Data Wrangling Assignment(Group 6)

Akshara Joshi, Ashish Gupta, Arjun J Bhatnagar,Naman Dhameja, Palak Bansal, Sudhanshu Ranjan

20/10/2021

Importing Data and removing the first Row

```
library(readxl)
data <- read_excel("G:/GithubBITS/Assignment3_DataWrangling/India Credit Data.xlsx", skip = 1)
View(data)
# Remove first row
data<-data[-1,]
```

Removing "End of.." Using length logic 1st Method

```
# for(i in 1:range(nrow(data))){
#   if(nchar(data$`BORROWER ID`[i])>=18){
#     print(data$`BORROWER ID`[i])
#
#     data<-data[-c(i),]
#   }
# }
```

Removing "End of.." Using Space logic 2nd method

```
# for(i in 1:range(nrow(data))){
#   text = data$`BORROWER ID`[i]
#    if(str_count(text, ' ')>0){
#      data<-data[-c(i),]
#    }
# }
```

Remove columns with "End of..." using row numbers 3rd method

```
# 3rd method
data<-data[-c(98,197,298,398,498,1006),]
View(data)
```

Converting different formats of string to one.

```
unique(data$`Bank Account`) # Finding the unique Bank Account entries detected by the computer
```

```
##  [1] "Inactive"        "No Bank Account" "In-active"       "Do not Know"
##  [5] "in-active"       "DNK"             "Active"          "active"
##  [9] NA                "NBA"             "inactive"
```

```
tolower(unique( data$`Bank Account`))# Converting them to lowercase
```

```
##  [1] "inactive"        "no bank account" "in-active"       "do not know"
##  [5] "in-active"       "dnk"             "active"          "active"
##  [9] NA                "nba"             "inactive"
```

```
trial<-data$`Bank Account` # Trial list for experimenting the operations on the column
trial<-str_replace_all(tolower(data$`Bank Account`),"-","")
unique(trial)
```

```
## [1] "inactive"        "no bank account" "do not know"     "dnk"
## [5] "active"          NA                "nba"
```

```
# Replacing similar meaning columns entered differently
data$`Bank Account`<-str_replace_all(tolower(data$`Bank Account`),"-","")
data$`Bank Account`<-replace(data$`Bank Account`,data$`Bank Account`=="dnk","do not know")

data$`Bank Account`<-replace(data$`Bank Account`,data$`Bank Account`=="nba","no bank account")
unique(data$`Bank Account`) # Again checking if now the data is cleaned or not
```

```
## [1] "inactive"        "no bank account" "do not know"     "active"
## [5] NA
```

```
View(data)
```

Gender column to M,F,O only

```
# Categorizing Gender Data into "Male(M)", "Female(F)", "Other(O)"
data$Gender<-str_replace_all(tolower(data$Gender),"-","")

unique(data$Gender)
```

```
## [1] "male"   "female" "other"  "f"      NA       "o"      "m"
```

```
data$Gender<-replace(data$Gender,data$Gender=="male","M")
data$Gender<-replace(data$Gender,data$Gender=="m","M")

data$Gender<-replace(data$Gender,data$Gender=="female","F")
data$Gender<-replace(data$Gender,data$Gender=="f","F")

data$Gender<-replace(data$Gender,data$Gender=="other","O")
data$Gender<-replace(data$Gender,data$Gender=="o","O")

unique(data$Gender)
```

```
## [1] "M" "F" "O" NA
```

```
View(data)
```

Converting multiple NAs to “Missing”

```
data$`Bank Account`[which(is.na(data$`Bank Account`))]="Missing"
data$Purpose[which(is.na(data$Purpose))]="Missing"
data$`Marital status`[which(is.na(data$`Marital status`))]="Missing"
data$SECURITY[which(is.na(data$SECURITY))]="Missing"
data$Occupation[which(is.na(data$Occupation))]="Missing"
```

Replace ‘-’ characters in Months column

```
data$Months <- str_replace_all(as.character(data$Months),"-","")
data$Age<-as.numeric(data$Age)
data$Months<-as.numeric(data$Months)

View(data)
```

