

Trabalho de Introdução ao Pré-processamento de Dados com Pandas

Nivelamento Técnico - Pós-Dev. IA - IFES - TRE

Nome do Aluno/a: _____

Data de Entrega: **05 de Julho de 2023**

Professor: Sérgio N. Simões

Oi pessoal. O objetivo deste trabalho é praticar suas habilidades na utilização da biblioteca **Pandas**. Para isso, vocês deverão criar um Notebook e realizar os exercícios abaixo aplicados ao dataset que receberam com dados reais de COVID.

Exercícios de Fixação de comandos do Pandas

Exercício 1:

- a) Carregue o dataset em um DataFrame do Pandas.
- b) Verifique a quantidade de linhas e colunas presentes no dataset.
- c) Exiba as primeiras 3 (três) linhas do DataFrame para entender a estrutura dos dados.
- d) Exiba as últimas 4 (quatro) linhas do DataFrame.

Exercício 2:

- a) Exiba os nomes das colunas do DataFrame.
- b) Verifique os tipos de dados presentes em cada coluna.
- c) Em alguns casos, pode ser necessário fazer conversões de tipo de dados adequadas. Por exemplo: um dado que deveria ser inteiro e foi lido como *string*. Pesquise como fazer esta conversão, caso necessária.

Exercício 3:

- a) Verifique se há valores ausentes (**NaN**) no DataFrame. Observe que alguns bancos de dados utilizam números para representar dados ausentes.
- b) Identifique as **colunas** que possuem valores ausentes.
- c) Identifique e conte quantas **linhas** possuem valores ausentes.
- d) **Discuta** uma decisão apropriada para lidar com os valores ausentes. Por exemplo, preencher com média, mediana ou remover as linhas. E em alguns casos pode até mesmo ser necessário remover uma coluna (caso ela tem um percentual muito alto de valores nulos).

Exercício 4:

- a) Realize uma análise estatística básica do DataFrame, exibindo medidas como média, mediana, desvio padrão, etc.
- b) Identifique a coluna que possui o maior valor máximo e a coluna com o menor valor mínimo.

Exercício 5:

- a) Crie uma nova coluna no DataFrame que represente a taxa de letalidade do COVID, calculada como o número de óbitos dividido pelo número de casos confirmados, multiplicado por 100.
- b) Exiba as primeiras cinco linhas do DataFrame para verificar se a coluna foi adicionada corretamente.

Exercício 6:

- a) Realize uma filtragem no DataFrame para exibir apenas as linhas referentes aos pacientes que foram para casa.
- b) Verifique quantas linhas foram filtradas.
- c) A partir desses dados filtrados, verifique se todos possuem o valor NULO (97, 98, 99) a coluna referente a 'intubados' e a coluna referente a UTI.
- d) Repita o item (a) e (b) para os pacientes que foram hospitalizados.
- e) Conte qual o percentual de pacientes grávidas dentre as mulheres que foram hospitalizadas.

Exercício 7:

- a) Ordene o DataFrame em ordem decrescente com base em uma coluna numérica de sua escolha.
- b) Exiba as primeiras 10 (dez) linhas do DataFrame ordenado.
- c) Agora selecione duas colunas numéricas de sua escolha e ordene o DataFrame em ordem decrescente pela primeira coluna e, em caso de empate, em ordem crescente da segunda coluna.
- d) Agora verifique o resultado exibindo as primeiras 10 (dez) linhas do DataFrame ordenado por duas colunas.

Exercício 8:

- a) Agrupe os dados por uma determinada coluna e calcule a soma, média e contagem de outra coluna numérica.
- b) Exiba o resultado do agrupamento de forma organizada.

Exercício 9:

- a) Utilize a função `pivot_table` para criar uma tabela que resuma os dados de duas colunas do DataFrame.
- b) Exiba a tabela criada.

Exercício 10:

- a) Salve o DataFrame final em um arquivo CSV para futuras referências. b) Verifique se consegue ler corretamente o arquivo salvo.

Lembrem-se de documentar cada etapa do trabalho e, se necessário, fornecer explicações claras para cada exercício.

OBS:

Este trabalho tem duas partes e esta é a primeira.
A segunda parte será desenvolvida em sala de aula no dia 05/07/2023.

Bom trabalho!
