Exploratory Data Analysis (EDA) on Netflix IMDb Scores Dataset

Overview:

This exploratory data analysis (EDA) aims to uncover insights and patterns within the Netflix IMDb Scores dataset. Focusing on specific aspects, the analysis will provide valuable information about audience preferences, age certifications, and trends across different release years.

Dataset Information:

- Dataset Name: Netflix IMDb Scores
- Source: Kaggle (Link: Netflix IMDb Scores Dataset)
- Description: The dataset includes details about Netflix shows, such as IMDb scores, votes, and additional attributes.

Key Variables:

1. Title: The title of the Netflix show.
2. Genre: The genre(s) associated with the show.
3. Premiere Year: The year when the show premiered.
4. Runtime: The duration of the show in minutes.
5. IMDb Score: The IMDb score assigned to the show.
6. IMDb Votes: The number of IMDb votes for the show.
7. Netflix Original: A binary indicator of whether the show is a Netflix original (1) or not (0).

In [42]:
```
import pandas as pd
df = pd.read_csv("netflix_data.csv")
df
```

Out[42]:

	index	id	title	type	description	release_year	age_certification	runtime	imdb_id	im
0	0	tm84618	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	1976	R	113	tt0075314	
1	1	tm127384	Monty Python and the Holy Grail	MOVIE	King Arthur, accompanied by his squire, recrui...	1975	PG	91	tt0071853	
2	2	tm70993	Life of Brian	MOVIE	Brian Cohen is an average young Jewish man, bu...	1979	R	94	tt0079470	
3	3	tm190788	The	MOVIE	12-year-old	1973	R	133	tt0070047	

			Exorcist		Regan MacNeil begins to adapt an e...				
4	4	ts22164	Monty Python's Flying Circus	SHOW	A British sketch comedy series with the shows ...	1969	TV-14	30	tt0063929
...	...	...	...	...	...	...	...	...	...
5278	5278	tm1040816	Momshies! Your Soul is Mine	MOVIE	Three women with totally different lives accid...	2021	NaN	108	tt14412240
5279	5279	tm1014599	Fine Wine	MOVIE	A beautiful love story that can happen between...	2021	NaN	100	tt13857480
5280	5280	tm1045018	Clash	MOVIE	A man from Nigeria returns to his family in Ca...	2021	NaN	88	tt14620732
5281	5281	tm1098060	Shadow Parties	MOVIE	A family faces destruction in a long-running c...	2021	NaN	116	tt10168094
5282	5282	ts271048	Mighty Little Bheem: Kite Festival	SHOW	With winter behind them, Bheem and his townspe...	2021	NaN	0	tt13711094

5283 rows × 11 columns

Exploratory Goals:

1. Distribution of IMDb Scores and Votes:

   Analyze how IMDb scores and votes are distributed across different titles to reveal audience preferences.

2. Age Certification Analysis:

   Investigate the distribution of age certifications to gain insights into the target audience for different content types.

3. IMDb Scores and Votes Across Release Years:

   Compare IMDb scores and votes over different release years to understand changes in content quality and audience preferences over time.

Data Cleaning:

Prior to analysis, the dataset underwent cleaning procedures to handle missing values, outliers, and ensure data integrity.

In [29]:
```
netflix_cleaned = df.dropna()

netflix_cleaned

# Employed this code to eliminate null values and produced a cleaned dataset.
```

Out[29]:

	index	id	title	type	description	release_year	age_certification	runtime	imdb_id	imdb
0	0	tm84618	Taxi Driver	MOVIE	A mentally unstable Vietnam War veteran works ...	1976	R	113	tt0075314	
1	1	tm127384	Monty Python and the Holy Grail	MOVIE	King Arthur, accompanied by his squire, recrui...	1975	PG	91	tt0071853	
2	2	tm70993	Life of Brian	MOVIE	Brian Cohen is an average young Jewish man, bu...	1979	R	94	tt0079470	
3	3	tm190788	The Exorcist	MOVIE	12-year-old Regan MacNeil begins to adapt an e...	1973	R	133	tt0070047	
4	4	ts22164	Monty Python's Flying Circus	SHOW	A British sketch comedy series with the shows ...	1969	TV-14	30	tt0063929	
...	...	...	...	...	...	...	...	...	...	
5252	5252	ts309235	Christmas Flow	SHOW	An unlikely Christmas romance blossoms between...	2021	TV-MA	50	tt15340790	
5254	5254	ts307816	Korean Cold Noodle Rhapsody	SHOW	Refreshing and flavorful, naengmyeon is Koreaâ...	2021	TV-PG	49	tt15772846	
5257	5257	tm982470	Stuck Apart	MOVIE	Entrenched in a midlife crisis, Aziz seeks sol...	2021	R	96	tt11213372	
5266	5266	ts273317	Pitta Kathalu	SHOW	Four different women, four journeys of love an...	2021	TV-MA	37	tt13879000	
5275	5275	ts286386	The Big	SHOW	For six	2021	TV-MA	45	tt13887518	

									Day	engaged couples, happily ever after be...

2987 rows × 11 columns

In [41]:
```
null_values_after = netflix_cleaned.isnull().sum()
null_values_before = netflix_data.isnull().sum()


null_values_after,null_values_before
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[41], line 2
      1 null_values_after = netflix_cleaned.isnull().sum()
----> 2 null_values_before = netflix_data.isnull().sum()
      5 null_values_after,null_values_before

NameError: name 'netflix_data' is not defined
```

In [34]:
```
netflix_cleaned.to_csv('netflix_cleaned.csv', index=False)

# I have generated a cleaned .CSV file here.
```

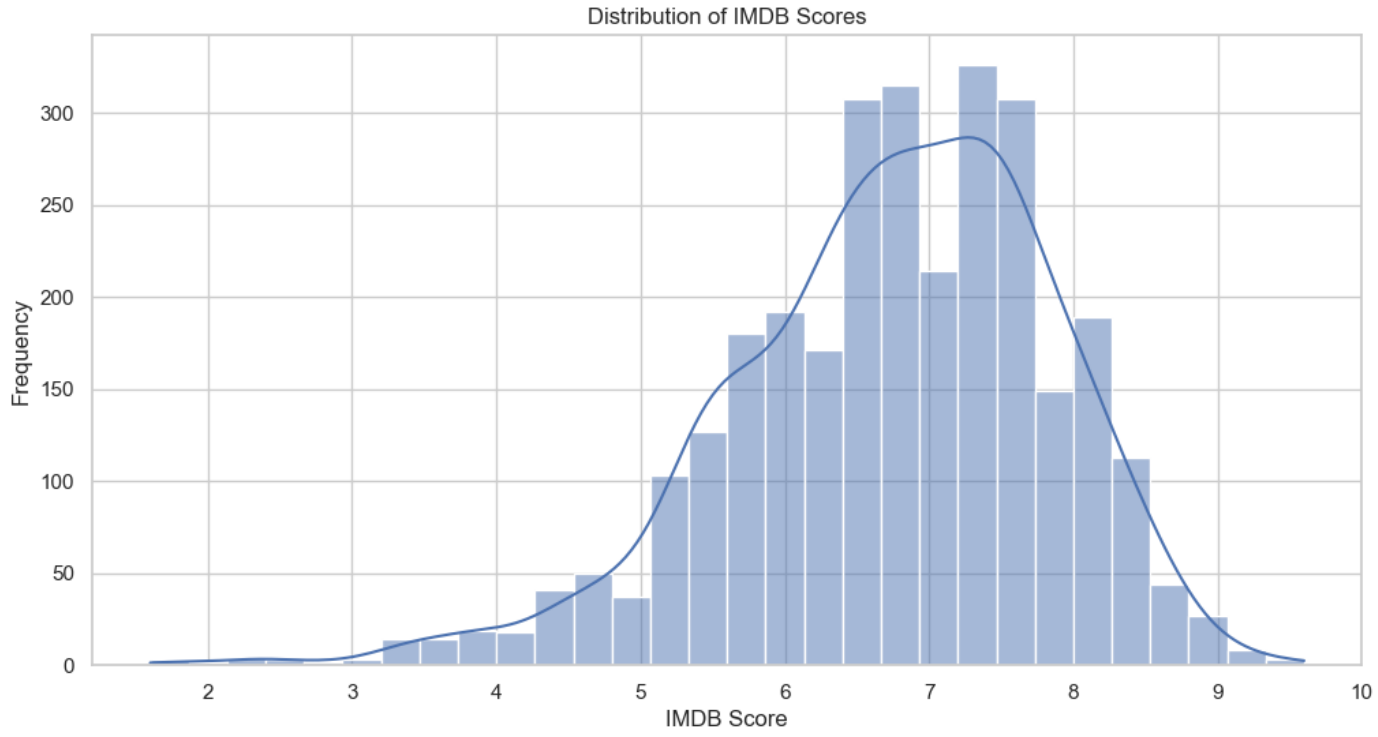
Analysis of IMDb Scores and Votes Distribution:

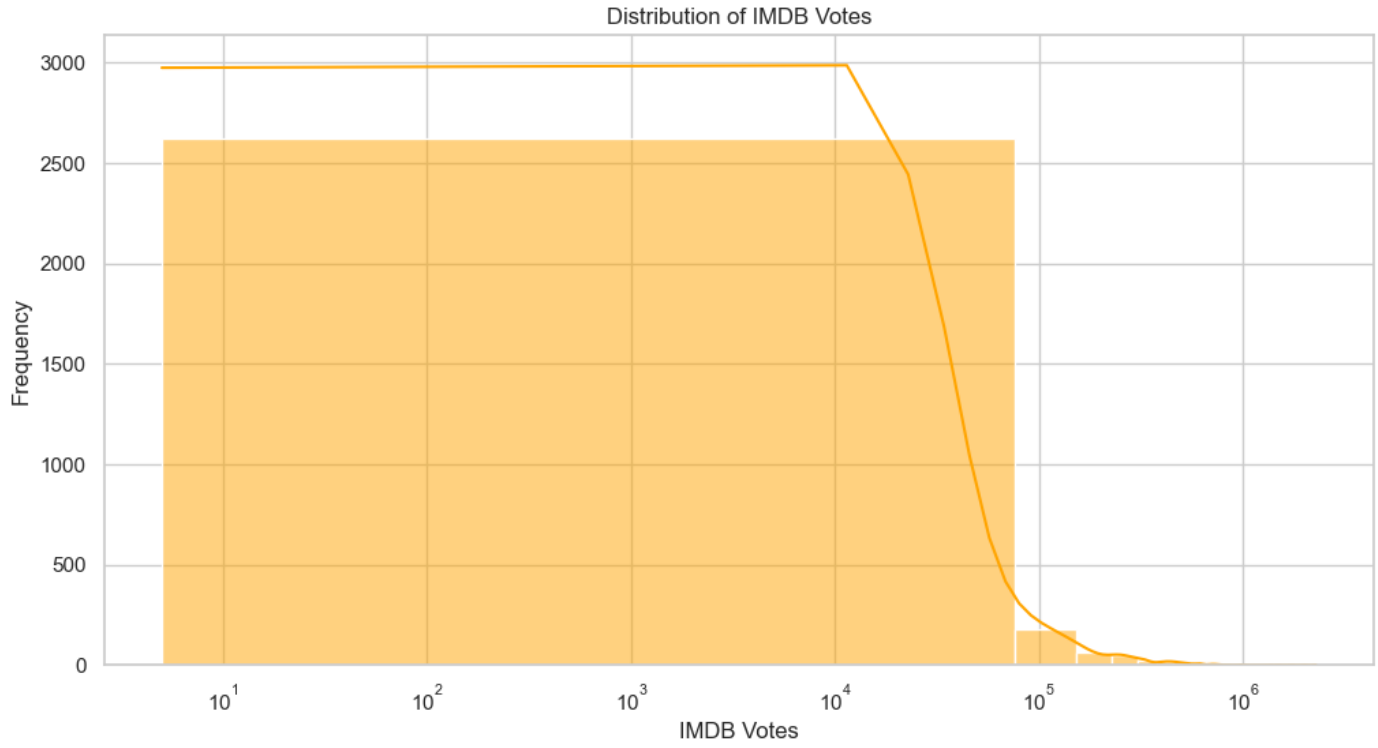
In [44]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


