DESAFIO 1:

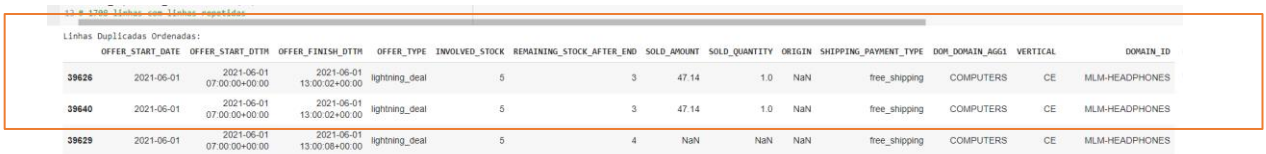
Considerando que a possível descrição dos campos são:

- OFFER_START_DATE: Data de início da oferta.
- OFFER_START_DTTM: Data e hora de início da oferta.
- OFFER_FINISH_DTTM: Data e hora de término da oferta.
- OFFER_TYPE: Tipo de oferta ("lightning_deal" neste caso).
- INVOLVED_STOCK: Estoque envolvido na oferta.
- REMAINING_STOCK_AFTER_END: Estoque remanescente após o término da oferta.
- SOLD_AMOUNT: Valor total vendido.
- SOLD_QUANTITY: Quantidade total vendida.
- ORIGIN: Origem da oferta (“A” ou “NA”)
- SHIPPING_PAYMENT_TYPE: Tipo de pagamento do frete.
- DOM_DOMAIN_AGG1: Categoria ou domínio do produto.
- VERTICAL: Categoria vertical do produto.
- DOMAIN_ID: ID do domínio do produto.

A princípio, minhas dúvidas são referentes a integridade dos dados, não parecem estar confiáveis.

PERGUNTA 1

No arquivo **ofertas_relampago.csv** pode conter registros repetidos duplicadas ou triplicadas, ou seja, registros idênticos em todas as colunas?



Linhas Duplicadas Ordenadas:

	OFFER_START_DATE	OFFER_START_DTTM	OFFER_FINISH_DTTM	OFFER_TYPE	INVOLVED_STOCK	REMAINING_STOCK_AFTER_END	SOLD_AMOUNT	SOLD_QUANTITY	ORIGIN	SHIPPING_PAYMENT_TYPE	DOM_DOMAIN_AGG1	VERTICAL	DOMAIN_ID
39626	2021-06-01	2021-06-01 07:00:00+00:00	2021-06-01 13:00:02+00:00	lightning_deal	5	3	47.14	1.0	NaN	free_shipping	COMPUTERS	CE	MLM-HEADPHONES
39640	2021-06-01	2021-06-01 07:00:00+00:00	2021-06-01 13:00:02+00:00	lightning_deal	5	3	47.14	1.0	NaN	free_shipping	COMPUTERS	CE	MLM-HEADPHONES
39629	2021-06-01	2021-06-01 07:00:00+00:00	2021-06-01 13:00:08+00:00	lightning_deal	5	4	NaN	NaN	NaN	free_shipping	COMPUTERS	CE	MLM-HEADPHONES
39630	2021-06-01	2021-06-01 07:00:00+00:00	2021-06-01 13:00:08+00:00	lightning_deal	5	4	NaN	NaN	NaN	free_shipping	COMPUTERS	CE	MLM-HEADPHONES
40007	2021-06-01	2021-06-01 07:00:00+00:00	2021-06-01 13:00:08+00:00	lightning_deal	15	15	NaN	NaN	NaN	free_shipping	APPAREL ACCESORIES	APP & SPORTS	MLM-SUNGLASSES
40012	2021-06-01	2021-06-01 07:00:00+00:00	2021-06-01 13:00:08+00:00	lightning_deal	15	15	NaN	NaN	NaN	free_shipping	APPAREL ACCESORIES	APP & SPORTS	MLM-SUNGLASSES
40275	2021-06-01	2021-06-01 13:00:00+00:00	2021-06-01 13:00:01+00:00	lightning_deal	5	5	NaN	NaN	NaN	free_shipping	COMPUTERS	CE	MLM-COMPUTER_MONITORS
40276	2021-06-01	2021-06-01 13:00:00+00:00	2021-06-01 13:00:01+00:00	lightning_deal	5	5	NaN	NaN	NaN	free_shipping	COMPUTERS	CE	MLM-COMPUTER_MONITORS
39794	2021-06-01	2021-06-01 13:00:00+00:00	2021-06-01 19:00:07+00:00	lightning_deal	5	5	NaN	NaN	NaN	free_shipping	APPAREL	APP & SPORTS	MLM-SHIRTS
39802	2021-06-01	2021-06-01 13:00:00+00:00	2021-06-01 19:00:07+00:00	lightning_deal	5	5	NaN	NaN	NaN	free_shipping	APPAREL	APP & SPORTS	MLM-SHIRTS
39823	2021-06-01	2021-06-01 13:00:00+00:00	2021-06-01 19:00:07+00:00	lightning_deal	5	5	NaN	NaN	NaN	free_shipping	APPAREL ACCESORIES	APP & SPORTS	MLM-WRISTWATCHES
39831	2021-06-01	2021-06-01 13:00:00+00:00	2021-06-01 19:00:07+00:00	lightning_deal	5	5	NaN	NaN	NaN	free_shipping	APPAREL ACCESORIES	APP & SPORTS	MLM-WRISTWATCHES

Pergunta: Posso seguir com a deleção desses registros repetidos? Ou tem alguma justificativa para essas replicações?

PERGUNTA 2:

Os valores negativos em REMAINING_STOCK_AFTER_END parecem inconsistentes com a interpretação de que representa o estoque remanescente após o término da oferta. Normalmente, não seria possível ter um estoque remanescente negativo, pois implicaria em vender mais unidades do que o estoque original envolvido. uma tem alguma interpretação diferente do que REMAINING_STOCK_AFTER_END representa?

Ex.:

Evelyn Dias Santos 12 99604-8050

	OFFER_START_DATE	OFFER_START_DTTM	OFFER_FINISH_DTTM	OFFER_TYPE	INVOLVED_STOCK	REMAINING_STOCK_AFTER_END	SO
32275	2021-07-15	2021-07-15 19:00:00+00:00	2021-07-15 21:50:19+00:00	lightning_deal	15	-192	
11448	2021-07-27	2021-07-27 12:00:00+00:00	2021-07-27 15:50:59+00:00	lightning_deal	124	-81	
41045	2021-06-15	2021-06-15 13:00:00+00:00	2021-06-15 13:26:57+00:00	lightning_deal	5	-70	
45648	2021-07-26	2021-07-26 22:00:00+00:00	2021-07-26 23:05:13+00:00	lightning_deal	200	-41	
33627	2021-06-28	2021-06-28 13:00:00+00:00	2021-06-28 14:04:09+00:00	lightning_deal	15	-40	
32845	2021-07-15	2021-07-15 13:00:00+00:00	2021-07-15 17:10:31+00:00	lightning_deal	15	-39	
42715	2021-07-02	2021-07-02 13:00:00+00:00	2021-07-02 18:40:45+00:00	lightning_deal	128	-39	
86	2021-06-22	2021-06-22 16:00:00+00:00	2021-06-22 16:38:21+00:00	lightning_deal	7	-33	

Pergunta: Devo considerar os valores negativos como “0” ou poddo manter como estão, considerando que vendas superiores ao esperado resultem em reposições de estoque? (Apesar de que isso pode estar envolvendo possível "perda" de margem de lucro, pois estão vendendo a mais do q se propuseram a vender na campanha! Mas não vou entrar nesse mérito nesse momento, mas seria uma possível abordagem)

PERGUNTA 3:

Mesmo após o tratamento, o campo `SOLD_AMOUNT` tem valores nulos, com uma representação significativa na base:

```
CAMPO              QTD REGISTRO NULOS     % DA BASE APÓS AJUSTES
SOLD_AMOUNT              22001            47.26%
```

O campo SOLD_AMOUNT é importante para análises de desempenho. Temos algumas abordagens possíveis para tratá-lo:

1- **Excluir os dados nulos da tabela**
   Isso evitará que os dados nulos afetem os resultados
2- **Substituir os dados nulos pela média dos preenchidos**
3- **Apresentar os dados nulos separadamente**
4- **Excluir a coluna das análises**

Pergunta: Poderíamos entrar em abordagens detalhadas sobre as possibilidades desses campos, mas por se tratar de um teste, considera válido seguirmos com o preenchimento dos dados pela média dos dados não nulos por categoria do produto?

PERGUNTA 4:

O Campo `SOLD_QUANTITY` também tem valores nulos

CAMPO	QTD REGISTRO NULOS	% DA BASE
SOLD_QUANTITY	23272	48.66%

Gostaria de entender melhor se o valor do SOLD_QUANTITY deveria ser um campo calculado de

SOLD_QUANTITY = INVOLVED_STOCK - REMAINING_STOCK_AFTER_END

Se for derivado da fórmula, percebo algumas inconsistências como nulos, e valores que não batem, por exemplo, no primeiro registro, o valor deveria ser 5 e não 4, assim por diante:

```
dados_filtrados_positivos[['INVOLVED_STOCK','REMAINING_STOCK_AFTER_END','SOLD_QUANTITY', 'SOLD_QUANTITY_CA
```

	INVOLVED_STOCK	REMAINING_STOCK_AFTER_END	SOLD_QUANTITY	SOLD_QUANTITY_CALCULATE	DIFF_QUANTITY
33	15	11	5	4	1
37	10	8	3	2	1
93	10	9	2	1	1
153	5	3	3	2	1
230	5	3	3	2	1
253	15	11	5	4	1
254	5	4	2	1	1
367	5	-1	8	6	2
394	40	40	3	0	3
401	10	2	10	8	2
468	286	189	171	97	74
470	1000	982	19	18	1
472	2000	1909	556	91	465
473	1200	1114	224	86	138

(Pode ser também que o campo REMAINNING_STOCK_AFTER_END esteja com os valores inconsistentes, mas nesse caso, vamos considerar que estão corretos.)

Temos 3 abordagens:

1- **Atualizar o valor do campo com base na fórmula**
2- **Excluir os nulos dos gráficos**
3- **Substituir os dados nulos pela média dos preenchidos**

Pergunta: Poderíamos entrar em abordagens detalhadas, mas por se tratar de um teste, considera válido seguirmos com a abordagem 1, atualização do valor pela fórmula apresentada acima para deixar o campo com os valores coerentes?

Evelyn Dias Santos 12 99604-8050

PERGUNTA 5:

Oferta com 0 horas de duração são atípicas. Entendo que as ofertas relâmpagos geralmente têm uma duração positiva, mesmo de definida.

Acredito que a presença de ofertas com 0 horas de duração pode ser um erro na captura dos dados deixando esses registros incorretos

```
Número de casos com duração 0: 1314
Percentual em relação à base: 2.70%
```

Pergunta: Dado o baixo número desses registros, podemos excluí-los? Há alguma informação relevante sobre essa situação que devo saber?