
Carlos Andrés Luna Leyva

Data analysis of movies in the streaming platforms Netflix and Disney+

13 January 2025

Table of Contents

1. Introduction

As the digital age advances, the rise of streaming platforms has cemented its spot at the forefront of media consumption in the average household. Of which, Netflix cemented itself as the strongest competitor by also being the first popular platform to gain traction as a household name. In an attempt to capitalize on the new market, the Disney Corporation launched their own streaming platform, Disney+, in which they hold a monopoly on their own franchises, seating themselves at the top of the popularity rankings due to their exclusive catalog. This data analysis aims to compare the movie catalogs of these two streaming platforms, and compare the overall ratings of their selections as well as their stance on age ratings, since Disney has been traditionally a family-oriented media company.

The data set was selected and provided by the Technical University of Dortmund, via the website Kaggle.com, initially uploaded by the user Ruchi Bhatia to the website (Bhatia, 2021). The approach for this analysis project will comprise of cleaning the data set for the wanted variables, separating movie titles and their respective data into platform-specific data set for independent assessment, and finally comparing the results of these assessments to their respective counterparts to formulate results. Data set manipulation is achieved using the following Python libraries: Pandas (McKinney et al, 2010) for data set manipulation, NumPy (Harris et al, 2020) for array methods, Matplotlib (Hunter, 2007) for creating plots, and SciPy (Virtanen et al, 2020) for the comparison method. This project will focus on the use of the Mann-Whitney U-Test (Mann & Whitney, 1947) to compare any significant differences between Netflix and Disney+ as streaming platforms.

2. Detailed Description of the Problem

2.1- Objectives

Netflix's popularity is mostly due to being one of the first media streaming platforms that provided the service of digitally streaming movies and shows on a subscription service, however competition has forced larger corporations who own the rights to franchises and media, such as Disney among other traditional media companies. Having being forced out of the monopoly over digital streaming, Netflix has lost portions of its movie selection to other streaming platforms, possibly affecting the overall quality of its catalog. One of the core objectives of this statistical analysis is to compare the Netflix catalog's ratings to Disney+, the relatively newer streaming platform born out of the Disney Corporation's own catalog, arguably one of the most recognizable media companies in the world.

Disney+ has created a brand image of family-oriented content first and foremost, however with recent acquisitions of franchises such as Star Wars and Marvel, as well as the desire to retain adult audiences in their streaming platform, not all of its content can be appropriately promoted to the youngest audiences. The second core objective of this statistical analysis will be to analyze any bias in the age ratings of the Disney+ catalog, as well as compare these to the Netflix catalog, a more diverse and well-rounded selection of movies marketed mostly to adult audiences.

2.2 – First Look at the Data

The data set used for this project was gathered from the website Kaggle.com, initially uploaded by the user Ruchi Bhatia. The original upload of this data to the website contains a description stating how the data was gathered using web scraping methods, however there is no mention of what methods were used or the direct sources of this data. While the publication has a usability score of 10.00 by the website, maximum score reviewed by Kaggle, there is no real assurance of the validity of how or where this data was gathered from. In a real-world scenario this could bring attention to the credibility of the data set and any statistical analysis gathered from it, however for the sake of this project as an exercise of data analytics the data set will be treated as legitimate and complete in all regards.

The data set contains 9515 different movie titles available across the platforms Netflix, Hulu, Amazon Prime Video, and Disney+, along with the movie name, and identifier number (ID), year of release, Rotten Tomatoes score, and an age rating. Due to this project focusing entirely on the platforms Netflix and Disney+, the columns Hulu and Amazon Prime Video will be discarded, as well as any movie titles not belonging to the Netflix or Disney+ catalogs.

PLATFORM	CATALOG COUNT
Netflix	3695
Disney+	922
Hulu	1047
Prime Video	4113

Figure 1: Total entries for each streaming platform in the given data set. It is worth noting the total adds up to 9777 due to some movies being shared across platforms.

