



Data Visualize

# 515K HOTEL REVIEWS IN EUROPE

Giảng viên hướng dẫn: TS. Nguyễn An Tế

Nhóm sinh viên thực hiện:

- Nguyễn Phúc Minh Trâm - 31221024796
- Nguyễn Thành Vinh - 31221025662
- Trầm Thái Tú - 31221022394
- Trần Vọng Triển - 31221021725
- Nguyễn Văn Phi Yến - 31221021785

# table of content

01 Giới thiệu đề tài

02 Tiền xử lý dữ liệu

03 Exploratory Data Analysis

04 Biểu diễn trực quan

# TỔNG QUAN ĐỀ TÀI

## 515K HOTEL REVIEWS DATA IN EUROPE

- Được đăng tải trên [Kaggle](#) bởi tác giả Jiashen Liu
- Đánh giá của khách hàng** tại các **khách sạn cao cấp** trong khu vực châu Âu
- Bao gồm **515,738 dòng**, với **17 thuộc tính**

## MỤC TIÊU

Cung cấp thông tin tổng quát hỗ trợ du khách châu Á chọn lựa các khách sạn phù hợp dựa trên đánh giá và trải nghiệm thực tế của khách hàng.

RangeIndex: 515738 entries, 0 to 515737

Data columns (total 17 columns):

#	Column	Non-Null Count	Dtype
0	Hotel_Address	515738	non-null object
1	Additional_Number_of_Scoring	515738	non-null int64
2	Review_Date	515738	non-null object
3	Average_Score	515738	non-null float64
4	Hotel_Name	515738	non-null object
5	Reviewer_Nationality	515738	non-null object
6	Negative_Review	515738	non-null object
7	Review_Total_Negative_Word_Counts	515738	non-null int64
8	Total_Number_of_Reviews	515738	non-null int64
9	Positive_Review	515738	non-null object
10	Review_Total_Positive_Word_Counts	515738	non-null int64
11	Total_Number_of_Reviews_Reviewer_Has_Given	515738	non-null int64
12	Reviewer_Score	515738	non-null float64
13	Tags	515738	non-null object
14	days_since_review	515738	non-null object
15	lat	512470	non-null float64
16	lng	512470	non-null float64

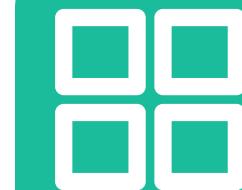
dtypes: float64(4), int64(5), object(8)

# CHỈNH DẠNG DỮ LIỆU

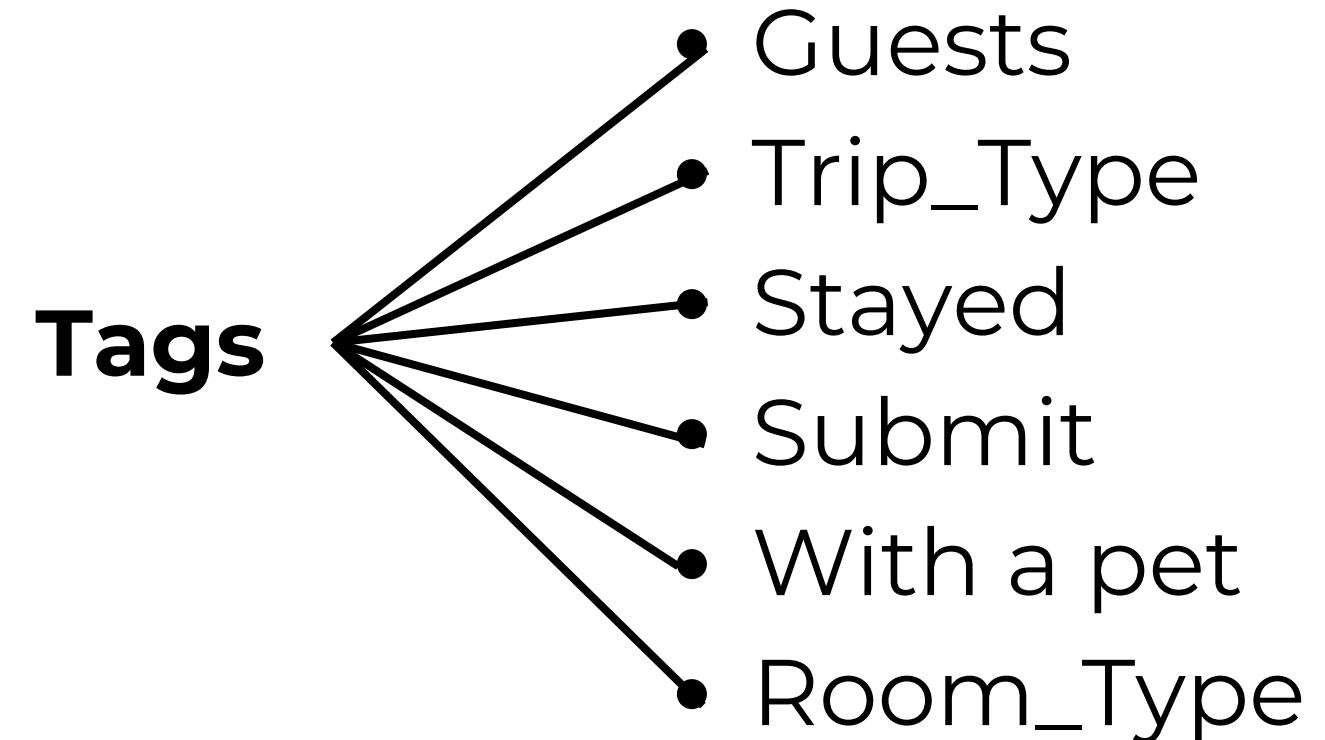


## Thêm thuộc tính

- **Country** (Hotel\_Address)
- **Reviewer\_Regions**  
(Reviewer\_Nationality)
  - **Asia**
  - **Non-Asia**



## Tách thuộc tính



# KIỂM TRA DỮ LIỆU

## Missing Values

	Missing Values	Percentage (%)
Hotel_Address	0	0.0
Review_Date	0	0.0
Hotel_Name	0	0.0
Reviewer_Nationality	0	0.0
Negative_Review	0	0.0
Review_Total_Negative_Word_Counts	0	0.0
Positive_Review	0	0.0
Review_Total_Positive_Word_Counts	0	0.0
Reviewer_Score	0	0.0
Country	0	0.0
Reviewer_Regions	0	0.0
Guests	0	0.0
Trip_Type	0	0.0
Stayed	0	0.0
Submit	0	0.0
With a pet	0	0.0
Room_Type	0	0.0

## Duplicate Values

# Kiểm tra giá trị trùng lặp  
df\_copy.duplicated().sum()

✓ 0.8s

527

# Xóa giá trị trùng lặp  
df\_copy.drop\_duplicates(inplace=True, ignore\_index=True)

✓ 0.9s

TIỀN XỬ LÝ

EDA

BIỂU ĐIỂN

# DỮ LIỆU SAU KHI TIỀN XỬ LÝ

## Quan sát

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 515211 entries, 0 to 515210
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Hotel_Address    515211 non-null   object  
 1   Review_Date      515211 non-null   object  
 2   Hotel_Name       515211 non-null   object  
 3   Reviewer_Nationality 515211 non-null   object  
 4   Negative_Review   515211 non-null   object  
 5   Review_Total_Negative_Word_Counts 515211 non-null   int64  
 6   Positive_Review   515211 non-null   object  
 7   Review_Total_Positive_Word_Counts 515211 non-null   int64  
 8   Reviewer_Score     515211 non-null   float64 
 9   Country          515211 non-null   object  
 10  Reviewer_Regions 515211 non-null   object  
 11  Guests            515211 non-null   object  
 12  Trip_Type         515211 non-null   object  
 13  Stayed            515211 non-null   object  
 14  Submit             515211 non-null   object  
 15  With a pet        515211 non-null   object  
 16  Room_Type         515211 non-null   object  
dtypes: float64(1), int64(2), object(14)
memory usage: 66.8+ MB
```

TIỀN XỬ LÝ

EDA

BIỂU DIỄN

Numerical

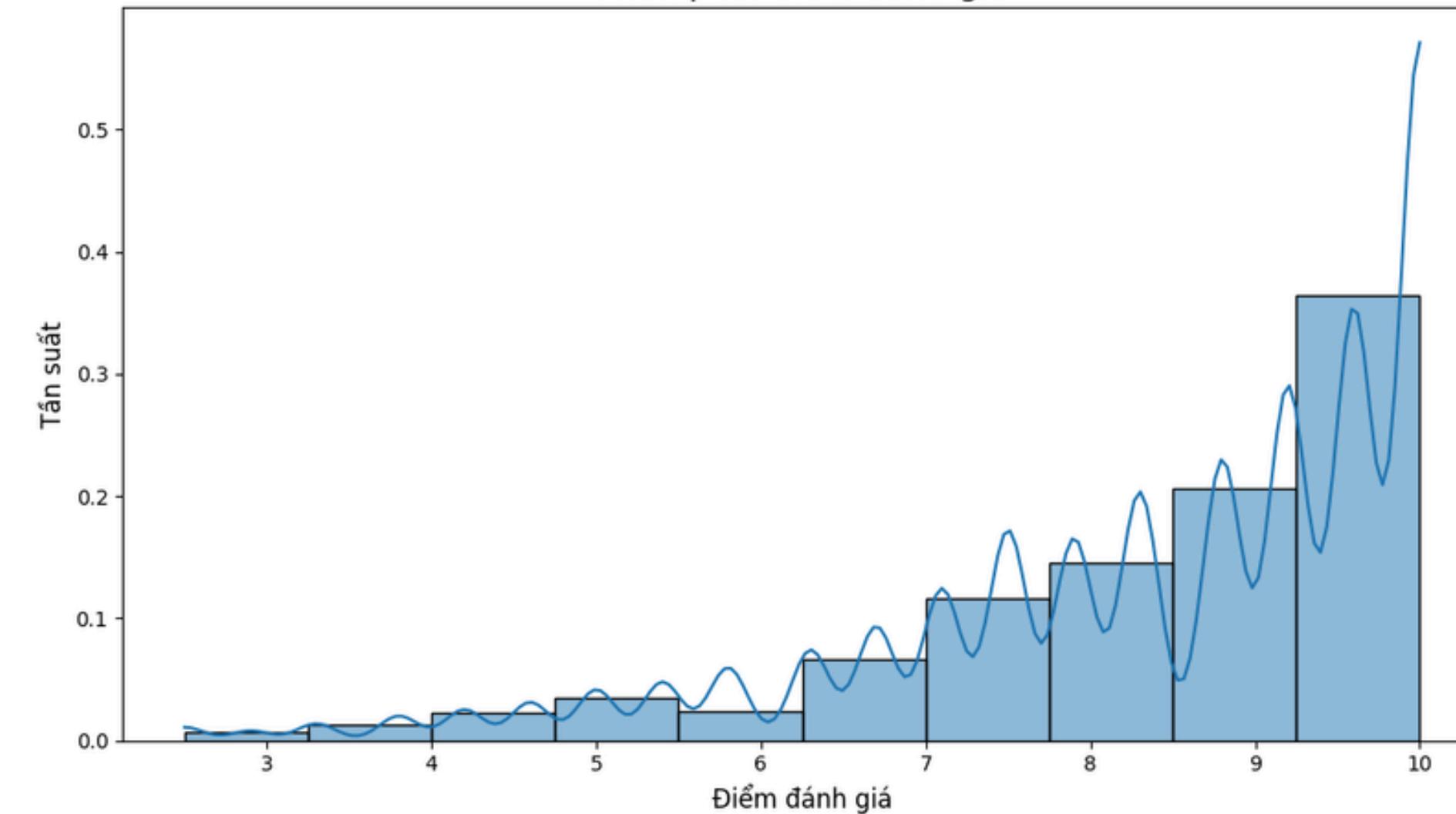
	Rows	Mean	Standard Deviation	Min	Max
Review_Total_Negative_Word_Counts	515211	18.540811	29.694018	0.0	408.0
Review_Total_Positive_Word_Counts	515211	17.778268	21.804561	0.0	395.0
Reviewer_Score	515211	8.395530	1.637468	2.5	10.0

Categorical

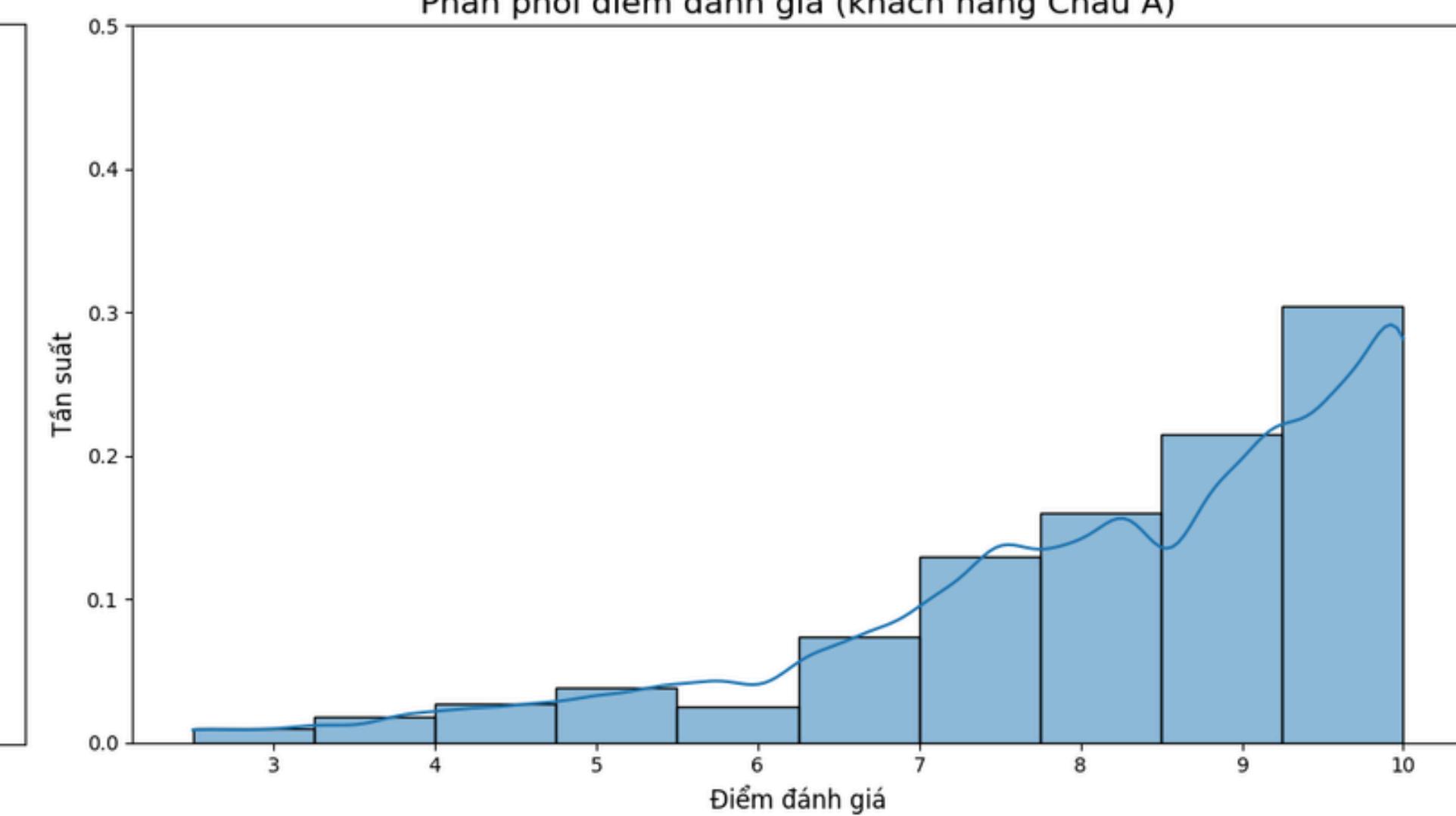
	count	unique	top	freq
Hotel_Address	515211	1493	163 Marsh Wall Docklands Tower Hamlets London ...	4789
Review_Date	515211	731	8/2/2017	2584
Hotel_Name	515211	1492	Britannia International Hotel Canary Wharf	4789
Reviewer_Nationality	515211	227	United Kingdom	245110
Negative_Review	515211	330011	No Negative	127757
Positive_Review	515211	412601	No Positive	35904
Country	515211	6	United Kingdom	262297
Reviewer_Regions	515211	2	Non-Asia	482694
Guests	515211	6	Couple	252005
Trip_Type	515211	3	Leisure trip	417355
Stayed	515211	32	1	193497
Submit	515211	2	from mobile	307355
With a pet	515211	2	Not Mention	513806
Room_Type	515211	2382	Double Room	34027

# PHÂN PHỐI ĐIỂM ĐÁNH GIÁ CỦA KHÁCH HÀNG

Phân phối điểm đánh giá



Phân phối điểm đánh giá (khách hàng Châu Á)



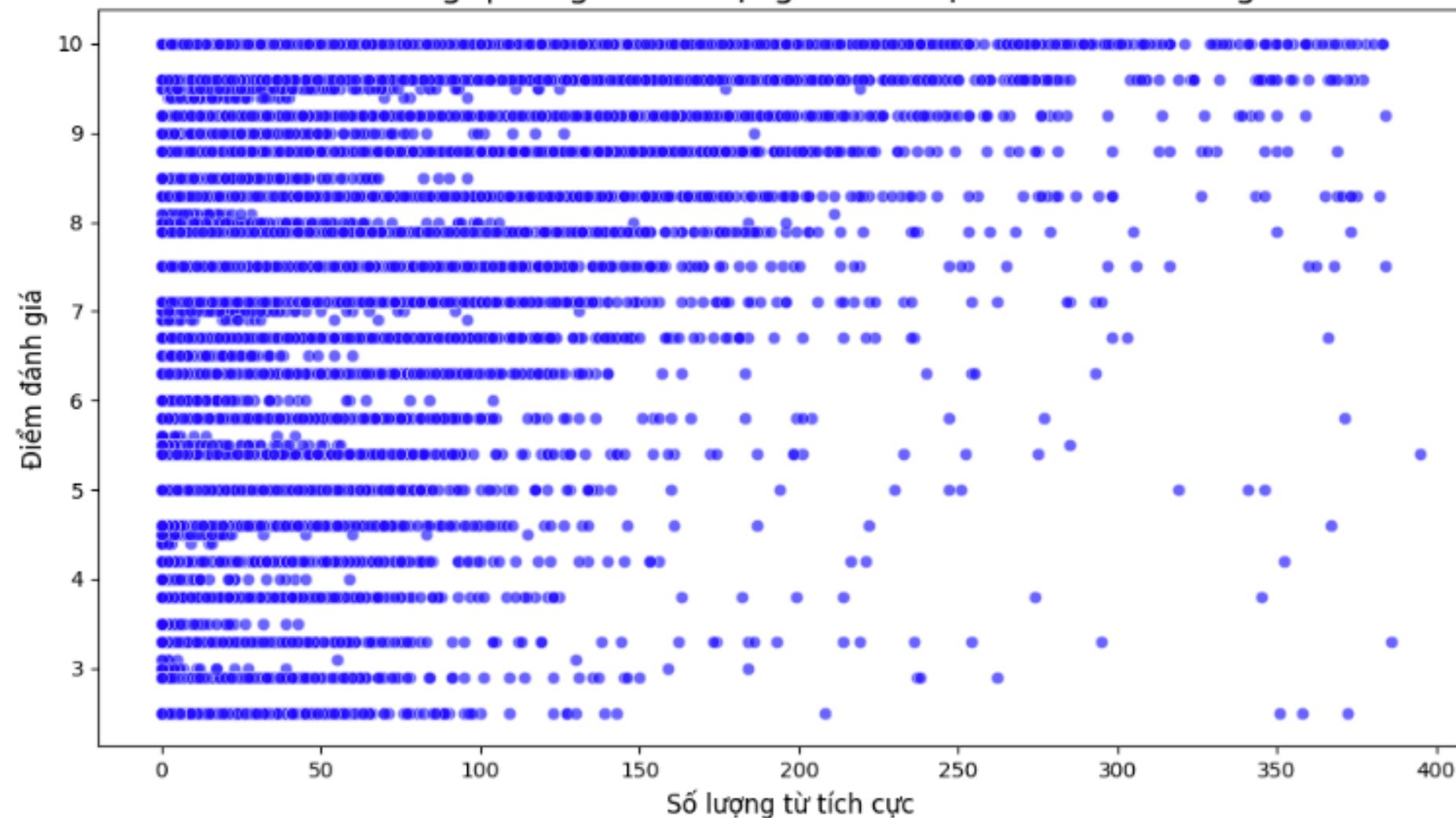
TIỀN XỬ LÝ

EDA

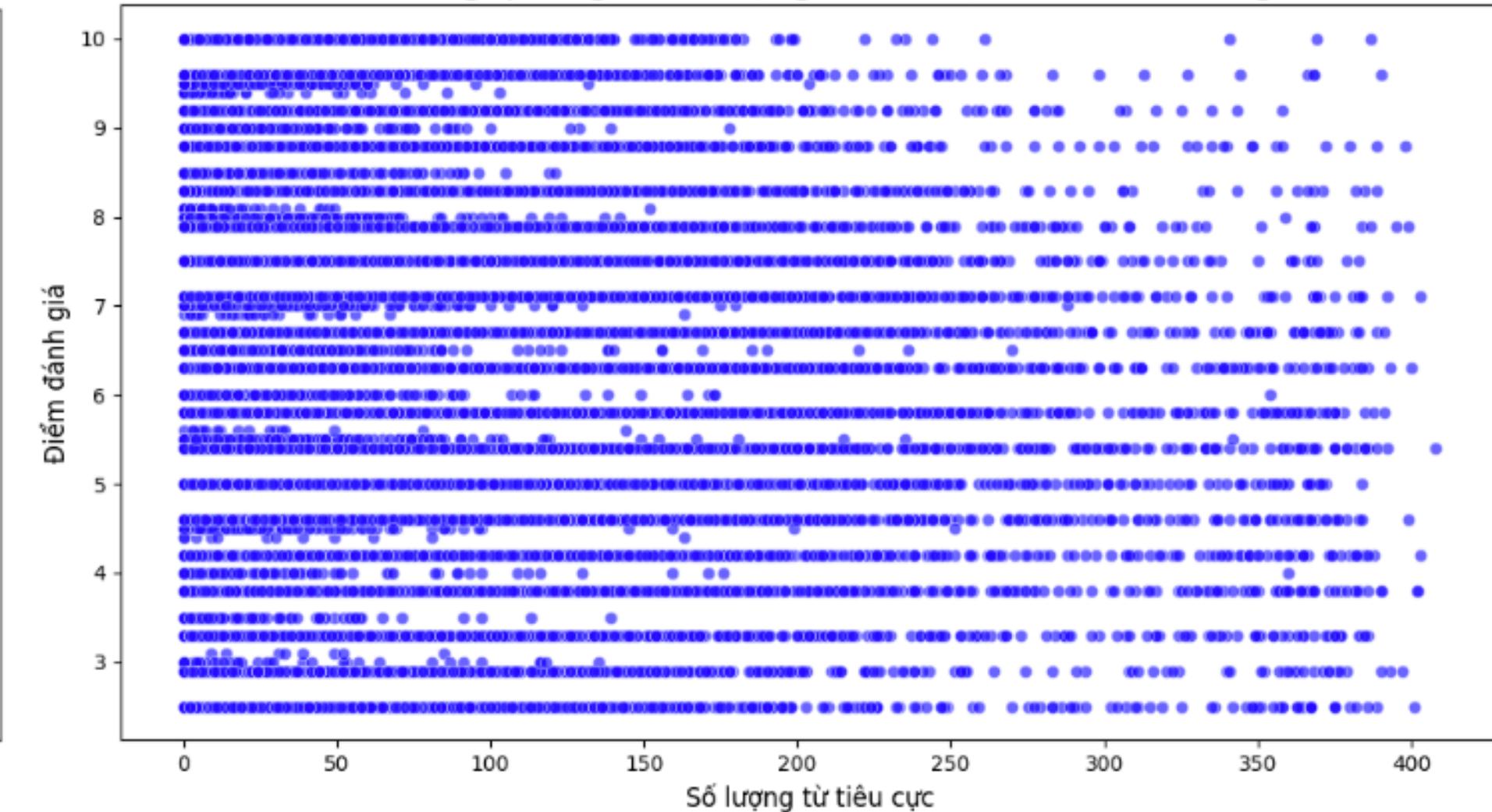
BIỂU DIỄN

# TƯƠNG QUAN GIỮA ĐIỂM SỐ VÀ BÌNH LUẬN

Mối tương quan giữa số lượng từ tích cực và điểm đánh giá



