In [1]:
```
#Netflix data analysis by Josh Wong
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#The following data is the raw data that unmodified, the first step I will clean the da
data = pd.read_csv('./NetflixData.csv')
data
```

Out[1]:

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
0	215309	Ace Ventura: Pet Detective	Comedy	Comedy, Mystery, US	1994.0	A	Hindi, English [Original]	70184054, 60001650, 70112729, 70027007, 115246...
1	215318	Ace Ventura: When Nature Calls	Comedy	Comedy, Action & Adventure, US	1995.0	U/A 16+	Hindi, English [Original]	70184054, 60001650, 70112729, 70027007, 115246...
2	217258	The Addams Family	Comedy	Comedy, US	1991.0	U/A 13+	English [Original], Hindi, English - Audio Des...	81156676, 81231974, 70027007, 80049939, 702179...
3	217303	Addams Family Values	Comedy	Comedy, US	1993.0	U/A 13+	English [Original], Hindi, English - Audio Des...	81156676, 70044593, 81231974, 70027007, 800500...
4	235527	Agneepath	Drama	Hindi-Language, Bollywood, Crime, Drama	1990.0	U/A 16+	Hindi [Original]	17517355, 80158546, 80158395, 80074065, 702042...
...	...	...	...	...	...	...	...	...
6398	81988312	Laila Majnu	Romance	Hindi-Language, Bollywood, Drama, Romantic, Ba...	2018.0	U/A 13+	Hindi [Original]	80065328, 81994054, 80087743, 81423081, 819940...
6399	81988313	Veere Di Wedding	Comedy	Hindi-Language, Bollywood, Comedy	2018.0	U/A 16+	Hindi [Original]	70181653, 80065328, 81672746, 80032081, 703034...

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
6400	81992621	Paw Patrol Holiday Fireplace	Kids	Kids Music, Special Interest	2024.0	U	No Dialogue [Original]	81294811, 81154166, 81272431, 81640914, 815003...
6401	81994051	Notebook	Drama	Hindi-Language, Bollywood, Drama, Romantic, So...	2019.0	U/A 7+	Hindi [Original]	80080110, 80065328, 81994054, 80087743, 819883...
6402	81994054	Loveyatri	Romance	Hindi-Language, Bollywood, Drama, Romantic	2018.0	U/A 7+	Hindi [Original]	80080110, 80065328, 81988312, 70303428, 800877...

6403 rows × 8 columns

In [2]:
```
#Data Cleaning 1
#Remove the Maturing Rating with the U/A at the beginning, with only numbers, A or U ra
data['Maturity Rating'] = data['Maturity Rating'].str.replace('U/A ', '', regex=False)
data
```

Out[2]:

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
0	215309	Ace Ventura: Pet Detective	Comedy	Comedy, Mystery, US	1994.0	A	Hindi, English [Original]	70184054, 60001650, 70112729, 70027007, 115246...
1	215318	Ace Ventura: When Nature Calls	Comedy	Comedy, Action & Adventure, US	1995.0	16+	Hindi, English [Original]	70184054, 60001650, 70112729, 70027007, 115246...
2	217258	The Addams Family	Comedy	Comedy, US	1991.0	13+	English [Original], Hindi, English - Audio Des...	81156676, 81231974, 70027007, 80049939, 702179...
3	217303	Addams Family Values	Comedy	Comedy, US	1993.0	13+	English [Original], Hindi, English - Audio Des...	81156676, 70044593, 81231974, 70027007, 800500...
4	235527	Agneepath	Drama	Hindi-Language, Bollywood,	1990.0	16+	Hindi [Original]	17517355, 80158546, 80158395, 80074065, 702042...

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
				Crime, Drama				
...	...	...	...	...	...	...	...	...
6398	81988312	Laila Majnu	Romance	Hindi-Language, Bollywood, Drama, Romantic, Ba...	2018.0	13+	Hindi [Original]	80065328, 81994054, 80087743, 81423081, 819940...
6399	81988313	Veere Di Wedding	Comedy	Hindi-Language, Bollywood, Comedy	2018.0	16+	Hindi [Original]	70181653, 80065328, 81672746, 80032081, 703034...
6400	81992621	Paw Patrol Holiday Fireplace	Kids	Kids Music, Special Interest	2024.0	U	No Dialogue [Original]	81294811, 81154166, 81272431, 81640914, 815003...
6401	81994051	Notebook	Drama	Hindi-Language, Bollywood, Drama, Romantic, So...	2019.0	7+	Hindi [Original]	80080110, 80065328, 81994054, 80087743, 819883...
6402	81994054	Loveyatri	Romance	Hindi-Language, Bollywood, Drama, Romantic	2018.0	7+	Hindi [Original]	80080110, 80065328, 81988312, 70303428, 800877...

