Projeto Final
MC536

Daniel Credico de Coimbra – 155077
Gabriel Bonfim Silva de Moraes – 216111
Victor Durço Gomes Bijos – 206508

Resumo do Projeto

Dataset que integra informações de diversas fontes sobre os 25 filmes com maior bilheteria de cada um dos últimos 50 anos.

Cada filme está relacionado a gêneros e estúdios, além de ser informado características como ano, país de origem, avaliação crítica (IMDb e Metacritic), bilheteria, bilheteria ajustada pela inflação, número de bilhetes vendidos, e código IMDb.

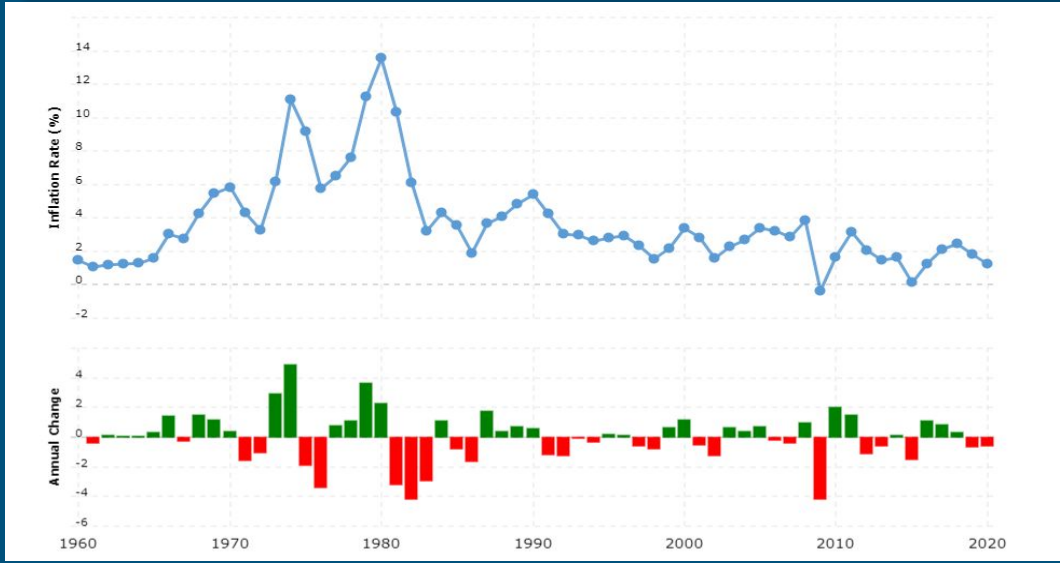
Evolução do Projeto

- *Brainstorm* - Pandemia, jogos, segurança pública e enfim, filmes;
- Projeto inicial - apenas filmes e bilheteria, depois acrescentamos número de bilheteria, gêneros, estúdios e correção da bilheteria baseada na inflação;
- Complicações na hora de conseguir uma API completa;
- Inconsistência da obtenção do HTML pelo *Webscraping* e falta de informações das API's com acesso;
- Utilização do Python e JavaScript para obtenção e limpeza de dados em CSV e TSV, assim como o Microsoft SQL Server Management para dar JOIN nas tabelas e adquirir enfim o *dataset*.

Bases de Dados

Título de Base	*Link*	*Breve Descrição*
The Movie Database	https://www.themoviedb.org/ (API)	O Movie Database (TMDB) é um banco de dados popular e editável pelo usuário para filmes e programas de TV.
IMDB Database	https://datasets.imdbws.com/ (API)	API oficial do IMDB (com grandes restrições de uso)
Metacritic (site)	https://www.metacritic.com/ (Web scraping)	Popular site para reviews de jogos, filmes e séries
The Numbers (site)	https://www.the-numbers.com/ (Web scraping)	Site com útil serviço de dados financeiros sobre filmes.

Bases de Dados



Título de Base	*Link*	*Breve Descrição*
Macro Trends (site)	https://www.macrotrends.net/countries/USA/united-states/inflation-rate-cpi	Taxa de inflação anual dos EUA.

Dataset Final

Tabela única contendo uma linha por cada filme no nosso recorte, informando: código IMDb, título, ano, bilheteria, número de ingressos vendidos, avaliação IMDb, avaliação Metacritic, box office ajustado e número de tickets ajustado.

	imdb_id	title	year	boxOffice	numTickets	boxOfficeAdjusted	ticketPriceAdjusted
2	tt0078721	10	1979	52134699	20770796	30031508.640552998	1.445852563404551
3	tt0078723	1941	1979	34175000	13615537	19686059.9078341	1.445852624676801
4	tt0322259	2 Fast 2 Furious	2003	127120058	21081269	28897489.88406456	1.370766147145343
5	tt0298203	8 Mile	2002	115270265	19839976	26794575.77870758	1.350534687073592
6	tt0075784	A Bridge Too Far	1977	50800000	22780269	35058661.145617664	1.5389924124959922

Dataset Final

	imdb_id	studio	country
2	tt1386588	Columbia Pictures	US
3	tt1386588	Gary Sanchez Productions	US
4	tt1386588	Mosaic Media Group	
5	tt1386588	Sony Pictures	US
6	tt0242653	Village Roadshow Pictures	US
7	tt0242653	NPV Entertainment	US
8	tt0242653	Silver Pictures	US
9	tt0190641	OLM	JP
10	tt0190641	Shogakukan Production	JP

Estúdios

	imdb_id	genre
2	tt0078721	Comedy
3	tt0078721	Romance
4	tt0078723	Action
5	tt0078723	Comedy
6	tt0078723	War
7	tt0322259	Action
8	tt0322259	Crime
9	tt0322259	Thriller
10	tt0298203	Drama

Gêneros

	imdb_id	source	rating
2	tt0078721	IMDb	6.1
3	tt0078723	IMDb	5.8
4	tt0078723	Metacritic	34.0
5	tt0322259	IMDb	5.9
6	tt0322259	Metacritic	58.0

Resenhas

Pergunta 1

- **Quais estúdios mais presentes na produção de filmes de alta bilheteria nos últimos 50 anos?**

Com os dados obtidos nas tabelas: *studios_table* e *films_table*, podemos obter o significativo resultado da companhia com maior presença no setor cinematográfico.

Como hipótese, já podemos pensar que alguns resultados podem acabar surgindo, como Warner Brothers ou Walt Disney Studios, porém podemos ter certeza desses números através do *boxOfficeAdjusted*, isto é, a bilheteria corrigida pela inflação.

Análise - Pergunta 1

Podemos usar a *SUM* dos *box_office_adjusted* junto com um JOIN de ambas as tabelas para nos mostrar uma lista ordenada de estúdios que mais lucraram nos últimos cinquenta anos.

```
SELECT SUM(movies.box_office_adjusted),  studios.studio
FROM movies
JOIN studios ON movies.imdb_id = studios.imdb_id
GROUP BY studios.studio
ORDER BY SUM(movies.box_office_adjusted) DESC;
```

	A	B
1	sum	studio
2	4152346066.035898542	Warner Bros. Pictures
3	3971413512.980648547	Universal Pictures
4	3122034554.4847508905	Paramount
5	2746732252.377001057	Columbia Pictures
6	2456785777.690848536	Walt Disney Pictures
7	2337264715.667633240	20th Century Fox

Pergunta 2

- **Quais filmes possuem maior colaboração internacional?**

Nosso modelo hierárquico associa cada filme a um documento, que contém o campo “studio” listando seus estúdios e países de produção. Nosso script MongoDB descobriu haver três filmes que se destacam por ter quatro países de origem: **Casino Royale** (tt0381061); **Gladiator** (tt0172495; e **Troy** (tt0332452).

```
db.getCollection('data').find( { 'studios' : { $size : { $gte : 4 } } } )
```

Pergunta 3

- **Qual a evolução temporal da bilheteria nominal média, corrigida pela inflação, dos filmes de maior sucesso ao longo dos anos?**

Com o modelo relacional é fácil obter dados para essa questão. Em primeiro lugar obtivemos o valor das bilheterias corrigidas pela inflação, em seguida adquirimos uma média da bilheteria nominal e depois agrupamos todos os filmes conforme o ano em que foram lançados. Enfim, ordenamos pelo ano para observar a mudança no valor da média.

Exemplo de SQL Query

Análise - Pergunta 3

Aqui temos a *query* que utilizamos, um exemplo de como foi a resposta do nosso Dataset (a lista continua até o ano de 1972 e está completa no repositório).

- *AVG* representa a bilheteria nominal média em dólares

```
SELECT AVG(box_office_adjusted), year
FROM movie_table
GROUP BY YEAR
ORDER BY year DESC;
```

avg	year
15573544.250062426977	2021
9823963.979872123464	2020
45706148.544324499950	2019
44716670.163752427524	2018
38515404.717595607136	2017
42657857.673079834000	2016
40508701.452599388333	2015
35164790.857033506208	2014
38226738.209516675739	2013