ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

TRỰC QUAN HÓA DỮ LIỆU

< Đồ án Quan hóa Dữ liệu >

Sinh viên thực hiện: 21127115 - Trần Thanh Ngân
21127229 - Dương Trường Bình
21127616 - Lê Phước Quang Huy

Giảng viên hướng dẫn: TS. Bùi Tiến Lên

Lớp: 21KHDL

Mục lục

Thông tin nhóm và phân công công việc

MSSV	Họ và tên	Công việc được phân công	Mức độ hoàn thành
21127115	Trần Thanh Ngân	• IV. EDA 3D • V. Insight	100%
21127229	Dương Trường Bình	• I. Data Understanding • II. EDA 1D • V. Insight	100%
21127616	Lê Phước Quang Huy	• III. EDA 2D • IV. EDA 3D • V. Insight	100%

Tiến độ công việc

Phần	Nội dung	Mức độ hoàn thành
I. Data Understanding	1. Đếm số dòng và số cột.	100%
	2. Viết bảng mô tả về các cột.	100%
	3. Lấy 5 điểm dữ liệu ra làm mẫu.	100%
	4. Phân tích tỷ lệ missing rate.	100%
	5. Phân tích tỷ lệ duplicate.	100%
	6. Fill missing rate.	100%
II. EDA 1D	1. Chia loại dữ liệu num hoặc cate.	100%
	2. Phân tích tỷ lệ đó với biến cate.	100%
	3. Phân tích phân phối đối biến num.	100%
III. EDA 2D	1. Phân tích hệ số tương quan giữa các biến num.	100%
	2. Sử dụng Scatter plot để phân tích dữ liệu 2D.	100%
	3. Sử dụng bar chart để phân tích dữ liệu num và cate.	100%
	4. Tính tỷ trọng đối với hai biến cate.	100%
IV. EDA 3D	1. Sử dụng Scatter plot để phân tích dữ liệu 3D cho ba biến num.	100%
	2. Sử dụng Scatter plot 2D và màu đối với hai biến num và cate.	100%
	3. Tính tỷ trọng theo bin chia theo thể loại với hai biến cate.	100%
V. Insight	Rút ra insight, kết luận từ những phân tích đã thực hiện ở các phần trước.	100%

1 Data Understanding

Bộ dữ liệu nhóm chọn: `hotel_bookings.csv`

- Tập dữ liệu gồm 11930 dòng và 32 cột. Mỗi dòng trong bộ dữ liệu gồm thông tin về một lượt đặt phòng khách sạn. Các cột trong bảng dữ liệu được mô tả trong bảng 1
- Trong đó có 2 cột thiếu dữ liệu nhiều nhất với cột `company` hơn 94% dòng dữ liệu bị thiếu và cột ‘agent‘ là hơn 13
- Số trường dữ liệu bị thiếu ở mỗi dòng: Trong đó 91% các dòng bị thiếu 1 trường dữ liệu, 8% các dòng bị thiếu 2 trường dữ liệu.
- Tỷ lệ dòng dữ liệu bị trùng lặp là 26.8
- Nhóm đã điền tất cả các giá trị thiếu trong bộ dữ liệu bằng giá trị `-1` để có thể dùng đến các cột có giá trị thiếu trong quá trình phân tích dữ liệu
- Sau khi điền giá trị thiếu, xóa những dòng dữ liệu trùng lặp, bộ dữ liệu còn lại gồm 87396 dòng và 32 cột

Bảng 1: Ý nghĩa các cột trong tập dữ liệu

Tên cột	**Ý nghĩa**
hotel	Tên khách sạn (Resort Hotel hoặc City Hotel)
is_canceled	1 nếu lượt đặt bị hủy, 0 nếu không
lead_time	Số ngày giữa ngày đặt phòng và ngày nhận phòng
arrival_date_year	Năm của ngày nhận phòng
arrival_date_month	Tháng của ngày nhận phòng
arrival_date_week_number	Số tuần trong năm của ngày nhận phòng
arrival_date_day_of_month	Ngày nhận phòng
stays_in_weekend_nights	Số đêm cuối tuần (Thứ 7 và Chủ Nhật) mà khách ở hoặc đặt phòng để ở
stays_in_week_nights	Số đêm trong tuần (từ Thứ 2 đến Thứ 6) mà khách ở hoặc đặt phòng để ở
adults	Số người lớn
children	Số trẻ em
babies	Số em bé
meal	Loại bữa ăn được đặt (Undefined/SC – không bao gồm bữa ăn, BB – Bữa sáng, HB – Bữa sáng và bữa tối, FB – Bữa sáng, bữa trưa và bữa tối)
country	Mã quốc gia của khách (dựa trên ISO 3155-3:2013)
market_segment	Phân khúc thị trường của khách hàng thuộc về
distribution_channel	Kênh phân phối mà qua đó lượt đặt được thực hiện
is_repeated_guest	1 nếu khách đã từng đặt phòng trước đó, 0 nếu không
previous_cancellations	Số lần hủy đặt phòng trước đó của khách
previous_bookings_not_canceled	Số lần đặt phòng trước đó mà không bị hủy của khách
reserved_room_type	Mã loại phòng đã đặt
assigned_room_type	Mã loại phòng được gán cho lượt đặt
booking_changes	Số lần thay đổi thông tin đặt phòng trước khi nhận phòng hoặc trước khi hủy
deposit_type	Loại tiền đặt cọc đã đặt
agent	ID của đại lý đặt phòng
company	ID của công ty (thực thể) đặt phòng hoặc trả tiền cho lượt đặt
days_in_waiting_list	Số ngày mà lượt đặt đã nằm trong danh sách chờ trước khi được xác nhận
customer_type	Loại đặt phòng (Contract, Group, Transient, Transient-Party)
adr	Tỉ lệ giá phòng mỗi đêm
required_car_parking_spaces	Số lượng chỗ đậu xe mà khách yêu cầu
total_of_special_requests	Số lượng yêu cầu đặc biệt từ khách
reservation_status	Trạng thái cuối cùng của lượt đặt (Canceled, Check-Out, No-Show)
reservation_status_date	Ngày cuối cùng mà trạng thái cuối cùng được cập nhật

2 EDA 1D

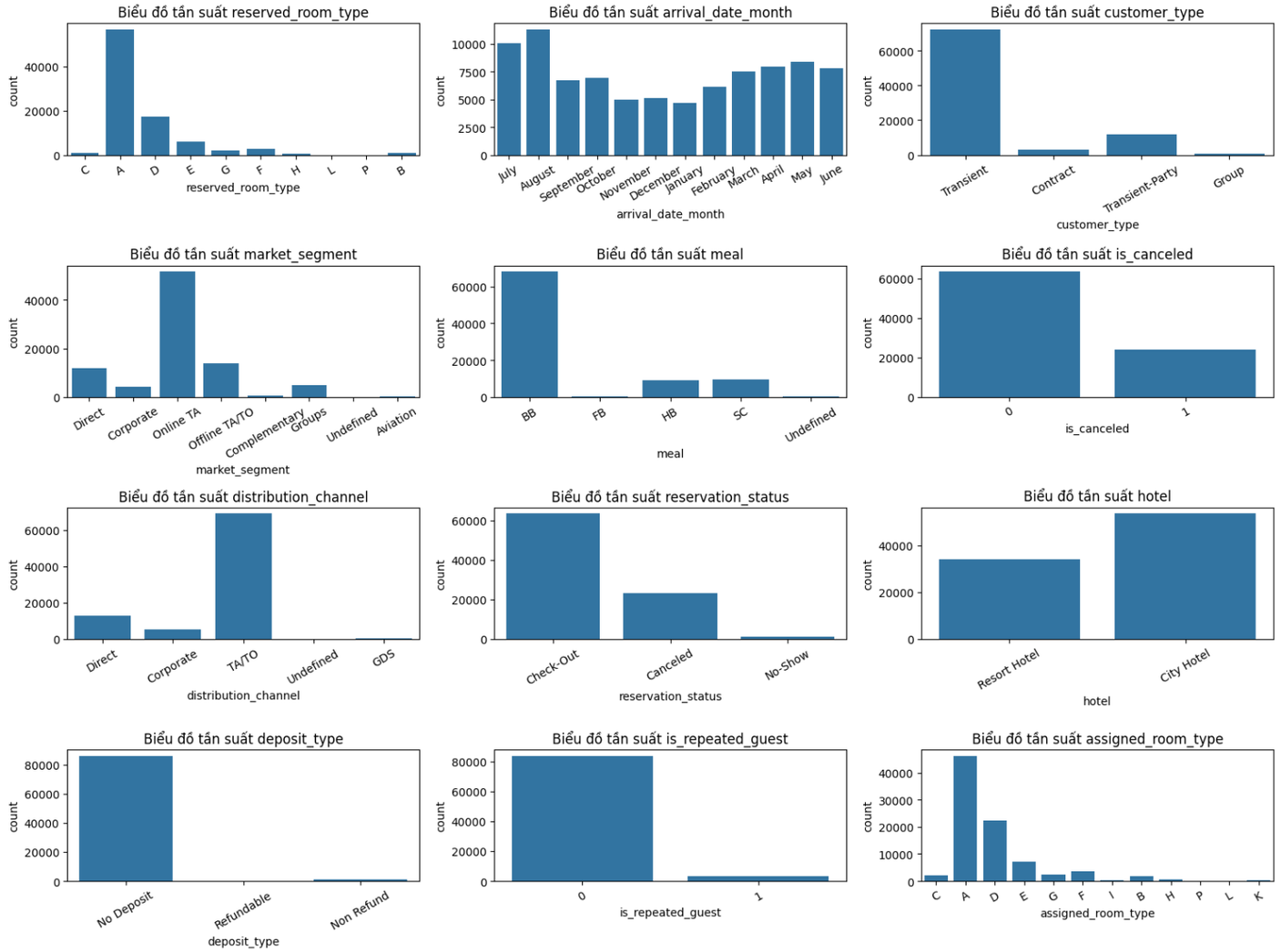
2.1 Chia loại dữ liệu numerical hoặc categorical

