

Relatório

icmc
júnior

Análise de dados de músicas populares em 2023 no Spotify

Sumário

1	Objetivos	3
2	Desenvolvimento	3
2.1	Bases de Dados	3
2.2	Tratamento de Dados	3
2.2.1	Most Streamed Spotify Songs 2023	3
2.2.2	Spotify daily top 200 (2017-2021)	4
2.2.3	API Lyrics.ovh, IA Gemini - Pro e Google Translator	4
2.3	Apresentação de Gráficos e Métodos	4
2.3.1	Gráfico de dispersão	4
2.3.2	Histogramas	4
2.3.3	Gráfico de Barras e Gráfico de Colunas	5
2.3.4	Gráfico de Linha	5
2.3.5	Word Cloud	5
2.3.6	Análise de Sentimentos com VADER	5
2.3.7	Testes de Normalidade	5
2.3.8	Matriz de correlação de Spearman	5
2.3.9	Análise Temporal	6
2.3.10	Regressão de Quadrados Mínimos Ordinários	6
2.3.11	Análise de Componente Principais	6
2.3.12	Probabilidade	6
3	Resultados	6
3.1	Análise exploratória dos dados	6
3.1.1	Gráficos de Dispersão	6
3.1.2	Histogramas	8
3.2	Word Cloud	9
3.3	Análise de sentimentos	10
3.4	Análise de correlação de Spearman	11
3.5	Análise temporal	12
3.5.1	Streams por data de lançamento	12
3.5.2	Streams por mês de lançamento	13
3.5.3	Streams por dia de lançamento	14
3.5.4	5 artistas mais ouvidos	15
3.5.5	Músicas mais ouvidas por década	15
3.6	Análise de Regressão de Quadrados Mínimos Ordinários	15
3.7	Análise de Componentes Principais	17
3.8	Análise de Probabilidade	18
4	Conclusão	19

1 Objetivos

O projeto tem por objetivo analisar a base de dados Most Streamed Spotify Songs 2023, de modo a compreender a relação entre aspectos das músicas e suas popularidades. Dentre os objetivos específicos, algumas questões foram levantadas para estudo e análise empírica:

- Quais os padrões e características em comum das músicas mais populares no Spotify em 2023?
- Quais músicas foram mais populares?
- Características individuais da música, por exemplo energia, valência , possuem correlação direta com a popularidade da música? E se analisadas em conjunto, possuem relação com o número de streams?
- Músicas mais recentes tendem a serem mais populares no ano? Se sim, quando as músicas tendem a perder popularidade?
- Existe algum período do ano e/ou de cada mês que as músicas mais populares foram lançadas?
- Quais palavras mais apareceram nas músicas com mais streams em 2023?
- O quanto a popularidade do artista influencia na popularidade da música?
- Existe algum sentimento transmitido pelas canções com mais streams que seja mais recorrente?

2 Desenvolvimento

O projeto foi desenvolvido sob manipulação da linguagem de programação Python e bibliotecas voltadas a ciência de dados. Os algoritmos, o código-fonte e as análises estão disponíveis no repositório on-line do Github: [Análise de dados de músicas populares em 2023 no Spotify](#).

O desenvolvimento prático contou com as etapas desenvolvidas abaixo: coleta e seleção das bases de dados, tratamento de dados, análise exploratória, geração de gráficos, correlações e probabilidades, por fim a obtenção e análise dos resultados.

2.1 Bases de Dados

- **Most Streamed Spotify Songs 2023:** contém uma lista das músicas mais tocadas pelo Spotify em 2023, com dados gerais de nome da música e dos artistas, data de lançamento, dados sobre influência no Spotify, Deezer e Apple Music, além de estatísticas sobre o áudio, como porcentagem de energia e instrumentalismo da música.
- **Spotify daily top 200 (2017-2021):** contempla as 200 músicas mais tocadas do Spotify de 2017 a 2021 por dia, com gêneros da música incluso.
- **API Lyrics.ovh e IA Gemini - Pro:** foi usada uma API (Application Programming Interface) e uma Inteligência Artificial (IA) para a coleta das letras das músicas. Os resultados foram inseridos em colunas novas no próprio conjunto de dados principal (Most Streamed Spotify Songs 2023).

2.2 Tratamento de Dados

2.2.1 Most Streamed Spotify Songs 2023

- Retirada de uma linha que possuía valor de erro na coluna "streams";
- Correção de valores decimais para inteiros para as colunas: "in_deezer_playlists" e "in_shazam_charts";

- Troca de valores categóricos na coluna "key" para sua respectiva frequência em Hertz. Para esse caso , foi utilizada a quarta oitava, por ser a oitava central, da tabela de frequências do piano.
- Troca dos valores categóricos da coluna "mode" para valores 0 ou 1, utilizando a técnica One-hot Encoding.

2.2.2 Spotify daily top 200 (2017-2021)

A partir dessa base de dados, criou-se a coluna "popularidade_artista" no conjunto de dados Most Streamed Spotify Songs 2023, contendo quantos dias cada artista produtor da música popular teve algum som, entre 2017 e 2021, no top 200 do Spotify. O objetivo dessa interação entre base de dados diferentes é analisar o quanto um artista ser famoso previamente impacta na popularidade da música nos próximos anos.

2.2.3 API Lyrics.ovh, IA Gemini - Pro e Google Translator

Dado o nome da música e de seus artistas, era necessário pesquisar a letra de 952 músicas de forma automática. A API Lyrics.ovh permitia requisições automáticas por busca, e foi possível extrair 589 letras, com 38% de dados faltantes, uma alta quantidade para a proposta. Para complementar a coleta, utilizou-se a IA generativa Gemini - Pro para gerar, automaticamente, as novas letras, com apenas 15 dados faltantes (1.6%).

Uma vez com o objetivo de fazer uma nuvem de palavras mais tocadas nas músicas famosas, era necessário considerar "amor" e "love" a mesma palavra, portanto traduzir as músicas ao inglês. As letras foram traduzidas pela API do Google Translator, que apresentou algumas falhas no processamento de algumas músicas, mas foi capaz de traduzir a maioria.

Dada as diferentes fontes de coleta e tradução foram criadas 4 colunas no conjunto de dados principal "Most Streamed Spotify Songs 2023":

- letter: letras colhidas pela API Lyrics.ovh;
- translated_letter: letras da coluna"letter" traduzidas;
- translated_lyrics_IA: letras colhidas pela IA generativa (que traduziu a maioria das letras);
- final_letter: letras da "translated_lyrics_IA" traduzidas novamente pela API do Google Translator.

2.3 Apresentação de Gráficos e Métodos

2.3.1 Gráfico de dispersão

Os diagramas de dispersão ou gráficos de dispersão são representações de dados de duas (tipicamente) ou mais variáveis que são organizadas em um gráfico. O gráfico de dispersão utiliza coordenadas cartesianas para exibir valores de um conjunto de dados. Os dados são exibidos como uma coleção de pontos, cada um com o valor de uma variável determinando a posição no eixo horizontal e o valor da outra variável determinando a posição no eixo vertical (em caso de duas variáveis).

2.3.2 Histogramas

Usado normalmente para observar a distribuição dos dados, organiza-os em classes de tal forma que o eixo horizontal (x) representa os intervalos dos dados e o eixo vertical (y) a frequência ou densidade, ou seja, o número de ocorrências de cada classe.

2.3.3 Gráfico de Barras e Gráfico de Colunas

O gráfico de barras é um gráfico com barras retangulares horizontais e comprimento proporcional aos valores que ele apresenta. Um eixo do gráfico mostra especificamente o que está sendo comparado enquanto o outro eixo representa valores discretos. O gráfico de colunas possui a mesma lógica, a diferença é que as barras retangulares são verticais.

