# X18105149 Assignment

*Amarnath V*

*October 15, 2018*

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

## Abstract:

This report dicusses about the KDD [Fayyad, Piatetsky-Shapiro, and Smyth (1996)) applied on the financial data. This report covers the data exploration and transformation of financial data. It also explains the patterens of data which were identified by appying the KDD (Fayyad, Piatetsky-Shapiro, and Smyth (1996)) process with the help of data quality report.

## Motivation of Datasets:

### Dataset 01: Credit Card Default prediction

Dataset 01 - "Credit card default prediction as a classification problem" (Soui et al., 2018) is the Taiwanese data on which predictive analysis was performed to predict the response variable "default payment next month" which is a binary categorical data that denotes whether credible ot non credible clients. Using the novel sorting smoothing method, the probability of default is predicted. The prediction of response variable has been performed using artificial neural network (ANN).Artificial Neural Network (ANN) are the machine learning model which was insipired the behaviour and structure of human brain [Olden et al., 2008]. This type of ANN focuses on monitored learning, that allows the utlization of input and output datasets to continously varies the weights until the simulated output is as same as the predicted ones [Olaya M E.J., 2013] .This data has been been primarily selected to satisfy this assignment's requirements as mentioned below, so I will prefer to choose dataset 02 - bank marketing for the project rather than this dataset.

1. This dataset is relevant to financial sector
2. This dataset consists of more than 10 features and 20000 instances.

Dataset source: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/home

### Details of rows and columns of the dataset 01 ( 25 features and 30000 instances)

ID: ID of each client LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit SEX: Gender (1=male, 2=female) EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown) MARRIAGE: Marital status (1=married, 2=single, 3=others) AGE: Age in years PAY_1: Repayment status in September, 2005 (-2 = Balance paid in full and no transactions this period, -1= Balance paid in full, but account has a positive balance at end of period due to recent transactions for which payment has not yet come due, 0= Customer paid the minimum due amount, but not the entire balance, 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)- As per comment in the kaggle website - https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset/discussion/34608 PAY_2: Repayment status in August, 2005 (scale same as above) PAY_3: Repayment status in July, 2005 (scale same as above) PAY_4: Repayment status in June, 2005 (scale same as above) PAY_5: Repayment status in May, 2005 (scale same as above) PAY_6: Repayment status in April, 2005 (scale same as above) BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar) BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar) BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar) BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar) BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)

BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar) PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar) PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar) PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar) PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar) PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar) PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar) default.payment.next.month: Default payment (1=yes, 0=no)

**Dataset 02 - Bank Marketing**

This dataset - Bank Marketing [Moro et al., 2014] is related to "Bank Marketing". This dataset is improvised by adding the 5 new social and economic features in order to improve the predicton of success. This dataset has been selected because it can provide answer for comparative effectiveness research (CER).Usually among various marketing domains, customer segmentation(analysis) is considered important sector in research and organization practices. Different data mining techniques will be used to perform efficient marketing. RFM (Recency, frequence and monetary methods) technique is one of those technique used to perform customer segmentation which is most useful to produce marketing as discussed before [Olson, D.L. & Chae, B., 2012]. This dataset has been analysed by applying CRISP-DM methodology [Moro S., 2011].

**Details of rows and columns of the dataset 01 ( 21 features and 41188 instances)**

1 - age (numeric) 2 - job : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","s employed","services","student","technician","unemployed","unknown") 3 - marital : marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed) 4 - education (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown") 5 - default: has credit in default? (categorical: "no","yes","unknown") 6 - housing: has housing loan? (categorical: "no","yes","unknown") 7 - loan: has personal loan? (categorical: "no","yes","unknown") # related with the last contact of the current campaign: 8 - contact: contact communication type (categorical: "cellular","telephone") 9 - month: last contact month of year (categorical: "jan", "feb", "mar", . . . , "nov", "dec") 10 - day_of_week: last contact day of the week (categorical: "mon","tue","wed","thu","fri") 11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. # other attributes: 12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact) 13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted) 14 - previous: number of contacts performed before this campaign and for this client (numeric) 15 - poutcome: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success") # social and economic context attributes 16 - emp.var.rate: employment variation rate - quarterly indicator (numeric) 17 - cons.price.idx: consumer price index - monthly indicator (numeric)
18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19 - euribor3m: euribor 3 month rate - daily indicator (numeric) 20 - nr.employed: number of employees - quarterly indicator (numeric) 21 - y: response variable

Dataset source: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

**Reading the Datasets:**

**Dataset 01:**

```
data1 <- read.csv(file="C://Users//admin//Desktop//Data Analytics//Data1.csv")
```

**Dataset 02:**

```r
data2 <- read.csv(file="C://Users//admin//Desktop//Data Analytics//Data2a.csv", na.strings = c("unknown
```

## Exploration:

### Verifying metadata of datasets:

**Dataset 01:**

As per repository page, dataset01 must have 25 features and 30000 instances. Response variable is "default.payment.next.month" which is binary categorical.

X (X1,. . . .,X24) is, (ID,LIMIT_BAL,SEX,EDUCATION,MARRIAGE,AGE,PAY_0,PAY_2,PAY_3,PAY_4,PAY_5,PAY_6,B

Among 25 features, SEX, EDUCATION, MARRIAGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, default.payment.next.month must be categorical features.

### Structure of the Datasets:

**Dataset 01:**

```r
str(data1)
```

```
## 'data.frame':    30000 obs. of  25 variables:
##  $ ID                        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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                 : int  2 2 2 2 2 1 1 2 3 3 ...
##  $ MARRIAGE                  : int  1 2 2 1 1 2 2 2 1 2 ...
##  $ 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.payment.next.month: int  1 1 0 0 0 0 0 0 0 0 ...
```

**Dataset 02:**

```r
str(data2)
```

```
## 'data.frame':    41188 obs. of  21 variables:
##  $ age           : int  56 57 37 40 56 45 59 41 24 25 ...
##  $ job           : Factor w/ 11 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ...
```

```
##  $ marital       : Factor w/ 3 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...
##  $ education     : Factor w/ 7 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 NA 6 4 ...
##  $ default       : Factor w/ 2 levels "no","yes": 1 NA 1 1 1 NA 1 NA 1 1 ...
##  $ housing       : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 1 1 2 2 ...
##  $ loan          : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 1 1 1 1 ...
##  $ contact       : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...
##  $ month         : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ day_of_week   : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ duration      : int  261 149 226 151 307 198 139 217 380 50 ...
##  $ campaign      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays         : int  999 999 999 999 999 999 999 999 999 999 ...
##  $ previous      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome      : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ emp.var.rate  : num  1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
##  $ cons.price.idx: num  94 94 94 94 94 ...
##  $ cons.conf.idx : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
##  $ euribor3m     : num  4.86 4.86 4.86 4.86 4.86 ...
##  $ nr.employed   : num  5191 5191 5191 5191 5191 ...
##  $ y             : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

**Dataset 01:**

Dataset01 has 25 features and 30000 instances. By looking at the struture of dataset01 we can see that features, SEX, EDUCATION, MARRIAGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, default.payment.next.month are not categorical features which are supposed to be the categorical features. So,we must convert these features into categorical using the function as.factor(). In addition to these, the feature "ID" acts as just serial number which is not contributing anything to predict the predictive variable so we will go with 24 features by eliminating it.

```r
data1 <- data1[,2:ncol(data1)]
str(data1)    #structure of data1 after removing "ID"
```

```
## '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                 : int  2 2 2 2 2 1 1 2 3 3 ...
##  $ MARRIAGE                  : int  1 2 2 1 1 2 2 2 1 2 ...
##  $ 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.payment.next.month: int  1 1 0 0 0 0 0 0 0 0 ...
```

```r
data1$SEX <- as.factor(data1$SEX)
data1$EDUCATION <- as.factor(data1$EDUCATION)
data1$MARRIAGE <- as.factor(data1$MARRIAGE)
data1$PAY_0 <- as.factor(data1$PAY_0)
data1$PAY_2 <- as.factor(data1$PAY_2)
data1$PAY_3 <- as.factor(data1$PAY_3)
data1$PAY_4 <- as.factor(data1$PAY_4)
data1$PAY_5 <- as.factor(data1$PAY_5)
data1$PAY_6 <- as.factor(data1$PAY_6)
data1$default.payment.next.month <- as.factor(data1$default.payment.next.month)
str(data1)
```

```
## 'data.frame':    30000 obs. of  24 variables:
##  $ LIMIT_BAL                 : int  20000 120000 90000 50000 50000 50000 500000 100000 140000 20000
##  $ SEX                       : Factor w/ 2 levels "1","2": 2 2 2 2 1 1 1 2 2 1 ...
##  $ EDUCATION                 : Factor w/ 7 levels "0","1","2","3",..: 3 3 3 3 3 2 2 3 4 4 ...
##  $ MARRIAGE                  : Factor w/ 4 levels "0","1","2","3": 2 3 3 2 2 3 3 3 2 3 ...
##  $ AGE                       : int  24 26 34 37 57 37 29 23 28 35 ...
##  $ PAY_0                     : Factor w/ 11 levels "-2","-1","0",..: 5 2 3 3 2 3 3 3 3 1 ...
##  $ PAY_2                     : Factor w/ 11 levels "-2","-1","0",..: 5 5 3 3 3 3 3 2 3 1 ...
##  $ PAY_3                     : Factor w/ 11 levels "-2","-1","0",..: 2 3 3 3 2 3 3 2 5 1 ...
##  $ PAY_4                     : Factor w/ 11 levels "-2","-1","0",..: 2 3 3 3 3 3 3 3 3 1 ...
##  $ PAY_5                     : Factor w/ 10 levels "-2","-1","0",..: 1 3 3 3 3 3 3 3 3 2 ...
##  $ PAY_6                     : Factor w/ 10 levels "-2","-1","0",..: 1 4 3 3 3 3 3 2 3 2 ...
##  $ 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.payment.next.month: Factor w/ 2 levels "0","1": 2 2 1 1 1 1 1 1 1 1 ...
```

Now, we can check the structure of the dataset 01, all those 10 features became categorical features as expected. Here, responding variable is "default payment next month" which is a binary categorical data. X(X1,X2,......,X25) are, (ID,LIMIT_BAL,SEX,EDUCATION,MARRIAGE,AGE,PAY_0,PAY_2,PAY_3,PAY_4,PAY_5,PA Among these SEX, EDUCATION, MARRIAGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, default.payment.next.month are categorical features. ###Dataset 02: #Dataset02 has 21 features and 41188 instances. #####Here, responding variable is "Y" which is a binary categorical data.

**X(X1,........,X31)are, (age,job,marital,education,default,housing,loan,contact,month,day_of_week,duration this dataset exhibit almost all fundamental data types.**

**Treating Missing Attribute Values:**

**There are several missing values in some categorical attributes. These missing values can be treated using imputation techniques.**

**Missing value:**

A missing value is one whose value is unknown. Missing values are represented in R by the NA symbol. NA is a special value whose properties are different from other values. NA is one of the very few reserved words in R: you cannot give anything this name. (Source: http://faculty.nps.edu/sebuttre/home/R/missings.html)

**Finding missing value for dataset 01:**

In the Dataset 01, the "education" feature has categorical values from 1 to 6 among these 5 and 6 are unknown and in addition to these 0 also exists in education feature which is also a unknown value. Like dataset 02, dataset 01's source website doesn't mentioned direcly that unknown values are missing values. So, let's make an assumption that unknown values here are missing values and process further to generate Data Quality Report by performing transformation and also generate another DQR without transforming data.

**Tranformation of data for dataset 01**

```
data1a <- data1
#Converting 0's,5's and 6's to NA's in the feature data1a$EDUCATION
data1a$EDUCATION <- sapply(data1a$EDUCATION, FUN = function(x) {if(x == 0 | x == 5 | x == 6) {x <- NA} e
#Conerting the EDUCATION feature back to factor datatype
data1a$EDUCATION <- as.factor(data1a$EDUCATION)
#Summarise of EDUCATION feature of Dataset 01
summary(data1a$EDUCATION)
```

```
##     2     3     4     5  NA's
## 10585 14030  4917   123   345
```

```
#Converting 0's to 3's
data1a$MARRIAGE <- sapply(data1a$MARRIAGE, FUN = function(x) {if(x == 0) {x <- NA} else {x <- x}})
data1a$MARRIAGE <- as.factor(data1a$MARRIAGE)
summary((data1a$MARRIAGE))
```

```
##     2     3     4  NA's
## 13659 15964   323    54
```

So, Here we have transformed unknown values(0, 5, 6) of the feature EDUCATION to NA's. And also for a feature MARRIAGE we have converted unknow value "0's" to "NA's".

**Volume of data missing in the dataset 01:**

```
missing1 <- sapply(data1a, FUN = function(x) {sum(is.na(x) / length(x) * 100)})
#Volume of missing value
missing1
```

```
##                  LIMIT_BAL                        SEX
##                       0.00                       0.00
##                  EDUCATION                   MARRIAGE
##                       1.15                       0.18
##                        AGE                      PAY_0
##                       0.00                       0.00
##                      PAY_2                      PAY_3
##                       0.00                       0.00
##                      PAY_4                      PAY_5
##                       0.00                       0.00
##                      PAY_6                   BILL_AMT1
##                       0.00                       0.00
```

```
##                   BILL_AMT2                  BILL_AMT3
##                        0.00                       0.00
##                   BILL_AMT4                  BILL_AMT5
##                        0.00                       0.00
##                   BILL_AMT6                   PAY_AMT1
##                        0.00                       0.00
##                    PAY_AMT2                   PAY_AMT3
##                        0.00                       0.00
##                    PAY_AMT4                   PAY_AMT5
##                        0.00                       0.00
##                    PAY_AMT6 default.payment.next.month
##                        0.00                       0.00
```

**Finding missing value for dataset 02:**

As mentioned in the source website, in this dataset missing values are labelled as "unknown" which were replaced by "N/A" while reading the dataset itself. Now, let us find the features that have missing value using the below command.

**Volume of data missing in the dataset 02:**

```
missing2 <- sapply(data2, FUN = function(x) {sum(is.na(x) / length(x) * 100)})
#Volume of missing value
missing2
```

```
##           age           job       marital     education       default
##     0.0000000     0.8012042     0.1942313     4.2026804    20.8725842
##       housing          loan       contact         month   day_of_week
##     2.4036127     2.4036127     0.0000000     0.0000000     0.0000000
##      duration      campaign         pdays      previous      poutcome
##     0.0000000     0.0000000     0.0000000     0.0000000     0.0000000
##  emp.var.rate cons.price.idx cons.conf.idx     euribor3m   nr.employed
##     0.0000000     0.0000000     0.0000000     0.0000000     0.0000000
##             y
##     0.0000000
```

## Transformation:

**Treating the missing value of dataset 01:**

```
MaxTable <- function(x){
     dd <- unique(x)
     dd[which.max(tabulate(match(x,dd)))]
}
```

Using this MaxTable function we can replace missing values with the most frequent value in categorical data. As we got features Education and MARRIAGE (Both are Categorical data) in the dataset 01 so we can use this imputation technique.

```
data1a$EDUCATION[is.na(data1a$EDUCATION)] <- MaxTable(data1a$EDUCATION)
data1a$MARRIAGE[is.na(data1a$MARRIAGE)] <- MaxTable(data1a$MARRIAGE)
summary(data1a$EDUCATION)
```

```
##     2     3     4     5
## 10585 14375  4917   123
```

```r
summary(data1a$MARRIAGE)
```

```
##     2     3     4
## 13659 16018   323
```

```r
summary(data1a)
```

```
##    LIMIT_BAL       SEX       EDUCATION MARRIAGE       AGE
##  Min.   :  10000   1:11888   2:10585   2:13659   Min.   :21.00
##  1st Qu.:  50000   2:18112   3:14375   3:16018   1st Qu.:28.00
##  Median : 140000             4: 4917   4:  323   Median :34.00
##  Mean   : 167484             5:  123             Mean   :35.49
##  3rd Qu.: 240000                                 3rd Qu.:41.00
##  Max.   :1000000                                 Max.   :79.00
##
##      PAY_0           PAY_2           PAY_3           PAY_4
##  0      :14737   0      :15730   0      :15764   0      :16455
##  -1     : 5686   -1     : 6050   -1     : 5938   -1     : 5687
##  1      : 3688   2      : 3927   -2     : 4085   -2     : 4348
##  -2     : 2759   -2     : 3782   2      : 3819   2      : 3159
##  2      : 2667   3      :  326   3      :  240   3      :  180
##  3      :  322   4      :   99   4      :   76   4      :   69
##  (Other):  141   (Other):   86   (Other):   78   (Other):  102
##      PAY_5           PAY_6         BILL_AMT1         BILL_AMT2
##  0      :16947   0      :16286   Min.   :-165580   Min.   :-69777
##  -1     : 5539   -1     : 5740   1st Qu.:   3559   1st Qu.:  2985
##  -2     : 4546   -2     : 4895   Median :  22382   Median : 21200
##  2      : 2626   2      : 2766   Mean   :  51223   Mean   : 49179
##  3      :  178   3      :  184   3rd Qu.:  67091   3rd Qu.: 64006
##  4      :   84   4      :   49   Max.   : 964511   Max.   :983931
##  (Other):   80   (Other):   80
##    BILL_AMT3         BILL_AMT4         BILL_AMT5        BILL_AMT6
##  Min.   :-157264   Min.   :-170000   Min.   :-81334   Min.   :-339603
##  1st Qu.:   2666   1st Qu.:   2327   1st Qu.:  1763   1st Qu.:   1256
##  Median :  20089   Median :  19052   Median : 18105   Median :  17071
##  Mean   :  47013   Mean   :  43263   Mean   : 40311   Mean   :  38872
##  3rd Qu.:  60165   3rd Qu.:  54506   3rd Qu.: 50191   3rd Qu.:  49198
##  Max.   :1664089   Max.   : 891586   Max.   :927171   Max.   : 961664
##
##     PAY_AMT1         PAY_AMT2          PAY_AMT3         PAY_AMT4
##  Min.   :     0   Min.   :      0   Min.   :     0   Min.   :     0
##  1st Qu.:  1000   1st Qu.:    833   1st Qu.:   390   1st Qu.:   296
##  Median :  2100   Median :   2009   Median :  1800   Median :  1500
##  Mean   :  5664   Mean   :   5921   Mean   :  5226   Mean   :  4826
##  3rd Qu.:  5006   3rd Qu.:   5000   3rd Qu.:  4505   3rd Qu.:  4013
##  Max.   :873552   Max.   :1684259   Max.   :896040   Max.   :621000
##
##     PAY_AMT5           PAY_AMT6         default.payment.next.month
##  Min.   :     0.0   Min.   :     0.0   0:23364
##  1st Qu.:   252.5   1st Qu.:   117.8   1: 6636
##  Median :  1500.0   Median :  1500.0
##  Mean   :  4799.4   Mean   :  5215.5
##  3rd Qu.:  4031.5   3rd Qu.:  4000.0
##  Max.   :426529.0   Max.   :528666.0
```

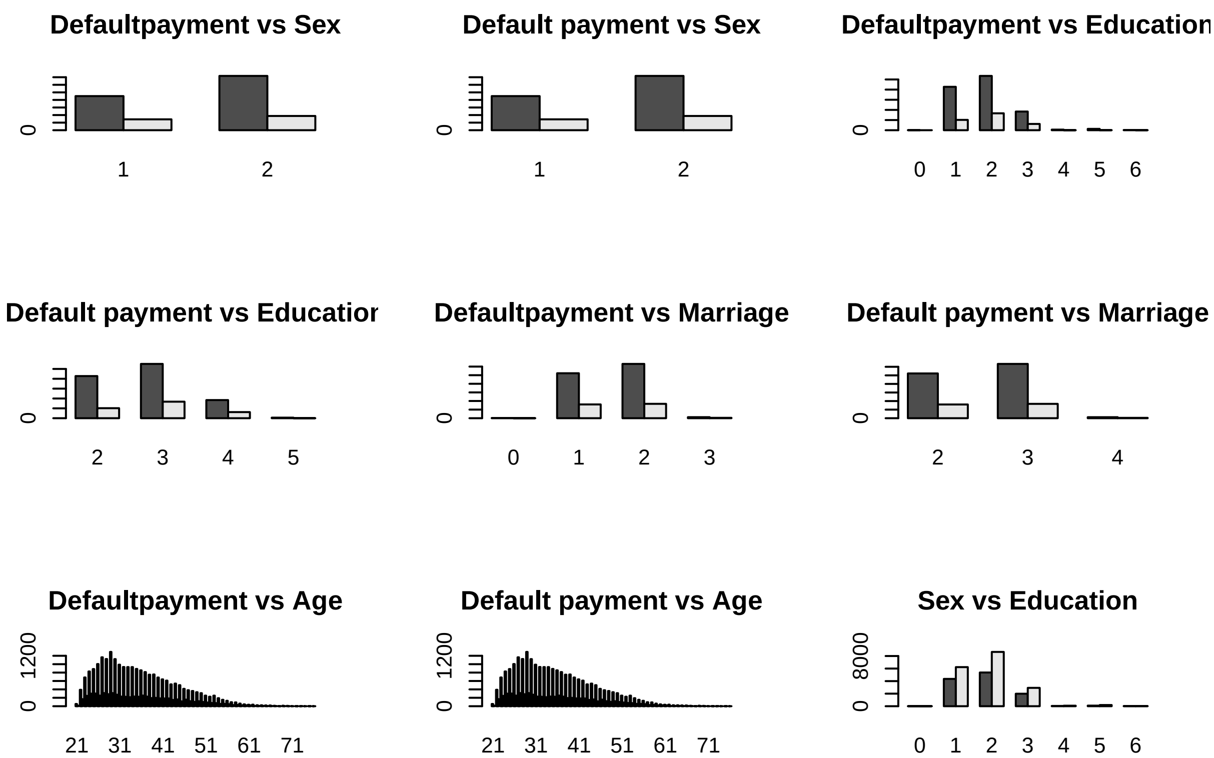
```
##
```

Now, by checking the summary of the features EDUCATION and MARRIAGE as well as whole dataset01 we can ensure that all assumed missing vlaues were replaced with most frequent value in the categorical features EDUCATION and MARRIAGE and there is no missing value in the dataset 01.

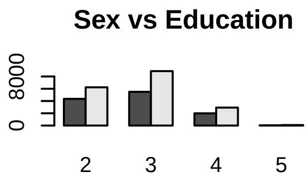
**Relationship between features of the Dataset 01 with and without transformation:**

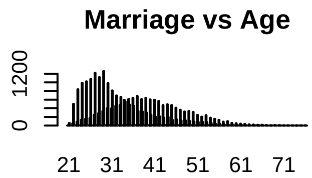
Let's discuss about the relationship between various features of the dataset 01 using graphs.

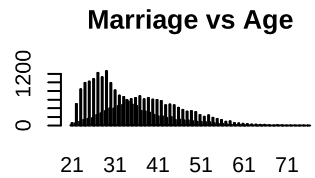
```
par(mfrow = c(3,3))
#Default payment vs Sex
barplot(table(data1$default.payment.next.month,data1$SEX),beside = T, main = "Defaultpayment vs Sex")
barplot(table(data1a$default.payment.next.month,data1a$SEX),beside = T, main = "Default payment vs Sex")

#Default payment vs Education
barplot(table(data1$default.payment.next.month,data1$EDUCATION),beside = T, main = "Defaultpayment vs Ed
barplot(table(data1a$default.payment.next.month,data1a$EDUCATION),beside = T, main = "Default payment vs

#Default payment vs Marriage status
barplot(table(data1$default.payment.next.month,data1$MARRIAGE),beside = T, main = "Defaultpayment vs Mar
barplot(table(data1a$default.payment.next.month,data1a$MARRIAGE),beside = T, main = "Default payment vs

#Default payment vs Age
barplot(table(data1$default.payment.next.month,data1$AGE),beside = T, main = "Defaultpayment vs Age")
barplot(table(data1a$default.payment.next.month,data1a$AGE),beside = T, main = "Default payment vs Age")

#Sex vs Education
barplot(table(data1$SEX,data1$EDUCATION),beside = T, main = "Sex vs Education")
```




```
barplot(table(data1a$SEX,data1a$EDUCATION),beside = T, main = "Sex vs Education")

#Marriage status vs Age
barplot(table(data1$MARRIAGE,data1$AGE),beside = T, main = "Marriage vs Age")
barplot(table(data1a$MARRIAGE,data1a$AGE),beside = T, main = "Marriage vs Age")
```




In the above first four graphs, dark shaded are customers not opted for default payment and light shaded are customer opted for default payment.In the fifth graph dark shade represents male (1) and light shade represents female. In the sixth graph, dark shaded represents married(1), light shaded represents single(2) and white represents others(3).

**Transformation of data for dataset 02**

So here, we can see 6 features namely job, martial, education, default, housing and loan are having missing values. All these features are categorized variable so we can replace missing value with the most frequent class value. To find the most frequent class value let us apply the MaxTable function for each of the features that has missing value,

So, here we going to replace missing value with the value returned by using the MaxTable function to the features that have missing value. Firstly, will have a look at the dataset before performing imputation.

