

ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ





# BÁO CÁO

## ĐỒ ÁN (TIỂU LUẬN, BÀI TẬP LỚN)

**Học kỳ I, năm học 2021 - 2022**

**Học phần:**

### Phân tích dữ liệu với R

***Đề tài:*** Phân tích dữ liệu sản phẩm Lazada

**Số phách**

*(Do hội đồng chấm thi ghi)*

Thừa Thiên Huế, ngày 14 tháng 12 năm 2021



ĐẠI HỌC HUẾ

KHOA KỸ THUẬT VÀ CÔNG NGHỆ





# BÁO CÁO

## ĐỒ ÁN (TIỂU LUẬN, BÀI TẬP LỚN)

**Học kỳ I, năm học 2021 - 2022**

**Học phần:**

### Phân tích dữ liệu với R

***Đề tài:*** Phân tích dữ liệu sản phẩm Lazada

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

###### Sinh viên thực hiện: Trần Lý Bửu

###### Lớp: Khoa học dữ liệu và Trí tuệ nhân tạo – K1

**Số phách**

*(Do hội đồng chấm thi ghi)*

Thừa Thiên Huế, ngày 14 tháng 12 năm 2021

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

[BÁO CÁO 1](#_Toc92636123)

[Mục lục 2](#_Toc92636129)

[Chương 1: Tổng quan 3](#_Toc92636130)

[I/ Ngôn ngữ lập trình R 3](#_Toc92636131)

[II/ Ngôn ngữ lập trình Python 3](#_Toc92636132)

[Chương 2: Crawl và xử lý dữ liệu 5](#_Toc92636133)

[I/ Scrapy và Selenium 5](#_Toc92636134)

[II/ Xử lý dữ liệu 6](#_Toc92636135)

[Chương 3: Mô tả dữ liệu sản phẩm Lazada 8](#_Toc92636136)

[I/ Mô tả dữ liệu 8](#_Toc92636137)

[II/ Phân tích số liệu 16](#_Toc92636138)

[Danh mục tham khảo 20](#_Toc92636139)

# Chương 1: Tổng quan

## I/ Ngôn ngữ lập trình R

**R là ngôn ngữ lập trình thường được sử dụng trong học máy, thống kê, phân tích dữ liệu và nghiên cứu khoa học. Đây là một trong những ngôn ngữ phổ biến nhất được sử dụng bởi các nhà thống kê, phân tích dữ liệu, nhà nghiên cứu và tiếp thị, phân tích, trực quan hóa và trình bày dữ liệu. Nó bao gồm các thuật toán học máy, hồi quy tuyến tính, chuỗi thời gian, suy luận thống kê. Hầu hết các thư viện R được viết bằng R, nhưng đối với các tác vụ tính toán nặng, mã C, C ++ và Fortran được ưu tiên.**

Bạn có thể chạy code mà không cần đến bất cứ compiler nào - R là ngôn ngữ thông dịch ( interpreted language ). Do đó code có thể chạy mà không cần compiler. R thông dịch code và làm cho việc viết code đơn giản hơn, dễ phát triển hơn.

Bất kỳ một phép tính nào cũng có thể thực hiện trên vectors - R là một vector-language, do đó chúng ta có thể dùng bất kỳ function nào trên một vector mà không cần phải dùng vòng lặp.

Ưu điểm của ngôn ngữ R:

* R có những package thống kê toàn diện nhất với công nghệ mới nhất, những ý tưởng mới thường xuất hiện đầu tiên trên R.
* R là nguồn mở và miễn phí nên bất kỳ ai cũng có thể sử dụng và cải tiến nó.
* Ngôn ngữ R là mã nguồn mở nên ai trong chúng ta cũng có thể phân tích source code để hiểu được chính xác cách R vận hành. Bất kỳ ai cũng có thể thêm tính năng và fix bug mà không cần chờ nhà phát hành ra bản vá. Đồng thời, R có thể tích hợp được với ngôn ngữ khác (C,C++). Nó cũng cho phép chúng ta tương tác với nhiều nguồn dữ liệu và các gói thống kê (SAS, SPSS).
* R có thể chạy trên bất kỳ hệ điều hành nào.
* Hệ sinh thái trọn gói mạnh mẽ

Hạn chế của ngôn ngữ R:

* Một vài package của R có thể không hoàn hảo và còn lỗi
* Không có ai để "complain" cho việc code không chạy
* Quản lý bộ nhớ kém
* Thiếu tính năng bảo mật.

## II/ Ngôn ngữ lập trình Python

Python là [ngôn ngữ lập trình](https://magenest.com/vi/ngon-ngu-lap-trinh/) hướng đối tượng đa năng. Ngôn ngữ này sở hữu cấu trúc dữ liệu cấp cao mạnh mẽ và hệ thống thư viện lớn. Python sử dụng cơ chế cấp phát bộ nhớ tự động với cú pháp đơn giản và rõ ràng, giúp người học dễ tiếp cận và làm quen, kể cả đối với những người mới bắt đầu học lập trình.

Với đặc điểm là một ngôn ngữ dễ học, dễ đọc với bộ mã nguồn và bố cục trực quan dễ hiểu, về cơ bản ngôn ngữ Python có các điểm sau:

* Từ khóa: Ngôn ngữ Python sử dụng chủ yếu các từ khóa bằng tiếng Anh, và hạn chế sử dụng các từ khóa thuộc ngôn ngữ khác. Python cũng phân biệt các kiểu chữ hoa hay chữ thường.
* Khối lệnh: Python sử dụng lùi đầu dòng (indentation) như dấu cách hay tab để tạo các khối lệnh. Các khối lệnh con sẽ ở vị trí thụt lùi đầu dòng vào sâu hơn so với khối lệnh cha mà nó trực thuộc.
* Python có khả năng mở rộng linh hoạt: Nếu bạn biết ngôn ngữ lập trình C, bạn có thể dễ dàng tích hợp ngôn ngữ Python vào việc viết các đoạn hàm khác nhau đáp ứng nhu cầu của bạn. Bạn có thể mở rộng chức năng của trình thông dịch, hoặc liên kết trình thông dịch Python với các ứng dụng viết từ ngôn ngữ C để mở rộng hoặc tạo lệnh phụ trợ cho ứng dụng đó.

Bộ thư viện chuẩn rộng lớn của Python, thường được cho là một trong những điểm mạnh lớn nhất của nó. cung cấp các công cụ phù hợp cho nhiều công việc khác nhau. Với các ứng dụng giao tiếp với Internet, nhiều giao thức và định dạng chuẩn chẳng hạn như [MIME](https://vi.wikipedia.org/wiki/MIME) and [HTTP](https://vi.wikipedia.org/wiki/HTTP) được hỗ trợ. Nó cũng có chứa các mô đun đành cho việc tạo lập [giao diện người dùng đồ hoạ](https://vi.wikipedia.org/wiki/Giao_di%E1%BB%87n_ng%C6%B0%E1%BB%9Di_d%C3%B9ng_%C4%91%E1%BB%93_ho%E1%BA%A1), kết nối với [cơ sở dữ liệu quan hệ](https://vi.wikipedia.org/wiki/C%C6%A1_s%E1%BB%9F_d%E1%BB%AF_li%E1%BB%87u_quan_h%E1%BB%87), [sinh số giả ngẫu nhiên](https://vi.wikipedia.org/wiki/B%E1%BB%99_sinh_s%E1%BB%91_gi%E1%BA%A3_ng%E1%BA%ABu_nhi%C3%AAn), tính toán với số thập phân có độ chính xác tuỳ ý, thao tác với [biểu thức chính quy](https://vi.wikipedia.org/wiki/Bi%E1%BB%83u_th%E1%BB%A9c_ch%C3%ADnh_quy) và kiểm thử đơn vị.

# Chương 2: Crawl và xử lý dữ liệu

## I/ Scrapy và Selenium

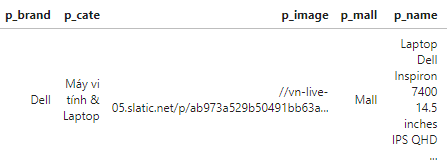
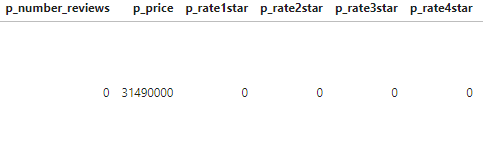
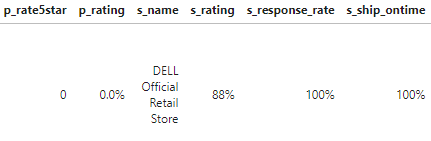
Sự bùng nổ của Internet đã mang lại lợi ích cho những người đam mê [khoa học dữ liệu](http://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=WebScrapinginPythonarticle) . Sự đa dạng và số lượng dữ liệu có sẵn ngày nay thông qua internet giống như một kho tàng bí mật và bí ẩn đang chờ được giải đáp. Bạn hoàn toàn có thể sử dụng Scrapy để lấy số dữ liệu đấy.

