Sean McLean.

Catherine Smereena Dommaty.

Susheel Reddy Nelvoy.

ALY 6040 –Data Mining Applications.

Module 1 Final Project

Title:Exploring Content Trends and User Preferences through Data Analysis

Introduction:

Netflix is a well-known streaming service that provides its users with a huge selection of films and TV series. In this paper, we examine a dataset that we downloaded from Kaggle that includes details on the Netflix material that is available. The dataset has 8,807 entries with 12 variables, each containing details about a movie or television show's director, release year, rating, runtime, genre, and synopsis. We want to learn about user behaviour and preferences, spot content trends, and investigate how Netflix uses this information to create recommendation engines and draw in new users through exploratory data analysis, or EDA.

Exploration of the Data:

In the context of exploration, we loaded the dataset, examined its structure, and performed data preprocessing tasks. We aimed to understand the data's characteristics, detect any missing values, and clean the data appropriately. After the initial exploration, we conducted EDA to delve deeper into the dataset.

After loading and exploring the Netflix dataset, we performed data preprocessing, converting 'type' and 'rating' to factor variables and 'date_added' to Date type. Missing values in the 'director' and 'cast' columns were handled by replacing them with "Unknown". The dataset comprises 8,807 entries without any duplications. Through exploratory data analysis (EDA), it was determined that the data is relatively clean, with no significant outliers. However, some unusual entries, such as 0 minutes of runtime, were identified. The EDA revealed valuable insights into the Netflix streaming dataset. The majority of content on Netflix consists of movies, with a smaller portion being TV shows. The most common ratings are TV-MA, TV-14, and TV-PG. The number of releases on Netflix has generally increased over the years, with a significant surge after 2010. Most content on Netflix has a runtime between 90 to 120 minutes. Raúl Campos, Jan Suter, and Marcus Raboy are among the directors with the most content on Netflix, and International Movies, Dramas, and Comedies are the most common genres available on the platform. These findings are significant as they provide insights into user preferences, enabling Netflix to enhance its recommendation system, tailor its content offerings, and increase user engagement and subscriber retention.

Exploratory Data Analysis (EDA):

1.Distribution of Content Type (Movie/TV Show):
The distribution of content type on Netflix reveals that the majority of content consists of movies, with a smaller portion being TV shows. In this distribution, movies represent the majority, while TV shows make up a smaller portion. The graph indicates that Netflix has a more extensive collection of movies compared to TV shows. This insight is crucial for understanding the composition of content available on Netflix. With movies being the dominant content type, it suggests that users may prefer movies over TV shows. Understanding this distribution enables Netflix to tailor its content offerings to meet user preferences,

enhancing its recommendation system and increasing user engagement and subscriber retention.

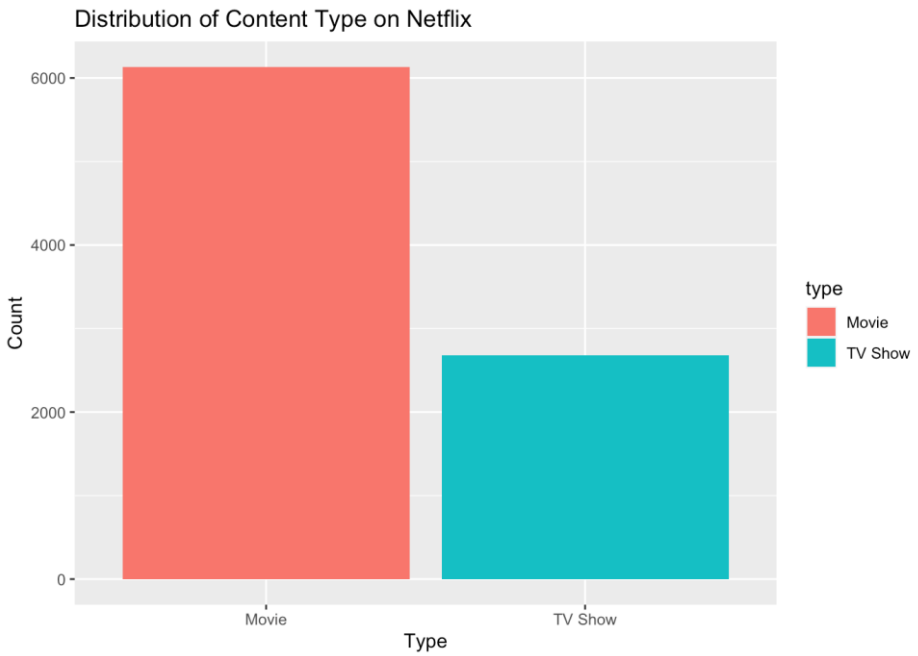

Figure1: Distribution of Content Type On Netflix(Movie/Tv Show)

2.Distribution of Ratings:

The Netflix rating distribution shows how frequently certain content ratings are offered on the network. The most common ratings are TV-MA, TV-14, and TV-PG, which suggests that there is more content aimed at older audiences. The graph illustrates the range of content ratings, offering viewers a glimpse into the assortment of Netflix series and films. Comprehending this distribution facilitates the classification of material and recommendation algorithms, enabling Netflix to customise content suggestions according to users' inclinations and suitability for their age. By making tailored recommendations and making sure that the content is in line with the tastes and sensitivities of the audience, this method improves the user experience.

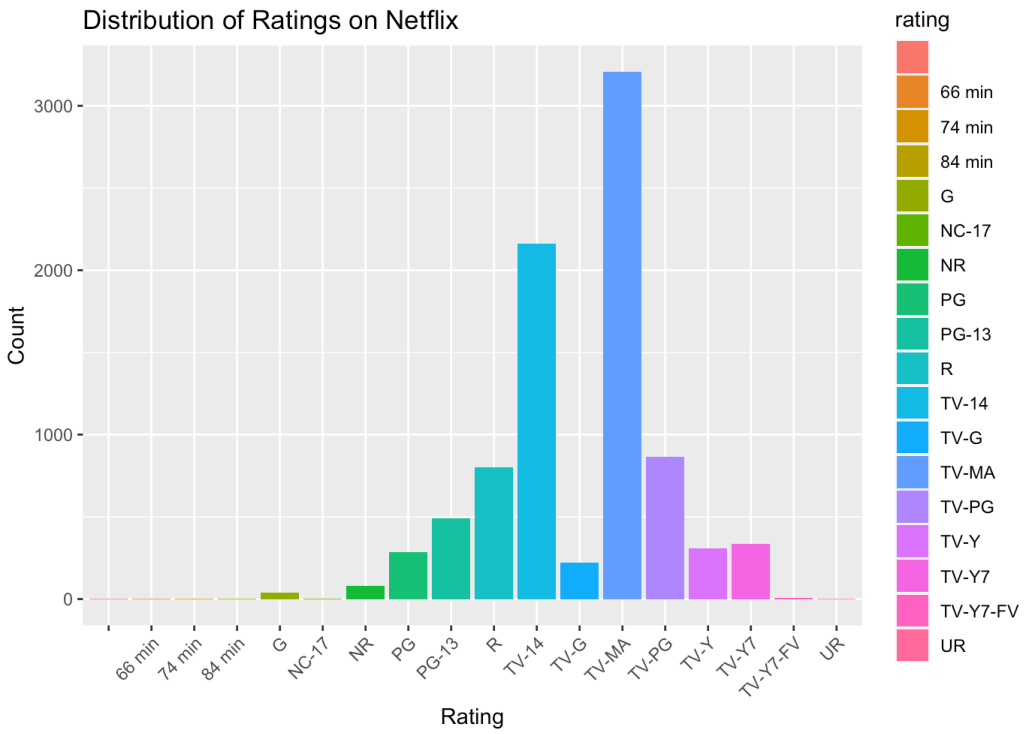

Figure2:Distribution of Ratings on netflix.

3. Release Year Distribution:

The release year distribution illustrates the number of content releases on Netflix over the years. There has been a general increase in the number of releases over time, with a significant surge after 2010. The graph provides valuable insights into the evolution of content available on Netflix, indicating a growing collection over the years.

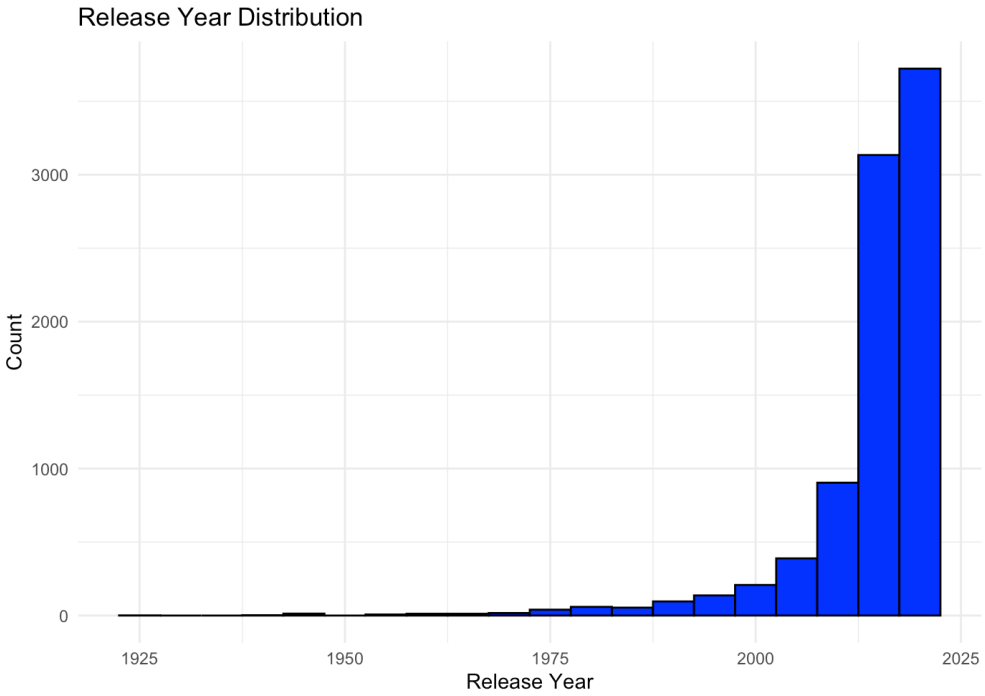

Figure3:Release Year Distribution,their Count.

4. Runtime Distribution:

The frequency of various content runtimes that are available on Netflix is shown by the runtime distribution. Most of the content is between ninety and one hundred and twenty minutes long. Comprehending the runtime distribution facilitates content classification and enhances recommendation systems.