2.3.4 Gráfico de Linha

Os gráficos de linhas mostram as informações como uma série de pontos de dados que estão conectadas por segmentos de linha reta. As categorias são mostradas ao longo do eixo horizontal (x) e as estatísticas são mostradas ao longo do eixo vertical (y).

2.3.5 Word Cloud

Também chamado de nuvem de palavras, o Word Cloud é uma representação visual de palavras com tamanhos proporcionais a sua frequência no conjunto de dados. Nessa aplicação, quanto maior a palavra, mais ocorrências ela teve em músicas populares do Spotify.

2.3.6 Análise de Sentimentos com VADER

O VADER (Valence Aware Dictionary and sEntiment Reasoner) é uma ferramenta de análise de sentimentos que lida com textos com linguagem natural, por exemplo letras de música. O algoritmo do VADER quantifica a emoção de um texto, com uma pontuação de -1 a +1, em que -1 representa sentimentos negativos, 0 neutros e +1 positivos. Em seu funcionamento, usa um dicionário de valência: cada palavra tem uma "pontuação" que reflete um sentimento. Por exemplo, "bom" pode valer +2, enquanto "ruim" -2. Fora isso, o VADER interpreta a frase com modificadores, como "muito", que intensifica sentimentos ou "pouco", que suaviza os sentimentos, bem como "não", que inverte sentimentos.

Os resultados do algoritmo são os seguintes:

- neg: pontuação de sentimento negativo (0 a 1);
- neu: pontuação de sentimento neutro (0 a 1);
- pos: pontuação de sentimento positivo (0 a 1);
- compound: calculada a partir das 3 pontuações, resume o sentimento (-1 a 1).

2.3.7 Testes de Normalidade

A curva de distribuição normal ou gaussiana é uma curva simétrica em torno de seu ponto médio, similar a um formato de sino. Vários testes estatísticos, como alguns testes de correlação de duas variáveis, assumem a distribuição normal ou não normal dos dados, e por isso é necessário validá-la. Para isso, fora a visualização a partir dos histogramas (percebida a partir da linha vermelha), foi utilizado o teste de Shapiro-Wilk. Nesse teste estatístico, ideal para amostras pequenas de 50 a 2000 dados, é possível a partir de um teste de hipóteses verificar a normalidade de cada distribuição de variável.

2.3.8 Matriz de correlação de Spearman

Uma matriz de correlação é um gráfico cujos elementos no seu interior representam o índice de correlação entre da variável na linha do elemento e a variável da coluna do elemento. No índice de correlação utilizado, chamado de Spearman, a interpretação é dada por:

- Próximo a 1: correlação positiva, ambas variáveis crescem de forma diretamente proporcional (crescem juntas).

- Próximo a -1: correlação negativa, ambas variáveis crescem de forma inversamente proporcional (enquanto uma cresce, a outra decresce).
- Próximo a 0: correlação muito baixa, significa que as variáveis não tem relação entre si. Uma ser alta ou baixa não impacta diretamente na outra.

2.3.9 Análise Temporal

A análise temporal é um método que consiste em avaliar o comportamento de uma variável ao longo do tempo (aumento, diminuição ou estabilidade). Para possibilitar essa visualização, foram utilizados gráficos de linha, gráficos de dispersão, gráficos de colunas, gráficos de barras e tabela simples.

2.3.10 Regressão de Quadrados Mínimos Ordinários

A Regressão de Quadrados Mínimos Ordinários (OLS, do inglês Ordinary Least Squares) é uma estratégia de otimização que ajuda a encontrar uma linha reta o mais próxima possível dos seus pontos de dados em um modelo de regressão linear. O OLS é considerado a estratégia de otimização mais útil para modelos de regressão linear, pois pode ajudar a encontrar estimativas não tendenciosas para os valores reais dos seus parâmetros alfa e beta.

2.3.11 Análise de Componente Principais

A Análise de Componentes Principais ou PCA (Principal Component Analysis) é uma técnica de análise multivariada que pode ser usada para analisar inter-relações entre um grande número de variáveis e explicar essas variáveis em termos de suas dimensões inerentes (componentes). O objetivo é encontrar um meio de condensar a informação contida em várias variáveis originais em um conjunto menor de variáveis estatísticas (componentes) com uma perda mínima de informação.

2.3.12 Probabilidade

A probabilidade é uma parte da Estatística que estuda a possibilidade de ocorrência de ou mais possíveis resultados dentro de um experimento aleatório. Dentro da probabilidade, a probabilidade condicional analisa a possibilidade de ocorrer um evento dado um outro acontecimento anterior.

3 Resultados

3.1 Análise exploratória dos dados

Serão apresentadas algumas técnicas para compreender o comportamento dos dados e a distribuição deles, com o objetivo de identificar padrões, testar hipóteses e verificar possíveis anomalias.

3.1.1 Gráficos de Dispersão

Dada a necessidade de perceber a correlação de duas variáveis, os gráficos de dispersão mostram o comportamento dos dados. A partir da análise da Figura 1, algumas variáveis não apresentaram correlação alta direta com o número de streams, como o número de artistas (artist_count), rank da música nas plataformas Deezer, Shazam, Spotify e Apple Music, nomeadas por "in_(nome da plataforma)_charts". Também apresentaram baixa correlação visual as variáveis relativas às características das músicas, como dançabilidade, valência, entre outros, variáveis com "%" na terceira e quarta linha de gráficos.

Entretanto, algumas variáveis apresentaram correlação positiva, como a presença das músicas em playlists da Apple Music, da Deezer, mas principalmente do Spotify, enquanto a correlação da plataforma Shazam aparentou ser mais baixa. Essas variáveis estão visíveis nos gráficos de dispersão pelos nomes "in_(nome da plataforma)_playlists".

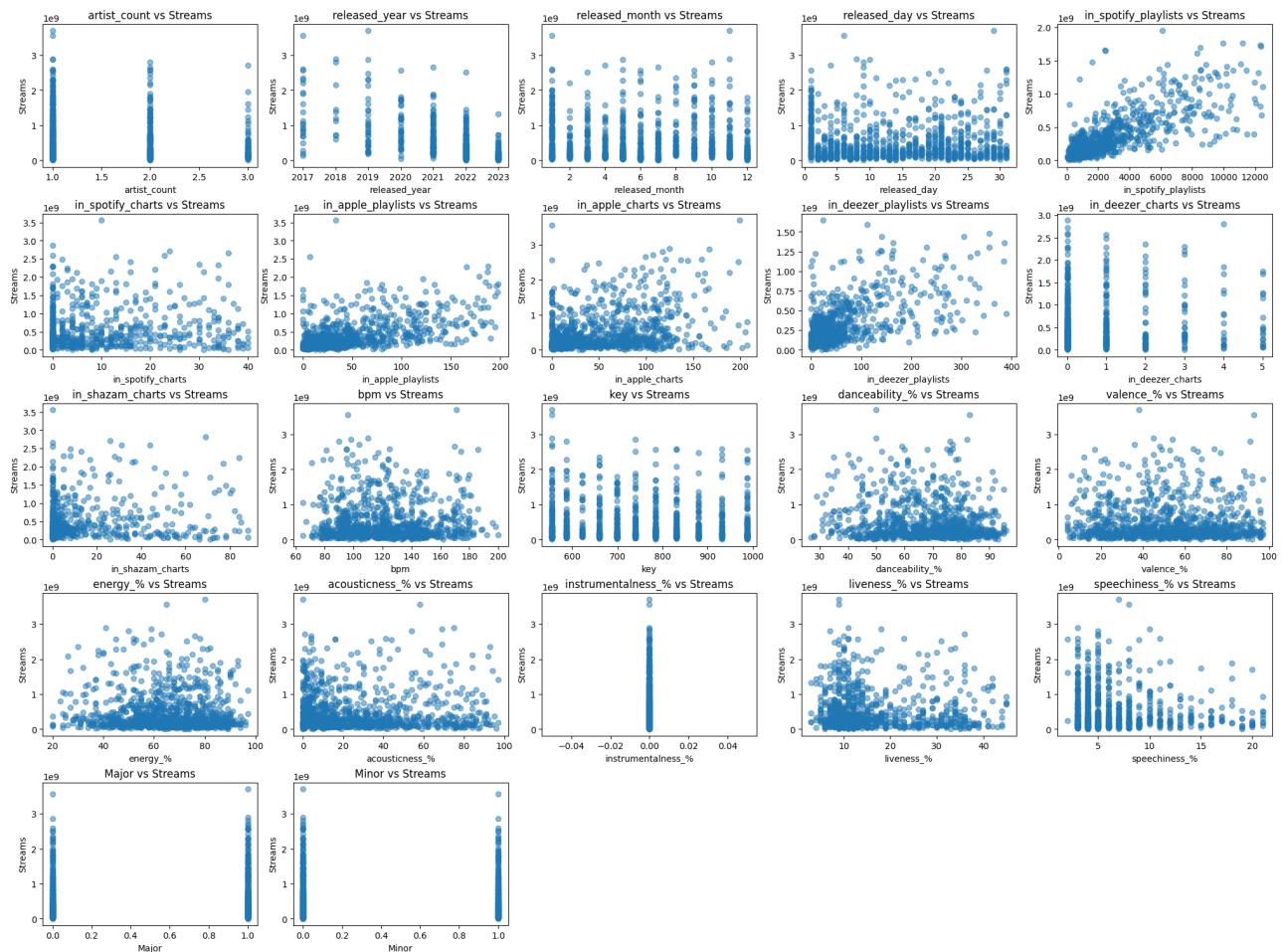


Figura 1: Gráficos de dispersão de número de streams x variáveis dos dados.

3.1.2 Histogramas

Para cada variável do conjunto de dados foi criado um histograma, para identificação de alguns padrões. Percebe-se, a partir da Figura 2, que a maioria das músicas mais populares do Spotify em 2023 possuem apenas 1 artista, e a proporção cai mais que a metade para 2 , 3 e 4 artistas (artist_count).

Além disso, a "energia" das músicas tem uma distribuição com destaque à faixa entre 60 e 80%, que em geral apresenta músicas com uma alta energia. Ainda que com menor destaque, essa mesma faixa se destaca em "danceability", propensão de uma música a ser dançada.

A análise de "bpm", batidas por minuto, nos fornece uma distribuição relativamente equilibrada entre músicas agitadas e mais calmas, com um pouco de tendência para as mais movimentadas, haja vista que na faixa de 90 e 120 são músicas moderadas, e tem um pico de frequência, bem como entre 130 e 150, considerado BPM alto.

A distribuição de "valênci" refere-se medida do tom emocional da música, mais próximo de 0% uma música negativa , mais próxima do 100% uma música positiva. Nota-se uma distribuição de dados que tende levemente a sentimentos positivos, mas não nota-se uma distorção muito desproporcional. Esse resultado será reforçado na seção 3.3.

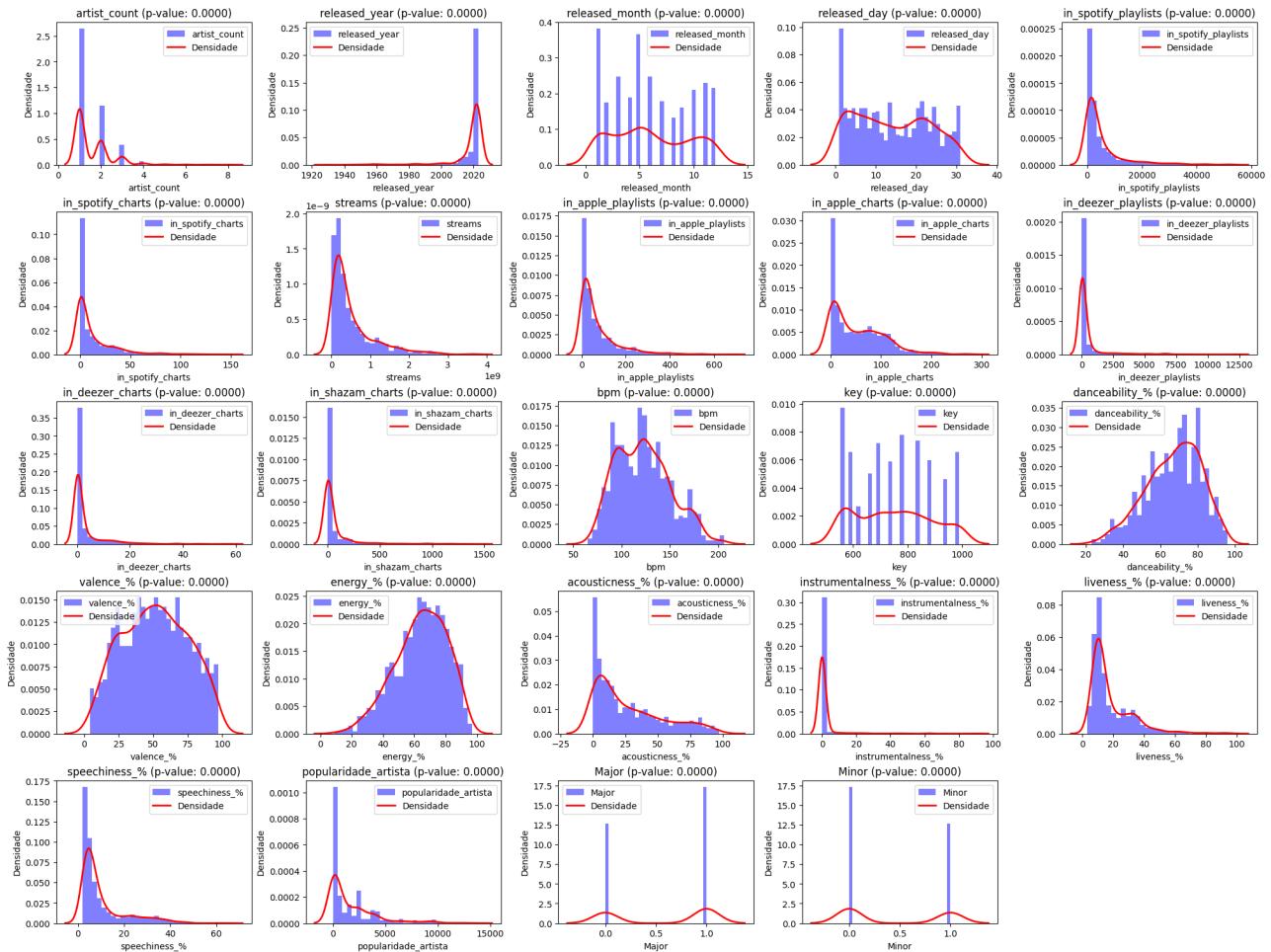


Figura 2: Histogramas de número de streams x variáveis dos dados.

Entretanto, para "acousticness", ou seja, músicas com menos distorções eletrônicas , mais acústicas, mantiveram-se a maioria em uma faixa de 0 a 25%, muitas com 0, demonstrando que músicas mais acústicas não tendem a ter tanta popularidade quanto outras presentes nas mais populares do Spotify. A mesma lógica para "instrumentalness", que se refere a músicas instrumentais, sem vocal humano, a grande maioria das músicas populares da base de dados possui 0%, o que evidencia mais um padrão observado.

Além disso, a maioria das músicas possui "liveness" (probabilidade da música ser gravada ao vivo) entre 0 e

20%, com alguns dados entre 20 e 40% e pouquíssimos maiores que 40%. Ou seja, demonstra um padrão das músicas mais populares do Spotify terem tendência a não serem gravadas ao vivo.

Em complemento, a maioria dos dados de "speechiness", que se refere à probabilidade da música conter fala humana ou grande quantidade de fala, como rap, discurso ou podcasts distribui-se na faixa de 0 a 20%, um padrão de baixa "speechiness" para as músicas populares.

Além disso, note que nenhuma distribuição se aproxima da distribuição normal, o que foi comprovado pelos testes estatísticos de normalidade. Essa informação será utilizada na seção 3.4.

3.2 Word Cloud



Figura 3: Letras da API [Lyrics.ovh](#).



Figura 4: Letras da API Lyrics.ovh para 10% de músicas mais ouvidas.



Figura 5: Letras de IA generativa.



Figura 6: Letras de IA generativa para 10% de músicas mais ouvidas.

Figura 7: Comparação de Word Cloud para diferentes frações do conjunto de dados e para diferentes tipos de coleta.

De acordo com as nuvens de palavras apresentadas acima, note que 3 palavras são sempre destacadas em uma proporção grande: "love" (amor) , "know" (saber) e "like" (gostar). Sob análise mais específica, os Word Cloud das Figuras 3 e 5 , testados com 2 tipos de coletas diferentes de dados, destacaram mais ainda as palavras "love" e "know", com ocorrência evidente também das palavras "want" (querer), "make"(fazer) e got/get(pegou/pegar).

Quando analisadas apenas os 10% de músicas mais ouvidas, a coleta pela API Lyrics.ovh destacou mais a palavra "like" do que a "love", resultado contrário para o teste com IA generativa, mas em ambos as duas palavras foram extremamente significativas. Por outro lado, outras palavras foram destacadas, além de novamente as palavras "want" e "make", foram destacadas as palavras "feel" (sentir) e "wanna" (quero).

3.3 Análise de sentimentos

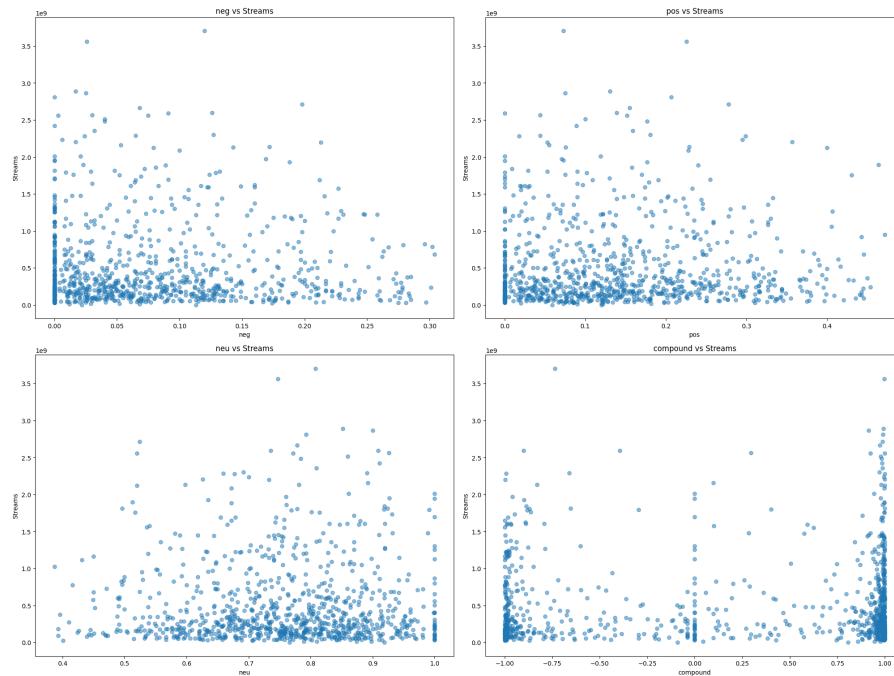


Figura 8: Gráficos de dispersão de número de streams x sentimentos. Da esquerda para a direita, de cima para baixo: neg, pos, neu, compound.

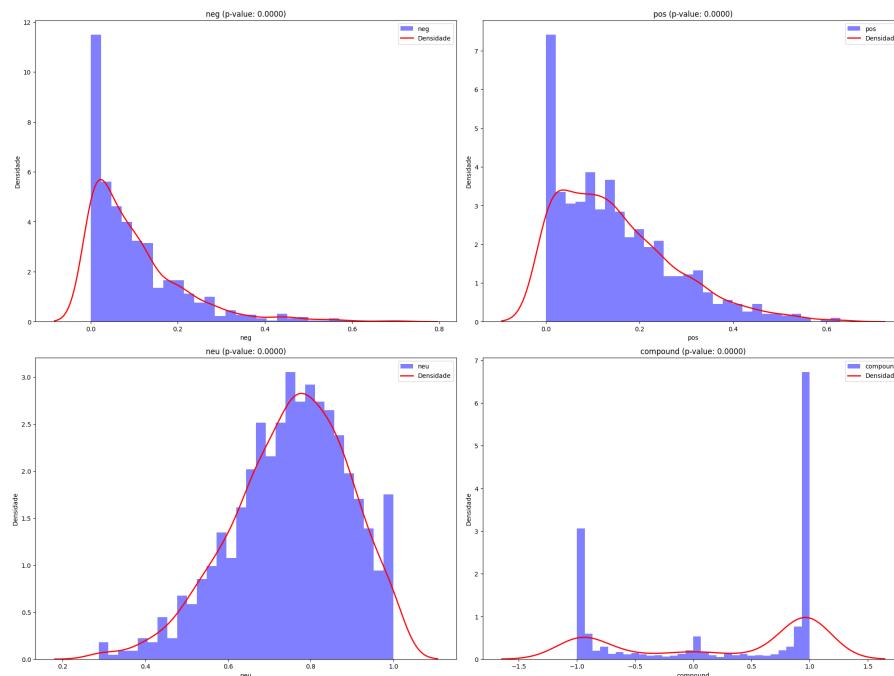


Figura 9: Histogramas de número de streams x sentimentos. Segue a mesma distribuição da Figura 8

A partir da análise de sentimentos utilizando VADER, pode-se perceber, a partir dos gráficos de dispersão e histogramas das Figuras 8 e 9, que não há correlação visível entre o número de streams e as variáveis que representam sentimentos negativos, positivos ou neutros.

Entretanto, note que há uma maior predominância de sentimentos positivos em relação aos negativos, de acordo com os histogramas, embora os sentimentos neutros predominem. Ainda assim, note que em "compound" há mais distribuição de valores próximos do 1.0, denotando mais predominância de sentimentos positivos. Note também que no gráfico de dispersão de "compound", os sentimentos negativos e positivos apresentaram maior destaque no número de streams, mas principalmente os positivos, percebido pela quantidade de pontos com valores altos no eixo y, por exemplo acima de 1.5×10^9 streams.

Note que a análise da "valência" na seção 3.1.2 apresenta um comportamento similar: uma tendência tênue das músicas apresentarem sentimentos mais voltados aos positivos.

3.4 Análise de correlação de Spearman

Sob uso do teste de normalidade de Shapiro-Wilk, verificou-se que todas as variáveis estudadas no problema assumiam distribuição não normal. Haja vista que o índice de correlação de Spearman performa melhor para distribuições não normais, em relação ao índice de Pearson, o de Spearman foi escolhido para analisar a correlação entre 2 variáveis.

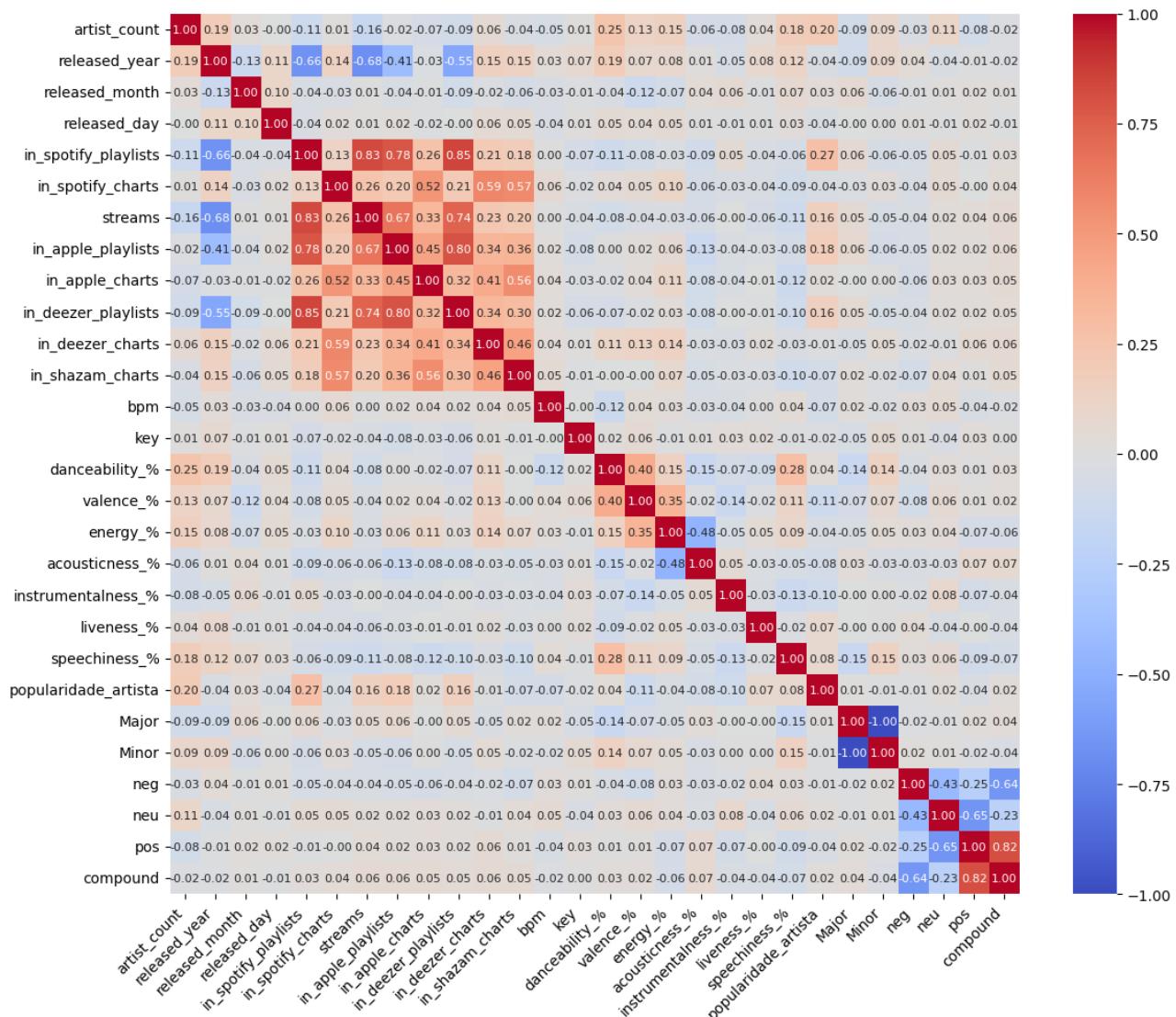


Figura 10: Matriz de correlação com índice de Spearman.

A variável que queremos avaliar é "streams", ao compará-la com as outras, obtemos algumas correlações em destaque:

- **Presença em playlists do Spotify (0.83), do Apple Music (0.67) e Deezer (0.74):** embora seja um resultado diretamente esperado, haja vista que os dados são do Spotify, é notável uma correlação mais forte de músicas com altas streams no Spotify com a presença em playlists do Deezer em comparação à Apple Music, ainda que pequena. Esse resultado numérico é compatível visualmente pelos gráficos de dispersão da Figura 1.
- **Ano de lançamento da música (-0.68):** de acordo com a correlação quase forte apresentada, quanto mais recente a música, no que tange o ano de lançamento, mais ela tende a estar popular naquele ano. Essa análise será melhor apresentada na próxima seção "Análise Temporal".
- **Popularidade do Artista (0.16):** testado com os 10% de músicas mais famosas, a proporção subiu para 0.22. Ainda assim, é uma correlação baixa, o que contradiz a hipótese levantada de que teria uma alta correlação. É provável que essa variável de popularidade tenham outros fatores a serem ponderados, o que pode alterar os resultados.
- **Características da música:** variáveis terminadas com "%", além de bpm (batidas por minuto), escala maior ou menor e nota predominante da música apresentaram índices muito próximos de 0, denotando quase inexistência de correlação direta entre essas variáveis, individualmente, e o número de streams.
- **Variáveis de sentimentos:** tal qual as variáveis de características das músicas, apresentaram correlação muito próxima de 0, denotando que o número de streams e essas variáveis não possuem correlação direta aparente.

Uma observação necessária a ser pontuada nesses testes de correlação é que os dados disponibilizados são referentes apenas às músicas com mais streams do Spotify, com um total de 952. Devido a esse fato, muitas características sobre as músicas apresentam valores numéricos quase invariáveis ou pouco variáveis, como "instrumentalness", que apresenta a maioria de seus valores próximos a 0%. Por esse fator, a correlação entre "instrumentalness" e "streams" foi 0, mesmo que quase todas as músicas populares no Spotify em 2023 não sejam instrumentais, segundo os dados. Uma forma de verificar mais profundamente essas correlações seria a comparação com músicas que não atingiram um número alto de streams ou análises comparativas ou de distribuição, como foi feito nos histogramas.

Nesse contexto, valores muito próximos de 0 não necessariamente indicam que aquela variável é desprezível para que uma música tenha muitas streams.

3.5 Análise temporal

Os resultados obtidos da análise temporal podem ser observados nos gráficos abaixo, que destacam as principais tendências e padrões nos dados ao longo do tempo.

3.5.1 Streams por data de lançamento

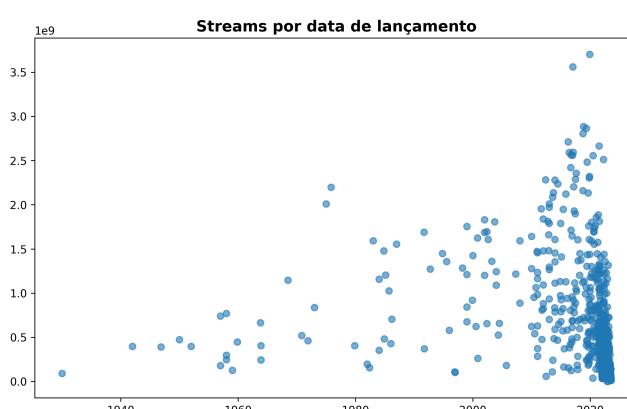


Figura 11: Gráfico de dispersão.

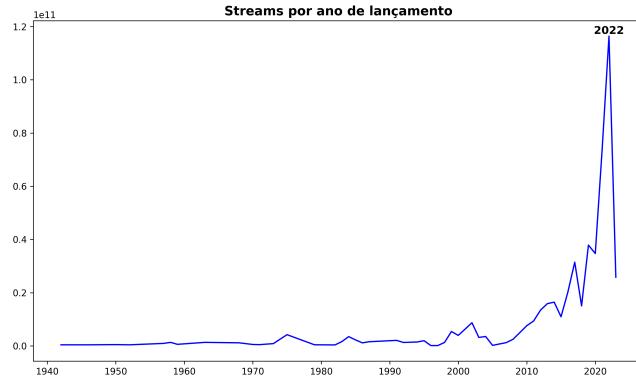


Figura 12: Gráfico de linha.

Os gráficos acima indicam um aumento significativo no número de streams de músicas lançadas a partir de 2000, refletindo a popularização das plataformas de streaming digital. Antes de 1980, o número de streams é consideravelmente menor, sugerindo que músicas mais antigas são menos acessadas ou procuradas nas plataformas atuais. A partir de 2010, há uma explosão no número de músicas com alta quantidade de streams, que obtém seu auge em 2022, indicando um crescimento da disponibilidade e do consumo de músicas digitais nos últimos anos. Isso sugere uma preferência por lançamentos recentes e o impacto das tecnologias de streaming no comportamento de consumo musical.

3.5.2 Streams por mês de lançamento

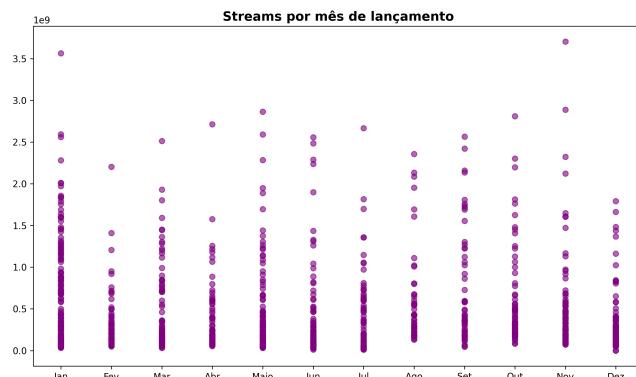


Figura 13: Gráfico de dispersão.

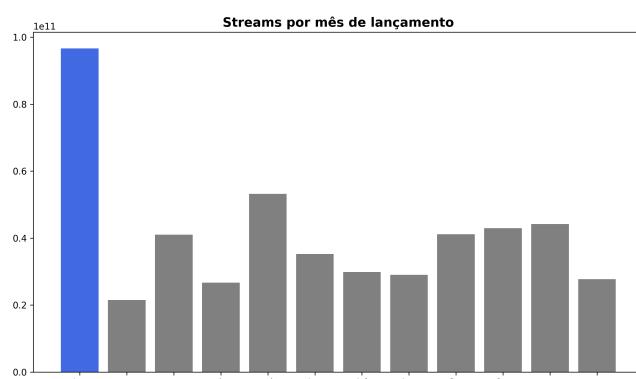


Figura 14: Gráfico de colunas.

Em relação à análise mensal, é possível observar que janeiro é o mês com a maior concentração de streams, destacando-se de maneira significativa em comparação aos outros meses. No entanto, não há uma tendência clara de crescimento ou declínio de streams ao longo do ano, o que sugere que os picos estão mais relacionados a fatores pontuais, como eventos específicos, datas comemorativas, ou calendários de lançamentos de álbuns em vez de fatores sazonais.

3.5.3 Streams por dia de lançamento

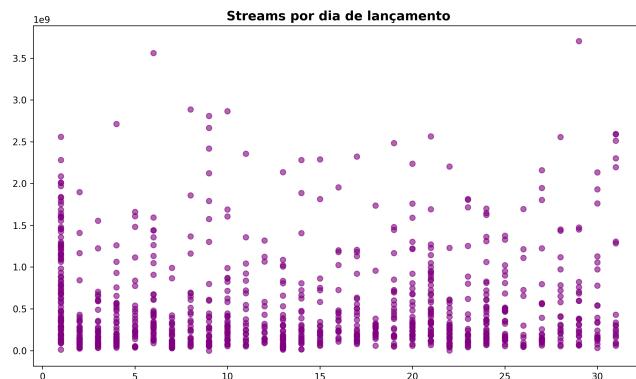


Figura 15: Gráfico de dispersão.

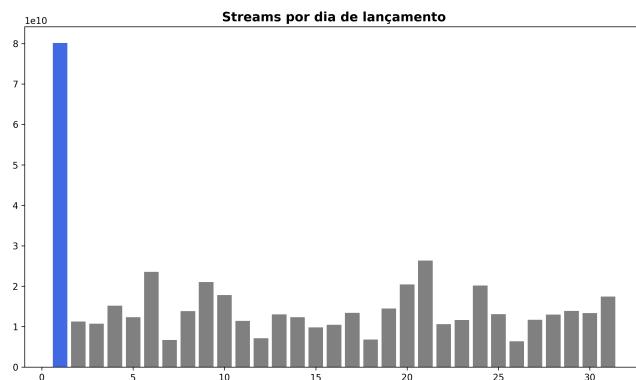


Figura 16: Gráfico de colunas.

Os gráficos acima revelam que o dia 1 é o principal dia no que diz respeito às concentrações de streams, com um volume significativamente superior aos demais. Picos menores ocorrem nos dias 6, 9 e 21, enquanto os outros dias mostram uma distribuição mais equilibrada e baixa. Após o dia 20, há uma leve elevação no número de streams. De modo geral, o gráfico sugere que o início do mês é um momento estratégico para lançamentos, possivelmente por gerar maior visibilidade ou interesse dos usuários, com oportunidades adicionais ao longo do mês.

3.5.4 5 artistas mais ouvidos

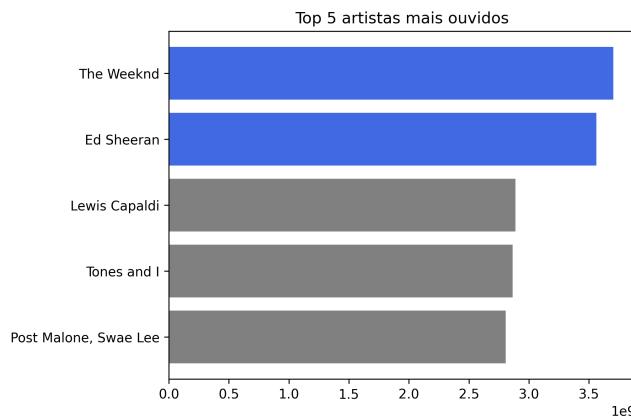


Figura 17: Gráfico de barras.

O gráfico da Figura 17 explicita os cinco artistas mais ouvidos, com "The Weeknd" liderando, seguido por "Ed Sheeran". Ambos têm audiências próximas, cada um com mais de 3 bilhões de reproduções, enquanto os demais possuem em torno de 2,5 bilhões. É possível observar que o gênero predominante é o pop, o que reflete o gosto popular e as tendências de consumo musical atuais. O pop é conhecido por sua capacidade de adaptação, incorporando elementos de outros estilos, como R&B, hip hop, eletrônica e até folk, o que o torna mais versátil e acessível para uma variada classe de ouvintes.

3.5.5 Músicas mais ouvidas por década

Década	Música	Streams
2020	STAY (with Justin Bieber)	2.67B
2010	Blinding Lights	3.70B
2000	Lose Yourself - Soundtrack Version	1.83B
1990	Yellow	1.76B
1980	Every Breath You Take - Remastered 2003	1.59B
1970	Bohemian Rhapsody - Remastered 2011	2.20B
1960	Have You Ever Seen The Rain?	1.15B
1950	Rockin' Around The Christmas Tree	0.77B
1940	White Christmas	0.40B

Figura 18: Músicas mais ouvidas por década e seus respectivos números de streams.

As músicas mais populares por década revelam duas tendências distintas: enquanto as faixas mais ouvidas nas décadas de 2010 e 2020, como "Blinding Lights" e "STAY" refletem um aumento significativo em streams devido ao crescimento das plataformas de streaming, músicas antigas como "Bohemian Rhapsody" e "Every Breath You Take" mantêm sua relevância ao longo do tempo, gerando milhões de streams, especialmente após remasterizações. Isso evidencia um sentimento de nostalgia do público, em que o catálogo antigo se beneficia das novas tecnologias e alcança novos usuários.

3.6 Análise de Regressão de Quadrados Mínimos Ordinários

Dado o objetivo principal do projeto de pesquisar a correlação entre características musicais e a popularidade da música, medida pelo número de streams, estrutura-se então uma correlação múltipla entre algumas variáveis independentes e o número de stream como variável dependente.

Diferente da análise de correlação de Spearman, que analisa pares de variáveis, a análise de Regressão de Quadrados Mínimos Ordinários busca compreender a relação entre um conjunto de variáveis e o número de streams.

Para esta primeira análise utilizou-se o método de Quadrados Mínimos Ordinários (OLS), cujo algoritmo não necessitou de mudanças e transformações dos dados fornecidos, pois as colunas de interesse nessa análise não possuem valores nulos ou fora do tipo. Foram consideradas as variáveis referentes às características musicais: danceability, valence, etc. Na primeira rodada de regressões utilizamos 2 variáveis independentes e o número de streams como variável dependente.

O resultado obtido não correspondeu às hipóteses de possível correlação, uma vez que quase todas as regressões descreveram uma porcentagem muito baixa da variância dos dados de streams, o maior valor para R^2 obtido foi 0.02. Além disso, neste modelo as variáveis independentes que possuem significância estatística apresentam coeficientes negativos.

OLS Regression Results						
Dep. Variable:	streams	R-squared:	0.020			
Model:	OLS	Adj. R-squared:	0.018			
Method:	Least Squares	F-statistic:	9.710			
Date:	Thu, 31 Oct 2024	Prob (F-statistic):	6.69e-05			
Time:	19:19:31	Log-Likelihood:	-20529.			
No. Observations:	952	AIC:	4.106e+04			
Df Residuals:	949	BIC:	4.108e+04			
Df Model:	2					
Covariance Type:	nonrobust					
<hr/>						
	coef	std err	t	P> t	[0.025	0.975]
<hr/>						
const	7.975e+08	8.54e+07	9.336	0.000	6.3e+08	9.65e+08
danceability	-3.399e+06	1.27e+06	-2.683	0.007	-5.88e+06	-9.13e+05
speechiness	-5.497e+06	1.87e+06	-2.941	0.003	-9.17e+06	-1.83e+06
<hr/>						
Omnibus:		382.767	Durbin-Watson:			0.040
Prob(Omnibus):		0.000	Jarque-Bera (JB):			1366.930
Skew:		1.967	Prob(JB):			1.50e-297
Kurtosis:		7.358	Cond. No.			325.
<hr/>						

Figura 19: Relatório da regressão OLS para as variáveis "danceability" e "speechiness".

Em continuação, usando regressão OLS, foi adicionada mais uma variável independente a esta análise. De forma semelhante à regressão anterior, não foi possível obter uma boa descrição dos dados com este modelo. As regressões apresentaram resultados semelhantes, com valor máximo para R^2 de 0.024. De forma similar à análise anterior, não apenas considerando o valor de R^2 , as variáveis independentes que apresentaram significância estatística não apresentaram coeficientes positivos.

OLS Regression Results						
Dep. Variable:	streams	R-squared:	0.024			
Model:	OLS	Adj. R-squared:	0.021			
Method:	Least Squares	F-statistic:	7.702			
Date:	Thu, 31 Oct 2024	Prob (F-statistic):	4.36e-05			
Time:	19:19:31	Log-Likelihood:	-20527.			
No. Observations:	952	AIC:	4.106e+04			
Df Residuals:	948	BIC:	4.108e+04			
Df Model:	3					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	8.188e+08	8.6e+07	9.518	0.000	6.5e+08	9.88e+08
danceability	-3.581e+06	1.27e+06	-2.823	0.005	-6.07e+06	-1.09e+06
instrumentalness	-4.147e+06	2.18e+06	-1.906	0.057	-8.42e+06	1.23e+05
speechiness	-5.742e+06	1.87e+06	-3.069	0.002	-9.41e+06	-2.07e+06
Omnibus:	379.107	Durbin-Watson:	0.048			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1341.953			
Skew:	1.949	Prob(JB):	3.97e-292			
Kurtosis:	7.317	Cond. No.	328.			

Figura 20: Relatório da regressão OLS, adicionando a variável "instrumentalness".

3.7 Análise de Componentes Principais

A Análise de Componentes Principais (ACP) organiza as colunas numéricas para entender e compreender melhor a variabilidade dos dados a serem analisados. O primeiro componente principal é a direção que maximiza a variância dos dados, enquanto o segundo componente é ortogonal ao primeiro e maximiza a variância restante. Foram utilizados nesta tabela apenas 2 componentes principais.

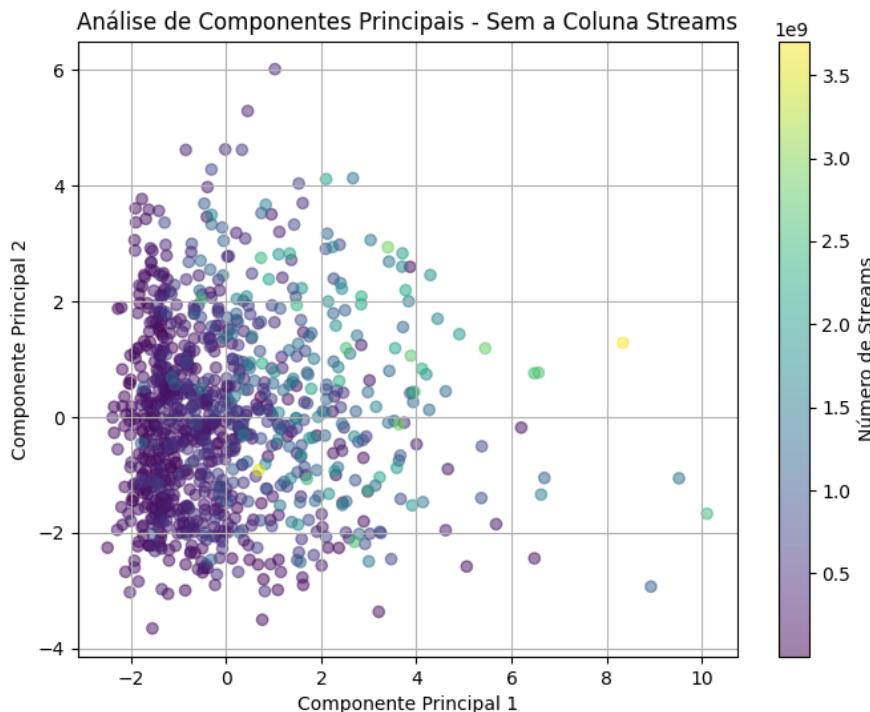


Figura 21: Gráfico de dispersão da análise multivariável.

Com análises mais específicas, disponíveis no repositório on-line, é possível encontrar mais precisamente quais características descrevem cada componente:

- Componente principal 1: inappleplaylists, inapplecharts, inspotifyplaylists, inspotifycharts.
- Componente principal 2: acousticness, instrumentalness.

Entretanto este modelo explica pouco da variância:

- Componente principal 1: 0.139
- Componente principal 2: 0.111

3.8 Análise de Probabilidade

Por último, é possível entender alguma probabilidade dos dados. Primeiramente, analisando novamente apenas as características utilizadas na OLS, separamos os dados em classes para facilitar a interpretação das porcentagens:

- 0% – 20%: 1
- 21% – 40%: 2
- 41% – 60%: 3
- 61% – 80%: 4
- 81% – 100%: 5

E buscou-se compreender o comportamento dos 10% de músicas mais ouvidas, cujo número de streams é maior que 1302145668.5, e o conjunto de dados por completo.

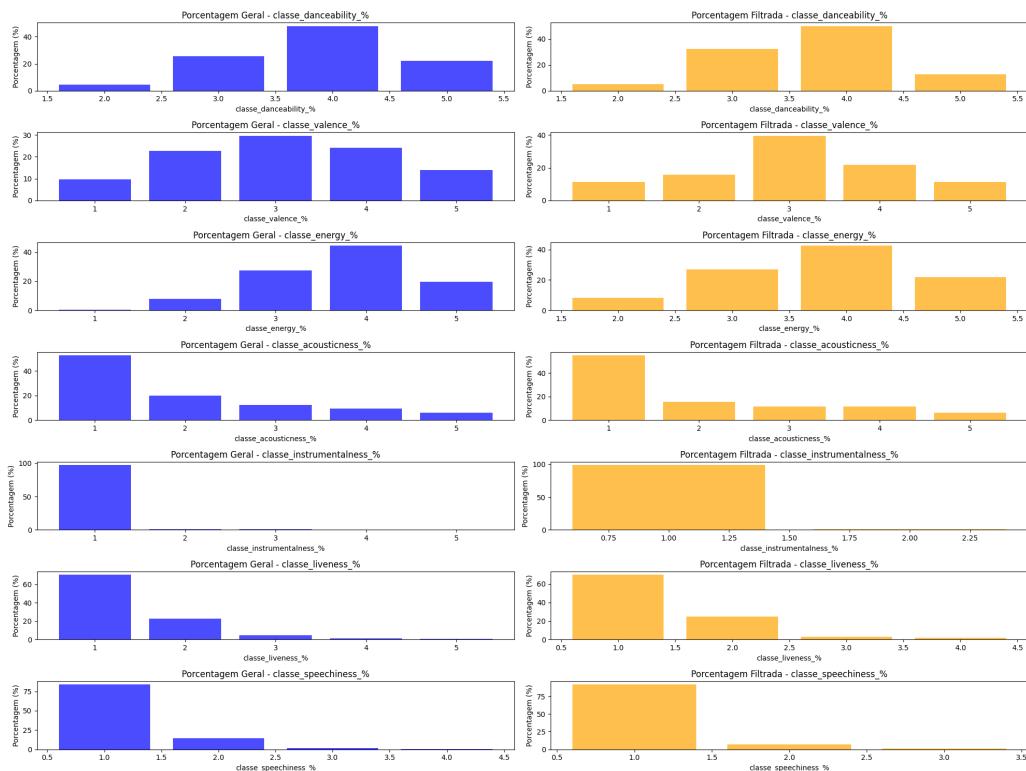


Figura 22: Gráficos de colunas. Em azul e à esquerda, fração dos 10% mais ouvidas, à direita o conjunto inteiro.

Em seguida buscamos ver quais músicas do conjunto de dados seguiam esse comportamento:

```

Música: Mine (Taylor's Version), Streams: 36912123
Música: Mejor Que Yo, Streams: 50847624
Música: VOID, Streams: 67070410
Música: Hold Me Closer, Streams: 284216603
Música: Crazy What Love Can Do, Streams: 286739476
Música: You Belong With Me (Taylor Swift's Ve, Streams: 350381515
Música: Acapulco, Streams: 363467642
Música: Take My Breath, Streams: 403097450
Música: Take My Breath, Streams: 432702334
Música: Angels Like You, Streams: 570515054
Música: Yandel 150, Streams: 585695368
Música: I Like You (A Happier Song) (with Doja Cat), Streams: 609293408
Música: The Motto, Streams: 656013912
Música: Style, Streams: 786181836
Música: Sure Thing, Streams: 950906471
Música: Anti-Hero, Streams: 999748277
Música: Dandelions, Streams: 1116995633
Música: Astronaut In The Ocean, Streams: 1138474110
Música: Kill Bill, Streams: 1163093654
Música: Set Fire to the Rain, Streams: 1163620694
Música: Pepas, Streams: 1309887447
Música: Blank Space, Streams: 1355959075
Música: Adore You, Streams: 1439191367
Música: Rolling in the Deep, Streams: 1472799873
Música: Counting Stars, Streams: 2011464183
Música: Something Just Like This, Streams: 2204080728
Total de músicas na classe de maior porcentagem para todas as colunas: 26
Número de linhas da planilha: 952

```

Figura 23: Relatório de músicas e número de streams com comportamento descrito.

Note que a partir desta lista podemos notar que apenas 6 músicas no top 10% estão nas classes com maior porcentagem, e apenas 26 de todas as musicas estão nessas classes. Assim, podemos fazer a probabilidade de uma música estar no top 10% sabendo que está nas classes com maior porcentagem:

- *A*: Estar nas classes com maior porcentagem.
- *B*: Estar no top 10% de streams.

De tal forma, é possível calcular:

$$\begin{aligned}
 P(A \cap B) &= \frac{6}{952} = 0,00630 \\
 P(A) &= \frac{26}{952} = 0,0273 \\
 P(B|A) &= \frac{0,00630}{0,0273} = 0,2307
 \end{aligned}$$

Ou seja, se uma música seguir o padrão de características expressos acima, ela tende, analisando exclusivamente as características técnicas da música, a ter uma probabilidade de 23,7% de estar no top 10% músicas mais ouvidas do ano.

4 Conclusão

Dada as análises dos resultados, percebe-se que, embora muitos testes de correlação e regressão apresentassem uma correlação baixa ou neutra entre o número de streams e características das músicas, artistas e datas de lançamento, alguns padrões conseguiram ser percebidos.

De tal forma, é possível perceber uma tendência de músicas populares do Spotify em 2023 de não serem a versão "ao vivo" da música, nem acústicas e muito menos instrumentais, além de terem pouca porcentagem de "fala" nas músicas (speechiness).

Além disso, palavras comuns nas músicas costumam ser "amor", "saber", "gostar", "gostar" e "fazer", e os sentimentos refletidos pelas músicas tendem suavemente a serem positivos, mas com quase equilíbrio entre sentimentos positivos e negativos.

Também foi possível perceber um padrão de lançamento das músicas no primeiro dia dos meses, além de, separadamente, muitas serem lançadas em janeiro. Em complemento, a maioria das músicas com popularidade alta em 2023 tendem a ser mais novas, maioria lançadas em 2022, com queda brusca para 2021 e os anos anteriores.

Por fim, foi possível concluir, estatisticamente, que a presença das músicas em playlists ou popularidade das músicas em outras plataformas de música senão o Spotify também influenciam positivamente em sua popularidade na plataforma estudada.