Data 608 HW 4 LQ

Layla Quinones

11/10/2021

Libraries

```
library(tidyverse)
library(ggplot2)
library(VIM)
library(GGally)
library(caret)
library(broom)
library(naniar)
library(stringr)
```

EDA

```
# Load data
# Training
rawTrain <- read.csv("https://raw.githubusercontent.com/MsQCompSci/Data621Group4/main/HW4/insurance_tra
```

```
# check to see if we need to clean the data
glimpse(rawTrain)
```

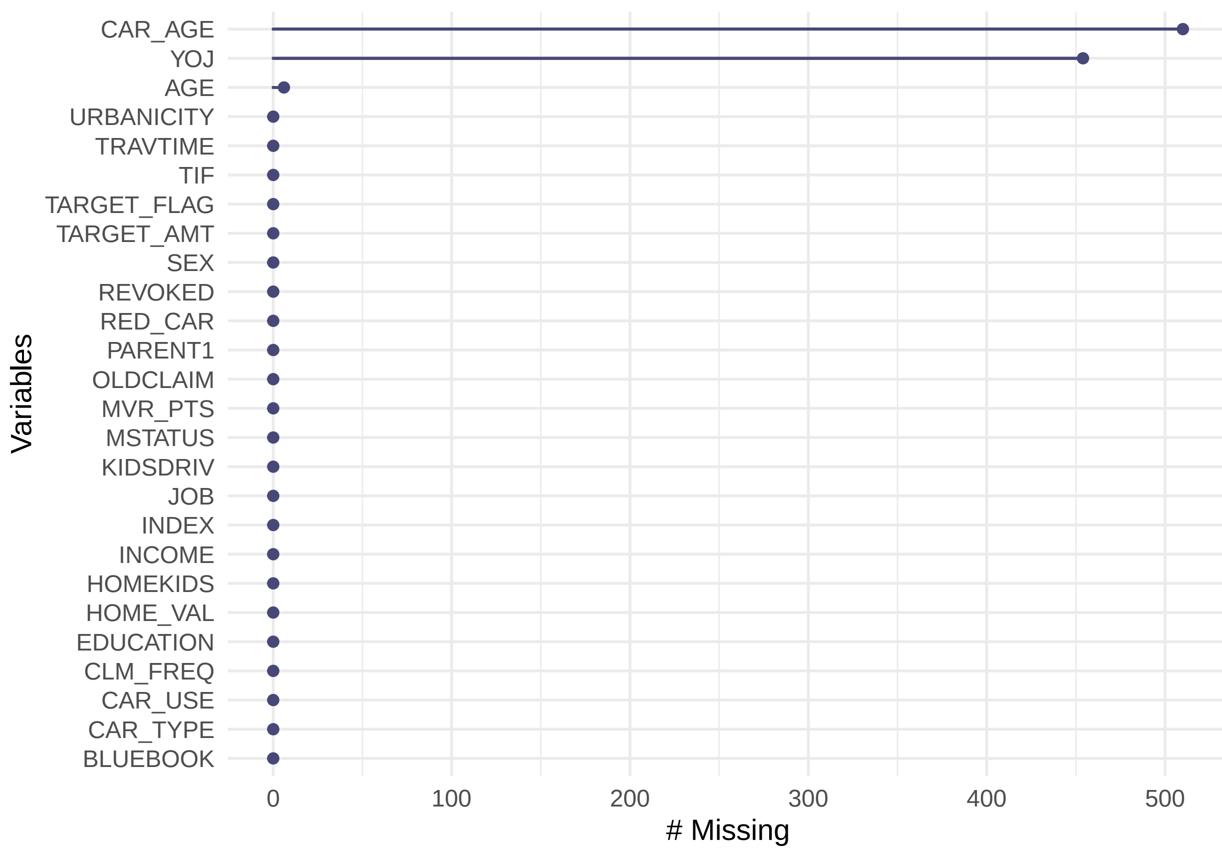
```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20...
## $ TARGET_FLAG <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0...
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 402...
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53,...
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2...
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0...
## $ INCOME      <chr> "$67,349", "$91,449", "$16,039", "", "$114,986", "$125,...
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", ...
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,925", "...
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes", "Ye...
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z_F", ...
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High School"...
## $ JOB         <chr> "Professional", "z_Blue Collar", "Clerical", "z_Blue Co...
```

```
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, ...
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Private...
## $ BLUEBOOK    <chr> "$14,230", "$14,940", "$4,010", "$15,440", "$18,000", "...
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, ...
## $ CAR_TYPE    <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV", "Spo...
## $ RED_CAR     <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes", "no...
## $ OLDCLAIM    <chr> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", "$0",...
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0...
## $ REVOKED     <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No",...
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, ...
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, ...
## $ URBANICITY  <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Highly U...
```

There are 8161 observations in this data set and 26 columns. We know that `INDEX`, `TARGET_FLAG` and `TARGET_AMT` are not predictor variables. This gives us 8161 observations with 23 predictors that are a combination of int, double and character data types. We also see that the character variables will have to converted to factors in order for us to explore their distributions. Variables such and `INCOME`, `HOME_VAL`, `BLUEBOOK`, `OLDCLAIM` will be converted to numeric because they are numbers with values that have meaning in their heirarchy.

Missing Values

```r
#plot missing values using VIM package
gg_miss_var(rawTrain)
```



There are missing variables in the columns `Car_AGE`, `AGE` and `YOJ`. None of these exceed the 10% missing data so we will continue with all variables for noe (not dropping any of them due to missing data)

DATA CLEANING - CONVERTING DATA TYPES

```
#lets remove the $ and , and put in a different variable name from numeric strings
rawTrain <- rawTrain %>%
  mutate(INCOME = gsub("\\$", "", INCOME),     #Remove $
         HOME_VAL = gsub("\\$", "", HOME_VAL),
         BLUEBOOK = gsub("\\$", "", BLUEBOOK),
         OLDCLAIM = gsub("\\$", "", OLDCLAIM),
         MSTATUS = gsub("z_", "", MSTATUS),
         SEX = gsub("z_", "", SEX),
         EDUCATION= gsub("z_", "", EDUCATION),
         JOB= gsub("z_", "", JOB),
         CAR_TYPE= gsub("z_", "", CAR_TYPE),
         URBANICITY= gsub("z_", "", URBANICITY),
         INCOME = as.numeric(gsub(",", "", INCOME)),     #remove , and cast to numeric
         HOME_VAL = as.numeric(gsub(",", "", HOME_VAL)),
         BLUEBOOK = as.numeric(gsub(",", "", BLUEBOOK)),
         OLDCLAIM = as.numeric(gsub(",", "", OLDCLAIM)),
         TARGET_FLAG = as.factor(TARGET_FLAG))

#lets also change all other character variables into factors
rawTrain[sapply(rawTrain, is.character)] <- lapply(rawTrain[sapply(rawTrain, is.character)],
                                                   as.factor)

#display summary statistics again to confirm
summary(rawTrain)
```

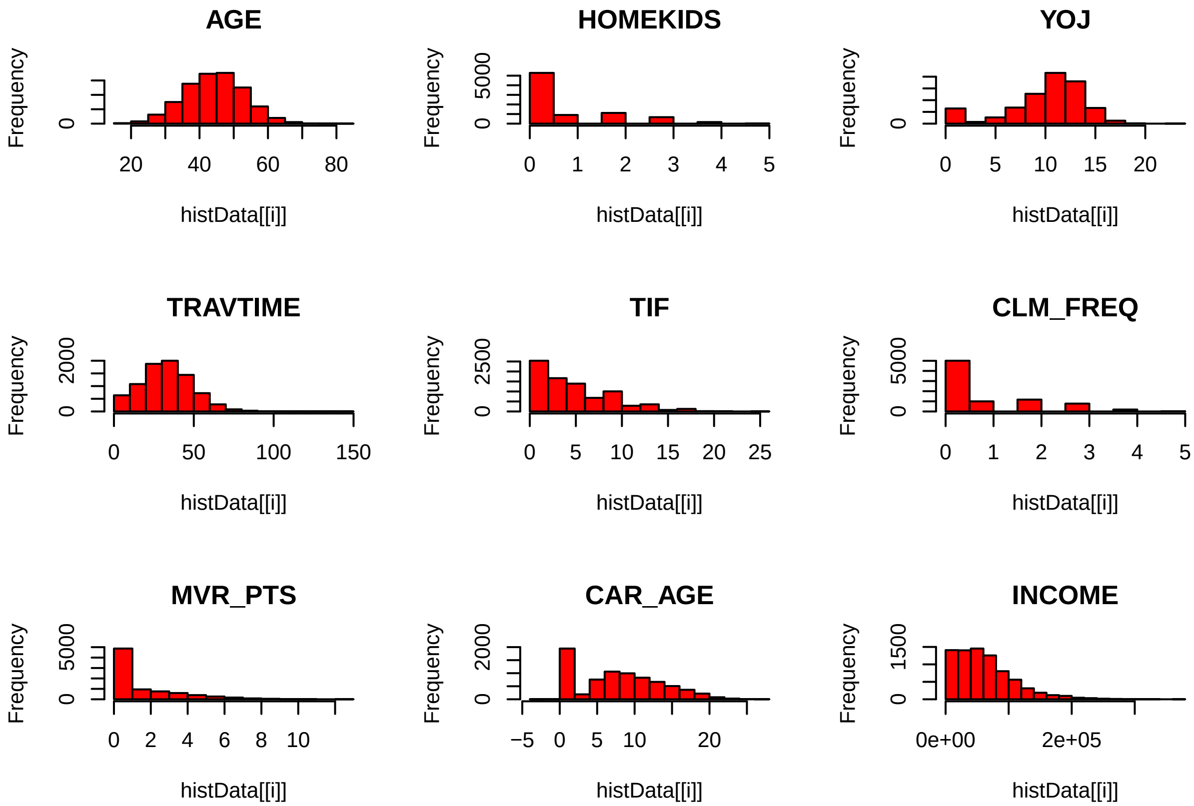
```
##      INDEX       TARGET_FLAG   TARGET_AMT        KIDSDRIV           AGE
##  Min.   :    1   0:6008      Min.   :     0   Min.   :0.0000   Min.   :16.00
##  1st Qu.: 2559   1:2153      1st Qu.:     0   1st Qu.:0.0000   1st Qu.:39.00
##  Median : 5133               Median :     0   Median :0.0000   Median :45.00
##  Mean   : 5152               Mean   :  1504   Mean   :0.1711   Mean   :44.79
##  3rd Qu.: 7745               3rd Qu.:  1036   3rd Qu.:0.0000   3rd Qu.:51.00
##  Max.   :10302               Max.   :107586   Max.   :4.0000   Max.   :81.00
##                                                                NA's   :6
##     HOMEKIDS           YOJ          INCOME        PARENT1       HOME_VAL
##  Min.   :0.0000   Min.   : 0.0   Min.   :     0   No :7084   Min.   :     0
##  1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097   Yes:1077   1st Qu.:     0
##  Median :0.0000   Median :11.0   Median : 54028              Median :161160
##  Mean   :0.7212   Mean   :10.5   Mean   : 61898              Mean   :154867
##  3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986              3rd Qu.:238724
##  Max.   :5.0000   Max.   :23.0   Max.   :367030              Max.   :885282
##                   NA's   :454    NA's   :445                 NA's   :464
##  MSTATUS    SEX            EDUCATION             JOB          TRAVTIME
##  No :3267   F:4375   <High School:1203   Blue Collar :1825   Min.   :  5.00
##  Yes:4894   M:3786   Bachelors   :2242   Clerical    :1271   1st Qu.: 22.00
##                      High School :2330   Professional:1117   Median : 33.00
##                      Masters     :1658   Manager     : 988   Mean   : 33.49
##                      PhD         : 728   Lawyer      : 835   3rd Qu.: 44.00
```

