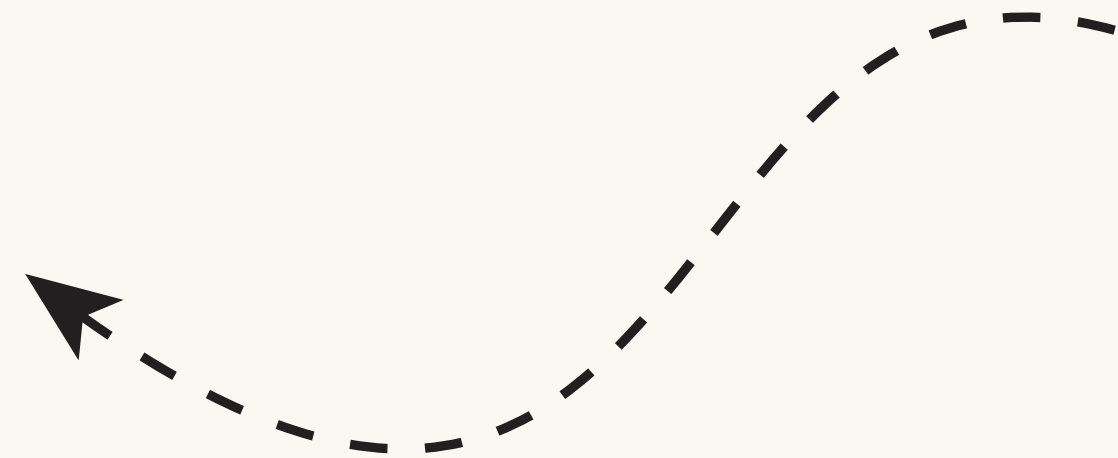
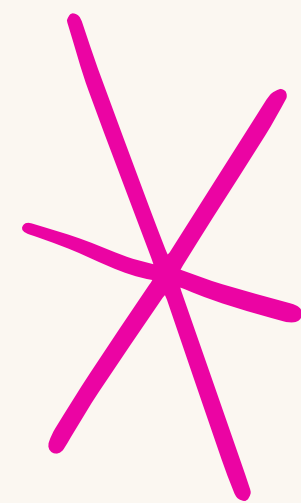



By Kem



System

RECOMMENDATION



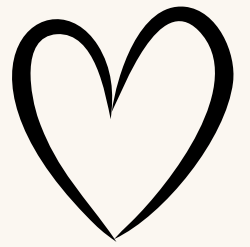


Agenda

1. Business Objective
2. Data Acquirement
3. Data preparation steps
4. Cosine_similarity
5. Genism
6. ALS
7. Insight Business
8. Conclusion



Business Objective



- Giả sử Agoda chưa triển khai hệ thống Recommender System giúp đề xuất khách sạn/ resort phù hợp tới người dùng và bạn được yêu cầu triển khai hệ thống này, bạn sẽ làm gì?
- Chủ khách sạn đặt trên Agoda muốn nắm rõ insight dựa trên thông tin khách hàng, bạn sẽ đem đến cho họ những gì

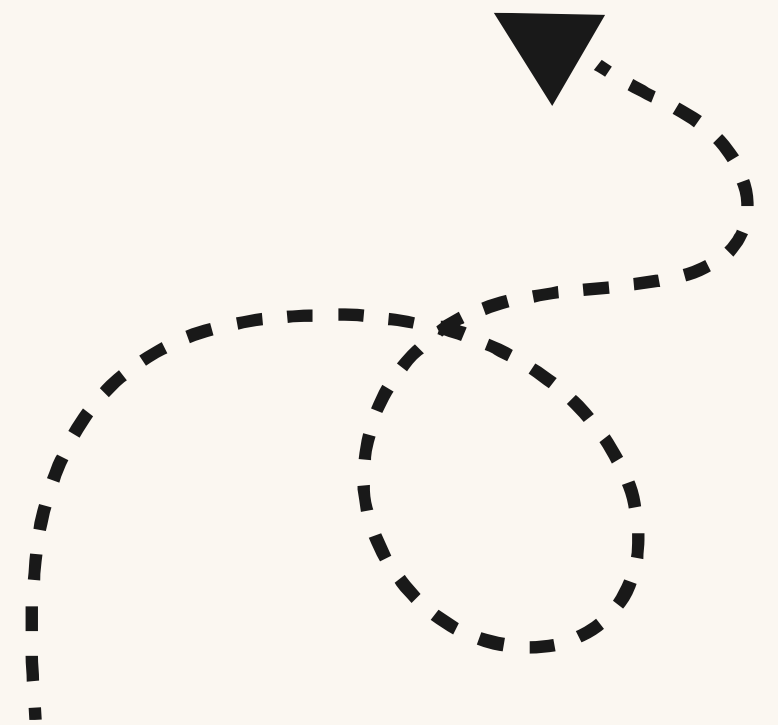




Data Acquirement

hotel_info.csv: chứa
thông tin các địa
điểm lưu trú

hotel_comment.csv:
chứa nhận xét, đánh giá
các địa điểm cư trú



Data preparation steps

Chuyển các dirty
value như No
information, none
thành NaN

Đổi các cột số
dạng object sang
float

Fill các cột float
có null sang 0

Data preparation steps

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 740 entries, 0 to 739
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype
---  -
0   num                    740 non-null   int64
1   Hotel_ID               740 non-null   object
2   Hotel_Name             740 non-null   object
3   Hotel_Rank             267 non-null   object
4   Hotel_Address          740 non-null   object
5   Total_Score            740 non-null   float64
6   Location               740 non-null   float64
7   Cleanliness            740 non-null   float64
8   Service                740 non-null   float64
9   Facilities              740 non-null   float64
10  Value_for_money         740 non-null   float64
11  Comfort_and_room_quality 740 non-null   float64
12  comments_count         740 non-null   int64
13  Hotel_Description       739 non-null   object
14  Hotel_Rank_Num          740 non-null   float64
dtypes: float64(8), int64(2), object(5)
memory usage: 86.8+ KB
```

Bảng Hotel_info.

- 1.Fill các cột Total Score, Location, Cleanliness, Service, Facilities, Value for money, Comfort null = 0
- 2.Tách cột Hotel_Rank sang cột Hotel_Rank_Num để lấy số sao
- 3.Fill cột Hotel_Description = “-”

Data preparation steps

