Question 2 - EDA of Junk Datasets

Anil Akyildirim

3/25/2020

Introduction

Our client gave us two data sets for exploratory analysis. We will provide explanotory data analysis and provide insights and actions.

Data Colleciton

```r
#load libraries
library(ggplot2)
library(ggcorrplot)
library(statsr)
```

```
## Warning: package 'statsr' was built under R version 3.6.2

## Loading required package: BayesFactor

## Warning: package 'BayesFactor' was built under R version 3.6.2

## Loading required package: coda

## Loading required package: Matrix

## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richarddmorey
##
## Type BFManual() to open the manual.
## ************
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(dplyr)
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.6.2

## corrplot 0.84 loaded
```

```r
library(PerformanceAnalytics)
```

```
## Warning: package 'PerformanceAnalytics' was built under R version 3.6.2

## Loading required package: xts

## Warning: package 'xts' was built under R version 3.6.2

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last

##
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':
##
##     legend
```

Analysis of the first file

```r
# Load First dataset in R
df_1 <- read.table("junk1.txt", header = TRUE, sep = "", dec = ".")
head(df_1)
```

```
##           a         b class
## 1 1.6204214 3.0036241     1
## 2 1.4340220 0.7852487     1
## 3 2.4766615 0.9367761     1
## 4 0.5283093 0.1196222     1
## 5 1.0054081 0.7872866     1
## 6 1.1032636 0.7330594     1
```

```r
#Look at the structure
str(df_1)
```

```
## 'data.frame':    100 obs. of  3 variables:
##  $ a    : num  1.62 1.434 2.477 0.528 1.005 ...
##  $ b    : num  3.004 0.785 0.937 0.12 0.787 ...
##  $ class: int  1 1 1 1 1 1 1 1 1 1 ...
```

```r
# look at descriptive statistics
metastats_1 <- data.frame(describe(df_1))
metastats_1 <- tibble::rownames_to_column(metastats_1, "STATS")
metastats_1["pct_missing"] <- round(metastats_1["n"]/100, 3)
head(metastats_1)
```

```
##   STATS vars   n       mean        sd      median     trimmed      mad
## 1     a    1 100 0.04757654 1.2677402 -0.04753700  0.03368056 1.465745
## 2     b    2 100 0.01324258 1.4460671 -0.07455614 -0.01777082 1.485025
## 3 class    3 100 1.50000000 0.5025189  1.50000000  1.50000000 0.741300
##         min      max    range      skew   kurtosis         se pct_missing
## 1 -2.298540 3.006037 5.304577 0.1285786 -0.7976884 0.12677402           1
## 2 -3.171737 3.102297 6.274034 0.1178428 -0.5289542 0.14460671           1
## 3  1.000000 2.000000 1.000000 0.0000000 -2.0199000 0.05025189           1
```

```r
#look for missing values
missing_values_1 <- metastats_1 %>%
  filter(pct_missing < 1) %>%
  dplyr::select(STATS, pct_missing) %>%
  arrange(pct_missing)

missing_values_1
```

```
## [1] STATS       pct_missing
## <0 rows> (or 0-length row.names)
```

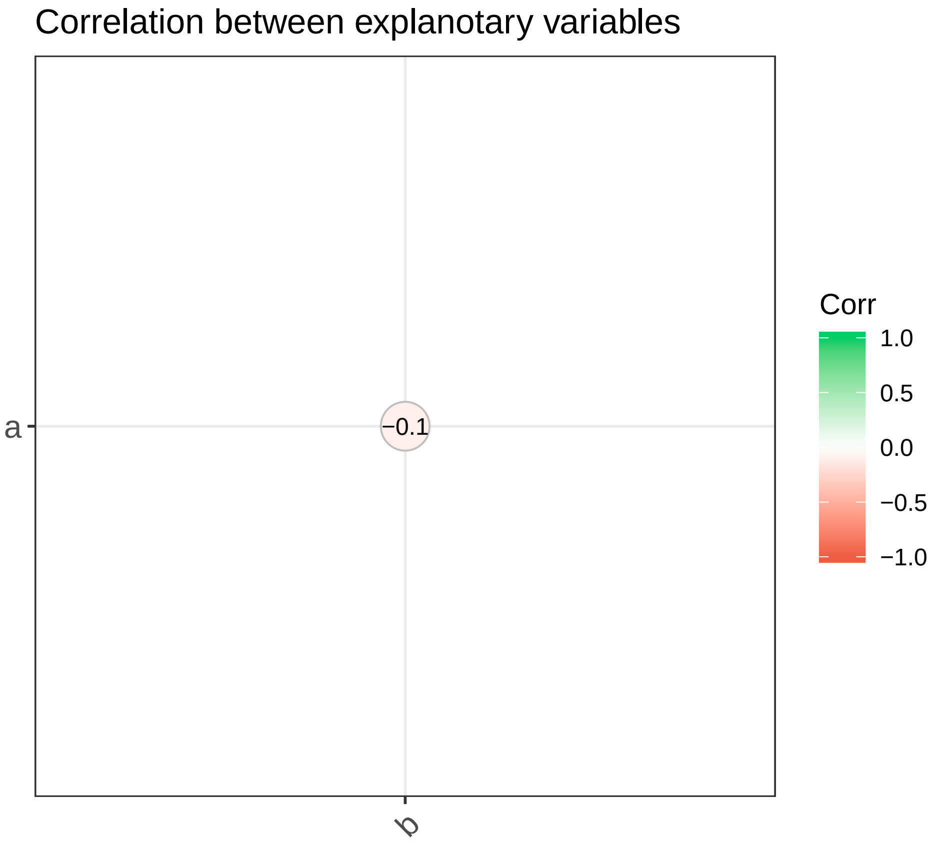
```r
unique(df_1$class)
```

```
## [1] 1 2
```

