IMDB Top 250 Films

Examining Trends

What are the trends of the most popular films on IMDB?

This analysis starts with the Top 250 Films Dataset from the IMDB and looks for trends among the films over time.

Questions include:

- What are the most popular genres within this dataset?
- Do ratings increase or decrease over time?
- How does film runtime change over time?
- What is the relationship between Budget and Revenue over time?
- How does the number of ratings change over time?
- How do the Top 10 films perform in terms of Budget and Revenue?
- Are movies becoming more profitable as budgets increase?

Datasets Used:

This analysis draws from 2 datasets, both of which are hosted through Kaggle. The data provided includes films from 1921 through 2017.

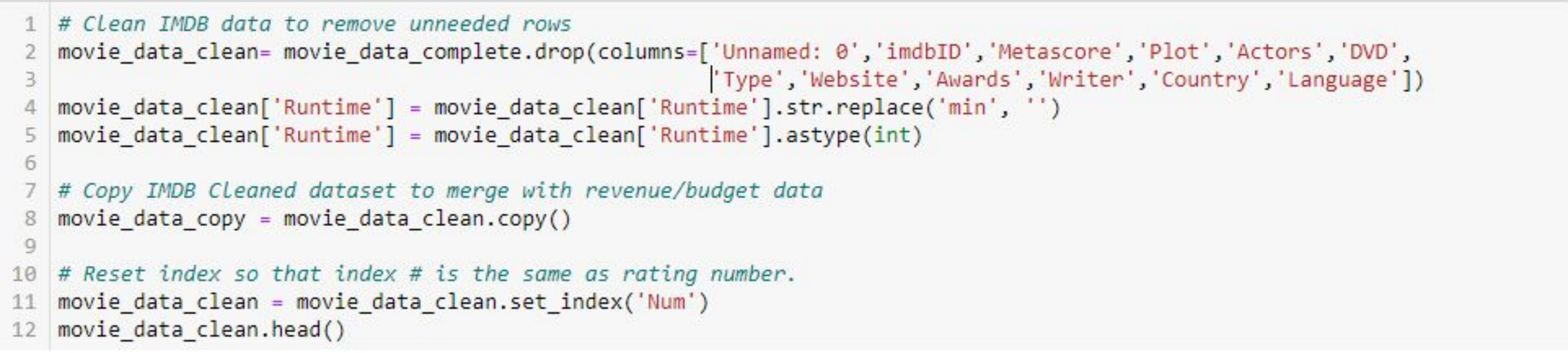
1. "IMDB Top 250 Movies": https://www.kaggle.com/datasets/yehorkorzh/imdb-top-250-movies

2. MovieLens "The Movies Dataset"(used to supplement Revenue and add Budget Data):

   https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

Data Cleaning

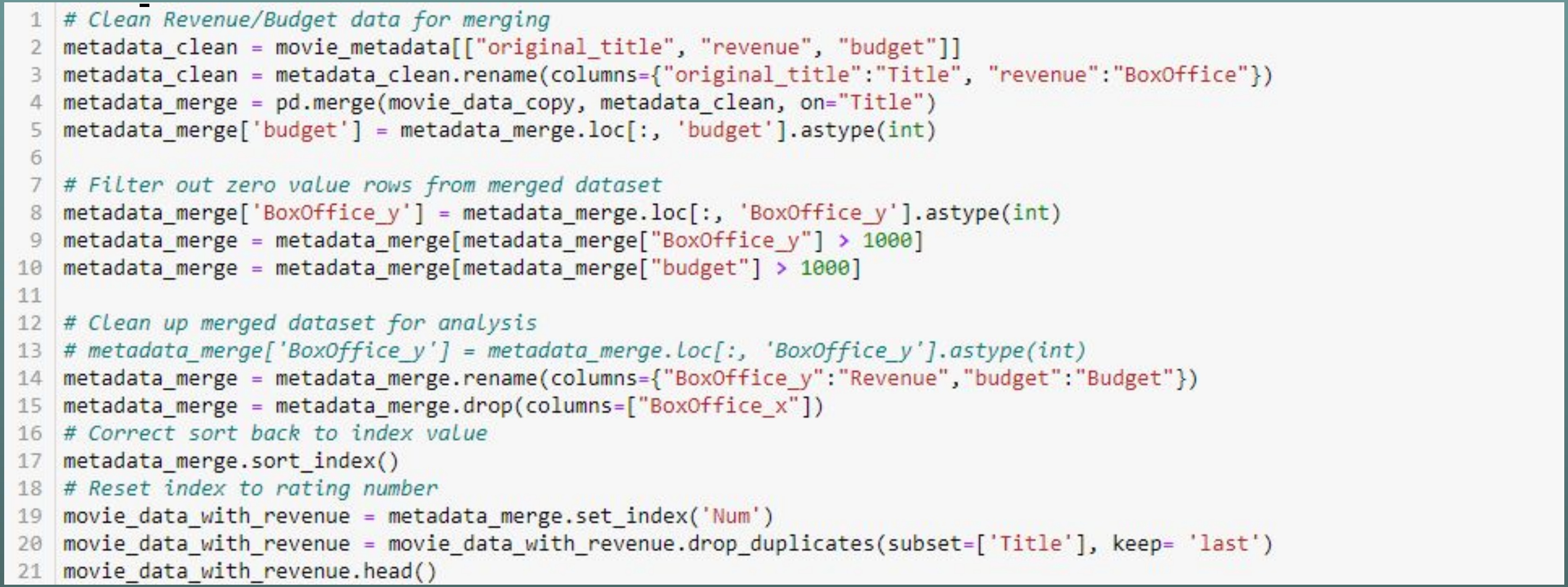
- Datasets had unneeded columns and missing values removed
- Index of datasets reset
- Prepwork for merging sets



```python
# Clean IMDB data to remove unneeded rows
movie_data_clean= movie_data_complete.drop(columns=['Unnamed: 0','imdbID','Metascore','Plot','Actors','DVD',
                                                    'Type','Website','Awards','Writer','Country','Language'])
movie_data_clean['Runtime'] = movie_data_clean['Runtime'].str.replace('min', '')
movie_data_clean['Runtime'] = movie_data_clean['Runtime'].astype(int)

# Copy IMDB Cleaned dataset to merge with revenue/budget data
movie_data_copy = movie_data_clean.copy()

# Reset index so that index # is the same as rating number.
movie_data_clean = movie_data_clean.set_index('Num')
movie_data_clean.head()
```

