Primeiro trabalho prático de ICD
2020/1

Ábner de Marcos Neves, Leonardo de Almeida Brito

Departamento de Ciência da Computação da UFMG

1. Introdução

Hot 100 é uma tabela musical mantido pela *Billboard* que elenca as cem músicas mais vendidas ao longo de uma semana. O *dataset* deste trabalho consiste em todos esses charts semanais de 1958 a 2019. Entretanto, para termos mais informações acerca das músicas presentes nesse chart, acrescentamos também um *dataset* oriundo do Spotify, que possui métricas mais detalhadas de cada música.

Iremos analisar as *features* das músicas que entraram no *Hot 100*, assim como sua relação com a posição delas no rank.

As perguntas que tentamos responder foram as seguintes:

* Alguma característica da música pode influenciá-la a ficar em #1?

* O gênero da música influencia o tempo (BPM) dela?

* Como foi a evolução das características das músicas mais ouvidas ao longo dos anos?

2. Metodologia

2.1. Base de dados

A base de dados sobre a qual se desenvolve esse trabalho é composta por duas tabelas: todos os *Hot 100* semanais de 02/08/1958 até 28/12/2019 e as características das músicas que aparecem nesse chart.

As informações mais relevantes da primeira tabela, com nomes autoexplicativos:

* SongID

* Song name

* Performer name

* Week position

Na segunda tabela, proveniente do Spotify, há muitas informações, dentre elas:

* spotify_genre

* duration

* danceability

* energy

* acousticness

* tempo

Algumas dessas *features* são peculiares ao Spotify. Descrições mais detalhadas de cada uma podem ser lidas neste link.

2.2. Tratamento dos dados

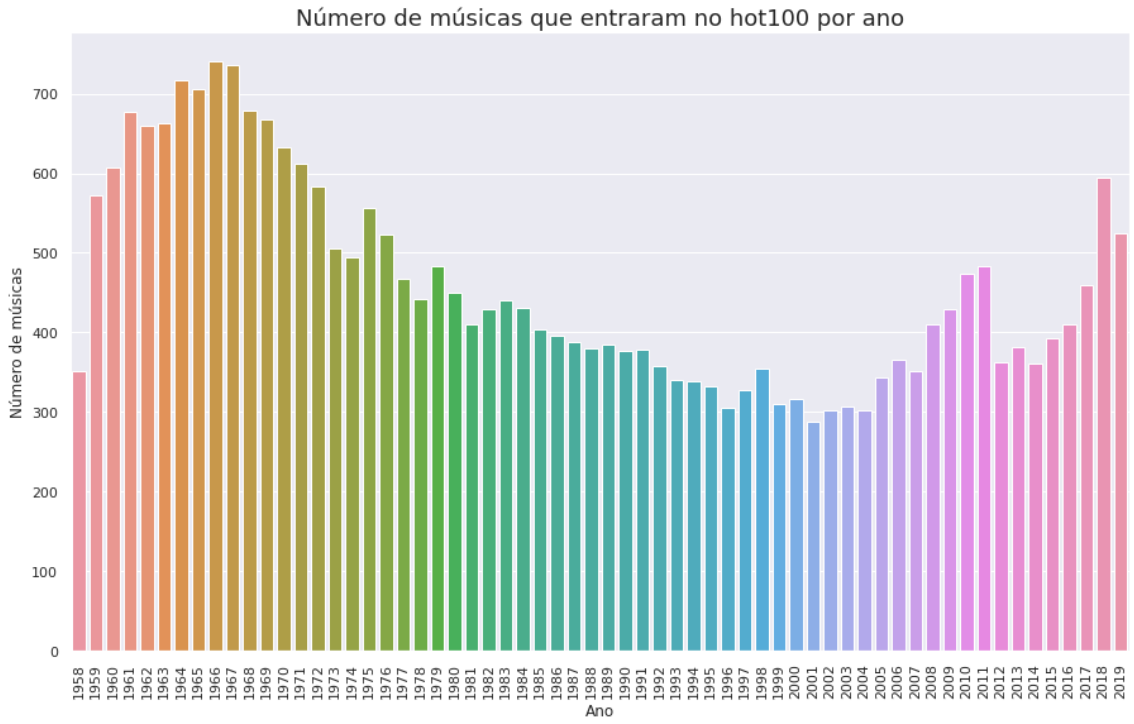
Para responder às perguntas foram necessários alguns ajustes nas tabelas, como a eliminação de colunas desnecessárias ao nosso objetivo, sendo elas as de *url*, *spotify_track_preview_url*, *spotify_track_album*. E, consequentemente, algumas colunas foram adicionadas para facilitar a análise dos dados: *ano*, *weeks on top1*, *weeks on chart*, *chart debut*, *melhor posicao*.

Em determinado ponto, os gêneros das músicas foram quebrados e transformados em colunas categóricas, a fim de possibilitar uma regressão linear e outras análises.

