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)
```

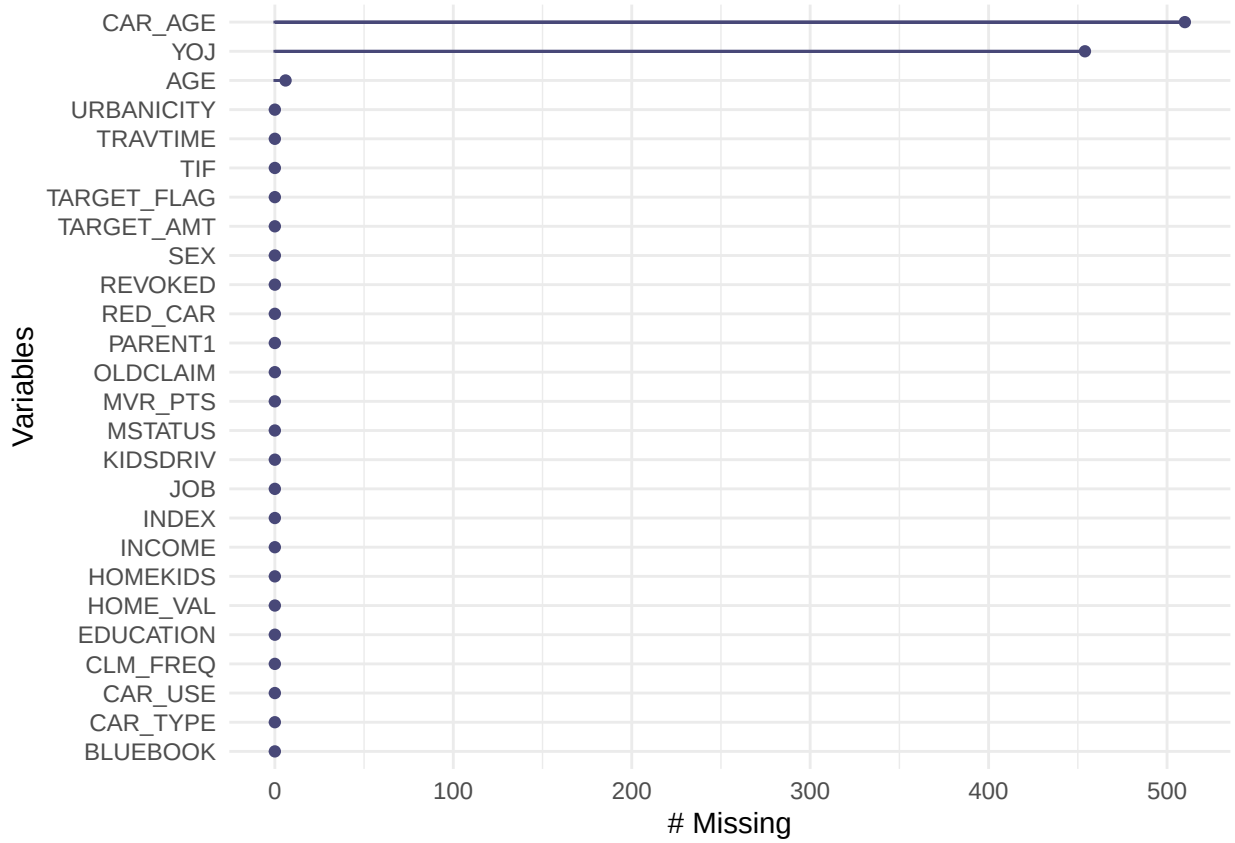
```
## Rows: 8,161
## Columns: 26
## $ INDEX       <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20...
## $ TARGET_FLAG <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0...
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 402...
## $ KIDSDRIV    <int> 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

```
#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)) %>%
  mutate(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)))

#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)
```

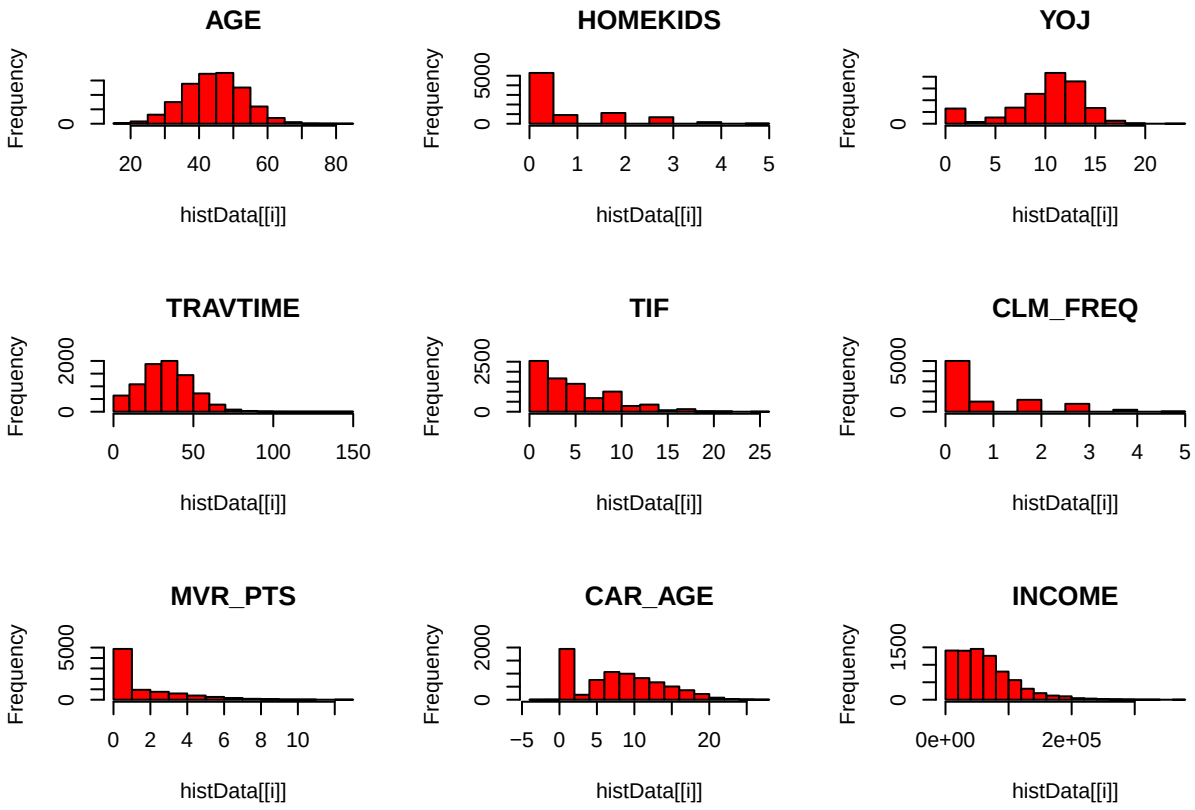
```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :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       PARENT1
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Min.   :     0   No :7084
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   1st Qu.: 28097   Yes:1077
##  Median :45.00   Median :0.0000   Median :11.0   Median : 54028
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5   Mean   : 61898
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   3rd Qu.: 85986
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   Max.   :367030
##  NA's   :6                        NA's   :454    NA's   :445
##     HOME_VAL      MSTATUS    SEX             EDUCATION              JOB
##  Min.   :     0   No :3267   F:4375   <High School:1203   Blue Collar :1825
##  1st Qu.:     0   Yes:4894   M:3786   Bachelors   :2242   Clerical    :1271
##  Median :161160                       High School :2330   Professional:1117
##  Mean   :154867                       Masters     :1658   Manager     : 988
##  3rd Qu.:238724                       PhD         : 728   Lawyer      : 835
##  Max.   :885282                                           Student     : 712
```

