

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 các khách sạn ở

Vũng Tàu từ Tripadvisor

**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:*** Phân tích dữ liệu các khách sạn ở

Vũng Tàu từ Tripadvisor

**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:*** Phân tích dữ liệu các khách sạn ở

Vũng Tàu từ Tripadvisor

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

**Sinh viên thực hiện: Phan Văn Duy Phúc**

**Lớp: KDL&TTNT K1**

**Số phách**

*(Do hội đồng chấm thi ghi)*

**Thừa Thiên Huế, ngày 12 tháng 01 năm 2021**

Contents

[BÁO CÁO 1](#_Toc92783768)

[**ĐỒ ÁN (TIỂU LUẬN, BÀI TẬP LỚN)** 1](#_Toc92783769)

[**PHÂN TÍCH DỮ LIỆU VỚI R** 1](#_Toc92783770)

[BÁO CÁO ĐỒ ÁN 2](#_Toc92783771)

[**PHÂN TÍCH DỮ LIỆU VỚI R** 2](#_Toc92783772)

[BÁO CÁO ĐỒ ÁN 3](#_Toc92783773)

[**PHÂN TÍCH DỮ LIỆU VỚI R** 3](#_Toc92783774)

[CHƯƠNG 1: TỔNG QUAN VỀ NGÔN NGỮ LẬP TRÌNH R 6](#_Toc92783775)

[1. **Sơ lược về ngôn ngữ lập trình R** 6](#_Toc92783776)

[2. **Đặc điểm ngôn ngữ lập trình R** 6](#_Toc92783777)

[**2.1.** **Lý do R mạnh mẽ cho việc phân tích dữ liệu** 6](#_Toc92783778)

[**2.2.** **Lý do R được dùng nhiều trong kinh doanh** 6](#_Toc92783779)

[**2.3.** **Khía cạnh thống kê** 6](#_Toc92783780)

[**2.4.** **Khía cạnh lập trình** 6](#_Toc92783781)

[**2.5.** **Ưu, nhược điểm** 6](#_Toc92783782)

[CHƯƠNG 2: MÔ TẢ DỮ LIỆU SỬ DỤNG CRAWL LẤY VỀ (SỬ DỤNG NNLT PYTHON) 7](#_Toc92783783)

[**1. Mô tả về dữ liệu sử dụng crawl lấy về** 7](#_Toc92783784)

[**2. Danh sách các trường dữ liệu (tên cột), kiểu dữ liệu (số, văn bản…)** 7](#_Toc92783785)

[CHƯƠNG 3: PHÂN TÍCH THỐNG KÊ MÔ TẢ SỐ LIỆU (SỬ DỤNG NNLT R) 8](#_Toc92783786)

[**1.1. Tính các chỉ số thống kê mô tả như mean, mode, median:** 8](#_Toc92783787)

[**1.2. Tạo hàm để tính mode:** 9](#_Toc92783788)

[**1.3. Tóm tắt các trường dữ liệu:** 9](#_Toc92783789)

[**2. Vẽ các biểu đồ để trực quan hóa dữ liệu và nhận xét về số liệu:** 10](#_Toc92783790)

[**1.1 Biểu đồ thể hiện “Tổng giá trị của dữ liệu định lượng”** 10](#_Toc92783791)

[**1.2 Biểu đồ thể hiện “Số lượng khách sạn với mức giá theo cấp độ”** 11](#_Toc92783792)

[**1.3. Biểu đồ thể hiện “Đánh giá chất lượng của từng khách sạn”** 12](#_Toc92783793)

[**1.4. Biểu đồ thể hiện “Số lượng khách sạn sử dụng ngôn ngữ giống nhau”** 13](#_Toc92783794)

[CHƯƠNG 4: ĐỀ XUẤT Ý TƯỞNG VÀ PHƯƠNG ÁN XỬ LÝ SỐ LIỆU (SỬ DỤNG NNLT PYTHON) 13](#_Toc92783795)

[**1.** **Thực hiện thống kê mô tả với tập dữ liệu dulieucrawl1.csv** 13](#_Toc92783796)

[Chương 5: Kiểm tra đạo văn: 20](#_Toc92783797)

# CHƯƠNG 1: TỔNG QUAN VỀ NGÔN NGỮ LẬP TRÌNH R

1. **Sơ lược về ngôn ngữ lập trình R**

* R là một công cụ rất mạnh cho học máy, thống kê và phân tích dữ liệu. Nó là một ngôn ngữ lập trình. Ngôn ngữ R là một platform-independent do đó chúng ta có thể sử dụng nó cho bất kỳ hệ điều hành nào.
* 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.
* 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 một cộng đồng phát triển mạnh mẽ.

1. **Đặc điểm ngôn ngữ lập trình R**
   1. **Lý do R mạnh mẽ cho việc phân tích dữ liệu**

* Bạn có thể chạy code mà không cần đến bất cứ compiler nào
* Bất một phép tính nào cũng có thể thực hiện trên vectors
* Đây là Statistical-Language
  1. **Lý do R được dùng nhiều trong kinh doanh**
* Là một open-source, nên R cực kỳ “kinh tế”. Đồng thời R rất phù hợp cho việc mô phỏng dữ liệu qua bảng biểu. Nhờ một cộng đồng phát triển và hơn 15000+ packages trong mọi lĩnh vực nghiên cứu. Hiện tại, khó có một công cụ nào có thể theo kịp R.
* Trong việc nghiên cứu dữ liệu, việc khan hiếm nhân tài là một vấn đề rất lớn. Các công ty có thể dùng ngôn ngữ R để làm nền tảng và training nhân viên sử dụng nó.
  1. **Khía cạnh thống kê**
* Một số điểm nổi bật:
* Nó là một open-source và miễn phí hoàn toàn.
* R có một kho package phong phú và rất mạnh mẽ, cùng một cộng đồng người dùng vô cùng lớn mạnh.
* Một vài tính năng thống kê:
* Basic Statistics – Mean, variance, median.
* Static graphics – Basic plots, graphic maps.
* Probability distributions – Beta, Binomial.
  1. **Khía cạnh lập trình**
* Một điều nổi bật:
* Data inputs – data type, importing data, keyboard typing.
* Data Management – data variables, operators.
* Một vài tính năng lập trình:
* **Tính toán phân tán (Distributed-Computing)**
* **R packages** (một tập các R functions, compiled code và dữ liệu mẫu)
  1. **Ưu, nhược điểm**
* Ưu điểm:
* 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 open-source nên bất kỳ ai cũng có thể sử dụng và cải tiến nó.
* Vì là open-source nên R có thế được dùng mọi lúc mọi nơi cho bất cứ việc gì, kể cả bán các sản phẩm từ R theo điều kiện của giấy phép.
* R có thể chạy trên bất kỳ hệ điều hành nào.
* Bất kỳ ai cũng được hỗ trợ để đưa ra ý tưởng phát triển, fix bug, phát triển package mới.
* Nhược điểm:
* Một vài package của R có thể không hoàn hảo và còn lỗi
* Không có ai để “complain” cho việc code không chạy
* R có thể chiếm dụng hết “available memory”.

# CHƯƠNG 2: MÔ TẢ DỮ LIỆU SỬ DỤNG CRAWL LẤY VỀ (SỬ DỤNG NNLT PYTHON)

## **1. Mô tả về dữ liệu sử dụng crawl lấy về**

-Dữ liệu được crawl về để sử dụng là dữ liệu liên quan đến số lượng khách sạn ở Vũng Tàu.

-Dữ liệu gồm có 1241 dòng dữ liệu tương ứng với các dịch vụ và chất lượng của từng khách sạn.

