Netflix Data Analysis and Visualization

Project Title : Netflix Content Cleaning, Analysis, and Visualization

Project Description

This project focuses on the analysis and visualization of Netflix content data, with a particular emphasis on the global catalog as well as specific markets such as India, South Korea, and the USA. The work involves processing, cleaning, and analyzing datasets related to movies and TV shows available on Netflix. The main objectives are to uncover trends, compute key statistics, and present insights through clear visualizations.

Approach

1. Data Cleaning and Preprocessing

Before proceeding, first run data_cleaning.py . This script performs essential data cleaning steps on netflix.csv and generates a cleaned dataset called cleaned_netflix.csv for future use.The initial phase involves robust data cleaning to ensure the reliability of the analysis. This includes:

- **Handling Missing Values:** Identifying and removing or imputing rows with null or NaN values in critical columns.
- **Removing Duplicates:** Eliminating any duplicate entries to maintain data integrity.
- **Standardization:** Ensuring consistent data formats across all relevant columns.

2. Data Analysis and Visualization

The project employs a structured approach to analyze content trends and generate statistical insights:

- **Content Trends by Year (TV shows per year,Movies per year):**
    - Separate analyses are conducted for TV shows and movies.
    - Visualizations are generated to show the number of releases per year, both for the overall global catalog and specifically for content originating from the USA, South Korea, and India. This helps in understanding content growth and regional focus over time.

- **Combined Content Trends (merged data for Movies and TV shows per year):**
    - Yearly trends for movies and TV shows are merged to provide a comprehensive overview of Netflix’s content acquisition patterns.
    - Visualizations are generated to illustrate the total number of content additions per year, combining both movies and TV shows. This helps in understanding the overall growth trajectory and strategic shifts in Netflix's content library over time.

- **Statistical Analysis (statistics):**
  - Analysis is performed on date_added (the date content was added to Netflix) and release_year (the original public release date of the Movie or TV show).
  - This helps in understanding the distribution of content additions over time versus the original production timelines, providing insights into Netflix's content licensing and production strategies.

Project Structure

```
Netflix
├── data_cleaning.py
├── img
│   ├── mergered_graph
│   ├── movies_per_year
│   ├── statistics
│   └── TV_Shows_per_year
├── mergered_graph
│   ├── India
│   ├── mergered_graph.py
│   ├── South_Korea
│   └── USA
├── movies_per_year
│   ├── India
│   ├── movies_per_year.py
│   ├── South_Korea
│   └── USA
├── netflix.csv
├── README.md
├── statistics
│   ├── date_added.py
│   ├── India
│   ├── release_year.py
│   ├── South_Korea
│   └── USA
└── TV_Shows_per_year
    ├── India
    ├── South_Korea
    ├── TV_Shows_per_year.py
    └── USA
```

Explanation of Folders:

- netflix_titles.csv: The original, raw dataset containing Netflix content information.
- cleaned_netflix.csv: The output file from data_cleaning.py, containing the processed and cleaned data ready for analysis.
- data_cleaning.py: The Python script responsible for all initial data cleaning and preprocessing steps.
- India/: This folder contains India.py, which is specifically dedicated to processing and generating visualizations for Netflix content data related to India.
- South_Korea/: This folder contains South_Korea.py, which is specifically dedicated to processing and generating visualizations for Netflix content data related to South Korea.
- USA/: This folder contains USA.py, which is specifically dedicated to processing and generating visualizations for Netflix content data related to the USA.

Technologies Used

- **Programming Language:** Python
- **Key Libraries (Anticipated):** numpy, pandas for data manipulation, matplotlib and seaborn for visualization.