Vì hai cột dữ liệu `is_canceled` và `is_repeated_guest` chỉ có giá trị 0 và 1 nên nhóm sẽ coi chúng là biến categorical. Sau đó nhóm sẽ phân tách các cột dữ liệu thành 2 nhóm: numerical và categorical

```
df['is_canceled'] = df['is_canceled'].astype('object')
df['is_repeated_guest'] = df['is_repeated_guest'].astype('object')
cate_list = list(df.dtypes[df.dtypes == 'object'].index)
num_list = list(df.dtypes[df.dtypes != 'object'].index)
```

2.2 Phân tích phân phối đối với biến categorical

- Nhóm dùng hàm `describe` để tính toán các giá trị thống kê mô tả cho các biến: `unique`, `top`, `freq`.
- Nhóm đã sử dụng biểu đồ cột và biểu đồ tròn để trực quan phân phối và tỉ lệ của các giá trị của các biến categorical ở Hình 1 và Hình 2

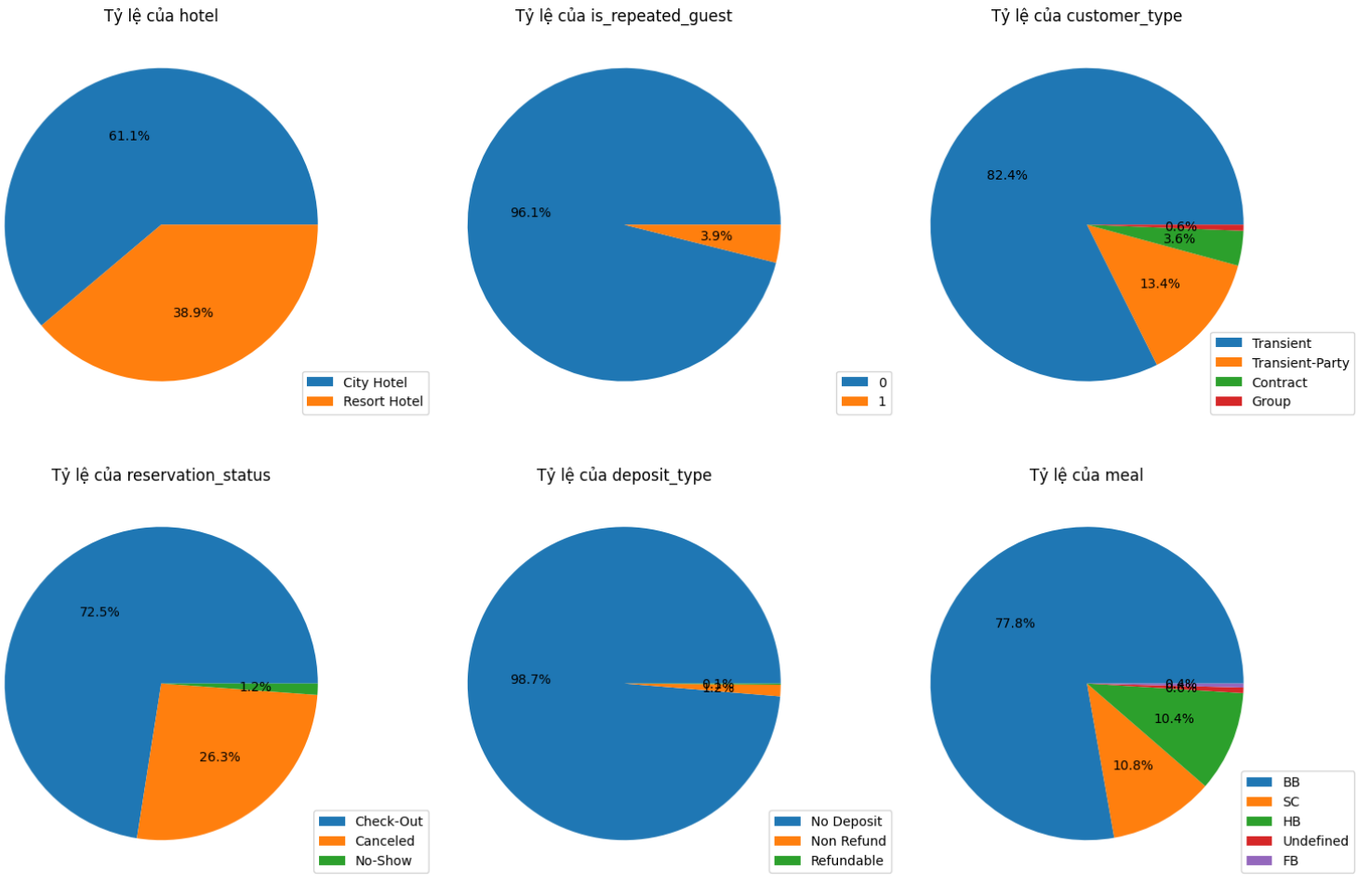


Hình 1: Biểu đồ thể hiện giá trị của các biến categorical

Nhận xét:

- **is_repeated_guest**: Hầu hết khách hàng đặt phòng không phải là khách hàng quay lại.
- **is_canceled**: Có một số lượng lớn đặt phòng đã được hủy (xấp xỉ $\frac{1}{3}$ số lượng đặt phòng).
- **reservation_status**: Đa số các đặt phòng đã được thanh toán và trải qua quá trình *check-out*, một lượng đáng kể bị hủy và chỉ một phần nhỏ là không có mặt (*no-show*).
- **hotel**: Có nhiều đặt phòng tại khách sạn thành phố (*City Hotel*) hơn là khách sạn nghỉ dưỡng (*Resort Hotel*) (gấp 1.5 lần).

- **distribution_channel**: Kênh phân phối chủ yếu cho đặt phòng là qua đại lý (*TA/TO*) với số lần đặt phòng nhiều nhất.
- **reserved_room_type**: Loại phòng được đặt nhiều nhất là loại A, tiếp theo là loại D và các loại phòng khác có số lượng đặt ít hơn rất nhiều.
- **arrival_date_month**: Tháng có số lượng đặt phòng cao nhất là tháng 8 (*August*), và thấp nhất vào tháng 1 (*January*). Ta có thể suy ra mùa cao điểm là mùa hè, còn mùa thấp điểm là mùa đông.
- **market_segment**: Phân khúc thị trường chính là *Online TA* với số lượng đặt phòng nổi bật hơn hẳn so với các phân khúc khác.
- **deposit_type**: Đa số các đặt phòng không đặt cọc (*No Deposit*). Có rất ít đặt phòng đặt cọc hoàn lại (*Refundable*) hoặc không hoàn lại (*Non Refund*).
- **customer_type**: Khách hàng chủ yếu là khách hàng lưu trú qua đêm (*Transient*), số lượng lớn hơn nhiều so với các loại khách hàng khác như hợp đồng (*Contract*) hay nhóm (*Group*).
- **assigned_room_type**: Giống như với loại phòng được đặt, loại phòng được chỉ định cũng chủ yếu là loại A, sau đó là loại D và các loại phòng khác có số lượng được chỉ định ít hơn nhiều.
- **meal**: Hầu hết khách hàng chọn gói ăn sáng (*BB* - Bed & Breakfast), tiếp theo là HB (bán pension) và FB (đầy đủ pension) với số lượng thấp hơn.



Hình 2: Biểu đồ thể hiện giá trị của các biến categorical

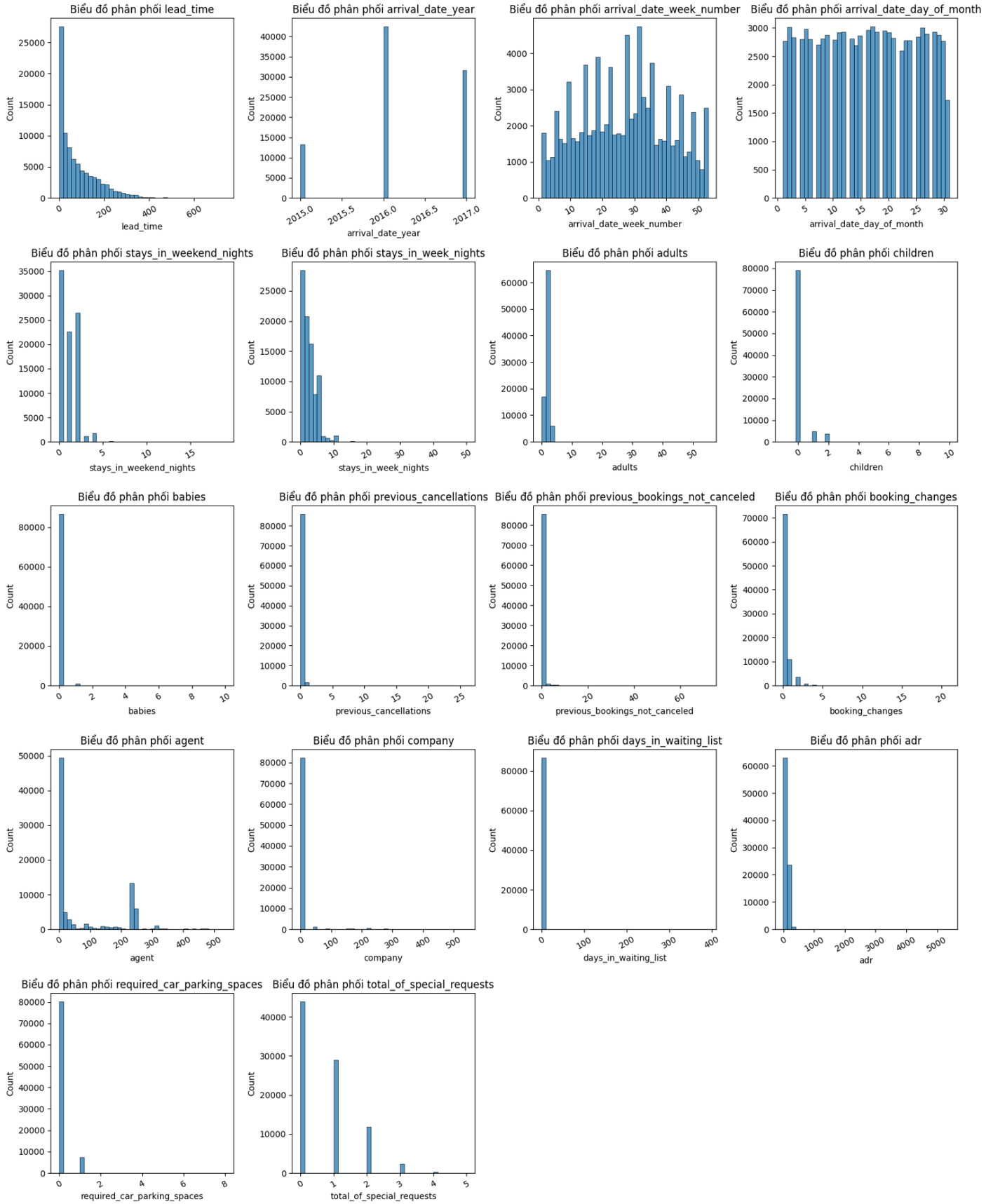
Nhận xét:

- Tỷ lệ đặt phòng tại **City Hotel** trong tập dữ liệu lớn hơn nhiều so với **Resort Hotel**, cụ thể là gấp 1.5 lần.
- Đa phần khách hàng đặt phòng khách sạn đều là khách mới, chỉ có khoảng 4% là khách cũ.
- Khách hàng chủ yếu thuộc loại (*Transient*), một phần rất ít thuộc loại *Group* và *Contract*.
- Hơn 70% trạng thái đặt phòng được *Check-Out*, khoảng 26% còn lại là *Canceled*. Vẫn có hơn 1% đặt phòng nhưng không có mặt (*No-Show*).
- Hầu hết (gần 99%) đặt phòng đều thuộc loại không đặt cọc (*No Deposit*).
- Hơn 3/4 khách hàng chọn gói ăn sáng (*BB* - Bed & Breakfast), còn lại là gói SC và HB chiếm

khoảng 20%.

2.3 Phân tích phân phối đối với biến numerical

- Nhóm dùng hàm `describe` để tính toán các giá trị thống kê mô tả cho các biến: `min, max, mean, std, 25%,50%,75%,`
- Nhóm sử dụng histogram để trực quan phân phối cho tất cả các biến numerical được thể hiện ở Hình 3



Hình 3: Histogram của các biến numerical

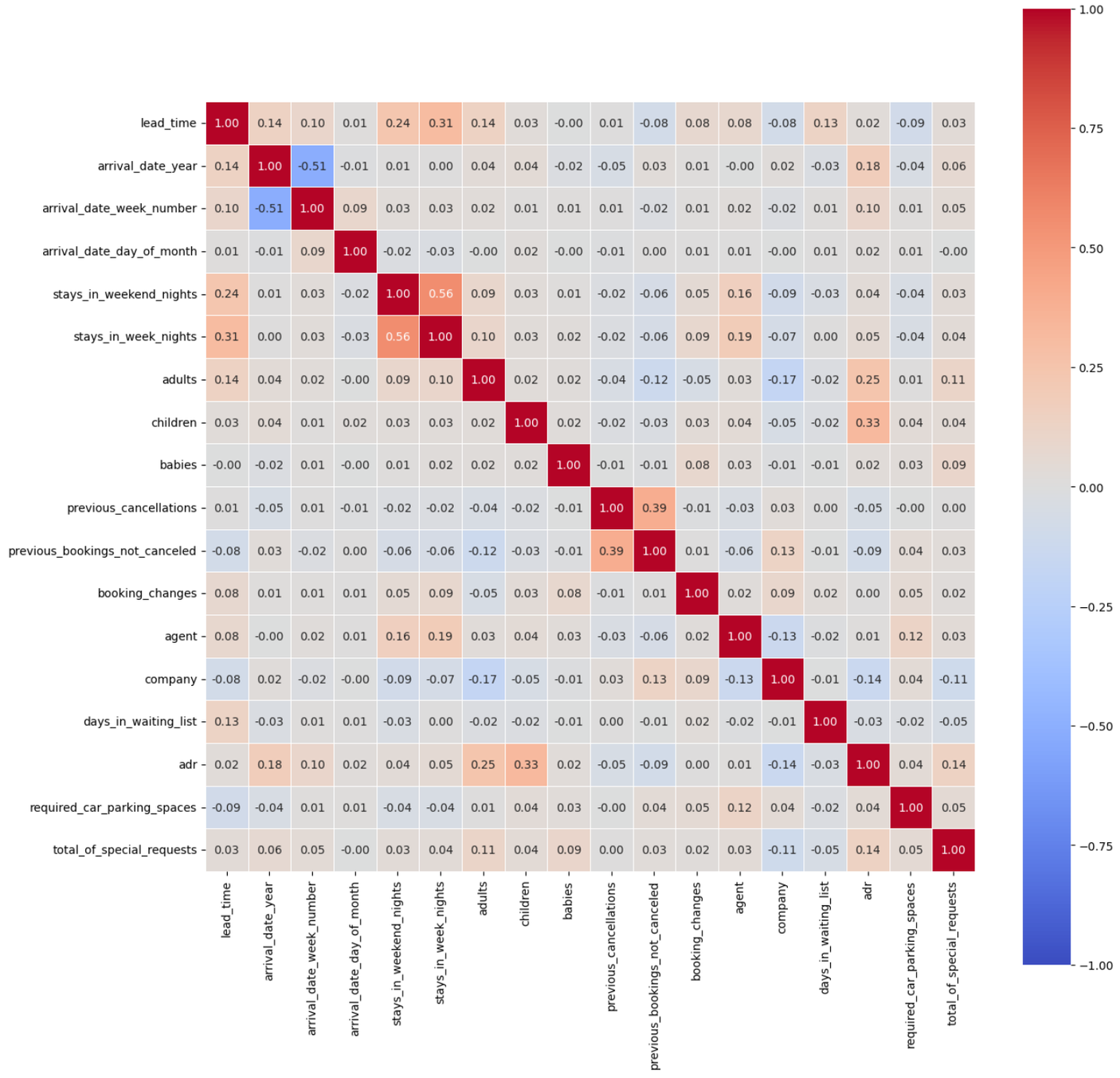
Nhận xét:

- **lead_time**: Đa số các đặt phòng được thực hiện với thời gian chờ dẫn đầu (lead time) khá ngắn, chỉ khoảng vài ngày đến một tuần. Có một lượng nhỏ các đặt phòng được thực hiện với thời gian chờ dài hơn.
- **arrival_date_year**: Số lượng đặt phòng tăng đáng kể trong năm 2016 và 2017 so với năm 2015.
- **arrival_date_week_number** và **arrival_date_day_of_month**: Không có sự biến đổi lớn về mùa vụ hoặc ngày cụ thể trong tháng - phân phối khá đồng đều qua các tuần và ngày trong tháng.
- **stays_in_weekend_nights** và **stays_in_week_nights**: Số đêm nghỉ phổ biến nhất là từ 0 đến 2 đêm.
- **adults**, **children**, và **babies**: Phần lớn các đặt phòng được làm cho người lớn, với số lượng trẻ em và trẻ sơ sinh ít hơn nhiều.
- **previous_cancellations** và **previous_bookings_not_canceled**: Đa số khách hàng chưa từng hủy bỏ đặt phòng trước đó hoặc có lịch sử đặt phòng mà không hủy bỏ.
- **booking_changes**: Hầu hết đặt phòng không có sự thay đổi sau khi được đặt.
- **days_in_waiting_list**: Phần lớn khách hàng không phải chờ đợi để đặt phòng sau khi yêu cầu.
- **required_car_parking_spaces** và **total_of_special_requests**: Hầu hết các đặt phòng không yêu cầu chỗ đậu xe và có ít hoặc không có yêu cầu đặc biệt.

3 EDA 2D

Nhóm dùng heatmap để trực quan ma trận hệ số tương quan cho các biến numerical ở Hình 4

3.1 Phân tích hệ số tương quan giữa các biến numerical



Hình 4: Ma trận tương quan giữa các biến numerical

Nhận xét

- **stays_in_week_nights** và **stays_in_weekend_nights** có tương quan dương lớn (0.56). Điều này cho thấy khách hàng có xu hướng lưu trú cả trong tuần lẫn cuối tuần. Nói cách khác, những

khách hàng chọn lưu trú nhiều đêm trong tuần thường cũng chọn lưu trú nhiều đêm vào cuối tuần. Điều này có thể phản ánh một xu hướng chung trong việc đặt phòng dài hạn hơn, không chỉ giới hạn ở việc lưu trú ngắn ngày.

