# Project Part 3

Gideon French

Is there a relationship between movie scores and which streaming platform it is available on?

## Description of Data

### Netflix, Hulu, Disney Plus, Amazon Prime Movies and TV Shows

These datasets were all created by the same user on Kaggle. They contain data on the movies and shows available on four of the biggest streaming platforms for movies: Netflix, Hulu, Disney Plus, and Amazon Prime. We will only be using the data relating to the movies on these platforms. The data is sourced from Wikipedia, Unofficial DB Search, and FlixDatabase, and was obtained through web scraping and API calls. This data will be used to subset the other datasets by streaming platform. Each dataset contains over a thousand entries. The data is of the entire population of movies available on each streaming platform up to the end of 2021.

### Movie Ratings Dataset

The Movie Ratings Dataset on Kaggle consists of information relating to the movies on the IMDB website since 2000 by IMDB score. This data is normally distributed, as opposed to the IMDB top 1000 dataset. The data in the dataset was obtained by web scraping the IMDB website. IMDB (Internet Movie Data Base) is a website for aggregating, among other data, a movie's critical reception using scores. IMDB scores on movies are a weighted average of user votes on the quality of the movie from 1 to 10. I subset this data by the movies available on select streaming platforms for analysis. I removed outliers.

Sources:

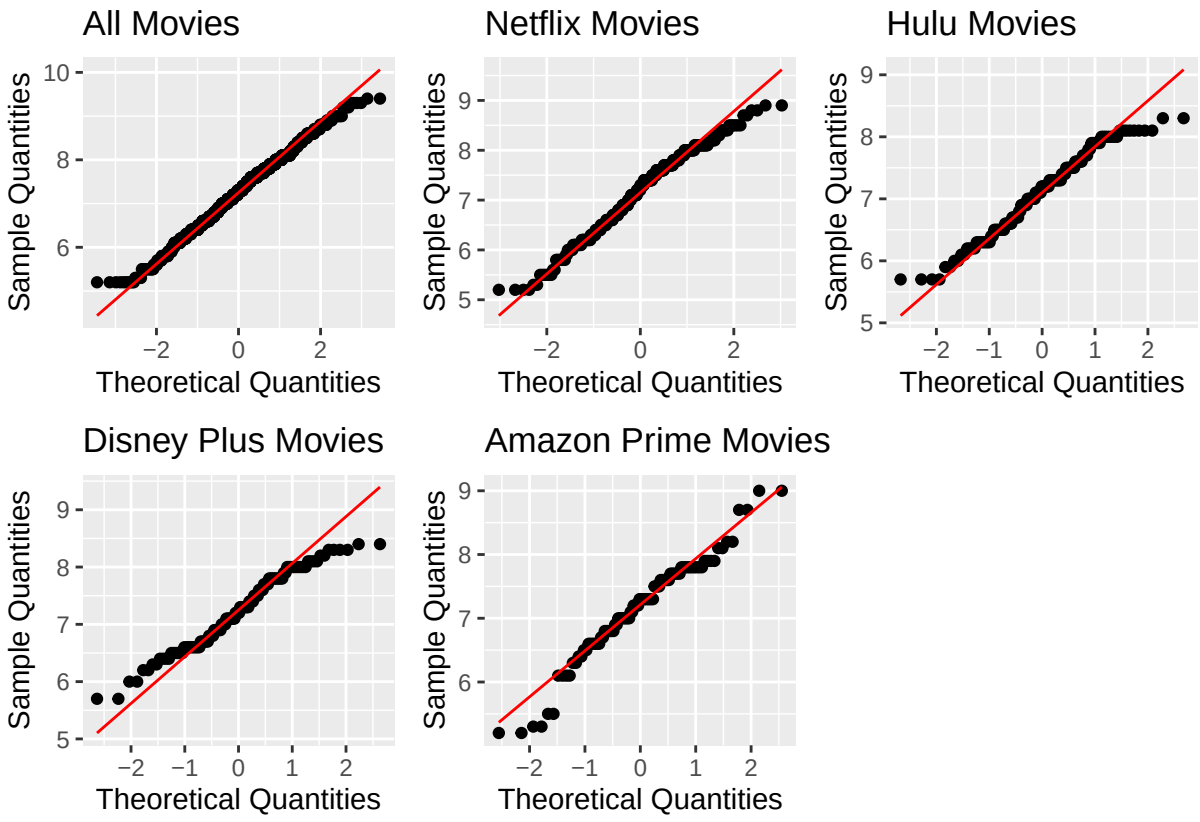
- "Movie Ratings Dataset" on Kaggle by user *Devesh Kumar Rai*

- "Disney+ Movies and TV Shows" on Kaggle by user *Shivam Bansal*
- "Netflix Movies and TV Shows" on Kaggle by user *Shivam Bansal*
- "Hulu Movies and TV Shows" on Kaggle by user *Shivam Bansal*
- "Amazon Prime Movies and TV Shows" on Kaggle by user *Shivam Bansal*

## Test

I will perform a two-sample t-test to test whether the mean for movie scores and revenue on each streaming platform is significantly different from the mean for all movies (excluding each streaming platform). This test is appropriate for my research question and data because I am investigating the difference between average scores in two groups, based on the movies' presence on a platform or not. I do not know the full population standard deviation, making a t-test the ideal choice. The two-sample t-test can be used to compare the means of two groups that are independent, approximately normally distributed, and have a similar amount of variance within each group, with no excessive outliers.

