ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN THỰC HÀNH

MÔN: TRỰC QUAN HOÁ DỮ LIỆU

GVHD: Tiết Gia Hồng

Phạm Minh Tú

Thành phố Hồ Chí Minh, tháng 12 năm 2023

Mục lục

A. Thông tin nhóm và phân công công việc

NHÓM 8			
MSSV	**Họ tên**	**Công việc**	**Đánh giá**
20127350	Phan Thanh Thúy	Data profiling	100%
		Data abstraction	
		Task abstraction	
		Thiết kế idiom và charts	
21127234	Nguyễn Lê Anh Chi	Data profiling	100%
		Data abstraction	
		Task abstraction	
		Thiết kế idiom và charts	
21127235	Nguyễn Xuân Quỳnh Chi	Data profiling	100%
		Data abstraction	
		Task abstraction	
		Thiết kế idiom và charts	

B. Yêu cầu đồ án

Loại bài tập	**• Lý thuyết □ Thực hành • Bài tập □ Đồ án**
Ngày bắt đầu	**04/10/2023**
Ngày kết thúc	**05/01/2024**

Với tập dữ liệu tự xác định, thực hiện các yêu cầu sau:

- Profiling tập dữ liệu trên và đưa ra nhận xét về tập dữ liệu này.
- Thực hiện giai đoạn data abstraction
- Đưa ra các câu hỏi cần khai thác trên tập dữ liệu
- Thực hiện giai đoạn task abstraction
- Thiết kế các Idiom
- Sử dụng D3.js để cài đặt thiết kế Idiom
- Đánh giá biểu đồ đã cài

C. Kết quả đồ án

I. Data profiling

Field Name	NULL	Missing	Actual	Completeness	Cardinality	Uniqueness	Distinctness
InvoiceID	0	0	1016	100.00%	1016	100.00%	100.00%
Branch	0	0	1016	100.00%	3	0.30%	0.30%
Customer type	0	0	1016	100.00%	2	0.20%	0.20%
Gender	0	0	1016	100.00%	2	0.20%	0.20%
ProductID	0	0	1016	100.00%	943	92.81%	92.81%
Quantity	0	0	1016	100.00%	10	0.98%	0.98%
Tax 5%	0	0	1016	100.00%	990	97.44%	97.44%
Total	0	0	1016	100.00%	990	97.44%	97.44%
Date	0	0	1016	100.00%	103	10.14%	10.14%
Time	0	0	1016	100.00%	506	49.80%	49.80%
Payment	0	0	1016	100.00%	3	0.30%	0.30%
Cogs	0	0	1016	100.00%	990	97.44%	97.44%
Gross margin percentage	0	0	1016	100.00%	1	0.10%	0.10%
Gross income	0	0	1016	100.00%	990	97.44%	97.44%
Rating	0	0	1016	100.00%	61	6.00%	6.00%

Đánh giá:

- Giá trị null và missing không tồn tại ở tất cả các thuộc tính.
- Có tất cả 1016 hoá đơn được xuất ra. Mỗi hoá đơn có một mã hoá đơn riêng, không trùng nhau.
- Có tất cả 943 sản phẩm khác nhau được bán ra. Mỗi sản phẩm có một mã sản phẩm duy nhất.
- Hoá đơn được lập đến từ một trong 3 chi nhánh: A, B, C.
- Có 2 loại khách hàng: Khách hàng thành viên (Member) và Khách hàng bình thường (Normal).
- Khách hàng chỉ có thể thuộc về giới tính nam hoặc nữ.
- Có 3 phương thức thanh toán: Cash, Credit Card, Ewallet.

- Các ngày ghi nhận hóa đơn chỉ nằm trong 3 tháng: Một, Hai, Ba trong 2 năm là 2019 và 2020
- Số lượng sản phẩm của 1 hóa đơn chỉ nằm trong khoảng từ 1 đến 10.
- Thang điểm đánh giá nằm từ khoảng 4.0 đến 10.0.
- Tỷ số lợi nhuận gộp giống nhau ở mọi hóa đơn.

1. Thuộc tính 1: InvoiceID

Input Metadata	
Field Name	InvoiceID
Field Data Type	CHAR
Field Length	11
Data Profiling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	1016
Uniqueness	100.00%
Distinctness	100.00%
Data Profiling Additional Statistics	
Field Data Types	1
Field Length (MIN)	11
Field Length (MAX)	11
Field Value (MIN)	NULL
Field Value (MAX)	NULL
Field Formats	1

InvoiceID (Field Data Types)	Count	Percentage
CHAR	1016	100.00%
InvoiceID (Field Formats)	**Count**	**Percentage**
nnn-nn-nnnn	1016	100.00%

2. Thuộc tính 2: Branch

Input Metadata	
Field Name	Branch
Field Data Type	CHAR
Field Length	1
Data Profiling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	3
Uniqueness	0.30%
Distinctness	0.30%
Data Profiling Additional Statistics	
Field Data Types	1
Field Length (MIN)	1
Field Length (MAX)	1
Field Value (MIN)	NULL
Field Value (MAX)	NULL
Field Formats	3

Branch (Field Data Types)	Count	Percentage
CHAR	1016	100.00%
Branch (Field Formats)	**Count**	**Percentage**
A	349	34.35%
B	338	33.27%
C	329	32.38%

