Descrição do projeto

DATASETS

- instacart_orders.csv
- products.csv
- order_products.csv
- aisles.csv
- departments.csv

Parabéns por concluir o sprint sobre AED! É hora de aplicar o conhecimento e as habilidades que você adquiriu a um estudo de caso analítico.

Quando terminar o projeto, envie seu trabalho ao revisor na plataforma para avaliação. Você vai receber feedback em até 48 horas. Use o feedback para fazer alterações e, em seguida, envie a nova versão de volta ao revisor do projeto.

Você talvez receba mais feedback referente à nova versão. Isso é completamente normal. Não é incomum passar por vários ciclos de feedback e revisão.

Seu projeto será considerado concluído assim que o revisor do projeto o aprovar.

Descrição do projeto

Neste projeto, você vai trabalhar com dados da Instacart.

A Instacart é uma plataforma de entrega de supermercado onde os clientes podem fazer um pedido de supermercado e receber a compra em casa, semelhante ao funcionamento do Uber Eats e do iFood. Esse conjunto de dados específico foi **lançado publicamente** *(os materiais estão em inglês) pela Instacart em 2017 para uma **competição Kaggle** *(os materiais estão em inglês)*. Os dados reais podem ser baixados na página da Kaggle.

O conjunto de dados que fornecemos foi modificado a partir do original. Reduzimos o tamanho dele para que seus cálculos sejam executados mais rapidamente e incluímos valores ausentes e duplicados. Também tivemos o cuidado de preservar as distribuições dos dados originais quando fizemos as alterações.

Sua missão é limpar os dados e preparar um relatório que forneça informações sobre os hábitos de compra dos clientes da Instacart. Após responder a cada

Neste projeto, você vai precisar criar gráficos para apresentar seus resultados. Certifique-se de que todos os gráficos que você criar têm um título, eixos rotulados e uma legenda, se necessário; e inclua `plt.show()` no final de cada célula com um gráfico.

Assista a este vídeo para mais dicas sobre o projeto:

Dicionário de dados

Há cinco tabelas no conjunto de dados, e você vai precisar usar todas elas para pré-processar os dados e fazer a AED. Abaixo está um dicionário que lista as colunas de cada tabela e descreve os dados contidos nelas.

- `instacart_orders.csv` : cada linha corresponde a um pedido no aplicativo da Instacart
  - `'order_id'` : é o número de identificação exclusivo de cada pedido
  - `'user_id'` : é o número de identificação exclusivo da conta de cada cliente
  - `'order_number'` : é o número de vezes que o cliente fez um pedido
  - `'order_dow'` : é o dia da semana em que o pedido foi feito (0 é domingo)
  - `'order_hour_of_day'` : é a hora do dia em que o pedido foi feito
  - `'days_since_prior_order'` : é o número de dias desde que o cliente fez seu pedido anterior
- `products.csv` : cada linha corresponde a um produto exclusivo que os clientes podem comprar
  - `'product_id'` : é o número de identificação exclusivo de cada produto
  - `'product_name'` : é o nome do produto
  - `'aisle_id'` : é o número de identificação exclusivo de cada categoria de corredor do supermercado
  - `'department_id'` : é o número de identificação exclusivo de cada categoria de departamento do supermercado
- `order_products.csv` : cada linha corresponde a um item incluído em um pedido
  - `'order_id'` : é o número de identificação exclusivo de cada pedido
  - `'product_id'` : é o número de identificação exclusivo de cada produto
  - `'add_to_cart_order'` : é a ordem sequencial em que cada item foi colocado no carrinho
  - `'reordered'` : 0 se o cliente nunca comprou o produto antes, 1 se já o comprou
- `aisles.csv`
  - `'aisle_id'` : é o número de identificação exclusivo de cada categoria de corredor do supermercado
  - `'aisle'` : é o nome do corredor
- `departments.csv`

- `'department_id'` : é o número de identificação exclusivo de cada categoria de departamento do supermercado
- `'department'` : é o nome do departamento

Instruções para concluir o projeto

Etapa 1: abra os arquivos de dados (`/datasets/instacart_orders.csv` , `/datasets/products.csv` , `/datasets/aisles.csv` , `/datasets/departments.csv` e `/datasets/order_products.csv`) e dê uma olhada no conteúdo geral de cada tabela.

Observe que os arquivos têm formatação fora do padrão, então você vai precisar definir certos argumentos em `pd.read_csv()` para ler os dados corretamente. Dê uma olhada nos arquivos CSV para ter uma ideia de quais deveriam ser esses argumentos.

Observe que `order_products.csv` contém *muitas* linhas de dados. Quando um DataFrame tem muitas linhas, `info()` não imprime as contagens de valores não nulos por padrão. Se quiser imprimi-las, inclua `show_counts=True` quando chamar `info()` .

Etapa 2: faça o pré-processamento dos dados da seguinte maneira:

- Verifique e corrija os tipos de dados (por exemplo, certifique-se de que as colunas de ID sejam números inteiros)
- Identifique e preencha valores ausentes
- Identifique e remova valores duplicados

Certifique-se de explicar que tipos de valores ausentes e duplicados você encontrou, como você os preencheu ou removeu, por que escolheu esses métodos e por que você acha que esses valores ausentes e duplicados estavam presentes no conjunto de dados.

Etapa 3: quando os dados estiverem processados e prontos, execute a seguinte análise:

[A] (é necessário concluir tudo para passar)

1. Verifique se os valores nas colunas `'order_hour_of_day'` e `'order_dow'` na tabela `orders` fazem sentido (ou seja, os valores da coluna `'order_hour_of_day'` variam de 0 a 23 e os da `'order_dow'` variam de 0 a

6).

2. Crie um gráfico que mostre quantas pessoas fazem pedidos a cada hora do dia.
3. Crie um gráfico que mostre em que dia da semana as pessoas fazem compras.
4. Crie um gráfico que mostre quanto tempo as pessoas esperam até fazer seu próximo pedido e comente sobre os valores mínimo e máximo.

[B] (é necessário concluir tudo para passar)

1. Há diferenças entre as distribuições de `'order_hour_of_day'` na quarta-feira e no sábado? Construa histogramas para ambos os dias no mesmo gráfico e descreva as diferenças que você notou.
2. Construa um gráfico de distribuição para o número de pedidos que os clientes fazem (ou seja, quantos clientes fizeram apenas 1 pedido, quantos fizeram apenas 2, quantos apenas 3, etc.)
3. Quais são os 20 produtos comprados com mais frequência? Exiba os números de identificação (ID) e nomes.

[C] (é necessário concluir pelo menos duas perguntas para passar)

1. Quantos itens as pessoas normalmente compram em um pedido? Como fica a distribuição?
2. Quais são os 20 principais itens incluídos mais frequentemente em pedidos repetidos? Exiba os números de identificação (ID) e nomes.
3. Para cada produto, que proporção de pedidos em que ele aparece são pedidos repetidos? Crie uma tabela com colunas para ID do produto, nome do produto e a proporção de pedidos repetidos.
4. Para cada cliente, que proporção dos produtos comprados são pedidos repetidos?
5. Quais são os 20 principais itens que as pessoas colocam nos carrinhos antes de todos os outros? Exiba o ID do produto, nome e o número de vezes que ele foi o primeiro a ser adicionado a um carrinho.

Ou faça o projeto em seu computador e faça upload de seu trabalho quando finalizar. **Fazer upload e enviar**

O que você achou do projeto? 😊 😕