Estudo para a solução do problema

1.Qual é a média de tempo de entrega para pedidos feitos na plataforma da Olist?

```
[ ] media_dias_entrega = merged_df_2['days_between'].mean()
    media_dias_entrega
```

```
11.96588279759216
```

2.Existe uma correlação entre o valor do frete e o tempo de entrega?

```
[ ] correlacao = merged_df_2[['days_between', 'freight_value']].corr()
    print("A correlação entre as colunas 'days_between' e 'frete' é:\n", correlacao)
```

```
A correlação entre as colunas 'days_between' e 'frete' é:
                days_between  freight_value
days_between        1.000000       0.215359
freight_value       0.215359       1.000000
```

Nem sempre existe correlação entre o valor do frete e tempo de entrega. Visto que o valor do frete não é só em relação à distância, mas tambem ao peso e volume dos produtos.

3.Quais áreas geográficas (cidades ou estados) têm os vendedores com os maiores atrasos na entrega?

```
[ ] total_linhas = merged_df_2['seller_state'].value_counts()
    df_filtrado = merged_df_2[merged_df_2['days_behind'] > 0]
    linhas_positivas = df_filtrado['seller_state'].value_counts()
    percentual = (linhas_positivas / total_linhas) * 100
    resultado = pd.DataFrame({
        'Total de Linhas': total_linhas,
        'Total de Linhas Positivas': linhas_positivas,
        'Percentual (%)': percentual
    })
    resultado = resultado.sort_values('Percentual (%)', ascending=False)
    print(resultado)
```

```
    Total de Linhas  Total de Linhas Positivas  Percentual (%)
AM                3                        1.0       33.333333
MA              396                       72.0       18.181818
RN               56                        4.0        7.142857
SP            77372                     5391.0        6.967637
RJ             4586                      306.0        6.672481
CE               90                        6.0        6.666667
DF              874                       53.0        6.064073
PR             8453                      437.0        5.169762
ES              352                       18.0        5.113636
SC             3969                      189.0        4.761905
MG             8378                      391.0        4.666985
BA              624                       27.0        4.326923
MT              144                        6.0        4.166667
MS               49                        2.0        4.081633
RS             2115                       68.0        3.215130
PE              443                       14.0        3.160271
PB               33                        1.0        3.030303
GO              499                       13.0        2.605210
PA                8                        NaN             NaN
PI               11                        NaN             NaN
RO               14                        NaN             NaN
SE               10                        NaN             NaN
```

Amazonas, Maranhão e Rio grande do Norte, tem o maior percentual de vendedores com entregas em atraso, porém os mais significativos são São Paulo, Rio de Janeiro e Paraná, por terem o maior numero de vendas.

4.Em quais cidades moram os compradores que mais são afetados pelos atrasos das entregas.

