BÁO CÁO CASE STUDY 07

PHÂN TÍCH DỮ LIỆU KHÁCH SẠN / DU LỊCH VIỆT NAM

1. Phương pháp xử lý và phân tích dữ liệu

Dữ liệu trong Case Study 07 được xử lý và phân tích bằng ngôn ngữ **Python** với thư viện **Pandas**. Quy trình thực hiện gồm các bước chính sau:

1.1. Làm sạch dữ liệu

- Chuẩn hóa các trường dữ liệu dạng text bằng `str.strip()` và `str.title()` nhằm loại bỏ khoảng trắng thừa và thống nhất định dạng chữ.
- Chuẩn hóa tên thành phố để tránh sai lệch khi thực hiện thống kê và tổng hợp.
- Xử lý cột `base_price` bằng **Regular Expression (Regex)** để loại bỏ ký tự tiền tệ và ký tự không hợp lệ.
- Chuyển đổi dữ liệu sang kiểu số bằng `pd.to_numeric(errors="coerce")` để xử lý các giá trị sai định dạng.
- Chuẩn hóa cột ngày tháng bằng `pd.to_datetime()`.
- Xử lý các giá trị thiếu (`NaN`) và giá trị không hợp lệ (ví dụ: số đêm lưu trú âm).

1.2. Truy vấn và thống kê mô tả

- Thống kê số lượt booking theo từng thành phố.
- Xác định các booking chưa có đánh giá của khách hàng.
- Tính số đêm lưu trú trung bình cho mỗi booking.
- Tính số lượng khách trung bình theo từng loại phòng.

1.3. Tổng hợp dữ liệu (GroupBy)

- Sử dụng `groupby()` để tính giá phòng trung bình theo thành phố và loại phòng.
- Tính điểm rating trung bình theo từng khách sạn.
- Phân tích số lượt booking theo khách sạn để phát hiện các khách sạn có lượng đặt phòng thấp.

1.4. Merge dữ liệu

- Thực hiện merge giữa dữ liệu khách sạn, dữ liệu booking và dữ liệu đánh giá khách hàng.
- Đảm bảo các khóa nối (`hotel_id`, `booking_id`) được chuẩn hóa nhằm tránh mất dữ liệu khi kết hợp bảng.

1.5. Pivot Table và Stack/Unstack

- Sử dụng **Pivot Table** để phân tích dữ liệu theo nhiều chiều như: thành phố, loại phòng, giá phòng và rating.
- Áp dụng `stack()` và `unstack()` để chuyển đổi cấu trúc bảng dữ liệu và làm việc với **MultiIndex**.

2. Các bảng kết quả chính

Một số bảng kết quả tiêu biểu thu được từ quá trình phân tích bao gồm:

- Bảng thống kê số lượt booking theo từng thành phố.
- Bảng số đêm lưu trú trung bình và số khách trung bình theo loại phòng.
- Bảng giá phòng trung bình theo thành phố và loại phòng (GroupBy).
- Bảng rating trung bình của từng khách sạn.
- Bảng dữ liệu tổng hợp sau khi merge (hotel + booking + review).
- Pivot Table thể hiện mối quan hệ giữa thành phố, loại phòng và mức giá/rating.

(Các bảng chi tiết được trình bày trong các notebook tương ứng với từng Task.)

3. Nhận xét dựa trên kết quả phân tích

- Các thành phố lớn như **Hà Nội** và **TP.HCM** có số lượt booking cao hơn đáng kể so với các khu vực khác.
- Số đêm lưu trú trung bình ở mức ổn định, cho thấy nhu cầu lưu trú ngắn hạn chiếm ưu thế.
- Phòng **Standard** có số lượt đặt nhiều nhất do mức giá phù hợp, trong khi phòng **Suite** có giá cao nhất nhưng ít booking hơn.
- Một số khách sạn có **rating trung bình cao** nhưng số lượt booking còn thấp, cho thấy tiềm năng cải thiện doanh thu nếu được quảng bá tốt hơn.

4. Kết luận

Qua Case Study 07, nhóm đã:

- Xây dựng được bộ dữ liệu sạch và nhất quán từ dữ liệu thô ban đầu.
- Áp dụng thành công các kỹ thuật xử lý và phân tích dữ liệu bằng Pandas.
- Rút ra được những nhận xét có ý nghĩa hỗ trợ cho việc đánh giá và ra quyết định trong lĩnh vực kinh doanh khách sạn.

Bài làm giúp nhóm củng cố kiến thức về **Pandas**, đồng thời nâng cao tư duy phân tích dữ liệu trong bối cảnh thực tế.

5. Nhật ký nhóm

Thành viên	Công việc tham gia
Nguyễn Đức Hiếu	Làm sạch dữ liệu, Task 01
Trần Văn Hiếu	Truy vấn & thống kê, Task 02
Phạm Thùy Anh	GroupBy & tổng hợp, Task 03
Bùi Tuấn Anh	Merge dữ liệu
Nguyễn Viết Trung Hiếu	Pivot table + Stack/Unstack

Các thành viên phối hợp làm việc theo đúng phân công, đảm bảo tiến độ và chất lượng bài làm.