HW4_YQ

Youqing Xiang

July 7, 2016

Data Exploration

```r
library(PerformanceAnalytics)
```

```
## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
##
##     legend
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.4
```

```r
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.2.4
```

```r
library(knitr)
library(lattice)
library(caret)
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following objects are masked from 'package:xts':
##
##     first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
library(car)

data <- read.csv('insurance_training_data.csv')
summary(data)
```

```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ            INCOME
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   $0      : 615
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0           : 445
##  Median :45.00   Median :0.0000   Median :11.0   $26,840 :   4
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5   $48,509 :   4
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   $61,790 :   4
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   $107,375:   3
##  NA's   :6                        NA's   :454    (Other) :7086
##  PARENT1        HOME_VAL      MSTATUS      SEX              EDUCATION
##  No :7084   $0      :2294   Yes :4894   M  :3786   <High School :1203
##  Yes:1077           : 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
##
##
##
##
##
```

```r
dim(data)
```

```
## [1] 8161   26
```

Data Preparation

Deal with Missing Data and Nonsense Data

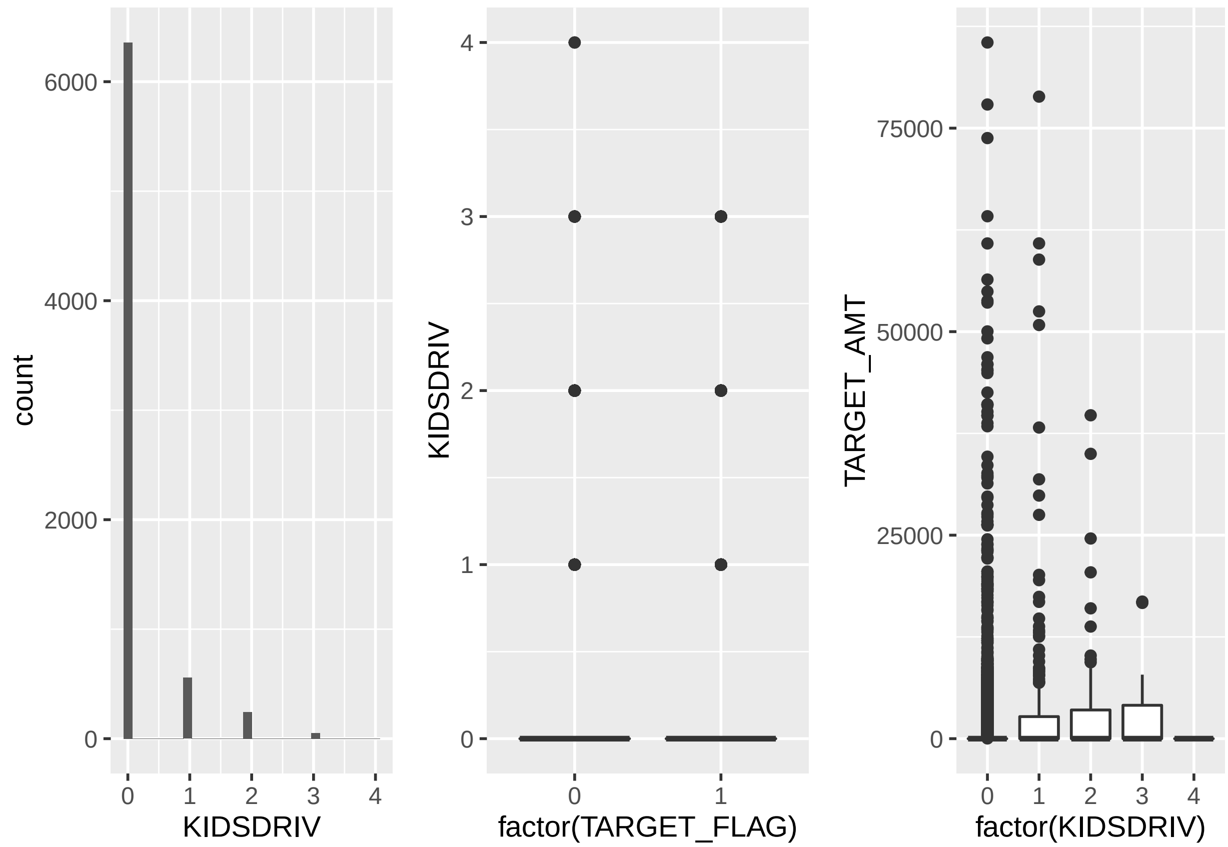
```r
# AGE
data <- data[!is.na(data$AGE),]

# YOJ
dataN <- data[!is.na(data$YOJ),]

# CAR_AGE
dataN <- dataN[!is.na(dataN$CAR_AGE),]
dataN <- dataN[dataN$CAR_AGE >= 0,]
```

Data Transformation

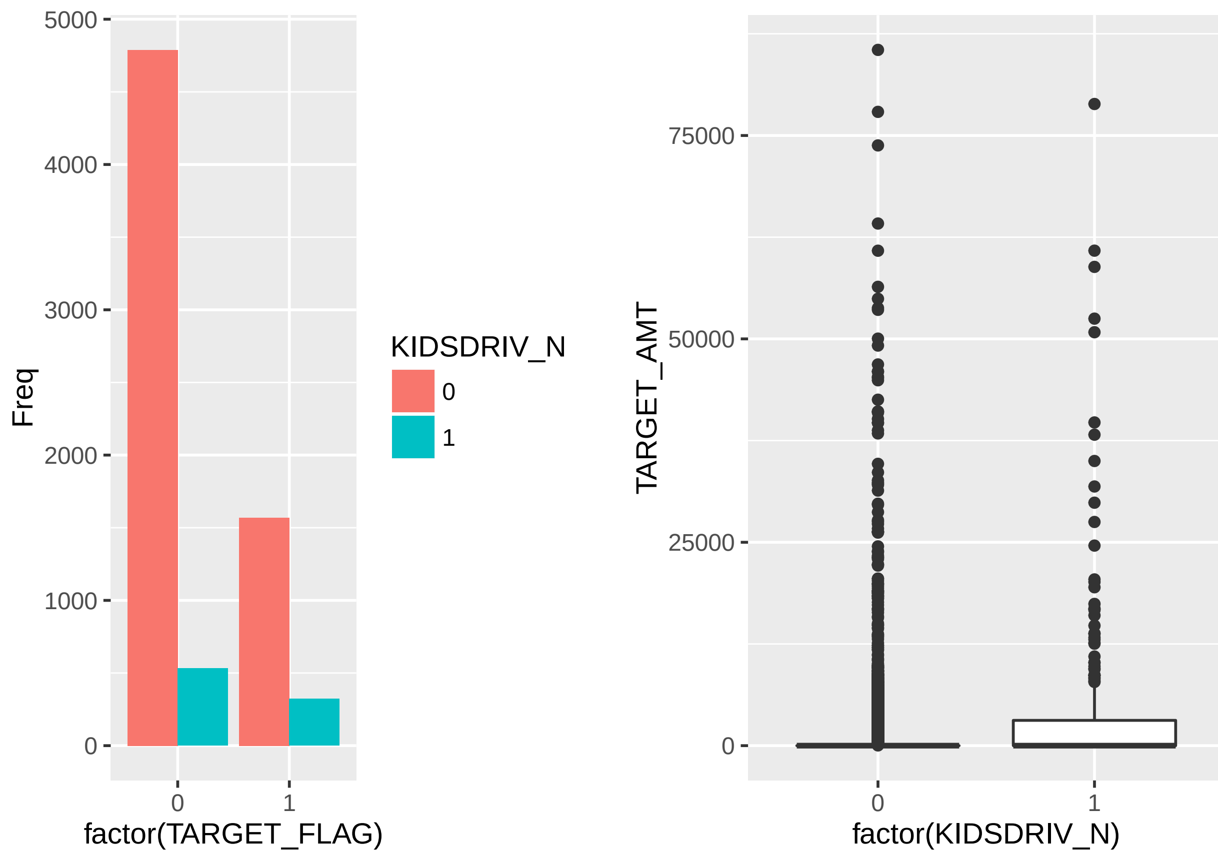
KIDSDRIV

```r
# Before transformation
p1 <- ggplot(dataN, aes(KIDSDRIV)) + geom_histogram()
p2 <- ggplot(dataN, aes(factor(TARGET_FLAG), KIDSDRIV)) + geom_boxplot()
p3 <- ggplot(dataN, aes(factor(KIDSDRIV), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,p3,ncol=3,nrow=1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



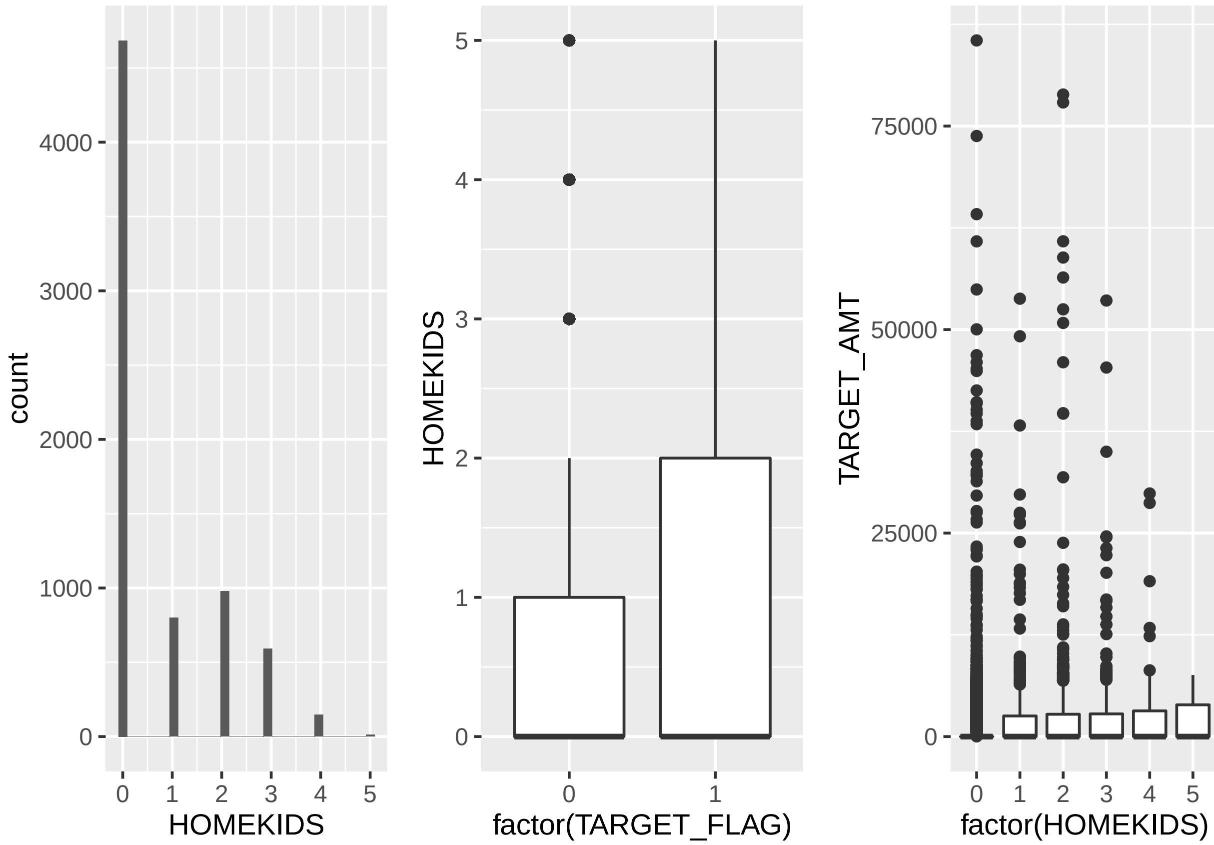

```r
# Data transformation
dataN$KIDSDRIV_N <- ifelse(dataN$KIDSDRIV == 0, 0, 1)
dataN$KIDSDRIV_N <- as.factor(dataN$KIDSDRIV_N)

# After transformation
t <- as.data.frame(table(KIDSDRIV_N=dataN$KIDSDRIV_N, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=KIDSDRIV_N),stat='identity',positio
p2 <- ggplot(dataN, aes(factor(KIDSDRIV_N), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```



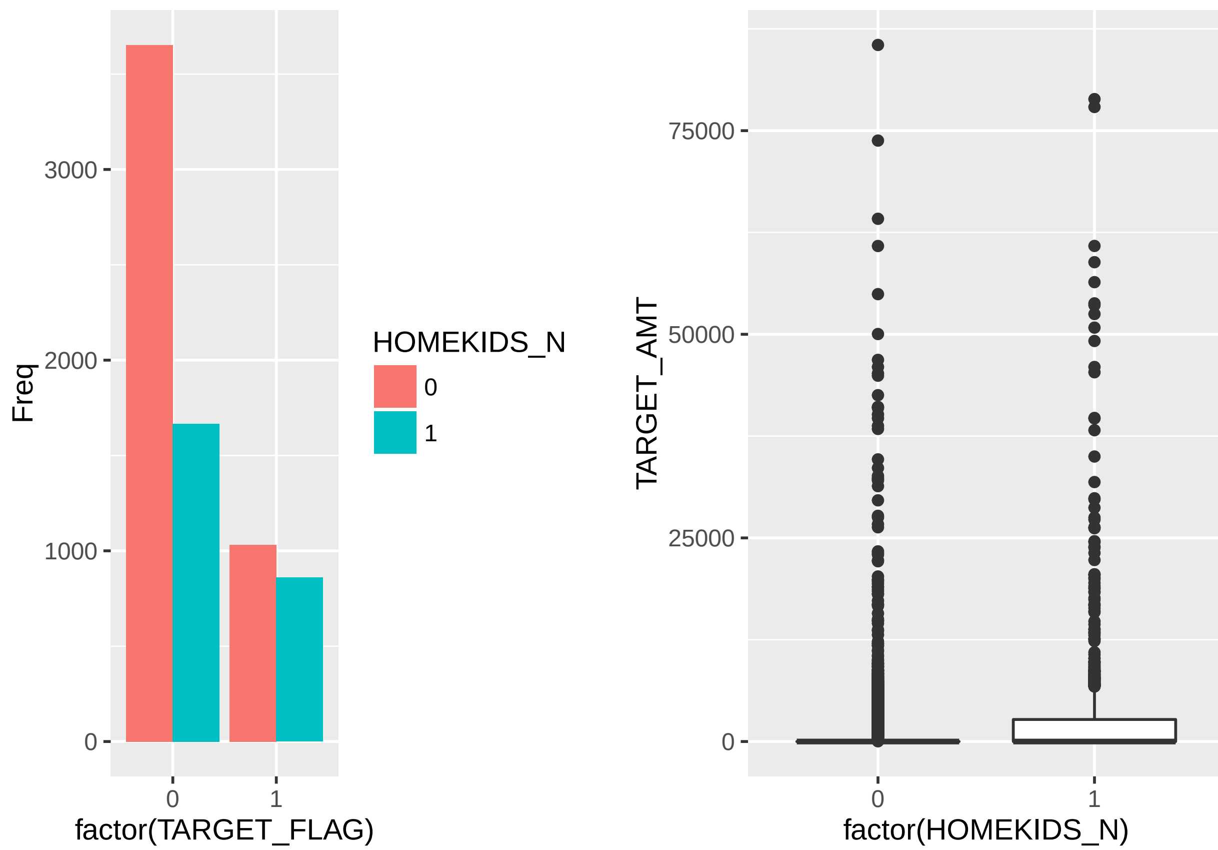
HOMEKIDS

```r
# Before transformation
p1 <- ggplot(dataN, aes(HOMEKIDS)) + geom_histogram()
p2 <- ggplot(dataN, aes(factor(TARGET_FLAG), HOMEKIDS)) + geom_boxplot()
p3 <- ggplot(dataN, aes(factor(HOMEKIDS), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,p3,ncol=3,nrow=1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Data transformation
dataN$HOMEKIDS_N <- ifelse(dataN$HOMEKIDS == 0, 0, 1)
dataN$HOMEKIDS_N <- as.factor(dataN$HOMEKIDS_N)

# After transformation
t <- as.data.frame(table(HOMEKIDS_N=dataN$HOMEKIDS_N, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=HOMEKIDS_N),stat='identity', positi
p2 <- ggplot(dataN, aes(factor(HOMEKIDS_N), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```



INCOME

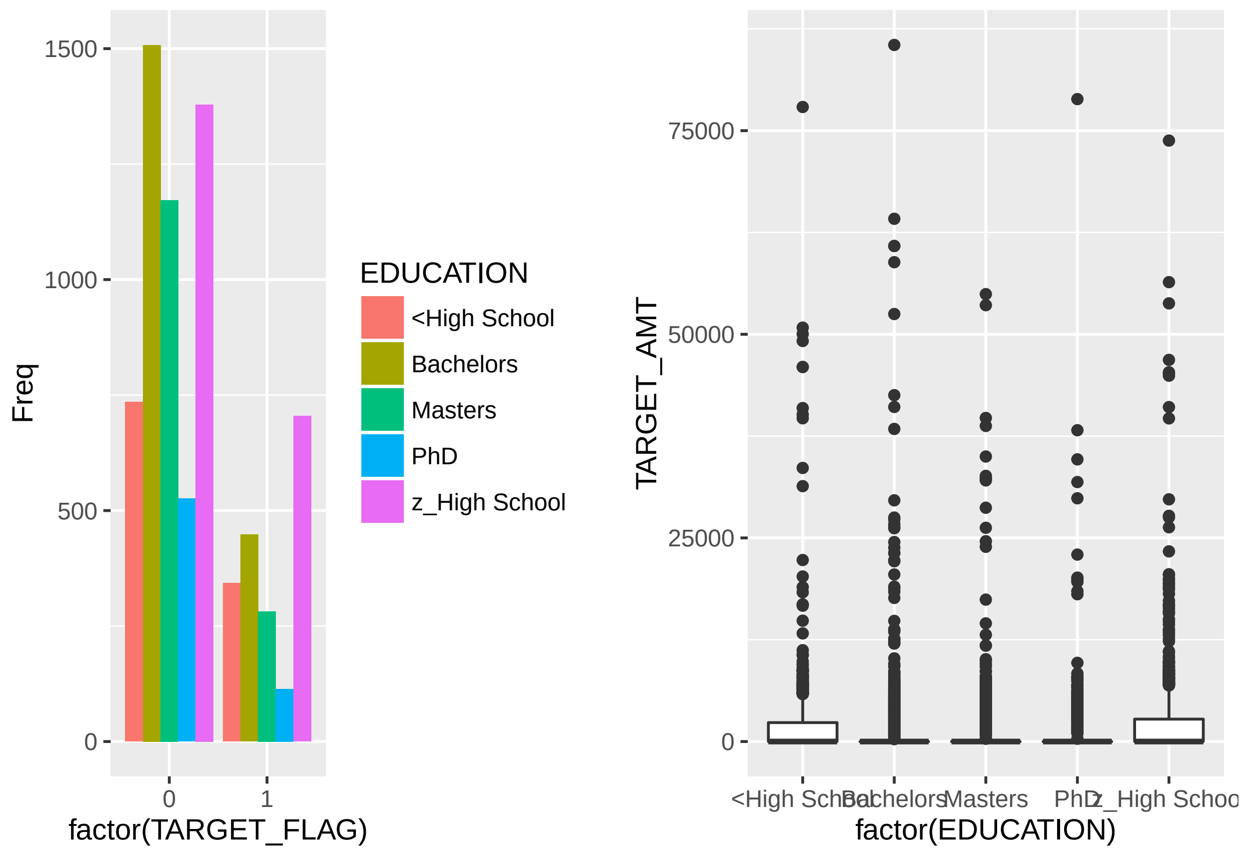
```
dataN$INCOME <- as.numeric(dataN$INCOME)
```

HOME_VAL

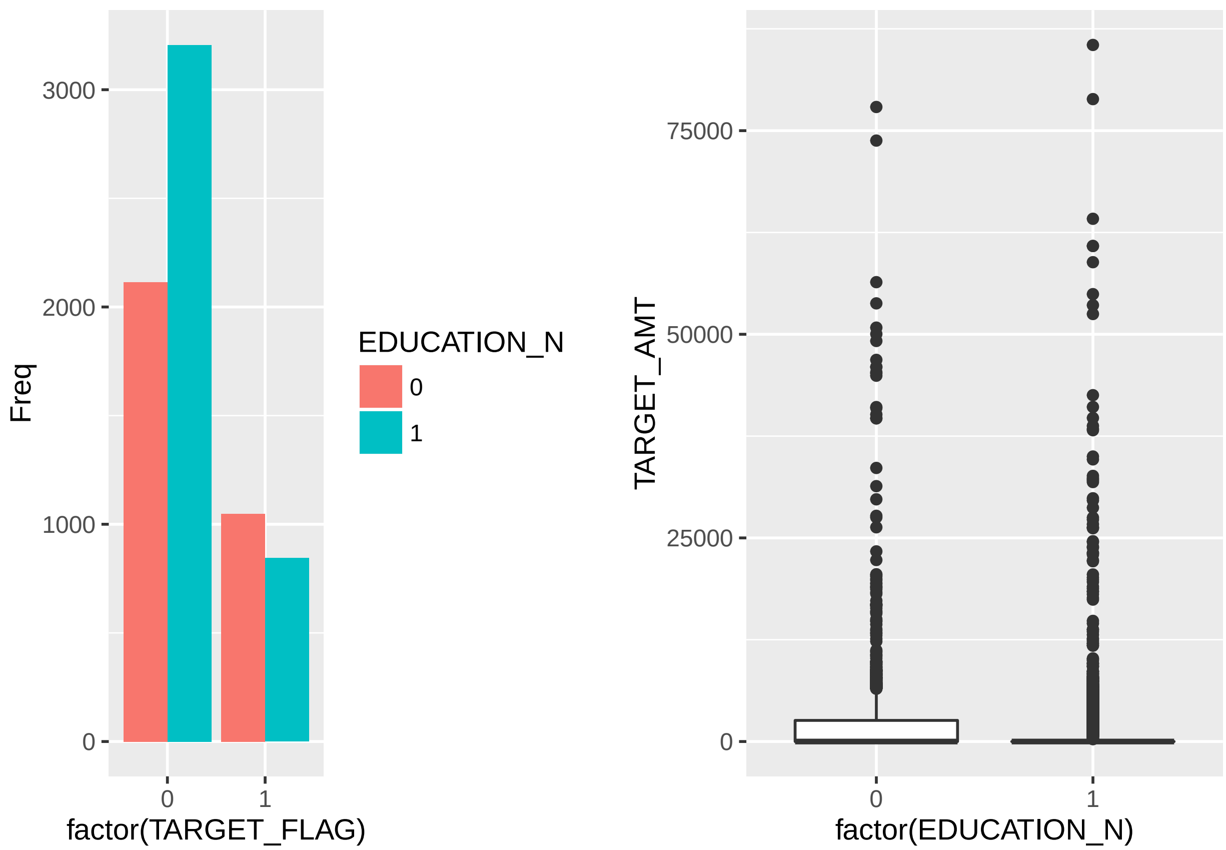
```
dataN$HOME_VAL <- as.numeric(dataN$HOME_VAL)
```

EDUCATION

