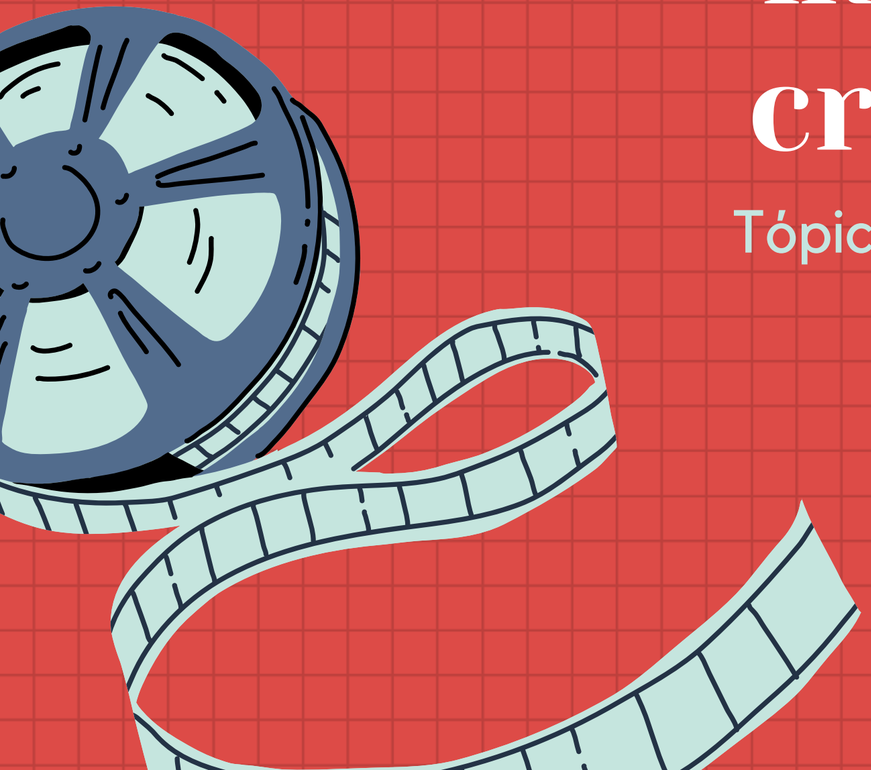
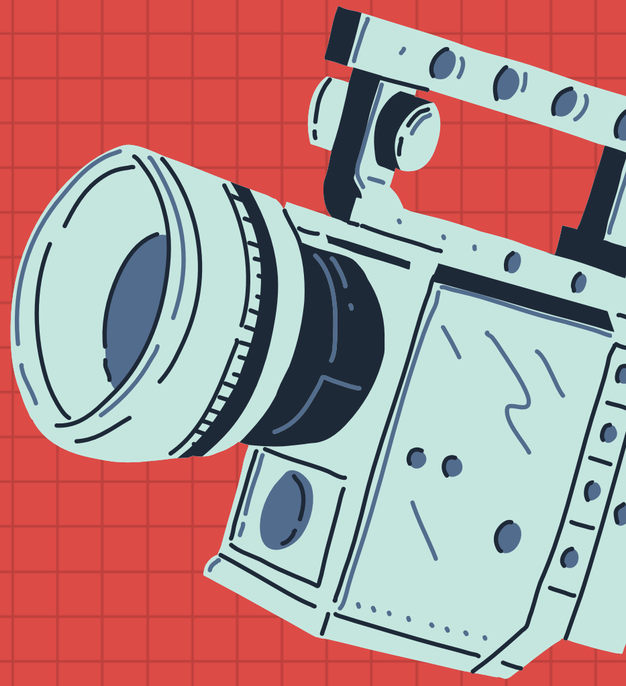
INE5454

Web Scraping dos maiores sites de críticas de filme

Tópicos Especiais em Gerência de Dados





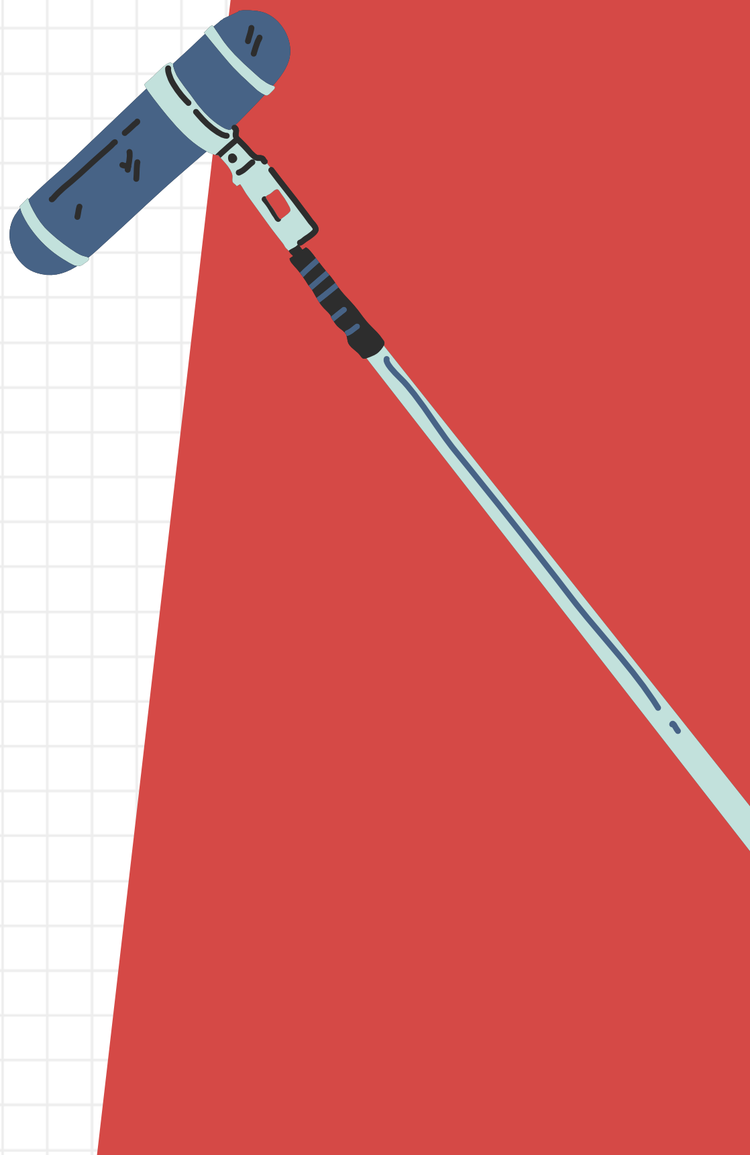
Cleberton Oliveira
Kevin Rafael
Lucas Martins

Introdução

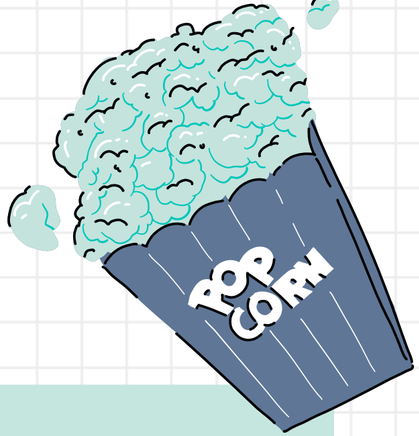
O tema escolhido foram os **sites de crítica de filme** onde especialistas e entusiastas do cinema discutem e avaliam com resenhas ou notas os filmes, fornecendo análises detalhadas sobre aspectos como roteiro, direção, atuação, cinematografia e música.

Esses sites influenciam as percepções do público e ajudam na tomada de decisão na hora de assistir a filmes.

O Objetivo desse trabalho é fazer a **extração de dados** que decidimos como mais importantes dos **maiores sites de critica de cinema.**



Definições



Crawler

Programa automatizado para navegar e coletar informações de forma sistemática como as páginas.

Domínio

Filmes e críticas relacionados ao universo cinematográfico

- Grande volume de dados textuais (críticas e análises).
- Formatos diversos entre os sites.

Entidades do Mundo Real

- Nome do Filme
- Nota Crítica
- Nota Audiencia
- Nome do critico
- Sinopse

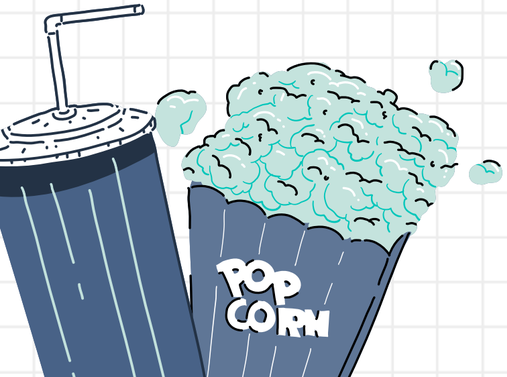
Público Interessado

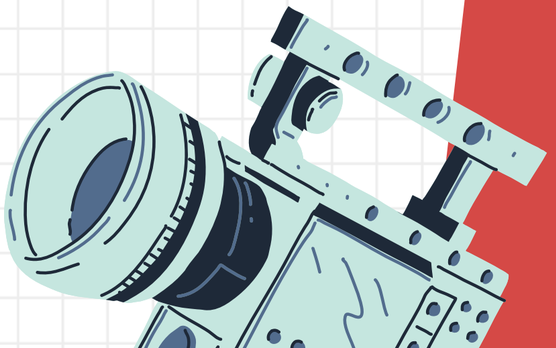
- Leitores e espectadores (buscam recomendações)
- Estúdios e produtores de filme (para insights de público).
- Diretores, atores e equipe de produção
- Plataforma de Streaming para melhorar seus catálogos

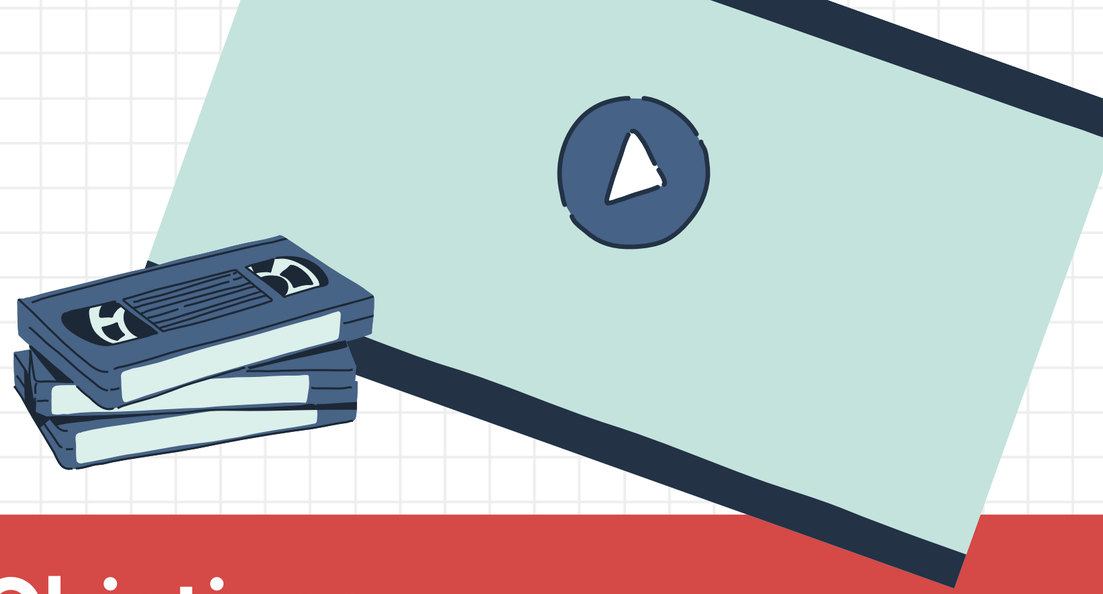
Motivação

Centralizar, organizar e sintetizar dados dispersos para facilitar análises, como:

- Identificar filmes subestimados
- Visualizar diferenças de gosto entre críticos e público para analisar tendencias ou discutir teorias cinematográficas
- Entender a recepção de filmes ou gêneros de filmes no mercado