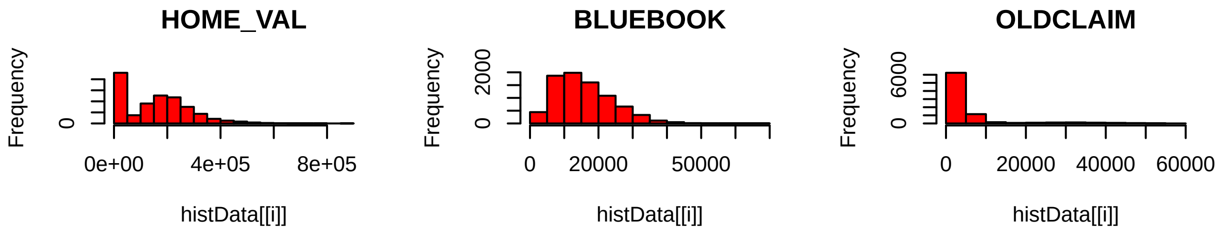
```
##                                           Student     : 712   Max.   :142.00
##                                           (Other)     :1413
##       CAR_USE        BLUEBOOK          TIF              CAR_TYPE
##  Commercial:3029   Min.   : 1500   Min.   : 1.000   Minivan    :2145
##  Private   :5132   1st Qu.: 9280   1st Qu.: 1.000   Panel Truck: 676
##                    Median :14440   Median : 4.000   Pickup     :1389
##                    Mean   :15710   Mean   : 5.351   Sports Car : 907
##                    3rd Qu.:20850   3rd Qu.: 7.000   SUV        :2294
##                    Max.   :69740   Max.   :25.000   Van        : 750
##
##  RED_CAR       OLDCLAIM        CLM_FREQ      REVOKED       MVR_PTS
##  no :5783   Min.   :    0   Min.   :0.0000   No :7161   Min.   : 0.000
##  yes:2378   1st Qu.:    0   1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000
##             Median :    0   Median :0.0000              Median : 1.000
##             Mean   : 4037   Mean   :0.7986              Mean   : 1.696
##             3rd Qu.: 4636   3rd Qu.:2.0000              3rd Qu.: 3.000
##             Max.   :57037   Max.   :5.0000              Max.   :13.000
##
##     CAR_AGE                     URBANICITY
##  Min.   :-3.000   Highly Rural/ Rural:1669
##  1st Qu.: 1.000   Highly Urban/ Urban:6492
##  Median : 8.000
##  Mean   : 8.328
##  3rd Qu.:12.000
##  Max.   :28.000
##  NA's   :510
```

We get a better sense of the information available in each variable now with the data type change.

```r
#histagrams for only the numerical data
histData <- rawTrain %>%
  select(AGE, HOMEKIDS, YOJ,TRAVTIME, TIF, CLM_FREQ, MVR_PTS, CAR_AGE, INCOME, HOME_VAL, BLUEBOOK, OLDCI

par(mfrow = c(3,3))
for(i in 1:ncol(histData)) {#distribution of each variable
  hist(histData[[i]], main = colnames(histData[i]), col = "red")
}
```


AGE
HOMEKIDS
YOJ
TRAVTIME
TIF
CLM_FREQ
MVR_PTS
CAR_AGE
INCOME
Frequency
histData[[i]]

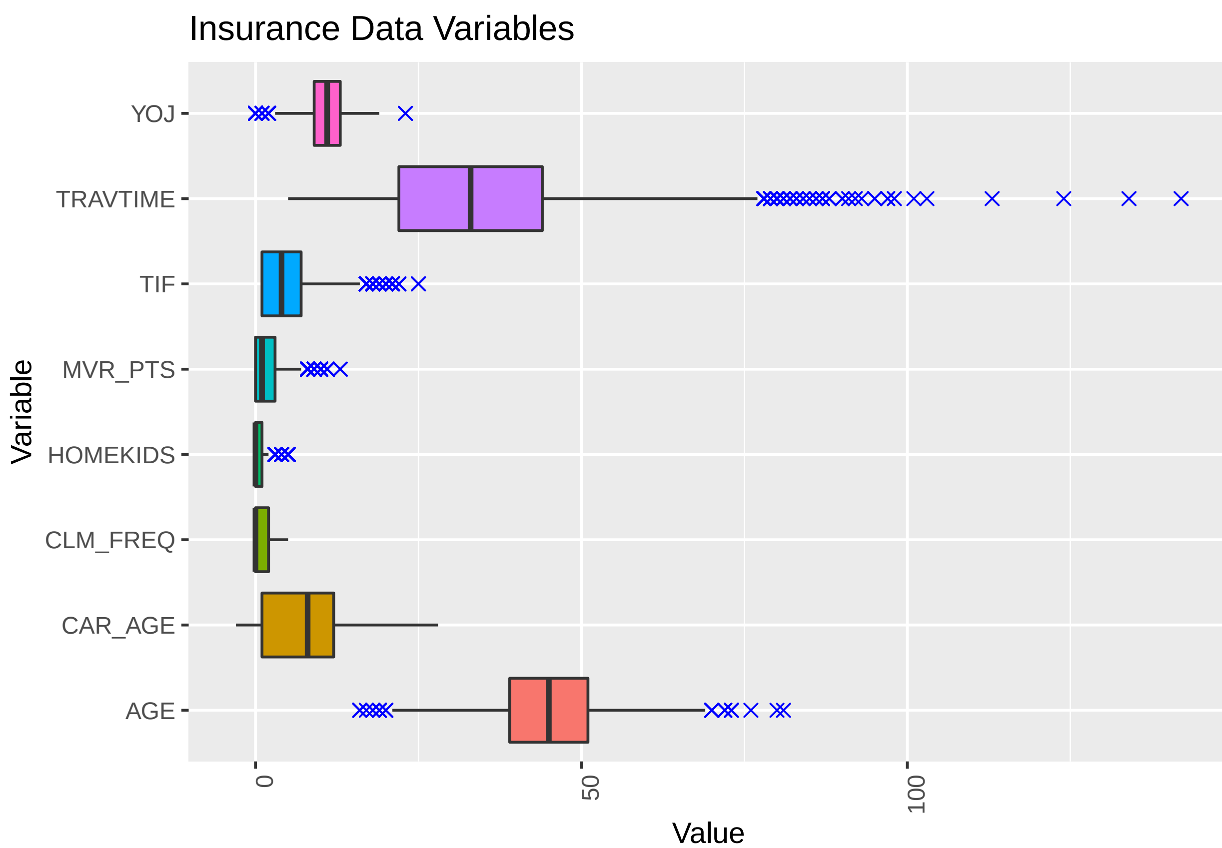


From the above histagrams of numerical data we can see that mose numerical variables have a right skew which may indicate that a transformation will be helpful for these variables.

```
longData <- histData %>%
  select(-HOME_VAL, -INCOME, -BLUEBOOK, -OLDCLAIM) %>%  # remove this for scale issue will plot below
  gather(key = Variable, value = Value)

# generate boxplot to identify outliers
ggplot(longData, aes(Variable, Value, fill = Variable)) +geom_boxplot(outlier.colour="blue",
               outlier.shape=4,
               outlier.size=2,
               show.legend=FALSE) +
   theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
   coord_flip()+
  labs(title="Insurance Data Variables", y="Value")
```

