KHOA HỌC WEB

PROJECT 2 - MỐI QUAN HỆ CỦA DỮ LIỆU

Giảng Viên: **thầy Lê Ngọc Thành**

Thành viên nhóm:

20424027 – Phạm Thi Minh Hậu

20424056 – Nguyễn Thế Ngọc

20424058 – Nguyễn Văn Nhật

Mục lục

1. Môi trường, thư viện

- Link drive 3 đồ án của nhóm: https://drive.google.com/drive/folders/1HXb2XEt5mfVIRtzOZarSnEXAKTaTR6_y?usp=sharing
- Ngôn ngữ: Python 3.8.5
- Định dạng file: .py và . ipynb
- Phần mềm sử dụng: JupyterLab 2.2.6 trong Anaconda
- Hướng dẫn chạy: cài đầy đủ các thư viện bên dưới (bản mới nhất càng tốt), sau đó mở notebook, chọn Reset kernal and run all cells
- Định dạng Dữ liệu nhóm tổ chức lưu: JSON
- Các thư viện nhóm sử dụng:
  - json
  - os
  - time
  - statistics
  - numpy
  - pandas
  - matplotlib (matplotlib.pyplot) (yêu cầu phiên bản 3.4.2 trở lên)
  - scipy (scipy.stats)
  - collections
  - tabulate để in ra cho đẹp khi thống kê như bên dưới (**KHÔNG PHẢI PHẦN MỀM TABLEAU**)

Có phải mall	Số sản phẩm	chiếm %
1	927	56.5589
0	711	43.3801

2. Công việc của mỗi thành viên

MSSV – Họ Tên	Công việc
20424027 – Phạm Thị Minh Hậu	Tiền xử lý dữ liệu Liệt kê một vài quan hệ đơn Trực quan dữ liệu nhóm 1
20424056 – Nguyễn Thế Ngọc	Bổ sung dữ liệu và thêm trường dữ liệu mới Thống kê dữ liệu, tổ chức dữ liệu trước khi trực quan Trực quan dữ liệu nhóm 2, 3 và 4
20424058 – Nguyễn Văn Nhật	Tiền xử lý dữ liệu Liệt kê một vài quan hệ giữa các trường Trực quan vài dữ liệu nhóm 2

3. Mức độ hoàn thành tổng thể và đánh giá của nhóm

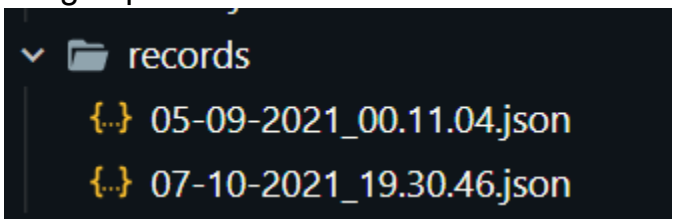
Các tiêu chí đánh giá	Điểm tối đa (%)	Nhóm đánh giá (%)
Tiền xử lý dữ liệu	5%	5%
Thống kê dữ liệu	10%	10%
Chọn, giải thích, trực quan các trường và các mối quan hệ giữa chúng	35%	35%
Rút ra ý nghĩa hợp lý sau mỗi dự liệu được trực quan	20%	20%
Xem xét trên nhiều quan hệ/góc nhìn khác nhau	10%	10%
Báo cáo trình bày có bố cục, định dạng hợp lý, rõ ràng	20%	20%
Tổng	100%	100%

4. Thu thập dữ liệu

- Nhóm thu thập dữ liệu trên trang tiki và shopee, tại các sản phẩm sale thông qua project 1, kỹ thuật không thay đổi, tại shopee dùng selenium đi dần, tại tiki sẽ dùng request giả trình duyệt với cookie giả để yêu cầu API.
- Dữ liệu nhóm thu thạp được sẽ tổ chức ở dạng JSON, các dữ liệu này chỉ được tiền xử lý cơ bản, như loại bỏ các chữ cái “đ”, “VND” trong giá tiền….
- Ở project 2 này, nhóm đã dùng lại project 1 crawl thêm khoảng 1000 data và bổ sung thêm 2 trường dữ liệu là ship và mall. Trường ship thể hiện sản phẩm này sẽ được gửi từ đâu (trong nước hoặc quốc tế), trường mall để biết sản phẩm có phải được bán từ 1 cửa hàng mall hay không, vì người dùng thường ưu tiên mua tại các cửa hàng mall.
- Nhóm có tổng cộng khoảng 4000 dữ liệu.
- Mỗi dữ liệu có các thuộc tính **chính** như: tên sản phẩm, giá gốc, % giảm giá, số đã bán được, danh mục (mảng và có phân cấp), điểm đánh giá, ship từ đâu, có phải bán từ cửa hàng mall không.

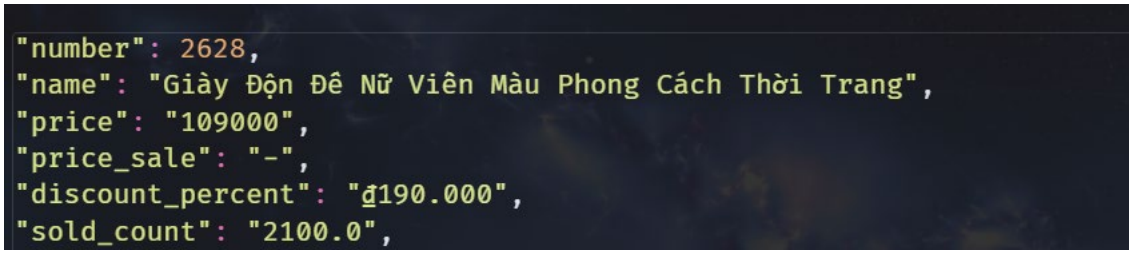
5. Tiền xử lý dữ liệu

- Nhóm thu thập dữ liệu 2 lần nên khi tiền xử lý, nhóm cũng sẽ gọp 2 file từ project 1 cung cấp.

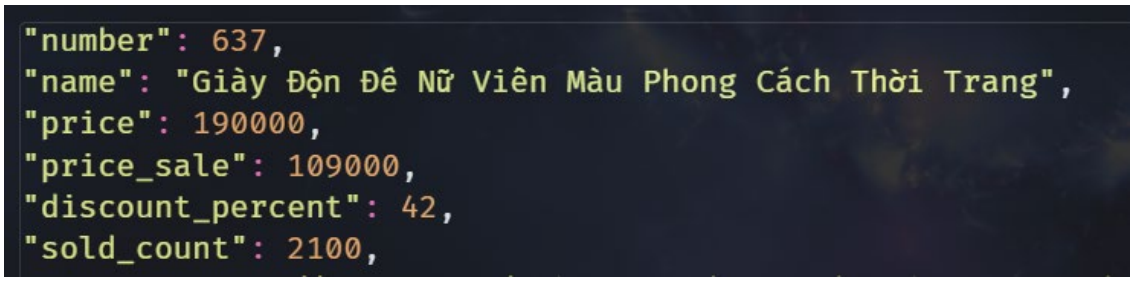


- Nhóm tiền xử lý dữ liệu theo 2 bước, tại bước 2 sẽ chia làm 3 hướng đi:
  - Bước 1: Sau khi gọp dữ liệu từ các file JSON, nhóm thực hiện:
    - Chuẩn hóa các dữ liệu số cho đồng nhất với nhau
    - Loại bỏ 1 số dữ liệu không thể dùng nữa vì thiếu quá nhiều trường (không có trường danh mục...)
    - Sắp xếp tìm dữ liệu cho hợp lý với lỗi bị lộn trường

    Trước:

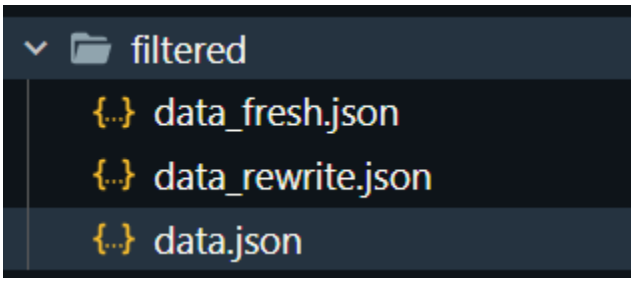


    Sau tìm trường sai và sắp lại:



  - Bước 2: Hướng số 1:
    - Nhóm sẽ bổ sung các trường có thể như điểm đánh giá, % giảm giá, số đã bán bằng con số 0.
  - Bước 2: Hướng số 2:
    - Nhóm bổ sung các trường có thể bằng giá trị trung bình của trường đó.
  - Bước 2: **Hướng số 3** (Hướng chính nhóm sử dụng, hướng 1 và 2 sẽ dùng trong phần góc nhìn khác):

- Loại bỏ hoàn toàn các dữ liệu bị lỗi các trường số.



data_fresh.sjon sẽ dành cho hướng 1

data_rewrite.json sẽ dành cho hướng 2

data.json dành cho hướng 3, và là hướng nhóm dùng chính để trực quan.

