Netflix Data Analysis (SQL and Python)

Netflix is a popular streaming service that offers a vast catalog of movies, TV shows, and original content. The data consists of contents added to Netflix from 2008 to 2021. This dataset is a cleaned version. This dataset will be visualized using python. The purpose of this dataset is to analyze, clean and visualize the data.

Data Cleaning

While cleaning data, these are the things that are followed:

1. Treat the Nulls
2. Treat the duplicates
3. Populate missing rows
4. Drop unneeded columns
5. Split columns

Procedure

1. Here, we are importing libraries in order to connect to databases, and for visualization purposes.

```
#Netflix_Data_Analysis
```

```
import sqlite3
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

2. Here, we are reading the file into the data frame and displaying the contents in the dataframe.

```
#Read the file into the dataframe
df = pd.read_csv(r'C:\Users\14695\OneDrive\Desktop\Netflix_data_analysis(sql &python).csv', encoding='unicode_escape')
```

```
df
```

:

	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_i
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021	2020	PG-13	90 min	Documentarie
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act.
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysterie
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021	2021	TV-PG	91 min	Children & Family Movies, Comedie
4	s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021	1993	TV-MA	125 min	Dramas, Independent Movies International Movie
...	...	...	...	...	...	...	...	...	...	.
8785	s8797	TV Show	Yunus Emre	Not Given	Turkey	1/17/2017	2016	TV-PG	2 Seasons	International TV Shows, TV Drama
8786	s8798	TV Show	Zak Storm	Not Given	United States	9/13/2018	2016	TV-Y7	3 Seasons	Kids' T

3. Here, we are saving the data frame into the SQLite database and executing sql query to fetch the data.

```
# Save the DataFrame to an SQLite database
df.to_sql(name='Netflix', con=connection, index=False, if_exists='replace')
```

: 8790

```
import sqlite3
import pandas as pd

# Connect to the SQLite database
database_path = r'C:\Users\14695\OneDrive\Desktop\python_projects\travel.sqlite'
connection = sqlite3.connect(database_path)
cursor = connection.cursor()

# Execute SQL query to fetch all data from the specified table
query = "SELECT * FROM Netflix"
df_from_sql = pd.read_sql_query(query, connection)

# Display the DataFrame
df_from_sql
```

:

	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	9/25/2021	2020	PG-13	90 min	Documentaries
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries

4. Here, we are checking if the Data is cleaned or not by checking if there are any duplicates present on the show_id column which is the unique id column in the dataset.

```
#The show_id column is the unique id for the dataset, therefore we are going to check for duplicates
import sqlite3
import pandas as pd

# Connect to the SQLite database
database_path = r'C:\Users\14695\OneDrive\Desktop\python_projects\travel.sqlite'
connection = sqlite3.connect(database_path)
cursor = connection.cursor()

# Execute SQL query to fetch all data from the specified table
query = """
    SELECT show_id, COUNT(*) as Count
    FROM Netflix
    GROUP BY show_id
    ORDER BY show_id DESC
"""

# Use pd.read_sql_query to execute the query and fetch the results into a DataFrame
df_from_sql = pd.read_sql_query(query, connection)

# Display the DataFrame
df_from_sql
```

3]:

	show_id	Count
0	s999	1
1	s998	1

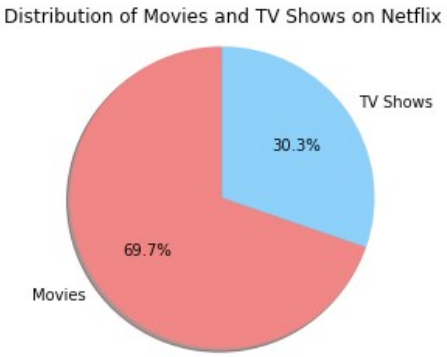
5. Here, we are checking if there are any null values present across the columns in the dataset.

