STATS 419 Survey of Multivariate Analysis

Week 03 Assignment 02_datasets Revisited

Josh Pickel
(joshua.pickel@wsu.edu)

Instructor: Monte J. Shaffer

21 September 2020

0.1 Get Everyhting Set Up

```
library(devtools)
github.path = ’https://raw.githubusercontent.com/JoshuaPickel/WSU_STATS419_FALL2020/’;
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"))
# Create myMatrix
myMatrix = matrix ( c (
1, 0, 2,
0, 3, 0,
4, 0, 5
), nrow=3,
byrow=T);
```

1.1 Matrix 90 Degrees

```
# Rotating clockwise.
rotateMatrix90(myMatrix)
```

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

1.2 Matrix 180 Degrees

```
# Rotating Clockwise
rotateMatrix180(myMatrix)
```

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

1.3 Matrix 270 Degrees

```
# Rotating Clockwise
rotateMatrix270(myMatrix)
```

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

2 IRIS Plot

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

2.1 Plot (See next page for plot)

```
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)
```

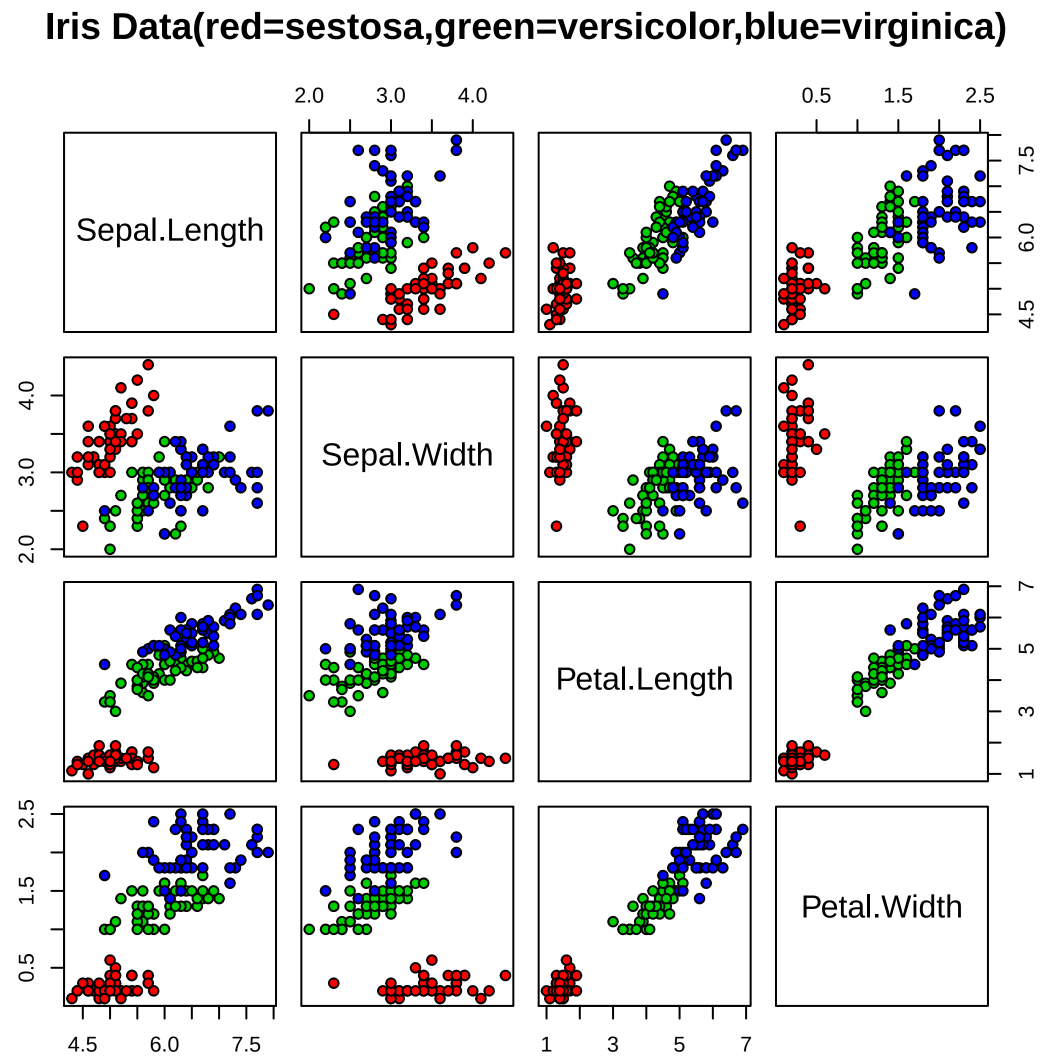

Figure 1: Plot From Plotting IRIS Data

3 IRIS Summary

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.

3.1 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.

4 Personality Data

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.

4.1 Clean Personality Data

```
source_url(paste0(github.path,"master/functions/functions-cleanup.R"))
files = paste0(github.path,"master/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)
head(personality_clean,3)
```

```
##                              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
##     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
##     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
##     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
```

4.2 Raw Dataset Dimensions

```
dim(personality_raw)
```

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

4.3 Clean Dataset Dimensions

```
dim(personality_clean)
```