Data Cleaning 2

- Merge Datasets based on Title, clean up naming, drop duplicate entries from Metadata Dataframe.



```python
# Clean Revenue/Budget data for merging
metadata_clean = movie_metadata[["original_title", "revenue", "budget"]]
metadata_clean = metadata_clean.rename(columns={"original_title":"Title", "revenue":"BoxOffice"})
metadata_merge = pd.merge(movie_data_copy, metadata_clean, on="Title")
metadata_merge['budget'] = metadata_merge.loc[:, 'budget'].astype(int)

# Filter out zero value rows from merged dataset
metadata_merge['BoxOffice_y'] = metadata_merge.loc[:, 'BoxOffice_y'].astype(int)
metadata_merge = metadata_merge[metadata_merge["BoxOffice_y"] > 1000]
metadata_merge = metadata_merge[metadata_merge["budget"] > 1000]

# Clean up merged dataset for analysis
# metadata_merge['BoxOffice_y'] = metadata_merge.loc[:, 'BoxOffice_y'].astype(int)
metadata_merge = metadata_merge.rename(columns={"BoxOffice_y":"Revenue","budget":"Budget"})
metadata_merge = metadata_merge.drop(columns=["BoxOffice_x"])
# Correct sort back to index value
metadata_merge.sort_index()
# Reset index to rating number
movie_data_with_revenue = metadata_merge.set_index('Num')
movie_data_with_revenue = movie_data_with_revenue.drop_duplicates(subset=['Title'], keep= 'last')
movie_data_with_revenue.head()
```

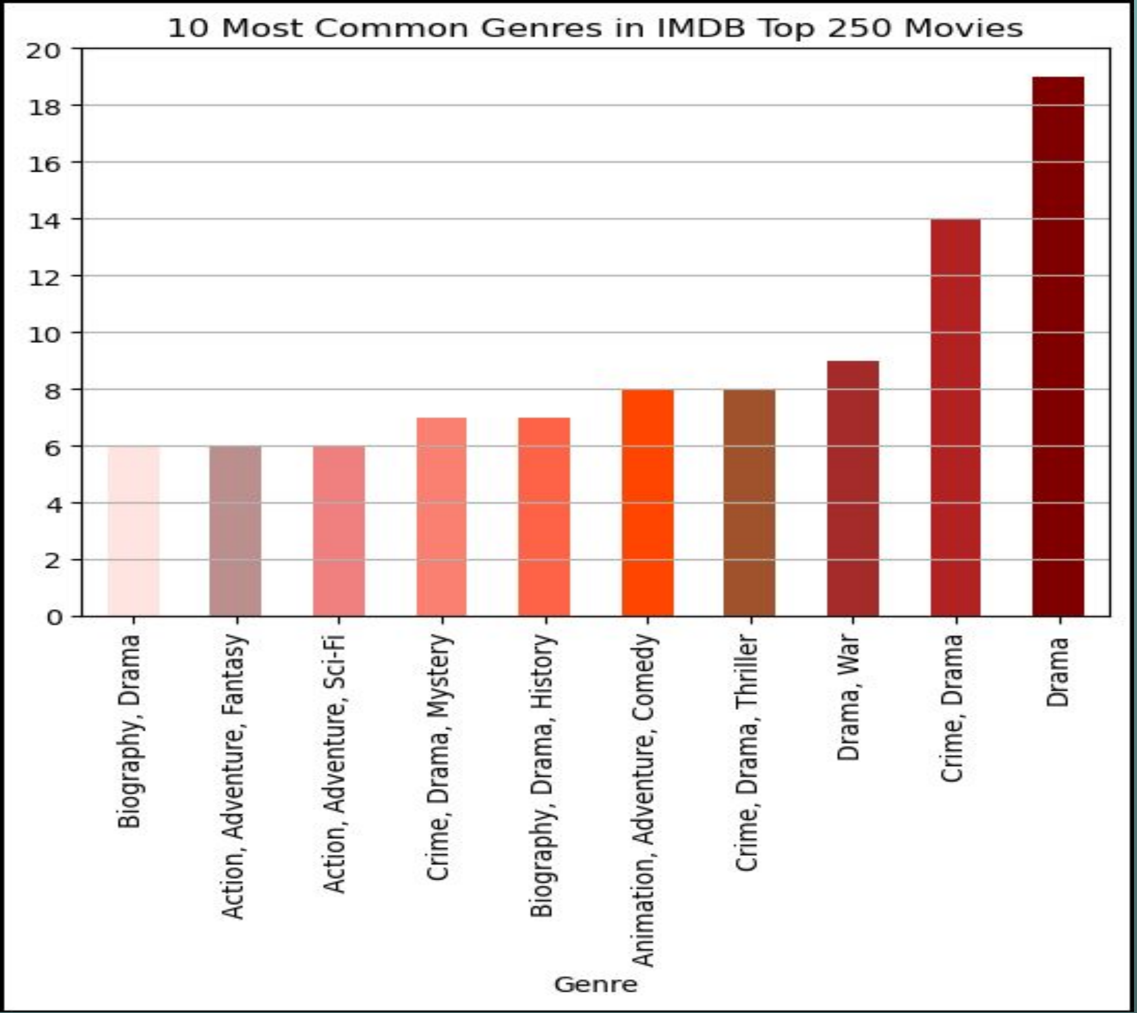
Dataset Averages for the Top 250 Films

The **Average Score** is **8.25 (out of 10)**

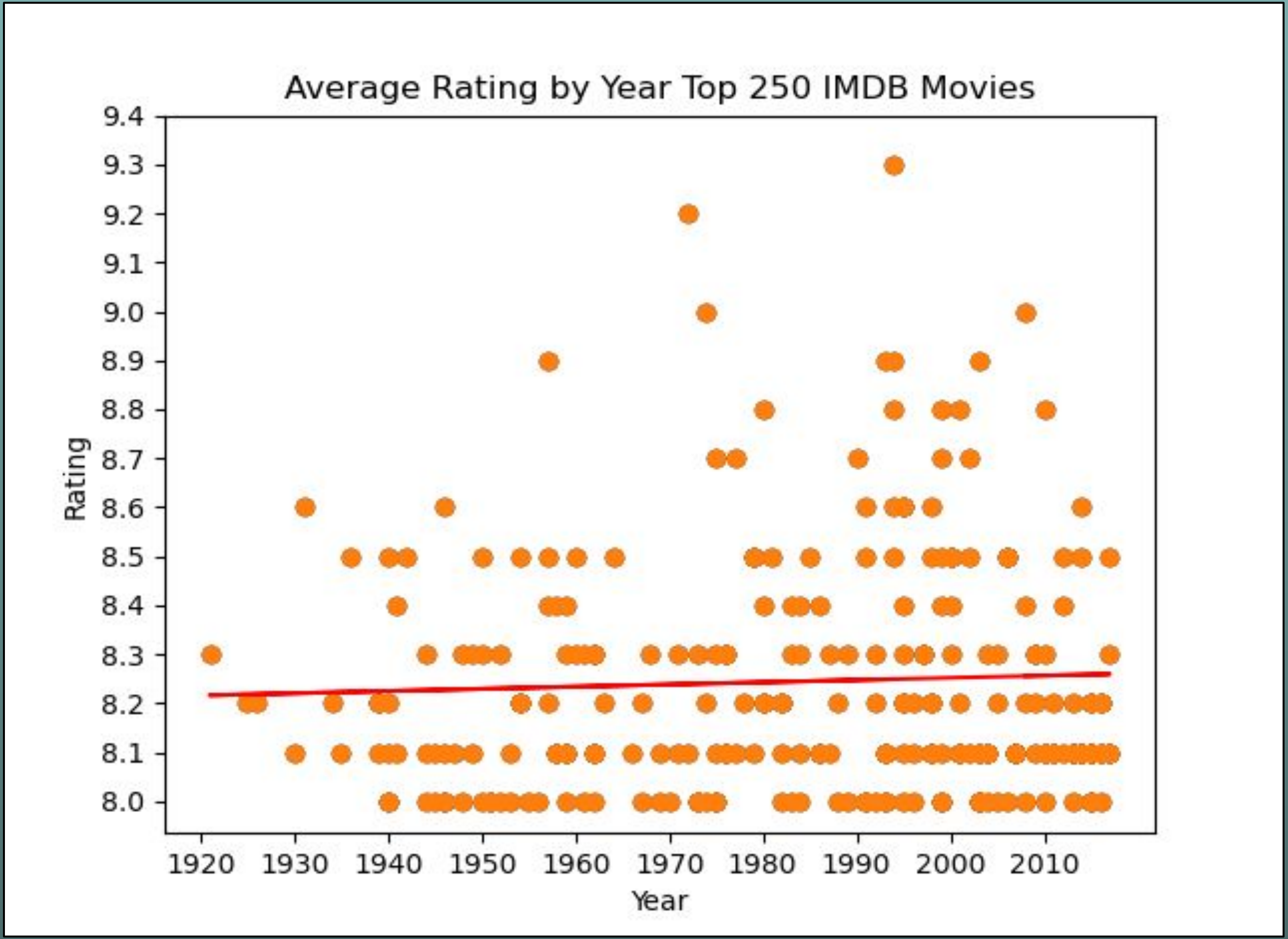
The **Average Vote Count** is **478151.68**

The **Average Budget** is **$67.61 Million**

The **Average Revenue** is **$232.95 Million**



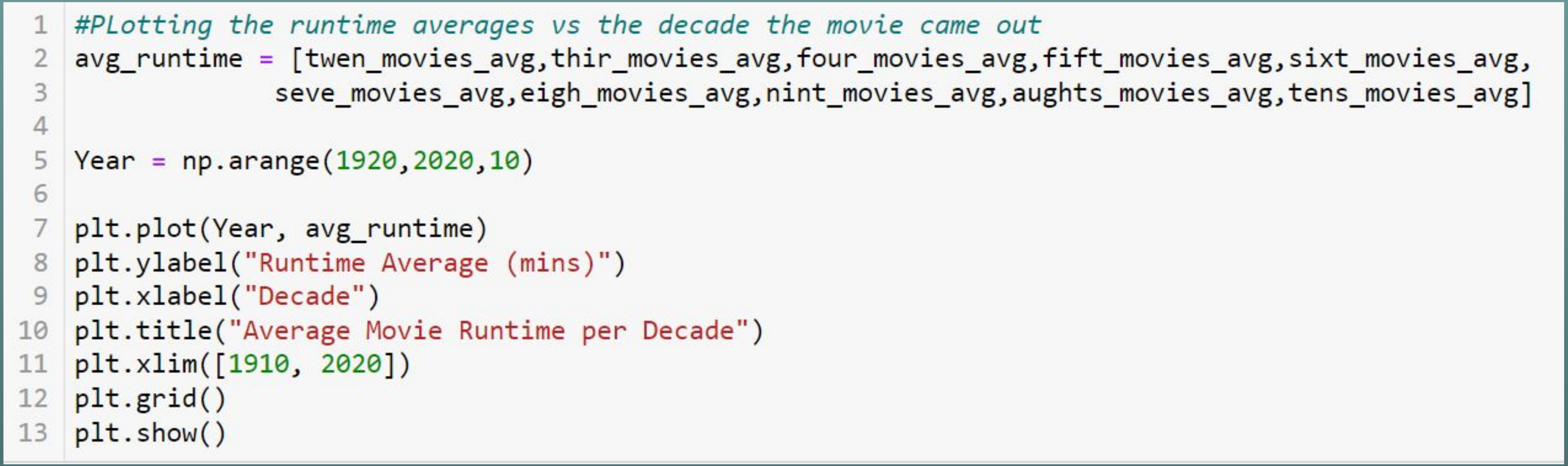
- Drama occurs most frequently
- Crime appears second most
- Adventure appears third most



- There appears to be a very small correlation between year and average rating
- Overall, one could say that more recent movies have higher ratings.
- There are some issues:
    - More movies released in modern time
    - Older movies may not have the same amount of viewers who review as new movies