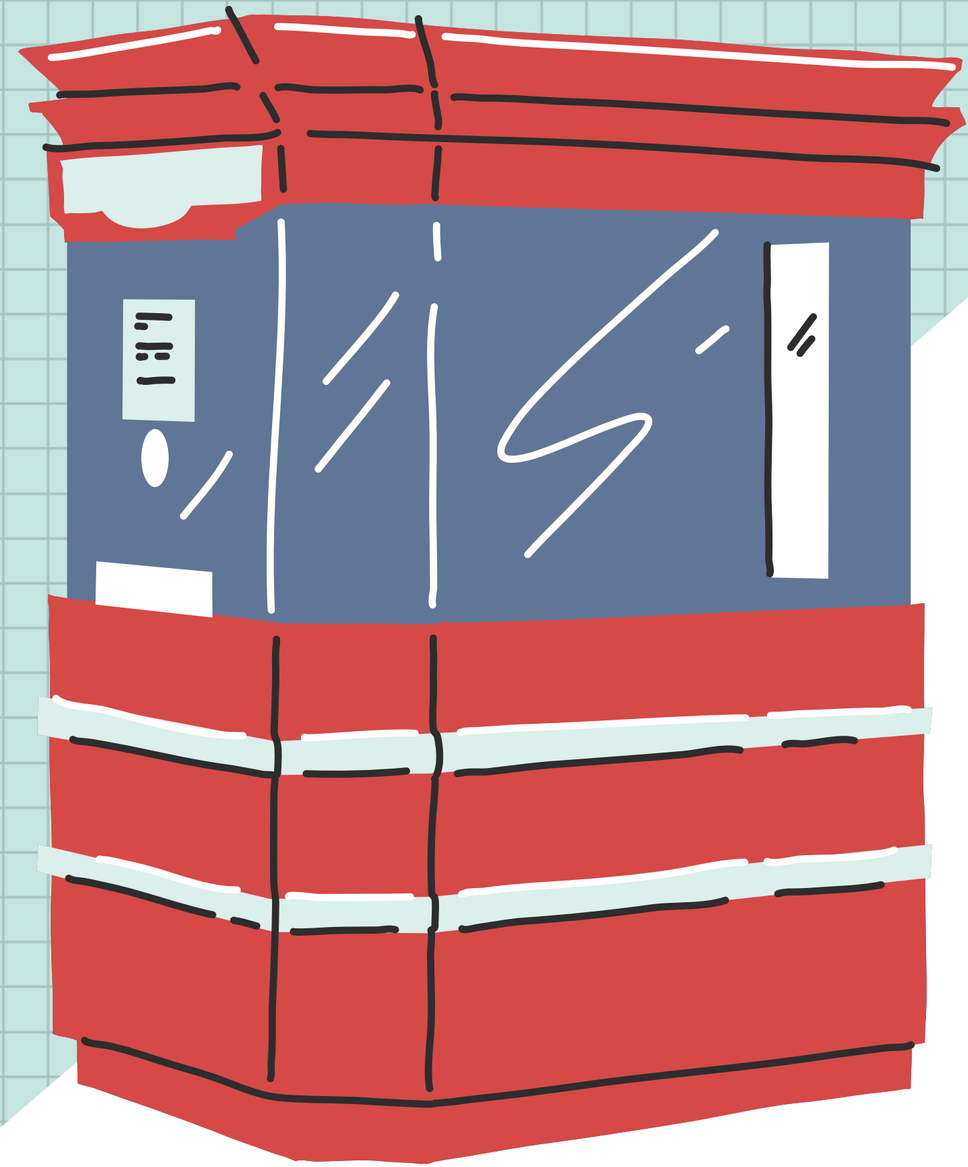


Objetivos

- Construir um dataset consolidado com informações de críticas e avaliações de filmes.
- Aplicar análises comparativas e sintetizar dados para melhor visualização e acompanhamento de tendencias ou estatísticas no mundo cinematográfico

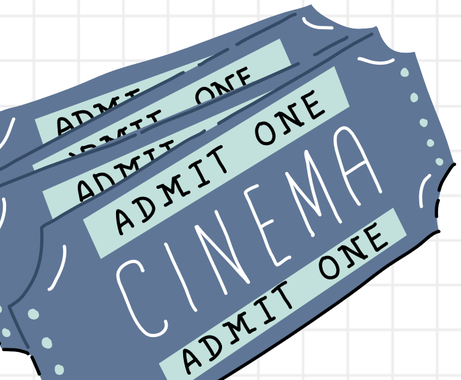
Sites extraídos

1. AdoroCinema
2. RottenTomatoes
3. RogerEbert



Exemplo de Entrada

Página HTML do RottenTomatoes




```
  slot="image"
  loading="lazy"
  src="https://resizing.flixster.com/yZBsfbpElnPRs59-m52IoUn97Wk=/206x305/v2/https://resizir
  alt="Moana 2 poster"
></rt-img>

  <button
    aria-label="Play Moana 2 trailer"
    class="transparent unset"
    data-content-type="movie"
    data-disable-ads=""
    data-ems-id="4f945245-5bc5-3121-abf2-2b23228ac6ec"
    data-mpx-id="2375559235526"
    data-position="2"
    data-public-id="h6X1HgXu_nik"

    data-title="Moana 2"
    data-track="poster"
    data-type="Movie"
    data-VideoPlayerOverlayManager="btnVideo:click"
    data-video-list="rt-hp-poster-list-coming-soon"
    slot="imageAction"

  >
    <span class="sr-only">Moana 2</span>
  </button>

  <a

    data-track="scores"
    href="/m/moana_2"
    slot="caption"
  >


  <score-pairs-deprecated>
    <score-icon-critics
      slot="criticsScoreIcon"
      certified="false"
      size="1"
      sentiment="positive"
    ></score-icon-critics>
    <rt-text slot="criticsScore" context="label" size="1"> 62%</rt-text>
    <score-icon-audience
      slot="audienceScoreIcon"
      certified="false"
      size="1"
      sentiment="positive"
    ></score-icon-audience>
    <rt-text slot="audienceScore" context="label" size="1"> 87%</rt-text>
  </score-pairs-deprecated>


  <span class="p--small">Moana 2</span>
```

Exemplo de Saida

Objeto JSON gerado


```
{
    "NUMERO": "92",
    "TITULO": "Paper Moon",
    "CRITIC SCORE": " 90%",
    "AUDIENCE SCORE": " 94%",
    "CRITICO": "",
    "SINOPSE": "",
    "LINK": "https://www.rottentomatoes.com/m/paper_moon"
},
{
    "NUMERO": "93",
    "TITULO": "Paris, Texas",
    "CRITIC SCORE": " 95%",
    "AUDIENCE SCORE": " 93%",
    "CRITICO": "",
    "SINOPSE": "",
    "LINK": "https://www.rottentomatoes.com/m/paris_texas"
},
{
    "NUMERO": "94",
    "TITULO": "Autumn Sonata",
    "CRITIC SCORE": " 85%",
    "AUDIENCE SCORE": " 92%",
    "CRITICO": "",
    "SINOPSE": "",
    "LINK": "https://www.rottentomatoes.com/m/autumn_sonata"
},
{
    "NUMERO": "95",
    "TITULO": "Trolls Band Together",
    "CRITIC SCORE": " 64%",
    "AUDIENCE SCORE": " 92%",
    "CRITICO": "",
    "SINOPSE": "",
    "LINK": "https://www.rottentomatoes.com/m/trolls_band_together"
},
```