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

5 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.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.

```
# Set up and get row for monte.shaffer@gmail.com
source_url(paste0(github.path,"master/functions/functions-custom.R"))
monte_row = personality_clean[which(personality_clean$md5_email ==
'b62c73cdaf59e0a13de495b84030734e'),]
```

5.1 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
```

5.2 doMode

```
doMode(monte_row)
```

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

5.3 doSampleVariance Naive

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

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

5.4 doSampleVariance Two Pass

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

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

5.5 Z-scores

```
zscores = zScores(monte_row)
plot(zscores$raw.scores,zscores$z.score,main = 'Raw Scores Vs Z-Scores',xlab = 'raw scores',ylab = 'z-s
```

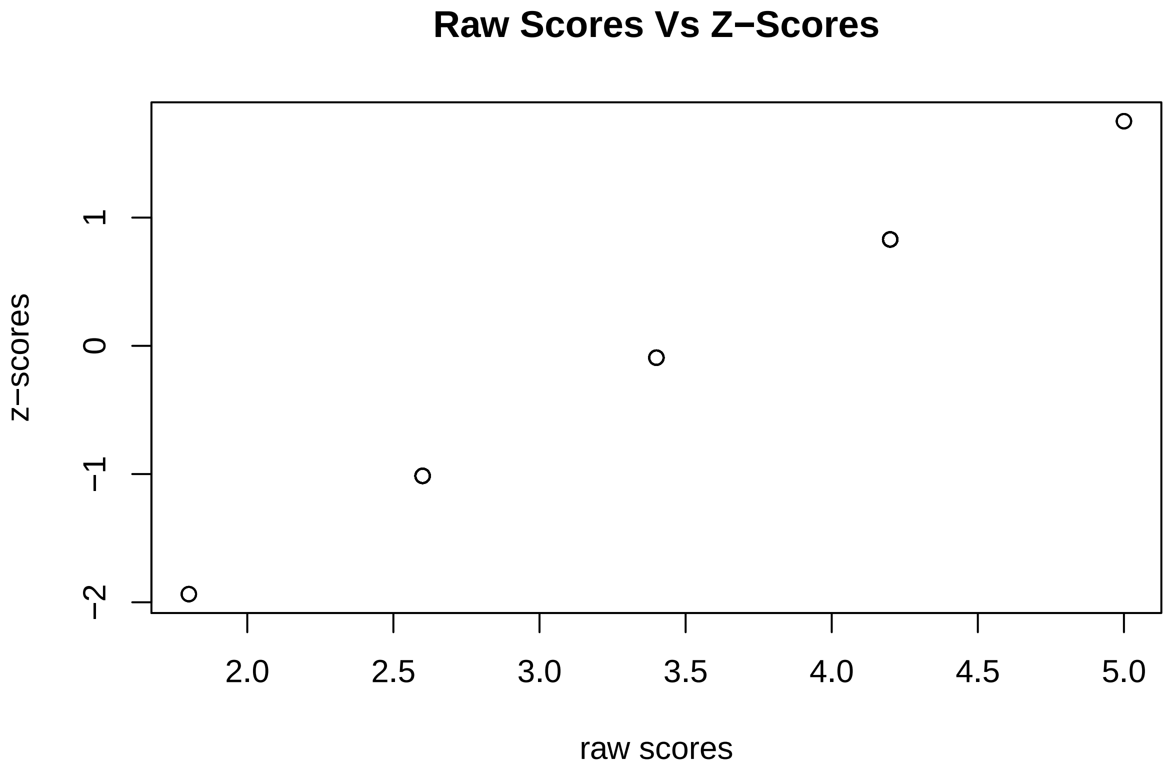

Figure 2: Plotting Raw Scores Against Z-scores

5.6 Response

After running the doSummary() on the record for 'monte.shaffer@gmail.com', it is apparent that the mode is 4.2, the mean is 3.48, there are no NAN values, the length is 64 columns, the two pass and naïve variances are both .7528136 and the standard deviation is .8676483. Considering the plot created from plotting raw scores against z-scores we can conclude from Figure 9 that there is a strong positive linear relationship between raw scores and z-scores. This is because the z-score is calculated from the raw values, meaning the two are highly correlated.

6 Will vs Denzel

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.

6.1 Get Actor Data And Compare Boxplots

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

#Get data for actors
nmid = "nm0000226"
will = grabFilmsForPerson(nmid)

nmid = "nm0000243"
denzel = grabFilmsForPerson(nmid)

#Plot actors
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" );
```

