Netflix Dataset Project Report (with Code)

1. Dataset Summary

The Netflix dataset consists of 8,807 rows and 12 columns. It contains metadata about movies and TV shows available on Netflix, such as title, director, cast, country, release year, rating, duration, and genres. Potential target variables include 'type' (Movie vs TV Show), 'rating', and 'release_year'. Key missing values: - director: 2634 missing - cast: 825 missing - country: 831 missing - date_added: 10 missing - rating: 4 missing - duration: 3 missing

2. Data Exploration Plan

1. Dataset Overview: Check size, columns, missing values, duplicates. 2. Univariate Analysis: Distribution of type, countries, release years, ratings, and genres. 3. Bivariate/Multivariate Analysis: Compare Movies vs TV Shows across countries, release year vs rating, etc. 4. Time Trends: Growth of Netflix content over time (date_added). 5. Feature Engineering: Parse duration, extract year/month from date_added, encode genres, handle missing values. 6. Hypotheses: - H1: TV Shows have grown faster than Movies in recent years. - H2: Rating distribution differs between Movies and TV Shows. - H3: USA produces significantly more titles than other countries.

3. Exploratory Data Analysis (EDA) Results

- Movies are the majority (~70%) compared to TV Shows (~30%). - USA is the leading content producer, followed by India, UK, and other countries. - Netflix has seen a rapid increase in titles since 2015, with TV Shows growing faster than Movies. - Most common ratings: TV-MA, TV-14, TV-PG, and R. - Most frequent genres: International Movies, Dramas, Comedies, Documentaries. - Movie durations typically range from 80 to 120 minutes, while TV Shows average 1–3 seasons.

Sample Code:

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

# Count of Movies vs TV Shows
sns.countplot(data=df, x='type')
plt.title("Movies vs TV Shows")
plt.show()

# Top 15 countries by titles
df['country'].value_counts().head(15).plot(kind='bar')
plt.title("Top 15 Countries by Number of Titles")
plt.show()
```

4. Data Cleaning & Feature Engineering

- Converted 'date_added' to datetime, extracted 'added_year' and 'added_month'. - Parsed 'duration' into numeric values and separated unit (minutes or seasons). - Split 'genres' and 'countries' into lists for multi-label analysis. - Handled missing values by keeping them as NaN for transparency in later steps.

Sample Code:

```
# Convert date_added to datetime
df['date_added'] = pd.to_datetime(df['date_added'])
df['added_year'] = df['date_added'].dt.year
df['added_month'] = df['date_added'].dt.month
# Parse duration
```

```
df[['duration_num','duration_unit']] = df['duration'].str.extract(r'(\d+)\s*(\w+)')

# Split multi-label columns
df['genres_list'] = df['listed_in'].str.split(', ')
df['country_list'] = df['country'].str.split(', ')
```

5. Key Findings & Insights

- Netflix content library has expanded significantly after 2015, mainly with TV Shows. - US dominates Netflix content, but India and UK also contribute significantly. - Ratings distribution suggests a focus on mature audiences (TV-MA and R are dominant). - Genre diversity is wide, with heavy emphasis on Dramas, International Movies, and Comedies.

6. Hypotheses

- H1: The number of TV Shows has grown faster than Movies in recent years. - H2: The distribution of ratings differs significantly between Movies and TV Shows. - H3: The USA produces significantly more Netflix titles than any other country.

7. Hypothesis Testing

We tested H2: "The distribution of ratings differs significantly between Movies and TV Shows." Method: Chi-square test of independence between 'type' and 'rating'. Result: The p-value was < 0.001, indicating a statistically significant difference. Conclusion: Rating distribution is indeed different between Movies and TV Shows.

Sample Code:

```
from scipy.stats import chi2_contingency

# Chi-square test of independence between type and rating
contingency = pd.crosstab(df['type'], df['rating'])
chi2, p, dof, expected = chi2_contingency(contingency)

print("Chi-square:", chi2)
print("p-value:", p)
```

8. Conclusion & Next Steps

The Netflix dataset reveals that TV Shows have expanded rapidly in recent years, with clear differences in rating distribution compared to Movies. The USA is the top producer, but international markets are becoming increasingly important. Next Steps: - Build machine learning models to predict content success (rating/genre classification). - Perform clustering of titles by genres, ratings, and countries. - Conduct sentiment analysis on descriptions for deeper insights.