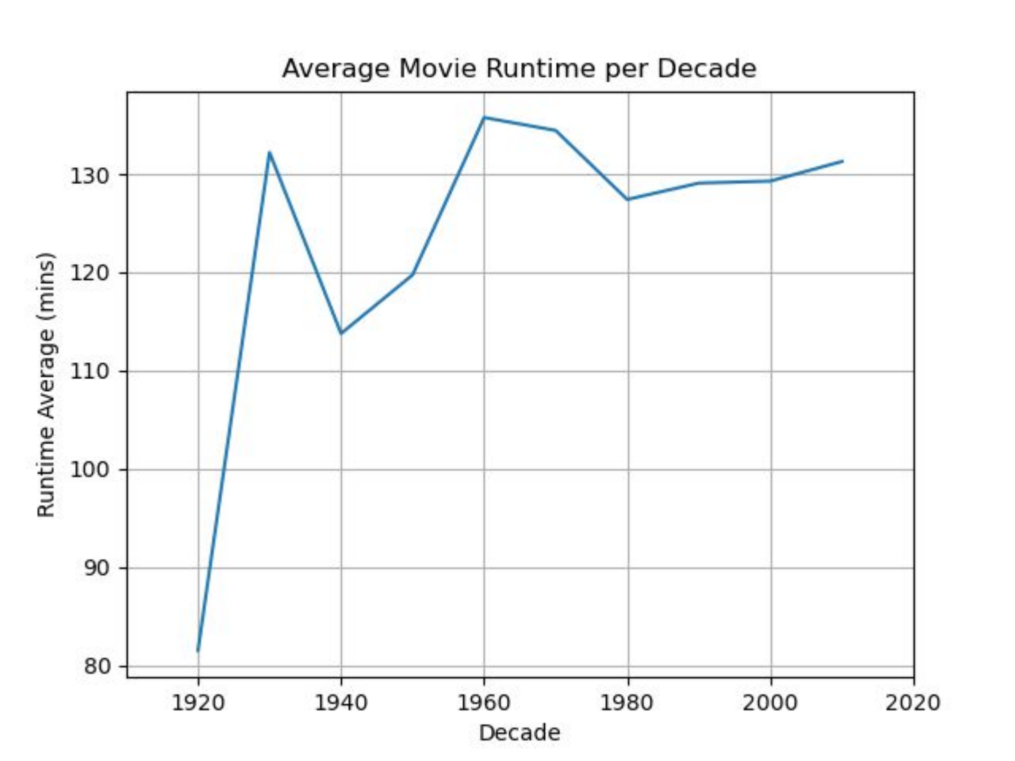
Sources:
https://sparkbyexamples.com/pandas/pandas-convert-string-to-integer/#:~:text=Alternatively%2C%20you%20can%20convert%20all,'Fee'%20column%20to%20int.
https://stackoverflow.com/questions/25668828/how-to-create-colour-gradient-in-python
https://www.tutorialspoint.com/matplotlib/matplotlib_setting_ticks_and_tick_labels.htm

Runtime vs Debut Decade



```python
#PLotting the runtime averages vs the decade the movie came out
avg_runtime = [twen_movies_avg,thir_movies_avg,four_movies_avg,fift_movies_avg,sixt_movies_avg,
             seve_movies_avg,eigh_movies_avg,nint_movies_avg,aughts_movies_avg,tens_movies_avg]

Year = np.arange(1920,2020,10)

plt.plot(Year, avg_runtime)
plt.ylabel("Runtime Average (mins)")
plt.xlabel("Decade")
plt.title("Average Movie Runtime per Decade")
plt.xlim([1910, 2020])
plt.grid()
plt.show()
```

Runtime vs Debut Decade



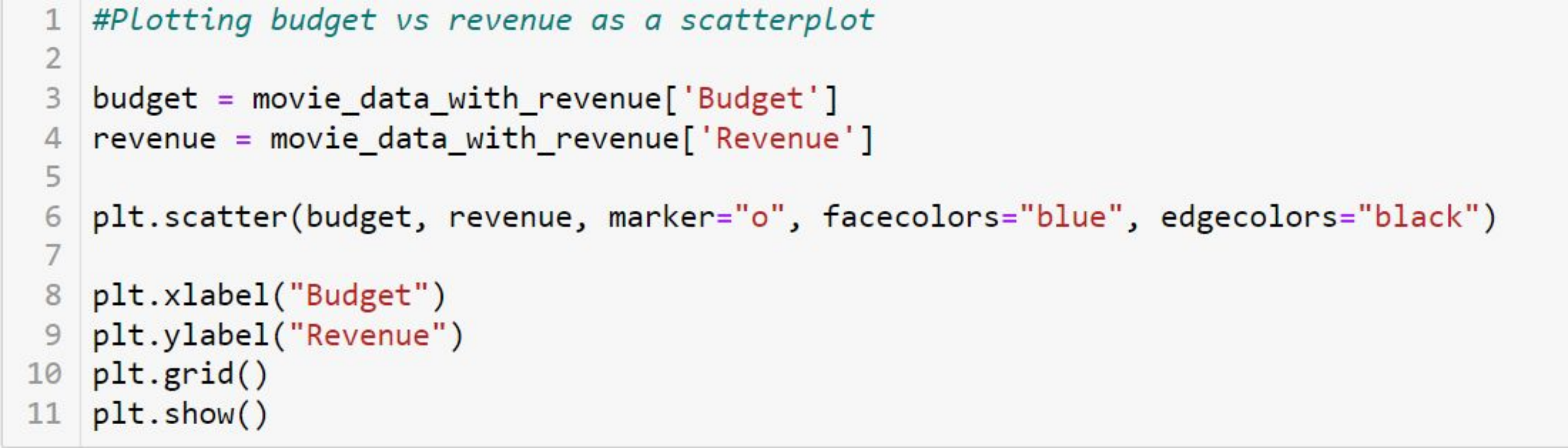
- Average runtime per each
- decade.

- Shortest runtime - 1920’s

- Longest runtime - 1960’s

- All movies remained above in hour in length.

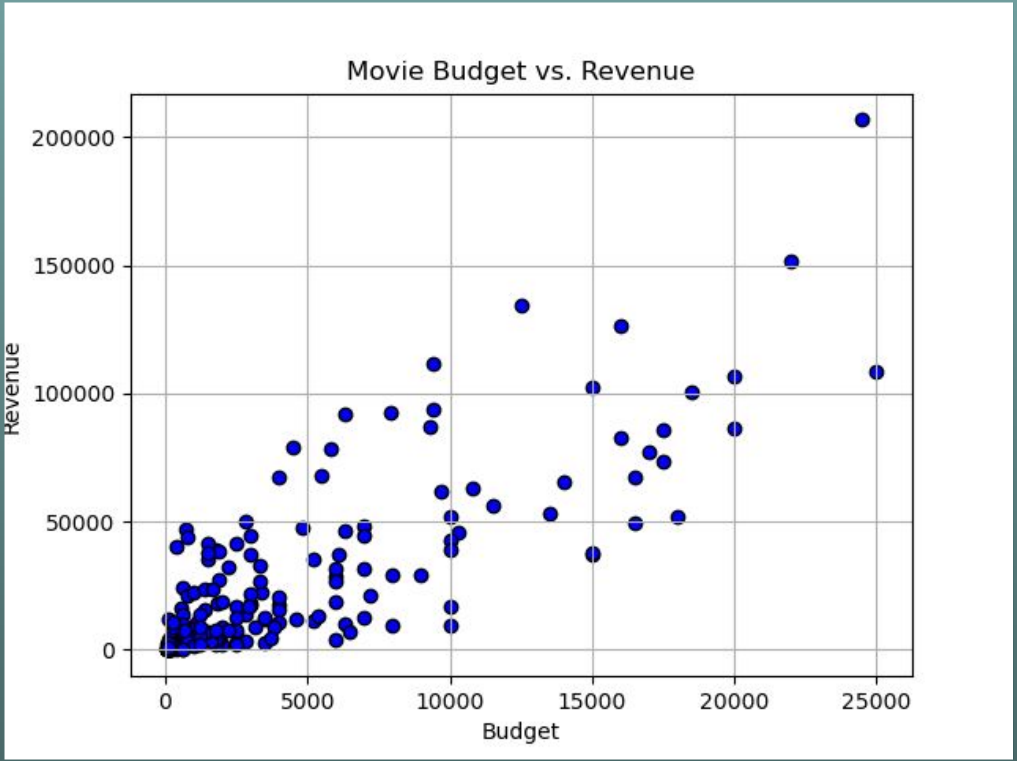
Budget vs. Revenue



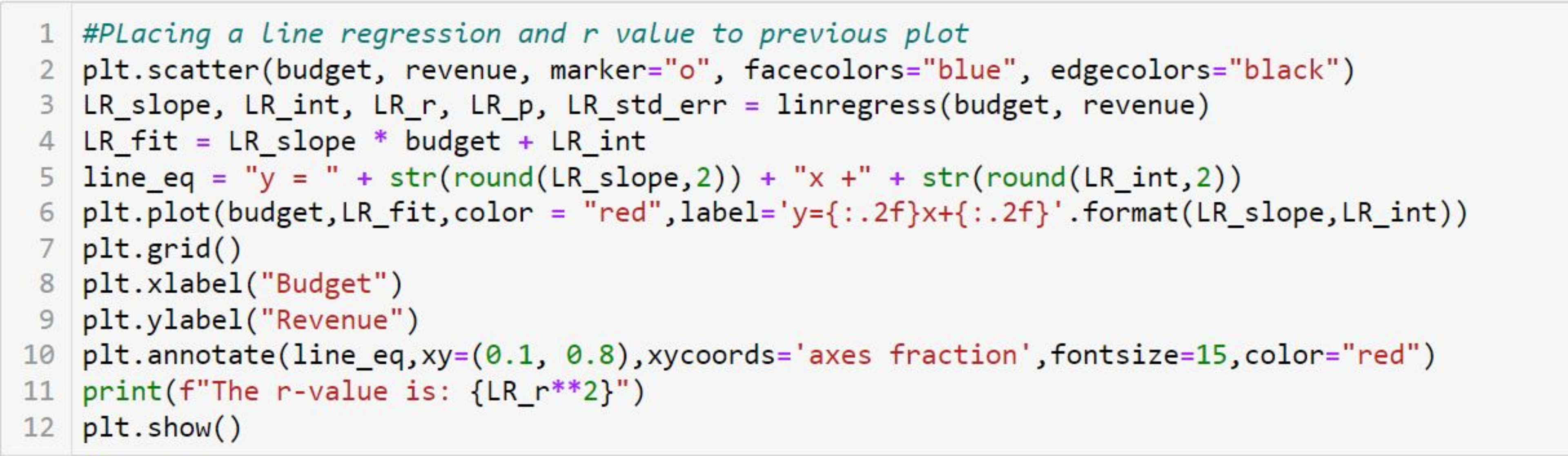
```python
#Plotting budget vs revenue as a scatterplot

budget = movie_data_with_revenue['Budget']
revenue = movie_data_with_revenue['Revenue']

plt.scatter(budget, revenue, marker="o", facecolors="blue", edgecolors="black")

plt.xlabel("Budget")
plt.ylabel("Revenue")
plt.grid()
plt.show()
```

Budget vs. Revenue

- All values divided by 10,000 for readability.
- Top earning movie earned over $2 billion, Star Wars:The Force Awakens (2015)
- Highest budget - $250 million The Dark Knight Rises (2012)
- Lowest earned $ 19,200, CIty Lights (1931)
- Lowest budget - $325,000, It Happened One Night (1934)