6. Một vài thống kê cơ bản

- Nhóm thống kê đếm số lượng, tìm max, min, mean và độ lệch chuẩn của các trường dữ liệu (chi tiết hơn ở file notebook).
- Ở đây nhóm sẽ liệt kê vài thống kê không quá dài:

  - **Thống Kê về số sản phẩm và nền tảng của chúng:**

```
Tên nền tảng      Số sản phẩm    chiếm %
--------------  -------------  ---------
tiki                      1056    64.4295
shopee                     582    35.5095
__________________Sản phẩm__________________
mean của data này: 819.0
độ lệch chuẩn(std): 237.0
max: 1056
min: 582
```

  - **Thống kê về số sản phẩm và hình thức ship của chúng:**

```
Cửa hàng tại      Số sản phẩm    chiếm %
--------------  -------------  ---------
oversea                    608    37.0958
vietnam                   1030    62.8432
____________________Sản phẩm____________________
mean của data này: 819.0
độ lệch chuẩn(std): 211.0
max: 1030
min: 608
```

- **Thống kê về sản phẩm và các điểm đánh giá:**

```
điểm đánh giá (0~5 điểm)      Số sản phẩm    chiếm %
-------------------------  -------------  ---------
0.0~0.9                                0   0
1.0~1.9                                5   0.305064
2.0~2.9                                5   0.305064
3.0~3.9                               64   3.90482
4.0~4.9                             1126  68.7004
5.0                                  438  26.7236
__________________Sản phẩm_____________________
mean của data này: 273.0
độ lệch chuẩn(std): 411.6414297257586
max: 1126
min: 0
__________________Điểm đánh giá_____________________
mean của data này: 4.676434676434677
độ lệch chuẩn(std): 0.4200197525083536
max: 5.0
min: 1.0
```

- **Thống kê giá gốc của danh mục cấp 1 có tên "Điện gia dụng" (201 sản phẩm)**

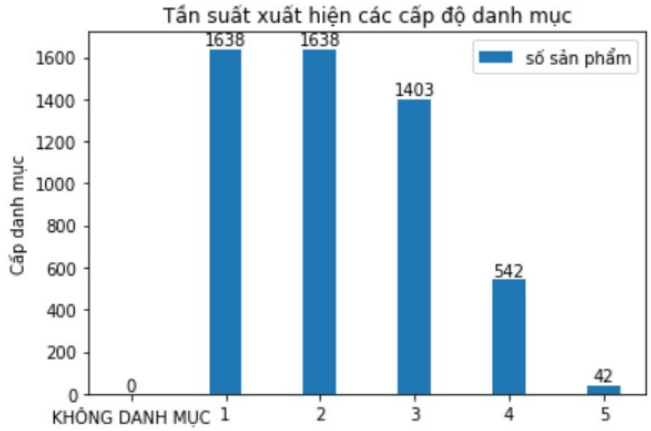
```
__________________Giá gốc 'Điện Gia Dụng'_____________________
mean của data này: 2780506.5621890547
độ lệch chuẩn(std): 3741818.7757000765
max: 33400000
min: 139000
```

7. Các kỹ thuật sử dụng

Nhóm sử dụng 4 trong 5 nhóm loại chart đã học ở lý thuyết. Mỗi loại nhóm sẽ dùng 1 hoặc nhiều hình để biểu diễn và giải thích, cũng như nêu ý nghĩa.

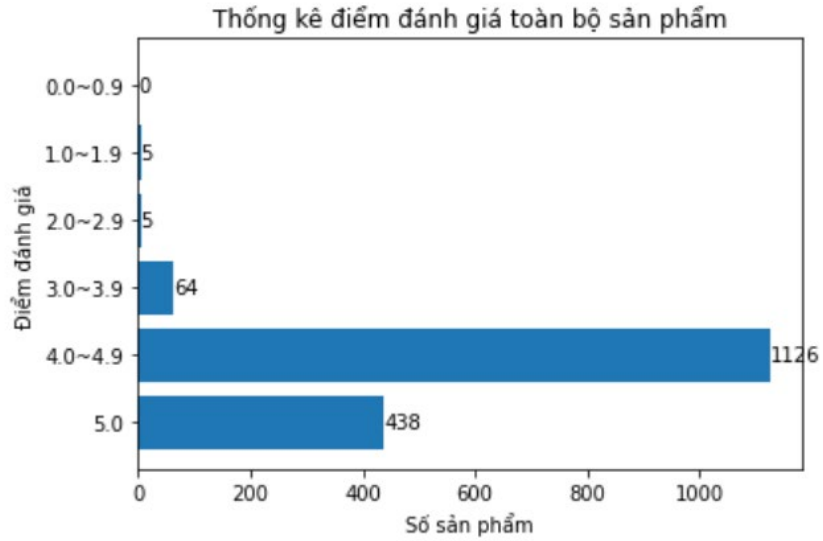
7.1. Nhóm 1: bar chart và line chart

Nhóm nhận thấy các sản phẩm đa số đều có từ 2 cho đến 5 danh mục, nếu từ danh mục ta biết danh mục nào chiếm nhiều, đánh giá cao thì có thể ta biết được nó có quan hệ với giá hoặc số bán ra, đầu tiên cần xác định có mấy cấp danh mục, mỗi cấp có bao nhiêu sản phẩm. (Cấp 1 lồng cấp 2, cấp 2 lồng cấp 3….)



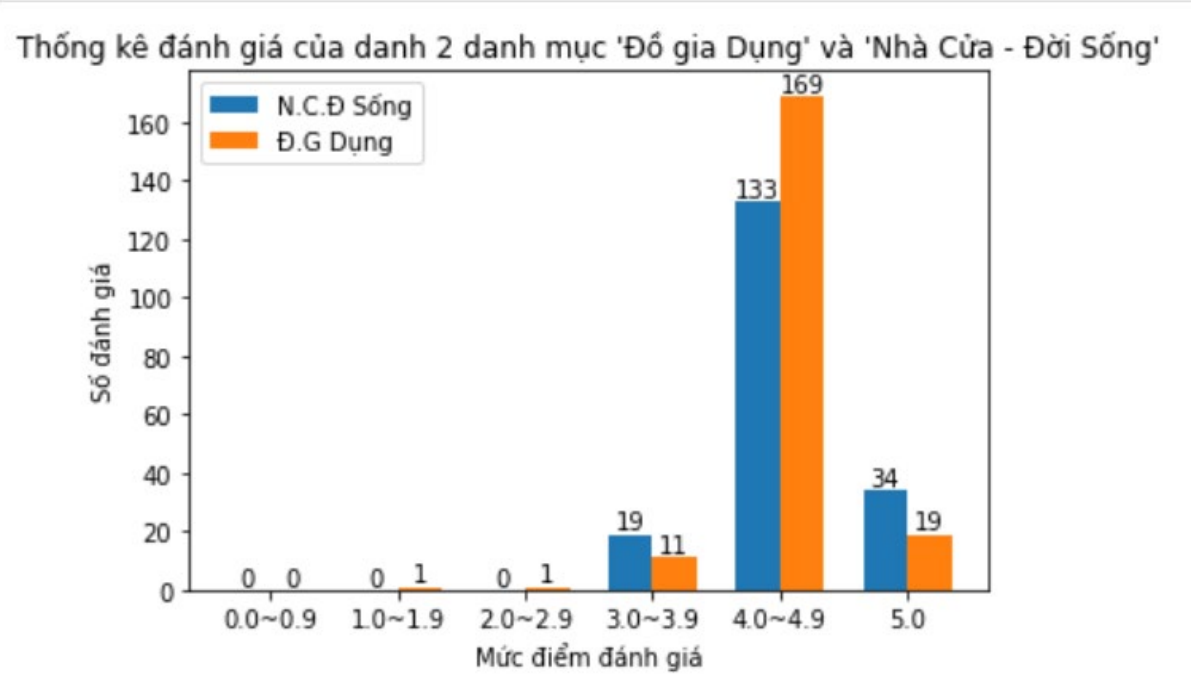
Sau khi trực quan, ta thấy rằng các sản phẩm có 5 cấp độ danh mục rất ít, ta nên để ý từ cấp độ 1 đến 3. Mặt khác nếu xét phương diện ở người dùng, khi mua sắm họ sẽ cần tìm kiếm ở danh mục lớn trước, nên ta đặt độ quan trọng danh mục tăng dần theo cấp 1 2 3.

Ta đi trực quan thử điểm đánh giá toàn bộ sản phẩm, vì khi mua hàng người dùng sẽ quan tâm đến điểm đánh giá của sản phẩm đó, nếu sản phẩm có đánh giá thấp thì khả năng cao số bán ra cũng thấp theo, và cũng có thể dự đoán đó không phải cửa hàng mall, ta đi trực quan thử:



Sau trực quan ta thấy các sản phẩm được đánh giá khá tốt, phần lớn được đánh giá 4 sao, 5 sao. chỉ có 10 sản phẩm thuộc 1~2.9 sao, ta sẽ chú ý hơn phần 4~4.9 sao.

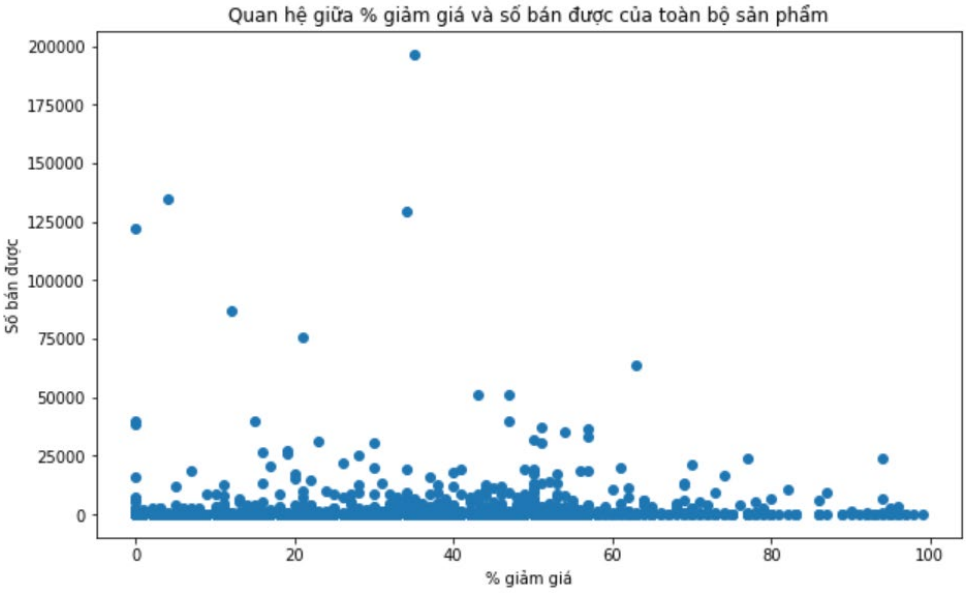
Ở bước thống kê (trong notebook), ta thấy Danh mục cấp 1 tên “Đồ gia dụng” và “Nhà cửa đời sống” có số sản phẩm khá nhiều, ta thử đi xem điểm đánh giá của 2 danh mục này. liệu các đánh giá xấu ở trên có năm ở đây không.



Ta thấy danh mục Đồ Gia dụng có đánh giá khá tốt, đánh giá mức 4 sao rất cao tuy nhiên, danh mục “Nhà cửa đời sống” hoàn hảo hơn hẳn, tuy ít đánh giá 4 và 5 sao hơn nhưng lại không có đánh giá nào dưới 3 sao, nếu là người dùng nhìn vào người ta sẽ ưu tiên an toàn ổn định hơn mà mua, từ đó có thể tạm nghĩ là % bán ra của đồ gia dụng cao hơn.

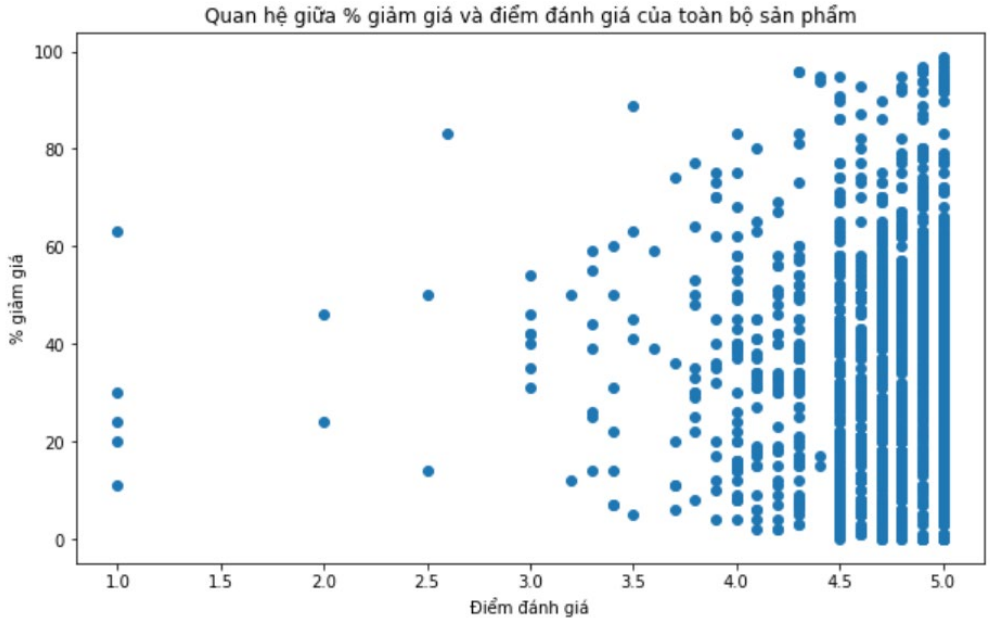
7.2. Nhóm 2: scatter plot, hubble plot, heatmap

Ở phương diện người dùng, nếu một sản phẩm nhiều người muốn mua giảm giá sâu, thì có thể doanh số bán ra của sán phẩm đó cao, ta đi xem thử quan hệ giữa % giảm giá và số bán được của toàn sản phẩm.



Ta thấy phần lớn tập trung ở số bán được rất thấp, có thể dữ liệu không đều, tuy nhiên nếu bỏ qua các sản phẩm có số bán ra bằng 0, thì ta thấy số bán được và % giảm giá khá tương quan nghịch nhau hơi ngược với ta kỳ vọng, ở mức giảm giá gần 100% thì chỉ bán được khoảng 25000, còn giảm giá khoảng 40% thì lại bản được rất nhiều. Có thể là các sản phẩm giảm sâu làm người dùng lo ngại chất lượng.

Ta đi xem thử quan hệ giữa %giảm giá và điểm đánh giá của sản phẩm, phương diện người dùng có thể nghĩ là %giảm cao thì có thể là xã hàng tồn kho, đánh giá chưa tốt.

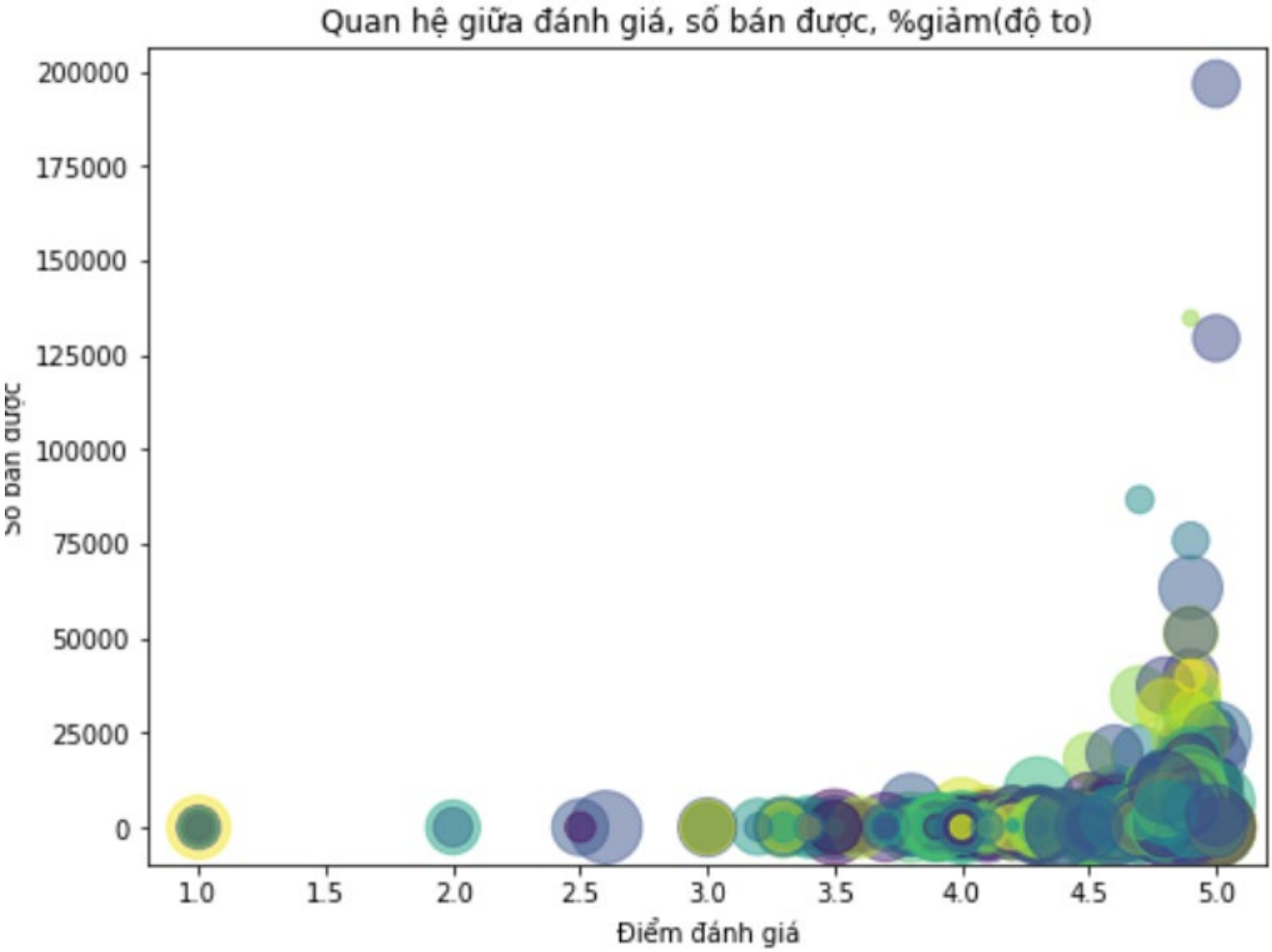


Phần lớn tập trung ở điểm đánh giá 4 và 5 sao, nhưng nhìn kỹ, ta thấy rằng chỉ có tại cột 4-5 sao thì % giảm giá mới có thể cao, còn tại cột 1 sao, 2 sao thì % giảm giá khá thấp. ta có thể tạm kết luận rằng chúng tương quan thuận, ngược với ta dự đoán trước khi trực quan.

Ta đi xem thử điểm đánh giá và số bán được, nếu đánh giá cao và tốt, có thể số bán được sẽ cao.

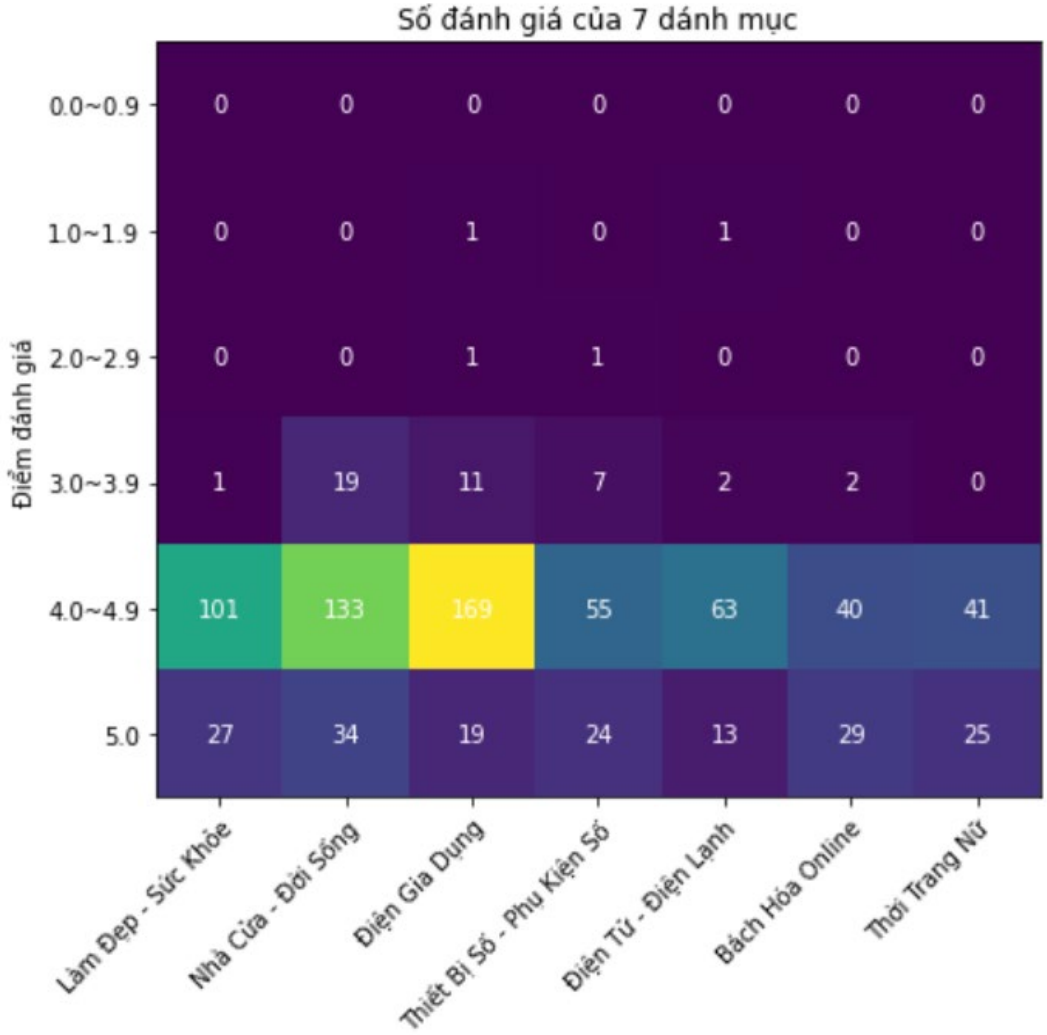
Lần này kết quả như ta dự đoán, tại các sản phẩm có đánh giá cao (4-5 sao) sẽ có số lượng bán ra rất cao, ta có thể tạm kết luận là chúng tương quan thuận nhau, nếu dùng để dự đoán, một sản phẩm bán rất nhiều nhiều thì ta cũng đoán được điểm chúng không thể dưới 3.

Bây giờ ta sẽ đi xét độ tương quan giữa điểm đánh giá, số bán được, % giảm giá (độ lớn bong bóng), chúng nó có cặp nghịch, có cặp thuận, bây giờ kết hợp xem sẽ ra sao.



Số bán được và % giảm giá, khi chỉ xét 2 trường này, chúng nó tương quan nghịch nhau, nhưng khi xét cùng với điểm đánh giá tại một vài trường hợp, cả 3 khá tương quan thuận nhưng ta chưa đủ dữ liệu để kết luận cả 3 tương quan thuận, tại hình này thì điểm đánh giá và số bán ra vẫn thể hiện độ tương quan thuận giữa chúng.