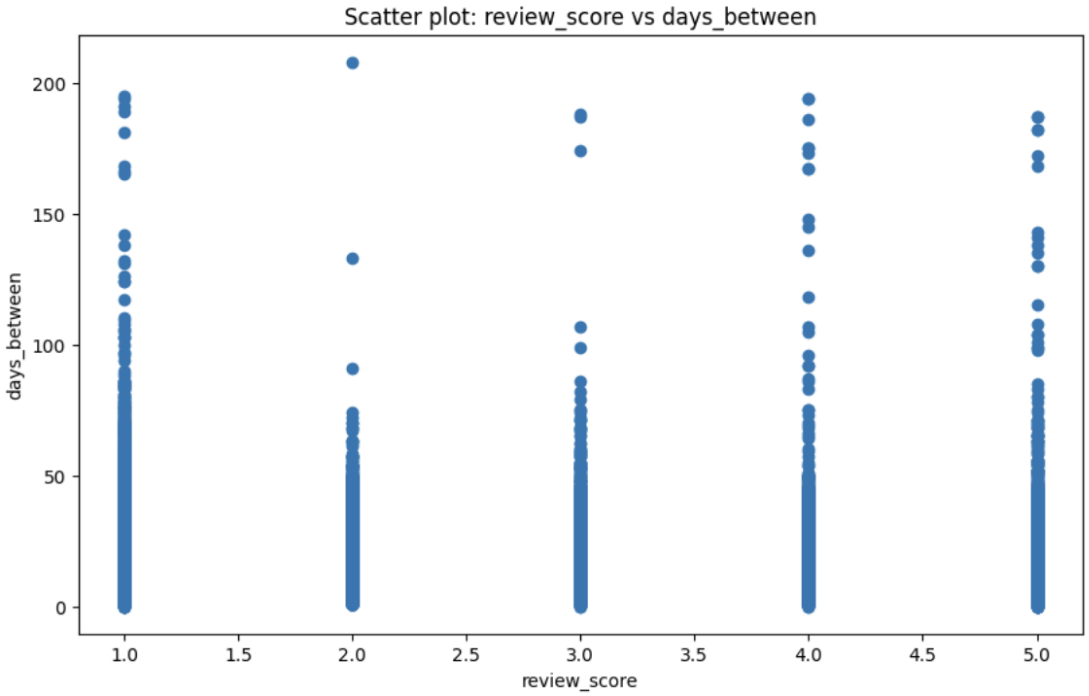
```
total_linhas = merged_df_2['customer_state'].value_counts()
df_filtrado = merged_df_2[merged_df_2['days_behind'] > 0]
linhas_positivas = df_filtrado['customer_state'].value_counts()
percentual = (linhas_positivas / total_linhas) * 100
resultado = pd.DataFrame({
    'Total de Linhas': total_linhas,
    'Total de Linhas Positivas': linhas_positivas,
    'Percentual (%)': percentual
})
resultado = resultado.sort_values('Total de Linhas Positivas', ascending=False)
print(resultado)
```

```
    Total de Linhas  Total de Linhas Positivas  Percentual (%)
SP            45801                       1969        4.299033
RJ            13848                       1591       11.489024
MG            12727                        547        4.297949
BA             3612                        424       11.738649
RS             6067                        349        5.752431
SC             4019                        317        7.887534
ES             2187                        226       10.333791
PR             5571                        213        3.823371
CE             1408                        186       13.210227
PE             1717                        150        8.736168
DF             2332                        141        6.046312
MA              790                        140       17.721519
GO             2209                        131        5.930285
PA             1023                        105       10.263930
AL              424                         86       20.283019
MS              812                         72        8.866995
PI              512                         69       13.476562
PB              572                         61       10.664336
MT             1018                         60        5.893910
SE              371                         59       15.902965
RN              514                         46        8.949416
TO              305                         30        9.836066
RO              266                         11        4.135338
RR               44                          5       11.363636
AM              161                          5        3.105590
AP               80                          3        3.750000
AC               89                          3        3.370787
```

Os estados onde os compradores são mais afetados são AL,SE e CE

5.Como as avaliações (review score) estão relacionadas ao tempo de entrega? Existe alguma correlação entre eles?

```
[ ]  import matplotlib.pyplot as plt
     import seaborn as sns
     plt.figure(figsize=(10, 6))
     plt.scatter(merged_df_2['review_score'], merged_df_2['days_between'])
     plt.title('Scatter plot: review_score vs days_between')
     plt.xlabel('review_score')
     plt.ylabel('days_between')
     plt.show()
```