```r
summary(data2)
```

```
##       age                 job            marital     
##  Min.   :17.00   admin.     :10422   divorced: 4612  
##  1st Qu.:32.00   blue-collar: 9254   married :24928  
##  Median :38.00   technician : 6743   single  :11568  
##  Mean   :40.02   services   : 3969   NA's    :   80  
##  3rd Qu.:47.00   management : 2924                   
##  Max.   :98.00   (Other)    : 7546                   
##                  NA's       :  330                   
##                education     default      housing        loan      
##  university.degree  :12168   no  :32588   no  :18622   no  :33950  
##  high.school        : 9515   yes :    3   yes :21576   yes : 6248  
```

```
##  basic.9y           : 6045   NA's: 8597   NA's:  990   NA's:  990
##  professional.course: 5243
##  basic.4y           : 4176
##  (Other)            : 2310
##  NA's               : 1731
##       contact           month       day_of_week    duration
##  cellular :26144   may    :13769   fri:7827    Min.   :   0.0
##  telephone:15044   jul    : 7174   mon:8514    1st Qu.: 102.0
##                    aug    : 6178   thu:8623    Median : 180.0
##                    jun    : 5318   tue:8090    Mean   : 258.3
##                    nov    : 4101   wed:8134    3rd Qu.: 319.0
##                    apr    : 2632               Max.   :4918.0
##                    (Other): 2016
##     campaign          pdays          previous           poutcome
##  Min.   : 1.000   Min.   :  0.0   Min.   :0.000   failure    : 4252
##  1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000   nonexistent:35563
##  Median : 2.000   Median :999.0   Median :0.000   success    : 1373
##  Mean   : 2.568   Mean   :962.5   Mean   :0.173
##  3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000
##  Max.   :56.000   Max.   :999.0   Max.   :7.000
##
##   emp.var.rate      cons.price.idx  cons.conf.idx     euribor3m
##  Min.   :-3.40000   Min.   :92.20   Min.   :-50.8   Min.   :0.634
##  1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344
##  Median : 1.10000   Median :93.75   Median :-41.8   Median :4.857
##  Mean   : 0.08189   Mean   :93.58   Mean   :-40.5   Mean   :3.621
##  3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961
##  Max.   : 1.40000   Max.   :94.77   Max.   :-26.9   Max.   :5.045
##
##   nr.employed     y
##  Min.   :4964   no :36548
##  1st Qu.:5099   yes: 4640
##  Median :5191
##  Mean   :5167
##  3rd Qu.:5228
##  Max.   :5228
##
```

As we mentioned above, we can see 6 features are having missing values. ####job:

```
data2$job[is.na(data2$job)] <- MaxTable(data2$job)
```

#### marital:

```
data2$marital[is.na(data2$marital)] <- MaxTable(data2$marital)
```

#### education:

```
data2$education[is.na(data2$education)] <- MaxTable(data2$education)
```

#### default:

```
data2$default[is.na(data2$default)] <- MaxTable(data2$default)
```

#### housing;

```r
data2$housing[is.na(data2$housing)] <- MaxTable(data2$housing)
```

**loan:**

```r
data2$loan[is.na(data2$loan)] <- MaxTable(data2$loan)
```

Let's have a look at the dataset after performing imputation to the missing values. Now there should not be any missing values in the dataset

```r
summary(data2)
```

```
##       age                  job            marital     
##  Min.   :17.00   admin.     :10752   divorced: 4612  
##  1st Qu.:32.00   blue-collar: 9254   married :25008  
##  Median :38.00   technician : 6743   single  :11568  
##  Mean   :40.02   services   : 3969                   
##  3rd Qu.:47.00   management : 2924                   
##  Max.   :98.00   retired    : 1720                   
##                  (Other)    : 5826                   
##                education     default      housing       loan      
##  basic.4y           : 4176   no :41185   no :18622   no :34940  
##  basic.6y           : 2292   yes:    3   yes:22566   yes: 6248  
##  basic.9y           : 6045                                      
##  high.school        : 9515                                      
##  illiterate         :   18                                      
##  professional.course: 5243                                      
##  university.degree  :13899                                      
##       contact          month        day_of_week    duration     
##  cellular :26144   may    :13769   fri:7827    Min.   :   0.0  
##  telephone:15044   jul    : 7174   mon:8514    1st Qu.: 102.0  
##                    aug    : 6178   thu:8623    Median : 180.0  
##                    jun    : 5318   tue:8090    Mean   : 258.3  
##                    nov    : 4101   wed:8134    3rd Qu.: 319.0  
##                    apr    : 2632               Max.   :4918.0  
##                    (Other): 2016                                
##     campaign          pdays          previous            poutcome    
##  Min.   : 1.000   Min.   :  0.0   Min.   :0.000   failure    : 4252  
##  1st Qu.: 1.000   1st Qu.:999.0   1st Qu.:0.000   nonexistent:35563  
##  Median : 2.000   Median :999.0   Median :0.000   success    : 1373  
##  Mean   : 2.568   Mean   :962.5   Mean   :0.173                      
##  3rd Qu.: 3.000   3rd Qu.:999.0   3rd Qu.:0.000                      
##  Max.   :56.000   Max.   :999.0   Max.   :7.000                      
##                                                                     
##   emp.var.rate      cons.price.idx  cons.conf.idx     euribor3m    
##  Min.   :-3.40000   Min.   :92.20   Min.   :-50.8   Min.   :0.634  
##  1st Qu.:-1.80000   1st Qu.:93.08   1st Qu.:-42.7   1st Qu.:1.344  
##  Median : 1.10000   Median :93.75   Median :-41.8   Median :4.857  
##  Mean   : 0.08189   Mean   :93.58   Mean   :-40.5   Mean   :3.621  
##  3rd Qu.: 1.40000   3rd Qu.:93.99   3rd Qu.:-36.4   3rd Qu.:4.961  
##  Max.   : 1.40000   Max.   :94.77   Max.   :-26.9   Max.   :5.045  
##                                                                    
##   nr.employed     y        
##  Min.   :4964   no :36548  
##  1st Qu.:5099   yes: 4640  
##  Median :5191              
```

```
##  Mean   :5167
##  3rd Qu.:5228
##  Max.   :5228
##
```

After performing imputation let's have a look at the datset to ensure whether fundamental datatypes of dataset are correct and the categorical data has appropriate labels. This can be done by looking into structure of this dataset,

```r
str(data2)
```

```
## 'data.frame':    41188 obs. of  21 variables:
##  $ age           : int  56 57 37 40 56 45 59 41 24 25 ...
##  $ job           : Factor w/ 11 levels "admin.","blue-collar",..: 4 8 8 1 8 8 1 2 10 8 ...
##  $ marital       : Factor w/ 3 levels "divorced","married",..: 2 2 2 2 2 2 2 2 3 3 ...
##  $ education     : Factor w/ 7 levels "basic.4y","basic.6y",..: 1 4 4 2 4 3 6 7 6 4 ...
##  $ default       : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ housing       : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 1 1 2 2 ...
##  $ loan          : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 1 1 1 1 ...
##  $ contact       : Factor w/ 2 levels "cellular","telephone": 2 2 2 2 2 2 2 2 2 2 ...
##  $ month         : Factor w/ 10 levels "apr","aug","dec",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ day_of_week   : Factor w/ 5 levels "fri","mon","thu",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ duration      : int  261 149 226 151 307 198 139 217 380 50 ...
##  $ campaign      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays         : int  999 999 999 999 999 999 999 999 999 999 ...
##  $ previous      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome      : Factor w/ 3 levels "failure","nonexistent",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ emp.var.rate  : num  1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 1.1 ...
##  $ cons.price.idx: num  94 94 94 94 94 ...
##  $ cons.conf.idx : num  -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 -36.4 ...
##  $ euribor3m     : num  4.86 4.86 4.86 4.86 4.86 ...
##  $ nr.employed   : num  5191 5191 5191 5191 5191 ...
##  $ y             : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

**Finding useful feature in the dataset 02:**

Let us find the feature with more unique values,

```r
use2 <- sapply(data2, FUN = function(x) {length(unique(x))})
use2
```

```
##            age            job        marital      education        default
##             78             11              3              7              2
##        housing           loan        contact          month    day_of_week
##              2              2              2             10              5
##       duration       campaign          pdays       previous       poutcome
##           1544             42             27              8              3
##   emp.var.rate cons.price.idx  cons.conf.idx      euribor3m    nr.employed
##             10             26             26            316             11
##              y
##              2
```

So, the "duration" feature has the most number of unique values in the dataset. When comparing "duration" with "y(response variable)" we can even find "duration" feature has serious impact on "y" by looking at the dataset itself. For instance, "y" will be always 0 when "duration" is 0.But with this information we can say that this is useful feature but not perfect feature because even higher "duration" value end up in no for "y".

**Outliers;**

An outlier is an observation in a data set that lies a substantial distance from other observations. These unusual observations can have a disproportionate effect on statistical analysis, such as the mean, which can lead to misleading results. Outliers can provide useful information about your data or process, so it's important to investigate them. Of course, you have to find them first.

**Outliers in the dataset 01:**

Dataset 01 has binary categorical response variable and most of its useful features are categorical data, so finding outliers in those categorical data is tricky, so let us assume the least occuring value in the categorical data as outliers and we will create the DQR with and without outlier from the tranformed data of dateset 01(data1a).

In this dataset we can identify the least occuring values of categorical features which is nothing but other categories ("4" in EDUCATION and "3" in the feature MARRIAGE), this can be ensured by checking the summary of the dataset.

```r
#Cloning the transformed dataset 01
data1b <- data1a
#Summary of clone of transformed dataset 01
summary(data1b)
```

```
##    LIMIT_BAL        SEX       EDUCATION MARRIAGE       AGE
##  Min.   :  10000   1:11888   2:10585   2:13659   Min.   :21.00
##  1st Qu.:  50000   2:18112   3:14375   3:16018   1st Qu.:28.00
##  Median : 140000             4: 4917   4:  323   Median :34.00
##  Mean   : 167484             5:  123             Mean   :35.49
##  3rd Qu.: 240000                                 3rd Qu.:41.00
##  Max.   :1000000                                 Max.   :79.00
##
##      PAY_0           PAY_2           PAY_3           PAY_4
##  0      :14737   0      :15730   0      :15764   0      :16455
##  -1     : 5686   -1     : 6050   -1     : 5938   -1     : 5687
##  1      : 3688   2      : 3927   -2     : 4085   -2     : 4348
##  -2     : 2759   -2     : 3782   2      : 3819   2      : 3159
##  2      : 2667   3      :  326   3      :  240   3      :  180
##  3      :  322   4      :   99   4      :   76   4      :   69
##  (Other):  141   (Other):   86   (Other):   78   (Other):  102
##      PAY_5           PAY_6          BILL_AMT1         BILL_AMT2
##  0      :16947   0      :16286   Min.   :-165580   Min.   :-69777
##  -1     : 5539   -1     : 5740   1st Qu.:   3559   1st Qu.:  2985
##  -2     : 4546   -2     : 4895   Median :  22382   Median : 21200
##  2      : 2626   2      : 2766   Mean   :  51223   Mean   : 49179
##  3      :  178   3      :  184   3rd Qu.:  67091   3rd Qu.: 64006
##  4      :   84   4      :   49   Max.   : 964511   Max.   :983931
##  (Other):   80   (Other):   80
##    BILL_AMT3         BILL_AMT4          BILL_AMT5        BILL_AMT6
##  Min.   :-157264   Min.   :-170000   Min.   :-81334   Min.   :-339603
##  1st Qu.:   2666   1st Qu.:   2327   1st Qu.:  1763   1st Qu.:   1256
##  Median :  20089   Median :  19052   Median : 18105   Median :  17071
##  Mean   :  47013   Mean   :  43263   Mean   : 40311   Mean   :  38872
##  3rd Qu.:  60165   3rd Qu.:  54506   3rd Qu.: 50191   3rd Qu.:  49198
##  Max.   :1664089   Max.   : 891586   Max.   :927171   Max.   : 961664
##
##     PAY_AMT1         PAY_AMT2          PAY_AMT3         PAY_AMT4
```

```
##  Min.   :     0   Min.   :      0   Min.   :     0   Min.   :     0
##  1st Qu.:  1000   1st Qu.:    833   1st Qu.:   390   1st Qu.:   296
##  Median :  2100   Median :   2009   Median :  1800   Median :  1500
##  Mean   :  5664   Mean   :   5921   Mean   :  5226   Mean   :  4826
##  3rd Qu.:  5006   3rd Qu.:   5000   3rd Qu.:  4505   3rd Qu.:  4013
##  Max.   :873552   Max.   :1684259   Max.   :896040   Max.   :621000
##
##     PAY_AMT5           PAY_AMT6        default.payment.next.month
##  Min.   :     0.0   Min.   :     0.0   0:23364
##  1st Qu.:   252.5   1st Qu.:   117.8   1: 6636
##  Median :  1500.0   Median :  1500.0
##  Mean   :  4799.4   Mean   :  5215.5
##  3rd Qu.:  4031.5   3rd Qu.:  4000.0
##  Max.   :426529.0   Max.   :528666.0
##
```

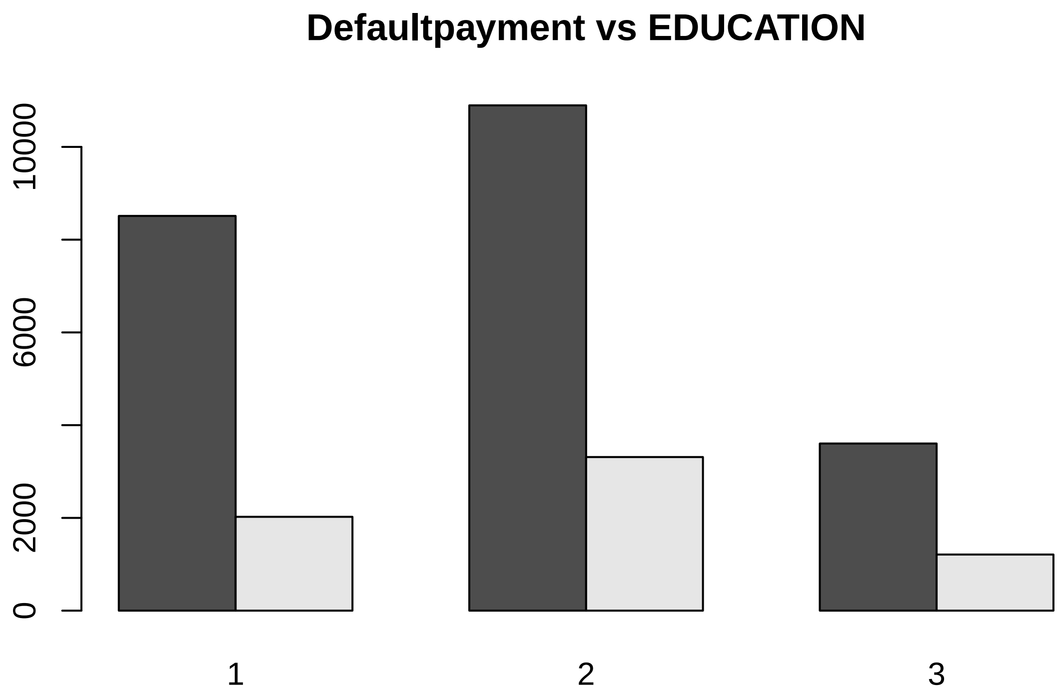
```r
#Declaration of function to return least frequent value in the categorical variable.
MinTable <- function(x){
      dd <- unique(x)
      dd[which.min(tabulate(match(x,dd)))]
}


#Transforming least frequent value to NA's for EDUCATION feature
data1b$EDUCATION <- sapply(data1b$EDUCATION, FUN = function(x) {if(x == MinTable(data1b$EDUCATION)) {x
data1b$EDUCATION <- as.factor(data1b$EDUCATION)   #converting back to categorical
summary(data1b$EDUCATION)  #summarizing to check
```

```
##     1     2     3  NA's
## 10585 14375  4917   123
```

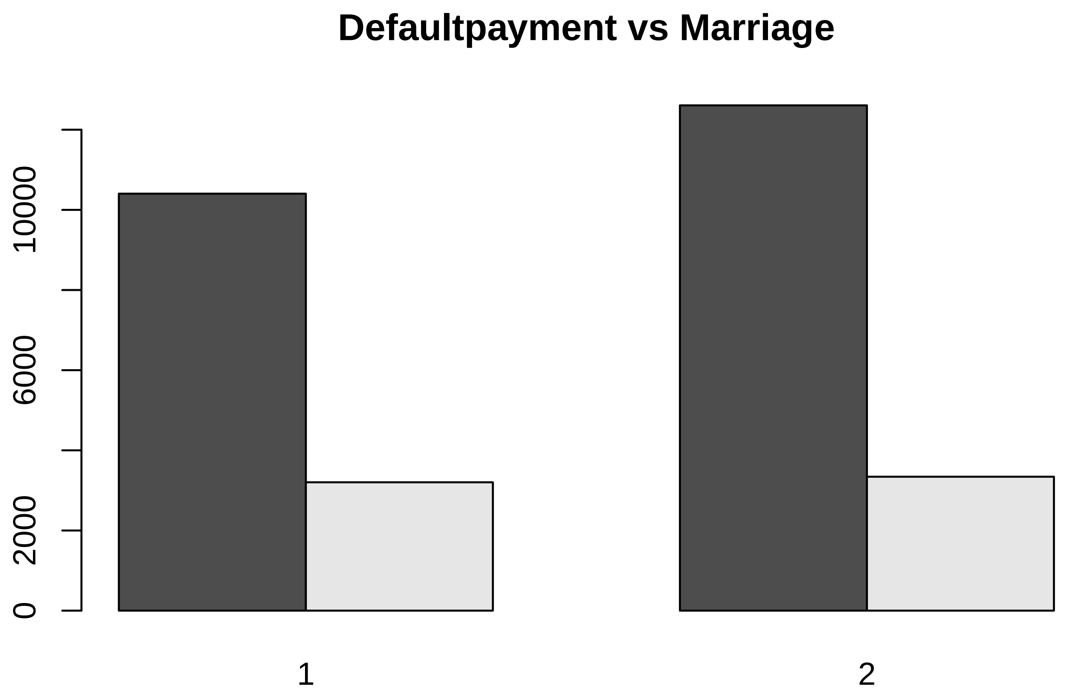
```r
#Transforming least frequent value to NA's for MARRIAGE feature
data1b$MARRIAGE <- sapply(data1b$MARRIAGE, FUN = function(x) {if(x == MinTable(data1b$MARRIAGE)) {x <- 
data1b$MARRIAGE <- as.factor(data1b$MARRIAGE)   #converting back to categorical
summary(data1b$MARRIAGE)   #summarizing to check
```

```
##     1     2  NA's
## 13659 16018   323
```

```r
#Removing entire rows in the dataset with NA's
data1b <- na.omit(data1b)

#Let's check whether outliers we removed via plotting graph
barplot(table(data1b$default.payment.next.month,data1b$EDUCATION),beside = T, main="Defaultpayment vs E
```



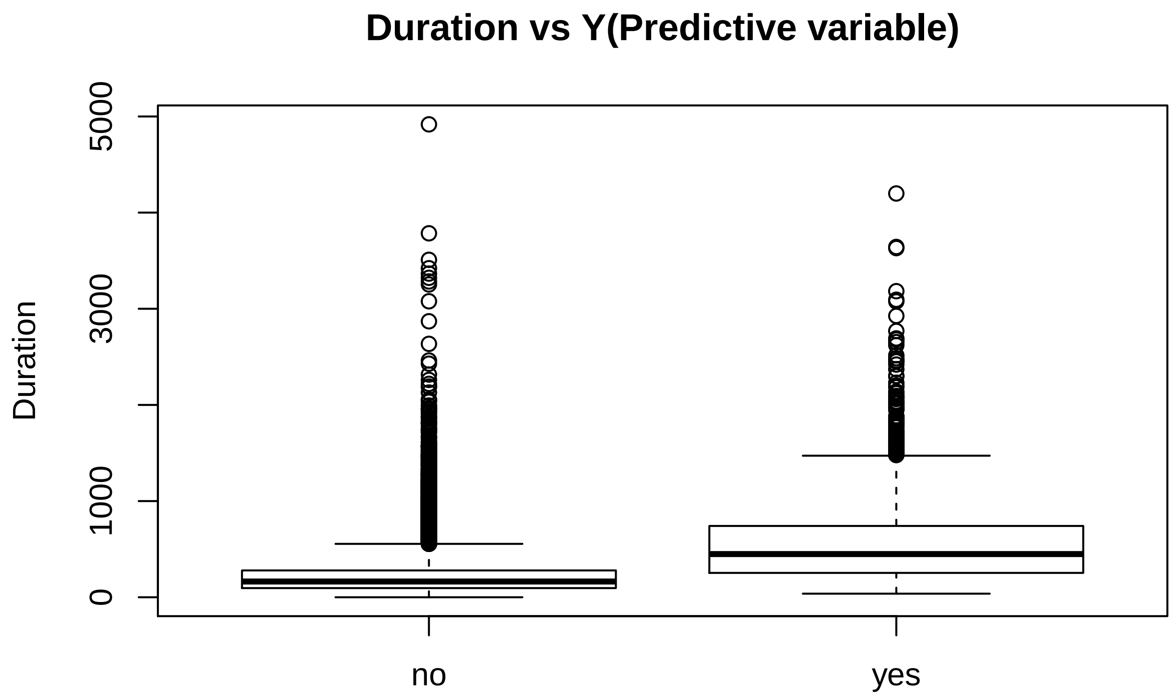

```
barplot(table(data1b$default.payment.next.month,data1b$MARRIAGE),beside = T, main="Defaultpayment vs Ma
```




**Outliers in the dataset 02:**

Let's visualise the relationship between "duration" and "y",

```
boxplot(data2$duration ~ data2$y, ylab='Duration', main='Duration vs Y(Predictive variable)')
```




We can see, most of the values stays between 0 to 2000 and there are more outliers in the box plot which has to be treated. In order to treat this, let's create a benchmark and remove the values that falls beyond the benchmark. Benchmark can be calucalted by the following formula, Benchmark = third-quantile + (1.5 * IQR(x)), where, IQR = Interquantile Range. Instead of disturbing existing data frame we can make a copy and remove outliers in it.

```
data2a <- data2
#to find third quantile
quantile(data2a$duration)
```

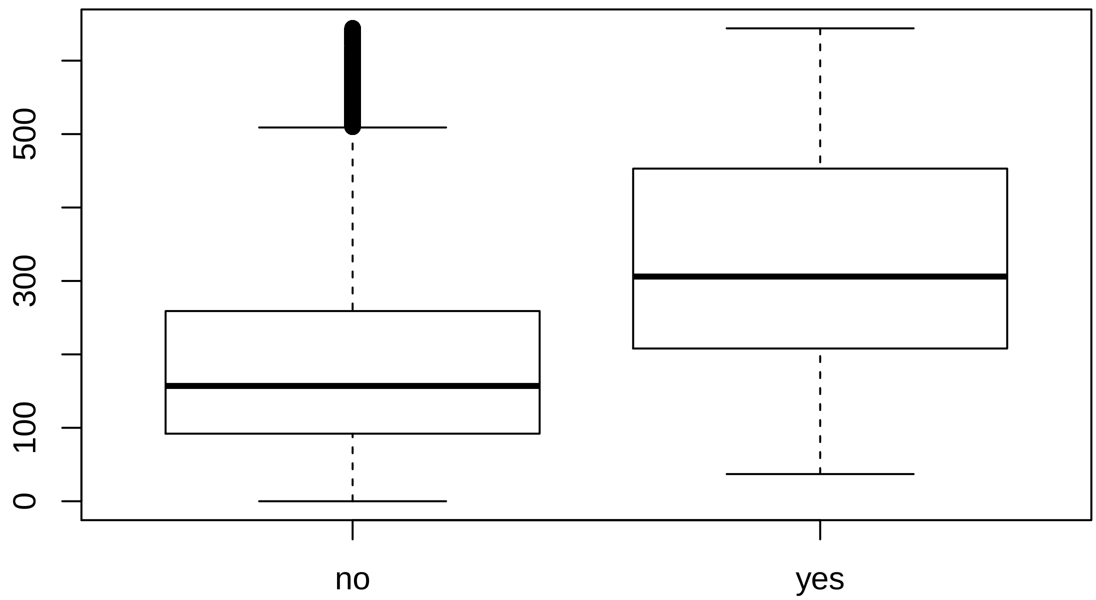
```
##   0%  25%  50%  75% 100%
##    0  102  180  319 4918
```

```
bench <- 319 + (1.5 * IQR(data2a$duration))
bench
```

```
## [1] 644.5
```

Now, we can treat outliers by replacing the values of "duration" feature in the dataset which are greater than the benchmark with "N/A" and then remove the entire rows that has "N/A".

```
treat <- data2a$duration > bench
data2a$duration[treat] <- NA
data2a <- na.omit(data2a)
boxplot(data2a$duration ~ data2a$y)
```




In the above boxplot we can see that significant amount of outliers are removed.