Loan Closed column into Y,N only

```
unique(data$`Loan Closed`)
```

```
## [1] "No"  "y"   "YES" "Y"   "NO"  "N"   "Yes" "n"   NA
```

```
data$`Loan Closed`<-replace(data$`Loan Closed`,data$`Loan Closed`=="No","N")
data$`Loan Closed`<-replace(data$`Loan Closed`,data$`Loan Closed`=="NO","N")
data$`Loan Closed`<-replace(data$`Loan Closed`,data$`Loan Closed`=="n","N")
data$`Loan Closed`<-replace(data$`Loan Closed`,data$`Loan Closed`=="N","N")

data$`Loan Closed`<-replace(data$`Loan Closed`,data$`Loan Closed`=="y","Y")
data$`Loan Closed`<-replace(data$`Loan Closed`,data$`Loan Closed`=="YES","Y")

data$`Loan Closed`<-replace(data$`Loan Closed`,data$`Loan Closed`=="Y","Yes")
data$`Loan Closed`<-replace(data$`Loan Closed`,data$`Loan Closed`=="N","No")

unique(data$`Loan Closed`)
```

```
## [1] "No"  "Yes" NA
```

Going through the data for any discrepancies

```
unique(data$Purpose)
```

```
##  [1] "household furniture/equipment" "vehicle purchase  (used)"
##  [3] "others"                        "vacation"
##  [5] "education"                     "domestic appliances"
##  [7] "vehicle purchase (new)"        "Missing"
##  [9] "business"                      "farming"
## [11] "work furniture/equipment"      "television"
## [13] "repairs"
```

```r
head(data$`Loan Amount`,10)
```

```
##  [1] 85000 41000 62000 28000 36000 25000 86000 41000 86000 34000
```

```r
head(data$`Annual Income`,10)
```

```
##  [1] 7e+05 3e+05 4e+05 2e+05 3e+05 6e+05 7e+05 4e+05 6e+05 1e+05
```

```r
data$`Annual Income`<-as.integer(data$`Annual Income`)
```

```r
unique(data$`Marital status`)
```

```
##  [1] "Widowed"   "Missing"   "Divorced"  "SEPARATED" "MARRIED"   "Married"
##  [7] "WIDOWED"   "DIVORCED"  "SINGLE"    "Single"    "Separated"
```

```r
data$`Marital status`<-tolower(data$`Marital status`)
unique(data$`Marital status`)
```

```
## [1] "widowed"   "missing"   "divorced"  "separated" "married"   "single"
```

```r
unique(data$SECURITY)
```

```
##  [1] "Farm-land"             "RESIDENTIAL LAND"      "Stocks"
##  [4] "COMMERCIAL Property"   "House"                 "Financial Instruments"
##  [7] "Missing"               "Third Person Gurantee" "Employer Gurantee"
## [10] "COMMERCIAL LAND"       "Gold"
```

Creating a new column of good and bad loans, according to the purpose. The depreciating assets are termed to be bad loans like: vehicles and repairs. The assets like education are considered to be good loans. According to this, we created 3 lists.

```r
trial<-c()# creating temporary empty list
unique(data$Purpose)
```

```
##  [1] "household furniture/equipment" "vehicle purchase  (used)"
##  [3] "others"                        "vacation"
##  [5] "education"                     "domestic appliances"
##  [7] "vehicle purchase (new)"        "Missing"
##  [9] "business"                      "farming"
## [11] "work furniture/equipment"      "television"
## [13] "repairs"
```

```r
# Categorizing the Purpose into Good, Bad and Others
other<-list("missing","others")
good<-list("education","business","farming")
bad<-list("household furniture/equipment","vehicle purchase  (used)","vacation" ,"domestic appliances",

# Making a column according to the critera mentioned above
```

```
for( i in 1:range(nrow(data))){
  if(data$Purpose[i] %in% bad ){
    trial<-c(trial,"Bad Loan")

  }else if(data$Purpose[i] %in% good){
    trial<-c(trial,"Good Loan")
  }else{
    trial<-c(trial,"Missing/Other")
  }
}
```

```
## Warning in 1:range(nrow(data)): numerical expression has 2 elements: only the
## first used
```

```
data$LoanType<-trial
View(data)
summary(as.factor(data$LoanType))
```

```
##      Bad Loan     Good Loan Missing/Other
##           596           288           116
```

