DATA621_Home_Work_4

Dilip Ganesan

7/5/2018

Home Work Assignment 4.

1.DATA EXPLORATION.(Exploratory Data Analysis EDA)

As first step in our EDA, let us load the train data and do a summary statistics on the loaded dataset.

```
# Let us load the train.csv data
train = read.csv('insurance_training_data.csv')
test = read.csv('insurance-evaluation-data.csv')

train = within(train, rm('INDEX'))
test = within(test, rm('INDEX'))

summary(train)
```

```
##   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              INCOME     PARENT1
##  Min.   :0.0000   Min.   : 0.0   $0       : 615   No :7084
##  1st Qu.:0.0000   1st Qu.: 9.0            : 445   Yes:1077
##  Median :0.0000   Median :11.0   $26,840 :   4
##  Mean   :0.7212   Mean   :10.5   $48,509 :   4
##  3rd Qu.:1.0000   3rd Qu.:13.0   $61,790 :   4
##  Max.   :5.0000   Max.   :23.0   $107,375:   3
##                   NA's   :454    (Other) :7086
##      HOME_VAL    MSTATUS      SEX               EDUCATION
##  $0      :2294   Yes :4894   M  :3786   <High School :1203
##          : 464   z_No:3267   z_F:4375   Bachelors    :2242
##  $111,129:   3                          Masters      :1658
##  $115,249:   3                          PhD          : 728
##  $123,109:   3                          z_High School:2330
##  $153,061:   3
##  (Other) :5391
##             JOB           TRAVTIME            CAR_USE        BLUEBOOK
##  z_Blue Collar:1825   Min.   :  5.00   Commercial:3029   $1,500 : 157
##  Clerical     :1271   1st Qu.: 22.00   Private   :5132   $6,000 :  34
##  Professional :1117   Median : 33.00                     $5,800 :  33
##  Manager      : 988   Mean   : 33.49                     $6,200 :  33
##  Lawyer       : 835   3rd Qu.: 44.00                     $6,400 :  31
##  Student      : 712   Max.   :142.00                     $5,900 :  30
##  (Other)      :1413                                      (Other):7843
```

```
##       TIF                CAR_TYPE    RED_CAR       OLDCLAIM
##  Min.   : 1.000   Minivan    :2145   no :5783   $0      :5009
##  1st Qu.: 1.000   Panel Truck: 676   yes:2378   $1,310 :   4
##  Median : 4.000   Pickup     :1389              $1,391 :   4
##  Mean   : 5.351   Sports Car : 907              $4,263 :   4
##  3rd Qu.: 7.000   Van        : 750              $1,105 :   3
##  Max.   :25.000   z_SUV      :2294              $1,332 :   3
##                                                 (Other):3134
##     CLM_FREQ      REVOKED       MVR_PTS          CAR_AGE
##  Min.   :0.0000   No :7161   Min.   : 0.000   Min.   :-3.000
##  1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.0000              Median : 1.000   Median : 8.000
##  Mean   :0.7986              Mean   : 1.696   Mean   : 8.328
##  3rd Qu.:2.0000              3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :5.0000              Max.   :13.000   Max.   :28.000
##                                               NA's   :510
##                 URBANICITY
##  Highly Urban/ Urban  :6492
##  z_Highly Rural/ Rural:1669
##
##
##
##
##
```

We will make sure there is no inappropriate distribution of target(response) variables in our training data.

```
knitr::kable(table(train$TARGET_FLAG))
```

Var1	Freq
0	6008
1	2153

Examination of Data Set.

1. There are 8161 rows and 25 columns(excluding the INDEX) in the train data set. The 2 columns are response variable and rest are predictor variables.

TARGET_FLAG is a binary variable where 1 means that person had a crash. 0 means that person did not had the crash.

TARGET_AMT is the expense because of the accident. 0 when there is no accident.

Statistical Summary of Data Set:

```
summary = describe(train, quant = c(.25,.75))
knitr::kable(summary)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	
TARGET_FLAG	1	8161	0.2638157	0.4407276	0	0.2047787	0.0000	0	1.0	
TARGET_AMT	2	8161	1504.3246481	4704.0269298	0	593.7121106	0.0000	0	107586.1	1
KIDSDRIV	3	8161	0.1710575	0.5115341	0	0.0252719	0.0000	0	4.0	

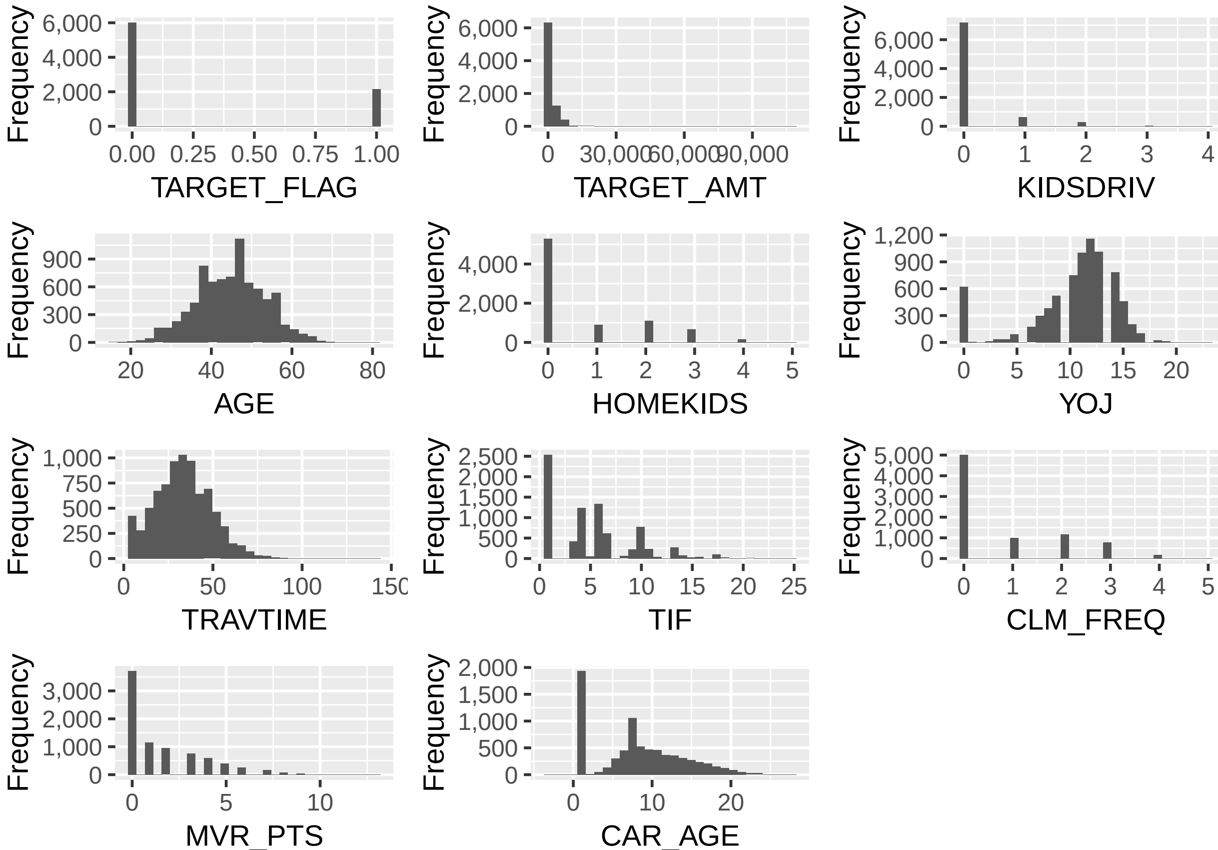
	vars	n	mean	sd	median	trimmed	mad	min	max
AGE	4	8155	44.7903127	8.6275895	45	44.8306513	8.8956	16	81.0
HOMEKIDS	5	8161	0.7212351	1.1163233	0	0.4971665	0.0000	0	5.0
YOJ	6	7707	10.4992864	4.0924742	11	11.0711853	2.9652	0	23.0
INCOME*	7	8161	2875.5505453	2090.6786785	2817	2816.9534385	2799.1488	1	6613.0
PARENT1*	8	8161	1.1319691	0.3384779	1	1.0399755	0.0000	1	2.0
HOME_VAL*	9	8161	1684.8931503	1697.3791897	1245	1516.4994639	1842.8718	1	5107.0
MSTATUS*	10	8161	1.4003186	0.4899929	1	1.3754021	0.0000	1	2.0
SEX*	11	8161	1.5360863	0.4987266	2	1.5451064	0.0000	1	2.0
EDUCATION*	12	8161	3.0906752	1.4448565	3	3.1133405	1.4826	1	5.0
JOB*	13	8161	5.6871707	2.6818733	6	5.8145198	2.9652	1	9.0
TRAVTIME	14	8161	33.4857248	15.9083334	33	32.9954051	16.3086	5	142.0
CAR_USE*	15	8161	1.6288445	0.4831436	2	1.6610507	0.0000	1	2.0
BLUEBOOK*	16	8161	1283.6185516	893.5117428	1124	1259.5665492	1132.7064	1	2789.0
TIF	17	8161	5.3513050	4.1466353	4	4.8402512	4.4478	1	25.0
CAR_TYPE*	18	8161	3.5297145	1.9653570	3	3.5371420	2.9652	1	6.0
RED_CAR*	19	8161	1.2913859	0.4544287	1	1.2392403	0.0000	1	2.0
OLDCLAIM*	20	8161	552.2714128	862.2006829	1	380.3196508	0.0000	1	2857.0
CLM_FREQ	21	8161	0.7985541	1.1584527	0	0.5886047	0.0000	0	5.0
REVOKED*	22	8161	1.1225340	0.3279216	1	1.0281820	0.0000	1	2.0
MVR_PTS	23	8161	1.6955030	2.1471117	1	1.3138306	1.4826	0	13.0
CAR_AGE	24	7651	8.3283231	5.7007424	8	7.9632413	7.4130	-3	28.0
URBANICITY*	25	8161	1.2045093	0.4033673	1	1.1306479	0.0000	1	2.0

1. CAR_AGE, YOJ and AGE has NA values.
2. CAR_AGE min value is -3. This needs some manipulation.
3. TARGET_AMT has a large skewness. We need to do some transformation on this variable.

Visual Exploration of Data set:

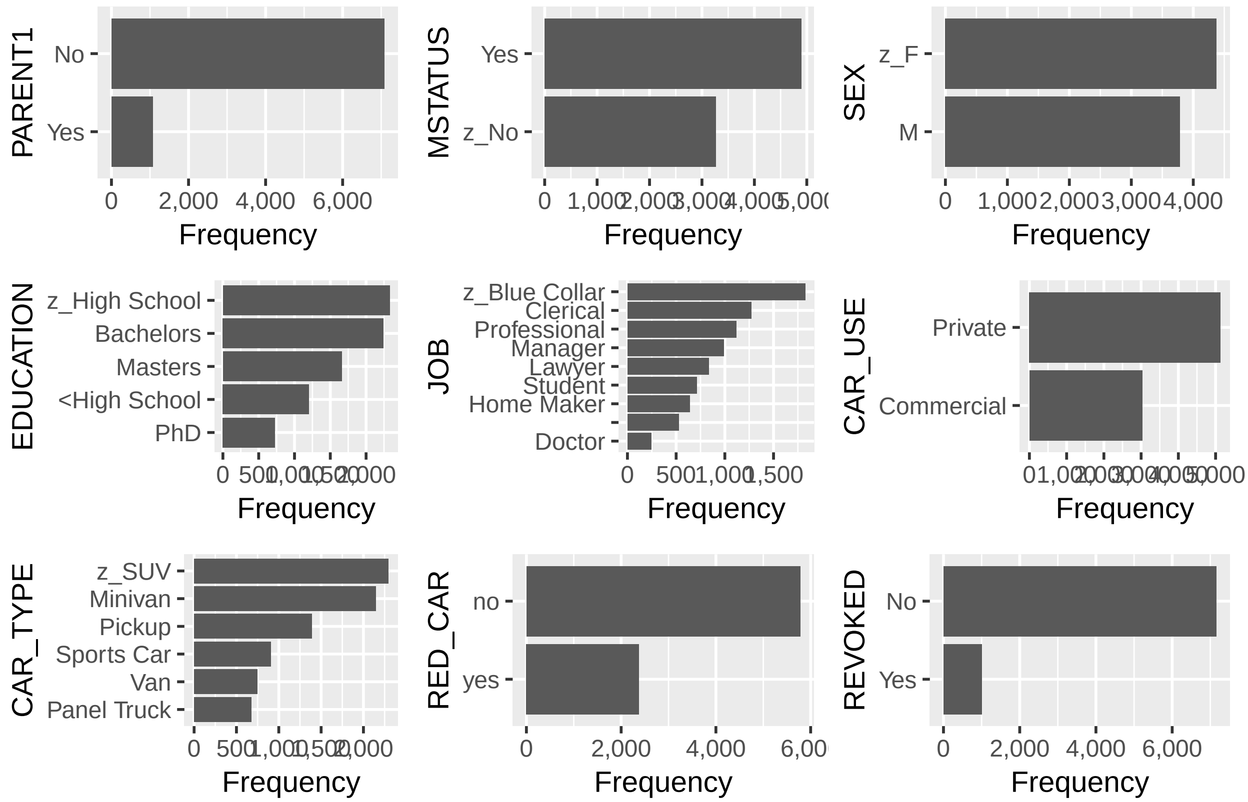
Histogram