3. Resultados

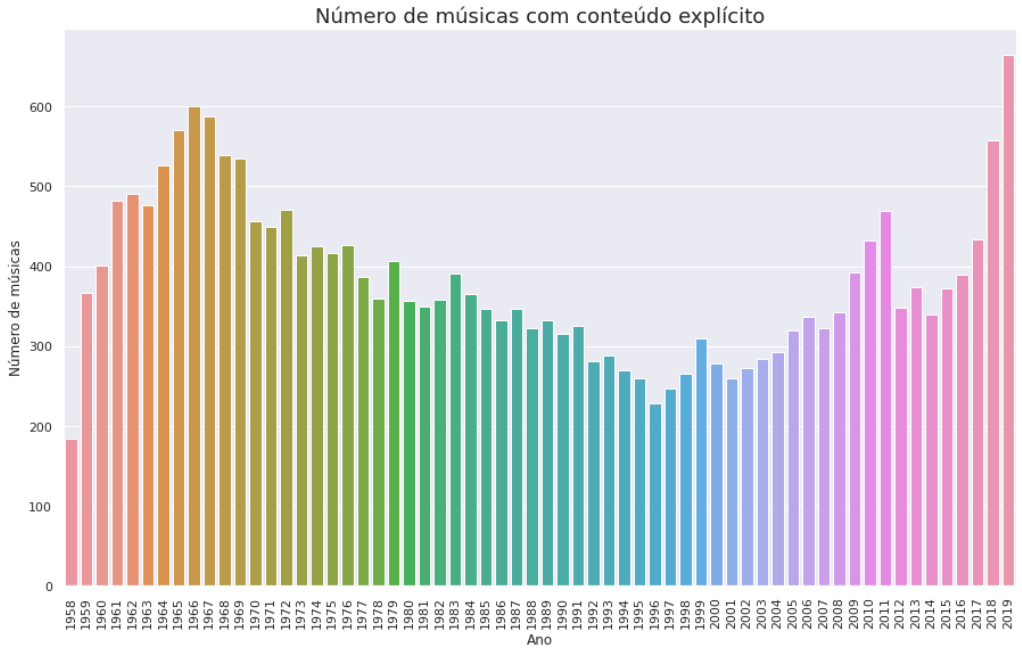
3.1. Análise exploratória

Quantas músicas entraram na tabela musical por ano?



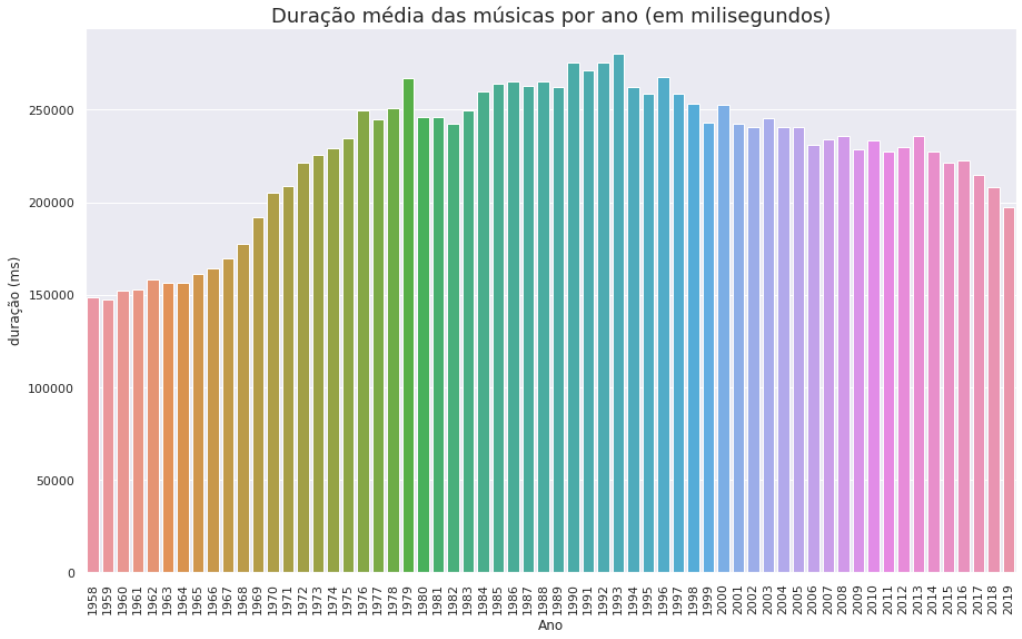
Durante a década de 60, um número maior de músicas conseguia entrar na tabela musical, enquanto que a partir da década de 70 até os anos 2000, as músicas tiveram uma permanência maior nos charts, barrando a entrada de outras. Na ultima década, houve um aumento no número de entrada de diferentes músicas.

Quantas músicas com conteúdo explícito entraram na tabela musical por ano?



Músicas com conteúdo explícito tiveram seu ápice de popularidade na década de 60 até pouco tempo, quando foi ultrapassada pela década atual em que voltaram a crescer e ganhar popularidade. A década de 90 foi o período ondem tiveram menos entrada de músicas com conteúdo explicito na tabela musical.

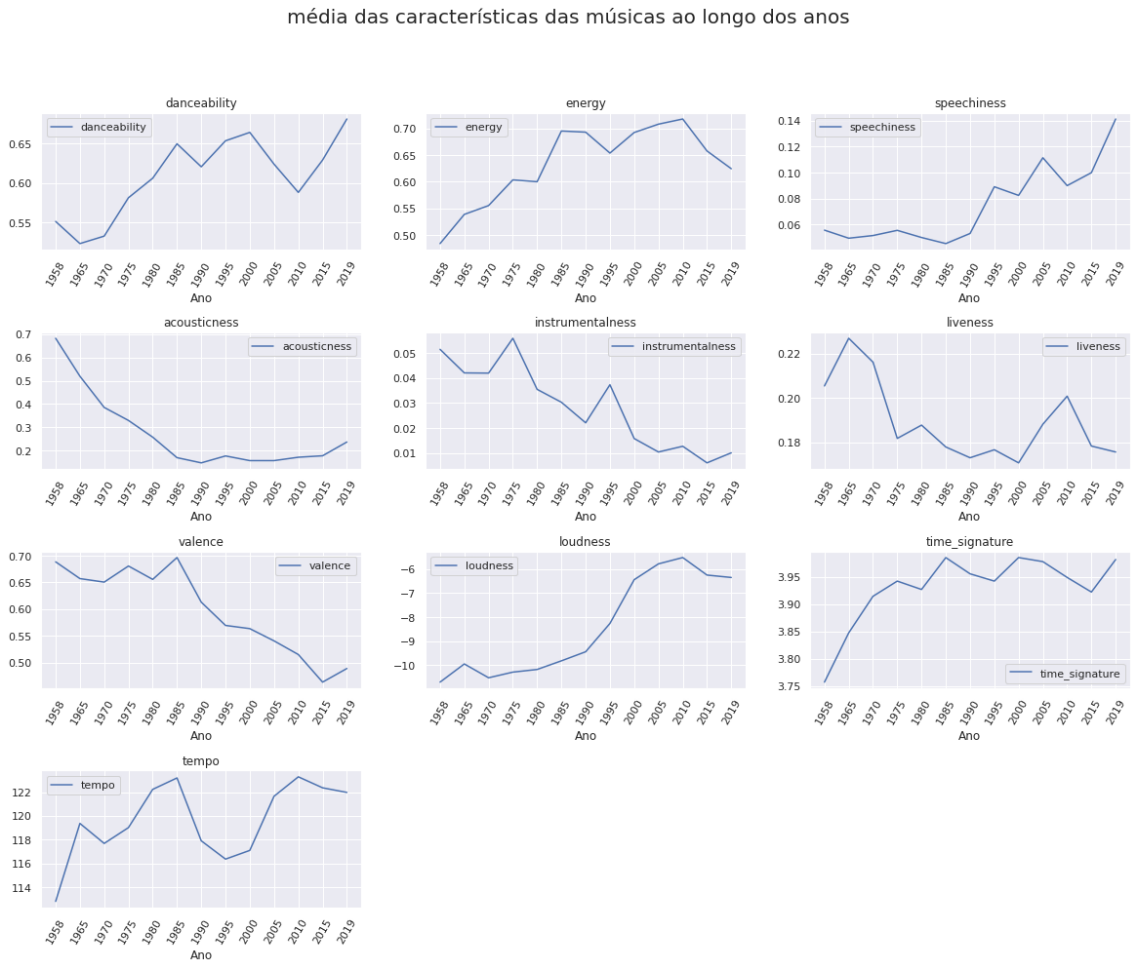
Qual a duração média das músicas que entraram na tabela musical por ano?



Durante as primeiras décadas registradas, a duração das músicas era bem menor

se comparando os outros períodos. Podendo ser observado um aumento gradual ao longo do anos até os anos 2000, onde se iniciou uma diminuição até o período atual.

Qual a média das características das músicas por ano?



danceability: cresceu ao longo dos anos.

energy: cresceu ao longo dos anos.

speechiness: cresceu ao longo dos anos.

acousticness: diminuiu ao longo dos anos.

instrumentalness: diminuiu ao longo dos anos.

liveness: diminuiu ao longo dos anos.

valence: diminuiu ao longo dos anos.

loudness: cresceu ao longo dos anos.

time_signature: cresceu ao longo dos anos.

tempo: teve momentos de altos e baixos, voltando a crescer recentemente.

3.2. Testes de hipótese

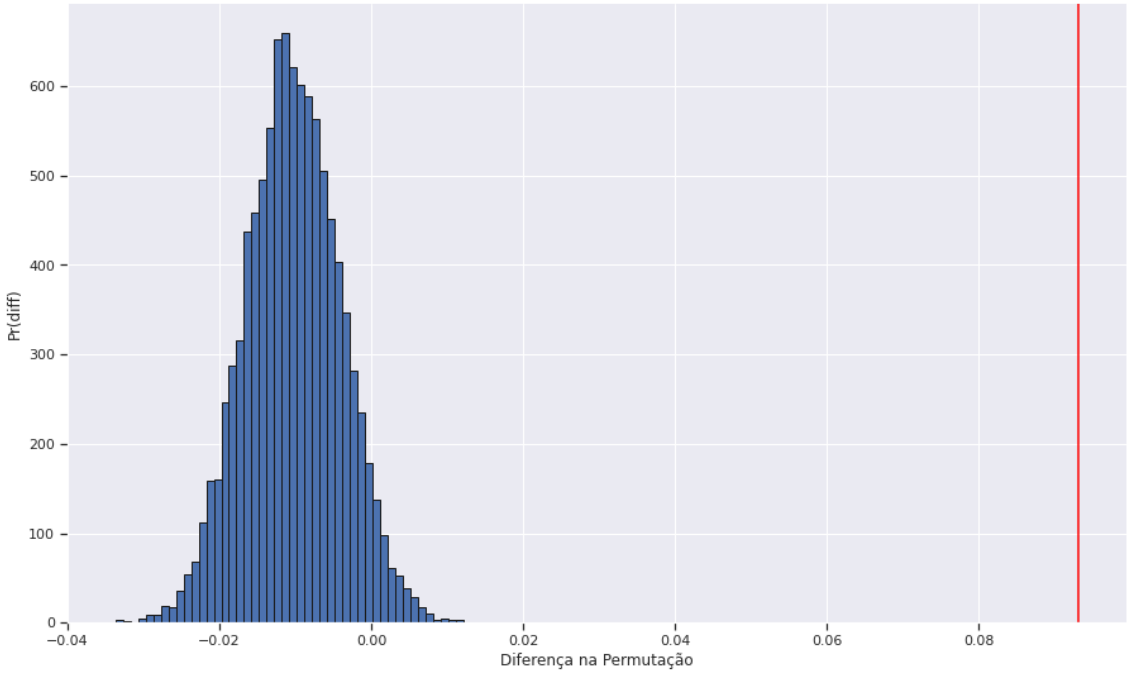
3.2.1. Como foi a evolução das características das músicas mais ouvidas ao longo dos anos?

