

ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ





**BÁO CÁO ĐỒ ÁN**

**Học kỳ I, năm học 2021 - 2022**

**Học phần:**

Phân tích dữ liệu với R

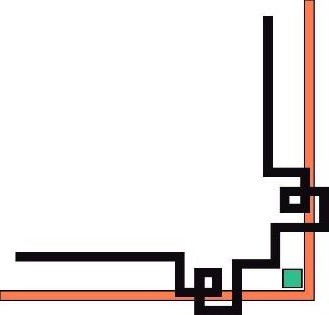
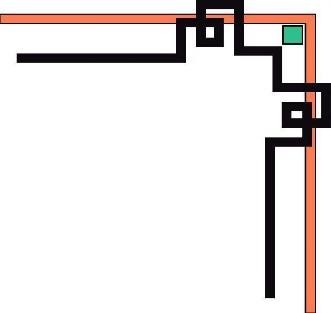
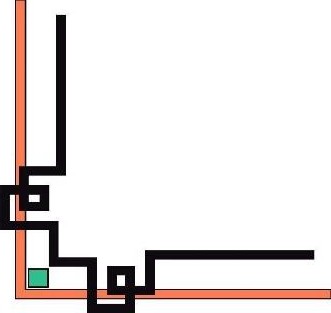
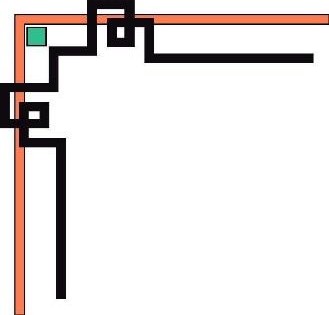
***Đề tài:*** Lazada - Ecommerce

(Sản phẩm áo Hoodie)

**Số phách**

*(Do hội đồng chấm thi ghi)*

Thừa Thiên Huế, ngày 12 tháng 01 năm 2022



ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ





**BÁO CÁO ĐỒ ÁN**

**Học kỳ I, năm học 2021 - 2022**

**Học phần:**

Phân tích dữ liệu với R

***Đề tài:*** Lazada - Ecommerce

(Sản phẩm áo Hoodie)

**Giảng viên hướng dẫn: Hồ Quốc Dũng**

**Sinh viên thực hiện: Lê Thị Ngọc Thắm**

**Lớp: Khoa học dữ liệu và trí tuệ nhân tạo**

**Số phách**

*(Do hội đồng chấm thi ghi)*

Thừa Thiên Huế, ngày 12 tháng 01 năm 2022

ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ



PHIẾU ĐÁNH GIÁ ĐỒ ÁN/TIỂU LUẬN/BÀI TẬP LỚN

**Học kỳ I, năm học 2021 - 2022**

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1** | **Cán bộ chấm thi 2** |
| **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBCT1:**  Bằng số: .........................................  Bằng chữ: ....................................... | **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBCT2:**  Bằng số: .........................................  Bằng chữ: ....................................... |

Điểm kết luận: ...........................................................................................................

Bằng số:.....................................................................................................................

Bằng chữ: .................................................................................................................

*Thừa Thiên Huế, ngày tháng năm 2021*

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1**  *(Ký và ghi rõ họ và tên)* | **Cán bộ chấm thi 2**  *(Ký và ghi rõ họ và tên)* |

**MỤC LỤC**

[I. MÔ TẢ VỀ DỮ LIỆU SỬ DỤNG CRAWL LẤY VỀ. 1](#_Toc92788037)

[1. Dữ liệu crawl lấy về: 1](#_Toc92788038)

[2. Số lượng dòng dữ liệu: 1](#_Toc92788039)

[3. Danh sách các trường dữ liệu : 1](#_Toc92788040)

[II. PHÂN TÍCH THỐNG KÊ MÔ TẢ DỮ LIỆU 1](#_Toc92788041)

[1.Chuẩn bị dữ liệu 1](#_Toc92788042)

[2.Tính các chỉ số thống kê mô tả 1](#_Toc92788043)

[3.Vẽ các biểu đồ trực quan hóa dữ liệu 3](#_Toc92788044)

[4.Một số nhận xét về số liệu 6](#_Toc92788045)

[III. PHÂN TÍCH DỰ ÁN KAGGLE 7](#_Toc92788046)

[1.Chuẩn bị dữ liệu: 7](#_Toc92788047)

[2.Xác định khung dữ liệu 7](#_Toc92788048)

[3. df\_reviews (Dataframe reviews) 8](#_Toc92788049)

[4. Data Cleaning 8](#_Toc92788050)

[IV. Visualisasi Data(Trực quan hóa dữ liệu) 15](#_Toc92788051)

# 

**I. MÔ TẢ VỀ DỮ LIỆU SỬ DỤNG CRAWL LẤY VỀ**.

**1. Dữ liệu crawl lấy về:**

Dữ liệu được lấy là thông tin về mặt hàng các sản phẩm áo Hoodie bán trên trang web Lazada.vn

**2. Số lượng dòng dữ liệu:**

Có tất cả 100 dòng dữ liệu.

**3. Danh sách các trường dữ liệu :**

Tên cột:

Thương hiệu (Brand)

Tỷ lệ giao hàng đúng hạn (delivered\_on\_time)

Thời gian giao hàng (delivery\_time)

Tỷ lệ giảm giá (discount\_price)

Ảnh sản phẩm (image)

Giá gốc sản phẩm (original\_price)

Tên sản phẩm (p\_name)

Tỷ lệ đánh giá tích cực (p\_rating)

Giá sản phẩm (price)

Tỷ lệ phản hồi (response\_rate)

Tên nhà bán hàng (s\_name)

Phí vận chuyển (shipping\_fee)

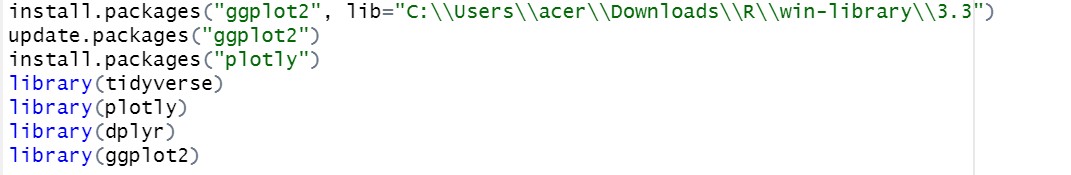
Tổng số lượng đánh giá (total\_reviews)

Kiểu dữ liệu số, văn bản

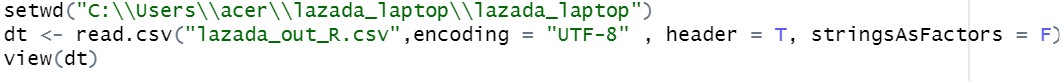
**II. PHÂN TÍCH THỐNG KÊ MÔ TẢ DỮ LIỆU**

**1.Chuẩn bị dữ liệu**

Sử dụng các thư viện

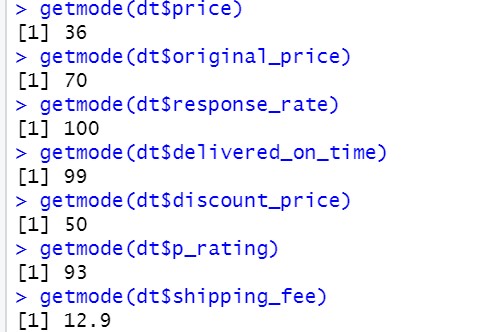
****

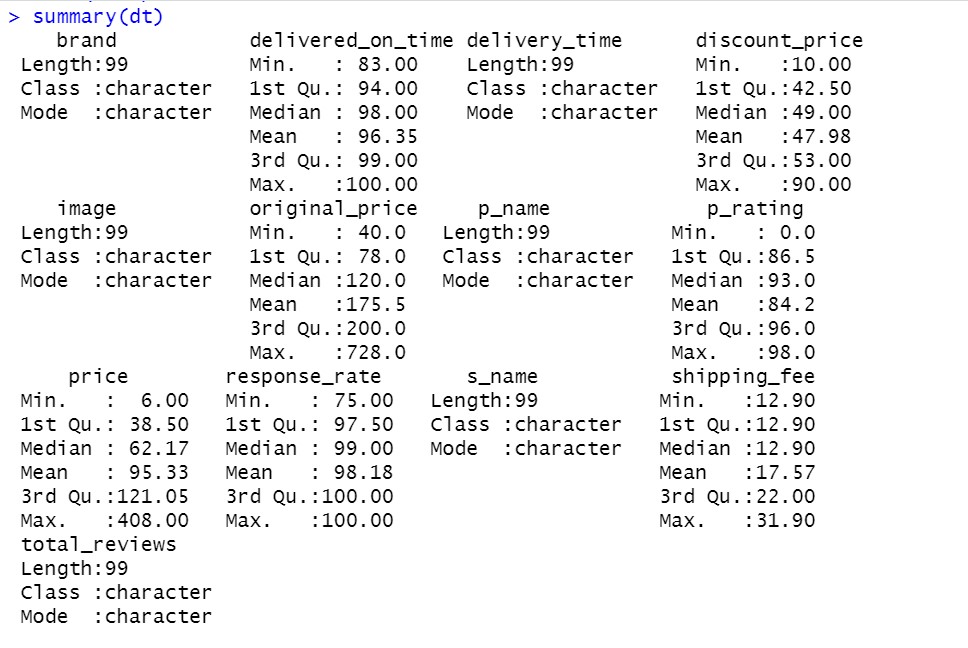
Đọc file dữ liệu:



**2.Tính các chỉ số thống kê mô tả**

Tổng quan về dữ liệu:

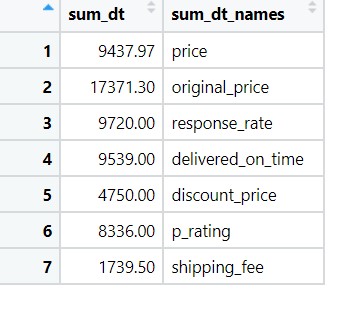
****



Trong đó:

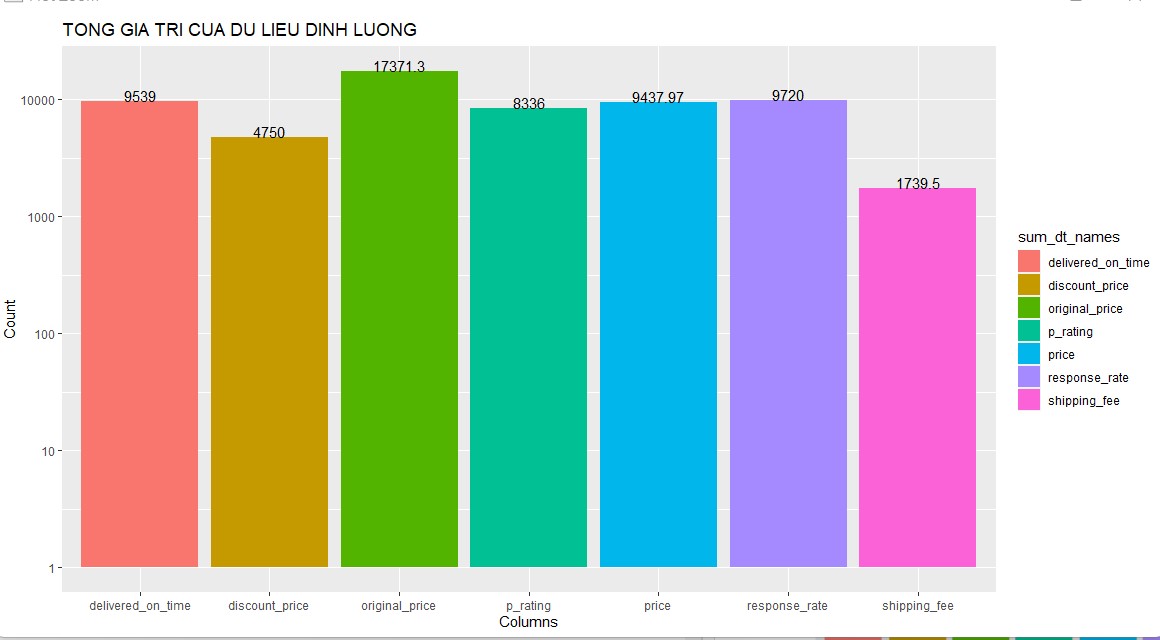
* Mean: trung bình. - Mode: yếu vị.
* Median: trung vị. - Var: phương sai.
* Sd: độ lệch chuẩn. - Min: giá trị thấp nhất.
* Max: giá trị lớn nhất. - Range: toàn cự.

Tổng của các cột trong tập dữ liệu:

****

**3.Vẽ các biểu đồ trực quan hóa dữ liệu**

* Biểu đồ 1: Biểu đồ cột

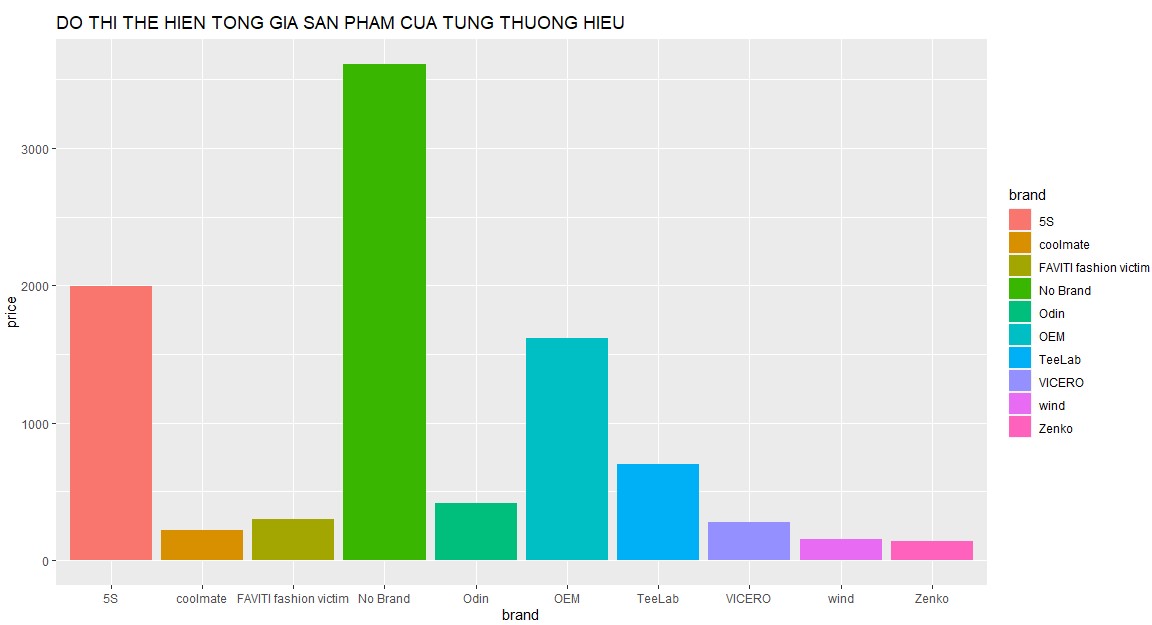
****

Biểu đồ thể hiển tổng giá trị của từng dữ liệu định lượng của tập dữ liệu .