```
ggtrain = split_columns(train)

plot_histogram(ggtrain$continuous)
```

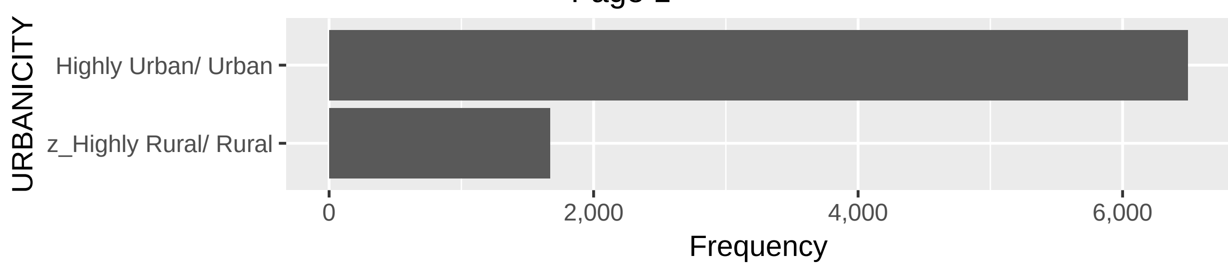



```
plot_bar(ggtrain$discrete)
```

```
## 4 columns ignored with more than 50 categories.
## INCOME: 6613 categories
## HOME_VAL: 5107 categories
## BLUEBOOK: 2789 categories
## OLDCLAIM: 2857 categories
```


PARENT1
No
Yes
0
2,000
4,000
6,000
Frequency
MSTATUS
Yes
z_No
Frequency
SEX
z_F
M
Frequency
EDUCATION
z_High School
Bachelors
Masters
<High School
PhD
Frequency
JOB
z_Blue Collar
Clerical
Professional
Manager
Lawyer
Student
Home Maker
Doctor
Frequency
CAR_USE
Private
Commercial
Frequency
CAR_TYPE
z_SUV
Minivan
Pickup
Sports Car
Van
Panel Truck
Frequency
RED_CAR
no
yes
0
2,000
4,000
Frequency
REVOKED
No
Yes
0
2,000
4,000
6,000
Frequency

Page 1


URBANICITY
Highly Urban/ Urban
z_Highly Rural/ Rural
0
2,000
4,000
6,000
Frequency

Page 2

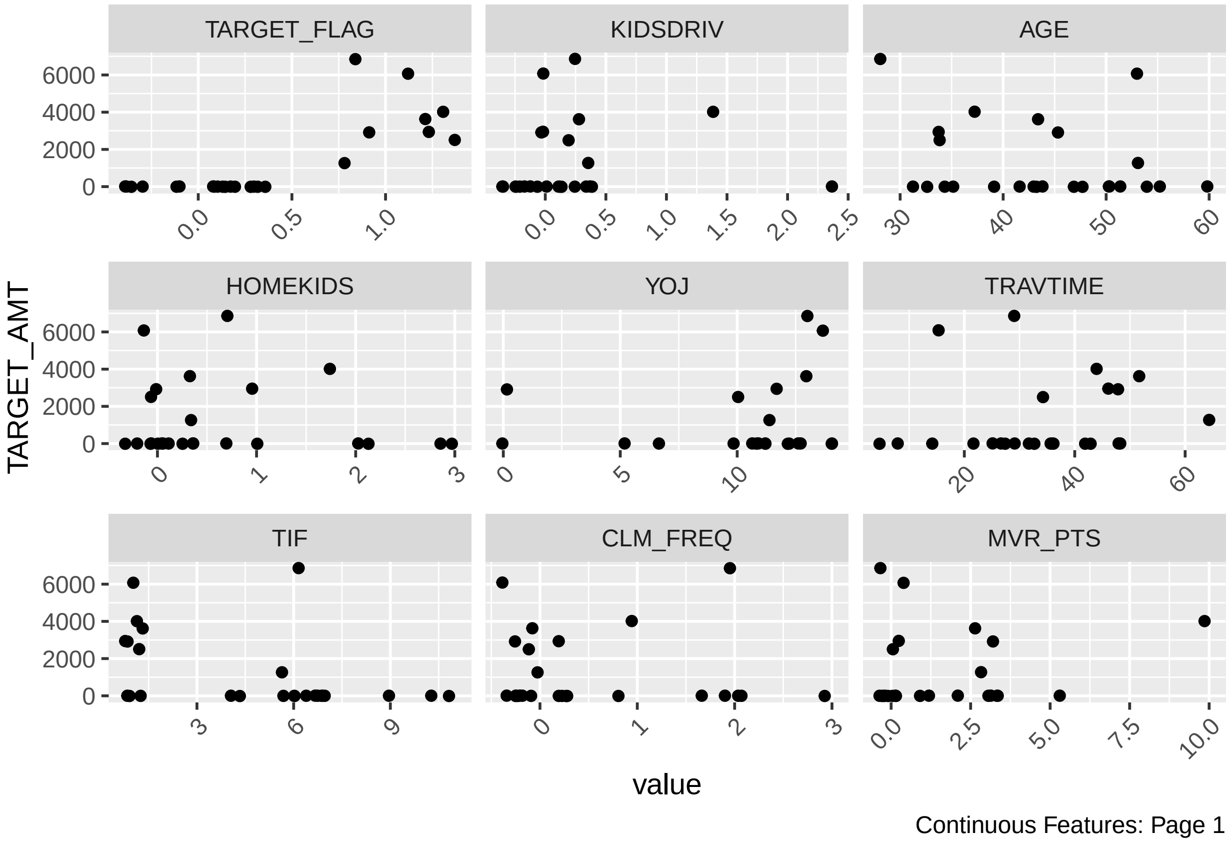
1. From the histogram we can see that variable TARGET_AMT is skewed. This is a ideal candidate for transformation.

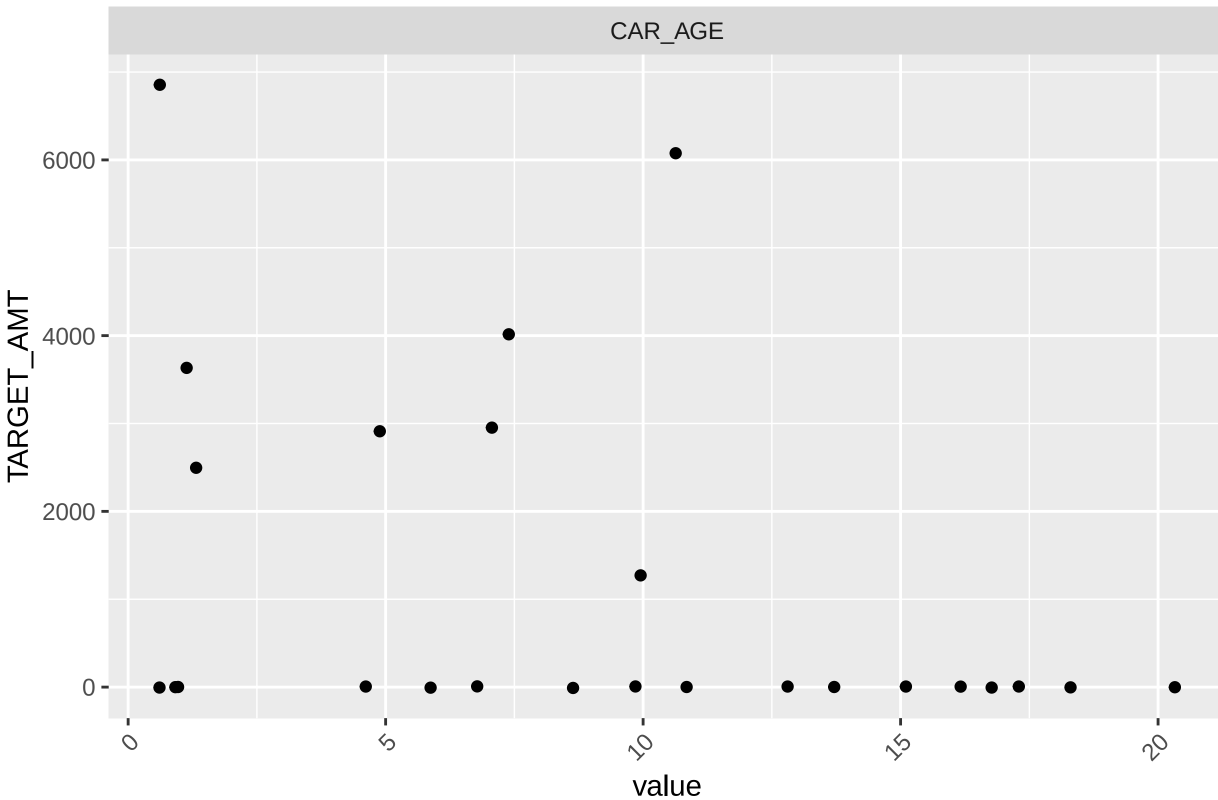
2. We can see TARGET_FLAG has more observations of not having accidents compared to having an accidents.

ScatterPlot

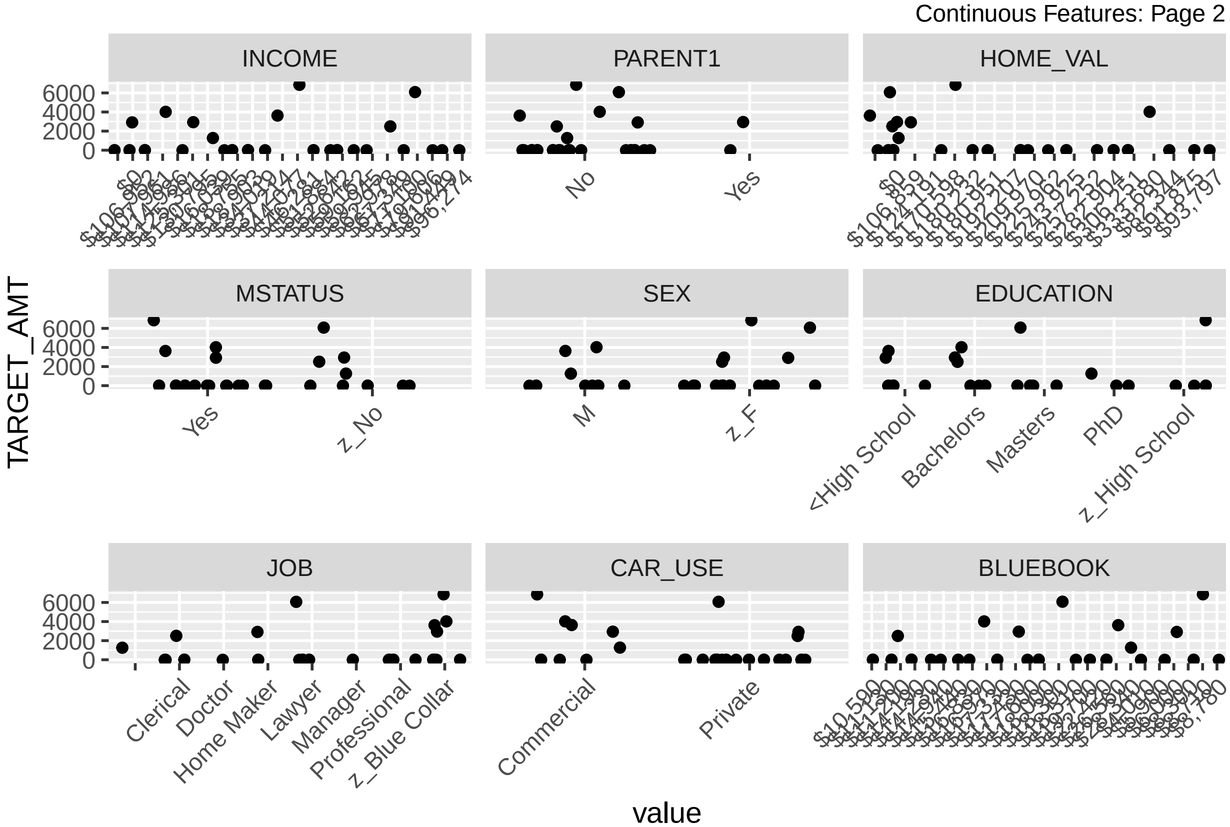
```
plot_scatterplot(train[1:25,], "TARGET_AMT", position = "jitter")
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

