TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

KHOA CÔNG NGHỆ THÔNG TIN

TRỰC QUAN HÓA DỮ LIỆU – 21KHDL

BÁO CÁO ĐỒ ÁN

PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN VIỆC ĐẶT PHÒNG KHÁCH SẠN

DANH SÁCH THÀNH VIÊN

Họ và tên	MSSV	Mức độ đóng góp	Tỉ lệ thực hiện
Võ Duy Anh	21127221	100%	33%
Nguyễn Mậu Gia Bảo	21127583	100%	34%
Vũ Minh Phát	21127739	100%	33%

GIẢNG VIÊN HƯỚNG DẪN: Bùi Tiến Lên

Lê Ngọc Thành

Lê Nguyễn Nhựt Trường

Thành phố Hồ Chí Minh, ngày 27 tháng 04 năm 2024

Mục lục

1. Thông tin chung

1.1. Thông tin nhóm, mức độ đóng góp và tỉ lệ thực hiện của mỗi thành viên

Họ và tên	MSSV	Mức độ đóng góp	Tỉ lệ thực hiện
Võ Duy Anh	21127221	100%	33%
Nguyễn Mậu Gia Bảo	21127583	100%	34%
Vũ Minh Phát	21127739	100%	33%

1.2. Các câu hỏi chưa làm được

Cả nhóm đã hoàn thành toàn bộ công việc và trả lời mọi câu hỏi được giao.

2. Giới thiệu về bộ dữ liệu Hotel Booking

Bộ dữ liệu này chứa thông tin đặt phòng cho các khách sạn thành phố (city hotel) và khách sạn nghỉ dưỡng (resort hotel). Bộ dữ liệu bao gồm các thông tin như: thời gian đặt phòng, thời gian lưu trú, số lượng người lớn, trẻ em và/hoặc em bé, và số lượng chỗ đỗ xe có sẵn hoặc cần tăng lên, v.v..

Vì bộ dữ liệu gốc là dữ liệu khách sạn thực tế, của khách hàng thật, nên tất cả các yếu tố liên quan đến nhận dạng khách sạn hoặc khách hàng đều đã được loại bỏ khỏi dữ liệu.

Dữ liệu ban đầu được lấy từ bài viết "Hotel Booking Demand Datasets" của các tác giả Nuno Antonio, Ana Almeida và Luis Nunes vào tháng 2 năm 2019.

Hiện tại, ta có thể tìm thấy bộ dữ liệu này từ nhiều nguồn khác nhau, và một trong những nguồn phổ biến nhất là Kaggle.

3. Tìm hiểu về bộ dữ liệu gốc

3.1. Đếm số lượng dòng và số lượng cột của bộ dữ liệu gốc

- Bộ dữ liệu gốc có 119390 dòng và 32 cột.

3.2. Ý nghĩa của mỗi dòng trong bộ dữ liệu gốc

- Mỗi dòng trong bộ dữ liệu tương ứng với một lần đặt phòng khách sạn từ khách hàng (hay hành khách).
- Quan sát bảng dữ liệu, ta thấy có vẻ như không có dòng nào "lạc loài" (hay bất thường). Đây là một dấu hiệu rất tốt chứng tỏ dữ liệu của ta đủ chất lượng để tiến hành các bước phân tích tiếp theo.

3.3. Viết bảng mô tả về các cột trong bộ dữ liệu gốc

STT	Tên cột	Mô tả
1	hotel	Loại khách sạn, nhận một trong hai giá trị: "City Hotel" hoặc "Resort Hotel".
2	is_canceled	Giá trị nhị phân cho biết việc đặt phòng có bị hủy hay không ("0": nếu không hủy, "1": nếu đã hủy).
3	lead_time	Số ngày (chênh lệch) giữa ngày đặt phòng và ngày đến.
4	arrival_date_year	Năm của ngày đến.
5	arrival_date_month	Tháng của ngày đến.
6	arrival_date_week_number	Tuần trong năm của ngày đến.
7	arrival_date_day_of_month	Ngày trong tháng của ngày đến.
8	stays_in_weekend_nights	Số đêm cuối tuần (Thứ bảy hoặc Chủ nhật) mà khách lưu trú hoặc đặt phòng lưu trú tại khách sạn.
9	stays_in_week_nights	Số đêm trong tuần (Thứ Hai đến Thứ Sáu) mà khách lưu trú hoặc đặt phòng lưu trú tại khách sạn.
10	adults	Số lượng người lớn.
11	children	Số lượng trẻ em.
12	babies	Số lượng em bé (trẻ sơ sinh).
13	meal	Loại bữa ăn đã đặt. Các gói bữa ăn khách sạn tiêu chuẩn: - "Undefined"/"SC": Không có gói bữa ăn.

		- "BB" (Bed & Breakfast): Nhà nghỉ phục vụ bữa sáng. - "HB" (Half board): Bữa sáng và một bữa khác (thường là bữa tối). - "FB" (Full board): Bao ăn trọn gói (bữa sáng, bữa trưa và bữa tối).
14	country	Cho biết khách hàng đến từ quốc gia nào. Các giá trị được thể hiện ở định dạng ISO 3155–3:2013.
15	market_segment	Phân khúc thị trường của khách hàng.
16	distribution_channel	Phương thức hoặc kênh mà thông qua đó yêu cầu đặt phòng được thực hiện.
17	is_repeated_guest	Giá trị nhị phân cho biết khách hàng đã từng lưu trú trước đó hay chưa ("1": đã từng; "0": chưa từng).
18	previous_cancellations	Số lần đặt phòng trước đó đã bị khách hàng hủy trước lượt đặt phòng hiện tại.
19	previous_bookings_not_canceled	Số lần đặt phòng trước đó không bị khách hàng hủy trước lượt đặt phòng hiện tại.
20	reserved_room_type	Mã loại phòng mà khách đã đặt.
21	assigned_room_type	Mã loại phòng được chỉ định khi khách nhận phòng. Đôi khi loại phòng được chỉ định khác với loại phòng đã đặt vì lý do vận hành khách sạn (ví dụ: đặt trước quá nhiều) hoặc theo yêu cầu của khách hàng.

22	booking_changes	Số lượng thay đổi/sửa đổi được thực hiện đối với việc đặt phòng kể từ thời điểm đặt phòng cho đến thời điểm nhận hoặc hủy phòng.
23	deposit_type	Cho biết liệu khách hàng có đặt cọc để đảm bảo việc đặt phòng hay không. Biến này có thể nhận một trong ba giá trị: - "No Deposit": Không đặt cọc. - "Non Refund": Khoản tiền đặt cọc bằng tổng chi phí lưu trú. - "Refundable": Khoản tiền đặt cọc có giá trị ít hơn tổng chi phí lưu trú.
24	agent	ID (mã số) của đại lý lữ hành đã đặt phòng.
25	company	ID (mã số) của công ty đã đặt phòng hoặc chịu trách nhiệm thanh toán việc đặt phòng.
26	days_in_waiting_list	Số ngày mà lượt đặt phòng nằm trong danh sách chờ trước khi được xác nhận với khách hàng.
27	customer_type	Loại đặt phòng. Biến này nhận một trong bốn giá trị: "Contract", "Group", "Transient", "Transient-party".
28	adr	Giá trung bình hàng ngày (Average Daily Rate) được xác định bằng cách chia tổng tất cả các giao dịch lưu trú cho tổng số đêm lưu trú.

29	required_car_parking_spaces	Số lượng chỗ đậu xe theo yêu cầu của khách hàng.
30	total_of_special_requests	Số lượng yêu cầu đặc biệt của khách hàng (ví dụ: giường đôi hoặc giường tầng).
31	reservation_status	Trạng thái đặt phòng cuối cùng. Biến này nhận một trong ba giá trị: - "Canceled": Việc đặt phòng đã bị khách hàng hủy. - "Check-Out": Khách hàng đã nhận phòng và trả phòng. - "No-Show": Khách hàng không nhận phòng và đã thông báo cho khách sạn lý do.
32	reservation_status_date	Ngày cập nhật trạng thái đặt phòng cuối cùng. Biến này có thể được sử dụng cùng với "reservation_status" để hiểu khi nào việc đặt phòng bị hủy hoặc khi nào khách hàng trả phòng khách sạn.

Bảng 3.3: Bảng mô tả về các cột.

3.4. Mục tiêu cần đạt được trong bài tập này

Chúng ta sẽ thực hiện quy trình phân tích khám phá dữ liệu (EDA) lên bộ dữ liệu "Hotel Booking" để đưa ra kết luận hữu ích về xu hướng chung trong việc đặt phòng khách sạn và cách các yếu tố chi phối việc đặt phòng khách sạn tương tác với nhau. Thông qua đó, ta có thể đưa ra những đề xuất hợp lý giúp nâng cao hiệu suất hoạt động của khách sạn.

4. Tiền xử lý dữ liệu thô

4.1. Phân tích tỷ lệ trùng lặp (duplicate) và xử lý các dòng bị trùng lặp (nếu cần)

4.1.1. Phân tích tỷ lệ các dòng bị trùng lặp

Ta sử dụng phương thức "duplicated()" của đối tượng DataFrame để kiểm tra xem có dòng nào xuất hiện nhiều hơn một lần hay không.

Nhận xét:

- Từ kết quả phân tích dữ liệu, ta phát hiện bộ dữ liệu thô có khá nhiều dòng bị trùng lặp (31994 dòng), tương ứng với tỷ lệ hơn 26%.
- Như vậy, ta cần phải xử lý các dòng dữ liệu bị trùng lặp trước khi tiến hành các bước phân tích tiếp theo.

4.1.2. Xử lý các dòng bị trùng lặp

Ta sử dụng phương thức "drop_duplicates()" của đối tượng DataFrame để tiến hành loại bỏ các dòng bị trùng lặp.

Nhận xét:

- Bộ dữ liệu sau khi loại bỏ các dòng bị trùng lặp có số cột không đổi (32) và số dòng giảm xuống còn 87396 (= 119390 - 31994). Kết quả này hoàn toàn trùng khớp với mong đợi ban đầu của chúng ta.
- Như vậy, bước "Xử lý các dòng bị trùng lặp" ở trên đã đạt được các kết quả khá tốt. Điều này sẽ giúp cải thiện phần nào chất lượng của bộ dữ liệu và giúp ta đưa ra các kết luận chính xác hơn trong các bước phân tích tiếp theo.

4.2. Phân tích tỷ lệ thiếu giá trị ở mỗi cột (missing rate)

Ta sử dụng phương thức "isnull()" của đối tượng DataFrame để kiểm tra xem có ô nào trong bảng dữ liệu bị thiếu giá trị hay không. Từ đó ta suy ra tỷ lệ thiếu giá trị ở mỗi cột.

STT	Tên cột	Số lượng giá trị bị thiếu	Tỷ lệ thiếu giá trị (%)
1	company	82137	93.982562
2	agent	12193	13.951439
3	country	452	0.517186
4	children	4	0.004577
5	arrival_date_month	0	0.000000
6	arrival_date_week_number	0	0.000000
7	hotel	0	0.000000
8	is_canceled	0	0.000000
9	stays_in_weekend_nights	0	0.000000
10	arrival_date_day_of_month	0	0.000000

Bảng 4.2: Bảng mô tả tỷ lệ thiếu giá trị của 10 cột.

Nhận xét:

- Từ kết quả phân tích dữ liệu, ta thấy có bốn cột bị thiếu giá trị là: "company" (khoảng 94%), "agent" (khoảng 14%), "country" (khoảng 0.5%) và "children" (khoảng 0.005%).
- Trong khi tình trạng thiếu giá trị ở ba cột "agent", "country" và "children" chỉ dừng lại ở mức trung bình - nhẹ (tỷ lệ thiếu giá trị dưới 15%), thì cột "company" gặp phải tình trạng thiếu giá trị rất nghiêm trọng, với tỷ lệ thiếu giá trị hơn 90%. Điều này có thể gây ra rất nhiều khó khăn trong việc lựa chọn phương pháp điền giá trị thích hợp vào các ô bị thiếu.
- Như vậy, ta sẽ cần lựa chọn các phương pháp phù hợp để xử lý tình trạng thiếu giá trị trong bốn cột bên trên trước khi tiến hành các bước phân tích tiếp theo.

4.3. Phân tích kiểu dữ liệu của mỗi cột và xử lý các cột có kiểu dữ liệu chưa phù hợp (nếu cần)

Thay vì sử dụng thuộc tính "dtypes" của đối tượng DataFrame, ta có thể phân tích dữ liệu trong mỗi cột để xác định kiểu dữ liệu của cột tương ứng.

STT	Tên cột	Kiểu dữ liệu
1	hotel	str
2	is_canceled	int64
3	lead_time	int64
4	arrival_date_year	int64
5	arrival_date_month	str
6	arrival_date_week_number	int64
7	arrival_date_day_of_month	int64
8	stays_in_weekend_nights	int64
9	stays_in_week_nights	int64
10	adults	int64
11	children	float64
12	babies	int64
13	meal	str
14	country	str
15	market_segment	str
16	distribution_channel	str
17	is_repeated_guest	int64
18	previous_cancellations	int64
19	previous_bookings_not_canceled	int64
20	reserved_room_type	str
21	assigned_room_type	str
22	booking_changes	int64
23	deposit_type	str

24	agent	float64
25	company	float64
26	days_in_waiting_list	int64
27	customer_type	str
28	adr	float64
29	required_car_parking_spaces	int64
30	total_of_special_requests	int64
31	reservation_status	str
32	reservation_status_date	str

Bảng 4.3.1: Bảng mô tả kiểu dữ liệu cho mỗi cột.

Nhận xét:

- Cột "is_canceled" (hiện đang có kiểu dữ liệu dạng số "int64"): là biến cờ lệnh giúp ta nhận biết liệu việc đặt phòng có bị huỷ hay không. Do đó, ta cần chuyển cột này sang kiểu dữ liệu "str" để phân vào nhóm "categorical".
- Các cột "arrival_date_year", "arrival_date_week_number", "arrival_date_day_of_month" (hiện đang có kiểu dữ liệu dạng số "int64"): cho biết thông tin về thời gian mà khách hàng đến nhận phòng. Ta thấy "độ lớn" (magnitude) của các giá trị này không có ý nghĩa. Và các giá trị này thực sự đại diện cho một khoảng thời gian thay vì một con số. Do đó, ta cần chuyển các cột này sang kiểu dữ liệu "str" để phân vào nhóm "categorical".
- Cột "children" (hiện đang có kiểu dữ liệu dạng số "float64"): là một biến đếm cho biết số lượng trẻ em. Do đó, ta cần chuyển cột này sang kiểu dữ liệu "int64" để phù hợp hơn với ý nghĩa thực sự của cột này. Tuy nhiên, vì cột này đang bị thiếu giá trị nên ta sẽ thực hiện bước chuyển đổi kiểu dữ liệu sau khi đã điền giá trị thích hợp vào những vị trí bị thiếu.

- Cột "is_repeated_guest" (hiện đang có kiểu dữ liệu dạng số "int64"): là biến cờ lệnh cho biết khách hàng đã từng lưu trú ở khách sạn trước đó hay chưa. Do đó, ta cần chuyển cột này sang kiểu dữ liệu "str" để phân vào nhóm "categorical".
- Các cột "agent", "company" (hiện đang có kiểu dữ liệu dạng số "float64"): lần lượt lưu trữ mã định danh (ID) của đại lý lữ hành và công ty đã đặt phòng. Tuy nhiên, đây là hai cột đang bị thiếu giá trị. Do đó, ta sẽ tạm thời chuyển các cột này sang kiểu dữ liệu "np.object_" để tiến hành phân tích phân bố và xử lý các giá trị bị thiếu. Sau cùng, ta mới chuyển các cột này sang kiểu dữ liệu "str" để phân vào nhóm "categorical" (nếu cần).
- Cột "reservation_status_date" (hiện đang có kiểu dữ liệu "str"): cho biết ngày cập nhật trạng thái đặt phòng cuối cùng. Do đó, ta cần chuyển cột này sang kiểu dữ liệu "datetime".
- Các cột còn lại đều đã có kiểu dữ liệu phù hợp nên ta không cần chuyển đổi.

4.4. Phân tích phân bố của các giá trị trong mỗi cột "thực sự" có kiểu dữ liệu dạng số (numerical) và xử lý các cột bị thiếu giá trị hoặc các cột có giá trị bất thường

4.4.1. Phân tích phân bố của các giá trị trong mỗi cột "thực sự" có kiểu dữ liệu dạng số

Với mỗi cột "thực sự" có kiểu dữ liệu dạng số, ta sẽ tính:

- Tỷ lệ thiếu giá trị (từ 0 đến 100).
- Giá trị tối thiểu.
- Giá trị tứ phân vị thứ nhất.
- Giá trị tứ phân vị thứ hai (giá trị trung vị).
- Giá trị tứ phân vị thứ ba.
- Giá trị tối đa.

Chỉ số thống kê Tên cột	Tỷ lệ thiếu giá trị (%)	Giá trị tối thiểu	Tứ phân vị thứ nhất (Q1)	Giá trị trung vị (Q2)	Tứ phân vị thứ ba (Q3)	Giá trị tối đa
lead_time	0.000	0.00	11.0	49.0	125.0	737.0
stays_in_weekend_nights	0.000	0.00	0.0	1.0	2.0	19.0
stays_in_week_nights	0.000	0.00	1.0	2.0	4.0	50.0
adults	0.000	0.00	2.0	2.0	2.0	55.0
children	0.005	0.00	0.0	0.0	0.0	10.0
babies	0.000	0.00	0.0	0.0	0.0	10.0
previous_cancellations	0.000	0.00	0.0	0.0	0.0	26.0
previous_bookings_not_canceled	0.000	0.00	0.0	0.0	0.0	72.0
booking_changes	0.000	0.00	0.0	0.0	0.0	21.0
days_in_waiting_list	0.000	0.00	0.0	0.0	0.0	391.0
adr	0.000	-6.38	72.0	98.1	134.0	5400.0
required_car_parking_spaces	0.000	0.00	0.0	0.0	0.0	8.0
total_of_special_requests	0.000	0.00	0.0	0.0	1.0	5.0

Bảng 4.4.1: Bảng mô tả các giá trị thống kê cho biến định lượng (ban đầu).

Nhận xét:

- Trong bộ dữ liệu mà ta đang xem xét, nhóm thuộc tính số bao gồm 13 cột.
- Ta thấy chỉ có cột "children" bị thiếu giá trị với tỷ lệ rất nhỏ (khoảng 0.005%). Thông qua kết quả phân tích phân bố, ta phát hiện có hơn 75% giá trị trong cột này bằng "0". Tức là hầu hết hành khách đến khách sạn đều không dẫn theo trẻ em. Do đó, ta sẽ điền giá trị yếu vị (mode) vào những vị trí bị thiếu của cột này.
- Các cột còn lại không bị thiếu giá trị và có vẻ như cũng không có gì bất thường.

4.4.2. Xử lý các cột bị thiếu giá trị

Ta xử lý tình trạng thiếu giá trị của cột "children" bằng cách điền giá trị yếu vị (mode) của cột vào những vị trí bị thiếu. Sau đó, ta sẽ chuyển cột "children" sang kiểu dữ liệu dạng số "int64" để phù hợp với mô tả của cột này.

4.4.3. Kết quả của quá trình tiền xử lý các cột "thực sự" có kiểu dữ liệu dạng số

Chỉ số thống kê Tên cột	Tỷ lệ thiếu giá trị (%)	Giá trị tối thiểu	Tứ phân vị thứ nhất (Q1)	Giá trị trung vị (Q2)	Tứ phân vị thứ ba (Q3)	Giá trị tối đa
lead_time	0.0	0.00	11.0	49.0	125.0	737.0
stays_in_weekend_nights	0.0	0.00	0.0	1.0	2.0	19.0
stays_in_week_nights	0.0	0.00	1.0	2.0	4.0	50.0
adults	0.0	0.00	2.0	2.0	2.0	55.0
children	0.0	0.00	0.0	0.0	0.0	10.0
babies	0.0	0.00	0.0	0.0	0.0	10.0
previous_cancellations	0.0	0.00	0.0	0.0	0.0	26.0
previous_bookings_not_canceled	0.0	0.00	0.0	0.0	0.0	72.0
booking_changes	0.0	0.00	0.0	0.0	0.0	21.0
days_in_waiting_list	0.0	0.00	0.0	0.0	0.0	391.0
adr	0.0	-6.38	72.0	98.1	134.0	5400.0
required_car_parking_spaces	0.0	0.00	0.0	0.0	0.0	8.0
total_of_special_requests	0.0	0.00	0.0	0.0	1.0	5.0

Bảng 4.4.3: Bảng mô tả các giá trị thống kê cho biến định lượng (sau khi tiền xử lý).

Nhận xét:

- Sau quá trình tiền xử lý dữ liệu, quan sát bảng mô tả cho các cột có kiểu dữ liệu dạng số, ta thấy không có cột nào bị thiếu giá trị và có vẻ như dữ liệu cũng không có gì bất thường.

4.5. Phân tích phân bố của các giá trị trong mỗi cột "thực sự" có kiểu dữ liệu dạng phân loại (categorical) và xử lý các cột bị thiếu giá trị hoặc các cột có giá trị bất thường

4.5.1. Phân tích phân bố của các giá trị trong mỗi cột "thực sự" có kiểu dữ liệu dạng phân loại

Với mỗi cột "thực sự" có kiểu dữ liệu dạng phân loại, ta sẽ tính:

- Tỷ lệ thiếu giá trị (từ 0 đến 100).
- Số lượng các giá trị khác nhau.
- Tỷ lệ xuất hiện (từ 0 đến 100) của mỗi giá trị.

Tên cột	Tỷ lệ thiếu giá trị (%)	Số lượng giá trị khác nhau	Tỷ lệ xuất hiện (%) của mỗi giá trị
hotel	0.0	2	{'City Hotel': 61.1, 'Resort Hotel': 38.9}
is_canceled	0.0	2	{'0': 72.5, '1': 27.5}
arrival_date_year	0.0	3	{'2016': 48.5, '2017': 36.3, '2015': 15.2}
arrival_date_month	0.0	12	{'August': 12.9, 'July': 11.5, 'May': 9.6, 'Ap...
arrival_date_week_number	0.0	53	{'33': 3.2, '34': 2.9, '32': 2.8, '28': 2.7, '...
arrival_date_day_of_month	0.0	31	{'17': 3.5, '2': 3.5, '26': 3.4, '5': 3.4, '16...
meal	0.0	5	{'BB': 77.8, 'SC': 10.8, 'HB': 10.4, 'Undefine...
country	0.5	177	{'PRT': 31.6, 'GBR': 12.0, 'FRA': 10.2, 'ESP':...
market_segment	0.0	8	{'Online TA': 59.1, 'Offline TA/TO': 15.9, 'Di...
distribution_channel	0.0	5	{'TA/TO': 79.1, 'Direct': 14.9, 'Corporate': 5...
is_repeated_guest	0.0	2	{'0': 96.1, '1': 3.9}
reserved_room_type	0.0	10	{'A': 64.7, 'D': 19.9, 'E': 6.9, 'F': 3.2, 'G'...
assigned_room_type	0.0	12	{'A': 53.0, 'D': 25.7, 'E': 8.2, 'F': 4.2, 'G'...
deposit_type	0.0	3	{'No Deposit': 98.7, 'Non Refund': 1.2, 'Refun...
agent	14.0	333	{'9.0': 38.2, '240.0': 17.3, '14.0': 4.5, '7.0...
company	94.0	352	{'40.0': 16.2, '223.0': 9.6, '45.0': 4.5, '153...
customer_type	0.0	4	{'Transient': 82.4, 'Transient-Party': 13.4, '...
reservation_status	0.0	3	{'Check-Out': 72.5, 'Canceled': 26.3, 'No-Show...

Bảng 4.5.1: Bảng mô tả các giá trị thống kê cho biến định tính (ban đầu).

Nhận xét:

- Trong bộ dữ liệu mà ta đang xem xét, nhóm thuộc tính phân loại bao gồm 18 cột.
- Cột "company" có tỷ lệ thiếu giá trị rất cao (hơn 90%). Với tình trạng thiếu dữ liệu nghiêm trọng như thế này thì có vẻ như "company" không phải là một thuộc tính đủ chất lượng để ta có thể phân tích và đưa ra các kết luận có độ tin cậy cao. Do đó, ta sẽ loại bỏ cột này khỏi bộ dữ liệu.
- Cột "agent" có tỷ lệ thiếu giá trị ở mức trung bình (khoảng 14%), không quá nghiêm trọng. Do đó, ta sẽ chọn phương pháp điền giá trị "-1" vào những vị trí bị thiếu. Vì "-1" là giá trị không xuất hiện trong bộ dữ liệu, việc dùng giá trị này vừa giúp ta dễ dàng nhận biết những vị trí bị thiếu dữ liệu vừa tránh làm ảnh hưởng đến phân bố của các giá trị còn lại.
- Cột "country" có tỷ lệ thiếu giá trị rất thấp, chỉ khoảng 0.5% (không đáng kể). Do đó, ta sẽ tiền xử lý cột "country" bằng phương pháp giống với khi tiền xử lý cột "agent" nhưng thay giá trị "-1" bằng giá trị "others" để phù hợp với ngữ cảnh của dữ liệu.
- Các cột còn lại không bị thiếu giá trị và có vẻ như cũng không có gì bất thường.

4.5.2. Xử lý các cột bị thiếu giá trị

- Do cột "company" có tình trạng thiếu giá trị quá nghiêm trọng, nên ta tiến hành loại bỏ cột này khỏi bảng dữ liệu.
- Với cột "agent", ta điền "-1" vào những vị trí bị thiếu giá trị. Sau đó, ta chuyển cột này sang kiểu dữ liệu "str" để phân vào nhóm "categorical".
- Với cột "country", ta điền "others" vào những vị trí bị thiếu giá trị.

4.5.3. Kết quả của quá trình tiền xử lý các cột "thực sự" có kiểu dữ liệu dạng phân loại

Tên cột	Tỷ lệ thiếu giá trị (%)	Số lượng giá trị khác nhau	Tỷ lệ xuất hiện (%) của mỗi giá trị
hotel	0.0	2	{'City Hotel': 61.1, 'Resort Hotel': 38.9}
is_canceled	0.0	2	{'0': 72.5, '1': 27.5}
arrival_date_year	0.0	3	{'2016': 48.5, '2017': 36.3, '2015': 15.2}
arrival_date_month	0.0	12	{'August': 12.9, 'July': 11.5, 'May': 9.6, 'Ap...
arrival_date_week_number	0.0	53	{'33': 3.2, '34': 2.9, '32': 2.8, '28': 2.7, '...
arrival_date_day_of_month	0.0	31	{'17': 3.5, '2': 3.5, '26': 3.4, '5': 3.4, '16...
meal	0.0	5	{'BB': 77.8, 'SC': 10.8, 'HB': 10.4, 'Undefine...
country	0.0	178	{'PRT': 31.4, 'GBR': 11.9, 'FRA': 10.1, 'ESP':...
market_segment	0.0	8	{'Online TA': 59.1, 'Offline TA/TO': 15.9, 'Di...
distribution_channel	0.0	5	{'TA/TO': 79.1, 'Direct': 14.9, 'Corporate': 5...
is_repeated_guest	0.0	2	{'0': 96.1, '1': 3.9}
reserved_room_type	0.0	10	{'A': 64.7, 'D': 19.9, 'E': 6.9, 'F': 3.2, 'G'...
assigned_room_type	0.0	12	{'A': 53.0, 'D': 25.7, 'E': 8.2, 'F': 4.2, 'G'...
deposit_type	0.0	3	{'No Deposit': 98.7, 'Non Refund': 1.2, 'Refun...
agent	0.0	334	{'9.0': 32.9, '240.0': 14.9, '-1': 14.0, '14.0...
customer_type	0.0	4	{'Transient': 82.4, 'Transient-Party': 13.4, '...
reservation_status	0.0	3	{'Check-Out': 72.5, 'Canceled': 26.3, 'No-Show...