Hipótese nula: a característica *danceability* das músicas ao longo dos anos é igual/a característica *danceability* não se alterou ao longo dos anos.

Vamos comparar a *danceability* de dois anos: 2010 e 2019.

Inicialmente, exploramos os dados, selecionando as características de todas as musicas que apareceram na tabela nesses dois anos. Em seguida, pegamos o *danceability* médio de cada ano e calculamos sua diferença que foi em torno de 0.09. Tal estatística será a nossa *t_obs*.

Em seguida, realizamos o teste de permutação que nos gerou o seguinte gráfico para 10000 amostras.

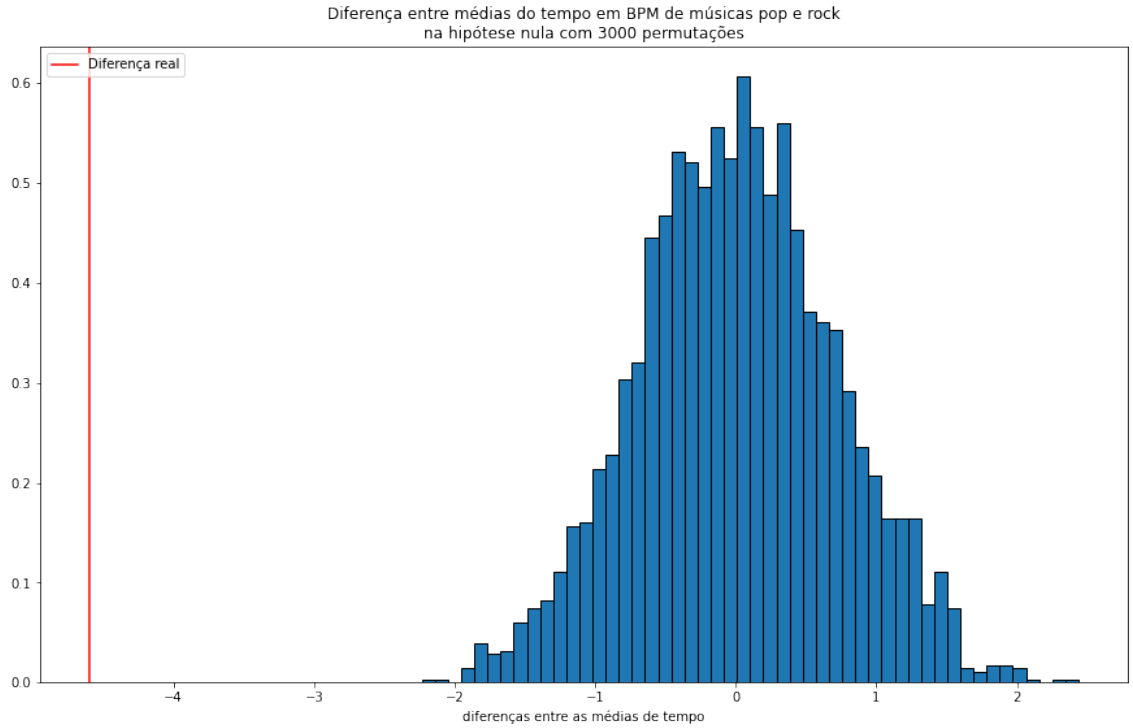


O traço vermelho mostra o valor que observamos, as barras mostram as diferentes amostras.

t_obs é bastante raro, logo rejeitamos a hipótese nula e indicamos que a variação dos valores de *danceability* não pode ser explicado pelo acaso. Sendo observado uma alteração no valor da característica em questão entre os anos 2010 e 2019.

3.2.2. Há uma diferença de tempo (bpm) entre músicas de pop e rock?

Para responder a essa pergunta, fizemos um teste de permutação que avaliasse a diferença entre as médias de tempo das músicas categorizadas nesses dois gêneros.

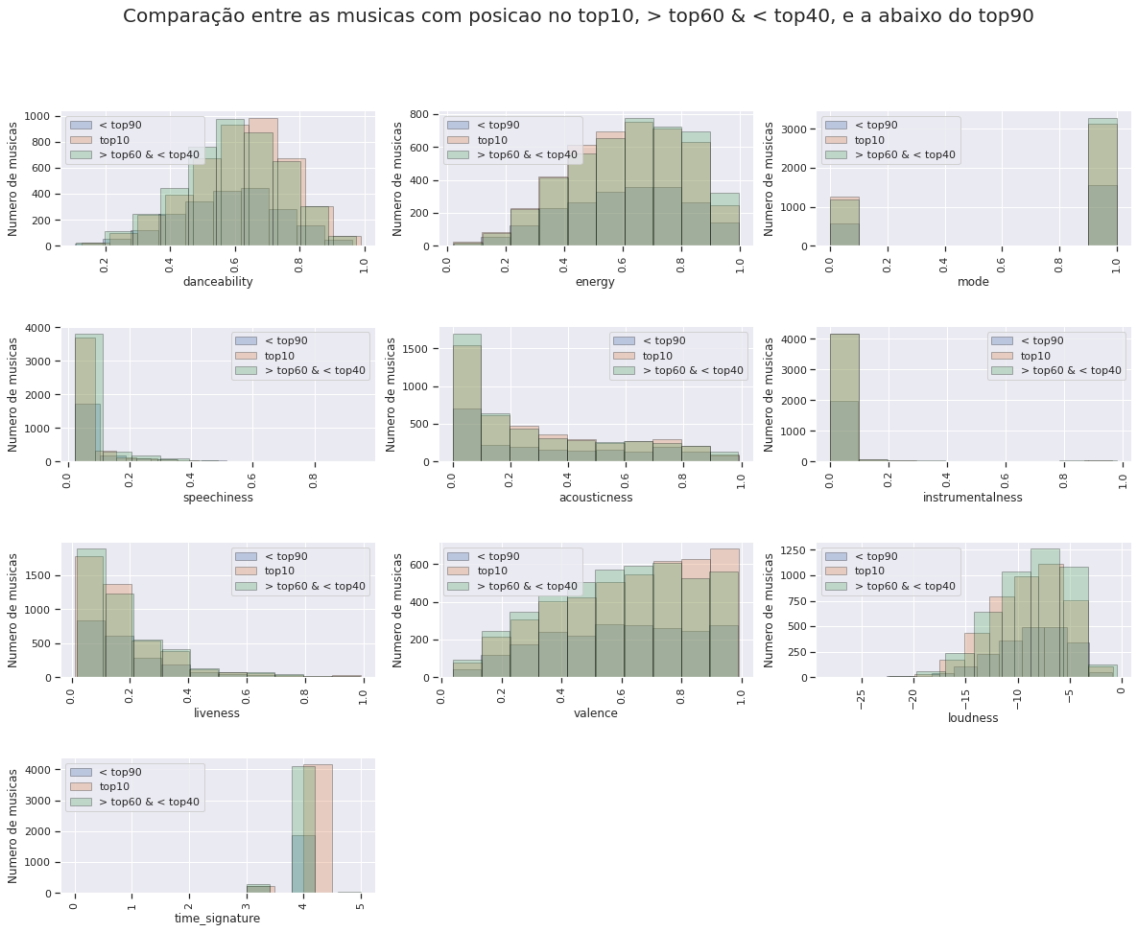


O histograma mostra que as diferenças entre as médias de BPM das músicas categorizadas como pop e as categorizadas como rock giram em torno de 0 quando se supõe que não há relação entre esses dois gêneros e seus respectivos tempos. Entretanto, a diferença observada é completamente disjunta do que se esperava na hipótese nula. Logo, a hipótese nula não é aceita: há, sim, nas músicas de pop e rock, uma relação entre o tempo e seus gêneros.

3.3. IC - Intervalo de confiança

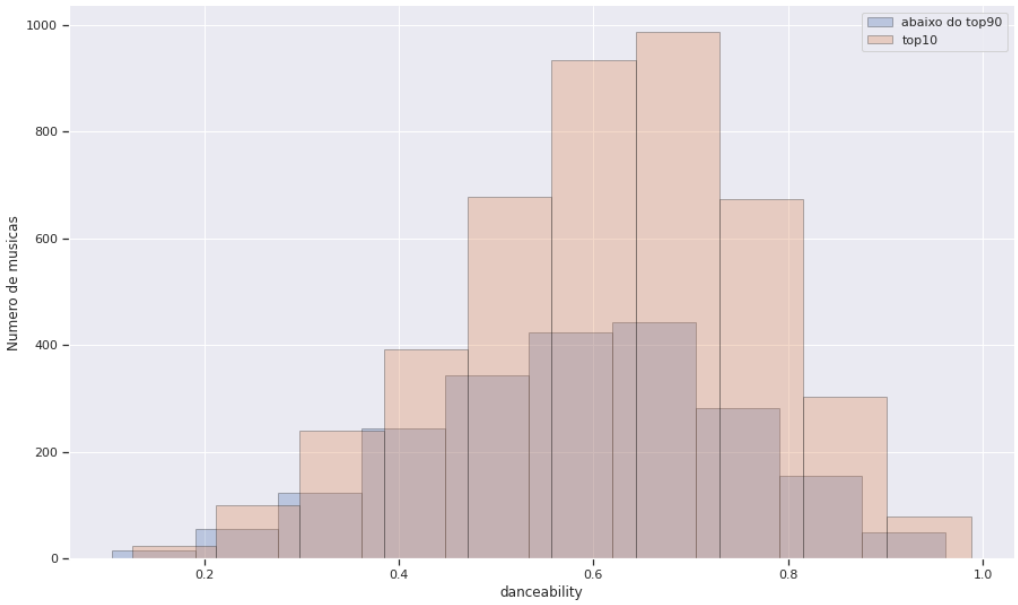
Alguma característica da música pode influenciá-la a ficar em #1?

Vamos comparar as características das músicas que ficaram no top10 com as que chegaram apenas até a 90 posição e com as que ficaram entre as posições 60 e 40. Para isso fizemos um tratamento dos dados, adicionando uma coluna da melhor posição de cada música.



Pelos resultados, as características não aparentam influenciar na posição da música, com exceção dos gráficos para *danceability*, *speechiness* e *loudness* que apresentam uma pequena diferença entre esses conjuntos de dados.

Vamos selecionar uma característica para aprofundar o resultado.



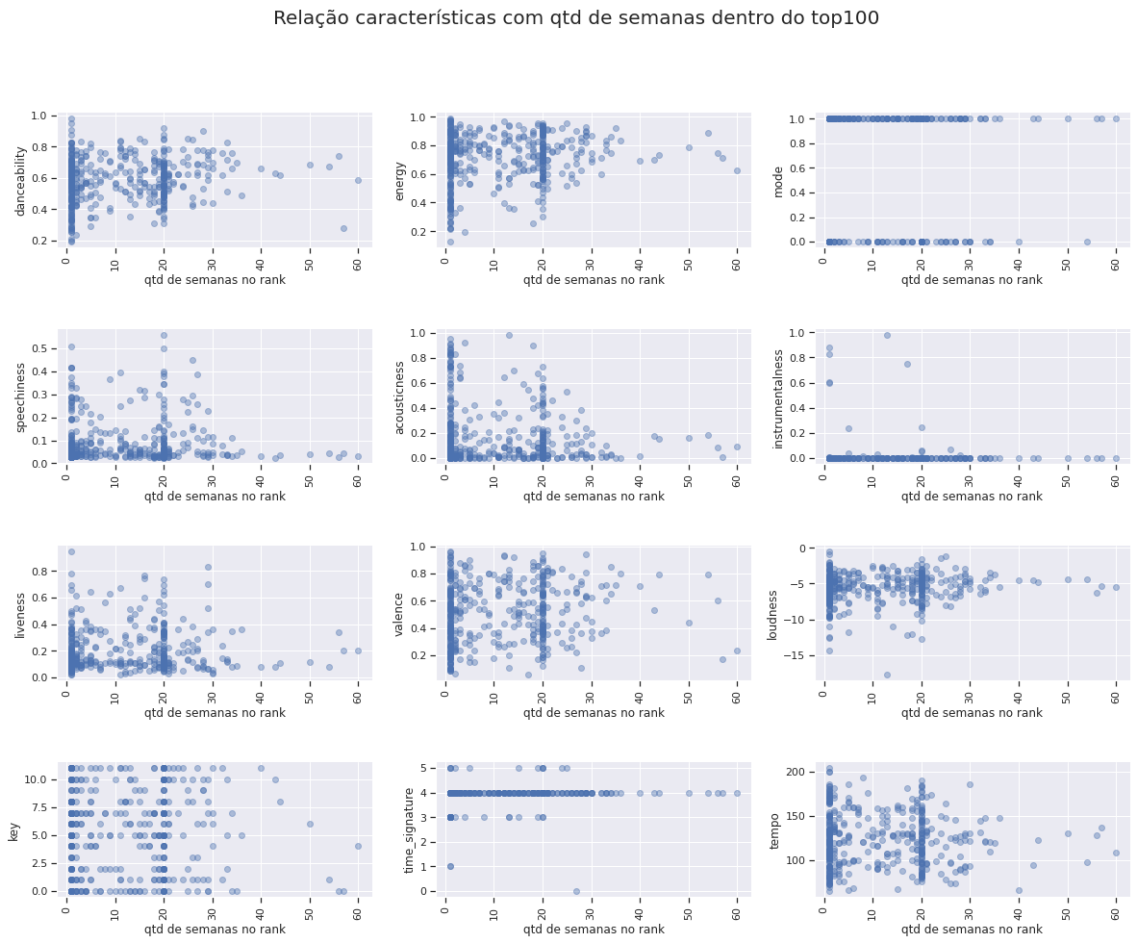
Casos:

top10: (0.6124385242756285, 0.6212301454289129)

Entre as posicoes 40 e 60: (0.5892004713150517, 0.5981854192748449)

Abaixo do top90: (0.5767036895130537, 0.5899340972222219)

Pelo gráfico, músicas com uma posição melhor no ranking top100 da Billboard, tendem a ter uma *danceability* maior.



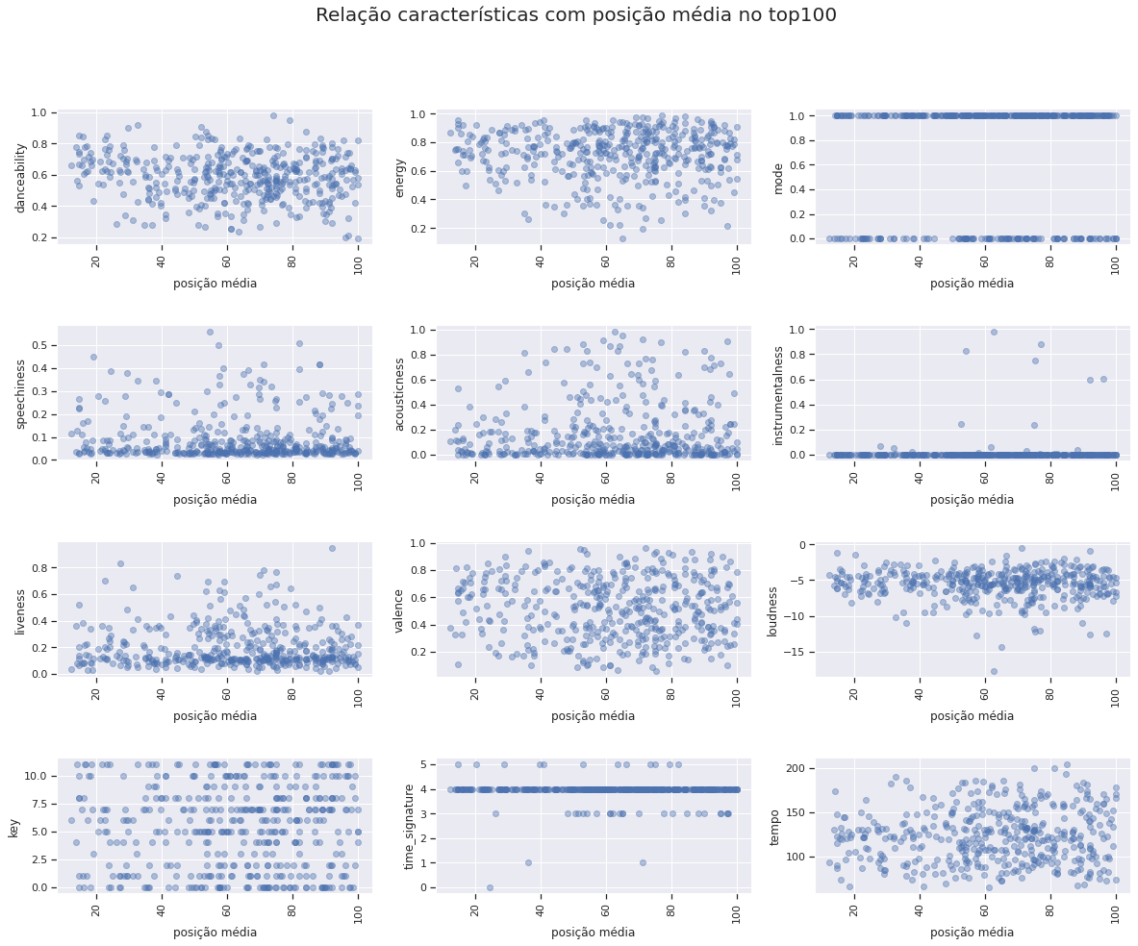
Pelos gráficos de dispersão, pode-se inferir que:

* músicas com *danceability* muito alta ou muito baixa tendem a não permanecer muito tempo no rank

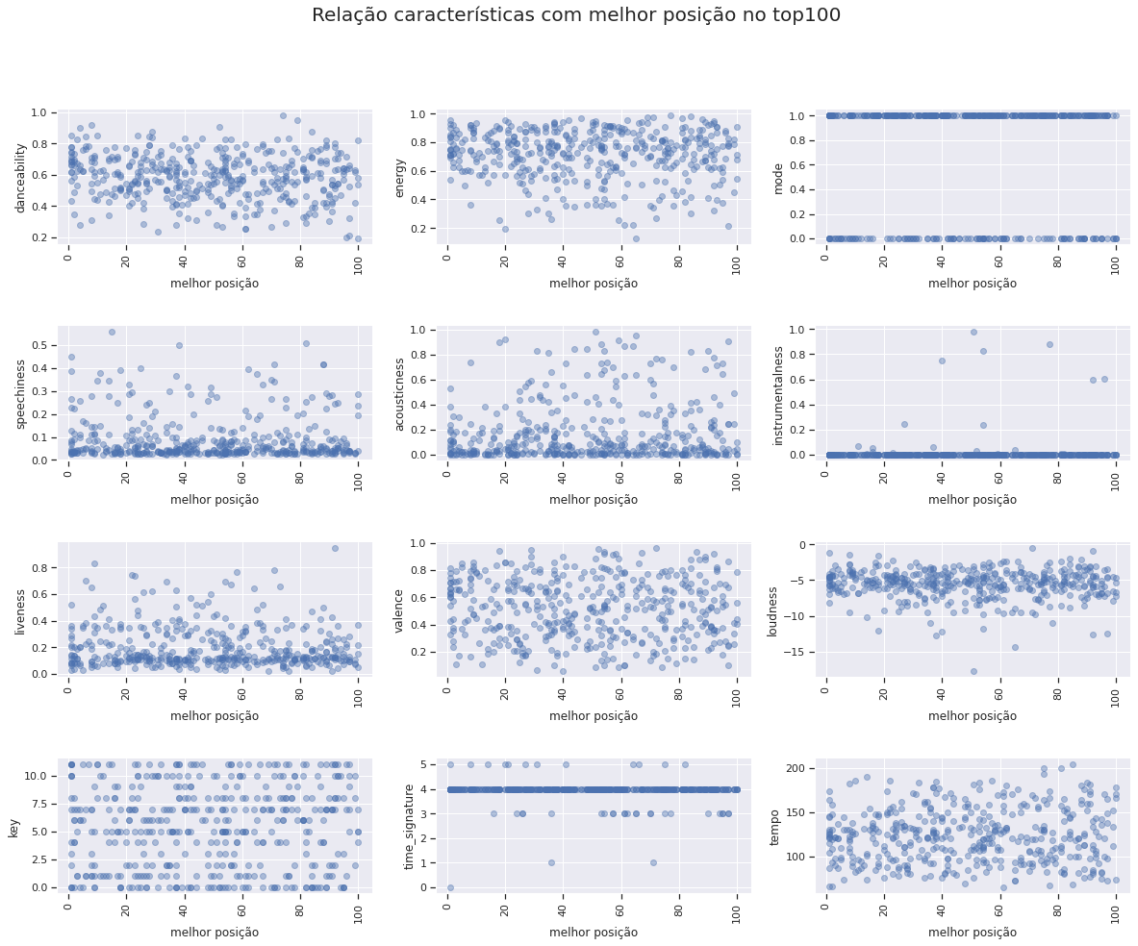
* músicas com *energy* baixa tendem a não permanecer muito tempo no rank

* músicas com alta *speechiness* tendem a não permanecer muito tempo no rank

* músicas com alta *acoustiness* tendem a não permanecer muito tempo no rank



Pelos gráficos de dispersão, não é possível inferir alguma relação entre as características das músicas e sua posição média no rank da Billboard.



Pelos gráficos de dispersão, não é possivel inferir alguma relação entre as características das músicas e sua melhor posição no rank da Billboard.

3.4. Caracterização

3.5. Previsão classificatória

A primeira previsão foi uma tentativa de classificar as músicas numa posição do chart tendo como variáveis independentes as *features* do Spotify. Para isso, utilizamos a abordagem de K-Nearest-Neighbors: músicas com características parecidas talvez ocupassem posições parecidas.

O K mais adequado foi encontrado pela função GridSearchCV e, então, o modelo de previsão foi treinado com 75% dos dados. Os resultados demonstraram que há algum poder de previsão nesses dados, mas que ele não é tão efetivo: as métricas de precisão, revocação, acurácia e F1 ficaram em 0.48.

Intuitivamente, até faz sentido que não seja uma previsão muito boa, pois parece haver músicas de vertentes muito distintas nos charts. Ainda mais quando se considera os dados de vários anos, que englobam muitas mudanças sociais e culturais: os gostos mudam, as formas de ouvir música mudam, as formas de fazer música mudam.

A princípio, imaginamos que a coluna ¡i¿spotify_track_popularity¡/i¿ conseguiria exercer grande influência positiva sobre os resultados, por já ser, em si, uma forma de

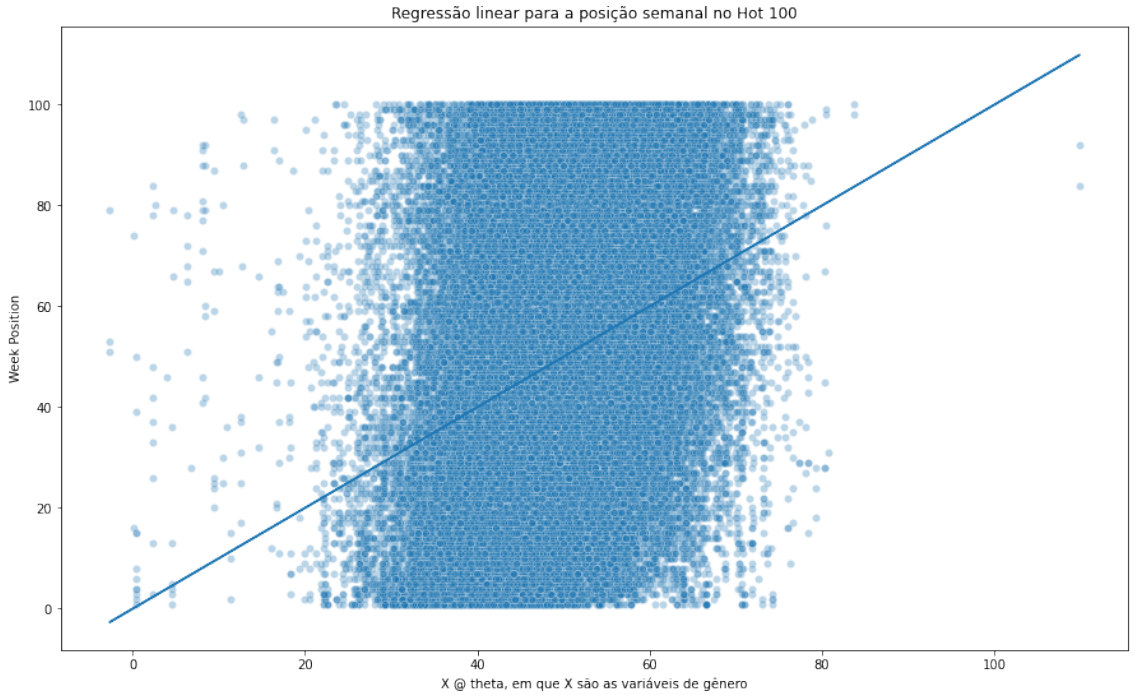
chartear as músicas, mesmo que reflita a popularidade em apenas uma plataforma. Entretanto, isso não se mostrou verdadeiro.

3.6. Previsão numérica

A métrica *tempo* captura a quantidade de batidas por minuto (BPM) numa música: um tempo maior indica uma música mais rápida. Assim, é natural conjecturar se existe um tempo característico de cada gênero. Isto é, se, sabendo o gênero de uma música, conseguiríamos estimar seu tempo.

O primeiro problema surgiu na representação dos gêneros, fornecidos como uma única string para cada música, mas felizmente separados por vírgula. A solução foi quebrar essas strings e transformar cada gênero numa variável categórica. Assim, uma música com os gêneros *"[pop, dance pop]"* passaria a ter o valor 1 nas novas colunas *pop* e *dance pop*.

Resolvidos os problemas de gênero e retiradas as linhas com valores nulos, o próximo passo era encontrar os coeficientes θ. Esses parâmetros refletiriam o peso de cada gênero na determinação do tempo. Para isso, foi utilizada a função *LinearRegression* da biblioteca *scikit-learn* e o modelo gerado foi treinado com 75% dos dados. Um plot desses dados segue abaixo:



O eixo-x é a multiplicação da matriz com as colunas de gêneros pelos parâmetros θ encontrados pela regressão linear. E, como o eixo-y é a posição da música no *Hot 100*, quanto menor, melhor.

4. Conclusão

As conclusões a respeito de cada hipótese foram discutidas ao longo do documento. No geral, as músicas que ocupam a base de dados são muito diversas. Pode haver certa dominância de uma ou outra característica, mas não de forma gritante.