ALLIAN UNIVERSI

Private University Estd. in Karnataka State by Act No. 34

Project Report

Introduction to Data Science

Semester – 2

“Collection of IMDb top 5000 movies data”

GITHUB LINK : piyushT3003/IDS

By

Piyush Tandale

Reg no: 2411021240006

Department of Computer Application

Alliance University Chandapura — Anekal Main Road,

Anekal Bengaluru — 562 106

April 2025

Project Overview

This project analyzes a dataset of popular movies from IMDb. The data includes movie metadata such as title, release year, rating, votes, genre, runtime, and associated crew members (directors and writers). Through data visualization and statistical exploration, the project seeks to uncover patterns and trends in movie popularity and quality.

Introduction:

The dataset provided appears to be a collection of IMDb movie data, containing information about popular movies. Each row represents a movie and includes the following attributes:

- tconst: Unique identifier for the movie (IMDb ID).
- primaryTitle: The title of the movie.
- startYear: The year the movie was released.
- rank: The ranking of the movie based on its popularity or rating.
- averageRating: The average IMDb rating of the movie.
- numVotes: The number of votes the movie has received on IMDb.
- runtimeMinutes: The runtime of the movie in minutes.
- directors: The director(s) of the movie.
- writers: The writer(s) of the movie.
- genres: The genres associated with the movie (e.g., Drama, Action).
- IMDbLink: A clickable link to the movie's IMDb page. Title_IMDb_Link: Another clickable link to the movie's IMDb page with the title.

Project Goals

1. **Understand Key Attributes – Gain insights into movie characteristics that correlate with higher IMDb ratings.**
2. **Visualize Top Performers – Identify and visualize the top 10 highest-rated movies.**
3. **Explore Genre Trends – Examine the distribution of movie genres to identify the most common types.**
4. **Runtime Analysis – Investigate the distribution of movie lengths and assess if runtime has any relationship with ratings or genre.**
5. **Use Visualizations – Employ bar plots, pie charts, and histograms for a more intuitive understanding of the data.**

