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 KHÁM PHÁ DỮ LIỆU ĐẶT PHÒNG KHÁCH SẠN

DANH SÁCH THÀNH VIÊN

Họ và tên	MSSV	Email	Tỉ lệ thực hiện
Võ Duy Anh	21127221	vdanh21@clc.fitus.edu.vn	35%
Nguyễn Mậu Gia Bảo	21127583	nmgbao21@clc.fitus.edu.vn	35%
Vũ Minh Phát	21127739	vmphat21@clc.fitus.edu.vn	30%

GIẢNG VIÊN HƯỚNG DẪN: Bùi Tiến Lên

Lê Ngọc Thành

Lê Nguyễn Nhựt Trường

Tp. Hồ Chí Minh, ngày 06 tháng 04 năm 2024

Mục lục

I. Thông tin nhóm

1. Danh sách thành viên và tỉ lệ thực hiện của mỗi thành viên

Họ và tên	MSSV	Email	Tỉ lệ thực hiện
Võ Duy Anh	21127221	vdanh21@clc.fitus.edu.vn	35%
Nguyễn Mậu Gia Bảo	21127583	nmgbao21@clc.fitus.edu.vn	35%
Vũ Minh Phát	21127739	vmphat21@clc.fitus.edu.vn	30%

2. Các câu hỏi chưa làm được

Nhóm đã hoàn thành công việc và trả lời tất cả câu hỏi được giao.

II. Data Understanding

1. Đếm số dòng và số cột

- Sử dụng thuộc tính shape để trả về một tuple chứa số lượng dòng và số lượng cột trong DataFrame.
- Dữ liệu gốc có **119390 dòng** và **32 cột**.

2. Viết bảng mô tả về các cột

Stt	Cột	Ý nghĩa
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.
9	stays_in_week_nights	Số đêm lưu trú vào các ngày trong tuần (thứ hai đến thứ sáu).
10	adults	Số lượng người lớn (được bao gồm trong mỗi lần đặt phòng).
11	children	Số lượng trẻ em (được bao gồm trong mỗi lần đặt phòng).

12	babies	Số lượng em bé (được bao gồm trong mỗi lần đặt phòng).
13	meal	Loại bữa ăn đã đặt.
14	country	Quốc gia của khách hàng.
15	market_segment	Phân khúc thị trường của khách hàng.
16	distribution_channel	Phương thức hoặc kênh mà 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 hủy đặt phòng trước đó.
19	previous_bookings_not_canceled	Số lần đặt phòng trước đó nhưng không bị hủy.
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.
22	booking_changes	Số lượng thay đổi/sửa đổi được thực hiện đối với việc đặt phòng.
23	deposit_type	Loại tiền đặt cọc (phải trả để đặt phòng).
24	agent	ID (mã số) của đại lý đặt phòng.
25	company	ID (mã số) của công ty đặt phòng.
26	days_in_waiting_list	Số ngày mà một lượt đặt phòng đã nằm trong danh sách chờ của khách sạn.
27	customer_type	Loại khách hàng.
28	adr	Giá trung bình hàng đêm (Average Daily Rate).

29	required_car_parking_spaces	Số lượng chỗ đậu xe yêu cầu.
30	total_of_special_requests	Tổng số yêu cầu đặc biệt.
31	reservation_status	Trạng thái đặt phòng cuối cùng.
32	reservation_status_date	Ngày cập nhật trạng thái đặt phòng cuối cùng.

3. Phân tích tỷ lệ missing rate và fill missing rate

- Cột company có tỉ lệ thiếu dữ liệu rất cao, lên đến hơn 90% (93.9826%), do đó nhóm sẽ tiến hành loại bỏ cột này.
- Cột children với tỉ lệ thiếu nhỏ, nhóm điền bằng giá trị mode của thuộc tính này.
- Cột agent với tỉ lệ thiếu khá cao, nhóm điền bằng giá trị median của thuộc tính này do giá trị của nó có phân bố rộng.
- Cột country với tỉ lệ thiếu là 0.5172%, không đáng kể, do đó nhóm sẽ điền giá trị thiếu bằng giá trị mode của cột.

4. Phân tích tỷ lệ duplicate

- Sử dụng duplicated(keep='first') để xem mỗi dòng có bị trùng lặp không. Tham số keep='first' chỉ ra rằng chỉ các dòng được coi là trùng lặp nếu chúng trùng với hàng đầu tiên xuất hiện trước đó.
- detectDupSeries là Series kết quả, trong đó mỗi giá trị True cho biết dòng tương ứng là một dòng trùng lặp.
- Dữ liệu có **31994 dòng bị trùng lặp** sau khi thực hiện hai đoạn code trên để khám phá.
- Thực hiện phép chia để tính tỷ lệ giữa số lượng dòng trùng lặp với tổng số dòng trong DataFrame, ta biết có bao nhiêu phần trăm các dòng trong DataFrame là các dòng bị trùng lặp. Tỷ lệ trùng lặp của dữ liệu là: 26.80%.

III. EDA 1D

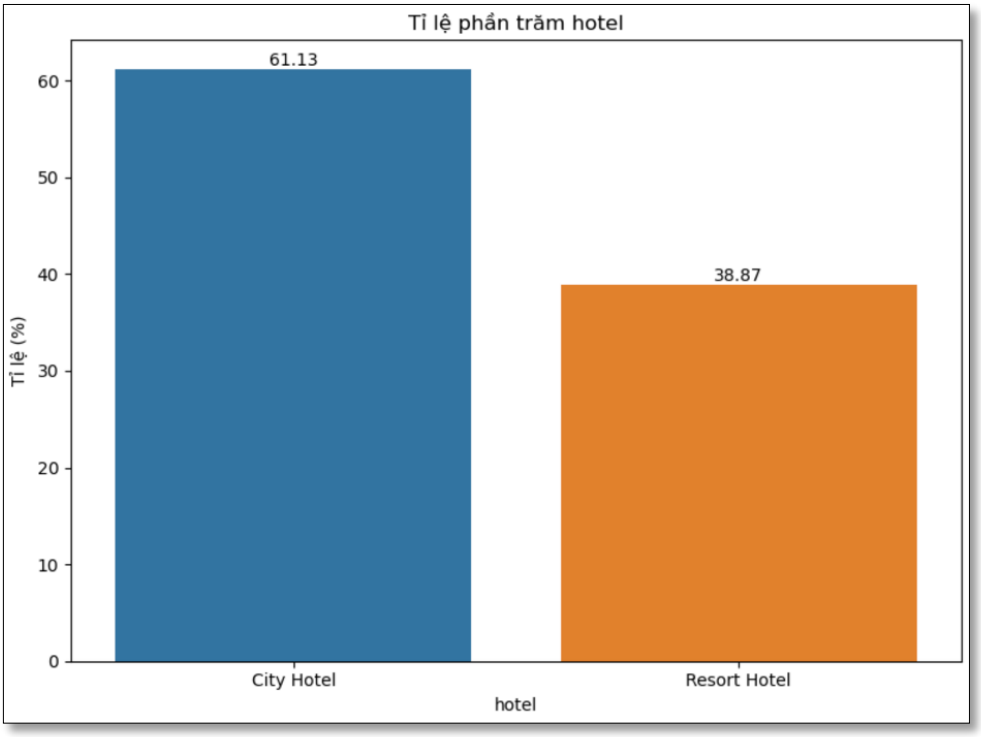
1. Chia dữ liệu theo kiểu Numerical và Categorical

- Sử dụng select_dtypes để chọn các cột dựa trên kiểu dữ liệu của chúng.
- 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 số.
  - cat_col_df là DataFrame chứa các cột chỉ có kiểu dữ liệu không phải số.

2. Phân tích tỷ lệ đối với biến Cate

2.1. Đối với thuộc tính "hotel"

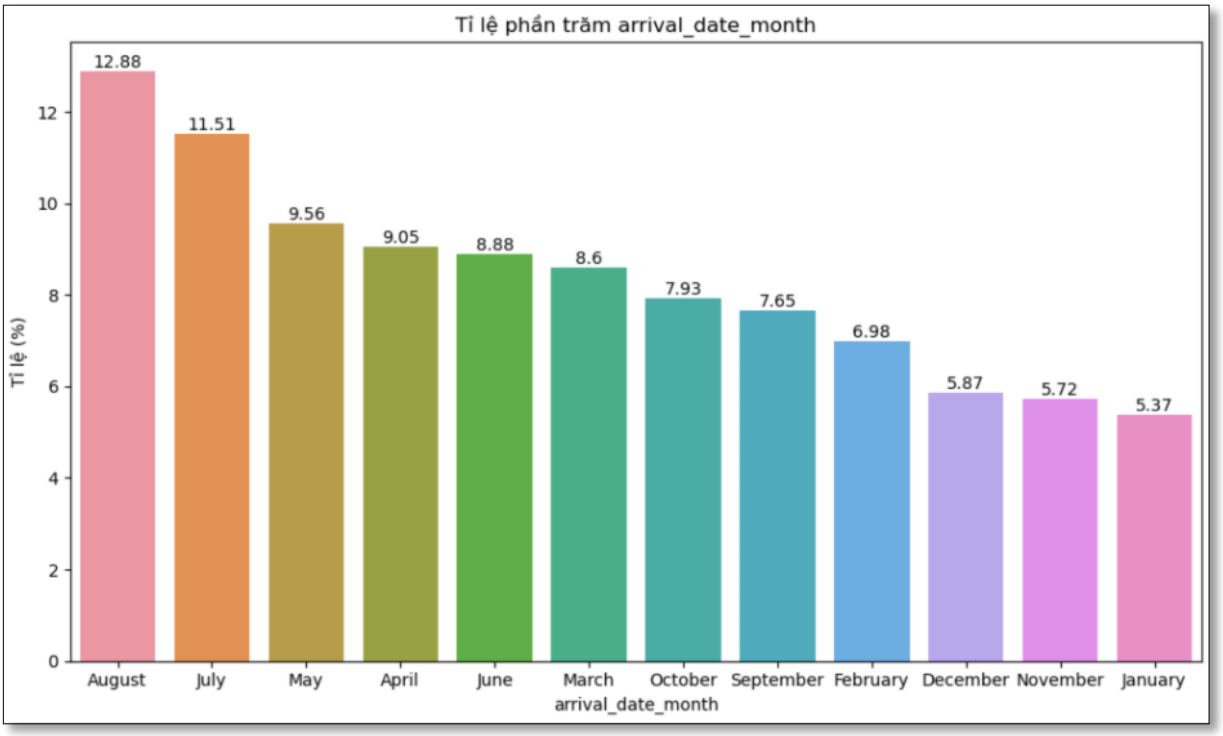
- Sử dụng value_counts để đếm số lần mỗi giá trị xuất hiện trong cột hotel của DataFrame cat_col_df và chuẩn hóa bằng cách chia cho tổng số lượng các giá trị trong cột. Kết quả sẽ là một Series với các giá trị là tỷ lệ phần trăm của mỗi giá trị so với tổng số lượng các giá trị.
- Sử dụng reset_index để thiết lập lại chỉ số của Series và chuyển đổi thành DataFrame với cột đầu tiên chứa các giá trị của cột hotel và cột thứ hai chứa tỷ lệ phần trăm đã được làm tròn.
- Sử dụng barplot của thư viện Seaborn để trực quan bằng biểu đồ cột.



