Movie Popularity and Rating Analysis

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

# Load dataset
file_path = "movies.csv"
df = pd.read_csv(file_path)

# Convert release_date to datetime
df['release_date'] = pd.to_datetime(df['release_date'],
errors='coerce')
df['release_year'] = df['release_date'].dt.year

# Set style
sns.set_style("whitegrid")
```

1. Distribution of vote averages

This histogram visualizes the distribution of movie vote averages. The x-axis represents vote average scores, while the y-axis shows the count of movies within each bin. A KDE curve is overlaid to show the density of scores.

```python
plt.figure(figsize=(10, 5))
sns.histplot(df['vote_average'], bins=30, kde=True)
plt.title('Distribution of Vote Averages')
plt.xlabel('Vote Average')
plt.ylabel('Count')
plt.show()
```
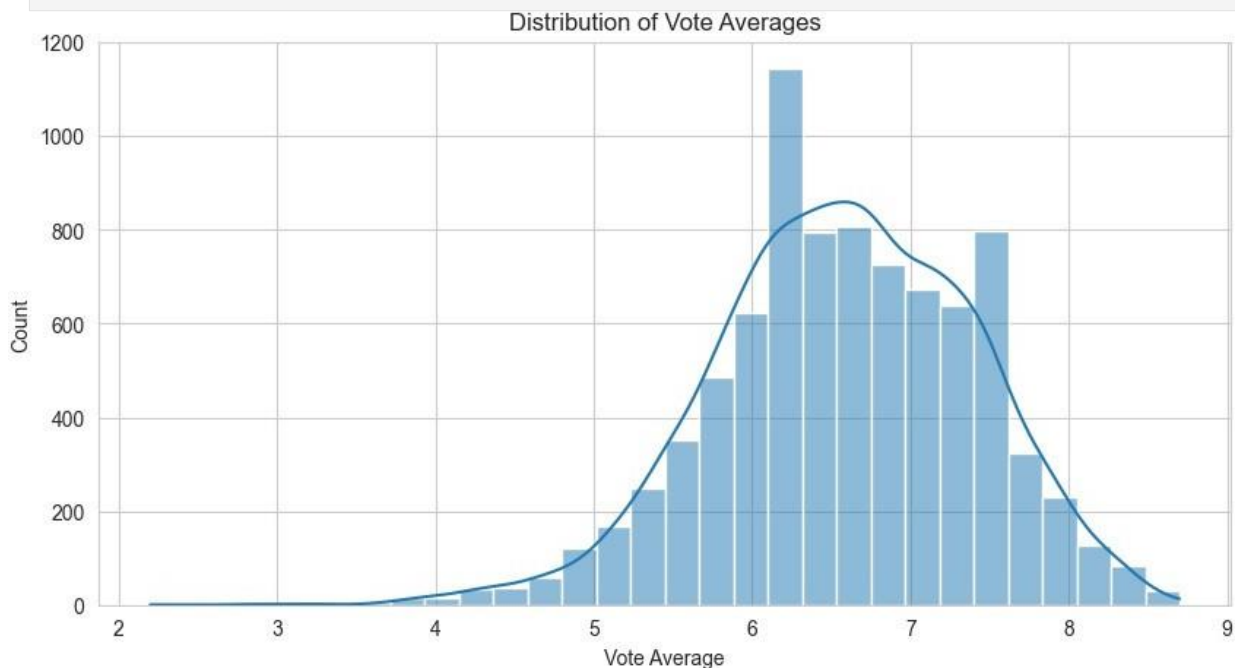
2. Number of movies released per year

This line graph illustrates the trend of movie releases over time. The x-axis represents the release years, and the y-axis shows the number of movies released each year. It helps identify patterns or spikes in film production.

```
plt.figure(figsize=(12, 5))
df['release_year'].value_counts().sort_index().plot(kind='line')
plt.title('Number of Movies Released Per Year')
plt.xlabel('Year')
plt.ylabel('Number of Movies')
plt.show()
```
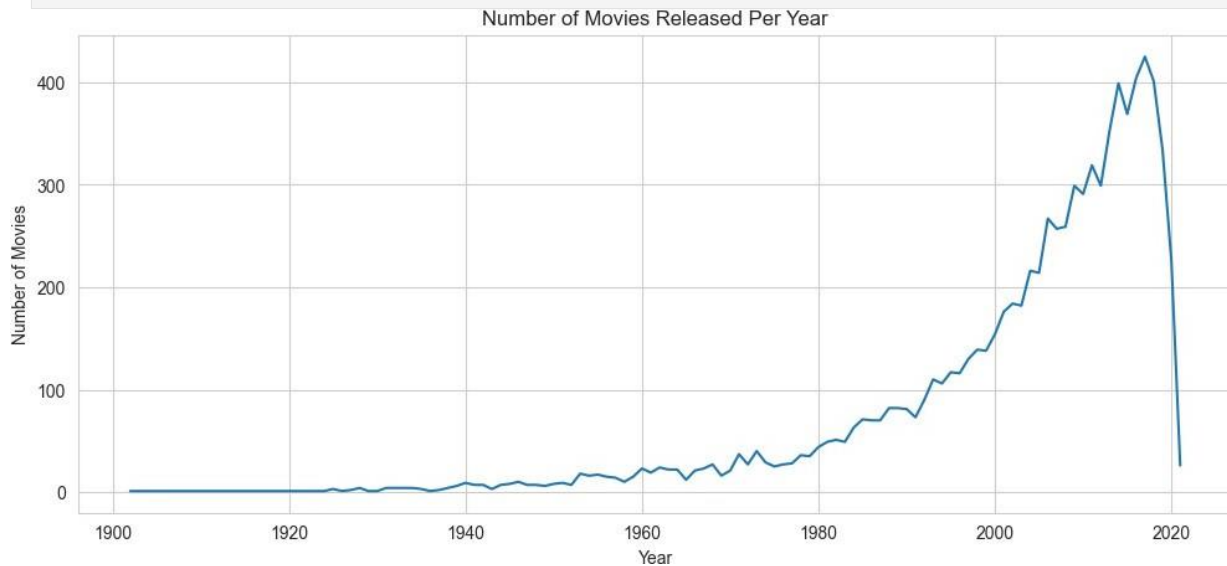
3. Popularity vs Vote Average

This scatter plot examines the relationship between vote average and popularity. Each point represents a movie, with vote averages on the x-axis and popularity on the y-axis. The alpha value is used to prevent overlapping and improve visibility.

```python
plt.figure(figsize=(10, 5))
sns.scatterplot(data=df, x='vote_average', y='popularity', alpha=0.5)
plt.title('Popularity vs. Vote Average')
plt.xlabel('Vote Average')
plt.ylabel('Popularity')
plt.show()
```
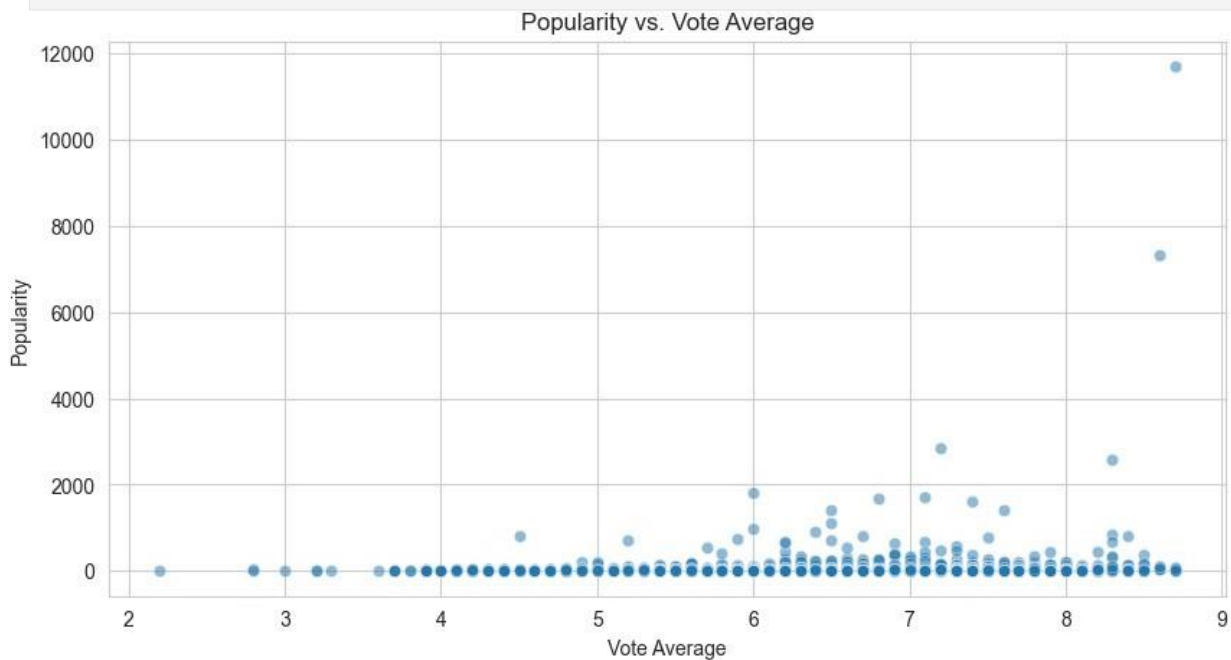
4. Top 10 highest-rated movies

This horizontal bar chart highlights the top 10 highest-rated movies based on vote averages. The x-axis represents vote averages, and the y-axis lists the movie titles. The viridis color palette enhances visual appeal.

```
top_movies = df.nlargest(10, 'vote_average')
plt.figure(figsize=(12, 5))
sns.barplot(data=top_movies, x='vote_average', y='title',
palette='viridis')
plt.title('Top 10 Highest Rated Movies')
plt.xlabel('Vote Average')
plt.ylabel('Movie Title')
plt.show()

C:\Users\nikhil\AppData\Local\Temp\ipykernel_1504\3518270814.py:3:
FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `y` variable to `hue` and set
`legend=False` for the same effect.

  sns.barplot(data=top_movies, x='vote_average', y='title',
palette='viridis')
```
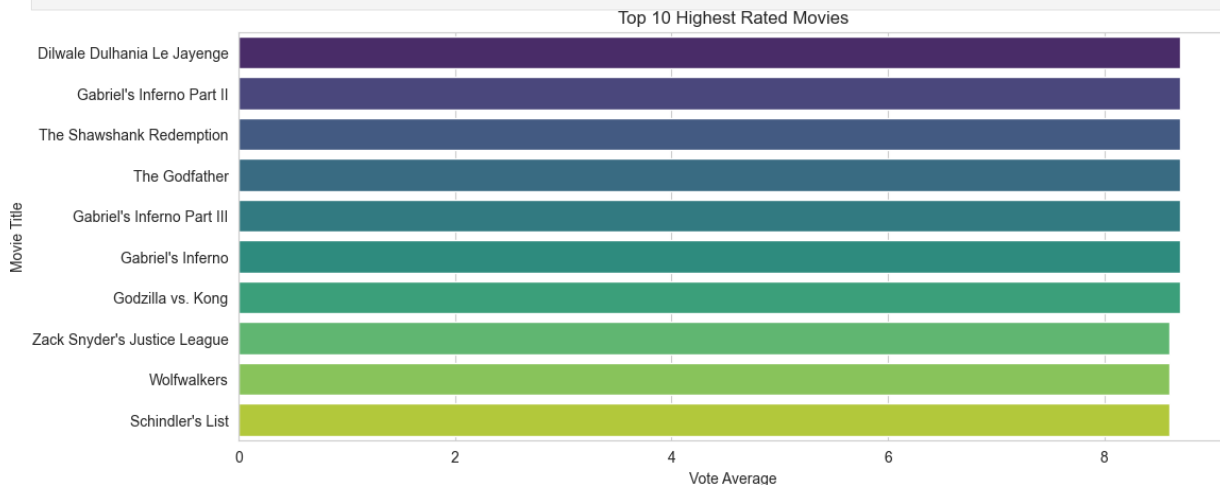
5. Correlation heatmap of numerical features

This heatmap displays the correlation between key numerical variables such as popularity, vote average, and vote count. The values indicate the strength and direction of the relationships, with colors representing positive (red) and negative (blue) correlations.

