1 DATA VISUALIZATION BUAN-665-BB

1.0.1 MOHAMMED AFTAB HUSSAIN - 0961370

1.0.2 VISUALIZATION #5

2 Netflix Movies and TV Shows Analysis

2.1 DATASET OVERVIEW

-

2.1.1 I choose this Netflix Movies and TV Shows dataset because I'm interested in understanding trends in Netflix content over time. The dataset includes information such as title, type (movie or TV show), release year, genre, and more. and in my other course i have don a presentation on it.

2.2 KEY QUESTIONS

-

2.2.1 1. How has the number of movies and TV shows on Netflix changed over time?

-

2.2.2 2. Which genres are the most popular on Netflix?

-

2.2.3 3. Is there a relationship between release year and duration for movies?

2.2.4 I plan to create visualizations to answer these questions and to find interesting insights about Netflix's content strategy.

2.3 LIBRARIES & DATASET

[9]:

```
import pandas as pd

## For loading the dataset you have to use df = pd.read_csv() for selecting the␣
 ↪data set but make sure you have to upload your dataset in jupyter notebook␣
 ↪to read it otherwise it shows error or else you have to use path its a long␣
 ↪procees this one is easy.
df = pd.read_csv('netflix_titles.csv')

## For displaying the first few rows you have use this code which will help you␣
 ↪to check is it right dataset your working or not.
df.head()
```

[9]:

```
  show_id     type                  title          director  \
0      s1    Movie   Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show          Blood & Water              NaN
2      s3  TV Show              Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans              NaN
4      s5  TV Show           Kota Factory              NaN

                                                cast        country  \
0                                                NaN  United States
1  Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban…   South Africa
2  Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi…            NaN
3                                                NaN            NaN
4  Mayur More, Jitendra Kumar, Ranjan Raj, Alam K…          India

           date_added  release_year rating   duration  \
0  September 25, 2021          2020  PG-13     90 min
1  September 24, 2021          2021  TV-MA  2 Seasons
2  September 24, 2021          2021  TV-MA   1 Season
3  September 24, 2021          2021  TV-MA   1 Season
4  September 24, 2021          2021  TV-MA  2 Seasons

                                           listed_in  \
0                                      Documentaries
1    International TV Shows, TV Dramas, TV Mysteries
2  Crime TV Shows, International TV Shows, TV Act…
3                             Docuseries, Reality TV
4  International TV Shows, Romantic TV Shows, TV …
```

```
                                         description
0  As her father nears the end of his life, filmm…
1  After crossing paths at a party, a Cape Town t…
2  To protect his family from a powerful drug lor…
3  Feuds, flirtations and toilet talk go down amo…
4  In a city of coaching centers known to train I…
```

2.4 Creating Calculated Columns

[2]:

```
import pandas as pd
df = pd.read_csv('netflix_titles.csv')

## Adding a decade column to the table for knowing movies release year.
df['decade'] = (df['release_year'] // 10) * 10

## Adding a duration min column for knowing movies time duration.
df['duration_min'] = df.apply(
    lambda row: int(row['duration'].split()[0]) if row['type'] == 'Movie' and␣
  ↪pd.notna(row['duration']) else None,
    axis=1
)

## Here this will show the updated dataframe which we have command to insert.
print(df[['title', 'type', 'duration', 'duration_min', 'decade']].head())
```

```
                 title     type   duration  duration_min  decade
0    Dick Johnson Is Dead    Movie     90 min          90.0    2020
1           Blood & Water  TV Show  2 Seasons           NaN    2020
2              Ganglands  TV Show   1 Season           NaN    2020
3  Jailbirds New Orleans  TV Show   1 Season           NaN    2020
4           Kota Factory  TV Show  2 Seasons           NaN    2020
```