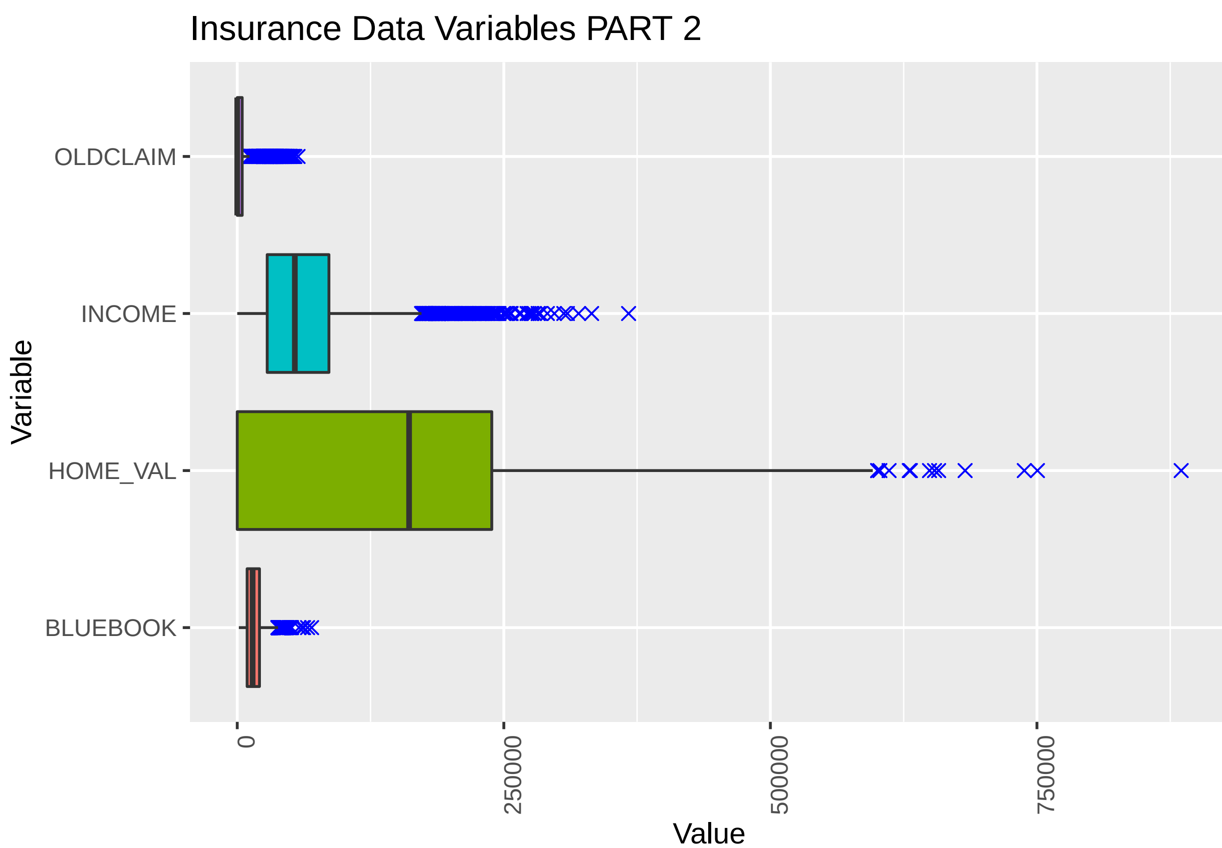
```
## Warning: Removed 970 rows containing non-finite values (stat_boxplot).
```




```r
longData2 <- histData %>%
  select(HOME_VAL, INCOME, BLUEBOOK, OLDCLAIM) %>%  # remove this for scale issue will plot below
  gather(key = Variable, value = Value)

# generate boxplot to identify outliers
ggplot(longData2, aes(Variable, Value, fill = Variable)) +geom_boxplot(outlier.colour="blue",
               outlier.shape=4,
               outlier.size=2,
               show.legend=FALSE) +
   theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
   coord_flip()+
  labs(title="Insurance Data Variables PART 2", y="Value")
```

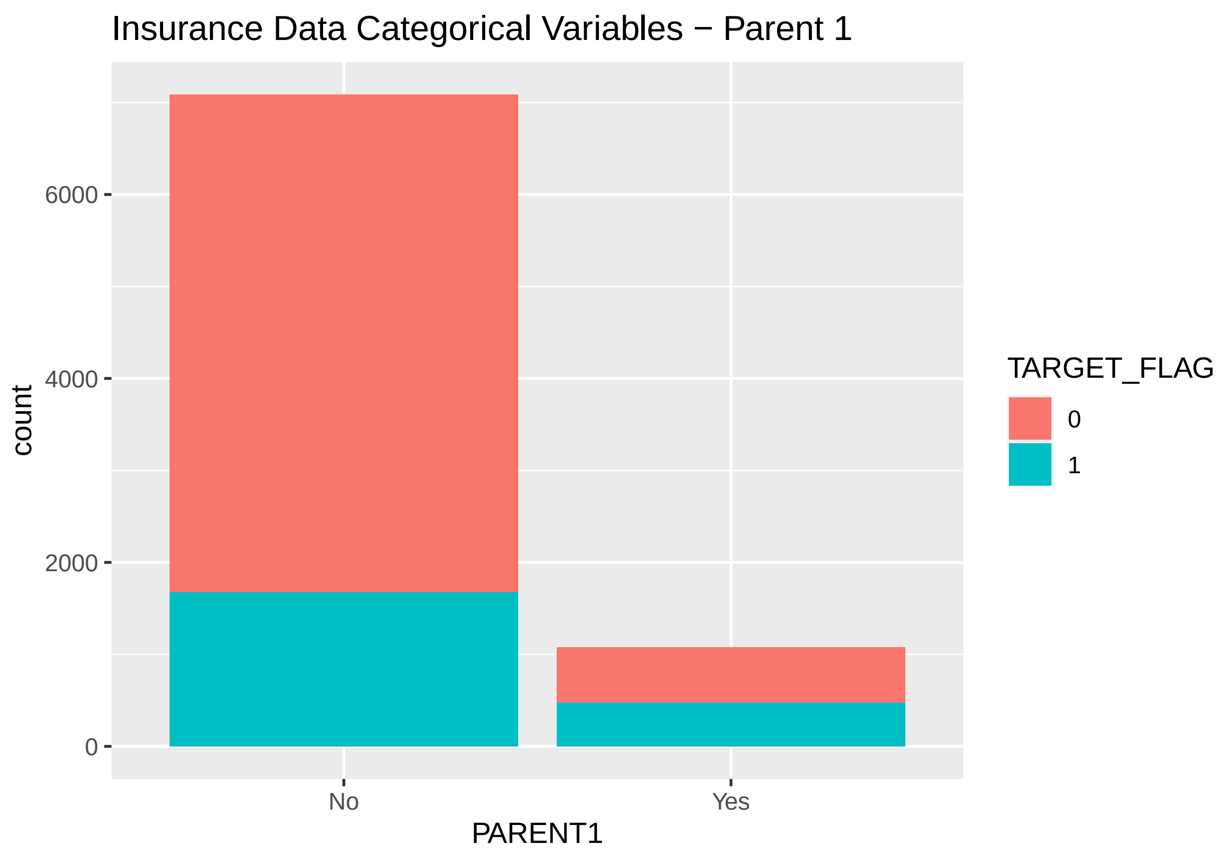
```
## Warning: Removed 909 rows containing non-finite values (stat_boxplot).
```



From these initial box plots we can see that there are outliers specifically `TRAVTIME`, `INCOME`, `HOME_VAL` has many outliers more spread out compared to the other variables.

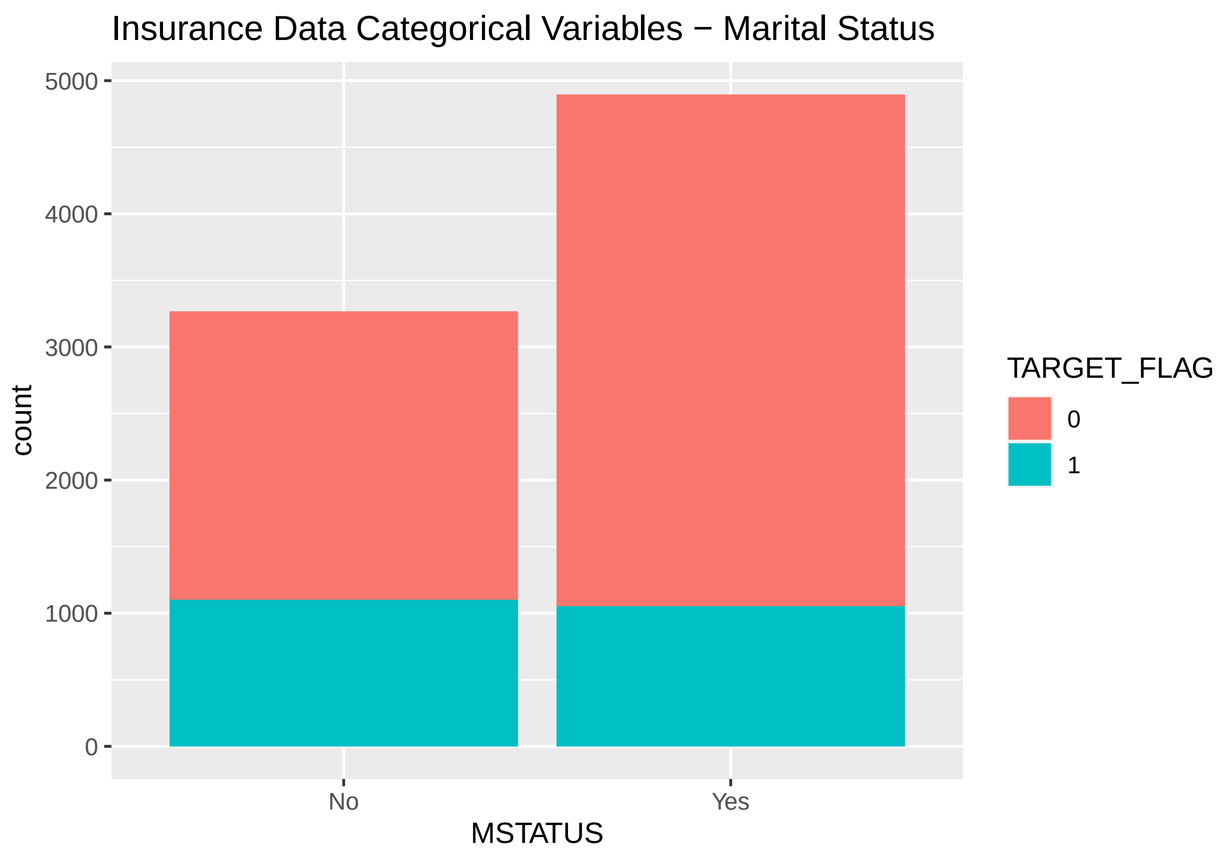
Categorical Predictors - with target variable

```
#plot
ggplot(rawTrain, aes(x = PARENT1, fill = TARGET_FLAG)) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - Parent 1")
```



