|  |
| --- |
| **ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**  **KHOA HỆ THỐNG THÔNG TIN** |
| Logo UIT Web Transparent |
| **BÁO CÁO ĐỒ ÁN**  **KHO DỮ LIỆU VÀ OLAP**  **ĐỀ TÀI:**  **XÂY DỰNG KHO DỮ LIỆU ĐẶT PHÒNG KHÁCH SẠN** |
| **GIẢNG VIÊN HƯỚNG DẪN**   |  | | --- | | ThS. Nguyễn Thị Kim Phụng |   **SINH VIÊN THỰC HIỆN**   |  | | --- | | Trần Phương Anh – 21520595  Trần Thị Luyến – 21521107 |   Thành phố Hồ Chí Minh, tháng 07 năm 2024 |

LỜI CẢM ƠN

Lời đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến toàn thể giảng viên trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM cũng như là nhà trường vì đã giúp nhóm chúng em có những kiến thức cơ bản làm nền tảng để thực hiện đồ án môn học này.

Đặc biệt hơn nữa, chúng em xin gửi lời cảm ơn chân thành đến cô Nguyễn Thị Kim Phụng – Giảng viên môn Kho dữ liệu và OLAP. Thời gian vừa qua, cô đã trực tiếp giảng dạy, truyền đạt kinh nghiệm, hướng dẫn chúng em một cách tận tình giúp nhóm chúng em hoàn thành tốt đồ án môn học của mình. Chúc cô sẽ luôn dồi dào sức khỏe, tràn đầy nhiệt huyết để có thể tiếp tục giảng dạy, dìu dắt những thế hệ sinh viên tiếp theo.

Và để đồ án này được hoàn thành thì không thể nào không nhắc đến những người đã làm ra nó. Cảm ơn thành viên trong nhóm đã chăm chỉ và chịu khó hoàn thành nhiệm vụ đúng tiến độ để đồ án của chúng ta được hoàn thành đúng hạn.

Mặc dù đã vận dụng tối đa những gì đã học được nhưng chúng em vẫn khó có thể tránh khỏi những sai sót. Chính vì vậy, nhóm chúng em rất mong nhận được sự góp ý từ phía cô để có thể hoàn thiện một cách tốt nhất có thể. Qua đó cũng tích lũy và học hỏi kinh nghiệm để làm hành trang cho tương lai.

Lời cuối cùng, chúng em một lần nữa xin được chân thành cảm ơn.

|  |
| --- |
| Thành phố Hồ Chí Minh, tháng 06 năm 2024  Nhóm sinh viên thực hiện |

NHẬN XÉT CỦA GIẢNG VIÊN

MỤC LỤC

[CHƯƠNG 1. GIỚI THIỆU 1](#_Toc171110141)

[1.1. Tổng quan đề tài 1](#_Toc171110142)

[1.2. Phát biểu về dữ liệu 1](#_Toc171110143)

[1.2.1. Mô tả dữ liệu 1](#_Toc171110144)

[1.2.2. Thuộc tính kho dữ liệu 2](#_Toc171110145)

[1.3. Xây dựng kho dữ liệu 4](#_Toc171110146)

[1.3.1. Lược đồ bông tuyết 4](#_Toc171110147)

[1.3.2. Dim\_Guest 4](#_Toc171110148)

[1.3.3. Dim\_Country 5](#_Toc171110149)

[1.3.4. Dim\_Reservation 5](#_Toc171110150)

[1.3.5. Dim\_Payment 5](#_Toc171110151)

[1.3.6. Dim\_Room 6](#_Toc171110152)

[1.3.7. Dim\_Date 6](#_Toc171110153)

[1.3.8. Fact 6](#_Toc171110154)

[CHƯƠNG 2. TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS) 8](#_Toc171110155)

[2.1. Tạo Project và thực hiện kết nối 8](#_Toc171110156)

[2.1.1. Tạo Project 8](#_Toc171110157)

[2.1.2. Tạo cơ sở dữ liệu 9](#_Toc171110158)

[2.2. Chuẩn bị dữ liệu 10](#_Toc171110159)

[2.3. Quá trình lọc dữ liệu 13](#_Toc171110160)

[2.4. Quá trình tạo các bảng Dimension 16](#_Toc171110161)

[2.5. Quá trình tạo bảng Fact 20](#_Toc171110162)

[2.6. Quy trình và lược đồ dữ liệu 23](#_Toc171110163)

[2.6.1. Quy trình ETL 23](#_Toc171110164)

[2.6.2. Lược đồ bông tuyết 24](#_Toc171110165)

[CHƯƠNG 3. PHÂN TÍCH DỮ LIỆU TRONG KHO (SSAS) 24](#_Toc171110166)

[3.1. Quá trình SSAS 24](#_Toc171110167)

[3.1.1. Quá trình thực hiện trên Visual Data Studio 24](#_Toc171110168)

[3.1.2. Điều chỉnh và phân cấp các Dimension 34](#_Toc171110169)

[3.2. Truy vấn MDX 38](#_Toc171110170)

[3.2.1. Doanh thu từng loại phòng qua các năm 38](#_Toc171110171)

[3.2.2. Doanh thu từng nguồn đặt phòng của từng quý qua các năm 38](#_Toc171110172)

[3.2.3. Danh sách khách hàng đặt phòng loại ‘Standard’ 39](#_Toc171110173)

[3.2.4. Tổng doanh thu theo loại phòng và tháng năm 2023 40](#_Toc171110174)

[3.2.5. Tổng số lượng khác theo quốc gia và loại phòng năm 2023 40](#_Toc171110175)

[3.2.6. Số lượng đặt phòng theo từng tháng và từng loại phòng 41](#_Toc171110176)

[3.2.7. Thông tin khách từ Thụy Sĩ (Switzerland) , chỉ hiển thị các cột có giá trị tổng số tiền chi tiêu lớn hơn 1000 42](#_Toc171110177)

[3.2.8. Số lượng khách nữ từ Thụy Sĩ (Switzerland) với tổng số tiền chi tiêu lớn hơn 1000 43](#_Toc171110178)

[3.2.9. Số lượng đặt phòng qua các nguồn đặt phòng qua các năm 44](#_Toc171110179)

[3.2.10. Thông tin khách hàng nam ở phòng loại ‘Deluxe’, chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 740 45](#_Toc171110180)

[3.2.11. Tìm các khách hàng đặt phòng có số lượng đêm ở lớn hơn 4 trong quý 4/2023 46](#_Toc171110181)

[3.2.12. Mỗi quý trong năm 2023, cho biết tháng nào có tổng doanh thu cao nhất 47](#_Toc171110182)

[3.2.13. Doanh thu của loại phòng ‘Deluxe’ qua từng năm 48](#_Toc171110183)

[3.2.14. Phần trăm số lượng đặt phòng có sử dụng bữa sáng 48](#_Toc171110184)

[3.2.15. Phần trăm doanh thu từ mỗi nguồn đặt phòng năm 2023 50](#_Toc171110185)

[3.2.16. Top 3 doanh thu theo quốc gia từ các đơn đặt phòng có sử dụng dịch vụ đưa đón sân bay 52](#_Toc171110186)

[3.2.17. Số lượng đặt phòng của từng loại phòng theo mỗi nguồn đặt phòng 53](#_Toc171110187)

[3.2.18. Tổng số lượng đặt phòng theo từng loại phòng và tầng 54](#_Toc171110188)

[3.2.19. Trung bình số đêm ở theo từng tháng trong năm 2023 55](#_Toc171110189)

[3.2.20. Tổng doanh thu theo loại phòng 56](#_Toc171110190)

[3.2.21. Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng 57](#_Toc171110191)

[3.2.22. Liệt kê top 2 loại phòng có doanh thu nhiều nhất trong cả hai năm 2022 và năm 2023 58](#_Toc171110192)

[3.3. Thực hiện trên Excel 59](#_Toc171110193)

[3.3.1. Tổng doanh thu theo loại phòng qua các năm (group by) 59](#_Toc171110194)

[3.3.2. Tổng số đơn đặt phòng theo nguồn đặt phòng (group by) 59](#_Toc171110195)

[3.3.3. Số lượng khách từ Thụy Sĩ (Switzerland) , chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 1000 (table) 60](#_Toc171110196)

[3.3.4. Tạo ma trận hiển thị số lượng khách hàng theo loại phòng và nguồn đặt phòng (matrix) 62](#_Toc171110197)

[3.3.5. Tạo biểu đồ hiển thị tổng doanh thu theo tháng năm 2023 (chart) 63](#_Toc171110198)

[3.3.6. Tạo biểu đồ hiển thị top 2 doanh thu theo tháng năm 2023 (chart) 63](#_Toc171110199)

[3.3.7. Số lượng đặt phòng theo từng tháng và từng loại phòng trong năm 2023 65](#_Toc171110200)

[3.3.8. Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng 67](#_Toc171110201)

[3.4. Thực hiện trên Power BI 69](#_Toc171110202)

[3.4.1. Tổng doanh thu theo loại phòng qua các năm 70](#_Toc171110203)

[3.4.2. Tổng số lượng khách hàng theo nguồn đặt phòng 72](#_Toc171110204)

[3.4.3. Số lượng khách từ Thụy Sĩ (Switzerland), chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 1000 74](#_Toc171110205)

[3.4.4. Tạo ma trận hiển thị số lượng đặt phòng theo loại phòng và nguồn đặt phòng 75](#_Toc171110206)

[3.4.5. Tạo biểu đồ hiển thị tổng doanh thu theo tháng năm 2023 76](#_Toc171110207)

[3.4.6. Tạo biểu đồ hiển thị top 2 doanh thu theo tháng năm 2023 79](#_Toc171110208)

[3.4.7. Số lượng đặt phòng theo từng tháng và từng loại phòng trong năm 2023 80](#_Toc171110209)

[3.4.8. Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng 81](#_Toc171110210)

[CHƯƠNG 4. QUÁ TRÌNH DATA MINING 84](#_Toc171110211)

[4.1. Tổng quan đề tài 84](#_Toc171110212)

[4.2. Lý thuyết mô hình phân cụm 85](#_Toc171110213)

[4.2.1. K-Means 85](#_Toc171110214)