## **2. Danh sách các trường dữ liệu (tên cột), kiểu dữ liệu (số, văn bản…)**

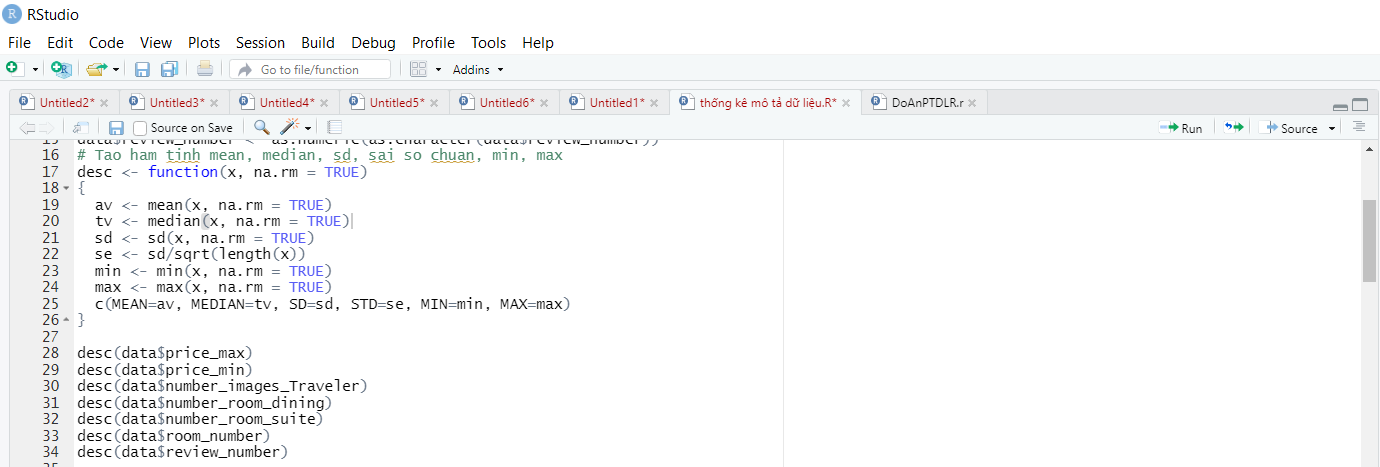
-Dữ liệu gồm 12 trường dữ liệu và có 2 kiểu dữ liệu.

|  |  |
| --- | --- |
| **Tên Cột** | **Kiểu Dữ Liệu** |
| Tên khách sạn (Hotel\_name) | Văn bản |
| Địa chỉ của khách sạn (address) | Văn bản |
| Ngôn ngữ khách sạn đó sử dụng (languages) | Văn bản |
| Số lượng ảnh chụp không gian phòng ăn (number\_room\_dinning) | Số (number) |
| Số lượng ảnh chụp không gian phòng ngủ (number\_room\_suite) | Số (number) |
| Tổng số phòng của khách sạn (room\_number) | Số (number) |
| Số lượng ảnh chụp toàn cảnh khách sạn đó  (number\_images\_Traveler) | Số (number) |
| Xếp hạng của từng khách sạn (rank) | Số (number) |
| Đánh giá chất lượng khách sạn (review) | Văn bản |
| Số lượng người đánh giá (review\_number) | Số (number) |
| Gía cao nhất | Số (number) |
| Gía thấp nhất | Số (number) |

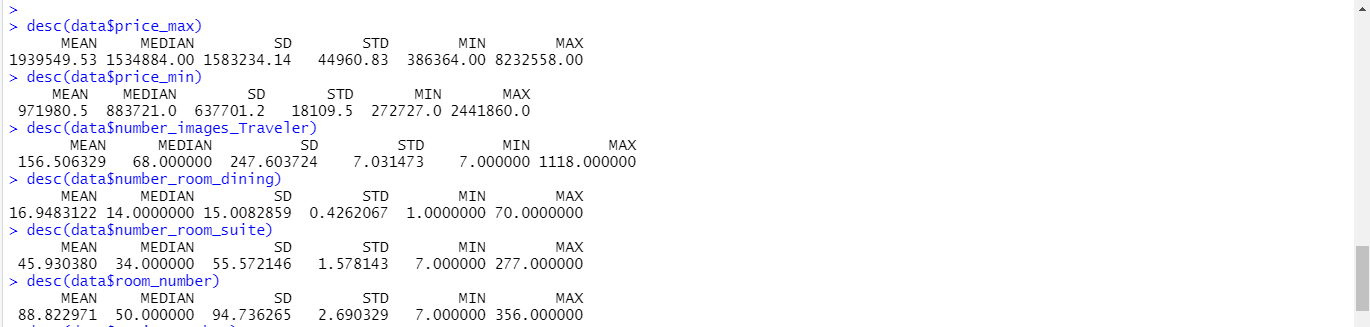
# CHƯƠNG 3: PHÂN TÍCH THỐNG KÊ MÔ TẢ SỐ LIỆU (SỬ DỤNG NNLT R)

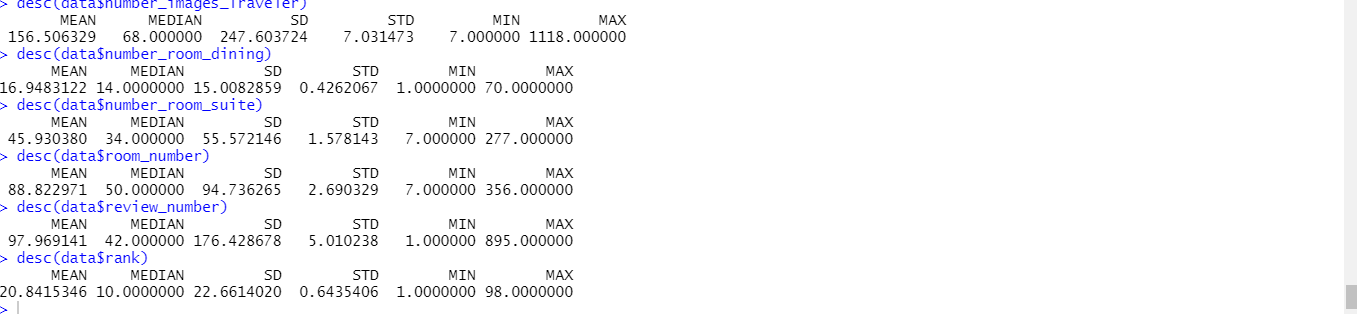
### **1.1. Tính các chỉ số thống kê mô tả như mean, mode, median:**

-Tạo các hàm để tính mean, median:

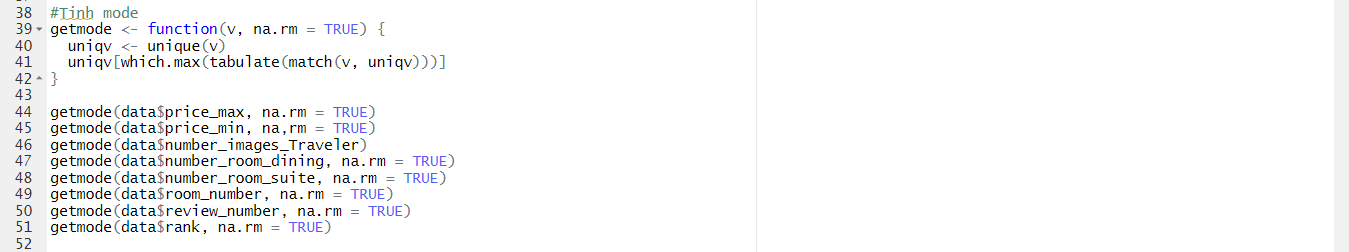


-Kết quả của hàm mean, median:





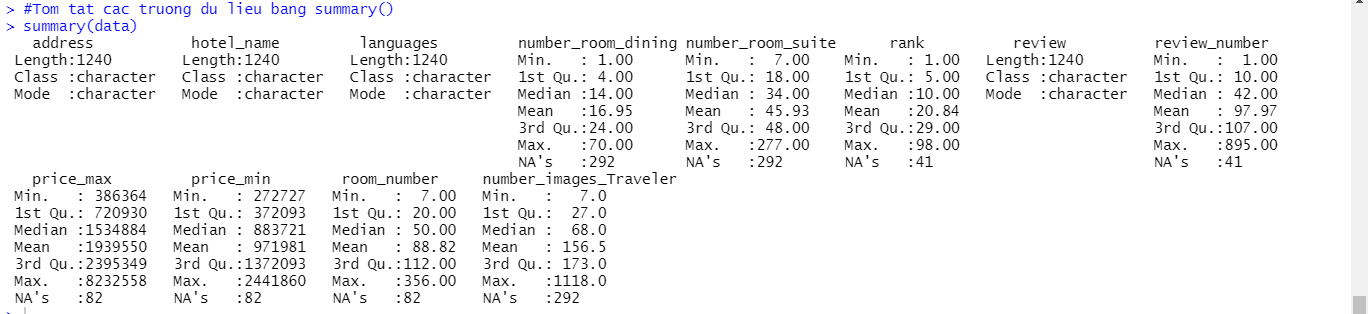
### **1.2. Tạo hàm để tính mode:**



-Kết quả:

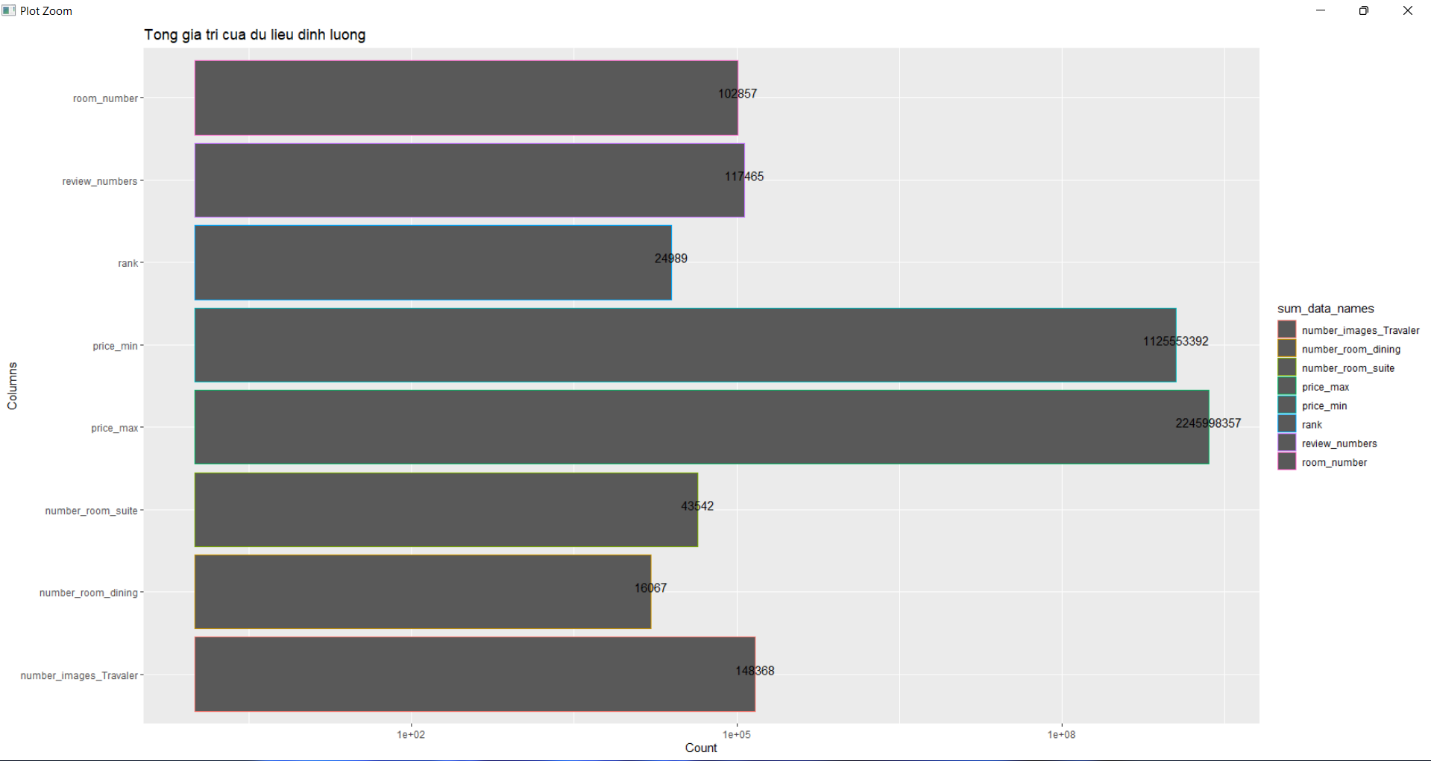


### **1.3. Tóm tắt các trường dữ liệu:**



## **2. Vẽ các biểu đồ để trực quan hóa dữ liệu và nhận xét về số liệu:**

### **1.1 Biểu đồ thể hiện “Tổng giá trị của dữ liệu định lượng”**



**+Nhận xét:**

-Cột có tổng giá trị định lượng lớn nhất là cột price\_max (2245998357) gấp đôi so với cột price\_min (1125553392).

-Cột có tổng giá trị định lượng thấp nhất là số lượng ảnh chụp phòng ăn number\_room\_dinning (16067).

-Số lượng phòng các khách sạn ở Vùng Tàu là: 102857, số lượng người đánh giá: 117485.

-Số lượng ảnh chụp các khách sạn ở Vũng Tàu khá nhiều: 148368.

🡺Giá cả các khách sạn ở Vũng Tàu có sự chênh lệch lớn suy ra giá cả khách sạn ở đây rất đa dạng với nhìu mức giá khác nhau.

🡺Có nhiều thông tin để cho du khách tham khảo nếu muốn đến du lịch và ở lại các khách sạn tại Vũng Tàu.

### **1.2 Biểu đồ thể hiện “Số lượng khách sạn với mức giá theo cấp độ”**



+Nhận xét:

-Số lượng khách sạn có mức giá cao là: 701

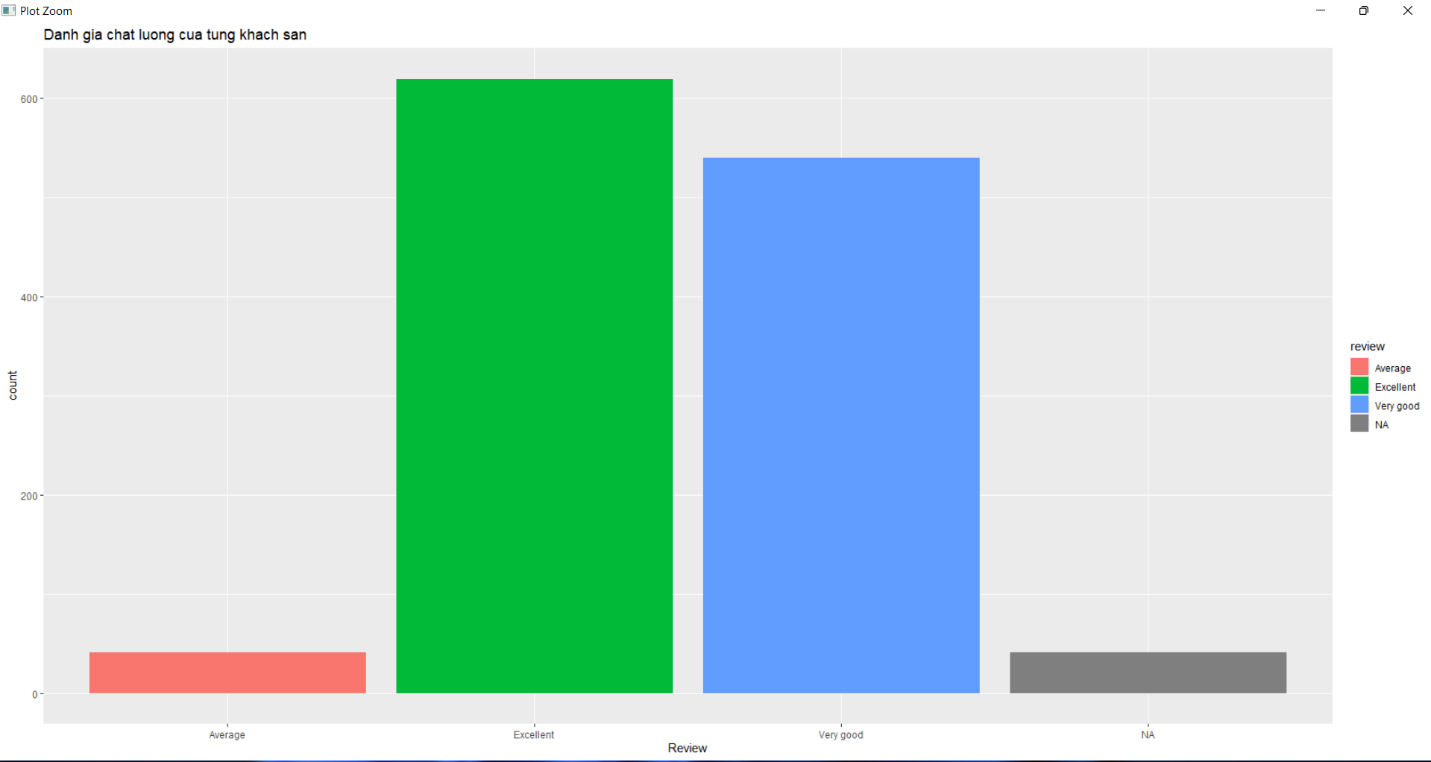
-Số lượng khách sạn có mức giá trung bình là: 415

-Số lượng khách sạn có mức giá thấp là: 42

-Số lượng khách sạn không công bố giá là: 82

🡺Mức giá của các khách sạn ở Vũng Tàu khá cao. Số lượng các khách sạn có mức giá trung bình và thấp chỉ bằng ½ so với các khách sạn có mức giá cao.

### **1.3. Biểu đồ thể hiện “Đánh giá chất lượng của từng khách sạn”**

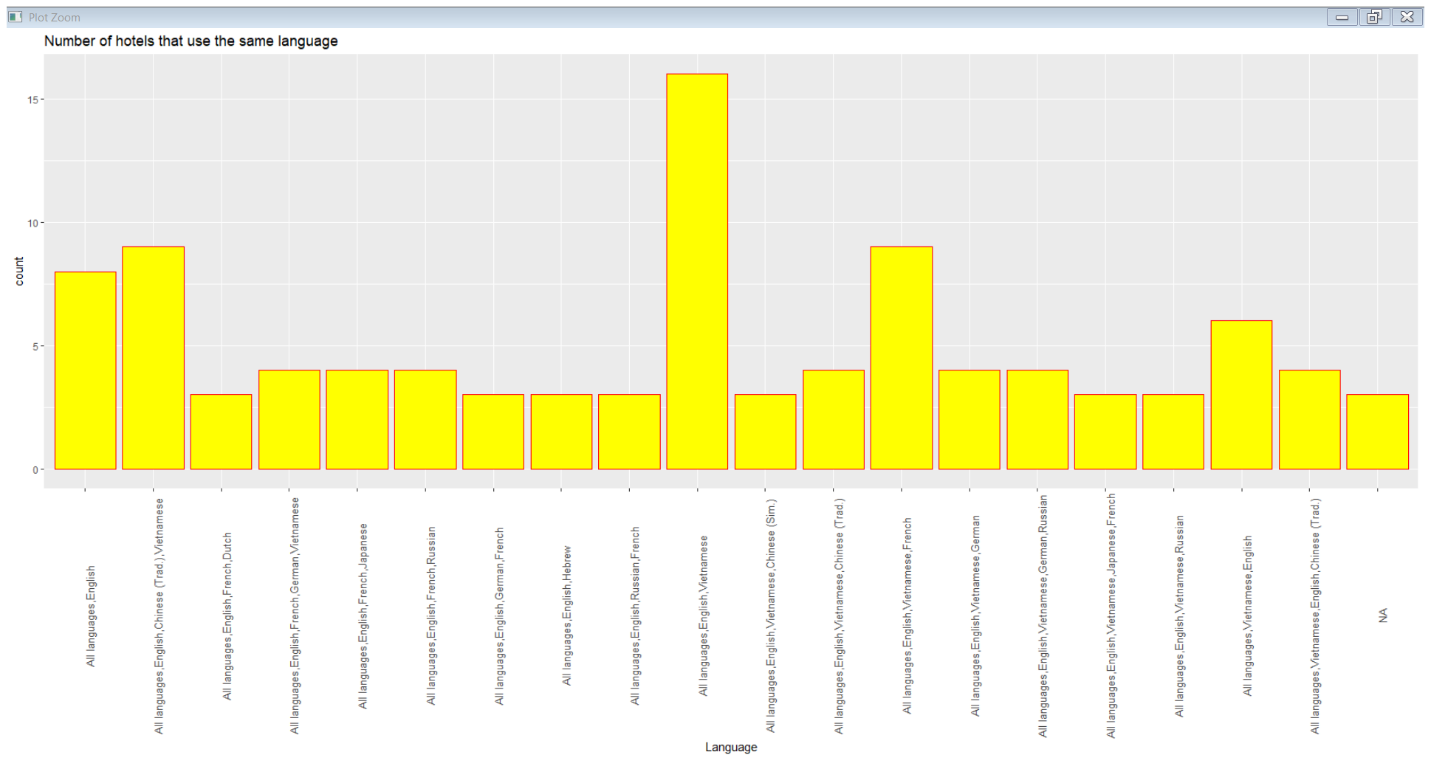


+Nhận xét:

-Số lượng khách sạn được đánh giá có chất lượng xuất sắc và rất tốt nằm ở mức rất cao so với các đánh giá có chất lượng trung bình và không đánh giá.

🡺Chất lượng các khách sạn ở Vũng Tàu được du khách đánh giá rất tốt và có chất lượng cao.

### **1.4. Biểu đồ thể hiện “Số lượng khách sạn sử dụng ngôn ngữ giống nhau”**



+Nhận xét:

-Đa số các khách sạn ở Vũng Tàu đều sử dụng dưới 5 ngôn ngữ, một số ngôn ngữ phổ biến như là: Tiếng Việt, Tiếng Anh, Tiếng Pháp, ….

-Chỉ có một số lượng ít các khách sạn sử dụng từ 5 đến 15 ngôn ngữ khác nhau.

🡺Số lượng ngôn ngữ sử dụng tại các khách sạn còn hạn chế sẽ dẫn đến một số khó khăn khi có nhiều du khách sử dụng các ngôn ngữ khác nhau đến tham quan và ở lại tại các khách sạn của Vũng Tàu.

