

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**ĐẠI HỌC QUỐC GIA TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN – KHÓA 2021**

**---------------o0o---------------**



**BÁO CÁO ĐỒ ÁN QUAN HÓA DỮ LIỆU**

Hotel bookings



**Giảng viên:** Bùi Tiến Lên

Thành phố Hồ Chí Minh, tháng 4 năm 2024

# **MỤC LỤC**

[**MỤC LỤC** 1](#_Toc163334547)

[**THÔNG TIN CHUNG** 2](#_Toc163334548)

[**NỘI DUNG CHÍNH** 3](#_Toc163334549)

[**I. Data understanding** 3](#_Toc163334550)

[1. Đếm số dòng và số cột 3](#_Toc163334551)

[2. Viết bảng mô tả về các cột 3](#_Toc163334552)

[3. Lấy 5 điểm dữ liệu ra làm mẫu 4](#_Toc163334553)

[4. Phân tích tỉ lệ missing rate 4](#_Toc163334554)

[5. Phân tích tỉ lệ duplicate 5](#_Toc163334555)

[6. Fill in missing rate 6](#_Toc163334556)

[**II. EDA 1D** 6](#_Toc163334557)

[1. Chia loại dữ liệu num hoặc cate 6](#_Toc163334558)

[2. Phân tích tỷ lệ đó với biến cate 6](#_Toc163334559)

[3. Phân tích phân phối đối biến num 6](#_Toc163334560)

[**III. EDA 2D** 10](#_Toc163334561)

[1. Phân tích hệ số tương quan giữa các biến num 10](#_Toc163334562)

[2. Sử dụng Scatter plot để phân tích dữ liệu 2D 11](#_Toc163334563)

[3. Sử dụng bar chart để phân tích dữ liệu num và cate 14](#_Toc163334564)

[4. Tính tỷ trọng đối với hai biến cate 15](#_Toc163334565)

[**IV. EDA 3D** 16](#_Toc163334566)

[1. Sử dụng Scatter plot để phân tích dữ liệu 3D cho ba biến num 16](#_Toc163334567)

[2. Sử dụng Scatter plot 2D và màu đối với hai biến num và cate 17](#_Toc163334568)

[3. Tính tỷ trọng theo bin chia theo thể loại với hai biến cate 17](#_Toc163334569)

[**V. Insight** 18](#_Toc163334570)

# **THÔNG TIN CHUNG**

**I. Chi tiết các thành viên**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | MSSV | Họ Tên | Công việc | Tỉ lệ thực hiện |
| 1 | 21127243 | Phùng Siêu Đạt | I. Data understading  II. EDA 1D  III. EDA 2D | 100% |
| 2 | 21127296 | Đặng Hà Huy | III. EDA 2D  IV. EDA 3D  V. Insight | 100% |

**II. Chi tiết công việc và mức độ hoàn thành**

|  |  |  |
| --- | --- | --- |
| **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% |

# **NỘI DUNG CHÍNH**

## **I. Data understanding**

### **1. Đếm số dòng và số cột**

Bộ dữ liệu được chọn: hotel\_bookings.csv

- Tập dữ liệu bao gồm có 119390 dòng và 32 cột. Mỗi dòng bao gồm toàn bộ thông tin về một lượt đặt phòng khách sạn. Nội dung các cột được mô tả bảng bên dưới

### **2. Viết bảng mô tả về các cột**

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Mô tả** |
| 1 | hotel | Tên khách sạn (gồm có City Hotel và Resort Hotel) |
| 2 | is\_canceled | Có hủy hay không (1 là có hủy 0 là ngược lại) |
| 3 | lead\_time | Số ngày tính từ ngày đặt phòng đến ngày nhận phòng |
| 4 | arrival\_date\_year | Năm nhận phòng |
| 5 | arrival\_date\_month | Tháng nhận phòng |
| 6 | arrival\_date\_week\_number | Tuần trong năm nhận phòng (1 đến 53) |
| 7 | arrival\_date\_day\_of\_month | Ngày nhận phòng trong tháng (1 đến 31) |
| 8 | stays\_in\_weekend\_nights | Số đêm cuối tuần (thứ Bảy và Chủ Nhật) mà khách hàng đã đặt |
| 9 | stays\_in\_week\_nights | Số đêm trong tuần (thứ Hai đến thứ Sáu) mà khách hàng đã đặt |
| 10 | adults | Số người lớn |
| 11 | childrens | Số trẻ em |
| 12 | babies | Số em bé |
| 13 | meal | Loại bữa ăn khách hàng đã đặt được phân loại bằng các gói bữa ăn khách sạn tiêu chuẩn gồm:   * BB – bữa sáng * HB – bữa sáng và bữa tối * FB – bữa sáng, bữa trưa và bữa tối |
| 14 | country | Quốc tịch của khách hàng (dưới dạng ISO 3155–3:2013) |
| 15 | market\_segment | Phân khúc thị trường |
| 16 | distribution\_channel | Kênh phân phối đặt phòng |
| 17 | is\_repeated\_guest | Là khách quen (0 là khách hàng lần đầu 1 là khách hàng đã đến nhiều lần) |
| 18 | previous\_cancellations | Số lượng hủy phòng của khách hàng trước khi đặt phòng hiện tại |
| 19 | previous\_bookings\_not\_canceled | Số lượng phòng khách hàng đã đặt mà không hủy trước khi đặt phòng hiện tại |
| 20 | reserved\_room\_type | Mã loại phòng đã đặt trước |
| 21 | assigned\_room\_type | Mã loại phòng được dành cho lượt đặt |
| 22 | booking\_changes | Số lần thay đổi phòng chi tiết đặt phòng trước khi nhận phòng |
| 23 | deposit\_type | Loại hình thanh toán |
| 24 | agent | Mã ID của đại lý du lịch đã đặt chỗ |
| 25 | company | Mã ID của công ty du lịch đã đặt chỗ hoặc chịu trách nhiệm thanh toán |
| 26 | days\_in\_waiting\_list | Số ngày trong danh sách chờ của lượt đặt phòng trước khi được xác nhận |
| 27 | customer\_type | Loại khách hàng |
| 28 | adr | Giá trung bình hằng ngày được tính 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 mà khách hàng yêu cầu |
| 30 | total\_of\_special\_requests | Số lượng những yêu cầu đặt biệt của khách hàng |
| 31 | reservation\_status | Tình trạng của phòng được đặt |
| 32 | reservation\_status\_date | Ngày cuối cùng mà tình trạng của phòng được được cập nhật |

### **3. Lấy 5 điểm dữ liệu ra làm mẫu**

- Để lấy ra 5 điểm dữ liệu đầu tiên để làm mẫu ta sử dụng hàm head() trong thư viện pandas. Số lượng dòng dữ liệu mặc định đối với hàm head() là 5.

A screenshot of a computer

Description automatically generated

### **4. Phân tích tỉ lệ missing rate**

- Tỉ lệ missing rate đề cập đến tỷ lệ phần trăm của dữ liệu thiếu hoặc không có sẵn trong một tập dữ liệu cụ thể.

- Để tính tỉ lệ missing rate, ta cần biết số lượng các giá trị thiếu trong tập dữ liệu và tổng số lượng quan sát. Được tính bằng cách chia số lượng giá trị thiếu cho tổng số lượng quan sát.

- Tỉ lệ missing rate của dataset theo từng thuộc tính là

A screen shot of a computer program

Description automatically generated

- Trong đó cột company thiếu nhiều dữ liệu nhất với hơn 94.3% dữ liệu bị thiếu, thứ hai là cột agent với 13.6%

### **5. Phân tích tỉ lệ duplicate**

- Tỉ lệ missing rate dùng để đo lường mức độ trùng lặp của các quan sát trong tập dữ liệu.

- Để tính tỉ lệ duplicate, ta cần biết số lượng các quan sát trùng lặp và tổng số lượng quan sát trong tập dữ liệu. Tỉ lệ duplicate được tính bằng cách chia số lượng quan sát trùng lặp cho tổng số lượng quan sát.

- Tỉ lệ duplicate của dataset là **26.8%**

### **6. Fill in missing rate**

- Để fill in missing rate trong tập dữ liệu ta sử dụng hàm fillna() của thư viện pandas với giá trị là -1 để có thể điền vào những ô dữ liệu trống giá trị -1

- Sau khi điền vào các ô bị thiếu và xóa những dòng dữ liệu bị trùng lặp thì tập dữ liệu còn lại 87396 dòng và 32 cột

## **II. EDA 1D**

### **1. Chia loại dữ liệu num hoặc cate**

- Num: bao gồm các cột dữ liệu có giá trị là kiểu dữ liệu số. Tập dữ liệu có 20 cột num



- Cate: bao gồm các cột dữ liệu có giá trị là kiểu dữ liệu object. Tập dữ liệu có 12 cột cate



- Bên cạnh đó, ta cũng tính số lần xuất hiện của mỗi giá trị trong từng cột của tập num\_list và chuẩn hóa kết quả theo tỉ lệ phần trăm.



### **2. Phân tích tỷ lệ đó với biến cate**

- Ta cũng làm điều tương tự đối với tập dữ liệu cate\_list tính số lần xuất hiện của mỗi giá trị trong từng cột và chuẩn hóa kết quả theo tỉ lệ phần trăm để có thể so sánh với tập num\_list



### **3. Phân tích phân phối đối biến num**

- Cột num đầu tiên mang ý nghĩa quyết định đến tập dữ liệu

A graph with a blue bar

Description automatically generated

- 6 cột dữ liệu kế tiếp của tập num\_list có độ đa dạng dữ liệu cao

A group of blue and white graphs

Description automatically generated with medium confidence

- 6 cột dữ liệu kế tiếp thì chỉ có dữ liệu ở một số giá trị nhất định

A screenshot of a graph

Description automatically generated

- Những cột dữ liệu còn lại thì một số có độ đa dạng dữ liệu cao, một số thì chỉ có dữ liệu ở một số giá trị nhất định

A group of white rectangular objects with blue and black lines

Description automatically generated with medium confidence

## **III. EDA 2D**

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

A chart with numbers and a red and blue gradient

Description automatically generated with medium confidence

Dựa vào heatmap trên ta có thể thấy được một vài điểm quan trọng như:

- **stay\_in\_week\_nights** có độ tương quan lớn với **stay\_in\_weekend\_nights** có vẻ như nếu khách hàng đã lựa chọn ở qua đêm tại khách sạn vào những đêm trong tuần thì họ cũng sẽ có xu hướng ở qua đêm tại khách sạn vào những đêm cuối tuần

- **adr** và **children** có độ tương quan lớn với nhau nên ta có thể suy đoán rằng số lượng trẻ em sẽ ảnh hưởng đến chi tiêu của khách hàng

- **previous\_bookings\_not\_cancel** có độ tương quan cao với **is\_repeated\_guest** cho ta thấy rằng những khách hàng quen thuộc thì có xu hướng hủy những lịc trình khác hơn so với những khách hàng lần đầu

- **arrival\_date\_week\_number** và **arrival\_day\_year** có độ tương quan thấp nhất thể hiện rằng năm nhận phòng gần như không hề có ảnh hưởng gì đến với quyết định khách hàng nhận phòng vào tuần nào trong năm đó

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

Tương quan giữa previous\_cancellations và previous\_bookings\_not\_canceled

A graph of blue dots

Description automatically generated

- Previoius bookings not canceled và Previous cancellations có tương quan thuận với nhau

- Một 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.

Tương quan giữa lead\_times và booking\_changesA graph with blue dots

Description automatically generated

- Có thể thấy Lead time và booking changes có tỉ lệ nghịch với nhau

- Có thể thấy lead time càng cao thì việc khách hàng thay đổi chi tiết đặt phòng càng giảm

Tương quan giữa stays\_in\_week\_nights và stays\_in\_weekend\_nights

A graph with blue dots

Description automatically generated

- Stays in weekend nights và Stays in week nights tỉ lệ thuận với nhau

- Khách hàng có xu hướng ở dài hạn từ giữa tuần đến cuối tuần hơn là ở ngắn hạn

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

A graph with blue squares

Description automatically generated

- Có thể thấy kênh phân phối đặt phòng đến từ các Đại lý du lịch hay các nhà điều hành tour chiếm đa số tỉ lệ đặt phòng

A graph with blue rectangles

Description automatically generated

- Có thể thấy PRT (Bồ Đào Nha) có lượng khách lặp lại nhiều nhất nên ta có thể suy đoán rằng cả 2 khách sạn đang khảo sát đều nằm ở Bồ Đào Nha

Phân bố loại khách hàng giữa 2 khách sạn City Hotel và Resort Hotel

A graph of different colored bars

Description automatically generated

- Loại khách hàng chiến tỉ lệ đa số là khách hàng riêng lẻ (như khách du lịch) tiếp đến là loại khách hàng đi theo nhóm (Transient-Party)

### **4. Tính tỷ trọng đối với hai biến cate**

A screenshot of a hotel

Description automatically generatedA screenshot of a reservation system

Description automatically generated

## **IV. EDA 3D**

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

- Phân tích dữ liệu giữa 3 biến num is\_repeated\_guest, previous\_booking\_not\_canceled và previous\_cancellations

A graph of a graph

Description automatically generated with medium confidence

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

A graph of a hotel

Description automatically generated

- Ta có thể thấy những khách hàng của City Hotel có số lượng đặt phòng trước đó mà không bị hủy cao hơn rất nhiều so với Resort Hotel. Còn Resort Hotel có lượng hủy phòng trước đó cao cho thấy rằng có lẽ như chất lượng của Resort Hotel thấp hơn so với City Hotel

### **3. Tính tỷ trọng theo bin chia theo thể loại với hai biến cate**

A screenshot of a phone

Description automatically generated

## **V. Insight**

Sau khi phân tích tập dữ liệu ta có thể rút ra được một số ý chính sau:

* Hầu hết khách hàng không hủy đặt phòng khi họ không có lịch sử hủy bỏ trước đó.
* Thời gian từ lúc đặt đến lúc nhận phòng càng cao thì khách hàng thường không có xu hướng thay đổi chi tiết đặt phòng của mình
* Khi đặt phòng, đa số khách hàng thường sẽ đặt thời gian ở tại khách sạn qua đêm dài hạn kéo dài từ trong tuần đến tận cuối tuần mà không chọn ngắn hạn
* Hầu hết khách hàng thông qua các kênh phân phối đại lý du lịch hoặc các nhà điều hành tour để có thể đặt phòng cho mình thay cho các lựa chọn khác
* Đa số các khách hàng trong tập dữ liệu trên đều có quốc tịch Bồ Đào Nha, từ đó ta có thể suy đoán rằng cả 2 khách sạn mà ta khảo sát có thể ở Bồ Đào Nha
* Loại khách hàng chiếm tỉ lệ đa số là loại khách hàng Transient (khách hàng riêng lẻ) có lẽ như đa số là khách du lịch và khách hàng nội địa
* Số lượt hủy phòng trược đó của City Hotel thấp hơn so với Resort Hotel. Ta có thể suy đoán rằng do chất lượng của City Hotel cao hơn so với Resort Hotel.