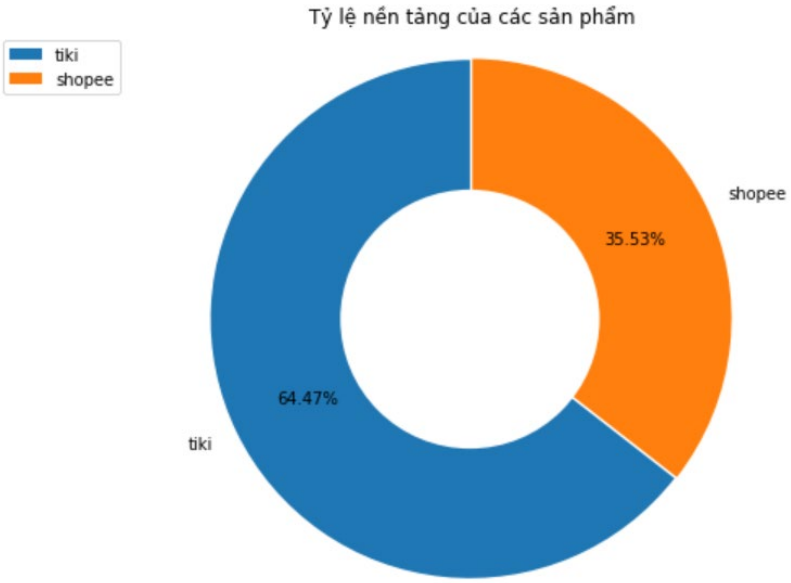
Nhận thấy việc điểm đánh giá khá “thân thiện”, dễ tương quan thuận, có thể giúp cho việc dự đoán sau này, ta sử dụng heatmap để xem các điểm đánh giá của 7 danh mục cấp 1 có nhiều sản phẩm nhất.



Danh mục Điện Gia Dụng có số đánh giá 4 sao khá cao, tuy nhiên danh mục Làm dep sức khỏe và Thời trang Nữ hoàn hảo hơn, vì rất ít hoặc không có đánh giá dưới 4 sao, có vẻ như các sản phẩm càng có nhiều đánh giá, sẽ có khả năng có vài đánh giá thấp xuất hiện.

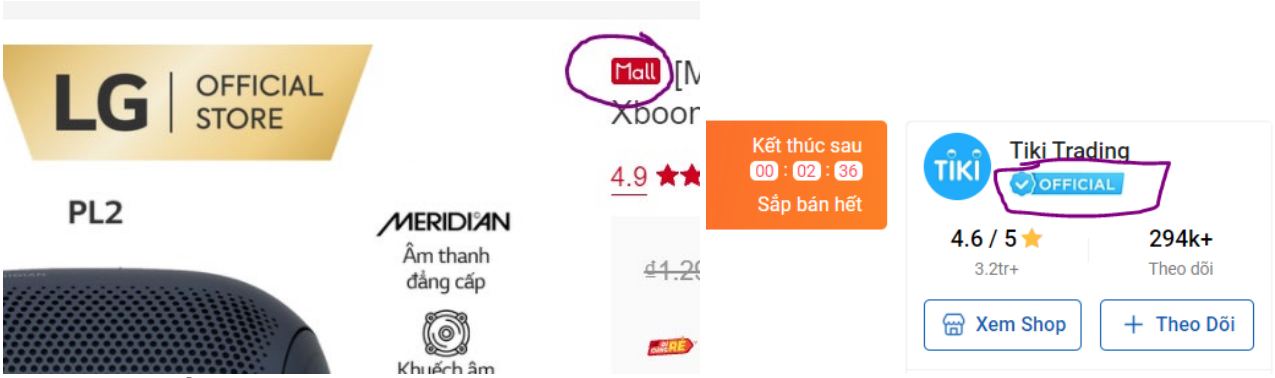
7.3. Nhóm 3: pie chart, donut chart, 100% stack bar chart

Nhóm đánh giá nền tảng các sản phẩm mà nhóm thu thập được.

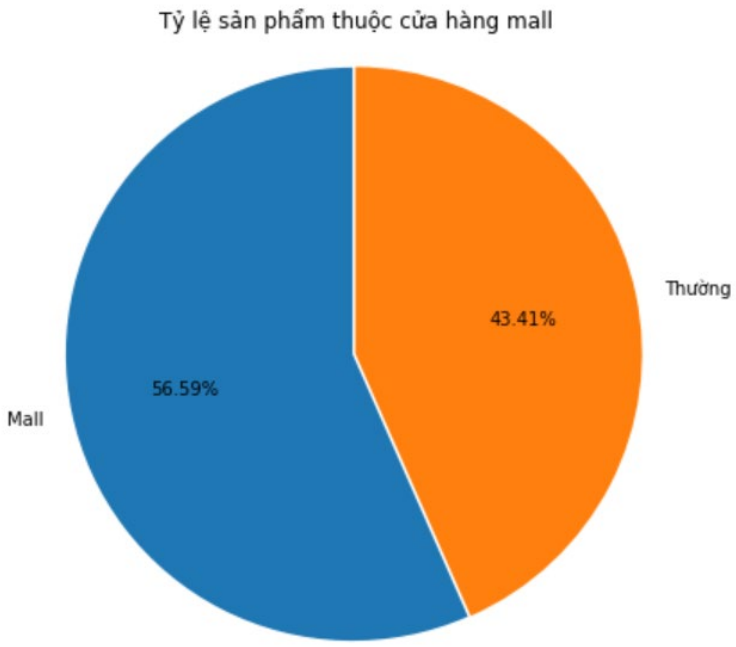


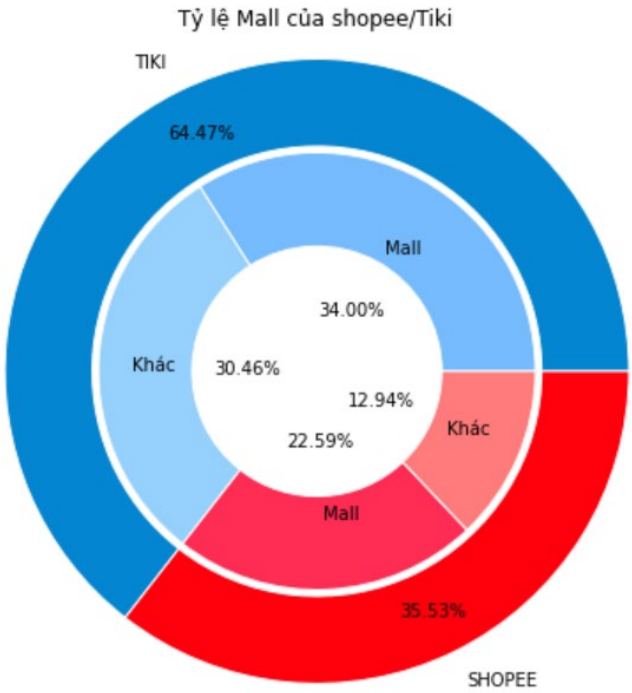
Hiện tại cả Shopee, Tiki (ngoài ra có Lazada) đang là các sàn thương mại lớn cạnh tranh nhau, ở hình trên ta thấy tiki chiếm phần nhiều so với shopee, có lẽ vì lúc crawl dữ liệu, thời gian phản hồi shopee chậm hơn, tuy nhiên bảo mật hơn, tiki kém bảo mật hơn nên nhóm đã crawl nhiều dữ liệu hơn.

Khi mua hàng người dùng sẽ khá chú trọng đến shop mall, hình bên dưới sẽ giải thích shop mall của tiki và shopee là gì.



Nếu sản phẩm thuộc về 1 cửa hàng mall (phân phối chính hãng), có lẽ người dùng sẽ tin tưởng hơn, sản phẩm bán ra cũng nhiều hơn, ta cùng đi xem thống kê số sản phẩm bán từ các cửa hàng mall của 2 nền tảng này.

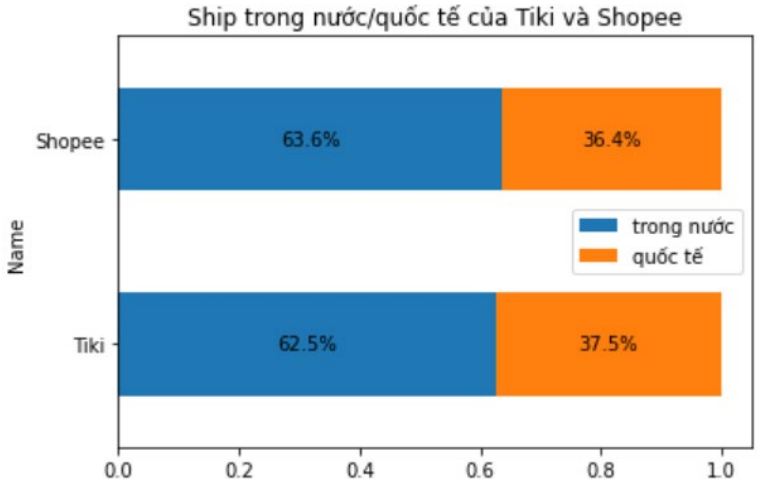




Ở hình tổng thể, số lượng sản phẩm của hàng mall chiếm 56%, có thể đây là lý do giải thích vì sao đa số sản phẩm có đánh giá từ 3~5 sao rất nhiều.
Hình đánh giá chi tiết, ta thấy rằng Tiki sẽ có 34% cửa hàng là tiki mall, cũng khá dễ hiểu vì khi phân tích trang web ở project 1, mua hàng ở tiki phần lớn sẽ thông qua “tiki trading”, còn các cửa hàng thường sẽ khả năng cao là cửa hàng ở nước ngoài. Riêng shopee tuy có số sản phẩm ít nhưng tỷ lệ mall tốt hơn tiki, 22% so với 12% là gần gấp

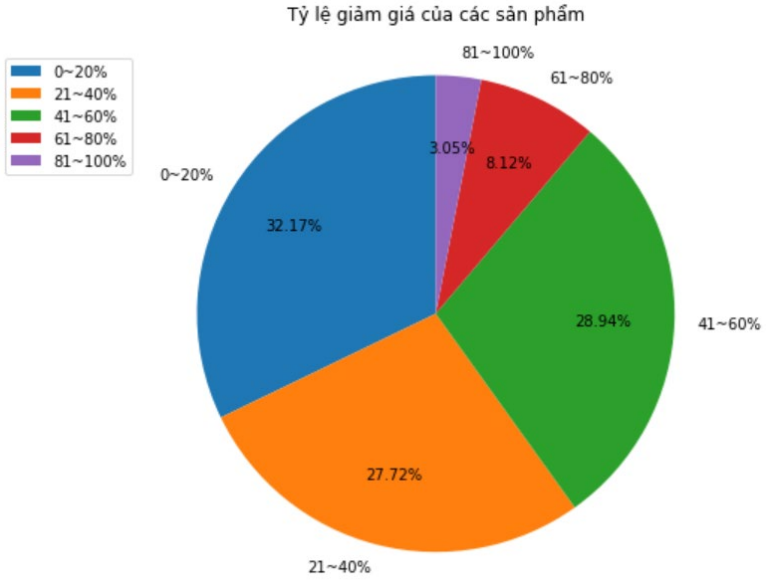
đôi, có vẻ như nếu 1 sản phẩm đưa vào Shopee, được đánh giá tốt nhiều, thì có khả năng đó là 1 sản phẩm của cửa hàng mall.

Ship từ đâu, hay là cửa hàng đặt ở đâu cũng sẽ ảnh hưởng tới người dùng quyết định mua hay không, từ đó ảnh hưởng đến số bán ra cũng như điểm đánh giá, đôi khi vận chuyển quá xa gây ảnh hưởng xấu sản phẩm, cũng kéo theo điểm đánh giá đi xuống, ta cùng xem thử 2 nền tảng này có tỷ lệ cửa hàng ngoài nước thế nào.



Có vẻ như cả 2 nền tảng có chỉ số cửa hàng trong ngoài nước khá đồng đều, nhưng ta thấy trong nước chiếm phần lớn, gần gấp đôi, có thể tạm cho là trong nước đánh giá sẽ tốt hơn.

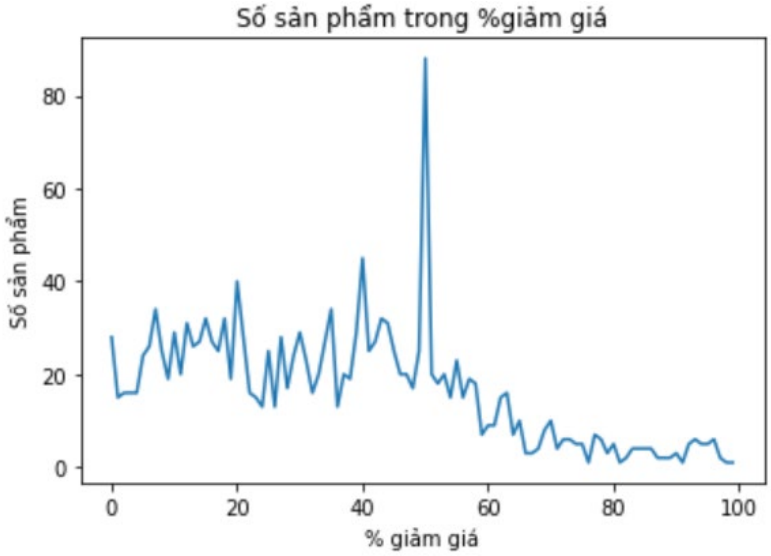
Hình này sẽ thống kể tỷ lệ giảm giá của toàn bộ sản phẩm



Phần lớn các sản phẩm sẽ được giảm từ 0~60%, điều này cũng lý giải vì sao lúc nãy ta thấy rằng % giảm giá cao chưa chắc số sản phẩm sẽ bán được ra cao.

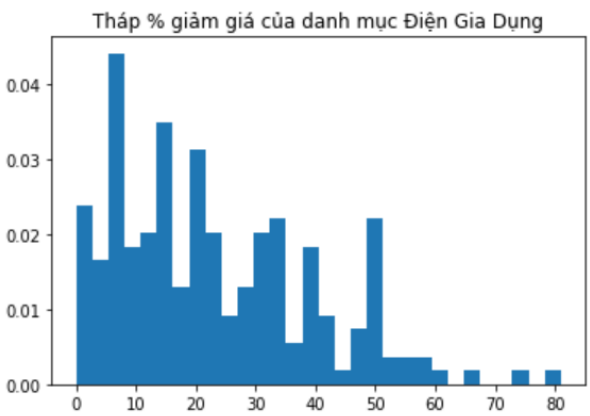
7.4. Nhóm 4: histogram, density

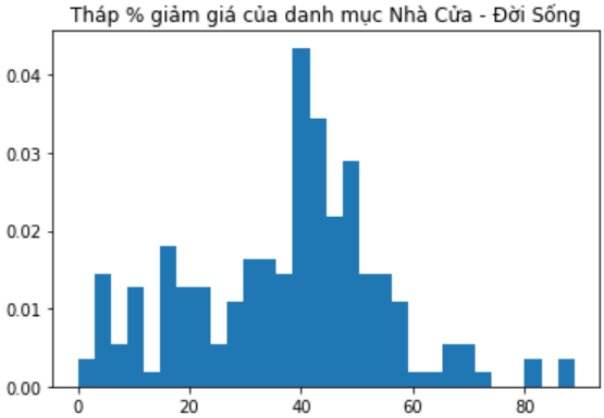
Ta dùng lại % giảm giá cho histogram/density

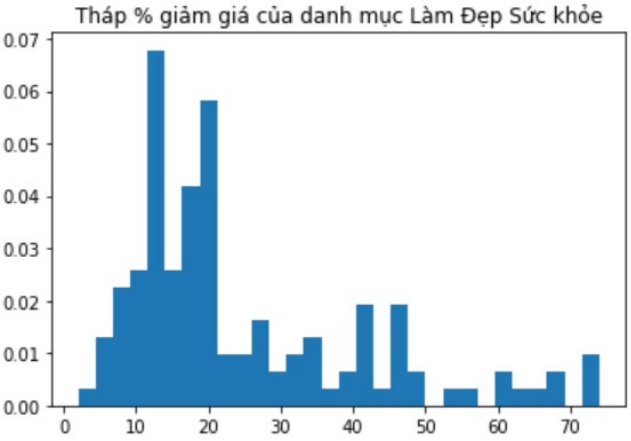


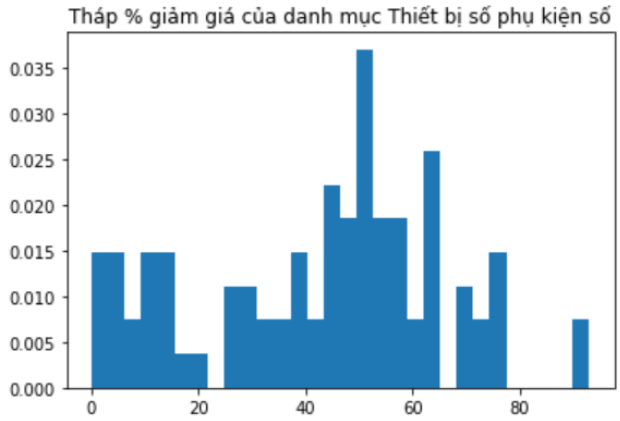
Giảm giá tập trung hầu hết và cao nhất ở khoảng 50%, có vẻ vì giảm quá thấp sẽ ít người mua, giảm quá sâu sẽ làm người ta quan ngại việc mua, nên các cửa hàng bán sẽ giảm mức vừa phải từ 20~60, đủ để thu hút người dùng mua, tăng doanh số bán ra.

Ta sẽ đi phân tích sâu hơn về % giảm giá của 4 danh mục sau:







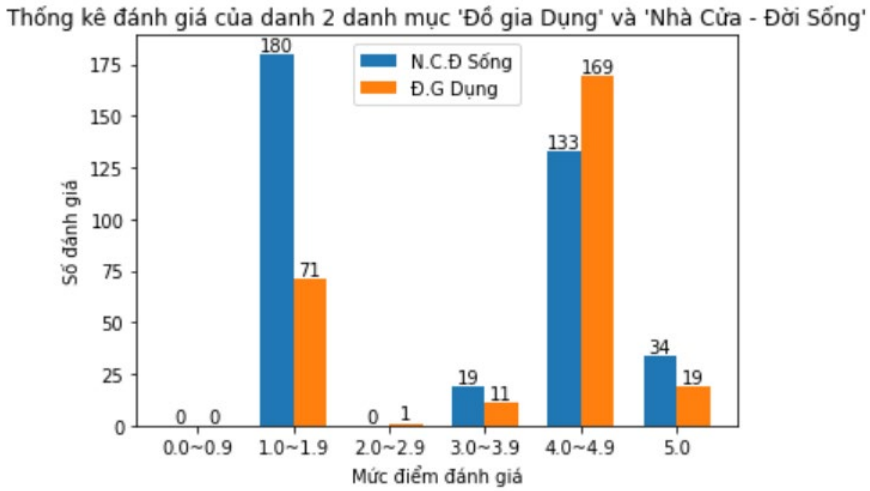


Nhận xét: cả 4 danh mục đều có rất ít sản phẩm giảm giá nhiều, phần lớn các sản phẩm sẽ tập trung giảm giá ở phần 10~60%, điều này giải thích vì sao lúc nãy ta phân tích quan hệ giữa % giảm giá với số bán ra có tương quan nghịch,vì hầu hết %giảm giá tập trung từ đầu hoặc giữa, giảm dần về sau.

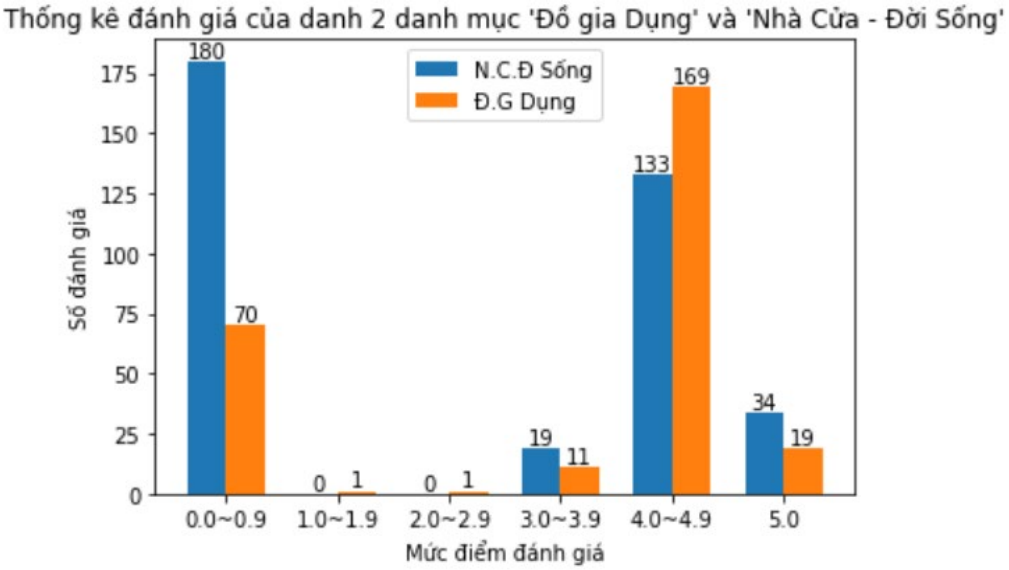
8. Góc nhìn khác

Khi nãy ở đầu báo cáo, nhóm có chỉ ra 3 cách lọc dữ liệu ở bước 2 nhỏ, các biểu đồ nãy giờ nhóm trình bày là sử dụng cách 3 (loại bỏ hẳn dữ liệu lỗi, thiếu trường), bây giờ nếu sử dụng cách tính mean cho các dữ liệu có khả năng sửa chữa (thiếu điểm đánh giá,…) thì sẽ thế nào, hoặc gán cho bằng 0 các dữ trường dữ liệu số bị thiếu sẽ như thế nào, ta sẽ đi so sánh vài biểu đồ giữa 3 cách.

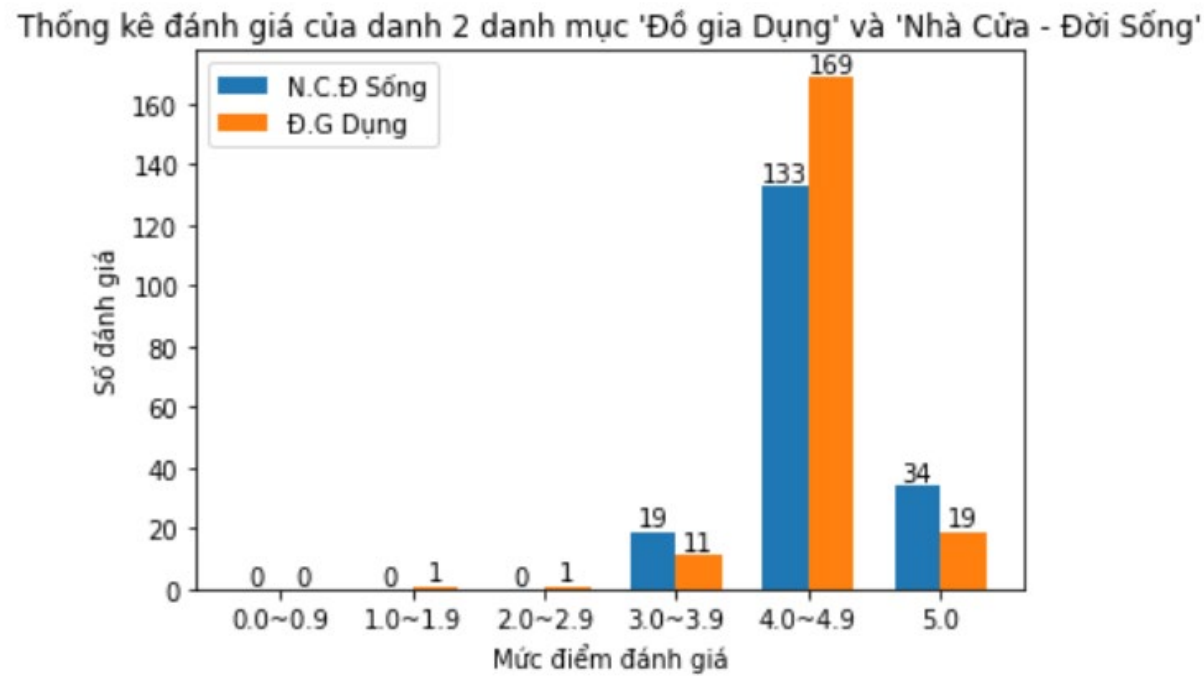
Cách thay bằng mean



Cách gán 0

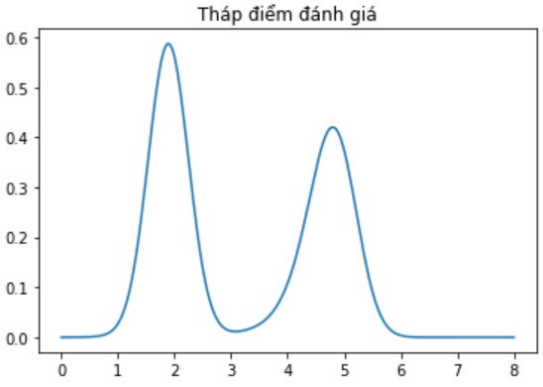


Cách loại bỏ hẳn

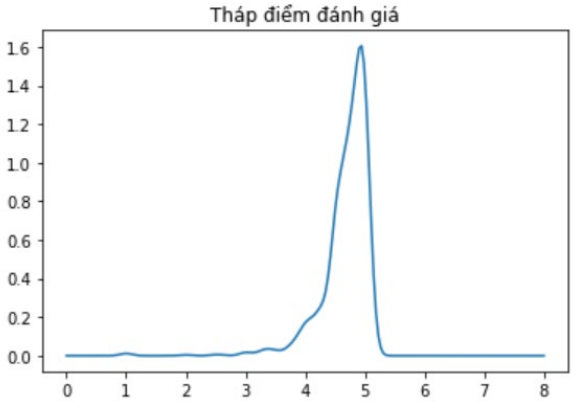


Biểu đồ tháp điểm.

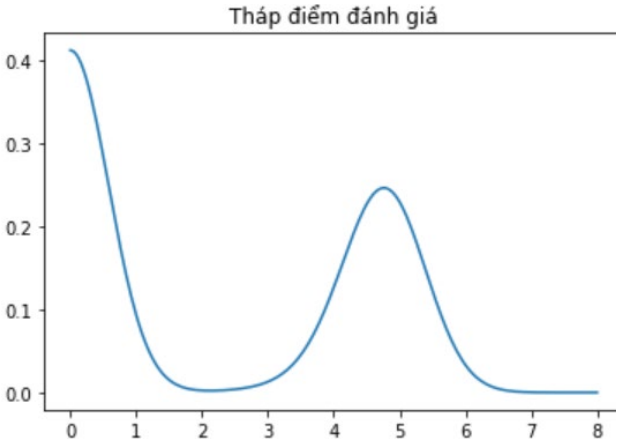
Cách gán mean



Cách loại bỏ hẳn

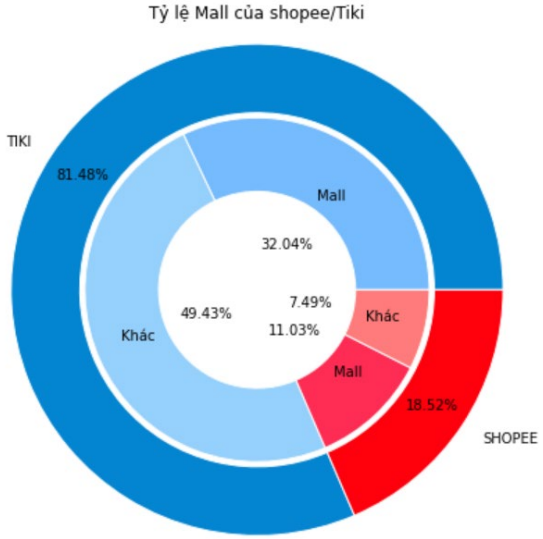


Cách gán 0

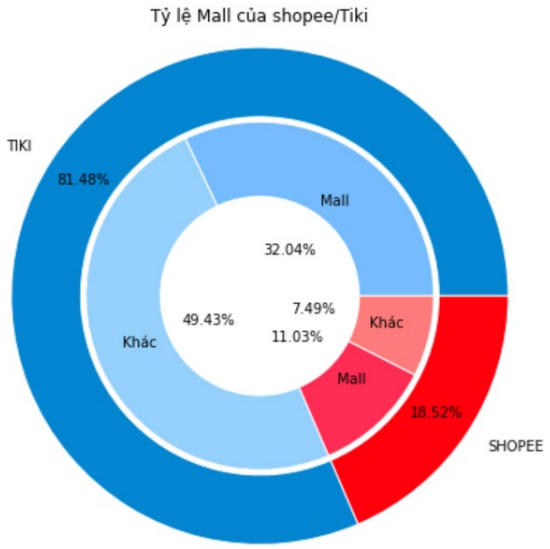


Biểu đồ tỷ lệ mall

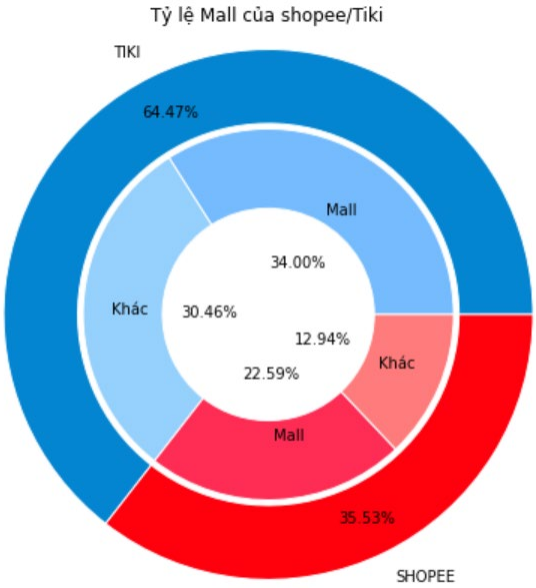
Cách gán mean



Cách gán 0



Cách loại bỏ hẳn

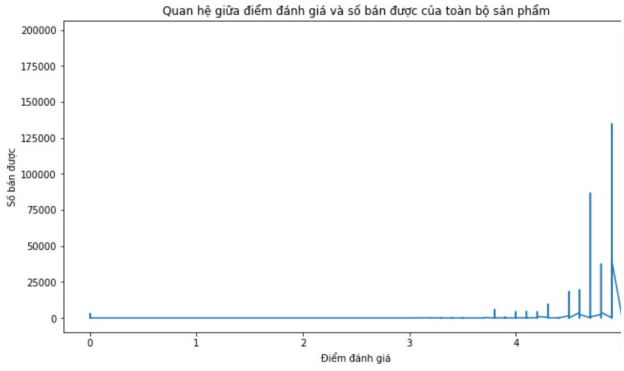


Biểu đồ tương quan điểm đánh giá và số bán được

Cách gán mean

Cách loại bỏ hẳn

Cách gán 0



9. Các nguồn tham khảo

- https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/#32.-Pie-Chart
- https://towardsdatascience.com/basics-of-donut-charts-with-pythons-matplotlib-100cf71b259d
- https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_and_donut_labels.html
- https://www.analyticsvidhya.com/blog/2021/06/donut-plots-data-visualization-with-python/
- https://www.geeksforgeeks.org/donut-chart-using-matplotlib-in-python/
- https://thecleverprogrammer.com/2020/12/16/donut-plot-with-python/
- https://towardsdatascience.com/how-to-create-and-customize-venn-diagrams-in-python-263555527305
- https://www.w3schools.com/python/matplotlib_line.asp
- https://pypi.org/project/matplotlib-venn/
- https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.sort_values.html
- https://stackabuse.com/matplotlib-line-plot-tutorial-and-examples/
- https://matplotlib.org/stable/gallery/statistics/hist.html
- https://www.geeksforgeeks.org/plotting-histogram-in-python-using-matplotlib/
- https://www.w3schools.com/python/matplotlib_histograms.asp
- https://matplotlib.org/stable/gallery/statistics/hist.html
- https://stackabuse.com/matplotlib-scatterplot-tutorial-and-examples/
- https://towardsdatascience.com/bubble-plots-in-matplotlib-3f0b3927d8f9
- https://matplotlib.org/devdocs/gallery/misc/packed_bubbles.html
- https://nirpyresearch.com/nir-data-correlograms-seaborn-python/
- https://matplotlib.org/stable/gallery/images_contours_and_fields/image_annotated_heatmap.html
- https://stackoverflow.com/questions/43566956/100-stacked-bar-chart-in-matplotlib
- https://stackoverflow.com/questions/53293382/how-to-make-subplots-in-donut-pie-chart-in-matplotlib-pyhon
- https://stackoverflow.com/questions/19841535/python-matplotlib-venn-diagram
- https://stackoverflow.com/questions/33203645/how-to-plot-a-histogram-using-matplotlib-in-python-with-a-list-of-data
- https://stackoverflow.com/questions/4150171/how-to-create-a-density-plot-in-matplotlib