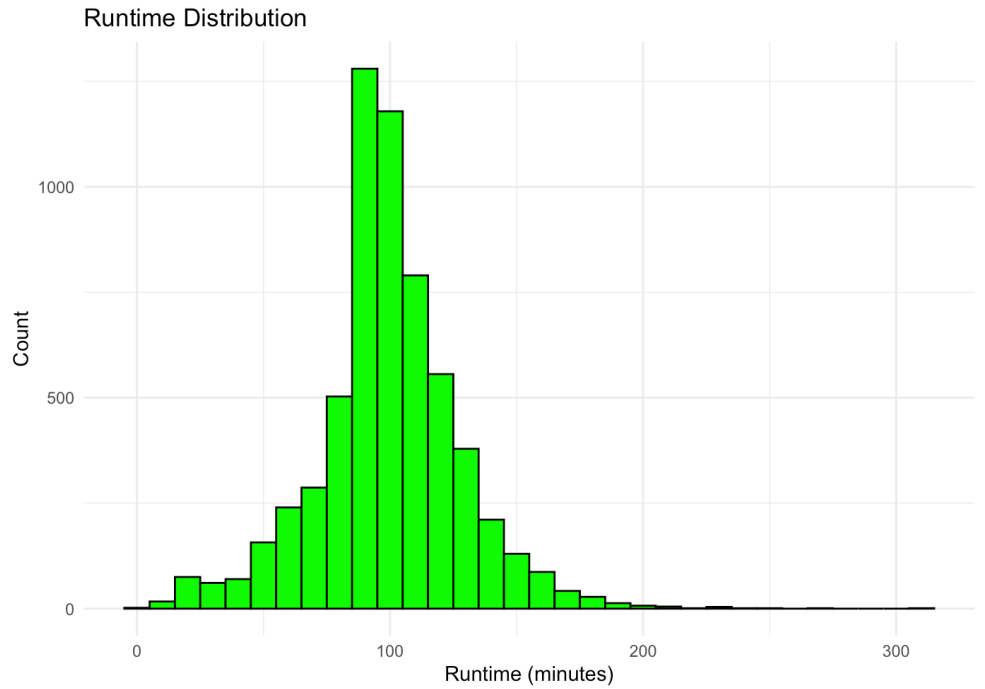

Figure4:Runtime Distribution of Netflix.

5. Top 10 Directors with Most Content:

The top ten filmmakers with the most Netflix material are shown. Among the most content directors are Jan Suter, Marcus Raboy, and Raúl Campos. This information can help with content acquisition and recommendation methods and is essential for comprehending the impact of directors on the platform

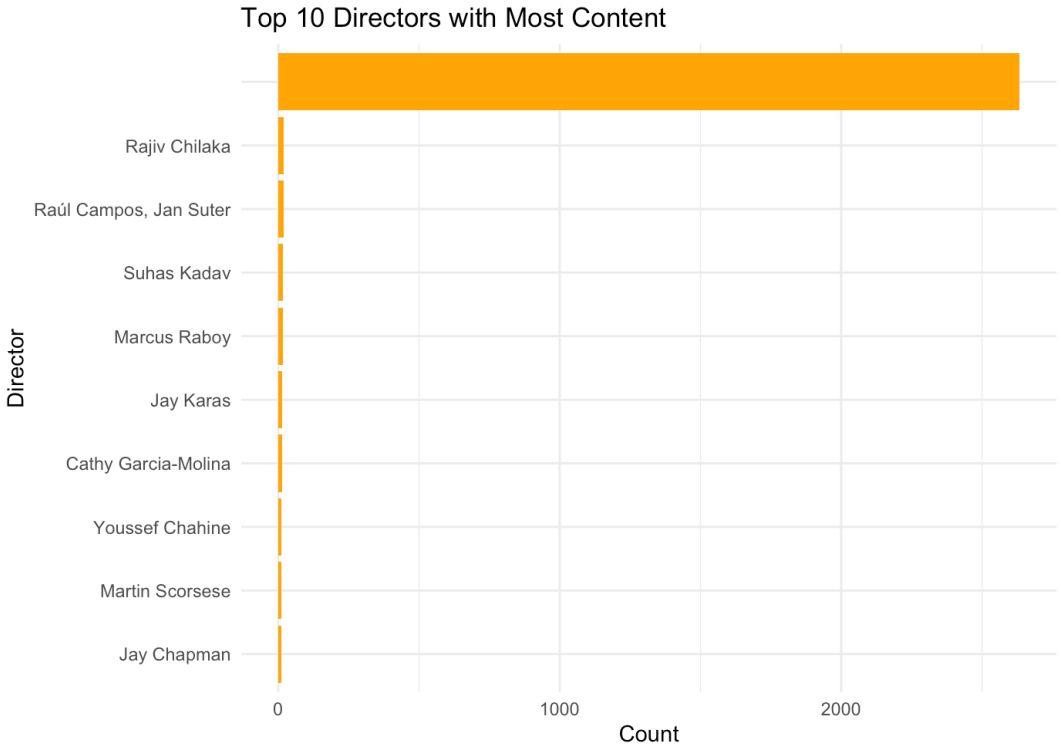

Figure5: Top 10 Directors with most Content

6.The Top Ten Genre:

Here is a list of the top 10 genres that Netflix offers. Comedies, dramas, and foreign films are the most popular genres on the platform. Understanding user preferences and content categorization requires this understanding.

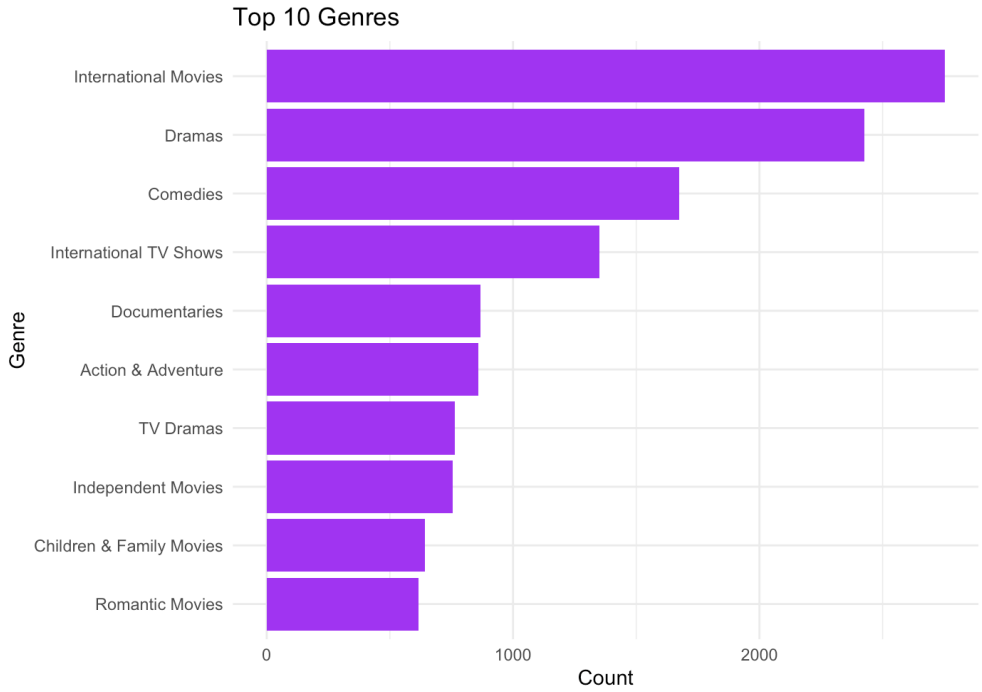

Figure6:Top 10 Genres According to User.

Key Findings:

1.Content Type Distribution:

The dataset includes a substantially higher proportion of movies than TV shows. The majority of the 8,807 entries that make up Netflix's content are movies. This is an interesting discovery because it reveals that Netflix subscribers prefer movies to TV series. Netflix and content makers need to understand this distribution in order to invest and buy more movies than TV series.

2.Distribution of Ratings:

On Netflix, TV-MA, TV-14, and TV-PG are the most popular ratings. The distribution of ratings provides information about the target demographic and age suitability of the Netflix material. This discovery is crucial for enhancing recommendation engines and classifying material. Netflix may adjust its content offerings to reflect the tastes and sensitivities of its user base by knowing the most popular ratings.

3.Distribution by Release Year: Over time, there has been a noticeable upsurge of releases, with a notable peak occurring after 2010. The distribution of release years sheds light on how Netflix's content has changed over time. The notable increase of releases subsequent to 2010 suggests the platform's growth and heightened expenditure on content generation and procurement. This discovery is noteworthy since it demonstrates Netflix's ongoing efforts to update and improve its catalogue.

4.Runtime Distribution: Ninety to one hundred and twenty minutes is the average runtime for content on Netflix. Comprehending the runtime distribution facilitates content classification and enhances recommendation systems. This knowledge is crucial for both content producers and users who have preferences for particular content durations, as it helps them create content that caters to the tastes of the platform's audience.

5.Top 10 Directors with Most Content:

Raúl Campos, Jan Suter, and Marcus Raboy are among the directors with the most content on Netflix. This finding emphasises the influence of directors on the platform and can aid in content recommendation and acquisition strategies. Identifying the top directors can help Netflix understand which directors are most appealing to its audience, allowing them to acquire more content directed by these individuals.
6.Top 10 Genres: The most popular genres on Netflix are comedies, dramas, and foreign films. Understanding user preferences and content categorization depends on this result. By identifying the most popular genres, Netflix can increase its investment in these areas and better cater its content offerings to its user base's tastes.

What really caught my attention was how much more common movies were than TV series. For Netflix and content providers to invest and acquire more movies than TV series, they must comprehend this inclination. Furthermore, the growing number of releases after 2010 points to Netflix's ongoing efforts to expand its library, which is fascinating to consider from the perspectives of corporate strategy and content acquisition.
Deeper study would be required in the next stages to comprehend the relationships between the various features. For example, determining whether runtime and user ratings are correlated, or whether the content's rating is influenced by the year of release. Further insights can be gained by investigating whether specific directors have a preference for any given genre. Sentiment analysis on the descriptions might also be done to comprehend user preferences more deeply.

Here are some Business Questions that can be Answered with Data Mining:
1.Content Strategy: What is the best mix of content (movies vs. TV shows) to attract and retain subscribers?
2.Recommendation System Improvement: How can Netflix enhance its recommendation system based on user ratings and preferred genres?
3.Content Acquisition: Which directors and genres should Netflix invest in to maintain and grow its subscriber base?
4.Platform Growth: How can Netflix continue its growth trajectory by leveraging the insights from the data to bring in new television shows and movies and attract new subscribers?

Conclusion:

In conclusion, the exploratory data analysis (EDA) of the Netflix dataset provided valuable insights into the content landscape of the streaming platform. The distribution of content types, ratings, release years, runtimes, top directors, and genres revealed patterns and trends that shed light on user preferences and content strategy. The prevalence of movies over TV shows, the dominance of certain ratings, and the surge in releases post-2010 highlight Netflix's evolving content strategy to cater to diverse audience tastes. Furthermore, the identification of top directors and popular genres underscores the importance of content acquisition and recommendation algorithms in maintaining and growing the platform's subscriber base. Overall, this EDA serves as a foundation for further analysis and decision-making processes, enabling Netflix to optimise its content offerings, enhance user experience, and drive platform growth in the highly competitive streaming landscape.

References:

1.Baek, H., Kim, J., & Ahn, J. (2019). Predicting Netflix Movies' Success: An Exploratory Data Analysis. arXiv preprint arXiv:1907.11294.
2.Bhatt, S., Aggarwal, A., & Kumar, S. (2020). Netflix Show and Movie Analysis. https://www.kaggle.com/shivamb/netflix-shows-and-movies-exploratory-analysis.
3.Davis, J. (2020). Netflix EDA and Visualization. https://www.kaggle.com/johndavisjr/netflix-eda-and-visualization.
4.Gilbert, P. (2019). Netflix Originals: Movies and TV Shows. https://www.kaggle.com/pavansubhasht/netflix-shows-and-movies.
5.Harish, G. (2020). An In-depth Analysis of Netflix Movies and TV Shows. https://www.kaggle.com/gshanbhag/an-in-depth-analysis-of-netflix.
6.Ji, J. (2020). Netflix Show and Movie Analysis. https://www.kaggle.com/jerryji/netflix-show-and-movie-analysis.
7.Kaggle. (n.d.). Netflix Movies and TV Shows. https://www.kaggle.com/shivamb/netflix-shows-and-movies.
8.Netflix. (n.d.). About Netflix. https://about.netflix.com/en.
Taseen, A. (2020). Netflix Exploratory Data Analysis (EDA). https://www.kaggle.com/taseendoshi/netflix-exploratory-data-analysis-eda.
9.Watson, A. (2020). Netflix Shows and Movies: EDA and WordCloud. https://www.kaggle.com/aldiwatson/netflix-shows-and-movies-eda-and-wordcloud.