6403 rows × 8 columns

In [3]:
```
#Data cleaning 2
#Remove the decimal from the Release Year

#I found that only convert the data type to Int64 could fix the issue
data['Release Year'] = data['Release Year'].astype('Int64')
data
```

Out[3]:

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
0	215309	Ace Ventura: Pet Detective	Comedy	Comedy, Mystery, US	1994	A	Hindi, English [Original]	70184054, 60001650, 70112729, 70027007, 115246...
1	215318	Ace Ventura: When Nature Calls	Comedy	Comedy, Action & Adventure, US	1995	16+	Hindi, English [Original]	70184054, 60001650, 70112729, 70027007, 115246...

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
2	217258	The Addams Family	Comedy	Comedy, US	1991	13+	English [Original], Hindi, English - Audio Des...	81156676, 81231974, 70027007, 80049939, 702179...
3	217303	Addams Family Values	Comedy	Comedy, US	1993	13+	English [Original], Hindi, English - Audio Des...	81156676, 70044593, 81231974, 70027007, 800500...
4	235527	Agneepath	Drama	Hindi-Language, Bollywood, Crime, Drama	1990	16+	Hindi [Original]	17517355, 80158546, 80158395, 80074065, 702042...
...	...	...	...	...	...	...	...	...
6398	81988312	Laila Majnu	Romance	Hindi-Language, Bollywood, Drama, Romantic, Ba...	2018	13+	Hindi [Original]	80065328, 81994054, 80087743, 81423081, 819940...
6399	81988313	Veere Di Wedding	Comedy	Hindi-Language, Bollywood, Comedy	2018	16+	Hindi [Original]	70181653, 80065328, 81672746, 80032081, 703034...
6400	81992621	Paw Patrol Holiday Fireplace	Kids	Kids Music, Special Interest	2024	U	No Dialogue [Original]	81294811, 81154166, 81272431, 81640914, 815003...
6401	81994051	Notebook	Drama	Hindi-Language, Bollywood, Drama, Romantic, So...	2019	7+	Hindi [Original]	80080110, 80065328, 81994054, 80087743, 819883...
6402	81994054	Loveyatri	Romance	Hindi-Language, Bollywood, Drama, Romantic	2018	7+	Hindi [Original]	80080110, 80065328, 81988312, 70303428, 800877...

6403 rows × 8 columns

In [4]:

```
#Data Cleaning 3
#Remove the duplicated value in Sub Genres

def remove_duplicate_genre(row):
```

```
    main_genre = row['Main Genre']
    sub_genres = row['Sub Genres'].split(',') #Split the value by the ','

    # Remove the sub genres value if it exists in main genres
    sub_genres = [genre for genre in sub_genres if genre.strip().lower() != main_genre.

    return ','.join(sub_genres)

# Apply the function
data['Sub Genres'] = data.apply(remove_duplicate_genre, axis=1)

data
```

Out[4]:

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
0	215309	Ace Ventura: Pet Detective	Comedy	Mystery, US	1994	A	Hindi, English [Original]	70184054, 60001650, 70112729, 70027007, 115246...
1	215318	Ace Ventura: When Nature Calls	Comedy	Action & Adventure, US	1995	16+	Hindi, English [Original]	70184054, 60001650, 70112729, 70027007, 115246...
2	217258	The Addams Family	Comedy	US	1991	13+	English [Original], Hindi, English - Audio Des...	81156676, 81231974, 70027007, 80049939, 702179...
3	217303	Addams Family Values	Comedy	US	1993	13+	English [Original], Hindi, English - Audio Des...	81156676, 70044593, 81231974, 70027007, 800500...
4	235527	Agneepath	Drama	Hindi-Language, Bollywood, Crime	1990	16+	Hindi [Original]	17517355, 80158546, 80158395, 80074065, 702042...
...	...	...	...	...	...	...	...	...
6398	81988312	Laila Majnu	Romance	Hindi-Language, Bollywood, Drama, Romantic, Ba...	2018	13+	Hindi [Original]	80065328, 81994054, 80087743, 81423081, 819940...
6399	81988313	Veere Di Wedding	Comedy	Hindi-Language, Bollywood	2018	16+	Hindi [Original]	70181653, 80065328, 81672746, 80032081, 703034...

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
6400	81992621	Paw Patrol Holiday Fireplace	Kids	Kids Music, Special Interest	2024	U	No Dialogue [Original]	81294811, 81154166, 81272431, 81640914, 815003...
6401	81994051	Notebook	Drama	Hindi-Language, Bollywood, Romantic, Social Is...	2019	7+	Hindi [Original]	80080110, 80065328, 81994054, 80087743, 819883...
6402	81994054	Loveyatri	Romance	Hindi-Language, Bollywood, Drama, Romantic	2018	7+	Hindi [Original]	80080110, 80065328, 81988312, 70303428, 800877...

6403 rows × 8 columns

In [5]:
```
#Find out all null values in the data frame
missing_values = data.isna().sum()
print(missing_values)

missing_data_df = data[data.isnull().any(axis=1)]
missing_data_df
```

```
N_id                0
Title               0
Main Genre          0
Sub Genres          0
Release Year        1
Maturity Rating     0
Original Audio     2636
Recommendations      11
dtype: int64
```

Out[5]:

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
256	70136117	House, M.D.	Drama	Medical TV Shows, TV Dramas, US TV Shows	2004	A	NaN	70195800, 70281312, 81667161, 70143836, 802411...
257	70136120	The Office (U.S.)	Comedy	Sitcoms, TV Comedies, US TV Shows	2005	13+	NaN	70153373, 81468289, 81021929, 70143830, 600333...

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
258	70136126	Dexter	Thriller	TV Dramas, TV Shows Based on Books, TV Mysteri...	2006	A	NaN	80021955, 81287562, 80201500, 70143836, 815549...
264	70140373	CSI: Crime Scene Investigation	Drama	TV Dramas, TV Mysteries, Crime TV Shows, US TV...	2000	A	NaN	70281312, 80241181, 81164276, 70142386, 701573...
265	70140375	Deadliest Catch	Reality TV	Science & Nature TV, TV Action & Adventure, US...	2005	16+	NaN	81666400, 81196690, 81518623, 81780339, 812753...
...	...	...	...	...	...	...	...	...
6388	81954670	The Story of Pearl Girl	Drama	TV Dramas, Period Pieces, Chinese TV Shows	2024	16+	NaN	81718224, 81605075, 80987113, 81633653, 810195...
6389	81954820	Kill Me Love Me	Drama	Romantic TV Dramas, TV Dramas, TV Shows Based ...	2024	16+	NaN	81605075, 81622849, 81019520, 81954670, 816897...
6391	81967459	BORDERLESS Ae! group's Debut Tour	Documentary	Japanese, Docuseries	2024	U	NaN	81705443, 81901457, 81587828, 81219073, 817768...
6392	81970550	When the Stars Gossip	Romance	Romantic TV Dramas, TV Dramas, Korean, Sci-Fi TV	2025	16+	NaN	81012551, 81736915, 80123798, 81942170, 819129...
6394	81971071	Black Warrant	Drama	TV Dramas, TV Shows Based on Books, Crime TV S...	2025	16+	NaN	81154455, 81555298, 81732726, 80065328, 811834...

2644 rows × 8 columns

In [6]:

```
#Create a new dataframe for Main Genre

# Count number of each Main Genre
genre_counts = data['Main Genre'].value_counts()
print(genre_counts)



# Extract all genres beyond the top 10
genres_after_top_10 = genre_counts.iloc[10:]

print(genres_after_top_10)

#Calculate the total number of after top 10
others_total = genre_counts.iloc[10:].sum()

#print(others_total)

genre_df = genre_counts.reset_index()
genre_df.columns = ['Main Genre', 'Count']

# Extract the top 10
top_10_df = genre_df.head(10)

print('Top 10:',top_10_df)

# Add a new row of Others into the dataframe
new_row = pd.DataFrame({'Main Genre': ['Others'], 'Count': others_total})
genre_df = pd.concat([top_10_df, new_row], ignore_index=True)

#Convert the dataframe to pandas format
genre_df = pd.DataFrame(genre_df)
```

```
Drama               1639
Comedy              1259
Documentary          730
Kids                 566
Action               506
Romance              369
Thriller             346
Anime                267
Reality TV           259
Horror               225
Sci-Fi                97
Fantasy               66
Music                 27
Talk Show             14
Sports                11
Variety TV             8
Special Interest       5
Musical                4
Western                4
Friendship             1
Name: Main Genre, dtype: int64
Sci-Fi              97
Fantasy             66
```

```
Music               27
Talk Show           14
Sports              11
Variety TV           8
Special Interest     5
Musical              4
Western              4
Friendship           1
Name: Main Genre, dtype: int64
Top 10:     Main Genre  Count
0         Drama   1639
1        Comedy   1259
2   Documentary    730
3          Kids    566
4        Action    506
5       Romance    369
6      Thriller    346
7         Anime    267
8    Reality TV    259
9        Horror    225
```

