Bootcamp: Engenheiro(a) de Dados

Desafio Prático

Módulo 2 - Linguagem Python aplicada a Engenharia de Dados

Objetivos de Ensino

Exercitar os seguintes conceitos trabalhados no Módulo:

1. Pandas;
2. Numpy;
3. Classes, Atributos e Métodos;
4. Lógica de programação.

Enunciado

Neste trabalho prático utilizaremos o dataset presente no link a seguir: https://www.kaggle.com/datasets/ramjasmaurya/1-gb-internet-price

Baixe o dataset antes de prosseguir.

Dentro do dataset existem 4 arquivos.

Nosso objetivo com esse desafio prático é simular uma das principais atividades no escopo da engenharia de dados e da engenharia de analíticas, que é o tratamento de bases de dados para o trabalho de analistas e cientistas de dados.

Atividades

1. Vamos criar uma classe para isso. Comece criando um método para abrir os arquivos e armazená-los como atributos da classe. Para isso, abra os arquivos 'worldwide internet users - users.csv', 'worldwide internet prices in 2022 - IN 2022.csv' e 'worldwide internet speed in 2022 - avg speed.csv' usando PANDAS e os deixe armazenados em atributos: usuarios, precos e velocidades, respectivamente.

2. Analise o shape e as colunas de cada um. Observe também o tipo de cada coluna e identifique qual é a coluna em comum entre os arquivos.

3. Escreva um método para unir os arquivos em um único arquivo, usando a coluna de nome do país com índice para unificá-los. Esse método não precisa de parâmetros, apesar de você poder fazê-lo recebendo a lista da coluna que é o nome do país de cada arquivo. Salve o dataframe unificado no atributo 'dados_unificados'.

4. Analise os dados faltantes em todas as colunas do atributo ‘dados_unificados’. Comece a criar um método de tratamento de dados e implemente nele a remoção dos itens que não possuírem informações sobre a população (NaN).

5. Neste mesmo método, trate todas as colunas que possuem números para ficarem com tipos numéricos (use a função 'to_numeric' do Pandas). Juntamente com essa ação, remova os itens que tiverem dados textuais e NaN nessas colunas. Por enquanto, evite a coluna 'Avg \n(Mbit/s)Ookla', iremos trabalhar com ela na sequência.

6. Agora que quase todas as colunas estão tratadas, precisamos tratar a coluna de média de velocidade. Para isso, vamos fazer alguns processos, simulando a atuação de um time de negócio que está responsável pela demanda da base de dados que estamos criando. Faça uma análise da quantidade de planos ('NO. OF Internet Plans') e

implemente um método para criar uma coluna chamada 'AnaliseQuantidadePlanos' que recebe o texto ‘POUCO’ se a quantidade de planos for menor que a mediana ou 'MUITO' se a quantidade for maior ou igual a mediana.

7. Faça uma análise do preço médio de 1GB ('Average price of 1GB (USD)') e implemente outro método para criar uma coluna chamada 'AnalisePrecoMedio' que recebe o texto 'BARATO' se o preço for menor que a mediana ou 'CARO' se o preço for maior ou igual a mediana.

8. Por fim, crie um método que chame os outros dois métodos e que combine as colunas 'AnaliseQuantidadePlanos' e 'AnalisePrecoMedio' em uma nova coluna com o nome 'PerfilPlanoPreco' que será a concatenação do valor das duas colunas usadas.

9. Com os métodos criados nas 3 atividades anteriores, implemente, volte no método de tratamento de dados e atribua um valor para a coluna de 'Avg \n(Mbit/s)Ookla' nos casos em que ela esteja NaN. Para isso, usando os itens que possuem algum valor atribuído a essa coluna, calcule a mediana da coluna, separando por grupos. Cada grupo será um valor da coluna 'PerfilPlanoPreco'. A ideia aqui é atribuir ao valor da velocidade de um país a mediana das velocidades dos países que possuem o mesmo perfil de quantidade de planos e de preço de internet que ele.

10. Os valores monetários até o momento estão em dólares americanos. Considerando a conversão de 1 dólar para 5 reais, crie um método que altere os valores das colunas que possuem valores financeiros para que elas fiquem com valores em reais. Lembre-se de chamá-lo no método de tratamento de dados.

11. Pensando em facilitar a vida de quem vai utilizar esses dados, vamos renomear as colunas. Implemente um método que faça a renomeação

das colunas conforme a indicação abaixo e adicione-o ao método de tratamento de dados.

Country code: Codigo
Continental region: Regiao Continental
NO. OF Internet Plans : Numero de Planos
Average price of 1GB (USD): Preco medio de 1GB
Cheapest 1GB for 30 days (USD): 1GB mais barato
Most expensive 1GB (USD): 1GB mais caro
Average price of 1GB (USD at the start of 2021): Preco medio de 1GB desde 2021
Average price of 1GB (USD – at start of 2020): Preco medio de 1GB desde 2020
Subregion: Subcontinente
Region: Continente
Internet users: Numero de Usuarios
Population: Populacao
Avg \n(Mbit/s)Ookla: Velocidade media
AnaliseQuantidadePlanos: Numero de Plano Qualitativo
AnalisePrecoMedio: Preco medio de 1GB Qualitativo
PerfilPlanoPreco: Perfil de Plano e Preco

12. Finalizando, crie um método para salvar (ou persistir) essa versão final dos dados. O método deve receber o nome do arquivo de saída e também um parâmetro indicando se os dados devem ser salvos em csv ou em pickle. Salve os dados em três versões:

    a. CSV;

    b. Pickle com a extensão ".pkl";

    c. Pickle com a extensão ".gz".