IJCRT.ORG ISSN : 2320-2882
INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

NETFLIX DATA FOR ANALYSIS USING EXPLORATORY DATA ANALYSIS

Kshitija Marke[1] , Suhas Mache[2]
Department of Computer Application, JSPM University, Pune, India

ABSTRACT –
Netflix is one of the largest global streaming services with extensive library of movies, television series and original content available in various countries around the world. This paper employs Exploratory Data Analysis (EDA) approaches to Netflix data to discover the patterns related to the distribution of content, ratings of the material, tendencies in productions, as well as the contribution of different regions. First of all, the dataset is preprocessed to deal with missing information, to clean the data, as well as to standardize its format. Then, statistical analysis and visualization methods are applied to get insightful information. It was found out that movies prevail on the Netflix, and the most popular ratings are TV-MA and TV-14, which means that the company addresses the adult and teenage audience. The tendency towards the development of the Netflix is also clearly seen from the year 2010 onwards. In addition, important patterns are found related to sentiment distribution and movie duration.

Keywords: Netflix, Exploratory Data Analysis, Content Trends, Data Visualization, Streaming Analytics

I. INTRODUCTION

The development of streaming platforms has led to tremendous changes in the entertainment industry. Netflix, one of the major streaming platforms, offers its users access to numerous movies, TV series, and other types of media across more than 190 countries. The switch from traditional media and cable channels to streaming media has changed consumer behavior. The availability of large datasets and advancements in data analysis make it possible to perform detailed exploration of customer preferences and provide suggestions for improvement. While there is extensive literature related to building recommendation engines and prediction models, there are only a few studies regarding exploratory analysis of Netflix content.
The goal of this paper is to fill this gap and conduct Exploratory Data Analysis (EDA) of Netflix movies.

II. LITERATURE REVIEW

Data analytics have been instrumental in studying the consumption of media products and user behavior. In their study, [1] Gentzkow and Shapiro (2010) focused on analyzing the effect of data on media content and audiences opinions regarding such content. In another study by [2] Sun and Lim (2007), movie rating systems and users' reviews about movies were considered. [3] Jadhav et al. (2023) utilized association rule mining technique on data sets to discover patterns, which helps to understand users' behavior.

In the same line, [4] Maddodi and Prasad (2019) discussed the use of big data analytics by Netflix and its impact on recommendation systems to improve user experience. [5] McAuley and Leskovec (2013) analyzed rating systems and the way in which they represent different dimensions of opinions about products or services. [6] Balabanovic and Shoham (1997) introduced the concept of content-based recommendation systems, the base of today's streaming platforms. [8] Boukhelifa and Zhang (2015) explained the significance of using visualization techniques to interpret complex data sets.

III. OBJECTIVE

Objective

- Conducting an exploratory data analysis to investigate the trends related to the distribution of movies and TV shows on Netflix
- Conducting a sentiment analysis of Netflix content and identification of content rating and audience segmentation trends
- Conducting exploratory analysis of the changes in production trends over time
- Conducting exploratory analysis for identifying the trend of contributions made by each country
- Conducting exploratory analysis for investigating the trends of movie durations and sentiment analysis.

Research Questions

- What type of content is popular in Netflix dataset?
- Changes in Netflix content productions over time?
- Which countries contribute more to Netflix than others?
- Average movie durations in Netflix dataset?
- What would be concluded after sentiment analysis?

IV. PROBLEM STATEMENT

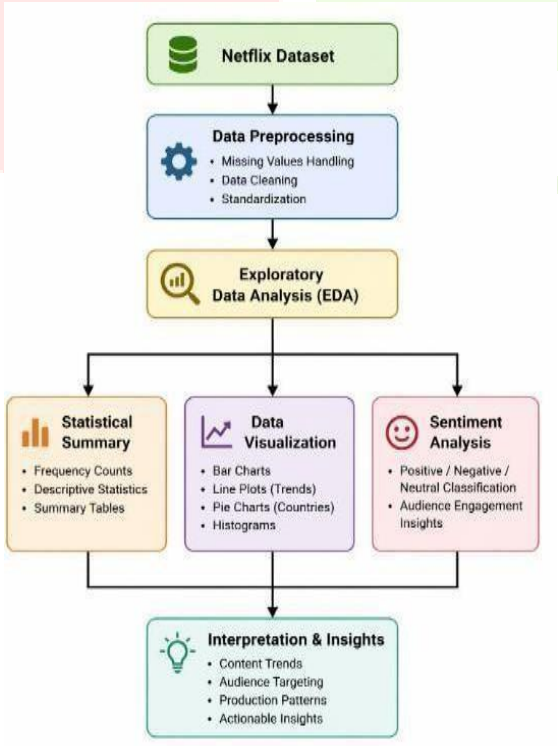
With increasing popularity of streaming platforms, there growing content on these platforms single day. This makes it hard to determine the type of content that audiences like and how it is distributed. Thus, the research on these trends and preferences should be conducted.

