**BÁO CÁO DỰ ÁN**

**PHÂN TÍCH DỮ LIỆU ĐẶT PHÒNG KHÁCH SẠN**

**Người thực hiện:**

1. Nguyễn Văn Huỳnh

MỤC LỤC

**[I. GIỚI THIỆU DỰ ÁN](#_heading=h.1fob9te)**

[1. MÔ TẢ DỮ LIỆU](#_heading=h.3znysh7)3

[2. YÊU CẦU CỦA CÔNG TY](#_heading=h.tyjcwt)3

**[II.](#_heading=h.2s8eyo1)****PHÂN TÍCH KHÁCH HÀNG**

[1. PHÂN TÍCH YÊU CẦU](#_heading=h.17dp8vu) 4

[2. CÂU CHUYỆN DỮ LIỆU](#_heading=h.3rdcrjn) 4

3[. XÁC ĐỊNH ĐỐI TƯỢNG](#_heading=h.2jxsxqh) 6

**[III. KIẾN TRÚC HỆ THỐNG](#_heading=h.1y810tw)**

[1. KIẾN TRÚC](#_heading=h.4i7ojhp) 6

[2. GIẢI THÍCH](#_heading=h.2xcytpi) 6

[3. GIẢI THÍCH BỘ DỮ LIỆU](#_heading=h.1ci93xb) 7

**[IV. LÀM SẠCH VÀ CHUYỂN ĐỔI DỮ LIỆU](#_heading=h.2bn6wsx)**

[1. CHUẨN BỊ DỮ LIỆU](#_heading=h.qsh70q) 8

[2. LÀM SẠCH DỮ LIỆU](#_heading=h.ihv636) 8

[3. CHUYỂN ĐỔI DỮ LIỆU](#_heading=h.vx1227) 9

**[V. XỬ LÝ DỮ LIỆU](#_heading=h.2u6wntf)**

[1. CHUẨN HÓA DỮ LIỆU](#_heading=h.19c6y18) 9

[2. MÔ HÌNH HÓA DỮ LIỆU](#_heading=h.28h4qwu) 10

[3. XỬ LÝ DỮ LIỆU DAX](#_heading=h.1mrcu09) 11

**[VI. TRỰC QUAN DỮ LIỆU](#_heading=h.3ygebqi)**

[1. CÁC KỸ THUẬT TRỰC QUAN HÓA](#_heading=h.2dlolyb) 12

[2. CÁC NGUYÊN TẮC TRỰC QUAN HÓA](#_heading=h.34g0dwd) 14

[3. TRÌNH BÀY TẠO CÁC VISUAL CHO DỰ ÁN](#_heading=h.1jlao46)14

**[VIII. KẾT LUẬN](#_heading=h.40ew0vw)**

[1. CÁC BƯỚC VIẾT BÁO CÁO](#_heading=h.2fk6b3p) 18

[2. KHÓ KHĂN](#_heading=h.upglbi) 20

[4. THUẬN LỢI](#_heading=h.1tuee74) 20

[5. TỔNG KẾT](#_heading=h.meukdy) 20

# **I. GIỚI THIỆU DỰ ÁN**

## 1. MÔ TẢ DỮ LIỆU

Bạn được thuê để phân tích dữ liệu đặt phòng trong 2 năm qua cho một chuỗi khách sạn cao cấp. Hiện tại, khách sạn đang gặp một số vấn đề và cần bạn tìm ra giải pháp để tăng doanh thu, giảm tỷ lệ hủy phòng và nâng cao trải nghiệm khách hàng.

## 2. YÊU CẦU CỦA CÔNG TY

* Về mặt dữ liệu: Cần chính xác và minh bạch đã được qua kiểm chứng.
* Quản lý và lưu trữ: Power BI Service, SQL Server, Google Drive.
* Mục tiêu: Chỉ ra các thông số từ nguồn dữ liệu thu thập được, trực quan hoá và đưa ra báo cáo dễ đọc dễ hình dung.
* Các công cụ và ngôn ngữ được sử dụng: Python, SQL Server, Power BI, …

# 

# **II. PHÂN TÍCH KHÁCH HÀNG**

### 1. PHÂN TÍCH YÊU CẦU

- Dữ liệu: Cuộc thi thực chiến dữ liệu cùng dứa  
- Quản lý và lưu trữ: SQL Server

### 2. CÂU CHUYỆN DỮ LIỆU

### Chuỗi khách sạn **Luxury Stay** đang gặp khó khăn. Doanh thu không ổn định, tỷ lệ hủy phòng tăng, và khách hàng chưa hài lòng với dịch vụ. Ban lãnh đạo cần một giải pháp dựa trên dữ liệu để cải thiện tình hình.

### Bạn, một chuyên gia phân tích dữ liệu, bắt tay vào khám phá hệ thống đặt phòng và dịch vụ của khách sạn. Dữ liệu bao gồm các thông tin quan trọng như:

### **Đặt phòng**: Ai đã đặt, ngày check-in/check-out, tình trạng phòng.

### **Thanh toán**: Số tiền, phương thức thanh toán, ngày thanh toán.

### **Phòng**: Loại phòng, giá theo đêm, tình trạng phòng.

### **Khách hàng**: Thông tin liên hệ, lịch sử đặt phòng.

### **Dịch vụ**: Các dịch vụ sử dụng trong kỳ lưu trú.

##### 3.1. MÔ TẢ THỰC TRẠNG:

##### Chuỗi khách sạn cao cấp đang gặp một số vấn đề ảnh hưởng đến hiệu suất kinh doanh. Cụ thể:

##### Doanh thu không ổn định, đặc biệt trong mùa thấp điểm.

##### Tỷ lệ hủy phòng cao, gây thất thoát doanh thu và ảnh hưởng đến công suất phòng.

##### Dịch vụ khách sạn chưa tối ưu, nhiều khách hàng không sử dụng hết các dịch vụ bổ sung.

##### Trải nghiệm khách hàng chưa đạt kỳ vọng, một số phản hồi tiêu cực về chất lượng phòng, quy trình đặt phòng và thanh toán.

##### Những vấn đề này đặt ra yêu cầu cấp thiết phải phân tích dữ liệu để tìm ra nguyên nhân và đề xuất giải pháp hiệu quả.