```
.. <class 'pandas.core.frame.DataFrame'>
Index: 25906 entries, 0 to 80185
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Hotel_ID              25906 non-null  object
1   Reviewer_Name         25906 non-null  object
2   Nationality           25906 non-null  object
3   Group_Name            25906 non-null  object
4   Room_Type             25906 non-null  object
5   Score                 25906 non-null  float64
6   Score_Level           25906 non-null  object
7   Title                 25906 non-null  object
8   Body                  25906 non-null  object
9   Review_Date           25906 non-null  datetime64[ns]
10  Hotel_Name            25906 non-null  object
11  Days                  25906 non-null  int32
12  Month_Stay            25906 non-null  int32
13  Mean_Reviewer_Score   25906 non-null  float64
14  Review_ID_real        25906 non-null  object
dtypes: datetime64[ns](1), float64(2), int32(2), object(10)
memory usage: 3.0+ MB
```

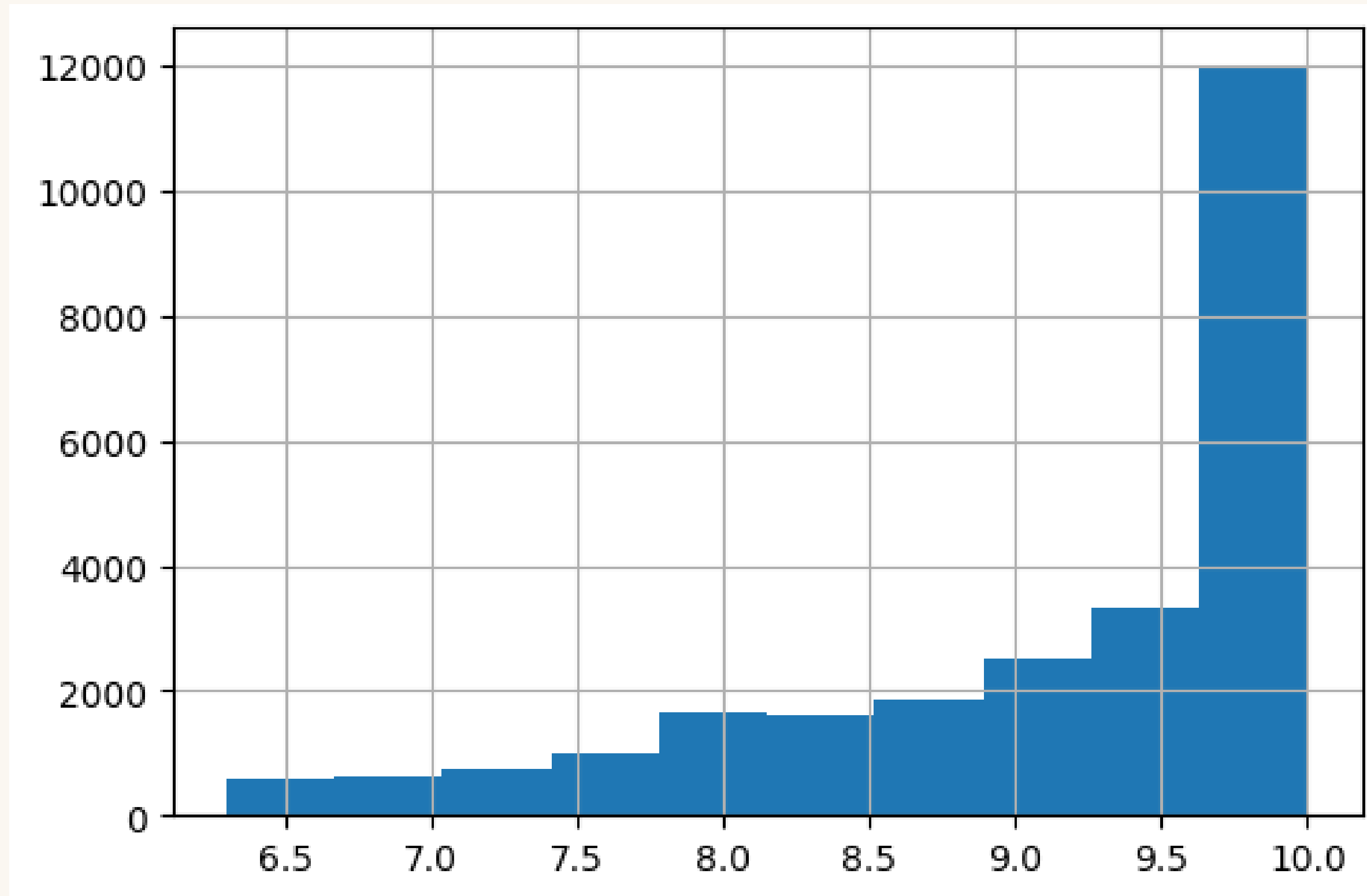
Bảng Hotel_comment.

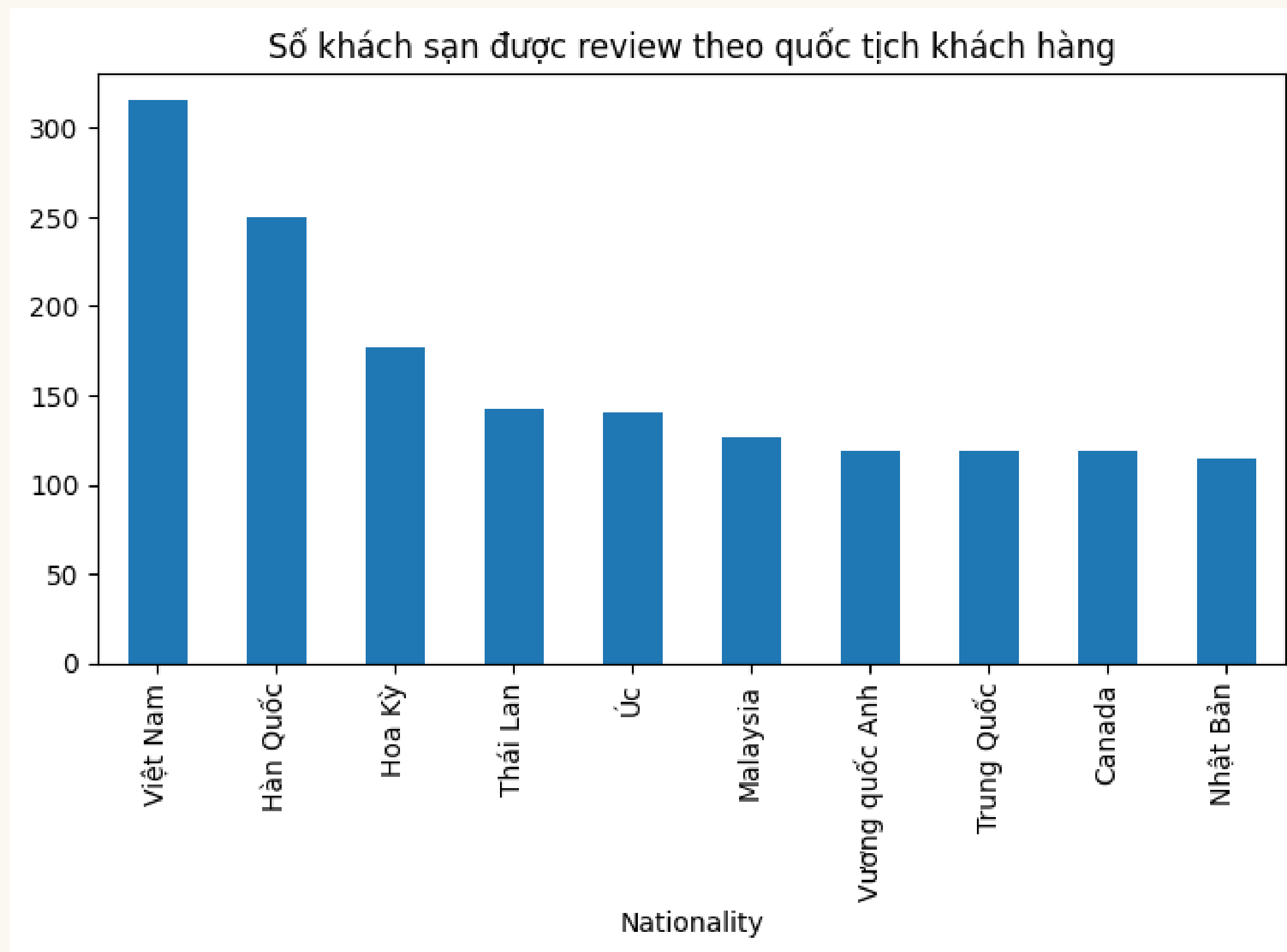
1. Đổi cột Score từ object sang float
2. Drop những hàng mà cột Review_Name và cột Body null
3. Tách cột Stay Detail thành cột Days và Month_Stay để lấy số ngày ở vào tháng ở của khách.
4. Đổi cột Review_Date sang dạng datetime
5. Tính Mean_Reviewer_Score cho từng Hotel_ID
6. Merge cột Hotel_Name của Hotel_info vào bảng Hotel_comment
7. Drop những hàng không có Hotel_Name
8. Do Review_ID không đáng tin tưởng nên Review_ID_Real sẽ là sự kết hợp giữa Hotel_ID, Review_Name và Review_Date. Vì 1 người chỉ nên có 1 cmt về 1 Hotel_ID trong 1 ngày. Các cmt cùng ngày cùng tên và cùng Hotel_ID được xem xét là cmt trùng
9. Cột Body remove wrong word, stopword, replace teencode, emoji, english

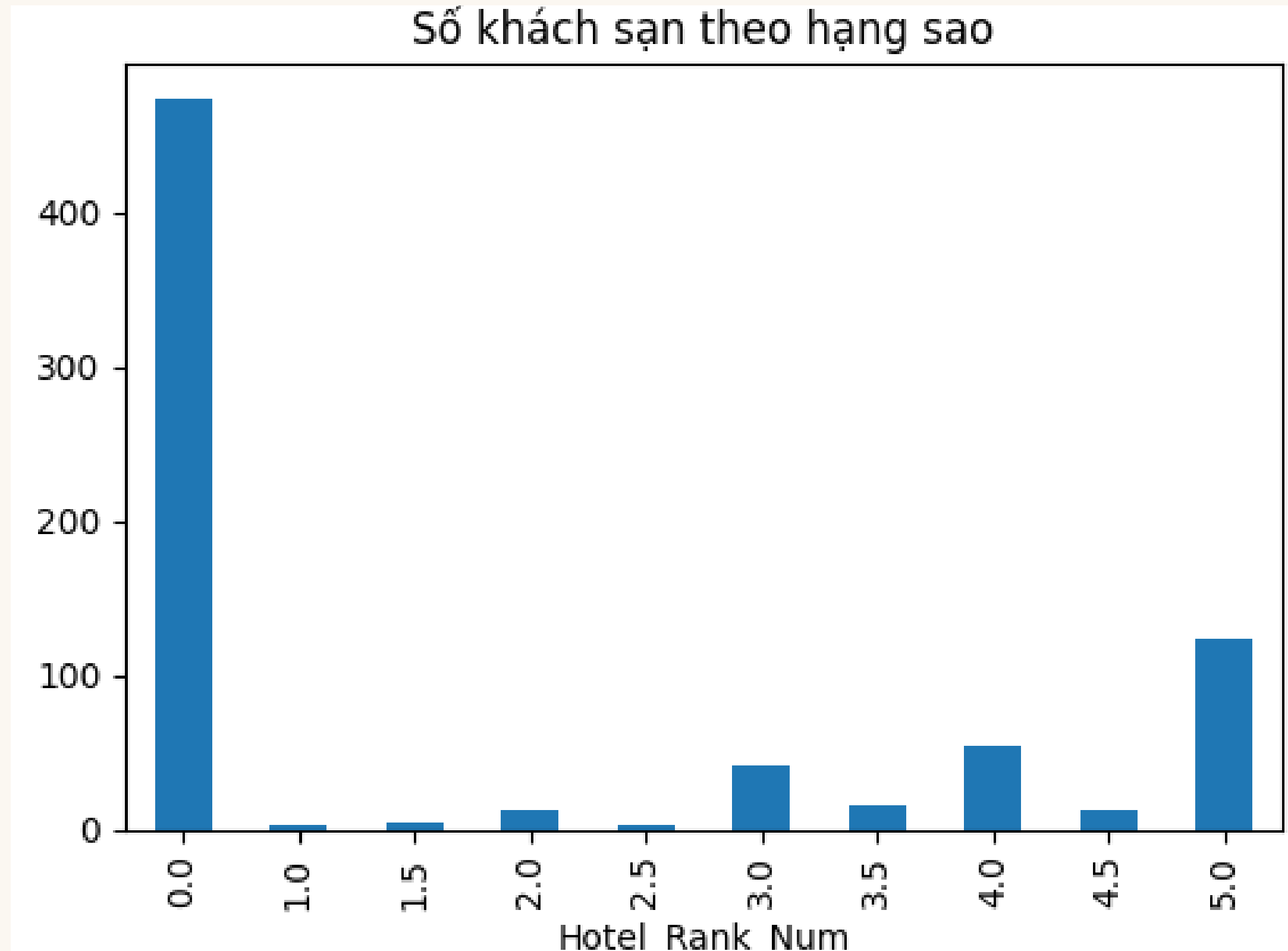
=> Data từ 80168 dòng xuống còn 25906 dòng

EDA

Số khách đánh giá theo thang điểm

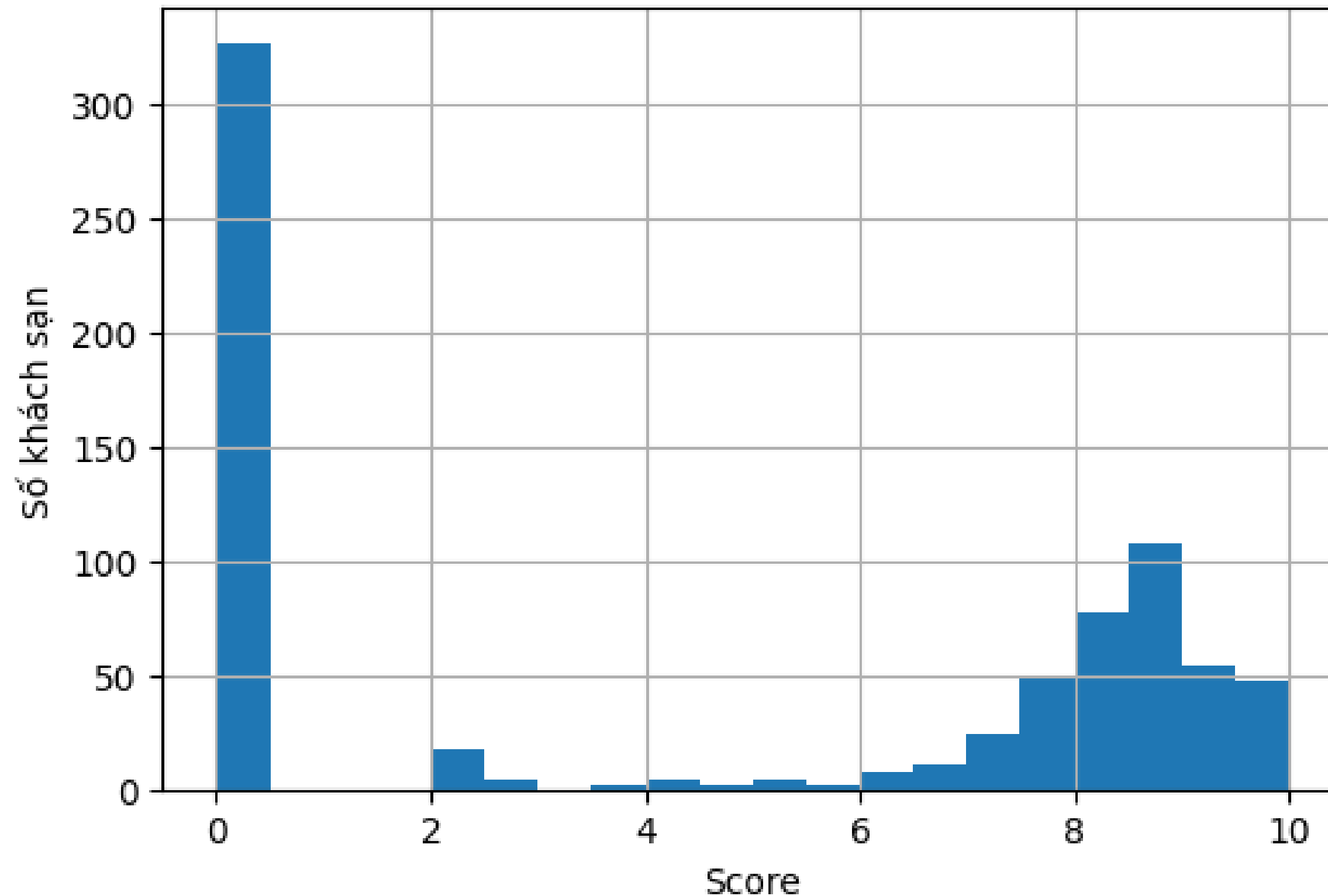






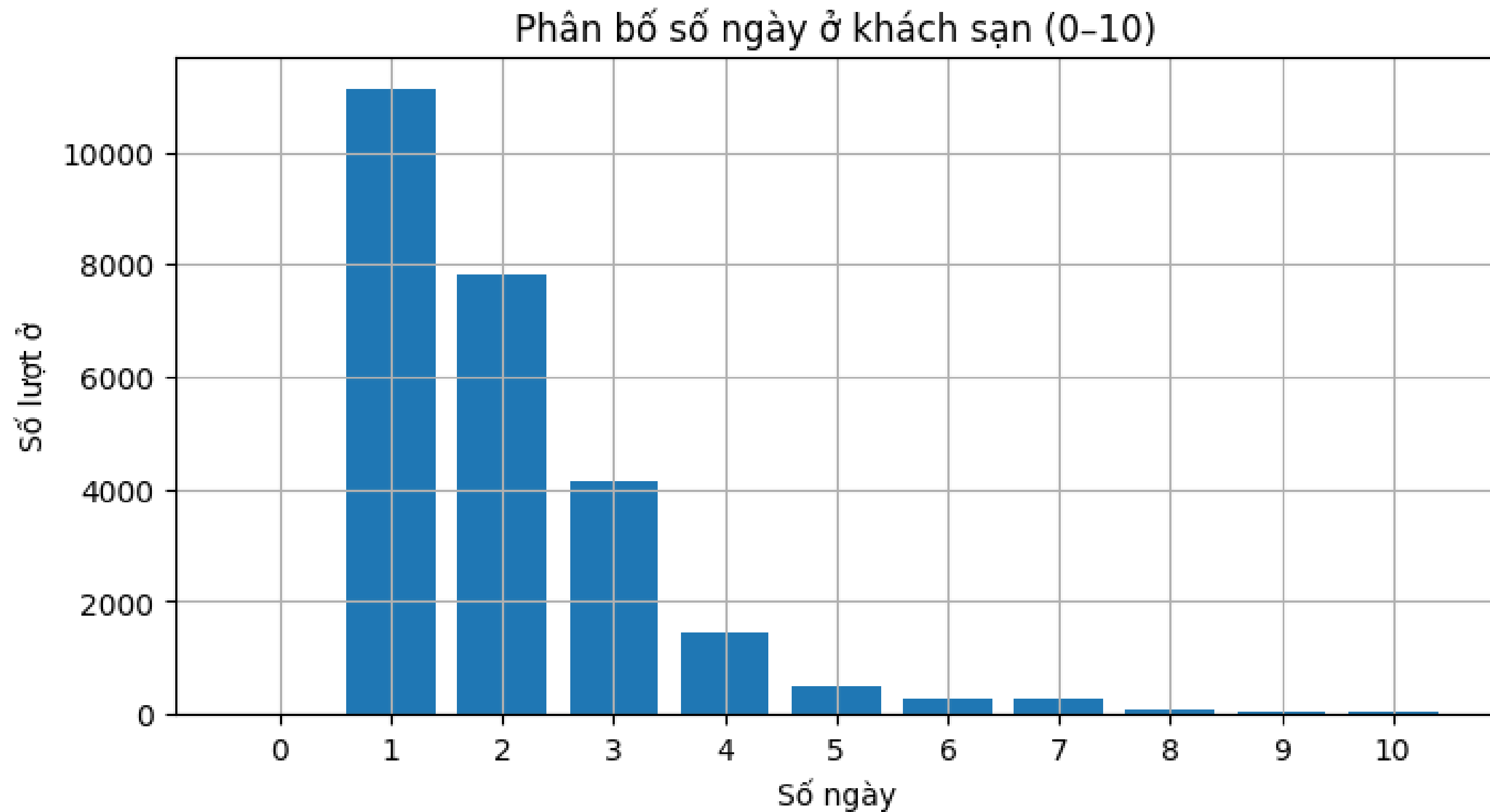
Những khách sạn 0 sao có thể là những khách sạn có Hotel_Rank là null
=> Như vậy có khoảng gần 500 khách sạn không có sao

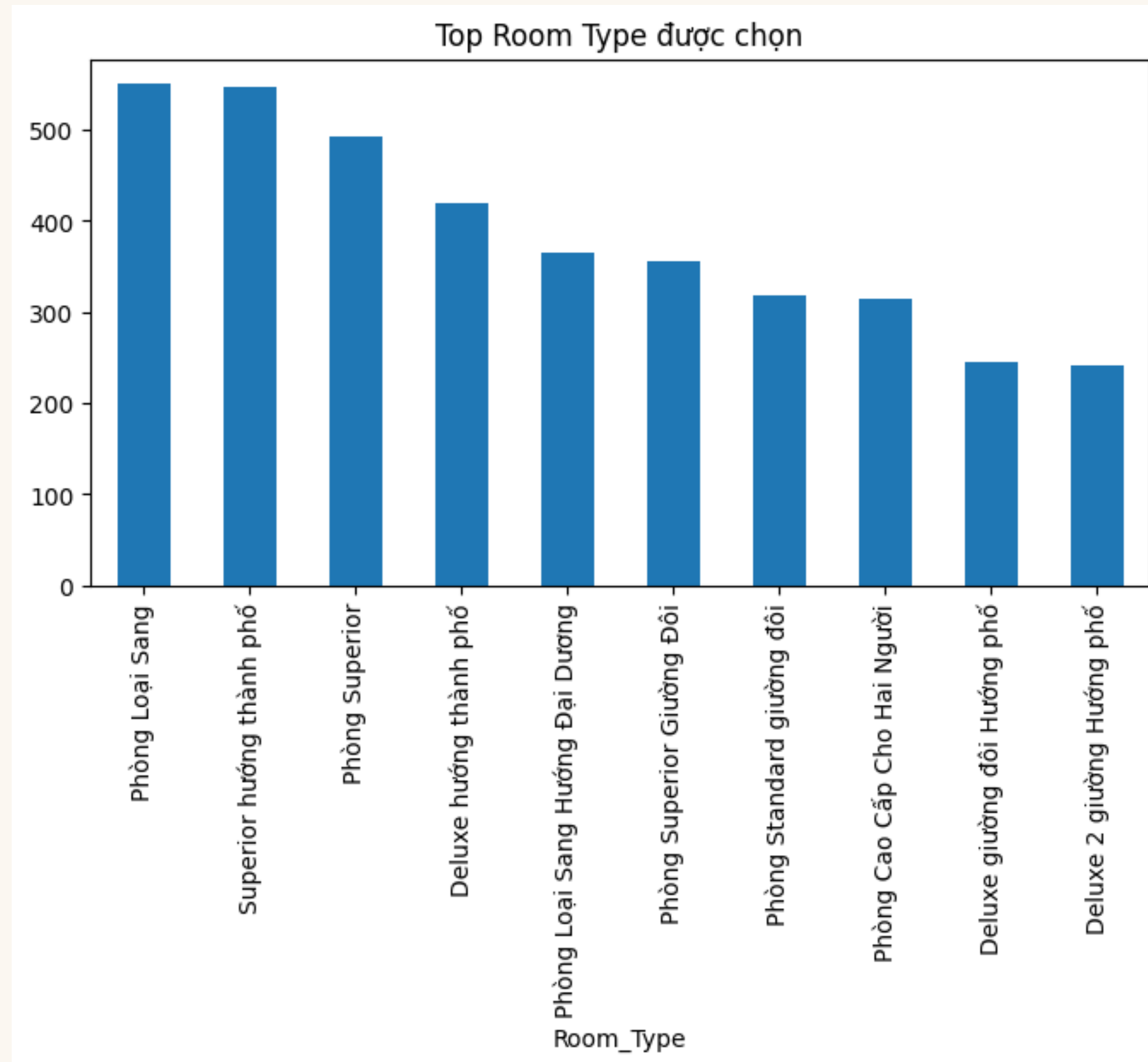
Phân bố điểm trung bình khách sạn (Total_Score)



Những khách sạn Total Score = 0 có thể là những khách sạn có Total Score là null

=> Như vậy có khoảng gần 350 khách sạn không có Total_Score





Content-based Filtering

Cosine_similarity

