Final Project by Muthukumar Kadhirvel

Muthukumar Kadhirvel

2022-06-02

# Introduction

## We are using the TMDB 5000 Movies dataset and OMDBAPI and Wikipedia link of highest grossing films to collect more information about the various movies and the correlation between the various parameters.

# Preparation

## We take the following steps on the API/CSV/WEB files

## Drop unnecessary columns

## Rearrange and replace headers

## Find Duplicates

## Format data into readable format

## Identify outliers

## Then we combine the datasets into an SQLite table via SQL JOINS and then put it into a pandas data frame and create data visualizations.

# Challenges

## The scraping of the website data took the most time as the table was in a different format and we had to undertake additional steps to achieve the necessary output.

## Learning

# Getting data from APIs was a huge learning along with scraping data from webpages. Reading CSV files was something that we had done in previous courses also and hence it was easier.

# Ethical Implications

# Since our topic was with respect to movies the ethical implications were less severe compared to more sensitive topics like healthcare/customer information. But still scraping website data is not ethical and we have seen web scraping return information having warnings that high usage of web scraping can lead to the IPs blocked from retrieving more data.

# Conclusion

## Using publicly available data from various sources like API/CSV/WEB we were able to build a huge database of movies released and retrieve various parameters like Release Date/Budget/Revenue/Runtime/Rating/Votes/Language/Country etc. We were able to use that information to visualize the data and arrive at conclusions like how budget of a movie affects the revenue. We were able to see that budget and runtime are more connected than budget and revenue which has more outliers. We can perform much more visualization and analysis with the data we have and something that we can revisit in future courses.

# References

https://www.omdbapi.com

https://en.wikipedia.org/wiki/List\_of\_highest-grossing\_films

https://en.wikipedia.org/wiki/List\_of\_highest-grossing\_non-English\_films

https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata