STATS 419 Survey of Multivariate Analysis

Week 03 Assignment 02_datasets

Isaac Schultz
(isaac.schultz@wsu.edu)
[11583435]

Instructor: Monte J. Shaffer

21 September 2020

0.1 R Markdown

```
library(devtools);  # devtools is required for function source_url() to work ...
my.source = 'github';
github.path = "https://raw.githubusercontent.com/IsaacMSchultz/CptS419/";
source_url( paste0(github.path,"master/functions/libraries.R") );
source_url( paste0(github.path,"master/functions/functions-imdb.R") );
```

1 Matrix

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

```
source_url( paste0(github.path,"master/functions/functions-matrix.R") );


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

rotateMatrix90 = function(mat)
{
    reversed = apply(mat, 2, rev); # reverse the matrix order going through the columns using apply
    t(reversed); # Transpose the results and return
}

rotateMatrix180 = function(mat)
{
    rotateMatrix90(rotateMatrix90(mat));
}
```

```
rotateMatrix270 = function(mat)
{
    rotateMatrix180(rotateMatrix90(mat));
}
```

```
transposeMatrix(myMatrix);
```

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

```
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
```

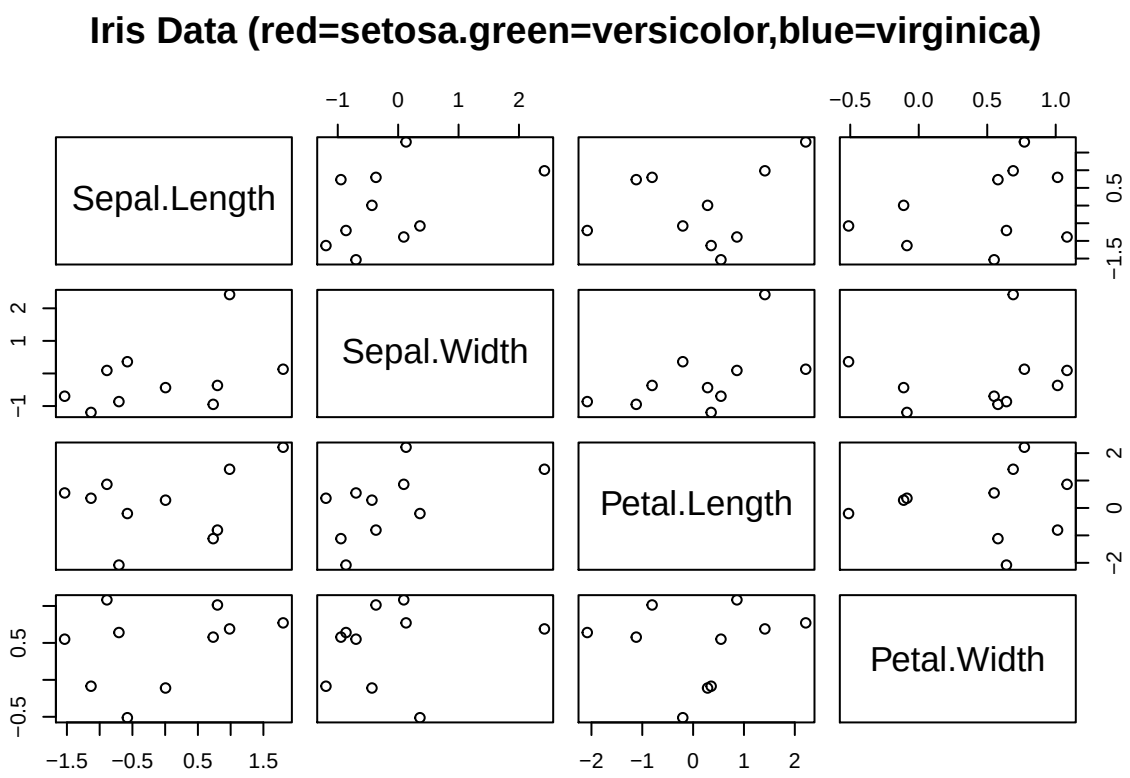
2 IRIS

Recreate the graphic for the IRIS Data Set using R. Same titles, same scales, same colors. See: https://en.wikipedia.org/wiki/Iris_flower_data_set#/media/File:Iris_dataset_scatterplot.svg

```
Sepal.Length <- rnorm(10);
Sepal.Width <- rnorm(10);
Petal.Length <- rnorm(10);
Petal.Width <- rnorm(10);

df = data.frame(Sepal.Length, Sepal.Width, Petal.Length, Petal.Width);

plot(df, main="Iris Data (red=setosa,green=versicolor,blue=virginica)", );
```