[4.2.2. DBSCAN (Density-Based Clustering) 86](#_Toc171110215)

[4.2.3. Agglomerative Clustering 87](#_Toc171110216)

[4.2.4. Thống kê mô tả 88](#_Toc171110217)

[4.3. Tiền xử lý dữ liệu 90](#_Toc171110218)

[4.3.1. Xử lý giá trị thiếu 90](#_Toc171110219)

[4.3.2. Tính toán các chỉ số RFM 91](#_Toc171110220)

[4.3.3. Khám phá và xử lý dữ liệu RFM 92](#_Toc171110221)

[4.3.4. Chuẩn hóa dữ liệu 94](#_Toc171110222)

[4.4. Xây dựng mô hình phân cụm 95](#_Toc171110223)

[4.4.1. Xác định số lượng cụm (k) tối ưu bằng kỹ thuật Elbow 95](#_Toc171110224)

[4.4.2. Xây dựng các mô hình 96](#_Toc171110225)

[4.4.3. Đánh giá mô hình và trực quan hóa kết quả 96](#_Toc171110226)

[4.4.4. Phân tích & phân cụm khách hàng 97](#_Toc171110227)

DANH MỤC CHỮ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| Từ viết tắt | Từ đầy đủ | Ý nghĩa |
| DB | Database | Cơ sở dữ liệu |
| SSMS | SQL Server Management Studio |  |
| SSIS | SQL Server Integration Services |  |
| SSAS | SQL Server Analysis Services |  |
| ETL | Extract, Transform, Load | Trích xuất, Chuyển đổi, Nạp dữ liệu |
| MDX | Multidimensional Expressions | Ngôn ngữ truy vấn trong SSAS |

BẢNG PHÂN CÔNG NHIỆM VỤ

|  |  |  |
| --- | --- | --- |
| Công việc | Trần Phương Anh | Trần Thị Luyến |
| Xây dựng bố cục bài báo cáo | Checkmark with solid fill |  |
| Chương 1. Giới thiệu | Checkmark with solid fill | Checkmark with solid fill |
| Chương 2. SSIS | Checkmark with solid fill |  |
| Chương 3. SSAS | Checkmark with solid fill |  |
| Chương 4. Data Mining |  | Checkmark with solid fill |
| Tổng hợp và chỉnh sửa báo cáo | Checkmark with solid fill | Checkmark with solid fill |
| Hoàn thành | 100% | 100% |

NỘI DUNG

# GIỚI THIỆU

## Tổng quan đề tài

Du lịch và lữ hành chiếm hơn 10% GDP trên toàn thế giới và đang có xu hướng chiếm được phần lớn hơn trong chiếc bánh toàn cầu. Đồng thời, đây là ngành tạo ra khối lượng dữ liệu khổng lồ và việc tận dụng lợi thế của nó có thể giúp các doanh nghiệp nổi bật giữa đám đông.

Đề tài "Xây dựng Kho dữ liệu OLAP về đặt phòng khách sạn" nhấn mạnh vào việc áp dụng công nghệ thông tin để cải thiện quản lý thông tin đặt phòng trong ngành du lịch và khách sạn. Trong bối cảnh ngày nay, việc quản lý thông tin khách hàng, dự đoán nhu cầu và tối ưu hóa dịch vụ là rất quan trọng để cung cấp trải nghiệm tốt nhất cho khách hàng và tăng cường hiệu suất kinh doanh cho các doanh nghiệp trong ngành.

Hệ thống đặt phòng khách sạn truyền thống thường gặp phải nhiều thách thức, bao gồm việc xử lý lượng dữ liệu lớn từ các giao dịch đặt phòng, theo dõi xu hướng và sở thích của khách hàng, cũng như dự đoán nhu cầu đặt phòng trong tương lai. Để giải quyết những thách thức này, việc xây dựng một Kho dữ liệu OLAP (Online Analytical Processing) là cần thiết.

Chúng em sẽ sử dụng Microsoft SQL Server Integration Services (SSIS) để triển khai quy trình ETL (Extract, Transform, Load) để lấy dữ liệu từ các nguồn khác nhau và chuyển đổi chúng thành định dạng phù hợp cho việc phân tích. Sau đó, sử dụng Microsoft SQL Server Analysis Services (SSAS) để xây dựng các cube và dimension, tạo ra một môi trường phân tích dữ liệu linh hoạt và mạnh mẽ. Bên cạnh đó, chúng em sử dụng công cụ Power BI để phát triển các báo cáo và trực quan hóa dữ liệu giúp người dùng dễ dàng tìm hiểu và phân tích thông tin.

Cuối cùng, khám phá ứng dụng của kỹ thuật Data Mining trong việc phân tích dữ liệu đặt phòng khách sạn, từ việc phát hiện mẫu tự nhiên trong dữ liệu đến việc dự đoán xu hướng và hành vi của khách hàng.

## Phát biểu về dữ liệu

### Mô tả dữ liệu

Bộ dữ liệu cung cấp dữ liệu đặt phòng trong hai mùa liên tiếp (2021 - 2023) của một khách sạn hạng sang

Bộ dữ liệu gồm 32 thuộc tính và 9974 dòng

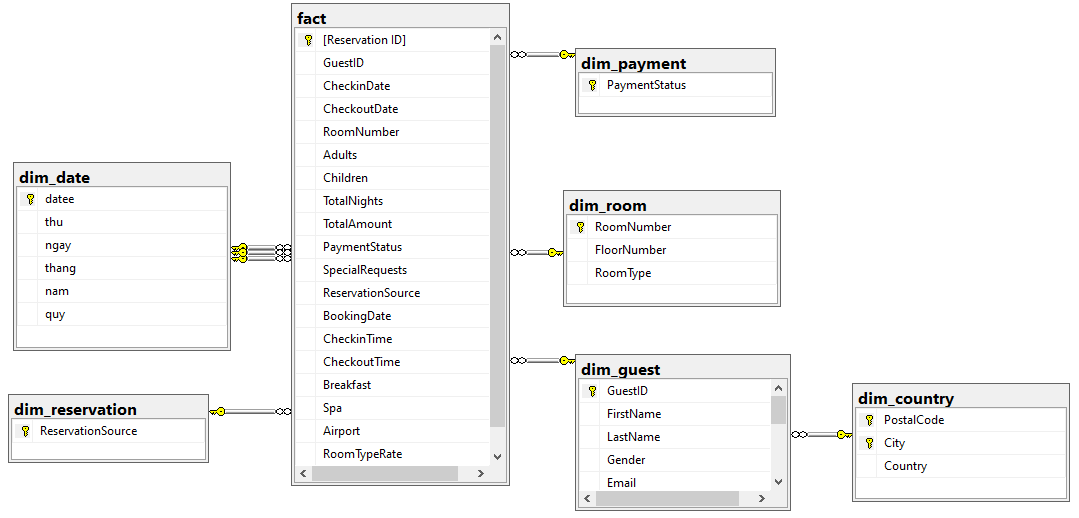
Link dữ liệu: <https://www.kaggle.com/datasets/dimitrisangelide/hotel-reservations-data>

### Thuộc tính kho dữ liệu

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ý nghĩa |
| 1 | Reservation ID | int | Mã định danh duy nhất cho đặt phòng |
| 2 | Guest ID | int | Mã định danh duy nhất của người đã đặt phòng |
| 3 | First Name | string | Tên của khách |
| 4 | Last Name | string | Họ của khách |
| 5 | Gender | string | Giới tính |
| 6 | Email | string | Địa chỉ email |
| 7 | Phone | string | Số điện thoại |
| 8 | Nationality | string | Quốc tịch |
| 9 | Birthdate | date | Ngày sinh |
| 10 | Address | string | Địa chỉ nhà |
| 11 | City | string | Thành phố mà khách sống |
| 12 | Postal Code | int | Mã bưu chính |
| 13 | Country | string | Quốc gia cư trú |
| 14 | Check-in Date | date | Ngày dự kiến ​​nhận phòng |
| 15 | Check-out Date | date | Ngày dự kiến ​​trả phòng |
| 16 | Room Number | int | Số phòng đã đặt |
| 17 | Floor Number | int | Tầng của phòng đã đặt |
| 18 | Room Type | string | Loại phòng đã đặt |
| 19 | Adults | int | Số người lớn ở trong phòng |
| 20 | Children | int | Số trẻ em ở trong phòng |
| 21 | Total Nights | int | Tổng số đêm đã đặt |
| 22 | Total Amount | int | Tổng số tiền đã thanh toán cho đặt phòng |
| 23 | Payment Status | string | Trạng thái thanh toán vào thời điểm đặt phòng |
| 24 | Special Requests | string | Yêu cầu đặc biệt của khách |
| 25 | Reservation Source | string | Nguồn thông tin mà khách sử dụng để đặt phòng |
| 26 | Booking Date | date | Ngày đặt phòng diễn ra |
| 27 | Check-in Time | time | Thời gian dự kiến ​​nhận phòng |
| 28 | Check-out Time | time | Thời gian dự kiến ​​trả phòng |
| 29 | Breakfast Included | boolean | Nếu bữa sáng đã bao gồm trong đặt phòng |
| 30 | Spa Package Included | string | Nếu gói dịch vụ spa đã bao gồm trong đặt phòng. |
| 31 | Airport Pickup Included | string | Nếu khách yêu cầu đưa đón sân bay |
| 32 | Room Type Rate | int | Giá của mỗi loại phòng tại thời điểm đặt phòng |

## Xây dựng kho dữ liệu

### Lược đồ bông tuyết



### Dim\_Guest

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | GuestID | int | Primary key | Mã số của khách |
| 2 | FirstName | string |  | Tên của khách |
| 3 | LastName | string |  | Họ của khách |
| 4 | Gender | string |  | Giới tính |
| 5 | Email | string |  | Địa chỉ email |
| 6 | Phone | string |  | Số điện thoại |
| 7 | Nationality | string |  | Quốc tịch |
| 8 | Birthdate | date |  | Ngày sinh |
| 9 | Address | string |  | Địa chỉ nhà |
| 10 | City | string |  | Thành phố mà khách sống |
| 11 | PostalCode | int |  | Mã bưu chính |

