EDA ON NETFLIX MOVIES AND TV SHOWS DATASET

PRESENTED BY:- ALISHA RANI
GUIDED BY:- HIMANSHU SOUDA
COLLEGE:- TECHNO INDIA UNIVERSITY

INTRODUCTION:

This project focuses on performing Exploratory Data Analysis (EDA) on the Netflix Movies and TV Shows dataset to understand the structure, patterns, and trends within the platform's content. By analyzing features such as type, genre, release year, country, duration, and ratings, the project aims to uncover useful insights about Netflix's catalog and how it has evolved over time. The analysis helps transform raw data into meaningful information that supports better understanding of user preferences and Netflix's content strategy.

LIBRARIES AND TECHNOLOGIES USED:

- Programming Language: Python

- Libraries Used:
  - Numpy for data manipulation
  - Matplotlib
  - Sckit learn for predictive modeling
  - Joblib for model serialization

- Deployment and Interface:
  - Streamlit for rapid development
  - Render for cloud deployment

- Dataset Source:
  - KaggleMymoviedb.csv provides rich features for Netflix shows EDA
  - Development in Jupyter Notebook and version control with GitHub

ALGORITHM:

❑ Import Required Libraries

- Load Pandas, NumPy, Matplotlib, and Seaborn.

❑ Load the Dataset

- Read the CSV file into a DataFrame.

❑ Explore Dataset Structure

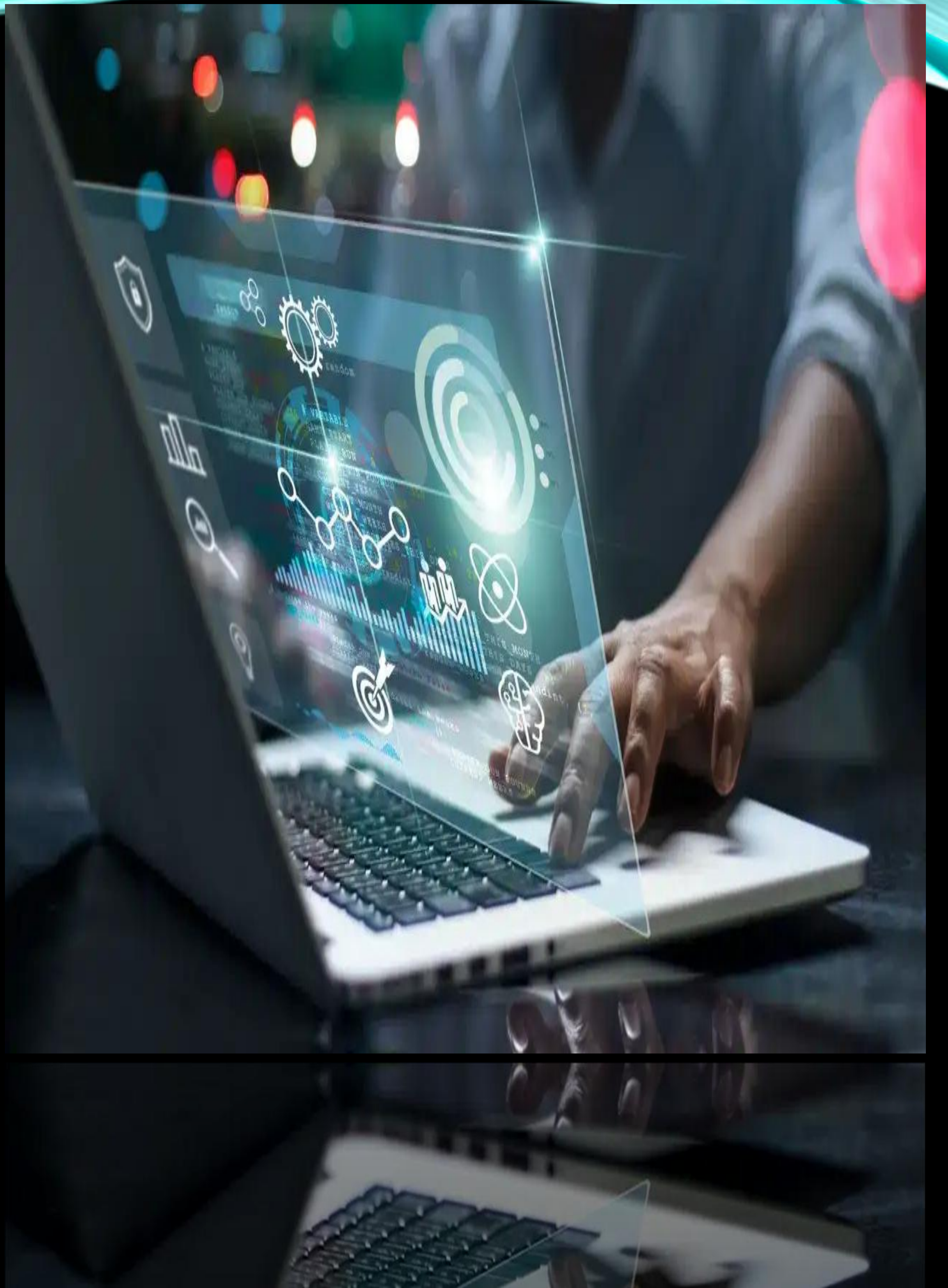
- Check shape, data types, missing values, and preview sample records.