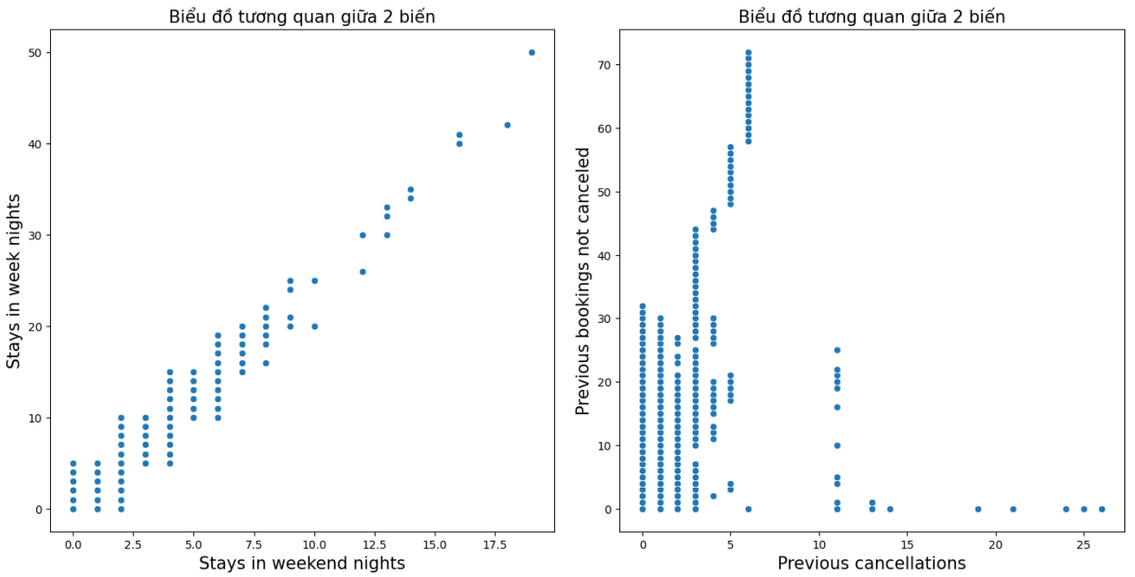
- **lead_time** và **stays_in_week_nights** có tương quan dương vừa phải (0.31), có thể cho thấy khách hàng thường đặt phòng trước càng lâu thì càng có xu hướng lưu trú nhiều đêm trong tuần hơn.
- **adr** (Average Daily Rate) có một tương quan dương mạnh với **adults** (0.25) và **children** (0.33), nhưng lại có mức tương quan thấp với **babies** (0.04). Điều này cho thấy giá trung bình hàng ngày có liên quan đến số người lớn và trẻ em, nhưng không nhiều đối với số lượng em bé.
- **arrival_date_year** và **lead_time** có tương quan dương yếu (0.14), có thể cho thấy rằng việc đặt phòng trước không có sự thay đổi đáng kể theo năm.
- **agent** có tương quan nhỏ với hầu hết các biến khác, và một tương quan dương yếu với **booking_changes** (0.13), có thể là do việc đặt phòng qua đại lý có những điều chỉnh nhất định.
- **required_car_parking_spaces** và **adr** có một tương quan lên tới (0.14), cho thấy có thể khách đặt phòng với giá cao hơn có xu hướng yêu cầu nhiều chỗ đậu xe hơn.
- **total_of_special_requests** có một tương quan dương yếu với **booking_changes** (0.10), điều này có thể đến từ việc những khách hàng có nhiều yêu cầu đặc biệt hơn cũng thường xuyên thực hiện thay đổi đối với đặt phòng của họ.

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

Vì bộ dữ liệu có khá nhiều biến nhưng độ tương quan của chúng đa số khá thấp nên nhóm sẽ chỉ dùng scatter plot để phân tích 2 biến numeric có độ lớn hệ số tương quan >=0.35 từ biểu đồ heatmap ở phần trên

- **stay_in_week_nights** và **stay_in_weekend_nights**

- **previous_bookings_not_canceled** và **previous_cancellations**
- **arrival_date_year** và **arrival_date_week_number**: Tuy hai biến này có độ lớn hệ số tương quan cao nhưng đề là biến về thời gian nên không có ý nghĩa khi phân tích, nhóm sẽ bỏ qua chúng.



Hình 5: 2 biểu đồ scatter cho 2 cặp biến có độ tương quan cao

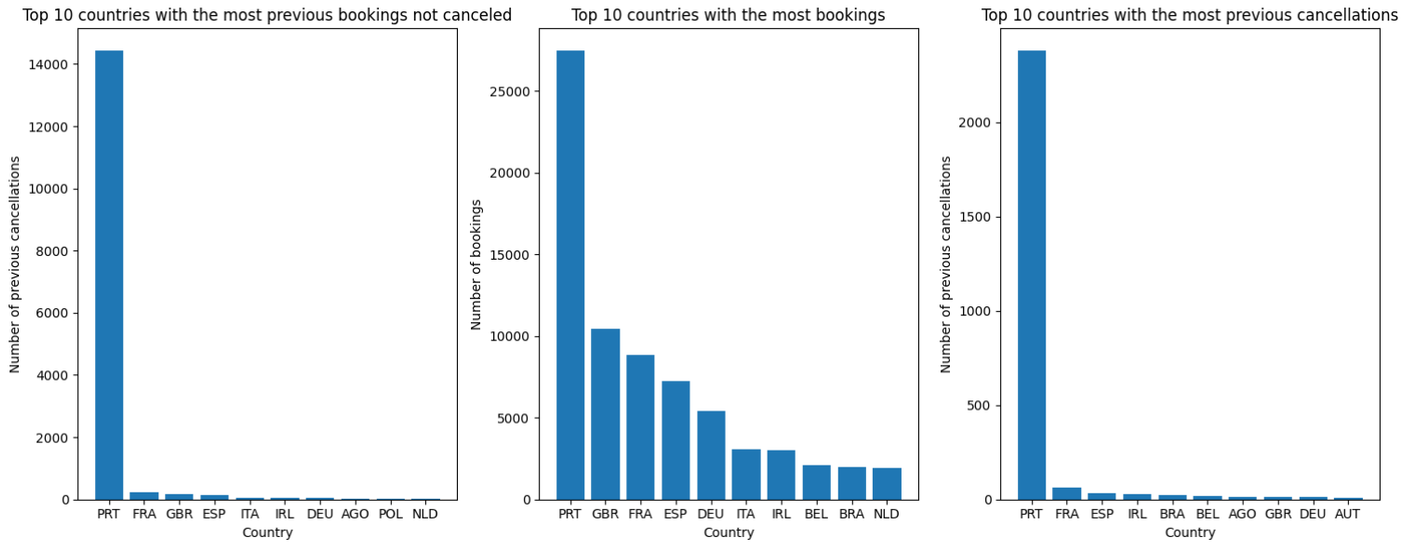
Nhận xét:

- Dựa vào biểu đồ ta thấy có tương quan dương cao giữa số đêm lưu trú trong tuần và số đêm lưu trú vào cuối tuần.
- Khi số đêm lưu trú trong tuần tăng thì số đêm lưu trú vào cuối tuần cũng tăng theo. Điều này cũng hợp lí ở các lượt đặt phòng dài hạn thì họ phải lưu trú cả trong tuần lẫn cuối tuần.
- Do đó biểu đồ cũng có thể nói lên xu hướng khách hàng thường chọn cả lưu trú dài hạn hơn chứ không chỉ giới hạn ở việc lưu trú ngắn ngày.
- Có một số lượng lớn các điểm dữ liệu tập trung ở góc dưới bên trái của biểu đồ, cho thấy rằng hầu hết khách hàng không hủy đặt phòng khi họ không có lịch sử hủy bỏ trước đó.
- Khi số lượng hủy bỏ trước đó tăng lên từ số lần thứ 10, số lượng đặt phòng không hủy giảm

xuống rõ rệt.

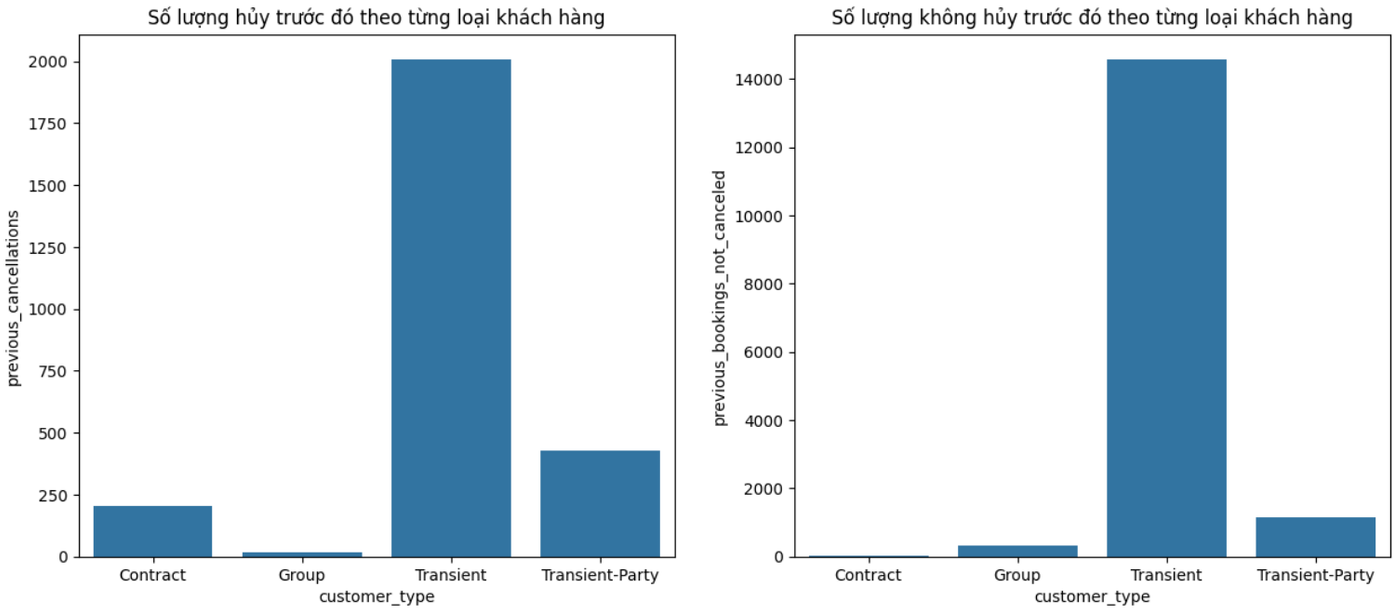
3.3 Sử dụng bar chart để phân tích dữ liệu num và cate

- Nhóm sử dụng `groupby` để gom nhóm các quốc gia sau đó tính tổng các biến: `previous_bookings_not_canceled`. Rồi lấy 10 giá trị lớn nhất.
- Sau đó sử dụng biểu đồ cột để trực quan các giá trị trên ở Hình 6
- Tương tự ta cũng trực quan cho biến Customer Type ở hình 7 và biến Markey ở hình 8



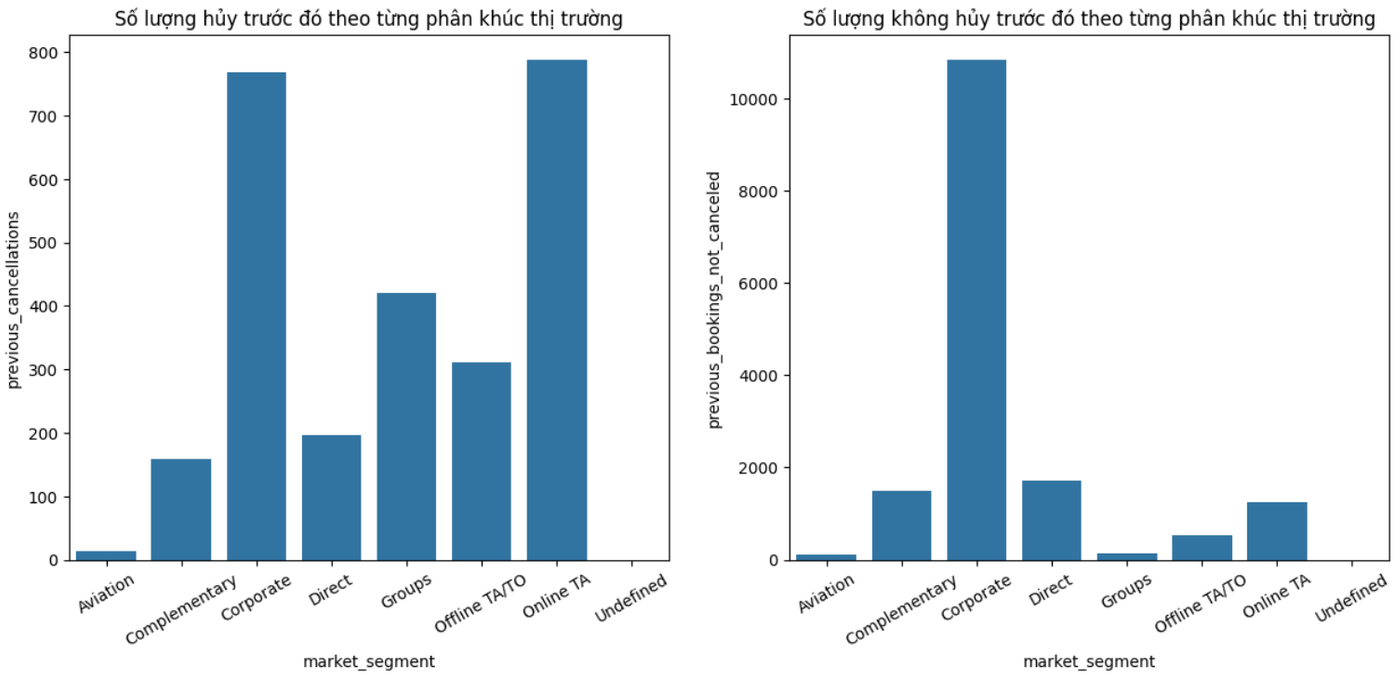
Hình 6: Biểu đồ cột về country

- Dựa vào biểu đồ trên ta thấy quốc gia PRT (Bồ Đào Nha) dẫn đầu cả ba biểu đồ, điều này cho thấy rằng không chỉ có số lượng đặt phòng mà cũng có tỷ lệ hủy phòng cao đến từ quốc gia này.
- Trong top 5 quốc gia có số lượng đặt phòng lớn nhất gồm GBR (Anh Quốc), FRA (Pháp), ESP (Tây Ban Nha), DEU (Đức) nhưng số lượng hủy đặt phòng trước đó có tỷ lệ rất thấp.



Hình 7: Biểu đồ cột về customer_type

- Khách hàng loại `Transient` có tỷ lệ hủy đặt phòng cao nhất so với các loại khách hàng khác. Điều này thể hiện qua cột màu xanh lá cây cao vượt trội trong biểu đồ hủy đặt phòng.
- Mặc dù tỷ lệ hủy đặt phòng của khách hàng `Transient` cao, nhưng số lượng đặt phòng không hủy của loại khách hàng này cũng là cao nhất. Điều này có thể là do tổng số lượng đặt phòng của khách hàng `Transient` rất lớn, do đó cả số lượng hủy đặt và số lượng không hủy đặt đều cao.
- Khách hàng loại `Contract` có số lượng hủy đặt phòng trên tổng số lượng đặt phòng loại `Contract` khá là lớn



Hình 8: Biểu đồ cột về market

- Phân khúc Corporate có số lượng hủy đặt phòng cao nhất trong tất cả các phân khúc thị trường được nêu, với gần 800 hủy đặt phòng.
- Tuy nhiên, phân khúc Corporate cũng có lượng đặt phòng không bị hủy cao nhất, với hơn 10.000 đặt phòng không hủy.
- Phân khúc Groups có số lượng hủy đặt phòng khá cao, nhưng số lượng đặt phòng không hủy lại rất thấp

3.4 Tính tỷ trọng đối với hai biến cate

- Ở phần này nhóm cũng đã sử dụng `groupby` để gom nhóm 2 biến categorical. Và thực hiện phép tính `count` cho 1 trong 2 biến categorical đó.
- Ta sẽ được 2 bảng như sau ở Hình 9 và ở Hình 10

		reservation_status
hotel	reservation_status	
City Hotel	Canceled	17.51
	Check-Out	42.77
	No-Show	0.86
Resort Hotel	Canceled	8.82
	Check-Out	29.74
	No-Show	0.30

Hình 9: Tỷ trọng giữa 2 biến hotel và reservation_status

- Ở cả 2 loại hotel status Check-Out đều chiếm phần lớn cụ thể: ở Resort Hotel chiếm hơn 75% và ‘City Hotel‘ chếm hơn 70%

		deposit_type
hotel	deposit_type	
City Hotel	No Deposit	60.15
	Non Refund	0.97
	Refundable	0.02
Resort Hotel	No Deposit	38.54
	Non Refund	0.22
	Refundable	0.11

Hình 10: Tỷ trọng giữa 2 biến hotel và deposit_type

- Như bảng đã cho thì ta thấy được tỷ lệ của các loại deposit_type đều tương đối như nhau ở cả

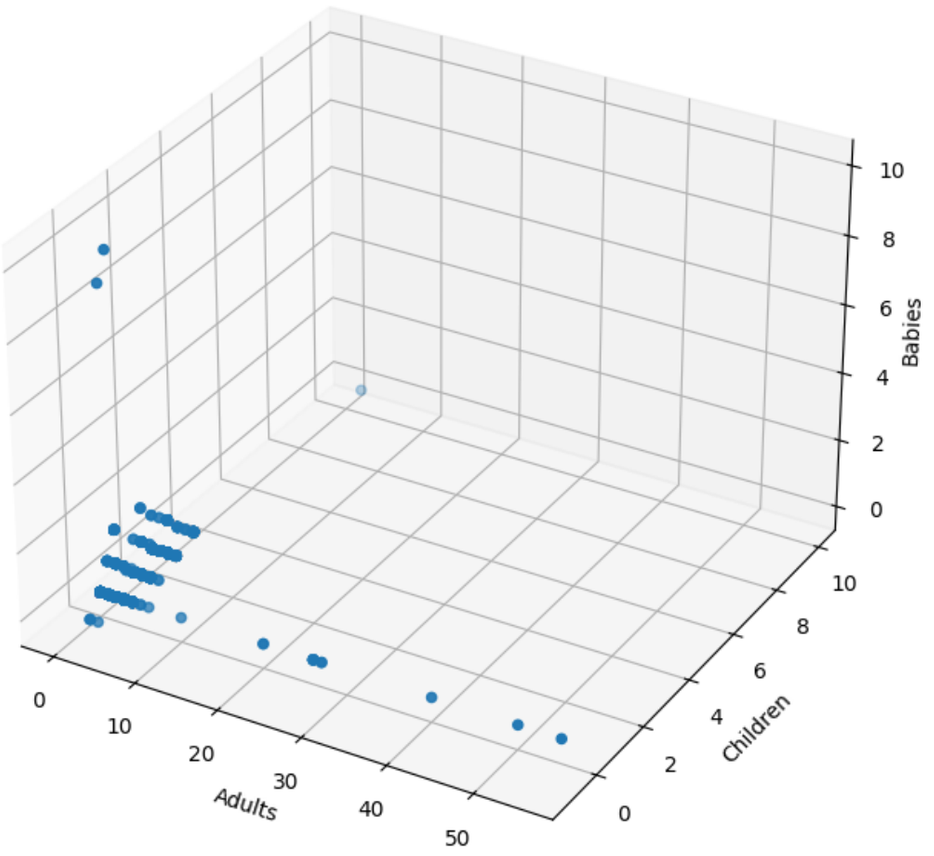
2 loại hotel. Tỷ lệ cao nhất ở City Hotel và Resort Hotel đều là No Deposit với tỷ lệ lần lượt là 60% và 38%.

