

**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](#_Toc163244052)

[**THÔNG TIN NHÓM** 2](#_Toc163244053)

[**NỘI DUNG CHÍNH** 3](#_Toc163244054)

[**I. Data understanding** 3](#_Toc163244055)

[1. Đếm số dòng và số cột 3](#_Toc163244056)

[2. Viết bảng mô tả về các cột 3](#_Toc163244057)

[3. Lấy 5 điểm dữ liệu ra làm mẫu 4](#_Toc163244058)

[4. Phân tích tỉ lệ missing rate 4](#_Toc163244059)

[5. Phân tích tỉ lệ duplicate 4](#_Toc163244060)

[6. Fill in missing rate 4](#_Toc163244061)

[**II. EDA 1D** 4](#_Toc163244062)

[1. Fill in missing rate 4](#_Toc163244063)

[**III. EDA 2D** 4](#_Toc163244064)

[**IV. EDA 3D** 4](#_Toc163244065)

# **THÔNG TIN NHÓM**

|  |  |  |  |
| --- | --- | --- | --- |
| STT | MSSV | Họ Tên | Tỉ lệ thực hiện |
| 1 | 21127243 | Phùng Siêu Đạt | 50% |
| 2 | 21127296 | Đặng Hà Huy | 50% |

# **NỘI DUNG CHÍNH**

## **I. Data understanding**

### **1. Đếm số dòng và số cột**

- Bộ dữ liệu được chọn: hotel\_bookings.csv

- Dataset gồm có 119390 dòng và 32 cột

### **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 |
| 14 | country | Quốc tịch của khách hàng (được biểu diễn dưới dạng ISO 3155–3:2013 format) |
| 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 khách hàng |
| 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 mà tình trạng của phòng được đặt bị sửa đổi lần cuối cùng |

### **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 được lấy ra 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

### **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 **0.7320211072954184**

### **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

## **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 tổng cộng 20 cột num

A computer screen shot of a computer code

Description automatically generated

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

A computer code with text

Description automatically generated with medium confidence

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

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

A graph with a blue bar

Description automatically generated

A group of blue and white graphs

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated

## **III. EDA 2D**

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

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

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

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

## **IV. EDA 3D**

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

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

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

## **V. Insight**