

# Web Scraping - **As músicas mais ouvidas de 1950 até 2022**

Tópicos especiais em gerência de dados – INE5454

Amanda Santiago da Costa (19201689) e Erich Hanemann Jr (20100842)

# Extração e análise de dados musicais de 1950 até 2022

- Crawler e Extractor.
- O domínio envolve a coleta de dados históricos sobre as músicas mais ouvidas ao longo de várias décadas, de 1950 até 2022.
- Entidades do mundo real envolvidas: Músicas, artistas, anos, e informações adicionais da Wikipedia sobre cada música.
- Público interessado: Pesquisadores musicais, entusiastas e profissionais de música e desenvolvedores de aplicativos de música.

# Resultado esperado

- Ao final do trabalho, espera-se ter um conjunto de dados estruturado (JSON) contendo informações sobre as músicas mais populares de cada ano entre 1950 e 2022.

# Escolha do tema

A música é um aspecto cultural significativo e entender suas tendências ao longo do tempo pode proporcionar insights valiosos. A extração de dados históricos musicais pode beneficiar várias áreas, incluindo pesquisa acadêmica, desenvolvimento de software musical e análise de mercado para a indústria da música.

# Objetivo

Entrada analisada:

- <https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/>
- <https://www.revistabula.com/15910-2-a-musica-mais-tocada-no-ano-em-que-voce-nasceu/>

Extração de dados do valores coletados da entrada:

- [www.wikipedia.org](http://www.wikipedia.org)

Saída


Um arquivo JSON (musicas\_final.json) que contém uma lista de 2 dicionários. Cada dicionário representa uma url e contém outra lista, que inclui detalhes como o ano da música, nome da música, artista, informações adicionais extraídas da Wikipedia sobre a música e o artista, e a URL da página de origem

# Dados de entrada

https://www.fatosdesconhecidos.com.br/essas-sai 60%

Veja a lista das **músicas** mais ouvidas no Reino Unido abaixo.

## 1950's



1952 – Singin' In The Rain by Gene Kelly

1953 – That's Amore by Dean Martin

1954 – I've Got A Woman by Ray Charles

1955 – Tutti Frutti by Little Richard

1956 – I Walk The Line by Johnny Cash

1957 – Jailhouse Rock by Elvis Presley

1958 – Johnny B. Goode by Chuck Berry

1959 – Put Your Head On My Shoulder by Paul Anka

revistabula.com/15910-2-a-musica-mais-tocada-no-ano-em-que-voce-nasceu/

## Bula

INÍCIO SEÇÕES ANU

1950 – Goodnight Irene (Gordon Jenkins e The Weavers)

1951 – Too Young (Nat King Cole)

1952 – Blue Tango (Leroy Anderson)

1953 – Música de Moulin Rouge (Percy Faith)

1954 – Little Things Mean A Lot (Kitty Kallen)

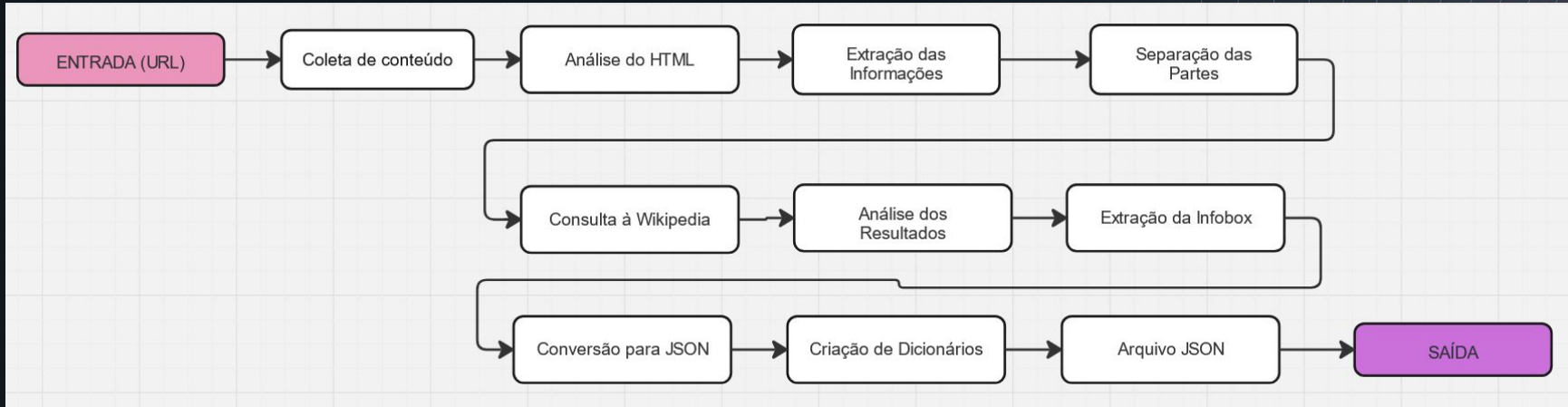
1955 – Cherry Pink And Apple Blossom White (Perez Prado)

