Xử lý dữ liệu data OLIST

```
import pandas as pd
import os
from google.colab import drive
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
```

```
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
## Import dữ liệu từ drive
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
customers=pd.read_excel('/content/drive/MyDrive/OLIST/datafinal.xlsx','customers')
geolocation=pd.read_excel('/content/drive/MyDrive/OLIST/datafinal.xlsx','geolocation')
order_items=pd.read_excel('/content/drive/MyDrive/OLIST/datafinal.xlsx','order_items')
order_payments=pd.read_excel('/content/drive/MyDrive/OLIST/datafinal.xlsx','order_payments')
order_reviews=pd.read_excel('/content/drive/MyDrive/OLIST/datafinal.xlsx','order_reviews')
orders=pd.read_excel('/content/drive/MyDrive/OLIST/datafinal.xlsx','orders')
products=pd.read_excel('/content/drive/MyDrive/OLIST/datafinal.xlsx','products')
```

```
customers.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Data columns (total 5 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   customer_id               99441 non-null  object
 1   customer_unique_id        99441 non-null  object
 2   customer_zip_code_prefix  99441 non-null  int64 
 3   customer_city             99441 non-null  object
 4   customer_state            99441 non-null  object
dtypes: int64(1), object(4)
memory usage: 3.8+ MB
```

```
geolocation.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27912 entries, 0 to 27911
Data columns (total 3 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   geolocation_zip_code_prefix  27912 non-null  int64 
 1   geolocation_city             27912 non-null  object
 2   geolocation_state            27912 non-null  object
dtypes: int64(1), object(2)
memory usage: 654.3+ KB
```

```
order_items.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 112650 entries, 0 to 112649
Data columns (total 7 columns):
 #   Column               Non-Null Count   Dtype         
---  ------               --------------   -----         
 0   order_id             112650 non-null  object        
 1   order_item_id        112650 non-null  int64         
 2   product_id           112650 non-null  object        
 3   seller_id            112650 non-null  object        
 4   shipping_limit_date  112650 non-null  datetime64[ns]
 5   price                112650 non-null  int64         
 6   freight_value        112650 non-null  int64         
dtypes: datetime64[ns](1), int64(3), object(3)
memory usage: 6.0+ MB
```

```
order_payments.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103886 entries, 0 to 103885
Data columns (total 5 columns):
 #   Column                Non-Null Count   Dtype 
---  ------                --------------   ----- 
 0   order_id              103886 non-null  object
 1   payment_sequential    103886 non-null  int64 
 2   payment_type          103886 non-null  object
 3   payment_installments  103886 non-null  int64 
 4   payment_value         103886 non-null  int64 
dtypes: int64(3), object(2)
memory usage: 4.0+ MB
```

```
order_reviews.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99224 entries, 0 to 99223
Data columns (total 7 columns):
 #   Column                   Non-Null Count  Dtype         
---  ------                   --------------  -----         
 0   review_id                99224 non-null  object        
 1   order_id                 99224 non-null  object        
 2   review_score             99224 non-null  int64         
 3   review_comment_title     11565 non-null  object        
 4   review_comment_message   40947 non-null  object        
 5   review_creation_date     99224 non-null  datetime64[ns]
 6   review_answer_timestamp  99224 non-null  datetime64[ns]
dtypes: datetime64[ns](2), int64(1), object(4)
memory usage: 5.3+ MB
```

```
orders.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99441 entries, 0 to 99440
Data columns (total 8 columns):
 #   Column                         Non-Null Count  Dtype         
---  ------                         --------------  -----         
 0   order_id                       99441 non-null  object        
 1   customer_id                    99441 non-null  object        
 2   order_status                   99441 non-null  object        
 3   order_purchase_timestamp       99441 non-null  datetime64[ns]
 4   order_approved_at              99441 non-null  datetime64[ns]
 5   order_delivered_carrier_date   99441 non-null  datetime64[ns]
 6   order_delivered_customer_date  99441 non-null  datetime64[ns]
 7   order_estimated_delivery_date  99441 non-null  datetime64[ns]
dtypes: datetime64[ns](5), object(3)
memory usage: 6.1+ MB
```

```
products.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32951 entries, 0 to 32950
Data columns (total 9 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   product_id                  32951 non-null  object 
 1   product_category_name       32341 non-null  object 
 2   product_name_lenght         32341 non-null  float64
 3   product_description_lenght  32341 non-null  float64
 4   product_photos_qty          32341 non-null  float64
 5   product_weight_g            32949 non-null  float64
 6   product_length_cm           32949 non-null  float64
 7   product_height_cm           32949 non-null  float64
 8   product_width_cm            32949 non-null  float64
dtypes: float64(7), object(2)
memory usage: 2.3+ MB
```

```
order_reviews[['review_comment_title', 'review_comment_message']] = order_reviews[['review_comment_title', 'review_comment_message']].f
order_reviews.head()
```

	review_id	order_id	review_score	review_comment_title	review_comment_message
0	7ea98afafdbe948b63e8007a5cba824c	723f4709e0dc0b2fc27c8d68792c2ba6	5	No comment	No comment
1	3279a7d666c135d654d1c1a395b34d20	b7d87e06832b932da4cd4ae7dee2974f	5	No comment	No comment
2	0300525ef63e9c77a5a798fd312e13f1	08fc1aa23fef732befd3559729544a6a	5	No comment	No comment
3	936cc8ba40587f4c477d6c538e924012	a58e64aaf179aee88af27f932d4da7b2	5	No comment	No comment
4	396a511f928dcbb73efeebab1586eab0	4e839b1ce670c701bd83a9d9aeb58969	5	No comment	No comment

Next steps: Generate code with order_reviews   View recommended plots   New interactive sheet

```
products = products.dropna(subset=['product_category_name'])
products.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 32341 entries, 0 to 32340
Data columns (total 9 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   product_id                  32341 non-null  object 
 1   product_category_name       32341 non-null  object 
 2   product_name_lenght         32341 non-null  float64
 3   product_description_lenght  32341 non-null  float64
 4   product_photos_qty          32341 non-null  float64
 5   product_weight_g            32340 non-null  float64
 6   product_length_cm           32340 non-null  float64
 7   product_height_cm           32340 non-null  float64
 8   product_width_cm            32340 non-null  float64
dtypes: float64(7), object(2)
memory usage: 2.5+ MB
```

```
null_values_df = products[products[['product_weight_g', 'product_length_cm','product_height_cm','product_width_cm']].isnull().any(axis=1
null_values_df
```

	product_id	product_category_name	product_name_lenght	product_description_lenght	product_photos_qty	product_weight_g	product_

```
# Chỉ chọn các cột số
numeric_columns = ['product_weight_g', 'product_length_cm', 'product_height_cm', 'product_width_cm']

# Tính giá trị trung bình cho các cột số trong category "bebes"
mean_values = products[products['product_category_name'] == 'bebes'][numeric_columns].mean()

# Áp dụng giá trị trung bình cho các giá trị null
for column in numeric_columns:
    products.loc[(products['product_category_name'] == 'bebes') & (products[column].isnull()), column] = mean_values[column]
products.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 32341 entries, 0 to 32340
Data columns (total 9 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   product_id                  32341 non-null  object 
 1   product_category_name       32341 non-null  object 
 2   product_name_lenght         32341 non-null  float64
 3   product_description_lenght  32341 non-null  float64
 4   product_photos_qty          32341 non-null  float64
 5   product_weight_g            32341 non-null  float64
 6   product_length_cm           32341 non-null  float64
 7   product_height_cm           32341 non-null  float64
 8   product_width_cm            32341 non-null  float64
dtypes: float64(7), object(2)
memory usage: 3.5+ MB
```