Cho biết tổng tỷ lệ giao hàng đúng hạn là 9539, tỷ lệ giảm giá là 4750, tỷ lệ phản hồi là 9720,

giá gốc là 173713, đánh giá tích cực là 8336, giá bán là 943794 và phí vận chuyển là 1739.5.

* Biểu đồ 2: Biểu đồ cột



Biểu đồ thể hiện tổng giá sản phẩm của từng thương hiệu

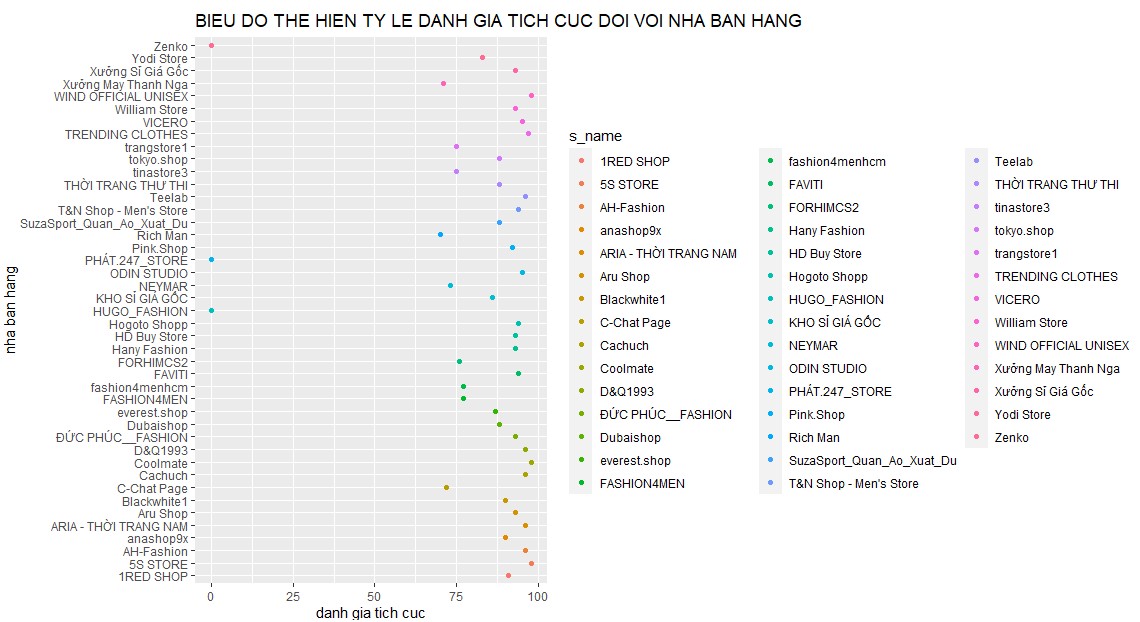
Qua biểu đồ tên cho thấy được tổng giá sản phẩm cao nhất là của thương hiệu No Brand là

hơn 3000đ, thấp nhất là Wind và Zenko dưới 1000đ.

Qua những số liệu trên thì cho ta biết được số lượng bán các mặt hàng sản phẩm của No

Brand nhiều hơn so với các thương thiệu còn lại.

* Biểu đồ 3: Biểu đồ điểm



Biểu đồ thể hiện tỷ lệ đánh giá tích cực của khách hàng đối với nhà bán hàng

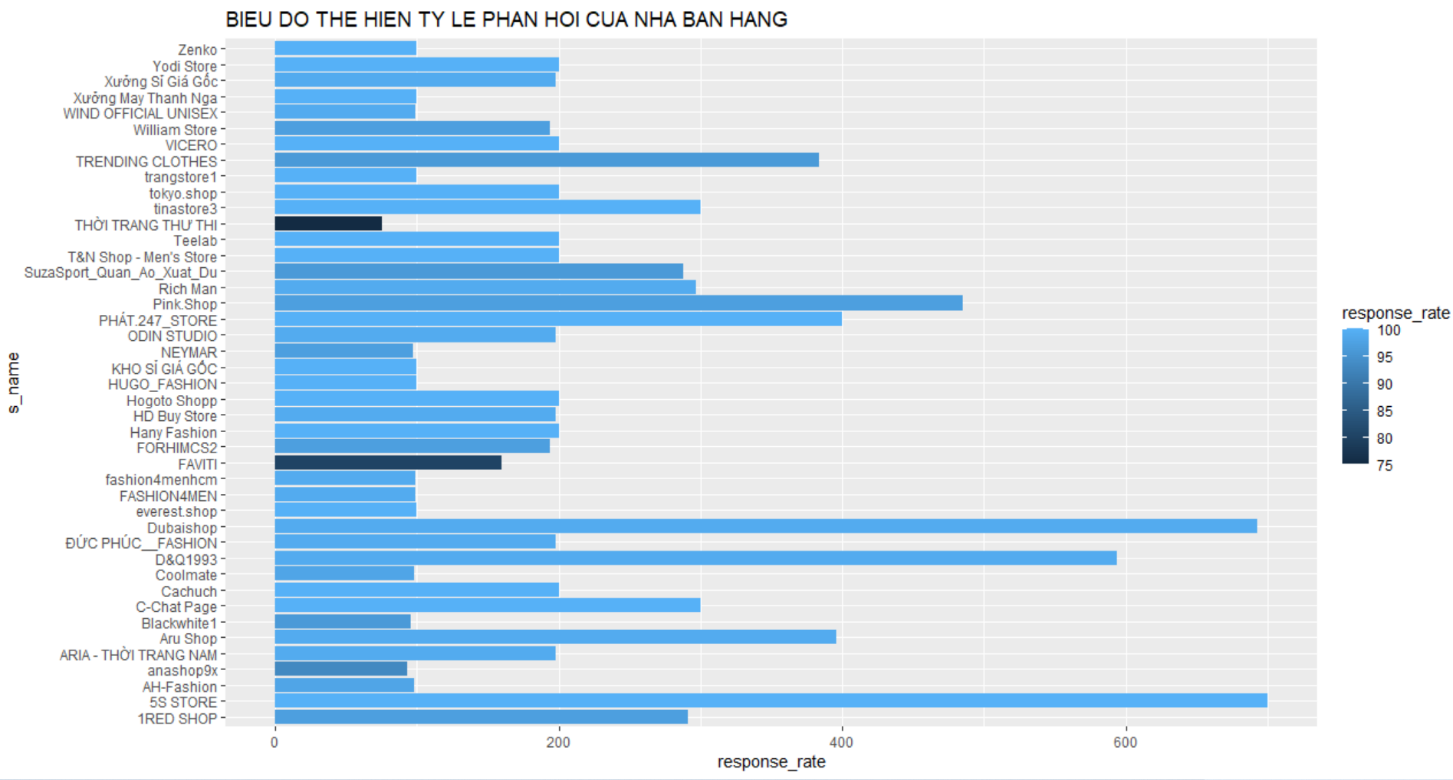
Nhìn vào biểu đồ trên ta có thể biết được hầu hết tỷ lệ đánh giá tích của khách hàng

đối với nhà bán hàng từ khoảng 74%-98%, 2 thương hiệu không có tỷ lệ đánh giá tích

cực đó là Phát.247\_STORE.

Qua đó cho thấy được các sự hài lòng của khách hàng đối với sản phẩm và nhà bán hàng.

* Biểu đồ 4: Biểu đồ cột



Biểu đồ thể hiện tỷ lệ phản hồi của nhà bán hàng đối với khách hàng

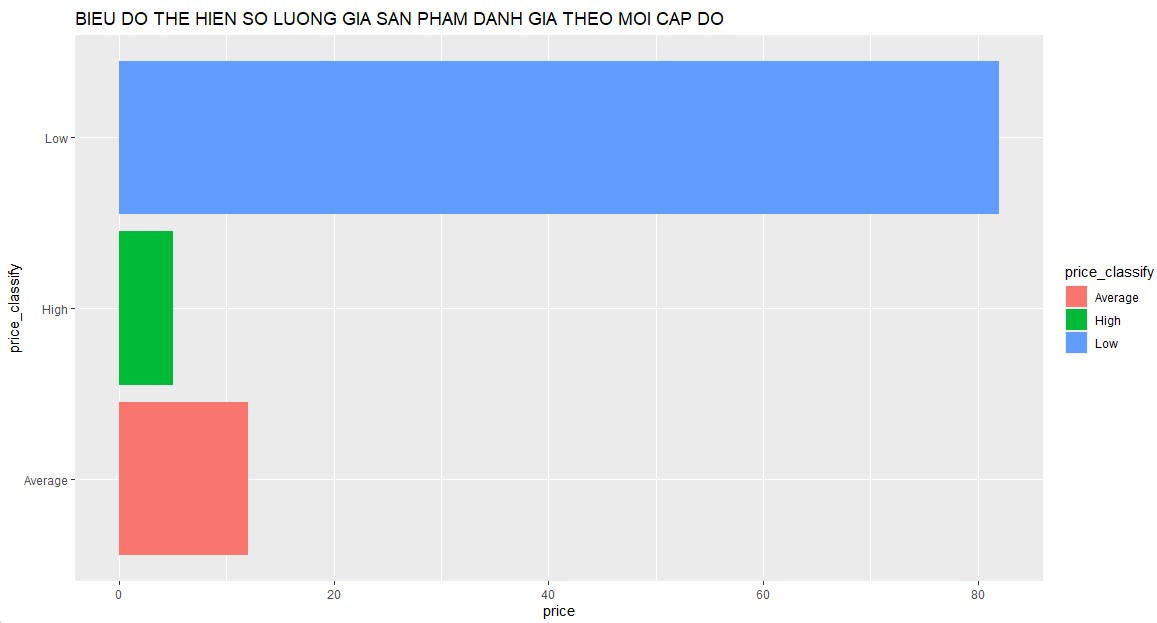
Nhìn vào biểu đồ có thể biết được tổng tỷ lệ phản hồi của nhà bán hàng đối với khách

hàng cao nhất là 5S STORE, Dubaishop và thấp nhất là Thời trang thư thi và Fativi.

Qua biểu đồ đó ta có thể biết được lượng tương tác, phản hồi của các nhà bán hàng

đối với khách hàng.

* Biểu đồ 5: Biểu đồ cột



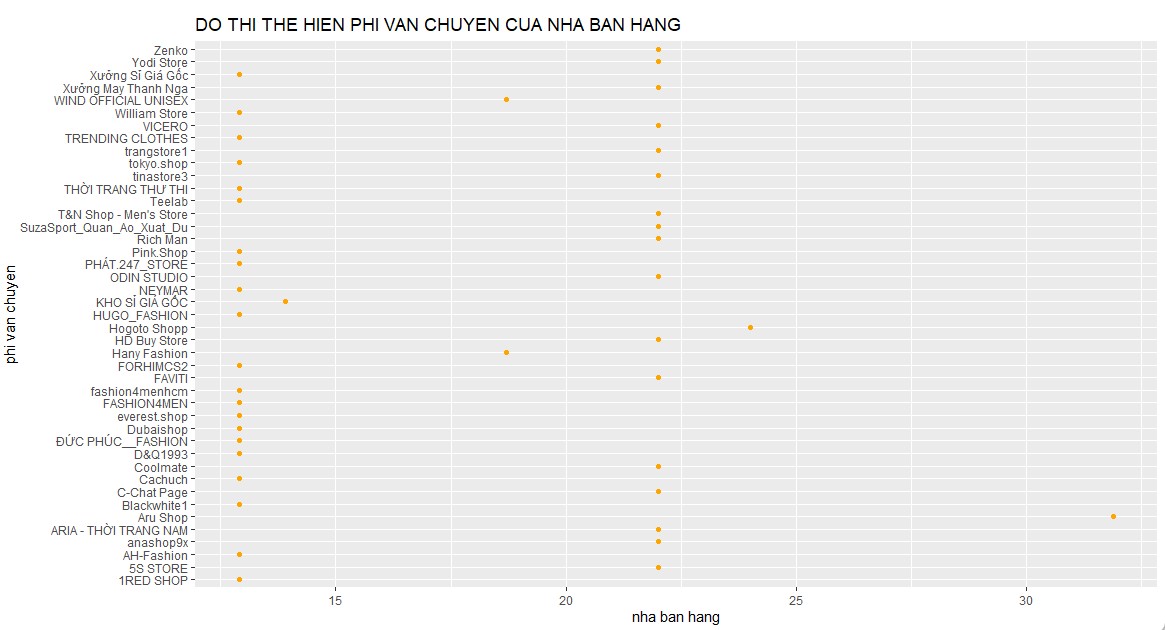
Biểu đồ thể hiện số lượng giá sản phẩm theo mỗi cấp độ

Giá sản phẩm <=150 thấp , > 150 đến <= 300 là trung bình, > 300 là cao

Nhìn vào biểu đồ thì có thể thấy được giá sản phẩm thấp có 82 sản phẩm, giá sản

phẩm trung bình có 12 sản phẩm, giá sản phẩm cao có 5 sản phẩm.

* Biểu đồ 6: Biểu đồ điểm

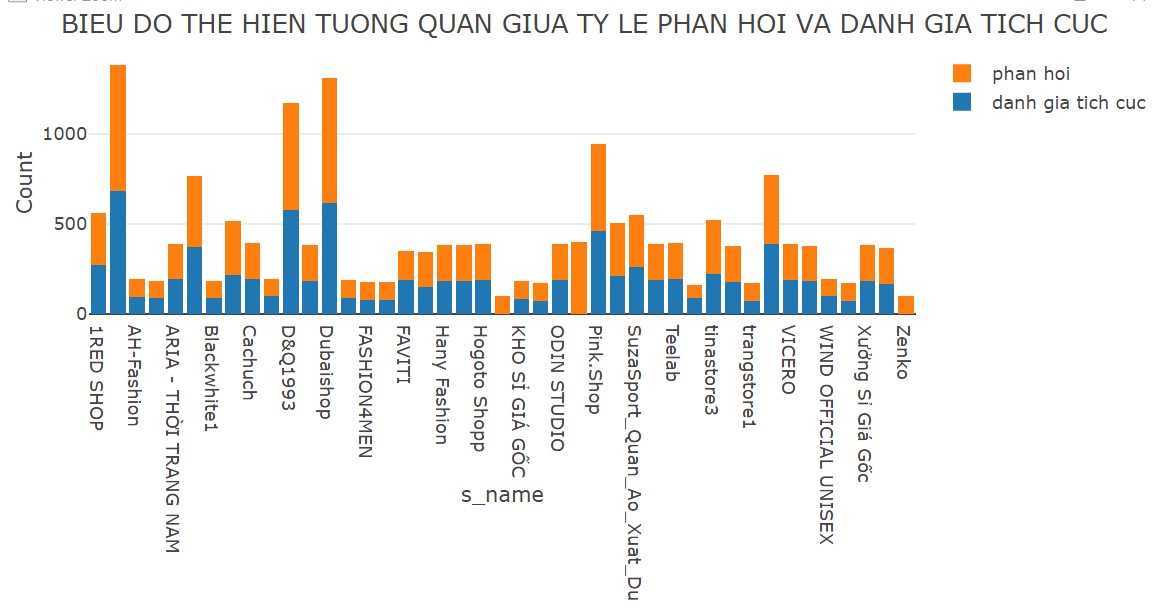


Biểu đồ thể hiện phí vận chuyển của nhà bán hàng đối với khách hàng

Từ biểu đồ trên có thể biết được hầu hết phí vận chuyển mà các nhà bán hàng thu là

12đ và 23đ và cao nhất nhà bán hàng Aru Shop với phí là 32 đ.

* Biểu đồ 7: Biểu đồ cột chồng



Biểu đồ thể hiện tương quan giữa tỷ lệ phản hồi và đánh giá tích cực.

Từ biểu đồ trên có thể thấy được tỷ lệ phản hồi của nhà bán hàng và tỷ lệ đánh giá tích

cực của khách hàng không đồng đều, nhà bán hàng không có đánh giá tích cực là: Zenko, Hugo Shop, Phát.247\_Store.

**4.Một số nhận xét về số liệu**

* Dữ liệu các sản phẩm áo Hoodie sau khi crawl về qua xử lý và phân tích thì có một
* số nhận xét như:
* Các tỷ lệ phần trăm như giao hàng đúng hạn, đánh giá tích cực và tỷ lệ phản hồi đều
* có tỷ lệ từ 85% trở lên.
* Tổng số lượng đánh giá của mỗi sản phẩm không đồng đều.
* Mỗi thương hiệu và tên nhà bán hàng có sự lặp lại qua đó thấy được số lượng mỗi sản phẩm bán trên lazada khá nhiều.

**III. PHÂN TÍCH DỰ ÁN KAGGLE**

**1.Chuẩn bị dữ liệu:**

Sử dụng các thư viện :



Các chức năng của một số thư viện được sử dụng bao gồm:

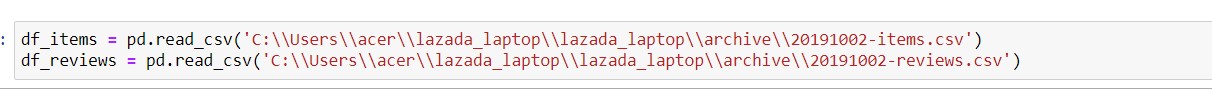
Pandas : Thư viện này được sử dụng để chuyển đổi tập dữ liệu thành khung dữ liệu / bảng để được xử lý và hiển thị

Numpy : Thư viện xử lý dữ liệu số được sử dụng để xử lý và tính toán dữ liệu

Matplotlib : Thư viện cơ bản để trực quan hóa dữ liệu

Seaborn : Một thư viện hoạt động trên đầu thư viện Matplotlib được sử dụng cho các quá trình trực quan hóa dữ liệu. Thư viện này có các tính năng bổ sung như bầy đàn, v.v.

Quá trình tiếp theo là tải tập dữ liệu được sử dụng sang Python bằng cách sử dụng pandas.read\_csv()



**2.Xác định khung dữ liệu**

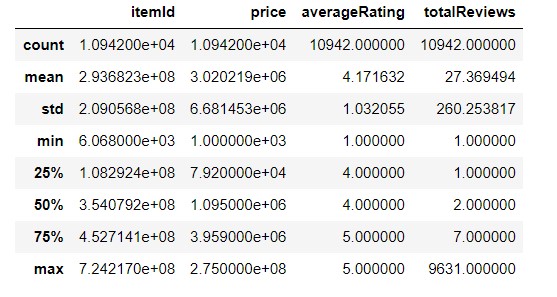
Giai đoạn này nhằm mục đích xác định từng khung dữ liệu để có được ý tưởng về khung dữ liệu được sử dụng và những gì có thể được thực hiện từ khung dữ liệu dựa trên ý định.

**df\_items (Mục khung dữ liệu)**

Điều đầu tiên cần làm là mô tả khung dữ liệu để lấy thông tin về thành phần và nội dung của các cột và hàng của khung dữ liệu.

Đọc file:



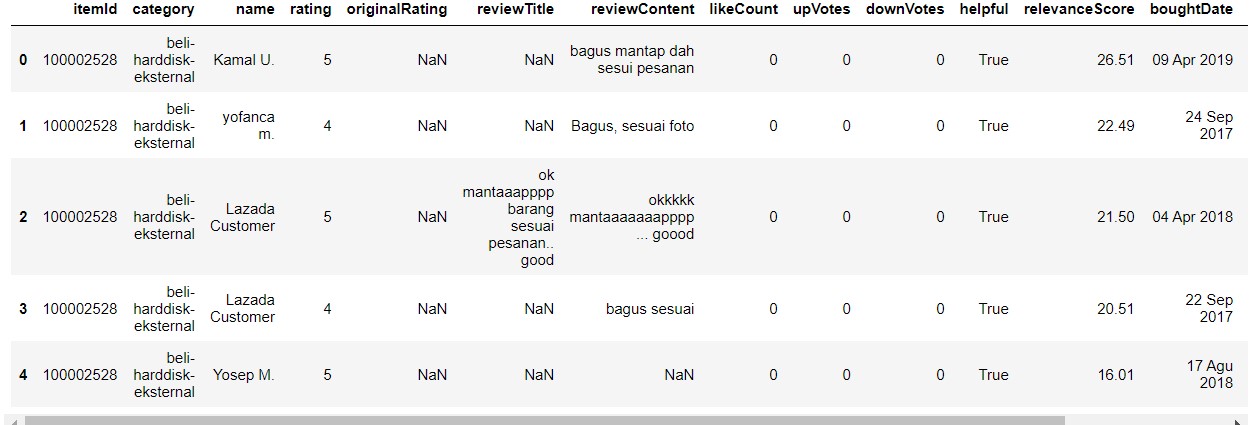


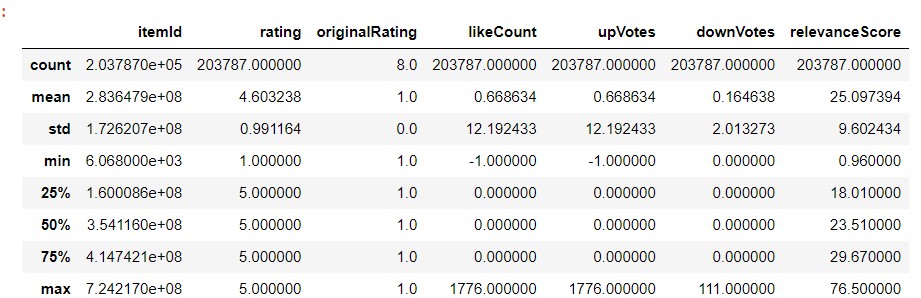
Dựa trên mô tả ở trên, chúng tôi nhận được số cột từ df\_items là 9 và hàng là 10942. Trong cột giá, giá cao nhất là 275000000, giá thấp nhất là 1000 và giá trung bình là 3020219.

Trong cột xếp hạng trung bình, xếp hạng trung bình là 4,17, xếp hạng cao nhất là 5 và xếp hạng thấp nhất là 1. Cột TotalReviews có xếp hạng trung bình là 27, thấp nhất là 1 và cao nhất là 9631.

**3. df\_reviews (Dataframe reviews)**

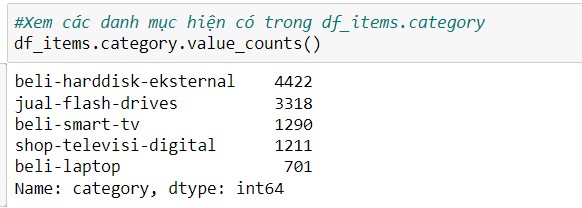
Đọc file:

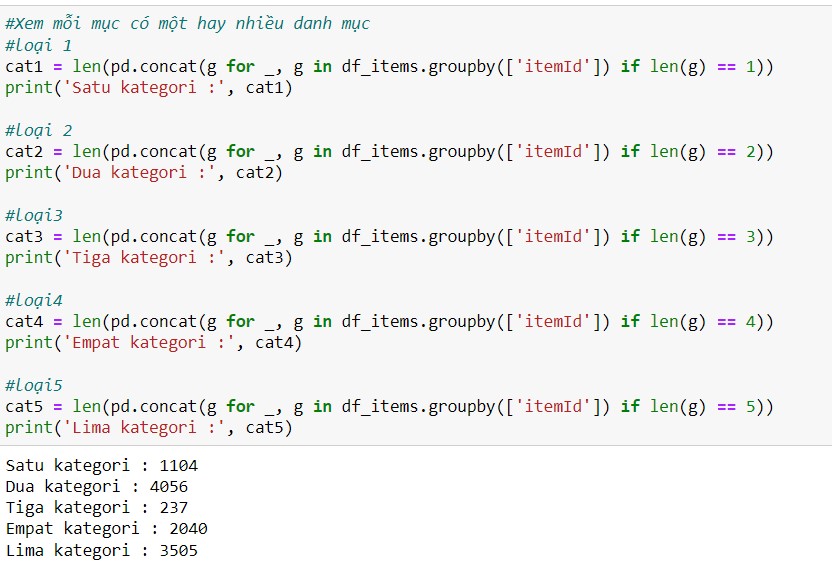




**4. Data Cleaning**

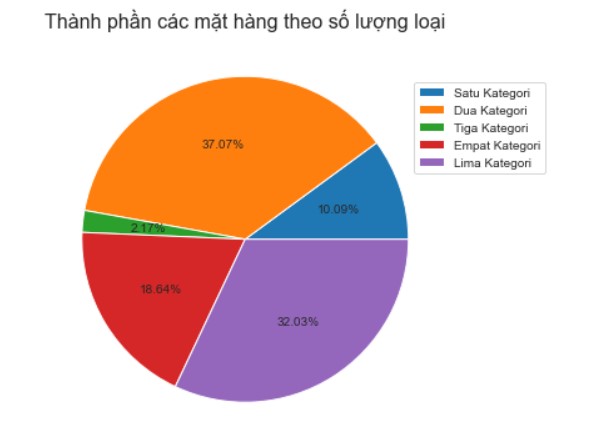
Các giai đoạn làm sạch, sửa đổi và chuẩn bị khung dữ liệu của chúng tôi khi cần thiết



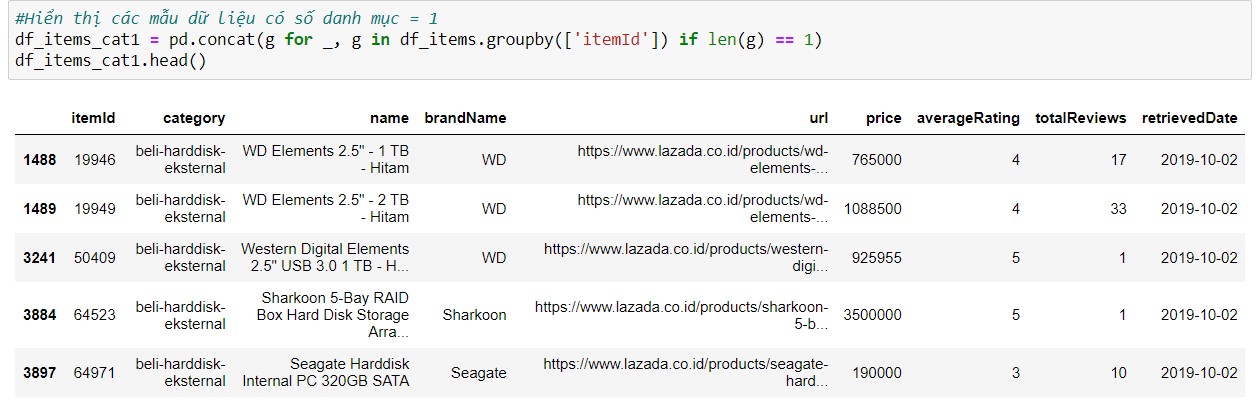


Từ quy trình trên, mỗi mục có thể có nhiều hơn một danh mục.

Vẽ biểu đồ tròn:



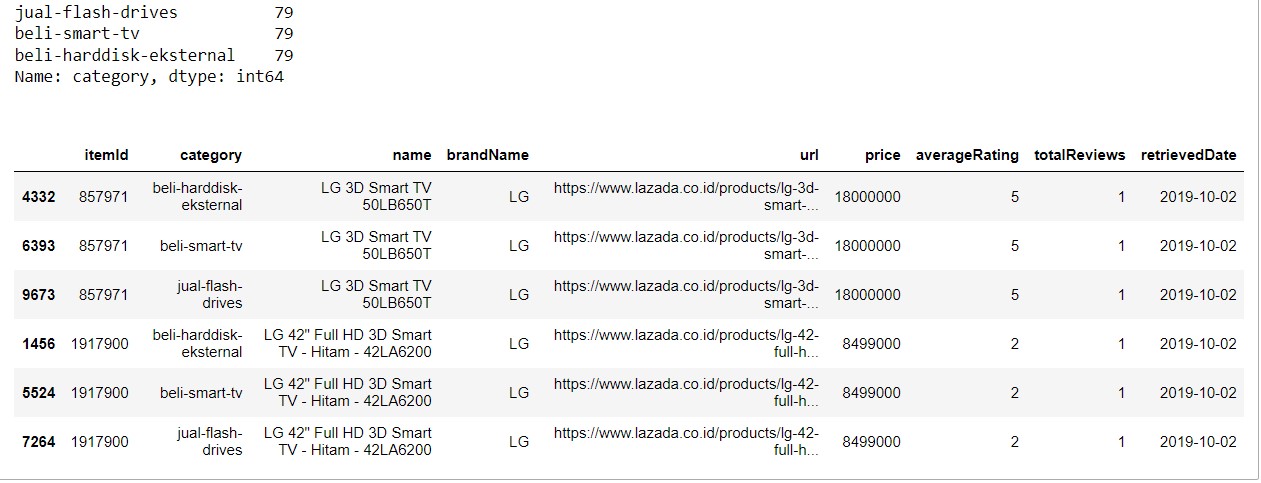
Từ hình ảnh bên trên, chúng ta có thể thấy thành phần của các mặt hàng dựa trên số lượng danh mục. Thành phần lớn nhất nằm trong hai danh mục với 37.07% trên tổng số 10942 hàng trong df\_items và thấp nhất là 2.17% trong ba danh mục. Chúng tôi cần khám phá thêm để tìm ra cách sao chép các mục trong tập dữ liệu df\_items.

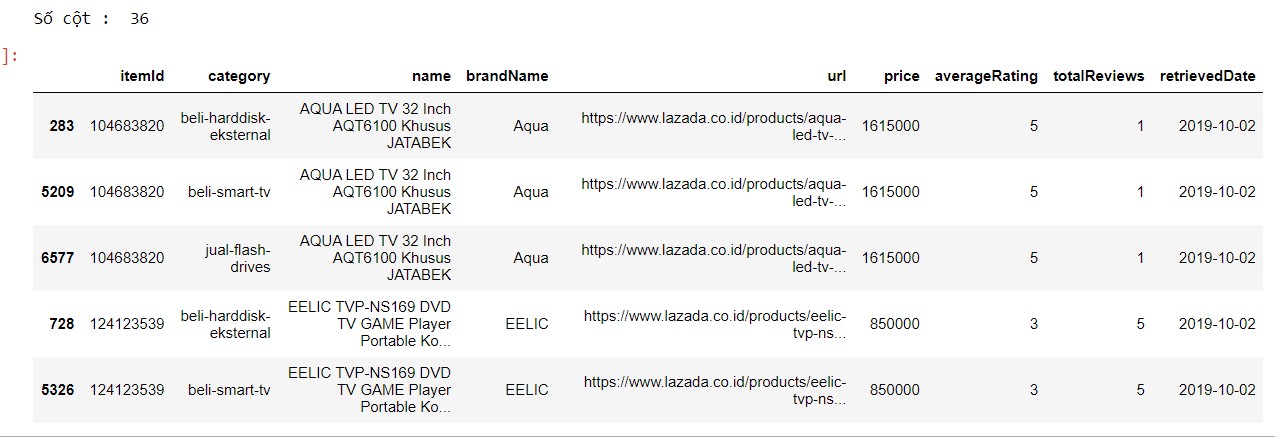




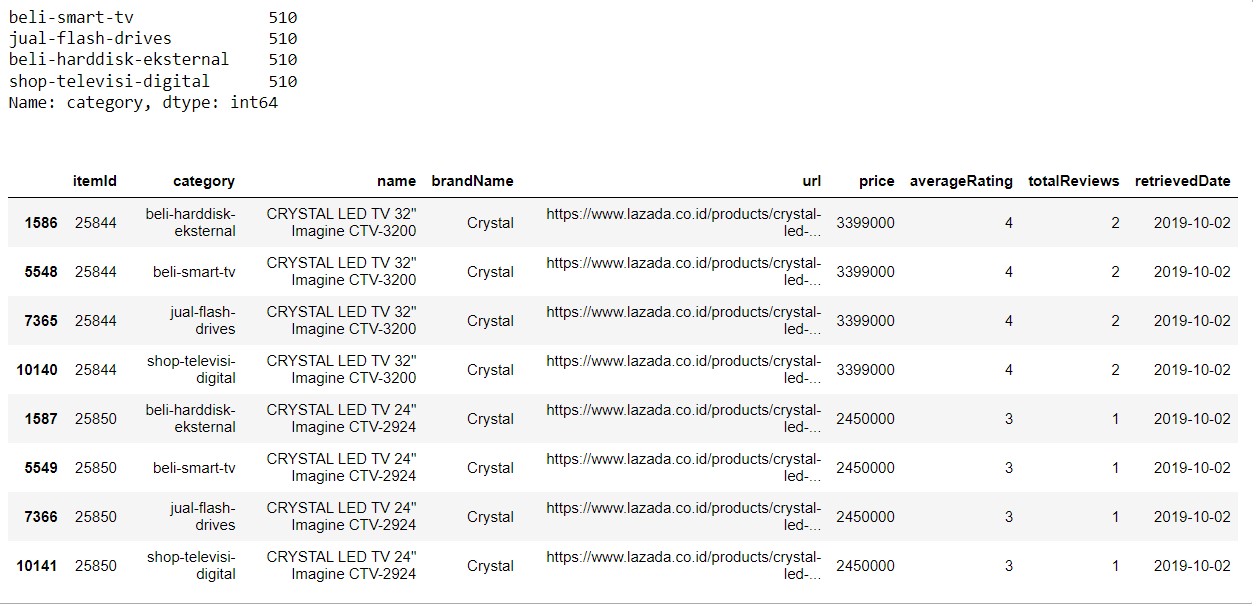
Dựa trên một mẫu mục có số danh mục = 2, chúng ta có thể thấy rằng tên mục phù hợp với danh mục ổ đĩa flash hơn là ổ đĩa cứng.

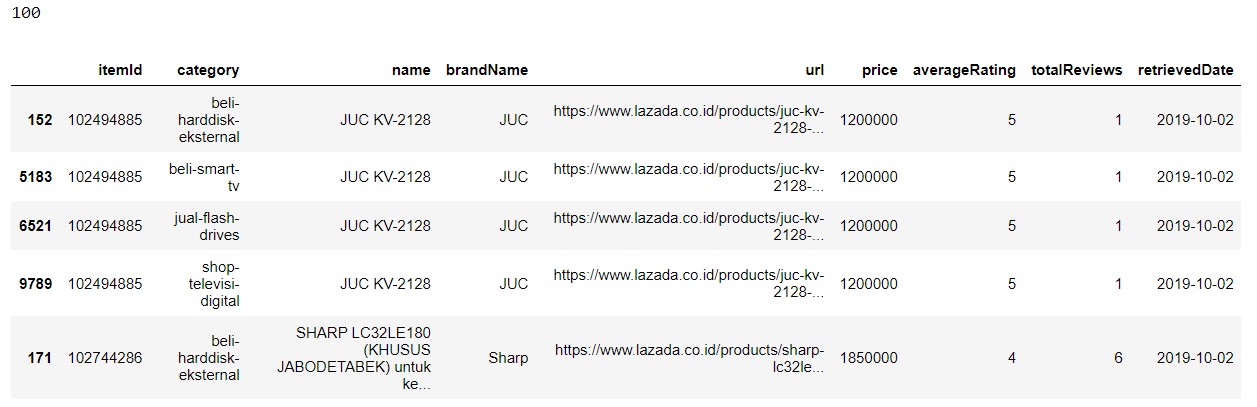
Hiển thị các mẫu dữ liệu có số danh mục = 3



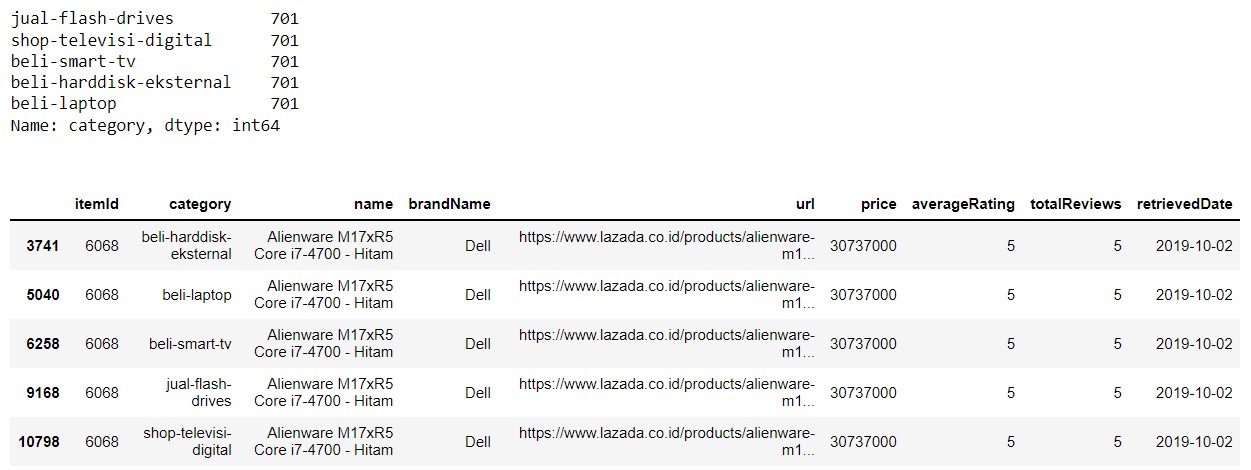


Trong ví dụ trên, chúng ta có thể thấy rằng dựa trên tên mặt hàng, danh mục phù hợp là mua-smart-TV với từ khóa 'thông minh'. Tuy nhiên, có một số mục nằm ngoài các từ khóa này. Những mục này vẫn được đưa vào danh mục mua-smart-tv vì chúng vẫn liên quan đến danh mục này.



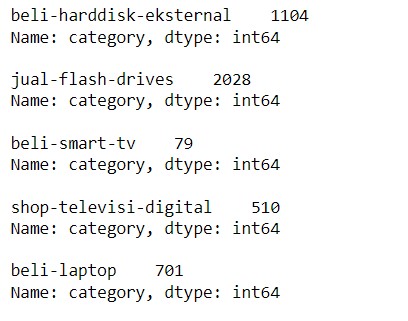


Từ mẫu và kết quả thăm dò của các hàng có số danh mục = 4, chúng tôi nhận thấy rằng hầu hết các tên mục có từ khóa tv và dẫn đầu, và chỉ 100 hàng không có từ khóa. Mặc dù vậy, tên của các sản phẩm này vẫn có sự gắn bó với danh mục cửa hàng-truyền hình-kỹ thuật số. Do đó, chúng tôi giả định nó là một danh mục cửa hàng-truyền hình-kỹ thuật số.



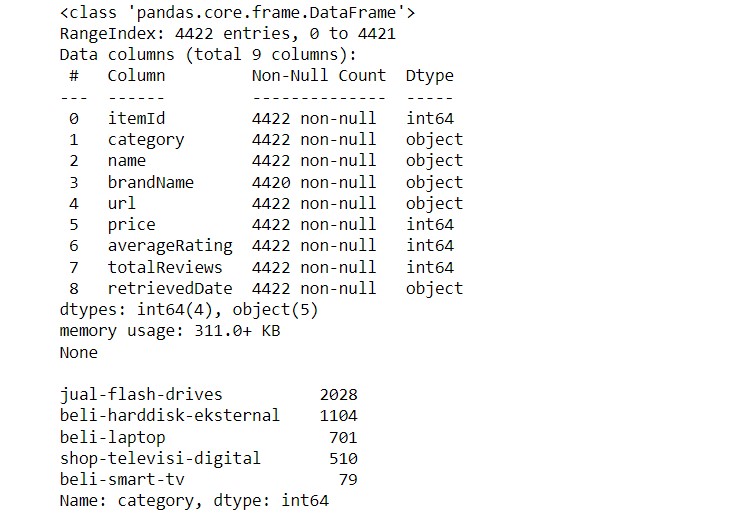
Dựa trên mẫu và phân phối tên thương hiệu từ df\_items với số danh mục = 5 ở trên, chúng tôi có ý kiến ​​rằng tên và thương hiệu liên quan đến danh mục mua máy tính xách tay nhiều hơn so với các danh mục khác.

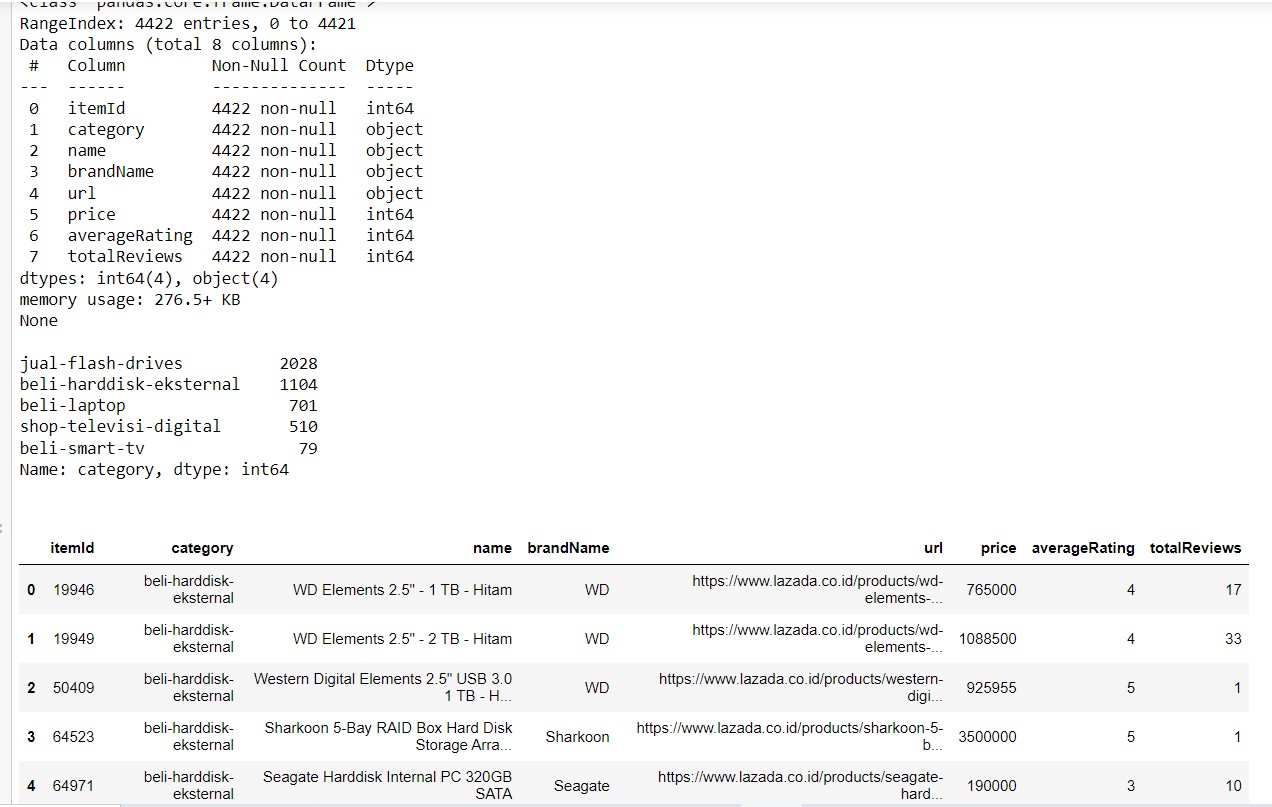
Tiếp theo, chúng tôi sẽ loại bỏ các hàng có giá trị trùng lặp được lọc theo danh mục, với kết quả mong muốn của mỗi mục chỉ có một danh mục cụ thể dựa trên khung dữ liệu đã được chia cho số danh mục.



Mỗi khung dữ liệu trước đây bao gồm nhiều hơn một danh mục hiện là duy nhất, có nghĩa là mỗi mục chỉ có một danh mục. Quá trình này được thực hiện để đơn giản hóa tập dữ liệu vì sự trùng lặp ảnh hưởng đến số hàng của tập dữ liệu và kết quả sẽ được tìm kiếm sau này. Sự trùng lặp này sẽ được giải thích khi chúng ta đến với quá trình làm sạch dữ liệu trong phần đánh giá khung dữ liệu.

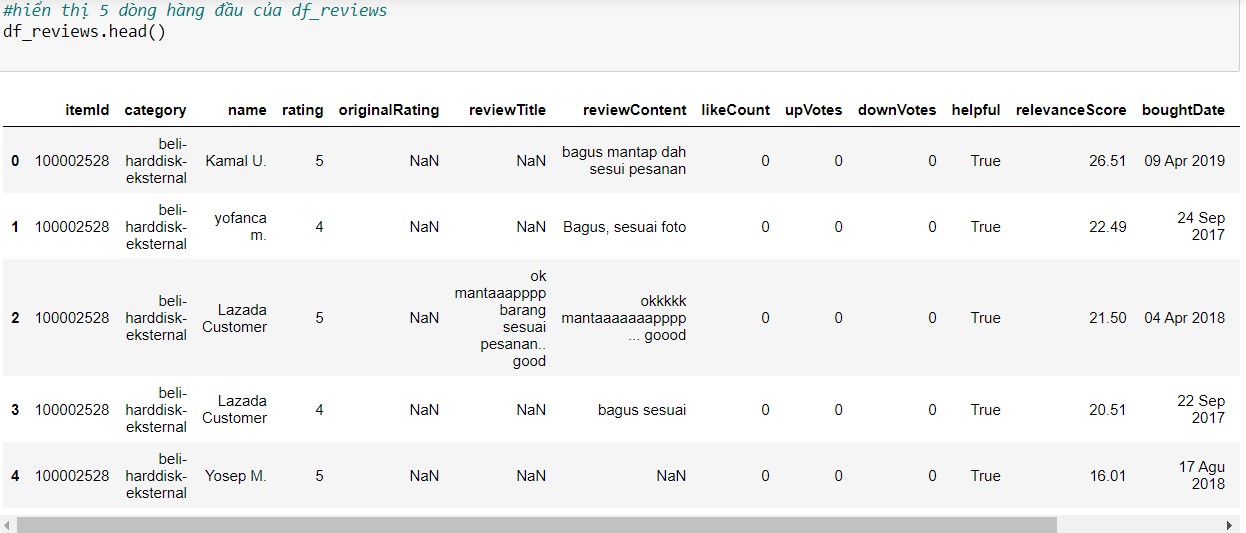
Bước tiếp theo là kết hợp năm khung dữ liệu duy nhất thành một khung dữ liệu mới và loại bỏ cột Ngày truy xuất.

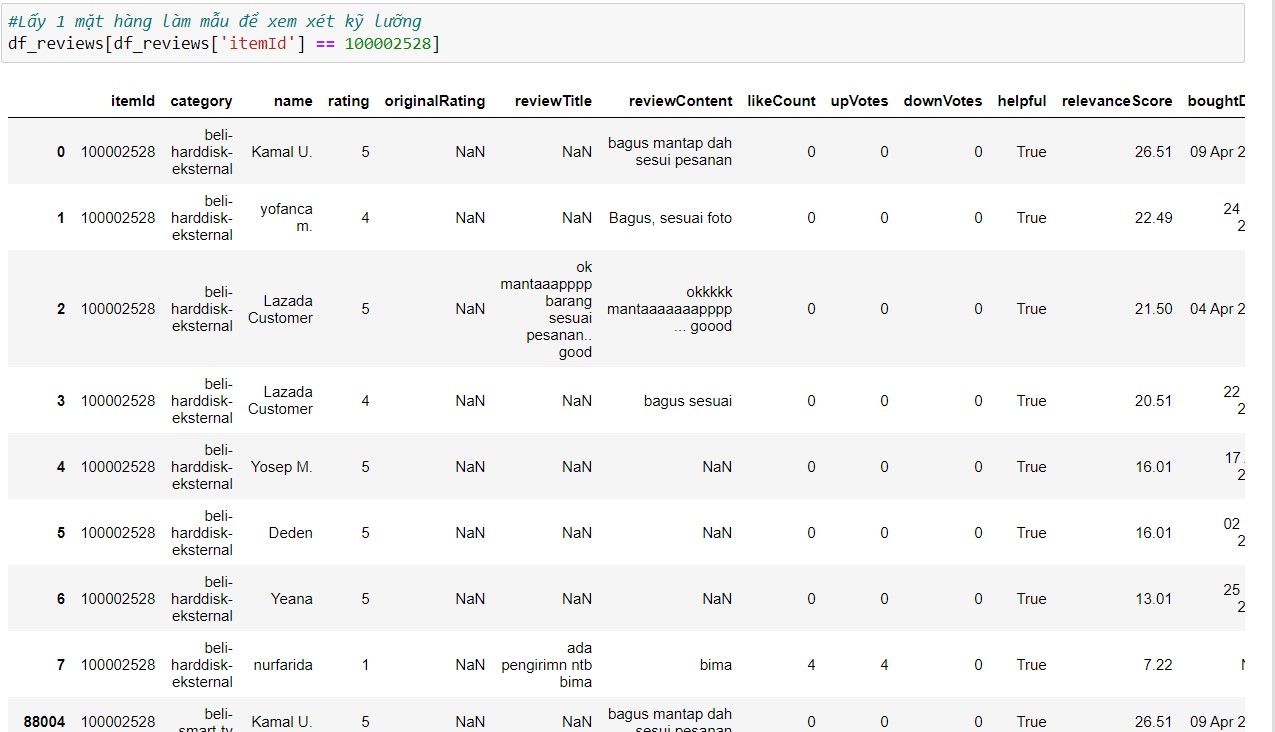




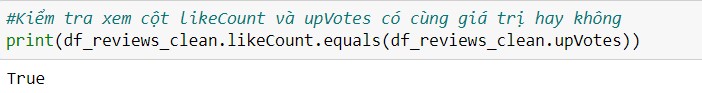
Từ quá trình trên, chúng ta đã làm sạch dữ liệu trên khung dữ liệu df\_items và lưu nó vào đối tượng df\_items\_clean. Tiếp theo, hãy chuyển sang phần đánh giá khung dữ liệu. Quá trình này ít nhiều giống với df\_items.

**df\_reviews (Dataframe reviews)**



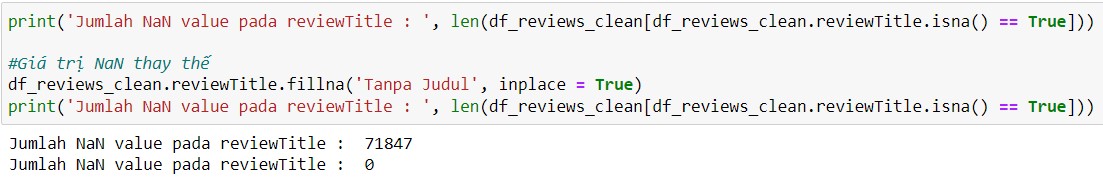


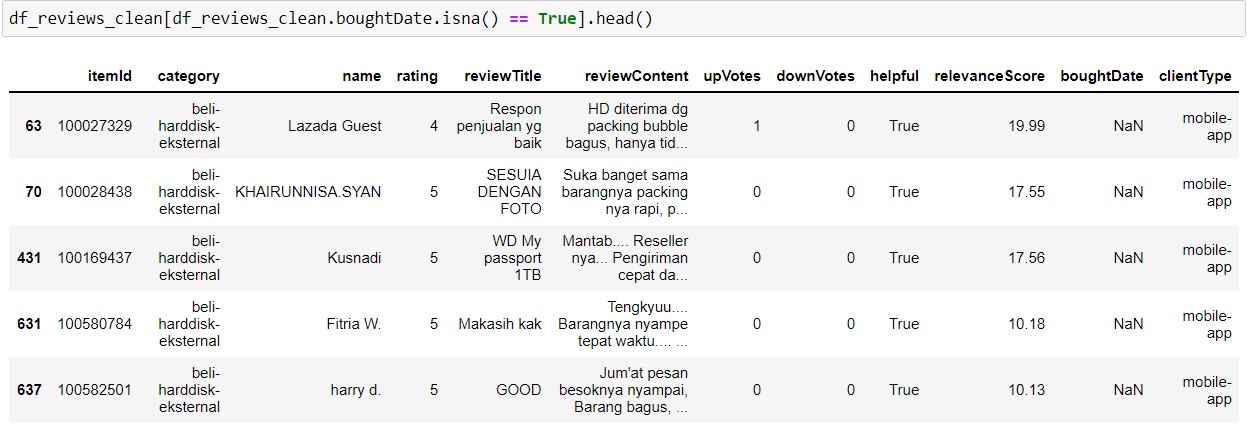
Từ một trong các mục được lấy mẫu ở trên, chúng ta có thể thấy rằng đặc điểm của df\_reviews cũng giống như df\_items, đó là trùng lặp theo danh mục, sao cho mỗi hàng đều giống nhau, chỉ phân biệt theo danh mục nên cần loại bỏ sự trùng lặp.





Tiếp theo, thay thế giá trị NaN trên reviewTitle bằng chuỗi 'Untitled' và kiểm tra giá trị NaN vào ngày mua





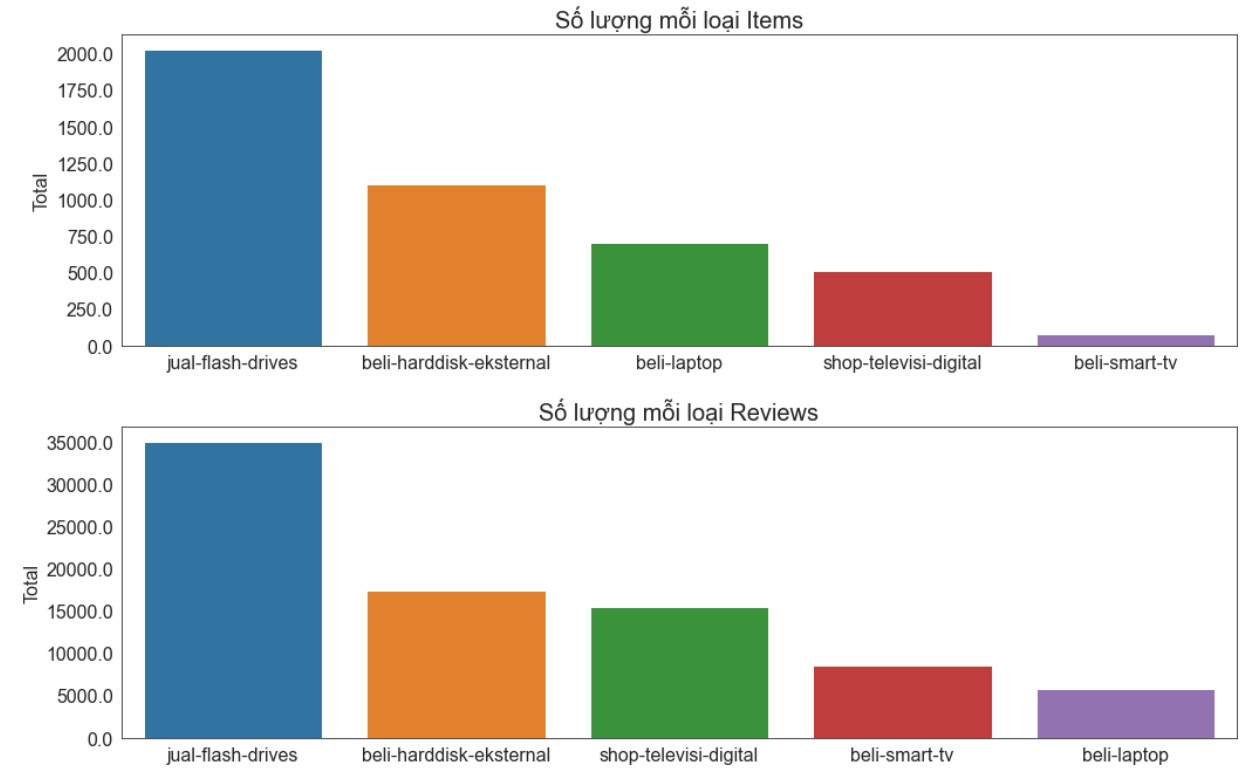
Ngoài việc chọn danh mục df\_items và df\_reviews theo cách thủ công, còn là giá trị NaN trong cột Ngày mua vì nó sẽ ảnh hưởng đến phân tích Chuỗi thời gian nếu phân tích được thực hiện. Tuy nhiên, những hạn chế này có thể được sửa chữa bằng cách sử dụng phân tích Nội dung đánh giá để cuối cùng xác định xem ai đó có mua nó hay không ngay cả khi không có ngày mua được ghi lại. Bước tiếp theo là kết hợp hai khung dữ liệu thành một để phân tích thêm.

**df\_merged (Merged Dataframe)**

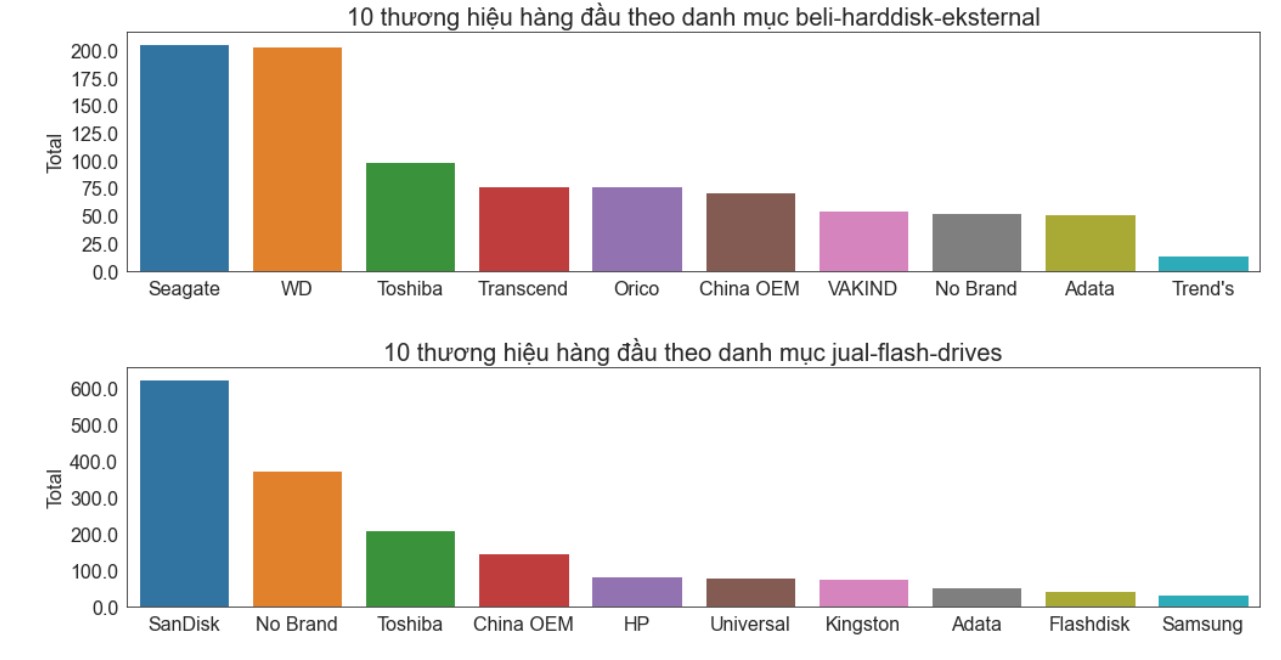
****

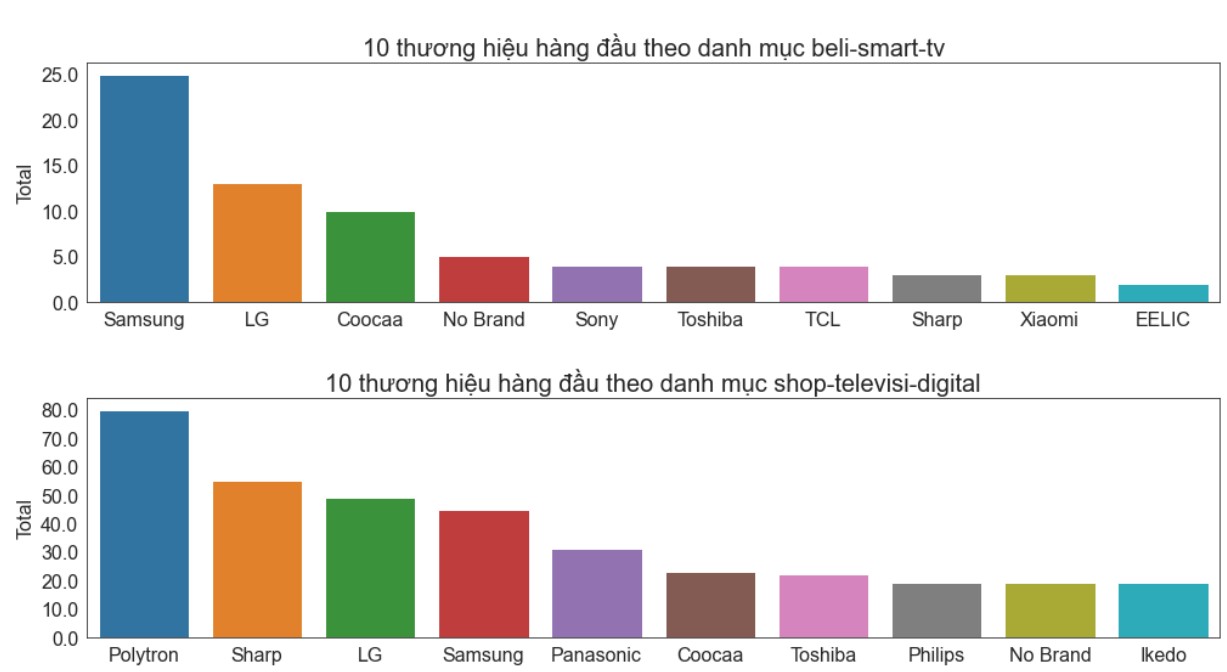
# IV. Visualisasi Data(Trực quan hóa dữ liệu)

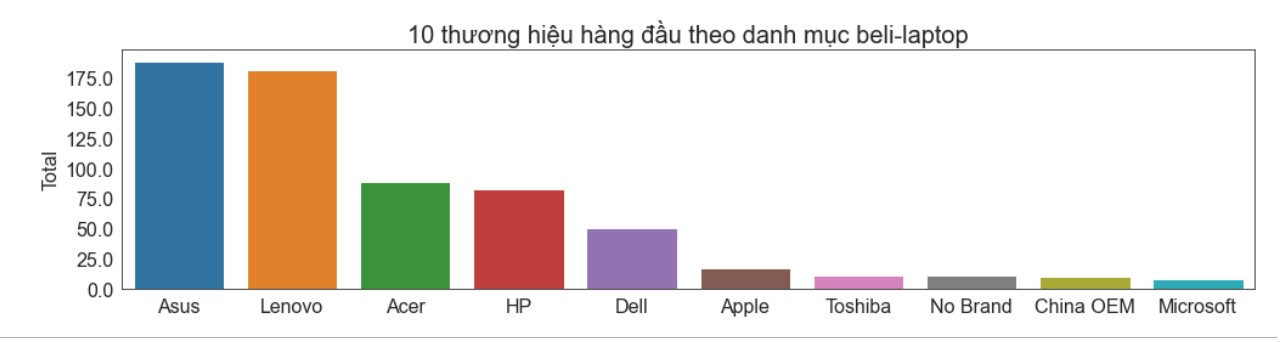
Số lượng của từng danh mục trong df\_items\_clean và df\_reviews\_clean



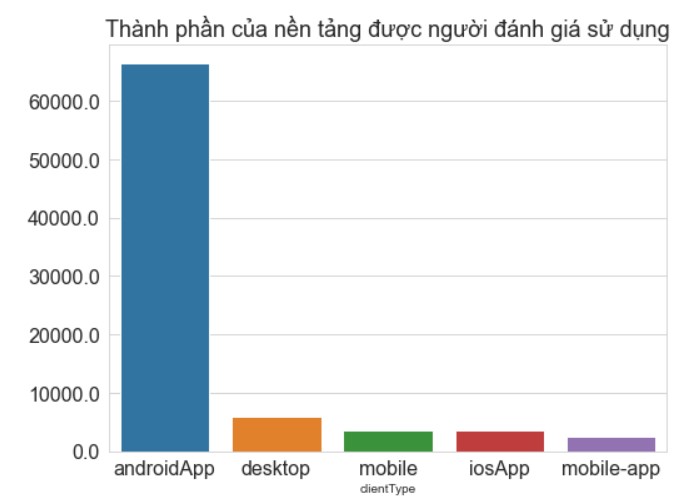
10 thương hiệu có số lượng mặt hàng cao nhất từ ​​mỗi danh mục trên df\_items\_clean



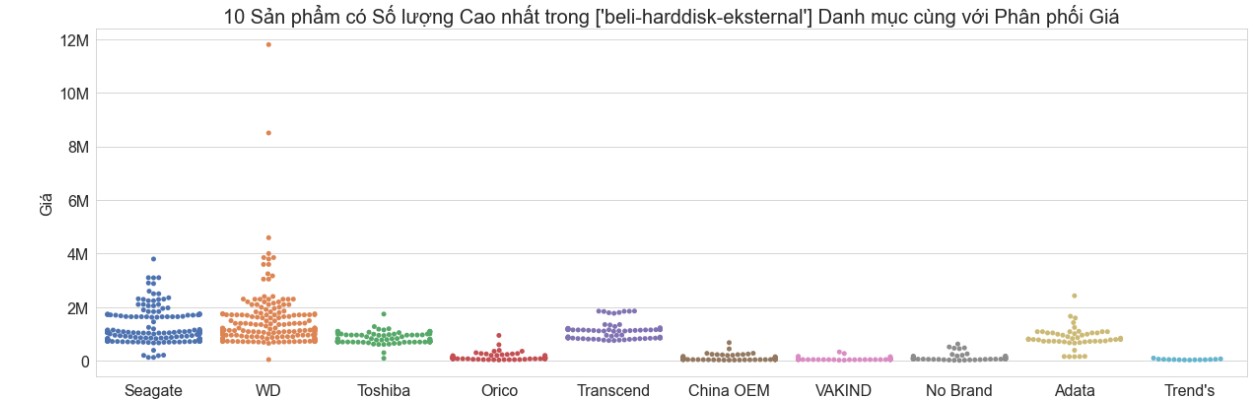


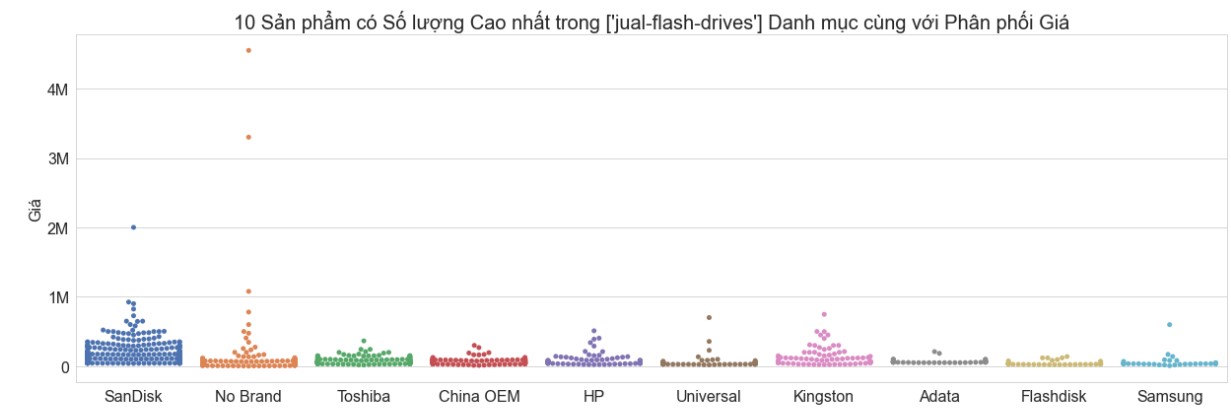


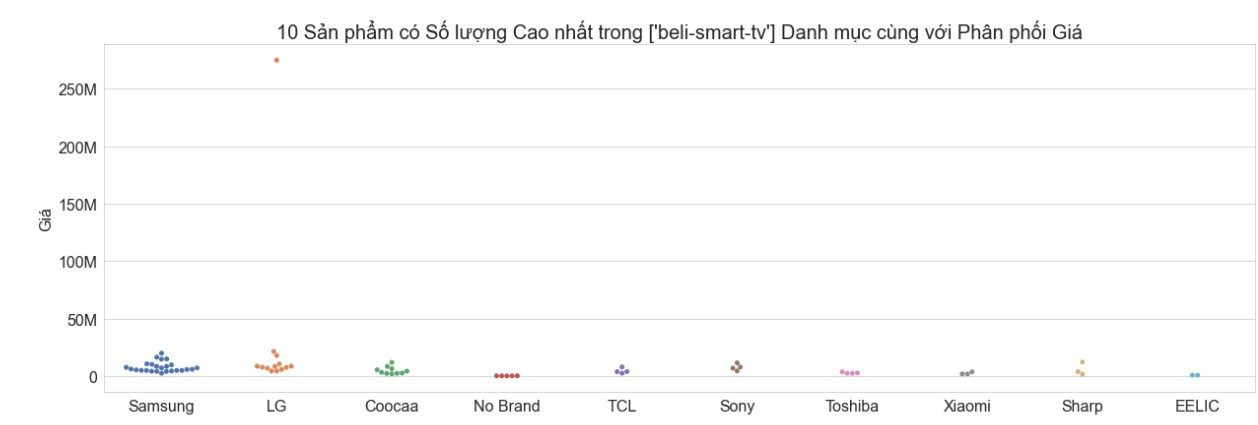
Trực quan hóa ClientType

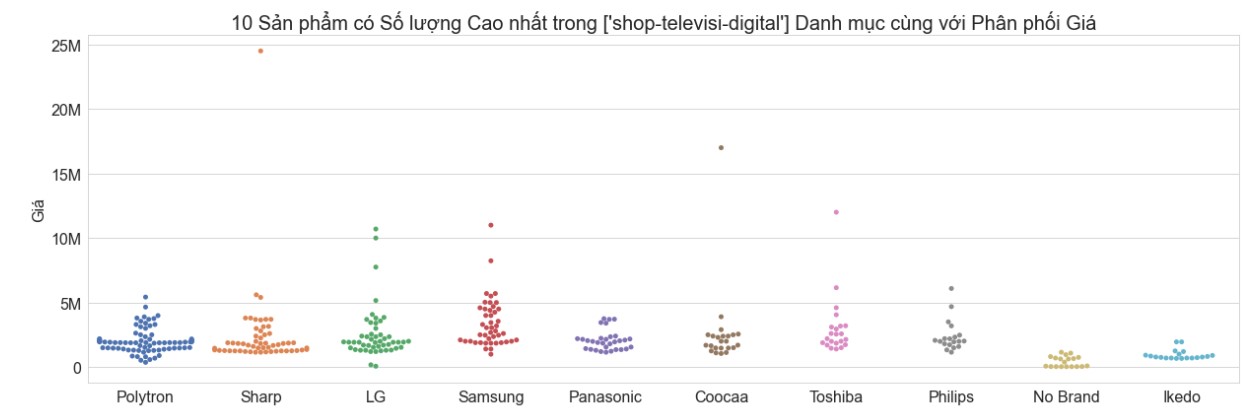


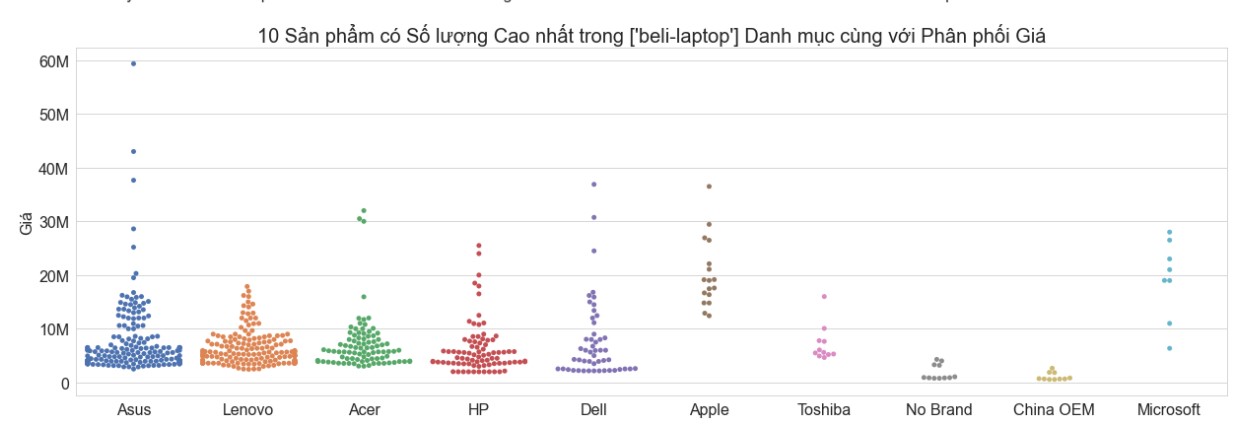
Phân bổ giá của 10 thương hiệu có số lượng cao nhất dựa trên từng danh mục





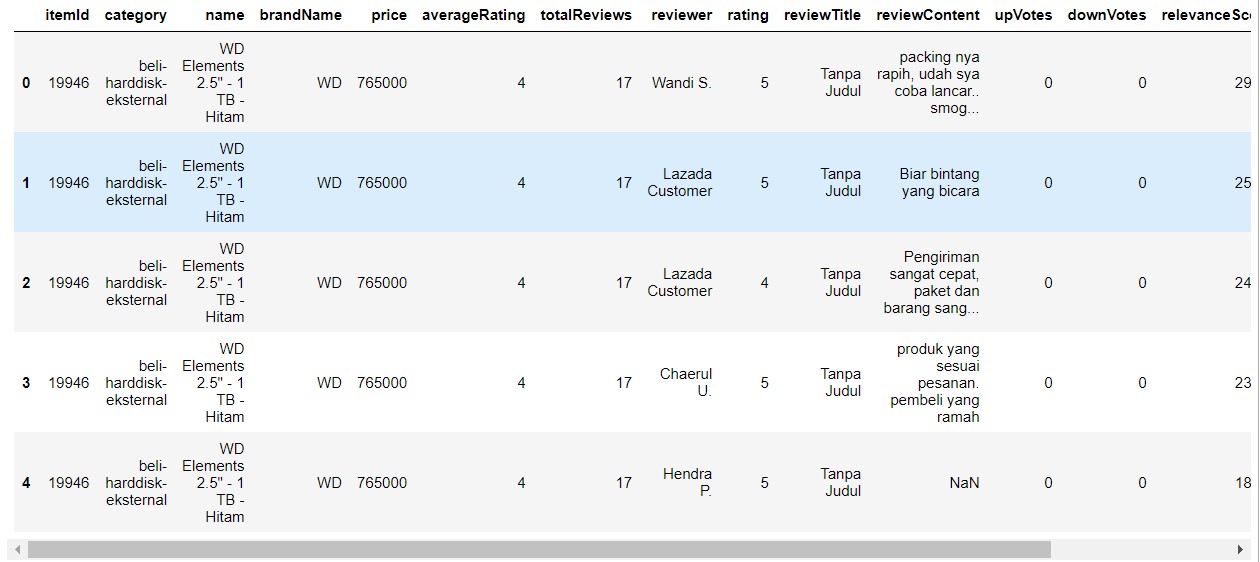




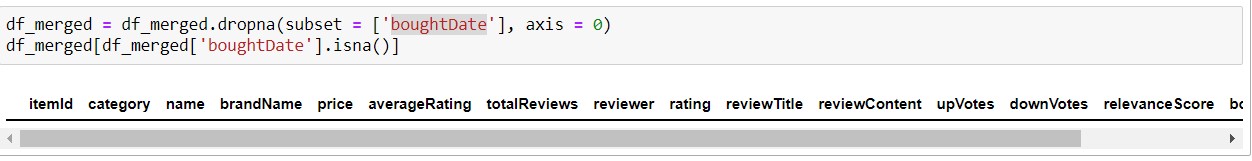


**Phân tích chuỗi thời gian**

Phân tích chuỗi thời gian này được thực hiện để có được cái nhìn sâu về xu hướng mua hàng trong một khoảng thời gian nhất định, dựa trên số lượng giao dịch cũng như tổng giá trị của các giao dịch.



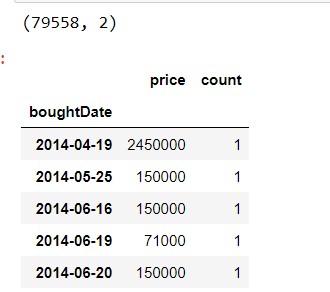
Tiếp theo dọn dẹp boughtDate là NaN



Bước tiếp theo là lấy các cột cần được phân tích trong chuỗi thời gian, đó là ngày mua và giá.



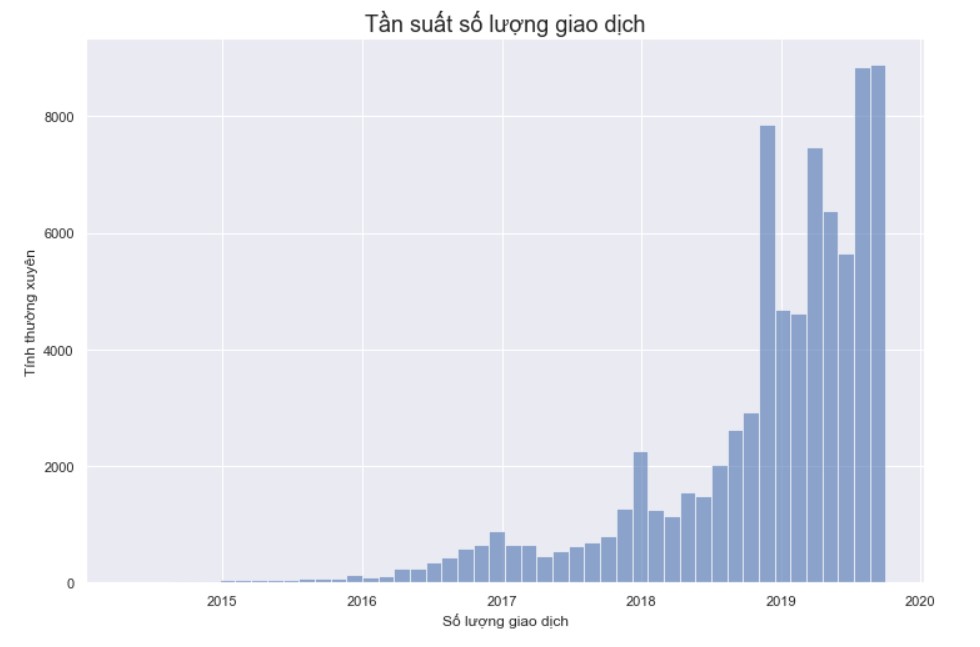
Tạo một cột mới = count



Chi tiết kho dữ liệu

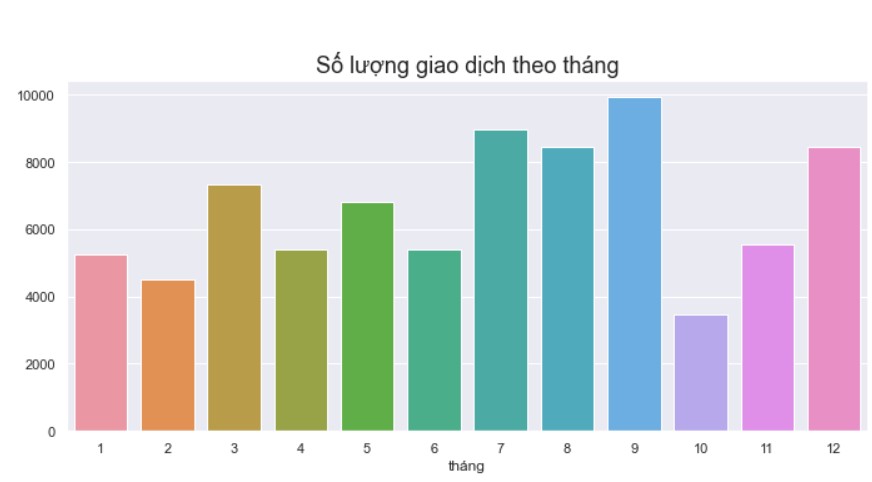
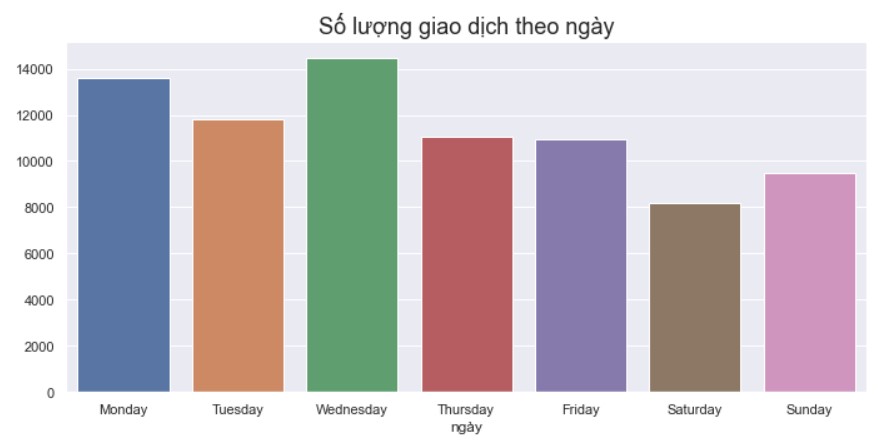


Biểu đồ số giao dịch

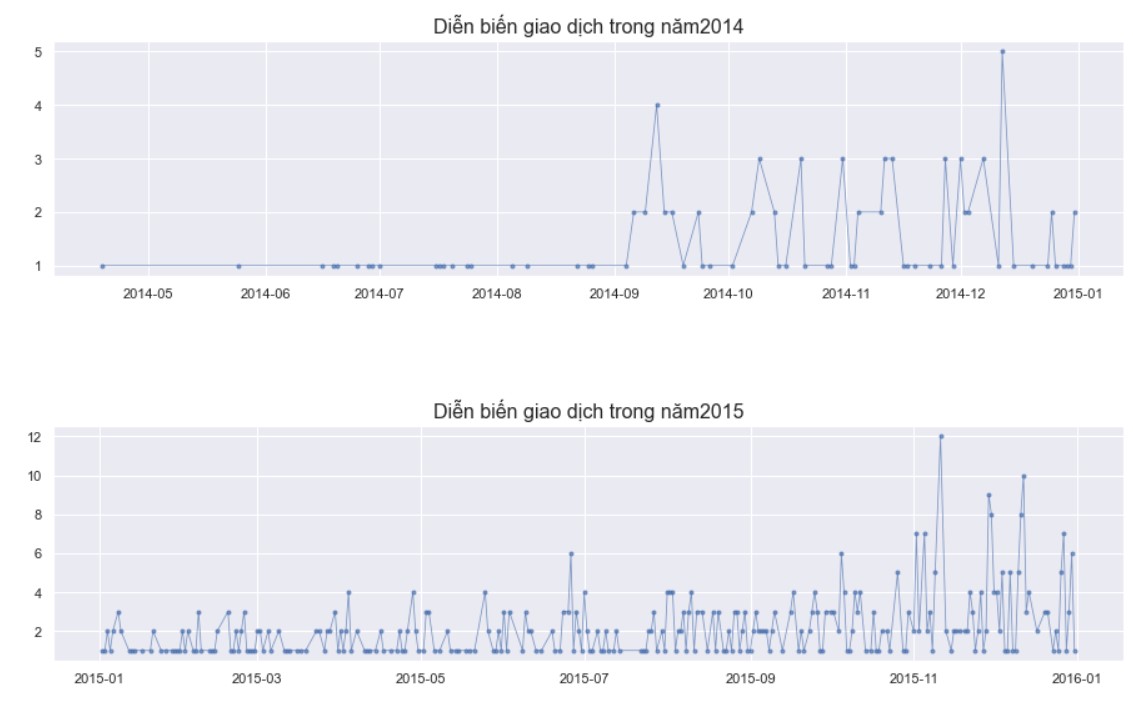


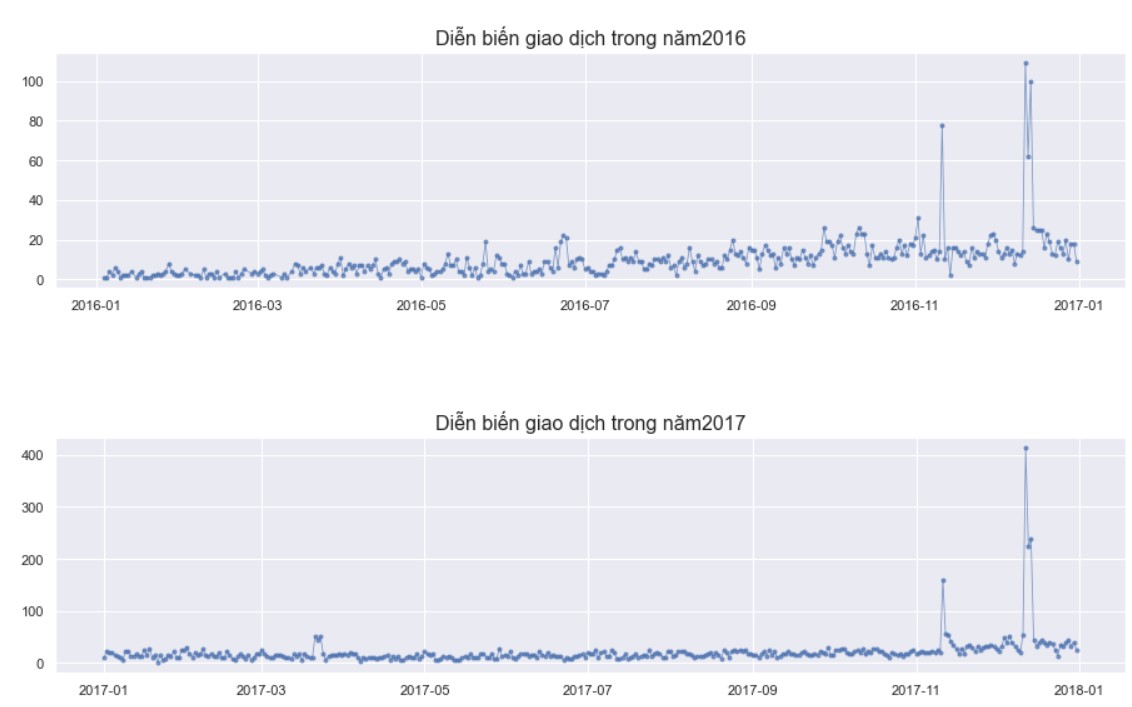
Dựa trên biểu đồ trên, có sự gia tăng cao về số lượng giao dịch vào thời điểm gần cuối năm, kể từ năm 2017. Tiếp theo, chúng ta sẽ hình dung chi tiết hơn về khung dữ liệu.

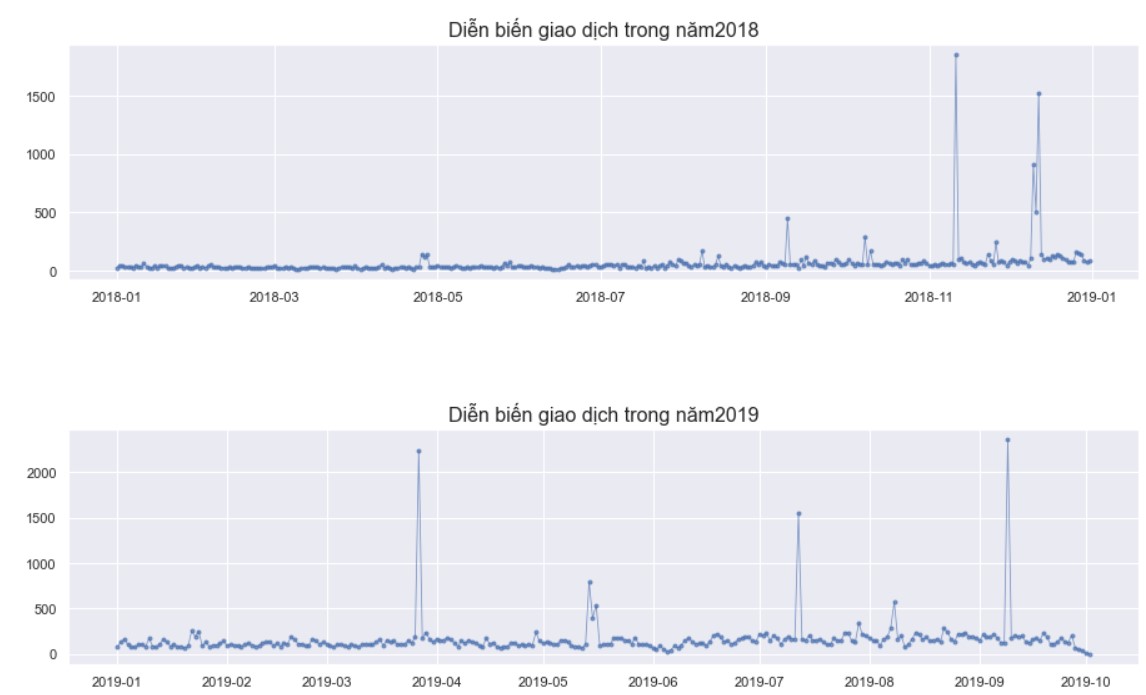




Chúng tôi sẽ lập sơ đồ chi tiết hơn, tức là so sánh các lô hàng năm





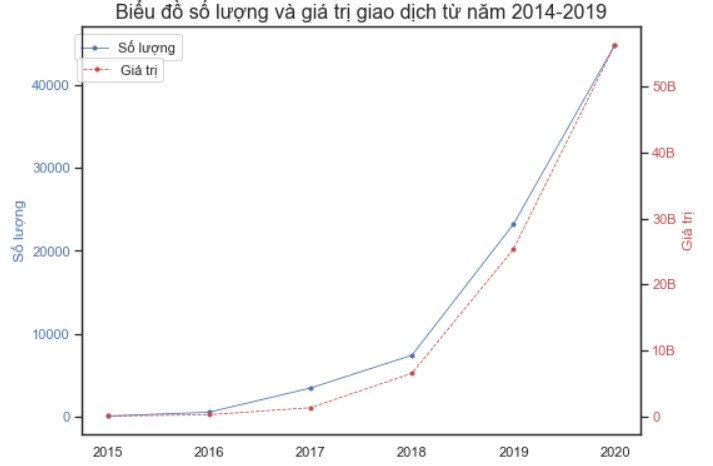


**Resampling(Lấy mẫu lại)**

****

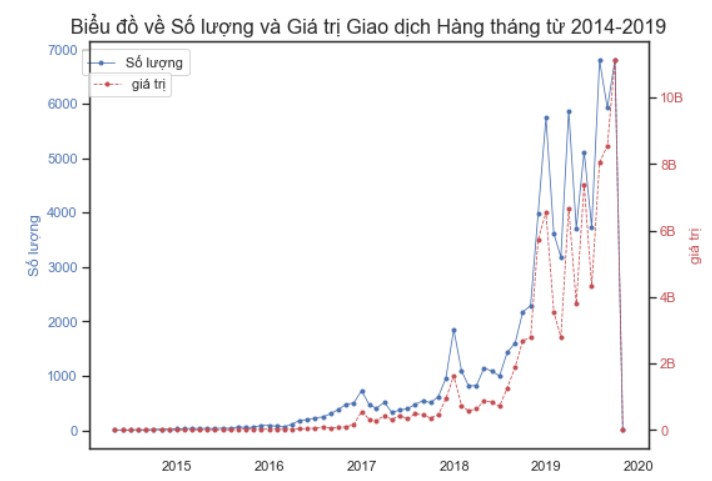
**Dựa trên giá trị giao dịch và số tiền giao dịch**

Chúng tôi sẽ trực quan hóa giá trị giao dịch dựa trên chuỗi thời gian để có thể tìm ra biến động của giá trị giao dịch trong một thời gian nhất định

****

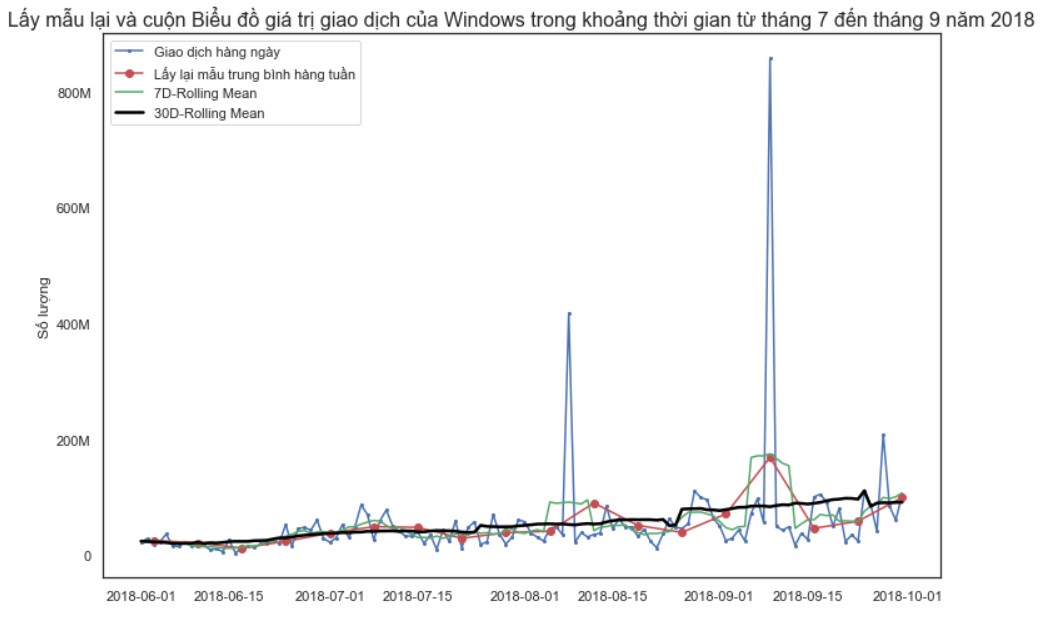
Lấy mẫu 3 tháng: 2018-06 đến 2018-09

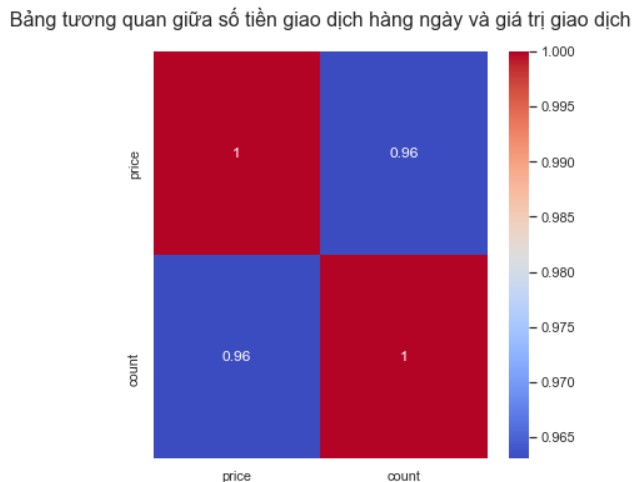
Lấy mẫu lại và luân phiên hàng tuần có nghĩa là 7, 30 ngày: số lượng giao dịch





Lấy mẫu lại và luân phiên hàng tuần có nghĩa là 7, 30 ngày: giá trị giao dịch





Nhận xét: Dựa trên hình ảnh hóa lấy mẫu lại, sự di chuyển của dữ liệu giữa số lượng giao dịch và giá trị giao dịch gần như giống nhau. Cả hai cả hai đều tăng rất đáng kể vào năm 2019. Khi chúng tôi hình dung chi tiết về mẫu cho khoảng thời gian từ tháng 7 đến tháng 9 năm 2018, cả hai đều có xu hướng tương tự. Tiếp theo, chúng tôi cố gắng tương quan các cột số lượng và giá cả và tạo ra mức tương quan là 0,96, có nghĩa là nó có mối tương quan thuận rất mạnh.

Nguồn Kaggle:

https://www.kaggle.com/hanafication/lazada-data-visualization