Netflix Data Trends

By Raj Shah

Introduction

- In this slideshow I will present my analysis of the Netflix dataset
- The analysis was performed using Python Pandas, Seaborn, SciKit, and Jupyter Notebook
- This analysis includes manipulation of quantitative data and visualizations of relationships between the data

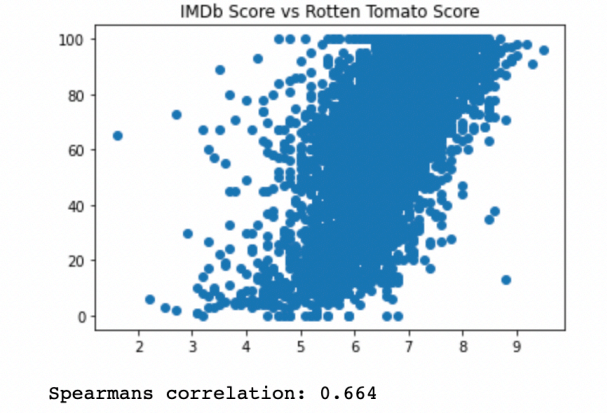
Table of Contents:

1 Correlation Between Scoring Systems

Correlation Between Scoring Systems Introduction

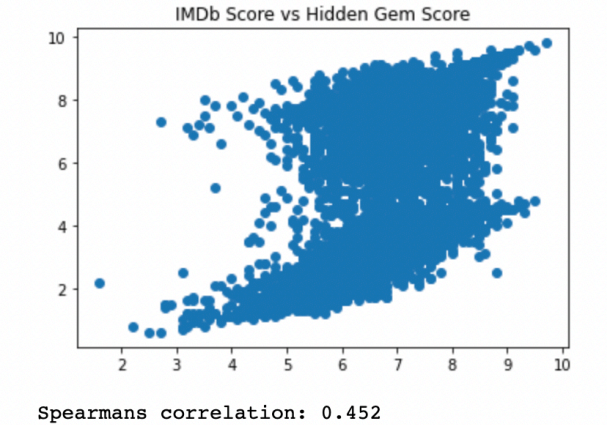
- The four title scoring systems included in the data set were:
  - Rotten Tomato Score
  - IMDb Score
  - Hidden Gem Score
  - Metacritic Score
- To measure each scoring system’s relationship to one another, a scatter plot was created and a correlation coefficient was calculated for each relationship.
- Correlation scores are given from −1 to 1.
  - 1 or −1 meaning strongest positive or negative correlations, respectively
  - 0 meaning no correlation

Scoring Correlation Scatter Plots

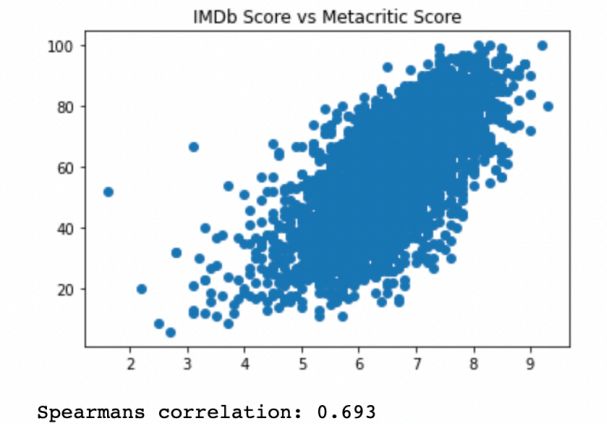


Spearmans correlation: 0.664

- IMDb Score has a relatively strong correlation with the other Scoring Systems.
- IMDb Score has the **strongest correlation** with Rotten Tomato and Metacritic Score.
- IMDb Score has a **weaker correlation** with Hidden Gem Score.

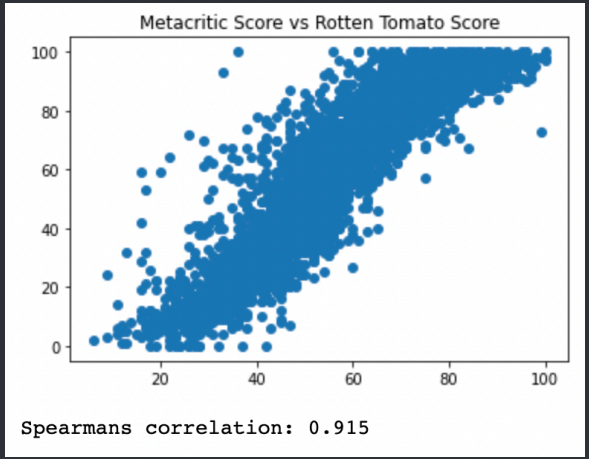


Spearmans correlation: 0.452



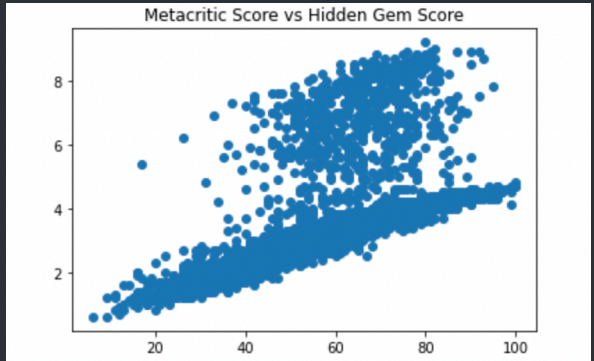
Spearmans correlation: 0.693

Scoring Correlation Scatter Plots

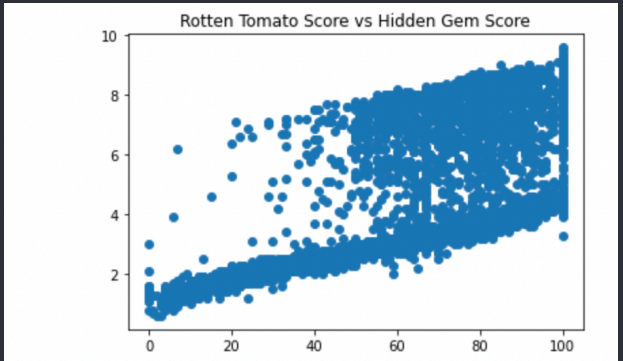


Spearmans correlation: 0.915

- In general, Rotten Tomato, Metacritic, and Hidden Gem Scores have higher correlations with one another than IMDb Score.
- Metacritic Score and Rotten Tomato Score have the **strongest correlation.**
- The strong correlation makes sense because Metacritic and Rotten Tomato ratings are given by professional critics.



Spearmans correlation: 0.812



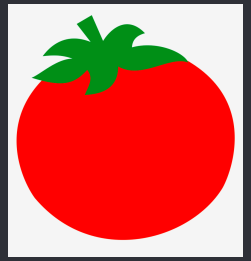
Spearmans correlation: 0.693

2

Composite Score

What is the Composite Score

- The dataset provided four Scoring Systems per movie.
- Such a high number of rating systems makes it difficult to provide an effective rating for each movie.
- To tackle this issue, I developed a composite score for each title.



How to Handle Missing Data in Calculations?

- Virtually all titles in the data set had an IMDb Score.
- However, many of the titles had null values for the other scoring metrics.
- To compensate for the missing data, the values for the other scoring. metrics were predicted using linear regression with the IMDb Score.

How is the Composite Score Calculated?

- The correlation coefficients of each rating system were averaged out to create a rating constant.
- Each rating metric per title is multiplied by its rating constant.
- After each metric is weighted with their rating constant, the rating metrics are added together to create the raw composite score.

How is the Composite Score Calculated?

- The raw composite score doesn’t exhibit a normal distribution.
- The raw composite score is scaled using a square root function to normalize the data.
- The range of the final composite score is from 10 – 80.
- The composite score is strongly negatively correlated with the other scoring systems, meaning that 10 is the best score and 80 is the worst score.

