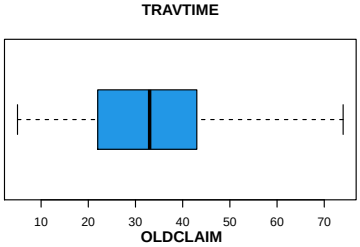


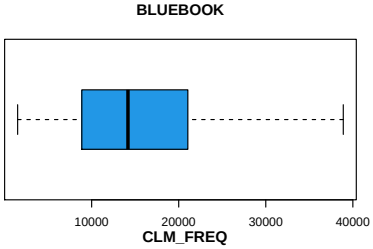


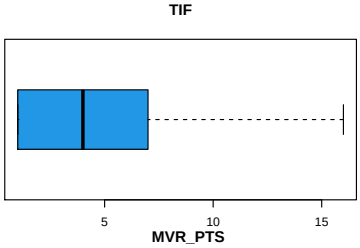


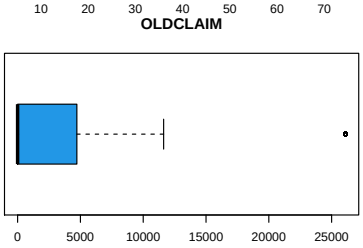


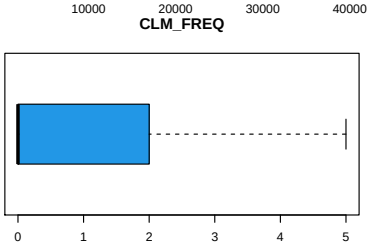


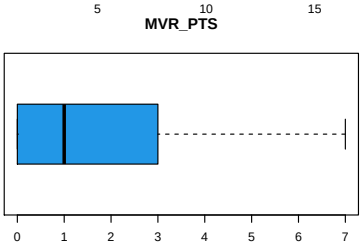


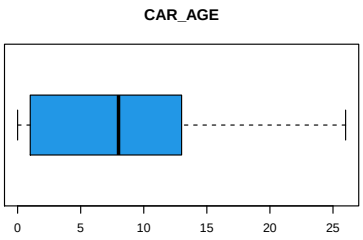












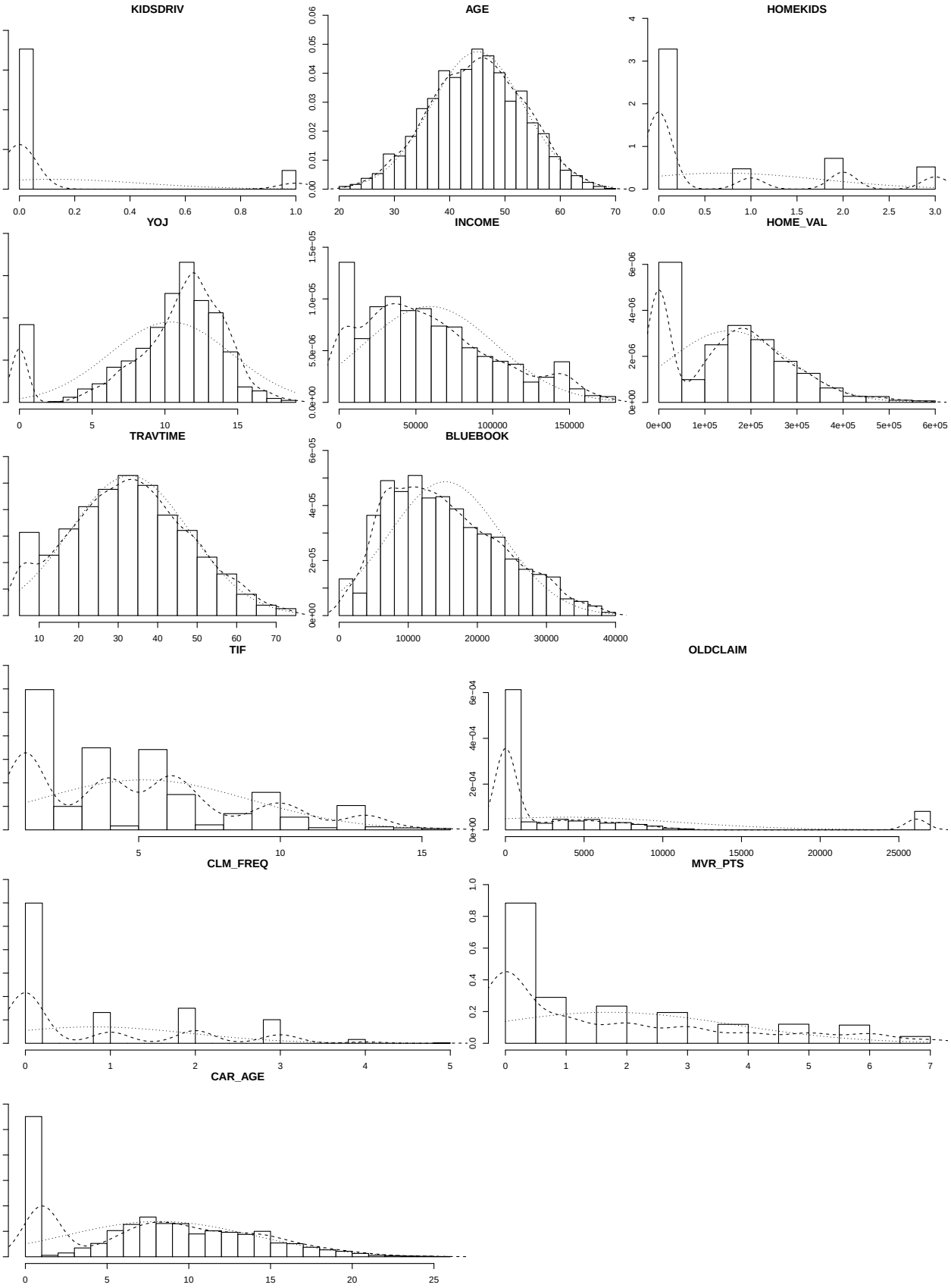
The fields AGE, HOMEKIDS, INCOME, HOME_VAL, TRVTIME, BLUEBOOK, TIF, CLM_FREQ, MVR_PTS, CAR_AGE have higher variance.

Let's ignore the boxplots for TARGET_FLAG and TARGET_AMT.

We'll do the boxplots differently, with gglplot, to check if there are any differences.

Histograms

Histograms tell us how the data is distributed in the dataset (numeric fields).


KIDSDRIV
AGE
HOMEKIDS
YOJ
INCOME
HOME_VAL
TRAVTIME
BLUEBOOK
TIF
OLDCLAIM
CLM_FREQ
MVR_PTS
CAR_AGE

The histograms show that AGE, YOJ, HOME_VAL, TRAVTIME, BLUEBOOK and CAR_AGE are approximately normally distributed. HOME_VALUE, CAR_AGE and CLM_FREQ are quite dispersed.

Categorical variables

Now, we'll explore the Categorical variables.

```
## PARENT1:

##
##   No  Yes
## 1875  266

## MSTATUS:

##
##   No  Yes
##  847 1294

## SEX:

##
##    F    M
## 1170  971

## EDUCATION:

##
## <High School    Bachelors  High School      Masters          PhD
##          312          581          622          420          206

## JOB:

##
##               Blue Collar     Clerical       Doctor   Home Maker       Lawyer
##          139          463          319           75          202          196
##      Manager Professional      Student
##          269          291          187

## CAR_USE:

##
## Commercial    Private
##        760       1381

## CAR_TYPE:

##
##     Minivan Panel Truck      Pickup  Sports Car         SUV         Van
##         549         177         383         272         589         171

## RED_CAR:

##
##   no  yes
## 1543  598

## REVOKED:

##
##   No  Yes
## 1880  261

## URBANICITY:

##
## Highly Rural/ Rural Highly Urban/ Urban
```

```
##                 403                1738
```

Observation: In JOB columns, 139 rows are empty. So, we' have to'll impute them with "Unknown".

```
## JOB:
```

