

A close-up of a logo

Description automatically generated

Mục lục

[I.Thông tin của nhóm và tiến độ đồ án 3](#_heading=h.gjdgxs)

[II. Giai đoạn 1 Data abstraction 3](#_heading=h.30j0zll)

[A. Data forms 3](#_heading=h.1fob9te)

[B.Data Profiling 6](#_heading=h.3dy6vkm)

[1. Invoice ID 7](#_heading=h.1t3h5sf)

[2. Branch: 7](#_heading=h.4d34og8)

[3. Customer type: 8](#_heading=h.2s8eyo1)

[4. Gender: 8](#_heading=h.17dp8vu)

[5. ProductID: 9](#_heading=h.3rdcrjn)

[6. Quantity: 9](#_heading=h.26in1rg)

[7. Tax 5%: 10](#_heading=h.lnxbz9)

[8. Total: 11](#_heading=h.35nkun2)

[9. Date: 11](#_heading=h.1ksv4uv)

[10. Time: 12](#_heading=h.44sinio)

[11. Payment: 12](#_heading=h.2jxsxqh)

[12. cogs: 13](#_heading=h.z337ya)

[13. gross margin percentage 14](#_heading=h.3j2qqm3)

[14. gross income 14](#_heading=h.1y810tw)

[15. Rating 15](#_heading=h.4i7ojhp)

[16.Đánh giá 15](#_heading=h.2xcytpi)

[III. Giai đoạn 2 Task abstraction 16](#_heading=h.1ci93xb)

[1.Khái niệm: 16](#_heading=h.3whwml4)

[2. Các Task Abstraction: 17](#_heading=h.2bn6wsx)

[So sánh số lượng hóa đơn bán hàng giữa loại khách hàng ‘Normal’ và ‘Member’ theo từng chi nhánh 17](#_heading=h.qsh70q)

[Thống kê tổng doanh thu bán hàng của các chi nhánh 17](#_heading=h.3as4poj)

[Tìm hóa đơn của khách hàng có tổng tiền thanh toán ( Tính cả thuế) cao nhất 17](#_heading=h.1pxezwc)

[Tính đánh giá trung bình cho toàn bộ đơn hàng của chi nhánh B 17](#_heading=h.49x2ik5)

[Sản phẩm nào bán chạy nhất ? 17](#_heading=h.2p2csry)

[So sánh phương thức thanh toán nào được sử dụng nhiều hơn giữa Ewallet và Credit card ? 17](#_heading=h.147n2zr)

[Có sự chênh lệch nào đáng kể giữa mua sắm của nam và nữ không? 17](#_heading=h.3o7alnk)

[So sánh doanh thu giữa 3 tháng đầu năm 2019 và 3 tháng đầu năm 2020? 17](#_heading=h.23ckvvd)

[IV. Giai đoạn 3: Idiom and Implementation 17](#_heading=h.ihv636)

[A. Idiom 17](#_heading=h.32hioqz)

[1. So sánh số lượng hóa đơn bán hàng giữa loại khách hàng ‘Normal’ và ‘Member’ theo từng chi nhánh 17](#_heading=h.8i3jhfxr575)

[2. Thống kê tổng doanh thu bán hàng của các chi nhánh 18](#_heading=h.ozijsytgu5ds)

[3. Tìm hóa đơn của khách hàng có tổng tiền thanh toán ( Tính cả thuế) cao nhất 18](#_heading=h.24w3ky4l1c83)

[4. Tính đánh giá trung bình cho toàn bộ đơn hàng của chi nhánh B 18](#_heading=h.3t5lko6pgp1v)

[5. Sản phẩm nào bán chạy nhất ? 18](#_heading=h.ob0revr7rf0y)

[6. So sánh phương thức thanh toán nào được sử dụng nhiều hơn giữa Ewallet và Credit card ? 19](#_heading=h.nj80wfjl7ebd)

[7. Có sự chênh lệch nào đáng kể giữa mua sắm của nam và nữ không? 19](#_heading=h.mn5m16qqnzo)

[8. So sánh doanh thu giữa 3 tháng đầu năm 2019 và 3 tháng đầu năm 2020? 19](#_heading=h.nf4i83q4rfey)

[B. Implementation 20](#_heading=h.1hmsyys)

[1. So sánh số lượng hóa đơn bán hàng giữa loại khách hàng ‘Normal’ và ‘Member’ theo từng chi nhánh 20](#_heading=h.ysv5pfox0x7n)

[2. Thống kê tổng doanh thu bán hàng của các chi nhánh 21](#_heading=h.fe4nb907m08)

[3. Tìm hóa đơn của khách hàng có tổng tiền thanh toán ( Tính cả thuế) cao nhất 22](#_heading=h.ydmz95gsbhne)

[4. Tính đánh giá trung bình cho toàn bộ đơn hàng của chi nhánh B 23](#_heading=h.17r6izl6m0jj)

[5. Sản phẩm nào bán chạy nhất ? 23](#_heading=h.jh27gxhpxc59)

[6. So sánh phương thức thanh toán nào được sử dụng nhiều hơn giữa Ewallet và Credit card ? 23](#_heading=h.s23q7o66vhhr)

[7. Có sự chênh lệch nào đáng kể giữa mua sắm của nam và nữ không? 24](#_heading=h.9gmvanmw7h98)

[8. So sánh doanh thu giữa 3 tháng đầu năm 2019 và 3 tháng đầu năm 2020? 25](#_heading=h.3chynoaftych)

[V. Reference 25](#_heading=h.41mghml)

# I.Thông tin của nhóm và tiến độ đồ án

| Bảng phân công | | |
| --- | --- | --- |
| **Công việc thực hiện** | **Người thực hiện** | **Tiến độ hoàn thành** |
| Data Abstraction, TaskAbtraction, 2 IDIOM, 2 CHARTS (D3.JS) | 21127705 – Từ Phước Toàn | 100% |
| Data Abstraction, TaskAbtraction, 2 IDIOM, 2 CHARTS (D3.JS) | 21127620 – Trần Hoàng Khải | 100% |
| Data Abstraction, TaskAbtraction, 2 IDIOM, 2 CHARTS (D3.JS) | 20127361 – Nguyễn Ngọc Diệu Trang | 100% |
| Data Abstraction, TaskAbtraction, 2 IDIOM, 2 CHARTS (D3.JS) | 20127632 – Trần Đình Thiện | 100% |

# II. Giai đoạn 1 Data abstraction

Dữ liệu được lấy từ File: [supermarket\_sales.xls](https://courses.ctda.hcmus.edu.vn/mod/resource/view.php?id=75144)

*Conceptual Model  
 Khái quát về bộ dữ liệu:*Bộ dữ liệu ghi nhận lại lịch sử giao dịch khách hàng của ba chi nhánh A, B, C

## Data forms

1. Dataset Type & Data types:

* Dataset Type: Table.
* Loại Dataset: Static. Vì có thể truy xuất một lần  
  được hết tất cả các dữ liệu.
* Data types:
* Mỗi dòng là 1 hóa đơn mua sản phẩm của một khách hàng .
* Mỗi cột là một attribute bao gồm:  
  { Invoice ID, Branch, Customer type, Gender, ProductID, Quantity, Tax 5%, Total, Date, Time, Payment, cogs, gross margin percentage, gross income, Rating}.
* Tổng số Thuộc tính: 15.
* Tổng số items: 1016.