Now, let us create Data Quality Report (DQR) each for data2(with outliers) and data2a(without outliers).

## Data Quality Report:

**Declaration of function to generate Numeric Data Quality Report:**

```
library(ISLR)
dataQualityNum <- function(df) {
#Filteration of numeic values in the dataset
n <- sapply(df, function(x) {is.numeric(x)})
df_num <- df[, n]
# Number of numeric rows
instances <- sapply(df_num, FUN=function(x) {length(x)})
# Number of missing values (It must be zero for all numeric features as we are generating DQR for trans
missing <- sapply(df_num, FUN=function(x) {sum(is.na(x))})
missing <- missing / instances * 100
# Length of the vector of unique values
unique <- sapply(df_num, FUN=function(x) {length(unique(x))})
# Calculation of the quantiles
quantiles <- t(sapply(df_num, FUN=function(x) {quantile(x)}))
# Calculation of the mean
means <- sapply(df_num, FUN=function(x) {mean(x)})
# Calculation of the standard deviation
sds <- sapply(df_num, FUN=function(x) {sd(x)})
# Build a dataframe of all components of the DQR
```

```
df_frame <- data.frame(Feature=names(df_num),
Instances=instances,
Missing=missing,
Cardinality=unique,
Min=quantiles[,1],
Q1=quantiles[,2],
Feature=names(df_num),
Median=quantiles[,3],
Q3=quantiles[,4],
Max=quantiles[,5],
Mean=means,
Stdev=sds)
#To fit the table on the page, the abovecolumns were slightly renamed.
# Removal of rownames -- as they have no meaning here
rownames(df_frame) <- NULL
return(df_frame)
}
```

**Declaring a function to generate Categorical Data Quality Report:**

```
dataQualityCat <- function(df) {
# Filteration of categorical data from dataset 2 without outliers
n <- sapply(df, function(x) {is.numeric(x)})
df_categoricals <- df[, !n]
# Number of categorical rows in each feature
instances <- sapply(df_categoricals, FUN=function(x) {length(x)})
# Number of missing values (It must be zero for all numeric features as we are generating DQR for trans
missing <- sapply(df_categoricals, FUN=function(x) {sum(is.na(x))})
missing <- missing / instances * 100
# Length of the vector of unique values
unique <- sapply(df_categoricals, FUN=function(x) {length(unique(x))})
# Finding the most frequent categorical level
modeFreqs <- sapply(df_categoricals, FUN=function(x) {
t <- table(x)
modeFreq <- max(t)
return(modeFreq)
})
# For all modes, get their frequency
modes <- sapply(df_categoricals, FUN=function(x) {
t <- table(x)
modeFreq <- max(t)
mode <- names(t)[t==modeFreq]
return(mode)
})
# Now throw away the mode and repeat for the second mode
modeFreqs2 <- sapply(df_categoricals, FUN=function(x) {
t <- table(x)
modeFreq <- max(t)
mode <- names(t)[t==modeFreq]
# we remove the 1st mode here
x <- x[x != mode]
t <- table(x)
mode2Freq <- max(t)
```

```r
return(mode2Freq)
})
modes2 <- sapply(df_categoricals, FUN=function(x) {
t <- table(x)
modeFreq <- max(t)
mode <- names(t)[t==modeFreq]
# we remove the 1st mode here
x <- x[x != mode]
t <- table(x)
mode2Freq <- max(t)
mode2 <- names(t)[t==mode2Freq]
return(mode2)
})
# Build data.frame as before, but also derive the mode frequenies
df_categorical <- data.frame(Feature=names(df_categoricals),
Inst=instances,
Miss=missing,
Card=unique,
FstMod=modes,
FstModFrq=modeFreqs,
Feature=names(df_categoricals),
FstModPnt=modeFreqs/instances*100,
SndMod=modes2,
SndModFrq=modeFreqs2,
SndModPnt=modeFreqs2/instances*100)
#To fit the table on the page, the above columns were slightly renamed.
rownames(df_categorical) <- NULL
return(df_categorical)
}
```

**Data Quality report for numeric data of Dataset 1(without assuming missing value):**

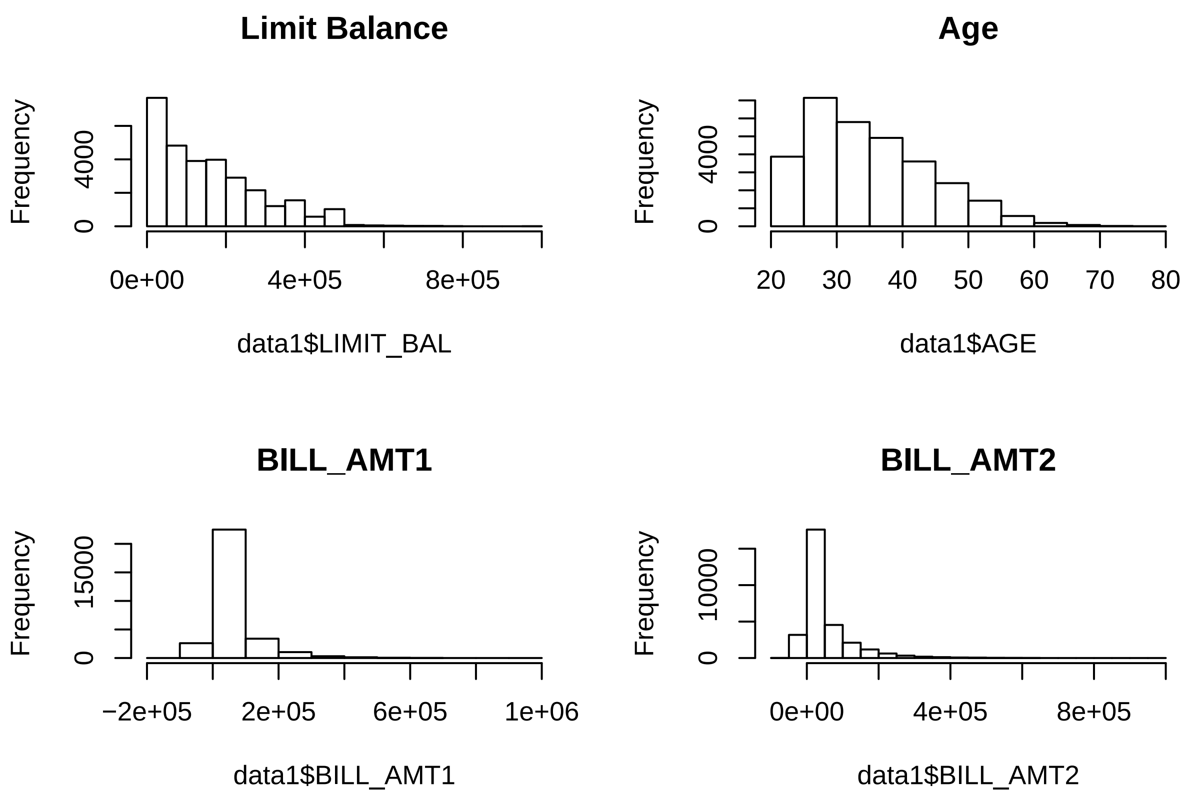
```r
# Calling a function to create a DQR for Dataset 1 without assuming missing value:
df1_dqrnum <- dataQualityNum(data1)
library(pander)
pandoc.table(df1_dqrnum, style = "grid", caption = "Numeric DQR for datset 1")
```

```
##
##
## +-----------+-----------+---------+-------------+---------+-------+
## |  Feature  | Instances | Missing | Cardinality |   Min   |  Q1   |
## +===========+===========+=========+=============+=========+=======+
## | LIMIT_BAL |   30000   |    0    |     81      |  10000  | 50000 |
## +-----------+-----------+---------+-------------+---------+-------+
## |    AGE    |   30000   |    0    |     56      |   21    |  28   |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT1 |   30000   |    0    |    22723    | -165580 | 3559  |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT2 |   30000   |    0    |    22346    | -69777  | 2985  |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT3 |   30000   |    0    |    22026    | -157264 | 2666  |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT4 |   30000   |    0    |    21548    | -170000 | 2327  |
```

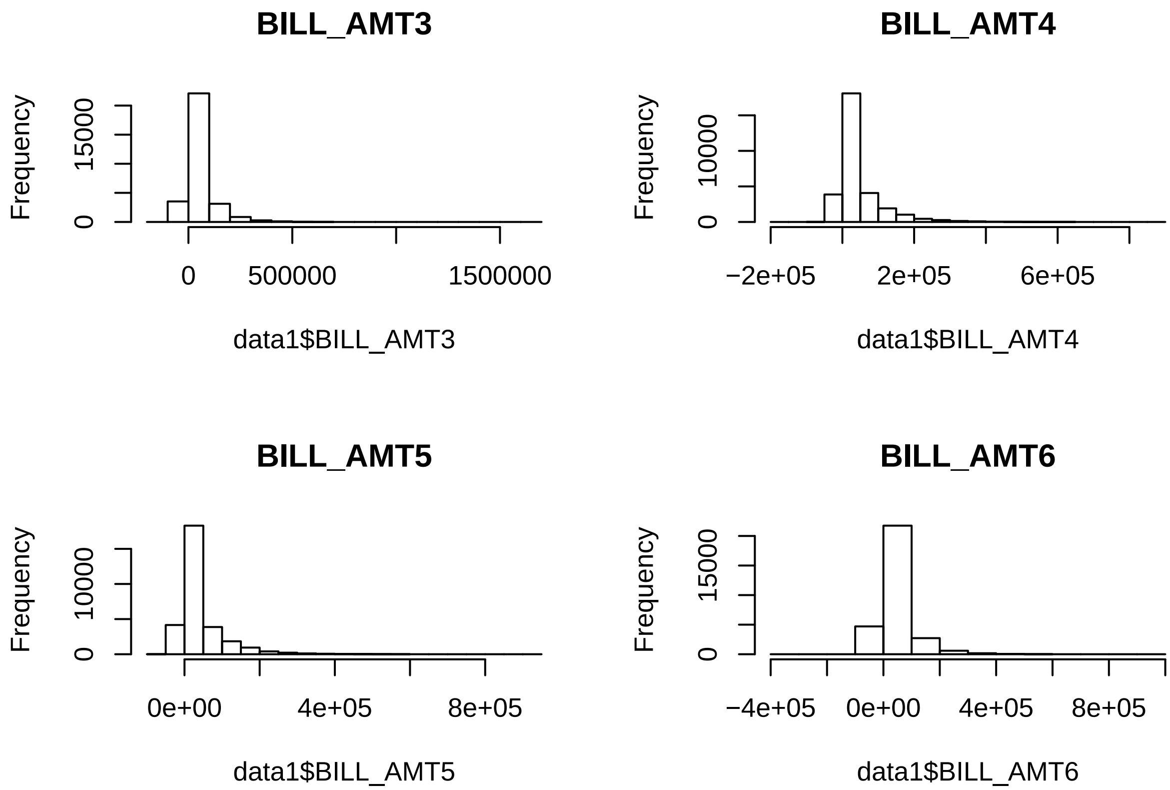
```
## +-----------+----------+---------+-------------+--------+-------+
## | BILL_AMT5 |  30000   |    0    |    21010    | -81334 | 1763  |
## +-----------+----------+---------+-------------+--------+-------+
## | BILL_AMT6 |  30000   |    0    |    20604    | -339603 | 1256  |
## +-----------+----------+---------+-------------+--------+-------+
## | PAY_AMT1  |  30000   |    0    |    7943     |   0    | 1000  |
## +-----------+----------+---------+-------------+--------+-------+
## | PAY_AMT2  |  30000   |    0    |    7899     |   0    |  833  |
## +-----------+----------+---------+-------------+--------+-------+
## | PAY_AMT3  |  30000   |    0    |    7518     |   0    |  390  |
## +-----------+----------+---------+-------------+--------+-------+
## | PAY_AMT4  |  30000   |    0    |    6937     |   0    |  296  |
## +-----------+----------+---------+-------------+--------+-------+
## | PAY_AMT5  |  30000   |    0    |    6897     |   0    | 252.5 |
## +-----------+----------+---------+-------------+--------+-------+
## | PAY_AMT6  |  30000   |    0    |    6939     |   0    | 117.8 |
## +-----------+----------+---------+-------------+--------+-------+
##
## Table: Numeric DQR for datset 1 (continued below)
##
##
##
## +-----------+--------+--------+---------+--------+--------+
## | Feature.1 | Median |   Q3   |   Max   |  Mean  | Stdev  |
## +===========+========+========+=========+========+========+
## | LIMIT_BAL | 140000 | 240000 |  1e+06  | 167484 | 129748 |
## +-----------+--------+--------+---------+--------+--------+
## |    AGE    |   34   |   41   |   79    | 35.49  | 9.218  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT1 | 22382  | 67091  | 964511  | 51223  | 73636  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT2 | 21200  | 64006  | 983931  | 49179  | 71174  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT3 | 20089  | 60165  | 1664089 | 47013  | 69349  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT4 | 19052  | 54506  | 891586  | 43263  | 64333  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT5 | 18105  | 50191  | 927171  | 40311  | 60797  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT6 | 17071  | 49198  | 961664  | 38872  | 59554  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT1  |  2100  |  5006  | 873552  |  5664  | 16563  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT2  |  2009  |  5000  | 1684259 |  5921  | 23041  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT3  |  1800  |  4505  | 896040  |  5226  | 17607  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT4  |  1500  |  4013  | 621000  |  4826  | 15666  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT5  |  1500  |  4032  | 426529  |  4799  | 15278  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT6  |  1500  |  4000  | 528666  |  5216  | 17777  |
## +-----------+--------+--------+---------+--------+--------+
```

**Plotting the features' distribution of Numeric data of dataset 1 without assumed missing value:**

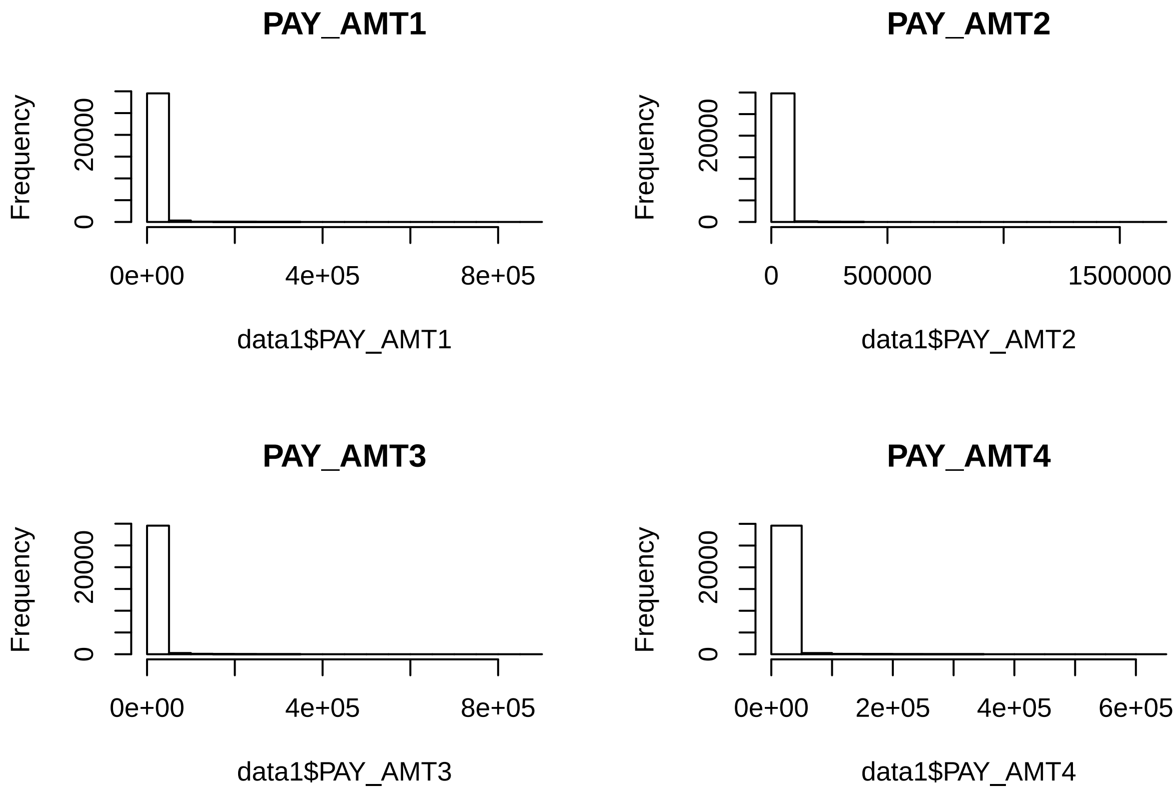
```
par(mfrow = c(2,2))
hist(data1$LIMIT_BAL, main="Limit Balance")
hist(data1$AGE, main="Age")
hist(data1$BILL_AMT1, main="BILL_AMT1")
hist(data1$BILL_AMT2, main="BILL_AMT2")
```



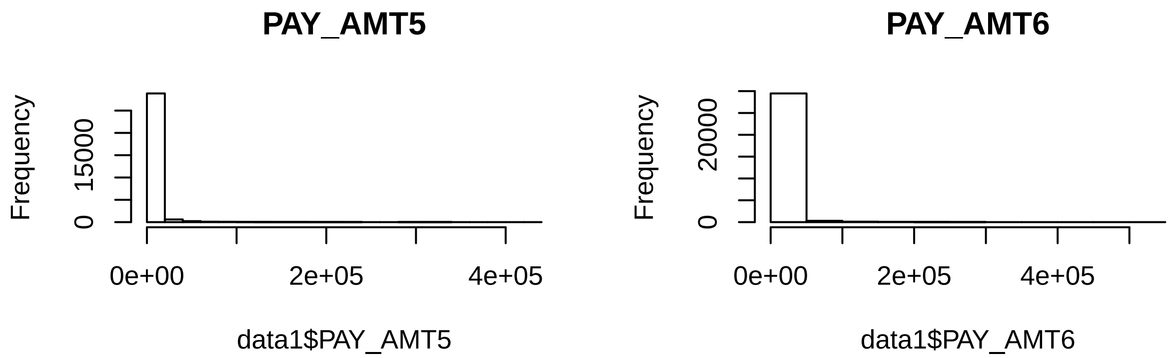

```
hist(data1$BILL_AMT3, main="BILL_AMT3")
hist(data1$BILL_AMT4, main="BILL_AMT4")
hist(data1$BILL_AMT5, main="BILL_AMT5")
hist(data1$BILL_AMT6, main="BILL_AMT6")
```




```r
hist(data1$PAY_AMT1, main="PAY_AMT1")
hist(data1$PAY_AMT2, main="PAY_AMT2")
hist(data1$PAY_AMT3, main="PAY_AMT3")
hist(data1$PAY_AMT4, main="PAY_AMT4")
```




```r
hist(data1$PAY_AMT5, main="PAY_AMT5")
hist(data1$PAY_AMT6, main="PAY_AMT6")
```




Most of the numeric features are not normally distributed,

Alomst all of the numeric features of dataset 01 without assumed missing value are right skewed.

