Rui_Peng_HMEQ.R

raypeng

2025-03-23

```
#Step 1: Read in the Data
#Read the data into R
PATH = "/Users/raypeng/Documents/IS 5213 Data science and big data/Insurance"

FILE_NAME <- "HMEQ_Loss.csv"
OUT_NAME <- "HMEQ_Loss_Scrubbed.csv"

INFILE <- file.path(PATH, FILE_NAME)
OUTFILE <- file.path(PATH, OUT_NAME)

df = read.csv (INFILE)
head(df)
```

```
##   TARGET_BAD_FLAG TARGET_LOSS_AMT LOAN MORTDUE  VALUE  REASON    JOB  YOJ DEROG
## 1               1             641 1100   25860  39025 HomeImp  Other 10.5     0
## 2               1            1109 1300   70053  68400 HomeImp  Other  7.0     0
## 3               1             767 1500   13500  16700 HomeImp  Other  4.0     0
## 4               1            1425 1500      NA     NA                  NA    NA
## 5               0              NA 1700   97800 112000 HomeImp Office  3.0     0
## 6               1             335 1700   30548  40320 HomeImp  Other  9.0     0
##   DELINQ     CLAGE NINQ CLNO  DEBTINC
## 1      0  94.36667    1    9       NA
## 2      2 121.83333    0   14       NA
## 3      0 149.46667    1   10       NA
## 4     NA        NA   NA   NA       NA
## 5      0  93.33333    0   14       NA
## 6      0 101.46600    1    8 37.11361
```

```
setwd(PATH)
df <- read.csv(FILE_NAME)

#List the structure of the data (str)
str(df)
```

```
## 'data.frame':    5960 obs. of  14 variables:
##  $ TARGET_BAD_FLAG: int  1 1 1 1 0 1 1 1 1 1 ...
##  $ TARGET_LOSS_AMT: int  641 1109 767 1425 NA 335 1841 373 1217 1523 ...
##  $ LOAN           : int  1100 1300 1500 1500 1700 1700 1800 1800 2000 2000 ...
##  $ MORTDUE        : num  25860 70053 13500 NA 97800 ...
##  $ VALUE          : num  39025 68400 16700 NA 112000 ...
##  $ REASON         : chr  "HomeImp" "HomeImp" "HomeImp" "" ...
```

```
##  $ JOB            : chr  "Other" "Other" "Other" "" ...
##  $ YOJ            : num  10.5 7 4 NA 3 9 5 11 3 16 ...
##  $ DEROG          : int  0 0 0 NA 0 0 3 0 0 0 ...
##  $ DELINQ         : int  0 2 0 NA 0 0 2 0 2 0 ...
##  $ CLAGE          : num  94.4 121.8 149.5 NA 93.3 ...
##  $ NINQ           : int  1 0 1 NA 0 1 1 0 1 0 ...
##  $ CLNO           : int  9 14 10 NA 14 8 17 8 12 13 ...
##  $ DEBTINC        : num  NA NA NA NA NA ...
```

```
#Execute a summary of the data
summary(df)
```

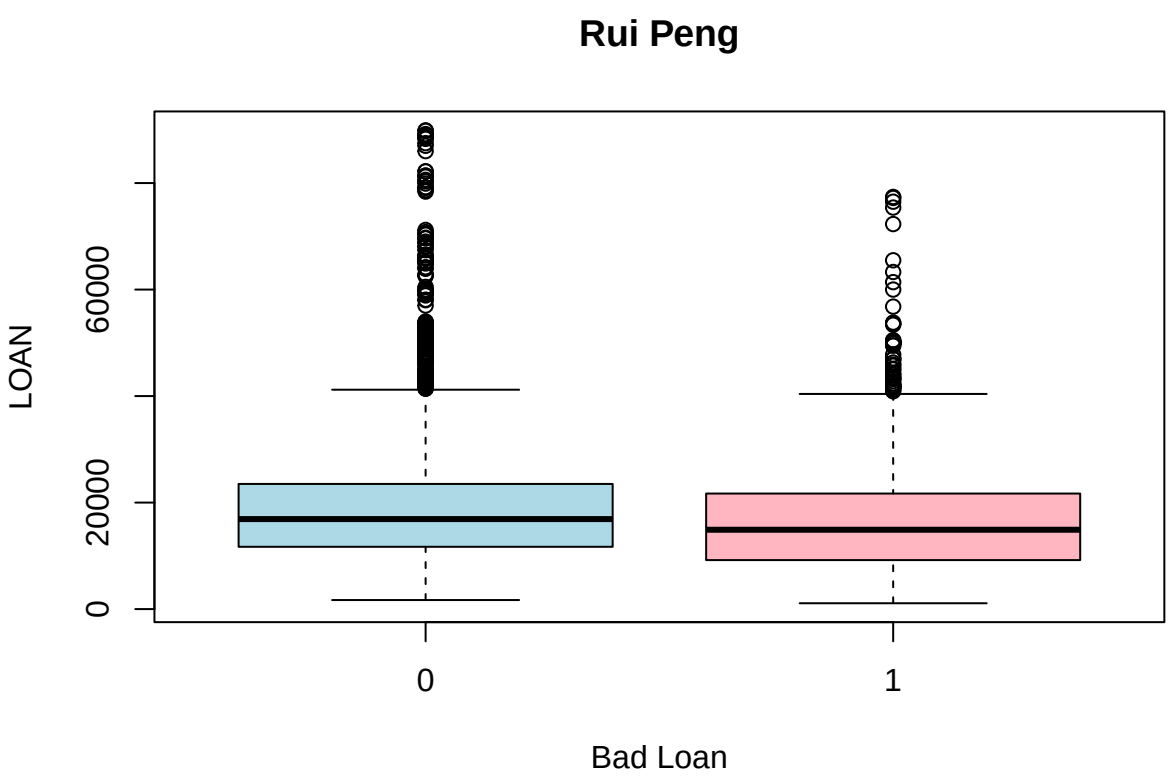
```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN          MORTDUE
##  Min.   :0.0000   Min.   :  224   Min.   : 1100   Min.   :  2063
##  1st Qu.:0.0000   1st Qu.: 5639   1st Qu.:11100   1st Qu.: 46276
##  Median :0.0000   Median :11003   Median :16300   Median : 65019
##  Mean   :0.1995   Mean   :13415   Mean   :18608   Mean   : 73761
##  3rd Qu.:0.0000   3rd Qu.:17634   3rd Qu.:23300   3rd Qu.: 91488
##  Max.   :1.0000   Max.   :78987   Max.   :89900   Max.   :399550
##                   NA's   :4771                    NA's   :518
##      VALUE           REASON              JOB                YOJ
##  Min.   :  8000   Length:5960        Length:5960        Min.   : 0.000
##  1st Qu.: 66076   Class :character   Class :character   1st Qu.: 3.000
##  Median : 89236   Mode  :character   Mode  :character   Median : 7.000
##  Mean   :101776                                         Mean   : 8.922
##  3rd Qu.:119824                                         3rd Qu.:13.000
##  Max.   :855909                                         Max.   :41.000
##  NA's   :112                                            NA's   :515
##      DEROG             DELINQ            CLAGE             NINQ
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :   0.0   Min.   : 0.000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 115.1   1st Qu.: 0.000
##  Median : 0.0000   Median : 0.0000   Median : 173.5   Median : 1.000
##  Mean   : 0.2546   Mean   : 0.4494   Mean   : 179.8   Mean   : 1.186
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 231.6   3rd Qu.: 2.000
##  Max.   :10.0000   Max.   :15.0000   Max.   :1168.2   Max.   :17.000
##  NA's   :708       NA's   :580       NA's   :308      NA's   :510
##       CLNO         DEBTINC
##  Min.   : 0.0   Min.   :  0.5245
##  1st Qu.:15.0   1st Qu.: 29.1400
##  Median :20.0   Median : 34.8183
##  Mean   :21.3   Mean   : 33.7799
##  3rd Qu.:26.0   3rd Qu.: 39.0031
##  Max.   :71.0   Max.   :203.3121
##  NA's   :222    NA's   :1267
```

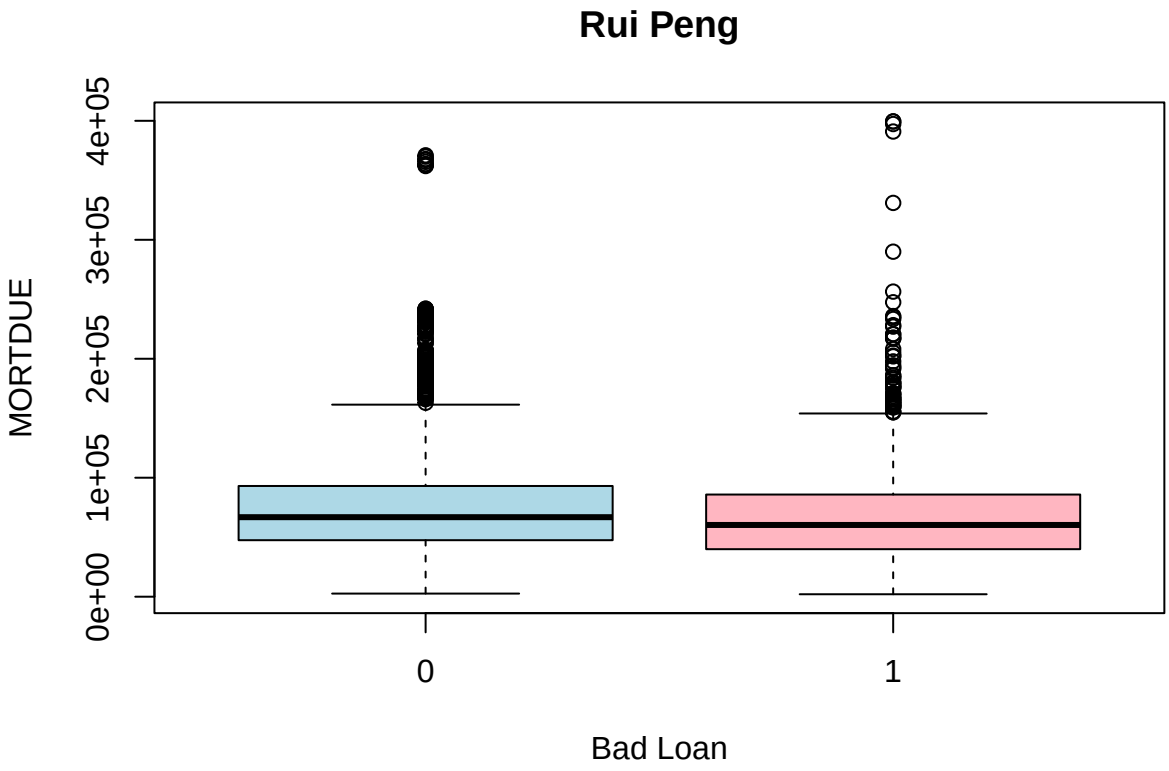
```
#Print the first six records
head (df)
```

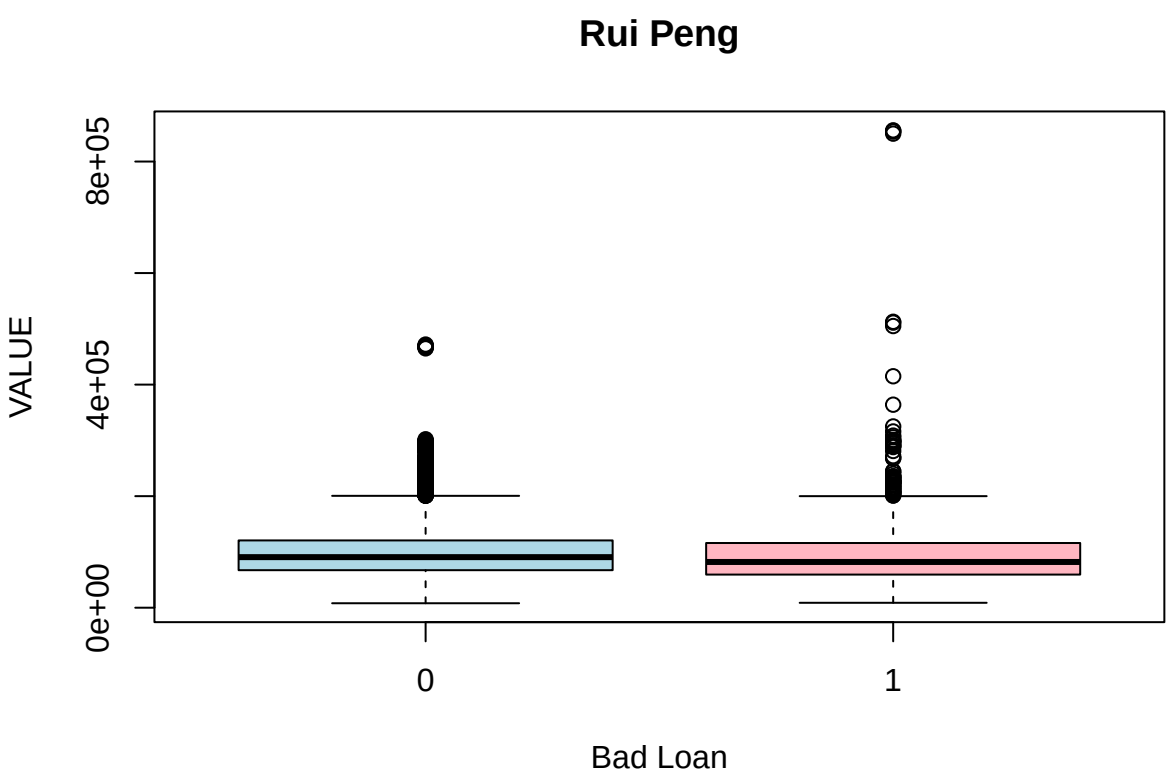
```
##   TARGET_BAD_FLAG TARGET_LOSS_AMT LOAN MORTDUE  VALUE  REASON    JOB  YOJ DEROG
## 1               1             641 1100   25860  39025 HomeImp  Other 10.5     0
## 2               1            1109 1300   70053  68400 HomeImp  Other  7.0     0
## 3               1             767 1500   13500  16700 HomeImp  Other  4.0     0
## 4               1            1425 1500      NA     NA                  NA    NA
```

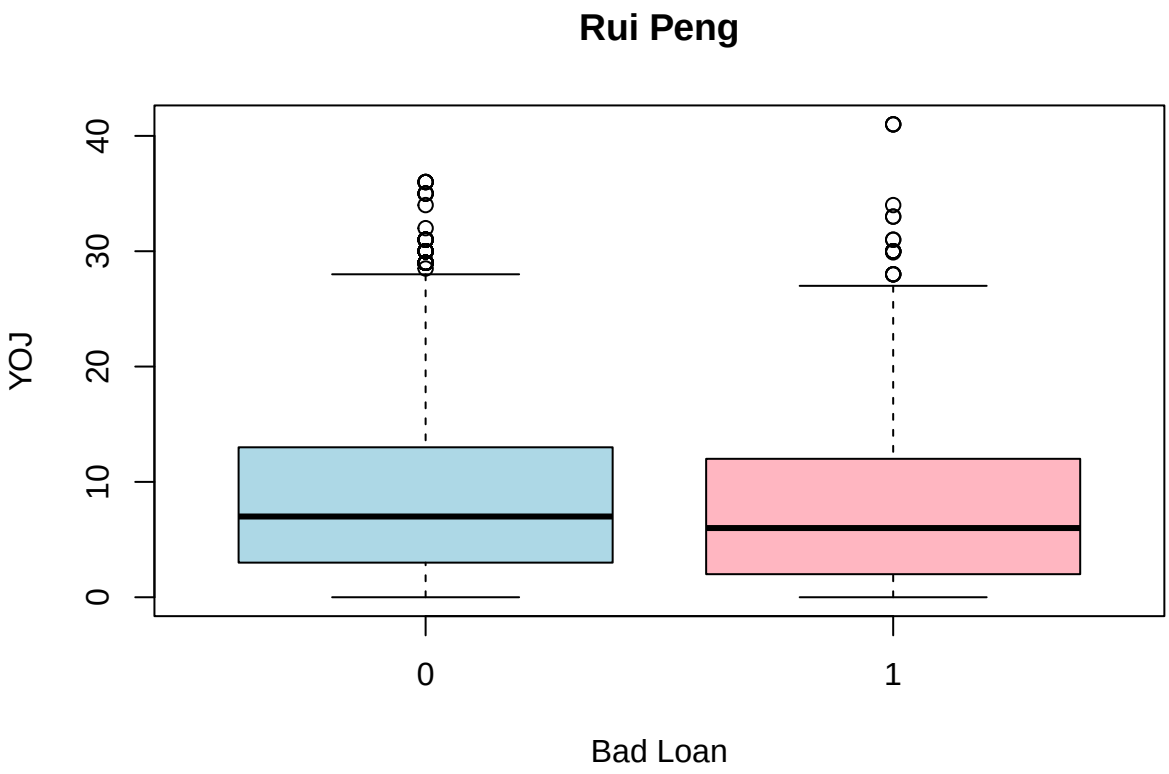
```
## 5              0             NA 1700   97800 112000 HomeImp Office  3.0     0
## 6              1            335 1700   30548  40320 HomeImp  Other  9.0     0
##   DELINQ     CLAGE NINQ CLNO  DEBTINC
## 1      0  94.36667    1    9       NA
## 2      2 121.83333    0   14       NA
## 3      0 149.46667    1   10       NA
## 4     NA        NA   NA   NA       NA
## 5      0  93.33333    0   14       NA
## 6      0 101.46600    1    8 37.11361
```

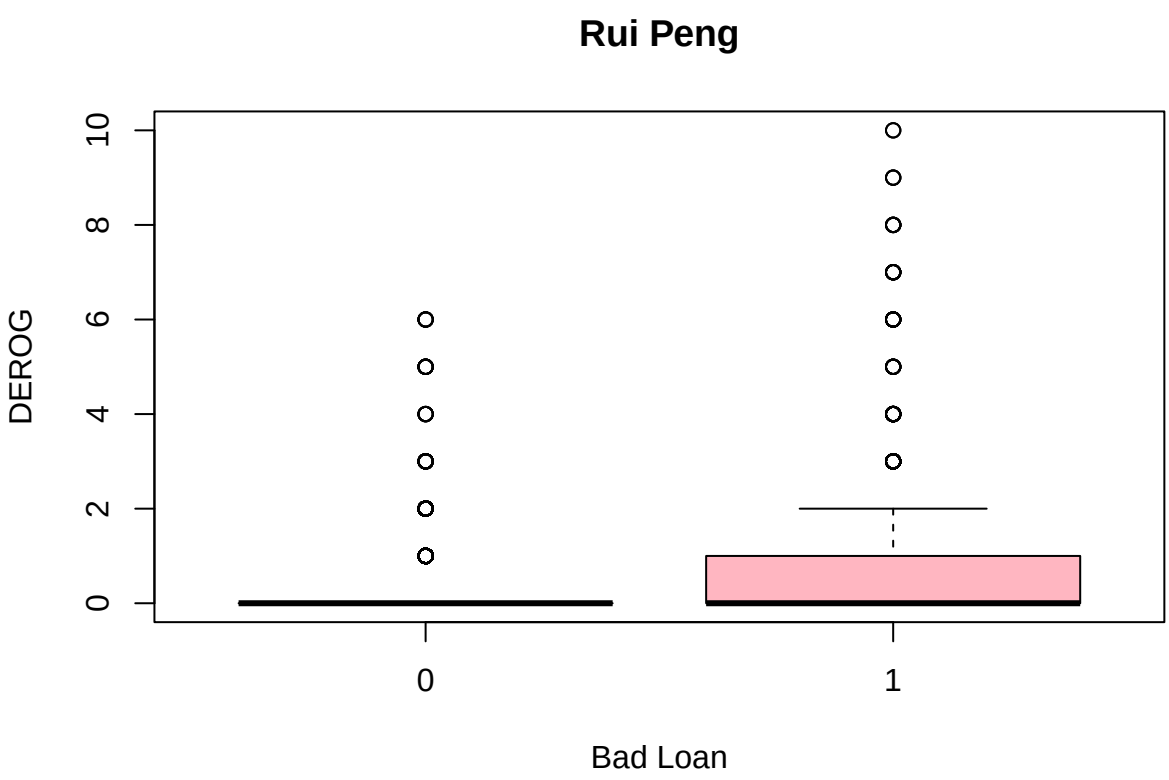
```r
#Step 2: Box-Whisker Plots
#Plot a box plot of all the numeric variables split by the grouping variable.
group_var <- df$TARGET_BAD_FLAG

numeric_vars <- c("LOAN", "MORTDUE", "VALUE", "YOJ", "DEROG",
                  "DELINQ", "CLAGE", "NINQ", "CLNO", "DEBTINC")

for (var in numeric_vars) {
  boxplot(
    df[[var]] ~ group_var,
    #The MAIN TITLE of the box plot should be set to your name
    main = "Rui Peng",
    xlab = "Bad Loan",
    ylab = var,

    #Add color to the boxes
    col = c("lightblue", "lightpink")
  )
}
```

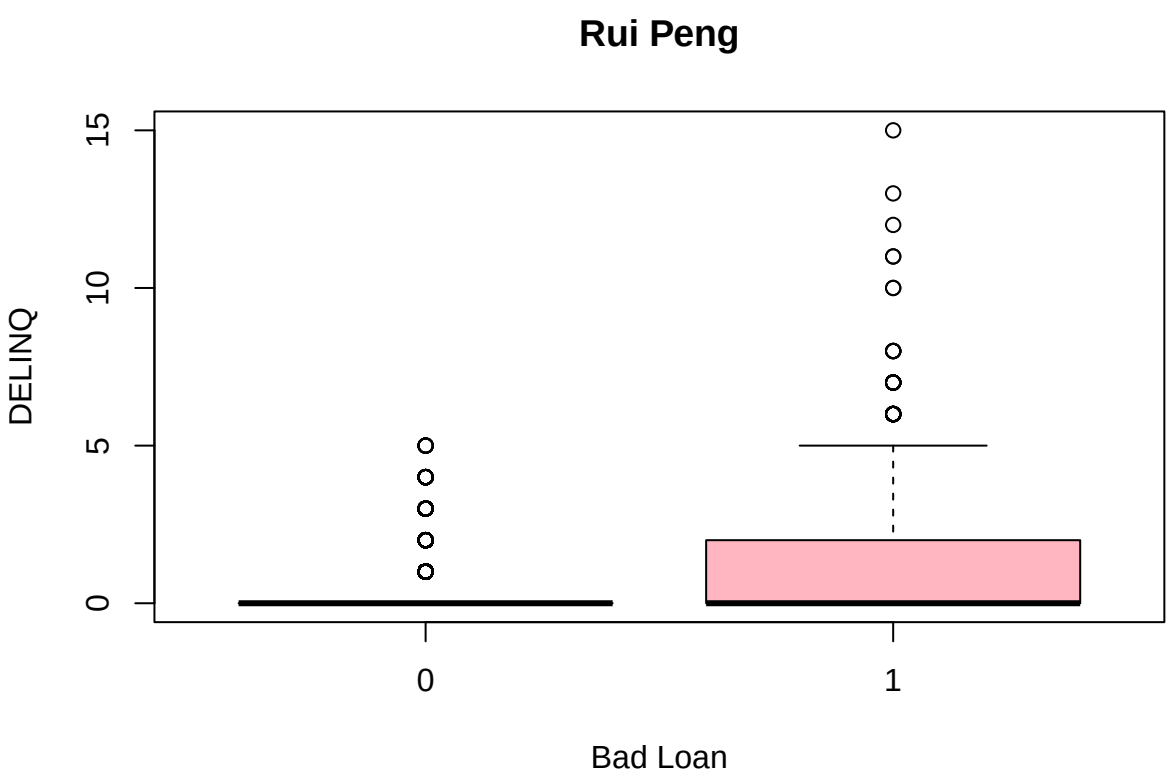

Rui Peng
LOAN
0
20000
60000
0
1
Bad Loan

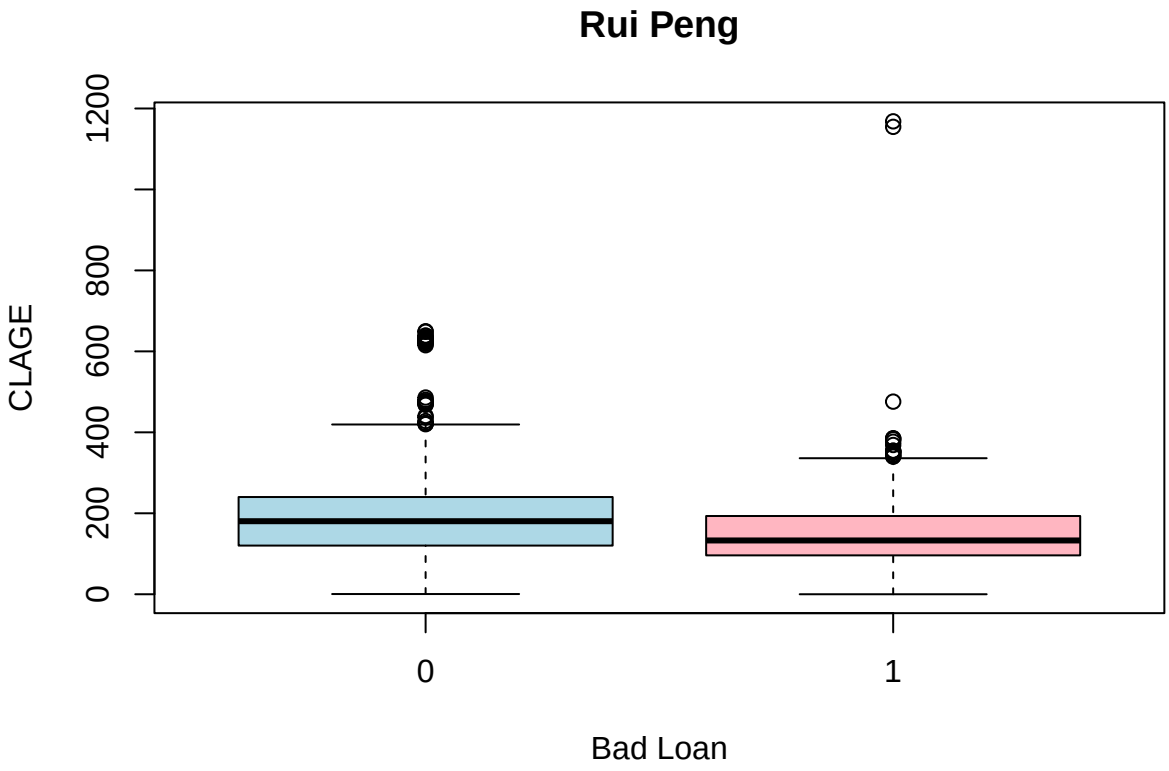

Rui Peng
MORTDUE
0e+00
1e+05
2e+05
3e+05
4e+05
0
1
Bad Loan

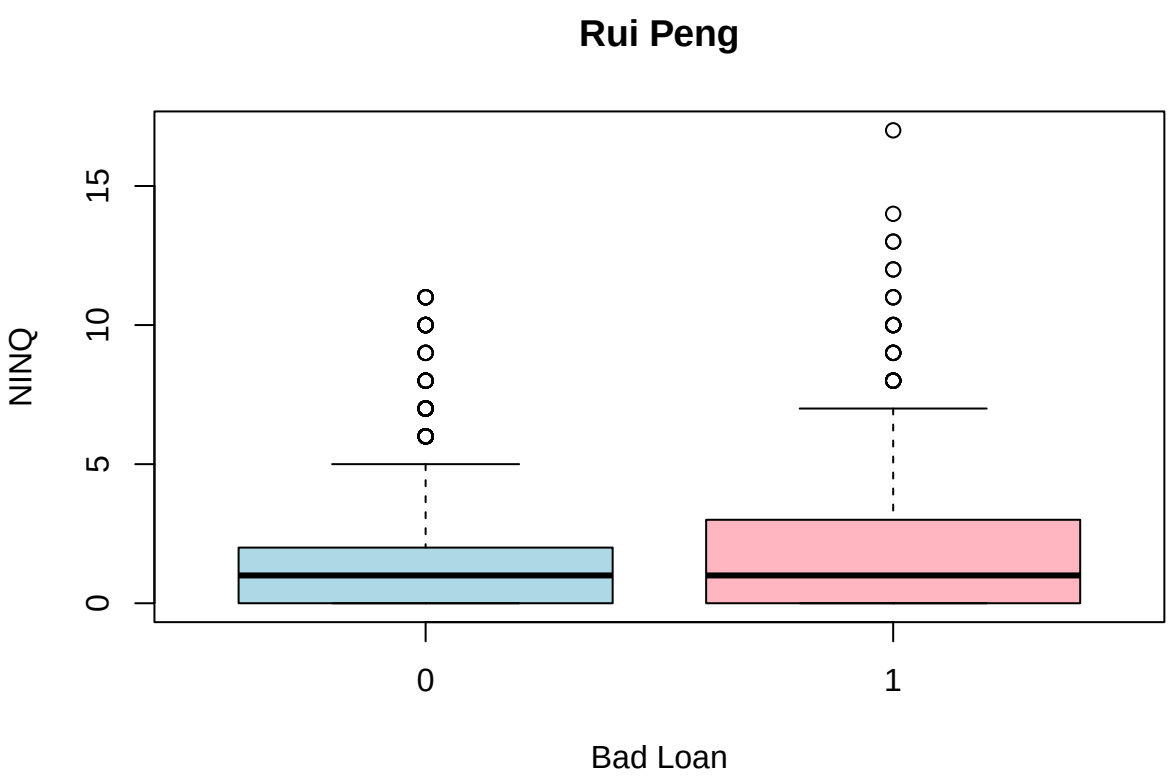

Rui Peng
VALUE
0e+00
4e+05
8e+05
0
1
Bad Loan

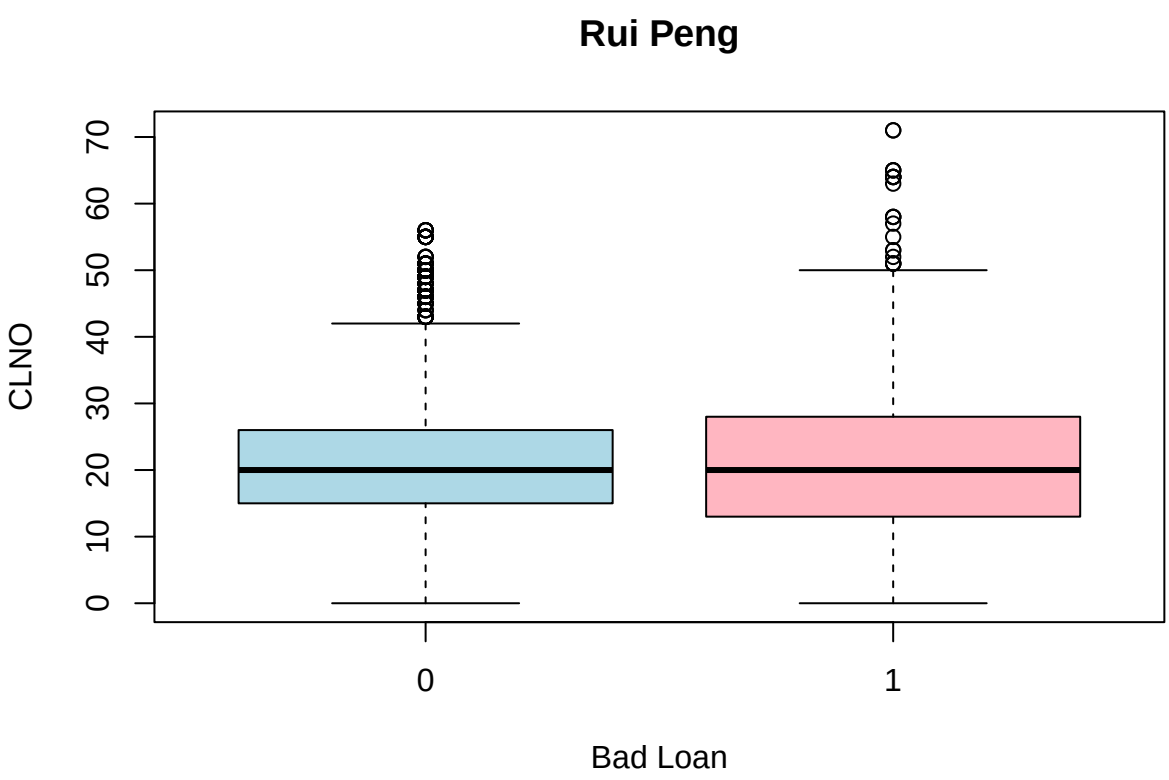

Rui Peng
YOJ
0
10
20
30
40
0
1
Bad Loan

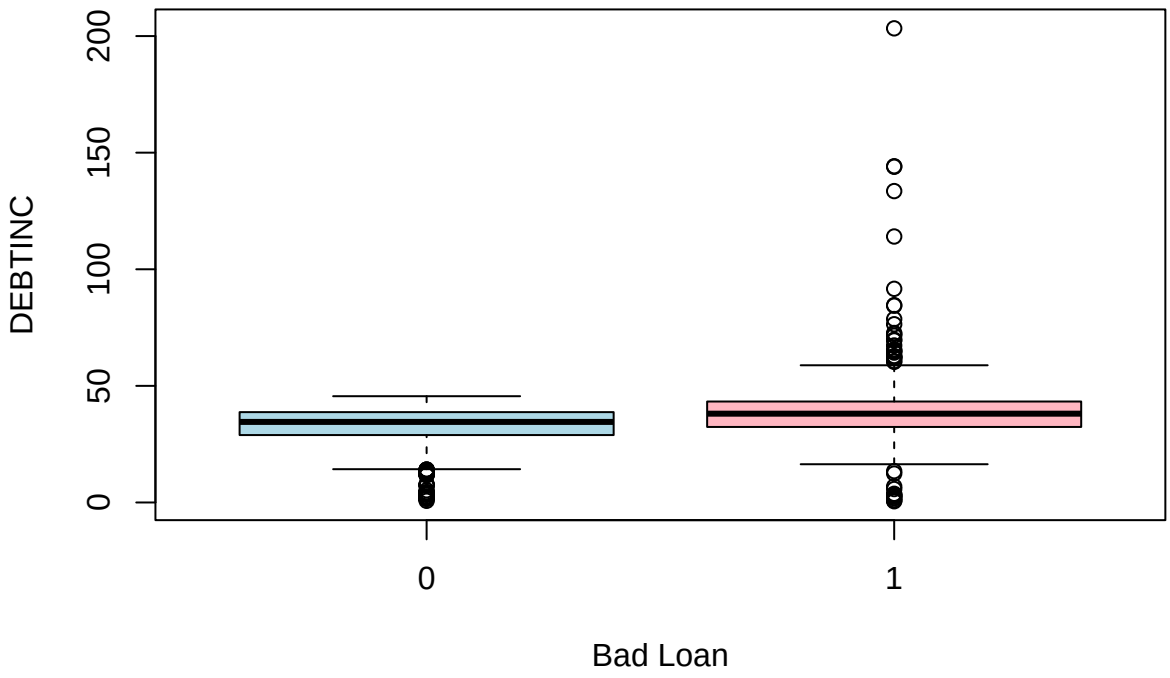

Rui Peng
DEROG
0
2
4
6
8
10
0
1
Bad Loan


Rui Peng
DELINQ
0
5
10
15
0
1
Bad Loan


Rui Peng
CLAGE
0
200
400
600
800
1200
0
1
Bad Loan

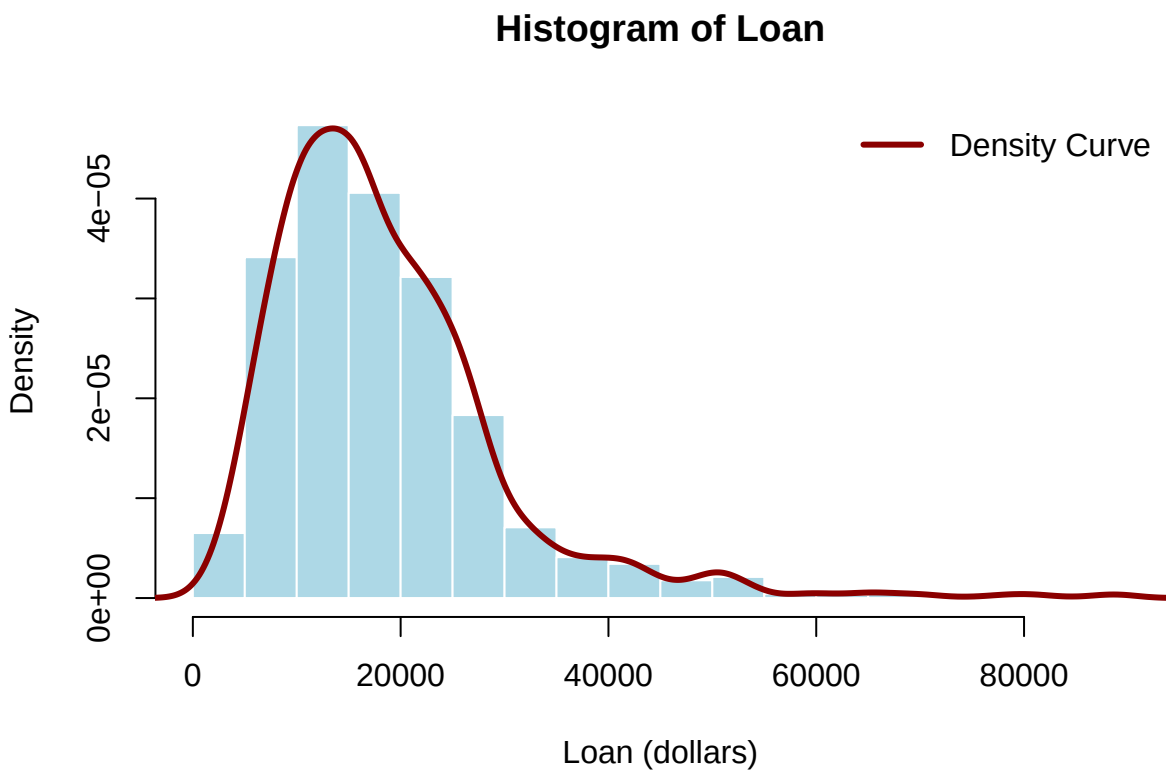

Rui Peng
NINQ
0
5
10
15
0
1
Bad Loan


Rui Peng
CLNO
0
10
20
30
40
50
60
70
0
1
Bad Loan




```
#Comment on whether or not there are any observable differences in the box plots between the two groups
#So according to my observation, those people tend to have bad loan:
#less loan amount, less mortgage due, less house value, less year of job,less credit line age
#who borrowed money more often, who have a higher debt to income ratio

#Step 3: Histograms
#Plot a histogram of at least one of the numeric variables
hist(
  df$LOAN,
  #Manually set the number of breaks to a value that makes sense
  breaks = 20,
  col = "lightblue",
  border = "white",
  xlab = "Loan (dollars)",
  ylab = "Density",
  main = "Histogram of Loan",
  freq = FALSE
)

#Superimpose a density line to the graph
lines(
  density(df$LOAN, bw = 2000),
  col = "darkred",
  lwd = 3
)
```

```
legend(
  "topright",
  legend = c("Density Curve"),
  col = "darkred",
  lwd = 3,
  bty = "n"
)
```




```
#Step 4: Impute "Fix" all the numeric variables that have missing values
#For the missing Target variables, simply set the missing values to zero

#Var 1: TARGET_LOSS_AMT
df$TARGET_LOSS_AMT[is.na(df$TARGET_LOSS_AMT)] <- 0

#Var 2: MORTDUE
summary(df$MORTDUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2063   46276   65019   73761   91488  399550     518
```

```
#The median or mean value will be useful in most cases.
median(df$MORTDUE, na.rm = TRUE)
```

```
## [1] 65019
```

```r
#Create two new variables: #One variable beginning with IMP_ and the second value beginning with M_.
df$IMP_MORTDUE <- df$MORTDUE
df$IMP_MORTDUE[ is.na(df$MORTDUE) ] = 65019
df$M_MORTDUE = is.na(df$MORTDUE) + 0

#Compute a sum for all the M_ variables
sum(df$M_MORTDUE)
```

```
## [1] 518
```

```r
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN          MORTDUE
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Min.   :  2063
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   1st Qu.: 46276
##  Median :0.0000   Median :    0   Median :16300   Median : 65019
##  Mean   :0.1995   Mean   : 2676   Mean   :18608   Mean   : 73761
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300   3rd Qu.: 91488
##  Max.   :1.0000   Max.   :78987   Max.   :89900   Max.   :399550
##                                                   NA's   :518
##      VALUE           REASON              JOB                YOJ
##  Min.   :  8000   Length:5960        Length:5960        Min.   : 0.000
##  1st Qu.: 66076   Class :character   Class :character   1st Qu.: 3.000
##  Median : 89236   Mode  :character   Mode  :character   Median : 7.000
##  Mean   :101776                                         Mean   : 8.922
##  3rd Qu.:119824                                         3rd Qu.:13.000
##  Max.   :855909                                         Max.   :41.000
##  NA's   :112                                            NA's   :515
##      DEROG             DELINQ            CLAGE             NINQ
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :   0.0   Min.   : 0.000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 115.1   1st Qu.: 0.000
##  Median : 0.0000   Median : 0.0000   Median : 173.5   Median : 1.000
##  Mean   : 0.2546   Mean   : 0.4494   Mean   : 179.8   Mean   : 1.186
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 231.6   3rd Qu.: 2.000
##  Max.   :10.0000   Max.   :15.0000   Max.   :1168.2   Max.   :17.000
##  NA's   :708       NA's   :580       NA's   :308      NA's   :510
##       CLNO          DEBTINC          IMP_MORTDUE       M_MORTDUE
##  Min.   : 0.0   Min.   :  0.5245   Min.   :  2063   Min.   :0.00000
##  1st Qu.:15.0   1st Qu.: 29.1400   1st Qu.: 48139   1st Qu.:0.00000
##  Median :20.0   Median : 34.8183   Median : 65019   Median :0.00000
##  Mean   :21.3   Mean   : 33.7799   Mean   : 73001   Mean   :0.08691
##  3rd Qu.:26.0   3rd Qu.: 39.0031   3rd Qu.: 88200   3rd Qu.:0.00000
##  Max.   :71.0   Max.   :203.3121   Max.   :399550   Max.   :1.00000
##  NA's   :222    NA's   :1267
```

```r
#Delete the original variable after it has been imputed.
df$MORTDUE = NULL
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN           VALUE
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Min.   :  8000
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   1st Qu.: 66076
```

```
##  Median :0.0000   Median :    0   Median :16300   Median : 89236
##  Mean   :0.1995   Mean   : 2676   Mean   :18608   Mean   :101776
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300   3rd Qu.:119824
##  Max.   :1.0000   Max.   :78987   Max.   :89900   Max.   :855909
##                                                   NA's   :112
##     REASON              JOB                YOJ             DEROG
##  Length:5960        Length:5960        Min.   : 0.000   Min.   : 0.0000
##  Class :character   Class :character   1st Qu.: 3.000   1st Qu.: 0.0000
##  Mode  :character   Mode  :character   Median : 7.000   Median : 0.0000
##                                        Mean   : 8.922   Mean   : 0.2546
##                                        3rd Qu.:13.000   3rd Qu.: 0.0000
##                                        Max.   :41.000   Max.   :10.0000
##                                        NA's   :515      NA's   :708
##      DELINQ            CLAGE             NINQ             CLNO
##  Min.   : 0.0000   Min.   :   0.0   Min.   : 0.000   Min.   : 0.0
##  1st Qu.: 0.0000   1st Qu.: 115.1   1st Qu.: 0.000   1st Qu.:15.0
##  Median : 0.0000   Median : 173.5   Median : 1.000   Median :20.0
##  Mean   : 0.4494   Mean   : 179.8   Mean   : 1.186   Mean   :21.3
##  3rd Qu.: 0.0000   3rd Qu.: 231.6   3rd Qu.: 2.000   3rd Qu.:26.0
##  Max.   :15.0000   Max.   :1168.2   Max.   :17.000   Max.   :71.0
##  NA's   :580       NA's   :308      NA's   :510      NA's   :222
##     DEBTINC          IMP_MORTDUE       M_MORTDUE
##  Min.   :  0.5245   Min.   :  2063   Min.   :0.00000
##  1st Qu.: 29.1400   1st Qu.: 48139   1st Qu.:0.00000
##  Median : 34.8183   Median : 65019   Median :0.00000
##  Mean   : 33.7799   Mean   : 73001   Mean   :0.08691
##  3rd Qu.: 39.0031   3rd Qu.: 88200   3rd Qu.:0.00000
##  Max.   :203.3121   Max.   :399550   Max.   :1.00000
##  NA's   :1267
```

```
#Try one complex imputation like the one described in the lectures.
#Var 3: VALUE
summary(df$VALUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    8000   66076   89236  101776  119824  855909     112
```

```
a = aggregate(x=df$VALUE, by=list(df$JOB), na.rm = TRUE, FUN = median)
a = a[ order(a$x, decreasing = TRUE), ]

df$IMP_VALUE = df$VALUE
df$IMP_VALUE[ is.na(df$VALUE) & (df$JOB == "Self")] = 130631
df$IMP_VALUE[ is.na(df$VALUE) & (df$JOB == "ProfExe")] = 110007

df$IMP_VALUE[ is.na(df$VALUE) & (df$JOB == "Mgr")] = 101258
df$IMP_VALUE[ is.na(df$VALUE) & (df$JOB == "Office")] = 89094.5

df$IMP_VALUE[ is.na(df$VALUE) & (df$JOB == "Sales")] = 84473.5
df$IMP_VALUE[ is.na(df$VALUE) & (is.na(df$JOB))] = 78227
df$IMP_VALUE[ is.na(df$VALUE) & (df$JOB == "Other")] = 76599.5

df$IMP_VALUE[ is.na(df$IMP_VALUE)] = 89236
df$M_VALUE = is.na(df$VALUE) + 0
sum(df$M_VALUE)
```

```
## [1] 112
```

```
df$VALUE = NULL
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN          REASON
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Length:5960
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   Class :character
##  Median :0.0000   Median :    0   Median :16300   Mode  :character
##  Mean   :0.1995   Mean   : 2676   Mean   :18608
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300
##  Max.   :1.0000   Max.   :78987   Max.   :89900
##
##      JOB                 YOJ             DEROG             DELINQ
##  Length:5960        Min.   : 0.000   Min.   : 0.0000   Min.   : 0.0000
##  Class :character   1st Qu.: 3.000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Mode  :character   Median : 7.000   Median : 0.0000   Median : 0.0000
##                     Mean   : 8.922   Mean   : 0.2546   Mean   : 0.4494
##                     3rd Qu.:13.000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
##                     Max.   :41.000   Max.   :10.0000   Max.   :15.0000
##                     NA's   :515      NA's   :708       NA's   :580
##      CLAGE             NINQ             CLNO         DEBTINC
##  Min.   :   0.0   Min.   : 0.000   Min.   : 0.0   Min.   :  0.5245
##  1st Qu.: 115.1   1st Qu.: 0.000   1st Qu.:15.0   1st Qu.: 29.1400
##  Median : 173.5   Median : 1.000   Median :20.0   Median : 34.8183
##  Mean   : 179.8   Mean   : 1.186   Mean   :21.3   Mean   : 33.7799
##  3rd Qu.: 231.6   3rd Qu.: 2.000   3rd Qu.:26.0   3rd Qu.: 39.0031
##  Max.   :1168.2   Max.   :17.000   Max.   :71.0   Max.   :203.3121
##  NA's   :308      NA's   :510      NA's   :222    NA's   :1267
##   IMP_MORTDUE       M_MORTDUE         IMP_VALUE         M_VALUE
##  Min.   :  2063   Min.   :0.00000   Min.   :  8000   Min.   :0.00000
##  1st Qu.: 48139   1st Qu.:0.00000   1st Qu.: 66490   1st Qu.:0.00000
##  Median : 65019   Median :0.00000   Median : 89236   Median :0.00000
##  Mean   : 73001   Mean   :0.08691   Mean   :101585   Mean   :0.01879
##  3rd Qu.: 88200   3rd Qu.:0.00000   3rd Qu.:119145   3rd Qu.:0.00000
##  Max.   :399550   Max.   :1.00000   Max.   :855909   Max.   :1.00000
##
```

```
#Var 4: YOJ
summary(df$YOJ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   3.000   7.000   8.922  13.000  41.000     515
```

```
df$IMP_YOJ <- df$YOJ
df$IMP_YOJ[is.na(df$YOJ)] <- 7
df$M_YOJ <- is.na(df$YOJ) + 0

sum(df$M_YOJ)
```

```
## [1] 515
```

```r
df$YOJ = NULL
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN          REASON
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Length:5960
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   Class :character
##  Median :0.0000   Median :    0   Median :16300   Mode  :character
##  Mean   :0.1995   Mean   : 2676   Mean   :18608
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300
##  Max.   :1.0000   Max.   :78987   Max.   :89900
##
##      JOB                DEROG              DELINQ             CLAGE
##  Length:5960        Min.   : 0.0000   Min.   : 0.0000   Min.   :   0.0
##  Class :character   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 115.1
##  Mode  :character   Median : 0.0000   Median : 0.0000   Median : 173.5
##                     Mean   : 0.2546   Mean   : 0.4494   Mean   : 179.8
##                     3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 231.6
##                     Max.   :10.0000   Max.   :15.0000   Max.   :1168.2
##                     NA's   :708       NA's   :580       NA's   :308
##       NINQ             CLNO          DEBTINC          IMP_MORTDUE
##  Min.   : 0.000   Min.   : 0.0   Min.   :  0.5245   Min.   :  2063
##  1st Qu.: 0.000   1st Qu.:15.0   1st Qu.: 29.1400   1st Qu.: 48139
##  Median : 1.000   Median :20.0   Median : 34.8183   Median : 65019
##  Mean   : 1.186   Mean   :21.3   Mean   : 33.7799   Mean   : 73001
##  3rd Qu.: 2.000   3rd Qu.:26.0   3rd Qu.: 39.0031   3rd Qu.: 88200
##  Max.   :17.000   Max.   :71.0   Max.   :203.3121   Max.   :399550
##  NA's   :510      NA's   :222    NA's   :1267
##    M_MORTDUE         IMP_VALUE         M_VALUE           IMP_YOJ
##  Min.   :0.00000   Min.   :  8000   Min.   :0.00000   Min.   : 0.000
##  1st Qu.:0.00000   1st Qu.: 66490   1st Qu.:0.00000   1st Qu.: 3.000
##  Median :0.00000   Median : 89236   Median :0.00000   Median : 7.000
##  Mean   :0.08691   Mean   :101585   Mean   :0.01879   Mean   : 8.756
##  3rd Qu.:0.00000   3rd Qu.:119145   3rd Qu.:0.00000   3rd Qu.:12.000
##  Max.   :1.00000   Max.   :855909   Max.   :1.00000   Max.   :41.000
##
##      M_YOJ
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.08641
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
```

```r
#Var 5: DEROG
summary(df$DEROG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.2546  0.0000 10.0000     708
```

```r
df$IMP_DEROG <- df$DEROG
df$IMP_DEROG[is.na(df$DEROG)] <- 0
```

```r
df$M_DEROG <- is.na(df$DEROG) + 0

sum(df$M_DEROG)
```

```
## [1] 708
```

```r
df$DEROG = NULL
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN          REASON
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Length:5960
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   Class :character
##  Median :0.0000   Median :    0   Median :16300   Mode  :character
##  Mean   :0.1995   Mean   : 2676   Mean   :18608
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300
##  Max.   :1.0000   Max.   :78987   Max.   :89900
##
##      JOB                DELINQ            CLAGE             NINQ
##  Length:5960        Min.   : 0.0000   Min.   :   0.0   Min.   : 0.000
##  Class :character   1st Qu.: 0.0000   1st Qu.: 115.1   1st Qu.: 0.000
##  Mode  :character   Median : 0.0000   Median : 173.5   Median : 1.000
##                     Mean   : 0.4494   Mean   : 179.8   Mean   : 1.186
##                     3rd Qu.: 0.0000   3rd Qu.: 231.6   3rd Qu.: 2.000
##                     Max.   :15.0000   Max.   :1168.2   Max.   :17.000
##                     NA's   :580       NA's   :308      NA's   :510
##       CLNO          DEBTINC          IMP_MORTDUE       M_MORTDUE
##  Min.   : 0.0   Min.   :  0.5245   Min.   :  2063   Min.   :0.00000
##  1st Qu.:15.0   1st Qu.: 29.1400   1st Qu.: 48139   1st Qu.:0.00000
##  Median :20.0   Median : 34.8183   Median : 65019   Median :0.00000
##  Mean   :21.3   Mean   : 33.7799   Mean   : 73001   Mean   :0.08691
##  3rd Qu.:26.0   3rd Qu.: 39.0031   3rd Qu.: 88200   3rd Qu.:0.00000
##  Max.   :71.0   Max.   :203.3121   Max.   :399550   Max.   :1.00000
##  NA's   :222    NA's   :1267
##    IMP_VALUE         M_VALUE           IMP_YOJ           M_YOJ
##  Min.   :  8000   Min.   :0.00000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.: 66490   1st Qu.:0.00000   1st Qu.: 3.000   1st Qu.:0.00000
##  Median : 89236   Median :0.00000   Median : 7.000   Median :0.00000
##  Mean   :101585   Mean   :0.01879   Mean   : 8.756   Mean   :0.08641
##  3rd Qu.:119145   3rd Qu.:0.00000   3rd Qu.:12.000   3rd Qu.:0.00000
##  Max.   :855909   Max.   :1.00000   Max.   :41.000   Max.   :1.00000
##
##    IMP_DEROG          M_DEROG
##  Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 0.0000   Median :0.0000
##  Mean   : 0.2243   Mean   :0.1188
##  3rd Qu.: 0.0000   3rd Qu.:0.0000
##  Max.   :10.0000   Max.   :1.0000
##
```

```r
#Var 6: DELINQ
summary(df$DELINQ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.4494  0.0000 15.0000     580
```

```
df$IMP_DELINQ <- df$DELINQ
df$IMP_DELINQ[is.na(df$DELINQ)] <- 0
df$M_DELINQ <- is.na(df$DELINQ) + 0

sum(df$M_DELINQ)
```

```
## [1] 580
```

```
df$DELINQ = NULL
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN          REASON
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Length:5960
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   Class :character
##  Median :0.0000   Median :    0   Median :16300   Mode  :character
##  Mean   :0.1995   Mean   : 2676   Mean   :18608
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300
##  Max.   :1.0000   Max.   :78987   Max.   :89900
##
##      JOB                CLAGE             NINQ             CLNO
##  Length:5960        Min.   :   0.0   Min.   : 0.000   Min.   : 0.0
##  Class :character   1st Qu.: 115.1   1st Qu.: 0.000   1st Qu.:15.0
##  Mode  :character   Median : 173.5   Median : 1.000   Median :20.0
##                     Mean   : 179.8   Mean   : 1.186   Mean   :21.3
##                     3rd Qu.: 231.6   3rd Qu.: 2.000   3rd Qu.:26.0
##                     Max.   :1168.2   Max.   :17.000   Max.   :71.0
##                     NA's   :308      NA's   :510      NA's   :222
##     DEBTINC          IMP_MORTDUE       M_MORTDUE         IMP_VALUE
##  Min.   :  0.5245   Min.   :  2063   Min.   :0.00000   Min.   :  8000
##  1st Qu.: 29.1400   1st Qu.: 48139   1st Qu.:0.00000   1st Qu.: 66490
##  Median : 34.8183   Median : 65019   Median :0.00000   Median : 89236
##  Mean   : 33.7799   Mean   : 73001   Mean   :0.08691   Mean   :101585
##  3rd Qu.: 39.0031   3rd Qu.: 88200   3rd Qu.:0.00000   3rd Qu.:119145
##  Max.   :203.3121   Max.   :399550   Max.   :1.00000   Max.   :855909
##  NA's   :1267
##     M_VALUE           IMP_YOJ           M_YOJ           IMP_DEROG
##  Min.   :0.00000   Min.   : 0.000   Min.   :0.00000   Min.   : 0.0000
##  1st Qu.:0.00000   1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.: 0.0000
##  Median :0.00000   Median : 7.000   Median :0.00000   Median : 0.0000
##  Mean   :0.01879   Mean   : 8.756   Mean   :0.08641   Mean   : 0.2243
##  3rd Qu.:0.00000   3rd Qu.:12.000   3rd Qu.:0.00000   3rd Qu.: 0.0000
##  Max.   :1.00000   Max.   :41.000   Max.   :1.00000   Max.   :10.0000
##
##     M_DEROG         IMP_DELINQ         M_DELINQ
##  Min.   :0.0000   Min.   : 0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.00000
##  Median :0.0000   Median : 0.0000   Median :0.00000
##  Mean   :0.1188   Mean   : 0.4057   Mean   :0.09732
##  3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :15.0000   Max.   :1.00000
##
```

```
#Var 7: CLAGE
summary(df$CLAGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   115.1   173.5   179.8   231.6  1168.2     308
```

```
df$IMP_CLAGE <- df$CLAGE
df$IMP_CLAGE[is.na(df$CLAGE)] <- 173.5
df$M_CLAGE <- is.na(df$CLAGE) + 0

sum(df$M_CLAGE)
```

```
## [1] 308
```

```
df$CLAGE = NULL
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN          REASON
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Length:5960
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   Class :character
##  Median :0.0000   Median :    0   Median :16300   Mode  :character
##  Mean   :0.1995   Mean   : 2676   Mean   :18608
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300
##  Max.   :1.0000   Max.   :78987   Max.   :89900
##
##      JOB                 NINQ             CLNO         DEBTINC
##  Length:5960        Min.   : 0.000   Min.   : 0.0   Min.   :  0.5245
##  Class :character   1st Qu.: 0.000   1st Qu.:15.0   1st Qu.: 29.1400
##  Mode  :character   Median : 1.000   Median :20.0   Median : 34.8183
##                     Mean   : 1.186   Mean   :21.3   Mean   : 33.7799
##                     3rd Qu.: 2.000   3rd Qu.:26.0   3rd Qu.: 39.0031
##                     Max.   :17.000   Max.   :71.0   Max.   :203.3121
##                     NA's   :510      NA's   :222    NA's   :1267
##   IMP_MORTDUE       M_MORTDUE         IMP_VALUE         M_VALUE
##  Min.   :  2063   Min.   :0.00000   Min.   :  8000   Min.   :0.00000
##  1st Qu.: 48139   1st Qu.:0.00000   1st Qu.: 66490   1st Qu.:0.00000
##  Median : 65019   Median :0.00000   Median : 89236   Median :0.00000
##  Mean   : 73001   Mean   :0.08691   Mean   :101585   Mean   :0.01879
##  3rd Qu.: 88200   3rd Qu.:0.00000   3rd Qu.:119145   3rd Qu.:0.00000
##  Max.   :399550   Max.   :1.00000   Max.   :855909   Max.   :1.00000
##
##     IMP_YOJ           M_YOJ           IMP_DEROG          M_DEROG
##  Min.   : 0.000   Min.   :0.00000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 7.000   Median :0.00000   Median : 0.0000   Median :0.0000
##  Mean   : 8.756   Mean   :0.08641   Mean   : 0.2243   Mean   :0.1188
##  3rd Qu.:12.000   3rd Qu.:0.00000   3rd Qu.: 0.0000   3rd Qu.:0.0000
##  Max.   :41.000   Max.   :1.00000   Max.   :10.0000   Max.   :1.0000
##
##    IMP_DELINQ         M_DELINQ         IMP_CLAGE         M_CLAGE
##  Min.   : 0.0000   Min.   :0.00000   Min.   :   0.0   Min.   :0.00000
##  1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.: 117.4   1st Qu.:0.00000
```

```
##  Median : 0.0000   Median :0.00000   Median : 173.5   Median :0.00000
##  Mean   : 0.4057   Mean   :0.09732   Mean   : 179.4   Mean   :0.05168
##  3rd Qu.: 0.0000   3rd Qu.:0.00000   3rd Qu.: 227.1   3rd Qu.:0.00000
##  Max.   :15.0000   Max.   :1.00000   Max.   :1168.2   Max.   :1.00000
##
```

```
#Var 8: NINQ
summary(df$NINQ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.000   1.000   1.186   2.000  17.000     510
```

```
df$IMP_NINQ <- df$NINQ
df$IMP_NINQ[is.na(df$NINQ)] <- 1
df$M_NINQ <- is.na(df$NINQ) + 0

sum(df$M_NINQ)
```

```
## [1] 510
```

```
df$NINQ = NULL
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN          REASON
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Length:5960
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   Class :character
##  Median :0.0000   Median :    0   Median :16300   Mode  :character
##  Mean   :0.1995   Mean   : 2676   Mean   :18608
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300
##  Max.   :1.0000   Max.   :78987   Max.   :89900
##
##      JOB                CLNO         DEBTINC          IMP_MORTDUE
##  Length:5960        Min.   : 0.0   Min.   :  0.5245   Min.   :  2063
##  Class :character   1st Qu.:15.0   1st Qu.: 29.1400   1st Qu.: 48139
##  Mode  :character   Median :20.0   Median : 34.8183   Median : 65019
##                     Mean   :21.3   Mean   : 33.7799   Mean   : 73001
##                     3rd Qu.:26.0   3rd Qu.: 39.0031   3rd Qu.: 88200
##                     Max.   :71.0   Max.   :203.3121   Max.   :399550
##                     NA's   :222    NA's   :1267
##    M_MORTDUE         IMP_VALUE         M_VALUE           IMP_YOJ
##  Min.   :0.00000   Min.   :  8000   Min.   :0.00000   Min.   : 0.000
##  1st Qu.:0.00000   1st Qu.: 66490   1st Qu.:0.00000   1st Qu.: 3.000
##  Median :0.00000   Median : 89236   Median :0.00000   Median : 7.000
##  Mean   :0.08691   Mean   :101585   Mean   :0.01879   Mean   : 8.756
##  3rd Qu.:0.00000   3rd Qu.:119145   3rd Qu.:0.00000   3rd Qu.:12.000
##  Max.   :1.00000   Max.   :855909   Max.   :1.00000   Max.   :41.000
##
##      M_YOJ           IMP_DEROG          M_DEROG         IMP_DELINQ
##  Min.   :0.00000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median :0.00000   Median : 0.0000   Median :0.0000   Median : 0.0000
##  Mean   :0.08641   Mean   : 0.2243   Mean   :0.1188   Mean   : 0.4057
```

```
##  3rd Qu.:0.00000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000
##  Max.   :1.00000   Max.   :10.0000   Max.   :1.0000   Max.   :15.0000
##
##     M_DELINQ         IMP_CLAGE         M_CLAGE           IMP_NINQ
##  Min.   :0.00000   Min.   :   0.0   Min.   :0.00000   Min.   : 0.00
##  1st Qu.:0.00000   1st Qu.: 117.4   1st Qu.:0.00000   1st Qu.: 0.00
##  Median :0.00000   Median : 173.5   Median :0.00000   Median : 1.00
##  Mean   :0.09732   Mean   : 179.4   Mean   :0.05168   Mean   : 1.17
##  3rd Qu.:0.00000   3rd Qu.: 227.1   3rd Qu.:0.00000   3rd Qu.: 2.00
##  Max.   :1.00000   Max.   :1168.2   Max.   :1.00000   Max.   :17.00
##
##      M_NINQ
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.08557
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
```

```
#Var 9: CLNO
summary(df$CLNO)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0    15.0    20.0    21.3    26.0    71.0     222
```

```
df$IMP_CLNO <- df$CLNO
df$IMP_CLNO[is.na(df$CLNO)] <- 20
df$M_CLNO <- is.na(df$CLNO) + 0

sum(df$M_CLNO)
```

```
## [1] 222
```

```
df$CLNO = NULL
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN          REASON
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Length:5960
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   Class :character
##  Median :0.0000   Median :    0   Median :16300   Mode  :character
##  Mean   :0.1995   Mean   : 2676   Mean   :18608
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300
##  Max.   :1.0000   Max.   :78987   Max.   :89900
##
##      JOB               DEBTINC          IMP_MORTDUE       M_MORTDUE
##  Length:5960        Min.   :  0.5245   Min.   :  2063   Min.   :0.00000
##  Class :character   1st Qu.: 29.1400   1st Qu.: 48139   1st Qu.:0.00000
##  Mode  :character   Median : 34.8183   Median : 65019   Median :0.00000
##                     Mean   : 33.7799   Mean   : 73001   Mean   :0.08691
##                     3rd Qu.: 39.0031   3rd Qu.: 88200   3rd Qu.:0.00000
##                     Max.   :203.3121   Max.   :399550   Max.   :1.00000
```

```
##                   NA's   :1267
##    IMP_VALUE         M_VALUE           IMP_YOJ           M_YOJ
##  Min.   :  8000   Min.   :0.00000   Min.   : 0.000   Min.   :0.00000
##  1st Qu.: 66490   1st Qu.:0.00000   1st Qu.: 3.000   1st Qu.:0.00000
##  Median : 89236   Median :0.00000   Median : 7.000   Median :0.00000
##  Mean   :101585   Mean   :0.01879   Mean   : 8.756   Mean   :0.08641
##  3rd Qu.:119145   3rd Qu.:0.00000   3rd Qu.:12.000   3rd Qu.:0.00000
##  Max.   :855909   Max.   :1.00000   Max.   :41.000   Max.   :1.00000
##
##    IMP_DEROG          M_DEROG         IMP_DELINQ         M_DELINQ
##  Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000   Min.   :0.00000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.00000
##  Median : 0.0000   Median :0.0000   Median : 0.0000   Median :0.00000
##  Mean   : 0.2243   Mean   :0.1188   Mean   : 0.4057   Mean   :0.09732
##  3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.:0.00000
##  Max.   :10.0000   Max.   :1.0000   Max.   :15.0000   Max.   :1.00000
##
##    IMP_CLAGE         M_CLAGE           IMP_NINQ         M_NINQ
##  Min.   :   0.0   Min.   :0.00000   Min.   : 0.00   Min.   :0.00000
##  1st Qu.: 117.4   1st Qu.:0.00000   1st Qu.: 0.00   1st Qu.:0.00000
##  Median : 173.5   Median :0.00000   Median : 1.00   Median :0.00000
##  Mean   : 179.4   Mean   :0.05168   Mean   : 1.17   Mean   :0.08557
##  3rd Qu.: 227.1   3rd Qu.:0.00000   3rd Qu.: 2.00   3rd Qu.:0.00000
##  Max.   :1168.2   Max.   :1.00000   Max.   :17.00   Max.   :1.00000
##
##     IMP_CLNO         M_CLNO
##  Min.   : 0.00   Min.   :0.00000
##  1st Qu.:15.00   1st Qu.:0.00000
##  Median :20.00   Median :0.00000
##  Mean   :21.25   Mean   :0.03725
##  3rd Qu.:26.00   3rd Qu.:0.00000
##  Max.   :71.00   Max.   :1.00000
##
```

```
#Var 10: DEBTINC
summary(df$DEBTINC)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##   0.5245  29.1400  34.8183  33.7799  39.0031 203.3121     1267
```

```
df$IMP_DEBTINC <- df$DEBTINC
df$IMP_DEBTINC[is.na(df$DEBTINC)] <- 34.8183
df$M_DEBTINC <- is.na(df$DEBTINC) + 0

sum(df$M_DEBTINC)
```

```
## [1] 1267
```

```
df$DEBTINC = NULL

#Run a summary to prove that all the variables have been imputed
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN          REASON
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Length:5960
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   Class :character
##  Median :0.0000   Median :    0   Median :16300   Mode  :character
##  Mean   :0.1995   Mean   : 2676   Mean   :18608
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300
##  Max.   :1.0000   Max.   :78987   Max.   :89900
##      JOB             IMP_MORTDUE       M_MORTDUE         IMP_VALUE
##  Length:5960        Min.   :  2063   Min.   :0.00000   Min.   :  8000
##  Class :character   1st Qu.: 48139   1st Qu.:0.00000   1st Qu.: 66490
##  Mode  :character   Median : 65019   Median :0.00000   Median : 89236
##                     Mean   : 73001   Mean   :0.08691   Mean   :101585
##                     3rd Qu.: 88200   3rd Qu.:0.00000   3rd Qu.:119145
##                     Max.   :399550   Max.   :1.00000   Max.   :855909
##     M_VALUE           IMP_YOJ           M_YOJ           IMP_DEROG
##  Min.   :0.00000   Min.   : 0.000   Min.   :0.00000   Min.   : 0.0000
##  1st Qu.:0.00000   1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.: 0.0000
##  Median :0.00000   Median : 7.000   Median :0.00000   Median : 0.0000
##  Mean   :0.01879   Mean   : 8.756   Mean   :0.08641   Mean   : 0.2243
##  3rd Qu.:0.00000   3rd Qu.:12.000   3rd Qu.:0.00000   3rd Qu.: 0.0000
##  Max.   :1.00000   Max.   :41.000   Max.   :1.00000   Max.   :10.0000
##     M_DEROG          IMP_DELINQ         M_DELINQ         IMP_CLAGE
##  Min.   :0.0000   Min.   : 0.0000   Min.   :0.00000   Min.   :   0.0
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.: 117.4
##  Median :0.0000   Median : 0.0000   Median :0.00000   Median : 173.5
##  Mean   :0.1188   Mean   : 0.4057   Mean   :0.09732   Mean   : 179.4
##  3rd Qu.:0.0000   3rd Qu.: 0.0000   3rd Qu.:0.00000   3rd Qu.: 227.1
##  Max.   :1.0000   Max.   :15.0000   Max.   :1.00000   Max.   :1168.2
##     M_CLAGE           IMP_NINQ         M_NINQ           IMP_CLNO
##  Min.   :0.00000   Min.   : 0.00   Min.   :0.00000   Min.   : 0.00
##  1st Qu.:0.00000   1st Qu.: 0.00   1st Qu.:0.00000   1st Qu.:15.00
##  Median :0.00000   Median : 1.00   Median :0.00000   Median :20.00
##  Mean   :0.05168   Mean   : 1.17   Mean   :0.08557   Mean   :21.25
##  3rd Qu.:0.00000   3rd Qu.: 2.00   3rd Qu.:0.00000   3rd Qu.:26.00
##  Max.   :1.00000   Max.   :17.00   Max.   :1.00000   Max.   :71.00
##      M_CLNO          IMP_DEBTINC         M_DEBTINC
##  Min.   :0.00000   Min.   :  0.5245   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.: 30.7632   1st Qu.:0.0000
##  Median :0.00000   Median : 34.8183   Median :0.0000
##  Mean   :0.03725   Mean   : 34.0007   Mean   :0.2126
##  3rd Qu.:0.00000   3rd Qu.: 37.9499   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :203.3122   Max.   :1.0000
```

```
#Step 5: One Hot Encoding
#For char/category variables, perform one hot encoding. For this create a Flag for each categories.

#Char 1: REASON
table(df$REASON)
```

```
##
##         DebtCon HomeImp
##     252    3928    1780
```

```r
df$FLAG.Reason.DebtCon = (df$REASON == "DebtCon") + 0
df$FLAG.Reason.HomeImp = (df$REASON == "HomeImp") + 0

#Delete the original class variable
df$REASON = NULL

#Run a summary to show that the category variables have been replaced by Flag variables.
summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN           JOB
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Length:5960
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   Class :character
##  Median :0.0000   Median :    0   Median :16300   Mode  :character
##  Mean   :0.1995   Mean   : 2676   Mean   :18608
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300
##  Max.   :1.0000   Max.   :78987   Max.   :89900
##   IMP_MORTDUE       M_MORTDUE         IMP_VALUE         M_VALUE
##  Min.   :  2063   Min.   :0.00000   Min.   :  8000   Min.   :0.00000
##  1st Qu.: 48139   1st Qu.:0.00000   1st Qu.: 66490   1st Qu.:0.00000
##  Median : 65019   Median :0.00000   Median : 89236   Median :0.00000
##  Mean   : 73001   Mean   :0.08691   Mean   :101585   Mean   :0.01879
##  3rd Qu.: 88200   3rd Qu.:0.00000   3rd Qu.:119145   3rd Qu.:0.00000
##  Max.   :399550   Max.   :1.00000   Max.   :855909   Max.   :1.00000
##     IMP_YOJ           M_YOJ           IMP_DEROG          M_DEROG
##  Min.   : 0.000   Min.   :0.00000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 3.000   1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 7.000   Median :0.00000   Median : 0.0000   Median :0.0000
##  Mean   : 8.756   Mean   :0.08641   Mean   : 0.2243   Mean   :0.1188
##  3rd Qu.:12.000   3rd Qu.:0.00000   3rd Qu.: 0.0000   3rd Qu.:0.0000
##  Max.   :41.000   Max.   :1.00000   Max.   :10.0000   Max.   :1.0000
##    IMP_DELINQ         M_DELINQ         IMP_CLAGE         M_CLAGE
##  Min.   : 0.0000   Min.   :0.00000   Min.   :   0.0   Min.   :0.00000
##  1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.: 117.4   1st Qu.:0.00000
##  Median : 0.0000   Median :0.00000   Median : 173.5   Median :0.00000
##  Mean   : 0.4057   Mean   :0.09732   Mean   : 179.4   Mean   :0.05168
##  3rd Qu.: 0.0000   3rd Qu.:0.00000   3rd Qu.: 227.1   3rd Qu.:0.00000
##  Max.   :15.0000   Max.   :1.00000   Max.   :1168.2   Max.   :1.00000
##     IMP_NINQ          M_NINQ           IMP_CLNO          M_CLNO
##  Min.   : 0.00   Min.   :0.00000   Min.   : 0.00   Min.   :0.00000
##  1st Qu.: 0.00   1st Qu.:0.00000   1st Qu.:15.00   1st Qu.:0.00000
##  Median : 1.00   Median :0.00000   Median :20.00   Median :0.00000
##  Mean   : 1.17   Mean   :0.08557   Mean   :21.25   Mean   :0.03725
##  3rd Qu.: 2.00   3rd Qu.:0.00000   3rd Qu.:26.00   3rd Qu.:0.00000
##  Max.   :17.00   Max.   :1.00000   Max.   :71.00   Max.   :1.00000
##   IMP_DEBTINC         M_DEBTINC      FLAG.Reason.DebtCon FLAG.Reason.HomeImp
##  Min.   :  0.5245   Min.   :0.0000   Min.   :0.0000      Min.   :0.0000
##  1st Qu.: 30.7632   1st Qu.:0.0000   1st Qu.:0.0000      1st Qu.:0.0000
##  Median : 34.8183   Median :0.0000   Median :1.0000      Median :0.0000
##  Mean   : 34.0007   Mean   :0.2126   Mean   :0.6591      Mean   :0.2987
##  3rd Qu.: 37.9499   3rd Qu.:0.0000   3rd Qu.:1.0000      3rd Qu.:1.0000
##  Max.   :203.3122   Max.   :1.0000   Max.   :1.0000      Max.   :1.0000
```

```r
#Char 2: JOB
table(df$JOB)
```

```
##
##             Mgr  Office   Other ProfExe   Sales    Self
##     279     767     948    2388    1276     109     193
```

```r
df$FLAG.Job.Mgr = (df$JOB == "Mgr") + 0
df$FLAG.Job.Office = (df$JOB == "Office") + 0

df$FLAG.Job.Other = (df$JOB == "Other") + 0
df$FLAG.Job.ProfExe = (df$JOB == "ProfExe") + 0

df$FLAG.Job.Sales = (df$JOB == "Sales") + 0
df$FLAG.Job.Self = (df$JOB == "Self") + 0

df$FLAG.Job.Salary = (df$JOB %in% c("Self", "ProfExe", "Mgr")) + 0
df$JOB = NULL

summary(df)
```

```
##  TARGET_BAD_FLAG  TARGET_LOSS_AMT      LOAN        IMP_MORTDUE
##  Min.   :0.0000   Min.   :    0   Min.   : 1100   Min.   :  2063
##  1st Qu.:0.0000   1st Qu.:    0   1st Qu.:11100   1st Qu.: 48139
##  Median :0.0000   Median :    0   Median :16300   Median : 65019
##  Mean   :0.1995   Mean   : 2676   Mean   :18608   Mean   : 73001
##  3rd Qu.:0.0000   3rd Qu.:    0   3rd Qu.:23300   3rd Qu.: 88200
##  Max.   :1.0000   Max.   :78987   Max.   :89900   Max.   :399550
##    M_MORTDUE         IMP_VALUE         M_VALUE           IMP_YOJ
##  Min.   :0.00000   Min.   :  8000   Min.   :0.00000   Min.   : 0.000
##  1st Qu.:0.00000   1st Qu.: 66490   1st Qu.:0.00000   1st Qu.: 3.000
##  Median :0.00000   Median : 89236   Median :0.00000   Median : 7.000
##  Mean   :0.08691   Mean   :101585   Mean   :0.01879   Mean   : 8.756
##  3rd Qu.:0.00000   3rd Qu.:119145   3rd Qu.:0.00000   3rd Qu.:12.000
##  Max.   :1.00000   Max.   :855909   Max.   :1.00000   Max.   :41.000
##      M_YOJ           IMP_DEROG          M_DEROG         IMP_DELINQ
##  Min.   :0.00000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median :0.00000   Median : 0.0000   Median :0.0000   Median : 0.0000
##  Mean   :0.08641   Mean   : 0.2243   Mean   :0.1188   Mean   : 0.4057
##  3rd Qu.:0.00000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000
##  Max.   :1.00000   Max.   :10.0000   Max.   :1.0000   Max.   :15.0000
##     M_DELINQ         IMP_CLAGE         M_CLAGE           IMP_NINQ
##  Min.   :0.00000   Min.   :   0.0   Min.   :0.00000   Min.   : 0.00
##  1st Qu.:0.00000   1st Qu.: 117.4   1st Qu.:0.00000   1st Qu.: 0.00
##  Median :0.00000   Median : 173.5   Median :0.00000   Median : 1.00
##  Mean   :0.09732   Mean   : 179.4   Mean   :0.05168   Mean   : 1.17
##  3rd Qu.:0.00000   3rd Qu.: 227.1   3rd Qu.:0.00000   3rd Qu.: 2.00
##  Max.   :1.00000   Max.   :1168.2   Max.   :1.00000   Max.   :17.00
##      M_NINQ           IMP_CLNO          M_CLNO         IMP_DEBTINC
##  Min.   :0.00000   Min.   : 0.00   Min.   :0.00000   Min.   :  0.5245
##  1st Qu.:0.00000   1st Qu.:15.00   1st Qu.:0.00000   1st Qu.: 30.7632
##  Median :0.00000   Median :20.00   Median :0.00000   Median : 34.8183
```

```
##  Mean   :0.08557   Mean   :21.25   Mean   :0.03725   Mean   : 34.0007
##  3rd Qu.:0.00000   3rd Qu.:26.00   3rd Qu.:0.00000   3rd Qu.: 37.9499
##  Max.   :1.00000   Max.   :71.00   Max.   :1.00000   Max.   :203.3122
##    M_DEBTINC      FLAG.Reason.DebtCon FLAG.Reason.HomeImp  FLAG.Job.Mgr
##  Min.   :0.0000   Min.   :0.0000      Min.   :0.0000      Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000      1st Qu.:0.0000      1st Qu.:0.0000
##  Median :0.0000   Median :1.0000      Median :0.0000      Median :0.0000
##  Mean   :0.2126   Mean   :0.6591      Mean   :0.2987      Mean   :0.1287
##  3rd Qu.:0.0000   3rd Qu.:1.0000      3rd Qu.:1.0000      3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000      Max.   :1.0000      Max.   :1.0000
##  FLAG.Job.Office  FLAG.Job.Other   FLAG.Job.ProfExe FLAG.Job.Sales
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.00000
##  Mean   :0.1591   Mean   :0.4007   Mean   :0.2141   Mean   :0.01829
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##  FLAG.Job.Self     FLAG.Job.Salary
##  Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000
##  Mean   :0.03238   Mean   :0.3752
##  3rd Qu.:0.00000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.0000
```

```r
write.csv(df, OUTFILE, row.names = FALSE)
```