ĐỒ ÁN LÝ THUYẾT

Nhóm 5

TRỰC QUAN HÓA DỮ LIỆU 19CQ

Contents

Dataset: Airbnb_NYC_2019

Danh sách thành viên, phân công và đánh giá đồ án

Giai đoạn 1

Thông tin nhóm và đánh giá:

Nhóm 5	MSSV	Họ tên	Đánh giá cá nhân	Tỉ lệ đóng góp
	19120423	Phạm Sơn Tùng(*)	100%	20%
	19120585	Nguyễn Hải Nhật Minh	100%	20%
	19120529	Nguyễn Phước Huy	100%	20%
	19120261	Nguyễn Hữu Khôi	100%	20%
	18120357	Bùi Hoàn Hảo	100%	20%

Phân công:

Data abstraction và profiling cơ bản cho từng thuộc tính. Dựa vào task abstraction, ai có sử dụng thuộc tính nào thì sẽ profiling chi tiết thêm.

Hảo	Data abstraction cho dataset (dataset type, dataset availability, item semantic…) ID Name
Huy	Host ID Host Name Calculated_host_listing_count
Khôi	Neighbourhood_group Neighbourhood Latitude Longitude Availability_365
Minh	Number_of_reviews Last_review Reviews_per_month
Tùng	Room_type price Minimum_nights

Giai đoạn 2

Nhóm 5	MSSV	Họ tên	Đánh giá cá nhân	Tỉ lệ đóng góp
	19120423	Phạm Sơn Tùng	100%	20%
	19120585	Nguyễn Hải Nhật Minh	100%	20%
	19120529	Nguyễn Phước Huy	100%	20%
	19120261	Nguyễn Hữu Khôi	100%	20%
	18120357	Bùi Hoàn Hảo	100%	20%

Phân công:

Tùng	Phân bố phòng cho thuê của từng khu vực dựa trên giá phòng, số đêm tối thiểu và loại phòng
Huy	Loại phòng có giá trung bình thấp nhất trên mỗi khu vực. Loại phòng nào là nhiều nhất trên mỗi khu vực.
Hảo	Tìm khách sạn có số đánh giá nhiều nhất Loại phòng có giá trung bình cao nhất cho tất cả khu vực Thống kê số lượng thuê phòng trên mỗi quận của New York
Minh	Tìm loại phòng có số đánh giá trung bình mỗi tháng nhiều nhất
Khôi	Loại phòng nào có số ngày có phòng trung bình cao nhất trên mỗi khu vực

Giai đoạn 3

Nhóm 5	MSSV	Họ tên	Đánh giá cá nhân	Tỉ lệ đóng góp
	19120423	Phạm Sơn Tùng	100%	20%
	19120585	Nguyễn Hải Nhật Minh	100%	20%
	19120529	Nguyễn Phước Huy	100%	20%
	19120261	Nguyễn Hữu Khôi	100%	20%
	18120357	Bùi Hoàn Hảo	100%	20%

Phân công:

- mỗi thành viên xây dựng sheet trên Tableau và đánh giá dựa trên yêu cầu đã đặt ra ở giai đoạn Task abstraction
- Profiling chi tiết cho các thuộc tính được sử dụng:
    - Tùng: Room type, Minimum_nights
    - Khôi: neighbourhood_group, availability_365

Đánh giá chung

Nhóm 5	MSSV	Họ tên	Đánh giá cá nhân	Tỉ lệ đóng góp
	19120423	Phạm Sơn Tùng	100%	20%
	19120585	Nguyễn Hải Nhật Minh	100%	20%
	19120529	Nguyễn Phước Huy	100%	20%
	19120261	Nguyễn Hữu Khôi	100%	20%
	18120357	Bùi Hoàn Hảo	100%	20%

Báo cáo giai đoạn 1. Profiling tập dữ liệu và thực hiện data abstraction

1. Profiling tập dữ liệu

a. Kết quả profiling:

- Xem trong file Data Profiling đính kèm (Link dự phòng: DataProfiling.xlsx)

b. Nhận xét:

- Ở cột minimum_nights nhận 2 giá trị lớn nhất là 1250 và 1000 cho thấy khách hàng có nhu cầu thuê trong 1 khoảng thời gian dài (hơn 3 năm).
- Cột ID có distinctness đạt 100% cho thấy thuộc tính này có thể là thuộc tính khóa của dataset.
- Tại sao số lượng missing value trong "last_review" và "reviews_per_month" lại nhiều như vậy?
  - Airbnb là 1 dịch vụ du lịch khách sạn, và "last_review" và "reviews_per_month" được ghi lại dựa trên đánh giá của khách hàng về dịch vụ. Sở dĩ missing value của "last_review" và "reviews_per_month" nhiều như vậy là vì nó phụ thuộc hoàn toàn vào người dùng. Và đối với 1 khách sạn ,homestay, … Nếu người dùng không đánh giá sau khi sử dụng dịch vụ thì sẽ không có bất kì record nào về đánh giá của người dùng. Điều đó dẫn đến sự thiếu hụt về dữ liệu trong trường "last_review" và "review_per_month".
  - Lấy ví dụ trong thực tế, chúng ta cũng không thường xuyên đánh giá khi chúng ta vào các nhà hàng, cà phê hay chỉ đơn giản là sử dụng các ứng dụng trên CH PLAY, App Store. Trên phương diện cá nhân chúng ta thường chỉ xài app và bỏ qua phần đánh giá.
- Tại sao Actual value của trường "number_of_reviews" lại đạt giá trị tuyệt đối?
  - Như đã nói ở trên, do "number_of_reviews" không phụ thuộc vào khách hàng. Airbnb có thể tự ghi lại được. Do đó Actual value của trường "number_of_reviews" đạt giá trị tuyệt đối

c. Attribute profiling chi tiết:

Input Metadata

Field Name	Room_type
Field Data Type	String
Field Length	15
Data Profiling Summary Statistics	
NULL	0
Missing	0
Actual	48895
Completeness	100%
Cardinality	3
Uniqueness	0
Distinctness	0
Data Profiling Additional Statistics	
Field Data types	1
Field Length (MIN)	11
Field Length (MAX)	15
Field Value (MIN)	Entire home/apt
Field Value (MAX)	Shared room
Field Format	XXXXXXXXXXXXXXX

Room_type (Top 3 Field Values)	Count	Percentage
Entire home/apt	25409	51.97%
Private room	22326	45,67%
Shared room	1160	2.37%

Field Name	Price
Field Data Type	Number
Field Length	3
Data Profiling Summary Statistics	
NULL	0
Missing	11
Actual	48884
Completeness	99.98%
Cardinality	674
Uniqueness	1.38%
Distinctness	1.38%
Data Profiling Additional Statistics	
Field Data types	1
Field Length (MIN)	1
Field Length (MAX)	3
Field Value (MIN)	0

Field Value (MAX)	10000
Field Format	11111

Price (Top 10 Field Values)	Count	Percentage
100	2051	4%
150	2047	4%
50	1534	3%
60	1458	3%
200	1401	3%
75	1370	3%
80	1272	3%
65	1190	2%
70	1170	2%
120	1130	2%

Field Name	**Minimum_nights**
Field Data Type	Number
Field Length	3
Data Profiling Summary Statistics	
NULL	0
Missing	0
Actual	48895
Completeness	100%
Cardinality	109
Uniqueness	0.22%
Distinctness	0.22%
Data Profiling Additional Statistics	
Field Data types	1
Field Length (MIN)	1
Field Length (MAX)	3
Field Value (MIN)	1
Field Value (MAX)	365
Field Format	111

Minimum_nights (Top 10 Field Values)	Count	Percentage
1	12720	26%
2	11696	24%
3	7999	16%
30	3760	8%
4	3303	7%

5	3034	6%
7	2058	4%
6	752	2%
14	562	1%
10	483	1%

Field Name	**Neighbourhood_group**
Field Data Type	String
Field Length	20
Data Profiling Summary Statistics	
NULL	0
Missing	0
Actual	48895
Completeness	100%
Cardinality	5
Uniqueness	0%
Distinctness	0.01%
Data Profiling Additional Statistics	
Field Data types	1
Field Length (MIN)	5
Field Length (MAX)	13
Field Value (MIN)	Bronx
Field Value (MAX)	Staten Island
Field Format	XXXXXX