Most variables come in a string format, such as Rotten Tomatoes score following the format of “XX/100” or age rating having a special character to denote the appropriate minimal age such as “13+”, making direct manipulation of the data difficult. To solve this, further cleaning of the data set will be required to allow for direct numerical manipulation.

Looking directly at the entire data set, the only missing data can be located in the Rotten Tomatoes and age rating columns. In the case of Rotten Tomatoes, less than 0.07% of the values are registered as NULL, allowing for the use of simple data imputation to recreate these values, introducing minimal data bias. However, in the age rating column a massive 44% of the values are registered as NULL. While there is a non-zero chance these missing elements could be part of the streaming platforms that will not be taken into consideration for this analysis, it is highly likely a significant portion of this column for either Netflix or Disney+ catalogs will be missing. Missing data makes any attempts of data imputation more complex and over the scope of this statistical analysis. To continue with the statistical analysis exercise, this column will be manipulated and compared as two separate instances: with a label for no age rating given, and without these missing elements. The purpose of this decision is to compare if any bias is formed or if any misleading inference could be made from either scenario.

3. Methods

3.1 – Data Cleaning

In order to facilitate data manipulation, the data set will be cleaned to only include the relevant columns and their information, as well as remove any unnecessary columns as to not bloat the data set with information that will not be used. The first removal is the Title column, as the actual name of individual movies will not be used for this analysis. Following this decision, the column ID will be kept in case further study would require a cross-reference to the original data set for the movie’s title, however for this statistical analysis this value will not be used. Similarly, the Year column will be removed as this data does not require to be used to meet our objectives. Lastly in regards to removal, any data entries that do not belong to the Netflix or Disney+ catalogs will be removed from the data set, along with the columns for Hulu and Amazon Prime Video availability.

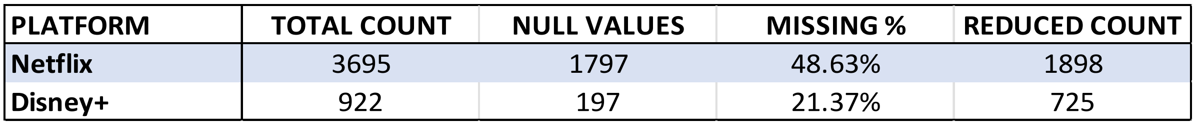
To allow for numerical manipulation of the data, the columns for Rotten Tomatoes score and age rating were modified. For the Rotten Tomatoes score, the suffix “(…)/100” was removed and the data type was set to Int64, treating the numbers as integers and allowing for the existence of NULL values for the time being. For the age ratings, the values were converted into categorical labels utilizing the numbers 1 to 5, ranging from younger to more adult audiences, and similarly converted into Int64 to allow for the NULL values. The final step would be separating this data set into two independent sets, one for Netflix and one for Disney+, to independently assess the information, however it is important to address the issue of the missing values.

3.2 – Data Imputation

Starting with the column for Rotten Tomatoes, since the missing values only consist of 0.07% of the total data set, recreating these values using a simple data imputation method would have no effect on creating a bias or skewing the data. To recreate this missing data, the method used was median imputation, a simple data imputation method described by Van Buuren (2018), the missing values are recreated using the median of the whole column. The imputation of these values was made utilizing the original data set, before the titles belonging to Hulu or Amazon Prime Video were removed, to increase the pool of candidates for movie titles, further decreasing any probability for a bias when working on the separate catalogs as these missing values belonged to the Netflix catalog.

The same method cannot be used for age rating, as the data set as a whole is missing 43.9% of these values, corresponding to 48.63% of Netflix entries and 21.37% of Disney+ entries. Any methods of imputation as described by Van Buuren (2018) would require methodologies closer to a prediction model, going over the scope of statistical analysis of the data set. These methodologies would also require a larger data set to train on and the imputed data is not recommended to exceed 20%, reserved only for drastic exceptions. As both Netflix and Disney+ exceed this metric, any attempts to recreate the missing data would lead to an inevitable bias.

In an attempt to retain as much data as possible, the age rating column will be kept as is, replacing all missing values with the categorical label of 0 to indicate a lack of information. The addition of this new label allows for numerical manipulation of the data without directly sacrificing nearly half of the relevant data. To further compare for a fair analysis of the data, a new data set will be created for both Netflix and Disney+ catalogs where this missing data is removed entirely, effectively reducing the total data set to work entirely with known information and offer a new comparison point.



PLATFORM	TOTAL COUNT	NULL VALUES	MISSING %	REDUCED COUNT
Netflix	3695	1797	48.63%	1898
Disney+	922	197	21.37%	725

Figure 2: Total count of entries in Netflix and Disney+, along with missing values in the Age Rating column and count for the reduced data set once the missing values are removed.

3.3 – Independent Assessment

Individual assessment of the Netflix and Disney+ platforms, with and without missing values, follows simple statistical analysis to find any particular outliers or trends at a surface level. With the separation of entries into their respective platform, any statistical analysis is ensured to result entirely independent from external factors such as any difference in total entries or significant trends noticeable only when the data is together. Throughout this section, each streaming platform will be referred to by its name as well as the added differentiator of "Full" or "Reduced" to signify if the data collected includes the missing values in the Age Rating column as the category 0, or if the missing value entries are removed entirely. Each data set will be assessed independently, and the values used to compare the data sets will be mean, median, standard deviation, minimum, maximum, and quarterly percentiles.

3.4 – Direct Comparison

Comparison of both Netflix and Disney+ streaming platforms will be calculated utilizing the Mann-Whitney U-Test. This test is used due to our data being independent to any other variables, and our data sets not belonging to a normal distribution. If the data had shown to belong to a normal distribution, a Student's T-Test could be used for the same purpose, however since our data sets rely on more than half of the values in the Age Rating column to be entirely missing, no assumption of normal distribution can be made safely without risking a significant error degree, placing the validity of any findings at severe risk of being incorrect.

The Mann-Whitney U-Test can be defined as the probability of both data sets existing within the Null Hypothesis of no difference among the two sets, in terms of central tendency. The calculation also settles the Alternative Hypothesis, dictating that the two data sets differ in terms of central tendency, which will be accepted if the resulting p-value (probability of the obtained results assuming the Null Hypothesis is true) is below the critical value of $\alpha = 0.05$ (5%), as per regular statistical standards, and any other value below that as the value approaches 0 will further eliminate the uncertainty of randomness or errors in the evaluations.

The calculation is based on the overall rank sum of each entry, separated by the category it belongs to (in this case either Netflix or Disney+). Ranking begins from the smallest value with the rank of 1, increasing until the *n*-th value with the rank of *n* (the value *n* is the total amount of entries in both data sets). The ranks are added to formulate the variables R_1 and R_2, pertaining to our compared data sets, Netflix and Disney+. Assignation of R_1 and R_2 is arbitrary and does not affect the results of the standard Mann-Whitney U-Test, although some variants that look for specific differences in the data sets such as increments or decrements toward central tendency are reliant on which data set is assigned to the appropriate *R* value for evaluation.

After calculating the rank sums for each data set, the following formula is used to obtain the test statistics U_1 and U_2, of which the smallest value is used as the test statistic *U*:

$$U = \min\left(U_1 = R_1 - \frac{n_1(n_1+1)}{2}\ , U_2 = R_2 - \frac{n_2(n_2+1)}{2}\right)$$

Since our sample sizes are larger than *n* = 20, a normal distribution is used to calculate the critical value using the following formula:

$$z = \frac{\left(U - \left(\frac{n_1 \cdot n_2}{2}\right)\right)}{\sqrt{\frac{n_1 \cdot n_2(n_1 + n_2 + 1)}{12}}}$$

The variable *z* is then used in the Normal Cumulative Distribution Function (Normal CDF) to determine the probability of observability within a normal distribution. This *z* value, multiplied by 2, gives us the resulting *p*-value of our test, allowing for direct comparison to the critical value α. A case of $p > \alpha$ signifies our Null Hypothesis is true and both data sets have no difference in their central tendency of distribution, and a case of $p < \alpha$ signifies the rejection of the Null Hypothesis, how both data sets are different in their central tendency of distribution.

4. Evaluation

4.1 – Descriptive Analysis of the Data Sets

After separating the data set into its individual streaming platforms, these can be used for individual assessment for any superficial distinctions. Figure 3 displays a descriptive analysis of the Rotten Tomatoes scores for each specific platform, further separated into the "Full" and "Reduced" variants by including or removing any missing values respectively from the Age Rating column, to further explore the possibility of bias created upon artificially imputing the missing data values.

PLATFORM	COUNT	MEAN	MEDIAN	STANDARD DEVIATION
Netflix (Full)	3695	54.44	53	13.83
Netflix (Reduced)	1898	60.5	60	12.81
Disney+ (Full)	922	58.31	57.5	13.95
Disney+ (Reduced)	725	61.77	61	12.98

Figure 3: Descriptive analysis of Rotten Tomatoes Score of streaming platforms, data sets separated to individual streaming platforms and "Full" and "Reduced" variants to account for keeping and removing any missing entries.

Although the missing values were only present in the column for Age Rating, the analysis for Rotten Tomatoes scores was also calculated as a split between the Full and Reduced variants as an exercise to explore how data analysis changes when data is removed. Both streaming platforms returned minimal changes to the analysis when the entries with missing values were removed, but it is worth noting the differences such as the shift of the mean and median values in both streaming platforms. Figure 4 plots the difference in the overall curve before and after data is removed from the data sets. While Disney+ keeps its overall shape since the missing values only accounted to a 21.37% of the data set, a more drastic change can be observed in the Netflix data set, where 48.63% of the data was categorized as missing and removed.

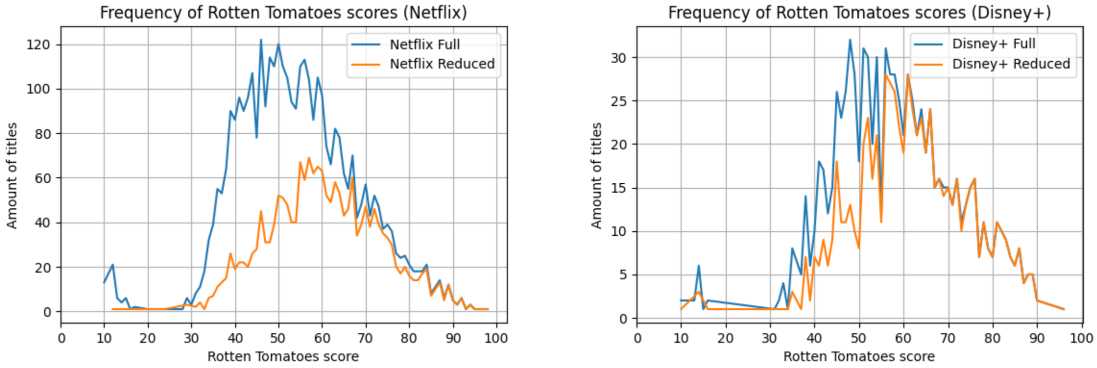
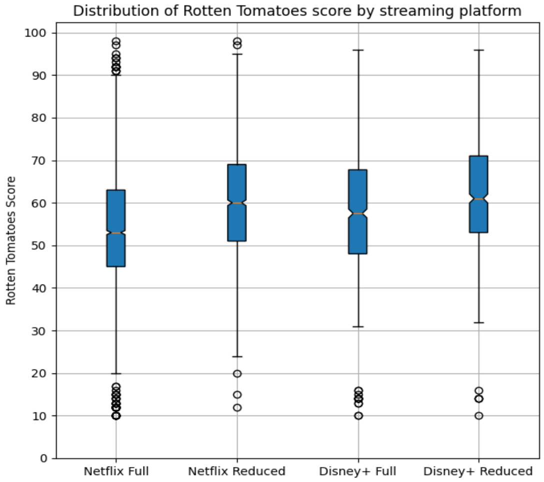


Figure 4: Frequency of Rotten Tomatoes scores on Netflix and Disney+ before and after missing value entries were removed

Analysis of the quartile values of the streaming platforms in both Full and Reduced variants as shown in Figure 5 further emphasizes the shift into higher values after the missing value entries are removed from the data set. In a more extreme case, removal of missing value entries could lead to a more significant difference in central tendency, making any statistical analysis or test on the pruned data set lead closer to an artificial bias introduced by the lack of critical information. The data set of streaming platforms respected the overall central tendency after the removal of missing value entries; however, the principle of artificial bias still exists as there is no proof indicating the missing values could not be significant enough where the central tendency is already shifted by the missing values (this is more prominent in the Age Rating column due to the missing values belonging to that column, whereas in this case Rotten Tomatoes scores were already considered complete even before the imputation method was utilized and any entries removed are a consequence of missing values in Age Rating).



	Netflix (Full)	Netflix (Reduced)	Disney+ (Full)	Disney+ (Reduced)
MIN	10.00	12.00	10.00	10.00
25%	45.00	51.00	48.00	53.00
50%	53.00	60.00	57.50	61.00
75%	63.00	69.00	67.75	71.00
MAX	98.00	98.00	96.00	96.00

Figure 5: Quartile distributions of Rotten Tomatoes scores on Netflix and Disney+ before and after missing value entries were removed

The column for Age Rating is where all the missing values are present, meaning this column would be affected the most compared to any other variable of the data set. Since our missing values represent 48.63% and 21.37% of the Netflix and Disney+ data sets respectively, any form of data imputation or prediction leads to an artificial bias, whether over-reinforcing the current central tendency or distributing values in a seemingly random distribution or following a normal distribution, any attempt would lead to an inevitable alteration of the original data. Taking advantage of Age Rating being measured as categorical data, keeping all the missing value entries by creating a new category labeled as “Not Given” allows for complete analysis of the data sets without the need to impute or remove the missing value entries.

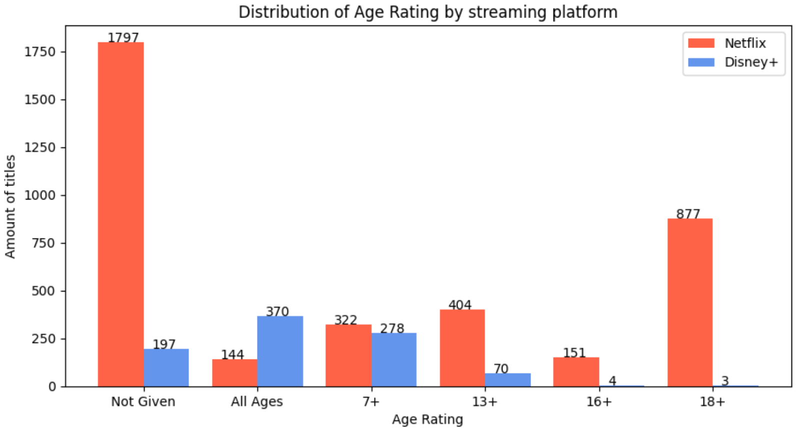


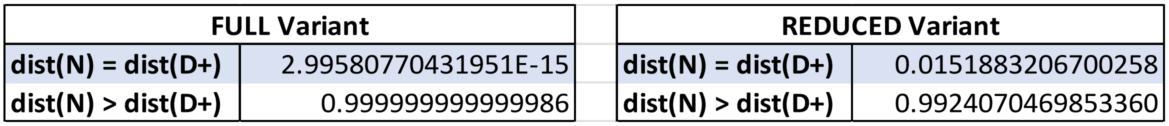
Figure 6: Side-by-side comparison of distribution of Age Rating by streaming platform

For the purpose of this analysis project, the label “Unrated” was not used since most film rating associations use this label to categorize films that have yet to be rated due to pre-production early releases or a film that has yet to be submitted for rating. For this reason, the label “Not Given” was chosen instead of “Unrated” to avoid any confusion with the existing label while emphasizing the relevance of the data as missing from the data set.

4.2 – Inferential Analysis

The Mann-Whitney U-Test was applied to the streaming platforms’ data sets to compare their distributions. Two different variants were applied: 1) Two-sided comparison to test for equality in distributions for any given value, represented as *dist(N) = dist(D+)*; 2) Greater-than comparison to test if the Netflix data set is stochastically greater than the Disney+ data set, represented as *dist(N) > dist(D+)*. Both tests were applied to both variants, Full and Reduced, to test for any significant differences before and after pruning the missing value entries from the data sets. The purpose of this test to compare the data sets is to analyze and evaluate the central tendency of distribution in regards to Rotten Tomatoes scores for each platform.

In both variants, the test proved both data sets are stochastically different from each other. It is worth noting the Full variant had a >3.0E^{-15}% probability of occurring given the equality hypothesis being true, while the Reduced variant shared the same result however with a probability of 1.5%, well under the accepted error of 5% but magnitudes larger than its variant counterpart. Testing for a test hypothesis of greater distribution returned results for both variants with over 99% accuracy that Netflix’s distribution is stochastically larger than Disney+’s distribution.



FULL Variant	
dist(N) = dist(D+)	2.99580770431951E-15
dist(N) > dist(D+)	0.999999999999986

REDUCED Variant	
dist(N) = dist(D+)	0.0151883206700258
dist(N) > dist(D+)	0.9924070469853360

Figure 7: Mann-Whitney U-Test p-value results for column Rotten Tomatoes Score. Displayed number is probability of comparison between the data sets distributions being true given the assigned comparison.

5. Conclusion

Analysis of the data sets illustrates the overall tendencies and trends these streaming platforms align towards. Although it is true that our sample sizes vary significantly (roughly 4:1 when comparing the Full variants and roughly 2:1 for the Reduced variants), the observed trends do not change the aim the streaming platforms' catalogs are marketed. Netflix dominating the sample size is natural due to being the older platform and aiming to fill its catalog with third party properties while Disney+ falls behind with a more select catalog of purely first party ownership of their franchises, or at least a very tight knit business relationship with the Disney corporation.

From these results it is concluded how, although both streaming platforms share similar middle points in their Rotten Tomatoes score by hovering near the 60 points mark, Netflix is proven to be more reliable with higher scores mainly due to its sheer catalog size advantage as well as less volatile distribution as demonstrated by the Mann-Whitney U-Tests. Disney's advantage, however, is its dedicated effort to keeping the vast majority of its catalog available to a wider family audience, dominating the "All Ages" and "7+" labels while anything above "13+" appears to be minimal. Compared to Netflix who despite its larger catalog and broader aim at licensing streaming rights to movies, the label "18+" overwhelms the platform.

The separation of data into variants before and after data entries were removed was a deliberate choice to illustrate the importance of tracking what kind of values as missing as well as try to replicate an extreme case where nearly half of a column's data is missing. For this project alone the analysis and calculations on the data sets were minimally affected due to the overall shapes of the data sets remaining relatively equal. Further study of this data set will need to inquire as to why the Age Rating column, a vital variable when releasing media to the public, was not included as part of the data acquisition. It is also worth noting the only reference to how the data was acquired is the comment that it was scraped from the web, claiming to be from American sources, but no further information is provided as to what kind of scraping algorithm was used or whether or not different regions of the same streaming platforms were taken into consideration when creating the original data set.

6. Bibliography

Bhatia, R. (2021). Movies on Netflix, Prime Video, Hulu and Disney+. Retrieved from https://www.kaggle.com/datasets/ruchi798/movies-on-netflix-prime-video-hulu-and-disney?resource=download

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature, 585*, 357-362. doi:https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering, 9*, 90-95.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random varialbes is stochastically larger than the other. *The Annals of Mathematical Statistics, 18*(1), 50-60. doi:https://doi.org/10.1214/aoms/1177730491

McKinney, W., & others. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, pp. 51-56.

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman & Hall/CRC.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods, 17*, 261-272. doi:https://doi.org/10.1038/s41592-019-0686-2