```
#To Check null values across columns
import sqlite3
import pandas as pd

# Connect to the SQLite database
database_path = r'C:\Users\14695\OneDrive\Desktop\python_projects\travel.sqlite'
connection = sqlite3.connect(database_path)
cursor = connection.cursor()

# Execute SQL query to fetch counts of NULL values for each column
query = """
    SELECT
        COUNT(*) FILTER(WHERE show_id IS NULL) AS showid_nulls,
        COUNT(*) FILTER(WHERE type IS NULL) AS type_nulls,
        COUNT(*) FILTER(WHERE title IS NULL) AS title_nulls,
        COUNT(*) FILTER(WHERE director IS NULL) AS director_nulls,
        COUNT(*) FILTER(WHERE country IS NULL) AS country_nulls,
        COUNT(*) FILTER(WHERE date_added IS NULL) AS date_addes_nulls,
        COUNT(*) FILTER(WHERE release_year IS NULL) AS release_year_nulls,
        COUNT(*) FILTER(WHERE rating IS NULL) AS rating_nulls,
        COUNT(*) FILTER(WHERE duration IS NULL) AS duration_nulls,
        COUNT(*) FILTER(WHERE listed_in IS NULL) AS listed_in_nulls
    FROM Netflix
"""

# Use pd.read_sql_query to execute the query and fetch the results into a DataFrame
df_from_sql = pd.read_sql_query(query, connection)

# Display the DataFrame
df_from_sql
```

	showid_nulls	type_nulls	title_nulls	director_nulls	country_nulls	date_addes_nulls	release_year_nulls	rating_nulls	duration_nulls	listed_in_nulls
0	0	0	0	0	0	0	0	0	0	0

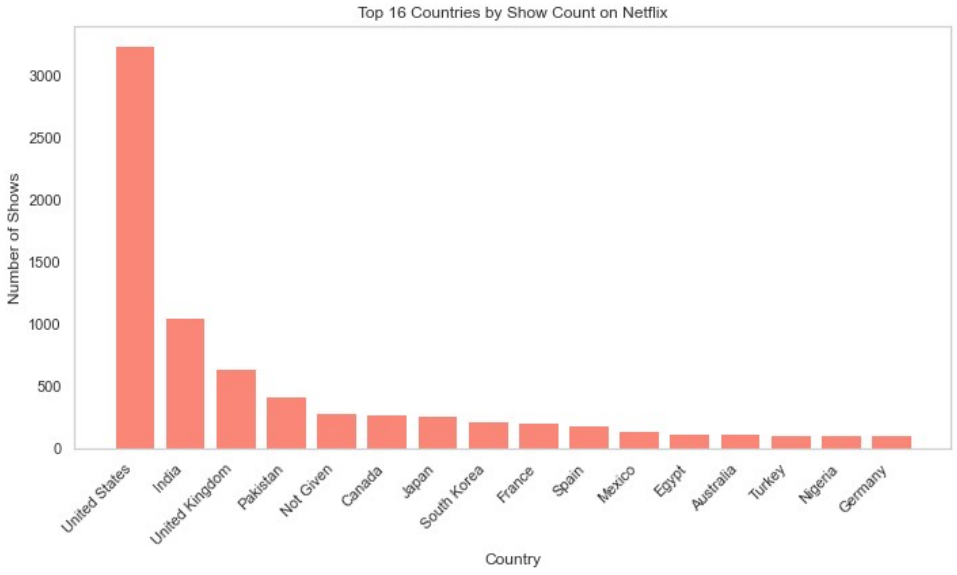
Data Visualization: -

Fig 1: Shows the distribution of Movies and Tv shows on Netflix.



From the above pie chart, it is understood that Movies cover the highest percentage on Netflix i.e 69.7%.

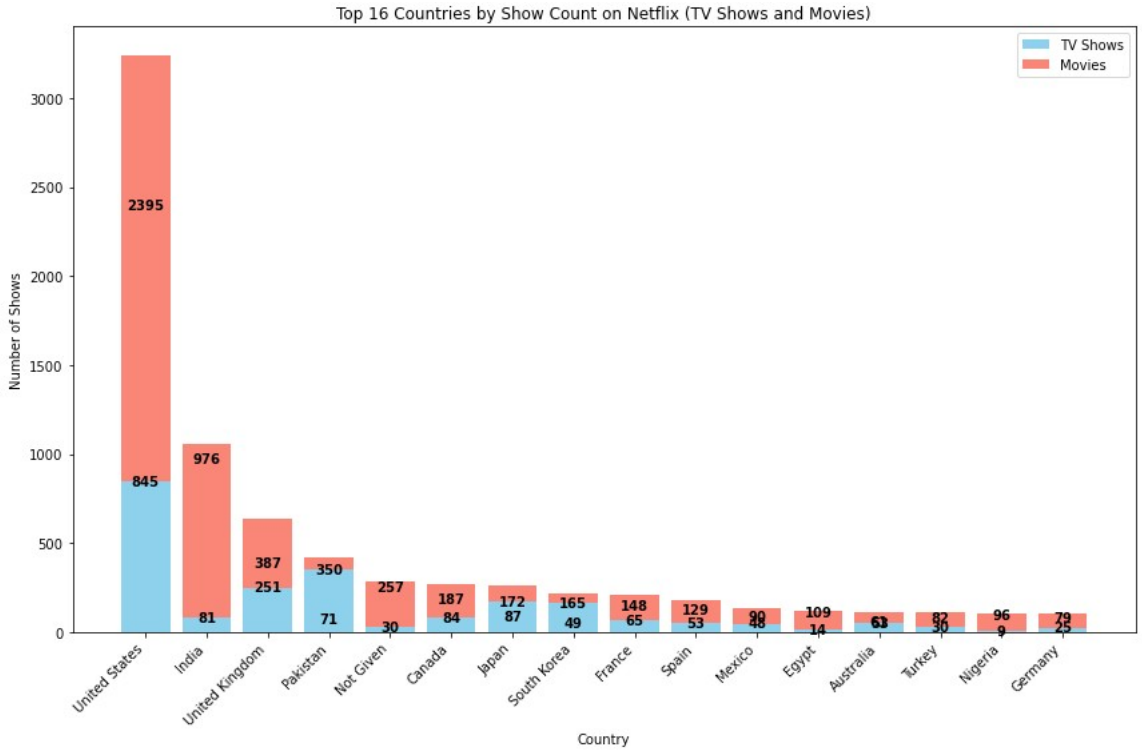
Fig 2: Shows the Top countries by show count on Netflix.



Observation:

1)From the above figures it is clear that United States has highest number of shows on Netflix.

Fig 3: Shows the top countries by show count (Movies and Tv) on Netflix.

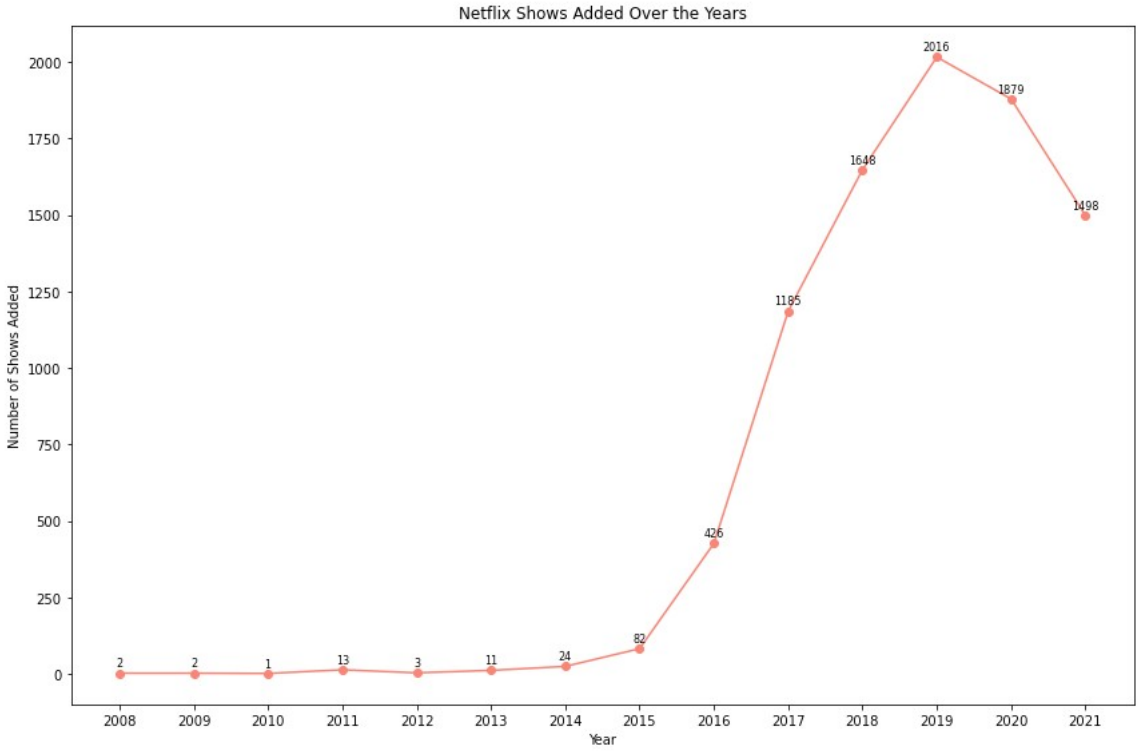


Observation:

1)From the above figures it is clear that United States has highest number of shows on Netflix.

2)The count of Movies is higher than that of Tv shows in majority of countries.

