Data Analysis Project

Name: Shubham Murali Prajapati

Semester: 5 (B.Sc. Data Science)

College Name: Viva

Roll No: 22

Introduction

In today's digital era, online streaming platforms like Netflix have revolutionized the entertainment industry by offering thousands of movies and TV shows across various genres and languages. This project focuses on performing an in-depth analysis of Netflix's content database using various Data Science tools and techniques. It involves cleaning, processing, visualizing, and interpreting the data to explore trends related to content type, release year, genres, countries, and much more. Netflix uses dozens of tools in a layered architecture for data analysis. If you're looking for a number: it's 30+ core tools across categories, but it depends on how granular you get.

Problem Statement

Thousands of shows and movies available on Netflix make it difficult to identify trends without data analysis. This project helps answer questions like:

- What types of content (movies vs. shows) are most common?
- Which countries produce the most content on Netflix?
- Who are the top actors and directors featured on Netflix?
- To analyze Netflix's dataset and find trends in movies and TV shows.
- To visualize Netflix's content growth over the years.
- To explore the distribution of Netflix content by genre and country.
- To study Netflix's data for creating basic content recommendations.
- To build visual dashboards showing Netflix content trends.

Tools Used

Language: Python 3.x

Libraries: Pandas, NumPy, Matplotlib, Seaborn, WordCloud

Environment: Jupyter Notebook / Google Colab

Dataset and Attributes

We have used the Netflix dataset which contains 12 attributes: show_id, type, title, director, cast, country, date_added, release_year, rating, duration, listed_in, and description.

Methods

- Data Cleaning: Identifying and handling missing values by dropping or filling them.
- Data Preprocessing: Converting date columns into proper formats and analyzing categorical variables.
- Data Visualization: Creating graphs, bar charts, histograms, and word clouds to explore data trends.
- Statistical Analysis: Calculating counts, frequencies, and distributions of different attributes.
- Algorithms: This project does not involve complex machine learning algorithms as the main goal is to understand and analyze the data.

In the future, algorithms like Clustering, Classification, or Collaborative Filtering can be used to build recommendation systems based on this data.

References

https://www.kaggle.com/datasets/shivamb/netflix-shows