UNIVERSIDADE PRESBITERIANA MACKENZIE

Faculdade de Computação e Informática
Tecnologia em Ciência de Dados
Projeto Aplicado 1

Aplicando Conhecimento 1 e 2

Análise de pontuação da Netflix

Grupo Projeto Aplicado 1

Cristiano Prado do Carmo	10720249
Felipe Garcia Pereira Brathwaite	10408441
Ingryd Cristine Hidalgo Sella	10424934
Pablo Rodrigo Dias	10414537
Thiago Godeguesi	10408130

Sumário

Seção 1 - Apresentação

Título do Trabalho

Análise de pontuação da Netflix.

Membros do Grupo

Aluno	RA
Cristiano Prado do Carmo	10720249
Felipe Garcia Pereira Brathwaite	10408441
Ingryd Cristine Hidalgo Sella	10424934
Pablo Rodrigo Dias	10414537
Thiago Godeguesi	10408130

Objetivo do Projeto

Analisar o comportamento dos usuários com as votações, buscando as pontuações de cada filme, análise média de pontuação por ano, pontuação por usuário.

Contexto de Estudo

A Netflix promoveu uma competição aberta chamada "Netflix Prize", em 2006, para que um novo algoritmo de predição de avaliação de filmes pelos seus assinantes fosse criado.

Foi oferecido um prêmio de U$ 1.000.000,00 para o vencedor: o grupo conhecido por **BellKor's Pragmatic Chaos**[1], em 2009.

A título de curiosidade, a Netflix nunca chegou a usar realmente o algoritmo vencedor[2].

Descrição da Origem

A origem dos dados é o Kaggle, que é um repositório de milhares de datasets, contendo os mais variados temas.

Descrição do Dataset

O dataset escolhido é formado por um conjunto de 5 (cinco) arquivos em formato "*.txt*" e 1 (um) arquivo em formato "*.csv*", detalhados a seguir.

Os arquivos "*combined_data_1.txt*", "*combined_data_2.txt*", "*combined_data_3.txt*" e "*combined_data_4.txt*" contêm os dados de avaliação realizada pelo assinante, distribuídos da seguinte forma:

- Uma linha contendo o Id do filme seguido de dois pontos e, cada linha subsequente, correspondendo a uma avaliação de um assinante, obedecendo o seguinte formato, separado por vírgula:

[1] Saiba mais sobre a competição e sobre o time vencedor em https://www2.seas.gwu.edu/~simhaweb/champalg/cf/papers/KorenBellKor2009.pdf e https://www.asc.ohio-state.edu/statistics/statgen/joul_aut2009/BigChaos.pdf

[2] https://thenextweb.com/news/remember-netflixs-1m-algorithm-contest-well-heres-why-it-didnt-use-the-winning-entry

- Id do assinante, anonimizado
- Avaliação, sendo ela um número inteiro em uma escala de 1 a 5
- Data da avaliação, no formato YYYY-MM-DD

Exemplo do formato nos arquivos *"combined_data_1.txt"*, *"combined_data_2.txt"*, *"combined_data_3.txt"* e *"combined_data_4.txt"*:

```
MovieID1:
CustomerID11,Rating11,Date11
CustomerID12,Rating12,Date12
…
MovieID2:
CustomerID21,Rating21,Date21
CustomerID22,Rating22,Date22
…
```

- O arquivo *movie_titles.csv* contém os dados dos filmes avaliados e disponíveis no catálogo da Netflix, no seguinte formato, separado por vírgula:
  - Id do filme, anonimizado
  - Ano de lançamento[3], variando de 1890 a 2005
  - Título[4]

Exemplo do formato no arquivo "*movie_titles.csv*":

```
MovieID1,YearOfRelease1,Title1
MovieID2,YearOfRelease2,Title2
MovieID3,YearOfRelease3,Title3
MovieID4,YearOfRelease4,Title4
…
```

- O arquivo *"qualifying.txt"* consiste em linhas indicando o Id do filme avaliado seguido por dois-pontos. As linhas subsequentes contêm os dados do assinante e a data de avaliação, obedecendo ao seguinte formato, separado por vírgula:
  - Id do assinante, anonimizado
  - Data da avaliação, no formato YYYY-MM-DD

Exemplo do formato no arquivo "*qualifying.txt*":

```
MovieID1:
CustomerID11,Date11
CustomerID12,Date12
…
MovieID2:
CustomerID21,Date21
CustomerID22,Date22
…
```

[3] O ano de lançamento não corresponde ao ano de lançamento do filme nos cinemas, podendo ser o seu ano de lançamento em DVD

[4] Os títulos dos filmes estão em inglês e correspondem aos títulos como são encontrados na Netflix

- O arquivo *"probe"* não será utilizado neste projeto. Ele contém dados de treinamento do modelo desenvolvido e, modelagem e treinamento não faz parte do escopo deste projeto.

Seção 2 – Proposta Analítica

Empresa

A Netflix foi fundada em 1997 por Reed Hastings e Marc Randolph em Scotts Valley, Califórnia, EUA. Inicialmente, operava como um serviço de aluguel de DVDs pelo correio e, em 2007, introduziu o serviço de streaming de vídeo.

A missão[5] da Netflix é proporcionar entretenimento de qualidade globalmente, oferecendo uma ampla variedade de conteúdos que atendem aos diversos gostos de seus assinantes.

Os valores[5] fundamentais da Netflix incluem:

- Integridade e Transparência: Manter altos padrões éticos e uma comunicação aberta.
- Inovação e Excelência Criativa: Investir no desenvolvimento de conteúdo e tecnologia de ponta.
- Diversidade e Inclusão Global: Promover uma cultura inclusiva e representativa.
- Foco no Cliente: Priorizar a experiência e satisfação dos assinantes.
- Compromisso com a Sustentabilidade: Adotar práticas que minimizem o impacto ambiental.

Desde que deixou de ser uma empresa de aluguel de DVDs, a Netflix passou a atuar exclusivamente no segmento de streaming de vídeo por assinatura, disponível em diversas plataformas. Inicialmente, seu catálogo era composto por conteúdos de outras produtoras, mas, devido ao grande sucesso do streaming, a empresa passou a produzir suas próprias obras, conquistando enorme reconhecimento com diversas produções originais.

Atualmente, estima-se que cerca de 14.000 pessoas trabalham na Netflix (dados de 2024). Esse número representa o dobro da quantidade de funcionários em 2018, evidenciando um crescimento significativo em um curto período.

A Netflix mantém investimentos contínuos em *Analytics*[6], *Engenharia e Ciência de Dados*[7], buscando aprimorar seus algoritmos de recomendação e oferecer uma experiência cada vez mais personalizada aos seus assinantes.

De fato, o nosso projeto utiliza o dataset disponibilizado pela Netflix em uma ação para aprimorar seu mecanismo de sugestão de conteúdo para seus assinantes.

Problema de Estudo

Nossa proposta de estudo baseia-se em encontrar variações de avaliações de acordo com o tempo em que o filme está disponível, ou seja, qual a taxa de variação na quantidade e na qualidade das avaliações ao longo do tempo que o filme se encontra disponível no catálogo.

[5] https://dcfmodeling.com/pt/blogs/vision/nflx-mission-vision?utm_source=chatgpt.com
[6] https://research.netflix.com/research-area/analytics
[7] https://jobs.netflix.com/team?slug=data-science-and-engineering

Proposta Analítica

O objetivo do nosso trabalho é explorar o dataset para encontrar medidas que descrevam claramente os dados e a sua distribuição, buscando compreender o comportamento das avaliações de filmes realizadas pelos usuários.

A partir da amostra de dados fornecida, que contém informações sobre filmes, usuários, notas e datas de avaliação, pretende-se identificar padrões, tendências e possíveis insights que possam embasar futuras decisões, como recomendações personalizadas e análises de qualidade de filmes.

As etapas que seguiremos serão as seguintes:

1. Análise Descritiva das Avaliações:
    a. Distribuição das notas (1 a 5 estrelas)
    b. Filmes mais e menos avaliados
    c. Média de avaliações por filme
2. Análise Temporal:
    a. Como a distribuição das notas evoluiu ao longo do tempo?
    b. Houve anos em que os usuários foram mais críticos ou mais generosos?
    c. Existe sazonalidade nas avaliações? (ex.: mais avaliações em determinados meses ou anos)
3. Análise do Comportamento dos Usuários:
    a. Identificação de padrões de avaliação: há usuários que tendem a dar notas mais altas ou mais baixas?
    b. Quantidade média de avaliações por usuário
    c. Existe um grupo pequeno de usuários que avalia a maioria dos filmes?
4. Relação entre Ano de Lançamento e Notas:
    a. Filmes mais antigos recebem notas mais baixas ou mais altas?
    b. Existe um padrão de notas para filmes lançados em anos específicos?
5. Visualizações e Insights Gerais:
    a. Histogramas das avaliações por nota
    b. Linha do tempo da média de avaliações por ano
    c. Scatter plot relacionando a quantidade de avaliações e a nota média dos filmes
    d. Análise de outliers: filmes com notas extremamente altas ou baixas com base no volume de avaliações

Este projeto permitirá não apenas compreender a distribuição das avaliações de filmes, mas também levantar hipóteses relevantes sobre o comportamento dos usuários e a qualidade percebida dos filmes ao longo do tempo. A partir dessa EDA, é possível gerar insights para otimizar recomendações e melhorar a experiência do usuário em plataformas de streaming.

Seção 3 – Narrativa do Projeto

Apresentação do Grupo

Olá! Somos o Grupo 1 de Projeto Aplicado. Nosso grupo é formado por 5 alunos que estão neste semestre totalmente empenhados em construir um grande projeto.

Aluno	RA
Cristiano Prado do Carmo	10720249
Felipe Garcia Pereira Brathwaite	10408441
Ingryd Cristine Hidalgo Sella	10424934
Pablo Rodrigo Dias	10414537
Thiago Godeguesi	10408130

Nome do Projeto

Nosso projeto se chama Análise de Pontuação da Netflix, e queremos analisar o comportamento dos usuários com as votações, buscando as pontuações de cada filme, análise média de pontuação por ano, pontuação por usuário.

Descrição do Problema

A Netflix promoveu uma competição aberta chamada "Netflix Prize", em 2006, para que um novo algoritmo de predição de avaliação de filmes pelos seus assinantes fosse criado.

Foi oferecido um prêmio de U$ 1.000.000,00 para o vencedor: o grupo conhecido por **BellKor's Pragmatic Chaos**[8], em 2009.

A título de curiosidade, a Netflix nunca chegou a usar realmente o algoritmo vencedor[9].

A Empresa Escolhida

Fundada em 1997, a Netflix teve um papel decisivo na transformação do consumo de mídia, começando como um serviço de aluguel de DVDs e evoluindo para se tornar uma das maiores plataformas de streaming do mundo.

Desde que deixou de ser uma empresa de aluguel de DVDs, a Netflix passou a atuar exclusivamente no segmento de streaming de vídeo por assinatura, disponível em diversas plataformas. Inicialmente, seu catálogo era composto por conteúdos de outras produtoras, mas, devido ao grande sucesso do streaming, a empresa passou a produzir suas próprias obras, conquistando enorme reconhecimento com diversas produções originais.

[8] Saiba mais sobre a competição e sobre o time vencedor em https://www2.seas.gwu.edu/~simhaweb/champalg/cf/papers/KorenBellKor2009.pdf e https://www.asc.ohio-state.edu/statistics/statgen/joul_aut2009/BigChaos.pdf

[9] https://thenextweb.com/news/remember-netflixs-1m-algorithm-contest-well-heres-why-it-didnt-use-the-winning-entry

Atualmente, estima-se que cerca de 14.000 pessoas trabalham na Netflix (dados de 2024). Esse número representa o dobro da quantidade de funcionários em 2018, evidenciando um crescimento significativo em um curto período.

A Netflix mantém investimentos contínuos em *Analytics*[10], *Engenharia e Ciência de Dados*[11], buscando aprimorar seus algoritmos de recomendação e oferecer uma experiência cada vez mais personalizada aos seus assinantes.

De fato, o nosso projeto utiliza o dataset disponibilizado pela Netflix em uma ação para aprimorar seu mecanismo de sugestão de conteúdo para seus assinantes.

Problema de Estudo

Nosso projeto foi inspirado em um desafio real lançado pela empresa Netflix, conhecido como Netflix Prize, uma competição de ciência de dados voltada à melhoria do seu sistema de recomendação.

O Gap

Em 2006, com o objetivo de aprimorar a personalização das recomendações de filmes para seus usuários, a Netflix lançou um desafio global: equipes do mundo inteiro foram convidadas a desenvolver algoritmos que superassem em pelo menos 10% a acurácia do sistema de recomendação interno da empresa, conhecido como Cinematch.

Dados do Desafio

Para isso, a Netflix disponibilizou uma base de dados com mais de 100 milhões de avaliações anônimas de usuários reais sobre cerca de 18 mil filmes. Os participantes deveriam prever a nota que um usuário daria a um filme com base em seu histórico de avaliações e no comportamento de outros usuários semelhantes.

O critério de avaliação era baseado no erro quadrático médio (RMSE), e o time vencedor recebeu um prêmio de US$ 1 milhão em 2009, após atingir o objetivo proposto.

Este desafio se tornou um marco na história da ciência de dados e dos sistemas de recomendação, incentivando a criação de técnicas mais robustas de filtragem colaborativa e aprendizado de máquina.

Proposta Analítica

O objetivo do nosso trabalho é explorar o dataset para encontrar medidas que descrevam claramente os dados e a sua distribuição, buscando compreender o comportamento das avaliações de filmes realizadas pelos usuários.

[10] https://research.netflix.com/research-area/analytics
[11] https://jobs.netflix.com/team?slug=data-science-and-engineering

A partir da amostra de dados fornecida, que contém informações sobre filmes, usuários, notas e datas de avaliação, pretende-se identificar padrões, tendências e possíveis insights que possam embasar futuras decisões, como recomendações personalizadas e análises de qualidade de filmes.

As etapas que seguiremos serão as seguintes:

6. Análise Descritiva das Avaliações:
    a. Distribuição das notas (1 a 5 estrelas)
    b. Filmes mais e menos avaliados
    c. Média de avaliações por filme
7. Análise Temporal:
    a. Como a distribuição das notas evoluiu ao longo do tempo?
    b. Houve anos em que os usuários foram mais críticos ou mais generosos?
    c. Existe sazonalidade nas avaliações? (ex.: mais avaliações em determinados meses ou anos)
8. Análise do Comportamento dos Usuários:
    a. Identificação de padrões de avaliação: há usuários que tendem a dar notas mais altas ou mais baixas?
    b. Quantidade média de avaliações por usuário
    c. Existe um grupo pequeno de usuários que avalia a maioria dos filmes?
9. Relação entre Ano de Lançamento e Notas:
    a. Filmes mais antigos recebem notas mais baixas ou mais altas?
    b. Existe um padrão de notas para filmes lançados em anos específicos?
10. Visualizações e Insights Gerais:
    a. Histogramas das avaliações por nota
    b. Linha do tempo da média de avaliações por ano
    c. Scatter plot relacionando a quantidade de avaliações e a nota média dos filmes
    d. Análise de outliers: filmes com notas extremamente altas ou baixas com base no volume de avaliações

Este projeto permitirá não apenas compreender a distribuição das avaliações de filmes, mas também levantar hipóteses relevantes sobre o comportamento dos usuários e a qualidade percebida dos filmes ao longo do tempo. A partir dessa EDA, é possível gerar insights para otimizar recomendações e melhorar a experiência do usuário em plataformas de streaming.

Resultados Pretendidos

Explorar os dados e observar resultados da análise, como notas médias, mínimas, máximas, dispersão dos dados, evolução temporal etc. É esperado que esses resultados permitam entender melhor a percepção e comportamento dos assinantes, o que pode levar empresas como a Netflix a ajustar melhor seu portfólio e proporcionar uma melhor experiência e satisfação a seus clientes.

Seção 4 – Scripts da Análise Exploratória de Dados

https://github.com/thiagogodeguesi/ProjetoAplicado1/tree/main/scr

Seção 5 – Análise Exploratória

Nossa Análise Exploratória de Dados buscou conhecer o dataset que escolhemos, em vários aspectos, tais como:

- Esquema
- Cabeçalho
- Formato
- Contagem de valores nulos
- Descrição de todas as variáveis contendo:
    - Contagem
    - Média
    - Desvio padrão
    - Valor mínimo
    - Valor máximo
    - Valor nulos
    - 1º quartil
    - Mediana
    - 3º quartil

Também utilizamos gráficos para visualização dos dados. Nesta fase, informações visuais são um auxílio enorme na compreensão. Até o momento, temos as seguintes visualizações:

- Histograma dos filmes agrupado pela média de visualizações
- Popularidade e a média de avaliação dos filmes
- Total e média das avaliações agrupadas por ano, com o desvio padrão agrupado por ano

Todos os dados coletados do dataset, bem como os gráficos usados para melhor visualização e compreensão dos dados, estão disponíveis no arquivo "*PA1_A2_AplicandoConhecimento_Grupo.ipynb"* que se encontra na pasta "*scr*" do projeto no GutHub. Os arquivos com os dados estão na pasta "*data/input*" do projeto.

