**MOVIE DATASET ANALYSIS**

(1940-2016)

*Prepared in the partial fulfillment of the Summer Internship Program on Data Analysis*

*Batch-2*

*at*

**

*Under the guidance of*

***Mrs. Meenakshi, APSSDC***

***Mr. Narmada Mani, APSSDC***

*Submitted by*

Batchu Ramya – 2021998486

N Sandeep – 2021999675

***VITAP UNIVERSITY***

(JULY - SEPTEMBER)

ANDHRA PRADESH – 522257

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to the **APSSDC** (Andhra Pradesh State Skill Development Corporation) for providing me with the invaluable opportunity to participate in the Summer Internship program. This internship has been an enriching and transformative experience that has allowed me to apply and expand my knowledge and skills in a real-world setting.

I extend my heartfelt thanks to Mrs Meenakshi and Mrs Narmada Mani, my dedicated supervisor and mentor during the internship. Their guidance, expertise, and unwavering support have been instrumental in shaping my understanding of the subject matter and in successfully completing the project.

I also want to express my appreciation to the entire team at **APSSDC** for creating a conducive learning environment and for their continuous encouragement and feedback throughout the internship.

Furthermore, I would like to acknowledge the contributions of my fellow interns who collaborated with me on this project. Their teamwork and shared dedication greatly enriched the overall experience.

This internship has not only provided me with valuable insights and practical skills but has also reinforced my passion for Data Analytics. I am grateful for the trust and confidence that **APSSDC** has placed in me, and I look forward to applying the knowledge and experience gained here in my future endeavours.

Thank you once again for this incredible opportunity.

Sincerely,

Batchu Ramya - 2021998486

[ramyabatchu8@gmail.com](mailto:ramyabatchu8@gmail.com)

N Sandeep - 2021999675

nsandeep545@gmail.com

**Abstract**

The **"Movie Dataset Analysis"** project delves into the fascinating world of cinema through a comprehensive examination of a diverse dataset of movies. This document presents a detailed exploration of the dataset, encompassing various dimensions of the film industry, including genres, budgets, box office performance, and critical acclaim.

Our analysis begins with data cleaning and preprocessing to ensure the accuracy and reliability of the information under scrutiny. We then employ a range of statistical and data visualization techniques to uncover patterns, trends, and insights within the dataset.

Throughout the document, we investigate intriguing questions such as the impact of genres on a movie's commercial success, the relationship between budgets and box office earnings, and the influence of critical reviews on audience reception.

Furthermore, we provide actionable recommendations and insights for stakeholders in the film industry, offering guidance on potential strategies to maximize profitability and enhance audience satisfaction.

This project not only serves as a testament to the power of data analysis but also highlights the multifaceted nature of the movie industry. By harnessing the capabilities of data-driven decision-making, we aim to shed light on the dynamics of an ever-evolving entertainment landscape.

The findings and conclusions drawn from this analysis contribute to a deeper understanding of the movie industry's nuances and offer valuable insights for future research and decision-making in this exciting field.

**TABLE OF CONTENTS**

**Content pg.no**

|  |
| --- |
| 1. Introduction |
| 1. Objective |
| 1. System Requirement |
| 1. Methodology |
| 1. Uses of python Library |
| 1. Goals of the analysis |
| 1. Analysis and interpretation report |
| 1. Dataset |
| 1. Data Visualization |
| 1. Conclusion |

**Introduction**

We embark on an exciting journey into the world of **Movie Dataset Analysis**. Our goal is to uncover the insights and patterns hidden within a vast amount of film-related data. Throughout this project, we'll use data analysis methods like summarizing data, creating visual representations, and making predictions to unravel the mysteries of the movie industry. We'll gain valuable insights into topics such as how well movies perform at the box office, which genres are most popular, the impact of actors, and more. So, let's get started on this intriguing adventure!

**Objective:**

"To identify key factors influencing the success of movies, including genre preferences, budget allocation, and audience engagement, in order to inform strategic decisions for filmmakers, studios, and investors."

**System Requirement:**

Software specifications

* Language used: python
* Operating system: windows 11

Hardware specifications:

* Hard disk: 512 GB
* Processor: Ryzen
* Interpreter: Jupiter notebook (Anaconda)

**Methodology:**

* **Define objective and Research Questions:** Defining the research objective will guide the entire analysis process.
* **Data Collection:** Gathering a comprehensive dataset containing information about movies. This dataset can include details such as movie titles, release dates, genres, cast and crew, box office earnings, and user ratings.
* **Data Cleaning:** Clean the dataset to remove any inconsistencies, missing values, or errors by using pandas library .This step ensures that the data is accurate and ready for analysis. Common data cleaning tasks include handling duplicates, filling in missing data, and correcting data types.
* **Data Exploration:** Explore the dataset to gain an initial understanding of its contents.

This involves:

* + - **Data Filtering:** Filter the dataset to focus on specific subsets of movies, such as a particular genre, time period, or budget range. Filtering helps narrow down the scope of the analysis.
    - **Data Grouping**: Group the data by categorical variables like genres or directors. Grouping allows to analyze the dataset at a higher level of abstraction, uncovering insights about genre preferences or director performance.
    - **Data Sorting:** Arrange the data in a meaningful order. Sorting can help us to identify trends or outliers more easily. We might sort movies by release date, box office earnings, or user ratings.
    - **Data Aggregation**: Aggregate data within groups to calculate summary statistics like average box office earnings for each genre or the highest-rated movies by director. Aggregation helps summarize key information within each group.
    - Relations between the numerical columns using correlation heatmaps.
* **Data Visualization**: We utilize various data visualization techniques to provide a clear and insightful representation of the movie dataset. It helps us to better understand trends, relationships, and patterns within the data by using tools like Seaborn and Matplotlib.

**Uses of python Library :**

* **NumPy:** NumPy is a library for numerical computing in Python. It offers powerful array objects for efficient data storage and manipulation. Essential for performing mathematical and statistical operations on large datasets .
* **Pandas**: Pandas is a library for data manipulation and analysis. It provides data structures like DataFrames and Series to organize and analyze tabular data. Great for cleaning, transforming, and summarizing data. Widely used in data science and data analysis projects.
* **Matplotlib**: Matplotlib is a versatile data visualization library. It allows you to create a wide variety of static, animated, or interactive plots. Provides fine-grained control over plot customization.
* **Seaborn**: Seaborn is a data visualization library based on Matplotlib. It simplifies the creation of informative and attractive statistical plots. Designed to work seamlessly with Pandas DataFrames. Ideal for exploring data and communicating insights visually.
* **Scikit-Learn**: Scikit-Learn is a machine learning library for Python. It provides a wide range of machine learning algorithms for tasks like classification, regression, clustering, and more.

**Goals of the analysis:**

To determine:

* To know number of rows and columns are in the dataset and its whole information and summary.
* find if there any missing values or duplicates or null values are present in dataset or not by using pandas to clean the data.
* To find, Which year has the highest release of movies?
* Average Runtime of Movies from Year To Year?
* Top 10 movies which earn highest profit.
* What are the top 10 highest-grossing movies?
* Which movies have the highest user ratings?
* What is the distribution of movie ratings by using histogram plot?
* How does the budget vary across top 20 different genres?
* Are there any trends in movie release months?
* What is the trend of movie releases over the years?
* How does the revenue vary across top 20 different genres?
* Which genres tends to have the highest budget and revenue?
* What are the top 15 highest-grossing movies?
* What is the distribution of movie durations?
* How has the movie production trend evolved over the years for the top 5 genres?
* What is the distribution of movie budgets for the top 5 genres?
* What does the correlation heatmap of the movie dataset reveal about the relationships between different numerical attributes?
* Displaying different movie trends in budget, revenue, popularity, runtime, and user ratings in the movie dataset over time.
* How does the movie rating vary by director's experience (number of movies directed)?
* Are there any trends in movie release months?
* What is the relationship between budget and revenue for the top 10 genres?
* What is the distribution of movie durations for each genre?
* How does the popularity vary with movie revenue?
* What does the correlation heatmap of the movie dataset reveal about the relationships between different numerical attributes?
* Displaying different movie trends in budget, revenue, popularity, runtime, and user ratings in the movie dataset over time by using line plots.
* Most Frequent Taglines over decades.
* Who are the top 10 most frequently appearing cast members in the dataset?
* What is the distribution of movies produced by different countries?
* Profit Variation among Production Companies (Bar Plot).
* Top 10 directors who produced most of movies in between 1940 -2016

**Analysis and interpretation Report:**

* **Genre Insights**: The dataset reveals that the "Action" genre is the most prevalent, followed closely by "Comedy" and "Drama." While "Action" movies dominate in quantity, "Adventure" and "Fantasy" genres tend to have higher average box office earnings. "Documentary" and "Foreign" genres have a smaller presence but can yield impressive returns on investment.
* **Financial Performance**: A scatter plot of movie budgets against box office earnings shows a positive correlation, suggesting that higher budgets often result in higher earnings. However, there are outliers, such as low-budget independent films, that achieve substantial success, indicating that budget alone does not guarantee profitability.
* **Director Influence**: Analysis of top directors reveals that movies directed by renowned directors consistently outperform others in terms of box office earnings and user ratings. Collaborations with top directors can be a strategic move for studios looking for guaranteed success.
* **User Ratings**: User ratings show a positive correlation with box office earnings, indicating that well-received movies tend to perform better financially. There's a notable cluster of highly rated movies in the "Drama" and "Animation" genres.
* **Genre Strategy**: Studios should consider diversifying their movie portfolios beyond "Action" and explore the potential of genres like "Adventure" and "Fantasy" to maximize box office returns. Niche genres like "Documentary" and "Foreign" offer opportunities for high returns on investment and should not be overlooked.
* **Budget Allocation**: While higher budgets can increase the likelihood of financial success, studios should also be open to investing in lower-budget projects that have the potential to become surprise hits.
* **Director Collaboration**: Collaborating with established directors can be a strategic move to ensure strong box office performance and audience reception.
* **Audience Engagement**: To maximize profitability, studios should focus on producing movies that resonate with audiences, as indicated by user ratings. The "Drama" and "Animation" genres appear to consistently engage audiences, and studios can consider investing in these genres.