V. METHODOLOGY

The methodology adopted for pattern recognition in Netflix dataset will rely on a scientific and systematic approach line in Exploratory Data Analysis (EDA). For this purpose, the data set used in this study will comprise of more than 8000 records with variables such as title, type, director, cast, country, release year, rating, duration, and genre.

Preprocessing of data is performed initially because it is important in enhancing the quality of data set. Furthermore, standardization of data set is essential and involves application of proper data processing techniques. After the above mentioned steps, the next stage involves conducting EDA using programming tools such as Pandas, NumPy, Matplotlib, and Seaborn. At this point, initial statistical analyses are conducted and include mean, median, and frequency distributions. Apart from the above mentioned parameters, other processes include correlation analyses to determine relationships between variables and detecting outliers through graphical representation. Some of the visualization methods adopted during EDA include bar charts for studying the content type and rating distribution. In addition, line graphs are adopted for identifying production trends over time. Moreover, pie charts are applied for identifying the contribution of countries, while histograms are used to identify the duration of movies. Ultimately, the last process involved in the analysis includes sentiment analysis by categorizing content into three categories; namely, positive, negative, and neutral.

Flowchart Of EDA Process



VI. HISTORY OF NETFLIX

In 1997, Netflix was founded by Reed Hastings and Marc Randolph is a DVD renting company. In 1999, Netflix introduced a subscription-based model of renting DVDs without any late penalties. Netflix started a streaming platform in 2007, where people could stream contents right away from their computer. Since then, it spread all around the world, and now produces its own content.

VII. DATASET DESCRIPTION

The dataset employed in this study consists of extensive details concerning the Netflix contents such as show ID, title, type, director, actors, country, release year, rating, duration and genre. Movies and TV series are available in the dataset.

Fig. 1: Netflix Dataset Structure

Figure 1 shows the structure of the Netflix dataset utilized in this study. In the dataset there is listed various attributes of the contents including show ID, title, type, director, actors, country, release year, rating, duration, and genre. There are movies as well as TV series available in the dataset.

show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
s2	Movie	7.19	Demian Bichir	Hector Bonilla, Oscar Serrano	Mexico	December 23, 2016	2016	TV-MA	93 min	Dramas, International Movies	After a devastating earthquake hits Mexico City...
s3	Movie	23:59	Gilbert Chan	Tedd Chan, Stella Chung	Singapore	December 20, 2018	2011	R	78 min	Horror Movies, International Movies	When an army recruit is found dead...
s4	Movie	9	Shane Acker	Elijah Wood, John C. Reilly	United States	November 16, 2017	2009	PG-13	80 min	Action & Adventure, Sci-Fi	In a postapocalyptic world...
s5	Movie	21	Robert Luketic	Kevin Spacey, Kate Bosworth	United States	January 1, 2020	2008	PG-13	123 min	Dramas	A brilliant group of students...
s7	Movie	122	Yousri Nasrallah	Amina Khalil, Ahmed Dawood	Egypt	June 1, 2020	2019	TV-MA	95 min	Horror Movies, International Movies	After an awful accident...

This structured dataset will be helpful in conducting exploratory data analysis.

VIII. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis plays an essential role in summarization and visualization of the data to discover its inherent patterns. While statistical methods like the average and frequency distributions help us learn more about the distribution of the data, visualizations can help identify some trends and patterns that we might be unaware of. Missing value detection, outlier identification, and correlations are also part of EDA.

Table 1: Rating Distribution

Rating	Numbers of Titles	Percentage (%)
TV- MA	1200	35
TV-14	950	28
PG-13	450	13
R	400	12
G	150	4
Others	100	8

As per the exploratory data analysis of Netflix's data, there are several important trends regarding the distribution of the content, target audiences, and patterns of productions.

