STATS 419 Survey of Multivariate Analysis

03_Datasets_Writeup

Nahom Debela
(nahom.debela@wsu.edu)
[NA]

Instructor: Monte J. Shaffer

03 November 2020

```
library(devtools)
my.source = 'local';
local.path = "C:/Users/nahom/_git_/WSU_STATS419_FALL2020/";
local.data.path = ""
source( paste0(local.path, "functions/libraries.R"), local = T);
```

1 Matrix

```
source( paste0(local.path,"WEEK-03/functions/functions-personality.R"), local=T );

myMatrix = matrix ( c (
                                                1, 0, 2,
                                                0, 3, 0,
                                                4, 0, 5
                                                ), nrow=3, byrow=T);
#Transpose matrix
transposeMatrix(myMatrix)
```

```
##      [,1] [,2] [,3]
## [1,]    1    0    4
## [2,]    0    3    0
## [3,]    2    0    5
```

```
#Rotate by 90 degrees
rotate90(myMatrix)
```

```
##      [,1] [,2] [,3]
## [1,]    4    0    1
## [2,]    0    3    0
## [3,]    5    0    2
```

```
#Rotate by 180 degrees
rotate180(myMatrix)
```

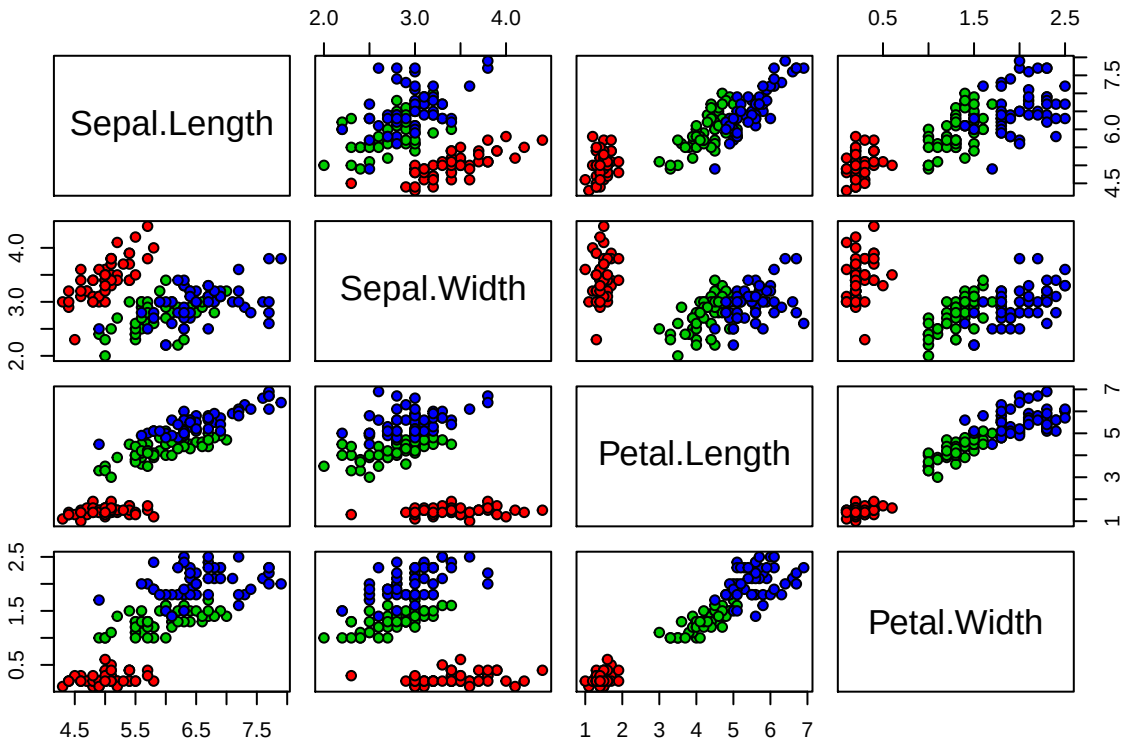
```
##      [,1] [,2] [,3]
## [1,]    5    0    4
## [2,]    0    3    0
## [3,]    2    0    1
```

```
#Rotate by 270 degrees
rotate270(myMatrix)
```

```
##      [,1] [,2] [,3]
## [1,]    2    0    5
## [2,]    0    3    0
## [3,]    1    0    4
```

2 IRIS

```
pairs(iris[1:4], pch = 21, bg = c("red", "green3", "blue") [unclass(iris$Species)])
```




```
#DESCRIPTION

#The Iris dataset contains data about the iris flower and was created by Ronald Fisher. The
#multivariate data set consists of 3 different species and has dimensions of 5 columns and 150
#rows.
```

3 Personality

```
source( paste0(local.path,"WEEK-03/functions/functions-personality.R"), local=T );


myFile = paste0(local.path,"datasets/personality/personality-raw.txt");

my_data <- read.table(myFile, header = TRUE, sep = "|");
# Deleted column V00
my_data$V00 <- NULL

#Strips time apart so we can parse it
date = strptime(my_data$date_test, format = '%m/%d/%Y %H:%M');

new_df = cbind(my_data, date);

#Parse the date

yr <- year(date);
new_df$year <- yr

wk <- week(date);
new_df$week <- wk

new_df$date_test <- NULL


#sort dataframe by date and week, descending
new_df <- new_df[
 order(-new_df$year,-new_df$week),]

# Delete duplicate rows by md5_email column
new_df <- unique(new_df, by = "md5_email")

#write a pipeline delimited dataframe to text
#write.table(new_df, file = "personality-clean.txt", sep = "|")

# newFile = paste0(local.path,"datasets/personality/personality-clean.txt");
# write.table( myData.cleansed , file=newFile, quote=FALSE, col.names=TRUE, row.names=FALSE, sep="|");

dim(my_data)
```

```
## [1] 838  62
```

```
dim(new_df)
```

```
## [1] 822  64
```

```
mshafer_data <- new_df[1,]

doSummary(mshafer_data)
```

```
## length 64
```

```
## Number of NA 0
## Mean 36.17419
## Median 3.4
## Mode 4.2
## Variance 0.7528136
## Standard Deviation 0.8676483
```

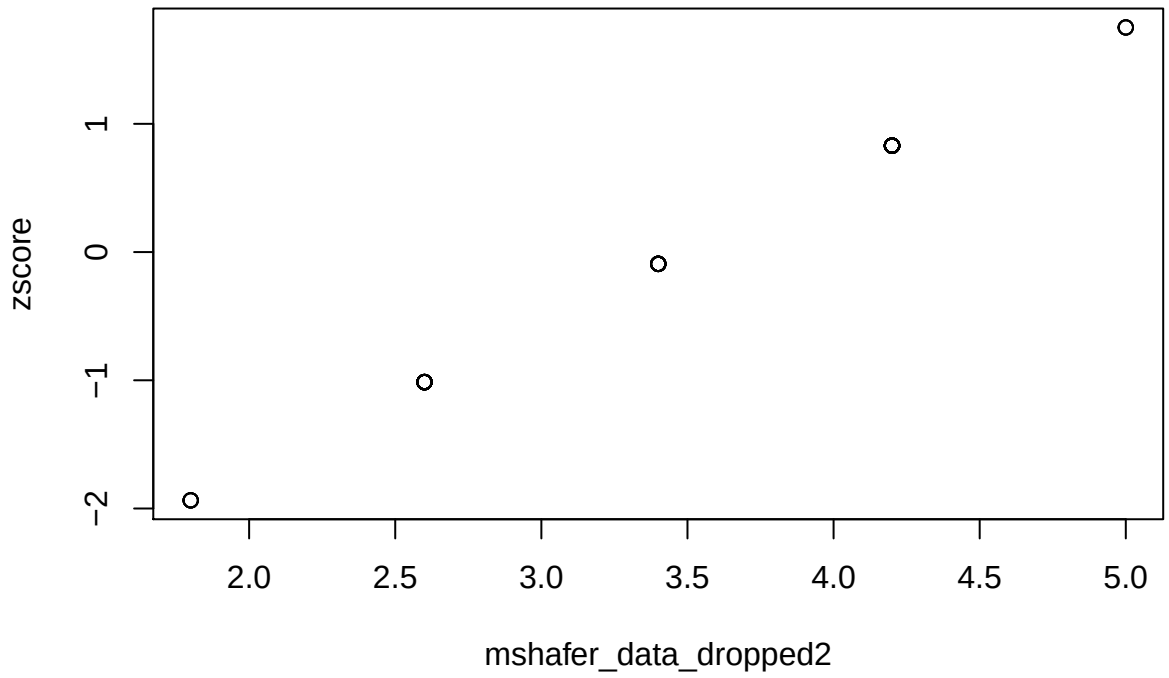
```
#doVariance(mshafer_data)
doMode(mshafer_data)
```

```
##     V02
## 838 4.2
```

4 Variance and Z-Scores

```
source( paste0(local.path,"WEEK-03/functions/functions-personality.R"), local=T );

zscore_plot(mshafer_data)
```




```
##  [1] -0.09220326  0.82982933 -1.01423585  0.82982933 -1.01423585 -1.01423585
##  [7]  0.82982933 -1.01423585 -0.09220326  0.82982933  0.82982933 -0.09220326
## [13] -0.09220326  0.82982933  1.75186192 -0.09220326  1.75186192 -0.09220326
## [19] -1.93626843 -1.01423585 -1.01423585 -1.01423585  0.82982933 -0.09220326
## [25]  1.75186192 -1.01423585  0.82982933 -0.09220326 -1.01423585 -1.01423585
```

```
## [31]  0.82982933 -1.93626843 -0.09220326  0.82982933  0.82982933  0.82982933
## [37] -1.01423585  0.82982933 -1.01423585  0.82982933  0.82982933  0.82982933
## [43]  0.82982933 -1.01423585  0.82982933  0.82982933 -1.01423585 -0.09220326
## [49] -1.01423585  0.82982933 -1.93626843  0.82982933 -1.01423585 -0.09220326
## [55]  0.82982933  0.82982933 -1.93626843  0.82982933 -1.01423585  0.82982933
```

```
#The pattern is obvious in the picture. There is a positive correlation between z score and the raw
#data and this is because z score is directly related to how many standard deviations it is from
#the mean. This is shown in the graph: the higher the z score the farther from the mean and the
#same is true for a low z score. Additionally the mean is directly at where the z score is 0 which
#should happen.
```