3. Thuộc tính 3: Customer type

Input Metadata	
Field Name	Customer type
Field Data Type	CHAR
Field Length	6
Data Profilling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	2
Uniqueness	0.20%
Distinctness	0.20%
Data Profilling Additional Statistics	
Field Data Types	1
Field Length (MIN)	6
Field Length (MAX)	6
Field Value (MIN)	NULL
Field Value (MAX)	NULL
Field Formats	2

Customer type (Field Data Types)	Count	Percentage
CHAR	1016	100.00%
Customer type (Field Formats)	**Count**	**Percentage**
Member	507	49.90%
Normal	509	50.10%

4. Thuộc tính 4: Gender

Input Metadata	
Field Name	Gender
Field Data Type	VARCHAR
Field Length	6
Data Profilling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	2
Uniqueness	0.20%
Distinctness	0.20%
Data Profilling Additional Statistics	
Field Data Types	1
Field Length (MIN)	6
Field Length (MAX)	4
Field Value (MIN)	NULL
Field Value (MAX)	NULL
Field Formats	2

Gender (Field Data Types)	Count	Percentage
VARCHAR	1016	100.00%
Gender (Field Formats)	**Count**	**Percentage**
Female	507	49.90%
Male	509	50.10%

5. Thuộc tính 5: ProductID

Input Metadata	
Field Name	ProductID
Field Data Type	VARCHAR
Field Length	6
Data Profilling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	943
Uniqueness	92.81%
Distinctness	92.81%
Data Profilling Additional Statistics	
Field Data Types	1
Field Length (MIN)	6
Field Length (MAX)	5
Field Value (MIN)	NULL
Field Value (MAX)	NULL
Field Formats	2

ProductID (Field Data Types)	Count	Percentage
VARCHAR	1016	100.00%
ProductID (Top 3 Field Values)	**Count**	**Percentage**
PID897	3	0.30%
PID943	3	0.30%
PID942	2	0.20%
ProductID (Field Formats)	**Count**	**Percentage**
PIDnn	102	10.04%
PIDnnn	914	89.96%

6. Thuộc tính 6: Quantity

Input Metadata	
Field Name	Quantity
Field Data Type	CHAR
Field Length	2
Data Profiling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	10
Uniqueness	0.98%
Distinctness	0.98%
Data Profiling Additional Statistics	
Field Data Types	1
Field Length (MIN)	2
Field Length (MAX)	1
Field Value (MIN)	1
Field Value (MAX)	10
Field Formats	2

Quantity (Field Data Types)	Count	Percentage
CHAR	1016	100.00%
Quantity (Top 5 Values)	**Count**	**Percentage**
10	121	11.91%
1	113	11.12%
4	111	10.93%
5	105	10.33%
7	103	10.14%
Quantity (Field Formats)	**Count**	**Percentage**
n	895	88.09%
nn	121	11.91%

7. Thuộc tính 7: Tax 5%

Input Metadata	
Field Name	Tax 5%
Field Data Type	FLOAT
Field Length	7
Data Profilling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	990
Uniqueness	97.44%
Distinctness	97.44%
Data Profilling Additional Statistics	
Field Data Types	1
Field Length (MIN)	7
Field Length (MAX)	2
Field Value (MIN)	0.5085
Field Value (MAX)	49.65
Field Formats	1

Tax5% (Field Data Types)	Count	Percentage
FLOAT	1016	100.00%
Tax5% (Top 5 Values)	**Count**	**Percentage**
41.54	2	0.20%
24.1255	2	0.20%
22.428	2	0.20%
21.036	2	0.20%
21.783	2	0.20%
Tax5% (Field Formats)	**Count**	**Percentage**
FLOAT	1016	100.00%

8. Thuộc tính 8: Total

Input Metadata	
Field Name	Total
Field Data Type	FLOAT
Field Length	8
Data Profilling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	990
Uniqueness	97.44%
Distinctness	97.44%
Data Profilling Additional Statistics	
Field Data Types	1
Field Length (MIN)	2
Field Length (MAX)	8
Field Value (MIN)	10.6785
Field Value (MAX)	1042.65
Field Formats	1

Total (Field Data Types)	Count	Percentage
FLOAT	1016	100.00%
Total (Top 5 Values)	**Count**	**Percentage**
87.234	2	0.20%
506.6355	2	0.20%
470.988	2	0.20%
441.756	2	0.20%
457.443	2	0.20%
Total (Field Formats)	**Count**	**Percentage**
FLOAT	1016	100.00%

9. Thuộc tính 9: Date

Input Metadata	
Field Name	Date
Field Data Type	DATE
Field Length	10
Data Profiling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	103
Uniqueness	10.14%
Distinctness	10.14%
Data Profiling Additional Statistics	
Field Data Types	1
Field Length (MIN)	9
Field Length (MAX)	10
Field Value (MIN)	NULL
Field Value (MAX)	NULL
Field Formats	2

Date (Field Data Types)	Count	Percentage
DATE	1016	100.00%
Date (Top 5 Values)	**Count**	**Percentage**
07/02/2019	20	1.97%
15/02/2019	19	1.87%
08/01/2019	18	1.77%
14/03/2019	18	1.77%
02/03/2019	18	1.77%
Date (Field Formats)	**Count**	**Percentage**
dd/mm/yyyy	1016	100.00%

