DADOS, PYTHON E ESTATÍSTICA: CAMINHOS PARA STEM

Ana Júlia Amaro Pereira Rocha

SEMANA 2

DIA 1

DADOS NA PRÁTICA

Dados quase nunca vêm prontos. Antes de usá-los, precisamos entendê-los e organizá-los.

Importar dados é trazer informações para serem analisadas. Esses dados podem vir de tabelas, formulários ou arquivos. Antes de qualquer análise, precisamos saber de onde eles vêm.

O que vamos aprender hoje no papel, vamos fazer em Python depois.

DADOS NA PRÁTICA

Todo dado foi coletado por alguém, em algum contexto. A forma de coleta influencia a qualidade do dado. Entender a origem é essencial para interpretar corretamente. Por exemplo, no caso de uma contagem de ovos da dengue em uma cidade, suponhamos que estejam faltando dados curiosamente nos feriados. Isso quer dizer que a mosquita não bota no Natal ou que as pessoas não contaram os ovos nessa data?

Na prática, dados raramente são perfeitos. Eles podem ter erros, valores faltantes ou inconsistências. **Identificar esses problemas é parte do trabalho com dados.**

DADOS NA PRÁTICA

Alguns problemas aparecem com frequência: **valores ausentes**, **erros de digitação** e **formatos misturados**. Esses problemas dificultam análises e conclusões corretas.

valores ausentes: Às vezes, uma informação não foi coletada. Isso pode acontecer por erro ou porque a pessoa não respondeu. Precisamos decidir como lidar com esses casos.

DADOS NA PRÁTICA

Incosistências: Um mesmo tipo de informação pode aparecer de formas diferentes. Por exemplo, “F”, “Fem” e “Feminino”. Isso dificulta contagens e comparações.

Erros de digitação: Dados digitados por pessoas podem conter erros. Um número a mais ou uma letra errada muda o significado. Esses erros precisam ser identificados com cuidado.

Formatos diferentes: Às vezes, números aparecem como texto. Datas podem estar em formatos diferentes. Isso impede cálculos corretos.

LIMPEZA DE DADOS

Limpar dados não significa mudar resultados. Significa tornar os dados coerentes e utilizáveis. **Toda ação precisa de um motivo claro.**

Substituir dados sem critério pode distorcer a realidade. Isso é considerado manipulação de dados. Ciência exige transparência e justificativa.

Ao limpar dados, sempre perguntamos: “Por que estou fazendo isso?”
Essa pergunta protege a análise de erros e vieses.

LIMPEZA DE DADOS

Podemos remover uma linha se o dado for impossível. Podemos padronizar categorias iguais escritas de formas diferentes. Às vezes, removemos até uma coluna inteira por ter pouquíssimos dados. Ou, ainda, adicionamos a média de uma coluna nas linhas em que faltam dados para conseguir fazer uma análise mais coerente. Porém, **sempre explicando o motivo.**

Trabalhar com dados é uma responsabilidade. Dados contam histórias sobre pessoas e realidades. Nosso papel é tratá-los com cuidado. Lembrando que eles são a base para IA também, o que nos faz ter que tomar ainda mais cuidado para não gerar vieses e equívocos.