### Normality



Using the Q-Q plot, the points should align in a straight line if the data is normally distributed. The data from each streaming platform appears to be roughly normally distributed and the points appear aligned on the red line.

### Independence

The observations in one sample must be independent of the other sample. We guarantee this by drawing out two samples from one group from that streaming platform and one from a group of all movies excluding those available on the streaming platform.

We will be using the Welch t-test which does not require homogeneity of variances.

### Outliers

I removed outliers from the total scores when filtering the data.

### Homogeneity of Variance

The t-test being used by default, Welch's t-test, does not require homogeneity of variance.

### Test Hypotheses and Results

We will test the difference in means between all movies (excluding those available on each streaming platform) and movies available on each streaming platform. Our null hypothesis is that the difference in means is equal to 0. Our alternative hypothesis is that the difference in means is not equal to 0.

```
n_p <- t.test(netflixratings$imdb,notnetflixratings$imdb)$p.value
h_p <- t.test(huluratings$imdb,nothuluratings$imdb)$p.value
d_p <- t.test(disneyratings$imdb,notdisneyratings$imdb)$p.value
a_p <- t.test(amazonratings$imdb,notamazonratings$imdb)$p.value
ptable <- data.frame(PValue=c(n_p,h_p,d_p,a_p))
rownames(ptable) <- c("Netflix","Hulu","Disney Plus","Amazon Prime")
ptable
```

```
##                   PValue
## Netflix      0.011346174
```

```
## Hulu         0.009447451
## Disney Plus  0.754724611
## Amazon Prime 0.305765623
```

## Conclusions

We will use a significance level of 0.05.

With a p-value of 0.0113, we reject the null hypothesis that the difference between the mean IMDB score of movies on Netflix and the mean score of movies not available on Netflix is 0. There is evidence of a significant difference between the mean IMDB score of movies available on Netflix and those not.

With a p-value of 0.0094, we reject the null hypothesis that the difference between the mean IMDB score of movies on Hulu and the mean score of movies not available on Hulu is 0. There is evidence of a significant difference between the mean IMDB score of movies available on Hulu and those not.

With a p-value of 0.7547, we fail to reject the null hypothesis that the difference between the mean IMDB score of movies on Disney Plus and the mean score of movies not available on Disney Plus is 0. There is not evidence of a significant difference between the mean IMDB score of movies available on Disney Plus and those not.

With a p-value of 0.3058, we fail to reject the null hypothesis that the difference between the mean IMDB score of movies on Amazon Prime and the mean score of movies not available on Amazon Prime is 0. There is not evidence of a significant difference between the mean IMDB score of movies available on Amazon Prime and those not.

Using the two-sample t-test, we found evidence of a difference in the mean scores of movies on Netflix and Hulu to those movies not available on those platforms. We did not find evidence of a difference for Disney Plus or Amazon Prime. This means that the movies available on Netflix or Hulu are on average scored differently from movies not available on those platforms. Since we used a two-tailed test, these scores could be on average higher or lower.

These results should not necessarily be generalized to the entire population because my data is only concerned with movies released since the year 2000.

## References


1. IMDB Scoring Information https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV#.
2. Netflix Movies and TV Shows https://www.kaggle.com/datasets/shivamb/netflix-shows.
3. Amazon Prime Movies and TV Shows https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows.
4. Hulu Movies and TV Shows https://www.kaggle.com/datasets/shivamb/hulu-movies-and-tv-shows.
5. Disney+ Movies and TV Shows https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows.
6. Top Box Office Revenue Data - English Movies https://www.kaggle.com/datasets/kalilurrahman/top-box-office-revenue-data-english-movies.
7. IMDB Movie Dataset https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows.
8. Two-Sample T-Test Assumptions https://www.statology.org/t-test-assumptions.
9. Superimpose Normal Curve https://stackoverflow.com/questions/6967664/ggplot2-histogram-with-normal-curve.
10. Movie Ratings Dataset https://www.kaggle.com/datasets/raidevesh05/movie-ratings-dataset.
11. Outlier Detection https://www.r-bloggers.com/2020/08/outliers-detection-in-r/.
12. T-test in R https://www.datanovia.com/en/lessons/t-test-in-r/.