Scrapy là một Python framework để quét web với quy mô lớn. Nó cung cấp cho bạn tất cả các công cụ bạn cần để trích xuất dữ liệu từ các trang web một cách hiệu quả, xử lý chúng theo ý muốn và lưu trữ chúng ở định dạng và cấu trúc mà bạn muốn.

Selenium là một công cụ tự động hóa dựa trên web mã nguồn mở. Ngôn ngữ Python được sử dụng với Selenium để tự động hóa. Các API Python cho phép bạn kết nối với trình duyệt thông qua Selenium. Vậy tại sao phải sử dụng cả Scrapy và Selenium? Một số trang web được đội ngũ lập trình viên sử dụng ngôn ngữ lập trình Javascript và trong đó có Lazada.

Scrapy-selenium là một phần mềm trung gian được sử dụng trong việc quét web. scrapy không hỗ trợ tìm kiếm các trang web hiện đại sử dụng khung javascript và đây là lý do mà phần mềm trung gian này được sử dụng với scrapy để quét các trang web hiện đại đó. Scrapy-selenium cung cấp các chức năng của selen giúp làm việc với các trang web javascript. Những lợi thế khác được cung cấp bởi điều này là trình điều khiển mà chúng ta cũng có thể thấy những gì đang xảy ra đằng sau hậu trường. Vì selen là công cụ tự động nên nó cũng cung cấp cho chúng ta cách xử lý các thẻ đầu vào và quét theo những gì bạn chuyển vào trường nhập. Việc chuyển đầu vào trong các trường đầu vào trở nên dễ dàng hơn bằng cách sử dụng selen. Scrapy-selenium lần đầu tiên được giới thiệu vào năm 2018 và là một nguồn mở của nó.

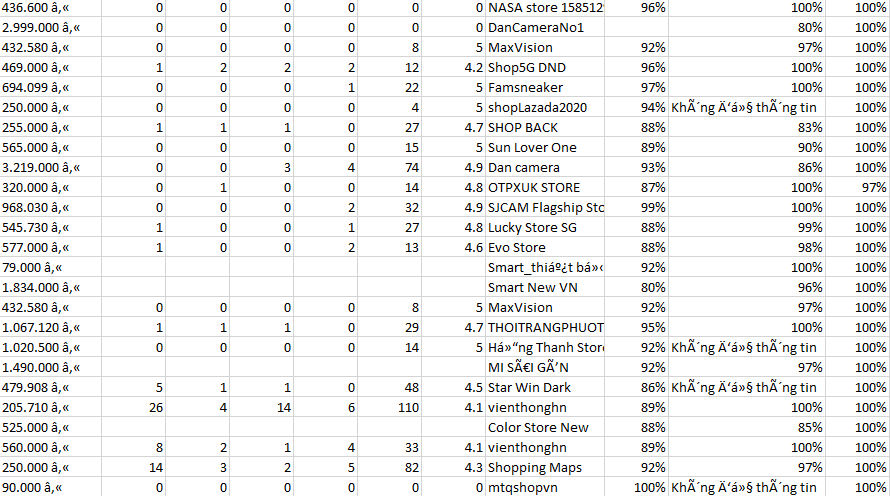
Kết quả thu được từ việc crawl dữ liệu chính là các file csv. Dữ liệu lấy được có 17 giá trị:

* **p\_name** - Tên sản phẩm
* **p\_brand** - Tên thương hiệu sản phẩm
* **p\_cate** - Danh mục sản phẩm
* **p\_price** - Giá sản phẩm
* **p\_rating** - Đánh giá trung bình của sản phẩm
* **p\_image** - Đường dẫn ảnh sản phẩm
* **p\_mall** - Chứng nhận Lazada Mall
* **p\_number\_reviews** - Tổng số lượt đánh giá sản phẩm
* **p\_rate1star** - Tổng số đánh giá 1 sao cho sản phẩm
* **p\_rate2star** - Tổng số đánh giá 2 sao cho sản phẩm
* **p\_rate3star** - Tổng số đánh giá 3 sao cho sản phẩm
* **p\_rate4star** - Tổng số đánh giá 4 sao cho sản phẩm
* **p\_rate5star** - Tổng số đánh giá 5 sao cho sản phẩm
* **s\_name** - Tên nhà bán hàng
* **s\_rating** - Tỉ lệ đánh giá cho nhà bán hàng
* **s\_response\_rate** - Tỉ lệ nhà bán hàng phản hồi
* **s\_ship\_ontime** - Tỉ lệ nhà bán hàng giao hàng đúng giờ

Source code: <https://github.com/tranlybuu/Data_Lazada/tree/master/lazada>

## II/ Xử lý dữ liệu

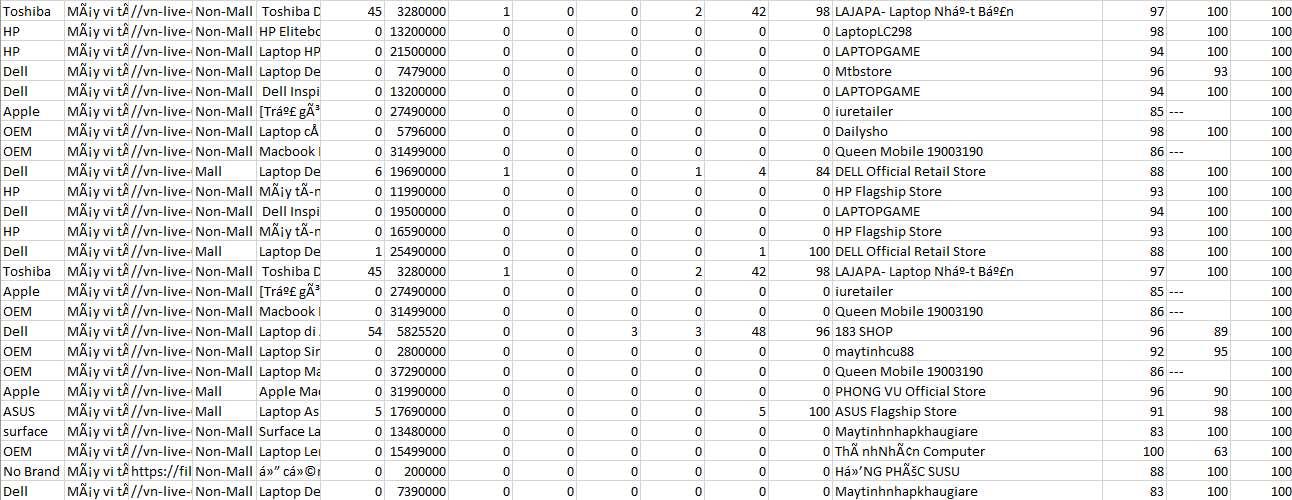
Dữ liệu crawl được không phải lúc nào cũng đầy đủ và việc mất dữ liệu sẽ xảy ra thường xuyên. Việc bạn cần làm là xử lý dữ liệu sẽ giúp việc thống kê và mô tả dữ liệu sau này chính xác hơn



Dữ liệu trước khi xử lý

Việc mất dữ liệu sẽ xuất phát từ 2 nguyên nhân chính. Đầu tiên, trang web sẽ ngẫu nhiên xuất hiện quảng cáo pop-up dẫn đến những đoạn địa chỉ chúng ta code trong scrapy sẽ không tìm được đến dữ liệu ta cần. Một số cửa hàng mới mở trên Lazada nên hệ thống chưa thể thống kê được và kết quả trả về sẽ là “Không có thông tin”.

Để thống nhất trong kết quả xuất ra thì tôi đã chuyển những đánh giá được tính trung bình dành cho sản phẩm và cửa hàng thành thang điểm 100 và xóa dấu %. Giá tiền sản phẩm sẽ có ký tự ₫ và có thể mỗi thư viện hoặc phần mềm sẽ phiên dịch ký tự ₫ thành các ký tự khác nhau. Xóa các dòng bị thiếu dữ liệu lúc crawl sẽ giúp mô tả dữ liệu chính xác hơn và chuyển những ô dữ liệu “Không có thông tin” thành “---”. Ngoài ra, một số sản phẩm có dấu phẩy “,” trong tên sản phẩm có thể làm dữ liệu bị sai với cấu trúc ban đầu dẫn đến quá trình mô tả dữ liệu sẽ không chính xác.



Dữ liệu sau khi xử lý

Với việc mỗi lần sử dụng bot crawl để lấy dữ liệu sản phẩm trên Lazada sẽ xuất ra một file csv riêng nên tôi đã viết một Python script để đọc toàn bộ các file csv dữ liệu chưa qua xử lý để tiến hành lọc và xử lý dữ liệu. Kết quả đầu ra sẽ là một file csv có tên **LazData.csv**. Mỗi lần có them file dữ liệu chưa xử lý thì tôi chỉ cần chạy script trên và dữ liệu trong **LazData.csv** sẽ tự được cập nhật thêm và có thể dung file csv này để mô tả dữ liệu.