❑ Data Cleaning

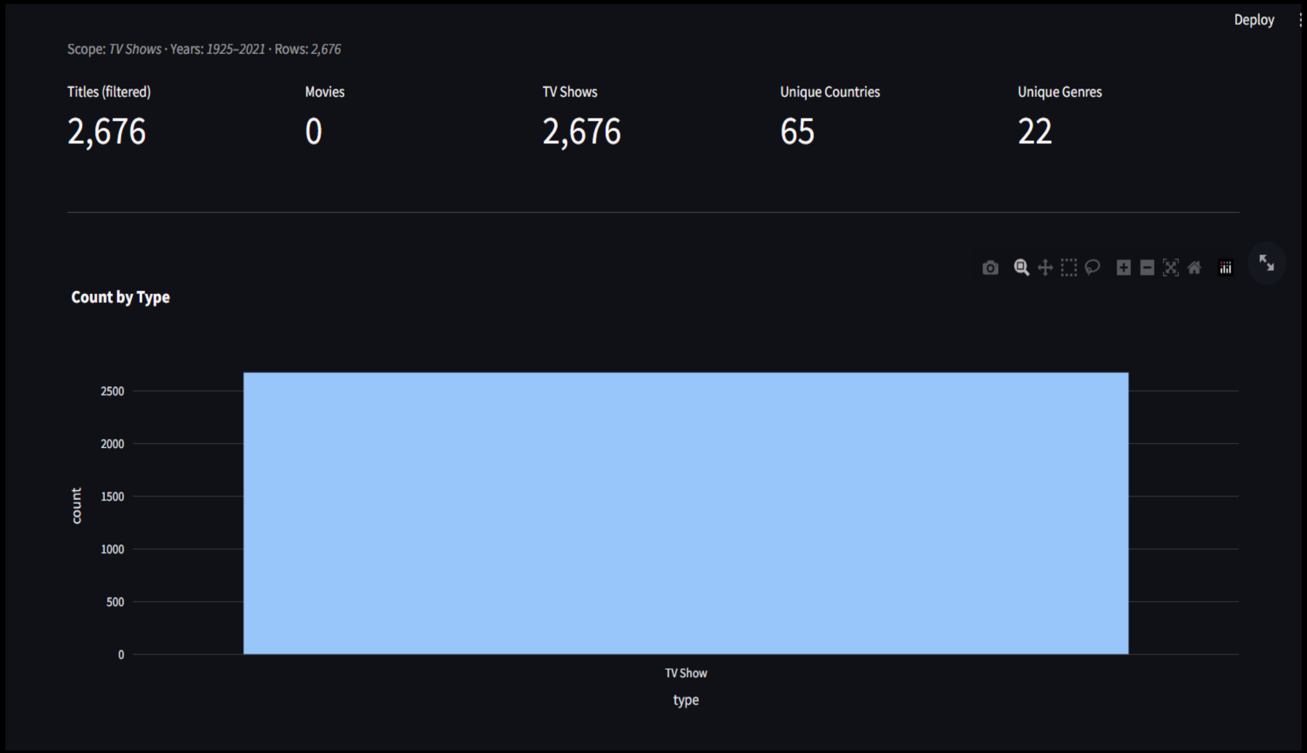
- Remove duplicates, handle missing data, and format columns properly.

❑ Data Visualization

- Create bar charts, histograms, count plots, heatmaps, and word clouds.