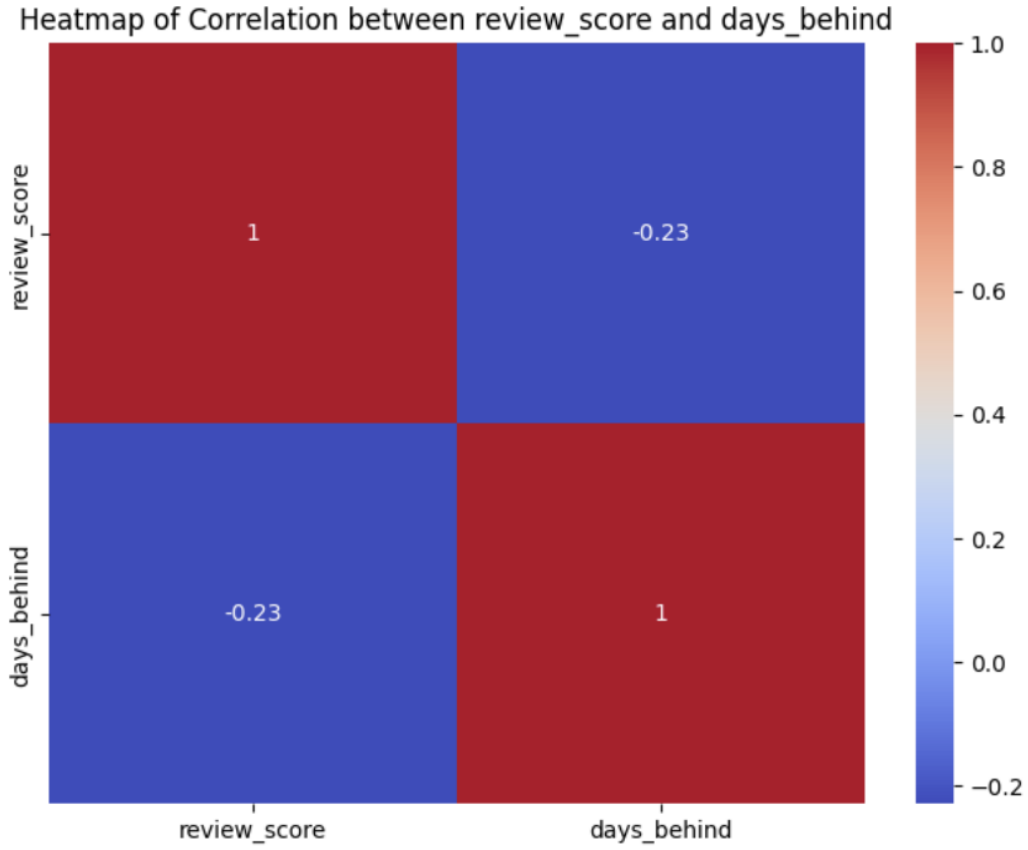


Não parece ter uma relação clara, pois o review score não depende só do tempo de entrega. Depende também da qualidade dos produtos, atendimento ao cliente, embalagem e pós-venda, entre outros.

6.Qual é a correlação entre review score e dias de atraso com as entregas?

```
[ ]  # Calcular a correlação
     corr = merged_df_2[['review_score', 'days_behind']].corr()

     # Criar um mapa de calor
     plt.figure(figsize=(8, 6))
     sns.heatmap(corr, annot=True, cmap='coolwarm')
     plt.title('Heatmap of Correlation between review_score and days_behind')
     plt.show()
```



Não existe correlação entre review score e atraso de entrega.

7.Qual o percentual de clientes da Olist que pode ser considerado satisfeito (review score >= 4)?

```
[ ] percentage = (merged_df_2[merged_df_2['review_score'] >= 4].shape[0] / merged_df_2.shape[0]) * 100
    print(f"O percentual de clientes onde o review_score é maior ou igual a 4 é {percentage:.2f}%")
```

```
O percentual de clientes onde o review_score é maior ou igual a 4 é 76.84%
```

Desempenho da categoria do Produto e dos vendedores:

8.Quais são as categorias de produtos mais vendidos na plataforma da Olist?

```
grouped = merged_df_2.groupby('product_category_name').agg({'review_score': ['count', 'mean']})
grouped.columns = ['count', 'mean_review_score']
grouped = grouped.sort_values(by='count', ascending=False)
grouped
```

product_category_name	count	mean_review_score
cama_mesa_banho	10985	3.920983
beleza_saude	9458	4.190738
esporte_lazer	8436	4.165955
moveis_decoracao	8159	3.950116
informatica_acessorios	7671	3.985139
...	...	...
portateis_cozinha_e_preparadores_de_alimentos	14	3.428571
la_cuisine	13	4.000000
pc_gamer	8	3.625000
fashion_roupa_infanto_juvenil	7	5.000000
seguros_e_servicos	2	2.500000

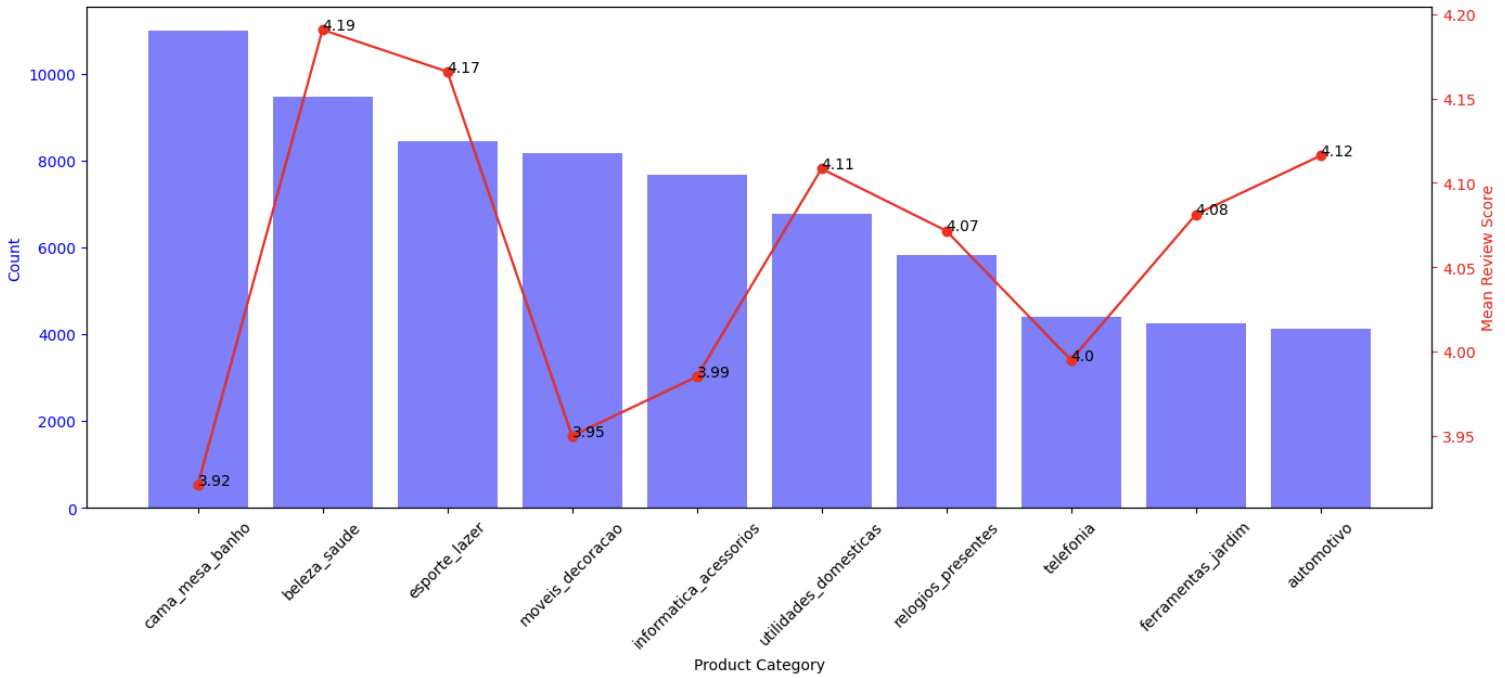
73 rows × 2 columns

```
[ ] top_10 = grouped.head(10)

    fig, ax1 = plt.subplots(figsize=(16,6))

    ax1.bar(top_10.index, top_10['count'], color='b', alpha=0.5)
    ax1.set_xlabel('Product Category')
    ax1.set_ylabel('Count', color='b')
    ax1.tick_params('y', colors='b')
    plt.xticks(rotation=45)
    ax2 = ax1.twinx()
    ax2.plot(top_10.index, top_10['mean_review_score'], color='r', marker="o")
    for i, txt in enumerate(top_10['mean_review_score']):
        ax2.annotate(round(txt, 2), (top_10.index[i], top_10['mean_review_score'].iloc[i]))
    ax2.set_ylabel('Mean Review Score', color='r')
    ax2.tick_params('y', colors='r')

    plt.show()
```



