Data Analysis using Python as a tool

Netflix's top 500 movies in the Dataset

Tools Used: Python

1. Which year has the most TV Shows?
2. How many Movies are made for kids?
3. Which director has the most films and TV shows produced?
4. How many trillers are produced all over?
5. How many movies are produced in the USA?
6. **Average Duration by Genre:** Compute each genre's average duration of movies and TV shows.
7. Which actors appear most frequently in Netflix movies?
8. Are there certain actors associated with higher-rated movies?

Results Displayed:

-We have found the Year with the most TV Shows, 2021 with 118 shows.

- Number of movies made for kids: 51

- The director with the most films and TV shows is Rajiv Chilaka with 16 productions.

- Number of thrillers produced: 40

- Number of movies produced in the USA: 180

- Count of Movies and TV Shows:

Type

Movie 330

TV Show 169

- Average Duration by Genre:

Listed\_in

Action & Adventure 101.000000

Action & Adventure, Anime Features 84.000000

Action & Adventure, Anime Features, Children & Family Movies 59.000000

Action & Adventure, Anime Features, International Movies 94.333333

Action & Adventure, Children & Family Movies, Comedies 95.000000

...

TV Dramas, TV Mysteries, TV Sci-Fi & Fantasy 3.000000

TV Dramas, TV Sci-Fi & Fantasy 3.000000

TV Dramas, Teen TV Shows 4.000000

TV Shows 1.000000

Thrillers 100.571429

- Top 5 Most Common Words in Descriptions:

a: 708

the: 459

to: 411

and: 352

of: 293

-Statistical measures: Relationship Between the Duration and Years

(See the Figure1 file attached to this project)

**Data Analysis Project Overview**

**Project Overview**

This project analyzes a dataset of Netflix movies and TV shows using Python. The primary goal was to perform various data analysis tasks, such as determining trends, exploring relationships, and generating statistical insights from the dataset.

**Steps Taken**

1. **Setup and Data Preparation**
   * Installed necessary Python libraries for data manipulation and analysis.
   * Loaded the dataset from a CSV file into a Pandas DataFrame.
2. **Data Exploration**
   * Examined the dataset to understand its structure and contents.
   * Cleaned the data by handling missing values and ensuring proper data types.
3. **Data Analysis Tasks**
   * Determined which year had the most TV shows.
   * Counted the number of movies made for kids.
   * Analyzed the count of movies and TV shows.
   * Explored statistical measures like the average and median duration of movies.
   * Identified the top 5 most common words in the descriptions of movies and TV shows.
   * Investigated the relationship between movie duration and release year.
4. **Visualization**
   * Created scatter plots to visualize relationships between different columns in the dataset.

**Libraries Used**

* pandas: For data manipulation and analysis.
* matplotlib: For data visualization and plotting graphs.
* seaborn: For statistical data visualization.
* shutil: For file operations like copying files.

**Major Issues and Solutions**

1. **Issue with File Path Handling**
   * **Problem**: Encountered issues with file paths and Unicode errors.
   * **Solution**: Used raw strings (prefix r) for file paths to avoid escape character issues.
2. **Library Installation Issues**
   * **Problem**: Missing libraries like matplotlib and seaborn.
   * **Solution**: Installed the required libraries using pip.
3. **Data Column Issues**
   * **Problem**: Mismatch in column names and data types.
   * **Solution**: Corrected column names and converted data types to match analysis requirements.
4. **Permission Errors During File Operations**
   * **Problem**: Permission denied errors while copying files.
   * **Solution**: Ensured correct file permissions and avoided system-specific directories.

**Summary**

This project successfully demonstrated the use of Python for data analysis, leveraging libraries such as Pandas for data manipulation and Matplotlib/Seaborn for visualization. Through various tasks and troubleshooting steps, the project provided valuable insights into the dataset, including identifying trends, understanding data distribution, and visualizing relationships.