Netflix Movies and TV Shows – Data Analysis using Exploratory Data Analysis (EDA)

Akhil Surnedi(C0865688)
Supriya Karnam(C0867461)
Artificial Intelligence and Machine Learning
Lambton College in Toronto

***Abstract*—The primary objective of the project is to explore the dataset to gain insights into the trends and patterns of movies and TV shows available on Netflix. The EDA techniques such as data cleaning, data visualization, and statistical analysis are applied to the dataset to gain a better understanding of the data. The project also aims to identify the most popular genres, countries, and directors based on the data available in the dataset**

Keywords— Netflix, Movies, TV Shows, Data Analysis, Exploratory Data Analysis (EDA), Data Cleaning, Data Visualization, Statistical Analysis, Genres, Countries, Directors.

I. Introduction

Netflix is a popular streaming platform that offers a wide variety of movies and TV shows to its viewers. This project aims to study the dataset of movies and TV shows available on Netflix using a method called Exploratory Data Analysis (EDA). This method helps us to understand the data better by cleaning, visualizing, and analyzing it. By analyzing the data, we will be able to identify the most popular genres, countries, and directors of movies and TV shows on Netflix. The findings of the project can be helpful for those who create and distribute movies and TV shows to understand what the viewers like and create content that matches their interests. The project aims to provide insights into the trends and patterns of movies and TV shows available on Netflix and to demonstrate the effectiveness of EDA in analyzing large datasets.

II. Objectives

The primary objectives of the project are as follows:

1. To explore the dataset using EDA techniques to gain insights into the trends and patterns of movies and TV shows available on Netflix.
2. To identify the most popular genres, countries, and directors based on the data available in the dataset.
3. To analyze the ratings of the movies and TV shows and identify the factors that affect the ratings.
4. To create visualizations that provide a better understanding of the data and facilitate the interpretation of the results.

III. Methodology

The project will involve the following steps:

1. *Data cleaning:*

   The first step in the methodology is data cleaning, which involves removing irrelevant or inconsistent data to ensure the accuracy of the analysis. The dataset is then transformed into a suitable format that can be analyzed using EDA techniques.

2. Data visualization:

   Data visualization techniques such as bar charts, line charts, scatter plots, and heatmaps are used to gain insights into the data. These visualizations help identify trends, patterns, and outliers in the data, which can be used to draw conclusions and make recommendations.

3. Statistical analysis:

   Statistical analysis is then performed to validate the findings obtained from the visualizations. This involves calculating the mean, median, mode, and standard deviation of the data. The analysis helps identify statistical relationships between different variables in the dataset, such as the relationship between the rating of a movie or TV show and its release year or genre.

IV. Data Set

The dataset used in this project has 12 columns and 7789 rows. Each row represents a movie or a TV show available on Netflix, and the columns provide information about different aspects of the content.

1. **show_id**: This variable represents the unique ID for each TV show or movie available on Netflix. It is an alphanumeric variable.
2. **type**: This variable represents whether the entry is a TV show or movie. The possible values are "TV Show" and "Movie".
3. **title**: This variable represents the title of the TV show or movie. It is a text variable
4. **director**: This variable represents the name of the director of the TV show or movie. It is a text variable.
5. **cast**: This variable represents the names of the main actors and actresses in the TV show or movie. It is a text variable.
6. **country**: This variable represents the country or countries where the TV show or movie was produced. It is a text variable.
7. **date_added**: This variable represents the date when the TV show or movie was added to Netflix's catalog. It is a date variable.

8. **release_year**: This variable represents the year when the TV show or movie was released. It is a numeric variable.
9. **rating**: This variable represents the maturity rating of the TV show or movie. The possible values are "TV-MA", "TV-14", "TV-PG", "R", "PG-13", "PG", "G", "NR", and "UR".
10. **duration**: This variable represents the length of the TV show or movie in minutes or seasons. It is a text variable.
11. **listed_in**: This variable represents the categories or genres that the TV show or movie falls under. It is a text variable.
12. **description**: This variable represents a brief summary or description of the TV show or movie.

The overall purpose of this dataset is to provide information on the TV shows and movies available on Netflix, including their titles, directors, casts, release years, ratings, durations, genres, and descriptions. This dataset could be used to analyze the popularity of different genres, the frequency of certain directors or actors, or the relationship between release year and popularity.

V. Planning

Week	Task	Responsibility
1	Familiarize with the dataset and understand the objectives of the project.	Akhil Supriya
2	Data cleaning: remove duplicates, handle missing values, and correct errors. Analyze the distribution of each column.	Akhil Supriya
3	Data visualization: create charts and graphs to explore the trends and patterns in the data. Explore the relationships between different columns.	Akhil Supriya
4	Statistical analysis: perform statistical tests to validate the observations made in the previous weeks. Analyze the statistical significance of the trends and patterns observed in the data.	Akhil Supriya
5	Identify the most popular genres, countries, and directors based on the data available in the dataset. Draw insights from the data to understand the preferences of the viewers.	Akhil Supriya
6	Prepare the final report and presentation of the project. Discuss the findings and conclusions of the project. Reflect on the learnings and challenges faced during the project.	Akhil Supriya

VI. Results

The project is expected to provide insights into the trends and patterns of movies and TV shows available on Netflix. The analysis of the ratings is expected to identify the factors that affect the ratings and provide recommendations to improve the ratings. The visualizations will facilitate the interpretation of the results and provide a better understanding of the data.

VII. References

[1] S. Bansal, “Netflix Movies and TV Shows,” Kaggle dataset, 2021. [Online]. Available: https://www.kaggle.com/shivamb/netflix-shows.

[2] S. Pachori, "Performing EDA of Netflix Dataset with Plotly," Analytics Vidhya, Sep. 14, 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/09/performing-eda-of-netflix-dataset-with-plotly/.

[3] L. Tagliaferri, "Exploratory Data Analysis in Python," DigitalOcean, Aug. 7, 2019. [Online]. Available: https://www.digitalocean.com/community/tutorials/exploratory-data-analysis-python.

[4] R. M. A. Heeren, "Imaging Mass Spectrometry: A New Tool to Investigate the Spatial Organization of Metabolites within Biological Tissues," Journal of the American Society for Mass Spectrometry, vol. 23, no. 3, pp. 497-501, Mar. 2012. doi: 10.1007/s13361-012-0516-6

[5] S. K. Mukhiya and U. Ahmed, "Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize and investigate your data," Packt Publishing, Mar. 2021.