```
# Before transformation
t <- as.data.frame(table(EDUCATION=dataN$EDUCATION, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=EDUCATION),stat='identity',
                                                            position=position_dodge())
p2 <- ggplot(dataN, aes(factor(EDUCATION), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```

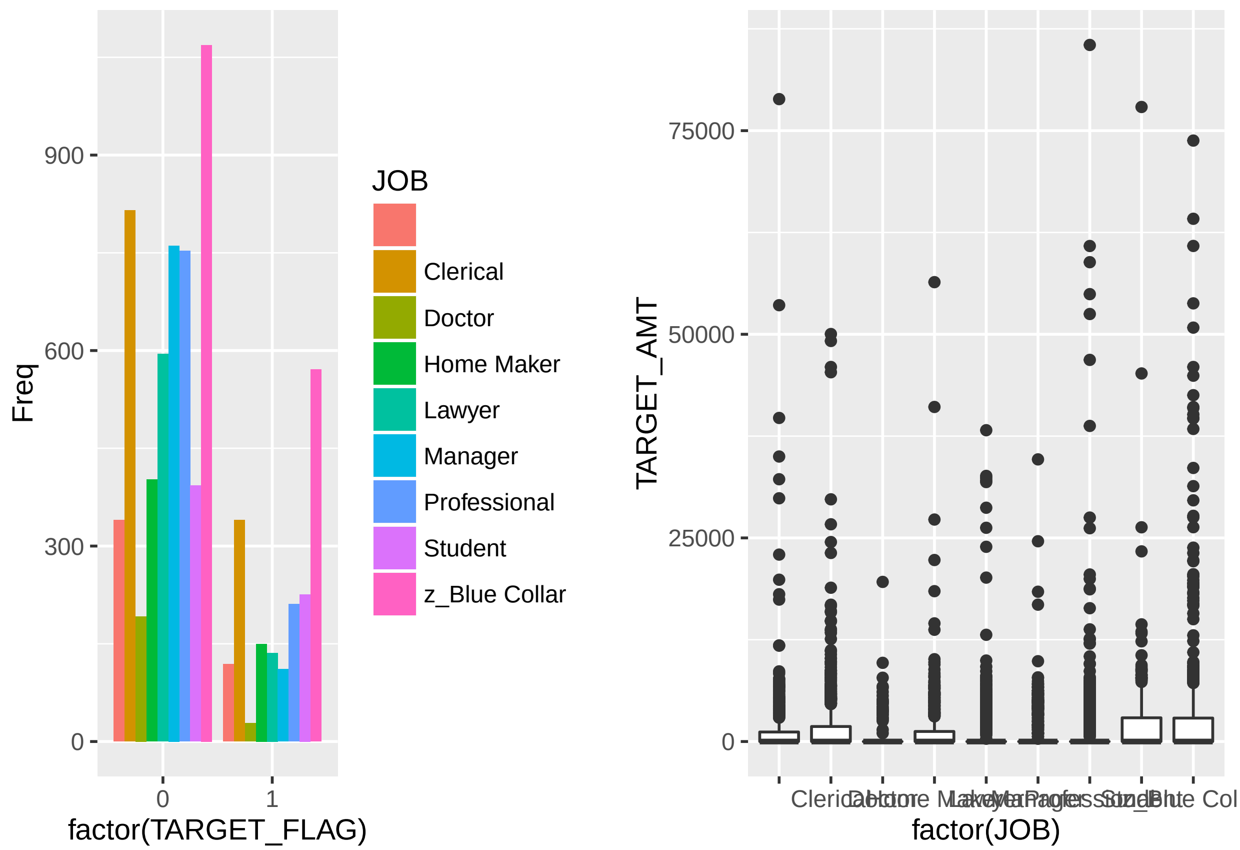



```
# Data transformation
dataN$EDUCATION_N <- ifelse(dataN$EDUCATION %in% c('<High School','z_High School'), 0, 1)
dataN$EDUCATION_N <- as.factor(dataN$EDUCATION_N)

# After transformation
t <- as.data.frame(table(EDUCATION_N=dataN$EDUCATION_N, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=EDUCATION_N),stat='identity',
                                                            position=position_dodge())
p2 <- ggplot(dataN, aes(factor(EDUCATION_N), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```

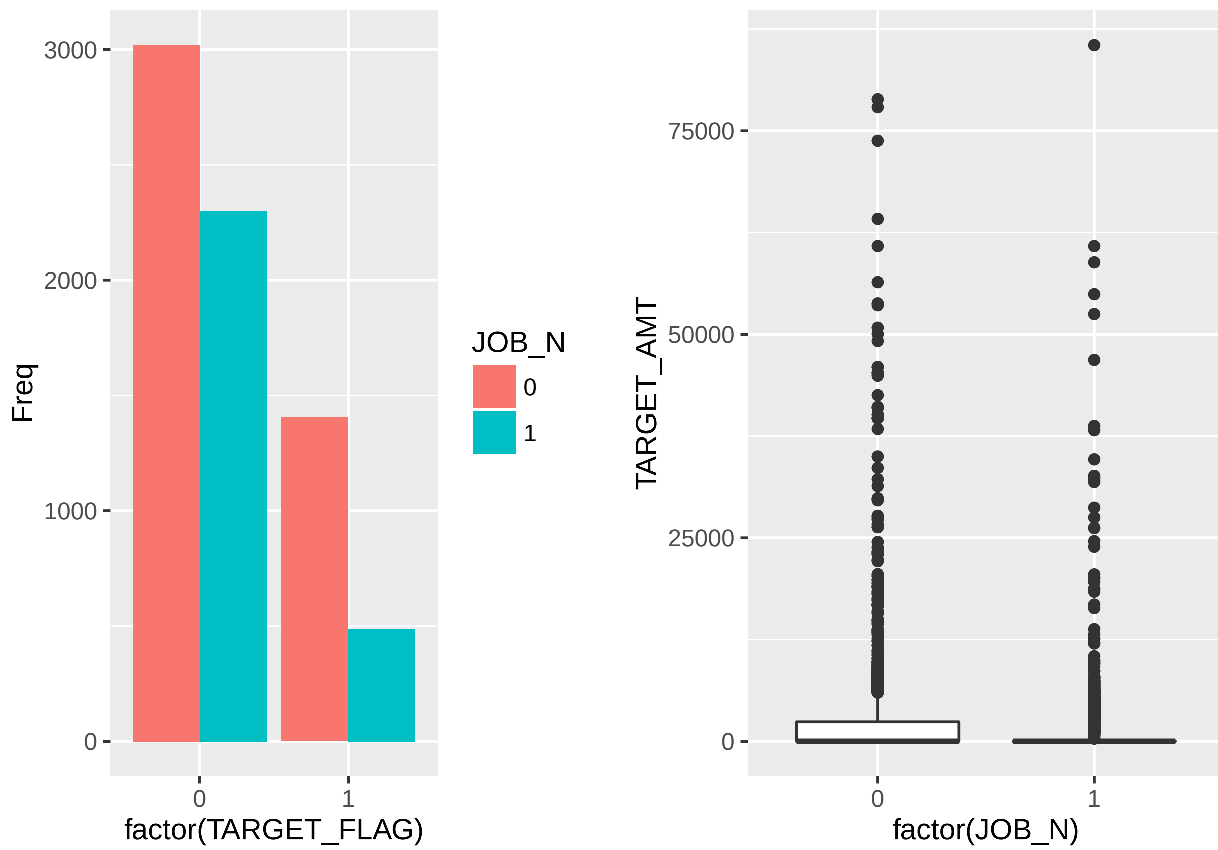


JOB

```
# Before transformation
t <- as.data.frame(table(JOB=dataN$JOB, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=JOB),stat='identity',
                                                           position=position_dodge())
p2 <- ggplot(dataN, aes(factor(JOB), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```




```r
# Data transformation
dataN$JOB_N <- ifelse(dataN$JOB %in% c('Doctor','Lawyer','Manager','Professional'), 1, 0)
dataN$JOB_N <- as.factor(dataN$JOB_N)

# After transformation
t <- as.data.frame(table(JOB_N=dataN$JOB_N, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=JOB_N),stat='identity',
                                                    position=position_dodge())
p2 <- ggplot(dataN, aes(factor(JOB_N), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```



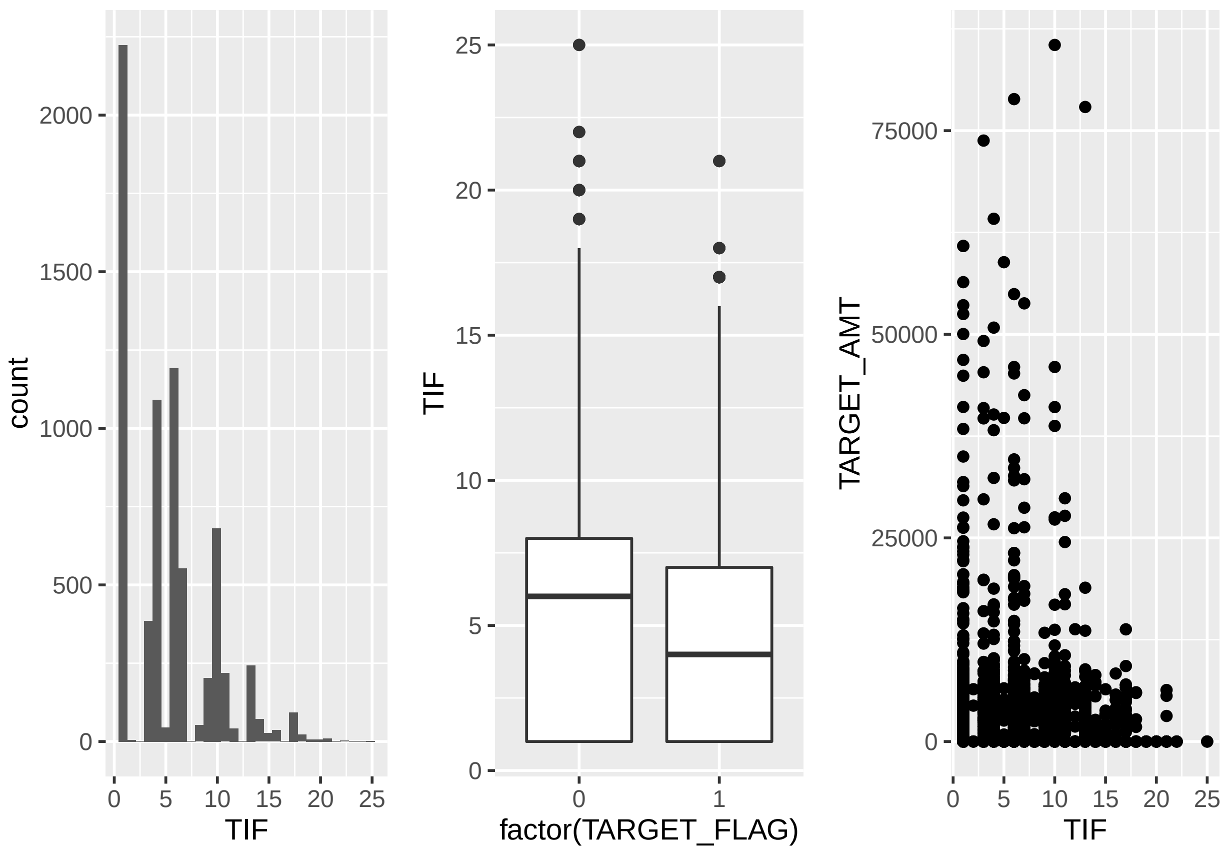
BLUEBOOK

```
dataN$BLUEBOOK <- as.numeric(dataN$BLUEBOOK)
```

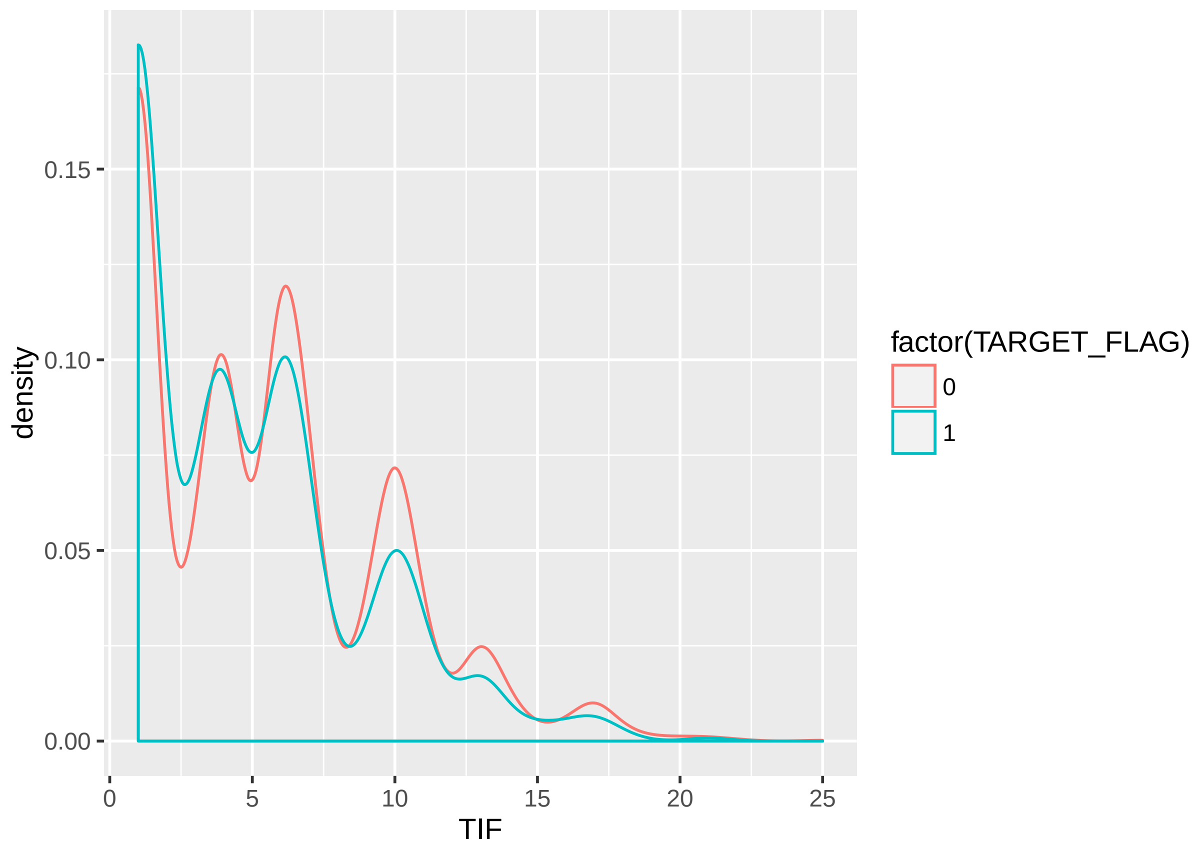
TIF

