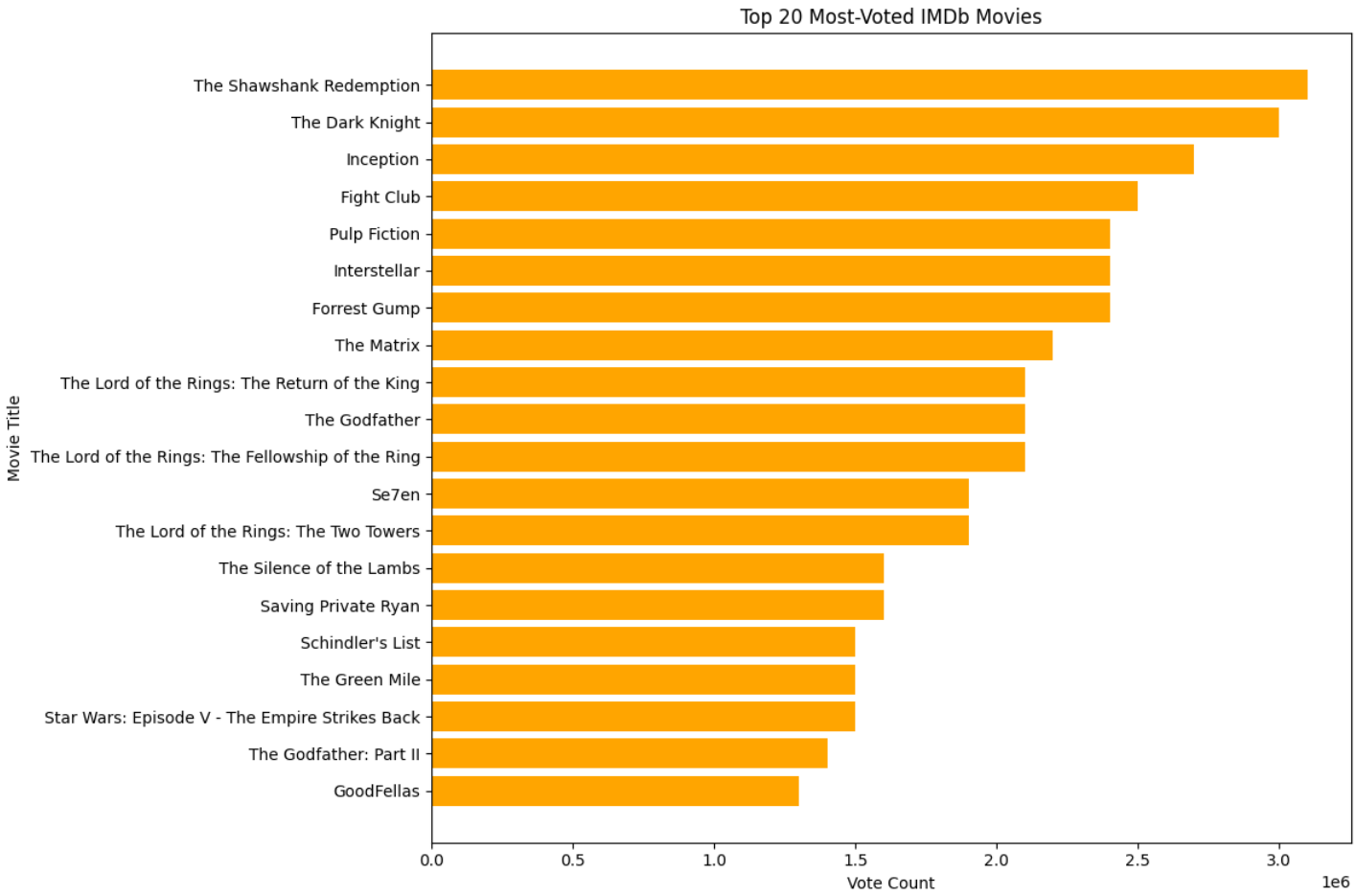
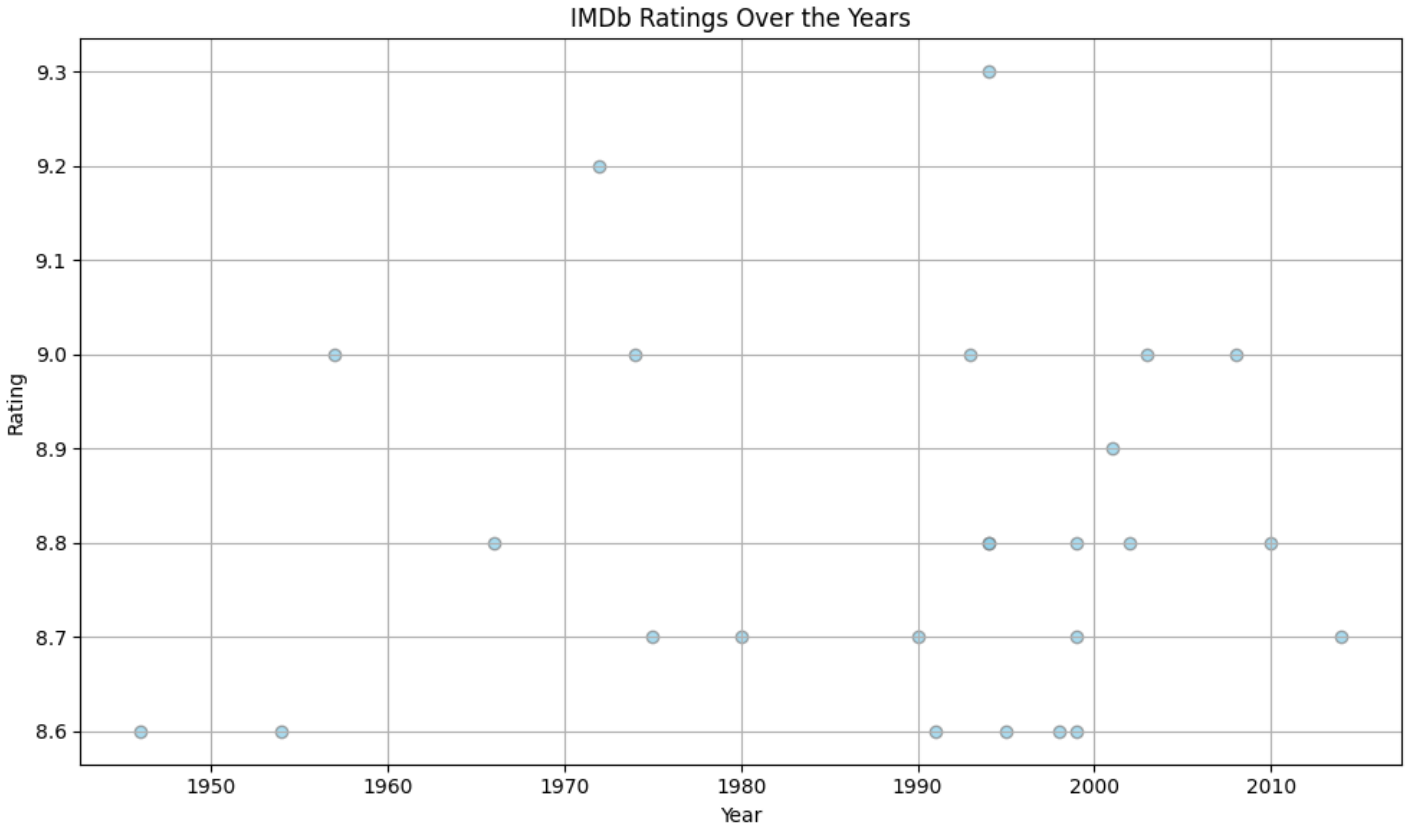
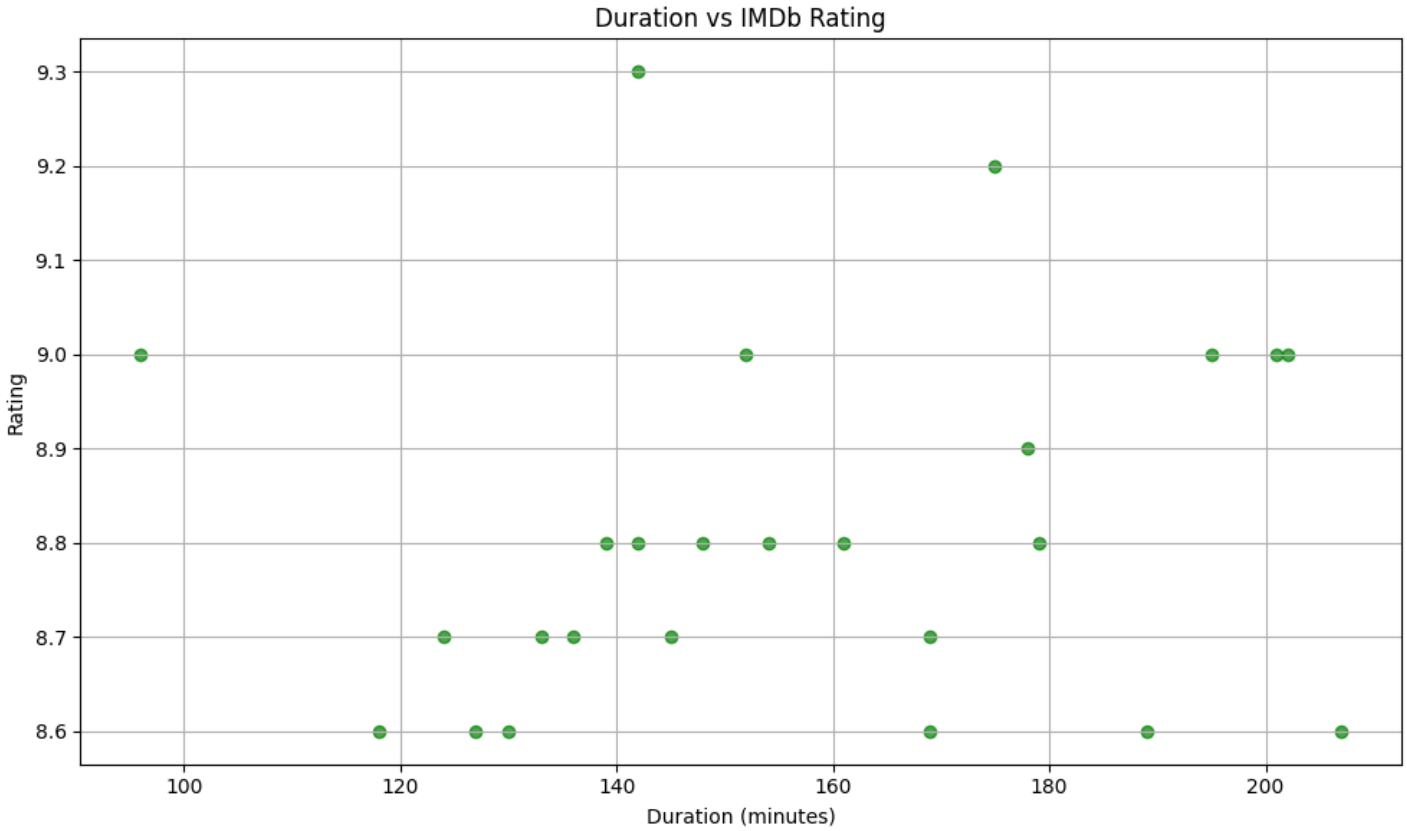
IMDb Top 250 Movies Analysis

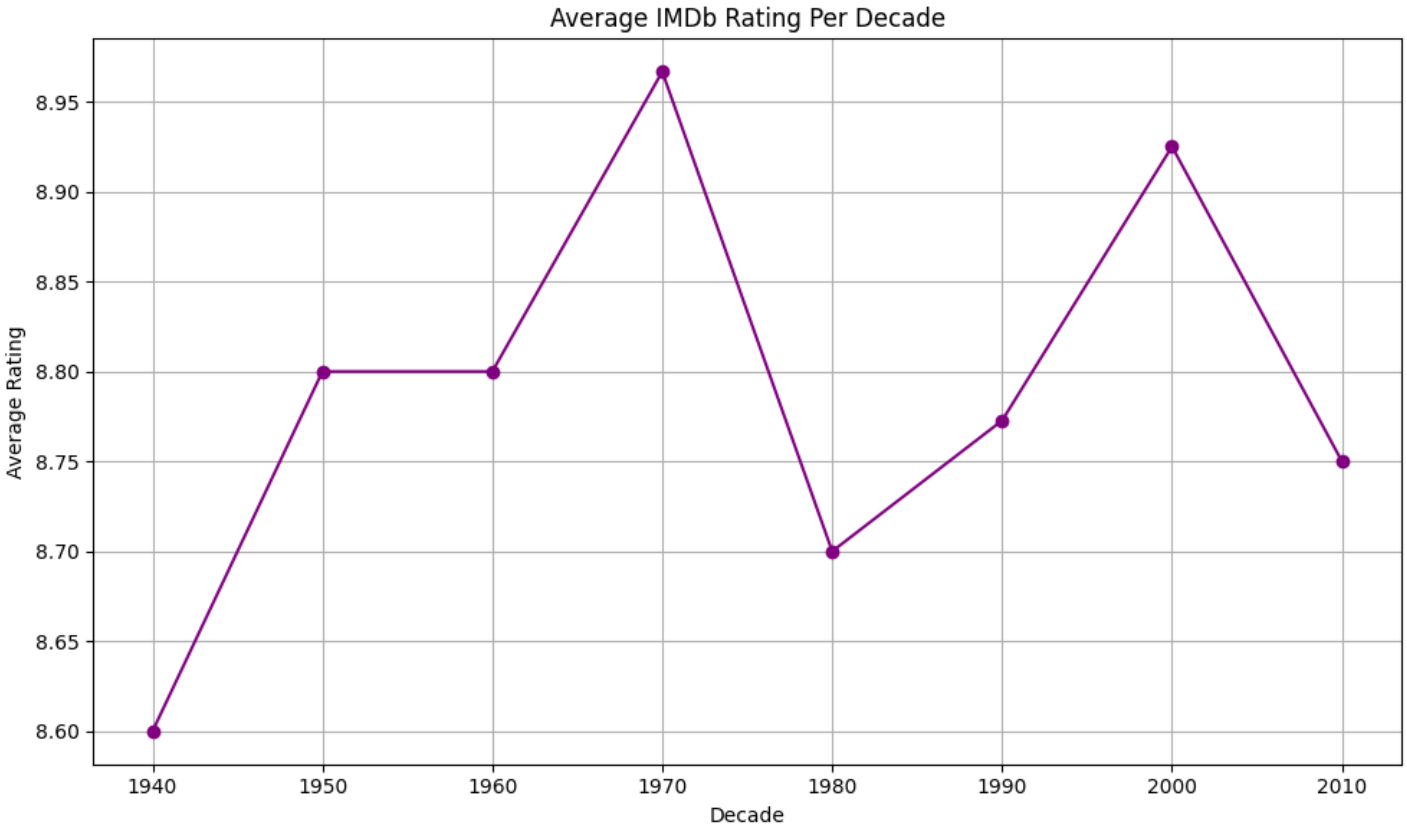
```
import pandas as pd
import matplotlib.pyplot as plt
import re

# Load Excel file
df = pd.read_excel("IMDb_Top_250.xlsx")
df.head()


# Function to extract numeric rating and vote count
def extract_rating_votes(rating_str):
    rating_str = rating_str.replace('\xa0', '')  # Clean non-breaking space
    match = re.match(r"([\d.]+)\s*\(([\d.,]+)([MK])\)", rating_str)
    if match:
        rating = float(match.group(1))
        count = float(match.group(2).replace(',', ''))
        multiplier = {'K': 1e3, 'M': 1e6}
        votes = count * multiplier[match.group(3)]
        return rating, int(votes)
    return None, None

df[['RatingValue',        'VoteCount']]        =        df['Rating'].apply(lambda        x:
pd.Series(extract_rating_votes(x)))

# Convert Duration (e.g., "2h 22m") to minutes
def duration_to_minutes(duration_str):
    match = re.match(r'(\d+)h\s*(\d+)m', duration_str)
    if match:
        return int(match.group(1)) * 60 + int(match.group(2))
    return None

df['DurationMin'] = df['Duration'].apply(duration_to_minutes)
df.head()


plt.figure(figsize=(10, 6))
plt.scatter(df['Year'], df['RatingValue'], alpha=0.7, color='skyblue', edgecolor='gray')
plt.title("IMDb Ratings Over the Years")
plt.xlabel("Year")
plt.ylabel("Rating")
plt.grid(True)
plt.tight_layout()
plt.show()


top_voted = df.sort_values(by='VoteCount', ascending=False).head(20)

plt.figure(figsize=(12, 8))
plt.barh(top_voted['Title'][::-1], top_voted['VoteCount'][::-1], color='orange')
plt.title("Top 20 Most-Voted IMDb Movies")
plt.xlabel("Vote Count")
plt.ylabel("Movie Title")
plt.tight_layout()
```

```
plt.show()


plt.figure(figsize=(10, 6))
plt.scatter(df['DurationMin'], df['RatingValue'], alpha=0.7, color='green')
plt.title("Duration vs IMDb Rating")
plt.xlabel("Duration (minutes)")
plt.ylabel("Rating")
plt.grid(True)
plt.tight_layout()
plt.show()


df['Decade'] = (df['Year'] // 10) * 10
decade_rating = df.groupby('Decade')['RatingValue'].mean().reset_index()

plt.figure(figsize=(10, 6))
plt.plot(decade_rating['Decade'],        decade_rating['RatingValue'],        marker='o',
linestyle='-', color='purple')
plt.title("Average IMDb Rating Per Decade")
plt.xlabel("Decade")
plt.ylabel("Average Rating")
plt.grid(True)
plt.tight_layout()
plt.show()
```


Top 20 Most-Voted IMDb Movies
The Shawshank Redemption
The Dark Knight
Inception
Fight Club
Pulp Fiction
Interstellar
Forrest Gump
The Matrix
The Lord of the Rings: The Return of the King
The Godfather
The Lord of the Rings: The Fellowship of the Ring
Se7en
The Lord of the Rings: The Two Towers
The Silence of the Lambs
Saving Private Ryan
Schindler's List
The Green Mile
Star Wars: Episode V - The Empire Strikes Back
The Godfather: Part II
GoodFellas
Movie Title
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Vote Count
1e6


IMDb Ratings Over the Years
Rating
9.3
9.2
9.1
9.0
8.9
8.8
8.7
8.6
1950
1960
1970
1980
1990
2000
2010
Year


Duration vs IMDb Rating
Rating
9.3
9.2
9.1
9.0
8.9
8.8
8.7
8.6
100
120
140
160
180
200
Duration (minutes)


Average IMDb Rating Per Decade
Average Rating
8.95
8.90
8.85
8.80
8.75
8.70
8.65
8.60
1940
1950
1960
1970
1980
1990
2000
2010
Decade