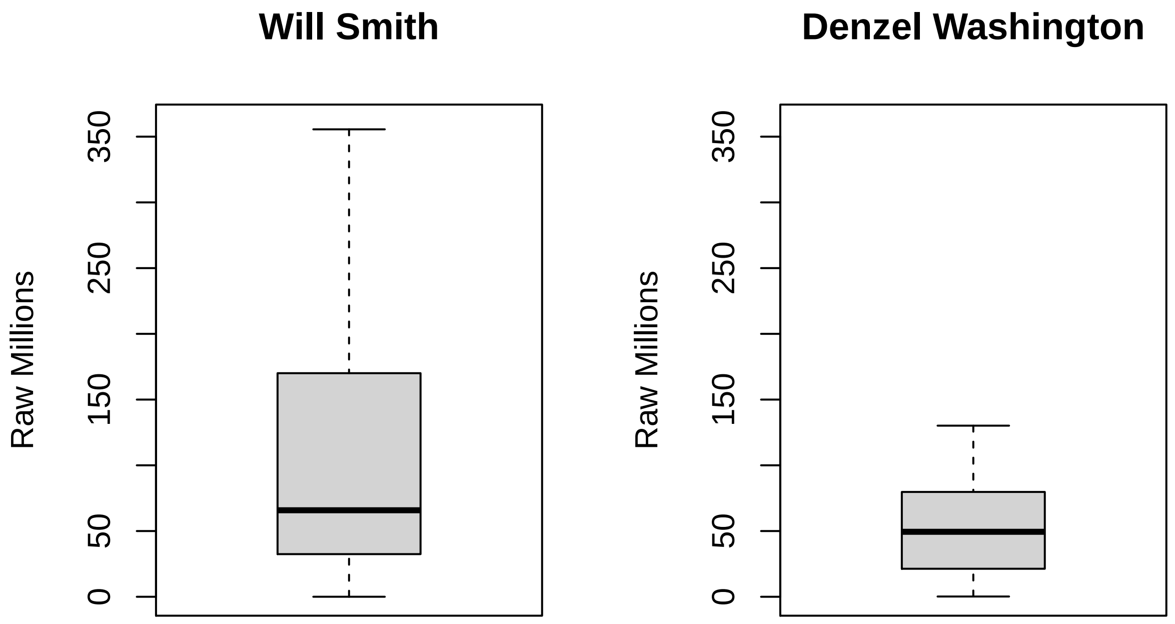

Figure 3: Denzel and Will Raw Millions

6.2 Response

Comparing Will SMith and Denzel Washington's raw million revenue values we can tell that 50% of the movies Will Smith is in covers a bigger range than 50% of the movies Denzel is featured in. We can also tell that Will Smith was in a movie that generated over $300 million. In general, we can see that the median revenue is the similar between Will Smith and Denzel Washington with Will SMith having a higher median. It can also be determined that Will Smith has a higher maximum revenue value than Denzel Washington.

6.3 Adjust Denzel and Will Million Values

```
# Get inflation data
source_url(paste0(github.path,"master/functions/functions-inflation.R"))
inflation_data = grabInflation()

will = convertDollars(will,inflation_data)
denzel = convertDollars(denzel,inflation_data)

head(will$movies.50)
```

```
##   rank                     title      ttid year rated minutes
## 1    1              I Am Legend tt0480249 2007 PG-13     101
## 2    2            Suicide Squad tt1386697 2016 PG-13     123
## 3    3         Independence Day tt0116629 1996 PG-13     145
## 4    4             Men in Black tt0119654 1997 PG-13      98
## 5    5                 I, Robot tt0343818 2004 PG-13     115
## 6    6 The Pursuit of Happyness tt0454921 2006 PG-13     117
##                        genre ratings metacritic  votes millions mill.2000
## 1   Action, Adventure, Drama     7.2         65 675726   256.39  212.9349
## 2 Action, Adventure, Fantasy     6.0         40 588655   325.10  233.2524
## 3  Action, Adventure, Sci-Fi     7.0         59 520939   306.17  336.0260
## 4  Action, Adventure, Comedy     7.3         71 507974   250.69  268.9646
## 5      Action, Drama, Sci-Fi     7.1         59 491852   144.80  131.9987
## 6           Biography, Drama     8.0         64 438552   163.57  139.7160
```

7 Will Vs Denzel

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.

7.1 Comparing Will vs Denzel Year 2000 Millions

```
boxplot(denzel$movies.50$mill.2000,will$movies.50$mill.2000,
        main = 'Will vs Denzel Year 2000 Revenue (in millions)',
        names = c('Denzel Washington', 'Will Smith'),
        xlab = 'Actor' , ylab = 'Revenue(Millions)')
```