Price (Top 5 Field Values)	**Count**	**Percentage**
Manhattan	21,661	44.30%
Brooklyn	20,104	41.12%
Queens	5,666	11.59%
Bronx	1,091	2.23%
Staten Island	373	0.76%

Field Name	**Availability 365**
Field Data Type	Int64
Field Length	
Data Profiling Summary Statistics	
NULL	0
Missing	0
Actual	48895

Completeness	100%
Cardinality	366
Uniqueness	0.74%
Distinctness	0.74%
Data Profiling Additional Statistics	
Field Data types	[<class 'int'>]
Field Length (MIN)	1
Field Length (MAX)	3
Field Value (MIN)	0
Field Value (MAX)	365
Field Format	XXX

Availability 365 (Top 10 Field Values)	**Count**	**Percentage**
0	17533	35.9
365	1295	2.6
364	491	1.0
1	408	0.8
89	361	0.73
5	340	0.69
3	306	0.62
179	301	0.61
90	290	0.59
2	270	0.55

2. Data abstraction

- Dataset type: Table
- Dataset availability: Static type
- Item semantic: Thông tin về 1 lượt cho thuê phòng ở New York với các thông tin chi tiết
    - ID của mỗi hóa đơn thuê phòng
    - Mã định danh của Host
    - Tọa độ của phòng
    - Giá phòng
    - Số đêm tối thiểu phải đặt
    - Số lượng đánh gia của khách hàng
    - Ngày đánh giá gần nhất của khác hàng
    - Số lượng đánh giá mỗi tháng
    - Số danh sách host lưu trữ
    - Số ngày có phòng trong năm

- Tên của hợp đồng thuê phòng
- Tên của host
- Nhóm khu vực của phòng (quận)
- Khu phố
- Loại phòng

- Attribute Abstraction: xem trong file AttAbstract.xlsx đính kèm (Link dự phòng: AttAbstract.xlsx)

3. Xử lý missing value

- Tất cả các missing value của "last_review" và "reviews_per_month" đều có number_of_review = 0, tuy nhiên không phải tất cả các record có number_of_review = 0 thì "last_review" và "reviews_per_month" đều là missing value.
  - Kết luận: Đây là missing value MAR.
  - Xử lý missing value: Tỉ lệ missing value chiếm cực kỳ cao là 20.56% (10052/48895) > 3% tương ứng với các trường "last_review" và "review_per_month". Do đó chúng ta không thể bỏ các dữ liệu này. Chúng ta cần thay thế các missing value.
    - Last_review: Các giá trị của trường "last_review" lệch nhau rất xa, do đó chúng ta phải sử dụng dữ liệu trung vị để thay thế. Giá trị trung vị tìm được là "19/05/2019"
    - Review_per_month: Các giá trị của trường "review_per_month" khá gần nhau. Do vậy chúng ta dùng giá trị trung bình(MEAN) để thay thế cho các missing value. Giá trị trung bình để thay thế là 1,37.
- Thuộc tính Price có 11 item nhận giá trị 0 (chiếm tỉ lệ rất ít) nên có thể xóa bỏ các dòng có Price = 0.

Báo cáo giai đoạn 2. Xác định yêu cầu khai thác và thực hiện giai đoạn task asbtraction

Yêu cầu 1. Phân bố phòng cho thuê của từng khu vực dựa trên giá phòng, số đêm tối thiểu và loại phòng

- Bước 1: Khái quát lại yêu cầu: thể hiện phân bố của các item dựa trên 3 thuộc tính price, minimum_nights và room_type
- Bước 2: Phân rã tác vụ:
  - Thể hiện phân bố của 3 thuộc tính trên biểu đồ.
    - Analyze → Consume → Present
    - Target: Many attributes → Correlations

Yêu cầu 2. Loại phòng có giá trung bình thấp nhất trên mỗi khu vực

- Bước 1: Khái quát yêu cầu: Tìm giá trị của thuộc tính “room_type" có trung bình của thuộc tính “price" thấp nhất theo từng giá trị phân biệt của thuộc tính “neighbourhood_group"
- Bước 2: Phân rã tác vụ:
  - Thêm thuộc tính “AVG_Price”
  - Tìm kiếm các dòng thông tin có thuộc tính “neighbourhood_group" và “room_type" nhận giá trị lần lượt thoả các giá trị phân biệt của 2 thuộc tính này
    - Mid-level -> Search -> Browse
  - Tổng hợp các giá trị phân biệt của “neighbourhood_group" và “room_type" cùng giá trị trung bình tương ứng.
    - Low level à Query → Summarize.
    - Target: Attributes → One attribute → Distribution.

