DMPM assignment 1

Name:: Saniya Inamdar

SRN:: 201900913

Roll.no:: 17 Division : TY A

1. Read the file **"pva97nk.csv"** that is supplied to you.

```
data <- read.csv("pva97nk.csv")
head(data)
```

TargetB	ID	TargetD	GiftCnt36	GiftCntAll	GiftCntCard36	GiftCntCardAll	GiftAvgLast	GiftAvg36	GiftAvgAll	...	PromCntCardAll	StatusCat96NK	StatusCa
0	14974	NA	2	4	1	3	17	13.50	9.25	...	13	A	
0	6294	NA	1	8	0	3	20	20.00	15.88	...	24	A	
1	46110	4	6	41	3	20	6	5.17	3.73	...	22	S	
1	185937	10	3	12	3	8	10	8.67	8.50	...	16	E	
0	29637	NA	1	1	1	1	20	20.00	20.00	...	6	F	
1	112632	11	3	11	2	9	11	10.33	8.27	...	22	S	

2. Identify the variables in the file "pva97nk.csv" and determine whether any variable has any missing values.

```
summary(data)
```

```
    TargetB             ID             TargetD         GiftCnt36     
 Min.   :0.0   Min.   :    12   Min.   :  1.00   Min.   : 0.000  
 1st Qu.:0.0   1st Qu.: 48836   1st Qu.: 10.00   1st Qu.: 2.000  
 Median :0.5   Median : 99106   Median : 13.00   Median : 3.000  
 Mean   :0.5   Mean   : 97975   Mean   : 15.62   Mean   : 3.205  
 3rd Qu.:1.0   3rd Qu.:148539   3rd Qu.: 20.00   3rd Qu.: 4.000  
 Max.   :1.0   Max.   :191779   Max.   :200.00   Max.   :16.000  
                                NA's   :4843                     
   GiftCntAll     GiftCntCard36   GiftCntCardAll    GiftAvgLast     
 Min.   : 1.00   Min.   :0.000   Min.   : 0.000   Min.   :  0.00  
 1st Qu.: 4.00   1st Qu.:1.000   1st Qu.: 2.000   1st Qu.: 10.00  
 Median : 8.00   Median :1.000   Median : 4.000   Median : 15.00  
 Mean   :10.51   Mean   :1.857   Mean   : 5.582   Mean   : 16.02  
 3rd Qu.:15.00   3rd Qu.:3.000   3rd Qu.: 8.000   3rd Qu.: 20.00  
 Max.   :91.00   Max.   :9.000   Max.   :41.000   Max.   :450.00  

   GiftAvg36        GiftAvgAll      GiftAvgCard36     GiftTimeLast   
 Min.   :  0.00   Min.   :  1.50   Min.   :  1.33   Min.   : 4     
 1st Qu.:  9.60   1st Qu.:  7.75   1st Qu.:  8.67   1st Qu.:16     
 Median : 13.50   Median : 10.71   Median : 12.50   Median :18     
 Mean   : 14.88   Mean   : 12.49   Mean   : 14.22   Mean   :18     
 3rd Qu.: 18.50   3rd Qu.: 15.00   3rd Qu.: 18.00   3rd Qu.:20     
 Max.   :260.00   Max.   :450.00   Max.   :260.00   Max.   :27     
                                   NA's   :1780                    
 GiftTimeFirst      PromCnt12        PromCnt36        PromCntAll     
 Min.   : 15.0   Min.   : 2.00   Min.   : 4.00   Min.   :  5.00  
 1st Qu.: 36.0   1st Qu.:11.00   1st Qu.:25.00   1st Qu.: 29.00  
 Median : 68.0   Median :12.00   Median :31.00   Median : 48.00  
 Mean   : 71.1   Mean   :12.99   Mean   :29.35   Mean   : 48.48  
 3rd Qu.:105.0   3rd Qu.:13.00   3rd Qu.:33.00   3rd Qu.: 65.00  
 Max.   :260.0   Max.   :59.00   Max.   :78.00   Max.   :174.00  
```

```
 PromCntCard12    PromCntCard36   PromCntCardAll  StatusCat96NK
 Min.   : 0.000   Min.   : 2.00   Min.   : 2.00   A:5826
 1st Qu.: 5.000   1st Qu.: 7.00   1st Qu.:12.00   E: 227
 Median : 6.000   Median :13.00   Median :19.00   F: 660
 Mean   : 5.392   Mean   :11.95   Mean   :19.01   L:  34
 3rd Qu.: 6.000   3rd Qu.:16.00   3rd Qu.:26.00   N: 574
 Max.   :17.000   Max.   :28.00   Max.   :56.00   S:2365

 StatusCatStarAll   DemCluster        DemAge      DemGender DemHomeOwner
 Min.   :0.0000   Min.   : 0.00   Min.   : 0.00   F:5223    H:5377
 1st Qu.:0.0000   1st Qu.:14.00   1st Qu.:47.00   M:3925    U:4309
 Median :1.0000   Median :27.00   Median :60.00   U: 538
 Mean   :0.5406   Mean   :27.15   Mean   :59.15
 3rd Qu.:1.0000   3rd Qu.:40.00   3rd Qu.:73.00
 Max.   :1.0000   Max.   :53.00   Max.   :87.00
                                  NA's   :2407
 DemMedHomeValue  DemPctVeterans  DemMedIncome
 Min.   :     0   Min.   : 0.0   Min.   :     0
 1st Qu.: 52300   1st Qu.:25.0   1st Qu.: 24464
 Median : 76900   Median :31.0   Median : 43100
 Mean   :110986   Mean   :30.6   Mean   : 40491
 3rd Qu.:128175   3rd Qu.:37.0   3rd Qu.: 56876
 Max.   :600000   Max.   :85.0   Max.   :200001
```