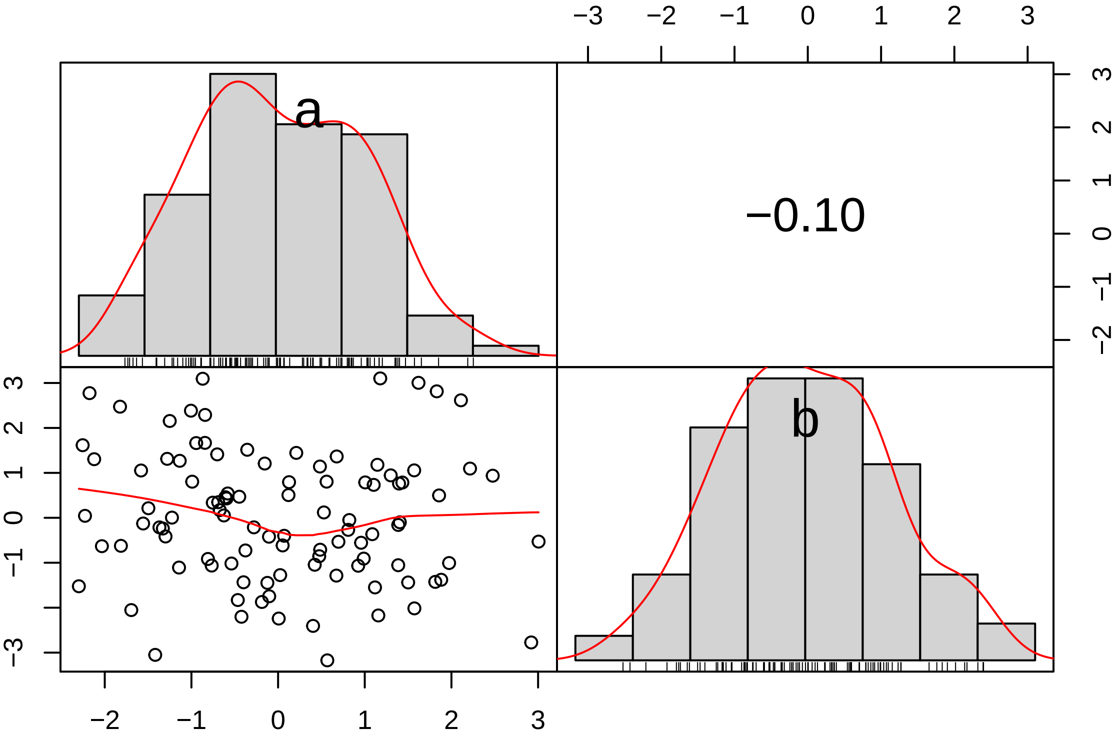
```r
df_class_1 <- df_1$class
df_exp_1 <- subset(df_1, select = -class)
head(df_exp_1)
```

```
##           a         b
## 1 1.6204214 3.0036241
## 2 1.4340220 0.7852487
## 3 2.4766615 0.9367761
## 4 0.5283093 0.1196222
## 5 1.0054081 0.7872866
## 6 1.1032636 0.7330594
```

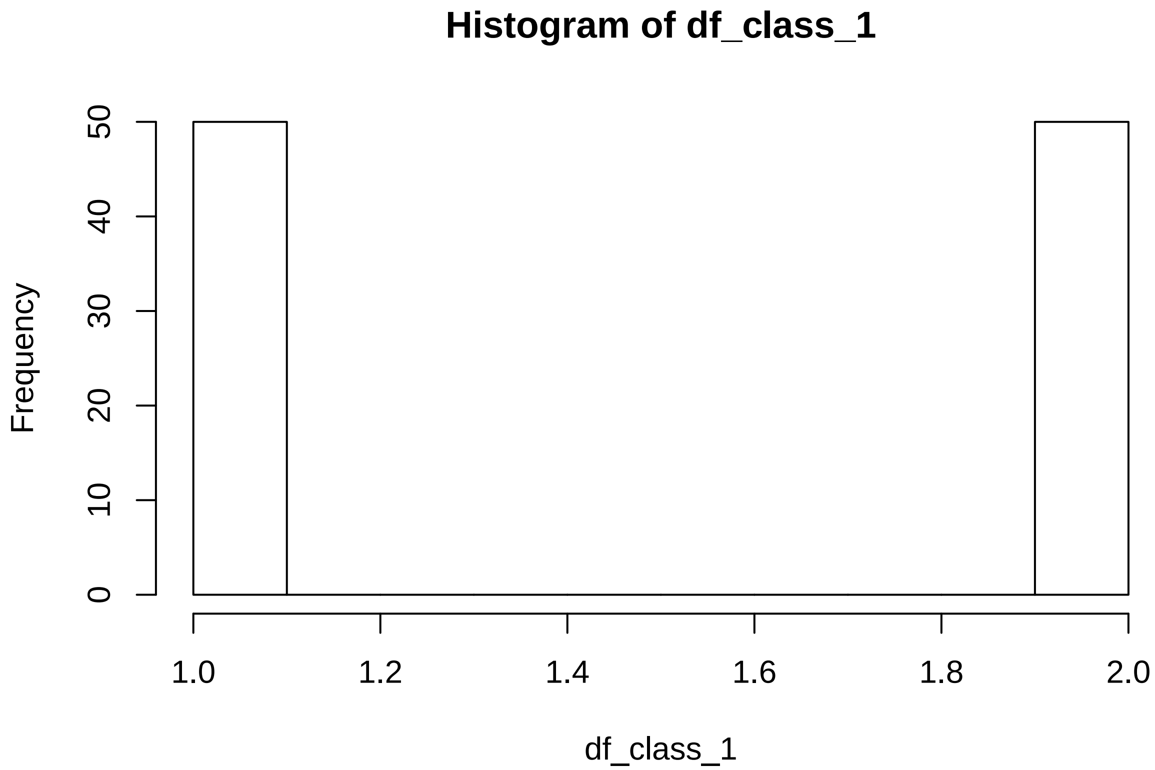
```r
# Look at correlation

corr_1 <- round(cor(df_exp_1), 1)

ggcorrplot(corr_1,
           type="lower",
           lab=TRUE,
           lab_size=3,
           method="circle",
           colors=c("tomato2", "white", "springgreen3"),
           title="Correlation between explanotary variables",
           ggtheme=theme_bw)
```



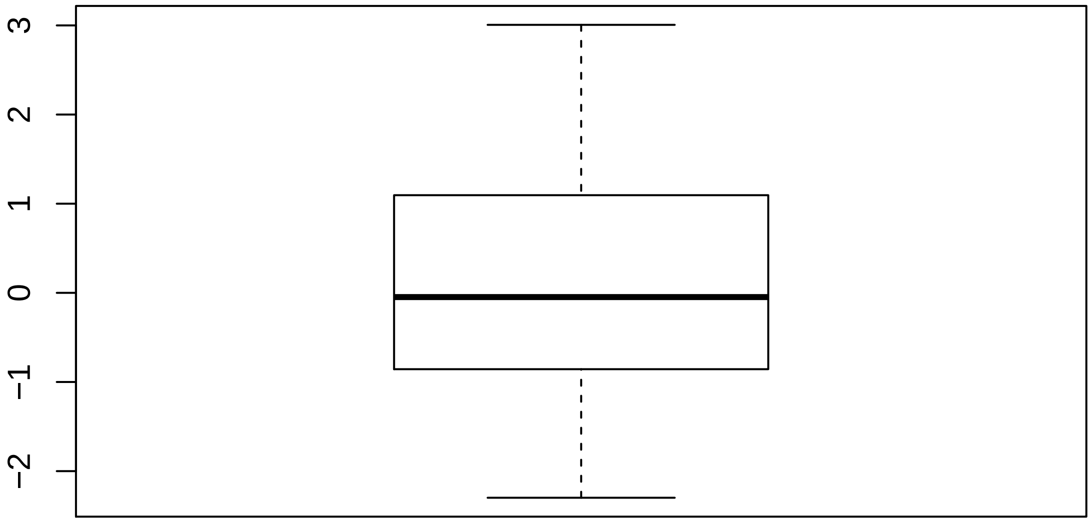
```
# look at correlation and distribution
chart.Correlation(df_exp_1, histogram=TRUE, pch=19)
```



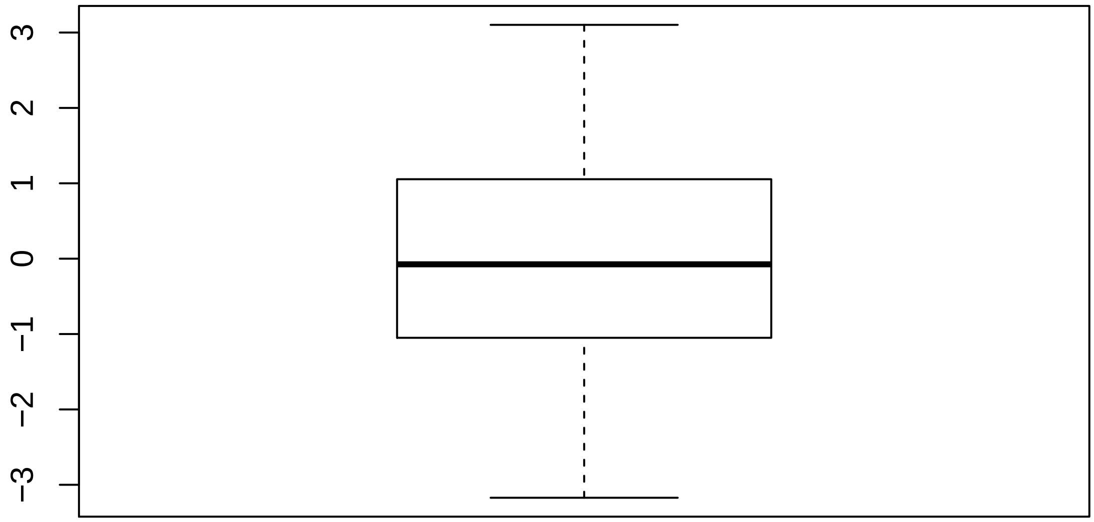
```
#look at distribution of the class
hist(df_class_1)
```



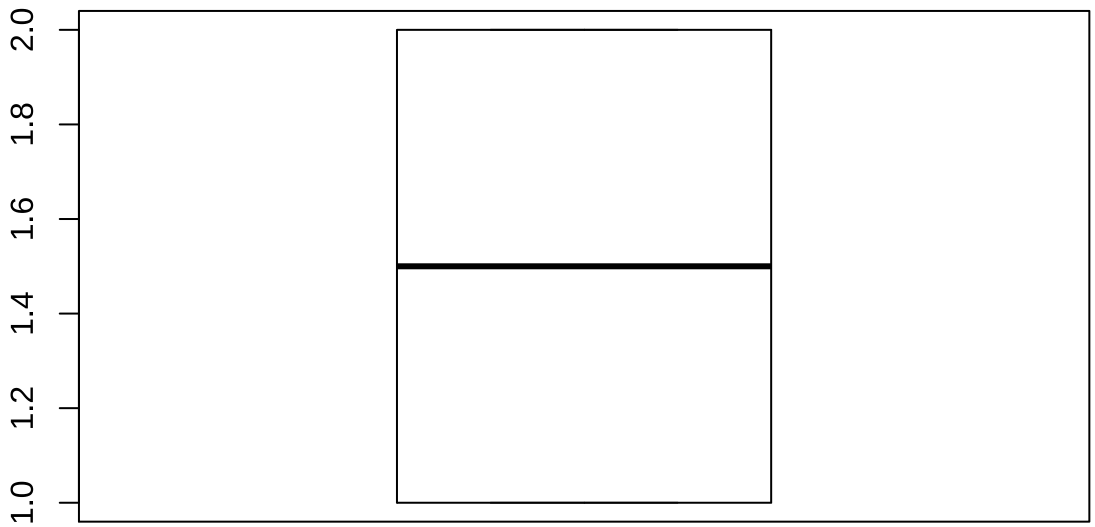
```r
# look at outliers for a
boxplot(df_exp_1$a)
```



```
# look at outliers for b
boxplot(df_exp_1$b)
```



```
boxplot(df_class_1)
```



Second Data File (csv)

```
# Load second data set in R
df_2 <- read.csv("junk2.csv", header = TRUE, sep = ",", dec = ".")
head(df_2)
```

```
##            a           b class
## 1  3.1886481  0.92917735     0
## 2  0.8224527  0.04760314     0
## 3  0.8147247  0.02910931     0
## 4 -1.5065362  3.13231360     0
## 5  0.4426887  2.84942822     0
## 6  0.8564405 -0.66143851     0
```

```
#Look at the structure
str(df_2)
```

```
## 'data.frame':    4000 obs. of  3 variables:
##  $ a    : num  3.189 0.822 0.815 -1.507 0.443 ...
##  $ b    : num  0.9292 0.0476 0.0291 3.1323 2.8494 ...
##  $ class: int  0 0 0 0 0 0 0 0 0 0 ...
```

```r
# look at descriptive statistics
metastats_2 <- data.frame(describe(df_2))
metastats_2 <- tibble::rownames_to_column(metastats_2, "STATS")
metastats_2["pct_missing"] <- round(metastats_2["n"]/4000, 3)
head(metastats_2)
```

```
##   STATS vars    n        mean        sd      median     trimmed      mad
## 1     a    1 4000 -0.05125461 1.2980758  0.08754417 -0.02350871 1.404696
## 2     b    2 4000  0.05624118 1.3143855 -0.08357556  0.02514754 1.388799
## 3 class    3 4000  0.06250000 0.2420917  0.00000000  0.00000000 0.000000
##         min      max    range       skew   kurtosis          se
## 1 -4.165048 4.626473 8.791521 -0.1723838 -0.3420351 0.020524380
## 2 -3.904721 4.310516 8.215237  0.2136746 -0.3462411 0.020782260
## 3  0.000000 1.000000 1.000000  3.6134290 11.0596342 0.003827806
##   pct_missing
## 1           1
## 2           1
## 3           1
```

```r
#look for missing values
missing_values_2 <- metastats_2 %>%
  filter(pct_missing < 1) %>%
  dplyr::select(STATS, pct_missing) %>%
  arrange(pct_missing)

missing_values_2
```

```
## [1] STATS       pct_missing
## <0 rows> (or 0-length row.names)
```

```r
unique(df_2$class)
```