- **Nhận xét**:
  - "City Hotel" chiếm tỷ lệ lớn hơn khá nhiều so với "Resort Hotel" trong bộ dữ liệu. Điều này có thể phản ánh sự phổ biến của "City Hotel" hơn so với "Resort Hotel" trong thị trường hoặc trong bộ dữ liệu này.
  - "City Hotel" có tỷ lệ cao hơn có thể do nó thường được đặt 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 nghỉ dưỡng, hướng tới những khách hàng muốn có một kỳ nghỉ thư giãn và tiện nghi tại các khu nghỉ dưỡng. Điều này cho thấy về nhu cầu du lịch cũng ảnh hưởng trong việc chọn khách sạn.

2.2. Đối với thuộc tính “arrival_date_month”

- Sử dụng value_counts để đếm số lần mỗi giá trị xuất hiện trong cột arrival_date_month của DataFrame cat_col_df và chuẩn hóa bằng cách chia cho tổng số lượng các giá trị trong cột. Kết quả sẽ là một Series với các giá trị là tỷ lệ phần trăm của mỗi giá trị so với tổng số lượng các giá trị.
- Sử dụng reset_index để thiết lập lại chỉ số của Series và chuyển đổi thành DataFrame với cột đầu tiên chứa các giá trị của arrival_date_month và cột thứ hai chứa tỷ lệ phần trăm đã được làm tròn.
- Sử dụng barplot của thư viện Seaborn để trực quan bằng biểu đồ cột.



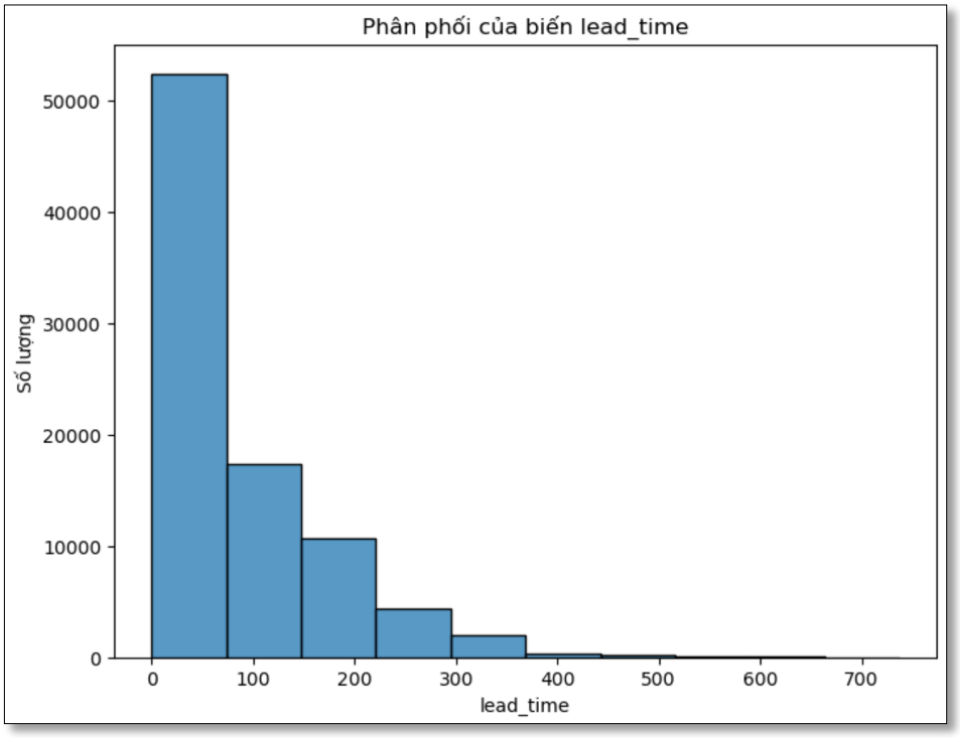
- **Nhận xét**:
  - Có sự phân bố không đồng đều về tỷ lệ phần trăm của các tháng. August (tháng 8) là tháng có tỷ lệ phần trăm cao nhất với 12.88%, tiếp theo là July (tháng 7) với 11.51%, và May (tháng 5) với 9.56%.

- Có nhiều yếu tố ảnh hưởng đến phân bố của các tháng như mùa du lịch, thời tiết, kỳ nghỉ học..., nhu cầu đặt khách sạn khi ấy sẽ nhiều. Tháng 8 và 7 thường là thời gian cao điểm du lịch trong mùa hè, khi mà nhiều người nghỉ hè hoặc có kỳ nghỉ lễ dài. Tháng 5 cũng thường là thời điểm du lịch bận rộn do thời tiết ấm áp và đẹp đẽ ở nhiều địa điểm. Trong khi đó, tháng 1 và 11 thường là thời gian ít du khách du lịch hơn do thời tiết lạnh hơn, bắt đầu vào năm học mới.
- Nhìn chung thì các tháng vào mùa hè (tháng 6 đến tháng 8) sẽ có lượng khách đến khách sạn nhiều hơn và sẽ ít hơn ở các tháng vào mùa đông (tháng 11 đến tháng 1). Điều này đã phản ánh xu hướng tổng quát về nhu cầu du lịch theo mùa trong năm.

3. Phân tích phân phối đối với biến Num

3.1. Đối với thuộc tính “lead_time”

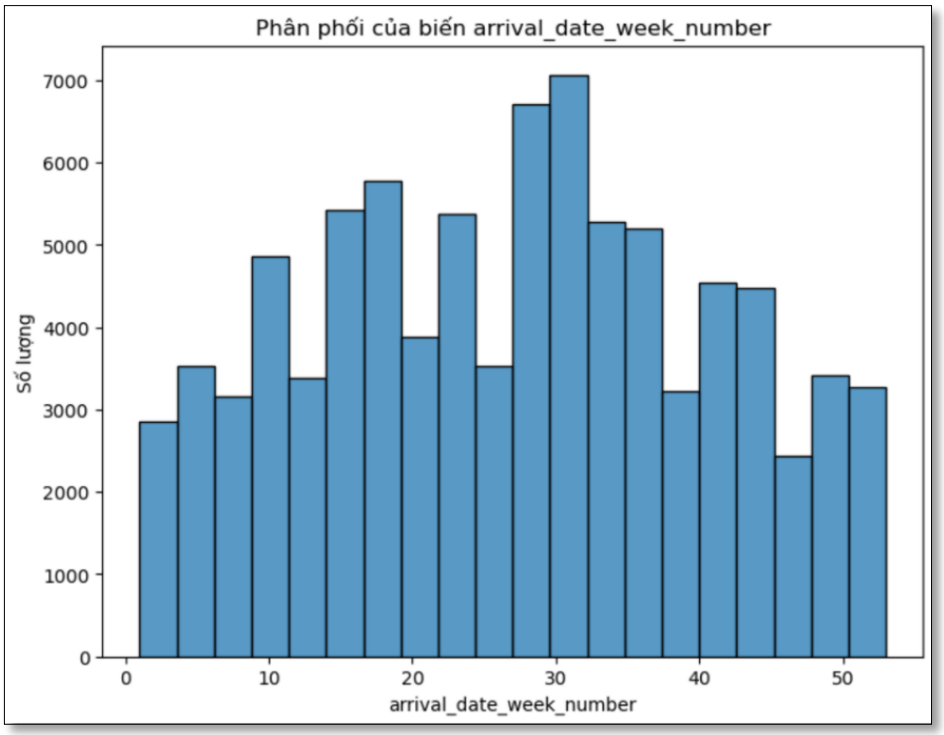
- Sử dụng describe để tạo một tóm tắt thống kê của dữ liệu, trong trường hợp này là thuộc tính lead_time. Bao gồm các thống kê cơ bản như số lượng, giá trị trung bình, độ lệch chuẩn, giá trị tối thiểu, các phân vị và giá trị tối đa.
- Sử dụng reset_index để thiết lập lại chỉ số của Series và chuyển đổi thành DataFrame.
- Sử dụng histplot của thư viện Seaborn để trực quan bằng biểu đồ histogram.



- **Nhận xét**: Đồ thị có một phân phối không đối xứng, với đỉnh của phân phối lệch về phía bên phải. Có thể thấy, khoảng thời gian đặt phòng và ngày đến 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, 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.

3.2. Đối với thuộc tính “arrival_date_week_number”

- Sử dụng describe để tạo một tóm tắt thống kê của dữ liệu của thuộc tính arrival_date_week_number. Bao gồm các thống kê cơ bản như số lượng, giá trị trung bình, độ lệch chuẩn, giá trị tối thiểu, các phân vị và giá trị tối đa.
- Sử dụng reset_index để thiết lập lại chỉ số của Series và chuyển đổi thành DataFrame.
- Sử dụng histplot của thư viện Seaborn để trực quan bằng biểu đồ histogram.



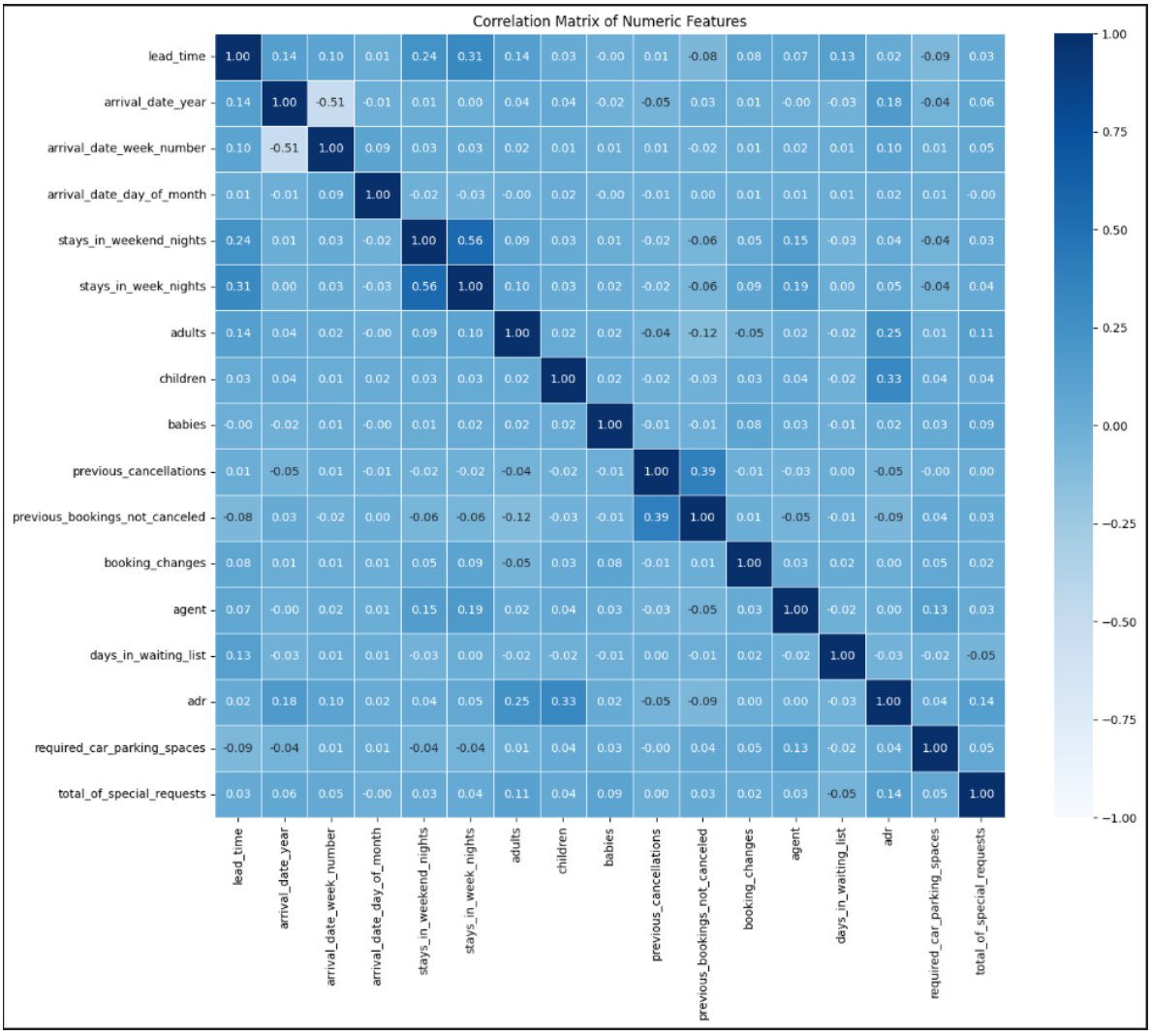
- **Nhận xét**:
  - Nhìn chung, sự phân phối của biến arrival_date_week_number khá đối xứng xung quanh giá trị trung bình.
  - Dựa vào biểu đồ ta có thể thấy, số lượng đặt khách sạn tập trung nhiều vào các tuần giữa năm. Đây là thời điểm hè và lượng đặt khách sạn cho du lịch, nghỉ dưỡng sẽ là rất nhiều. Có sự giảm dần khi càng về hai biên của đồ thị. Đây là các thời điểm đầu năm và cuối năm, do đó thường sẽ ít số lượng đặt khách sạn

do thời điểm này thường là thường điểm mà học tập, công việc nhiều, cùng với đó là thời tiết cũng không phù hợp cho việc nghỉ dưỡng, du lịch, ...

IV. EDA 2D

1. Phân tích hệ số tương quan giữa các biến num

a) Sử dụng heat map để phân tích hệ số tương quan giữa tất cả các biến num



Hình 4.1: Heat map hệ số tương quan của các biến num.

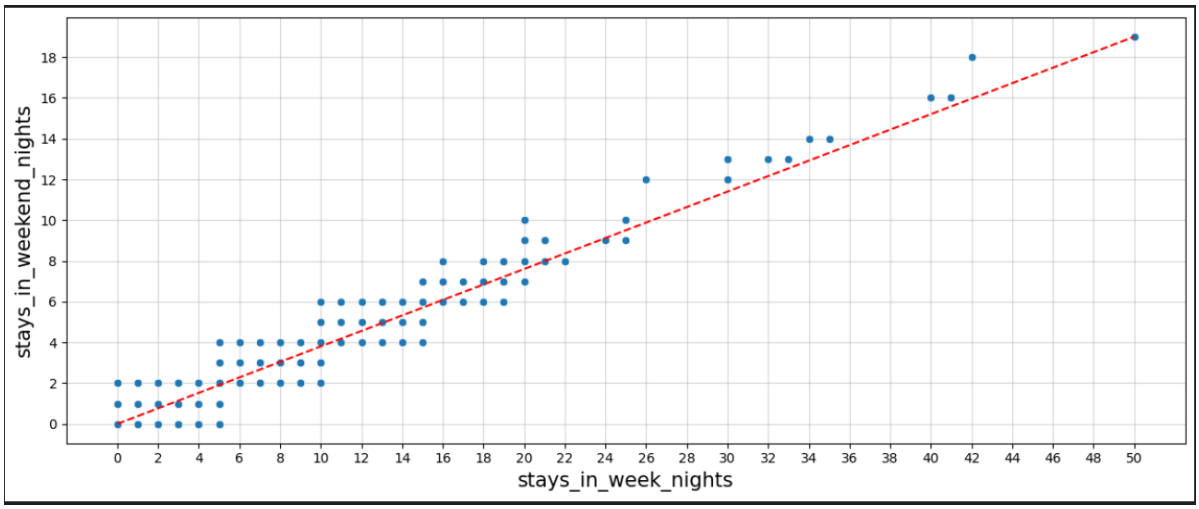
b) Rút ra insight và kết luận từ phân tích trước đó

Sau khi tiến hành trực quan bằng heatmap và quan sát, có thể đưa ra một số nhận xét sau:

- Nhìn chung, không có mối quan hệ tương quan hoàn toàn giữa 2 thuộc tính khác nhau. Giá trị tương quan nằm trong khoảng từ -0.5 đến 0.5.
- Mối tương quan thuận lớn nhất là 0.56 giữa thuộc tính stays_in_weekend_nights và stays_in_week_nights.
- Mối tương quan nghịch lớn nhất là -0.51 giữa thuộc tính arrival_date_year và arrival_date_week_number.
- Ngoài ra còn một số mối quan hệ giữa các biến:
  - Giữa previous_bookings_not_canceled với previous_cancellations: 0.39.
  - Giữa adr với children: 0.33.
  - Giữa stays_in_week_nights với lead_time: 0.31.
- Hầu hết các giá trị tương quan của các thuộc tính còn lại đều xấp xỉ bằng 0, chứng tỏ các thuộc tính này không có mối quan hệ quá chặt chẽ với nhau.

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

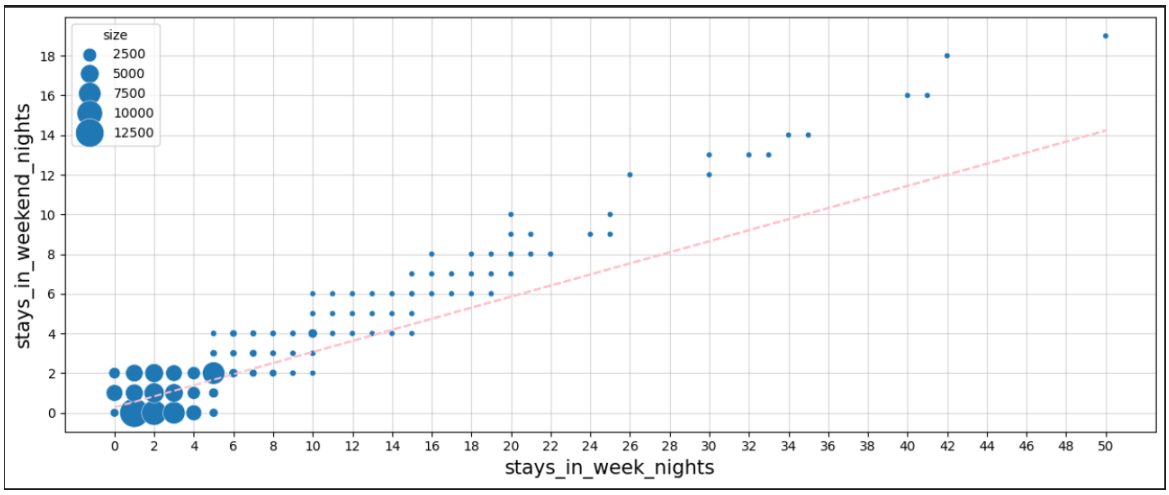
a) Sử dụng scatter plot để phân tích mối tương quan giữa biến num "stays_in_weekend_nights" và "stays_in_week_nights"



Hình 4.2: Scatter plot của biến num stays_in_weekend_nights và stays_in_week_nights.

Nhận xét sơ bộ:

- Giá trị của stays_in_week_nights phần lớn tập trung trong khoảng từ 0 đến 22.
- Giá trị của stays_in_weekend_nights phần lớn 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 outlier.
- Tuy nhiên, để chắc chắn hơn, chúng ta sẽ tiến hành tìm đường thẳng tuyến tính cho nó.



Hình 4.3: Scatter plot của biến num stays_in_weekend_nights, stays_in_week_nights và đường hồi quy tuyến tính.

- 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 tạo 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 lên biểu đồ.

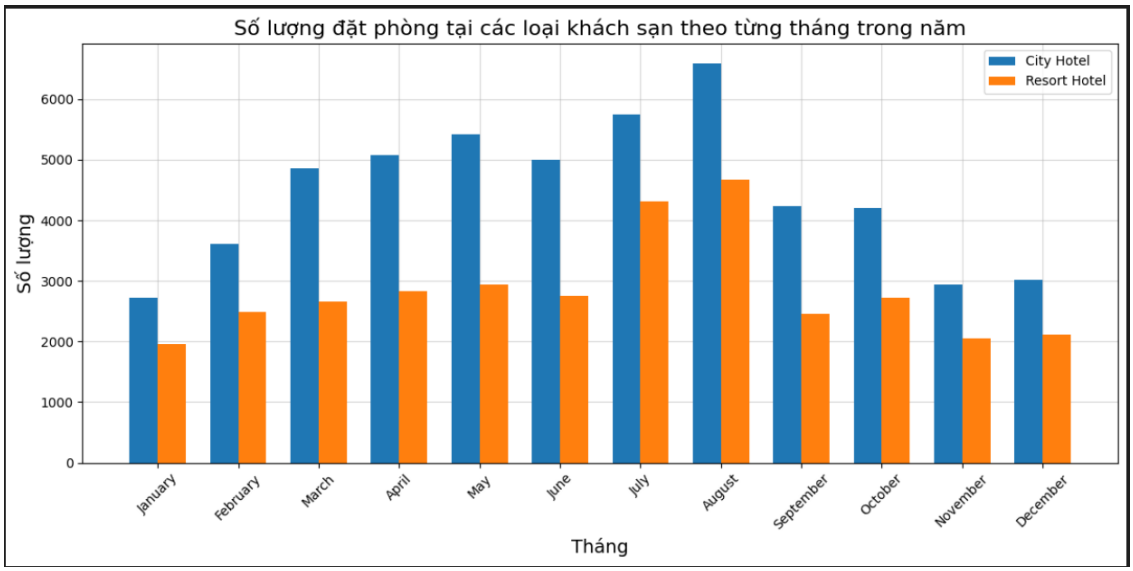
b) Rút ra insight và kết luận từ phân tích trước đó

- Các điểm dữ liệu không còn thực sự khớp với đuòng thẳng hồi quy mới.
- Hầu hết các điểm dữ liệu stays_in_week_nights nằm trong đoạn [0, 6].
- Hầu hết các điểm dữ liệu stays_in_weekend_nights nằm trong đoạn [0, 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; 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 không tăng nhiều bằng.
- Khách hàng không hoàn toàn chỉ chọn một trong hai dịch vụ trên.

3. Sử dụng bar chart để phân tích dữ liệu num và cate

3.1. Số lượng đặt phòng tại các loại khách sạn theo các tháng trong năm

a) Sử dụng bar chart để phân tích dữ liệu cate "hotel" và "arrival_date_month"



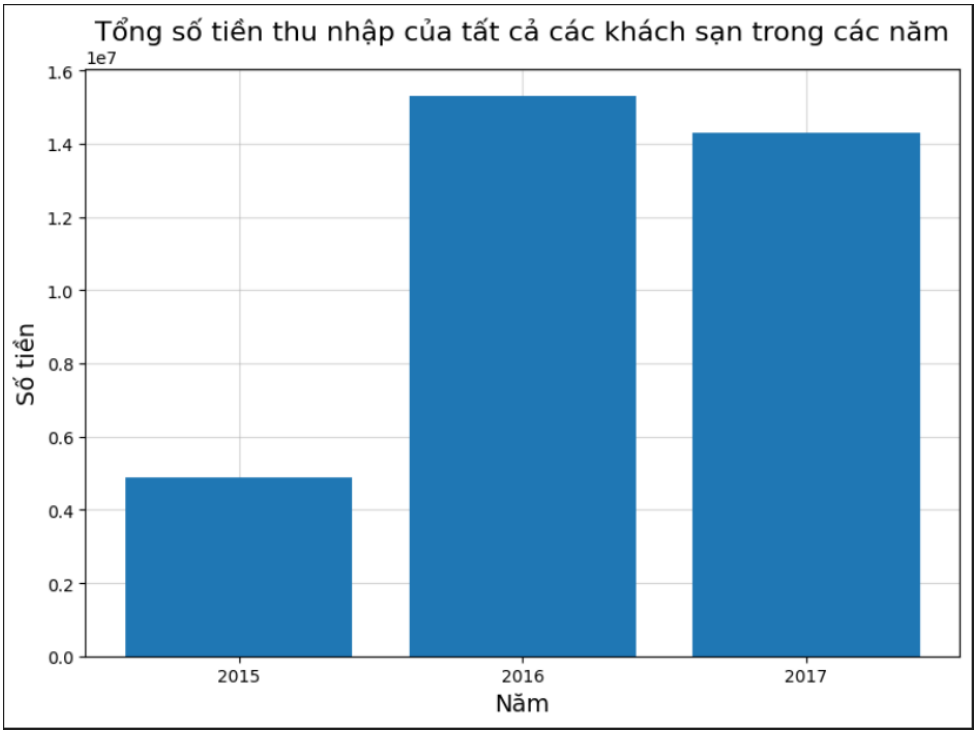
Hình 4.4: Biểu đồ cột về sự phân bố số lượng đặt phòng tại các loại khách sạn theo các tháng trong năm.

b) Rút ra insight và kết luận từ phân tích trước đó

- 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.
- Lượng khách hàng đặt phòng khách sạn ít nhất là vào tháng 1, 11, 12 với khoảng 5000 lượt đặt 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 về tháng 1 năm sau.
- Xu hướng trên là khá phù hợp vì mùa hè và thu là thời điểm phổ biến mà mọi người quyết định đi du lịch nhiều hơn.

3.2. Tổng số tiền mà tất cả các khách sạn thu được trong các năm

a) Sử dụng bar chart để phân tích dữ liệu num "adr" và "arrival_date_year"



Hình 4.5: Biểu đồ cột về thu nhập hằng năm của tất cả khách sạn.

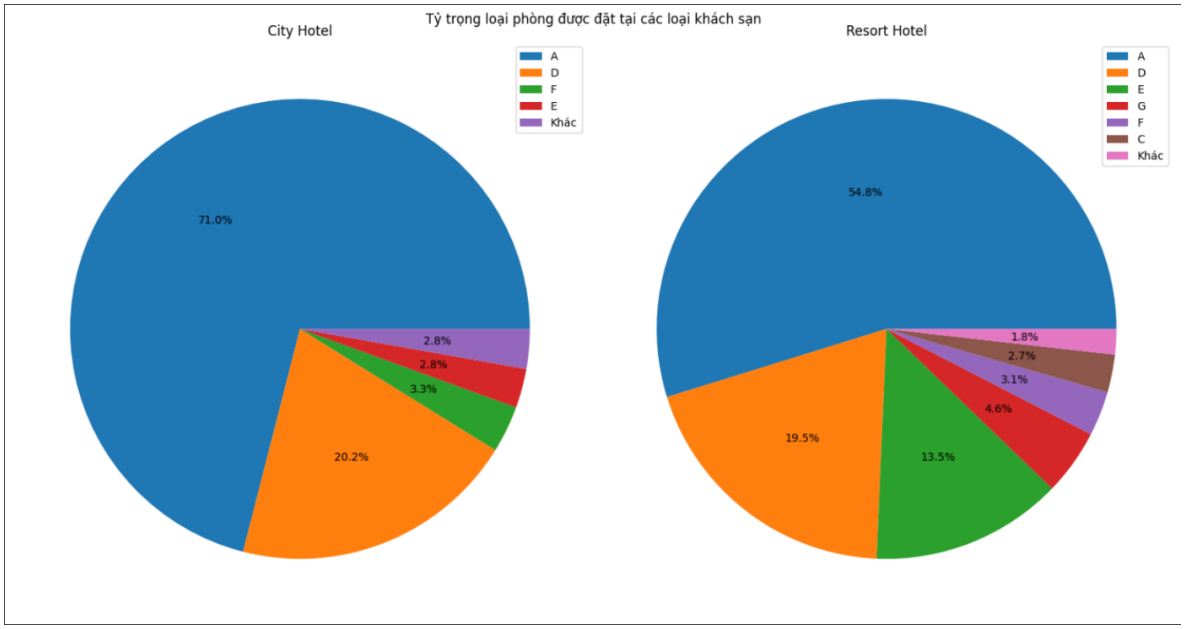
b) Rút ra insight và kết luận từ phân tích trước đó

- Có sự chênh lệch lớn giữa tổng số tiền thu được giữa các năm.
- Năm nhiều nhất là 2016 với 15 000 000.
- Năm ít nhất là 2015 với 5 000 000.
- Chênh lệch giữa năm 2016 và 2015 là khoảng 10 000 000, 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.
- 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, 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.

4. Tính tỉ trọng đối với hai biến cate

4.1. Tỷ trọng loại phòng được đặt tại các loại khách sạn khác nhau

a) Sử dụng biểu đồ tròn để phân tích tỉ trọng của biến "hotel" và "reserved_room_type"



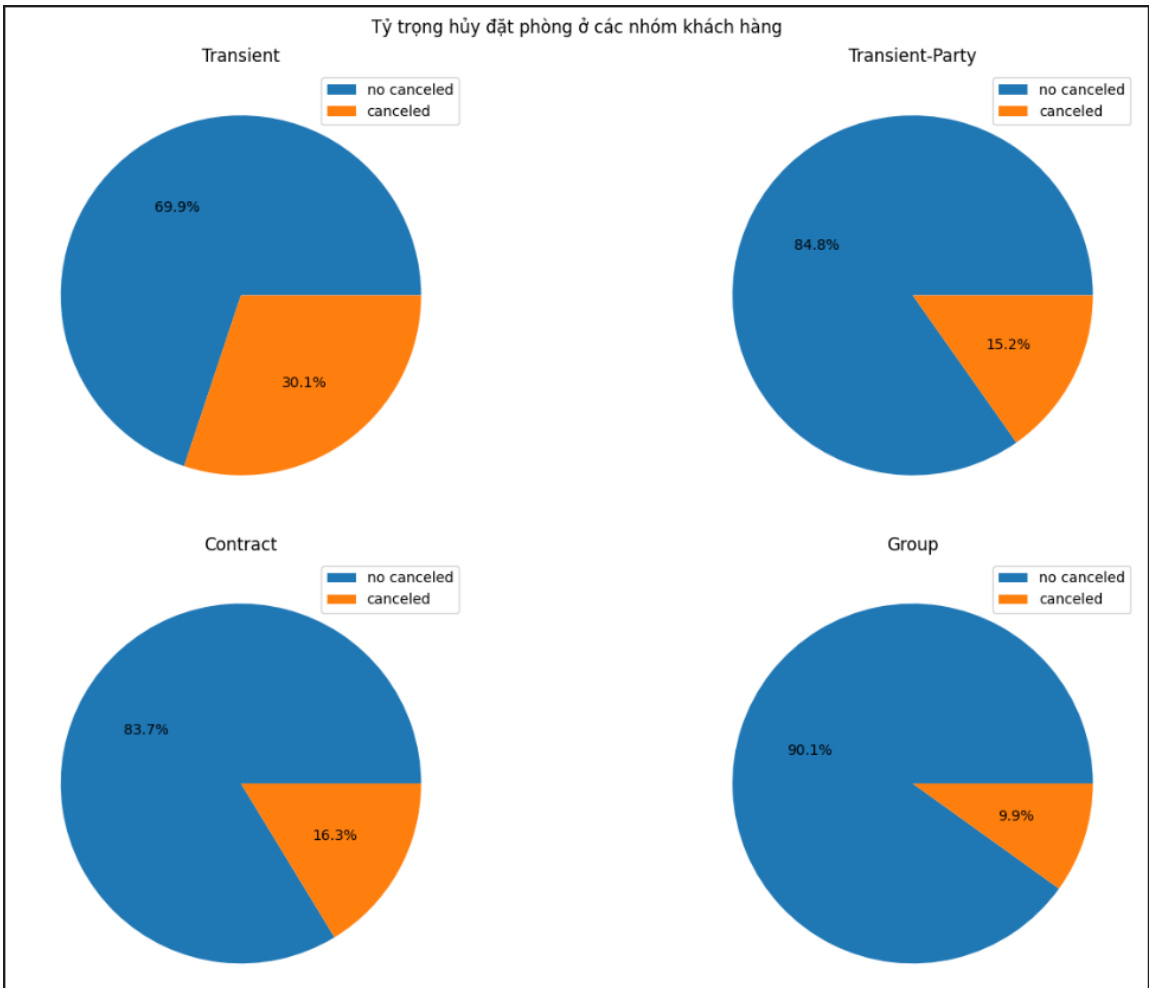
Hình 4.6: Biểu đồ tròn thể hiện tỷ trọng của các loại phòng được đặt tại các loại khách sạn khác nhau.

b) Rút ra insight và kết luận từ phân tích trước đó

- Nhìn chung, tại các loại khách sạn, 2 loại phòng được yêu thích nhất lần lượt là loại A, loại D.
- Trong khi tại "City Hotel" thì phòng loại A chiếm tỉ lệ rất lớn 71%, phòng loại D chiếm 20% và 9% cho các loại còn lại. Thì bên "Resort Hotel" tỉ lệ giữa các loại phòng căn bằng hơn, loại A vẫn nhiều nhất nhưng chỉ chiếm 55%, loại D 19.5%, loại E 13.5% và 12% cho các loại còn lại.

4.2. Tỷ trọng hủy đặt phòng ở các nhóm khách hàng

a) Sử dụng biểu đồ tròn để phân tích tỉ trọng của biến "customer_type" và "is_canceled"



Hình 4.5: Biểu đồ tròn thể hiện tỷ trọng về tỉ lệ hủy phòng ở các nhóm khách hàng khác nhau.

b) Rút ra insight và kết luận từ phân tích trước đó

- Tỉ lệ hủy 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 tỉ lệ hủy phòng cao.
- Tỉ lệ hủy 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 tỉ lệ hủy phòng thấp.
- Ở nhóm “transient-party” và “contract” là xấp xỉ nhau, khoảng 15-16%. Tỉ lệ hủy phòng 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 tỉ lệ hủy phòng cao.

V. EDA 3D

1. Tóm tắt quá trình tiền xử lý dữ liệu

- Đầu tiên, ta tạo ra một DataFrame mới để tránh làm ảnh hưởng đến tập dữ liệu trước đó, đặt tên là EDA_3D_df.
- 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
- Loại bỏ các dòng dữ liệu có giá trị tại thuộc tính total_people bằng 0.

2. Xử lý giá trị ngoại lai trên thuộc tính adr (sử dụng phương pháp IQR)

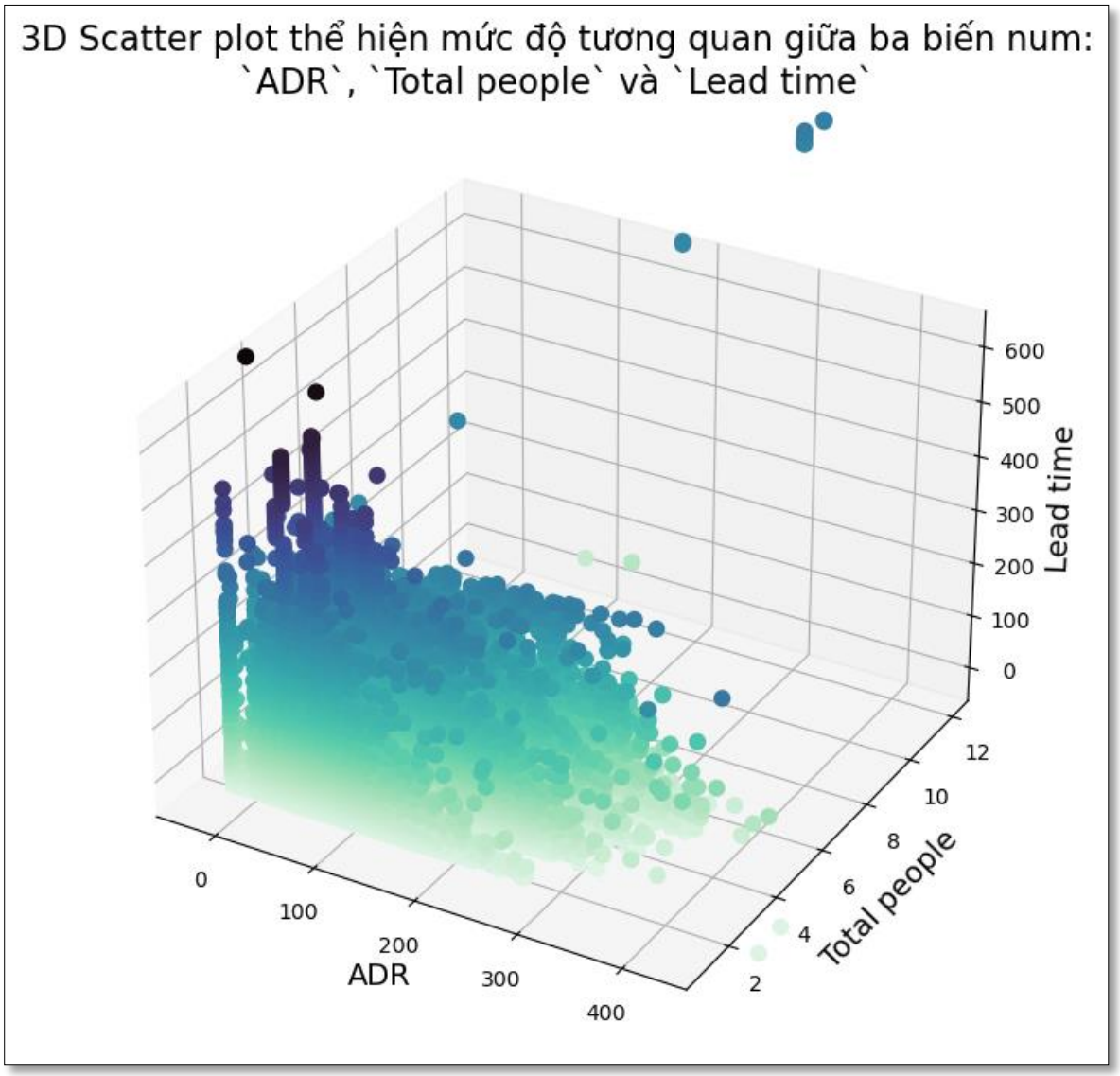
	adr
48515	5400.000000
111403	510.000000
15083	508.000000
103912	451.500000
13142	450.000000

Hình 5.1: 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 kết quả bên trên, 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.

