

# Análise Estatística de Dados do Spotify utilizando Mineração de Dados com Python

Levi Ferreira de O. Baptista<sup>1</sup>, Rafael Jesus B. Cerqueira<sup>2</sup>

<sup>1</sup> Centro Universitário - Centro Universitário de Excelência (UNEX)  
Feira de Santana – BA – Brazil

{Levi, Rafael}24103588@aluno.unex.edu.br, 242030195@aluno.unex.edu.br

**Abstract.** *This report describes the implementation and application of a statistical library developed in Python to analyze a dataset of over 8,700 Spotify tracks. The study focuses on descriptive statistics, such as mean, median, and quartiles, while addressing technical challenges in Python environment configuration and file handling.*

**Resumo.** *Este relatório descreve a implementação e aplicação de uma biblioteca estatística desenvolvida em Python para analisar um conjunto de dados com mais de 8.700 faixas do Spotify. O estudo foca em estatística descritiva, como média, mediana e quartis, abordando desafios técnicos de configuração de ambiente Python e manipulação de arquivos.*

## 1. Informações Gerais

O presente trabalho foi desenvolvido seguindo os padrões de formatação da SBC, utilizando folha A4 com coluna única e margens laterais de 3,0 cm. O objetivo principal é a aplicação prática de conceitos de estatística computacional sobre dados reais extraídos da plataforma Spotify.

No projeto não poderia ser utilizado bibliotecas externas, somente as funções nativas da linguagem, utilizamos uma classe com os métodos que implementamos para usá-la no estudo.

## 2. Fundamentação Matemática

A análise exploratória feita no dataset do Spotify, baseia-se em cinco medidas estatísticas para descrever o comportamento dos dados das faixas musicais. Seguindo as normas de precisão decimal e formatação de fórmulas:

### 2.1. Média Aritmética

Representa o valor médio do conjunto de dados. É calculada pela soma de todos os valores dividida pela quantidade total de elementos ( $n$ ).

## 2.2. Mediana

É o valor que separa a metade superior da metade inferior de um conjunto de dados ordenado.

## 2.3. Moda

Representa o valor (ou valores) com a maior contagem de ocorrências no conjunto.

## 2.4. Frequência Absoluta

É a contagem do número de vezes que cada valor individual aparece no dataset.

## 2.5. Quartis (Q1, Q2, Q3)

Dividem os dados ordenados em quatro partes iguais. O cálculo segue o método de fatiamento:

- Q1: Mediana da primeira metade dos dados. (25%)
- Q2: Mediana de todo o conjunto de dados. (50%)
- Q3: Mediana da segunda metade dos dados. (75%)

## 3. Comportamento do Python na Manipulação de Arquivos

Um dos desafios técnicos encontrados foi o erro **FileNotFoundError**. No Python, o comportamento padrão ao buscar um arquivo com **open()** é procurar no **diretório de trabalho atual (CWD)**.

Se o terminal estiver aberto em **C:\Projeto\** e o arquivo CSV estiver em **C:\Projeto\dados\**, a chamada **open('arquivo.csv')** falhará. É necessário fornecer o caminho relativo correto ou o caminho absoluto utilizando prefixos de *raw string* (ex: **r'C:\caminho\arquivo.csv'**) para evitar problemas com caracteres de escape no Windows.

.

#### 4. Resultados

Nas tabelas a seguir estão representados os resultados conseguidos em relação a popularidade das músicas feitas através dos métodos anteriormente citados

| Indicador       | Valor | Descrição                                   |
|-----------------|-------|---|
| Média           | 52,36 | Valor médio da popularidade das faixas.     |
| Mediana (Q2)    | 58    | O ponto central dos dados (50% das faixas). |
| Moda            | 0     | O valor que mais se repete no conjunto.     |
| 1º Quartil (Q1) | 39    | Limite para os 25% de menor popularidade.   |
| 3º Quartil (Q3) | 71    | Limite para os 25% de maior popularidade.   |

**Padronização de Decimais:** Todos os valores foram reportados com duas casas decimais, conforme a norma de não utilizar mais dígitos do que a precisão permite.

**A Moda:** Embora a média seja 52, o "valor típico" de muitas músicas novas no dataset é zero, indicando uma barreira de entrada no sucesso da plataforma.

**Amplitude Interquartil:** 50% das suas músicas estão concentradas entre 39 e 71 pontos de popularidade.

## **5. Referências**

**Bussab, W. O. e Morettin, P. A. (2017)**, Estatística Básica, 9ª Edição, Saraiva.

**SBC (2024)**, Instructions for Authors of SBC Conferences Papers and Abstracts, Sociedade Brasileira de Computação.

**Python Documentation**, The Python Standard Library - Statistics.