Sentences: [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.]

3 Personality

3.1 Cleanup RAW

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.

```
remote_file <- url(paste0(github.path,"master/datasets/personality/personality-raw.txt"));
my_data <- read.table(remote_file, sep ="|", header = TRUE, dec =".");
df = subset(my_data, select = -c(V00));

new_data <- within(df, {
    dates <- strptime(date_test, format="%m/%d/%Y %H:%M");
    years <- format(dates, "%Y");
    weeks <- format(dates, "%V");
```

```
});

clean_data <- new_data[order(new_data$years, new_data$weeks),];
clean_data = subset(clean_data, select = -c(dates));
unique_data <- clean_data[!duplicated(clean_data[,"md5_email"]),];

write.table(unique_data, file = "personality-clean.txt", sep = "|", row.names = TRUE, col.names = NA);
```

4 Variance and Z-scores

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.shaffer@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.

4.1 Variance

4.1.1 Naive

Problems with this approach

4.1.2 Traditional Two Pass

4.2 Z-Scores

5 Will vs Denzel

```
source_url( paste0(github.path,"master/functions/functions-imdb.R"));
```

Compare Will Smith and Denzel Washington. You will have to create a new variable $millions.2000 that converts each movie's $millions based on the $year of the movie, so all dollars are in the same time frame. You will need inflation data from about 1980-2020 to make this work.

5.1 Will Smith

```
nmid = "nm0000226";
    will = grabFilmsForPerson(nmid);
    plot(will$movies.50[,c(1,6,7:10)]);
```

```
    boxplot(will$movies.50$millions);
```

```
        widx =  which.max(will$movies.50$millions);
    will$movies.50[widx,];
```

```
##    rank   title      ttid year rated minutes                       genre ratings
## 15   15 Aladdin tt6139732 2019    PG     128 Adventure, Family, Fantasy       7
##    metacritic  votes millions
## 15         53 216916   355.56
```

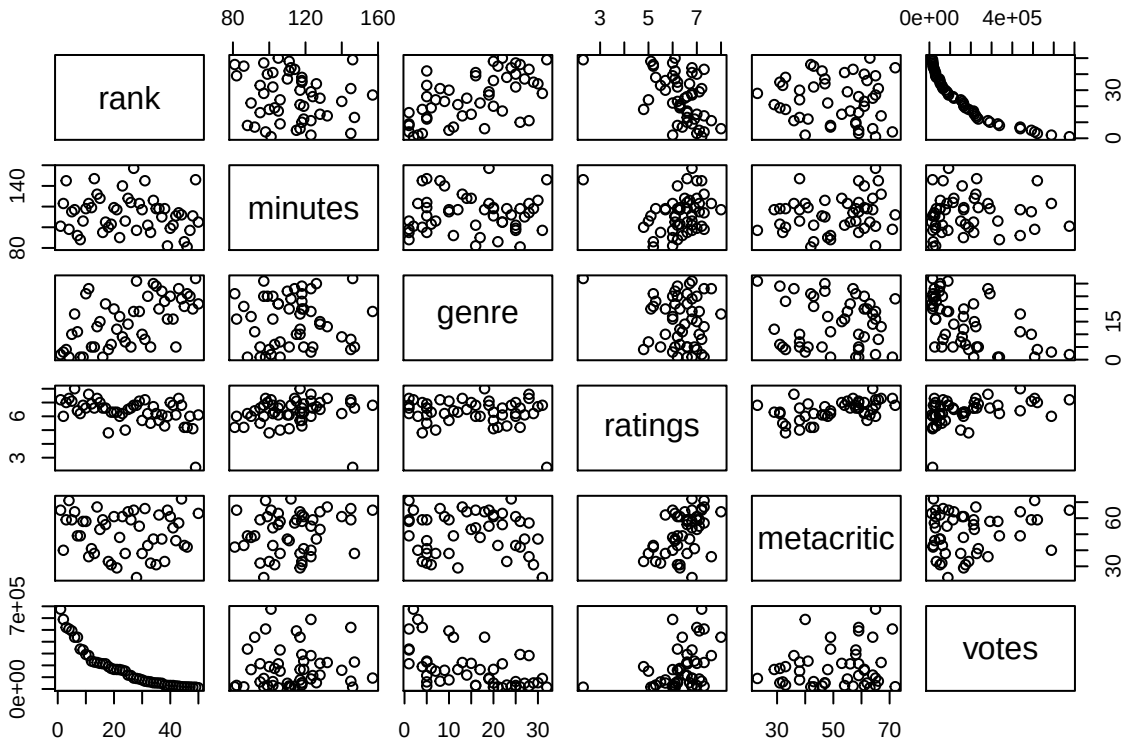


Figure 1: Will Smith scatterplot: IMDB(2020)

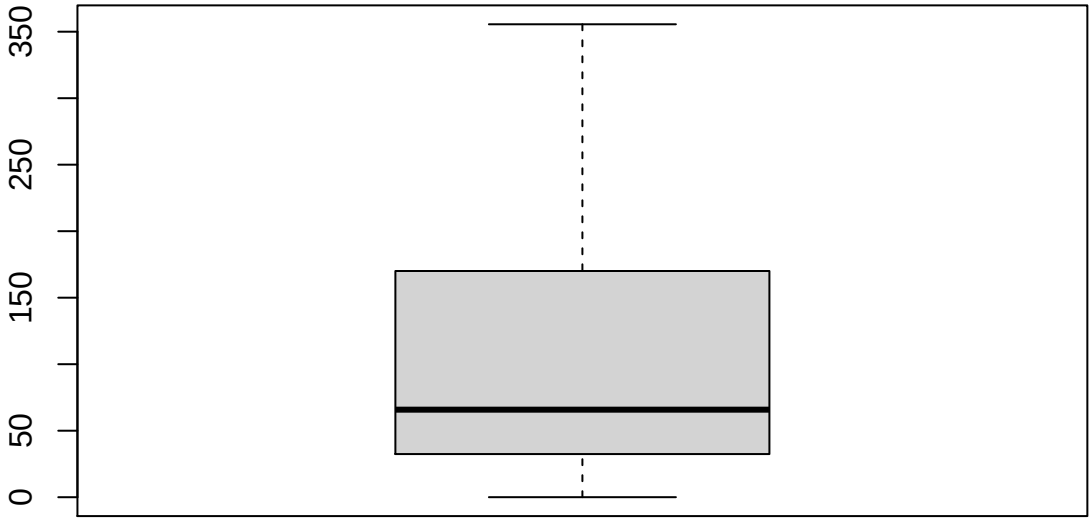


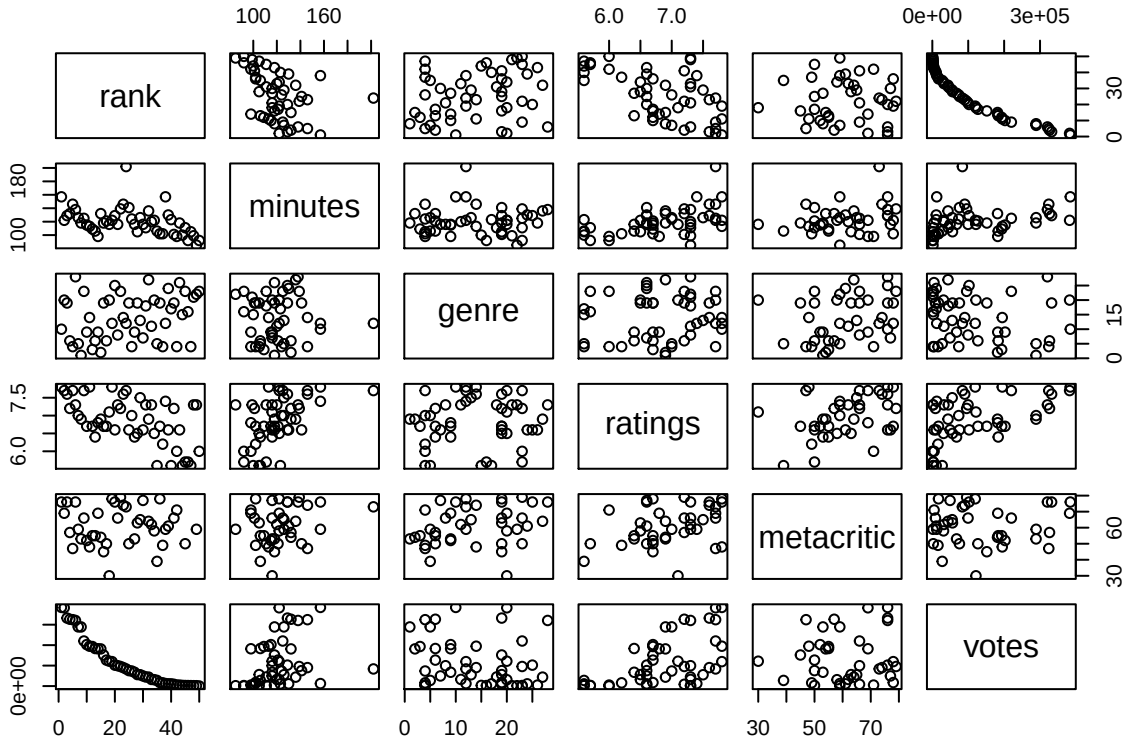
Figure 2: Will Smith boxplot raw millions: IMDB(2020)

```
    summary(will$movies.50$year);  # bad boys for life ... did data change?
```

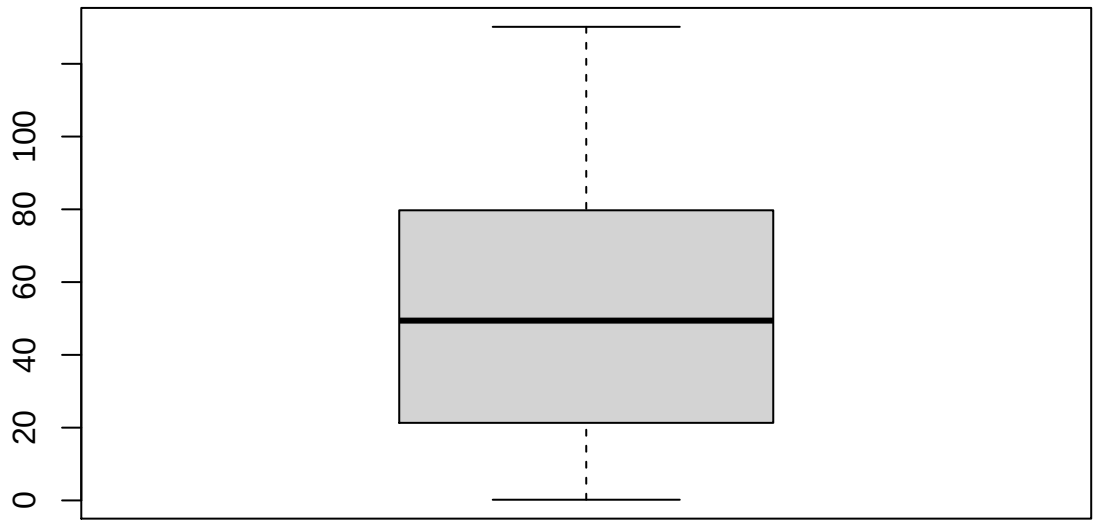
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1993    2001    2006    2007    2014    2020
```

5.2 Denzel Washington