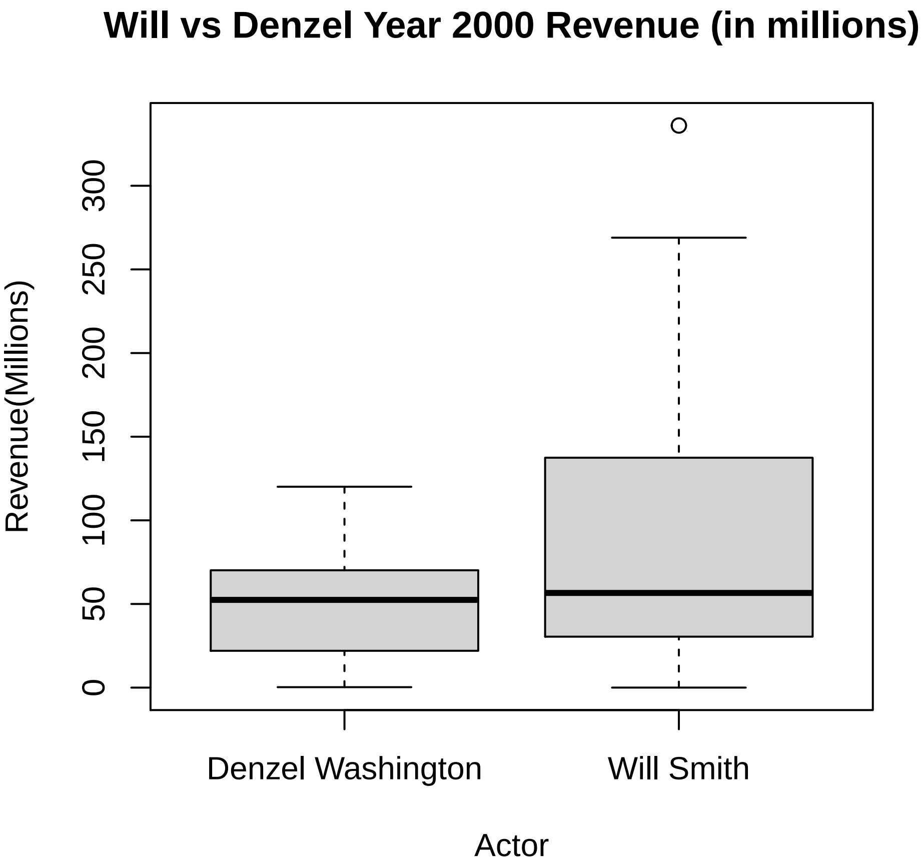

Figure 4: Boxplot Comparing Will vs Denzel Using Adjusted Millions

7.1.1 Response

Building a boxplot comparinga djusted revenues between Will Smith and Denzel Washington (Figure 4) we can tell that Will Smith has a slightly higher median revenue than Denzel Washington. We can also conclude that fifty percent of Will Smith's revenue values are between $30.45 million and $137.66 million and fifty percent of Denzel Washington's revenue values are between $22.39 million and $70.06 million. This shows us that Will Smith's revenue values are spread out more when compared to Denzel Washington's' values.

7.2 Comparing Will vs Denzel Years

```
boxplot(denzel$movies.50$year,will$movies.50$year,
        main = 'Will vs Denzel Year',
        names = c('Denzel Washington', 'Will Smith'),
        xlab = 'Actor' , ylab = 'Year')
```

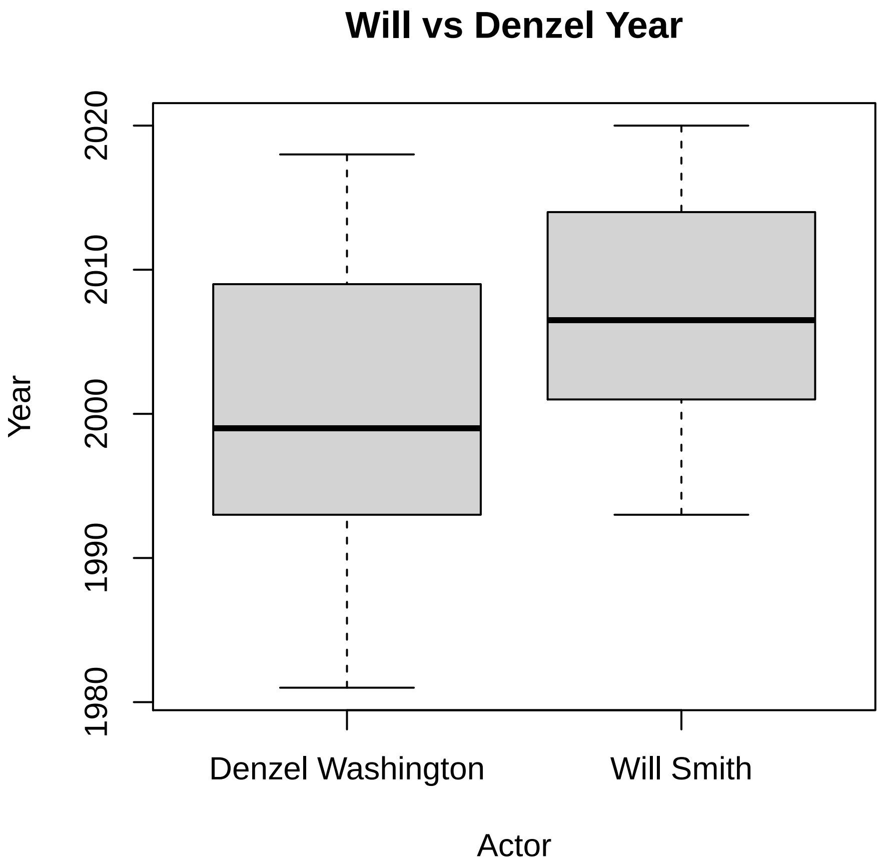

Figure 5: Boxplot Showing Will vs. Denzel Years

7.2.1 Response

Comparing Will Smith and Denzel Washington's years using a boxplot (Figure 11), there are serval key findings. We can conclude that, for this data set, Denzel Washington appears in older movies when compared to Will Smith. We can also determine that Denzel Washington's most recent film was before 2020, while Will Smith's most recent film was in 2020. We can also conclude that half of the movies Will Smith has been featured in were after 2006, and the other half before 2006. Likewise, we can determine that half of the movies Denzel Washington has appeared in were after 1999, and the other half before 1999.

7.3 Comparing Will vs Denzel Runtime

```
boxplot(denzel$movies.50$minutes,will$movies.50$minutes,
        main = 'Will vs Denzel Runtime',
        names = c('Denzel Washington', 'Will Smith'),
        xlab = 'Actor' , ylab = 'Runtime')
```

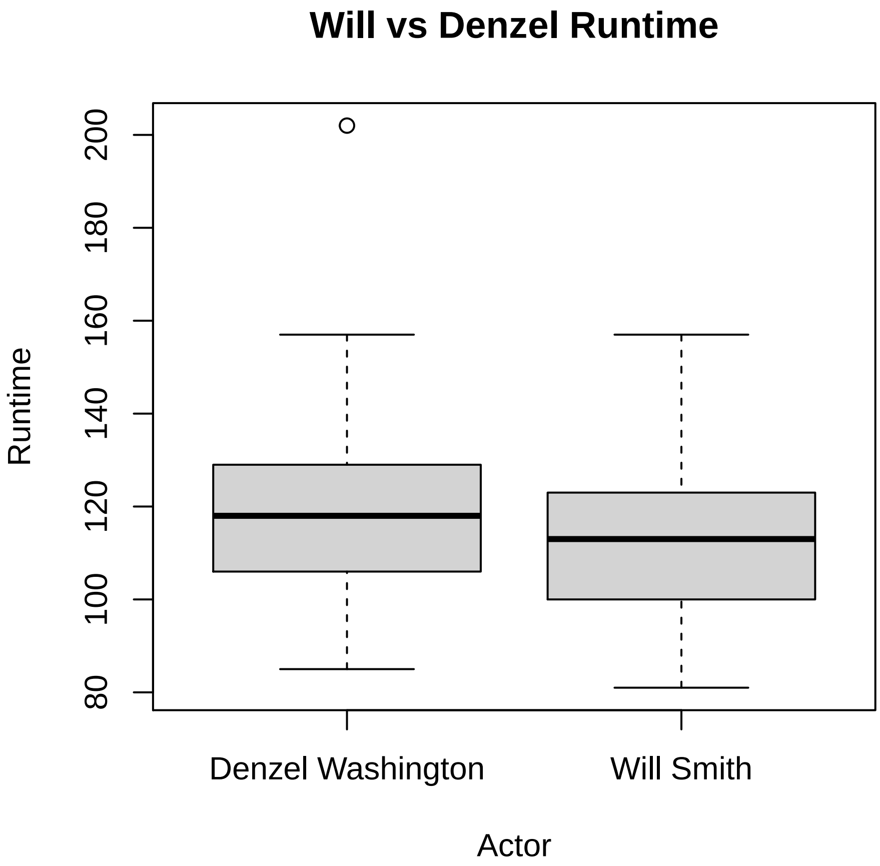

Figure 6: Boxplot Comparing Will vs Denzel Movie Runtime

7.3.1 Response

Using a boxplot to compare Will Smith and Denzel Washington's movie runtime(Figure 12), we can conclude that Denzel Washington has appeared in the longest movie in the data set. We can also conclude that Will Smith appeared in the shortest movie in the data set. Furthermore, we can conclude that the upper half of Denzel Washington's runtime values are higher than Will Smith's runtime values.