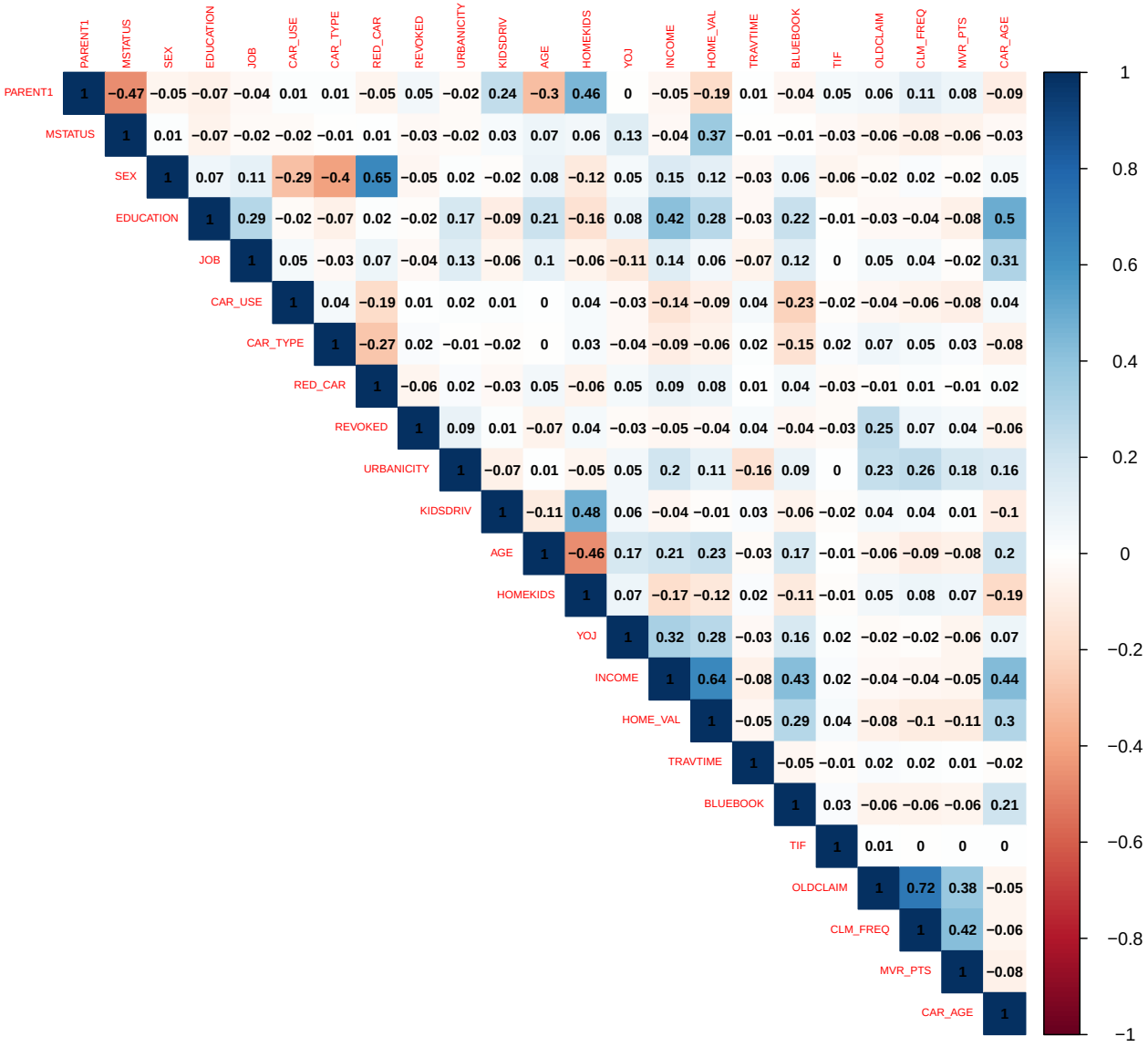
```
##
##  Blue Collar     Clerical       Doctor   Home Maker       Lawyer      Manager
##          463          319           75          202          196          269
## Professional      Student      Unknown
##          291          187          139
```

Correlations.

At this point the data is prepared. So, we'll explore the top correlated variables.

There are 25 variables, among which 15 are numeric and 10 are non-categorical. In order to find pai-wise correlations, we'll give numerical values to the correlated variables.

Now, we'll look at the correlation matrix of the variables.



At this point exploration, preparation and pair-wise correlations of insurance_evaluation_data.csv are done. So, We'll begin the model building process.

Building Models

Now, we are ready to build models. We will build model using **insurance_training_data.csv** and compare the models. Then we will use the best model to predict on **insurance-evaluation-data.csv**

Our task is to classify the TARGET_FLAG variable using logistic regression and predict the value of TARGET_AMT using Linear Regression.

We split Ins_train_cap_imputed into training and test in 80/20 ratio and named as **Ins_train_cap_imputed_trn** and **Ins_train_cap_imputed_tst**.

Logistic Regression Model

Here we will build our first logistic regression model

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT, family = binomial,
##     data = Ins_train_cap_imputed_trn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2811  -0.7141  -0.3957   0.6546   3.1779
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.307e+00  3.194e-01  -7.223 5.08e-13 ***
## PARENT1Yes                     4.255e-01  1.243e-01   3.423 0.000620 ***
## MSTATUSYes                    -4.953e-01  9.438e-02  -5.248 1.54e-07 ***
## SEXM                           5.822e-02  1.250e-01   0.466 0.641461
## EDUCATIONBachelors            -3.678e-01  1.293e-01  -2.845 0.004444 **
## EDUCATIONHigh School          -4.105e-02  1.069e-01  -0.384 0.700977
## EDUCATIONMmasters             -2.518e-01  2.007e-01  -1.254 0.209750
## EDUCATIONPhD                  -1.472e-01  2.353e-01  -0.625 0.531701
## JOBClerical                    5.569e-02  1.201e-01   0.464 0.642861
## JOBDoctor                     -5.723e-01  3.115e-01  -1.837 0.066213 .
## JOBHome Maker                 -1.221e-01  1.734e-01  -0.704 0.481280
## JOBLawyer                     -1.831e-01  2.116e-01  -0.866 0.386725
## JOBManager                    -8.237e-01  1.566e-01  -5.258 1.45e-07 ***
## JOBProfessional               -1.006e-01  1.348e-01  -0.746 0.455389
## JOBStudent                    -1.626e-01  1.458e-01  -1.116 0.264578
## JOBUnknown                    -3.681e-01  2.083e-01  -1.767 0.077250 .
## CAR_USEPrivate                -8.013e-01  1.024e-01  -7.822 5.19e-15 ***
## CAR_TYPEPanel Truck            5.489e-01  1.813e-01   3.027 0.002467 **
## CAR_TYPEPickup                 5.341e-01  1.125e-01   4.746 2.07e-06 ***
## CAR_TYPESports Car             1.046e+00  1.445e-01   7.238 4.54e-13 ***
## CAR_TYPESUV                    7.814e-01  1.237e-01   6.316 2.68e-10 ***
## CAR_TYPEVan                    7.143e-01  1.404e-01   5.087 3.63e-07 ***
## RED_CARyes                    -1.677e-02  9.685e-02  -0.173 0.862543
## REVOKEDYes                     7.893e-01  1.029e-01   7.671 1.70e-14 ***
## URBANICITYHighly Urban/ Urban  2.443e+00  1.256e-01  19.446  < 2e-16 ***
## KIDSDRIV                       6.028e-01  1.092e-01   5.521 3.38e-08 ***
## AGE                           -3.527e-03  4.611e-03  -0.765 0.444233
## HOMEKIDS                       4.525e-02  4.474e-02   1.011 0.311861
```

```
## YOJ                          -1.539e-02  9.321e-03  -1.651 0.098730 .
## INCOME                       -3.104e-06  1.373e-06  -2.261 0.023748 *
## HOME_VAL                     -1.551e-06  3.804e-07  -4.077 4.56e-05 ***
## TRAVTIME                      1.629e-02  2.182e-03   7.468 8.12e-14 ***
## BLUEBOOK                     -2.282e-05  6.096e-06  -3.744 0.000181 ***
## TIF                          -6.065e-02  8.543e-03  -7.099 1.25e-12 ***
## OLDCLAIM                     -1.301e-05  5.396e-06  -2.410 0.015939 *
## CLM_FREQ                      1.816e-01  3.286e-02   5.525 3.30e-08 ***
## MVR_PTS                       1.088e-01  1.653e-02   6.580 4.69e-11 ***
## CAR_AGE                      -1.344e-02  8.110e-03  -1.657 0.097469 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7535.7  on 6528  degrees of freedom
## Residual deviance: 5821.5  on 6491  degrees of freedom
## AIC: 5897.5
##
## Number of Fisher Scoring iterations: 5
```

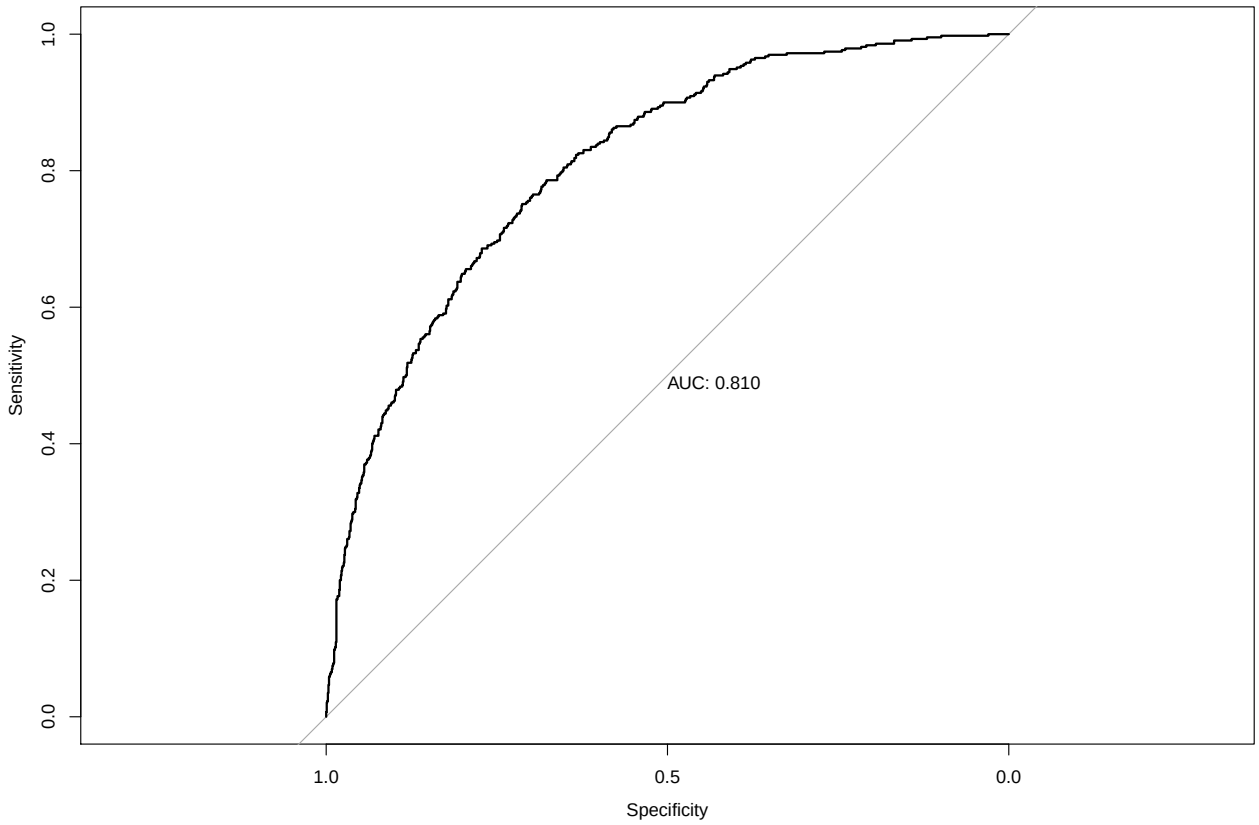
The important metric is AIC: 5899.5.

Now, we'll predict on the test set **Ins__train__cap__imputed__tst**.

Let's look at the Confusion Matrix.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    0
##          1  182   96
##          0  248 1106
##
##                Accuracy : 0.7892
##                  95% CI : (0.7686, 0.8088)
##     No Information Rate : 0.7365
##     P-Value [Acc > NIR] : 4.456e-07
##
##                   Kappa : 0.3874
##
##  Mcnemar's Test P-Value : 3.908e-16
##
##             Sensitivity : 0.4233
##             Specificity : 0.9201
##          Pos Pred Value : 0.6547
##          Neg Pred Value : 0.8168
##              Prevalence : 0.2635
##          Detection Rate : 0.1115
##    Detection Prevalence : 0.1703
##       Balanced Accuracy : 0.6717
##
##        'Positive' Class : 1
##
```

Plotting the AUC under roc curve.



Here we note the following important metrics.

Accuracy of this model is 0.7806.

AUC of the model is 0.801.

Logistic Regression Model

We'll build our second Logistic Regression model

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE +
##     URBANICITY, family = binomial, data = Ins_train_cap_imputed_trn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3165  -0.7197  -0.4068   0.6584   3.1706
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.433e+00  2.126e-01 -11.446  < 2e-16 ***
## KIDSDRIV                       5.869e-01  1.069e-01   5.492 3.98e-08 ***
## HOMEKIDS                       5.551e-02  4.051e-02   1.370 0.170651
## INCOME                        -4.371e-06  1.205e-06  -3.627 0.000286 ***
## PARENT1Yes                     4.267e-01  1.228e-01   3.476 0.000508 ***
## HOME_VAL                      -1.462e-06  3.686e-07  -3.967 7.27e-05 ***
## MSTATUSYes                    -5.216e-01  9.317e-02  -5.599 2.16e-08 ***
## EDUCATIONBachelors            -5.137e-01  1.175e-01  -4.373 1.23e-05 ***
## EDUCATIONHigh School          -1.060e-01  1.035e-01  -1.024 0.305820
## EDUCATIONMasters              -4.986e-01  1.483e-01  -3.361 0.000776 ***
## EDUCATIONPhD                  -5.015e-01  1.813e-01  -2.766 0.005671 **
## TRAVTIME                       1.693e-02  2.165e-03   7.819 5.32e-15 ***
## CAR_USEPrivate                -8.847e-01  8.096e-02 -10.927  < 2e-16 ***
## BLUEBOOK                      -2.468e-05  5.456e-06  -4.524 6.06e-06 ***
## TIF                           -6.054e-02  8.482e-03  -7.138 9.45e-13 ***
## CAR_TYPEPanel Truck            4.804e-01  1.600e-01   3.003 0.002675 **
## CAR_TYPEPickup                 4.675e-01  1.093e-01   4.278 1.88e-05 ***
## CAR_TYPESports Car             9.879e-01  1.182e-01   8.355  < 2e-16 ***
## CAR_TYPESUV                    7.340e-01  9.487e-02   7.737 1.02e-14 ***
## CAR_TYPEVan                    6.730e-01  1.321e-01   5.094 3.51e-07 ***
## CLM_FREQ                       1.392e-01  2.848e-02   4.887 1.02e-06 ***
## REVOKEDYes                     6.767e-01  8.987e-02   7.529 5.10e-14 ***
## MVR_PTS                        1.093e-01  1.629e-02   6.710 1.94e-11 ***
## CAR_AGE                       -1.325e-02  8.064e-03  -1.643 0.100374
## URBANICITYHighly Urban/ Urban  2.396e+00  1.252e-01  19.142  < 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: 7535.7  on 6528  degrees of freedom
## Residual deviance: 5875.8  on 6504  degrees of freedom
## AIC: 5925.8
##
## Number of Fisher Scoring iterations: 5
```

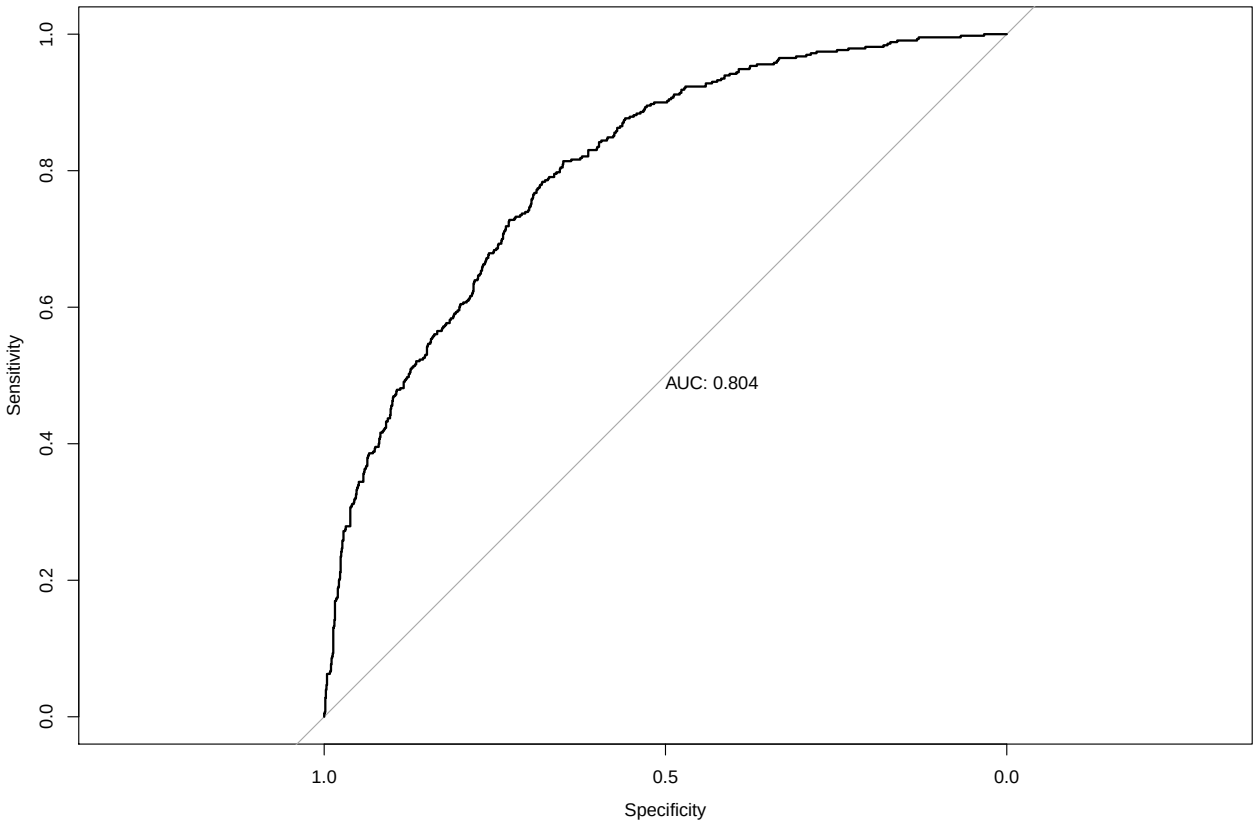
The important metric is AIC: 5930.1.

Now, we'll predict on the test set **Ins__train__cap__imputed__tst**.

Let's look at the Confusion Matrix.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    0
##          1  173   97
##          0  257 1105
##
##                Accuracy : 0.7831
##                  95% CI : (0.7623, 0.8029)
##     No Information Rate : 0.7365
##     P-Value [Acc > NIR] : 7.596e-06
##
##                   Kappa : 0.3653
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4023
##             Specificity : 0.9193
##          Pos Pred Value : 0.6407
##          Neg Pred Value : 0.8113
##              Prevalence : 0.2635
##          Detection Rate : 0.1060
##    Detection Prevalence : 0.1654
##       Balanced Accuracy : 0.6608
##
##        'Positive' Class : 1
##
```

Plotting the AUC under roc curve.