3 Visualizations of Trends on Netflix

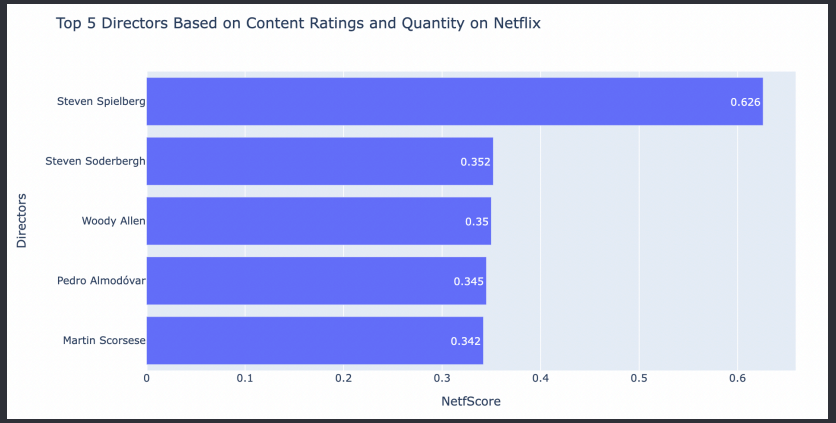
List of Visualizations

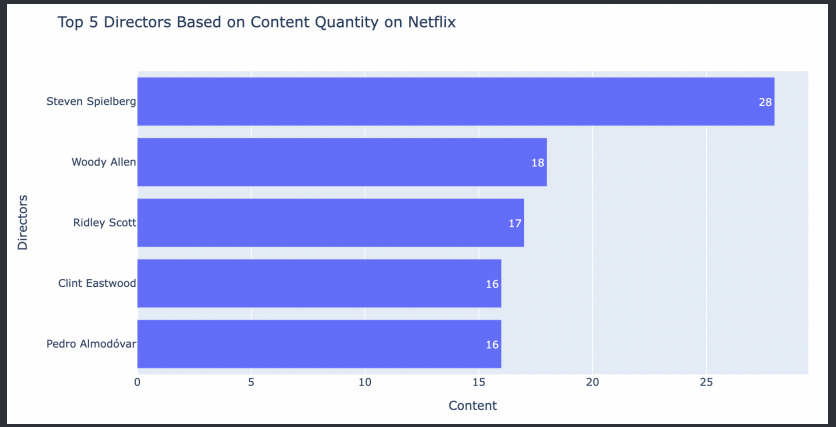
- Top 5 Directors on Netflix (Netf and Content)
- Top 5 Countries on Netflix (Netf and Content)
- Top 5 Genres on Netflix (Netf and Content)
- Top 5 Movie/TV Ratings & Languages on Netflix
- Trend of Content Produced Over the Past 20 Years on Netflix
- Series/Movies Rating/Content (Composite and Content)

NetfScore

- Generally, when elements within a category are ranked on Netflix, it is based on the popularity of that element on Netflix.
- To factor in the ratings of content within an element with the element’s prelevance, I developed the NetfScore.
- The NetfScore is calculated by dividing a category’s instances by its average composite score (no score range as it is just a ratio).
- The NetfScore allows us to determine the highest rated and most prelevant elements in each category.