PROJECT WORKING:



Scope: *TV Shows* · Years: *1925–2021* · Rows: *2,676*

Titles (filtered) 2,676

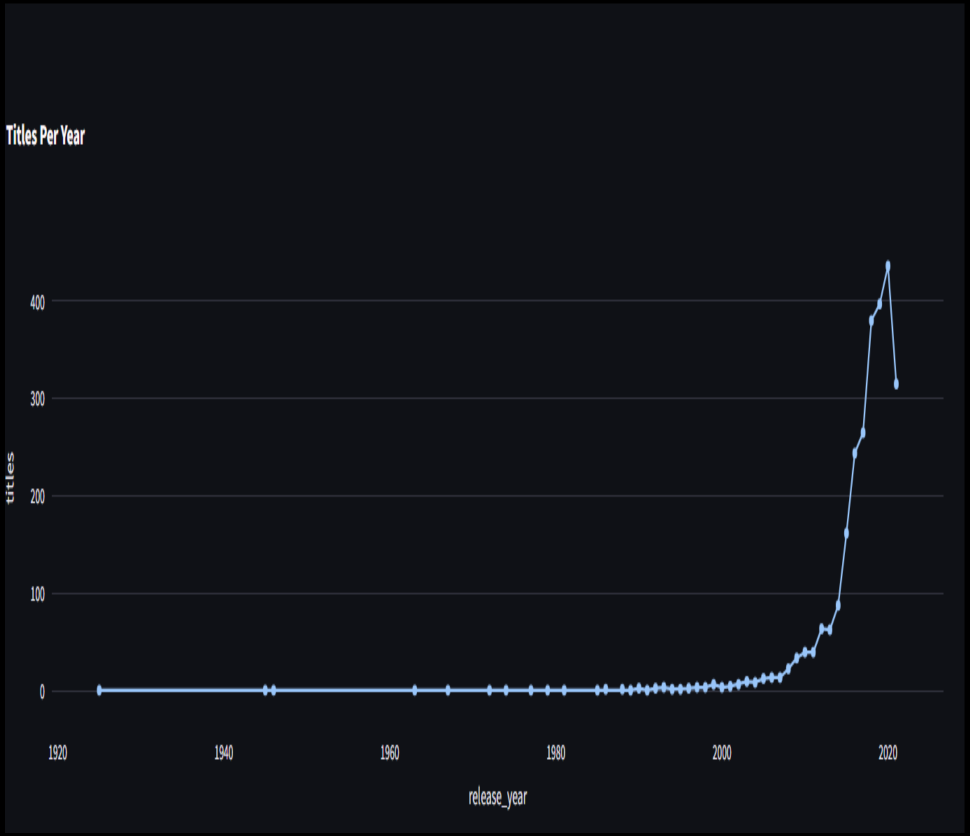
Movies 0

TV Shows 2,676

