netflix-data-analysis

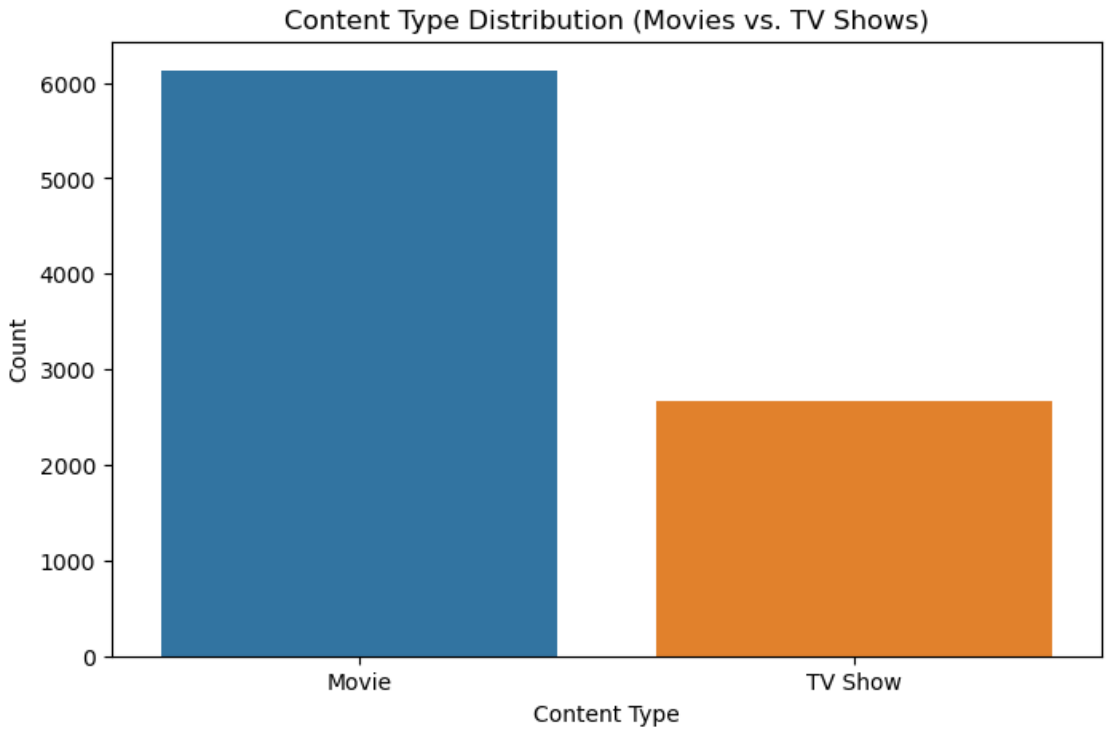
October 14, 2024

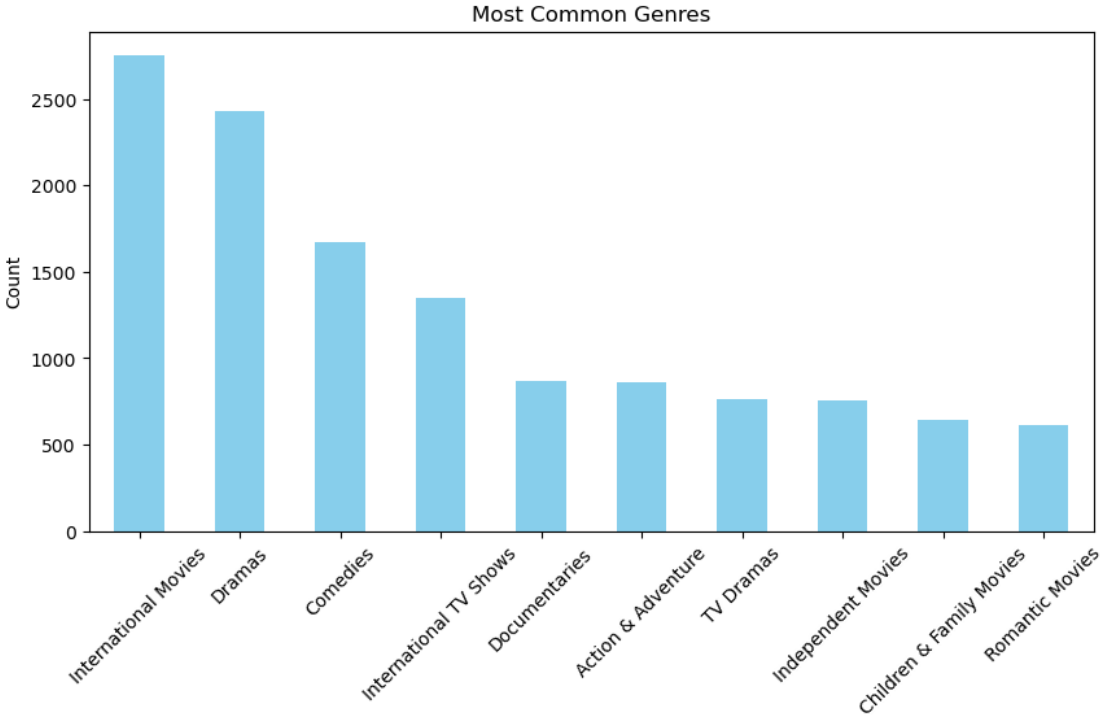
[4]:
```
# Step 1: Import Required Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
```

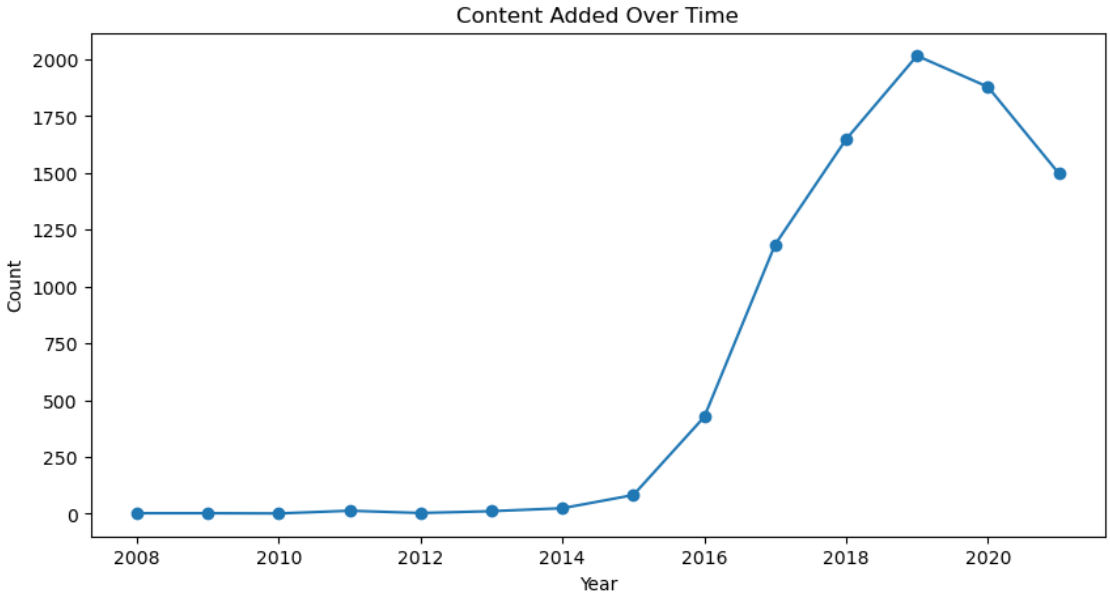
[5]:
```
# Step 2: Load the Dataset
df = pd.read_csv('netflix1.csv')  # Update with your dataset path
```

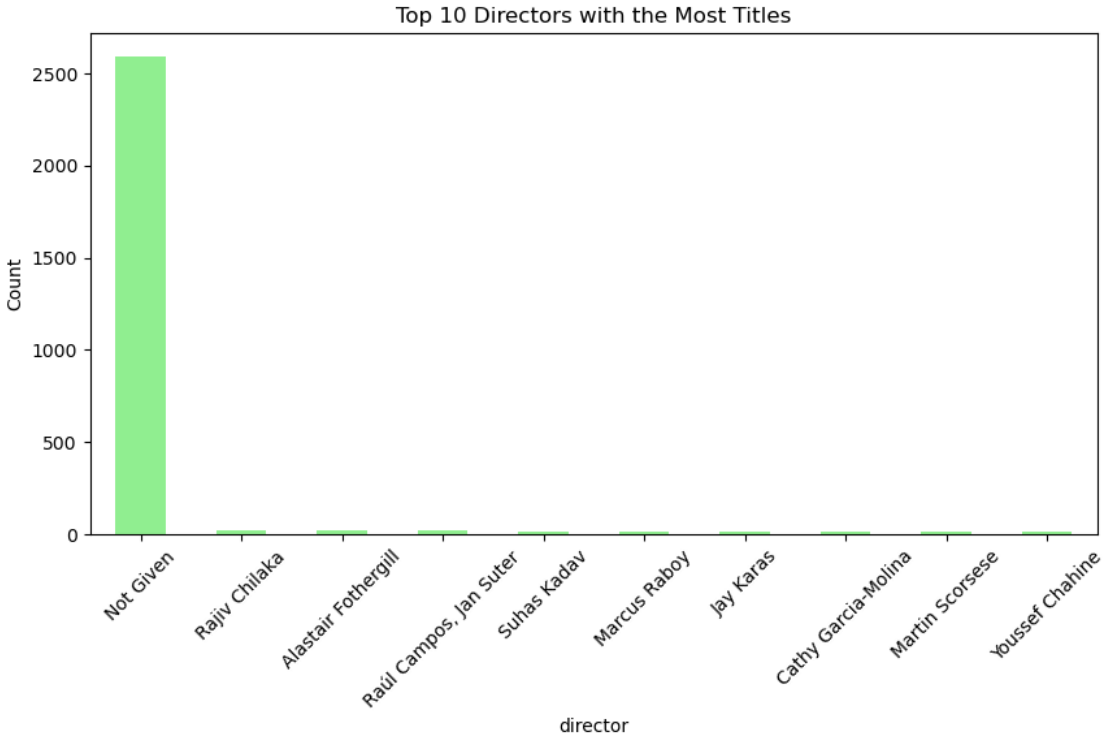
[7]:
```
# Step 3: Data Cleaning
# Handle missing values
df.dropna(subset=['title', 'type'], inplace=True)  # Dropping rows with missing␣
 ↪titles or types
df.drop_duplicates(inplace=True)  # Removing duplicates

# Convert 'date_added' to datetime
df['date_added'] = pd.to_datetime(df['date_added'])

# Handle 'duration' column
def convert_duration(duration):
    if 'min' in duration:
        return int(duration.replace(' min', ''))
    elif 'Season' in duration:
        return 0  # or you could return a different value indicating seasons
    return np.nan  # Handle any unexpected format

df['duration'] = df['duration'].apply(convert_duration)  # Convert duration to␣
 ↪int

# Step 4: Exploratory Data Analysis (EDA)
# 1. Content Type Distribution (Movies vs. TV Shows)
plt.figure(figsize=(8, 5))
sns.countplot(x='type', data=df)
plt.title('Content Type Distribution (Movies vs. TV Shows)')
plt.xlabel('Content Type')
plt.ylabel('Count')
```

```
plt.show()

# 2. Most Common Genres
# Exploding genres to get counts
genres = df['listed_in'].str.get_dummies(sep=', ')
most_common_genres = genres.sum().sort_values(ascending=False).head(10)
most_common_genres.plot(kind='bar', figsize=(10, 5), color='skyblue')
plt.title('Most Common Genres')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()

# 3. Content Added Over Time
df['year_added'] = df['date_added'].dt.year
content_added_over_time = df['year_added'].value_counts().sort_index()
content_added_over_time.plot(kind='line', figsize=(10, 5), marker='o')
plt.title('Content Added Over Time')
plt.xlabel('Year')
plt.ylabel('Count')
plt.show()

# 4. Top 10 Directors with the Most Titles
top_directors = df['director'].value_counts().head(10)
top_directors.plot(kind='bar', figsize=(10, 5), color='lightgreen')
plt.title('Top 10 Directors with the Most Titles')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()

# 5. Word Cloud of Movie Titles
wordcloud = WordCloud(width=800, height=400, background_color='white').
 ↪generate(' '.join(df['title']))
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Word Cloud of Movie Titles')
plt.show()
```


Content Type Distribution (Movies vs. TV Shows)
Count
6000
5000
4000
3000
2000
1000
0
Movie
TV Show
Content Type


Most Common Genres
Count
2500
2000
1500
1000
500
0
International Movies
Dramas
Comedies
International TV Shows
Documentaries
Action & Adventure
TV Dramas
Independent Movies
Children & Family Movies
Romantic Movies


Content Added Over Time
Count
Year
2000
1750
1500
1250
1000
750
500
250
0
2008
2010
2012
2014
2016
2018
2020


Top 10 Directors with the Most Titles
Count
director
2500
2000
1500
1000
500
0
Not Given
Rajiv Chilaka
Alastair Fothergill
Raúl Campos, Jan Suter
Suhas Kadav
Marcus Raboy
Jay Karas
Cathy Garcia-Molina
Martin Scorsese
Youssef Chahine

Word Cloud of Movie Titles


```
[ ]:
```