### Dim\_Country

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | City | string | Primary key | Thành phố mà khách sống |
| 2 | PostalCode | int | Primary key | Mã bưu chính |
| 3 | Country | string |  | Quốc gia cư trú |

### Dim\_Reservation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | ReservationSource | string | Primary key | Nguồn thông tin mà khách sử dụng để đặt phòng |

### Dim\_Payment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | PaymentStatus | string | Primary key | Trạng thái thanh toán vào thời điểm đặt phòng |

### Dim\_Room

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | RoomNumber | int | Primary key | Số phòng đã đặt |
| 2 | FloorNumber | int |  | Tầng của phòng đã đặt |
| 3 | RoomType | string |  | Loại phòng đã đặt |

### Dim\_Date

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | Date\_ngay | date | Primary key | Ngày dạng dd/MM/yyyy |
| 2 | Thu | string |  | Thứ trong tuần của ngày |
| 3 | Ngay | int |  | Ngày |
| 4 | Thang | int |  | Tháng |
| 5 | Nam | int |  | Năm |
| 6 | Quy | int |  | Quý |

### Fact

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Ràng buộc | Ý nghĩa |
| 1 | ReservationID | int | Primary key | Mã định danh duy nhất cho đặt phòng |
| 2 | GuestID | int | Foreign Key | Mã số của khách |
| 3 | CheckinDate | date | Foreign Key | Ngày dự kiến ​​nhận phòng |
| 4 | CheckoutDate | date | Foreign Key | Ngày dự kiến ​​trả phòng |
| 5 | RoomNumber | int | Foreign Key | Số phòng đã đặt |
| 6 | Adults | int |  | Số người lớn ở trong phòng |
| 7 | Children | int |  | Số trẻ em ở trong phòng |
| 8 | TotalNights | int |  | Tổng số đêm đã đặt |
| 9 | TotalAmount | int |  | Tổng số tiền đã thanh toán cho đặt phòng |
| 10 | PaymentStatus | string | Foreign Key | Trạng thái thanh toán vào thời điểm đặt phòng |
| 11 | SpecialRequests | string |  | Yêu cầu đặc biệt của khách |
| 12 | ReservationSource | string | Foreign Key | Nguồn thông tin mà khách sử dụng để đặt phòng |
| 13 | BookingDate | date | Foreign Key | Ngày đặt phòng diễn ra |
| 14 | CheckinTime | time |  | Thời gian dự kiến ​​nhận phòng |
| 15 | CheckoutTime | time |  | Thời gian dự kiến ​​trả phòng |
| 16 | Breakfast | boolean |  | Nếu bữa sáng đã bao gồm trong đặt phòng |
| 17 | Spa | string |  | Nếu gói dịch vụ spa đã bao gồm trong đặt phòng. |
| 18 | Airport | string |  | Nếu khách yêu cầu đưa đón sân bay |
| 19 | RoomTypeRate | int |  | Giá của mỗi loại phòng tại thời điểm đặt phòng |

# TÍCH HỢP DỮ LIỆU VÀO KHO (SSIS)

## Tạo Project và thực hiện kết nối

### Tạo Project

**Bước 1**: Mở Visual Studio (2022) và tạo mới dự án

**Bước 2:** Để tạo dự án SSIS, ta sử dụng công cụ “Integration Services Project”

A screenshot of a computer

Description automatically generated

**Bước 3**: Điền thông tin về dự án và “Create” để tạo mới

A screenshot of a black screen

Description automatically generated

**Bước 4**: Sau khi hoàn thành, ta có giao diện của SSIS

A screenshot of a computer

Description automatically generated

### Tạo cơ sở dữ liệu

Ta sử dụng công cụ SQL Server Management Studio (SSMS) để tạo Database cho dự án

Tên Database: SSIS

SSIS là nơi lưu trữ dữ liệu của các bảng Dim và Fact

* Fact
* Dim\_Guest
* Dim\_Date
* Dim\_Room
* Dim\_Country
* Dim\_Payment
* Dim\_Reservation

## Chuẩn bị dữ liệu

**Bước 1**: Trong SSIS Toolbox ở phía trái màn hình, tìm mục Other Sources à chọn Flat File Source và kéo vào Data Flow

A screenshot of a computer

Description automatically generated

**Bước 2**: Nhấp chuột phải vào Flat File Source vừa kéo, chọn Edit, hiển thị Editor để đổ dữ liệu vào và chọn New để thêm kết nối với DB

A screenshot of a computer

Description automatically generated

**Bước 3**: Điền tên Connection và chọn file dữ liệu (CSV)

A screenshot of a computer

Description automatically generated

**Bước 4**: Ở mục Advanced, chỉnh sửa kiểu dữ liệu cho khớp với dữ liệu trong file CSV (hệ thống mặc định kiểu dữ liệu là string, độ dài 50 ký tự)

A screenshot of a computer

Description automatically generated

**Bước 5:** Trong mục “Common”, chọn thành phần Multicast

Multicast giúp sao chép dữ liệu từ một nguồn đến nhiều đích khác nhau, giúp tạo ra nhiều nhánh luồng dữ liệu từ một luồng dữ liệu duy nhất (flat file source)

A screenshot of a computer

Description automatically generated

Nhóm em xử lý thuộc tính Ngày (CheckinDate, CheckoutDate, BookingDate) là gộp tất cả hàng/giá trị của 3 thuộc tính lại, tách riêng thành 1 sheet và lưu thành 1 file CSV riêng nên sẽ đổ dữ liệu bằng một Flat File Source khác, cách làm cũng tương tự như trên.

## Quá trình lọc dữ liệu

**Bước 1:** Trong mục “Common” kéo thành phần Sort để thực hiện sắp xếp và lọc dữ liệu, ở đây ta sẽ tạo 5 Sort cho 5 Dimension (Country, Room, Payment, Reservation và Date).

A diagram of a computer

Description automatically generated with medium confidence

**Bước 2**: Nhấp chuột phải vào thành phần Sort muốn xử lý (ở đây chọn sort\_payment) và chọn “Edit”

A screenshot of a computer

Description automatically generated

**Bước 3**: Chọn thuộc tính (PaymentStatus) muốn sắp xếp (tăng/giảm dần) và lọc dữ liệu trùng (Remove rows with duplicates)

A screenshot of a computer

Description automatically generated

**Bước 4**: Thực hiện tương tự với các Sort khác

**Bước 5:** Nhấn “Start” để tiến hành đổ dữ liệu vào

A diagram of rows and arrows

Description automatically generated

## Quá trình tạo các bảng Dimension

**Bước 1:** Trong mục “Other Destinations”, chọn thành phần “OLE DB Destination”. Đây là thành phần được sử dụng để nạp dữ liệu vào các bảng đích (bảng Dim) trong cơ sở dữ liệu

A screenshot of a computer

Description automatically generated

**Bước 2**: Kéo mũi tên xanh từ Sort qua Dim để đổ dữ liệu từ Sort. Nhấp chuột phải vào bảng Dim muốn chỉnh sửa (ở đây chọn dim\_country) và chọn “Edit”

A screenshot of a computer

Description automatically generated

**Bước 3**: Chọn OLE DB connection manager (nhấn “New”) và chọn Connection vừa tạo

A screenshot of a computer

Description automatically generated

**Bước 4:** Tạo bảng dim\_country

A screenshot of a computer program

Description automatically generated

**Bước 5**: Kiểm tra mapping và chọn “OK” để hoàn tất

A screenshot of a computer

Description automatically generated

**Bước 6:** Thực hiện tương tự với các bảng Dim khác

**Bước 7**: Kéo thả thành phần Execute SQL Task vào không gian Control Flow

A screenshot of a computer

Description automatically generated

**Bước 8**: Nhấp phải chọn “Edit”, điền thông tin cần thiết và chọn “Build Query” cho câu SQL vừa tạo (DELETE FROM)

A screenshot of a computer

Description automatically generated

**Bước 9**: Nhấn “Start” để đổ dữ liệu vào bảng Dim

A diagram of rows and rows

Description automatically generated

Do bảng dim\_guest có khóa ngoại từ các bảng Dim khác nên phải tạo thành công các bảng Dim đó mới tạo được dim\_guest

A diagram of a hotel

Description automatically generated

Thực hiện tương tự với bảng dim\_date

## Quá trình tạo bảng Fact

**Bước 1:** Kéo thành phần OLE DB Destination vào Data Flow

A screenshot of a computer

Description automatically generated

**Bước 2**: Kết nối bằng Connection Manager đến Database

**Bước 3:** Tạo bảng Fact bằng câu lệnh CREATE TABLE

A screenshot of a computer

Description automatically generated

**Bước 4:** Kiểm tra Mappings

A screenshot of a computer

Description automatically generated

**Bước 5:** Chỉnh sửa câu lệnh SQL trong Execute SQL Task và “Build Query”

A screenshot of a computer

Description automatically generated

**Bước 6:** Nhấn “Start” để tiến hành đổ dữ liệu vào bảng Fact

A diagram of a row

Description automatically generated with medium confidence

## Quy trình và lược đồ dữ liệu

### Quy trình ETL

* Extract (Trích xuất): Sử dụng thành phần Flat File Source để đọc dữ liệu từ flat file
* Transform (Chuyển đổi):
* Dữ liệu được làm sạch, chuyển đổi và chuẩn hóa để phù hợp với mô hình dữ liệu
* Các bước chuyển đổi bao gồm: lọc và sắp xếp dữ liệu, loại bỏ giá trị trùng lặp
* Load (Nạp): Sử dụng thành phần OLE DB Destination để nạp dữ liệu vào các bảng trong cơ sở dữ liệu đích gồm bảng fact (sự kiện) hoặc bảng dimension (chiều)

A diagram of a flowchart

Description automatically generated

### Lược đồ bông tuyết

A screenshot of a computer

Description automatically generated

# PHÂN TÍCH DỮ LIỆU TRONG KHO (SSAS)

## Quá trình SSAS

### Quá trình thực hiện trên Visual Data Studio

**Bước 1:** Tạo project SSAS trong Visual Studio 2022

**Bước 2:** Chọn Analysis Services Multidimensional Project và đặt tên

A screenshot of a computer

Description automatically generated

**Bước 3:** Chọn Data Source để chọn Database đã tạo ở quá trình SSIS

**A screenshot of a computer

Description automatically generated**

**Bước 4:** Chọn “Create a data source based on an existing or new connection” và chọn “New” để tạo Data Connection mới

A screenshot of a computer

Description automatically generated

**Bước 5:** Điền tên Server, tên Database muốn kết nối và chọn “OK”

A computer screen shot of a server

Description automatically generated

**Bước 6:** Điền username và password của Windows để kết nối

A screenshot of a computer

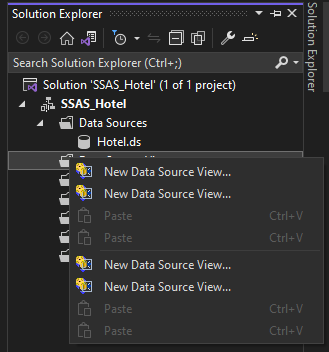
Description automatically generated

**Bước 7:** Đặt tên cho Data source và chọn “Finish”

A screenshot of a computer

Description automatically generated

**Bước 8:** Trong mục Data Source Views, chọn “New Data source view”



**Bước 9:** Chọn Data source vừa tạo và bấm “Next”

A screenshot of a computer

Description automatically generated

**Bước 10:** Chọn các bảng Dim và Fact cần sử dụng và chọn “Next”

A screenshot of a computer

Description automatically generated

**Bước 11:** Đặt tên tùy thích và “Finish”

A screenshot of a computer program

Description automatically generated

**Bước 12:** Trong mục Cubes, chọn “New cube”

A screenshot of a computer program

Description automatically generated

**Bước 13**: Chọn “Use existing tables”

A screenshot of a computer program

Description automatically generated

**Bước 14:** Chọn bảng có chứa Measures dùng để tính toán, ở đây là bảng Fact

A screenshot of a computer

Description automatically generated

**Bước 15:** Chọn các thuộc tính dùng để tính toán, mặc định sẽ chọn hết

A screenshot of a computer

Description automatically generated

**Bước 16:** Chọn các bảng Dim

A screenshot of a computer

Description automatically generated

**Bước 17:** Đặt tên cho Cube và xem lại cấu trúc

A screenshot of a computer program

Description automatically generated

**Bước 18**: Sau khi hoàn thành, màn hình sẽ hiển thị lược đồ Data Warehouse của Database

A screenshot of a computer

Description automatically generated

### Điều chỉnh và phân cấp các Dimension

**Bước 1:** Tại Dim Room, nhấp chuột hai lần để mở Dim Room

A screenshot of a computer menu

Description automatically generated

**Bước 2:** Kéo thả các thuộc tính của Dim Room vào Attributes

A screen shot of a computer

Description automatically generated

**Bước 3:** Thực hiện tương tự với bảng dim\_guest, dim\_country, dim\_payment, dim\_reservation

**Bước 4:** Với bảng dim\_date, kéo thả các thuộc tính vào Attributes (tương tự bước 2)

A screenshot of a computer

Description automatically generated

**Bước 5:** Kéo các thuộc tính vừa để vào trong Attributes vào Hierachies để tạo cây phân cấp

A screenshot of a computer

Description automatically generated

Khi kéo thả các thuộc tính vào trong Hierarchies để tạo cây phân cấp, hệ thống sẽ có thể xác định rõ ràng từng mức phân cấp mà không bị nhầm lẫn bởi các giá trị trùng lặp. Việc tạo mới "Ngay, Thang, Quy, Nam" nhằm đảm bảo tính duy nhất của các giá trị trong cây phân cấp, tránh sự trùng lặp và đảm bảo hệ thống có thể phân tích dữ liệu một cách chính xác và hiệu quả

**Bước 6:** Nhấp chuột phải vào thư mục gốc “SSAS\_Hotel” và chọn

Properties

A screenshot of a computer program

Description automatically generated

**Bước 7:** Trong mục Deployment, sửa tên Server và chọn “OK”

A screenshot of a computer

Description automatically generated

**Bước 8:** Nhấn chuột phải và chọn Deploy trong “SSAS\_Hotel” để triển khai dự án và kiểm tra các cube, dimension

A screenshot of a computer program

Description automatically generated

**Bước 9:** Trong Hotel.cube, chuột phải chọn Browse

A screenshot of a computer menu

Description automatically generated

**Bước 10:** Mặc định màn hình sẽ không hiển thị gì hết, cần load lại trang để hiện giao diện dùng để truy vấn

A screen shot of a computer

Description automatically generated

Sau khi load, hiển thị giao diện dùng để thực hiện các công việc như kéo thả, truy vấn MDX

A screenshot of a computer

Description automatically generated

## Truy vấn MDX

### Doanh thu từng loại phòng qua các năm

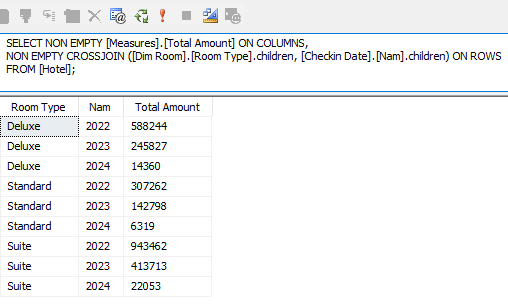
* Sử dụng Roll up
* Câu lệnh truy vấn MDX:

SELECT NON EMPTY[Measures].[Total Amount] ON COLUMNS,

NON EMPTY CROSSJOIN ([Dim Room].[Room Type].children, [Checkin Date].[Nam].children) ON ROWS

FROM [Hotel];

* Kết quả:



### Doanh thu từng nguồn đặt phòng của từng quý qua các năm

* Sử dụng Drill down
* Câu lệnh truy vấn MDX:

SELECT NON EMPTY [Measures].[Total Amount] ON COLUMNS,

NON EMPTY CROSSJOIN ([Dim Reservation].[Reservation Source].children, [Checkin Date].[Nam].children, [Checkin Date].[Quy].children) ON ROWS

FROM [Hotel];

* Kết quả:

A screenshot of a computer

Description automatically generated

### Danh sách khách hàng đặt phòng loại ‘Standard’

* Sử dụng Slice
* Câu lệnh truy vấn MDX:

SELECT NON EMPTY { [Measures].[Adults], [Measures].[Children], [Measures].[Total Nights] } ON COLUMNS,

NON EMPTY { ([Fact].[Reservation ID].[Reservation ID].ALLMEMBERS \* [Dim Guest].[Guest ID].[Guest ID].ALLMEMBERS ) } ON ROWS

FROM [Hotel]

WHERE [Dim Room].[Room Type].&[Standard];

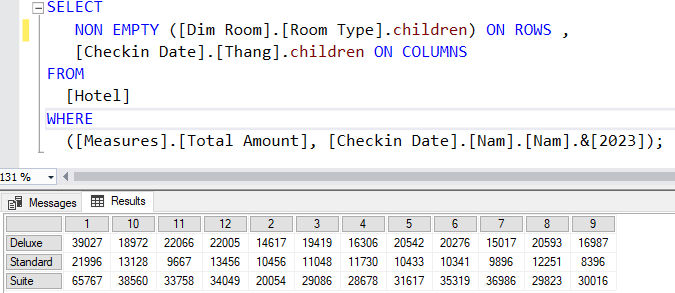
* Kết quả:

A screenshot of a data

Description automatically generated

### Tổng doanh thu theo loại phòng và tháng năm 2023

* Sử dụng Pivot
* Câu lệnh truy vấn MDX và kết quả:



### Tổng số lượng khác theo quốc gia và loại phòng năm 2023

* Tạo Calculate member [SumAdultsChildren]: Tính tổng số lượng khách gồm người lớn (Adults) và trẻ em (Children)

A screenshot of a computer

Description automatically generated

* Câu lệnh truy vấn MDX:

WITH

MEMBER [Measures].[SumAdultsChildren] AS

[Measures].[Adults] + [Measures].[Children]

SELECT

NON EMPTY ([Dim Room].[Room Type].children) ON COLUMNS ,

NON EMPTY ([Dim Guest].[Country].children) ON ROWS

FROM [Hotel]

WHERE

([Measures].[SumAdultsChildren], [Checkin Date].[Nam].[Nam].&[2023]);

* Kết quả:

A screenshot of a table

Description automatically generated

### Số lượng đặt phòng theo từng tháng và từng loại phòng

* Sử dụng Pivot
* Câu lệnh truy vấn MDX:

SELECT

NON EMPTY [Measures].[Fact Count] ON COLUMNS,

NON EMPTY CROSSJOIN(

[Checkin Date].[Thang].[Thang].Members,

[Dim Room].[Room Type].[Room Type].Members

) ON ROWS

FROM [Hotel]

WHERE ([Checkin Date].[Nam].&[2023]);

* Kết quả:

A screenshot of a computer

Description automatically generated

### Thông tin khách từ Thụy Sĩ (Switzerland) , chỉ hiển thị các cột có giá trị tổng số tiền chi tiêu lớn hơn 1000

* Sử dụng Slice
* Câu lệnh truy vấn MDX:

SELECT

{ [Measures].[Total Amount] } ON COLUMNS,

FILTER(

NONEMPTY(

[Dim Guest].[Guest ID].[Guest ID].MEMBERS \* [Dim Guest].[First Name].[First Name].MEMBERS \* [Dim Guest].[Last Name].[Last Name].MEMBERS,

[Measures].[Total Amount]

),

[Measures].[Total Amount] > 1000

) ON ROWS

FROM [Hotel]

WHERE ([Dim Guest].[Country].&[Switzerland]);

* Kết quả:

A screenshot of a computer

Description automatically generated

### Số lượng khách nữ từ Thụy Sĩ (Switzerland) với tổng số tiền chi tiêu lớn hơn 1000

* Sử dụng Dice
* Câu lệnh truy vấn MDX:

SELECT

{ [Measures].[Total Amount] } ON COLUMNS,

FILTER(

NONEMPTY(

[Dim Guest].[Guest ID].[Guest ID].MEMBERS \* [Dim Guest].[First Name].[First Name].MEMBERS \* [Dim Guest].[Last Name].[Last Name].MEMBERS,

[Measures].[Total Amount]

),

[Measures].[Total Amount] > 1000

) ON ROWS

FROM [Hotel]

WHERE ([Dim Guest].[Country].&[Switzerland], [Dim Guest].[Gender].&[Female]);

* Kết quả:

A screenshot of a computer

Description automatically generated

### Số lượng đặt phòng qua các nguồn đặt phòng qua các năm

* Sử dụng Roll up
* Câu lệnh truy vấn MDX:

SELECT NON EMPTY [Measures].[Fact Count] ON COLUMNS,

NON EMPTY CROSSJOIN ([Dim Reservation].[Reservation Source].children, [Checkin Date].[Nam].children) ON ROWS

FROM [Hotel];

* Kết quả:

A screenshot of a computer

Description automatically generated

### Thông tin khách hàng nam ở phòng loại ‘Deluxe’, chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 740

* Sử dụng Dice
* Câu lệnh truy vấn MDX:

SELECT

{ [Measures].[Total Amount] } ON COLUMNS,

FILTER(

NONEMPTY(

[Dim Guest].[Guest ID].[Guest ID].MEMBERS \* [Dim Guest].[First Name].[First Name].MEMBERS \* [Dim Guest].[Last Name].[Last Name].MEMBERS,

[Measures].[Total Amount]

),

[Measures].[Total Amount] > 740

) ON ROWS

FROM [Hotel]

WHERE ([Dim Room].[Room Type].&[Deluxe], [Dim Guest].[Gender].&[Male]);

* Kết quả:

A screenshot of a number

Description automatically generated

### Tìm các khách hàng đặt phòng có số lượng đêm ở lớn hơn 4 trong quý 4/2023

* Tạo Name set [SoDemLonHon4]: Lọc ra các Guest ID có số lượng đêm ở lớn hơn 4

A screenshot of a computer

Description automatically generated

* Câu lệnh truy vấn MDX:

SELECT NON EMPTY { [Measures].[Total Nights] } ON COLUMNS,

NON EMPTY { ([Dim Guest].[Guest ID].[Guest ID].ALLMEMBERS \* [Dim Guest].[First Name].[First Name].ALLMEMBERS \* [Dim Guest].[Last Name].[Last Name].ALLMEMBERS ) } ON ROWS

FROM ( SELECT ( { [Checkin Date].[date].[Nam].&[2023], [Checkin Date].[date].[Quy].&[4]&[2023] } ) ON COLUMNS

FROM ( SELECT ( [SoDemLonHon4] ) ON COLUMNS FROM [Hotel]))

WHERE ( [Checkin Date].[date].CurrentMember );

* Kết quả:

A screenshot of a computer

Description automatically generated

### Mỗi quý trong năm 2023, cho biết tháng nào có tổng doanh thu cao nhất

* Câu lệnh truy vấn MDX:

SELECT [Measures].[Total Amount] ON COLUMNS,

NON EMPTY GENERATE(

[Checkin Date].[Quy].children,

TOPCOUNT([Checkin Date].[Quy].CURRENTMEMBER \* [Checkin Date].[Thang].children,1,[Measures].[Total Amount])

) ON ROWS

FROM [Hotel]

WHERE [Checkin Date].[Nam].&[2023];

* Kết quả:

A screenshot of a computer

Description automatically generated

### Doanh thu của loại phòng ‘Deluxe’ qua từng năm

* Câu lệnh truy vấn MDX:

SELECT NON EMPTY [Checkin Date].[Nam].children ON ROWS, [Measures].[Total Amount] ON COLUMNS

FROM [Hotel]

WHERE [Dim Room].[Room Type].&[Deluxe];

* Kết quả:

A screenshot of a computer

Description automatically generated

### Phần trăm số lượng đặt phòng có sử dụng bữa sáng

* Tạo Calculate member:
* [TotalBreakfast]: Tính tổng số lượng đơn đặt phòng có sử dụng bữa sáng

A screenshot of a computer

Description automatically generated

* [PercentBreakfast]: Tính phần trăm đơn đặt phòng có sử dụng bữa sáng

A screenshot of a computer

Description automatically generated

* Câu lệnh truy vấn MDX:

WITH

MEMBER [Measures].[TotalBreakfast] AS

([Fact].[Breakfast].[Yes], [Measures].[Fact Count])

MEMBER [Measures].[PercentBreakfast] AS

([Measures].[TotalBreakfast] / [Measures].[Fact Count]) \* 100,

FORMAT\_STRING = "0.00"

SELECT

{[Measures].[TotalBreakfast], [Measures].[PercentBreakfast]} ON COLUMNS

FROM [Hotel];

* Kết quả:

A screenshot of a computer

Description automatically generated

### Phần trăm doanh thu từ mỗi nguồn đặt phòng năm 2023

* Tạo Calculate member:
* [TotalRevenueByReservation]: Tổng doanh thu một nguồn đặt phòng

A screenshot of a computer

Description automatically generated

* [TotalRevenueAllReservationSources]: Tổng doanh thu tất cả các nguồn đặt phòng

A screenshot of a computer

Description automatically generated

* [PercentRevenueByReservationSource]: Tính phần trăm doanh thu nguồn đặt phòng, lấy tổng doanh thu một nguồn chia cho tổng doanh thu của tất cả các nguồn đặt phòng

A screenshot of a computer

Description automatically generated

* Câu lệnh truy vấn MDX:

WITH

MEMBER [Measures].[TotalRevenueByReservationSource] AS

([Dim Reservation].[Reservation Source].CurrentMember, [Measures].[Total Amount])

MEMBER [Measures].[TotalRevenueAllReservationSources] AS

SUM(

[Dim Reservation].[Reservation Source].[Reservation Source].MEMBERS,

[Measures].[Total Amount]

)

MEMBER [Measures].[PercentRevenueByReservationSource] AS

([Measures].[TotalRevenueByReservationSource] / [Measures].[TotalRevenueAllReservationSources]) \* 100,

FORMAT\_STRING = "0.00"

SELECT

{[Measures].[TotalRevenueByReservationSource], [Measures].[PercentRevenueByReservationSource]} ON COLUMNS,

NON EMPTY [Dim Reservation].[Reservation Source].children ON ROWS

FROM [Hotel]

WHERE [Checkin Date].[Nam].&[2023];

* Kết quả:

A screenshot of a computer

Description automatically generated

### Top 3 doanh thu theo quốc gia từ các đơn đặt phòng có sử dụng dịch vụ đưa đón sân bay

* Tạo Calculate member [RevenueByAirport]: Tính doanh thu các đơn đặt phòng có sử dụng dịch vụ đưa đón sân bay

A screenshot of a computer

Description automatically generated

* Câu lệnh truy vấn MDX:

WITH

MEMBER [Measures].[RevenueByAirport] AS

([Dim Guest].[Country].CurrentMember, [Fact].[Airport].[Yes], [Measures].[Total Amount])

SELECT

{[Measures].[RevenueByAirport]} ON COLUMNS,

TOPCOUNT(

NONEMPTY([Dim Guest].[Country].Children, [Measures].[RevenueByAirport]),

3,

[Measures].[RevenueByAirport]

) ON ROWS

FROM [Hotel]

WHERE ([Checkin Date].[Nam].&[2023] );

* Kết quả:

A screenshot of a computer

Description automatically generated

### Số lượng đặt phòng của từng loại phòng theo mỗi nguồn đặt phòng

* Câu lệnh truy vấn MDX:

SELECT

NON EMPTY {[Measures].[Fact Count]} ON COLUMNS,

NON EMPTY CROSSJOIN(

[Dim Room].[Room Type].[Room Type].Members,

[Dim Reservation].[Reservation Source].[Reservation Source].Members

) ON ROWS

FROM [Hotel];

* Kết quả:

A screenshot of a hotel

Description automatically generated

### Tổng số lượng đặt phòng theo từng loại phòng và tầng

* Câu lệnh truy vấn MDX:

SELECT

{[Measures].[Fact Count]} ON COLUMNS,

NON EMPTY CROSSJOIN(

[Dim Room].[Room Type].[Room Type].Members,

[Dim Room].[Floor Number].[Floor Number].Members

) ON ROWS

FROM [Hotel];

* Kết quả:

A screenshot of a computer

Description automatically generated

### Trung bình số đêm ở theo từng tháng trong năm 2023

* Tạo Calculate member [AverageNightsStay]: Tính trung bình số đêm ở

A screenshot of a computer

Description automatically generated

* Câu lệnh truy vấn MDX:

WITH

MEMBER [Measures].[AverageNightsStay] AS

IIF(

[Measures].[Fact Count] = 0,

NULL,

[Measures].[Total Nights] / [Measures].[Fact Count]

),

FORMAT\_STRING = "0.00"

SELECT

{[Measures].[AverageNightsStay]} ON COLUMNS,

NON EMPTY [Checkin Date].[Thang].[Thang].Members ON ROWS

FROM [Hotel]

WHERE ([Checkin Date].[Nam].&[2023]);

* Kết quả:

A screenshot of a computer

Description automatically generated

### Tổng doanh thu theo loại phòng

* Câu lệnh truy vấn MDX:

SELECT

{[Measures].[Total Amount]} ON COLUMNS,

NON EMPTY [Dim Room].[Room Type].[Room Type].Members ON ROWS

FROM [Hotel];

* Kết quả:

A screenshot of a computer

Description automatically generated

### Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng

* Câu lệnh truy vấn MDX:

SELECT [Measures].[Fact Count] ON COLUMNS,

NON EMPTY Generate(

[Dim Reservation].[Reservation Source].children,

TopCount(

[Dim Reservation].[Reservation Source].currentmember \* [Dim Room].[Room Type].children,

2,

[Measures].[Fact Count]

)

) ON ROWS

FROM [Hotel];

* Kết quả:

A screenshot of a computer

Description automatically generated

### Liệt kê top 2 loại phòng có doanh thu nhiều nhất trong cả hai năm 2022 và năm 2023

* Câu lệnh truy vấn MDX:
* Cách 1: Dùng hàm Intersect

SELECT {

([Measures].[Total Amount],[Checkin Date].[Nam].&[2022]),

([Measures].[Total Amount],[Checkin Date].[Nam].&[2023])

} ON COLUMNS,

INTERSECT(

{TOPCOUNT([Dim Room].[Room Type].children,

2,

([Measures].[Total Amount],[Checkin Date].[Nam].&[2022])

)},

{TOPCOUNT([Dim Room].[Room Type].children,

2,

([Measures].[Total Amount],[Checkin Date].[Nam].&[2023])

)}

) ON ROWS

FROM [Hotel];

* Cách 2: Tạo các Member mới và dùng hàm Intersect

WITH

-- Tính tổng doanh thu cho năm 2022

MEMBER [Measures].[Total Amount 2022] AS

([Measures].[Total Amount], [Checkin Date].[Nam].&[2022])

-- Tính tổng doanh thu cho năm 2023

MEMBER [Measures].[Total Amount 2023] AS

([Measures].[Total Amount], [Checkin Date].[Nam].&[2023])

-- Tạo tập hợp các loại phòng có doanh thu cao nhất trong cả hai năm

SET [Top Rooms 2022] AS

TOPCOUNT (

[Dim Room].[Room Type].children,

2,

[Measures].[Total Amount 2022]

)

SET [Top Rooms 2023] AS

TOPCOUNT (

[Dim Room].[Room Type].children,

2,

[Measures].[Total Amount 2023]

)

SET [Intersected Top Rooms] AS

Intersect([Top Rooms 2022], [Top Rooms 2023])

SELECT

{[Measures].[Total Amount 2022], [Measures].[Total Amount 2023]} ON COLUMNS,

[Intersected Top Rooms] ON ROWS

FROM [Hotel];

* Kết quả:

A screenshot of a computer

Description automatically generated

## Thực hiện trên Excel

Chọn Excel để phân tích Pivot trong Excel

A screenshot of a computer

Description automatically generated

### Tổng doanh thu theo loại phòng qua các năm (group by)

* Kéo [Total Amount] trong ‘Measures’ vào Values
* Kéo [Room Type] trong ‘Dim Room’ vào Rows
* Kéo [Nam] trong ‘Checkin Date’ vào Columns

A screenshot of a computer

Description automatically generated

### Tổng số đơn đặt phòng theo nguồn đặt phòng (group by)

* Kéo [Fact Count] trong ‘Measures’ vào Values
* Kéo [Reservation Source] trong ‘Dim Reservation’ vào Rows

A screenshot of a computer

Description automatically generated

### Số lượng khách từ Thụy Sĩ (Switzerland) , chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 1000 (table)

* Kéo [Guest ID] từ Dimension ‘Dim Guest’ vào Rows
* Kéo [Total Amount] từ Measures vào Values
* Kéo [Country] từ Dimension ‘Dim Guest’ vào Filters

A screenshot of a computer

Description automatically generated

* Lọc để chỉ hiển thị khách hàng từ Thụy Sĩ bằng cách chọn Switzerland trong Filter Country

A screenshot of a computer

Description automatically generated

* Click vào mũi tên thả xuống bên cạnh Row Labels (Guest ID)
* Chọn Value Filters > Greater Than....

A screenshot of a computer

Description automatically generated

* Nhập 400 để lọc các giá trị Total Amount lớn hơn 1000 và “OK”

A screenshot of a computer

Description automatically generated

* Kết quả:

A screenshot of a computer

Description automatically generated

### Tạo ma trận hiển thị số lượng khách hàng theo loại phòng và nguồn đặt phòng (matrix)

* Kéo [Room Type] trong Dim Room vào Columns
* Kéo [Reservation Source] trong Dim Reservation vào Rows
* Kéo [Fact Count] trong Measures vào Values

A screenshot of a computer

Description automatically generated

### Tạo biểu đồ hiển thị tổng doanh thu theo tháng năm 2023 (chart)

* Kéo [Nam] trong ‘Checkin Date’ vào Filters
* Kéo [Thang] trong ‘Checkin Date’ vào Columns
* Kéo [Total Amount] trong ‘Measures’ vào Values
* Chọn 2023 để lọc doanh thu vào năm 2023

A screenshot of a computer

Description automatically generated

* Kết quả:

A graph of different colored bars

Description automatically generated

### Tạo biểu đồ hiển thị top 2 doanh thu theo tháng năm 2023 (chart)

* Trong mục [Thang] à chọn Value Filters à chọn Top 10

A screenshot of a computer

Description automatically generated

* Chỉnh sửa thành top 2 theo doanh thu (Total Amount)

A screenshot of a computer

Description automatically generated

* Kết quả:

A graph with blue and orange bars

Description automatically generated

### Số lượng đặt phòng theo từng tháng và từng loại phòng trong năm 2023

* Kéo [Room Type] trong ‘Dim Room’ vào Columns
* Kéo [Thang] trong ‘Checkin Date’ vào Rows
* Kéo [Fact Count] trong Measures vào Values
* Kéo [Nam] trong ‘Checkin Date’ vào Filters

A screenshot of a computer

Description automatically generated

* Lọc dữ liệu của Nam = 2023

A screenshot of a computer

Description automatically generated

* Kết quả:

A screenshot of a data

Description automatically generated

### Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng

* Kéo [Room Type] trong ‘Dim Room’ vào Columns
* Kéo [Reservation Source] trong ‘Dim Reservation’ vào Rows
* Kéo [Fact Count] trong Measures vào Values

A screenshot of a computer

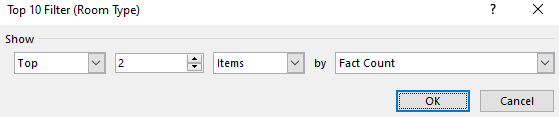
Description automatically generated

* Ở Column labels, chọn Sort 🡪 Value Filters 🡪 Top 10

A screenshot of a computer screen

Description automatically generated

* Chọn Top 2 theo số lượt thuê để liệt kê top 2 loại phòng



* Kết quả:

A screenshot of a computer

Description automatically generated

## Thực hiện trên Power BI

**Bước 1:** Đưa dữ liệu vào Power BI: Trong thẻ Home à chọn Get data à Analysis Services

A screenshot of a computer

Description automatically generated

**Bước 2**: Điền tên Server chứa Database

A screenshot of a computer

Description automatically generated

**Bước 3:** Chọn Cube chứa bảng Dim và Fact dùng để trực quan dữ liệu

A screenshot of a computer

Description automatically generated

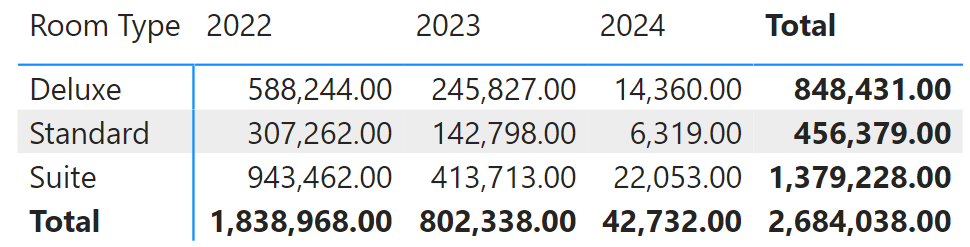
### Tổng doanh thu theo loại phòng qua các năm

* Column Chart

A graph of different colored columns

Description automatically generated

* Table



* Line chart

A graph of different colored lines

Description automatically generated

Từ kết quả trên, ta có:

* Cả 3 năm (2022, 2023, 2024) thì loại phòng Suite có doanh thu cao nhất trong các loại phòng
* Loại phòng Standard có doanh thu thấp nhất ở cả 3 năm

### Tổng số lượng khách hàng theo nguồn đặt phòng

* Sử dụng Calculate member [SumAdultsChildren] đã tạo trước trong SSAS
* Column chart

A graph of a bar chart

Description automatically generated with medium confidence

* Table

A screenshot of a computer

Description automatically generated

* Pie chart

A pie chart with numbers and text

Description automatically generated

### Số lượng khách từ Thụy Sĩ (Switzerland), chỉ hiển thị khách hàng có tổng số tiền chi tiêu lớn hơn 1000

* Filter Total Amount > 1000

A screenshot of a computer

Description automatically generated

* Filter Country = ‘Switzerland’

A screenshot of a computer

Description automatically generated

* Kết quả:

A screenshot of a table

Description automatically generated

### Tạo ma trận hiển thị số lượng đặt phòng theo loại phòng và nguồn đặt phòng

* Matrix

A table with numbers and text

Description automatically generated

* Line chart

A graph of different types of service

Description automatically generated

### Tạo biểu đồ hiển thị tổng doanh thu theo tháng năm 2023

* Filter [Nam] = 2023

A screenshot of a computer

Description automatically generated

* Bar chart

A graph of blue bars

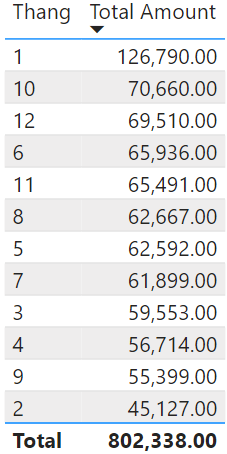
Description automatically generated

* Pie chart

A colorful pie chart with numbers and text

Description automatically generated

* Table



Từ kết quả trên, ta có:

* Tháng 1/2023 có doanh thu cao nhất tất cả các tháng trong năm
* Tháng 2 có doanh thu thấp nhất năm 2023

### Tạo biểu đồ hiển thị top 2 doanh thu theo tháng năm 2023

* Filter Top 2 doanh thu (Total Amount) theo tháng à Apply filter

A screenshot of a computer

Description automatically generated

* Table

A screenshot of a computer screen

Description automatically generated

* Bar chart

A graph with blue rectangles

Description automatically generated

### Số lượng đặt phòng theo từng tháng và từng loại phòng trong năm 2023

* Matrix

A screenshot of a number

Description automatically generated

* Column chart

A graph of different colored bars

Description automatically generated

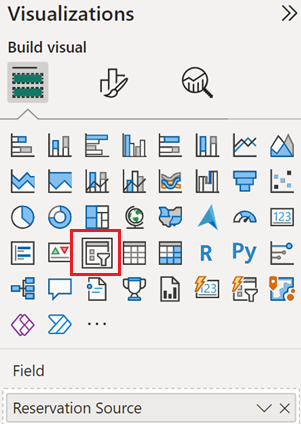
### Liệt kê top 2 loại phòng được thuê nhiều nhất theo nguồn đặt phòng

* Lọc ra top 2 loại phòng theo số lượt thuê

A screenshot of a computer

Description automatically generated

* Chọn Slicer trong Visualizations để tạo bộ lọc dữ liệu cần lọc (theo Reservation Source)



* Chọn Reservation Source là Booking Platform



* Chọn Reservation Source là Phone

A screenshot of a survey

Description automatically generated

* Chọn Reservation Source là Walk-in

A screenshot of a computer

Description automatically generated

* Chọn Reservation Source là Website

A screenshot of a survey

Description automatically generated

Từ kết quả trên, ta có:

* Reservation Source là Booking Platform và Walk-in sẽ có loại phòng *Deluxe và Suite* được thuê nhiều nhất, trong đó cả hai nguồn đặt chỗ đều có xu hướng đặt loại phòng Deluxe nhiều hơn so với Suite
* Reservation Source là Phone và Website sẽ có loại phòng *Standard và Suite* được thuê nhiều nhất, trong đó:

+ Với nguồn đặt chỗ Phone, số lượng thuê phòng loại Standard nhiều hơn so với Suite

+ Với nguồn đặt chỗ Website, số lượng thuê phòng loại Suite nhiều hơn so với Standard

# QUÁ TRÌNH DATA MINING

## Tổng quan đề tài

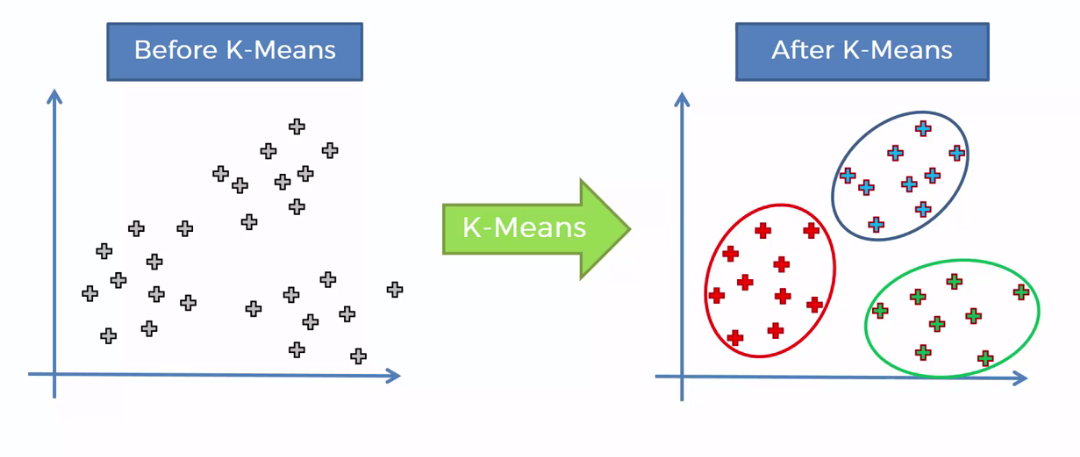
**Đề tài: Phân tích và phân cụm khách hàng dựa trên dữ liệu đặt phòng khách sạn.**

Mục tiêu của dự án là áp dụng phân tích RFM (Recency, Frequency và Monetary) và các phương pháp phân cụm để hiểu và phân loại khách hàng từ các quốc gia khác nhau dựa trên hành vi đặt phòng khách sạn. Qua đó, xác định các nhóm khách hàng tiềm năng từ từng quốc gia và phát triển chiến lược tiếp thị hiệu quả hơn để tối đa hóa giá trị từ từng nhóm khách hàng.

## Lý thuyết mô hình phân cụm

### K-Means

K-means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát(tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm k-means là phân chia 1 bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau.



Thuật toán phân cụm k-means thường được sử dụng trong các ứng dụng cỗ máy tìm kiếm, phân đoạn khách hàng, thống kê dữ liệu,…

**Thuật toán k-means có thể được chia thành các bước như sau:**

* Bước 1: Tạo các trung tâm ngẫu nhiên

****

* Bước 2: Gán các điểm dữ liệu vào các cụm

****

Với mỗi điểm dữ liệu, ta sẽ tính khoảng cách của nó tới các trung tâm và sẽ gán chúng vào trung tâm gần nhất. Tập hợp các điểm được gán vào cùng 1 trung tâm sẽ tạo thành cụm.

* Bước 3:Cập nhật trung tâm

Với mỗi cụm đã tìm được ở bước 2, trung tâm mới sẽ là trung bình cộng của các điểm dữ liệu trong cụm đó.

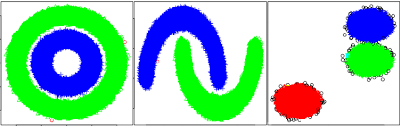
**A black background with a black square

Description automatically generated with medium confidence**

* Thuật toán sẽ lặp lại các bước trên cho tới khi đạt được kết quả chấp nhận được.

### DBSCAN (Density-Based Clustering)

Giải thuật DBSCAN (Density Based Spatial Clustering of Application with Noise) được Ester, Kriegel và Sander đề xuất năm 1996 khi nghiên cứu các thuật toán gom cụm dữ liệu không gian dựa trên định nghĩa cụm là tập tối đa các điểm liên thông về mật độ. Giải thuật DBSCAN phát hiện các cụm có hình dạng tùy ý, khả năng phát hiện nhiễu tốt. DBSCAN thực hiện tốt trên không gian nhiều chiều; thích hợp với cơ sở dữ liệu có mật độ phân bố dày đặc kể cả có phần tử nhiễu.



**Nguyên lý hoạt động của DBSCAN:**

DBSCAN dựa trên ý tưởng rằng một cụm dữ liệu là một vùng mật độ cao được tách biệt bởi các vùng mật độ thấp. Cụ thể, nó xác định các cụm dựa trên hai tham số chính:

* Epsilon (ε): Bán kính của vùng lân cận xung quanh một điểm dữ liệu.
* MinPts: Số lượng điểm tối thiểu cần thiết trong vùng lân cận của một điểm để xác định nó là một điểm lõi.

**Phân loại dạng điểm trong DBSCAN:**

A diagram of a diagram

Description automatically generated

Căn cứ vào vị trí của các điểm dữ liệu so với cụm chúng ta có thể chia chúng thành ba loại:

* Điểm lõi (Core Point): Một điểm được gọi là điểm lõi nếu trong vùng lân cận ε của nó có ít nhất MinPts điểm.
* Điểm biên (Border Point): Một điểm được gọi là điểm biên nếu nó nằm trong vùng lân cận của một điểm lõi nhưng không phải là một điểm lõi.
* Điểm nhiễu (Noise Point): Một điểm được gọi là điểm nhiễu nếu nó không phải là điểm lõi và không nằm trong vùng lân cận của bất kỳ điểm lõi nào.

### Agglomerative Clustering

Agglomerative Clustering, hay còn gọi là phân cụm tập hợp, là một thuật toán phân cụm phân cấp từ dưới lên. Thuật toán này bắt đầu với mỗi điểm dữ liệu là một cụm riêng lẻ, sau đó lặp lại quá trình hợp nhất các cụm nhỏ thành các cụm lớn hơn cho đến khi đạt được một tiêu chí dừng cụ thể (ví dụ: số lượng cụm mong muốn).

**Quá trình của Agglomerative Clustering bao gồm các bước:**

A collage of diagrams

Description automatically generated

* Bắt đầu với n cụm, mỗi cụm chỉ chứa một điểm dữ liệu.
* Tính toán ma trận khoảng cách giữa các cụm.
* Tìm cặp cụm gần nhau nhất (có khoảng cách nhỏ nhất) và hợp nhất chúng thành một cụm mới.
* Cập nhật ma trận khoảng cách bằng cách tính khoảng cách giữa cụm mới với các cụm còn lại.
* Lặp lại bước 3-4 cho đến khi đạt được số lượng cụm mong muốn hoặc một tiêu chí dừng khác.

**Các định nghĩa khoảng cách giữa các cụm, bao gồm:**

* Khoảng cách liên kết tối thiểu (single linkage): Khoảng cách giữa hai cụm là khoảng cách nhỏ nhất giữa các điểm trong hai cụm.
* Khoảng cách liên kết tối đa (complete linkage): Khoảng cách giữa hai cụm là khoảng cách lớn nhất giữa các điểm trong hai cụm.
* Khoảng cách liên kết trung bình (average linkage): Khoảng cách giữa hai cụm là trung bình khoảng cách giữa các điểm trong hai cụm.

Agglomerative Clustering có thể tạo ra các cụm có hình dạng và kích thước khác nhau, phù hợp với nhiều loại dữ liệu. Phân tích dataset gốc

### Thống kê mô tả

Tính toán đại lượng thống kê mô tả: Count, Min, Max, Mean, Mode,… trên tập dữ liệu.

* Thêm thư viện và đọc file excel của tập dữ liệu, kiểm tra 5 dòng đầu tiên dữ liệu.

A screenshot of a computer

Description automatically generated

Tập dữ liệu đặt phòng khách sạn với các thông tin mã đặt phòng, thông tin khách đặt phòng, thời gian đặt/giao/trả phòng, thông tin dịch vụ đi kèm, thông tin về phòng đã đặt.

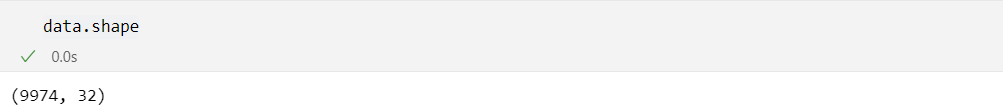
* Kiểm tra kiểu dữ liệu.

A screenshot of a computer

Description automatically generated

Tập dữ liệu bao gồm các kiểu dữ liệu như: datetime, int và object.

* Kiểm tra số lượng dòng và số lượng thuộc tính của tập dữ liệu cho thấy tập dữ liệu gồm 9974 dòng và 32 thuộc tính.



* Sử dụng phương thức describe() để tính toán thống kê mô tả tất cả các cột có kiểu dữ liệu là có dạng numberic trong toàn tập dữ liệu.

A screenshot of a computer

Description automatically generated

Qua bảng mô tả dữ liệu:

* Các ngày sinh (Birthdate) trải dài từ năm 1943 đến 2006, điều này cho thấy sự đa dạng về độ tuổi của khách hàng.
* Số người lớn (Adults) trung bình là 2.50, phần lớn các đặt phòng dành cho 2 hoặc 4 người lớn.
* Số trẻ em (Children) đi cùng gia đình với trung bình là 0.99, với nhiều đặt phòng không có trẻ em hoặc chỉ có 1-2 trẻ em.
* Tổng số đêm (Total Nights) trung bình là 2.83, phần lớn các đặt phòng có thời gian lưu trú từ 2-3 đêm.
* Tổng số tiền phải trả (Total Amount) trung bình là 408.76, với số tiền cho mỗi lần chi trả thấp nhất là 60$, số tiền lớn nhất lên đến 1250$, cho thấy sự đa dạng trong mức chi tiêu của khách hàng.

## Tiền xử lý dữ liệu

### Xử lý giá trị thiếu

* Kiểm tra dữ liệu rỗng trong tập, tiến hành xử lý bằng cách thay thế giá trị trung bình, trung vị,… hoặc loại bỏ hoàn toàn dòng chứa dữ liệu rỗng.

A screenshot of a computer

Description automatically generated

Tập dữ liệu hoàn toàn đầy đủ, không chứa giá trị nul trong tập.

### Tính toán các chỉ số RFM

A screenshot of a computer

Description automatically generated

Tính toán các giá trị của mô hình phân tích RFM, trong bài thực hiện tính toán RFM theo từng quốc gia.

* Recency: khoảng thời gian giữ thời gian đặt phòng khách sạn đến ngày đặt phòng cuối cùng ghi nhận trong tập dữ liệu.
* Frequency: Số lần đặt phòng.
* Monetary: Tổng số tiền đã chi tiêu.

### Khám phá và xử lý dữ liệu RFM

* Kiểm tra phân phối dữ liệu

A graph of a number of data

Description automatically generated with medium confidence

Phân phối dữ liệu lệch phải, nếu xử dụng dữ liệu cho việc chạy model sẽ không mang lại hiệu quả cao 🡺 tiến hành chuyển đổi dữ liệu.

* Dùng hàm analyze\_skewness() kiểm tra xem áp dụng phương pháp transform phù hợp với mô hình sau đó thực hiện transform.

A white screen with a black text

Description automatically generated with medium confidence

* Tiến hành kiểm tra trên cột Frequency cho thấy chuyển đổi dữ liệu theo kiểu boxcox là tốt nhất (do độ lệch thấp nhất là 0.02).

A screenshot of a graph

Description automatically generated

* Tiến hành kiểm tra trên cột Recency cho thấy chuyển đổi dữ liệu theo kiểu boxcox là tốt nhất (do độ lệch thấp nhất là 0.14).

A group of blue lines

Description automatically generated

* Tương tự tiến hành kiểm tra trên cột MonetaryValue cho thấy chuyển đổi dữ liệu theo kiểu boxcox là tốt nhất (do độ lệch thấp nhất là -0.01)

A screenshot of a graph

Description automatically generated

* Transform dữ liệu theo các kỹ thuật phù hợp.

A computer screen shot of a computer code

Description automatically generated

### Chuẩn hóa dữ liệu

Sử dụng phương pháp StandScaler là một bộ biến đổi (transformer) được sử dụng để chuẩn hóa dữ liệu bằng cách loại bỏ giá trị trung bình và chia cho độ lệch chuẩn của mỗi đặc trưng.

A white rectangular object with a black border

Description automatically generated with medium confidence

## Xây dựng mô hình phân cụm

### Xác định số lượng cụm (k) tối ưu bằng kỹ thuật Elbow

Kỹ thuật Elbow là một phương pháp phổ biến để xác định số lượng cụm (k) tối ưu trong các thuật toán phân cụm. Cách thức hoạt động của kỹ thuật Elbow:

* Chạy thuật toán phân cụm (ví dụ K-Means) với các giá trị k khác nhau, thường từ 1 đến 10 hoặc nhiều hơn tùy thuộc vào dữ liệu.
* Tính toán một chỉ số đánh giá chất lượng phân cụm, thường là Inertia (tổng bình phương khoảng cách các điểm đến tâm cụm gần nhất).
* Vẽ một đồ thị biểu diễn giá trị Inertia theo số lượng cụm k.
* Tìm "khuỷu tay" (elbow) trên đồ thị, tức là điểm mà sau đó giá trị Inertia giảm chậm hơn. Điểm này được coi là số lượng cụm tối ưu.

A graph of a number of clusters

Description automatically generated

🡺Qua biểu đồ: elbow xuất hiện ở k = 4, vì sau đó inertia giảm chậm hơn. Do đó, số lượng cụm tối ưu là 4.

### Xây dựng các mô hình

Xây dựng mô hình K-Means, DBSCAN và Agglomerative Clustering. K-Means và Agglomerative Clustering phân chia theo số cụm là 4 như đã xác định ở trên.

A white background with a black and blue object

Description automatically generated with medium confidence

### Đánh giá mô hình và trực quan hóa kết quả

Sử dụng Silhouette Score xác định chất lượng của việc phân cụm. Silhouette Score tính toán độ tương đồng của một điểm dữ liệu với các điểm dữ liệu trong cùng cụm (cohesion) so với các điểm dữ liệu trong các cụm khác (separation).

A group of graphs with different colored dots

Description automatically generated

* K-means: 0.37, các cụm được tạo ra bởi K-means có độ tương đồng trong nội bộ cụm và độ phân biệt giữa các cụm ở mức trung bình.
* DBSCAN: 0.31, hơn so với K-means 🡺 độ phân biệt giữa các cụm được DBSCAN tạo ra không tốt bằng K-means.
* Agglomerative Clustering: 0.40, giá trị cao nhất trong 3 phương pháp.

🡺 Dựa trên các giá trị Silhouette Score, Agglomerative Clustering cho kết quả phân cụm tốt nhất, tiếp đến là K-means, và DBSCAN cho kết quả phân cụm kém hơn. Sử dụng thuật toán Agglomerative Clustering tiến hành phân tích phân cụm khách hàng.

### Phân tích & phân cụm khách hàng

* Sử dụng thuật toán Agglomerative Clustering tiến hành gắn nhãn phân cụm cho các dòng dữ liệu:

A white rectangular object with a black border

Description automatically generated with medium confidence

* Trực quan hóa kết quả phân cụm bằng các cặp thuộc tính khác nhau:

A group of graphs with different colored dots

Description automatically generated

* Tính toán giá trị trung bình RFM tìm ra các đặc trưng của từng nhóm:

A screenshot of a graph

Description automatically generated

**Nhận xét:**

* Cluster 0: Khách hàng VIP có giá trị cao nhất với tần suất và giá trị chi tiêu rất cao.
* Cluster 1: Khách hàng trung thành với tần suất và giá trị chi tiêu cao.
* Cluster 2: Khách hàng có giá trị cao nhưng không hoạt động gần đây.
* Cluster 3: Khách hàng thường xuyên với giá trị chi tiêu trung bình.

Đặt tên các phân loại dựa theo đặc điểm chung:

A screenshot of a computer

Description automatically generated

* Vẽ biểu đồ cho thấy tỉ lệ phân phối của từng nhóm khách hành trên tập dữ liệu:

A pie chart with text on it

Description automatically generated

🡪Hiển thị các quốc gia thuộc từng nhóm:

A close up of a text

Description automatically generated

* Các chiến lược phù hợp cho từng nhóm khách hàng

****High-Value Inactive Customers (32%):****

* + Chiếm tỷ lệ lớn nhất trong các cụm khách hàng. Các quốc gia trong cụm này có giá trị chi tiêu cao nhưng tần suất đặt phòng giảm sút.
  + **Mục tiêu:** Kích thích hoạt động trở lại của khách hàng từ các quốc gia này bằng cách: gửi các chiến dịch tiếp thị và ưu đãi đặc biệt, khảo sát để hiểu nguyên nhân giảm sút tần suất đặt phòng và cải thiện dịch vụ.

**VIP Customers (30%):**

* + Chiếm tỷ cao thứ 2 lệ gần như tương đương với nhóm High-Value Inactive. Các quốc gia này có khách hàng có giá trị chi tiêu cao và thường xuyên đặt phòng.
  + **Mục tiêu:** cần duy trì và nâng cao trải nghiệm để giữ chân nhóm khách hàng này.
  + **Chiến lược:**
    - Cung cấp các chương trình ưu đãi đặc biệt và dịch vụ cao cấp.
    - Tạo các chương trình khách hàng thân thiết và tích điểm.
    - Thường xuyên gửi lời cảm ơn và đánh giá dịch vụ từ khách hàng.

**Regular Customers (25%):**

* + Chiếm tỷ lệ khá đáng kể. Các quốc gia này có khách hàng thường xuyên đặt phòng nhưng với giá trị chi tiêu trung bình.
  + **Mục tiêu:** Tăng tần suất đặt phòng và giá trị chi tiêu.
  + **Chiến lược:**
    - Cung cấp các chương trình khuyến mãi định kỳ.
    - Tăng cường chăm sóc khách hàng để nâng cao mức độ hài lòng.
    - Tạo các gói dịch vụ hấp dẫn để khách hàng chi tiêu nhiều hơn.

**Loyal Customers (13%):**

* + Chiếm tỷ lệ nhỏ nhất trong các cụm khách hàng. Nhóm khách hàng trung thành, thường xuyên đặt phòng với giá trị chi tiêu cao.
  + **Mục tiêu:** Khuyến khích tiếp tục đặt phòng thường xuyên và biến họ thành VIP Customers.
  + **Chiến lược:** khuyến khích họ tiếp tục đặt phòng bằng các chương trình tích điểm hoặc giảm giá, tăng cường chăm sóc khách hàng và cá nhân hóa trải nghiệm.