The movies dataset

The questions that I asked and answered

1. How has the popularity of various movie genres evolved over the years in terms of revenue and ratings?

2. Which production company holds the highest popularity, revenue, and vote average?

3. Which movie genres produce the most revenue and receive the best audience ratings?

4. Which movies rank highest for revenue, ratings, and average votes within each genre?

What I did to analyze the data

Firstly (Data Wrangling):

I import the libraries that I will use and import the dataset

Check the type of the data and its shape to know how the data is formed

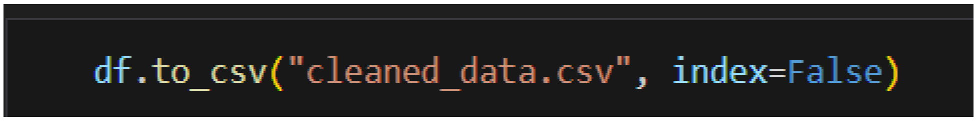
Secondly (Data Wrangling):

I check the null values the 0 values the outliers and the duplicated values

Thirdly (Data Cleaning):

After I finished the data wrangling, I entered the data cleaning stage data cleaning I deleted the outliers (because when I tried to replace them with the mean the outliers were changing every time I ran the code and the median too, so I deleted them)

Delete the unused columns (id, imdb_id, budget_adj, revenue_adj, tagline, and release date) remove the duplicated column replace and delete the null values, and delete and replace the 0 values. at the end of this stage, I made a backup for the cleaned data



```
df.to_csv("cleaned_data.csv", index=False)
```

Because when I used it in the next stage the data was changed (the columns were changing)
So I make this backup to make another backup to work on it without changing the original data and to avoid the accident

Fourthly (Exploratory Data Analysis):
Before I worked on any questions, I made a copy from the original data to avoid accidents and changed the original data

The limitations:

Limitations

During the data analysis process, several limitations could impact the findings' accuracy and reliability.

1. **Data Quality Issues:** The dataset contained many missing values, particularly in critical columns such as "release date" which limited its usability in the analysis. Additionally, columns like "budget"

and "revenue" had numerous zero values, making it difficult to derive meaningful insights.

2. **Data Consistency:** There were inconsistencies in the dataset, with some values appearing unreliable or incomplete. This required additional cleaning steps and, in some cases, led to the removal of certain data points.
3. **Sample Size Limitations:** The dataset may not be large enough to fully support deep analysis, which could affect the generalizability of the conclusions.
4. **Potential Bias in Data Collection:** Since the dataset was pre-collected from various sources, there is a possibility of inherent biases that might have influenced the results.

These limitations should be considered when interpreting the findings, and future research could benefit from a more comprehensive and higher-quality dataset.

Conclusion

In conclusion, this analysis provided insights into key trends in the movie industry based on the available data. However, it is important to acknowledge the limitations of the dataset. One major limitation was the presence of missing or zero values in critical financial columns, which may have affected the accuracy of the findings. Future research could benefit from a more comprehensive dataset that includes qualitative factors such as audience sentiment and critic reviews to enhance the depth of analysis.

Another way:

Limitations

This analysis encountered several data quality challenges that significantly impacted the accuracy and depth of insights. One of the critical issues was the presence of numerous missing or zero values in key financial columns, such as revenue and budget. These anomalies distorted statistical measures and made it difficult to extract meaningful insights regarding a movie's financial success.

Additionally, the dataset size may not have been large or diverse enough to capture broader industry trends accurately. The results might vary if a different sample was used, introducing uncertainty in the estimates. Moreover, there may be biases in data collection, as some independent or low-budget films could be underrepresented, affecting the generalizability of our findings.

Another key limitation is that this analysis primarily focused on numerical and quantitative aspects. Incorporating qualitative factors, such as audience sentiment, critic reviews, and genre-specific preferences, could have provided deeper insights into movie performance. However, due to the structure of the dataset, these elements were not explored in depth.

Lastly, while correlations between variables were analyzed, it is important to acknowledge that correlation does not imply causation. The findings do not establish a direct cause-and-effect relationship between variables, and further research would be required to determine any causal links.

The References:

Kaggle.com

Pandas documentation

Numpy documentation

Matplotlib and Seaborn documentation

Cheat sheet from Google

Stack Overflow