Centro Universitário de Excelência
Sistemas de Informação

ANÁLISE EXPLORATÓRIA DE DADOS EQUIPE LAURO

Autor: Carlos Henrique de Souza Santana Santiago
Gustavo Bezerra Nonato
Hudnei Sued Passos Santana
João Guilherme Gonçalves Pinheiro

Feira de Santana, 2026

Agenda

O objetivo desta apresentação é demonstrar o ciclo de vida da Análise Exploratória de Dados, desde a infraestrutura de código até a geração de insights reais.

1. Item: Tipos de Operações e de Dados:
Resumo: Separação entre dados numéricos e categóricos e suas respectivas operações.

2. Item: Equações e Casos de Uso:
Resumo: A matemática aplicada à Média e à Frequência Relativa.

3. Item: Implementação e Limitações:
Resumo: Lógica em Python nativo e os pontos fracos de cada métrica.

4. Item: . Aplicação no Spotify:
Resumo: Resultados da nossa Análise Exploratória no dataset.

5. Item: Considerações Finais:
Resumo: Conclusões gerais do projeto.

Classificação e Tipos de Dados

- Tipos de Dados no Dataset (Spotify):
  - Numéricos: Permitem operações matemáticas diretas (Ex: track_duration_min, artist_followers).
  - Categóricos: Representam qualidades, classes ou agrupamentos (Ex: album_type, explicit).
- Operações para Dados Numéricos:
  - Tendência Central: Média e Mediana.
  - Dispersão e Posição: Variância, Desvio Padrão e Quartis.
  - Requisito: Exigem validação estrita de tipo (método _is_numeric).
- Operações para Dados Categóricos:
  - Frequências: Absoluta, Relativa e Acumulada.
  - Análise de Conjunto: Moda, Valores Únicos (Itemset) e Probabilidade Condicional.
  - Foco: Baseiam-se em contagem, agrupamento e ocorrência.

Equações e Exemplos de Uso

Operação Numérica (Média):

- Equação:

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

- Exemplo de Uso: Calcular a duração média das faixas musicais do Spotify para entender o padrão de consumo.

Operação Categórica (Frequência Relativa):

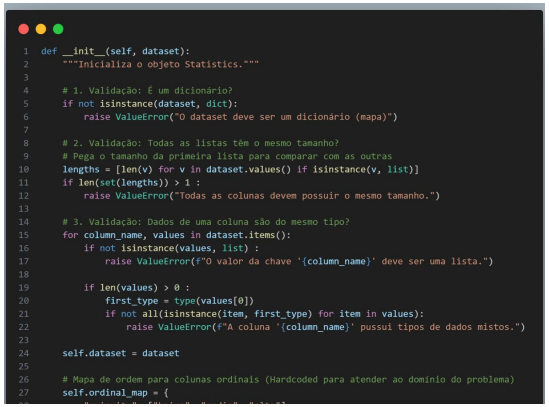
- Equação:

$$f_r = \frac{f_i}{n}$$

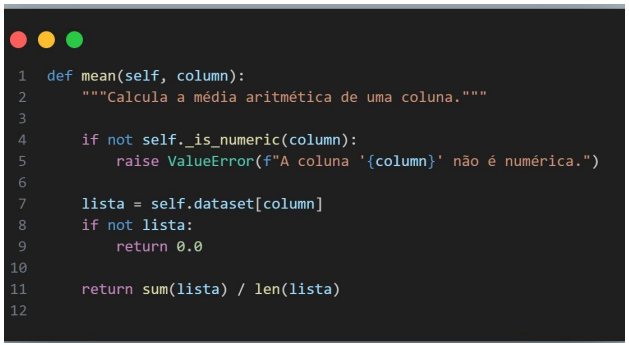
- Exemplo de Uso: Descobrir a porcentagem exata de músicas que possuem conteúdo explícito (explicit = True) no dataset.