There are few instances where all installments are paid but still loan isn't mentioned close, so for that we have changed certain entries where number of paid installments are equal to 48.

```
for( i in 1:range(nrow(data))){
  if(data$`# paid`[i]==48){
    data$`Loan Closed`[i]<-"Yes"
  }
  }
```

```
## Warning in 1:range(nrow(data)): numerical expression has 2 elements: only the
## first used
```

Calculated % paid amount and updated the %paid column in dataset.

```
View(data)
for( i in 1:range(nrow(data))){
  data$`% Paid`[i]<-round((data$`# paid`[i]/48)*100,2)
  }
```

```
## Warning in 1:range(nrow(data)): numerical expression has 2 elements: only the
## first used
```

##Potential Defaulter Criteria decided for potential defaulter is: 1. %Paid amount is less than 40. 2. The skipped installments are atleast 30% of the number of installments paid.

```
trial<-c()

for( i in 1:range(nrow(data))){
  if(data$`% Paid`[i]<=40 && data$`# skipped`[i]>=round(data$`# paid`[i]*0.3,0) ){
```

```
    trial<-c(trial,"Yes")

  }else{
    trial<-c(trial,"No")
  }
}
```

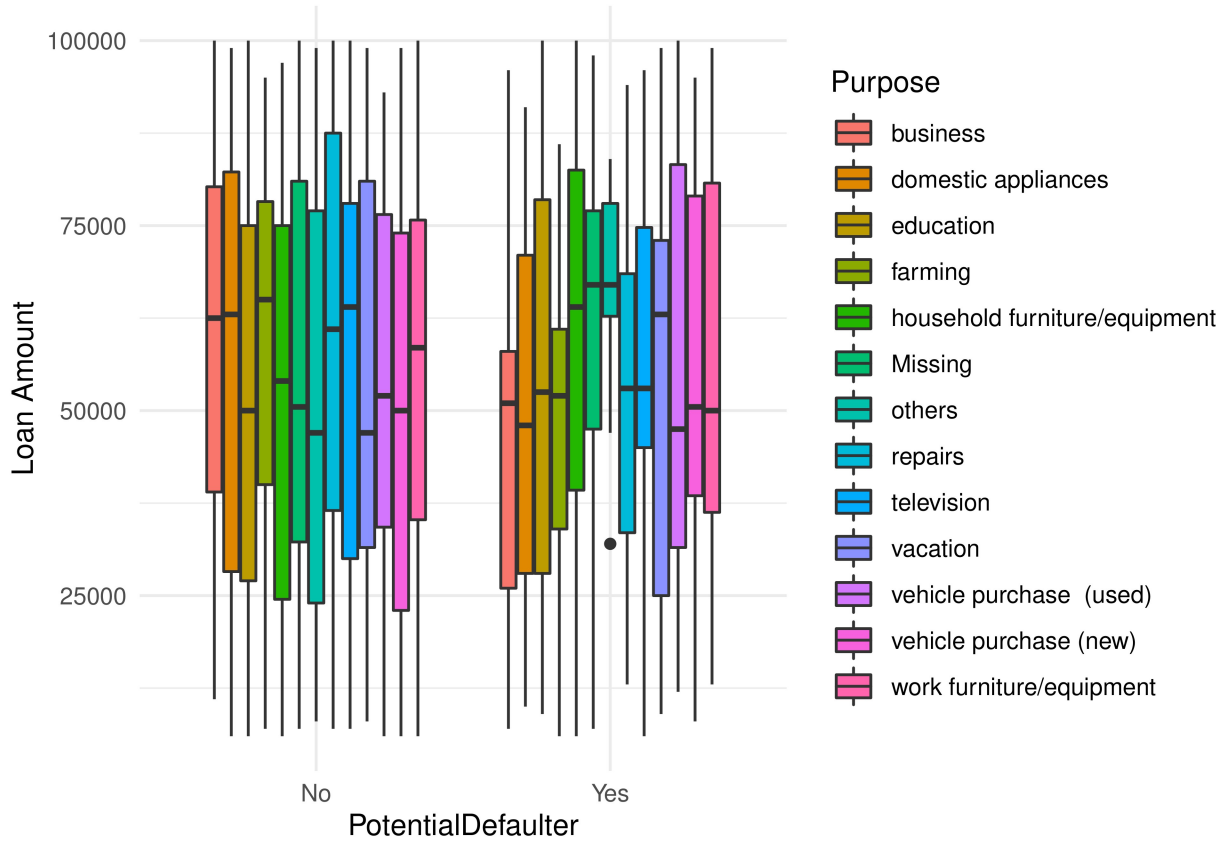
```
## Warning in 1:range(nrow(data)): numerical expression has 2 elements: only the
## first used
```

```
data$PotentialDefaulter<-trial
View(data)
summary(as.factor(data$PotentialDefaulter))
```

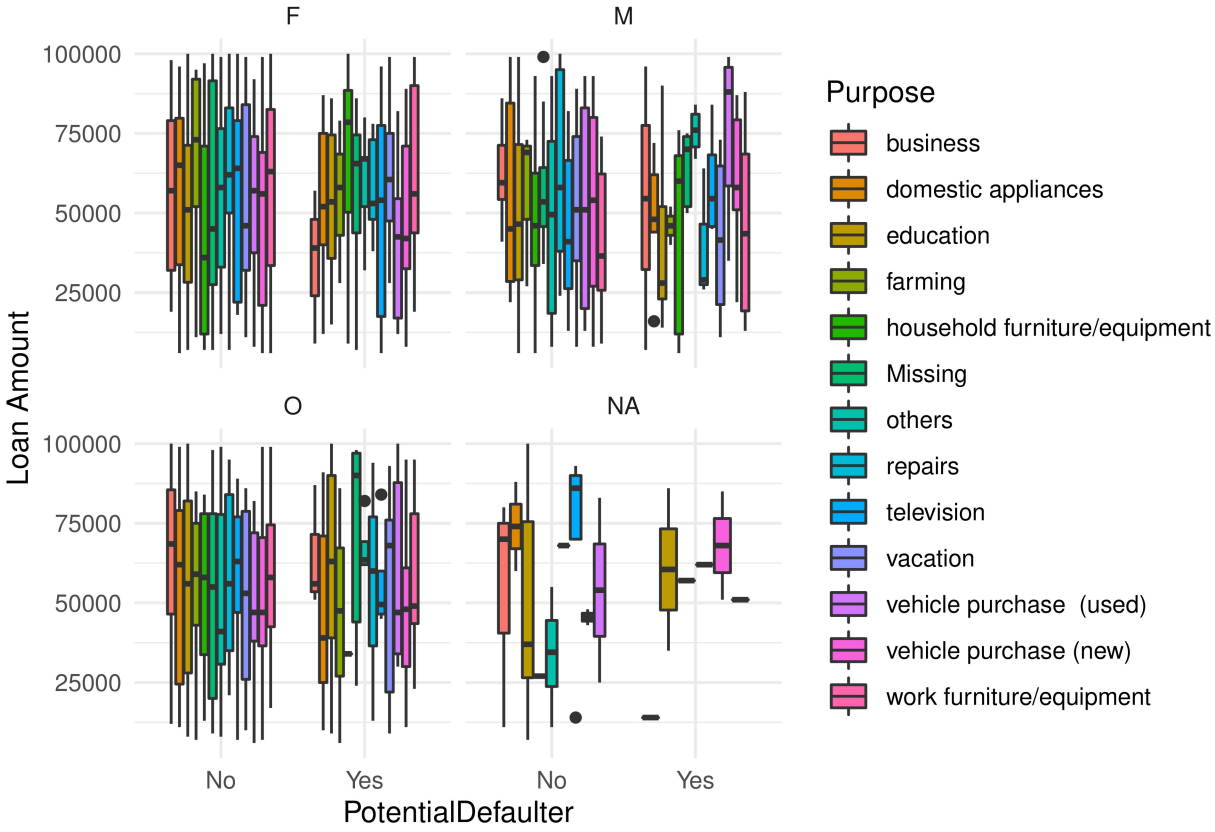
```
##  No Yes
## 718 282
```

##Data Visualization

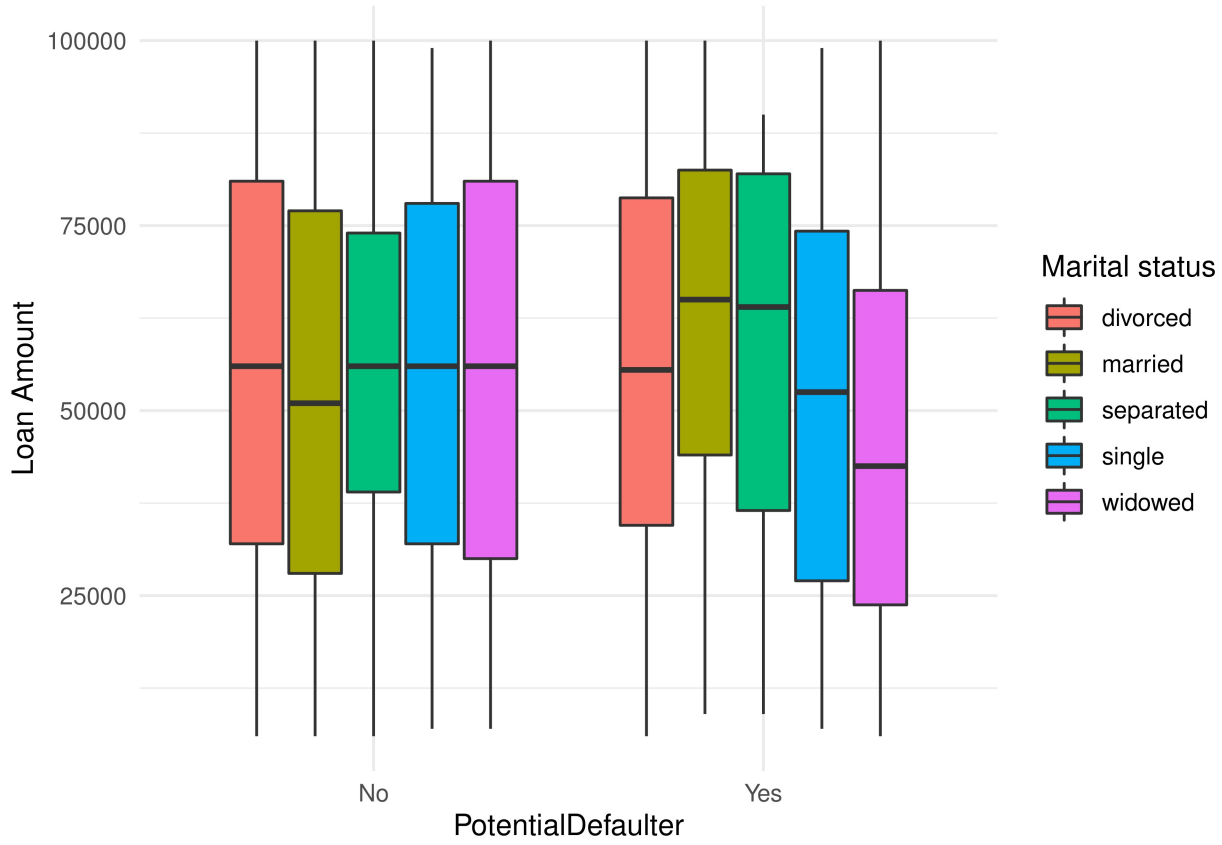
```
library(ggplot2)

#potential defaulters and loan amount and purpose
ggplot(data) +
 aes(x = PotentialDefaulter, y = `Loan Amount`, fill = Purpose) +
 geom_boxplot(shape = "circle") +
 scale_fill_hue(direction = 1) +
 theme_minimal()
```



