Technical Analysis Report

Netflix Content Analysis: Trends, Insights, and Strategic Insights

1. Introduction

This document provides an in-depth analysis of the Netflix dataset using Python. It covers data cleaning, exploratory data analysis (EDA), and key insights derived from the dataset. The purpose of this analysis is to uncover trends in Netflix's content distribution, genres, country-wise availability, and key contributors such as directors and actors. This document also explores how Netflix's content has evolved over time, what types of shows dominate the platform, and the overall diversity of content offered to users globally.

With the rise of online streaming services, understanding content trends is crucial for business strategies. Netflix, as a pioneer in this industry, constantly updates its content library to cater to changing viewer preferences. This analysis aims to highlight patterns that drive content strategies and provide recommendations based on data-driven insights.

2. Dataset Overview

The dataset used in this analysis is netflix_titles.csv, which contains information about movies and TV shows available on Netflix. The dataset includes the following attributes:

- show_id: Unique ID for each show
- type: Movie or TV Show
- title: Name of the show
- director: Director(s) of the show
- cast: Cast members
- country: Country of production
- date_added: Date added to Netflix
- release_year: Year of release
- rating: Content rating
- duration: Duration of the show/movie
- listed_in: Genres
- description: Description of the show/movie

The dataset is essential for understanding the content available on Netflix, patterns in content additions, and distribution based on different factors such as country, genre, and duration. By analyzing these aspects, we can gain a deeper insight into Netflix’s content strategy and how it caters to various audience demographics.

3. Data Preprocessing

3.1 Handling Duplicate Records

Checked for duplicate records and removed them to maintain data integrity.

```
import pandas as pd

df = pd.read_csv("netflix_titles.csv")
df.drop_duplicates(inplace=True)
```

3.2 Handling Missing Values

Identified missing values across columns and visualized them using a heatmap.

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10,6))
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.show()
```

Missing values were then handled appropriately based on their relevance. For instance, missing director names were left as is, while missing country values were replaced with "Unknown" to avoid loss of data. Additionally, missing values in the date_added column were imputed with the mode to ensure that analysis based on content addition dates remains meaningful.

Handling missing values is crucial to ensure that our insights remain reliable. Without proper preprocessing, missing values can lead to biased interpretations, affecting trend analyses and forecasting.

4. Exploratory Data Analysis (EDA)

4.1 Year-wise Content Release

Analyzed the year with the highest number of movie and TV show releases.

```
df['release_year'].value_counts().head(10).plot(kind='bar', figsize=(10,5), color='red')
plt.xlabel('Release Year')
plt.ylabel('Count')
plt.title('Top 10 Years with Most Releases')
plt.show()
```

From the visualization, we observed that recent years (2018-2020) had the highest number of content additions, indicating an aggressive expansion strategy by Netflix. This aligns with Netflix's focus on producing and acquiring a large number of titles to compete with emerging streaming platforms.

4.2 Content Type Distribution

Counted the number of movies vs. TV shows and displayed insights using a count plot.

```
sns.countplot(data=df, x='type', palette='coolwarm')
plt.title("Distribution of Movies and TV Shows")
plt.show()
```

The dataset revealed that Netflix has a higher proportion of movies than TV shows. This suggests that Netflix primarily focuses on movies, although TV shows have seen significant growth over the past few years.

4.3 Country-Specific Analysis

Identified TV shows released exclusively in India.

```
df_india = df[(df['type'] == 'TV Show') & (df['country'] == 'India')]
print(df_india[['title', 'release_year']])
```

This helped analyze regional content production trends, showcasing Netflix's investment in localized content. India has emerged as a key market, and Netflix continues to increase its regional productions to cater to the local audience.

4.4 Top 10 Directors on Netflix

Extracted and ranked directors based on the number of movies/TV shows contributed.

```
top_directors = df['director'].value_counts().head(10)
top_directors.plot(kind='barh', color='blue')
plt.xlabel('Count')
plt.ylabel('Director')
plt.title('Top 10 Directors on Netflix')
plt.show()
```

4.5 Genre-Specific Filtering

Extracted movies categorized under "Comedies".

```
df_comedy = df[(df['type'] == 'Movie') & (df['listed_in'].str.contains('Comedies', na=False))]
print(df_comedy[['title', 'release_year']])
```

4.6 Actor-Based Analysis

Counted the number of movies and TV shows featuring Tom Cruise.

```
tom_cruise_movies = df[df['cast'].str.contains('Tom Cruise', na=False, case=False)]
print(tom_cruise_movies[['title', 'release_year']])
```

4.7 Content Ratings

Identified different ratings available on Netflix.

```
print(df['rating'].unique())
```

This provided insights into the classification of Netflix content, showing a significant number of PG-13 and TV-MA rated content.

4.8 Duration Analysis

Determined the longest movie or TV show available on Netflix.

```
df['duration'] = df['duration'].str.extract(r'(\d+)').astype(float)
print(df.nlargest(1, 'duration'))
```

4.9 Sorting Data by Year

Sorted and analyzed content release trends by year.

```
df_sorted = df.sort_values(by='release_year', ascending=False)
print(df_sorted[['title', 'release_year']].head(10))
```

5. Conclusion

This analysis provided valuable insights into Netflix's content library. The dataset was cleaned, visualized, and explored to identify trends in content distribution, genres, and country-wise availability. Some of the major findings include:

- Recent years (2018-2020) had the highest content additions.
- Netflix has a higher number of movies than TV shows.
- India has a growing number of Netflix-exclusive TV shows.
- "Comedies" and "Dramas" are among the most popular genres.

6. Future Scope

- Perform sentiment analysis on movie descriptions to understand content themes.
- Predict trends in content production using machine learning models.
- Analyze user preferences and recommendations based on viewing patterns.
- Conduct a deeper study on the popularity of content by region and genre.

This document serves as a comprehensive technical reference for analyzing Netflix's content using Python and data science techniques. The insights derived from this study can help streaming platforms refine their content strategy, improve recommendation systems, and understand global content trends effectively.