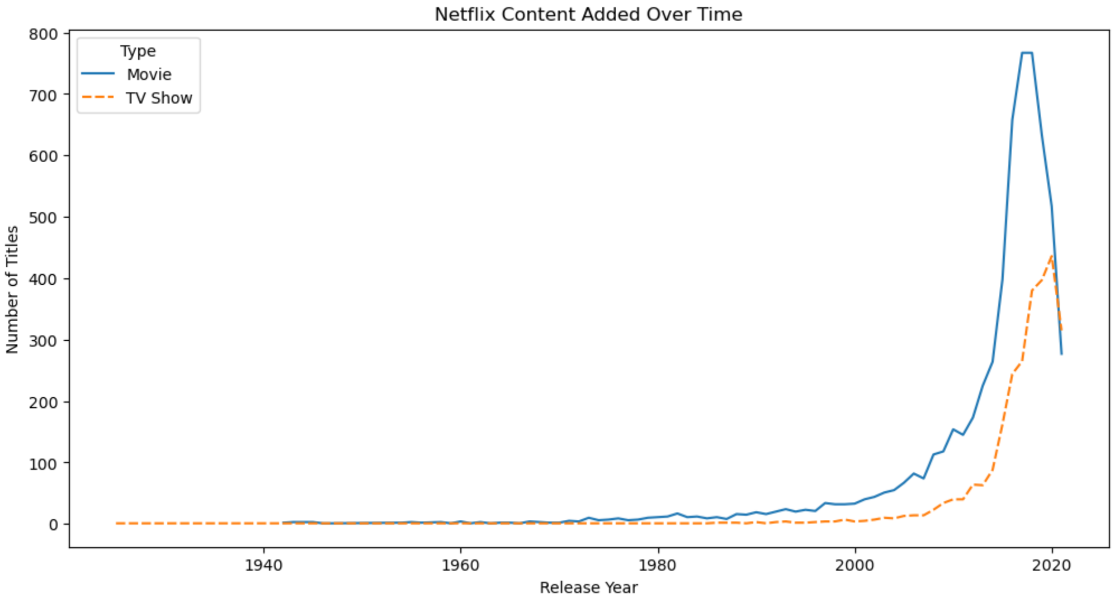
VISUALIZATION 1 - LINE CHART

[6]:

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

content_by_year = df.groupby(['release_year', 'type']).size().unstack()

plt.figure(figsize=(12, 6))
sns.lineplot(data=content_by_year)
plt.title('Netflix Content Added Over Time')
plt.xlabel('Release Year')
plt.ylabel('Number of Titles')
plt.legend(title='Type')
plt.show()
```



2.4.1 LINE CHART : NETFLIX CINTENT ADDED OVER TIME

Goal of the Visualization The goal is to show how the number of movies and TV shows on Netflix has changed over time.

Key Insight The line chart reveals that Netflix has significantly increased its content library, especially after 2015. Movies dominate the library, but TV shows are also growing rapidly.

Why a Line Chart? A line chart is ideal for showing trends over time, making it easy to compare the growth of movies and TV shows.

Improvements

- Add annotations for key events (e.g., Netflix's expansion into original content).
- Use a log scale if the growth is exponential.

2.5 Visualization 2 - Bar Chart

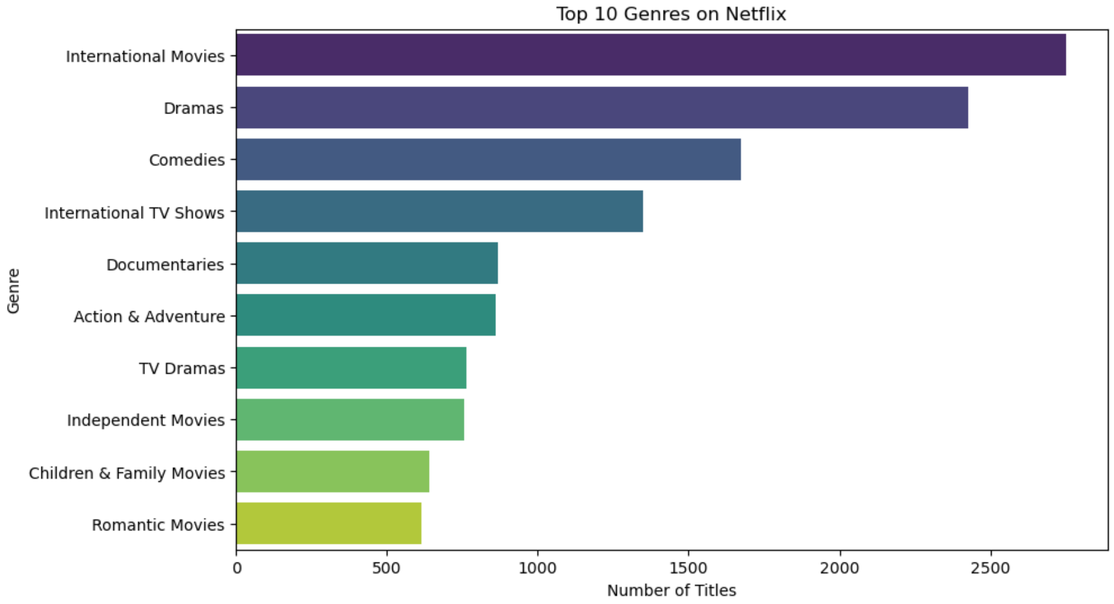
[8]:

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

df = pd.read_csv('netflix_titles.csv')

##
## Plotting the data as required.
plt.figure(figsize=(10, 6))
```

```
sns.barplot(x=top_genres.values, y=top_genres.index, hue=top_genres.index,␣
 ↪palette='viridis', legend=False)
plt.title('Top 10 Genres on Netflix')
plt.xlabel('Number of Titles')
plt.ylabel('Genre')
plt.show()
```



2.5.1 Bar Chart: Top 10 Genres on Netflix

Goal of the Visualization The goal is to identify the most popular genres on Netflix.

Key Insight The bar chart shows that **Dramas**, **Comedies**, and **Documentaries** are the most common genres on Netflix.

Why a Bar Chart? A bar chart is effective for comparing categories (genres) and their counts.

Improvements

- Group similar genres (ex: “International Movies” and “International TV Shows”).
- Add percentages to show the proportion of each genre.

2.6 Visualization 3 - Scatter Plot

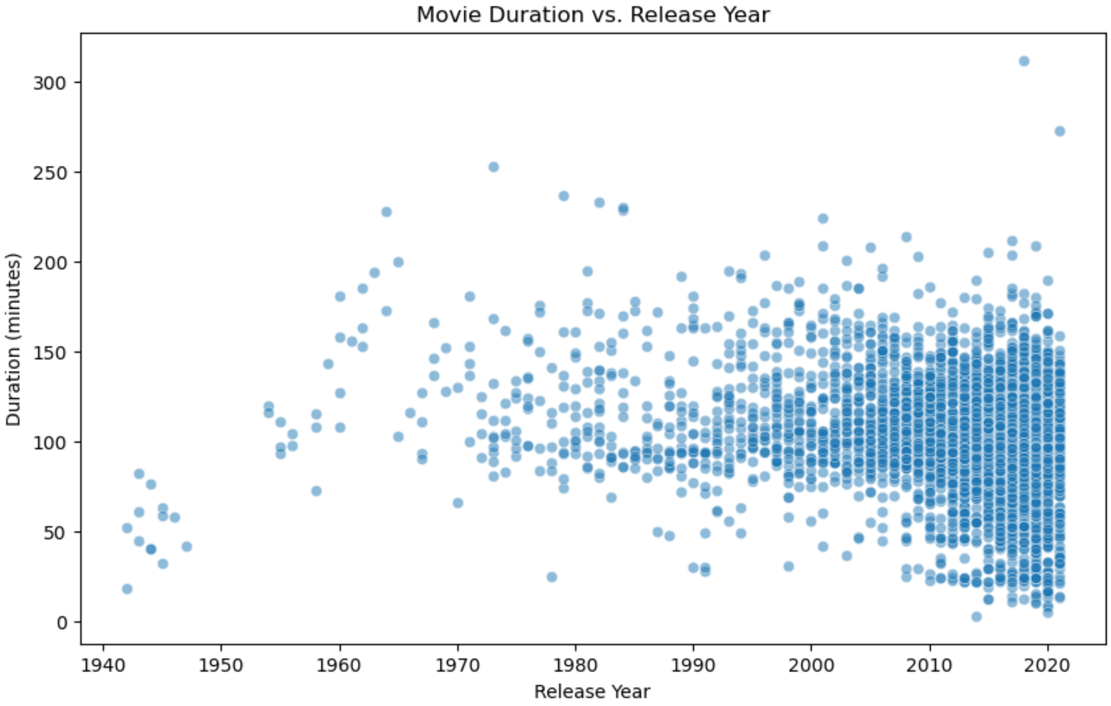
[3]:

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

## Filter for movies
```

```
movies = df[df['type'] == 'Movie']

## Plotting the data
plt.figure(figsize=(10, 6))
sns.scatterplot(x=movies['release_year'], y=movies['duration_min'], alpha=0.5)
plt.title('Movie Duration vs. Release Year')
plt.xlabel('Release Year')
plt.ylabel('Duration (minutes)')
plt.show()
```



2.7 Visualization 3 scatter plot Explanation

2.7.1 Scatter Plot: Movie Duration vs. Release Year

Goal of the Visualization The goal is to explore the relationship between movie duration and release year.

Key Insight The scatter plot shows that older movies tend to be longer, while newer movies are more consistent in duration.

Why a Scatter Plot? A scatter plot is ideal for showing relationships between two numerical variables.

Improvements

- Add a trendline to highlight the overall pattern.
- Use color to differentiate genres.

3 DONE