```
# Before transformation
p1 <-  ggplot(dataN, aes(TIF)) + geom_histogram()
p2 <- ggplot(dataN, aes(factor(TARGET_FLAG), TIF)) + geom_boxplot()
p3 <- ggplot(dataN, aes(TIF, TARGET_AMT)) + geom_point()
grid.arrange(p1,p2,p3,ncol=3,nrow=1)
```

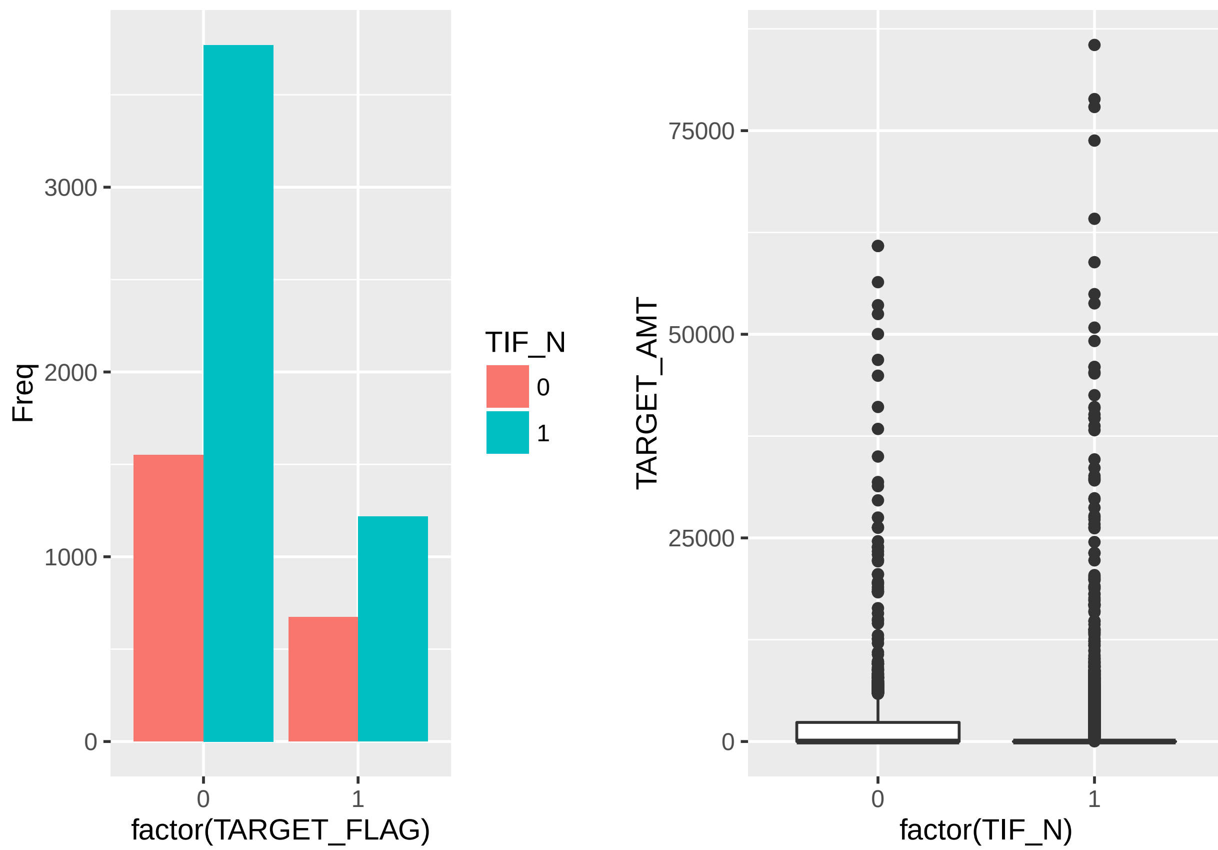
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(dataN, aes(x=TIF)) + geom_density(aes(colour=factor(TARGET_FLAG)))
```

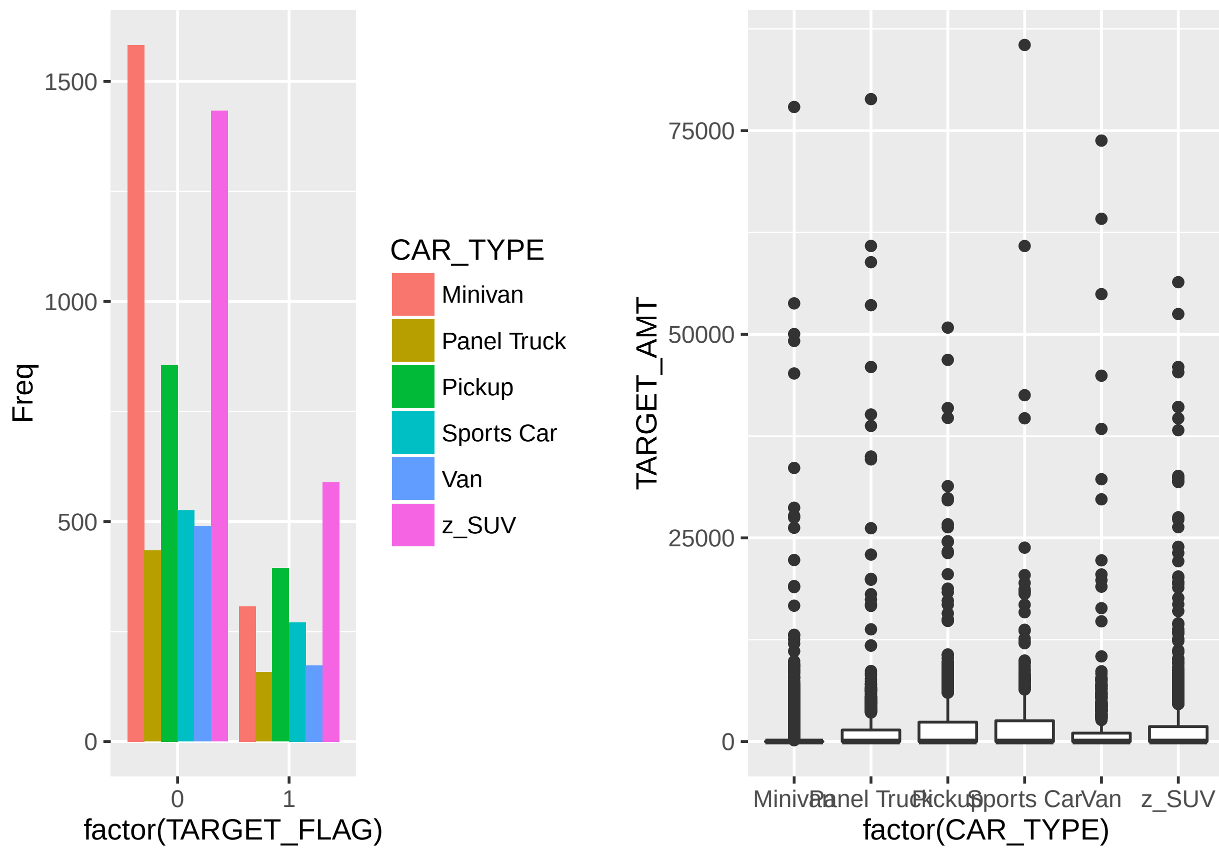



```r
# Data transformation
dataN$TIF_N <- ifelse(dataN$TIF > 1, 1, 0)
dataN$TIF_N <- as.factor(dataN$TIF_N)

# After transformation
t <- as.data.frame(table(TIF_N=dataN$TIF_N, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=TIF_N),stat='identity',
                                                           position=position_dodge())
p2 <- ggplot(dataN, aes(factor(TIF_N), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```

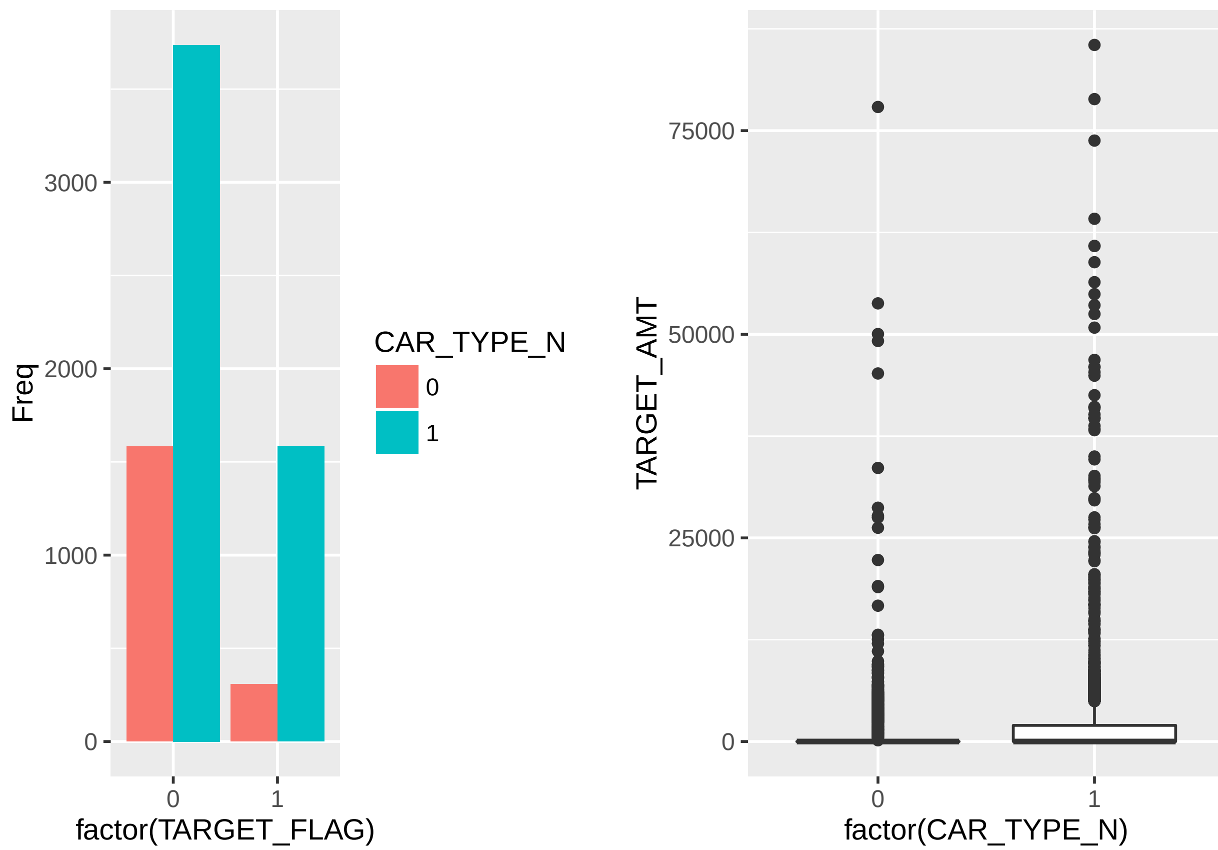


CAR_TYPE

```
# Before transformation
t <- as.data.frame(table(CAR_TYPE=dataN$CAR_TYPE, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=CAR_TYPE),stat='identity',
                                                           position=position_dodge())
p2 <- ggplot(dataN, aes(factor(CAR_TYPE), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```



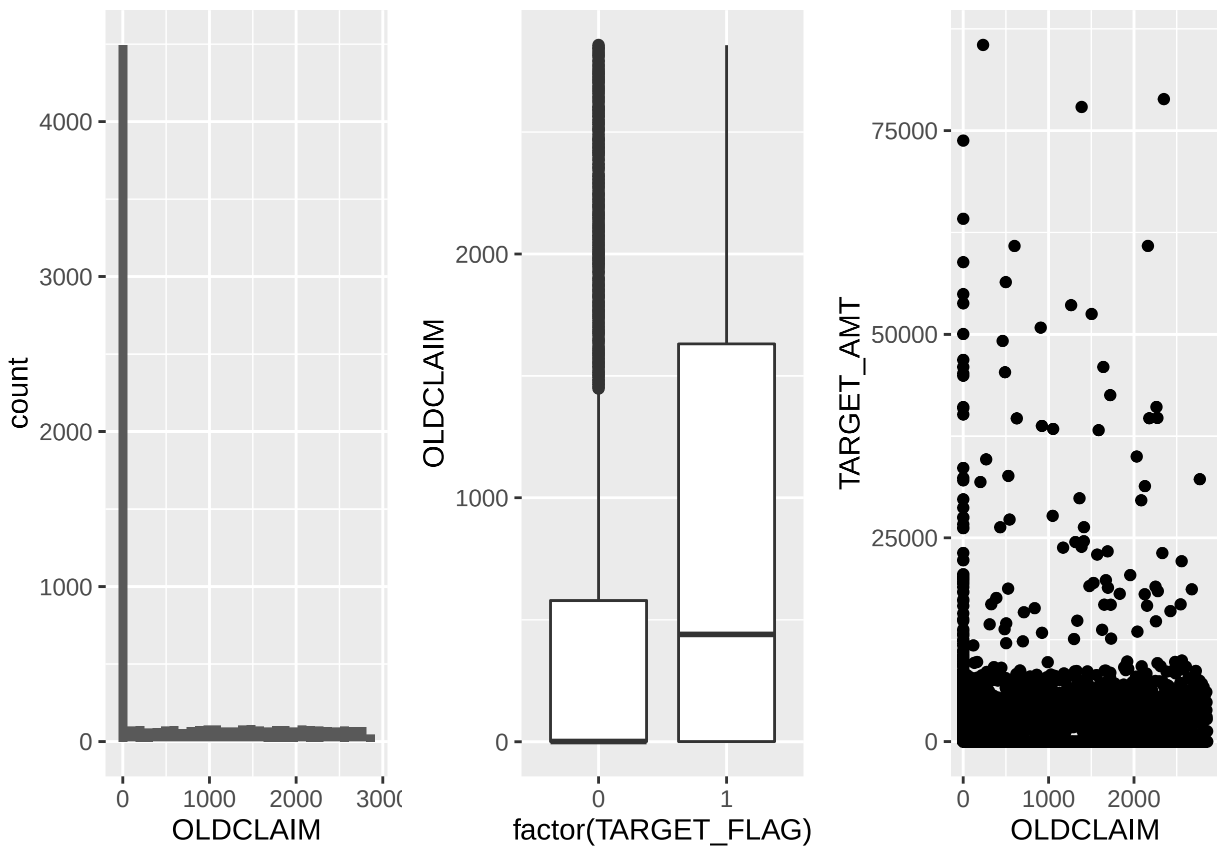

```
# Data transformation
dataN$CAR_TYPE_N <- ifelse(dataN$CAR_TYPE == 'Minivan', 0, 1)
dataN$CAR_TYPE_N <- as.factor(dataN$CAR_TYPE_N)

# After transformation
t <- as.data.frame(table(CAR_TYPE_N=dataN$CAR_TYPE_N, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=CAR_TYPE_N),stat='identity',
                                                            position=position_dodge())
p2 <- ggplot(dataN, aes(factor(CAR_TYPE_N), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```



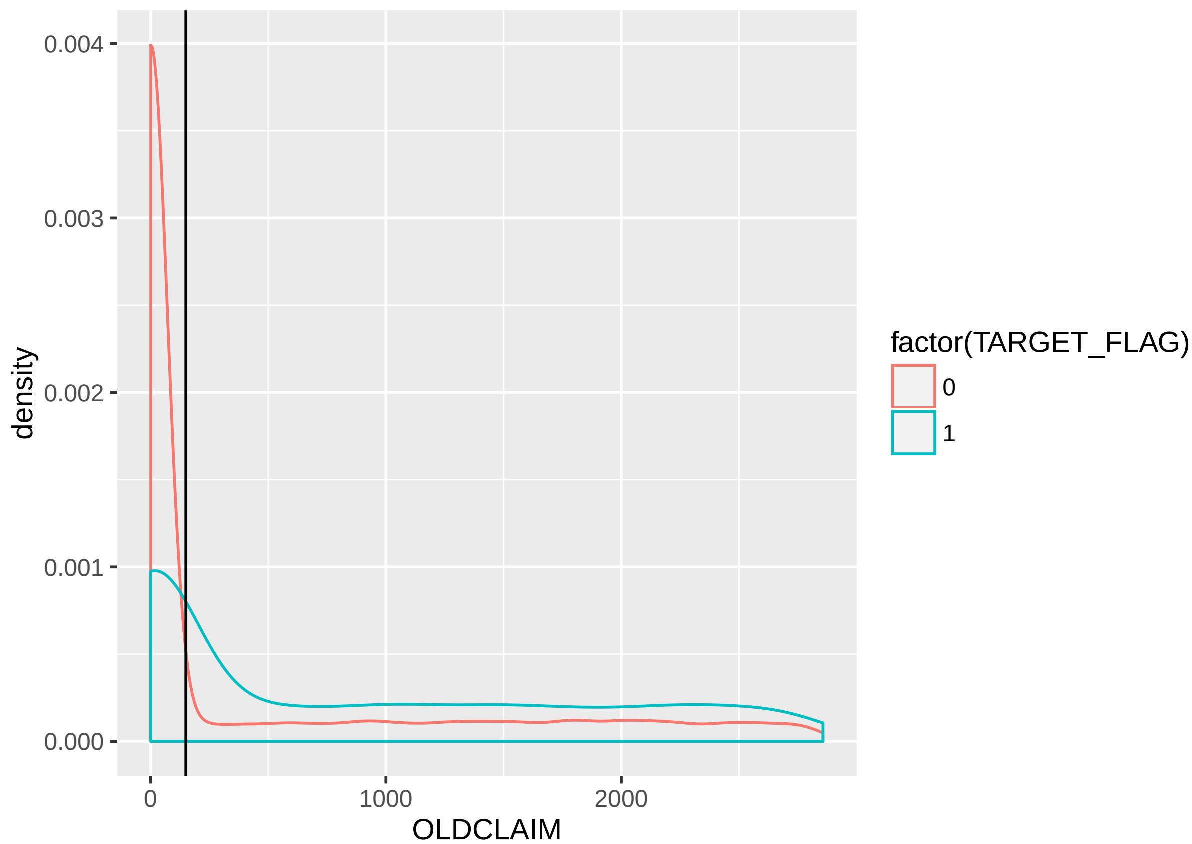
OLDCLAIM

