Project: An investigation into a TMDB dataset

Project Overview

In this project, I analyzed a TMDB dataset. I used the Python libraries NumPy, pandas, and Matplotlib to make the analysis easier.

The following libraries are installed:

- pandas
- NumPy
- Matplotlib
- csv

I also used Anaconda, which comes with all of the necessary packages, as well as Jupyer notebook.

The TMDb movie dataset Contain:
Total Rows = 10866
Total Columns = 21

Data

This project contains 3 files:

- tmdb-movies.csv : The dataset file containing 10,000 entries of movies.
- report.ipynb: The investigation of the dataset has been done in this jupyter notebook file.
- export: PDF file of notebook.

Dataset file

This data set contains information about 10,000 movies collected from TMDb. Contains data such as title, cast, director, runtime, budget, revenue, release year etc.

- Certain columns, like ‘cast’ and ‘genres’, contain multiple values separated by pipe (|) characters.
- There are some odd characters in the ‘cast’ column. Nothing to care much of, I leave them as is.
- The following column 'imdb_id', 'vote_count', 'budget_adj', 'revenue_adj', 'homepage', 'keywords', 'overview', 'tagline', 'production_companies', and 'vote_average' are deleted from the dataset.

The following Questions are asked

- Which year has the highest release of movies?
- Which movie has the Highest or lowest profit?
- Which movie has the Highest or lowest Budget?
- Which movie has the Highest or lowest revenue?
- Which Year Has The Highest Profit Rate?
- Average Runtime of movies?
- Which genre has the highest release of movies?
- Most frequent star cast?