O gráfico apresenta as vendas e as médias de review score de top 10 das categorias de produtos mais vendidos. As barras azuis representam as vendas totais para cada categoria de produto, enquanto a linha vermelha representa a média do review score para cada categoria. A categoria com as maiores vendas é “Cama mesa e banho” e ao mesmo tempo é a categoria com a menor review score das categorias mais vendidas. A categoria com maior review score, resultou ser "beleza e saude" sendo a segunda categoria mais vendida a segunda.

9.Quais são as categorias de produtos com menor review score na plataforma da Olist?

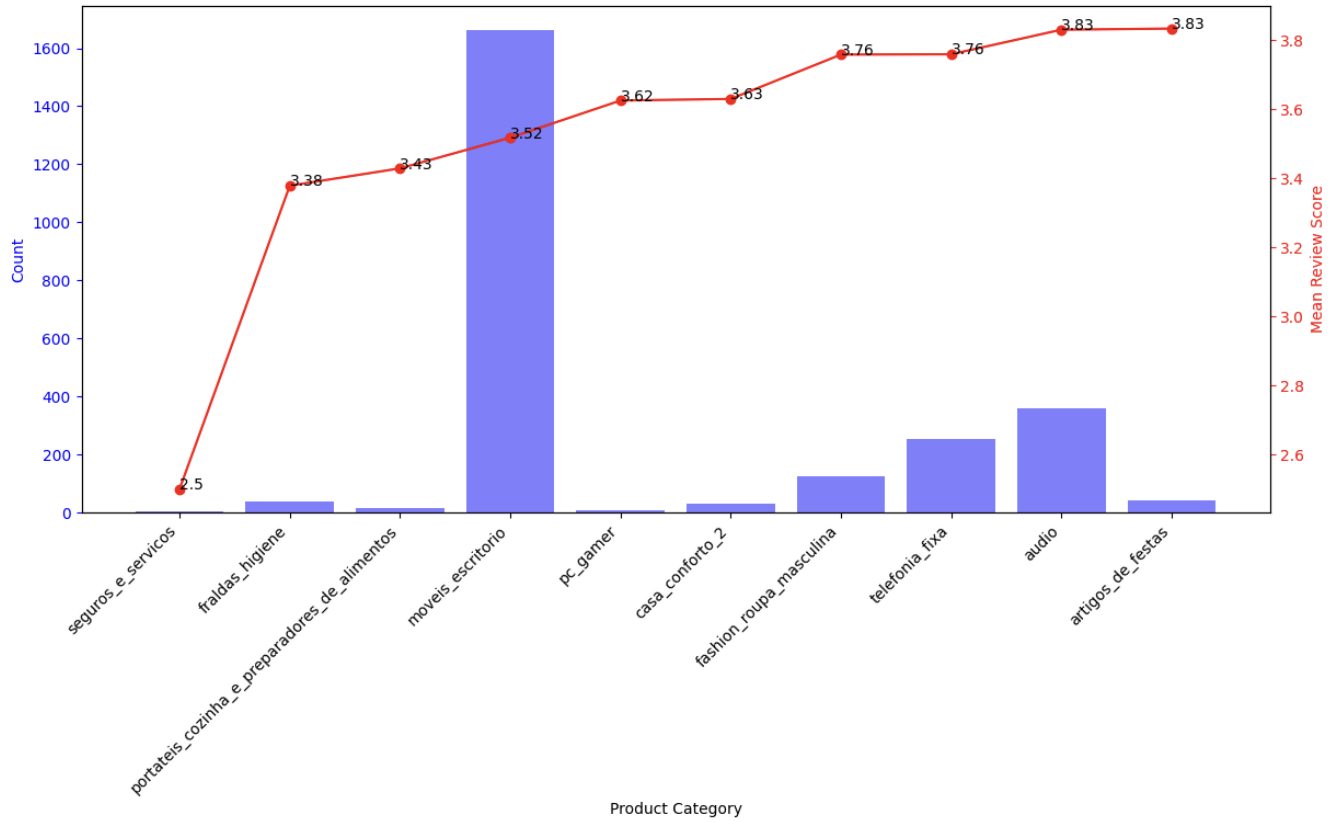
```
[ ] grouped1 = merged_df_2.groupby('product_category_name').agg({'review_score': ['count', 'mean']})
    grouped1.columns = ['count', 'mean_review_score']
    grouped1 = grouped1.sort_values(by='mean_review_score', ascending=True)

    top_10_1 = grouped1.head(10)

    fig, ax1 = plt.subplots(figsize=(14,6))

    ax1.bar(top_10_1.index, top_10_1['count'], color='b', alpha=0.5)
    ax1.set_xlabel('Product Category')
    ax1.set_ylabel('Count', color='b')
    ax1.tick_params('y', colors='b')
    plt.xticks(rotation=45,ha="right")
    ax2 = ax1.twinx()
    ax2.plot(top_10_1.index, top_10_1['mean_review_score'], color='r', marker="o")
    for i, txt in enumerate(top_10_1['mean_review_score']):
        ax2.annotate(round(txt, 2), (top_10_1.index[i], top_10_1['mean_review_score'].iloc[i]))
    ax2.set_ylabel('Mean Review Score', color='r')
    ax2.tick_params('y', colors='r')

    plt.show()
```