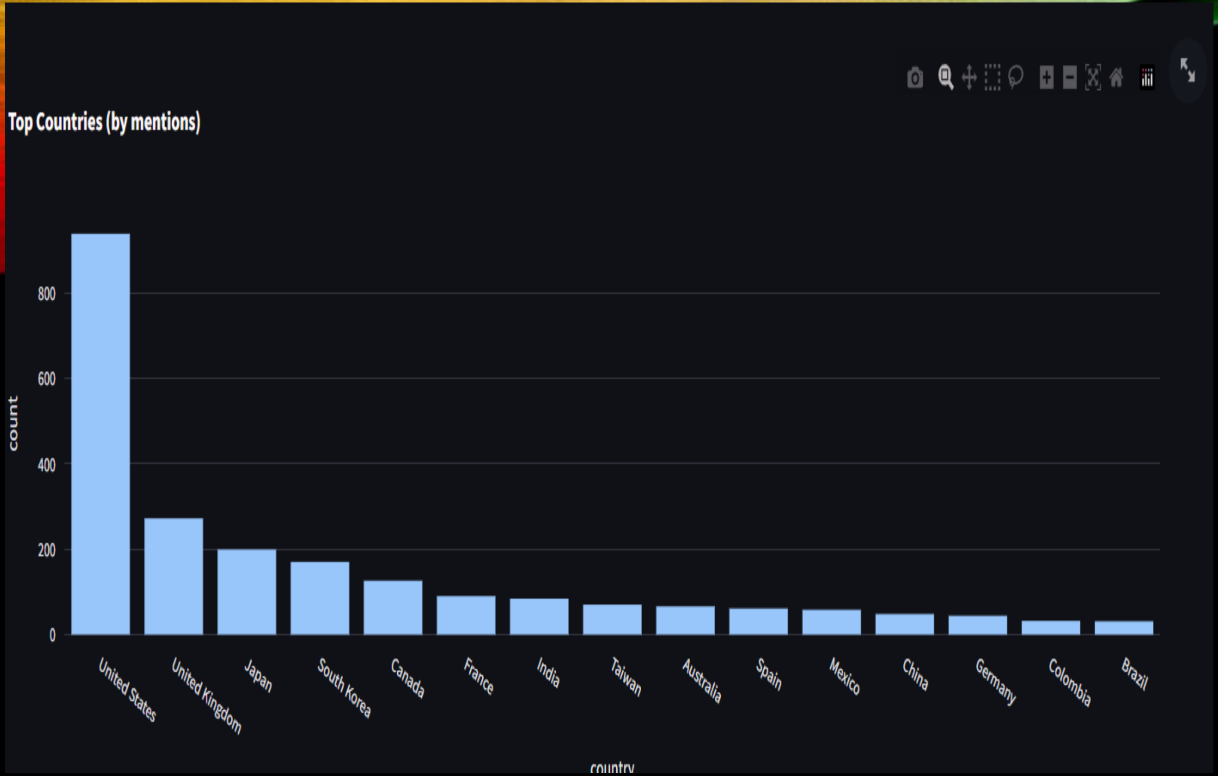
Unique Countries 65

Unique Genres 22

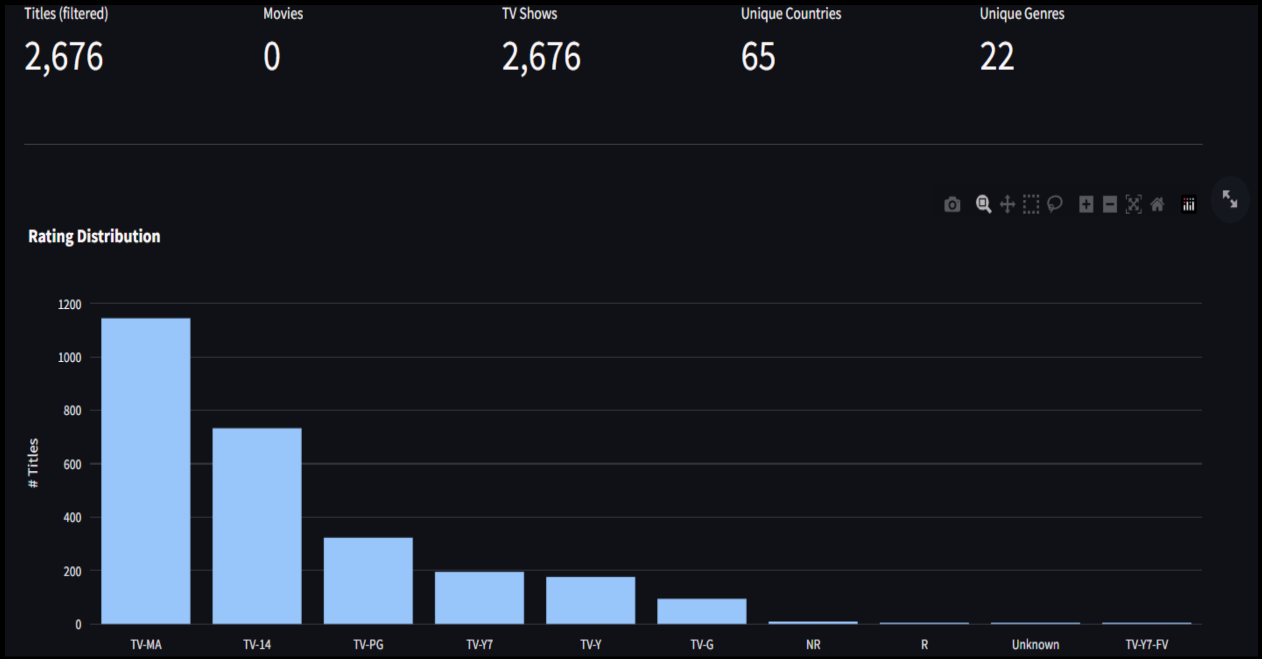
Count by Type



Titles Per Year

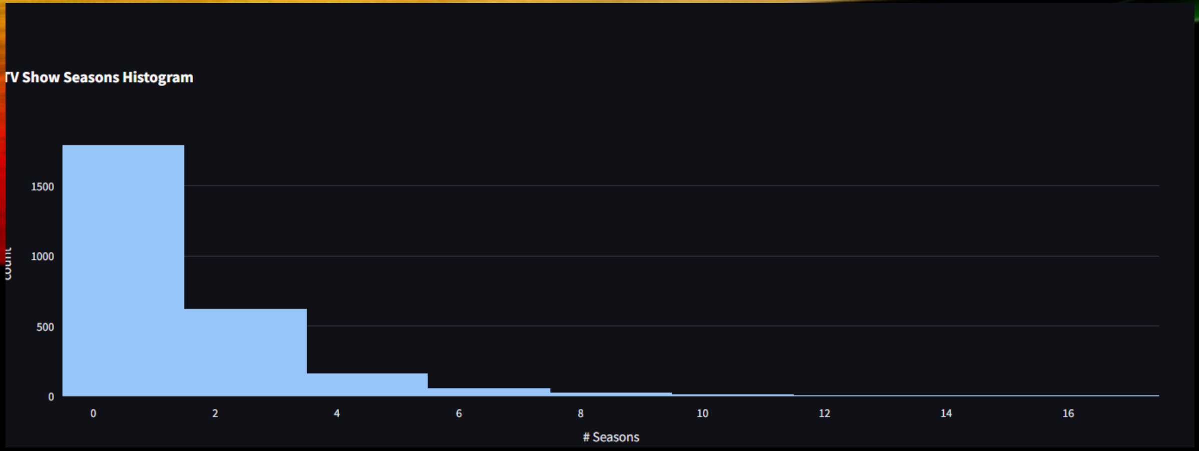