Utilizando uma amostra de 134.630 registros, obtivemos:

- 35 registros sem ano de lançamento
- Validamos a distribuição dessas avaliações
- A média geral das avaliações é 3,22
- Desenvolvemos um gráfico de Histograma de filmes por média de avaliações.
- Desenvolvemos um gráfico de Popularidade vs Média de Avaliação dos Filmes.
- Desenvolvemos um gráfico do Total e média das avaliações por ano da votação, considerando o desvio padrão.

Esquema dos Dados

```
Schema([
('MovieID', Int32), ('CustomerID', Int32), ('Rating', Int32),
('Date', String),
('Year', String),
('Title', String)])
```

Cabeçalho do Dataframe

MovieID	CustomerID	Rating	Date	Year
i32	i32	i32	str	str
1	1508350	4	"2005-06-27"	"2003"
1	2165002	4	"2004-04-06"	"2003"
1	1604707	4	"2005-10-17"	"2003"
1	2088415	4	"2005-02-01"	"2003"
1	818416	3	"2005-07-27"	"2003"

Formato do Dataframe

(134630, 6)

Valores Nulos

MovieID	CustomerID	Rating	Date	Year	Title
---	---	---	---	---	---
u32	u32	u32	u32	u32	u32
0	0	0	0	35	0

Descrição do Dataframe

statistic	MovieID	CustomerID	Rating	Date	Year	Title	DateParsed	AnoAvaliação	MesAvaliação
str	f64	f64	f64	str	str	str	str	f64	f64
"count"	134630.0	134630.0	134630.0	"134630"	"134595"	"134630"	"134630"	134630.0	134630.0
"null_count"	0.0	0.0	0.0	"0"	"35"	"0"	"0"	0.0	0.0
"mean"	2241.776974	1.3296e6	3.220181	"2004-10-10 21:48:42.409000"	null	null	"2004-10-10 21:48:42.409000"	2004.244388	6.917522
"std"	1297.966552	765081.369003	1.225771	null	null	null	null	1.013084	3.348264
"min"	1.0	6.0	1.0	"1999-12-30"	"1915"	"'N Sync: 'N the Mix"	"1999-12-30"	1999.0	1.0
"25%"	1114.0	664090.0	2.0	"2004-04-10"	null	null	"2004-04-10"	2004.0	4.0
"50%"	2236.0	1.328713e6	3.0	"2005-01-25"	null	null	"2005-01-25"	2005.0	7.0
"75%"	3365.0	1.996999e6	4.0	"2005-07-26"	null	null	"2005-07-26"	2005.0	10.0
"max"	4499.0	2.649388e6	5.0	"2005-12-31"	"2005"	"s-Cry-ed"	"2005-12-31"	2005.0	12.0

Distribuição das Avaliações

Rating

struct[2]

{1,15413}

{2,20270}

{3,41282}

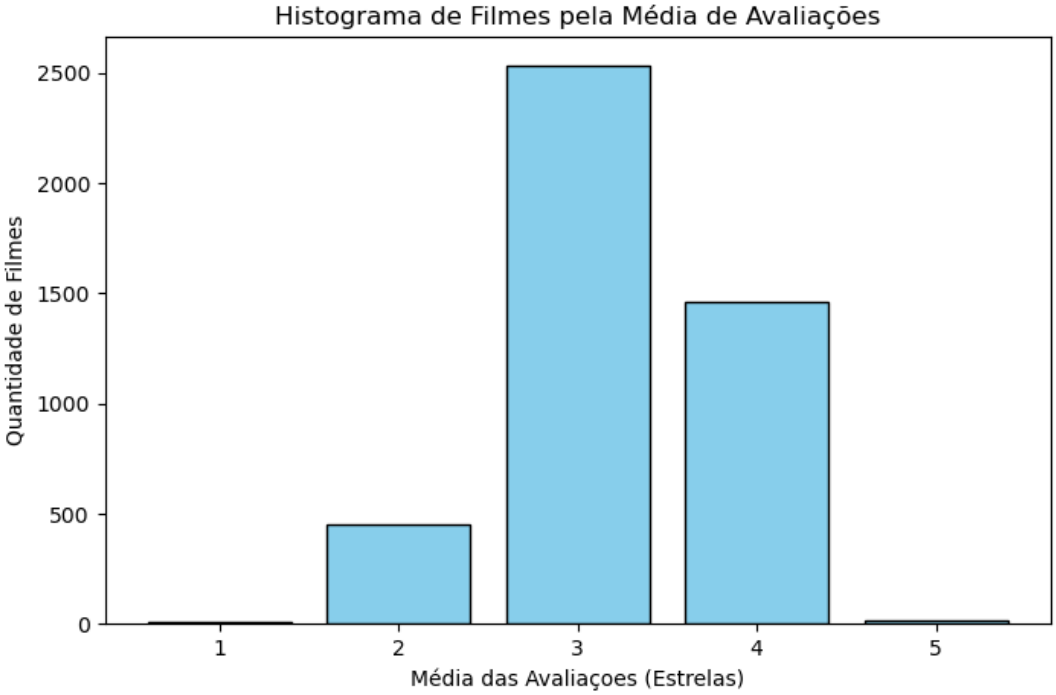
{4,34591}

{5,23074}

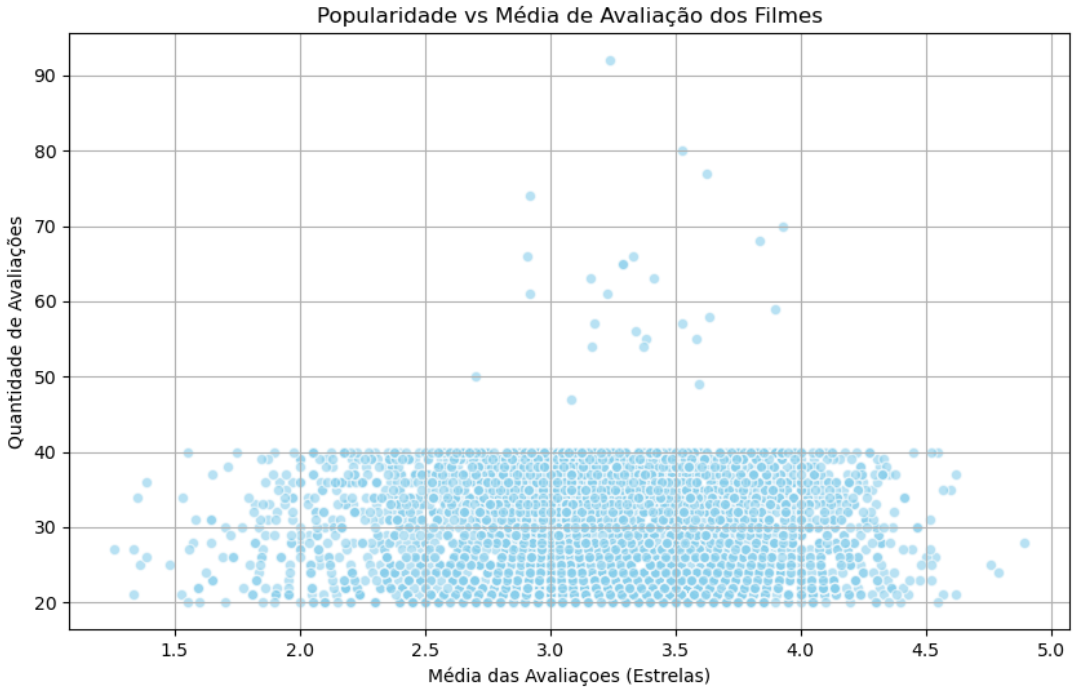
Média Geral das Avaliações

```
3.2201812374656464
```

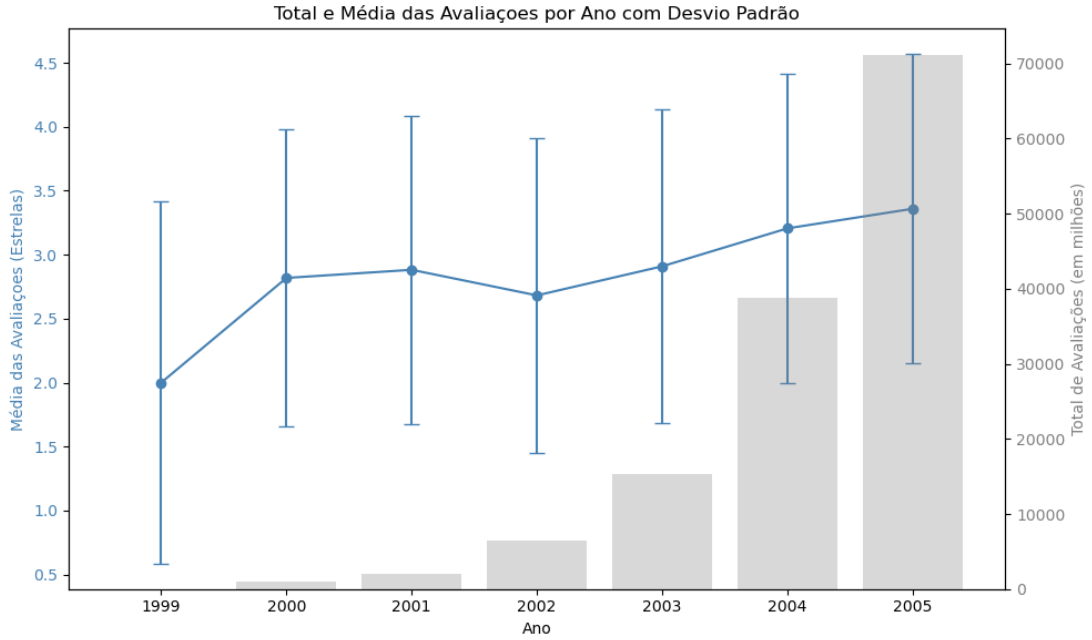
Histórico de Filmes pela Média de Avaliações



Popularidade vs Média de Avaliação dos Filmes



Total e Média das Avaliações por Ano com Desvio Padrão



Conforme a imagem acima, no período considerado (1999-2005), o número de avaliações por ano aumentou exponencialmente a partir de 2002. No geral, observa-se que a média das notas dos filmes por ano também foi aumentando ao longo do período avaliado.

Dessa forma, o cenário da amostra nos permite delinear um padrão: quanto maior foi o número de avaliações a cada ano, maior também tendeu a ser a nota média dos filmes.

No sentido contrário, o único ano da amostra que parece contrariar esse padrão é 2002, quando houve um considerável aumento no volume de avaliações comparado aos anos anteriores, mas a nota média teve uma pequena queda.

Outro ponto interessante a ser notado são os desvios-padrão das notas, que se mantiveram relativamente constantes ao longo dos anos da análise. Assim, temos um centro de notas que vai aumentando no período, com uma dispersão em torno desse centro que se mantém mais ou menos igual.

Por fim, é interessante notar uma variação mais significativa na nota média dos filmes do ano de 1999 para 2000, os dois anos com os menores números de avaliações na amostra.

Embora a nota atribuída a um filme possa depender de diferentes fatores, tais como a qualidade intrínseca dos filmes e a percepção subjetiva dos assinantes, podemos também considerar a estatística e supor que o tamanho pequeno das amostras (pequeno número de avaliações) desses anos possa ter interferido nos resultados, distorcendo-os em alguma medida.

Seção 6 - Bibliografia

Referências

GODEGUESI, Thiago; CARMO, Cristiano Prado do; BRATHWAITE, Felipe Garcia Pereira; SELLA, Ingryd Cristine Hidalgo; DIAS, Pablo Rodrigo. **Projeto Aplicado 1**. Disponível em: <https://github.com/thiagogodeguesi/ProjetoAplicado1>. Acesso em: 03 mar. 2025.
NETFLIX INC. Netflix Prize Data. Disponível em: <https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data>. Acesso em: 28 fev. 2025.

KOREN, Yehuda. The BellKor Solution to the Netflix Grand Prize. 2009. Disponível em https://www2.seas.gwu.edu/~simhaweb/champalg/cf/papers/KorenBellKor2009.pdf. Acesso em 02 mar. 2025.

TÖSCHER, Andreas; JAHRER, Michael; BELL, Robert M. The BigChaos Solution to the Netflix Grand Prize. Commendo Research & Consulting, Neuer Weg 23, A-8580 Köflach, Áustria; AT&T Labs - Research, Florham Park, NJ, 2009. Disponível em https://www.asc.ohio-state.edu/statistics/statgen/joul_aut2009/BigChaos.pdf. Acesso em 02 mar. 2025.

SAWERS, Paul. Remember Netflix's $1m algorithm contest? Well, here's why it didn't use the winning entry. 2012. Disponível em https://thenextweb.com/news/remember-netflixs-1m-algorithm-contest-well-heres-why-it-didnt-use-the-winning-entry. Acesso em 02 mar. 2025.