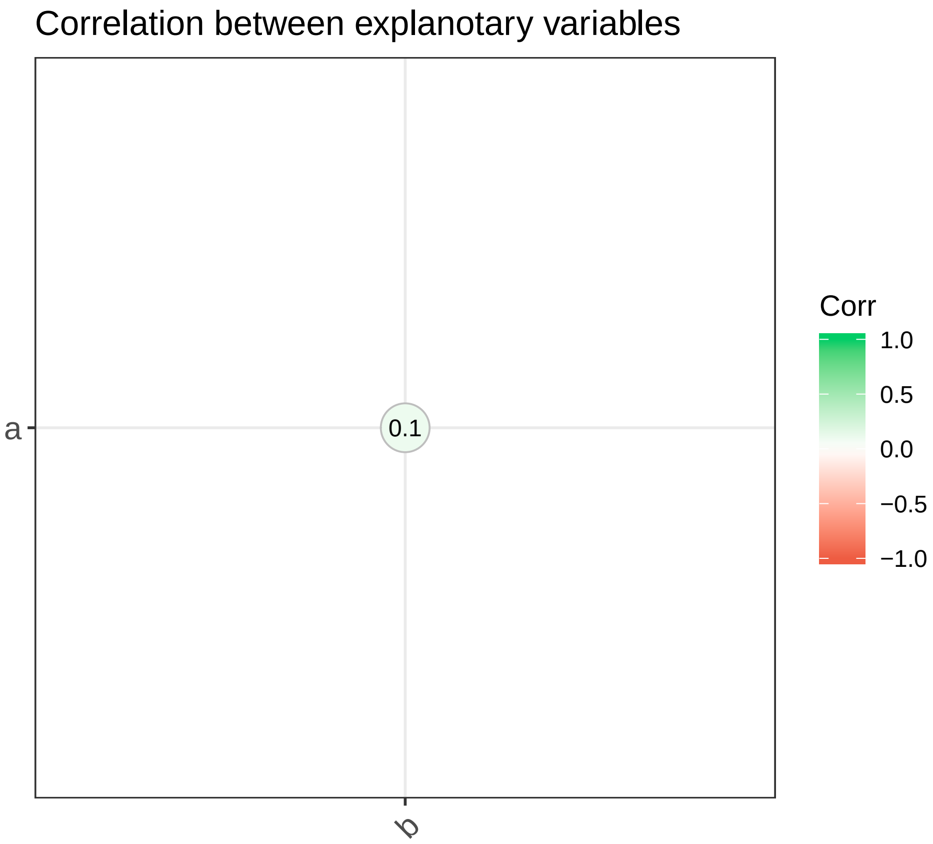
```
## [1] 0 1
```

```r
df_class_2 <- df_2$class
df_exp_2 <- subset(df_2, select = -class)
head(df_exp_2)
```

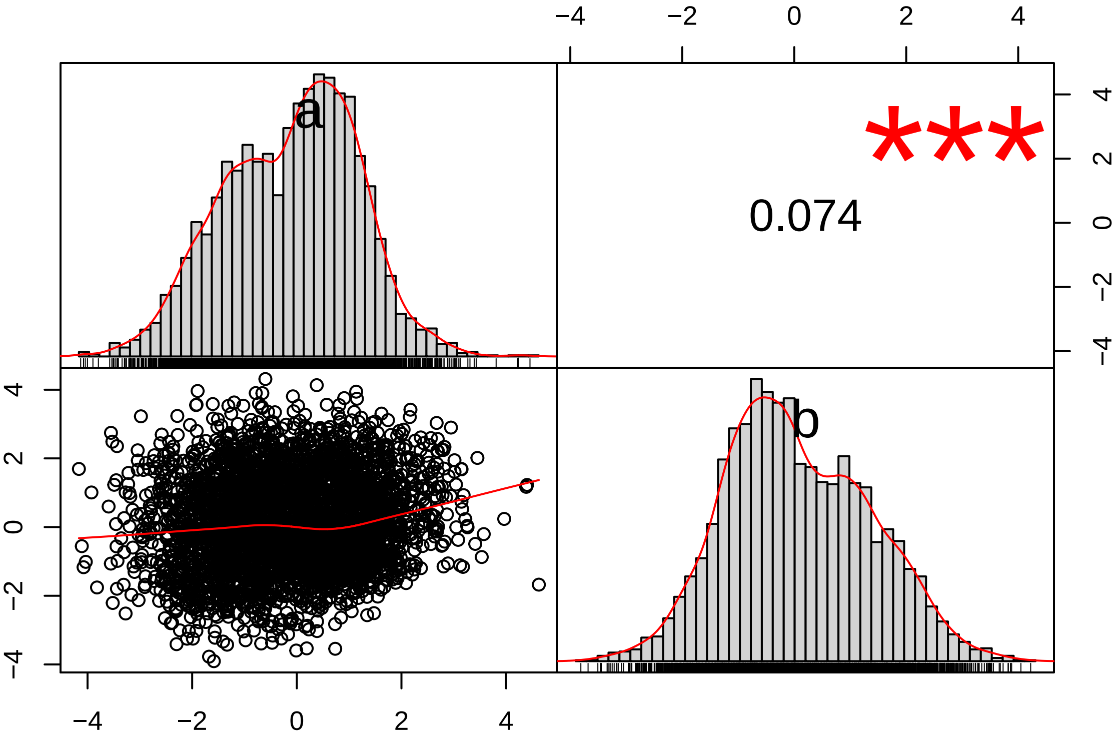
```
##            a           b
## 1  3.1886481  0.92917735
## 2  0.8224527  0.04760314
## 3  0.8147247  0.02910931
## 4 -1.5065362  3.13231360
## 5  0.4426887  2.84942822
## 6  0.8564405 -0.66143851
```

```r
# Look at correlation

corr_2 <- round(cor(df_exp_2), 1)
```

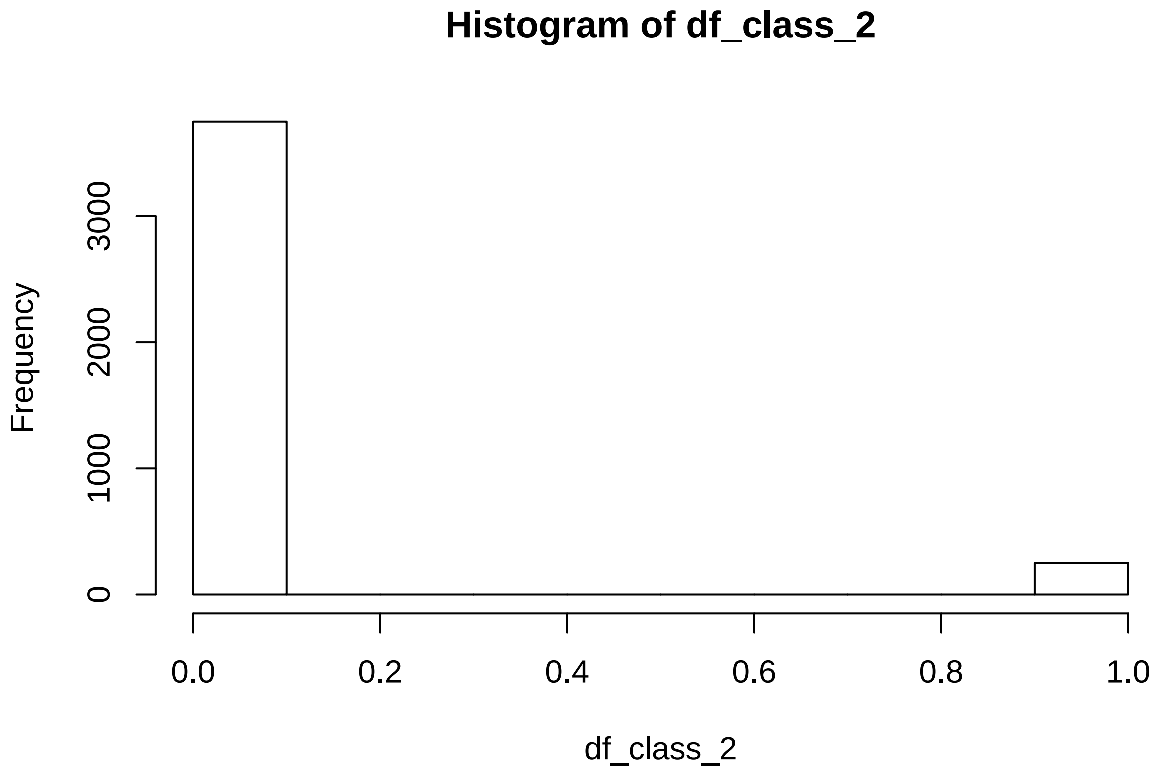
```r
ggcorrplot(corr_2,
           type="lower",
           lab=TRUE,
           lab_size=3,
           method="circle",
           colors=c("tomato2", "white", "springgreen3"),
           title="Correlation between explanotary variables",
           ggtheme=theme_bw)
```



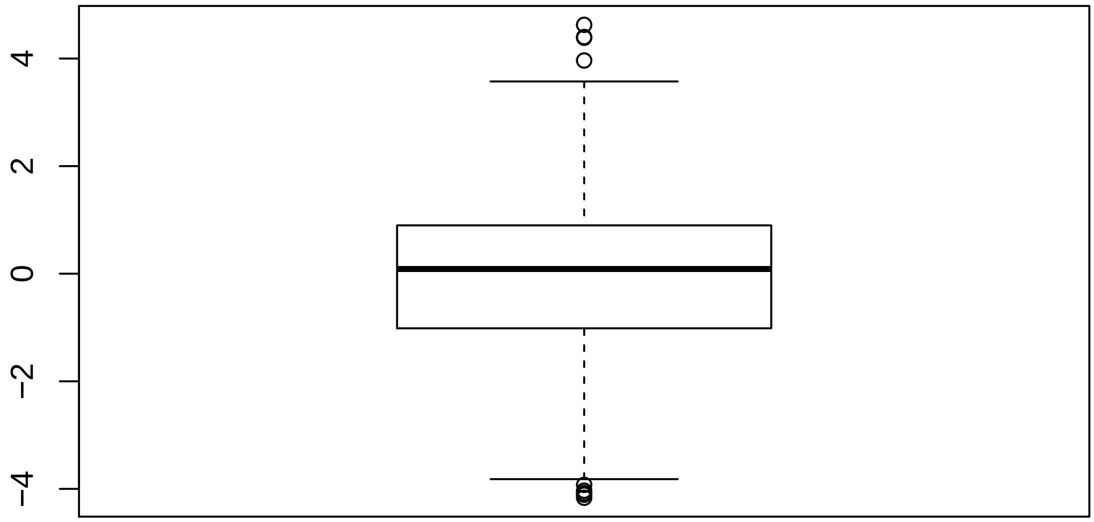
```r
# look at correlation and distribution
chart.Correlation(df_exp_2, histogram=TRUE, pch=19)
```



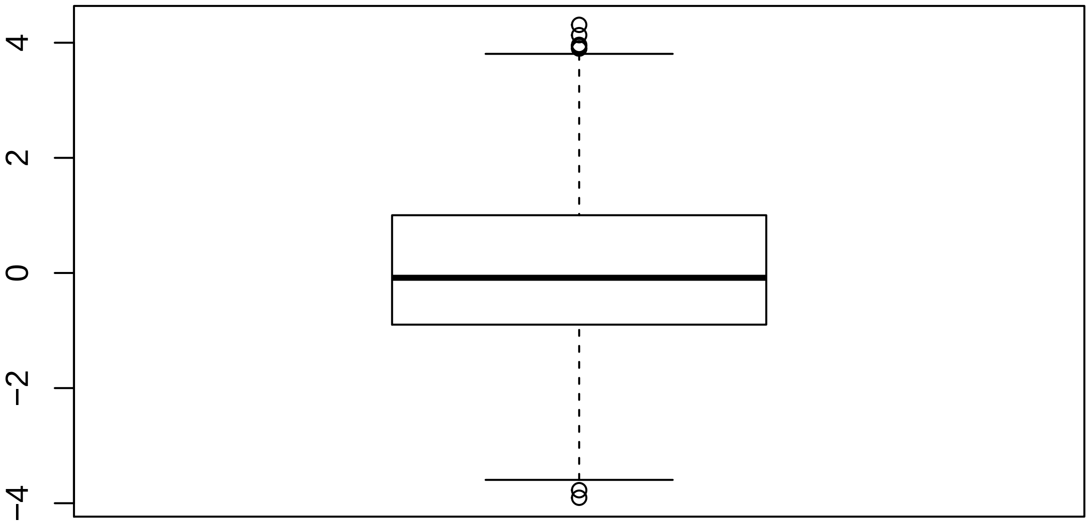
```r
#look at distribution of the class
hist(df_class_2)
```



