ITEC 4210

Information Analytics

Digital Story Report

The Movie Database

Spring Semester

2025

**Summary of Data Source**

The data source chosen for this project was “The Ultimate 1 Million Movies Dataset (TMDB + IMDb).” It contains all of the films in The Movie Database (TMDB), along with their titles, release information, genres, ratings and popularity (from IMDb, the Internet Movie Database), production companies, budget, revenue, etc. The reason this dataset was chosen was threefold: First, the vast amount of records gives a large sample size to work with and create accurate conclusions. Second, there are enough continuous data points to have concrete and quantifiable analysis. Lastly, there are a plethora of nominal data points on which to analyze.

**Summary of Data Cleaning & Formatting**

In order to get a dataset for analysis, there was a lot of cleaning and transforming. First, I dropped all the columns that would not be used in this analysis, like poster path, overview, tagline, id, etc. Then, I changed the release date to just show the year instead of being formatted as yyyy/mm/dd. I then turned the “genres,” “production\_companies,” and “production\_countries” columns into lists. After that, I qualified the dataset in order to weed out outliers and films that should not be used for this analysis. Every film has to have over 20,000 ratings, must be released after 1990 to ensure that these trends are current, the runtime must be over 60 minutes to ensure that only feature-length films are included, and “Documentary” must not be a genre. From this qualified dataset, I made three different dataframes, where in each I “exploded” either the genre, production company, or production country column. Exploding a column makes it so for each different value in the list, it creates a new record for that film with that value, to allow for accurate genre/production analysis. Lastly, I sent the original qualified dataset along with the three different analysis data frames to CSVs.

**Main Objective Explored**

Explore and analyze profits in the film industry based on genre, production company, and production country. With this project we aimed to give a hypothetical film director insights into profit trends to highlight which companies and countries would be better to pursue.

**Summary of Visuals**

For the visuals of our digital story, we created a dashboard mock-up to encompass all visuals as a whole, this way the message could be conveyed in one single view.

We added some cards near the top to display some basic metrics like unique movies analyzed from our final subset, unique genres, and unique producers. We then proceeded to display 4 different visuals.The first visual was a table including top 10 production companies by average profit. The visual per-se is a table, although we did some formatting to the data bars to make it display as a type of horizontal bar chart. Then we displayed a bar chart to show the top 3 producing countries by profit from our dataset, very simple and straightforward. Followed by a pie chart of top 5 movie genres by profit. Here we see the profit brought in by these top genres, as well as the percentage they account for. Finally a scatter plot to show some sort of relation, if any, between budget and revenue, to which we could see that there was an existing positive correlation between these two variables in the cleaned dataset.

**Contributions**

On this project each team member contributed in their own important way, which was the plan agreed upon by the team initially when the work was distributed for improved efficiency. While John did an outstanding job with the vast majority of data cleaning, Caleb focused on putting the visuals together from the resulting clean dataset. It is important to point out that both team members were very much present with feedback and ideas throughout all stages of this digital story. Deliverable work (digital story presentation & digital story report) was done by both individuals to be able to have the work done by the stipulated due date.