Here we note the following important metrics.

Accuracy of this model is 0.7855.

AUC of the model is 0.802.

Logistic Regression Model

In the third Logistic Regression model, we'll do stepwise model selection.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + EDUCATION + TRAVTIME + CAR_USE + BLUEBOOK + TIF +
##     CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY,
##     family = binomial, data = Ins_train_cap_imputed_trn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3546  -0.7204  -0.4077   0.6547   3.1609
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.414e+00  2.120e-01 -11.388  < 2e-16 ***
## KIDSDRIV                      6.438e-01  9.870e-02   6.523 6.90e-11 ***
## INCOME                       -4.350e-06  1.204e-06  -3.611 0.000305 ***
## PARENT1Yes                    5.140e-01  1.050e-01   4.895 9.81e-07 ***
## HOME_VAL                     -1.501e-06  3.680e-07  -4.081 4.49e-05 ***
## MSTATUSYes                   -4.821e-01  8.851e-02  -5.447 5.12e-08 ***
## EDUCATIONBachelors           -5.240e-01  1.172e-01  -4.471 7.78e-06 ***
## EDUCATIONHigh School         -1.117e-01  1.034e-01  -1.080 0.280101
## EDUCATIONMasters             -5.147e-01  1.478e-01  -3.482 0.000498 ***
## EDUCATIONPhD                 -5.171e-01  1.809e-01  -2.859 0.004251 **
## TRAVTIME                      1.686e-02  2.164e-03   7.788 6.78e-15 ***
## CAR_USEPrivate               -8.860e-01  8.096e-02 -10.944  < 2e-16 ***
## BLUEBOOK                     -2.492e-05  5.454e-06  -4.569 4.89e-06 ***
## TIF                          -6.039e-02  8.480e-03  -7.122 1.07e-12 ***
## CAR_TYPEPanel Truck           4.820e-01  1.599e-01   3.014 0.002580 **
## CAR_TYPEPickup                4.653e-01  1.093e-01   4.258 2.06e-05 ***
## CAR_TYPESports Car            9.889e-01  1.182e-01   8.365  < 2e-16 ***
## CAR_TYPESUV                   7.384e-01  9.479e-02   7.790 6.68e-15 ***
## CAR_TYPEVan                   6.717e-01  1.321e-01   5.084 3.70e-07 ***
## CLM_FREQ                      1.394e-01  2.847e-02   4.896 9.78e-07 ***
## REVOKEDYes                    6.797e-01  8.983e-02   7.567 3.82e-14 ***
## MVR_PTS                       1.098e-01  1.629e-02   6.743 1.55e-11 ***
## CAR_AGE                      -1.336e-02  8.061e-03  -1.658 0.097342 .
## URBANICITYHighly Urban/ Urban  2.393e+00  1.252e-01  19.122  < 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: 7535.7  on 6528  degrees of freedom
## Residual deviance: 5877.7  on 6505  degrees of freedom
## AIC: 5925.7
##
## Number of Fisher Scoring iterations: 5
```

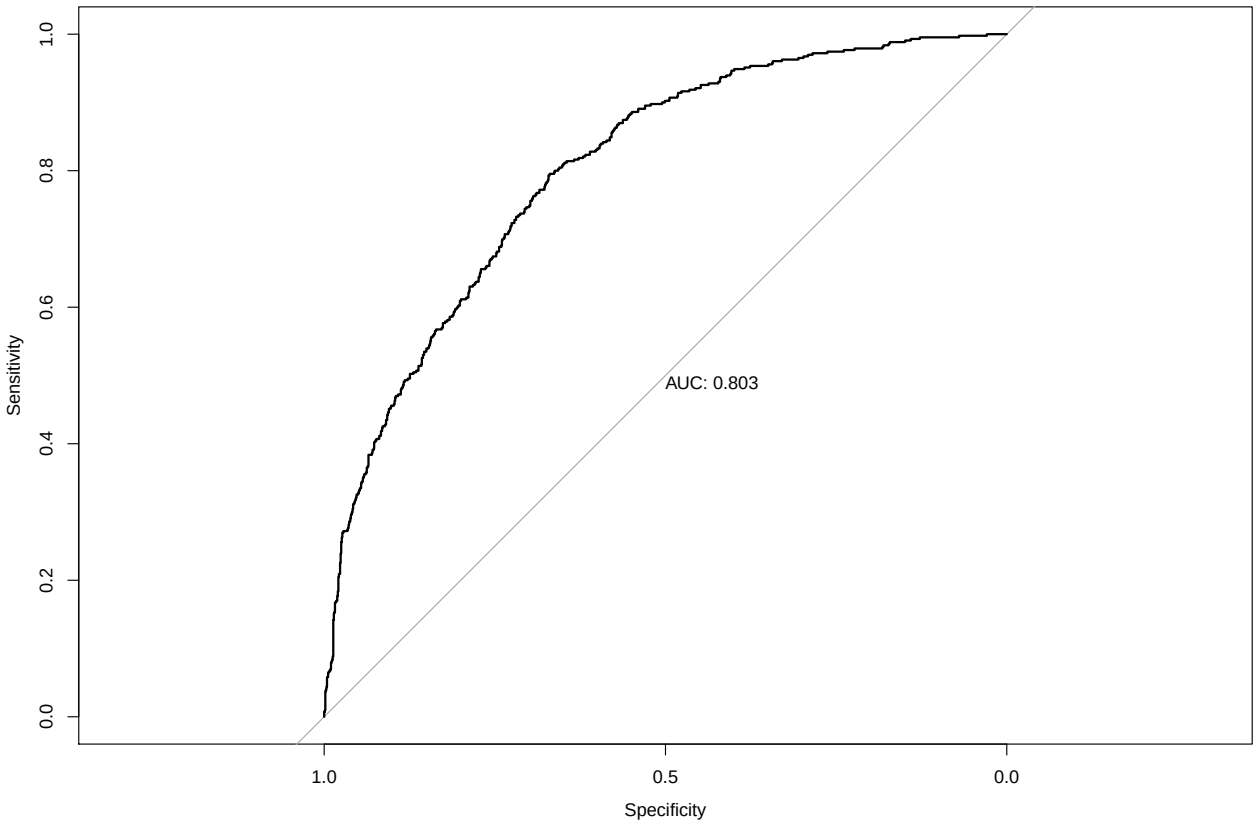
The important metric is AIC: 5959.5.

Now, we'll predict on the test set **Ins_train_cap_imputed_tst**.

Creation of Confusion Matrix.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    1    0
##          1  175   95
##          0  255 1107
##
##                Accuracy : 0.7855
##                  95% CI : (0.7648, 0.8052)
##     No Information Rate : 0.7365
##     P-Value [Acc > NIR] : 2.548e-06
##
##                   Kappa : 0.3724
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.4070
##             Specificity : 0.9210
##          Pos Pred Value : 0.6481
##          Neg Pred Value : 0.8128
##              Prevalence : 0.2635
##          Detection Rate : 0.1072
##    Detection Prevalence : 0.1654
##       Balanced Accuracy : 0.6640
##
##        'Positive' Class : 1
##
```

Plotting the AUC under roc curve.



Here we note the following important metrics.

Accuracy of this model is 0.7874.

AUC of the model is 0.802.

At this point three Logistic Regression models were built. The accuracy was highest in the third model. We will use the third model on evaluation dataset, for classification.

Linear Regression model

First Linear Regression model

Since our goal is to predict the TARGET_AMT, and not classify (as we did in Logistic Regression), we'll build linear regression model for TARGET_AMT.

```
##
## Call:
## lm(formula = TARGET_AMT ~ . - TARGET_FLAG, data = Ins_train_cap_imputed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5161  -1704   -755    361 103695
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.232e+01  4.763e+02   0.110 0.912536
## PARENT1Yes                      5.458e+02  2.042e+02   2.672 0.007546 **
## MSTATUSYes                     -5.391e+02  1.445e+02  -3.732 0.000191 ***
## SEXM                            3.584e+02  1.833e+02   1.955 0.050616 .
## EDUCATIONBachelors             -2.454e+02  2.051e+02  -1.196 0.231612
## EDUCATIONHigh School           -8.887e+01  1.720e+02  -0.517 0.605496
## EDUCATIONMasters                5.224e+01  3.002e+02   0.174 0.861878
## EDUCATIONPhD                    2.658e+02  3.533e+02   0.753 0.451749
## JOBClerical                     1.336e+01  1.933e+02   0.069 0.944904
## JOBDoctor                      -9.781e+02  4.354e+02  -2.246 0.024711 *
## JOBHome Maker                  -1.658e+02  2.704e+02  -0.613 0.539794
## JOBLawyer                      -2.785e+02  3.108e+02  -0.896 0.370237
## JOBManager                     -9.825e+02  2.344e+02  -4.191 2.80e-05 ***
## JOBProfessional                -4.277e+01  2.131e+02  -0.201 0.840933
## JOBStudent                     -2.407e+02  2.354e+02  -1.023 0.306564
## JOBUnknown                     -5.104e+02  3.215e+02  -1.588 0.112429
## CAR_USEPrivate                 -7.941e+02  1.644e+02  -4.831 1.38e-06 ***
## CAR_TYPEPanel Truck             2.682e+02  2.765e+02   0.970 0.332104
## CAR_TYPEPickup                  3.729e+02  1.708e+02   2.184 0.028989 *
## CAR_TYPESports Car              1.015e+03  2.178e+02   4.662 3.18e-06 ***
## CAR_TYPESUV                     7.497e+02  1.794e+02   4.179 2.95e-05 ***
## CAR_TYPEVan                     5.171e+02  2.132e+02   2.425 0.015332 *
## RED_CARyes                     -3.898e+01  1.491e+02  -0.261 0.793715
## REVOKEDYes                      5.895e+02  1.743e+02   3.382 0.000724 ***
## URBANICITYHighly Urban/ Urban   1.670e+03  1.392e+02  12.001  < 2e-16 ***
## KIDSDRIV                        6.055e+02  1.790e+02   3.383 0.000720 ***
## AGE                             6.590e+00  7.187e+00   0.917 0.359220
## HOMEKIDS                        7.647e+01  6.991e+01   1.094 0.274073
## YOJ                            -2.183e+00  1.463e+01  -0.149 0.881384
## INCOME                         -4.026e-03  2.094e-03  -1.923 0.054526 .
## HOME_VAL                       -9.639e-04  5.898e-04  -1.634 0.102233
## TRAVTIME                        1.272e+01  3.339e+00   3.810 0.000140 ***
## BLUEBOOK                        1.408e-02  8.977e-03   1.569 0.116739
## TIF                            -5.099e+01  1.294e+01  -3.940 8.22e-05 ***
## OLDCLAIM                       -1.678e-02  9.120e-03  -1.840 0.065821 .
## CLM_FREQ                        1.648e+02  5.687e+01   2.898 0.003767 **
## MVR_PTS                         1.733e+02  2.790e+01   6.213 5.45e-10 ***
## CAR_AGE                        -2.881e+01  1.233e+01  -2.335 0.019548 *
```

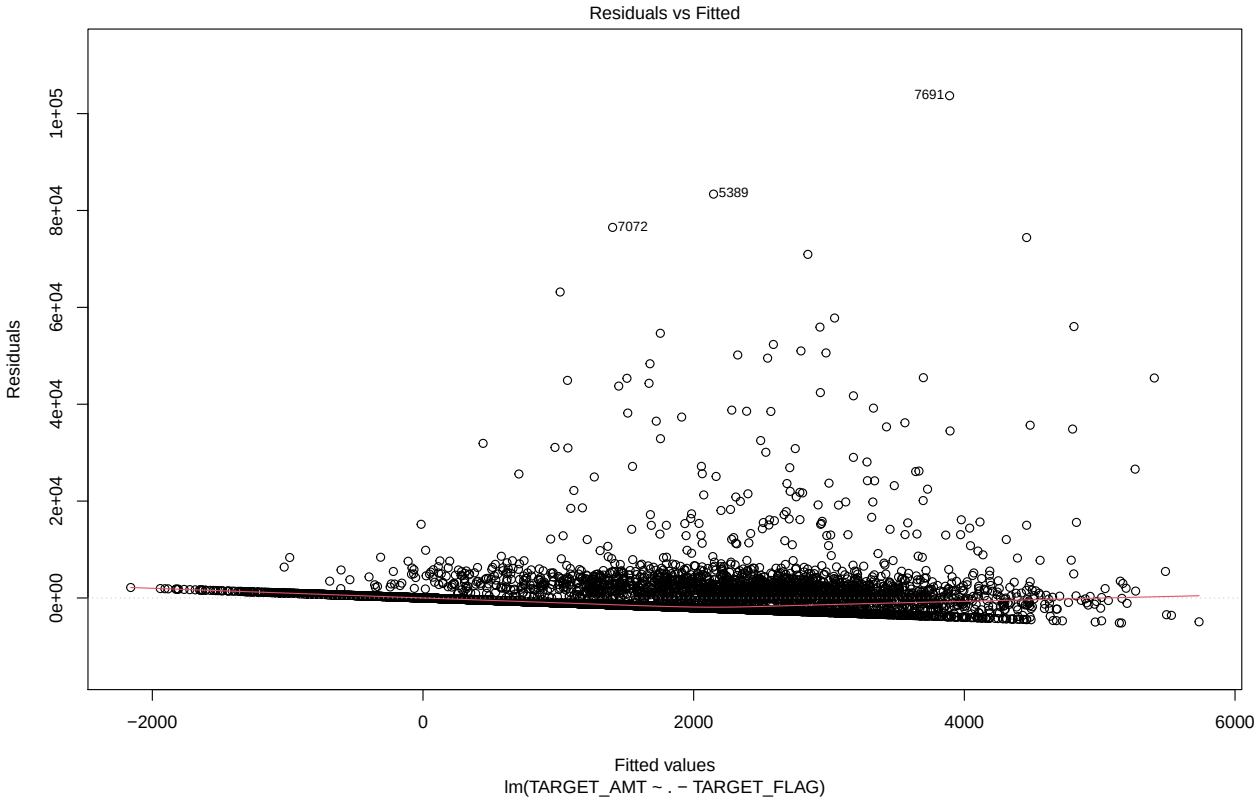
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4544 on 8123 degrees of freedom
## Multiple R-squared:  0.07111,    Adjusted R-squared:  0.06688
## F-statistic: 16.81 on 37 and 8123 DF,  p-value: < 2.2e-16
```

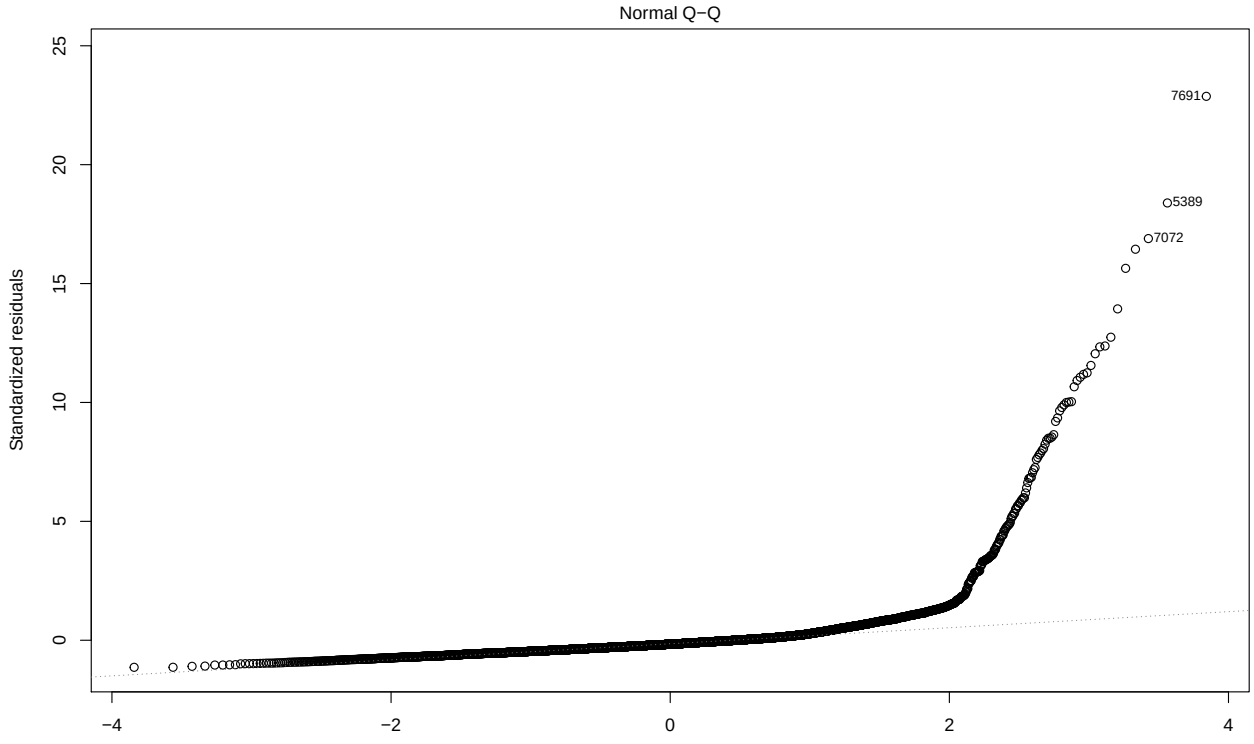
Here we note the following important metrics.

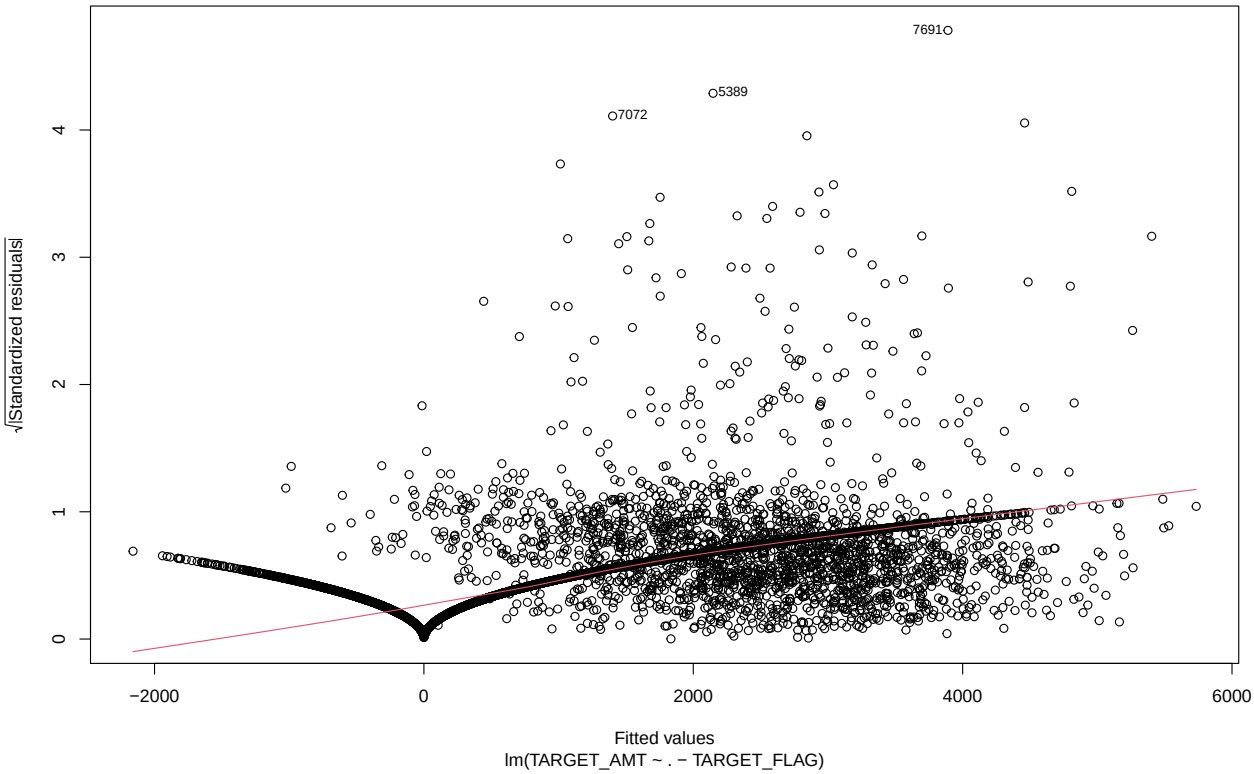
R-squared: 0.07094.

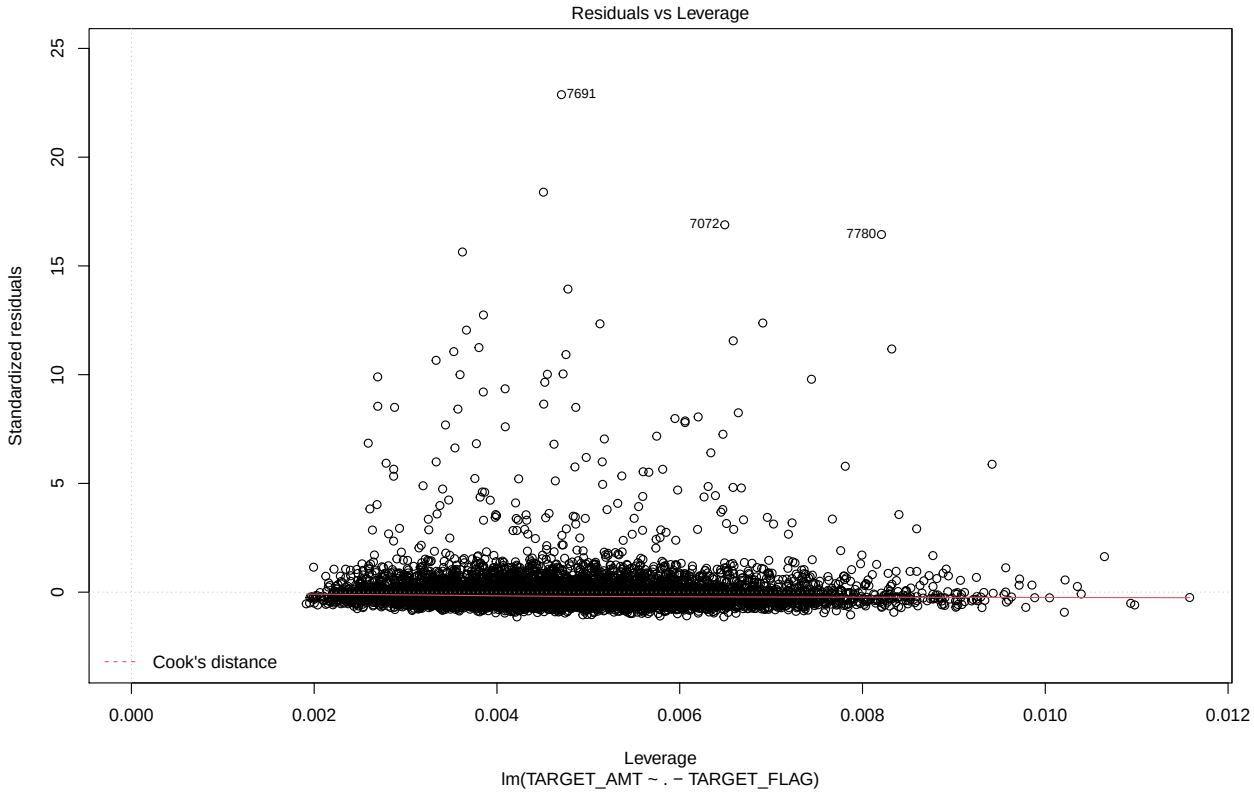
Adjusted R-squared: 0.06671.

The R-squared values are far from 1. So, the model is not good. In below plot, the points are widely scattered, but not linearly.




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TARGET_AMT ~ . − TARGET_FLAG)
7691
5389
7072


Scale−Location
√|Standardized residuals|
Fitted values
lm(TARGET_AMT ~ . − TARGET_FLAG)
7691
5389
7072


Residuals vs Leverage
Standardized residuals
25
20
15
10
5
0
7691
7072
7780
Cook's distance
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Leverage
lm(TARGET_AMT ~ . − TARGET_FLAG)

Linear Regression model

In Second Linear Regression model, we'll do stepwise AIC.

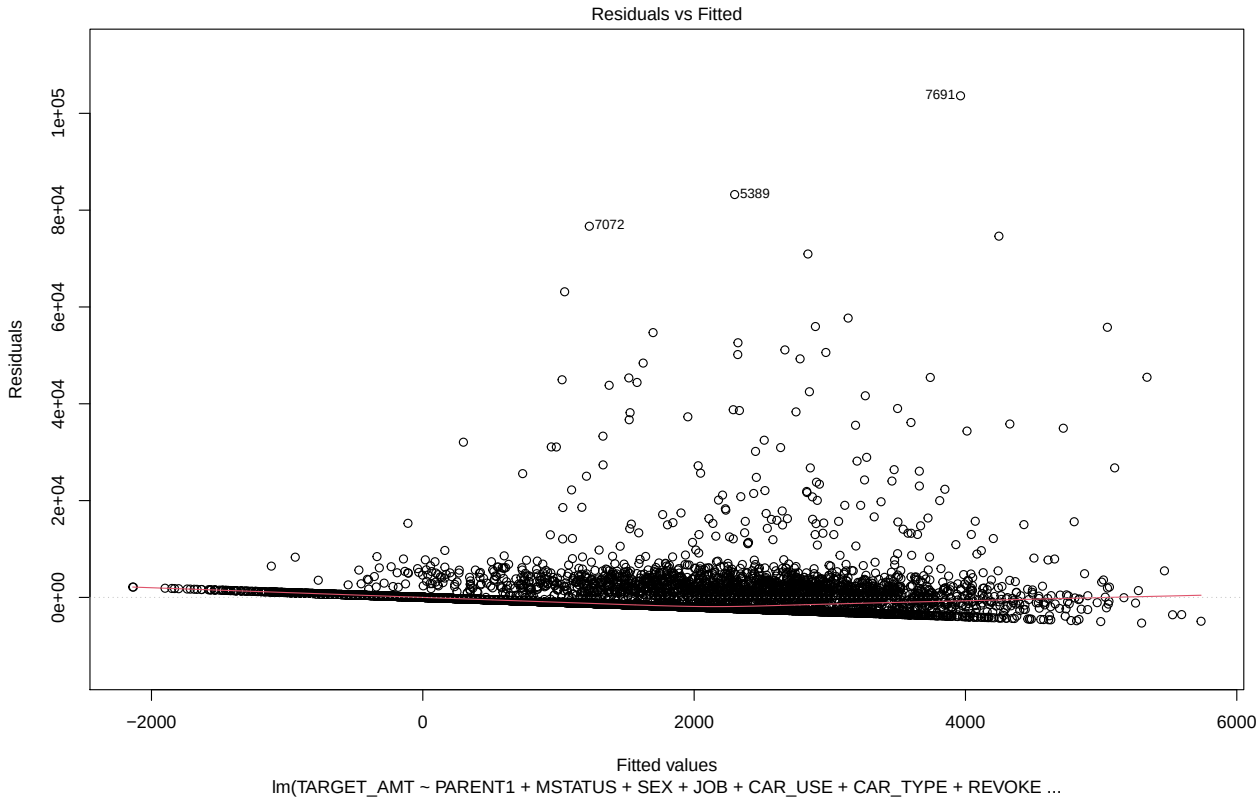
```
##
## Call:
## lm(formula = TARGET_AMT ~ PARENT1 + MSTATUS + SEX + JOB + CAR_USE +
##     CAR_TYPE + REVOKED + URBANICITY + KIDSDRIV + INCOME + HOME_VAL +
##     TRAVTIME + BLUEBOOK + TIF + OLDCLAIM + CLM_FREQ + MVR_PTS +
##     CAR_AGE, data = Ins_train_cap_imputed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5298  -1692   -757    350 103623
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.822e+02  3.406e+02   0.535 0.592812
## PARENT1Yes                     6.013e+02  1.783e+02   3.372 0.000751 ***
## MSTATUSYes                    -5.045e+02  1.383e+02  -3.647 0.000267 ***
## SEXM                           3.313e+02  1.604e+02   2.065 0.038932 *
## JOBClerical                    2.587e+01  1.922e+02   0.135 0.892928
## JOBDoctor                     -6.198e+02  3.548e+02  -1.747 0.080664 .
## JOBHome Maker                 -1.189e+02  2.484e+02  -0.479 0.632102
## JOBLawyer                     -1.196e+02  2.414e+02  -0.496 0.620213
## JOBManager                    -9.589e+02  2.124e+02  -4.515 6.44e-06 ***
## JOBProfessional               -1.153e+02  1.974e+02  -0.584 0.559182
## JOBStudent                    -2.013e+02  2.250e+02  -0.895 0.370923
## JOBUnknown                    -2.803e+02  2.671e+02  -1.049 0.294155
## CAR_USEPrivate                -7.477e+02  1.568e+02  -4.768 1.89e-06 ***
## CAR_TYPEPanel Truck            3.096e+02  2.733e+02   1.133 0.257379
## CAR_TYPEPickup                 3.991e+02  1.694e+02   2.356 0.018511 *
## CAR_TYPESports Car             1.035e+03  2.164e+02   4.781 1.77e-06 ***
## CAR_TYPESUV                    7.580e+02  1.785e+02   4.245 2.21e-05 ***
## CAR_TYPEVan                    5.347e+02  2.120e+02   2.522 0.011695 *
## REVOKEDYes                     5.923e+02  1.742e+02   3.399 0.000680 ***
## URBANICITYHighly Urban/ Urban  1.666e+03  1.391e+02  11.976  < 2e-16 ***
## KIDSDRIV                       6.902e+02  1.620e+02   4.260 2.07e-05 ***
## INCOME                        -3.883e-03  2.021e-03  -1.921 0.054754 .
## HOME_VAL                      -9.638e-04  5.864e-04  -1.643 0.100320
## TRAVTIME                       1.257e+01  3.337e+00   3.768 0.000165 ***
## BLUEBOOK                       1.466e-02  8.879e-03   1.651 0.098745 .
## TIF                           -5.045e+01  1.293e+01  -3.901 9.67e-05 ***
## OLDCLAIM                      -1.687e-02  9.115e-03  -1.851 0.064255 .
## CLM_FREQ                       1.654e+02  5.684e+01   2.910 0.003629 **
## MVR_PTS                        1.732e+02  2.786e+01   6.216 5.35e-10 ***
## CAR_AGE                       -2.867e+01  1.091e+01  -2.627 0.008632 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4544 on 8131 degrees of freedom
## Multiple R-squared:  0.07035,    Adjusted R-squared:  0.06703
## F-statistic: 21.22 on 29 and 8131 DF,  p-value: < 2.2e-16
```

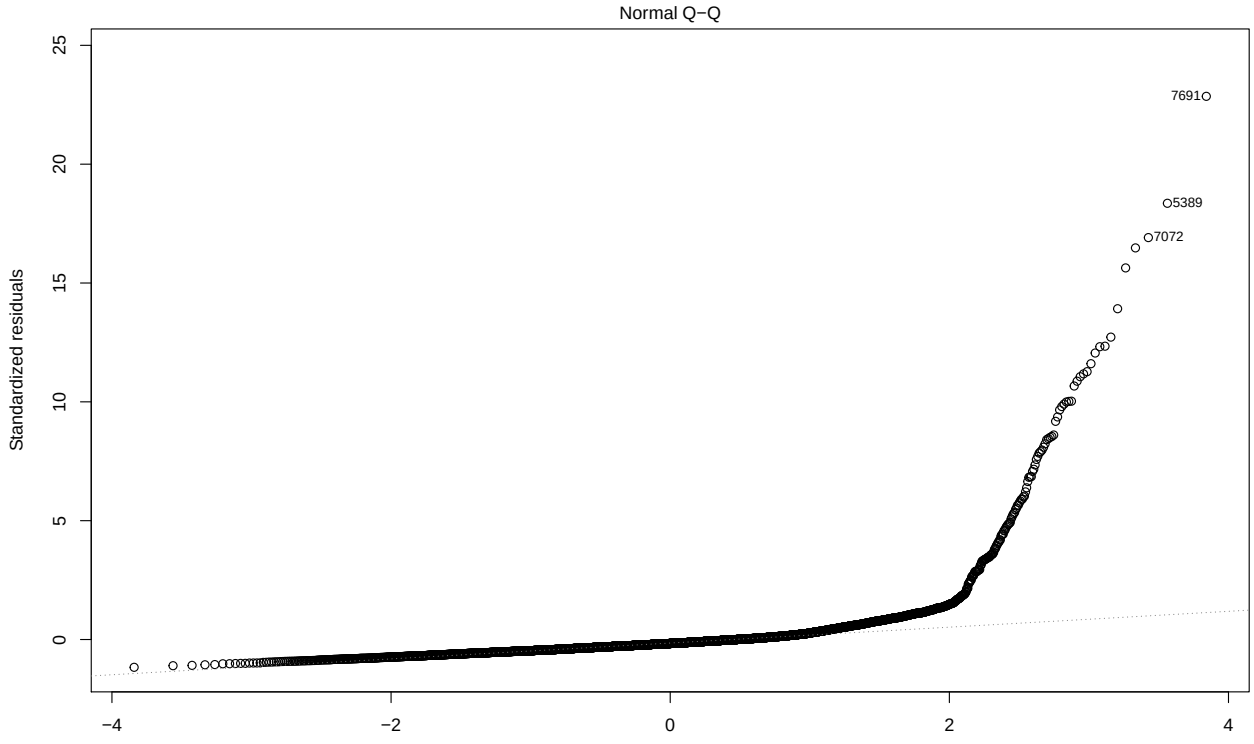
Here we note the following important metrics.

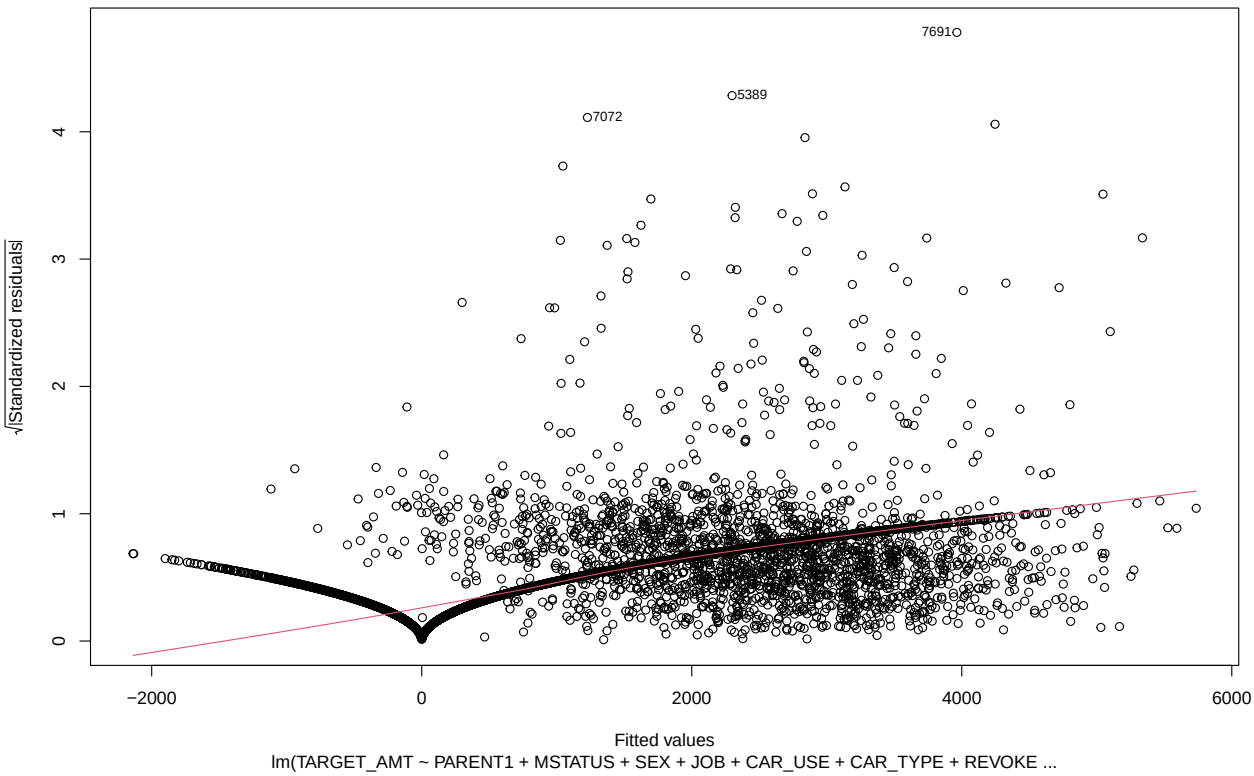
R-squared: 0.07012.

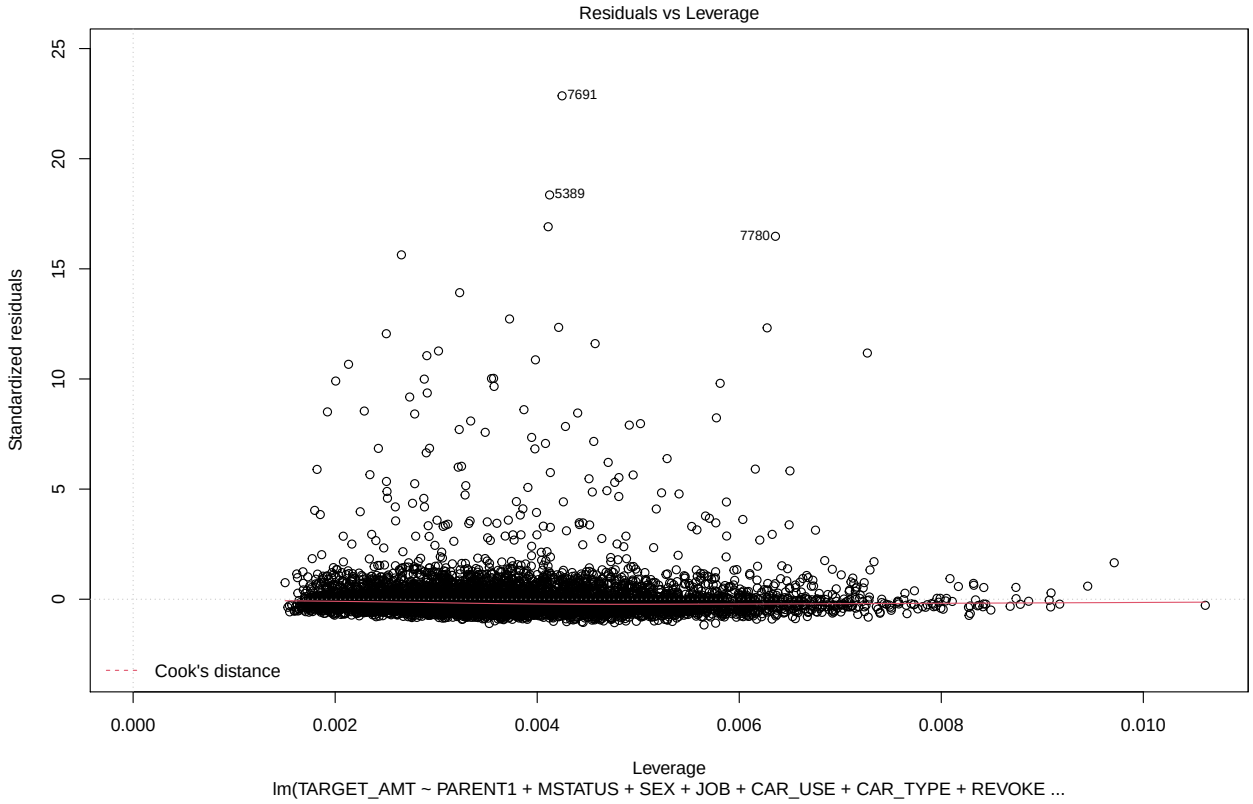
Adjusted R-squared: 0.06692.

The second model slightly improved over the first model. The plot also suggest no proper linear regression.




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(TARGET_AMT ~ PARENT1 + MSTATUS + SEX + JOB + CAR_USE + CAR_TYPE + REVOKE ...
7691
5389
7072


Scale−Location
√|Standardized residuals|
Fitted values
lm(TARGET_AMT ~ PARENT1 + MSTATUS + SEX + JOB + CAR_USE + CAR_TYPE + REVOKE ...
7691
5389
7072


Residuals vs Leverage
Standardized residuals
25
20
15
10
5
0
7691
5389
7780
Cook's distance
0.000
0.002
0.004
0.006
0.008
0.010
Leverage
lm(TARGET_AMT ~ PARENT1 + MSTATUS + SEX + JOB + CAR_USE + CAR_TYPE + REVOKE ...

Model Selection.

We ran three Logistic Regression models and two Linear Regression models. The third Logistic Regression model did best based on the Accuracy and AUC. And the second Linear Regression model did best based on R-squared value.

We'll use these models to predict the evaluation dataset **insurance-evaluation-data.csv**.

Head of the predicted data.

