

# **MINERAÇÃO DE DADOS**

**Autor:** Levi Ferreira de Oliveira Baptista  
Rafael Jesus Bezerra Cerqueira

# TÓPICOS

O objetivo desta apresentação é explicar de maneira clara e simples nosso entendimento e raciocínio para desenvolver o trabalho

## **1. Fundamentos :**

A base para a criação do algoritmo

## **3. Métodos:**

Abordamos os 3 métodos principais usados para fazer as estatísticas

## **2. Filtragem:**

A discriminação de dados por valor ou tipo

## **4. Resultado:**

A demonstração empírica dos tópicos anteriores

# Fundamentos

- Funções nativas da linguagem

`type()``sum()``min()``len()``sorted()``max()`

# Fundamentos

- Fórmulas

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$$S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

**Primeiro Quartil ( $Q_1$ ):**  $i = \frac{1(n+1)}{4}$

**Segundo Quartil ( $Q_2$  ou Mediana):**  $i = \frac{2(n+1)}{4}$

**Terceiro Quartil ( $Q_3$ ):**  $i = \frac{3(n+1)}{4}$

# Filtragem

- A filtragem de dados ocorreu tanto por números quanto por tipos:
  - Moda
  - Mediana
  - Quartis

# Métodos

- Os 3 principais métodos usados no dataset do Spotify
  - Média
  - Frequência Absoluta
  - Mediana

# Métodos

Média

```
dados = self.dataset[column]
QuantantidaDados = len(dados)
Soma = sum(dados)

Media = Soma/QuantantidaDados

return Media
```

# Métodos

## Mediana

```
mediana = 0
if isinstance(column, str):
    dados = self.dataset[column]
else:
    dados = column

if type(dados[0]) == str:
    prioridade = {"baixa": 1, "media": 2, "alta": 3}
    dadosOrdenados = sorted(dados, key=prioridade.get)
    mediana = dadosOrdenados[len(dadosOrdenados)//2-1]
    return mediana

else:
    dadosOrdenados = sorted(dados)
    if len(dadosOrdenados)%2 == 0:
        mediana = (dadosOrdenados[(len(dadosOrdenados)//2)-1]+ dadosOrdenados[(len(dadosOrdenados)//2)])//2
        return mediana

    else:
        mediana = dadosOrdenados[len(dadosOrdenados)//2]
        return mediana
```

# Métodos

## Frequência Absoluta

```
dados = self.dataset[column]
frequenciaAbsoluta = {}
for i in dados:
    if i not in frequenciaAbsoluta :
        frequenciaAbsoluta[i] = 0
    frequenciaAbsoluta[i] += 1

return frequenciaAbsoluta
```

# Resultados

## Terminal

```
[ Análise de Frequência: track_popularity ]
Quantidade de músicas com popularidade 0 (Inativas): 503
Quantidade de músicas com popularidade 50 (Média): 107
Quantidade de músicas com popularidade 100 (Hits Máximos): 0
Valor mais frequente (Moda): [0]
```

```
[ Estatísticas de track_duration_min ]
{'Q1': 2.88, 'Q2': 3.445000000000003, 'Q3': 3.99}
Duração Média: 3.49 min
Duração Mediana: 3.45 min
Moda de Duração: [3.67] min
```

# Referências

- <https://www.w3schools.com/python/default.asp>
- <https://www.todamateria.com.br/matematica/estatistica/>