ĐỒ ÁN THỰC HÀNH

Nhóm 5

TRỰC QUAN HÓA DỮ LIỆU 19CQ

Contents

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

Giai đoạn 1 – Data profiling, data abstraction

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 chi tiết:

Data profiling và data abstraction cơ bản, xử lý missing value (nếu có) cho từng thuộc tính

Tùng	CLIENTNUM Attrition_Flag Customer_Age Gender Dependent_count Education_Level
Huy	Marital_Status Income_Category Card_Category Months_on_book Total_Relationship_Count
Hảo	Months_Inactive_12_mon Contacts_Count_12_mon Credit_Limit Total_Revolving_Bal
Minh	Dataset abstraction(dataset type, dataset availability, item semantic…) Avg_Open_To_Buy Total_Amt_Chng_Q4_Q1

	Total_Trans_Amt Total_Trans_Ct
Khôi	Total_Ct_Chng_Q4_Q1 Avg_Utilization_Ratio Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1 Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2

Data profiling chi tiết cho từng thuộc tính:

Tùng: Attrition_flag, Education_level, Customer_age, Total_Trans_Amt, Total_Trans_Ct

Giai đoạn 2 – task abstraction

Bảng 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: mỗi thành viên tự đề xuất 1-2 yêu cầu và trình bày các bước task abstraction.

Tùng	Độ tương quan giữa quyết định rời đi/ở lại của khách hàng với số lượng giao dịch đã thực hiện và tổng số tiền đã giao dịch
Huy	Thống kê tỷ lệ khách hàng sử dụng thẻ tín dụng dựa trên mức thu nhập của họ. Thống kê loại thẻ tín dụng mà khách hàng đang nắm giữ.
Hảo	Thống kê số lượng khách hàng trên từng độ tuổi Tìm khác hàng có tuổi nhỏ nhất/ lớn nhất Thống kê độ tuổi trung bình của mỗi giới tính
Minh	Tìm trình độ học vấn có tổng số lượng giao dịch nhiều nhất trong 12 tháng gần nhất?
Khôi	Thống kê trung bình số tiền giao dịch trên mỗi giao dịch trong 12 tháng so với mức thu nhập của họ

Giai đoạn 3- Visualization

Bảng 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: mỗi thành viên tự thực hiện code trình bày biểu đồ và đánh giá dựa trên yêu cầu đã đặt ra ở giai đoạn Task abstraction

Đánh giá chung

Bảng 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%

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 DataProfiling đính kèm (Link Onedrive dự phòng: DataProfiling.xlsx)

b. Nhận xét:

- Nhìn vào tỉ lệ Completeness và Uniqueness của CLIENTNUM có thể đoán được đây là primary key của tập dữ liệu này không?
  - Ta thấy Distinctness của CLIENTNUM = 100%, tức là CLIENTNUM nhận giá trị phân biệt tại mọi dòng, đồng nghĩa với việc CLIENTNUM đảm bảo được vai trò là khóa của dataset.
- Nhận xét riêng cho các trường Avg_Open_To_Buy, Total_Amt_Chng_Q4_Q1, Total_Trans_Amt, Total_Trans_Ct thì dữ liệu không có bất thường nào.

c. Attribute profiling:

	Input Metadata
Field Name	Attrition_Flag
Field Data Type	String
Field Length	17
	Data Profiling Summary Statistics
NULL	0
Missing	0
Actual	10127
Completeness	100%
Cardinality	5033
Uniqueness	0
Distinctness	0
	Data Profiling Additional Statistics
Field Data types	1
Field Length (MIN)	16
Field Length (MAX)	17
Field Format	XXXXXXXXXX

Room_type (Top 2 Field Values)	**Count**	**Percentage**
Existing Customer	8500	83.93%