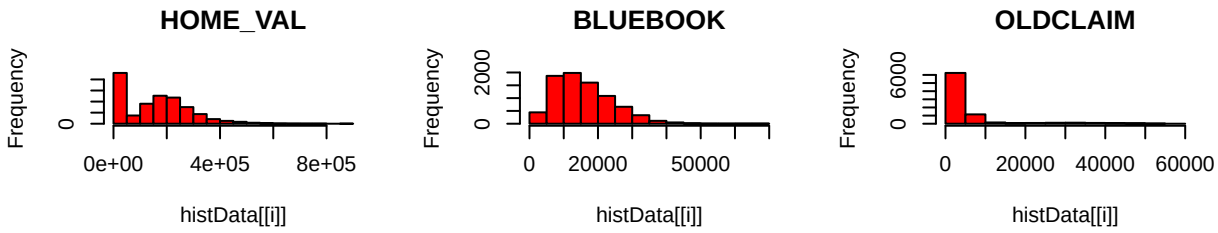
```
##  NA's   :464                                                (Other)     :1413
##     TRAVTIME           CAR_USE        BLUEBOOK          TIF
##  Min.   :  5.00   Commercial:3029   Min.   : 1500   Min.   : 1.000
##  1st Qu.: 22.00   Private   :5132   1st Qu.: 9280   1st Qu.: 1.000
##  Median : 33.00                     Median :14440   Median : 4.000
##  Mean   : 33.49                     Mean   :15710   Mean   : 5.351
##  3rd Qu.: 44.00                     3rd Qu.:20850   3rd Qu.: 7.000
##  Max.   :142.00                     Max.   :69740   Max.   :25.000
##
##         CAR_TYPE    RED_CAR       OLDCLAIM        CLM_FREQ      REVOKED
##  Minivan    :2145   no :5783   Min.   :    0   Min.   :0.0000   No :7161
##  Panel Truck: 676   yes:2378   1st Qu.:    0   1st Qu.:0.0000   Yes:1000
##  Pickup     :1389              Median :    0   Median :0.0000
##  Sports Car : 907              Mean   : 4037   Mean   :0.7986
##  SUV        :2294              3rd Qu.: 4636   3rd Qu.:2.0000
##  Van        : 750              Max.   :57037   Max.   :5.0000
##
##     MVR_PTS          CAR_AGE                     URBANICITY
##  Min.   : 0.000   Min.   :-3.000   Highly Rural/ Rural:1669
##  1st Qu.: 0.000   1st Qu.: 1.000   Highly Urban/ Urban:6492
##  Median : 1.000   Median : 8.000
##  Mean   : 1.696   Mean   : 8.328
##  3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :13.000   Max.   :28.000
##                   NA's   :510
```

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

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

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
Frequency
0
20
40
60
80
histData[[i]]
HOMEKIDS
Frequency
0
5000
0
1
2
3
4
5
histData[[i]]
YOJ
Frequency
0
0
5
10
15
20
histData[[i]]
TRAVTIME
Frequency
0
2000
0
50
100
150
histData[[i]]
TIF
Frequency
0
2500
0
5
10
15
20
25
histData[[i]]
CLM_FREQ
Frequency
0
5000
0
1
2
3
4
5
histData[[i]]
MVR_PTS
Frequency
0
5000
0
2
4
6
8
10
histData[[i]]
CAR_AGE
Frequency
0
2000
−5
0
5
10
20
histData[[i]]
INCOME
Frequency
0
1500
0e+00
2e+05
histData[[i]]

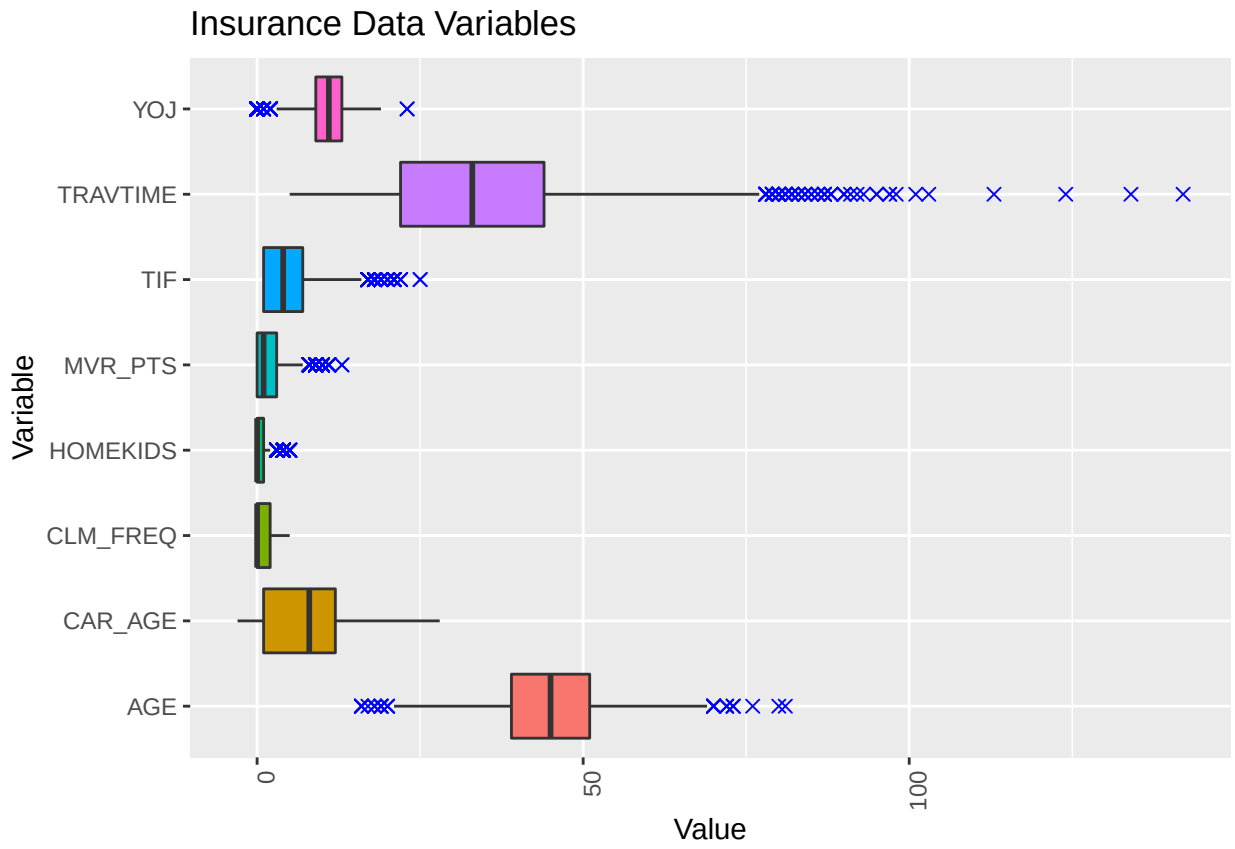


From the above histagrams of numerical data we can see that mose numerical variables have a right skew which may indicate that a transformation will be helpful for these variables.

```
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")
```

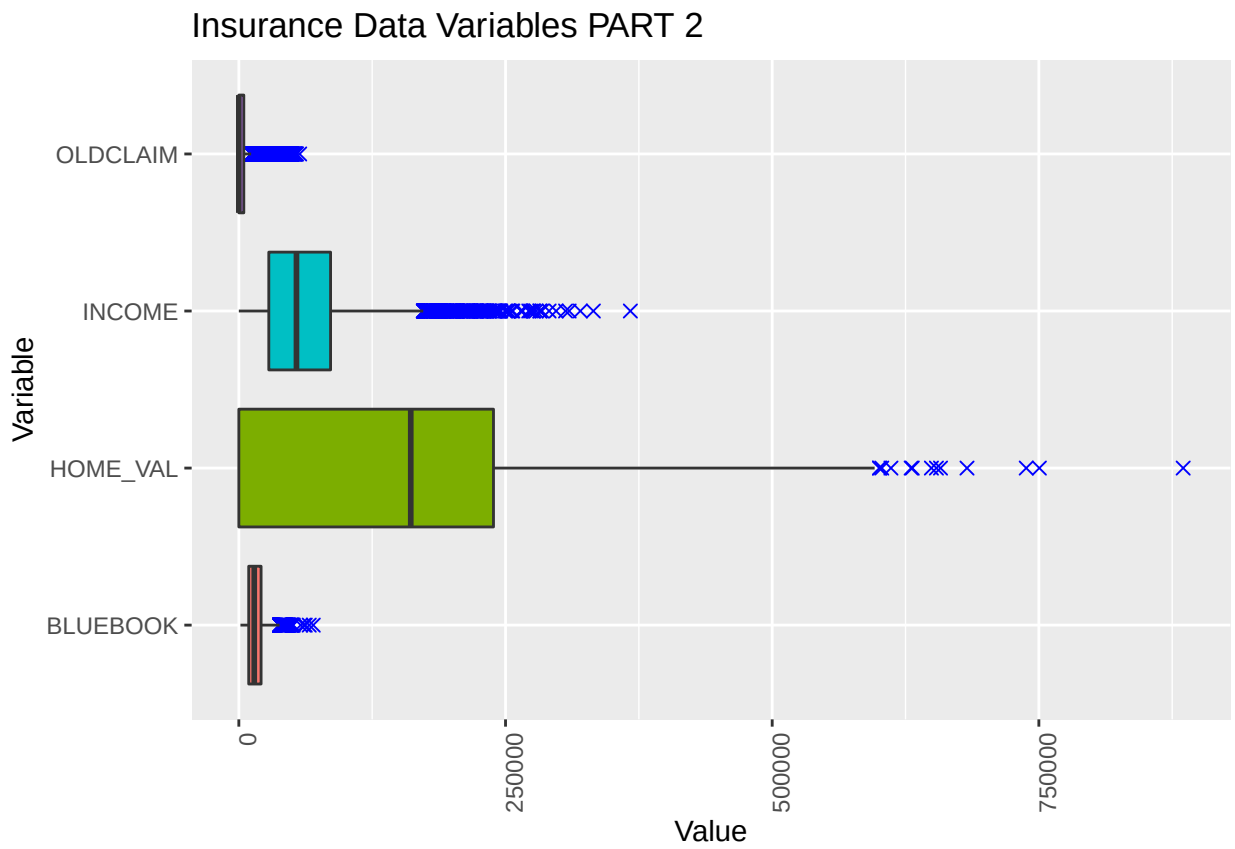
```
## Warning: Removed 970 rows containing non-finite values (stat_boxplot).
```




```
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).
```

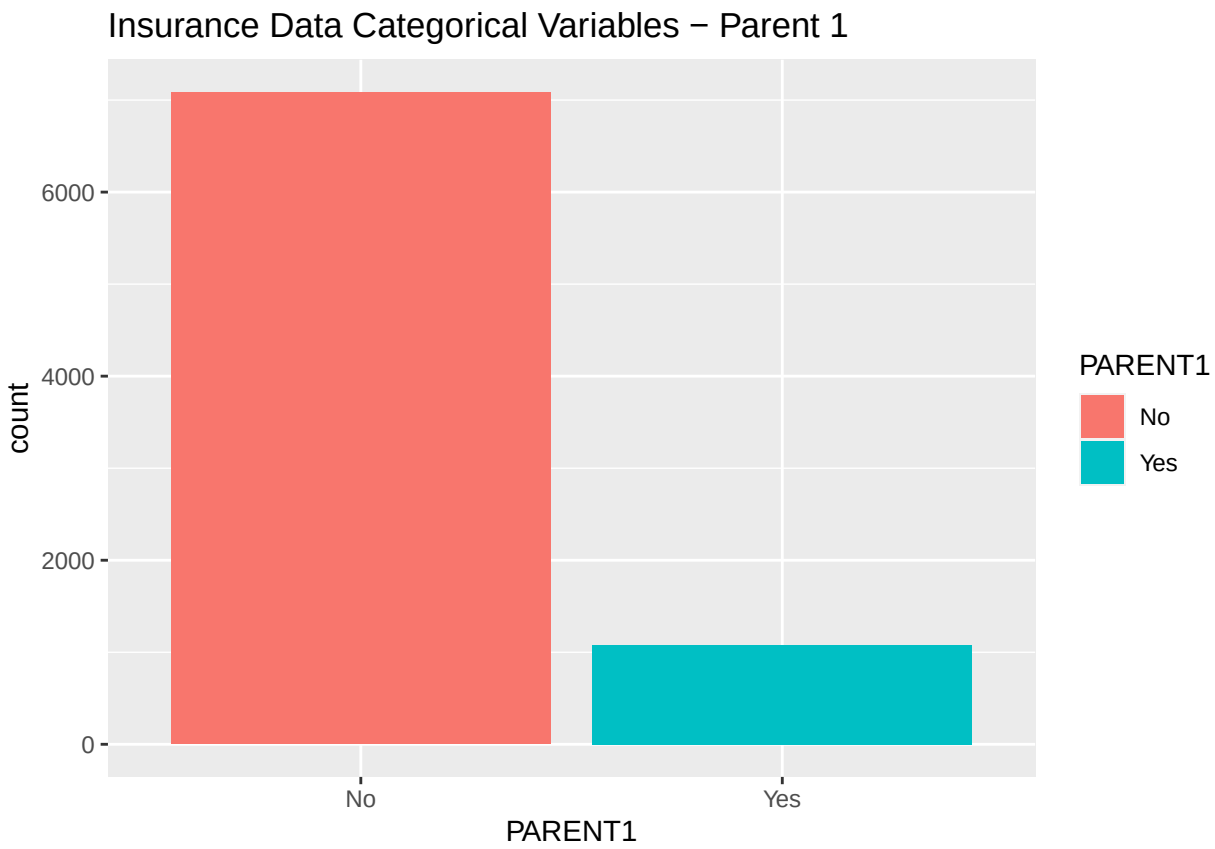


From these initial box plots we can see that there are outliers specifically `TRAVTIME`, `INCOME`, `HOME_VAL` has many outliers more spread out compared to the other variables.

Categorical Predictors

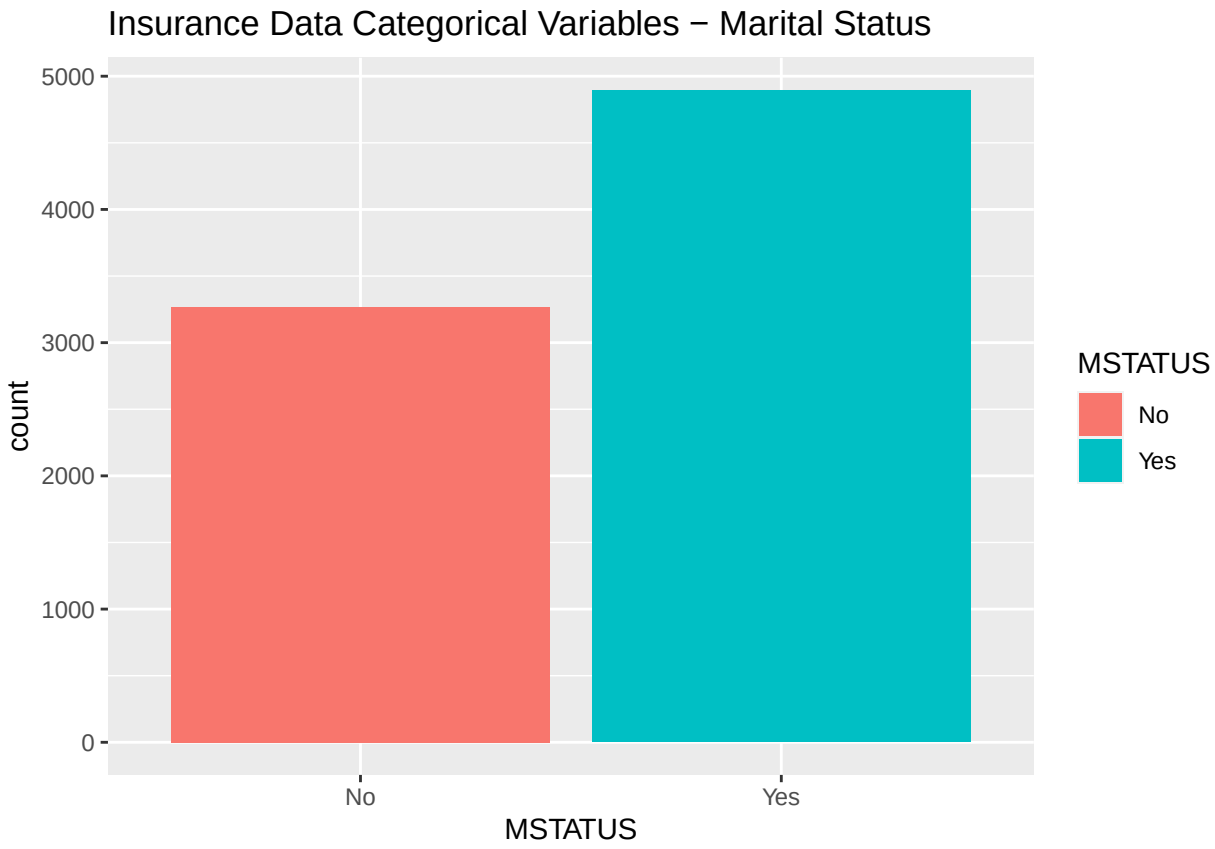
```
#select categorical data only
barData <- rawTrain %>%
  select(PARENT1, MSTATUS:JOB, CAR_USE, CAR_TYPE, RED_CAR,REVOKED, URBANICITY)

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



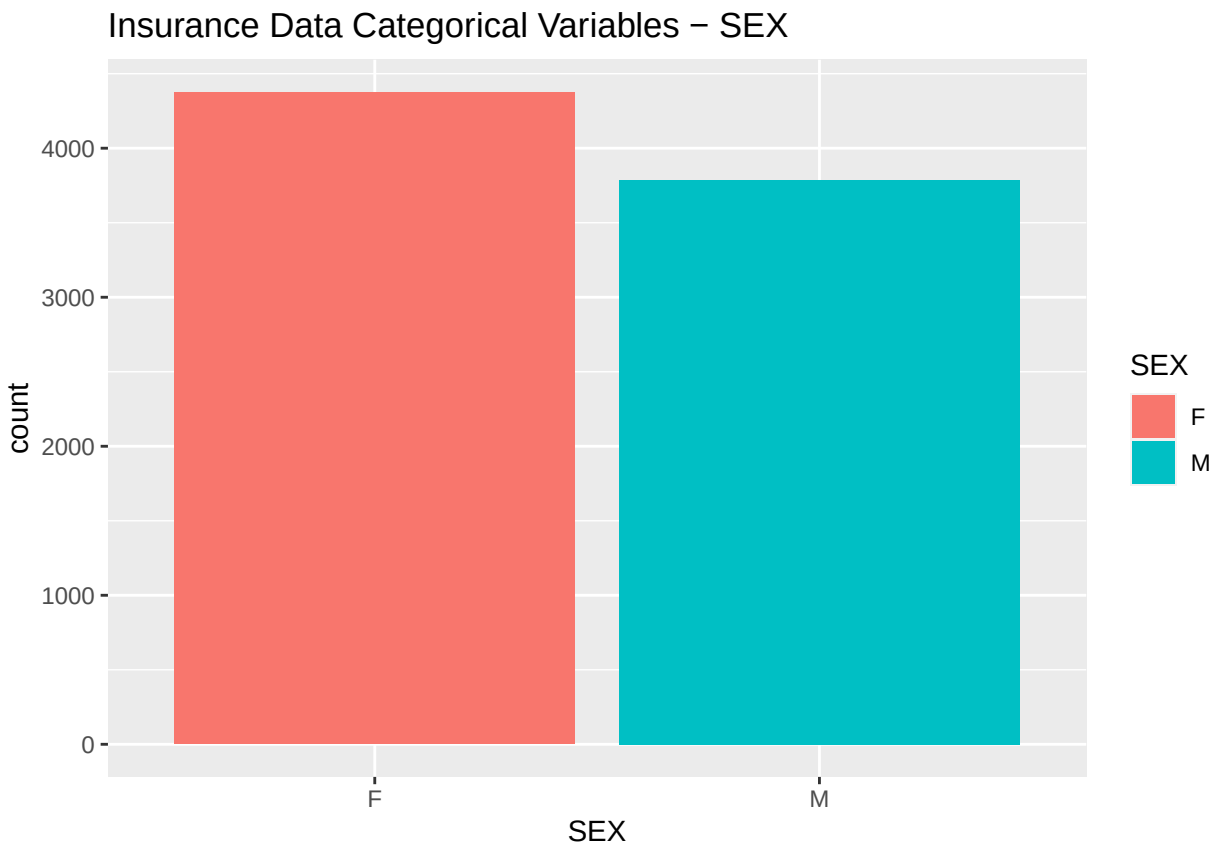

```
#imbalanced here
```

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



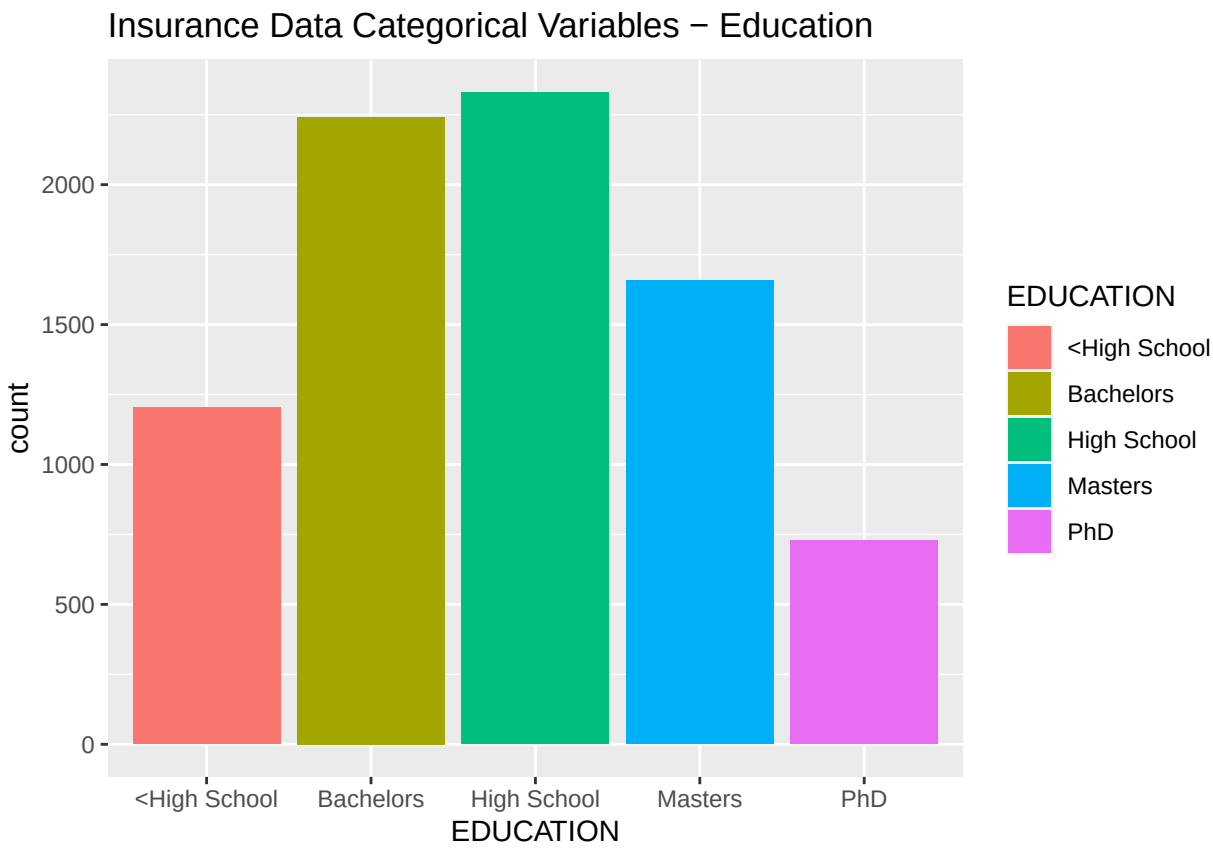