Source code: <https://github.com/tranlybuu/Data_Lazada/blob/master/update_data.py>

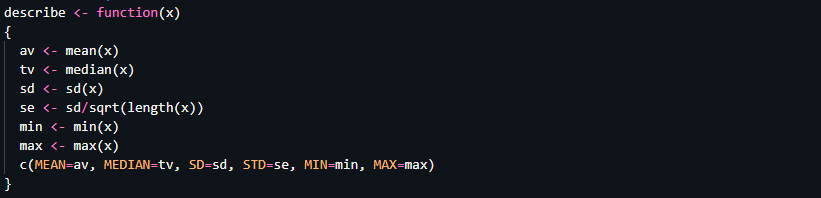
Dataset: <https://www.kaggle.com/tranlybuu/lazada-product-data>

# Chương 3: Mô tả dữ liệu sản phẩm Lazada

## I/ Mô tả dữ liệu

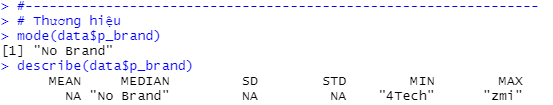
Hàm tính mode

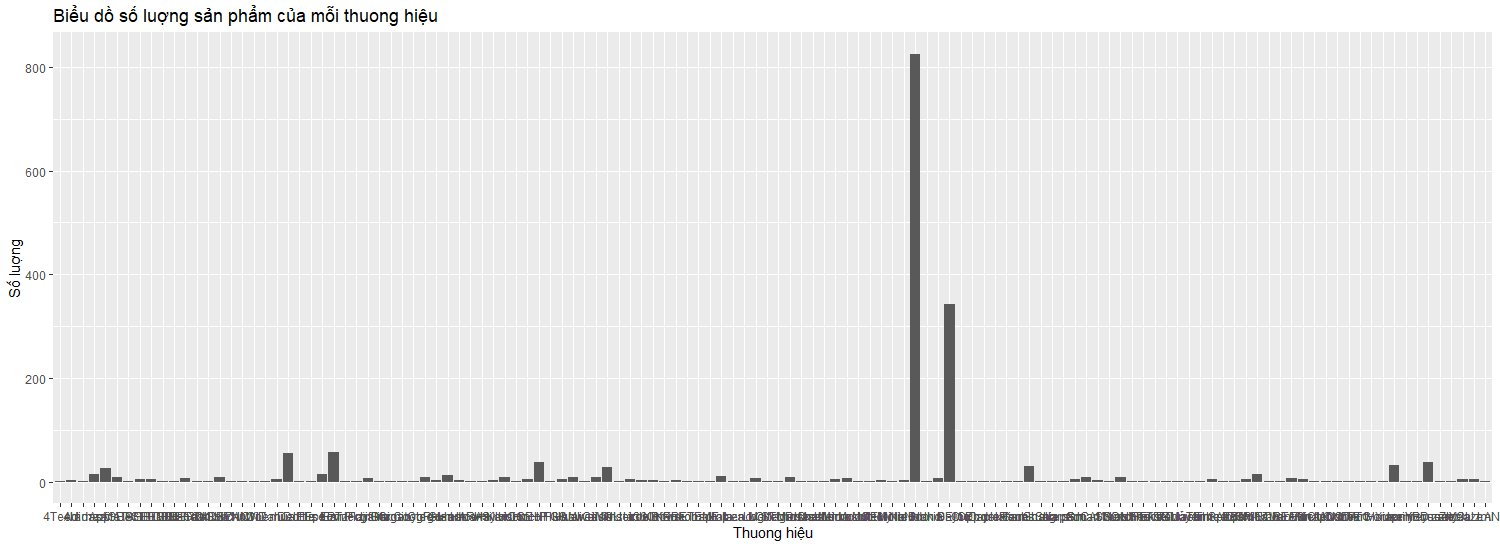


Hàm mô tả tổng qua về dữ liệu gồm: Mean,Median, Số trung vị, Min,… 

Tên sản phẩm (p\_name) và đường dẫn ảnh sản phẩm (p\_image) không mô tả

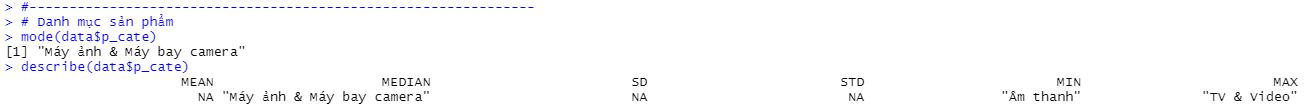
* **p\_brand** - Tên thương hiệu sản phẩm

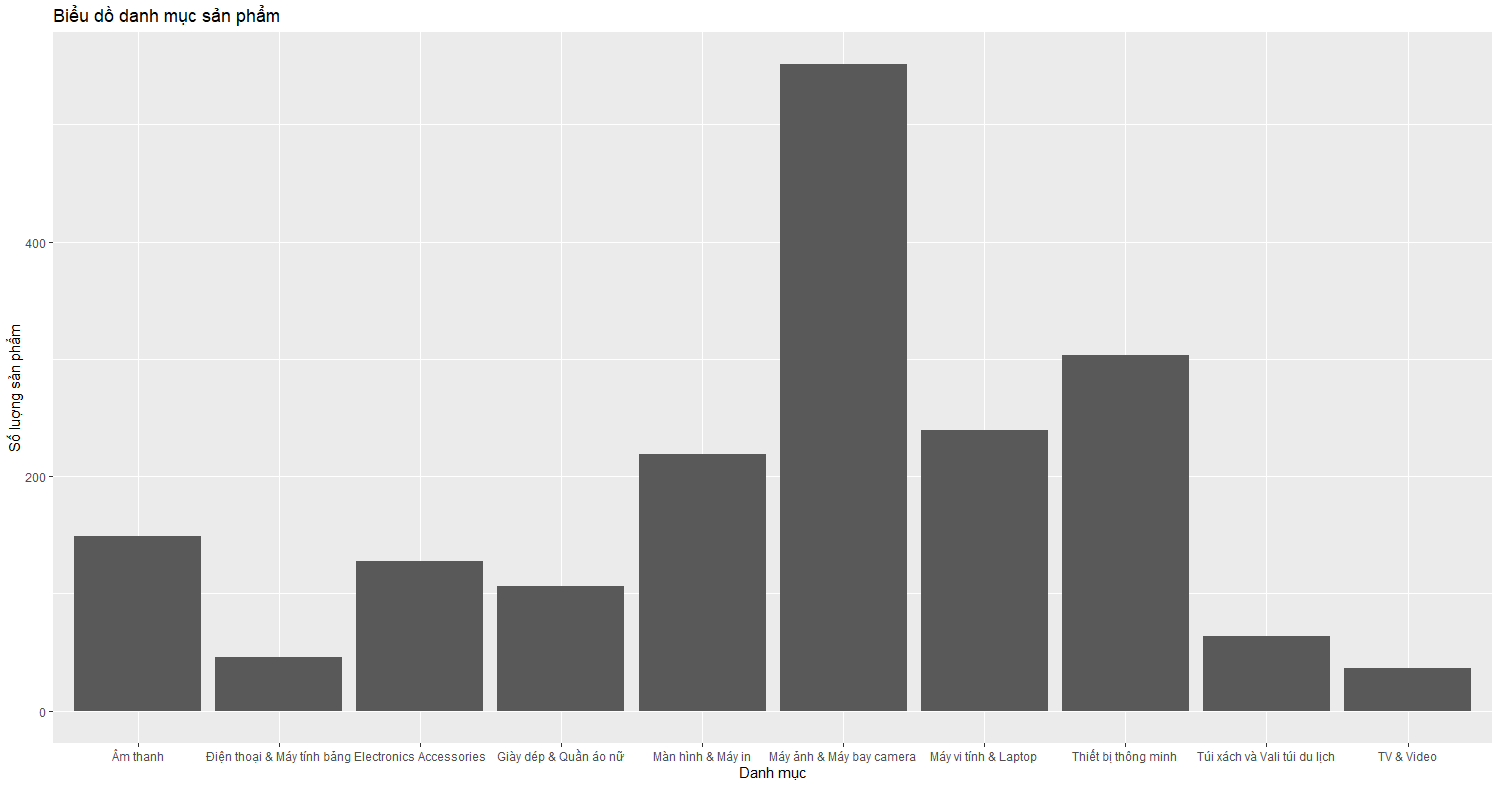




Các sản phẩm trên Lazada chủ yếu không có thương hiệu rõ ràng, hàng OEM và các thương hiệu nhỏ

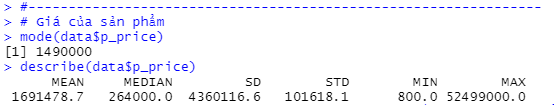
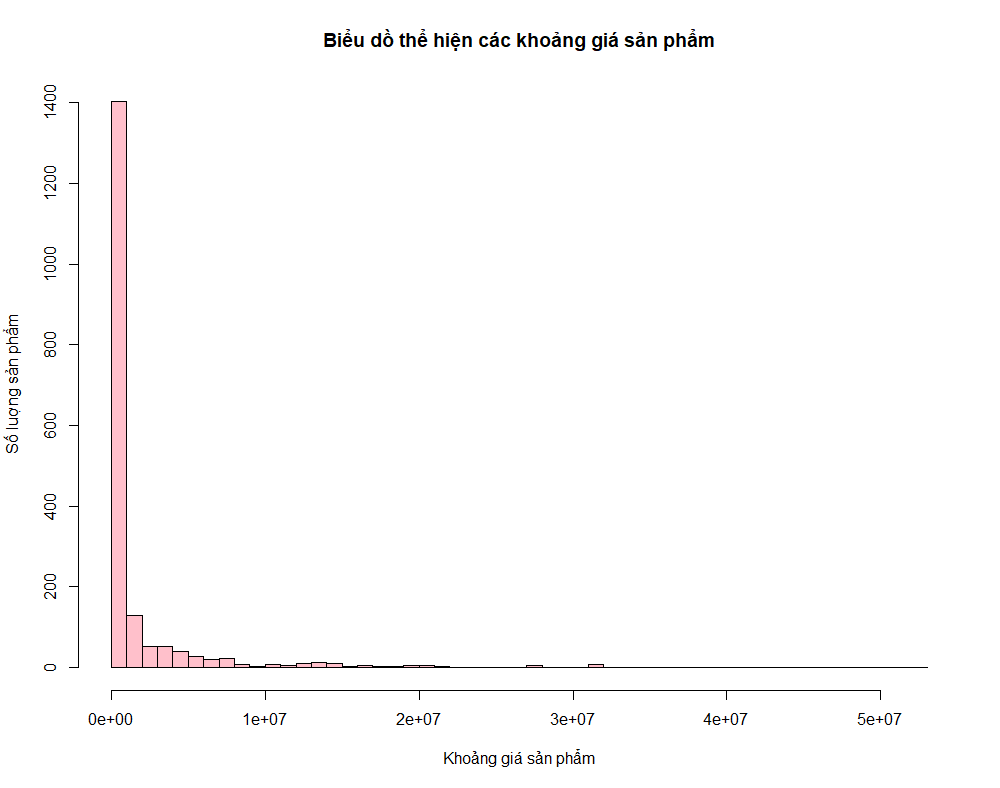
* **p\_cate** - Danh mục sản phẩm





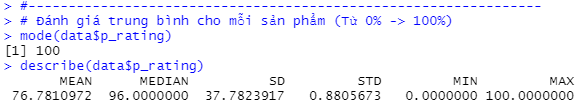
Danh mục được nhiều cửa hàng đăng bán là Camera và thiết bị thông minh

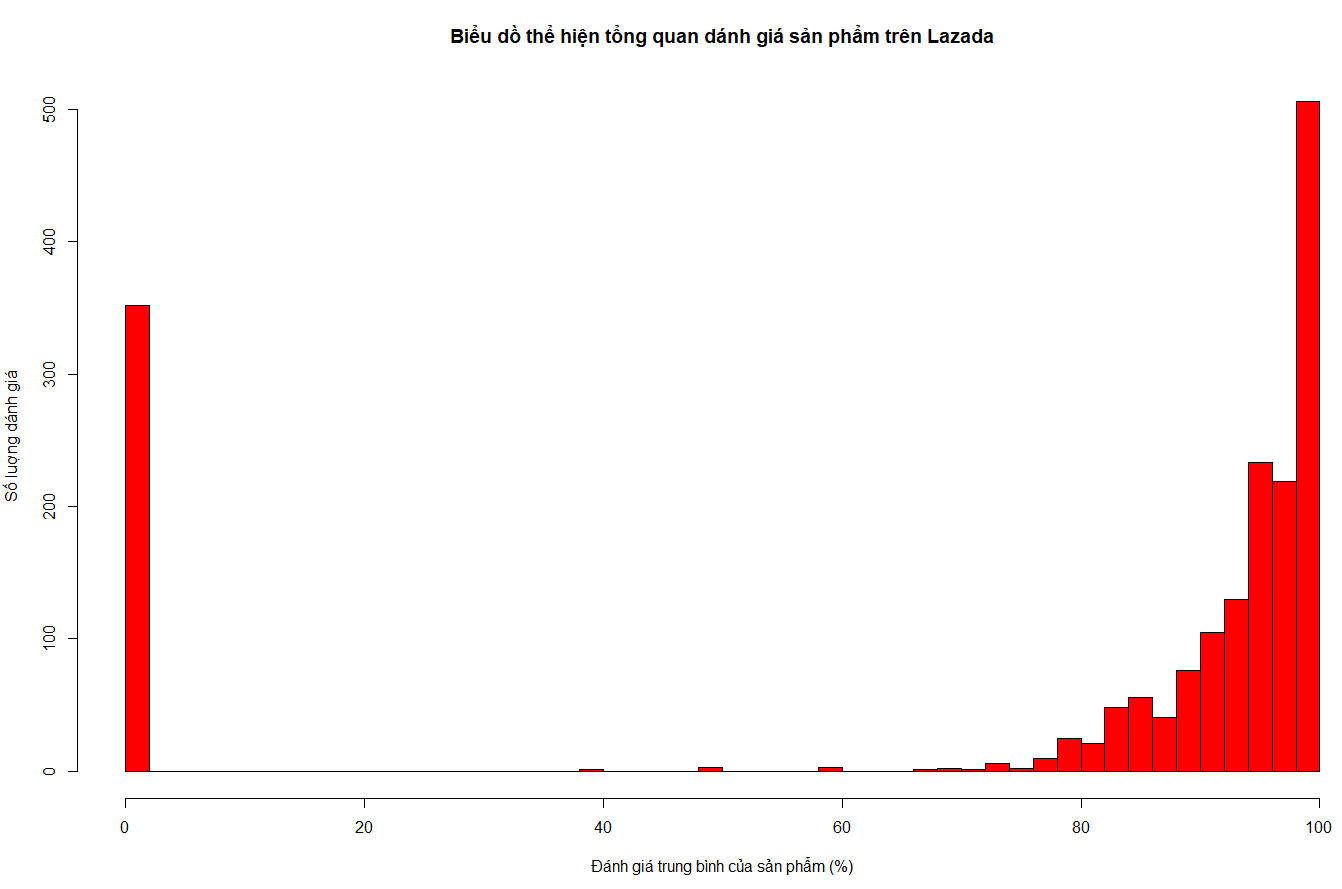
* **p\_price** - Giá sản phẩm

Đa số người dùng tìm đến Lazada để mua các sản phẩm có khoảng tiền nhỏ. Đó là khoảng tiền họ có thể đánh đổi việc xem mặt hàng trực tiếp để lấy giá thành rẻ hơn và tiết kiệm thời gian để đi nhiều cửa hàng lựa chọn

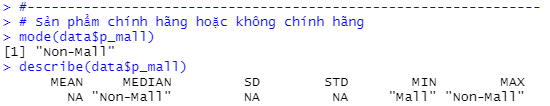
* **p\_rating** - Đánh giá trung bình của sản phẩm

