TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO BÀI TẬP LỚN

ĐỀ TÀI: PHÂN TÍCH NHU CẦU ĐẶT CHỖ KHÁCH SẠN

**Môn học: Phân tích dữ liệu**

**GVHD: ThS. Hồ Hướng Thiên**

**LỚP:**

SINH VIÊN THỰC HIỆN:

**Võ Huỳnh Thanh Phương - 2151013072**

**MỤC LỤC**

[CHƯƠNG 1: GIỚI THIỆU 2](#__RefHeading___Toc318_754522617)

[1. Giới thiệu đề tài 2](#__RefHeading___Toc330_754522617)

[2. Mô tả dữ liệu 2](#__RefHeading___Toc332_754522617)

[CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU 4](#__RefHeading___Toc328_754522617)

[CHƯƠNG 3: KHÁM PHÁ DỮ LIỆU 8](#__RefHeading___Toc326_754522617)

[1. Tổng quan dữ liệu 8](#__RefHeading___Toc346_754522617)

[2. Mối tương quan dữ liệu 11](#__RefHeading___Toc344_754522617)

[CHƯƠNG 4: GOM CỤM K-MEANS 13](#__RefHeading___Toc326_754522617_Copy_1)

[1. Tổng quan dữ liệu 13](#__RefHeading___Toc346_754522617_Copy_1)

[KẾT LUẬN 13](#__RefHeading___Toc324_754522617)

[1. Kết quả đạt được 13](#__RefHeading___Toc338_754522617)

[2. Hạn chế của đề tài 13](#__RefHeading___Toc336_754522617)

[TÀI LIỆU THAM KHẢO 13](#__RefHeading___Toc322_754522617)

[BẢNG PHÂN CÔNG CÔNG VIỆC 14](#__RefHeading___Toc320_754522617)

# CHƯƠNG 1: GIỚI THIỆU

## Giới thiệu đề tài

Ngành du lịch là một ngành trọng điểm của các quốc gia (trong đó có Việt Nam). Trong đó vấn đề quản lý khách sạn là một vấn đề thú vị. Thông qua đề tài, ta sẽ được thấy sự khác biệt của nhu cầu đặt chỗ resort và khách sạn trong thành phố là như thế nào qua hai datasets được cung cấp.

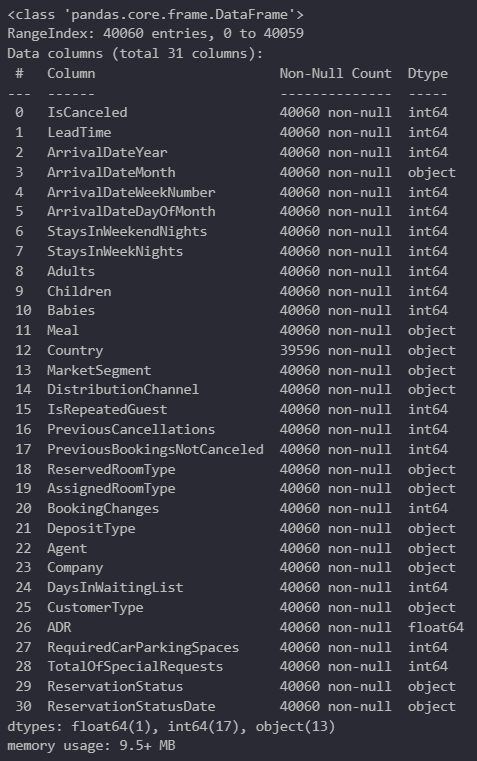
## Mô tả dữ liệu

Báo cáo sử dụng 2 datasets [1] mô tả nhu cầu đặt chỗ khách sạn tên là “H1” và “H2”. Dataset “H1” gồm các thông tin về nhu cầu đặt chỗ một resort, trong khi đó “H2” gồm các thông tin về nhu cầu đặt chỗ một khách sạn trong thành phố. Cả “H1” và “H2” đều lấy thông tin bắt đầu từ ngày 1/7/2015 tới ngày 31/8/2017. Tuy “H1” có 40060 quan sát còn “H2” có 79330 quan sát (1 quan sát là 1 lần đặt chỗ), nhưng các thuộc tính của chúng thì tương tự [1]:

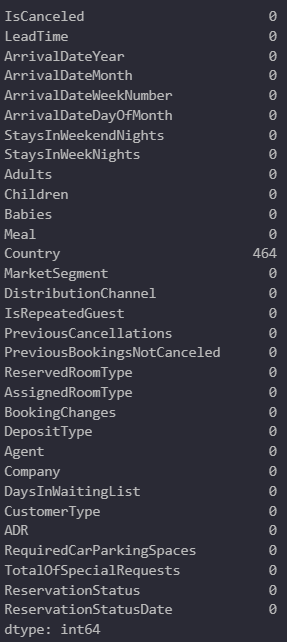
* IsCanceled: thuộc tính chỉ ra việc đặt chỗ có bị hủy (1) hoặc không (0)
* LeadTime: số ngày trôi qua kể từ khi nhập ngày đặt chỗ vào hệ thống quản lý khách sạn (PMS) và ngày hẹn
* ArrivalDateYear: năm diễn ra ngày hẹn
* ArrivalDateMonth: tháng diễn ra ngày hẹn
* ArrivalDateWeekNumber: tuần diễn ra ngày hẹn (trong một năm)
* ArrivalDateDayOfMonth: ngày của tháng diễn ra ngày hẹn
* StaysInWeekendNights: số đêm cuối tuần (thứ bảy hoặc Chúa Nhật) mà khách hàng ở lại hoặc đặt chỗ để ở lại khách sạn
* StaysInWeekNights: số đêm trong tuần (từ thứ hai đến thứ sáu) mà khách hàng ở lại hoặc đặt chỗ để ở lại khách sạn
* Adults: số lượng khách là người lớn tính theo lần đặt chỗ
* Children: số lượng khách là trẻ em tính theo lần đặt chỗ
* Babies: số lượng khách là trẻ sơ sinh của tính theo lần đặt chỗ
* Meal: loại bữa ăn được đặt theo lần đặt chỗ (không đặt, chỉ bữa sáng, bữa sáng và một bữa khác, cả 3 bữa)
* Country: quốc gia của khách đặt chỗ
* MarketSegment: chỉ định phân khúc thị trường (Travel Agent hoặc Tour Operator)
* DistributionChannel: kệnh phân phối đặt chỗ (qua Travel Agent hoặc Tour Operator)
* IsRepeatedGuest: thuộc tính chỉ ra việc đặt chỗ là từ khách quen (1) hoặc không (0)
* PreviousCancellations: số lần đặt chỗ trong quá khứ bị hủy bởi khách hàng trước lần đặt chỗ hiện tại
* PreviousBookingsNotCanceled: số lần đặt chỗ trong quá khứ không bị hủy bởi khách hàng trước lần đặt chỗ hiện tại
* ReservedRoomType: mã của loại phòng đã đặt (không nêu rõ vì nhu cầu ẩn danh của khách sạn)
* AssignedRoomType: mã của phòng được giao cho khách hàng trong thực tế (không nêu rõ vì nhu cầu ẩn danh của khách sạn)
* BookingChanges: số lượng những thay đổi/chỉnh sửa từ khi việc đặt chỗ được lưu vào hệ thống tới khi khách hàng nhận chỗ hoặc hủy việc đặt chỗ
* DepositType: loại đặt cọc (không đặt cọc, đặt cọc không hoàn lại, đặt cọc hoàn lại)
* Agent: số ID của hãng du lịch thực hiện đặt chỗ
* Company: mã định danh của công ty/thực thể đặt chỗ hoặc có trách nhiệm trong việc chi trả cho lần đặt chỗ
* DaysInWaitingList: số ngày lần đặt chỗ được đưa vào hàng đợi trước khi được xác nhận với khách hàng
* CustomerType: loại đặt chỗ (theo hợp đồng, theo nhóm, tạm thời, tạm thời nhưng có liên kết với các lần đặt loại tạm thời khác)
* ADR: giá trung bình hàng ngày đo lường doanh thu cho thuê trung bình kiếm được từ một phòng đã có khách mỗi ngày
* RequiredCarParkingSpaces: số lượng chỗ đậu xe mà khách hàng yêu cầu
* TotalOfSpecialRequests: số lượng yêu cầu đặt biệt mà khách hàng đưa ra
* ReservationStatus: trạng thái mới nhất của lần đặt chỗ (đã huỷ, đã trả phòng, vắng mặt)
* ReservationStatusDate: ngày mà trạng thái mới nhất của lần đặt chỗ được cập nhật.

# CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU

Đầu tiên, ta nhìn qua các thông số dữ liệu của dataset “H1”.



