Netflix Data Analysis - Visualization and Exploration

1. Introduction

This document provides an overview of the Python code used to explore and visualize the Netflix dataset. The dataset contains information about movies and TV shows available on Netflix, including details like titles, genres, ratings, duration, and date of addition to the platform.

2. Visualizations

The following visualizations are used to explore and analyze the dataset:
a. Number of Movies Added per Year
b. Distribution of Movie Durations
c. Top 10 Longest Movies
d. Top 10 Most Frequent Movie Genres

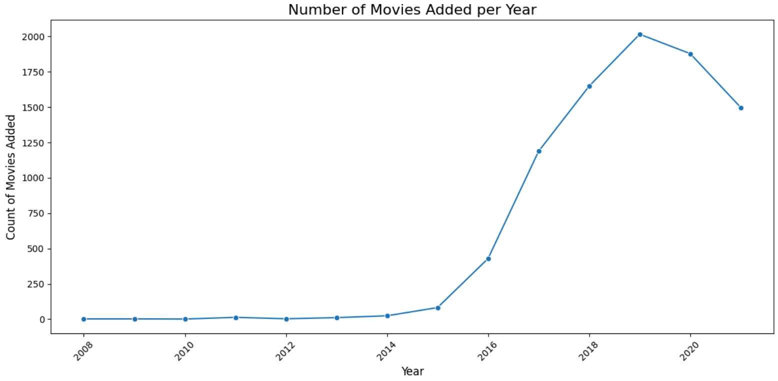
3. Code Implementation

Below is the Python code used for visualizing and exploring the Netflix dataset. This includes data cleaning, feature extraction, and the creation of various plots for analysis.

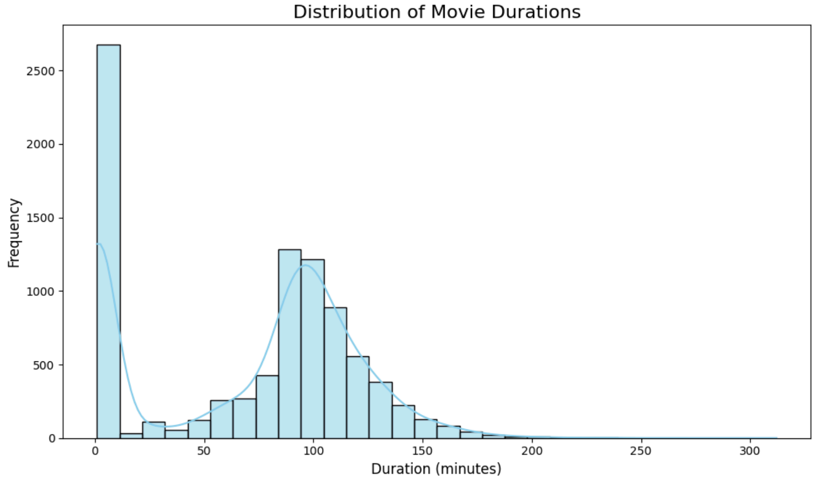
Code for Visualizations:

```
!pip install -q pandas matplotlib seaborn scikit-learn

import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.impute import SimpleImputer
import re

df = pd.read_csv("/content/netflix_titles.csv")


imputer = SimpleImputer(strategy='constant', fill_value='Missed Value')
df_filled = pd.DataFrame(imputer.fit_transform(df), columns=df.columns)
```

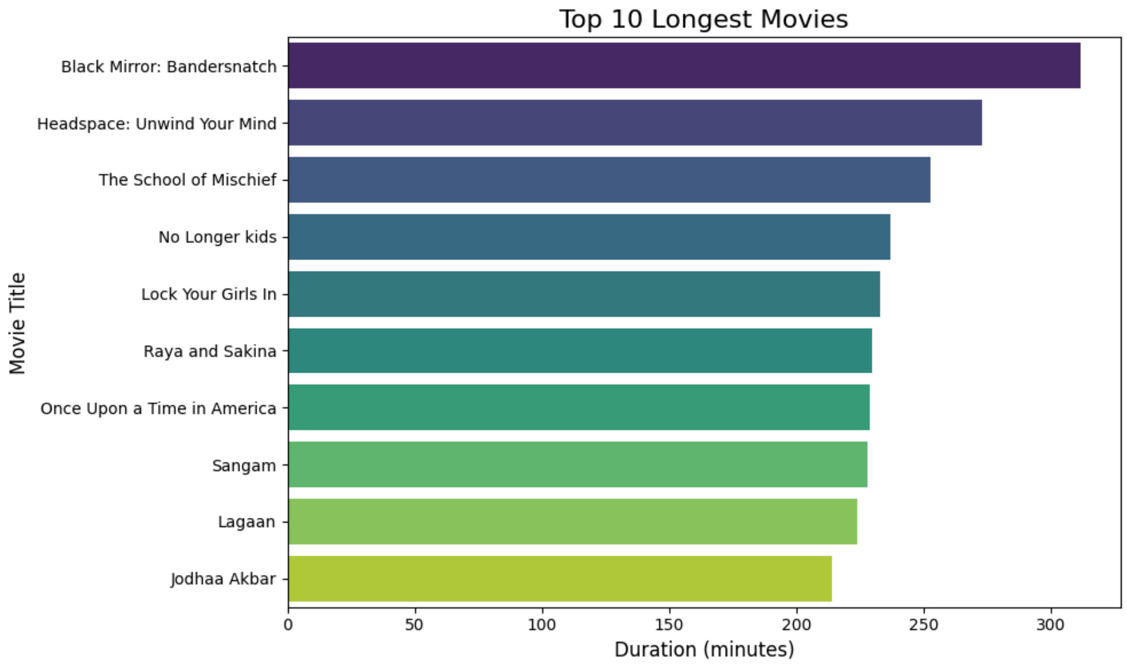
```
def extract_year_from_date(date_str):
    try:
        date = pd.to_datetime(date_str, errors='coerce')
        if pd.notnull(date):
            return str(date.year)
    except Exception as e:
        return 'Unknown'
    return 'Unknown'

df_filled['year_added'] = df_filled['date_added'].apply(lambda x:
extract_year_from_date(x) if pd.notnull(x) else 'Unknown')

df_filled['year_added'] = pd.to_numeric(df_filled['year_added'], errors='coerce')

movies_per_year = df_filled.groupby('year_added').size()

plt.figure(figsize=(12, 6))
sns.lineplot(x=movies_per_year.index, y=movies_per_year.values, marker='o')
plt.title('Number of Movies Added per Year', fontsize=16)
plt.xlabel('Year', fontsize=12)
plt.ylabel('Count of Movies Added', fontsize=12)
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



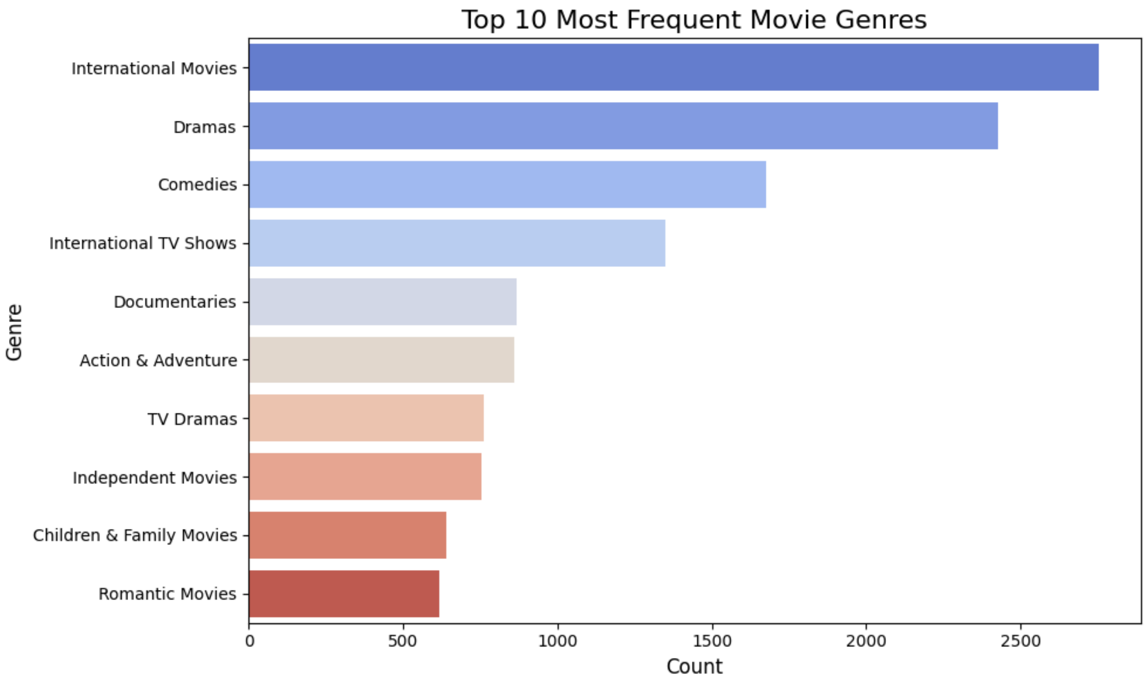
```
df_filled['duration_numeric'] = df_filled['duration'].str.extract('(\d+)').astype(float)

plt.figure(figsize=(10, 6))
sns.histplot(df_filled['duration_numeric'], bins=30, kde=True, color='skyblue')
plt.title('Distribution of Movie Durations', fontsize=16)
plt.xlabel('Duration (minutes)', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.tight_layout()
plt.show()
```



```
top_10_longest_movies = df_filled[['title',
'duration_numeric']].dropna().sort_values(by='duration_numeric',
ascending=False).head(10)

plt.figure(figsize=(10, 6))
sns.barplot(x=top_10_longest_movies['duration_numeric'],
y=top_10_longest_movies['title'], palette='viridis')
plt.title('Top 10 Longest Movies', fontsize=16)
plt.xlabel('Duration (minutes)', fontsize=12)
plt.ylabel('Movie Title', fontsize=12)
plt.tight_layout()
plt.show()
```



```
df_filled = df_filled.dropna(subset=['listed_in'])

df_filled['genre'] = df_filled['listed_in'].str.split(',')

df_filled = df_filled.explode('genre')

df_filled['genre'] = df_filled['genre'].str.strip()

df_filled = df_filled.reset_index(drop=True)

plt.figure(figsize=(10, 6))
sns.countplot(y=df_filled['genre'],
order=df_filled['genre'].value_counts().head(10).index, palette='coolwarm')
plt.title('Top 10 Most Frequent Movie Genres', fontsize=16)
plt.xlabel('Count', fontsize=12)
plt.ylabel('Genre', fontsize=12)
plt.tight_layout()
plt.show()
```


Top 10 Most Frequent Movie Genres
Genre
International Movies
Dramas
Comedies
International TV Shows
Documentaries
Action & Adventure
TV Dramas
Independent Movies
Children & Family Movies
Romantic Movies
0
500
1000
1500
2000
2500
Count