In [7]:

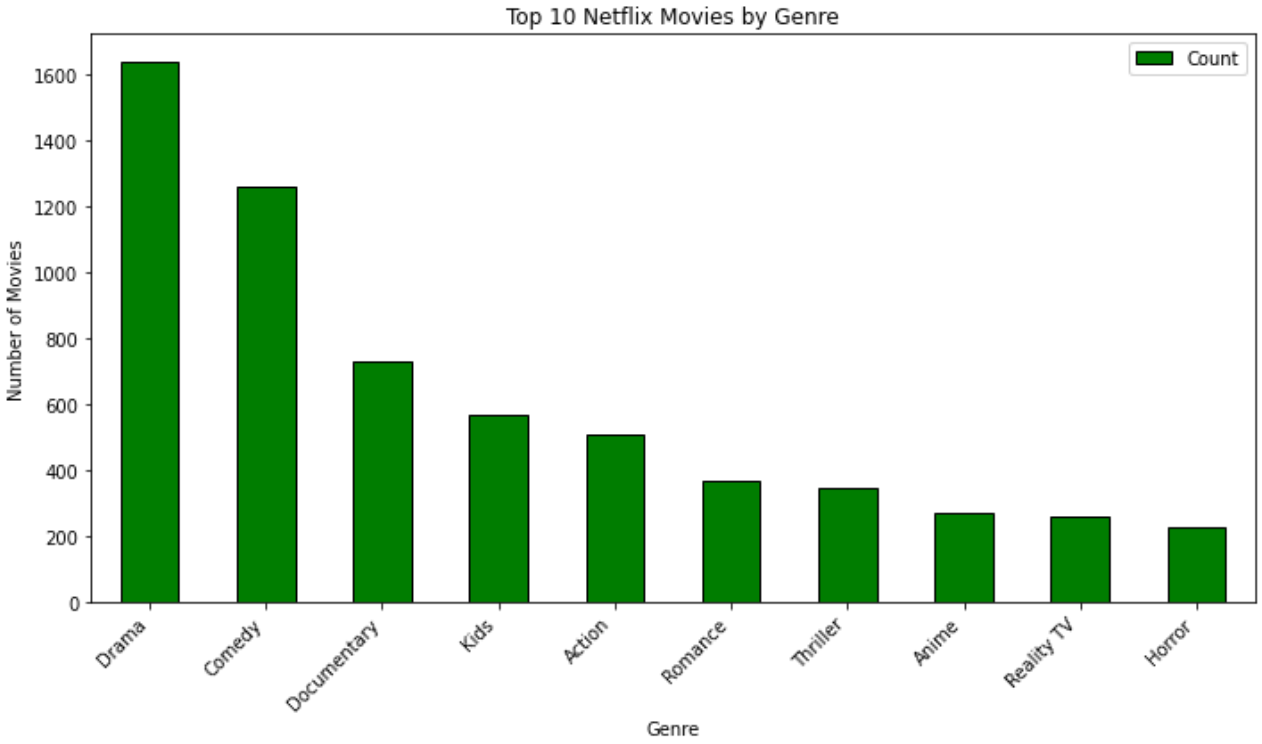
```python
#Plot the top_10 in dataframe

top_10_df.plot.bar(
    x='Main Genre',      # X-axis: genres
    y='Count',           # Y-axis: counts
    figsize=(10, 6),     # Size of the chart (width, height)
    color='green',     # Bar color
    edgecolor='black'    # Border color for bars
)

# Add labels and title
plt.xlabel('Genre')
plt.ylabel('Number of Movies')
plt.title('Top 10 Netflix Movies by Genre')

# Rotate x-axis labels for better readability
plt.xticks(rotation=45, ha='right')

# Display the plot
plt.tight_layout()  # Adjust layout to prevent label cutoff
plt.show()
```



In [8]:
```
#The new Main Genre dataframe
genre_df
```

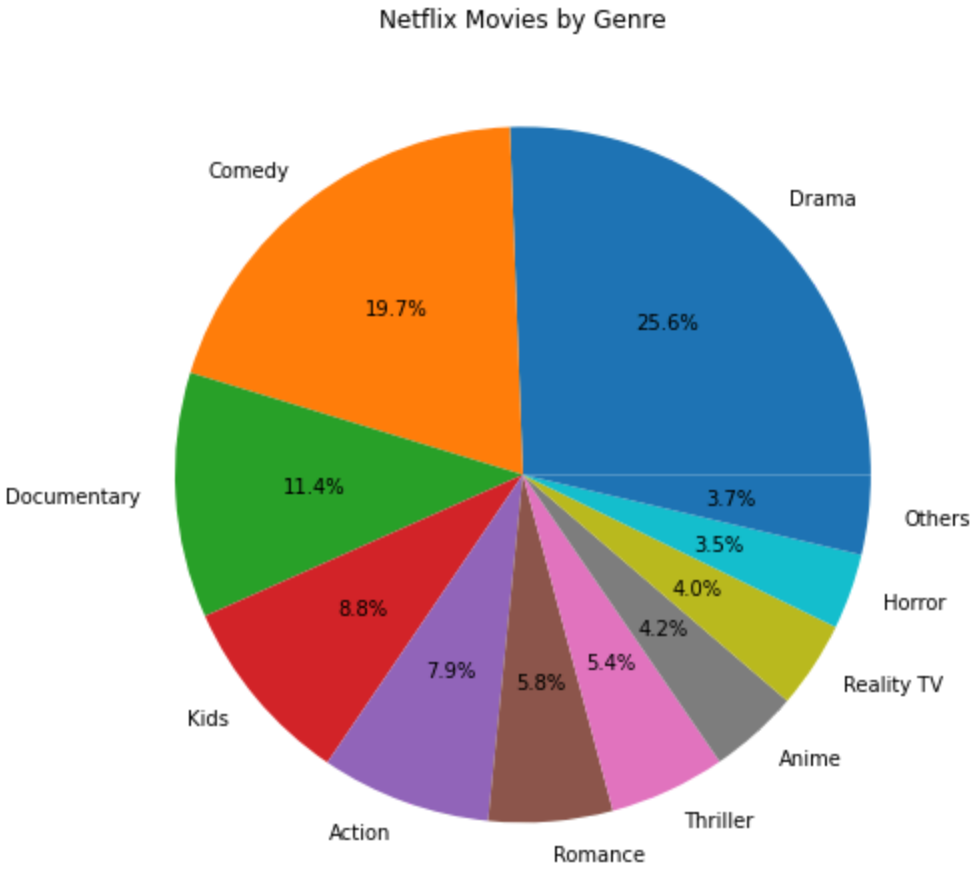
Out[8]:

	Main Genre	Count
0	Drama	1639
1	Comedy	1259
2	Documentary	730
3	Kids	566
4	Action	506
5	Romance	369
6	Thriller	346
7	Anime	267
8	Reality TV	259
9	Horror	225
10	Others	237

In [9]:
```
#Create a pie chart to show netflix movies Main Genre
plt.figure(figsize=(8, 8))
plt.pie(genre_df['Count'], labels=genre_df['Main Genre'], autopct='%1.1f%%')

plt.title('Netflix Movies by Genre')

# Display the plot
plt.show()
```



In [10]:
```
#Show all unique Release Year in the data base in ascending order
unique_years = sorted(data['Release Year'].dropna().unique())
print(unique_years)
```

```
[1962, 1966, 1969, 1971, 1972, 1973, 1974, 1975, 1977, 1978, 1979, 1980, 1981, 1982, 198
3, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1
998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012,
2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025]
```

In [11]:
```
#Find out the row which the Release Year value is NA
missing_years_df = data[data['Release Year'].isna()]
missing_years_df
```

Out[11]:

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
4689	81505899	Lady Tamara	Reality TV	Food & Travel TV, Spanish, Lifestyle	<NA>	13+	NaN	81720895, 70153388, 81512574, 81462121, 813125...