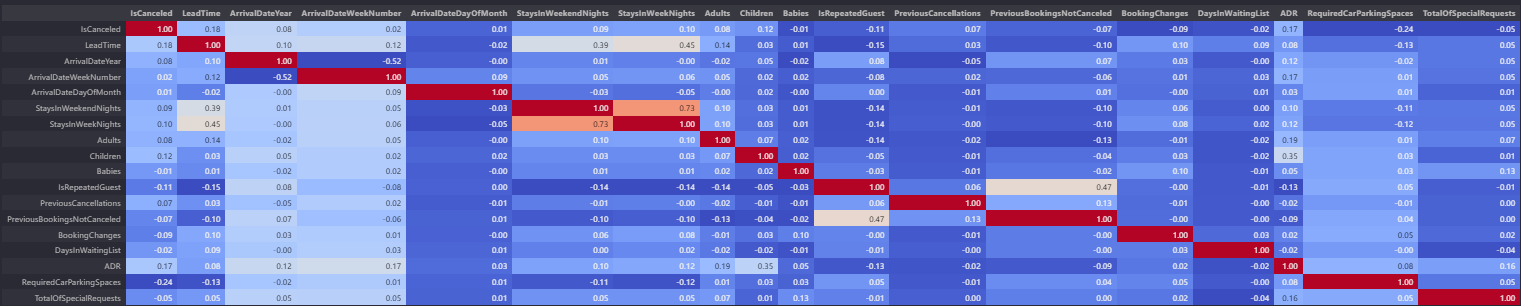
Tiếp đến, ta kiểm tra xem có thuộc tính nào có giá trị rỗng hay không, và đây là kết quả:



Ta thấy rằng, thuộc tính “Country” có 464 giá trị rỗng (chiếm 1%). Ta xóa các quan sát có những dữ liệu bị thiếu vì không ảnh hưởng nhiều đến dữ liệu.

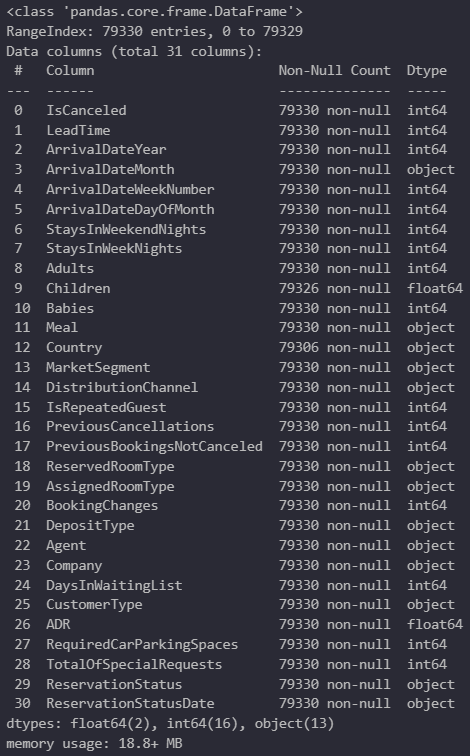
Tiếp đến, ta kiểm tra xem có quan sát nào bị trùng với nhau không. Và kết quả là có 6074 quan sát bị trùng (khoảng 15% dataset “H1”). Ta xóa những quan sát này để không còn giá trị trùng nữa.

Ta kiểm tra xem có thuộc tính nào của “H1” là dư thừa không. Kết quả là không có thuộc tính nào dư thừa.

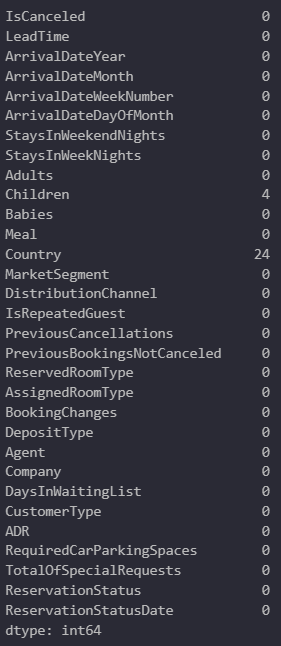


Ta thấy thuộc tính “ADR” của df1 có min < 0 (vô lý vì đây là giá cho thuê phòng nên không thể < 0). Kiểm tra có bao nhiêu quan sát có giá trị của “ADR” < 0 thì cho ra 1 kết quả duy nhất. Ta loại bỏ quan sát này.

Ta nhìn qua các thông số dữ liệu của dataset “H2”.



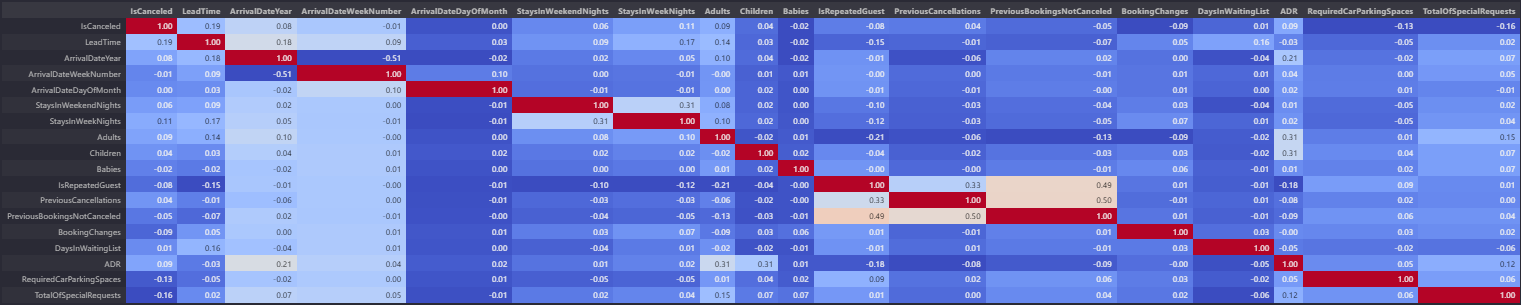
Tiếp đến, ta kiểm tra xem có thuộc tính nào có giá trị rỗng hay không, và đây là kết quả:

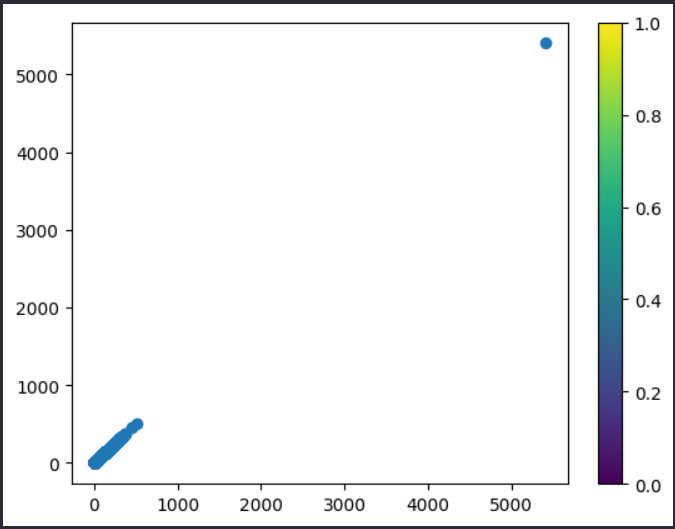


Ta thấy rằng, thuộc tính “Country” có 24 giá trị rỗng và thuộc tính “Children” có 4 giá trị rỗng. Ta xóa các quan sát có những dữ liệu bị thiếu vì không ảnh hưởng nhiều đến dữ liệu.

Kế đến, ta kiểm tra xem có quan sát nào bị trùng với nhau không. Và kết quả là có 25884 quan sát bị trùng (khoảng 33% dataset “H2”). Ta xóa những quan sát này để không còn giá trị trùng nữa.

Ta kiểm tra xem có thuộc tính nào của “H2” là dư thừa không. Kết quả là không có thuộc tính nào dư thừa.

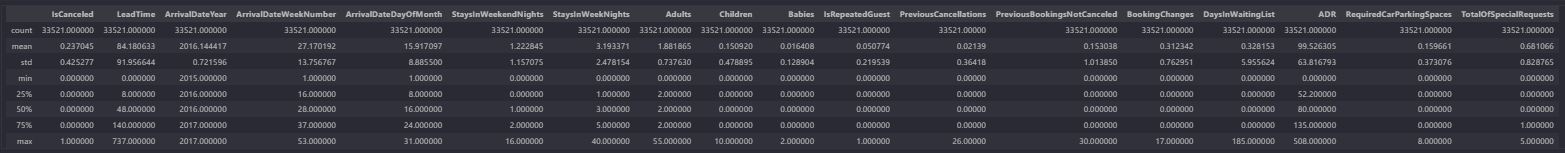
Ta thấy thuộc tính “ADR” có 1 quan sát cao bất thường. Ta loại bỏ quan sát này.



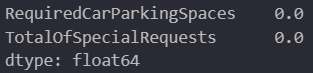
# CHƯƠNG 3: KHÁM PHÁ DỮ LIỆU

## Tổng quan dữ liệu

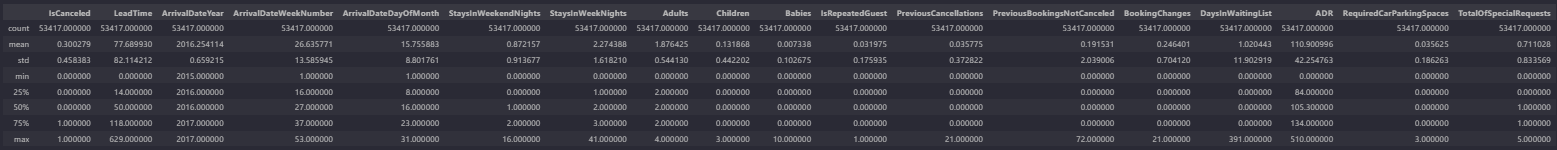
Nhìn vào các thông số của các thuộc tính trong dataset “H1”, rút ra được các nhận xét sau:



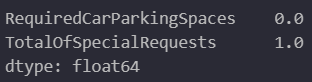
* Khả năng một lần đặt chỗ bị hủy là khoảng 24% (dựa theo mean của thuộc tính “IsCanceled”)
* Trong các lần đặt chỗ, thời gian đặt trước lớn nhất được ghi nhận là 737 ngày, trong khi đó thời gian đặt trước trung bình là khoảng 84 ngày (dựa theo max và mean của “LeadTime”)
* Trung bình một lần đặt chỗ sẽ đặt 1 ngày trong tuần, trong khi đó số ngày cuối tuần trung bình mà lần đặt chỗ đạt được là 3 ngày (dựa theo mean của “StaysInWeekNights” và “StaysInWeekendNights”)
* Khách hàng là người lớn chiếm đa số, trong khi đó trẻ em và trẻ sơ sinh là thiểu số. Dù vậy, cũng có trường hợp lần đặt chỗ không có khách hàng là người lớn (min của “Adults” = 0)
* Khả năng khách hàng đã từng đặt chỗ là 5% (tính theo mean của “IsRepeatedGuest”)
* Những trường hợp khách hàng từng hủy hoặc không hủy các lần đặt chỗ trước là thấp, vì thường khách hàng là những khách mới, dù vậy có những khách đã từng hủy đặt chỗ rất nhiều lần và ngược lại, có những khách nhiều lần đặt chỗ và không hủy (dựa theo mean và max của “PreviousCancellations” và “PreviousBookingsNotCanceled”)
* Có những trường hợp khách hàng cần chỉnh sửa lại thông tin lần đặt chỗ, nhưng hầu hết trường hợp là không chỉnh sửa gì cả (mean của “BookingChanges” < 1)
* Các trường hợp thường được xử lý nhanh gọn, trong chưa đầy 1 ngày (mean của “DaysInWaitingList” = 0.3) nhưng cũng nên xét đến những trường hợp ngoại lệ (max của “DaysInWaitingList” = 185)
* Giá cho thuê phòng trung bình là khoảng 99.5 đơn vị tiền, có thể lên đến tối đa là 508 đơn vị tiền, dù vậy cũng có những trường hợp miễn phí (dựa theo mean, max và min của “ADR”)
* Thông thường, khách hàng không cần chỗ đậu xe hay có yêu cầu đặt biệt nào (median của “RequiredCarParkingSpaces” = 0, median của “TotalOfSpecialRequests” = 0).



Nhìn vào các thông số của các thuộc tính trong dataset “H2”, rút ra được các nhận xét sau:



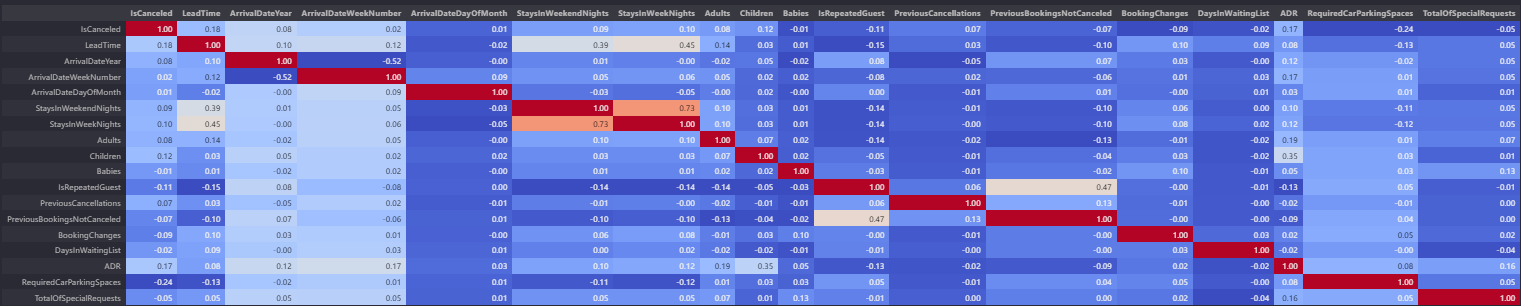
* Khả năng một lần đặt chỗ bị hủy là khoảng 30% (dựa theo mean của thuộc tính “IsCanceled”)
* Trong các lần đặt chỗ, thời gian đặt trước lớn nhất được ghi nhận là 629 ngày, trong khi đó thời gian đặt trước trung bình là khoảng 77 ngày (dựa theo max và mean của “LeadTime”)
* Trung bình một lần đặt chỗ sẽ đặt 2 ngày trong tuần, trong khi đó số ngày cuối tuần trung bình mà lần đặt chỗ đạt được là khoảng 1 ngày (dựa theo mean của “StaysInWeekNights” và “StaysInWeekendNights”)
* Khách hàng là người lớn chiếm đa số, trong khi đó trẻ em và trẻ sơ sinh là thiểu số. Dù vậy, cũng có trường hợp lần đặt chỗ không có khách hàng là người lớn (min của “Adults” = 0)
* Khả năng khách hàng đã từng đặt chỗ là 3% (tính theo mean của “IsRepeatedGuest”)
* Những trường hợp khách hàng từng hủy hoặc không hủy các lần đặt chỗ trước là thấp, vì thường khách hàng là những khách mới, dù vậy có những khách đã từng hủy đặt chỗ rất nhiều lần và ngược lại, có những khách nhiều lần đặt chỗ và không hủy (dựa theo mean và max của “PreviousCancellations” và “PreviousBookingsNotCanceled”)
* Có những trường hợp khách hàng cần chỉnh sửa lại thông tin lần đặt chỗ, nhưng hầu hết trường hợp là không chỉnh sửa gì cả (mean của “BookingChanges” < 1)
* Các trường hợp thường được xử lý trung bình khoảng 1 ngày nhưng cũng nên xét đến những trường hợp ngoại lệ (max của “DaysInWaitingList” = 391)
* Giá cho thuê phòng trung bình là khoảng 111 đơn vị tiền, có thể lên đến tối đa là 510 đơn vị tiền, dù vậy cũng có những trường hợp miễn phí (dựa theo mean, max và min của “ADR”)
* Thông thường, khách hàng không cần chỗ đậu xe nhưng có thể có yêu cầu đặt biệt (median của “RequiredCarParkingSpaces” là 0, median của “TotalOfSpecialRequests” là 1).



## Mối tương quan dữ liệu

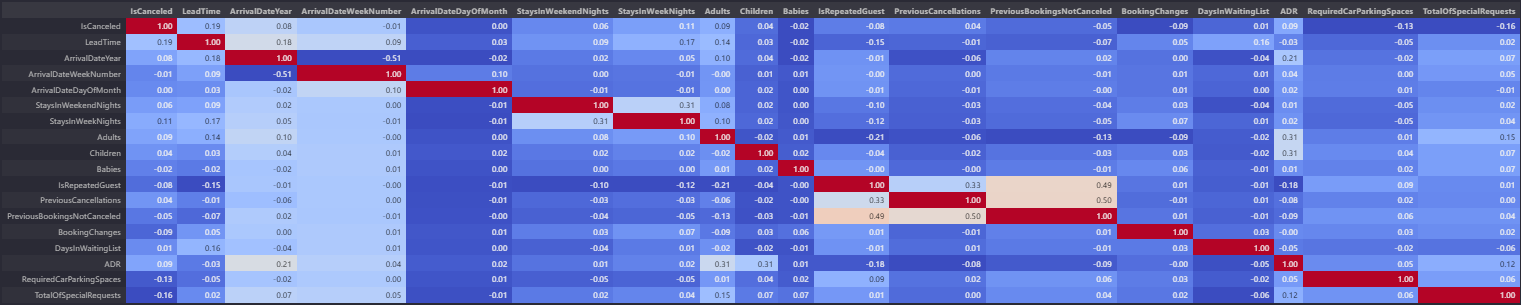
* + 1. Mối tương quan giữa các thuộc tính

Dưới đây là bảng tương quan giữa các thuộc tính của “H1”



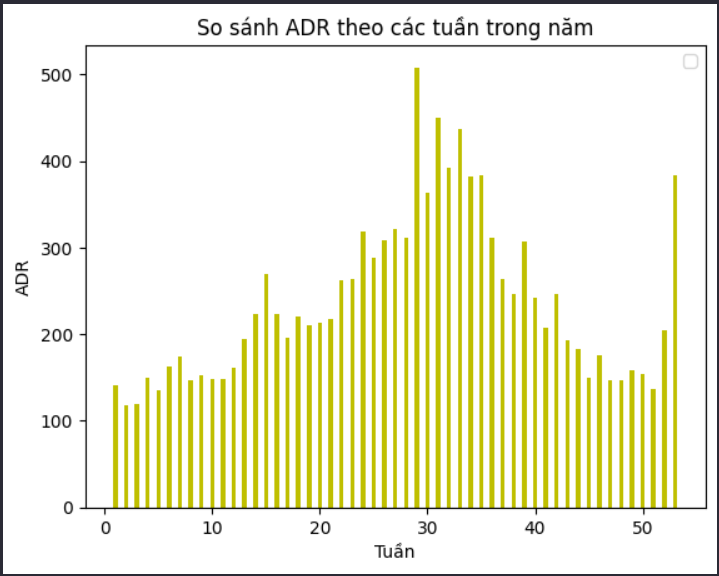
* Ta thấy số ngày được đặt chỗ trong tuần và số ngày được đặt chỗ cuối tuần có mối tương quan đồng biến mạnh nhất. Điều này là hợp lý vì chúng đều biểu thị cho số đêm khách hàng trải nghiệm tại khách sạn.
* Ta thấy tuần được đặt chỗ và năm được đặt chỗ có mối tương quan nghịch biến mạnh nhất.

Dưới đây là bảng tương quan giữa các thuộc tính của “H2”

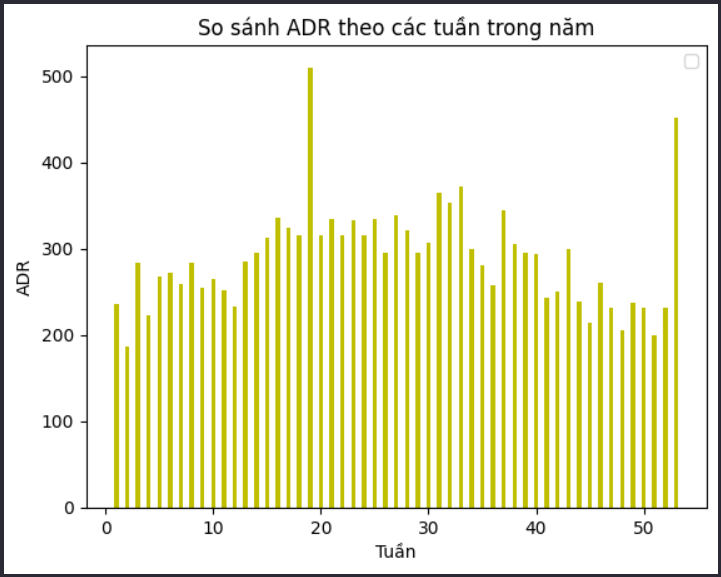


* Ta thấy số lần đặt chỗ trong quá khứ bị hủy bởi khách hàng trước lần đặt chỗ hiện tại và số lần đặt chỗ trong quá khứ không bị hủy bởi khách hàng trước lần đặt chỗ hiện tại có mối tương quan đồng biến mạnh nhất.
* Ta thấy tuần được đặt chỗ và năm được đặt chỗ có mối tương quan nghịch biến mạnh nhất.
  + 1. Tương quan giữa giá thuê phòng và tuần được hẹn

Đối với bộ dữ liệu H1 thì ta thấy tuần có giá trung bình cao nhất là tuần thứ 29 và thấp nhất là tuần 2. Điều này là hợp lý vì những tuần có giá cao đều là khoảng thời gian nghỉ lễ hoặc nghỉ hè. Ngoài ra, biểu đồ có độ dốc cao biểu thị cho xu hướng du lịch của khách hàng.



Đối với bộ dữ liệu H2 thì ta thấy tuần có giá trung bình cao nhất là tuần thứ 19 và thấp nhất là tuần 2. Điều này là hợp lý vì những tuần có giá cao đều là khoảng thời gian nghỉ lễ hoặc nghỉ hè. Biểu đồ có độ dốc thấp. Điều này có nghĩa giá phòng được giữ khá là ổn định.

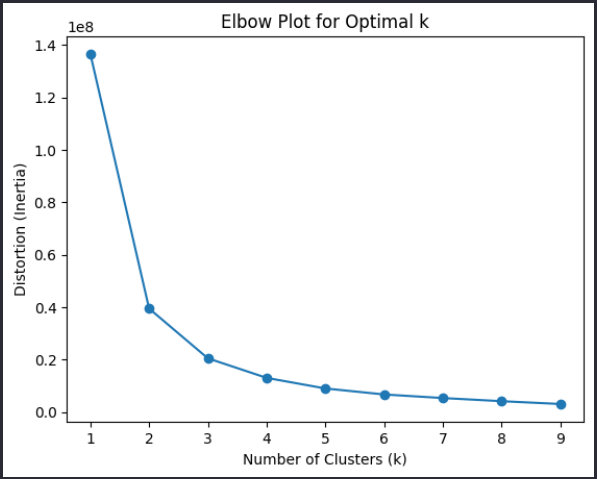


# CHƯƠNG 4: GOM CỤM K-MEANS

## Bộ dữ liệu “H1”: mối liên hệ giữa tuần trong năm và giá cho thuê phòng

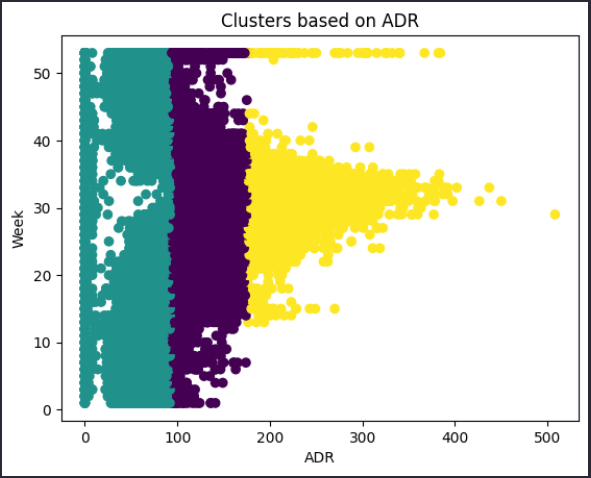
## Tìm số cụm

Ta sử dụng phương pháp Elbow để tìm số cụm:



Ta chọn số cụm k = 3 vì các giá trị thay đổi ít kể từ khi số cụm bằng 3.

* 1. Biểu đồ K-Means

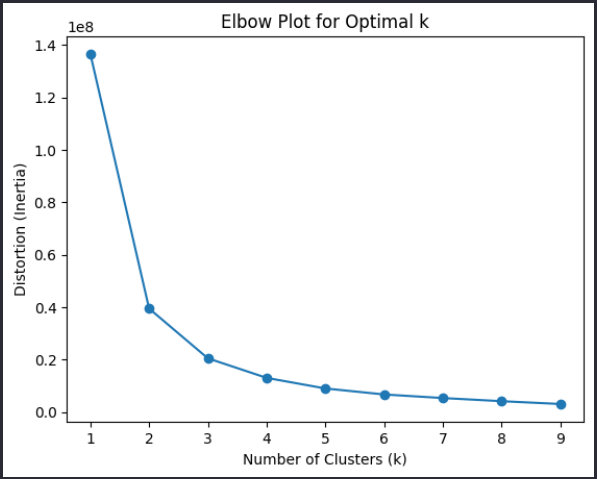


Ta nhận thấy cụm 1 và cụm 2 khá giống nhau, đều trải dài theo các tuần. Điều này chứng tỏ lượng khách đến để đặt chỗ phòng giá rẻ luôn ổn định.

Ngoài ra, cụm 3 tập trung vào cuối năm và mùa hè, là thời gian lễ hội. Điều này chứng tỏ lượng khách hàng cao cấp không phải lúc nào cũng có, mà chỉ trong một vài thời điểm trong năm.

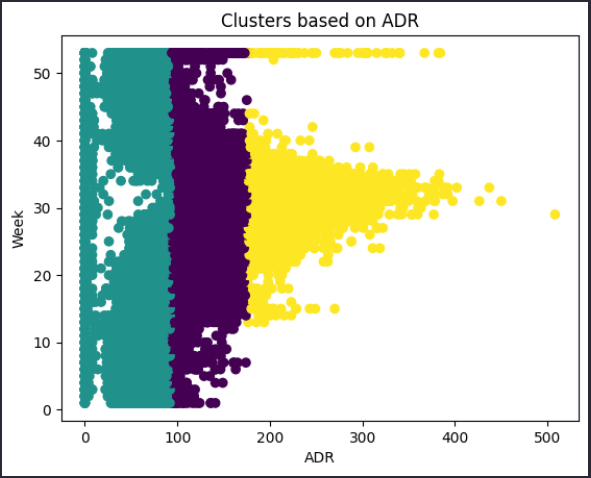
* 1. Bộ dữ liệu “H2”: mối liên hệ giữa tuần trong năm và giá cho thuê phòng
  2. Tìm số cụm

Ta sử dụng phương pháp Elbow để tìm số cụm:

1. 

Ta chọn số cụm k = 3 vì các giá trị thay đổi ít kể từ khi số cụm bằng 3.

* + 1. Biểu đồ K-Means

1. 
2. Ta nhận thấy cụm 1 và cụm 2 khá giống nhau, đều trải dài theo các tuần. Điều này chứng tỏ lượng khách đến để đặt chỗ phòng giá rẻ luôn ổn định.

Ngoài ra, cụm 3 tập trung vào cuối năm và mùa hè, là thời gian lễ hội. Điều này chứng tỏ lượng khách hàng cao cấp không phải lúc nào cũng có, mà chỉ trong một vài thời điểm trong năm.

# KẾT LUẬN

## Kết quả đạt được

Nhận biết được sự khác biệt của 2 loại khách sạn: resort (dataset “H1”) và khách sạn trong thành phố (dataset “H2”)

## Hạn chế của đề tài

Chưa triển khai sâu vào vấn đề

# TÀI LIỆU THAM KHẢO

[1] Nuno Antonio, Ana de Almeida, Luis Nunes, “Hotel booking demand datasets”, Tháng 2 năm 2019 [Trực tuyến]. Địa chỉ: https://www.sciencedirect.com/science/article/pii/S2352340918315191#t0005

# BẢNG PHÂN CÔNG CÔNG VIỆC

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Công việc** | **Mức độ hoàn thành** |
| Võ Huỳnh Thanh Phương | Mô tả dữ liệu  Tiền xử lý dữ liệu  Khám phá dữ liệu |  |