O gráfico apresenta as vendas e as médias de review score das 10 categorias com pior review score. Em geral os piores review score pertencem a categorias com baixo numero de vendas. Exeto a categoria moveis_escritorio que mesmo tendo um review baixo vende muito.

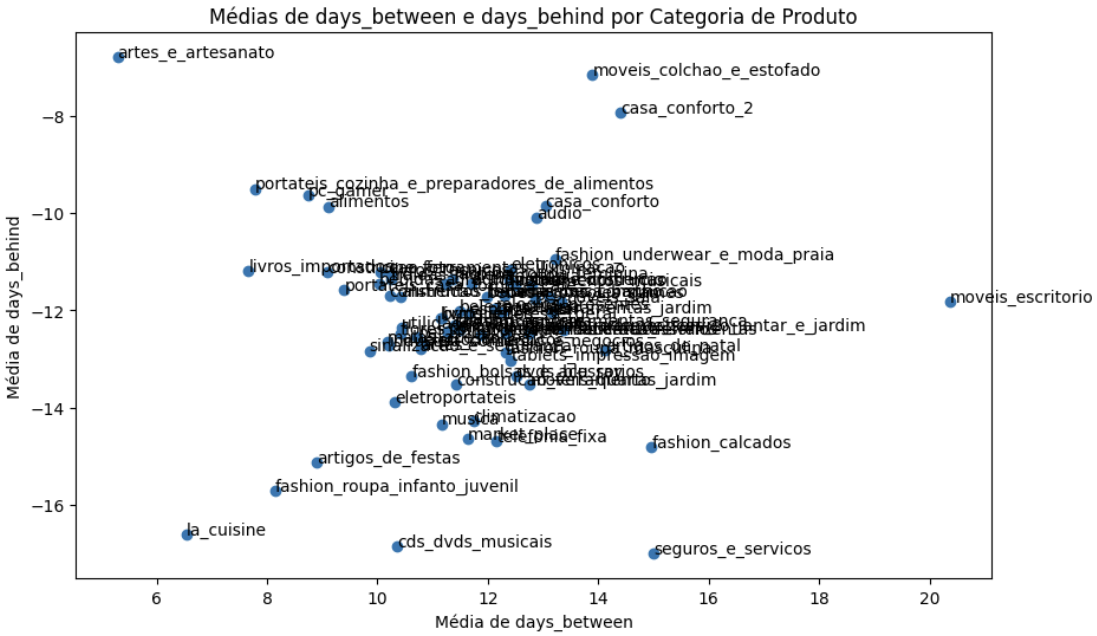
10.Existe uma correlação entre a categoria do produto (como "esporte e lazer", "utilidades domésticas" etc.) e o tempo médio de entrega (days_between)? Algumas categorias de produtos têm tempos de entrega significativamente mais longos ou curtos do que outras?

```
[ ]
    # Calcule as médias
    grouped = merged_df_2.groupby('product_category_name').agg({'days_between': 'mean', 'days_behind': 'mean'})

    # Crie o gráfico de dispersão
    plt.figure(figsize=(10,6))
    plt.scatter(grouped['days_between'], grouped['days_behind'])

    # Adicione anotações para cada ponto
    for i in range(len(grouped)):
        plt.text(grouped.iloc[i, 0], grouped.iloc[i, 1], grouped.index[i])

    plt.xlabel('Média de days_between')
    plt.ylabel('Média de days_behind')
    plt.title('Médias de days_between e days_behind por Categoria de Produto')
    plt.show()
```



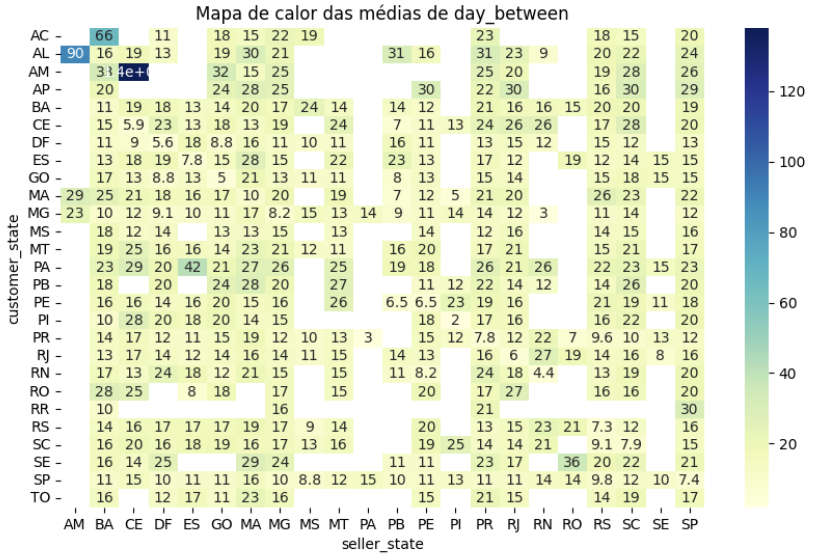
Existem categorias que demoram mais a ser entregues, como por exemplo moveis_escritorio e categorias que demoram menos como La cuisine. Na média, os produtos de todas as categorias são entregues sem atraso, mesmo antes da data prevista.