Top 5 Directors on Netflix

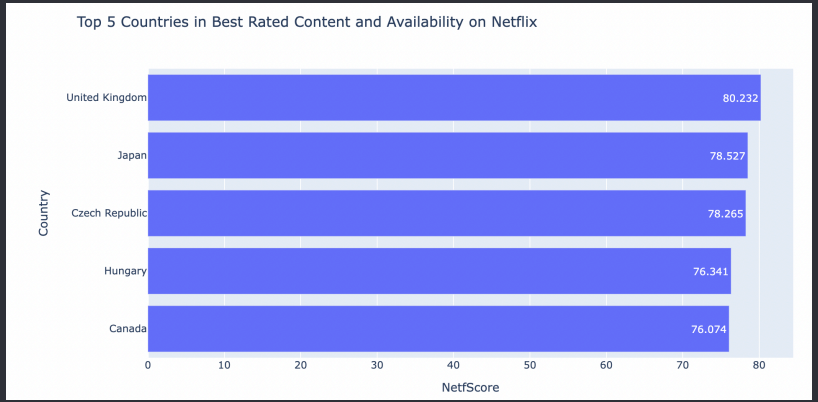


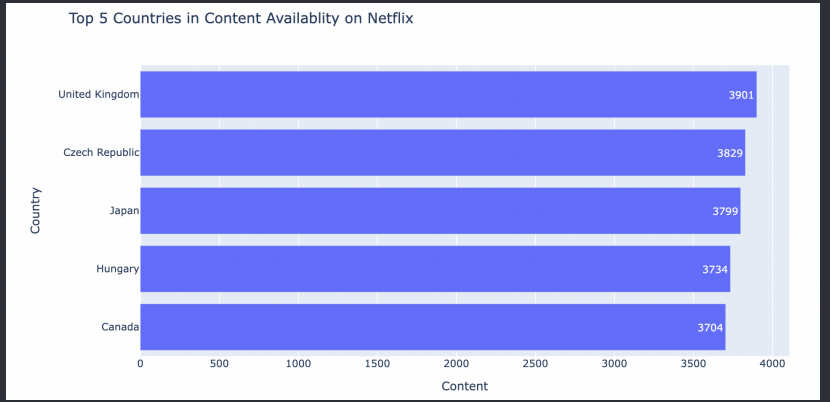


- The top 5 directors ranking based on NetfScore is quite different from the top 5 director ranking based on content quantity.
- Pedro Almodovar and Martin Scorsese were not ranked top 5 based on their content quantity, but were ranked top 5 based on their NetfScore.
- On the other hand, Clint Eastwood and Steven Soderbergh were no longer ranked top 5 when their NetfScore was factored in.

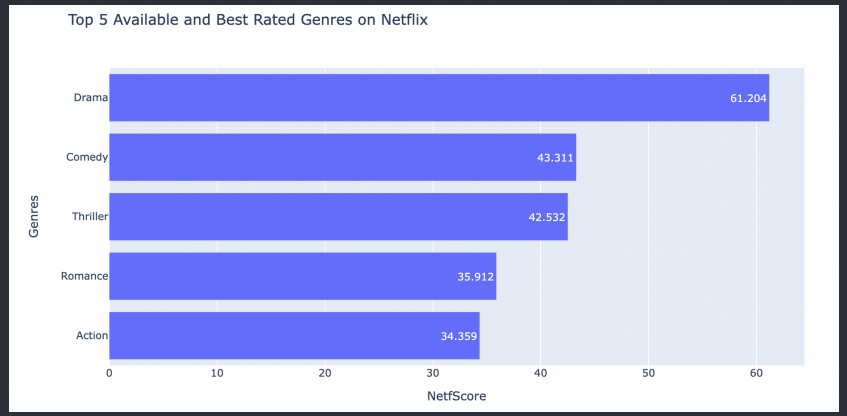
Top 5 Countries on Netflix

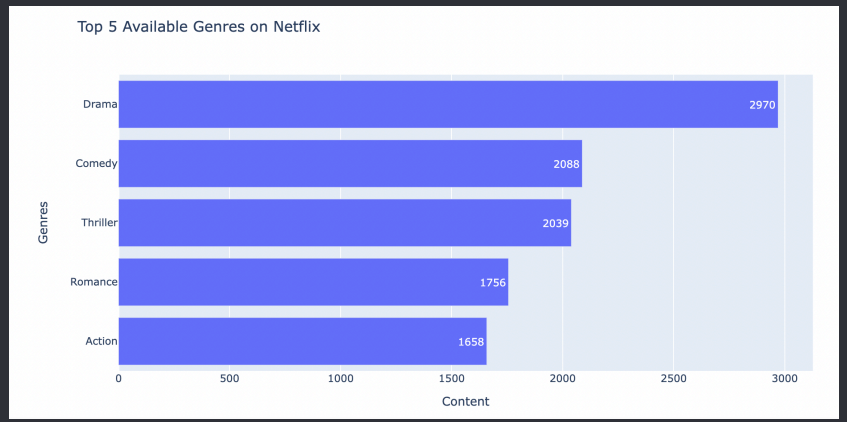
- The top 5 countries ranking based on NetfScore is not too different from the top 5 countries ranking based on content quantity.
- Based on content availability, Czech Republic is ranked second with Japan ranked third, while based on NetfScore, Japan is second, while Czech Republic is third.
- Rankings aren’t as different on both lists because the composite scores of content from each country likely aren’t too different.





Top 5 Genres on Netflix



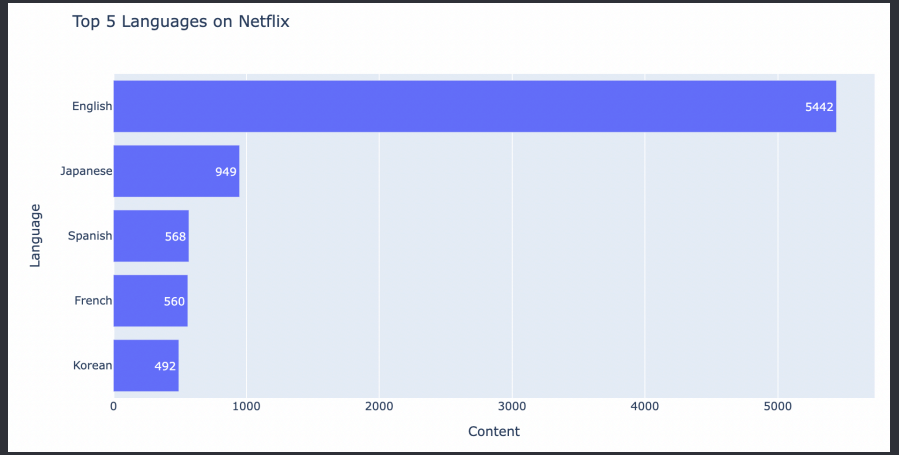


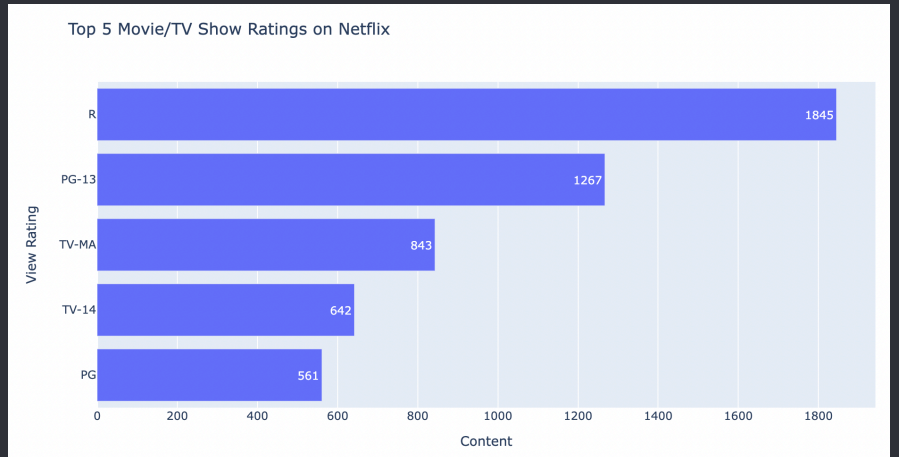
- The top 5 genres ranking based on NetfScore is identical from the top 5 genres ranking based on content quantity.
- The rankings are likely very similar as the content available per category is very different per category and thus influences the NetfScore heavily.

Top 5 Languages and Ratings on Netflix

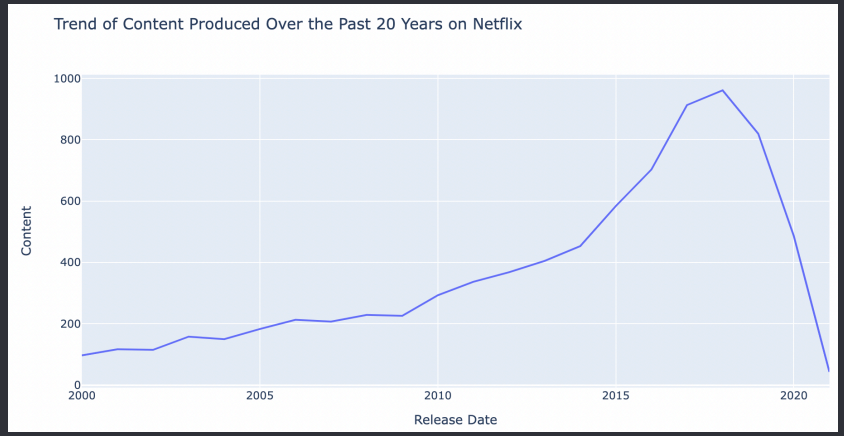
English is the most popular language on Netflix beating the second most popular language, Japanese by almost 4000 titles.

Movie/TV Show Ratings for teens and older (PG-13+) are among the most popular ratings on Netflix.





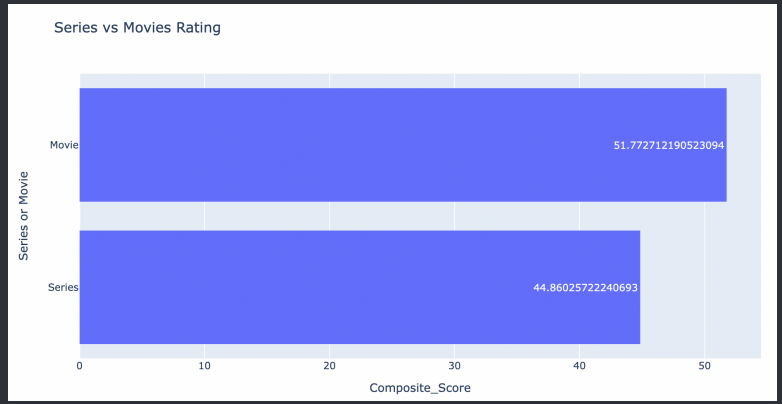
Trend of Content Produced in the Past 20 Years on Netflix

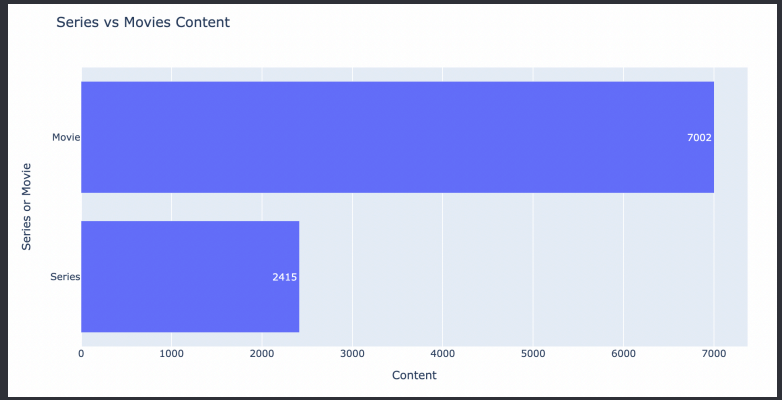


- The graph demonstrates a gradually increasing upward trend in content on Netflix till 2018.
- The amount of content on Netflix produced after 2018 decreases sharply year by year.
- The dataset is from 2021, so it is conceivable why Netflix has such a low amount of content produced in 2021.

Series vs Movies Rating/Content on Netflix

- In this comparison, composite score was used instead of NetfScore because there is such a drastic difference in the content quantity of movies and series.
- While there are a lot more movies on Netflix than TV shows, the composite score of TV shows is lower than that of movies.
- This means that on average TV shows are higher rated than movies.





4

GitHub Link to Jupyter Notebook Repo

Github Link to Jupyter Notebook Repo

Link: https://github.com/rajshah0904/Netflix-Data-Analysis/blob/main/SIFAnalysis%20(2).ipynb

Thank You!