Attrited Customer	1627	16.07%

Input Metadata

Field Name	Education_Level
Field Data Type	String
Field Length	13
	Data Profiling Summary Statistics
NULL	0
Missing	0
Actual	8608
Completeness	85%
Cardinality	7
Uniqueness	0.07%
Distinctness	0.08%
	Data Profiling Additional Statistics
Field Data types	1
Field Length (MIN)	1
Field Length (MAX)	13
Field Format	XXXXXXXXXXXXX

Education_Level (Top 10 Field Values)	**Count**	**Percentage**
Graduate	298	2.94%
High School	212	2.09%
Uneducated	148	1.46%
Unknown	146	1.44%
College	97	0.96%
Post-Graduate	56	0.55%
Doctorate	43	0.42%

Input Metadata

Field Name	Customer_Age
Field Data Type	Number
Field Length	2
	Data Profiling Summary Statistics
NULL	0
Missing	0
Actual	10127
Completeness	100%
Cardinality	45
Uniqueness	0.02%

Distinctness	0.44%
	Data Profiling Additional Statistics
Field Data types	1
Field Length (MIN)	2
Field Length (MAX)	2
Field Value (MIN)	26
Field Value (MAX)	73
Field Format	NN

Customer_Age (Top 10 Field Values)	**Count**	**Percentage**
59	157	1.55%
34	146	1.44%
33	127	1.25%
60	127	1.25%
32	106	1.05%
65	101	1.00%
61	93	0.92%
62	93	0.92%
31	91	0.90%
26	78	0.77%

	Input Metadata
Field Name	Total_Trans_Amt
Field Data Type	Integer
Field Length	5
	Data Profiling Summary Statistics
NULL	0
Missing	0
Actual	10127
Completeness	100%
Cardinality	5033
Uniqueness	49.7%
Distinctness	79.7%
	Data Profiling Additional Statistics
Field Data types	1
Field Length (MIN)	3
Field Length (MAX)	5
Field Value (MIN)	510
Field Value (MAX)	18484
Field Format	NNN, NNNN, NNNNN

Total_Trans_Amt (Top 10 Field Value)	Count	Percentage
4253	11	0.11%
4509	11	0.11%
2229	10	0.10%
4518	10	0.10%
4869	9	0.09%
4220	9	0.09%
4498	9	0.09%
4037	9	0.09%
4313	9	0.09%
4042	9	0.09%

	Input Metadata
Field Name	Total_Trans_Ct
Field Data Type	Integer
Field Length	3
	Data Profiling Summary Statistics
NULL	0
Missing	0
Actual	10127
Completeness	100%
Cardinality	126
Uniqueness	1.24%
Distinctness	1.24%
	Data Profiling Additional Statistics
Field Data types	1
Field Length (MIN)	1
Field Length (MAX)	3
Field Value (MIN)	10
Field Value (MAX)	139
Field Format	N, NN, NNN

Total_Trans_Ct value	Count	Percentage
81	208	2.05%
75	203	2.00%
71	203	2.00%
69	202	1.99%
82	202	1.99%
76	198	1.96%

77	197	1.95%
70	193	1.91%
74	190	1.88%
78	190	1.88%

Input Metadata

Field Name	Total_Ct_Chng_Q4_Q1
Field Data Type	Float64
Field Length	-
	Data Profiling Summary Statistics
NULL	0
Missing	0
Actual	10127
Completeness	100%
Cardinality	830
Uniqueness	0.07%
Distinctness	0.08%
	Data Profiling Additional Statistics
Field Data types	1
Field Length (MIN)	-
Field Length (MAX)	-
Field Format	X.XXXXXXXXXX

Total_Ct_Chng_Q4_Q1	**Count**	**Percentage**
0.667	171	1.688555%
1.000	166	1.639182%
0.500	161	1.589809%
0.750	156	1.540436%
0.600	113	1.115829%
0.800	101	0.997334%
0.714	92	0.908463%
0.833	85	0.839340%
0.778	69	0.681347%
0.625	63	0.622099%

Input Metadata

Field Name	Income_Category
Field Data Type	object
Field Length	-
	Data Profiling Summary Statistics