```
# Before transformation
dataN$OLDCLAIM <- as.numeric(dataN$OLDCLAIM)
p1 <-  ggplot(dataN, aes(OLDCLAIM)) + geom_histogram()
p2 <- ggplot(dataN, aes(factor(TARGET_FLAG), OLDCLAIM)) + geom_boxplot()
p3 <- ggplot(dataN, aes(OLDCLAIM, TARGET_AMT)) + geom_point()
grid.arrange(p1,p2,p3,ncol=3,nrow=1)
```

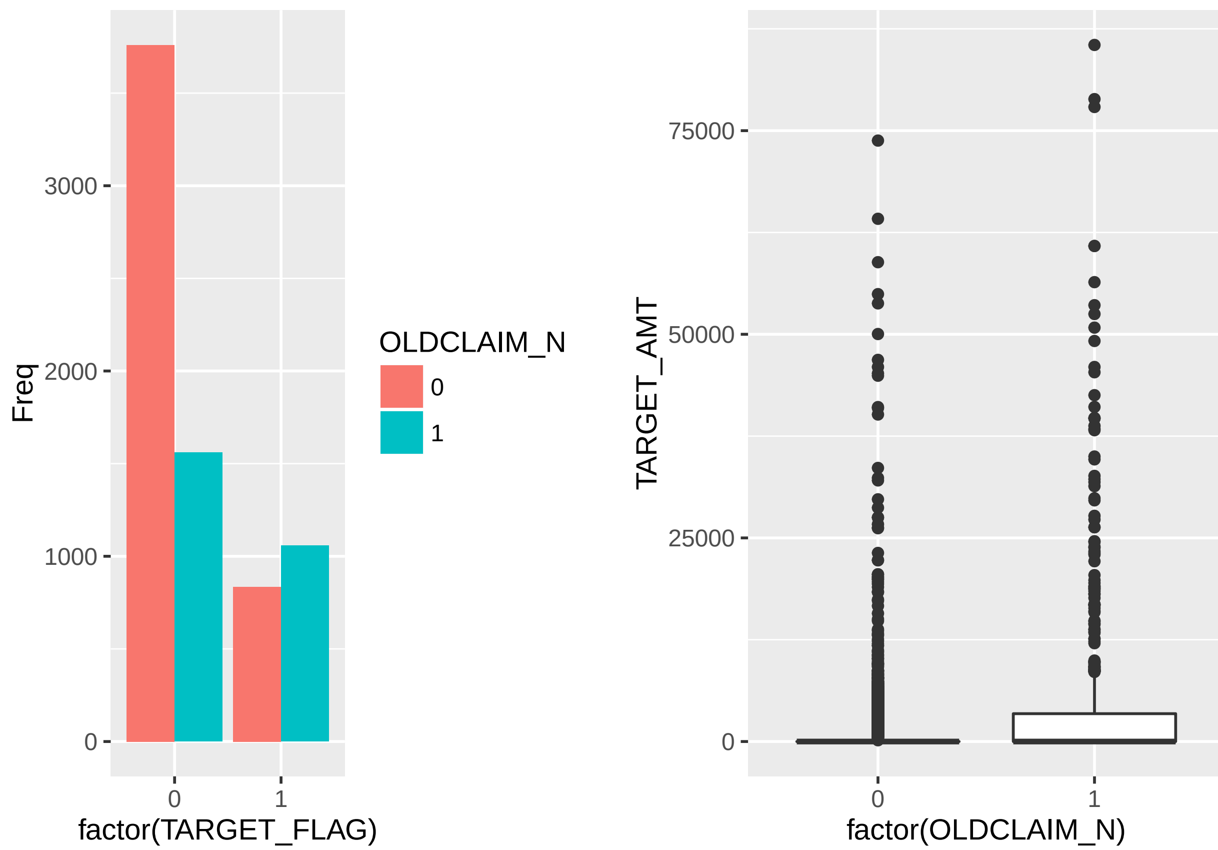
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(dataN, aes(x=OLDCLAIM)) + geom_density(aes(colour=factor(TARGET_FLAG))) +
  geom_vline(xintercept = 150)
```

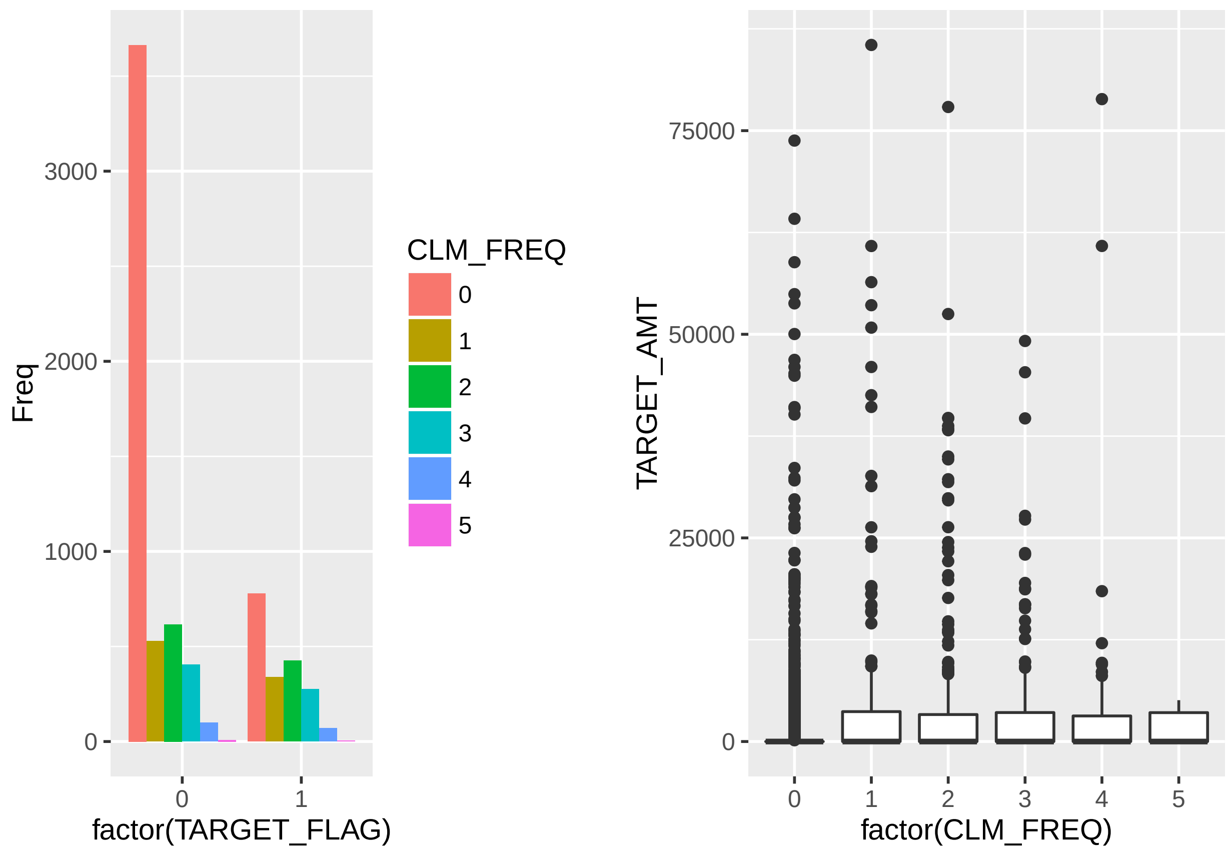



```
# Data transformation
dataN$OLDCLAIM_N <- ifelse(dataN$OLDCLAIM > 150, 1, 0)
dataN$OLDCLAIM_N <- as.factor(dataN$OLDCLAIM_N)

# After transformation
t <- as.data.frame(table(OLDCLAIM_N=dataN$OLDCLAIM_N, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=OLDCLAIM_N),stat='identity',positio

p2 <- ggplot(dataN, aes(factor(OLDCLAIM_N), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```

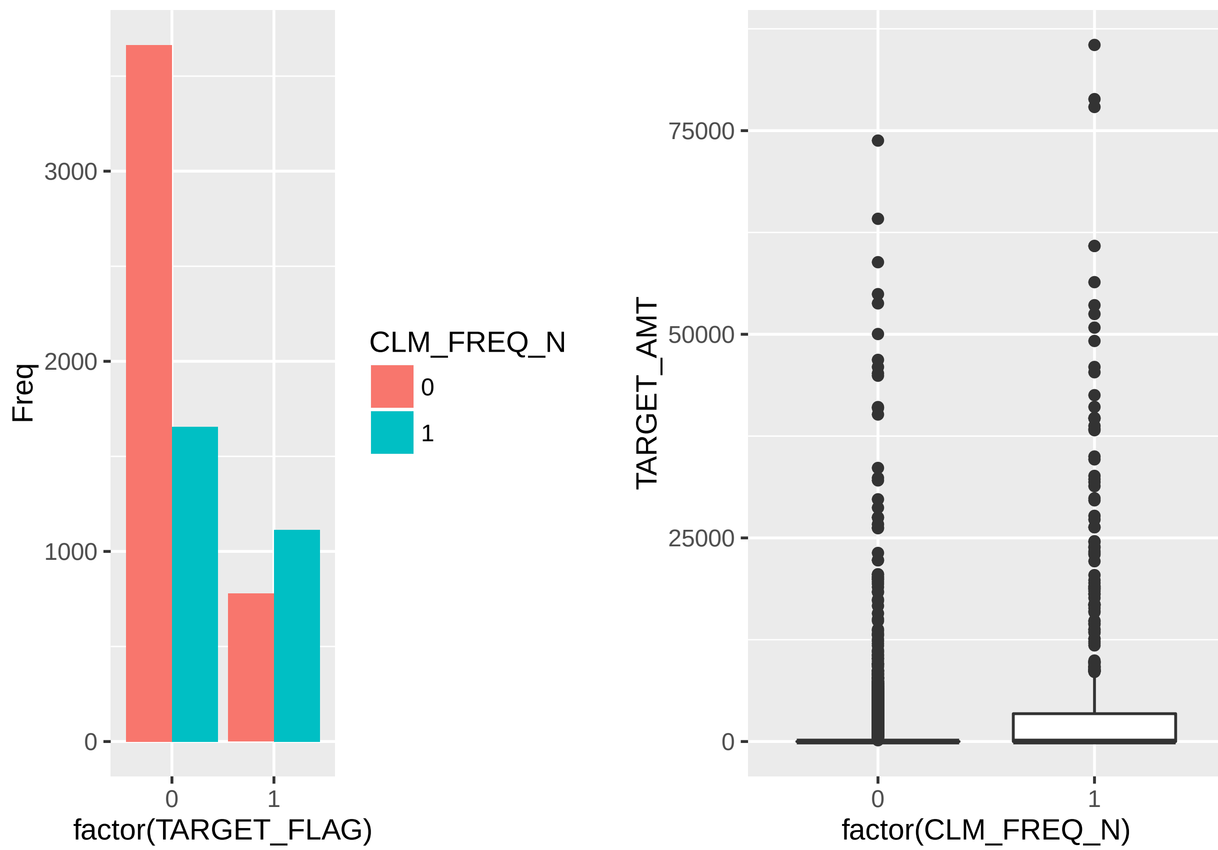


CLM_FREQ

```
# Before transformation
t <- as.data.frame(table(CLM_FREQ=dataN$CLM_FREQ, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=CLM_FREQ),stat='identity',
                                                           position=position_dodge())
p2 <- ggplot(dataN, aes(factor(CLM_FREQ), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```

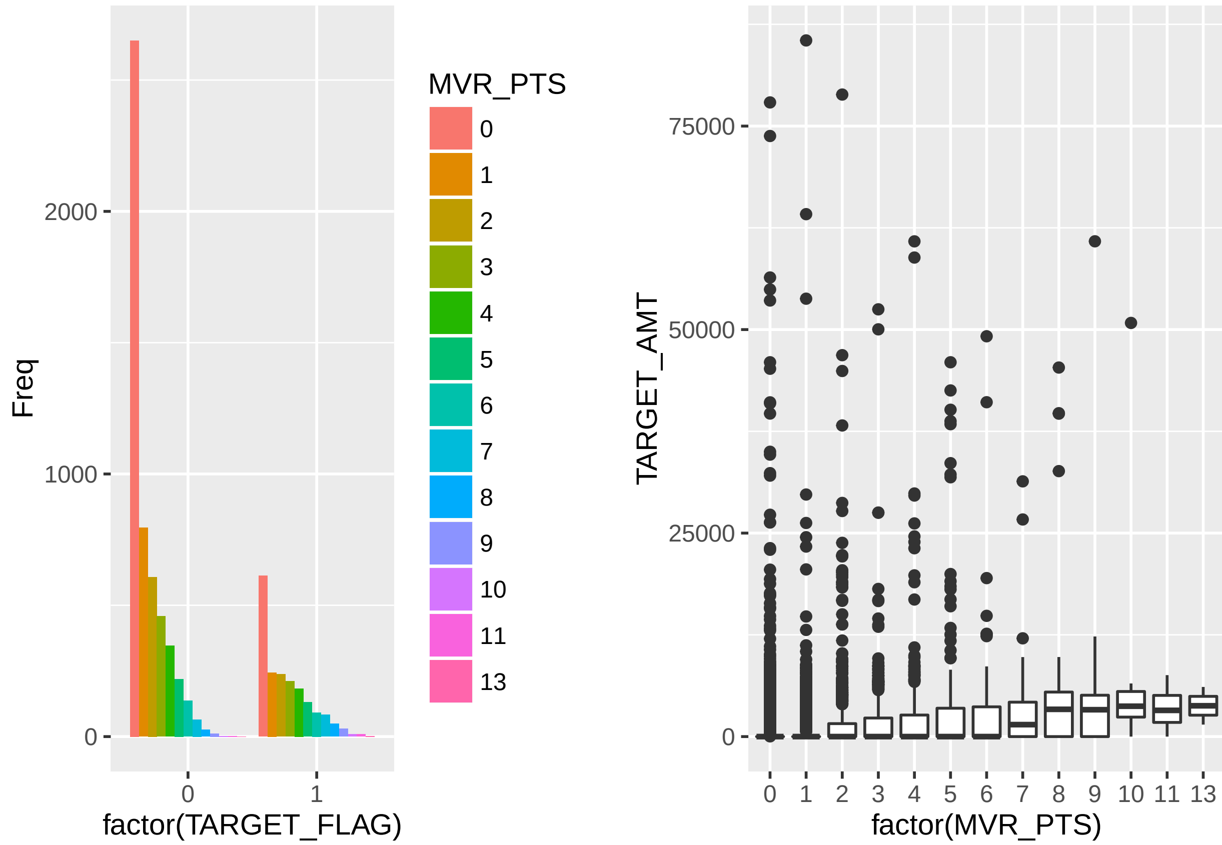



```
# Data transformation
dataN$CLM_FREQ_N <- ifelse(dataN$CLM_FREQ == 0, 0, 1)
dataN$CLM_FREQ_N <- as.factor(dataN$CLM_FREQ_N)

# After transformation
t <- as.data.frame(table(CLM_FREQ_N=dataN$CLM_FREQ_N, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=CLM_FREQ_N),stat='identity',positio
p2 <- ggplot(dataN, aes(factor(CLM_FREQ_N), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```

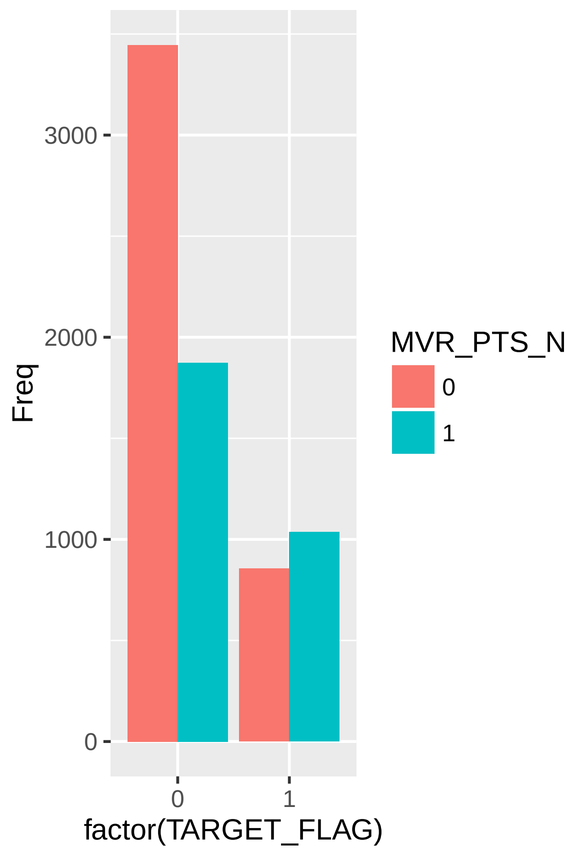


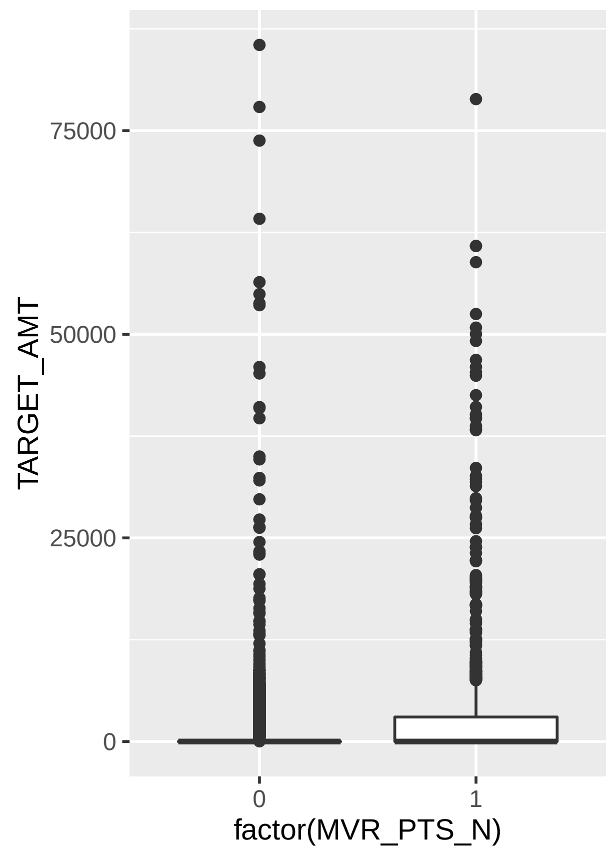
MVR_PTS

```
# Before transformation
t <- as.data.frame(table(MVR_PTS=dataN$MVR_PTS, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=MVR_PTS),stat='identity',
                                                    position=position_dodge())
p2 <- ggplot(dataN, aes(factor(MVR_PTS), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```




```
# Data transformation
dataN$MVR_PTS_N <- ifelse(dataN$MVR_PTS %in% c(0,1), 0, 1)
dataN$MVR_PTS_N <- as.factor(dataN$MVR_PTS_N)

# After transformation
t <- as.data.frame(table(MVR_PTS_N=dataN$MVR_PTS_N, TARGET_FLAG=dataN$TARGET_FLAG))
p1 <- ggplot(t, aes(factor(TARGET_FLAG), Freq)) + geom_bar(aes(fill=MVR_PTS_N),stat='identity',
                                                        position=position_dodge())
p2 <- ggplot(dataN, aes(factor(MVR_PTS_N), TARGET_AMT)) + geom_boxplot()
grid.arrange(p1,p2,ncol=2,nrow=1)
```





Modeling

TARGET_FLAG

Splitting data for training and testing models

```
set.seed(45)
inTrain_1 <- createDataPartition(y=dataN$TARGET_FLAG, p=0.7,list=FALSE)
training_1 <- dataN[inTrain_1,]
testing_1 <- dataN[-inTrain_1,]
```

Model 1 - Using original variables

```
training_1a <- select(training_1, -c(KIDSDRIV_N,HOMEKIDS_N,EDUCATION_N,JOB_N,TIF_N,CAR_TYPE_N,OLDCLAIM_
m11 <- glm(TARGET_FLAG ~ . -INDEX-TARGET_AMT, data=training_1a,family = binomial(link='probit'))
#summary(m11)
m12 <- update(m11, .~. -AGE-INCOME-BLUEBOOK-RED_CAR-CAR_AGE-CLM_FREQ-TIF)
#summary(m12)
TARGET_FLAG_m1 <- m12
summary(TARGET_FLAG_m1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + PARENT1 +
##     HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     CAR_TYPE + OLDCLAIM + REVOKED + MVR_PTS + URBANICITY, family = binomial(link = "probit"),
##     data = training_1a)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6419  -0.7341  -0.4183   0.6332   3.4313
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -8.743e-01  1.879e-01  -4.652 3.29e-06 ***
## KIDSDRIV                         2.430e-01  4.467e-02   5.441 5.31e-08 ***
## HOMEKIDS                         4.514e-02  2.533e-02   1.782 0.074783 .
## YOJ                             -8.602e-03  6.081e-03  -1.415 0.157209
## PARENT1Yes                       2.294e-01  8.006e-02   2.866 0.004163 **
## HOME_VAL                        -4.942e-05  1.492e-05  -3.313 0.000923 ***
## MSTATUSz_No                      2.581e-01  5.666e-02   4.555 5.23e-06 ***
## SEXz_F                          -2.060e-01  6.452e-02  -3.193 0.001409 **
## EDUCATIONBachelors              -3.586e-01  7.817e-02  -4.587 4.49e-06 ***
## EDUCATIONMasters                -4.169e-01  1.156e-01  -3.607 0.000309 ***
## EDUCATIONPhD                    -3.871e-01  1.378e-01  -2.810 0.004958 **
## EDUCATIONz_High School          -8.491e-02  6.955e-02  -1.221 0.222140
## JOBClerical                      1.532e-01  1.399e-01   1.095 0.273612
## JOBDoctor                       -1.634e-01  1.831e-01  -0.893 0.372115
## JOBHome Maker                    3.207e-01  1.449e-01   2.214 0.026825 *
## JOBLawyer                        8.849e-02  1.207e-01   0.733 0.463544
## JOBManager                      -3.654e-01  1.225e-01  -2.982 0.002863 **
## JOBProfessional                  6.245e-02  1.274e-01   0.490 0.623926
## JOBStudent                       1.307e-01  1.507e-01   0.867 0.385910
## JOBz_Blue Collar                 1.167e-01  1.336e-01   0.874 0.382269
## TRAVTIME                         8.334e-03  1.376e-03   6.059 1.37e-09 ***
## CAR_USEPrivate                  -4.292e-01  6.730e-02  -6.378 1.79e-10 ***
## CAR_TYPEPanel Truck              8.583e-02  1.040e-01   0.826 0.409041
## CAR_TYPEPickup                   3.291e-01  7.185e-02   4.580 4.65e-06 ***
## CAR_TYPESports Car               6.937e-01  8.783e-02   7.898 2.83e-15 ***
## CAR_TYPEVan                      1.477e-01  8.742e-02   1.690 0.091030 .
## CAR_TYPEz_SUV                    5.142e-01  7.363e-02   6.984 2.88e-12 ***
## OLDCLAIM                         1.465e-04  2.495e-05   5.871 4.32e-09 ***
## REVOKEDYes                       4.120e-01  5.983e-02   6.886 5.73e-12 ***
## MVR_PTS                          6.983e-02  1.006e-02   6.939 3.94e-12 ***
## URBANICITYz_Highly Rural/ Rural -1.213e+00  7.263e-02 -16.701  < 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: 5772.6  on 5048  degrees of freedom
## Residual deviance: 4576.8  on 5018  degrees of freedom
## AIC: 4638.8
##
## Number of Fisher Scoring iterations: 5
```

Model 2 - Using transformed variables

```
training_1b <- select(training_1, -c(KIDSDRIV,HOMEKIDS,EDUCATION,JOB,TIF,CAR_TYPE,OLDCLAIM,CLM_FREQ,MVR_
m21 <- glm(TARGET_FLAG ~ . -INDEX-TARGET_AMT, data=training_1b,family = binomial(link='probit'))
#summary(m21)
m22 <- update(m21, .~. -AGE-INCOME-PARENT1-SEX-BLUEBOOK-RED_CAR-CAR_AGE-OLDCLAIM_N)
#summary(m22)
TARGET_FLAG_m2 <- m22
summary(TARGET_FLAG_m2)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ YOJ + HOME_VAL + MSTATUS + TRAVTIME +
##     CAR_USE + REVOKED + URBANICITY + KIDSDRIV_N + HOMEKIDS_N +
##     EDUCATION_N + JOB_N + TIF_N + CAR_TYPE_N + CLM_FREQ_N + MVR_PTS_N,
##     family = binomial(link = "probit"), data = training_1b)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2747  -0.7465  -0.4216   0.7186   3.5148
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -8.503e-01  1.104e-01  -7.702 1.34e-14 ***
## YOJ                            -1.382e-02  5.296e-03  -2.610 0.009061 **
## HOME_VAL                       -5.350e-05  1.431e-05  -3.739 0.000185 ***
## MSTATUSz_No                     3.283e-01  4.697e-02   6.990 2.75e-12 ***
## TRAVTIME                        8.599e-03  1.370e-03   6.275 3.49e-10 ***
## CAR_USEPrivate                 -2.512e-01  4.659e-02  -5.391 7.00e-08 ***
## REVOKEDYes                      3.780e-01  5.942e-02   6.361 2.01e-10 ***
## URBANICITYz_Highly Rural/ Rural -1.136e+00  7.230e-02 -15.708  < 2e-16 ***
## KIDSDRIV_N1                     3.594e-01  6.983e-02   5.146 2.65e-07 ***
## HOMEKIDS_N1                     2.293e-01  5.033e-02   4.557 5.19e-06 ***
## EDUCATION_N1                   -3.507e-01  5.064e-02  -6.926 4.32e-12 ***
## JOB_N1                         -2.476e-01  5.614e-02  -4.410 1.03e-05 ***
## TIF_N1                         -2.234e-01  4.500e-02  -4.963 6.92e-07 ***
## CAR_TYPE_N1                     3.707e-01  5.280e-02   7.021 2.21e-12 ***
## CLM_FREQ_N1                     3.920e-01  4.646e-02   8.437  < 2e-16 ***
## MVR_PTS_N1                      2.086e-01  4.577e-02   4.559 5.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5772.6  on 5048  degrees of freedom
## Residual deviance: 4619.0  on 5033  degrees of freedom
## AIC: 4651
##
## Number of Fisher Scoring iterations: 5
```

Model 3 - Combining both of original and transformed variables

```
m31 <- glm(TARGET_FLAG ~ KIDSDRIV_N+HOMEKIDS_N+YOJ+HOME_VAL+PARENT1+MSTATUS+SEX+
             EDUCATION_N+JOB_N+TIF_N+CAR_USE+CAR_TYPE_N+OLDCLAIM+REVOKED+MVR_PTS+
             URBANICITY, data=training_1,family = binomial(link='probit'))
#summary(m31)
m32 <- update(m31, .~. -SEX-PARENT1)
#summary(m32)
TARGET_FLAG_m3 <- m32
summary(TARGET_FLAG_m3)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV_N + HOMEKIDS_N + YOJ + HOME_VAL +
##     MSTATUS + EDUCATION_N + JOB_N + TIF_N + CAR_USE + CAR_TYPE_N +
##     OLDCLAIM + REVOKED + MVR_PTS + URBANICITY, family = binomial(link = "probit"),
##     data = training_1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9602  -0.7385  -0.4321   0.7071   3.4550
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -5.233e-01  9.844e-02  -5.316 1.06e-07 ***
## KIDSDRIV_N1                     3.619e-01  6.964e-02   5.197 2.02e-07 ***
## HOMEKIDS_N1                     2.099e-01  5.013e-02   4.186 2.84e-05 ***
## YOJ                            -1.294e-02  5.293e-03  -2.445 0.014504 *
## HOME_VAL                       -5.223e-05  1.425e-05  -3.665 0.000248 ***
## MSTATUSz_No                     3.216e-01  4.673e-02   6.882 5.89e-12 ***
## EDUCATION_N1                   -3.430e-01  5.047e-02  -6.796 1.07e-11 ***
## JOB_N1                         -2.600e-01  5.588e-02  -4.654 3.26e-06 ***
## TIF_N1                         -2.249e-01  4.478e-02  -5.023 5.09e-07 ***
## CAR_USEPrivate                 -2.569e-01  4.645e-02  -5.531 3.19e-08 ***
## CAR_TYPE_N1                     3.586e-01  5.241e-02   6.841 7.84e-12 ***
## OLDCLAIM                        1.457e-04  2.474e-05   5.889 3.90e-09 ***
## REVOKEDYes                      4.183e-01  5.915e-02   7.071 1.54e-12 ***
## MVR_PTS                         7.647e-02  9.944e-03   7.691 1.47e-14 ***
## URBANICITYz_Highly Rural/ Rural -1.100e+00  7.047e-02 -15.607  < 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: 5772.6  on 5048  degrees of freedom
## Residual deviance: 4658.8  on 5034  degrees of freedom
## AIC: 4688.8
##
## Number of Fisher Scoring iterations: 5
```

TARGET_AMT

Splitting data for training and testing models

```
set.seed(1234)
inTrain_2 <- createDataPartition(y=dataN$TARGET_AMT, p=0.7,list=FALSE)
training_2 <- dataN[inTrain_2,]
testing_2 <- dataN[-inTrain_2,]
```

Model 1 - Using original variables

```
training_2a <- select(training_2, -c(KIDSDRIV_N,HOMEKIDS_N,EDUCATION_N,JOB_N,TIF_N,
                                     CAR_TYPE_N,OLDCLAIM_N,CLM_FREQ_N,MVR_PTS_N))
M11 <- lm( TARGET_AMT~ .-TARGET_FLAG-INDEX, data=training_2a)
#summary(M11)
M12 <- update(M11,.~.-AGE-HOMEKIDS-YOJ-INCOME-SEX-EDUCATION-BLUEBOOK-RED_CAR-OLDCLAIM-CLM_FREQ)
#summary(M12)
TARGET_AMT_m1 <- M12
summary(TARGET_AMT_m1)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + PARENT1 + HOME_VAL + MSTATUS +
##     JOB + TRAVTIME + CAR_USE + TIF + CAR_TYPE + REVOKED + MVR_PTS +
##     CAR_AGE + URBANICITY, data = training_2a)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5982  -1709   -741    394  82925
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.536e+03  4.442e+02   3.458 0.000550 ***
## KIDSDRIV                        4.767e+02  1.308e+02   3.645 0.000270 ***
## PARENT1Yes                      5.503e+02  2.229e+02   2.469 0.013587 *
## HOME_VAL                       -1.167e-01  4.463e-02  -2.614 0.008976 **
## MSTATUSz_No                     5.234e+02  1.640e+02   3.191 0.001428 **
## JOBClerical                    -1.804e+02  3.605e+02  -0.501 0.616713
## JOBDoctor                      -7.379e+02  4.773e+02  -1.546 0.122149
## JOBHome Maker                  -1.841e+02  3.972e+02  -0.464 0.642904
## JOBLawyer                      -3.720e+02  3.661e+02  -1.016 0.309589
## JOBManager                     -1.237e+03  3.422e+02  -3.614 0.000304 ***
## JOBProfessional                -3.047e+02  3.383e+02  -0.901 0.367801
## JOBStudent                     -2.367e+02  3.867e+02  -0.612 0.540533
## JOBz_Blue Collar                9.903e+01  3.344e+02   0.296 0.767104
## TRAVTIME                        1.531e+01  4.071e+00   3.760 0.000172 ***
## CAR_USEPrivate                 -5.031e+02  1.989e+02  -2.529 0.011454 *
## TIF                            -6.142e+01  1.554e+01  -3.953 7.81e-05 ***
## CAR_TYPEPanel Truck             7.286e+02  3.094e+02   2.355 0.018575 *
## CAR_TYPEPickup                  4.767e+02  2.105e+02   2.264 0.023609 *
```

```
## CAR_TYPESports Car               9.798e+02  2.357e+02   4.157 3.28e-05 ***
## CAR_TYPEVan                      4.385e+02  2.567e+02   1.708 0.087659 .
## CAR_TYPEz_SUV                    5.593e+02  1.767e+02   3.165 0.001560 **
## REVOKEDYes                       4.980e+02  1.941e+02   2.566 0.010323 *
## MVR_PTS                          2.071e+02  3.045e+01   6.800 1.17e-11 ***
## CAR_AGE                         -3.640e+01  1.391e+01  -2.616 0.008911 **
## URBANICITYz_Highly Rural/ Rural -1.690e+03  1.733e+02  -9.753  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4541 on 5024 degrees of freedom
## Multiple R-squared:  0.07502,    Adjusted R-squared:  0.0706
## F-statistic: 16.98 on 24 and 5024 DF,  p-value: < 2.2e-16
```

Model 2 - Using transformed variables

```
training_2b <- select(training_2, -c(KIDSDRIV,HOMEKIDS,EDUCATION,JOB,TIF,CAR_TYPE,
                                     OLDCLAIM,CLM_FREQ,MVR_PTS))
M21 <- lm( TARGET_AMT~ .-TARGET_FLAG-INDEX, data=training_2b)
#summary(M21)
M22 <- update(M21, .~.-AGE-YOJ-INCOME-PARENT1-SEX-BLUEBOOK-RED_CAR-HOMEKIDS_N-
                EDUCATION_N-OLDCLAIM_N-CLM_FREQ_N)
#summary(M22)
TARGET_AMT_m2 <- M22
summary(TARGET_AMT_m2)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ HOME_VAL + MSTATUS + TRAVTIME + CAR_USE +
##     REVOKED + CAR_AGE + URBANICITY + KIDSDRIV_N + JOB_N + TIF_N +
##     CAR_TYPE_N + MVR_PTS_N, data = training_2b)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4785  -1732   -807    351  83476
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.503e+03  2.802e+02   5.362 8.59e-08 ***
## HOME_VAL                        -1.093e-01  4.296e-02  -2.544 0.010984 *
## MSTATUSz_No                      7.531e+02  1.435e+02   5.249 1.59e-07 ***
## TRAVTIME                         1.654e+01  4.081e+00   4.052 5.15e-05 ***
## CAR_USEPrivate                  -5.600e+02  1.448e+02  -3.867 0.000111 ***
## REVOKEDYes                       5.098e+02  1.944e+02   2.623 0.008747 **
## CAR_AGE                         -3.843e+01  1.253e+01  -3.066 0.002183 **
## URBANICITYz_Highly Rural/ Rural -1.729e+03  1.695e+02 -10.197  < 2e-16 ***
## KIDSDRIV_N1                      9.363e+02  1.990e+02   4.705 2.61e-06 ***
## JOB_N1                          -5.980e+02  1.601e+02  -3.734 0.000190 ***
## TIF_N1                          -4.763e+02  1.382e+02  -3.447 0.000572 ***
## CAR_TYPE_N1                      5.886e+02  1.494e+02   3.939 8.29e-05 ***
## MVR_PTS_N1                       7.281e+02  1.322e+02   5.505 3.86e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4559 on 5036 degrees of freedom
## Multiple R-squared:  0.06539,    Adjusted R-squared:  0.06316
## F-statistic: 29.36 on 12 and 5036 DF,  p-value: < 2.2e-16
```

Model 3 - Combining both of original and transformed variables

```
M31 <- lm( TARGET_AMT~HOME_VAL+MSTATUS+TRAVTIME+CAR_USE+REVOKED+CAR_AGE+URBANICITY+
             KIDSDRIV+JOB_N+TIF+CAR_TYPE_N+MVR_PTS+PARENT1, data=training_2)
#summary(M31)
TARGET_AMT_m3 <- M31
summary(TARGET_AMT_m3)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ HOME_VAL + MSTATUS + TRAVTIME + CAR_USE +
##     REVOKED + CAR_AGE + URBANICITY + KIDSDRIV + JOB_N + TIF +
##     CAR_TYPE_N + MVR_PTS + PARENT1, data = training_2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6055  -1695   -765    306  83683
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.435e+03  2.752e+02   5.215 1.91e-07 ***
## HOME_VAL                     -1.056e-01  4.288e-02  -2.464 0.013779 *
## MSTATUSz_No                   5.450e+02  1.623e+02   3.358 0.000791 ***
## TRAVTIME                      1.596e+01  4.072e+00   3.919 9.01e-05 ***
## CAR_USEPrivate               -5.449e+02  1.445e+02  -3.771 0.000165 ***
## REVOKEDYes                    4.973e+02  1.938e+02   2.566 0.010325 *
## CAR_AGE                      -3.437e+01  1.252e+01  -2.744 0.006084 **
## URBANICITYz_Highly Rural/ Rural -1.649e+03  1.698e+02  -9.706  < 2e-16 ***
## KIDSDRIV                      4.675e+02  1.307e+02   3.575 0.000353 ***
## JOB_N1                       -5.738e+02  1.599e+02  -3.588 0.000336 ***
## TIF                          -6.140e+01  1.553e+01  -3.954 7.79e-05 ***
## CAR_TYPE_N1                   5.654e+02  1.491e+02   3.793 0.000151 ***
## MVR_PTS                       2.164e+02  3.034e+01   7.133 1.13e-12 ***
## PARENT1Yes                    5.510e+02  2.224e+02   2.478 0.013248 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4547 on 5035 degrees of freedom
## Multiple R-squared:  0.0707, Adjusted R-squared:  0.06831
## F-statistic: 29.47 on 13 and 5035 DF,  p-value: < 2.2e-16
```

Model Selection

TARGET_FLAG

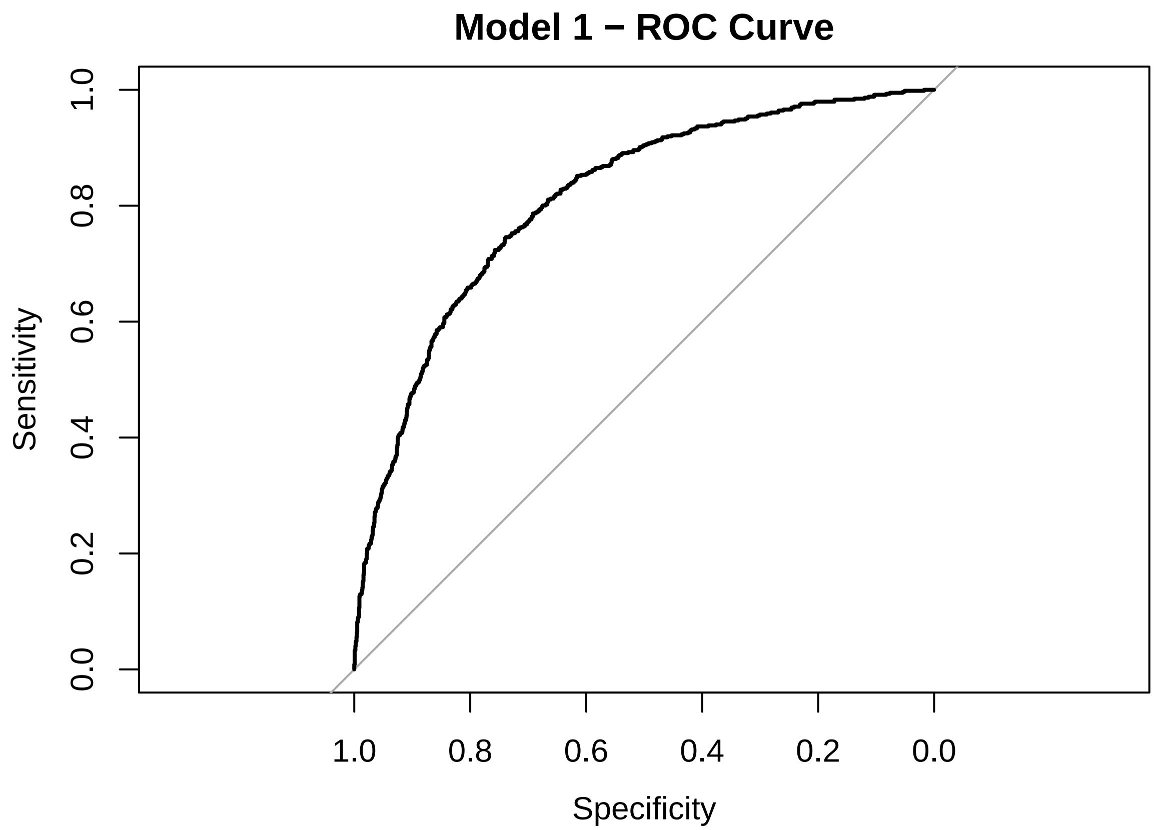
key model statistics measurements

```r
# Model1
predict_1 <- predict(TARGET_FLAG_m1, newdata=testing_1, type='response')
glm.pred1 = ifelse(predict_1 > 0.5, 1, 0)
cM1 <- confusionMatrix(glm.pred1, testing_1$TARGET_FLAG, positive = "1")

# Model2
predict_2 <- predict(TARGET_FLAG_m2, newdata=testing_1, type='response')
glm.pred2 = ifelse(predict_2 > 0.5, 1, 0)
cM2 <- confusionMatrix(glm.pred2, testing_1$TARGET_FLAG, positive = "1")

# Model3
predict_3 <- predict(TARGET_FLAG_m3,newdata=testing_1,type='response')
glm.pred3 = ifelse(predict_3 > 0.5, 1, 0)
cM3 <- confusionMatrix(glm.pred3, testing_1$TARGET_FLAG, positive = "1")
```

	Model1	Model2	Model3
Accuracy	0.7762367	0.7753121	0.7762367
Kappa	0.3426382	0.3390993	0.3351919
AccuracyLower	0.7580741	0.7571250	0.7580741
AccuracyUpper	0.7936478	0.7927501	0.7936478
AccuracyNull	0.7290800	0.7290800	0.7290800
AccuracyPValue	0.0000003	0.0000005	0.0000003
McnemarPValue	0.0000000	0.0000000	0.0000000
Sensitivity	0.3668942	0.3634812	0.3515358
Specificity	0.9283450	0.9283450	0.9340520
Pos Pred Value	0.6554878	0.6533742	0.6645161
Neg Pred Value	0.7978202	0.7969516	0.7949271
Prevalence	0.2709200	0.2709200	0.2709200
Detection Rate	0.0993990	0.0984743	0.0952381
Detection Prevalence	0.1516412	0.1507166	0.1433195
Balanced Accuracy	0.6476196	0.6459131	0.6427939

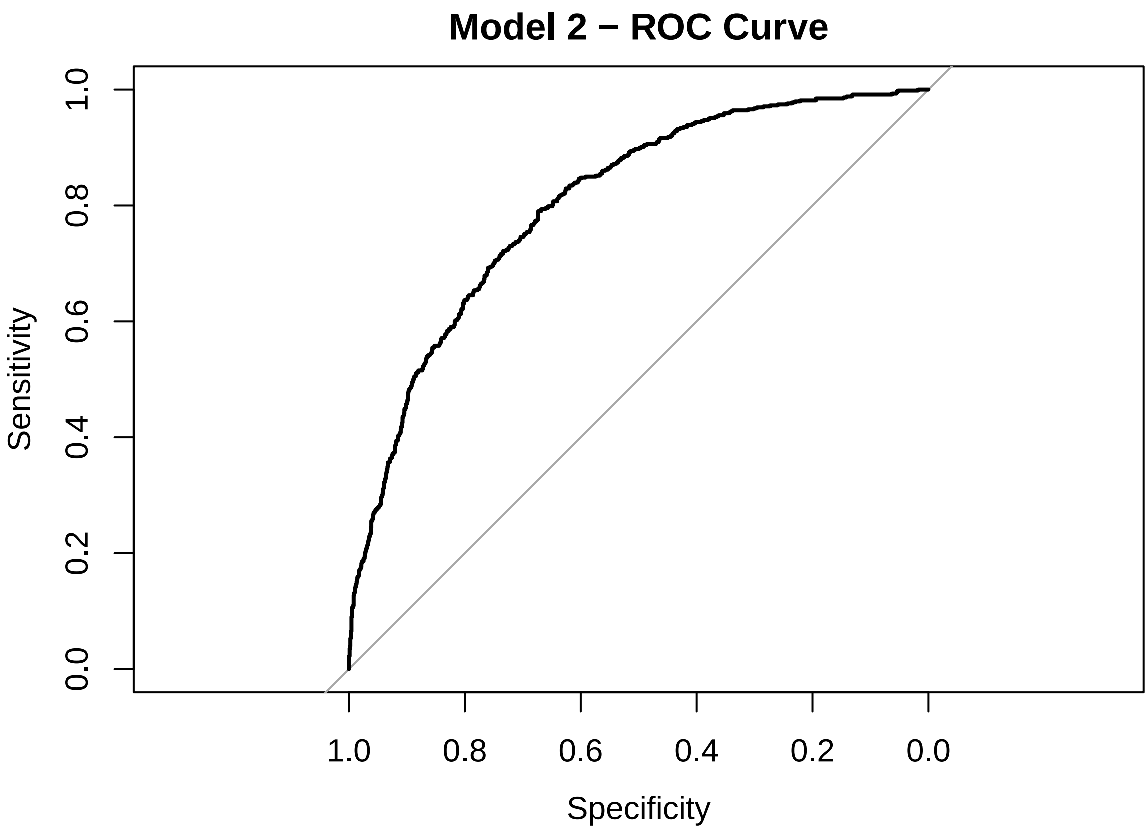
ROC Curve and Area Under the Curve

```r
rc1 <- roc(factor(TARGET_FLAG) ~ predict_1, data=testing_1)
rc2 <- roc(factor(TARGET_FLAG) ~ predict_2, data=testing_1)
rc3 <- roc(factor(TARGET_FLAG) ~ predict_3, data=testing_1)

plot(rc1,main='Model 1 - ROC Curve')
```



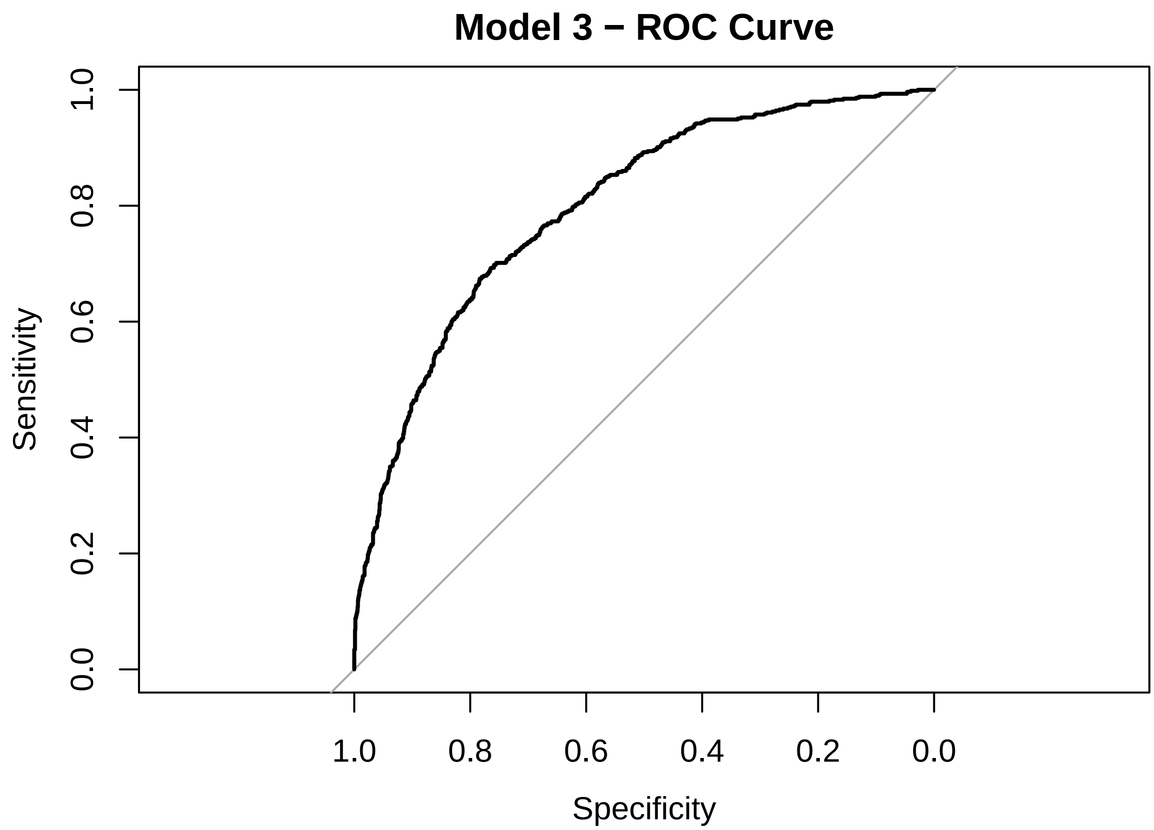

```
##
## Call:
## roc.formula(formula = factor(TARGET_FLAG) ~ predict_1, data = testing_1)
##
## Data: predict_1 in 1577 controls (factor(TARGET_FLAG) 0) < 586 cases (factor(TARGET_FLAG) 1).
## Area under the curve: 0.8108
```

```
plot(rc2,main='Model 2 - ROC Curve')
```




```
##
## Call:
## roc.formula(formula = factor(TARGET_FLAG) ~ predict_2, data = testing_1)
##
## Data: predict_2 in 1577 controls (factor(TARGET_FLAG) 0) < 586 cases (factor(TARGET_FLAG) 1).
## Area under the curve: 0.8028
```

```
plot(rc3,main='Model 3 - ROC Curve')
```




```
##
## Call:
## roc.formula(formula = factor(TARGET_FLAG) ~ predict_3, data = testing_1)
##
## Data: predict_3 in 1577 controls (factor(TARGET_FLAG) 0) < 586 cases (factor(TARGET_FLAG) 1).
## Area under the curve: 0.7986
```

```r
model <- c('Model 1', 'Model 2', 'Model 3')
area <- c(auc(rc1),auc(rc2),auc(rc3))
df <- data.frame(Model=model,AUC=area)
kable(df,caption='Area under the curve')
```

Table 2: Area under the curve

Model	AUC
Model 1	0.8108161
Model 2	0.8028366
Model 3	0.7986023

Log-likelihood/AIC/BIC

```r
# Log-likelihood
LL.1 <- logLik(TARGET_FLAG_m1)
LL.2 <- logLik(TARGET_FLAG_m2)
LL.3 <- logLik(TARGET_FLAG_m3)
LL <- rbind(LL.1, LL.2, LL.3) %>% round(2)
```

```r
# Akaike Information Criterion
AIC.1 <- AIC(TARGET_FLAG_m1)
AIC.2 <- AIC(TARGET_FLAG_m2)
AIC.3 <- AIC(TARGET_FLAG_m3)
AIC <- rbind(AIC.1, AIC.2, AIC.3) %>% round(2)
```

```r
# BIC
BIC.1 <- BIC(TARGET_FLAG_m1)
BIC.2 <- BIC(TARGET_FLAG_m2)
BIC.3 <- BIC(TARGET_FLAG_m3)
BIC <- rbind(BIC.1, BIC.2, BIC.3) %>% round(2)
```

```r
eval.table <- cbind(LL, AIC, BIC)

rownames(eval.table) <- c("Model 1", "Model 2", "Model 3")
colnames(eval.table) <- c("Log Likelihood", "AIC", "BIC")

kable(eval.table)
```

	Log Likelihood	AIC	BIC
Model 1	-2288.39	4638.78	4841.11
Model 2	-2309.50	4651.00	4755.44
Model 3	-2329.41	4688.82	4786.73

Checking variance inflation factors

```r
V1 <- vif(TARGET_FLAG_m1)
V2 <- vif(TARGET_FLAG_m2)
V3 <- vif(TARGET_FLAG_m3)
V1; V2; V3
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## KIDSDRIV    1.300978  1        1.140604
## HOMEKIDS    1.875707  1        1.369564
## YOJ         1.406953  1        1.186151
## PARENT1     1.907995  1        1.381302
## HOME_VAL    1.356386  1        1.164640
## MSTATUS     1.765394  1        1.328681
## SEX         2.309900  1        1.519836
## EDUCATION   7.393259  4        1.284117
```

```
## JOB        19.641345  8        1.204545
## TRAVTIME    1.037600  1        1.018626
## CAR_USE     2.460266  1        1.568523
## CAR_TYPE    3.658248  5        1.138485
## OLDCLAIM    1.177148  1        1.084965
## REVOKED     1.015057  1        1.007501
## MVR_PTS     1.165121  1        1.079408
## URBANICITY  1.160564  1        1.077295

##          YOJ    HOME_VAL     MSTATUS    TRAVTIME     CAR_USE     REVOKED
##     1.073966    1.264598    1.224222    1.034798    1.189132    1.005713
##   URBANICITY  KIDSDRIV_N  HOMEKIDS_N EDUCATION_N       JOB_N       TIF_N
##     1.154341    1.314100    1.351015    1.434041    1.592444    1.006717
##   CAR_TYPE_N  CLM_FREQ_N   MVR_PTS_N
##     1.043254    1.209704    1.170746

##   KIDSDRIV_N  HOMEKIDS_N         YOJ    HOME_VAL     MSTATUS EDUCATION_N
##     1.313487    1.351394    1.074426    1.265339    1.222378    1.437123
##        JOB_N       TIF_N     CAR_USE  CAR_TYPE_N    OLDCLAIM     REVOKED
##     1.596840    1.006400    1.192129    1.043165    1.175010    1.010591
##      MVR_PTS  URBANICITY
##     1.156696    1.108070
```

TARGET_AMT

Key model statistics results

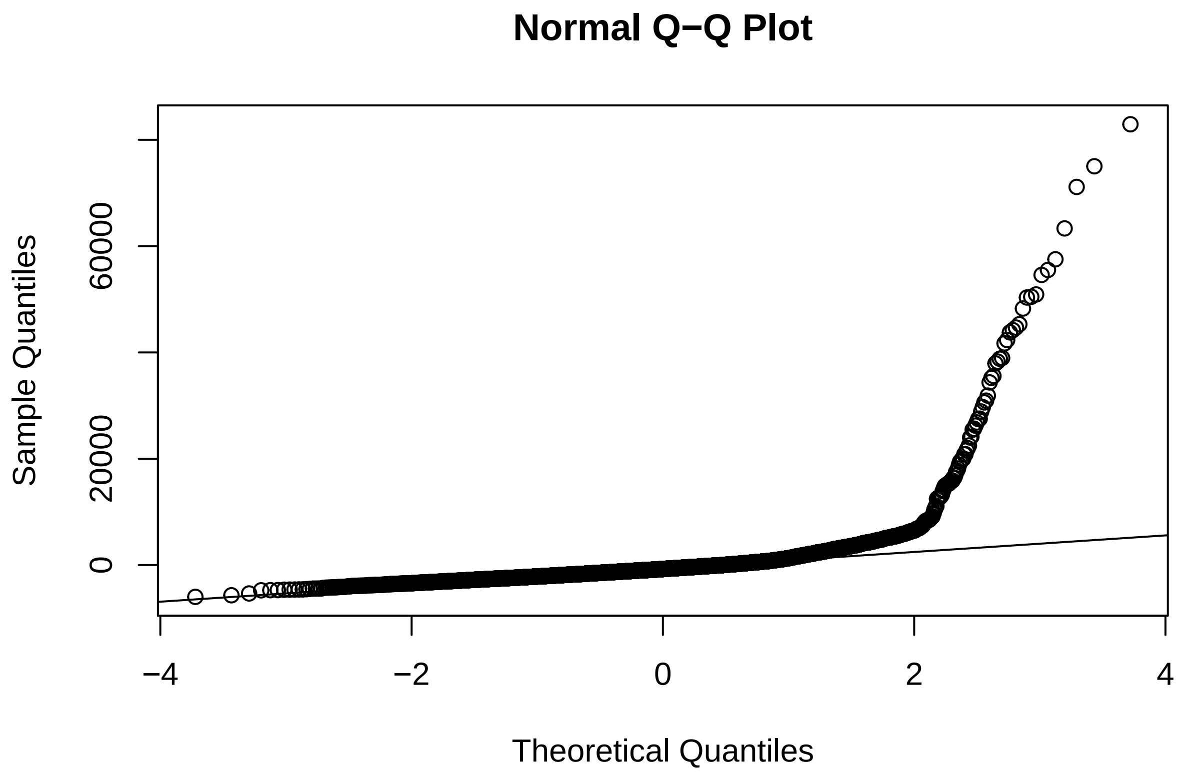
```
col_mdl_names <- c("Model 1", "Model 2", "Model 3")

#Calculate mean squared errors for each model
mse <- function(sm)
    mean(sm$residuals^2)

if(FALSE){
  cat("Mean Squared Error of Model 1:")
  mse(TARGET_AMT_m1)
  cat("Mean Squared Error of Model 2:")
  mse(TARGET_AMT_m2)
  cat("Mean Squared Error of Model 3:")
  mse(TARGET_AMT_m3)
}

col_mse <- c(mse(TARGET_AMT_m1),mse(TARGET_AMT_m2),mse(TARGET_AMT_m3))

# Calculate R^2 for each model:
if(FALSE){
  cat("R Squared of Model 1:")
  summary(TARGET_AMT_m1)$r.squared
  cat("R Squared of Model 2:")
  summary(TARGET_AMT_m2)$r.squared
  cat("R Squared of Model 3")
```

```
  summary(TARGET_AMT_m3)$r.squared
}

col_r_sq <- c(summary(TARGET_AMT_m1)$r.squared, summary(TARGET_AMT_m2)$r.squared, summary(TARGET_AMT_m3)

if(FALSE){
  cat("F-Stat of Model 1:")
  summary(aov(TARGET_AMT_m1))[[1]]$F[1]
  cat("F-Stat of Model 2:")
  summary(aov(TARGET_AMT_m2))[[1]]$F[1]
  cat("F-Stat of Model 3")
  summary(aov(TARGET_AMT_m1))[[1]]$F[1]
}

col_f_stat <- c(summary(aov(TARGET_AMT_m1))[[1]]$F[1], summary(aov(TARGET_AMT_m2))[[1]]$F[1], summary(ac
summary_df <- data.frame(cbind(col_mdl_names, col_mse, col_r_sq, col_f_stat))

colnames(summary_df) <- c("Model Name", "Mean Sq. Error", "R Squared", "F Stat")
kable(summary_df)
```

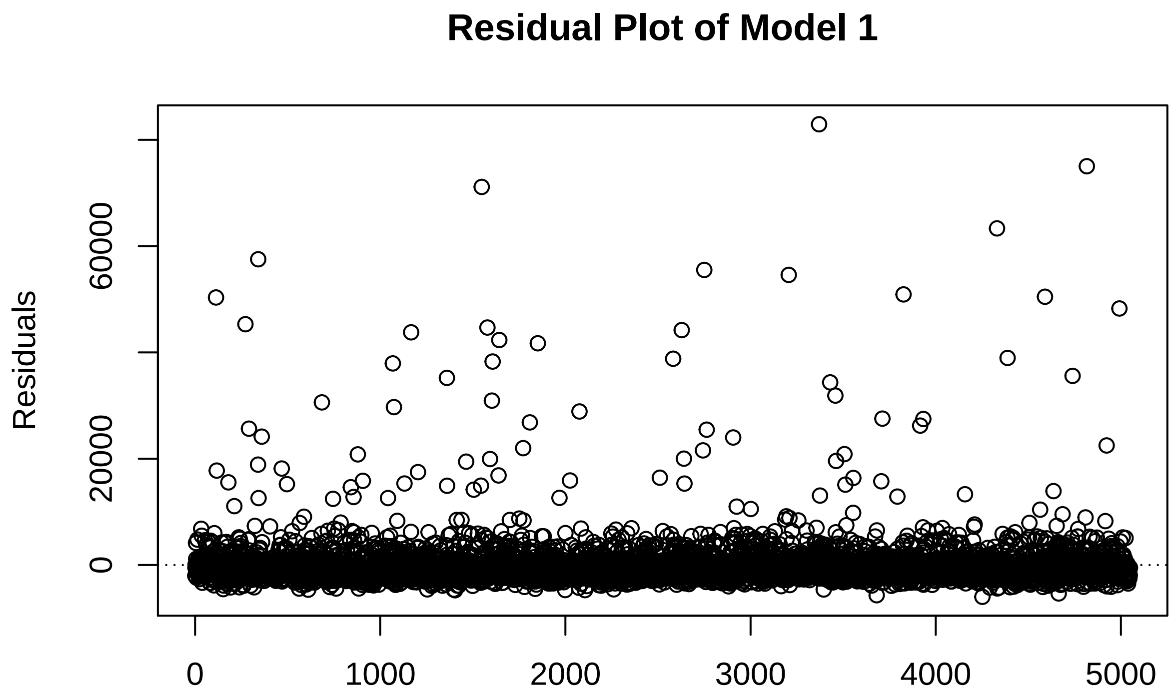
Model Name	Mean Sq. Error	R Squared	F Stat
Model 1	20520094.8851512	0.0750156588036813	20.9661466170798
Model 2	20733655.2432116	0.0653889982930091	38.8780437203011
Model 3	20615735.0702516	0.0707044860677789	39.09265849377

Residual plots for each model

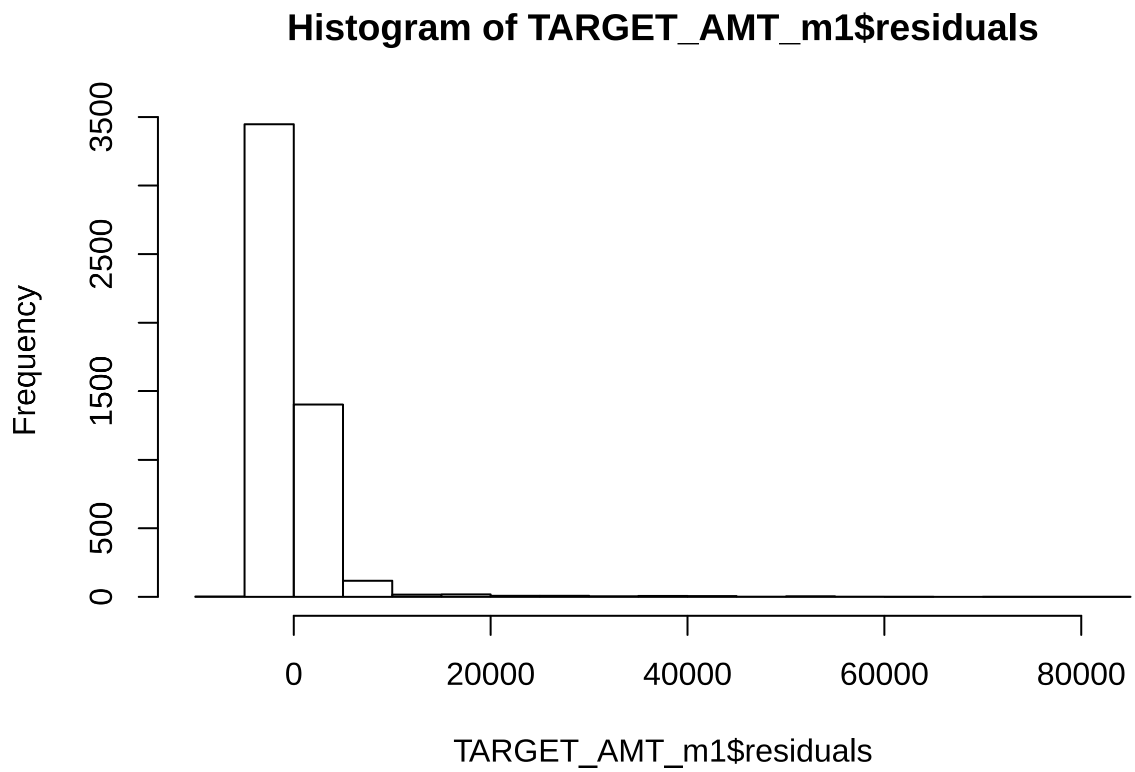
```
# Model 1
qqnorm(TARGET_AMT_m1$residuals)
qqline(TARGET_AMT_m1$residuals)
```



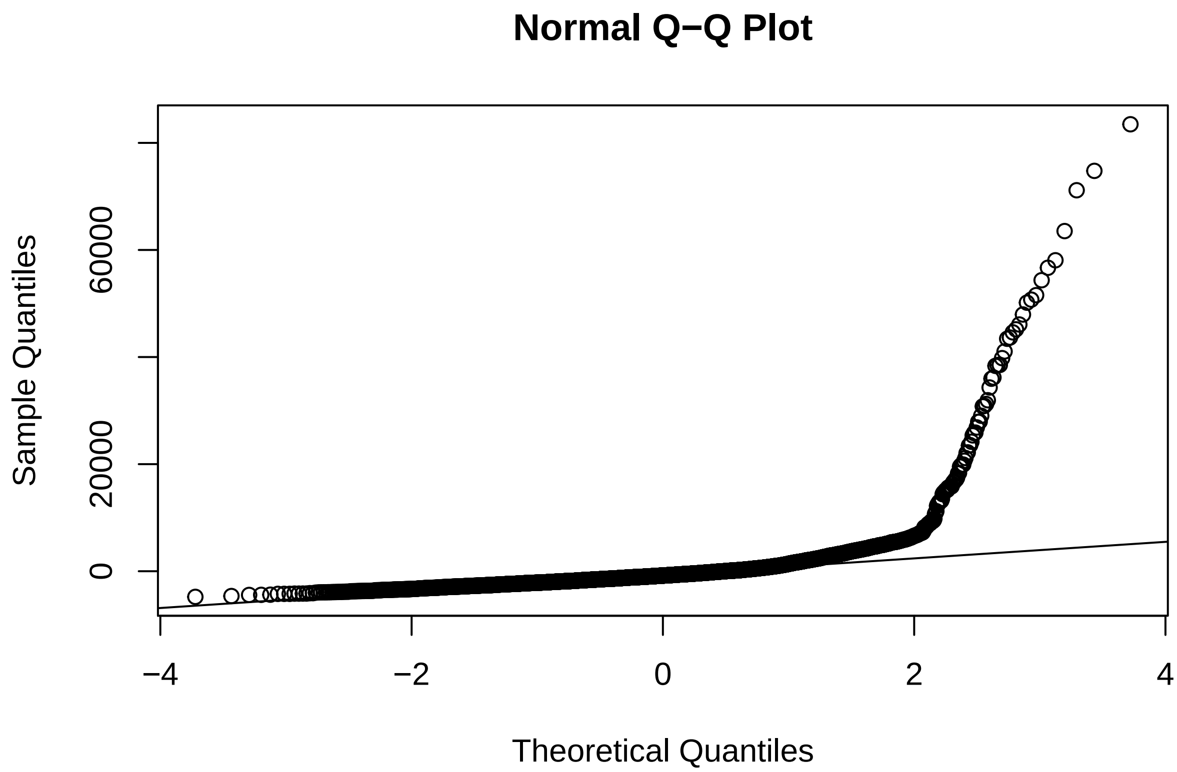

```r
INDEX <- seq(1,length(training_2$TARGET_AMT))
plot(TARGET_AMT_m1$residuals ~ INDEX,
     xlab='',
     ylab='Residuals',
     main='Residual Plot of Model 1')
abline(h=0,lty=3)
```



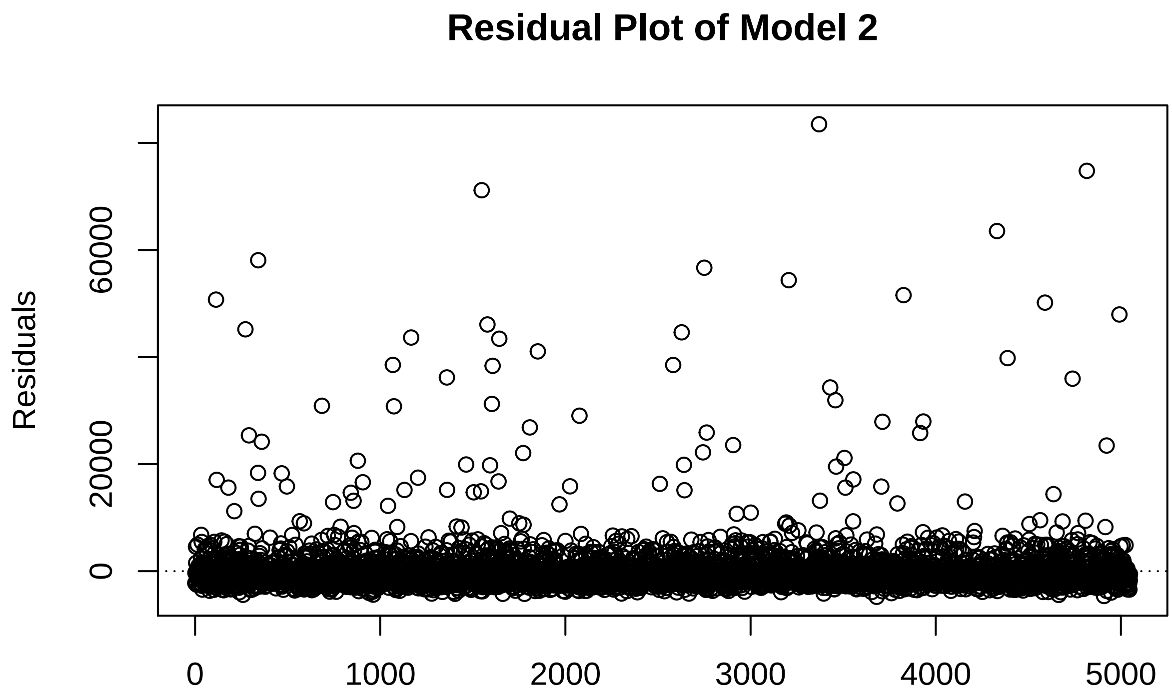

```r
hist(TARGET_AMT_m1$residuals)
```



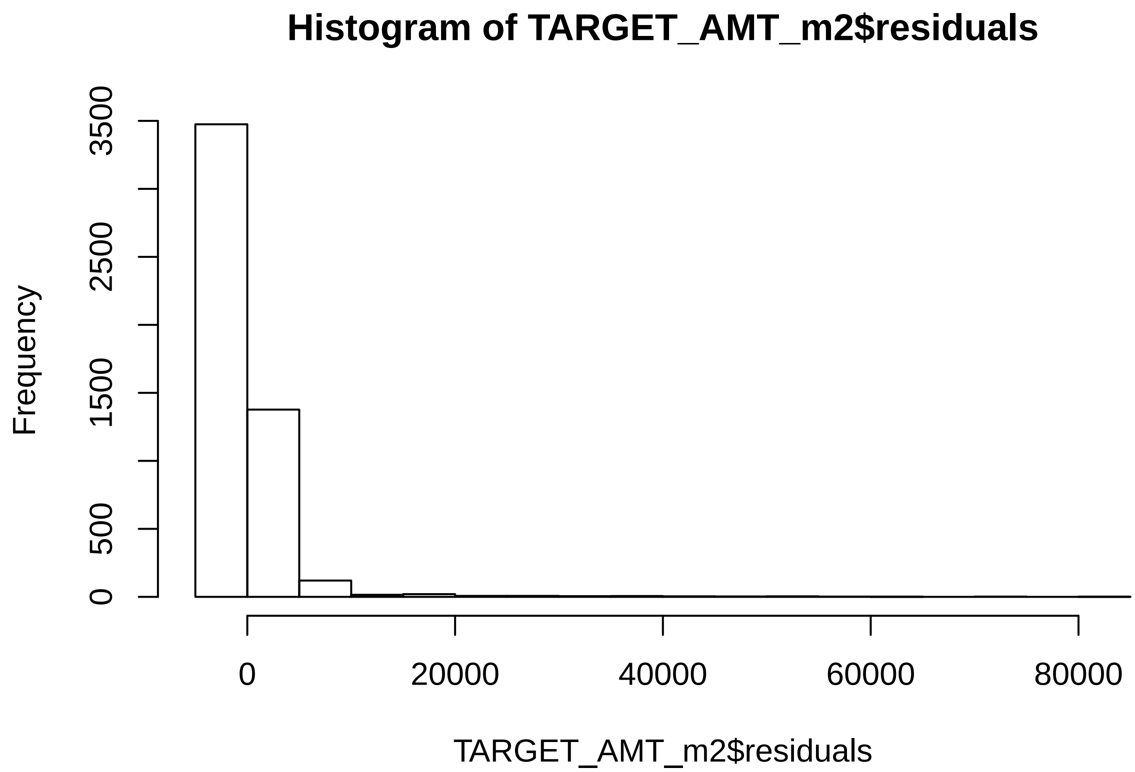

```r
# Model 2
qqnorm(TARGET_AMT_m2$residuals)
qqline(TARGET_AMT_m2$residuals)
```



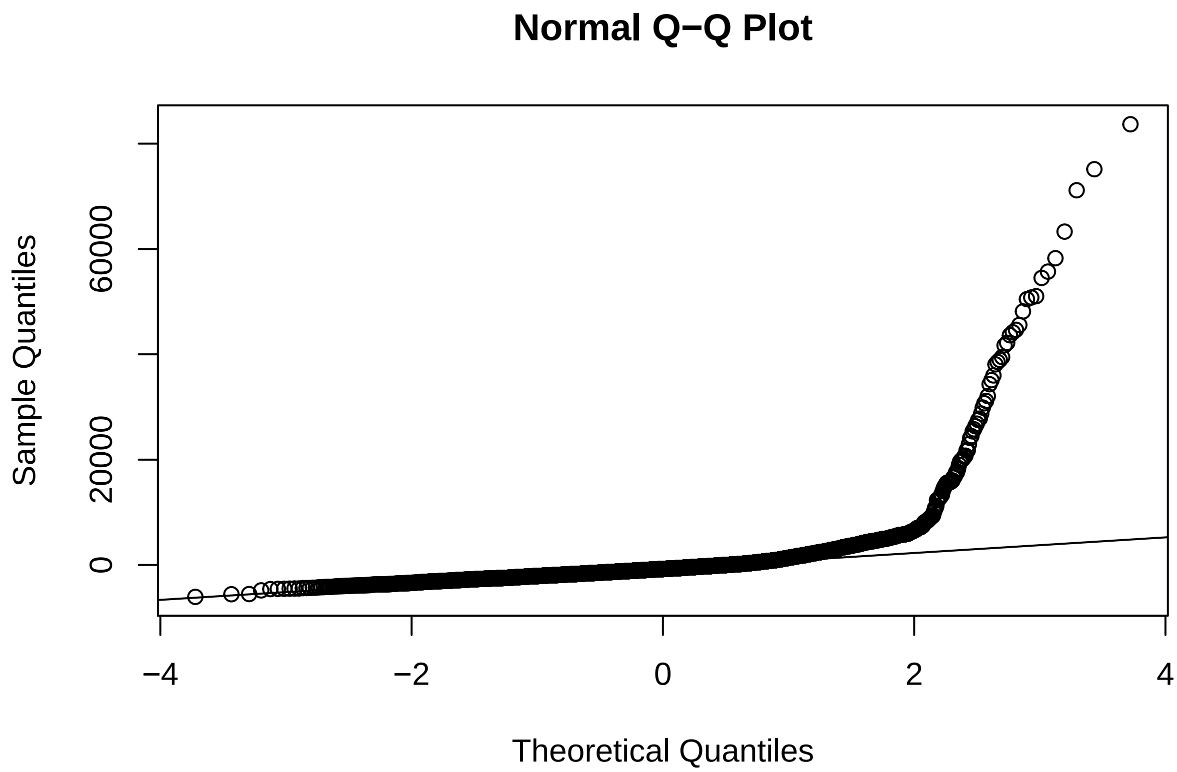

```
INDEX <- seq(1,length(training_2$TARGET_AMT))
plot(TARGET_AMT_m2$residuals ~ INDEX,
     xlab='',
     ylab='Residuals',
     main='Residual Plot of Model 2')
abline(h=0,lty=3)
```