4 EDA 3D

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

Nhóm chọn ra các bộ 3 biến sau để trực quan 3D:

- **adults**, **children** và **babies**



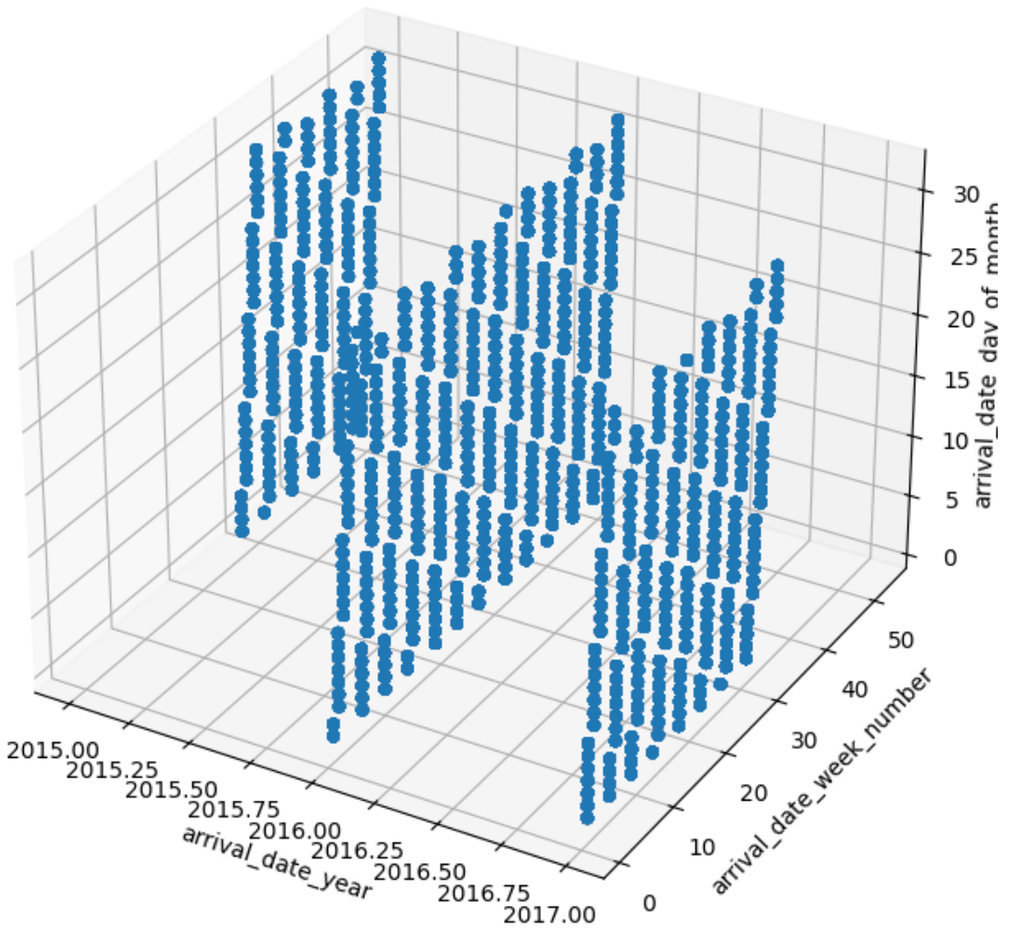
Hình 11: Scatter plot 3D của adults, children và babies

Nhận xét:

– Từ biểu đồ ta có thể thấy chủ yếu các điểm dữ liệu tập trung ở góc trái phía dưới (gần gốc tọa độ), cho thấy phần lớn các đặt phòng đều là cho người lớn, và chỉ một số ít đặt phòng cho trẻ em hoặc em bé.

– Chủ yếu các đặt phòng cho trẻ em và em bé đều là khi có người lớn đi kèm, và số lượng trẻ em và em bé không quá nhiều. Đa số rơi vào 1-4 trẻ em và 1-2 em bé.

• **arrival_date_year**, **arrival_date_month** và **arrival_date_day_of_month**



Hình 12: Scatter plot 3D của arrival_date_year, arrival_date_month và arrival_date_day_of_month

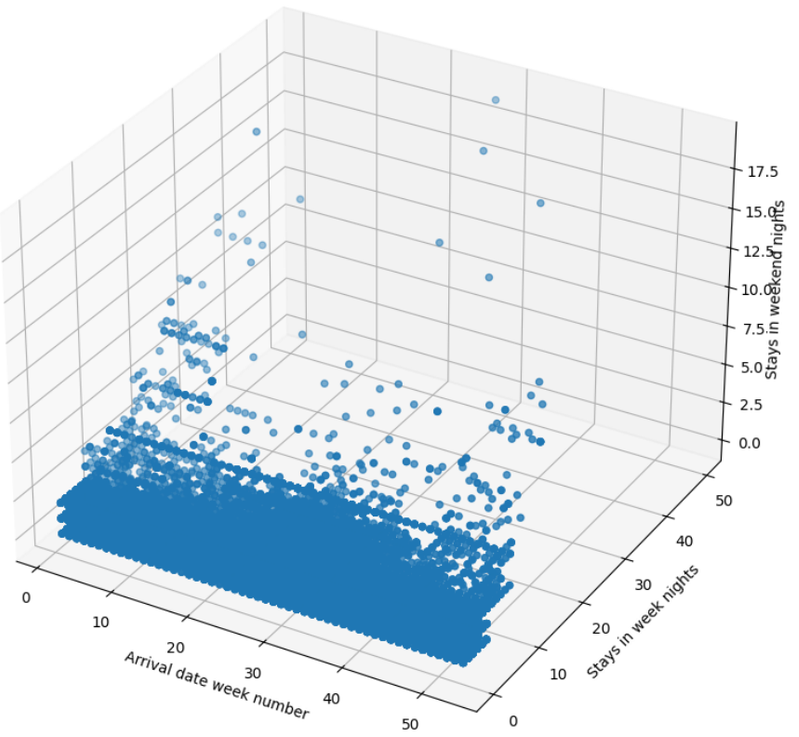
Nhận xét:

– Biểu đồ có 3 trục tọa độ: trục x (năm), trục y(số tuần trong năm) và trục z (ngày trong

tháng)

- Từ biểu đồ ta thấy rằng số điểm dữ liệu bắt đầu từ giữa năm 2015 đến giữa năm 2017, và phân bố khá đều qua các tuần trong năm và các ngày trong tháng.
- Mỗi ngày trong tháng có số lượng đặt phòng khá đồng đều, không có sự biến đổi lớn qua các ngày trong tháng

- **arrival_date_week_number**, **stays_in_weekend_nights** và **stays_in_week_nights**



Hình 13: Scatter plot 3D của arrival_date_week_number, stays_in_weekend_nights và stays_in_week_nights

Nhận xét:

- Nhìn vào biểu đồ ta có thể thấy một dải các điểm dữ liệu dày đặc tập trung dọc gần trục *arrival_date_week_number*, cho thấy rằng số lượng đặt phòng khá nhiều xuyên suốt các tuần trong năm.

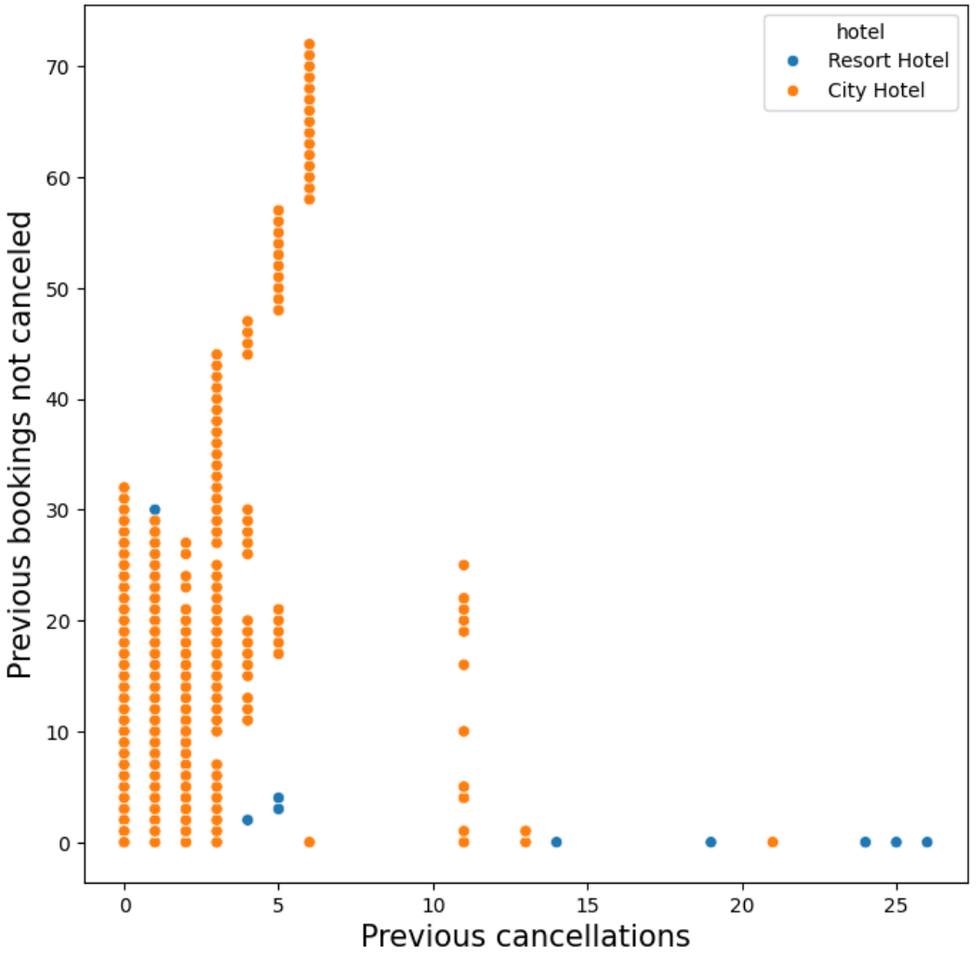
– Trục *stays_in_weekend_nights* và *stays_in_week_nights* cho biết thời gian ở lại cuối tuần và trong tuần, và ta thấy rằng có chủ yếu các lượt đặt phòng ở lại khoảng 1-15 ngày trong tuần và 1-4 ngày vào cuối tuần.