Challenges

- **Missing or Incomplete Data: Some entries may have missing values (especially in runtime or crew information), which could impact analysis accuracy.**
- **Genre Complexity: Movies often belong to multiple genres, but they may be listed as comma-separated strings, making precise categorization difficult.**
- **Outliers: Extremely high or low values in runtime or ratings might skew results and require careful handling.**
- **Data Cleaning: Ensuring the dataset is clean and well-structured for analysis took effort, particularly due to formatting in fields like directors, genres, or IMDb links.**

```
import pandas as pd

df=pd.read_csv(r"C:\Users\piyus\Desktop\ids project\results_with_crew.csv")

df
```

```
         tconst                                    primaryTitle  startYear  \
0      tt0111161                     The Shawshank Redemption       1994
1      tt0068646                                The Godfather       1972
2      tt0468569                              The Dark Knight       2008
3      tt0167260  The Lord of the Rings: The Return of the King       2003
4      tt0108052                              Schindler's List       1993
...          ...                                          ...        ...
4995   tt0880578                                Untraceable       2008
4996  tt27459160             Teri Baaton Mein Aisa Uljha Jiya       2024
4997   tt3174376                                Before I Wake       2016
4998   tt5177088                The Girl in the Spider's Web       2018
4999   tt8368408                           Gunpowder Milkshake       2021

      rank  averageRating  numVotes  runtimeMinutes                directors
\
0        1            9.3   3029407             142           Frank Darabont
1        2            9.2   2114342             175     Francis Ford Coppola
2        3            9.0   3005759             152        Christopher Nolan
3        4            9.0   2068876             201            Peter Jackson
4        5            9.0   1515877             195         Steven Spielberg
...    ...            ...       ...             ...                      ...
4995  4996            6.2     54012             101           Gregory Hoblit
4996  4997            6.2     53979             141  Amit Joshi, Aradhana Sah
4997  4998            6.2     53934              97             Mike Flanagan
4998  4999            6.1     53870             115              Fede Alvarez
```

```
4999  5000            6.1     53710             114           Navot Papushado

                                                  writers  \
0                           Stephen King, Frank Darabont
1                         Mario Puzo, Francis Ford Coppola
2     Jonathan Nolan, Christopher Nolan, David S. Go...
3     J.R.R. Tolkien, Fran Walsh, Philippa Boyens, P...
4                         Thomas Keneally, Steven Zaillian
...                                                   ...
4995     Robert Fyvolent, Mark Brinker, Allison Burnett
4996                             Amit Joshi, Aradhana Sah
4997                           Mike Flanagan, Jeff Howard
4998  Jay Basu, Fede Alvarez, Steven Knight, Stieg L...
4999                          Navot Papushado, Ehud Lavski

                          genres  \
0                          Drama
1                  Crime, Drama
2           Action, Crime, Drama
3     Adventure, Drama, Fantasy
4     Biography, Drama, History
...                         ...
4995   Crime, Mystery, Thriller
4996     Comedy, Drama, Romance
4997     Drama, Fantasy, Horror
4998   Action, Adventure, Crime
4999    Action, Crime, Thriller

                                                 IMDbLink  \
0     <a href="https://www.imdb.com/title/tt0111161"...
1     <a href="https://www.imdb.com/title/tt0068646"...
2     <a href="https://www.imdb.com/title/tt0468569"...
3     <a href="https://www.imdb.com/title/tt0167260"...
4     <a href="https://www.imdb.com/title/tt0108052"...
...                                                 ...
4995  <a href="https://www.imdb.com/title/tt0880578"...
4996  <a href="https://www.imdb.com/title/tt27459160...
4997  <a href="https://www.imdb.com/title/tt3174376"...
4998  <a href="https://www.imdb.com/title/tt5177088"...
4999  <a href="https://www.imdb.com/title/tt8368408"...

                                          Title_IMDb_Link
0     <a href="https://www.imdb.com/title/tt0111161"...
1     <a href="https://www.imdb.com/title/tt0068646"...
2     <a href="https://www.imdb.com/title/tt0468569"...
3     <a href="https://www.imdb.com/title/tt0167260"...
4     <a href="https://www.imdb.com/title/tt0108052"...
...                                                 ...
4995  <a href="https://www.imdb.com/title/tt0880578"...
```

```
4996  <a href="https://www.imdb.com/title/tt27459160...
4997  <a href="https://www.imdb.com/title/tt3174376"...
4998  <a href="https://www.imdb.com/title/tt5177088"...
4999  <a href="https://www.imdb.com/title/tt8368408"...

[5000 rows x 12 columns]
```

```
df.head()
```

```
      tconst                                     primaryTitle  startYear  rank
\
0  tt0111161                       The Shawshank Redemption       1994     1
1  tt0068646                                  The Godfather       1972     2
2  tt0468569                                The Dark Knight       2008     3
3  tt0167260  The Lord of the Rings: The Return of the King       2003     4
4  tt0108052                                Schindler's List       1993     5

   averageRating  numVotes  runtimeMinutes             directors  \
0            9.3   3029407             142        Frank Darabont
1            9.2   2114342             175  Francis Ford Coppola
2            9.0   3005759             152    Christopher Nolan
3            9.0   2068876             201         Peter Jackson
4            9.0   1515877             195      Steven Spielberg

                                        writers  \
0                    Stephen King, Frank Darabont
1                  Mario Puzo, Francis Ford Coppola
2  Jonathan Nolan, Christopher Nolan, David S. Go...
3  J.R.R. Tolkien, Fran Walsh, Philippa Boyens, P...
4                   Thomas Keneally, Steven Zaillian

                      genres  \
0                      Drama
1               Crime, Drama
2       Action, Crime, Drama
3  Adventure, Drama, Fantasy
4  Biography, Drama, History

                                             IMDbLink  \
0  <a href="https://www.imdb.com/title/tt0111161"...
1  <a href="https://www.imdb.com/title/tt0068646"...
2  <a href="https://www.imdb.com/title/tt0468569"...
3  <a href="https://www.imdb.com/title/tt0167260"...
4  <a href="https://www.imdb.com/title/tt0108052"...

                                    Title_IMDb_Link
0  <a href="https://www.imdb.com/title/tt0111161"...
1  <a href="https://www.imdb.com/title/tt0068646"...
2  <a href="https://www.imdb.com/title/tt0468569"...
```

```
3  <a href="https://www.imdb.com/title/tt0167260"...
4  <a href="https://www.imdb.com/title/tt0108052"...
```

```
df.tail()
```

```
          tconst                      primaryTitle  startYear  rank  \
4995   tt0880578                       Untraceable       2008  4996
4996  tt27459160  Teri Baaton Mein Aisa Uljha Jiya       2024  4997
4997   tt3174376                     Before I Wake       2016  4998
4998   tt5177088      The Girl in the Spider's Web       2018  4999
4999   tt8368408               Gunpowder Milkshake       2021  5000

      averageRating  numVotes  runtimeMinutes                directors  \
4995            6.2     54012             101            Gregory Hoblit
4996            6.2     53979             141  Amit Joshi, Aradhana Sah
4997            6.2     53934              97             Mike Flanagan
4998            6.1     53870             115              Fede Alvarez
4999            6.1     53710             114           Navot Papushado

                                          writers  \
4995    Robert Fyvolent, Mark Brinker, Allison Burnett
4996                          Amit Joshi, Aradhana Sah
4997                        Mike Flanagan, Jeff Howard
4998  Jay Basu, Fede Alvarez, Steven Knight, Stieg L...
4999                      Navot Papushado, Ehud Lavski

                      genres  \
4995  Crime, Mystery, Thriller
4996    Comedy, Drama, Romance
4997    Drama, Fantasy, Horror
4998  Action, Adventure, Crime
4999   Action, Crime, Thriller

                                          IMDbLink  \
4995  <a href="https://www.imdb.com/title/tt0880578"...
4996  <a href="https://www.imdb.com/title/tt27459160...
4997  <a href="https://www.imdb.com/title/tt3174376"...
4998  <a href="https://www.imdb.com/title/tt5177088"...
4999  <a href="https://www.imdb.com/title/tt8368408"...

                                    Title_IMDb_Link
4995  <a href="https://www.imdb.com/title/tt0880578"...
4996  <a href="https://www.imdb.com/title/tt27459160...
4997  <a href="https://www.imdb.com/title/tt3174376"...
4998  <a href="https://www.imdb.com/title/tt5177088"...
4999  <a href="https://www.imdb.com/title/tt8368408"...
```

```
df.info
```

```
<bound method DataFrame.info of           tconst
primaryTitle  startYear  \
0      tt0111161                       The Shawshank Redemption       1994
1      tt0068646                                  The Godfather       1972
2      tt0468569                                The Dark Knight       2008
3      tt0167260  The Lord of the Rings: The Return of the King       2003
4      tt0108052                                  Schindler's List       1993
...          ...                                                ...        ...
4995   tt0880578                                     Untraceable       2008
4996  tt27459160                Teri Baaton Mein Aisa Uljha Jiya       2024
4997   tt3174376                                   Before I Wake       2016
4998   tt5177088                      The Girl in the Spider's Web       2018
4999   tt8368408                                Gunpowder Milkshake       2021

      rank  averageRating  numVotes  runtimeMinutes                 directors
\
0        1            9.3   3029407             142            Frank Darabont
1        2            9.2   2114342             175      Francis Ford Coppola
2        3            9.0   3005759             152         Christopher Nolan
3        4            9.0   2068876             201             Peter Jackson
4        5            9.0   1515877             195          Steven Spielberg
...    ...            ...       ...             ...                       ...
4995  4996            6.2     54012             101            Gregory Hoblit
4996  4997            6.2     53979             141  Amit Joshi, Aradhana Sah
4997  4998            6.2     53934              97             Mike Flanagan
4998  4999            6.1     53870             115              Fede Alvarez
4999  5000            6.1     53710             114           Navot Papushado

                                            writers  \
0                        Stephen King, Frank Darabont
1                     Mario Puzo, Francis Ford Coppola
2     Jonathan Nolan, Christopher Nolan, David S. Go...
3     J.R.R. Tolkien, Fran Walsh, Philippa Boyens, P...
4                      Thomas Keneally, Steven Zaillian
...                                                 ...
4995     Robert Fyvolent, Mark Brinker, Allison Burnett
4996                           Amit Joshi, Aradhana Sah
4997                         Mike Flanagan, Jeff Howard
4998  Jay Basu, Fede Alvarez, Steven Knight, Stieg L...
4999                       Navot Papushado, Ehud Lavski

                         genres  \
0                         Drama
1                 Crime, Drama
2           Action, Crime, Drama
3     Adventure, Drama, Fantasy
4     Biography, Drama, History
...                         ...
4995   Crime, Mystery, Thriller
4996     Comedy, Drama, Romance
```

```
4997     Drama, Fantasy, Horror
4998   Action, Adventure, Crime
4999    Action, Crime, Thriller

                                                IMDbLink  \
0     <a href="https://www.imdb.com/title/tt0111161"...
1     <a href="https://www.imdb.com/title/tt0068646"...
2     <a href="https://www.imdb.com/title/tt0468569"...
3     <a href="https://www.imdb.com/title/tt0167260"...
4     <a href="https://www.imdb.com/title/tt0108052"...
...                                                 ...
4995  <a href="https://www.imdb.com/title/tt0880578"...
4996  <a href="https://www.imdb.com/title/tt27459160...
4997  <a href="https://www.imdb.com/title/tt3174376"...
4998  <a href="https://www.imdb.com/title/tt5177088"...
4999  <a href="https://www.imdb.com/title/tt8368408"...

                                      Title_IMDb_Link
0     <a href="https://www.imdb.com/title/tt0111161"...
1     <a href="https://www.imdb.com/title/tt0068646"...
2     <a href="https://www.imdb.com/title/tt0468569"...
3     <a href="https://www.imdb.com/title/tt0167260"...
4     <a href="https://www.imdb.com/title/tt0108052"...
...                                                 ...
4995  <a href="https://www.imdb.com/title/tt0880578"...
4996  <a href="https://www.imdb.com/title/tt27459160...
4997  <a href="https://www.imdb.com/title/tt3174376"...
4998  <a href="https://www.imdb.com/title/tt5177088"...
4999  <a href="https://www.imdb.com/title/tt8368408"...

[5000 rows x 12 columns]>
```

```
df.describe()
```

```
         startYear         rank  averageRating      numVotes  runtimeMinutes
count  5000.000000  5000.000000    5000.000000  5.000000e+03      5000.00000
mean   2002.021200  2500.500000       7.137140  1.630208e+05       114.56380
std      18.262844  1443.520003       0.597229  2.403863e+05        23.21629
min    1915.000000     1.000000       5.900000  2.500000e+04        25.00000
25%    1994.000000  1250.750000       6.700000  4.008425e+04        99.00000
50%    2007.000000  2500.500000       7.100000  7.438750e+04       111.00000
75%    2015.000000  3750.250000       7.600000  1.780925e+05       126.00000
max    2025.000000  5000.000000       9.300000  3.029407e+06       374.00000
```