Desenvolvimento do trabalho

1º Passo

Definição do Tema

2º Passo

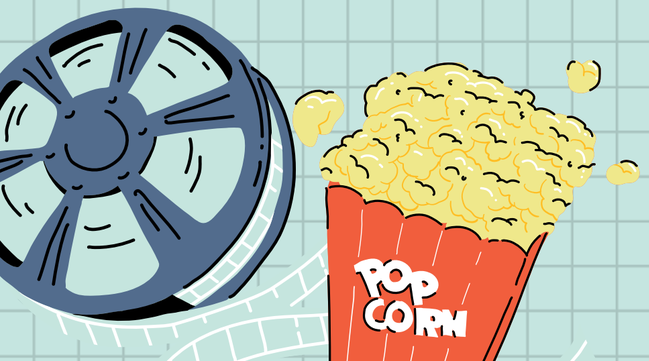
Escolha dos links

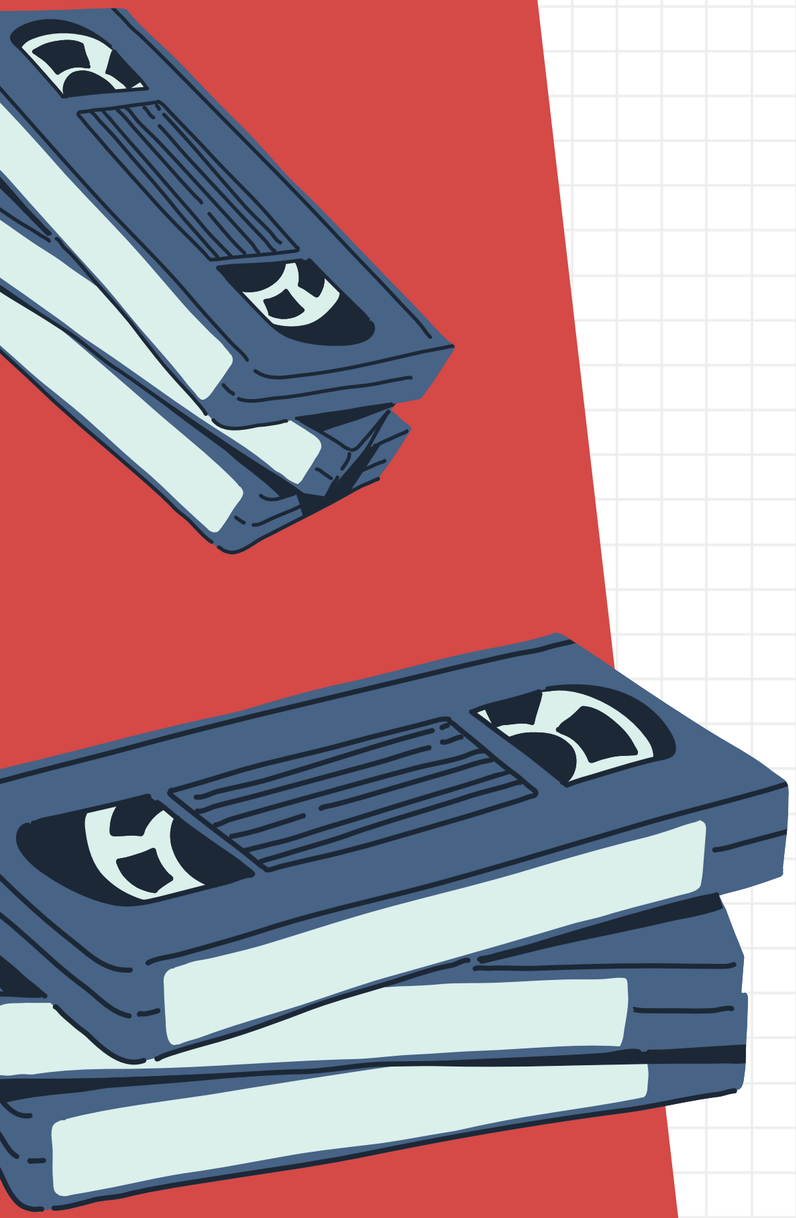
3º Passo

Extração de dados de cada site separadamente

4º Passo

Junção e normalização dos dados





Infraestutura tecnológica

- **Linguagem:** Python
- **Pacotes:** BeautifulSoup, json, requests
- **Frameworks:** nenhum
- **Plugins:** nenhum