2. Attributes

* Branch: Chi nhánh của siêu thị (có 3 chi nhánh được xác định là A, B và C).
* cogs: giá vốn hàng bán
* Customer type: Loại khách hàng, ghi nhận bởi Members cho khách hàng sử dụng thẻ thành viên và Normal cho những người không có thẻ thành viên.
* Date: Ngày mua hàng (Bản ghi có sẵn từ tháng 1 năm 2019 đến tháng 3 năm 2019)
* Gender: Giới tính của khách hàng
* gross income: lợi nhuận sau khi trừ đi giá vốn hàng bán từ doanh thu
* gross margin percentage: Tỷ lệ lợi nhuận gộp
* Invoice ID: Số nhận dạng hóa đơn hóa đơn bán hàng được tạo bởi máy tính
* Payment: Phương thức thanh toán được sử dụng bởi khách hàng cho việc mua sắm (có 3 phương thức có sẵn - Tiền mặt, Thẻ tín dụng và Ví điện tử)
* ProductID: Mã sản phẩm
* Quantity: Số lượng sản phẩm được mua bởi khách hàng
* Rating: Đánh giá phân loại của khách hàng về trải nghiệm mua sắm tổng thể của họ (trên một thang điểm từ 1 đến 10)
* Tax 5%: Phí thuế 5% đối với khách hàng mua sắm
* Time: Thời gian mua hàng (từ 10 giờ sáng đến 9 giờ tối)
* Total: Tổng giá bao gồm cả thuế

3. Attributes Type

| | Tên thuộc tính | Attribute types | Ordered attribute type | | Hierarchical | Direction | | | Attribute characteristics | | Bin | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Interval | Raito | Sequential | Diverging | Cyclic | continuous | Discrete | |  | | Invoice ID | c |  |  |  |  |  |  |  | x | 1016 |  | | Branch | c |  |  |  |  |  |  |  | x | 3 |  | | Customer type | c |  |  |  |  |  |  |  | x | 2 |  | | Gender | c |  |  |  |  |  |  |  | x | 2 |  | | ProductID | c |  |  | x |  |  |  | x |  | 943 |  | | Quantity | q | x |  |  | x |  |  | x |  | 10 |  | | Tax 5% | q | x |  |  | x |  |  | x |  | 990 |  | | Total | q |  | x |  | x |  |  |  | x | 990 |  | | Date | q | x |  | x |  | x |  |  | x | 103 |  | | Time | o | x |  | x |  |  | x |  | x | 506 |  | | Payment | c |  |  |  |  |  |  |  | x | 3 |  | | cogs | q | x |  |  | x |  |  | x |  | 990 |  | | gross margin percentage | q | x |  |  | x |  |  | x |  | 1 |  | | gross income | q | x |  |  | x |  |  | x |  | 990 |  | | Rating | o |  | x |  |  | x |  |  | x | 61 |  | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |
| * Cột dữ liệu tên Date không thống nhất về định dạng có 2 kiểu là dd-mm-yy và dd-mmm-yy Khắc phục chỉnh toàn bộ về dd-mm-yy * Cột dữ liệu tên gender không thông nhất về định dang, có 4 kiểu lựa chọn giới tính(Female, Male, F, M). Khắc Phục F về Female, M về Male |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

## B.Data Profiling

| Field Name | NULL | Missing | Actual | Completenes | Cardinality | Uniqueness | Distinctness |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Invoice ID | 0 | 0 | 1016 | 100% | 1016 | 100.00% | 100.00% |
| Branch | 0 | 0 | 1016 | 100% | 3 | 0.29% | 0.29% |
| Customer type | 0 | 0 | 1016 | 100% | 2 | 0.19% | 0.19% |
| Gender | 0 | 0 | 1016 | 100% | 2 | 0.20% | 0.20% |
| ProductID | 0 | 0 | 1016 | 100% | 943 | 92.80% | 92.80% |
| Quantity | 0 | 0 | 1016 | 100% | 10 | 1% | 1% |
| Tax 5% | 0 | 0 | 1016 | 100% | 990 | 97.44% | 97.44% |
| Total | 0 | 0 | 1016 | 100% | 990 | 97.44% | 97.44% |
| Date | 0 | 0 | 1016 | 100% | 103 | 10.10% | 10.10% |
| Time | 0 | 0 | 1016 | 100% | 506 | 49.80% | 49.80% |
| Payment | 0 | 0 | 1016 | 100% | 3 | 0.30% | 0.30% |
| cogs | 0 | 0 | 1016 | 100% | 990 | 97.44% | 97.44% |
| gross margin percentage | 0 | 0 | 1016 | 100% | 1 | 100% | 100% |
| gross income | 0 | 0 | 1016 | 100% | 990 | 97.44% | 97.44% |
| Rating | 0 | 0 | 1016 | 100% | 61 | 6% | 6% |

### Invoice ID

| **Input Metadata** | |  | **Invoice ID (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Invoice ID |  | varchar | 1016 | 100% |
| Field Data Type | varchar |  | **Thứ Hạng (Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 11 |  | 750-67-8428 | 1 | 0.09% |
| **Data Profilling Summary Statistics** | |  | 123-19-1176 | 1 | 0.09% |
| NULL | 0 |  | 355-53-5943 | 1 | 0.09% |
| Missing | 0 |  | 665-32-9167 | 1 | 0.09% |
| Actual | 1016 |  | 656-95-9349 | 1 | 0.09% |
| Completeness | 100% |  | **Thứ Hạng (Field Formats)** | **Count** | **Percentage** |
| Cardinality | 1016 |  | nnn-nn-nnnn | 1016 | 100.00% |
| **Data Profilling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 11 |  |  |  |  |
| Field Length (MAX) | 11 |  |  |  |  |
| Field Value (MIN) |  |  |  |  |  |
| Field Value (MAX) |  |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 2. Branch:

| **Input Metadata** | |  | **Branch (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Branch |  | char | 1016 | 100% |
| Field Data Type | char |  | **Tên đăng nhập (Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 1 |  | A | 349 | 34.30% |
| **Data Profilling Summary Statistics** | |  | B | 338 | 33.26% |
| NULL | 0 |  | C | 320 | 31.49% |
| Completeness | 100% |  | **Tên đăng nhập (Field Formats)** | **Count** | **Percentage** |
| Cardinality | 3 |  | n | 1016 | 100.00% |
| Uniqueness | 0,29% |  |  |  |  |
| Distinctness | 0.29% |  |  |  |  |
| **Data Profilling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 1 |  |  |  |  |
| Field Length (MAX) | 1 |  |  |  |  |
| Field Value (MIN) |  |  |  |  |  |
| Field Value (MAX) |  |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 3. Customer type:

| **Input Metadata** | |  | **Quality (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Customer type |  | varchar | 1016 | 100% |
| Field Data Type | varchar |  | **Quality (Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 25 |  | Normal | 509 | 50.10% |
| **Data Profilling Summary Statistics** | |  | Member | 507 | 49.90% |
| NULL | 0 |  | **Quality (Field Formats)** | **Count** | **Percentage** |
| Missing | 0 |  | n | 1016 | 100.00% |
| Actual | 1016 |  |  |  |  |
| Completeness | 100% |  |  |  |  |
| Cardinality | 2 |  |  |  |  |
| Uniqueness | 0.19% |  |  |  |  |
| Distinctness | 0.19% |  |  |  |  |
| **Data Profilling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 6 |  |  |  |  |
| Field Length (MAX) | 6 |  |  |  |  |
| Field Value (MIN) |  |  |  |  |  |
| Field Value (MAX) |  |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 4. Gender:

| **Input Metadata** | |  | **Quality (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Quality |  | varchar | 1016 | 100% |
| Field Data Type | varchar |  | **Quality (Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 6 |  | female | 507 | 49.90% |
| **Data Profilling Summary Statistics** | |  | male | 509 | 50.10% |
| NULL | 0 |  | **Quality (Field Formats)** | **Count** | **Percentage** |
| Missing | 0 |  | nnnn | 509 | 50.10% |
| Actual | 1016 |  | nnnnnn | 507 | 49.90% |
| Completeness | 100% |  |  |  |  |
| Cardinality | 2 |  |  |  |  |
| Uniqueness | 0.20% |  |  |  |  |
| Distinctness | 0.20% |  |  |  |  |
| **Data Profilling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 4 |  |  |  |  |
| Field Length (MAX) | 6 |  |  |  |  |
| Field Value (MIN) | Male |  |  |  |  |
| Field Value (MAX) | Female |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 5. ProductID:

| **Input Metadata** | |  | **Quality (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Quality |  | varchar | 1016 | 100% |
| Field Data Type | varchar |  | **Quality (Top 2 Values)** | **Count** | **Percentage** |
| Field Length | 6 |  | pid943 | 3 | 0.30% |
| **Data Profilling Summary Statistics** | |  | pid897 | 3 | 0.30% |
| NULL | 0 |  | **Quality (Field Formats)** | **Count** | **Percentage** |
| Missing | 0 |  | nnnnnn | 917 | 90,26% |
| Actual | 1016 |  | nnnnn | 99 | 9,74% |
| Completeness | 100% |  |  |  |  |
| Cardinality | 943 |  |  |  |  |
| Uniqueness | 92.80% |  |  |  |  |
| Distinctness | 92.80% |  |  |  |  |
| **Data Profilling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 5 |  |  |  |  |
| Field Length (MAX) | 6 |  |  |  |  |
| Field Value (MIN) | PID01 |  |  |  |  |
| Field Value (MAX) | PID934 |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 6. Quantity:

| **Input Metadata** | |  | **Quality (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Quality |  | interger | 1016 | 100% |
| Field Data Type | interger |  | **Quality (Top 2 Values)** | **Count** | **Percentage** |
| Field Length | 1 |  | 10 | 121 | 11.90% |
| **Data Profilling Summary Statistics** | |  | 1 | 113 | 11.10% |
| NULL | 0 |  | **Quality (Field Formats)** | **Count** | **Percentage** |
| Missing | 0 |  | n | 1016 | 100% |
| Actual | 1016 |  |  |  |  |
| Completeness | 100% |  |  |  |  |
| Cardinality | 10 |  |  |  |  |
| Uniqueness | 1.00% |  |  |  |  |
| Distinctness | 1.00% |  |  |  |  |
| **Data Profilling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 1 |  |  |  |  |
| Field Length (MAX) | 1 |  |  |  |  |
| Field Value (MIN) | 1 |  |  |  |  |
| Field Value (MAX) | 10 |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 7. Tax 5%:

| **Input Metadata** | |  | **Quality (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Quality |  | float | 1016 | 100% |
| Field Data Type | float |  | **Quality (Top 2 Values)** | **Count** | **Percentage** |
| Field Length | 1 |  | 28.116 | 2 | 0.20% |
| **Data Profilling Summary Statistics** | |  | 10.3635 | 2 | 0.20% |
| NULL | 0 |  | **Quality (Field Formats)** | **Count** | **Percentage** |
| Missing | 0 |  | nn.nnnn | 130 | 13% |
| Actual | 1016 |  | nn.nnn | 335 | 33% |
| Completeness | 100% |  | nn.nn | 99 | 10% |
| Cardinality | 990 |  | nn.n | 11 | 1% |
| Uniqueness | 97.40% |  | nn | 1 | 0% |
| Distinctness | 97.40% |  | n.nnnn | 108 | 11% |
| **Data Profilling Additional Statistics** | |  | n.nnn | 276 | 27% |
| Field Data Types | 1 |  | n.nn | 50 | 5% |
| Field Length (MIN) | 3 |  | n.n | 6 | 1% |
| Field Length (MAX) | 7 |  |  |  |  |
| Field Value (MIN) | 0.5085 |  |  |  |  |
| Field Value (MAX) | 49.65 |  |  |  |  |
| Field Formats | 9 |  |  |  |  |

### 8. Total:

| **Input Metadata** | |  | **Total(Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Total |  | REAL | 1016 | 100% |
| Field Data Type | REAL |  | **Total(Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 8 |  | 590.436 | 2 | 0.20% |
| **Data Profiling Summary Statistics** | |  | 217.6335 | 2 | 0.20% |
| NULL | 0 |  | 87.234 | 2 | 0.20% |
| Missing | 0 |  | 69.72 | 2 | 0.20% |
| Actual | 1016 |  | 277.137 | 2 | 0.20% |
| Completeness | 100% |  | **Total (Field formats)** | **Count** | **Percentage** |
| Cardinarity | 990 |  | nnn | 1 | 0.098425 |
| Uniqueness | 97.44% |  | nnnn | 1 | 0.098425 |
| Distinctness | 97.44% |  | nnnnn | 35 | 3.444882 |
| **Data Profiling Additional Statistics** | |  | nnnnnn | 246 | 24.2126 |
| Field Data Types | 1 |  | nnnnnnn | 561 | 55.21654 |
| Field Length(Min) | 3 |  | nnnnnnnn | 172 | 16.92913 |
| Field Length(Max) | 8 |  |  |  |  |
| Field Value(MIN) | 10.6785 |  |  |  |  |
| Field Value(MAX) | 1042.65 |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 9. Date:

| **Input Metadata** | |  | **Date(Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Date |  | Datetime | 1016 | 100% |
| Field Data Type | Datetime |  | **Date(Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 8 |  | 07-02-19 | 20 | 1.97% |
| **Data Profiling Summary Statistics** | |  | 15-02-19 | 19 | 1.87% |
| NULL | 0 |  | 02-03-19 | 18 | 1.77% |
| Missing | 0 |  | 08-01-19 | 18 | 1.77% |
| Actual | 1016 |  | 14-03-19 | 18 | 1.77% |
| Completeness | 100% |  | **Date (Field formats)** | **Count** | **Percentage** |
| Cardinarity | 103 |  | dd-mm-yy | 1016 | 100% |
| Uniqueness | 10.10% |  |  |  |  |
| Distinctness | 10.10% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length(Min) | 8 |  |  |  |  |
| Field Length(Max) | 8 |  |  |  |  |
| Field Value(MIN) | ###### |  |  |  |  |
| Field Value(MAX) | ###### |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 10. Time:

| **Input Metadata** | |  | **Time(Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Time |  | Datetime | 1016 | 100% |
| Field Data Type | Datetime |  | **Time(Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 5 |  | 20:00 - 20:59 | 114 | 11.22% |
| **Data Profiling Summary Statistics** | |  | 16:00 - 16:59 | 103 | 10.14% |
| NULL | 0 |  | 11:00 - 11:59 | 101 | 9.94% |
| Missing | 0 |  | 14:00 - 14:59 | 100 | 9.84% |
| Actual | 1016 |  | 19:00 - 19:59 | 92 | 9.06% |
| Completeness | 100% |  | **Time (Field formats)** | **Count** | **Percentage** |
| Cardinarity | 506 |  | hh:mm | 1016 | 100% |
| 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) | 10:00 |  |  |  |  |
| Field Value(MAX) | 20:59 |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 11. Payment:

| **Input Metadata** | |  | **Payment(Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Time |  | String | 1016 | 100% |
| Field Data Type | String |  | **Payment(Top 3 Values)** | **Count** | **Percentage** |
| Field Length | 5 |  | Ewallet | 352 | 34.60% |
| **Data Profiling Summary Statistics** | |  | Cash | 348 | 34.30% |
| NULL | 0 |  | Credit card | 316 | 31.10% |
| Missing | 0 |  |  |  |  |
| Actual | 1016 |  |  |  |  |
| Completeness | 100% |  | **Payment (Field formats)** | **Count** | **Percentage** |
| Cardinarity | 3 |  | nnnn | 352 | 34.60% |
| Uniqueness | 0.30% |  | nnnnnnn | 348 | 34.30% |
| Distinctness | 0.30% |  | nnnnnnnnnnn | 316 | 31.10% |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length(Min) | 4 |  |  |  |  |
| Field Length(Max) | 11 |  |  |  |  |
| Field Value(MIN) | Cash |  |  |  |  |
| Field Value(MAX) | Credit card |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 12. cogs:

| **Input Metadata** | |  | **cogs (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Cogs |  | Float | 1016 | 100% |
| Field Data Type | Float |  | **cogs (Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 6 |  | 993 | 1 | 0.01% |
| **Data Profiling Summary Statistics** | |  | 989.8 | 1 | 0.01% |
| NULL | 0 |  | 985.2 | 1 | 0.01% |
| Missing | 0 |  | 975 | 1 | 0.01% |
| Actual | 1016 |  | 973.8 | 1 | 0.01% |
| Completeness | 100% |  | **Thứ Hạng (Field Formats)** | **Count** | **Percentage** |
| Cardinality | 990 |  | nnn.nn | 543 |  |
| Uniqueness | 97.44% |  | nnn.n | 216 |  |
| Distinctness | 97.44% |  | nnn | 30 |  |
| **Data Profiling Additional Statistics** | |  | nn.nn | 186 |  |
| Field Data Types | 1 |  | nn.n | 37 |  |
| Field Length (MIN) | 4 |  | nn | 4 |  |
| Field Length (MAX) | 6 |  |  |  |  |
| Field Value (MIN) | 10.17 |  |  |  |  |
| Field Value (MAX) | 993 |  |  |  |  |
| Field Formats | 6 |  |  |  |  |

### 13. gross margin percentage

| **Input Metadata** | |  | **Gross margin percentage (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Gross margin percentage |  | Float | 1016 | 100% |
| Field Data Type | Float |  | **Gross margin percentage (Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 10 |  | 4.761904762 | 1016 | 100% |
| **Data Profiling Summary Statistics** | |  |  |  |  |
| NULL | 0 |  |  |  |  |
| Missing | 0 |  |  |  |  |
| Actual | 1016 |  |  |  |  |
| Completeness | 100% |  | **Gross margin percentage (Field Formats)** | **Count** | **Percentage** |
| Cardinality | 1 |  | n.nnnnnnnn | 1016 | 100% |
| Uniqueness | 100% |  |  |  |  |
| Distinctness | 100% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 10 |  |  |  |  |
| Field Length (MAX) | 10 |  |  |  |  |
| Field Value (MIN) | 4.761904762 |  |  |  |  |
| Field Value (MAX) | 4.761904762 |  |  |  |  |
| Field Formats | 1 |  |  |  |  |

### 14. gross income

| **Input Metadata** | |  | **Gross income (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Gross income |  | Float | 1016 | 100% |
| Field Data Type | Float |  | **Gross income (Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 7 |  | 49.65 | 1 | 0.01% |
| **Data Profiling Summary Statistics** | |  | 49.49 | 1 | 0.01% |
| NULL | 0 |  | 49.26 | 1 | 0.01% |
| Missing | 0 |  | 48.75 | 1 | 0.01% |
| Actual | 1016 |  | 48.69 | 1 | 0.01% |
| Completeness | 100% |  | **Gross income (Field Formats)** | **Count** | **Percentage** |
| Cardinality | 990 |  | nn.nnnn | 130 | 13% |
| Uniqueness | 97.44% |  | nn.nnn | 335 | 33% |
| Distinctness | 97.44% |  | nn.nn | 99 | 10% |
| **Data Profiling Additional Statistics** | |  | nn.n | 11 | 1% |
| Field Data Types | 1 |  | nn | 1 | 0% |
| Field Length (MIN) | 3 |  | n.nnnn | 108 | 11% |
| Field Length (MAX) | 7 |  | n.nnn | 276 | 27% |
| Field Value (MIN) | 0.5085 |  | n.nn | 50 | 5% |
| Field Value (MAX) | 49.65 |  | n.n | 6 | 1% |
| Field Formats | 9 |  |  |  |  |

### 15. Rating

| **Input Metadata** | |  | **Rating (Field Data Types)** | **Count** | **Percentage** |
| --- | --- | --- | --- | --- | --- |
| Field Name | Rating |  | Float | 1016 | 100% |
| Field Data Type | Float |  | **Rating (Top 5 Values)** | **Count** | **Percentage** |
| Field Length | 3 |  | 10 | 5 | 0.49% |
| **Data Profiling Summary Statistics** | |  | 9.9 | 17 | 1.67% |
| NULL | 0 |  | 9.8 | 19 | 1.87% |
| Missing | 0 |  | 9.7 | 14 | 1.38% |
| Actual | 1016 |  | 9.6 | 17 | 1.67% |
| Completeness | 100% |  | **Rating (Field Formats)** | **Count** | **Percentage** |
| Cardinality | 61 |  | n | 120 | 11.81% |
| Uniqueness | 6% |  | n.n | 896 | 88.19% |
| Distinctness | 6% |  |  |  |  |
| **Data Profiling Additional Statistics** | |  |  |  |  |
| Field Data Types | 1 |  |  |  |  |
| Field Length (MIN) | 1 |  |  |  |  |
| Field Length (MAX) | 3 |  |  |  |  |
| Field Value (MIN) | 4 |  |  |  |  |
| Field Value (MAX) | 10 |  |  |  |  |
| Field Formats | 2 |  |  |  |  |

### 16.Đánh giá

- Invoice ID có completeness, Uniqueness, Distinctness là 100% cho nên Invoice ID là thuộc tính làm khóa chính

- Số lượng khách hàng mua ở chi nhánh A là nhiều nhất và ở chi nhánh C là thấp nhất

- Không có sự chênh lệch nhiều về số lượng loại khách hàng mua.

- Khung giờ cao điểm với nhiều khách hàng mua nhất là từ 20:00 đến 20:59

- Chỉ có 3 phương thức thanh toán và phương thích thực hiện nhiều nhất là ví điện tử với 352 lần chiếm 34.6% và ít nhất là thẻ tín dụng với 316 lần chiếm 31.1%

- Thời gian được bán được nhiều sản phẩm nhất là 7/2/2019 với 20 mới lượt mua sản phẩm

- Giá cao nhất là 1042.65, Giá thấp nhát là 10.6785, Giá trung bình là 322.90439, Tổng tiền toàn bộ khách hàng là 328070.86

- Đánh giá thấp nhất của khách hàng là 4 và cao nhất là 10.

- Tất cả các giá trị của cột “gross margin percentage” đều giống nhau, điều này có thể cho thấy rằng tỷ lệ lợi nhuận gộp của công ty đối với tất cả các sản phẩm hoặc dịch vụ là nhất quán.

# III. Giai đoạn 2 Task abstraction

## 1.Khái niệm:

Phân tích thành dạng trừu tượng của một hoặc nhiều công việc để tìm ra những điểm giống nhau hoặc khác nhau của những công việc đó. Nếu chỉ dựa vào ngôn ngữ bình thường, mặc dù các công việc đó có thể giống nhau nhưng lại được thể hiện bởi những thuật ngữ khác nhau thì chúng ta sẽ khó tìm ra được điểm giống và khác nhau.

→ Task abstraction là sử dụng một ngôn ngữ chung để thể hiện, từ đó dễ dàng tìm ra những điểm giống nhau và khác nhau của nhiều công việc với nhau.

A screenshot of a computer

Description automatically generated

*Hình 2.1 Hình ảnh về Task Abstraction*

## Các Task Abstraction:

### So sánh số lượng hóa đơn bán hàng giữa loại khách hàng ‘Normal’ và ‘Member’ theo từng chi nhánh

Produce -> browse -> compare

### Thống kê tổng doanh thu bán hàng của các chi nhánh

Produce -> explore -> Sum

### Tìm hóa đơn của khách hàng có tổng tiền thanh toán ( Tính cả thuế) cao nhất

Product -> locate -> identify

### Tính đánh giá trung bình cho toàn bộ đơn hàng của chi nhánh B

Produce -> browse -> Summarize

### Sản phẩm nào bán chạy nhất ?

produce -> explore -> identify

### So sánh phương thức thanh toán nào được sử dụng nhiều hơn giữa Ewallet và Credit card ?

produce -> explore -> compare

### Có sự chênh lệch nào đáng kể giữa mua sắm của nam và nữ không?

Produce -> locate -> compare

### So sánh doanh thu giữa 3 tháng đầu năm 2019 và 3 tháng đầu năm 2020?

Produce -> lookup -> Sum

# IV. Giai đoạn 3: Idiom and Implementation

## Idiom

## So sánh số lượng hóa đơn bán hàng giữa loại khách hàng ‘Normal’ và ‘Member’ theo từng chi nhánh

Derived attribute: totalInvoice

| Idiom | Stacked Bar Charts |
| --- | --- |
| Data Type | 2 categorical: Customer type, branch  1 quantitative: totalInvoice |
| Encoding | Mark: bar mark (vertical)   * Glyph: multiple sub-bars are stacked vertically   Channel:   * Position: vertical and horizontal * Length, color HUE * Spatial region: glyph * Align: full glyph, lowest bar component * Separate: horizotal position |
| Task | Part-to-whole relationship |
| Scale | Stacked key attribute: 2 levels  Main key attribute: 3 levels |

## Thống kê tổng doanh thu bán hàng của các chi nhánh

Derived column: totalSales

| Idiom | Bar Charts |
| --- | --- |
| Data Type | 1 categorical: Branch  1 quantitative: totalSales |
| Encoding | Mark: line  Channel:   * Position: vertical and horizontal * Size: length, area * Express: quantitative (length) * Spatial region: categorical * Separate: horizontal position * Align: vertical position * Order: label (alphabeltic) * Color: HUE |
| Task | Compare, lookup |
| Scale | Key attribute: 3 levels |

## Tìm hóa đơn của khách hàng có tổng tiền thanh toán ( Tính cả thuế) cao nhất

| idioms | Bar chart | |
| --- | --- | --- |
| Data Type | 1 categorical: invoice ID  1quantitative: Total | |
| Encoding | Mark: line  Channel:   * Position:Both * Length, color HUE * Size: length, Area * Spatial region: Categorical * Separate: horizotal position | |
| Task | Find the highest bill | |
| scale | top 10 items | |

## Tính đánh giá trung bình cho toàn bộ đơn hàng của chi nhánh B

| idioms | Bar chart | |
| --- | --- | --- |
| Data Type | 1 categorical: Branch  1quantitative: Rating | |
| Encoding | Mark:line (vertical)  Channel:   * Position: Both * Length, color HUE * Spatial region: Categorical | |
| Task | Compare, lookup value | |
| scale | All items | |

## Sản phẩm nào bán chạy nhất ?

| Idiom | Bar Charts |
| --- | --- |
| Data Type | 1 categorical: ProductID  1 quantitative: Quantity |
| Encoding | Express: quantitative  Spatial region: categorical  Separate: vertical position  Align: horizontal position  Order:length |
| Task | Compare, lookup value |

## So sánh phương thức thanh toán nào được sử dụng nhiều hơn giữa Ewallet và Credit card ?

| Idiom | Bar Charts |
| --- | --- |
| Data Type | 1 categorical: YearWorked  1 quantitative: Payment |
| Encoding | Mark: line  Channel:  · Position: horizontal and vertical  · Align: lengths ( quantitative)   * Separate, order: horizontal region (ordered) |
| Task | Compare |
| Scale | 2 items |

## Có sự chênh lệch nào đáng kể giữa mua sắm của nam và nữ không?

| idioms | Bar chart | |
| --- | --- | --- |
| Data Type | 1 categorical: Gender  1quantitative: Total | |
| Encoding | Mark: line  Channel:   * Position:Both * Length, color HUE * Size: length, Area * Spatial region: Categorical * Separate: horizotal position | |
| Task | compare | |
| scale | All items | |

## So sánh doanh thu giữa 3 tháng đầu năm 2019 và 3 tháng đầu năm 2020?

| idioms | Bar chart | |
| --- | --- | --- |
| Data Type | 1 categorical: Year  1quantitative: Gross income | |
| Encoding | Mark: line  Channel:   * Position:Both * Length, color HUE * Size: length, Area * Spatial region: Categorical * Separate: horizotal position | |
| Task | compare | |
| scale | All items | |

## Implementation

## So sánh số lượng hóa đơn bán hàng giữa loại khách hàng ‘Normal’ và ‘Member’ theo từng chi nhánh

A graph of blue and red squares

Description automatically generated

| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout |  | x |
| Perceptual grouping | x |  |

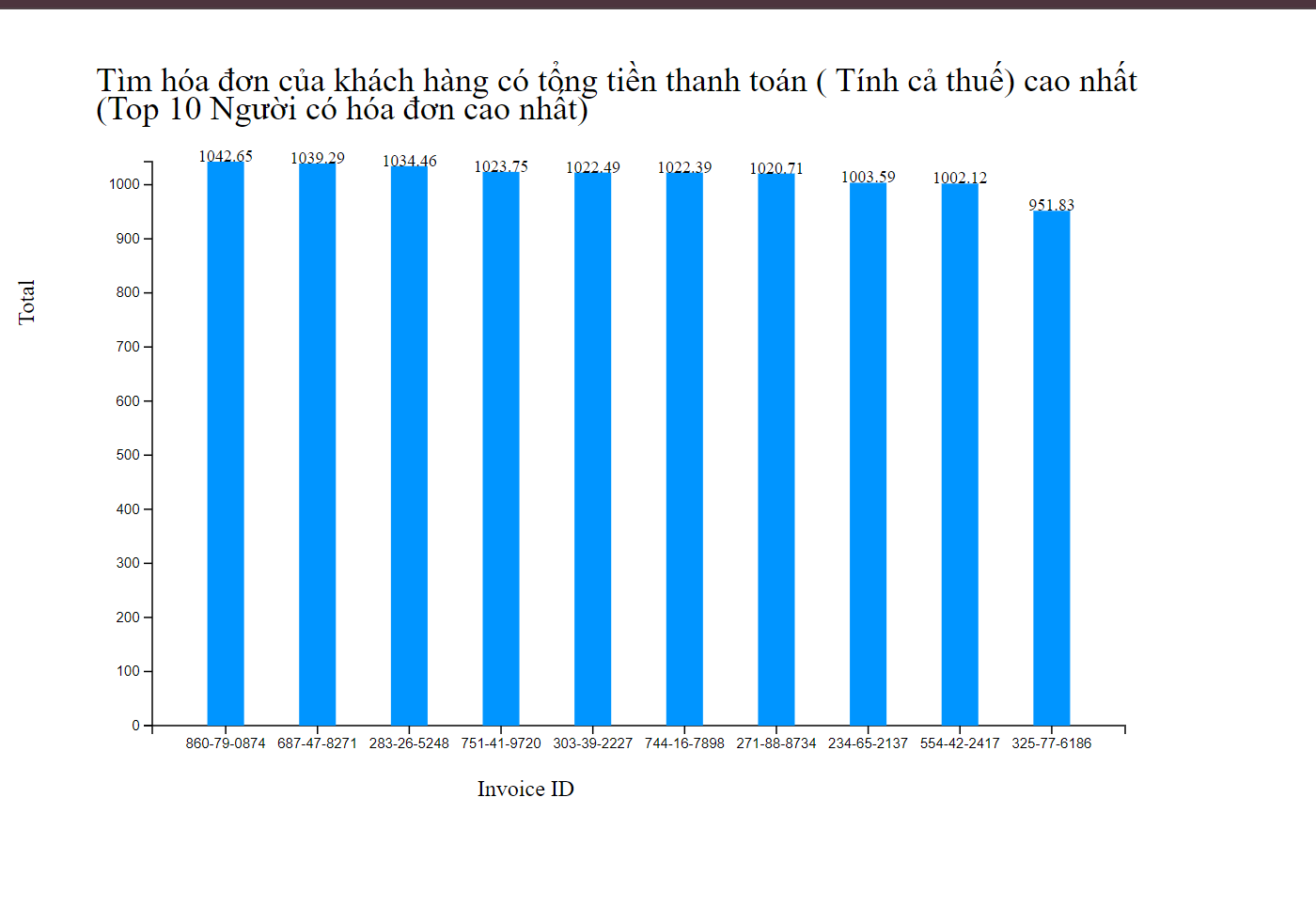
## Thống kê tổng doanh thu bán hàng của các chi nhánh

A graph with different colored bars

Description automatically generated

| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout |  | x |
| Perceptual grouping | x |  |

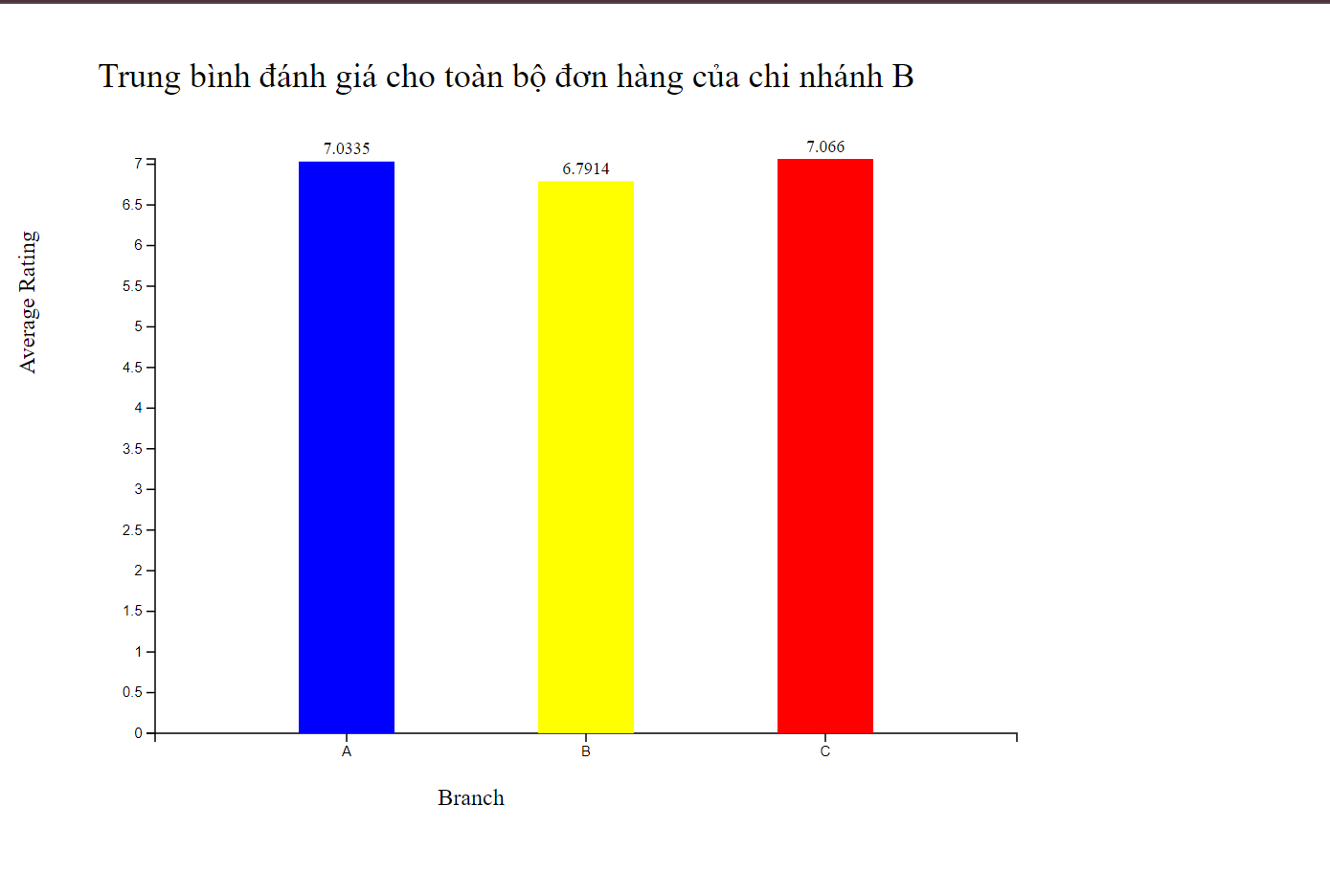
## Tìm hóa đơn của khách hàng có tổng tiền thanh toán ( Tính cả thuế) cao nhất



| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout |  | x |
| Perceptual grouping | x |  |

Đánh Giá:

## Tính đánh giá trung bình cho toàn bộ đơn hàng của chi nhánh B



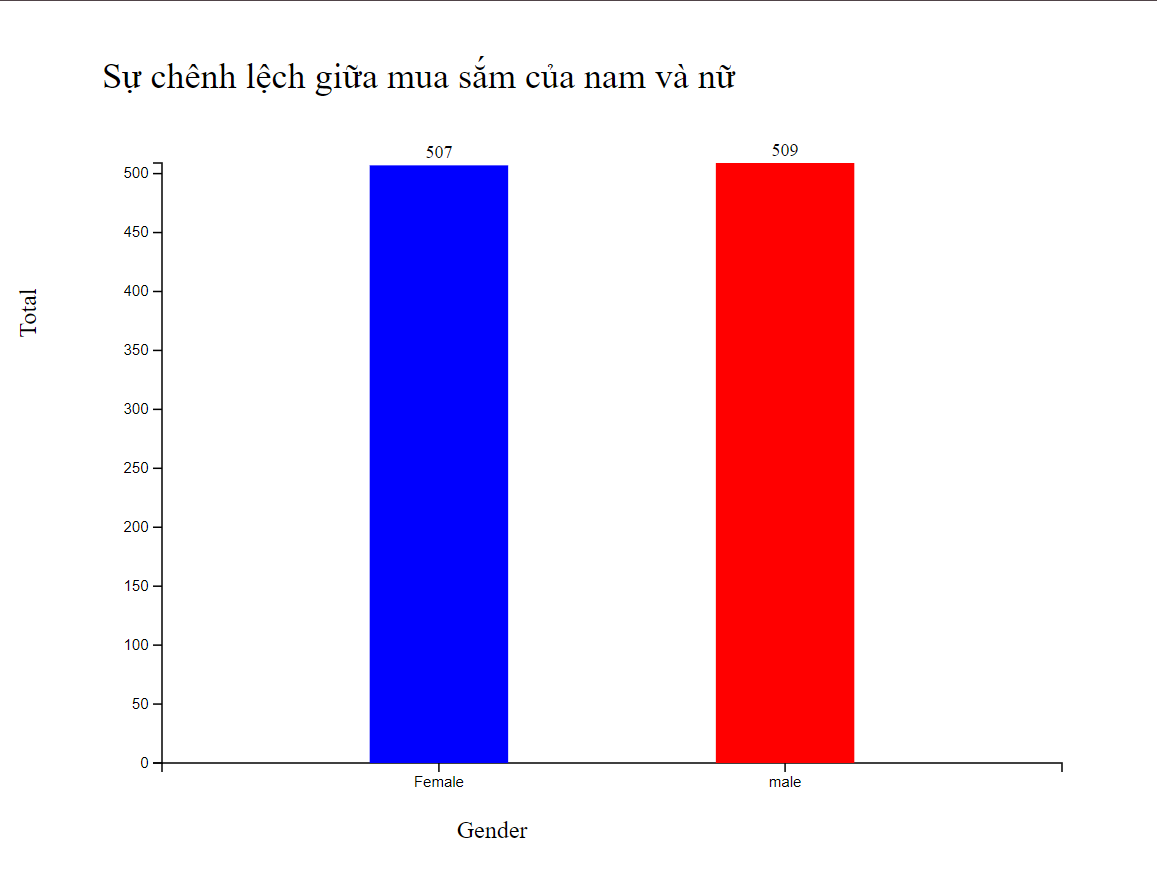
| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout |  | x |
| Perceptual grouping | x |  |

Đánh Giá:

## Sản phẩm nào bán chạy nhất ?

## So sánh phương thức thanh toán nào được sử dụng nhiều hơn giữa Ewallet và Credit card ?

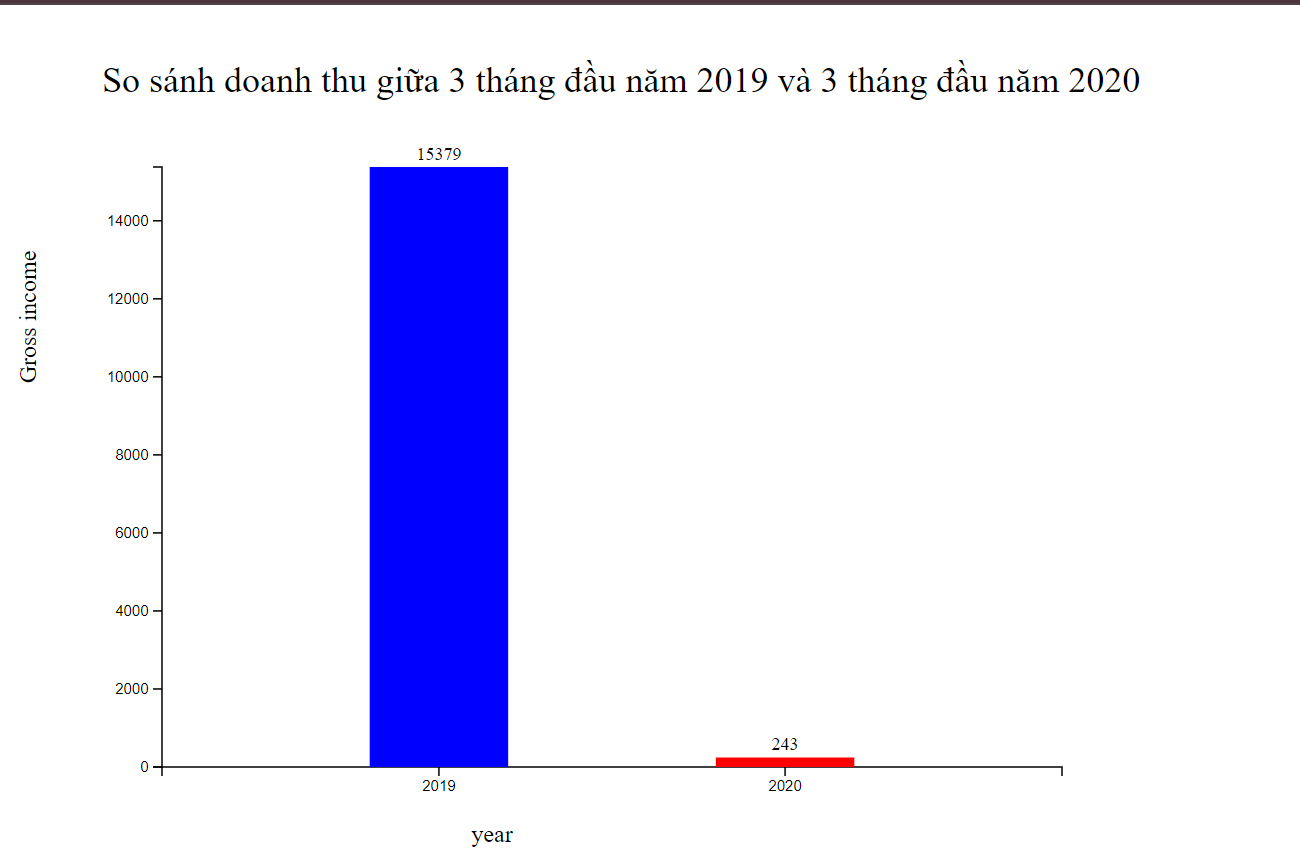
## Có sự chênh lệch nào đáng kể giữa mua sắm của nam và nữ không?



| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout |  | x |
| Perceptual grouping | x |  |

Đánh Giá: Nhìn Biểu đồ ta thấy sự chênh lệch mua saswsm giữa nam và nữ không cs sự chênh lệnh cao, gần như là bằng nhau

## So sánh doanh thu giữa 3 tháng đầu năm 2019 và 3 tháng đầu năm 2020?



| Effectiveness | Đạt | Không đạt |
| --- | --- | --- |
| Accuracy | x |  |
| Discriminability | x |  |
| Separability | x |  |
| Visual popout |  | x |
| Perceptual grouping | x |  |

# 

Đánh Giá: Nhìn biều đồ ta thấy sự chênh lệch doanh thu giữa 3 tháng đầu năm 2019 và 2020 là rất lớn và theo số liệu, lượng mua hàng 3 tháng đầu năm 2020 rất thấp(14 khách hàng trên tổng 1017 khách hàng) nên doanh thu chênh lệch rất nhiều

# V. Reference

<https://drive.google.com/drive/folders/14t8jrlHysQawx5R5opvRfMtF9FRPRXoH>

<https://drive.google.com/file/d/1TBRaHCOakafAyb9We9NvbkT-Dg-MPC9h/view>

<https://courses.ctda.hcmus.edu.vn/mod/folder/view.php?id=75246>