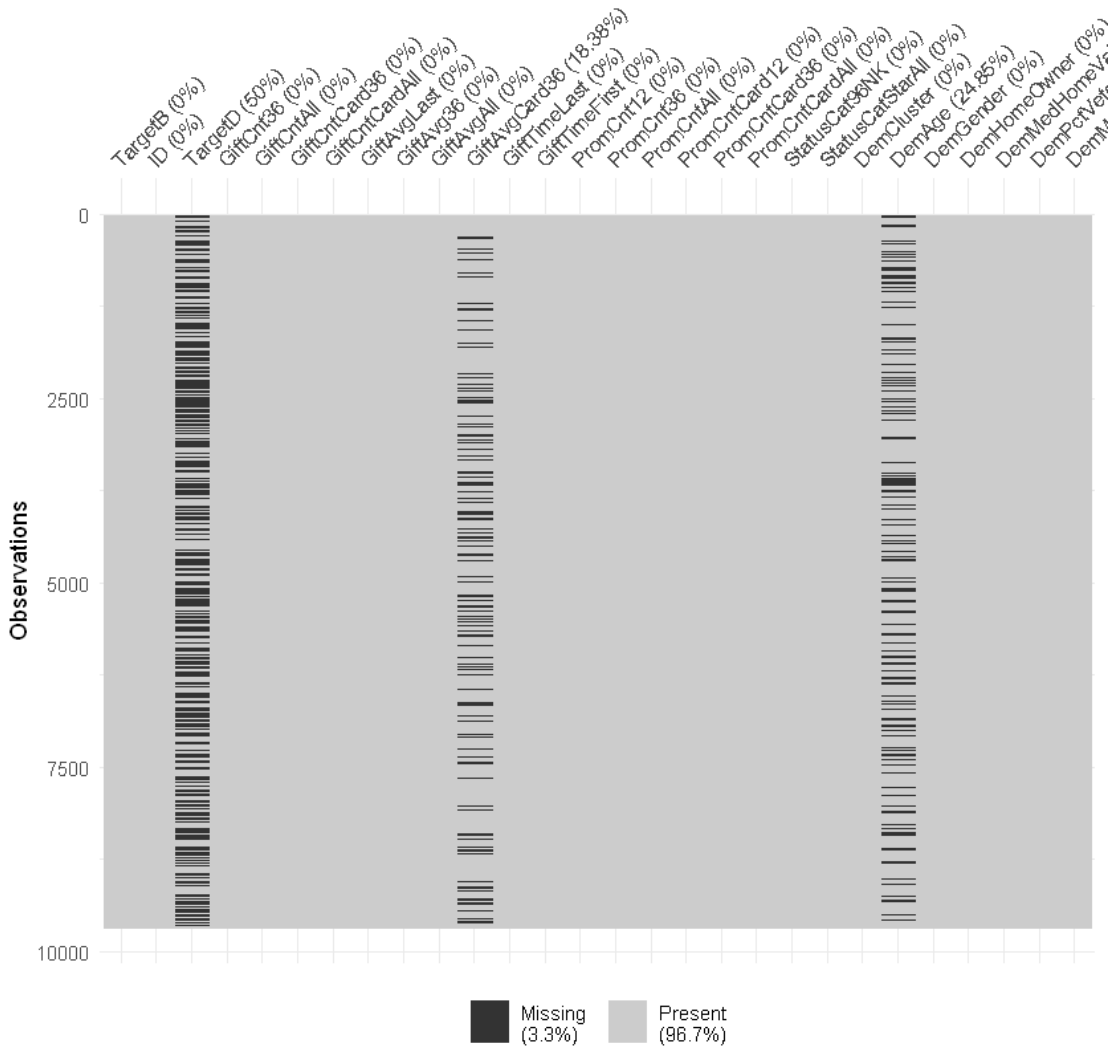
```
missing<-colnames(data)[colSums(is.na(data)) > 0]
```

```
missing
```

'TargetD' 'GiftAvgCard36' 'DemAge'

There are three missing variables as observed above and also visualised from the graph below.

There are many NaN values in TargetID variable as compared to GiftAvgCard36 and DemAge.



---

3. Impute some of the variables that have missing values using their corresponding mean values. Verify whether your task has been correctly done.

   In TargetD variable which is basically the value of donation received so we cannot impute it as mean, since No donations are set as NaN, so we change them to 0.

   Where as we impute the other two variables to their mean values.

```
data$DemAge[is.na(data$DemAge)] <- mean(data$DemAge, na.rm=TRUE)
data$GiftAvgCard36[is.na(data$GiftAvgCard36)] <- mean(data$GiftAvgCard36, na.rm=TRUE)
```

```
data$TargetD[is.na(data$TargetD)] <- 0
```

   Now we check if there are any missing values in all the columns as a form of verification, as observed there aren't any missing values after imputation.

```
sapply(data, function(x) sum(is.na(x)))
```

TargetB	0
ID	0
TargetD	0
GiftCnt36	0
GiftCntAll	0
GiftCntCard36	0
GiftCntCardAll	0
GiftAvgLast	0
GiftAvg36	0
GiftAvgAll	0
GiftAvgCard36	0
GiftTimeLast	0
GiftTimeFirst	0
PromCnt12	0
PromCnt36	0
PromCntAll	0
PromCntCard12	0
PromCntCard36	0
PromCntCardAll	0
StatusCat96NK	0
StatusCatStarAll	0
DemCluster	0
DemAge	0
DemGender	0
DemHomeOwner	0
DemMedHomeValue	0
DemPctVeterans	0

4. Compute the Kurtosis and Skewness of the variables and interpret the results obtained.

```
data2 <- select_if(data, is.numeric)            # Subset numeric columns with dplyr
```

```
sapply(data2, function(x) kurtosis(x))
```

TargetB	1
ID	1.76499388908142
TargetD	44.6852040150102
GiftCnt36	5.04573866222823
GiftCntAll	9.04402476685775
GiftCntCard36	4.49347650451663
GiftCntCardAll	5.02319982552531
GiftAvgLast	248.922802332594
GiftAvg36	80.0595530934173
GiftAvgAll	564.464672945906
GiftAvgCard36	110.349361431637
GiftTimeLast	5.46718249607676
GiftTimeFirst	1.75215759300893
PromCnt12	14.9885668393101
PromCnt36	5.17259965595257
PromCntAll	3.21586433137725
PromCntCard12	8.79507311881507
PromCntCard36	2.01304446955049
PromCntCardAll	2.21946780973605
StatusCatStarAll	1.02651462219208
DemCluster	1.87733646410766
DemAge	3.35582970621106
DemMedHomeValue	9.44741599847037
DemPctVeterans	4.27313325355918
DemMedIncome	3.6358994929709

```
sapply(data2, function(x) skewness(x))
```

TargetB	0
ID	-0.0576037092707591
TargetD	4.1700379279157
GiftCnt36	1.28815342983378
GiftCntAll	1.86282019252729
GiftCntCard36	1.17227085325347
GiftCntCardAll	1.33114710860084
GiftAvgLast	9.91735696141465
GiftAvg36	5.62691999751938
GiftAvgAll	14.4842458420671
GiftAvgCard36	6.69685963638864
GiftTimeLast	-0.777926772105199
GiftTimeFirst	0.195368919079909
PromCnt12	2.87327793766051
PromCnt36	0.261917077109174
PromCntAll	0.460694020376593
PromCntCard12	0.684887899967141
PromCntCard36	-0.426533547242262
PromCntCardAll	0.142833837257773
StatusCatStarAll	-0.162833111473325
DemCluster	-0.086701309265669
DemAge	-0.447383274082047
DemMedHomeValue	2.37784234048721
DemPctVeterans	-0.207026460895079
DemMedIncome	0.309976854620842

I have computed te skewness and kurtosis of all the columns. It basically helps us to check weather our data is normally distributed, or a measure of symmetry or asymmetry of data distribution.

As observed the data is not properly distributed, the most common way to tackle this is by taking log.

5. Determine the "summary" information for the numerical variables.

```
summary(data2)
```

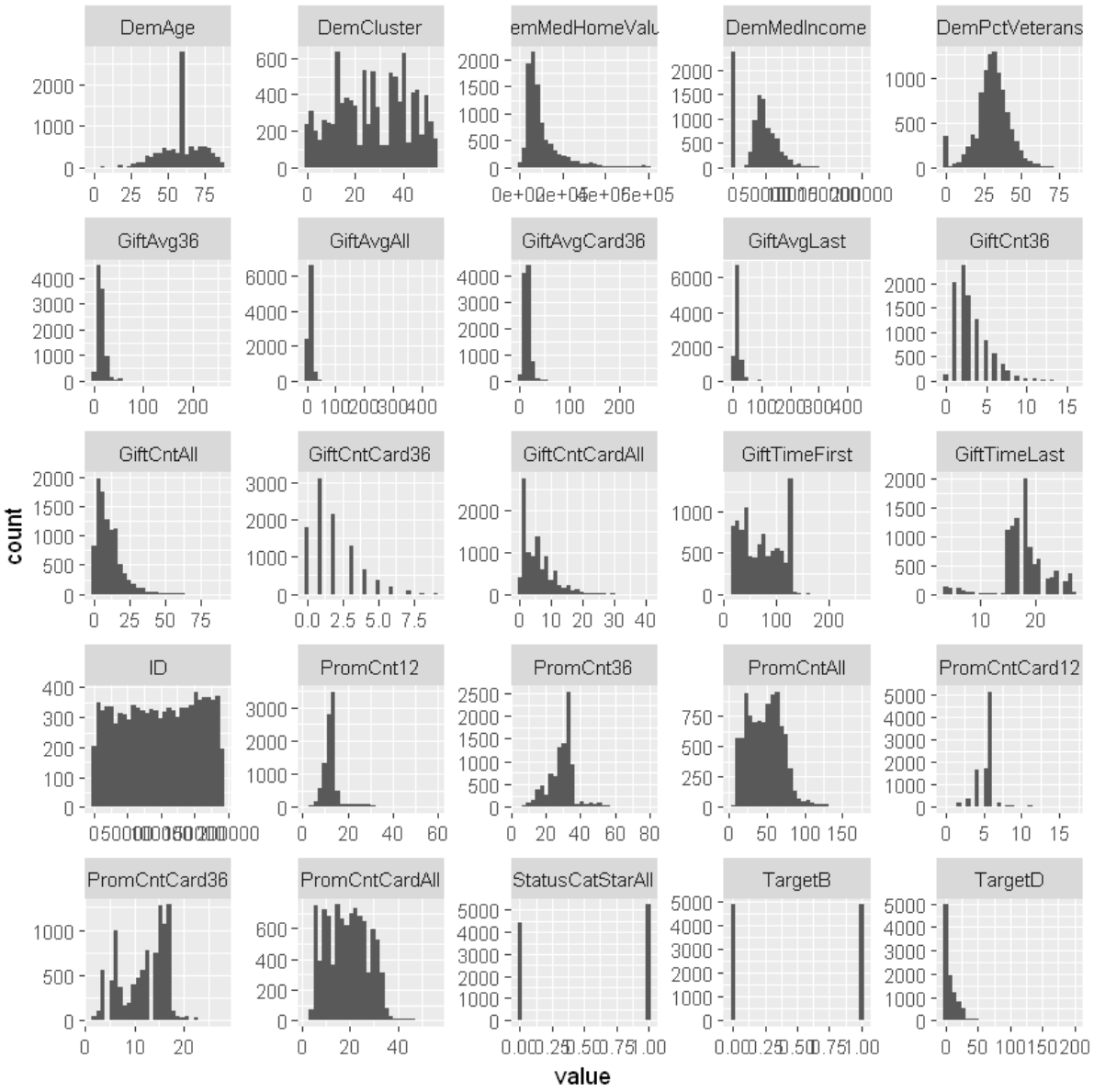
```
    TargetB             ID             TargetD          GiftCnt36     
 Min.   :0.0   Min.   :    12   Min.   :  0.000   Min.   : 0.000  
 1st Qu.:0.0   1st Qu.: 48836   1st Qu.:  0.000   1st Qu.: 2.000  
 Median :0.5   Median : 99106   Median :  0.500   Median : 3.000  
 Mean   :0.5   Mean   : 97975   Mean   :  7.812   Mean   : 3.205  
 3rd Qu.:1.0   3rd Qu.:148539   3rd Qu.: 13.000   3rd Qu.: 4.000  
 Max.   :1.0   Max.   :191779   Max.   :200.000   Max.   :16.000  
   GiftCntAll     GiftCntCard36   GiftCntCardAll    GiftAvgLast     
 Min.   :  1.00   Min.   :0.000   Min.   :  0.000   Min.   :  0.00  
 1st Qu.:  4.00   1st Qu.:1.000   1st Qu.:  2.000   1st Qu.: 10.00  
 Median :  8.00   Median :1.000   Median :  4.000   Median : 15.00  
 Mean   : 10.51   Mean   :1.857   Mean   :  5.582   Mean   : 16.02  
 3rd Qu.: 15.00   3rd Qu.:3.000   3rd Qu.:  8.000   3rd Qu.: 20.00  
 Max.   :91.00   Max.   :9.000   Max.   :41.000   Max.   :450.00  
   GiftAvg36        GiftAvgAll      GiftAvgCard36     GiftTimeLast   
 Min.   :  0.00   Min.   :  1.50   Min.   :  1.33   Min.   :  4  
 1st Qu.:  9.60   1st Qu.:  7.75   1st Qu.: 10.00   1st Qu.: 16  
 Median : 13.50   Median : 10.71   Median : 14.22   Median : 18  
 Mean   : 14.88   Mean   : 12.49   Mean   : 14.22   Mean   : 18  
 3rd Qu.: 18.50   3rd Qu.: 15.00   3rd Qu.: 15.38   3rd Qu.: 20  
 Max.   :260.00   Max.   :450.00   Max.   :260.00   Max.   : 27  
 GiftTimeFirst     PromCnt12        PromCnt36       PromCntAll     
 Min.   : 15.0   Min.   :  2.00   Min.   :  4.00   Min.   :  5.00  
 1st Qu.: 36.0   1st Qu.: 11.00   1st Qu.: 25.00   1st Qu.: 29.00  
 Median : 68.0   Median : 12.00   Median : 31.00   Median : 48.00  
 Mean   : 71.1   Mean   : 12.99   Mean   : 29.35   Mean   : 48.48  
 3rd Qu.:105.0   3rd Qu.: 13.00   3rd Qu.: 33.00   3rd Qu.: 65.00  
 Max.   :260.0   Max.   : 59.00   Max.   : 78.00   Max.   :174.00  
 PromCntCard12    PromCntCard36    PromCntCardAll   StatusCatStarAll 
 Min.   :  0.000   Min.   :  2.00   Min.   :  2.00   Min.   :0.0000  
 1st Qu.:  5.000   1st Qu.:  7.00   1st Qu.: 12.00   1st Qu.:0.0000  
 Median :  6.000   Median : 13.00   Median : 19.00   Median :1.0000  
 Mean   :  5.392   Mean   : 11.95   Mean   : 19.01   Mean   :0.5406  
 3rd Qu.:  6.000   3rd Qu.: 16.00   3rd Qu.: 26.00   3rd Qu.:1.0000  
 Max.   : 17.000   Max.   : 28.00   Max.   : 56.00   Max.   :1.0000  

  DemCluster        DemAge       DemMedHomeValue  DemPctVeterans  
 Min.   : 0.00   Min.   : 0.00   Min.   :     0   Min.   : 0.0  
 1st Qu.:14.00   1st Qu.:51.00   1st Qu.: 52300   1st Qu.:25.0  
 Median :27.00   Median :59.15   Median : 76900   Median :31.0  
 Mean   :27.15   Mean   :59.15   Mean   :110986   Mean   :30.6  
 3rd Qu.:40.00   3rd Qu.:69.00   3rd Qu.:128175   3rd Qu.:37.0  
 Max.   :53.00   Max.   :87.00   Max.   :600000   Max.   :85.0  
  DemMedIncome  
 Min.   :     0  
 1st Qu.: 24464  
 Median : 43100  
 Mean   : 40491  
 3rd Qu.: 56876  
 Max.   :200001  
```

