STATS 419 Survey of Multivariate Analysis

Week 03 Assignment 02_datasets

Nathan Shine
(nathan.shine@wsu.edu)

Instructor: Monte J. Shaffer

16 September 2020

1 Matrix

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

```
myMatrix;
```

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

```
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
```

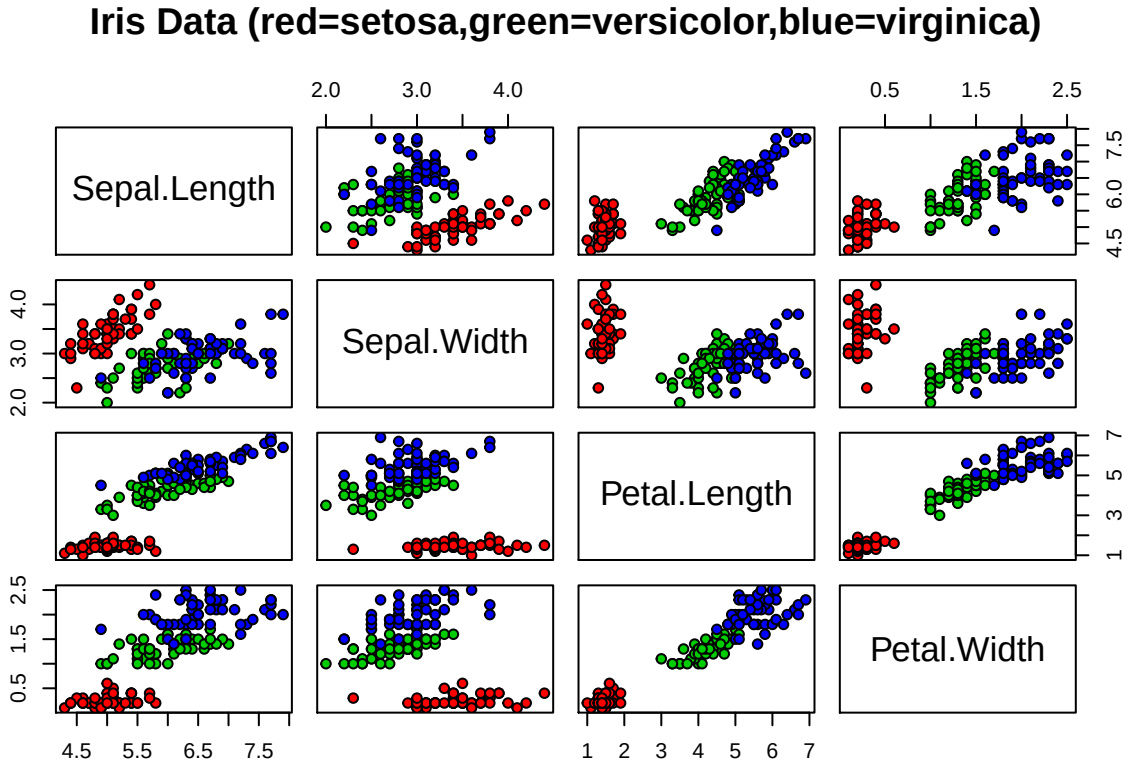
```
rotateMatrix270(myMatrix);
```

```
##      [,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

```
pairs(iris[1:4],main="Iris Data (red=setosa,green=versicolor,blue=virginica)", pch=21,
  bg=c("red","green3","blue")[unclass(iris$Species)])
```



The iris dataset contains 50 observations from three types of the iris flower (Setosa, Versicolor, and Virginica). Each observation has the petal length, petal width, sepal length, and sepal width recorded in centimeters. Ideas came from: https://www.kaggle.com/arshid/iris-flower-dataset

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.

Raw Personality records:

```
dim(df_raw)[1];
```

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

Cleaned Personality records:

```
dim(df)[1];
```

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

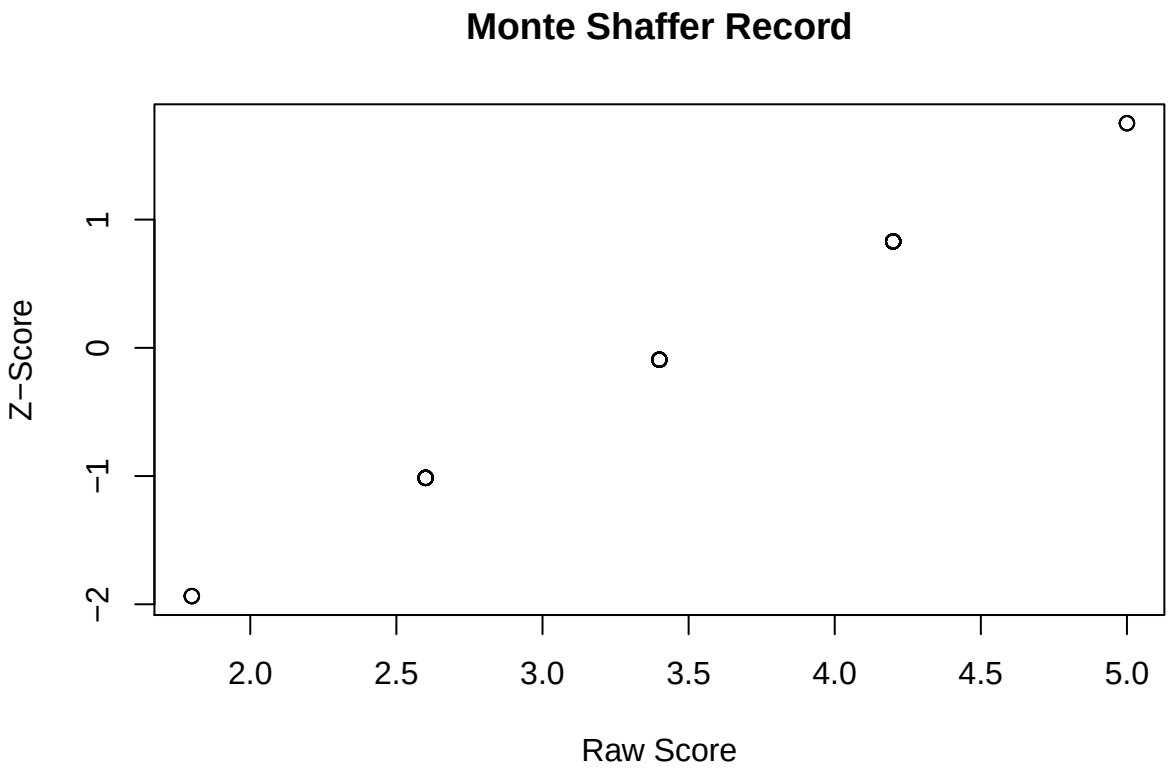
4 Variance and Z-scores

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

```
##           Length            NumNA             Mean           Median
##       60.0000000        0.0000000        3.4800000        3.4000000
##             Mode              Sum            SumSq         Variance
##        4.2000000      208.8000000      771.0400000        0.7528136
##        BuildInSD CustomSD.Variance
##        0.8676483        0.8676483
```

4.2 Z-Scores



There appear to be not many points on this graph, but many of the z-scores are the same so there are many points on top of one another. The reason many have the same z-scores is because the user can only choose between six values on the survey, even though on this record only five of the values were picked.

5 Will vs. Denzel

5.1 BoxPlot of Top-50 movies using Raw Dollars

5.2 Side-by-Side Comparisions

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.

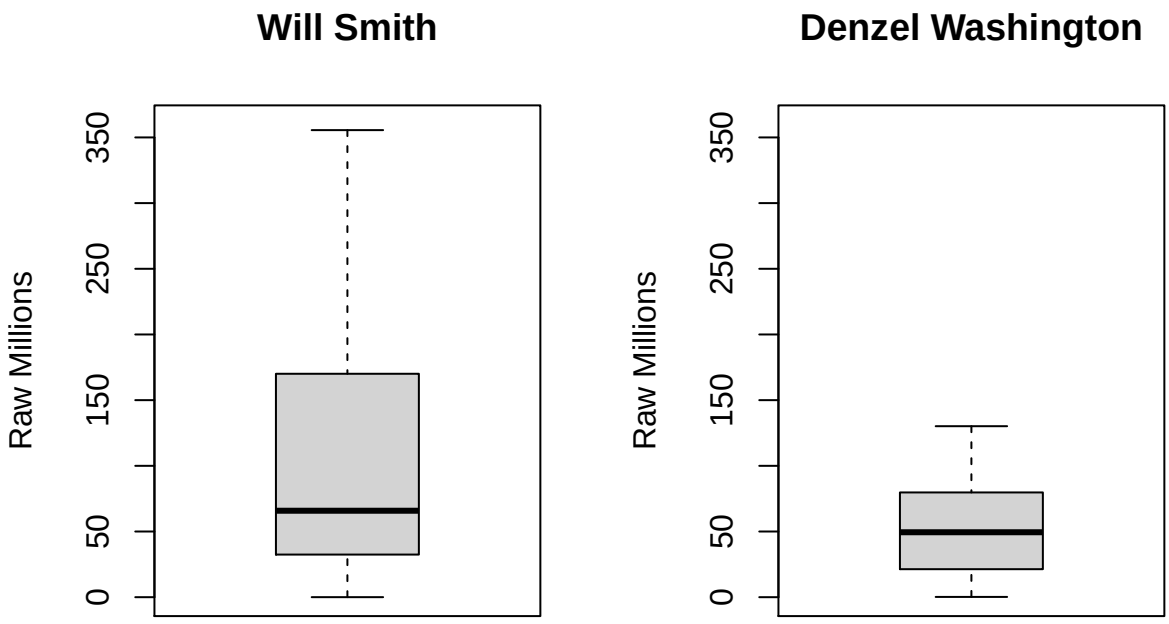
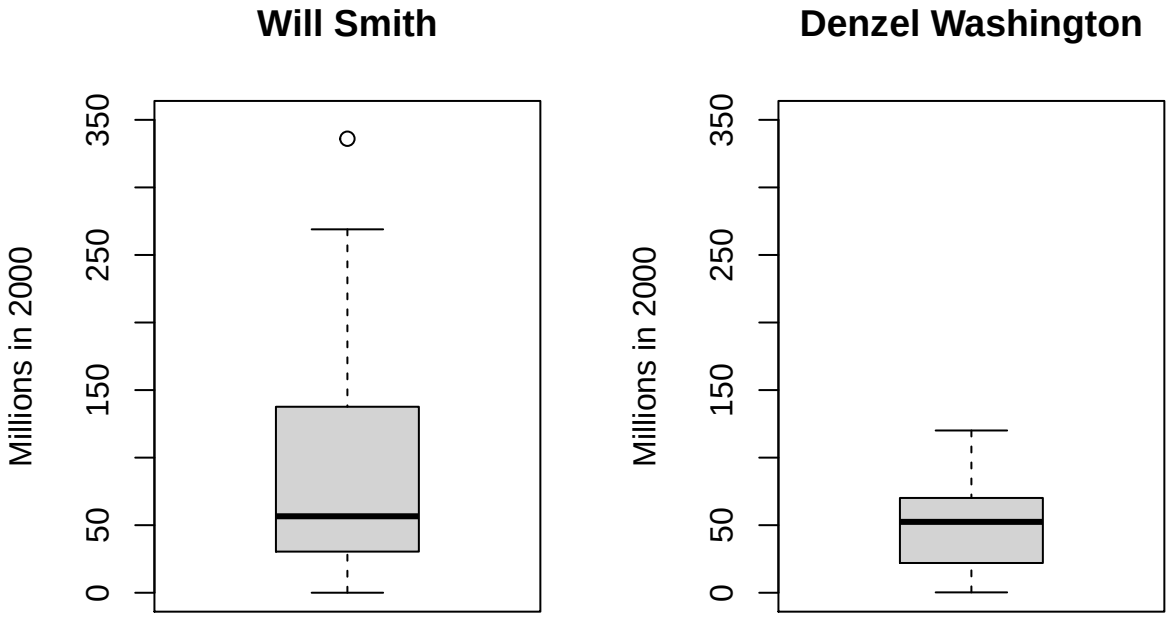


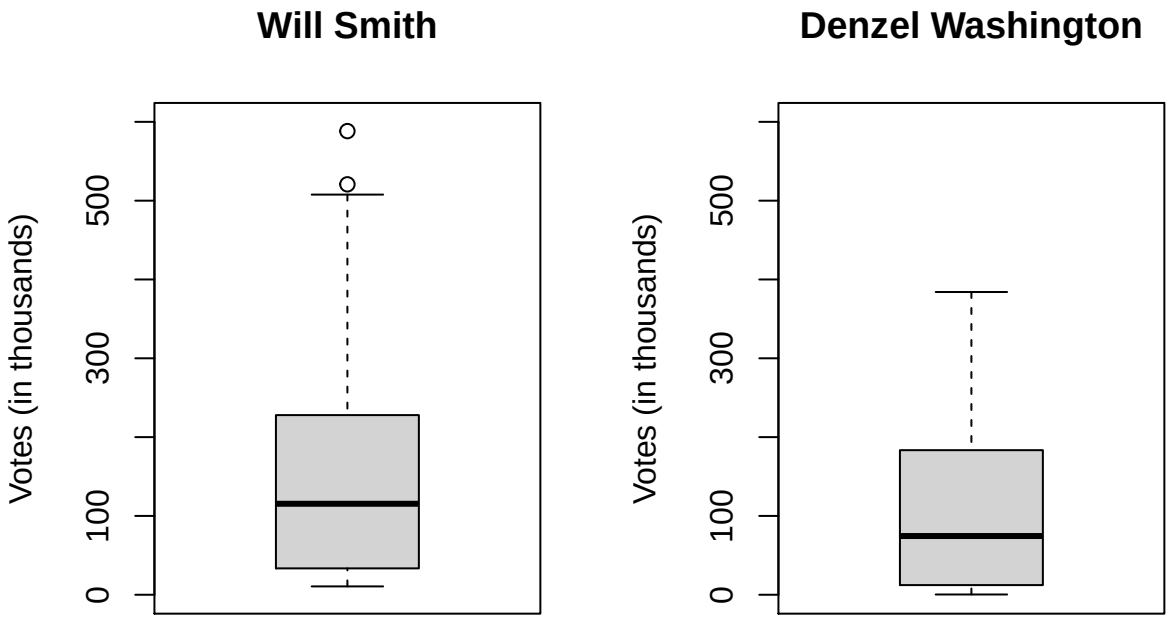
Figure 1: Raw Millions for Will Smith and Denzel Washington: IMDB(2020)

5.2.1 Adjusted Dollars (2000)



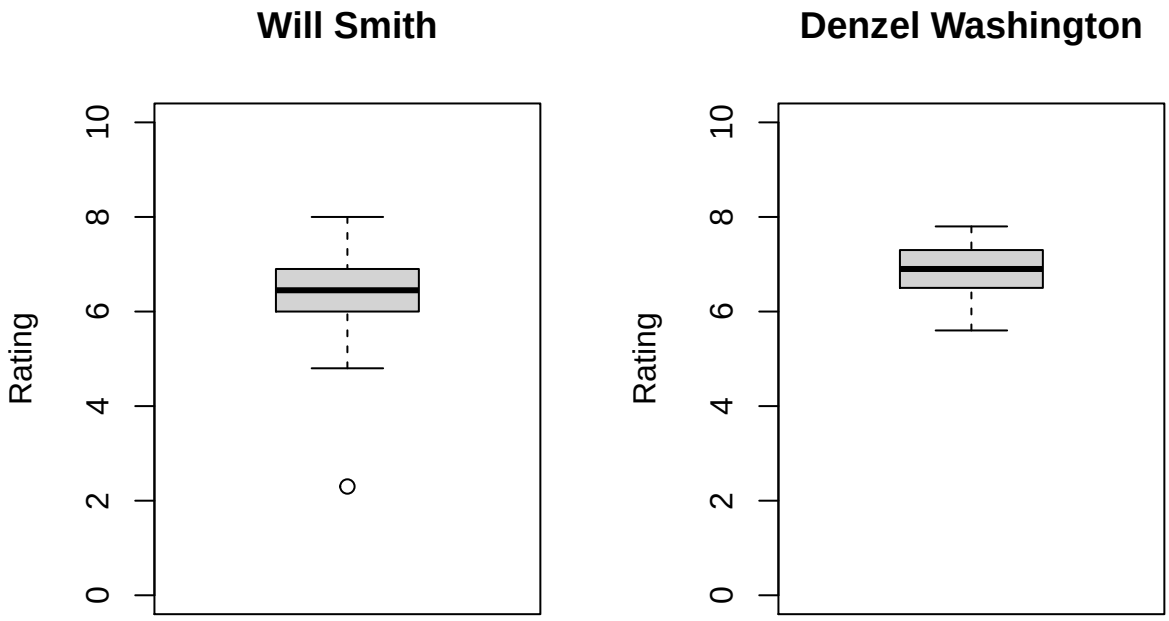
Here we can see that although they have around the same average, the movies Will Smith has starred in have a lot more box office potential, than the movies of Denzel Washington. The upper quartile of Smith is higher than the maximum of Washington. The lower quartiles for both are probably comparable as well as their minimums.

5.2.2 Total Votes (Divide by 1,000)



From this plot we can see that the movies of Will Smith generally receive more votes than the movies of Denzel Washington. We see this because the mean and upper quartile of Smith's movies are notably higher than those of Washington's.

5.2.3 Average Ratings

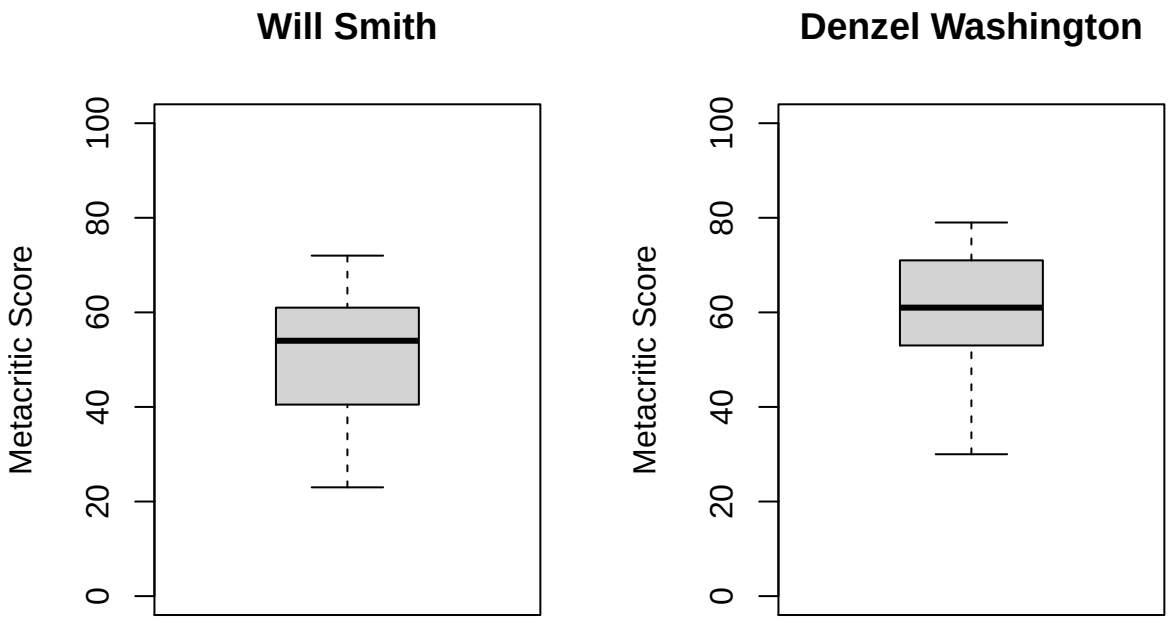


