ROBERT H. SMITH
SCHOOL OF BUSINESS

BUDT704: DATA PROCESSING AND ANALYSIS IN PYTHON

"Exploratory Data Analysis of Netflix Streaming Data"

Report By

Akshay Sharma
Harish Bhupathiraju
Solayappan Ganesaan
Srishti Gupta
Sai Sumanth Devara

1 INTRODUCTION

This study aims to look into the Netflix data to understand the viewing preferences of customers using Python for exploratory data analysis (EDA). The objective is to provide insights that drive strategic growth recommendations for the streaming platform. We use a comprehensive research methodology, including statistical techniques and visualizations, ensuring our findings serve as a basis for strategic recommendations.

The report's deliverables encompass diverse facets of the Netflix dataset, covering content trends, user preferences, regional variations, and performance metrics. The objective is to understand the factors contributing to Netflix's success comprehensively.

The dataset, sourced from Kaggle, includes essential information like genre, language, region, runtime, cast, ratings, box office performance, and awards. Before analysis, meticulous data cleaning ensures the accuracy and reliability of our insights.

The EDA covers correlation matrix analysis, content type trends, runtime patterns, genre preferences, regional content availability, language distribution, directorial achievements, production house performance, and recognition based on awards and ratings.

Ultimately, the exploration seeks to formulate strategic recommendations to meaningfully contribute to Netflix's growth trajectory. The objective of understanding content preferences, regional dynamics, and performance metrics is to equip Netflix with actionable insights for informed decision-making.

2 BACKGROUND

In the dynamic landscape of streaming platforms, understanding viewer preferences and content dynamics is pivotal for the sustained success of platforms like Netflix.

As viewers increasingly turn to streaming services for entertainment, streaming platforms face the challenge of delivering diverse and engaging content and keeping a close tab on ever-evolving customer preferences. The rise of on-demand streaming has transformed the traditional media landscape, making it imperative for platforms to navigate the complexities of user behavior, content popularity, and strategic decision-making.

This exploration exercise is set against the backdrop of Netflix's prominence in the streaming industry. As one of the leading platforms globally, Netflix's success is closely tied to its ability to offer a vast array of content and tailor its offerings to the diverse tastes of its global audience. The project recognizes Netflix's position as a trendsetter and seeks to contribute to its ongoing success by uncovering nuanced insights within its streaming data.

The exploration aims to contribute to the broader conversation on streaming analytics. The project aspires to offer actionable insights beyond mere data analysis by focusing on decoding patterns within the Netflix dataset. It is a venture into the narratives behind the numbers, aiming to unravel the stories of viewer choices, content popularity, and the strategic decisions that shape the Netflix experience.

3 RESEARCH METHOD SUMMARY

The research methodology involves a comprehensive examination of the Netflix dataset, rigorous data cleaning, and in-depth exploratory data analysis (EDA). We employ various statistical techniques and visualizations to extract meaningful insights from the dataset. This thorough approach ensures that our findings foster a robust, data-driven decision-making process.

- The initial data-cleaning phase is crucial to enhance the reliability and accuracy of the analysis. This involves addressing discrepancies, removing duplications, handling missing values, and creating a clean and dependable dataset.

- Thereafter, the exploratory data analysis phase delves into the heart of the Netflix data, uncovering patterns, correlations, and trends. The underlying narratives within the data are deciphered through strategic statistical techniques, providing the decision-makers with a nuanced understanding of the customers' viewing preferences, content dynamics, and potential growth areas.
- Visualizations, such as word clouds, charts, and graphs, are used to present complex insights in an accessible manner.
- The conclusion of these efforts results in a comprehensive and insightful analysis that informs strategic recommendations and instills confidence in decision-making processes.

By embracing a data-driven approach, the methodology ensures that recommendations are grounded in empirical evidence, which the stakeholders can use to make informed choices for the strategic advancement of the Netflix platform.

4 DATASET

Sourced from Kaggle, our dataset encompasses vital information such as genre, category classification, language and region availability, runtime, cast details, rating metrics, box office performance, and awards. Prioritizing data cleaning is imperative to ensure the integrity of our findings, enhancing the reliability of the insights derived.

4.1 DATA CLEANING

Dropped Duplicate Values:

- Removed duplicate rows to ensure unique data entries.

Filled Missing Values:

- Categorical Columns: Filled missing values with "Unknown" for consistency.
- Numerical Columns: Filled missing values with " to avoid biasing calculations.
- Dropped Unnecessary Columns: Removed irrelevant columns like links, summaries, and images.

Handled Date Columns:

- Converted "Release Date" and "Netflix Release Date" to DateTime format and extracted year and month. Handled missing dates by filling with NaN and creating additional columns for year and month (filled with 0 for missing dates).
- Handled anomalies in date data, such as years greater than 2021(the dataset was last updated in April 2021)

Handled Awards Columns:

- Converted "Awards Received" and "Awards Nominated For" to integers and filled missing values with 0.
- Created a new column "Total Awards" as the sum of awards received and nominated.

Handled Boxoffice Column:

- Removed non-numeric characters and converted to numeric values.
- Filled missing values with "Unknown".

Handled Runtime Column:

- Created a new column "Runtime (mins)" by converting runtime strings to specific ranges (0-30 mins, 30-60 mins, 60-120 mins, 120+ mins, Unknown).

Rationale for Data Cleaning:

- Removing duplicates ensures data integrity and prevents bias in analysis.
- Filling missing values avoids errors and allows for complete data analysis.
- Dropping unnecessary columns focuses the analysis on relevant features.
- Handling awards columns provides additional data points for insights into critical reception.
- Handling box office columns allows for financial analysis.

Impact of Data Cleaning:

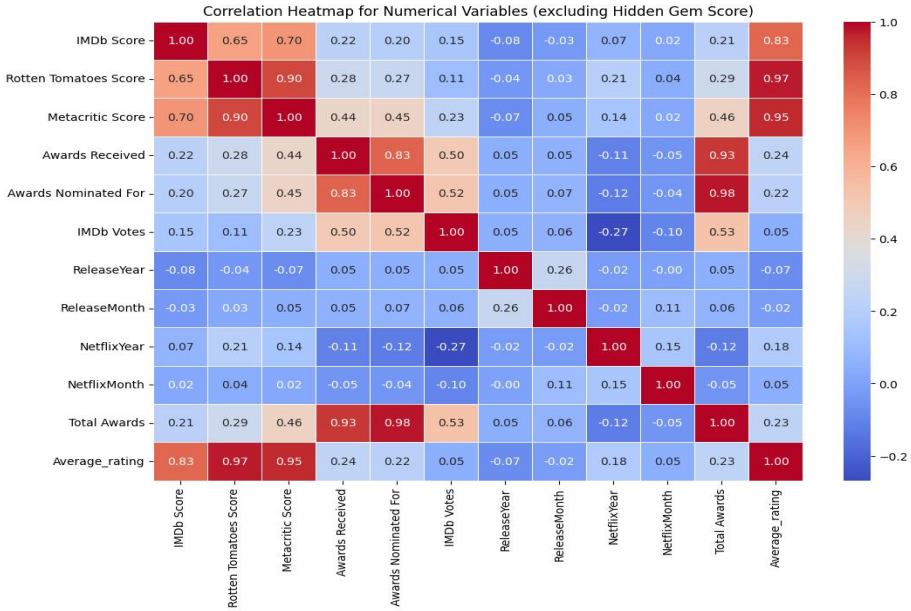
By implementing these data cleaning steps, the data becomes reliable, consistent, and ready for further analysis. This ensures the accuracy and validity of the findings presented in the final report.

5 EXPLORATORY DATA ANALYSIS

The data analysis uncovers various facets, including correlation matrix analysis, content type trends, runtime patterns, genre preferences, regional content availability, language distribution, directorial achievements, production house performance, and recognition based on awards and ratings.

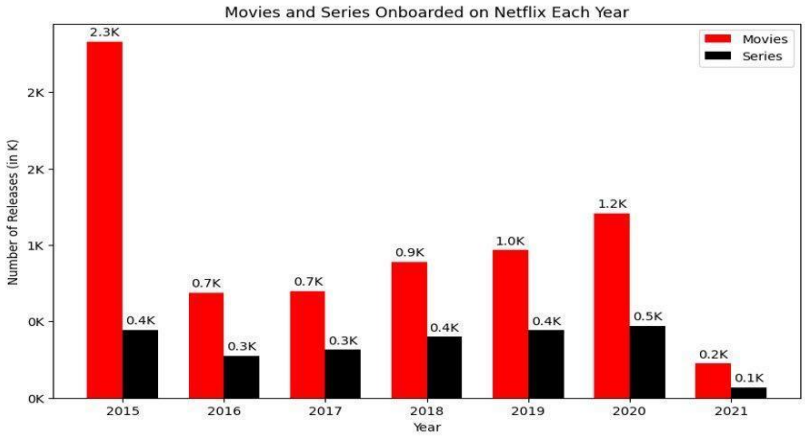
5.1 CORRELATION MATRIX CONTENT ANALYSIS

The resulting heatmap provides a visual representation of how strongly or weakly numerical variables in the dataset are correlated. The color intensity and the numerical values help interpret the direction (positive or negative) and strength of the correlations. The heatmap is a valuable tool for identifying patterns and dependencies among numerical features in the dataset. It is worthwhile to focus on the relationship between variables IMDB score, Rotten Tomatoes score, Metacritic, Netflix year, and Release Year. We observe that the correlation between the scores and Netflix Year are 0.07,0.21, 0.14 respectively while the correlation between the scores and Release Year is -0.08, -0.04, and -0.07 respectively. This means that Netflix has been onboarding better content over the years while there has been no clear uptick in the quality of content released over the years.



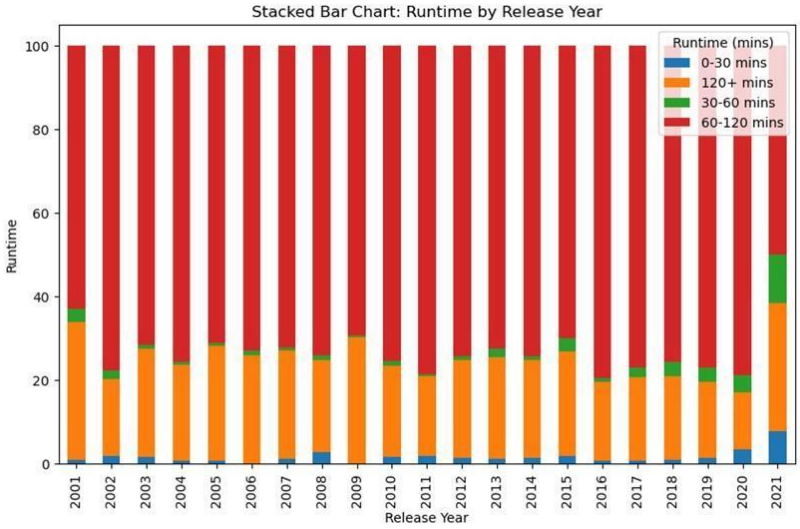
5.2 CONTENT ANALYSIS

The trend of the percentage mix of movies and series released on Netflix over the years was analyzed. The exercise is aimed at providing insights into Netflix's acquisition strategy. Excluding 2015(Netflix boom) and 2021 (Data available till April 2021), content acquisition peaked in 2020 in both the series and the movies categories.Although the increase is gradual in the series category, this could be attributed to the content availability in the market as opposed to movies.



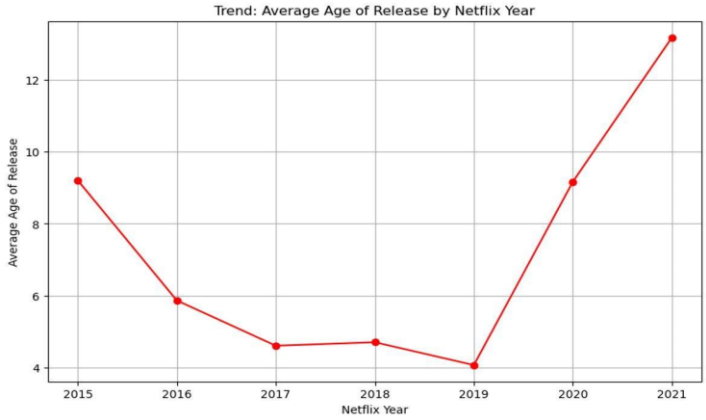
5.3 RUNTIME ANALYSIS

Although some of the movies in the dataset even date back to the 1920s, the focus was narrowed to those released post-2000. This selective approach aims to contrast patterns among movie watchers and to understand Netflix's strategy in acquiring movies and TV series concerning their runtime. We can observe that there has been no stark change in the trend of the run time of movies released over the years.



5.4 HAS NETFLIX BEEN ONBOARDING RETRO CONTENT LATELY?

A new column named "Age" was defined, which represents the difference between the release year on Netflix (i.e., the year the content was onboarded) and the original release year. Consistent with the earlier exclusion of 2015 and 2021 for analytical purposes, the observation that emerged was that during the onset of the COVID-19 pandemic, Netflix embarked on an extensive retro content acquisition phase, contributing to user acquisition and the production of exclusive Netflix shows and films.

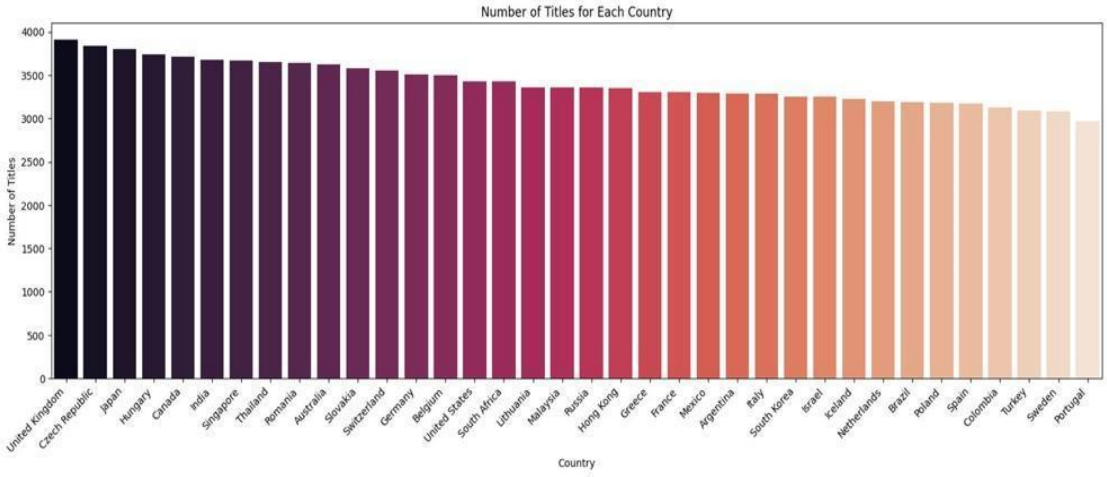


5.5 GENRE PREFERENCES

Genres are pivotal in shaping user preferences and streaming behavior, making this exploration important to understanding the viewer's choices. It is observed that the genre preferences exhibit a notable variation across different years. The popularity of specific genres directly influences the company's content acquisition strategy. Drama has been the top genre, followed by Comedy and Thriller. It is interesting to note that the top genre changes every year.

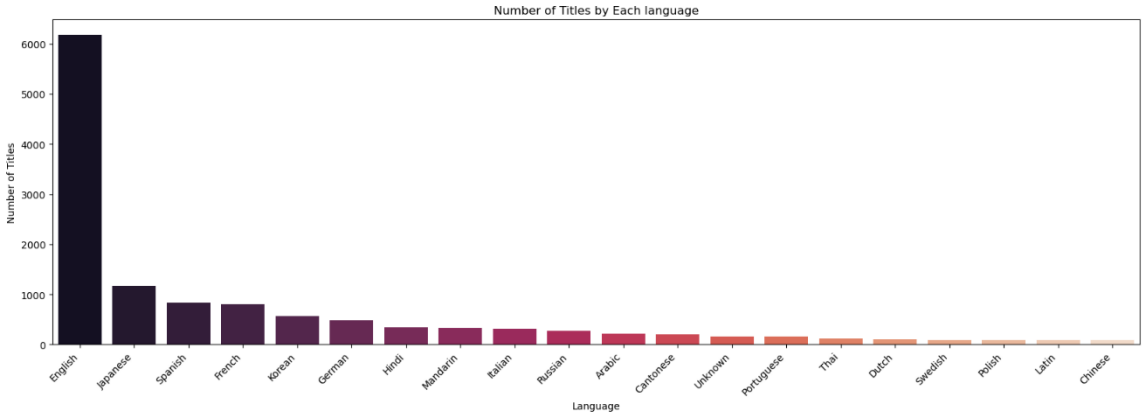
5.6 TITLE AVAILABILITY PER COUNTRY

Understanding regional content offerings is vital for appreciating Netflix's global strategy and commitment to tailoring content to local audiences.The analysis pinpoints the countries that enjoy a rich selection of titles, reflecting Netflix's investment in diverse content offerings. Observing how the number of titles varies from country to country is fascinating. This variation may be influenced by multiple factors such as licensing agreements, regional preferences, and the streaming platform's strategic decisions. UK has the highest number of titles followed by the Czech Republic and then Japan, US is in the middle of the list, despite Netflix being a US-based company and this could be because of regional restrictions.



5.7 TITLES AVAILABLE PER LANGUAGE:

The analysis derives insights from the Netflix dataset, particularly highlighting the distribution of titles across various languages. As per the findings, English emerges as the predominant language, having over 6000 titles, followed by Japanese with around 1100 titles. Netflix has strategically utilized localization strategies to accommodate varying linguistic preferences. By providing content in multiple languages, Netflix ensures a globally dispersed user base can access and enjoy content in their preferred language.



5.8 TOP SERIES AND MOVIES BASED ON AWARDS

Title	Series or Movie	Languages
Sesame Street	Series	English, Spanish, American Sign Language
American Horror Story	Series	English
Modern Family	Series	English, Spanish, Chinese
Will and Grace	Series	English
Breaking Bad	Series	English, Spanish
The Big Bang Theory	Series	English, Hindi, Italian, Russian, Mandarin, Klingon
Greys Anatomy	Series	English
The Walking Dead	Series	English
Friends	Series	English, Dutch, Italian, French
Glee	Series	English
12 Years a Slave	Movie	English
Parasite	Movie	Korean, English
La La Land	Movie	English, Cantonese
Moonlight	Movie	English
Once Upon a Time in Hollywood	Movie	English, Italian, Spanish, German
Birdman or (The Unexpected Virtue of Ignorance)	Movie	English
The Shape of Water	Movie	English, American Sign Language, Russian, French
Mad Max: Fury Road	Movie	English, Russian
ROMA	Movie	Spanish, Mixtec, English, Japanese, German, French, Norwegian
Gravity	Movie	English, Greenlandic

5.9 TOP SERIES AND MOVIES BASED ON AVG RATING

```
# Top movies and series based on Overall Net rating across IMDB,Rotten Tomatoes and Meta Critic
df['Rotten Tomatoes Score'] = df['Rotten Tomatoes Score']/10
df['Metacritic Score'] = df['Metacritic Score']/10
df['Average_rating'] = df[['IMDb Score','Rotten Tomatoes Score','Metacritic Score']].mean(axis=1, skipna=True)
```

Title	Series or Movie	Languages
Breaking Bad	Series	English, Spanish
Flavours of Romania	Series	English
A Lion in the House	Series	English
Our Planet	Series	English
Im Sorry	Series	Unknown
The Last Dance	Series	English
Avatar: The Last Airbender	Series	English
Rick and Morty	Series	English
Reply 1988	Series	Korean
Sherlock	Series	English
No Festival	Movie	Romanian
The Godfather	Movie	English, Italian, Latin
In the Shadow of the Moon	Movie	Unknown
The Dream House	Movie	English, Spanish
City Lights	Movie	None, English
Mission: Impossible - Ghost Protocol	Movie	Unknown
Trois Couleurs - Rouge	Movie	French
Modern Times	Movie	English
Schindlers List	Movie	English, Hebrew, German, Polish, Latin
Parasite	Movie	Korean, English

5.10 TOP SERIES AND MOVIES BASED ON BOXOFFICE COLLECTION

Title	Genre	Languages
Titanic	Drama, Romance	English, Swedish, Italian, French
Jurassic World	Action, Adventure, Sci-Fi	English
Avengers Assemble	Action, Adventure, Sci-Fi	English, Russian, Hindi
The Dark Knight	Action, Crime, Drama, Thriller	English, Mandarin
Avengers: Age of Ultron	Action, Adventure, Sci-Fi	English, Korean
The Dark Knight Rises	Action, Adventure	English, Arabic
E.T. the Extra-Terrestrial	Family, Sci-Fi	English
The Hunger Games: Catching Fire	Action, Adventure, Mystery, Sci-Fi, Thriller	English
Jurassic World: Fallen Kingdom	Action, Adventure, Sci-Fi	English, Russian
Честное пионерское 3	Animation, Adventure, Comedy, Family, Fantasy	English, Spanish

6 RECOMMENDATIONS

Calculation of our Recommendation Score:
The code aims to compute a recommendation rating to identify hidden gems in films or series. Hidden gems are characterized by high ratings and low user reviews. To ensure a fair comparison, the number of IMDb votes is normalized on a scale of 1 to 10, contributing to a comprehensive recommendation metric.

Calculation Steps:
Normalization of IMDb Votes:

- The formula is applied to normalize the IMDb votes on a scale of 1 to 10.
- This normalization is designed to create a fair comparison among different films or series.

Calculation of Recommendation Rating:

- The 'normalized_IMDB_Votes' and 'IMDb Score' are combined using the formula to create the 'Recommendation_rating.'
- The combination of normalized votes and IMDb score ensures a balanced metric for identifying hidden gems.

```
# Calculate Recommendation rating
df['normalized_IMDB_Votes'] = 10 - ((df['IMDb Votes'] - df['IMDb Votes'].min())/(df['IMDb Votes'].max() - df['IMDb Votes'].min())*10)
df['Recommendation_rating'] = df['normalized_IMDB_Votes']*0.5 + df['IMDb Score']*0.5
```

Title	Series or Movie	Languages	IMDb Score
Flavours of Romania	Series	English	9.5
Im Sorry	Series	Unknown	9.2
Our Planet	Series	English	9.3
Regiment Diaries	Series	Hindi	9.1
Irmão do Jorel	Series	Portuguese	9.1
The World Between Us	Series	Mandarin	9.1
My Mister	Series	Korean	9.1
Reply 1988	Series	Korean	9.1
Leah Remini: Scientology and the Aftermath	Series	English	9.1
Stranger	Series	French	9.0
No Festival	Movie	Romanian	9.7
The Dream House	Movie	English, Spanish	9.4
Conspiracy	Movie	English	9.2
The Consuls Son	Movie	English	9.1
One Girl	Movie	Romanian, Arabic, Finnish, English, Acholi	9.1
Green Gold	Movie	Spanish	9.0
Bye Bye London	Movie	Arabic	9.0
No Longer kids	Movie	Arabic	9.0
David Attenborough: A Life on Our Planet	Movie	English	9.0
Hole in the Wall	Movie	Unknown	8.9

RECOMMENDATION

Strategy Recommendations for Netflix:
Diversify Content Portfolio:

- Continue offering a diverse range of content to cater to a global audience.
- Invest in a mix of genres, languages, and regional content to appeal to a wide spectrum of viewers.

User Engagement Strategies:

- Emphasize the creation of engaging and interactive content to enhance user experience.
- Explore interactive storytelling formats, interactive episodes, and user-driven narratives.

Personalization Algorithms:

- Strengthen personalization algorithms to provide more accurate content recommendations.
- Leverage machine learning and AI to understand individual viewing preferences and tailor recommendations accordingly.

Global Localization:

- Enhance content localization strategies to better serve diverse linguistic and cultural preferences.
- Invest in region-specific content creation and marketing to strengthen the global footprint.

Strategic Partnerships:

- Form strategic partnerships with regional content creators and production houses.
- Collaborate with local talent to produce exclusive content that resonates with specific audiences.

Data Limitations:

Genre Classification:

- Genre classification may need improvement for better content categorization.
- Consider refining genre labels and acquiring more accurate genre data for enhanced recommendation systems.

Missing Values:

- Address missing values in critical data fields for a more comprehensive analysis.
- Implement strategies to reduce missing data and improve data completeness.
- Fields such as Box office collection were sparsely populated, a detailed study on that field could offer insights into the popular matchups for Box office hits in the past thereby providing plausible advice for future investments in Netflix's production.

Streaming Statistics:

- Lack of detailed streaming statistics has limited our study over the depth of content performance analysis.
- Seek partnerships with third-party analytics providers or internal tools to gather more granular streaming data.