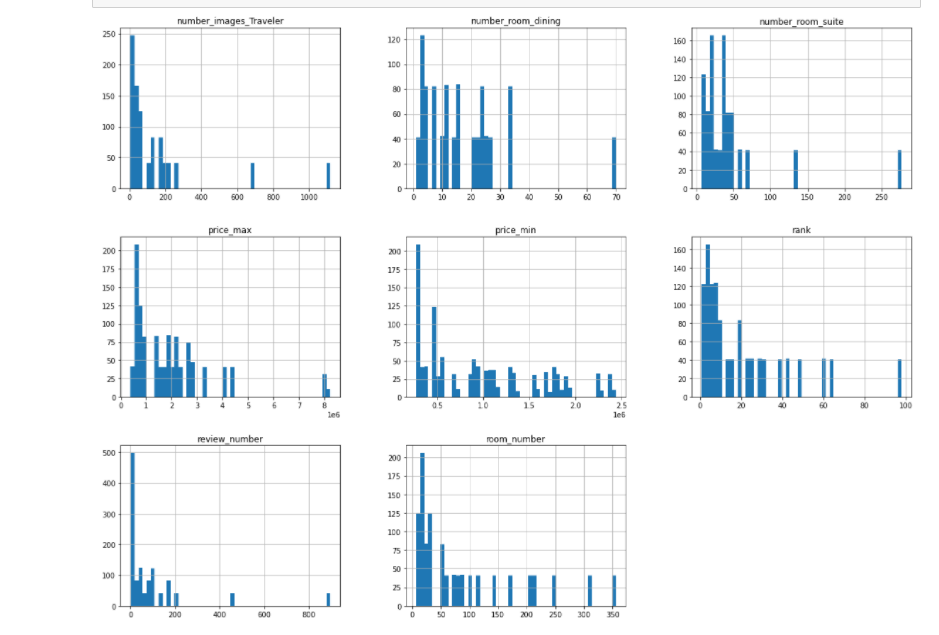
# CHƯƠNG 4: ĐỀ XUẤT Ý TƯỞNG VÀ PHƯƠNG ÁN XỬ LÝ SỐ LIỆU (SỬ DỤNG NNLT PYTHON)

1. **Thực hiện thống kê mô tả với tập dữ liệu dulieucrawl1.csv**

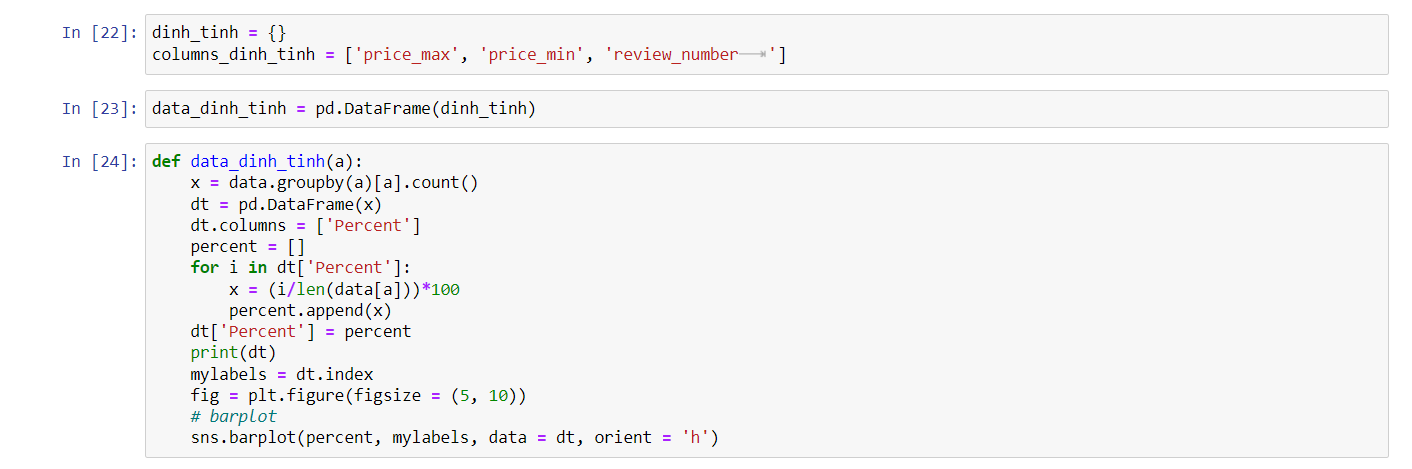
* Vẽ bảng thống kê cơ bản cho 8 cột dữ liệu trong tập dulieucrawl1.csv

****

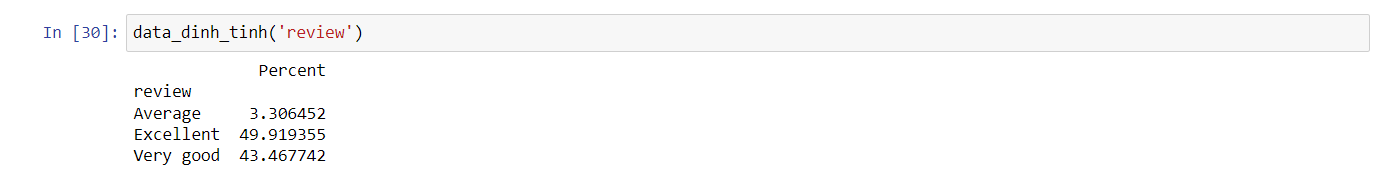
* Vẽ một số biểu đồ cột cho các dữ liệu trên