Mối tương quan giữa số lượng từ tiêu cực và điểm đánh giá



TIỀN XỬ LÝ

EDA

BIỂU DIỄN

# KIỂM ĐỊNH TƯƠNG QUAN GIỮA ĐIỂM SỐ VÀ BÌNH LUẬN

## KIỂM ĐỊNH SPEARMAN

- **H<sub>0</sub>:** Không có mối tương quan giữa điểm đánh giá và số lượng từ tích cực/tiêu cực trong bình luận (p-value  $\geq 0.05$ )
- **H<sub>1</sub>:** Tồn tại mối tương quan giữa điểm đánh giá và số lượng từ tích cực/tiêu cực trong bình luận. (p-value  $< 0.05$ )

## HỆ SỐ TƯƠNG QUAN

- >0: tương quan thuận
- <0: tương quan nghịch
- =0: không có tương quan rõ rệt

TIỀN XỬ LÝ

EDA

BIỂU DIỄN

# KẾT QUẢ KIỂM ĐỊNH

**Mối tương quan giữa số lượng từ tiêu cực và điểm đánh giá:**

- p-value = 0.000
- hệ số tương quan: -0.47

**Kết luận:** tồn tại mối tương quan nghịch.

**Mối tương quan giữa số lượng từ tích cực và điểm đánh giá:**

- p-value = 0.000
- hệ số tương quan: 0.31

**Kết luận:** tồn tại mối tương quan thuận.

**Điểm đánh giá thể hiện được phần nào ý kiến của khách hàng đối với dịch vụ của khách sạn**

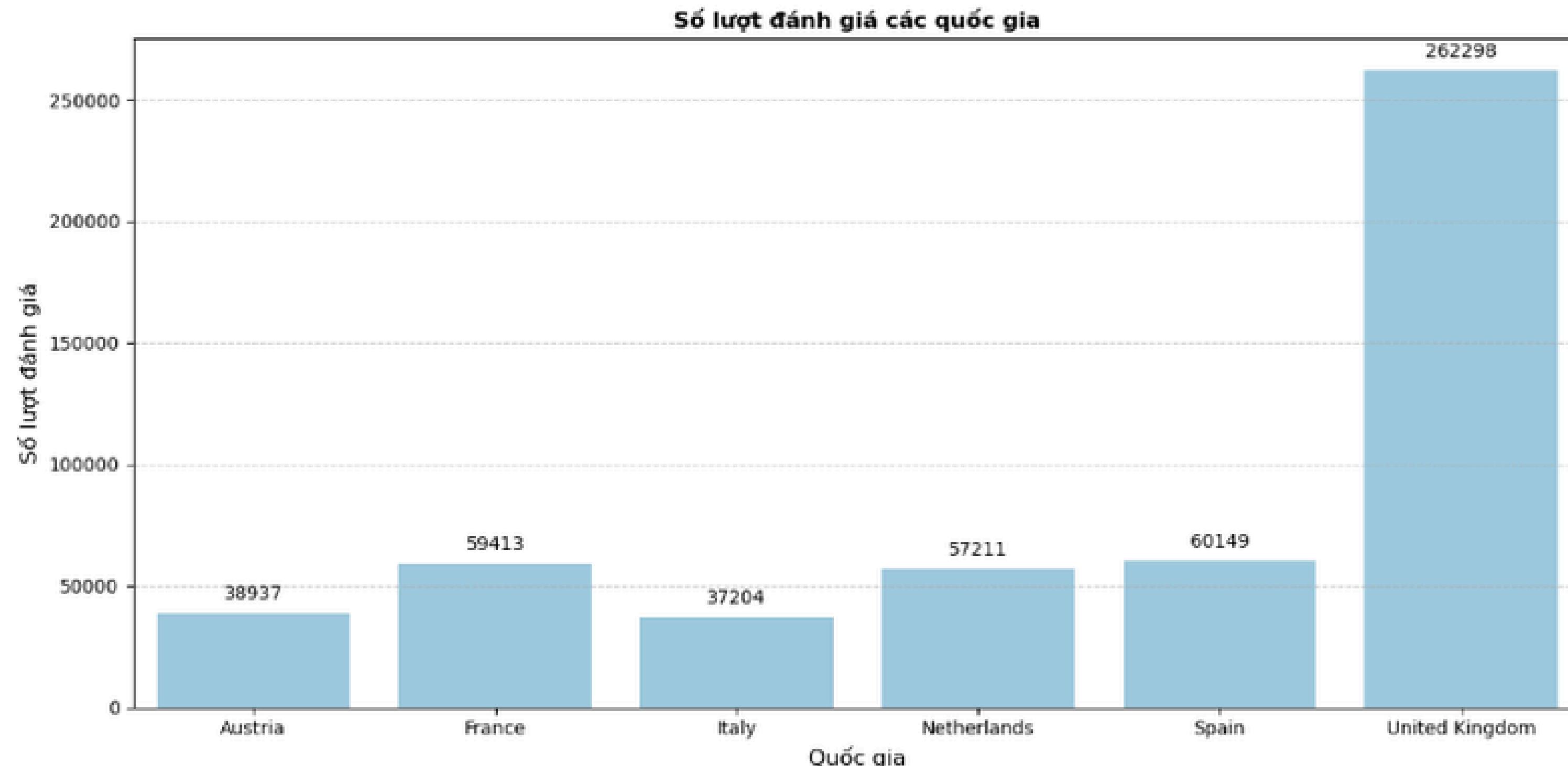
TIỀN XỬ LÝ

EDA

BIỂU DIỄN

# BIỂU ĐIỂN TRỰC QUAN

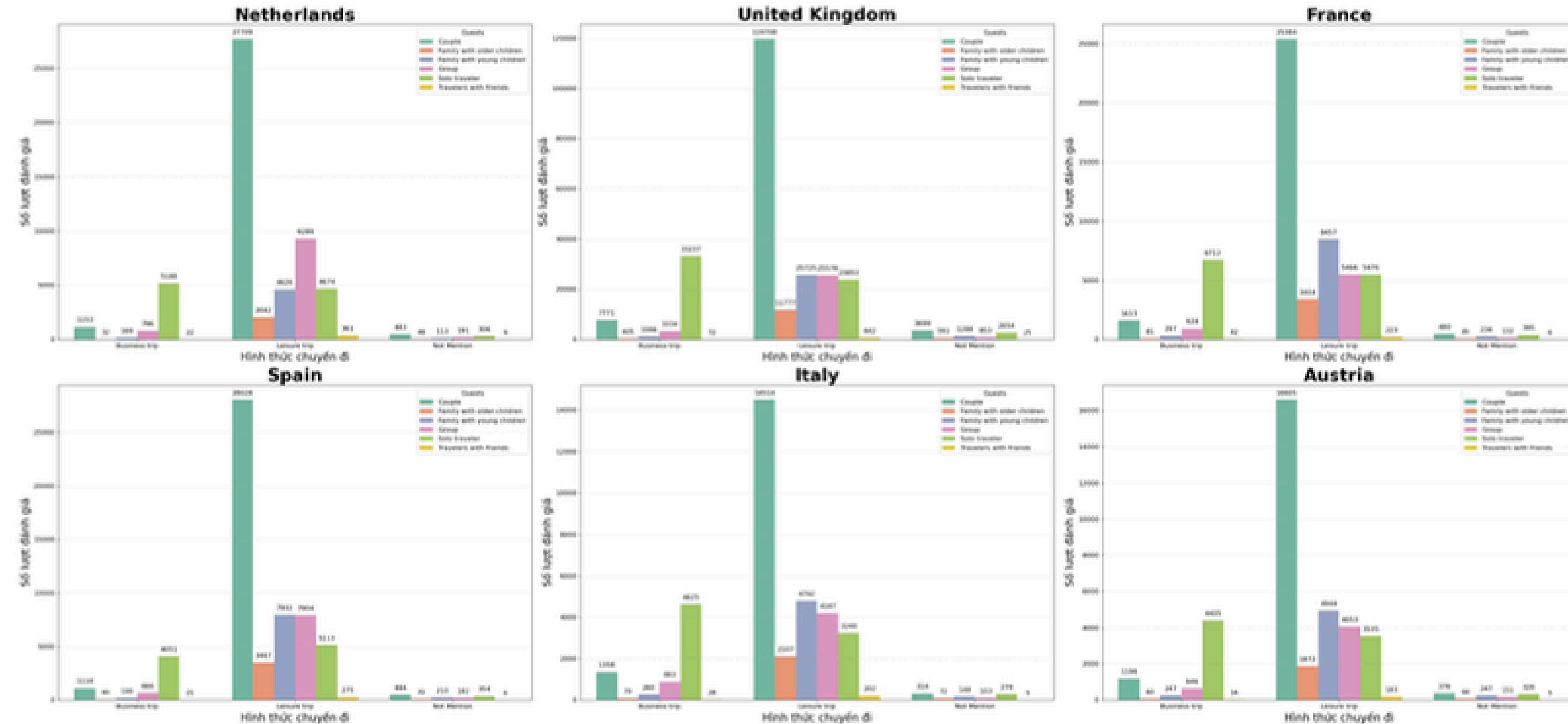
*Số lượng đánh giá của các quốc gia*



# BIỂU DIỄN TRỰC QUAN

Nhóm theo hình thức chuyến đi của từng quốc gia

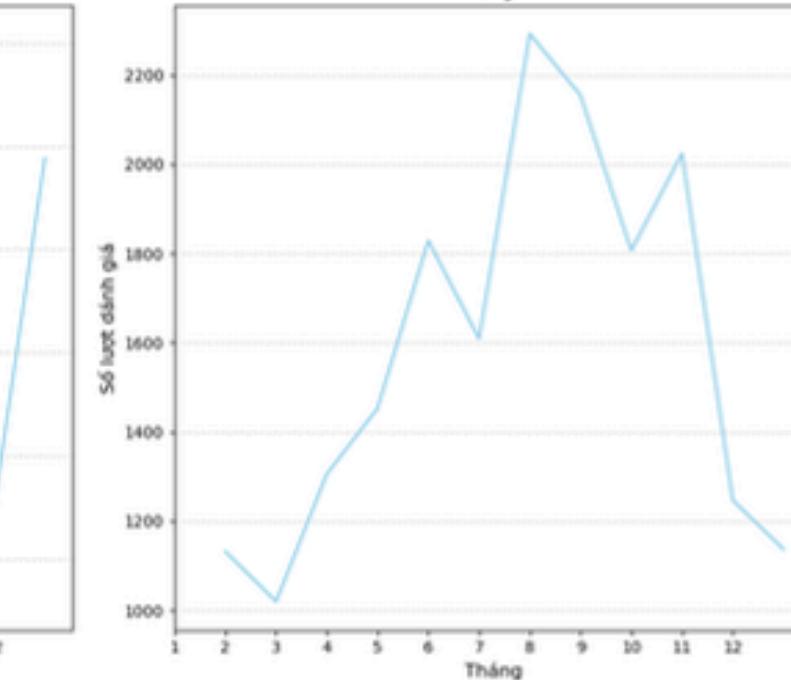
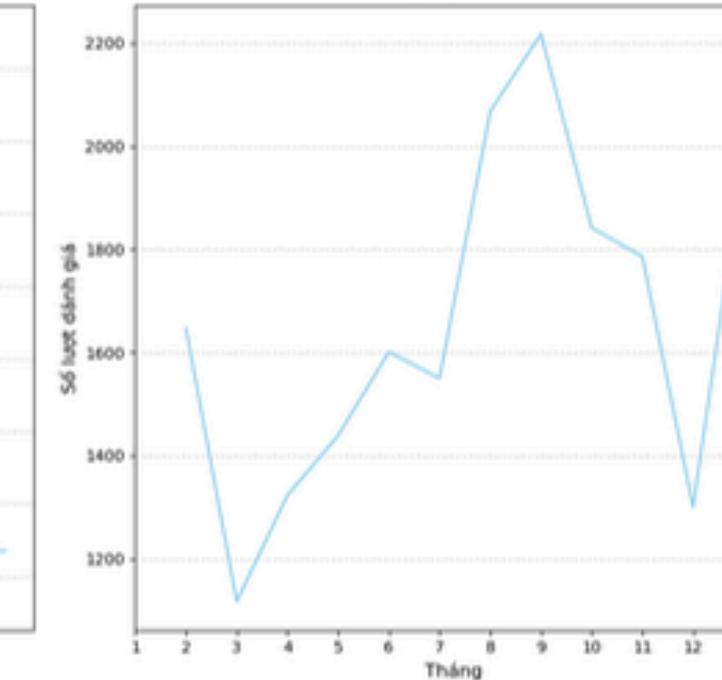
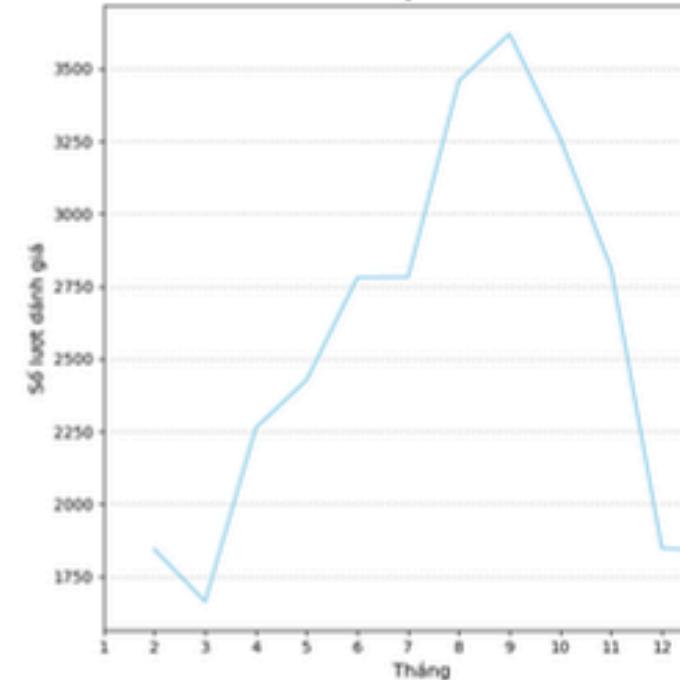
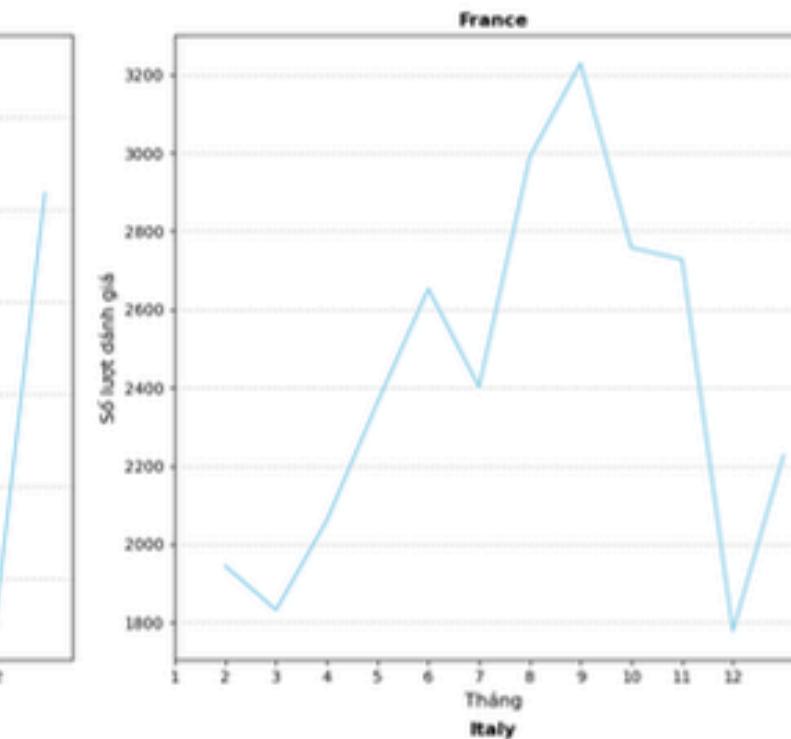
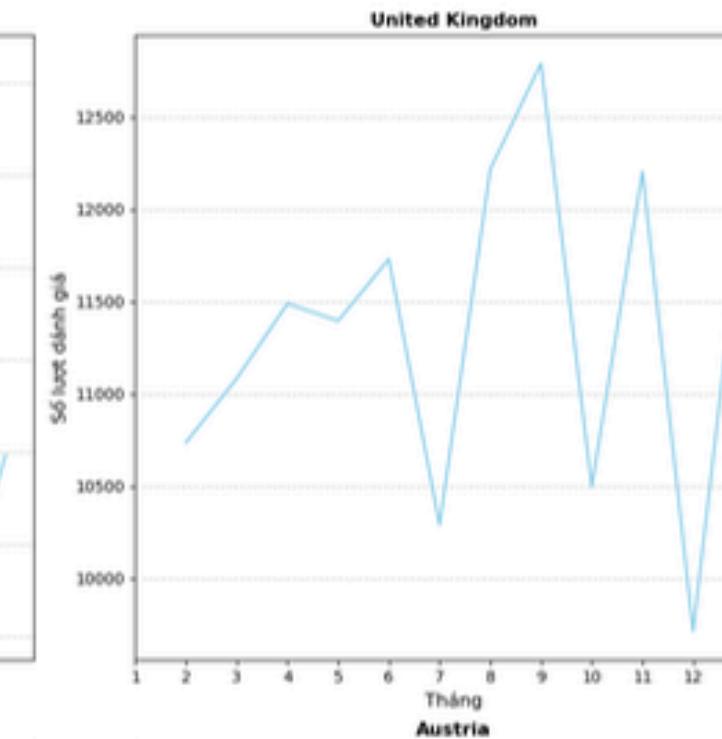
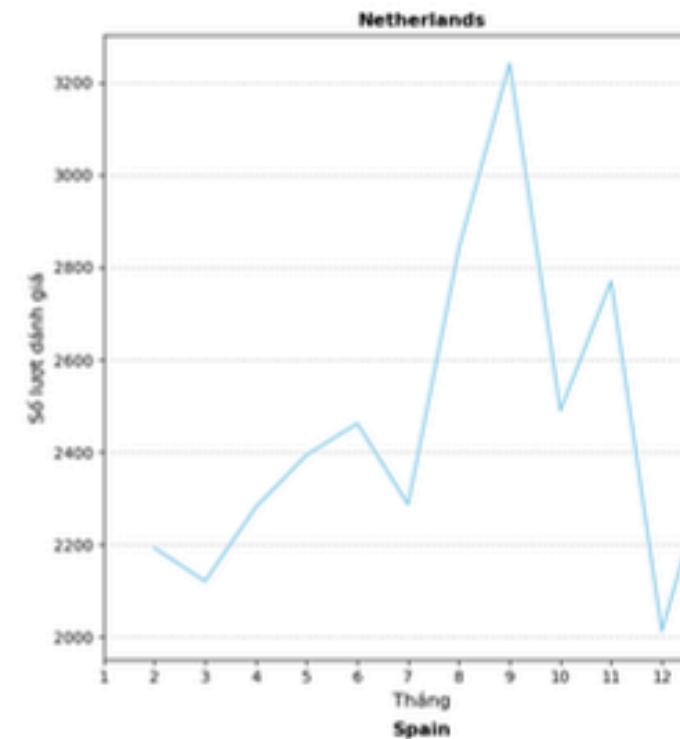
## Nhóm theo hình thức chuyến đi của người đánh giá cho từng quốc gia



# BIỂU ĐIỂN TRỰC QUAN

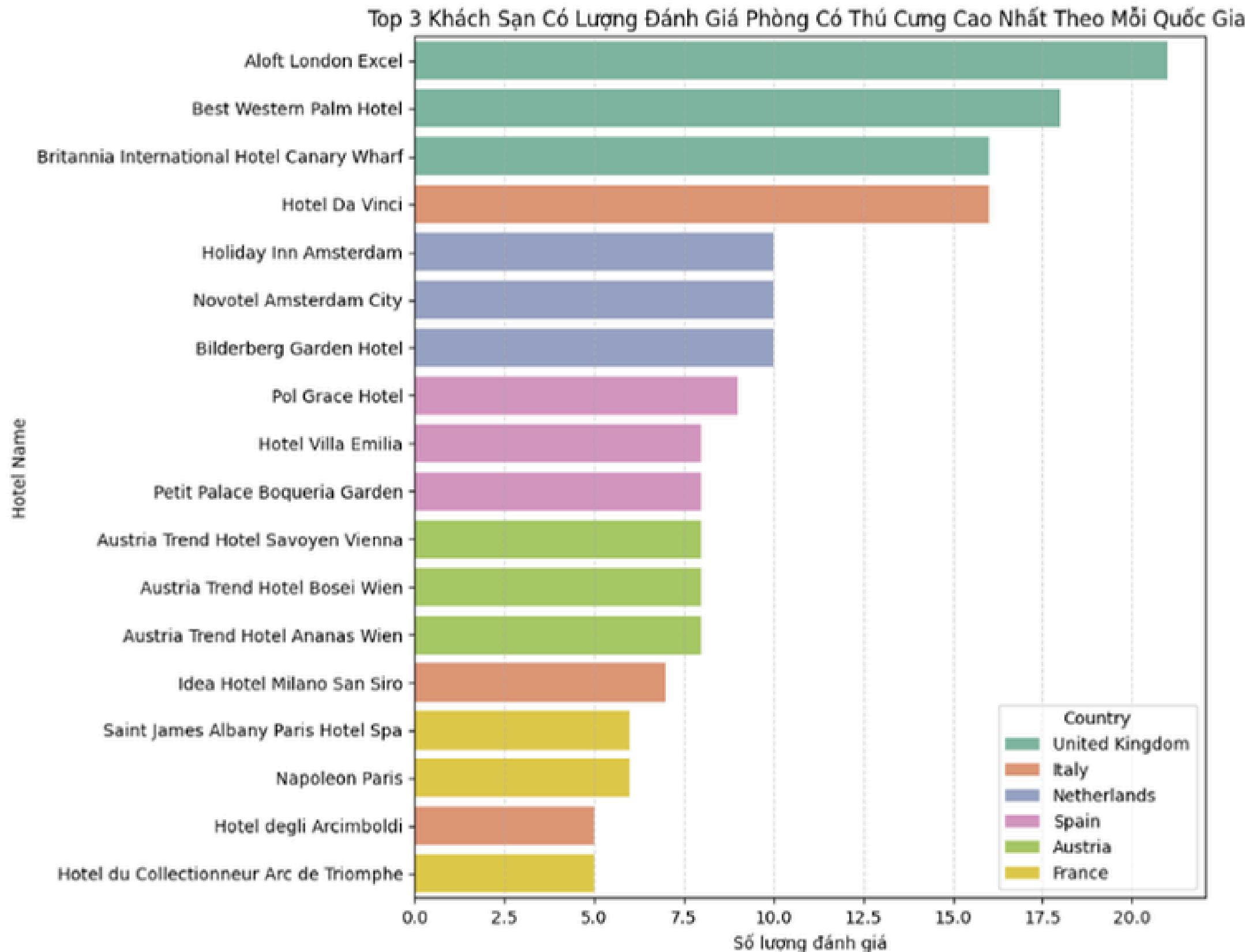
Số lượng đánh giá của các quốc gia theo từng tháng năm 2016

Số lượt đánh giá theo tháng cho từng quốc gia (Năm 2016)



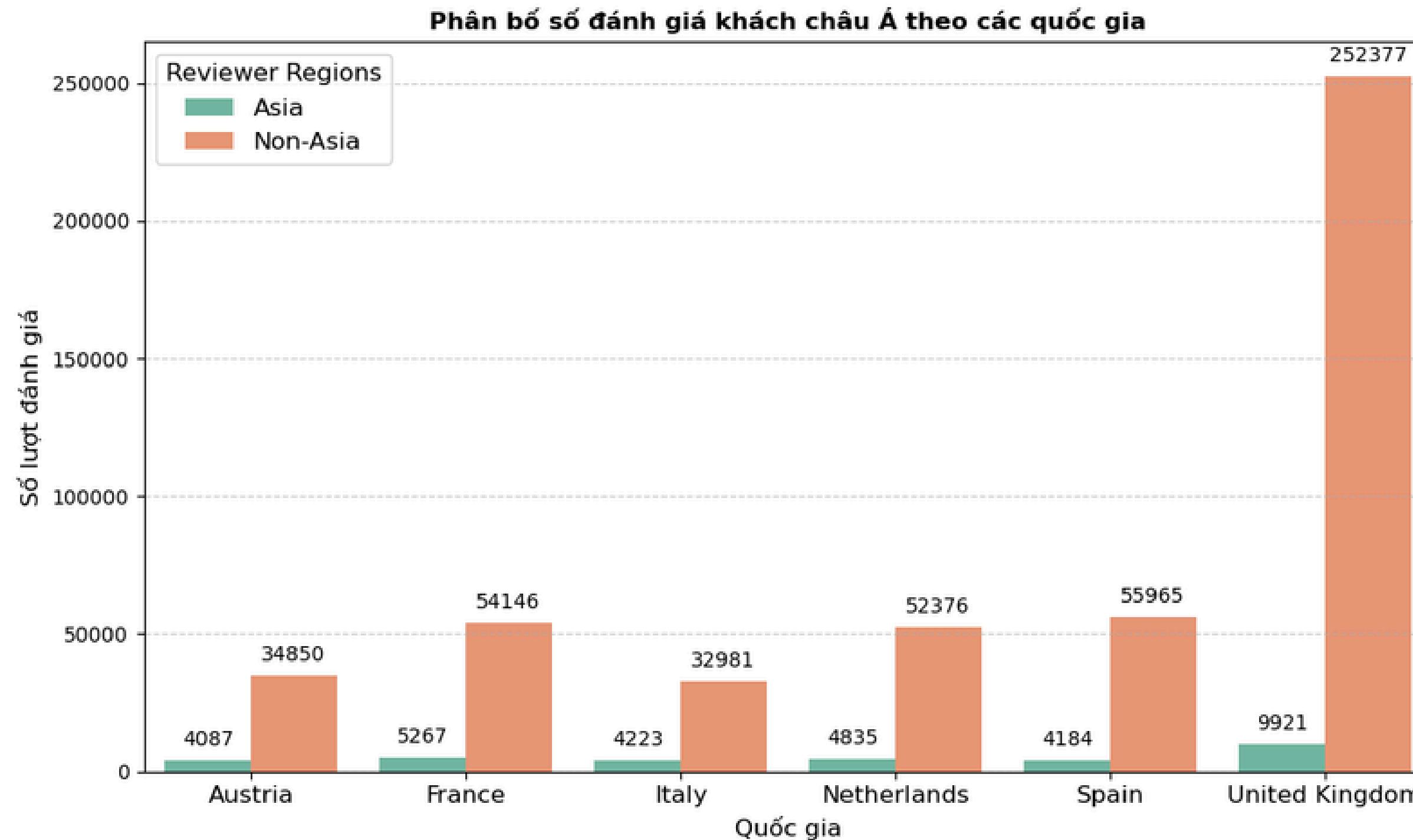
# BIỂU ĐIỂN TRỰC QUAN

*Top 3 khách sạn có dịch vụ cho mang theo thú cưng và có  
nhiều lượt đánh giá ở mỗi quốc gia*



# BIỂU ĐIỂN TRỰC QUAN

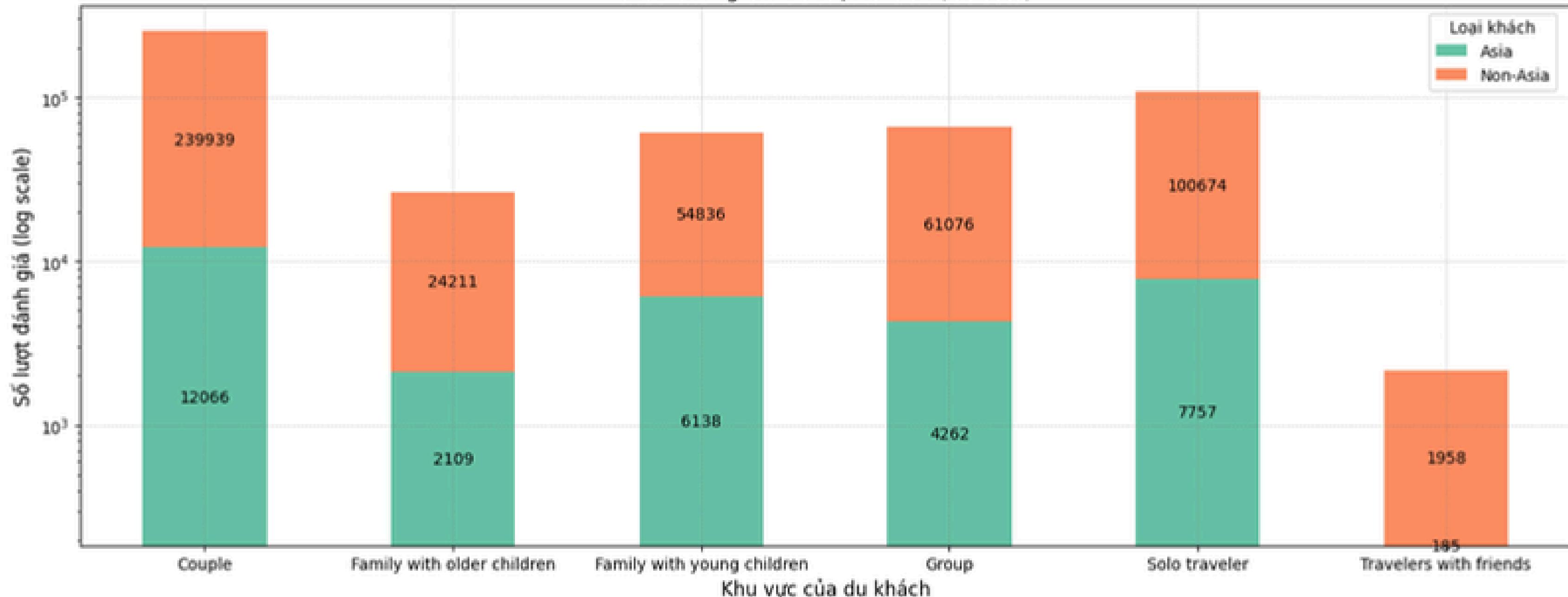
*Lượng đánh giá của khách châu Á ở từng quốc gia*



# BIỂU ĐIỂN TRỰC QUAN

*Số lượt đánh giá của du khách từ Châu Á (Asia) và từ khu vực khác (Non-Asia) theo thông tin về loại khách (Guests)*

Biểu đồ thể hiện số lượt đánh giá của Du khách từ Châu Á (Asia) và từ khu vực khác (Non-Asia) theo thông tin về loại khách (Guests)



# BIỂU DIỄN TRỰC QUAN

*Top 3 khách sạn đối với khách châu Á ở từng quốc gia*

## TIÊU CHÍ:

- Điểm đánh giá trung bình
- Số lượng đánh giá ( $\geq 150$ )



Ý

1. Room Mate Giulia
2. Hotel Berna
3. The Square Milano Duomo



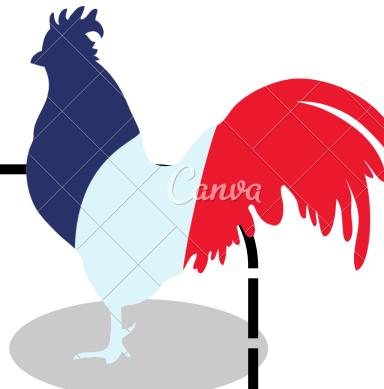
## VƯƠNG QUỐC ANH

1. The Montcalm Marble Arch
2. Amba Hotel Marble Arch
3. Royal Garden Hotel



## PHÁP

1. Pullman Paris Tour Eiffel
2. Warwick Paris Former Warwick Champs Elysees
3. Hotel California Champs Elysees



# BIỂU DIỄN TRỰC QUAN

*Top 3 khách sạn đối với khách châu Á ở từng quốc gia*

## TIÊU CHÍ:

- Điểm đánh giá trung bình
- Số lượng đánh giá ( $\geq 150$ )



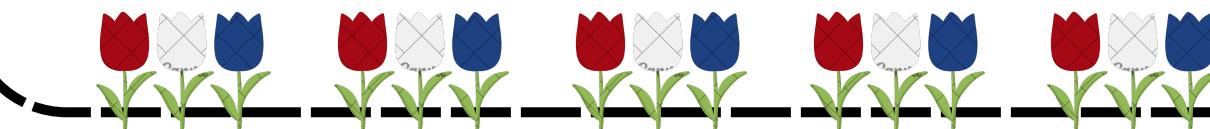
## ÁO

1. Austria Trend Hotel Europa Wien
2. Hotel de France Wien
3. Hilton Vienna



## HÀ LAN

1. art otel Amsterdam
2. Urban Lodge Hotel
3. The Student Hotel Amsterdam City



## TÂY BAN NHA

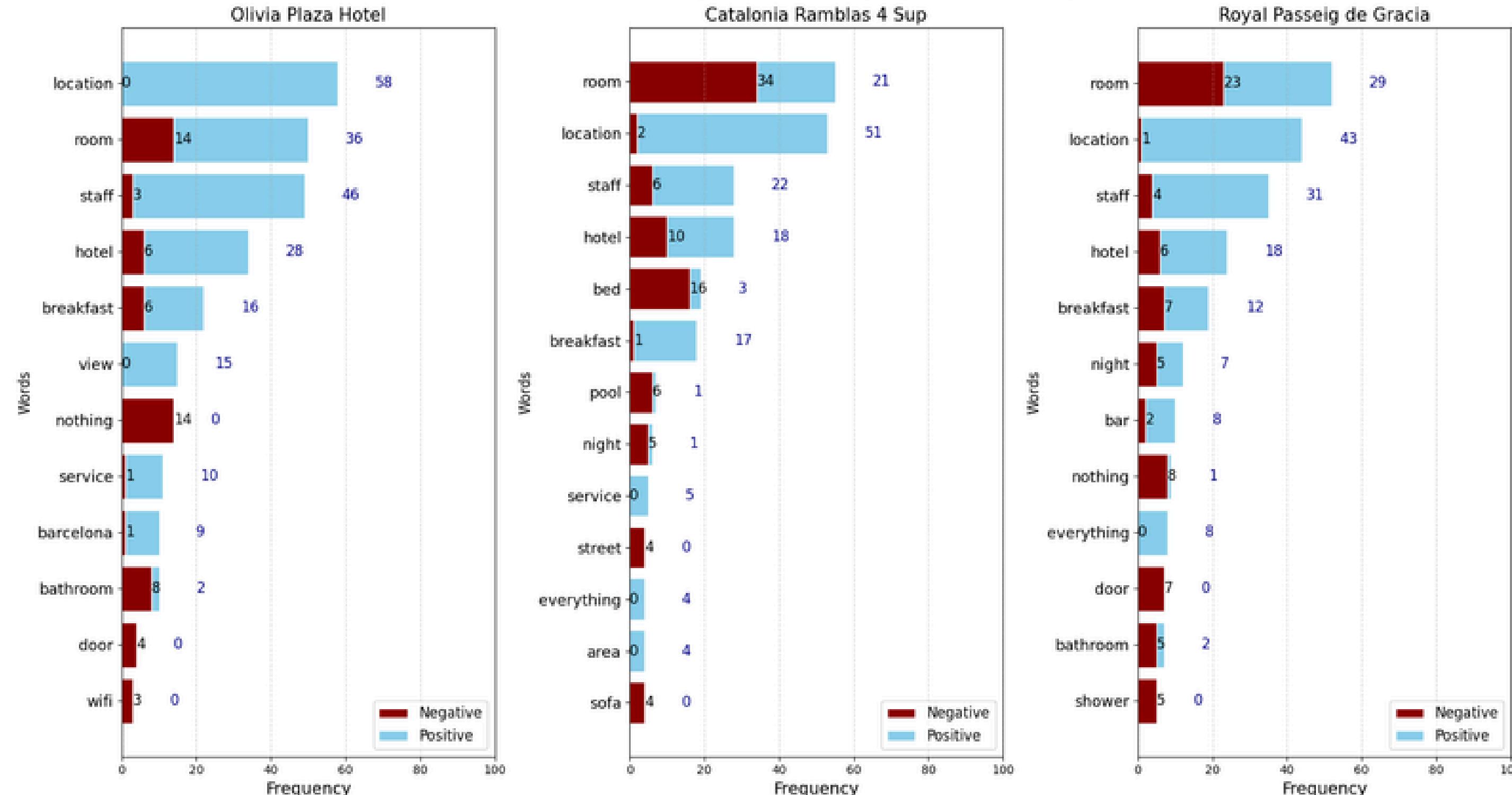
1. Olivia Plaza Hotel
2. Majestic Hotel Spa Barcelona GL
3. Eurostars Grand Marina Hotel GL



# BIỂU ĐIỂN TRỰC QUAN

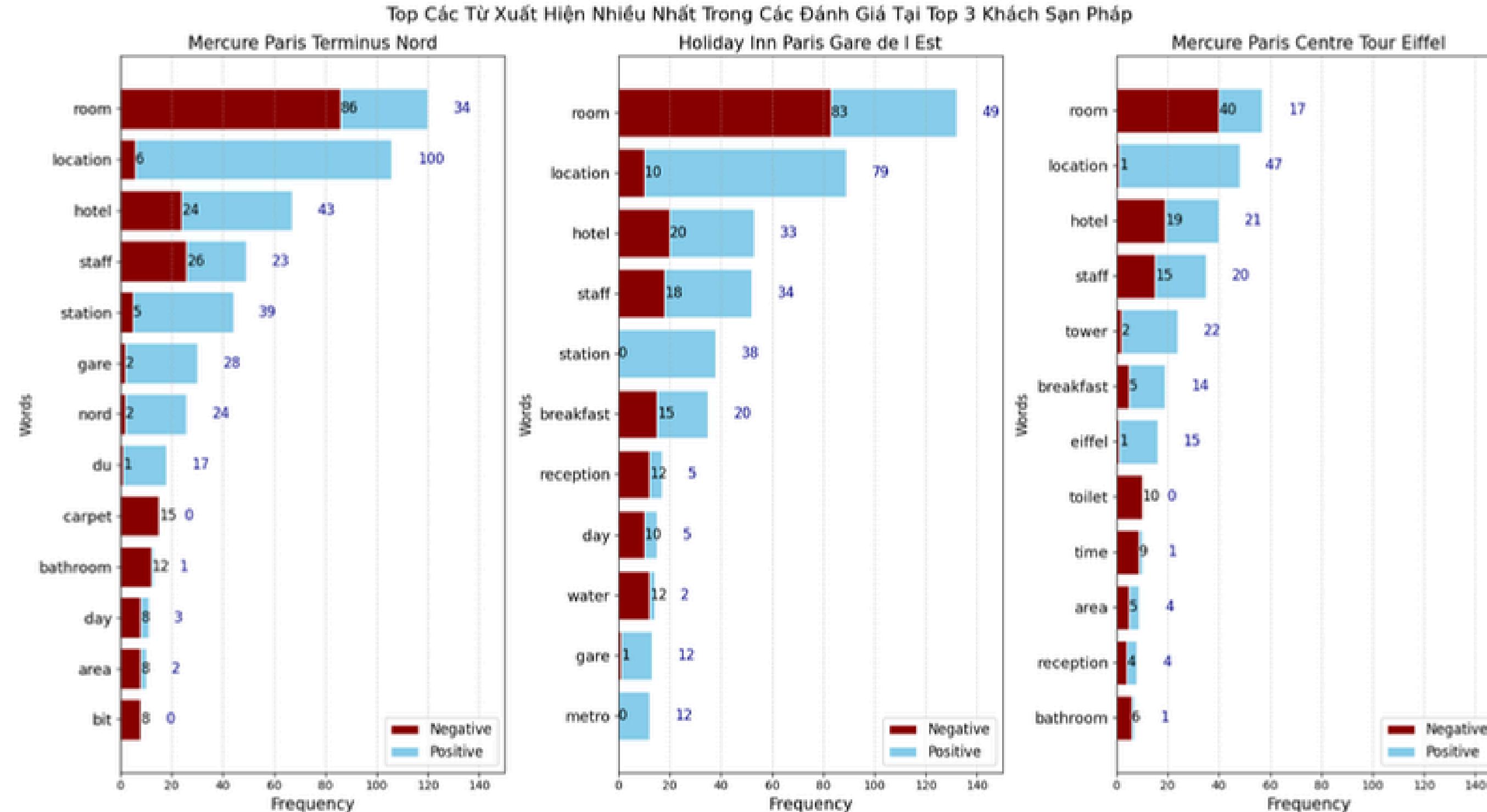
## Top 3 mà khách châu Á đánh giá nhiều nhất ở Tây Ban Nha

Các Từ Xuất Hiện Nhiều Nhất Trong Các Đánh Giá Tại Top 3 Khách Sạn Tây Ban Nha



# BIỂU ĐIỂN TRỰC QUAN

*Top 3 mà khách châu Á đánh giá nhiều nhất ở Pháp*



THANK FOR  
*thank for*  
WATCHING