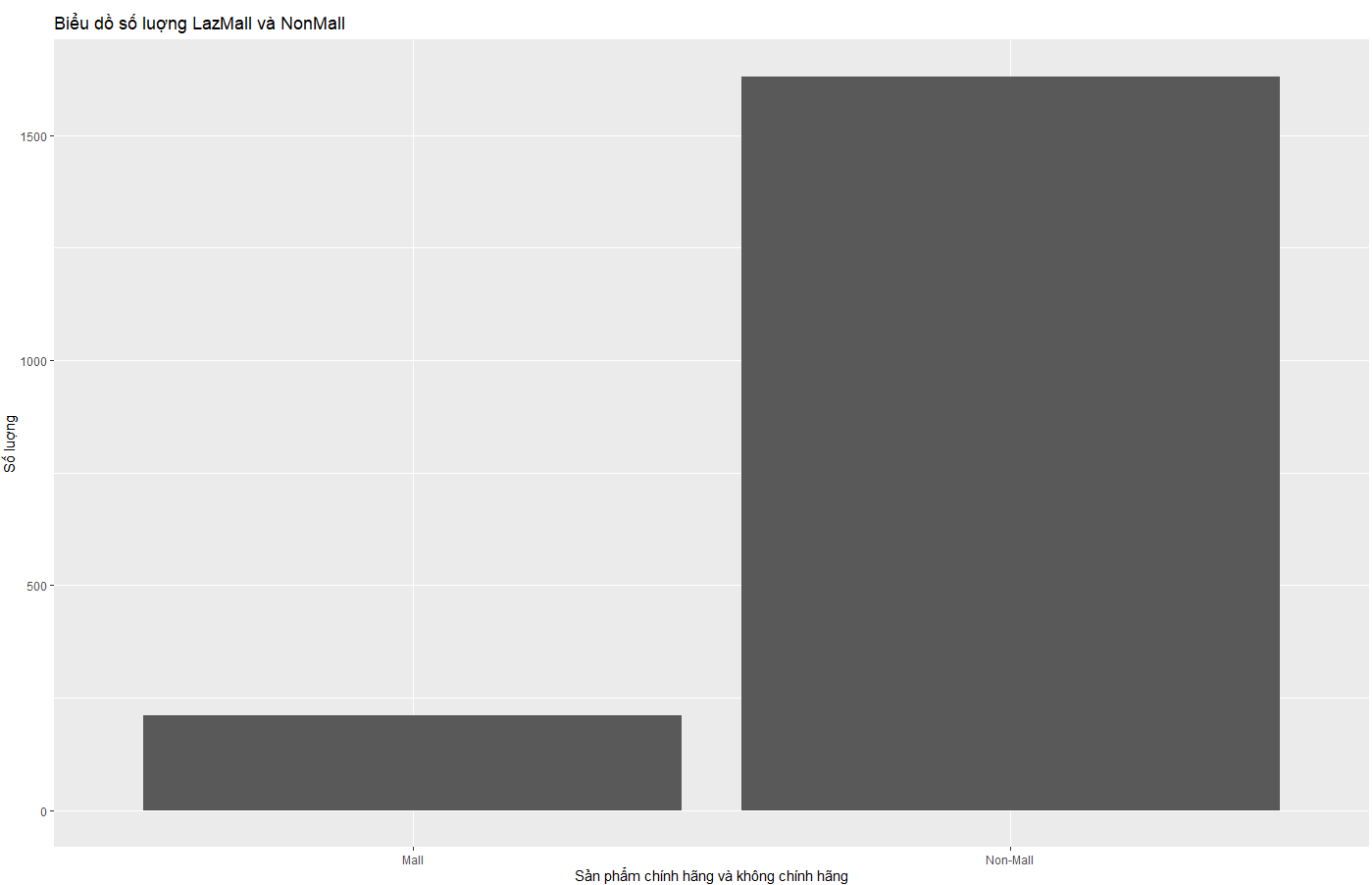




Một số cửa hàng bán sản phẩm không chất lượng nhưng đa số sản phẩm trên Lazada đều được mọi người đánh giá cao.

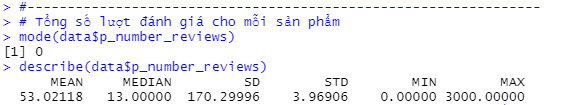
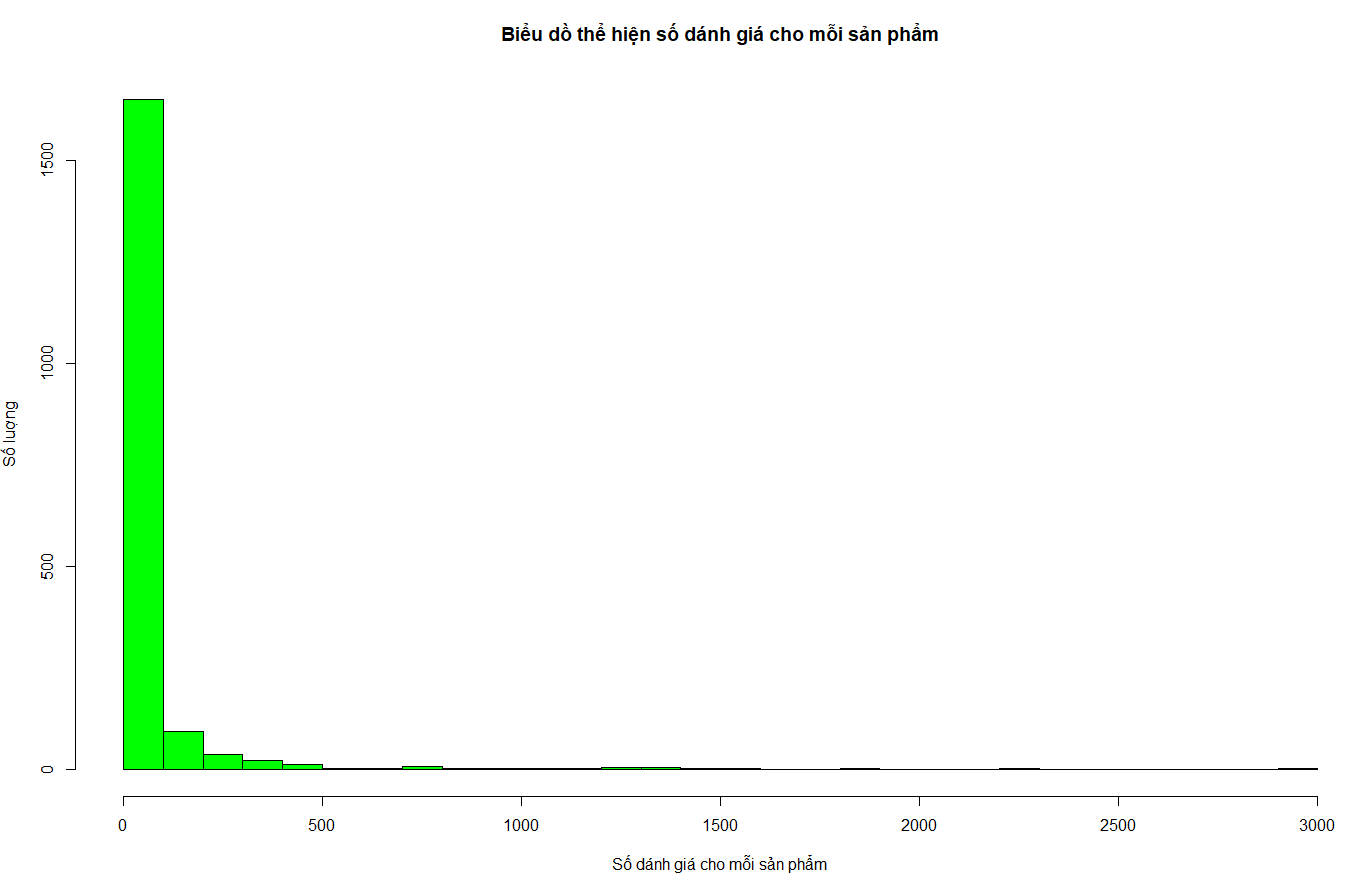
* **p\_mall** - Chứng nhận Lazada Mall





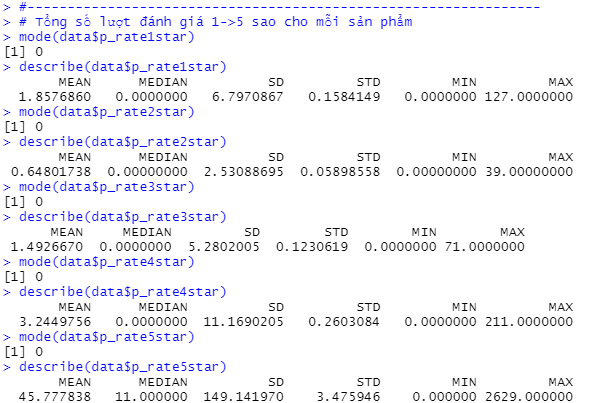
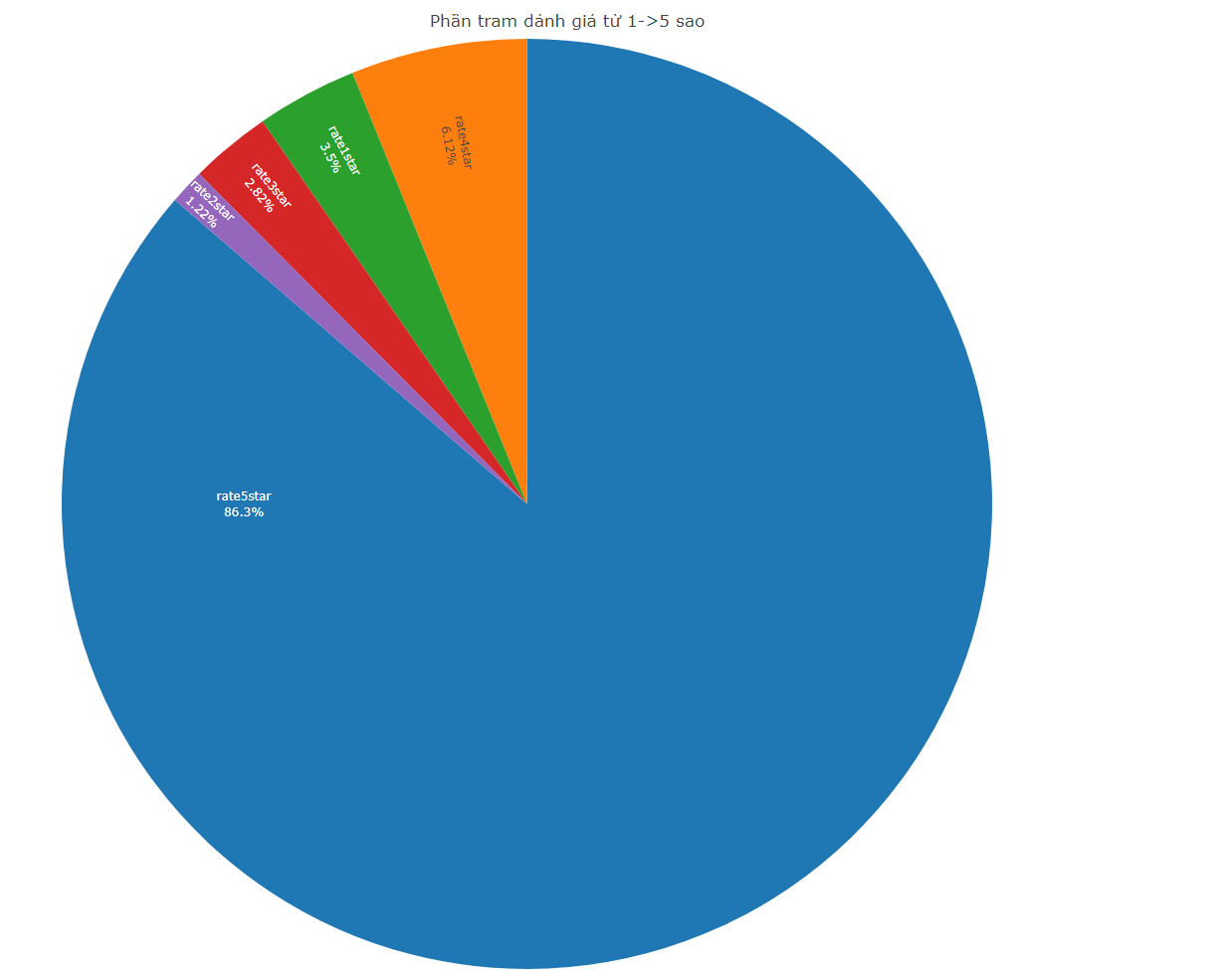
Các sản phẩm đang được đăng bán trên Lazada chủ yếu là sản phẩm không chính hãng

* **p\_number\_reviews** - Tổng số lượt đánh giá sản phẩm

Nếu quy ước mỗi lượt đánh giá là một lần mua sản phẩm thì hầu hết sản phẩm đều ít được khách hàng tin tưởng mua hàng trên Lazada

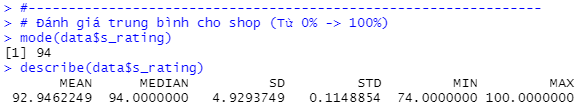
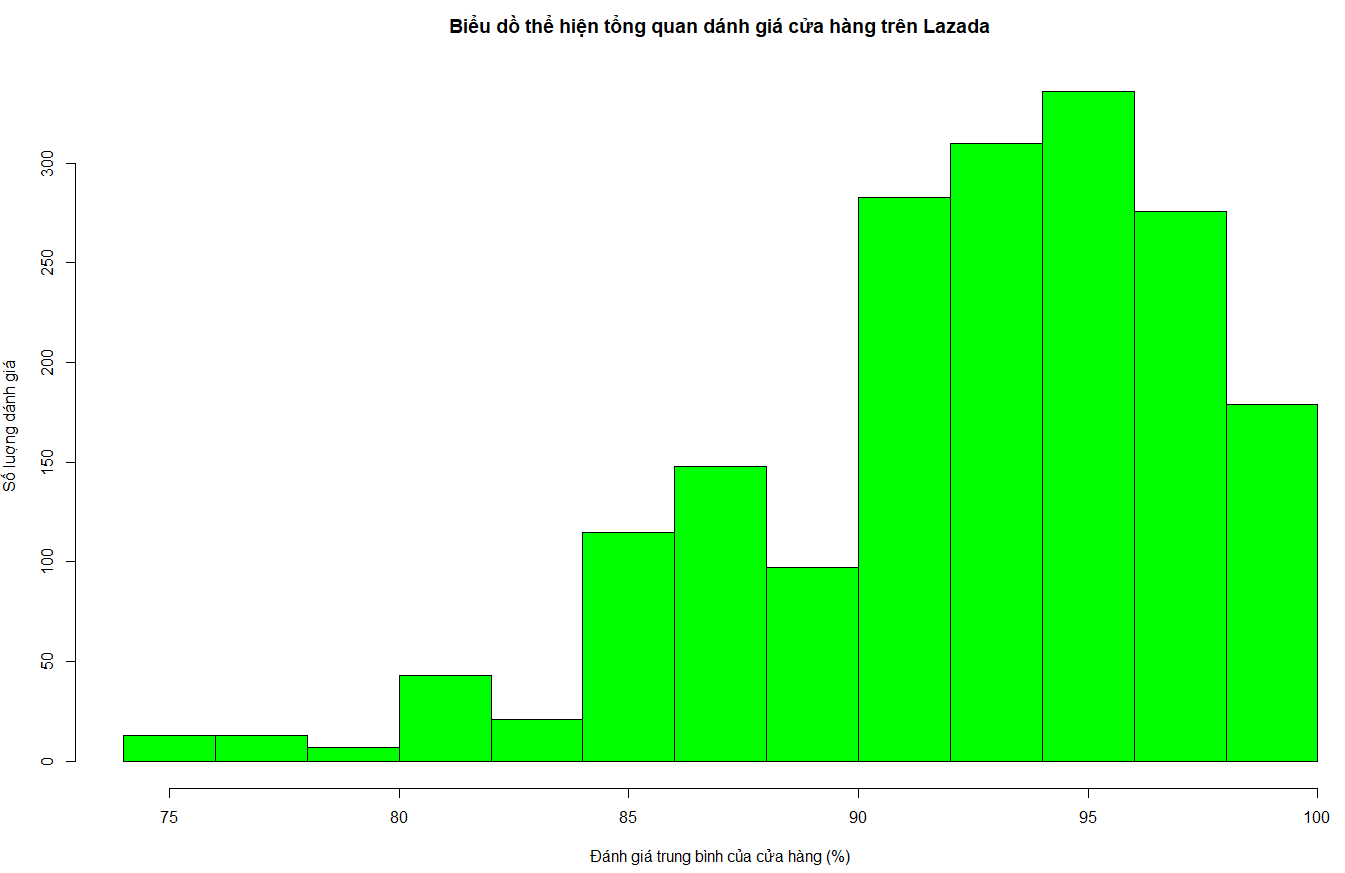
* **p\_rate1=>5star** - Tổng số đánh giá 1=>5 sao cho sản phẩm

* **s\_name** - Tên nhà bán hàng

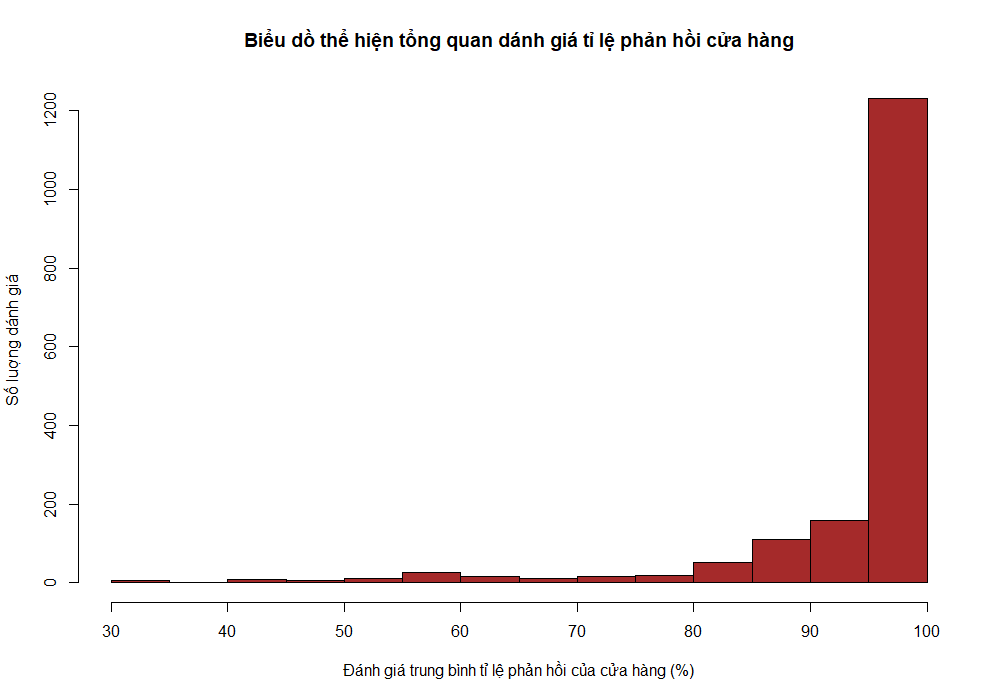
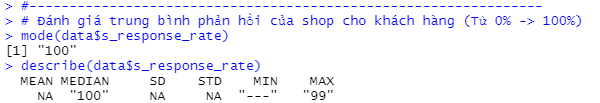


* **s\_rating** - Tỉ lệ đánh giá cho nhà bán hàng

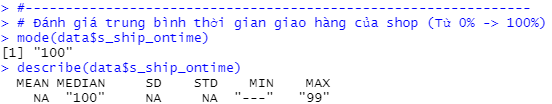
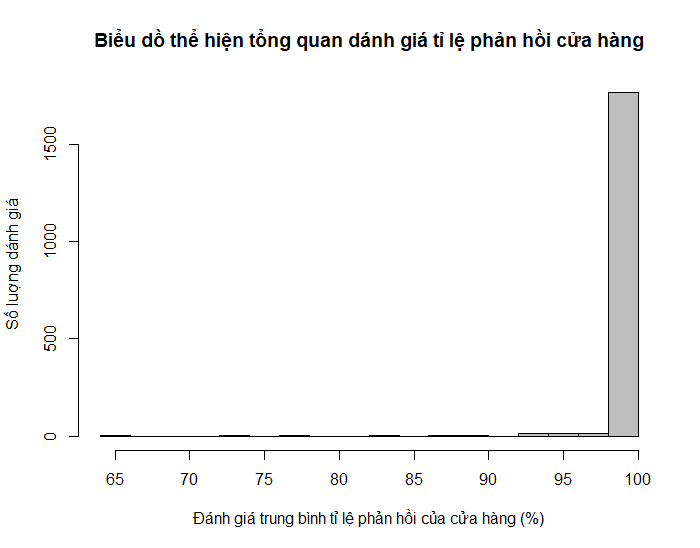
Một số sản phẩm kém chất lượng không thể đánh giá cửa hàng đấy bán toàn bộ mặt hàng đều không chất lượng. Biểu đồ này cho thấy đa phần các cửa hàng trên Lazada vẫn được khách hàng tin tưởng.

* **s\_response\_rate** - Tỉ lệ nhà bán hàng phản hồi



Một số cửa hàng mới mở nên chưa đủ thông tin để đánh giá tỉ lệ và tốc độ phản hồi của cửa hàng cho khách hàng nhưng đa số cửa hàng đều phản hồi rất nhanh

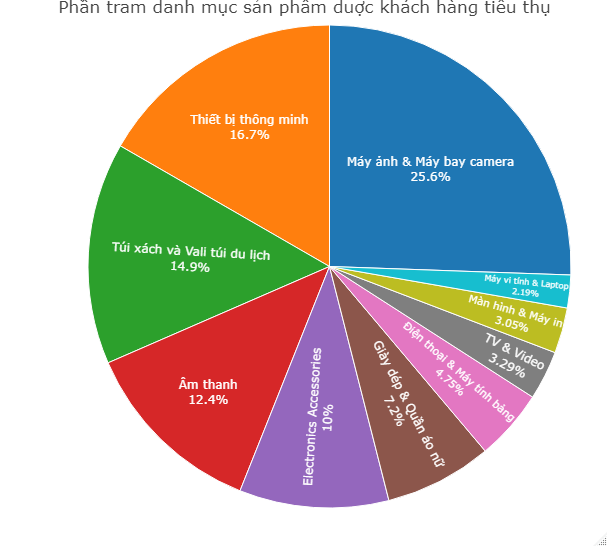
* **s\_ship\_ontime** - Tỉ lệ nhà bán hàng giao hàng đúng giờ

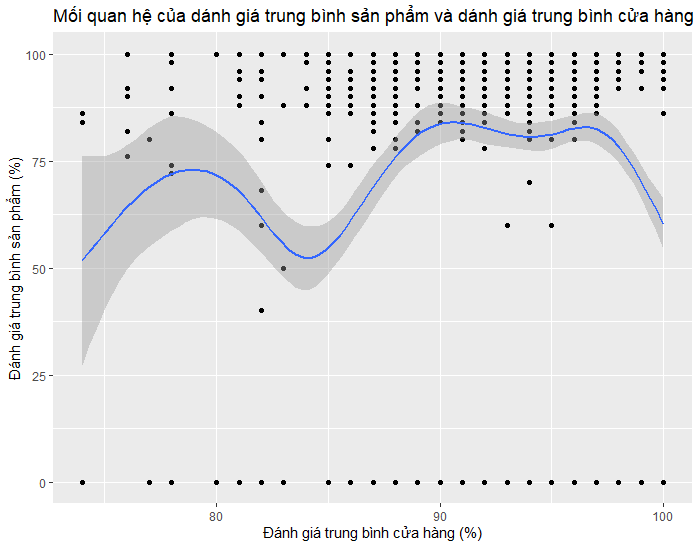
Thời gian giao hàng có tác động từ phía bên vận chuyển trung gian nhưng qua biểu đồ trên cho thấy đây là kết quả của nhà bán hàng và đơn vị vận chuyển trung gian của Lazada làm việc khá tốt, thời gian giao hàng nhanh.

## II/ Phân tích số liệu

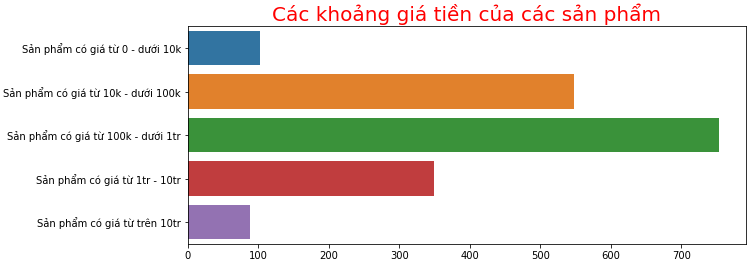
Dựa vào dữ liệu từ các sản phẩm được bán và đánh giá của khách hàng trên Lazada, để tối ưu lượt mua hàng và tang lợi nhuận kinh doanh thì bạn có thể áp dụng vào các biểu đồ sau đây.

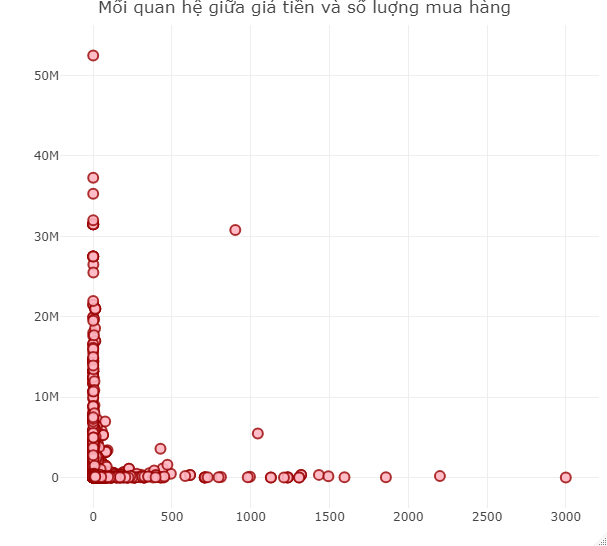


Đầu tiên bạn phải chọn một danh mục được nhiều người quan tâm và được nói đến ở đây là Camera và thiết bị thông minh.