Fig 4: Shows the Netflix shows (both Tv shows and Movies) added over the years.

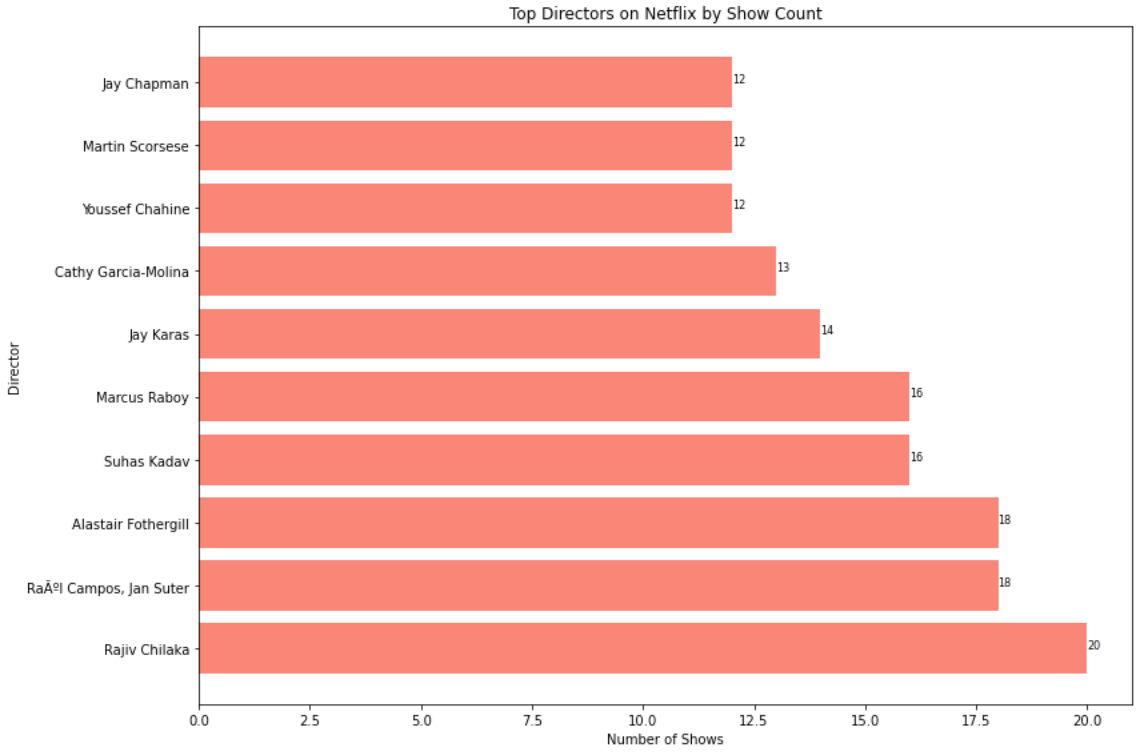


Observation:

1)This time series chart shows the total number of contents added to Netflix all through the given years (2008 - 2021).

2)It shows that most movies and tv shows on Netflix were added in 2019.

Fig 5: Shows the Top Directors on Netflix



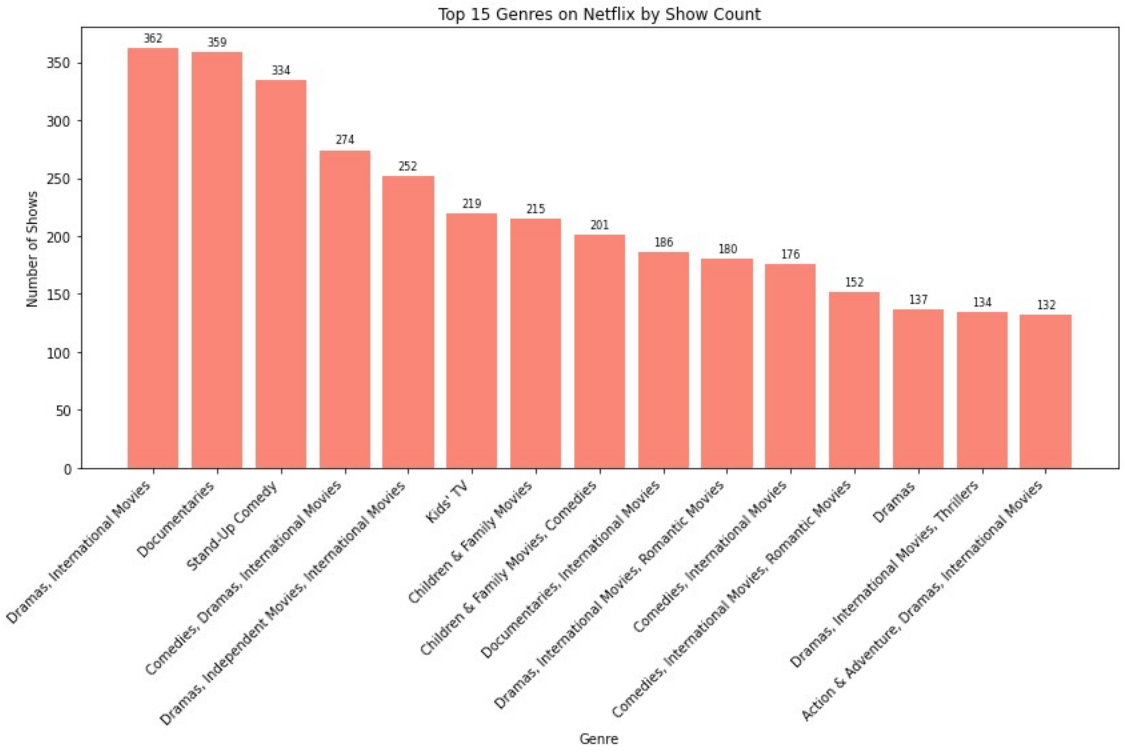
Observation:

1)The chart shows the top 10 directors with the most contents on Netflix.

2)Rajiv Chilaka is the top director with most shows on Netflix.

3)We can also note that the duo of Raul Campos and Jan Suter are fond of working together and have directed 18 movies on Netflix.

Fig 6: Shows the Top 15 Genres on Netflix

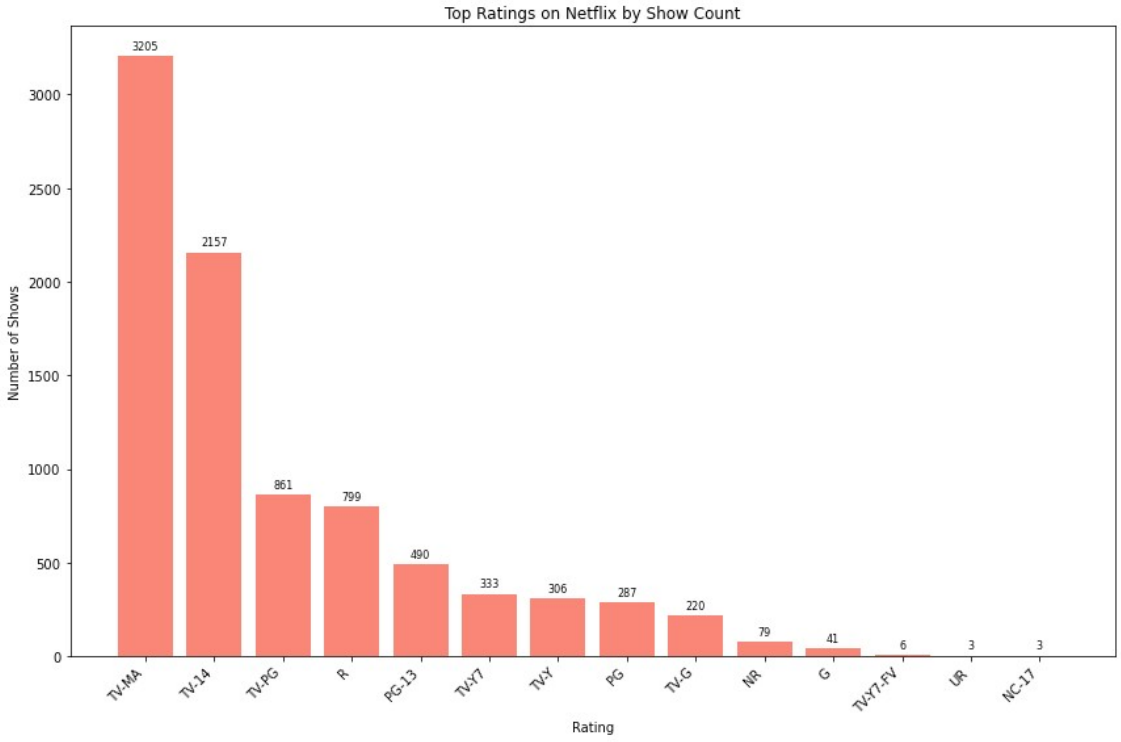


Observation:

1)This chart shows the genres with the highest numbers on Netflix.

2)We can see that Drama & International movies followed by Documentary have the highest number of contents on Netflix within the period.