Outras informações

Qual foi o formato de dados coletado/extraídos: HTML

De quais locais ou áreas geográficas: Brasil

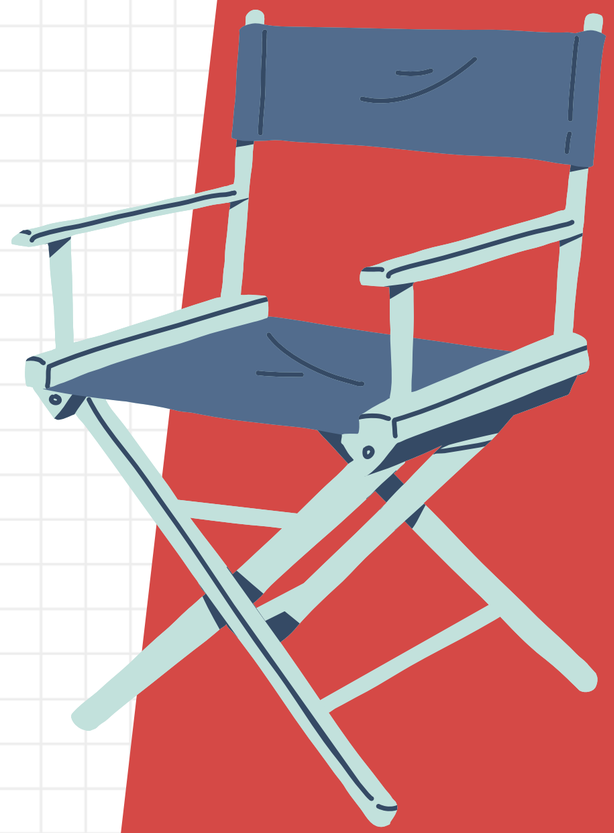
Qual o idioma: Português / Inglês

Quais foram os atributos coletados/extraídos:

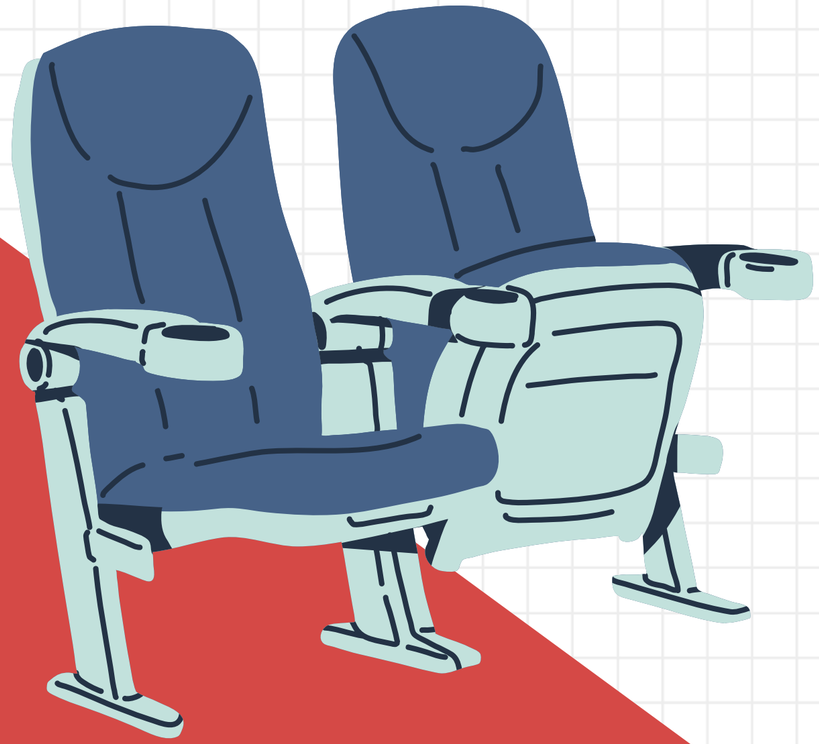
- Numero (quantitativo)
- Titulo (qualitativo)
- Critic score (quantitativo)
- Audience score (quantitativo)
- Critico (qualitativo)
- Sinopse (qualitativo)
- Link (qualitativo)

Data da coleta: 16 de novembro de 2024

Por quanto tempo os dados foram coletados: Apenas um dia



Casos de falha



- **Falhas**

Tivemos problemas ao tentar coletar informações dos sites IMDB, MetaCritic porque ambos os sites não deram permissão para a coleta.

- **Dificuldades**

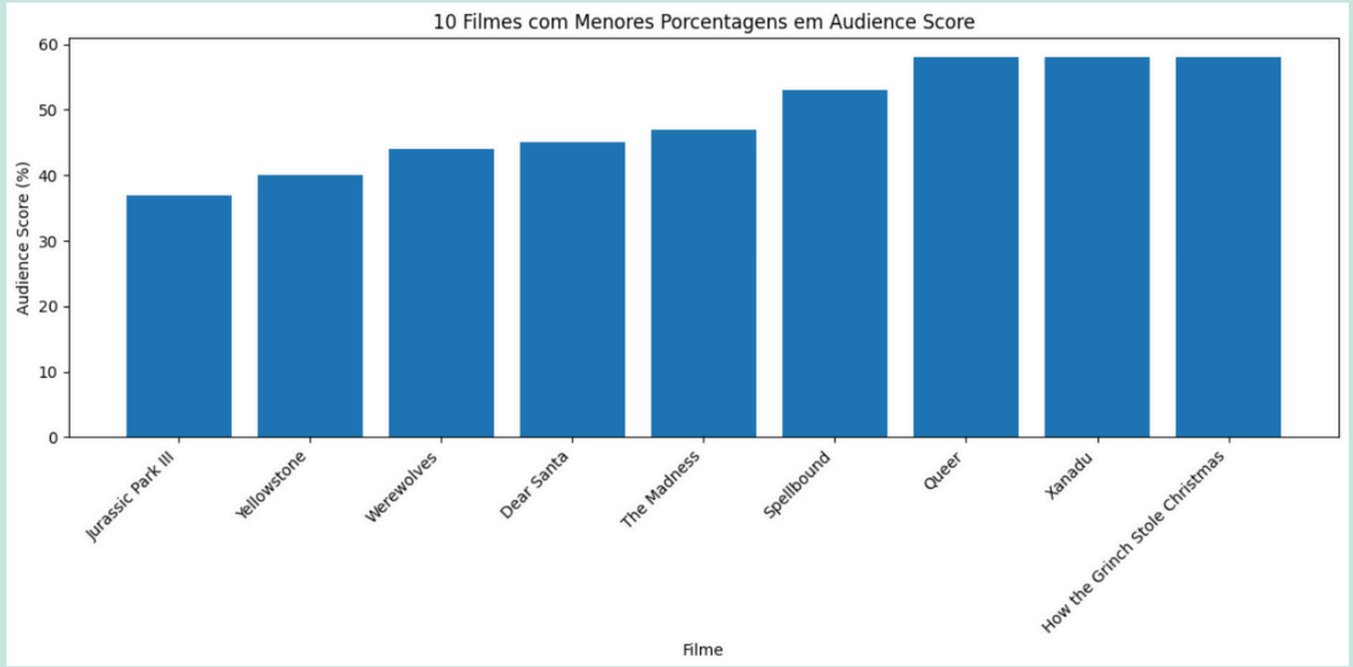
Na coleta do site AdoroCinema tivemos complicaçõe em relação a estruturação da página inicial.

- **Minerou coisas que não deveria**

Minerou notícias.

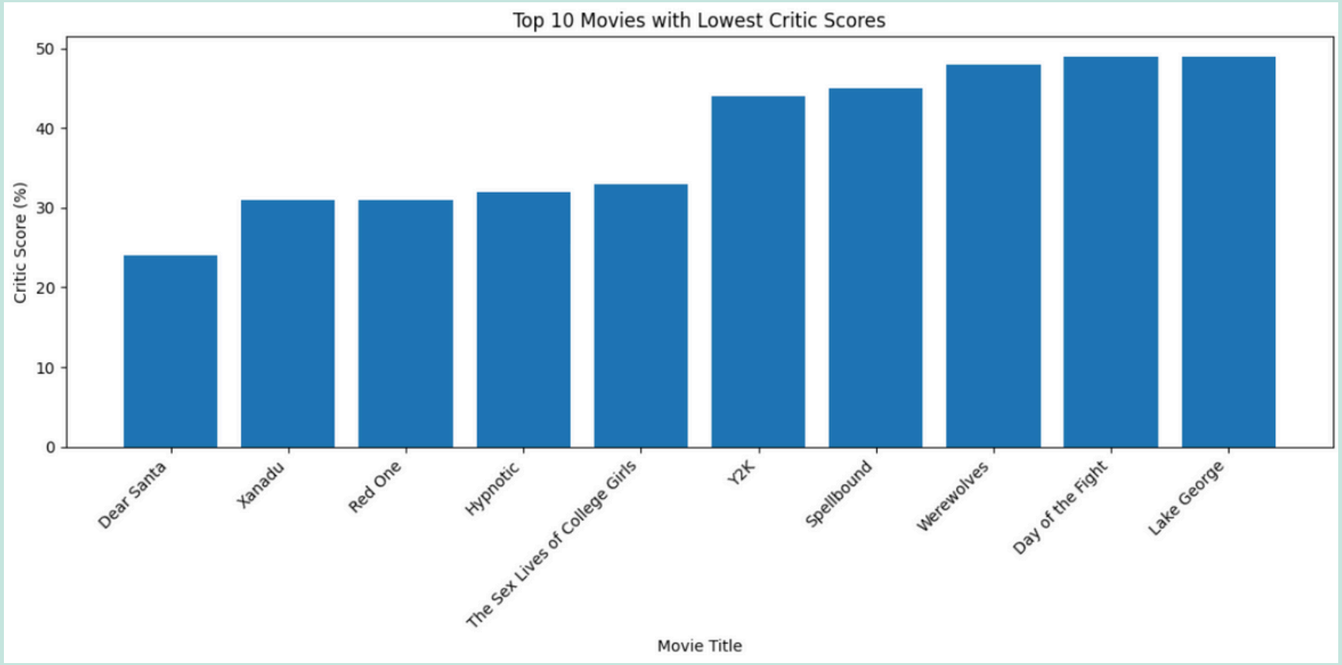
Resultados finais

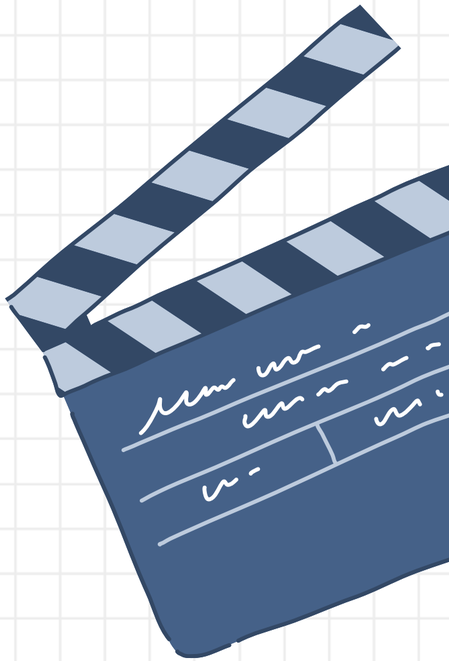
Os 10 piores filmes segundo a audiência

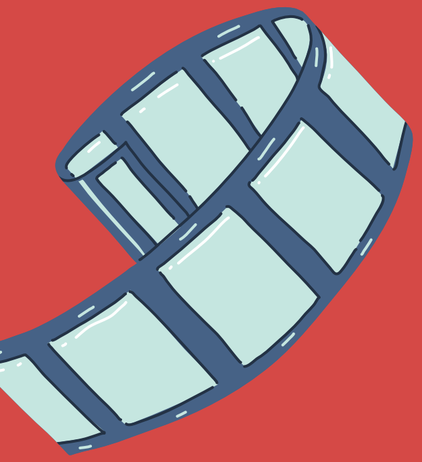


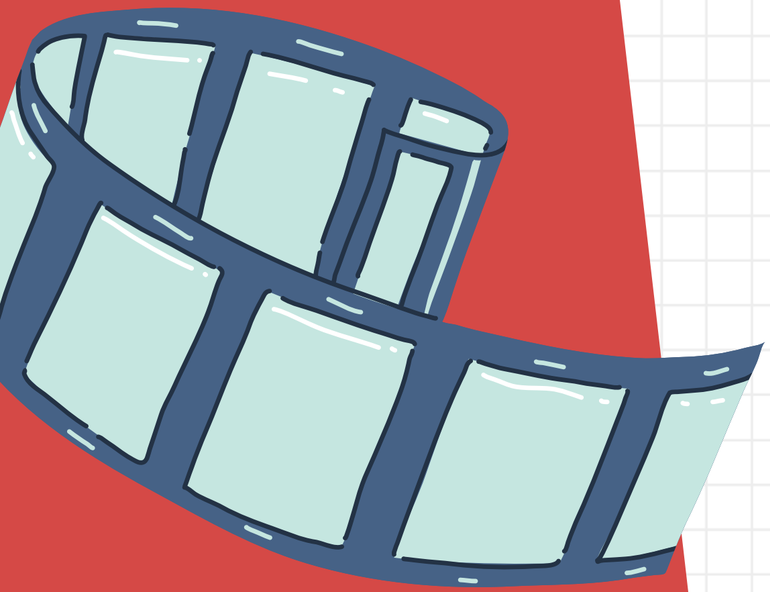
Resultados finais

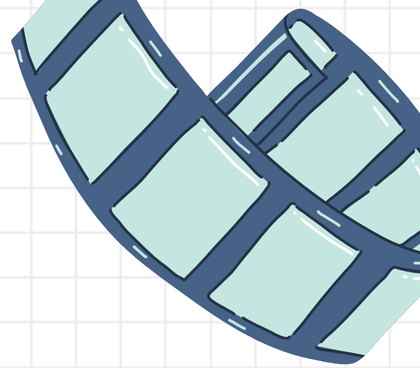
Os 10 piores filmes segundo a crítica











Conclusão

O que teriam feito diferente?

O trabalho tem potencial para ser algo visto com bons olhos pelo mercado? Porque? O que faltaria desenvolver para ser um produto?