## Dados de saída

```
main.py x {} musicas.json x
{} musicas.json > {} 1 > {} Wikipedia Info
1 [
2 {
3   "Ano": "1952",
4   "Nome da Música": "Singin' In The Rain",
5   "Artista": "Gene Kelly",
6   "Wikipedia Info": {
7     "Language": "English",
8     "Published": "1929(1929)",
9     "Songwriter(s)": "Arthur Freed",
10    "Composer(s)": "Nacio Herb Brown"
11  },
12  "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
13 },
14 {
15   "Ano": "1953",
16   "Nome da Música": "That's Amore",
17   "Artista": "Dean Martin",
18   "Wikipedia Info": {
19     "B-side": "\"You're The Right One\"",
20     "Released": "November 7, 1953",
21     "Recorded": "August 13, 1953 at Capitol Studios, Hollywood",
22     "Genre": "Pop",
23     "Length": "3:05",
24     "Label": "Capitol",
25     "Composer(s)": "Harry Warren",
26     "Lyricist(s)": "Jack Brooks"
27   },
28   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
29 },
30 {
31   "Ano": "1954",
32   "Nome da Música": "The Way We Were",
33   "Artista": "Barbra Streisand",
34   "Wikipedia Info": {
35     "B-side": "None",
36     "Released": "February 1, 1968",
37     "Recorded": "1967",
38     "Genre": "Pop",
39     "Length": "3:50",
40     "Label": "Capitol",
41     "Composer(s)": "Barbra Streisand",
42     "Lyricist(s)": "Barbra Streisand"
43   },
44   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
45 },
46 {
47   "Ano": "1955",
48   "Nome da Música": "The Way We Were",
49   "Artista": "Barbra Streisand",
50   "Wikipedia Info": {
51     "B-side": "None",
52     "Released": "February 1, 1968",
53     "Recorded": "1967",
54     "Genre": "Pop",
55     "Length": "3:50",
56     "Label": "Capitol",
57     "Composer(s)": "Barbra Streisand",
58     "Lyricist(s)": "Barbra Streisand"
59   },
60   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
61 },
62 {
63   "Ano": "1956",
64   "Nome da Música": "The Way We Were",
65   "Artista": "Barbra Streisand",
66   "Wikipedia Info": {
67     "B-side": "None",
68     "Released": "February 1, 1968",
69     "Recorded": "1967",
70     "Genre": "Pop",
71     "Length": "3:50",
72     "Label": "Capitol",
73     "Composer(s)": "Barbra Streisand",
74     "Lyricist(s)": "Barbra Streisand"
75   },
76   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
77 },
78 {
79   "Ano": "1957",
80   "Nome da Música": "The Way We Were",
81   "Artista": "Barbra Streisand",
82   "Wikipedia Info": {
83     "B-side": "None",
84     "Released": "February 1, 1968",
85     "Recorded": "1967",
86     "Genre": "Pop",
87     "Length": "3:50",
88     "Label": "Capitol",
89     "Composer(s)": "Barbra Streisand",
90     "Lyricist(s)": "Barbra Streisand"
91   },
92   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
93 },
94 {
95   "Ano": "1958",
96   "Nome da Música": "The Way We Were",
97   "Artista": "Barbra Streisand",
98   "Wikipedia Info": {
99     "B-side": "None",
100    "Released": "February 1, 1968",
101    "Recorded": "1967",
102    "Genre": "Pop",
103    "Length": "3:50",
104    "Label": "Capitol",
105    "Composer(s)": "Barbra Streisand",
106    "Lyricist(s)": "Barbra Streisand"
107  },
108  "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
109 },
110 {
111   "Ano": "1959",
112   "Nome da Música": "The Way We Were",
113   "Artista": "Barbra Streisand",
114   "Wikipedia Info": {
115     "B-side": "None",
116     "Released": "February 1, 1968",
117     "Recorded": "1967",
118     "Genre": "Pop",
119     "Length": "3:50",
120     "Label": "Capitol",
121     "Composer(s)": "Barbra Streisand",
122     "Lyricist(s)": "Barbra Streisand"
123   },
124   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
125 },
126 {
127   "Ano": "1960",
128   "Nome da Música": "The Way We Were",
129   "Artista": "Barbra Streisand",
130   "Wikipedia Info": {
131     "B-side": "None",
132     "Released": "February 1, 1968",
133     "Recorded": "1967",
134     "Genre": "Pop",
135     "Length": "3:50",
136     "Label": "Capitol",
137     "Composer(s)": "Barbra Streisand",
138     "Lyricist(s)": "Barbra Streisand"
139   },
140   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
141 },
142 {
143   "Ano": "1961",
144   "Nome da Música": "The Way We Were",
145   "Artista": "Barbra Streisand",
146   "Wikipedia Info": {
147     "B-side": "None",
148     "Released": "February 1, 1968",
149     "Recorded": "1967",
150     "Genre": "Pop",
151     "Length": "3:50",
152     "Label": "Capitol",
153     "Composer(s)": "Barbra Streisand",
154     "Lyricist(s)": "Barbra Streisand"
155   },
156   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
157 },
158 {
159   "Ano": "1962",
160   "Nome da Música": "The Way We Were",
161   "Artista": "Barbra Streisand",
162   "Wikipedia Info": {
163     "B-side": "None",
164     "Released": "February 1, 1968",
165     "Recorded": "1967",
166     "Genre": "Pop",
167     "Length": "3:50",
168     "Label": "Capitol",
169     "Composer(s)": "Barbra Streisand",
170     "Lyricist(s)": "Barbra Streisand"
171   },
172   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
173 },
174 {
175   "Ano": "1963",
176   "Nome da Música": "The Way We Were",
177   "Artista": "Barbra Streisand",
178   "Wikipedia Info": {
179     "B-side": "None",
180     "Released": "February 1, 1968",
181     "Recorded": "1967",
182     "Genre": "Pop",
183     "Length": "3:50",
184     "Label": "Capitol",
185     "Composer(s)": "Barbra Streisand",
186     "Lyricist(s)": "Barbra Streisand"
187   },
188   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
189 },
190 {
191   "Ano": "1964",
192   "Nome da Música": "The Way We Were",
193   "Artista": "Barbra Streisand",
194   "Wikipedia Info": {
195     "B-side": "None",
196     "Released": "February 1, 1968",
197     "Recorded": "1967",
198     "Genre": "Pop",
199     "Length": "3:50",
200     "Label": "Capitol",
201     "Composer(s)": "Barbra Streisand",
202     "Lyricist(s)": "Barbra Streisand"
203   },
204   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
205 },
206 {
207   "Ano": "1965",
208   "Nome da Música": "The Way We Were",
209   "Artista": "Barbra Streisand",
210   "Wikipedia Info": {
211     "B-side": "None",
212     "Released": "February 1, 1968",
213     "Recorded": "1967",
214     "Genre": "Pop",
215     "Length": "3:50",
216     "Label": "Capitol",
217     "Composer(s)": "Barbra Streisand",
218     "Lyricist(s)": "Barbra Streisand"
219   },
220   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
221 },
222 {
223   "Ano": "1966",
224   "Nome da Música": "The Way We Were",
225   "Artista": "Barbra Streisand",
226   "Wikipedia Info": {
227     "B-side": "None",
228     "Released": "February 1, 1968",
229     "Recorded": "1967",
230     "Genre": "Pop",
231     "Length": "3:50",
232     "Label": "Capitol",
233     "Composer(s)": "Barbra Streisand",
234     "Lyricist(s)": "Barbra Streisand"
235   },
236   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
237 },
238 {
239   "Ano": "1967",
240   "Nome da Música": "The Way We Were",
241   "Artista": "Barbra Streisand",
242   "Wikipedia Info": {
243     "B-side": "None",
244     "Released": "February 1, 1968",
245     "Recorded": "1967",
246     "Genre": "Pop",
247     "Length": "3:50",
248     "Label": "Capitol",
249     "Composer(s)": "Barbra Streisand",
250     "Lyricist(s)": "Barbra Streisand"
251   },
252   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
253 },
254 {
255   "Ano": "1968",
256   "Nome da Música": "The Way We Were",
257   "Artista": "Barbra Streisand",
258   "Wikipedia Info": {
259     "B-side": "None",
260     "Released": "February 1, 1968",
261     "Recorded": "1967",
262     "Genre": "Pop",
263     "Length": "3:50",
264     "Label": "Capitol",
265     "Composer(s)": "Barbra Streisand",
266     "Lyricist(s)": "Barbra Streisand"
267   },
268   "Fonte URL": "https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/"
269 },
270 {
271   "Ano": "1969",
272   "Nome da Música": "The Way We Were",
273   "Artista": "Barbra Streisand",
274   "Wikipedia Info": {
275     "B-side": "None",
276     "Released": "February 1, 1968",
277     "Recorded": "1967",
278     "Genre": "Pop",
279     "Length":
```