Bảng 4.5.3: Bảng mô tả các giá trị thống kê cho biến định lượng (sau khi tiền xử lý).

Nhận xét:

- Sau quá trình tiền xử lý dữ liệu, số lượng thuộc tính phân loại đã giảm từ 18 xuống còn 17 (do ta loại bỏ cột "company"). Kết quả này hoàn toàn trùng khớp với dự đoán ban đầu của chúng ta.
- Quan sát bảng mô tả cho các cột có kiểu dữ liệu dạng phân loại, ta thấy không có cột nào bị thiếu giá trị và có vẻ như dữ liệu cũng không có gì bất thường.

4.6. Tổng kết quá trình tiền xử lý dữ liệu

Sau khi thực hiện rất nhiều bước phân tích và tiền xử lý trên bộ dữ liệu gốc, ta sẽ sử dụng phương thức "info()" của đối tượng DataFrame để tạo ra một bảng mô tả các thông tin tổng quát nhất về bộ dữ liệu (đã tiền xử lý) được dùng trong các bước phân tích tiếp theo.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 87396 entries, 0 to 87395
Data columns (total 31 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   hotel                           87396 non-null  object
 1   is_canceled                     87396 non-null  object
 2   lead_time                       87396 non-null  int64
 3   arrival_date_year               87396 non-null  object
 4   arrival_date_month              87396 non-null  object
 5   arrival_date_week_number        87396 non-null  object
 6   arrival_date_day_of_month       87396 non-null  object
 7   stays_in_weekend_nights         87396 non-null  int64
 8   stays_in_week_nights            87396 non-null  int64
 9   adults                          87396 non-null  int64
 10  children                        87396 non-null  int64
 11  babies                          87396 non-null  int64
 12  meal                            87396 non-null  object
 13  country                         87396 non-null  object
 14  market_segment                  87396 non-null  object
 15  distribution_channel            87396 non-null  object
 16  is_repeated_guest               87396 non-null  object
 17  previous_cancellations          87396 non-null  int64
 18  previous_bookings_not_canceled  87396 non-null  int64
 19  reserved_room_type              87396 non-null  object
 20  assigned_room_type              87396 non-null  object
 21  booking_changes                 87396 non-null  int64
 22  deposit_type                    87396 non-null  object
 23  agent                           87396 non-null  object
 24  days_in_waiting_list            87396 non-null  int64
 25  customer_type                   87396 non-null  object
 26  adr                             87396 non-null  float64
 27  required_car_parking_spaces     87396 non-null  int64
 28  total_of_special_requests       87396 non-null  int64
 29  reservation_status              87396 non-null  object
 30  reservation_status_date         87396 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(1), int64(12), object(17)
memory usage: 20.7+ MB
```

Bảng 4.6: Bảng mô tả dữ liệu sau quá trình tiền xử lý.

5. EDA 1D

5.1. Chia loại dữ liệu theo kiểu "numerical" và "categorical"

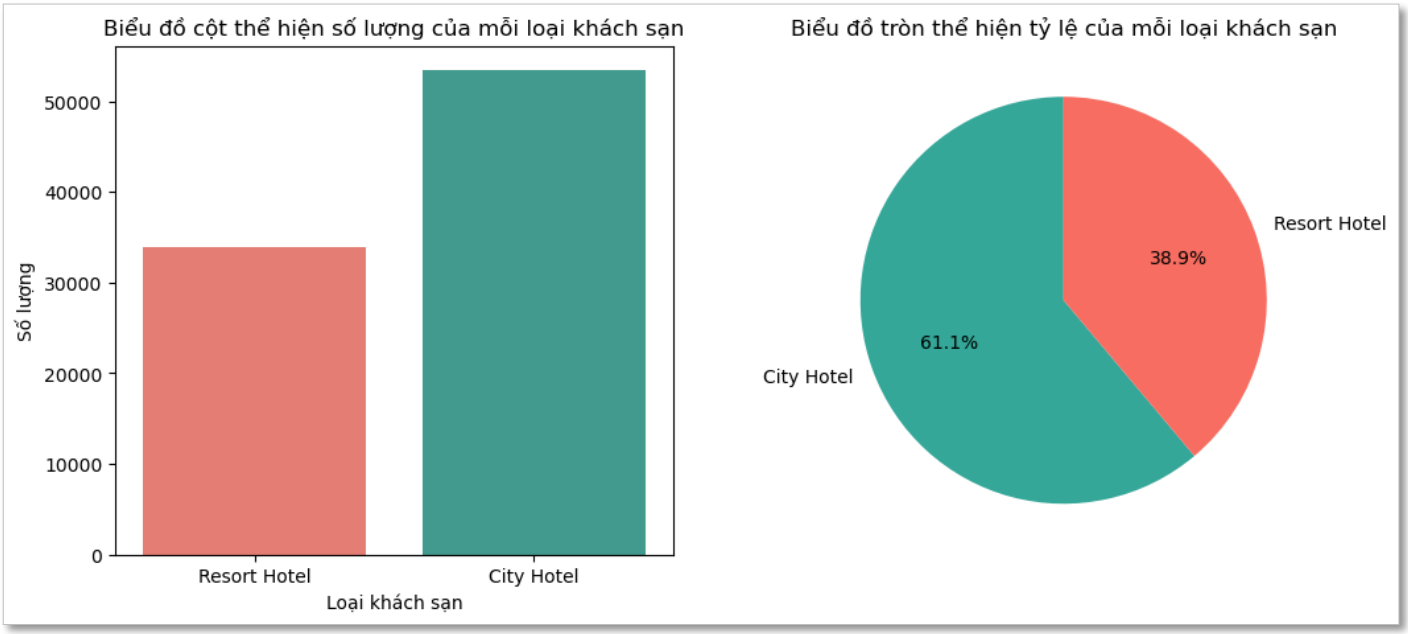
Ta sử dụng phương thức "select_dtypes()" của đối tượng DataFrame với các tham số "include" thích hợp để tiến hành chia dữ liệu theo kiểu "numerical" và "categorical".

Nhận xét:

- Ta tạo thành hai DataFrame mới là "num_col_df" và "cat_col_df". Với:
  - "num_col_df" là DataFrame chứa các cột chỉ có kiểu dữ liệu dạng số (biến định lượng).
  - "cat_col_df" là DataFrame chứa các cột chỉ có kiểu dữ liệu dạng phân loại (biến định tính).

5.2. Phân tích tỷ lệ cho các biến định tính (categorical)

5.2.1. Phân tích tỷ lệ đối với thuộc tính "hotel"



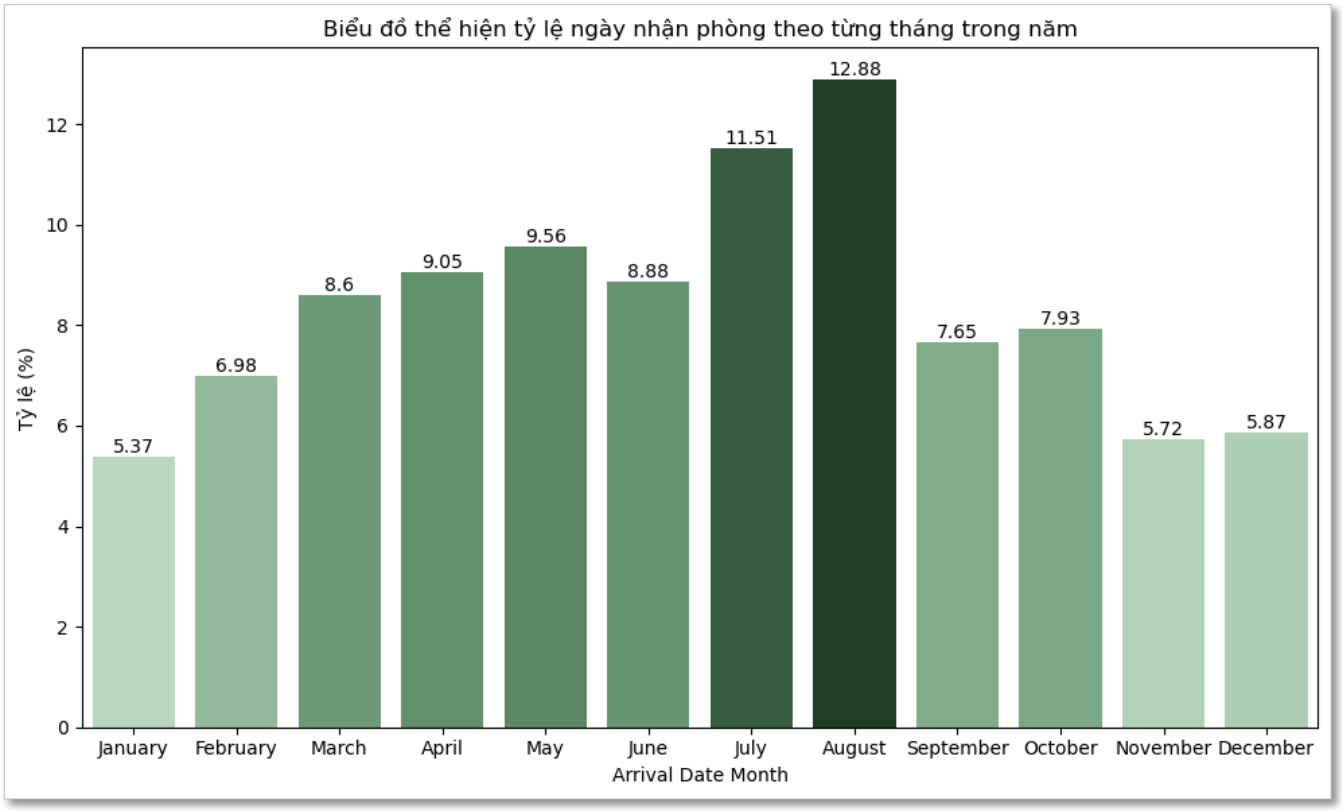
Biểu đồ 5.2.1: Biểu đồ thể hiện phân bố của thuộc tính "hotel".

Nhận xét:

- Từ kết quả phân tích dữ liệu, ta thấy "City Hotel" chiếm tỷ lệ lớn hơn khá nhiều so với "Resort Hotel" (61.1% của "City Hotel" so với 38.9% của "Resort Hotel"). Điều này cho thấy "City Hotel" có thể có độ phổ biến cao hơn "Resort Hotel" trong ngành khách sạn hoặc ít nhất là trong bộ dữ liệu này. Tuy nhiên, để có thể đưa ra kết luận có độ tin cậy cao hơn về xu hướng chung trong việc lựa chọn loại khách sạn thì ta cần phải xem xét thêm các yếu tố khác cũng ảnh hưởng đến một chuyến đi, như: số lượng trẻ em đi cùng hoặc thời điểm đi du lịch, v.v..
- Các "City Hotel" thường nhận được nhiều lượt đặt phòng hơn có thể là do chúng nằm ở các khu vực đô thị, nơi có sự tiện lợi trong việc truy cập các điểm du lịch, kinh doanh và mua sắm. Trong khi đó, "Resort Hotel" thường được đặt ở những khu vực biển hoặc khu nghỉ dưỡng, đây là điểm đến lý tưởng cho các khách hàng muốn tận hưởng một kỳ nghỉ thư giãn và tiện nghi sau những tháng ngày làm việc vất vả. Như vậy, nhu cầu và mục tiêu của mỗi chuyến đi du lịch cũng có ảnh hưởng phần nào đến việc chọn lựa loại khách sạn.
- Bên cạnh đó, các "Resort Hotel" thường có chi phí đắt đỏ hơn khá nhiều so với "City Hotel". Đây là yếu tố có ảnh hưởng rất lớn đến quyết định lựa chọn loại khách sạn của khách hàng, đặc biệt là đối với những khách hàng có nguồn kinh tế không quá dư dả. Điều này có thể phản ánh một xu hướng khá thú vị: những khách du lịch có thể lựa chọn lưu trú tại "City Hotel" (với chi phí rẻ hơn "Resort Hotel") và dành phần tiền dư ra cho việc ăn uống, mua sắm quà lưu niệm, v.v.. Tuy nhiên, để có thể đưa ra kết luận có độ tin cậy cao hơn về thói quen của khách hàng thì ta cần thực hiện thêm nhiều phân tích chuyên sâu hơn.

5.2.2. Phân tích tỷ lệ đối với thuộc tính "arrival_date_month"

Đầu tiên, ta sẽ tính tỷ lệ phần trăm của mỗi giá trị có trong cột "arrival_date_month". Sau đó, ta dùng biểu đồ cột (đứng) để thể hiện phân bố của ngày nhận phòng theo từng tháng trong năm.



Biểu đồ 5.2.2: Biểu đồ thể hiện phân bố của thuộc tính "arrival_date_month".

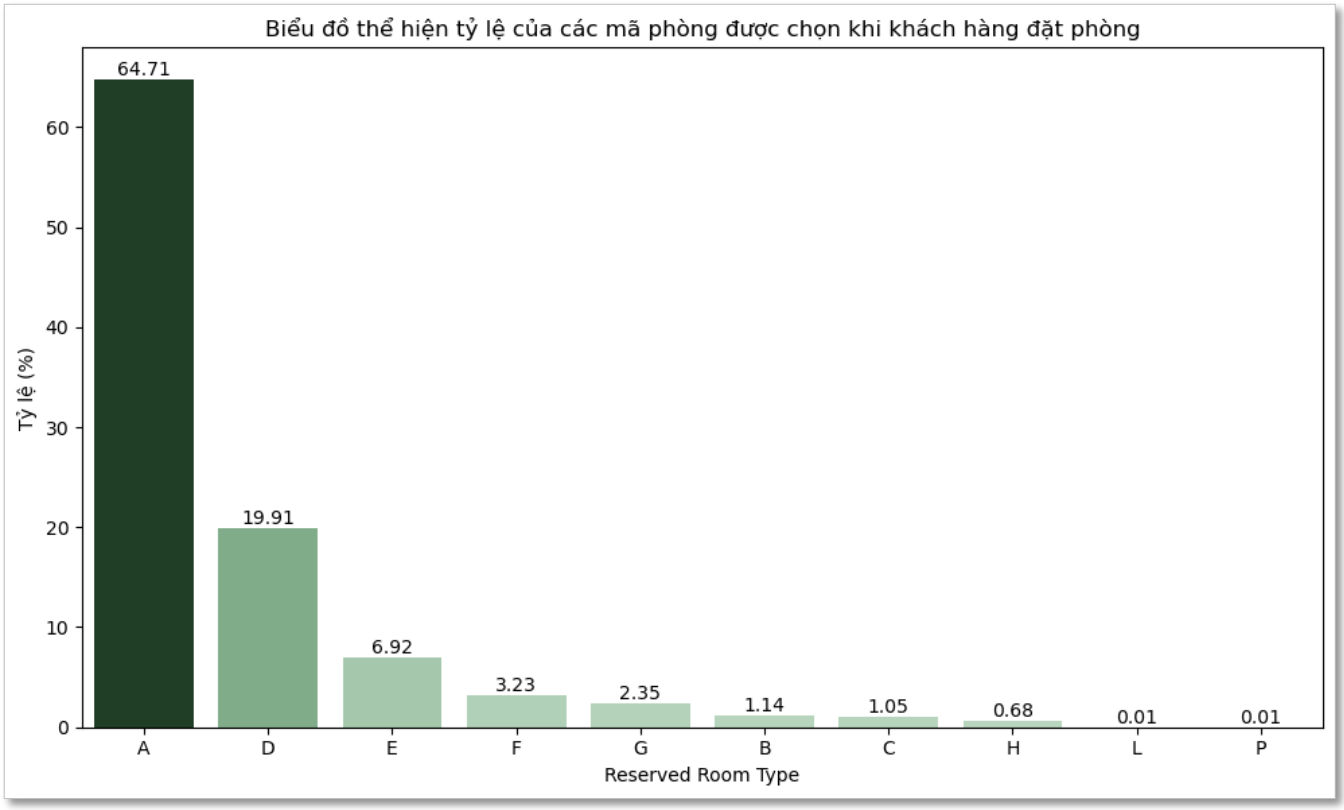
Nhận xét:

- Từ kết quả phân tích dữ liệu, ta thấy có sự phân bố không đồng đều về tỷ lệ ngày đến nhận phòng theo các tháng trong năm. "August" (tháng 8) là tháng có tỷ lệ khách hàng đến nhận phòng cao nhất với khoảng 12.88%, tiếp theo là "July" (tháng 7) với tỷ lệ khoảng 11.51%, và "May" (tháng 5) với tỷ lệ khoảng 9.56%.
- Có rất nhiều yếu tố có thể ảnh hưởng đến phân bố của ngày nhận phòng như: thời tiết, mùa du lịch trong năm, khoảng thời gian nghỉ lễ, v.v.. Nhìn chung, các tháng trùng

với các kỳ nghỉ thường có tỷ lệ khách đến nhận phòng cao hơn các tháng khác do nhiều người có nhu cầu đi du lịch vào khoảng thời gian này.

- Khoảng thời gian từ tháng 7 đến tháng 8 thường là thời gian cao điểm của các chuyến đi du lịch trong mùa hè. Vì đây là lúc mà nhiều học sinh, sinh viên có kỳ nghỉ hè hoặc có kỳ nghỉ lễ dài, v.v..
- Bên cạnh đó, tháng 5 cũng thường là thời điểm ngành du lịch bắt đầu bận rộn. Do khoảng thời gian này thường có thời tiết ấm áp và có các sự kiện văn hóa, lễ hội thú vị ở nhiều địa điểm.
- Ngược lại, khoảng thời gian từ tháng 11 năm trước đến tháng 1 năm sau thường chứng kiến số lượng người đi du lịch ít hơn đáng kể so với các khoảng thời gian khác trong năm. Điều này có thể lý giải phần nào bởi thời tiết trong khoảng thời gian này thường rất lạnh, tuyết phủ kín ở nhiều nơi gây ra khó khăn trong việc đi lại nên mọi người thường không thích đi du lịch vào thời gian này. Bên cạnh đó, cuối năm cũng là khoảng thời gian mà tất cả mọi người chuẩn bị cho dịp lễ Giáng sinh. Đây là một dịp lễ lớn, rất quan trọng (đặc biệt là với các quốc gia phương Tây) nên mọi người thường chuẩn bị cho dịp lễ từ rất sớm (từ cuối tháng 11) để có thể tổ chức các bữa tiệc và quây quần bên người thân, bạn bè.
- Như vậy, các tháng vào mùa hè (từ tháng 6 đến tháng 8) là thời điểm có rất nhiều người đi du lịch nên số lượng khách đến nhận phòng khách sạn là rất lớn (thường là nhiều nhất trong năm). Ngược lại, các tháng vào mùa đông (từ tháng 11 năm trước đến tháng 1 năm sau) có thể xem là thời điểm "đóng băng" của ngành du lịch khi số lượng người đi du lịch là rất ít (thường là ít nhất trong năm). Điều này có thể phản ánh phần nào xu hướng tổng quát về nhu cầu đi du lịch của mọi người theo từng mùa trong năm. Việc đưa ra các chính sách khuyến mãi, kích cầu du lịch vào mùa hè sẽ là một trong các chiến lược mà các khách sạn cần ưu tiên nghiên cứu để có thể thu hút nhiều khách hàng sử dụng dịch vụ của mình và thu được lợi nhuận cao hơn.

5.2.3. Phân tích tỷ lệ đối với thuộc tính "reserved_room_type"



Biểu đồ 5.2.3: Biểu đồ thể hiện phân bố của thuộc tính "reserved_room_type".

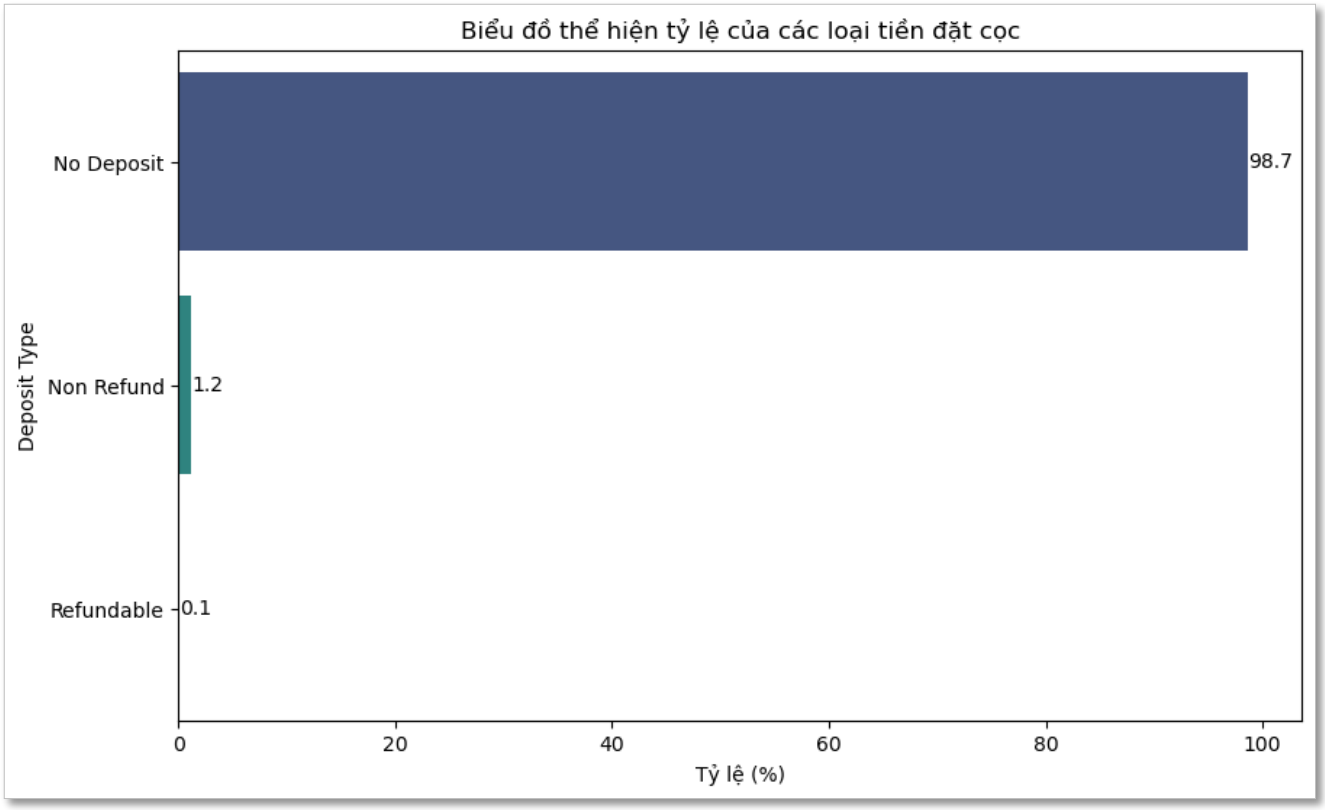
Nhận xét:

- Từ kết quả phân tích dữ liệu, ta thấy có sự phân bố không đồng đều trong các mã phòng được khách hàng lựa chọn khi tiến hành đặt phòng khách sạn. Nhìn chung, loại phòng được khách hàng yêu cầu nhiều nhất là loại "A" với tỷ lệ khoảng 65%, kế đến là các phòng thuộc loại "D" với tỷ lệ khoảng 20% và loại "E" với tỷ lệ khoảng 7%. Các loại phòng còn lại thường có tỷ lệ được lựa chọn rất thấp nên ta không cần phân tích chi tiết.
- Việc các phòng thuộc loại "A" thường được khách hàng ưa chuộng có thể đến từ việc giá phòng rẻ hơn so với các mã phòng khác nhưng vẫn đáp ứng được nhu cầu nghỉ dưỡng cơ bản cho khách hàng. Nếu giả thuyết này là đúng thì ta có thể suy ra rằng: phòng loại "A" đem lại lợi ích kinh tế tốt nhất cho khách hàng, sau đó là đến phòng

loại "D" và "E". Tuy nhiên, đây chỉ là giả thuyết được đặt ra để cố gắng giải thích cho xu hướng lựa chọn của khách hàng chứ ta không có một cơ sở vững chắc nào để chứng minh điều này là đúng. Do đó, ta cần được cung cấp thêm các thông tin mô tả về mỗi loại phòng để có thể đưa ra các kết luận chính xác hơn, phù hợp với ngữ cảnh thực tế hơn.

- Như vậy, dựa trên nhu cầu thị trường, việc gia tăng số lượng phòng thuộc loại "A", "D" và "E" tại các khách sạn là giải pháp hiệu quả để đáp ứng nhu cầu đa dạng của du khách, đồng thời tối ưu hóa doanh thu, đặc biệt trong các mùa cao điểm du lịch. Còn đối với các loại phòng có tỷ lệ lựa chọn thấp như "H", "L" và "P", ta sẽ cần kết hợp thuộc tính đang phân tích với một số thuộc tính khác (như: doanh thu trung bình, v.v.) để có thể lý giải vì sao các mã phòng này thường ít được lựa chọn. Khi đó, ta mới có đủ cơ sở để đưa ra các đề xuất tiếp theo giúp cải thiện hiệu suất hoạt động của khách sạn.

5.2.4. Phân tích tỷ lệ đối với thuộc tính "deposit_type"



Biểu đồ 5.2.4: Biểu đồ thể hiện phân bố của thuộc tính "deposit_type".

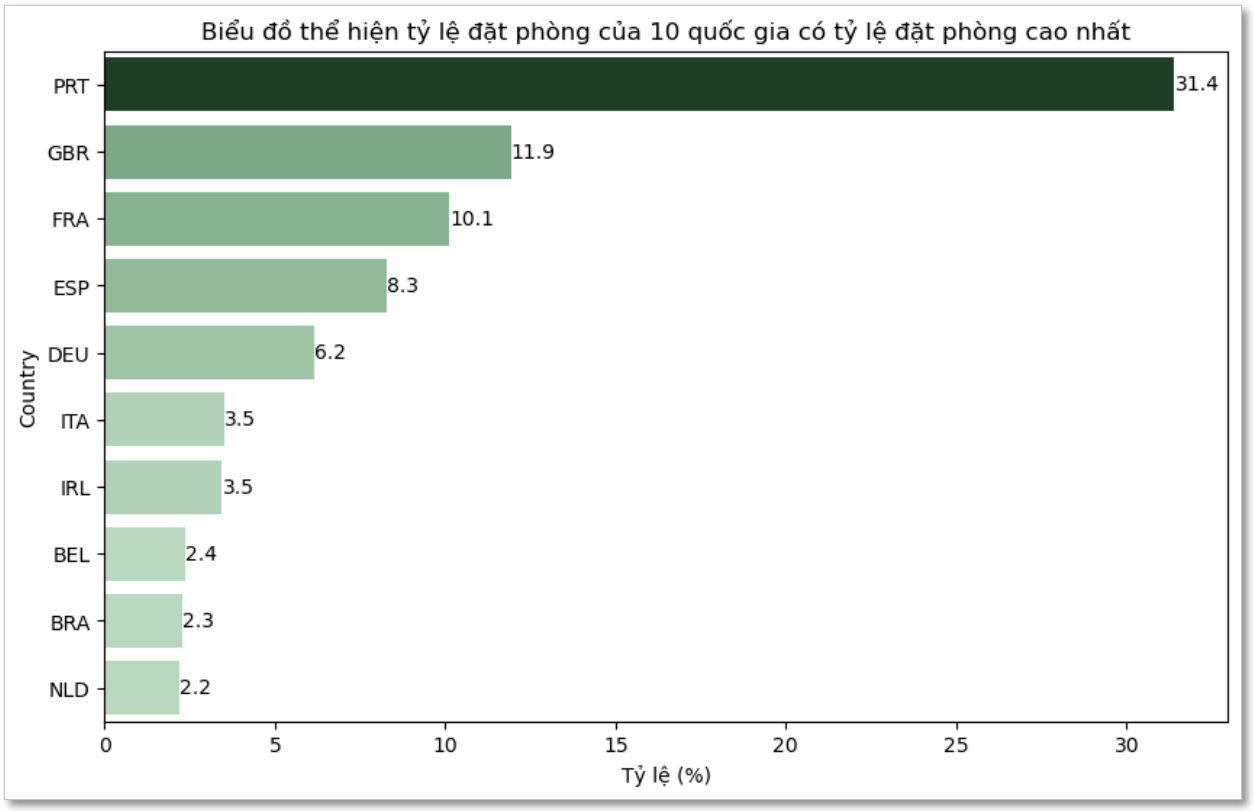
Nhận xét:

- Từ kết quả phân tích dữ liệu, ta thấy rằng hầu hết khách hàng đều không đặt cọc khi đặt phòng khách sạn (No Deposit). Xu hướng này có thể xuất phát từ việc khách hàng sợ mất tiền đặt cọc trong trường hợp xuất hiện một việc bận đột xuất buộc họ phải hủy chuyến đi của mình. Điều này cũng bình thường và không có gì khó hiểu.
- Ta thấy chỉ có một số ít khách hàng lựa chọn đặt cọc để đảm bảo không gặp phải tình trạng hết phòng vào các mùa cao điểm du lịch. Nhưng điều thú vị nhất nằm ở chỗ: nếu đã quyết định đặt cọc, thì hầu hết khách hàng sẽ lựa chọn trả trước toàn bộ chi phí lưu trú trong chuyến đi (Non Refund) và có rất ít khách hàng lựa chọn phương án trả trước một phần chi phí lưu trú (Refundable). Điều này cho thấy các khách hàng đã thực hiện đặt cọc là những người "rất quyết tâm đi du lịch" và thường không hủy đặt

phòng khách sạn. Họ sẽ sẵn sàng thanh toán toàn bộ chi phí để đảm bảo có chỗ ở. Đây chính là nhóm khách hàng mà ta cần dành nhiều sự quan tâm để có thể tạo ra nguồn doanh thu tốt hơn cho khách sạn.