```
##   PARENT1 MSTATUS SEX   EDUCATION          JOB    CAR_USE    CAR_TYPE RED_CAR
## 1      No      No   M   Bachelors      Manager    Private         Van     yes
## 2     Yes      No   M High School      Manager    Private     Minivan      no
## 3     Yes      No   F High School  Blue Collar Commercial         SUV      no
## 4     Yes      No   M High School     Clerical    Private      Pickup      no
## 5      No      No   M High School      Manager    Private     Minivan     yes
## 6      No     Yes   M   Bachelors Professional Commercial Panel Truck      no
##   REVOKED          URBANICITY KIDSDRIV AGE HOMEKIDS YOJ INCOME HOME_VAL
## 1      No Highly Urban/ Urban        0  48        0  11  52881        0
## 2      No Highly Urban/ Urban        1  40        1  11  50815        0
## 3      No Highly Rural/ Rural        0  44        2  12  43486        0
## 4     Yes Highly Rural/ Rural        0  35        2   0  21204        0
## 5      No Highly Urban/ Urban        0  59        0  12  87460        0
## 6      No Highly Urban/ Urban        0  46        0  14  77012   207519
##   TRAVTIME BLUEBOOK TIF OLDCLAIM CLM_FREQ MVR_PTS CAR_AGE TARGET_FLAG
## 1       26    21970   1        0        0       2      10           0
## 2       21    18930   6     3295        1       2       1           0
## 3       30     5900  10        0        0       0      10           0
## 4       74     9230   6        0        0       0       4           0
## 5       45    15420   1    26114        2       4       1           0
## 6        7    25660   1     2119        1       2      12           0
##   TARGET_AMT
## 1          0
## 2          0
## 3          0
## 4          0
## 5          0
## 6          0
```

Write the data to a CSV file