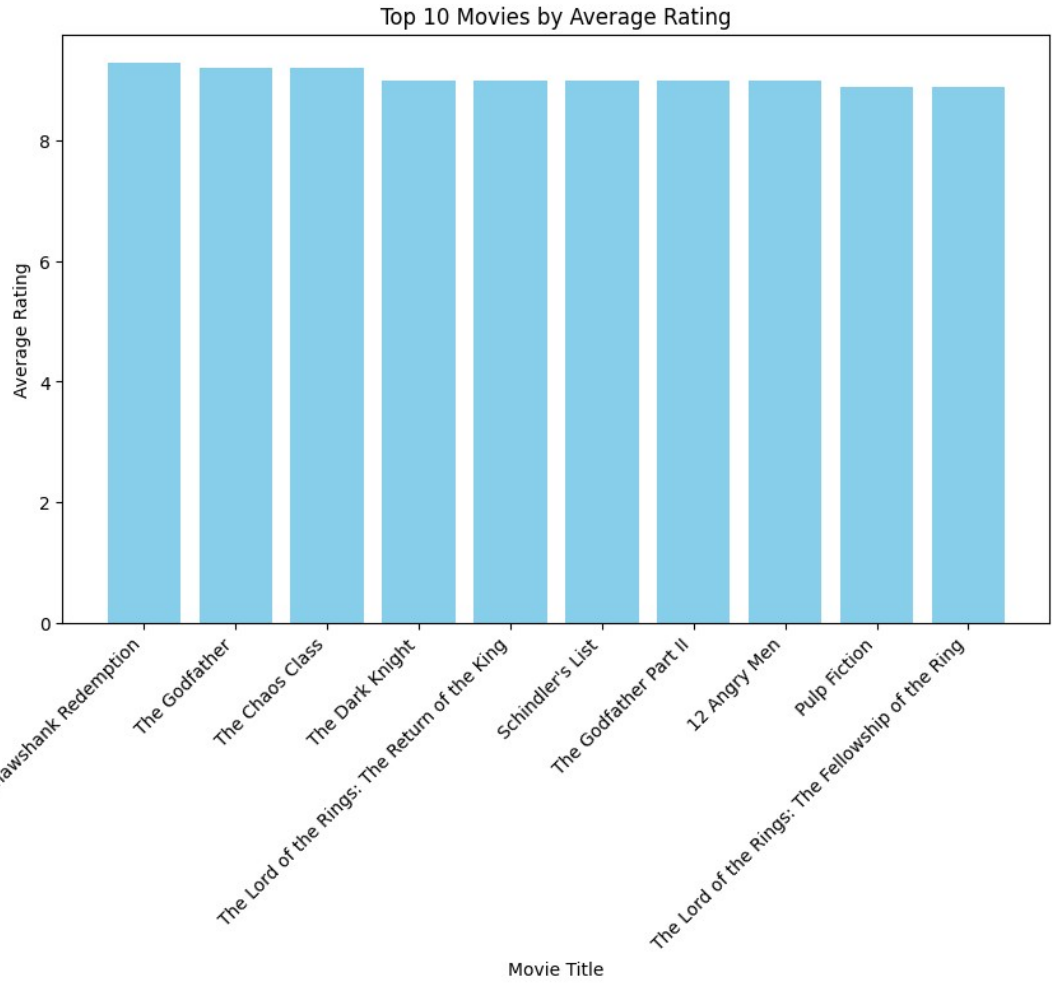
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv(r"C:\Users\piyus\Desktop\ids project\results_with_crew.csv")
```

```
df
```

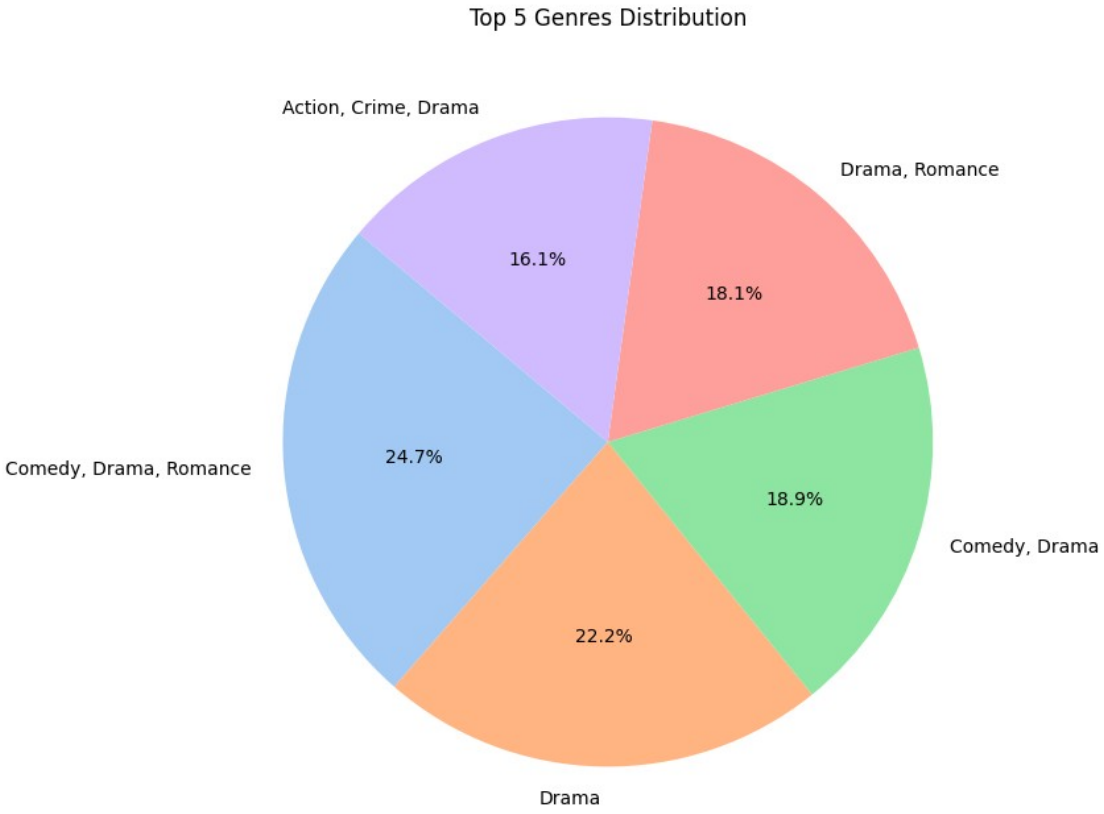
```
         tconst                                     primaryTitle  startYear  \
0      tt0111161                       The Shawshank Redemption       1994
1      tt0068646                                  The Godfather       1972
2      tt0468569                                The Dark Knight       2008
3      tt0167260  The Lord of the Rings: The Return of the King       2003
4      tt0108052                                Schindler's List       1993
...          ...                                            ...        ...
4995   tt0880578                                     Untraceable       2008
4996  tt27459160                Teri Baaton Mein Aisa Uljha Jiya       2024
4997   tt3174376                                  Before I Wake       2016
4998   tt5177088                      The Girl in the Spider's Web       2018
4999   tt8368408                               Gunpowder Milkshake       2021

      rank  averageRating  numVotes  runtimeMinutes                 directors
\
0        1            9.3   3029407             142            Frank Darabont
1        2            9.2   2114342             175      Francis Ford Coppola
2        3            9.0   3005759             152         Christopher Nolan
3        4            9.0   2068876             201             Peter Jackson
4        5            9.0   1515877             195          Steven Spielberg
...    ...            ...       ...             ...                       ...
4995  4996            6.2     54012             101            Gregory Hoblit
4996  4997            6.2     53979             141  Amit Joshi, Aradhana Sah
4997  4998            6.2     53934              97             Mike Flanagan
4998  4999            6.1     53870             115              Fede Alvarez
4999  5000            6.1     53710             114           Navot Papushado

                                            writers  \
0                       Stephen King, Frank Darabont
1                     Mario Puzo, Francis Ford Coppola
2     Jonathan Nolan, Christopher Nolan, David S. Go...
3     J.R.R. Tolkien, Fran Walsh, Philippa Boyens, P...
4                      Thomas Keneally, Steven Zaillian
...                                                 ...
4995    Robert Fyvolent, Mark Brinker, Allison Burnett
4996                          Amit Joshi, Aradhana Sah
4997                        Mike Flanagan, Jeff Howard
4998  Jay Basu, Fede Alvarez, Steven Knight, Stieg L...
4999                        Navot Papushado, Ehud Lavski

                        genres  \
0                        Drama
1                Crime, Drama
2          Action, Crime, Drama
3     Adventure, Drama, Fantasy
4     Biography, Drama, History
...                         ...
4995   Crime, Mystery, Thriller
```

```
4996     Comedy, Drama, Romance
4997     Drama, Fantasy, Horror
4998   Action, Adventure, Crime
4999    Action, Crime, Thriller

                                                  IMDbLink  \
0     <a href="https://www.imdb.com/title/tt0111161"...
1     <a href="https://www.imdb.com/title/tt0068646"...
2     <a href="https://www.imdb.com/title/tt0468569"...
3     <a href="https://www.imdb.com/title/tt0167260"...
4     <a href="https://www.imdb.com/title/tt0108052"...
...                                                 ...
4995  <a href="https://www.imdb.com/title/tt0880578"...
4996  <a href="https://www.imdb.com/title/tt27459160...
4997  <a href="https://www.imdb.com/title/tt3174376"...
4998  <a href="https://www.imdb.com/title/tt5177088"...
4999  <a href="https://www.imdb.com/title/tt8368408"...

                                       Title_IMDb_Link
0     <a href="https://www.imdb.com/title/tt0111161"...
1     <a href="https://www.imdb.com/title/tt0068646"...
2     <a href="https://www.imdb.com/title/tt0468569"...
3     <a href="https://www.imdb.com/title/tt0167260"...
4     <a href="https://www.imdb.com/title/tt0108052"...
...                                                 ...
4995  <a href="https://www.imdb.com/title/tt0880578"...
4996  <a href="https://www.imdb.com/title/tt27459160...
4997  <a href="https://www.imdb.com/title/tt3174376"...
4998  <a href="https://www.imdb.com/title/tt5177088"...
4999  <a href="https://www.imdb.com/title/tt8368408"...

[5000 rows x 12 columns]
```

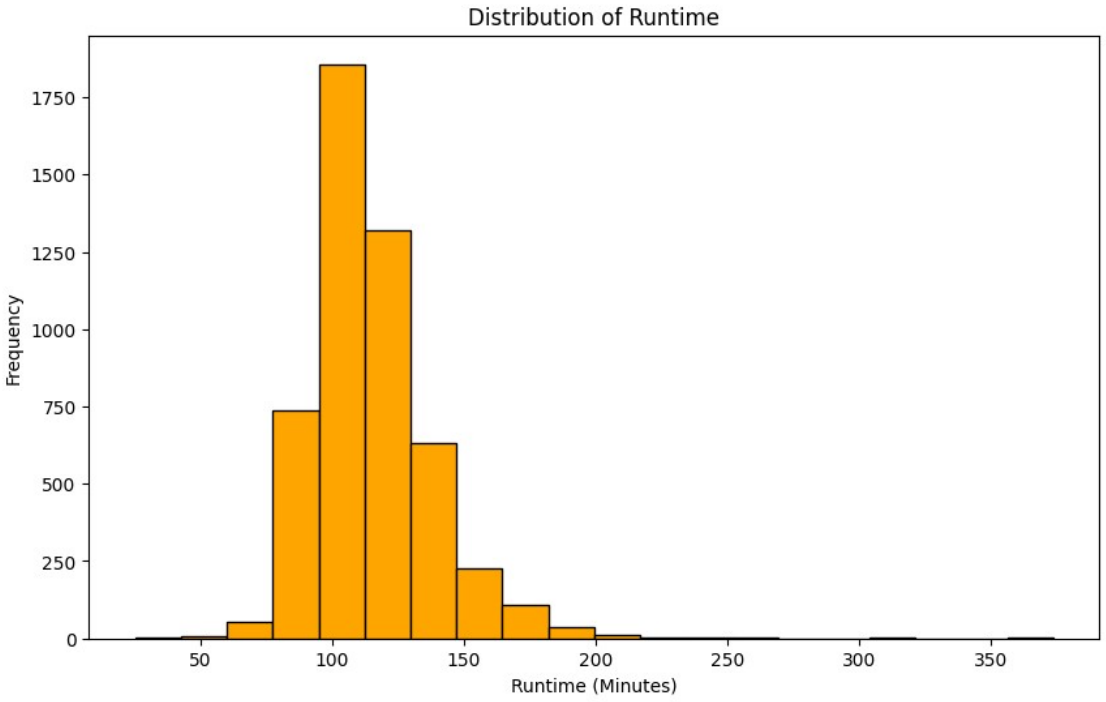
```
# Bar Plot: Top 10 movies by average rating
top_10_movies = df.nlargest(10, 'averageRating')
plt.figure(figsize=(10, 6))
plt.bar(top_10_movies['primaryTitle'], top_10_movies['averageRating'],
color='skyblue')
plt.xticks(rotation=45, ha='right')
plt.title('Top 10 Movies by Average Rating')
plt.xlabel('Movie Title')
plt.ylabel('Average Rating')
plt.show()
```