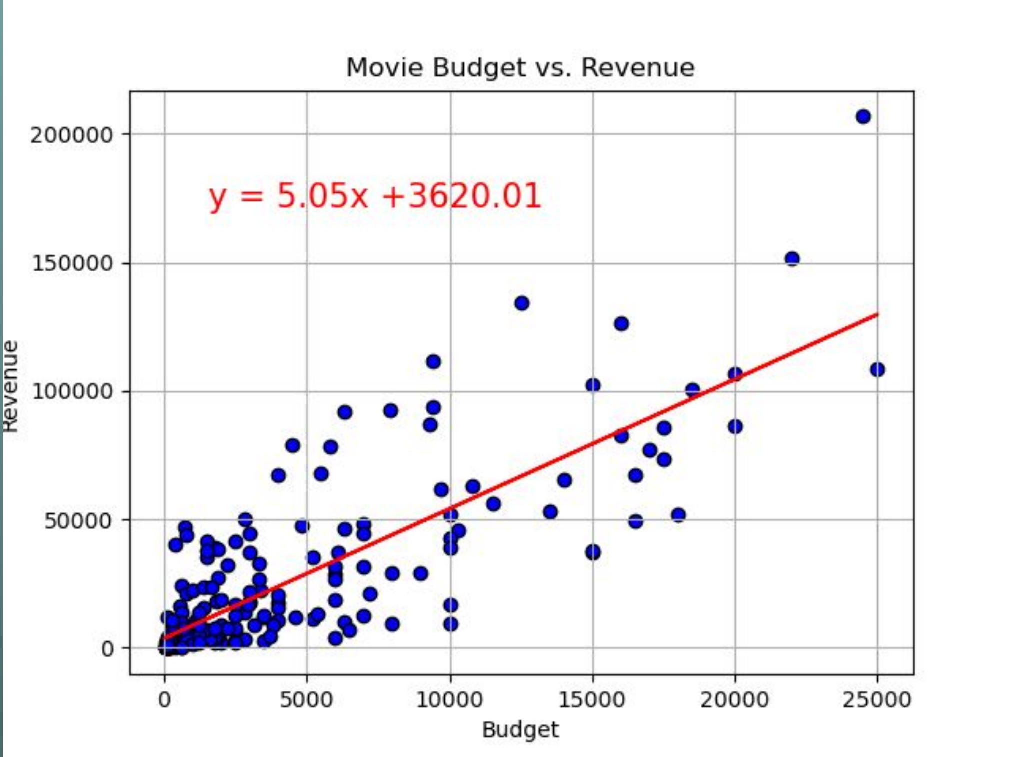
Budget vs. Revenue Line Regression

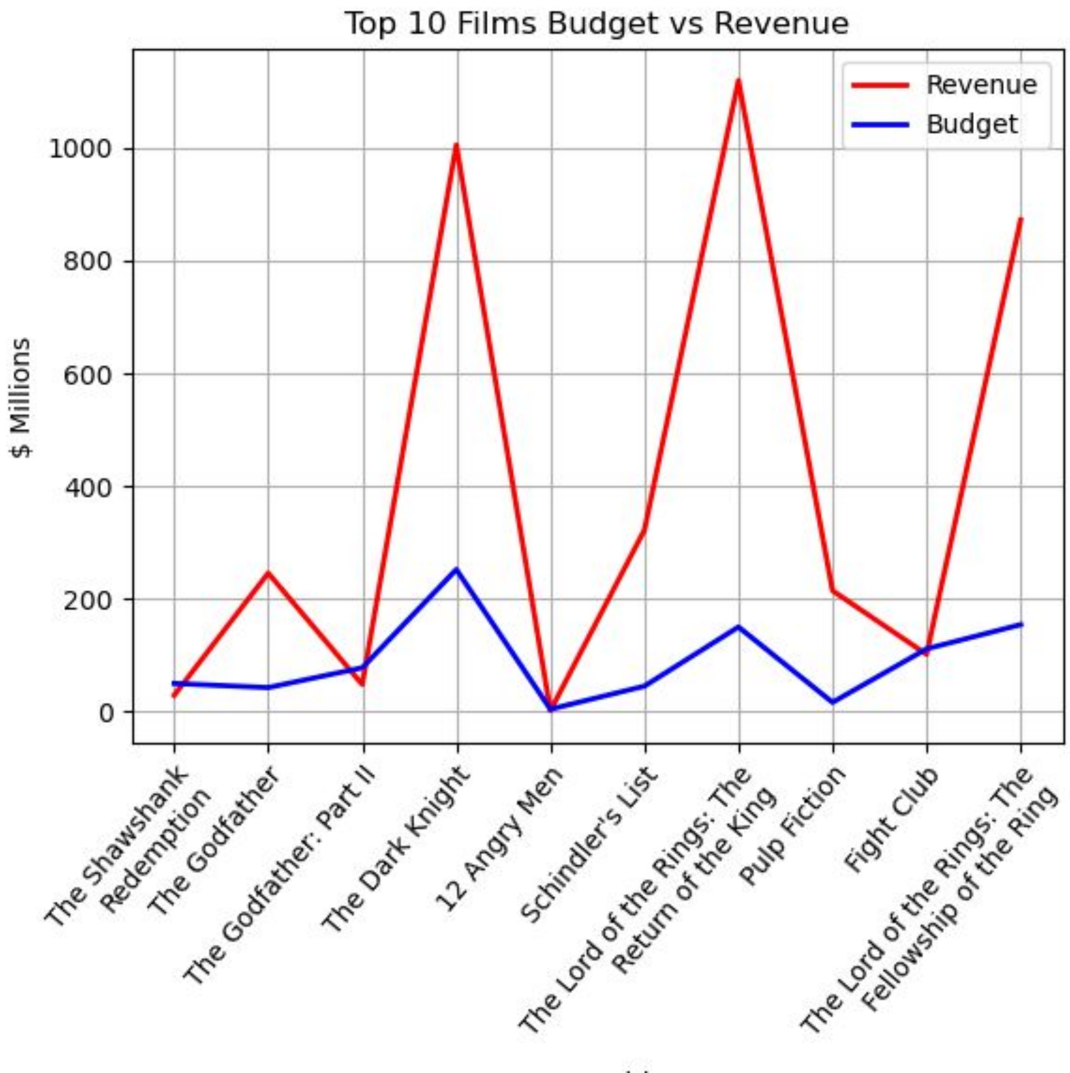


```
#PLacing a line regression and r value to previous plot
plt.scatter(budget, revenue, marker="o", facecolors="blue", edgecolors="black")
LR_slope, LR_int, LR_r, LR_p, LR_std_err = linregress(budget, revenue)
LR_fit = LR_slope * budget + LR_int
line_eq = "y = " + str(round(LR_slope,2)) + "x +" + str(round(LR_int,2))
plt.plot(budget,LR_fit,color = "red",label='y={:.2f}x+{:.2f}'.format(LR_slope,LR_int))
plt.grid()
plt.xlabel("Budget")
plt.ylabel("Revenue")
plt.annotate(line_eq,xy=(0.1, 0.8),xycoords='axes fraction',fontsize=15,color="red")
print(f"The r-value is: {LR_r**2}")
plt.show()
```

Budget vs. Revenue Line Regression

- Most movies had a budget of $ 50 million or less and revenue of $500 million or less
- R^2 value is 0.678
- Average positive correlation between budget and revenue
- As budget grew revenue also grew




Top 10 Films Budget vs Revenue
Revenue
Budget
$ Millions
0
200
400
600
800
1000
The Shawshank Redemption
The Godfather
The Godfather: Part II
The Dark Knight
12 Angry Men
Schindler's List
The Lord of the Rings: The Return of the King
Pulp Fiction
Fight Club
The Lord of the Rings: The Fellowship of the Ring

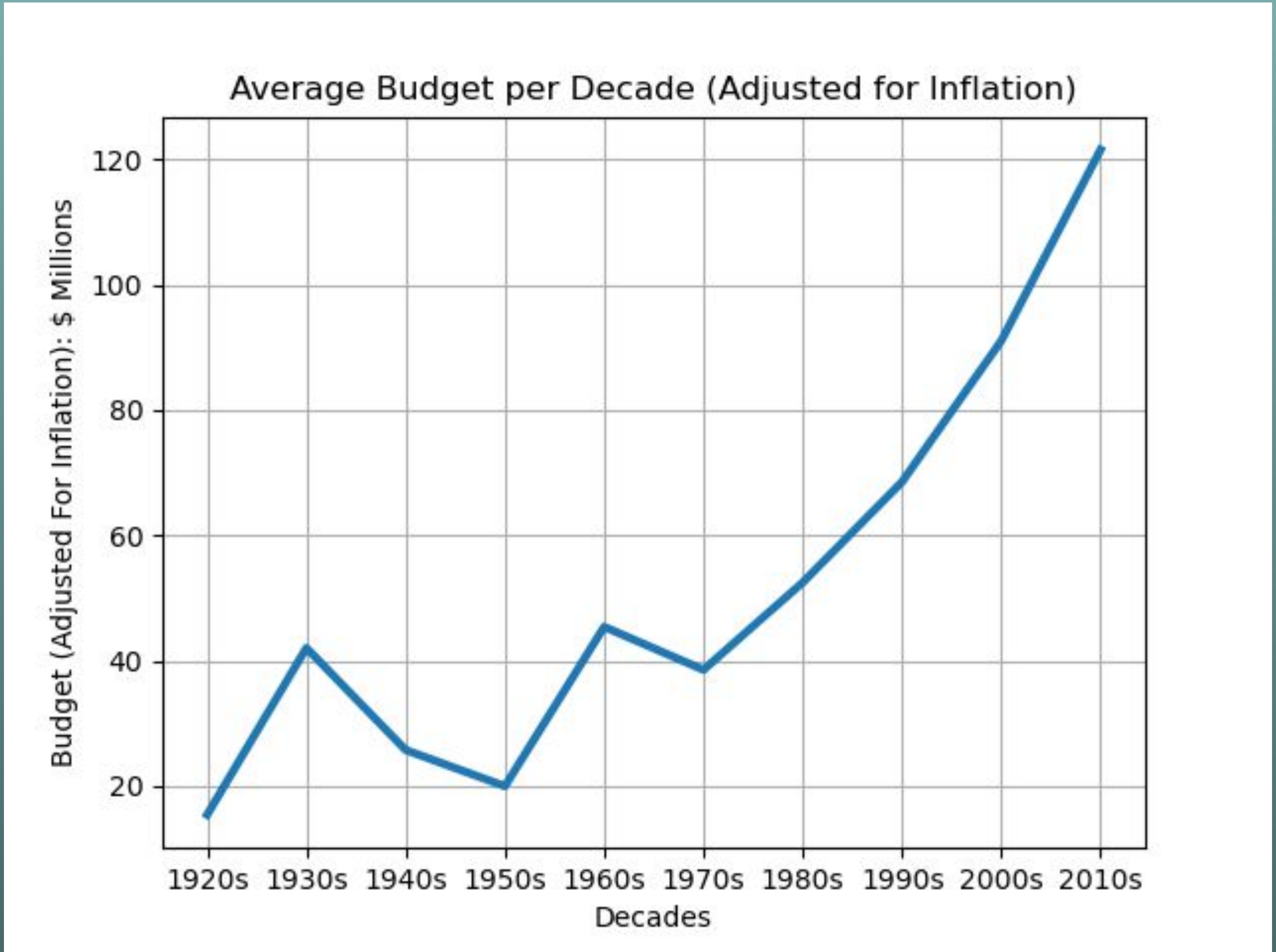
Revenue vs. Budget for the Top 10 Films

Examining the Budget vs. Revenue of just the top 10 films shows an interesting slice of films, from blockbusters, to classics to cult films.

- 12 Angry Men is considered by some to be the greatest film of all time, though it's production budget and revenue are very low in comparison to the other highest rated films.

- "Fight Club" could be considered more of a "Cult Classic". It was did earn high revenue in relation to its production cost but it is still among the top 10 films.

- "The Lord of the Rings" films were Blockbusters, earning huge revenue, far over production costs and also ranking among the top 10.

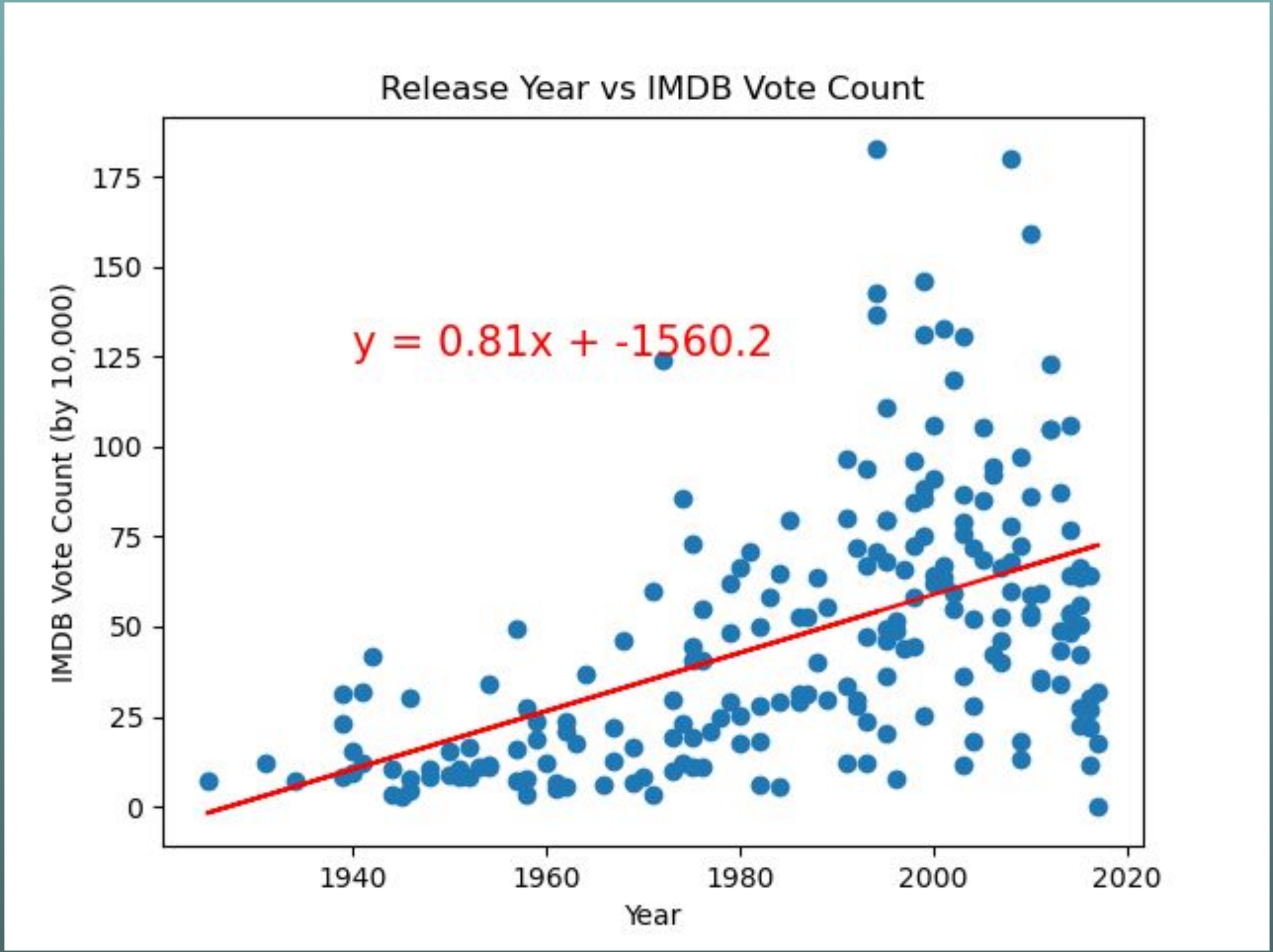

Average Budget per Decade (Adjusted for Inflation)
Budget (Adjusted For Inflation): $ Millions
120
100
80
60
40
20
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2010s
Decades

Average Budget by Decade

Historical budget numbers were adjusted for inflation using the CPI library but evenso, movies are getting more expensive with time. This could be attributed to a variety of factors, including digital effects and other new technologies as well as changes in the industry as films become more international and rental/streaming options give opportunities to earn additional revenue post-box office release.

Budget Numbers also seem to align with runtime around the 1940s, when both budget and runtime decreased and then increased to new highs in the 1960s.

Further information regarding specifics of film budgets over time could give greater insight into what is driving these trends.


Release Year vs IMDB Vote Count
y = 0.81x + -1560.2
IMDB Vote Count (by 10,000)
175
150
125
100
75
50
25
0
1940
1960
1980
2000
2020
Year

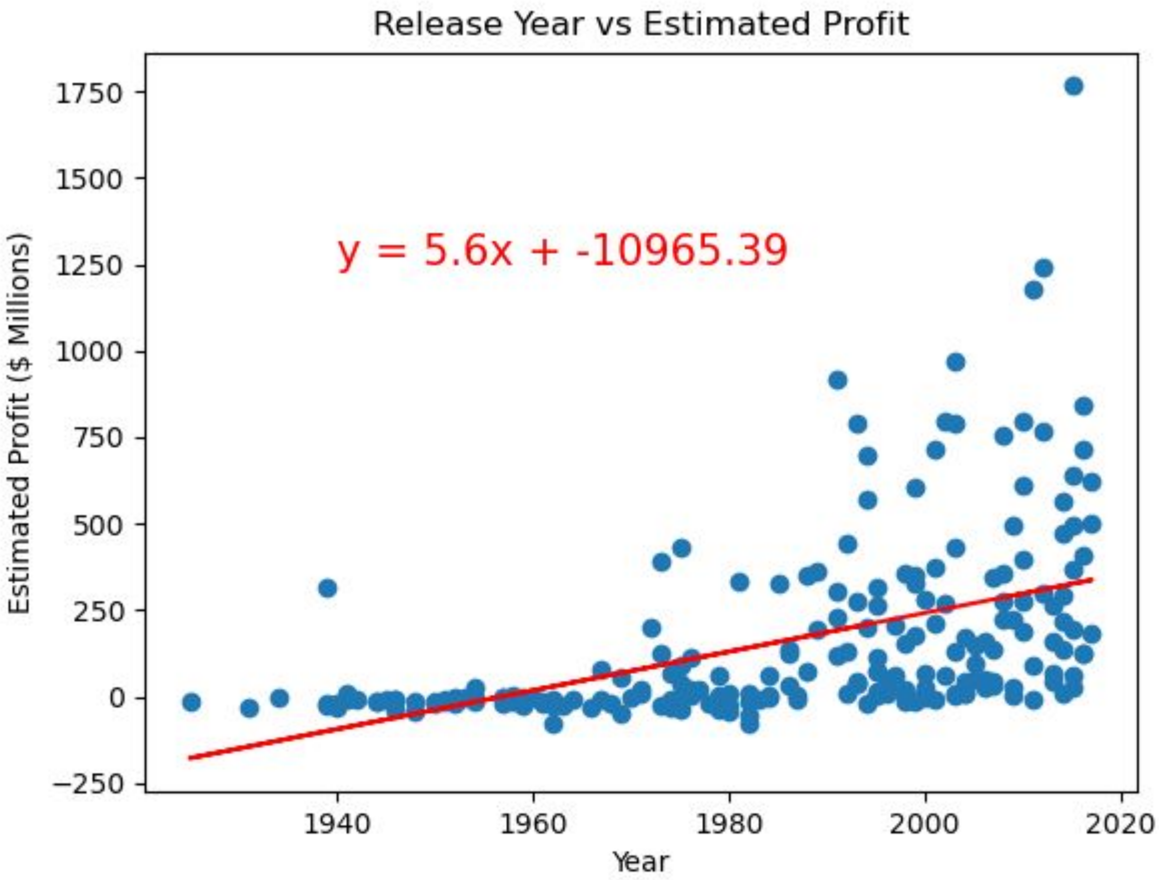
Release Date vs. IMDB Vote Count

Generally, the vote count increases with more recent films.

The r-value of .512 reflects a positive correlation but not a strong enough relationship to assume that the 2 are completely linked.

This could be possibly attributed to availability of technology and the history of IMDB.

IMDB was started in 1990, so any films released prior to 1990 had to be rated retroactively, whereas films released after 1990, could begin to be rated upon their release, which could possible garner more reviews. With time, internet access and access to films has only continued to increase, with more cable channels, DVDs, torrenting and streaming. Overall, newer releases are receiving more votes than legacy releases.


Release Year vs Estimated Profit
y = 5.6x + -10965.39
Estimated Profit ($ Millions)
1750
1500
1250
1000
750
500
250
0
−250
1940
1960
1980
2000
2020
Year

Estimating Profit Over Time

Subtracting Budget from Revenue gives a profit estimate for each film in the Top 250. The scatter plot shows increasing profits over time, including increasing instances of very high revenue films, including franchise films like "Star Wars", "The Lord of the Rings" and "Harry Potter".

The r-value of .475 implies some positive correlation between the values but it is not strong enough to assume a definite trend.

This “profit” calculation is just comparing Budget and Revenue numbers from the datasets. The actual financial information will be much more complicated and there are likely many additional expenses that could bring profit down such as marketing and promotion.

A more complete breakdown of budget and revenue numbers would be necessary to draw more than general conclusions.

Conclusion

The Top 250 Films dataset provides a useful path to examine the movie industry. Since the rating is based on viewer ratings, movies can be measured more on popularity than just on revenue, as is often the case.

Examining the most popular films over time delivers some interesting insights on Genre and Runtime.

Comparing the vote data with financials shows that there is a wide variety of combinations of Budget and Revenue for the Top 250 Films, and even for the Top 10. Overall, Budget and Revenue are climbing with the passing of time (even when adjusted for inflation) and newer Blockbuster Franchises are ranking high in both profit and viewer votes.