10. Thuộc tính 10: Time

Input Metadata	
Field Name	Time
Field Data Type	TIME
Field Length	5
Data Profiling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	506
Uniqueness	49.80%
Distinctness	49.80%
Data Profiling Additional Statistics	
Field Data Types	1
Field Length (MIN)	5
Field Length (MAX)	5
Field Value (MIN)	NULL
Field Value (MAX)	NULL
Field Formats	1

Time (Field Data Types)	Count	Percentage
TIME	1016	100.00%
Time (Top 5 Values)	**Count**	**Percentage**
14:42	7	0.69%
19:48	7	0.69%
17:38	6	0.59%
17:36	6	0.59%
19:20	6	0.59%
Time (Field Formats)	**Count**	**Percentage**
hh:mm	1016	100.00%

11. Thuộc tính 11: Payment

Input Metadata	
Field Name	Payment
Field Data Type	VARCHAR
Field Length	11
Data Profilling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	3
Uniqueness	0.30%
Distinctness	0.30%
Data Profilling Additional Statistics	
Field Data Types	1
Field Length (MIN)	4
Field Length (MAX)	11
Field Value (MIN)	NULL
Field Value (MAX)	NULL
Field Formats	3

Payment (Field Data Types)	Count	Percentage
VARCHAR	1016	100.00%
Payment (Field Formats)	**Count**	**Percentage**
Cash	348	34.25%
Credit card	316	31.10%
Ewallet	352	34.65%

12. Thuộc tính 12: Cogs

Input Metadata	
Field Name	cogs
Field Data Type	FLOAT
Field Length	3
Data Profilling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	990
Uniqueness	97.44%
Distinctness	97.44%
Data Profilling Additional Statistics	
Field Data Types	1
Field Length (MIN)	2
Field Length (MAX)	6
Field Value (MIN)	993
Field Value (MAX)	10.17
Field Formats	1

cogs (Field Data Types)	Count	Percentage
FLOAT	1016	100.00%
cogs (Field Formats)	**Count**	**Percentage**
FLOAT	1016	100.00%
cogs (Top 5 Field Values)	**Count**	**Percentage**
993	1	0.10%
989.8	1	0.10%
985.2	1	0.10%
975	1	0.10%
973.8	1	0.10%
cogs (Top Popular Field Values)	**Count**	**Percentage**
33.5	2	0.20%
66.4	2	0.20%
80.6	2	0.20%
83.1	2	0.20%
89.3	2	0.20%
102.0	2	0.20%
164	2	0.20%
167.5	2	0.20%
172.8	2	0.20%
180.1	2	0.20%
206.5	2	0.20%
207.3	2	0.20%
234.8	2	0.20%
251.4	2	0.20%
263.8	2	0.20%
263.9	2	0.20%
265.9	2	0.20%
420.7	2	0.20%
430.2	2	0.20%
431.9	2	0.20%
435.7	2	0.20%
448.6	2	0.20%
482.5	2	0.20%
562.3	2	0.20%
713.8	2	0.20%
789.6	2	0.20%

13. Thuộc tính 13: Gross margin percentage

Input Metadata	
Field Name	gross margin percentage
Field Data Type	FLOAT
Field Length	11
Data Profilling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	1
Uniqueness	0.10%
Distinctness	0.10%
Data Profilling Additional Statistics	
Field Data Types	1
Field Length (MIN)	11
Field Length (MAX)	11
Field Value (MIN)	4.761904762
Field Value (MAX)	4.761904762
Field Formats	1

gross margin percentage (Field Data Types)	Count	Percentage
FLOAT	1016	100.00%
gross margin percentage (Field Formats)	**Count**	**Percentage**
FLOAT	1016	100.00%

14. Thuộc tính 14: Gross income

Input Metadata	
Field Name	gross income
Field Data Type	FLOAT
Field Length	7
Data Profilling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	990
Uniqueness	97.44%
Distinctness	97.44%
Data Profilling Additional Statistics	
Field Data Types	1
Field Length (MIN)	2
Field Length (MAX)	7
Field Value (MIN)	0.5085
Field Value (MAX)	49.65
Field Formats	1

gross income (Field Data Types	Count	Percentage
FLOAT	1016	100.00%
gross income (Field Formats)	**Count**	**Percentage**
FLOAT	1016	100.00%
gross income (Top Popular Fiel	**Count**	**Percentage**
1.68	2	0.20%
3.32	2	0.20%
4.03	2	0.20%
4.15	2	0.20%
4.46	2	0.20%
5.1	2	0.20%
8.2	2	0.20%
8.38	2	0.20%
8.64	2	0.20%
9	2	0.20%
10.33	2	0.20%
10.36	2	0.20%
11.74	2	0.20%
12.57	2	0.20%
13.19	2	0.20%
13.2	2	0.20%
13.29	2	0.20%
21.04	2	0.20%
21.51	2	0.20%
21.6	2	0.20%
21.78	2	0.20%
22.43	2	0.20%
24.13	2	0.20%
28.12	2	0.20%
35.69	2	0.20%
39.48	2	0.20%