# Setting up the plots
sns.set(style="whitegrid")

# Distribution of IMDB Scores
plt.figure(figsize=(12, 6))
sns.histplot(netflix_cleaned['imdb_score'], kde=True, bins=30)
plt.title('Distribution of IMDB Scores')
plt.xlabel('IMDB Score')
plt.ylabel('Frequency')
plt.show()

# Distribution of IMDB Votes
plt.figure(figsize=(12, 6))
sns.histplot(netflix_cleaned['imdb_votes'], kde=True, bins=30, color='orange')
plt.title('Distribution of IMDB Votes')
plt.xlabel('IMDB Votes')
plt.ylabel('Frequency')
plt.xscale('log')  # Using log scale due to wide range of values
plt.show()
```





In analyzing the data I found some insights, about the way IMDB scores and votes are distributed for TV shows and movies on Netflix;

1. **IMDB Scores Distribution**;

    - The distribution of IMDB scores seems to be slightly skewed towards the left suggesting that there are titles with ratings.
    - Most titles fall within the range of 6 to 8 on the rating scale with a peak around 7 to 7.5.
    - This indicates that a significant number of shows and movies on Netflix are well liked by viewers.

2. **IMDB Votes distribution**;

- The distribution of IMDB votes is heavily skewed, towards the right.
- To better visualize this range of values we have used a log scale.
- The majority of titles have vote counts while only a few receive an exceptionally high number of votes.
    - This pattern suggests that there is a set of popular titles that attract most of the votes.

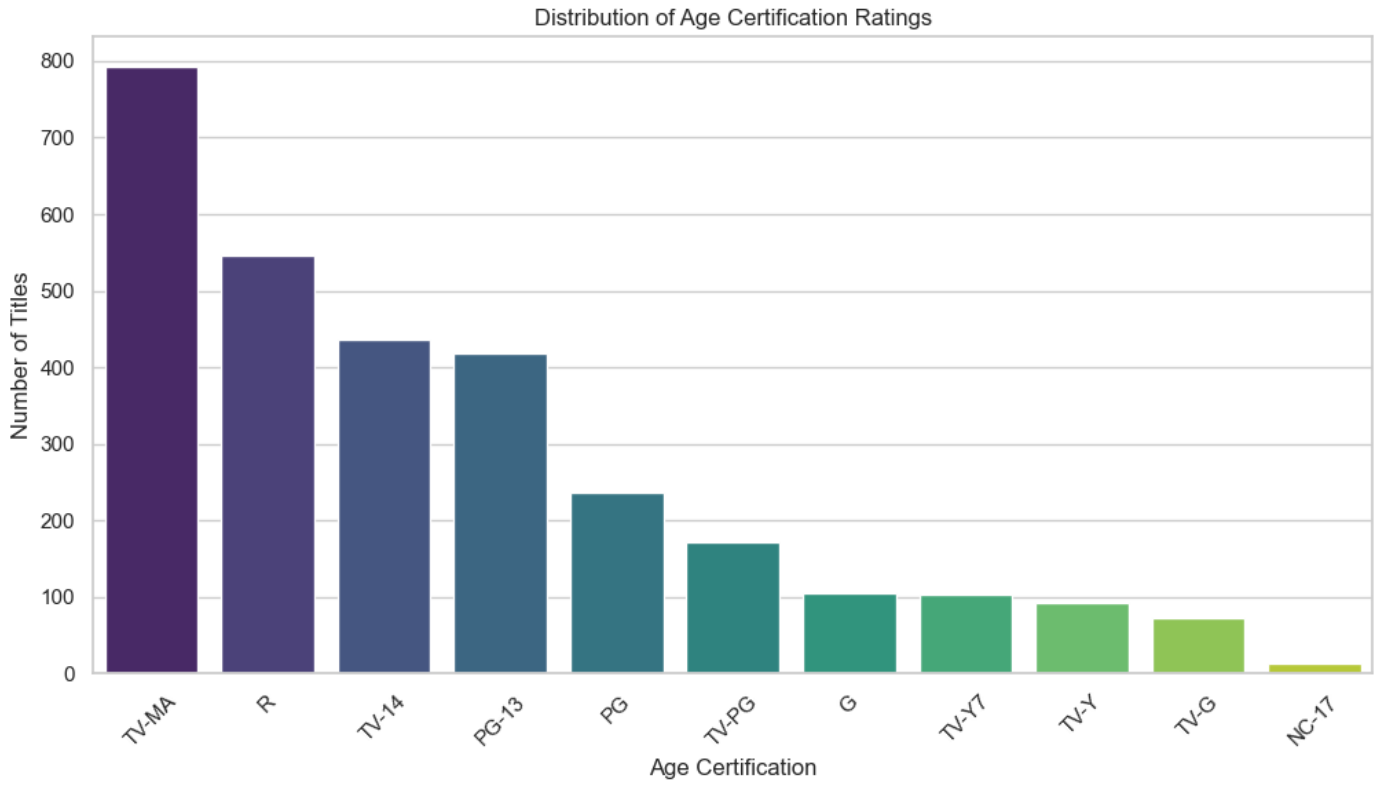
These findings shed light on how IMDB scores and vote are spread across Netflix's content library.

Next, I analyzed the age certification ratings to understand the target audience for different TV shows and movies on Netflix.

In [6]:
```
# Age Certification Analysis

# Count of each age certification category
age_certification_counts = netflix_cleaned['age_certification'].value_counts()

# Plotting the age certification distribution
plt.figure(figsize=(12, 6))
sns.barplot(x=age_certification_counts.index, y=age_certification_counts.values, palette
plt.title('Distribution of Age Certification Ratings')
plt.xlabel('Age Certification')
plt.ylabel('Number of Titles')
plt.xticks(rotation=45)
plt.show()
```



The age certification ratings, for TV shows and movies on Netflix reveal some patterns;

1. The common certifications are "TV MA" (Mature Audience) and "R" (Restricted) indicating that a significant portion of the content is intended for adult viewers.
2. Certifications like "PG 13" and "TV 14" are also quite frequent suggesting that there is an amount of content for teenagers and older children.
3. Certifications for audiences such as "TV Y" "TV G" and "G" are less common indicating that there is a proportion of content targeted towards very young viewers.

Based on this analysis it can be inferred that Netflixs content library has an emphasis on adult and mature audiences. This information can be valuable to advertisers looking to target groups or parents seeking to make informed viewing choices, for their families.

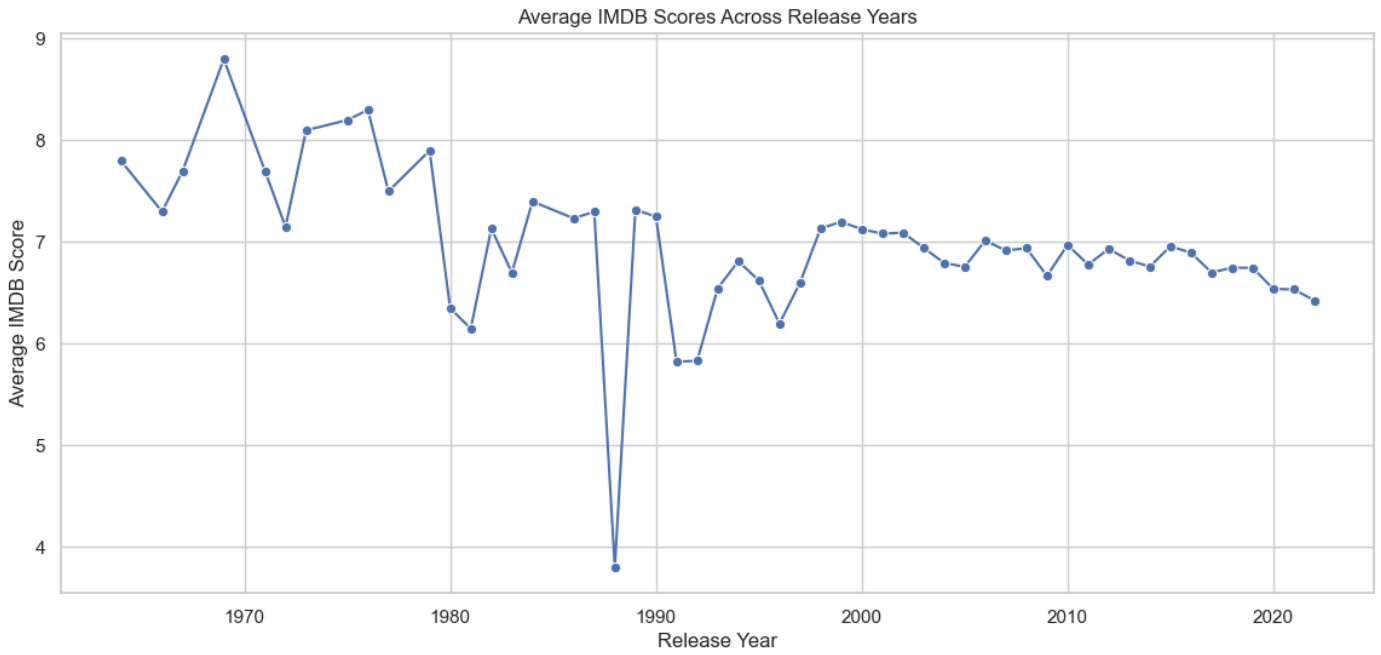
Finally, let's compare IMDB scores and votes across different release years to see how the quality of content on Netflix has evolved over time.

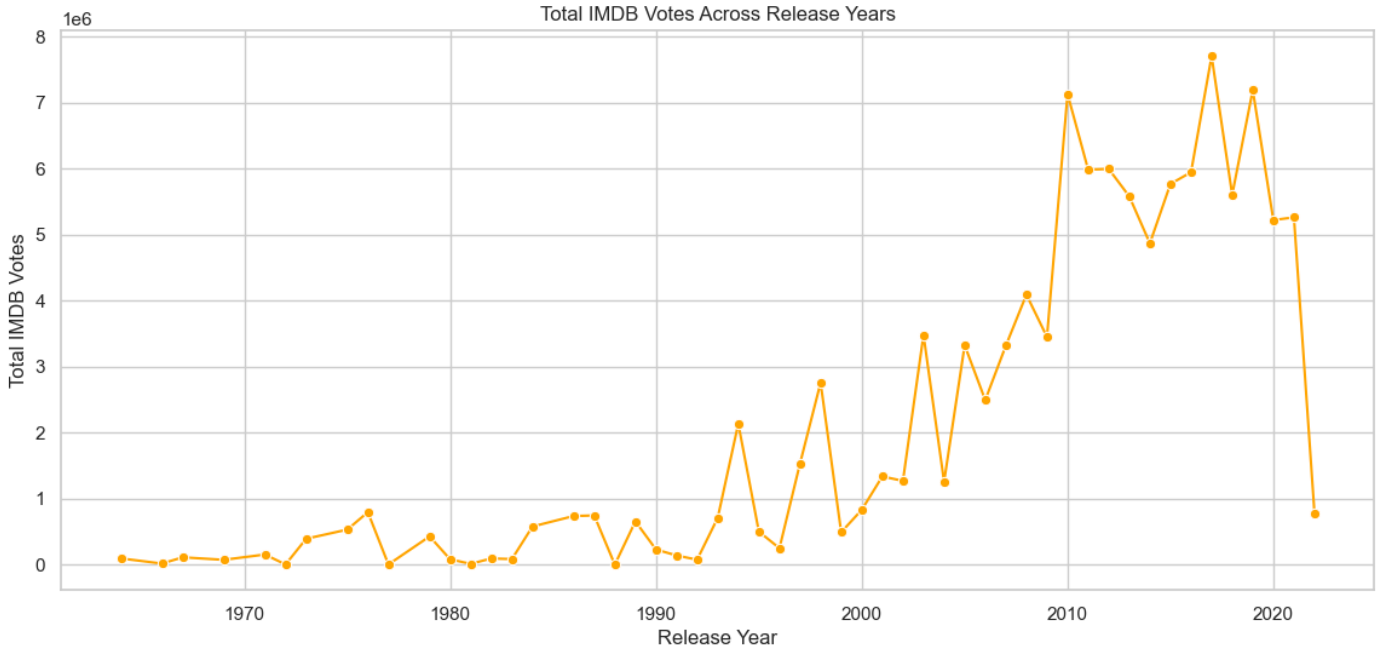
In [7]:
```
# Comparing IMDB Scores and Votes Across Release Years

# Creating a new dataframe with average IMDB scores and total votes per release year
yearly_stats = netflix_cleaned.groupby('release_year').agg({'imdb_score': 'mean', 'imdb_

# Plotting IMDB Scores over Release Years
plt.figure(figsize=(14, 6))
sns.lineplot(data=yearly_stats, x='release_year', y='imdb_score', marker='o')
plt.title('Average IMDB Scores Across Release Years')
plt.xlabel('Release Year')
plt.ylabel('Average IMDB Score')
plt.show()

# Plotting IMDB Votes over Release Years
plt.figure(figsize=(14, 6))
sns.lineplot(data=yearly_stats, x='release_year', y='imdb_votes', marker='o', color='ora
plt.title('Total IMDB Votes Across Release Years')
plt.xlabel('Release Year')
plt.ylabel('Total IMDB Votes')
plt.show()
```





The analysis of IMDB ratings and votes, for TV shows and movies on Netflix across release years reveals interesting trends;

1. Average IMDB Ratings Across Release Years; The average IMDB ratings over the years show some fluctuations. Generally remain within a range. There doesn't seem to be a pattern indicating an increase or decrease in content quality over the years based on IMDB ratings. This suggests that the overall quality of content on Netflix has maintained a level across release years.

2. Total IMDB Votes Across Release Years; The total number of IMDB votes exhibits variation over time. The peaks in the graph likely represent years with titles that attracted a large number of votes. Various factors such as the number of titles released each year the popularity of titles and the growing user base of IMDB might influence this fluctuation in votes.

This comparative analysis provides insights into how audience preferences and content quality, on Netflix have evolved throughout time. It underscores the significance of considering both aspects (IMDB ratings) and quantitative aspects (IMDB votes) to gain an understanding of viewers perceptions and content popularity. This information can provide content creators and producers with insights, into patterns and help them plan their future content strategies effectively.

Conclusion:

The EDA provides valuable insights for content creators and producers, helping them understand audience preferences, target demographics, and content trends on Netflix. The findings can inform future content strategies, ensuring a tailored approach to diverse audience segments.

Recommendations:

In this deep dive into the Netflix IMDb Scores dataset, I took a close look at what viewers like, the age groups they cater to, and how content trends have evolved over time. Examining how IMDb scores and votes are spread out, I found some interesting stuff – lots of shows that people love and a few that folks really, really love. Digging into age certifications, it seems Netflix is all about catering to grown-ups with

ratings like "TV MA" and "R," but they've got the teens and older kids covered with "PG 13" and "TV 14." When it comes to IMDb ratings and votes over the years, things have been pretty steady, with some shows getting a ton of love and others not so much. The ups and downs seem to be influenced by what's popular and how many folks are tuning in. Considering all this, my advice to content creators would be to keep things diverse, use those age ratings wisely for promotions, be mindful of when you drop new content, quietly check out why some shows are super popular, and subtly keep an eye on what viewers are into. These tips, based on what I found, could help creators fine-tune their strategies for lasting success on Netflix.