UNIVERSIDADE FEDERAL DO CEARÁ - CAMPUS CRATEÚS
CIÊNCIA DOS DADOS (2025.2)

RELATÓRIO DO TRABALHO PRÁTICO 02

Análise Exploratória de Dados e Visualização de Dados.

Docente: Renan Gomes Vieira

Discentes:
Ana Larissa Teixeira Dantas - 536615
Francisco Enzo Sousa Oliveira - 556192
Gabriela Silva Ximenes - 556297
Wagner Vasconcelos Dias - 538314

1. Objetivo do Trabalho

O objetivo deste trabalho foi realizar uma **Análise Exploratória de Dados (EDA)** utilizando o dataset MovieLens, identificando padrões, tendências, relações entre variáveis e questões analíticas relacionadas a filmes, notas e comportamento dos usuários. A análise foi conduzida em Python, com foco em técnicas estatísticas e visualização de dados.

2. Dados Utilizados

Os dados utilizados neste trabalho fazem parte do MovieLens, um sistema de recomendação de filmes mantido pelo GroupLens Research, da Universidade de Minnesota. O conjunto específico usado aqui é o MovieLens 32M (ml-32m).
Esse dataset reúne informações sobre filmes e avaliações feitas por usuários da plataforma ao longo de quase três décadas.
O MovieLens 32M inclui avaliações registradas entre 09 de janeiro de 1995 e 12 de outubro de 2023.

A base MovieLens contém:

- **ratings.csv** – informações das avaliações (ID do usuário, ID do filme, nota e timestamp).
- **movies.csv** – títulos e gêneros dos filmes (ID, título e gêneros).

O dataset completo contém:

- 32.000.204 avaliações;
- 87.585 filmes;
- 200.948 usuários.

3. Processamento realizado

Nesta etapa, foram aplicadas diversas verificações e transformações para garantir a integridade e consistência dos dados utilizados.

- Verificação de dados faltantes.
- Verificação de duplicatas.
- Consistência interna (Verifica se há avaliações que não estão em movies).
- Análise de valores fora do domínio (Notas entre 0,5 e 5).
- Amostragem estratificada para redução do conjunto (Mantendo 10% das avaliações de cada usuário, reduzindo de 32M amostras para 3M, aproximadamente).
- Feature Engineering:
    - Ratings
        - year: Criada a partir do timestamp.
        - month: Criada a partir do timestamp.
        - movie_rating_count: número de avaliações do filme.
        - movie_rating_mean: média das notas.
        - movie_rating_std: variabilidade das notas.
        - user_rating_count: número de avaliações feita por usuário.
        - user_rating_mean: médias das notas dadas por um usuário.
        - user_rating_std: dispersão das notas.

- Movies
  - release_year: Extraída do title que na maioria das vezes tem a seguinte forma: “Toy Story (1995)”.

Essas etapas permitiram consolidar um conjunto de atributos enriquecido, adequado para análises numéricas, temporais e categóricas.

4. Principais Achados da Análise

4.1 Distribuição das Notas

- **Média**: 3.54
- **Mediana**: 3.5
- **Desvio padrão**: 1.06
- **Intervalo**: 0.5 – 5.0
- 75% das notas **≥** 3.0

A maior parte das avaliações está entre **3.0 e 4.5**, indicando tendência positiva.

4.2 Filmes Mais Bem Avaliados

Ao analisar apenas filmes com quantidade mínima de avaliações, foi possível identificar um conjunto consistente de títulos bem avaliados, geralmente obras reconhecidas ou clássicas.

4.3 Popularidade dos Filmes

A distribuição de popularidade é bastante **assimétrica**: poucos filmes concentram a maior parte das avaliações, enquanto a maioria recebe atenção limitada. Essa concentração demonstra que o comportamento dos usuários se volta fortemente para obras mais conhecidas.

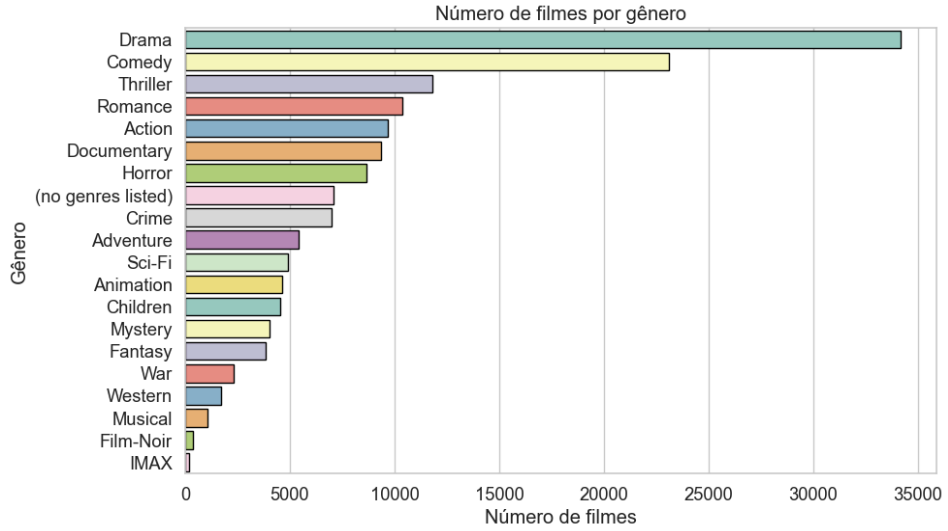
4.4 Padrões Temporais nas Avaliações

Ao longo dos anos, as notas médias atribuídas se mantêm relativamente estáveis, sugerindo que o comportamento dos usuários não mudou de forma significativa com o tempo. Já o volume de avaliações apresenta oscilações mais marcantes, com picos de participação em determinados períodos, possivelmente associados ao crescimento da base de usuários da plataforma.

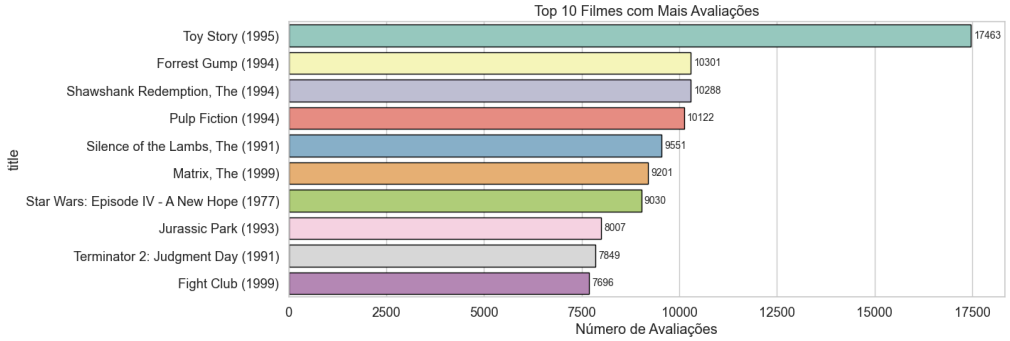
4.5 Distribuição e Preferência por Gêneros

Drama, Comédia e Ação são os gêneros mais frequentes, refletindo grande presença no catálogo. A análise por gênero também mostrou diferenças na forma como cada categoria é avaliada, embora sem grandes extremos, indicando que a percepção de qualidade varia, mas não de maneira abrupta entre os principais gêneros.

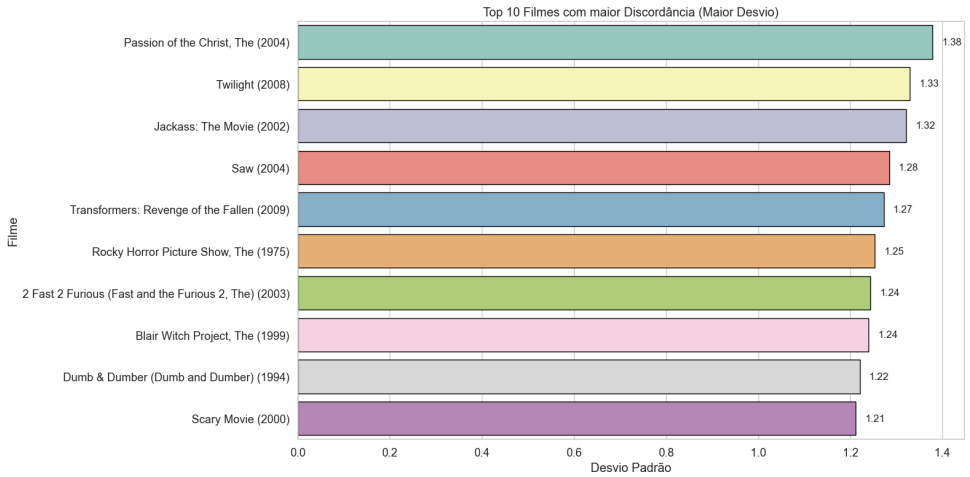
5. Principais Achados Gráficos



O gráfico apresenta a distribuição do número de filmes por gênero, evidenciando que **Drama** e **Comedy** são, de longe, os gêneros mais frequentes, com mais de 20 mil filmes cada.



O gráfico apresenta os **10 filmes com maior número de avaliações**, destacando títulos amplamente conhecidos do público.



O gráfico mostra os **10 filmes com maior discordância entre as avaliações**, medida pelo desvio padrão.

6. Limitações do Estudo

- Desbalanceamento no número de avaliações entre filmes
- Dataset limitado ao universo MovieLens
- Computador limitado à quantidade de dados
- Ausência de dados demográficos dos usuários

7. Recomendações

- Utilizar métodos ponderados ou filtros de volume mínimo de avaliações
- Explorar análises temporais mais complexas
- Investir em visualizações interativas
- Avaliar relações entre usuários e gêneros utilizando técnicas mais avançadas

8. Desafios Encontrados

- Manipulação correta da coluna de gêneros
- Processamento eficiente de grandes volumes de dados
- Criação de critérios justos para ranqueamento
- Construção de gráficos informativos sem distorção visual

9. Aprendizados Obtidos

- Proficiência maior em Pandas, Matplotlib e Seaborn
- Experiência completa em pipeline de EDA
- Importância da limpeza e validação de dados
- Aplicação prática de estatística descritiva
- Melhora na organização e comunicação de achados por meio de visualização gráfica

10. Referências

- **GroupLens Research – MovieLens Dataset**
- **Documentação oficial do Pandas**
- **Documentação do Matplotlib**
- **Documentação do Seaborn**
- Notas de aula e orientações fornecidas na disciplina

11. Link do repositório no GitHub

- **Enzooliveira29/Trabalho-de-CD**