**Data Quality report for numeric data of Dataset 1(with treated assumed missing values and without outliers):**

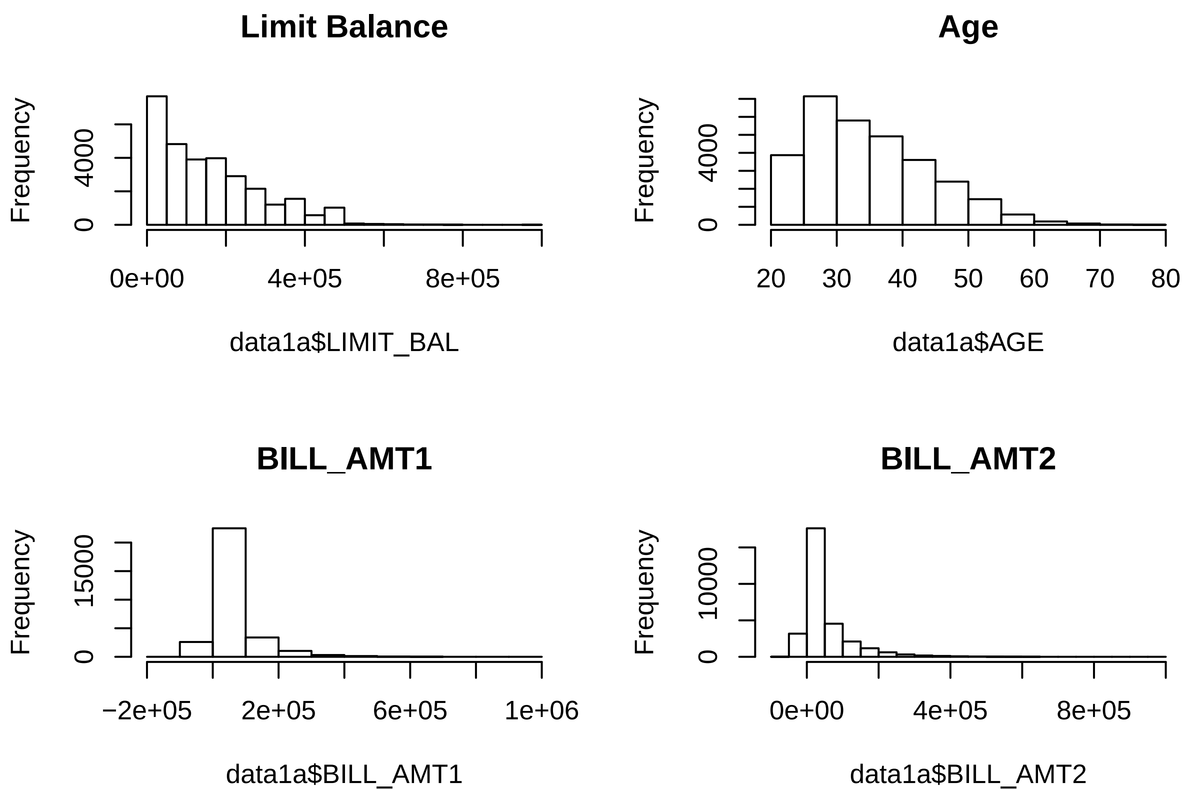
```
# Calling a function to create a DQR for Dataset 1 with treated assumed missing values and without outl
df1a_dqrnum <- dataQualityNum(data1a)
library(pander)
pandoc.table(df1a_dqrnum, style = "grid", caption = "Numeric DQR for datset 1a")
```

```
##
##
## +-----------+-----------+---------+-------------+---------+-------+
## |  Feature  | Instances | Missing | Cardinality |   Min   |  Q1   |
## +===========+===========+=========+=============+=========+=======+
## | LIMIT_BAL |   30000   |    0    |     81      |  10000  | 50000 |
## +-----------+-----------+---------+-------------+---------+-------+
## |    AGE    |   30000   |    0    |     56      |   21    |  28   |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT1 |   30000   |    0    |    22723    | -165580 | 3559  |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT2 |   30000   |    0    |    22346    | -69777  | 2985  |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT3 |   30000   |    0    |    22026    | -157264 | 2666  |
## +-----------+-----------+---------+-------------+---------+-------+
```

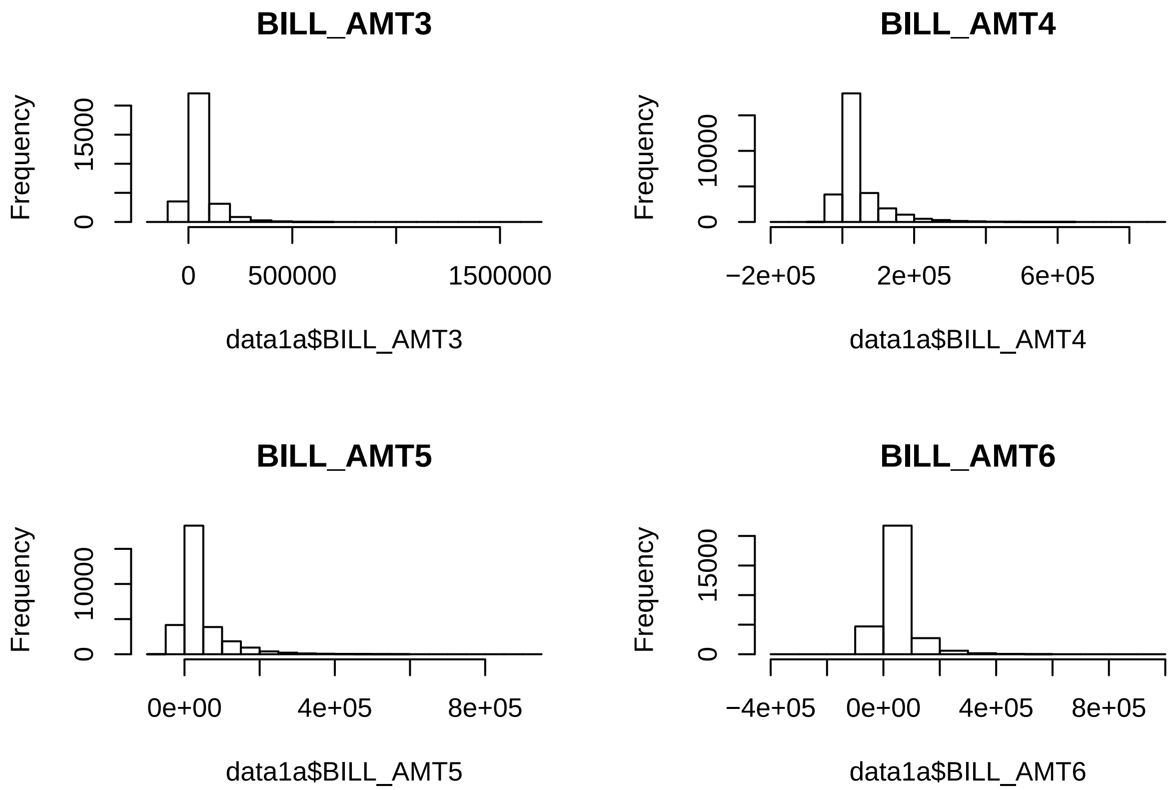
```
## | BILL_AMT4 |   30000   |    0    |     21548     | -170000 | 2327  |
## +-----------+-----------+---------+---------------+---------+-------+
## | BILL_AMT5 |   30000   |    0    |     21010     | -81334  | 1763  |
## +-----------+-----------+---------+---------------+---------+-------+
## | BILL_AMT6 |   30000   |    0    |     20604     | -339603 | 1256  |
## +-----------+-----------+---------+---------------+---------+-------+
## | PAY_AMT1  |   30000   |    0    |     7943      |    0    | 1000  |
## +-----------+-----------+---------+---------------+---------+-------+
## | PAY_AMT2  |   30000   |    0    |     7899      |    0    |  833  |
## +-----------+-----------+---------+---------------+---------+-------+
## | PAY_AMT3  |   30000   |    0    |     7518      |    0    |  390  |
## +-----------+-----------+---------+---------------+---------+-------+
## | PAY_AMT4  |   30000   |    0    |     6937      |    0    |  296  |
## +-----------+-----------+---------+---------------+---------+-------+
## | PAY_AMT5  |   30000   |    0    |     6897      |    0    | 252.5 |
## +-----------+-----------+---------+---------------+---------+-------+
## | PAY_AMT6  |   30000   |    0    |     6939      |    0    | 117.8 |
## +-----------+-----------+---------+---------------+---------+-------+
##
## Table: Numeric DQR for datset 1a (continued below)
##
##
##
## +-----------+--------+--------+---------+--------+--------+
## | Feature.1 | Median |   Q3   |   Max   |  Mean  | Stdev  |
## +===========+========+========+=========+========+========+
## | LIMIT_BAL | 140000 | 240000 |  1e+06  | 167484 | 129748 |
## +-----------+--------+--------+---------+--------+--------+
## |    AGE    |   34   |   41   |   79    | 35.49  | 9.218  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT1 | 22382  | 67091  | 964511  | 51223  | 73636  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT2 | 21200  | 64006  | 983931  | 49179  | 71174  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT3 | 20089  | 60165  | 1664089 | 47013  | 69349  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT4 | 19052  | 54506  | 891586  | 43263  | 64333  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT5 | 18105  | 50191  | 927171  | 40311  | 60797  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT6 | 17071  | 49198  | 961664  | 38872  | 59554  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT1  |  2100  |  5006  | 873552  |  5664  | 16563  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT2  |  2009  |  5000  | 1684259 |  5921  | 23041  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT3  |  1800  |  4505  | 896040  |  5226  | 17607  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT4  |  1500  |  4013  | 621000  |  4826  | 15666  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT5  |  1500  |  4032  | 426529  |  4799  | 15278  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT6  |  1500  |  4000  | 528666  |  5216  | 17777  |
## +-----------+--------+--------+---------+--------+--------+
```

**Plotting the features' distribution of Numeric data of dataset 1 with treated assumed missing values and without outliers:**

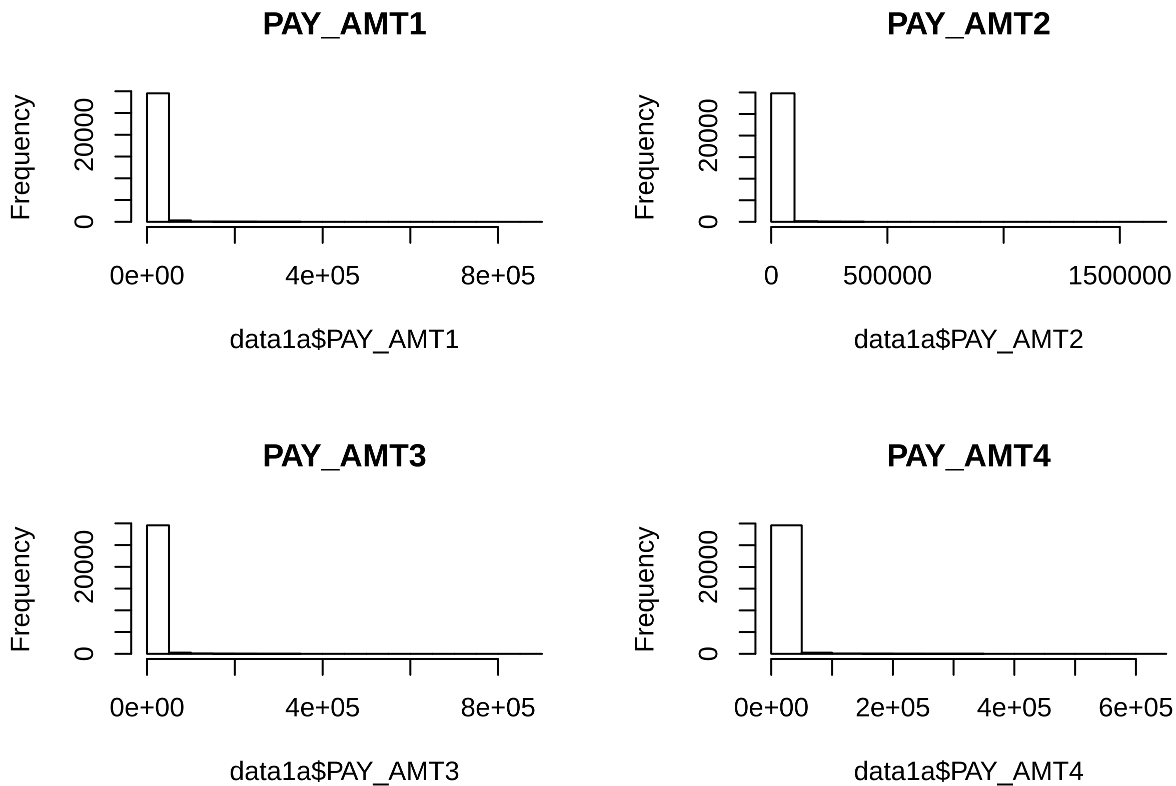
```r
par(mfrow = c(2,2))
hist(data1a$LIMIT_BAL, main="Limit Balance")
hist(data1a$AGE, main="Age")
hist(data1a$BILL_AMT1, main="BILL_AMT1")
hist(data1a$BILL_AMT2, main="BILL_AMT2")
```



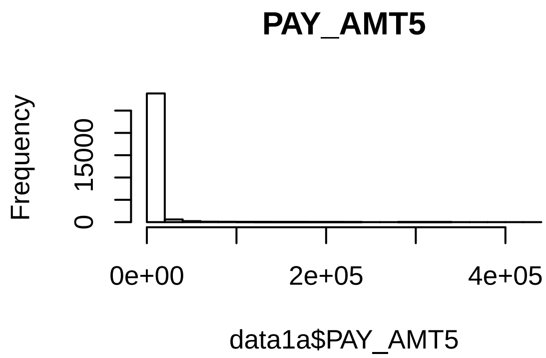

```r
hist(data1a$BILL_AMT3, main="BILL_AMT3")
hist(data1a$BILL_AMT4, main="BILL_AMT4")
hist(data1a$BILL_AMT5, main="BILL_AMT5")
hist(data1a$BILL_AMT6, main="BILL_AMT6")
```

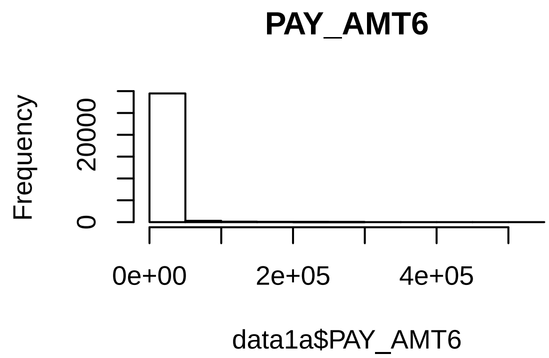



```r
hist(data1a$PAY_AMT1, main="PAY_AMT1")
hist(data1a$PAY_AMT2, main="PAY_AMT2")
hist(data1a$PAY_AMT3, main="PAY_AMT3")
hist(data1a$PAY_AMT4, main="PAY_AMT4")
```




```r
hist(data1a$PAY_AMT5, main="PAY_AMT5")
hist(data1a$PAY_AMT6, main="PAY_AMT6")
```




Even after transforming data with assumed missing value the data looks same as before but outliers(least frequent value) were removed.

**Data Quality report for numeric data of Dataset 1(with treated assumed missing values and outliers):**

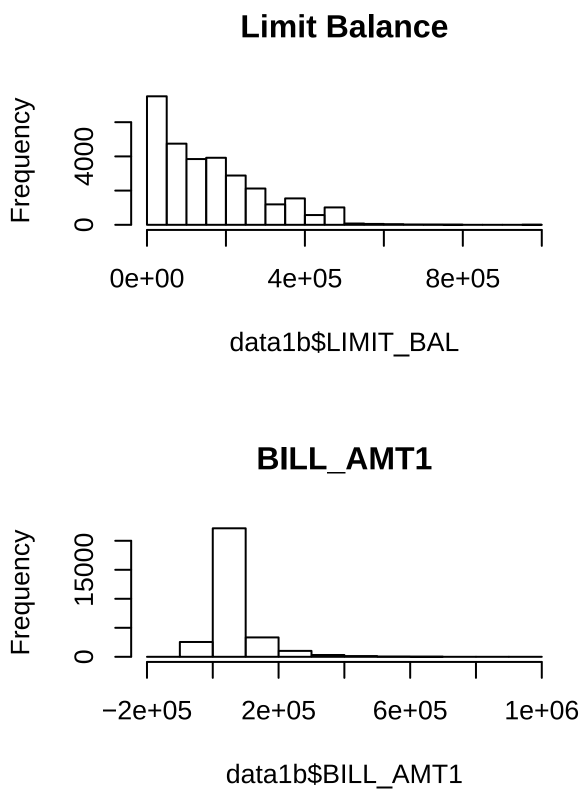
```r
# Calling a function to create a DQR for Dataset 1 with treated assumed missing values and outliers:
df1b_dqrnum <- dataQualityNum(data1a)
library(pander)
pandoc.table(df1b_dqrnum, style = "grid", caption = "Numeric DQR for datset 1b")
```

```
##
##
## +-----------+-----------+---------+-------------+---------+-------+
## |  Feature  | Instances | Missing | Cardinality |   Min   |  Q1   |
## +===========+===========+=========+=============+=========+=======+
## | LIMIT_BAL |   30000   |    0    |     81      |  10000  | 50000 |
## +-----------+-----------+---------+-------------+---------+-------+
## |    AGE    |   30000   |    0    |     56      |   21    |  28   |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT1 |   30000   |    0    |    22723    | -165580 | 3559  |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT2 |   30000   |    0    |    22346    | -69777  | 2985  |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT3 |   30000   |    0    |    22026    | -157264 | 2666  |
## +-----------+-----------+---------+-------------+---------+-------+
```

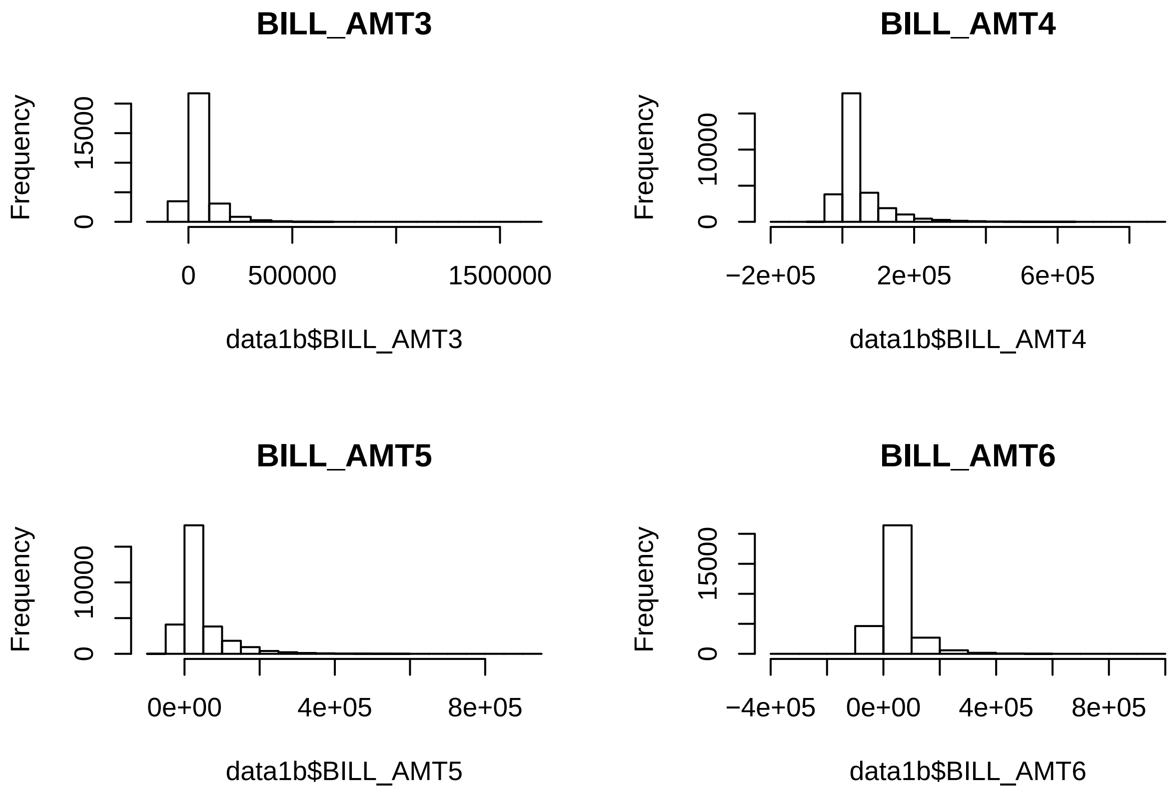
```
## | BILL_AMT4 |   30000   |    0    |    21548    | -170000 | 2327  |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT5 |   30000   |    0    |    21010    | -81334  | 1763  |
## +-----------+-----------+---------+-------------+---------+-------+
## | BILL_AMT6 |   30000   |    0    |    20604    | -339603 | 1256  |
## +-----------+-----------+---------+-------------+---------+-------+
## | PAY_AMT1  |   30000   |    0    |    7943     |    0    | 1000  |
## +-----------+-----------+---------+-------------+---------+-------+
## | PAY_AMT2  |   30000   |    0    |    7899     |    0    |  833  |
## +-----------+-----------+---------+-------------+---------+-------+
## | PAY_AMT3  |   30000   |    0    |    7518     |    0    |  390  |
## +-----------+-----------+---------+-------------+---------+-------+
## | PAY_AMT4  |   30000   |    0    |    6937     |    0    |  296  |
## +-----------+-----------+---------+-------------+---------+-------+
## | PAY_AMT5  |   30000   |    0    |    6897     |    0    | 252.5 |
## +-----------+-----------+---------+-------------+---------+-------+
## | PAY_AMT6  |   30000   |    0    |    6939     |    0    | 117.8 |
## +-----------+-----------+---------+-------------+---------+-------+
##
## Table: Numeric DQR for datset 1b (continued below)
##
##
##
## +-----------+--------+--------+---------+--------+--------+
## | Feature.1 | Median |   Q3   |   Max   |  Mean  | Stdev  |
## +===========+========+========+=========+========+========+
## | LIMIT_BAL | 140000 | 240000 |  1e+06  | 167484 | 129748 |
## +-----------+--------+--------+---------+--------+--------+
## |    AGE    |   34   |   41   |   79    | 35.49  | 9.218  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT1 | 22382  | 67091  | 964511  | 51223  | 73636  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT2 | 21200  | 64006  | 983931  | 49179  | 71174  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT3 | 20089  | 60165  | 1664089 | 47013  | 69349  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT4 | 19052  | 54506  | 891586  | 43263  | 64333  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT5 | 18105  | 50191  | 927171  | 40311  | 60797  |
## +-----------+--------+--------+---------+--------+--------+
## | BILL_AMT6 | 17071  | 49198  | 961664  | 38872  | 59554  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT1  |  2100  |  5006  | 873552  |  5664  | 16563  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT2  |  2009  |  5000  | 1684259 |  5921  | 23041  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT3  |  1800  |  4505  | 896040  |  5226  | 17607  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT4  |  1500  |  4013  | 621000  |  4826  | 15666  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT5  |  1500  |  4032  | 426529  |  4799  | 15278  |
## +-----------+--------+--------+---------+--------+--------+
## | PAY_AMT6  |  1500  |  4000  | 528666  |  5216  | 17777  |
## +-----------+--------+--------+---------+--------+--------+
```

**Plotting the feature's distribution of Numeric data of dataset 1 with treated assumed missing values and outliers:**

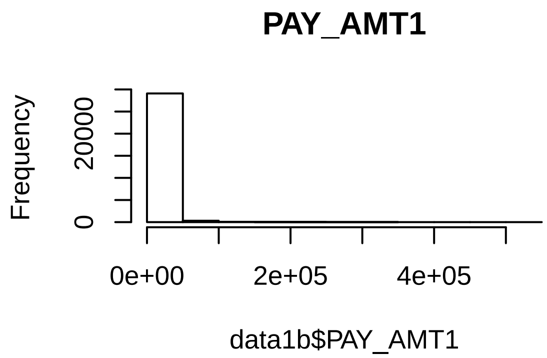
```
par(mfrow = c(2,2))
hist(data1b$LIMIT_BAL, main="Limit Balance")
hist(data1b$AGE, main="Age")
hist(data1b$BILL_AMT1, main="BILL_AMT1")
hist(data1b$BILL_AMT2, main="BILL_AMT2")
```

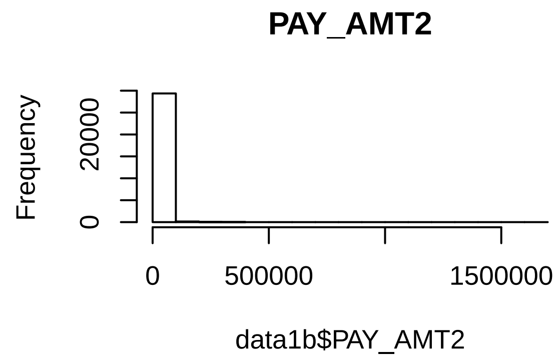


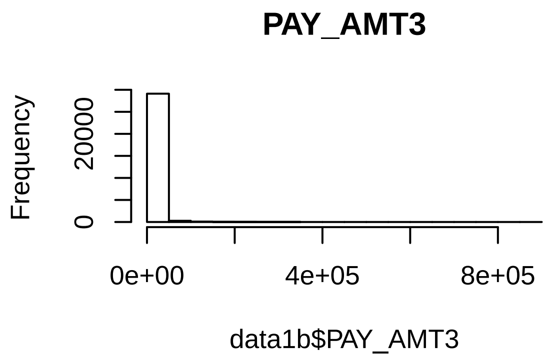

```
hist(data1b$BILL_AMT3, main="BILL_AMT3")
hist(data1b$BILL_AMT4, main="BILL_AMT4")
hist(data1b$BILL_AMT5, main="BILL_AMT5")
hist(data1b$BILL_AMT6, main="BILL_AMT6")
```

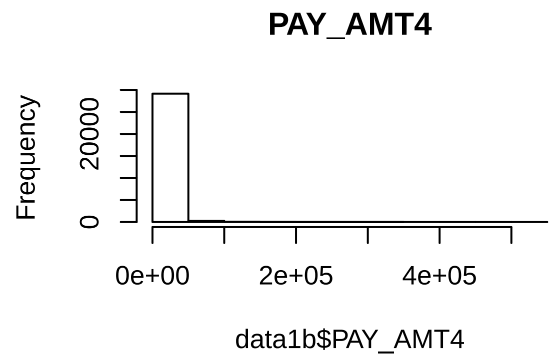



```r
hist(data1b$PAY_AMT1, main="PAY_AMT1")
hist(data1b$PAY_AMT2, main="PAY_AMT2")
hist(data1b$PAY_AMT3, main="PAY_AMT3")
hist(data1b$PAY_AMT4, main="PAY_AMT4")
```

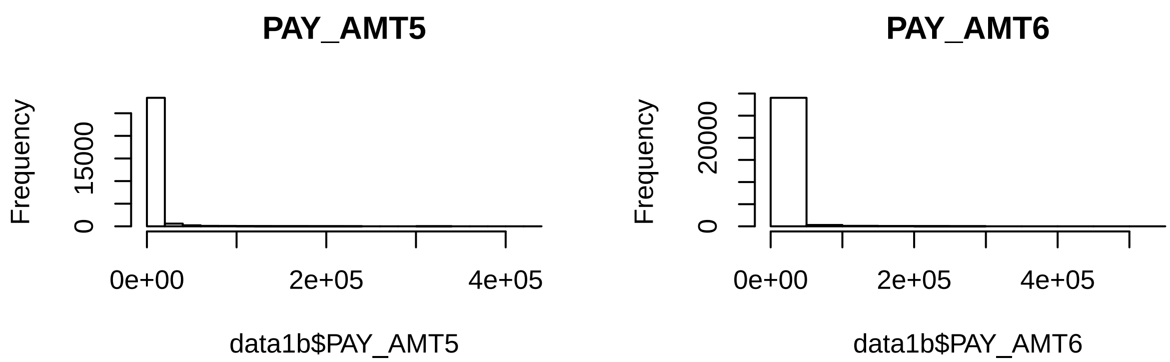



```r
hist(data1b$PAY_AMT5, main="PAY_AMT5")
hist(data1b$PAY_AMT6, main="PAY_AMT6")
```




Even after transforming data with assumed missing value the data looks same as before, all the numeric data looks right skewed.