- Như vậy, ta có thể tạo ra chính sách ưu đãi dành cho các khách hàng đặt cọc trước, đặc biệt là các khách hàng lựa chọn phương án trả trước toàn bộ chi phí lưu trú (Non Refund). Vì đây là nhóm khách hàng thường có mức độ cam kết cao, nên việc triển khai các chương trình ưu đãi như: giảm giá, tặng quà, nâng hạng phòng, v.v. sẽ khuyến khích khách hàng đặt cọc, tăng tỷ lệ đặt phòng và doanh thu cho khách sạn. Mặt khác, khi khách hàng đã đặt cọc trước thì rủi ro mà họ hủy đặt phòng sẽ giảm đi đáng kể. Đây sẽ là một mối quan hệ có lợi cho cả bên cung cấp dịch vụ (khách sạn) và bên sử dụng dịch vụ (khách hàng). Do đó, việc nghiên cứu và phát triển các chính sách ưu đãi cho khách hàng đặt cọc trước sẽ là một công việc cần được các người chủ khách sạn đầu tư nhiều hơn. Việc có được các chính sách ưu đãi hấp dẫn sẽ có thể thu hút nhiều du khách hơn và có tiềm năng đem lại lợi nhuận khổng lồ cho khách sạn.

5.2.5. Phân tích tỷ lệ đối với thuộc tính "country"



Biểu đồ 5.2.5: Biểu đồ thể hiện phân bố của thuộc tính "country".

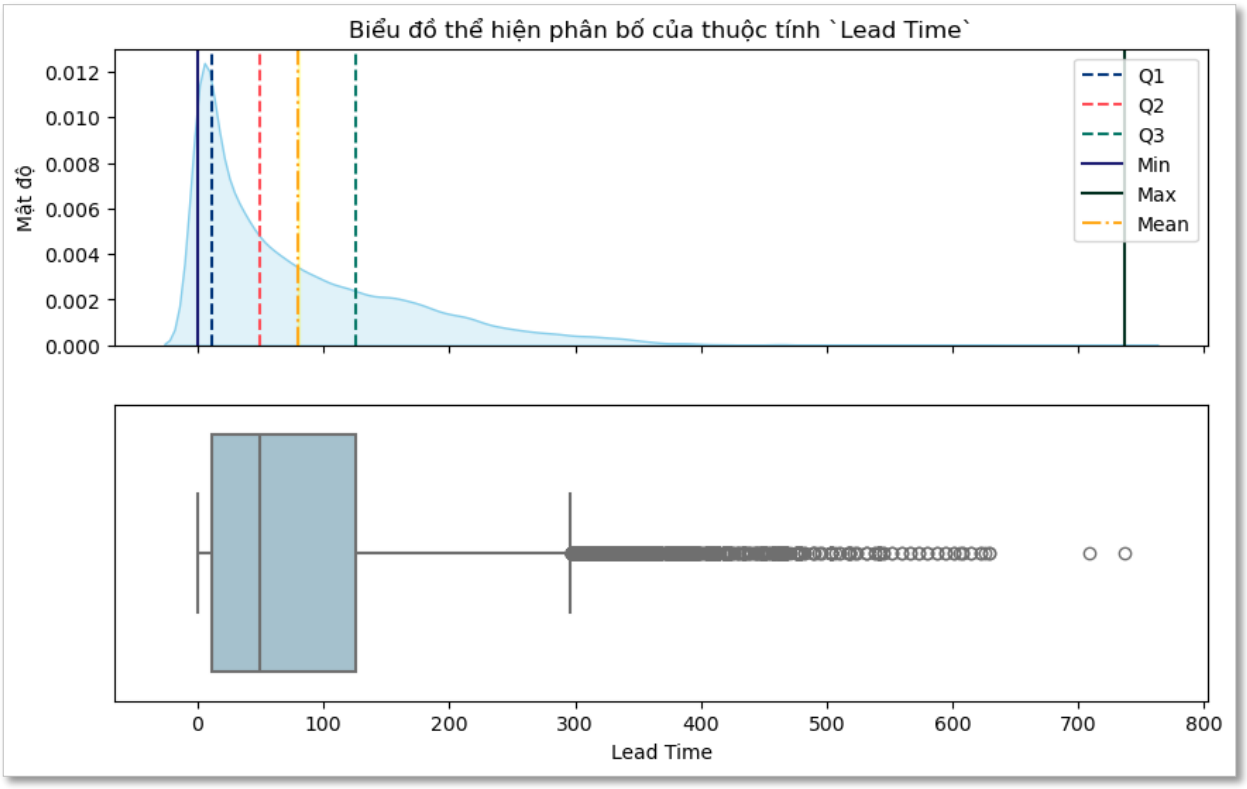
Nhận xét:

- Từ kết quả phân tích dữ liệu, ta thấy các khách hàng đến từ đất nước "Portugal" (Bồ Đào Nha) có tỷ lệ đặt phòng cao nhất với tỷ lệ hơn 30%. Xếp ngay sau đó là các khách hàng đến từ "Great Britain" (Đảo Anh) với tỷ lệ khoảng 12% và ở trí thứ ba là đất nước "France" (Pháp) với tỷ lệ hơn 10%. Nếu tiếp tục quan sát tên của các quốc gia xuất hiện trong biểu đồ, ta phát hiện bộ dữ liệu này chủ yếu được thu thập từ các khách sạn ở châu Âu khi số lượng quốc gia đến từ châu Âu chiếm hơn một nửa số mẫu dữ liệu ta quan sát được.
- Như vậy, các khách sạn có thể tạo ra các chính sách ưu đãi dành cho các khách hàng đến từ các quốc gia thuộc châu Âu để có thể thu hút thêm các khách hàng đến nghỉ ngơi tại khách sạn của họ. Việc tập trung phân tích thói quen của các khách du lịch

đến từ châu Âu trong bộ dữ liệu này có thể giúp ta phát hiện ra các mối tương quan thú vị giữa các biến, từ đó giúp ta đề xuất ra các chiến lược kinh doanh tốt hơn dành cho các khách sạn mà bộ dữ liệu này hướng đến.

5.3. Phân tích phân phối cho các biến định lượng (numerical)

5.3.1. Phân tích phân phối đối với thuộc tính "lead_time"



Biểu đồ 5.3.1: Biểu đồ thể hiện phân phối của thuộc tính "lead_time".

Đại lượng thống kê cho `Lead Time`	Giá trị
min	0.000000
lower_quartile	11.000000
median	49.000000
upper_quartile	125.000000
max	737.000000
mean	79.891368
std	86.052325
skew	1.431774
kurt	2.126343

Bảng 5.3.1: Bảng mô tả các đại lượng thống kê cho cột "lead_time".

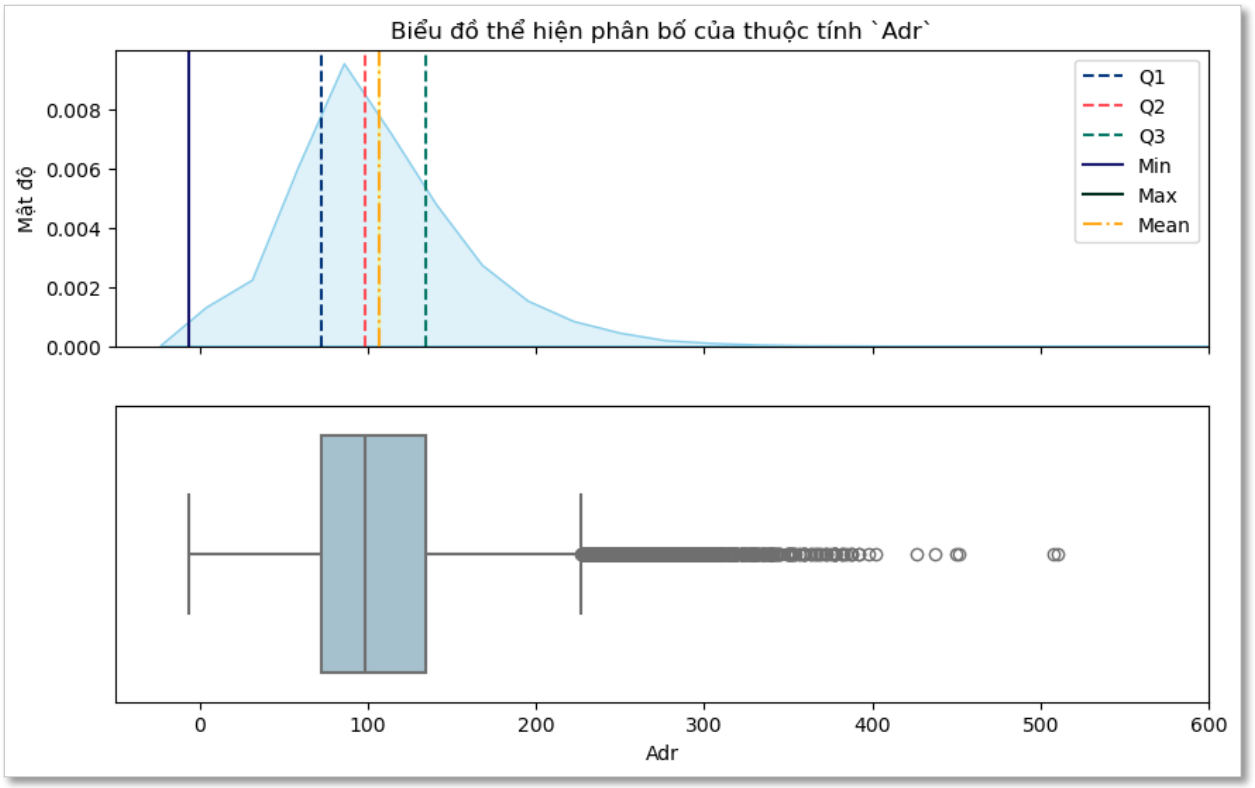
Nhận xét:

- Ta thấy thuộc tính "lead_time" có phân bố lệch phải khá nặng (skew = 1.43 > 0.5) và có phần đuôi rất đậm (kurt = 2.13 > 0):
  - Các giá trị ta quan sát được sẽ có phạm vi phân bố nằm trong đoạn [0; 737] (đơn vị: ngày).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [11; 125] (đơn vị: ngày).
  - Phạm vi phân bố của 25% giá trị lớn nhất rộng hơn rất nhiều so với phạm vi phân bố của 25% giá trị nhỏ nhất và 50% giá trị ở khu vực trung tâm là một điều đáng chú ý.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị trung bình của thuộc tính "lead_time" là (79.32; 80.46) (đơn vị: ngày).
  - Quan sát biểu đồ hộp ở ngay bên dưới, ta phát hiện có rất nhiều điểm dữ liệu ngoại lai nằm lệch về phía bên phải của phân phối. Điều này có thể lý giải phần nào cho việc biểu đồ mật độ lại có phần đuôi trải rất dài về phía bên phải. Việc có quá nhiều giá trị ngoại lai có thể là một xu hướng thú vị mà ta cần tiến hành phân tích chi tiết hơn để tìm hiểu nguyên nhân sâu xa của hiện tượng này. Đồng thời, ta cũng có thể thực hiện các phương pháp giúp tiền xử lý giá trị ngoại lai nếu cần thiết để các phân tích của ta có tính tổng quát cao hơn.
- Như vậy, phân phối của thuộc tính "lead_time" không đối xứng, tập trung khá nhiều ở các khoảng giá trị nhỏ và thưa thớt dần khi tiến về phần đuôi phía bên phải. Ta có thể thấy rằng khoảng thời gian chênh lệch giữa ngày đặt phòng và ngày đến nhận phòng của khách hàng thường không quá lớn. Điều này phản ánh rằng trong phần lớn các trường hợp, khách hàng thường đặt phòng trong khoảng thời gian ngắn trước khi đến ngày nhận phòng. Điều này có thể là do họ thích lên kế hoạch gần với thời điểm thực hiện chuyến đi hoặc có sự linh hoạt trong việc thay đổi kế hoạch. Việc tập trung

phân tích các mẫu dữ liệu có giá trị "lead_time" ở mức trung bình - thấp có thể giúp ta phát hiện mối tương quan giữa các yếu tố chi phối việc đặt phòng khách sạn.

- Tuy nhiên, tình trạng một số lượng không nhỏ các điểm dữ liệu nằm lệch rất xa về phía bên phải của phân phối cũng có thể phản ánh thói quen trong một nhóm khách hàng thú vị. Ta có thể tiến hành phân tích chuyên sâu trên các mẫu dữ liệu này để tìm ra đặc điểm chung giữa các mẫu dữ liệu. Từ đó, ta có thể rút ra các quy luật chung giúp tăng hiệu suất hoạt động của khách sạn trên các nhóm khách hàng đặc thù.

5.3.2. Phân tích phân phối đối với thuộc tính "adr"



Biểu đồ 5.3.2: Biểu đồ thể hiện phân phối của thuộc tính "adr".

Đại lượng thống kê cho `Adr`	Giá trị
min	-6.380000
lower_quartile	72.000000
median	98.100000
upper_quartile	134.000000
max	5400.000000
mean	106.337246
std	55.013953
skew	10.921447
kurt	981.620772

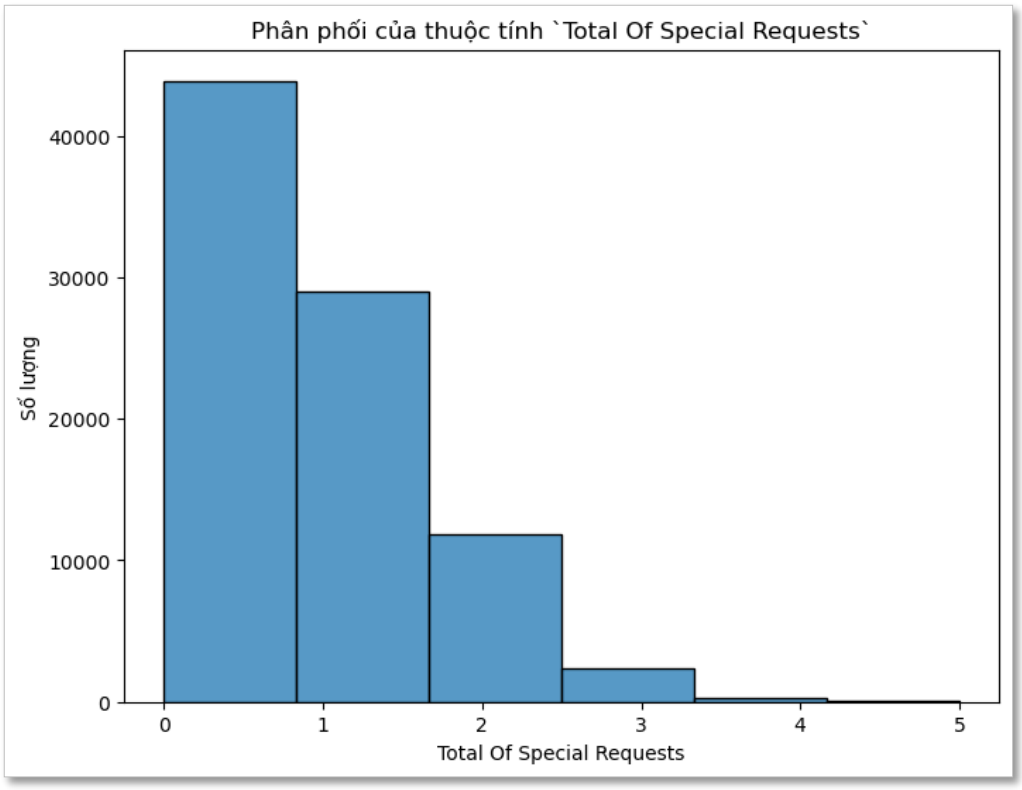
Bảng 5.3.2: Bảng mô tả các đại lượng thống kê cho cột "adr".

Nhận xét:

- Ta thấy thuộc tính "adr" có phân bố lệch phải rất nghiêm trọng (skew = 10.92 > 0.5) và có phần đuôi cực kỳ đậm (kurt = 981.62 > 0):
  - Các giá trị ta quan sát được sẽ có phạm vi phân bố nằm trong đoạn [-6.38; 5400] (đơn vị tiền tệ).
  - Khoảng 50% điểm dữ liệu ở khu vực trung tâm sẽ có giá trị nằm trong đoạn [72; 134] (đơn vị tiền tệ).
  - Phạm vi phân bố của 25% giá trị lớn nhất rộng hơn rất nhiều so với phạm vi phân bố của 25% giá trị nhỏ nhất và 50% giá trị ở khu vực trung tâm là một điều đáng chú ý.
  - Trong bộ dữ liệu này, với độ tin cậy 95%, khoảng tin cậy cho giá trị trung bình của thuộc tính "adr" là (105.97; 106.70) (đơn vị tiền tệ).
- Quan sát biểu đồ hộp ở ngay bên dưới, ta phát hiện có rất nhiều điểm dữ liệu ngoại lai nằm lệch về phía bên phải của phân phối. Điều này có thể lý giải phần nào cho việc biểu đồ mật độ lại có phần đuôi trải rất dài về phía bên phải. Trong đó, ta quan sát thấy có một điểm dữ liệu với giá trị "không tưởng" là 5000. Đây chính là giá trị ngoại lai đã kéo lệch toàn bộ phân bố của thuộc tính "adr". Do đó, ta cần có phương án tiền xử lý phù hợp đối với mẫu dữ liệu này để giúp các phân tích có độ tin cậy cao hơn.
- Như vậy, doanh thu trung bình của các khách sạn trong bộ dữ liệu này thường không quá cao và tập trung chủ yếu ở mức trung bình. Điều này có thể cung cấp một cơ sở mạnh mẽ cho nhận định về việc mã phòng loại "A" được lựa chọn nhiều vì có chi phí rẻ mà ta đã đề cập bên trên. Tuy nhiên, ta cần tiến hành các phân tích trên dữ liệu để làm rõ nhận định này.
- Tuy nhiên, có một số lượng không nhỏ các mẫu dữ liệu nằm lệch về phía bên phải của phân phối. Đó chính là các khách hàng tiềm năng sẵn sàng chi tiêu nhiều hơn cho các dịch vụ của khách sạn. Việc tập trung phân tích đặc điểm chung giữa các mẫu dữ liệu này có thể giúp khách sạn kiếm được doanh thu nhiều hơn từ các khách hàng này.

- Do đó, khách sạn một mặt cần tiếp tục duy trì chất lượng cho các dịch vụ có chi phí ở mức trung bình - khá để tạo ra nguồn doanh thu ổn định từ đại đa số các khách hàng đến đặt phòng. Mặt khác, khách sạn cũng cần nâng cao chính sách chăm sóc các khách hàng sử dụng dịch vụ cao cấp, vì đây chính là cơ hội để tạo ra nguồn thu nhập khổng lồ cho khách sạn.

5.3.3. Phân tích phân phối đối với thuộc tính "total_of_special_requests"



Biểu đồ 5.3.3: Biểu đồ thể hiện phân phối của thuộc tính "total_of_special_requests".

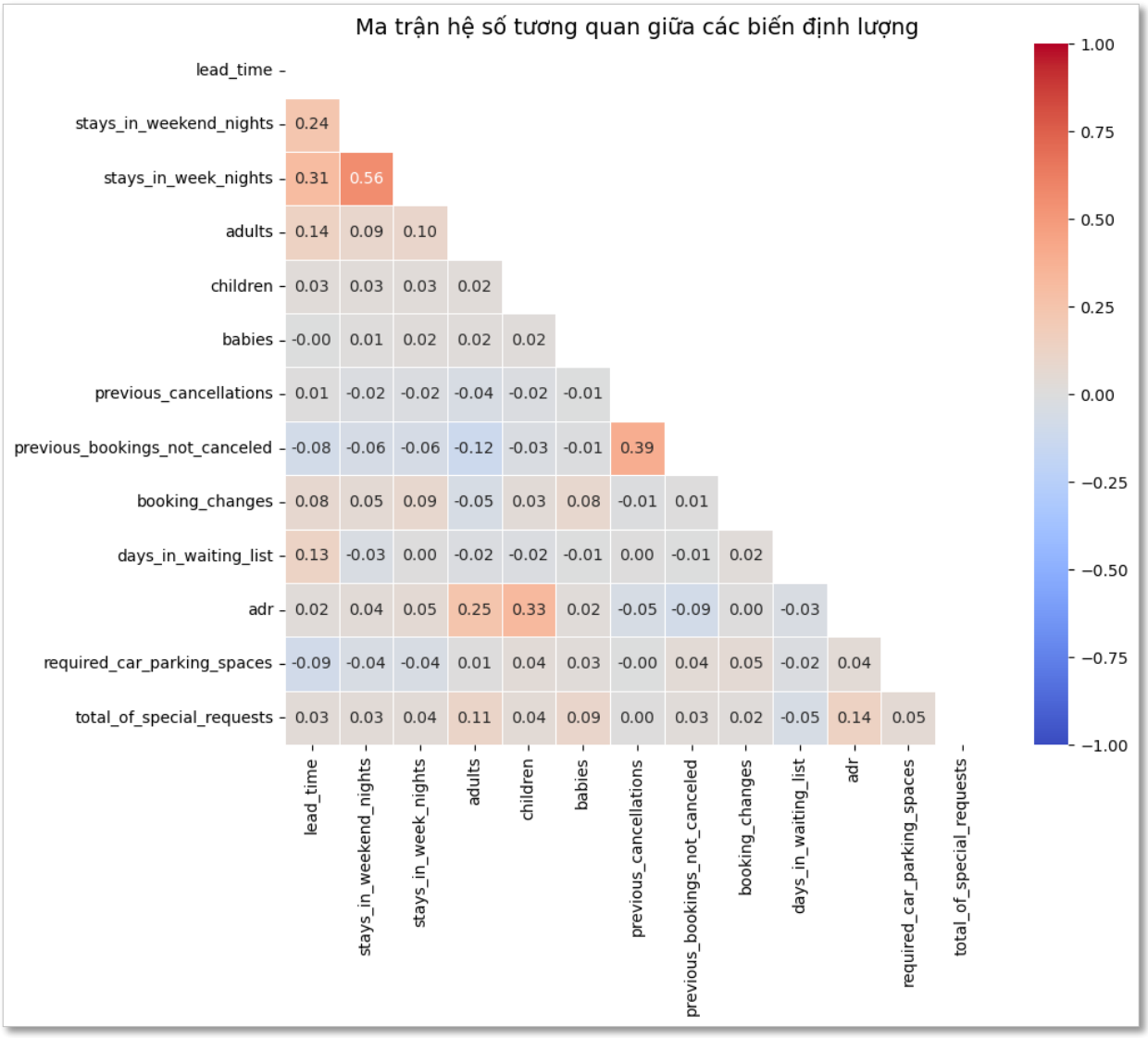
Nhận xét:

- Từ kết quả phân tích dữ liệu, ta thấy thuộc tính "total_of_special_requests" có giá trị nhỏ nhất là 0 (tức là khách hàng không có yêu cầu đặc biệt) và giá trị lớn nhất là 5. Phần lớn các điểm dữ liệu sẽ tập trung ở các giá trị nhỏ như 0 và 1. Khi số lượng yêu cầu vượt qua giá trị 1 thì số lượng quan sát sẽ giảm đi đáng kể.
- Như vậy, ta thấy hầu hết các khách hàng sẽ không đưa ra bất kỳ yêu cầu đặc biệt gì cho khách sạn. Vì các yêu cầu đặc biệt thường có tính đặc thù và không phải ai cũng có yêu cầu đặc biệt như vậy. Do đó, kết quả mà ta quan sát được là hoàn toàn hợp lý và không có gì bất thường.
- Tuy nhiên, số lượng yêu cầu đặc biệt ("total_of_special_requests") chỉ mang tính tổng quát giúp ta nhận biết xu hướng chung trong việc đặt phòng của khách hàng chứ không

mô tả nội dung chi tiết của các yêu cầu. Do đó, ta cần phải biết nội dung chi tiết hoặc ít nhất là biết được yêu cầu thuộc vào nhóm nào để có thể đưa ra các phân tích cụ thể, chi tiết hơn. Chẳng hạn như nếu ta biết xuất hiện một lượng lớn khách hàng với mong muốn được cung cấp cà phê hòa tan của một hãng nào đó, thì ta có thể đề xuất cho khách sạn kết hợp với nhãn hàng đó trong dài hạn để có thể mua được sản phẩm với giá cả phải chăng hơn. Điều này vừa giúp đáp ứng được yêu cầu của khách hàng, giúp khách hàng thoải mái hơn khi sử dụng dịch vụ của khách sạn. Đồng thời cũng giúp khách sạn giảm thiểu chi phí và tối đa hóa lợi nhuận từ khách hàng của mình.

6. EDA 2D

6.1. Phân tích hệ số tương quan giữa các biến định lượng (numerical)



Biểu đồ 6.1: Biểu đồ thể hiện hệ số tương quan giữa các biến định lượng.

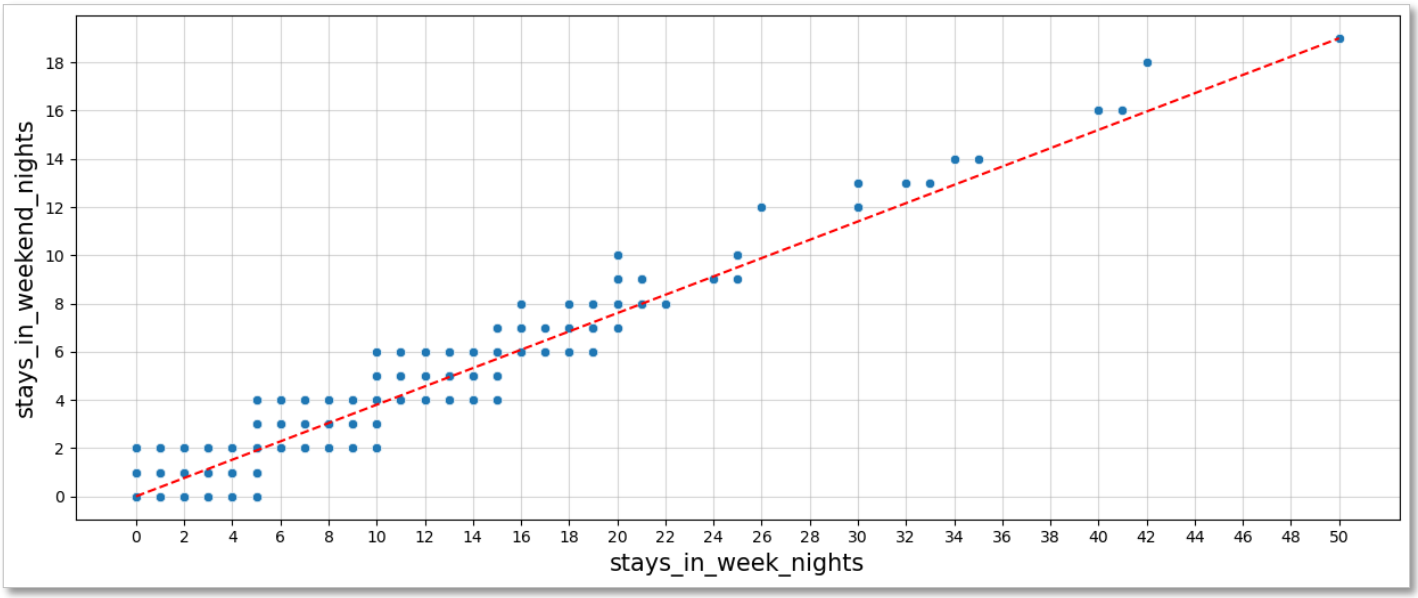
Nhận xét:

- Từ kết quả phân tích dữ liệu, ta thấy không có mối quan hệ tương quan "hoàn hảo" giữa hai thuộc tính khác nhau. Giá trị hệ số tương quan Pearson giữa các cặp thuộc tính số trong bộ dữ liệu ta đang phân tích có giá trị nằm trong khoảng (-0.2; 0.6).
- Nổi bật nhất trên biểu đồ là mối tương quan thuận khá mạnh giữa hai thuộc tính "stays_in_weekend_nights" và "stays_in_week_nights" (với giá trị của hệ số tương quan là 0.56).
- Ngoài ra, ta còn nhận thấy một số mối quan hệ nổi bật giữa các biến:
  - "previous_bookings_not_canceled" và "previous_cancellations" có mức độ tương quan khá cao (0.39).
  - "adr" và "children" có mức độ tương quan khá cao (0.33).
  - "stays_in_week_nights" và "lead_time" có mức độ tương quan khá cao (0.31).
  - "adr" và "adults" có mức độ tương quan vừa phải (0.25).
  - "stays_in_weekend_nights" và "lead_time" có mức độ tương quan vừa phải (0.24).
- Hầu hết giá trị hệ số tương quan giữa các cặp thuộc tính còn lại đều có độ lớn (giá trị tuyệt đối) bé hơn 0.2. Điều này cho thấy các thuộc tính này có mức độ tương quan khá thấp hoặc thậm chí là không tương quan với nhau. Do đó, ta có thể không cần phân tích mối tương quan giữa các cặp thuộc tính đó.
- Như vậy, ta có thể thấy rằng mức độ tương quan giữa các thuộc tính số không quá cao, đa phần dừng ở mức trung bình - thấp. Điều này cho thấy các biến thường độc lập với nhau, không có sự phụ thuộc lẫn nhau quá nhiều. Ta sẽ cần thực hiện thêm nhiều phân tích để có nhìn nhìn chính xác hơn về mối quan hệ giữa các biến. Tuy nhiên, trong bộ dữ liệu vẫn tồn tại một vài cặp biến có mức tương quan cao. Việc tập trung phân tích mối quan hệ giữa chúng có thể giúp ta phát hiện ra cách chúng tương tác với nhau và ảnh hưởng của chúng đến việc đặt phòng khách sạn.

6.2. Sử dụng Scatter plot để phân tích dữ liệu 2D

6.2.1. Phân tích mối quan hệ giữa "stays_in_weekend_nights" và "stays_in_ week_nights"

Sau khi đã có cái nhìn tổng quát về mối quan hệ giữa các thuộc tính định lượng, ta sẽ dùng Scatter plot để phân tích dữ liệu. Kết quả từ quá trình phân tích dữ liệu có thể giúp ta hiểu rõ hơn về mối quan hệ giữa hai thuộc tính "stays_in_weekend _nights" và "stays_in_week_nights".



Biểu đồ 6.2.1.a: Biểu đồ thể hiện mối tương quan giữa "stays_in_weekend_nights" và "stays_in_week_nights".

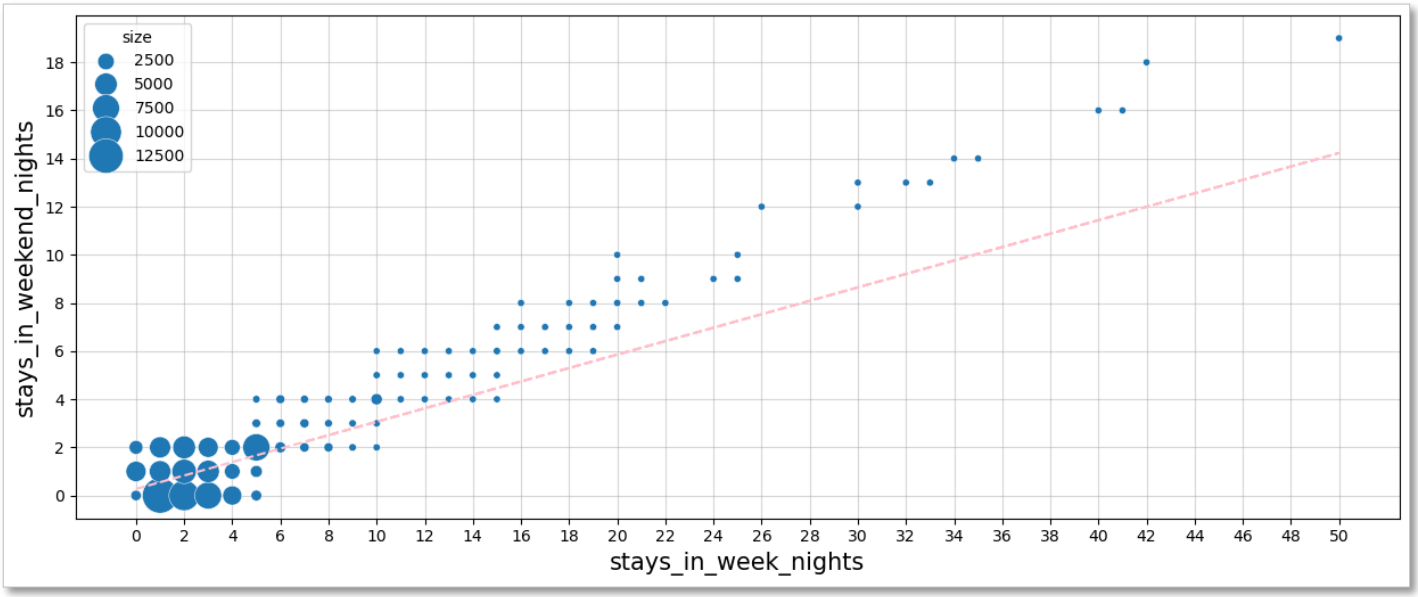
Nhận xét:

- Phần lớn giá trị của thuộc tính "stays_in_week_nights" sẽ tập trung trong khoảng từ 0 đến 22.
- Phần lớn giá trị của thuộc tính "stays_in_weekend_nights" sẽ tập trung trong khoảng từ 0 đến 10.
- Chúng ta sẽ vẽ một đường thẳng đi qua điểm nhỏ nhất và lớn nhất trên biểu đồ để dễ quan sát. Qua quan sát, có vẻ như hai thuộc tính này có mối quan hệ rất chặt chẽ với

nhau khi các điểm dữ liệu hầu hết tập trung xung quanh đường thẳng và không có điểm ngoại lai (outlier).

- Tuy nhiên, để chắc chắn hơn, chúng ta sẽ tiến hành tìm đường thẳng hồi quy tuyến tính cho nó.

Lần này, chúng ta sẽ thêm vào biểu đồ số lượng của các điểm dữ liệu được biểu thị thông qua độ lớn của chấm tròn. Ta đồng thời tạo ra một đường thẳng hồi quy cho tập dữ liệu của hai thuộc tính "stays_in_weekend_nights" và "stays_in_week _nights", sau đó trực quan kết quả lên biểu đồ.



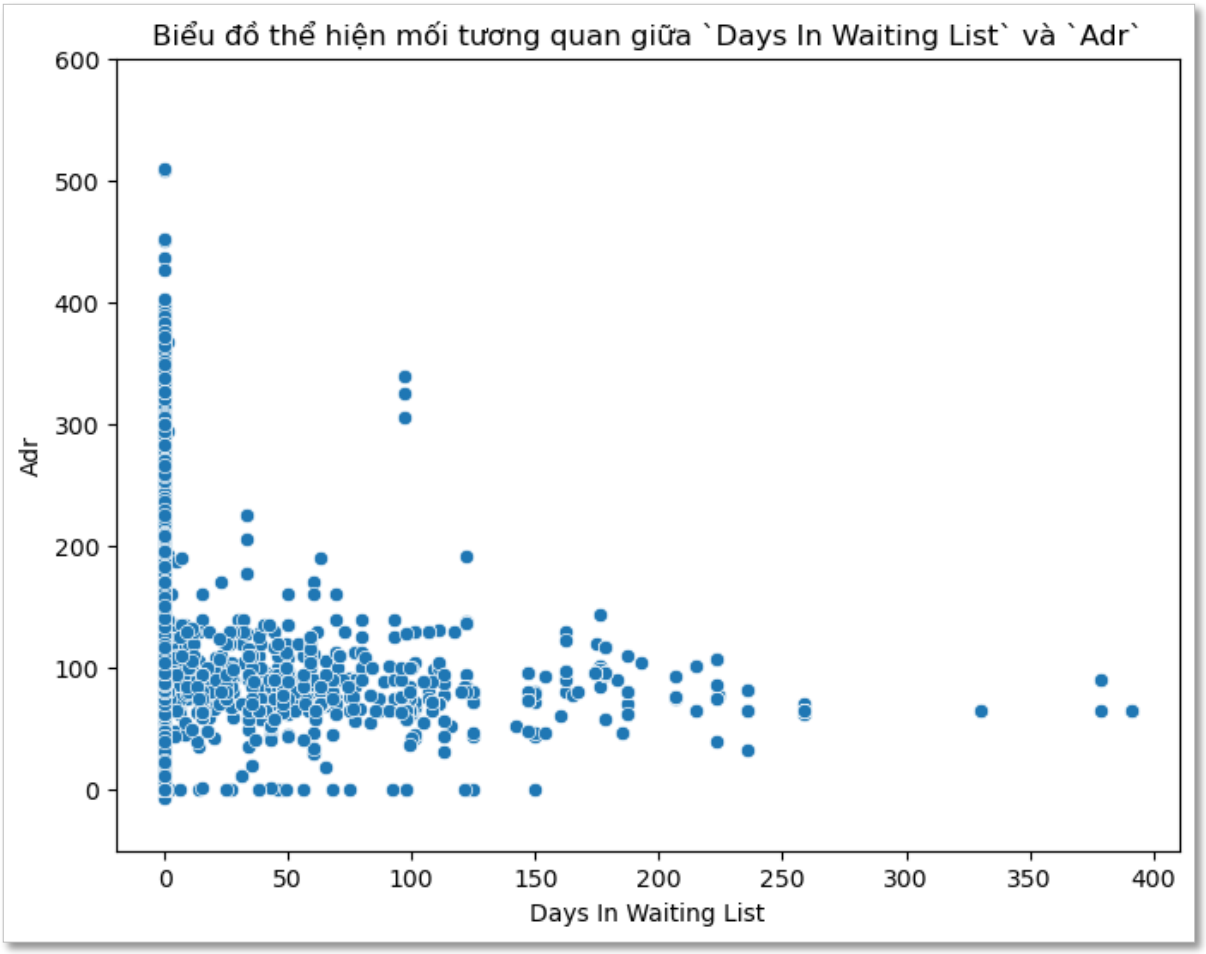
Biểu đồ 6.2.1.b: Biểu đồ thể hiện mối tương quan giữa "stays_in_weekend_nights" và "stays_in_week_nights" với đường hồi quy.

Nhận xét:

- Các điểm dữ liệu không còn thực sự khớp với đường thẳng hồi quy mới.
- Hầu hết các điểm dữ liệu tập trung trong khoảng: 0 <= "stays_in_week_nights" <= 6; 0 <= "stays_in_weekend_nights" <= 3.

- Do số lượng các điểm dữ liệu phân bố không đều, nên đường thẳng chỉ khớp ở các dữ liệu có giá trị nhỏ; khi có giá trị càng lớn, dữ liệu càng không khớp với đường thẳng.
- Có thể thấy, "stays_in_weekend_nights" và "stays_in_week_nights" có mối quan hệ đồng biến nhưng không hoàn toàn.
- Khách hàng có xu hướng chọn dịch vụ ở qua đêm vào các ngày trong tuần nhiều hơn các ngày cuối tuần.
- Khi số đêm ở lại thuộc các ngày trong tuần tăng thì số đêm ở lại thuộc các ngày cuối tuần cũng tăng (nhưng tăng ít hơn).
- Khách hàng không hoàn toàn chỉ chọn một trong hai dịch vụ trên.

6.2.2. Phân tích mối quan hệ giữa "days_in_waiting_list" và "adr"



Biểu đồ 6.2.2: Biểu đồ thể hiện mối tương quan giữa "days_in_waiting_list" và "adr".

Nhận xét:

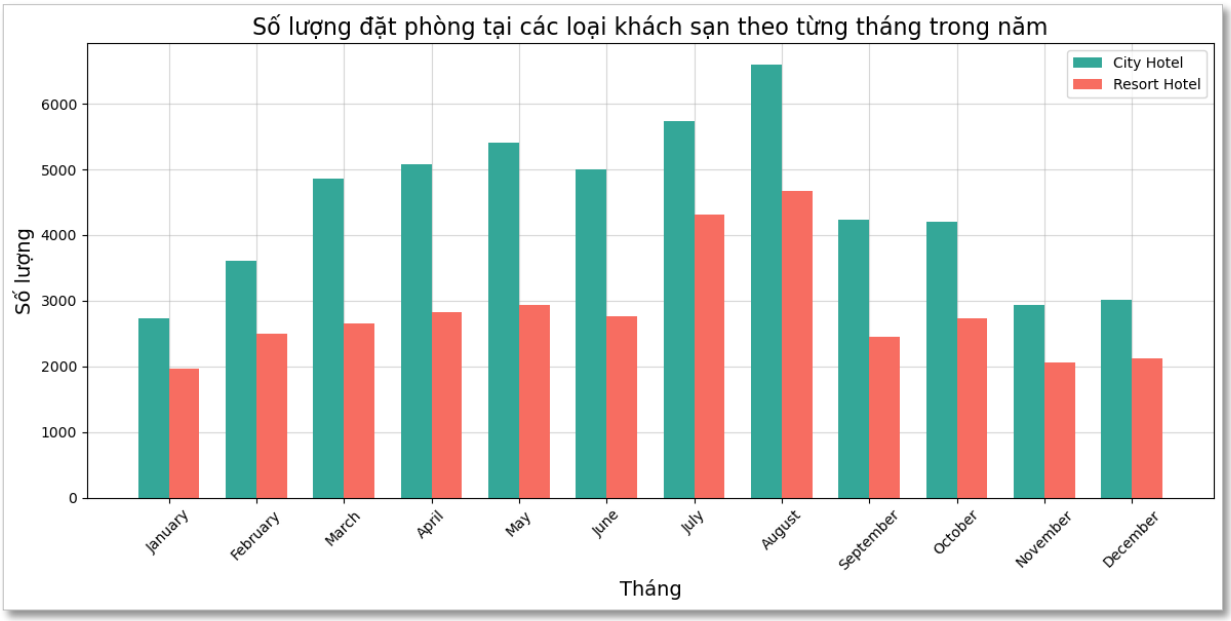
- Từ kết quả phân tích dữ liệu, ta thấy hai thuộc tính "days_in_waiting_list" và "adr" có mối tương quan nghịch không quá rõ ràng. Nhìn chung, khi giá trị của thuộc tính "days_in_waiting_list" tăng lên thì giá trị của thuộc tính "adr" sẽ có xu hướng giảm xuống (và ngược lại). Điều này phản ánh một xu hướng là: khi một lượt đặt phòng có thời gian nằm trong danh sách chờ càng ngắn thì khách hàng sẽ sẵn lòng chi trả nhiều hơn cho dịch vụ của khách sạn, khi này doanh thu trung bình của khách sạn sẽ tăng

lên. Điều này cũng không quá khó hiểu vì khi khách hàng phải chờ đợi lâu thì tâm trạng của họ cũng ít nhiều bị ảnh hưởng.

- Hiểu được mong đợi của khách hàng, phía khách sạn cần cố gắng đưa ra các biện pháp nhằm hạn chế tình trạng khách hàng phải chờ đợi quá lâu. Một trong những cách đầu tiên mà ta có thể nghĩ ngay đến là việc mở rộng số lượng phòng để có thể cung cấp chỗ ở đồng thời cho nhiều người khách hàng hơn. Tuy nhiên cách làm này thường gặp nhiều khó khăn trong thực tế và hiếm khi có thể hoàn tất trong một khoản thời gian ngắn.
- Thay vào đó, phía khách sạn có thể cải thiện chính sách chăm sóc khách hàng của mình bằng cách tạo ra một số ưu đãi nhỏ dành tặng cho các khách hàng nằm trong danh sách chờ quá lâu. Điều này có thể phần nào cải thiện tâm trạng của khách hàng, giúp cuộc trò chuyện giữa hai bên vui vẻ hơn. Khi này có thể khách hàng sẽ sẵn sàng lựa chọn các gói dịch vụ cao cấp cho chuyến đi du lịch, giúp công ty thu được nhiều lợi nhuận hơn. Do đó, việc không ngừng cải thiện và nâng cao chính sách chăm sóc khách hàng sẽ luôn là chìa khóa giúp khách sạn có thể thu được lợi nhuận tốt hơn.

6.3. Sử dụng bar chart để phân tích dữ liệu "numerical" và "categorical"

6.3.1. Phân tích số lượng đặt phòng tại các loại khách sạn theo các tháng trong năm

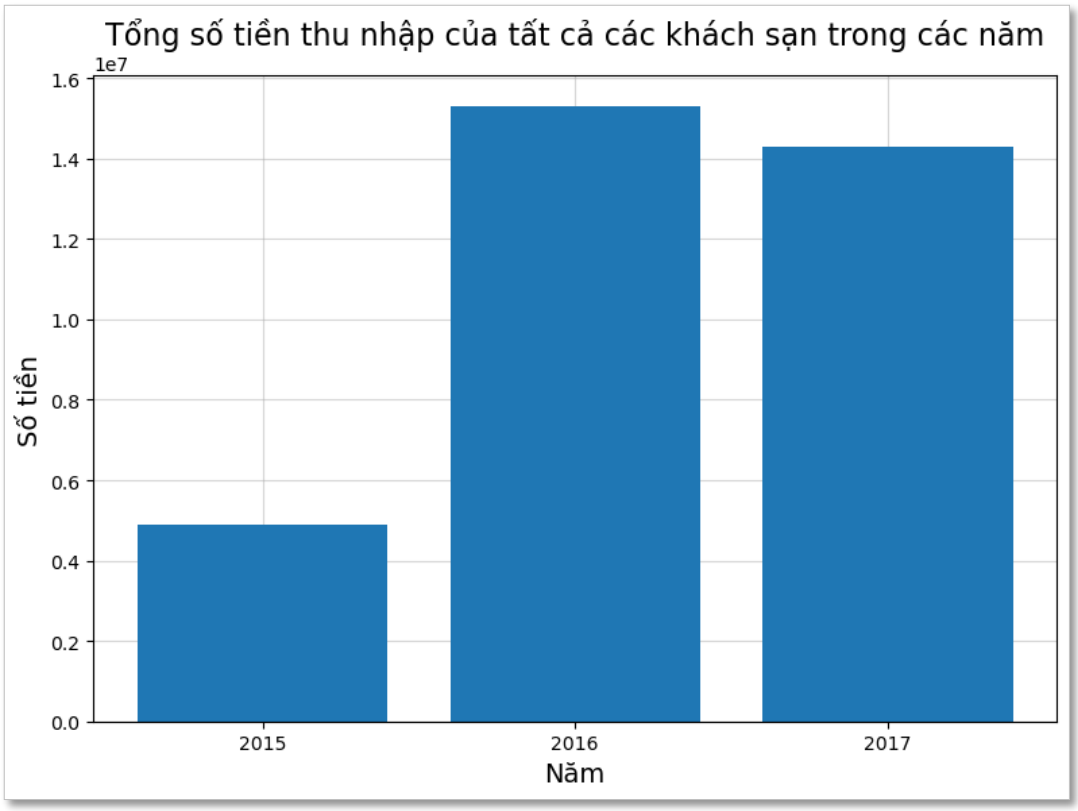


Biểu đồ 6.3.1: Biểu đồ thể hiện số lượng đặt phòng tại các loại khách sạn theo các tháng trong năm.

Nhận xét:

- Khách hàng đặt phòng trong tất cả các tháng trong năm, tuy nhiên, số lượng phân bố không đều.
- Lượng khách hàng đặt phòng khách sạn nhiều nhất là vào tháng 8 với tổng cộng khoảng 11000 lượt đặt phòng.
- Lượng khách hàng đặt phòng khách sạn ít nhất là vào tháng 1, 11, 12 với khoảng 5000 lượt đặt phòng mỗi tháng.
- Trong tất cả các tháng, "City Hotel" luôn được lựa chọn nhiều hơn so với "Resort Hotel".
- Lượng đặt phòng có xu hướng tăng từ tháng 1 đến 8, sau đó lại giảm dần từ tháng 8 về tháng 1 năm sau.
- Xu hướng trên là khá phù hợp với thực tế vì mùa hè và mùa thu là thời điểm phổ biến mà mọi người thường tổ chức nhiều chuyến đi du lịch.

6.3.2. Phân tích tổng số tiền mà tất cả các khách sạn thu được trong các năm



Biểu đồ 6.3.2: Biểu đồ thể hiện tổng số tiền mà tất cả các khách sạn thu được trong các năm.

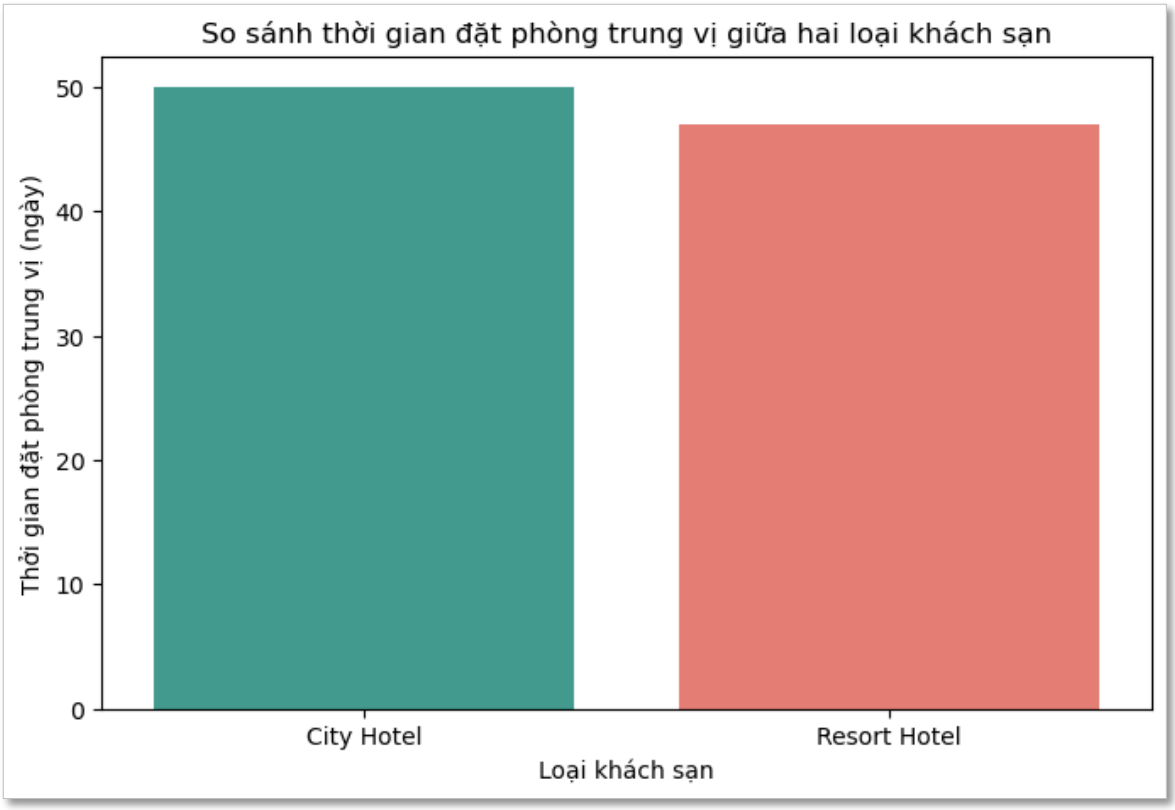
Nhận xét:

- Có sự chênh lệch lớn giữa tổng số tiền thu được giữa các năm.
- Năm nhiều nhất là 2016 với 15000000 (đơn vị tiền tệ).
- Năm ít nhất là 2015 với 5000000 (đơn vị tiền tệ).
- Chênh lệch giữa năm 2016 và 2015 là khoảng 10000000 (đơn vị tiền tệ), gấp 3 lần.
- Tổng số tiền thu được tăng đột biến từ năm 2015 sang năm 2016 và giảm nhẹ từ năm 2016 sang 2017.
- Có thể thấy, năm 2016 là một bước nhảy vọt thu nhập của các khách sạn, có thể do nhiều nguyên nhân thúc đẩy. Năm 2017 tuy doanh thu có giảm nhưng không quá

nghiêm trọng, đây vẫn là một dấu hiệu tiềm năng về thu nhập cho tất cả các khách sạn.

- Xu hướng sử dụng các dịch vụ khách sạn của khách hàng ngày càng nhiều, là một lĩnh vực phù hợp để đầu tư và phát triển.

6.3.3. Phân tích thời gian đặt phòng giữa hai loại khách sạn



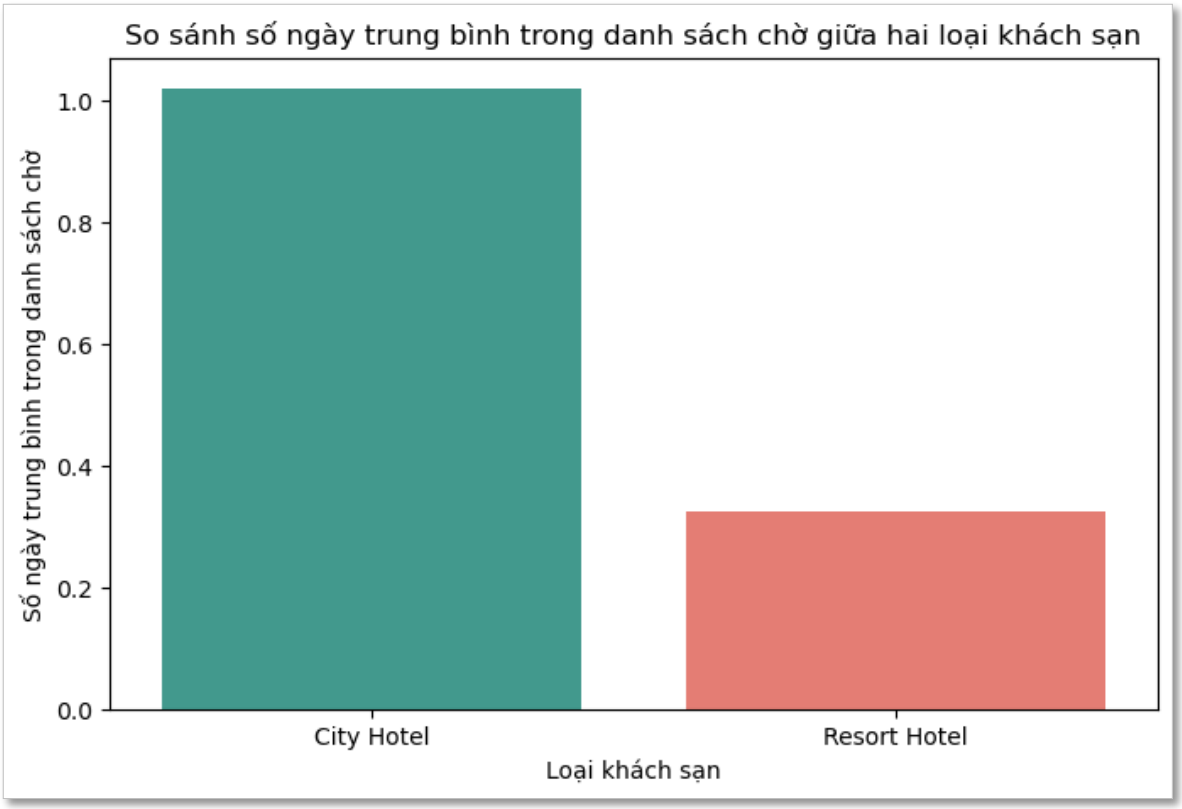
Biểu đồ 6.3.3: Biểu đồ thể hiện thời gian đặt phòng giữa hai loại khách sạn.

Nhận xét:

- Nhìn chung, thời gian đặt phòng trung vị ("median_lead_time") giữa hai loại khách sạn không có quá nhiều sự chênh lệch. Tuy nhiên, các "City Hotel" vẫn thường có thời gian đặt phòng sớm hơn một chút so với "Resort Hotel". Mặt khác, do số lượng mẫu dữ liệu ở hai nhóm khách sạn là không tương đồng nên có thể các phân tích của ta chưa thể phản ánh đúng tình trạng trong thực tế. Do đó, ta có thể tiến hành thêm phân tích trên các bộ dữ liệu về đặt phòng khách sạn ở nhiều khu vực khác trên thế giới để có cái nhìn tổng quát hơn.
- Cả hai loại khách sạn đều có thời gian đặt phòng trung vị vào khoảng sớm hơn 50 ngày so với ngày chính thức nhận phòng. Ta nhận thấy đây là một giá trị khá lớn, cho thấy các khách hàng dù đến loại khách sạn nào thì cũng thường lên kế hoạch từ khá

sớm (khoảng trước 2 tháng). Đây cũng là một xu hướng dễ hiểu vì nếu không đặt phòng khách sạn từ sớm thì rất dễ xảy ra khả năng đến lúc ta tới nơi thì không còn phòng nào trống để thuê.

6.3.4. Phân tích số ngày trong danh sách chờ giữa hai loại khách sạn

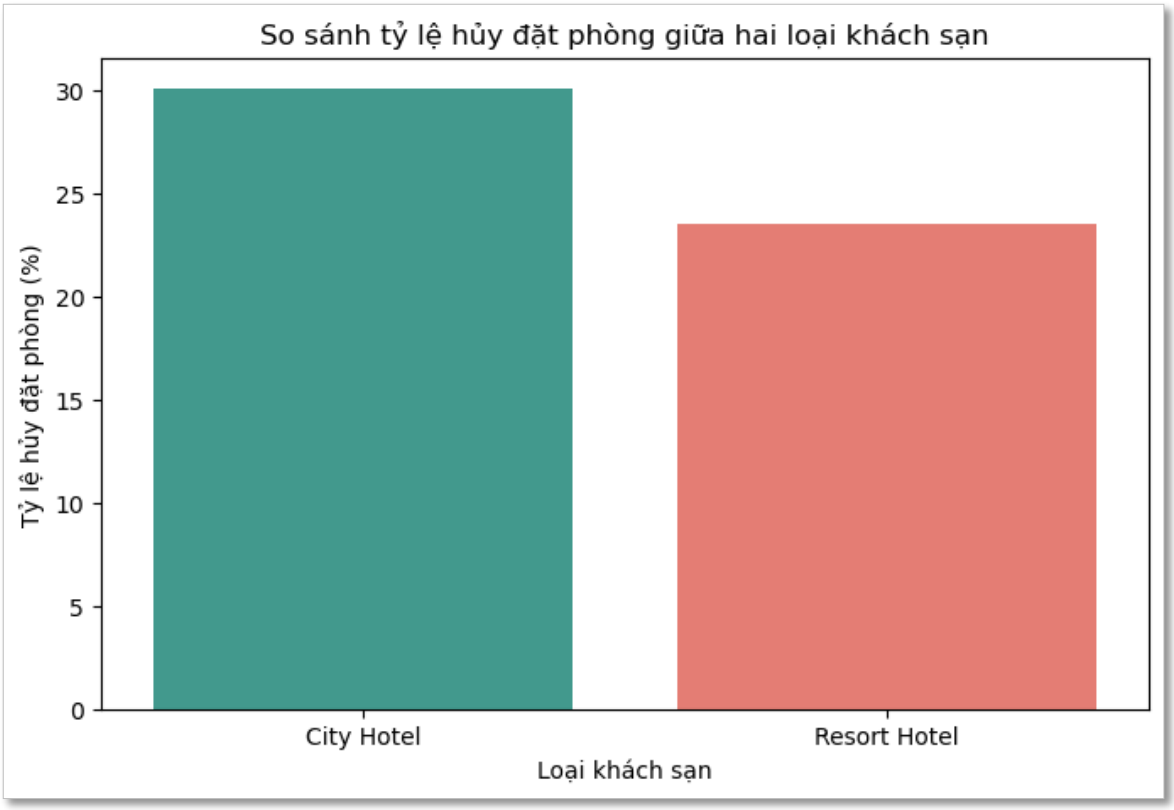


Biểu đồ 6.3.4: Biểu đồ thể hiện số ngày trong danh sách chờ giữa hai loại khách sạn.

Nhận xét:

- Nhìn chung, khách hàng của các "City Hotel" thường có thời gian trung bình trong danh chờ lâu hơn so với các khách hàng lựa chọn nghỉ dưỡng tại các "Resort Hotel". Điều này cho thấy các "City Hotel" thường "bận rộn" hơn so với các "Resort Hotel".
- Điều này cũng phần nào phản ánh thực tế khách quan. Các khách sạn "Resort Hotel" có chi phí đắt đỏ nên thường có ít khách hàng đến đặt phòng. Trong khi chi phí phải trả cho các "City Hotel" thường thoải mái hơn nên tỷ lệ người sử dụng (và muốn sử dụng) "City Hotel" sẽ nhiều hơn đáng kể. Có thể vì điều này đã làm cho các "City Hotel" thường rơi vào tình trạng quá tải và làm gia tăng thời gian khách hàng ở trong danh sách chờ khi tiến hành đặt phòng.

6.3.5. Phân tích tỷ lệ hủy đặt phòng giữa hai loại khách sạn



Biểu đồ 6.3.5: Biểu đồ thể hiện tỷ lệ hủy đặt phòng giữa hai loại khách sạn.

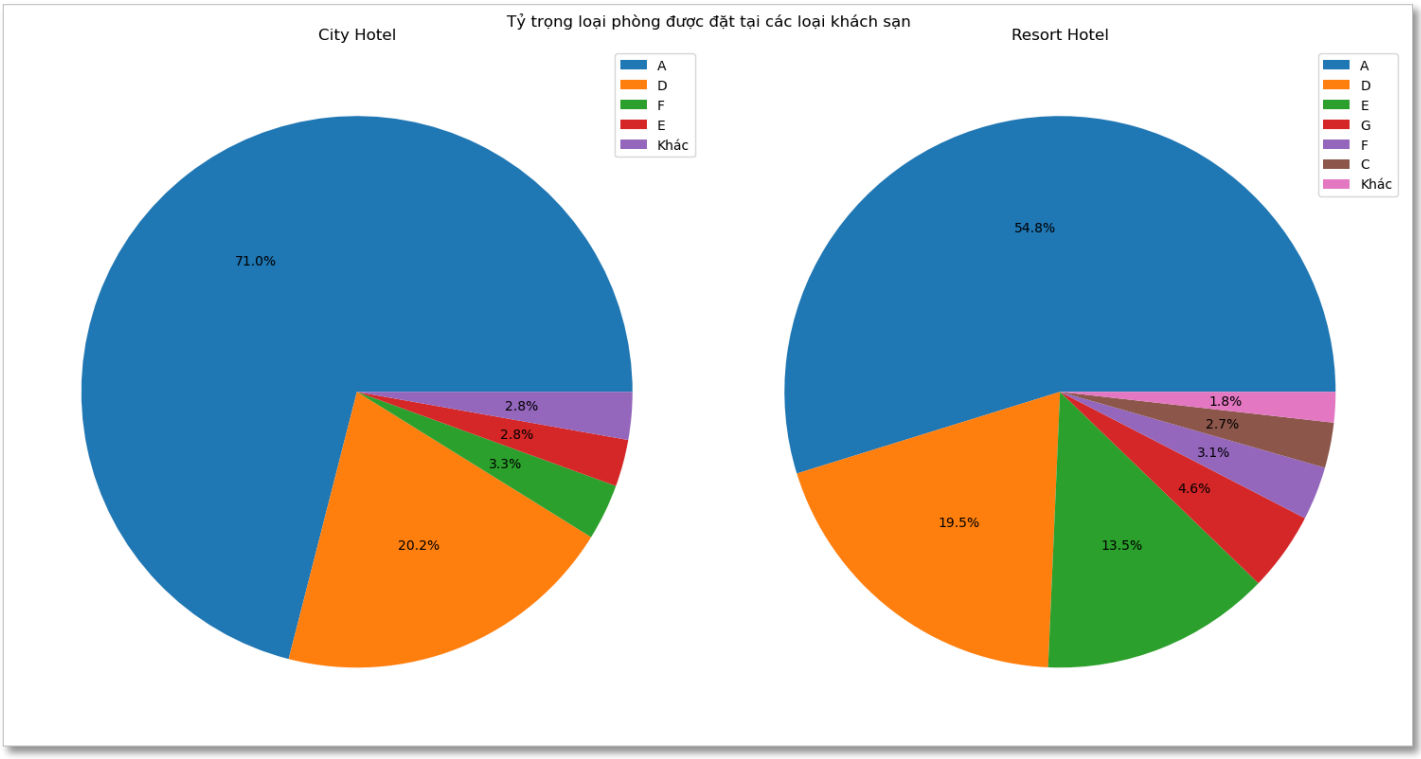
Nhận xét:

- Như vậy, ta thấy rằng "City Hotel" vừa có tỷ lệ đặt phòng cao hơn vừa có tỷ lệ hủy đặt phòng cao hơn "Resort Hotel". Cụ thể, theo kết quả thống kê từ dữ liệu, tỷ lệ hủy đặt phòng của "City Hotel" là khoảng 30%, nhiều hơn tương đối so với tỷ lệ chưa đến 25% của "Resort City".
- Tuy nhiên, việc cả hai loại khách sạn đều có tỷ lệ hủy đặt phòng khá lớn cũng cho thấy tình trạng hủy đặt phòng không thực sự tập trung ở một nhóm khách hàng cụ thể nào. Điều này đặt ra một thách thức, một bài toán kinh tế cần giải quyết để hạn chế tình trạng hủy đặt phòng và tối đa hóa doanh thu cho khách sạn. Có rất nhiều phương pháp được áp dụng và một trong các phương pháp phổ biến nhất là sử dụng một mô hình học máy giúp xác định các khách hàng có rủi ro cao và không cho phép họ đặt

phòng trước. Đây là một bài toán lớn và đòi hỏi ta phải đầu tư nhiều công sức mới mong thu được các kết quả tích cực.

6.4. Tính tỷ trọng đối với hai biến "categorical"

6.4.1. Phân tích tỷ trọng loại phòng được đặt tại các loại khách sạn khác nhau

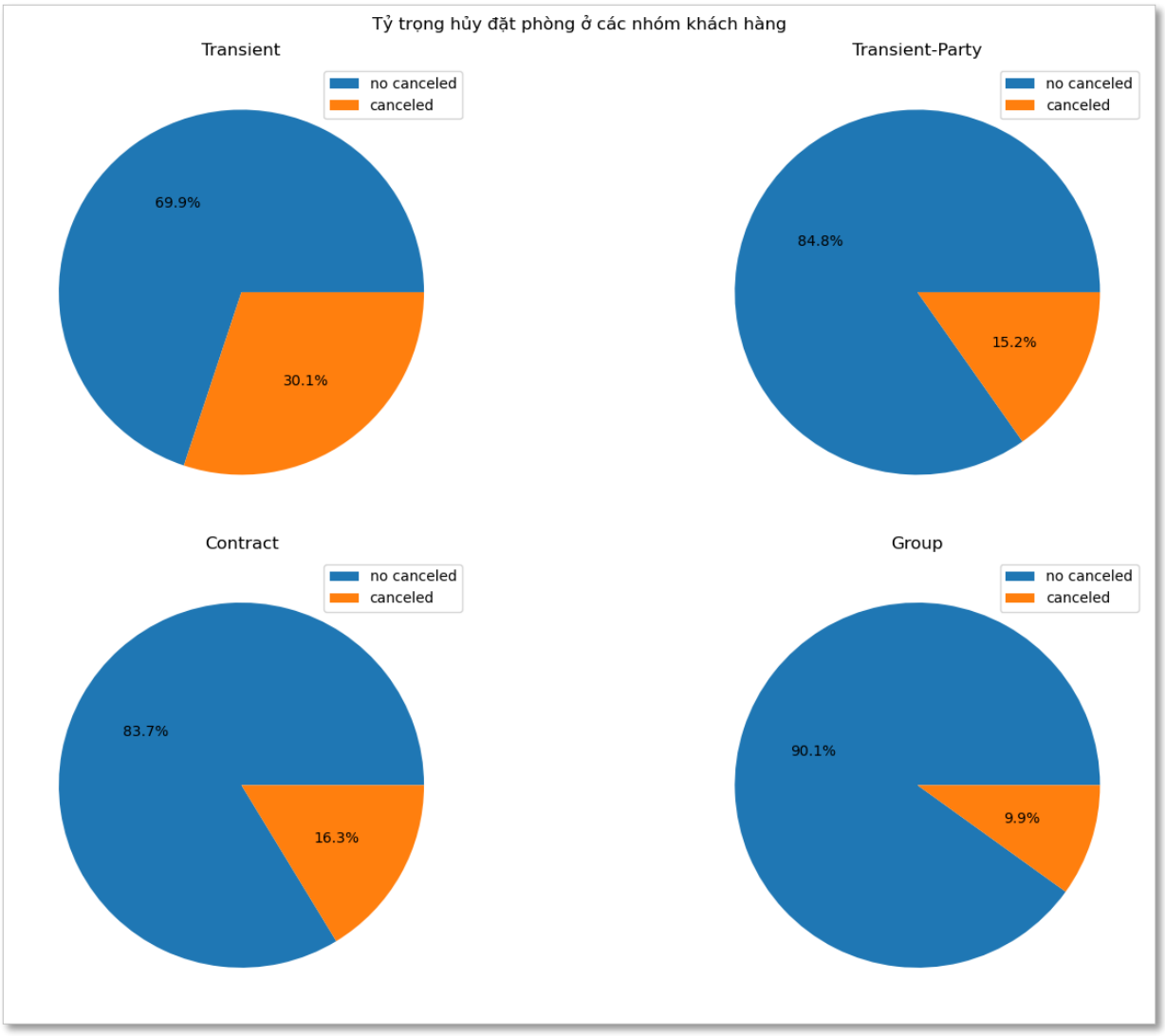


Biểu đồ 6.4.1: Biểu đồ thể hiện tỷ trọng loại phòng được đặt tại các loại khách sạn khác nhau.

Nhận xét:

- Nhìn chung, tại các loại khách sạn, hai loại phòng được yêu thích nhất lần lượt là loại "A" và loại "D".
- Tại "City Hotel", phòng loại "A" chiếm tỉ lệ rất lớn (khoảng 71%), phòng loại "D" chiếm khoảng 20% và khoảng 9% chia cho các loại phòng còn lại.
- Còn đối với "Resort Hotel", ta thấy tỷ lệ giữa các loại phòng cân bằng hơn, loại "A" vẫn nhiều nhất nhưng chỉ chiếm khoảng 55%, loại "D" chiếm khoảng 19.5%, loại "E" chiếm khoảng 13.5% và khoảng 12% chia cho các loại phòng còn lại.

6.4.2. Phân tích tỷ trọng hủy đặt phòng ở các nhóm khách hàng khác nhau



Biểu đồ 6.4.2: Biểu đồ thể hiện tỷ trọng hủy đặt phòng ở các nhóm khách hàng khác nhau.

Nhận xét:

- Tỷ lệ hủy đặt phòng ở nhóm khách hàng "Transient" là cao nhất với khoảng 30%, do đặc thù của nhóm khách hàng này là mọi thứ không có sự sắp xếp trước nên tỷ lệ hủy đặt phòng cao.

- Tỷ lệ hủy đặt phòng ở nhóm khách hàng "Group" là thấp nhất với khoảng 10%, do đặc thù của nhóm khách hàng này là có sự sắp xếp trước và có sự tham gia của nhiều người nên tỷ lệ hủy đặt phòng thấp.
- Tỷ lệ hủy đặt phòng ở nhóm "Transient-party" và "Contract" là xấp xỉ nhau, khoảng 15-16% (không quá cao).
- Các khách sạn cần có quy định phù hợp cũng như tạo điều kiện thuận lợi cho khách hàng để đảm bảo quyền lợi của khách sạn mà vẫn đảm bảo sự hài lòng của khách hàng khi hủy đặt phòng ở các nhóm khách hàng có tỷ lệ hủy đặt phòng cao.

7. EDA 3D

7.1. Tiền xử lý dữ liệu

(1) Đầu tiên, ta tạo ra một DataFrame mới để tránh làm ảnh hưởng đến tập dữ liệu bên trên.

(2) Sau đó, ta tạo thêm thuộc tính mới để phục vụ cho quá trình phân tích dữ liệu:

- "total_stays" = "stays_in_weekend_nights" + "stays_in_week_nights"
- "total_people" = "adults" + "children" + "babies"

(3) Ta thấy có một vài điểm dữ liệu có thuộc tính "total_people" bằng 0, điều này khá kỳ lạ. Do đó, ta sẽ xem đây là các dữ liệu nhiễu là loại bỏ các dòng này để tránh làm ảnh hưởng đến kết quả phân tích.

(4) Phân tích và xử lý giá trị ngoại lai trên thuộc tính "adr" (sử dụng phương pháp IQR).

Bước 1: Đầu tiên, ta xác định giới hạn dưới và giới hạn trên của thuộc tính "adr" theo phương pháp IQR.

Bước 2: Sau đó, ta kiểm tra xem thuộc tính "adr" có giá trị ngoại lai hay không? Nếu có thì đó là (những) giá trị nào?

	adr
38749	5400.000000
80728	510.000000
11661	508.000000
74461	451.500000
9984	450.000000

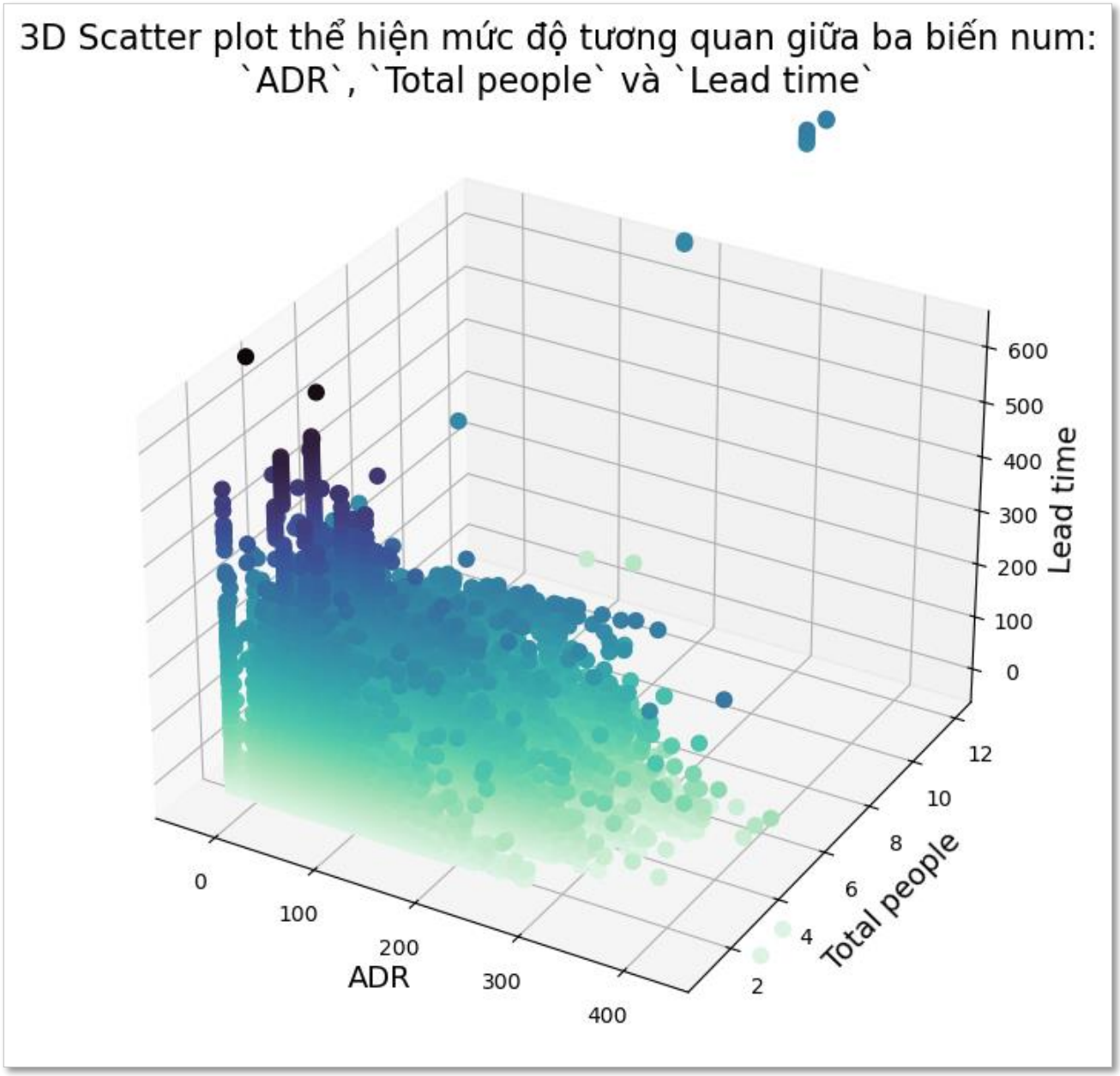
Hình 7.1: Kết quả phát hiện các giá trị ngoại lai của thuộc tính "adr".

Nhận xét:

- Ta biết rằng giá trị ngoại lai không phải lúc nào cũng mang ý nghĩa tiêu cực và điều đó cũng đúng phần nào trong trường hợp này. Từ bảng kết quả, ta thấy chỉ có một điểm dữ liệu có giá trị tại thuộc tính "adr" lớn hơn 5000, cách rất xa phạm vi phân bố của các điểm dữ liệu còn lại. Trong khi đó, các giá trị ngoại lai khác thường có khoảng cách không quá xa so với giới hạn trên mà ta tìm được. Do đó, ta sẽ xử lý trường hợp này bằng cách: loại bỏ điểm dữ liệu có giá trị tại thuộc tính "adr" lớn hơn 5000 và giữ nguyên các điểm dữ liệu còn lại.

Bước 3: Loại bỏ (các) điểm dữ liệu có giá trị tại thuộc tính "adr" lớn hơn 5000.

7.2. Sử dụng Scatter plot để phân tích dữ liệu 3D cho ba biến định lượng (numerical)



Biểu đồ 7.2: Biểu đồ thể hiện mức độ tương quan giữa ba thuộc tính:
"adr", "total_people" và "lead_time".

Nhận xét:

(1) Giữa "adr" và "total_people":

- Ta thấy có mối tương quan theo chiều dương không quá mạnh giữa hai thuộc tính "adr" và "total_people". Nhìn chung, khi số lượng người trong nhóm hành khách tăng lên thì khách sạn cũng thu được nhiều lợi nhuận hơn, do đó giá trị "adr" cũng cao hơn.
- Tuy nhiên mối liên hệ này không phải lúc nào cũng đúng. Khi số lượng người trong nhóm hành khách vượt qua con số 5 thì lợi nhuận mà khách sạn thu được thường không quá ấn tượng, giá trị "adr" thường thấp hơn 100. Quan sát biểu đồ, ta thấy các nhóm hành khách đi từ 1 đến 5 người là nhóm khách hàng chủ yếu và đóng góp rất nhiều vào nguồn doanh thu của khách sạn.
- Việc các nhóm hành khách đông người thường đem lại giá trị doanh thu không quá tốt có thể còn phụ thuộc vào các yếu tố khác như: loại bữa ăn, loại phòng, loại khách sạn, v.v. mà họ đã lựa chọn. Ngoài ra, số điểm dữ liệu thuộc vào nhóm này cũng không quá nhiều, nên ta cần được cung cấp thêm dữ liệu để có thể phân tích và làm sáng tỏ các xu hướng thú vị trong nhóm hành khách đông người.

(2) Giữa "adr" và "lead_time":

- Mối tương quan tuyến tính giữa hai thuộc tính "adr" và "lead_time" không quá rõ ràng. Nhìn chung, những hành khách có hẹn nhận phòng trong vòng một năm kể từ ngày đặt lịch ("lead_time" < 365) thường sẽ tạo ra doanh thu nhiều hơn cho khách sạn. Trong trường hợp khoảng thời gian giữa ngày đặt phòng và ngày nhận phòng vượt quá một năm, khi thời gian kéo dài càng lâu, ta thấy doanh thu của khách sạn có xu hướng giảm xuống.
- Từ kết quả thống kê mà ta quan sát được, ta có thể đặt ra một câu hỏi: "Liệu rằng các hành khách có lịch hẹn sớm hơn một năm ("lead_time" > 365) sẽ nhận được nhiều chính sách ưu đãi hơn từ phía khách sạn?". Đây chỉ là một câu hỏi phỏng đoán để giúp ta lý giải mối tương quan giữa hai thuộc tính bên trên. Và ta chỉ có thể trả lời câu hỏi này nếu được cung cấp thông tin về chính sách ưu đãi của khách sạn. Như vậy,

"lead_time" có thể không phải là một thuộc tính đủ tốt để sử dụng trong bài toán dự đoán doanh thu ("adr") của khách sạn.

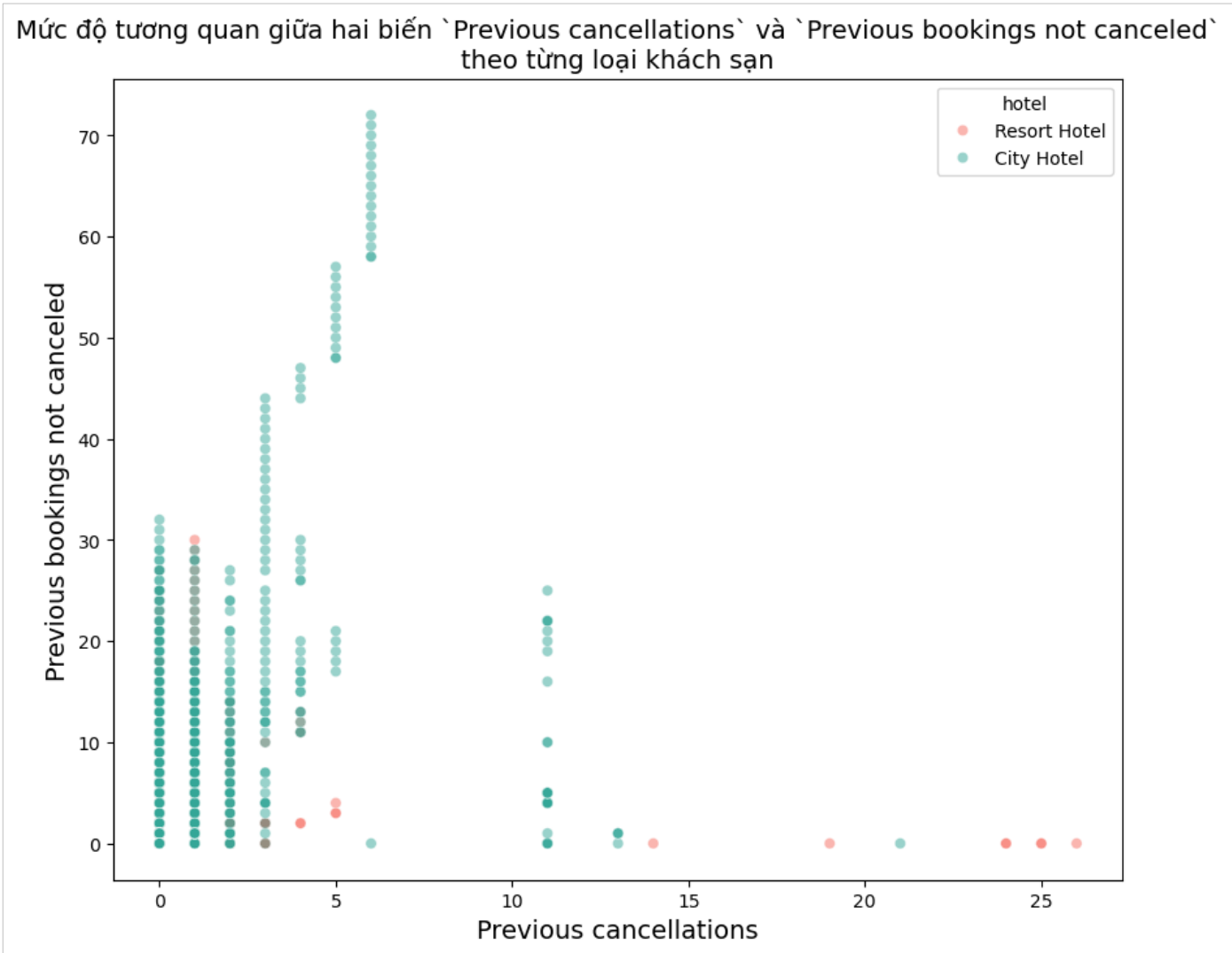
(3) Giữa "total_people" và "lead_time":

- Ta thấy có mối tương quan theo chiều dương khá yếu giữa hai thuộc tính "total_people" và "lead_time". Nhìn chung, khi số lượng người trong nhóm hành khách tăng lên thì họ cũng có xu hướng đặt phòng sớm hơn ("lead_time" lớn). Điều này cũng không quá khó hiểu vì các nhóm hành khách đông người thường sẽ là: một đại gia đình, một nhóm bạn bè hoặc là các nhân viên trong một công ty, v.v.. Khi đi du lịch đông người mà ta không đặt hẹn với khách sạn từ sớm thì sẽ rất dễ xuất hiện tình trạng thiếu phòng và làm ảnh hưởng đến kế hoạch của cả nhóm. Do đó, để tránh những tình huống không may xảy ra, trong các chuyến du lịch đông người, hành khách sẽ có xu hướng đặt phòng khách sạn từ rất sớm, thường là gần một năm trước khi buổi du lịch diễn ra.

7.3. Sử dụng Scatter plot 2D và màu đối với hai biến num và cate

Ta sẽ sử dụng "hotel" là biến cate để phân tích các đặc điểm khác nhau giữa hai loại khách sạn.

7.3.1. Kết hợp "hotel" với hai biến num: "previous_cancellations" và "previous_bookings_not_canceled"



Biểu đồ 7.3.1: Biểu đồ thể hiện mức độ tương quan giữa hai thuộc tính "previous_cancellations" và "previous_bookings_not_canceled" theo từng loại khách sạn.

Nhận xét:

(1) Nhìn chung, mối tương quan giữa hai biến "previous_cancellations" và "previous_bookings_not_canceled" có nhiều nét tương đồng ở cả hai nhóm khách sạn "City Hotel" và "Resort Hotel":

- Đối với các hành khách chưa từng hủy đặt phòng hoặc có số lần hủy đặt phòng ít hơn 10 lần, ta thấy hai thuộc tính "previous_cancellations" và "previous_bookings_not_canceled" có mối tương quan theo chiều dương khá mạnh. Nhìn chung, các khách hàng đã nhiều lần hủy đặt phòng thì cũng có nhiều lần không hủy đặt phòng (tức là họ vẫn đến ở khách sạn như lịch hẹn từ trước). Do đó, việc hủy đặt phòng trong trường hợp này không thực sự đồng nghĩa với tình trạng hành khách đang "rời bỏ" khách sạn - tức là không tiếp tục sử dụng khách sạn này mà chuyển sang các khách sạn khác có nhiều chính sách, dịch vụ tốt hơn. Việc hành khách hủy đặt phòng có thể chỉ đơn giản là do họ bận một việc gì đó và không thể đến như lịch hẹn, v.v.. Và các người chủ khách sạn không nên quá lo lắng về chất lượng dịch vụ mà khách sạn của mình cung cấp.
- Đối với các hành khách có số lần hủy đặt phòng nhiều hơn 10 lần, ta thấy hai thuộc tính "previous_cancellations" và "previous_bookings_not_canceled" có mối tương quan theo chiều âm không quá mạnh. Tức là, khi hành khách đã hủy đặt phòng quá nhiều lần ở một khách sạn thì họ có xu hướng là chưa từng đến khách sạn đó, hoặc chỉ mới đến khách sạn đó một vài lần (con số này nhỏ hơn rất nhiều so với số lần hủy đặt phòng). Nguyên nhân cho hiện tượng này có thể đến từ việc các hành khách có việc bận đột xuất, không thể tiến hành buổi đi chơi theo kế hoạch và phải hủy đặt phòng tại khách sạn. Hoặc cũng có thể là do hành khách đã tìm được một khách sạn có chất lượng dịch vụ tốt hơn, giá cả cạnh tranh hơn, v.v. đáp ứng được các yêu cầu của họ và họ quyết định hủy lịch hẹn tại khách sạn đã đặt lịch trước đó. Tuy nhiên, các nhận định bên trên chỉ mang tính giả thuyết chứ không có một cơ sở cụ thể nào cả. Nhưng việc phân tích các hành khách có nhiều lần hủy đặt phòng sẽ là một bài

toán thú vị mà các người chủ dịch vụ khách sạn có thể cân nhắc và tiến hành phân tích trong tương lai để làm rõ nguyên nhân dẫn đến hiện tượng này.

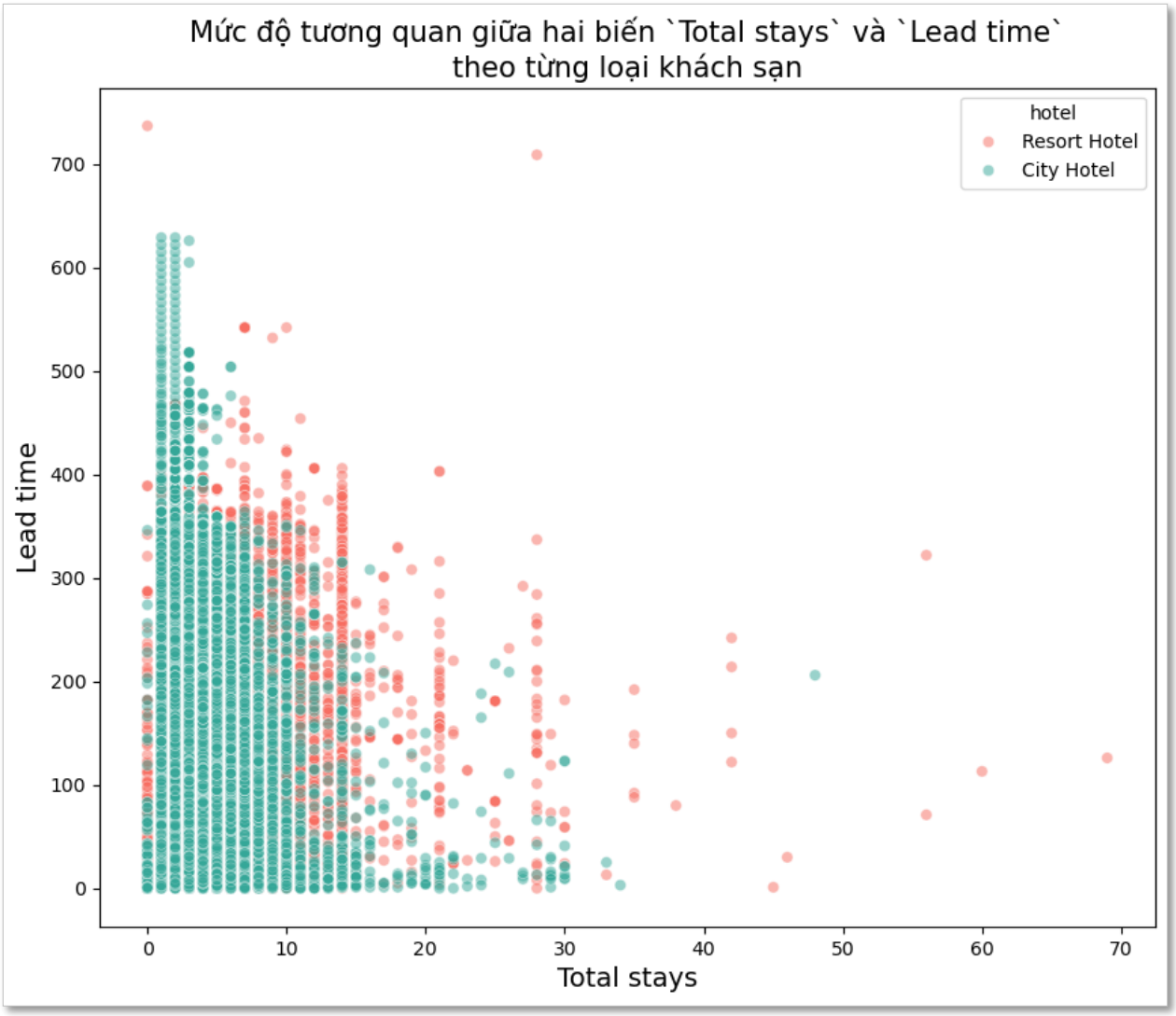
(2) Điểm khác nhau cơ bản giữa hai nhóm "City Hotel" và "Resort Hotel" đến từ phạm vi phân bố của các điểm dữ liệu:

- Đối với thuộc "previous_cancellations", tuy các điểm dữ liệu của hai nhóm đều tập trung chủ yếu trong đoạn [0, 5], nhưng nhóm "Resort Hotel" có phạm vi phân bố rộng hơn một chút so với nhóm "City Hotel". Tuy nhiên, trung bình "previous_ cancellations" của "City Hotel" cao hơn trung bình "previous_cancellations" của "Resort Hotel". Điều này cho thấy tình trạng hành khách hủy đặt phòng sẽ diễn ra phổ biến hơn ở các khách sạn "City Hotel". Ngược lại, "Resort Hotel" sẽ có số lần hủy đặt phòng bởi cùng một khách hàng nhiều hơn.
- Đối với thuộc "previous_bookings_not_canceled", các điểm dữ liệu của nhóm "Resort Hotel" thường tập trung trong đoạn [0, 30]. Trong khi đó, nhóm "City Hotel" lại có phạm vi phân bố rộng hơn khá nhiều, ta thấy các điểm dữ liệu phân bố khá đều trong đoạn [0, 70].

(3) Như vậy, thông qua tập dữ liệu, ta thấy:

- Các khách sạn thuộc loại "Resort Hotel" thường gặp phải tình trạng một hành khách hủy đặt phòng rất nhiều lần. Đồng thời, các hành khách đã ở khách sạn thuộc loại "City Hotel" thường có xu hướng quay trở lại khách sạn này vào các lần tiếp theo. Các "City Hotel" thường tọa lạc ở khu vực thành thị, chẳng hạn như các quận trung tâm thành phố hoặc khu thương mại của các thành phố lớn, v.v.. Có thể chính sự tập nập, náo nhiệt ở nơi đây là một điểm cộng rất lớn và thu hút du khách trở lại nơi đây trong tương lai.

7.3.2. Kết hợp "hotel" với hai biến num: "total_stays" và "lead_time"



Biểu đồ 7.3.2: Biểu đồ thể hiện mức độ tương quan giữa hai thuộc tính "total_stays" và "lead_time" theo từng loại khách sạn.

Nhận xét:

(1) Ở cả hai nhóm "City Hotel" và "Resort Hotel":

- Nhìn chung, hai thuộc tính "total_stays" và "lead_time" có mối tương quan nhẹ theo chiều dương. Nghĩa là, nếu các hành khách có ý định đi du lịch thì họ thường có xu hướng lên kế hoạch và đặt phòng khách sạn từ rất sớm để tránh tình trạng hết phòng, dẫn đến giá trị "lead_time" trong các mẫu dữ liệu này cũng lớn hơn.

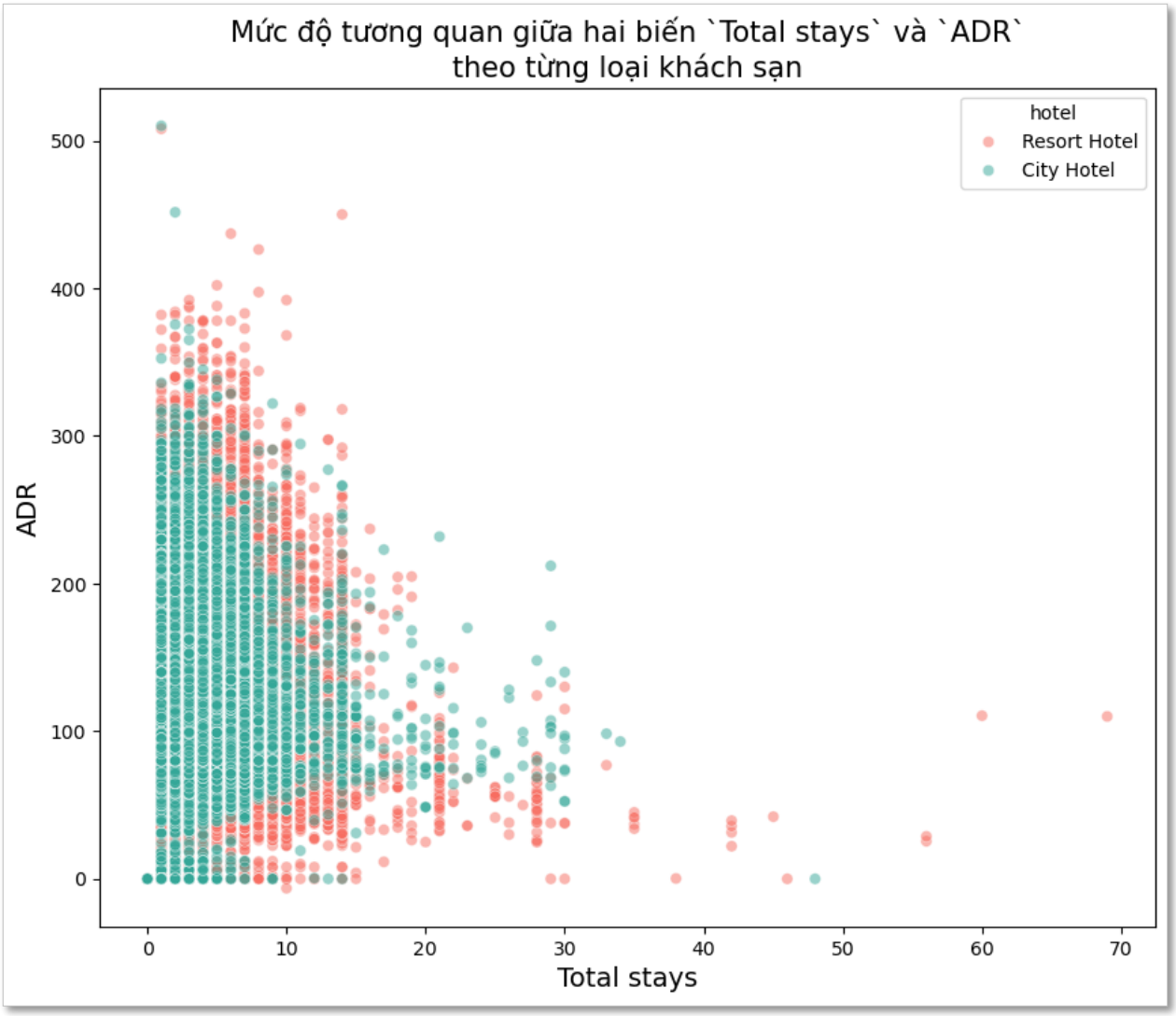
(2) Xét nhóm "City Hotel":

- Phần lớn hành khách ở các "City Hotel" thường có thời gian lưu trú trong khoảng 10 ngày, tuy xuất hiện một vài hành khách có thời gian lưu trú dài hơn một chút (khoảng hơn 1 tháng) nhưng số lượng mẫu dữ liệu này là khá ít. Đồng thời, ta cũng phát hiện rằng khoảng thời gian kể từ ngày đặt phòng đến ngày nhận phòng của các hành khách này thường không vượt quá một năm (365 ngày).
- Tuy nhiên, ta thấy xuất hiện khá nhiều hành khách có thời gian đặt phòng sớm hơn ít nhất 400 ngày. Đây là một nhóm các khách hàng thú vị mà ta nên dành thêm thời gian để phân tích lý do vì sao mà họ lại đặt phòng từ rất sớm như vậy. Đó có thể là các công ty muốn đặt phòng để chuẩn bị cho chuyến đi du lịch của nhân viên, v.v.. Hiểu rõ hơn về đặc điểm của nhóm hành khách này có thể giúp khách sạn đưa ra nhiều chính sách ưu đãi hấp dẫn để thu hút sự quan tâm từ các công ty lớn, từ đó giúp gia tăng doanh thu cho khách sạn.

(3) Xét nhóm "Resort Hotel":

- Nhìn chung, thời gian lưu trú của hành khách ở các khách sạn thuộc loại "Resort Hotel" thường sẽ kéo dài lâu hơn so với "City Hotel". Phần lớn hành khách ở các "Resort Hotel" sẽ có thời gian nghỉ dưỡng kéo dài trong khoảng nửa tháng (15 ngày). Tuy nhiên, cũng có rất nhiều hành khách lựa chọn lưu trú tại khách sạn từ 1 đến 2 tháng. Điều này cũng hoàn toàn hợp lý với mục đích "nghỉ dưỡng" như trong tên gọi của loại khách sạn này ("Resort Hotel").
- So với "City Hotel", thời gian đặt phòng của hành khách ở các khách sạn "Resort Hotel" thường có độ dao động không quá lớn. Gần như toàn bộ các hành khách đều có lịch đặt phòng ít hơn 400 ngày ("lead_time" < 400). Tuy có một vài điểm dữ liệu vượt qua khỏi ngưỡng 400, nhưng số lượng này là không đáng kể.

7.3.3. Kết hợp "hotel" với hai biến num: "total_stays" và "adr"

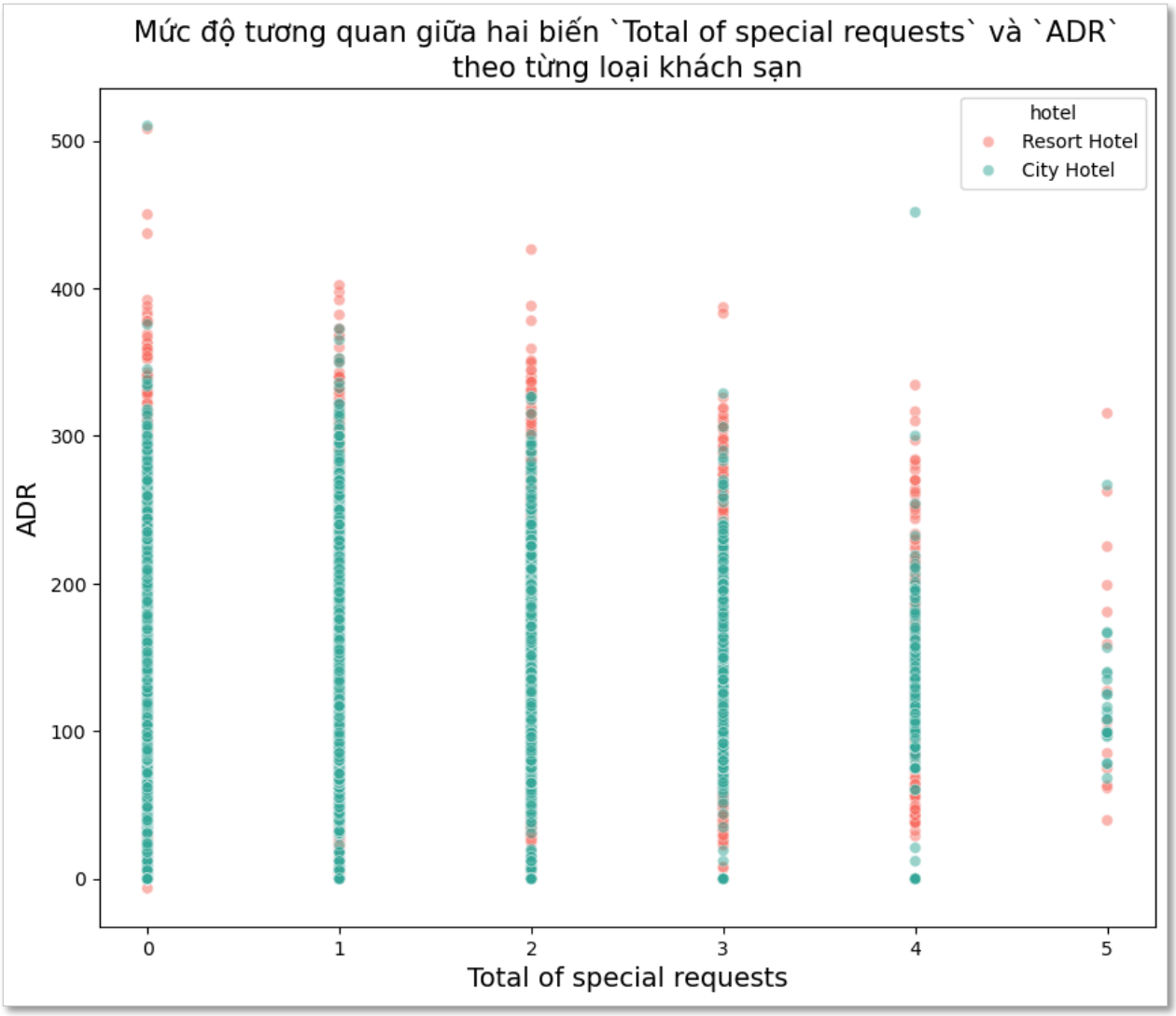


Biểu đồ 7.3.3: Biểu đồ thể hiện mức độ tương quan giữa hai thuộc tính "total_stays" và "adr" theo từng loại khách sạn.

Nhận xét:

- Quan sát biểu đồ phân tán (Scatter plot), ta thấy rằng khi giá trị của thuộc tính "total_stays" tăng lên thì giá trị của thuộc tính "adr" có xu hướng giảm xuống. Nghĩa là, khi các hành khách có thời gian lưu trú lâu hơn tại khách sạn thì lợi nhuận mà khách sạn thu được sẽ có xu hướng giảm xuống.
- Câu "lợi nhuận có xu hướng giảm xuống" không đồng nghĩa với việc khách sạn bị lỗ vốn khi cung cấp dịch vụ cho một khách hàng nào đó. Mà câu bên trên nên được hiểu theo nghĩa là khách sạn thu được số tiền ít hơn (từ hành khác có thời gian lưu trú dài) so với số tiền mà khách sạn kiếm được từ các hành khách có thời gian lưu trú ngắn hơn.
- Như vậy, các hành khách ở lại khách sạn lâu hơn thường có được các "thỏa thuận" tốt hơn từ phía khách sạn. Do đó, nếu một đại gia đình muốn tổ chức chuyến đi du lịch cho tất cả thành viên thì nên lựa chọn các chuyến đi dài ngày và ở lại một khách sạn lâu hơn để có thể tiết kiệm chi phí.
- Mặt khác, nhìn vào thuộc tính "total_stays", ta thấy hành khách thường lựa chọn các "City Hotel" trong các chuyến đi ngắn ngày. Nhưng đối với các chuyến đi dài ngày, các khách sạn thuộc nhóm "Resort Hotel" thường được ưu ái hơn.

7.3.4. Kết hợp "hotel" với hai biến num: "total_of_special_requests" và "adr"



Biểu đồ 7.3.4: Biểu đồ thể hiện mức độ tương quan giữa hai thuộc tính "total_of_special_requests" và "adr" theo từng loại khách sạn.

Nhận xét:

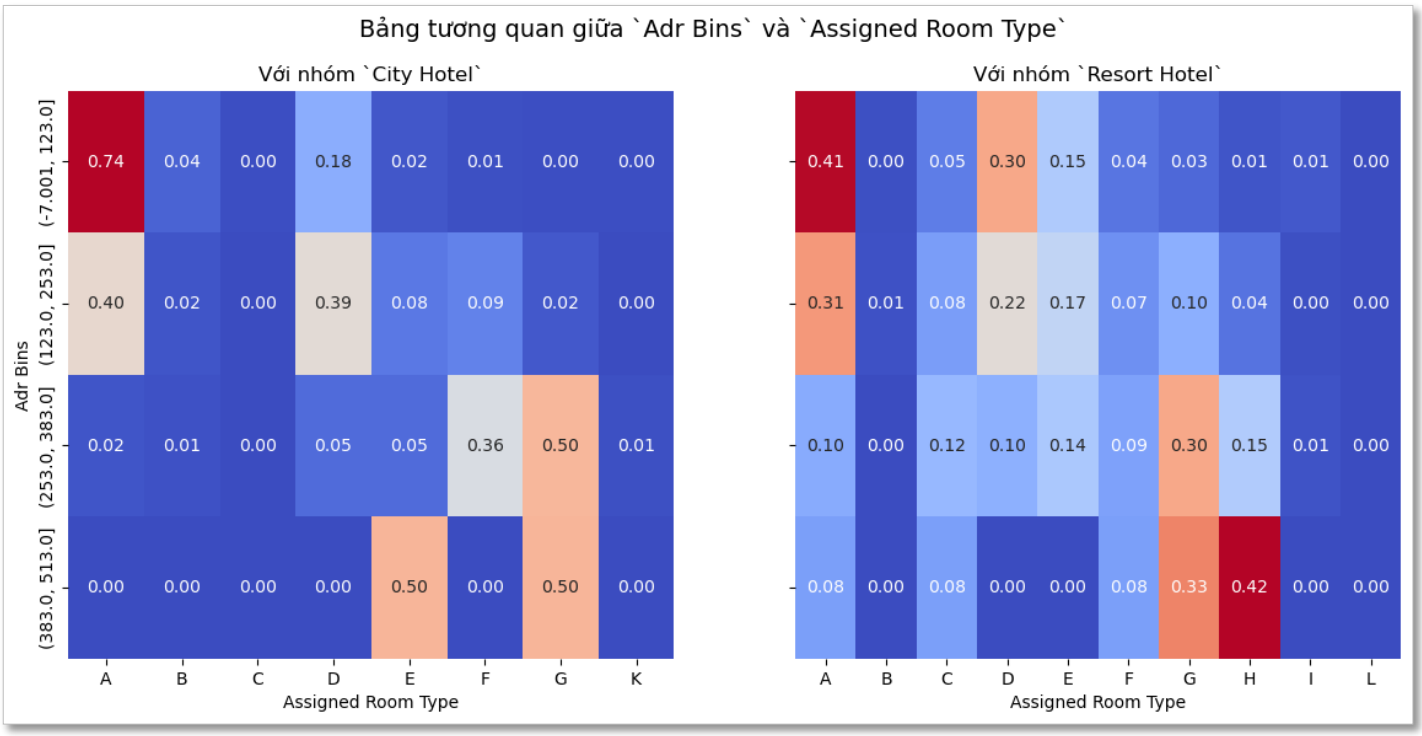
