Credit Card Data Exploration

Attribute Information:

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables: X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit. X2: Gender (1 = male; 2 = female). X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others). X4: Marital status (1 = married; 2 = single; 3 = others). X5: Age (year). X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above. X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005. X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

```
library(tidyverse)
library(MASS)
library(e1071)
library(nnet)
library(class)
library(psych)
library(caret)
library(ROSE)
library(klaR)
library(RColorBrewer)
library(gridExtra)
library(corrplot)
```

Reading the data

```
dd <- read.table(file = "C://Users/cs_mo/Downloads/ISYE7406/ProjectCreditCard/creditcards.csv",
 sep =",",header = TRUE, skip =1)
head(dd)
```

```
##   ID LIMIT_BAL SEX EDUCATION MARRIAGE AGE PAY_0 PAY_2 PAY_3 PAY_4 PAY_5 PAY_6
## 1  1     20000   2         2        1  24     2     2    -1    -1    -2    -2
## 2  2    120000   2         2        2  26    -1     2     0     0     0     2
## 3  3     90000   2         2        2  34     0     0     0     0     0     0
## 4  4     50000   2         2        1  37     0     0     0     0     0     0
## 5  5     50000   1         2        1  57    -1     0    -1     0     0     0
## 6  6     50000   1         1        2  37     0     0     0     0     0     0
##   BILL_AMT1 BILL_AMT2 BILL_AMT3 BILL_AMT4 BILL_AMT5 BILL_AMT6 PAY_AMT1 PAY_AMT2
## 1      3913      3102       689         0         0         0        0      689
## 2      2682      1725      2682      3272      3455      3261        0     1000
## 3     29239     14027     13559     14331     14948     15549     1518     1500
## 4     46990     48233     49291     28314     28959     29547     2000     2019
## 5      8617      5670     35835     20940     19146     19131     2000    36681
## 6     64400     57069     57608     19394     19619     20024     2500     1815
##   PAY_AMT3 PAY_AMT4 PAY_AMT5 PAY_AMT6 default.payment.next.month
## 1        0        0        0        0                          1
## 2     1000     1000        0     2000                          1
## 3     1000     1000     1000     5000                          0
## 4     1200     1100     1069     1000                          0
## 5    10000     9000      689      679                          0
## 6      657     1000     1000      800                          0
```

The data consists of 25 variables and 30,000 records. Removing the ID column.

```
names(dd)[25] <- 'DEFAULT'
data <- dd[,2:25]
dim(data)
```

```
## [1] 30000    24
```

The Marriage and Education has far more category than what was mentioned in the data set explanation.

```
data$MARRIAGE[data$MARRIAGE == "0"] <- "3"
data$EDUCATION[data$EDUCATION== "6"]<-"4"
data$EDUCATION[data$EDUCATION== "5"]<-"4"
data$EDUCATION[data$EDUCATION== "0"]<-"4"
data$DEFAULT[data$DEFAULT=="0"] <- "ND"
data$DEFAULT[data$DEFAULT=="1"] <- "DEF"
```

```
table(data$EDUCATION)
```

```
## 
##     1     2     3     4 
## 10585 14030  4917   468
```

```
table(data$MARRIAGE)
```

```
##
##     1     2     3
## 13659 15964   377
```

Checking the structure of the data.

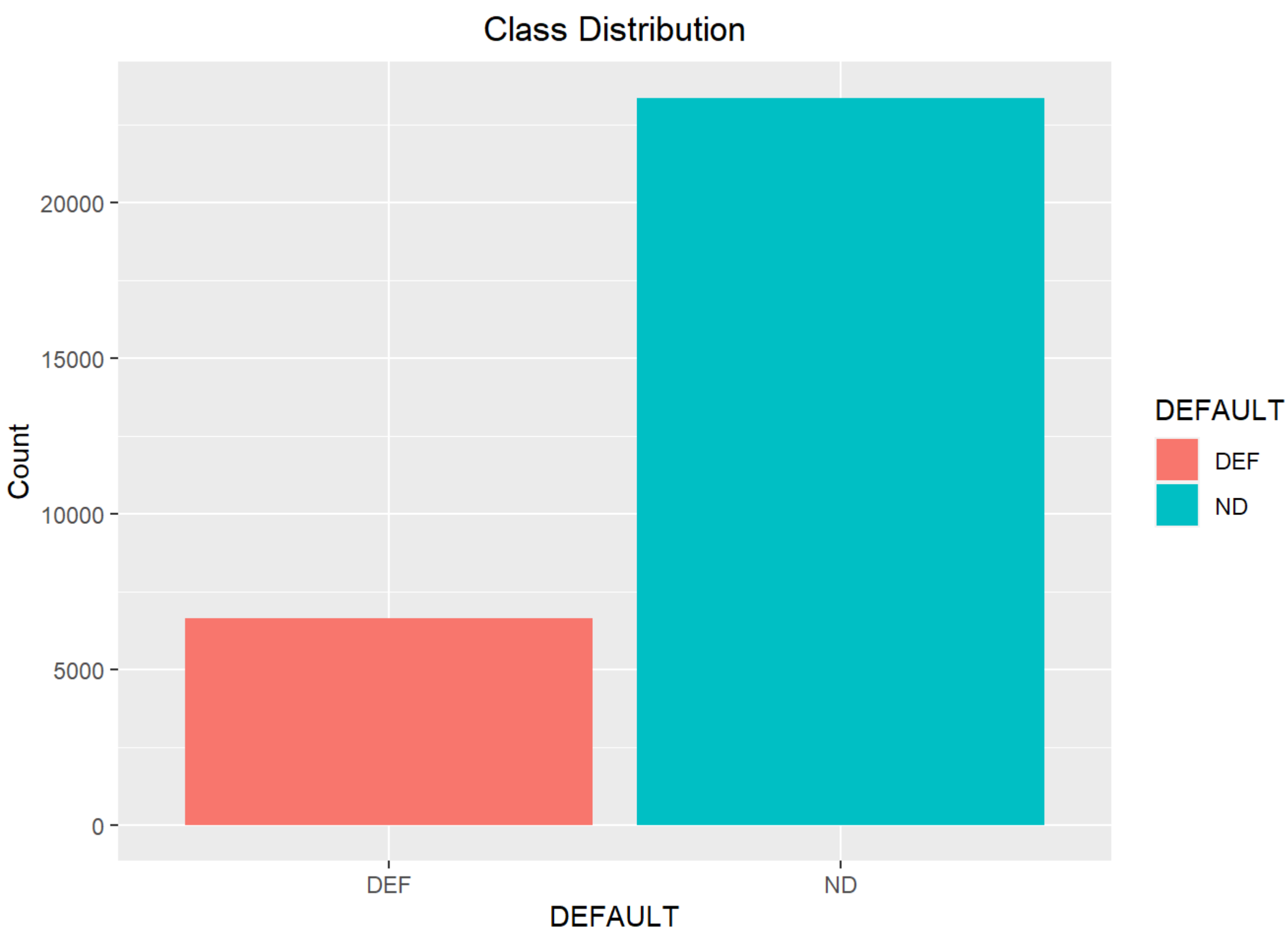
```
str(data)
```

```
## 'data.frame':    30000 obs. of  24 variables:
##  $ LIMIT_BAL: int  20000 120000 90000 50000 50000 50000 500000 100000 140000 20000 ...
##  $ SEX      : int  2 2 2 2 1 1 1 2 2 1 ...
##  $ EDUCATION: chr  "2" "2" "2" "2" ...
##  $ MARRIAGE : chr  "1" "2" "2" "1" ...
##  $ AGE      : int  24 26 34 37 57 37 29 23 28 35 ...
##  $ PAY_0    : int  2 -1 0 0 -1 0 0 0 0 -2 ...
##  $ PAY_2    : int  2 2 0 0 0 0 0 -1 0 -2 ...
##  $ PAY_3    : int  -1 0 0 0 -1 0 0 -1 2 -2 ...
##  $ PAY_4    : int  -1 0 0 0 0 0 0 0 0 -2 ...
##  $ PAY_5    : int  -2 0 0 0 0 0 0 0 0 -1 ...
##  $ PAY_6    : int  -2 2 0 0 0 0 0 -1 0 -1 ...
##  $ BILL_AMT1: int  3913 2682 29239 46990 8617 64400 367965 11876 11285 0 ...
##  $ BILL_AMT2: int  3102 1725 14027 48233 5670 57069 412023 380 14096 0 ...
##  $ BILL_AMT3: int  689 2682 13559 49291 35835 57608 445007 601 12108 0 ...
##  $ BILL_AMT4: int  0 3272 14331 28314 20940 19394 542653 221 12211 0 ...
##  $ BILL_AMT5: int  0 3455 14948 28959 19146 19619 483003 -159 11793 13007 ...
##  $ BILL_AMT6: int  0 3261 15549 29547 19131 20024 473944 567 3719 13912 ...
##  $ PAY_AMT1 : int  0 0 1518 2000 2000 2500 55000 380 3329 0 ...
##  $ PAY_AMT2 : int  689 1000 1500 2019 36681 1815 40000 601 0 0 ...
##  $ PAY_AMT3 : int  0 1000 1000 1200 10000 657 38000 0 432 0 ...
##  $ PAY_AMT4 : int  0 1000 1000 1100 9000 1000 20239 581 1000 13007 ...
##  $ PAY_AMT5 : int  0 0 1000 1069 689 1000 13750 1687 1000 1122 ...
##  $ PAY_AMT6 : int  0 2000 5000 1000 679 800 13770 1542 1000 0 ...
##  $ DEFAULT  : chr  "DEF" "DEF" "ND" "ND" ...
```

```
married <- c('1' = "Married",
             "2"= "Single",
             "3" = "others")
gender <- c("1"= "Male",
            "2"= "Female")
edu <- c("1"= "Graduate",
         "2" = "University",
         "3" = "High School",
         "4" = "Others")
```

The data has 23364 non-defaulters and 6636 defaulters. This is imbalanced data. However, in real life this very common with the credit card data as most of the customers do not default on their payments. We can try oversampling or under sampling to balance the data.

```
data %>%
  ggplot(aes(DEFAULT, fill = DEFAULT))+
  geom_bar()+
  labs(title = "Class Distribution", x= "DEFAULT",y= "Count")+
  theme(plot.title = element_text(hjust = 0.5) )
```

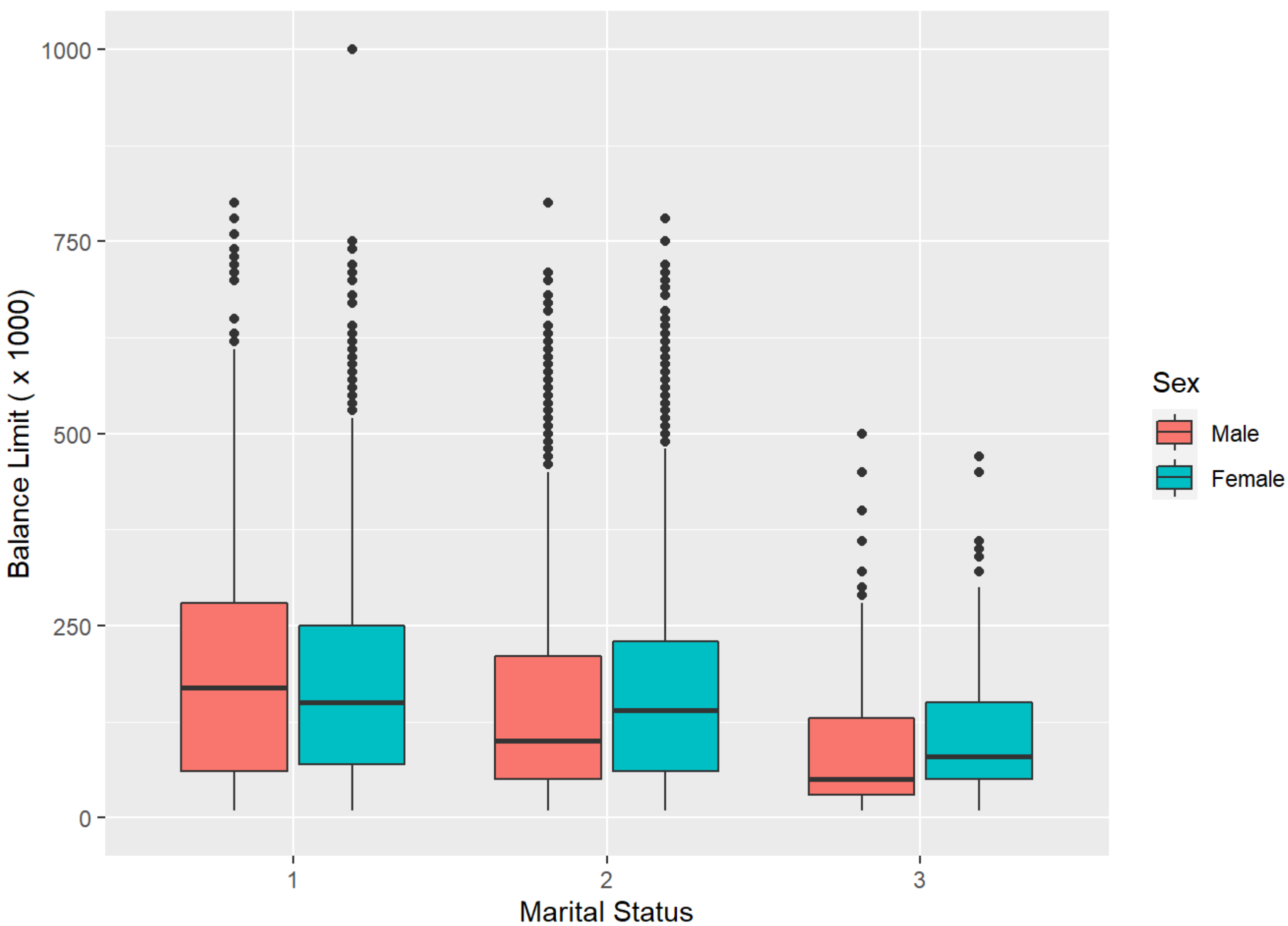



```
table(data$DEFAULT)
```

```
##
##   DEF    ND
##  6636 23364
```

