Trabalho final

Data de entrega: 10/11

Trabalho Individual ou em grupo (de 3 pessoas)

Objetivo: Aplicar os conhecimentos adquiridos na disciplina em uma análise de três bancos de dados.

Instruções:

- A análise de cada banco deve ser realizada em notebooks **separados**
- Utilizar markdown (escrita no notebook) para deixar claro o que está sendo realizado em cada passo.
- Utilize de comentários e docstring para deixar o raciocínio no código.
- O arquivo final deve conter o nome **TP_NomeDoAluno_NumerodoBanco.ipynb**
- Pode ser utilizada uma amostra do banco para fins de otimização de tempo (Porém deve ser deixado explícito o critério utilizado para seleção da amostra)

1 Análise

Fonte de dados:
Sales_Transactions_Dataset_Weekly.csv

Este banco de dados contém o valor de transações em vendas de alguns produtos durante 52 semanas. Iremos analisar as séries temporais das transações.

Colunas:

Product_Code: código do produto

W[0-51]: valor de transação da semana

Realize os seguintes processos de análises:

1. Transforme o data frame do formato wide para long
   (Aplicar função para transformar o dado W[0-51] em dia da semana disponibilizada no arquivo week_day.py)
2. Escolha três produtos (de forma aleatória; pode utilizar semente) e construa um gráfico com o valor de transação de cada um dos produtos selecionados ao longo do tempo.
3. Por produto calcule a diferença entre as semanas e gere um gráfico da série temporal da diferença ao longo das semanas.
4. Calcule as descritivas de cada produto (média, desvio-padrão, mínimo, mediana, máximo)

5. Calcule a média móvel de cada 3 semanas e gere um gráfico contendo as médias móveis de um produto.

2ª Análise

Fonte de dados:
Womens Clothing E-Commerce Reviews.csv

Este banco de dados contém review de produtos de e-commerce de vestuário feminino.

Colunas:
Review text: texto da review
RecommendedIND: variável binária onde 1 indica que a pessoa recomenda o produto e 0 caso contrário.

Realize os seguintes processos de análises:

1. Calcule a distribuição da variável de recomendação (**RecommendedIND**)
2. Construa uma função para normalizar o texto (passar tudo para minúsculo e remover espaços duplos) e aplique ao texto (**ReviewText**)

TEXTO NORMALIZADO PARA 3 E 4

3. Crie um wordcloud para as frases em que a recomendação do produto é 1
4. Crie um dataframe contendo o número de frases em que uma palavra aparece. As colunas são
    a. Palavra: a palavra identificada
    b. Valor: número de frases com a palavra.
5. Obtenha as 5 palavras mais frequentes nos textos (utilize o dataframe criado no processo 4).

3ª Análise

Fonte de dados: visitas.csv
Este banco contém as visitas que um cliente realizou a um site de compra.

Colunas:
id_visita: Identificação da visita
id_cliente: Identificação do cliente

Fonte de dados: receita.csv
Este banco contém a receita de produtos comprados durante a visita de um usuário.

Colunas:
id_visita: Identificação da visita
receita: Receita gera pelos produtos comprados naquela visita.

1. Unifique os dois dataframes em um utilizando o campo id_visita, preencha os dados faltantes com valor 0.
2. Calcule as descritivas a seguir por id_cliente:
    a. Média incluindo visitas sem receita
    b. Média excluindo visitas sem receita
    c. Percentual de visitas com receita
3. Crie um algoritmo que faça os seguintes passos:

Para B indo de 1 até 1000:

1. Gere uma amostra (com reposição) do dataframe original. (df.sample(n="tamanho do banco", replace=True))
2. Calcule a média de receita da amostra (considerando visitas sem receita)
3. Armazene esse valor em uma lista com o valor de B e uma lista com os valores de média.

Criar um DataFrame utilizando essas listas.

4. Crie uma visualização para a distribuição da coluna media_amostrada no passo 3.