NULL	1112
Missing	1112
Actual	9015
Completeness	89%
Cardinality	5
Uniqueness	0.06%
Distinctness	0.05%
	Data Profiling Additional Statistics
Field Data types	object
Field Length (MIN)	3
Field Length (MAX)	14
Field Format	XXXXXXXXXXXXX

Income_Category	**Count**	**Percentage**
Less than $40K	3561	39.500832%
$40K - $60K	1790	19.855796%
$80K - $120K	1535	17.027177%
$60K - $80K	1402	15.551858%
$120K +	727	8.064337%

2. Data abstraction

- Dataset type: Table
- Dataset availability: Static type
- Item semantic:
  - BankChurners là thuật ngữ dùng để chỉ những người rời bỏ ngân hàng. Tức là họ ngưng sử dụng dịch vụ của ngân hàng đó
  - Môi item trong trong dataset cung cấp các thông tin về khách hàng. Một số thông tin cơ bản như tuổi, giới tính hay trình độ học vấn.
  - Sau đó là thông tin sử dụng tài khoản của khách hàng(còn sử dụng hay đã ngưng sử dụng, đang nuôi con hay không, khoảng thu nhập,….).
  - Và sau đó là các chỉ số sử dụng tài khoản của khách hàng như average_open_buy hay tổng số tiền đã thay đổi từ quý 1 sang quý 4.
- Attribute Abstraction: xem trong file AttAbstract.xlsx đính kèm (link Onedrive dự phòng): AttAbstract.xlsx

3. Xử lý missing value

- Thuộc tính Education_Level nhận giá trị Unknown tại gần 1500 dòng (chiếm 15% tổng số dòng) → Dùng phương thức điền missing value thay vì xóa dòng và xóa cột.
    - Phương thức điền: điền bằng giá trị Mode trong các dòng có cùng giá trị Customer_Age (dùng giá trị Customer_Age phổ biến nhất trong những người cùng độ tuổi).
- Thuộc tính Income_Category nhận giá trị Unknow tại 1112 dòng (chiếm 11% tổng số dòng) nhưng vì Income_Category là thuộc tính Categorical nên xem giá trị Unknown này là giá trị có ý nghĩa.
- Thuộc tính Marital_Status nhận giá trị Unknow tại 749 dòng (chiếm 7% tổng số dòng) nhưng vì Marital_Status là thuộc tính Categorical nên xem giá trị Unknown này là giá trị có ý nghĩa.

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

Yêu cầu 1. Độ tương quan giữa quyết định rời đi/ở lại của khách hàng với số lượng giao dịch đã thực hiện và tổng số tiền đã giao dịch.

- Bước 1: Khái quát lại yêu cầu: biểu hiện độ tương quan giữa 3 thuộc tính "Attrition_flag", "Total_Trans_Amt" và "Total_Trans_Ct"
- Bước 2: Phân rã tác vụ:
    - Biểu thị 1 item là 1 đối tượng trên biểu đồ với 3 thuộc tính được chọn.
        - Analyze → Consume → Present
        - Target: Many attributes → Correlations

Yêu cầu 2. Thống kê độ tuổi trung bình của mỗi giới tính

- Bước 1: Khái quát lại yêu cầu: Tìm giá trị của thuộc tính "gender" có trung bình của thuộc tính "customer_age"
- Bước 2: Phân rã tác vụ:
    - Thêm thuộc tính "AVG_Price"
        - High level → Analyze → Produce → Derive
    - Tìm kiếm các dòng có thuộc tính "gender" nhận giá trị lần lượt thỏa các giá trị phân biệt của thuộc tính này
        - Mid level → Search → Browse
    - Tổng hợp các "gender" có độ tuổi trung bình tương ứng
        - Low level → Query → Summarize.
        - Target: Attributes → One attribute → Distribution.

Yêu cầu 3. Tìm số tuổi lớn nhất/nhỏ nhất của khách hàng.

- Bước 1: Khái quát lại yêu cầu: Tìm giá trị của thuộc tính "customer_age" lớn nhất nhỏ nhất với từng khách hàng "clientnum"
- Bước 2: Phân rã tác vụ:
  - Tìm kiếm các dòng thông tin của thuộc tính "customer_age" nhận giá trị lần lượt thoả các giá trị phân biệt của thuộc tính "clientnum"
    - Low level: Mid level -> Search -> Browse
    - Target: Attributes → One attribute → Distribution.

Yêu cầu 4. Tìm trình độ học vấn có tổng số lượng giao dịch nhiều nhất trong 12 tháng gần nhất?

- Bước 1: Khái quát lại yêu cầu: Tìm trình độ học vấn "Education_Level" có tổng số lượng giao dịch "Total_Trans_Ct" nhiều nhất.
- Bước 2: Phân rã tác vụ:
  - Tính tổng số lượng giao dịch của từng trình độ học vấn => Thêm cột tổng số lượng giao dịch "Total_Trans_Ct _Sum".
    - Action: analyze -> produce -> derive
    - Target: Attribute -> One attribute (Total_Trans_Ct _Sum)
  - Tìm trình độ học vấn có tổng số lượng giao dịch nhiều nhất
    - Action: query -> compare
    - Target: Attributes -> One attribute -> Extremes (cực trị)

Yêu cầu 5. Thống kê tỷ lệ khách hàng sử dụng thẻ tín dụng dựa trên mức thu nhập của họ.

- Bước 1: Khái quát lại yêu cầu: Thống kê số lượng khách hàng "clientnum" và gom nhóm dựa trên "income_category"
- Bước 2: Phân rã tác vụ:
  - Thêm thuộc tính "number_of_user":
    - High level: Analyze -> Produce -> Derive
  - Lần lượt tìm các item thỏa từng giá trị phân biệt của thuộc tính "income_category" và đếm số item có thuộc tính "clientnum"
    - Mid level: Search à Browse
  - Tổng hợp các giá trị phân biệt của thuộc tính "income_category" và "clientnum" cùng kết quả đếm được tương ứng:
    - Low level: Query -> Summarize

- Target: Attributes -> One attribute à Distribution.

Yêu cầu 6. Thống kê loại thẻ tín dụng mà khách hàng đang nắm giữ.

- Bước 1: Khái quát lại yêu cầu: Thống kê số lượng khách hàng “clientnum" và gom nhóm dựa trên "card_category"
- Bước 2: Phân rã tác vụ:
  - Thêm thuộc tính “number_of_user”:
    - High level: Analyze ->Produce -> Derive
  - Lần lượt tìm các item thỏa từng giá trị phân biệt của thuộc tính “card_category” và đếm số item có thuộc tính “clientnum”
    - Mid level: Search à Browse
  - Tổng hợp các giá trị phân biệt của thuộc tính “card_category” và “clientnum” cùng kết quả đếm được tương ứng:
    - Low level: Query ->àummarize
    - Target: Attributes -> One attribute à Distribution.

Yêu cầu 7. Thống kê trung bình số tiền trên mỗi giao dịch so với mức thu nhập của khách hàng.

- Bước 1: Khái quát lại yêu cầu: Thống kê tính thương số giữa “Total_Trans_Amt” và “Total_Trans_Ct", gom nhóm dựa trên “Income_Category”
- Bước 2: Phân rã tác vụ:
  - Thêm thuộc tính “avg_trans_amt”:
    - High level: Analyze -> Produce -> Derive
  - Lần lượt tìm kiếm các item thoả từng giá trị phân biệt của thuộc tính “Income_Category" và tính trung bình giá trị thuộc tính “Total_Trans_Amt".
    - Mid level: Search và Browse
  - Tổng hợp các giá trị phân biệt của thuộc tính “Education_level" và “Total_Trans_Amt”, “Total_Trans_Ct" cùng kết quả tính được tương ứng:
    - Low level: Query -> Summarize
    - Target: Attribute -> One attribute -> Distribution

Yêu cầu 9.Thống kê số lượng khách hàng theo từng độ tuổi?

- Bước 1: Khái quát lại yêu cầu: Thống kê số lượng khách hàng “CLIENT_NUM” và gom nhóm dựa trên độ tuổi (Customer_Age)
- Bước 2: Phân rã tác vụ:

  - Thêm thuộc tính “number_of_customer”:
    - High level: Analyze -> produce -> Derive

- Lần lượt tìm các item thỏa từng giá trị phân biệt của từng độ tuổi và đếm số item có thuộc tính "CLIENT_NUM"
  - Mid level: Search -> Browse
- Tổng hợp các giá trị phân biệt của từng độ tuổi và "CLIENT_NUM" 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: Độ tương quan giữa quyết định rời đi/ở lại của khách hàng với số lượng giao dịch đã thực hiện và tổng số tiền đã giao dịch.

Idiom	**Scatter-plot**
Data	• Total_Trans_Amt: quantitative • Total_Trans_Ct: quantitative • Attrition_flag: categorical • Keys: 0 → Scatter plot

Encode	• Mark: point
	• Channel:
	  ○ Total_Trans_Amt: Vertical spatial position (trục y)
	  ○ Total_Trans_Ct: Horizontal spatial postion (trục x)
	  ○ Attrition_flag: Color Hue
	• Align: Không
	• Sort: Không

Manipulate	• Selection → Highlight (change color to Red và phóng to) • Navigate → Attribute Reduction → Slice (Attrition_flag)
Task	Correlation
Scalability	Số item được biểu hiện: 10127

Yêu cầu 4: Tìm trình độ học vấn có tổng số lượng giao dịch nhiều nhất trong 12 tháng gần nhất?

Idiom	**Bar chart**
Data	• Total_Trans_Ct _Sum: quantitative • Education_level: categorical • Keys: 1(Education_level) → Bar chart
Encode	Mark: Line

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

Task	Compare, lookup value
Scalability	Số item được biểu đạt trên biểu đồ: 7

Yêu cầu 5: Thống kê tỷ lệ khách hàng sử dụng thẻ tín dụng dựa trên mức thu nhập của họ.

Idiom	**Pie chart**
Data	Income_Category: categorical

Encode	• Mark: area
	• Channel:
	  ○ Income_Category:: color
	  ○ Income_Category:: angle

Task	Distribution
Scalability	Số item được biểu đạt trên biểu đồ: 5

Yêu cầu 7: Thống kê trung bình số tiền trên mỗi giao dịch so với mức thu nhập của khách hàng.

Idiom	**Bar chart**
Data	Income_Category: categorical
Encode	• Mark: area • Channel: ○ Income_Category:: length
Task	Distribution
Scalability	Số item được biểu đạt trên biểu đồ: 5

Yêu cầu 9: Thống kê số lượng khách hàng theo từng độ tuổi?

Idiom	**Pie chart**
Data	Customer_Age: categorical

Encode	• Mark: area
	• Channel:
	  ○ Customer_Age: color
	  ○ Customer_Age: angle

Task	Distribution
Scalability	Số item được biểu đạt trên biểu đồ: 6

Biểu đồ trực quan

Xem trong source code đính kèm (Link Github dự phòng: Click here)

Nhận xét:

Yêu cầu 1:

- Nguyên tắc biểu đạt:
  - Mục tiêu của biểu đồ là thể hiện độ tương quan giữa 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,…
- 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 (Total_Trans_amt, Total_Trans_ct)
    - Color hue là kênh hiệu quả dành cho thuộc tính phân loại (attrition_flag)
  - Discriminability:
    - Số lượng item trên biểu đồ quá nhiều dẫn đến sự chồng chất → Một vài mark bị đè lên → 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 attrited Customer có số lượng ít thì được tô màu cam để làm nổi bật hơn so với các điểm thuộc phân loại còn lại.

Yêu cầu 4:

- Nguyên tắc biểu đạt:
  - Mục tiêu của biểu đồ là thể sự so sánh giữa số lượng các giao dịch → Dùng biểu đồ cột (bar chart) là hợp lý nhất vì có thể giúp người xem dễ dàng so sánh số lượng của các giao dịch, dễ dàng tìm ra max cũng như min của thuộc tính key.
- Nguyên tắc hiệu quả:
  - Accuracy (độ chính xác)
    - Align: vertical position (trục y) là cách hiệu quả để so sánh mức độ, bên cạnh đó, biểu đồ hỗ trợ tương tác để khi hover có thể nhìn thấy giá trị chính xác của cột.
  - Discriminability:

        - Các item (bar) được phân biệt rõ ràng nhờ tách biệt về vị trí

    - Visual popout:
        - Sử dụng màu giúp làm nổi bật các cột, hỗ trợ đổi màu(làm mờ cột) khi hover vào cột trong biểu đồ.

Yêu cầu 5:

- Nguyên tắc biểu đạt: Mục tiêu của biểu đồ là thể hiện sự phân phối của dữ liệu trong biến Income_category. Sử dụng biểu đồ pie_chart là phù hợp vì nó thường được sử dụng để biểu diễn tỷ lệ phần trăm hoặc phân phối của các mục trong một tập dữ liệu.
- Nguyên tắc hiệu quả:
    - Accuracy (độ chính xác):
        - Biểu đồ pie chart sử dụng kênh diện tích (area) của các phần để biểu thị tỷ lệ phần trăm của mỗi mức thu nhập. Kênh này có thể giúp người đọc nhận biết được sự khác biệt trong phân phối các mức thu nhập.
        - Color (màu sắc) và angle (góc) cũng được sử dụng để phân biệt các phần trong biểu đồ. Màu sắc được sử dụng để định danh các nhóm tuổi khác nhau, trong khi góc của các phần thể hiện tỷ lệ phần trăm tương ứng của mỗi nhóm.
    - Discriminability:
        - Các phần trong biểu đồ pie chart được phân biệt rõ ràng nhờ sự khác biệt về diện tích, màu sắc và góc. Điều này giúp người đọc dễ dàng nhận ra tỷ lệ phần trăm của mỗi mức thu nhập.
    - Separability:
        - Biểu đồ sử dụng các channel position và color để biểu thị dữ liệu. Không có tương tác lý thuyết giữa các channel này, tuy nhiên, trong trường hợp số lượng item trên biểu đồ là nhỏ (trong trường hợp này là 5), việc sử dụng cả hai kênh position và color vẫn đảm bảo tính phân tách giữa các phần.
    - Visual popout:
        - Màu sắc được sử dụng trong biểu đồ để làm nổi bật các phần. Màu sắc bão hòa và sự tương phản giữa các màu giúp các phần trở nên dễ nhìn thấy và nổi bật.

Yêu cầu 6:

- Nguyên tắc biểu đạt:

  - Mục tiêu của biểu đồ là thể sự so sánh giữa số lượng các đánh giá của từng loại phòng à Dùng biểu đồ cột (bar chart) là hợp lý nhất vì có thể giúp người xem dễ dàng so sánh số lượng của các đánh giá, dễ dàng tìm ra max cũng như min của thuộc tính key.
- Nguyên tắc hiệu quả:
  - Accuracy (độ chính xác)
    - Align: vertical position (trục y) là cách hiệu quả để so sánh mức độ, bên cạnh đó, biểu đồ hỗ trợ tương tác để khi hover có thể nhìn thấy giá trị chính xác của cột.
  - Discriminability:
    - Các item (bar) được phân biệt rõ ràng nhờ tách biệt về vị
  - Separability:
    - Các loại phòng trong từng khu vực được sử dụng các màu khác nhau, điều này làm tăng khả năng nhận biết giữa các loại phòng riêng biệt
  - Visual popout:
    - Sử dụng màu giúp làm nổi bật các cột, mỗi loại phòng tương ứng 1 màu.

Yêu cầu 7:

- Nguyên tắc biểu đạt:
  - Mục tiêu của biểu đồ là thể sự so sánh giữa số tiền trung bình trên mỗi giao dịch → Dùng biểu đồ cột (bar chart) là hợp lý nhất vì có thể giúp người xem dễ dàng so sánh số lượng của các giao dịch, dễ dàng tìm ra max cũng như min của thuộc tính key.
- Nguyên tắc hiệu quả:
  - Accuracy (độ chính xác)
    - Align: vertical position (trục y) là cách hiệu quả để so sánh mức độ, bên cạnh đó, biểu đồ hỗ trợ tương tác để khi hover có thể nhìn thấy giá trị chính xác của cột.
  - Discriminability:
    - Các item (bar) được phân biệt rõ ràng nhờ tách biệt về vị trí
  - Visual popout:
    - Sử dụng màu giúp làm nổi bật các cột, hỗ trợ đổi màu(làm mờ cột) khi hover vào cột trong biểu đồ.

Yêu cầu 9:

- Nguyên tắc biểu đạt: Mục tiêu của biểu đồ là thể hiện phân phối của dữ liệu trong biến Customer_Age. Sử dụng biểu đồ pie chart là phù hợp vì nó thường được sử dụng để biểu diễn tỷ lệ phần trăm hoặc phân phối của các mục trong một tập dữ liệu.
- Nguyên tắc hiệu quả:
    - Accuracy (độ chính xác): Biểu đồ pie chart sử dụng kênh diện tích (area) của các phần để biểu thị tỷ lệ phần trăm của mỗi nhóm tuổi. Kênh này có thể giúp người đọc nhận biết được sự khác biệt trong phân phối các nhóm tuổi.
    - Color (màu sắc) và angle (góc) cũng được sử dụng để phân biệt các phần trong biểu đồ. Màu sắc được sử dụng để định danh các nhóm tuổi khác nhau, trong khi góc của các phần thể hiện tỷ lệ phần trăm tương ứng của mỗi nhóm.
    - Discriminability: Các phần trong biểu đồ pie chart được phân biệt rõ ràng nhờ sự khác biệt về diện tích, màu sắc và góc. Điều này giúp người đọc dễ dàng nhận ra tỷ lệ phần trăm của mỗi nhóm tuổi.
    - Separability: Biểu đồ sử dụng các channel position và color để biểu thị dữ liệu. Không có tương tác lý thuyết giữa các channel này, tuy nhiên, trong trường hợp số lượng item trên biểu đồ là nhỏ (trong trường hợp này là 6), việc sử dụng cả hai kênh position và color vẫn đảm bảo tính phân tách giữa các phần.
    - Visual popout: Màu sắc được sử dụng trong biểu đồ để làm nổi bật các phần. Màu sắc bão hòa và sự tương phản giữa các màu giúp các phần trở nên dễ nhìn thấy và nổi bật.
- Tổng quan về mã code j3: Mã code j3 sử dụng thư viện D3.js để tạo biểu đồ pie chart. Nó đọc dữ liệu từ một URL (trong trường hợp này là URL tới file CSV) và sau đó tính toán phần trăm của mỗi nhóm tuổi. Dữ liệu được sắp xếp và truyền cho pie generator của D3 để tạo các cung pie tương ứng với phần trăm. Mỗi cung pie được vẽ bằng cách sử dụng arc generator và được tô màu theo một màu được xác định trước. Biểu đồ cũng bao gồm chú thích và tooltip để cung cấp thông tin chi tiết cho người dùng khi di chuột qua các phần của biểu đồ.
    - Tổng thể, mã code j3 cung cấp một cách đơn giản và hiệu quả để tạo biểu đồ pie chart và biểu diễn phân phối độ tuổi trong tập dữ liệu.