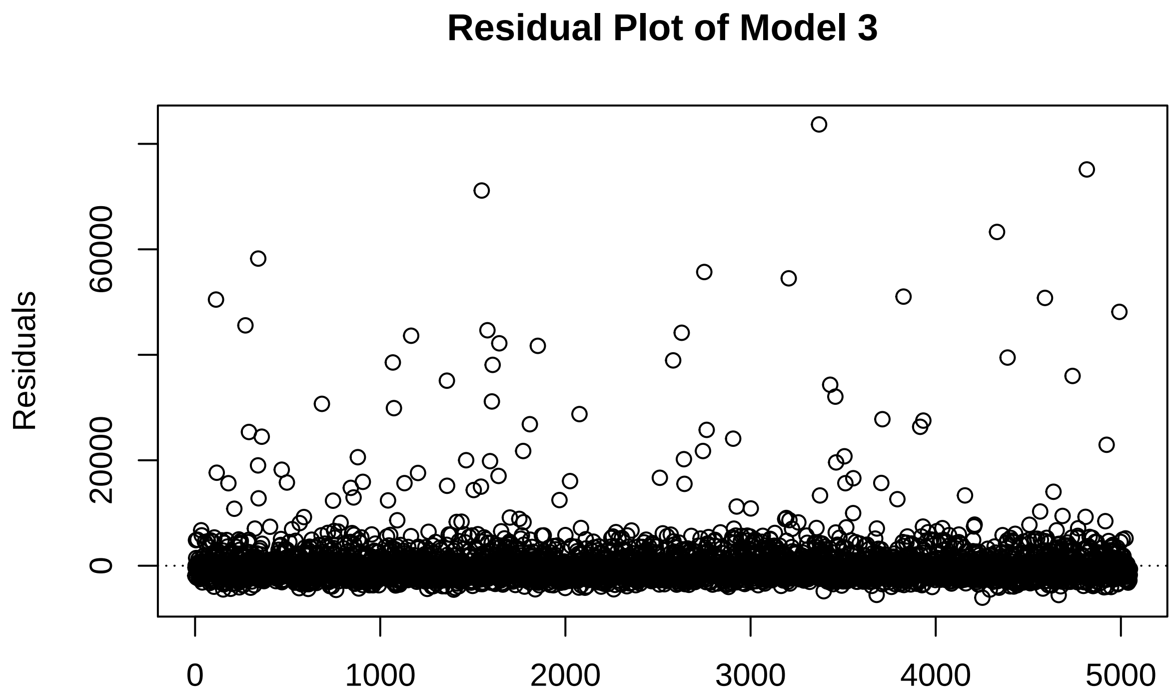

```r
hist(TARGET_AMT_m2$residuals)
```



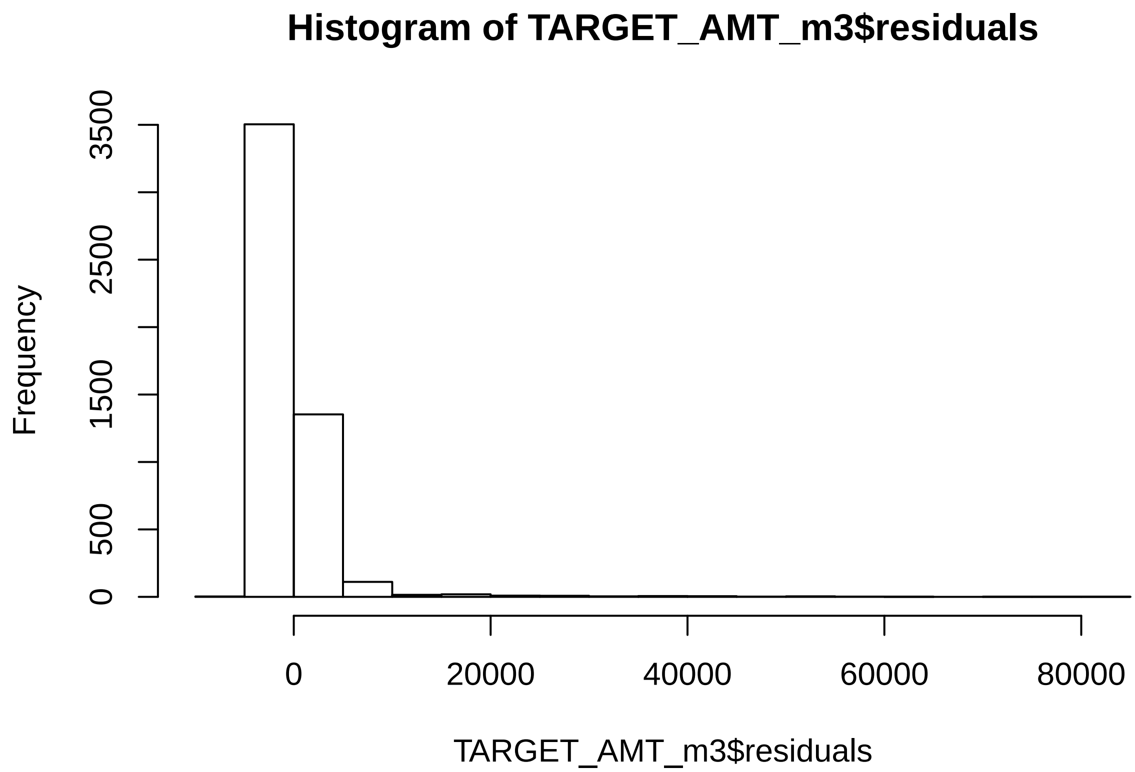

```
# Model 3
qqnorm(TARGET_AMT_m3$residuals)
qqline(TARGET_AMT_m3$residuals)
```




```r
INDEX <- seq(1,length(training_2$TARGET_AMT))
plot(TARGET_AMT_m3$residuals ~ INDEX,
     xlab='',
     ylab='Residuals',
     main='Residual Plot of Model 3')
abline(h=0,lty=3)
```

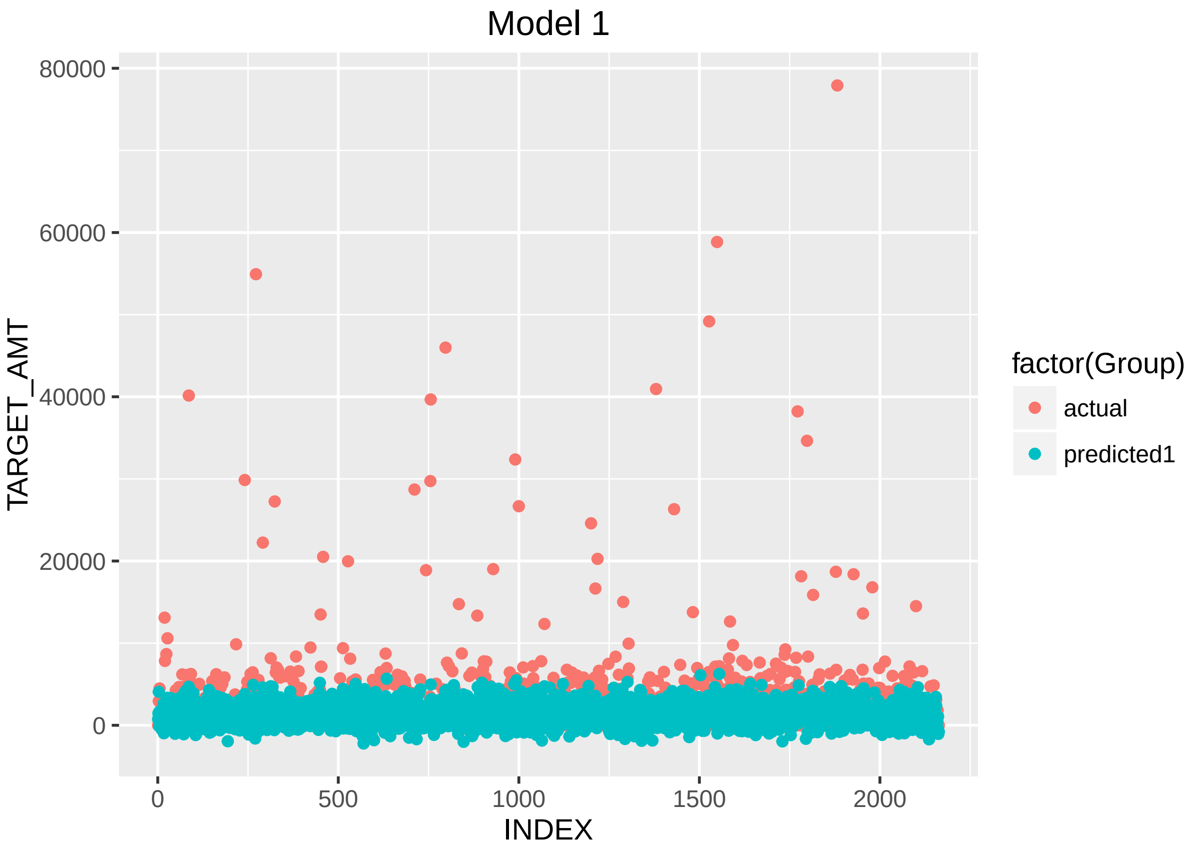


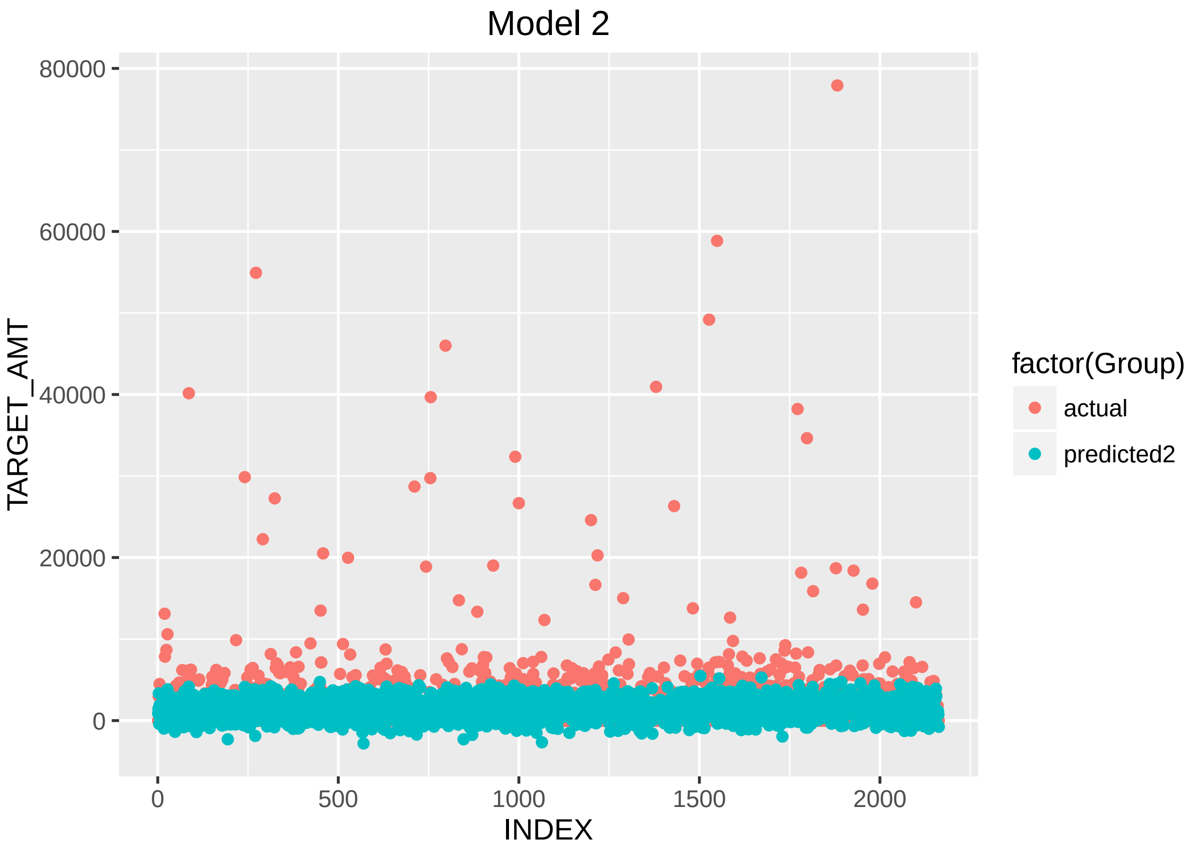

```
hist(TARGET_AMT_m3$residuals)
```

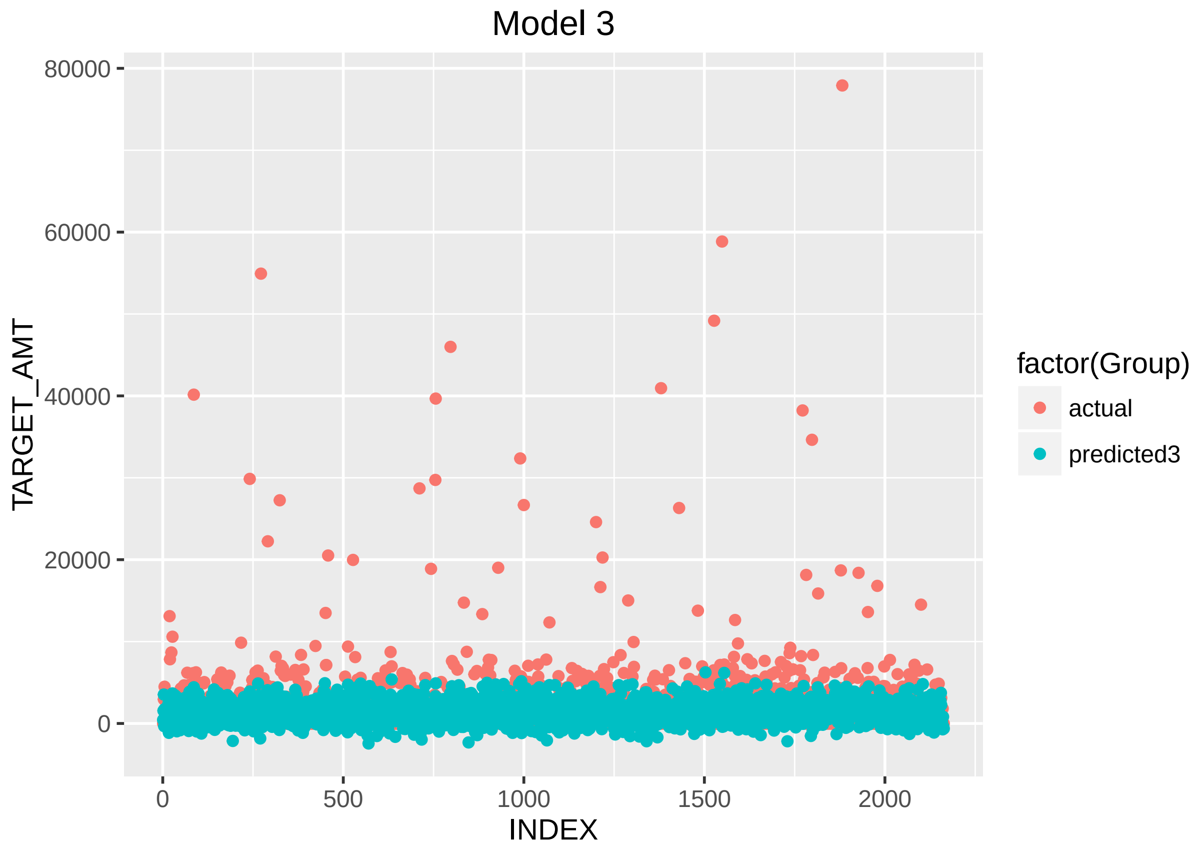


Using testing data to check the predicting result

```
actual <- testing_2$TARGET_AMT
new_data = select(testing_2, -TARGET_AMT)
predicted1 <- predict(TARGET_AMT_m1,newdata=new_data)
predicted2 <- predict(TARGET_AMT_m2,newdata=new_data)
predicted3 <- predict(TARGET_AMT_m3,newdata=new_data)

INDEX <- seq(1,length(testing_2$TARGET_AMT))
result1 <- data.frame(INDEX=INDEX, actual=actual, predicted1=predicted1)
result2 <- data.frame(INDEX=INDEX, actual=actual, predicted2=predicted2)
result3 <- data.frame(INDEX=INDEX, actual=actual, predicted3=predicted3)
result1 <- gather(result1,Group,TARGET_AMT,2:3)
result2 <- gather(result2,Group,TARGET_AMT,2:3)
result3 <- gather(result3,Group,TARGET_AMT,2:3)
p1 <- ggplot(data=result1,aes(INDEX,TARGET_AMT)) +
  geom_point(aes(colour = factor(Group))) + ggtitle('Model 1')
p2 <- ggplot(data=result2,aes(INDEX,TARGET_AMT)) +
  geom_point(aes(colour = factor(Group))) + ggtitle('Model 2')
p3 <- ggplot(data=result3,aes(INDEX,TARGET_AMT)) +
  geom_point(aes(colour = factor(Group))) + ggtitle('Model 3')
p1; p2; p3
```


Model 1
80000
60000
40000
20000
0
TARGET_AMT
0
500
1000
1500
2000
INDEX
factor(Group)
actual
predicted1


Model 2
80000
60000
40000
20000
0
TARGET_AMT
0
500
1000
1500
2000
INDEX
factor(Group)
actual
predicted2



Overall, for predicting `TARGET_FLAG` Model 2 handled **Multicollinearity** issuse appropriately. Meanwhile, Model 2 slightly stands out after considering Accuracy, Sensitivity, Specificity, AUC, Log-likelihood number, AIC and BIC.

For predicting `TARGET_AMT`, all of three models are not good enough in this case. More advanced modeling techniques might be helpful, although intution tells me that it would be very difficult to build a good model to predict `TARGET_AMT`. How much the cost would be in car accident is a highly random event.