```
# Pie Chart: Distribution of genres
genre_counts = df['genres'].value_counts().head(5)  # Top 5 genres
plt.figure(figsize=(8, 8))
plt.pie(genre_counts, labels=genre_counts.index, autopct='%1.1f%%',
startangle=140, colors=sns.color_palette("pastel"))
plt.title('Top 5 Genres Distribution')
plt.show()
```



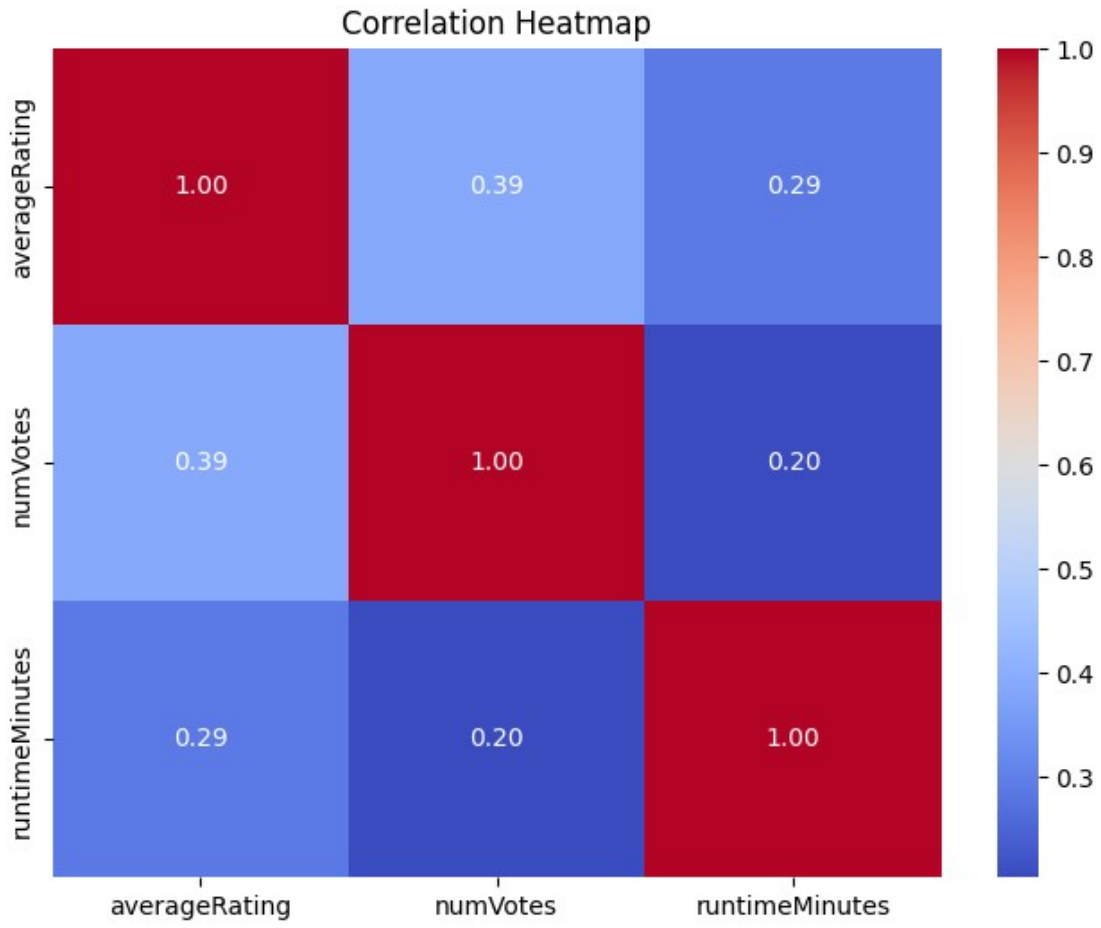

```
# Histogram: Distribution of runtime
plt.figure(figsize=(10, 6))
plt.hist(df['runtimeMinutes'].dropna(), bins=20, color='orange',
edgecolor='black')
plt.title('Distribution of Runtime')
plt.xlabel('Runtime (Minutes)')
plt.ylabel('Frequency')
plt.show()
```



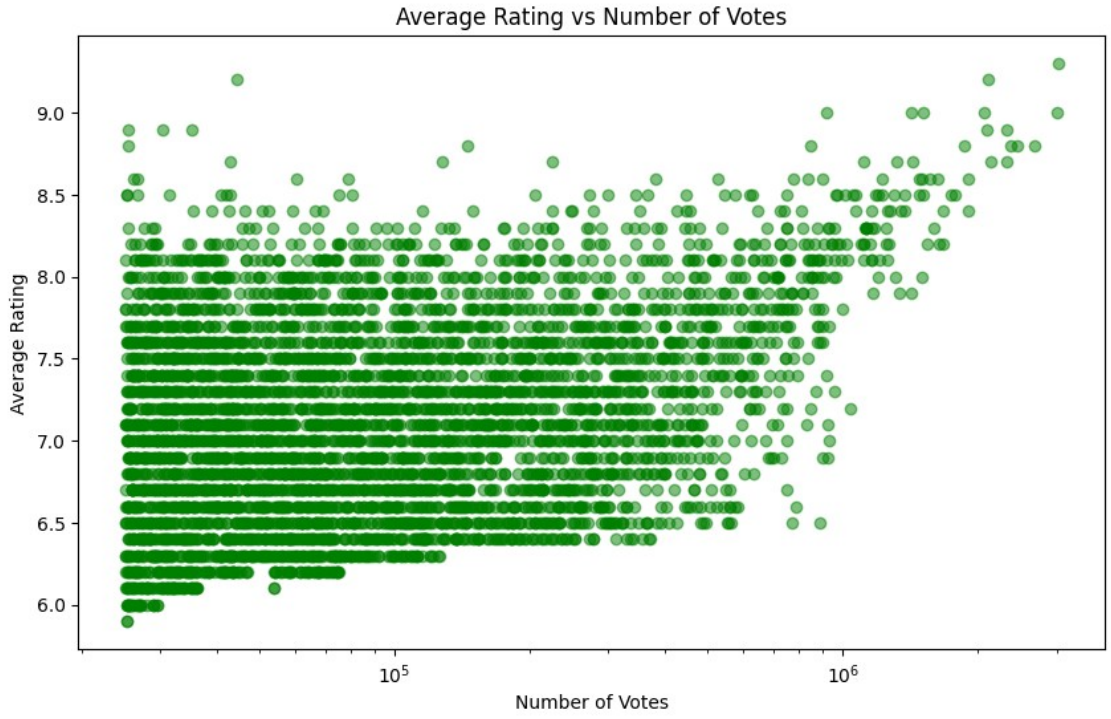

```
# Correlation Matrix
correlation_matrix = df[['averageRating', 'numVotes',
'runtimeMinutes']].corr()
print(correlation_matrix)
```

```
                averageRating  numVotes  runtimeMinutes
averageRating        1.000000  0.387595        0.287495
numVotes             0.387595  1.000000        0.203039
runtimeMinutes       0.287495  0.203039        1.000000
```

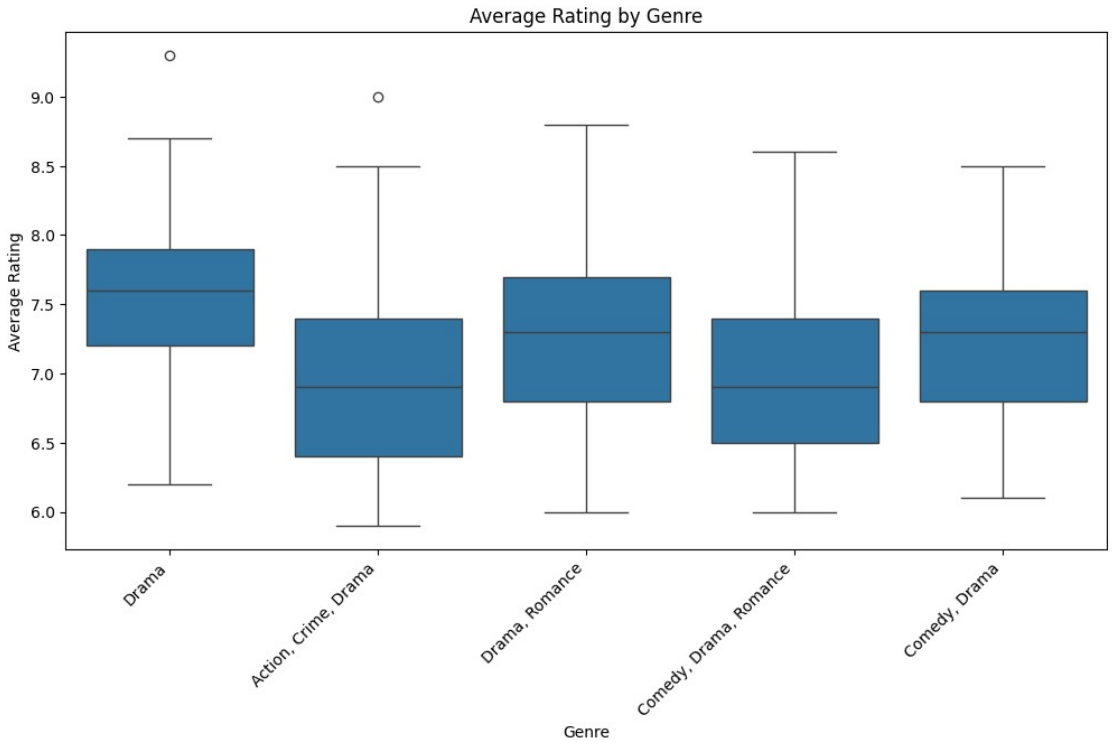
```
# Correlation Heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()
```



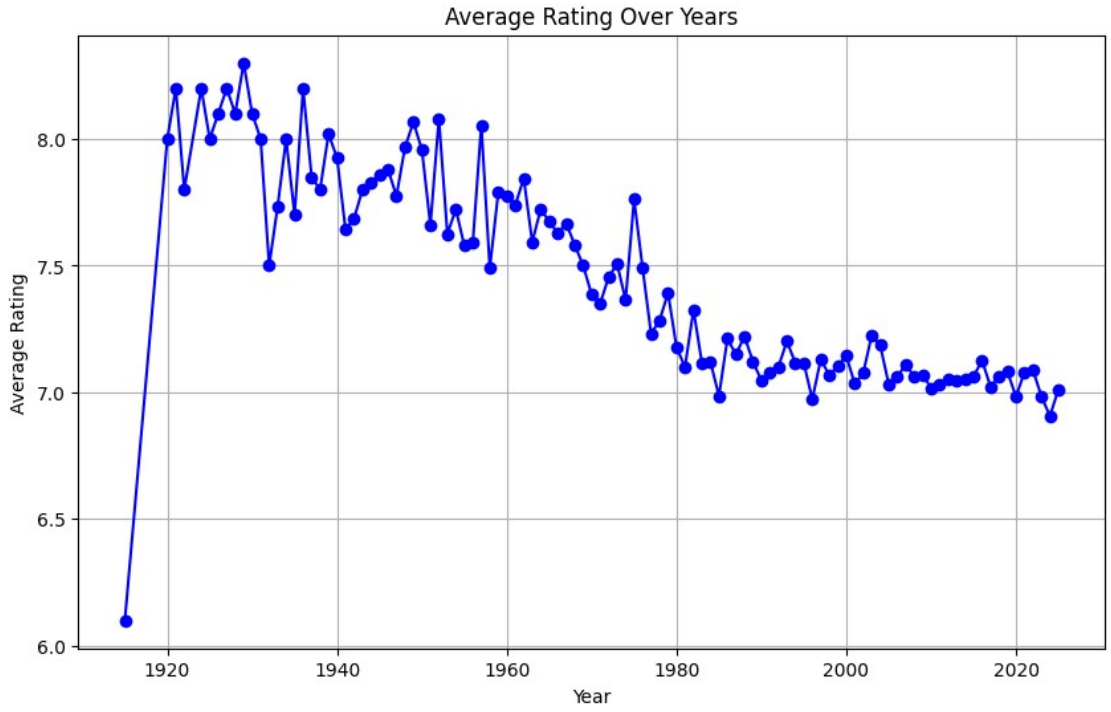

```
# Scatter Plot: Average Rating vs Number of Votes
plt.figure(figsize=(10, 6))
plt.scatter(df['numVotes'], df['averageRating'], alpha=0.5, color='green')
plt.title('Average Rating vs Number of Votes')
plt.xlabel('Number of Votes')
plt.ylabel('Average Rating')
plt.xscale('log')  # Log scale for better visualization
plt.show()
```




```
# Boxplot: Average Rating by Genre
plt.figure(figsize=(12, 6))
sns.boxplot(x='genres', y='averageRating',
data=df[df['genres'].isin(genre_counts.index)])
plt.xticks(rotation=45, ha='right')
plt.title('Average Rating by Genre')
plt.xlabel('Genre')
plt.ylabel('Average Rating')
plt.show()
```




```
# Time-Series Analysis: Average Rating over Years
df['startYear'] = pd.to_numeric(df['startYear'], errors='coerce')  # Ensure
numeric year
yearly_avg_rating = df.groupby('startYear')['averageRating'].mean().dropna()
plt.figure(figsize=(10, 6))
plt.plot(yearly_avg_rating.index, yearly_avg_rating.values, marker='o',
linestyle='-', color='blue')
plt.title('Average Rating Over Years')
plt.xlabel('Year')
plt.ylabel('Average Rating')
plt.grid()
plt.show()
```




```
# Data Cleaning & Preprocessing
# Handle missing values and ensure correct data types
df['startYear'] = pd.to_numeric(df['startYear'], errors='coerce')  # Convert
startYear to numeric
df['runtimeMinutes'] = pd.to_numeric(df['runtimeMinutes'], errors='coerce')
# Convert runtimeMinutes to numeric
df['averageRating'] = pd.to_numeric(df['averageRating'], errors='coerce')  #
Convert averageRating to numeric
df['numVotes'] = pd.to_numeric(df['numVotes'], errors='coerce')  # Convert
numVotes to numeric

# Drop rows with missing critical values
df = df.dropna(subset=['averageRating', 'numVotes', 'runtimeMinutes',
'startYear'])
```

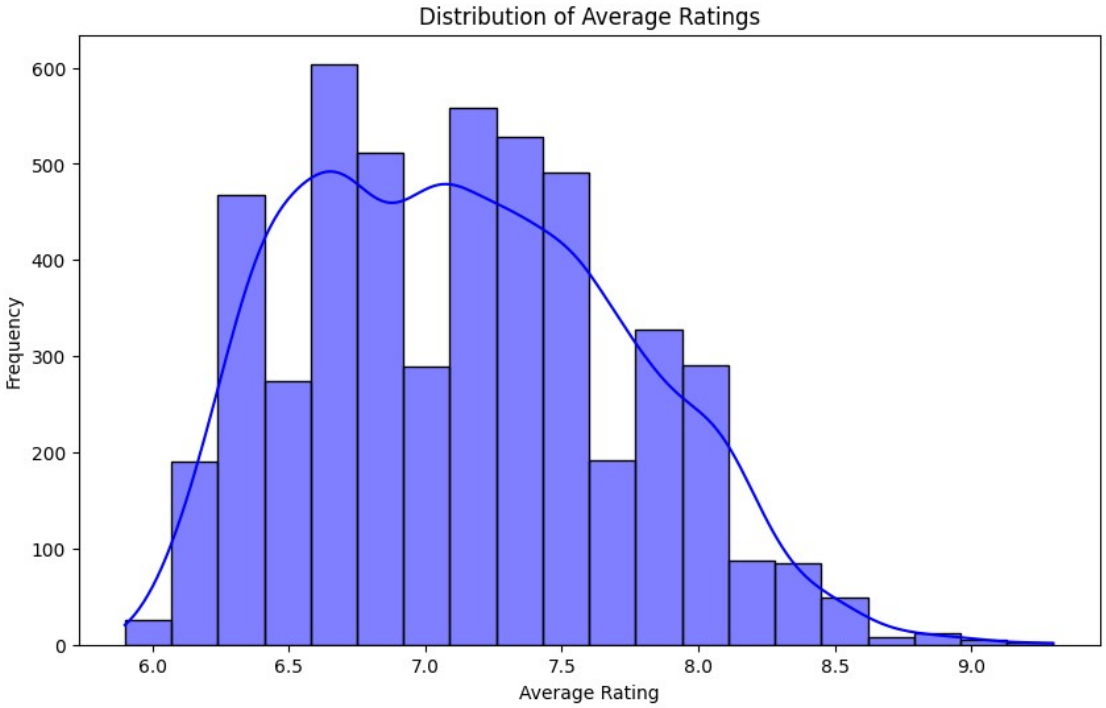
```
df
```

```
          tconst                                     primaryTitle  startYear  \
0      tt0111161                          The Shawshank Redemption       1994
1      tt0068646                                     The Godfather       1972
2      tt0468569                                   The Dark Knight       2008
3      tt0167260  The Lord of the Rings: The Return of the King       2003
4      tt0108052                                  Schindler's List       1993
...          ...                                               ...        ...
4995   tt0880578                                       Untraceable       2008
4996  tt27459160                    Teri Baaton Mein Aisa Uljha Jiya       2024
4997   tt3174376                                       Before I Wake       2016
4998   tt5177088                       The Girl in the Spider's Web       2018
4999   tt8368408                                  Gunpowder Milkshake       2021
```

```
      rank  averageRating  numVotes  runtimeMinutes                directors
\
0        1            9.3   3029407             142           Frank Darabont
1        2            9.2   2114342             175     Francis Ford Coppola
2        3            9.0   3005759             152        Christopher Nolan
3        4            9.0   2068876             201            Peter Jackson
4        5            9.0   1515877             195         Steven Spielberg
...    ...            ...       ...             ...                      ...
4995  4996            6.2     54012             101           Gregory Hoblit
4996  4997            6.2     53979             141  Amit Joshi, Aradhana Sah
4997  4998            6.2     53934              97            Mike Flanagan
4998  4999            6.1     53870             115             Fede Alvarez
4999  5000            6.1     53710             114           Navot Papushado

                                              writers  \
0                          Stephen King, Frank Darabont
1                        Mario Puzo, Francis Ford Coppola
2     Jonathan Nolan, Christopher Nolan, David S. Go...
3     J.R.R. Tolkien, Fran Walsh, Philippa Boyens, P...
4                         Thomas Keneally, Steven Zaillian
...                                                  ...
4995     Robert Fyvolent, Mark Brinker, Allison Burnett
4996                             Amit Joshi, Aradhana Sah
4997                           Mike Flanagan, Jeff Howard
4998  Jay Basu, Fede Alvarez, Steven Knight, Stieg L...
4999                         Navot Papushado, Ehud Lavski

                         genres  \
0                         Drama
1                  Crime, Drama
2           Action, Crime, Drama
3     Adventure, Drama, Fantasy
4     Biography, Drama, History
...                         ...
4995   Crime, Mystery, Thriller
4996     Comedy, Drama, Romance
4997     Drama, Fantasy, Horror
4998   Action, Adventure, Crime
4999    Action, Crime, Thriller

                                              IMDbLink  \
0     <a href="https://www.imdb.com/title/tt0111161"...
1     <a href="https://www.imdb.com/title/tt0068646"...
2     <a href="https://www.imdb.com/title/tt0468569"...
3     <a href="https://www.imdb.com/title/tt0167260"...
4     <a href="https://www.imdb.com/title/tt0108052"...
...                                                 ...
4995  <a href="https://www.imdb.com/title/tt0880578"...
```

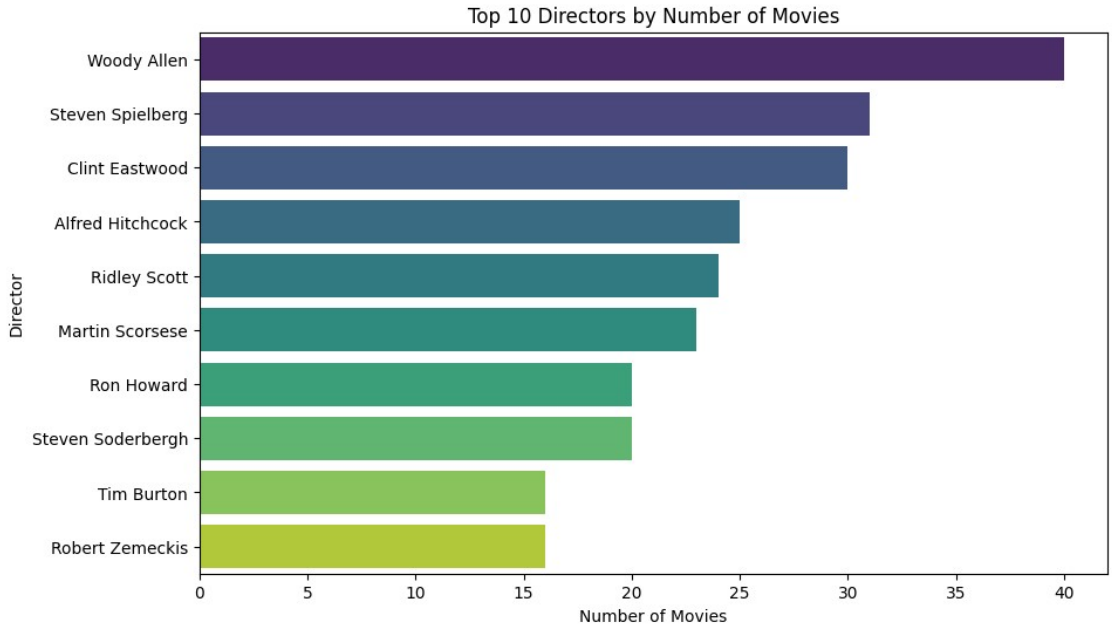
```
4996  <a href="https://www.imdb.com/title/tt27459160...
4997  <a href="https://www.imdb.com/title/tt3174376"...
4998  <a href="https://www.imdb.com/title/tt5177088"...
4999  <a href="https://www.imdb.com/title/tt8368408"...

                                        Title_IMDb_Link
0     <a href="https://www.imdb.com/title/tt0111161"...
1     <a href="https://www.imdb.com/title/tt0068646"...
2     <a href="https://www.imdb.com/title/tt0468569"...
3     <a href="https://www.imdb.com/title/tt0167260"...
4     <a href="https://www.imdb.com/title/tt0108052"...
...                                                 ...
4995  <a href="https://www.imdb.com/title/tt0880578"...
4996  <a href="https://www.imdb.com/title/tt27459160...
4997  <a href="https://www.imdb.com/title/tt3174376"...
4998  <a href="https://www.imdb.com/title/tt5177088"...
4999  <a href="https://www.imdb.com/title/tt8368408"...

[5000 rows x 12 columns]
```

```python
# Extract primary genre for simplicity
df['primaryGenre'] = df['genres'].str.split(',').str[0]

# EDA with plots
import matplotlib.pyplot as plt
import seaborn as sns

# Distribution of average ratings
plt.figure(figsize=(10, 6))
sns.histplot(df['averageRating'], bins=20, kde=True, color='blue')
plt.title('Distribution of Average Ratings')
plt.xlabel('Average Rating')
plt.ylabel('Frequency')
plt.show()
```

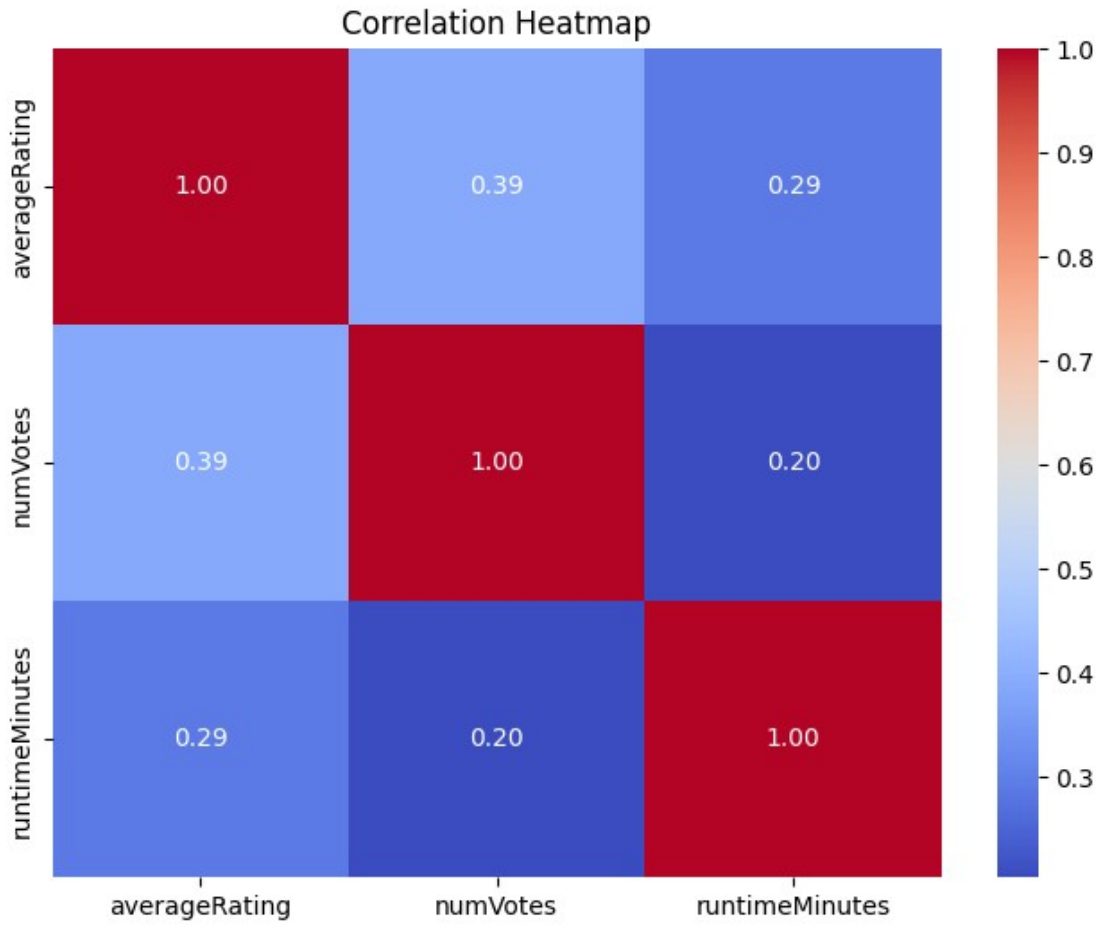



```
# Top 10 directors by number of movies
top_directors = df['directors'].value_counts().head(10)
plt.figure(figsize=(10, 6))
sns.barplot(x=top_directors.values, y=top_directors.index, palette='viridis')
plt.title('Top 10 Directors by Number of Movies')
plt.xlabel('Number of Movies')
plt.ylabel('Director')
plt.show()
```

```
C:\Users\piyus\AppData\Local\Temp\ipykernel_26520\4062912145.py:4:
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(x=top_directors.values, y=top_directors.index,
palette='viridis')
```




```
# Correlation heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(df[['averageRating', 'numVotes', 'runtimeMinutes']].corr(),
annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()
```




```
# Feature Engineering
# Create a new feature: log-transformed numVotes for better scaling
import numpy as np
df['log_numVotes'] = np.log1p(df['numVotes'])

# One-hot encode primaryGenre
df = pd.get_dummies(df, columns=['primaryGenre'], drop_first=True)

# Select features and target
X = df[['runtimeMinutes', 'log_numVotes'] + [col for col in df.columns if
col.startswith('primaryGenre_')]]
y = df['averageRating']

# Train-Test Split
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Train a Linear Regression Model
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train, y_train)
```

```
LinearRegression()

# Evaluate the model
from sklearn.metrics import mean_squared_error, r2_score

# Predictions
y_pred = model.predict(X_test)

# Metrics
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
print(f"R-squared: {r2}")

Mean Squared Error: 0.2848424333814623
R-squared: 0.24909020219910427
```

Conclusion

This project successfully used Python's data analysis libraries (Pandas, Matplotlib, Seaborn) to extract insights from an IMDb movie dataset. It highlighted the characteristics of top-rated films, identified dominant genres, and explored runtime distributions. Despite challenges with data quality and formatting, the analysis provided a deeper understanding of what factors may influence movie ratings and popularity, offering valuable insights for further exploration or recommendation system development.