I would strongly recommend that you provide an overview of the two actors, one strong summary table, and one graphic that aligns with your final answer.

# \Abstract

# \Introduction

The IMDB database is a public repository containing information related to movies and actors, including cast, crew, ratings, and earnings data. This project attempts to investigate, compare, and draw conclusions regarding two well-known actors by performing multivariate analysis on the data gleaned from the IMDB website. Previous work into the subject has attempted to evaluate this by considering box office sales and IMDB movie rankings. It was shown that the median box office sales, when adjusted for inflation, were about the same for the two stars. The median Metacritic ratings for each actors films were also similar and did not enable a meaningful distinction. As the results were inconclusive, further investigation is needed to provide deterministic insights into which actor is "better". This project will look at multiple variables and perform a mathematical analysis to attempt to quantitatively evaluate which actor is better using the available data features. This an important demonstration of how data analysis can be used to empirically inform a response to an otherwise hard to answer and somewhat subjective question. The author has no predisposition toward either actor, and the information presented is intended to provide an unbiased and data driven perspective.

# \Overview

Denzel Washington and Will Smith are two well-known movie stars with comparable ratings and extensive filmographies. Denzel is 66 years old and made his first movie in 1981, while Will Smith is 52 and made his first movies in 1992. Will Smith has 111 films listed on IMDB with a mean rating of 6.2, while Denzel has 61 films listed with a mean of 6.8. Both actors are listed in the top 500 on IMDB.

The IMDB data contains metrics on movies in which the actors have appeared including ratings, box office sales, and Metacritic scores, as well as info on the cast and crew. For actors, their rating within each movie is available, allowing for the determination of whether their role was a lead or whether they played a less prominent character. Demographic data and star meter rank was also available for each actor.

The concept of which actor is "better" is an inherently subjective measure, and the choice of which factors to consider in the analysis was given careful consideration. Calculating a diversity score from an analysis of the gender breakdown of the cast and crew for each movie was suggested. However, different genres can have different casting requirements, and the outcome of any analysis may be more indicative of which genre an actor prefers than anything else. A preliminary assessment of the diversity index of Will Smith movies, for example, showed a wide variation in score across genres. A strong correlation between gender diversity and acting acumen appears unlikely, and gender/diversity score was not selected as a component of the analysis.

Another available factor was the "star meter" ranking, a measure of popularity created by IMDB that is a function of the number of credits a person has, popularity of the work a star appears in, and traffic to the celebrity's profile. While this initially seemed promising, after closer inspection, many of the actors used to build the data set had identical ratings. This rating also appears as a static snapshot that reflects the score today, rather than at the time a particular movie was released, and it's unclear how these ratings might change over time. This metric was deemed unreliable and non-deterministic, and was not included in the final analysis.

The analysis required building a dataframe for each actor with the desired covariates. For each film an actor was associated with, the movie ratings, Metacritic score, box office sales, and actor rank were selected. The decision was made to only consider movies where the star had been a headliner, defined as having an actor rank of 1, 2, or 3 for a given movie. This was done in an attempt to make sure the covariates were a mainly a function of the actor being considered, not some one else who may have been in the movie and had a larger role. The assumption was made that a "better" actor would be appear as the lead more often than in a supporting role, so the ratio of leading roles to total movies was also considered. Box office sales were adjusted to 2020 dollars, and the mean of the ratings, scores, and sales for each movie were calculated.

To obtain a more robust analysis, Will and Denzel should be compared to more actors than just each other. A pool of 48 contemporaries was drawn from the actorRank2000.rds provided by the instructor, ensuring only candidates from the modern era were considered. Actor ID's were randomly selected and screened for quality. The criteria were at least 10 movies where the actor ranked 3 or lower, and complete data for ratings, Metacritic scores, and box office sales.

Using the entire pool of 50 actors, the data was standardized by dividing each value for a covariate by the maximum value for that feature found in the pool. The result was a 50 X 4 matrix where each value was a positive integer between 0 and 1. An adjacency matrix was calculated where the value of the eigenvector in each row/column is the rating of the actor evaluated against the other actors in the pool for that factor. From these eigen-ratings, an empirical evaluation of which actor is "better", Will or Denzel, was able to be obtained.

actor rank in movie rank, box office, movie rating

Movies where the ators role was significant enough to assign a rank.

Chose to look at z score of covariates where will had a major role ( 1, 2,or 3)

Movies where each star had a a leading role (rank <3) were considered. The attribites of metacritic score, rating, and inflation adjusted box office sales were considered relevant to determining which actor is "better". The table shows the mean of each category for comparison, demonstrating there is no clear winner between the two actors.

The z scores for these covariates were calculated for the pool consisting of both actors movies. The average of each actor's z scores for each factor were found. Would it be better to find the z score of these samples when compared to the mean of the entire population (the large pool of movies from the database?)

From the popular 50 list, I got all the movie ttids from the last 50 most popular movies in the last 20 years. I merged this with the all actors movies list to get the metacritic socres, box office info, etc for thos movies.

# \Key Findings

actor rank in movie rank, box office, movie rating

Movies where the ators role was significant enough to assign a rank.

Chose to look at z score of covariates where will had a major role ( 1, 2,or 3)

Decided cast size wasn't significant

# \Table

Data from 1980 - 2018 (subset by year, 40 years is 1980 to 2018?)

merged cast to wills movies based on movie name to get all the people will has worked with. Called unique, appended the biography. Summarize the cast data in a table. What is the score/ranking for his costars? Average score/star ranking for each movie? Box office earnings for each movie? Box plot of earnings for each movie? Can look at gender disparity by cast?

# \Results

# \Conclusion

For further research: (in the discussion or conclusion)

Could you do a sentiment analysis of comments on will and Denzel's pages?

# \Brainstorming

# will.movies; has all the rankings of his movies in several ways, but not rankings of his costars or gender diversity score.

#From Will's rankings on some films, can we make use of how many times he is number 1 or 2?

could look at subset of how many times each was one or two: Could look at proportion of leading roles.

For the subset of movies where the actors were the lead, we could look at average adjusted box office of those movies, total adjusted box office, average gender equity score, average cast size? and average movie rating, The result would be an eigenscore.

Could build an adjacency matrix of the features I pick,

Can compare my results to the actor actor or actor movie adjacency matrix results and discuss.

take the dataframe and do z scores on each column so it's "normalized". Instead of giving will smith a 60 of 110 , you could use a z score of all the actors and how they compare the top movies.

actor rank in movie rank, box office, movie rating

Movies where the ators role was significant enough to assign a rank.

Chose to look at z score of covariates where will had a major role ( 1, 2,or 3)

can go to headliners dataframe and see how many each one were headliners.

The main event:

Fish in a barrel, need to provide context:

, merge data frame with any variables like cast, or movie dollars, and do rankings and create Z scores. (The zscore is a measure of how far from the mean a data point is, to phrase it

## informally) Compare will vs Denzel in the "pond" of the last 40 years. The example compares cast, I can do it with other variables.

line 517 in IMDB graph ranks, I have the ttids, can pick whatever.

Get all the coluns, get all the ranking dollars that we've looked at in the past for the entire pond, create z scores, normalize it, and then look at will versus denzel in the context of the envrinment in which they work,,

We will build MM a matrix of movies linked to movies (via actors).

A movie has a direct link is an actor connects two movies together, this is "zero-degrees" of separation. ``MM%\*%MM`` would represent "one-degrees" of separation;

movieranks2000.rds: we are seeing top movies ranked by the "star power" on a given film. It's googles page rank essentially applied to actors. So Tom cruise has a 100 because he's been in the most popular movies and is connected to other people who have been in the most popular movies and so on.

We could use this result and see the movies that are linked to Will vs. Denzel. Plus we can transform "money" to a common year and do rankings as discussed. In this larger "pond" and see where Will vs Denzel fit

Could discuss the average move rank from the adjacency matrix. Or combine the movie rank from the adjacency matrix with the will.movie data.

Could build a table with two rows, with the columns I decide, one row for each actor. Columns could have a 0 or 1, 1 for the "winner". Columns could be diversity score, average movie rank, highest movie rank, ratio of leading roles, etc. other interesting variables. Then could do an adjacency matrix similar to the GRIMM stuff to see which actor ranks higher when all those factors are considered.

entire film crew? Do films with more female stars have higher diversity scores? What about films with female producers? What factors allow you to predict diversity? If Will or Denzel were the a predictor, would that make them a better actor? It could be surpised that if that were the case, a higher diversity score would indicate that the leading man had more influence over production, whcih could be considered a measure of sucess. But do we know that Will or Denzel are feminist and would advocate for equality even if they had the "pull"? There are issues to consider here that are beyond the scope of this assignment. Plenty of opportunities for interesting analysis exist, but for determining who is a "better" actor, the issue of diversity was not considered.

Intro

Overview

overview of the actors

Maybe the means table goes here?

overview of how I define "best"

why I chose to look at the factors I chose

Why I didn't choose gender, etc

A description of the process, and what I did

Results

Discussion of results

show some of the output?