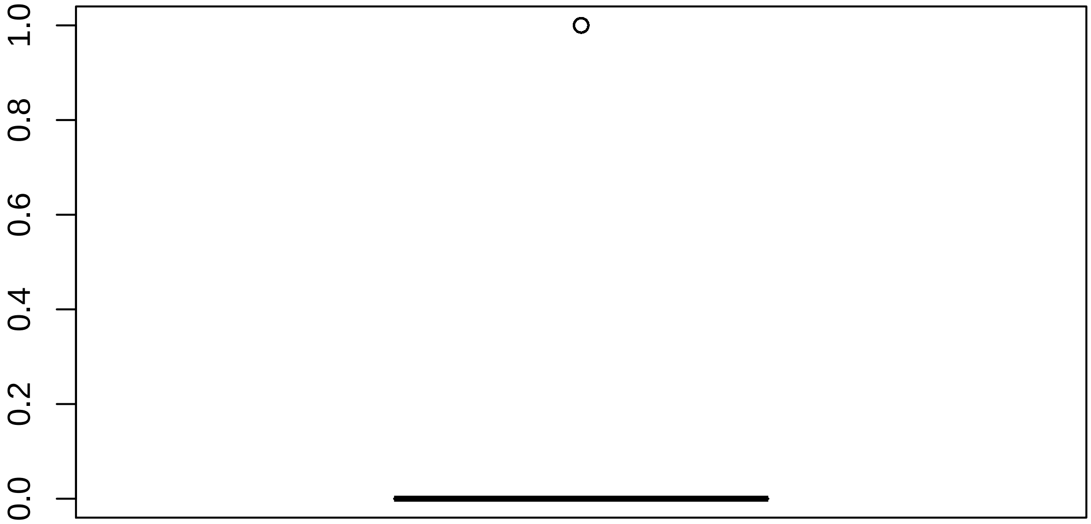
```r
# look at outliers for a
boxplot(df_exp_2$a)
```



```
# look at outliers for b
boxplot(df_exp_2$b)
```



```
boxplot(df_class_2)
```



Combination of both datasets

```
# Combine the two datasets

df <- rbind(df_1, df_2)
head(df)
```

```
##           a         b class
## 1 1.6204214 3.0036241     1
## 2 1.4340220 0.7852487     1
## 3 2.4766615 0.9367761     1
## 4 0.5283093 0.1196222     1
## 5 1.0054081 0.7872866     1
## 6 1.1032636 0.7330594     1
```

Descriptive Statistics

```
#Look at the structure
str(df)
```

```
## 'data.frame':    4100 obs. of  3 variables:
```

```
##  $ a    : num  1.62 1.434 2.477 0.528 1.005 ...
##  $ b    : num  3.004 0.785 0.937 0.12 0.787 ...
##  $ class: int  1 1 1 1 1 1 1 1 1 1 ...
```

```
# look at descriptive statistics
metastats <- data.frame(describe(df))
metastats <- tibble::rownames_to_column(metastats, "STATS")
metastats["pct_missing"] <- round(metastats["n"]/4100, 3)
head(metastats)
```

```
##   STATS vars    n        mean        sd      median    trimmed      mad
## 1     a    1 4100 -0.04884410 1.2972826  0.08446581 -0.0220621 1.408149
## 2     b    2 4100  0.05519243 1.3175778 -0.08357556  0.0242726 1.390850
## 3 class    3 4100  0.09756098 0.3353513  0.00000000  0.0000000 0.000000
##         min      max    range       skew   kurtosis          se
## 1 -4.165048 4.626473 8.791521 -0.1658534 -0.3486218 0.020260151
## 2 -3.904721 4.310516 8.215237  0.2102902 -0.3481641 0.020577108
## 3  0.000000 2.000000 2.000000  3.6298635 13.4744500 0.005237308
##   pct_missing
## 1           1
## 2           1
## 3           1
```

```
#look for missing values
missing_values <- metastats %>%
  filter(pct_missing < 1) %>%
  dplyr::select(STATS, pct_missing) %>%
  arrange(pct_missing)

missing_values
```

```
## [1] STATS       pct_missing
## <0 rows> (or 0-length row.names)
```

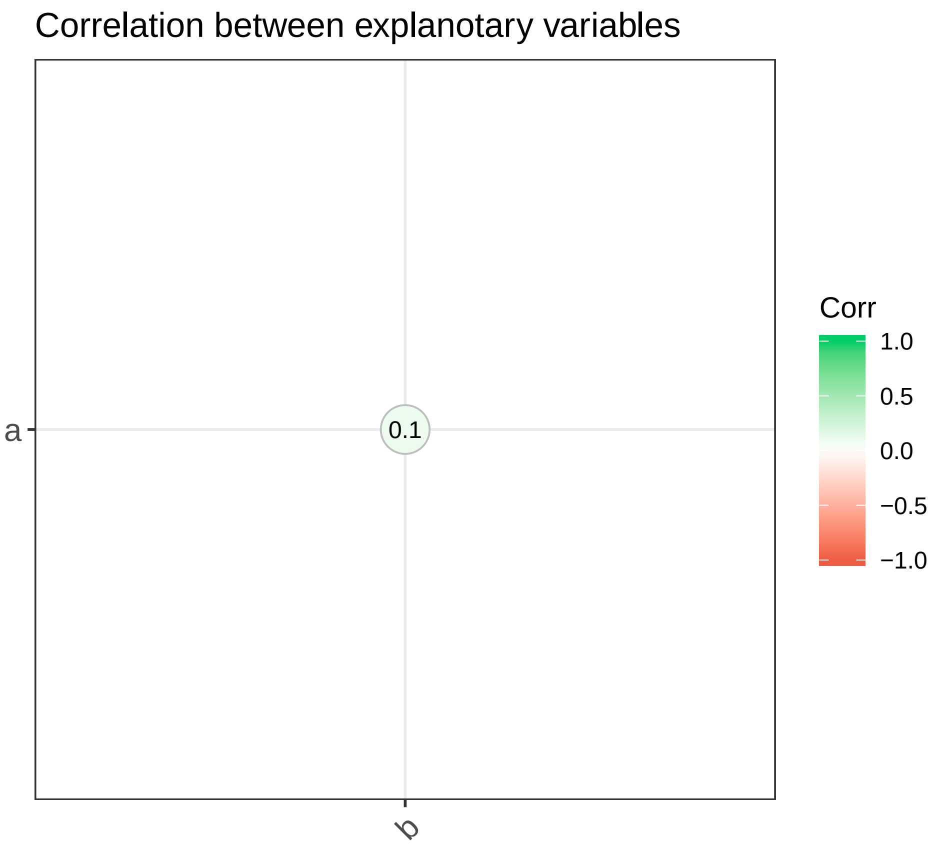
```
unique(df$class)
```

```
## [1] 1 2 0
```