```
plt.figure(figsize=(8, 6))
sns.heatmap(df[['popularity', 'vote_average', 'vote_count']].corr(),
annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()
```
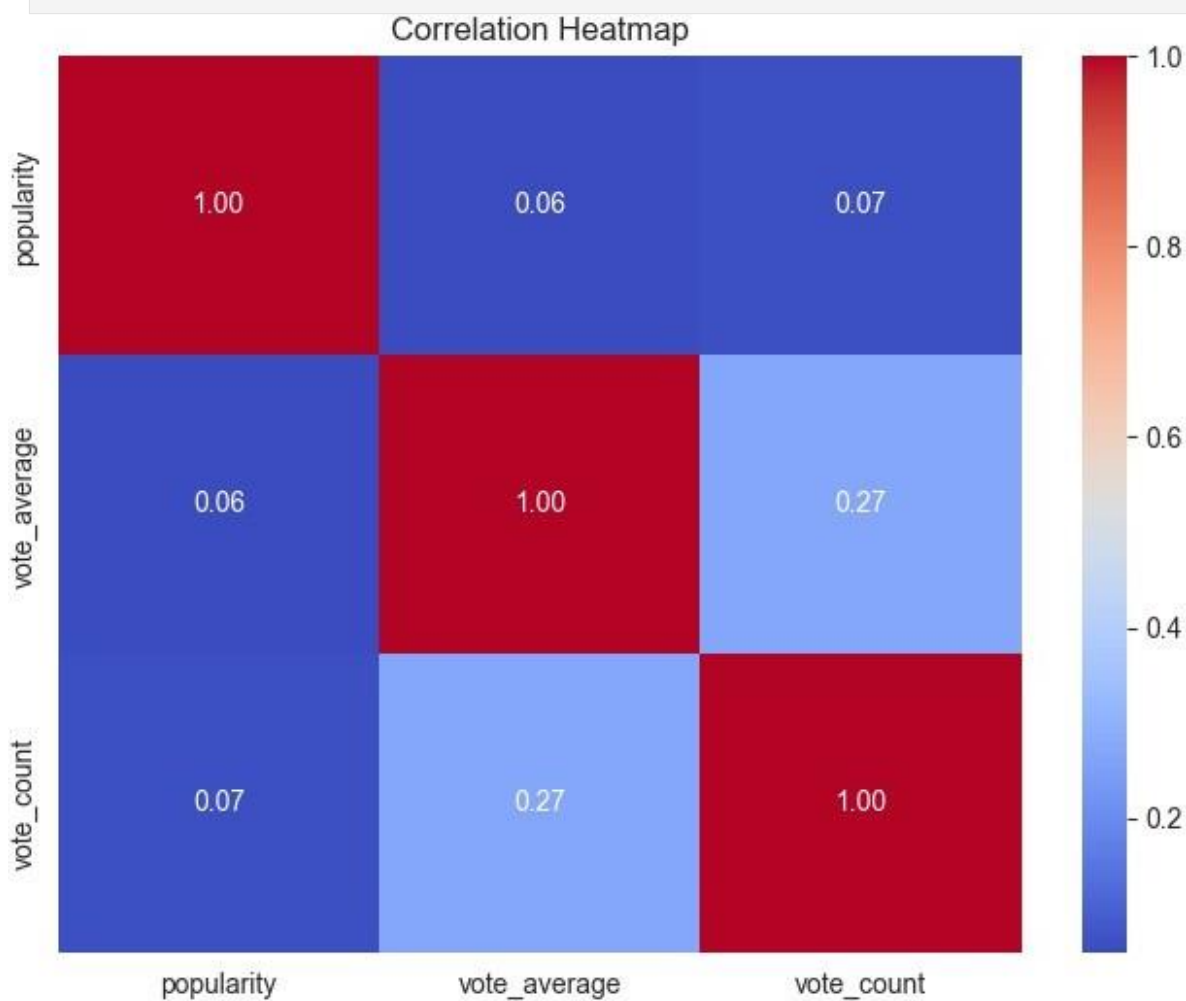
6. Popularity trend over time

This line graph shows how the average popularity of movies has changed over time. The x-axis represents the release year, and the y-axis shows the average popularity score. It helps identify trends in audience interest.

```
plt.figure(figsize=(12, 5))
df.groupby('release_year')['popularity'].mean().plot()
plt.title('Popularity Trend Over Time')
plt.xlabel('Year')
plt.ylabel('Average Popularity')
plt.show()
```
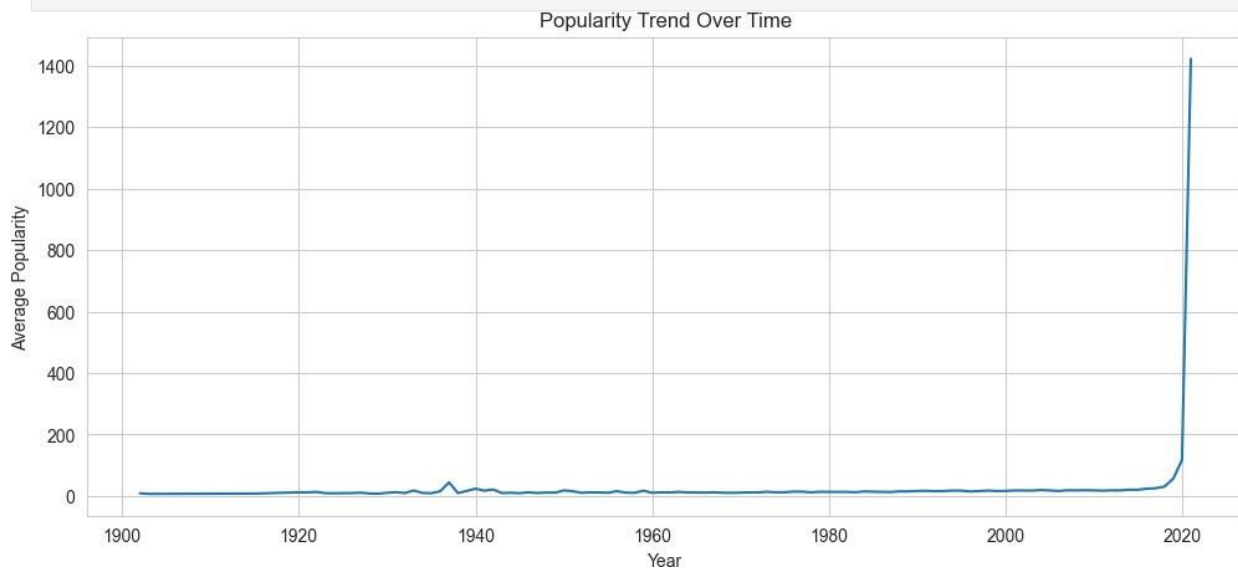
7. Vote count distribution

This box plot visualizes the distribution of vote counts for movies. The x-axis represents vote counts, and the plot highlights the spread, median, and potential outliers.

```python
plt.figure(figsize=(10, 5))
sns.boxplot(x=df['vote_count'])
plt.title('Vote Count Distribution')
plt.xlabel('Vote Count')
plt.show()
```