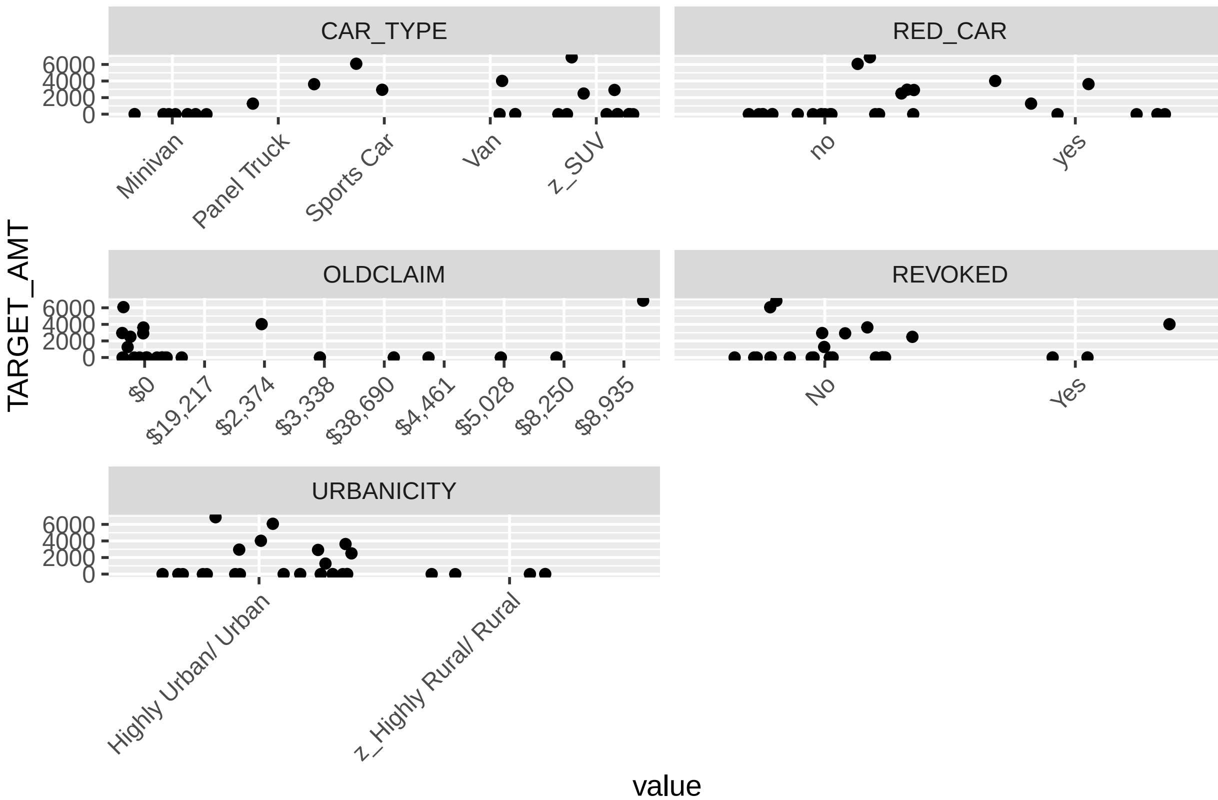




Continuous Features: Page 2



Discrete Features: Page 1

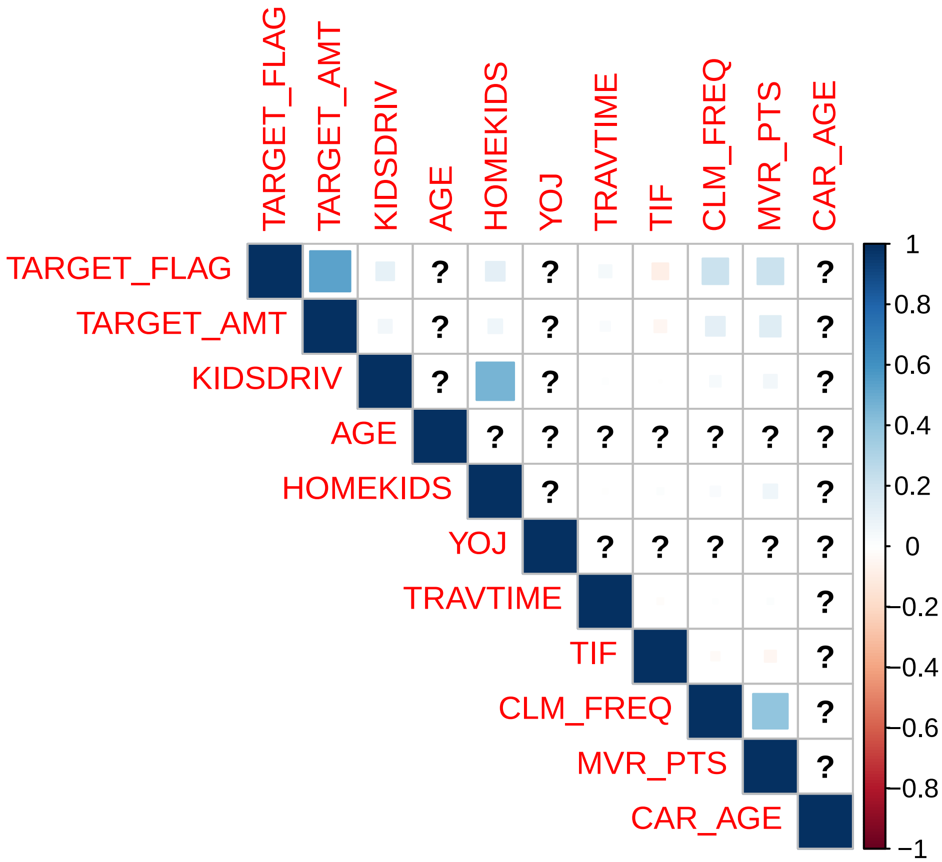


Discrete Features: Page 2

On the analysis of Scatter plot between TARGET_AMT and other Predictor variables, we do not see any pronounced positive or negative relationship.

MultiCollinearity between predictor variables and also with response variables

```
#TARGET_AMT,KIDSDRIV,AGE,YOJ,TRAVTIME,TIF,CLM_FREQ,MVR_PTS,CAR_AGE, INCOME,HOME_VAL,BLUEBOOK,OLDCLAIM



cordata = cor(ggtrain$continuous)
corrplot(cordata, method = "square", type = "upper")
```



From the corrplot we can see that KIDSDRIV and HOMEKIDS has a litte positive correlation.

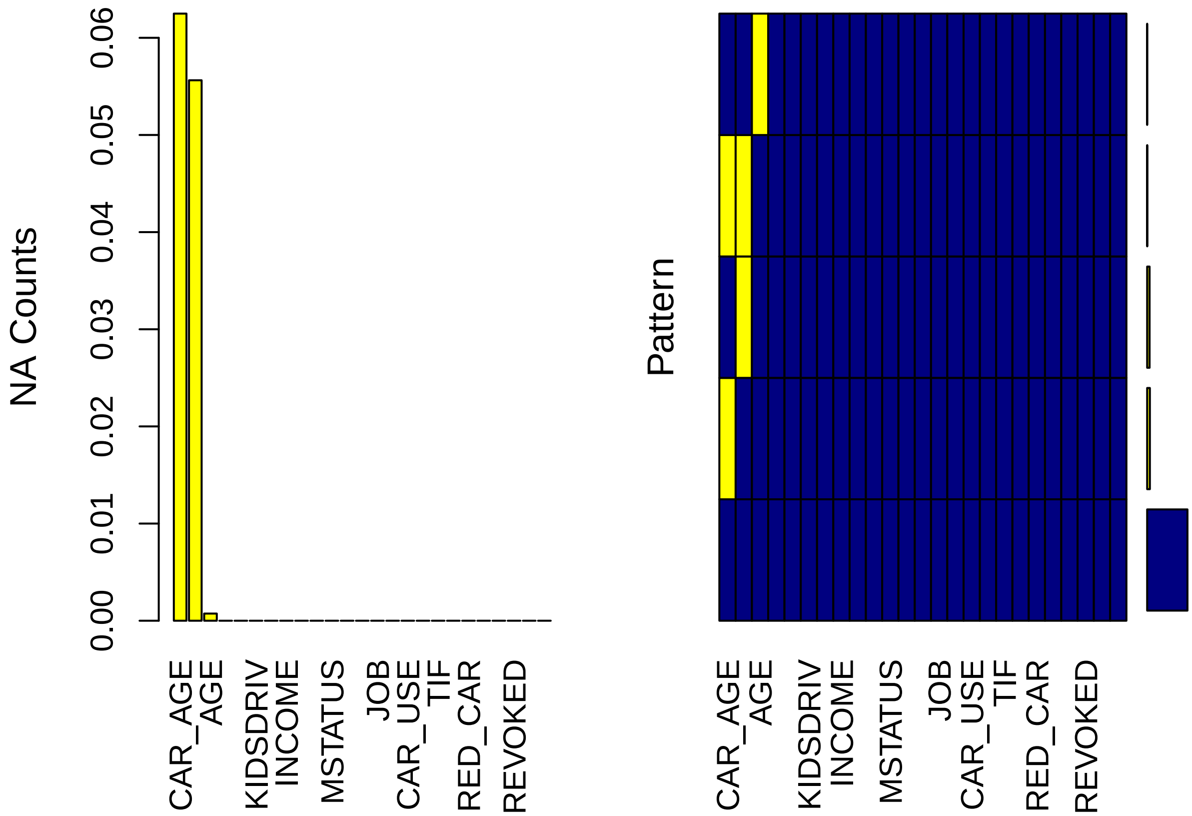
With respect to response and predictor variables, We do not see much of a bigger correlation.

Missing Value

On anlysis of missing values, we see CAR_AGE, YOJ and AGE has missing values in the respective order.

```
VIM::aggr(train, col=c('navyblue','yellow'),
                      numbers=TRUE, sortVars=TRUE,
                       labels=names(train),
                       ylab=c("NA Counts","Pattern"))
```

```
## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```




```
##
##  Variables sorted by number of missings:
##     Variable       Count
##      CAR_AGE 0.062492342
##          YOJ 0.055630437
##          AGE 0.000735204
##  TARGET_FLAG 0.000000000
##   TARGET_AMT 0.000000000
##     KIDSDRIV 0.000000000
##     HOMEKIDS 0.000000000
##       INCOME 0.000000000
##      PARENT1 0.000000000
##     HOME_VAL 0.000000000
##      MSTATUS 0.000000000
##          SEX 0.000000000
##    EDUCATION 0.000000000
##          JOB 0.000000000
##     TRAVTIME 0.000000000
##      CAR_USE 0.000000000
##     BLUEBOOK 0.000000000
##          TIF 0.000000000
##     CAR_TYPE 0.000000000
##      RED_CAR 0.000000000
##     OLDCLAIM 0.000000000
##     CLM_FREQ 0.000000000
##      REVOKED 0.000000000
##      MVR_PTS 0.000000000
##   URBANICITY 0.000000000
```

2.DATA PREPARATION

From our visual exploration we have identified few variables to go through transformation based on the dataset.

1. We will make Home Kids as a Boolean instead of Factor.
2. Will reset the -ve values of CAR_AGE to 0.
3. We will change the Jobs and Education Levels.
4. For the variables CAR_AGE, AGE, YOJ we fill those missing values with Median/Mean.

We tried executing imputing these random missing values using MICE package.

But running the package lead to crashing of R server multiple times. So for this project dropped from using mice.

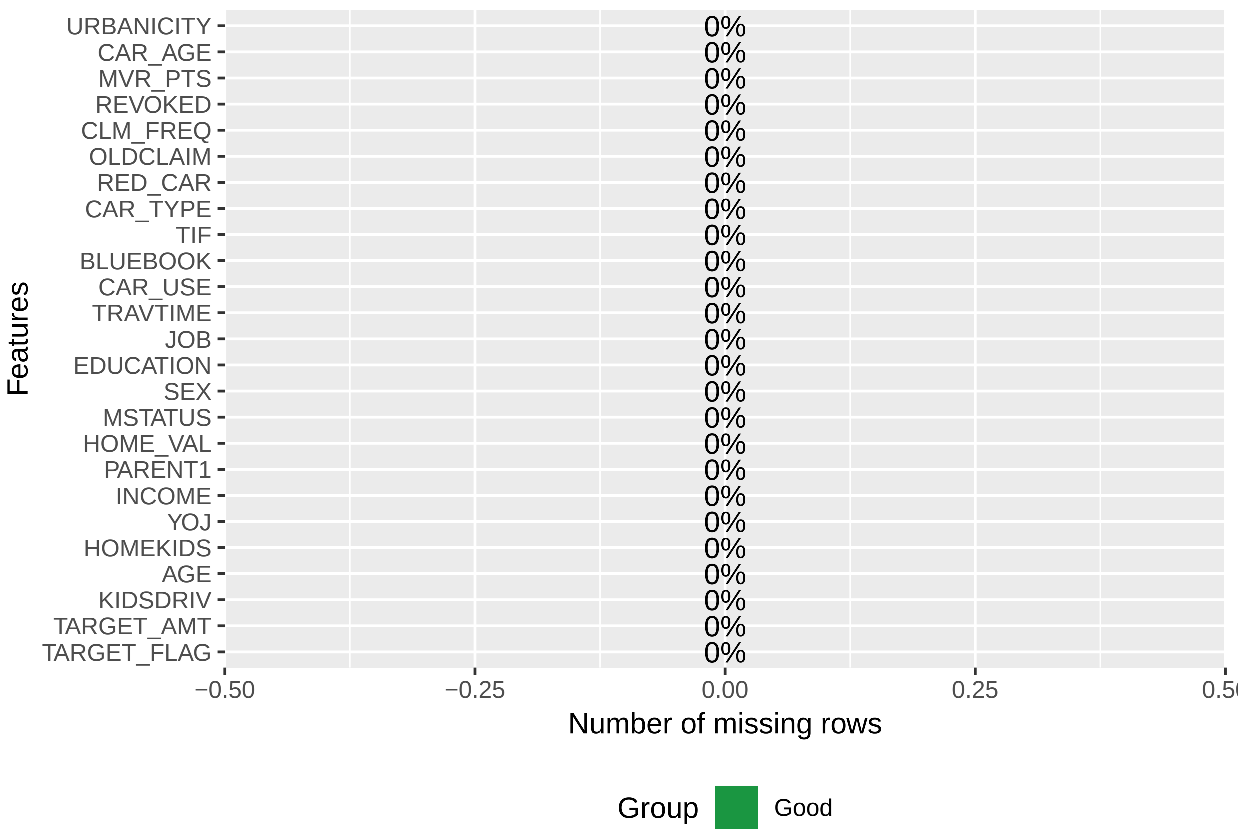
```r
train$HOMEKIDS[train$HOMEKIDS != 0 ] = 1