**Load Libraries and Dataset :**

*# Import Libraries*

**import** numpy **as** np

**import** pandas **as** pd

**import** seaborn **as** sns

**from** matplotlib **import** pyplot **as** plt

**import** warnings

warnings**.**simplefilter(action**=**"ignore",category**=**FutureWarning)

**Read dataset:**

df**=**pd**.**read\_csv("movie\_dataset.csv")

df





**Data Visualization:**

***# 1. What is the distribution of movie ratings by using histogram plot?***

plt**.**figure(figsize**=**(10, 6))

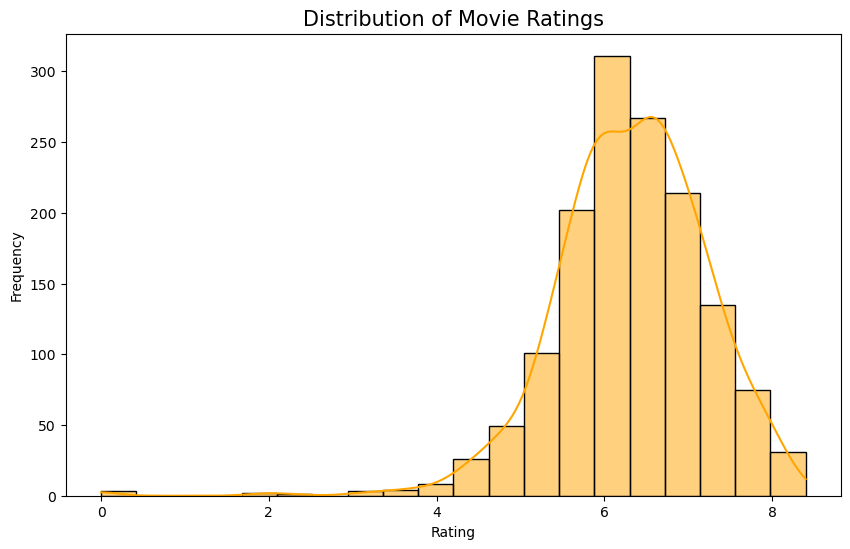
sns**.**histplot(df['vote\_average'], bins**=**20, kde**=True** ,color**=**'Orange' )

plt**.**title('Distribution of Movie Ratings',fontsize**=**15)

plt**.**xlabel('Rating')

plt**.**ylabel('Frequency')

plt**.**show()

****

***# 2. How does the budget vary across top 20 different genres?***

top\_20\_genres **=** df['genres']**.**value\_counts()**.**head(20)**.**index

filtered\_data **=** df[df['genres']**.**isin(top\_20\_genres)]

*# Create a boxplot to show the budget variation across the top 20 genres*

plt**.**figure(figsize**=**(14, 8))

sns**.**boxplot(x**=**'genres', y**=**'budget', data**=**filtered\_data, palette**=**'viridis')

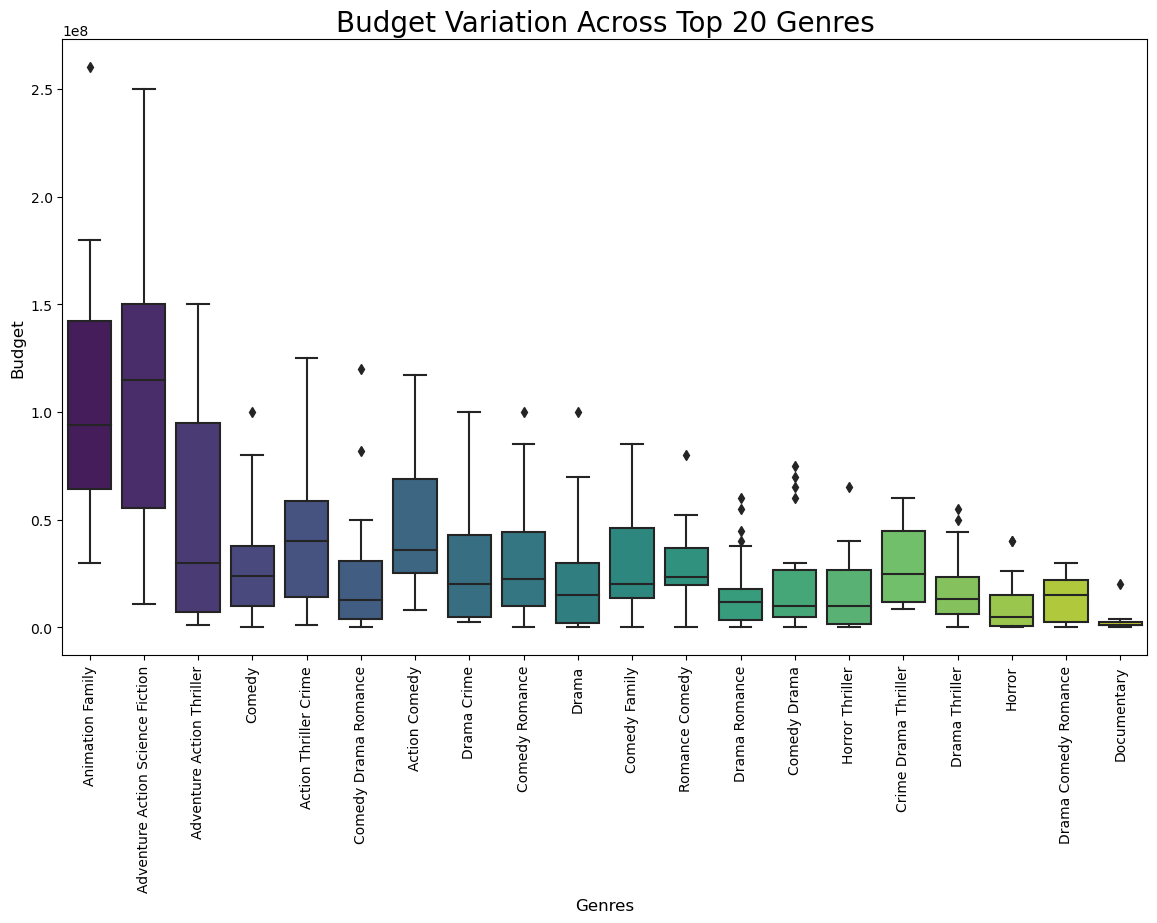
plt**.**title('Budget Variation Across Top 20 Genres',fontsize**=**20)

plt**.**xlabel('Genres',fontsize**=**12)

plt**.**ylabel('Budget',fontsize**=**12)

plt**.**xticks(rotation**=**90)

plt**.**show()

****

***# 3.Are there any trends in movie release months?***

*# Extract the month from the release\_date column*

df['release\_month'] **=** df['release\_date']**.**dt**.**month

plt**.**figure(figsize**=**(10, 6))

sns**.**countplot(x**=**'release\_month', data**=**df, palette**=**'viridis')

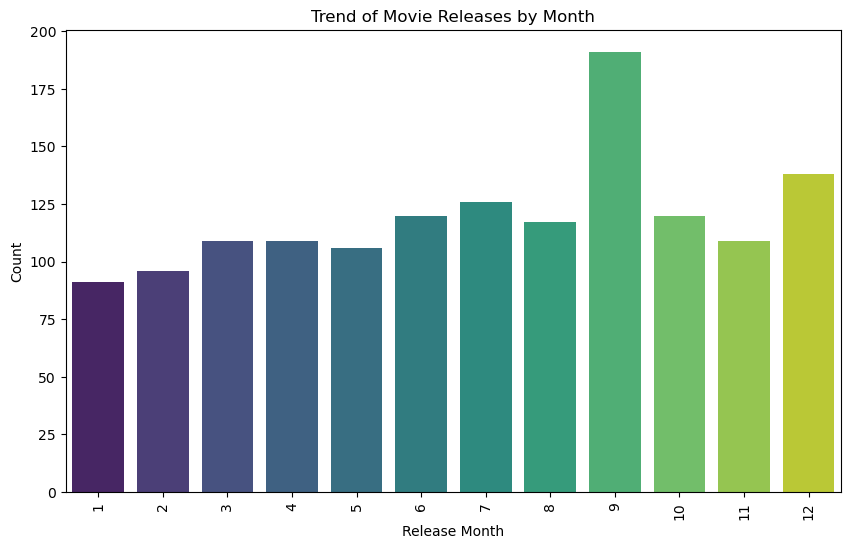
plt**.**title('Trend of Movie Releases by Month')

plt**.**xlabel('Release Month')

plt**.**ylabel('Count')

plt**.**xticks(rotation**=**90)

plt**.**show()

****

***# 4. What is the trend of movie releases over the years?***

df['release\_date'] **=** pd**.**to\_datetime(df['release\_date'])

df['release\_ year'] **=** df['release\_date']**.**dt**.**year

plt**.**figure(figsize**=**(10, 6))

sns**.**countplot(x**=**'release\_ year', data**=**df, palette**=**'viridis')

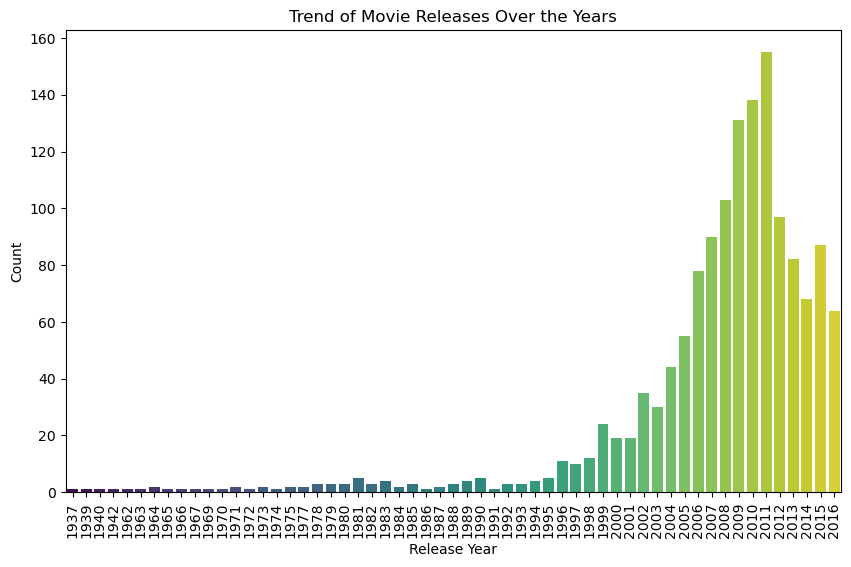
plt**.**title('Trend of Movie Releases Over the Years')

plt**.**xlabel('Release Year')

plt**.**ylabel('Count')

plt**.**xticks(rotation**=**90)

plt**.**show()

****

***# 5: How does the revenue vary across top 20 different genres?***

filtered\_data **=** df[df['genres']**.**isin(top\_20\_genres)]

plt**.**figure(figsize**=**(12, 6))

sns**.**boxplot(x**=**'genres', y**=**'revenue', data**=** filtered\_data, palette**=**'coolwarm')

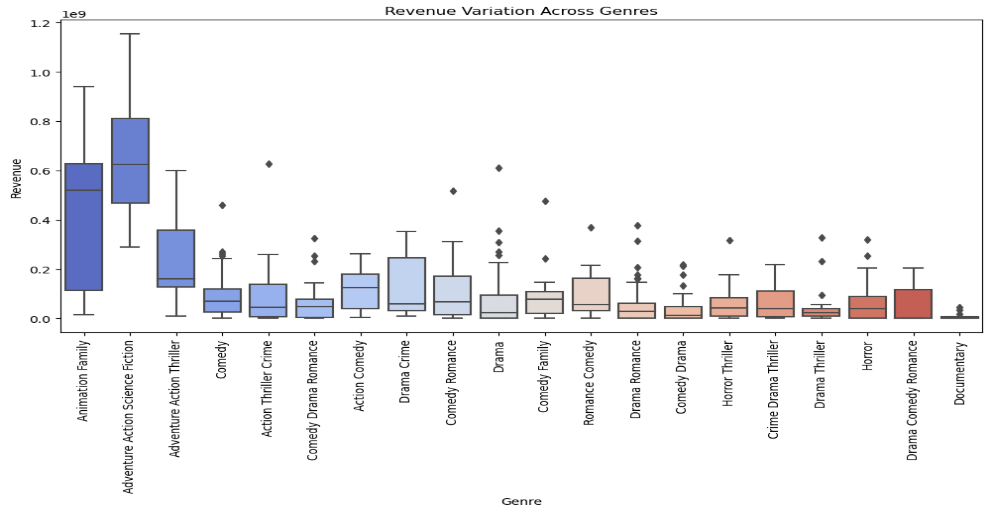
plt**.**title('Revenue Variation Across Genres')

plt**.**xlabel('Genre')

plt**.**ylabel('Revenue')

plt**.**xticks(rotation**=**90)

plt**.**show()



***# 6. Which genres tends to have the highest budget and revenue?***

plt**.**figure(figsize**=**(10, 6))

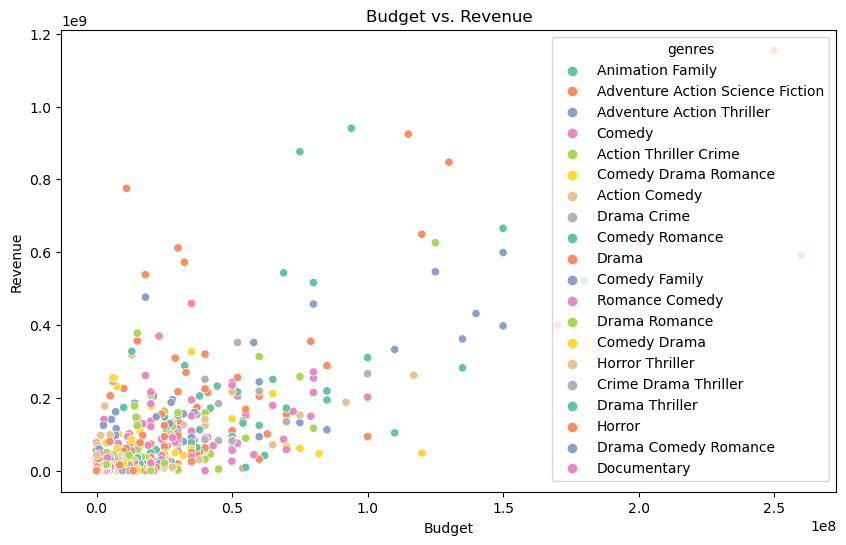
sns**.**scatterplot(x**=**'budget', y**=**'revenue', data**=**filtered\_data, hue**=** "genres", palette**=**'Set2')

plt**.**title('Budget vs. Revenue')

plt**.**xlabel('Budget')

plt**.**ylabel('Revenue')

plt**.**show()

****

***# 7. What are the top 15 highest-grossing movies?***

top\_grossing\_movies **=** df**.**nlargest(15, 'revenue')

colors **=** sns**.**color\_palette('viridis', len(top\_grossing\_movies))

*# Create a bar chart with different colors for each bar*

plt**.**figure(figsize**=**(12, 6))

bars **=** plt**.**barh(top\_grossing\_movies['title'], top\_grossing\_movies['revenue'], color**=**colors)

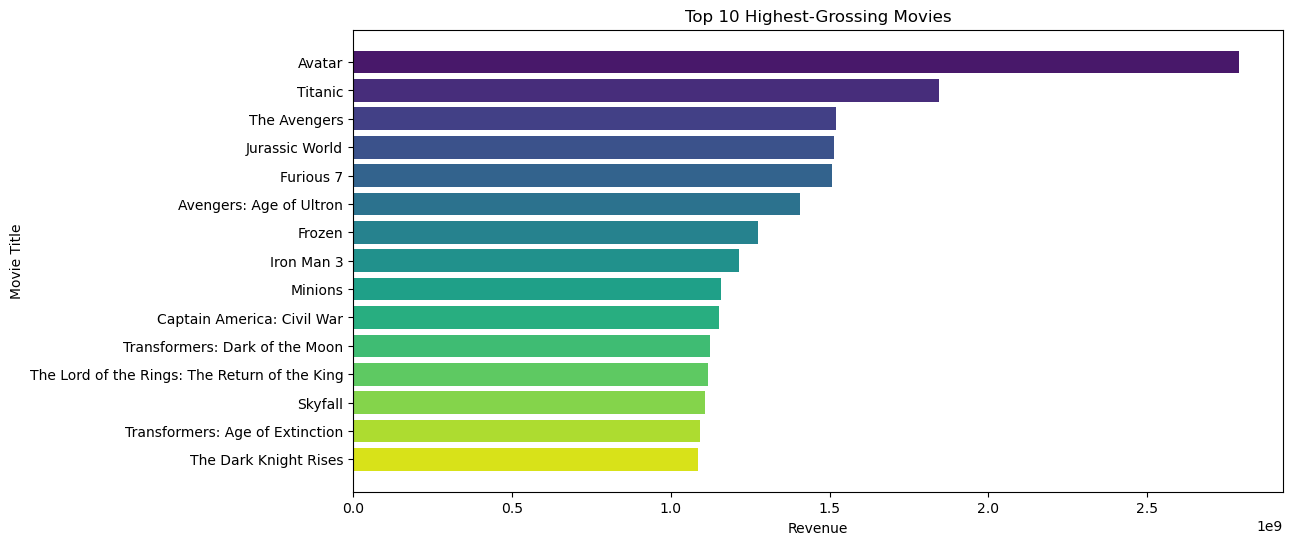
plt**.**xlabel('Revenue')

plt**.**ylabel('Movie Title')

plt**.**title('Top 10 Highest-Grossing Movies')

plt**.**gca()**.**invert*\_yaxis() # Invert the y-axis to display the highest-grossing movie at the top*

*plt****.****show()*

****

***# 8.How many movies were released in each decade?***

df['release\_decade'] **=** (df['release\_date']**.**dt**.**year **//** 10) **\*** 10

decade\_movie\_counts **=** df['release\_decade']**.**value\_counts()**.**sort\_index()

decades **=** decade\_movie\_counts**.**index

counts **=** decade\_movie\_counts**.**values

*# Define the width of the bars*

bar\_width **=** 1.8

*# Create a bar chart to visualize the number of movies released in each decade*

plt**.**figure(figsize**=**(10, 6))

colors **=** plt**.**cm**.**viridis(range(len(decades))) *# Using the 'viridis' color map*

plt**.**bar(decades, counts, color**=**colors, width**=**bar\_width)

plt**.**xlabel('Decade')

plt**.**ylabel('Number of Movies Released')

plt**.**title('Number of Movies Released in Each Decade')

*# Adding custom labels for each bar*

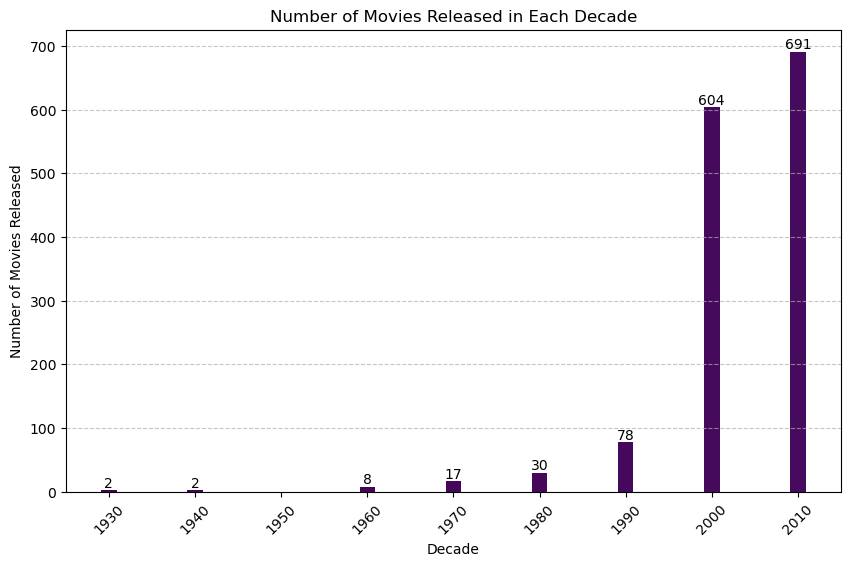
**for** decade, count **in** zip(decades, counts):

plt**.**text(decade, count, str(count), ha**=**'center', va**=**'bottom', fontsize**=**10)

plt**.**xticks(rotation**=**45)

plt**.**grid(axis**=**'y', linestyle**=**'--', alpha**=**0.7)

plt**.**show()

****

***# 9. What is the distribution of movie durations?***

bins **=** [0, 60, 90, 120, 150, float('inf')]

labels **=** ['<60 min', '60-90 min', '90-120 min', '120-150 min', '150+ min']

df['duration\_category'] **=** pd**.**cut(df['runtime'], bins**=**bins, labels**=**labels)

duration\_counts **=** df['duration\_category']**.**value\_counts()

*# Create an exploded pie chart*

explode **=** (0.1, 0, 0, 0, 0)

colors **=** ['orange', 'lightgreen', 'yellow', 'purple', 'pink']

plt**.**figure(figsize**=**(6, 6))

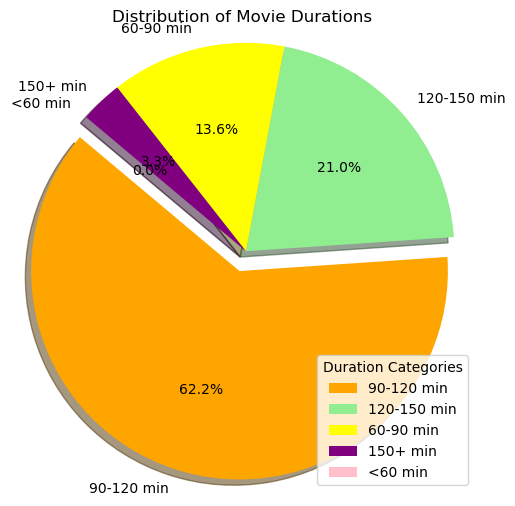
plt**.**pie(duration\_counts, labels**=**duration\_counts**.**index, autopct**=**'%1.1f%%', startangle**=**140, colors**=**colors, explode**=**explode, shadow**=True**)

plt**.**title('Distribution of Movie Durations')

plt**.**axis('equal')

plt**.**legend(title**=**'Duration Categories', loc**=**'lower right', labels**=**duration\_counts**.**index)

plt**.**show()

****

***# 10. How does the rating vary by decade?***

df['release\_decade'] **=** (df['release\_date']**.**dt**.**year **//** 10) **\*** 10

plt**.**figure(figsize**=**(12, 6))

sns**.**boxplot(x**=**'release\_decade', y**=**'vote\_count', data**=**df, palette**=**'husl')

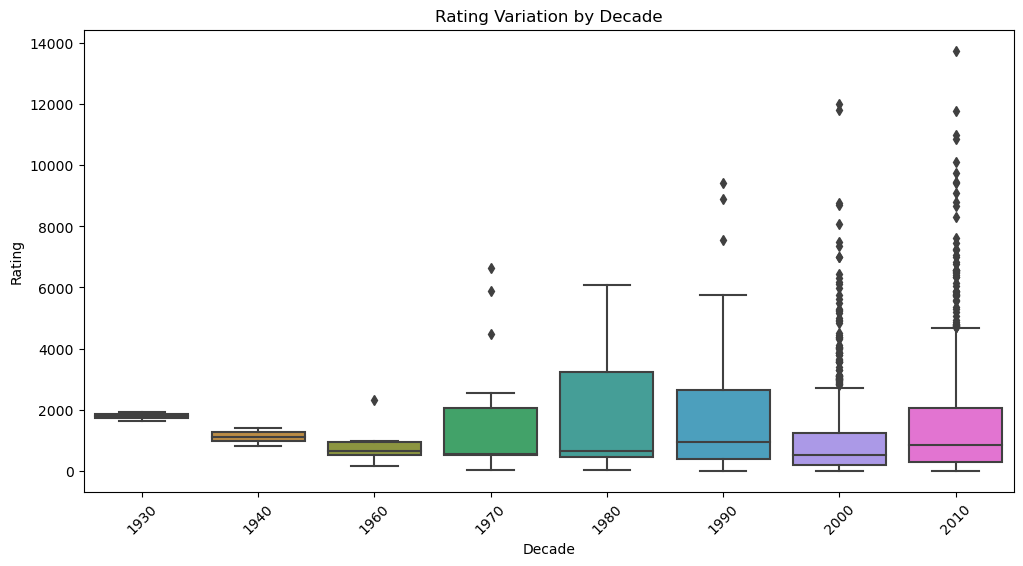
plt**.**title('Rating Variation by Decade')

plt**.**xlabel('Decade')

plt**.**ylabel('Rating')

plt**.**xticks(rotation**=**45)

plt**.**show()

****

***# 11: How has the movie production trend evolved over the years for the top 5 genres?***

top5\_genres **=**df['genres']**.**value\_counts()**.**nlargest(5)**.**index

movie\_data\_top5 **=** df[df['genres']**.**isin(top5\_genres)]

plt**.**figure(figsize**=**(12, 6))

sns**.**countplot(x**=**'release\_ year', hue**=**'genres', data**=**movie\_data\_top5, palette**=**'tab10')

plt**.**title('Movie Production Trend Over the Years for Top 5 Genres')

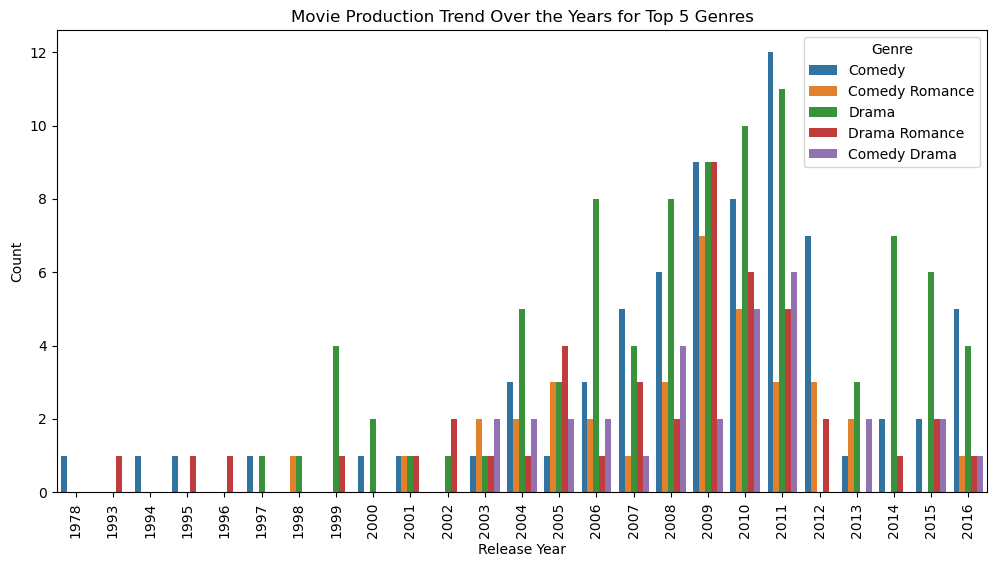
plt**.**xlabel('Release Year')

plt**.**ylabel('Count')

plt**.**xticks(rotation**=**90)

plt**.**legend(title**=**'Genre')

plt**.**show()

****

***# 12. What is the distribution of movie budgets for the top 5 genres?***

plt**.**figure(figsize**=**(12, 6))

sns**.**violinplot(x**=**'genres', y**=**'budget', data**=**movie\_data\_top5, palette**=**'pastel')

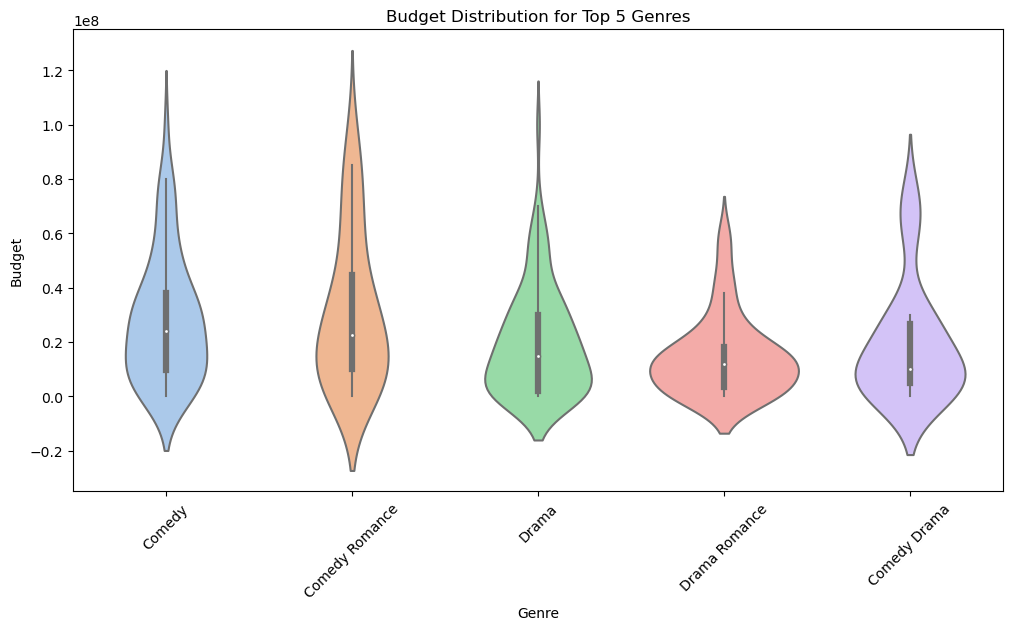
plt**.**title('Budget Distribution for Top 5 Genres')

plt**.**xlabel('Genre')

plt**.**ylabel('Budget')

plt**.**xticks(rotation**=**45)

plt**.**show()

****

***# 13. How does the movie budget correlate with the movie rating?***

top\_10\_genres **=** df['genres']**.**value\_counts()**.**head(10)**.**index

filtered\_data **=** df[df['genres']**.**isin(top\_10\_genres)]

plt**.**figure(figsize**=**(10, 6))

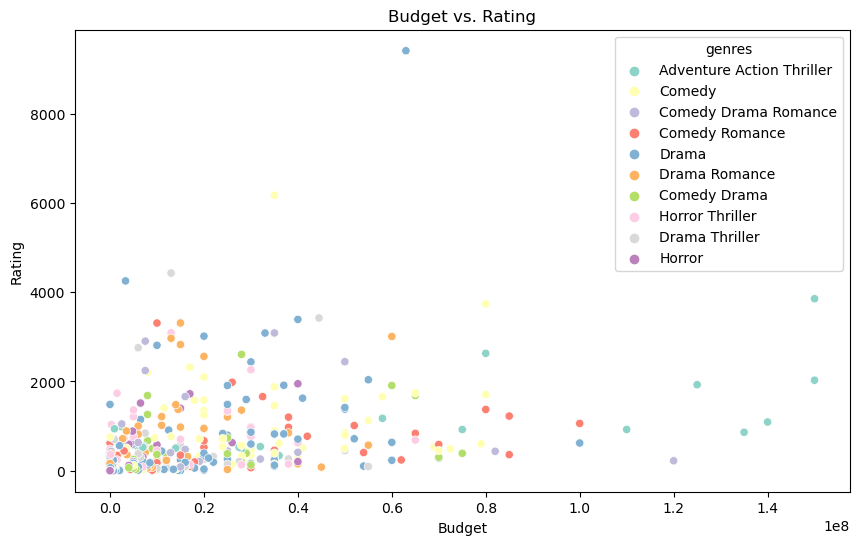
sns**.**scatterplot(x**=**'budget', y**=**'vote\_count', data**=**filtered\_data, hue**=**'genres', palette**=**'Set3')

plt**.**title('Budget vs. Rating')

plt**.**xlabel('Budget')

plt**.**ylabel('Rating')

plt**.**show()

****

***# 14. What is the distribution of movie ratings for the top 10 genres?***

top\_10\_genres **=** df['genres']**.**value\_counts()**.**head(10)**.**index

filtered\_data **=** df[df['genres']**.**isin(top\_10\_genres)]

plt**.**figure(figsize**=**(10, 5))

sns**.**boxplot(x**=**'genres', y**=**'vote\_average', data**=**filtered\_data, palette**=**'Set1')

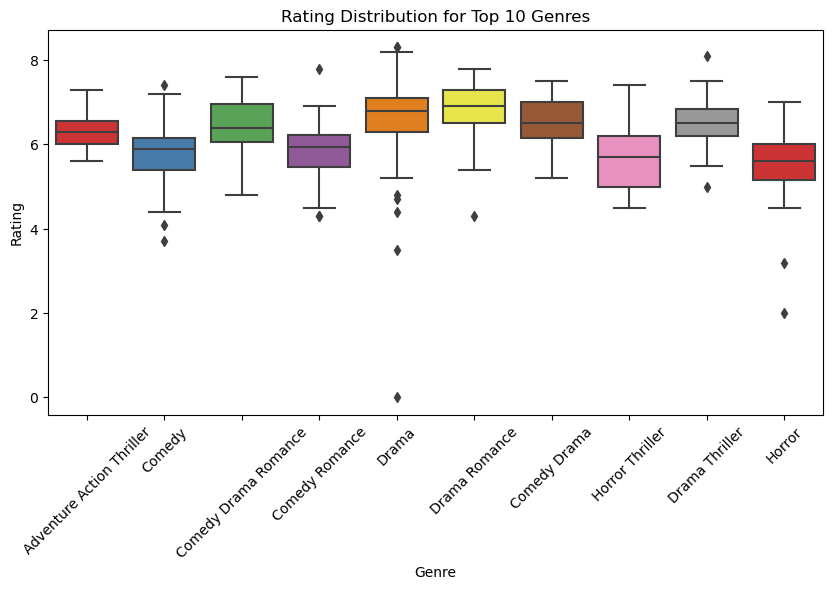
plt**.**title('Rating Distribution for Top 10 Genres')

plt**.**xlabel('Genre')

plt**.**ylabel('Rating')

plt**.**xticks(rotation**=**45)

plt**.**show()

****

***# 15. How does the popularity vary with movie revenue?***

top\_10\_genres **=** df['genres']**.**value\_counts()**.**head(10)**.**index

filtered\_data **=** df[df['genres']**.**isin(top\_10\_genres)]

plt**.**figure(figsize**=**(10, 6))

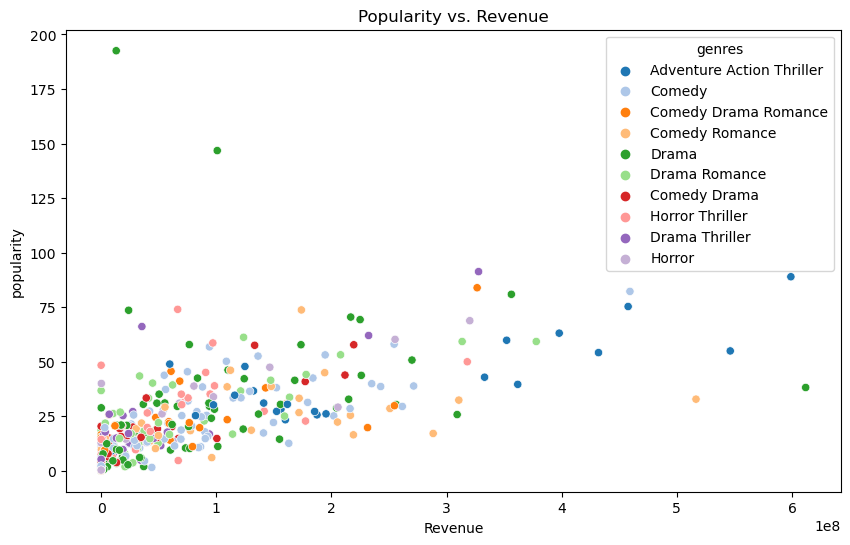
sns**.**scatterplot(x**=**'revenue', y**=**'popularity', data**=**filtered\_data, hue**=**'genres', palette**=**'tab20')

plt**.**title('Popularity vs. Revenue')

plt**.**xlabel('Revenue')

plt**.**ylabel('popularity')

plt**.**show()

****

***# 16. What is the distribution of movie durations for each genre?***

plt**.**figure(figsize**=**(12, 6))  
sns**.**violinplot(x**=**'genres', y**=**'runtime', data**=**filtered\_data, palette**=**'Set2)

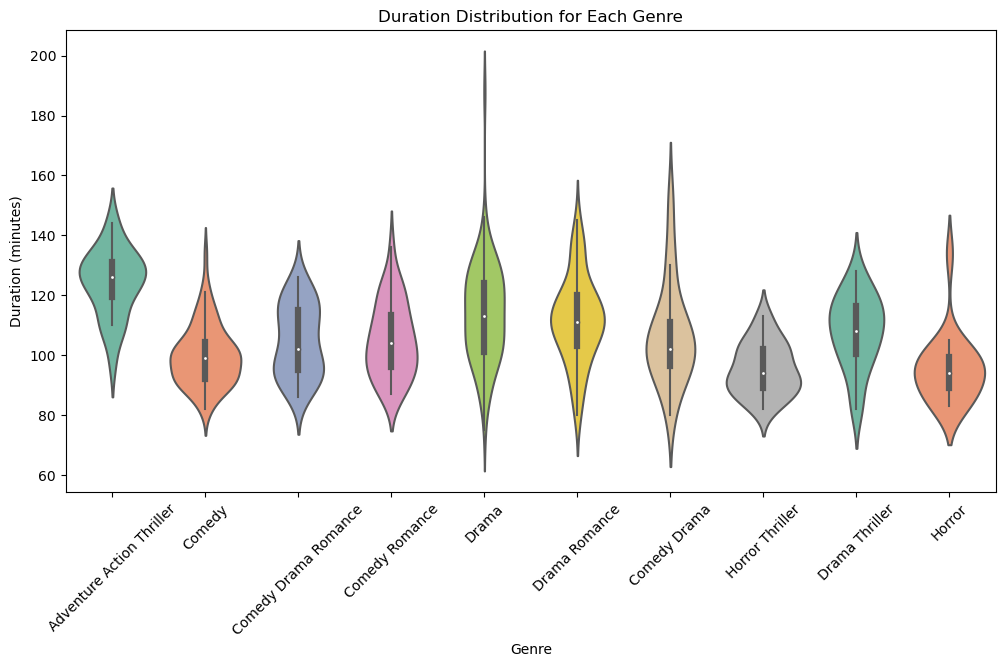
plt**.**title('Duration Distribution for Each Genre')

plt**.**xlabel('Genre')

plt**.**ylabel('Duration (minutes)')

plt**.**xticks(rotation**=**45)

plt**.**show()

****

***# 17. What is the relationship between budget and revenue for the top 10 genres?***

plt**.**figure(figsize**=**(12, 6))

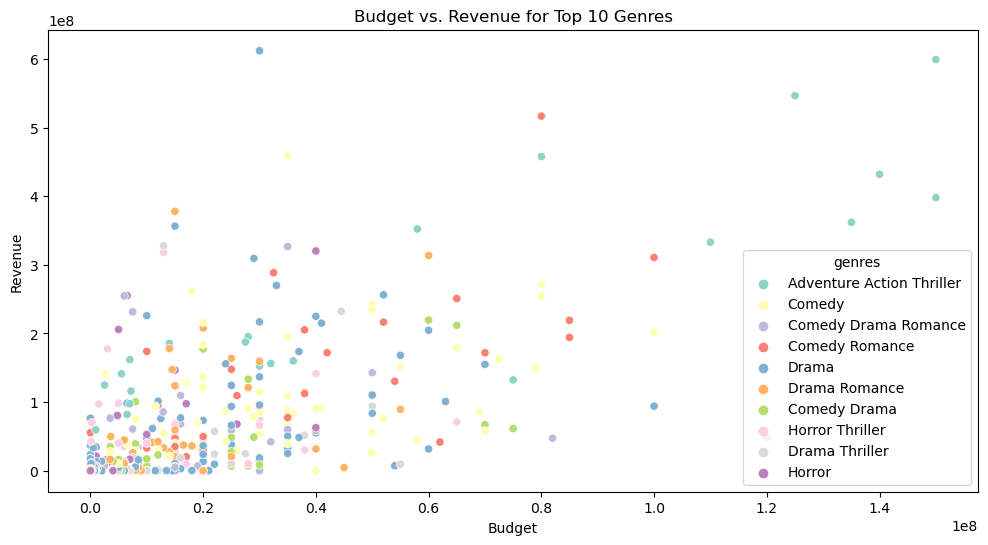
sns**.**scatterplot(x**=**'budget', y**=**'revenue', data**=**filtered\_data, hue**=**'genres', palette**=**'Set3')

plt**.**title('Budget vs. Revenue for Top 10 Genres')

plt**.**xlabel('Budget')

plt**.**ylabel('Revenue')

plt**.**show()

****

***# 18. What are top 20 the most common movie genres?***

top\_20\_genres **=** df['genres']**.**value\_counts()**.**head(20)**.**index

plt**.**figure(figsize**=**(10, 5))

sns**.**countplot(x**=**'genres', data**=**df, order**=**top\_20\_genres, palette**=**'viridis')

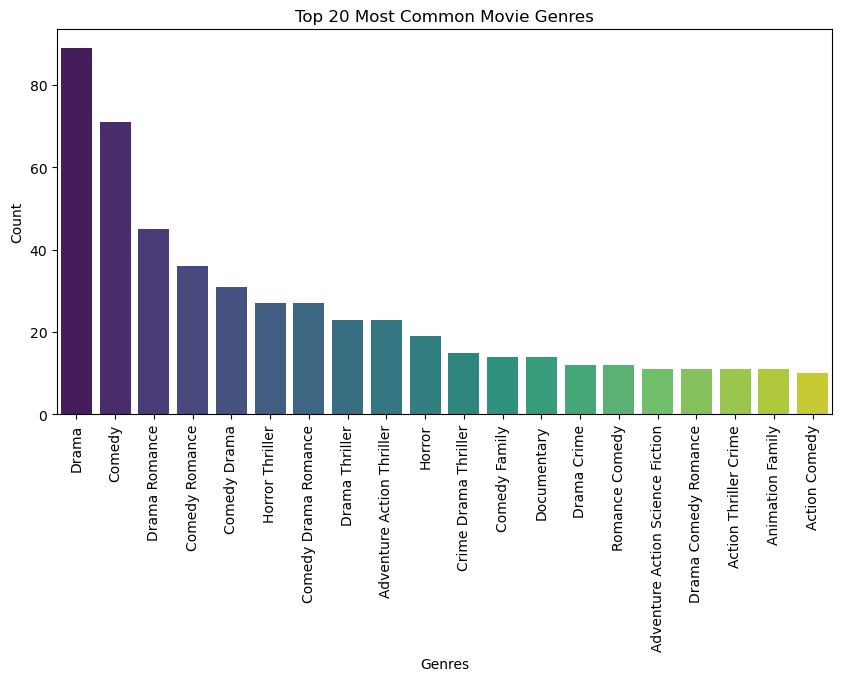
plt**.**title('Top 20 Most Common Movie Genres')

plt**.**xlabel('Genres')

plt**.**ylabel('Count')

plt**.**xticks(rotation**=**90)

plt**.**show()

****

***# 18. How does the movie rating vary by director's experience (number of movies directed)?***

director\_experience **=** df**.**groupby('director')['id']**.**count()**.**reset\_index()

director\_experience**.**columns **=** ['director', 'num\_movies']

movie\_data\_with\_experience **=** pd**.**merge(filtered\_data, director\_experience, on**=**'director', how**=**'left')

plt**.**figure(figsize**=**(12, 6))

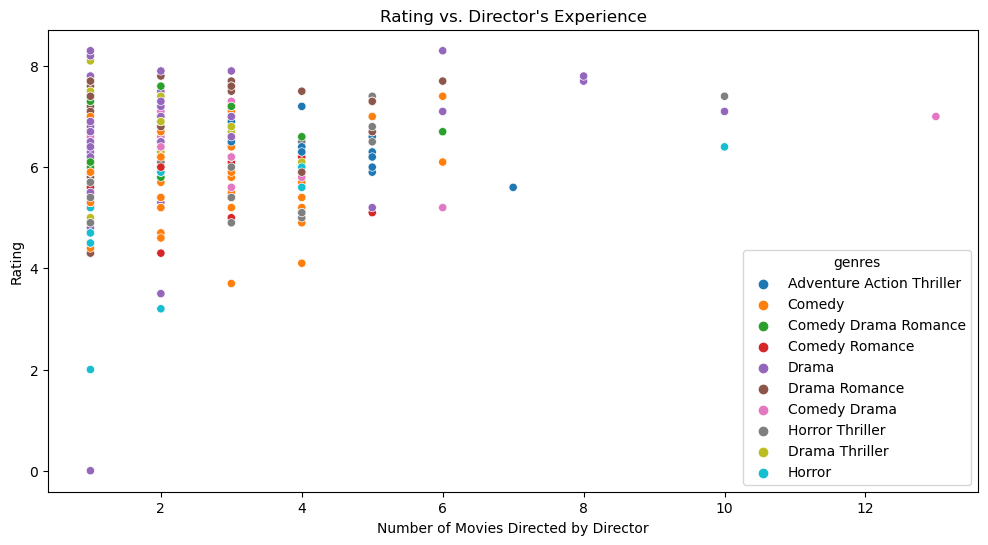
sns**.**scatterplot(x**=**'num\_movies', y**=**'vote\_average', data**=**movie\_data\_with\_experience, hue**=**'genres', palette**=**'tab10')

plt**.**title('Rating vs. Director\'s Experience')

plt**.**xlabel('Number of Movies Directed by Director')

plt**.**ylabel('Rating')

plt**.**show()

****

*#* ***19. What does the correlation heatmap of the movie dataset reveal about the relationships between different numerical attributes?***

numerical\_columns **=**df**.**select\_dtypes(include**=**['int', 'float'])

*# Create a correlation matrix*

correlation\_matrix **=** numerical\_columns**.**corr()

*# Set the size of the heatmap*

plt**.**figure(figsize**=**(10, 8))

*# Create the heatmap*

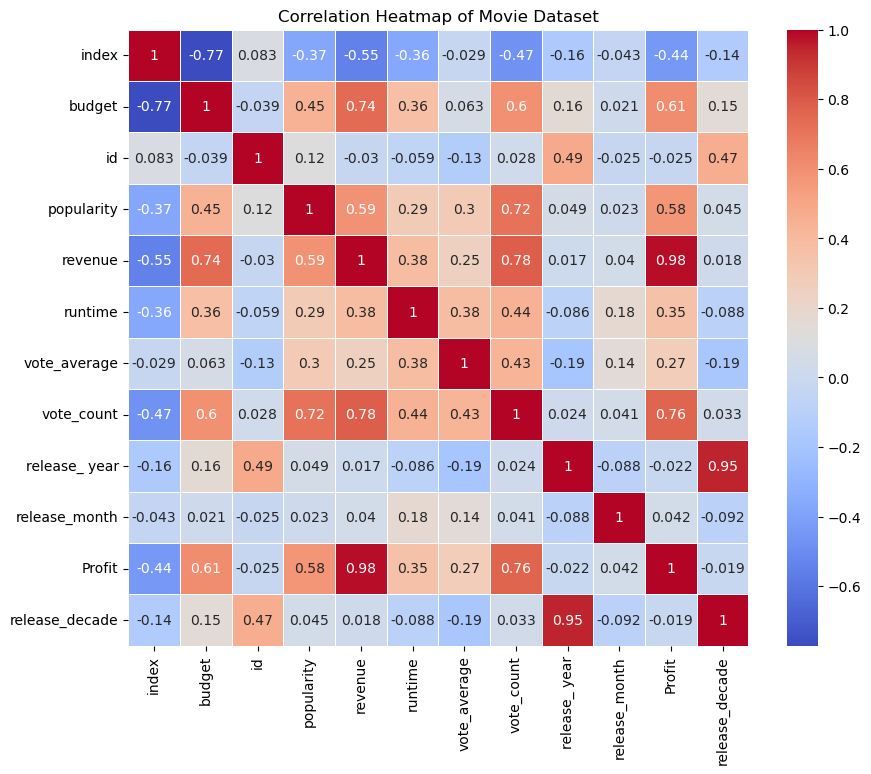
sns**.**heatmap(correlation\_matrix, annot**=True**, cmap**=**'coolwarm', linewidths**=**.5)

*# Set the title*

plt**.**title('Correlation Heatmap of Movie Dataset')

*# Display the heatmap*

plt**.**show()

****

***# 20. Displaying different movie trends in budget, revenue, popularity, runtime, and user ratings in the movie dataset over time by using line plots***

custom\_palette **=** ['royalblue', 'forestgreen', 'tomato', 'gold', 'purple']

fig, axes **=** plt**.**subplots(5, 1, figsize**=**(12, 20))

*# Plot 1: Budget Trends Over Time*

sns**.**lineplot(x**=**'release\_ year', y**=**'budget', data**=**df, ax**=**axes[0], color**=**custom\_palette[0])

axes[0]**.**set\_title('Average Movie Budget Over Time')

axes[0]**.**set\_xlabel('Release Year')

axes[0]**.**set\_ylabel('Budget')

*# Plot 2: Revenue Trends Over Time*

sns**.**lineplot(x**=**'release\_ year', y**=**'revenue', data**=**df, ax**=**axes[1], color**=**custom\_palette[1])

axes[1]**.**set\_title('Average Movie Revenue Over Time')

axes[1]**.**set\_xlabel('Release Year')

axes[1]**.**set\_ylabel('Revenue')

*# Plot 3: Popularity Trends Over Time*

sns**.**lineplot(x**=**'release\_ year', y**=**'popularity', data**=**df, ax**=**axes[2], color**=**custom\_palette[2])

axes[2]**.**set\_title('Movie Popularity Over Time')

axes[2]**.**set\_xlabel('Release Year')

axes[2]**.**set\_ylabel('Popularity')

*# Plot 4: Runtime Trends Over Time*

sns**.**lineplot(x**=**'release\_ year', y**=**'runtime', data**=**df, ax**=**axes[3], color**=**custom\_palette[3])

axes[3]**.**set\_title('Average Movie Runtime Over Time')

axes[3]**.**set\_xlabel('Release Year')

axes[3]**.**set\_ylabel('Runtime (minutes)')

*# Plot 5: User Rating Trends Over Time*

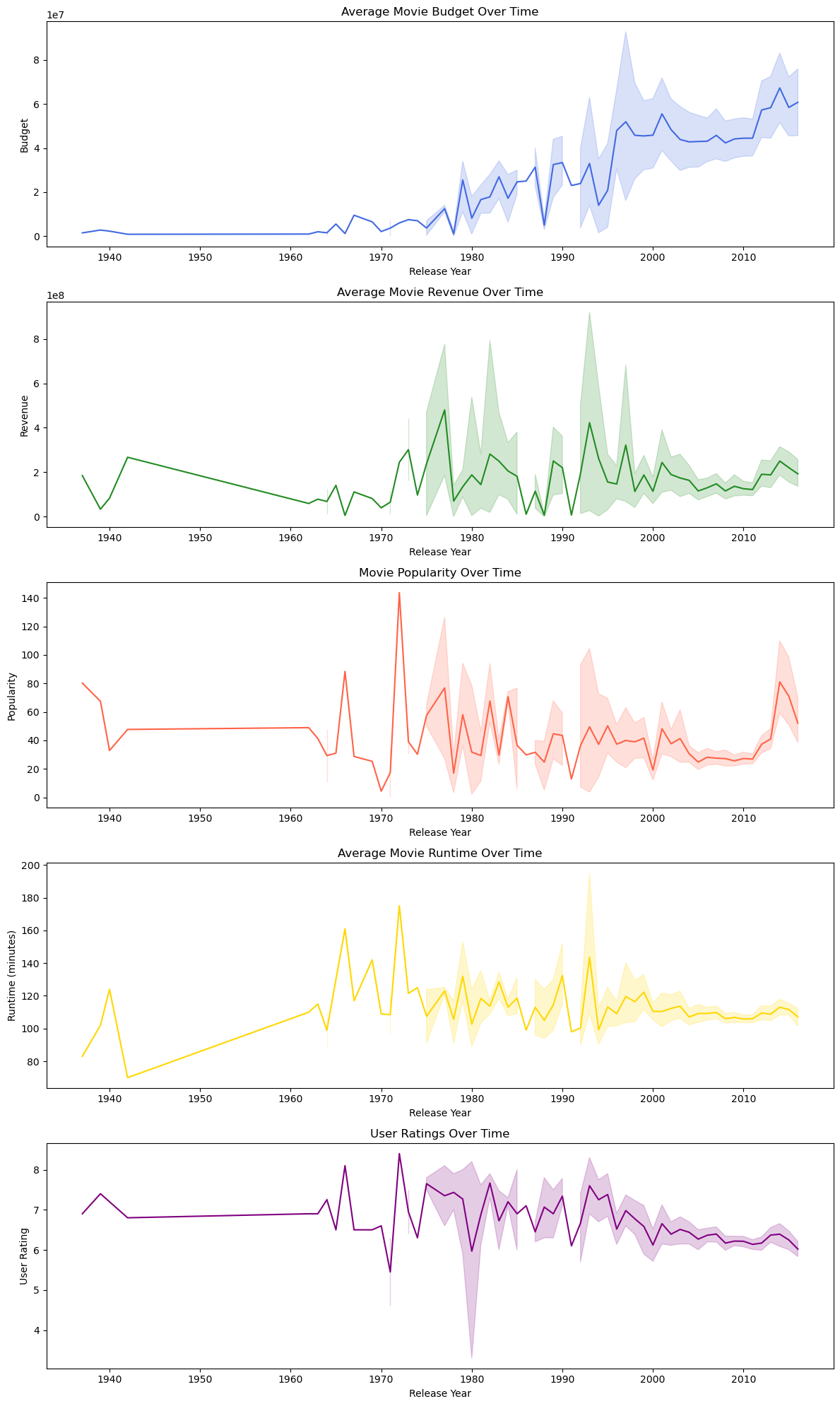
sns**.**lineplot(x**=**'release\_ year', y**=**'vote\_average', data**=**df, ax**=**axes[4], color**=**custom\_palette[4])

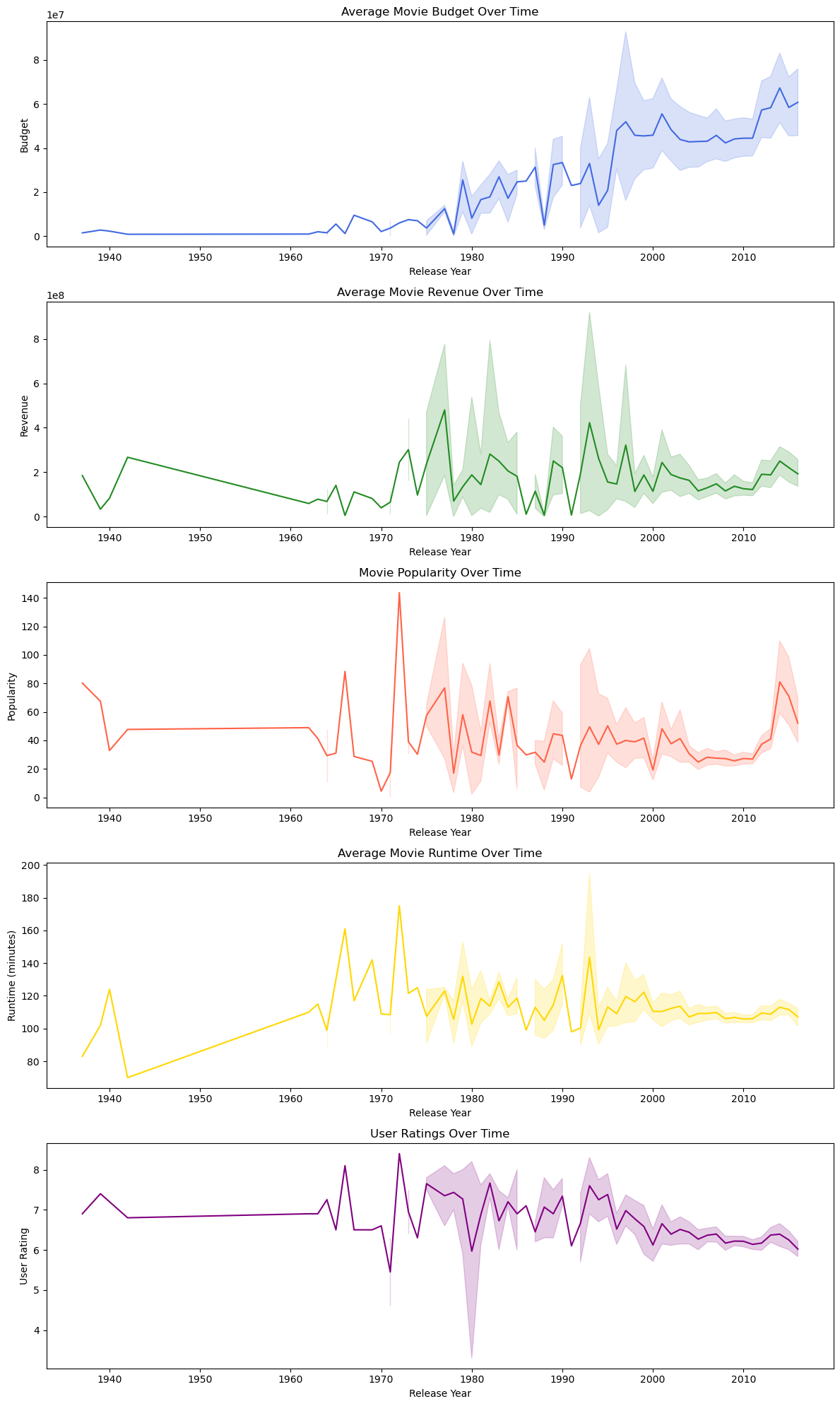
axes[4]**.**set\_title('User Ratings Over Time')

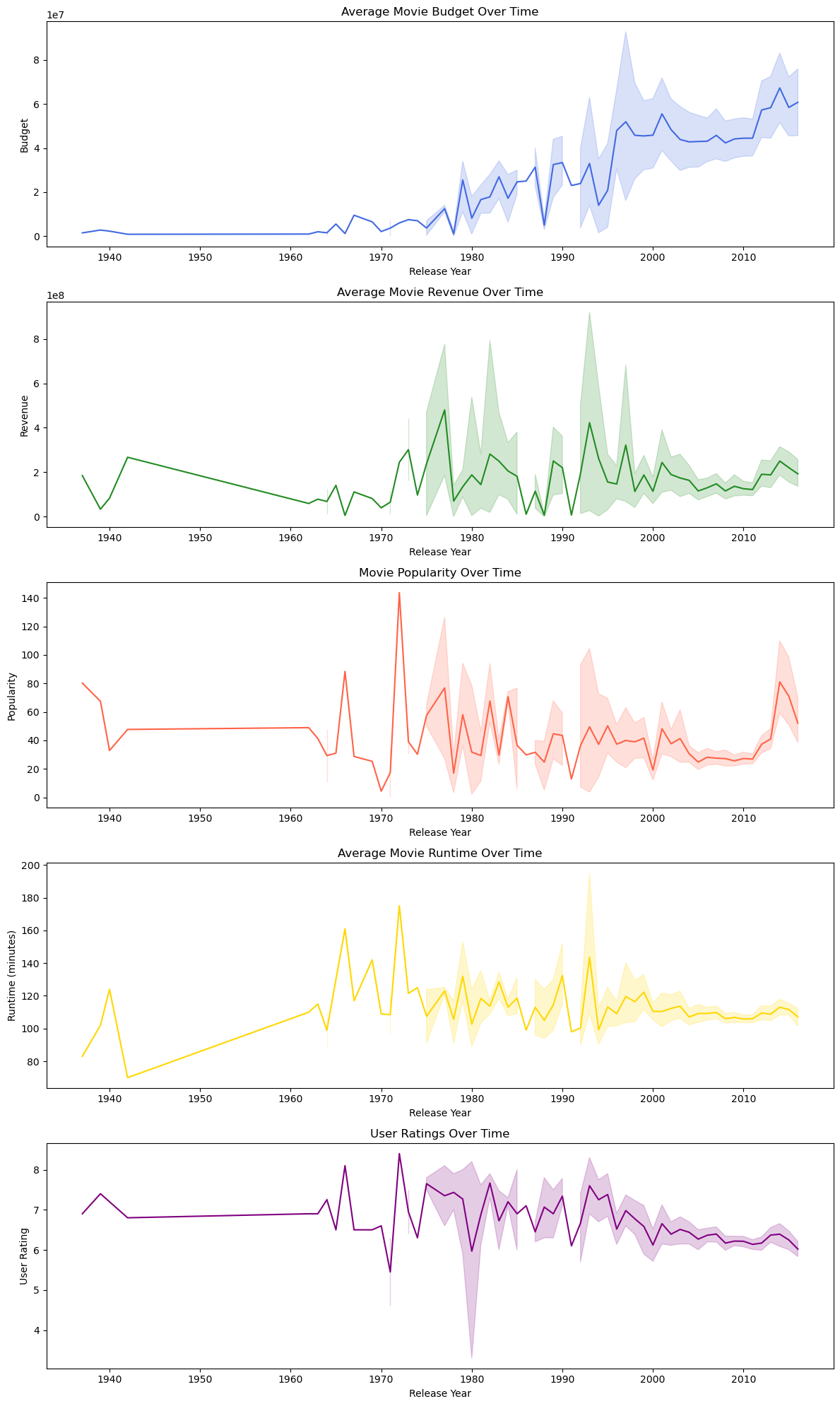
axes[4]**.**set\_xlabel('Release Year')

axes[4]**.**set\_ylabel('User Rating')

plt**.**tight\_layout()

plt**.**show()****

****

****

**# Most Frequent Taglines over decades**

df['release\_decade'] **=** (df['release\_date']**.**dt**.**year **//** 10) **\*** 10

top\_taglines **=** df**.**groupby('release\_decade')['tagline']**.**apply(**lambda** x: x**.**value\_counts()**.**idxmax())

plt**.**figure(figsize**=**(12, 6))

sns**.**countplot(x**=**'release\_decade', hue**=**'tagline', data**=**df, palette**=**'viridis')

plt**.**title('Most Frequent Taglines by Decade')

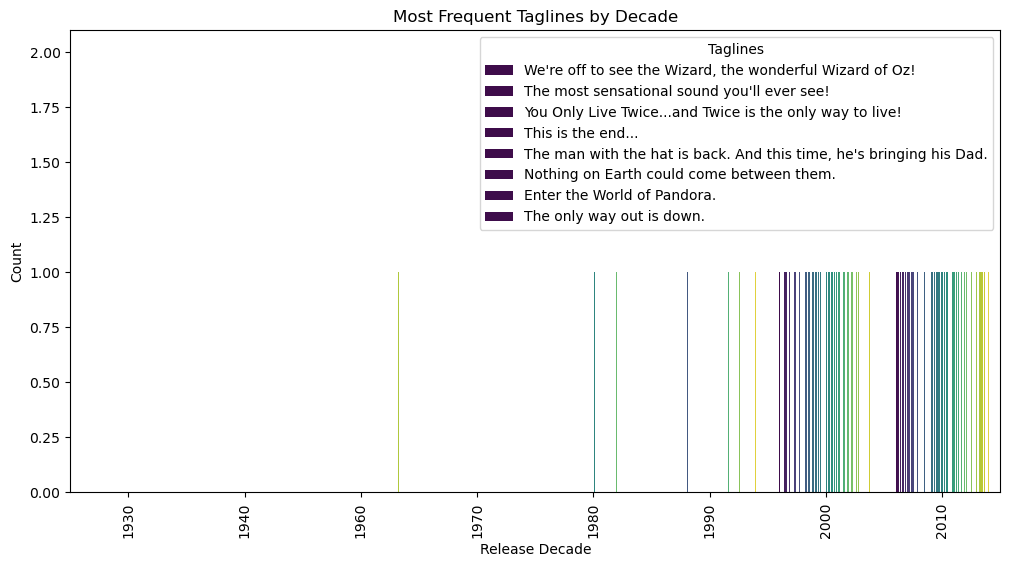
plt**.**xlabel('Release Decade')

plt**.**ylabel('Count')

plt**.**xticks(rotation**=**90)

plt**.**legend(title**=**'Taglines', loc**=**'upper right', labels**=**top\_taglines)

plt**.**show()



In [81]:

In [82]:

**# Who are the top 10 most frequently appearing cast members in the dataset?**

plt**.**figure(figsize**=**(12, 6))

top\_10\_cast **=** df['cast']**.**str**.**split('|')**.**explode()**.**value\_counts()**.**head(7)

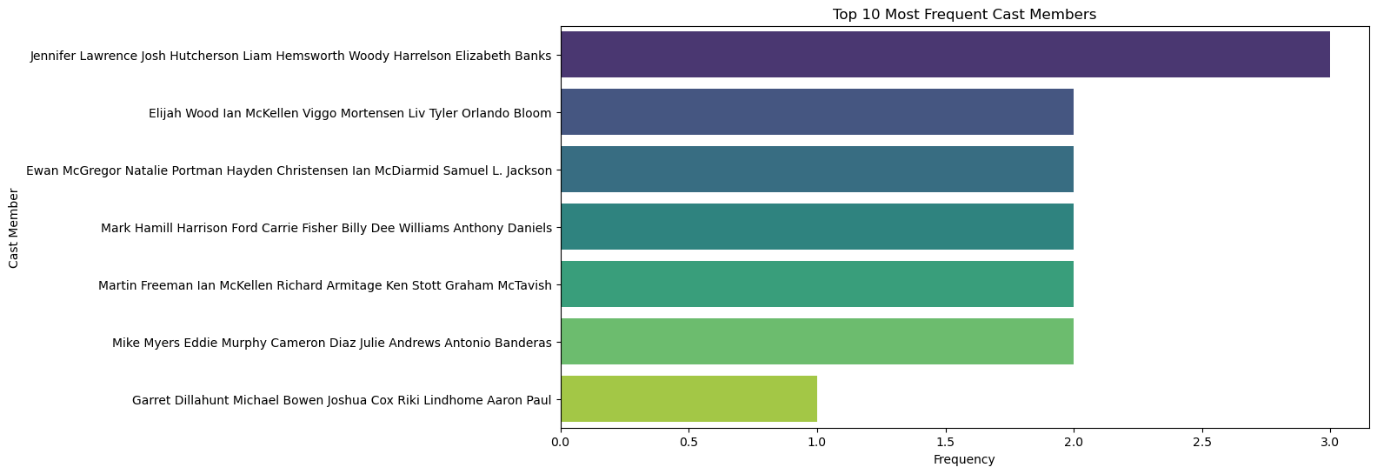
sns**.**barplot(x**=**top\_10\_cast**.**values, y**=**top\_10\_cast**.**index, palette**=**'viridis')

plt**.**title('Top 10 Most Frequent Cast Members')

plt**.**xlabel('Frequency')

plt**.**ylabel('Cast Member')

plt**.**show()



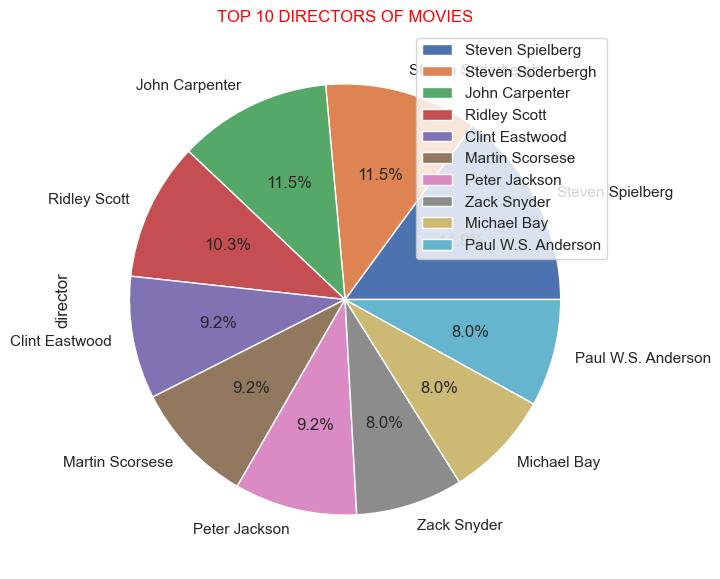
**# Top 10 directors who produced most of movies in between 1940 -2016**

df**.**director**.**value\_counts()[:10]**.**plot**.**pie(autopct**=**'%1.1f%%',figsize**=**(7,7))

plt**.**title('TOP 10 DIRECTORS OF MOVIES',color**=**'Red')

plt**.**legend()

plt**.**show()



In [89]:

**Conclusion:**

The Movie Dataset Analysis project has been a valuable learning experience, emphasizing the power of data-driven decision-making. It serves as a guide for future endeavours, whether in data analysis or other pursuits.

I have presented the findings and experiences gained from the Movie Dataset Analysis project.This project aimed to explore the movie industry's dynamics through data analysis and draw actionable insights for stakeholders.

By understanding industry trends, optimize financial performance, and enhance audience engagement. We used data exploration and data visualization to create meaningful insights.

I hadlearned analytical skills and enhanced ability to translate data into insights. I also developed communication of complex findings.

I am grateful for this opportunity to explore the world of data analysis and look forward to applying these skills in future projects.

**Reference:**

* **Kaggle**

[**www.kaggle.com**](http://www.kaggle.com)

* **YouTube**

[**www.youtube.com**](http://www.youtube.com)