

# MVP - Sprint: Engenharia de Dados

Cássia Francine Novello  
Setembro de 2023

## Objetivo

O objetivo do trabalho é realizar uma análise sobre as músicas mais famosas do Spotify no ano de 2023. A análise é interessante tanto para fãs de música quanto para profissionais da indústria que se interessam por uma visão mais detalhada do cenário atual. Algumas perguntas que podem ser levantadas sobre esses dados são:

- Músicas de anos anteriores foram marcantes neste período? Ou apenas lançamentos recentes?
- Qual é o mínimo e máximo de bpm (batidas por minuto) que está na lista? Qual a média?
- Os tons maiores e menores estão presentes proporcionalmente?
- Alguma música se destaca, com muito mais streams que as outras?

## Coleta e Modelagem

Kaggle: <https://www.kaggle.com/datasets/nelgiriyeewithana/top-spotify-songs-2023>

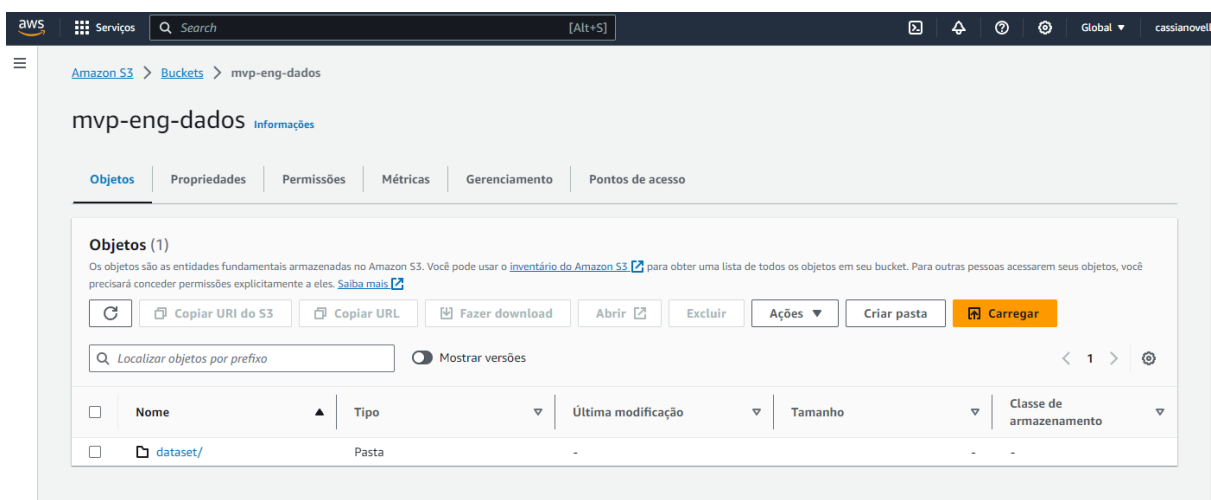
O dataset uma lista de músicas famosas no Spotify, durante o ano de 2023 até setembro. Abaixo segue a descrição de cada campo.




- **track\_name**: nome da canção
- **artist(s)\_name**: nome dos artistas
- **artist\_count**: número de artistas que contribuíram com a canção
- **released\_year**: ano de lançamento da canção (deve ser um ano válido)
- **released\_month**: mês de lançamento da canção (deve ser um mês válido)
- **released\_day**: dia de lançamento da canção (deve ser um dia de mês válido)
- **in\_spotify\_playlists**: Number of Spotify playlists the song is included in
  - número inteiro positivo
- **in\_spotify\_charts**: Presence and rank of the song on Spotify charts
  - número inteiro positivo
- **streams**: Total number of streams on Spotify
  - número inteiro positivo
- **in\_apple\_playlists**: Number of Apple Music playlists the song is included in
  - número inteiro positivo
- **in\_apple\_charts**: Presence and rank of the song on Apple Music charts
  - número inteiro positivo
- **in\_deezer\_playlists**: Number of Deezer playlists the song is included in
  - número inteiro positivo

- **in\_deezer\_charts**: *Presence and rank of the song on Deezer charts*
  - número inteiro positivo
- **in\_shazam\_charts**: *Presence and rank of the song on Shazam charts*
  - número inteiro positivo
- **bpm**: *Beats per minute, a measure of song tempo*
  - número inteiro positivo
- **key**: *Key of the song*
  - valor literal de acordo com a nomenclatura de cifras
- **mode**: *Mode of the song*
  - (major or minor) - Categórico
- **danceability\_%**: *Percentage indicating how suitable the song is for dancing*
  - (0-100)
- **valence\_%**: *Positivity of the song's musical content*
  - (0-100)
- **energy\_%**: *Perceived energy level of the song*
  - (0-100)
- **acousticness\_%**: *Amount of acoustic sound in the song*
  - (0-100)
- **instrumentalness\_%**: *Amount of instrumental content in the song*
  - (0-100)
- **liveness\_%**: *Presence of live performance elements*
  - (0-100)
- **speechiness\_%**: *Amount of spoken words in the song*
  - (0-100)

## Armazenamento

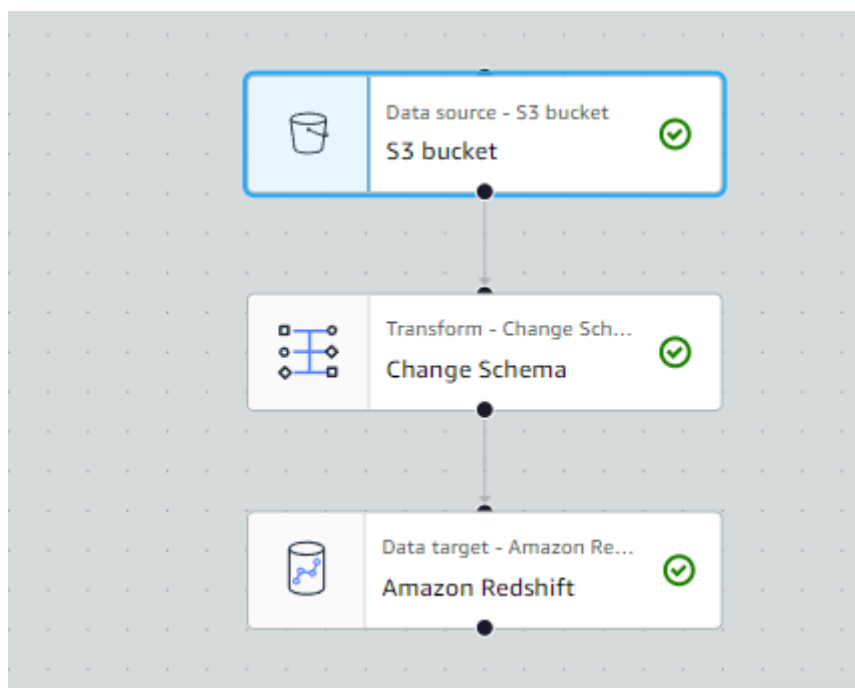
O arquivo csv foi armazenado no S3 da AWS, para ser importado em um banco de dados. Foi necessário criar um endpoint do tipo Gateway para permitir acesso ao arquivo pelo Glue.



vpce-03df587a842d5c620 / mvp-endpoint-s3			
<a href="#">Detalhes</a>	<a href="#">Tabelas de rotas</a>	<a href="#">Política</a>	<a href="#">Tags</a>
<b>Detalhes</b>			
ID do endpoint  vpce-03df587a842d5c620	Status  Disponível	Hora de criação sábado, 23 de setembro de 2023 às 13:57:10 BRT	Tipo de endpoint Gateway
ID da VPC <a href="#">vpc-0badbaacc34dd0431</a>	Mensagem de status -	Nome do serviço  com.amazonaws.us-east-1.s3	Nomes DNS privados habilitados Não

## Transformações

O AWS Glue foi utilizado para a criação de um job, responsável por transformar os dados armazenados no S3 no arquivo csv em registros de uma tabela num banco de dados RedShift.



Data source properties - S3

Output schema

Data preview

Name

S3 bucket

S3 source type

Info

☒ S3 location

Choose a file or folder in an S3 bucket.

☐ Data Catalog table

S3 URL

Q

s3://mvp-eng-dados/dataset/new\_dataset\_spo

X

View

☒ Recursive

Read files in all subdirectories.

Data format

CSV

O arquivo csv original baixado do Kaggle não pode ser importado devido a erros na execução do job. Alguns dos erros apresentados pela execução do job no Glue:

Ao analisar detalhadamente o log e investigar fóruns, foi identificado no arquivo csv original um problema com caracteres inválidos para a formação UTF-8. Foi necessário um pequeno código em python para eliminar os caracteres inválidos e permitir a carga para o Redshift.

Para mais detalhes, acesse o código em:

<https://colab.research.google.com/drive/1Pa5p5ov-NuMaz6eDUSnKaiOR3b8h-CRA?usp=sharing>

```
#open text file in read mode
text_file = open("spotify-2023.csv", "r", encoding='latin-1')

#read whole file to a string
data = text_file.read()

#close file
text_file.close()

[ ] my_bytes = data.encode('utf-8')

result = my_bytes.decode(
    'utf-8', errors='ignore'
).encode('utf-8')

print(result)

b'track_name,artist(s)_name,artist_count,released_year,released_month,r

[ ] file = open("new_dataset_spotify.csv", "w")
a = file.write(str(result, encoding='utf-8'))
file.close()
```

Outra transformação necessária foi a retirada dos caracteres por cento e parenteses dos nomes das colunas, que estavam causando erros. Além disso, alguns campos foram tipados como inteiro para melhor adequar-se ao conteúdo.

mode	mode	string	<input type="checkbox"/>
danceability_%	danceability	int	<input type="checkbox"/>
valence_%	valence	int	<input type="checkbox"/>
energy_%	energy	int	<input type="checkbox"/>
acousticness_%	acousticness	int	<input type="checkbox"/>
instrumentalness_%	instrumentalness	int	<input type="checkbox"/>
liveness_%	liveness	int	<input type="checkbox"/>
speechiness_%	speechiness	int	<input type="checkbox"/>

### Carga

Então, finalmente foi possível configurar o RedShift e realizar a carga dos dados. Foi criada a tabela “spotify” para receber os dados do arquivo csv. Além disso, foi criada uma conexão RedShift e um IAM Role para atribuir as permissões.

```
create table public.spotify (
```



**IAM** > Funções

**Funções (5)** [Informações](#)

Uma função do IAM é uma identidade que você pode criar que tem permissões específicas com credenciais válidas por curtos períodos. Funções podem ser assumidas por entidades em que você confia.

< 1 >

<input type="checkbox"/>	Nome da função ▲	Entidades confiáveis	Última atividade ▼
<input type="checkbox"/>	<a href="#">AmazonRedshift-CommandsAccessRole-20230921T172722</a>	Serviço da AWS: sagemaker, <a href="#">e mais 2</a>	-
<input type="checkbox"/>	<a href="#">AWSServiceRoleForRedshift</a>	Serviço da AWS: redshift (Função vinculada)	23 minutos atrás
<input type="checkbox"/>	<a href="#">AWSServiceRoleForSupport</a>	Serviço da AWS: support (Função vinculada)	-
<input type="checkbox"/>	<a href="#">AWSServiceRoleForTrustedAdvisor</a>	Serviço da AWS: trustedadvisor (Função vinculada)	-
<input type="checkbox"/>	<a href="#">mvp-iam-role-glue</a>	Serviço da AWS: glue	2 dias atrás

Após a execução do Job no Glue, os dados ficaram disponíveis para consulta no RedShift, permitindo a análise.

## Análise

### Qualidade dos dados

- Existem caracteres inválidos da formatação UTF-8 no dataset original. Isso pode gerar problemas para carga dos dados dependendo da plataforma utilizada. Ao remover os caracteres inválidos, alguns campos podem ficar ilegíveis ou incorretos Ex: "Frīĳ½ĳ½gil".
- Há 953 registros no total, uma quantidade suficiente para a análise.
- Não existem registros com campo "released\_date" nulo. Há uma única canção com lançamento em 1930 que parece ser um erro de registro. Depois deste ano, o próximo é de 1942.
- Não existem registros com campo "bpm" nulo.
- Não existem registros com campo "mode" nulo, nem diferente das categorias "Minor" e "Major".
- Há 26 registros com o campo "streams" nulo. Estes registros não serão considerados para a comparação do número de streams. Há um registro com número muito inferior aos demais (2762, o próximo registro é 1365184), que também será desconsiderado.

### Solução do problema

- **Músicas de anos anteriores foram marcantes neste período? Ou apenas lançamentos recentes?**

696 das 953 canções na lista foram lançadas no ano ou após 2021, mostrando que os lançamentos mais recentes tiveram um grande impacto na lista das canções mais ouvidas.

```
select released_year, count(*) from public.spotify
where released_year >=2021
group by released_year
order by released_year
```

Result 1 (3)		
<input type="checkbox"/>	released_year	count
<input type="checkbox"/>	2021	119
<input type="checkbox"/>	2022	402
<input type="checkbox"/>	2023	175

```
select released_year, count(*) from public.spotify
order by released_year
```



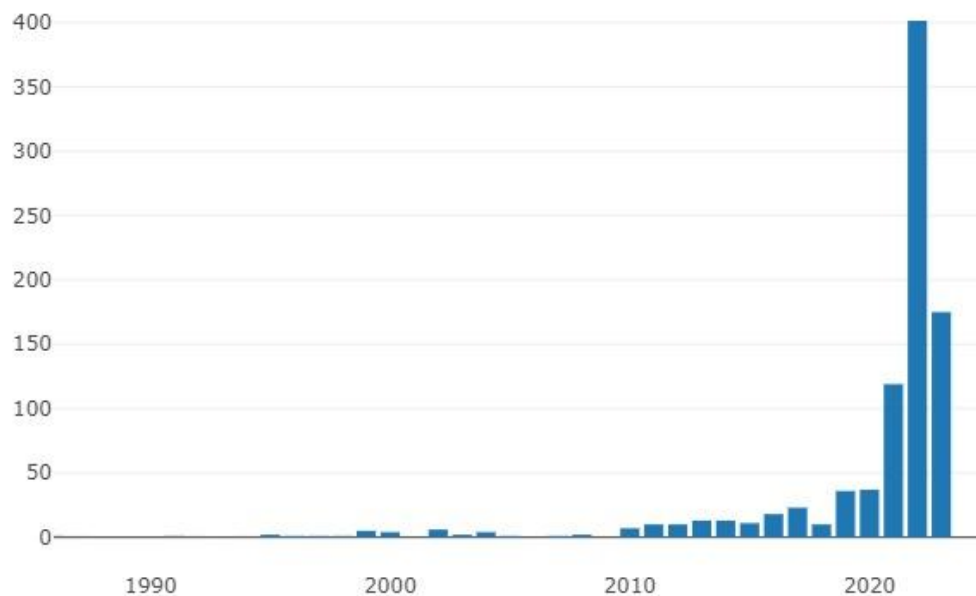


Gráfico de frequência por ano, gerado com o RedShift, usando a função Chart / Bar.

Entre os lançamentos de 1942 até 1973, a maioria são músicas sobre o Natal como Jingle Bell Rock, de Bobby Helms, e Rockin around the Christmas Tree, de Brenda Lee. As mais antigas na lista não relacionadas ao Natal são os hit “Have you ever seen the rain”, de Creedance, do ano de 1968, e “Dream On”, de Aerosmith, de 1973.

- Qual é o mínimo e máximo de bpm (batidas por minuto) que está na lista? Qual a média?

```
select max(bpm), min(bpm), avg(bpm) from public.spotify
```

<input type="checkbox"/>	max	min	avg
<input type="checkbox"/>	206	65	122

```
select bpm, count(*) from public.spotify
group by bpm
order by bpm
```

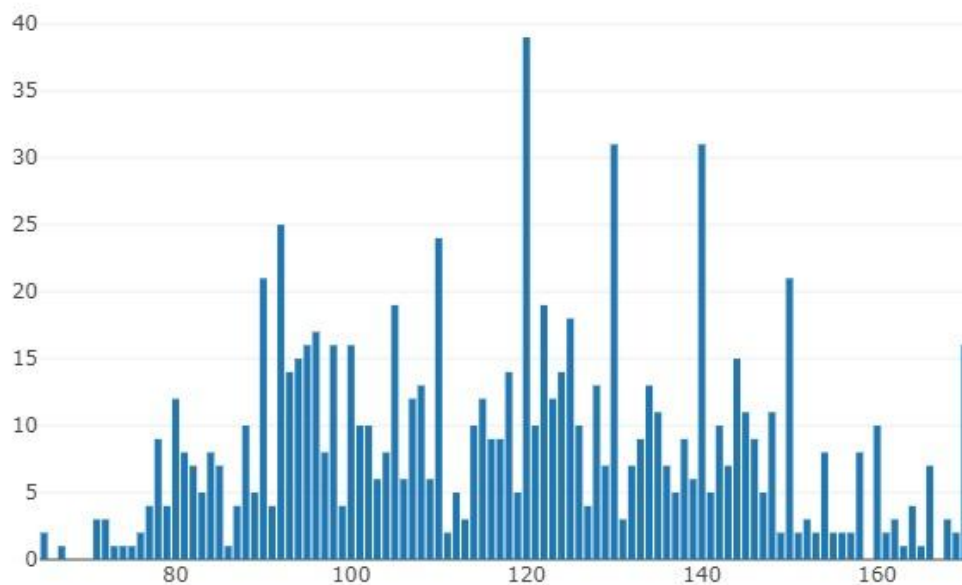


Gráfico de frequência por bpm gerado com o RedShift, usando a função Chart / Bar.

A Sociedade Artística Brasileira considera o andamento lento para bpm até aproximadamente 63, médio até 120 bpm e rápido até 208 bpm. É possível observar na lista uma maior incidência dos andamentos médios e rápidos, sendo a maior incidência (122) na fronteira entre a categoria média e rápida.

- Os tons maiores e menores estão presentes proporcionalmente?

```
select mode, count(*), cast(count(*) as decimal)/953 as "Perc" from
public.spotify
group by mode
```

<input type="checkbox"/>	mode	count	perc
<input type="checkbox"/>	Major	550	0.5771
<input type="checkbox"/>	Minor	403	0.4228

É possível observar uma maior ocorrência de músicas em tom maior, com diferença de aproximadamente 15%.

- **Alguma música se destaca, com muito mais streams que as outras?**

```
select track_name, artists_name, streams from public.spotify
where streams is not null
and streams > 100000
order by streams desc
```

É possível observar que a diferença entre o primeiro e segundo lugar da lista é de aproximadamente 2 milhões de streams, o que não é um valor muito relevante comparado ao total, que encontra-se por volta de 2 bilhões. Entretanto, se considerarmos o tempo médio de uma canção de 3 minutos, 2 milhões de streams equivalem a 11 anos!

<input type="checkbox"/>	track_name	artists_name	streams
<input type="checkbox"/>	Take Me To Church	Hozier	2135158446
<input type="checkbox"/>	Circles	Post Malone	2132335812
<input type="checkbox"/>	Love Yourself	Justin Bieber	2123309722
<input type="checkbox"/>	All of Me	John Legend	2086124197
<input type="checkbox"/>	Counting Stars	OneRepublic	2011464183
<input type="checkbox"/>	Riptide	Vance Joy	2009094673
<input type="checkbox"/>	Wake Me Up - Radio Edit	Avicii	1970673297
<input type="checkbox"/>	Can't Hold Us (feat. Ray ...	Ray Dalton, Ryan Lewis, ...	1953533826
<input type="checkbox"/>	The Hills	The Weeknd	1947371785
<input type="checkbox"/>	HUMBLE.	Kendrick Lamar	1929770265
<input type="checkbox"/>	One Kiss (with Dua Lipa)	Calvin Harris, Dua Lipa	1897517891
<input type="checkbox"/>	good 4 u	Olivia Rodrigo	1887039593
<input type="checkbox"/>	drivers license	Olivia Rodrigo	1858144199

## Conclusão e Autoavaliação

Apesar do dataset mostrar-se com nota máxima no kaggle, a codificação fora do padrão UTF-8 gerou diversos desafios ao usar o Glue. Os erros da plataforma AWS não foram diretos ao reportar este problema, sendo necessário uma pesquisa nos logs e fóruns da internet. Após a superação do problema da codificação, o dataset mostrou-se com uma boa qualidade, poucos valores nulos e inválidos, rendendo uma análise interessante.

Outras análises que poderiam ser realizadas no dataset poderiam utilizar os campos relativos a avaliação da canção como **danceability**, **valence**, **energy**, etc, para avaliar as tendências sonoras da época de 2023, que seriam bastante úteis para artistas e profissionais da indústria da música.