Fig 7: Shows the Top Ratings on Netflix



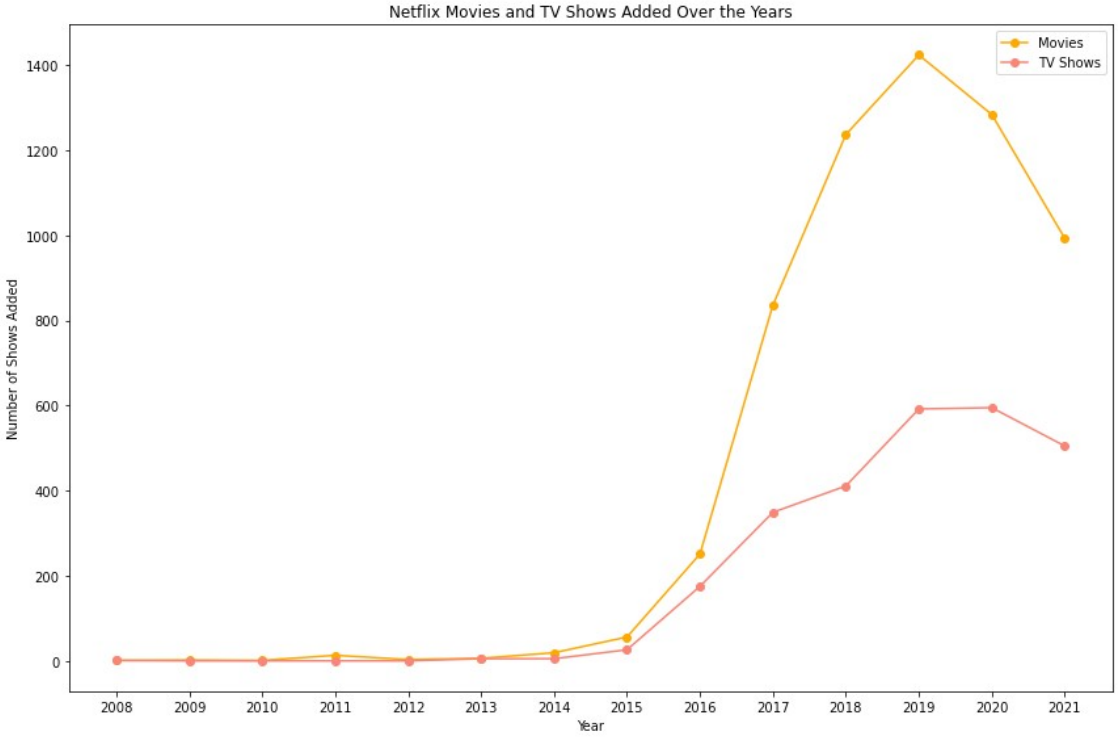
Observation:

The chart shows the top ratings on Netflix.

We can note that most content on Netflix is rated TV-MA.

TV-MA in the United States by the TV Parental Guidelines signifies content for mature audiences.

Fig 8: Shows the Movies and TV shows added over years on Netflix.



Observation:

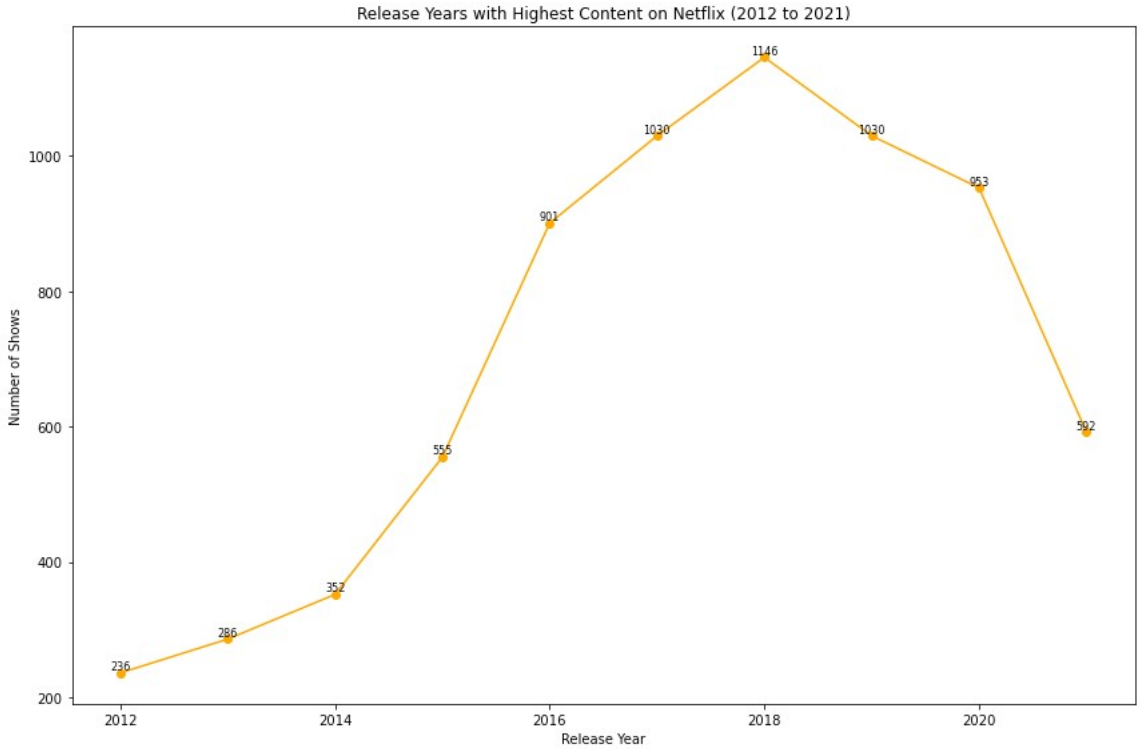
1)This line chart compares the Movie and Tv shows contents added to Netflix all through the years.

2)We can see that more movies have always been added and more movies and tv shows on Netflix were added in 2019.

3)In 2013, the number of contents added to Netflix for both were almost the same with Movies having 6 contents that year and Tv shows having 5.

4)It shows that in the first 5 years, only movies were added to Netflix.

