PROJETO DE CIÊNCIA DE DADOS: STREAMING DE MÚSICA

Nome: *Gustavo Navarro Felix*, Email: *gustavo.sempre.estude@gmail.com*

Introdução

Este projeto é uma iniciativa pessoal com o objetivo de praticar e aprofundar meus conhecimentos em ciência de dados.

Utilizei um conjunto de dados de atividades de streaming de música para este projeto, com o objetivo de passar por todas as etapas de um processo de análise de dados, desde a limpeza e preparação dos dados até a extração de insights valiosos.

As tecnologias que utilizei para desenvolver este projeto incluem:

Python 3 - Welcome to Python.org
Linguagem R - R: The R Project for Statistical Computing (r-project.org)
RStudio - RStudio IDE - RStudio
Anaconda - Anaconda | The World's Most Popular Data Science Platform
Jupyter Notebook - Project Jupyter | Home
Power Bi - Power BI - Visualização de Dados | Microsoft Power Platform

O código e arquivos relacionados ao projeto estão disponíveis no GitHub:
Fonte do dataset: Streaming Activity Dataset (kaggle.com)

Objetivo

O objetivo deste projeto é extrair valor significativo do conjunto de dados de atividades de streaming de música, realizando uma análise detalhada para descobrir padrões de comportamento do usuário e tendências de popularidade da música.

Análise exploratória

Iniciei a análise exploratória dos dados carregando-os no Microsoft Power BI Desktop. Com o auxílio da documentação do conjunto de dados, realizei uma análise preliminar.
O conjunto de dados é composto por duas tabelas principais: 'My Streaming Activity', que contém informações principais sobre a música, e 'Scrobble Features', que traz as

informações técnicas da música. A relação 1x1 entre essas duas tabelas pode ser visualizada por meio da exibição de modelo do Power BI.

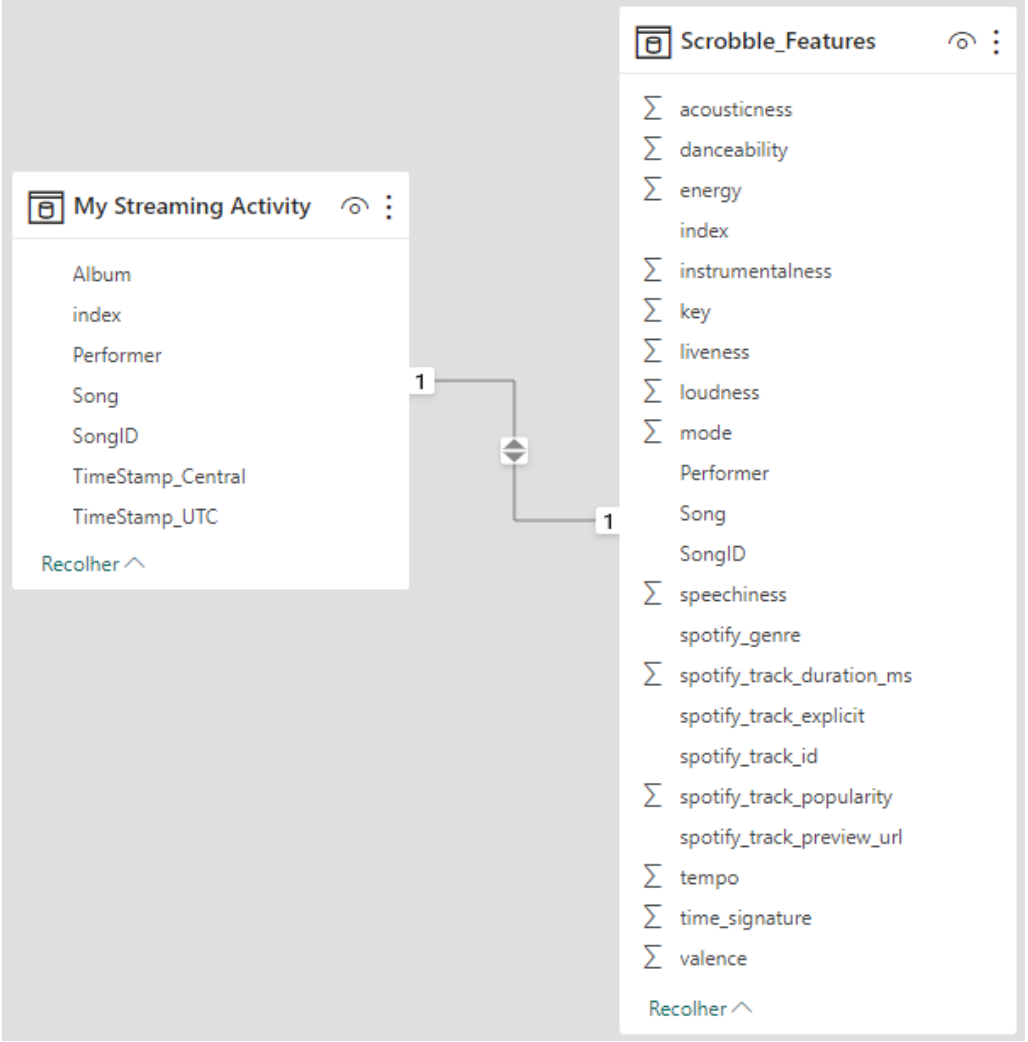

Figura 1: Exibição do Modelo

Para aprofundar a exploração dos dados, utilizei o Jupyter Notebook e a biblioteca pandas para leitura e análise dos dados. Verifiquei o número de observações para cada tabela, sendo 62907 para 'My Streaming Activity' e 13396 para 'Scrobble Features'.

Limpeza e Transformação

Após a análise exploratória, identifiquei alguns problemas nos conjuntos de dados que exigiam limpeza e transformação.

Utilizando o Jupyter Notebook, procurei por valores duplicados e encontrei alguns na coluna 'SongID' da tabela 'Musics'. No entanto, após uma análise cuidadosa, concluí que esses

valores duplicados são esperados, pois a tabela não contém apenas lançamentos únicos. Portanto, decidi manter a coluna inalterada.

Ao analisar os dados no Power Query, notei que algumas colunas continham valores nulos. Em muitos casos, várias colunas da mesma observação estavam nulas. Para investigar mais a fundo, levei os dados para o Jupyter Notebook e realizei uma contagem de valores nulos para todas as linhas. Descobri que uma parte considerável (11%) das observações estava comprometida. Esses dados são essenciais para a geração de insights. Portanto, após uma análise cuidadosa, decidi eliminar essas observações. Apesar de representarem uma fatia considerável dos dados, a falta de informações nessas linhas não pode ser simplesmente substituída. Acredito que a ausência dessas informações se deve ao fato de as músicas serem muito recentes e ainda não terem tido os dados coletados. Para garantir insights precisos e confiáveis, optei por eliminar esses valores nulos.

Além disso, o Power BI não conseguiu identificar corretamente a coluna de data. Para corrigir esse problema, alterei o tipo de dado com base na localidade, permitindo a formação de uma hierarquia por meio do Power Query. Também substituí os valores booleanos da coluna de conteúdo explícito por valores de texto.

Análise e Insights

O projeto em Power Bi está disponível no github para download.
Abaixo estão todos os dashboards e gráficos criados para análise do conjunto de dados.

Eu dividi o Dashboard em 3 Partes, essas que podem ser selecionadas pelo índice:

ANÁLISE DE DADOS
DADOS DE STREAMING DE VIDEO

ÍNDICE

Visão Geral

Relação de Duração

Relação de Conteúdo Explícito

Figura 2 - Índice

Figura 3 - Visão geral do conjunto de dados

Para o segundo dashboard, eu criei um nova coluna em Novofeatures, nomeada de “TempoMinutos”, e optei por fazer uma análise com Boxplot na linguagem R, o código está disponível no github. O resultado obtido segue abaixo.

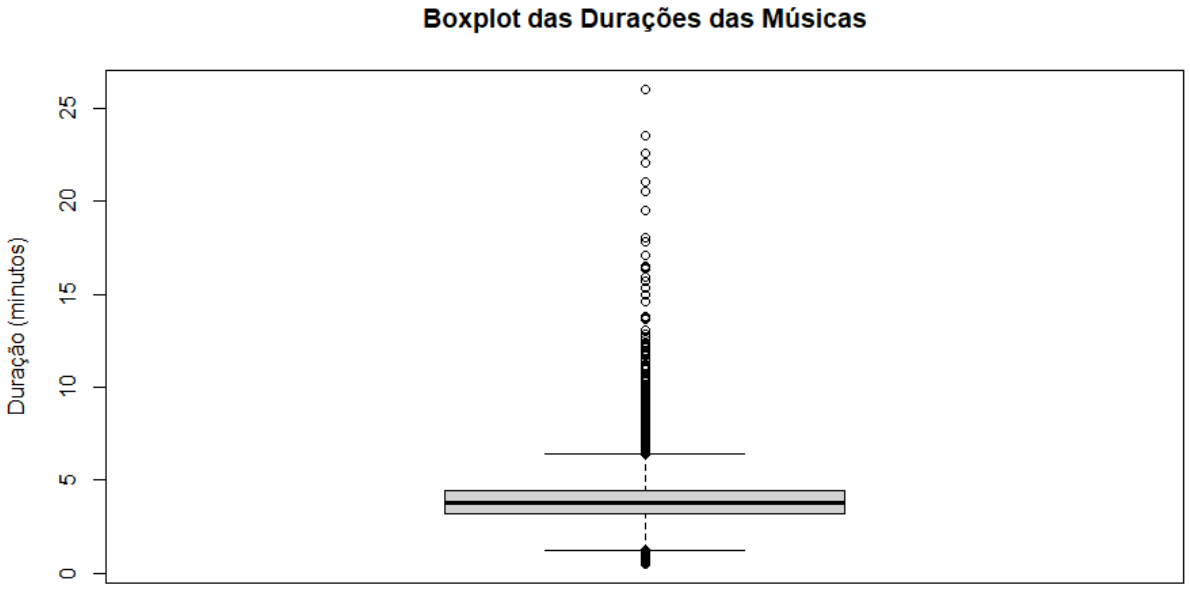

Figura 4 - Boxplot de duração das Músicas

Após realizar a análise, podemos perceber que há um grande número de valores consideráveis outliers. Entretanto, todos esses valores parecem válidos, sem qualquer tipo de anomalia, então eu escolhi por mantê-los.

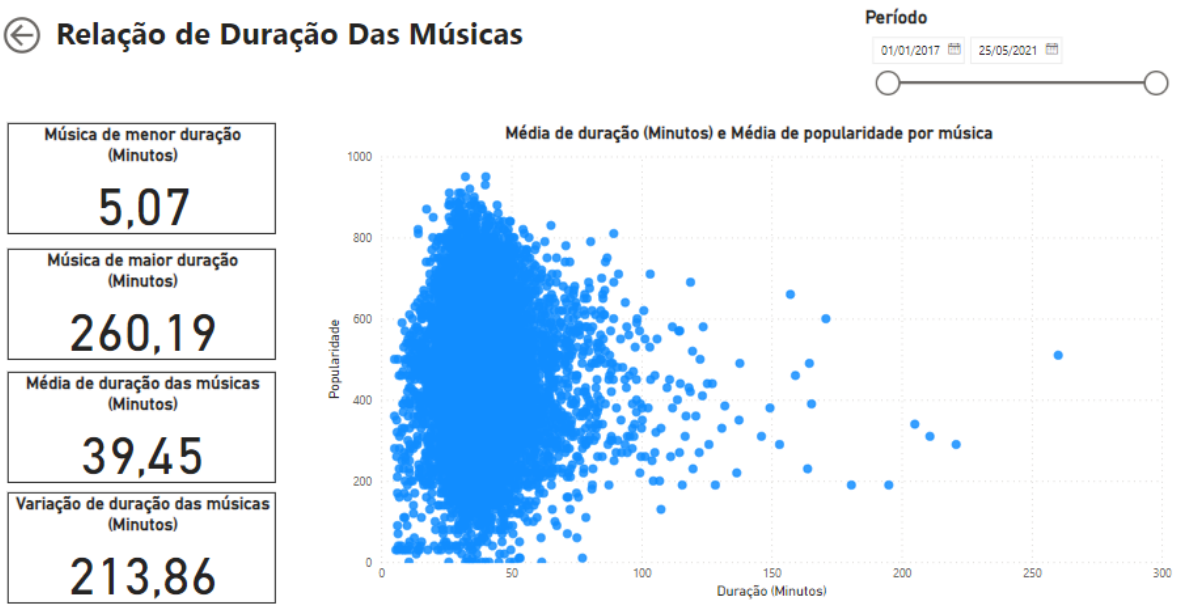

Figura 5 - Relação de Duração Das Músicas

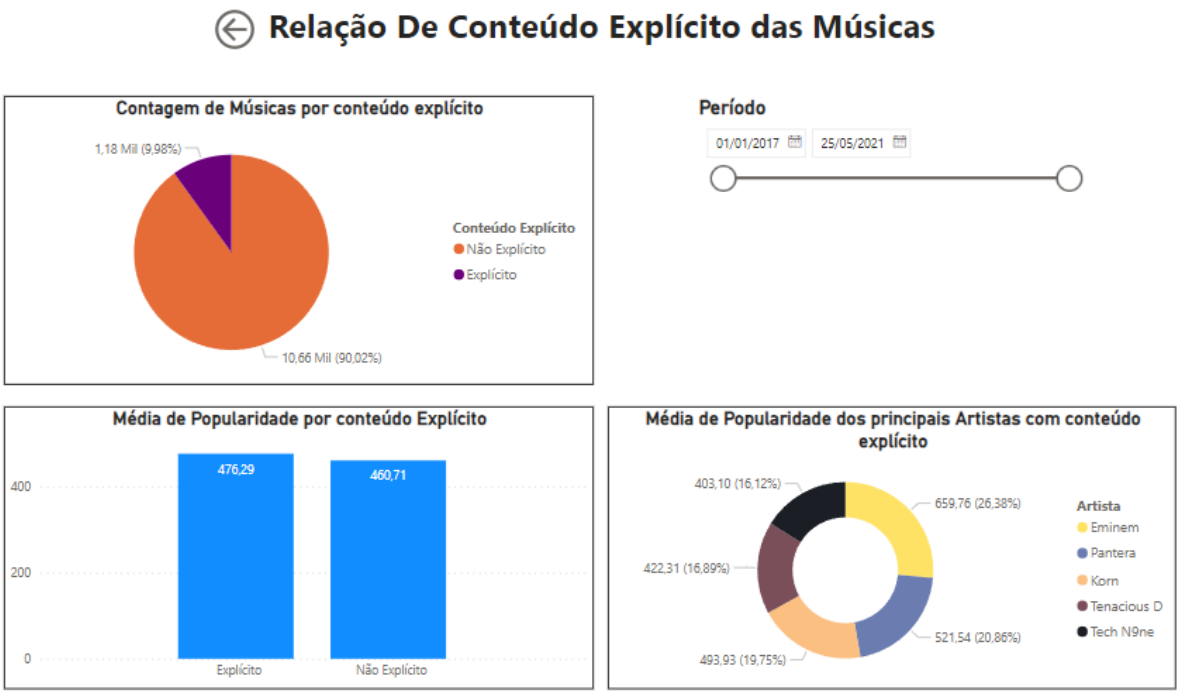

Figura 6 - Relação de Conteúdo Explícito