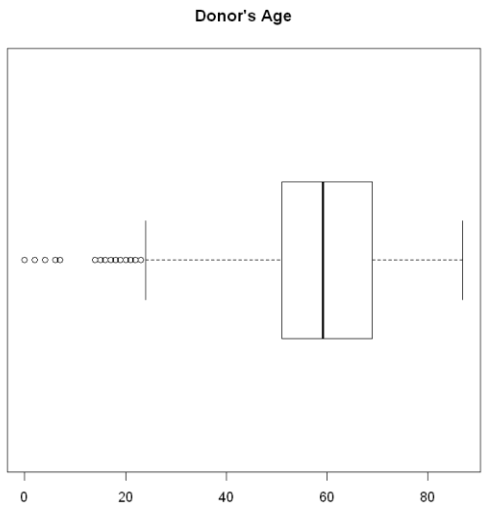
6. Identify the "distributions" of the numerical variables and plot the distributions.

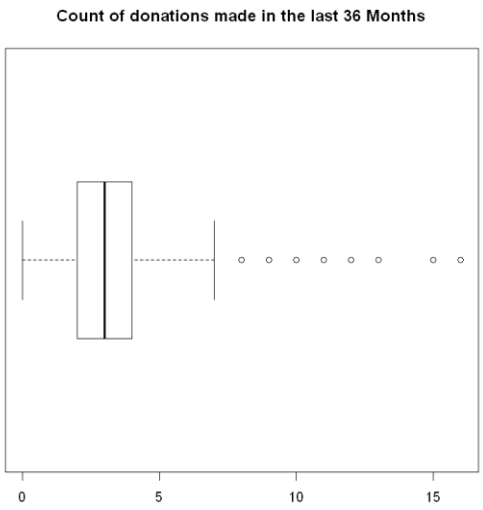
```
data %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
geom_histogram()
```

Distribution of data can be found out using various plots like histogram, density plots, bar plots,box plots. Here I have tried using a few of them




```
boxplot(data$DemAge, horizontal=TRUE, main="Donor's Age")
boxplot(data$GiftCnt36, horizontal=TRUE, main="Count of donations made in the last 36 Months")
```





7. Transform the numeric variables into their natural log values and scale [0 - 1] values.

```
data2<- log(data2)
head(data2)
```

TargetB	ID	TargetD	GiftCnt36	GiftCntAll	GiftCntCard36	GiftCntCardAll	GiftAvgLast	GiftAvg36	GiftAvgAll	...	PromCntAll	PromCntCard12	PromCnt
-Inf	9.614071	-Inf	0.6931472	1.386294	0.0000000	1.098612	2.833213	2.602690	2.224624	...	3.258097	1.0986123	2
-Inf	8.747352	-Inf	0.0000000	2.079442	-Inf	1.098612	2.995732	2.995732	2.765060	...	4.369448	1.6094379	1
0	10.738785	1.386294	1.7917595	3.713572	1.0986123	2.995732	1.791759	1.642873	1.316408	...	3.931826	1.6094379	2
0	12.133163	2.302585	1.0986123	2.484907	1.0986123	2.079442	2.302585	2.159869	2.140066	...	3.784190	0.6931472	1
-Inf	10.296779	-Inf	0.0000000	0.000000	0.0000000	0.000000	2.995732	2.995732	2.995732	...	2.564949	1.3862944	1
0	11.631881	2.397895	1.0986123	2.397895	0.6931472	2.197225	2.397895	2.335052	2.112635	...	3.806662	1.6094379	2

```
scale <- function(x) {
return ((x - min(x)) / (max(x) - min(x)))
}
```

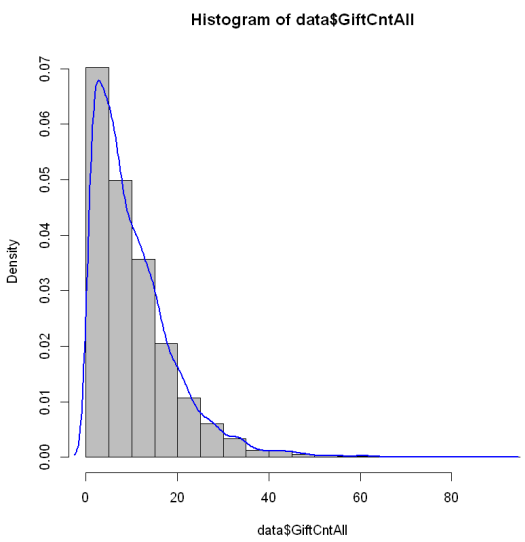
```
sapply(data2, function(x) scale(x))
```

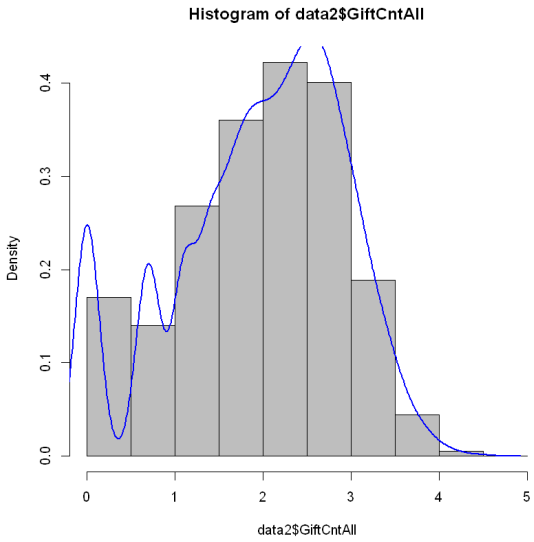
8. Check whether the numeric variables follow normality conditions.

After scaling the data, I have compared the the distributions of few variables before and after normalisation using Histogram, density plot and QQ Plot.

```
hist(data$GiftCntAll, prob=TRUE, col="grey")
lines(density(data$GiftCntAll), col="blue", lwd=2)
```

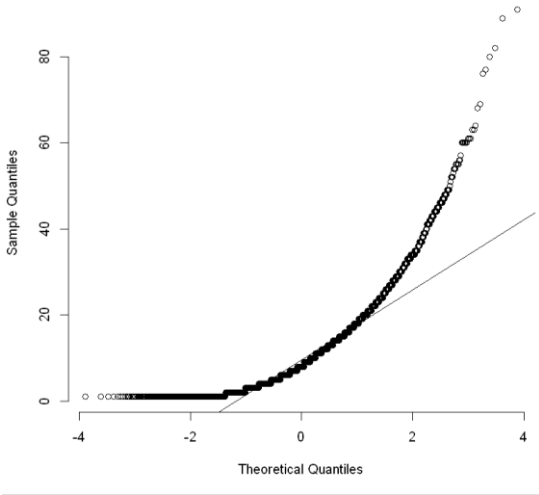
```
hist(data2$GiftCntAll, prob=TRUE, col="grey")
lines(density(data2$GiftCntAll), col="blue", lwd=2)
```