In [12]:
```
# Create new DataFrame by cleaning all rows of the Release Year where value is null
#The original dataframe has 6403 rows now after dropped the row value contains NA and
clean_year_df = data[data['Release Year'].notna()]
clean_year_df
```

Out[12]:

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
0	215309	Ace Ventura: Pet Detective	Comedy	Mystery, US	1994	A	Hindi, English [Original]	70184054, 60001650, 70112729, 70027007, 115246...
1	215318	Ace Ventura: When Nature Calls	Comedy	Action & Adventure, US	1995	16+	Hindi, English [Original]	70184054, 60001650, 70112729, 70027007, 115246...
2	217258	The Addams Family	Comedy	US	1991	13+	English [Original], Hindi, English - Audio Des...	81156676, 81231974, 70027007, 80049939, 702179...
3	217303	Addams Family Values	Comedy	US	1993	13+	English [Original], Hindi, English - Audio Des...	81156676, 70044593, 81231974, 70027007, 800500...
4	235527	Agneepath	Drama	Hindi-Language, Bollywood, Crime	1990	16+	Hindi [Original]	17517355, 80158546, 80158395, 80074065, 702042...
...	...	...	...	...	...	...	...	...
6398	81988312	Laila Majnu	Romance	Hindi-Language, Bollywood, Drama, Romantic, Ba...	2018	13+	Hindi [Original]	80065328, 81994054, 80087743, 81423081, 819940...
6399	81988313	Veere Di Wedding	Comedy	Hindi-Language, Bollywood	2018	16+	Hindi [Original]	70181653, 80065328, 81672746, 80032081, 703034...
6400	81992621	Paw Patrol Holiday Fireplace	Kids	Kids Music, Special Interest	2024	U	No Dialogue [Original]	81294811, 81154166, 81272431, 81640914, 815003...
6401	81994051	Notebook	Drama	Hindi-Language, Bollywood, Romantic, Social Is...	2019	7+	Hindi [Original]	80080110, 80065328, 81994054, 80087743, 819883...
6402	81994054	Loveyatri	Romance	Hindi-Language, Bollywood, Drama, Romantic	2018	7+	Hindi [Original]	80080110, 80065328, 81988312, 70303428, 800877...

6402 rows × 8 columns

In [13]:

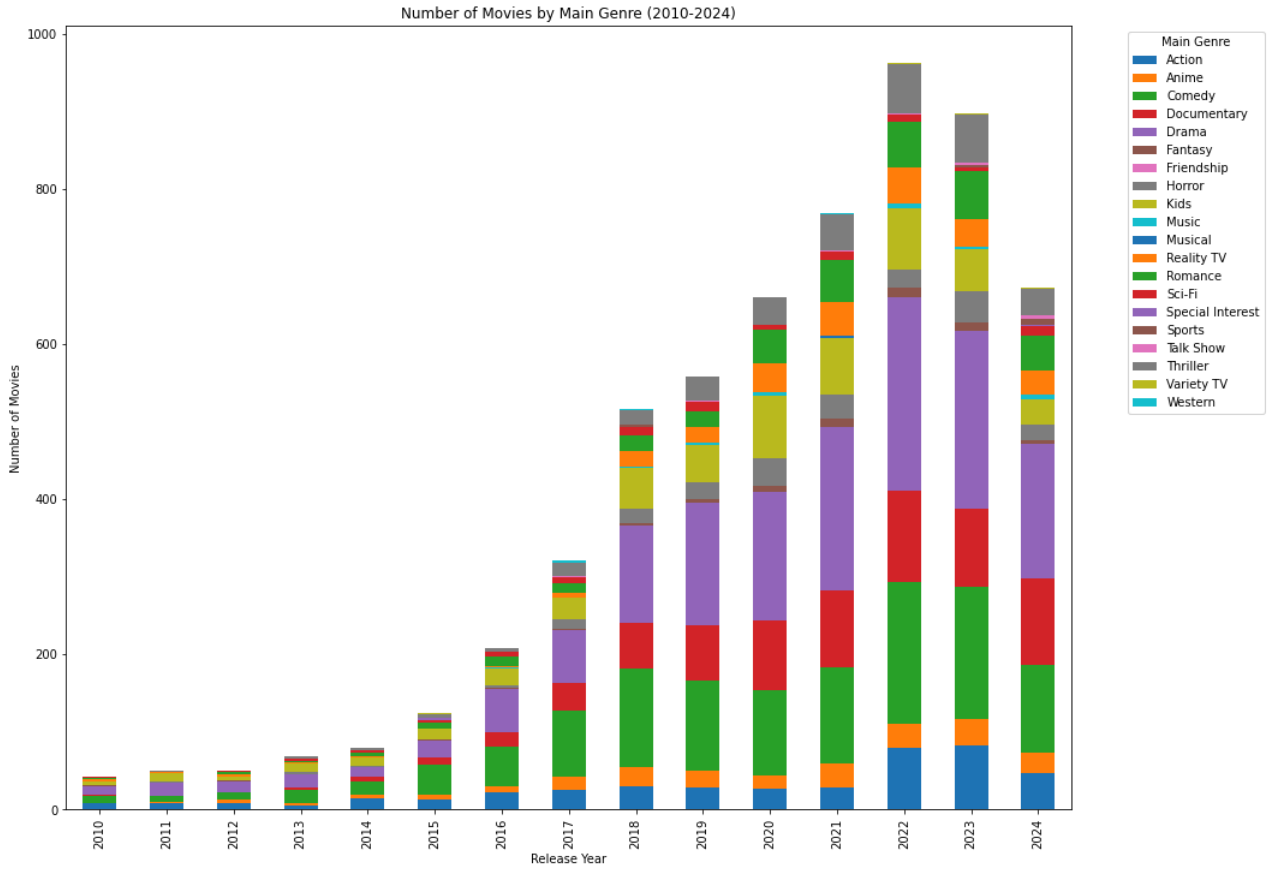
```
#Filter all movies where the release year between 2010 to 2024
df_filtered = clean_year_df[(clean_year_df['Release Year'] >= 2010) & (clean_year_df['R

# Count the number of movies per genre for each year
df_pivot = df_filtered.pivot_table(index='Release Year', columns='Main Genre', aggfunc=

# Plot Stacked Bar Chart
df_pivot.plot(kind='bar', stacked=True, figsize=(15, 12))

# Customize the plot
plt.title("Number of Movies by Main Genre (2010-2024)")
plt.xlabel("Release Year")
plt.ylabel("Number of Movies")
plt.legend(title="Main Genre", bbox_to_anchor=(1.05, 1), loc='upper left')

# Show the plot
plt.show()
```