Observation: Major of the content Netflix is either TV-MA or TV-14.

Table 2: Movies vs TV Shows

Content Type	Number of Titles	Percentage (%)
Movies	2500	60
TV Shows	1650	40

Observation: There are more movies than TV shows on Netflix which indicates their strategy of maintain large film library.

Figure 1: Comparison between Movies and TV Shows

Figure 6 above illustrates the fact that movies have a larger number than TV shows.

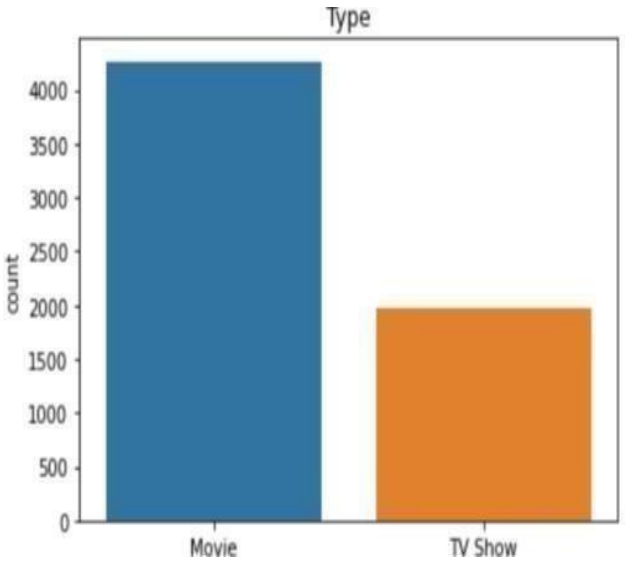


Fig 1: Comparison between Movies and TV Shows This could probably be due to the fact that Netflix focuses more on movies since they are more popular and less difficult to produce than other forms of media contents. However, this could probably also be due to the mechanism through which Netflix creates variations of movies.

Fig 2: Distribution of Movie and Show Ratings

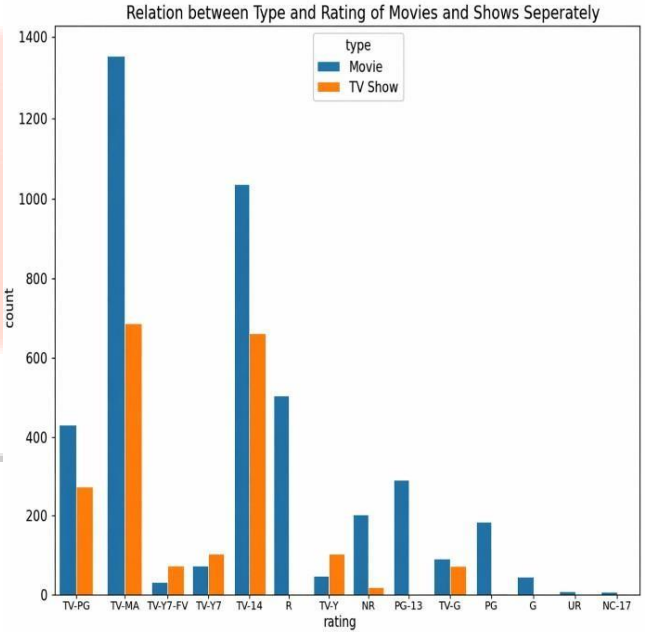


Fig 2: illustrates the distribution of ratings of the movies and shows offered by Netflix. It is observed that there are more frequencies of ratings like TV-MA and TV-14 in the dataset which means that these ratings occur most often.

It can be observed from above distribution that movies and shows produced by Netflix target adults and adolescents. Such ratings are indicative of the efforts put forth by Netflix to create superior-quality movies and shows, thereby increasing their viewer base.

Fig 3: Content Production Trend

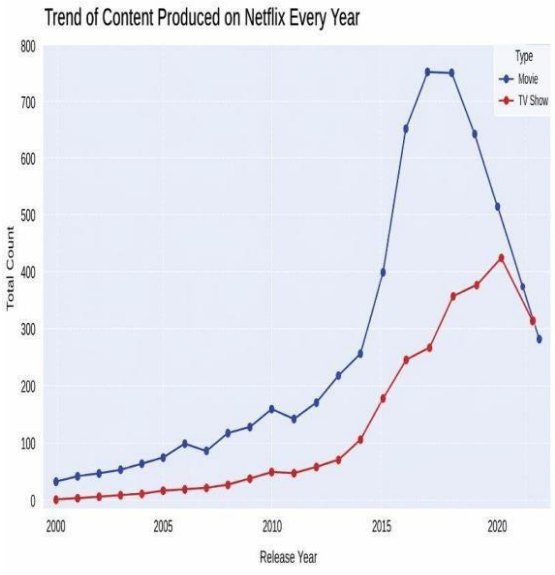


Fig 2: Content Production Trend

As shown above, rising trend signifies the quick growth of Netflix on a global scale. The declining trend after 2019 might be attributable to external factors.

Fig 4: Content Production Contribution by Countries

As illustrated in the figure below, the content production contribution by various countries in Netflix is presented. From Fig. 4, it is evident that the USA ranks first in content production at Netflix, followed by many other countries.

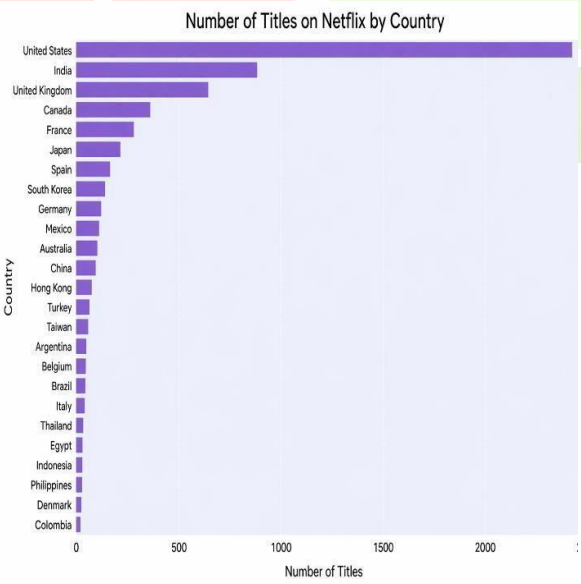


Fig 4: Distribution of Countries The largest portion of content is provided by the USA. However, there are some other countries apart from the USA, indicating the globalized approach followed by Netflix along with its presence in the US.

Fig 5: Sentiment Analysis

The sentiment analysis related to Netflix is illustrated in Fig. 5, wherein the number of contents according to their nature of sentiments is depicted in percentages.

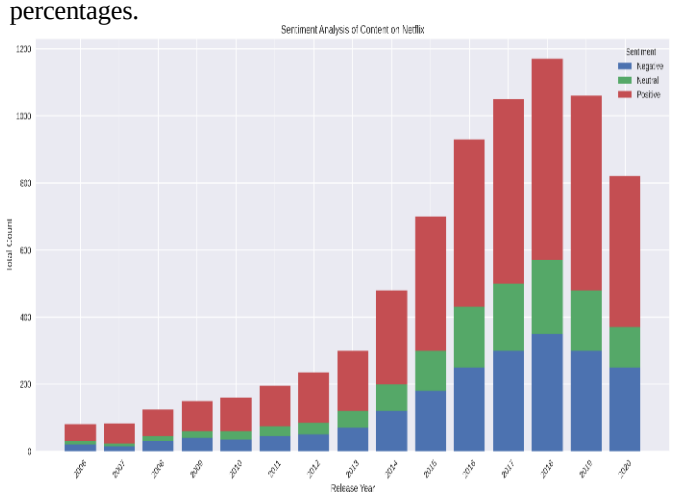


Fig 5: Sentiment Analysis

Based on the sentiment analysis, it may be seen that the percentage of positive contents is increasing. This indicates that Netflix has its focus on creating interesting contents for its viewers.

Figure 6: Comparison between Movies and TV Shows

Figure 6 above illustrates the fact that movies have a larger number than TV shows.

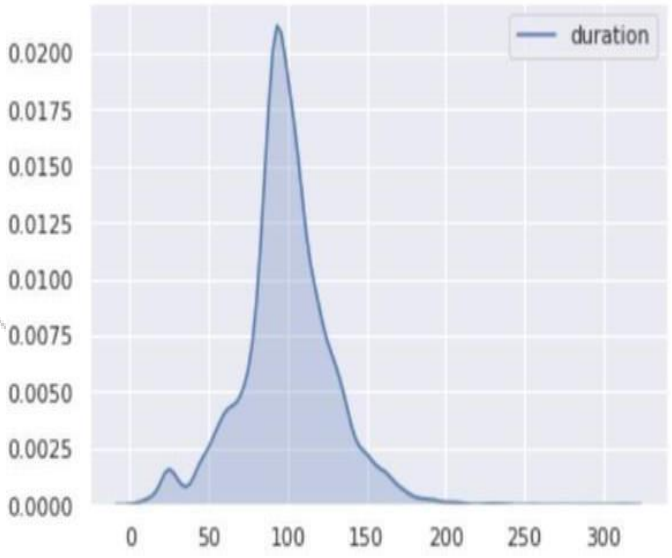


Fig 6: Movie Duration Distribution

Based on the above graph, it can be concluded that Netflix prefers ideal movie duration because of its convenience for viewing and higher viewership. In addition, moderating the movie duration ensures easier consumption of the movie. This is especially so because movies with durations that exceed the specified movie durations cater to specific types of audience groups.

X. LIMITATIONS

- There may be missing information on the latest content added to Netflix
- The dataset may contain some missing values which could affect analysis results
- The sentiment analysis could be influenced by the availability of relevant features

XI. FUTURE WORK

- Forecasting future trends through machine learning
- Creating recommendation engines for users
- Real-time streaming data analysis
- Using deep learning approaches for analysis

XII. CONCLUSION

This paper conducts an exhaustive examination of Netflix's content using Exploratory Data Analysis methods. Some major trends in content distribution, ratings, production increases, and viewer interactions were revealed.

It was found that movies make up the largest content on Netflix, and TV-MA and TV-14 are among the top-rated. It shows the increasing number of productions that the platform has achieved. Also, the international contributions indicate the expansion of Netflix beyond the United States.

XI. REFERENCES

[1] M. Gentzkow and J. M. Shapiro, "What drives media slant? Evidence from U.S. daily newspapers," *Quarterly Journal of Economics*, vol. 125, no. 1, pp. 35–71, 2010.

[2] L. Sun and E. P. Lim, "Movie rating and review aggregation in recommender systems," in *Proc. ACM SIGIR*, 2007, pp. 219–226.

[3] A. Jadhav, S. Patil, and R. Kulkarni, "Association rule mining in large datasets for pattern discovery," *SSRN Electronic Journal*, 2023.

[4] S. Maddodi and K. Prasad, "Big data analytics in Netflix: Improving recommendation systems and user experience," 2019.

[5] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proc. 7th ACM Conf. Recommender Systems*, 2013.

[6] M. Balabanović and Y. Shoham, "Content-based, collaborative recommendation," *Communications of the ACM*, vol. 40, no. 3, pp. 66–72, 1997.

[7] N. Castelo, "Content discovery and recommendation systems at Netflix," 2019.

[8] N. Boukhelifa and D. Zhang, "Visual analytics for data exploration: A survey," *IEEE Transactions on Visualization and Computer Graphics*, 2015.

[9] D. Sharma and A. B. Saxena, "Exploratory data analysis and visualization of Netflix data using Python libraries," *International Journal of Creative Research Thoughts*, vol. 11, no. 10, 2023.

[10] K. B. Vadloori and S. M. Sanghishetty, "Exploratory and sentiment analysis of Netflix data," *International Journal of Engineering Research & Technology*, vol. 10, no. 9, 2021.

[11] H. Trivedi and M. Shah, "An exploratory analysis on Indian Netflix's genres," *International Journal of Engineering Research & Technology*, vol. 10, no. 9, 2021.

[12] A. R. Jadhav *et al.*, "Netflix data analysis using exploratory data analysis," in *Lecture Notes in Networks and Systems*, 2024, pp. 309–317.

[13] S. R. Banu and B. Sravani, "Netflix movies and TV shows data analysis," *International Journal for Research in Applied Science & Engineering Technology*, 2023.

[14] J. Peng *et al.*, "DataPrep.EDA: Task-centric exploratory data analysis for statistical modeling in Python," 2021.

[15] "Netflix movies and TV shows dataset," Kaggle, 2021.