Netflix Data Analysis

Netflix is a global streaming service platform that offers a wide variety of movies, TV shows, documentaries, and more across a wide range of genres and languages. Netflix uses advanced streaming technology to deliver high-quality video content across different devices, including smart TVs, smartphones, tablets, and computers.

Breakdown of Jupyter Notebook

1) Importing Libraries
2) Loading Dataset
3) Understanding Dataset: By using info(), shape, size, describe(), duplicated().
4) We have columns: Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added', 'release_year', 'rating', 'duration', 'listed_in', 'description'], dtype='object')
5) Cleaning Dataset: Firstly, we check null values by df.isnull().sum().
    - The we clean NaN values by replacing it with ‘Unknown’ & ‘Mode’. Example- df.director.fillna(value='Unknown',inplace=True).
    - Dropping unwanted column like duration.
    - Some Transformations.

Data Visualization

1. Content Type on Netflix: Netflix offers a significant number of both movies and TV shows, allowing users to choose from a diverse range of content types.
2. Create visualizations to represent the distribution of content over different genres: Popular genres on Netflix include Drama, Internatinal Movies, Standup Comedy, Documentary, and others, providing insights into viewer preferences and content diversity.
3. Visualize the distribution of content across release years: The change in Netflix release years shows an increase in newer content and quantity.
4. Analyze the distribution of content ratings: It reveals TV-PG, TV-MA, R, TV-14 are most popular in audience preferences. Understanding the spread of ratings helps Netflix target specific viewer demographics more effectively.
5. Analyze trends in the popularity of different genres over time: It provides information about different genres like comedies, Drame, Sports, Movies, British TV shows, documentaries according to ratings who has the highest rating according to different rating categories.
6. Investigate potential correlations between variables (e.g., ratings and duration): It reveals different waiting categories and their duration in minutes PG -13, PG, TV- PG have 30000 to 50000 and the highest duration is TV-MA, TV-14 and R rating categories.
7. Explore the geographical distribution of content: Netflix content originates from various countries, with the USA leading in production, followed by other countries like India and the UK, Japan indicating Netflix's global reach and diversity in content sourcing.
8. Title of TV Show that were released in India only: This approach allows you to identify specific TV shows that were released exclusively in India like Kota Factory, Devlok with Devdutt Pattnayak, Classic Legend, Bhag re mann, Dar Sabko Lagta Hai etc.

Conclusion

Analyzing a Netflix CSV dataset provides valuable insights into the types of content available, trends in release dates, popular genres, geographic distribution, and content duration. These insights can inform decisions related to content acquisition, audience targeting, and platform strategy, enhancing the overall viewer experience and platform performance.