Yêu cầu 3. Loại phòng nào là nhiều nhất trên mỗi khu vực.

- Bước 1: Khái quát yêu cầu: Tìm giá trị của thuộc tính “room_type" có số lượng nhiều nhất theo từng giá trị phân biệt của thuộc tính “neighbourhood_group"
- Bước 2: Phân rã tác vụ:
  - Thêm thuộc tính “count_room_type"
  - Tìm kiếm các dòng thông tin có thuộc tính “neighbourhood_group" và “room_type" nhận giá trị lần lượt thoả các giá trị phân biệt của 2 thuộc tính này
    - Mid level -> Search -> Browse
  - Tổng hợp các giá trị phân biệt của “neighbourhood_group" và “room_type" cùng số lượng tương ứng.
    - Low level: Query → Summarize.
    - Target: Attributes → One attribute → Distribution.

Yêu cầu 4. Khách sạn có lượng đánh giá cao nhất.

- Bước 1: Khái quát yêu cầu: Tìm giá trị của thuộc tính “number_of_views" có số lượng cao nhất theo từng giá trị phân biệt của thuộc tính “id"
- Bước 2: Phân rã tác vụ:
  - Tìm kiếm các dòng thông tin có thuộc tính “number_of_reviews " và nhận giá trị lần lượt thoả các giá trị phân biệt của thuộc tính “id”
    - Mid level -> Search -> Browse
    - Target: Attributes → One attribute → Distribution.

Yêu cầu 5. Loại phòng có giá trung bình cao nhất cho tất cả khu vực

- Bước 1: Khái quát yêu cầu: Tìm giá trị của thuộc tính “room_type" có giá trung bình hay giá trị của thuộc tính “price" cao nhất
- Bước 2: Phân rã tác vụ:
  - Tìm kiếm các dòng thông tin có thuộc tính “price " và “room_type"
    - Mid level -> Search -> Browse
    - Target: Attributes → One attribute → Distribution.

Yêu cầu 6. Tìm loại phòng có số đánh giá trung bình mỗi tháng nhiều nhất theo từng khu vực?

- Bước 1: Khái quát yêu cầu: Tìm loại phòng “room_type” có số đánh giá trung bình mỗi tháng “reviews_per_month” nhiều nhất theo từng “neighborhood_group”.
- Bước 2: Phân rã tác vụ:
  - Tính số đánh giá trung bình của từng loại phòng => Thêm cột số đánh giá trung bình mỗi tháng “review_per_month_avg”.
    - Action: analyze -> produce -> derive
    - Target: Attribute -> One attribute (review_per_month_avg)

  - Tìm loại phòng có số đánh giá trung bình mỗi tháng nhiều nhất
    - Action: query -> compare
    - Target: Attributes -> One attribute -> Extremes (cực trị)

Yêu cầu 7. Tìm loại phòng có số ngày có phòng trung bình nhiều nhất mỗi khu vực

- Bước 1: Khái quát lại yêu cầu: Tìm loại phòng “room_type” có số ngày có phòng “availability_365” nhiều nhất theo từng “neighborhood_group”.
- Bước 2: Phân rã tác vụ:
  - Tính số đánh giá trung bình của từng loại phòng => Thêm cột số đánh giá trung bình mỗi tháng “avalability_avg”.
    - Action: analyze -> produce -> derive
    - Target: Attribute -> One attribute (avalability_avg)

  - Tìm loại phòng có số đánh giá trung bình mỗi tháng nhiều nhất
    - Action: query -> compare
    - Target: Attributes -> One attribute -> Extremes (cực trị)

Yêu cầu 8. Thống kê số lượng thuê phòng dựa trên mỗi quận của New York

- Bước 1: Khái quát lại yêu cầu: Thống kê số lượng thuê phòng dựa trên mỗi quận
- Bước 2: Phân rã tác vụ:
  - Tổng hợp các giá trị phân biệt của thuộc tính "Neighborhood_Group" và "ID" cùng kết quả đếm được tương ứng:
    - Low level: Query -> Summarize
    - Target: Attributes -> One attribute -> Distribution.

Báo cáo giai đoạn 3. Thiết kế Idiom và cài đặt thiết kế

Idiom cho các yêu cầu

Yêu cầu 1: Phân bố phòng cho thuê của từng khu vực dựa trên giá phòng, số đêm tối thiểu và loại phòng

Idiom	Scatter-plot
Data	Price: quantitative Minimum_nights: quantitative Room_type: categorical →Keys: 0 → Scatter plot
Encode	Mark: point Channel: Quantitative → Vertical spatial position (trục y) Minimum_nights → Horizontal spatial postion (trục x) Room_type: Color Hue Align: không Sort: không
Manipulate	Selection – Hover – Tooltip (Minimum nights, neighbourhood_group, neighbourhood, room_type, price) Navigate – Attribute Reduction – Slice (Neighbourhood_group, Room_type)
Task	Phân bố
Scalability	Số item được biểu hiện: 48895 chia ra cho từng level của Neighbourhood_group

Yêu cầu 2: Loại phòng có giá trung bình thấp nhất trên mỗi khu vực

Idiom	Bar chart
Data	Price _avg: quantitative Room_type: categorical Neighborhood_group: categorical →Keys: 1(Room_type) → Bar chart

Encode	Mark: Line
	Channel:
	• Express: quantitative (length)
	• Spatial region: categorical (mark)
	  ○ Separate: horizontal position (trục x)
	  ○ Align: vertical position (trục y)
	• Room_type: Color Hue

Task	Compare, lookup value
Scalability	Số item được biểu đạt trên biểu đồ: 15

Yêu cầu 6: Tìm loại phòng có số đánh giá trung bình mỗi tháng nhiều nhất theo từng khu vực?

Idiom	**Bar chart**
Data	Review_per_month_avg: quantitative Room_type: categorical Neighborhood_group: categorical →Keys: 1(Room_type) → Bar chart

Encode	Mark: Line
	Channel:
	• Express: quantitative (length)
	• Spatial region: categorical (mark)
	  ○ Separate: horizontal position (trục x)
	  ○ Align: vertical position (trục y)
	• Room_type: Color Hue

Task	Compare, lookup value
Scalability	Số item được biểu đạt trên biểu đồ: 15

Yêu cầu 8: Thống kê số lượng thuê phòng dựa trên mỗi quận của New York

Idiom	**Pie-chart**
Data	Neighborhood_group: categorical
Encode	Mark: area Channel: Neighborhood_group: color Neighborhood_group: angle
Task	Distribution
Scalability	Số item được biểu hiện: 5

Yêu cầu 7: Tìm loại phòng có số ngày có phòng trung bình nhiều nhất mỗi khu vực

Idiom	**Bar chart**
Data	availability_avg: quantitative Room_type: categorical Neighborhood_group: categorical

	→Keys: 1(Room_type) → Bar chart

Encode	Mark: Line
	Channel:
	• Express: quantitative (length)
	• Spatial region: categorical (mark)
	  ○ Separate: horizontal position (trục x)
	  ○ Align: vertical position (trục y)
	• Room_type: Color Hue

Task	Compare, lookup value
Scalability	Số item được biểu đạt trên biểu đồ: 15

Vẽ tableau:

Link: click here

Nhận xét:

Yêu cầu 1:

- Nguyên tắc biểu đạt: mục tiêu của biểu đồ là thể hiện phân bố của 3 thuộc tính → Dùng biểu đồ scatter plot là phù hợp vì scatter plot chuyên dùng cho những task liên quan tới độ tương quan, xu hướng, mật độ, phân bố,…
- Nguyên tắc hiệu quả:
  - Accuracy (độ chính xác)
    - Postion trên 2 trục x, y là kênh hiệu quả nhất cho 2 thuộc tính định lượng (Price, Minimum_nights)
    - Color hue là kênh hiệu quả dành cho thuộc tính phân loại (room_type)
  - Discriminability: các item được phân biệt rõ ràng nhờ tách biệt về vị trí và có sự slice giữa các level của neighbourhood_group, tuy nhiên ở một số level thì các item phân bố với mật độ cao dẫn đến sự chồng chất các điểm làm cho tính discriminability bị giảm.
  - Separability: Biểu đồ sử dụng 2 kênh là Position và color → Không có tương tác.
  - Visual popout:
    - Màu sử dụng trong biểu đồ là màu bão hòa giúp làm nổi bật các point trên biểu đồ.
    - Các point có phân loại là shared room có số lượng ít thì được tô màu đỏ để làm nổi bật hơn so với các điểm thuộc 2 phân loại còn lại.