Top Countries (by mentions)

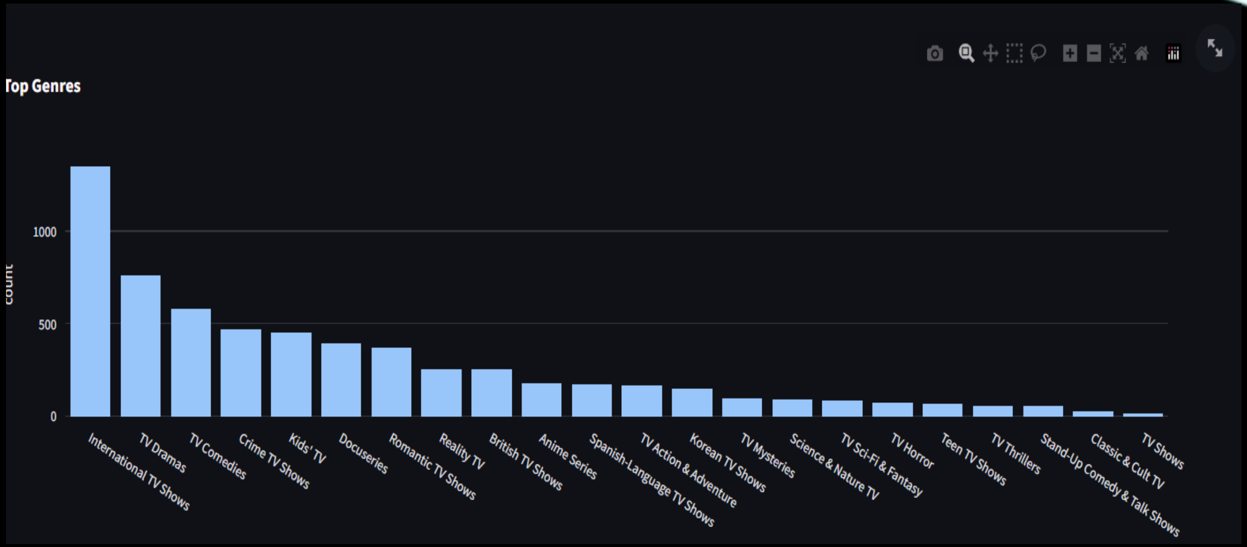


Titles (filtered)	Movies	TV Shows	Unique Countries	Unique Genres
2,676	0	2,676	65	22

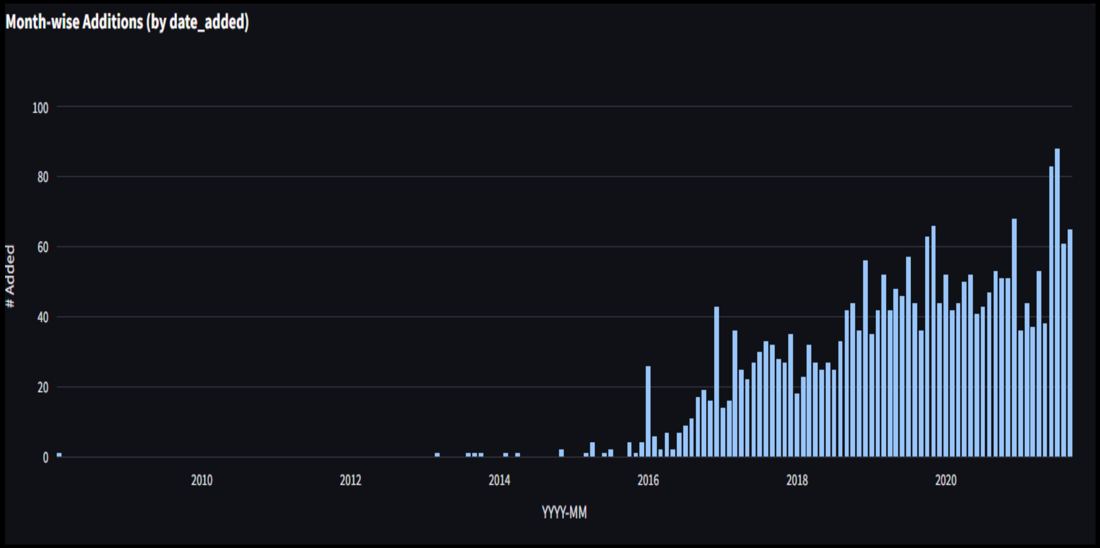
Rating Distribution



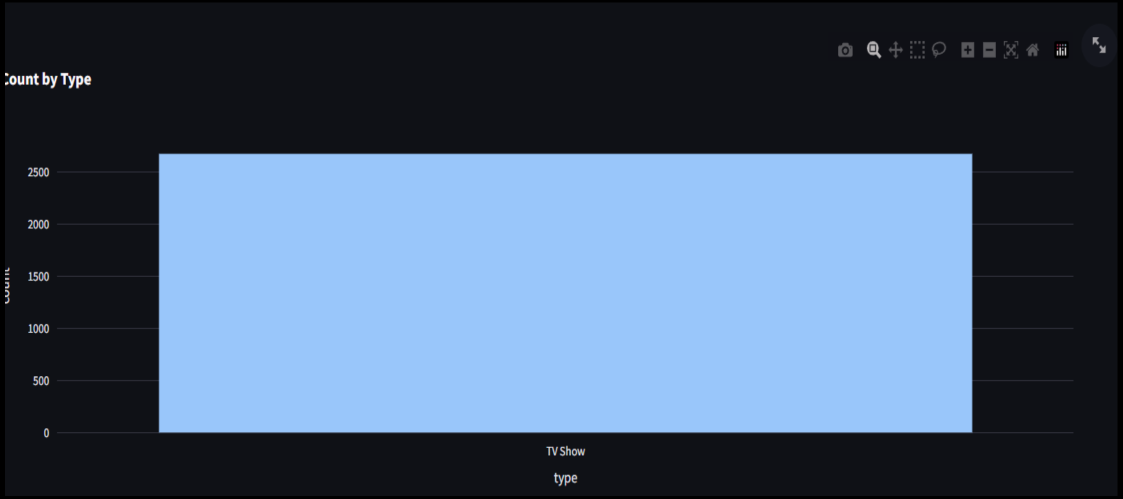
TV Show Seasons Histogram



Top Genres



Month-wise Additions (by date_added)



Count by Type

PROBLEM FACED AND SOLUTIONS:

CHALLENGES

- Missing values in important columns like director, cast, and country.
- Duplicate entries affecting the accuracy of analysis.
- Inconsistent formats in date and duration fields.
- Multiple values (genres/countries) stored in a single cell.
- Difficulty in creating clear visualizations due to overlapping labels.
- Unbalanced dataset with more movies than TV shows.

SOLUTIONS

- Filled or removed missing values to improve data completeness.
- Removed duplicate records using Pandas functions.
- Cleaned and converted date and duration columns to proper formats.
- Split columns containing multiple values (like genres/countries) for better analysis. Improved visualizations by adjusting chart size, label rotation, and formatting.
- Used percentage-based analysis to handle unbalanced data between movies and TV shows.

CONCLUSION

The EDA of the Netflix dataset helped uncover important patterns related to content types, genres, countries, and release trends. The analysis showed how Netflix’s library has grown over the years, which genres are most common, and how movies dominate over TV shows. By cleaning and visualizing the data, the project provided a clear understanding of Netflix’s content strategy and global distribution. Overall, the project transformed raw data into meaningful insights that highlight key trends within the platform.

THANK YOU