**Data Quality Report for categorical data of dataset 1 without assumed missing value:**

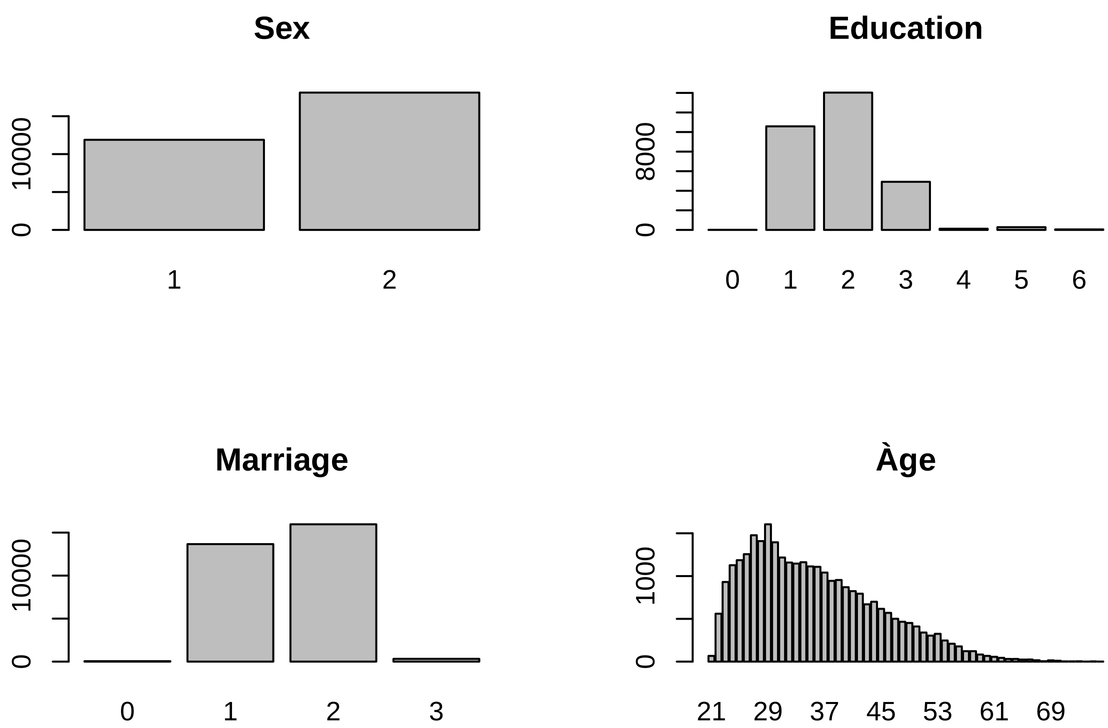
```r
# Calling a function to create a DQR for Dataset 1 without assumed missing value:
df1_categorical <- dataQualityCat(data1)
library(pander)
pandoc.table(df1_categorical, style = "grid", caption = "Categorical DQR for dataset 1")
```

```
##
##
## +-----------------------------+-------+------+------+--------+-----------+
## |           Feature           | Inst  | Miss | Card | FstMod | FstModFrq |
## +=============================+=======+======+======+========+===========+
## |             SEX             | 30000 |  0   |  2   |   2    |   18112   |
## +-----------------------------+-------+------+------+--------+-----------+
## |          EDUCATION          | 30000 |  0   |  7   |   2    |   14030   |
## +-----------------------------+-------+------+------+--------+-----------+
## |          MARRIAGE           | 30000 |  0   |  4   |   2    |   15964   |
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_0            | 30000 |  0   |  11  |   0    |   14737   |
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_2            | 30000 |  0   |  11  |   0    |   15730   |
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_3            | 30000 |  0   |  11  |   0    |   15764   |
```

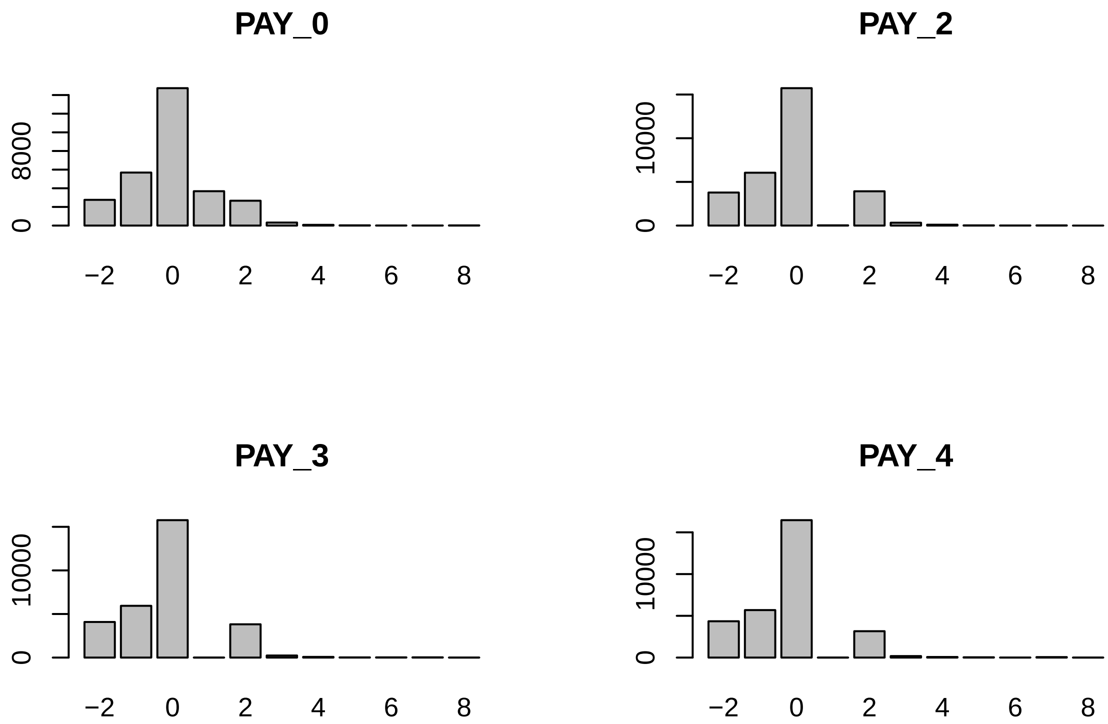
```
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_4            | 30000 |  0   |  11  |   0    |   16455   |
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_5            | 30000 |  0   |  10  |   0    |   16947   |
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_6            | 30000 |  0   |  10  |   0    |   16286   |
## +-----------------------------+-------+------+------+--------+-----------+
## | default.payment.next.month  | 30000 |  0   |  2   |   0    |   23364   |
## +-----------------------------+-------+------+------+--------+-----------+
##
## Table: Categorical DQR for dataset 1 (continued below)
##
##
##
## +-----------------------------+-----------+--------+-----------+-----------+
## |          Feature.1          | FstModPnt | SndMod | SndModFrq | SndModPnt |
## +=============================+===========+========+===========+===========+
## |             SEX             |   60.37   |   1    |   11888   |   39.63   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |          EDUCATION          |   46.77   |   1    |   10585   |   35.28   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |          MARRIAGE           |   53.21   |   1    |   13659   |   45.53   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_0            |   49.12   |   -1   |   5686    |   18.95   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_2            |   52.43   |   -1   |   6050    |   20.17   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_3            |   52.55   |   -1   |   5938    |   19.79   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_4            |   54.85   |   -1   |   5687    |   18.96   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_5            |   56.49   |   -1   |   5539    |   18.46   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_6            |   54.29   |   -1   |   5740    |   19.13   |
## +-----------------------------+-----------+--------+-----------+-----------+
## | default.payment.next.month  |   77.88   |   1    |   6636    |   22.12   |
## +-----------------------------+-----------+--------+-----------+-----------+
```

**Plotting the features' distribution of categorical data of dataset 1 without assuming missing value:**

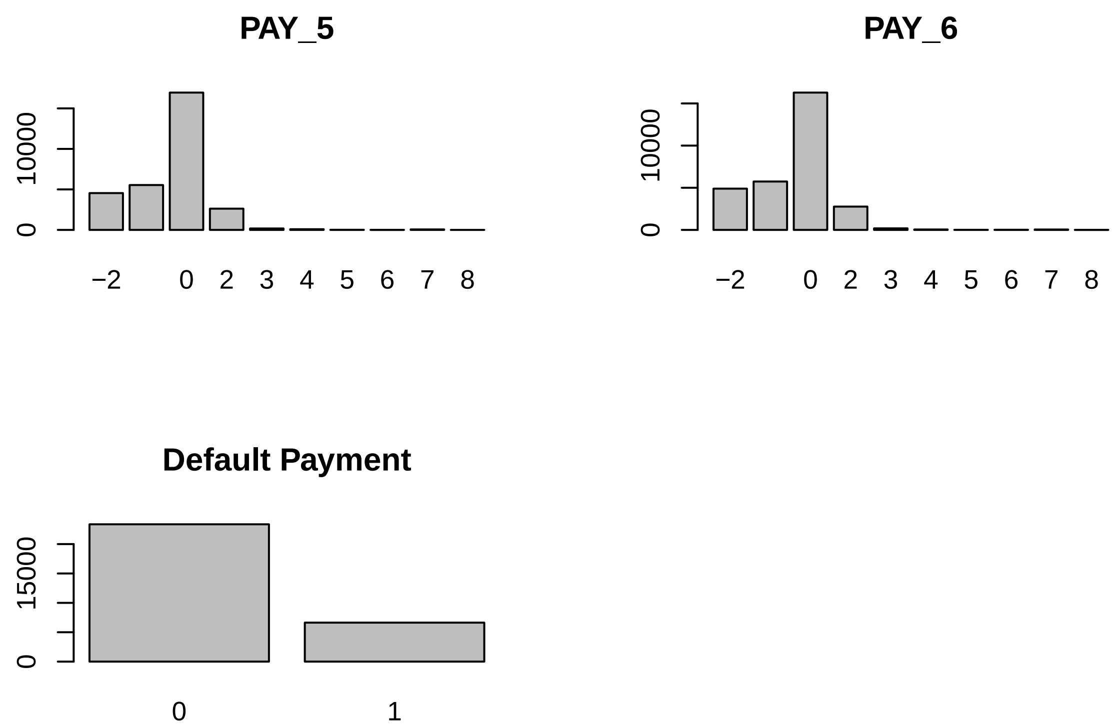
```r
par(mfrow = c(2,2))
barplot(table(data1$SEX), main="Sex")
barplot(table(data1$EDUCATION), main="Education")
barplot(table(data1$MARRIAGE), main="Marriage")
barplot(table(data1$AGE), main="Àge")
```




```
barplot(table(data1$PAY_0), main="PAY_0")
barplot(table(data1$PAY_2), main="PAY_2")
barplot(table(data1$PAY_3), main="PAY_3")
barplot(table(data1$PAY_4), main="PAY_4")
```




```r
barplot(table(data1$PAY_5), main="PAY_5")
barplot(table(data1$PAY_6), main="PAY_6")
barplot(table(data1$default.payment.next.month), main="Default Payment")
```




As numeric data, categorical data also not normally distributed and most of them are right skewed and the data are too irregular.

**Data Quality Report for categorical data of dataset 1 with treated assumed missing value and without outliers:**

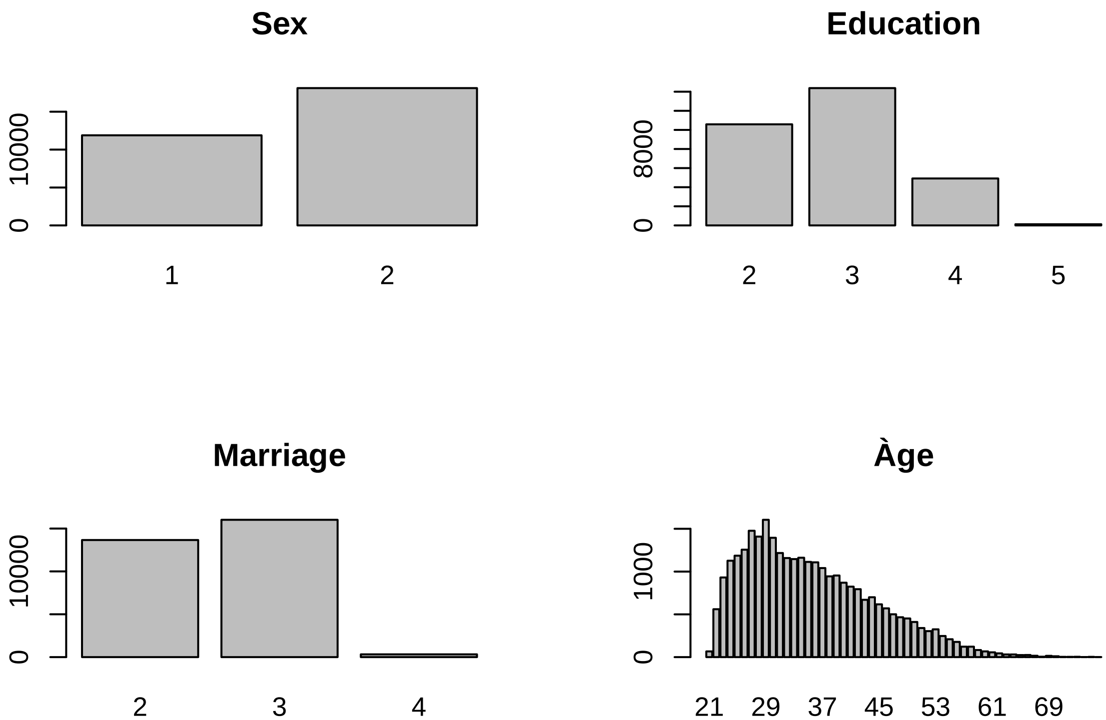
```r
# Calling a function to create a DQR for dataset 1 with treated assumed missing value and without outli
df1a_categorical <- dataQualityCat(data1a)
library(pander)
pandoc.table(df1a_categorical, style = "grid", caption = "categorical DQR for dataset 1a")
```

```
##
##
## +-----------------------------+-------+------+------+--------+-----------+
## |           Feature           | Inst  | Miss | Card | FstMod | FstModFrq |
## +=============================+=======+======+======+========+===========+
## |             SEX             | 30000 |  0   |  2   |   2    |   18112   |
## +-----------------------------+-------+------+------+--------+-----------+
## |          EDUCATION          | 30000 |  0   |  4   |   3    |   14375   |
## +-----------------------------+-------+------+------+--------+-----------+
## |          MARRIAGE           | 30000 |  0   |  3   |   3    |   16018   |
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_0            | 30000 |  0   |  11  |   0    |   14737   |
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_2            | 30000 |  0   |  11  |   0    |   15730   |
## +-----------------------------+-------+------+------+--------+-----------+
```

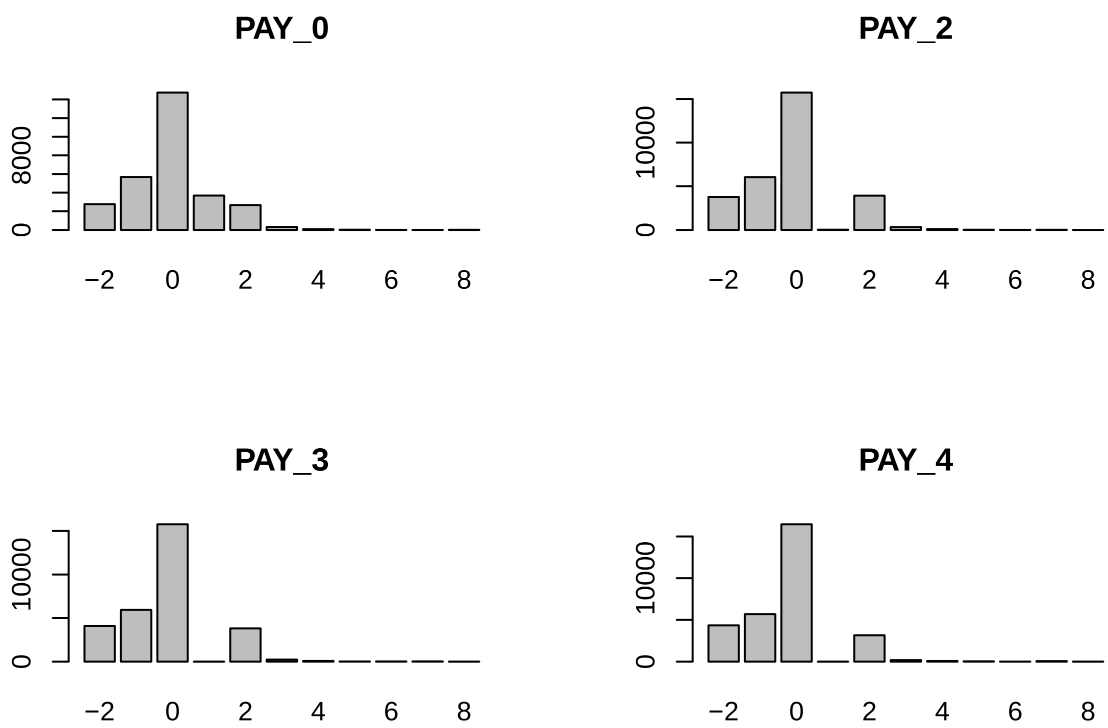
```
## |           PAY_3            | 30000 |  0   |  11  |    0    |   15764    |
## +----------------------------+-------+------+------+---------+------------+
## |           PAY_4            | 30000 |  0   |  11  |    0    |   16455    |
## +----------------------------+-------+------+------+---------+------------+
## |           PAY_5            | 30000 |  0   |  10  |    0    |   16947    |
## +----------------------------+-------+------+------+---------+------------+
## |           PAY_6            | 30000 |  0   |  10  |    0    |   16286    |
## +----------------------------+-------+------+------+---------+------------+
## | default.payment.next.month | 30000 |  0   |  2   |    0    |   23364    |
## +----------------------------+-------+------+------+---------+------------+
##
## Table: categorical DQR for dataset 1a (continued below)
##
##
##
## +----------------------------+-----------+--------+-----------+-----------+
## |         Feature.1          | FstModPnt | SndMod | SndModFrq | SndModPnt |
## +============================+===========+========+===========+===========+
## |            SEX             |   60.37   |   1    |   11888   |   39.63   |
## +----------------------------+-----------+--------+-----------+-----------+
## |         EDUCATION          |   47.92   |   2    |   10585   |   35.28   |
## +----------------------------+-----------+--------+-----------+-----------+
## |          MARRIAGE          |   53.39   |   2    |   13659   |   45.53   |
## +----------------------------+-----------+--------+-----------+-----------+
## |           PAY_0            |   49.12   |   -1   |   5686    |   18.95   |
## +----------------------------+-----------+--------+-----------+-----------+
## |           PAY_2            |   52.43   |   -1   |   6050    |   20.17   |
## +----------------------------+-----------+--------+-----------+-----------+
## |           PAY_3            |   52.55   |   -1   |   5938    |   19.79   |
## +----------------------------+-----------+--------+-----------+-----------+
## |           PAY_4            |   54.85   |   -1   |   5687    |   18.96   |
## +----------------------------+-----------+--------+-----------+-----------+
## |           PAY_5            |   56.49   |   -1   |   5539    |   18.46   |
## +----------------------------+-----------+--------+-----------+-----------+
## |           PAY_6            |   54.29   |   -1   |   5740    |   19.13   |
## +----------------------------+-----------+--------+-----------+-----------+
## | default.payment.next.month |   77.88   |   1    |   6636    |   22.12   |
## +----------------------------+-----------+--------+-----------+-----------+
```

**Plotting the features' distribution of categorical data of dataset 1 with treated assumed missing value and without outliers:**

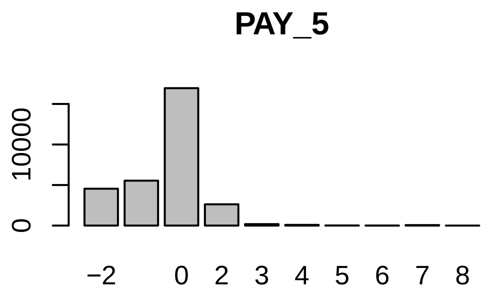
```r
par(mfrow = c(2,2))
barplot(table(data1a$SEX), main="Sex")
barplot(table(data1a$EDUCATION), main="Education")
barplot(table(data1a$MARRIAGE), main="Marriage")
barplot(table(data1a$AGE), main="Àge")
```

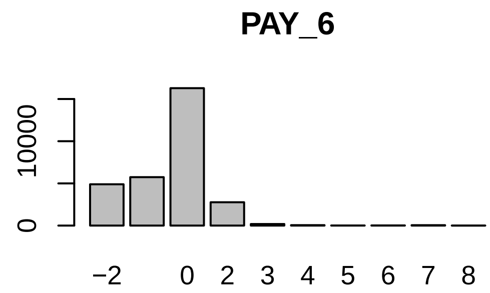


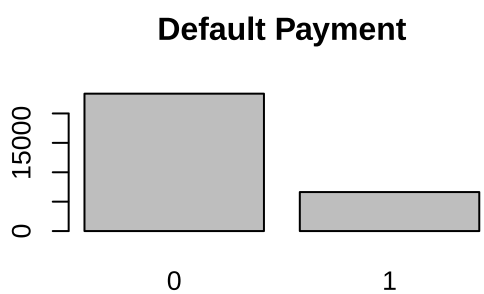

```r
barplot(table(data1a$PAY_0), main="PAY_0")
barplot(table(data1a$PAY_2), main="PAY_2")
barplot(table(data1a$PAY_3), main="PAY_3")
barplot(table(data1a$PAY_4), main="PAY_4")
```




```r
barplot(table(data1a$PAY_5), main="PAY_5")
barplot(table(data1a$PAY_6), main="PAY_6")
barplot(table(data1a$default.payment.next.month), main="Default Payment")
```




As numeric data, categorical data with assumed missing value and without outlier also not normally distributed and most of them are right skewed and the data are too irregular.

**Data Quality Report for categorical data of dataset 1 with treated assumed missing value and outliers:**

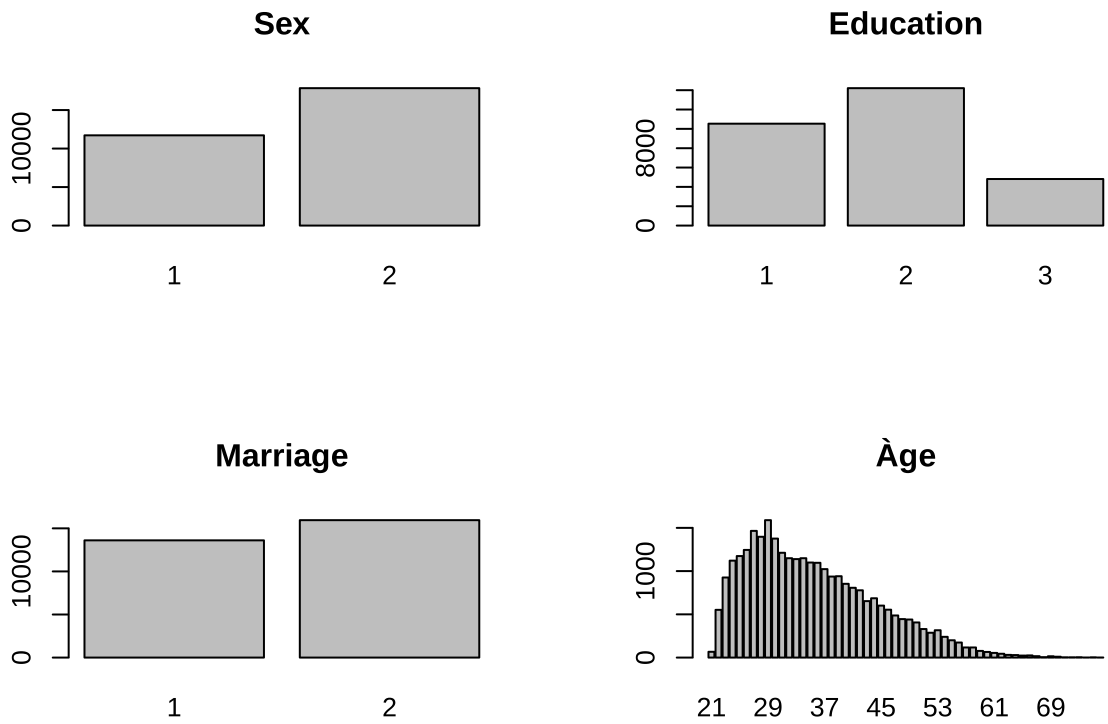
```
# Calling a function to create a DQR for dataset 1 with treated assumed missing value and outliers:
df1b_categorical <- dataQualityCat(data1b)
library(pander)
pandoc.table(df1b_categorical, style = "grid", caption = "categorical DQR for dataset 1b")
```

```
##
##
## +-----------------------------+-------+------+------+--------+-----------+
## |           Feature           | Inst  | Miss | Card | FstMod | FstModFrq |
## +=============================+=======+======+======+========+===========+
## |             SEX             | 29557 |  0   |  2   |   2    |   17841   |
## +-----------------------------+-------+------+------+--------+-----------+
## |          EDUCATION          | 29557 |  0   |  3   |   2    |   14208   |
## +-----------------------------+-------+------+------+--------+-----------+
## |          MARRIAGE           | 29557 |  0   |  2   |   2    |   15950   |
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_0            | 29557 |  0   |  11  |   0    |   14497   |
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_2            | 29557 |  0   |  11  |   0    |   15475   |
## +-----------------------------+-------+------+------+--------+-----------+
## |            PAY_3            | 29557 |  0   |  11  |   0    |   15510   |
```

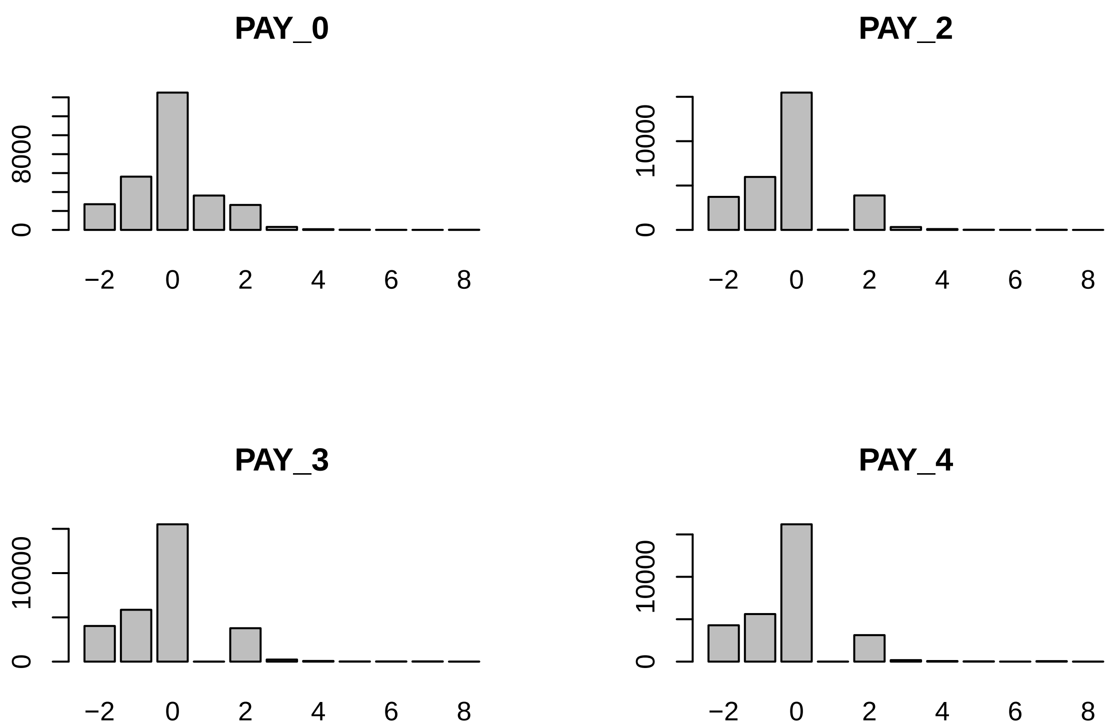
```
## +-----------------------------+-------+------+------+--------+-----------+
## |           PAY_4             | 29557 |  0   |  11  |   0    |   16193   |
## +-----------------------------+-------+------+------+--------+-----------+
## |           PAY_5             | 29557 |  0   |  10  |   0    |   16671   |
## +-----------------------------+-------+------+------+--------+-----------+
## |           PAY_6             | 29557 |  0   |  10  |   0    |   16024   |
## +-----------------------------+-------+------+------+--------+-----------+
## | default.payment.next.month  | 29557 |  0   |  2   |   0    |   23012   |
## +-----------------------------+-------+------+------+--------+-----------+
##
## Table: categorical DQR for dataset 1b (continued below)
##
##
##
## +-----------------------------+-----------+--------+-----------+-----------+
## |          Feature.1          | FstModPnt | SndMod | SndModFrq | SndModPnt |
## +=============================+===========+========+===========+===========+
## |             SEX             |   60.36   |   1    |   11716   |   39.64   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |          EDUCATION          |   48.07   |   1    |   10535   |   35.64   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |           MARRIAGE          |   53.96   |   1    |   13607   |   46.04   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_0            |   49.05   |   -1   |   5623    |   19.02   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_2            |   52.36   |   -1   |   5970    |   20.2    |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_3            |   52.47   |   -1   |   5855    |   19.81   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_4            |   54.79   |   -1   |   5609    |   18.98   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_5            |   56.4    |   -1   |   5460    |   18.47   |
## +-----------------------------+-----------+--------+-----------+-----------+
## |            PAY_6            |   54.21   |   -1   |   5657    |   19.14   |
## +-----------------------------+-----------+--------+-----------+-----------+
## | default.payment.next.month  |   77.86   |   1    |   6545    |   22.14   |
## +-----------------------------+-----------+--------+-----------+-----------+
```

**Plotting the features' distribution of categorical data of dataset 1 with treated assumed missing value and outliers:**

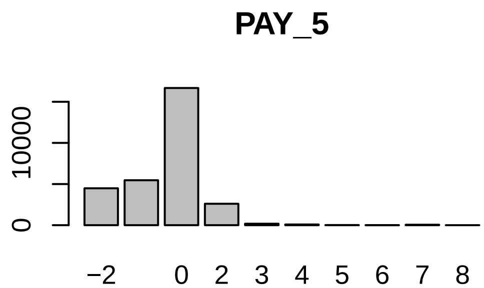
```r
par(mfrow = c(2,2))
barplot(table(data1b$SEX), main="Sex")
barplot(table(data1b$EDUCATION), main="Education")
barplot(table(data1b$MARRIAGE), main="Marriage")
barplot(table(data1b$AGE), main="Âge")
```

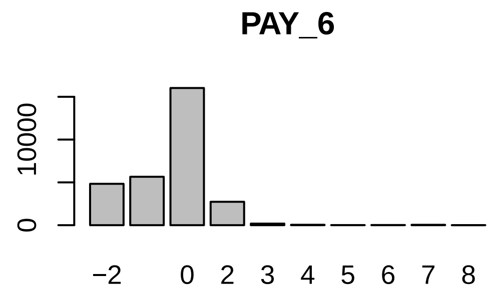


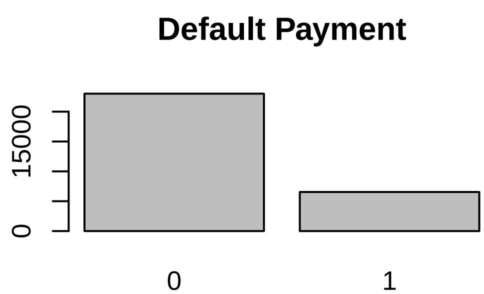

```r
barplot(table(data1b$PAY_0), main="PAY_0")
barplot(table(data1b$PAY_2), main="PAY_2")
barplot(table(data1b$PAY_3), main="PAY_3")
barplot(table(data1b$PAY_4), main="PAY_4")
```




```r
barplot(table(data1b$PAY_5), main="PAY_5")
barplot(table(data1b$PAY_6), main="PAY_6")
barplot(table(data1b$default.payment.next.month), main="Default Payment")
```




As numeric data, categorical data with assumed missing value and outlier also not normally distributed and most of them are right skewed and the data are too irregular.

**Data Quality report for numeric data of Dataset 2(with outliers):**

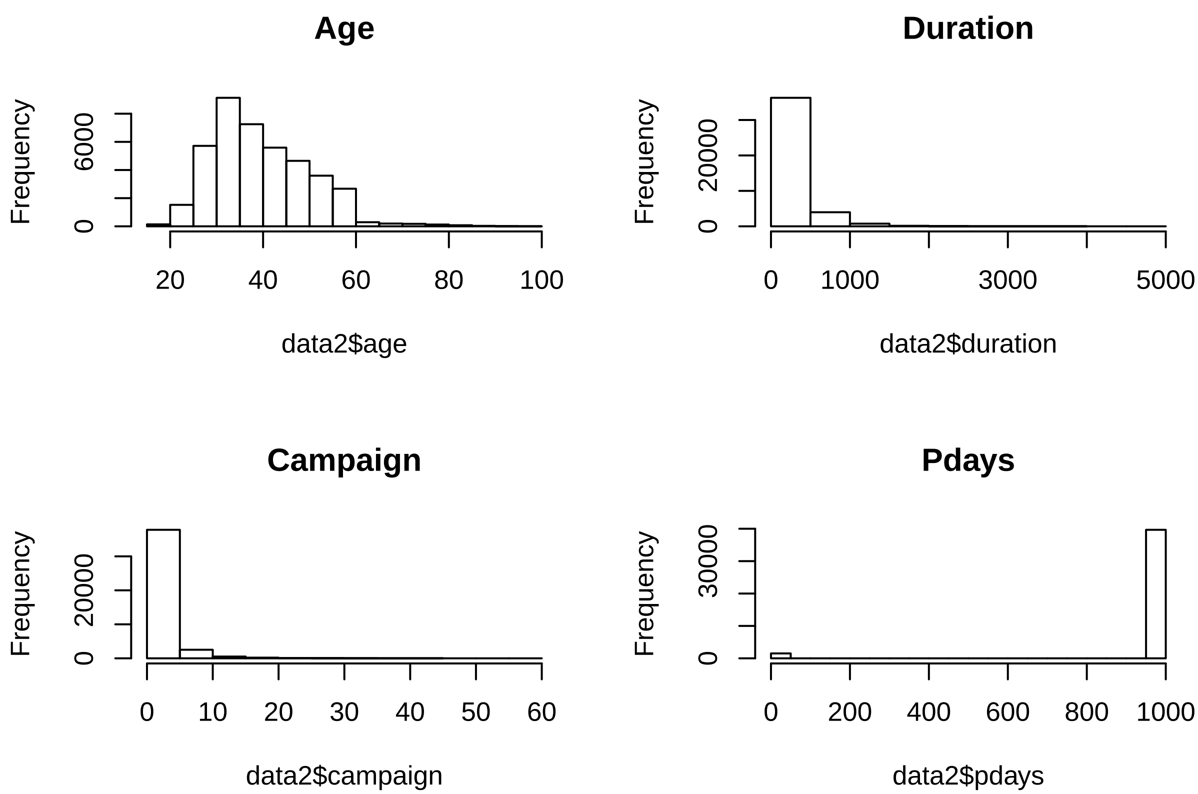
```r
# Calling a function to create a DQR for Dataset 2 with outlier:
df2_dqrnum <- dataQualityNum(data2)
library(pander)
pandoc.table(df2_dqrnum, style = "grid", caption = "Data Quality Report for numeric data of datset 2 wi
```

```
##
##
## +----------------+-----------+---------+-------------+-------+-------+
## |    Feature     | Instances | Missing | Cardinality |  Min  |  Q1   |
## +================+===========+=========+=============+=======+=======+
## |      age       |   41188   |    0    |     78      |  17   |  32   |
## +----------------+-----------+---------+-------------+-------+-------+
## |    duration    |   41188   |    0    |    1544     |   0   |  102  |
## +----------------+-----------+---------+-------------+-------+-------+
## |    campaign    |   41188   |    0    |     42      |   1   |   1   |
## +----------------+-----------+---------+-------------+-------+-------+
## |     pdays      |   41188   |    0    |     27      |   0   |  999  |
## +----------------+-----------+---------+-------------+-------+-------+
## |    previous    |   41188   |    0    |      8      |   0   |   0   |
## +----------------+-----------+---------+-------------+-------+-------+
## |  emp.var.rate  |   41188   |    0    |     10      | -3.4  | -1.8  |
```

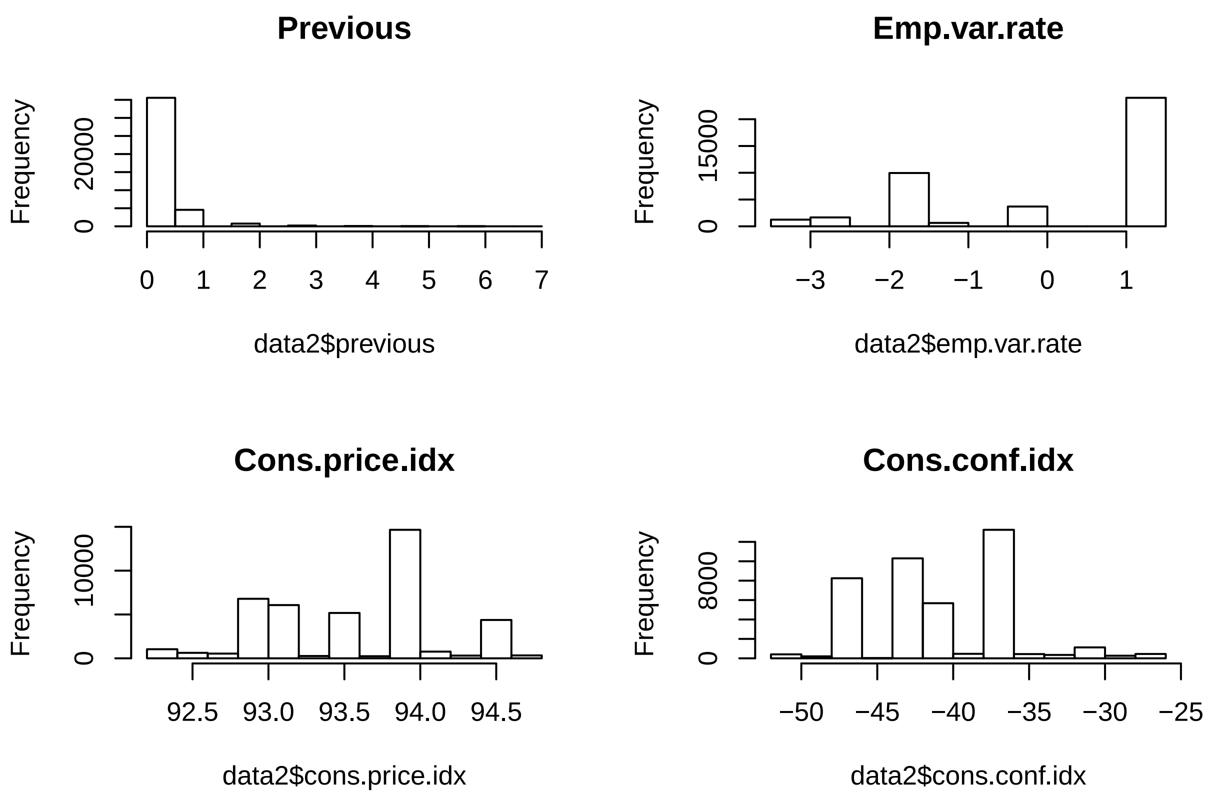
```
## +----------------+-----------+---------+-------------+-------+-------+
## | cons.price.idx |   41188   |    0    |     26      | 92.2  | 93.08 |
## +----------------+-----------+---------+-------------+-------+-------+
## | cons.conf.idx  |   41188   |    0    |     26      | -50.8 | -42.7 |
## +----------------+-----------+---------+-------------+-------+-------+
## |   euribor3m    |   41188   |    0    |     316     | 0.634 | 1.344 |
## +----------------+-----------+---------+-------------+-------+-------+
## |  nr.employed   |   41188   |    0    |     11      | 4964  | 5099  |
## +----------------+-----------+---------+-------------+-------+-------+
##
## Table: Data Quality Report for numeric data of datset 2 with outlier (continued below)
##
##
##
## +----------------+--------+-------+-------+---------+--------+
## |   Feature.1    | Median |  Q3   |  Max  |  Mean   | Stdev  |
## +================+========+=======+=======+=========+========+
## |      age       |   38   |  47   |  98   |  40.02  | 10.42  |
## +----------------+--------+-------+-------+---------+--------+
## |    duration    |  180   |  319  | 4918  |  258.3  | 259.3  |
## +----------------+--------+-------+-------+---------+--------+
## |    campaign    |   2    |   3   |  56   |  2.568  |  2.77  |
## +----------------+--------+-------+-------+---------+--------+
## |     pdays      |  999   |  999  |  999  |  962.5  | 186.9  |
## +----------------+--------+-------+-------+---------+--------+
## |    previous    |   0    |   0   |   7   |  0.173  | 0.4949 |
## +----------------+--------+-------+-------+---------+--------+
## |  emp.var.rate  |  1.1   |  1.4  |  1.4  | 0.08189 | 1.571  |
## +----------------+--------+-------+-------+---------+--------+
## | cons.price.idx | 93.75  | 93.99 | 94.77 |  93.58  | 0.5788 |
## +----------------+--------+-------+-------+---------+--------+
## | cons.conf.idx  | -41.8  | -36.4 | -26.9 |  -40.5  | 4.628  |
## +----------------+--------+-------+-------+---------+--------+
## |   euribor3m    | 4.857  | 4.961 | 5.045 |  3.621  | 1.734  |
## +----------------+--------+-------+-------+---------+--------+
## |  nr.employed   |  5191  | 5228  | 5228  |  5167   | 72.25  |
## +----------------+--------+-------+-------+---------+--------+
```

**Plotting the features' distribution of Numeric data of dataset 2 with outlier:**

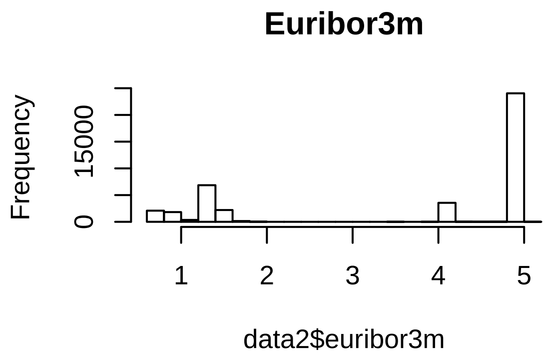
```r
par(mfrow = c(2,2))
hist(data2$age, main="Age")
hist(data2$duration, main="Duration")
hist(data2$campaign, main="Campaign")
hist(data2$pdays, main="Pdays")
```

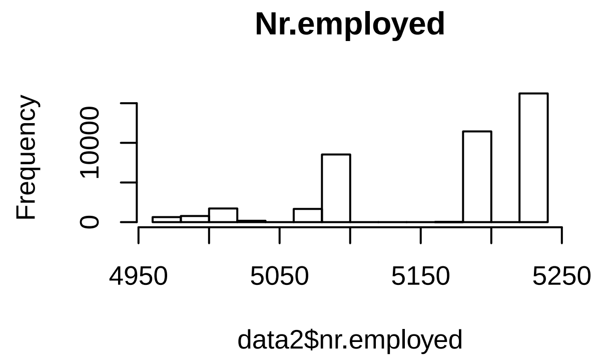



```r
hist(data2$previous, main="Previous")
hist(data2$emp.var.rate, main="Emp.var.rate")
hist(data2$cons.price.idx, main="Cons.price.idx")
hist(data2$cons.conf.idx, main="Cons.conf.idx")
```




```r
hist(data2$euribor3m, main="Euribor3m")
hist(data2$nr.employed, main="Nr.employed")
```




None of the numeric data of dataset 01 are uniformly distributed. Among these, agr, duration, campaignand previous are right skewed, pdays and emp.var.rate are right skewed and cons.price.idx, cons.conf.idx, euribor3m and nr.emplyeed are multimodal.

**Data Quality Report for numeric data of datset 2 without outlier:**

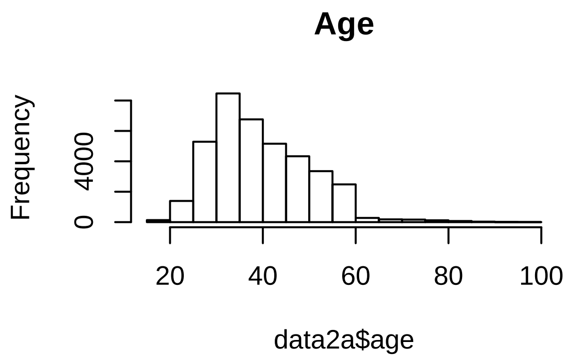
```
# Calling a function to create a DQR for Dataset 2 without outlier:
df2a_dqrnum <- dataQualityNum(data2a)
library(pander)
pandoc.table(df2a_dqrnum, style = "grid", caption = "Data Quality Report for numeric data of datset 2 wi
```

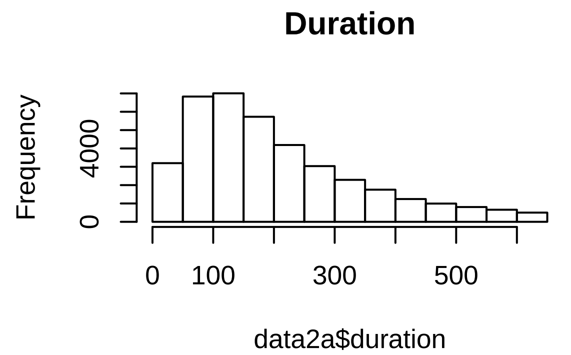
```
##
##
## +----------------+-----------+---------+-------------+-------+-------+
## |    Feature     | Instances | Missing | Cardinality |  Min  |  Q1   |
## +================+===========+=========+=============+=======+=======+
## |      age       |   38225   |    0    |     78      |  17   |  32   |
## +----------------+-----------+---------+-------------+-------+-------+
## |    duration    |   38225   |    0    |     645     |   0   |  97   |
## +----------------+-----------+---------+-------------+-------+-------+
## |    campaign    |   38225   |    0    |     42      |   1   |   1   |
## +----------------+-----------+---------+-------------+-------+-------+
## |     pdays      |   38225   |    0    |     27      |   0   |  999  |
## +----------------+-----------+---------+-------------+-------+-------+
## |    previous    |   38225   |    0    |      8      |   0   |   0   |
## +----------------+-----------+---------+-------------+-------+-------+
```

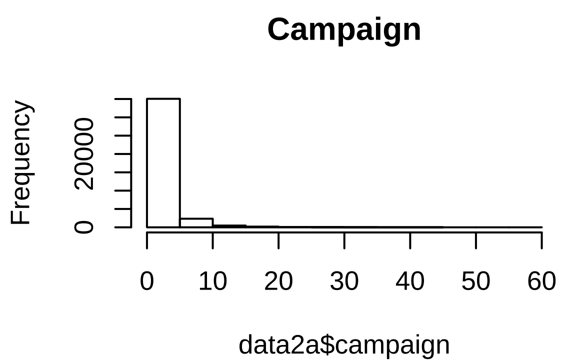
```
## |  emp.var.rate   |   38225   |    0    |      10      | -3.4  | -1.8  |
## +-----------------+-----------+---------+--------------+-------+-------+
## | cons.price.idx  |   38225   |    0    |      26      | 92.2  | 93.08 |
## +-----------------+-----------+---------+--------------+-------+-------+
## | cons.conf.idx   |   38225   |    0    |      26      | -50.8 | -42.7 |
## +-----------------+-----------+---------+--------------+-------+-------+
## |    euribor3m    |   38225   |    0    |     315      | 0.634 | 1.344 |
## +-----------------+-----------+---------+--------------+-------+-------+
## |  nr.employed    |   38225   |    0    |      11      | 4964  | 5099  |
## +-----------------+-----------+---------+--------------+-------+-------+
##
## Table: Data Quality Report for numeric data of datset 2 without outlier (continued below)
##
##
##
## +-----------------+--------+-------+-------+---------+--------+
## |    Feature.1    | Median |  Q3   |  Max  |  Mean   | Stdev  |
## +=================+========+=======+=======+=========+========+
## |       age       |   38   |  47   |  98   |  40.05  | 10.43  |
## +-----------------+--------+-------+-------+---------+--------+
## |    duration     |  167   |  277  |  644  |  203.3  |  141   |
## +-----------------+--------+-------+-------+---------+--------+
## |    campaign     |   2    |   3   |  56   |  2.575  |  2.81  |
## +-----------------+--------+-------+-------+---------+--------+
## |      pdays      |  999   |  999  |  999  |  963.3  | 184.8  |
## +-----------------+--------+-------+-------+---------+--------+
## |    previous     |   0    |   0   |   7   | 0.1732  | 0.4945 |
## +-----------------+--------+-------+-------+---------+--------+
## |  emp.var.rate   |  1.1   |  1.4  |  1.4  | 0.08181 | 1.572  |
## +-----------------+--------+-------+-------+---------+--------+
## | cons.price.idx  | 93.44  | 93.99 | 94.77 |  93.57  | 0.5798 |
## +-----------------+--------+-------+-------+---------+--------+
## | cons.conf.idx   | -41.8  | -36.4 | -26.9 | -40.48  | 4.632  |
## +-----------------+--------+-------+-------+---------+--------+
## |    euribor3m    | 4.857  | 4.961 | 5.045 |  3.623  | 1.734  |
## +-----------------+--------+-------+-------+---------+--------+
## |  nr.employed    |  5191  | 5228  | 5228  |  5167   | 72.08  |
## +-----------------+--------+-------+-------+---------+--------+
```

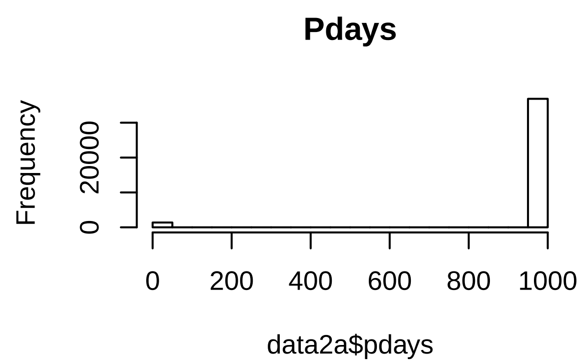
**Plotting the features' distribution of numeric data of dataset 2 without outlier:**

```r
par(mfrow = c(2,2))
hist(data2a$age, main="Age")
hist(data2a$duration, main="Duration")
hist(data2a$campaign, main="Campaign")
hist(data2a$pdays, main="Pdays")
```

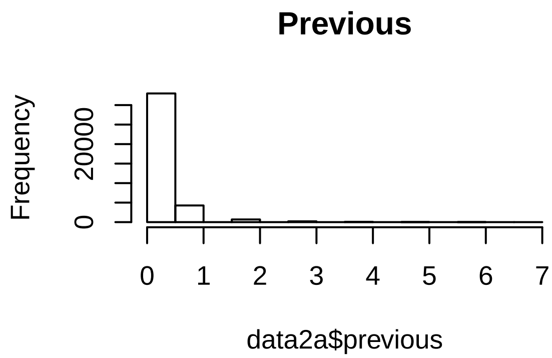


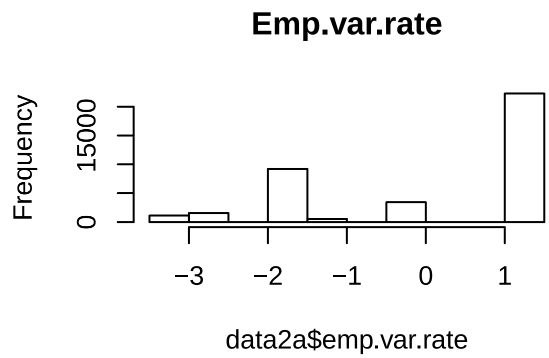



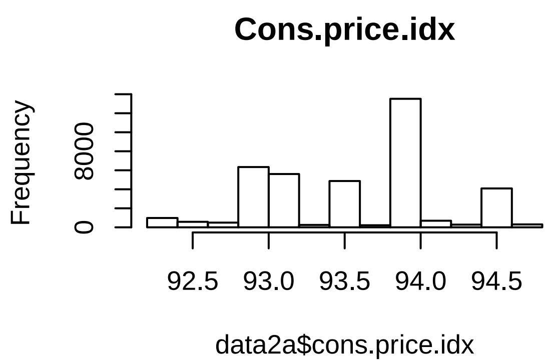



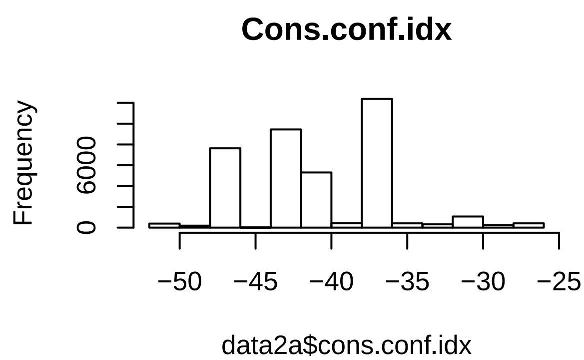




```
hist(data2a$previous, main="Previous")
hist(data2a$emp.var.rate, main="Emp.var.rate")
hist(data2a$cons.price.idx, main="Cons.price.idx")
hist(data2a$cons.conf.idx, main="Cons.conf.idx")
```

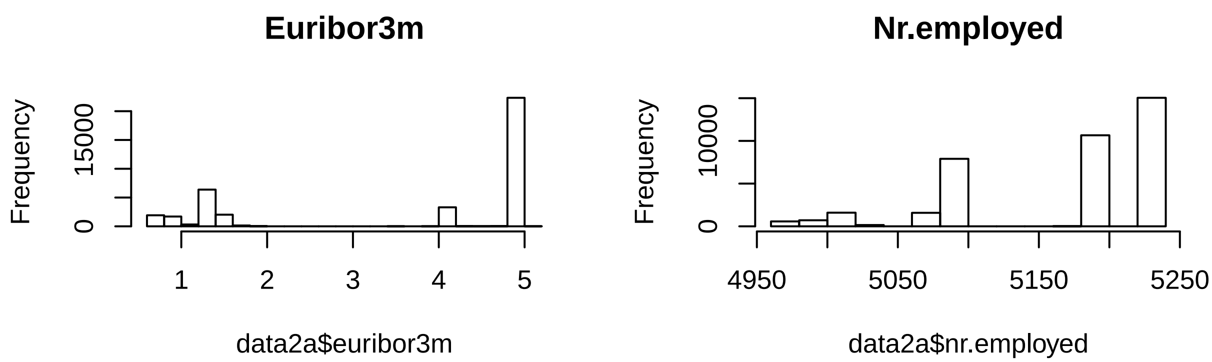



```
hist(data2a$euribor3m, main="Euribor3m")
hist(data2a$nr.employed, main="Nr.employed")
```




After removing outliers numeric variable duration became more accurate and clear and remains right skewed. All other numeric variables are also remain same.

**Data Quality Report for categorical data of dataset 2 with outlier:**

```
# Calling a function to create a DQR for Dataset 2 with outlier:
df2_categorical <- dataQualityCat(data2)
library(pander)
pandoc.table(df2_categorical, style = "grid", caption = "Data Quality Report for categorical data of da
```

```
##
##
## +-------------+-------+------+------+-------------------+-----------+
## |   Feature   | Inst  | Miss | Card |      FstMod       | FstModFrq |
## +=============+=======+======+======+===================+===========+
## |     job     | 41188 |  0   |  11  |      admin.       |   10752   |
## +-------------+-------+------+------+-------------------+-----------+
## |   marital   | 41188 |  0   |  3   |      married      |   25008   |
## +-------------+-------+------+------+-------------------+-----------+
## |  education  | 41188 |  0   |  7   | university.degree |   13899   |
## +-------------+-------+------+------+-------------------+-----------+
## |   default   | 41188 |  0   |  2   |        no         |   41185   |
## +-------------+-------+------+------+-------------------+-----------+
## |   housing   | 41188 |  0   |  2   |        yes        |   22566   |
## +-------------+-------+------+------+-------------------+-----------+
## |    loan     | 41188 |  0   |  2   |        no         |   34940   |
```

```
## +-------------+-------+------+------+------------------+-----------+
## |   contact   | 41188 |  0   |  2   |     cellular     |   26144   |
## +-------------+-------+------+------+------------------+-----------+
## |    month    | 41188 |  0   |  10  |       may        |   13769   |
## +-------------+-------+------+------+------------------+-----------+
## | day_of_week | 41188 |  0   |  5   |       thu        |   8623    |
## +-------------+-------+------+------+------------------+-----------+
## |  poutcome   | 41188 |  0   |  3   |   nonexistent    |   35563   |
## +-------------+-------+------+------+------------------+-----------+
## |      y      | 41188 |  0   |  2   |        no        |   36548   |
## +-------------+-------+------+------+------------------+-----------+
##
## Table: Data Quality Report for categorical data of datset 2 with outlier (continued below)
##
##
##
## +-------------+-----------+-------------+-----------+-----------+
## |  Feature.1  | FstModPnt |   SndMod    | SndModFrq | SndModPnt |
## +=============+===========+=============+===========+===========+
## |     job     |   26.1    | blue-collar |   9254    |   22.47   |
## +-------------+-----------+-------------+-----------+-----------+
## |   marital   |   60.72   |   single    |   11568   |   28.09   |
## +-------------+-----------+-------------+-----------+-----------+
## |  education  |   33.75   | high.school |   9515    |   23.1    |
## +-------------+-----------+-------------+-----------+-----------+
## |   default   |   99.99   |     yes     |     3     | 0.007284  |
## +-------------+-----------+-------------+-----------+-----------+
## |   housing   |   54.79   |     no      |   18622   |   45.21   |
## +-------------+-----------+-------------+-----------+-----------+
## |    loan     |   84.83   |     yes     |   6248    |   15.17   |
## +-------------+-----------+-------------+-----------+-----------+
## |   contact   |   63.47   |  telephone  |   15044   |   36.53   |
## +-------------+-----------+-------------+-----------+-----------+
## |    month    |   33.43   |     jul     |   7174    |   17.42   |
## +-------------+-----------+-------------+-----------+-----------+
## | day_of_week |   20.94   |     mon     |   8514    |   20.67   |
## +-------------+-----------+-------------+-----------+-----------+
## |  poutcome   |   86.34   |   failure   |   4252    |   10.32   |
## +-------------+-----------+-------------+-----------+-----------+
## |      y      |   88.73   |     yes     |   4640    |   11.27   |
## +-------------+-----------+-------------+-----------+-----------+
```

**Data Quality Report for categorical data of dataset 2 without outlier:**

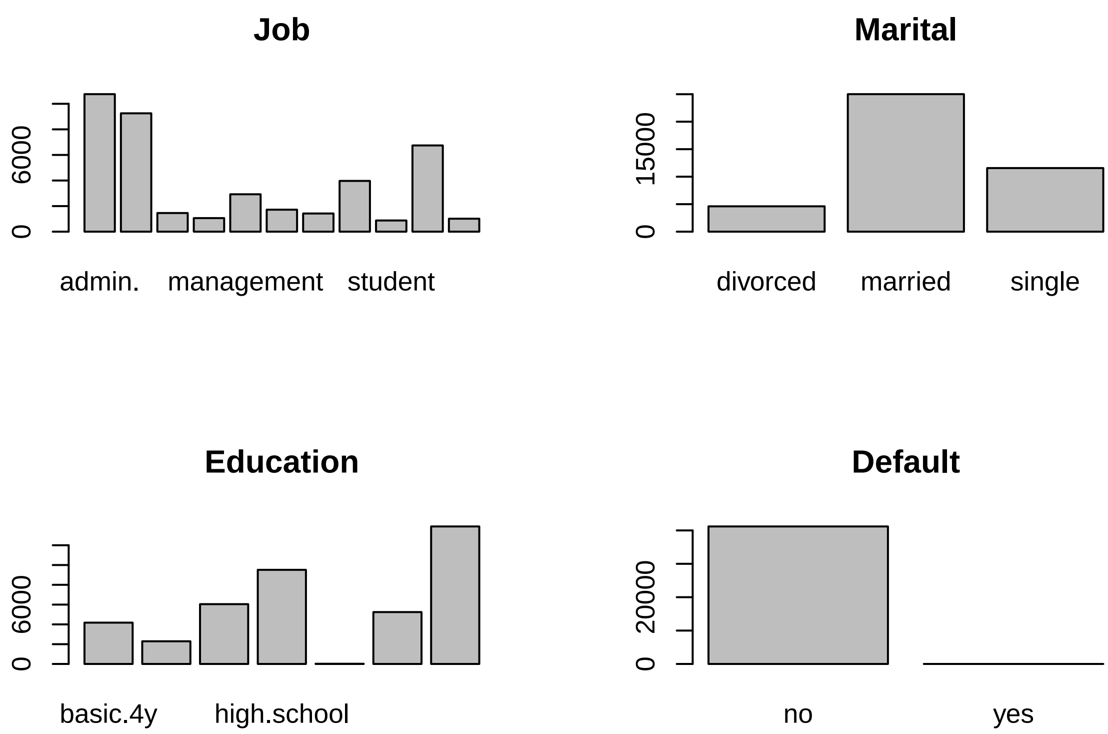
```r
# Calling a function to create a DQR for Dataset 2 without outlier:
df2a_categorical <- dataQualityCat(data2a)
library(pander)
pandoc.table(df2a_categorical, style = "grid", caption = "Data Quality Report for categorical data of da
```

```
##
##
## +-------------+-------+------+------+------------------+-----------+
## |   Feature   | Inst  | Miss | Card |      FstMod      | FstModFrq |
## +=============+=======+======+======+==================+===========+
```

```
## |     job     | 38225 |  0   |  11  |      admin.       |   10018   |
## +-------------+-------+------+------+-------------------+-----------+
## |   marital   | 38225 |  0   |  3   |      married      |   23222   |
## +-------------+-------+------+------+-------------------+-----------+
## |  education  | 38225 |  0   |  7   | university.degree |   12900   |
## +-------------+-------+------+------+-------------------+-----------+
## |   default   | 38225 |  0   |  2   |        no         |   38222   |
## +-------------+-------+------+------+-------------------+-----------+
## |   housing   | 38225 |  0   |  2   |        yes        |   20963   |
## +-------------+-------+------+------+-------------------+-----------+
## |    loan     | 38225 |  0   |  2   |        no         |   32442   |
## +-------------+-------+------+------+-------------------+-----------+
## |   contact   | 38225 |  0   |  2   |     cellular      |   24162   |
## +-------------+-------+------+------+-------------------+-----------+
## |    month    | 38225 |  0   |  10  |        may        |   12818   |
## +-------------+-------+------+------+-------------------+-----------+
## | day_of_week | 38225 |  0   |  5   |        mon        |   7992    |
## +-------------+-------+------+------+-------------------+-----------+
## |  poutcome   | 38225 |  0   |  3   |    nonexistent    |   32982   |
## +-------------+-------+------+------+-------------------+-----------+
## |      y      | 38225 |  0   |  2   |        no         |   35111   |
## +-------------+-------+------+------+-------------------+-----------+
##
## Table: Data Quality Report for categorical data of datset 2 without outlier (continued below)
##
##
##
## +-------------+-----------+-------------+-----------+-----------+
## |  Feature.1  | FstModPnt |   SndMod    | SndModFrq | SndModPnt |
## +=============+===========+=============+===========+===========+
## |     job     |   26.21   | blue-collar |   8565    |   22.41   |
## +-------------+-----------+-------------+-----------+-----------+
## |   marital   |   60.75   |   single    |   10701   |   27.99   |
## +-------------+-----------+-------------+-----------+-----------+
## |  education  |   33.75   | high.school |   8815    |   23.06   |
## +-------------+-----------+-------------+-----------+-----------+
## |   default   |   99.99   |     yes     |     3     | 0.007848  |
## +-------------+-----------+-------------+-----------+-----------+
## |   housing   |   54.84   |     no      |   17262   |   45.16   |
## +-------------+-----------+-------------+-----------+-----------+
## |    loan     |   84.87   |     yes     |   5783    |   15.13   |
## +-------------+-----------+-------------+-----------+-----------+
## |   contact   |   63.21   |  telephone  |   14063   |   36.79   |
## +-------------+-----------+-------------+-----------+-----------+
## |    month    |   33.53   |     jul     |   6535    |   17.1    |
## +-------------+-----------+-------------+-----------+-----------+
## | day_of_week |   20.91   |     thu     |   7942    |   20.78   |
## +-------------+-----------+-------------+-----------+-----------+
## |  poutcome   |   86.28   |   failure   |   3995    |   10.45   |
## +-------------+-----------+-------------+-----------+-----------+
## |      y      |   91.85   |     yes     |   3114    |   8.147   |
## +-------------+-----------+-------------+-----------+-----------+
```

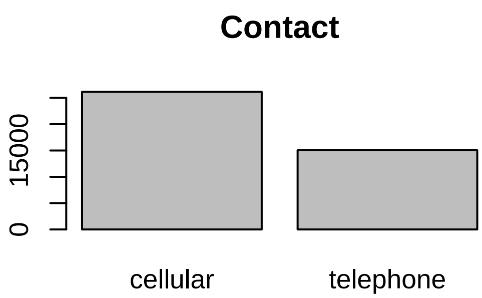
**Plotting the features' distribution of categorical data of dataset 2 with outlier:**

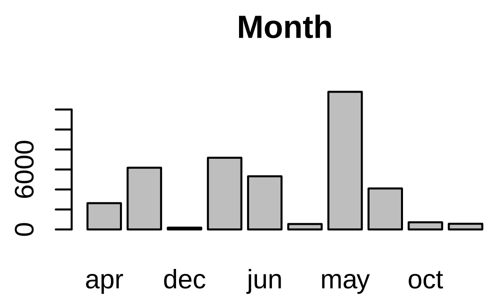
```
par(mfrow = c(2,2))
barplot(table(data2$job), main="Job")
barplot(table(data2$marital), main="Marital")
barplot(table(data2$education), main="Education")
barplot(table(data2$default), main="Default")
```



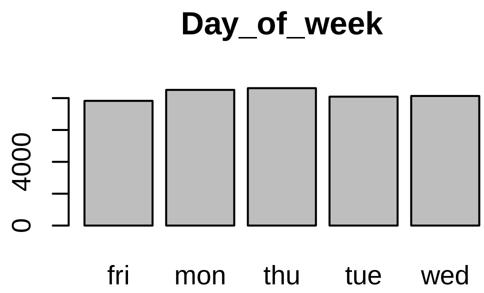

```
barplot(table(data2$housing), main="Housing")
barplot(table(data2$loan), main="Loan")
barplot(table(data2$contact), main="Contact")
barplot(table(data2$month), main="Month")
```

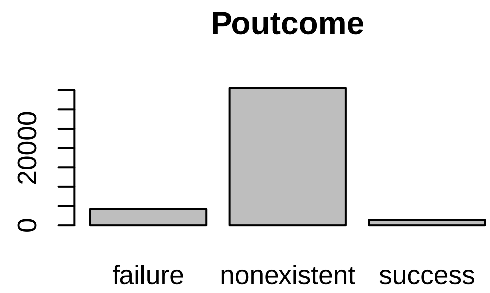


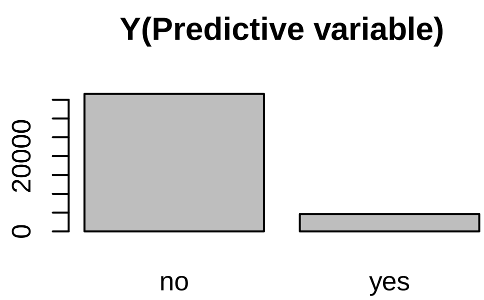



```r
barplot(table(data2$day_of_week), main="Day_of_week")
barplot(table(data2$poutcome), main="Poutcome")
barplot(table(data2$y), main="Y(Predictive variable)")
```



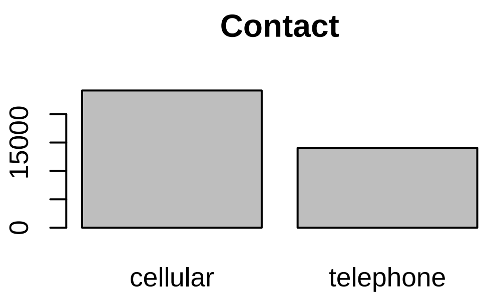

Among categorical variables of dataset 02 almost all are not normally ditributed. Among those categorical variable job, education, month are multi modal where as marital, poutcome are unimodal, where as other are either right or left skewed.

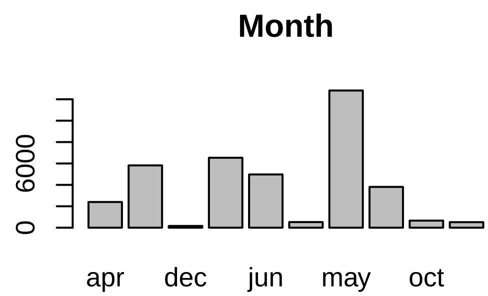
**Plotting the features' distribution of categorical data of dataset 2 without outlier:**

```
par(mfrow = c(2,2))
barplot(table(data2a$job), main="Job")
barplot(table(data2a$marital), main="Marital")
barplot(table(data2a$education), main="Education")
barplot(table(data2a$default), main="Default")
```

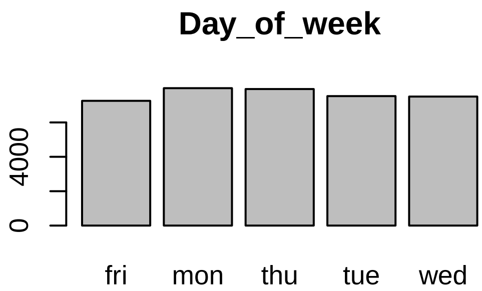

```
barplot(table(data2a$housing), main="Housing")
barplot(table(data2a$loan), main="Loan")
barplot(table(data2a$contact), main="Contact")
barplot(table(data2a$month), main="Month")
```

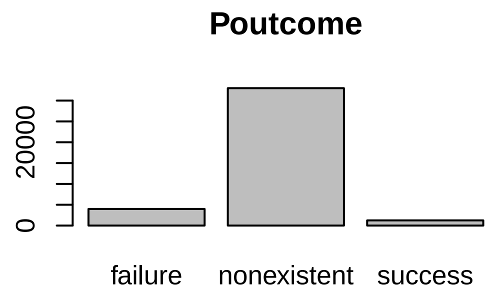


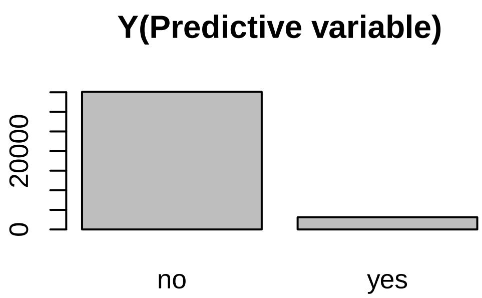



```r
barplot(table(data2a$day_of_week), main="Day_of_week")
barplot(table(data2a$poutcome), main="Poutcome")
barplot(table(data2a$y), main="Y(Predictive variable)")
```




The removal of outliers has no impact on categorical variables of dataset 02 so all of them remains same.

## Discussion:

Let's compare these two different datsets using common graphical visualisation.

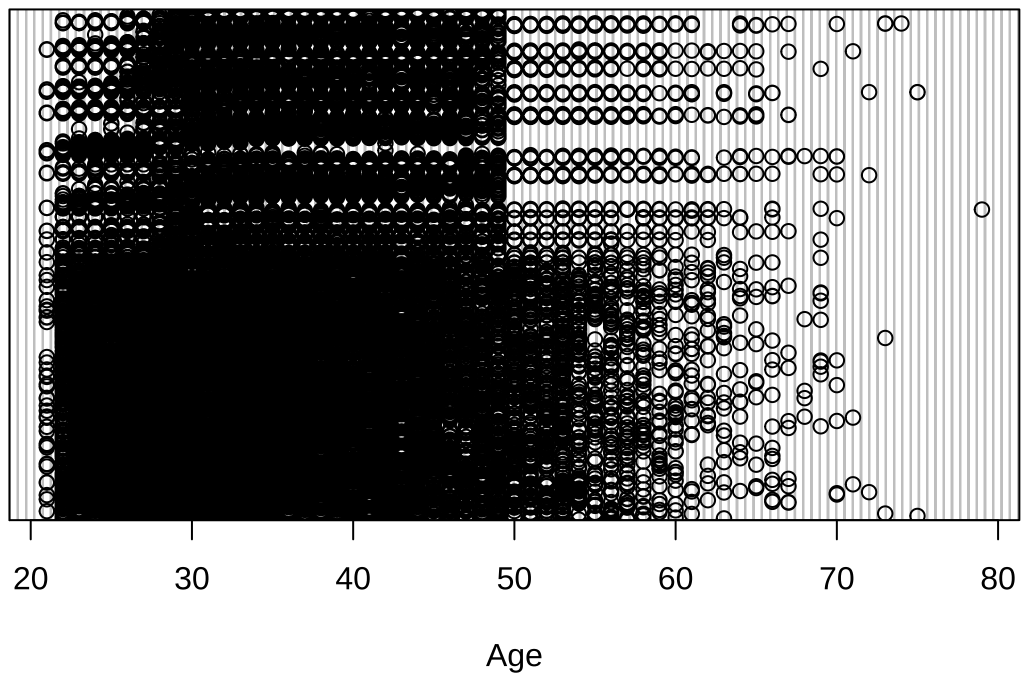
Firsly, we will compare the quanittative variables(numeric). Usually while comparing numeric variables of two datsets we will mostly focus on following 4 features namesly,

1. Center - Median of the variable
2. Spread - The range or Interquantile range
3. Shape - Symmetry, skewness, peaks
4. Unusual features - Gaps, clusters, outliers

First, let us discuss about a common numericl variable in both datsets "Age".

### Dot Plot:

```
#Generting dot plot for numerica vriable age of dataset 01
dotchart(data1$AGE, xlab="Age")
```

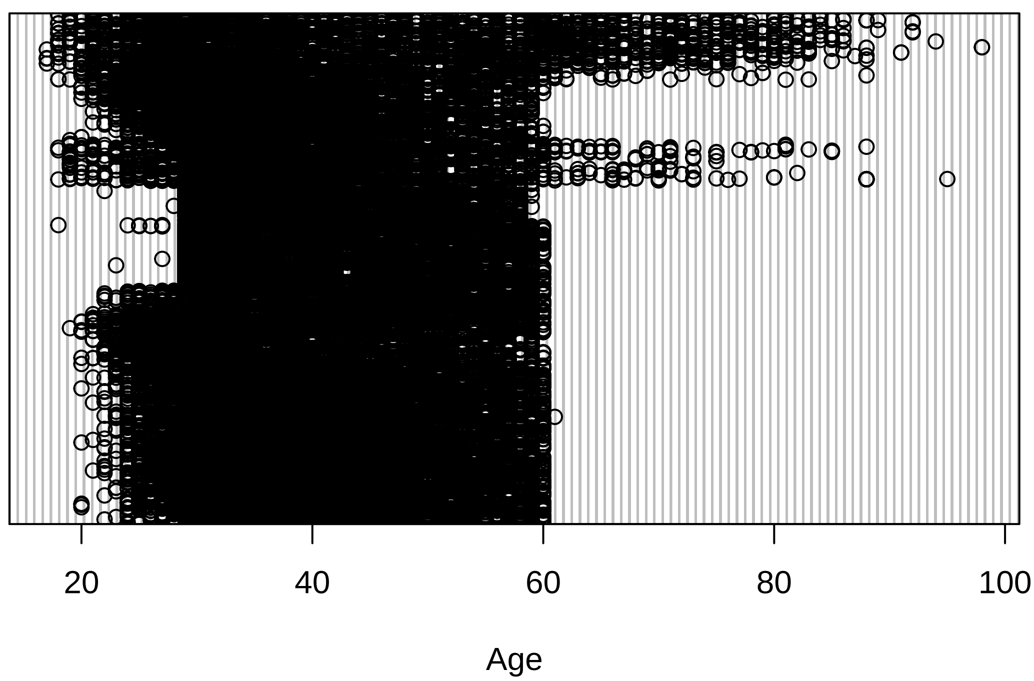



```
#Finding of 4 factors for numeric variable "Age" of dataset 01
#it is a asymmetric shape but when outliers are remove it almost forms bellshape
shape1 <- "Asymmetric bell curve"
uf1 <- "Gaps and Outliers"
range1 <- range(data1$AGE)
rl1 <- range1[1]
rh1 <- range1[2]
#let's create a dataframe in order to explain the above mentioned 4 factors in the table format.
cmp1 <- data.frame(Center = median(data1$AGE),
                                Spreadlowerrange = rl1,
                                Spreadhigherrange = rh1,
                                Shape = shape1,
                                UnusualFeature = uf1)


pandoc.table(cmp1, style = "grid", caption = "Factors of numerical variable Age of datset 01") #Generat
```

```
##
##
## +--------+------------------+-------------------+-----------------------+
## | Center | Spreadlowerrange | Spreadhigherrange |         Shape         |
## +========+==================+===================+=======================+
## |   34   |        21        |        79         | Asymmetric bell curve |
## +--------+------------------+-------------------+-----------------------+
##
## Table: Factors of numerical variable Age of datset 01 (continued below)
```

```
##
##
##
## +-------------------+
## |  UnusualFeature   |
## +===================+
## | Gaps and Outliers |
## +-------------------+
```

```
#Generting dot plot for numerica vriable age of dataset 02
dotchart(data2$age, xlab="Age")
```




```
#Finding of 4 factors for numeric variable "Age" of dataset 02
#it is a asymmetric shape but when outliers are remove it almost forms inverted bellshape
shape2 <- "Asymmetric inverted bell curve"
uf2 <- "Gaps and Outliers"
range2 <- range(data2$age)
rl2 <- range2[1]
rh2 <- range2[2]
#let's create a function in order to explain the above mentioned 4 factors in the table format.
cmp2 <- data.frame(Center = median(data2$age),
                                Spreadlowerrange = rl2,
                                Spreadhigherrange = rh2,
                                Shape = shape2,
                                UnusualFeature = uf2)
```

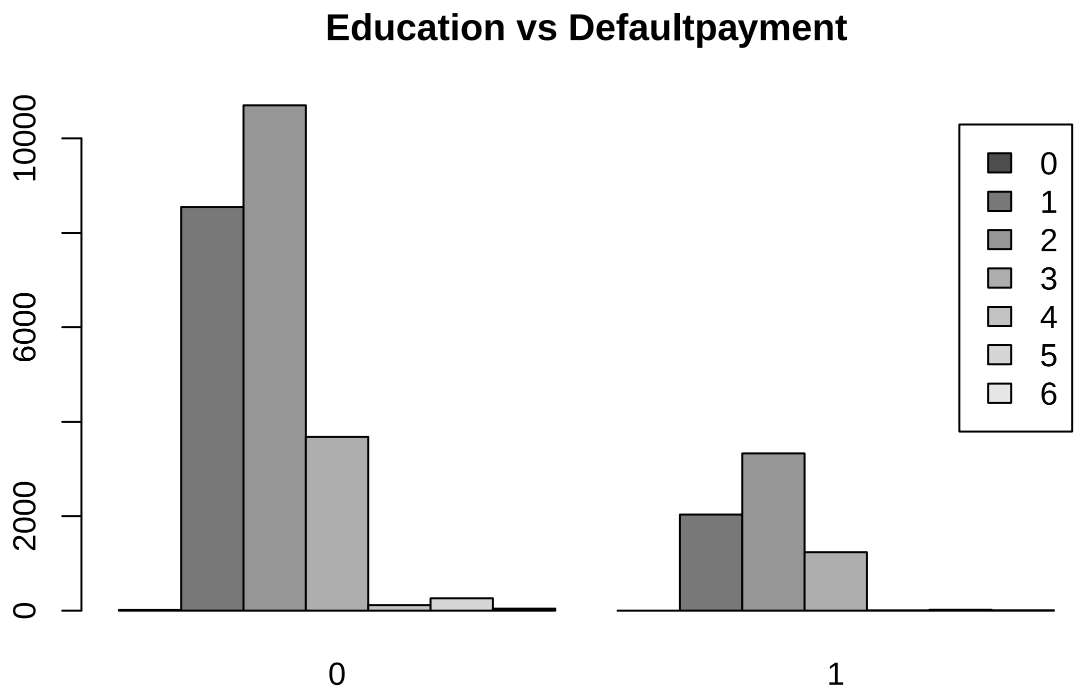
```r
pandoc.table(cmp2, style = "grid", caption = "Factors of numerical variable Age of datset 02") #Generat
```

```
##
##
## +---------+------------------+-------------------+
## | Center  | Spreadlowerrange | Spreadhigherrange |
## +=========+==================+===================+
## |   38    |        17        |        98         |
## +---------+------------------+-------------------+
##
## Table: Factors of numerical variable Age of datset 02 (continued below)
##
##
##
## +--------------------------------+-------------------+
## |             Shape              |  UnusualFeature   |
## +================================+===================+
## | Asymmetric inverted bell curve | Gaps and Outliers |
## +--------------------------------+-------------------+
```

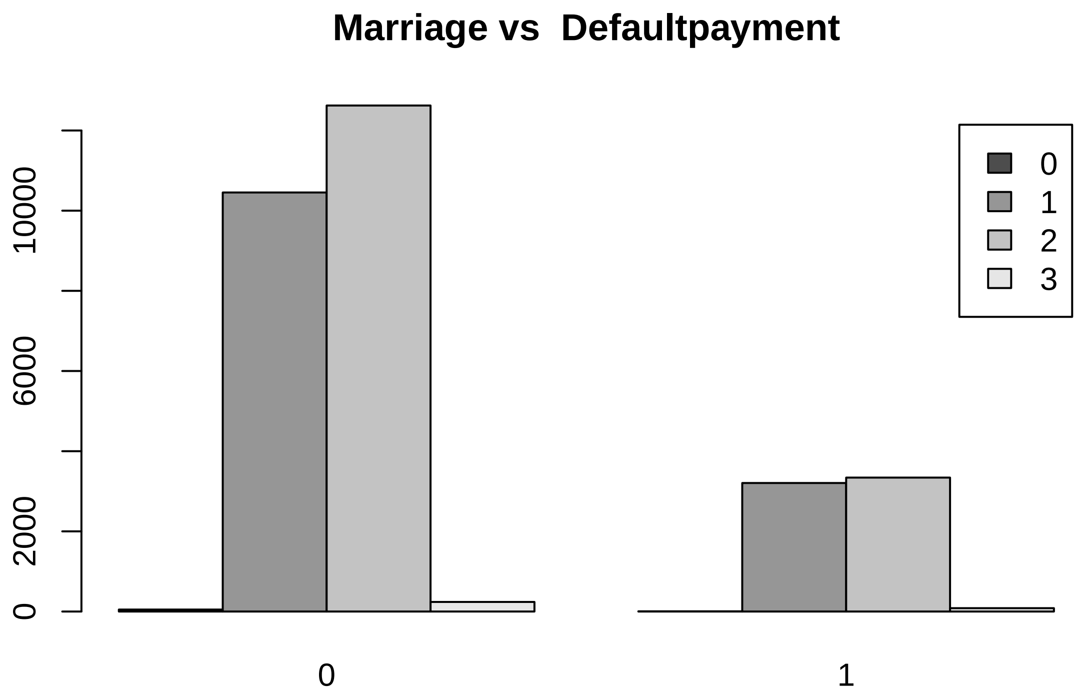
**Now, let's compare the common categorical variables**

```r
barplot(table(data1$EDUCATION,data1$MARRIAGE),beside = T,legend.text = T, main="Education vs Marry statu
```

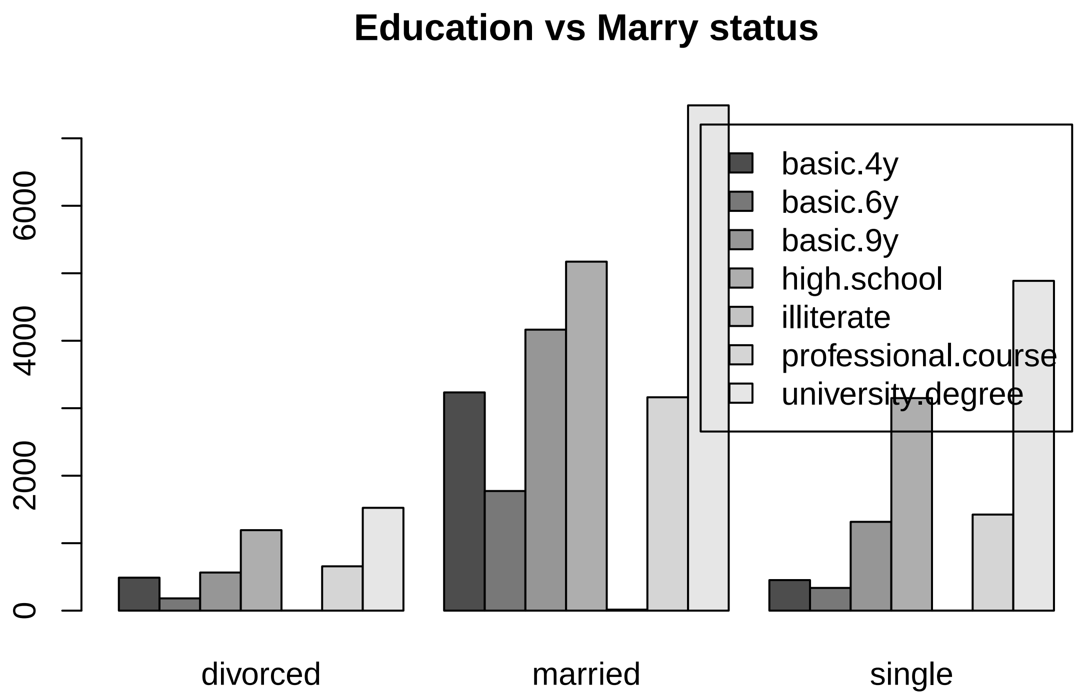

```r
barplot(table(data1$EDUCATION,data1$default.payment.next.month),beside = T,legend.text = T, main="Educat
```



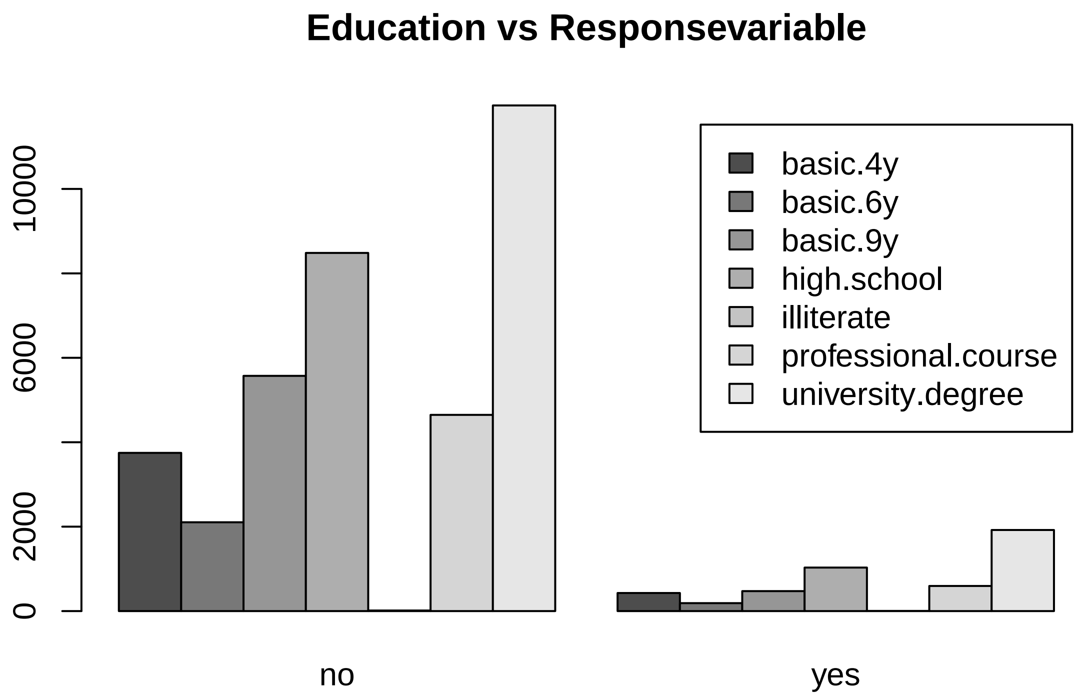

```
barplot(table(data1$MARRIAGE,data1$default.payment.next.month),beside = T,legend.text = T, main="Marria
```



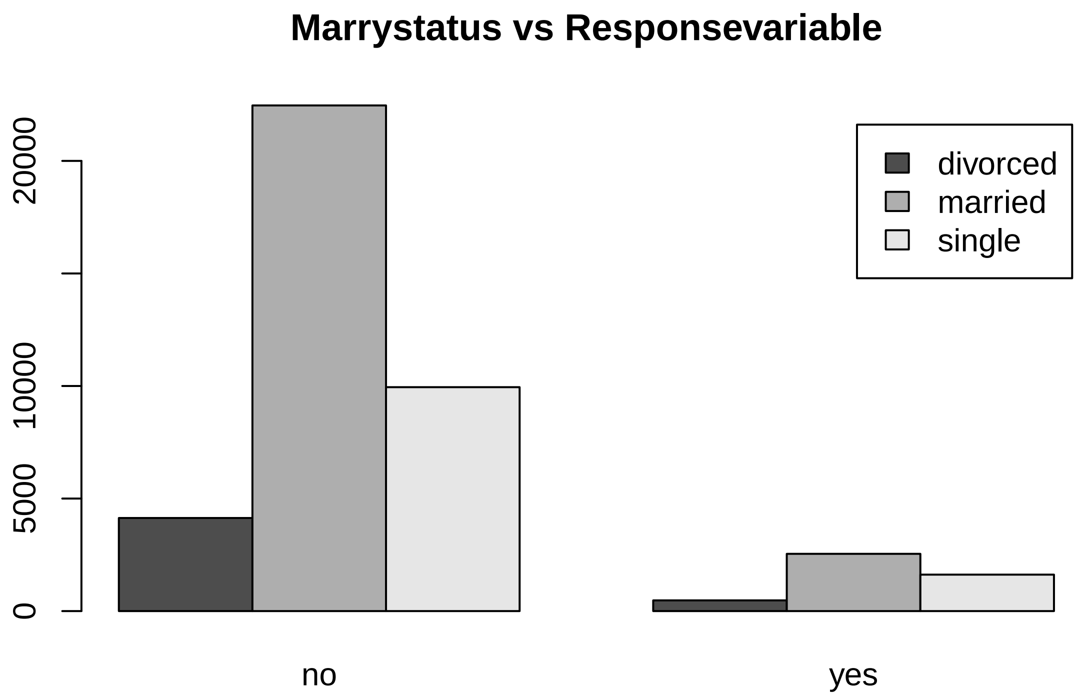

```
barplot(table(data2$education,data2$marital),beside = T,legend.text = T, main="Education vs Marry statu
```




```
barplot(table(data2$education,data2$y),beside = T,legend.text = T, main="Education vs Responsevariable")
```




```
barplot(table(data2$marital,data2$y),beside = T,legend.text = T, main="Marrystatus vs Responsevariable")
```




**Comparison and contrast of common numerical variable age in the both dataset**

In the predictive analysis of the credit card default payment(dataset 01) did not involved any teen ages but whereas in the predictive analysis of Bank marketing(dataset 02) teen ages were involved. In the both datasets most of the people aged between 30 to 40 were involved.

In the predictive analysis of both dataset married people opted options more than singles so as per analysis it seem married person are more responsible(subscribed for term deposit for future saving) than single people and they don't needanymore commitment and so they opted for default credible. But most of the educated people didn't subscribed term deposit where as more educated people opted for default payment.

The dataset 02 doesn't have any privacy data so there will not be any GDPR issue. As the dataset02 had being analysed using CRISP-DM methodology, it is in the detailed form with all required information which motivates me to select this dataset for this assignment and also I will select this dataset for my final project as well. Being the part of FinTech programme the banking dataset will be better suits the data analyst project.

## References:

[1] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. "The Kdd Process for Extracting Useful Knowledge from Volumes of Data." [2] Soui, M., Smiti, S., Bribech, S., Gasmi, I. Credit card default prediction as a classification problem (2018) Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10868 LNAI, pp. 88-100. [3] S. Moro, P. Cortez and P. Rita. 2014. "A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems", In press, http://dx.doi.org/10.1016/j.dss.2014.03.001 [4] Olson, D.L. & Chae, B. 2012, "Direct marketing decision support through predictive customer response modeling", Decision Support Systems, vol. 54, no. 1, pp. 443-451. [5] Olden J.D., Lawler J.J. and Poff N.L., 2008. Machine learning methods without tears: A primer for ecologists. Q. Rev. Biol., 83, 171-193. [6] OLAYA-MARÍN, E.J., MARTÍNEZ-CAPEL, F.

and VEZZA, P., 2013. A comparison of artificial neural networks and random forests to predict native fish species richness in Mediterranean rivers. Knowledge and Management of Aquatic Ecosystems, (409),. [7] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]