**Análise do grau de “danceability” musical usando regressão logística no Spotify**

*Carla Kim Gaieski, Giulia Braga Passarelli, Rebeca Moreno dos Santos*

**Introdução**

Imagine que você é um músico e quer compor uma música, porém quer que ela seja calma, qual a probabilidade de você usar o tom menor e qual nota provavelmente usaria? O objetivo desse projeto é responder esta pergunta prevendo qual seria a decisão que você tomaria nesta situação. Assim, temos que comparar os tons e as notas com o quão dançável as músicas são.

Nesse caso, o método escolhido foi a regressão pois é uma das ferramentas mais utilizadas para investigar a relação entre variáveis na estatística. Ela opera equacionando a variação de uma variável dependente por conta da alteração de uma independente. Para isso, há dois modelos que são mais comuns: a regressão linear e a regressão logística. A primeira é considerada quando a variável de saída é quantitativa, pois o gráfico apresenta um comportamento mais linear sendo, geralmente, mais fácil de encaixar uma função em cima. Já o segundo é próprio para a análise de uma variável de saída categórica, que não se porta da mesma forma que o último. Como queremos analisar as notas e os tons musicais (variáveis qualitativas), elas devem ser a saída em uma regressão logística.

A regressão logística possui uma função para a probabilidade de um ponto que pode ser definida como: .

**Procedimento:**

O Spotify é um serviço de música digital com um acervo de milhões de músicas, de forma a permitir o acesso, a qualquer momento, de uma coleção imensa de canções. Por conta disso, o projeto utiliza a base de dados do Spotify para realizar essa análise. Coletamos uma amostra contendo 500 músicas por meio do dataset “Spotify Song Attributes” do site kaggle.com, utilizando as colunas: “danceability” (valores contínuos de 0 a 1), “key” (notas- valores discretos de 0 a 11, sendo 0 a nota C e 11 a nota B -temos do inglês) e “mode” (tom- valor binário-, 0 -tom menor- ou 1 -tom maior).

Podemos relacionar tom e notas musicais para realizar a análise pois cada nota possui uma tonalidade correspondente que retorna um certo grau de “danceability”. E, assim, como estamos considerando a “danceability” das músicas, as duas variáveis possuem uma dependência nesse aspecto. Para realizar essa união de variáveis, criamos uma nova coluna no dataframe que calcula: . O motivo para multiplicarmos o tom por 12 é para ele possuir um intervalo que se assemelhe ao das notas.

Abaixo está uma tabela com a frequência de cada nota para diferentes tonalidades (tom maior ou tom menor):

Uma imagem contendo captura de tela

Descrição gerada com muito alta confiança

Quando geramos um gráfico que relaciona a nova coluna criada no dataframe com a coluna de “danceability”, obtemos:

Uma imagem contendo captura de tela

Descrição gerada com muito alta confiança

Para realizar uma regressão logística, o ideal (pois possui maior probabilidade de acerto) seria possuir uma variável dependente binária. Assim, como essa nova variável é categórica, podemos agrupar os valores que estão acima ou abaixo da média em 0 ou 1, respectivamente, para facilitar a regressão. Assim, geramos o gráfico a seguir:

Uma imagem contendo captura de tela

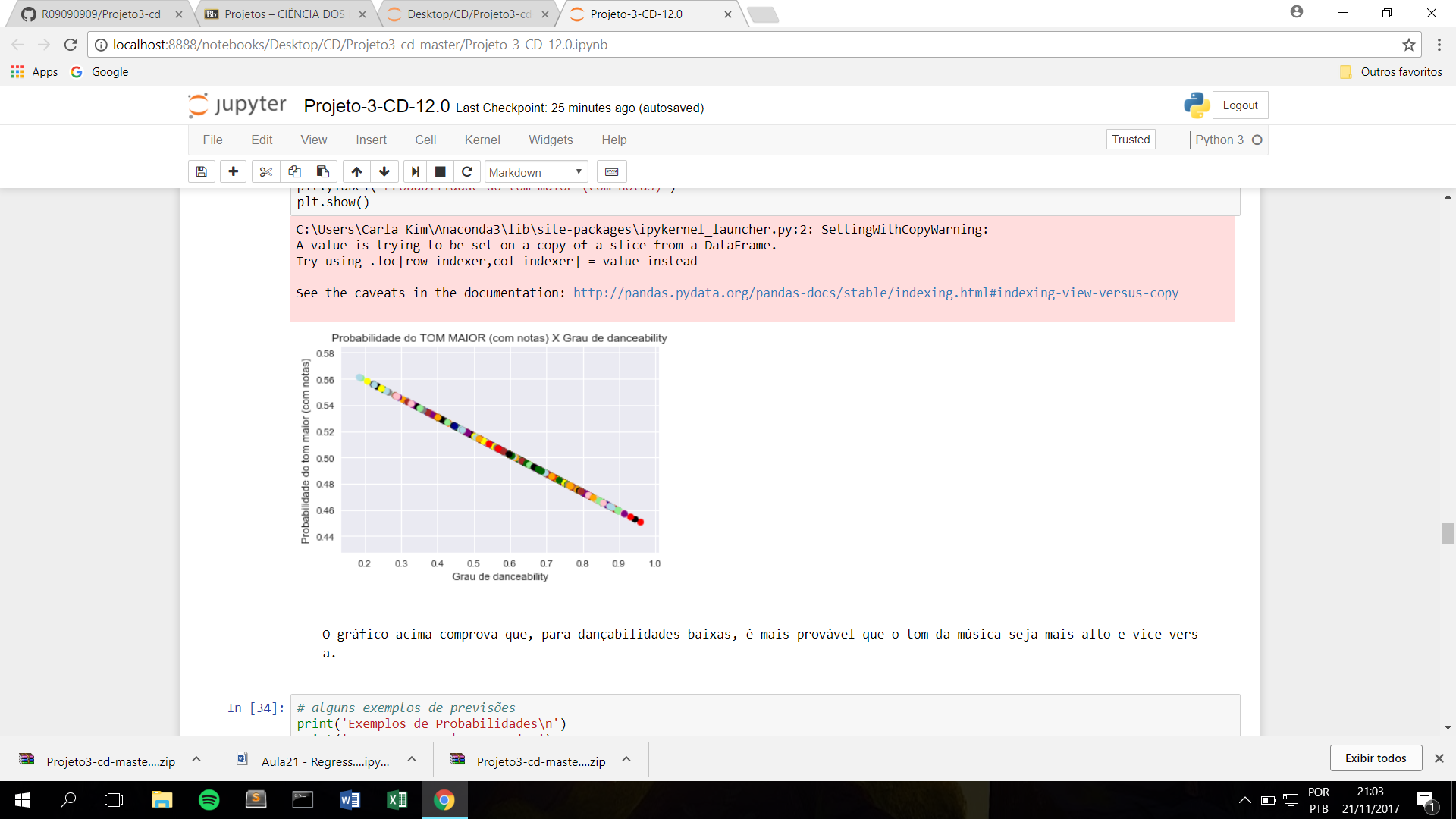
Descrição gerada com muito alta confiança­

Por conta do agrupamento que realizamos, muitos dos dados são perdidos, especialmente das notas, já que o tom da música é um valor binário. Assim, a regressão é obtida a partir dos pontos de cada nota individualmente.

Dessa forma, ao agruparmos







Criamos um novo dataframe, contendo todas as variáveis que utilizaríamos para realizar a regressão logística e posteriormente, na linear. As variáveis eram: 'mode' , 'danceability', 'energynaobin', 'loudnesS', 'livenesS', ou seja, tom, o quão dançável uma música é, energia de uma música descrita de maneira não binária, o quão alto e alegre uma música pode ser, respectivamente. Assim, poderíamos descobrir a tonalidade de uma música baseado nos outros dados. O resultado obtido no valor P nos informaria se a primeira hipótese, ou seja, de que o valor de beta era igual a zero, poderia ser rejeitada ou não.

No primeiro caso, regressão logística, obtivemos um valor P, em P>|z|

