ENUNCIADO TRABALHO DA DISCIPLINA DE TÓPICOS EM BIG DATA EM PYTHON

O grupo deverá utilizar um dataset para o trabalho. Uma boa fonte de datasets é o https://datasetsearch.research.google.com/.

O trabalho deve conter os seguintes itens:

Lista de bibliotecas usadas:

1. Indicação da lista de bibliotecas usadas no início do trabalho e a justificativa para uso

ETL (Extração - Transformação - Carga):

2. Indicação da fonte dos dados
3. Usar uma fonte com, pelo menos, 1 mil registros
4. Apresentar o metadado (dicionário de dados) do dataset
5. Escrever o código para a importação do dataset a partir do drive do Colab ou usar o Jupyter Notebook com o link para os dados (Google Drive ou outro)
6. Renomeie os nomes das colunas para nosso idioma, se necessário

Análise Exploratória de Dados básica:

7. Realizar uma visão inicial das 10 primeiras linhas do dataset
8. Apresentar uma amostra aleatória dos dados
9. Listar os nomes das colunas
10. Verificar a dimensão do dataset (total de linhas e colunas)
11. Contar o total de amostras por uma das variáveis categóricas tanto ordinal ou nominal (textual)
12. Destacar os valores máximos das 20 primeiras linhas
13. Destacar os valores mínimo das 20 últimas linhas
14. Realizar o destaque (highlight) dos dois itens anteriores
15. Apresentar a estatística básica para o dataset
16. Realizar a análise de correlação, usando o método Pearson
17. Contar o total de linhas para uma determinada variável categórica
18. Listar as informações sobre o dataset

Análise Exploratória de Dados com gráficos:

Análise de Correlação:

19. Criar um gráfico heatmap para a análise de correlação do item 16
20. Criar um scatterplot para o par de variáveis com maior correlação
21. Criar um correlograma
22. Realizar a análise bivariada por meio de scatterplots para exibir a distribuição dos dados entre as principais variáveis categóricas. Utilize cores e altere o tamanho dos pontos para facilitar a interpretação

Análise de Distribuição:

23. Realizar a análise univariada com um histograma para uma variável numérica
24. Apresentar em apenas um gráfico vários histogramas para as variáveis numéricas

25. Verificar com boxplots a presença de possíveis outliers
26. Remover os outliers, caso existam
27. Apresentar um gráfico de densidade para uma variável numérica
28. Apresentar um gráfico de densidade para mais de uma variável numérica

Gráficos de classificação oui rankeamento:

29. Criar um gráfico de barras ou colunas para exibir também o resultado do item 11
30. Criar um gráfico de nuvem palavras, caso seu dataset permita

Gráficos de partes ou setores:

31. Criar um gráfico de pizza ou setores para uma variável categórica com porcentagens

Gráficos de evolução:

32. Criar um gráfico de evolução para as variáveis numéricas

Análise dos Dados:

33. Realizar uma análise que pode ser uma nuvem de palavras, mineração de dados (associação, clusterização ou classificação)