```
.. <class 'pandas.core.frame.DataFrame'>
Index: 2088 entries, 0 to 25862
Data columns (total 8 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Hotel_ID               2088 non-null   object
1   Body_clean             2088 non-null   string
2   Hotel_Name             2088 non-null   string
3   Room_Type              2088 non-null   string
4   Hotel_Address          2088 non-null   string
5   Hotel_Description      2088 non-null   string
6   Content                2088 non-null   object
7   Content_wt             2088 non-null   string
dtypes: object(2), string(6)
memory usage: 146.8+ KB
```

1. Tạo 1 bảng Hotel_corpus được merge từ 2 bảng info và comment bao gồm các cột Hotel_ID, Hotel_Name, Hotel_Address, Hotel_Description, Room_Type, Body_Clean
2. Tạo 1 cột Content bao gồm các cột trên
3. Check content duplicate giảm từ 25906 dòng xuống 2088 dòng

Consimetheo profile



	0	1	2	3	4	5	6	7	8	9	...	2078	2079	2080	2081	2082	2083	2084	2085	2086
0	1.000000	0.025188	0.025183	0.025136	0.025140	0.025147	0.025183	0.025133	0.023904	0.023912	...	0.020384	0.020369	0.020377	0.020420	0.020360	0.020377	0.020362	0.020368	0.020371
1	0.025188	1.000000	0.999905	0.999911	0.999951	0.999923	0.999920	0.999864	0.416424	0.416652	...	0.461911	0.461299	0.462041	0.461940	0.461654	0.461609	0.461709	0.461627	0.461455
2	0.025183	0.999905	1.000000	0.999879	0.999856	0.999873	0.999888	0.999940	0.416288	0.416420	...	0.462006	0.461394	0.462158	0.462034	0.461771	0.461726	0.461826	0.461721	0.461571
3	0.025136	0.999911	0.999879	1.000000	0.999872	0.999888	0.999990	0.999829	0.416221	0.416353	...	0.462082	0.461470	0.462246	0.462121	0.461858	0.461802	0.461913	0.461797	0.461642
4	0.025140	0.999951	0.999856	0.999872	1.000000	0.999884	0.999871	0.999861	0.416875	0.417067	...	0.462555	0.461945	0.462663	0.462572	0.462297	0.462276	0.462331	0.462272	0.462100
...
2083	0.020377	0.461609	0.461726	0.461802	0.462276	0.461557	0.461715	0.462281	0.624754	0.624568	...	0.999468	0.999388	0.999539	0.999081	0.999473	1.000000	0.999460	0.999365	0.999481
2084	0.020362	0.461709	0.461826	0.461913	0.462331	0.461657	0.461816	0.462335	0.624505	0.624365	...	0.999809	0.999725	0.999920	0.999301	0.999987	0.999460	1.000000	0.999702	0.999971
2085	0.020368	0.461627	0.461721	0.461797	0.462272	0.461575	0.461711	0.462277	0.625306	0.625196	...	0.999712	0.999676	0.999781	0.999209	0.999688	0.999365	0.999702	1.000000	0.999721
2086	0.020371	0.461455	0.461572	0.461648	0.462100	0.461403	0.461562	0.462105	0.624638	0.624497	...	0.999801	0.999746	0.999898	0.999279	0.999965	0.999481	0.999978	0.999723	1.000000
2087	0.022008	0.461514	0.461608	0.461684	0.462159	0.461462	0.461598	0.462164	0.624340	0.624231	...	0.999458	0.999421	0.999527	0.999113	0.999434	0.999269	0.999447	0.999398	0.999460



Consine Similarity



	Source_Hotel_ID	Source_Hotel_Name	Recommended_Hotel_ID	Recommended_Hotel_Name	Recommended_Hotel_Address	Recommended_Hotel_Description	Similarity
0	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	60 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam	Khách sạn Mường Thanh Luxury Nha Trang - Nơi l...	1.000
1	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	3_10	Khách sạn Dendro Gold (Dendro Gold Hotel)	86/4 Tran Phu Street, Lộc Thọ, Nha Trang, Việt...	Khách sạn Dendro Gold - Kỳ nghỉ thú vị tại Nha...	0.823
2	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	2_11	Maris Hotel Nha Trang	27 Trần Quang Khải, Phường Lộc Thọ, Thành phố ...	Maris Hotel Nha Trang - Nơi lưu trú đẳng cấp t...	0.811
3	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	20_2	Khách Sạn MerPerle Beach (MerPerle Beach Hotel)	88A Tran Phu Street, Lộc Thọ, Nha Trang, Việt ...	Khách Sạn MerPerle Beach - Nơi Lý Tưởng Cho Kỳ...	0.808
4	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	4_25	Diamond Bay Hotel	20 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam	Diamond Bay Hotel - Nơi lưu trú sang trọng tại...	0.800
...

=> Điểm Similarity rơi vào khoảng 0.8 gần 80% giống



Content-based Filtering

Gensim

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2088 entries, 0 to 2087
Data columns (total 11 columns):
#   Column                Non-Null Count  Dtype
---  -
0   original_index         2088 non-null   int64
1   original_index         2088 non-null   int64
2   Hotel_ID               2088 non-null   object
3   Body_clean             2088 non-null   string
4   Hotel_Name             2088 non-null   string
5   Room_Type              2088 non-null   string
6   Hotel_Address          2088 non-null   string
7   Hotel_Description      2088 non-null   string
8   Content                2088 non-null   object
9   Content_wt             2088 non-null   object
10  original_index         2088 non-null   int64
dtypes: int64(3), object(3), string(5)
memory usage: 179.6+ KB
```

1. Tạo 1 bảng Hotel_corpus được merge từ 2 bảng info và comment bao gồm các cột Hotel_ID, Hotel_Name, Hotel_Address, Hotel_Description, Room_Type, Body_Clean
2. Tạo 1 cột Content bao gồm các cột trên
3. Check content duplicate giảm từ 25906 dòng xuống 2088 dòng
4. Tạo 1 cột Content_wt để làm Dictionary

Gensim

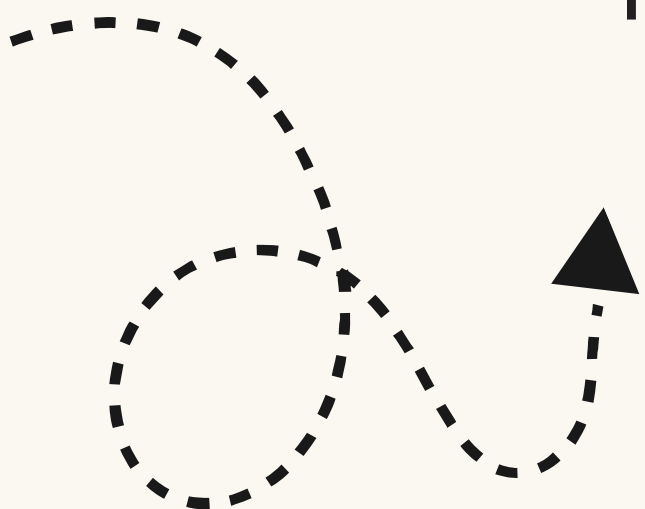
Content-based Filtering

Gensim

	Source_Hotel	Recommended_Hotel	Address	Description	Score
0	Oceanus Oasis Retreat Muong Thanh Vien Trieu	Oceanus Oasis Retreat Muong Thanh Vien Trieu	Phạm Văn Đồng, Vĩnh Phước, Nha Trang, Việt Nam...	Oceanus Oasis Retreat Muong Thanh Vien Trieu - ...	0.999987
1	Oasis Retreat - Muong Thanh Vien Trieu	Oasis Retreat - Muong Thanh Vien Trieu	3 Pham Van Dong, OC 1 B Muong Thanh Vien Trieu...	Oasis Retreat - Muong Thanh Vien Trieu\n\nLàm ...	0.999960
2	Oasis Retreat - Muong Thanh Vien Trieu	Muong Thanh Beachfront Apartment Nha Trang	03 Phạm Văn Đồng, Tòa Nhà OC2B - Mường Thanh V...	Muong Thanh Beachfront Apartment Nha Trang\n\n...	0.605492
3	Oasis Retreat - Muong Thanh Vien Trieu	HANZ Condo Hotel Muong Thanh Vien Trieu	5 Phạm Văn Đồng, Muong Thanh Vien Trieu, Nha T...	HANZ Condo Hotel Muong Thanh Vien Trieu\n\nThứ...	0.477664
4	Oasis Retreat - Muong Thanh Vien Trieu	HANZ Muong Thanh Vien Trieu Condo Hotel	5 Phạm Văn Đồng, Vĩnh Hải, Nha Trang, Khánh Hò...	HANZ Muong Thanh Vien Trieu Condo Hotel\n\nTận...	0.407923

=> Điểm Similarity của Gensim là khoảng 70%

=> Điểm này nhỏ hơn Điểm của Consine nên dùng Consine sẽ hiệu quả hơn



PYSPARK ALS

Collaborative~~X~~Filtering

1. Tạo bảng chạy ALS với các cột Nationality, Hotel_ID, Score, Hotel_Description và Body_clean,
2. Vì Nationality và Hotel_ID là object nên phải tạo 1 cột ID khác theo Nationality và Hotel_ID theo numeric

```
root
|-- nationality_id: double (nullable = false)
|-- hotel_numeric_id: double (nullable = false)
|-- Score: double (nullable = true)
|-- Body_clean: string (nullable = true)
|-- Hotel_Description: string (nullable = true)
```

PYSPARK ALS

Collaborative Filtering

RESULT



Hotel_ID	Hotel_Name	Hotel_Address
28_24	InterContinental Residences Nha Trang	32 Tran Phu Street, Lộc Thọ, Nha Trang, Việt Nam, 650000
8_27	Vinpearl Luxury Nha Trang	Đảo Hòn Tre, Phường Vĩnh Nguyên , Hòn Tre, Nha Trang, Việt N
5_1	Nomad Apartment	58 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam, 650000
12_7	Gran Meliá Nha Trang	Bãi Tiên, Duong De, Vinh Hoa Ward, Nha Trang City, Khanh Hoa
12_22	Khách Sạn Mojzo (Mojzo Inn)	65/07 Nguyen Thiet Thuat- Nha Trang- Khanh hoa, Lộc Thọ, Nha
7_29	Căn hộ 35 m² 1 phòng ngủ, 1 phòng tắm riêng ở Biển Bãi Dài (Angela Serviced Apartment (near Airport Cam Ranh))	The Arena Cam Ranh, Cam Nghia, Cam Ranh, Khanh Hoa, Viet Nam
5_14	Realus Sea-view Apartment in Cam Ranh Nha Trang	Nguyễn Tất Thành, Arena Cam Ranh , Cam Hải Đông, Nha Trang, V
19_4	Căn hộ studio 1100 m² có 1 phòng tắm riêng ở Vĩnh Phước (MCR Apartments Nha Trang)	Vĩnh Phước, Nha Trang, Việt Nam
37_28	Biệt thự 250 m² 6 phòng ngủ, 6 phòng tắm riêng ở Cam Hải Đông (Room#102- Garden view- H2B Homestay Bai Dai Beach)	Cam Hải Đông, Nha Trang, Việt Nam
12_17	The Signature Hotel Nha Trang	86A Trần Phú, Lộc Thọ, Lộc Thọ, Nha Trang, Việt Nam, 650000

	Hotel_Address	prediction
	32 Tran Phu Street, Lộc Thọ, Nha Trang, Việt Nam, 650000	10.077377
	Đảo Hòn Tre, Phường Vĩnh Nguyên , Hòn Tre, Nha Trang, Việt Nam, 650000	10.022418
	58 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam, 650000	9.965127
	Bãi Tiên, Duong De, Vinh Hoa Ward, Nha Trang City, Khanh Hoa Province , Vĩnh Hòa, Nha Trang, Việt Nam, 650000	9.899907
	65/07 Nguyen Thiet Thuat- Nha Trang- Khanh hoa, Lộc Thọ, Nha Trang, Việt Nam, 650000	9.888357
Biển Bãi Dài (Angela Serviced Apartment (near Airport Cam Ranh))	The Arena Cam Ranh, Cam Nghia, Cam Ranh, Khanh Hoa, Viet Nam, Cam Hải Đông, Nha Trang, Việt Nam, 650000	9.817557
	Nguyễn Tất Thành, Arena Cam Ranh , Cam Hải Đông, Nha Trang, Việt Nam, 57000	9.817557
Vĩnh Phước (MCR Apartments Nha Trang)	Vĩnh Phước, Nha Trang, Việt Nam	9.817557
Cam Hải Đông (Room#102- Garden view- H2B Homestay Bai Dai Beach)	Cam Hải Đông, Nha Trang, Việt Nam	9.817557
	86A Trần Phú, Lộc Thọ, Lộc Thọ, Nha Trang, Việt Nam, 650000	9.817557

PYSPARK ALS

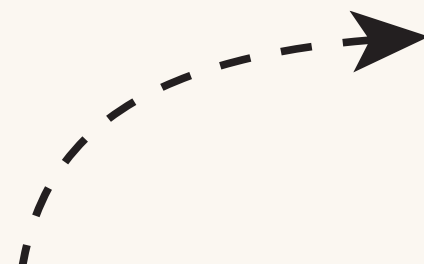
Collaborative Filtering



RMSE

0.9877484755582976

Nhận xét: Điểm RMSE khá cao. Tuy nhiên dùng ALS không được tối ưu hóa như Content-base - Consine vì không tính được những comment của tích cực tiêu cực của khách hàng mà chỉ dựa trên được Score, Hotel_ID, số cmt.



Insight Business



Hàm tìm hotel theo id hoặc key word trả ra thông tin dạng bảng

	Hotel_ID	Hotel_Name	Hotel_Rank_Num	Hotel_Address	Total_Score	Location	Cleanliness	Service	Facilities	Value_for_money	Comfort_and_room_quality	comments_count
0	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	5.0	60 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam	8.8	9.4	8.9	8.9	8.7	8.7	8.3	1269
1	18_11	Mường Thanh Nha Trang Poli Apartment	0.0	3 Phạm Văn Đồng, Vĩnh Hải, Vĩnh Phước, Nha Tra...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2	26_25	Chung cư 60 m² 2 phòng ngủ, 1 phòng tắm riêng ...	0.0	Vĩnh Phước, Nha Trang, Việt Nam	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	21_17	Căn hộ 45 m² 1 phòng ngủ, 1 phòng tắm riêng ở ...	0.0	Lộc Thọ, Nha Trang, Việt Nam	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
4	36_20	Căn hộ 50 m² 2 phòng ngủ, 1 phòng tắm riêng ở ...	0.0	Lộc Thọ, Nha Trang, Việt Nam	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
5	39_1	Căn hộ 60 m² 2 phòng ngủ, 1 phòng tắm riêng ở ...	0.0	Vĩnh Phước, Nha Trang, Việt Nam	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
6	40_15	Căn hộ 68 m² 2 phòng ngủ, 2 phòng tắm riêng ở ...	0.0	Xương Huân, Nha Trang, Việt Nam	10.0	10.0	10.0	10.0	10.0	10.0	0.0	1
7	40_16	Căn hộ 70 m² 3 phòng ngủ, 2 phòng tắm riêng ở ...	0.0	Vĩnh Phước, Nha Trang, Việt Nam	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

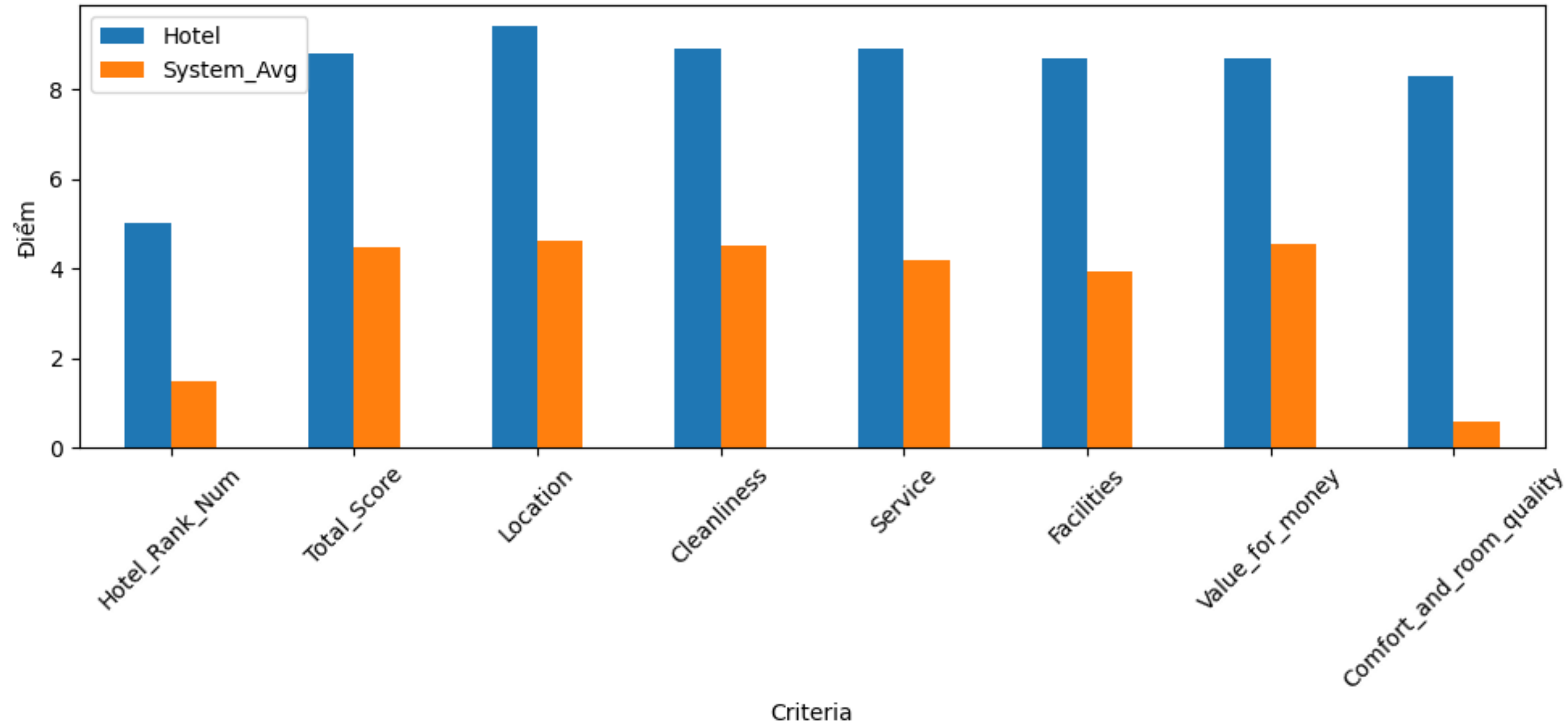
	Hotel_ID	Hotel_Name	Hotel_Rank_Num	Hotel_Address	Total_Score	Location	Cleanliness	Service	Facilities	Value_for_money	Comfort_and_room_quality	comments_count
0	1_1	Khách sạn Mường Thanh Luxury Nha Trang (Muong ...	5.0	60 Trần Phú, Lộc Thọ, Nha Trang, Việt Nam	8.8	9.4	8.9	8.9	8.7	8.7	8.3	1269



Insight Business

Hàm tìm khách sạn theo ID hoặc key word trả ra biểu đồ phân tích
VD: Theo keyword “Mường Thanh”

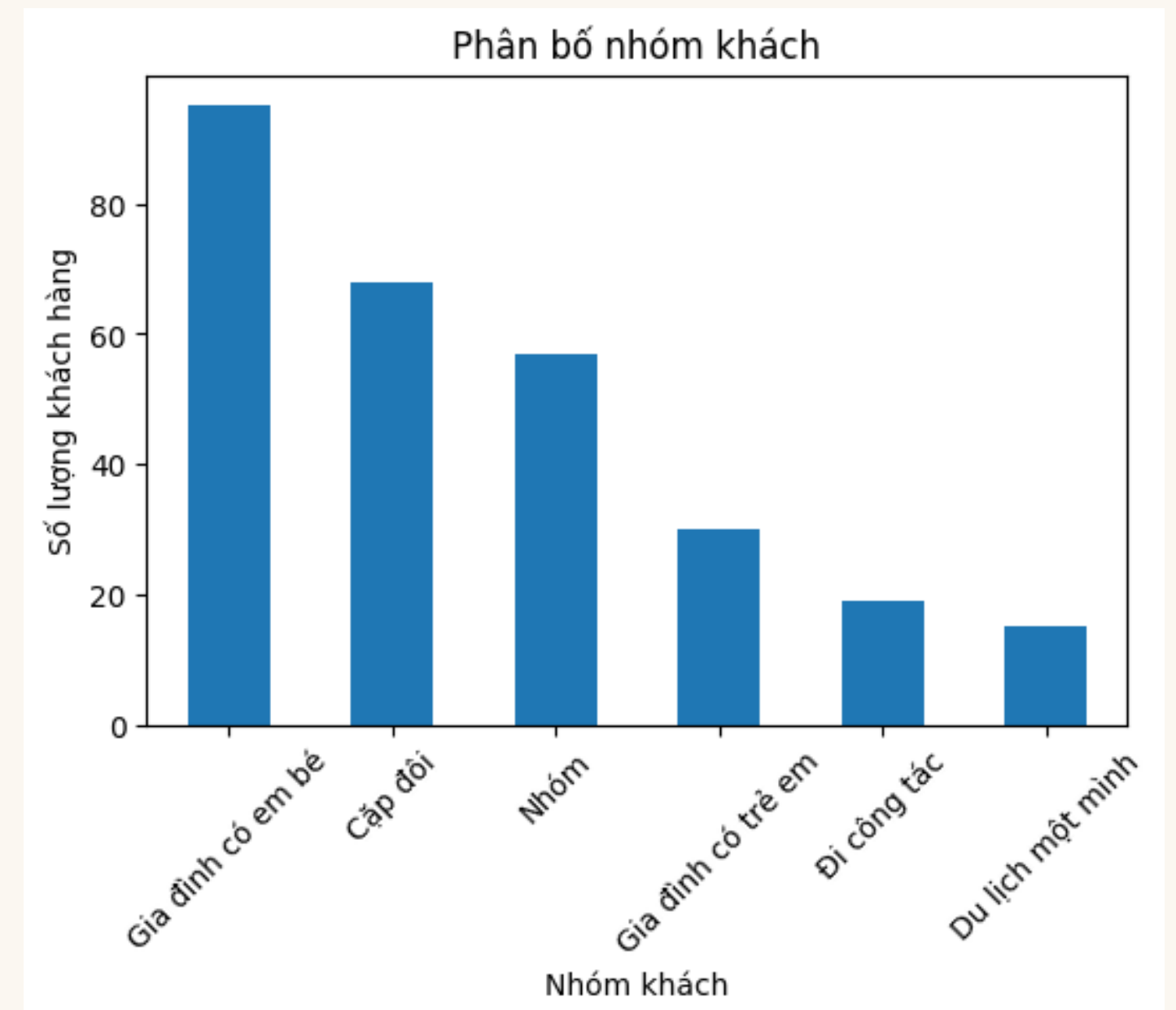
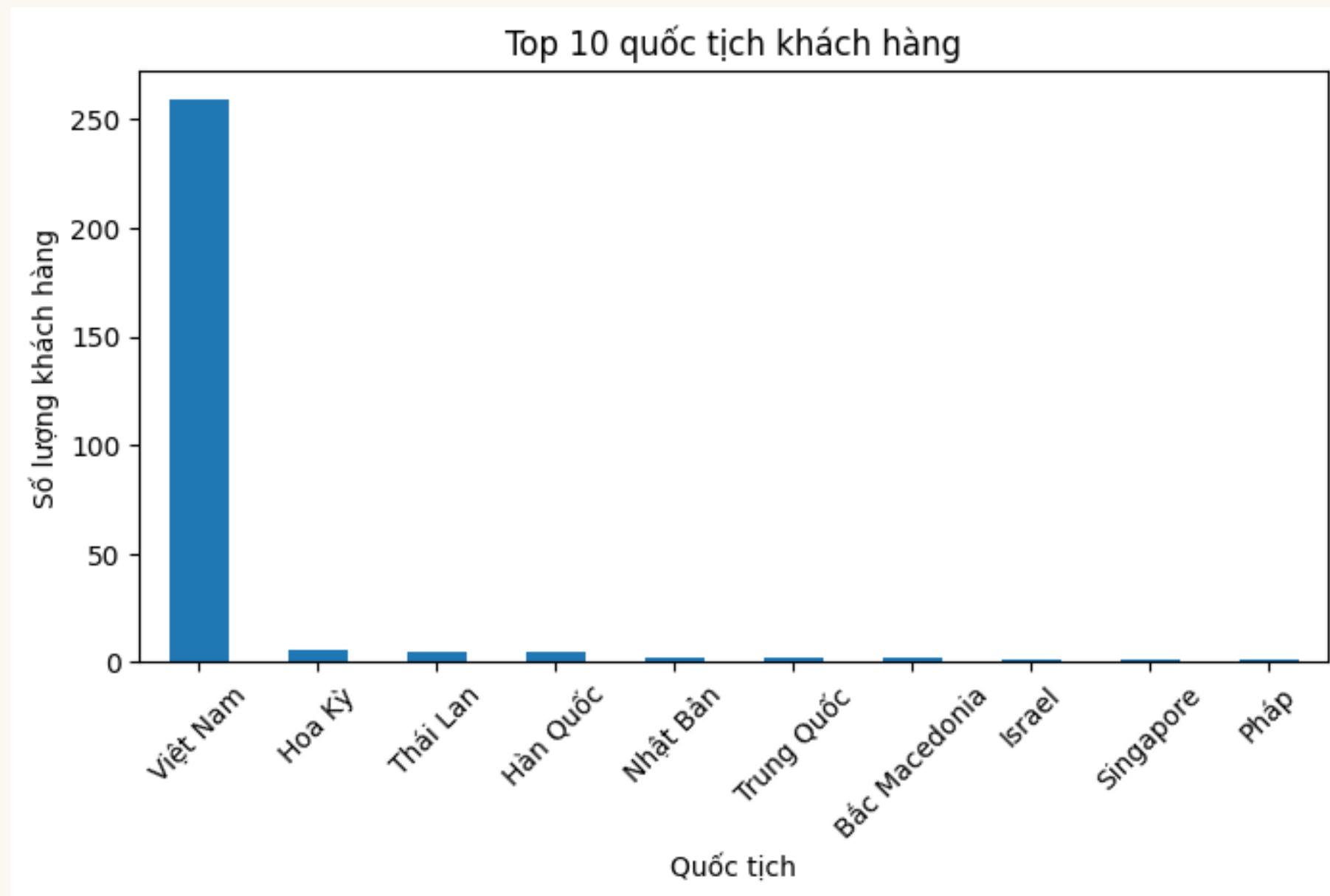
So sánh điểm khách sạn 'Khách sạn Mường Thanh Luxury Nha Trang (Muong Thanh Luxury Nha Trang Hotel)' với trung bình hệ thống



Insight Business

Hàm tìm theo ID hoặc key word cho chủ khách sạn, trả các biểu đồ thống kê cho khách sạn đó Quốc tịch, nhóm khách, xu hướng theo thời gian

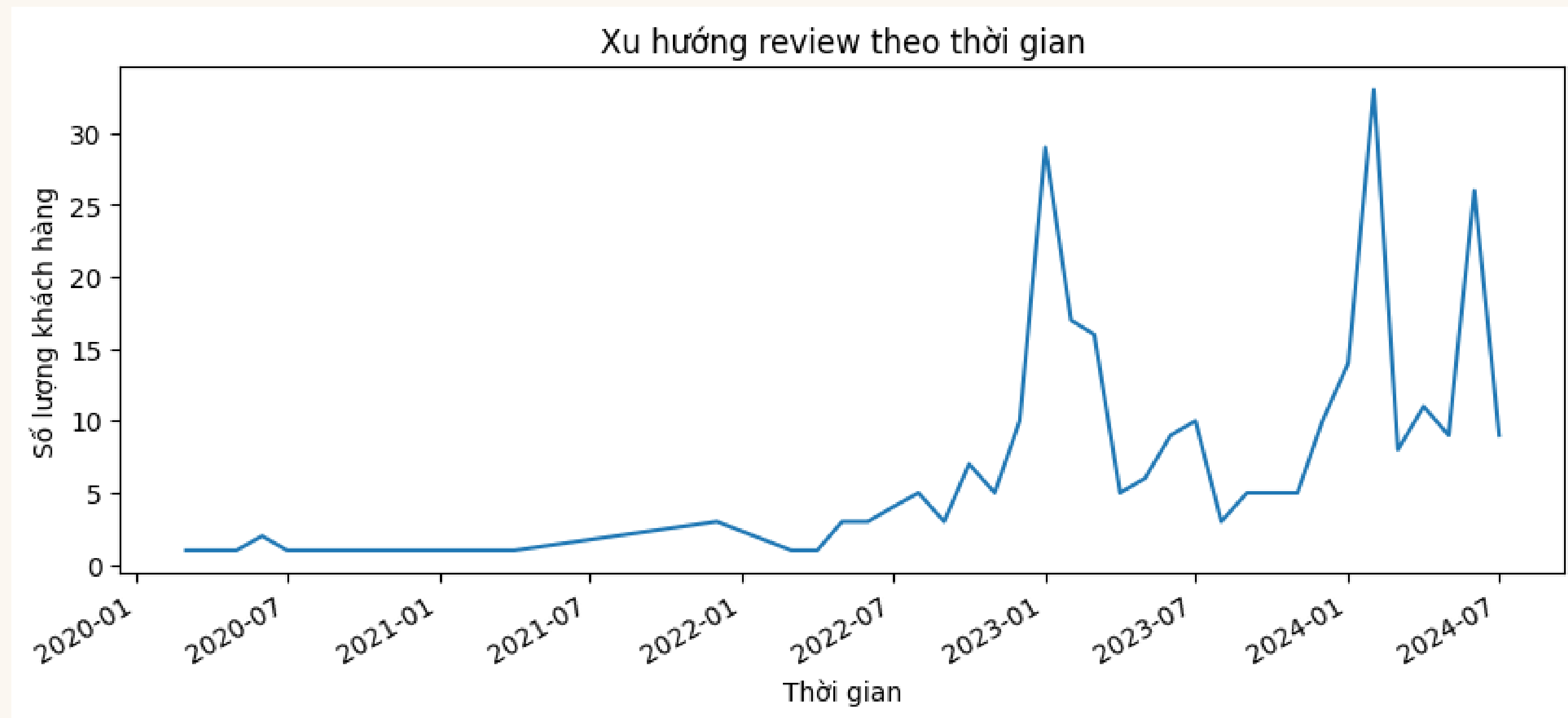
VD: Theo keyword “Mường Thanh”



Insight Business

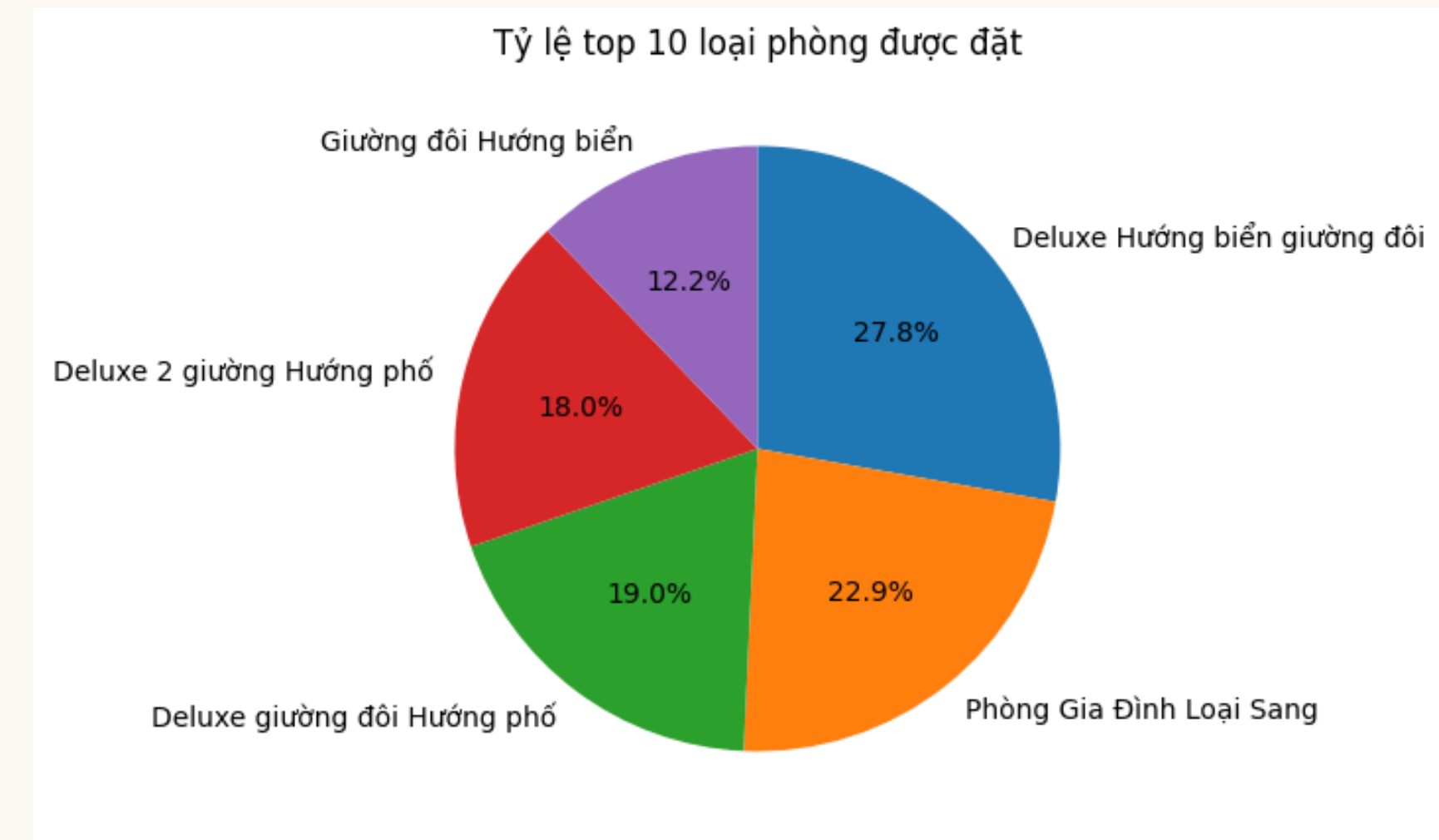
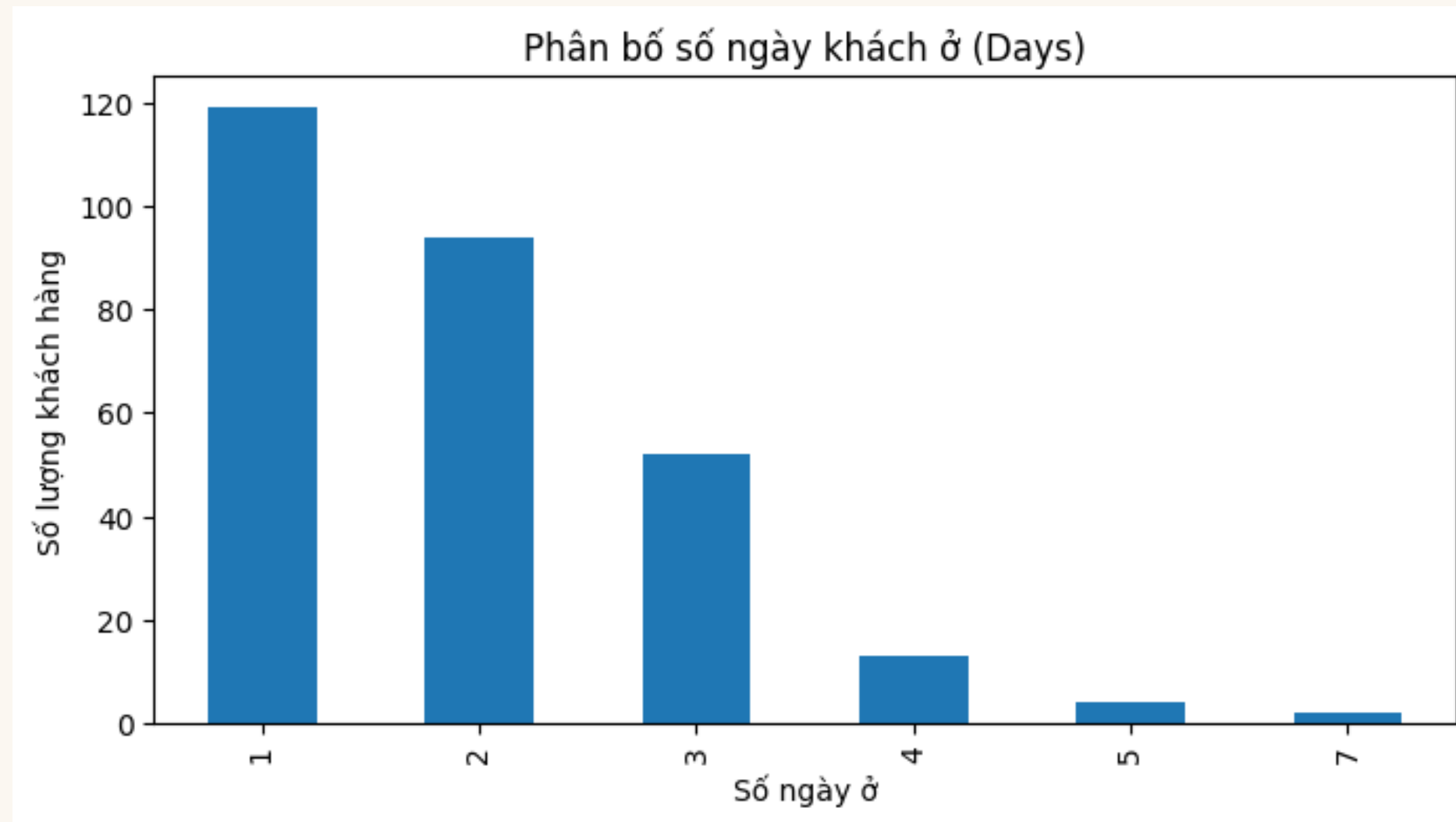
Hàm tìm theo ID hoặc key word cho chủ khách sạn, trả các biểu đồ thống kê cho khách sạn đó Quốc tịch, nhóm khách, xu hướng theo thời gian

VD: Theo keyword “Mường Thanh”



Insight Business

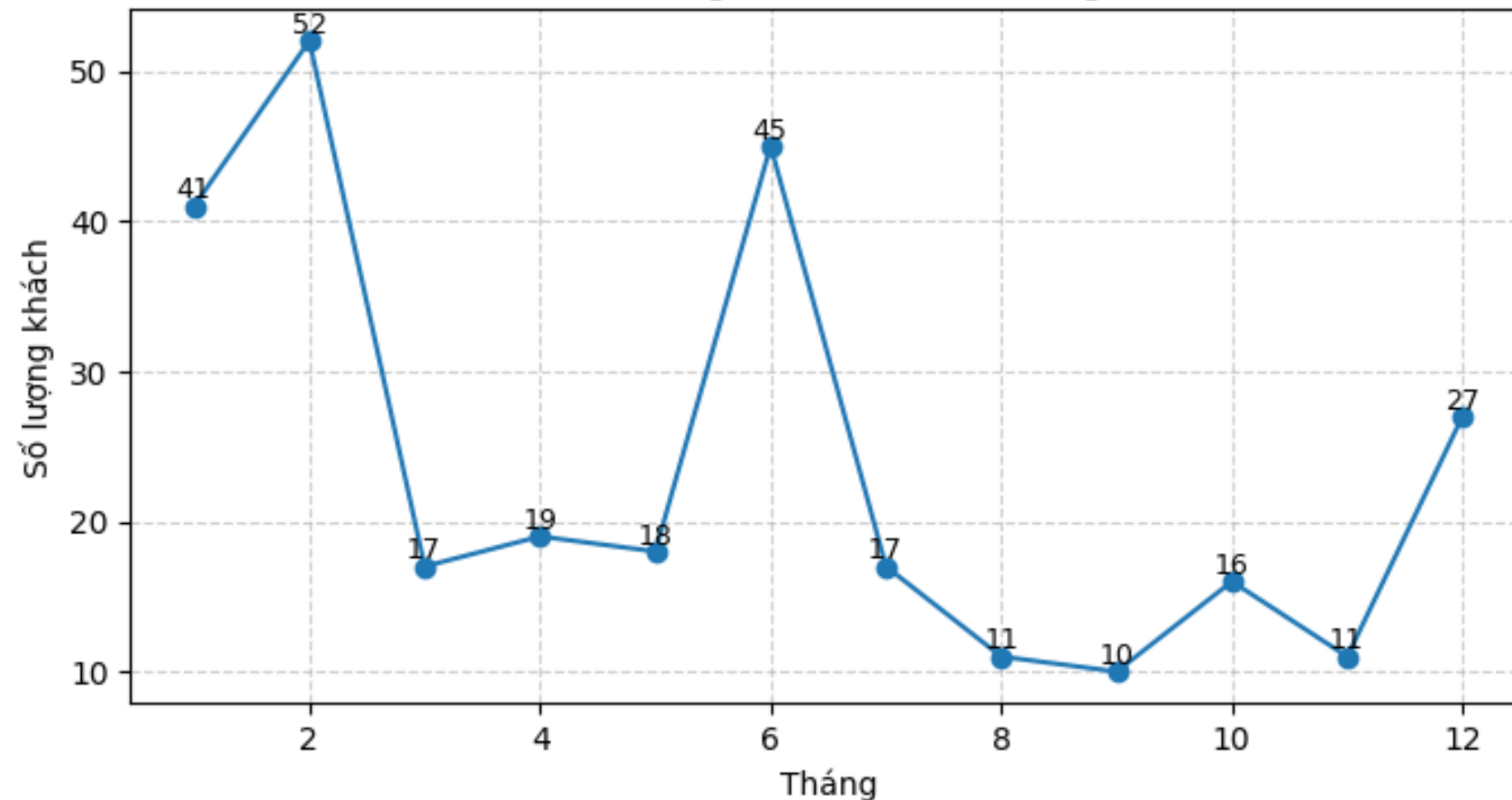
Hàm tìm theo ID hoặc key word cho chủ khách sạn, trả các biểu đồ thống kê cho khách sạn đó Quốc tịch, nhóm khách, xu hướng theo thời gian
VD: Theo keyword “Mường Thanh”



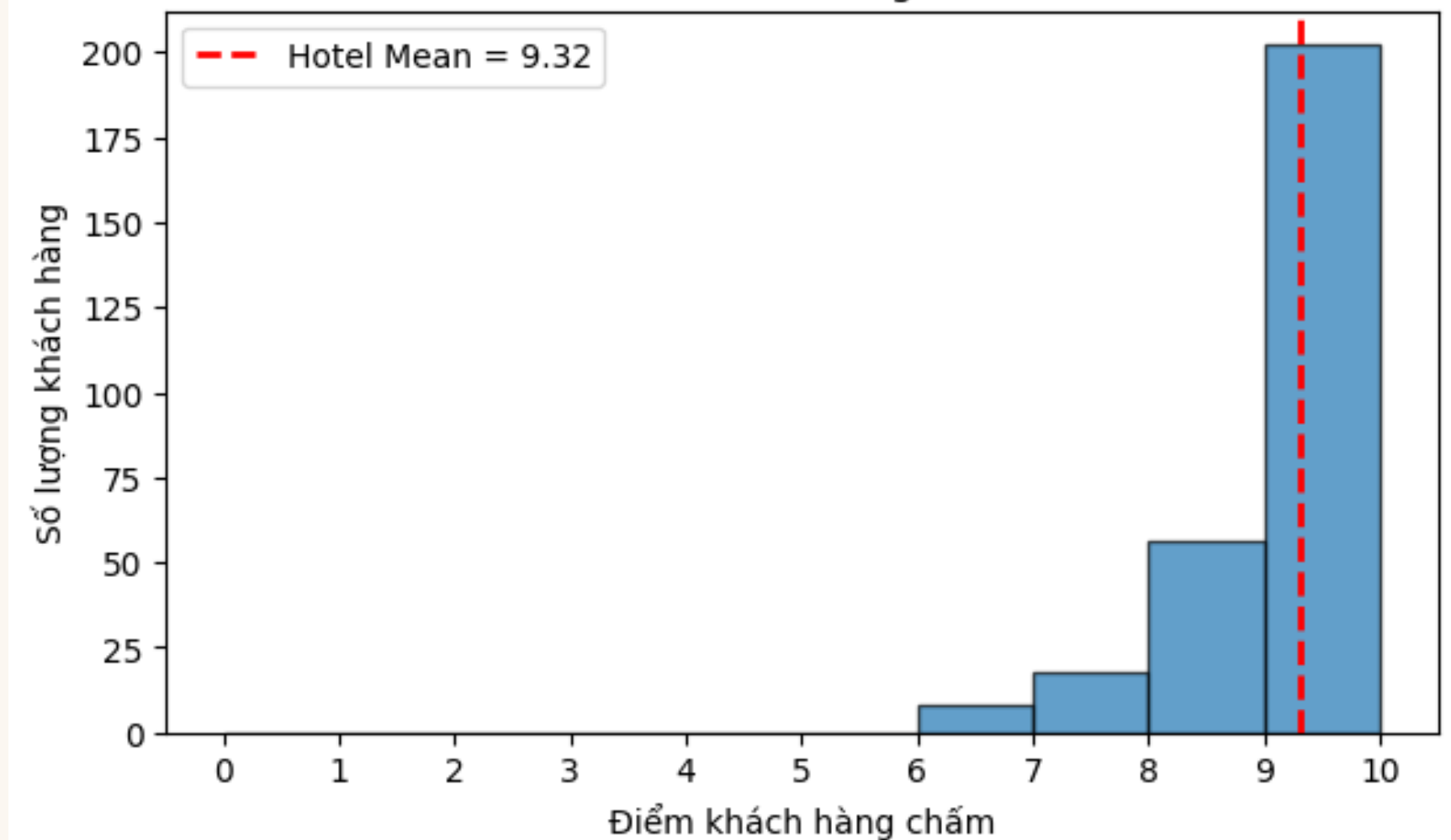
Insight Business

Hàm tìm theo ID hoặc key word cho chủ khách sạn, trả các biểu đồ thống kê cho khách sạn đó Quốc tịch, nhóm khách, xu hướng theo thời gian
VD: Theo keyword “Mường Thanh”

Xu hướng khách ở theo tháng



Phân bố điểm đánh giá (Score)



THANK
YOU!

