Análise Exploratória de Dados: Uma Abordagem Estatística Baseada em Python Nativo sobre o Dataset do Spotify

Carlos Henrique S. S. Santiago[1], Gustavo B. Nonato[1], Hudnei S. P. Santana[1], João Guilherme G. Pinheiro[1]

[1]Curso de Sistemas de Informação – Centro Universitário de Excelência (UNEX) Feira de Santana – BA – Brasil

{241031357}@aluno.unex.edu.br, 241030283@aluno.unex.edu.br, 242030118@aluno.unex.edu.br, 241031134@aluno.unex.edu.br}

Resumo. *Este artigo apresenta o desenvolvimento de uma biblioteca estatística em Python nativo e sua aplicação na Análise Exploratória de Dados (EDA) do dataset Spotify Songs. O projeto, desenvolvido como parte da avaliação OAT 1, implementa uma classe orientada a objetos (Statistics) com validações rigorosas de tipagem e integridade, dispensando o uso de bibliotecas de terceiros como Pandas ou NumPy. Os resultados evidenciam o comportamento das variáveis musicais, destacando a importância do tratamento de dados ausentes e a alta dispersão na popularidade dos artistas.*

Abstract. *This paper presents the development of a statistical library in native Python and its application in the Exploratory Data Analysis (EDA) of the Spotify Songs dataset. The project implements an object-oriented class (Statistics) with rigorous typing and integrity validations, avoiding third-party libraries such as Pandas or NumPy. The results highlight the behavior of musical variables, emphasizing the importance of handling missing data and the high dispersion in artist popularity.*

1. Introdução

A mineração de dados é uma etapa fundamental no desenvolvimento de sistemas inteligentes e modelos de Machine Learning. O presente trabalho descreve a concepção e implementação de uma ferramenta de estatística descritiva criada totalmente em linguagem Python nativa. O objetivo principal é realizar uma Análise Exploratória de Dados sobre o conjunto *Spotify Songs for ML & Analysis*, que contém mais de 8.500 registros. A construção manual dos algoritmos permite uma compreensão profunda das fórmulas estatísticas clássicas e dos desafios inerentes ao pré-processamento, como lidar com dicionários de dados mistos e valores inconsistentes.

Full papers must respect the page limits defined by the conference. Conferences that publish just abstracts ask for **one-**page texts.

2. Metodologia e Implementação

A ferramenta foi projetada utilizando o paradigma de Orientação a Objetos através da classe *Statistics*. Para garantir a integridade dos cálculos, o construtor da classe (__init__) foi programado para realizar três validações críticas: verificar se a estrutura de entrada é um mapa (dicionário), assegurar que todas as colunas possuam o mesmo comprimento e validar a consistência de tipos dentro de uma mesma coluna.

Os cálculos foram divididos para atender diferentes naturezas de dados. Para dados numéricos, foram implementadas as medidas de tendência central e dispersão. A fórmula da média populacional, por exemplo, foi codificada representando o somatório dos elementos dividido pelo número total de observações:

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

Para os dados categóricos, o sistema foi desenhado para extrair Frequências (Absoluta e Relativa), Moda e Probabilidade Condicional. A frequência relativa foi calculada através da razão entre a contagem absoluta de um item específico e o total de registros do dataset:

$$fr_i = \frac{f_i}{N}$$

3. Resultados e Discussões (Análise Exploratória do Spotify)

A ferramenta foi executada com sucesso sobre o arquivo spotify_data_clean.csv. O script principal converteu os dados textuais adequadamente, tratando falhas de conversão com valores padrão (default), o que evitou interrupções no processamento.

A análise revelou insights importantes sobre a base de dados:

- **Tendência Central:** A média de duração das faixas musicais globais ficou estabelecida em 3.49 minutos. Esse valor foi corroborado pela mediana, mostrando que a duração possui uma distribuição relativamente simétrica sem a influência extrema de grandes outliers.
- **Identificação de Dados Faltantes:** Ao tentar extrair a Moda da coluna referente aos Gêneros Musicais dos artistas, o algoritmo retornou o valor N/A. Isso não indica um erro no código, mas revela uma limitação do dataset original, que possui uma grande quantidade de *missing values* (dados ausentes) nessa categoria.
- **Dispersão:** A variância e o desvio padrão da quantidade de seguidores dos artistas demonstraram uma dispersão altíssima, o que reflete a realidade da indústria da música: poucos artistas concentram dezenas de milhões de seguidores, enquanto a grande maioria possui uma base muito pequena.

4. Considerações Finais

O desenvolvimento do zero da classe *Statistics* provou-se um excelente exercício arquitetural. A maior limitação encontrada não reside na matemática, mas na sensibilidade de algoritmos simples, como a média, a dados inconsistentes. A criação de métodos de proteção, como a verificação de tipo numérico antes de cada operação, foi vital para a estabilidade da análise. O sistema se mostrou eficiente e pronto para futuras

expansões nas próximas etapas de pré-processamento.

5. Referências

[Python 2026] Python Software Foundation. (2026). The Python Standard Library. Disponível em: https://docs.python.org/3/library/.

[Spotify 2024] Spotify Songs for ML & Analysis Dataset. Kaggle. Recuperado de repositório de dados públicos para estudo acadêmico.