Fig 9: Shows the release years with highest content on Netflix.



Observations:

1)This chart shows the Movies and Tv shows production year which has highest contents on Netflix.

2)We focus on the top 10 release year/production year.

3)We can see that from 2012 to 2018, Netflix added most recent contents, they made sure most recent contents per release year are higher than the older release year contents.

4)Then in 2019, it started dropping, this may be due to the Covid-19, but further analysis may be needed to determine this.

Recommendations:

Based on the observations, here are some recommendations that Netflix might consider improving content or quality:

1. While Drama, International movies, and Documentaries are popular, there may be an opportunity to diversify content across various genres to cater to a broader audience.
2. As TV-MA content is popular, Netflix may continue to invest in mature audience content. However, it's essential to ensure a balance and offer content across different age groups.
3. Continue collaborations with successful directors like Rajiv Chilaka and explore more partnerships with directors who have a proven track record of creating popular content.
4. While adding content from different production years, maintaining a focus on recent releases can keep the platform fresh.
5. It is good to understand the drop in the number of contents added in 2019. It might be associated with external factors such as market trends, competition, or the impact of COVID-19. Understanding these factors can help in making informed decisions.
6. Since movies dominate the platform, there could be an opportunity to enhance the TV shows catalog. Analyze viewer preferences and invest in high-quality TV show productions.
7. Encourage and listen to user feedback. Understand user preferences, conduct surveys, and use customer insights to shape the content strategy.

Thanks for following through. Cheers!