Yêu cầu 2:

- Nguyên tắc biểu đạt:
  - Mục tiểu của biểu đồ là thể hiện sự tương quan giữa giá trung bình (avg price), room_type và neighborhood_group.
  - Sử dụng Bar Chart là phù hợp vì Bar chart được dùng để so sánh giá trị giữa các loại khác nhau.
- Nguyên tắc hiệu quả:
  - Accuracy (độ chính xác):
    - Spatial region là kênh hiệu quả nhất cho thuộc tính phân loại .
    - Postion là kênh hiệu quả nhất cho thuộc tính định lượng.
  - Discriminability: các item (bar) được phân biệt rõ ràng nhờ tách biệt về vị trí và màu sắc.
  - Separability: biểu đồ sử dụng 2 kênh spatial region và color để biểu diễn => region và color chỉ tương tác khi kích thước của region nhỏ. Ta thấy, trên biểu đồ đã sử dụng màu để làm nổi bật region => dễ dàng nhận biết.
  - Visual popout: sử dụng màu sắc làm nổi bật các cột trong biểu đồ.

Yêu cầu 8:

- Nguyên tắc biểu đạt:
  - Mục tiêu của biểu đồ là thể hiện phân phối của biến Neighborhood_group.
  - Sử dụng biểu đồ Pie Chart là phù hợp với nhiệm vụ phân phối dữ liệu.
- Nguyên tắc hiệu quả:
  - Diện tích (area) được sử dụng làm đặc điểm biểu hiện (mark) cho các phần tử Pie Chart, giúp thể hiện tỷ lệ phần trăm mỗi nhóm trong phân phối.
  - Màu sắc (color) được sử dụng để phân biệt các phần tử Pie Chart thuộc các nhóm khác nhau.
  - Góc (angle) cũng được sử dụng để giúp phân biệt các phần tử.
  - Discriminability: các nhóm được phân biệt rõ ràng nhờ sự khác nhau về diện tích và màu sắc.
  - Separability:
    - Biểu đồ sử dụng màu sắc và góc, không có tương tác lý thuyết giữa chúng.
    - Tuy nhiên, do số lượng nhóm trên biểu đồ là 5, không thể gây sự chồng chất và làm tăng hiệu quả của chúng.
  - Visual popout:
    - Màu sắc được sử dụng trong biểu đồ giúp làm nổi bật các phần tử Pie Chart.
    - Các nhóm có màu sắc khác nhau giúp tạo sự tương phản và thu hút sự chú ý của người xem.

Yêu cầu 7:

- Nguyên tắc biểu đạt:
  - Mục tiêu của biểu đồ là thể hiện sự tương quan giữa số ngày còn phòng trung bình (availability_avg) và room_type, neighborhood_group.
  - Sử dụng biểu đồ Bar chart để so sánh các giá trị với nhau.
- Nguyên tắc hiệu quả:
  - Accuracy (độ chính xác):
    - Spatial region là kênh hiệu quả nhất cho thuộc tính phân loại .
    - Postion là kênh hiệu quả nhất cho thuộc tính định lượng.
  - Discriminability: các item (bar) được phân biệt rõ ràng nhờ tách biệt về vị trí và màu sắc.
  - Separability: biểu đồ sử dụng 2 kênh spatial region và color để biểu diễn => region và color chỉ tương tác khi kích thước của region nhỏ. Ta thấy, trên biểu đồ đã sử dụng màu để làm nổi bật region => dễ dàng nhận biết.
  - Visual popout: sử dụng màu sắc làm nổi bật các cột trong biểu đồ.