train$CAR_AGE[train$CAR_AGE < 0 ] = 0

train$JOB = as.character(train$JOB)
train$JOB[train$JOB == ""] = "Miscellaneous"
train$JOB <- as.factor(train$JOB)

train$EDUCATION <- ifelse(train$EDUCATION %in% c("PhD", "Masters"), 0, 1)


# ## Trying to use Mice package to fill in the missing values****
# mice_train =  mice(train, m = 1, maxit = 1, print = FALSE)
# train <- complete(mice_train)
#
#
# ###################


m = mean(train$AGE, na.rm = T)
train$AGE[is.na(train$AGE)] <- m

m = median(train$CAR_AGE, na.rm = T)
train$CAR_AGE[is.na(train$CAR_AGE)] = m

m = mean(train$YOJ, na.rm = T)
train$YOJ[is.na(train$YOJ)] = m

train$INCOME = as.numeric(train$INCOME)
train$HOME_VAL= as.numeric(train$HOME_VAL)
train$BLUEBOOK = as.numeric(train$BLUEBOOK)
train$OLDCLAIM= as.numeric(train$OLDCLAIM)

############## Transformation of Test Data####################
test$HOMEKIDS[test$HOMEKIDS != 0 ] = 1
test$CAR_AGE[test$CAR_AGE < 0 ] = 0

test$JOB = as.character(test$JOB)
test$JOB[test$JOB == ""] = "Miscellaneous"
```

```r
test$JOB <- as.factor(test$JOB)

test$EDUCATION <- ifelse(test$EDUCATION %in% c("PhD", "Masters"), 0, 1)

m = mean(test$AGE, na.rm = T)
test$AGE[is.na(test$AGE)] = m

m = median(test$CAR_AGE, na.rm = T)
test$CAR_AGE[is.na(test$CAR_AGE)] = m

m = mean(test$YOJ, na.rm = T)
test$YOJ[is.na(test$YOJ)] = m

test$INCOME = as.numeric(test$INCOME)
test$HOME_VAL= as.numeric(test$HOME_VAL)
test$BLUEBOOK = as.numeric(test$BLUEBOOK)
test$OLDCLAIM= as.numeric(test$OLDCLAIM)
```

We will plot and see the missing elements. This is after filling the missed values.

We are making sure that there are no missing values.

```r
plot_missing(train)
```



3. BUILDING MODELS

Classification:

As approach we are going to build the following models.

Each of our logistic regression models will use bionomial regression with a logit link function

Base Model and Transformed Variables

1. The first model will be Base Model. It will contain all transformed data. It contains all the 24 variables(excluding TARGET_AMT)

```
base_transform_model = glm(TARGET_FLAG ~ . -TARGET_AMT , family = binomial(link = 'logit'), data = trai

summary(base_transform_model)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5221  -0.7269  -0.4069   0.6513   3.1181
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -9.866e-01  3.062e-01  -3.222 0.001273 **
## KIDSDRIV                        3.485e-01  5.945e-02   5.862 4.59e-09 ***
## AGE                            -8.537e-04  4.058e-03  -0.210 0.833383
## HOMEKIDS                        3.039e-01  9.644e-02   3.151 0.001626 **
## YOJ                            -1.917e-02  8.620e-03  -2.223 0.026190 *
## INCOME                         -1.639e-05  1.581e-05  -1.037 0.299720
## PARENT1Yes                      2.237e-01  1.198e-01   1.867 0.061945 .
## HOME_VAL                       -9.019e-05  2.011e-05  -4.485 7.30e-06 ***
## MSTATUSz_No                     5.555e-01  8.036e-02   6.913 4.74e-12 ***
## SEXz_F                         -2.961e-01  1.027e-01  -2.882 0.003952 **
## EDUCATION                       1.582e-02  1.330e-01   0.119 0.905315
## JOBDoctor                      -1.143e+00  2.570e-01  -4.446 8.76e-06 ***
## JOBHome Maker                  -5.448e-02  1.407e-01  -0.387 0.698632
## JOBLawyer                      -5.424e-01  1.791e-01  -3.029 0.002454 **
## JOBManager                     -1.246e+00  1.368e-01  -9.107  < 2e-16 ***
## JOBMiscellaneous               -7.007e-01  1.903e-01  -3.683 0.000231 ***
## JOBProfessional                -5.218e-01  1.159e-01  -4.501 6.75e-06 ***
## JOBStudent                     -5.956e-02  1.280e-01  -0.465 0.641702
## JOBz_Blue Collar               -1.753e-01  1.060e-01  -1.654 0.098196 .
## TRAVTIME                        1.437e-02  1.872e-03   7.675 1.65e-14 ***
## CAR_USEPrivate                 -7.015e-01  8.655e-02  -8.105 5.29e-16 ***
## BLUEBOOK                        1.791e-05  3.358e-05   0.534 0.593654
## TIF                            -5.468e-02  7.297e-03  -7.493 6.71e-14 ***
## CAR_TYPEPanel Truck             2.149e-01  1.417e-01   1.517 0.129311
## CAR_TYPEPickup                  6.258e-01  1.011e-01   6.192 5.95e-10 ***
## CAR_TYPESports Car              1.224e+00  1.221e-01  10.023  < 2e-16 ***
## CAR_TYPEVan                     4.497e-01  1.205e-01   3.733 0.000190 ***
## CAR_TYPEz_SUV                   9.653e-01  1.027e-01   9.399  < 2e-16 ***
```

```
## RED_CARyes                        5.518e-03  8.596e-02   0.064 0.948817
## OLDCLAIM                          8.896e-05  4.232e-05   2.102 0.035548 *
## CLM_FREQ                          1.155e-01  3.201e-02   3.607 0.000310 ***
## REVOKEDYes                        7.451e-01  8.015e-02   9.296  < 2e-16 ***
## MVR_PTS                           1.060e-01  1.364e-02   7.774 7.61e-15 ***
## CAR_AGE                          -1.587e-02  6.867e-03  -2.312 0.020797 *
## URBANICITYz_Highly Rural/ Rural -2.338e+00  1.120e-01 -20.864  < 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: 7370.3  on 8126  degrees of freedom
## AIC: 7440.3
##
## Number of Fisher Scoring iterations: 5
```

```
knitr::kable(vif(base_transform_model))
```

	GVIF	Df	GVIF^(1/(2*Df))
KIDSDRIV	1.298221	1	1.139395
AGE	1.516209	1	1.231344
HOMEKIDS	2.671089	1	1.634347
YOJ	1.474231	1	1.214179
INCOME	1.266935	1	1.125582
PARENT1	2.350983	1	1.533292
HOME_VAL	1.309729	1	1.144434
MSTATUS	1.922674	1	1.386605
SEX	3.147536	1	1.774130
EDUCATION	4.059590	1	2.014842
JOB	12.381173	8	1.170301
TRAVTIME	1.036810	1	1.018239
CAR_USE	2.206358	1	1.485382
BLUEBOOK	1.121108	1	1.058824
TIF	1.010228	1	1.005101
CAR_TYPE	3.923111	5	1.146471
RED_CAR	1.836787	1	1.355281
OLDCLAIM	1.840900	1	1.356798
CLM_FREQ	1.850374	1	1.360285
REVOKED	1.016112	1	1.008024
MVR_PTS	1.183336	1	1.087813
CAR_AGE	1.672388	1	1.293209
URBANICITY	1.137012	1	1.066307

```
hoslem.test(train$TARGET_FLAG, fitted(base_transform_model))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  train$TARGET_FLAG, fitted(base_transform_model)
## X-squared = 5.5714, df = 8, p-value = 0.6951
```

```r
rocBaseTransformPlot = roc(base_transform_model$y, fitted(base_transform_model))
AUC = as.numeric(pROC::roc(base_transform_model$y, fitted(base_transform_model))$auc)
AUC
```

```
## [1] 0.8088218
```

Observations

From the above model, we see 8 out of the 25 variables has (stat-sig) p-values at a significance level greater than 0.05 These variable can be dropped in our next model to see how our model performs. The below are the variables which can be dropped in our next model. AGE
INCOME
EDUCATION
JOBHome Maker
JOBStudent
JOBz_Blue Collar
CAR_TYPEPanel Truck RED_CARyes

From the VIF function we can see 2 variables has VIF > 4. So this multicollinearity issue needs to be fixed. These can be removed from our model in future models.

From the above Hoslem test we can see the value of p = 0.6951, which is significantly greater than 0.05. Which says our model is not that good.

From the AUC(Area under the curve) values above of 'r AUC' is relatively high at .8088. As far as AUC, this model is good.

Considering the hoslem test result this model is creating, this will not be the ideal candidate.

Base Model Transformation pls Backward Elimination

2. For this model, we are going to use the Base Model(Transformation) and we are going to remove the variables which has higher p-value(>0.05). Also we are going to remove variables which has VIF > 4 from our Model 1.

The following variables have been removed from base_model variables.

```r
train_new = subset(train, select = -c(AGE, INCOME, EDUCATION, JOB, CAR_TYPE))

base_backward_model = glm(TARGET_FLAG ~ . -TARGET_AMT , family = binomial(link = 'logit'), data = train


summary(base_backward_model)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT, family = binomial(link = "logit"),
##     data = train_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3355  -0.7434  -0.4461   0.7224   2.9736
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -7.205e-01  1.549e-01  -4.653 3.27e-06 ***
## KIDSDRIV                        3.019e-01  5.700e-02   5.297 1.18e-07 ***
## HOMEKIDS                        4.008e-01  8.466e-02   4.735 2.19e-06 ***
## YOJ                            -4.022e-02  7.219e-03  -5.572 2.51e-08 ***
```

```
## PARENT1Yes                         1.993e-01  1.171e-01   1.702 0.088821 .
## HOME_VAL                          -1.379e-04  1.899e-05  -7.262 3.82e-13 ***
## MSTATUSz_No                        4.203e-01  7.766e-02   5.412 6.24e-08 ***
## SEXz_F                             2.683e-01  7.927e-02   3.385 0.000712 ***
## TRAVTIME                           1.476e-02  1.830e-03   8.066 7.28e-16 ***
## CAR_USEPrivate                    -7.550e-01  6.049e-02 -12.480  < 2e-16 ***
## BLUEBOOK                           7.733e-05  3.173e-05   2.437 0.014795 *
## TIF                               -5.141e-02  7.144e-03  -7.197 6.15e-13 ***
## RED_CARyes                        -1.136e-02  8.397e-02  -0.135 0.892424
## OLDCLAIM                           9.868e-05  4.152e-05   2.377 0.017452 *
## CLM_FREQ                           1.239e-01  3.130e-02   3.958 7.55e-05 ***
## REVOKEDYes                         7.701e-01  7.796e-02   9.878  < 2e-16 ***
## MVR_PTS                            1.127e-01  1.337e-02   8.426  < 2e-16 ***
## CAR_AGE                           -4.352e-02  5.403e-03  -8.056 7.88e-16 ***
## URBANICITYz_Highly Rural/ Rural -2.103e+00  1.102e-01 -19.087  < 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: 7646.7  on 8142  degrees of freedom
## AIC: 7684.7
##
## Number of Fisher Scoring iterations: 5
```

```
knitr::kable(vif(base_backward_model))
```

	x
KIDSDRIV	1.242013
HOMEKIDS	2.144087
YOJ	1.051585
PARENT1	2.359908
HOME_VAL	1.212542
MSTATUS	1.874480
SEX	1.952126
TRAVTIME	1.032653
CAR_USE	1.122006
BLUEBOOK	1.012350
TIF	1.005389
RED_CAR	1.818887
OLDCLAIM	1.844528
CLM_FREQ	1.854404
REVOKED	1.013313
MVR_PTS	1.179388
CAR_AGE	1.071260
URBANICITY	1.103281

```
hoslem.test(train$TARGET_FLAG, fitted(base_backward_model))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
```

```
## data:  train$TARGET_FLAG, fitted(base_backward_model)
## X-squared = 9.9022, df = 8, p-value = 0.272
```

```
rocBaseBackwardPlot = roc(base_backward_model$y, fitted(base_backward_model))
AUC = as.numeric(pROC::roc(base_backward_model$y, fitted(base_backward_model))$auc)
AUC
```

```
## [1] 0.7875901
```

Observations

From the above model, we see all of our variables has p-values at a significance level lesser than 0.05

From the VIF function we can see 0 variables has VIF > 4.

From the above Hoslem test we can see the value of p = 0.272, which is little greater than 0.05. Though this value is better compared to our MOdel 1, this is not the best Model for us to pick.

From the AUC(Area under the curve) values above of 'r AUC' has come down a littel compared to Model 1 at .787.

Step Model

3. For our final model, we will use the step function on the base model and transformation variables.

```
base_step_model = step(base_transform_model)
```

```
## Start:  AIC=7440.31
## TARGET_FLAG ~ (TARGET_AMT + KIDSDRIV + AGE + HOMEKIDS + YOJ +
##     INCOME + PARENT1 + HOME_VAL + MSTATUS + SEX + EDUCATION +
##     JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + RED_CAR +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY) -
##     TARGET_AMT
##
##              Df Deviance    AIC
## - RED_CAR     1   7370.3 7438.3
## - EDUCATION   1   7370.3 7438.3
## - AGE         1   7370.4 7438.4
## - BLUEBOOK    1   7370.6 7438.6
## - INCOME      1   7371.4 7439.4
## <none>            7370.3 7440.3
## - PARENT1     1   7373.8 7441.8
## - OLDCLAIM    1   7374.7 7442.7
## - YOJ         1   7375.2 7443.2
## - CAR_AGE     1   7375.6 7443.6
## - SEX         1   7378.6 7446.6
## - HOMEKIDS    1   7380.2 7448.2
## - CLM_FREQ    1   7383.2 7451.2
## - HOME_VAL    1   7390.6 7458.6
## - KIDSDRIV    1   7404.6 7472.6
## - MSTATUS     1   7418.2 7486.2
## - TIF         1   7428.5 7496.5
## - TRAVTIME    1   7429.4 7497.4
## - MVR_PTS     1   7431.2 7499.2
## - CAR_USE     1   7437.1 7505.1
## - REVOKED     1   7455.3 7523.3
## - JOB         8   7487.1 7541.1
## - CAR_TYPE    5   7501.0 7561.0
## - URBANICITY  1   7994.6 8062.6
```

```
##
## Step:  AIC=7438.31
## TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + CAR_AGE + URBANICITY
##
##              Df Deviance    AIC
## - EDUCATION   1   7370.3 7436.3
## - AGE         1   7370.4 7436.4
## - BLUEBOOK    1   7370.6 7436.6
## - INCOME      1   7371.4 7437.4
## <none>            7370.3 7438.3
## - PARENT1     1   7373.8 7439.8
## - OLDCLAIM    1   7374.7 7440.7
## - YOJ         1   7375.2 7441.2
## - CAR_AGE     1   7375.6 7441.6
## - HOMEKIDS    1   7380.2 7446.2
## - SEX         1   7381.8 7447.8
## - CLM_FREQ    1   7383.2 7449.2
## - HOME_VAL    1   7390.6 7456.6
## - KIDSDRIV    1   7404.6 7470.6
## - MSTATUS     1   7418.2 7484.2
## - TIF         1   7428.5 7494.5
## - TRAVTIME    1   7429.4 7495.4
## - MVR_PTS     1   7431.2 7497.2
## - CAR_USE     1   7437.1 7503.1
## - REVOKED     1   7455.3 7521.3
## - JOB         8   7487.1 7539.1
## - CAR_TYPE    5   7501.0 7559.0
## - URBANICITY  1   7994.7 8060.7
##
## Step:  AIC=7436.32
## TARGET_FLAG ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY
##
##              Df Deviance    AIC
## - AGE         1   7370.4 7434.4
## - BLUEBOOK    1   7370.6 7434.6
## - INCOME      1   7371.4 7435.4
## <none>            7370.3 7436.3
## - PARENT1     1   7373.8 7437.8
## - OLDCLAIM    1   7374.7 7438.7
## - YOJ         1   7375.3 7439.3
## - CAR_AGE     1   7376.5 7440.5
## - HOMEKIDS    1   7380.2 7444.2
## - SEX         1   7381.8 7445.8
## - CLM_FREQ    1   7383.2 7447.2
## - HOME_VAL    1   7390.6 7454.6
## - KIDSDRIV    1   7404.6 7468.6
## - MSTATUS     1   7418.2 7482.2
## - TIF         1   7428.6 7492.6
```

```
## - TRAVTIME    1   7429.4 7493.4
## - MVR_PTS     1   7431.2 7495.2
## - CAR_USE     1   7437.1 7501.1
## - REVOKED     1   7455.3 7519.3
## - JOB         8   7501.8 7551.8
## - CAR_TYPE    5   7501.0 7557.0
## - URBANICITY  1   7994.7 8058.7
##
## Step:  AIC=7434.37
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY
##
##              Df Deviance    AIC
## - BLUEBOOK    1   7370.7 7432.7
## - INCOME      1   7371.4 7433.4
## <none>            7370.4 7434.4
## - PARENT1     1   7373.8 7435.8
## - OLDCLAIM    1   7374.8 7436.8
## - YOJ         1   7375.6 7437.6
## - CAR_AGE     1   7376.6 7438.6
## - SEX         1   7381.8 7443.8
## - HOMEKIDS    1   7383.2 7445.2
## - CLM_FREQ    1   7383.3 7445.3
## - HOME_VAL    1   7390.7 7452.7
## - KIDSDRIV    1   7405.5 7467.5
## - MSTATUS     1   7419.0 7481.0
## - TIF         1   7428.6 7490.6
## - TRAVTIME    1   7429.4 7491.4
## - MVR_PTS     1   7431.4 7493.4
## - CAR_USE     1   7437.1 7499.1
## - REVOKED     1   7455.4 7517.4
## - JOB         8   7503.5 7551.5
## - CAR_TYPE    5   7501.2 7555.2
## - URBANICITY  1   7995.3 8057.3
##
## Step:  AIC=7432.66
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + JOB + TRAVTIME + CAR_USE + TIF +
##     CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE +
##     URBANICITY
##
##              Df Deviance    AIC
## - INCOME      1   7371.8 7431.8
## <none>            7370.7 7432.7
## - PARENT1     1   7374.1 7434.1
## - OLDCLAIM    1   7375.1 7435.1
## - YOJ         1   7375.9 7435.9
## - CAR_AGE     1   7377.0 7437.0
## - SEX         1   7382.1 7442.1
## - CLM_FREQ    1   7383.5 7443.5
## - HOMEKIDS    1   7383.6 7443.6
## - HOME_VAL    1   7391.0 7451.0
```

```
## - KIDSDRIV    1   7405.7 7465.7
## - MSTATUS     1   7419.5 7479.5
## - TIF         1   7428.8 7488.8
## - TRAVTIME    1   7429.7 7489.7
## - MVR_PTS     1   7431.6 7491.6
## - CAR_USE     1   7437.2 7497.2
## - REVOKED     1   7455.6 7515.6
## - JOB         8   7504.4 7550.4
## - CAR_TYPE    5   7505.4 7557.4
## - URBANICITY  1   7996.2 8056.2
##
## Step:  AIC=7431.79
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + JOB + TRAVTIME + CAR_USE + TIF + CAR_TYPE +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY
##
##              Df Deviance    AIC
## <none>            7371.8 7431.8
## - PARENT1     1   7375.2 7433.2
## - OLDCLAIM    1   7376.3 7434.3
## - CAR_AGE     1   7378.1 7436.1
## - YOJ         1   7378.3 7436.3
## - SEX         1   7383.2 7441.2
## - CLM_FREQ    1   7384.5 7442.5
## - HOMEKIDS    1   7384.9 7442.9
## - HOME_VAL    1   7391.9 7449.9
## - KIDSDRIV    1   7407.2 7465.2
## - MSTATUS     1   7420.5 7478.5
## - TIF         1   7430.2 7488.2
## - TRAVTIME    1   7430.8 7488.8
## - MVR_PTS     1   7432.4 7490.4
## - CAR_USE     1   7438.2 7496.2
## - REVOKED     1   7457.0 7515.0
## - JOB         8   7509.3 7553.3
## - CAR_TYPE    5   7506.8 7556.8
## - URBANICITY  1   7997.1 8055.1
```

```
summary(base_step_model)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + JOB + TRAVTIME + CAR_USE + TIF +
##     CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE +
##     URBANICITY, family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5258  -0.7270  -0.4063   0.6525   3.1130
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.007e+00  1.882e-01  -5.353 8.64e-08 ***
## KIDSDRIV                        3.472e-01  5.831e-02   5.954 2.62e-09 ***
## HOMEKIDS                        3.155e-01  8.694e-02   3.630 0.000284 ***
```

```
## YOJ                            -2.120e-02  8.331e-03  -2.544 0.010947 *
## PARENT1Yes                      2.225e-01  1.197e-01   1.858 0.063112 .
## HOME_VAL                       -8.995e-05  2.014e-05  -4.466 7.95e-06 ***
## MSTATUSz_No                     5.576e-01  7.995e-02   6.974 3.08e-12 ***
## SEXz_F                         -2.986e-01  8.887e-02  -3.360 0.000780 ***
## JOBDoctor                      -1.155e+00  2.302e-01  -5.019 5.20e-07 ***
## JOBHome Maker                  -4.573e-02  1.378e-01  -0.332 0.740007
## JOBLawyer                      -5.744e-01  1.374e-01  -4.180 2.92e-05 ***
## JOBManager                     -1.268e+00  1.288e-01  -9.847  < 2e-16 ***
## JOBMiscellaneous               -7.210e-01  1.537e-01  -4.691 2.72e-06 ***
## JOBProfessional                -5.476e-01  1.131e-01  -4.841 1.29e-06 ***
## JOBStudent                     -4.577e-02  1.272e-01  -0.360 0.719035
## JOBz_Blue Collar               -1.940e-01  1.047e-01  -1.852 0.063969 .
## TRAVTIME                        1.436e-02  1.871e-03   7.672 1.69e-14 ***
## CAR_USEPrivate                 -6.991e-01  8.649e-02  -8.082 6.35e-16 ***
## TIF                            -5.477e-02  7.293e-03  -7.510 5.92e-14 ***
## CAR_TYPEPanel Truck             2.322e-01  1.390e-01   1.671 0.094763 .
## CAR_TYPEPickup                  6.398e-01  9.865e-02   6.486 8.82e-11 ***
## CAR_TYPESports Car              1.231e+00  1.209e-01  10.186  < 2e-16 ***
## CAR_TYPEVan                     4.546e-01  1.204e-01   3.776 0.000159 ***
## CAR_TYPEz_SUV                   9.692e-01  1.022e-01   9.488  < 2e-16 ***
## OLDCLAIM                        8.957e-05  4.230e-05   2.118 0.034197 *
## CLM_FREQ                        1.145e-01  3.198e-02   3.581 0.000343 ***
## REVOKEDYes                      7.458e-01  8.013e-02   9.307  < 2e-16 ***
## MVR_PTS                         1.056e-01  1.362e-02   7.755 8.85e-15 ***
## CAR_AGE                        -1.626e-02  6.479e-03  -2.509 0.012103 *
## URBANICITYz_Highly Rural/ Rural -2.338e+00  1.120e-01 -20.874  < 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: 7371.8  on 8131  degrees of freedom
## AIC: 7431.8
##
## Number of Fisher Scoring iterations: 5
```

```
knitr::kable(vif(base_step_model))
```

	GVIF	Df	GVIF^(1/(2*Df))
KIDSDRIV	1.249841	1	1.117963
HOMEKIDS	2.171342	1	1.473547
YOJ	1.377388	1	1.173622
PARENT1	2.347636	1	1.532200
HOME_VAL	1.309494	1	1.144331
MSTATUS	1.903822	1	1.379790
SEX	2.355548	1	1.534780
JOB	4.706122	8	1.101644
TRAVTIME	1.036521	1	1.018097
CAR_USE	2.203776	1	1.484512
TIF	1.009478	1	1.004728
CAR_TYPE	3.597385	5	1.136577
OLDCLAIM	1.839934	1	1.356442

	GVIF	Df	GVIF^(1/(2*Df))
CLM_FREQ	1.847465	1	1.359215
REVOKED	1.015661	1	1.007800
MVR_PTS	1.181072	1	1.086771
CAR_AGE	1.488861	1	1.220189
URBANICITY	1.136417	1	1.066029

```
hoslem.test(train$TARGET_FLAG, fitted(base_step_model))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  train$TARGET_FLAG, fitted(base_step_model)
## X-squared = 4.1094, df = 8, p-value = 0.8471
```

```
rocBaseStepPlot = roc(base_step_model$y, fitted(base_step_model))
AUC = as.numeric(pROC::roc(base_step_model$y, fitted(base_step_model))$auc)
AUC
```

```
## [1] 0.8087917
```

Observations

From the above model, we see 8 variables dropped from 25.

From the VIF function we can see that the following variables has VIF > 4. JOB is the only variable which has a little higher vIF.

From the above Hoslem test we can see the value of p = 0.8471, which is more than 0.05. Which says our model is not the best.

From the AUC(Area under the curve) values above of 'r AUC' is relatively high at .808. This model is good at predicting the response variable.

Regression Analysis.

Base plus Transformed Variables.

For Linear Regression, we will first do as Base Model with transformed variables.

```
regbaseplustransform = lm(TARGET_AMT ~ .-TARGET_FLAG, data = train)
summary(regbaseplustransform)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ . - TARGET_FLAG, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5609  -1686   -766    344 103811
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.646e+03  5.299e+02   3.107 0.001895 **
## KIDSDRIV                        2.838e+02  1.113e+02   2.550 0.010801 *
## AGE                             9.591e+00  7.174e+00   1.337 0.181277
## HOMEKIDS                        3.319e+02  1.654e+02   2.007 0.044774 *
```

```
## YOJ                               -9.457e+00  1.521e+01  -0.622 0.534162
## INCOME                            -1.111e-02  2.677e-02  -0.415 0.678067
## PARENT1Yes                         4.684e+02  2.170e+02   2.158 0.030926 *
## HOME_VAL                          -7.358e-02  3.490e-02  -2.109 0.034999 *
## MSTATUSz_No                        5.710e+02  1.377e+02   4.147 3.41e-05 ***
## SEXz_F                            -2.987e+02  1.713e+02  -1.743 0.081357 .
## EDUCATION                         -2.512e+02  2.127e+02  -1.181 0.237634
## JOBDoctor                         -1.119e+03  3.858e+02  -2.900 0.003744 **
## JOBHome Maker                     -8.225e+01  2.464e+02  -0.334 0.738558
## JOBLawyer                         -4.800e+02  2.952e+02  -1.626 0.103983
## JOBManager                        -1.193e+03  2.256e+02  -5.286 1.28e-07 ***
## JOBMiscellaneous                  -7.284e+02  3.323e+02  -2.192 0.028434 *
## JOBProfessional                   -2.515e+02  2.022e+02  -1.244 0.213601
## JOBStudent                        -1.767e+02  2.333e+02  -0.757 0.448855
## JOBz_Blue Collar                  -7.443e+01  1.919e+02  -0.388 0.698070
## TRAVTIME                           1.192e+01  3.221e+00   3.701 0.000216 ***
## CAR_USEPrivate                    -7.385e+02  1.569e+02  -4.708 2.55e-06 ***
## BLUEBOOK                           2.008e-04  5.965e-02   0.003 0.997315
## TIF                               -4.774e+01  1.218e+01  -3.920 8.93e-05 ***
## CAR_TYPEPanel Truck                4.420e+02  2.520e+02   1.754 0.079432 .
## CAR_TYPEPickup                     3.763e+02  1.720e+02   2.188 0.028697 *
## CAR_TYPESports Car                 9.240e+02  2.058e+02   4.491 7.20e-06 ***
## CAR_TYPEVan                        5.785e+02  2.068e+02   2.797 0.005175 **
## CAR_TYPEz_SUV                      6.669e+02  1.655e+02   4.029 5.65e-05 ***
## RED_CARyes                        -5.013e+01  1.490e+02  -0.336 0.736580
## OLDCLAIM                          -2.493e-02  8.366e-02  -0.298 0.765702
## CLM_FREQ                           1.192e+02  6.303e+01   1.892 0.058565 .
## REVOKEDYes                         4.391e+02  1.556e+02   2.821 0.004796 **
## MVR_PTS                            1.753e+02  2.615e+01   6.704 2.17e-11 ***
## CAR_AGE                           -3.481e+01  1.188e+01  -2.930 0.003395 **
## URBANICITYz_Highly Rural/ Rural -1.660e+03  1.399e+02 -11.867  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4546 on 8126 degrees of freedom
## Multiple R-squared:  0.06986,    Adjusted R-squared:  0.06597
## F-statistic: 17.95 on 34 and 8126 DF,  p-value: < 2.2e-16
```

Observations *

1. Most of the variables has insignificant p=values. That is values greater than (0.05).
2. The Mutliple R-Squared and Adjusted R-Squared values are at 69% and 65% respectively. These values are not considered high for model selection.

BIC Step Model

Using the transformed values, we are going to do a BIC Forward and Backward selection with missing values imputed.

```
BICBasePlusTransform = step(regbaseplustransform)
```

```
## Start:  AIC=137499.6
## TARGET_AMT ~ (TARGET_FLAG + KIDSDRIV + AGE + HOMEKIDS + YOJ +
##     INCOME + PARENT1 + HOME_VAL + MSTATUS + SEX + EDUCATION +
##     JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE + RED_CAR +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY) -
```

```
##     TARGET_FLAG
##
##              Df  Sum of Sq        RSS    AIC
## - BLUEBOOK    1        234 1.6795e+11 137498
## - OLDCLAIM    1    1835544 1.6795e+11 137498
## - RED_CAR     1    2338867 1.6795e+11 137498
## - INCOME      1    3561586 1.6795e+11 137498
## - YOJ         1    7988239 1.6796e+11 137498
## - EDUCATION   1   28827474 1.6798e+11 137499
## - AGE         1   36942849 1.6799e+11 137499
## <none>                     1.6795e+11 137500
## - SEX         1   62796709 1.6801e+11 137501
## - CLM_FREQ    1   73962009 1.6802e+11 137501
## - HOMEKIDS    1   83259125 1.6803e+11 137502
## - HOME_VAL    1   91905585 1.6804e+11 137502
## - PARENT1     1   96285923 1.6805e+11 137502
## - KIDSDRIV    1  134356783 1.6808e+11 137504
## - REVOKED     1  164504877 1.6811e+11 137506
## - CAR_AGE     1  177481567 1.6813e+11 137506
## - TRAVTIME    1  283082422 1.6823e+11 137511
## - TIF         1  317580679 1.6827e+11 137513
## - MSTATUS     1  355392690 1.6830e+11 137515
## - CAR_TYPE    5  566183808 1.6852e+11 137517
## - CAR_USE     1  458060492 1.6841e+11 137520
## - JOB         8  825911570 1.6877e+11 137524
## - MVR_PTS     1  928773611 1.6888e+11 137543
## - URBANICITY  1 2910765489 1.7086e+11 137638
##
## Step:  AIC=137497.6
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - OLDCLAIM    1    1835435 1.6795e+11 137496
## - RED_CAR     1    2339758 1.6795e+11 137496
## - INCOME      1    3582469 1.6795e+11 137496
## - YOJ         1    7988321 1.6796e+11 137496
## - EDUCATION   1   28831412 1.6798e+11 137497
## - AGE         1   36976340 1.6799e+11 137497
## <none>                     1.6795e+11 137498
## - SEX         1   62817912 1.6801e+11 137499
## - CLM_FREQ    1   73969080 1.6802e+11 137499
## - HOMEKIDS    1   83259988 1.6803e+11 137500
## - HOME_VAL    1   91910716 1.6804e+11 137500
## - PARENT1     1   96286881 1.6805e+11 137500
## - KIDSDRIV    1  134356551 1.6808e+11 137502
## - REVOKED     1  164508257 1.6811e+11 137504
## - CAR_AGE     1  177679849 1.6813e+11 137504
## - TRAVTIME    1  283093012 1.6823e+11 137509
## - TIF         1  317581378 1.6827e+11 137511
## - MSTATUS     1  355671387 1.6830e+11 137513
## - CAR_TYPE    5  573792613 1.6852e+11 137515
```

```
## - CAR_USE     1  458072384 1.6841e+11 137518
## - JOB         8  826257525 1.6878e+11 137522
## - MVR_PTS     1  928882598 1.6888e+11 137541
## - URBANICITY  1 2913153695 1.7086e+11 137636
##
## Step:  AIC=137495.7
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     TIF + CAR_TYPE + RED_CAR + CLM_FREQ + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - RED_CAR     1    2335895 1.6795e+11 137494
## - INCOME      1    3544797 1.6795e+11 137494
## - YOJ         1    8096118 1.6796e+11 137494
## - EDUCATION   1   28867882 1.6798e+11 137495
## - AGE         1   36919716 1.6799e+11 137496
## <none>                     1.6795e+11 137496
## - SEX         1   62927085 1.6801e+11 137497
## - HOMEKIDS    1   83121358 1.6803e+11 137498
## - HOME_VAL    1   91326041 1.6804e+11 137498
## - PARENT1     1   96038794 1.6805e+11 137498
## - CLM_FREQ    1   99861679 1.6805e+11 137499
## - KIDSDRIV    1  134490431 1.6809e+11 137500
## - REVOKED     1  168949986 1.6812e+11 137502
## - CAR_AGE     1  177188946 1.6813e+11 137502
## - TRAVTIME    1  283513013 1.6823e+11 137507
## - TIF         1  317306476 1.6827e+11 137509
## - MSTATUS     1  355395909 1.6831e+11 137511
## - CAR_TYPE    5  572546911 1.6852e+11 137514
## - CAR_USE     1  458280356 1.6841e+11 137516
## - JOB         8  825961844 1.6878e+11 137520
## - MVR_PTS     1  938388177 1.6889e+11 137539
## - URBANICITY  1 2918808891 1.7087e+11 137634
##
## Step:  AIC=137493.8
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     TIF + CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE +
##     URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - INCOME      1    3484758 1.6796e+11 137492
## - YOJ         1    8176487 1.6796e+11 137492
## - EDUCATION   1   29045587 1.6798e+11 137493
## - AGE         1   37416303 1.6799e+11 137494
## <none>                     1.6795e+11 137494
## - SEX         1   69622628 1.6802e+11 137495
## - HOMEKIDS    1   83006540 1.6804e+11 137496
## - HOME_VAL    1   91219394 1.6804e+11 137496
## - PARENT1     1   96320600 1.6805e+11 137497
## - CLM_FREQ    1   99371631 1.6805e+11 137497
## - KIDSDRIV    1  135163951 1.6809e+11 137498
## - REVOKED     1  168858616 1.6812e+11 137500
```

```
## - CAR_AGE     1  177834066 1.6813e+11 137500
## - TRAVTIME    1  282937237 1.6824e+11 137506
## - TIF         1  317030400 1.6827e+11 137507
## - MSTATUS     1  354520507 1.6831e+11 137509
## - CAR_TYPE    5  573696832 1.6853e+11 137512
## - CAR_USE     1  458078467 1.6841e+11 137514
## - JOB         8  827852585 1.6878e+11 137518
## - MVR_PTS     1  938121111 1.6889e+11 137537
## - URBANICITY  1 2917448571 1.7087e+11 137632
##
## Step:  AIC=137492
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + TIF +
##     CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - YOJ         1   10593542 1.6797e+11 137491
## - EDUCATION   1   30757633 1.6799e+11 137492
## - AGE         1   37987609 1.6799e+11 137492
## <none>                     1.6796e+11 137492
## - SEX         1   69507901 1.6803e+11 137493
## - HOMEKIDS    1   83595729 1.6804e+11 137494
## - HOME_VAL    1   89935590 1.6805e+11 137494
## - PARENT1     1   96710630 1.6805e+11 137495
## - CLM_FREQ    1   98721276 1.6806e+11 137495
## - KIDSDRIV    1  136101820 1.6809e+11 137497
## - REVOKED     1  169325312 1.6813e+11 137498
## - CAR_AGE     1  179654190 1.6814e+11 137499
## - TRAVTIME    1  283032894 1.6824e+11 137504
## - TIF         1  317879367 1.6827e+11 137505
## - MSTATUS     1  354211793 1.6831e+11 137507
## - CAR_TYPE    5  576482166 1.6853e+11 137510
## - CAR_USE     1  457015853 1.6841e+11 137512
## - JOB         8  835230814 1.6879e+11 137516
## - MVR_PTS     1  936792019 1.6889e+11 137535
## - URBANICITY  1 2919406135 1.7088e+11 137631
##
## Step:  AIC=137490.5
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE + TIF +
##     CAR_TYPE + CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - EDUCATION   1   30982266 1.6800e+11 137490
## - AGE         1   32332917 1.6800e+11 137490
## <none>                     1.6797e+11 137491
## - SEX         1   70230257 1.6804e+11 137492
## - HOMEKIDS    1   76171307 1.6804e+11 137492
## - HOME_VAL    1   88077633 1.6806e+11 137493
## - PARENT1     1   98525929 1.6807e+11 137493
## - CLM_FREQ    1   99391919 1.6807e+11 137493
## - KIDSDRIV    1  138233850 1.6811e+11 137495
## - REVOKED     1  169443159 1.6814e+11 137497
## - CAR_AGE     1  180012536 1.6815e+11 137497
```

```
## - TRAVTIME    1  282074279 1.6825e+11 137502
## - TIF         1  318673159 1.6829e+11 137504
## - MSTATUS     1  370401544 1.6834e+11 137507
## - CAR_TYPE    5  581295512 1.6855e+11 137509
## - CAR_USE     1  459650051 1.6843e+11 137511
## - JOB         8  853406773 1.6882e+11 137516
## - MVR_PTS     1  943192821 1.6891e+11 137534
## - URBANICITY  1 2914045616 1.7088e+11 137629
##
## Step:  AIC=137490
## TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + JOB + TRAVTIME + CAR_USE + TIF + CAR_TYPE +
##     CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## - AGE         1   34809784 1.6803e+11 137490
## <none>                     1.6800e+11 137490
## - SEX         1   67147490 1.6807e+11 137491
## - HOMEKIDS    1   75738343 1.6807e+11 137492
## - HOME_VAL    1   85833400 1.6808e+11 137492
## - PARENT1     1   96938364 1.6810e+11 137493
## - CLM_FREQ    1   98686280 1.6810e+11 137493
## - KIDSDRIV    1  139886195 1.6814e+11 137495
## - CAR_AGE     1  150212709 1.6815e+11 137495
## - REVOKED     1  170125681 1.6817e+11 137496
## - TRAVTIME    1  281495113 1.6828e+11 137502
## - TIF         1  314984353 1.6831e+11 137503
## - MSTATUS     1  372449611 1.6837e+11 137506
## - CAR_TYPE    5  581195583 1.6858e+11 137508
## - CAR_USE     1  455654456 1.6845e+11 137510
## - JOB         8  827240101 1.6883e+11 137514
## - MVR_PTS     1  939420290 1.6894e+11 137534
## - URBANICITY  1 2920498332 1.7092e+11 137629
##
## Step:  AIC=137489.7
## TARGET_AMT ~ KIDSDRIV + HOMEKIDS + PARENT1 + HOME_VAL + MSTATUS +
##     SEX + JOB + TRAVTIME + CAR_USE + TIF + CAR_TYPE + CLM_FREQ +
##     REVOKED + MVR_PTS + CAR_AGE + URBANICITY
##
##              Df  Sum of Sq        RSS    AIC
## <none>                     1.6803e+11 137490
## - HOMEKIDS    1   47201032 1.6808e+11 137490
## - SEX         1   71319769 1.6810e+11 137491
## - HOME_VAL    1   85132097 1.6812e+11 137492
## - PARENT1     1   97906975 1.6813e+11 137492
## - CLM_FREQ    1  100739436 1.6813e+11 137493
## - CAR_AGE     1  147680577 1.6818e+11 137495
## - REVOKED     1  167260491 1.6820e+11 137496
## - KIDSDRIV    1  170735073 1.6820e+11 137496
## - TRAVTIME    1  283139671 1.6832e+11 137501
## - TIF         1  314035565 1.6835e+11 137503
## - MSTATUS     1  353418972 1.6839e+11 137505
## - CAR_TYPE    5  598879597 1.6863e+11 137509
## - CAR_USE     1  459509327 1.6849e+11 137510
```

```
## - JOB         8  804605159 1.6884e+11 137513
## - MVR_PTS     1  927418649 1.6896e+11 137533
## - URBANICITY  1 2909345072 1.7094e+11 137628
```

```r
summary(BICBasePlusTransform)
```

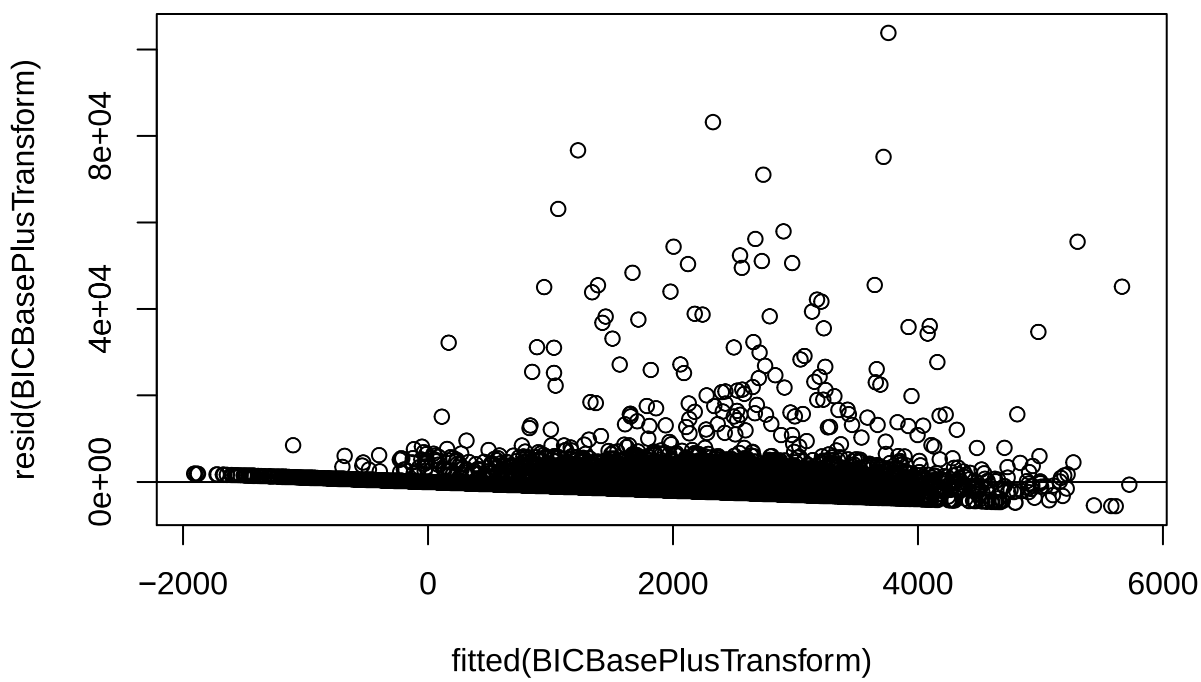
```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + HOMEKIDS + PARENT1 + HOME_VAL +
##     MSTATUS + SEX + JOB + TRAVTIME + CAR_USE + TIF + CAR_TYPE +
##     CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5615  -1693   -766    338 103828
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.641e+03  2.864e+02   5.730 1.04e-08 ***
## KIDSDRIV                         3.147e+02  1.095e+02   2.875 0.004055 **
## HOMEKIDS                         2.241e+02  1.483e+02   1.511 0.130704
## PARENT1Yes                       4.720e+02  2.168e+02   2.177 0.029518 *
## HOME_VAL                        -7.068e-02  3.482e-02  -2.030 0.042399 *
## MSTATUSz_No                      5.633e+02  1.362e+02   4.136 3.57e-05 ***
## SEXz_F                          -2.720e+02  1.464e+02  -1.858 0.063213 .
## JOBDoctor                       -8.557e+02  3.423e+02  -2.500 0.012447 *
## JOBHome Maker                    4.011e+01  2.295e+02   0.175 0.861271
## JOBLawyer                       -2.414e+02  2.346e+02  -1.029 0.303462
## JOBManager                      -1.092e+03  2.124e+02  -5.144 2.75e-07 ***
## JOBMiscellaneous                -5.044e+02  2.822e+02  -1.787 0.073912 .
## JOBProfessional                 -2.205e+02  1.975e+02  -1.117 0.264233
## JOBStudent                      -1.195e+02  2.201e+02  -0.543 0.587082
## JOBz_Blue Collar                -8.050e+01  1.899e+02  -0.424 0.671617
## TRAVTIME                         1.192e+01  3.220e+00   3.702 0.000215 ***
## CAR_USEPrivate                  -7.393e+02  1.568e+02  -4.716 2.45e-06 ***
## TIF                             -4.746e+01  1.217e+01  -3.899 9.75e-05 ***
## CAR_TYPEPanel Truck              4.580e+02  2.471e+02   1.854 0.063803 .
## CAR_TYPEPickup                   3.784e+02  1.680e+02   2.252 0.024324 *
## CAR_TYPESports Car               9.461e+02  2.038e+02   4.643 3.48e-06 ***
## CAR_TYPEVan                      5.835e+02  2.065e+02   2.826 0.004724 **
## CAR_TYPEz_SUV                    6.754e+02  1.648e+02   4.100 4.18e-05 ***
## CLM_FREQ                         1.078e+02  4.881e+01   2.208 0.027262 *
## REVOKEDYes                       4.409e+02  1.549e+02   2.845 0.004448 **
## MVR_PTS                          1.728e+02  2.580e+01   6.700 2.22e-11 ***
## CAR_AGE                         -2.983e+01  1.116e+01  -2.674 0.007520 **
## URBANICITYz_Highly Rural/ Rural -1.653e+03  1.393e+02 -11.867  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4545 on 8133 degrees of freedom
## Multiple R-squared:  0.0694, Adjusted R-squared:  0.06631
## F-statistic: 22.46 on 27 and 8133 DF,  p-value: < 2.2e-16
```

```r
results = NULL
modellist = list(m1 = regbaseplustransform, m2 = BICBasePlusTransform)
```

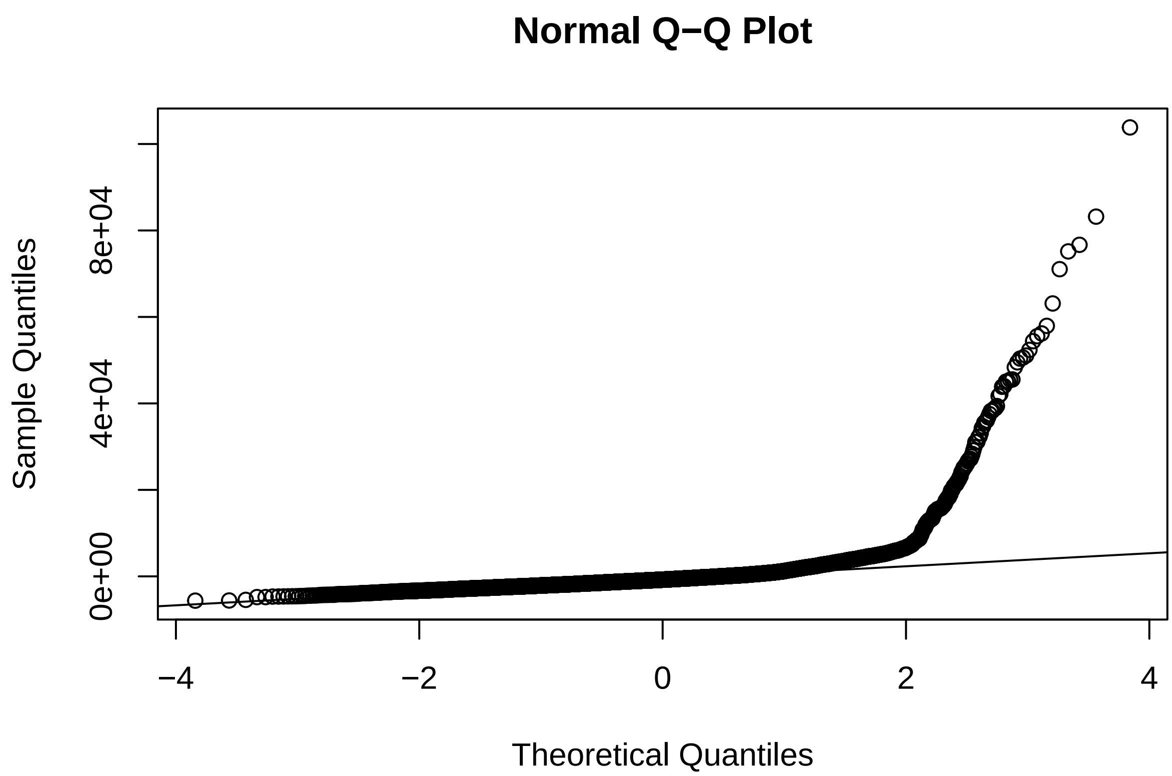
```r
for(i in names(modellist)){
    s = summary(modellist[[i]])
    name = i
    mse <- mean(s$residuals^2)
    r2 <- s$r.squared
    f <- s$fstatistic[1]
    k <- s$fstatistic[2]
    n <- s$fstatistic[3]
    results = rbind(results, data.frame(
        name = name, rsquared = r2, mse = mse, f = f,
        k = k, n = n))
}
rownames(results) = NULL
results
```

```
##   name   rsquared      mse        f  k    n
## 1   m1 0.06986176 20579455 17.95105 34 8126
## 2   m2 0.06939632 20589753 22.46253 27 8133
```

```r
###### Plots ######

plot(fitted(BICBasePlusTransform), resid(BICBasePlusTransform))
abline(h=0)
```




```r
qqnorm(BICBasePlusTransform$residuals)
qqline(BICBasePlusTransform$residuals)
```

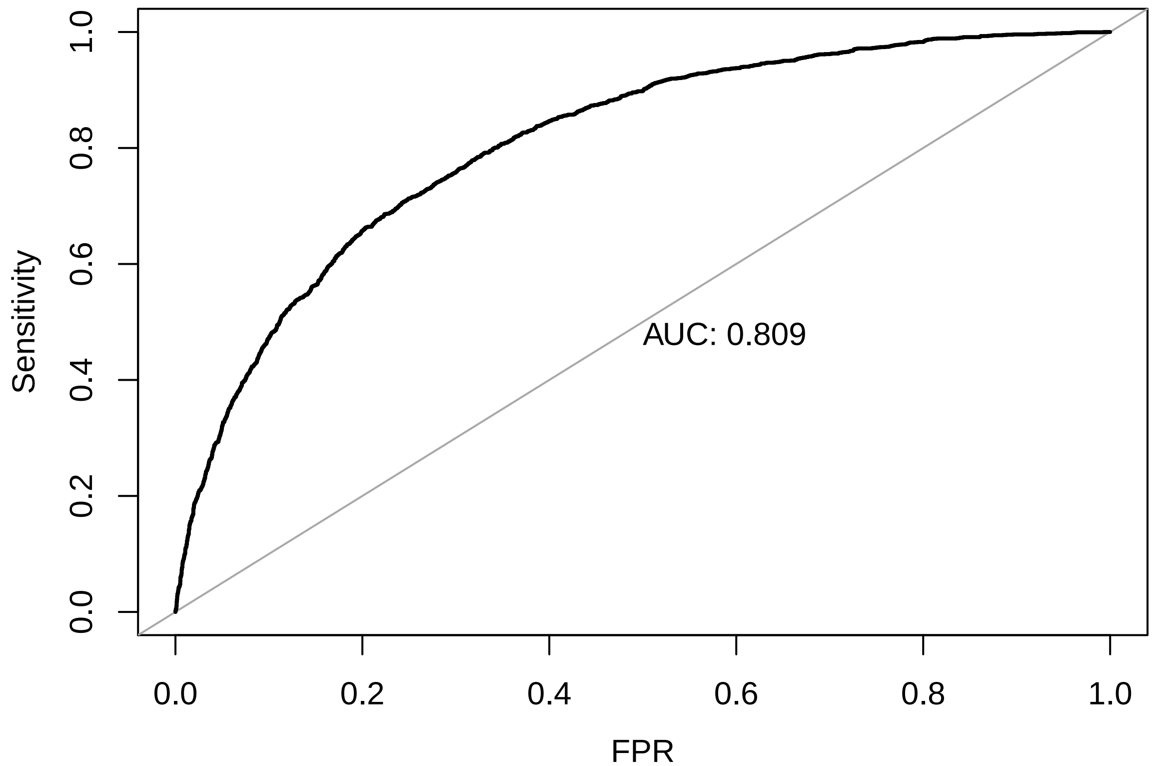


Comparing the base model and BICplusbase model, all the values are near identical from the above table. But the QQ plot is not normal as expected with heavier tails.

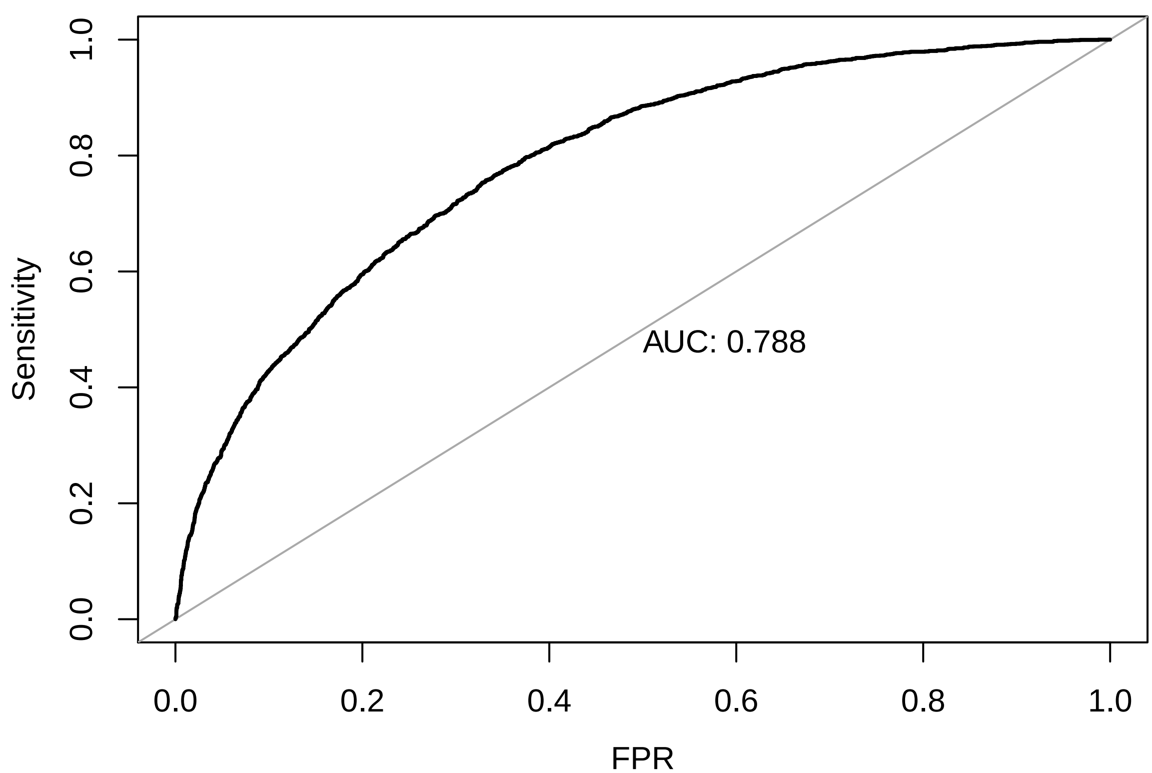
4. MODEL SELECTION:

From our 4 Model, we will first see the ROC and AUC curve.

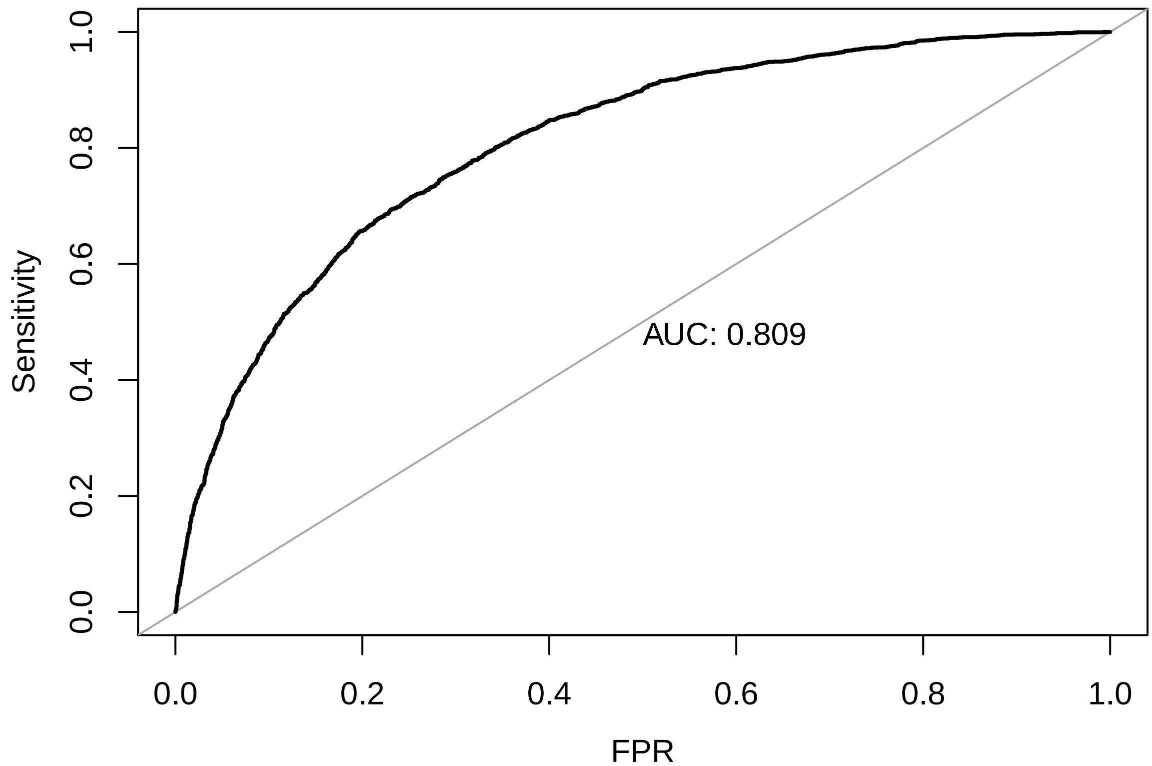
```
#plot(rocBasePlot, asp=NA, legacy.axes = TRUE, print.auc=TRUE, xlab="FPR")
plot(rocBaseTransformPlot, asp=NA, legacy.axes = TRUE, print.auc=TRUE, xlab="FPR")
```




```r
plot(rocBaseBackwardPlot, asp=NA, legacy.axes = TRUE, print.auc=TRUE, xlab="FPR")
```




```r
plot(rocBaseStepPlot, asp=NA, legacy.axes = TRUE, print.auc=TRUE, xlab="FPR")
```



confusion Matrix.

```
# Confusion Matrix of Base Transformation Model
baseTransformConfusion = as.factor(as.integer(fitted(base_transform_model ) > .5))
baseTransformCM = confusionMatrix(baseTransformConfusion, as.factor(base_transform_model$y), positive =
caretTransformResults = data.frame(Accurarcy = baseTransformCM$overall[['Accuracy']],
                              ClassErrorRate = 1 - baseTransformCM$overall[['Accuracy']],
                              Precision = baseTransformCM$byClass[['Precision']],
                              Senstivity = baseTransformCM$byClass[['Sensitivity']],
                              Specificity = baseTransformCM$byClass[['Specificity']],
                              F1 = baseTransformCM$byClass[['F1']])

# Confusion Matrix of Base Backward Elimination Model
baseBackwardConfusion = as.factor(as.integer(fitted(base_backward_model ) > .5))
baseBackwardCM = confusionMatrix(baseBackwardConfusion, as.factor(base_backward_model$y), positive = "1
caretBackwardResults = data.frame(Accurarcy = baseBackwardCM$overall[['Accuracy']],
                              ClassErrorRate = 1 - baseBackwardCM$overall[['Accuracy']],
                              Precision = baseBackwardCM$byClass[['Precision']],
                              Senstivity = baseBackwardCM$byClass[['Sensitivity']],
                              Specificity = baseBackwardCM$byClass[['Specificity']],
                              F1 = baseBackwardCM$byClass[['F1']])

# Confusion Matrix of Base Step Model
baseStepConfusion = as.factor(as.integer(fitted(base_step_model ) > .5))
baseStepCM = confusionMatrix(baseStepConfusion, as.factor(base_step_model$y), positive = "1")
caretStepResults = data.frame(Accurarcy = baseStepCM$overall[['Accuracy']],
                              ClassErrorRate = 1 - baseStepCM$overall[['Accuracy']],
                              Precision = baseStepCM$byClass[['Precision']],
                              Senstivity = baseStepCM$byClass[['Sensitivity']],
                              Specificity = baseStepCM$byClass[['Specificity']],
```

```
                          F1 = baseStepCM$byClass[['F1']])
TotalConMatrix = rbind(caretTransformResults, caretBackwardResults, caretStepResults)
knitr::kable(TotalConMatrix)
```

Accurarcy	ClassErrorRate	Precision	Senstivity	Specificity	F1
0.7872810	0.2127190	0.6545589	0.4101254	0.9224368	0.5042833
0.7789487	0.2210513	0.6438582	0.3627497	0.9280959	0.4640523
0.7869134	0.2130866	0.6528804	0.4105899	0.9217710	0.5041346

After analyzing all our four models, we will be using the Base_Step_Model for our prediction.

Since all the values are almost identical for all the models we will use Base_Step_Model for our prediction.

Evalution

Finally when we when we apply the BIC model to the evalution data, it predicts that there are 205 insurance customers that would have an auto accident and 1936 that would not.

```
eval_results = predict(base_step_model, newdata = test)
table(as.integer(eval_results > .5))
```

```
##
##    0    1
## 1936  205
```

```
eval_amount = predict(BICBasePlusTransform, test)
```

References

https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/