Implementação e Limitações

- A Implementação



```python
def __init__(self, dataset):
    """Inicializa o objeto Statistics."""

    # 1. Validação: É um dicionário?
    if not isinstance(dataset, dict):
        raise ValueError("O dataset deve ser um dicionário (mapa)")

    # 2. Validação: Todas as listas têm o mesmo tamanho?
    # Pega o tamanho da primeira lista para comparar com as outras
    lengths = [len(v) for v in dataset.values() if isinstance(v, list)]
    if len(set(lengths)) > 1 :
        raise ValueError("Todas as colunas devem possuir o mesmo tamanho.")

    # 3. Validação: Dados de uma coluna são do mesmo tipo?
    for column_name, values in dataset.items():
        if not isinstance(values, list) :
            raise ValueError(f"O valor da chave '{column_name}' deve ser uma lista.")

        if len(values) > 0 :
            first_type = type(values[0])
            if not all(isinstance(item, first_type) for item in values):
                raise ValueError(f"A coluna '{column_name}' pussui tipos de dados mistos.")

    self.dataset = dataset

    # Mapa de ordem para colunas ordinais (Hardcoded para atender ao domínio do problema)
    self.ordinal_map = {
```



```python
def mean(self, column):
    """Calcula a média aritmética de uma coluna."""

    if not self._is_numeric(column):
        raise ValueError(f"A coluna '{column}' não é numérica.")

    lista = self.dataset[column]
    if not lista:
        return 0.0

    return sum(lista) / len(lista)
```

- As Limitações:
- Limitação da Média: Altamente sensível a outliers. Uma única música de 40 minutos puxa toda a média do dataset para cima, distorcendo a realidade.
- Limitação da Moda: Em datasets não tratados, pode retornar ausência de dados (como 'N/A') ou ser bimodal (empate de frequências), exigindo tratamento prévio.

Aplicação da Operação no Dataset Spotify

- O Cenário:
- Preencher com: Script main.py validou e processou 8.582 linhas com sucesso.
- Resultados de Destaque (Cole os dados do terminal aqui):
- Preencher com: Duração Média (3.49 min).
- Preencher com: Moda dos Gêneros Musicais ('N/A' - Evidenciando dados faltantes).
- Preencher com: Alta dispersão na quantidade de seguidores (Desvio Padrão).

```
Carregando dados do Spotify...

Carregando dados de: src/data/spotify_data_clean.csv
Leitura concluída! 8582 linhas processadas.

--- Análise Exploratória: ---

1. TENDÊNCIA CENTRAL
Média de duração: 3.49 min
Mediana de duração: 3.45 min
Moda (Tipo de Álbum): ['album']

2. DISPERSÃO E POSIÇÃO
Desvio Padrão (Seguidores): 38029589.10
Quartis de Popularidade: {'Q1': 39.0, 'Q2': 58.0, 'Q3': 71.0}



3. FREQUÊNCIAS E CONJUNTOS
Tipos únicos (Itemset Explícito): {'TRUE', 'FALSE'}
Freq. Absoluta (Explícito): {'TRUE': 2148, 'FALSE': 6434}
Freq. Relativa (Explícito): {'TRUE': 0.25029130738755534, 'FALSE': 0.7497086926124447

4. RELAÇÕES E AVANÇADOS
Covariância (Pop. Faixa vs Artista): 218.43
Prob. Condicional (Explícito -> Explícito): 0.41945996275605213
Histograma de Duração (5 bins): {(0.07, 2.7579999999999996): 1793, (2.757999999999999
```

Considerações finais

- Estatística aplicada à análise de dados:
  - O desenvolvimento do zero (sem Pandas/Numpy) consolidou o entendimento da base algorítmica por trás das grandes ferramentas de Mineração de Dados.
- Desafios:
  - Garantir a consistência dos dicionários e evitar quebra de sistema por falha de tipagem.

Referências

- Documentação Oficial do Python.
- Dataset original "Spotify Songs for ML & Analysis".