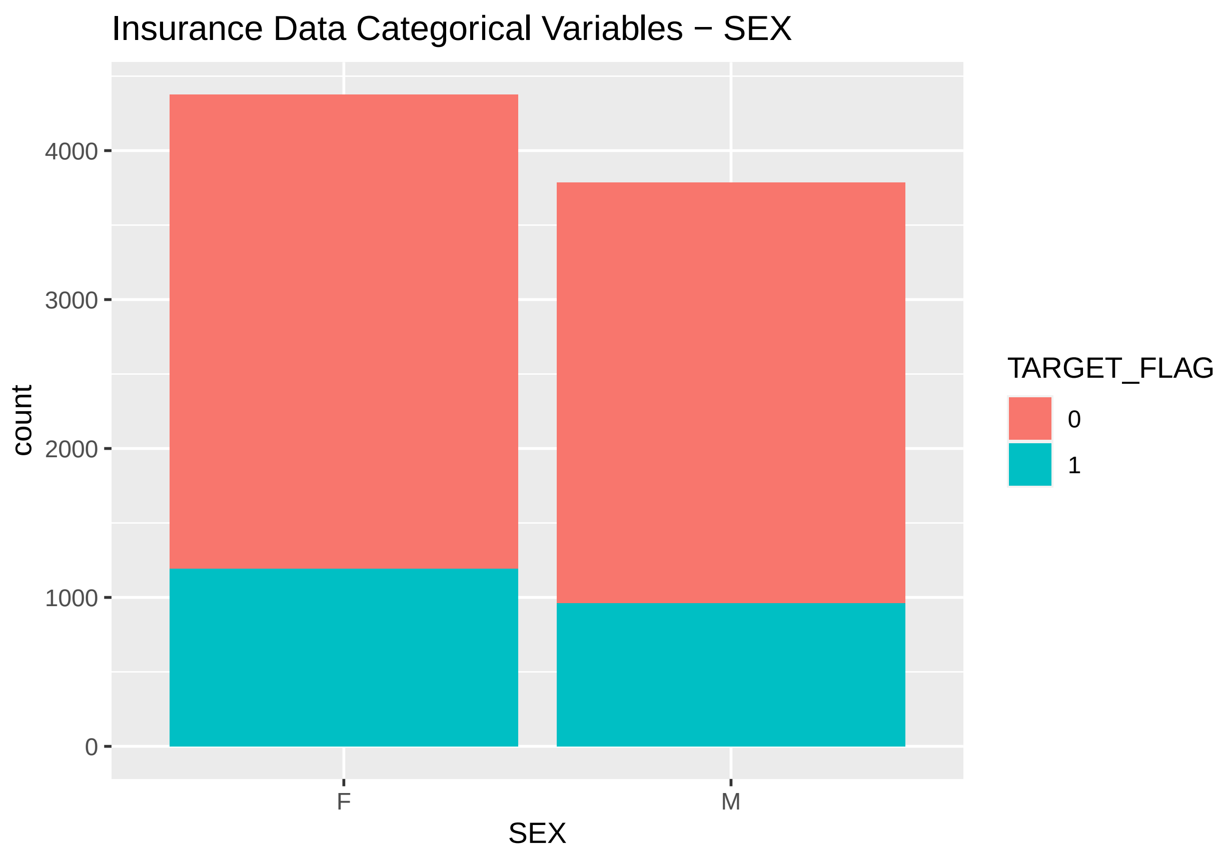

```
#imbalanced here
```

```
ggplot(rawTrain, aes(x = MSTATUS, fill = TARGET_FLAG)) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - Marital Status")
```



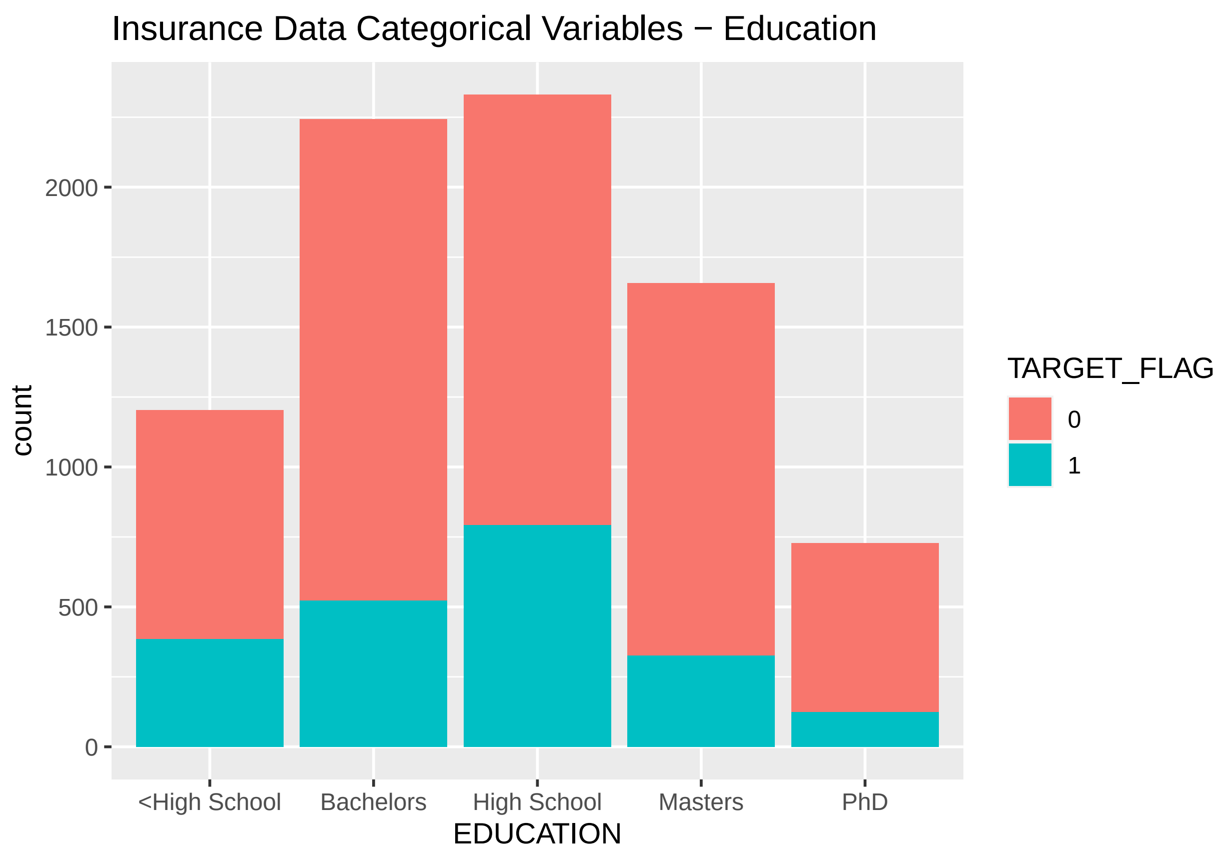

```r
#less imbalanced here
```

```r
ggplot(rawTrain, aes(x = SEX, fill = TARGET_FLAG)) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - SEX")
```



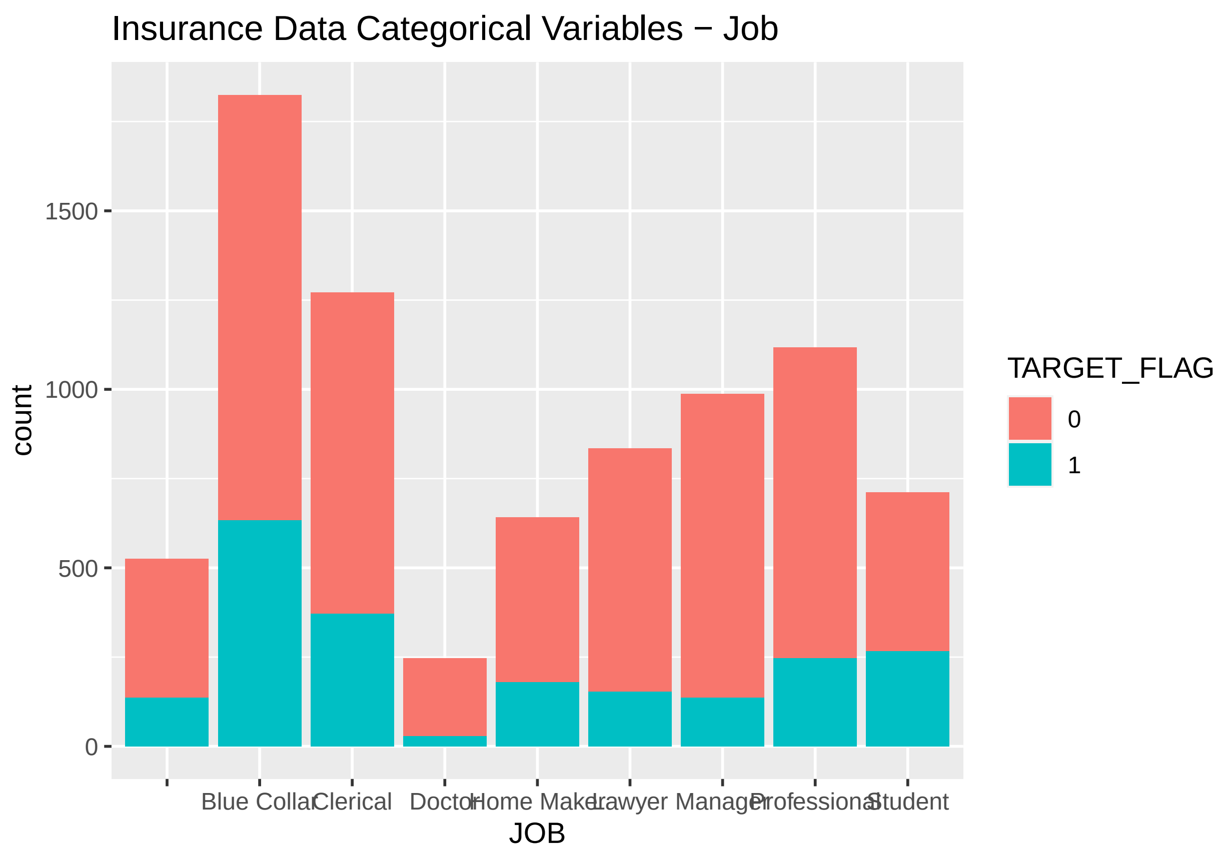

```
#I wouldnt consider this imbalanced but I am not sure what the threshold is for balance/imbalanced data
```

```
ggplot(rawTrain, aes(x = EDUCATION, fill = TARGET_FLAG)) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - Education")
```



