



UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS DE CRATEÚS
PROFESSOR: RENAN GOMES VIEIRA

TRABALHO 02 - CIÊNCIA DOS DADOS

Relatório: Análise Exploratória de Dados e Visualizações.

Equipe:

- Igor Santana Sampaio
- Jorge Roniel de Paula Souza
- Luis Eduardo Martins Barbosa
- Victor Lopes Mendes

Introdução

Este relatório documenta o processo de análise exploratória e visualização dos dados para a disciplina de ciência dos dados, referente ao trabalho T1. O projeto consiste na análise do dataset MovieLens que possui dados relativos a filmes e seus rankings e notas avaliados por fãs por meio de aplicativos como IMDB e TMDB. Nosso objetivo neste trabalho é tomar certo conhecimento dos dados para futuros trabalhos e responder perguntas sobre os dados através de visualizações gráficas.

Este documento apresenta a nossa metodologia e as dificuldades encontradas durante o desenvolvimento do trabalho.

Metodologia

Para atender os requisitos e objetivos do nosso trabalho, a equipe adotou um fluxo de trabalho, dividindo as tarefas para otimizar a execução dentro do prazo para a entrega.

1. Bibliotecas

Neste trabalho, utilizamos a linguagem Python que é bastante utilizada para trabalhos como este e, aliado a linguagem, utilizamos as seguintes bibliotecas que já são próprias para análise e visualização gráfica dos dados:

- Pandas
- Seaborn
- Matplotlib
- Numpy
- Regex (manipulação de strings)

Também utilizamos o GitHub para guardar nosso trabalho em um repositório e facilitar o trabalho em grupo.

2. Caracterização do Conjunto de Dados Escolhido

O conjunto de dados que utilizamos neste trabalho foi o *MovieLens* (ml-32m) , um recurso amplamente reconhecido na pesquisa de sistemas de recomendação, operado e mantido pelo *GroupLens Research* na Universidade de Minnesota

2.1. Apresentação e Estrutura dos Dados

- **Origem e descrição:** O conjunto descreve atividades de avaliação de 5 estrelas e *tagging* em texto livre de filmes, extraído do serviço de recomendação MovieLens.
- **Arquivos Fonte:** Os dados estão contidos em quatro arquivos CSV: links.csv, movies.csv, ratings.csv, e tags.csv.
- **Número de Instâncias**
 - **Avaliações:** 32.000.204
 - **Aplicações de Tags:** 2.000.072
 - **Usuários:** 200.948
 - **Filmes:** 87.585

- **Período de Tempo:** Os dados foram criados entre **9 de janeiro de 1995** e **12 de outubro de 2023**. O conjunto foi gerado em 13 de outubro de 2023.
- **Estrutura de linha:** Cada linha em *ratings.csv* é uma avaliação (*userId*, *movieId*, *rating*, *timestamp*). As avaliações são feitas em uma escala de 5 estrelas, com incrementos de meia estrela (0.5 - 5.0).

2.2. Licença de Uso e Questões Éticas/Privacidade

Licença de Uso

Os dados do *dataset* escolhido podem ser usados para quaisquer fins de pesquisa , contanto que sejam seguidas as seguintes condições:

- **Citação Obrigatória:** O uso do conjunto de dados deve ser reconhecido em publicações resultantes.
- **Não Endosso:** O usuário não pode declarar ou implicar qualquer endosso da Universidade de Minnesota ou do GroupLens.
- **Não Comercial:** O uso para fins comerciais ou geradores de receita é proibido sem permissão prévia.
- **Redistribuição:** O *dataset* pode ser redistribuído, incluindo transformações, desde que sob as mesmas condições de licença.

Questões Éticas e de Privacidade

Sim, questões éticas e de privacidade devem ser levantadas:

1. **Anonimato e Seleção de Usuários:** Os IDs de usuário foram anonimizados. Apenas o ID de usuário é fornecido, sem informações demográficas. Além disso, apenas usuários que avaliaram pelo menos 20 filmes foram incluídos, o que cria um viés de atividade no conjunto.
2. **Risco de Re-identificação:** Embora o anonimato seja um esforço importante, pesquisas passadas com conjuntos de dados de avaliação sugerem que, ao combinar as avaliações de um usuário com informações de outros *datasets* públicos, pode haver um risco de re-identificação (inferência da identidade real do usuário).
3. **Tags (Metadados Gerados pelo Usuário):** As *tags* são metadados gerados por usuários. O significado e o propósito de uma *tag* são determinados por cada usuário, o que pode introduzir subjetividade e viés de linguagem/cultural.

3. Análise Exploratória dos Dados

Para conhecer melhor os dados que temos, precisamos explorá-los e verificar como as amostras estão distribuídas. Para isso, fizemos os processamentos necessários desses dados e a AED em si.

3.1. Processamento dos dados

Para fazer uma boa análise, primeiro fizemos alguns processamento nos nossos dados. Primeiro, como os dados vieram divididos, tivemos que unir os dados em um *DataFrame* só, para isso, utilizamos a função *Merge* da biblioteca Pandas, porém, havia um atributo (Tags), que possuía muitos dados repetidos e ao tentar usar o merge data estouro de memória, então, agregamos as tags pelo atributo *MovieId* utilizando a função *GroupBy* do Pandas.

Fizemos também a conversão de *TimeStamp* para o *DateTime*, e o tratamento de dados faltantes utilizando o *left join* para unir movies com links e *tags_agregadas* garantindo que todos os filmes do dataset movies sejam representados no *info_filmes_completa*, mesmo que não tenham IDs externos (*tmdbId*) ou tags. Além disso, fizemos uma verificação de consistência usando um *inner join* com *Ratings* garantindo que apenas as avaliações que têm filmes correspondentes no nosso conjunto de informações sejam incluídas.

Dessa forma, com nossos dados mais limpos, podemos fazer algumas visualizações para entendermos melhor os dados.

3.2 Análise dos dados

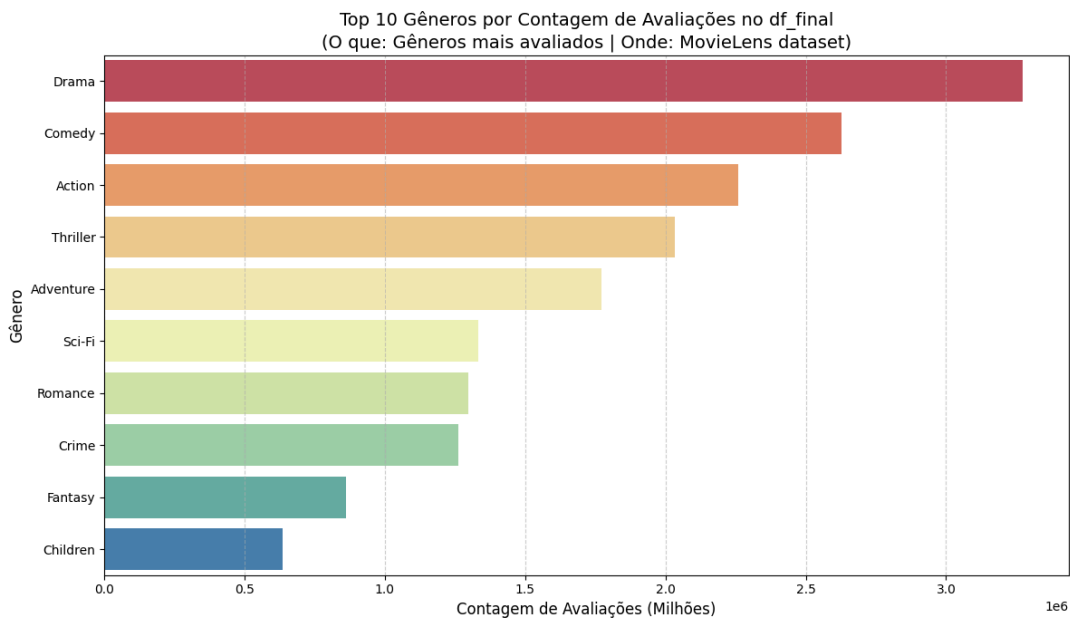
Após o tratamento dos dados utilizamos as bibliotecas Matplotlib e Seaborn para gerarmos visualizações gráficas para conhecermos melhor os dados do nosso conjunto, plotamos variados tipos de gráficos adequados para o que queríamos ver e entender. Depois, foi feita mais algumas visualizações para respondermos algumas questões e insights sobre os dados.

Foi uma tarefa interessante que nos possibilitou conhecer mais sobre o que estávamos lidando com aquele *Dataset*, e também nos ajudou a entender melhor sobre a parte de visualização dos dados. Assim, conseguimos obter alguns achados interessantes nos dados através de gráficos que podem ajudar, não só a nossa equipe, mas outras pessoas que se interessarem.

3.3 Principais Achados

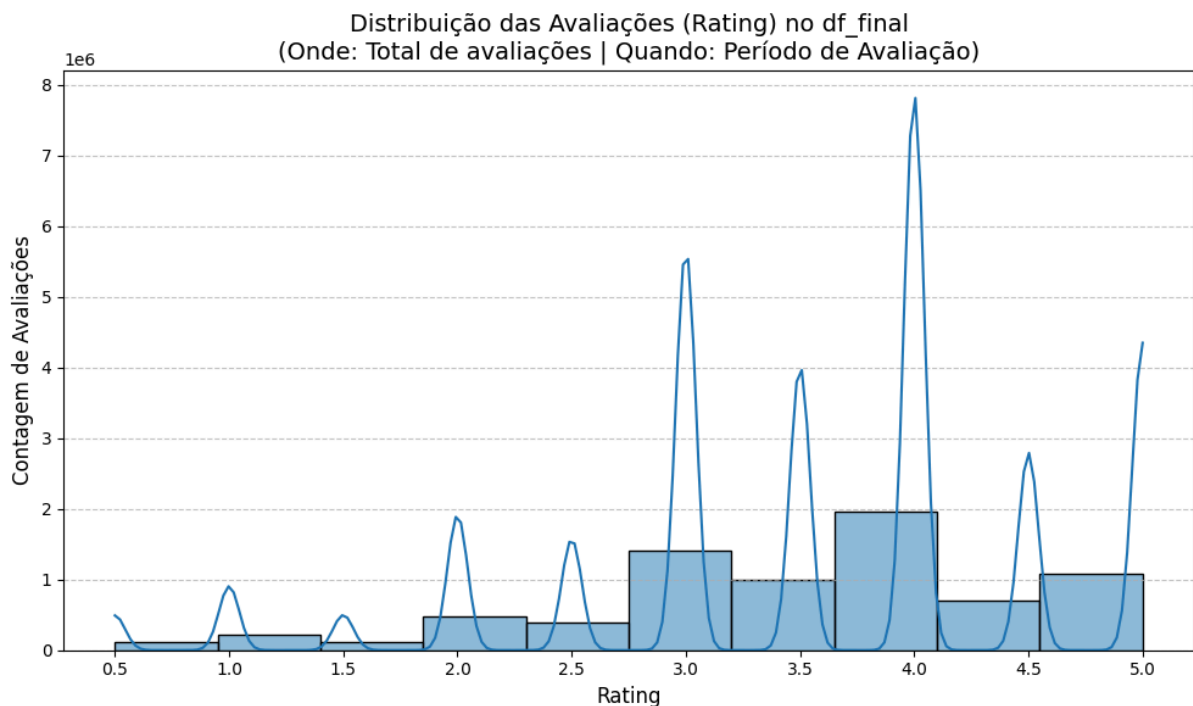
Durante a análise dos dados, conseguimos identificar alguns insights dos nossos dados que podem ser interessantes.

De acordo com o seguinte gráfico:



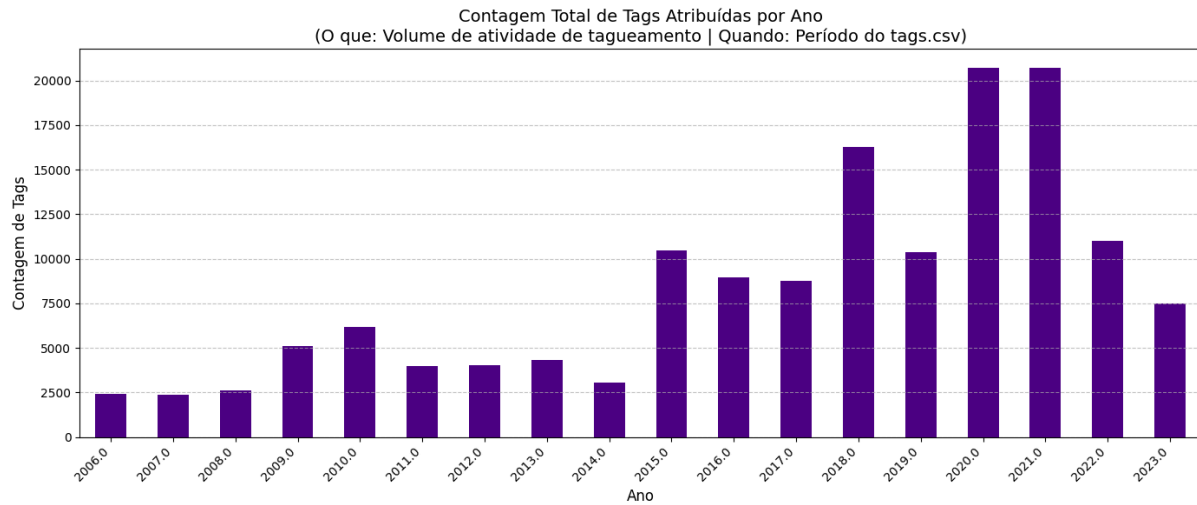
Podemos ver que o seguinte gráfico mostra os 10 gêneros cinematográficos com maior número total de avaliações no nosso dataset e que os gêneros de Drama e Comédia dominam com mais de 5,8 milhões de avaliações, isso pode ocorrer por serem gêneros mais amplos e diversificados, abrangendo muitos filmes e costuma agradar públicos variados.

Dado o seguinte gráfico:



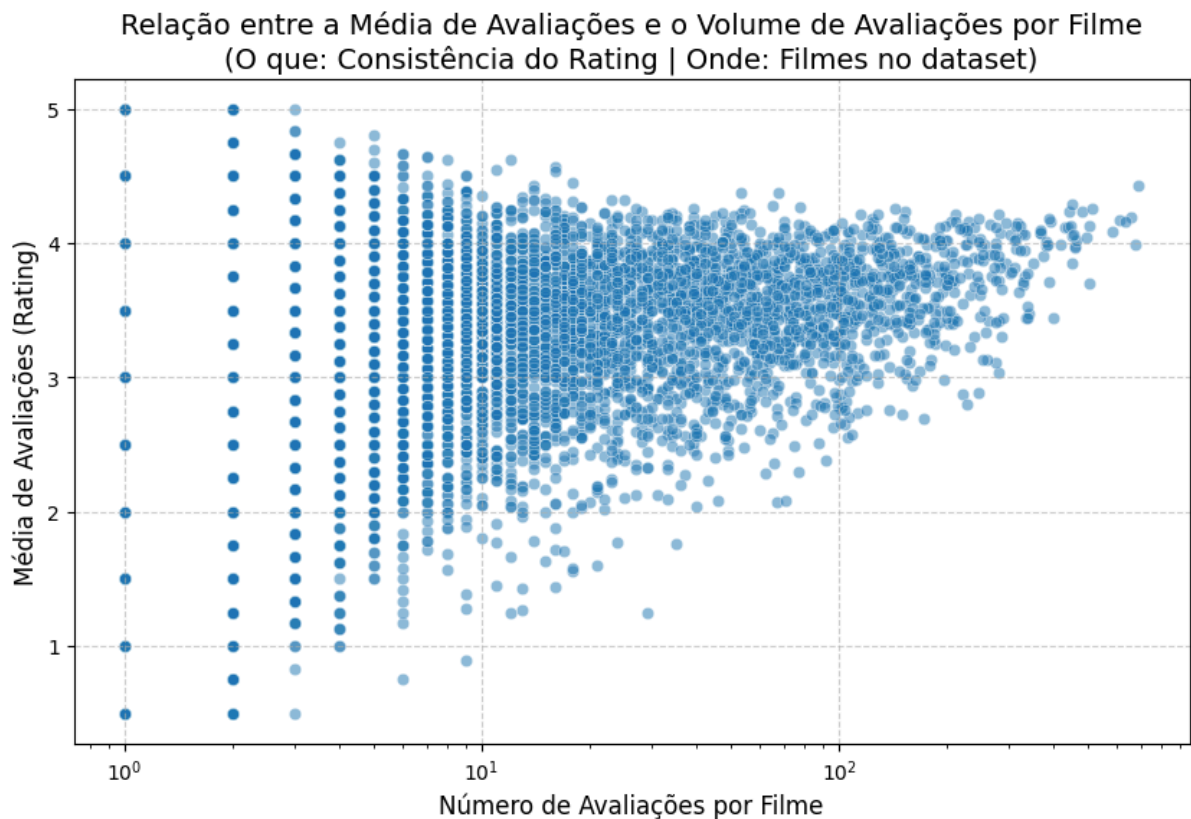
Podemos ver que há picos nítidos nas notas 3.0 e 4.0, e notas como 0.5, 1.0, são menos frequentes, o que pode indicar primeiramente que os usuários do *MovieLens* tendem a avaliar mais positivamente os filmes e que usuários avaliam filmes que tem afinidade. Uma curiosidade é que ,muitos usuários preferem dar notas “redondas” que é um comportamento comum em sistemas com notas discretas.

De acordo com este próximo gráfico:



Podemos perceber um aumento mais lento de 2006 à 2014 de tags atribuídas, mas em 2015 esse número começou a crescer mais, com saltos maiores, especialmente em 2018, 2020 e 2021, isso pode indicar aumento de usuários engajados depois de 2015 que podem ter sido provocadas por melhorias na interface da plataforma e o incentivo ao uso de tags.

E segundo o gráfico seguinte:



Temos que filmes com poucas avaliações apresentam grande variação da média, indo de muito baixas a muito altas e conforme o número de avaliações aumenta, a média converge para um intervalo mais estreito, entre 3.0 e 4.3. Isso pode implicar que filmes pouco avaliados podem sofrer com mais ruídos estatísticos, ou seja, uma única avaliação extrema

afeta a média, e filme muito avaliados tendem a ter médias mais sólidas, pois muitos usuários suavizam os extremos e há menor probabilidade de viés individual. Dessa forma, a reputação real de um filme no *MovieLens* aparece melhor quando o número de avaliações é alto.

Desafios Encontrados

Durante o processo de desenvolvimento do trabalho, foi encontrado algumas dificuldades que surtiram aprendizados para equipe:

- **Junção dos Dados:** Durante o desenvolvimento, ao utilizar os dados do MovieLens, tivemos a dificuldade de unir os 4 data frames em um só para melhor utilização. A falta de prática com a biblioteca pandas dificultou o começo do trabalho, mas após algumas pesquisas, conseguimos solucionar esse problema e demos continuidade aos trabalhos de AED com mais conhecimento sobre a biblioteca pandas.
- **Estouro de Memória:** Ao tentar gerar os gráficos, por serem muitos dados, as máquinas dos alunos apresentaram estouro de memória, para contornar isso, foi preciso utilizar o site Colab para gerarmos os gráficos por lá e utilizarmos eles em nosso trabalho, mostrando a importância de conhecermos outras plataformas que auxiliam no desenvolvimento.

Recomendações

Após a análise dos dados e o conhecimento do *DataSet* faz-se necessário algumas recomendações para trabalhos futuros:

- **Diminuição dos Dados:** Por o conjunto de dados completo ter muitas amostras, seria bom considerar a diminuição de algumas linhas, sem que os dados percam sua identidade, para um possível treinamento de um modelo nos próximos trabalhos, garantindo assim, eficiência e facilidade na hora de manipular tais dados.

Conclusão

Dessa forma, concluímos que este trabalho permitiu à equipe a colocar na prática conceitos de AED (Análise Exploratória dos Dados) e Visualização dos Dados vistos em sala de aula com datasets que apresentavam desafios que foram contornados pelos alunos. Assim, a análise feita pela equipe atende aos requisitos do trabalho e seus objetivos, auxiliando no aprendizado e desenvolvimento dos membros da equipe.