- Quan sát biểu đồ, ta thấy hai thuộc tính "total_of_special_requests" và "adr" có mối tương quan thuận không quá mạnh. Nhìn chung, khi số lượng yêu cầu đặc biệt của hành khách tăng lên thì khách sạn càng kiếm được nhiều doanh thu. Điều này cho thấy khách sạn nên tạo thêm điều kiện và cố gắng đáp ứng càng nhiều yêu cầu đặc biệt từ khách hàng càng tốt. Vì hành khách yêu cầu nhiều dịch vụ đặc biệt có thể sẵn

lòng trả giá cao hơn cho chỗ ở của họ. Những yêu cầu này có thể bao gồm nâng cấp phòng, tiện nghi đặc biệt hoặc dịch vụ cá nhân hóa, v.v.. Tất cả đều có thể tạo ra chi phí bổ sung và đóng góp vào doanh thu của khách sạn. Đây có thể xem là một loại hình dịch vụ đầy hứa hẹn và có thể đóng góp rất nhiều vào doanh thu của cả khách sạn.

- Nhóm "City Hotel" có giá trị "total_of_special_requests" trung bình cao hơn so với giá trị "total_of_special_requests" trung bình của nhóm "Resort Hotel". Tuy nhiên, về mặt doanh thu, các khách sạn thuộc nhóm "Resort Hotel" thường có xu hướng kiếm được nhiều tiền hơn từ các yêu cầu đặc biệt của hành khách so với các khách sạn trong nhóm "City Hotel". Việc này có thể xuất phát từ nhiều nguyên nhân, và một trong số đó có thể là do khách lưu trú tại "Resort Hotel" thường có những kỳ vọng và sở thích khác với những người khách lưu trú tại "City Hotel". Trong khách sạn nghỉ dưỡng, hành khách có thể đưa ra yêu cầu đặc biệt về các hoạt động, tiện nghi hoặc trải nghiệm độc đáo ở khu nghỉ dưỡng, chẳng hạn như trị liệu spa, thuê thiết bị thể thao dưới nước, các chuyến tham quan có hướng dẫn viên hoặc sử dụng các tiện nghi giải trí như sân gôn hoặc dốc trượt tuyết. Những dịch vụ và hoạt động chuyên biệt này có thể đi kèm với chi phí liên quan rất cao, do đó làm tăng giá trị "adr" lên đáng kể.
- Sau khi hiểu được tác động của các yêu cầu đặc biệt lên doanh thu của khách sạn, các quản lý khách sạn có thể tổng hợp tất cả yêu cầu đặc biệt từ những hành khách của mình trong quá khứ. Từ đó phân tích, xem xét những yêu cầu được xuất hiện phổ biến và đưa yêu cầu đó thành một trong những dịch vụ mà khách sạn cung cấp sẵn. Vì yêu cầu được xuất hiện nhiều lần đồng nghĩa với việc các hành khách cũ có vẻ quan tâm nhiều đến dịch vụ đó, và có thể các hành khách mới trong tương lai cũng sẽ thích dịch vụ này. Do đó, việc cung cấp sẵn dịch vụ từng được nhiều người yêu cầu sẽ trở thành một ưu điểm lớn trong việc thu hút các hành khách mới đến thuê phòng ở khách sạn. Đó chính là một món vũ khí bí mật giúp gia tăng sức cạnh tranh của khách sạn trên thương trường vô cùng khốc liệt.

7.4. Tính tỷ trọng theo bin chia theo thể loại với hai biến cate

7.4.1. Phân tích tỷ trọng theo bin của thuộc tính "adr" chia theo thể loại với hai biến cate là "hotel" và "assigned_room_type"



Biểu đồ 7.4.1: Biểu đồ thể hiện tỷ trọng theo bin của thuộc tính "adr" chia theo thể loại với hai biến cate là "hotel" và "assigned_room_type".

Nhận xét:

(1) Xét nhóm "City Hotel" (với 8 mã phòng khác nhau):

- Các phòng có mã phòng loại "A" sẽ chủ yếu đem lại doanh thu trong khoảng (-7.001; 123.0] cho khách sạn. Các phòng loại "D" cũng có tỷ lệ tạo ra doanh thu trong khoảng (-7.001; 123.0] khá ấn tượng nhưng vẫn không thể so sánh với phòng loại "A". Đúng như ta dự đoán, lý do mà phòng loại "A" có tỷ lệ đặt phòng rất cao là do nó có chi phí rẻ nhất, đem lại lợi ích kinh tế tốt nhất cho khách hàng.

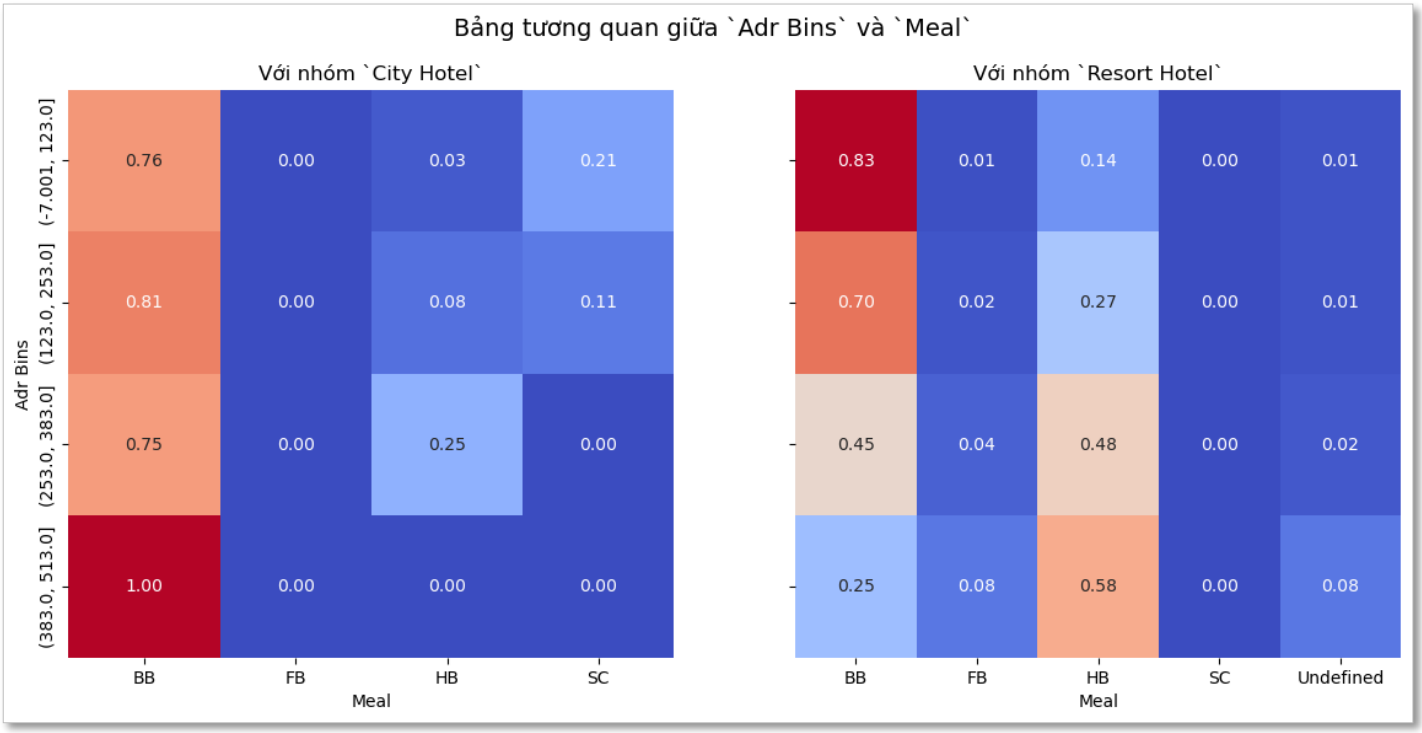
- Ở mức doanh thu trong khoảng (123.0; 253.0], ta thấy hai phòng loại "A" và "D" vẫn tiếp dẫn đầu trong cuộc đua doanh thu nhưng tỷ lệ giữa hai loại phòng này không có quá nhiều sự chênh lệch (khoảng 40%).
- Các phòng loại "G" thường đem lại doanh thu trong khoảng (253.0; 383.0]. Bên cạnh đó, các phòng loại "F" cũng đóng góp không ít vào mức doanh thu này cho khách sạn.
- Ở mức doanh thu cao nhất, khoảng (383.0; 513.0]: ta thấy diễn ra cuộc đua song mã giữa hai loại phòng "E" và "G". Các mã phòng khác hoàn toàn không thể tạo ra mức doanh thu khổng lồ như thế này cho khách sạn.
- Như vậy, các phòng loại "A" và loại "D" thường được lựa chọn bởi đa số khách hàng do các loại phòng này thường có chi phí khá dễ chịu, phù hợp với khả năng chi trả của đa số khách du lịch. Việc tăng cường, bổ sung các phòng loại "A" và "D" có thể giúp thu hút nhiều khách du lịch đến đặt phòng khách sạn, đặc biệt là trong các dịp cao điểm du lịch.
- Các phòng "E", "F", "G" tuy không được sử dụng quá phổ biến nhưng lại có khả năng tạo ra nguồn thu nhập khổng lồ cho khách sạn. Do đó, việc tiếp tục tăng cường chính sách chăm sóc cho các khách hàng đặt các phòng loại này sẽ giúp công ty tạo ra những bước nhảy vọt trong doanh thu.
- Với các mã phòng không được đề cập ở trên như "B", "C" và "K", ta thấy khách hàng không thực sự dành nhiều sự quan tâm cho các loại phòng này. Ta có thể thực hiện các khảo sát thu thập ý kiến khách hàng để có thể thay đổi các loại phòng này, nhằm thu hút khách hàng sử dụng các loại phòng này nhiều hơn.

(2) Xét nhóm "Resort Hotel" (với 10 mã phòng khác nhau):

- Các phòng loại "A", "D" và "E" là những quân bài đắt lực giúp khách sạn tạo ra nguồn thu nhập ổn định trong khoảng hai khoảng (-7.001; 123.0] và (123.0; 253.0]. Tuy các phòng loại "A" có được khách hàng ưu ái hơn một chút nhưng sự chênh lệch giữa chúng là không đáng kể.

- Ở các mức giá đắt đỏ dành cho các dịch vụ cao cấp, ta thấy các phòng loại "G" và "H" thường được khách hàng lựa chọn. Các phòng loại "G" có khả năng tạo ra doanh thu đồng đều ở hai khoảng (253.0; 383.0] và (383.0; 513.0]. Trong khi, các phòng loại "H" phần lớn đem lại doanh thu trong khoảng (383.0; 513.0].
- Như vậy, việc tiếp tục duy trì và nâng cao chất lượng dịch vụ trong các loại phòng "A", "D" và "E" vẫn là ưu tiên hàng đầu để khách sạn tạo ra nguồn thu nhập ổn định. Việc đầu tư vào chính sách chăm sóc khách hàng sử dụng các loại phòng cao cấp như "G" và "H" là cơ hội tốt để khách sạn tạo ra sự nhảy vọt trong doanh thu.
- Đồng thời, việc tiến hành phân tích và đưa ra các biện pháp xử lý phù hợp đối với các phòng ít được sử dụng như "I" và "L" có thể giúp thu hút khách hàng sử dụng các loại phòng này nhiều hơn.

7.4.2. Phân tích tỷ trọng theo bin của thuộc tính "adr" chia theo thể loại với hai biến cate là "hotel" và "meal"



Biểu đồ 7.4.2: Biểu đồ thể hiện tỷ trọng theo bin của thuộc tính "adr" chia theo thể loại với hai biến cate là "hotel" và " meal ".

Nhận xét:

(1) Xét nhóm "City Hotel":

- Ta thấy hầu hết khách hàng chỉ sử dụng bữa sáng ("BB") tại khách sạn và họ thường không có nhu cầu sử dụng các loại bữa ăn khác như "FB" và "HB".
- Như vậy, việc tập trung đầu tư vào chất lượng phục vụ bữa sáng tại nhà hàng của khách sạn sẽ là ưu tiên hàng đầu trong việc thu thú khách hàng quay trở lại khách sạn vào các lần du lịch tiếp theo.
- Đồng thời, khác sạn cũng có thể tiến hành khảo sát để tìm hiểu lý do vì sao khách hàng lại không ưa chuộng các loại bữa ăn khác (chẳng hạn như: cách chế biến không đẹp hay khẩu vị không phù hợp với khách hàng). Sau khi hiểu được lý do, khách sạn

có thể đưa ra các chính sách thay đổi để giúp nâng cao hiệu hoạt động tổng thể của khách sạn.

(2) Xét nhóm "Resort Hotel":

- Ta thấy các khách hàng tạo ra doanh thu trong hai khoảng đầu tiên là (-7.001; 123.0] và (123.0; 253.0] thường chỉ lựa chọn sử dụng bữa sáng tại khách sạn. Chỉ một ít trong số đó mới quyết định sử dụng thêm một bữa ăn khác được cung cấp từ nhà hàng của khách sạn. Trong phạm vi hai bin đầu tiên, ta thấy loại bữa ăn "BB" chiếm tỷ trọng lớn hơn đáng kể so với loại bữa ăn "HB".
- Tuy nhiên, ở nhóm khách hàng có thể tạo ra doanh thu cao (giá trị "adr" > 253) ta lại thấy một xu hướng ngược lại. Phần lớn khách hàng trong nhóm này sẽ lựa chọn việc dùng bữa sáng và một bữa ăn khác tại khách sạn. Tỷ trọng những khách hàng chỉ lựa chọn ăn sáng tại khách sạn và sẽ dùng bữa trưa và tối tại một nơi khác đã giảm đi đáng kể. Đây là một minh chứng cho thấy chất lượng nấu nướng ở các "Resort Hotel" là không phải bàn cải.
- Như vậy, việc đầu tư nhiều hơn vào chất lượng nấu ăn của nhà bếp sẽ là chìa khóa giúp khách sạn có thể thu hút khách du lịch quay lại khu nghỉ dưỡng vào các chuyến đi tiếp theo.
- Tuy nhiên, ta thấy loại bữa ăn trọn gói "FB" lại có tỷ trọng rất thấp trong tất cả các bin. Điều này cho thấy khách hàng không thích chỉ dùng món ăn tại nhà hàng của khách sạn. Thay vào đó, họ sẽ thích đi chơi vào buổi sáng và đến trưa sẽ dùng bữa ở một nhà hàng khác. Các người chủ khách sạn có thể đầu tư vào các chính sách ưu đãi để khuyến khích khách hàng dùng bữa nhiều hơn tại khách sạn. Đây có thể là một cách để giúp nâng cao doanh thu cho khách sạn.

8. Insight

Trong phần này, ta sẽ tóm tắt lại tất cả insight và kết luận mà ta đã rút ra từ những phân tích được thực hiện ở các phần trước đó.

8.1. Data Understanding

- Bộ dữ liệu gốc có 119390 dòng và 32 cột.
- Quan sát bảng dữ liệu, ta thấy có vẻ như không có dòng nào "lạc loài" (hay bất thường).
- Bộ dữ liệu thô có khá nhiều dòng bị trùng lặp (31994 dòng), tương ứng với tỷ lệ hơn 26%.
- Bộ dữ liệu sau khi loại bỏ các dòng bị trùng lặp có số cột không đổi (32) và số dòng giảm xuống còn 87396 (= 119390 - 31994).
- Có bốn cột bị thiếu giá trị là: "company" (khoảng 94%), "agent" (khoảng 14%), "country" (khoảng 0.5%) và "children" (khoảng 0.005%). Trong khi tình trạng thiếu giá trị ở ba cột "agent", "country" và "children" chỉ dừng lại ở mức trung bình - nhẹ, thì cột "company" gặp phải tình trạng thiếu giá trị rất nghiêm trọng.
- Ta phát hiện có khá nhiều cột có kiểu dữ liệu chưa phù hợp và ta cần phải tiền xử lý trước khi tiến hành các bước phân tích tiếp theo.
- Trong bộ dữ liệu mà ta đang xem xét, nhóm thuộc tính số bao gồm 13 cột.
- Ta xử lý tình trạng thiếu giá trị của cột "children" bằng cách điền giá trị yếu vị (mode) của cột vào những vị trí bị thiếu. Sau đó, ta sẽ chuyển cột "children" sang kiểu dữ liệu dạng số "int64" để phù hợp với mô tả của cột này.
- Trong bộ dữ liệu mà ta đang xem xét, nhóm thuộc tính phân loại bao gồm 18 cột.
- Cột "company" có tỷ lệ thiếu giá trị rất cao (hơn 90%). Với tình trạng thiếu dữ liệu nghiêm trọng như thế này thì có vẻ như "company" không phải là một thuộc tính đủ chất lượng để ta có thể phân tích và đưa ra các kết luận có độ tin cậy cao. Do đó, ta sẽ loại bỏ cột này khỏi bộ dữ liệu.

- Cột "agent" có tỷ lệ thiếu giá trị ở mức trung bình (khoảng 14%), không quá nghiêm trọng. Do đó, ta sẽ chọn phương pháp điền giá trị "-1" vào những vị trí bị thiếu.
- Cột "country" có tỷ lệ thiếu giá trị rất thấp, chỉ khoảng 0.5% (không đáng kể). Do đó, ta sẽ tiền xử lý cột "country" bằng phương pháp giống với khi tiền xử lý cột "agent" nhưng thay giá trị "-1" bằng giá trị "others" để phù hợp với ngữ cảnh của dữ liệu.
- Sau quá trình tiền xử lý dữ liệu, số lượng thuộc tính phân loại đã giảm từ 18 xuống còn 17 (do ta loại bỏ cột "company").
- Bộ dữ liệu sau quá trình tiền xử lý có 87396 dòng và 31 cột.

8.2. EDA 1D

8.2.1. Phân tích tỷ lệ cho các biến định tính (categorical)

Thông qua việc phân tích tỷ lệ đối với thuộc tính "hotel":

- "City Hotel" chiếm tỷ lệ lớn hơn khá nhiều so với "Resort Hotel" (61.1% của "City Hotel" so với 38.9% của "Resort Hotel"). Điều này cho thấy "City Hotel" có thể có độ phổ biến cao hơn "Resort Hotel" trong ngành khách sạn hoặc ít nhất là trong bộ dữ liệu này. Điều này có thể xuất phát từ các yếu tố về địa lý và mục đích đi du lịch của khách hàng.

Thông qua việc phân tích tỷ lệ đối với thuộc tính "arrival_date_month":

- Các tháng vào mùa hè (từ tháng 6 đến tháng 8) là thời điểm có rất nhiều người đi du lịch nên số lượng khách đến nhận phòng khách sạn là rất lớn (thường là nhiều nhất trong năm).
- Ngược lại, các tháng vào mùa đông (từ tháng 11 năm trước đến tháng 1 năm sau) có thể xem là thời điểm "đóng băng" của ngành du lịch khi số lượng người đi du lịch là rất ít (thường là ít nhất trong năm).
- Như vậy, các yếu tố như mùa, thời tiết, v.v. cũng có phần nào ảnh hưởng đến quyết định đi du lịch và đặt phòng khách sạn.

Thông qua việc phân tích tỷ lệ đối với thuộc tính "reserved_room_type":

- Loại phòng được khách hàng yêu cầu nhiều nhất là: "A", "D" và "E". Dựa trên nhu cầu thị trường, việc gia tăng số lượng phòng thuộc loại "A", "D" và "E" tại các khách sạn là giải pháp hiệu quả để đáp ứng nhu cầu đa dạng của du khách, đồng thời tối ưu hóa doanh thu, đặc biệt trong các mùa cao điểm du lịch.
- Việc các loại phòng này thường được khách hàng ưa chuộng có thể đến từ việc giá phòng rẻ hơn so với các mã phòng khác nhưng vẫn đáp ứng được nhu cầu nghỉ dưỡng cơ bản cho khách hàng. Tuy nhiên ta cần có những phân tích chi tiết hơn để làm rõ giả thuyết này.

Thông qua việc phân tích tỷ lệ đối với thuộc tính "deposit_type":

- Hầu hết khách hàng đều không đặt cọc khi đặt phòng khách sạn (No Deposit). Nếu đã quyết định đặt cọc, thì hầu hết khách hàng sẽ lựa chọn trả trước toàn bộ chi phí lưu trú trong chuyến đi (Non Refund) và có rất ít khách hàng lựa chọn phương án trả trước một phần chi phí lưu trú (Refundable).
- Khách sạn có thể tạo ra chính sách ưu đãi dành cho các khách hàng đặt cọc trước, đặc biệt là các khách hàng lựa chọn phương án trả trước toàn bộ chi phí lưu trú (Non Refund). Vì đây là nhóm khách hàng thường có mức độ cam kết cao, nên việc triển khai các chương trình ưu đãi như: giảm giá, tặng quà, nâng hạng phòng, v.v. sẽ khuyến khích khách hàng đặt cọc, tăng tỷ lệ đặt phòng và doanh thu cho khách sạn.

Thông qua việc phân tích tỷ lệ đối với thuộc tính "country":

- Các khách hàng đến từ đất nước "Portugal" (Bồ Đào Nha) có tỷ lệ đặt phòng cao nhất với tỷ lệ hơn 30%. Xếp ngay sau đó là các khách hàng đến từ "Great Britain" (Đảo Anh) với tỷ lệ khoảng 12% và ở trí thứ ba là đất nước "France" (Pháp) với tỷ lệ hơn 10%.
- Ta phát hiện bộ dữ liệu này chủ yếu được thu thập từ các khách sạn ở châu Âu khi số lượng quốc gia đến từ châu Âu chiếm hơn một nửa số mẫu dữ liệu ta quan sát được. Như vậy, các khách sạn có thể tạo ra các chính sách ưu đãi dành cho các khách hàng đến từ các quốc gia thuộc châu Âu để có thể thu hút thêm các khách hàng đến nghỉ ngơi tại khách sạn của họ.

8.2.2. Phân tích phân phối cho các biến định lượng (numerical)

Thông qua việc phân tích phân phối đối với thuộc tính "lead_time":

- Ta thấy thuộc tính "lead_time" có phân bố lệch phải khá nặng và có phần đuôi rất đậm. Các điểm dữ liệu tập trung khá nhiều ở các khoảng giá trị nhỏ và thưa thớt dần khi tiến về phần đuôi phía bên phải.
- Ta có thể thấy rằng khoảng thời gian chênh lệch giữa ngày đặt phòng và ngày đến nhận phòng của khách hàng thường không quá lớn. Điều này phản ánh rằng trong phần lớn các trường hợp, khách hàng thường đặt phòng trong khoảng thời gian ngắn trước khi đến ngày nhận phòng. Điều này có thể là do họ thích lên kế hoạch gần với thời điểm thực hiện chuyến đi hoặc có sự linh hoạt trong việc thay đổi kế hoạch.

Thông qua việc phân tích phân phối đối với thuộc tính "adr":

- Thuộc tính "adr" có phân bố lệch phải rất nghiêm trọng và có phần đuôi cực kỳ đậm.
- Doanh thu trung bình của các khách sạn trong bộ dữ liệu này thường không quá cao và tập trung chủ yếu ở mức trung bình. Điều này có thể cung cấp một cơ sở mạnh mẽ cho nhận định về việc mã phòng loại "A" được lựa chọn nhiều vì có chi phí rẻ mà ta đã đề cập bên trên.
- Có một số lượng không nhỏ các mẫu dữ liệu nằm lệch về phía bên phải của phân phối. Đó chính là các khách hàng tiềm năng sẵn sàng chi tiêu nhiều hơn cho các dịch vụ của khách sạn. Do đó, khách sạn một mặt cần tiếp tục duy trì chất lượng cho các dịch vụ có chi phí ở mức trung bình - khá để tạo ra nguồn doanh thu ổn định từ đại đa số các khách hàng đến đặt phòng. Mặt khác, khách sạn cũng cần nâng cao chính sách chăm sóc các khách hàng sử dụng dịch vụ cao cấp, vì đây chính là cơ hội để tạo ra nguồn thu nhập khổng lồ cho khách sạn.

Thông qua việc phân tích phân phối đối với thuộc tính "total_of_special_requests":

- Thuộc tính "total_of_special_requests" có giá trị nhỏ nhất là 0 (tức là khách hàng không có yêu cầu đặc biệt) và giá trị lớn nhất là 5. Phần lớn các điểm dữ liệu sẽ tập

trung ở các giá trị nhỏ như 0 và 1. Khi số lượng yêu cầu vượt qua giá trị 1 thì số lượng quan sát sẽ giảm đi đáng kể.

- Tuy nhiên, số lượng yêu cầu đặc biệt ("total_of_special_requests") chỉ mang tính tổng quát giúp ta nhận biết xu hướng chung trong việc đặt phòng của khách hàng chứ không mô tả nội dung chi tiết của các yêu cầu. Do đó, ta cần phải biết nội dung chi tiết hoặc ít nhất là biết được yêu cầu thuộc vào nhóm nào để có thể đưa ra các phân tích cụ thể, chi tiết hơn.

8.3. EDA 2D

8.3.1. Phân tích hệ số tương quan giữa các biến định lượng (numerical)

- Ta thấy không có mối quan hệ tương quan "hoàn hảo" giữa hai thuộc tính khác nhau. Giá trị hệ số tương quan Pearson giữa các cặp thuộc tính số trong bộ dữ liệu ta đang phân tích có giá trị nằm trong khoảng (-0.2; 0.6).
- Nổi bật nhất là mối tương quan thuận khá mạnh giữa hai thuộc tính "stays_in_weekend_nights" và "stays_in_week_nights" (với giá trị của hệ số tương quan là 0.56).
- Ta có thể thấy rằng mức độ tương quan giữa các thuộc tính số không quá cao, đa phần dừng ở mức trung bình - thấp. Điều này cho thấy các biến thường độc lập với nhau, không có sự phụ thuộc lẫn nhau quá nhiều. Ta sẽ cần thực hiện thêm nhiều phân tích để có nhìn nhìn chính xác hơn về mối quan hệ giữa các biến.

8.3.2. Sử dụng Scatter plot để phân tích dữ liệu 2D

Thông qua việc phân tích mối quan hệ giữa "stays_in_weekend_nights" và "stays_in_week_nights":

- Khách hàng có xu hướng sử dụng cả hai dịch vụ ở qua đêm các ngày trong tuần và ở qua đêm các ngày cuối tuần. Cả hai dịch vụ có xu hướng đồng biến với nhau nhưng không hoàn toàn tuyến tính.
- Khi khách hàng ở qua đêm các ngày trong tuần thì đồng thời cũng ở qua đêm các ngày cuối tuần, nhưng ở qua đêm các ngày cuối tuần tăng ít hơn. Do thường vào các ngày cuối tuần giá tiền phòng khách sạn sẽ tăng cao hơn so với các ngày trong tuần vì thế khách hàng ít đặt phòng vào dịp cuối tuần hơn.

Thông qua việc phân tích mối quan hệ giữa "days_in_waiting_list" và "adr":

- Ta thấy hai thuộc tính "days_in_waiting_list" và "adr" có mối tương quan nghịch không quá rõ ràng. Nhìn chung, khi giá trị của thuộc tính "days_in_waiting_list" tăng lên thì giá trị của thuộc tính "adr" sẽ có xu hướng giảm xuống (và ngược lại).

- Điều này phản ánh một xu hướng là: khi một lượt đặt phòng có thời gian nằm trong danh sách chờ càng ngắn thì khách hàng sẽ sẵn lòng chi trả nhiều hơn cho dịch vụ của khách sạn, khi này doanh thu trung bình của khách sạn sẽ tăng lên. Điều này cũng không quá khó hiểu vì khi khách hàng phải chờ đợi lâu thì tâm trạng của họ cũng ít nhiều bị ảnh hưởng.
- Khách sạn có thể cải thiện chính sách chăm sóc khách hàng của mình bằng cách tạo ra một số ưu đãi nhỏ dành tặng cho các khách hàng nằm trong danh sách chờ quá lâu. Điều này có thể phần nào cải thiện tâm trạng của khách hàng, giúp cuộc trò chuyện giữa hai bên vui vẻ hơn. Khi này có thể khách hàng sẽ sẵn sàng lựa chọn các gói dịch vụ cao cấp cho chuyến đi du lịch, giúp công ty thu được nhiều lợi nhuận hơn.

8.3.3. Sử dụng bar chart để phân tích dữ liệu "numerical" và "categorical"

Thông qua việc phân tích số lượng đặt phòng tại các loại khách sạn theo các tháng trong năm:

- Xu hướng đặt phòng khách sạn của khách hàng tăng dần từ tháng 1 đến tháng 8, cao điểm nhất là tháng 8. Và giảm dần từ tháng 9 đến tháng 1 năm sau. Do thời điểm mùa hè và thu là thời điểm thích hợp cho mọi người đi du lịch nên lượng nhu cầu tăng cao, các khách sạn nên tăng cường dịch vụ vào thời gian nàt để tăng trải nghiệm khách hàng. Đồng thời giảm giá, thu hút khách vào các thời điểm ít khách nhằm tối ưu hóa lợi nhuận.

Thông qua việc phân tích tổng số tiền mà tất cả các khách sạn thu được trong các năm:

- "City Hotel" luôn được yêu thích và lựa chọn nhiều hơn so với "Resort Hotel" trong mọi tháng trong năm. Có thể mô hình "City Hotel" có nhiều ưu điểm hơn nên được đặt phòng nhiều hơn, các khách sạn nên tiếp thu và tìm hiểu về mô hình khách sạn này.

Thông qua việc phân tích thời gian đặt phòng giữa hai loại khách sạn:

- Thời gian đặt phòng trung vị ("median_lead_time") giữa hai loại khách sạn không có quá nhiều sự chênh lệch. Tuy nhiên, các "City Hotel" vẫn thường có thời gian đặt phòng sớm hơn một chút so với "Resort Hotel".
- Cả hai loại khách sạn đều có thời gian đặt phòng trung vị vào khoảng sớm hơn 50 ngày so với ngày chính thức nhận phòng. Ta nhận thấy đây là một giá trị khá lớn, cho thấy các khách hàng dù đến loại khách sạn nào thì cũng thường lên kế hoạch từ khá sớm (khoảng trước 2 tháng). Đây cũng là một xu hướng dễ hiểu vì nếu không đặt phòng khách sạn từ sớm thì rất dễ xảy ra khả năng đến lúc ta tới nơi thì không còn phòng nào trống để thuê.

Thông qua việc phân tích phân tích số ngày trong danh sách chờ khi đặt phòng giữa hai loại khách sạn:

- Nhìn chung, khách hàng của các "City Hotel" thường có thời gian trung bình trong danh chờ lâu hơn so với các khách hàng lựa chọn nghỉ dưỡng tại các "Resort Hotel". Điều này cho thấy các "City Hotel" thường "bận rộn" hơn so với các "Resort Hotel".
- Các khách sạn "Resort Hotel" có chi phí đắt đỏ nên thường có ít khách hàng đến đặt phòng. Trong khi chi phí phải trả cho các "City Hotel" thường thoải mái hơn nên tỷ lệ người sử dụng (và muốn sử dụng) "City Hotel" sẽ nhiều hơn đáng kể. Có thể vì điều này đã làm cho các "City Hotel" thường rơi vào tình trạng quá tải và làm gia tăng thời gian khách hàng ở trong danh sách chờ khi tiến hành đặt phòng.

Thông qua việc phân tích tỷ lệ hủy đặt phòng giữa hai loại khách sạn:

- Ta thấy rằng "City Hotel" vừa có tỷ lệ đặt phòng cao hơn vừa có tỷ lệ hủy đặt phòng cao hơn "Resort Hotel". Cụ thể, theo kết quả thống kê từ dữ liệu, tỷ lệ hủy đặt phòng của "City Hotel" là khoảng 30%, nhiều hơn tương đối so với tỷ lệ chưa đến 25% của "Resort City".
- Việc cả hai loại khách sạn đều có tỷ lệ hủy đặt phòng khá lớn cũng cho thấy tình trạng hủy đặt phòng không thực sự tập trung ở một nhóm khách hàng cụ thể nào. Điều này

đặt ra một thách thức, một bài toán kinh tế cần giải quyết để hạn chế tình trạng hủy đặt phòng và tối đa hóa doanh thu cho khách sạn.

8.3.4. Tính tỷ trọng đối với hai biến "categorical"

Thông qua việc phân tích tỷ trọng loại phòng được đặt tại các loại khách sạn khác nhau:

- Các loại phòng "A" và "D" luôn được yêu thích nhất ở các loại khách sạn khác nhau, do giá cả phù hợp, dịch vụ phòng tiện ích. Vì thế, khách sạn nên tăng cường số lượng phòng loại "A" và "D" để phục vụ như cầu khách hàng, cũng như cải thiện các loại phòng khác để đảm bảo sự đa dạng cho khách sạn.

Thông qua việc phân tích tỷ trọng hủy đặt phòng ở các nhóm khách hàng khác nhau:

- Nhóm khách hàng "transient" có tỉ lệ hủy phòng cao do tính tạm thời, không có kế hoạch từ trước vì thế rủi ro hủy phòng cao. Khách sạn nên chú ý đến nhóm khách hàng này để đảm bảo quyền lợi cho khách sạn đồng thời tạo thuận lợi cho khách hàng, ví dụ như: đặt cọc, dời ngày đặt phòng, v.v..
- Đối với các nhóm khách hàng còn lại với tỉ lệ hủy phòng không quá cao thì khách sạn nên tạo những sự kiện thu hút, giữ chân khách hàng hơn là tập trung vào tỉ lệ hủy phòng của các nhóm khách hàng này.

8.4. EDA 3D

8.4.1. Sử dụng Scatter plot để phân tích dữ liệu 3D cho ba biến định lượng (numerical)

Thông qua việc phân tích mức độ tương quan giữa ba biến num "adr", "total_people" và "lead_time":

- Ta thấy có mối tương quan theo chiều dương không quá mạnh giữa hai thuộc tính "adr" và "total_people". Nhìn chung, khi số lượng người trong nhóm hành khách tăng lên thì khách sạn cũng thu được nhiều lợi nhuận hơn, do đó giá trị "adr" cũng cao hơn.
- Khi số lượng người trong nhóm hành khách vượt qua con số 5 thì lợi nhuận mà khách sạn thu được thường không quá ấn tượng, giá trị "adr" thường thấp hơn 100.
- Ta thấy các nhóm hành khách đi từ 1 đến 5 người là nhóm khách hàng chủ yếu và đóng góp rất nhiều vào nguồn doanh thu của khách sạn.
- Mối tương quan tuyến tính giữa hai thuộc tính "adr" và "lead_time" không quá rõ ràng.
- Trong trường hợp khoảng thời gian giữa ngày đặt phòng và ngày nhận phòng vượt quá một năm, khi thời gian kéo dài càng lâu, ta thấy doanh thu của khách sạn có xu hướng giảm xuống.
- Ta thấy có mối tương quan theo chiều dương khá yếu giữa hai thuộc tính "total_people" và "lead_time". Nhìn chung, khi số lượng người trong nhóm hành khách tăng lên thì họ cũng có xu hướng đặt phòng sớm hơn ("lead_time" lớn). Điều này cũng không quá khó hiểu vì các nhóm hành khách đông người thường sẽ là: một đại gia đình, một nhóm bạn bè hoặc là các nhân viên trong một công ty, v.v.. Khi đi du lịch đông người mà ta không đặt hẹn với khách sạn từ sớm thì sẽ rất dễ xuất hiện tình trạng thiếu phòng và làm ảnh hưởng đến kế hoạch của cả nhóm.

8.4.2. Sử dụng Scatter plot 2D và màu đối với hai biến num và cate

Thông qua việc phân tích mức độ tương quan giữa hai thuộc tính "previous_cancellations" và "previous_bookings_not_canceled" theo từng loại khách sạn:

- Đối với các hành khách chưa từng hủy đặt phòng hoặc có số lần hủy đặt phòng ít hơn 10 lần, ta thấy hai thuộc tính "previous_cancellations" và "previous_bookings_not_canceled" có mối tương quan theo chiều dương khá mạnh. Nhìn chung, các khách hàng đã nhiều lần hủy đặt phòng thì cũng có nhiều lần không hủy đặt phòng (tức là họ vẫn đến ở khách sạn như lịch hẹn từ trước). Do đó, việc hủy đặt phòng trong trường hợp này không thực sự đồng nghĩa với tình trạng hành khách đang "rời bỏ" khách sạn.
- Đối với các hành khách có số lần hủy đặt phòng nhiều hơn 10 lần, ta thấy hai thuộc tính "previous_cancellations" và "previous_bookings_not_canceled" có mối tương quan theo chiều âm không quá mạnh.
- Trung bình "previous_cancellations" của "City Hotel" cao hơn trung bình "previous_cancellations" của "Resort Hotel". Điều này cho thấy tình trạng hành khách hủy đặt phòng sẽ diễn ra phổ biến hơn ở các khách sạn "City Hotel". Ngược lại, "Resort Hotel" sẽ có số lần hủy đặt phòng bởi cùng một khách hàng nhiều hơn.
- Đồng thời, trung bình "previous_bookings_not_canceled" của "City Hotel" cũng cao hơn trung bình "previous_bookings_not_canceled" của "Resort Hotel". Điều này cho thấy các hành khách đã ở khách sạn thuộc loại "City Hotel" thường có xu hướng quay trở lại khách sạn này vào các lần tiếp theo. Do đó, các khách sạn thuộc nhóm "City Hotel" có thể tạo ra chính sách ưu đãi dành cho khách hàng thân thiết để khuyến khích các khách hàng cũ tiếp tục sử dụng dịch vụ khách sạn, từ đó tạo ra doanh thu tốt hơn.

Thông qua việc phân tích mức độ tương quan giữa hai thuộc tính "total_stays" và "lead_time" theo từng loại khách sạn:

- Nhìn chung, hai thuộc tính "total_stays" và "lead_time" có mối tương quan nhẹ theo chiều dương. Nghĩa là, nếu các hành khách có ý định đi du lịch thì họ thường có xu hướng lên kế hoạch và đặt phòng khách sạn từ rất sớm để tránh tình trạng hết phòng, dẫn đến giá trị "lead_time" trong các mẫu dữ liệu này cũng lớn hơn.
- Phần lớn hành khách ở các "City Hotel" thường có thời gian lưu trú trong khoảng 10 ngày. Đồng thời, ta cũng phát hiện rằng khoảng thời gian kể từ ngày đặt phòng đến ngày nhận phòng của các hành khách này thường không vượt quá một năm (365 ngày).
- Tuy nhiên, ta thấy xuất hiện khá nhiều hành khách có thời gian đặt phòng sớm hơn ít nhất 400 ngày. Đây là một nhóm các khách hàng thú vị mà ta nên dành thêm thời gian để phân tích lý do vì sao mà họ lại đặt phòng từ rất sớm như vậy.
- Thời gian lưu trú của hành khách ở các khách sạn thuộc loại "Resort Hotel" thường sẽ kéo dài lâu hơn so với "City Hotel". Phần lớn hành khách ở các "Resort Hotel" sẽ có thời gian nghỉ dưỡng kéo dài trong khoảng nửa tháng (15 ngày). Tuy nhiên, cũng có rất nhiều hành khách lựa chọn lưu trú tại khách sạn từ 1 đến 2 tháng. Điều này cũng hoàn toàn hợp lý với mục đích "nghỉ dưỡng" như trong tên gọi của loại khách sạn này (Resort Hotel).
- So với "City Hotel", thời gian đặt phòng của hành khách ở các khách sạn "Resort Hotel" thường có độ dao động không quá lớn.

Thông qua việc phân tích mức độ tương quan giữa hai thuộc tính "total_stays" và "adr" theo từng loại khách sạn:

- Khi giá trị của thuộc tính "total_stays" tăng lên thì giá trị của thuộc tính "adr" có xu hướng giảm xuống. Nghĩa là, khi các hành khách có thời gian lưu trú lâu hơn tại khách sạn thì lợi nhuận mà khách sạn thu được sẽ có xu hướng giảm xuống.
- Câu "lợi nhuận có xu hướng giảm xuống" không đồng nghĩa với việc khách sạn bị lỗ vốn khi cung cấp dịch vụ cho một khách hàng nào đó. Mà câu bên trên nên được hiểu theo nghĩa là khách sạn thu được số tiền ít hơn (từ hành khác có thời gian lưu trú dài) so với số tiền mà khách sạn kiếm được từ các hành khách có thời gian lưu trú ngắn hơn.
- Các hành khách ở lại khách sạn lâu hơn thường có được các "thỏa thuận" tốt hơn từ phía khách sạn. Do đó, nếu một đại gia đình muốn tổ chức chuyến đi du lịch cho tất cả thành viên thì nên lựa chọn các chuyến đi dài ngày và ở lại một khách sạn lâu hơn để có thể tiết kiệm chi phí.
- Ta thấy hành khách thường lựa chọn các "City Hotel" trong các chuyến đi ngắn ngày. Nhưng đối với các chuyến đi dài ngày, các khách sạn thuộc nhóm "Resort Hotel" thường được ưu ái hơn.

Thông qua việc phân tích mức độ tương quan giữa hai thuộc tính "total_of_special_requests" và "adr" theo từng loại khách sạn:

- Hai thuộc tính "total_of_special_requests" và "adr" có mối tương quan thuận không quá mạnh. Nhìn chung, khi số lượng yêu cầu đặc biệt của hành khách tăng lên thì khách sạn càng kiếm được nhiều doanh thu. Điều này cho thấy khách sạn nên tạo thêm điều kiện và cố gắng đáp ứng càng nhiều yêu cầu đặc biệt từ khách hàng càng tốt. Vì hành khách yêu cầu nhiều dịch vụ đặc biệt có thể sẵn lòng trả giá cao hơn cho chỗ ở của họ. Đây có thể xem là một loại hình dịch vụ đầy hứa hẹn và có thể đóng góp rất nhiều vào doanh thu của cả khách sạn.
- Nhóm "City Hotel" có giá trị "total_of_special_requests" trung bình cao hơn so với giá trị "total_of_special_requests" trung bình của nhóm Resort Hotel. Tuy nhiên, về mặt doanh thu, các khách sạn thuộc nhóm "Resort Hotel" thường có xu hướng kiếm được nhiều tiền hơn từ các yêu cầu đặc biệt của hành khách so với các khách sạn trong nhóm "City Hotel".
- Các quản lý khách sạn có thể tổng hợp các yêu cầu đặc biệt từ hành khách trong quá khứ, sau đó phân tích, xem xét những yêu cầu được xuất hiện phổ biến và đưa yêu cầu đó thành một trong những dịch vụ mà khách sạn cung cấp sẵn. Việc cung cấp sẵn dịch vụ từng được nhiều người yêu cầu sẽ trở thành một ưu điểm lớn trong việc thu hút các hành khách mới. Đó chính là một món vũ khí bí mật giúp gia tăng sức cạnh tranh của khách sạn trên thương trường vô cùng khốc liệt.

8.4.3. Tính tỷ trọng theo bin chia theo thể loại với hai biến cate

Thông qua việc phân tích tỷ trọng theo bin của thuộc tính "adr" chia theo thể loại với hai biến cate là "hotel" và "assigned_room_type":

- Ở nhóm "City Hotel":
  - Các phòng loại "A" và loại "D" thường được lựa chọn bởi đa số khách hàng do các loại phòng này thường có chi phí khá dễ chịu, phù hợp với khả năng chi trả của đa số khách du lịch. Việc tăng cường, bổ sung các phòng loại "A" và "D" có thể giúp thu hút nhiều khách du lịch đến đặt phòng khách sạn, đặc biệt là trong các dịp cao điểm du lịch.
  - Các phòng "E", "F", "G" tuy không được sử dụng quá phổ biến nhưng lại có khả năng tạo ra nguồn thu nhập khổng lồ cho khách sạn. Do đó, việc tiếp tục tăng cường chính sách chăm sóc cho các khách hàng đặt các phòng loại này sẽ giúp công ty tạo ra những bước nhảy vọt trong doanh thu.
  - Với các mã phòng "B", "C" và "K", ta thấy khách hàng không thực sự dành nhiều sự quan tâm cho các loại phòng này. Ta có thể thực hiện các khảo sát thu thập ý kiến khách hàng để có thể thay đổi các loại phòng này, nhằm thu hút khách hàng sử dụng các loại phòng này nhiều hơn.
- Ở nhóm "Resort Hotel":
  - Việc tiếp tục duy trì và nâng cao chất lượng dịch vụ trong các loại phòng "A", "D" và "E" vẫn là ưu tiên hàng đầu để khách sạn tạo ra nguồn thu nhập ổn định.
  - Việc đầu tư vào chính sách chăm sóc khách hàng sử dụng các loại phòng cao cấp như "G" và "H" là cơ hội tốt để khách sạn tạo ra sự nhảy vọt trong doanh thu.
  - Đồng thời, việc tiến hành phân tích và đưa ra các biện pháp xử lý phù hợp đối với các phòng ít được sử dụng như "I" và "L" có thể giúp thu hút khách hàng sử dụng các loại phòng này nhiều hơn.

Thông qua việc phân tích tỷ trọng theo bin của thuộc tính "adr" chia theo thể loại với hai biến cate là "hotel" và "meal":

- Ở nhóm "City Hotel":
  - Ta thấy hầu hết khách hàng chỉ sử dụng bữa sáng ("BB") tại khách sạn và họ thường không có nhu cầu sử dụng các loại bữa ăn khác như "FB" và "HB".
  - Như vậy, việc tập trung đầu tư vào chất lượng phục vụ bữa sáng tại nhà hàng của khách sạn sẽ là ưu tiên hàng đầu trong việc thu thú khách hàng quay trở lại khách sạn vào các lần du lịch tiếp theo.
  - Đồng thời, khác sạn cũng có thể tiến hành khảo sát để tìm hiểu lý do vì sao khách hàng lại không ưa chuộng các loại bữa ăn khác (chẳng hạn như: cách chế biến không đẹp hay khẩu vị không phù hợp với khách hàng). Sau khi hiểu được lý do, khách sạn có thể đưa ra các chính sách thay đổi để giúp nâng cao hiệu hoạt động tổng thể của khách sạn.
- Ở nhóm "Resort Hotel":
  - Ta thấy các khách hàng thường chỉ lựa chọn sử dụng bữa sáng tại khách sạn. Chỉ một ít trong số đó mới quyết định sử dụng thêm một bữa ăn khác được cung cấp từ nhà hàng của khách sạn. Trong phạm vi hai bin đầu tiên, ta thấy loại bữa ăn "BB" chiếm tỷ trọng lớn hơn đáng kể so với loại bữa ăn "HB".
  - Tuy nhiên, ở nhóm khách hàng có thể tạo ra doanh thu cao hơn, ta lại thấy một xu hướng ngược lại. Phần lớn khách hàng trong nhóm này sẽ lựa chọn việc dùng bữa sáng và một bữa ăn khác tại khách sạn. Tỷ trọng những khách hàng chỉ lựa chọn ăn sáng tại khách sạn và sẽ dùng bữa trưa và tối tại một nơi khác đã giảm đi đáng kể. Đây là một minh chứng cho thấy chất lượng nấu nướng ở các "Resort Hotel" là không phải bàn cải.
  - Như vậy, việc đầu tư nhiều hơn vào chất lượng nấu ăn của nhà bếp sẽ là chìa khóa giúp khách sạn có thể thu hút khách du lịch quay lại khu nghỉ dưỡng vào các chuyến đi tiếp theo.

- Tuy nhiên, ta thấy loại bữa ăn trọn gói "FB" lại có tỷ trọng rất thấp trong tất cả các bin. Điều này cho thấy khách hàng không thích chỉ dùng món ăn tại nhà hàng của khách sạn. Thay vào đó, họ sẽ thích đi chơi vào buổi sáng và đến trưa sẽ dùng bữa ở một nhà hàng khác. Các người chủ khách sạn có thể đầu tư vào các chính sách ưu đãi để khuyến khích khách hàng dùng bữa nhiều hơn tại khách sạn. Đây có thể là một cách để giúp nâng cao doanh thu cho khách sạn.

9. Tài liệu tham khảo

[1]: Hotel Bookings Exploratory Data Analysis - github.com/Neerajdataguy.

[2]: Hotel Booking Analysis - EDA Using Python - github.com/ajitmane36.

[3]: Hotel Booking project — Exploratory Data Analysis - medium.com/EthanDuong.

[4]: Hotel Booking Demand EDA/ Data Visualisation - kaggle.com/shrutidandagi.

[5]: Cancelaciones en Hoteles - kaggle.com/competitions/cancelaciones-en-hoteles.

[6]: File notebook được giảng viên cung cấp để làm tài liệu tham khảo cho bài tập này - github.com/truongcntn2017.

[7]: Tham khảo về mô tả của từng cột - Link.

[8]: Tham khảo về phương pháp khám phá và tiền xử lý dữ liệu: Slide bài giảng của môn học "Lập trình cho khoa học dữ liệu".