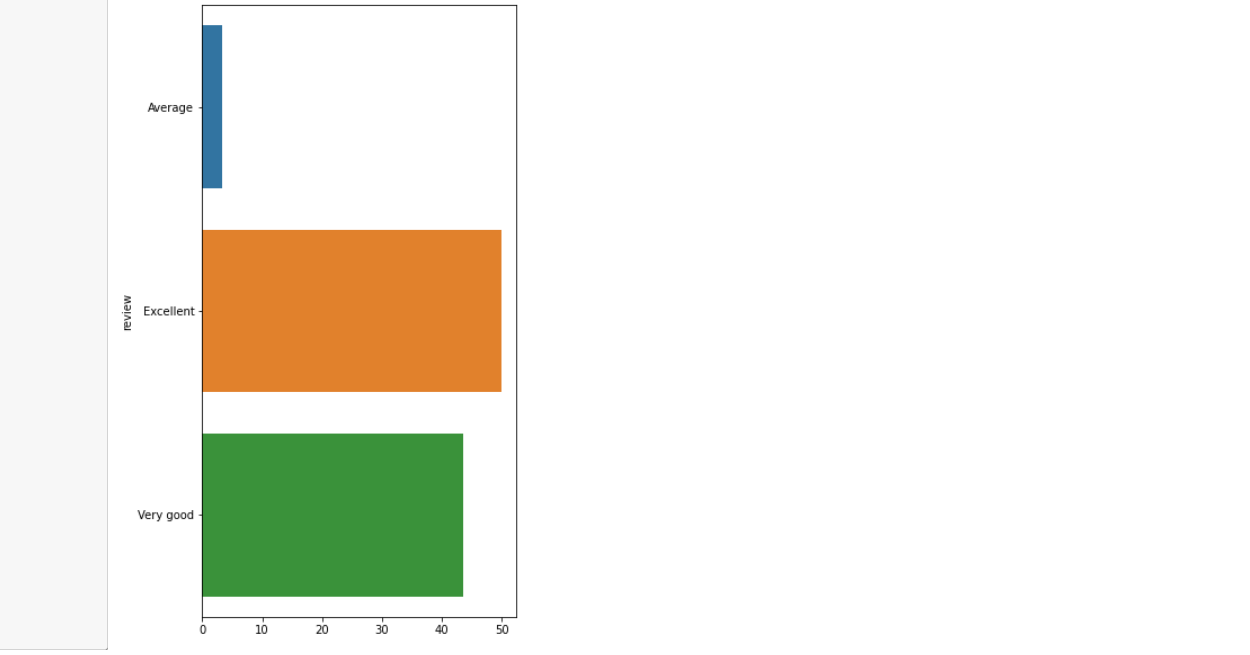
****

Tạo hàm để tính số phần trăm từng giá trị trong cột dữ liệu và vẽ biểu đồ

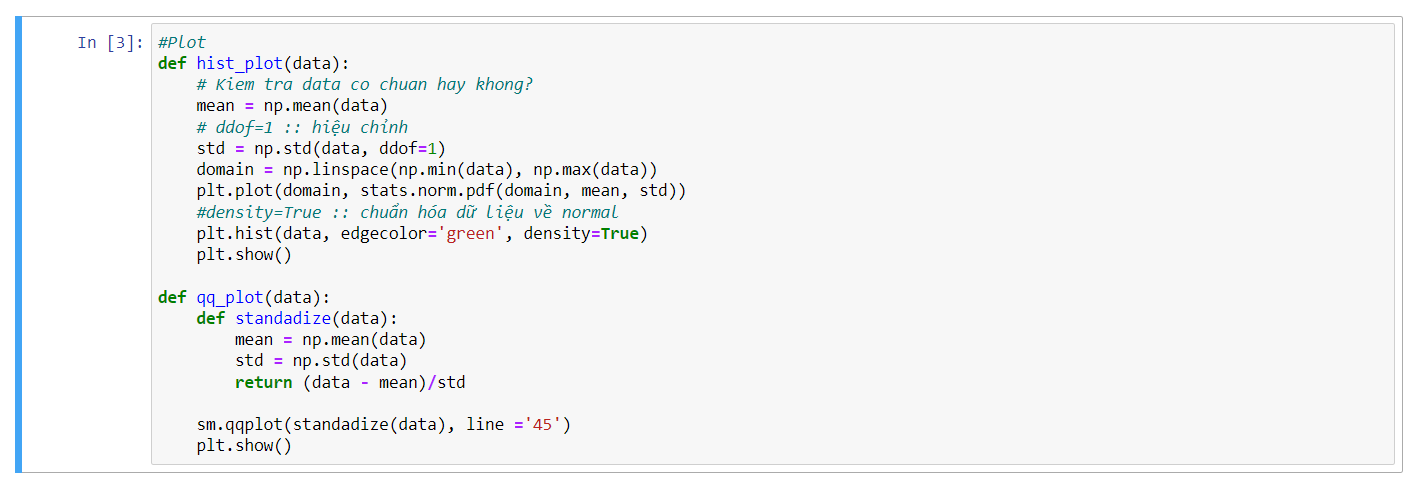
****

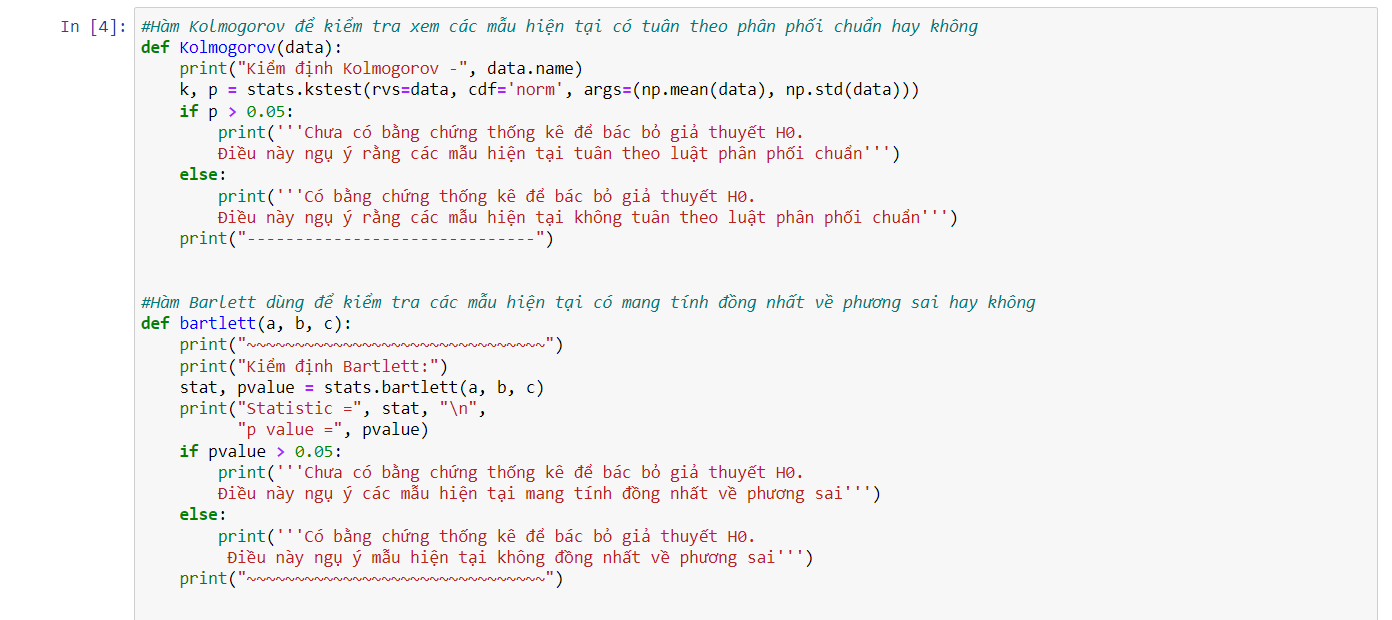
-Vẽ biểu đồ cho cột review

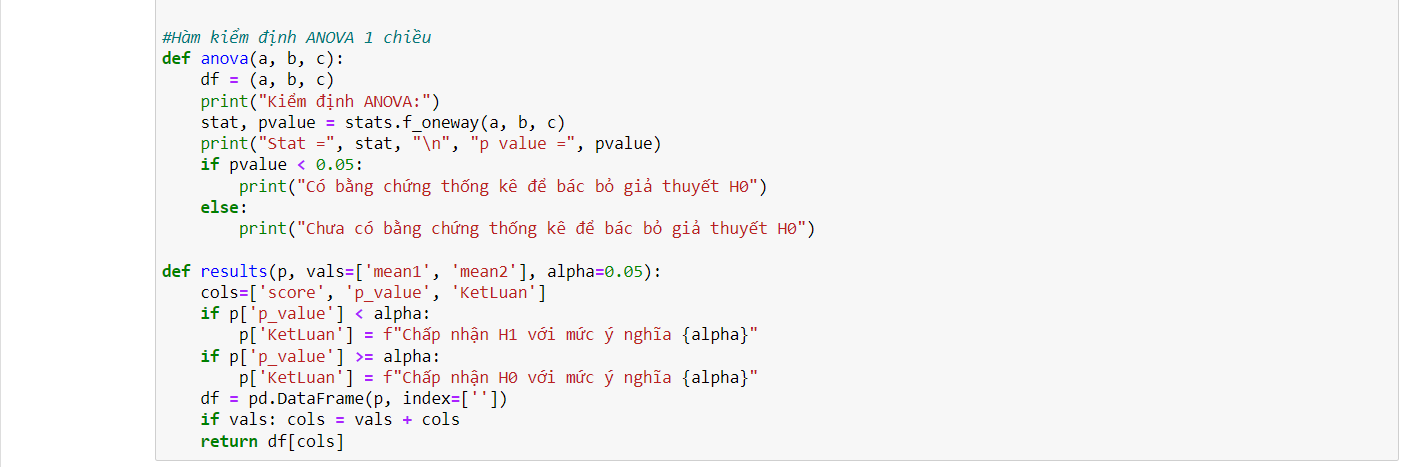




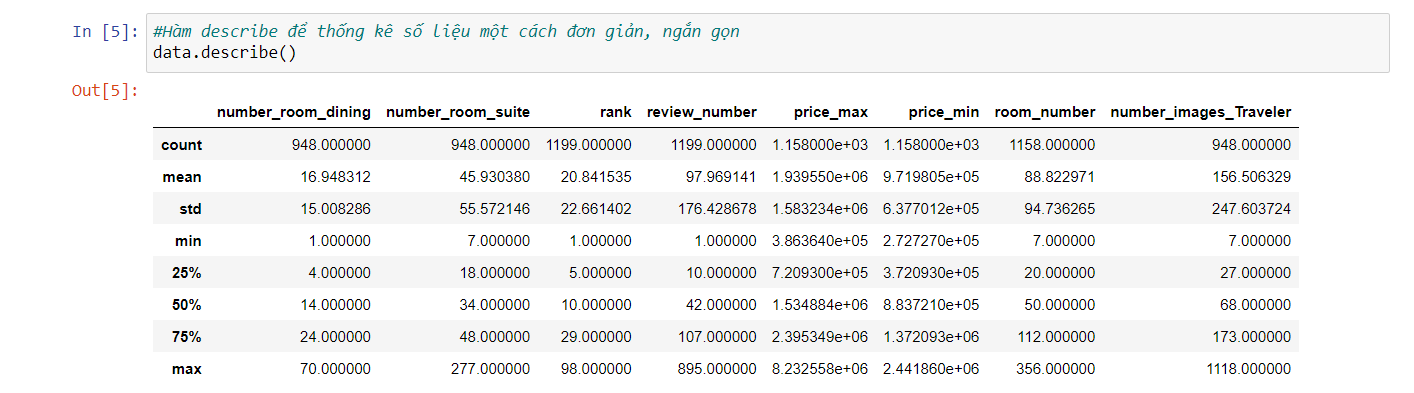
-Tạo các hàm để thực hiện kiểm tra các giả định



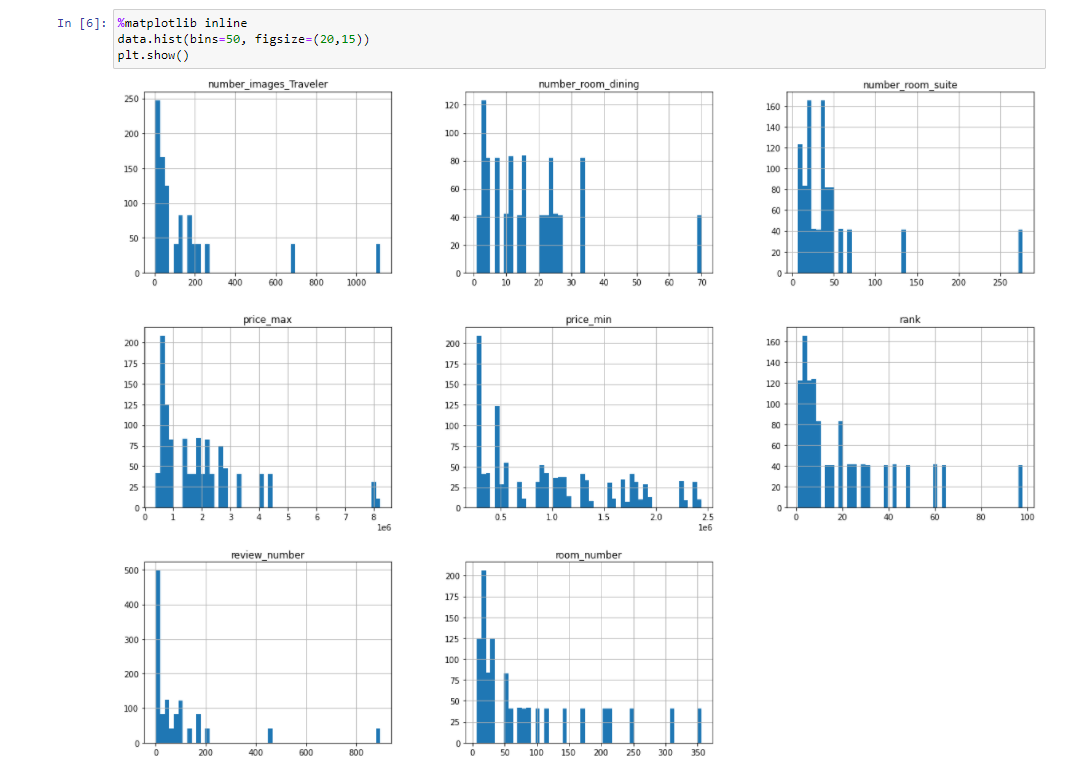




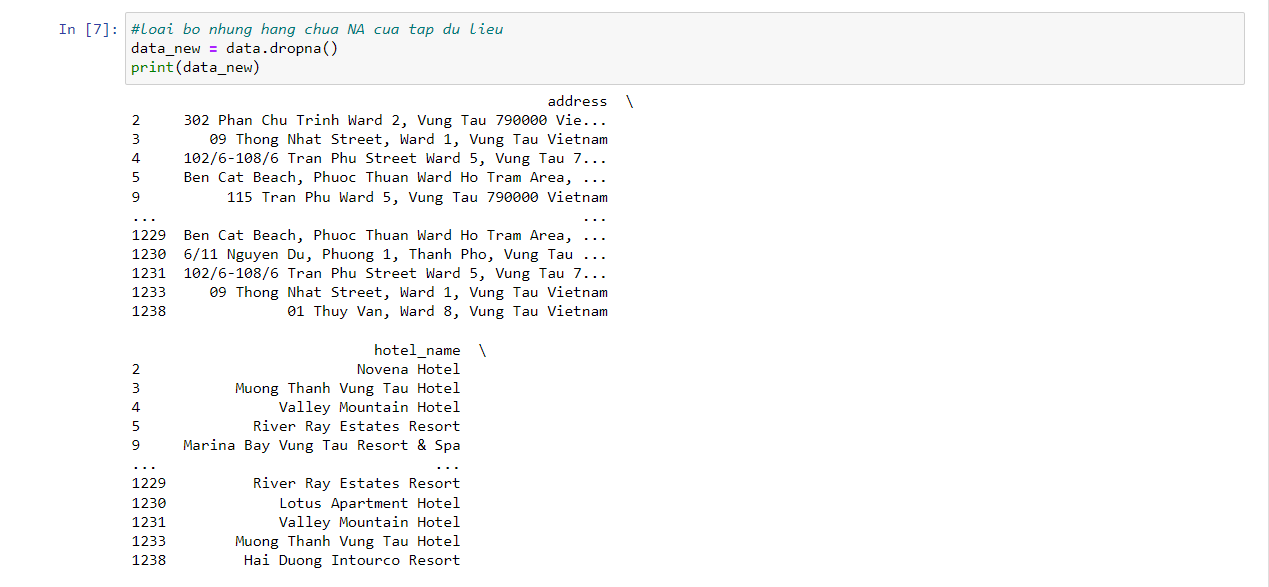
-Tạo hàm thống kê dữ liệu:

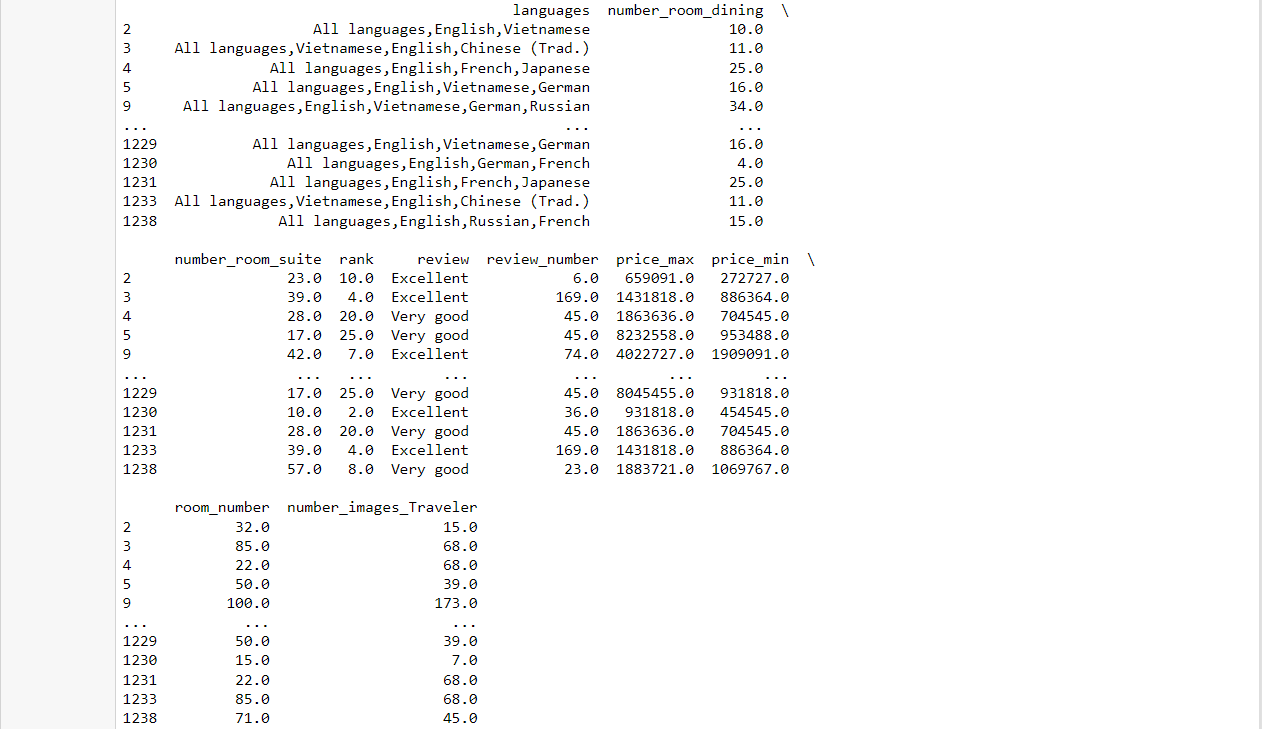


-Vẽ biều đồ cho các dữ liệu đó:

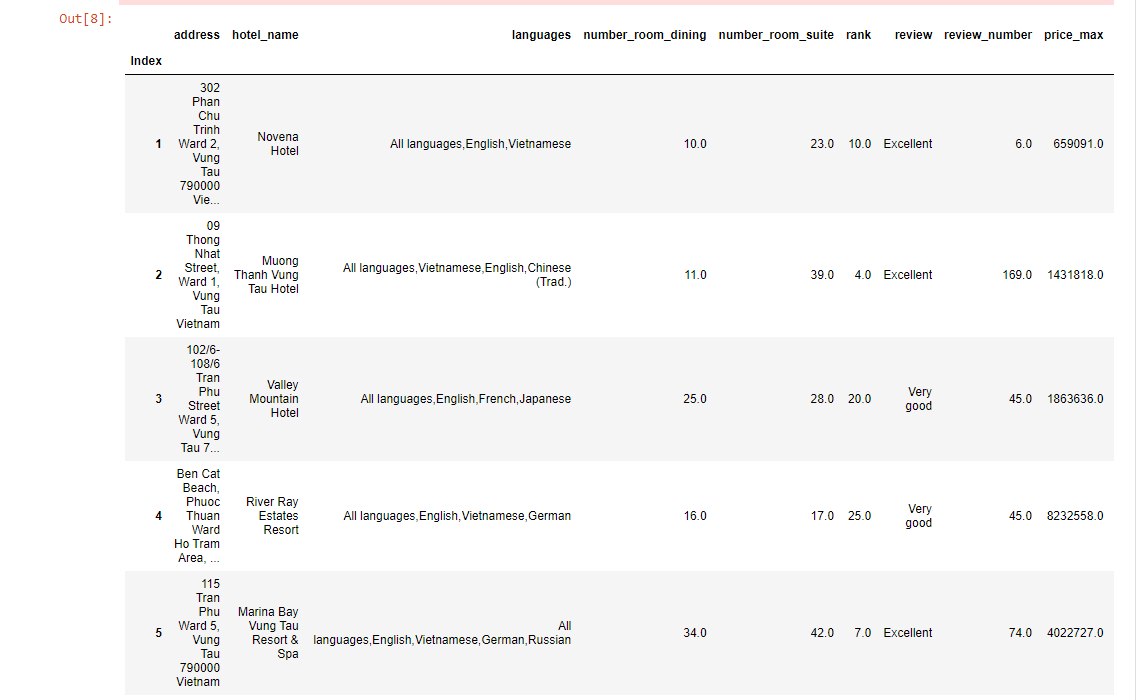


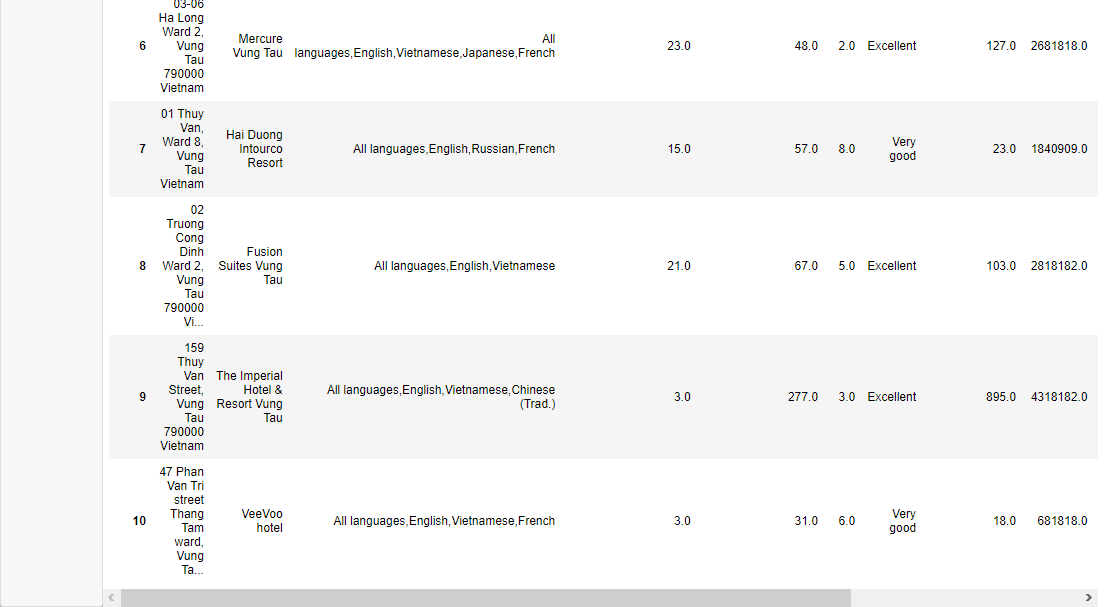
-Hàm để bỏ các dữ liệu NA

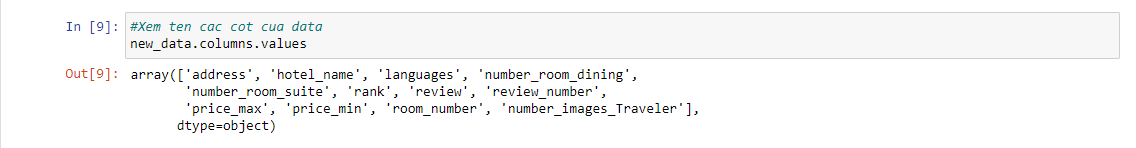












# Chương 5: Kiểm tra đạo văn: