STATS 419 Survey of Multivariate Analysis

Week 03 Assignment 02_datasets

Justin Pickel
(justin.pickel@wsu.edu)
[11592048]

Instructor: Monte J. Shaffer

10 September 2020

```
#library(devtools)
#github.path ="https://raw.githubusercontent.com/jrpickel/WSU_STATS419_FALL2020/";
#source_url(paste0(github.path,"master/functions/libraries.R"),);
#source_url( paste0(github.path,"master/functions/functions-imdb.R"),  );
```

```
library(devtools)
```

```
## Warning: package 'devtools' was built under R version 3.6.3

## Loading required package: usethis

## Warning: package 'usethis' was built under R version 3.6.3
```

```
github.path = "https://raw.githubusercontent.com/jrpickel/WSU_STATS419_FALL2020/";
source_url(paste0(github.path,"master/functions/libraries.R"),);
```

```
## SHA-1 hash of file is e70b26dff5f2b477307f40a1b6d6c0543b4da6f2

## Warning: package 'pryr' was built under R version 3.6.3

## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp

## Warning: package 'lmtest' was built under R version 3.6.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.6.3

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
## Warning: package 'MASS' was built under R version 3.6.3
```

```
## Warning: package 'tractor.base' was built under R version 3.6.3
```

```
## Warning: package 'stringr' was built under R version 3.6.3
```

```
## Warning: package 'rvest' was built under R version 3.6.3
```

```
## Loading required package: xml2
```

```
## Warning: package 'xml2' was built under R version 3.6.3
```

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

```
## SHA-1 hash of file is e4d1e96f5adb9e2e8c9c29eaaf19656fbace400b
```

1 Martrix

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

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

```
## SHA-1 hash of file is e71ab1e25a6edd60e7fc824957d6db4fd5bfee50
```

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

1.1 Matrix

```
rotateMatrix90(myMatrix)
```

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

1.2 Matrix

```
rotateMatrix180(myMatrix)
```

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

1.3 Matrix

```
rotateMatrix270(myMatrix)
```

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

2 Iris graphic

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

```
library(datasets)
data("iris")

plot(iris[1:4],main="Iris Data (red=setosa,green=versicolor,blue=virginica)",col=c("black"),
pch=21,bg=c("red","blue","#258425")[iris$Species])
```

3 Iris Summary

Write 2-3 sentences concisely defining the IRIS Data Set.

3.1 Iris Summary

The Iris flower dataset contains 150 total data points, 50 data points for each of the three Iris flowers (setosa, virginica, versicolor). The data set records the measurement of four variables (Petal Length, Pedal Width, Sepal Width, Sepal length) for each of the 50 samples. The data set also contains a variable "class" indicating what species the measurements belongs to.

4 Clean Personality dataset

In the homework, for these tasks, report how many records your raw dataset had and how many records your clean dataset has.

4.1 Clean Personality dataset

```
source_url( paste0(github.path,"master/WEEK-03/functions/functions-CleanData.R"),  )
```

```
## SHA-1 hash of file is efff475bf615a8dc113804f86a6f3ba0d84cb99b
```