```
nmid = "nm0000243";
    denzel = grabFilmsForPerson(nmid);
    plot(denzel$movies.50[,c(1,6,7:10)]);
```




```
    boxplot(denzel$movies.50$millions);
```




```
        didx =  which.max(denzel$movies.50$millions);
    denzel$movies.50[didx,];
```

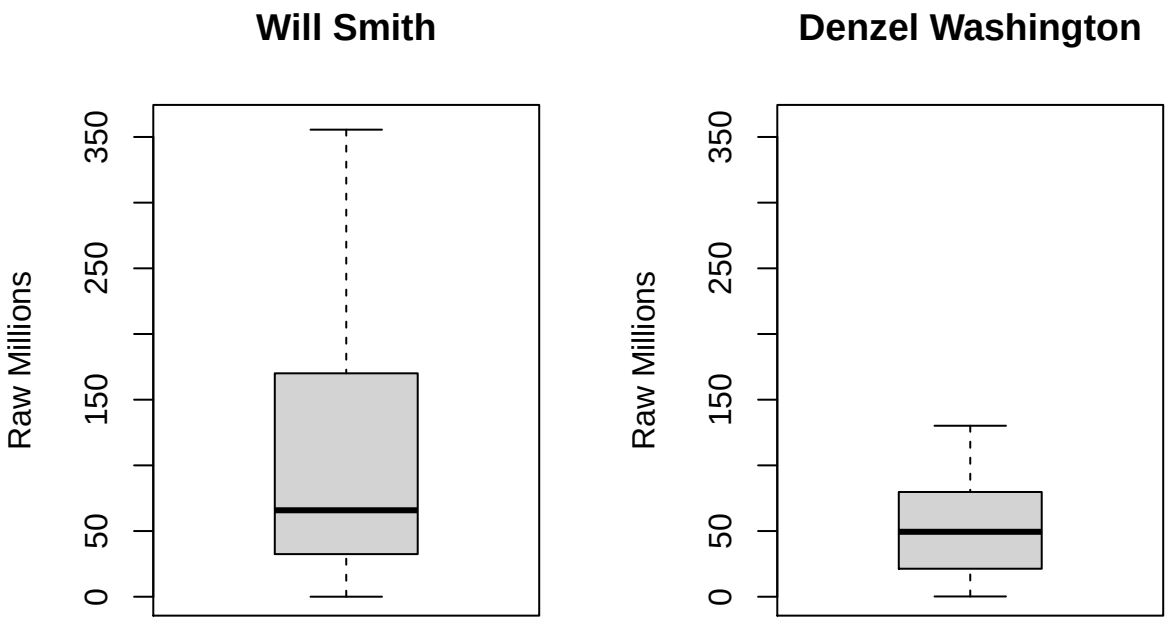
```
##   rank             title      ttid year rated minutes                    genre
## 1    1 American Gangster tt0765429 2007     R     157 Biography, Crime, Drama
##   ratings metacritic  votes millions
## 1     7.8         76 384284   130.16
```

```
        summary(denzel$movies.50$year);
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1981    1993    1999    2000    2008    2018
```

5.3 BoxPlot of Top-50 movies using Raw Dollars

```
par(mfrow=c(1,2));
    boxplot(will$movies.50$millions, main=will$name, ylim=c(0,360), ylab="Raw Millions" );
    boxplot(denzel$movies.50$millions, main=denzel$name, ylim=c(0,360), ylab="Raw Millions" );
```




```
par(mfrow=c(1,1));
```

5.4 Side-by-Side Comparisons

Build side-by-side box plots on several of the variables (including #6) to compare the two movie stars. After each box plot, write 2+ sentence describing what you are seeing, and what conclusions you can logically make. You will need to review what the box plot is showing with the box portion, the divider in the box, and the whiskers.

5.4.1 Adjusted Dollars (2000)

5.4.2 Total Votes (Divide by 1,000,000)

5.4.3 Average Ratings

5.4.4 Year? Minutes?

5.4.5 Metacritic (NA values)