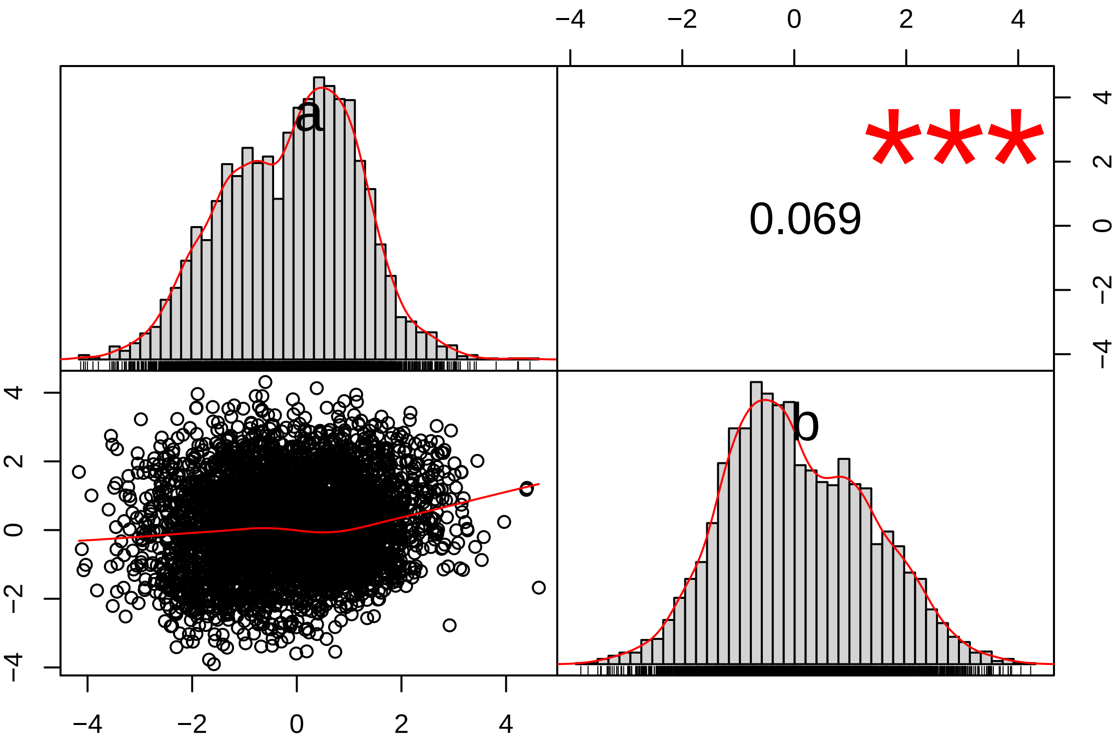
```
df_class <- df$class
df_exp <- subset(df, select = -class)
head(df_exp)
```

```
##           a         b
## 1 1.6204214 3.0036241
## 2 1.4340220 0.7852487
## 3 2.4766615 0.9367761
## 4 0.5283093 0.1196222
## 5 1.0054081 0.7872866
## 6 1.1032636 0.7330594
```

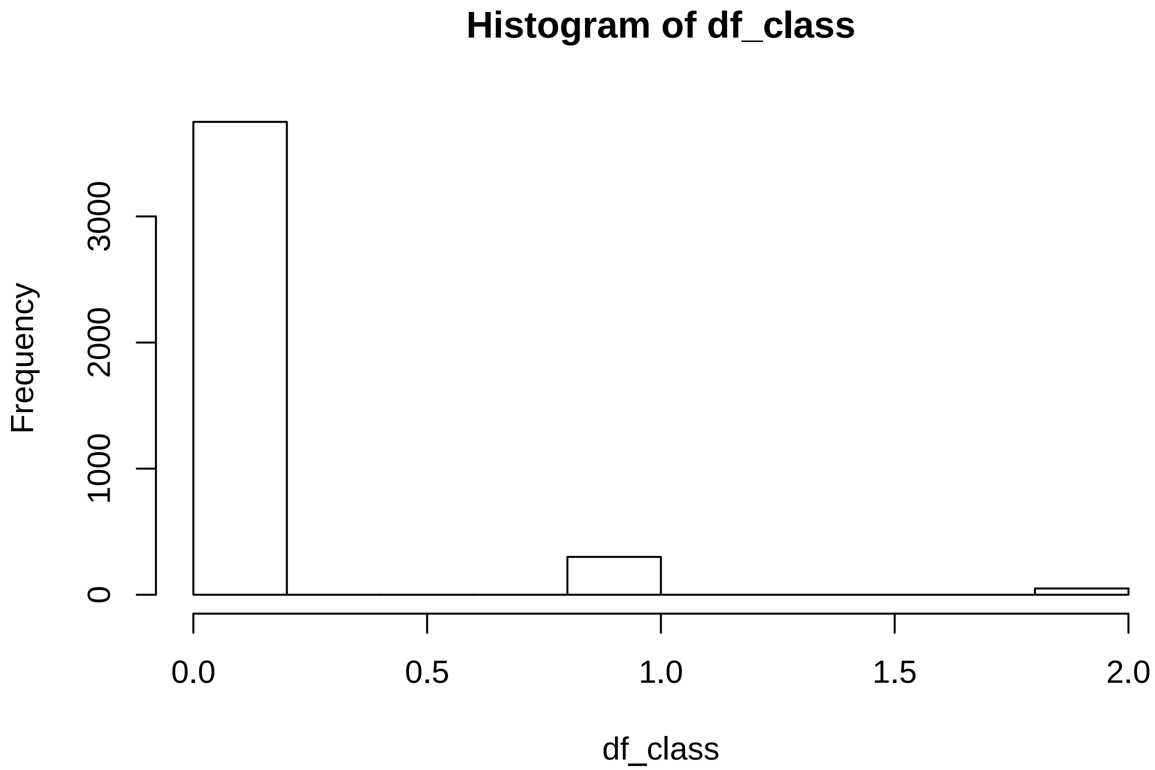
```
# Look at correlation

corr <- round(cor(df_exp), 1)

ggcorrplot(corr,
           type="lower",
           lab=TRUE,
           lab_size=3,
           method="circle",
           colors=c("tomato2", "white", "springgreen3"),
           title="Correlation between explanotary variables",
           ggtheme=theme_bw)
```



```
# look at correlation and distribution
chart.Correlation(df_exp, histogram=TRUE, pch=19)
```

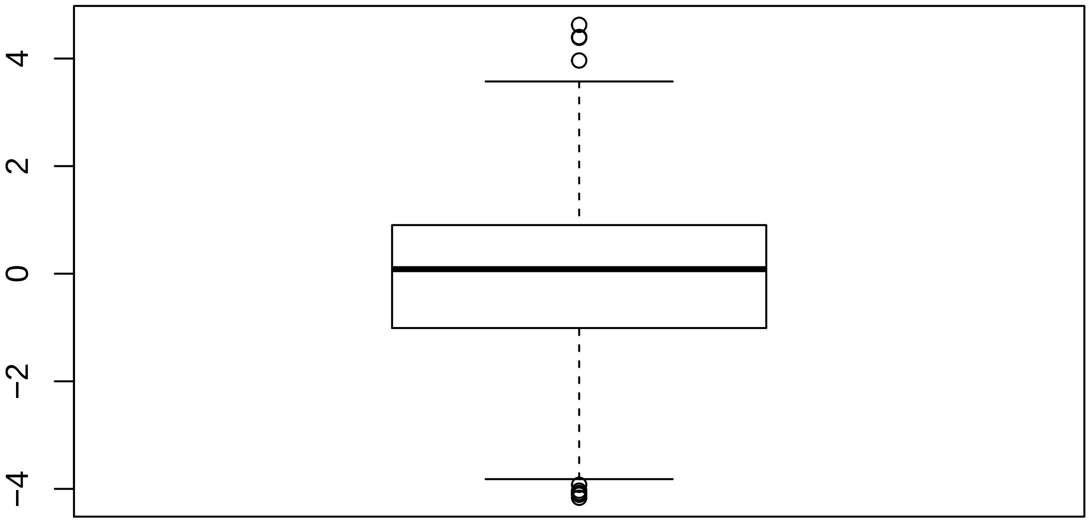


```
#look at distribution of the class
hist(df_class)
```

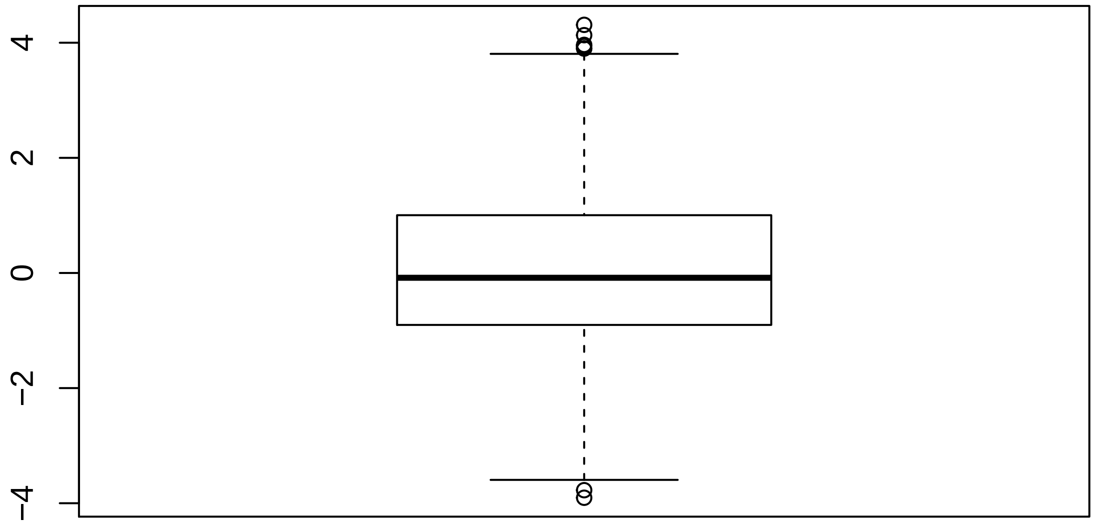


There is a very very weak positive correlation between a and b. a and b distribution is normal.

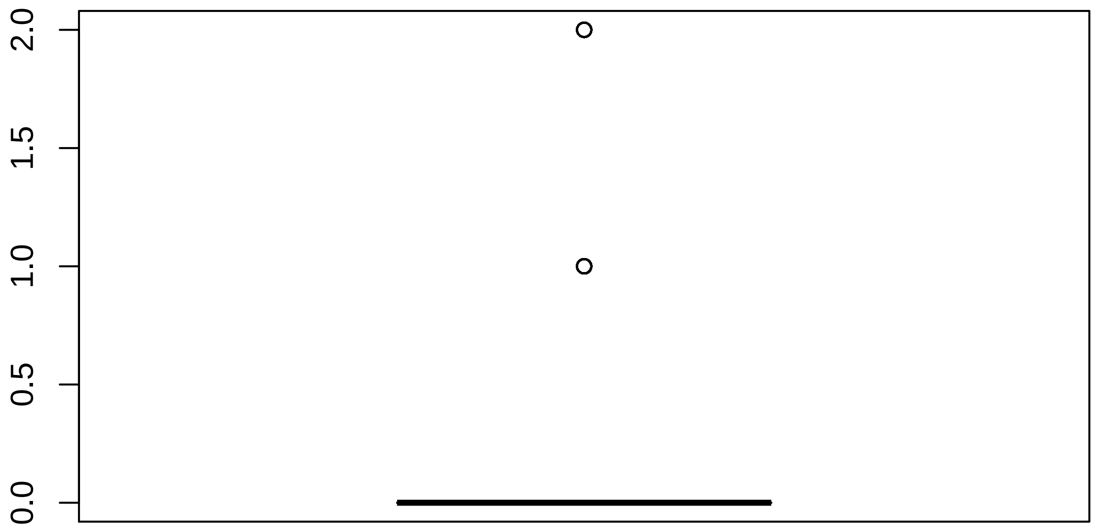
```
# look at outliers for a
boxplot(df_exp$a)
```



```
# look at outliers for b
boxplot(df_exp$b)
```



```
boxplot(df_class)
```



#Findings for All the Datasets.

In our first dataset, we have 100 observations with 3 variables. (a,b,c). We dont have any missing values. We have a and b columns as explanotary variables, c as the target variable (class variable). Mean of a is 0.047, mean of b is 0.013, min of a is -2.298, min of b is -3.1717. Max of a is 3.00, max of b is 3.1. There is a weak correlation between a and b. Both a an b are normally distributed. The c(class) variable has two values of 1 and 2 , unified distribution. There are no outliers in a,b,c.

In our second dataset, we have 4000 observations with 3 variables. (a,b,c). We have a and b columns as explanotary variables, c as the target variable(class variable). Mean of a is -0.051, mean of b is 0.056, min of a is -4.165, min of b is -3.904. Max of a is 4.626, max of b is 4.310. There are no missing values. We have 0 and 1 as classes in our second data set as c(class) variable. The correlation of a nd b is weak and same as the first data set. a and b has normal distribution. c distribution is not uniform. We can consider 1 as an outlier.

In the combination of the dataset. We have total of 4100 observations, the mean of a and b is similar to the first two datasets but the minimum and maximum values of a and b are around 4 and -4 range. These values above 4 and below -4 are outliers. The second dataset (junk.csv) is introducing another class(0) to the first dataset and the distribution of this class is much higher(due to the amount of observations) that the classes in the first dataset(junk.txt) becomes outliers.