R Notebook sandbox: Assignment "Datasets" (10 points)

Getting Everything Set Up

```
local.path = "C:/_git_/WSU_STATS419_FALL2020/";
source(paste0(local.path,"/functions/libraries.R"),local = T)
source( paste0(local.path,"functions/functions-imdb.R"), local=T );
```

Matrix

Create the "rotate matrix" functions as described in lectures. Apply to the example "myMatrix".

```
source(paste0(local.path,"/functions/functions-matrix.R"),local = T)
# Create myMatrix
myMatrix = matrix ( c (
1, 0, 2,
0, 3, 0,
4, 0, 5
), nrow=3,
byrow=T);

# Rotating clockwise.
rotateMatrix90(myMatrix)
```

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

```
rotateMatrix180(myMatrix)
```

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

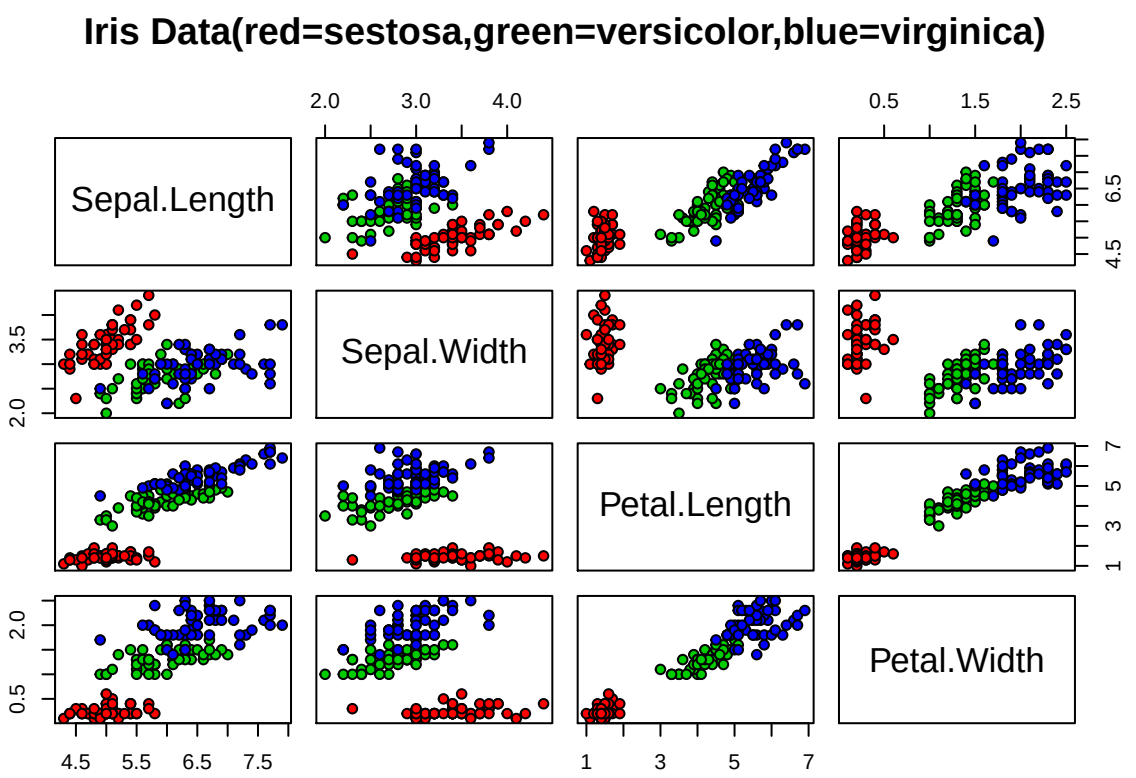
```
rotateMatrix270(myMatrix)
```

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

IRIS Question 2

Recreate the graphic for the IRIS Data Set using R. Same titles, same scales, same colors.

```
library(datasets)
plot(iris[1:4],pch = 21,cex = 1,col = 'black', bg= c("red", "green3",
"blue")[iris$Species],main = 'Iris Data(red=sestosa,green=versicolor,blue=virginica)',gap.axis = 1.5)
```



IRIS Question 3

Right 2-3 sentences concisely defining the IRIS Data Set. Maybe search KAGGLE for a nice template. Be certain the final writeup are your own sentences (make certain you modify what you find, make it your own, but also cite where you got your ideas from). NOTE: Watch the video, Figure 8 has a +5 EASTER EGG.