In [14]:

```
# Filter for Anime and non-null Release Year
anime_df = data[(data['Main Genre'] == 'Anime') & (data['Release Year'].notna())]
anime_df
```

Out[14]:

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
11	557010	Grave of the Fireflies	Anime	Japanese, Anime based on Books, Based on Books...	1988	13+	Japanese [Original]	80217130, 70106454, 60023642, 70045021, 817255...
49	28630857	Princess Mononoke	Anime	Sci-Fi & Fantasy Anime, Family, Action Anime, ...	1997	13+	Japanese [Original], English, Hindi	80217130, 60023642, 81725555, 70019062, 600322...
90	60023642	Spirited Away	Anime	Family, Japanese, Anime Features	2001	13+	Japanese [Original], English, Hindi	80217130, 70106454, 70045021, 70019058, 815657...
91	60024179	Mobile Suit Gundam: Char's Counterattack	Anime	Sci-Fi & Fantasy Anime, Action Anime, Sci-Fi, ...	1988	13+	English, Japanese [Original]	81217220, 80231373, 60024788, 81033445, 811868...
92	60024788	The End of Evangelion	Anime	Sci-Fi & Fantasy Anime, Action Anime, Sci-Fi, ...	1997	A	English, Japanese [Original]	81736884, 80179831, 80174974, 28630857, 800013...
...	...	...	...	...	...	...	...	...
6310	81910037	Failure Frame: I Became the Strongest and Anni...	Anime	Sci-Fi & Fantasy Anime, Action Anime, Japanese...	2024	16+	NaN	81028712, 81598010, 80196595, 81681485, 814747...
6311	81910168	Fairy Tail: 100 Years Quest	Anime	Shounen Anime, Sci-Fi & Fantasy Anime, Action ...	2024	13+	NaN	81028712, 81091393, 70204957, 81448990, 817003...
6343	81924850	SPY x FAMILY CODE: White	Anime	Shounen Anime, Action Anime, Japanese, Action ...	2023	16+	Japanese [Original], English, Hindi	81028791, 81091393, 81054849, 81736884, 812616...
6369	81943491	Dragon Ball DAIMA	Anime	Family Time TV, Shounen Anime,	2024	7+	NaN	81091393, 80117291, 81736884, 70204957, 812230...

	N_id	Title	Main Genre	Sub Genres	Release Year	Maturity Rating	Original Audio	Recommendations
				Action Anime, J...				
6380	81947352	One Piece Fan Letter	Anime	Shounen Anime, Action Anime, Japanese, Anime S...	2024	13+	Japanese [Original]	81091393, 80117291, 81736884, 81943491, 817003...