# Descrição - Fluxo





# Crawler extractor musics

A implementação do projeto foi desenvolvida em Python, para o Crawler foi utilizado a biblioteca Selenium, e para extração de dados BeautifulSoup.

Dados coletados/extraídos?

<https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/> - Página da web onde contém uma lista das músicas mais ouvidas, no Reino Unido, de 1950 até 2022

# Crawler extractor musics

Dados coletados/extraídos?

<https://www.revistabula.com/15910-2-a-musica-mais-tocada-no-ano-em-que-voce-nasceu/> - Página da web onde contém uma lista, elaborada pela revista americana Billboard, das músicas mais ouvidas de 1946 até 2022. Para ter uma comparação com a outra fonte de dados, resolvemos normalizar os dados partindo de 1950.

Wikipedia: Informações adicionais sobre as músicas e os artistas foram extraídas das páginas correspondentes no Wikipedia, acessadas via API e scraping direto.

# Crawler extractor music

Quem estaria interessado em comprar os dados coletados/extraídos?

Pesquisadores musicais, desenvolvedores de aplicativos musicais, plataformas de Streaming de música, acadêmicos e estudantes.

De quais locais ou áreas geográficas os dados foram coletados/extraídos?

Os dados foram coletados de duas páginas web brasileiras: fatosdesconhecidos, e revistabula, que são acessíveis globalmente. No entanto, a lista da fatosdesconhecidos é sobre as músicas mais ouvidas no Reino Unido e a lista da revista bula é com base em uma lista elaborada pela revista americana Billboard.

Qual o formato de dados coletados/extraídos?

Os dados coletados foram estruturados e armazenados no formato JSON. Cada entrada no arquivo JSON inclui informações como o ano, nome da música, artista, informações adicionais da Wikipedia e a URL da página de origem.

# Crawler extractor music – Atributos coletados/extraídos

## Ano

Tipo: Quantitativo

Descrição: O ano em que a música foi mais ouvida.

## Nome da Música

Tipo: Qualitativo

Descrição: O título da música.

## Artista

Tipo: Qualitativo

Descrição: O nome do artista ou banda que performa a música.

## Wikipedia Info

Tipo: Ambos (Qualitativo e Quantitativo)

Descrição: Informações adicionais extraídas da infobox da Wikipedia.

## Fonte URL

Tipo: Qualitativo

Descrição: A URL da página de origem de onde as informações foram coletadas.

# Crawler extractor music - Fontes de dados e Seeds

## Página da web

Seeds:

<https://www.fatosdesconhecidos.com.br/essas-sao-as-musicas-mais-ouvidas-de-1950-ate-2022/>

<https://www.revistabula.com/15910-2-a-musica-mais-tocada-no-ano-em-que-voce-nasceu/>

Descrição: Estas páginas contêm listas das músicas mais ouvidas de 1950 até 2022, que foram usadas como ponto de partida para coletar os nomes das músicas e dos artistas.

# Crawler extractor musics - Fontes de dados e Seeds

## Wikipedia:

Seeds: URL formatada para a consulta à API do Wikipedia baseada na combinação de nome da música e artista.

Descrição: Usada para obter informações adicionais sobre as músicas e artistas.

Exemplo: 

```
def get_wikipedia_info(artista, musica):  
    search_url = f"https://en.wikipedia.org/w/api.php?action=query&list=search&srsearch={musica} {artista} song&format=json"
```

## Locais que não eram Seeds e Foram Alcançados:

Páginas específicas da Wikipedia retornadas pela API. Por exemplo, uma busca na API do Wikipedia retorna a URL específica da página da música ou do artista que foi então acessada para extrair dados da infobox.

# Crawler extractor musics

Por quanto tempo os dados foram coletados (o escopo da coleta/extração de dados)?

As urls listam as músicas mais ouvidas dos últimos 72 anos.

Os dados foram extraídos apenas 1 vez por url, pois não mudaram após a primeira coleta.

Tempo de execução: 353.81 segundos. (2,4 segundos por música/item)

Data da coleta: 16/06/2024



# Análise dos resultados finais

- Os dados coletados mostram apenas quais músicas foram mais ouvidas, ano após ano, de 1950 até 2022, por isso a análise dos resultados foi relativamente simples.
- Após a extração de dados e geração do json, foi criado um csv para a análise dos dados obtidos.

# Análise dos resultados finais

Script `generate_csv.py` de geração do csv para análise.

```
generate_csv.py > ...
1 import requests
2 from selenium import webdriver
3 from bs4 import BeautifulSoup
4 import json
5 import re
6 import csv
7
8
9 def generate_csv(data: list, headers: list = ['Ano', 'Nome da Música', 'Artista', 'Gênero']):
10     with open('musicas.csv', 'w', encoding='utf-8', newline='') as f:
11         writer = csv.DictWriter(f, fieldnames=headers)
12         writer.writeheader()
13         for row in data:
14             writer.writerow({'Ano': row['Ano'], 'Nome da Música': row['Nome da Música'], 'Artista': row['Artista'], 'Gênero': row.get('Gênero', 'não encontrado')})
15
16 def extract_genre(data: list):
17     genre = data.get('Genre', 'não encontrado').split('\n')
18     genres = [i for i in genre if i != '']
19     cleaned_genres = [re.sub(r'[\d+]', '', genre) for genre in genres]
20     return cleaned_genres
21
22
23 def iterate_on_list(data: list):
24     formatted_data = []
25     for item in data:
26         extracted_gender = extract_genre(item['Wikipedia Info'])
27         print(('Ano: item["Ano"]', 'Nome da Música: item["Nome da Música"]', 'Artista: item["Artista"]', 'Gênero: extracted_gender'))
28         formatted_data.append(('Ano': item['Ano'], 'Nome da Música': item['Nome da Música'], 'Artista': item['Artista'], 'Gênero: extracted_gender'))
29     return formatted_data
30
31
32 def main():
33     with open('musicas.json', 'r', encoding='utf-8') as f:
34         data = json.load(f)
35         data_csv = iterate_on_list(data)
36         generate_csv(data_csv, headers=['Ano', 'Nome da Música', 'Artista', 'Gênero'])
37
38
39 if __name__ == '__main__':
40     main()
41
```

# Análise dos resultados finais

```
import pandas as pd
import re
from collections import Counter
from ast import literal_eval
from pathlib import Path
import matplotlib.pyplot as plt

def read_csv(file_path):
    return pd.read_csv(file_path)

def count_genres_and_artists(df):
    genre_counter = Counter()
    artist_counter = Counter()

    for _, row in df.iterrows():
        artist_counter[row['Artista']] += 1

        genres = literal_eval(row['Gênero'])
        if isinstance(genres, list):
            for sub_genre in re.split(r'[.,]', genre):
                cleaned_genre = sub_genre.strip().lower()
                if cleaned_genre != 'não encontrado':
                    genre_counter[cleaned_genre] += 1

    return genre_counter, artist_counter

def plot_most_common(counter, title, xlabel, ylabel, num_items=10):
    items = counter.most_common(num_items)
    labels, values = zip(*items)

    plt.figure(figsize=(12, 6))
    plt.bar(labels, values, color='skyblue')
    plt.title(title)
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.xticks(rotation=45, ha='right')
    plt.tight_layout()
    plt.savefig(f'{title}.png')
    plt.show()
```

Script *analise\_csv.py* para geração dos gráficos.

```
def main():
    path = Path(__file__).parent / 'musicas.csv'
    df = read_csv(path)

    genre_counter, artist_counter = count_genres_and_artists(df)

    print("Gêneros com mais músicas:")
    for genre, count in genre_counter.most_common():
        print(f"{genre}: {count}")

    print("\nArtistas com mais músicas:")
    for artist, count in artist_counter.most_common():
        print(f"{artist}: {count}")

    # Plot the results
    plot_most_common(genre_counter, 'Gêneros com mais músicas', 'Gêneros', 'Número de prêmios')
    plot_most_common(artist_counter, 'Artistas com mais músicas', 'Artistas', 'Número de prêmios')

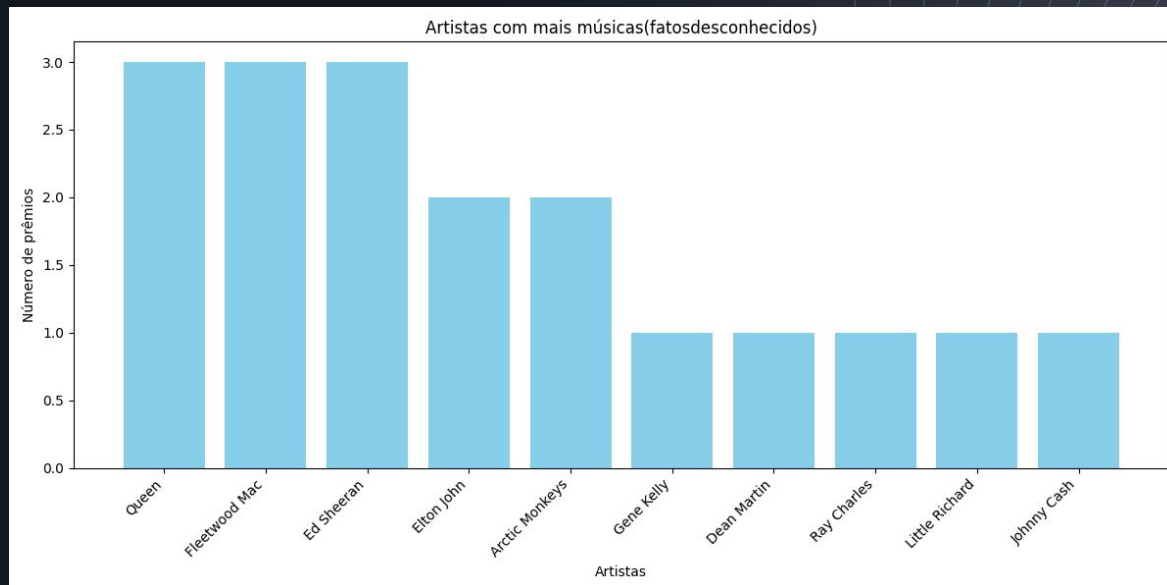
if __name__ == "__main__":
    main()
```

# Análise dos resultados finais

Considerando a lista de fatosdesconhecidos:

Empatados em primeiro lugar, os artistas que foram os mais ouvidos por mais anos foram, com 3 aparições na lista cada:

- Queen
- Fleetwood Mac
- Ed Sheeran

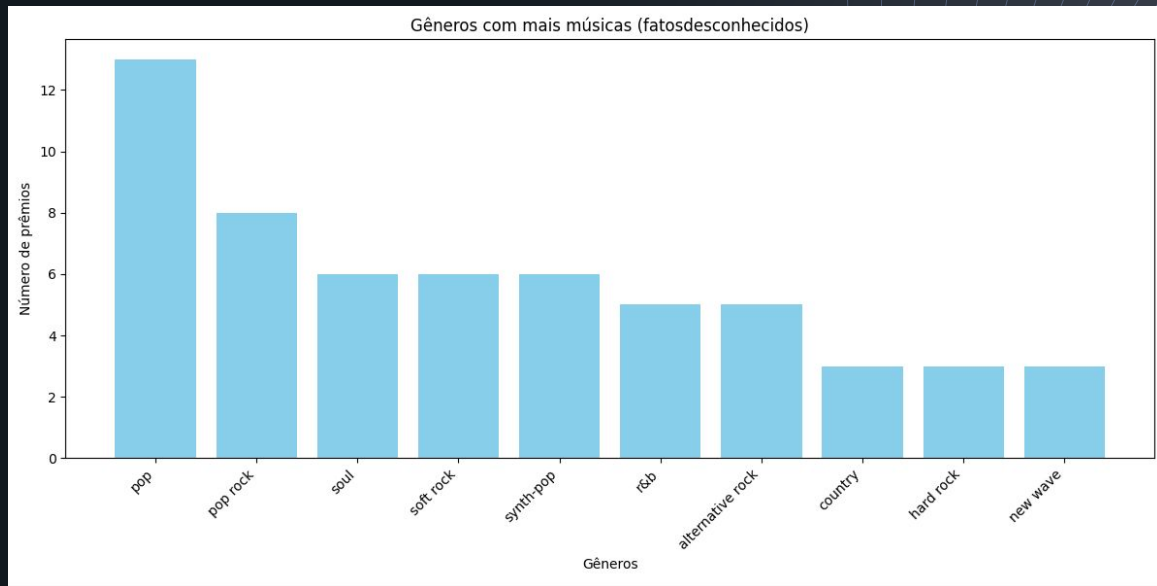


# Análise dos resultados finais

Considerando a lista de fatossdesconhecidos:

Os gêneros que mais vezes apareceram como 'o mais ouvido' por mais anos foram:

1. Pop (12 vezes!)
2. Pop Rock (8)
3. Soul (6)
4. Soft Rock (6)
5. Synth-pop (6)

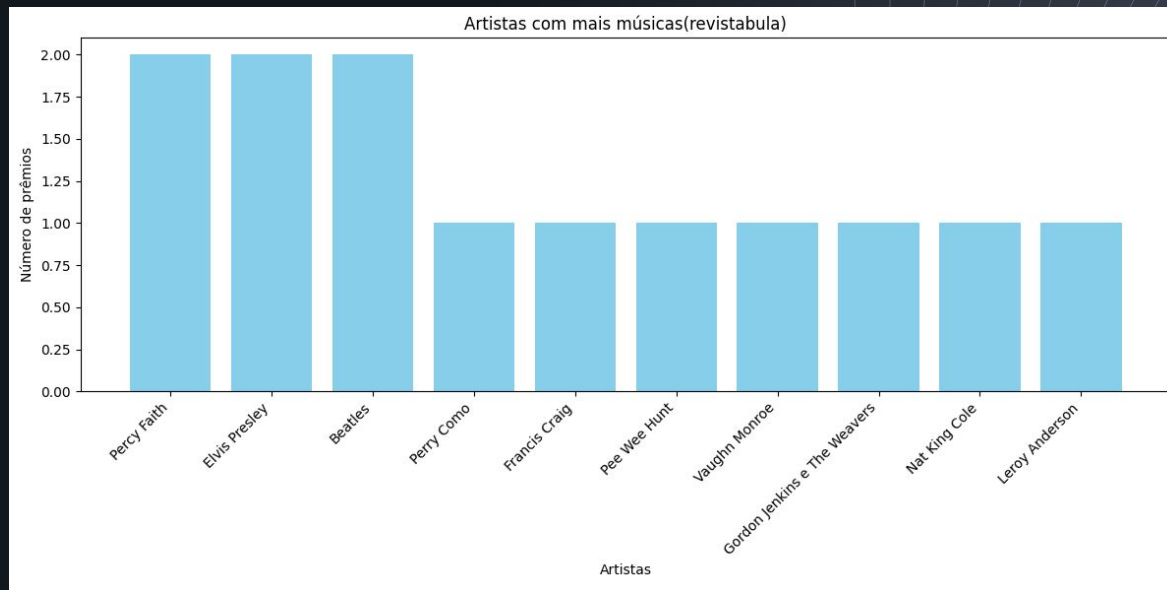


# Análise dos resultados finais

Considerando a lista de revistabula:

Empatados em primeiro lugar, os artistas que foram os mais ouvidos por mais anos foram, com 2 aparições na lista cada:

- Percy Faith
- Elvis Presley
- Beatles

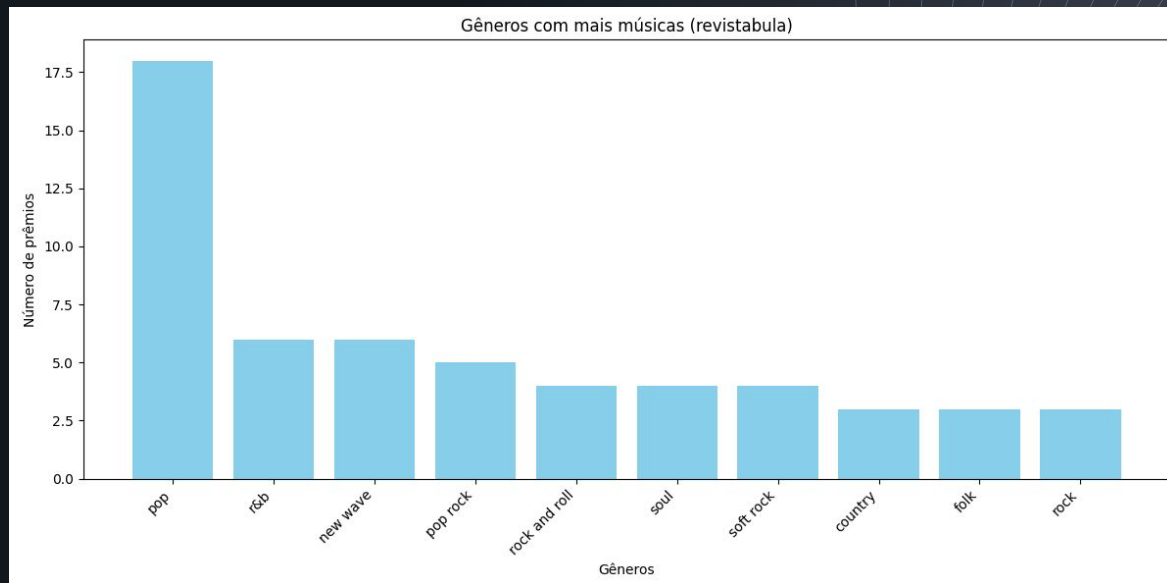


# Análise dos resultados finais

Considerando a lista de revistabula:

Os gêneros que mais vezes apareceram como 'o mais ouvido' por mais anos foram:

1. Pop (18 vezes!)
2. R&B (6)
3. New Wave (6)
4. Pop Rock (5)
5. Rock and Roll (5)





# Casos de falha

- Existiu um erro ao relacionar os dados coletados da entrada com a pesquisa no site da wikipedia, aonde a música pesquisada retornava uma página diferente porém com um nome parecido, portanto retornava uma infobox com informações completamente diferentes da música em questão.
- Porém isso era um erro da api da wikipedia que estava preparada para pesquisar no servidor norte americano ao invés de pesquisar no servidor brasileiro.
- Para a análise dos dados, percebemos que alguns gêneros de músicas eram retornados com nomes 'aglomerados'.

# Conclusão

Concluimos que as informações de entrada podem vir de diversas fontes, onde podemos realizar um trânsito de informações pelo nosso sistema controlando quais as informações são relevantes para nós e as separando em um arquivo ou enviando para algum outro sistema em tempo real.

Obrigado!