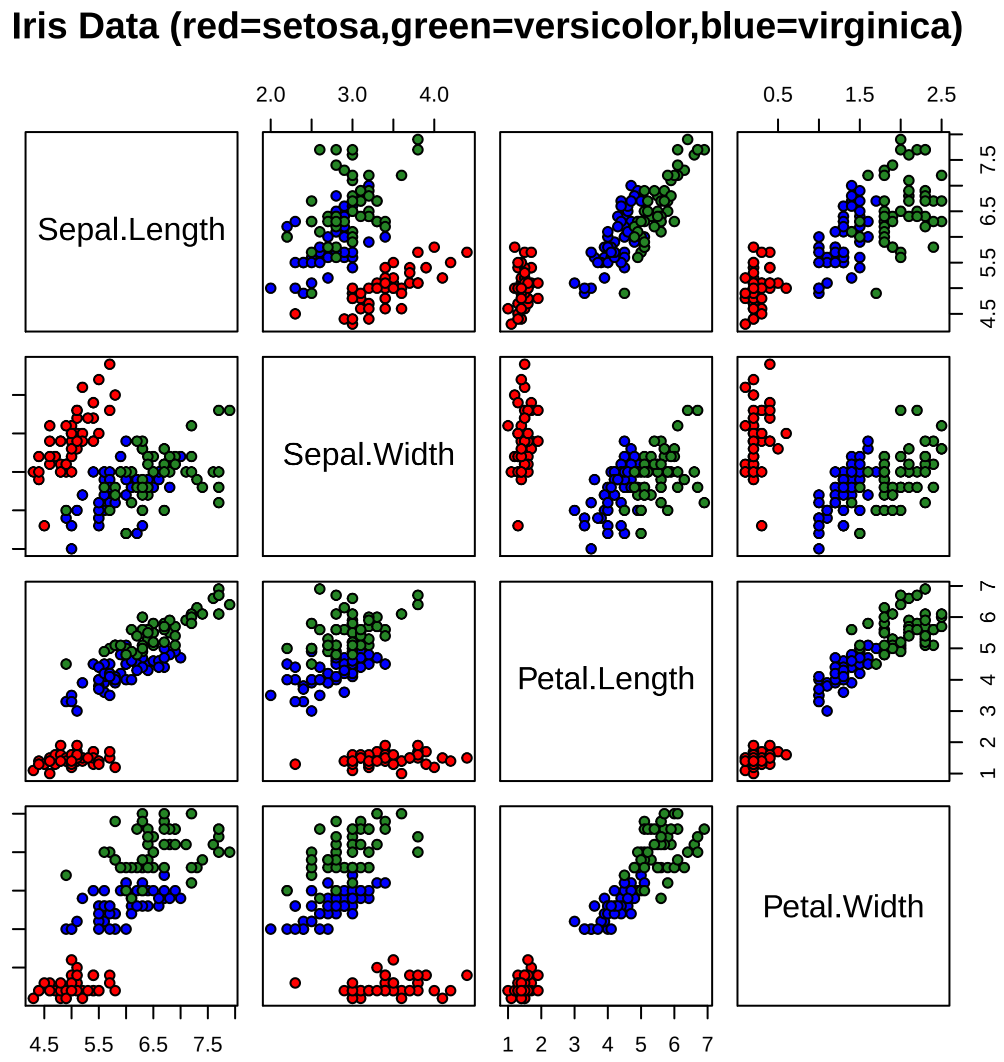


Figure 1: Iris data generated using `plot()`

```
personality.raw = readFile(paste0(github.path,"/master/datasets/personality-raw.txt"))

observations.raw = nrow(personality.raw)
observations.clean = nrow(cleanData(personality.raw))
```

The raw dataset has 838 reords
The clean dataset has 678 reords

5 Functions and Z-score

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.

5.1 Functions doSummary(), doMode()

```
source_url( paste0(github.path,"master/WEEK-03/functions/functions-CleanData.R"),  )
```

```
## SHA-1 hash of file is efff475bf615a8dc113804f86a6f3ba0d84cb99b
```

```
source_url( paste0(github.path,"master/WEEK-03/functions/functions-summary.R"),  )
```

```
## SHA-1 hash of file is 43c075a68a42055f1caba388a17d2bd654c0767c
```

```
personality.raw = readFile(paste0(github.path,"master/datasets/personality-raw.txt"))
personality.clean = cleanData(personality.raw)
record = retrieveTopRecord(personality.clean)

doSummary(record)
```

```
##   length countNA mean median NaiveVar.sum_ NaiveVar.sumSq NaiveVar.variance
## 1     60       0 3.48    3.4         208.8         771.04         0.7528136
##   twoPassVar.sum twoPassVar.sum2 twoPassVar.variance     stDev mode
## 1          208.8          44.416           0.7528136 0.8676483  4.2
```

```
doMode(record)
```

```
## [1] 4.2
```

5.2 Z-scores

```
z.score = (record-mean(record))/sd(record)
plot(record,z.score,main="Z-score and raw score relationship",xlab="Raw score",ylab="Z-score")
```

6 Denzel Washington vs. Will Smith

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.