267 rows × 8 columns

In [15]:
```
# Count the number of Animes Over the Years
anime_by_year = anime_df['Release Year'].value_counts()

anime_by_year = anime_by_year.reset_index()
anime_by_year.columns = ['Year', 'Count']

anime_by_year = anime_by_year.sort_values(by='Year', ascending=True)

anime_by_year
```

Out[15]:

	Year	Count
17	1981	2
31	1982	1
23	1988	2
25	1991	1
18	1992	2
27	1993	1
30	1994	1
32	1995	1
13	1997	3
20	1998	2
24	1999	2
28	2000	1
16	2001	2
26	2002	1
29	2003	1
19	2004	2
11	2006	4
22	2007	2

	Year	Count
21	2009	2
15	2011	2
12	2012	4
14	2013	3
9	2014	6
10	2015	6
8	2016	8
7	2017	17
4	2018	25
5	2019	22
6	2020	17
2	2021	31
1	2022	31
0	2023	34
3	2024	27
33	2025	1

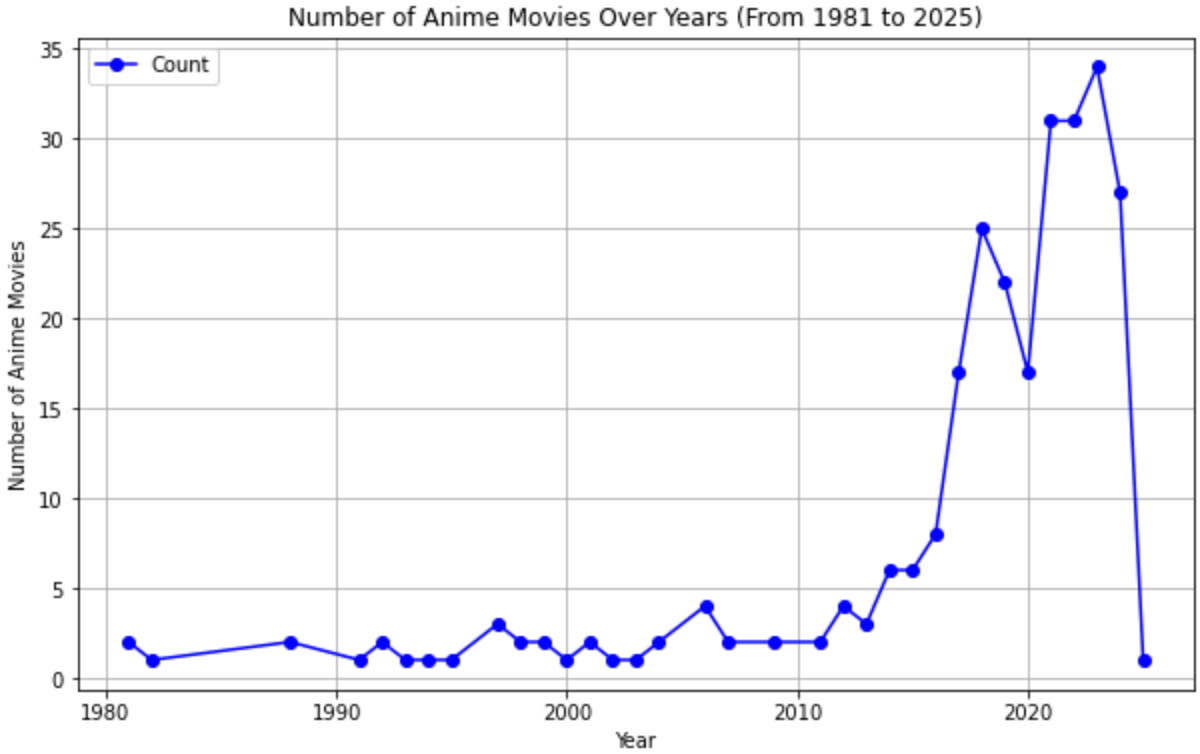
In [18]:

```
# Plot the line graph to show Animes over the years
anime_by_year.plot.line(
    x='Year',
    y='Count',
    figsize=(10, 6),  # Width, height in inches
    marker='o',       # Add markers at data points
    color='blue',     # Line color
    title='Number of Anime Movies Over Years (From 1981 to 2025)'  # Title
)

# Add labels
plt.xlabel('Year')
plt.ylabel('Number of Anime Movies')

# Add grid for readability
plt.grid(True)

# Display the plot
plt.show()
```


Number of Anime Movies Over Years (From 1981 to 2025)
Count
Number of Anime Movies
Year
0
5
10
15
20
25
30
35
1980
1990
2000
2010
2020

In [ ]: