Atividade 1 - Visualização de informação

Leonardo Cesar Silva dos Santos Fernando Augusto Cardoso Candalaft

1 Explorando o conjunto de dados

Nesta atividade exploramos o conjunto de dados *MovieLens Latest Datasets*. O mesmo era composto por diversas tabelas, cada uma representando um tipo de informação:

- *Links*: tabela contendo as chaves de ligação entre algumas tabelas.
- *Movies*: tabela com a lista de filmes e seus respectivos gêneros.
- *Ratings*: tabela com as notas de cada filme por usuário.
- *Tags*: tabela com as respectivas *tags* de cada filme.

Exploramos os dados focando principalmente na interação entre as tabelas *Movies* e *Ratings*.

Analisamos a estrutura dos dados em cada tabela e fizemos os respectivos tratamentos necessários de modo a extrair informações relevantes. Por exemplo, extrapolamos a lista de gêneros por filme na tabela *Movies* para obtermos a informação individual de cada gênero por filme e também extraímos a informação de quando cada filme foi produzido, quando a mesma estava disponível - poucos filmes não tinham essa informação com relação ao total de filmes na base. Já na tabela *Ratings* extraímos a informação de avaliação média, mediana, máxima e mínima por filme.

movieId	title	general_genres	genres	year
1	Toy Story	Adventure|Animation|Children|Comedy|Fantasy	Adventure	1995.0
1	Toy Story	Adventure|Animation|Children|Comedy|Fantasy	Animation	1995.0
1	Toy Story	Adventure|Animation|Children|Comedy|Fantasy	Children	1995.0
1	Toy Story	Adventure|Animation|Children|Comedy|Fantasy	Comedy	1995.0
1	Toy Story	Adventure|Animation|Children|Comedy|Fantasy	Fantasy	1995.0

(a) Tabela *Movies* ajustada.

title	rating_mean	rating_max	rating_min	rating_median
Karlson Returns	5.0	5.0	5.0	5.0
Adventures Of Sherlock Holmes And Dr. Watson: ...	5.0	5.0	5.0	5.0
Justice League: Doom	5.0	5.0	5.0	5.0
English Vinglish	5.0	5.0	5.0	5.0
Junior and Karlson	5.0	5.0	5.0	5.0

(b) Tabela *Ratings* ajustada.

Figura 1: Tabelas utilizadas durante as análises.

2 Analisando os dados visualmente

Ao analisar os dados buscamos responder às seguintes perguntas:

- Como a quantidade de filmes produzidos evolui ao longo dos anos? Temos mais ou menos filmes sendo produzidos?
- A qualidade dos filmes produzidos vem aumentando ou diminuindo ao longo dos anos?
- Dos filmes produzidos ao longo dos anos, algum gênero é mais frequente do que outro?
- Quais os melhores filmes e em quais anos eles foram produzidos?
- Quais os piores filmes avaliados? (Aqueles que não devemos assistir em termos de nota).

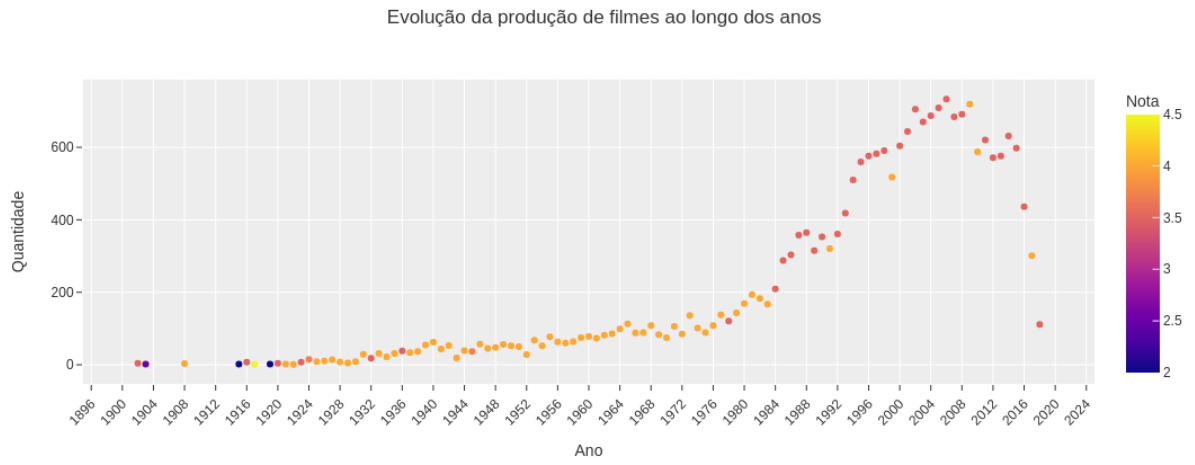

Figura 2: Evolução da quantidade de filmes produzidos e a respectiva qualidade dos mesmos ao longo dos anos.

2.1 Respondendo às perguntas via análise visual da informação

Buscamos responder às perguntas acima analisando as Figuras 2 a 6.

Na Figura 2 buscamos evidenciar a quantidade de filmes produzidos (eixo Y) ao longo dos anos (eixo X) juntamente com a qualidade de cada filme por ano (mapa de calor para as notas medianas de cada ano).

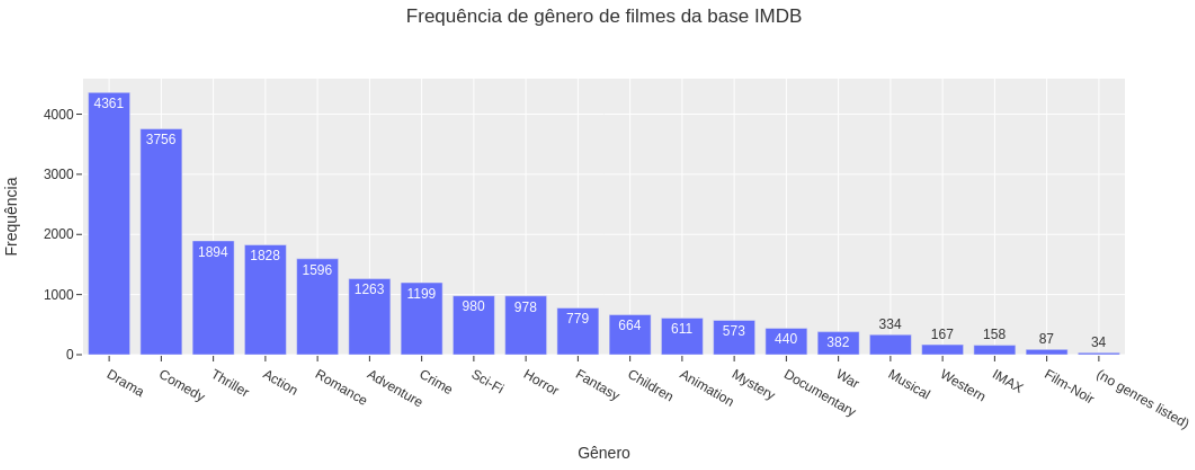

Figura 3: Frequência de cada gênero de filme presente na base IMDB.

Na Figura 3 analisamos a frequência de cada gênero presente em toda base (eixo Y) por gênero (eixo X) utilizando o auxílio de um gráfico de barras. Já na Figura 4 exibimos um gráfico do tipo *TreeMap* com os melhores filmes dentro dos gêneros mais relevantes exibidos na Figura 3.

Na Figura 5 utilizamos novamente um gráfico do tipo *TreeMap* para exibir o top 20 melhores filmes da base de dados analisada nos seus respectivos anos e com seus respectivos gêneros.

Por fim, na Figura 6 exibimos por meio de um gráfico de barras as notas dos piores 50 filmes (eixo Y) e seus respectivos títulos (eixo X).

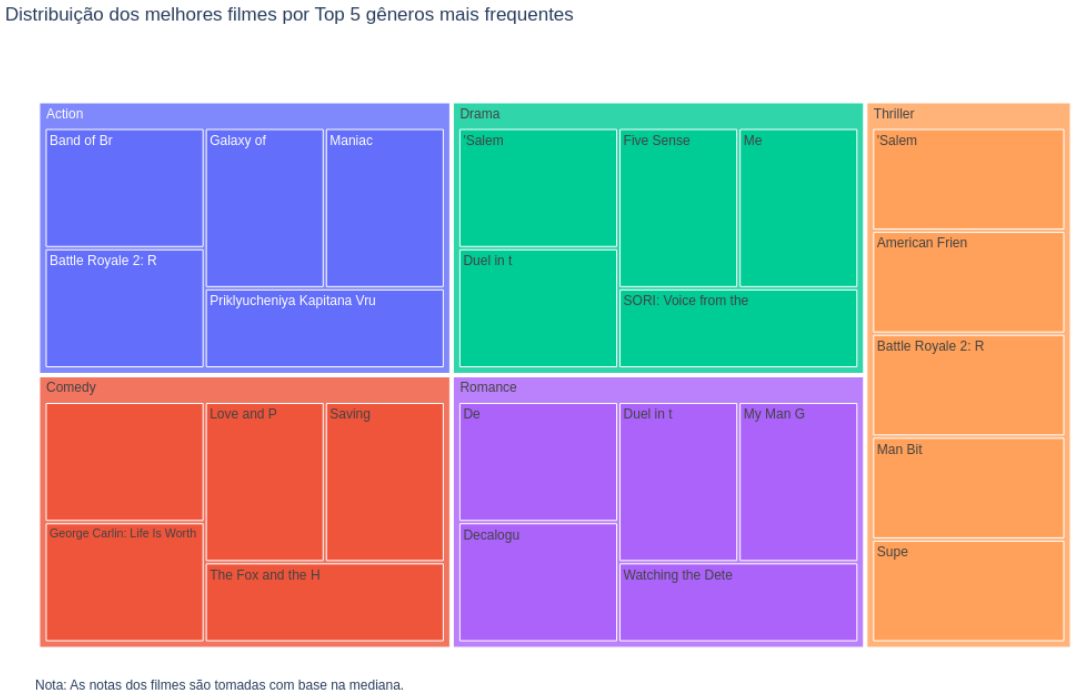

Figura 4: Melhores filmes dentro dos top 5 gêneros mais frequentes.

2.2 Resultados obtidos

- É possível notar pela Figura 2 que a quantidade de filmes produzidos teve um aumento considerável desde 1900, com um pico entre 2005 e 2010. Porém após esse período tivemos uma tendência de queda.
- É possível ver também pela Figura 2 que o período com maior qualidade de filmes está entre 1920 e 1980, com uma nota concentrada no valor 4. Após esse período temos uma leve queda nos valores das notas, 3.5, o que é inversamente proporcional a quantidade de filmes produzidos. Tal fato induz que produzir mais filmes não necessariamente está atrelado a produzir com mais qualidade.
- Dos filmes produzidos ao longo dos anos, os gêneros que mais aparecem são: Drama, Comédia, Suspense, Ação e Romance (Figura 3).
- Dentre os gêneros mais frequentes os melhores de cada gênero podem ser consultados na Figura 4.
- Os melhores filmes de acordo com a nota, seus respectivos anos e também os gêneros que o filme aborda são dados na Figura 5.
- Já os piores filmes avaliados pelos usuários são dados na Figura 6.

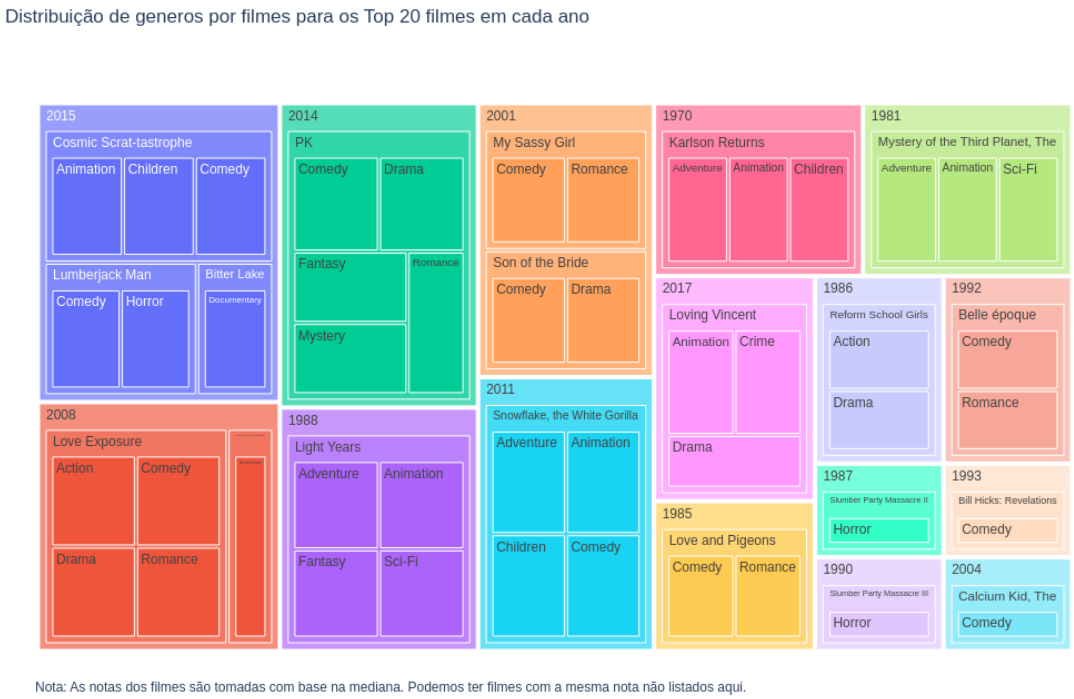

Figura 5: Melhores filmes por ano e seus respectivos gêneros.

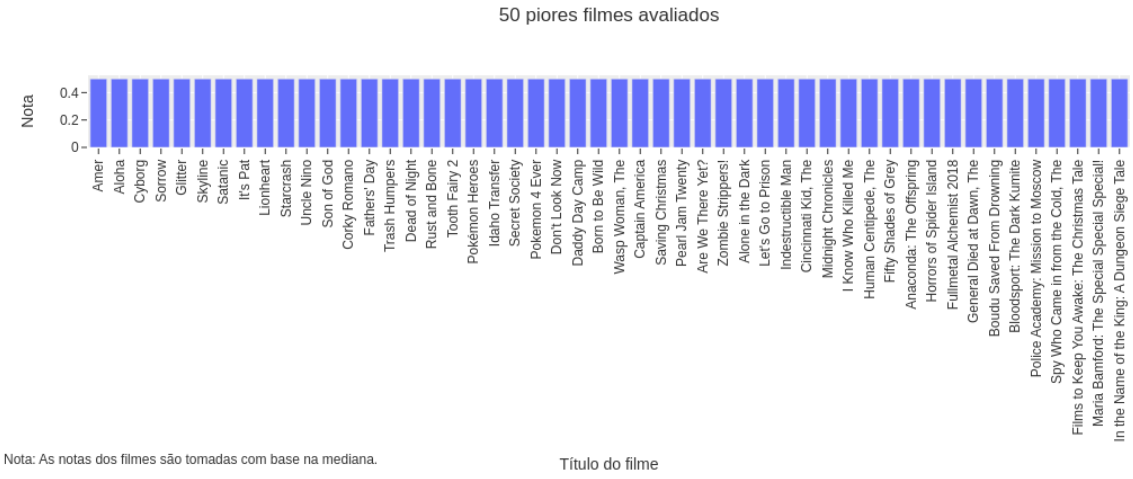

Figura 6: Melhores filmes por ano e seus respectivos gêneros.