3. Sử dụng Scatter plot để phân tích dữ liệu 3D cho ba biến num

a) Sử dụng Scatter plot 3D để phân tích mức độ tương quan giữa ba thuộc tính: "adr", "total_people" và "lead_time"



Hình 5.2: Scatter plot 3D thể hiện mức độ tương quan giữa ba thuộc tính: adr, total_people và lead_time.

b) Rút ra insight và kết luận từ phân tích trước đó

(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:

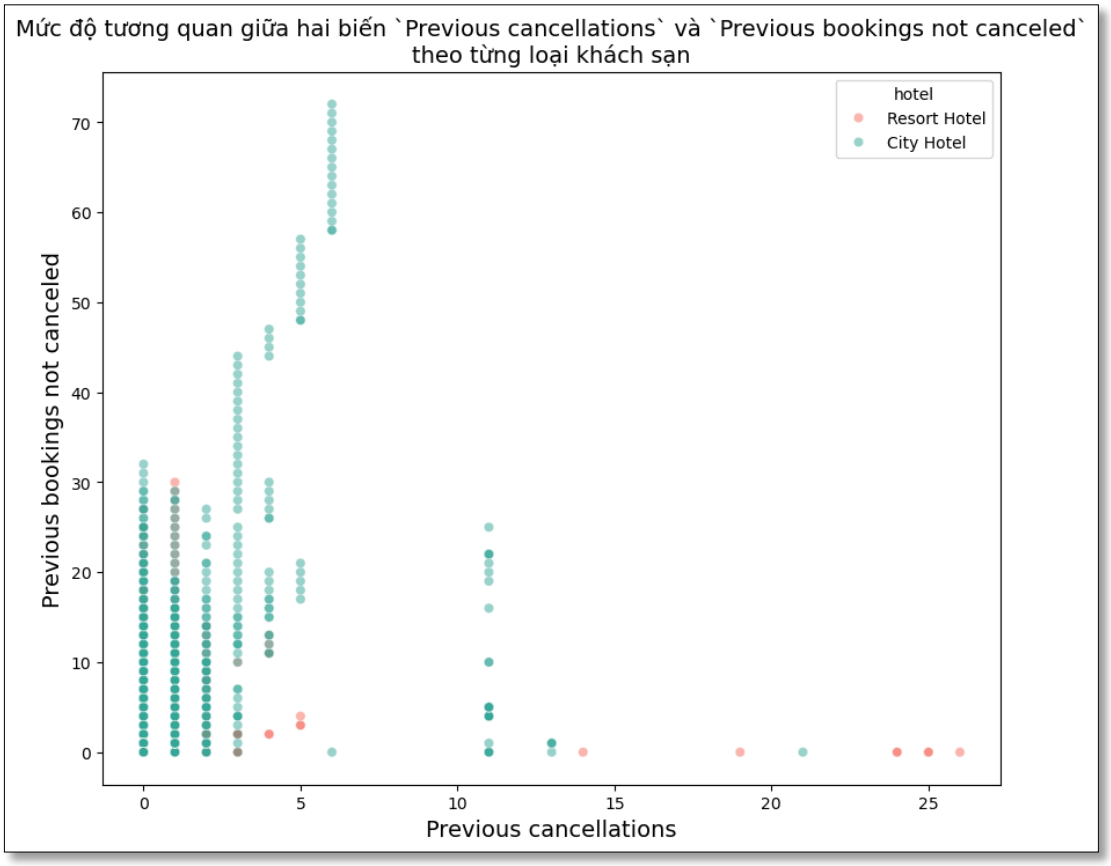
Ta thấy có mối tương quan theo chiều dương khá yếu giữa hai thuộc tính total_people và lead_time. Nhìn chung, khi số lượng người trong nhóm hành khách tăng lên thì họ cũng có xu hướng đặt phòng sớm hơn (lead_time lớn). Điều này cũng không quá khó hiểu vì các nhóm hành khách đông người thường sẽ là: một 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.

4. 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 điểm khác nhau giữa hai loại khách sạn.

4.1. Kết hợp "hotel" với hai biến "previous_cancellations" và "previous_bookings_not_canceled"

a) Phân tích mức độ tương quan giữa hai thuộc tính "previous_cancellations" và "previous_bookings_not_canceled" theo từng loại khách sạn



Hình 5.3: Biểu đồ Scatter plot thể hiện mối tương quan giữa hai thuộc tính previous_cancellations và previous_bookings_not_canceled theo từng loại khách sạn.

b) Rút ra insight và kết luận từ phân tích trước đó

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.

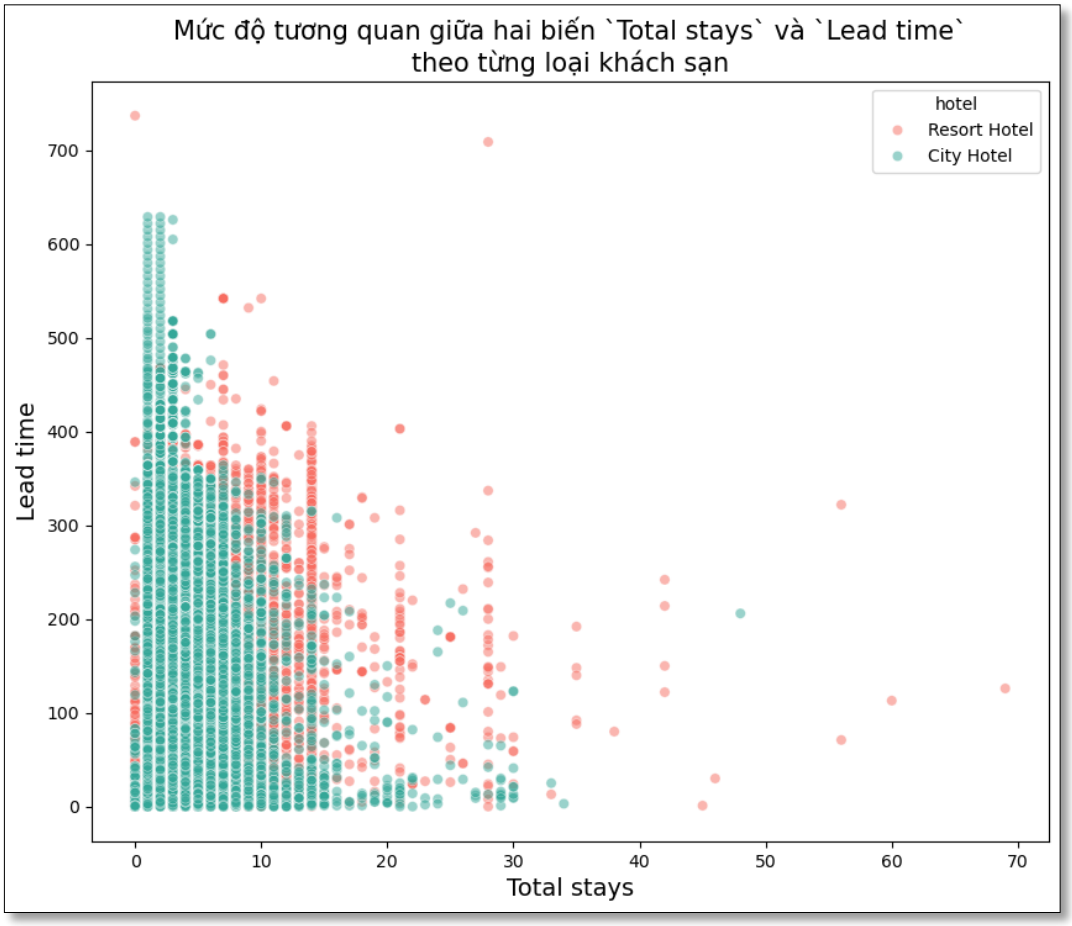
Điểm khác nhau cơ bản giữa hai nhóm "City Hotel" và "Resort Hotel" đến từ phạm vi phân bố của các điểm dữ liệu:

- 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, giá trị previous_cancellations trung bình của "City Hotel" cao hơn "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].

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.

4.2. Kết hợp “hotel” với hai biến “total_stays” và “lead_time”

a) Phân tích mức độ tương quan giữa hai thuộc tính “total_stays” và “lead_time” theo từng loại khách sạn



Hình 5.4: Biểu đồ Scatter plot thể hiện mối tương quan giữa hai thuộc tính total_stays và lead_time theo từng loại khách sạn.

b) Rút ra insight và kết luận từ phân tích trước đó

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.

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.

