HW1

Chanukya

9/25/2019

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
##
## 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
```

```
library(PerformanceAnalytics)
```

```
## Warning: package 'PerformanceAnalytics' was built under R version 3.5.2
```

```
## Loading required package: xts
```

```
## 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
```

1()

a) Yes

It is required to use data mining practices to forecast the future values based on previous data,it is not something which is database and retrieving the information.

b)No

since they vary drastically and they don’t depend on the previous data.

c) Yes

suppose we are able to derive relationship if particular students gets “A” in one subject then there is high probability on of getting “A” in operating system they we can predict the students who are likely to “A” in operating systems.

d)Yes

we do it by association method which is part data mining process

d)No

we are not forecasting anything, we are just retrieving the data which is present. we can do it by simple SQL query.

2)

Index Discrete or Continous quantitative or qualitative nominal or ordinal or interval or ration

a) Cellphone brands Discrete qualitative nominal

b) IQ levels continous quantitative Interval

C) The states of United Discrete qualitative nominal

D) The price of laptios Continous quantitative Interval

E) Pass or Fail Discrete qualitative nominal

3)

a) simple matching coefficient

x = c(1,0,1,1,0,1,0)

y = c(1,1,0,1,0,0,1)

result = same element at given index/total number of elements

result = 1+0+0+1+1+0+0/1+1+1+1+1+1+1

result = 3/7 = .42

b) Jaccard coefficient

M11= x =1 and y =1

M10= x=1 and y=0

M01= x=0 and y =1

result = M11/(M10+M01+M11)

result = (1+0+0+1+0+0+0)/(1+1+1+1+1+1) = 2/6 = 1/3 = .333

c) Cosine Correlation

sqrt = square root

x = (1,0,1,1,0,1,0), y = (1,1,0,1,0,0,1)

theta = cos^-1(|a|*|b|/sqrt(a^{2+b}2))

```
### theta = cos^(-1)((((1x1)+(0x1)+(1x0)+(1x1)+(0X0)+(1x0)+(1x1)))/sqrt((1^2+1
^2+1^2+1^2))++((1^2+1^2+1^2+1^2)))
```

theta = cos^-1(2/sqrt(8))

theta = cos^-1(2/2xsqrt(2))

theta = cos^-1(1/sqrt(2))

theta = 45 degree’s

d)Hamming Distance

x = (1,0,1,1,0,1,0), y = (1,1,0,1,0,0,1)

for hamming distance 1x0=1 0x1=0 0x0=1 1x1 =0

for hamming distance between x and y

x,y = ((1x1)+(0x1)+(1x0)+(1x1)+(0x0)+(1X0)+(0x1))

hamming distance between x,y = (1+1+1+1)

hamming distance between x,y = 4

2)

a.

(-2.05, 2.32), (-0.41, 5.36)

Euclidean distance

euclidean_distance = sqrt{(x2-x1)^2 - (y2-y1)^2}

euclidean_distance = sqrt{(-2.05-(-0.41))^2 + (2.32-5.36)^2}

euclidean_distance = sqrt(10.88)

euclidean_distance = 3.29

4

```
###4

#1

getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmean <- function(v) {
  uniqv <- sum(v)
  l <- length(v)
  return(uniqv/l)
}

#getmean(cr$X01)
cr <- read.csv("/Users/chanukya/Documents/GitHub/DataMining/HW1/crx.data",header=F)
summary(cr)
```

```
##  V1            V2            V3         V4       V5            V6
##  ?: 12   ?      : 12   Min.   : 0.000   ?:  6   ? :  6   c      :137
##  a:210   22.67  :  9   1st Qu.: 1.000   l:  2   g :519   q      : 78
##  b:468   20.42  :  7   Median : 2.750   u:519   gg:  2   w      : 64
##          18.83  :  6   Mean   : 4.759   y:163   p :163   i      : 59
##          19.17  :  6   3rd Qu.: 7.207                    aa     : 54
##          20.67  :  6   Max.   :28.000                    ff     : 53
##          (Other):644                                     (Other):245
##        V7            V8          V9      V10          V11       V12
##  v      :399   Min.   : 0.000   f:329   f:395   Min.   : 0.0   f:374
##  h      :138   1st Qu.: 0.165   t:361   t:295   1st Qu.: 0.0   t:316
##  bb     : 59   Median : 1.000                   Median : 0.0
##  ff     : 57   Mean   : 2.223                   Mean   : 2.4
##  ?      :  9   3rd Qu.: 2.625                   3rd Qu.: 3.0
##  j      :  8   Max.   :28.500                   Max.   :67.0
##  (Other): 20
##  V13          V14           V15           V16
##  g:625   00000  :132   Min.   :     0.0   -:383
##  p:  8   00120  : 35   1st Qu.:     0.0   +:307
##  s: 57   00200  : 35   Median :     5.0
##          00160  : 34   Mean   :  1017.4
##          00080  : 30   3rd Qu.:   395.5
##          00100  : 30   Max.   :100000.0
##          (Other):394
```

```
final <- function(cr){
cr$V1 <- as.factor(cr$V1)
cr$V2 <- as.numeric(cr$V2)
cr$V3 <- as.numeric(cr$V3)
cr$V4 <- as.factor(cr$V4)
cr$V5 <- as.factor(cr$V5)
cr$V6 <- as.character(cr$V6)
cr$V6 <- as.factor(cr$V6)
cr$V7<- as.character(cr$V7)
cr$V7 <- as.factor(cr$V7)
cr$V8 <- as.numeric(cr$V8)
cr$V9 <- as.factor(cr$V9)
cr$V10 <- as.factor(cr$V10)
cr$V11 <- as.numeric(cr$V11)
cr$V12 <- as.factor(cr$V12)
cr$V13 <- as.factor(cr$V13)
cr$V11 <-as.numeric(cr$V14)
cr$V15 <- as.numeric(cr$V15)
cr$V16 <- as.factor(cr$V16)


for (i in 1:length(colnames(cr))){
  for (j in 1:length(cr[,(colnames(cr)[i])])){
      if (class(cr[,(colnames(cr)[i])]) == "factor"){
         if (cr[j,i] == "?"){
          cr[j,i] <- getmode(cr[,(colnames(cr)[i])])
        }
      }
    if (class(cr[,(colnames(cr)[i])]) == "numeric"){
      if (cr[j,i] == "?"){
        cr[j,i] <- getmean(cr[,(colnames(cr)[i])])
      }
    }
  }
}
cr$V14[cr$V14 == "?"] <- getmean(cr$v14)
cr$V6[cr$V6 == "?"] <- getmode(cr$V6)
return(cr)
}
cr <- final(cr)
summary(cr)
```

```
##  V1            V2               V3           V4       V5            V6     
##  ?:  0   Min.   :  1.00   Min.   : 0.000   ?:  0   ? :  0   c      :146  
##  a:210   1st Qu.: 70.25   1st Qu.: 1.000   l:  2   g :525   q      : 78  
##  b:480   Median :130.00   Median : 2.750   u:525   gg:  2   w      : 64  
##          Mean   :146.44   Mean   : 4.759   y:163   p :163   i      : 59  
##          3rd Qu.:219.75   3rd Qu.: 7.207                    aa     : 54  
##          Max.   :350.00   Max.   :28.000                    ff     : 53  
##                                                             (Other):236  
##        V7            V8          V9      V10          V11         V12    
##  v      :408   Min.   : 0.000   f:329   f:395   Min.   :  1.00   f:374  
##  h      :138   1st Qu.: 0.165   t:361   t:295   1st Qu.: 19.00   t:316  
##  bb     : 59   Median : 1.000                   Median : 54.00          
##  ff     : 57   Mean   : 2.223                   Mean   : 58.17          
##  j      :  8   3rd Qu.: 2.625                   3rd Qu.: 95.00          
##  z      :  8   Max.   :28.500                   Max.   :171.00          
##  (Other): 12                                                            
##  V13          V14           V15           V16    
##  g:625   00000  :145   Min.   :     0.0   -:383  
##  p:  8   00120  : 35   1st Qu.:     0.0   +:307  
##  s: 57   00200  : 35   Median :     5.0          
##          00160  : 34   Mean   :  1017.4          
##          00080  : 30   3rd Qu.:   395.5          
##          00100  : 30   Max.   :100000.0          
##          (Other):381                             
```

```
cr$V2 <- as.numeric(as.character(cr$V2))
cr$V3 <- as.numeric(as.character(cr$V3))
cr$V8 <- as.numeric(as.character(cr$V8))
cr$V11 <-as.numeric(as.character(cr$V11))
cr$V14 <-as.numeric(as.character(cr$V14))
cr$V15 <-as.numeric(as.character(cr$V15))
mydata <- as.numeric(cr$V2, cr$V3, cr$V8, cr$V11, cr$V14, cr$V15)


result <- cr %>%select(V2, V3, V8, V11, V14, V15)
set.seed(100)
result <- data.frame(result)
#result
```

```
####2
#a
for ( i in 1:length(result)){
  print(class(result[,i]))
}
```

```
## [1] "numeric"
## [1] "numeric"
## [1] "numeric"
## [1] "numeric"
## [1] "numeric"
## [1] "numeric"
```

```
rando_sample <- result[sample(690, size=100, replace = T),]
#View(rando_sample)

#b

pairs(rando_sample, histogram=TRUE, pch=19)
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```

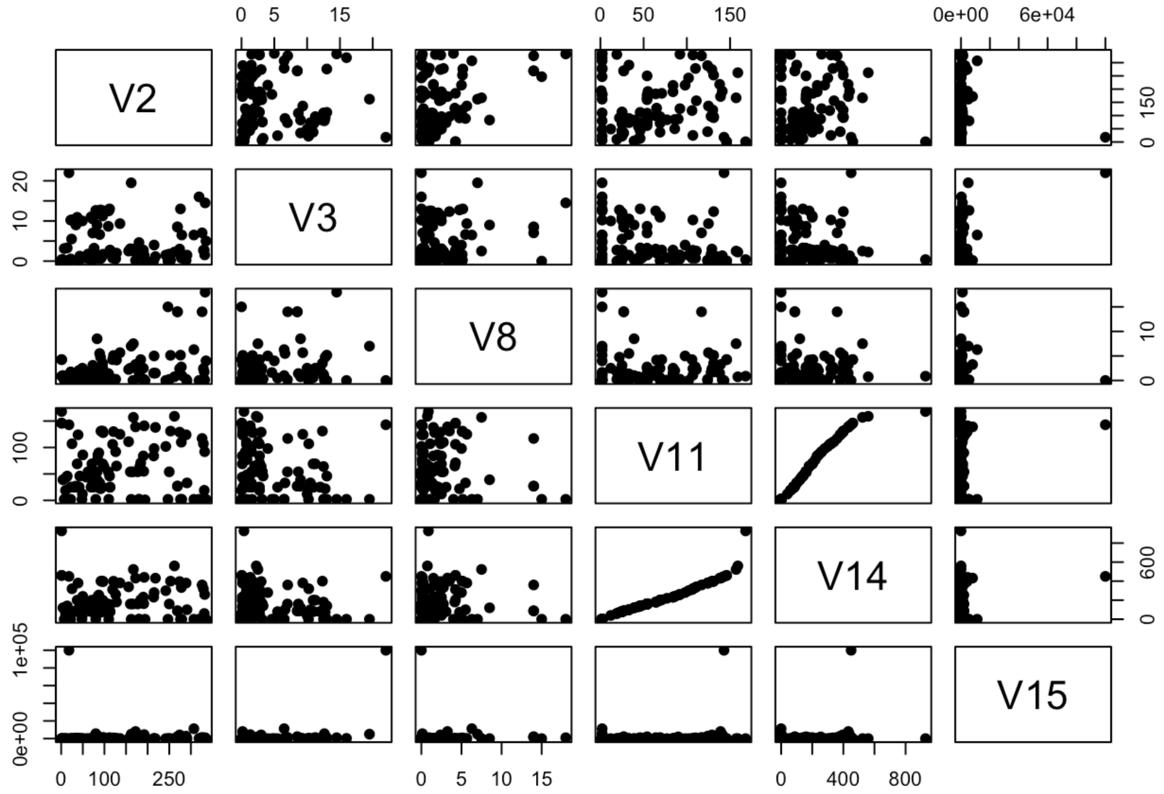
```
## Warning in axis(side = side, at = at, labels = labels, ...): "histogram" is
## not a graphical parameter
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "histogram" is not a
## graphical parameter
```

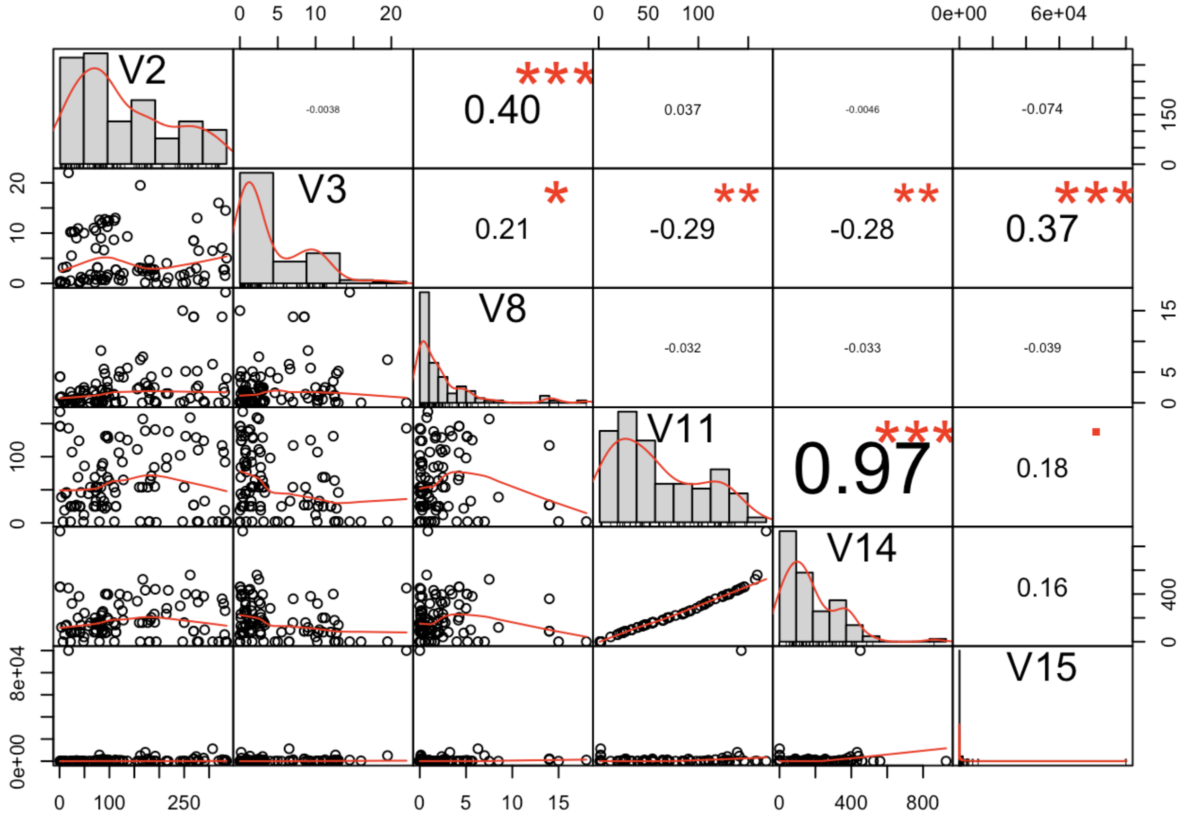
```
## Warning in plot.window(...): "histogram" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "histogram" is not a graphical parameter
```

```
## Warning in title(...): "histogram" is not a graphical parameter
```



```
chart.Correlation(rando_sample, histogram=TRUE, pch=19)
```

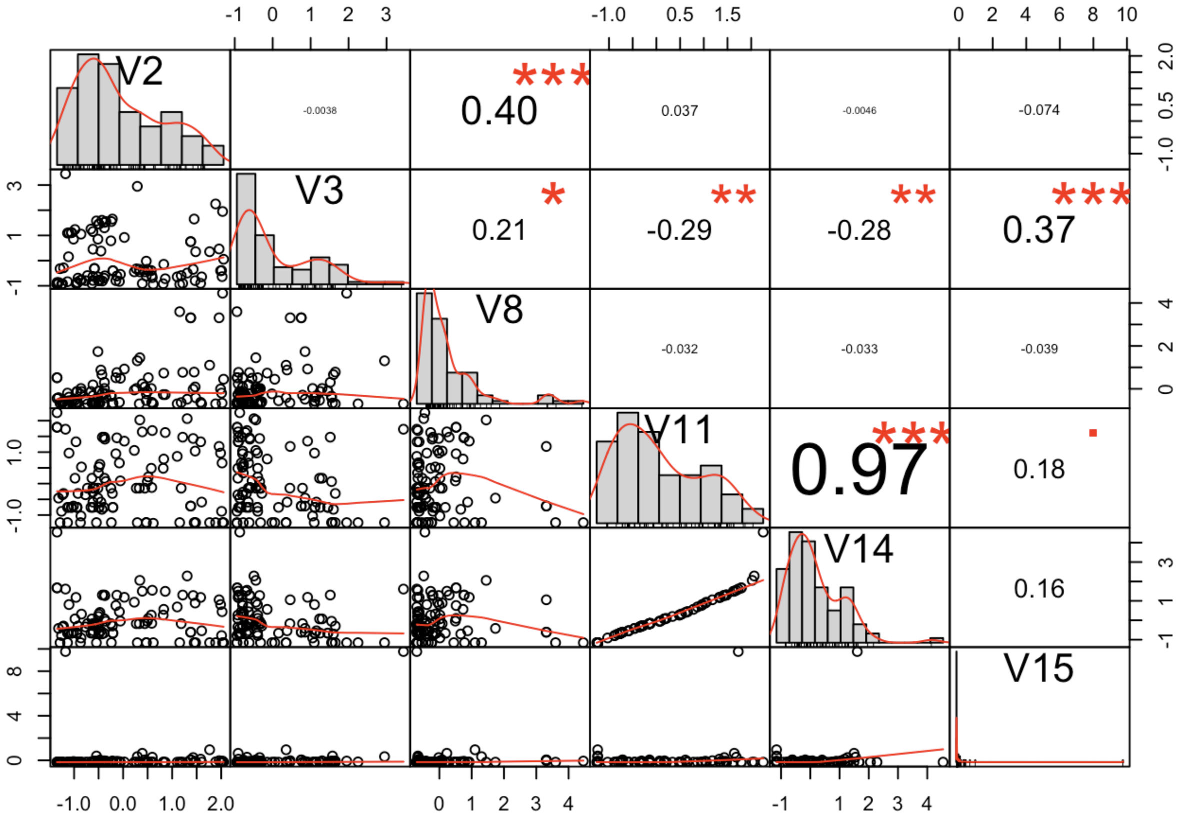


```
#cor(rando_Sample)

#d

Normalize <- function(s1){
  for ( i in 1:length(s1)){
     s1[i] = ((s1[i]-mean(s1))/sqrt(sum((s1[i]-mean(s1))^2)))
  }
  return(s1)
}
rando_sample <- scale(rando_sample)


#e
#pairs(rando_sample)
chart.Correlation(rando_sample, histogram=TRUE, pch=19)
```



```
#3) It didnot effect the correlation. We can see that before normalization and
after normalization, correlation in data is almost same.
```