11.Como a distância entre a cidade do cliente (customer_state) e a cidade do vendedor (seller_state) afeta a eficiência logística? Existem categorias de produtos que são frequentemente enviadas de estados ou cidades distantes, resultando em frete mais alto ou maior tempo de entrega?

```
[ ] matrix = merged_df_2.pivot_table(index='customer_state', columns='seller_state', values='days_between', aggfunc='mean')

    plt.figure(figsize=(10, 6))
    sns.heatmap(matrix, annot=True, cmap="YlGnBu")

    plt.title('Mapa de calor das médias de day_between')
    plt.show()
```

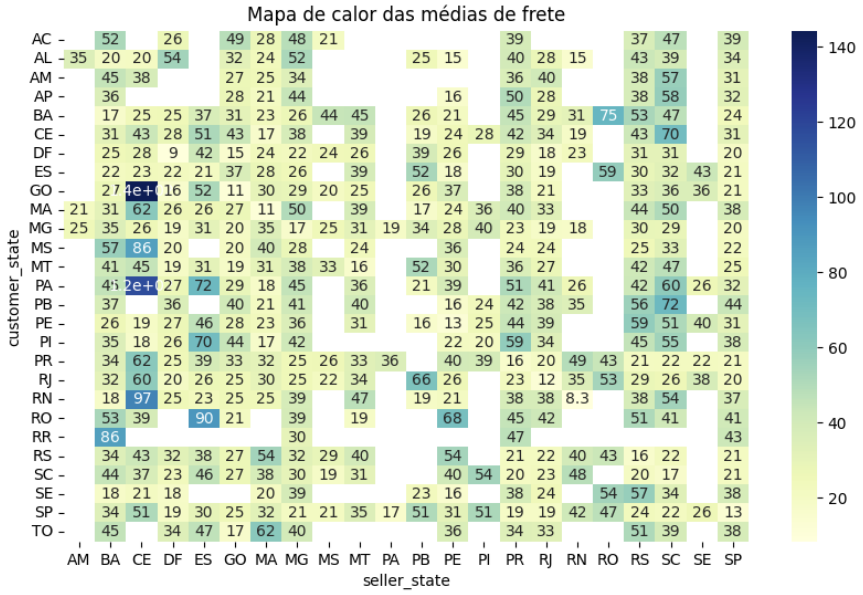


Existem alguns pares de estados com tempos de entrega particularmente altos. Por exemplo, pedidos de clientes em AL (Alagoas) para vendedores em AM (Amazonas) mostram um valor de 90. Em contraste, outros pares interestaduais têm tempos de entrega mais curtos. Por exemplo, pedidos de SP (São Paulo) para vendedores em RJ (Rio de Janeiro) têm uma média de 11.

```
[ ] matrix = merged_df_2.pivot_table(index='customer_state', columns='seller_state', values='freight_value', aggfunc='mean')

    plt.figure(figsize=(10, 6))
    sns.heatmap(matrix, annot=True, cmap="YlGnBu")

    plt.title('Mapa de calor das médias de frete')
    plt.show()
```



Quanto maior a distância entre cliente e vendedor, geralmente maior é o frete. O mapa mostra frete mais baixo entre estados próximos e mais alto entre estados distantes.

Auto-avaliação

Durante o desenvolvimento do meu MVP sobre o estudo da eficiência logística e satisfação do cliente no e-commerce da Olist, identifiquei com sucesso várias questões relevantes, abrangendo desde tempos médios de entrega até correlações entre variáveis. No entanto, ao abordar as questões 12,13 e 14 percebi que enfrentava desafios mais complexos, principalmente relacionados à disponibilidade de dados, análise multivariada e à detecção de tendências sazonais. A resposta a essas questões exige uma análise mais aprofundada, potencialmente envolvendo técnicas avançadas de machine learning e estatística. Também reconheço que pode haver uma necessidade de dados mais detalhados ou específicos para abordar com precisão tais questões. Este exercício ressaltou a importância de continuar aprimorando minhas habilidades em análise de dados e considerar a colaboração com especialistas em áreas que encontro mais desafiantes.