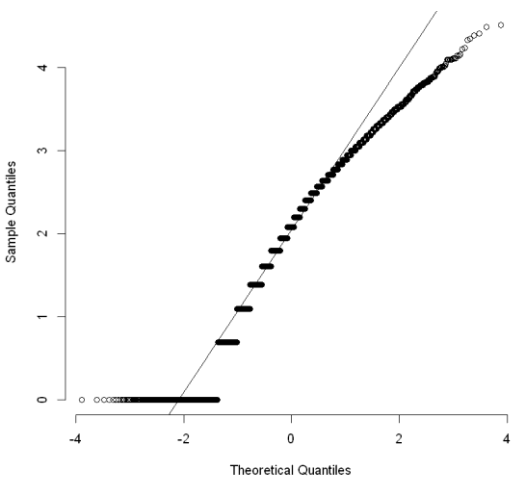





```
qqnorm(data$GiftCntAll, pch = 1, frame = FALSE)
qqline(data$GiftCntAll)
```

```
qqnorm(data2$GiftCntAll, pch = 1, frame = FALSE)
qqline(data2$GiftCntAll)
```



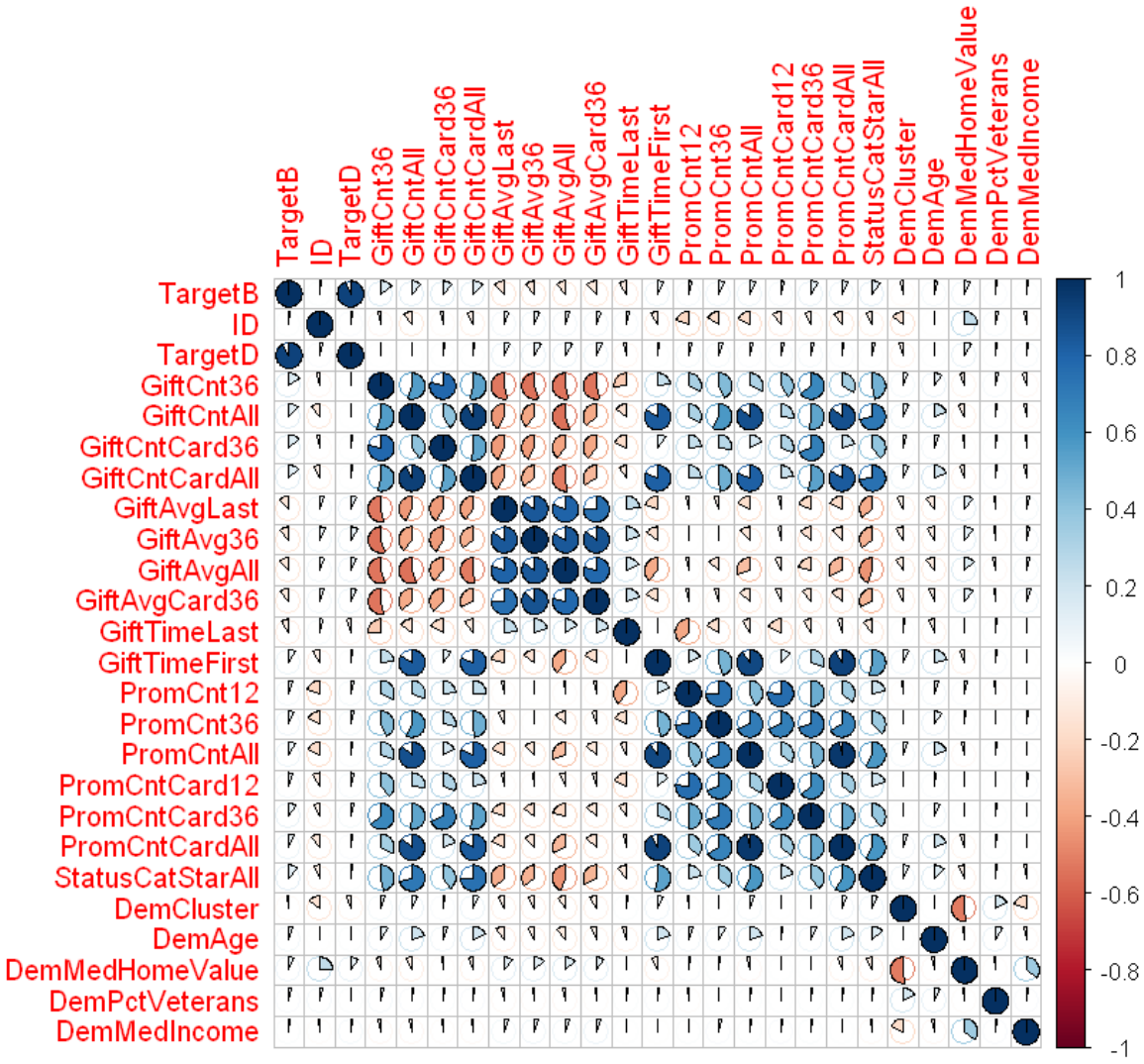


9. Find the correlation matrix for all the variables in the dataset and plot the graph of the correlation matrix.

```
cordata.cor = cor(data2, method = c("spearman"))
corrplot(cordata.cor,method='pie')
```

The graph below is a plot of the correlation matrix of all the numeric variables in the dataset;

Positive correlations are displayed in blue and negative correlations in red color. Color intensity and Pie charts are proportional to the correlation coefficients.



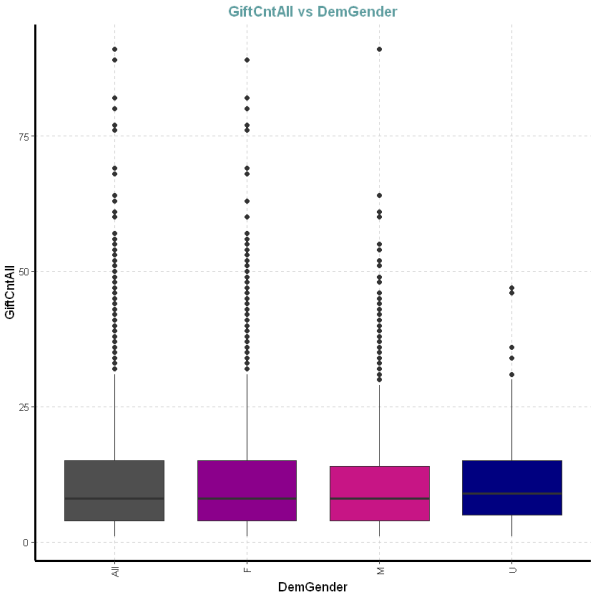
10. From the given dataset partition the data into 70-15-15 divisions so to construct the training, validation and test datasets.

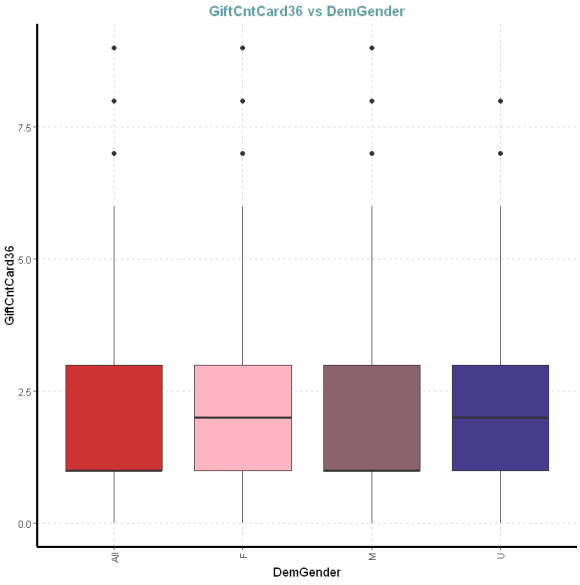
```
df = sort(sample(nrow(data2), nrow(data2)*.7))
train<-data[df,]
test<-data[-df,]
dfv = sort(sample(nrow(test), nrow(test)*.5))
test<-data[dfv,]
valid<-data[-dfv,]
```

11. **Any additional ways of Data Exploration & Visualization will be highly appreciated.**

```
ExpData(data, type=1)
```

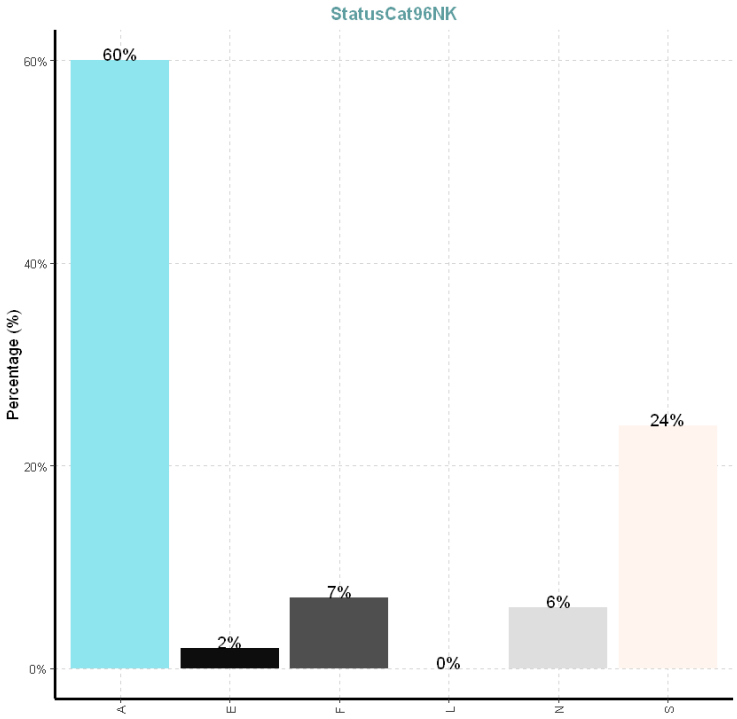
Descriptions	Value
Sample size (nrow)	9686
No. of variables (ncol)	28
No. of numeric/interger variables	25
No. of factor variables	3
No. of text variables	0
No. of logical variables	0
No. of identifier variables	1
No. of date variables	0
No. of zero variance variables (uniform)	0
%. of variables having complete cases	89.29% (25)
%. of variables having >0% and <50% missing cases	7.14% (2)
%. of variables having >=50% and <90% missing cases	3.57% (1)
%. of variables having >=90% missing cases	0% (0)

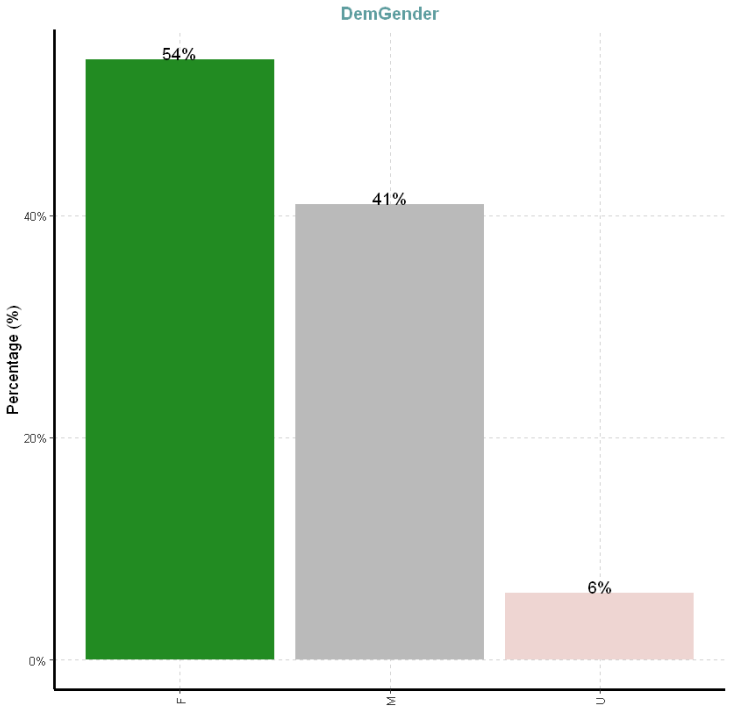


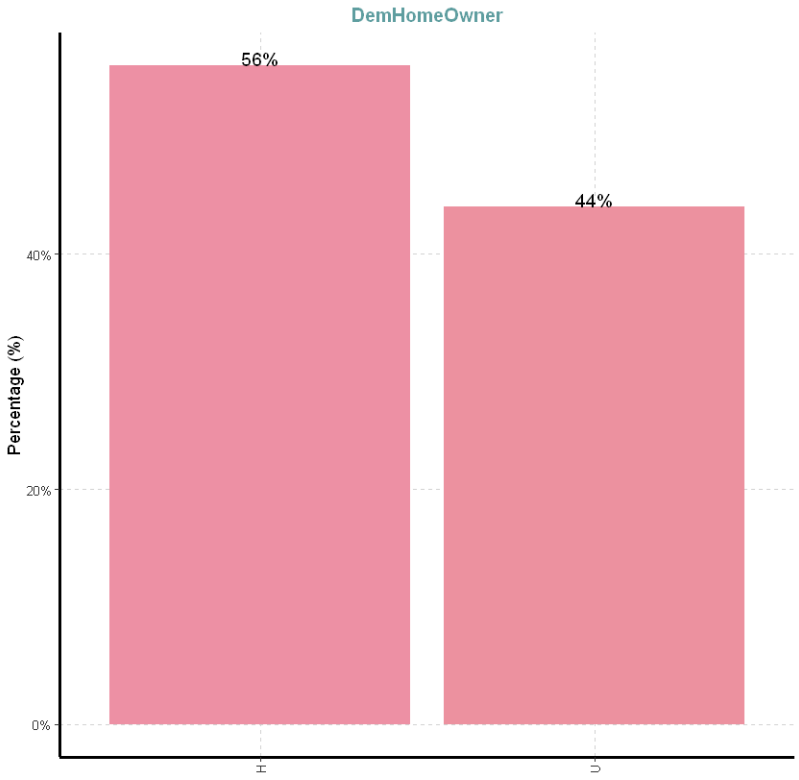


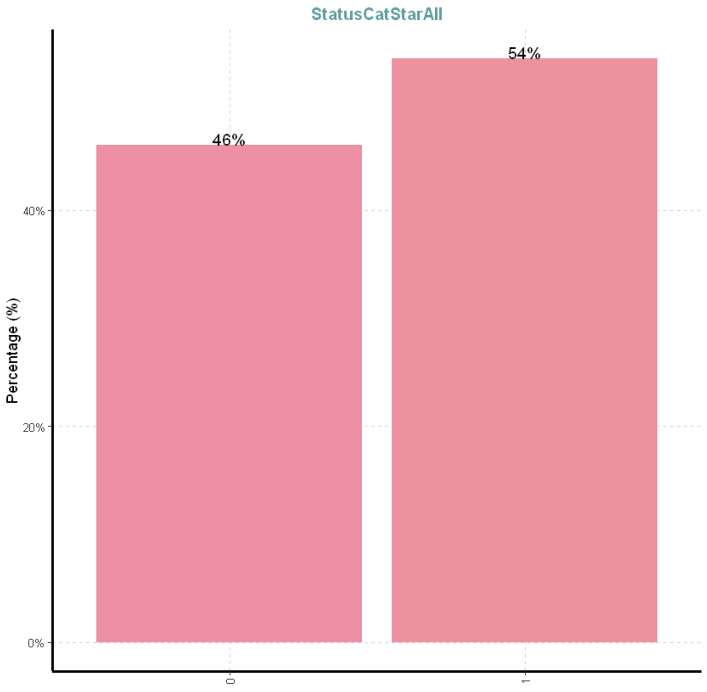
Above are the box plots, which are the distributions of variables , namely, GiftCntAll and GiftCntCard, according to the Gender category.

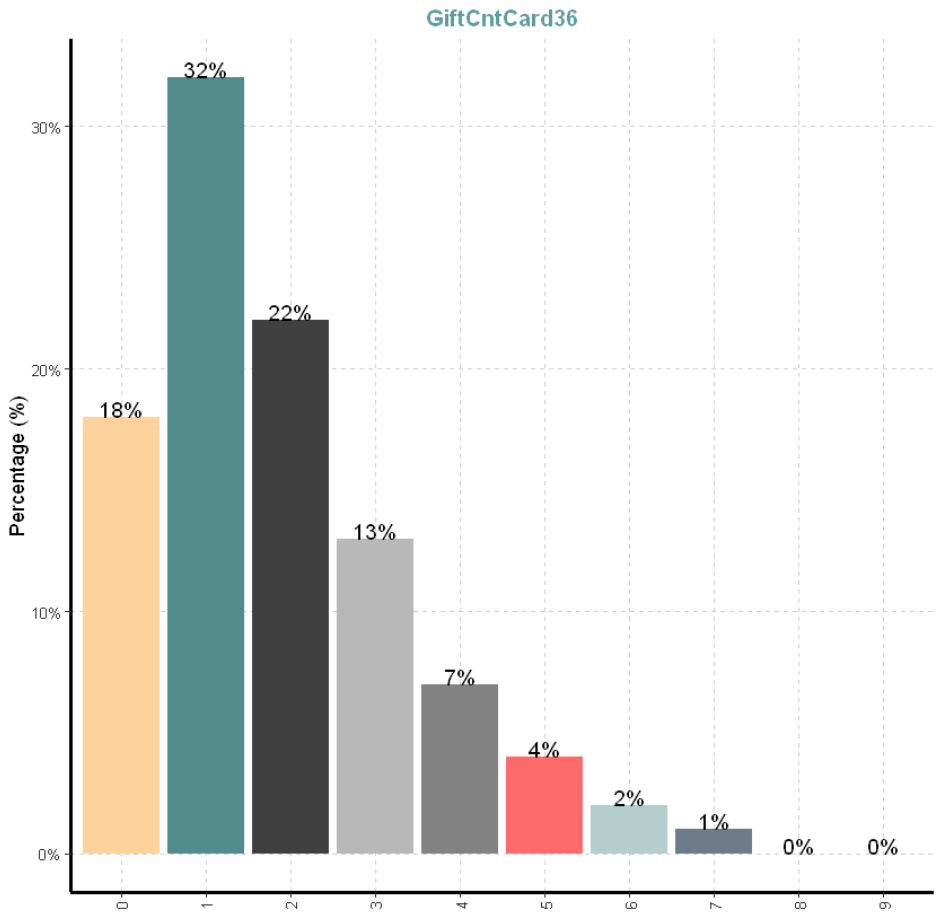
As observed there are more female Donors.


StatusCat96NK
Percentage (%)
60%
40%
20%
0%
60%
2%
7%
0%
6%
24%
A
E
F
L
N
S

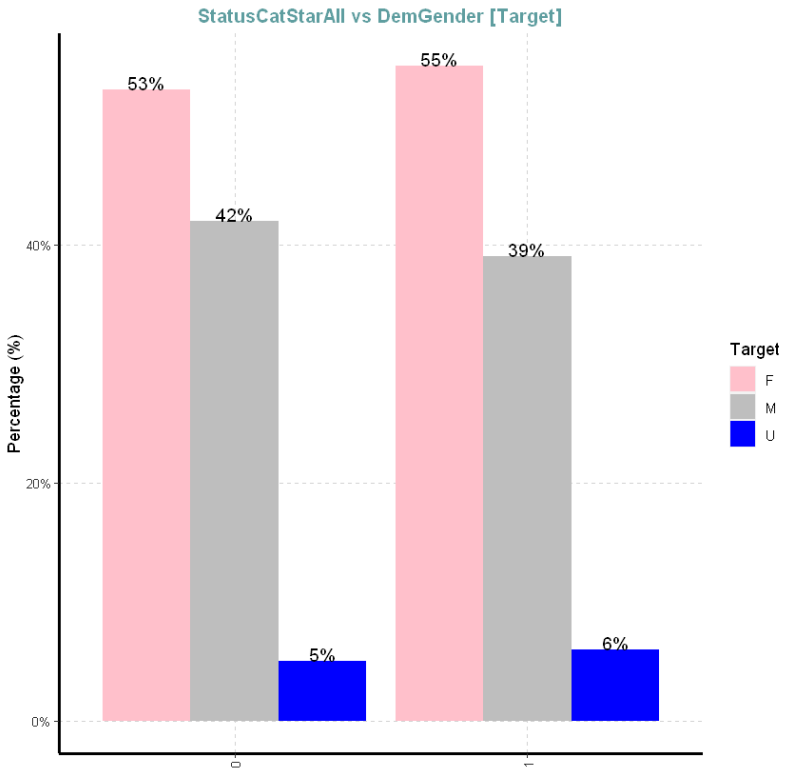

DemGender
Percentage (%)
40%
20%
0%
54%
41%
6%
F
M
U


DemHomeOwner
56%
44%
40%
20%
0%
Percentage (%)
H
U


StatusCatStarAll
54%
46%
40%
20%
0%
Percentage (%)
0
1

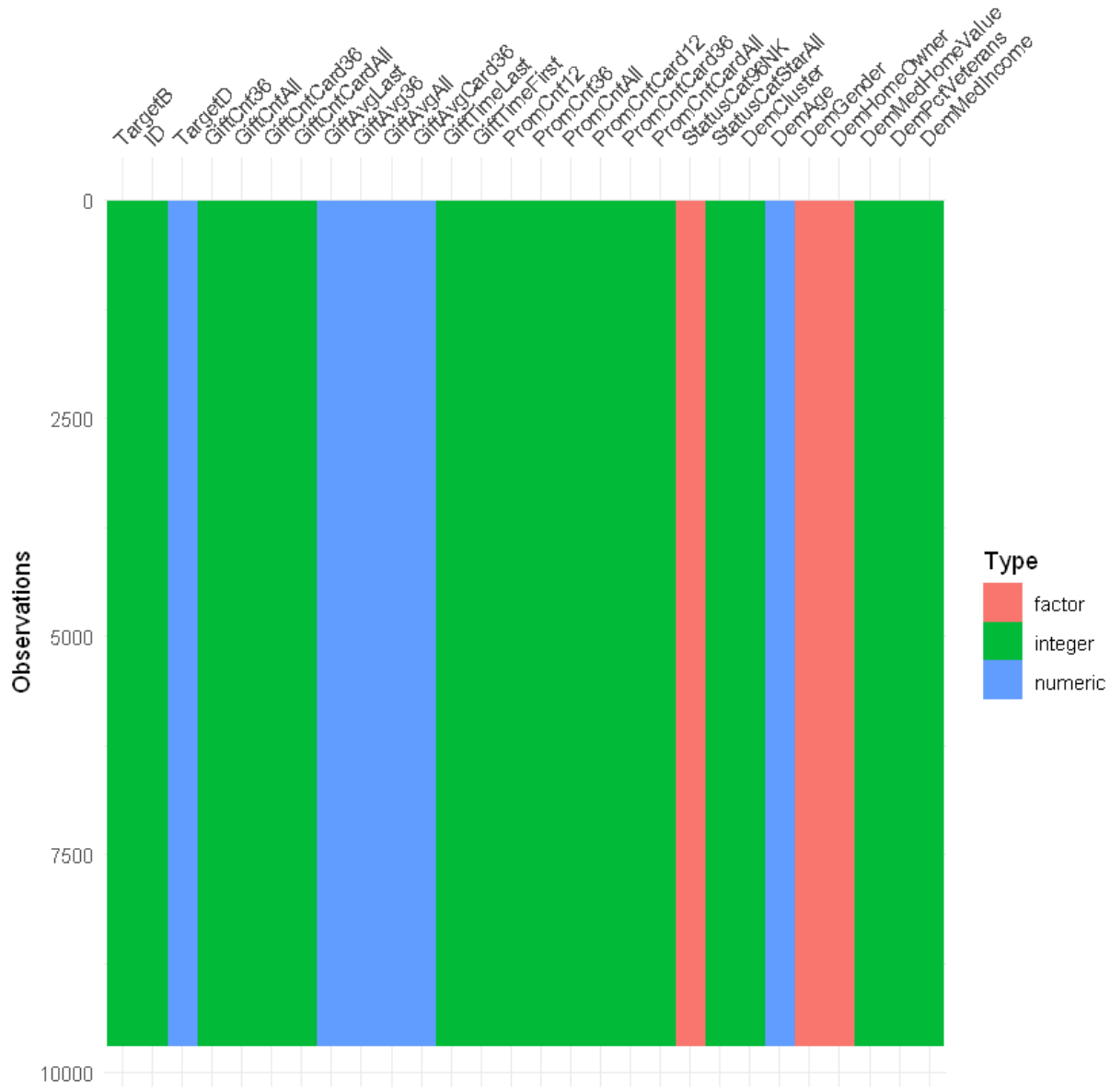


Above graphs depict a simple Distribution of Categorical variables in our dataset.

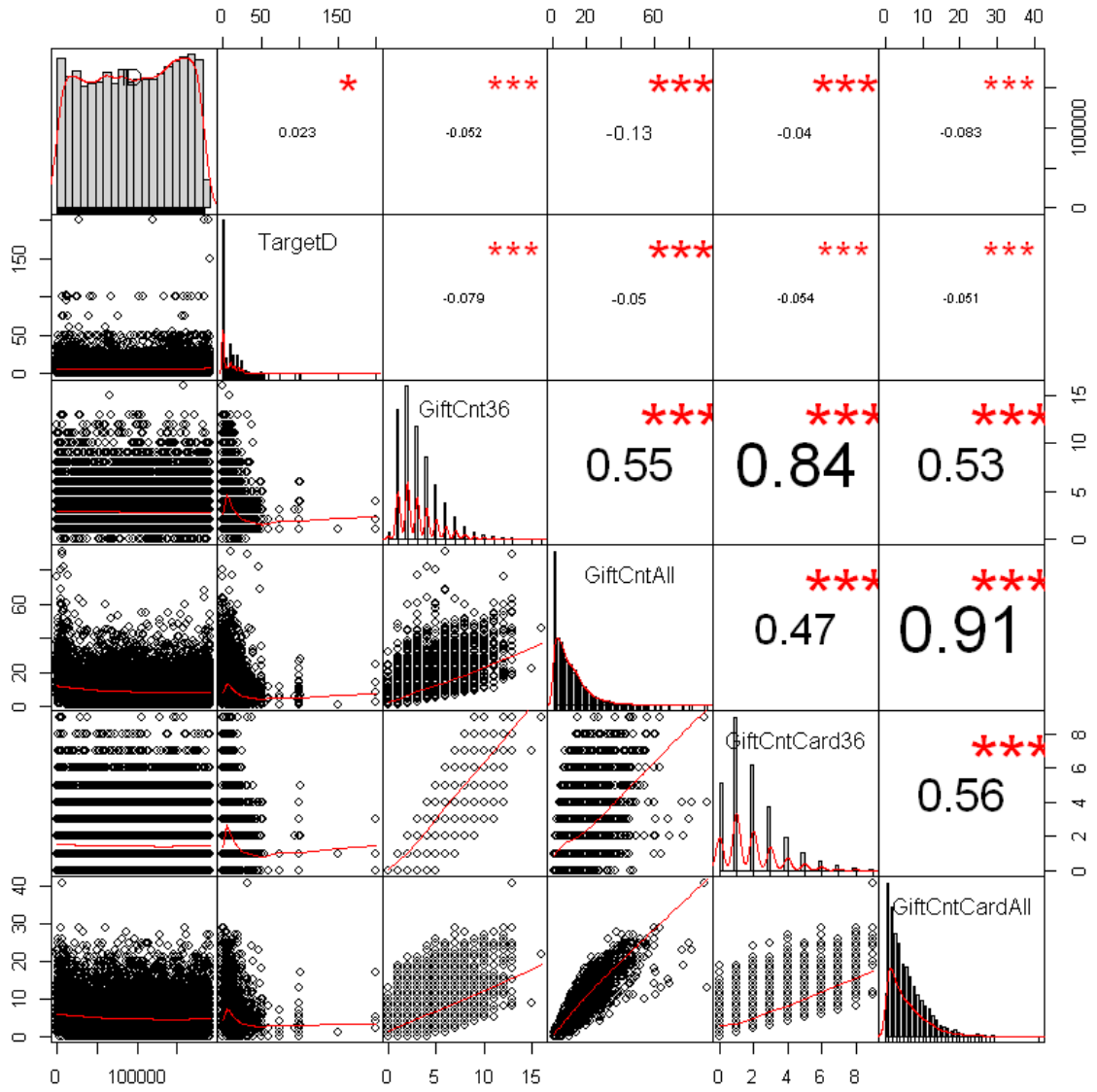


This is again a percentage distribution of StatusCatStarAll according to the gender roles;

Wherein Females are at the top.



The above Plot explains more about the dataset, its variables and the types of variables and the count of observations



This plot is similar to the Correlation Plot, But it gives more information about the:

- Numerical correlations (Pearson's coefficient) between numerical variables in the dataset, with larger sources for larger correlations.
- A mini-scatterplot between each of the pairs of variables.
- A histogram and density plot of each variable

END