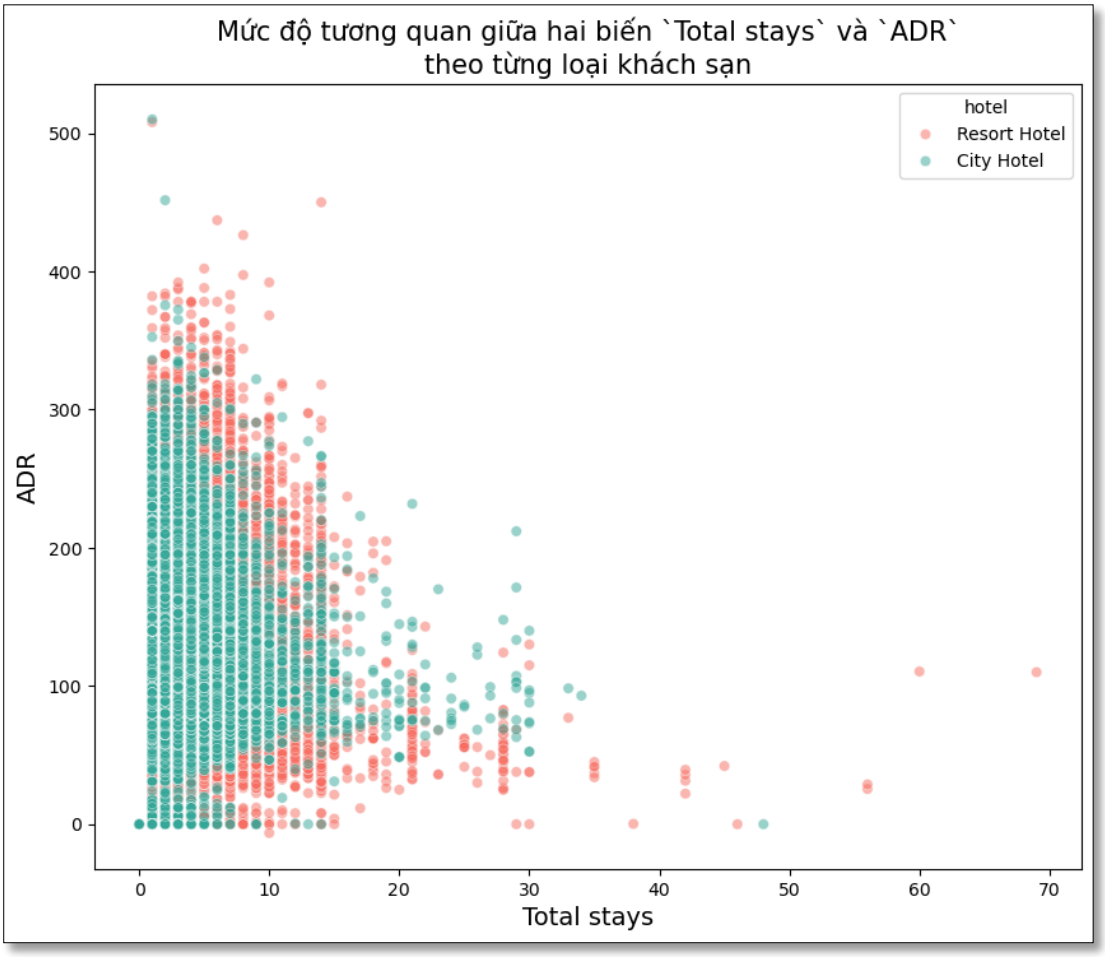
Xét nhóm "Resort Hotel":

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.

4.3. Kết hợp “hotel” với hai biến “total_stays” và “adr”

a) Phân tích mức độ tương quan giữa hai thuộc tính “total_stays” và “adr” theo từng loại khách sạn



Hình 5.5: Biểu đồ Scatter plot thể hiện mối tương quan giữa hai thuộc tính total_stays và adr theo từng loại khách sạn.

b) Rút ra insight và kết luận từ phân tích trước đó

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.

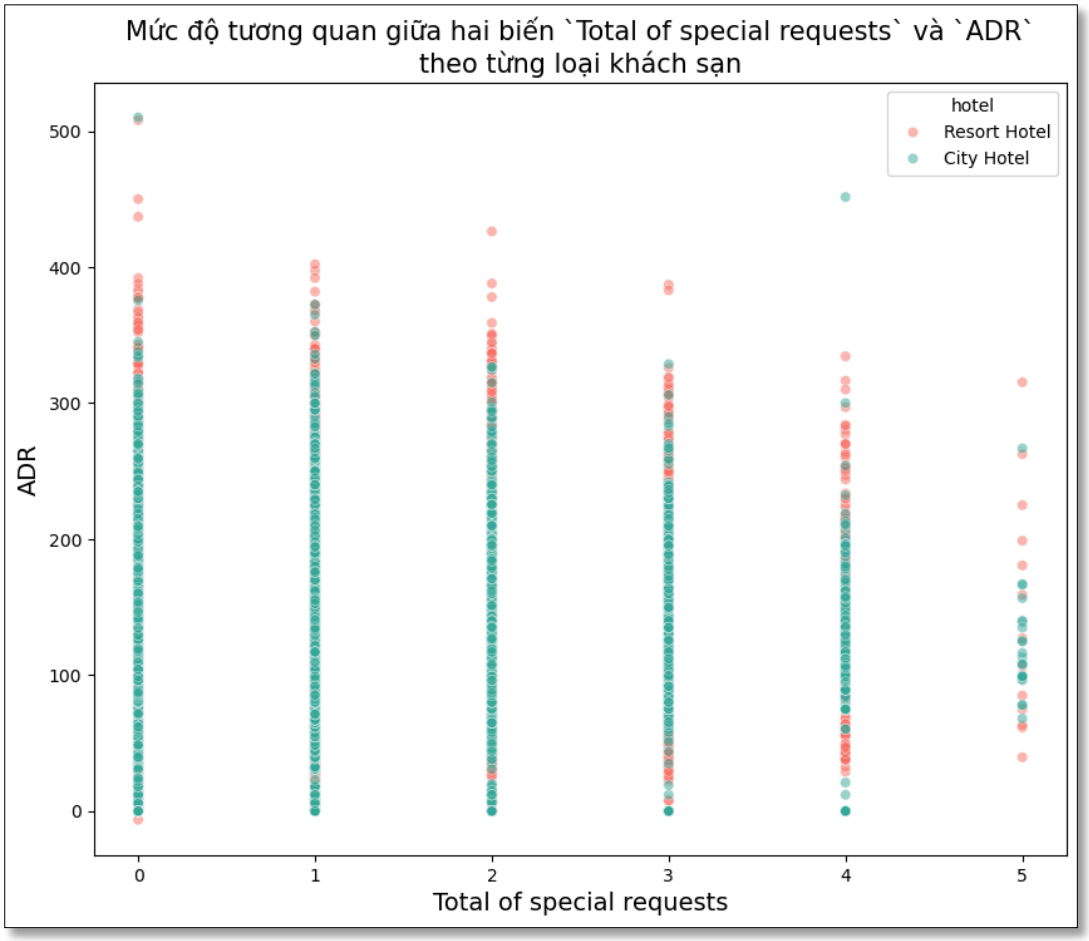
Câu "lợi nhuận có xu hướng giảm xuống" không đồng nghĩa với việc khách sạn bị lỗ vốn khi cung cấp dịch vụ cho một khách hàng nào đó. Mà câu bên trên nên được hiểu theo nghĩa là khách sạn thu được số tiền ít hơn (từ hành khác có thời gian lưu trú dài) so với số tiền mà khách sạn kiếm được từ các hành 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.

4.4. Kết hợp “hotel” với hai biến “total_of_special_requests” và “adr”

a) Phân tích mức độ tương quan giữa thuộc tính “total_of_special_requests” và “adr” theo từng loại khách sạn



Hình 5.6: Biểu đồ Scatter plot thể hiện mối tương quan giữa hai thuộc tính total_of_special_requests và adr theo từng loại khách sạn.

b) Rút ra insight và kết luận từ phân tích trước đó

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.

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

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

hotel	assigned_room_type	(-7.001, 123.0]	(123.0, 253.0]	(253.0, 383.0]	(383.0, 513.0]
City Hotel	A	0.787000	0.213000	0.000000	0.000000
	B	0.832000	0.166000	0.002000	0.000000
	C	0.766000	0.234000	0.000000	0.000000
	D	0.484000	0.515000	0.001000	0.000000
	E	0.333000	0.658000	0.008000	0.000000
	F	0.117000	0.830000	0.053000	0.000000
	G	0.225000	0.563000	0.210000	0.001000
	K	0.822000	0.168000	0.011000	0.000000
Resort Hotel	A	0.778000	0.216000	0.006000	0.000000
	B	0.679000	0.314000	0.006000	0.000000
	C	0.590000	0.362000	0.047000	0.000000
	D	0.783000	0.208000	0.009000	0.000000
	E	0.686000	0.292000	0.022000	0.000000
	F	0.593000	0.364000	0.043000	0.001000
	G	0.357000	0.506000	0.135000	0.002000
	H	0.333000	0.490000	0.170000	0.007000
	I	0.887000	0.096000	0.017000	0.000000
	L	1.000000	0.000000	0.000000	0.000000

Hình 5.7: Tỷ trọng theo bin của thuộc tính adr chia theo thể loại với hai biến cate là hotel và assigned_room_type.

b) Rút ra insight và kết luận từ phân tích trước đó

(1) Xét nhóm khách sạn "City Hotel":

Với các mã phòng A, B, C và K: Trong số các hành khách đến lưu trú tại khách sạn thuộc nhóm "City Hotel" và được sắp xếp vào ở một trong các phòng A, B, C hoặc K, ta thấy có khoảng 80% hành khách sẽ tạo ra giá trị doanh thu trung bình rơi vào khoảng

(-7.001, 123.0]. Có khoảng 20% hành khách tạo ra doanh thu trung bình trong khoảng (123.0, 253.0]. Và có rất ít hành khách (hoặc thậm chí không có hành khách nào) trong nhóm này tạo ra giá trị adr lớn hơn 253 cho các khách sạn "City Hotel".

Với các mã phòng D, E và F: Trong số các hành khách đến lưu trú tại khách sạn thuộc nhóm "City Hotel" và được sắp xếp vào ở một trong các phòng D, E hoặc F, ta thấy có hơn một nửa số hành khách tạo ra giá trị doanh thu trung bình rơi vào khoảng (123.0, 253.0]. Các hành khách còn lại chủ yếu sẽ tạo ra ít doanh thu hơn cho khách sạn, giá trị adr thuộc đoạn (-7.001, 123.0]. Tuy có một số ít hành khách tạo ra giá trị adr lớn hơn 253 nhưng số lượng mẫu dữ liệu này là quá ít nên ta sẽ không đề cập chi tiết.

Đặc biệt nhất là mã phòng G: Trong số các hành khách lưu trú tại khách sạn thuộc loại "City Hotel" và được sắp xếp ở phòng có mã G, ta thấy có hơn 56% hành khách tạo ra giá trị adr dao động trong khoảng (123.0, 253.0]. Nổi bật nhất là có hơn 20% hành khách tạo ra giá trị adr rơi vào khoảng (253.0, 383.0]. Phần lớn các hành khách còn lại sẽ đem lại doanh thu trung bình rơi vào khoảng (-7.001, 123.0]. Như vậy, mã phòng G là loại phòng có tiềm năng đem lại nhiều thu nhập cho khách sạn.

Như vậy, với các khách sạn trong nhóm "City Hotel", các chủ khách sạn có thể tăng thêm số lượng phòng có mã F và G để có thể tối đa hóa doanh thu cho khách sạn.

(2) Xét nhóm khách sạn "Resort Hotel":

Với các mã phòng A, B, D, E, I và L: Trong số các hành khách đến lưu trú tại khách sạn thuộc nhóm "Resort Hotel" và được sắp xếp vào ở một trong các phòng A, B, D, E, I hoặc L, ta thấy phần lớn hành khách sẽ tạo ra giá trị doanh thu trung bình ở bin thứ nhất, giá trị adr rơi vào khoảng (-7.001, 123.0]. Chỉ có một lượng nhỏ hành khách có thể tạo ra giá trị adr rơi vào khoảng (123.0, 253.0] và có rất ít hành khách có thể tạo ra giá trị adr lớn hơn 253.

Với các mã phòng C và F: Trong số các hành khách đến lưu trú tại khách sạn thuộc nhóm "Resort Hotel" và được sắp xếp vào ở một trong các phòng C hoặc F, ta thấy có khoảng 60% hành khách tạo ra giá trị adr trong khoảng (-7.001, 123.0]. Có khoảng hơn

30% hành khách tạo ra giá trị adr trong khoảng (123.0, 253.0]. Đồng thời, cũng có khoảng 5% hành khách tạo ra doanh thu nhiều hơn cho khách sạn, giá trị adr lớn hơn 253.

Đặc biệt nhất là mã phòng G và H: Phần lớn hành khách lưu trú tại các khách sạn thuộc nhóm "Resort Hotel" và ở một trong các mã phòng G hoặc H thường đem lại rất nhiều lợi nhuận cho khách sạn. Có khoảng 50% hành khách tạo ra lợi nhuận ở mức (bin) 2, giá trị adr thuộc đoạn (123.0, 253.0]. Đồng thời có khoảng 15% hành khách tạo ra lợi nhuận ở mức (bin) 3 và 4, giá trị adr lớn hơn 253.

Như vậy, với các khách sạn trong nhóm "Resort Hotel", các chủ khách sạn có thể tăng thêm số lượng phòng có mã H, G, C và F để có thể tối đa hóa doanh thu cho khách sạn.

VI. Insight

1. Data Understanding

- Dữ liệu có 119390 dòng và 32 cột.
- Nhóm đã thực hiện việc khám phá dữ liệu thông qua việc chia theo các biến dạng Num và Cate để khám phá. Đối với các biến dạng Num, cột company có tỉ lệ thiếu rất cao, lên đến hơn 90% nên nhóm đã loại bỏ cột này. Cột children và agent cũng có tỷ lệ thiếu nhưng không đáng kể, do đó sẽ được điền bằng giá trị mode và median. Đối với các biến dáng Cat, chỉ có cột country là có tỉ lệ thiếu nhưng cũng khá nhỏ và đã được điền bằng giá trị mode của cột.
- Tỷ lệ trùng lặp của dữ liệu là 26.80%, bao gồm 31994 dòng bị trùng lặp. Sau khi xử lý để loại bỏ các dòng trùng lặp, dữ liệu còn 87396 dòng và 32 cột.

2. EDA 1D

Thông qua việc tính toán tỉ lệ với các biến Cate, có thể thấy rằng:

- "City Hotel" có độ phổ biến hơn "Resort Hotel", điều này cho thấy xu hướng và sự ưu tiên của khách hàng trong đặt khách sạn. Có thể do nhiều yếu tố ảnh hưởng chẳng hạn như về vị trí và nhu cầu du lịch.
- Các tháng trong hè (tháng 6 đến tháng 8) thường có tỷ lệ đặt phòng cao hơn, điều này đã phần nào cho thấy nhu cầu về du lịch và kỳ nghỉ trong thời gian này. Trong khi đó các tháng cuối năm và đầu năm thường có tỷ lệ đặt phòng thấp, có thể các yếu tố như thời tiết, và những tháng này thường ít kỳ nghỉ nên đã phần nào làm giảm tỷ lệ đặt phòng. Có thể thấy, yếu tố về mùa và các tháng trong năm cũng phần nào ảnh hưởng đến việc đặt phòng khách sạn.

Thông qua việc phân tích phân phối của các biến Num, có thể thấy rằng:

- Đa phần khách hàng đều đặt phòng trong thời gian gần 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.

- Có một sự tập trung đáng kể của lượng đặt phòng khách sạn vào các tuần giữa năm, đặc biệt là trong mùa hè. Điều này có thể phản ánh mùa du lịch và kỳ nghỉ hè, khi người ta thường muốn tận hưởng thời gian nghỉ của họ trong khoảng thời gian này. Số lượng đặt phòng giảm dần khi di chuyển về hai biên của biểu đồ, đặc biệt là ở đầu năm và cuối năm. Có thể do yếu tố thời tiết không thuận lợi cho việc du lịch hoặc đây là thời điểm việc học và việc làm có lượng công việc và bài học lớn.

3. EDA 2D

Thông qua việc phân tích hệ số tương quan của các biến Numerical:

- 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 sử dụng bar chart để phân tích dữ liệu num và cate:

- 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.
- “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 sử dụng pie chart để phân tích dữ liệu cate:

- 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.
- 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, ...
- 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.

4. EDA 3D

Thông qua việc sử dụng Scatter plot 3D để phân tích mối tương quan giữa ba thuộc tính "adr", "total_people" và "lead_time":

- Thuộc tính adr và total_people có mối tương quan thuận không quá mạnh. 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.
- Mối tương quan giữa adr và lead_time không quá rõ ràng.
- 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 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.

Thông qua việc sử dụng Scatter plot 2D để phân tích mối tương quan giữa hai biến "previous_cancellations" và "previous_bookings_not_canceled" theo từng loại khách sạn:

- 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 thuận khá mạnh.
- Còn 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 nghịch 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, các khách sạn thuộc loại "Resort Hotel" thường gặp phải tình trạng một hành khách hủy đặt phòng rất nhiều lần.
- 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 sử dụng Scatter plot 2D để phân tích mối tương quan giữa hai biến "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 (tương quan thuận). 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.
- Phần lớn hành khách ở các "City Hotel" thường có thời gian lưu trú trong khoảng 10 ngày.
- 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.
- 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).

Thông qua việc sử dụng Scatter plot 2D để phân tích mối tương quan giữa hai biến "total_stays" và "adr" theo từng loại khách sạn:

- 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.
- 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.

- Trong các chuyến đi ngắn ngày (ít hơn 5 ngày), hành khách thường lựa chọn nghỉ ngơi tại "City Hotel". Nhưng trong các chuyến đi dài ngày, "Resort Hotel" lại thường được ưu ái hơn.
- Đồng thời, "City Hotel" cũng có giá trị trung bình của thuộc tính adr cao hơn so với "Resort Hotel". Điều này cho thấy các khách sạn thuộc nhóm "City Hotel" thường thu được nhiều lợi nhuận hơn.

Thông qua việc sử dụng Scatter plot 2D để phân tích mối tương quan giữa hai biến "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.

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

- Xét nhóm khách sạn "City Hotel":
  - Các mã phòng A, B, C và K: Có khoảng 80% hành khách sẽ tạo ra giá trị doanh thu trung bình rơi vào khoảng (-7.001, 123.0]. Có khoảng 20% hành khách tạo ra doanh thu trung bình trong khoảng (123.0, 253.0].
  - Với các mã phòng D, E và F: Có hơn một nửa số hành khách tạo ra giá trị doanh thu trung bình rơi vào khoảng (123.0, 253.0]. Các hành khách còn lại chủ yếu sẽ tạo ra ít doanh thu hơn cho khách sạn, giá trị adr thuộc đoạn (-7.001, 123.0].
  - Đặc biệt nhất là mã phòng G: Có hơn 56% hành khách tạo ra giá trị adr dao động trong khoảng (123.0, 253.0]. Nổi bật nhất là có hơn 20% hành khách tạo ra giá trị adr rơi vào khoảng (253.0, 383.0]. Như vậy, mã phòng G là loại phòng có tiềm năng đem lại nhiều thu nhập cho khách sạn.
  - Như vậy, các chủ khách sạn có thể tăng thêm số lượng phòng có mã F và G để có thể tối đa hóa doanh thu cho khách sạn.
- Xét nhóm khách sạn "Resort Hotel":
  - Với các mã phòng A, B, D, E, I và L: Phần lớn hành khách sẽ tạo ra giá trị doanh thu trung bình ở bin thứ nhất, giá trị adr rơi vào khoảng (-7.001, 123.0]. Chỉ có một lượng nhỏ hành khách có thể tạo ra giá trị adr rơi vào khoảng (123.0, 253.0] và có rất ít hành khách có thể tạo ra giá trị adr lớn hơn 253.
  - Với các mã phòng C và F: Có khoảng 60% hành khách tạo ra giá trị adr trong khoảng (-7.001, 123.0]. Có khoảng hơn 30% hành khách tạo ra giá trị adr trong khoảng (123.0, 253.0]. Đồng thời, cũng có khoảng 5% hành khách tạo ra doanh thu nhiều hơn cho khách sạn, giá trị adr lớn hơn 253.

- Đặc biệt nhất là mã phòng G và H: Có khoảng 50% hành khách tạo ra lợi nhuận ở mức (bin) 2, giá trị adr thuộc đoạn (123.0, 253.0]. Đồng thời có khoảng 15% hành khách tạo ra lợi nhuận ở mức (bin) 3 và 4, giá trị adr lớn hơn 253.
- Như vậy, các chủ khách sạn có thể tăng thêm số lượng phòng có mã H, G, C và F để có thể tối đa hóa doanh thu cho khách sạn.

VII. Tài liệu tham khảo

[1]: File notebook được cung cấp bởi giảng viên thực hành – link.

[2]: Hotel Bookings Exploratory Data Analysis – github.com/Neerajdataguy.

[3]: Hotel Booking Analysis – github.com/ajitmane36.

[4]: Tài liệu hướng dẫn (documentation) của các thư viện: NumPy, Pandas, Matplotlib, Seaborn, v.v.