15. Thuộc tính 15: Rating

Input Metadata	
Field Name	Rating
Field Data Type	FLOAT
Field Length	3
Data Profilling Summary Statistics	
NULL	0
Missing	0
Actual	1016
Completeness	100.00%
Cardinality	61
Uniqueness	6.00%
Distinctness	6.00%
Data Profilling Additional Statistics	
Field Data Types	1
Field Length (MIN)	1
Field Length (MAX)	3
Field Value (MIN)	4
Field Value (MAX)	10
Field Value (AVERAGE)	6.963
Field Formats	1

Rating (Field Data Types)	Count	Percentage
FLOAT	1016	100.00%
Rating (Field Formats)	**Count**	**Percentage**
FLOAT	1016	100.00%
Rating (Top 5 Popular Values)	**Count**	**Percentage**
6.00	27	2.66%
6.60	24	2.36%
4.20	22	2.17%
5.10	22	2.17%
9.50	22	2.17%

II. Data abstraction

- Dataset name: supermarket_sales
- Dataset type: Table
- Item: Mỗi item là thông tin chi tiết của một hoá đơn mua hàng trong tháng Một, Hai và Ba của năm 2019 và 2020.
- Dataset availability: Static
- Number of attributes: 15

	Semantics	Attribute type	Direction	Hierarchical	Discrete/ Continuous	Interval/ Ratio	Bin
InvoiceID	Mã hoá đơn	Categorical	x	Yes	Discrete	x	1
Branch	Chi nhánh xuất hoá đơn	Categorical	x	No	Discrete	x	3

	Semantics	Attribute type	Direction	Hierarchical	Discrete/ Continuous	Interval/ Ratio	Bin
Customer type	Loại khách hàng	Categorical	x	No	Discrete	x	2
Gender	Giới tính của khách hàng	Categorical	x	No	Discrete	x	2
ProductID	Mã sản phẩm	Categorical	x	No	Discrete	x	1
Quantity	Số lượng sản phẩm	Quantitative	x	No	Continuous	Ratio	10
Tax 5%	Phí tax	Quantitative	x	No	Continuous	Ratio	1
Total	Tổng tiền đã bao gồm tax	Quantitative	x	No	Continuous	Ratio	1
Date	Ngày mua hàng	Ordinal	Cyclic	Yes	Discrete	x	3
Time	Thời gian mua hàng	Ordinal	Cyclic	Yes	Discrete	x	24
Payment	Phương thức thanh toán	Categorical	x	No	Discrete	x	3
Cogs	Chi phí hàng bán (chi phí thực tế để sản xuất hoặc mua sản phẩm)	Quantitative	x	No	Continuous	Ratio	30
Gross margin percentage	Tỷ suất lợi nhuận gộp (lợi nhuận gộp / doanh số bán hàng)	Quantitative	Diverging	No	Continuous	Ratio	1
Gross income	Lợi nhuận gộp (doanh số bán hàng - chi phí hàng bán)	Quantitative	x	No	Continuous	Ratio	30
Rating	Điểm đánh giá chất lượng của giao dịch	Ordinal	Sequential	No	Discrete	x	10

III. Task abstraction

1. Task 1: Tìm chi nhánh có số lượng hoá đơn được xuất ra nhiều nhất.

Action	High-level: Analyze	Produce
	Mid-level: Search	Explore
	Low-level: Query	Identify

Idiom: Derive + Encode + Annotate

- Biểu đồ: Bar chart
- Mô tả biểu đồ:
    - Trục hoành là tên các chi nhánh.
    - Trục tung là số lượng hoá đơn mà chi nhánh đã xuất ra.

2. Task 2: So sánh số lượng khách hàng thành viên và khách hàng bình thường của chi nhánh A.

Action	High-level: Analyze	Produce
	Mid-level: Search	Browse
	Low-level: Query	Compare

Idiom: Derive + Encode

- Biểu đồ: Bar chart
- Mô tả biểu đồ:
    - Trục hoành là số lượng khách hàng mua hàng.
    - Trục tung là loại khách hàng.

3. Task 3: Thống kê doanh thu theo từng tháng của các năm.

Action	High-level: Analyze	Consume
	Mid-level: Search	Browse
	Low-level: Query	Summarize

Idiom: Present + Encode + Annotate

- Biểu đồ: Line chart
- Mô tả biểu đồ:
    - Trục hoành là tên tháng và năm.
    - Trục tung là doanh thu ở các mốc thời gian.

4. **Task 4: Tìm phương thức thanh toán được sử dụng nhiều nhất.**

Action	High-level: Analyze	Produce
	Mid-level: Search	Explore
	Low-level: Query	Identify

Idiom: Derive + Encode + Annotate

- Biểu đồ: Bar chart
- Mô tả biểu đồ:
    - Trục hoành là phương thức thanh toán.
    - Trục tung là số lần sử dụng của các phương thức thanh toán.

5. **Task 5: Thống kê lượt đánh giá chất lượng đơn hàng.**

Action	High-level: Analyze	Produce
	Mid-level: Search	Explore
	Low-level: Query	Identify

Idiom: Derive + Encode + Binning

- Biểu đồ: Bar chart
- Mô tả biểu đồ:
    - Trục hoành là điểm đánh giá đơn hàng (rating).
    - Trục tung là số lượng khách đánh giá đơn hàng.