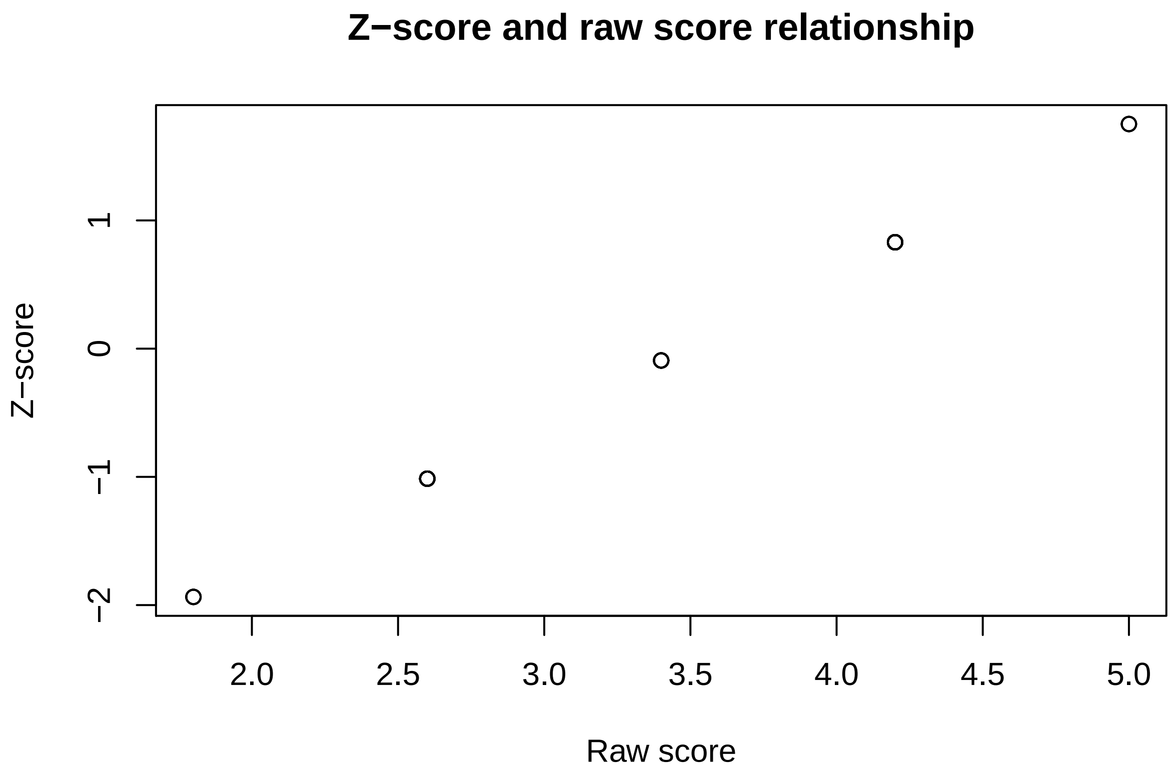


Figure 2: Relationship between z-scores and raw scores

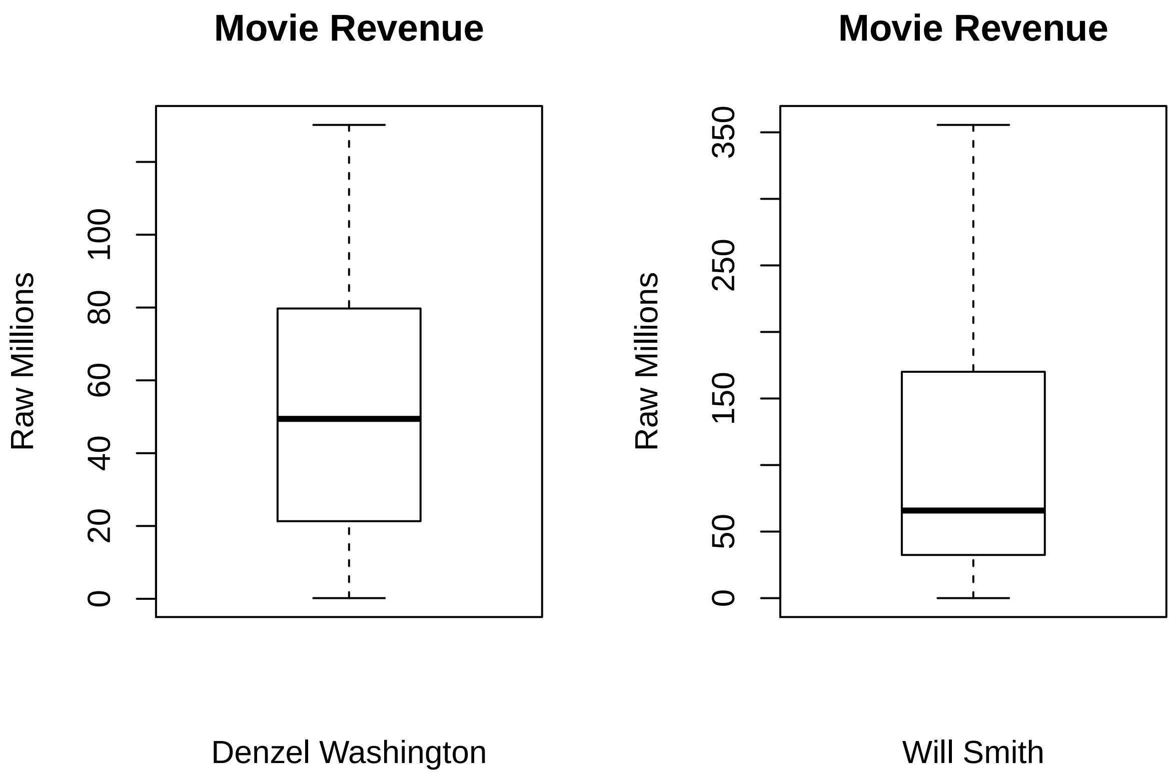


Figure 3: Side by Side comparison of Will Smith and Denzel Washington before conversion

6.1 Side by Side comparison before conversion

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

```
## SHA-1 hash of file is e4d1e96f5adb9e2e8c9c29eaaf19656fbace400b
```

```
nmid = "nm0000226";
will = grabFilmsForPerson(nmid);

nmid = "nm0000243";
denzel = grabFilmsForPerson(nmid);

par(mfrow=c(1,2))
boxplot(denzel$movies.50$millions,xlab="Denzel Washington",ylab="Raw Millions",main="Movie Revenue");
boxplot(will$movies.50$millions,xlab = "Will Smith",ylab="Raw Millions", main="Movie Revenue");
```

6.2 Converting millions into year 2000 dollars

```
source_url( paste0(github.path,"/master/WEEK-03/functions/functions-inflation.R"),  );
```

```
## SHA-1 hash of file is 9fb3081fd15dd3f9102428f5b0f24d1abf0420fa
```

```
will.converted = convertMillions(will,2000)
denzel.converted = convertMillions(denzel,2000)

# show top 5 rows
will.converted$movies.50[1:5,]
```

```
##   rank            title      ttid year rated minutes                      genre
## 1    1       I Am Legend tt0480249 2007 PG-13     101   Action, Adventure, Drama
## 2    2    Suicide Squad tt1386697 2016 PG-13     123 Action, Adventure, Fantasy
## 3    3 Independence Day tt0116629 1996 PG-13     145  Action, Adventure, Sci-Fi
## 4    4     Men in Black tt0119654 1997 PG-13      98  Action, Adventure, Comedy
## 5    5         I, Robot tt0343818 2004 PG-13     115      Action, Drama, Sci-Fi
##   ratings metacritic  votes millions millions.2000
## 1     7.2         65 674890   256.39      212.9349
## 2     6.0         40 587698   325.10      233.2524
## 3     7.0         59 520537   306.17      336.0260
## 4     7.3         71 507409   250.69      268.9646
## 5     7.1         59 491295   144.80      131.9987
```

By converting all the revenue dollars from various movies across different years to a constant year allows us to compare apples to apples. If we do not convert the revenue to the same year our data will be skewed. I used inflation data from 1980 to 2020 by scraping the data from the web using code provided by Monte J. Shaffer to convert every year into the year 2000 dollars. After the conversion step we see when comparing Denzel Washington to Will Smith we can see most movies featuring Will Smith revenue higher than Denzel Washington. Will Smith also has a higher max revenue than Denzel Washington and the median revenue between the two actors is almost equal. We can also see that Will Smith has an outlier, meaning one movie generated way more revenue when compared to all his other top 50 movies.

