Mestrado em Engenharia Informática
Ciência de Dados
Ano Letivo 2023/2024

Mini Projeto III

Trabalho Individual
Data Wrangling and Exploratory Data Analysis

© Ricardo Campos
ricardo.campos@ubi.pt

O trabalho prático é obrigatório para a obtenção de aprovação na unidade curricular. A não entrega durante o prazo previsto implica a automática reprovação dos alunos.

Objetivo: Familiarização com o processo de aquisição de dados através de ficheiros, packages e APIs.

Entrega: Os trabalhos (em formato notebook – devidamente documentados) devem ser inseridos na plataforma de e-learning (moodle) até 02/05/2024, 23h59. O nome do notebook a submeter no moodle deve cumprir com o seguinte formato: *XXXX.ipynb*, onde *XXXX* é o número do aluno (e.g., *10000.ipynb*).

Realização do trabalho: Os trabalhos devem ser realizados individualmente.

Tarefa 1: Familiarização com *Data Wrangling and Exploratory Data Analysis*
Considere um ficheiro *csv* à sua escolha. Proceda à descrição do *dataset* e das suas principais colunas. Responda às seguintes questões tendo por base o ficheiro. Nota: não serão admitidos trabalhos com base no mesmo ficheiro csv. Nesse sentido, solicita-se que indique a localização/origem do seu ficheiro (*first come – first serve*) no link abaixo: https://1drv.ms/x/s!AqbUf6ry5g9tlEG8UZfAp0ENctR5?e=DkxuHQ.

1. Proceda à importação do ficheiro para um Pandas *dataframe*.
2. Mostre os 5 primeiros registos.
3. Mostre o coeficiente de correlação de *pearson* entre cada par de atributos. Liste os valores de correlação de forma descendente para um atributo à sua escolha.

4. Devolva a mediana de um atributo à sua escolha (restringindo a um conjunto de dados específico. Exemplo: a mediana da idade das pessoas do sexo feminino).
5. Escreva o código que lhe permite contabilizar o número de registos *null* existente num conjunto de colunas à sua escolha.
6. Desenvolva uma função de imputação que proceda à substituição dos valores nulos de uma coluna à sua escolha com o valor da mediana desse atributo. Considere, sempre que possível, diferentes valores de mediana para cada classe (por exemplo, proceda à substituição dos valores nulos da coluna *Age* de acordo com a mediana da *Age* apurada para cada uma das três classes existentes (*Pclass* = 1, *Pclass* = 2, *Pclass* = 3)).
7. Crie novas colunas no seu *dataset*, potencialmente relacionadas com as colunas atuais (exemplo, a coluna *Title* (com os valores Mr; Miss; etc) a partir da coluna *Name* (que inclui os valores Mr. Santos; Miss Filipa).
8. Proceda a uma análise exploratória de dados que considere relevante no contexto do seu ficheiro.