– Ở các tuần đầu năm và cuối năm ta thấy số lượng đặt phòng có xu hướng ở lại nhiều hơn vào cuối tuần khi các điểm dữ liệu bắt đầu xuất hiện ở tọa độ cao hơn trên trục *stays_in_weekend_nights*.

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

Nhóm chọn ra lần lượt các bộ 2 biến num và 1 biến cate để trực quan và phân tích:

- **previous_bookings_not_canceled**, **previous_cancellations** và **hotel**



Hình 14: Scatter plot 3D của previous_cancellations, previous_bookings_not_canceled và hotel

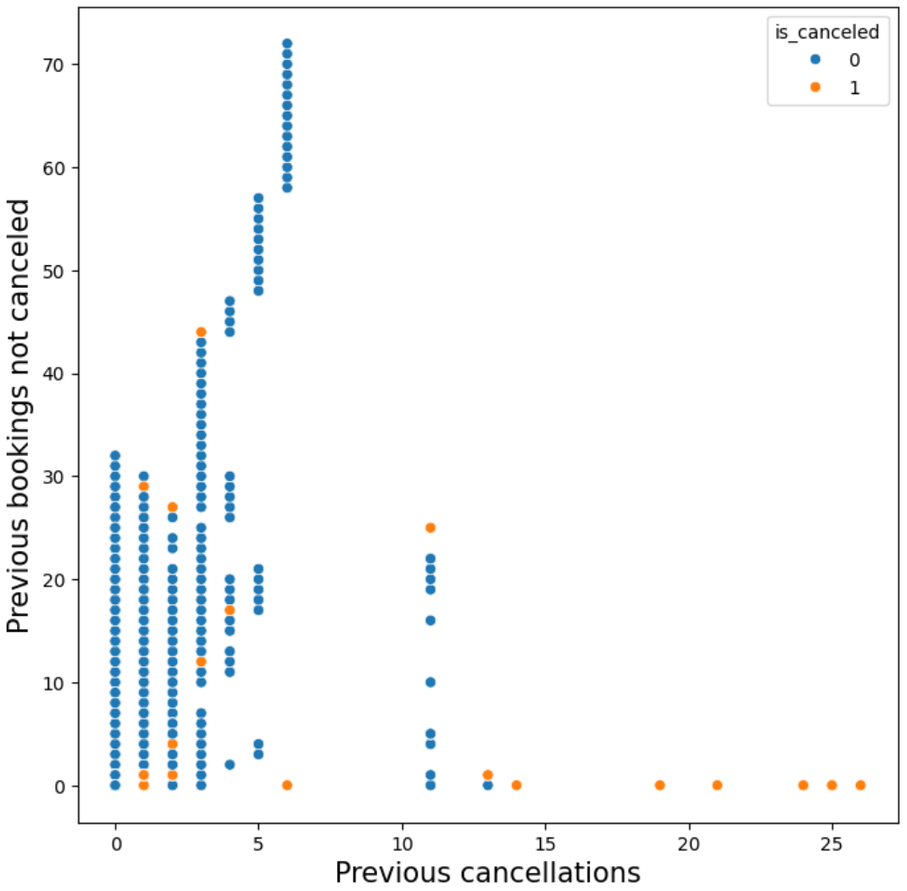
Nhận xét:

– **City Hotel** có số lượng đặt phòng không bị hủy nhiều hơn so với **Resort Hotel**.

– Ở phía trái trên của biểu đồ, chúng ta có thể thấy rằng với số lần hủy trước đó thấp, **City Hotel** thường có số lượng đặt phòng không bị hủy cao hơn so với **Resort Hotel**.

– Tại các giá trị thấp hơn của *Previous cancellations*, có nhiều đặt phòng không bị hủy hơn, trong khi ở các giá trị cao hơn, số lượng đặt phòng không bị hủy giảm mạnh cho thấy có thể có một xu hướng tiêu cực giữa hai biến được biểu diễn.

- **previous_bookings_not_canceled**, **previous_cancellations** và **is_canceled**



Hình 15: Scatter plot 3D của previous_cancellations, previous_bookings_not_canceled và is_canceled

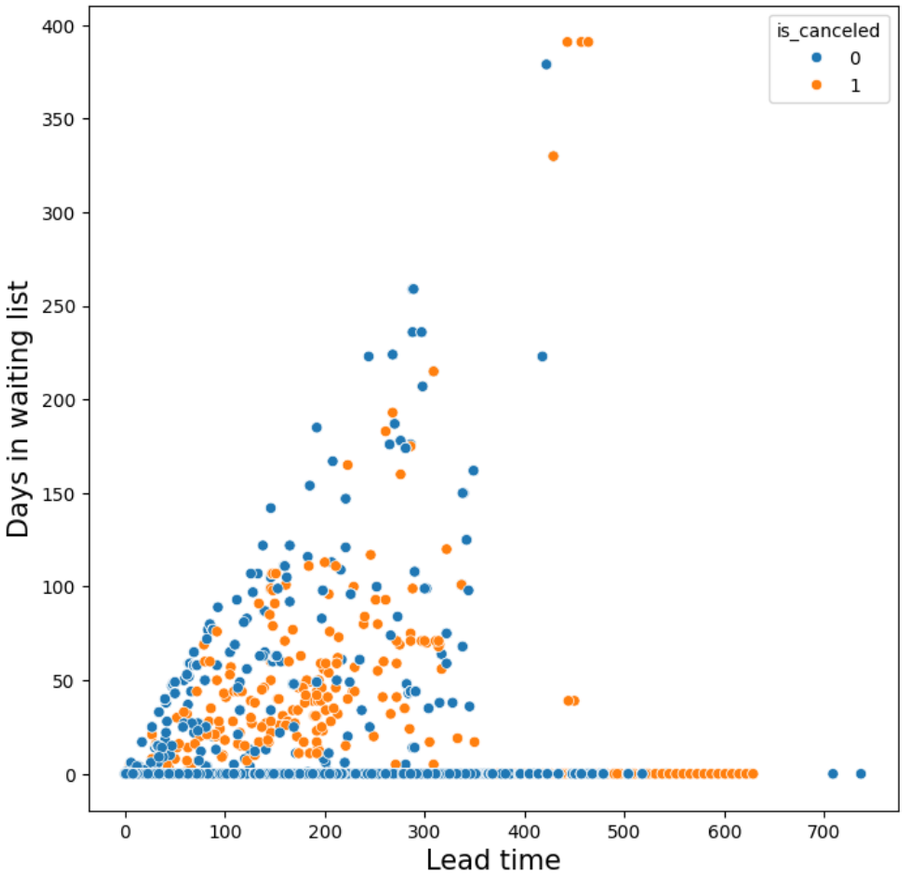
Nhận xét:

– Chúng ta có thể nhìn thấy một số điểm màu cam đại diện cho các trường hợp đặt phòng đã bị hủy bỏ (`is_canceled` = 1). Điều này cho thấy rằng khách hàng có một số lượng hủy bỏ đặt phòng trước đây vẫn có khả năng hủy bỏ đặt phòng trong tương lai.

– Ở phía trái, ta thấy các điểm với số lượt hủy đặt phòng trước đó ít và số lượng không hủy đặt phòng trước đó cao, thì phần lớn đều sẽ không hủy đặt phòng (`is_canceled` = 0).

– Mặc dù có vài trường hợp hủy bỏ (`is_canceled` = 1) nằm trong phạm vi có `Previous cancellations` cao, nhưng không thể kết luận chắc chắn về quy luật cố định nào từ biểu đồ này.

• **lead_time**,**days_in_waiting_list** và **is_canceled**



Hình 16: Scatter plot 3D của lead_time, days_in_waiting_list và is_canceled

Nhận xét:

– Đa số đặt phòng không bị hủy bỏ (điểm màu xanh) có `lead time` thấp và ít ngày nằm trong danh sách chờ.

– Cũng có một lượng đặt phòng đáng kể bị hủy bỏ (điểm màu cam) với `lead time` và số ngày nằm trong danh sách chờ thấp.

– Một số ít số lượng đặt phòng với `lead time` rất cao (trên 400 ngày) hoặc nằm rất lâu trong danh sách chờ (trên 250 ngày) thường bị hủy bỏ.

– Phần lớn dữ liệu tập trung vào khu vực có `lead time` từ thấp đến trung bình (khoảng 0-200 ngày) và số ngày nằm trong danh sách chờ thấp (dưới 50 ngày), trong đó số đặt phòng không bị hủy chiếm đa số.

4.3 Tính tỉ trọng theo bin chia theo thể loại với hai biến cate

Nhóm đã chọn các bộ gồm hai biến cate và một biến num

- **hotel** và**is_repeated_guest** và **arrival_date_week_number**

		arrival_date_week_number			
		min	mean	std	max
hotel	is_repeated_guest				
City Hotel	0	1	26.643774	13.504116	53
	1	1	26.450820	15.877587	53
Resort Hotel	0	1	27.414835	13.639095	53
	1	1	22.225542	15.876948	53

Hình 17: Tỷ trọng chia theo bin của các biến

Nhận xét:

– Bảng trên tính các giá trị thống kê của các tuần mà khách đặt phòng của 2 khách sạn (khách cũ và khách mới).

– Ta thấy rằng cả khách cũ và mới, cũng như cả hai khách sạn đều có thời gian đặt phòng trải dài từ tuần đầu đến tuần cuối của năm (tuần 1 - tuần 53).

– Cả khách quen và khách mới ở **City Hotel** đều có tuần trung bình đặt phòng rơi vào khoảng tuần 26-27 trong khi ở **Resort Hotel** thì có sự khác biệt giữa khách cũ và mới (tuần 27-28 và tuần 22-23).

– Độ lệch chuẩn của tuần đặt phòng của khách cũ của hai khách sạn đều có giá trị khoảng 15 tuần. Độ lệch chuẩn của tuần đặt phòng của khách mới cũng giống nhau ở hai khách sạn, khoảng 13 tuần.

• **hotel**, **is_canceled** và **lead_time**

		lead_time			
		min	mean	std	max
hotel	is_canceled				
City Hotel	0	0	67.387999	75.542625	518
	1	0	101.645710	91.294653	629
Resort Hotel	0	0	73.999115	89.519625	737
	1	0	113.915873	92.502393	471

Nhận xét:

– Bảng trên tính các giá trị thống kê của thời gian cách biệt giữa ngày đặt phòng và ngày đến của 2 khách sạn (lượt đặt phòng hủy và không hủy)

– Thời gian cách biệt giữa ngày đặt phòng và ngày đến của hai khách sạn trải khá rộng từ 0 đến hơn 700 ngày (khoảng 2 năm)

– Ta thấy trung bình cũng như độ lệch chuẩn của thời gian cách biệt giữa ngày đặt phòng và ngày đến của các lượt đặt phòng không bị hủy đều nhỏ hơn so với các lượt đặt phòng bị hủy (ở cả hai khách sạn). Điều này cũng có thể nói lên khi thời gian cách biệt giữa ngày đặt và

ngày đến càng lớn thì khả năng hủy đặt phòng càng cao.

5 Insight

Sau các phần trực quan và phân tích ở trên, nhóm đã rút kết ra được các insight về bộ dữ liệu:

5.1 Data Understanding

- Tập dữ liệu gồm 119390 dòng và 32 cột. Mỗi dòng trong bộ dữ liệu gồm thông tin về một lượt đặt phòng khách sạn
- Trong đó có 2 cột thiếu dữ liệu nhiều nhất với cột *company* hơn 94% dòng dữ liệu bị thiếu và cột *agent* là hơn 13%
- Số trường dữ liệu bị thiếu ở mỗi dòng: Trong đó 91% các dòng bị thiếu 1 trường dữ liệu, 8% các dòng bị thiếu 2 trường dữ liệu.
- Tỷ lệ dòng dữ liệu bị trùng lặp là 26.8%.
- Nhóm đã điền tất cả các giá trị thiếu trong bộ dữ liệu bằng giá trị '-1' để có thể dùng đến các cột có giá trị thiếu trong quá trình phân tích dữ liệu
- Sau khi điền giá trị thiếu, xóa những dòng dữ liệu trùng lặp, bộ dữ liệu còn lại gồm 87396 dòng và 32 cột

5.2 EDA 1D

- Dựa vào biểu đồ phân phối của trường *lead_time* ta thấy phần lớn người đặt phòng sẽ nhận phòng sớm trong vòng 10 ngày, trong đó hơn '6000' số lượng đặt phòng sẽ nhận phòng trong ngày (*lead_time* = 0)
- Ta cũng biết được năm 2016 là năm có số lượng đặt phòng lớn nhất trong giai đoạn giữa năm 2015 đến giữa năm 2017

- Số lượng đặt phòng ở City Hotel gấp 1.5 lần số lượng đặt Resort Hotel. Gần như toàn bộ khách hàng đều là khách mới, có một số lượng nhỏ là khách cũ và 82% khách hàng là dạng 'Transient' (Tạm thời)
- Trong tổng số lượng đặt phòng thì có 3/4 phòng đã check out và 1/3 còn lại đã 'canceled' và gần như toàn bộ số lượng đặt phòng đều là no deposit (Không đặt cọc)
- Từ biểu đồ số lượng đặt phòng ở các tháng ta thấy rằng số lượng đặt phòng vào tháng 7 và 8 là nhiều nhất. Theo thống kê thì đây là 2 tháng mà số lượng người đi du lịch là nhiều nhất nên số lượng đặt phòng cũng tăng theo.

5.3 EDA 2D

- **Những rút trích từ heatmap và scatter plot về tương quan giữa các biến numeric:**
  - *stays_in_week_nights* và *stays_in_weekend_nights* có tương quan dương lớn (0.56). Điều này cho thấy khách hàng có xu hướng lưu trú cả trong tuần lẫn cuối tuần. Điều này phản ánh khách hàng còn có thể đặt phòng dài hạn chứ không chỉ giới hạn ở việc lưu trú ngắn ngày.
  - *lead_time* và *stays_in_week_nights* có tương quan dương vừa phải (0.31), có thể cho thấy khách hàng thường đặt phòng trước càng lâu thì càng có xu hướng lưu trú nhiều đêm trong tuần hơn.
  - *adr* (Average Daily Rate) có một tương quan dương mạnh với *adults* (0.25) và *children* (0.33), nhưng lại có mức tương quan thấp với *babies* (0.04). Điều này cho thấy giá trung bình hàng ngày có liên quan đến số người lớn và trẻ em, nhưng không nhiều đối với số lượng em bé.
  - *arrival_date_year* và *lead_time* có tương quan dương yếu (0.14), có thể cho thấy rằng việc đặt phòng trước không có sự thay đổi đáng kể theo năm.
  - *agent* có tương quan nhỏ với hầu hết các biến khác, và một tương quan dương yếu với

booking_changes (0.13), có thể là do việc đặt phòng qua đại lý có những điều chỉnh nhất định.

- *required_car_parking_spaces* và *adr* có một tương quan lên tới (0.14), cho thấy có thể khách đặt phòng với giá cao hơn có xu hướng yêu cầu nhiều chỗ đậu xe hơn.
- *total_of_special_requests* có một tương quan dương yếu với *booking_changes* (0.10), điều này có thể đến từ việc những khách hàng có nhiều yêu cầu đặc biệt hơn cũng thường xuyên thực hiện thay đổi đối với đặt phòng của họ.
- *previous_bookings_not_canceled* và *previous_cancellations*: Có tương quan dương giữa 2 biến cho thấy khách hàng có lịch sử hủy bỏ đặt phòng có xu hướng tiếp tục hủy đặt phòng trong tương lai.

• **Bar chart giữa biến numerical và categorical:**

- Phân khúc *Corporate* có số lượng hủy đặt phòng cao nhất trong tất cả các phân khúc thị trường được nêu, với gần 800 hủy đặt phòng.
- Tỷ lệ hủy đặt phòng của khách hàng *Transient* cao, nhưng số lượng đặt phòng không hủy của loại khách hàng này cũng là cao nhất. Điều này có thể là do tổng số lượng đặt phòng của khách hàng *Transient* rất lớn, do đó cả số lượng hủy đặt và số lượng không hủy đặt đều cao.
- Quốc gia PRT (Bồ Đào Nha) dẫn đầu số lượng đặt phòng và cũng có tỷ lệ hủy phòng nhất.
- Trong top 5 quốc gia có số lượng đặt phòng lớn nhất gồm GBR (Anh Quốc), FRA (Pháp), ESP (Tây Ban Nha), DEU (Đức) và số lượng hủy đặt phòng trước đó của các nước này có tỷ lệ rất thấp.
- Số lượng đặt phòng của 2 loại *hotel* của các ngày trong tháng khá là đồng đều, ngoại trừ ngày 31. Có lẽ do tháng có ngày 31 ít hơn so với các ngày còn lại.
- Phần lớn số lượng đặt phòng không có bất cứ yêu cầu đặc biệt nào hoặc chỉ có 1 yêu cầu

đặc biệt.

- **Tỷ trọng giữa 2 biến categorical:**
  - Phần lớn số lượng đặt phòng đều không yêu cầu đặt cọc trước (No Deposit), và trong đó tỷ lệ *Check-Out* của loại này chiếm hơn 70%
  - Các loại còn lại chiếm tỷ trọng khá nhỏ trong đó có *Non Refund* với tỷ lệ Canceled lại khá cao khi chiếm gần như toàn bộ loại này.

5.4 EDA 3D

- **Phân loại đặt phòng theo đối tượng khách hàng:** Đa số đặt phòng là cho người lớn, chỉ có một số ít đặt phòng cho trẻ em hoặc em bé. Những đặt phòng cho trẻ em thường đi kèm với người lớn, và số lượng trẻ em và em bé thường không quá nhiều.
- **Phân bố thời gian đặt phòng qua các năm, tuần trong năm và ngày trong tháng:** Số lượng đặt phòng phân bố khá đều qua các năm, tuần trong năm và ngày trong tháng. Có xu hướng đặt phòng nhiều hơn vào giữa năm 2015 đến giữa năm 2017.
- **Thời gian lưu trú:** Đa số lượt đặt phòng lưu trú trong khoảng 1-15 ngày trong tuần và 1-4 ngày vào cuối tuần. Có xu hướng lưu trú nhiều hơn vào cuối tuần ở các tuần đầu và cuối năm.
- **Tỷ lệ hủy đặt phòng:** Số lượng đặt phòng không bị hủy nhiều hơn ở *City Hotel* so với *Resort Hotel*. Có một xu hướng tiêu cực giữa *Previous cancellations* và số lượng đặt phòng không bị hủy, nhưng không thể kết luận chắc chắn từ biểu đồ.
- **Lead time và số ngày nằm trong danh sách chờ:** Phần lớn đặt phòng không bị hủy có *lead time* và số ngày nằm trong danh sách chờ thấp, trong khi một số lượng đáng kể đặt phòng bị hủy có *lead time* và số ngày nằm trong danh sách chờ thấp.

Tài liệu tham khảo

[1] Dataset Hotel booking demand

https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand