Netflix Data : Cleaning, Analysis and Visualization

Explore the complexities of Netflix data analytics, focusing on essential tools like python , SQL, and Machine Learning for effective data cleaning, analysis and visualization.

~Ayush Raipure

The Importance of Data Analytics at Netflix

Key Methodologies and Insights

- Understanding Viewer Preferences:

Netflix uses data analytics to gain insights into viewer preferences, enhancing content curation.

- Optimizing Content Recommendations:

By analyzing user behavior, Netflix optimizes content recommendations tailored to individual tastes.

- Enhancing Use Experience:

Data analytics helps Netflix improve user experience through personal experience through personalized interfaces and suggestions.

- Advanced Data Cleaning Techniques:

Netflix employs advanced data cleaning methods to ensure the accuracy and quality of their datasets.
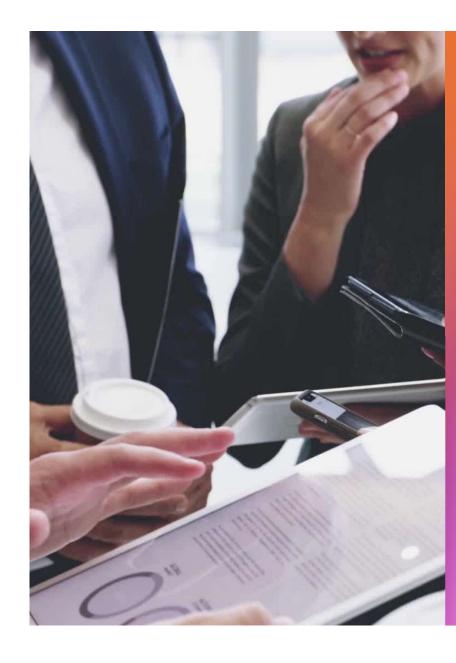
- Utilizing Machine Learning:

Machine learning algorithms are implemented to viewer habits and content popularity.

- Leveraging SQL for Data Management:

SQL databases are utilized for managing large volumes of data efficiently, supporting complex queries.

- Visualizing Data Insights:

Data visualization tools help Netflix present analytics insights clearly for informed decision-making.

Ensuring Data Quality

Data Cleaning Techniques at Netflix

- Importance of Data Cleaning:

Data cleaning is essential for accurate and reliable analytics.

- Handling Missing Values:

Imputation methods fills gaps to ensure data completeness.

- Outlier Detection:

Statistical methods identify anomalies that may skew analysis.

- Data Transformation:

Normalization and standardization maintain data consistency.\

- Duplicate Removal:

Algorithms eliminate rebundant records to enhance tha data quality.

Leveraging Python and SQL for Analysis

How Netflix Utilizes Python and SQL for Data Insights?

- Importance of Python in Data Analysis:

Python's libraries like Pandas and NumPy are vital for effective data manipulation and analysis.

- Visualization with Matplotlib:

Matplotlib enables clear visualization of complex data, enhancing and interpretability.

- SQL for Data Extraction:

SQL is essential for extracting and manipulating data from extensive databases at Netflix.

- Generating Reports with SQL:

SQL queries are crucial for generating reports that inform strategic decisions.

- Understanding Viewership Patterns:

Data analysis helps Netflix understand viewership patterns and user engagement.

- Actionable Insights:

Combining Python and SQL provides insights that influence content creation and engagement strategies.
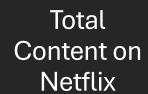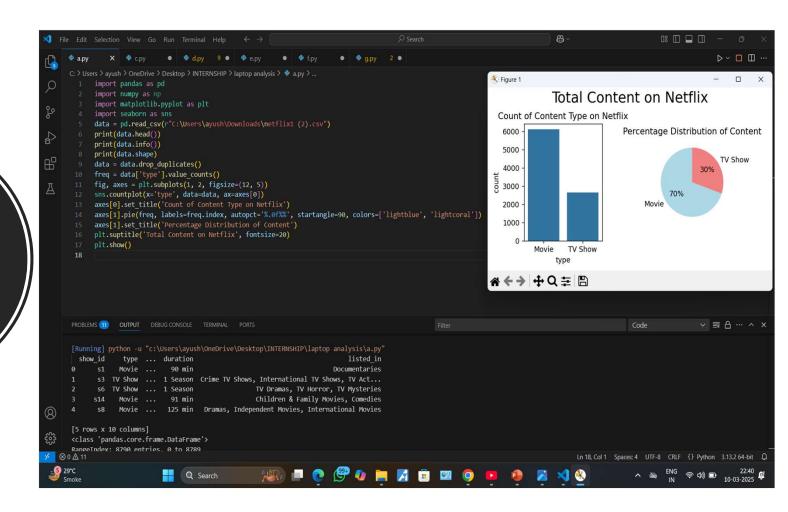
Data Cleaning & Analysis of Netflix Dataset

```
[Running] python -u "C:\Users\ayush\AppData\Local\Temp\tempCodeRunnerFile.python"
  show_id     type  ...  duration                                        listed_in
0      s1    Movie  ...    90 min                                     Documentaries
1      s3  TV Show  ...  1 Season  Crime TV Shows, International TV Shows, TV Act...
2      s6  TV Show  ...  1 Season             TV Dramas, TV Horror, TV Mysteries
3     s14    Movie  ...    91 min           Children & Family Movies, Comedies
4      s8    Movie  ...   125 min    Dramas, Independent Movies, International Movies

[5 rows x 10 columns]
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8790 entries, 0 to 8789
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   show_id       8790 non-null   object
 1   type          8790 non-null   object
 2   title         8790 non-null   object
 3   director      8790 non-null   object
 4   country       8790 non-null   object
 5   date_added    8790 non-null   object
 6   release_year  8790 non-null   int64
 7   rating        8790 non-null   object
 8   duration      8790 non-null   object
 9   listed_in     8790 non-null   object
dtypes: int64(1), object(9)
memory usage: 686.8+ KB
show_id         0
type            0
title           0
director        0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
dtype: int64
```
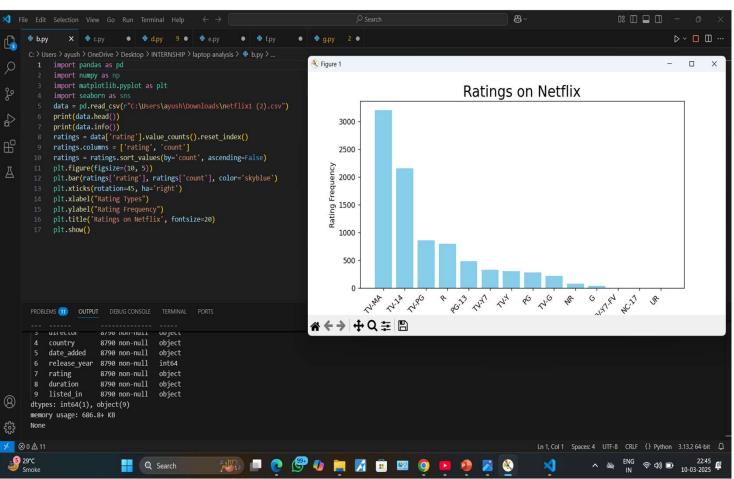
Data Visualisation
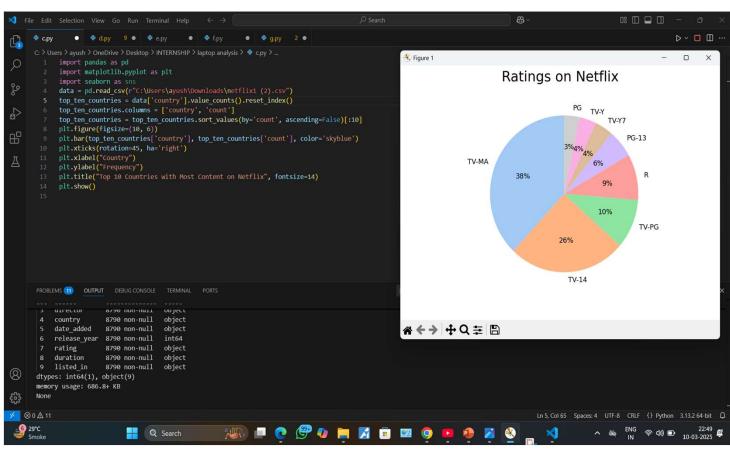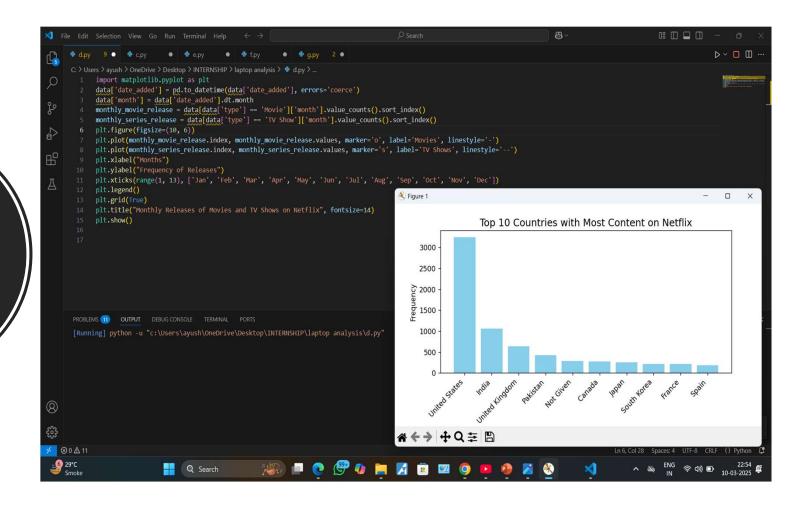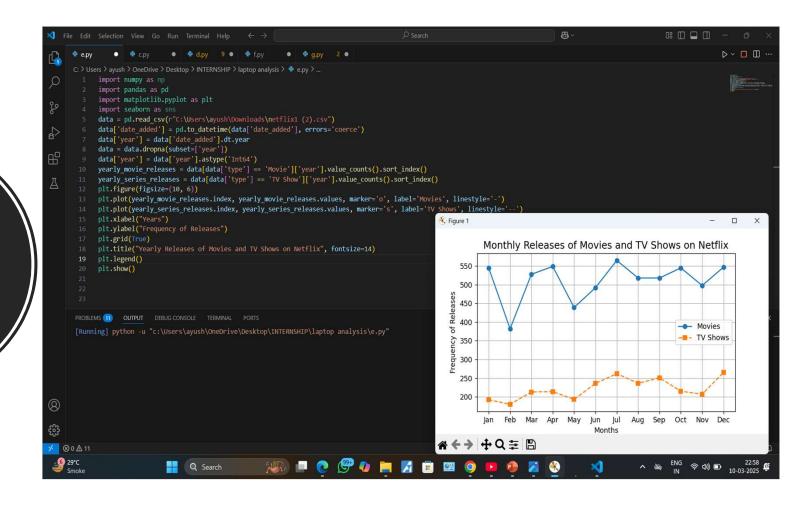
Total Content on Netflix

Ratings on Netflix

Ratings on Netflix

Top 10
Countries
with Most
Content on
Netflix

```python
import matplotlib.pyplot as plt
data['date_added'] = pd.to_datetime(data['date_added'], errors='coerce')
data['month'] = data['date_added'].dt.month
monthly_movie_release = data[data['type'] == 'Movie']['month'].value_counts().sort_index()
monthly_series_release = data[data['type'] == 'TV Show']['month'].value_counts().sort_index()
plt.figure(figsize=(10, 6))
plt.plot(monthly_movie_release.index, monthly_movie_release.values, marker='o', label='Movies', linestyle='-')
plt.plot(monthly_series_release.index, monthly_series_release.values, marker='s', label='TV Shows', linestyle='--')
plt.xlabel("Months")
plt.ylabel("Frequency of Releases")
plt.xticks(range(1, 13), ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov', 'Dec'])
plt.legend()
plt.grid(True)
plt.title("Monthly Releases of Movies and TV Shows on Netflix", fontsize=14)
plt.show()
```

C: > Users > ayush > OneDrive > Desktop > INTERNSHIP > laptop analysis > d.py > ...

Top 10 Countries with Most Content on Netflix

Frequency

United States, India, United Kingdom, Pakistan, Not Given, Canada, Japan, South Korea, France, Spain
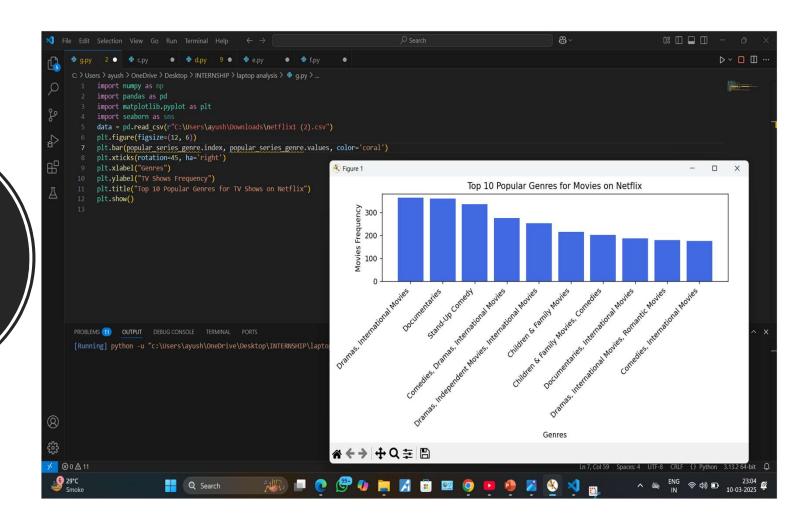
Monthly
Releases of
Movies and
TV Shows on
Netflix

Yearly Releases of Movies and TV Shows on Netflix
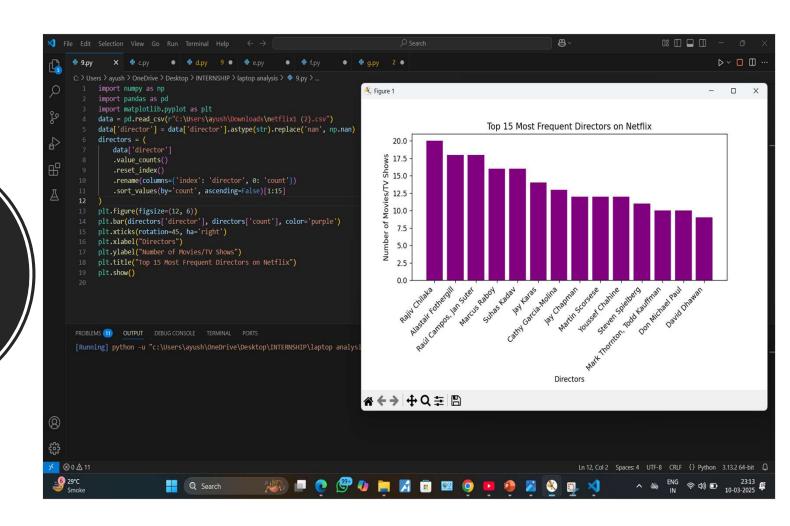
Top 10 Popular Genres for Movies on Netflix


```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
data = pd.read_csv(r"C:\Users\ayush\Downloads\netflix1 (2).csv")
plt.figure(figsize=(12, 6))
plt.bar(popular_series_genre.index, popular_series_genre.values, color='coral')
plt.xticks(rotation=45, ha='right')
plt.xlabel("Genres")
plt.ylabel("TV Shows Frequency")
plt.title("Top 10 Popular Genres for TV Shows on Netflix")
plt.show()
```
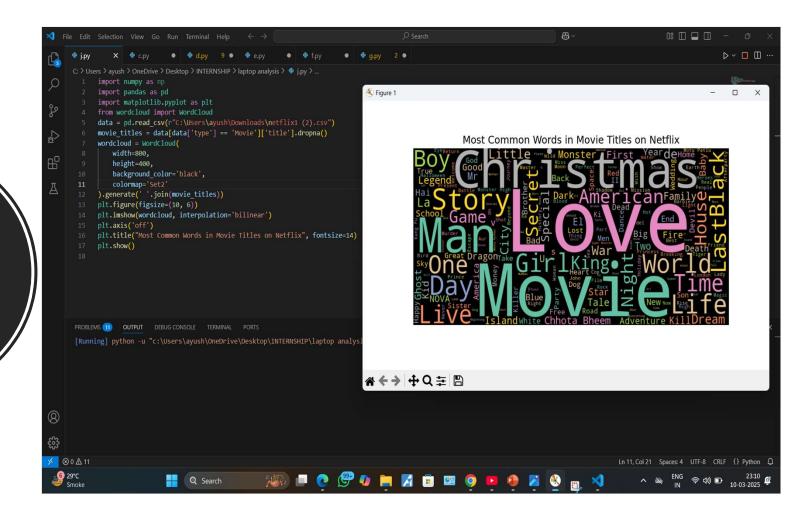
Top 10 Popular Genres for TV Shows on Netflix

Top 15 Most Frequent Directors on Netflix

Most Common Words in Movie Titles on Netflix

Key Insights from Netflix Analysis

Understanding Netflix's Data Driver Success

- Data Analytics as a Core Strategy

- Importance of Data Cleaning

- Visualization for Insight

- Machine Learning for Personalization

- Challenges in Data Management

Conclusion

This project involved **cleaning, visualizing, and analyzing** Netflix data to uncover content trends. We handled missing values, standardized formats, and ensured data consistency for accurate insights.

These insights help in content strategy and future recommendations. Further analysis with user engagement data could enhance decision-making.