##### 3.2. DỮ LIỆU LIÊN QUAN:

##### Để phân tích và đánh giá thực trạng, các dữ liệu sau được sử dụng:

##### Dữ liệu đặt phòng: thông tin khách hàng, ngày check-in/check-out, trạng thái đặt phòng.

##### Dữ liệu thanh toán: số tiền thanh toán, phương thức thanh toán, ngày giao dịch.

##### Dữ liệu phòng: loại phòng, giá phòng theo đêm, tình trạng phòng.

##### Dữ liệu khách hàng: thông tin cá nhân, lịch sử đặt phòng.

##### Dữ liệu dịch vụ: danh sách dịch vụ sử dụng, giá dịch vụ, số lượng sử dụng.

##### Dữ liệu này sẽ giúp phân tích xu hướng đặt phòng, tỷ lệ hủy, thói quen chi tiêu của khách hàng và hiệu suất kinh doanh của khách sạn.

##### 3.3. MỤC TIÊU:

Mục tiêu của phân tích này là:

* + Tối ưu doanh thu bằng cách xác định các yếu tố ảnh hưởng đến giá phòng và doanh thu dịch vụ.
  + Giảm tỷ lệ hủy phòng, tìm hiểu nguyên nhân và đưa ra giải pháp giảm thiểu tình trạng này nâng cao trải nghiệm khách hàng.
  + Xây dựng chiến lược kinh doanh hiệu quả, giúp khách sạn khai thác tối đa tiềm năng từ dữ liệu hiện có và nâng cao năng lực cạnh tranh trên thị trường.
  + Đề xuất giá phòng theo thời gian để tối ưu hoá doanh thu
  + Phân loại khách hàng và dự đoán khách hàng rời bỏ
  + Phát hiện các gian lận trong quá trình đặt phòng và thanh toán

### 

### 

### **4. XÁC ĐỊNH ĐỐI TƯỢNG**

### Trong phân tích này, các đối tượng chính được xác định bao gồm:

### Khách hàng

### Ban quản lý khách sạn

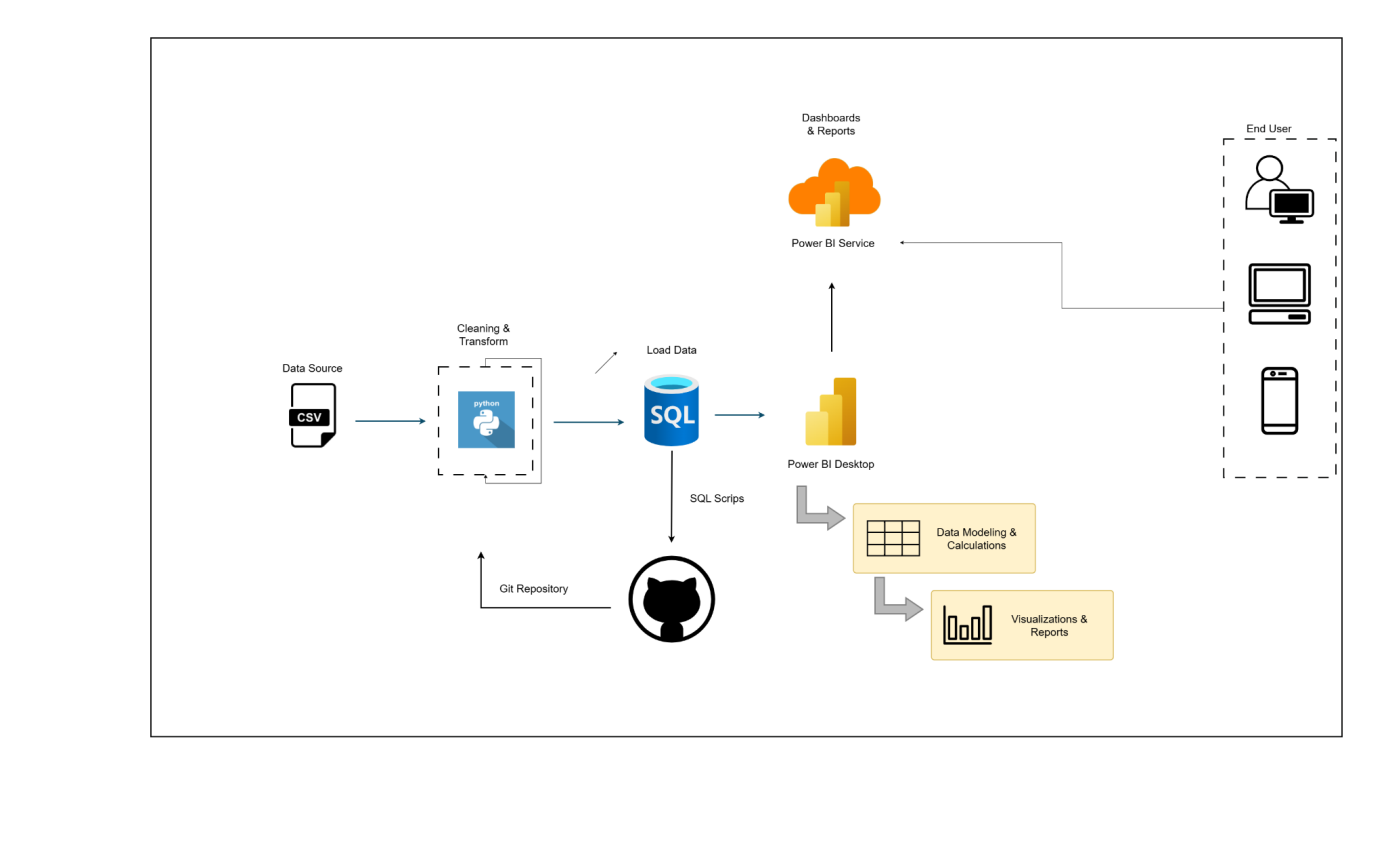
### Bộ phận tài chính & kinh doanh

### Quản lý dòng tiền, doanh thu, chi phí vận hành.

### Bộ phận dịch vụ khách hàng

# **III. KIẾN TRÚC HỆ THỐNG**

### 1. KIẾN TRÚC



*Hình ảnh 1.1*

### 2. GIẢI THÍCH

1. Từ file CSV làm sạch, chuẩn hoá bằng Python

2. Lưu trữ, truy vấn dữ liệu trên SQL Server

3. Trực quan bằng Power BI

4. Đưa ra các dự báo bằng các mô hình học máy

### 3. GIẢI THÍCH BỘ DỮ LIỆU

### 3.1. CÁC TRƯỜNG DỮ LIỆU

1. Bảng Customer

* customer\_id (PK) – Mã khách hàng (định danh duy nhất).
* full\_name – Họ và tên khách hàng.
* email – Địa chỉ email dùng để liên hệ.
* phone – Số điện thoại liên lạc.
* created\_at – Ngày đăng ký khách hàng.

2.Bảng Room (Phòng khách sạn)

* room\_id (PK) – Mã định danh của phòng.
* room\_number – Số phòng cụ thể.
* room\_type – Loại phòng (Deluxe, Suite, Standard...).
* price\_per\_night – Giá thuê phòng theo đêm.
* status – Trạng thái phòng (trống, đã đặt, đang dọn dẹp...).

3. Bảng Booking (Đặt phòng)

* booking\_id (PK) – Mã định danh của đơn đặt phòng.
* customer\_id (FK) – Mã khách hàng (liên kết với bảng Customer).
* room\_id (FK) – Mã phòng đặt (liên kết với bảng Room).
* check\_in – Ngày nhận phòng.
* check\_out – Ngày trả phòng.
* status – Trạng thái đặt phòng (đã đặt, hủy, hoàn thành...).
* created\_at – Ngày đặt phòng.

4. Bảng Payments (Thanh toán)

* payment\_id (PK) – Mã định danh thanh toán.
* booking\_id (FK) – Liên kết với đơn đặt phòng.
* amount – Số tiền thanh toán.
* payment\_method – Phương thức thanh toán (thẻ, tiền mặt, chuyển khoản...).
* payment\_date – Ngày thanh toán.

5. Bảng Service (Dịch vụ khách sạn)

* service\_id (PK) – Mã định danh dịch vụ.
* service\_name – Tên dịch vụ (spa, buffet, giặt là, đưa đón...).
* price – Giá dịch vụ.

6. Bảng Services Usage (Sử dụng dịch vụ)

* usage\_id (PK) – Mã định danh sử dụng dịch vụ.
* booking\_id (FK) – Đơn đặt phòng liên quan.
* service\_id (FK) – Dịch vụ đã sử dụng.
* quantity – Số lượng sử dụng.
* total\_price – Tổng tiền dịch vụ (số lượng \* đơn giá).

# **IV. LÀM SẠCH VÀ CHUYỂN ĐỔI DỮ LIỆU**

## 1. CHUẨN BỊ DỮ LIỆU

### 1.1 GIẢI PHÁP LƯU TRỮ DỮ LIỆU

* Tiêu chí so sánh: Chi phí và đầu tư ban đầu, Quản lý và bảo trì, Khả năng mở rộng và linh hoạt, Bảo mật và tuân thủ quy định, Hiệu suất và khả năng truy cập
* Quyết định : Sử dụng ứng dụng lưu trữ tại chỗ (on-premise) như SQL Server.

## 2. LÀM SẠCH DỮ LIỆU

### 2.1. CÁC VẤN ĐỀ ẢNH HƯỞNG DỮ LIỆU

* Thiếu dữ liệu tại các cột
* Kiểu dữ liệu không đồng nhất
* Các dữ liệu ngoại lệ và các giá trị bất thường
* Các dữ liệu trùng lặp
* Độ dài và kích thước của dữ liệu

### 2.2. CÁC TIÊU CHÍ ĐÁNH GIÁ CHẤT LƯỢNG DỮ LIỆU

* Độ chính xác của dữ liệu
* Dữ liệu phải đầy đủ không được thiếu và có tính nhất quán
* Dữ liệu phải tuân thủ các quy tắc và giới hạn đã được định sẵn
* Các bản ghi trong dữ liệu phải là duy nhất không được trùng lặp
* Độ an toàn và khả năng truy xuất của dữ liệu
* Mức độ truyền đạt của dữ liệu

### 2.3. CÁC BƯỚC LÀM SẠCH DỮ LIỆU

#### 2.3.1. TRÌNH BÀY CÁC BƯỚC LÀM SẠCH

* + Xử lý missing values
  + Xử lý dữ liệu trùng lặp
  + Gộp các dữ liệu để tạo data mới
  + Chuẩn hóa dữ liệu
  + Xử lý các ngoại lệ
  + Xử lý các dữ liệu được phân loại
  + Tạo hoặc tính toán các trường dữ liệu mới
  + Kiểm tra tính chính xác của dữ liệu

#### 2.3.2. TRÌNH BÀY CÁC BƯỚC LÀM SẠCH TRONG PHẠM VI DỰ ÁN

* + Xóa các cột không cần thiết
  + Replace các dữ liệu trống
  + Xóa các giá trị lỗi
  + Xóa các ký tự đặc biệt và bỏ các từ thừa

## 3. CHUYỂN ĐỔI DỮ LIỆU

### 3.1. CÁC TRƯỜNG HỢP CẦN CHUYỂN ĐỔI

* + Dữ liệu bị thiếu và không nhất quán
  + Dữ liệu thời gian cần phân tích theo chu kì, xu hướng
  + Dữ liệu quá lớn
  + Khi làm sạch các dữ liệu văn bản
  + Khi cần gộp hoặc tách các trường dữ liệu
  + Khi cần tính toán các trường dữ liệu mới

### 3.2. CÁC KỸ THUẬT CHUYỂN ĐỔI

* Loại bỏ giá trị thiếu
* Chuẩn hóa dữ liệu
* Thay đổi kiểu dữ liệu
* Phân nhóm dữ liệu
* Chuẩn hóa văn bản
* Ghép hoặc chia cột dữ liệu
* Tính toán các trường mới

**V. XỬ LÝ DỮ LIỆU**

## 1. CHUẨN HÓA DỮ LIỆU

### 1.1. TRÌNH BÀY CÁC BƯỚC CHUẨN HÓA

* Xử lý dữ liệu trống
* Xóa dữ liệu trùng lặp, không dùng đến
* Định dạng lại dữ liệu

## 

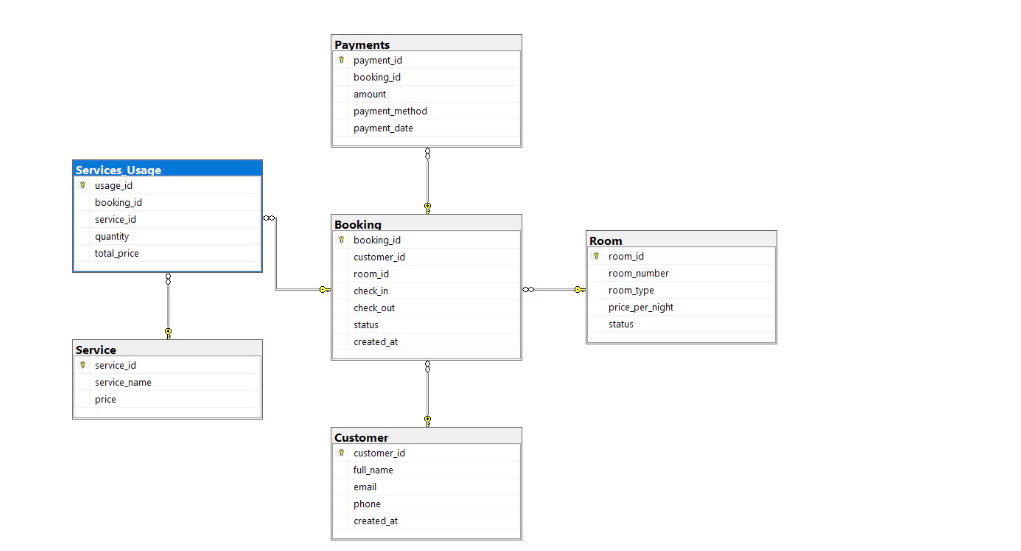
## 

## 

## 2. MÔ HÌNH HÓA DỮ LIỆU

### 2.1. CÁC LOẠI MÔ HÌNH DỮ LIỆU

1. Mô hình Star Schema
2. Mô hình Snowflake Schema
3. Mô hình Galaxy Schema
4. **Mô hình được dùng trong dự án**

* Mô hình Snowflake Schema****

*Hình 2.1*

### 2.2. CÁC TIÊU CHÍ ĐÁNH GIÁ MÔ HÌNH DỮ LIỆU

* Tính chính xác của mô hình
* Tính đầy đủ của dữ liệu
* Tính nhất quán, khả năng đo lường và tính khách quan
* Khả năng truyền đạt của mô hình, mô hình có tính cập nhật

## 

## 

## 3. XỬ LÝ DỮ LIỆU DAX

### 3.1. MEASURE

* Các measure được sử dụng
  + Booking By Season – Tổng số đặt phòng theo từng mùa.
  + Booking\_Per\_CustomerGroup – Số lượng đặt phòng theo nhóm khách hàng.
  + Cancelled Rate – Tỷ lệ đơn đặt phòng bị hủy.
  + Check\_RelatedTable – Kiểm tra mối quan hệ giữa các bảng dữ liệu.
  + Correlation\_Price\_Booking – Mức độ tương quan giữa giá phòng và số lượng đặt phòng.
  + Rate Cancelled – Tỷ lệ hủy đặt phòng (có thể theo từng danh mục cụ thể).
  + Rate Return – Tỷ lệ khách hàng quay lại đặt phòng.
  + Rate Room – Mức giá trung bình của phòng hoặc tỷ lệ phòng được đặt.
  + Return Customer – Số lượng khách hàng quay lại đặt phòng.
  + Revenue Per Customer – Doanh thu trung bình trên mỗi khách hàng.
  + Total Amount – Tổng doanh thu từ đặt phòng.
  + Total Customer – Tổng số khách hàng có đặt phòng.
  + Total Order – Tổng số đơn đặt phòng.
  + Total Usage Per Services – Tổng số lần sử dụng các dịch vụ bổ sung.
  + Total\_Payment – Tổng số tiền khách hàng đã thanh toán

##### 3.1.1. DASHBOARD OVERVIEW

* Total Amount – Tổng doanh thu từ đặt phòng.
* Total Order – Tổng số đơn đặt phòng.
* Total Usage per Services – Tổng số lần sử dụng các dịch vụ bổ sung.
* Revenue per Customer – Doanh thu của những khách hàng
* Rate Cancelled – Tỷ lệ hủy đặt phòng (có thể theo từng danh mục cụ thể).

##### 3.1.2. DASHBOARD DETAIL

* Rate Return – Tỷ lệ khách hàng quay lại đặt phòng.
* Correlation\_Price\_Booking – Mức độ tương quan giữa giá phòng và số lượng đặt phòng.
* Total Customer – Tổng số khách hàng có đặt phòng.
* Total\_Payment – Tổng số tiền khách hàng đã thanh toán

### 3.2. FILTER

* By year: theo năm
* By customer type: theo loại khách hàng

# **VI. TRỰC QUAN DỮ LIỆU**

## 1. CÁC KỸ THUẬT TRỰC QUAN HÓA

### 1.1. Biểu đồ cơ bản:

* **Bar Chart (Biểu đồ cột):**
  + Dùng để so sánh dữ liệu giữa các danh mục (categories).
  + Ví dụ: So sánh doanh thu theo từng sản phẩm hoặc quốc gia.
* **Column Chart (Biểu đồ cột dọc):**
  + Hiển thị các giá trị qua danh mục, tốt cho dữ liệu thời gian hoặc dữ liệu phân nhóm.
* **Line Chart (Biểu đồ đường):**
  + Dùng để phân tích xu hướng (trend) theo thời gian.
  + Ví dụ: Số lượng khách hàng mỗi tháng.
* **Pie Chart (Biểu đồ tròn):**
  + Biểu thị tỷ lệ phần trăm của các phần trong tổng thể.
  + Ví dụ: Tỷ lệ doanh thu đóng góp của từng sản phẩm.

1.2. Biểu đồ nâng cao:

* **Stacked Bar/Column Chart (Biểu đồ cột xếp chồng):**
  + Hiển thị tổng giá trị của các danh mục, đồng thời phân chia thành các thành phần con.
  + Ví dụ: Doanh thu từng năm, chia theo từng khu vực.
* **Area Chart (Biểu đồ vùng):**
  + Giống biểu đồ đường nhưng làm nổi bật phần diện tích dưới đường.
  + Dùng để biểu diễn xu hướng với sự nhấn mạnh vào tổng giá trị.
* **Scatter Plot (Biểu đồ phân tán):**
  + Hiển thị mối quan hệ giữa hai biến số.
  + Ví dụ: Mối quan hệ giữa chi tiêu quảng cáo và doanh thu.
* **Bubble Chart (Biểu đồ bong bóng):**
  + Giống Scatter Plot, nhưng thêm một biến nữa được biểu diễn qua kích thước bong bóng.

### 

### 1.3. Kỹ thuật hiển thị phân cấp:

* **Treemap (Biểu đồ cây):**
  + Biểu diễn dữ liệu phân cấp bằng các hình chữ nhật với kích thước tương ứng giá trị.
  + Ví dụ: Tỷ lệ đóng góp doanh thu theo danh mục sản phẩm.
* **Sunburst Chart (Biểu đồ ánh dương):**
  + Biểu diễn dữ liệu phân cấp bằng các vòng tròn đồng tâm.

### 1.4. Bản đồ :

* **Map (Bản đồ cơ bản):**
  + Hiển thị dữ liệu trên bản đồ địa lý.
  + Ví dụ: Doanh số theo khu vực hoặc quốc gia.
* **Filled Map (Bản đồ vùng):**
  + Làm nổi bật các khu vực theo giá trị (VD: dân số, doanh thu).

### 1.5. Bảng và KPI:

* **Table (Bảng dữ liệu):**
  + Hiển thị thông tin chi tiết với các hàng và cột.
  + Thích hợp để cung cấp dữ liệu ở dạng thô.
* **Matrix (Ma trận):**
  + Giống bảng, nhưng hỗ trợ phân cấp dữ liệu, giúp người dùng drill down chi tiết hơn.
* **KPI (Chỉ số hiệu suất chính):**
  + Biểu thị một giá trị đo lường cụ thể với ngữ cảnh (VD: đạt/không đạt mục tiêu doanh thu).

1.6. Biểu đồ tương tác và lọc:

* **Slicers (Bộ lọc):**
  + Dùng để lọc dữ liệu trực tiếp trên báo cáo, giúp người dùng tập trung vào một phần dữ liệu cụ thể.
* **Chỉ báo trạng thái (Gauge/Indicator):**
  + Thể hiện mức độ hoàn thành so với mục tiêu.

### 1.7. Biểu đồ dòng thời gian :

* **Gantt Chart:**
  + Dùng để quản lý dự án và biểu diễn thời gian thực hiện của từng nhiệm vụ.

**2. CÁC NGUYÊN TẮC TRỰC QUAN HÓA**

1. **Sự phù hợp và rõ ràng**

* Cần đảm bảo các dashboard và report phù hợp với đối tượng được truyền tải
* Các thông tin đưa ra cần được trình bày một cách chính xác, rõ ràng

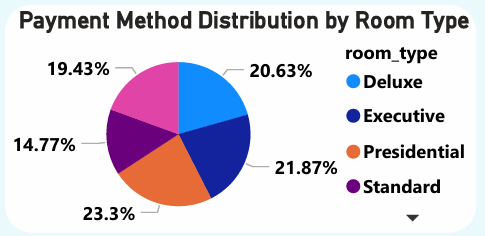
1. **Đơn giản, có tính thiết kế**

* Nên sử dụng các thiết kế đơn giản cho các chi tiết trên dashboard và cần chú ý vào cách chia bố cục, cách trình bày sao cho hợp lý
* Không đưa các dữ liệu vào một cách ngẫu nhiên mà các dữ liệu đó phải có mục đích cụ thể.

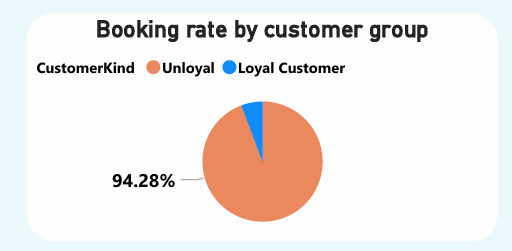
1. **Có thông điệp kể chuyện trên dữ liệu**

* Các dashboard/report đưa ra cần phải có thông điệp cụ thể, hỗ trợ đưa ra quyết định dựa trên số liệu, tránh mang các tính chất liệt kê
* Tránh việc làm dashboard nhưng chỉ đơn thuần là ghép các biểu đồ lại với nhau trên trang tổng thể

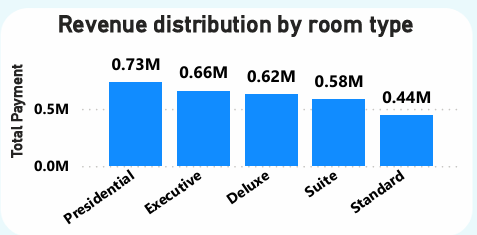
## 3. TRÌNH BÀY TẠO CÁC VISUAL CHO DỰ ÁN



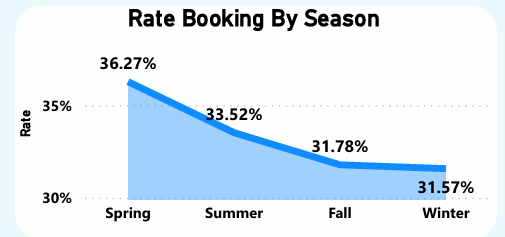
*Hình ảnh 3.1: Biểu đồ tỉ lệ phương thức thanh toán theo loại phòng*



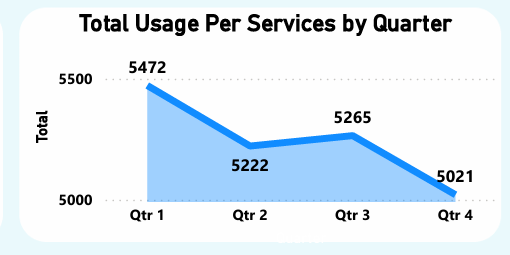
*Hình ảnh 3.2: Biểu đồ tỉ lệ đặt phòng theo nhóm khách hàng*

**

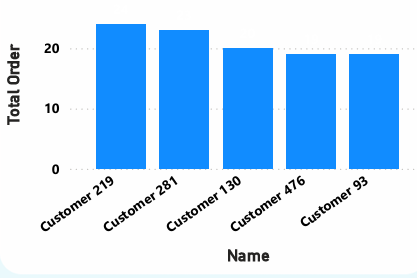
*Hình ảnh 3.3: Biểu đồ doanh thu theo loại phòng*

**

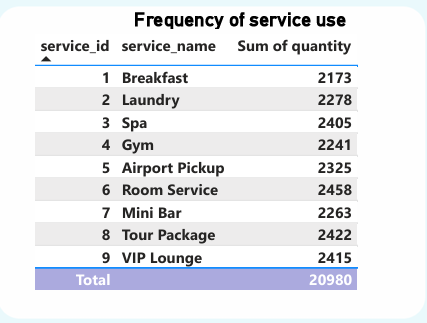
*Hình ảnh 3.4: Biểu đồ tỉ lệ đặt phòng theo mùa*

**

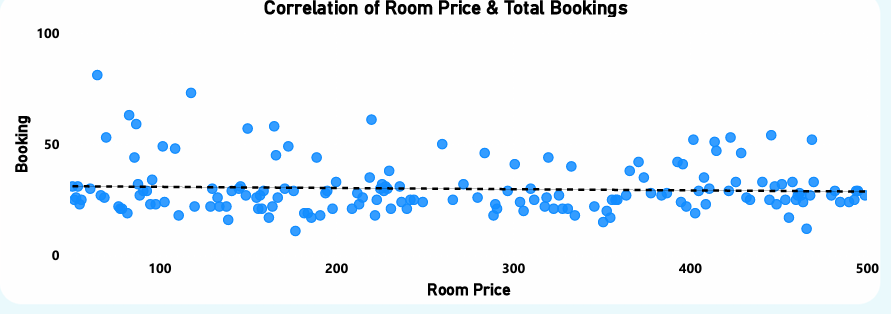
*Hình ảnh 3.5: Biểu đồ sử dụng dịch vụ theo quý*

**

*Hình ảnh 3.6: Biểu đồ top 5 khách hàng đặt phòng nhiều nhất*

**

*Hình ảnh 3.7: Biểu đồ tần suất sử dụng dịch vụ*

**

*Hình ảnh 3.8: Biểu đồ mối quan hệ giữa giá phòng và số lượng đặt phòng*

# 

# **VII. KẾT LUẬN**

## 1. CÁC BƯỚC VIẾT BÁO CÁO

### **1.1. Giới thiệu đề tài và câu chuyện dữ liệu**

* Trình bày bối cảnh và mục tiêu của dự án.
* Giải thích ý nghĩa của dữ liệu khách sạn và các chỉ số quan trọng.

### **1.2. Thiết kế khung báo cáo**

* Xác định bố cục báo cáo, các nội dung chính cần trình bày.
* Lựa chọn cách thức hiển thị thông tin để dễ hiểu và trực quan.

### **1.3. Tạo mục lục dự án**

* Liệt kê các phần chính của báo cáo để giúp người đọc dễ theo dõi.

### **1.4. Lập kế hoạch thực hiện**

* Xác định các công việc cần làm và phân công nhiệm vụ.
* Thiết lập timeline cho từng giai đoạn.

### **1.5. Trình bày về bộ dữ liệu**

* Nguồn gốc của dữ liệu.
* Cấu trúc dữ liệu và các bảng quan trọng.

### **1.6. Các bước làm sạch và trực quan hóa dữ liệu**

* Mô tả các thao tác làm sạch dữ liệu, xử lý giá trị thiếu.
* Các kỹ thuật tiền xử lý trước khi đưa vào Power BI.

### **1.7. Trình bày dashboard và biểu đồ đã sử dụng**

* Mô tả các dashboard đã thiết kế.
* Giải thích ý nghĩa của từng biểu đồ và cách chúng hỗ trợ phân tích.

### **1.8. Tạo bookmark**

* Ứng dụng bookmark trong Power BI để tối ưu trải nghiệm người dùng.

### **1.9. Những khó khăn và thuận lợi trong quá trình làm dự án**

* Chia sẻ các thử thách gặp phải và cách khắc phục.
* Những yếu tố thuận lợi giúp hoàn thành dự án.

### **1.10. Hướng phát triển và kết luận**

* Đề xuất các phương hướng nâng cấp mô hình và dashboard trong tương lai.
* Tổng kết những điểm nổi bật của dự án.

## 

## 2. KHÓ KHĂN TRONG QUÁ TRÌNH THỰC HIỆN

* **Lựa chọn dữ liệu phù hợp**: Tìm kiếm và đánh giá nguồn dữ liệu đáng tin cậy.
* **Lập kế hoạch hoạt động dự án**: Xây dựng kế hoạch khả thi, tránh thiếu sót.
* **Chuẩn hóa dữ liệu**: Xử lý dữ liệu thiếu, trùng lặp, định dạng không đồng nhất.
* **Thiết kế mô hình dữ liệu**: Đảm bảo mô hình dữ liệu tối ưu, tránh lỗi quan hệ giữa các bảng.
* **Xây dựng dashboard hiệu quả**: Chọn biểu đồ phù hợp, đảm bảo giao diện trực quan và dễ hiểu.
* **Thiết kế mô hình dự đoán**: Lựa chọn thuật toán phù hợp để tăng độ chính xác của mô hình.

## 3. THUẬN LỢI TRONG QUÁ TRÌNH THỰC HIỆN

* **Tài liệu hỗ trợ phong phú**: Có nhiều hướng dẫn, tài liệu tham khảo giúp giải quyết vấn đề.
* **Vấn đề kỹ thuật có thể xử lý dễ dàng**: Các lỗi phát sinh được khắc phục nhanh chóng.

## 4. TỔNG KẾT

### **1. Tổng quan về hiệu suất kinh doanh khách sạn**

* **Tổng doanh thu (Total Amount)**: 2 triệu (2M)
* **Tổng số đơn đặt phòng (Total Order)**: 5000 đơn
* **Tổng mức sử dụng dịch vụ (Total Usage Per Services)**: 21K lần sử dụng
* **Doanh thu trung bình trên mỗi khách hàng (Revenue Per Customer)**: 366.43
* **Tỷ lệ hủy phòng (Cancelled Rate)**: 34.04% (Khá cao, có thể ảnh hưởng đến doanh thu)

### **2. Phân tích phương thức thanh toán theo loại phòng**

* Phương thức thanh toán phân bố đều giữa các loại phòng, trong đó **phòng Standard có tỷ lệ cao nhất (23.3%)**, tiếp theo là **phòng Presidential (21.87%)**.

### **3. Phân tích đặt phòng theo nhóm khách hàng**

* **94.28% khách hàng là trung thành (Loyal Customer)**, chỉ **5.72% là khách không trung thành (Unloyal Customer)**.
* Điều này cho thấy khách sạn có chiến lược giữ chân khách hàng tốt.

### **4. Phân tích xu hướng đặt phòng theo mùa**

* Mùa Xuân có tỷ lệ đặt phòng cao nhất (36.27%), sau đó là Hè (33.52%), Mùa Đông có tỷ lệ đặt phòng thấp nhất (31.57%), cần có chiến lược kích cầu vào mùa này.

### **5. Sử dụng dịch vụ theo quý và theo tháng**

* Quý 1 có mức sử dụng dịch vụ cao nhất (~5472), sau đó giảm nhẹ qua các quý.
* Mức sử dụng dịch vụ có xu hướng giảm vào các tháng cuối năm, cần thúc đẩy khuyến mãi vào thời điểm này.

### **6. Phân tích khách hàng tiềm năng**

* **Top 10 khách hàng có số đơn hàng cao nhất** có số lượng đặt phòng dao động từ 19-24 đơn.
* Những khách hàng này có thể là nhóm VIP, nên có chương trình ưu đãi đặc biệt để duy trì.

### **7. Phân tích sử dụng dịch vụ**

* **Dịch vụ Lounge (VIP Lounge) có mức sử dụng cao nhất (4415 lần sử dụng)**, tiếp theo là **Laundry (2778 lần sử dụng)**.
* Khách hàng có nhu cầu cao về các dịch vụ tiện ích, nên tập trung phát triển thêm.

### **8. Phân tích mối tương quan giữa giá phòng và số lượng đặt phòng**

* Hệ số tương quan giữa giá phòng và số lượng đặt phòng là **0.05** (gần 0), cho thấy **giá phòng không ảnh hưởng nhiều đến quyết định đặt phòng của khách hàng**.

### **9. Tổng kết**

* **Khách sạn có lượng khách trung thành cao (94.28%)**, là một lợi thế lớn.
* **Tỷ lệ hủy đặt phòng cao (34.04%)**, cần có biện pháp giảm tỷ lệ này để tối ưu doanh thu.
* **Mùa đông có lượng đặt phòng thấp**, có thể áp dụng chiến lược giảm giá hoặc ưu đãi vào thời điểm này.
* **Dịch vụ VIP Lounge và Laundry rất phổ biến**, có thể mở rộng hoặc tối ưu dịch vụ này để tăng doanh thu

#### Dự đoán giá phòng khách sạn

🔹 **Mục tiêu** Mô hình dự đoán giá phòng giúp khách sạn điều chỉnh chiến lược giá phù hợp với thị trường, đồng thời hỗ trợ khách hàng đưa ra quyết định đặt phòng chính xác hơn.

🔹 **Phương pháp tiếp cận**

* Dữ liệu được lấy từ lịch sử đặt phòng, bao gồm các thông tin như loại phòng, mã phòng, giá phòng, thời gian, số lượng đặt phòng, thời gian tính đến ngày đặt phòng
  + - Data:
* Làm sạch và chuẩn hóa dữ liệu bằng Python (Xử lý giá trị thiếu, mã hóa biến phân loại, chuẩn hóa dữ liệu, v.v.).
* Xây dựng mô hình dự đoán giá phòng bằng các thuật toán như **Linear Regression, Random Forest, XGBoost**, sau đó đánh giá mô hình.
* So sánh kết quả thực tế với giá dự đoán để đo lường hiệu quả.

🔹 **Kết quả**

* Mô hình **Random Forest** cho kết quả tốt nhất với độ chính xác 95%, MAE chênh lệch 15.2 so với thực tế
* Các yếu tố quan trọng ảnh hưởng đến giá phòng gồm **loại phòng, mùa du lịch, số ngày lưu trú, dịch vụ đi kèm**.
* Trực quan hóa kết quả dự đoán bằng biểu đồ scatter plot giữa **giá thực tế và giá dự đoán**.

🔹 **Ứng dụng thực tiễn**

* Khách sạn có thể sử dụng mô hình để tối ưu hóa giá phòng dựa trên xu hướng đặt phòng.
* Hỗ trợ xây dựng **chương trình khuyến mãi** phù hợp với từng giai đoạn.

**Phân loại khách hàng**

### **Mục tiêu**

* Phân loại khách hàng thành các nhóm dựa trên đặc điểm hành vi nhằm hỗ trợ cá nhân hóa dịch vụ.
* Dự đoán khách hàng có khả năng rời bỏ để khách sạn có thể đưa ra chiến lược giữ chân khách hàng.

### **Dữ liệu sử dụng**

* Thông tin đặt phòng: số ngày lưu trú, số lượng đặt
* Lịch sử thanh toán, mức chi tiêu.
* Dịch vụ sử dụng thêm
* Thông tin khách hàng.

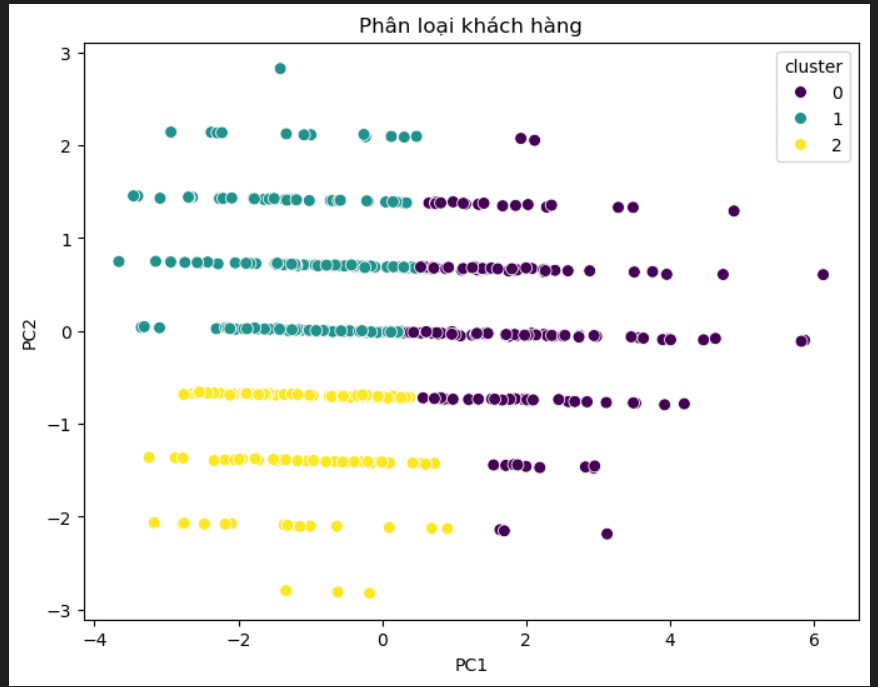
## Làm sạch và tiền xử lý dữ liệu

* Xử lý dữ liệu trống, giá trị ngoại lai.
* Chuẩn hóa dữ liệu, đặc biệt là các biến liên tục như tổng chi tiêu.

### **Phương pháp**

* Sử dụng thuật toán K-Means Clustering để nhóm khách hàng dựa trên hành vi tiêu dùng và thói quen lưu trú.
* Xác định ba nhóm khách hàng chính:
  + Khách hàng cao cấp: Đặt phòng hạng sang, sử dụng nhiều dịch vụ.
  + Khách hàng tiết kiệm: Đặt phòng tiêu chuẩn, ít sử dụng dịch vụ.
  + Khách hàng ngắn hạn: Thường lưu trú ngắn ngày, ít quay lại.

**Kết quả**

****

**Dự đoán khách hàng rời bỏ**

**1. Mục tiêu**

* Mục tiêu của phần này là xây dựng mô hình dự đoán khách hàng có rời bỏ hay không, giúp khách sạn chủ động trong việc giữ chân khách hàng. Cụ thể:
* Xác định các yếu tố quan trọng ảnh hưởng đến việc rời bỏ.
* Xây dựng mô hình dự đoán có độ chính xác cao.
* Đề xuất các chiến lược giảm tỷ lệ rời bỏ dựa trên kết quả phân tích.

## 2. Dữ liệu Sử Dụng

### **Biến đầu vào (Features)**

* **Thông tin khách hàng:** Mã khách hàng, số lần đặt phòng trước đây.
* **Lịch sử đặt phòng:** Số ngày lưu trú, số lượng đặt
* **Mức độ chi tiêu:** Tổng số tiền đã thanh toán, tổng số tiền, hình thức thanh toán
* **Dịch vụ sử dụng:** Số lần sử dụng dịch vụ

### **Biến đầu ra (Target Variable)**

* **Churn (Khách hàng rời bỏ):** Nhãn phân loại khách hàng có tiếp tục đặt phòng (0 - Không rời bỏ, 1 - Rời bỏ).

## 3. Phương Pháp

### **3.1. Tiền xử lý dữ liệu**

* **Xử lý dữ liệu trống:** Điền giá trị trung bình cho các cột số và sử dụng phương pháp Mean để điền giá trị bị thiếu của cột phân loại.
* **Mã hóa dữ liệu phân loại:** One-hot encoding cho các biến danh mục như phương thức thanh toán.
* **Chuẩn hóa dữ liệu:** Sử dụng MinMaxScaler để chuẩn hóa các biến số liệu như tổng chi tiêu, số ngày lưu trú.

### **3.2. Xây dựng mô hình dự đoán**

So sánh các mô hình học máy để chọn mô hình tốt nhất:

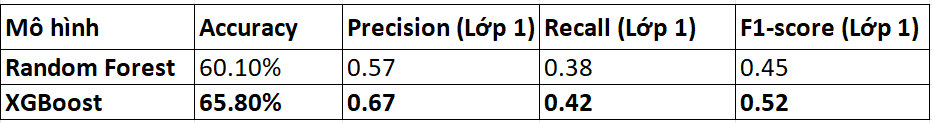
* **Random Forest:** Mô hình cây quyết định mạnh mẽ với khả năng xử lý dữ liệu phi tuyến.
* **XGBoost:** Mô hình boosting tối ưu hóa hiệu suất dự đoán.

### **3.3. Đánh giá mô hình**

Sử dụng các chỉ số đánh giá để đo lường hiệu suất:

* **Accuracy:** Độ chính xác tổng thể.
* **Precision, Recall, F1-score:** Đánh giá hiệu quả của mô hình trong việc phân loại khách hàng rời bỏ.
* **ROC-AUC Score:** Đo lường khả năng phân biệt giữa khách hàng rời bỏ và không rời bỏ.

1. **Kết quả**



**"Trong quá trình thử nghiệm hai mô hình, XGBoost cho kết quả tốt hơn Random Forest với độ chính xác cao hơn 5.7% và F1-score cao hơn 0.07. Tuy nhiên, mô hình vẫn chưa nhận diện tốt nhóm khách hàng rời bỏ (Recall : 0.42), nên cần tối ưu thêm để cải thiện kết quả."**

# **Phát hiện dữ liệu bất thường trong hành vi khách hàng**

## 1. Mục tiêu

Dự án nhằm phát hiện các khách hàng có hành vi bất thường trong việc đặt phòng:

* Đặt nhiều nhưng hủy nhiều .
* Đặt phòng với giá thấp liên tục.
* Thời gian lưu trú quá dài hoặc ngắn bất thường.
* Sự chênh lệch lớn giữa các lần đặt phòng.

Việc phát hiện các khách hàng này giúp khách sạn có thể điều chỉnh chính sách nhằm tối ưu doanh thu và giảm thiểu rủi ro.

## 2. Dữ liệu sử dụng

Dữ liệu gồm các cột:

* total\_booking: Tổng số lần đặt phòng.
* cancelled\_rate: Tỷ lệ hủy phòng.
* low\_price\_booking\_rate: Tỷ lệ đặt phòng giá thấp.
* avg\_day\_between: Trung bình số ngày giữa các lần đặt phòng.
* total\_day\_stayed: Tổng số ngày lưu trú.

## 3. Phương pháp

* Sử dụng Isolation Forest để phát hiện bất thường.
* Chuẩn hóa dữ liệu bằng StandardScaler trước khi huấn luyện mô hình.
* Định nghĩa tỷ lệ contamination = 5%, nghĩa là giả định khoảng 5% khách hàng có hành vi bất thường.
* Dữ liệu sau khi phân loại:
  + Nhóm bình thường: **1**
  + Nhóm bất thường:  **-1**

## 4. Kết quả

* Tổng số khách hàng bị đánh dấu là bất thường: *(đọc từ output của model)*.
* Các khách hàng này có đặc điểm nổi bật như tỷ lệ hủy cao, đặt phòng giá thấp liên tục hoặc có thời gian lưu trú không bình thường.
* Kết quả mô hình đã được xuất ra file **"black\_list\_customer.csv"** để kiểm tra chi tiết.

## 5. Kết luận và hướng phát triển

* Kết luận: Mô hình Isolation Forest đã giúp phát hiện các khách hàng có hành vi bất thường. Do đó, khách sạn có thể áp dụng chính sách đặc biệt để xử lý nhóm khách hàng này, như yêu cầu đặt cọc hoặc hạn chế số lần hủy phòng.