**Final Project: Data Preparation**

Kannur, Gyan

Bellevue university - Master of Science in Data Science

Course Name: DSC540

Assignment: Final Project Milestone 1

Instructor: Catherine Williams

# **Topic: Top 500 Movie Analysis**

Every year, thousands of movies get released but only a percentage of those become successful. The aim of this work is to analyze the movie profitability and commercial success. I have considered the top 500 movies and analyze on which occasions have these made it to the top.

## **Datasets**

As part of first milestone of this term project, I have considered three difference data sources that have different file types of information as mentioned below. I also established the relationship between them.

### **CSV file data source**

First data source is csv file containing top 500 movies between 2014 and 2023. Below is the datasource url

<https://www.kaggle.com/datasets/mitchellharrison/top-500-movies-budget>

### **API**

Second data source is API as mentioned below it returns the latest movie popularity, vote average and their counts.

[https://api.themoviedb.org/3/search/movie?<api\_key](https://api.themoviedb.org/3/search/movie?%3capi_key)>

### **Website data**

Third data source is website data. Take the root of the url, in this case <https://www.boxofficemojo.com> and add '/robots.txt' to the end. This will come up with a page that shows what type of web scraping is allowed or disallowed. Thankfully Box Office Mojo allows all.

Website Link: <https://www.boxofficemojo.com/weekend/by-year/><year>

## **Metadata and relationship for datasets**

### **CSV Dataset**

#### **Metadata**

The dataset has following fields

|  |  |
| --- | --- |
| **imdbID** | tt2953050 |
| **primaryTitle** | Encanto |
| **originalTitle** | Encanto |
| **startYear** | 2021 |
| **endYear** | \N |
| **runtimeMinutes** | 102 |
| **genres** | Animation,Comedy,Family |
| **averageRating** | 7.2 |
| **numVotes** | 254412 |

#### **Relationship**

“**originalTitle**” field which represents the movie title that is used as a common key to make a join with other two datasets.

### **API Dataset**

TBDB api key

#### **Metadata**

|  |  |  |
| --- | --- | --- |
| adult | integer | FALSE |
| backdrop\_path | string | /3G1Q5xF40HkUBJXxt2DQgQzKTp5.jpg |
| genre\_ids | list | [16, 35, 10751, 14] |
| id | integer | 568124 |
| original\_language | string | en |
| original\_title | string | Encanto |
| popularity | string | The tale of an extraordinary family, the Madrigals, who live hidden in the mountains of Colombia, in a magical house, in a vibrant town, in a wondrous, charmed place called an Encanto. The magic of the Encanto has blessed every child in the family—every child except one, Mirabel. But when she discovers that the magic surrounding the Encanto is in danger, Mirabel decides that she, the only ordinary Madrigal, might just be her exceptional family's last hope. |
| overview | integer | 246.406 |
| poster\_path | string | /4j0PNHkMr5ax3IA8tjtxcmPU3QT.jpg |
| release\_date | date | 10/13/2021 |
| video | string | Encanto |
| vote\_count | Integer | 7.599 |
| adult | boolean | FALSE |
| backdrop\_path | string | /3G1Q5xF40HkUBJXxt2DQgQzKTp5.jpg |

#### **Relationship**

“**originalTitle**” field represents the movie title which is used as a common key to make a join with other two datasets.

### **Website Dataset**

#### **Metadata**

The metadata for the website dataset has been mentioned below. This contains information about major cryptocurrencies.

|  |  |  |
| --- | --- | --- |
| date | date | Release year |
| occasion | String | Ex. Thanksgiving week |
| top10\_gross | String | Gross in USD |
| top10\_wow\_change | String | Represented as a % |
| overall\_gross | String | Gross in USD |
| overall\_wow\_change | String | Represented as a % |
| num\_releases | String | Number of releases |
|  |  |  |
| top\_release | String | The top movie name |
|  |  |  |
| week\_no | String | Week of the year |

The screenshot from website for the month of Dec 2023.

A screenshot of a computer

Description automatically generated

#### **Relationship**

“**#1 Release**” which is the movie title which is used as a common key to make a join with other two datasets.

## **Approach**

The main objective of this project is to analyze these datasets containing information movies for the last few years, collected from The Movie Database (TMDb) and including all films details such as the production cost, the revenue generated, rating information, etc. This work also tries to find answers to the questions below:

1- What kinds of properties are associated with most and less successful movies?

Which Movie had the highest or lowest profit?

Which year the movie industry\*\* made the highest profit?

Which Month the movie industry made the highest profit?

Do popular movies get higher profit?

What were the most or least expensive movies?

What is the statistical relationship between budget and profit?

Do movies with highest budget get highest rating?

Which Movie had the highest or lowest revenue?

Is there any statistical relationship between revenue and profit/ revenue and budget?

What is the movie length most liked by the audience?

Which movie was high or less rated?

Do high rated movies get higher profit?

2- What are the Top 10 movies according to different features ? in particular :

Profit

Budget

Revenue

Popularity

3- Which genres are most popular and profitable overall and overtime?

Which genres are more profitable overall?

Which genres are more profitable from year to year?

Which genres are more popular overall?

What is the evolution of the genres according to popularity from year to year?

4- What are the visuals which can best depict my analysis ?

## **Challenges**

Below is the list of challenges I faced while creating the datasets.

1. For CSV file, I struggled getting the top movies as some of the datasource just had 5000 movies without any budget of millions of records but nothing in terms of the gross turnover.
2. For the api, I initially signed up for an account in imdb for an api key.The api had limitations on the number of requests I can make, therefore I signed up with tmdb api key which was far more flexible .
3. Finding a suitable website and getting the HTML object was easy. But scanning through the HTML content and identifying the right tag to start scraping was initially a challenge
4. Either with scraping or with api, one of the major hurdles is to handle pagination.
5. The website has a year dropdown and it was observed that the html tags were different for some years due to which parsing the required data was a challenge.
6. Thankfully, every datasource had a title which is the main key for my entire analysis.

## **Learning**

I learnt that there are multiple options for datasources, not just to rely on flat file, if there is additional data required, I could always query the api. I also learnt that many websites do allow the option for webscraping. I intend to explore deeper into the python libraries for api and webscraping along with using pandas for my analysis.