NOVA IMS
Information Management School

INSIDE IMDb

DATA VISUALIZATION

GROUP AB

MARCH 2021

IMDb

MASTER IN DATA SCIENCE AND ADVANCED ANALYTICS
WITH A MAJOR IN BUSINESS ANALYTICS

CATARINA CANDEIAS, M20200656
CATARINA URBANO, M20200607
REBECA PINHEIRO, M20201096
RITA FERREIRA, M20200661

INTRODUCTION

The main objective of this project was to build a dashboard that provided interactive visualizations using *Plotly* with Dash software as our tool, in which the users could obtain information regarding the cinematographic world as well as movies suggestions according to their preferences.

The information obtained by the IMDb website was displayed by interactive visualizations that can be updated according to the user interaction. In short, this dashboard illustrates all the information associated with the movies as actors, directors, ratings, among others. To enable the creation of this interactive application, we made use of Kaggle's dataset which is described below in this report.

All the steps taken, from data treatment to *Plotly* interactive visualizations and app styling, can be found in the provided *GitHub* repository. To finalize the building of this application, we deployed it to Heroku to make it accessible to all the users.

LINK APP: https://inside-imdb.herokuapp.com/

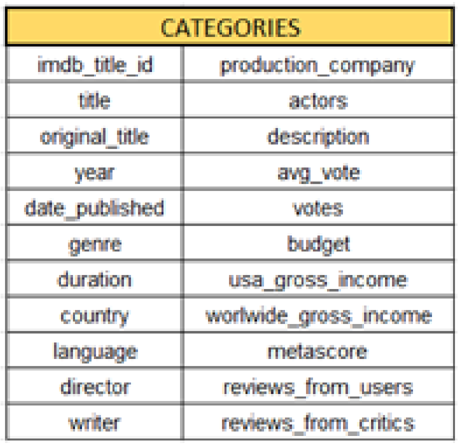
LINK GITHUB: https://github.com/rita-ilda/IMDB_Dash_App_DV_project

DATASET DESCRIPTION

LINK DATA SOURCE: https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

The *IMDb movies* csv file was the most used in the process, which included 85,855 movies from the year 1920 to 2020 (although we focused our attention from the year 1970 on) with attributes such as movie description, average rating, number of votes, genre, among others. The variables it holds are displayed in *Table 1*. It is important to note that in addition to having to preprocess and transform some of the data, this also required some data cleaning mainly to deal with missing information, being the reason why we disregarded the older movies, and some others were not included in our work.

Additionally, to get more information regarding certain movies, in specific to get the movies covers URL we used their IMDb's title id by applying the *IMDBpy* package.



CATEGORIES	
imdb_title_id	production_company
title	actors
original_title	description
year	avg_vote
date_published	votes
genre	budget
duration	usa_gross_income
country	worlwide_gross_income
language	metascore
director	reviews_from_users
writer	reviews_from_critics

Table 1 - The movies dataset variables

PROJECT INSPIRATION

Considering the IMDb's website as our role-model and the fact that this domain is one that we really appreciate and identify ourselves with, made it even more enjoyable to develop this application. Our application's main purpose is to analyze and provide different aspects about that matter, as the movies released per year(s), the main actors and directors, the most popular movie genre, the movies' ratings and the movies' profits and budget.

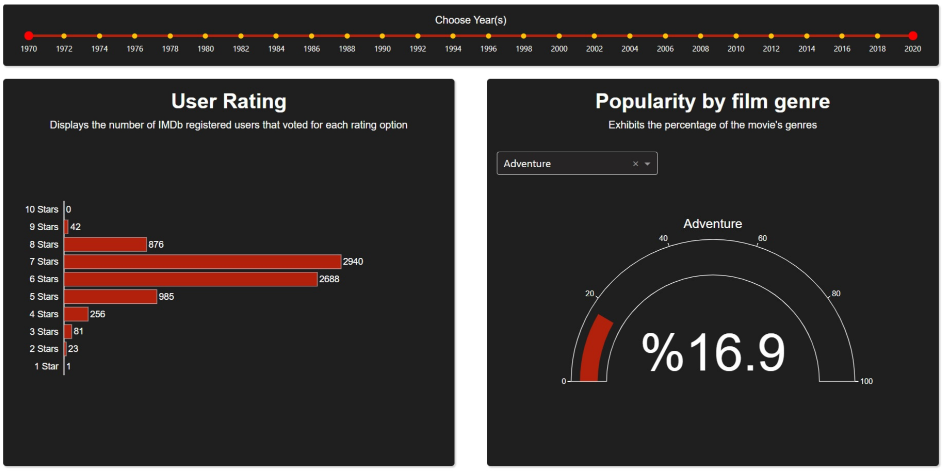
Our dashboard was designed to all movie lovers or even the ones who only want to explore the data available and get insights from it, keeping also the IMDb's main purpose. Additionally, our app provides a suggestion of five movies according to the user preference, being this achieved through the user filtration, which vary according to their curiosity and interests. It is also important to mention that the dashboard's colors were inspired by the IMDb website, to maintain the consistency and coherency.

TECHNICAL ASPECTS AND VISUALIZATIONS

The dashboard is divided into two principal parts separated by the tabs, *Discover IMDb* and *Search Movie*, respectively. The header is composed of a brief description regarding the usefulness of this application, in order to contextualize the user, the IMDb logo and the source where the data was acquired. Furthermore, one slight detail to highlight is the web page name and the icon displayed, which makes it even more associated with IMDb organization and the purpose of this application.

Concerning the first tab, *Discover IMDb*, it provides some insights about the movies covered on the IMDb website, which were released since 1970's years until now. As one crucial aspect is regarding the interactivity of the application, the user can decide if he/she wants to select the range of years or one year specifically to understand the data that IMDb provides.

The six visualizations available study different aspects concerning the years specified. The first one, *User rating*, shows how the users' movies classifications were distributed on a 1-10 star rating, with 1 as the worst one and 10 as the best one, in a bar chart. Regarding the second visualization, *Popularity by film genre*, it displays the percentage of movies released in each genre, in the format of an indicator graph, where the user has the possibility to select a category of its choice using the available dropdown.


Choose Year(s)

User Rating

Displays the number of IMDb registered users that voted for each rating option

Popularity by film genre

Exhibits the percentage of the movie's genres

Adventure

Image 1 - *User Rating* and *Popularity by film genre* visualizations

The third visualization, *Actors with the most movies made*, is a table that presents the ten actors that participated in more movies, in descending order, and the corresponding value. Additionally, in the same row of the dashboard, another visualization is presented, the fourth one, *Movies and Directors*, which shows the ten highest rated movies and their directors. By clicking in the center of this sunburst chart is possible to isolate the director and movie(s) pair for better visualization. Furthermore, by moving the cursor over the different movies and directors is possible to read the information regarding the average user's vote.

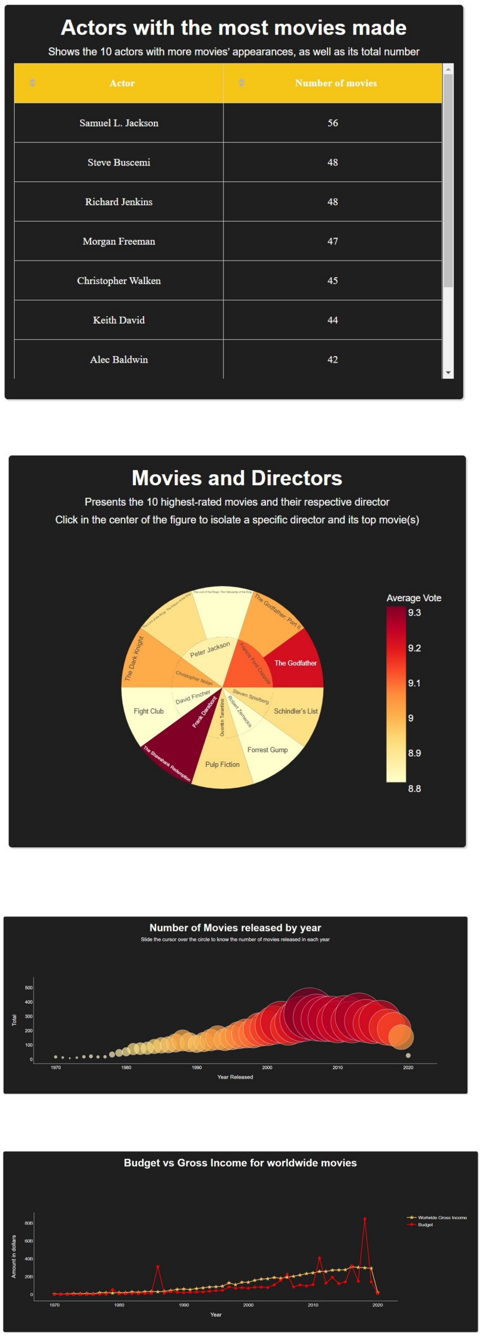
The fifth visualization, *Number of Movies released by year*, exhibits in a scatter graph the amount of movies released per year in the considered period of time, where the marker's sizes are proportional to that value (if the user moves the cursor over the bubble, this value is shown).

The last visualization developed, *Budget vs Gross Income for worldwide movies*, compares two different financial information, in a line chart, that is, the movies´ budget and their worldwide gross income, in billions of dollars, per year. If the user prefers to see only the budget or only the income, it is possible to inactivate one of them by clicking on their label located on the right. Following the same idea as the graph mentioned above, by passing the cursor over the star marker it is possible to discover the dollar amount in each year.

The second tab, *Appendix 1*, informs the movies released after the user fills all the filters according to their choice.

For the first and second filters, *Genre* and *Year*, it is possible to input multiple choices while in the third and fourth, *Rating* and *Sort*, only one option can be selected.

The five movies displayed are sorted according to the user selection and come with the following information: title, year, movie cover, description, genre, rating, duration and language.



Actors with the most movies made

Shows the 10 actors with more movies' appearances, as well as its total number

Actor	Number of movies
Samuel L. Jackson	56
Steve Buscemi	48
Richard Jenkins	48
Morgan Freeman	47
Christopher Walken	45
Keith David	44
Alec Baldwin	42

Movies and Directors

Presents the 10 highest-rated movies and their respective director

Click in the center of the figure to isolate a specific director and its top movie(s)

Number of Movies released by year

Slide the cursor over the circle to know the number of movies released in each year

Budget vs Gross Income for worldwide movies

Image 2 - The remaining visualizations

DISCUSSION AND FINAL CONSIDERATIONS

This project proved to be a challenging journey with many setbacks along the way. Nonetheless, we believe we achieved excellent results, having developed a user-friendly dashboard, with pertinent and interactive visualizations, a movies search engine, and an appealing design.

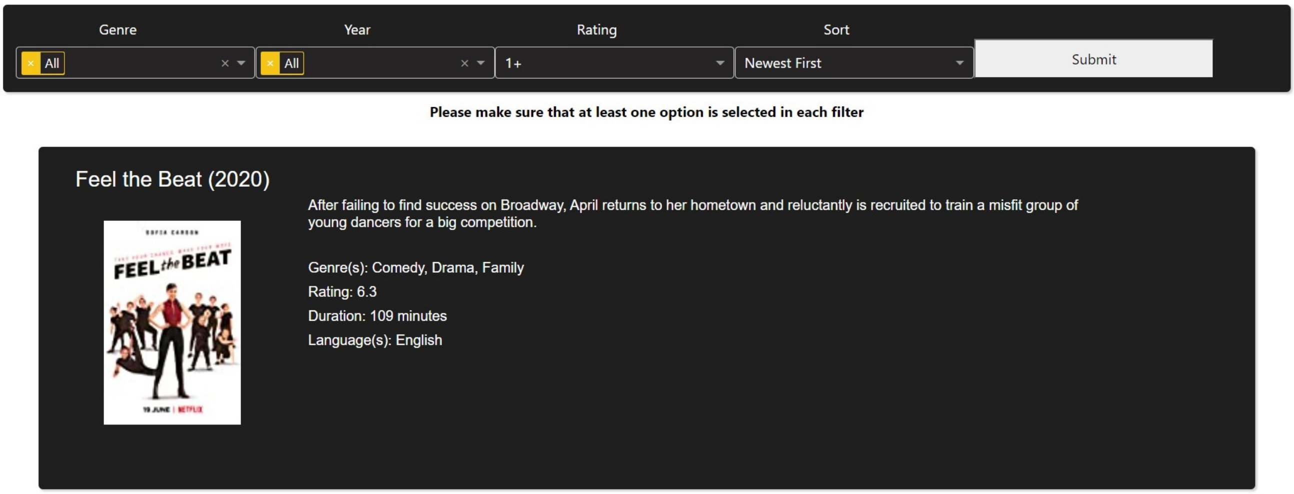
Some of the difficulties and limitations we faced were the replicability of our application in devices with different screen sizes, which in future work might be something to develop further (e.g., adapt it for small devices as smartphones). Also, the time it takes to run our application in the backend is something that could be optimized, along with the retrieved images' quality of the IMDb's movies' cover. An additional limitation we had, was the implementation of some ideas using libraries besides *Plotly* to make our representations.

Moreover, building this application made it possible to dive into a subject that everyone enjoys, which facilitated all the process involved and the intended storytelling.

REFERENCES

1. IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows
2. W3Schools Online Web Tutorials

APPENDIX



Genre: All

Year: All

Rating: 1+

Sort: Newest First

Please make sure that at least one option is selected in each filter

Feel the Beat (2020)

After failing to find success on Broadway, April returns to her hometown and reluctantly is recruited to train a misfit group of young dancers for a big competition.

Genre(s): Comedy, Drama, Family

Rating: 6.3

Duration: 109 minutes

Language(s): English

Appendix 1 - Movie Search Tab