Response

The Iris data set was created in 1936 by Edgar Anderson and is a multivariate data set that containing four different measurements of three different Iris flower species. The data set contains 50 records for each of the different flower species; Setosa, Virginica, and Versicolor recording measurements for petal length, petal width, sepal width and sepal length, all recorded in centimeters.

Cleaning data Question 4

Import "personality-raw.txt" into R. Remove the V00 column. Create two new columns from the current column "date_test": year and week. Stack Overflow may help: https://stackoverflow.com/questions/22439540/how-to-get-week-numbers-from-dates ... Sort the new data frame by YEAR, WEEK so the newest tests are first ... The newest tests (e.g., 2020 or 2019) are at the top of the data frame. Then remove duplicates using the unique function based on the column "md5_email". Save the data frame in the same "pipe-delimited format" (| is a pipe) with the headers. You will keep the new data frame as "personality-clean.txt" for

future work (you will not upload it at this time). In the homework, for this tasks, report how many records your raw dataset had and how many records your clean dataset has.

```
source(paste0(local.path,"/functions/functions-cleanup.R"),local = T)
files = paste0(local.path,"/datasets/personality-raw.txt")
personality_raw = read.csv(files,header = T, sep = '|')
personality_clean = removeColumn(personality_raw,'V00')
personality_clean = convertDates(personality_clean,personality_clean$date_test)
personality_clean = removeDuplicates(personality_clean,personality_clean$md5_email)
dim(personality_clean)
```

```
## [1] 678  64
```

```
dim(personality_raw)
```

```
## [1] 838  63
```

```
head(personality_clean)
```

```
##                            md5_email       date_test V01 V02 V03 V04 V05 V06
## 838 b62c73cdaf59e0a13de495b84030734e  4/6/2020 12:57 3.4 4.2 2.6 4.2 2.6 2.6
## 837 1358d38e6898b1a0e5940f8b99ba2325 12/1/2019 22:12 3.4 3.4 3.4 4.2 4.2 4.2
## 835 f529455e4400e76f323f8c68154e194b   5/6/2019 4:44 4.2 5.0 1.8 4.2 4.2 5.0
## 836 0445a05e751e17de30ebdcbcdb575d59  5/6/2019 10:32 1.8 2.6 3.4 4.2 5.0 3.4
## 828 bfd1c69406d322d17312e965752813c2  5/2/2019 10:26 2.6 4.2 1.0 4.2 4.2 2.6
## 829 9cf05d7d516099c9533b98beb91993b9  5/2/2019 10:48 5.0 5.0 1.8 5.0 5.0 4.2
##     V07 V08 V09 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25
## 838 4.2 2.6 3.4 4.2 4.2 3.4 3.4 4.2 5.0 3.4 5.0 3.4 1.8 2.6 2.6 2.6 4.2 3.4 5.0
## 837 5.0 3.4 4.2 3.4 2.6 3.4 3.4 4.2 4.2 4.2 4.2 4.2 3.4 2.6 3.4 4.2 4.2 4.2 2.6
## 835 3.4 3.4 4.2 3.4 2.6 2.6 4.2 5.0 3.4 4.2 5.0 4.2 2.6 2.6 1.8 3.4 5.0 3.4 1.8
## 836 2.6 2.6 5.0 3.4 2.6 4.2 2.6 3.4 4.2 3.4 4.2 4.2 3.4 1.8 2.6 3.4 4.2 5.0 1.8
## 828 3.4 1.8 4.2 4.2 1.8 2.6 3.4 5.0 4.2 4.2 5.0 4.2 3.4 1.8 1.0 3.4 4.2 3.4 1.8
## 829 1.0 5.0 5.0 5.0 1.0 5.0 5.0 5.0 5.0 5.0 5.0 3.4 3.4 3.4 4.2 5.0 5.0 3.4 1.8
##     V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38 V39 V40 V41 V42 V43 V44
## 838 2.6 4.2 3.4 2.6 2.6 4.2 1.8 3.4 4.2 4.2 4.2 2.6 4.2 2.6 4.2 4.2 4.2 4.2 2.6
## 837 4.2 4.2 3.4 2.6 4.2 4.2 3.4 4.2 3.4 4.2 5.0 3.4 4.2 4.2 4.2 4.2 4.2 4.2 4.2
## 835 4.2 3.4 5.0 1.8 5.0 4.2 1.8 4.2 3.4 2.6 3.4 2.6 3.4 3.4 5.0 3.4 3.4 3.4 3.4
## 836 3.4 2.6 3.4 2.6 4.2 5.0 5.0 5.0 5.0 5.0 5.0 3.4 5.0 5.0 5.0 4.2 5.0 5.0 5.0
## 828 4.2 5.0 3.4 1.8 4.2 3.4 4.2 4.2 3.4 4.2 3.4 1.8 5.0 3.4 4.2 1.8 2.6 4.2 4.2
## 829 5.0 5.0 4.2 3.4 5.0 5.0 4.2 5.0 5.0 5.0 5.0 2.6 3.4 5.0 4.2 5.0 5.0 3.4 5.0
##     V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56 V57 V58 V59 V60 week year
## 838 4.2 4.2 2.6 3.4 2.6 4.2 1.8 4.2 2.6 3.4 4.2 4.2 1.8 4.2 2.6 4.2   15 2020
## 837 3.4 4.2 4.2 2.6 3.4 4.2 3.4 4.2 4.2 4.2 4.2 3.4 4.2 4.2 3.4 3.4   48 2019
## 835 4.2 5.0 3.4 4.2 3.4 4.2 2.6 3.4 5.0 5.0 3.4 3.4 3.4 3.4 1.8 3.4   19 2019
## 836 5.0 5.0 5.0 4.2 4.2 3.4 3.4 5.0 5.0 5.0 2.6 5.0 5.0 5.0 4.2 5.0   19 2019
## 828 4.2 4.2 4.2 2.6 3.4 1.8 2.6 2.6 5.0 4.2 3.4 2.6 2.6 4.2 4.2 4.2   18 2019
## 829 5.0 5.0 5.0 3.4 5.0 5.0 5.0 5.0 5.0 5.0 1.0 2.6 3.4 5.0 5.0 3.4   18 2019
```

The raw dataset had 878 records, and the clean dataset has 673 records.

Custom Functions

Write functions for doSummary and sampleVariance and doMode ... test these functions in your homework on the "monte.shaffer@gmail.com" record from the clean dataset. Report your findings. For this "monte.shaf fer@gmail.com" record, also create z-scores. Plot(x,y) where x is the raw scores for "monte.shaffer@gmail.com" and y is the z-scores from those raw scores. Include the plot in your assignment, and write 2 sentences

describing what pattern you are seeing and why this pattern is present.

Setup and Get Row For "monte.shaffer@gmail.com"

```
source(paste0(local.path,"/functions/functions-custom.R"),local = T)
monte_row = personality_clean[which(personality_clean$md5_email ==
'b62c73cdaf59e0a13de495b84030734e'),]
```

doSummary

```
doSummary(monte_row)
```

```
##    SumSq   Sum  variance
## 1 771.04 208.8 0.7528136
```

```
##                           md5_email mode Mean naNum Length
## 838 b62c73cdaf59e0a13de495b84030734e  4.2 3.48     0     64
##     TwoPassVariance.variance NaiveVariance.variance        Sd
## 838                0.7528136              0.7528136 0.8676483
```