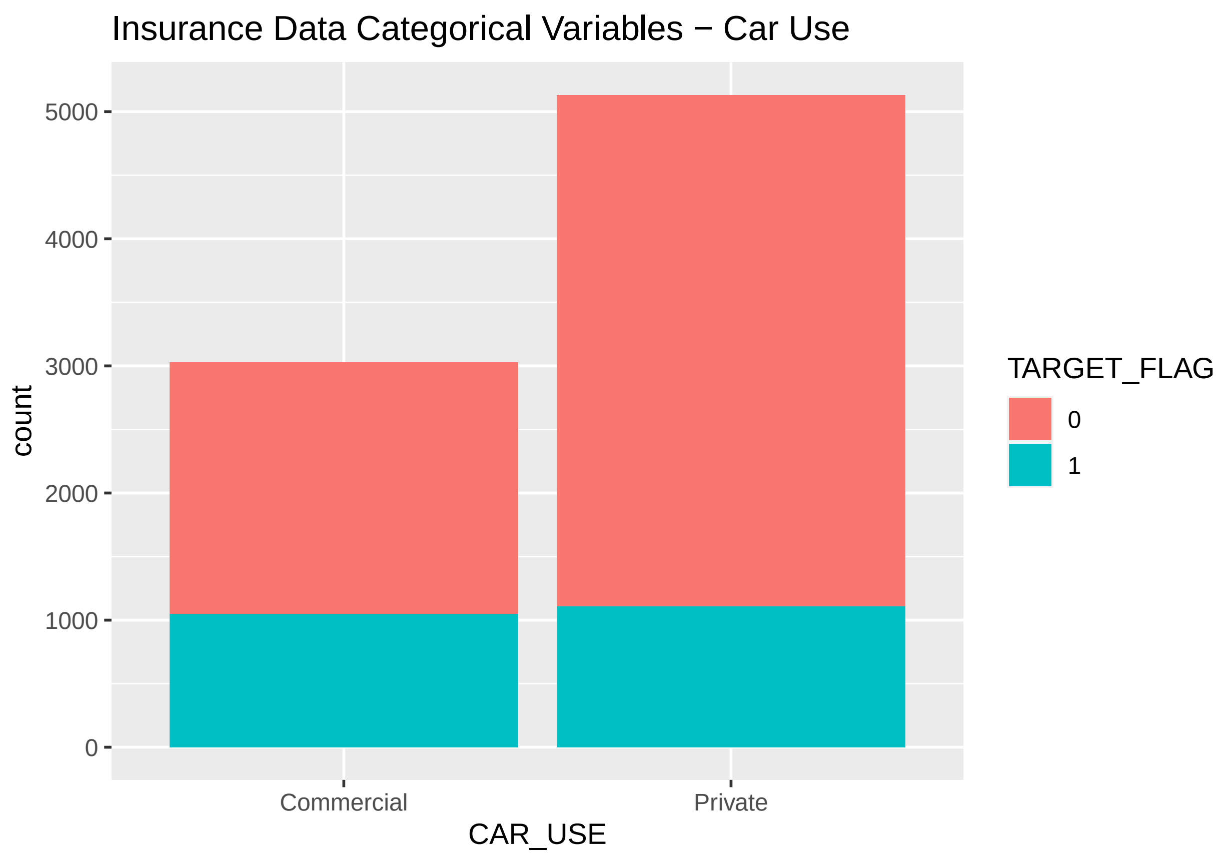

```
#I wouldnt consider this imbalanced but I am not sure what the threshold is for balance/imbalanced data
```

```
ggplot(rawTrain, aes(x = JOB, fill = TARGET_FLAG)) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - Job")
```



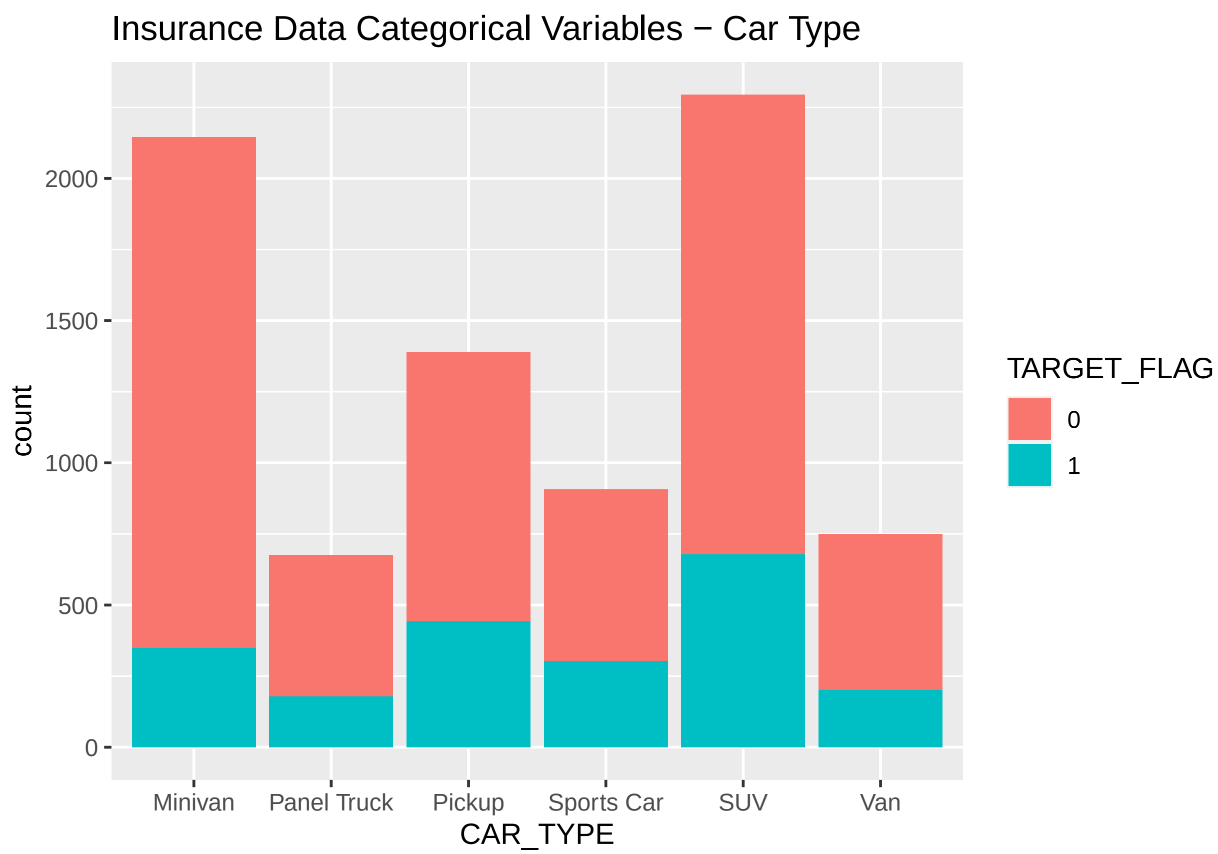

```
#I wouldnt consider this imbalanced but I am not sure what the threshold is for balance/imbalanced data
```

```
ggplot(rawTrain, aes(x = CAR_USE, fill = TARGET_FLAG)) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - Car Use")
```



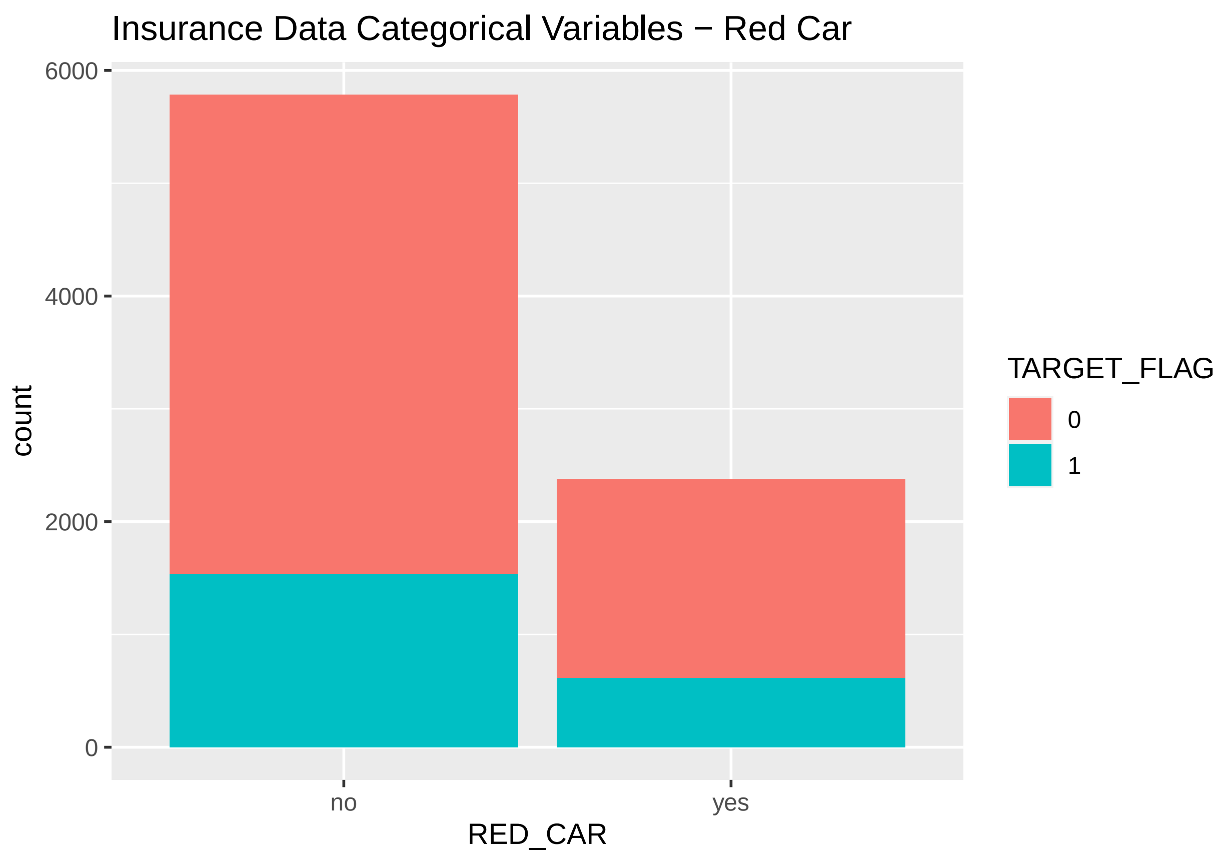

```r
#Imbalanced
```

```r
ggplot(rawTrain, aes(x = CAR_TYPE, fill = TARGET_FLAG)) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - Car Type")
```



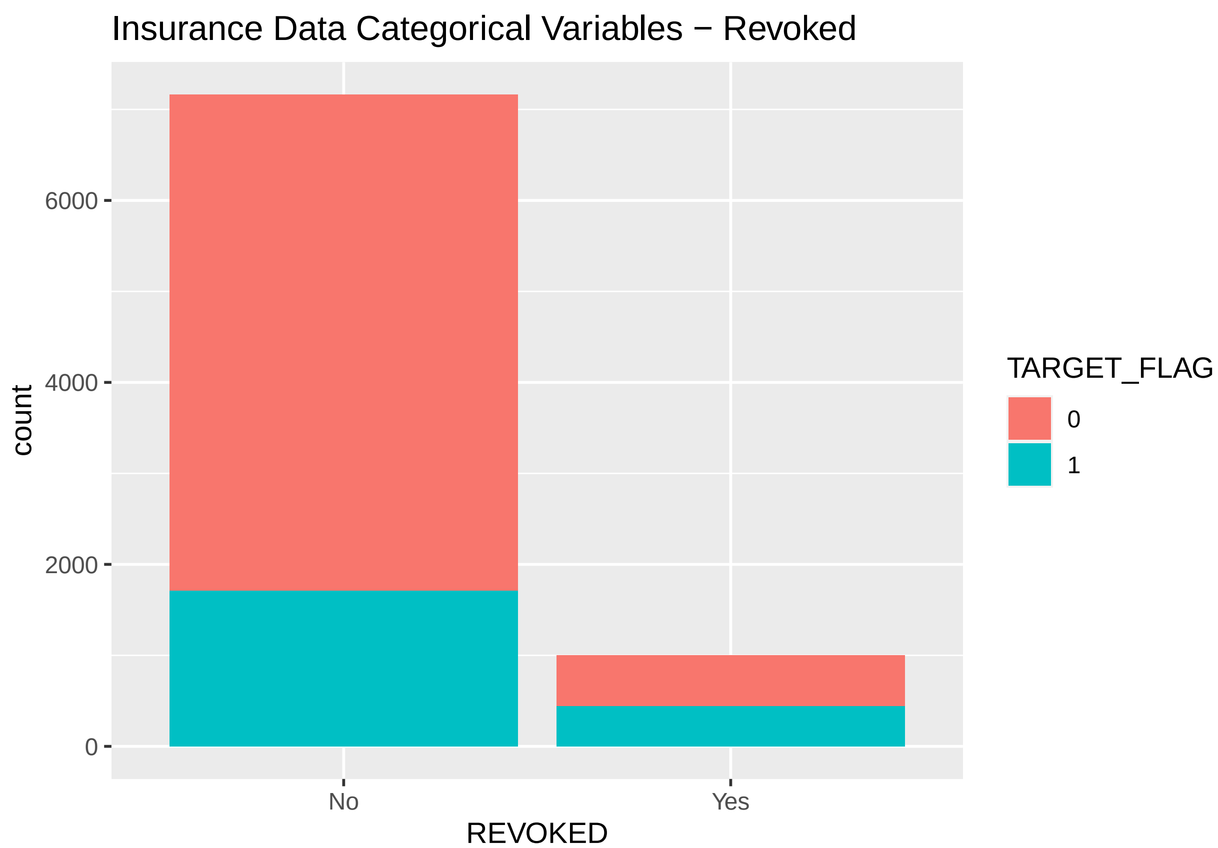

```
#Imbalanced
```

```
ggplot(rawTrain, aes(x = RED_CAR, fill = TARGET_FLAG)) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - Red Car")
```




```
#Imbalanced
```

```
ggplot(rawTrain, aes(x = REVOKED, fill = TARGET_FLAG)) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - Revoked")
```

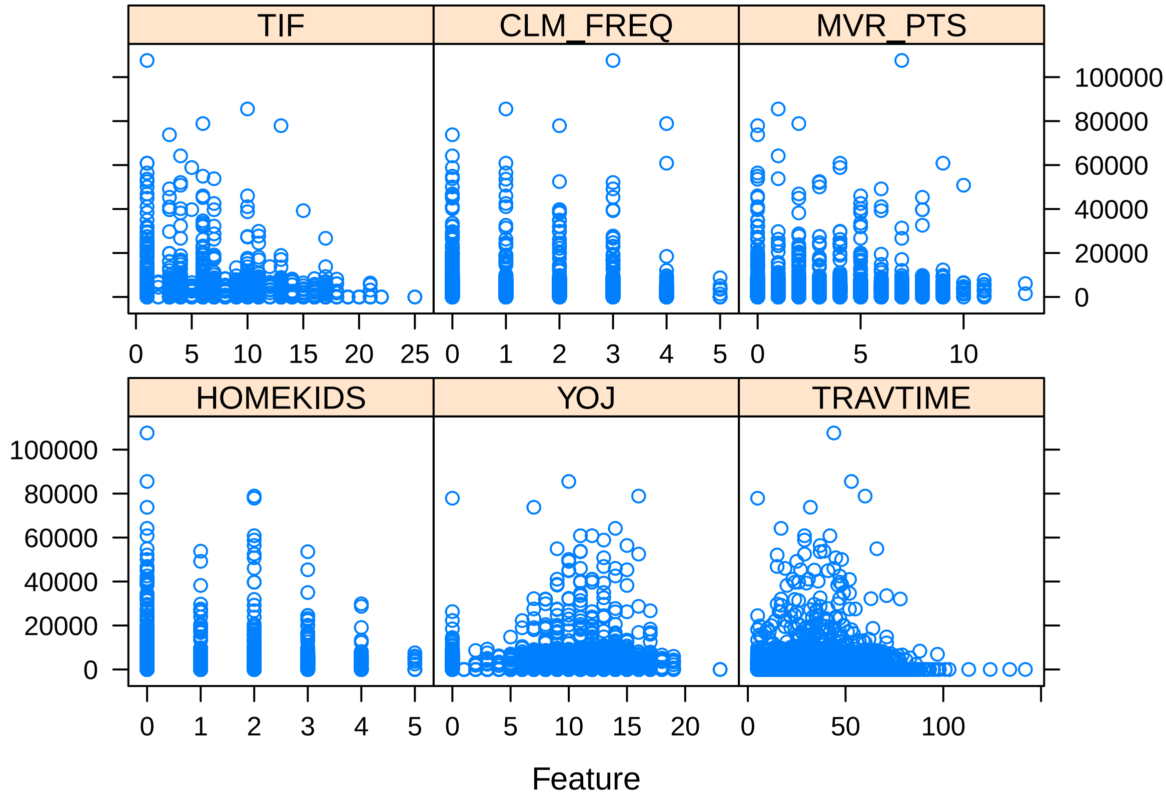



```
#Imbalanced
```

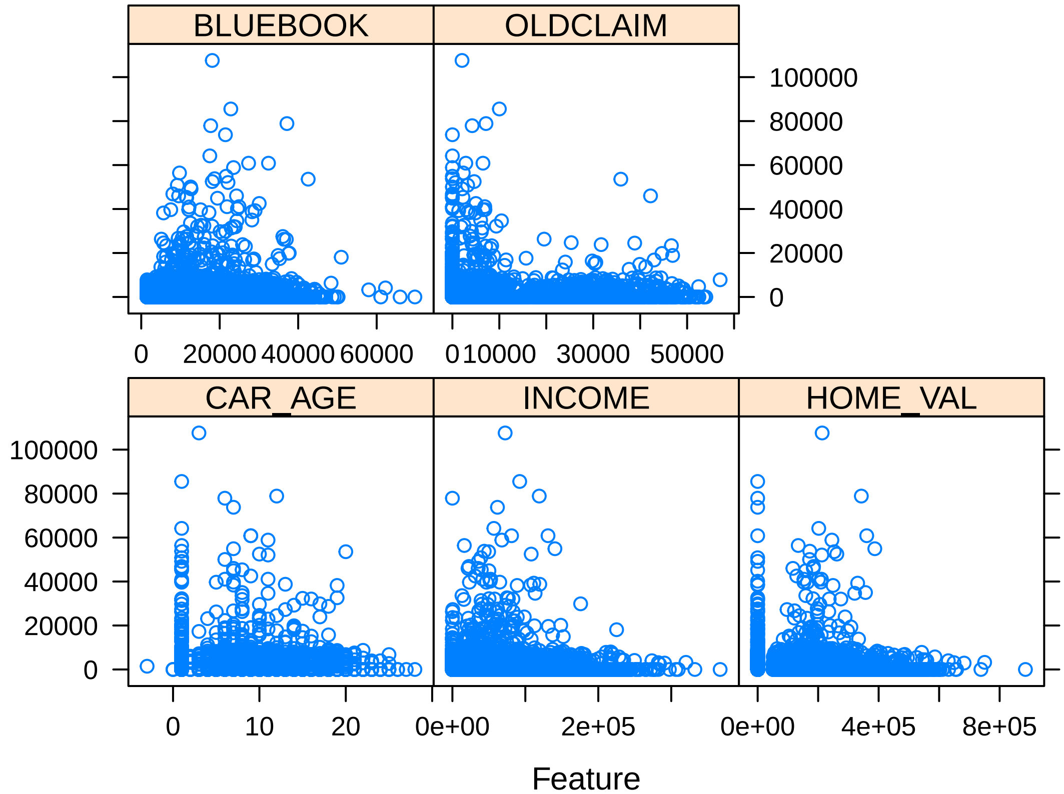
Numeric Data - Relationship to Target

```
#include target in the df for numeric data
histData <- rawTrain %>%
  select(TARGET_AMT, AGE, HOMEKIDS, YOJ,TRAVTIME, TIF, CLM_FREQ, MVR_PTS, CAR_AGE, INCOME, HOME_VAL, BLU

#How do I color by Target_flag
featurePlot(x= histData[3:8], y = histData[['TARGET_AMT']])
```




```r
featurePlot(x= histData[9:13], y = histData[['TARGET_AMT']])
```

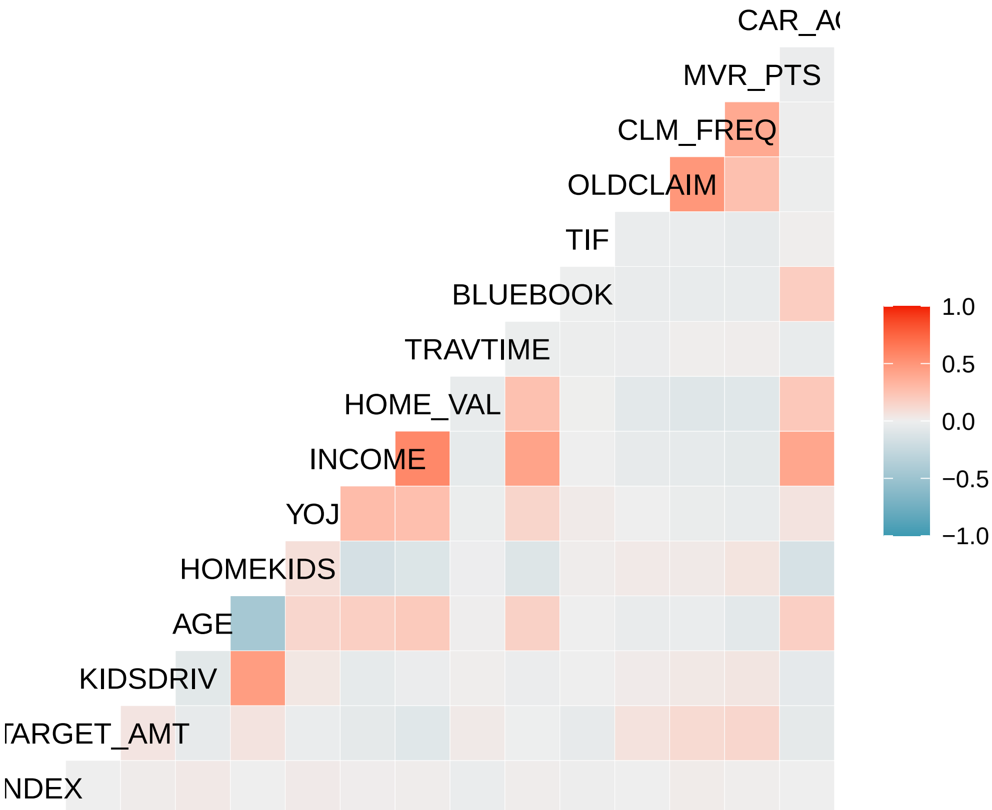



```r
#HOME KIDS and AGE NEED BAR CHARTS
```

Correlation

```r
#correlation matrix for predictors
ggcorr(rawTrain)
```

```
## Warning in ggcorr(rawTrain): data in column(s) 'TARGET_FLAG', 'PARENT1',
## 'MSTATUS', 'SEX', 'EDUCATION', 'JOB', 'CAR_USE', 'CAR_TYPE', 'RED_CAR',
## 'REVOKED', 'URBANICITY' are not numeric and were ignored
```




```
#Lets look at some highly correlated variables and drop them
findCorrelation(cor(histData),cutoff = 0.75, verbose = TRUE, names = TRUE)
```

```
## All correlations <= 0.75
```

```
## character(0)
```

```
# None of the numerical values are highly correlated
```

Data Cleaning

```
#due to skew home_val, income  will be imputed with median
#Age YOJ with the mean

#new DF
prepTrain <- rawTrain %>%
  select(-INDEX)

#impute NAs
prepTrain$AGE[is.na(prepTrain$AGE)] <- mean(prepTrain$AGE, na.rm=TRUE)
prepTrain$YOJ[is.na(prepTrain$YOJ)] <- mean(prepTrain$YOJ, na.rm=TRUE)
prepTrain$HOME_VAL[is.na(prepTrain$HOME_VAL)] <- median(prepTrain$HOME_VAL, na.rm=TRUE)
prepTrain$INCOME[is.na(prepTrain$INCOME)] <- median(prepTrain$INCOME, na.rm=TRUE)
```

```r
prepTrain$CAR_AGE[is.na(prepTrain$CAR_AGE)] <- mean(prepTrain$CAR_AGE, na.rm=TRUE)

# outlier detection and normalizing function
outlier_norm <- function(x){
  if (class(x) != "factor"){
    qntile <- quantile(x, probs=c(.25, .75))
     caps <- quantile(x, probs=c(.05, .95))
     H <- 1.5 * IQR(x, na.rm = T)
    x[x < (qntile[1] - H)] <- caps[1]
     x[x > (qntile[2] + H)] <- caps[2]
     return(x)
  }
}

#Aply the function to the columns in the dataframe
sapply(prepTrain, outlier_norm)
```

Variable Importance

```r
prepTrainA <- prepTrain %>%
  select(-TARGET_AMT)

prepTrainB <- prepTrain %>%
  select(-TARGET_FLAG)
```

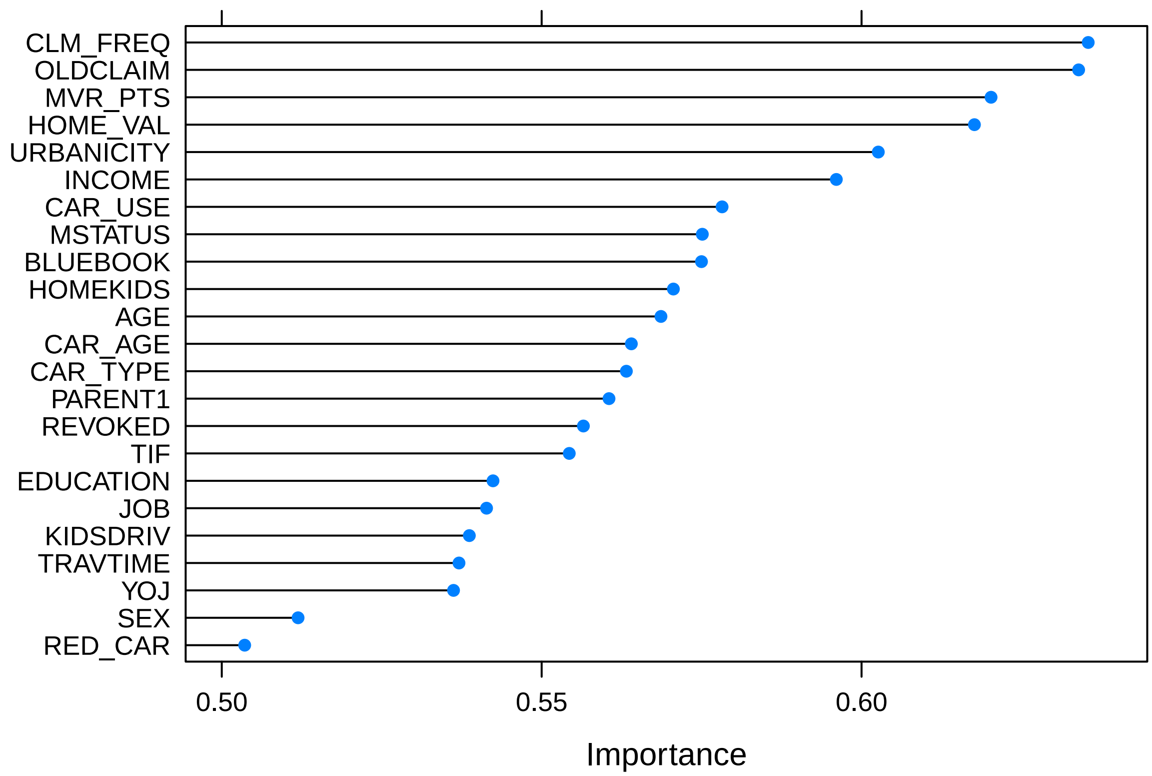
```r
# prepare training scheme
control <- trainControl(method="repeatedcv", number=10, repeats=3)

# train the model
modelA <- train(TARGET_FLAG~., data=prepTrainA, method="lvq", preProcess="scale", trControl=control)
# estimate variable importance
importance <- varImp(modelA, scale=FALSE)
# summarize importance
print(importance)
```

```
## ROC curve variable importance
##
##   only 20 most important variables shown (out of 23)
##
##              Importance
## CLM_FREQ         0.6354
## OLDCLAIM         0.6339
## MVR_PTS          0.6202
## HOME_VAL         0.6176
## URBANICITY       0.6026
## INCOME           0.5961
## CAR_USE          0.5782
## MSTATUS          0.5751
## BLUEBOOK         0.5750
## HOMEKIDS         0.5706
## AGE              0.5686
```

```
## CAR_AGE        0.5640
## CAR_TYPE       0.5632
## PARENT1        0.5605
## REVOKED        0.5565
## TIF            0.5543
## EDUCATION      0.5424
## JOB            0.5414
## KIDSDRIV       0.5387
## TRAVTIME       0.5371
```

```
# plot importance
plot(importance)
```

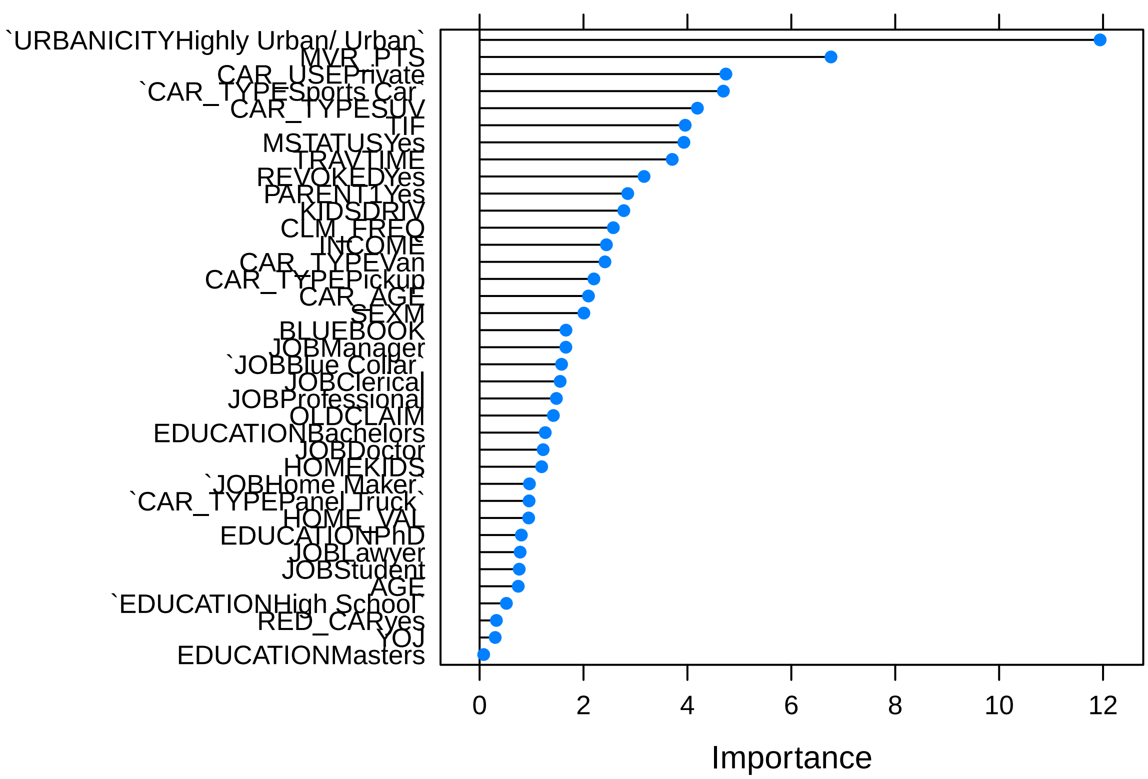



```
# train the model
modelB <- train(TARGET_AMT~., data=prepTrainB, method="glm", preProcess="scale", trControl=control)
# estimate variable importance
importance <- varImp(modelB, scale=FALSE)
# summarize importance
print(importance)
```

```
## glm variable importance
##
##   only 20 most important variables shown (out of 37)
##
##                                  Overall
```

```
## `URBANICITYHighly Urban/ Urban`  11.944
## MVR_PTS                           6.764
## CAR_USEPrivate                    4.741
## `CAR_TYPESports Car`              4.692
## CAR_TYPESUV                       4.193
## TIF                               3.958
## MSTATUSYes                        3.932
## TRAVTIME                          3.708
## REVOKEDYes                        3.166
## PARENT1Yes                        2.852
## KIDSDRIV                          2.776
## CLM_FREQ                          2.574
## INCOME                            2.441
## CAR_TYPEVan                       2.413
## CAR_TYPEPickup                    2.200
## CAR_AGE                           2.096
## SEXM                              2.007
## BLUEBOOK                          1.663
## JOBManager                        1.660
## `JOBBlue Collar`                  1.578
```

```r
# plot importance
plot(importance)
```



According to the plot above we can predict which variables would contribute best to the categorical predictions for `TARGET_FLAG`. We can use this to inform our data transformations.

Train Test Split

```r
## set the seed to make your partition reproducible
set.seed(123)
trainIndex<- sort(sample(nrow(prepTrain), nrow(prepTrain)*.8))

train <- prepTrain[trainIndex, ]
test <- prepTrain[-trainIndex, ]
```

Models

```r
##Baseline (logistic regression)
modelOne <- glm(formula = TARGET_FLAG ~ . - TARGET_AMT, data=train, family = "binomial" (link="logit"))
summary(modelOne)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ . - TARGET_AMT, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6207  -0.7138  -0.3982   0.6320   3.1760
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -2.794e+00  3.811e-01  -7.331 2.29e-13 ***
## KIDSDRIV                       3.954e-01  6.933e-02   5.703 1.18e-08 ***
## AGE                           -3.360e-03  4.509e-03  -0.745 0.456212
## HOMEKIDS                       2.628e-02  4.177e-02   0.629 0.529287
## YOJ                           -1.639e-02  9.646e-03  -1.699 0.089301 .
## INCOME                        -2.356e-06  1.194e-06  -1.972 0.048596 *
## PARENT1Yes                     4.746e-01  1.226e-01   3.871 0.000108 ***
## HOME_VAL                      -1.381e-06  3.795e-07  -3.640 0.000273 ***
## MSTATUSYes                    -4.922e-01  9.386e-02  -5.244 1.57e-07 ***
## SEXM                           6.883e-02  1.256e-01   0.548 0.583642
## EDUCATIONBachelors            -4.420e-01  1.295e-01  -3.413 0.000643 ***
## EDUCATIONHigh School          -5.567e-02  1.070e-01  -0.520 0.602836
## EDUCATIONMasters              -3.802e-01  2.010e-01  -1.891 0.058579 .
## EDUCATIONPhD                  -2.484e-01  2.370e-01  -1.048 0.294649
## JOBBlue Collar                 3.697e-01  2.081e-01   1.777 0.075644 .
## JOBClerical                    4.590e-01  2.202e-01   2.085 0.037058 *
## JOBDoctor                     -2.672e-01  2.901e-01  -0.921 0.357022
## JOBHome Maker                  3.097e-01  2.358e-01   1.314 0.188979
## JOBLawyer                      1.798e-01  1.916e-01   0.938 0.348195
## JOBManager                    -4.673e-01  1.928e-01  -2.424 0.015348 *
## JOBProfessional                2.623e-01  2.002e-01   1.310 0.190294
## JOBStudent                     2.746e-01  2.409e-01   1.140 0.254280
## TRAVTIME                       1.493e-02  2.105e-03   7.091 1.33e-12 ***
## CAR_USEPrivate                -7.869e-01  1.025e-01  -7.680 1.59e-14 ***
```

```
## BLUEBOOK                         -2.070e-05  5.921e-06  -3.496 0.000473 ***
## TIF                              -5.618e-02  8.141e-03  -6.901 5.17e-12 ***
## CAR_TYPEPanel Truck               5.310e-01  1.829e-01   2.903 0.003694 **
## CAR_TYPEPickup                    5.420e-01  1.125e-01   4.818 1.45e-06 ***
## CAR_TYPESports Car                1.067e+00  1.446e-01   7.377 1.62e-13 ***
## CAR_TYPESUV                       7.894e-01  1.239e-01   6.369 1.91e-10 ***
## CAR_TYPEVan                       7.015e-01  1.403e-01   5.002 5.68e-07 ***
## RED_CARyes                       -1.634e-02  9.674e-02  -0.169 0.865834
## OLDCLAIM                         -1.115e-05  4.394e-06  -2.537 0.011172 *
## CLM_FREQ                          1.718e-01  3.196e-02   5.377 7.55e-08 ***
## REVOKEDYes                        7.916e-01  1.026e-01   7.715 1.21e-14 ***
## MVR_PTS                           1.124e-01  1.523e-02   7.381 1.57e-13 ***
## CAR_AGE                          -3.696e-03  8.409e-03  -0.440 0.660251
## URBANICITYHighly Urban/ Urban  2.449e+00  1.263e-01  19.388  < 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: 7533.1  on 6527  degrees of freedom
## Residual deviance: 5827.2  on 6490  degrees of freedom
## AIC: 5903.2
##
## Number of Fisher Scoring iterations: 5
```

What is needed next is various models to be built after transforming some of these variables based on their shape (I would also play around with multiplying and dividing variables etc). One thing worth mentioning is that we have to predict two things. So essentially we have to come up with two types of models and test each of them. I was thinking like 3-4 models for each target showing how we are using the shape of the variables to determine transformation, feature engineering and feature selection.