🟡

Desafio: Prepare seu dataset para modelagem de dados

🚀 Desafio

Prepare seu dataset para modelagem de dados

📎 **Arquivo do Desafio: Dados.csv**

> Aplique os conhecimentos de limpeza e organização de dados (data cleaning & data wrangling) para estruturar uma base de dados para modelagem. Boa parte do dia de um cientista de dados é garantir a organização de suas bases para conseguir um modelo preciso, dominar isso é fundamental na carreira.

Contexto

Uma empresa do ramo de e-commerce contratou você para levantar os indicadores de recência, frequência e ticket médio (RFM) dos seus clientes.

A saber RFM:

- R (Recency): Tempo que o cliente realizou a última compra (em dias)
- F (Frequency): Quantidade de compras realizadas pelo cliente
- M (Monetary): Valor do ticket médio gasto pelo cliente

onde ticket médio = média do total gasto por pedido para cada cliente.

Para isso, vocês receberam uma base de dados (arquivo csv) e devem construir um código em Python que gera um output também csv, porém contendo apenas a identificação do cliente e métricas RFM.

Sobre os dados

A tabela contém informações de compras de um e-commerce em 37 países. Contém a identificação do cliente e os dados da compra.

Coluna	Descrição
CustomerID	Código de identificação do cliente
Description	Descrição do produto
InvoiceNo	Código da fatura
StockCode	Código de estoque do produto
Quantity	Quantidade do produto
InvoiceDate	Data do faturamento (compra)
UnitPrice	Preço unitário do produto
Country	País da compra

Como começar?

1. Importe o dataset para o colab
2. Entenda os dados
3. Trate os dados nulos
4. Trate os outliers

Desenvolva o algoritmo para receber o arquivo csv de entrada e retornar um algoritmo de saída com as seguintes colunas:

- **CustomerID: Código do cliente**
- **R: Recência**
- **F: Frequência**
- **M: Ticket médio**

🎯 Etapas de Desenvolvimento

Para te ajudar nesse processo, detalhar o processo

Etapa 01) Leia o arquivo e inspecione os dados

Pesquisar o mercado apontando os produtos que já existem hoje e seus diferenciais. É importante apresentar de forma clara os concorrentes diretos e indiretos, e destacar seus diferenciais e pontos fortes.

1. Leia o dataset
2. Utilize o describe para verificar a distribuição dos dados
3. Analise o tipo dos dados
4.

> 💡 **Dica:**
>
> Leia o dataset
>
> 1. Utilize o describe para verificar a distribuição dos dados
> 2. Analise o tipo dos dados

Etapa 02) Valores faltantes na identificação do cliente

> 💡 **Dica: Se sim, remova estas observações.**
>
> 1. Verifique os valores nulos com o isna e utilize a função sum para a somar a quantidade de nulos
> 2. Utilize a função dropna para remover os nulos

Etapa 03) Preços unitários e quantidade de produtos iguais ou inferior a 0

💡 **Dica:** Se sim, remova estas observações.

1. Realize um filtro para verificar se existem dados nulos ou menor que zero na coluna de preços
2. Filtre o dataset apenas para conter preços acima de zero
3. Realize um filtro para verificar se existem dados nulos ou menor que zero na coluna de quantidade
4. Filtre o dataset apenas para conter quantidade acima de zero

Etapa 04) Verifique se existem linhas duplicadas

💡 **Dica:** Se sim, remova estas observações (pois não faz sentido uma mesma compra para o mesmo cliente no mesmo horário, com mesmos valores etc.)

1. Verifique se tem linhas duplicadas com a função duplicated
2. Drope as linhas duplicadas

Etapa 05) Tipos de dados da coluna

💡 **Dica:**

1. Corrija o tipo de dado do CustomerID
2. Corrija o tipo de dado da InvoiceDate

Coluna	Tipo esperado
InvoiceNo	object / str
StockCode	object / str
Description	object / str
Quantity	int
InvoiceDate	datetime
UnitPrice	float
CustomerID	int
Country	object / str

Etapa 06) Tratando os outliers

> 💡 **Dica:** Vamos considerar estes valores como erro.**Visualize os outliers e remova os outliers extremos em que a quantidade do item na compra é superior a 10.000, e o preço unitário é maior que 5.000.**

Etapa 07) Crie uma coluna adicional

> 💡 **Dica:** Utilize as colunas Quantity e UnitPrice. **Crie uma coluna adicional com o preço total da compra**

Etapa 08) Última data

> 💡 **Dica:** Utilize a função max(). **Calcule a data da última compra no dataset como um todo, pois vamos utilizar este valor como data de comparação para cálculo da recência.**

Etapa 09) Plotando gráficos

- **Top 10 países com maior valor em vendas**

- **Top 10 produtos mais vendidos**
- **Valor de venda total por mês**
- **Valor de venda total por mês e por país (considere apenas os top 10)**

Etapa 10) Cálculo do RFM

> 💡 **Dica:** Agrupe os dados por cliente e pedido/compra (InvoiceNo) e obtenha a data e o preço total do pedido.
> Com isso, agrupe novamente apenas por cliente e calcule o RFM, onde:
>
> - R é a recência, diferença em dias da última compra do cliente e da última compra disponível no conjunto de dados, que calcularam previamente.
> - F é a frequência, ou seja, a quantidade de compras feitas pelo cliente;
> - M é o ticket médio, ou seja, a média das compras feitas pelo cliente.

 Critérios de Avaliação

Os critérios de avaliação mostram como você será avaliado em relação ao seu desafio.

Atendeu às Especificações		Pontos
Leia o arquivo e inspecione os dados	Ler o dataset, utilizar describe	5
Verifique se há valores faltantes na identificação do cliente	Remova estas observações.	5
Verifique se há produtos com preços unitários iguais ou inferior a 0	Remova estas observações.	10
Verifique se há produtos com quantidade igual ou	Remova estas observações.	10

inferior a 0		
Verifique se existem linhas duplicadas	Remova estas observações (pois não faz sentido uma mesma compra para o mesmo cliente no mesmo horário, com mesmos valores etc.)	10
Tipos de dados da coluna	1. Corrija o tipo de dado do CustomerID 2. Corrija o tipo de dado da InvoiceDate	10
Tratando os outliers	Considerou a quantidade acima 10000 e o preço acima 5000 como outlier	10
Crie uma coluna adicional	Utilizou as colunas Quantity e UnitPrice	10
Última data	Utilizou a função max() para achar a ultima data	10
Plotando gráficos	- Top 10 países com maior valor em vendas - Top 10 produtos mais vendidos - Valor de venda total por mês - Valor de venda total por mês e por país (considere apenas os top 10)	10
Cálculo do RFM	Agrupe os dados por cliente e pedido/compra (InvoiceNo) e obtenha a data e o preço total do pedido. Com isso, agrupe novamente apenas por cliente e calcule o RFM, onde: - R é a recência, diferença em dias da última compra do cliente e da última compra disponível no conjunto de dados, que calcularam previamente. - F é a frequência, ou seja, a quantidade de compras feitas pelo cliente; - M é o ticket médio, ou seja, a média das compras feitas pelo cliente.	10

📆 Entrega

📎 **Como entregar:** Você deverá submeter o link compartilhável do colab!