custom_mode

```
custom_mode(monte_row)
```

```
##                           md5_email mode
## 838 b62c73cdaf59e0a13de495b84030734e  4.2
```

doSampleVariance Naive

```
doSampleVariance(monte_row, 'naive')
```

```
##    SumSq   Sum  variance
## 1 771.04 208.8 0.7528136
```

doSampleVariance Two Pass

```
doSampleVariance(monte_row,'Two Pass')
```

```
##    Sum1   Sum2  variance
## 1 208.8 44.416 0.7528136
```

Zscores

```
num_vals = as.numeric(c(Filter(is.numeric,monte_row)))
z = (num_vals - mean(num_vals)/sd(num_vals))
plot(num_vals,z,main = 'Raw Scores Vs Z-Scores',xlab = 'raw scores',ylab = 'z-scores')
```

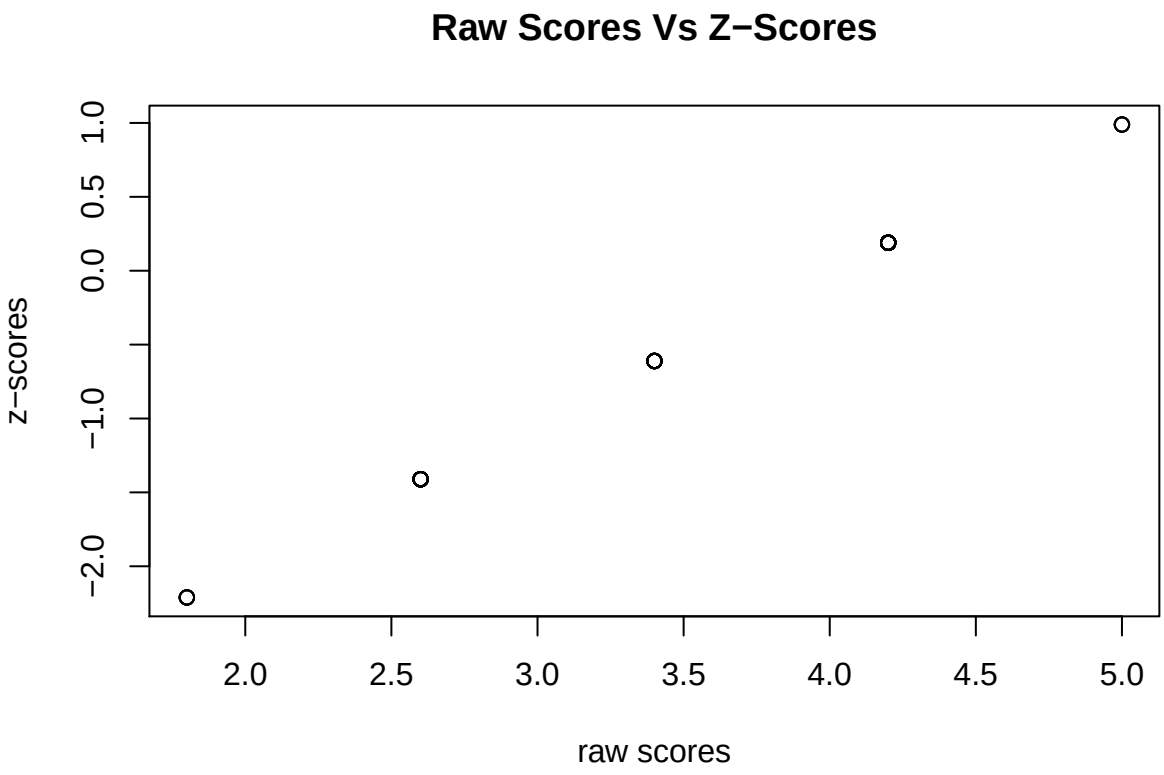


Figure 1: Plotting Raw Scores Against Z-scores