



CURSO: <BÁSICO EM MACHINE LEARNING>

Nome: Felipe Barbosa

- **Atividade 02 (ATIV-02)**

- Tipo: Somativa;
- Tema: Algoritmos básicos de programação em linguagem python e análise de dados.
- Conteúdo: Módulo 1 e 2.
- Participantes: Individual.
- Avaliação do aluno.
 - Objetivo: Avaliar desempenho do aluno sobre conhecimentos básicos de programação em linguagem python e visualização e análise de dados.
 - Nota: 0 a 3 supercrítico, 4 a 6 crítico, 5 a 7 razoável e 8 a 10 bom;
 - Critérios avaliados: Respostas com coerência, coesão e com exemplos.
- Informações adicionais: A atividade é composta por 5 questões dissertativas sobre python e 5 questões sobre visualização e análise de dados.
- **AO CONCLUIR A ATIVIDADE: ENVIAR APENAS O LINK DO REPOSITÓRIO GITHUB (ESPECIFICAR A BRANCH) PÚBLICO.**

1. Escreva uma função que receba uma lista de números e retorne outra lista com os números ímpares.
2. Escreva uma função que receba uma lista de números e retorne outra lista com os números primos presentes.
3. Escreva uma função que receba duas listas e retorne outra lista com os elementos que estão presentes em apenas uma das listas.
4. Dada uma lista de números inteiros, escreva uma função para encontrar o segundo maior valor na lista.
5. Crie uma função que receba uma lista de tuplas, cada uma contendo o nome e a idade de uma pessoa, e retorne a lista ordenada pelo nome das pessoas em ordem alfabética.
6. Como identificar e tratar outliers em uma coluna numérica usando desvio padrão ou quartis?

Para identificar e tratar outliers em uma coluna numérica, a pessoa pode usar o desvio padrão ou os quartis. A escolha do método depende da distribuição dos dados: o desvio padrão funciona melhor para distribuições normais, enquanto os quartis são mais robustos para distribuições assimétricas.

Identificação de outliers

Método do Z-Score (Desvio Padrão)

O Z-Score indica a quantos desvios padrão um ponto de dados está da média.

1. Calcule a média e o desvio padrão da coluna numérica
2. Para cada valor desvio, calcule o Z-Score: $Z = \frac{x - \mu}{\sigma}$
3. Identificação de outliers: Um valor é considerado outlier quando está muito distante da média normalmente acima de um certo número de desvios padrão.

Método do Intervalo Interquartil (IQR)

Esse método utiliza os quartis da distribuição e não é tão influenciado por valores extremos.

Passos:

Calcule os quartis:

1 passo:

Q1 (25%) -> valor abaixo do qual estão 25% dos dados.



Q3 (75%) → valor abaixo do qual estão 75% dos dados.

2 passo:

Calcule o IQR: $IQR = Q3 - Q1$

3 passo

Defina os limites:

Limite inferior = $Q1 - 1.5 \times IQR$

Limite superior = $Q3 + 1.5 \times IQR$

4 passo: Valores fora desse intervalo são considerados outliers.

Tratamento de outliers

Após identificar, existem diferentes formas de lidar com eles:

- **Remover:** excluir valores extremos (com cuidado para não perder informação importante).
- **Substituir:** usar estatísticas robustas (ex.: mediana).
- **Transformar:** aplicar logaritmo, raiz quadrada ou padronizações.
- **Manter:** se o outlier fizer sentido no contexto (ex.: faturamento muito alto de uma empresa real).

7. Como concatenar vários DataFrames (empilhando linhas ou colunas), mesmo que tenham colunas diferentes?

Dica: Utiliza-se `pd.concat()` especificando `axis=0` (linhas) ou `axis=1` (colunas). Quando há colunas diferentes, os valores ausentes são preenchidos com NaN.

A função `pd.concat()` da biblioteca Pandas pode concatenar vários DataFrames mesmo que eles tenham colunas diferentes, o segredo está em especificar o `axis` correto e entender como os valores ausentes são tratados.

8. Utilizando pandas, como realizar a leitura de um arquivo CSV em um DataFrame e exibir as primeiras linhas?

Para ler um arquivo .csv e exibir as primeiras linhas usando pandas, a pessoa precisa seguir os seguintes passos:

1. Importe a biblioteca Pandas
2. Leia o arquivo com `pd.read_csv()`
3. Exiba as primeiras linhas com `df.head()`

9. Utilizando pandas, como selecionar uma coluna específica e filtrar linhas em um "DataFrame" com base em uma condição?



No pandas posso selecionar uma coluna de um dataframe e ao mesmo tempo filtrar linhas com base em uma condição.

10. Utilizando pandas, como lidar com valores ausentes (NaN) em um DataFrame?

No pandas, os valores ausentes (NaN) podem ser tratados de diferentes formas dependendo da necessidade. Nesse sentido podemos remover valores ausentes, preencher valores ausentes, interpolar valores ou retorna uma máscara booleana onde existe NaN.