Turing Academy - 2023 - Entregável II

Análise/Limpeza de Dados

Chegou o momento! Os entregáveis que você, trainee, deverá realizar consistem em um conjunto de 3 exercícios que avaliarão o seu domínio nos tópicos estudados durante o Turing Academy até o momento, dentre os quais:

- Manipulação de DataFrame com pandas;
- Visualização;
- Limpeza;
- Análise;
- Pré-Processamento de dados;
- Básico de Machine Learning;

Para não sobrecarregá-los, soltaremos um exercício por vez (1 por semana). Recomendamos fortemente que vocês tentem fazer os exercícios à medida que eles saiam (não deixem acumular!!!).

Para a entrega dos exercícios, pediremos para que vocês criem um repositório no GitHub (ele deverá ser público) e adicionem o link para ele no espaço apropriado dessa planilha: https://docs.google.com/spreadsheets. Submetam a esse repositório 1 arquivo no formato .ipynp para cada entregável. O prazo para entrega de todos os exercícios é até dia 09/07 (não necessariamente eles precisam ser entregues na ordem).

Para a realização dos exercícios, não se limitem aos conhecimentos vistos durante as aulas ou ao material do TA - incentivamos que busquem conhecimento em outras fontes, como no DataCamp ou no Youtube. Não hesitem em chamar os seus mentores ou qualquer um dos professores do TA para tirar dúvidas. Estaremos sempre à disposição! Bom trabalho!

EX2 - O Peso de Turingópolis

O prefeito da cidade de Turingólis, Daniel Vidigal, decidiu realizar um censo com o objetivo de avaliar a saúde dos cidadãos. Para isso, foram entrevistados indivíduos de todas as faixas etárias, dos 14 até os 70 anos de idade. Durante o censo, foram feitas uma série de perguntas relevantes, abordando temas como a presença de doenças crônicas, o histórico de condições médicas na família, a situação de imunização contra raiva politécnica e outros aspectos relacionados à saúde.

Os dados coletados durante o censo foram divididos por temas e entregues a diferentes cientistas de dados. Você, um desses cientistas, recebeu acesso ao seguinte dataset: dataset2.xls, contendo informações do peso e da altura da população, e lhe foi pedido que o analisasse. Tendo isso em mente, responda o abaixo:

a) Esboce um gráfico de dispersão (scatterplot) do peso em função da altura. Que informações você tira desse gráfico? Você percebe algo unusual nele?

b) Identificou-se que, no processo de integração entre 2 conjuntos de dados distintos, gerou-se uma inconsistência em 2 features de nosso dataset. Identifique-as e corrija esses problemas.

   Dica: pode ser que as mesmas observações apresentem os mesmo problemas.

c) Crie uma nova coluna, IMC, a partir das colunas de peso e altura.

d) Explique a diferença de outliers uni e multivariados. Utilize a coluna IMC para identificar outliers em nosso conjunto de dados. Estabeleça um critério lógico e, a partir desse critério, remova os outliers que julgar cabível. Nesse contexto, seriam esses outliers uni ou multivariados?

e) Qual a porcentagem da população poderia se enquadrar fora da condição "saudável" de peso (com base no IMC)? Siga a tabela a seguir.

IMC	Classificação
< 18,5	Magreza
18,5 – 24,9	Saudável
25,0 – 29,9	Sobrepeso
30,0 – 34,9	Obesidade Grau I
35,0 – 39,9	Obesidade Grau II (severa)
? 40,0	Obesidade Grau III (morbida)

Tabela 1 - Classificação com base no Índice de Massa Corporal

f) Trace gráficos do tipo violino (violinplot) da altura e do peso em função do sexo e analise-os.