7 Denzel vs. Will Side by Side comparison of several Variables

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+ sentences 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.

7.1 Denzel vs. Will Side by Side comparison of revenue after conversion

Comparing the revenue generated from movies featuring Denzel Washington to Will Smith using a boxplot. We can see from Figure 4. the median for Will Smith is slightly higher than Denzel Washington. The actors have a median revenue of 56.5897 and 52.4536 million dollars, respectively. Fifty percent of the revenue for Denzel Washington is between 22.3942 and 70.0587 million dollars. Fifty percent of the revenue generated by Will smith is between 30.4543 and 137.6564 million dollars. The max revenue for Denzel is 120.0872 while Will Smith has a max of 336.0260 million dollars. From these statistics we can conclude Will Smith has been featured in more films with higher revenue than Denzel Washington.

```
# Revenue in year 2000 dollars comparison
boxplot(denzel.converted$movies.50$millions.2000,will.converted$movies.50$millions.2000,
names=c("Denzel Washington","Will Smith"),main="Movie revenue in year 2000 dollars",
ylab= "Revenue (millions)",xlab = "Actors");
```

7.2 Denzel vs. Will Side by Side comparison of ratings

Comparing the ratings from movies featuring Denzel Washington to Will Smith using a boxplot. We can see from Figure 5. the average ratings for films featuring Will Smith is slightly lower than Denzel Washington.

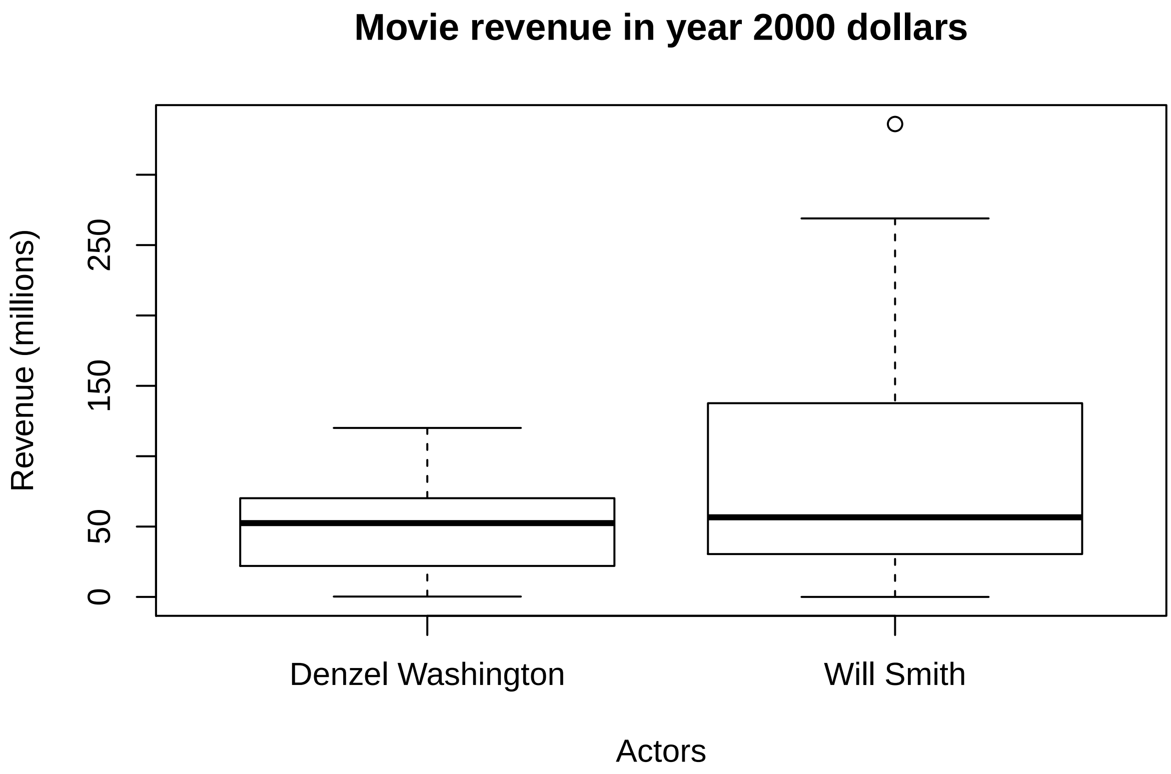


Figure 4: Side by Side comparison of Will and Denzel after conversion

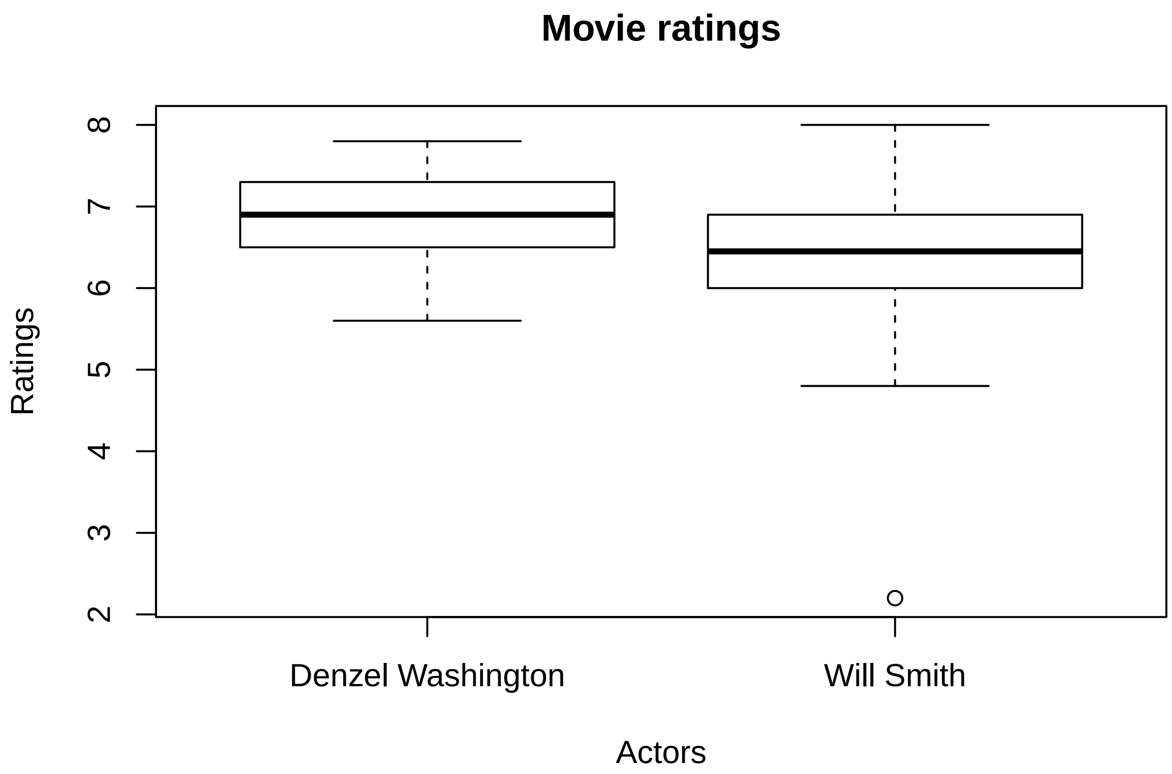


Figure 5: Side by Side comparison of ratings between Will and Denzel

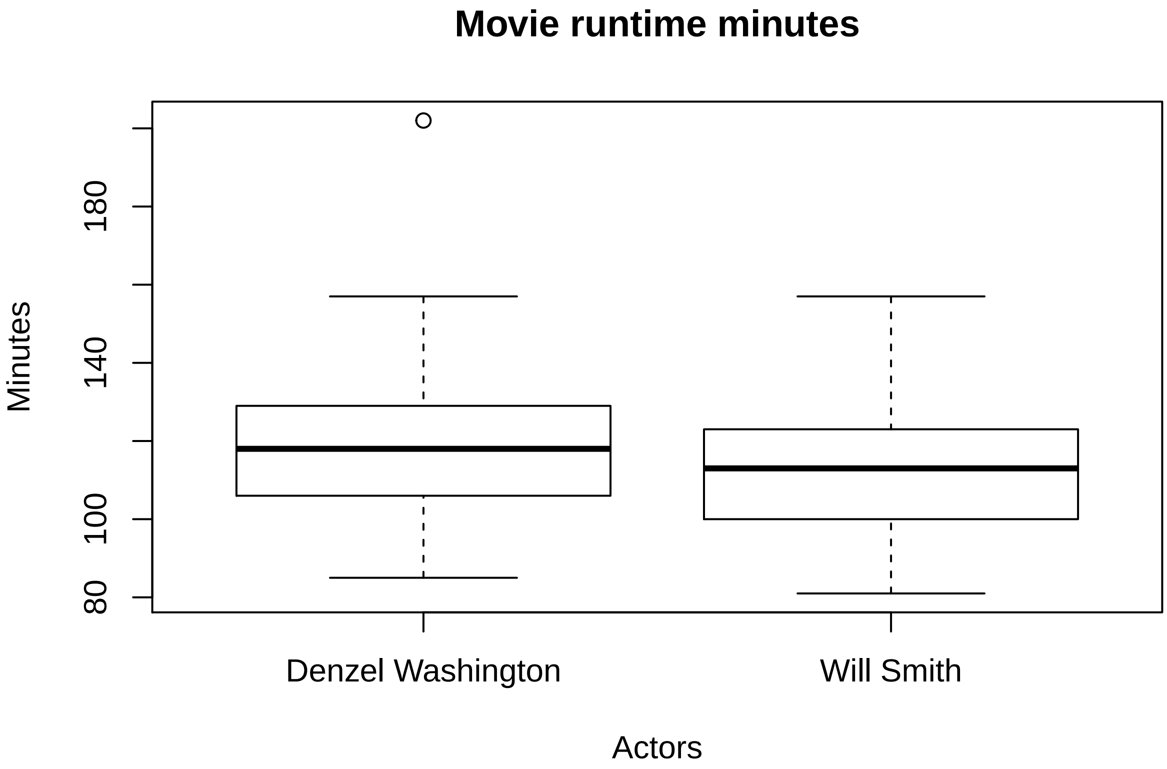
The actors have an average rating of 6.326 and 6.852, respectively. Fifty percent of ratings data for Denzel Washington is between 6.525 and 7.3. Fifty percent of the ratings data for Will smith is between 6.0 and 6.875. The max rating for Denzel is a 7.8 while Will Smith has a max rating of 8.0. From these statistics we can conclude Denzel Washington has been in more movies with higher ratings than Will Smith has.

```
# Ratings comparison
boxplot(denzel.converted$movies.50$ratings,will.converted$movies.50$ratings,
names=c("Denzel Washington","Will Smith"),main="Movie ratings",
ylab= "Ratings",xlab = "Actors")
```

7.3 Denzel vs. Will Side by Side comparison of runtime

Comparing the runtime from movies featuring Denzel Washington to Will Smith using a boxplot. We can see from Figure 6. the median for Will Smith is slightly lower than Denzel Washington. The actors have a median runtime of 113 and 118 minutes, respectively. Fifty percent of the runtime for Denzel Washington is between 106 and 129 minutes. Fifty percent of the runtime for Will smith is between 100.2 and 123 minutes. The max runtime for Denzel is 202 minutes while Will Smith has a max of 157 minutes. From these statistics we can conclude that Denzel Washington has been in movies with longer runtimes than movies featuring Will Smith.

```
# Runtime comparison
boxplot(denzel.converted$movies.50$minutes,will.converted$movies.50$minutes,
names=c("Denzel Washington","Will Smith"),main="Movie runtime minutes",
ylab= "Minutes",xlab = "Actors")
```



"'