6. **Task 6: Xem lợi nhuận gộp qua các năm (Gross Income).**

Action	High-level: Analyze	Produce
	Mid-level: Search	Browse
	Low-level: Query	Compare

Idiom: Derive + Encode + Annotate

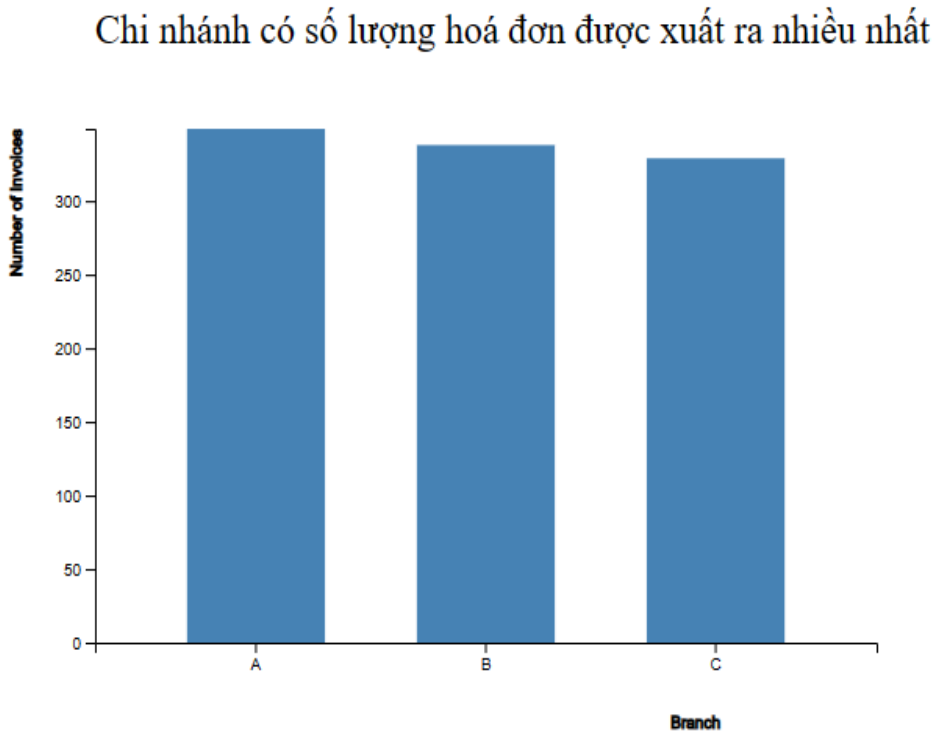
- Biểu đồ: Bar chart
- Mô tả biểu đồ:
    - Trục hoành là thời điểm bán hàng.
    - Trục tung là giá trị của Gross Income theo từng mốc thời gian (trung bình).

IV. Thiết kế idiom và cài đặt

Link github:

1. Task 1: Tìm chi nhánh có số lượng hoá đơn được xuất ra nhiều nhất.

Idiom: Bar chart



Đánh giá dữ liệu:

- Sự chênh lệch số hóa đơn giữa các chi nhánh không lớn, nhưng có thể dễ thấy được chi nhánh A là chi nhánh xuất được nhiều hóa đơn nhất.
- Từ dữ liệu, ta có thể đánh giá siêu thị có sự ổn định khách hàng ở tất cả các chi nhánh được ghi nhận.

Đánh giá tính biểu đạt:

- Phương thức Bar chart là một lựa chọn tương đối hợp lí trong trường hợp này vì có cho người xem có cái nhìn trực quan hơn về sự chênh lệch giữa số hóa đơn ở từng chi nhánh.
- Tuy nhiên, nếu để đưa ra đánh giá tổng quát thì Pie Chart có thể là một lựa chọn tốt hơn.

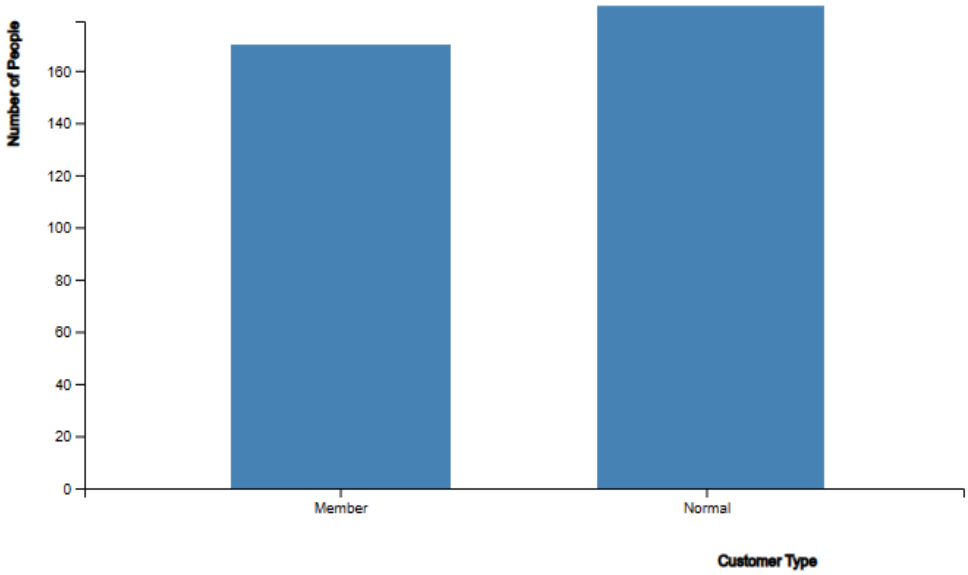
Đánh giá về hiệu quả:

- Về độ chính xác: Vì dữ liệu tương đối đồng đều và không có sự chênh lệch quá lớn nên không có thuộc tính nào khó quan sát.
- Về sự phân biệt: Khi nhìn vào biểu đồ ta dễ dàng phân biệt được ba chi nhánh A, B, C với nhau không bị lẫn lộn.
- Về sự phân tách: Biểu đồ sử dụng một mình bar chart nên có sự phân tách, không có sự ảnh hưởng lẫn nhau.
- Về khả năng biểu diễn: Có thể thấy được sự chênh lệch giữa các chi nhánh và từ đó tìm ra được chi nhánh xuất nhiều hóa đơn nhất.
- Về tính nổi bật: Sự chênh lệch giữa các chi nhánh được thể hiện một cách trực quan.

2. **Task 2: So sánh số lượng khách hàng thành viên và khách hàng bình thường của chi nhánh A.**

Idiom: Bar chart

Số lượng thành viên và khách hàng bình thường của chi nhánh A.



Đánh giá dữ liệu:

- Dễ dàng so sánh được sự khác nhau giữa 2 loại khách hàng thuộc chi nhánh A.
- Từ dữ liệu, ta có thể đánh giá được số khách hàng thông thường tương đối nhiều hơn khách hàng thành viên ở chi nhánh A.

Đánh giá tính biểu đạt:

- Sử dụng Bar Chart giúp người xem dễ so sánh sự khác nhau giữa 2 loại khách hàng ở chi nhánh A hơn với số liệu cụ thể.

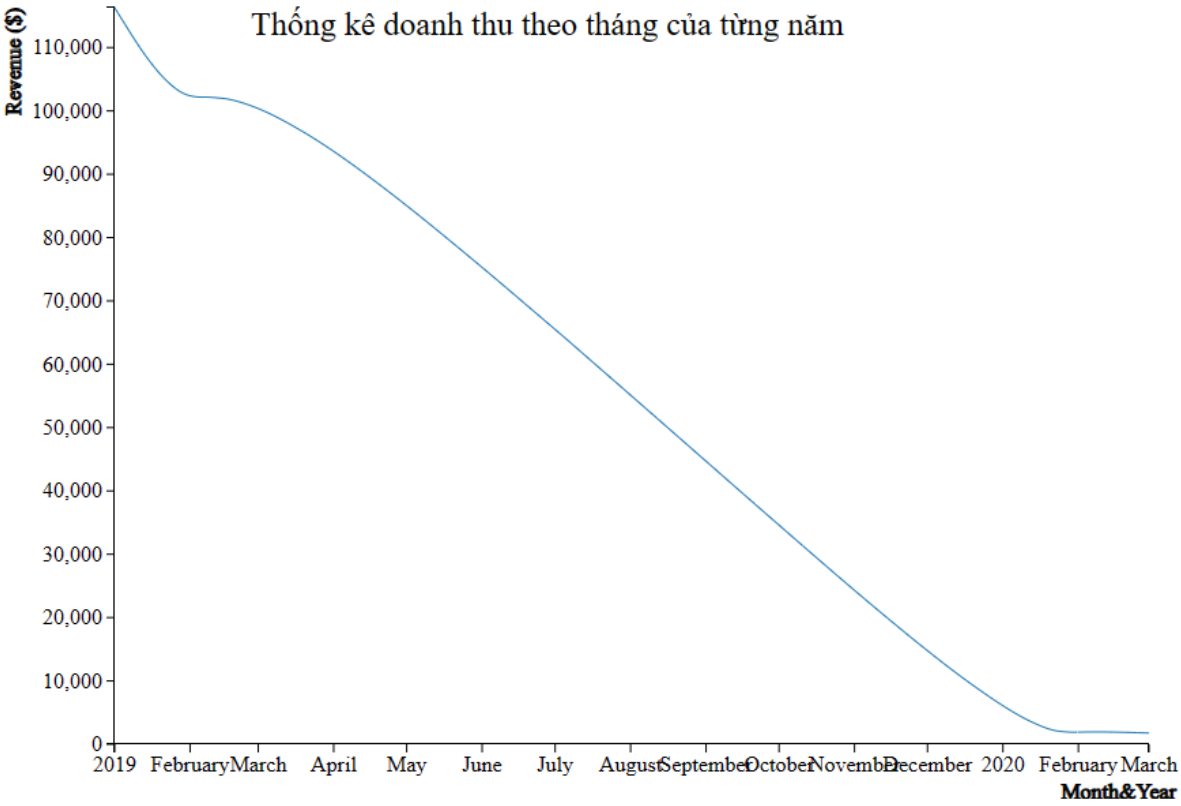
Đánh giá về hiệu quả:

- Về độ chính xác: Vì chỉ có 2 phân loại khách hàng và số liệu của 2 phân loại này không có sự cách biệt quá lớn nên dữ liệu được hiển thị khá chính xác.
- Về sự phân biệt: Có sự phân biệt rõ ràng do chỉ cần so sánh giữa 2 phân loại.
- Về sự phân tách: Biểu đồ sử dụng một mình Bar Chart nên có sự phân tách, không có sự ảnh hưởng lẫn nhau.

- Về khả năng biểu diễn: Có thể thấy được sự chênh lệch giữa hai phân loại khách hàng ở chi nhánh A.
- Về tính nổi bật: Dễ dàng thấy được sự chênh lệch giữa 2 phân loại khách hàng với số liệu cụ thể.

3. Task 3: Thống kê doanh thu theo từng tháng của các năm.

Idiom: Line chart



Đánh giá dữ liệu:

- Dữ liệu có sự chênh lệch lớn giữa năm 2019 và 2020 nên khó nhận thấy được sự cách biệt cụ thể.
- Từ dữ liệu, ta có thể thống kê được doanh thu của quý đầu tiên trong 2 năm 2019 và 2020.

Đánh giá tính biểu đạt:

- Sử dụng Line Chart có thể thể hiện được sự chênh lệch theo thời gian và là sự lựa chọn tốt nhất cho dạng câu hỏi này.
- Tuy nhiên, do thời gian ghi nhận là quý đầu của mỗi năm nên dẫn đến phần giữa bị trống và không thật sự khai thác được những lợi ích của Line Chart.

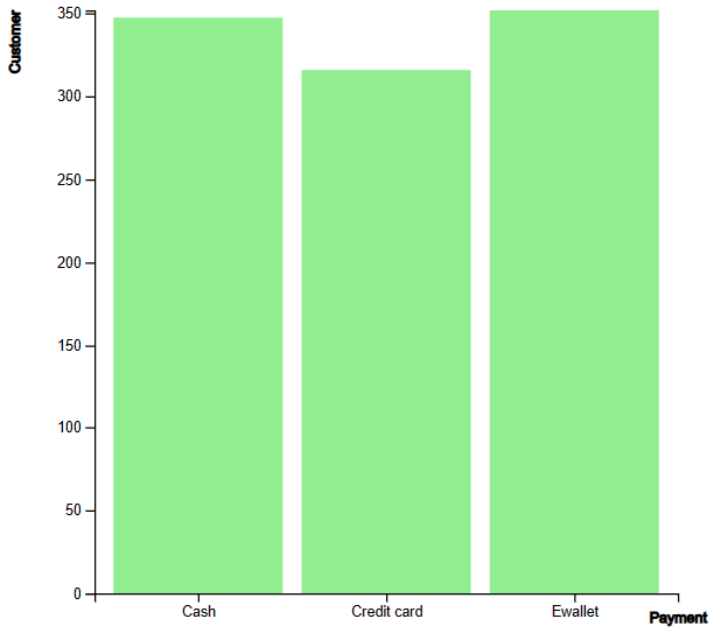
Đánh giá về hiệu quả:

- Về độ chính xác: Do có số liệu của năm 2019 gần gấp 10 lần số liệu năm 2020 nên khó quan sát được chính xác dữ liệu của năm 2020.
- Về sự phân biệt: Ở giữa khoảng thời gian có một khoảng trống lớn do dữ liệu chỉ nằm ở 2 đầu, dẫn đến các mốc thời gian ở trục hoành có vài chỗ bị đè lên nhau.
- Về sự phân tách: Do chỉ sự một loại biểu đồ là Line chart nên có sự phân tách, không gây ảnh hưởng lẫn nhau.
- Về khả năng biểu diễn: Sự giảm xuống rõ rệt giữa doanh thu của quý đầu tiên ở 2 năm 2019 và 2020 được thể hiện rõ ràng nhưng chênh lệch giữa những tháng trong quý thì khá nhỏ để thấy được chính xác.
- Về tính nổi bật: Thể hiện rõ ràng sự giảm xuống rõ rệt từ tháng 3 năm 2019 qua tháng 1 năm 2020.

4. Task 4: Tìm phương thức thanh toán được sử dụng nhiều nhất.

Idiom: Bar chart

Số người dùng theo phương thức thanh toán



Đánh giá dữ liệu:

- Có thể thấy các phương thức thanh toán được sử dụng khá đồng đều nhau, nhưng Ewallet là phương thức được sử dụng nhiều nhất và Cash cũng không kém là bao.
- Chênh lệch giữa các phương thức không quá lớn.
- Từ dữ liệu, ta có thể phương thức thanh toán bằng ví điện tử đang ngày càng phát triển và được sử dụng rộng rãi nhưng tiền mặt vẫn là lựa chọn không thể thiếu. Credit card có thể chưa phổ biến bằng do quá trình tạo thẻ phức tạp hơn so với ví điện tử.

Đánh giá nguyên lý biểu đạt:

- Để thể hiện rõ ràng và tương quan giữa phương thức thanh toán và số lượng người dùng thì bar chart là cách khá hợp lý vì có thể quan sát rõ từng số lượng người sử dụng phương thức nào.
- Tuy nhiên, ta có thể dùng pie chart để nắm tỷ lệ sử dụng trong tổng thể người dùng để có thể đánh giá rõ hơn.

Đánh giá về hiệu quả:

- Về độ chính xác: có thể dễ dàng nhìn rõ số lượng người sử dụng của các phương thức bằng cách di chuột vào các cột, từ đó dễ dàng kết luận phương thức được sử dụng nhiều nhất.
- Về sự phân biệt: do chỉ có 3 phương thức nên dễ phân biệt trong biểu đồ
- Về sự phân tách: do sử dụng bar chart nên có sự phân tách, không gây ảnh hưởng lẫn nhau trong dữ liệu.
- Về khả năng biểu diễn: có thể nắm được tình hình thanh toán.
- Về tính nổi bật: dễ tìm được phương thức sử dụng nhiều nhất.

5. **Task 5: Thống kê lượt đánh giá chất lượng đơn hàng.**

Idiom: Bar chart

Đánh giá dữ liệu:

- Ta có thể thấy được trong biểu đồ không có sự xuất hiện của điểm đánh giá nhỏ hơn 4, và đa số điểm cũng thường nằm ở mức 5 đến 10. Điều này chứng tỏ được phần lớn khách đã mua hàng đã hài lòng với trải nghiệm của họ.
- Mức điểm nhận được nhiều nhất là 7, và cũng không chênh lệch lắm so với các mức điểm khác từ 5 đến 10.

Đánh giá nguyên lý biểu đạt:

- Có thể dễ dàng thấy được sự chênh lệch giữa giá trị MAX và MIN, và cũng dễ dàng thấy được mức điểm đánh giá nhận được nhiều nhất.
- Vì số lượng đánh giá điểm ở các mức 1 đến 3 là 0 nên cũng bỏ qua để tránh gây ra độ lệch trong biểu đồ. Để thấy rõ được mối tương quan giữa các điểm đánh giá, ta nên bỏ đi cột điểm 4 và vẽ lại biểu đồ, lúc này dữ liệu của ta sẽ đồng đều và trực quan hơn. Tuy nhiên điều này có thể gây ra không rõ ở một vài trường hợp nên không khuyến khích cách này.

Đánh giá về hiệu quả:

- Về độ chính xác: do trong biểu đồ không biểu diễn các điểm đánh giá từ 1 đến 3 nên có thể gây ra sai sót.
- Về sự phân biệt: dễ dàng phân biệt được các mức điểm.
- Về sự phân tách: sử dụng bar chart nên có sự phân tách, không làm gây ảnh hưởng đến nhau.
- Về khả năng biểu diễn: nhìn nhận được tổng thể tình hình điểm đánh giá.
- Về tính nổi bật: dễ dàng thấy được mức điểm nhận được nhiều lượt đánh giá, và cũng dễ trong việc xem các mức điểm khác.

6. Task 6: Xem lợi nhuận gộp qua các năm (Gross Income).

Idiom: Bar chart

Đánh giá dữ liệu:

- Có thể thấy lợi nhuận vào tháng 1 thường cao hơn tháng 2, tháng 3, có thể đây là thời điểm trước Tết hoặc chuyển mùa nên nhu cầu mua sắm cao hơn. Và nhìn chung năm 2020 cũng có lợi nhuận cao hơn so với tổng thể năm 2019, có thể là người dùng đã quen thuộc hơn với thị trường này nên lượng người dùng vào năm 2020 cao hơn, và có thể cao hơn nữa vào năm 2021.
- Từ dữ liệu, ta có thể phân tích hành vi người dùng và điều chỉnh số lượng hàng cho phù hợp với từng thời điểm (ví dụ như nhập hàng nhiều vào giai đoạn chuyển giao mùa, hoặc năm)

Đánh giá tính biểu đạt:

- Việc sử dụng bar chart để xem lợi nhuận gộp ở các thời điểm nhìn chung khá hợp lý, tuy nhiên chưa có sự phân tách rõ ràng giữa mốc thời gian năm, điều này có thể gây nên hiểu lầm là một khoảng thời gian liên tục nếu không nhìn kỹ.
- Dễ dàng chỉ ra khoảng thời gian có lợi nhuận gộp cao nhất, thấp nhất và cũng dễ dàng so sánh được tình hình giữa hai năm 2019 và 2020.

Đánh giá về hiệu quả:

- Về độ chính xác: độ chênh lệch giữa các thời điểm không quá lớn, và cũng giúp ta nắm sơ lược tình hình năm 2019 và 2020.
- Về sự phân biệt: có thể dễ dàng phân biệt các mốc thời gian.
- Về sự phân tách: sử dụng bar chart nên biểu đồ có sự phân tách, không gây ảnh hưởng đến nhau.
- Về khả năng biểu diễn: có thể nắm được tình hình chung về lợi nhuận gộp của quý đầu năm 2019 và 2020.
- Về tính nổi bật: dễ dàng thấy được thời gian cùng lợi nhuận gộp.

D. Nguồn tham khảo

- ☐ Slide và code tham khảo từ ngày seminar.
- ☐ Multi-series line chart from a CSV file of technology stocks (github.com)
- ☐ d3.js - d3 show labels only for ticks with data in a bar chart - Stack Overflow
- ☐ Ticks | D3 by Observable (d3js.org)