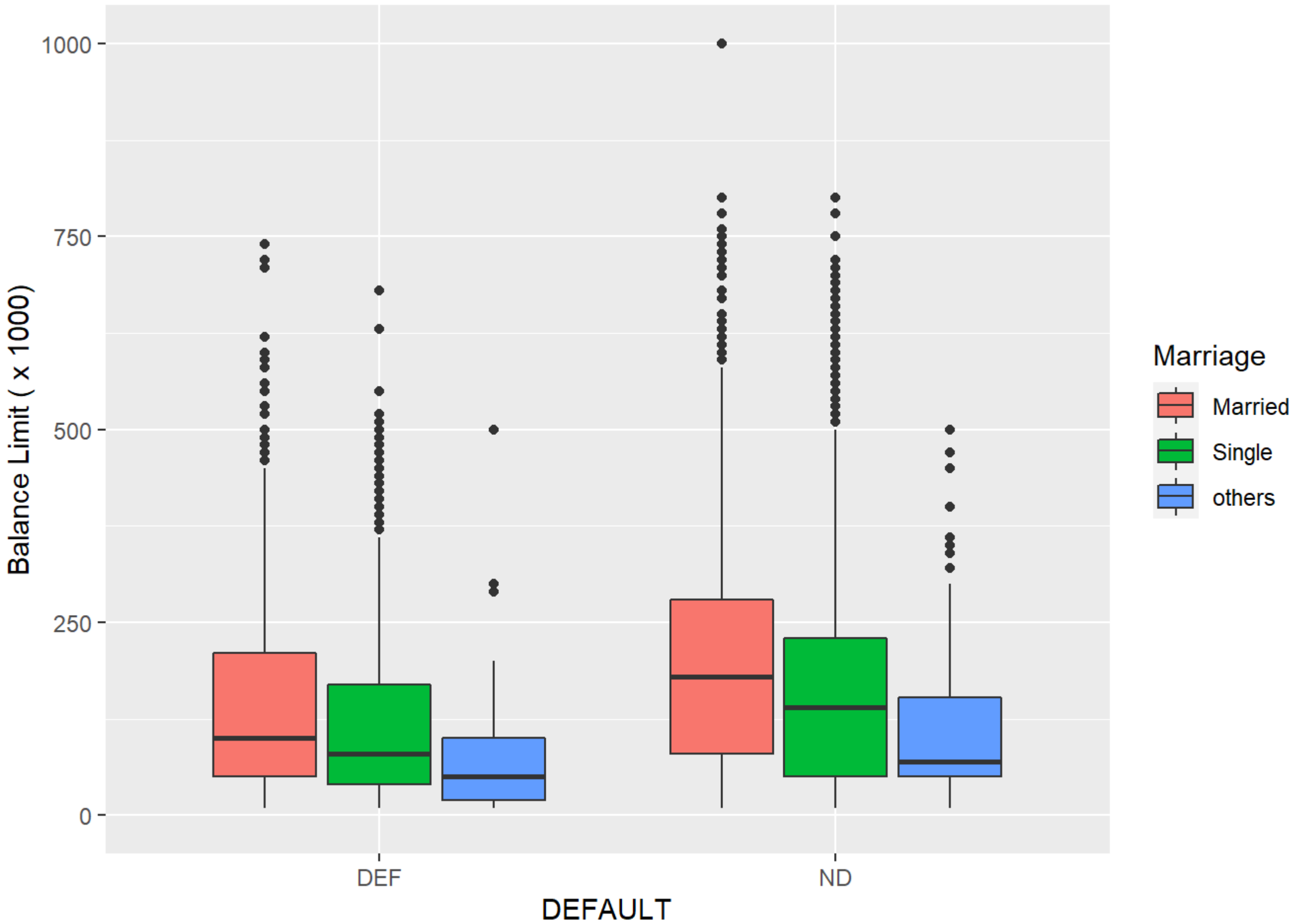
Below comparison indicates that married men and women have higher balance limit than single men and women. Also married men have little higher balance limit than married women.

```
bp<-ggplot(data, aes(factor(MARRIAGE), (LIMIT_BAL/1000), fill=as.factor(SEX))) +
  geom_boxplot() +
  xlab("Marital Status") +
  ylab("Balance Limit ( x 1000)") +
  coord_cartesian(ylim = c(0,1000))+
  scale_fill_discrete(name = "Sex", labels = gender)
bp
```



Below plot indicates that married, single or others defaulters have lower balance limit compare to non- defaulters.

```
bp0<-ggplot(data, aes(factor(DEFAULT), (LIMIT_BAL/1000), fill=as.factor(MARRIAGE))) +
  geom_boxplot() +
  xlab("DEFAULT") +
  ylab("Balance Limit ( x 1000)") +
  coord_cartesian(ylim = c(0,1000))+
  scale_fill_discrete(name = "Marriage", labels = married)
bp0
```

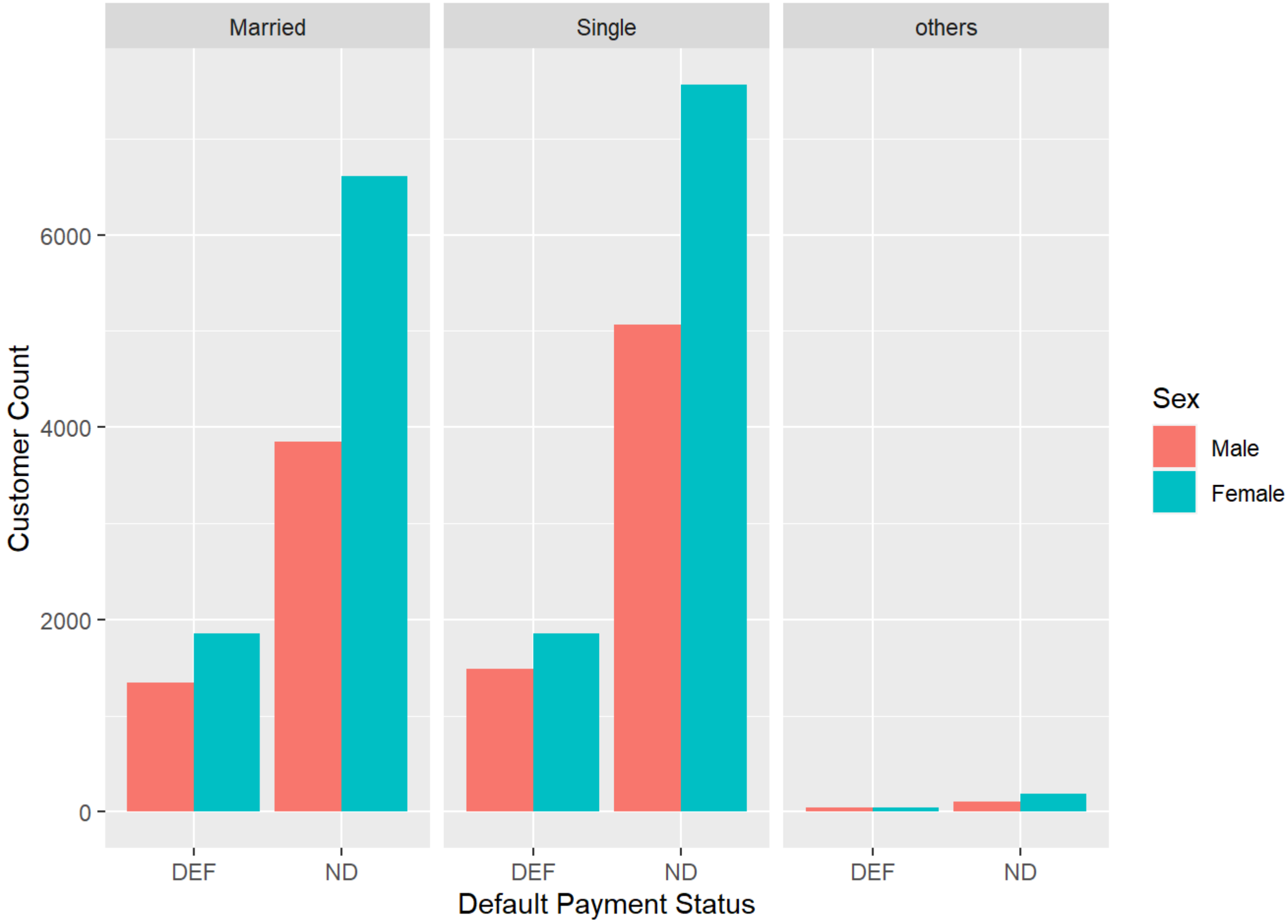


###

Below plot indicates that married or single females default on their payments more than married or single male. Even though male and female both have same limit balance.

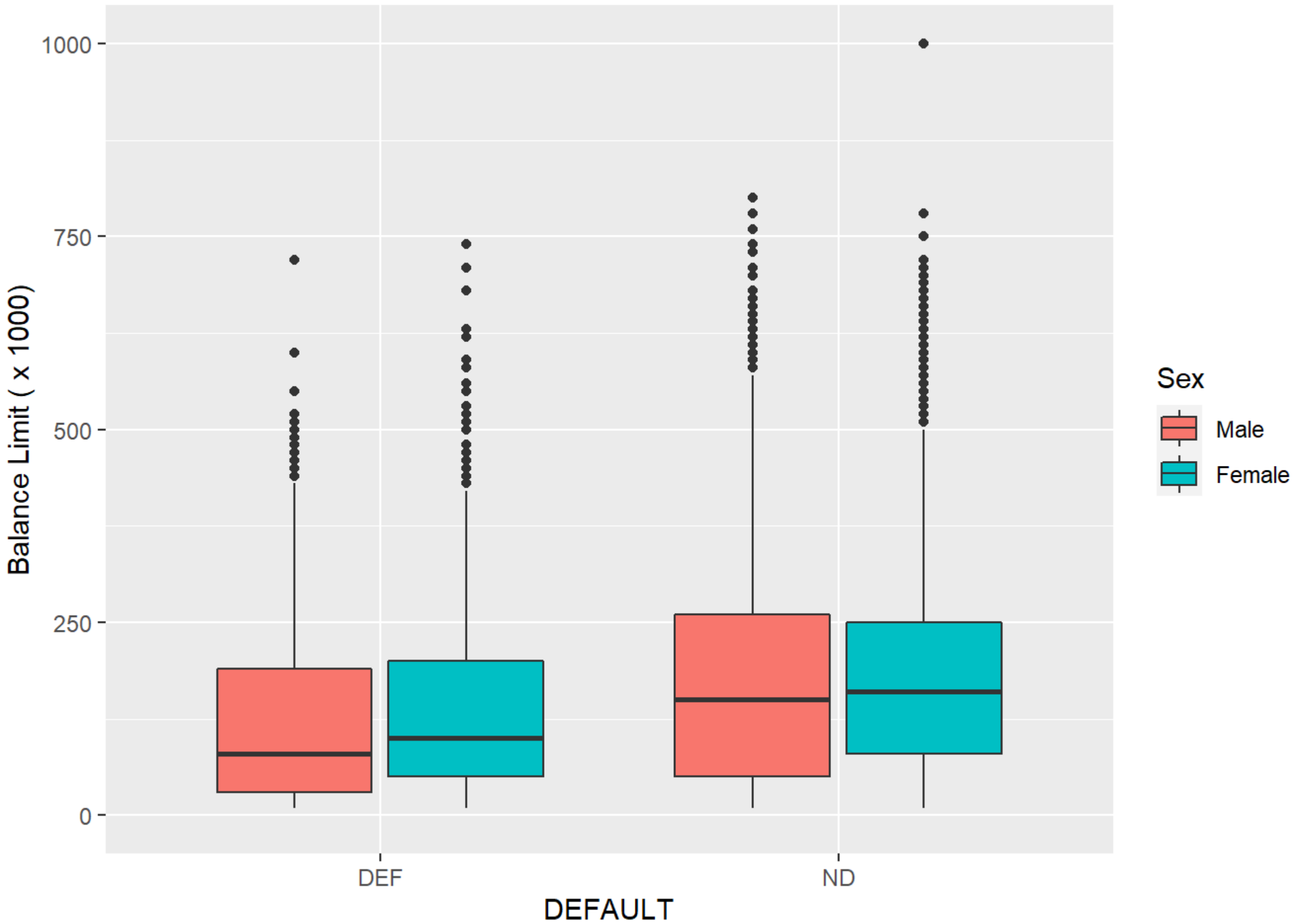
```
grm <- ggplot(data, aes(x=as.factor(DEFAULT), fill = as.factor(SEX), )) +
  geom_histogram(stat="count",position = "dodge") +
  xlab("Default Payment Status") + ylab("Customer Count") +
  facet_wrap(~MARRIAGE, labeller = as_labeller(married))+

  scale_fill_discrete(name = "Sex", labels = gender)
grm
```



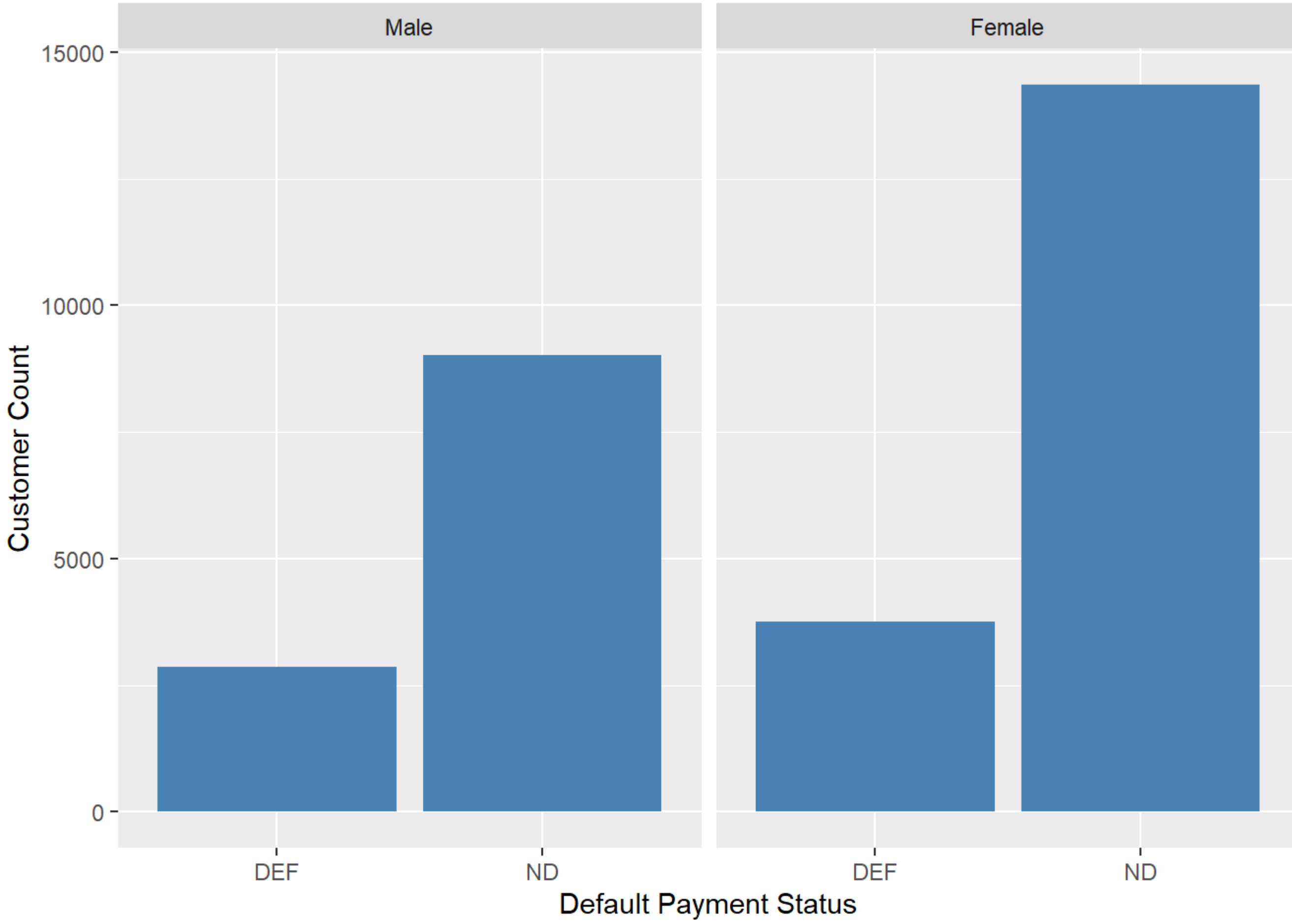
###Non-defaulter men and women both have higher balance limit than defaulter men and women. Which make sense as the higher limit balance mostly given based on the customer's credit and history of payment.

```
bp00<-ggplot(data, aes(factor(DEFAULT), (LIMIT_BAL/1000), fill=as.factor(SEX))) +
  geom_boxplot() +
  xlab("DEFAULT") +
  ylab("Balance Limit ( x 1000)") +
  coord_cartesian(ylim = c(0,1000))+
  scale_fill_discrete(name = "Sex", labels = gender)
bp00
```



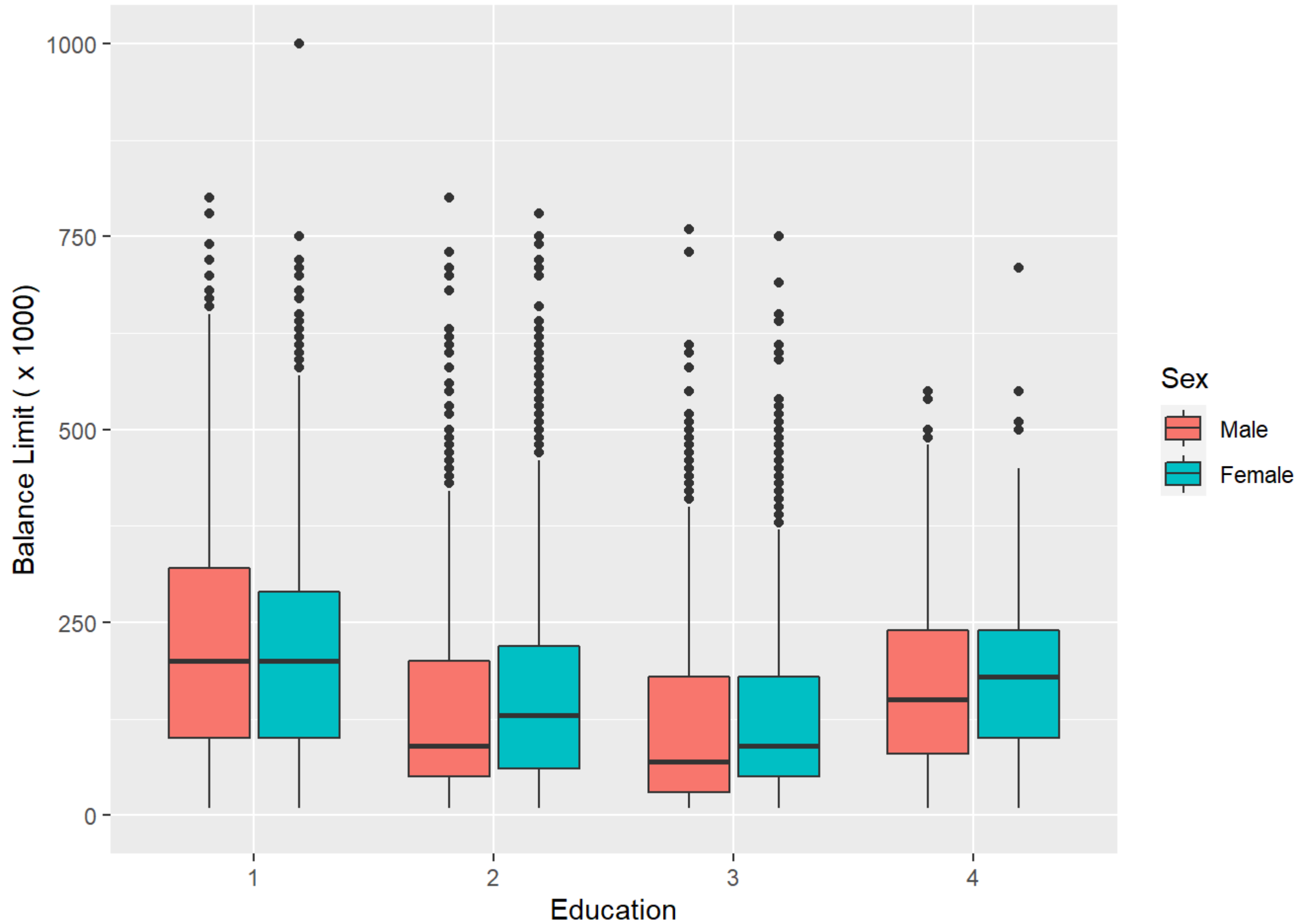
Below plot indicates that female are likely to default on their payment compare to male.

```
grs <- ggplot(data, aes(x=as.factor(DEFAULT))) +
  geom_histogram(stat="count",fill='steelblue') +
  xlab("Default Payment Status") + ylab("Customer Count") +
  facet_wrap(~SEX, labeller = as_labeller(gender))
grs
```



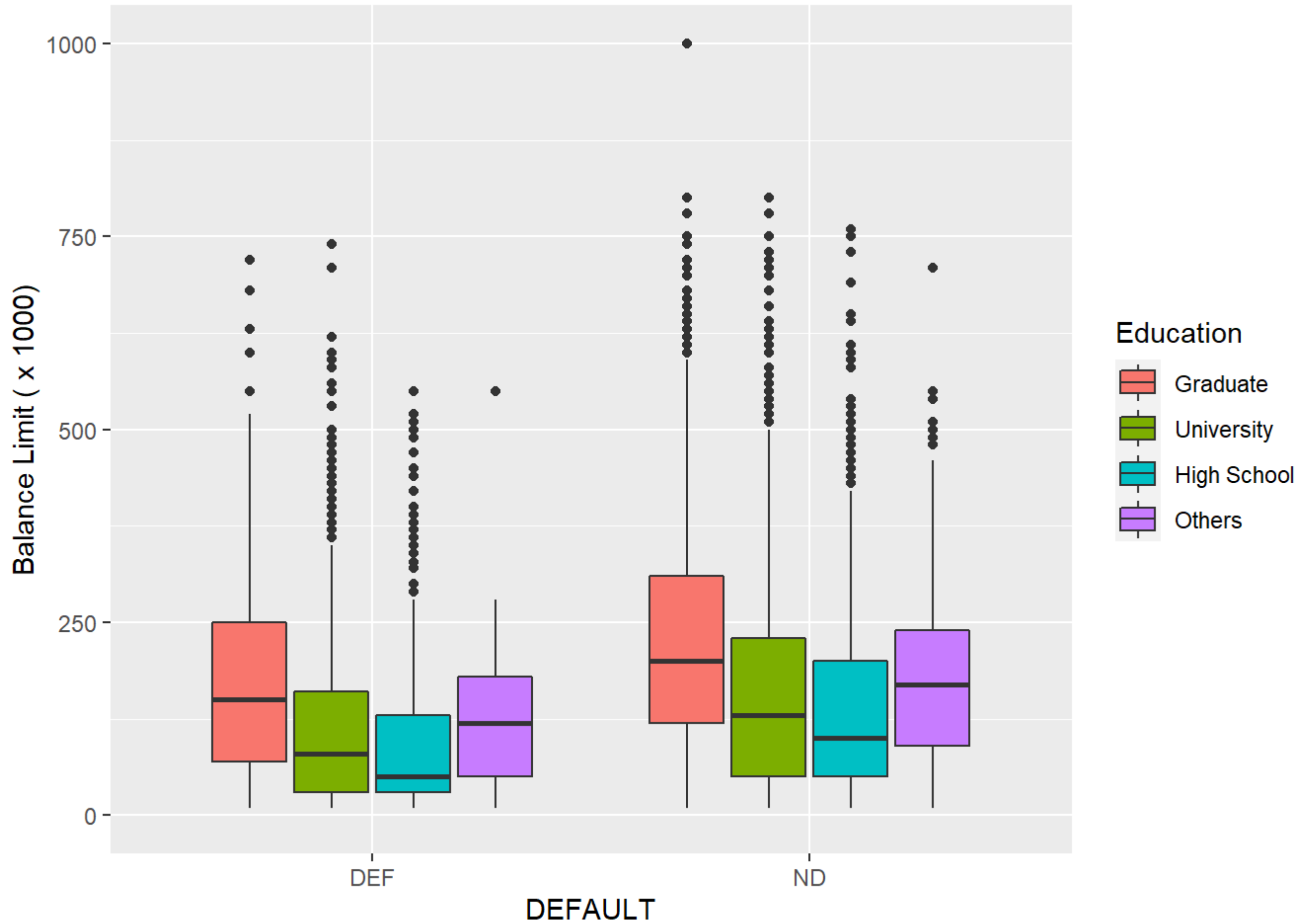
This plot indicates that graduate male and female have the highest limit balance than other groups.

```
bp2<-ggplot(data, aes(factor(EDUCATION), (LIMIT_BAL/1000), fill=as.factor(SEX))) +
  geom_boxplot() +
  xlab("Education") +
  ylab("Balance Limit ( x 1000)") +
  coord_cartesian(ylim = c(0,1000))+
  scale_fill_discrete(name = "Sex", labels = gender)
bp2
```



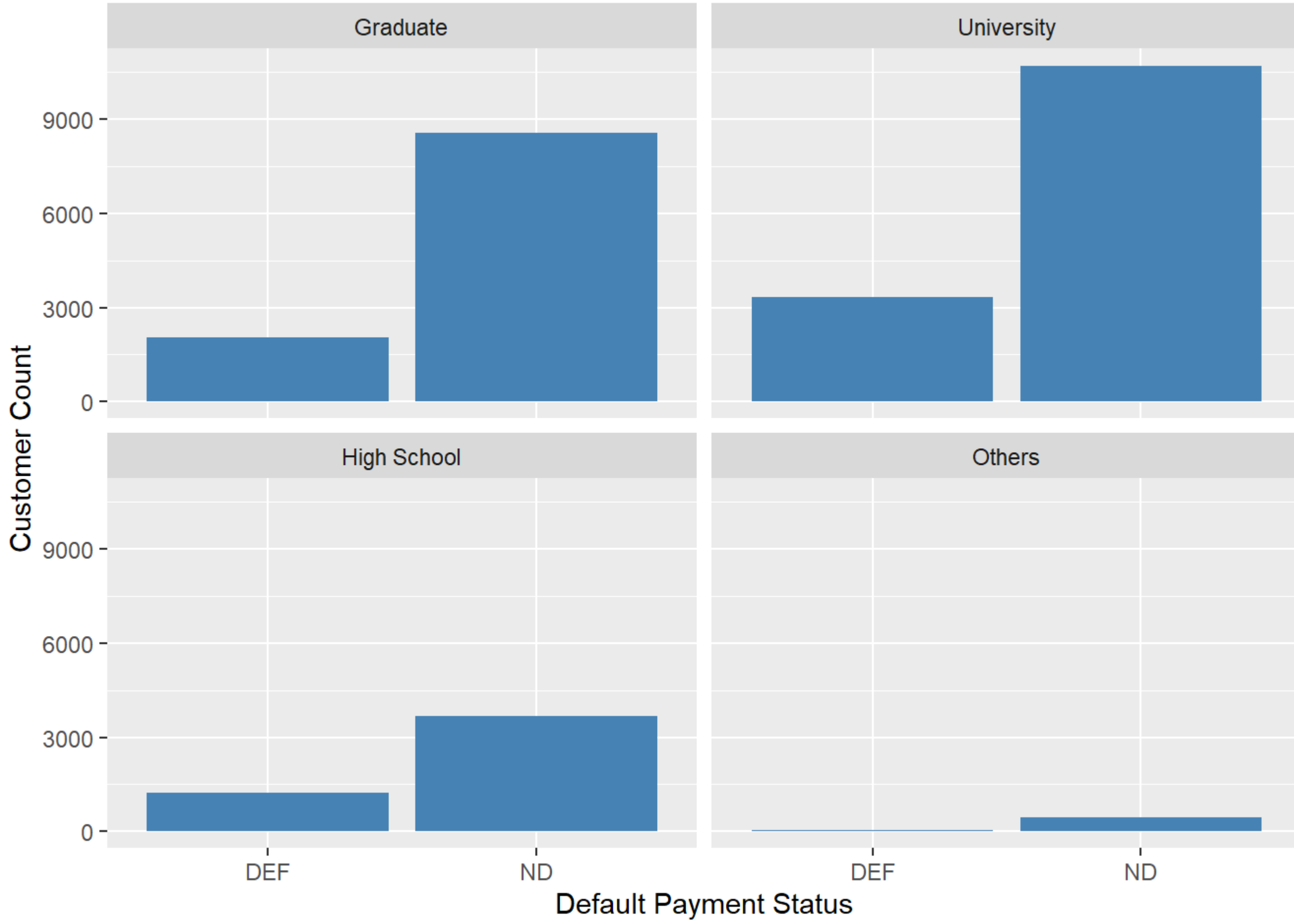
Below plot indicates that graduate persons are more likely to default on their payment, followed by high schoolers.

```
bp000<-ggplot(data, aes(factor(DEFAULT), (LIMIT_BAL/1000), fill=as.factor(EDUCATION))) +
  geom_boxplot() +
  xlab("DEFAULT") +
  ylab("Balance Limit ( x 1000)") +
  coord_cartesian(ylim = c(0,1000))+
  scale_fill_discrete(name = "Education", labels = edu)
bp000
```



This indicates that University graduates are the highest defaulters.

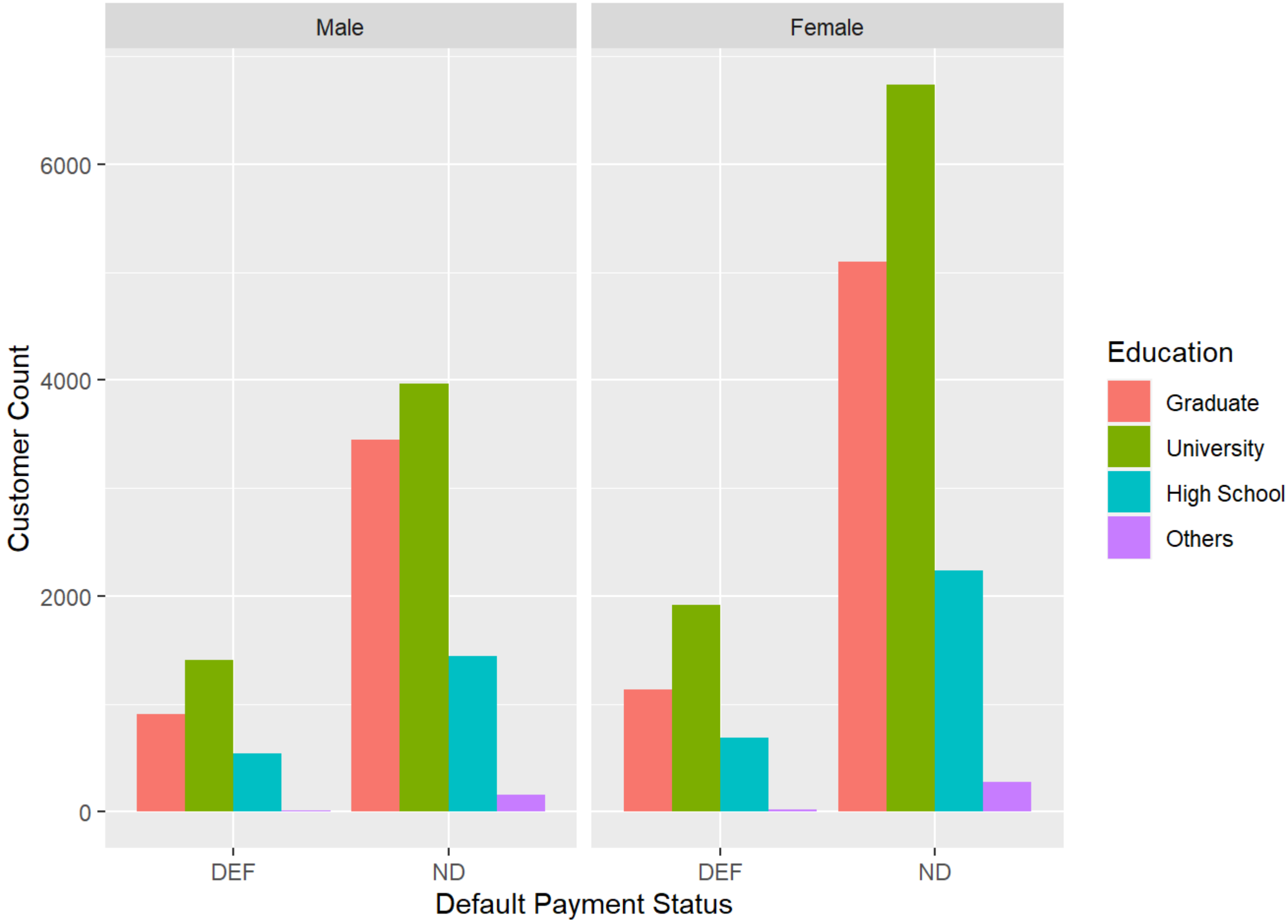
```
gr1 <- ggplot(data, aes(x=as.factor(DEFAULT))) +
  geom_histogram(stat="count",fill='steelblue') +
  xlab("Default Payment Status") + ylab("Customer Count") +
  facet_wrap(~EDUCATION, labeller = as_labeller(edu))
gr1
```



###

This plot indicates that the University graduate females are the highest defaulter among all categories, closely followed by university graduate males. The high school graduates are the least defaulter.

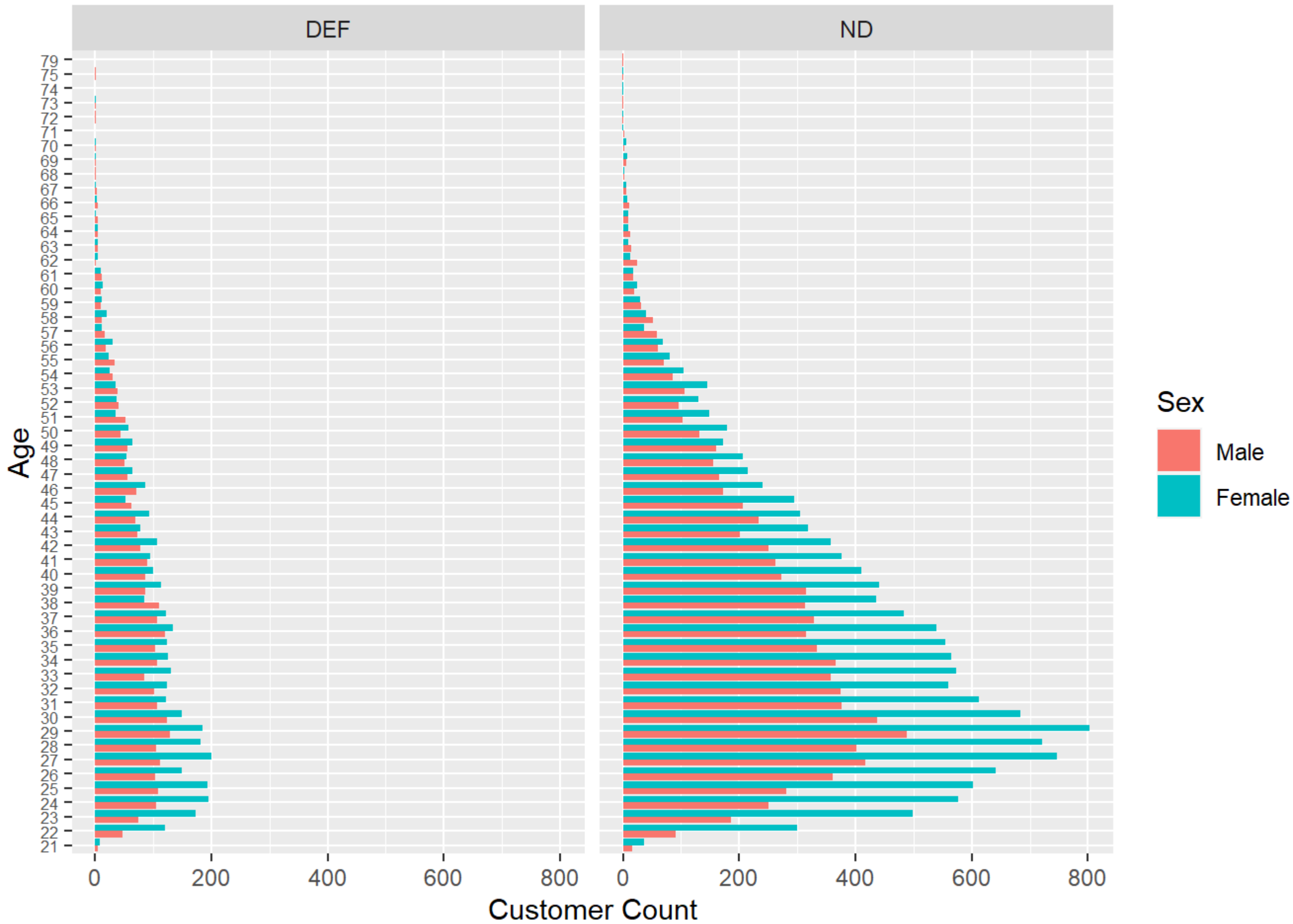
```
gr2 <- ggplot(data, aes(x=as.factor(DEFAULT)),aes(y=stat_count(SEX))) +
  geom_bar(aes(fill=factor(EDUCATION)), position = "dodge") +
  xlab("Default Payment Status")+ylab("Customer Count") +
  facet_wrap(~SEX, labeller = as_labeller(gender))+
  scale_fill_discrete(name="Education", labels = edu)
gr2
```



###

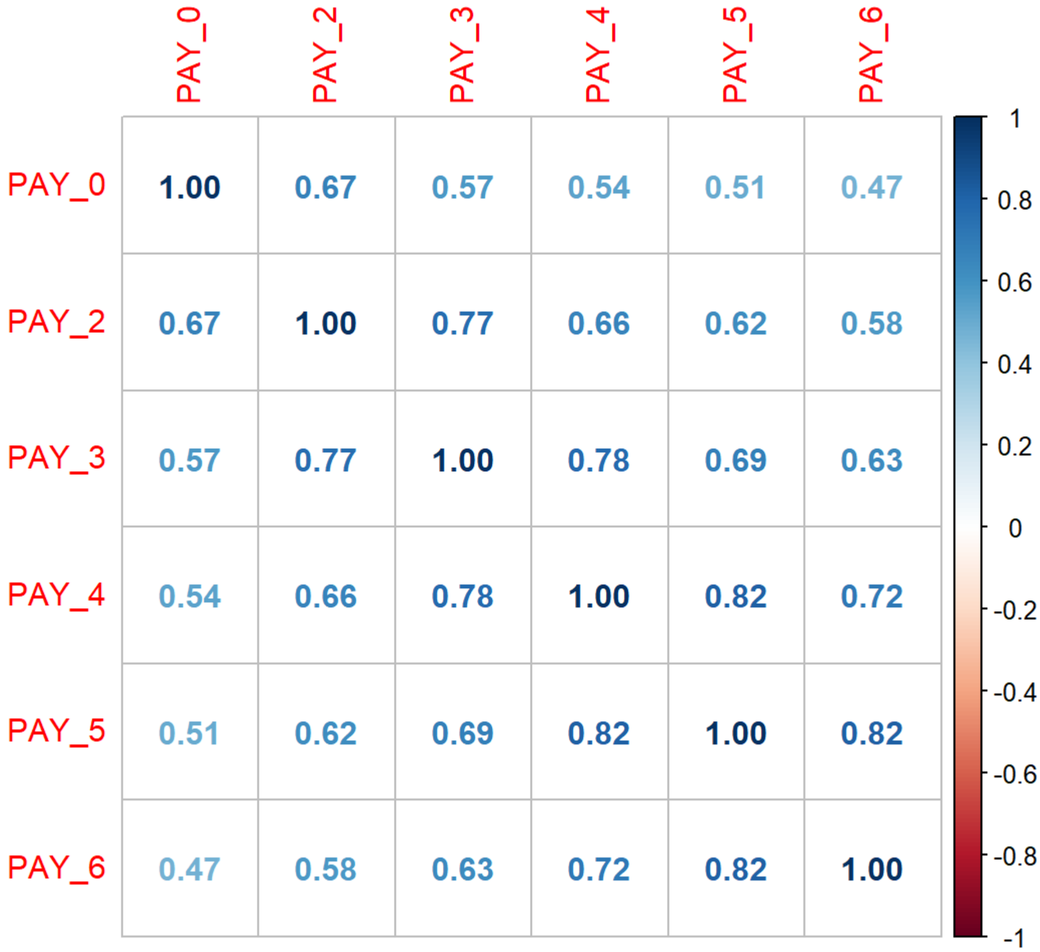
As we can see from blow plot females between age of 22 and 32 have tendency to default on their payment compared to males. As the customers age grow, they are less likely to default on their payment.

```
bp11<-ggplot(data, aes(x=as.factor(AGE))) +
  geom_bar(aes(fill=factor(SEX)), position = "dodge") +
  xlab("Age")+ylab("Customer Count") +
  facet_wrap(~DEFAULT)+
  scale_fill_discrete(name="Sex", labels = gender)+
  theme(axis.text.y = element_text(
                         size=6))+
  coord_flip()
bp11
```



Below plot indicates that as the month increases from April to September for the repayment status, they show medium to strong correlation.

```
df <- cor(data[,6:11])
corrplot(df, method = "number")
```



Bill Statement.

This plot shows that variables are highly correlated, and they might affect our results. Further exploration needs to be.

```
df <- cor(data[,12:17])
corrplot(df, method = "number")
```

###

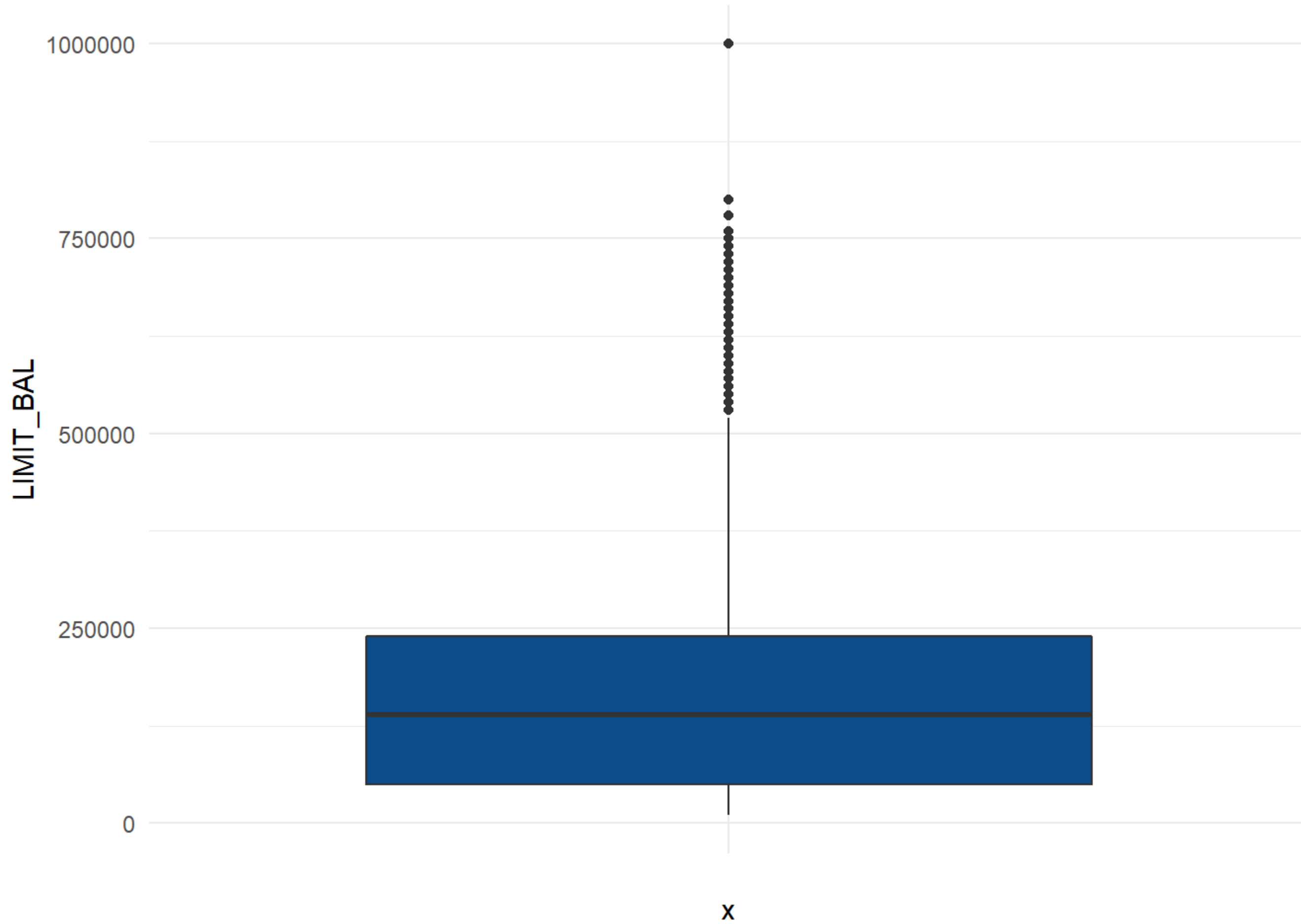
Below “Amount of Previous Payments” shows no correlation at all.

```
df <- cor(data[,18:23])
corrplot(df, method = "number")
```

###

As we saw in the above plots, the limit balance has lot of outliers. Those observations needs to be analyzed further. However, for this assignment I am going to remove them and see if that improves our results.

```
ggplot(data) +
  aes(x = "", y = LIMIT_BAL) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



removing outliers from the data……

```
out <- boxplot.stats(data$LIMIT_BAL)$out
out_ind <- which(data$LIMIT_BAL %in% c(out))
new_data <- data[-out_ind,]
dim(new_data)
```

```
## [1] 29833    24
```