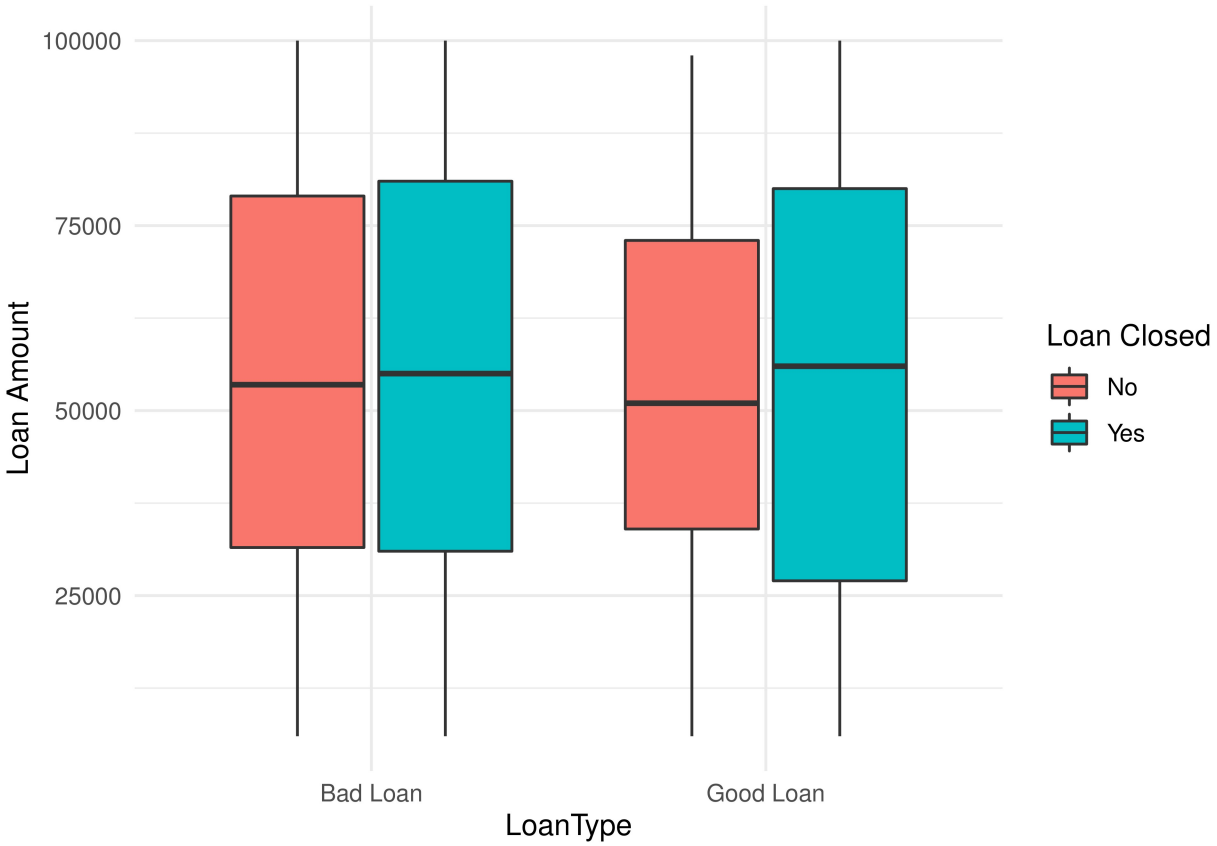

```
#potential defaulters and loan amount and purpose grouped by gender
ggplot(data) +
 aes(x = PotentialDefaulter, y = `Loan Amount`, fill = Purpose) +
 geom_boxplot(shape = "circle") +
 scale_fill_hue(direction = 1) +
 theme_minimal() +
 facet_wrap(vars(Gender))
```




```
library(dplyr)
library(ggplot2)
#potential defaulters and loan amount based on marital status
data %>%
 filter(!(`Marital status` %in% "missing")) %>%
 ggplot() +
 aes(x = PotentialDefaulter, y = `Loan Amount`, fill = `Marital status`) +
 geom_boxplot(shape = "circle") +
 scale_fill_hue(direction = 1) +
 theme_minimal()
```




```
#loan type based on loan closed or not
data %>%
 filter(!(`Marital status` %in% "missing")) %>%
 filter(!is.na(`Loan Closed`)) %>%
 filter(!(LoanType %in%
 "Missing/Other")) %>%
 ggplot() +
 aes(x = LoanType, y = `Loan Amount`, fill = `Loan Closed`) +
 geom_boxplot(shape = "circle") +
 scale_fill_hue(direction = 1) +
 theme_minimal()
```


100000
75000
50000
25000
Loan Amount
Bad Loan
Good Loan
LoanType
Loan Closed
No
Yes