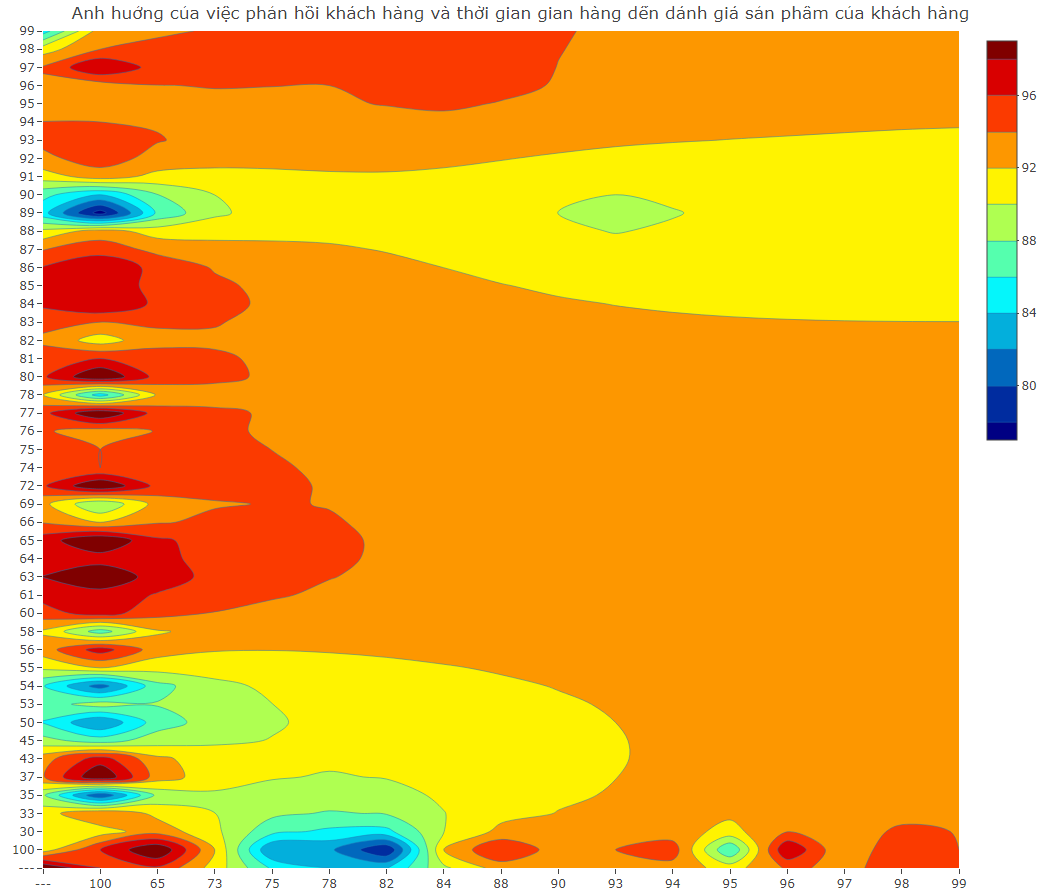


Một sản phẩm chất lượng sẽ mang lại tin tưởng cho khách hàng để khách hàng quay lại mua lần sau tại cửa hàng của bạn

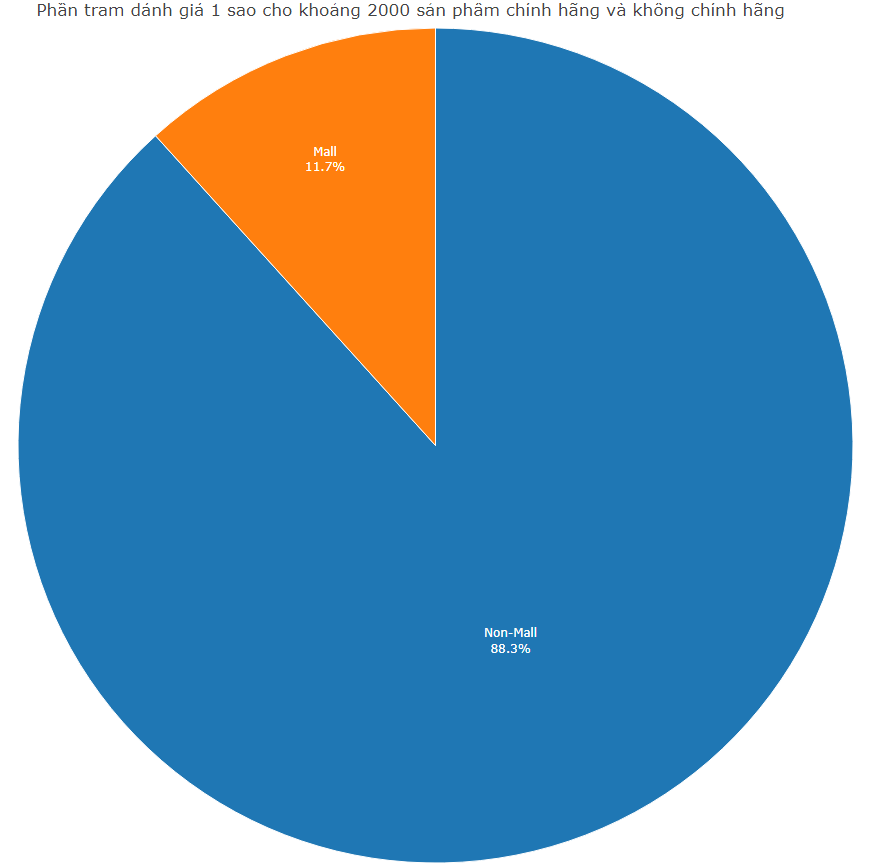
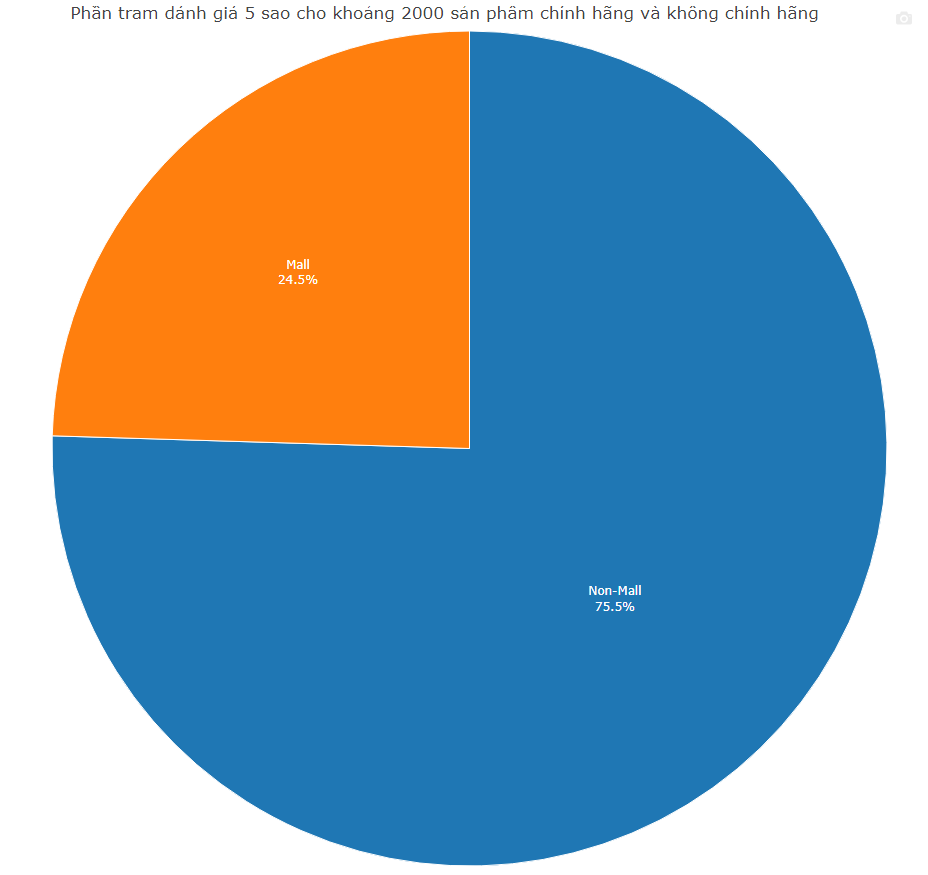




Giá sản phẩm cũng là một yếu tố quan trọng để quyết định bạn có bán được sản phẩm hay không. Đây cũng là lý do ngành hàng Laptop, Máy tính, Tv,… không được mua nhiều trên Lazada. Bạn nên chọn bán những sản phẩm có giá khoảng 50.000-1.000.000vnđ vì nếu sản phẩm quá ít tiền thì người dung cũng không sẵn sàng mua online những mặt hàng như vậy và họ sẽ ra những cửa hàng xung quanh mặc dù nó có thể sẽ đắt hơn một ít so với trên Lazada.



Phản hồi khách hàng và tốc độ chuẩn bị hàng để giao cho đơn vị vận chuyển cũng là một trong những yếu tố quyết định trong tư tưởng của khách hàng là mặt hàng của bạn có chất lượng hay không và khách hàng cũng dựa vào đó để tin tưởng mua hàng ở cửa hàng bạn hay cửa hàng khác trên Lazada hoặc nền tảng khác.

Các mặt hàng chính hãng có lượng đánh giá 5 sao khá cao so với 1 sao. Điều này thể hiện việc được Lazada kiểm chứng cửa hàng của bạn là cửa hàng chính hãng sẽ được khách hàng tin tưởng đảm bảo quyền lợi khi mua hàng online. Tuy chỉ khác sản phẩm bình thường khác trên Lazada là logo Mall  nhưng điều này có thể tang doanh số bán hàng của bạn.

Source code: <https://github.com/tranlybuu/Data_Lazada/blob/master/ThongKeR.R>

Source code: <https://github.com/tranlybuu/Data_Lazada/blob/master/ThongKePython.ipynb>

# Danh mục tham khảo

Darksider. (2020). *Có Nên Học Ngôn Ngữ Lập Trình R Không ?*

Muthukadan, B. (2011). *Selenium with Python¶*. Retrieved from https://selenium-python.readthedocs.io/

Nguyễn, D. (2020, 09 14). Retrieved from https://quantrimang.com/python-la-gi-tai-sao-nen-chon-python-140518

Sanad. (2017, July 25). *analytics vidhya*. Retrieved from https://www.analyticsvidhya.com/blog/2017/07/web-scraping-in-python-using-scrapy/