Desenvolvimento de uma Biblioteca para Análise Exploratória de Dados Utilizando Python

Bruno Sampaio[1], Levi Falcão[2], Móises Souza[1], Raphael Moreira[3]

[1]Instituto de Ensino – Universidade de Excelência (Unex)
Caixa Postal S/N – 44085-370 – Feira de Santana – BA – Brazil

{bruno.silva, levi.queiroz, moises.olveira, raphael.moreira}@aluno.unex.edu.br

Abstract. *This article presents the development of the dende_statistics library in Python, with a manual implementation of descriptive and relational statistical metrics using exclusively native language resources, without the use of external libraries. The system validates the dataset structure as a dictionary of lists, ensuring that each key represents a column and its values are stored as lists, and enables the calculation of measures such as mean, median, variance, and frequency distributions. The results obtained demonstrate mathematical consistency and alignment with the theoretical principles of Statistics.*

Resumo. *Este artigo apresenta o desenvolvimento da biblioteca dende_statistics em Python, com implementação manual de métricas estatísticas descritivas e relacionais utilizando exclusivamente recursos nativos da linguagem, sem o uso de bibliotecas externas. O sistema valida a estrutura do dataset como um dicionário de listas, garantindo que cada chave represente uma coluna e que seus valores estejam armazenados em listas, permitindo o cálculo de medidas como média, mediana, variância e distribuições de frequência. Os resultados obtidos demonstram consistência matemática e alinhamento com os princípios teóricos da Estatística.*

1. Introdução

A crescente geração de dados pela humanidade exige métodos eficientes para selecionar informações relevantes e transformá-las em conhecimento útil para a tomada de decisões em empresas e diversos setores da sociedade. Nesse contexto, a Análise Exploratória de Dados (AED) atua como uma etapa inicial na identificação de padrões, tendências e relações em conjuntos de dados.

Neste trabalho, foi desenvolvida a biblioteca *dende_statistics*, voltada à implementação de métricas estatísticas descritivas e relacionais aplicadas a um dataset real do Spotify. As funcionalidades implementadas permitem a análise de variáveis numéricas e categóricas por meio de medidas como média, variância, frequência e covariância, cujos fundamentos teóricos, metodologia de implementação e aplicação prática são apresentados ao longo do relatório.

2. Fundamentação Teórica

A Análise Exploratória de Dados (AED) é uma etapa essencial para resumir características, identificar padrões e compreender o comportamento de um conjunto de informações. Para ilustrar a aplicação prática, utilizaremos o conjunto abaixo:

$$X = \{10, 15, 13, 10, 13, 19, 10\}$$

Media Aritmética

A média aritmética representa uma medida de tendência central, correspondendo ao valor médio de um conjunto de dados numéricos, sendo definida pela seguinte fórmula:

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

No conjunto dado, primeiro deve ser feito a soma de todos os elementos e depois dividido pela quantidade de elementos.

$$\bar{x} = \frac{10 + 10 + 10 + 13 + 13 + 15 + 19}{7} = \frac{90}{7} \approx 12,86$$

Mediana

A mediana é uma medida de tendência central, que representa o número central de uma lista de dados organizados de forma crescente ou decrescente. No conjunto ordenado abaixo, a mediana está em negrito

Conjunto Ordenado: {10, 10, 10, **13**, 13, 15, 19}

Caso o número de elementos seja par, a mediana corresponde à média aritmética dos dois valores centrais.

Moda

A moda é uma **medida de tendência central** que corresponde ao valor que ocorre com maior frequência no conjunto de dados. No conjunto de exemplo, a moda corresponde ao valor 10

Variância

Sendo uma **medida de dispersão**, a variância mede o grau de afastamento dos dados em relação à média, dada pela fórmula:

$$\sigma^2 = \frac{\sum(x - \bar{x})^2}{N}$$

Sabemos que a média do conjunto de exemplo equivale a aproximadamente a 12,86. Aplicando a fórmula, temos:

$$\sigma^2 = \frac{(10 - 12,86)^2 + (15 - 12,86)^2 + (13 - 12,86)^2 + (10 - 12,86)^2 + (13 - 12,86)^2 + (19 - 12,86)^2 + (10 - 12,86)^2}{7}$$

$$\sigma^2 = \frac{8,18 + 4,59 + 0,02 + 8,18 + 0,02 + 37,70 + 8,18}{7} = \frac{66,87}{7} \approx 9,55$$

Desvio padrão

Classificado também como uma **medida de dispersão**, é o resultado obtido pela raiz quadrada da variância.

$$\sigma = \sqrt{\frac{\sum(x - \bar{x})^2}{N}}$$

O valor da variância no conjunto dado foi de aproximadamente 9,55. Obtendo a raiz quadrada desse valor teremos:

$$\sigma = \sqrt{\frac{66,87}{7}} = \sqrt{9,55} \approx 3,09$$

Covariância

A covariância é uma **medida de relação (dispersão conjunta)** que mede o grau de relação linear entre duas variáveis numéricas e é representada pela fórmula:

$$Cov(X,Y) = \frac{\sum(x - \bar{x})(y - \bar{y})}{N}$$

Considere os dois conjuntos:

- X={10,15,13,10,13,19,10}
- Y={1,2,3,1,3,4,1}

Primeiramente, calcula-se as médias:

$$\bar{x} = \frac{90}{7} \approx 12,86$$

$$\bar{y} = \frac{15}{7} \approx 2,14$$

Aplicando na fórmula:

$$\begin{aligned} Cov(X,Y) = \frac{1}{7}[&(10 - 12,86)(1 - 2,14) + (15 - 12,86)(2 - 2,14) + \\ &+ (13 - 12,86)(3 - 2,14) + (10 - 12,86)(1 - 2,14) + \\ &+ (13 - 12,86)(3 - 2,14) + (19 - 12,86)(4 - 2,14) + \\ &+ (10 - 12,86)(1 - 2,14)] \\ Cov(X,Y) = \frac{28,29}{7} &\approx 4,04 \end{aligned}$$

Frequência Absoluta

Classificada como uma **medida de frequência**, é a quantidade de vezes que um valor aparece em um conjunto.

Frequência Relativa

Também sendo uma **medida de frequência**, representa a proporção de vezes que um valor aparece em relação ao total de elementos do conjunto.

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

Frequência Acumulada

Outra **medida de frequência**, que consiste na soma progressiva das frequências absolutas ou relativas ao longo do conjunto.

Valor	f_i	$f_r(\%)$	F_i	$F_r(\%)$
10	3	$42,86$	3	$42,86$
13	2	$28,57$	5	$71,43$
15	1	$14,29$	6	$85,71$
19	1	$14,29$	7	$100,00$
Total	7	100%	$-$	$-$

Quartis

Os quartis são **medidas de posição (distribuição)** representadas por valores que dividem o conjunto de dados ordenado em quatro partes iguais. O Q1 representa os 25% dos dados, o Q2 a mediana (50%) e o Q3 os 75%.

Conjunto Ordenado: {10, **10**, 10, **13**, 13, **15**, 19}

$$Q_1 = 10 \quad | \quad Q_2 = 13 \quad | \quad Q_3 = 15$$

Histogramas por Buckets

Como uma ferramenta de **distribuição de dados**, o histograma organiza os dados em intervalos para mostrar a distribuição visual da frequência.

Considerando o conjunto {10, 15, 13, 10, 13, 19, 10}, a distribuição pode ser organizada em três intervalos:

- Intervalo [10 – 13]: 5 valores (10, 10, 10, 13, 13)
- Intervalo [14 – 17]: 1 valor (15)
- Intervalo [18 – 21]: 1 valor (19)

Probabilidade Condicional

Sendo uma **medida de probabilidade**, a probabilidade condicional representa a probabilidade de ocorrência de um evento **A**, dado que outro evento **B** já ocorreu. É definida pela seguinte expressão:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Utilizando o conjunto de exemplo, e considerando dois eventos distintos:

- Evento A: valor é igual a 13
- Evento B: valor menor ou igual a 13

Então temos o evento B 5 vezes e o evento A 2 vezes

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2}{5} = 0,4$$

3. Metodologia

A classe Statistics foi projetada para receber dados no formato de dicionários (mapas), onde cada chave representa uma coluna e seu valor associado é uma lista. O método construtor (__init__) garante a formatação correta dos dados exigindo três validações: verifica se a entrada é um dicionário, confirma se todas as listas de valores possuem o mesmo tamanho e testa que não há tipos mistos na mesma coluna.

3.1 Validação

Antes de realizar os cálculos, cada método estatístico verifica se a operação é permitida para o tipo de dado da coluna escolhida. Métodos como a média e mediana checam se a variável é numérica (int, float). Além disso, é necessário ter certeza que a coluna recebida pelo método existe no dataset, bem como ela não pode estar vazia

3.2 Implementação

As 13 operações foram desenvolvidas exclusivamente com recursos nativos da linguagem Python, sem o uso de bibliotecas externas de análise de dados, como Pandas ou NumPy. Os cálculos utilizaram métodos padrão do python (if else, for, while) e funções embutidas (len, sorted).

4. Resultados

A biblioteca foi validada em dois níveis: via testes automatizados e por meio do processamento de um volume real de dados (Spotify)

4.1 Teste Automatizado

Para assegurar a integridade matemática, foi utilizado o framework unittest. Foram criados cenários para todas as métricas. O sistema obteve **100% de aproveitamento** em todos os testes passados.

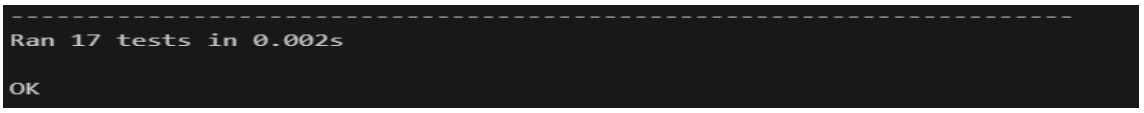

Figura 1. Sucesso no teste

4.2 Código da média

```python
def mean(self, column):
    if column not in self.dataset:              # Verifica se a coluna existe
        return None                             # Evita erro de chave inexistente

    dados = self.dataset[column]                # Obtém a lista de valores
    quantidade = len(dados)                     # Armazena o total de elementos

    if quantidade == 0:                         # Valida se a lista está vazia
        return 0                                # Retorna 0 para evitar divisão por zero

    if not isinstance(dados[0], (int, float)): # Valida se os dados são numéricos
        return None                             # Retorna nulo para colunas categóricas

    soma_total = 0                              # Inicializa acumulador nativo
    for valor in dados:                         # Percorre os valores da coluna
        soma_total += valor                     # Soma cada item ao total
```

Figura 2. Código da média comentado

4.3 Analise do dataset do Spotify

Ao aplicar a biblioteca sobre o dataset do Spotify, obtivemos operações relevantes e corretas sobre o dataset. Dentre elas é possível ver na imagem abaixo, resultados das operações feitas pela biblioteca, todas elas de tendência central.

```
[Running] python -u "c:\Users\Bruno\Downloads\mineracao-de-dados\open.py"
==================== DASHBOARD ====================

[SEÇÃO 01 - TENDÊNCIA CENTRAL]
 > Tempo Médio de Faixa: 3.49 min
 > Mediana de Duração:   3.45 min
 > Perfil de Álbum (Moda): ['album']
```

Figura 3. Uso da biblioteca no dataset

5. Considerações Finais

A construção de uma biblioteca estatística do zero sem *frameworks* externos permitiu consolidar tanto os conceitos matemáticos da Análise Exploratória de Dados quanto o domínio sobre estruturas de dados e controle de fluxo nativos da linguagem Python.

Durante o desenvolvimento, os principais desafios técnicos solucionados foram:

1. Tratamento de dados: lidar com colunas vazias e tipos primitivos.
2. Tipos Mistos: A coluna categórica track_name mistura textos e números.
3. Teste do Código: Identificar erros no teste e corrigi-los

Para trabalhos futuros, os pontos de melhoria incluem a otimização de performance para lidar com *datasets* maiores e implementação de novas métricas estatísticas.

6. Referencias

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Montgomery, D. C.; Runger, G. C. (2014). *Applied Statistics and Probability for Engineers*. 6th ed. Wiley.

Ross, S. M. (2014). *A First Course in Probability*. 9th ed. Pearson.