```
#less imbalanced here
```

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



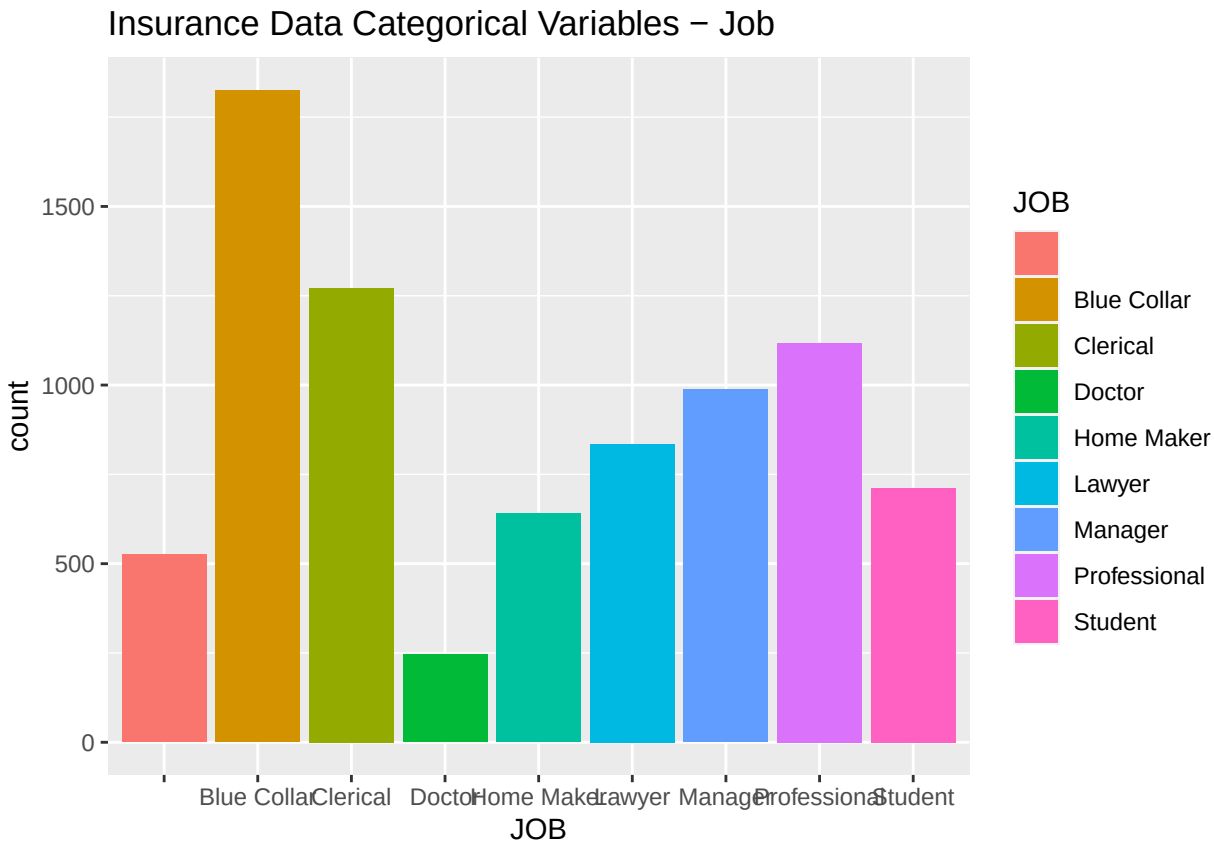

```
#I wouldnt consider this imbalanced but I am not sure what the threshold is for balance/imbalanced data
```

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



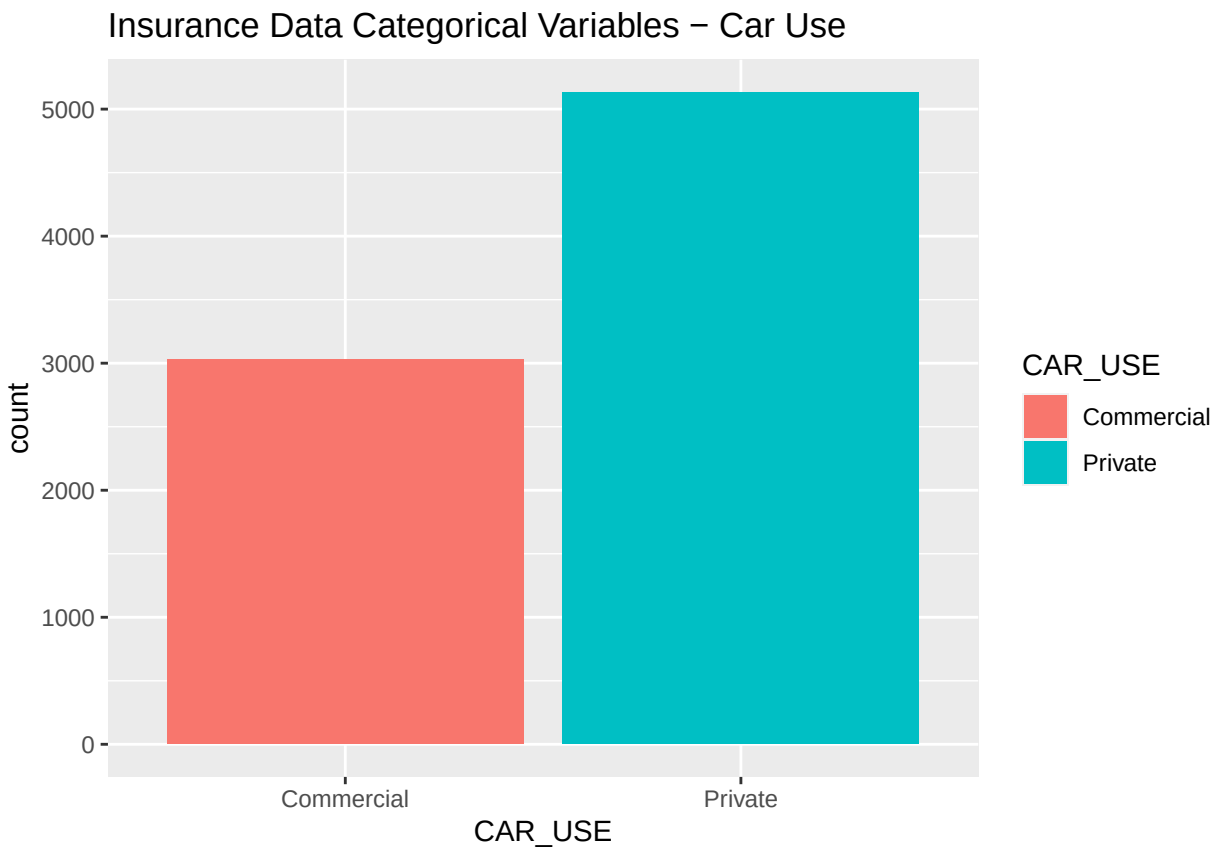

```
#I wouldnt consider this imbalanced but I am not sure what the threshold is for balance/imbalanced data
```

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



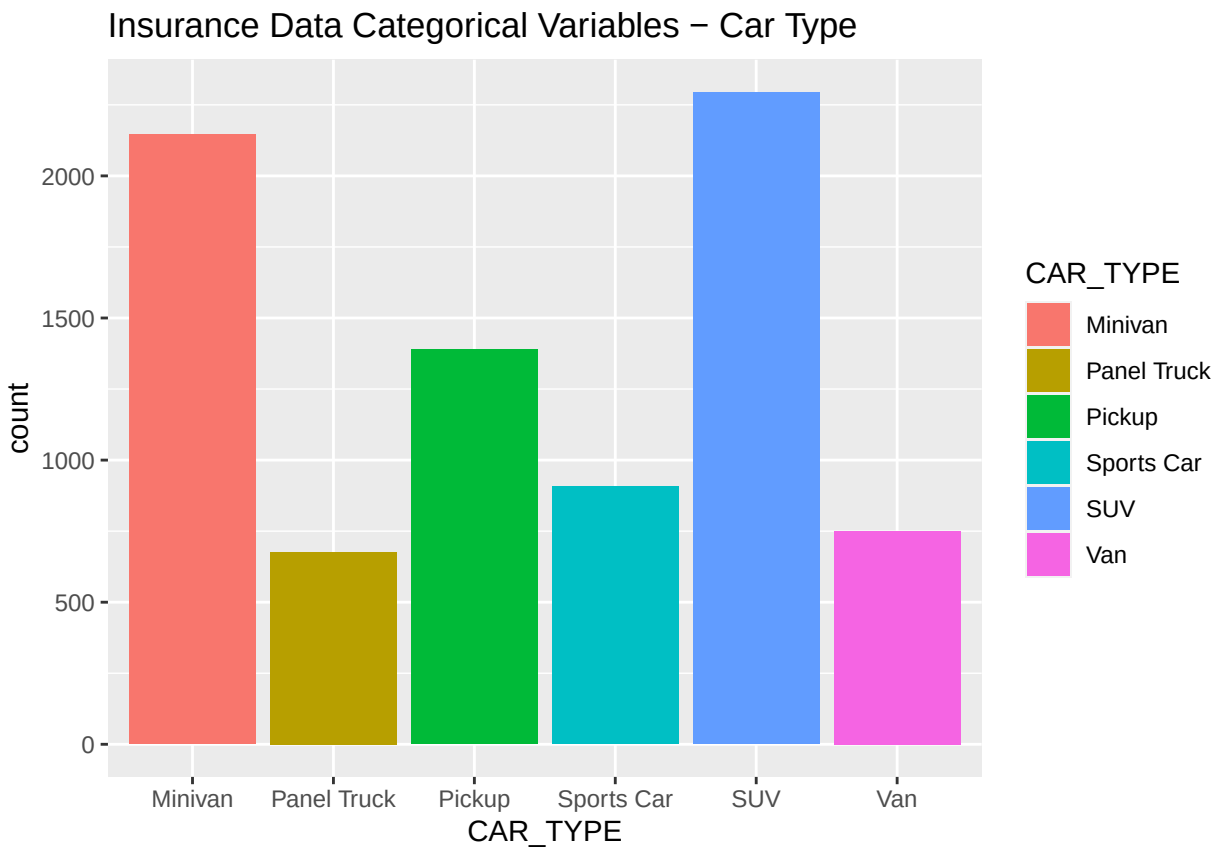

```
#I wouldnt consider this imbalanced but I am not sure what the threshold is for balance/imbalanced data
```

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



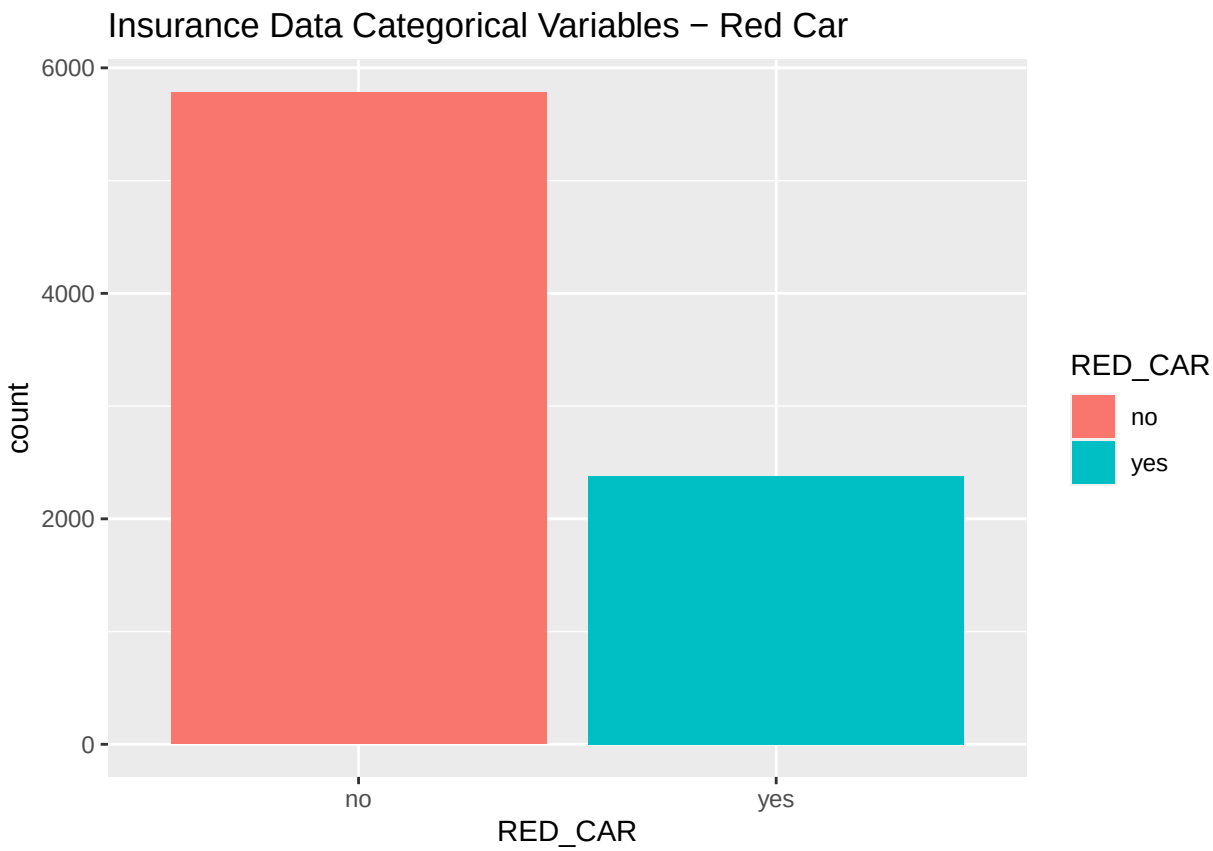

```
#Imbalanced
```

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



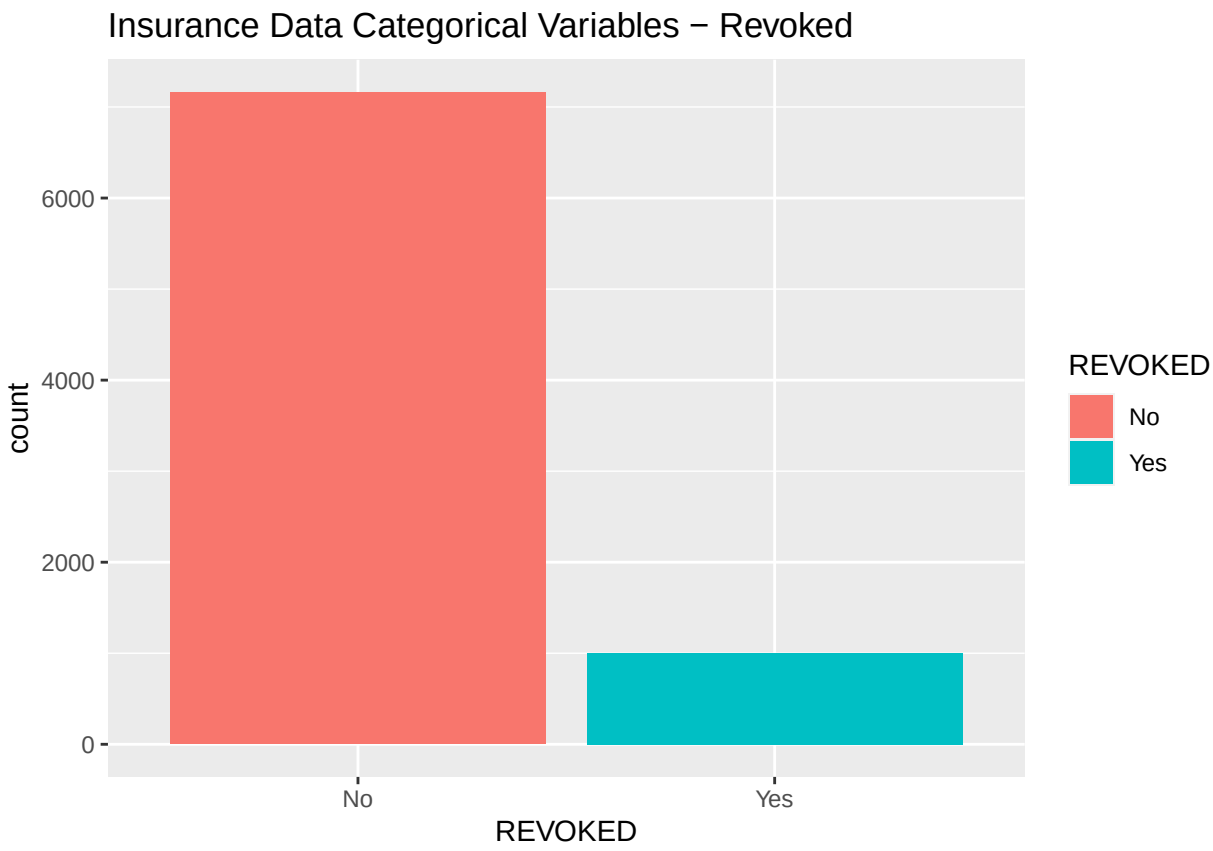

```
#Imbalanced
```

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




```
#Imbalanced
```

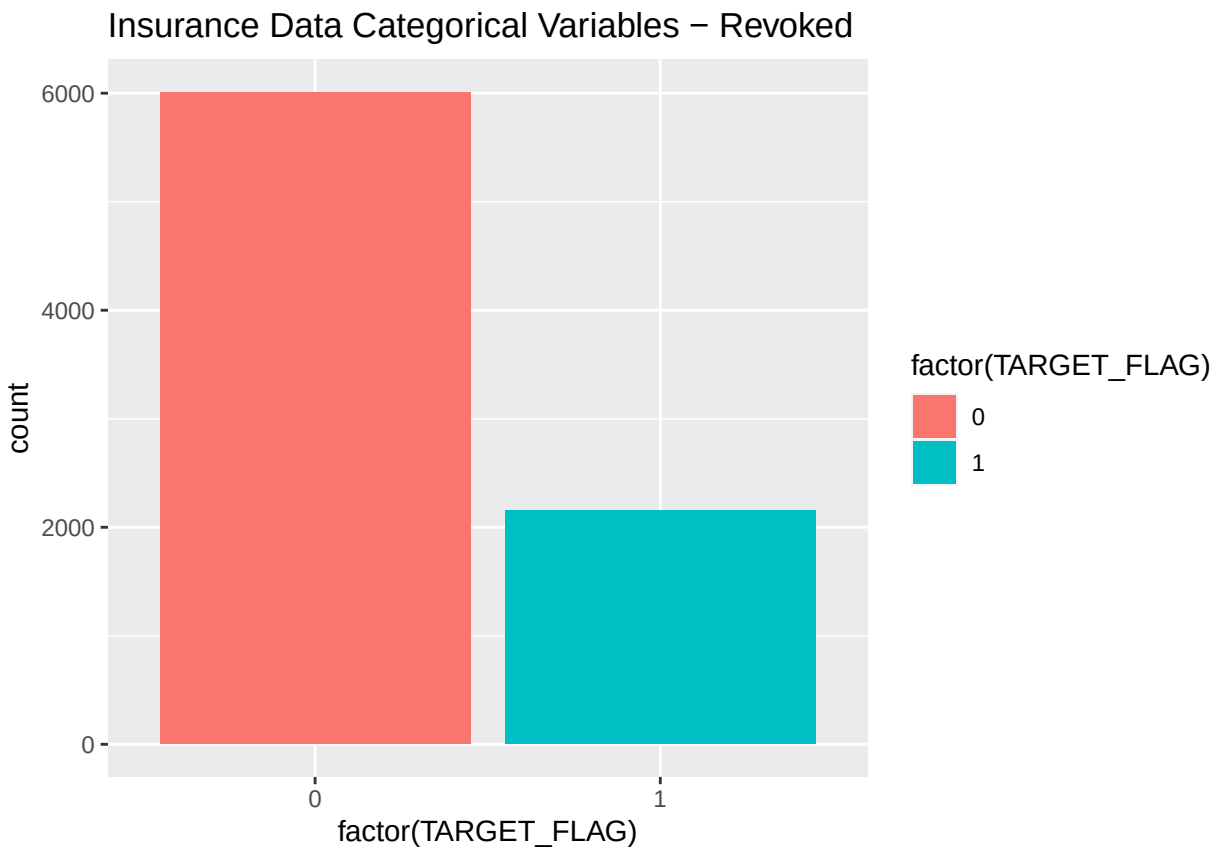
```
ggplot(barData, aes(x = REVOKED, fill = REVOKED)) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - Revoked")
```




```
#Imbalanced
```

TARGET VARIABLES

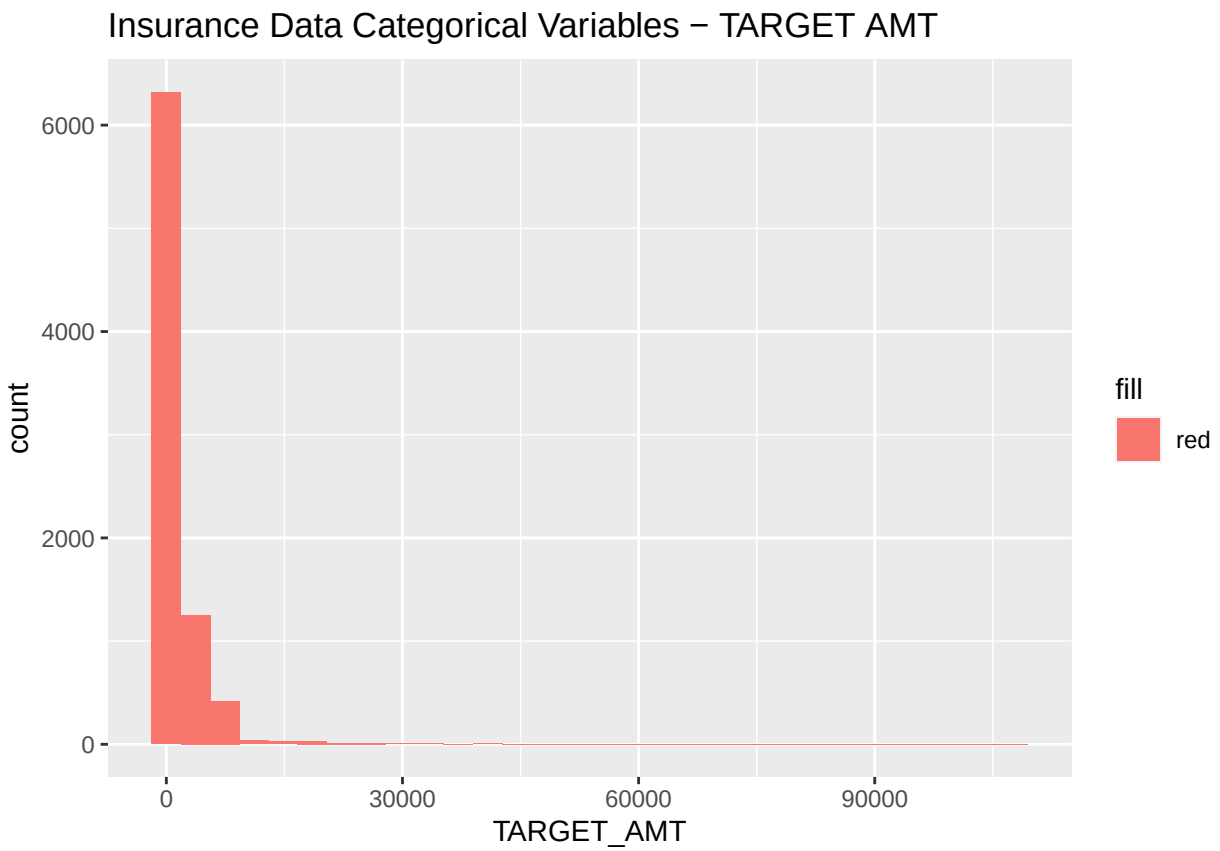
```
ggplot(rawTrain, aes(x = factor(TARGET_FLAG), fill =factor(TARGET_FLAG))) +
  geom_bar() +
  labs(title="Insurance Data Categorical Variables - Revoked")
```



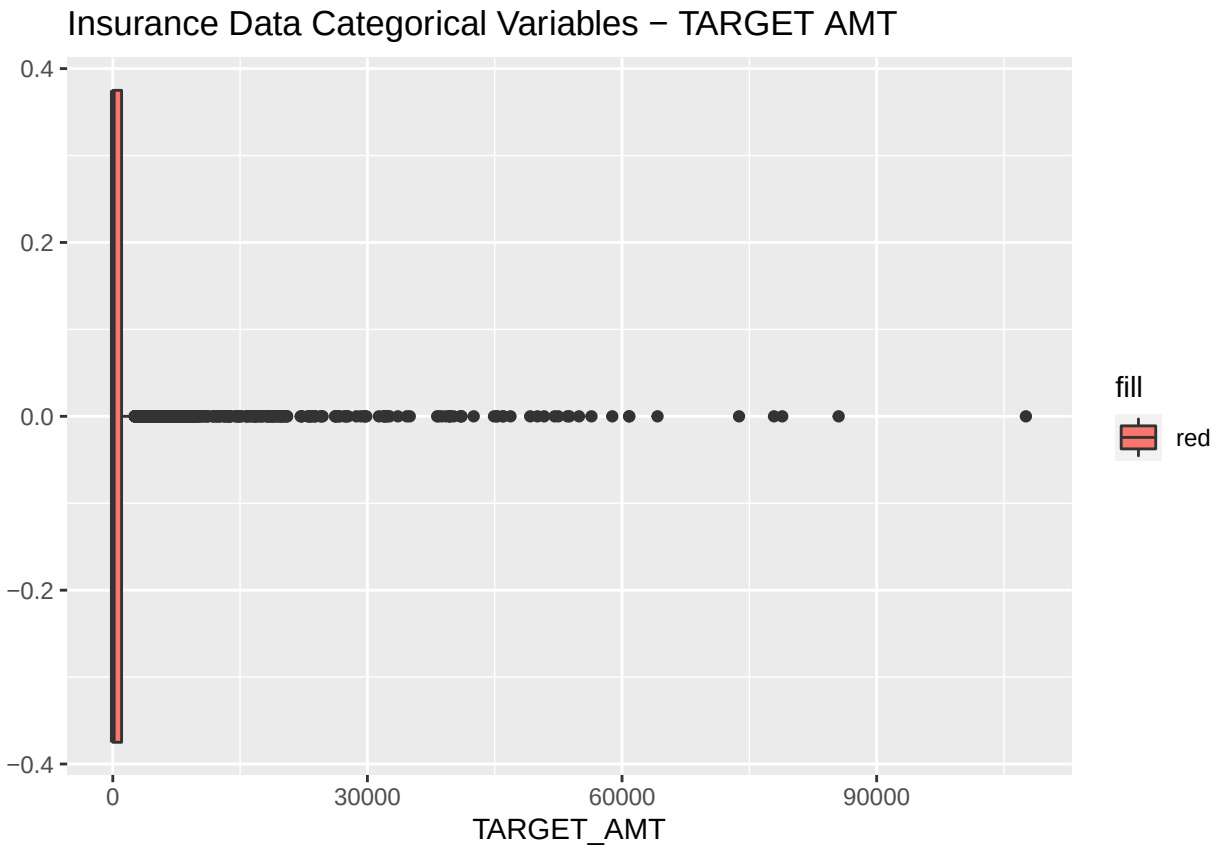

```
# Highly Skewed with lots of outliers

ggplot(rawTrain, aes(x = TARGET_AMT, fill = 'red')) + geom_histogram() +
  labs(title="Insurance Data Categorical Variables - TARGET AMT")
```

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




```
ggplot(rawTrain, aes(x = TARGET_AMT, fill = 'red')) + geom_boxplot() +
  labs(title="Insurance Data Categorical Variables - TARGET AMT")
```

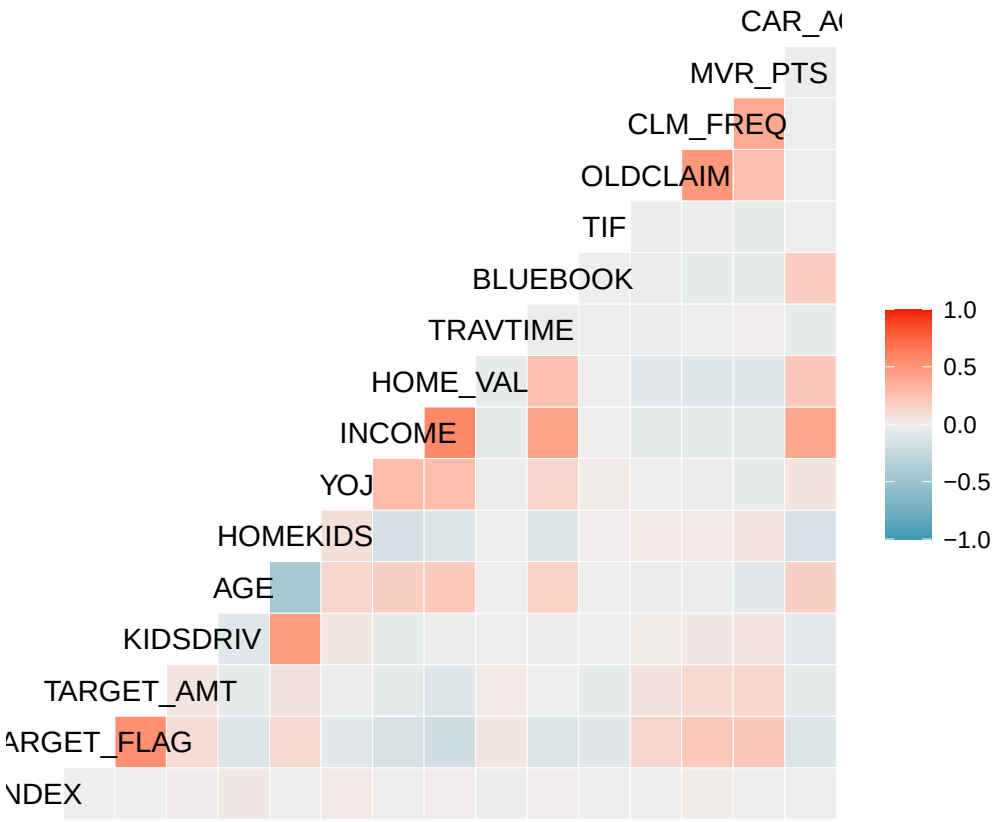



```
# Highly Skewed with lots of outliers
```

Correlation

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

```
## Warning in ggcorr(rawTrain): data in column(s) '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
```

—-I AM UP TO HERE———

Relationship to Target?

Use the appropriate column from the data set so you can plot a boxplot with target on the x-axis and variable on the y-axis.

Data Cleaning

- oulier and missing value imputing