Here we can see that Denzel Washington's movies have rated about a 7, a half point higher on average than Will Smith's movies which would be about 6.5. We can also tell that both actors movies generally do not rate too far apart, but the maximum and minimum for Will Smith have a great up and down. Also curious to note is the lowest rated movie "Student of the Year 2," which is an Indian movie that Will Smith made a special appearance in.

5.2.4 Runtime

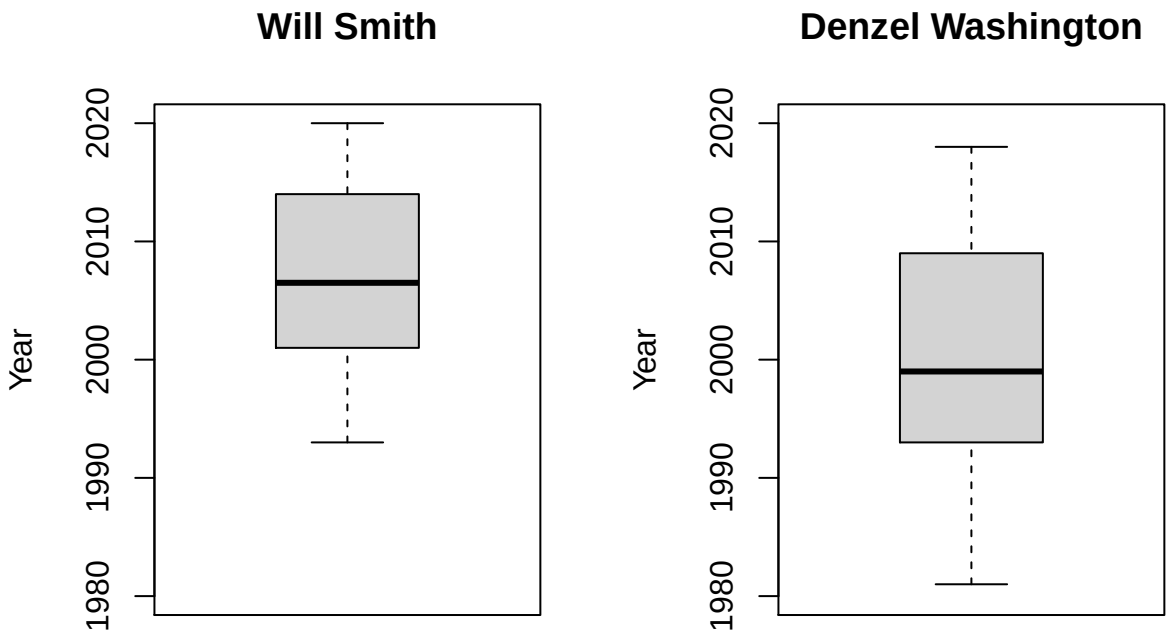
Lastly, we can look at the runtime to find that Denzel Washington's movies run about ten minutes longer than Will Smith's movies. However, both have a minimum and maximum that is relatively the same. The one exception to this is Washington's role in the movie "Malcolm X," which has a runtime of 202 minutes.

5.2.5 Metacritic



Looking at the Metacritic Scores confirms what was gathered from the IMDB ratings. An average Denzel Washington movie rates higher than the average Will Smith movie by about 5 points on a 100 point scale (or 0.5 on a 10 point scale). In addition Will Smith's movies only have a slightly higher upside than the average while having a much lower floor (lower quartile). The reverse is true for Denzel Washington's movies, which have an upside which is about equivalent with the highest rated Will Smith movie, and they have a higher floor which is about equivalent to the average of Will Smith.

5.2.6 Year



This plot shows us that Will Smith has been more active recently than Denzel Washington. It also shows that Denzel Washington has been in the movie business for much longer than Will Smith. Most of Will Smith's movies were released between 2000 and about 2015, while Denzel Washington's most active period was from the 1990s and 2000s.