****

**BÁO CÁO DỰ ÁN**

* **Chuyên ngành Xử Lý Dữ Liệu -**

**Customer Personality Analysis**

**Nhóm: Sắc màu**

Các thành viên:

Trần Thị Hưởng Thọ - PS41852 (Leader)

Trần Duy Trường - PS42545

Trịnh Thị Mỹ Huệ - PS42117

TP.HCM 11 - 2024

Contents

[1 Giới thiệu dự án 4](#_Toc183436989)

[1.1 Giới thiệu 4](#_Toc183436990)

[ **Tiềm năng của bộ dữ liệu:** 5](#_Toc183436991)

[ Thách thức: 6](#_Toc183436992)

[1.2 Yêu cầu của công ty 7](#_Toc183436993)

[1.3 Lập kế hoạch dự án 7](#_Toc183436994)

[2 Phân tích yêu cầu khách hàng 7](#_Toc183436995)

[2.1 Phân tích yêu cầu 7](#_Toc183436996)

[2.2 Câu chuyện dữ liệu 9](#_Toc183436997)

[2.2.1 Đặt vấn đề 9](#_Toc183436998)

[2.2.2 Xác định câu chuyện 11](#_Toc183436999)

[2.2.3 Xác định rõ đối tượng 13](#_Toc183437000)

[2.2.4 Xác định câu chuyện chi tiết 14](#_Toc183437001)

[2.2.5 Trình bày dữ liệu 16](#_Toc183437002)

[2.2.6 Những điều cần lưu ý 16](#_Toc183437003)

[2.3 Kiến trúc hệ thống 17](#_Toc183437004)

[2.3.1 Kiến trúc 17](#_Toc183437005)

[2.3.2 Giải thích 17](#_Toc183437006)

[2.4 Giải thích về bộ dữ liệu khách hàng 18](#_Toc183437007)

[2.4.1 Các khái niệm 18](#_Toc183437008)

[2.4.2 Các trường dữ liệu 19](#_Toc183437009)

[3 Làm sạch và chuyển đổi dữ liệu 19](#_Toc183437010)

[3.1 Chuẩn bị dữ liệu 19](#_Toc183437011)

[3.1.1 Giải pháp lưu trữ dữ liệu 20](#_Toc183437012)

[3.1.2 Giải pháp phân bố dữ liệu 21](#_Toc183437013)

[3.2 Làm sạch dữ liệu 23](#_Toc183437014)

[3.2.1 Các vấn đề ảnh hưởng tới dữ liệu 23](#_Toc183437015)

[3.2.2 Các tiêu chí đánh giá chất lượng dữ liệu 23](#_Toc183437016)

[3.2.3 Các bước làm sạch dữ liệu 24](#_Toc183437017)

[3.3 Chuyển đổi dữ liệu 27](#_Toc183437018)

[3.3.1 Các trường hợp cần chuyển đổi 27](#_Toc183437019)

[3.3.2 Các kỹ thuật chuyển đổi 27](#_Toc183437020)

[3.3.3 Trình bày các phép chuyển đổi trong dự án 27](#_Toc183437021)

[4 Xử lý dữ liệu 32](#_Toc183437022)

[4.1 Chuẩn hóa dữ liệu 32](#_Toc183437023)

[4.1.1 Trình bày các bước chuẩn hóa trong dự án 32](#_Toc183437024)

[4.2 Mô hình hóa dữ liệu 34](#_Toc183437025)

[4.2.1 Các loại mô hình hóa 34](#_Toc183437026)

[4.2.2 Các tiêu chí đánh giá mô hình dữ liệu 36](#_Toc183437027)

[4.2.3 Trình bày các bước mô hình hóa 37](#_Toc183437028)

[4.2.4 Trình bày các bước tạo bảng dữ liệu 38](#_Toc183437029)

[4.3 Xử lý dữ liệu DAX 43](#_Toc183437030)

[4.3.1 Measure 43](#_Toc183437031)

[4.3.2 Calculated column 44](#_Toc183437032)

[4.3.3 Filter 46](#_Toc183437033)

[5 Trực quan hóa dữ liệu 46](#_Toc183437034)

[5.1 Các kỹ thuật trực quan hóa 46](#_Toc183437035)

[5.2 Các nguyên tắc trực quan hóa 48](#_Toc183437036)

[5.3 Trình bày cách thêm visual mới 49](#_Toc183437037)

[5.4 Trình bày tạo các report cho dự án 52](#_Toc183437038)

[5.4.1 Tạo visual thống kê chi tiết 52](#_Toc183437039)

[5.4.2 Tạo visual thống kê tổng thể 53](#_Toc183437040)

[6 Xây dựng báo cáo 53](#_Toc183437041)

[6.1 Dashboard và report 53](#_Toc183437042)

[6.2 Xây dựng báo cáo 54](#_Toc183437043)

[6.2.1 Dashboard vs Report 54](#_Toc183437044)

[6.2.2 Dashboard 57](#_Toc183437045)

[6.2.3 Report 57](#_Toc183437046)

[6.2.4 Bookmark 57](#_Toc183437047)

[7 KẾT LUẬN 58](#_Toc183437048)

[7.1 Báo cáo 58](#_Toc183437049)

[7.1.1 Các bước viết báo cáo 58](#_Toc183437050)

[7.1.2 Tổng hợp 58](#_Toc183437051)

[7.2 Khó khăn 59](#_Toc183437052)

[7.3 Thuận lợi 59](#_Toc183437053)

[7.4 Hướng phát triển 59](#_Toc183437054)

# Giới thiệu dự án

## Giới thiệu

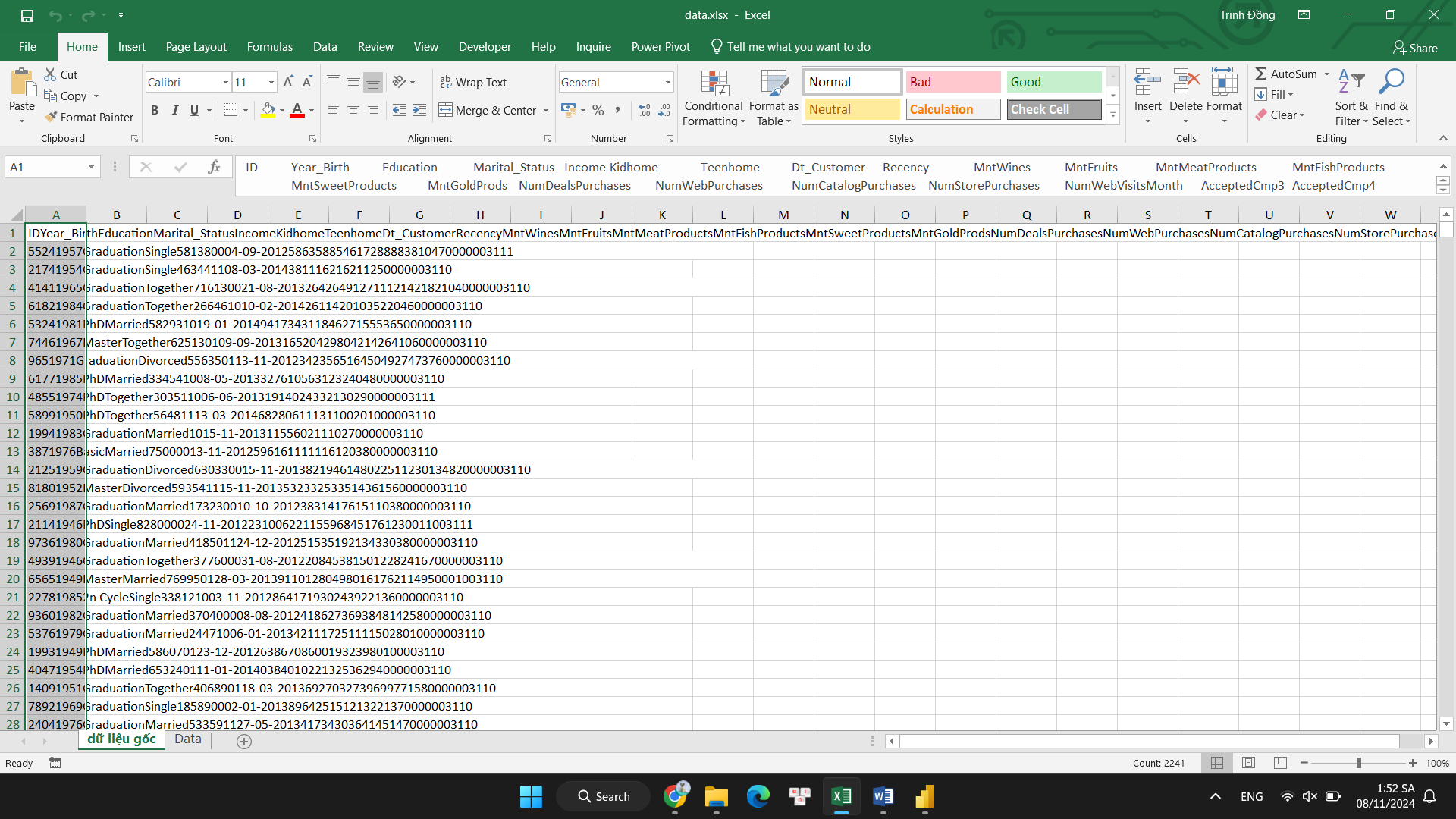
* **Customer Personality Analysis** phân tích hành vi khách hàng là phân tích chi tiết về khách hàng lý tưởng của công ty. Nó giúp doanh nghiệp hiểu rõ hơn về khách hàng của mình và giúp họ dễ dàng sửa đổi sản phẩm theo nhu cầu, hành vi và mối quan tâm cụ thể của các loại khách hàng khác nhau.
* Phân tích tính cách khách hàng giúp doanh nghiệp sửa đổi sản phẩm của mình dựa trên khách hàng mục tiêu từ các loại phân khúc khách hàng khác nhau. Giúp bán hàng và tiếp thị hiệu quả hơn.
* **Hiện trạng:**
* Bộ dữ liệu đã thu thập được các thông tin liên quan đến khách hàng như: MSKH, năm sinh, học vấn, lượt mua trên web và lượt mua trực tiếp, ….
* Dữ liệu được khảo sát từ năm 2012 đến 2014 tại một siêu thị ở Mỹ, bao gồm các thông tin không nhạy cảm về cá nhân và lịch sử mua sắm của khách hàng.
* Tuy nhiên dữ liệu hiện tại của dự án còn đang ở dạng thô, chưa được cấu trúc và phân tích.
* **Mục tiêu :**
* Xác định chân dung khách hàng lý tưởng (dựa trên thu nhập, tình trạng hôn nhân, số con trong gia đình, độ tuổi).
* Hiểu rõ hành vi chi tiêu: Khách hàng chi tiêu nhiều nhất vào danh mục nào (rượu, trái cây, thịt, cá…).
* Hiệu quả của các kênh mua sắm: Khách hàng ưa thích kênh mua hàng nào nhất (web, cửa hàng trực tiếp,…).
* Đánh giá các chiến dịch tiếp thị : Chiến dịch nào nhận được phản hồi tích cực từ khách hàng.
* Xác định tỷ lệ và lý do phàn nàn: Đưa ra các gợi ý để cải thiện chất lượng dịch vụ.
* **Thông tin bộ dữ liệu:**
* Bộ dữ liệu bao gồm 29 trường thông tin về khách hàng (năm sinh, trình độ học vấn, tình trạng hôn nhân, thu nhập hộ gia đình hằng năm, số con trong hộ gia đình, ngày đăng kí, số tiền chi các món hàng,…), phục vụ việc phân tích và dự đoán các hành vi và xu hướng của khách hàng.

|  |  |  |
| --- | --- | --- |
| STT | Tên trường dữ liệu | Nội dung |
| 1 | **ID** | Mã định danh khách hàng. |
| 2 | **Year\_Birth** | Năm sinh khách hàng. |
| 3 | **Education** | Trình độ học vấn của khách hàng. |
| 4 | **Marital\_Status** | Tình trạng hôn nhân của khách hàng. |
| 5 | **Income** | Thu nhập hàng năm của hộ gia đình khách hàng. |
| 6 | **Kidhome** | Số trẻ nhỏ trong hộ gia đình. |
| 7 | **Teenhome** | Số thanh thiếu niên trong hộ gia đình |
| 8 | **Dt\_Customer** | Ngày đăng ký làm khách hàng. |
| 9 | **Recency** | Số ngày từ lần mua gần nhất. |
| 10 | **Complain** | Phàn nàn của khách hàng (1: Có, 0: Không). |
| 11 | **MntWines** | Số tiền chi tiêu cho rượu trong 2 năm. |
| 12 | **MntFruits** | Số tiền chi cho trái cây trong 2 năm. |
| 13 | **MntMeatProducts** | Số tiền chi cho thịt trong 2 năm. |
| 14 | **MntFishProducts** | Số tiền chi cho cá trong 2 năm. |
| 15 | **MntSweetProducts** | Số tiền chi cho đồ ngọt trong 2 năm. |
| 16 | **MntGoldProds** | Số tiền chi cho trang sức trong 2 năm. |
| 17 | **NumDealsPurchases** | Số lần mua hàng với giảm giá. |
| 18 | **NumWebPurchases** | Số lần mua qua website. |
| 19 | **NumCatalogPurchases** | Số lần mua qua danh mục. |
| 20 | **NumStorePurchases** | Số lần mua tại cửa hàng. |
| 21 | **NumWebVisitsMonth** | Số lượt truy cập website hàng tháng. |
| 22-26 | **AcceptedCmp1 - AcceptedCmp5** | Khách hàng chấp nhận chiến dịch tiếp thị (1: Có, 0: Không). |
| 27 | **Response** | Phản hồi tích cực từ khách hàng cho chiến dịch mới nhất. |
| 28-29 | **Z\_CostContact, Z\_Revenue** | Không sử dụng. |

* **Tiềm năng của bộ dữ liệu:**
* **Phân tích hành vi khách hàng:**
  + Dữ liệu bao gồm thông tin nhân khẩu học (tuổi, tình trạng hôn nhân, thu nhập).
  + Các chỉ số mua sắm (mức chi tiêu cho rượu vang, thực phẩm, vàng,…) cho phép xác định mô hình hành vi khách hàng.
* **Cá nhân hóa tiếp thị:**
  + Có thể xác định nhóm khách hàng tiềm năng để cá nhân hóa các chiến dịch marketing.
  + Hiểu rõ hơn về sản phẩm nào được ưu tiên để tối ưu hóa các chiến lược bán hàng.
* **Dự đoán xu hướng:**
  + Sử dụng dữ liệu lịch sử để dự đoán xu hướng tiêu dùng hoặc khả năng phản hồi các chiến dịch trong tương lai.
* **Quản lý khách hàng hiệu quả hơn:**
  + Phân loại khách hàng theo nhóm giá trị để tập trung vào những nhóm có tiềm năng mang lại lợi nhuận cao nhất.
* **Phân tích hiệu quả chiến dịch:**
  + Các cột liên quan đến phản hồi chiến dịch (AcceptedCmp1, AcceptedCmp2, ...) cho phép đo lường hiệu quả và tối ưu hóa các chiến dịch tiếp theo.

### ****Thách thức:****

* **Chất lượng dữ liệu:**
  + Các trường hợp dữ liệu bị thiếu (Income), không nhất quán (định dạng ngày tháng).
  + Dữ liệu hiện tại chưa được cấu trúc, do đó cần được chuyển đổi sang định dạng có thể dễ dàng xử lý và phân tích.
* **Khả năng triển khai phân tích:**
  + Cần phát triển các thuật toán phân tích dữ liệu để trích xuất thông tin có giá trị từ dữ liệu thô
* Cần xây dựng các mô hình dữ liệu có thể dự đoán chính xác hướng phát triển
* **Dữ liệu gốc:**



* **Bộ dữ liệu này là một nguồn tài nguyên giá trị để doanh nghiệp nâng cao hiệu quả tiếp thị và quản lý khách hàng. Tuy nhiên, để phát huy tối đa lợi ích, cần xử lý dữ liệu cẩn thận để đạt được hiểu quả tốt nhất.**

## Yêu cầu của công ty

* **Về mặt dữ liệu:**
* Phân tích và làm rõ hành vi khách hàng, như xác định nhóm tuổi nào chi tiêu nhiều nhất, loại sản phẩm phổ biến, và khu vực có tỷ lệ mua hàng cao.
* Bám sát mục tiêu để đưa ra sự lựa chọn mua sắm tốt nhất cho khách hàng

Bên cạnh đó:

· **Sử dụng ngôn ngữ đơn giản, dễ hiểu:** Tránh sử dụng thuật ngữ chuyên ngành hay ngôn ngữ phức tạp, thay vào đó hãy sử dụng ngôn ngữ dễ hiểu và gần gũi với người nghe.

· **Kết hợp hình ảnh và ví dụ:** Sử dụng biểu đồ, đồ thị, hình ảnh và các ví dụ thực tế để minh họa cho thông tin, giúp người nghe dễ hình dung và ghi nhớ .Làm cho thông điệp muốn nói một cách rõ ràng và hấp dẫn.

· **Truyền tải thông điệp một cách súc tích, rõ ràng:** Nêu bật những điểm chính và thông tin quan trọng, tránh lan man hay sa đà vào chi tiết không cần thiết. Để hiểu được vấn đề nhanh nhất và chính xác nhất.

· **Sử dụng câu chuyện:** Kể những câu chuyện liên quan đến dữ liệu để thu hút sự chú ý và giúp người nghe ghi nhớ thông tin.

· **Nhấn mạnh vào lợi ích:** Giải thích cách thức mà thông tin từ dữ liệu có thể mang lại lợi ích cho công ty, nhân viên và khách hàng.

· **Sử dụng các công cụ trực quan:** Sử dụng các biểu đồ, đồ thị, hình ảnh và video hấp dẫn để trình bày dữ liệu một cách trực quan và dễ hiểu.

* **Quản lý và lưu trữ:**
* Yêu cầu lưu trữ dữ liệu an toàn, dễ dàng truy cập và mở rộng. Đề xuất lưu trữ trên nền tảng đám mây để tiện lợi trong việc quản lý dữ liệu và bảo mật.
* **Mục tiêu:**
* Xây dựng hồ sơ khách hàng lý tưởng từ các phân tích về hành vi mua sắm của khách hàng và khắc phục các tiêu cực của khách hàng đối với doanh nghiệp
* Tối ưu hóa chiến dịch tiếp thị và nâng cao hiệu quả bán hàng bằng cách nhắm vào các nhóm khách hàng với sản phẩm và dịch vụ phù hợp.
* **Năng lực (skill đang có):**
* **Kỹ năng xử lý dữ liệu:**
  + Có khả năng sử dụng Excel và SQL Server để xử lý và truy vấn dữ liệu.
  + Hiểu biết cơ bản để chuyển dữ liệu thô và xử lí để import qua Power BI.
* **Kỹ năng lập trình:**
  + Biết dùng Python trong việc quản lý và phân tích dữ liệu, bao gồm các thư viện như NumPy.
  + Biết cách tổ chức và viết chương trình với các hàm cơ bản Measure trong Power BI (liên quan đến quản lý nhân sự, tính toán).
* **Kỹ năng báo cáo:**
  + Sử dụng Power BI để tạo báo cáo và trực quan hóa dữ liệu.
  + Sử dụng Microsoft Work và Power Point,Canva để báo cáo dự án .
* **Năng lực (skill sẽ học thêm cho dự án):**

**Phân tích nâng cao:**

* + Học thêm về các kỹ thuật phân tích dữ liệu nâng cao, ví dụ: phân cụm (clustering), hồi quy (regression), và các thuật toán học máy cơ bản.
* **Xử lý dữ liệu lớn:**
  + Tìm hiểu các công cụ hỗ trợ xử lý dữ liệu lớn như Excel ,SQL nâng cao.
  + **Làm sạch và chuẩn hóa dữ liệu:** Các bộ dữ liệu thực tế thường có lỗi hoặc thiếu sót, cần biết cách xử lý các vấn đề nà
* **Kỹ năng trực quan hóa nâng cao:**
  + Nâng cao khả năng sử dụng Power BI với các DAX (Data Analysis Expressions) phức tạp và báo cáo động.
  + **Kỹ năng kể chuyện bằng dữ liệu (Data Storytelling):** Biết cách truyền tải thông tin từ dữ liệu một cách rõ ràng, dễ hiểu
  + Tạo báo cáo động và tương tác với công cụ như Power BI
* **Bảo mật và tuân thủ:**
  + Hiểu về các quy định bảo mật dữ liệu để đảm bảo việc phân tích dữ liệu khách hàng tuân thủ quy định pháp luật.
* **Kiến thức về ngành :**
* **Marketing và quản lý khách hàng:** Hiểu cách phân loại và đánh giá giá trị khách hàng theo các chiến lược kinh doanh.
* **Phân tích hành vi khách hàng:** Kiến thức về Customer Lifetime Value (CLV), Customer Retention, và Segmentation
* **Dự án này yêu cầu không chỉ kỹ năng kỹ thuật, mà còn cả tư duy phân tích, khả năng quản lý, và kiến thức chuyên môn sâu trong lĩnh vực liên quan. Việc kết hợp các kỹ năng hiện có và học hỏi những kỹ năng mới là rất cần thiết**

## Lập kế hoạch dự án

**Lập kế hoạch kế hoạch dự án theo mẫu sau**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **TT** | **Hạng mục** | **Bắt đầu** | **Kết thúc** | **Kết quả** |
| 1 | Giới thiệu dự án | 28/10/2024 | 30/10/2024 | Hoàn thành |
| 1.1 | Giới thiệu | 28/10/2024 | 30/10/2024 | Hoàn thành |
| 1.2 | Yêu cầu | 28/10/2024 | 30/10/2024 | Hoàn thành |
| 2 | Phân tích yêu cầu | 01/11/2024 | 04/11/2024 | Hoàn thành |
| 2.1 | Phân tích yêu cầu KH | 01/11/2024 | 04/11/2024 | Hoàn thành |
| 2.2 | Câu chuyện dữ liệu | 04/11/2024 | 06/11/2024 | Hoàn thành |
| 3 | Làm sạch và chuyển đổi dữ liệu | 04/11/2024 | 06/11/2024 | Hoàn thành |
| 3.1 | Chuẩn bị dữ liệu | 04/11/2024 | 06/11/2024 | Hoàn thành |
| 4 | Xử lý dữ liệu | 06/11/2024 | 08/11/2024 | Hoàn thành |
| 4.1 | Chuẩn hóa dữ liệu | 06/11/2024 | 08/11/2024 | Hoàn thành |
| 5 | Trực quan hóa dữ liệu | 11/11/2024 | 13/11/2024 | Hoàn thành |
| 5.1 | Các kĩ thuật trực quan hóa | 11/11/2024 | 13/11/2024 | Hoàn thành |
| 6 | Xây dựng báo cáo | 15/11/2024 | 18/11/2024 | Hoàn thành |
| 6.1 | Dashboard và Report | 18/11/2024 | 20/11/2024 | Hoàn thành |
| 7 | Kết luận | 22/11/2024 | 25/11/2024 | Hoàn thành |
| 7.1 | Báo cáo | 25/11/2024 | 27/11/2024 | Hoàn thành |

# Phân tích yêu cầu khách hàng

## Phân tích yêu cầu

* **Dữ liệu:** 
  + - * Dữ liệu cần được làm sạch và phân tích để đưa ra các đặc điểm chính của từng nhóm khách hàng.
      * Xác định bản chất, khối lượng và định dạng của dữ liệu.
      * Đánh giá chất lượng và độ nhất quán của dữ liệu.
      * Khám phá bất kỳ khoảng trống hoặc hạn chế tiềm ẩn nào của dữ liệu
* **Quản lý và lưu trữ:**
* Lưu trữ trên nền tảng đám mây để đảm bảo an toàn và truy cập nhanh chóng.
* Lập quy trình quản lý dữ liệu để đảm bảo tính toàn vẹn và bảo mật dữ liệu.Áp dụng các thực hành quản lý dữ liệu để điều chỉnh quyền truy cập và sử dụng dữ liệu
* **Công nghệ:**
* Power BI sẽ được sử dụng để trực quan hóa dữ liệu và phân tích sâu hơn. Ngôn ngữ DAX trong Power BI sẽ hỗ trợ tính toán các chỉ số phức tạp.
* **Sử dụng công nghệ:**

1. **Microsoft Power BI:**

* **Trực quan hóa dữ liệu mạnh mẽ:** Biểu đồ, báo cáo, bảng điều khiển dễ hiểu.
* Sử dụng Power Query để làm sạch và chuẩn hóa dữ liệu, đảm bảo chất lượng dữ liệu trước khi phân tích.
* Sử dụng Power BI và DAX để trực quan hóa và phân tích dữ liệu chuyên sâu, cung cấp các báo cáo và chỉ số giúp doanh nghiệp hiểu rõ hơn về tính cách và hành vi của khách hàng
* **Dễ sử dụng:** Giao diện trực quan cho người dùng mọi trình độ.
* **Kết nối dữ liệu đa dạng:** Tích hợp dữ liệu từ nhiều nguồn.
* **Chia sẻ và cộng tác:** Chia sẻ dễ dàng, cộng tác theo thời gian thực.
* **Mở rộng:** Đáp ứng nhu cầu dự án lớn, phức tạp.

**Phù hợp với yêu cầu:**

* **Trực quan:** Dữ liệu dễ hiểu.
* **Dễ sử dụng:** Phù hợp mọi trình độ.
* **Kết nối dữ liệu:** Tích hợp đa nguồn.
* **Chia sẻ/Cộng tác:** Chia sẻ dễ dàng, cộng tác hiệu quả.
* **Mở rộng:** Đáp ứng nhu cầu phát triển.

1. **PowerPoint:**

* **Mục đích:** Trình bày dữ liệu, kể chuyện dữ liệu, giao tiếp insights.
* **Lợi ích:** Công cụ trực quan, giao diện dễ sử dụng, cộng tác, tích hợp Power BI.

1. **Microsoft Word:**

* **Mục đích:** Tạo báo cáo, tài liệu hướng dẫn, giao tiếp bằng văn bản.
* **Lợi ích:** Chỉnh sửa văn bản mạnh mẽ, cộng tác, mẫu và định dạng, tích hợp Power BI.

1. **Microsoft Excel:**

Sử dụng Excel để làm sạch dữ liệu

**Phù hợp với yêu cầu:**

* Giao tiếp: Tạo bài thuyết trình, báo cáo và tài liệu hấp dẫn.
* Báo cáo: Tạo báo cáo chi tiết về dữ liệu.
* Cộng tác: Làm việc hiệu quả với nhóm.

## Câu chuyện dữ liệu

### Đặt vấn đề

* **Mô tả thực trạng:**
* Một chuỗi siêu thị lớn ở Mỹ đang tìm cách hiểu rõ hơn về khách hàng của mình. Bộ dữ liệu hiện tại bao gồm thông tin cá nhân và các hành vi mua sắm của khách hàng trong 2 năm (2020-2021). Mục tiêu của doanh nghiệp là tối ưu hóa các chiến dịch tiếp thị và cải thiện dịch vụ để tăng sự hài lòng và doanh thu.
* Dữ liệu này chứa thông tin cá nhân không nhạy cảm và lịch sử mua hàng, cung cấp nền tảng để hiểu sâu sắc hơn về sở thích và thói quen tiêu dùng của khách hàng

 Doanh nghiệp hiện tại có lượng dữ liệu lớn liên quan đến khách hàng, bao gồm thông tin nhân khẩu học, lịch sử mua hàng, và phản hồi từ các chiến dịch tiếp thị.

 Tuy nhiên, dữ liệu đang gặp một số vấn đề:

* Một số cột bị thiếu thông tin (như Income), có thể gây khó khăn trong việc phân tích chính xác.
* Dữ liệu chưa được chuẩn hóa, dẫn đến việc khó khai thác tiềm năng đầy đủ từ thông tin hiện có.
* Các chiến dịch trước đây chưa được tối ưu hóa, do thiếu khả năng dự đoán chính xác phản ứng của khách hàng.
* **Dữ liệu liên quan:**

Bộ dữ liệu bao gồm các thông tin chính:

* **Thông tin cá nhân:**
  + Tuổi (Year\_Birth), học vấn (Education), tình trạng hôn nhân (Marital\_Status), số con (Kidhome, Teenhome), thu nhập (Income).
* **Hành vi tiêu dùng:**
  + Chi tiêu vào các loại hàng hóa như rượu vang (MntWines), thực phẩm (MntFruits), vàng (MntGoldProds), và các mặt hàng khác.
* **Phản hồi từ chiến dịch:**
  + Kết quả của các chiến dịch tiếp thị trước (AcceptedCmp1, ..., AcceptedCmp5, Response).
* **Tương tác gần đây:**
  + Thời gian kể từ lần mua hàng gần nhất (Recency), ngày trở thành khách hàng (Dt\_Customer)
* **Mục tiêu:**

 **Hiểu rõ khách hàng:**

* Phân tích hành vi mua hàng, từ đó xác định nhóm khách hàng tiềm năng và các yếu tố ảnh hưởng đến quyết định mua sắm.

 **Tối ưu hóa chiến dịch tiếp thị:**

* Dự đoán khả năng khách hàng sẽ phản hồi tích cực với các chiến dịch trong tương lai.
* Phân bổ nguồn lực tiếp thị vào các phân khúc khách hàng hiệu quả nhất.

 **Cải thiện trải nghiệm khách hàng:**

* Cá nhân hóa sản phẩm và dịch vụ dựa trên sở thích và lịch sử mua sắm của khách hàng.

 **Hỗ trợ ra quyết định kinh doanh:**

* Tối ưu hóa danh mục sản phẩm và lập kế hoạch bán hàng dựa trên xu hướng tiêu dùng.
* Đưa ra dự đoán chính xác hơn về doanh thu và chi phí tiếp thị.

### Xác định câu chuyện

* **Dữ liệu công ty cung cấp cho thấy:**
* Mục tiêu là hiểu rõ hành vi khách hàng và xác định các yếu tố ảnh hưởng đến quyết định mua sắm của họ. Chúng tôi muốn giải thích vì sao một số khách hàng có xu hướng chi tiêu nhiều hơn, tham gia vào các chiến dịch tiếp thị nhiều hơn, hoặc có mức độ trung thành cao hơn so với những người khác.
* Câu chuyện dữ liệu sẽ cố gắng giải thích mối quan hệ giữa các yếu tố như thu nhập, độ tuổi, tình trạng hôn nhân và hành vi mua sắm. Ngoài ra, sẽ phân tích các yếu tố có thể tác động đến sự thành công của các chiến dịch marketing, như việc phản hồi của khách hàng đối với các chiến dịch tiếp thị trước đó.
* **Mục tiêu cụ thể của tôi khi kể câu chuyện dữ liệu này:**
* Mục tiêu chính là tối ưu hóa chiến lược tiếp thị bằng cách phân tích sâu hành vi tiêu dùng và phản hồi của khách hàng. Câu chuyện dữ liệu sẽ giúp xác định những khách hàng tiềm năng và những sản phẩm hoặc dịch vụ nào có khả năng mang lại lợi nhuận cao nhất.
* Ngoài ra, mục tiêu cũng bao gồm việc dự đoán hành vi của khách hàng trong tương lai, từ đó nâng cao khả năng phân bổ ngân sách marketing và tạo các chiến dịch quảng cáo hiệu quả hơn.
* **Và từ đó, tôi có vài giải pháp từ những phân tích này như sau :**
* Tập trung vào các nhóm khách hàng có giá trị cao: Sử dụng phân tích dữ liệu để tìm ra các phân khúc khách hàng có tỷ lệ mua sắm cao hoặc phản hồi tích cực với các chiến dịch trước.
* Cá nhân hóa các chiến dịch tiếp thị: Dựa trên hành vi và sở thích tiêu dùng của khách hàng, có thể đề xuất việc cá nhân hóa các chiến dịch marketing (ví dụ: gửi khuyến mãi phù hợp với sở thích của nhóm khách hàng cụ thể).
* Tối ưu hóa ngân sách tiếp thị: Đưa ra đề xuất về việc phân bổ ngân sách tiếp thị dựa trên những phân khúc khách hàng có phản hồi cao và doanh thu cao.
* **Một số cách tiếp cận dữ liệu**
* **Mối tương quan**

Tìm mối liên hệ giữa các yếu tố như thu nhập, độ tuổi, tình trạng hôn nhân và hành vi chi tiêu. Ví dụ, liệu khách hàng có thu nhập cao có xu hướng chi tiêu nhiều vào các sản phẩm cao cấp không?

Kiểm tra mối quan hệ giữa các chiến dịch tiếp thị và sự phản hồi của khách hàng. Liệu có sự tương quan giữa số lần tham gia chiến dịch và mức độ chi tiêu của khách hàng?

* **Xác định xu hướng:**

Phân tích xu hướng tiêu dùng theo thời gian: Khách hàng có xu hướng chi tiêu nhiều vào một số sản phẩm trong các mùa cụ thể không?

Xác định xu hướng liên quan đến các chiến dịch marketing: Các chiến dịch quảng cáo nào có xu hướng tạo ra sự tăng trưởng trong doanh thu?

* **So sánh:**

So sánh hành vi mua sắm giữa các nhóm khách hàng khác nhau, ví dụ: khách hàng có thu nhập cao và thấp, hoặc nhóm khách hàng đã phản hồi chiến dịch và nhóm chưa tham gia.

So sánh hiệu quả của các chiến dịch marketing theo thời gian, từ đó đánh giá chiến dịch nào đem lại hiệu quả nhất.

* **Phân tích phân khúc khách hàng:**

Sử dụng dữ liệu để phân chia khách hàng thành các nhóm dựa trên các yếu tố như độ tuổi, thu nhập, và mức độ tiêu dùng. Phân tích xem nhóm nào có giá trị cao nhất và có xu hướng mua sắm nhiều nhất.

* **Dự đoán hành vi khách hàng:**

Sử dụng các mô hình học máy để dự đoán khả năng phản hồi của khách hàng đối với các chiến dịch trong tương lai. Câu chuyện sẽ bao gồm việc tìm kiếm các yếu tố ảnh hưởng đến sự phản hồi của khách hàng (chẳng hạn như mức thu nhập hoặc độ tuổi).

* **Các kỹ thuật phân tích dữ liệu như tìm mối tương quan, xác định xu hướng, và phân khúc khách hàng sẽ cung cấp cái nhìn sâu sắc về hành vi tiêu dùng và khả năng tương tác của khách hàng, từ đó đưa ra các giải pháp nhằm tối ưu hóa chi phí và gia tăng lợi nhuận**

### Xác định rõ đối tượng

Trong câu chuyện dữ liệu này, **đối tượng nghiên cứu** chính là **khách hàng của doanh nghiệp**, đặc biệt là các khách hàng đã từng tham gia vào các chiến dịch tiếp thị trước đây và có dữ liệu mua sắm rõ ràng. Cụ thể, các nhóm khách hàng này có thể được phân loại dựa trên các yếu tố như:

* **Thông tin nhân khẩu học**: độ tuổi, thu nhập, tình trạng hôn nhân, số con cái, …
* **Hành vi tiêu dùng**: các sản phẩm mà khách hàng đã chi tiêu, mức độ chi tiêu vào từng loại sản phẩm (ví dụ: rượu vang, thực phẩm, vàng).
* **Tương tác với chiến dịch marketing**: các chiến dịch tiếp thị mà khách hàng đã tham gia, mức độ phản hồi, và kết quả của những chiến dịch này (ví dụ: "AcceptedCmp1", "AcceptedCmp2",...).

Việc chọn các khách hàng đã tham gia vào chiến dịch tiếp thị trước đó là bởi vì họ có **dữ liệu hành vi tiêu dùng và mức độ phản hồi** đã được ghi nhận. Điều này giúp việc phân tích trở nên chính xác hơn, vì ta có thể:

* Xem xét các hành vi tiêu dùng trong quá khứ của họ.
* Đánh giá sự thành công của các chiến dịch tiếp thị trước đó và xác định các yếu tố ảnh hưởng đến quyết định tham gia.

Ngoài ra, nhóm khách hàng có mức thu nhập cao hoặc chi tiêu nhiều sẽ giúp doanh nghiệp tập trung vào các phân khúc mang lại **lợi nhuận cao** nhất, tối ưu hóa chi phí marketing.

* **Các nhóm khách hàng cụ thể có thể xác định:**
* **Khách hàng tiềm năng**: Những khách hàng có **thu nhập cao**, nhưng lại không tham gia vào các chiến dịch trước. Việc này có thể giúp tìm ra những khách hàng tiềm năng cho các chiến dịch tiếp thị trong tương lai.
* **Khách hàng trung thành**: Các khách hàng **thường xuyên mua sắm** hoặc tham gia nhiều chiến dịch marketing, có thể sẽ tiếp tục phản hồi tốt với các chiến dịch trong tương lai.
* **Khách hàng không phản hồi**: Nhóm khách hàng này có thể có giá trị cao về mặt tiềm năng, nhưng chưa tham gia vào các chiến dịch tiếp thị trước đó. Phân tích nhóm này giúp tìm ra lý do tại sao họ không tham gia và cách thức thu hút họ trong tương lai.
* **Nhân viên của chuỗi cửa hàng:**
* Mức độ hiểu biết: hiểu biết cơ bản về cửa hàng
* Nhu cầu: Sử dụng dữ liệu để thực hiện công việc hàng ngày, chẳng hạn như phân tích hành vi khách hàng, đưa ra khuyến nghị đầu tư, cũng như chiến lược makerting để bán được nhiều hàng hơn.
* **Ban lãnh đạo của chuỗi cửa hàng:**
* Mức độ hiểu biết: hiểu rõ cửa hàng trong tình trạng như thế nào và định hướng phát triển.
* Nhu cầu: Sử dụng dữ liệu để đưa ra quyết định chiến lược tiếp thị và kinh doanh hiệu quả hơn, chẳng hạn như định hướng đầu tư, phát triển sản phẩm, mở rộng thị trường.
* **Nhà đầu tư tiềm năng:**
* Mức độ hiểu biết: Có kiến thức cơ bản về thị trường và đang tìm kiếm thông tin đầu tư.
* Nhu cầu: Đánh giá tiềm năng của chuỗi cửa hàng có phát triển hay không để quyết định đầu tư.
* **Những đối tượng này có hiểu biết nền tảng về thị trường và khách hàng nhưng cần thêm các phân tích chi tiết và minh họa từ dữ liệu để ra quyết định chính xác hơn.**

### Xác định câu chuyện chi tiết

* **Bối cảnh**:

Trong bối cảnh thị trường cạnh tranh cao, việc hiểu rõ khách hàng và tối ưu hóa chiến lược tiếp thị là cần thiết để chuỗi siêu thị tăng cường trải nghiệm khách hàng và cải thiện doanh thu.

* **Vấn đề cần giải quyết:**

Dữ liệu được thu thập từ các chiến dịch tiếp thị của doanh nghiệp, và mục tiêu chính là **tối ưu hóa các chiến lược marketing** và hiểu rõ hơn về hành vi tiêu dùng của khách hàng. Câu chuyện dữ liệu này muốn trả lời các câu hỏi chính như:

* Khách hàng là ai?
* Họ có xu hướng chi tiêu vào những sản phẩm gì?
* Làm thế nào để doanh nghiệp có thể **tối ưu hóa chiến dịch marketing** và **nhắm đến khách hàng tiềm năng** một cách hiệu quả hơn?
* **Lý do phân tích dữ liệu:**  
  Việc phân tích dữ liệu sẽ giúp doanh nghiệp hiểu rõ hơn về **hành vi của khách hàng** (từ các yếu tố như độ tuổi, thu nhập, mức độ chi tiêu, phản hồi chiến dịch marketing). Điều này sẽ giúp cải thiện các chiến lược tiếp thị, nâng cao hiệu quả và gia tăng doanh thu.
* **Mục tiêu của câu chuyện dữ liệu:**  
  Tìm ra các **phân khúc khách hàng** có khả năng mang lại lợi nhuận cao nhất, từ đó **đưa ra các chiến lược marketing nhắm mục tiêu** để tối ưu hóa hiệu quả chiến dịch và gia tăng doanh thu.
* **Đối tượng cần xem báo cáo :**
* **Quản lý tiếp thị và đội ngũ marketing:**  
  Báo cáo sẽ rất hữu ích cho những người làm trong lĩnh vực marketing để tối ưu hóa chiến dịch tiếp thị, hiểu được khách hàng của họ và xác định những phân khúc mang lại lợi nhuận cao nhất. Các nhà quản lý cũng có thể sử dụng thông tin này để **phân bổ ngân sách tiếp thị hợp lý**, đảm bảo chiến lược tiếp cận khách hàng được tối ưu.
* **Nhà quản lý cấp cao:**  
  Các giám đốc hoặc lãnh đạo doanh nghiệp sẽ quan tâm đến báo cáo này để **đưa ra quyết định chiến lược** về việc phát triển các sản phẩm, cải thiện dịch vụ khách hàng và tối ưu hóa các kênh marketing.
* **Phân tích dữ liệu (Data Analysts):**  
  Đội ngũ phân tích dữ liệu sẽ sử dụng báo cáo này để hiểu rõ hơn về cách thức dữ liệu được tổng hợp và phân tích, nhằm thực hiện các phân tích chi tiết hơn hoặc xây dựng các mô hình dự đoán dựa trên dữ liệu khách hàng.
* **Quy trình:**
* **Thu thập dữ liệu:** Thu thập dữ liệu từ nhiều nguồn khác nhau, đảm bảo tính chính xác và đầy đủ.
* **Làm sạch dữ liệu:** Loại bỏ dữ liệu lỗi, trùng lặp và không liên quan.
* **Chuyển đổi dữ liệu:** Chuyển đổi dữ liệu sang định dạng phù hợp cho việc phân tích.
* **Phân tích dữ liệu:** Sử dụng các kỹ thuật phân tích dữ liệu để tìm kiếm thông tin chi tiết có giá trị.
* **Trực quan hóa dữ liệu:** Biểu diễn dữ liệu dưới dạng biểu đồ, đồ thị, báo cáo dễ hiểu.

### Trình bày dữ liệu

 **Biểu đồ cột (Bar Chart):**So sánh giá trị giữa các nhóm (ví dụ: chi tiêu theo độ tuổi, thu nhập).

 **Biểu đồ tròn (Pie Chart):**Thể hiện tỷ lệ phần trăm của các thành phần (ví dụ: phân bổ chi tiêu vào các loại sản phẩm).

 **Biểu đồ đường (Line Chart):**Thể hiện xu hướng thay đổi theo thời gian (ví dụ: sự thay đổi trong chi tiêu theo tháng).

 **Biểu đồ phân tán (Scatter Plot)** Thể hiện mối quan hệ giữa hai biến số (ví dụ: thu nhập và chi tiêu).

 **Biểu đồ hộp (Box Plot):**Hiển thị sự phân bố dữ liệu và phát hiện giá trị ngoại lệ.

· Đảm bảo biểu đồ dễ đọc và dễ hiểu.

· Sử dụng chú thích và trục chú thích rõ ràng.

· Kết hợp màu sắc và biểu tượng một cách hiệu quả để làm nổi bật thông tin quan trọng.

### Những điều cần lưu ý

* **Chính xác và khách quan**: Đảm bảo rằng các biểu đồ và số liệu không gây nhầm lẫn và phản ánh đúng tình trạng thực tế.
* **Trực quan và dễ hiểu**: Các hình ảnh cần dễ đọc, rõ ràng, và có chú thích phù hợp để giúp người xem hiểu nhanh chóng.
* **Bám sát mục tiêu của câu chuyện**: Các phần của báo cáo cần hướng tới giải quyết vấn đề và đáp ứng nhu cầu của người đọc, không nên bao gồm các chi tiết không liên quan.
* **Chọn đúng loại biểu đồ:**Lựa chọn biểu đồ phù hợp với mục đích phân tích (ví dụ: biểu đồ cột cho so sánh, biểu đồ tròn cho tỷ lệ phần trăm, biểu đồ đường cho xu hướng theo thời gian).
* **Đảm bảo tính trực quan và dễ hiểu:**Biểu đồ cần dễ đọc, không quá phức tạp, và các yếu tố như tiêu đề, trục, và chú thích phải rõ ràng.
* **Giải thích dữ liệu một cách ngắn gọn:**Mỗi biểu đồ cần có phần giải thích ngắn gọn, làm rõ thông điệp mà dữ liệu muốn truyền tải.
* **Tránh lạm dụng quá nhiều biểu đồ:**Không sử dụng quá nhiều loại biểu đồ khác nhau trong một báo cáo, gây rối mắt và khó hiểu. Cần chọn lựa những biểu đồ thực sự cần thiết.
* **Chú ý đến độ chính xác của dữ liệu:**Đảm bảo dữ liệu được xử lý đúng cách và không có sai sót hoặc thiếu sót trong quá trình phân tích.
* **Tính nhất quán trong việc sử dụng màu sắc:**Sử dụng màu sắc một cách hợp lý và nhất quán để phân biệt các nhóm hoặc phân tích khác nhau, tránh gây nhầm lẫn cho người xem.

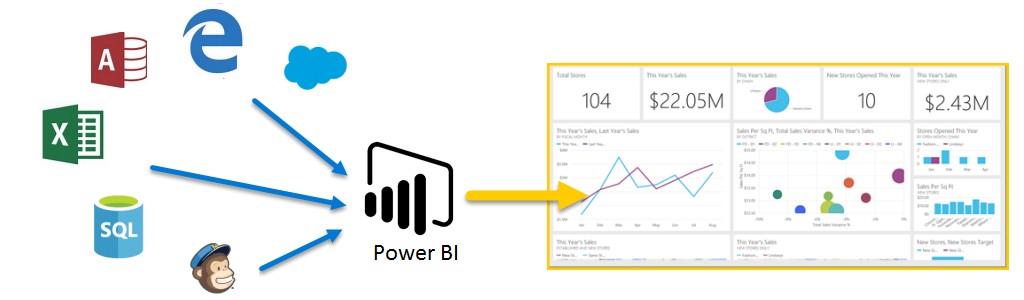
**Ngoài ra, cần lưu ý:**

* Hiểu rõ đối tượng mục tiêu và trình bày dữ liệu một cách phù hợp với nhu cầu và trình độ của họ.
* Xác định rõ mục đích sử dụng dữ liệu và trình bày dữ liệu một cách phục vụ mục đích đó.
* Thu thập phản hồi từ người xem và điều chỉnh cách trình bày dữ liệu cho phù hợp.

## Kiến trúc hệ thống

### Kiến trúc

Giữ liệu gốc trong file Excel 🡪 Power BI (phân tích dữ liệu) 🡪 Hoàn thiện báo cáo trên Word



### Giải thích

**Giải thích sơ đồ:**

* **Các nguồn dữ liệu (bên trái):**
  + **Excel :** Đây là những file Excel chứa dữ liệu nguồn. Có thể là các bảng tính chứa số liệu bán hàng, thông tin khách hàng, hoặc bất kỳ loại dữ liệu phân tích.
* **Power Query (biểu tượng ống nghiệm):**
  + Đây là một công cụ mạnh mẽ trong Power BI, cho phép bạn kết nối với các nguồn dữ liệu khác nhau, trích xuất dữ liệu, làm sạch và biến đổi dữ liệu trước khi đưa vào Power BI để tạo báo cáo.
  + Trong sơ đồ, Power Query đóng vai trò là một cầu nối, kết nối các nguồn dữ liệu đa dạng (Excel) và đưa chúng vào một môi trường thống nhất để xử lý.
* **Power BI (biểu tượng biểu đồ):**
  + Đây là công cụ trực quan hóa dữ liệu, cho phép bạn tạo các báo cáo, bảng điều khiển (dashboard) tương tác để khám phá và trình bày dữ liệu một cách hiệu quả.
  + Sau khi dữ liệu được chuẩn bị bởi Power Query, nó sẽ được đưa vào Power BI để tạo các biểu đồ, đồ thị, bảng và các hình thức trực quan khác nhau.

**Cách thức hoạt động:**

1. **Kết nối dữ liệu:** Power Query được sử dụng để kết nối với các nguồn dữ liệu khác nhau (Excel )và trích xuất dữ liệu.
2. **Làm sạch và biến đổi dữ liệu:** Dữ liệu được làm sạch và biến đổi để đảm bảo chất lượng và phù hợp với mục đích phân tích. Các hoạt động này có thể bao gồm:
   * Loại bỏ các hàng hoặc cột không cần thiết
   * Thay thế các giá trị thiếu
   * Tính toán các trường mới
   * Sắp xếp và lọc dữ liệu
3. **Tạo mô hình dữ liệu:** Dữ liệu được tổ chức và liên kết với nhau để tạo ra một mô hình dữ liệu. Mô hình dữ liệu này sẽ được sử dụng làm cơ sở để tạo các báo cáo và bảng điều khiển.
4. **Tạo báo cáo và bảng điều khiển:** Trong Power BI, bạn có thể tạo các báo cáo và bảng điều khiển tương tác để khám phá và trình bày dữ liệu. Các báo cáo này có thể bao gồm các biểu đồ, đồ thị, bảng, bản đồ và các hình thức trực quan khác nhau.
5. **Chia sẻ và hợp tác:** Các báo cáo và bảng điều khiển có thể được chia sẻ với các thành viên khác trong nhóm để cùng nhau phân tích và đưa ra quyết định.

## Giải thích về bộ dữ liệu khách hàng

### Các khái niệm

Giải thích các khái niệm và nghiệp vụ liên quan:

 **Dữ liệu khách hàng:** Là các thông tin về khách hàng được thu thập để phục vụ việc phân tích và hiểu rõ hành vi tiêu dùng của họ. Dữ liệu này có thể bao gồm thông tin cá nhân (như tên, tuổi, địa chỉ) và thông tin về giao dịch (như mức chi tiêu, sản phẩm mua).

 **Phân khúc khách hàng:** Là quá trình phân loại khách hàng thành các nhóm dựa trên các đặc điểm chung, giúp hiểu rõ hơn về nhu cầu và hành vi của từng nhóm.

 **Hành vi khách hàng:** Là các thói quen, xu hướng trong việc mua sắm và sử dụng sản phẩm/dịch vụ của khách hàng. Dữ liệu hành vi có thể bao gồm tần suất mua hàng, loại sản phẩm yêu thích, phản hồi từ khách hàng.

### Các trường dữ liệu

Giải thích các trường dữ liệu có trong bộ dữ liệu:

 **ID khách hàng (Customer ID):** Mã số duy nhất của khách hàng, dùng để phân biệt các khách hàng trong hệ thống.

 **Tên khách hàng (Name):** Tên đầy đủ của khách hàng.

 **Giới tính (Gender):** Giới tính của khách hàng (Nam, Nữ, Khác).

 **Độ tuổi (Age):** Tuổi của khách hàng, giúp phân tích hành vi và nhu cầu theo độ tuổi.

 **Địa chỉ (Address):** Địa chỉ nơi cư trú của khách hàng, giúp xác định vị trí địa lý và các chiến lược marketing phù hợp.

 **Số điện thoại (Phone):** Cung cấp phương thức liên lạc với khách hàng.

 **Email:** Địa chỉ email để gửi các thông báo hoặc quảng cáo.

 **Lịch sử giao dịch (Transaction History):** Dữ liệu về các lần mua hàng của khách hàng, bao gồm sản phẩm, số lượng, thời gian và giá trị giao dịch.

 **Tình trạng khách hàng (Customer Status):** Trạng thái của khách hàng, ví dụ: mới, khách hàng tiềm năng, khách hàng cũ, khách hàng VIP.

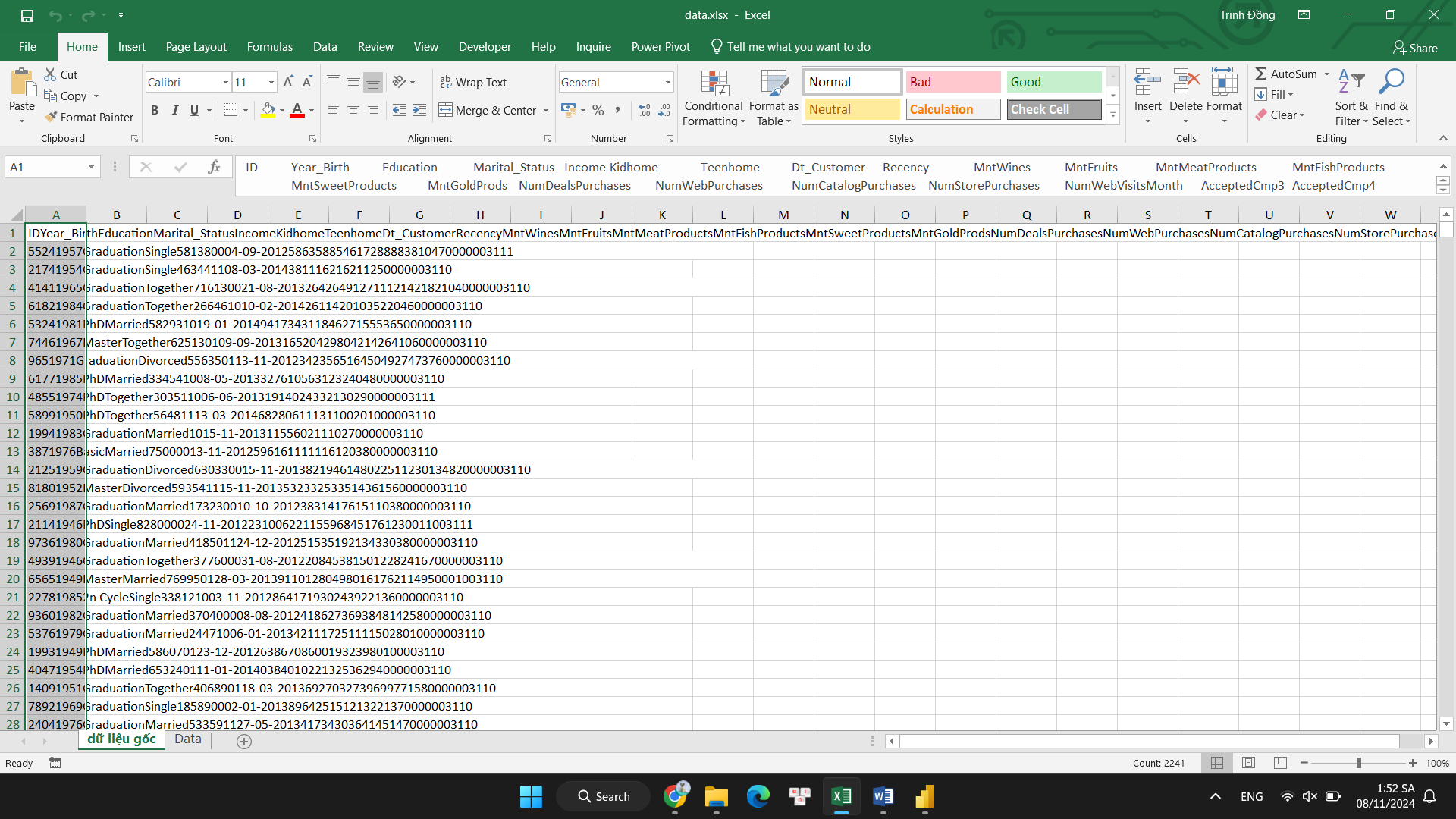
# Làm sạch và chuyển đổi dữ liệu

Bao gồm 2 bước:

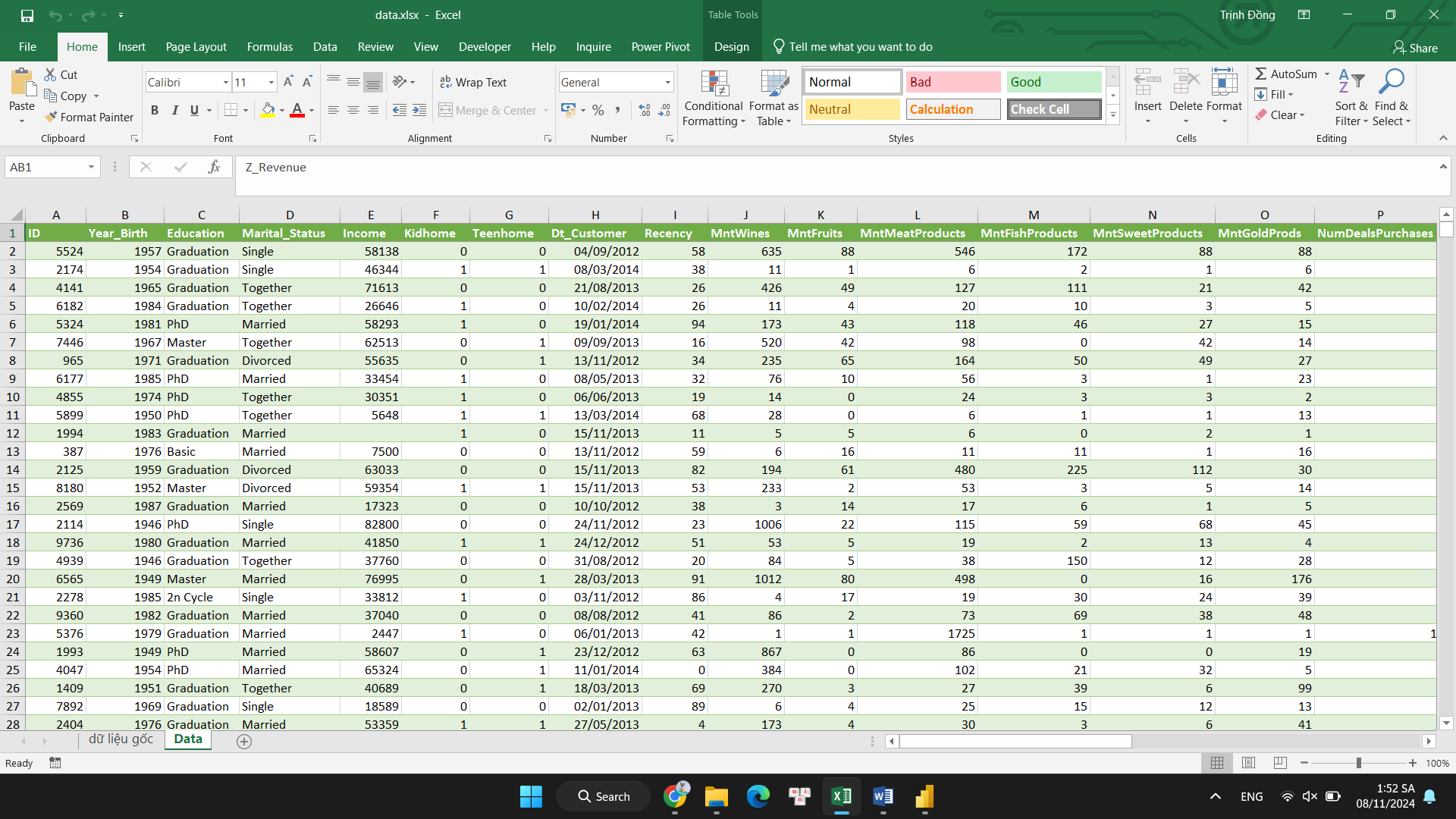
* Làm sạch dữ liệu: Xử lý các dữ liệu thiếu, sai lệch, trùng lặp và không chính xác để đảm bảo tính chất lượng của dữ liệu trước khi sử dụng.
* Chuyển đổi dữ liệu: thực hiện các thao tác chuyển đổi dữ liệu để đảm bảo dữ liệu ở dạng phù hợp cho phân tích và sử dụng trong mô hình.

## Chuẩn bị dữ liệu

Dữ liệu ban đầu:



Dữ liệu sau khi được làm sạch và tách cột:



### Giải pháp lưu trữ dữ liệu

**So sánh giữa giải pháp đám mây và on-premise**:

* **Giải pháp nền tảng đám mây (Cloud-based solution)**: cho phép lưu trữ và truy cập dữ liệu từ bất kỳ đâu, có khả năng mở rộng linh hoạt, không yêu cầu chi phí duy trì cơ sở hạ tầng. Các nhà cung cấp dịch vụ đám mây như Microsoft Azure, AWS, Google Cloud thường đảm bảo tính bảo mật cao.
* **Ứng dụng tại chỗ (On-premise)**: phù hợp cho những dự án cần quản lý dữ liệu nghiêm ngặt, yêu cầu quyền kiểm soát hoàn toàn đối với hạ tầng. Tuy nhiên, yêu cầu chi phí cao cho phần cứng và bảo trì.

**Quyết định**: Lựa chọn nền tảng đám mây, vì yêu cầu của dự án là an toàn, truy cập tiện lợi và khả năng mở rộng linh hoạt.

### Giải pháp phân bố dữ liệu

Phân bố dữ liệu là một kỹ thuật quan trọng giúp tối ưu hóa việc truy xuất, quản lý và bảo mật cơ sở dữ liệu (CSDL). Đặc biệt khi dữ liệu có khối lượng lớn hoặc yêu cầu truy xuất từ nhiều nguồn khác nhau. Nhân bản và phân phối dữ liệu hiệu quả giúp cải thiện hiệu suất và độ sẵn sàng của hệ thống

#### Ý nghĩa việc phân bố dữ liệu

Phân bố dữ liệu giúp cải thiện hiệu quả hoạt động của hệ thống CSDL, đặc biệt là trong các môi trường phân tán hoặc khi dữ liệu có khối lượng lớn. Việc phân bố dữ liệu mang lại những lợi ích sau:

1. **Tăng cường hiệu suất:**
   * Phân phối dữ liệu vào các phần nhỏ hơn, giúp cải thiện thời gian phản hồi và giảm độ trễ khi truy vấn dữ liệu.
2. **Dễ dàng quản lý và bảo trì:**
   * Việc phân chia dữ liệu theo từng khu vực hoặc từng nhóm giúp việc bảo trì và sao lưu dữ liệu dễ dàng hơn.
3. **Tối ưu hóa khả năng chịu tải (scalability):**
   * Khi cần tăng thêm dung lượng hoặc mở rộng, việc phân bố dữ liệu cho phép hệ thống mở rộng linh hoạt mà không làm giảm hiệu suất.
4. **Dự phòng và độ sẵn sàng cao:**
   * Nhân bản dữ liệu và phân bố trên các máy chủ khác nhau giúp đảm bảo dữ liệu luôn sẵn có, ngay cả khi một phần của hệ thống gặp sự cố

#### Trình bày cách phân bố dữ liệu

 **Phân bố dữ liệu theo vùng địa lý (Geographical Partitioning):**

* Dữ liệu có thể được phân phối và lưu trữ ở các vị trí khác nhau tùy theo yêu cầu về vị trí địa lý của người dùng hoặc nhóm người dùng (ví dụ: phân bố dữ liệu khách hàng theo khu vực).
* **Ưu điểm:** Tăng hiệu suất truy xuất và giảm độ trễ khi người dùng ở các khu vực khác nhau truy xuất dữ liệu.

 **Phân bố dữ liệu theo nhóm (Sharding):**

* Dữ liệu có thể được chia thành các mảnh nhỏ hơn và lưu trữ trên các máy chủ khác nhau, mỗi mảnh sẽ chứa một phần của dữ liệu. Ví dụ: dữ liệu của từng khách hàng hoặc sản phẩm có thể được phân bổ vào các bảng khác nhau.
* **Ưu điểm:** Giảm tải cho mỗi máy chủ, cải thiện khả năng mở rộng khi dữ liệu lớn.

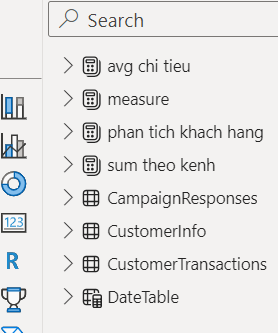
 **Nhân bản dữ liệu (Replication):**

* Tạo các bản sao của cơ sở dữ liệu hoặc các phần của cơ sở dữ liệu và phân phối chúng đến các máy chủ khác. Các bản sao giúp giảm tải cho hệ thống chính và nâng cao độ sẵn sàng của dữ liệu.
* **Ưu điểm:** Tăng cường độ tin cậy, đảm bảo dữ liệu có thể truy xuất từ các máy chủ khác nếu máy chủ chính gặp sự cố.

 **Phân bố theo loại dữ liệu (Data-based Partitioning):**

* Dữ liệu có thể được phân chia dựa trên loại dữ liệu hoặc các chỉ số phân loại, chẳng hạn như phân chia dữ liệu khách hàng theo độ tuổi hoặc các mức độ chi tiêu.
* **Ưu điểm:** Giúp truy vấn dữ liệu theo từng nhóm dễ dàng và hiệu quả hơn
* **Nhân bản dữ liệu** sẽ là giải pháp chính trong phạm vi dự án này. Việc nhân bản dữ liệu không chỉ giúp đảm bảo tính sẵn có mà còn giúp phân phối tải trong môi trường phân tán. Dữ liệu sẽ được phân bổ và nhân bản trên nhiều máy chủ để nâng cao hiệu suất truy cập và bảo mật.

**Phân bố theo bảng :**



## Làm sạch dữ liệu

### Các vấn đề ảnh hưởng tới dữ liệu

 **Dữ liệu thiếu và không đồng nhất:** Dữ liệu khách hàng có thể thiếu thông tin quan trọng như số điện thoại, email, hay dữ liệu giao dịch không đầy đủ, gây khó khăn trong việc phân tích chính xác.

 **Dữ liệu sai lệch và trùng lặp:** Các trường hợp khách hàng có thể được nhập nhiều lần dưới các định dạng khác nhau, dẫn đến sự không nhất quán trong thông tin.

### Các tiêu chí đánh giá chất lượng dữ liệu

1. **Độ chính xác (Accuracy):**
   * Đảm bảo dữ liệu chính xác và phản ánh đúng thực tế. Ví dụ: tên khách hàng, số điện thoại, và các thông tin quan trọng phải được nhập đúng và đầy đủ.
2. **Độ đầy đủ (Completeness):**
   * Dữ liệu phải đầy đủ, không có trường dữ liệu quan trọng nào bị thiếu hoặc bỏ sót. Ví dụ: mỗi khách hàng phải có đầy đủ thông tin như tên, địa chỉ, và thông tin giao dịch.
3. **Độ nhất quán (Consistency):**
   * Dữ liệu phải nhất quán trong suốt quá trình thu thập và lưu trữ. Không có sự khác biệt giữa các hệ thống hoặc bộ dữ liệu khác nhau.
4. **Độ hợp lệ (Validity):**
   * Dữ liệu phải hợp lệ và tuân thủ các quy chuẩn đã đặt ra, ví dụ như các trường số không được chứa chữ cái, các trường ngày phải tuân thủ định dạng hợp lệ.
5. **Độ truy cập (Accessibility):**
   * Dữ liệu phải có thể truy cập và sử dụng khi cần thiết. Các quyền truy cập phải được xác định rõ ràng và dễ dàng cho người dùng.
6. **Độ bảo mật (Security):**
   * Dữ liệu cần được bảo mật và bảo vệ khỏi truy cập trái phép, đặc biệt là thông tin nhạy cảm như dữ liệu khách hàng.

#### ****Tiêu chí áp dụng trong dự án này:****

* **Độ chính xác:** Cần đảm bảo thông tin khách hàng và các giao dịch là chính xác và không sai lệch.
* **Độ đầy đủ:** Dữ liệu cần có đầy đủ các trường thông tin để phục vụ cho phân tích.
* **Độ nhất quán:** Dữ liệu cần được chuẩn hóa và thống nhất giữa các nguồn khác nhau để đảm bảo dễ dàng phân tích và báo cáo.

### Các bước làm sạch dữ liệu

#### Trình bày các bước làm sạch

Các bước làm sạch cụ thể:

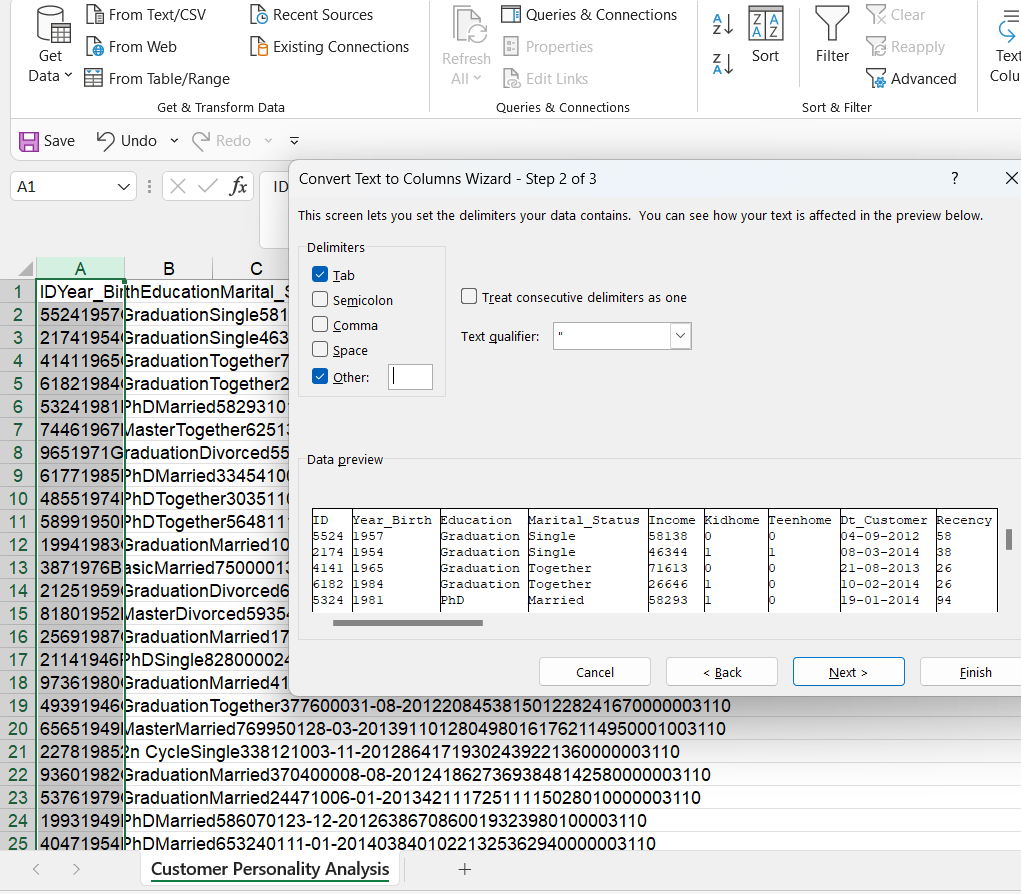
* Trong Excel sử dụng Text to column để tách các cột

**Kết quả :**

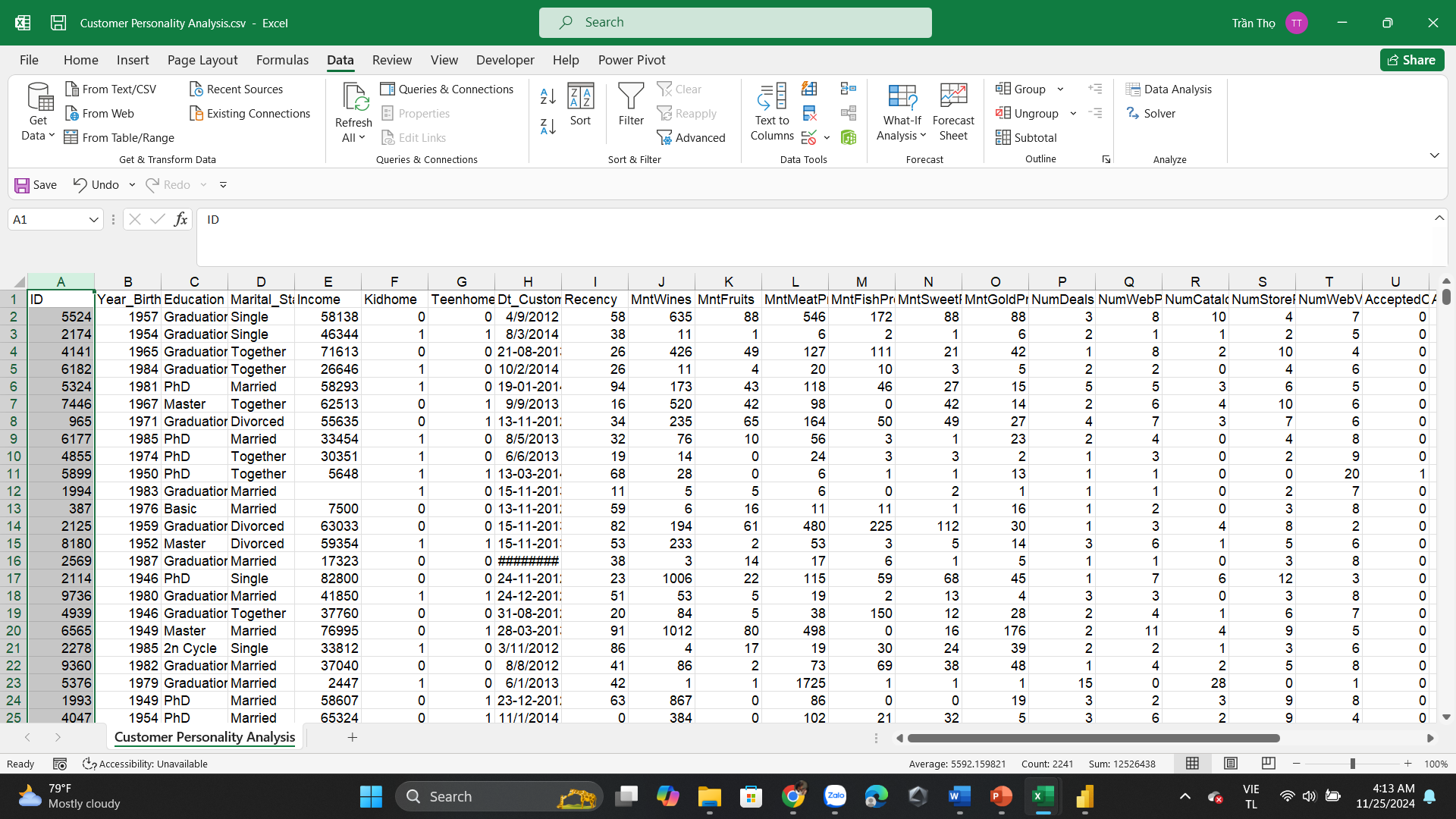
* Căn chỉnh các cột ,dòng, phông chữ,…
* Add dữ liệu vào Power BI để làm sạch và phân tích
* Xác định và xóa các bản ghi và xóa các cột không cần thiết.
* Tách cột để tiện theo dõi dữ liệu hơn
* Chuẩn hóa các trường dữ liệu như ngày tháng (đảm bảo cùng một định dạng).
* Chuyển các dạng DataType

#### Trình bày các bước làm sạch trong phạm vi dự án

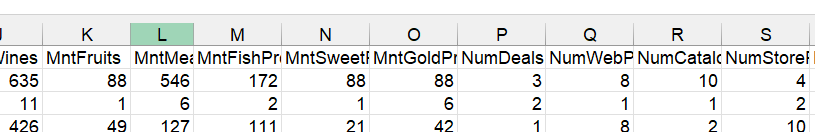
Dùng Text to column để tách dữ liệu :



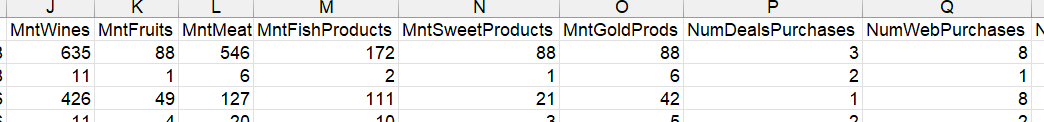
**Kết quả :**



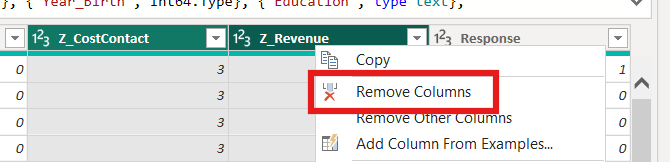
Căn chỉnh các cột cho đủ để hiện dữ liệu :



**Kết quả:**



* Xóa các cột không cần thiết: Xóa các cột **Z\_CostContact** và **Z\_Revenue** vì chúng không đóng góp thông tin cho phân tích.



## Chuyển đổi dữ liệu

### Các trường hợp cần chuyển đổi

Trường hợp cần chuyển đổi:

* Chuyển đổi các giá trị về giới tính, tình trạng hôn nhân thành dạng mã số để dễ phân tích.
* Chuyển đổi cột ngày tháng về định dạng chuẩn để dễ dàng phân tích theo thời gian.
* Chuyển đổi các cột định dạng không đồng nhất và dữ liệu thô thành dạng phù hợp.

### Các kỹ thuật chuyển đổi

Công cụ và kỹ thuật:

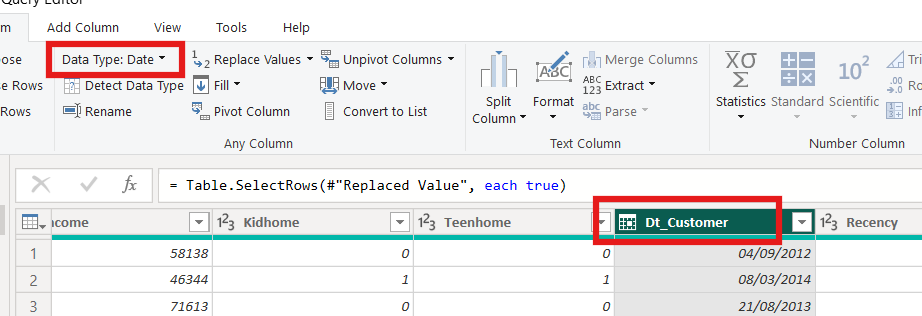
* Sử dụng Power Query trong Power BI để chuẩn hóa và sắp xếp lại các cột, loại bỏ các giá trị ngoại lệ.

Kỹ thuật chuyển đổi:

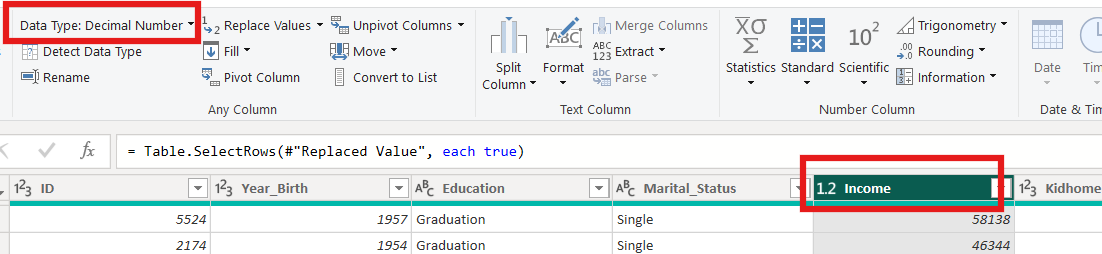
* Thay đổi kiểu dữ liệu
* Chuẩn hóa giá trị dữ liệu
* Loại bỏ các dữ liệu trùng lặp
* Sắp xếp và lặp chỉ mục

### Trình bày các phép chuyển đổi trong dự án

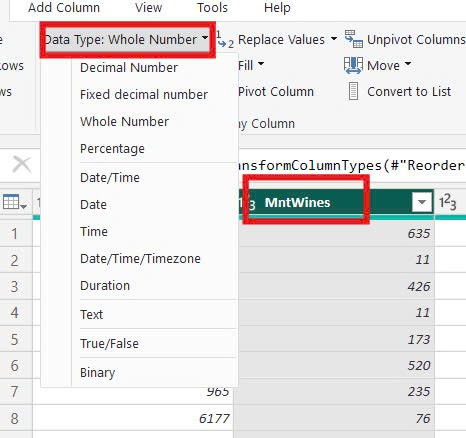
Đổi định dạng của cột Dt\_Customer thành Date để dễ tính toán khoảng thời gian khách hàng đã là thành viên.



Đổi các cột số như Income cần ở định dạng số để đảm bảo tính toán chính xác.

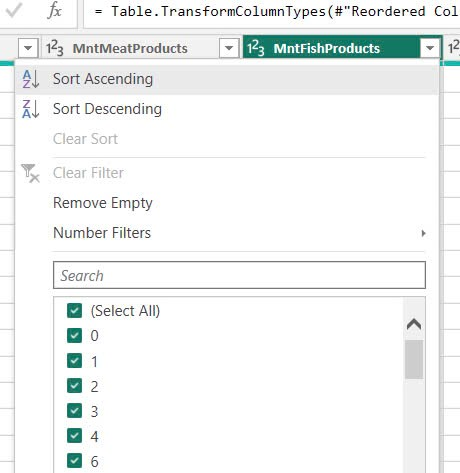


Đổi cột MntWines thành dạng số để chỉnh sửa dữ liệu hợp lý :



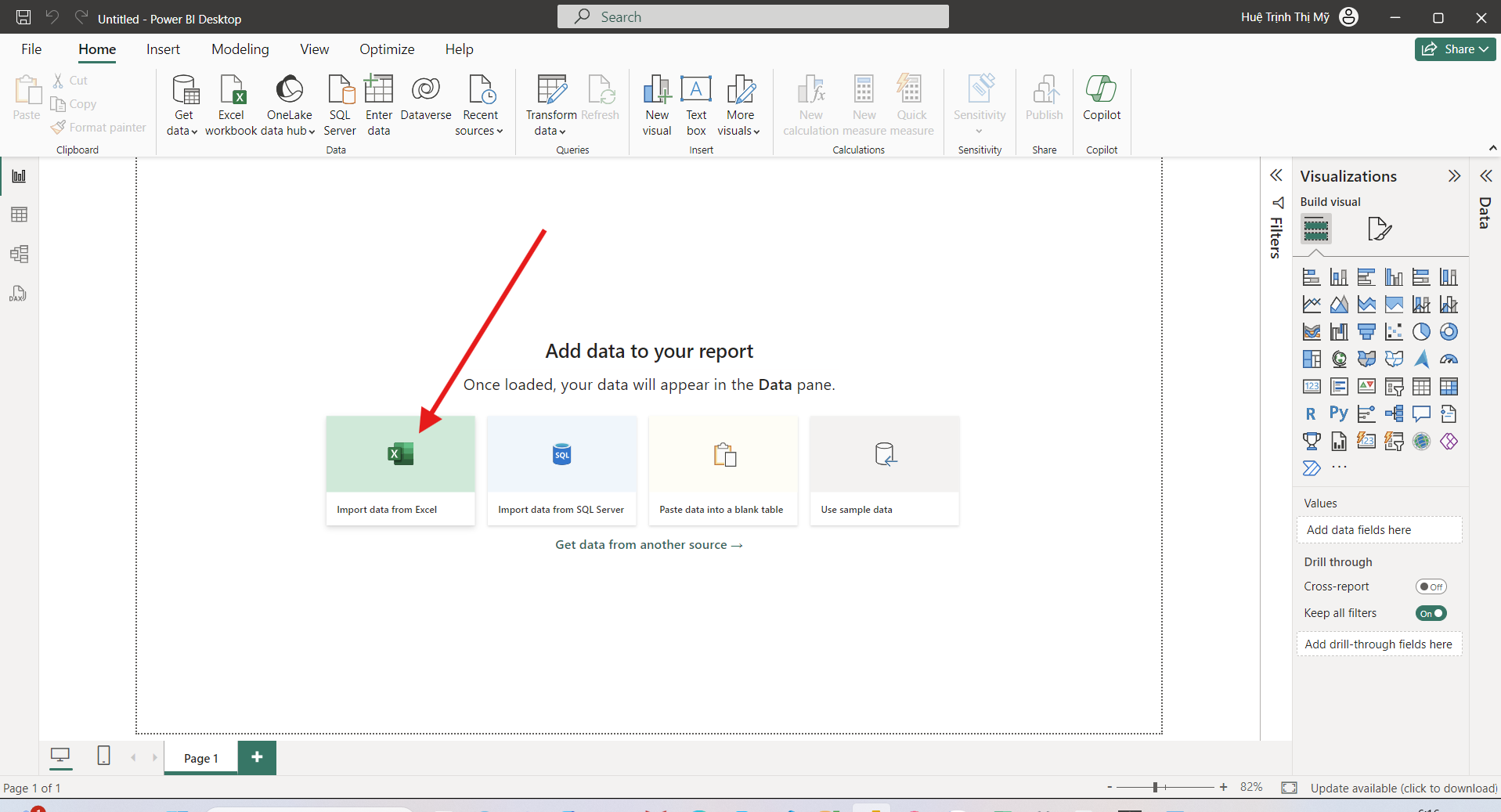
**Xử lý dữ liệu bị thiếu**:

* Các cột chi tiêu (MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds): Đảm bảo không có giá trị âm hoặc ngoại lai bất thường.

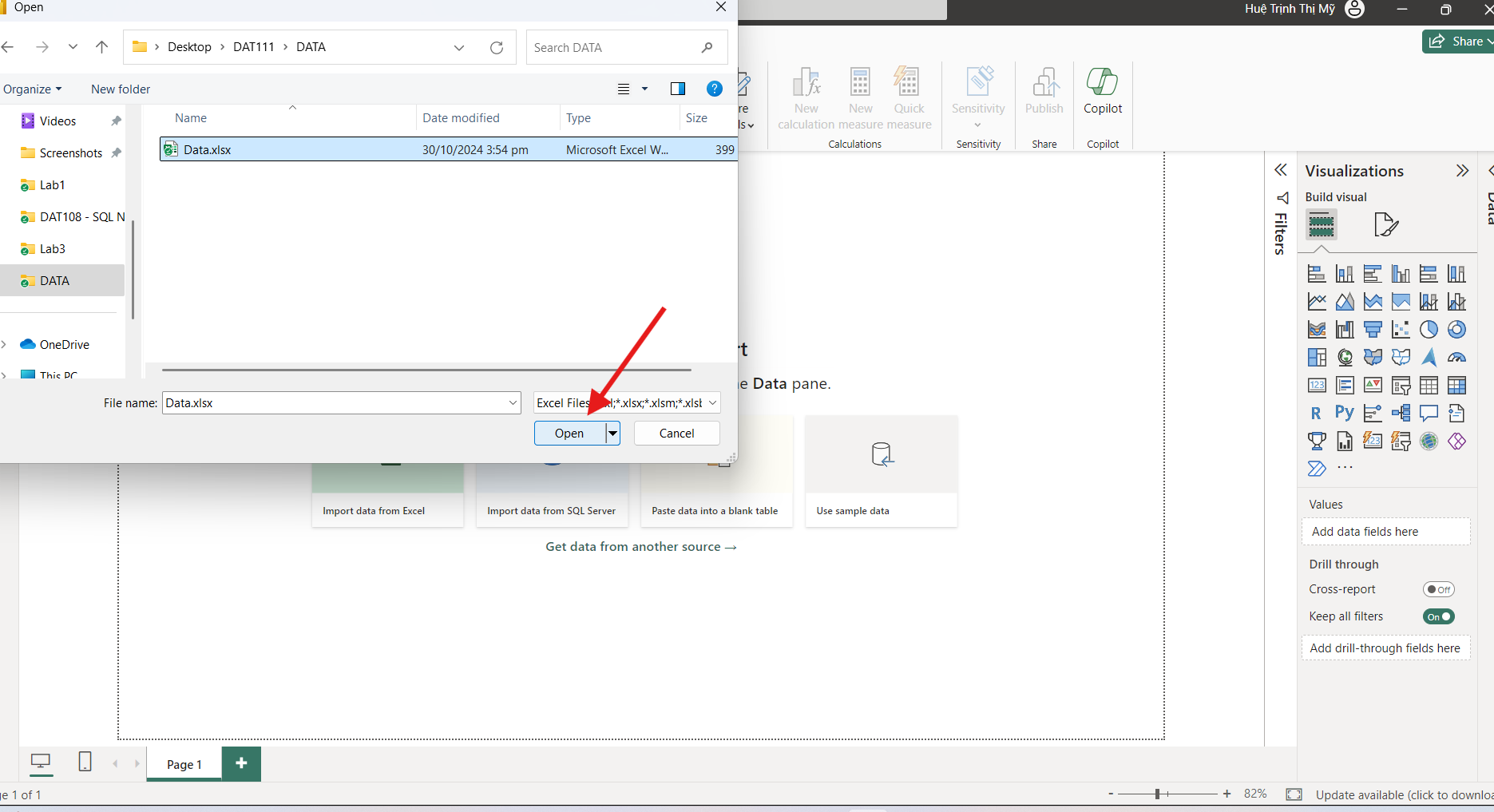


*Sau khi đã copy dữ liệu ở Excel để bảo toàn dữ liệu không bị thay đổi trong qua trình chỉnh sửa => Import Dữ liệu cần chỉnh sửa vào Power BI như sau:*

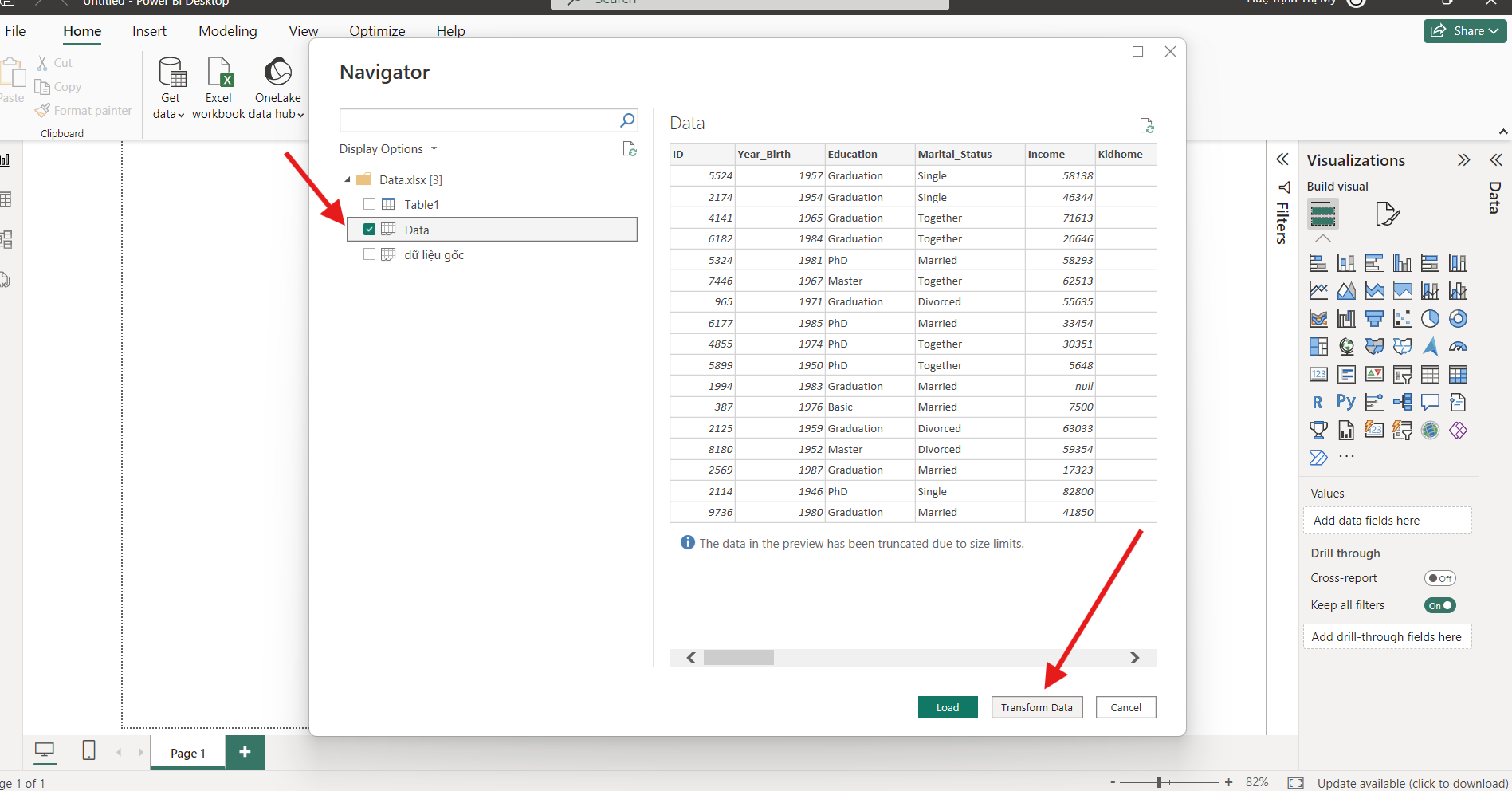
* Mở Power BI Desktop.
* Chọn Import data from Excel



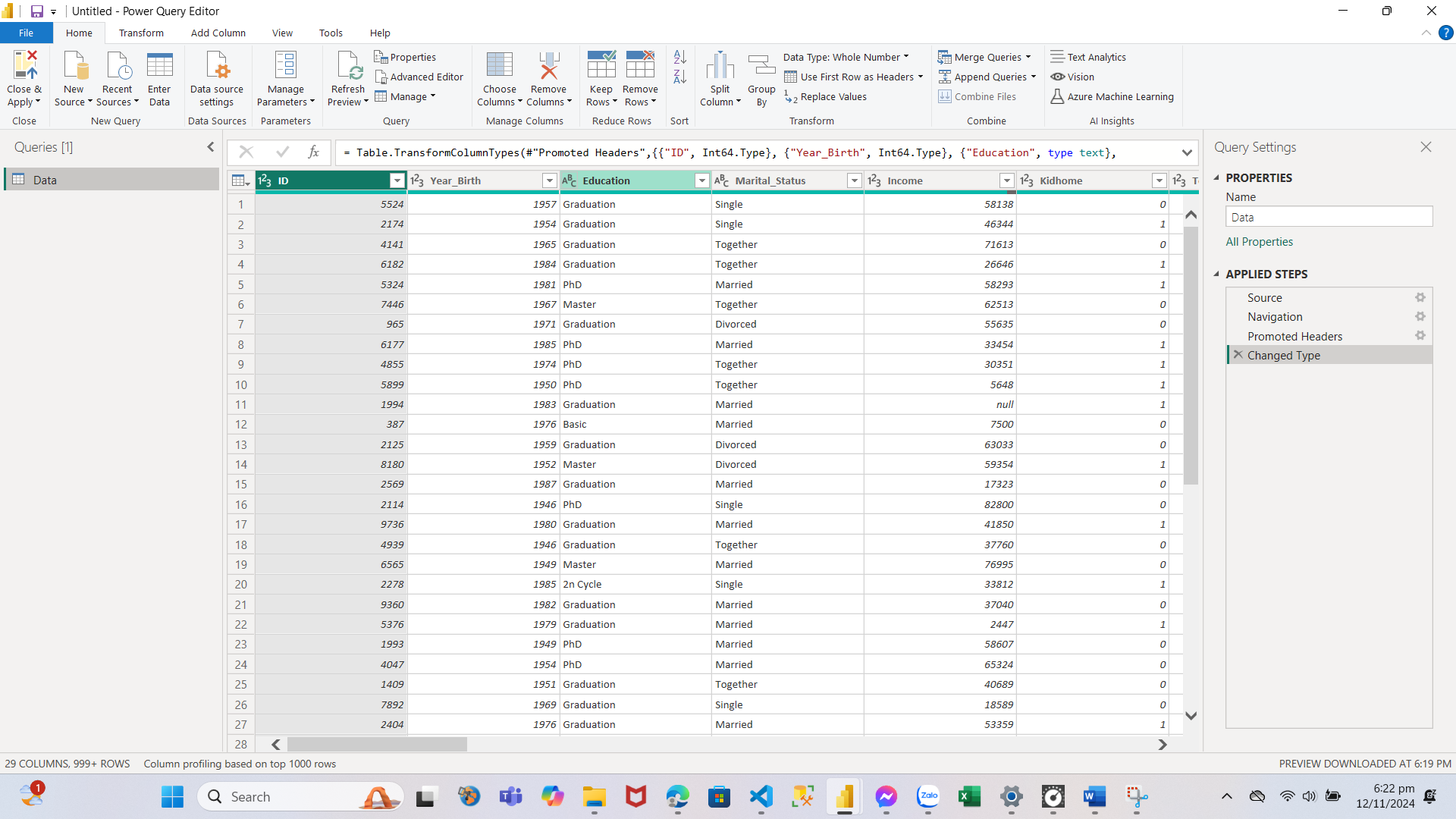
* Chọn File dữ liệu. Nhấn nút Open



* Chọn Data. Nhấn nút Transform Data.



Kết quả:



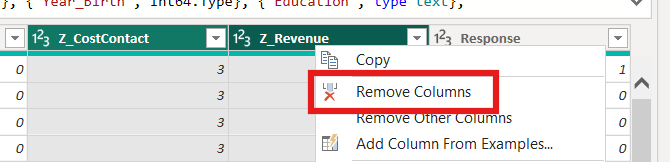
# Xử lý dữ liệu

## Chuẩn hóa dữ liệu

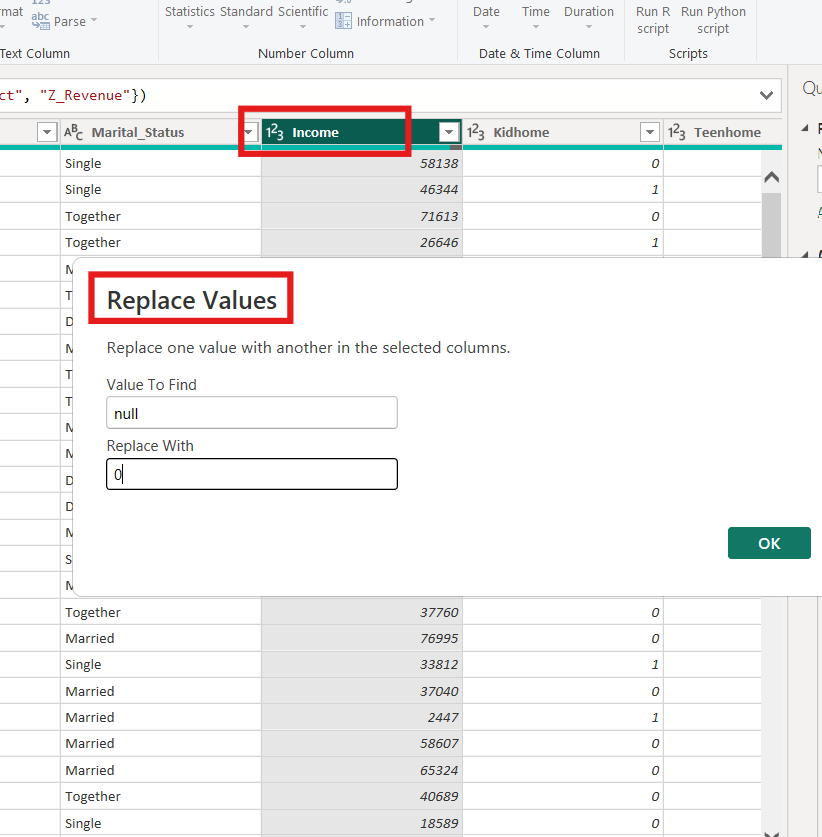
### Trình bày các bước chuẩn hóa trong dự án

Import data qua power BI để tiếp tục làm sạch, phân tích và trực quan:

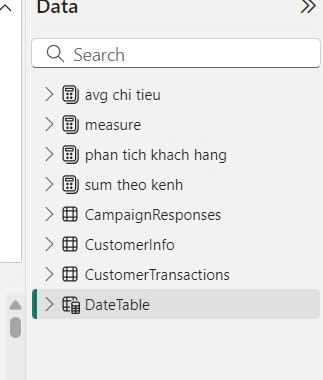
* Xóa các cột không cần thiết: Xóa các cột **Z\_CostContact** và **Z\_Revenue** vì chúng không đóng góp thông tin cho phân tích.



* **Cột Income**: Kiểm tra các giá trị bị thiếu, nếu cần thiết có thể thay bằng trung bình hoặc giá trị phổ biến nhất trong dữ liệu.Thay thế dữ liệu Null là 0.



Rename bảng dữ liệu gốc là Data để giữ nguyên dữ liệu gốc so sánh.

Từ Data gốc, tạo thành ba bảng để dễ dàng chỉnh sửa : 

## Mô hình hóa dữ liệu

* Mô hình dữ liệu đảm bảo tính nhất quán trong quy ước đặt tên, giá trị mặc định, ngữ nghĩa, bảo mật đồng thời đảm bảo chất lượng của dữ liệu.
* Data model (mô hình dữ liệu) là một mô hình trừu tượng tổ chức các phần tử của dữ liệu và chuẩn hóa cách chúng liên quan với nhau và với các thuộc tính của các thực thể trong thế giới thực.
* Mô hình hóa dữ liệu (data modeling) là quá trình tạo ra một mô hình dữ liệu để dữ liệu được lưu trữ trong CSDL. Mô hình dữ liệu này là một biểu diễn khái niệm của các đối tượng dữ liệu, sự liên kết giữa các đối tượng dữ liệu khác nhau và các quy tắc.

### Các loại mô hình hóa

* **Mô hình Dữ liệu Khái niệm (Conceptual Data Model)**
* **Khái niệm:** Như đã đề cập, đây là cấp độ trừu tượng nhất, tập trung vào việc xác định các thực thể chính và mối quan hệ giữa chúng trong hệ thống.
* **Đặc điểm:**
  + **Trừu tượng cao:** Không quan tâm đến chi tiết kỹ thuật như kiểu dữ liệu, độ dài trường.
  + **Dễ hiểu:** Dễ dàng cho người không chuyên về kỹ thuật nắm bắt.
  + **Ổn định:** Thường ít thay đổi so với các mô hình ở cấp độ thấp hơn.
* **2. Mô hình Dữ liệu Logic (Logical Data Model)**
* **Khái niệm:** Đây là bước chuyển tiếp từ cấp độ trừu tượng sang cấp độ cụ thể hơn. Mô hình này xác định cấu trúc của dữ liệu, bao gồm các thuộc tính, khóa chính, khóa ngoại và các ràng buộc.
* **Đặc điểm:**
  + **Cụ thể hơn:** Xác định rõ các thuộc tính của mỗi thực thể.
  + **Vẫn còn trừu tượng:** Không phụ thuộc vào một hệ quản trị cơ sở dữ liệu cụ thể.
  + **Cung cấp nền tảng:** Là cơ sở để xây dựng mô hình vật lý.
* **Ví dụ:** Tiếp tục ví dụ thư viện, thuộc tính của thực thể Sách có thể bao gồm: Mã sách (khóa chính), Tên sách, Tác giả (khóa ngoại), Năm xuất bản.
* **Biểu diễn:** Cũng sử dụng biểu đồ ER, nhưng chi tiết hơn mô hình khái niệm.
* **3. Mô hình Dữ liệu Vật lý (Physical Data Model)**
* **Khái niệm:** Đây là cấp độ chi tiết nhất, mô tả cách dữ liệu được lưu trữ trong cơ sở dữ liệu cụ thể.
* **Đặc điểm:**
  + **Cụ thể:** Xác định rõ kiểu dữ liệu, kích thước trường, chỉ số, ràng buộc, phân vùng, ...
  + **Phụ thuộc vào hệ quản trị:** Mỗi hệ quản trị có các đặc điểm riêng.
  + **Hiệu suất:** Mục tiêu là tối ưu hóa hiệu suất truy vấn và lưu trữ.
* **Ví dụ:** Trong SQL Server, bảng Sách sẽ được tạo với các cột tương ứng, kiểu dữ liệu và các chỉ số phù hợp.
* **Biểu diễn:** Thường sử dụng ngôn ngữ định nghĩa dữ liệu (DDL) của hệ quản trị cơ sở dữ liệu.

**Bổ sung các loại mô hình khác:**

* **Mô hình dữ liệu hướng đối tượng (Object-Oriented Data Model):** Dựa trên các khái niệm đối tượng, lớp, thừa kế, được sử dụng rộng rãi trong các ngôn ngữ lập trình hướng đối tượng.
* **Mô hình dữ liệu dạng mạng (Network Data Model):** Sử dụng cấu trúc mạng để biểu diễn mối quan hệ giữa các thực thể.
* **Mô hình dữ liệu phân cấp (Hierarchical Data Model):** Tổ chức dữ liệu theo cấu trúc cây.

### Các tiêu chí đánh giá mô hình dữ liệu

**Clearness (Tính rõ ràng)**

* Sự dễ hiểu đối với những người sử dụng.
* Hầu hết thời gian developers đọc mã thay vì viết.
* Vì vậy chúng ta cần hiểu rõ ràng những gì chúng ta đang làm với dữ liệu của mình.

**Flexibility (Tính linh hoạt)**

* Khả năng phát triển của mô hình mà không cần phải tác động quá lớn đến các đoạn code.
* Công ty startup mà bạn làm việc đang phát triển, vì vậy các hệ thống sẽ thay đổi và các mô hình dữ liệu đằng sau sẽ cần phải phát triển theo thời gian.

**Performance (Hiệu suất)**

* Đây là một chủ đề rất rộng.
* Trong môn học này sẽ không nói về các nhà cung cấp cơ sở dữ liệu (database vendors) hoặc một số chỉnh sửa kỹ thuật để cải thiện tốc độ đọc và ghi dữ liệu.
* Cách thức thiết kế data model đúng đắn cũng đem lại lợi ích về hiệu suất.

**Productivity (Năng suất)**

* Dưới góc nhìn của lập trình viên (developer), chắc hẳn bạn sẽ muốn có một mô hình dữ liệu dễ làm việc mà không cần sử dụng nhiều thời gian.

**Traceability (Khả năng truy xuất nguồn gốc)**

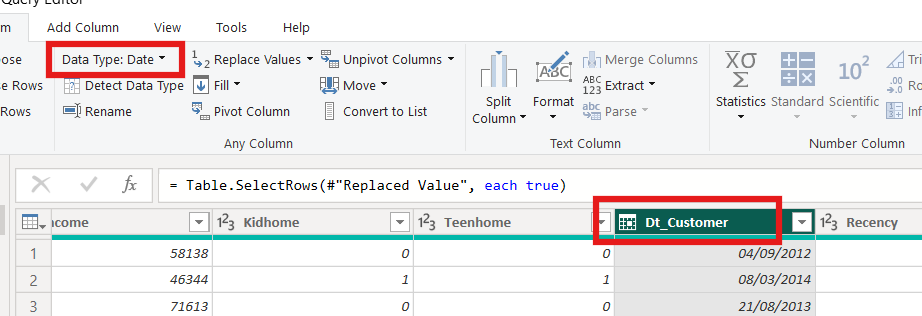
* Cuối cùng, các công ty không chỉ muốn có dữ liệu liên quan đến người dùng của mình mà còn có dữ liệu liên quan đến chính hệ thống.
* Dữ liệu có thể cung cấp thông tin những gì đã xảy ra trong quá khứ, những giá trị công ty có tại một thời điểm nào đó

### Trình bày các bước mô hình hóa

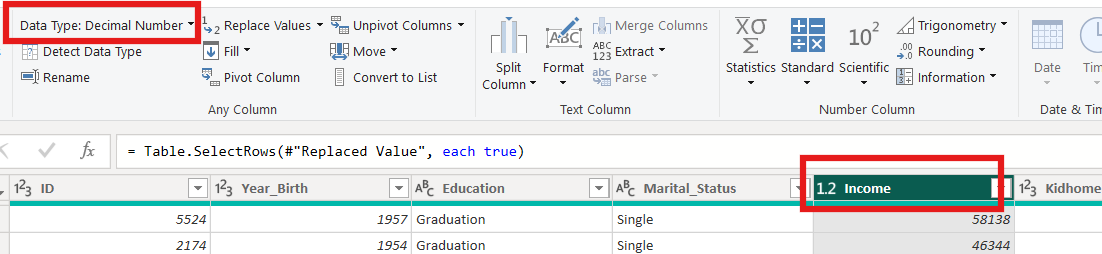
Import data qua power BI để tiếp tục làm sạch, phân tích và trực quan:

**Chuẩn hóa định dạng dữ liệu**:

* Đổi định dạng của cột Dt\_Customer thành Date để dễ tính toán khoảng thời gian khách hàng đã là thành viên.



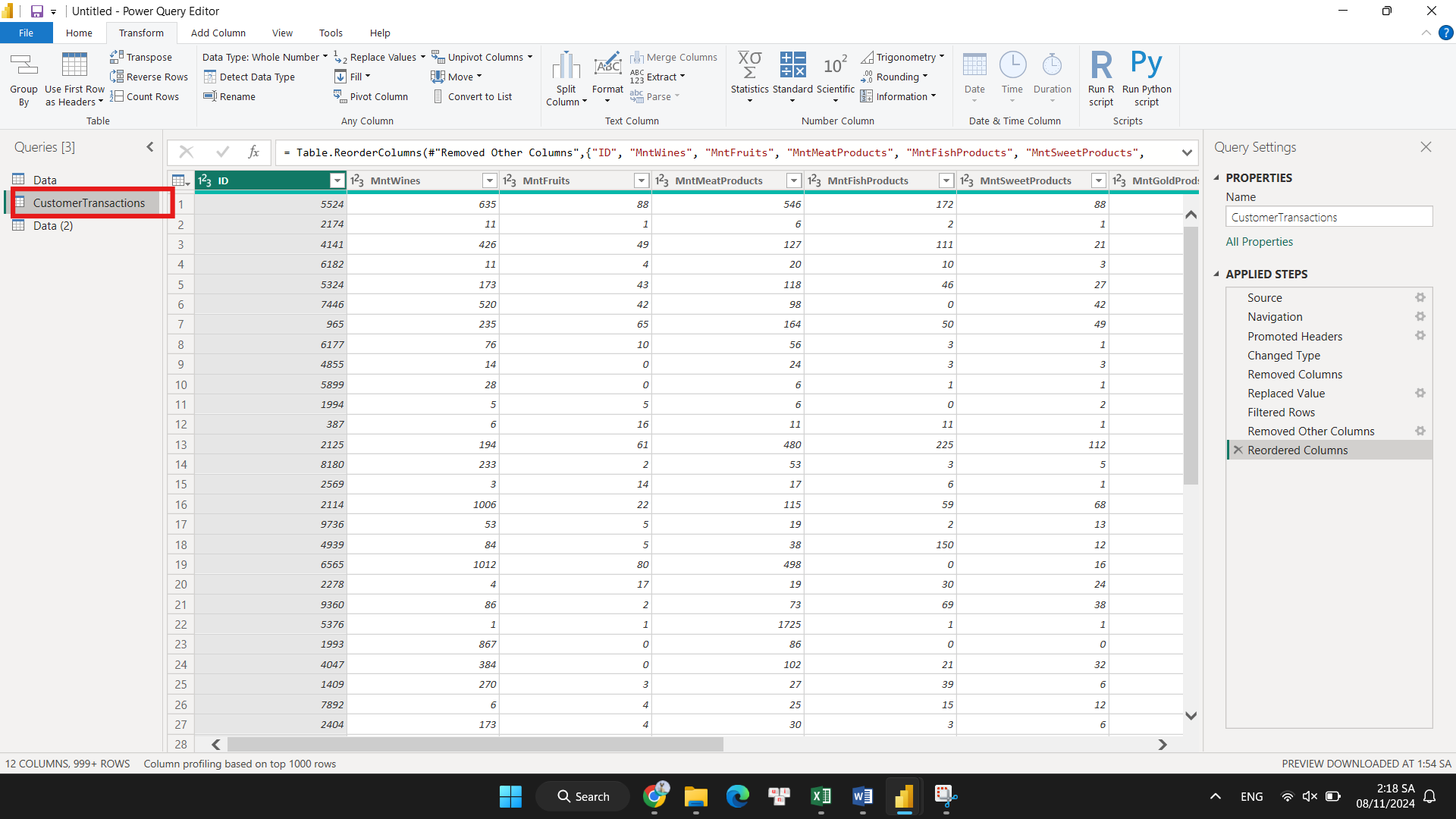
* Các cột số như Income, MntWines, NumWebPurchases,... cần ở định dạng số để đảm bảo tính toán chính xác.



### Trình bày các bước tạo bảng dữ liệu

**Bảng Fact chính: CustomerTransactions**

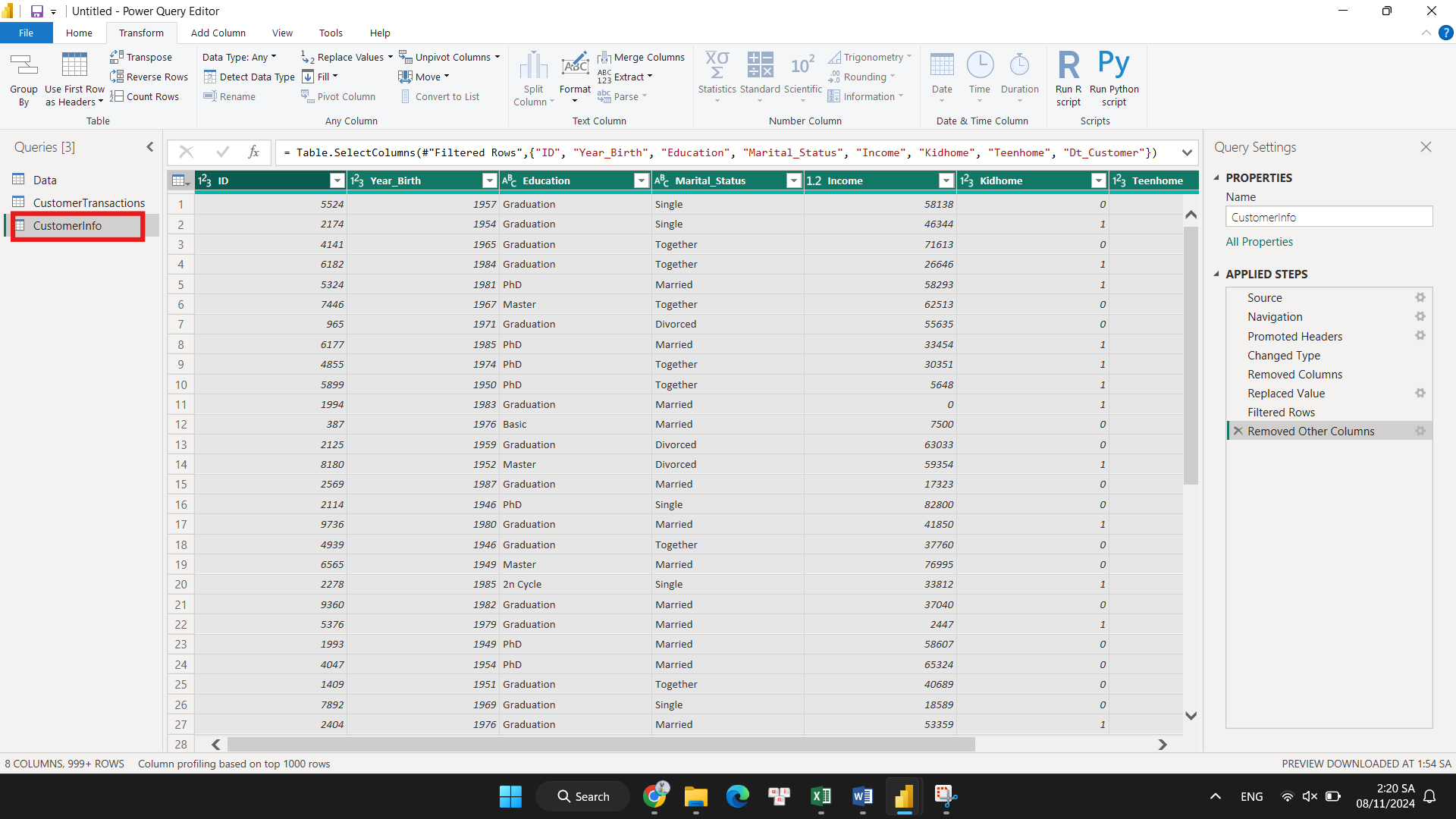
* **Các cột trong bảng**:
* ID (Mã định danh khách hàng)
* MntWines (Số tiền chi cho rượu trong 2 năm qua)
* MntFruits (Số tiền chi cho trái cây trong 2 năm qua)
* MntMeatProducts (Số tiền chi cho thịt trong 2 năm qua)
* MntFishProducts (Số tiền chi cho cá trong 2 năm qua)
* MntSweetProducts (Số tiền chi cho đồ ngọt trong 2 năm qua)
* MntGoldProds (Số tiền chi cho vàng trang sức trong 2 năm qua)
* NumDealsPurchases (Số lần mua hàng có khuyến mãi)
* NumWebPurchases (Số lần mua hàng qua web)
* NumCatalogPurchases (Số lần mua hàng qua danh mục)
* NumStorePurchases (Số lần mua hàng tại cửa hàng)
* NumWebVisitsMonth (Số lần truy cập trang web trong tháng trước)



**Bảng CustomerInfo (Thông tin Khách Hàng)**

Bảng này chứa các thông tin cá nhân của khách hàng, giúp phân tích sâu hơn về đặc điểm và hành vi của khách hàng.

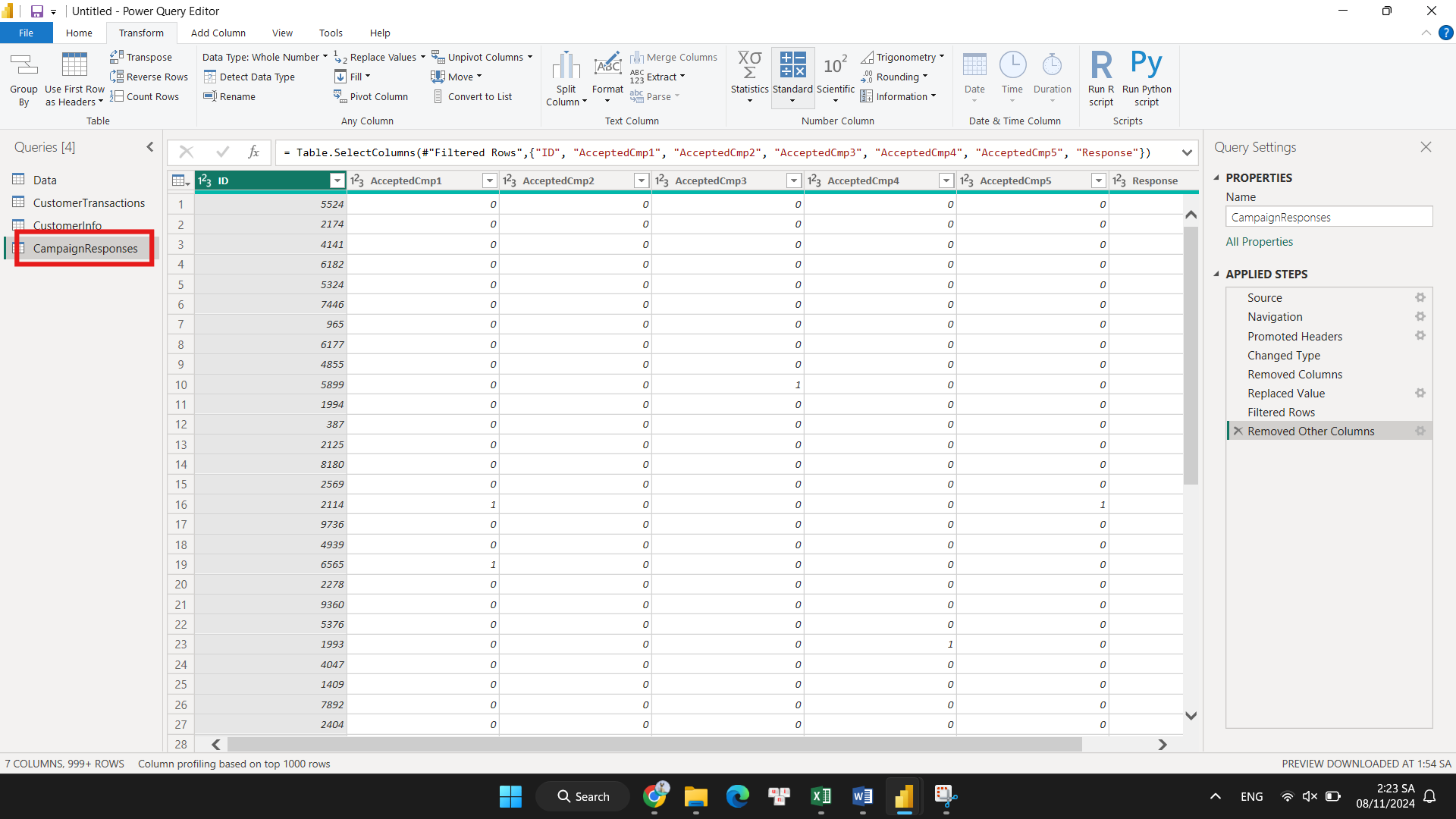
* **Các cột trong bảng**:
* ID (Khóa chính)
* Year\_Birth (Năm sinh của khách hàng)
* Education (Trình độ học vấn của khách hàng)
* Marital\_Status (Tình trạng hôn nhân của khách hàng)
* Income (Thu nhập hộ gia đình hằng năm)
* Kidhome (Số trẻ em trong hộ gia đình)
* Teenhome (Số thanh thiếu niên trong hộ gia đình)
* Dt\_Customer (Ngày khách hàng đăng ký thành viên)



**Bảng CampaignResponses (Phản Hồi Chiến Dịch)**

Bảng này giúp theo dõi các chiến dịch mà khách hàng đã phản hồi.

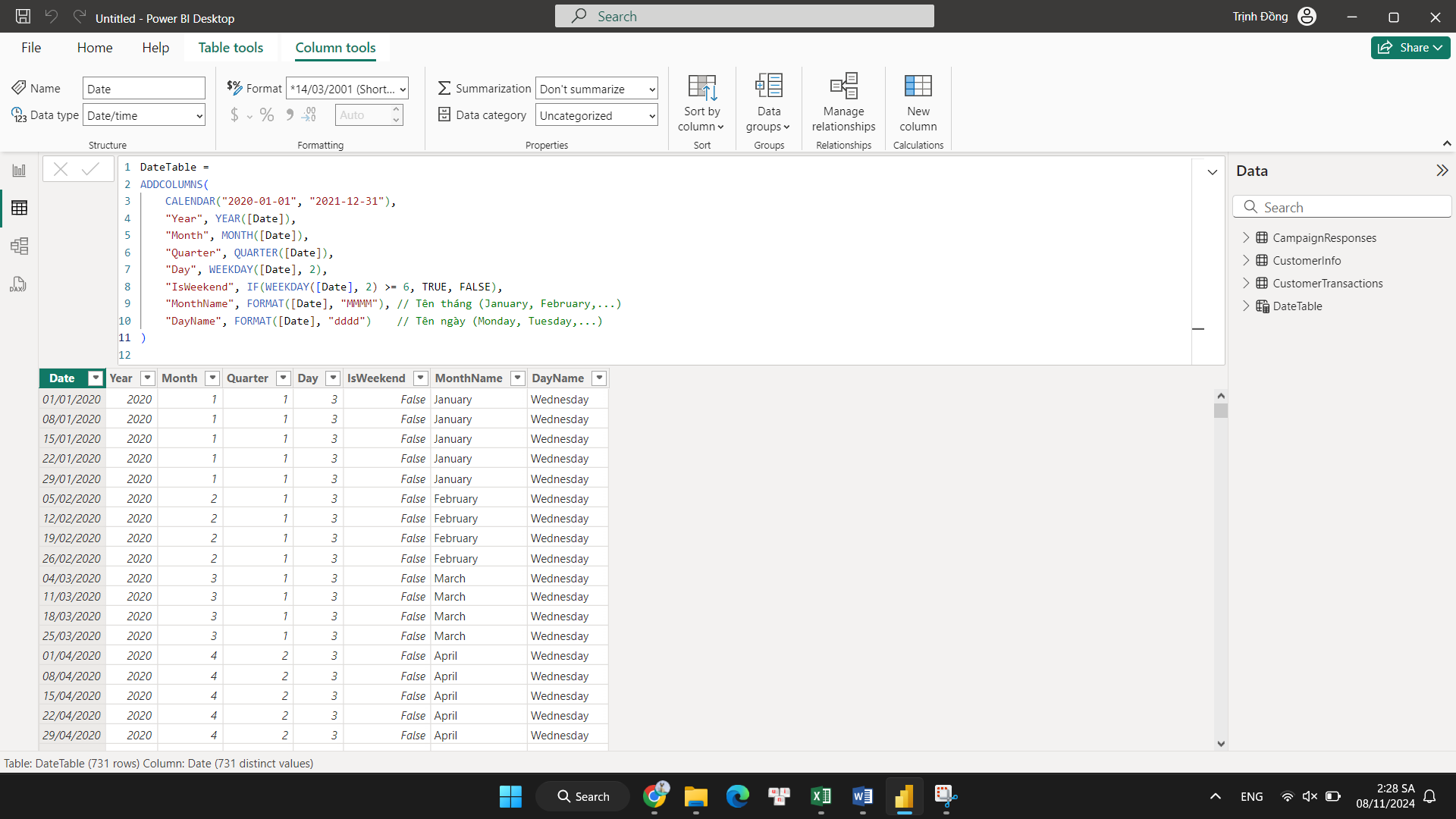
* **Các cột trong bảng**:
* ID (Khóa chính liên kết với bảng CustomerInfo)
* AcceptedCmp1 đến AcceptedCmp5 (Phản hồi của khách hàng với từng chiến dịch)
* Response (Phản hồi của khách hàng với chiến dịch trước)



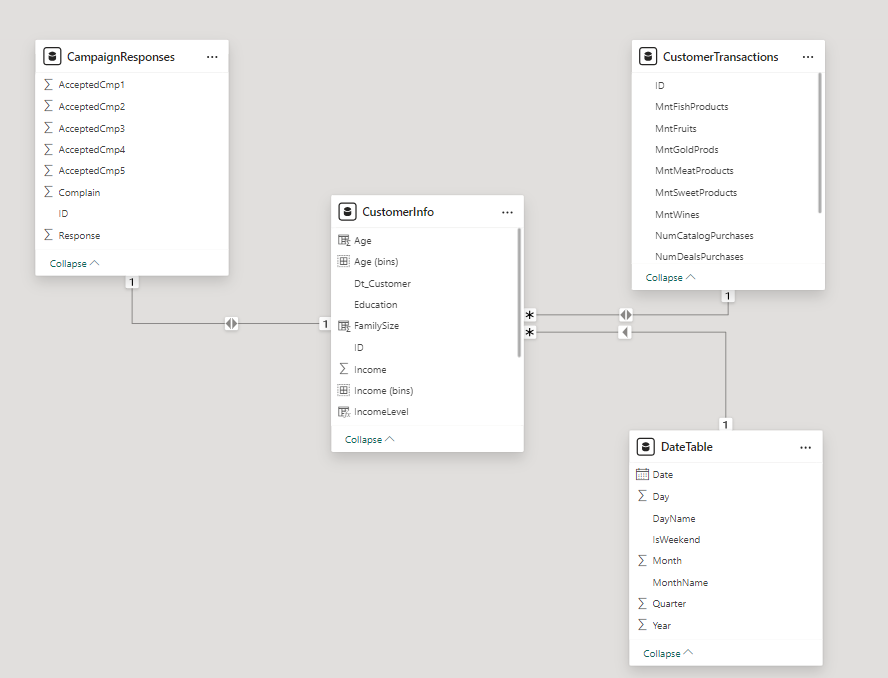
#### Bảng DateTable (Bảng Thời Gian)

Bảng này rất hữu ích để tạo các phân tích theo thời gian. Dùng để nối với cột Dt\_Customer trong bảng CustomerInfo.

* **Các cột trong bảng**:
* Date (Ngày)
* Year (Năm)
* Month (Tháng)
* Quarter (Quý)
* Day (Ngày trong tuần)
* IsWeekend (Xác định có phải ngày cuối tuần không)



**Sơ Đồ Quan Hệ Dữ Liệu (Data Model)**



## Xử lý dữ liệu DAX

### Measure

**Chi Tiêu Trung Bình Theo Danh Mục**:

* Rượu vang:



* Trái cây:



* Thịt:



* Cá:



* Đồ ngọt:



* Đồ trang sức:



**Tổng Mua Hàng Theo Kênh**:

* Tổng số lần mua hàng qua khuyến mãi:



* Tổng số lần mua hàng qua trang web:



* Tổng số lần mua hàng qua danh mục:



* Tổng số lần mua hàng tại cửa hàng:



### Calculated column

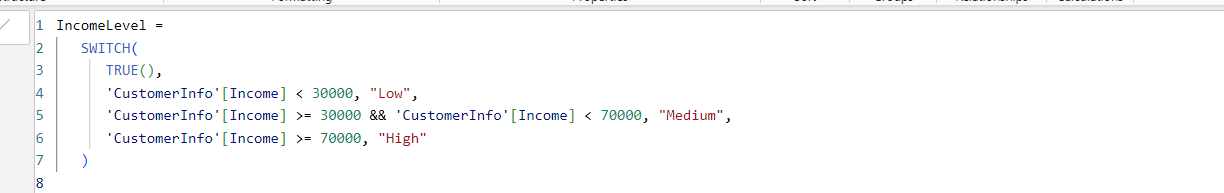
#### Tạo cột “Age” (Tuổi khách hàng)

Tính tuổi khách hàng từ cột Year\_Birth:



#### ****4.3.2.2 Tạo cột “IncomeLevel” (Phân loại Thu nhập)****

Phân loại khách hàng theo mức thu nhập:



#### ****4.3.2.3 Tạo cột “FamilySize” (Quy mô Gia đình)****

Tạo cột FamilySize bằng cách tính tổng số con nhỏ (Kidhome), thiếu niên (Teenhome), và chính khách hàng:



#### ****4.3.2.4 Tạo cột “Tenure” (Số năm làm khách hàng)****

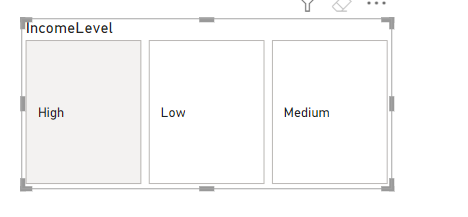
Tính số năm kể từ khi khách hàng đăng ký thành viên (Dt\_Customer):



### Filter

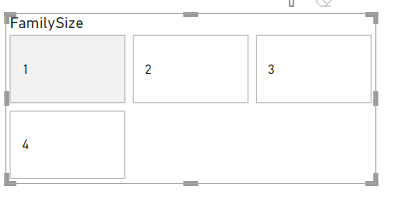
#### Tạo filter chọn ****“IncomeLevel” (Mức thu nhập)****

Sử dụng slicer cho cột IncomeLevel để lọc khách hàng theo các mức thu nhập: Low, Medium, High.



#### ****4.3.3.2 Tạo filter chọn “FamilySize” (Quy mô gia đình)****

Sử dụng slicer cho cột FamilySize để lọc các khách hàng theo quy mô gia đình.



# Trực quan hóa dữ liệu

## Các kỹ thuật trực quan hóa

Các Kĩ thuật trực quan hóa dữ liệu:

**❑ Temporal data visualization (Trực quan hóa dữ liệu chuỗi thời gian)**

❖ Được sử dụng để trình bày các đối tượng một chiều tuyến tính như đồ thị đường, biểu đồ đường hoặc dòng thời gian.

❖ Ví dụ: bạn có thể sử dụng biểu đồ đường để biểu thị các thay đổi xảy ra liên tục trong một khoảng thời gian nhất định.

**❑ Hierarchical data visualization (Trực quan hóa dữ liệu phân cấp)**

❖ Đề cập đến một nhóm hoặc tập hợp các mục có liên kết chung với một mục cha. Bạn có thể sử dụng những cây dữ liệu này để hiển thị các cụm thông tin.

❖ Ví dụ: bạn có thể biểu thị lượng dữ liệu về hàng tồn kho dưới dạng cây, trong đó có nút cha (quần áo) và các nút con (áo sơ mi, quần dài và tất).

**❑ Network data visualization (Trực quan hóa dữ liệu mạng)**

❖ Biểu đồ phân tán trình bày dữ liệu dưới dạng điểm trên một đồ thị.

❖ Biểu đồ bong bóng thêm một yếu tố dữ liệu thứ ba vào biểu đồ phân tán.

❖ Đám mây trình bày tần suất từ bằng cách sử dụng từ có các kích cỡ khác nhau.

**❑ Multidimensional data visualization (Trực quan hóa dữ liệu đa chiều)**

❖ Biểu đồ cột, biểu đồ tròn và đồ thị cột là những ví dụ phổ biến về các loại trực quan hóa này.

❖ Ví dụ: biểu đồ cột so sánh hai hoặc nhiều yếu tố dữ liệu và cho thấy các thay đổi của một biến trong một khoảng thời gian.

❖ Biểu đồ tròn trực quan hóa các phần của tổng thể thuộc từng danh mục.

**❑ Geospatial data visualization (Trực quan hóa dữ liệu không gian địa lý)**

❖ Bản đồ nhiệt, bản đồ mật độ hoặc bản đồ địa hình, trình bày dữ liệu liên quan đến các vị trí ngoài đời thực.

❖ Ví dụ: trực quan hóa dữ liệu biểu thị số lượng khách hàng ghé thăm các chi nhánh bán lẻ khác nhau

**Kĩ thuật được dùng trong dự án này là :**

* Trực quan hóa dữ liệu phân cấp
* Trực quan hóa dữ liệu chuỗi thời gian
* Trực quan hóa dữ liệu đa chiều

Vì  **Trực quan hóa dữ liệu phân cấp:**

* Giúp hiểu rõ hơn về cấu trúc và mối quan hệ giữa các thành phần trong hệ thống.
* Tìm ra các điểm nghẽn hoặc cơ hội cải tiến.

 **Trực quan hóa dữ liệu chuỗi thời gian:**

* Phát hiện xu hướng, số liệu, và các bất thường trong dữ liệu.
* Đánh giá hiệu quả của các chiến lược và biện pháp.

 **Trực quan hóa dữ liệu đa chiều:**

* Khám phá các mối quan hệ phức tạp giữa nhiều biến số.
* Tìm ra các yếu tố ảnh hưởng đến kết quả.
* Xây dựng các mô hình dự đoán.

## Các nguyên tắc trực quan hóa

**5 NGUYÊN TẮC TRỰC QUAN HÓA DỮ LIỆU**

**❑ #1 Chọn đúng loại biểu đồ**

❖ Điều này có vẻ hiển nhiên, nhưng thực tế vẫn nhiều người chọn biểu đồ theo thói quen hoặc cảm tính.

❖ “form follows function” (chức năng đi trước, trình bày đi sau). Phải tự hỏi mục đích của bạn là gì? Bạn muốn sánh các giá trị gì, mối liên hệ giữa các biến nào?

**❑ #2 Không phải tất cả data đều quan trọng**

❖ Không nên ném toàn bộ dataset vào một biểu đồ, hãy phân tích trước để biểu diễn/ highlight những thông tin quan trọng sao cho người xem dễ nắm bắt..

**❑ #3 Biểu đồ thể hiện đúng tương quan số liệu thực tế**

❖ Việc đồ thị hóa dữ liệu nên giúp thể hiện sự tương quan số liệu đúng nhất.

❖ Không nên cố tình lợi dụng các yếu tố về khung tham chiếu, tỷ lệ hình ảnh để gây nhầm lẫn cho người xem.

**❑ #4 Sử dụng màu sắc hợp lý khi chuyển dữ liệu sang dạng biểu đồ**

❖ Việc sử dụng màu sắc mang mục đích dễ nhìn và nhấn mạnh những yếu tố quan trọng trong biểu đồ.

❖ Tuy nhiên, cần chọn lựa màu sắc thông minh, không quá thừa thải, không gây rối và mất tập trung.

**❑ #5 Luôn đảm bảo dữ liệu được trình bày một cách đơn giản và hiệu quả**

❖ Các công cụ trực quan hóa dữ liệu có thể giúp chúng ta tạo ra những biểu đồ đẹp mắt và phức tạp với chỉ vài cái click chuột.

❖ Tuy nhiên, quá nhiều yếu tố về thẩm mỹ có thể làm người xem sao nhãng khỏi thông điệp chính. Hãy lược bỏ những yếu tố rườm rà không cần thiết

Trong dự án này , cho thấy nguyên tắc **chọn biểu đồ** là nguyên tắc quan trọng nhất , vì :

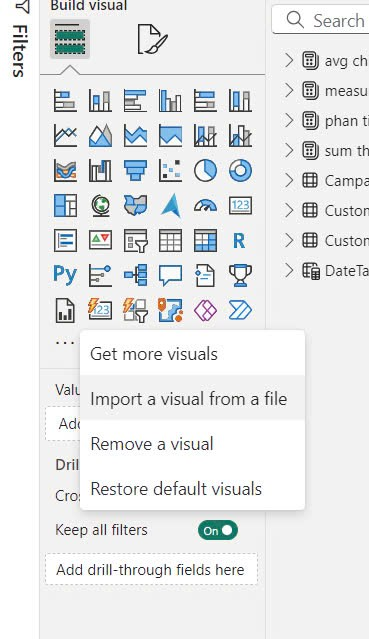
* Loại biểu đồ phù hợp sẽ giúp người xem dễ dàng nhận biết xu hướng, so sánh dữ liệu và rút ra kết luận.
* **Truyền tải thông điệp chính xác:** Mỗi loại biểu đồ có những ưu điểm riêng trong việc thể hiện các loại dữ liệu khác nhau. Chọn sai biểu đồ có thể dẫn đến việc hiểu sai về dữ liệu, gây ra những quyết định sai lầm.
* **Tối ưu hóa trải nghiệm người dùng:** Một biểu đồ được thiết kế tốt sẽ giúp người xem nhanh chóng nắm bắt thông tin, tiết kiệm thời gian và công sức. Ngược lại, một biểu đồ phức tạp hoặc không phù hợp sẽ khiến người xem cảm thấy khó hiểu và mất hứng thú.
* **Tăng tính thuyết phục:** Một biểu đồ trực quan, dễ hiểu sẽ giúp bạn thuyết phục người khác về những ý tưởng và kết luận của mình một cách hiệu quả hơn.

## Trình bày cách thêm visual mới

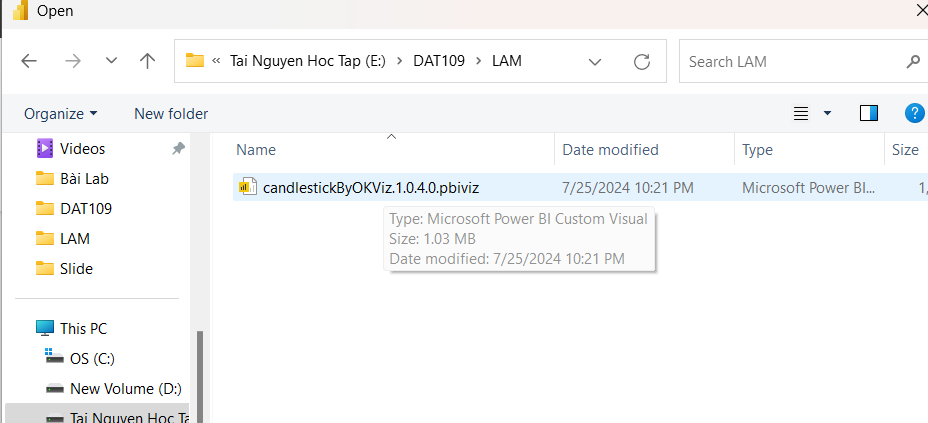
Trong Power BI, nếu visual có sẵn không đáp ứng nhu cầu, bạn có thể tải thêm visual mới từ **Power BI Marketplace**. Dưới đây là cách thêm visual mới:

1. **Mở Power BI Desktop** và vào tab **Insert > More visuals**.
2. **Chọn Get more visuals** để mở Power BI Marketplace.
3. **Tìm kiếm loại visual** mà bạn cần (ví dụ: biểu đồ phân tán, biểu đồ phân phối…).
4. **Chọn visual** mong muốn, sau đó nhấn **Add** để thêm vào Power BI của bạn.
5. Sau khi thêm, visual mới sẽ xuất hiện trong danh sách visual của Power BI, và bạn có thể kéo thả vào báo cáo như các visual khác.

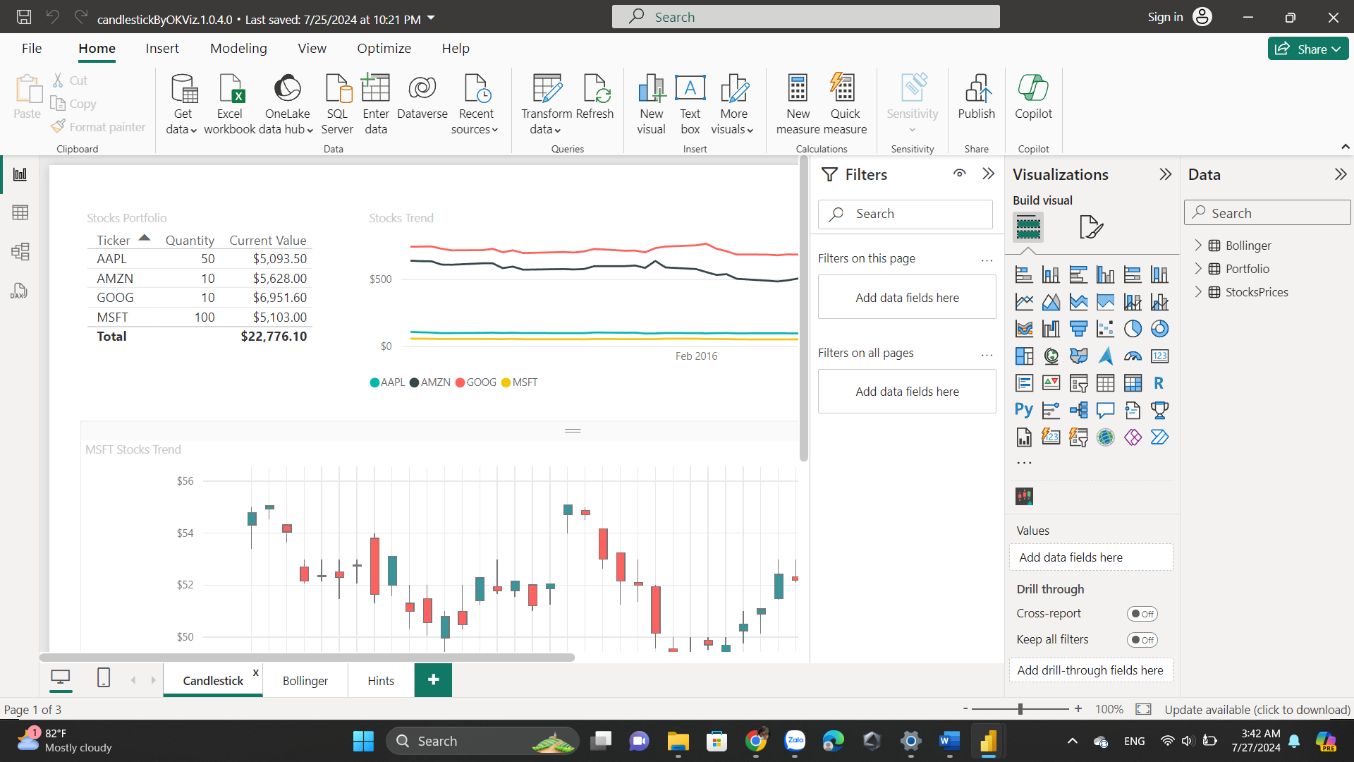
Ở thẻ Visualalzations => chọn vào dấu ba chấm=> Import a visual a file (đã có dữ liệu)



=> chọn tệp chứa visual



**Kết quả:**



## Trình bày tạo các report cho dự án

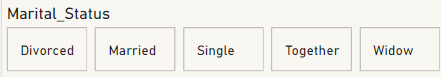
### Tạo visual thống kê chi tiết

#### Tạo visual filter theo năm để dễ theo dõi dữ liệu theo từng năm :



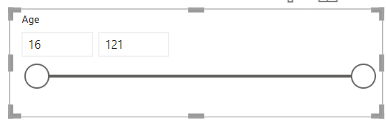
* **Mục đích**: Giúp người dùng thấy được sự khác biệt trong chi tiêu và hành vi theo mức thu nhập theo từng năm .

#### Tạo visual filter theo FamilySize:



* Dùng slicer để lọc các khách hàng theo quy mô gia đình.
* **Mục đích**: Giúp phân tích xu hướng chi tiêu của các hộ gia đình có quy mô khác nhau.

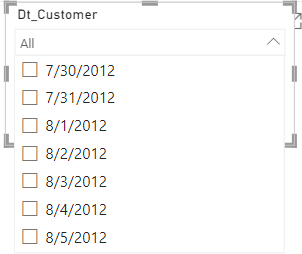
#### Tạo visual filter theo Age:



* Dùng slicer để lọc các khách hàng theo độ tuổi
* **Mục đích**: Thể hiện chi tiêu và hành vi của các độ tuổi khác nhau, giúp tối ưu hóa cho các chiến dịch tiếp thị.

### Tạo visual thống kê tổng thể

#### Tạo visual filter theo ngày giao dịch



* Dùng slicer để lọc theo thời gian, với các tùy chọn như tháng, quý, năm.
* **Mục đích**: Giúp người dùng xem xét xu hướng chi tiêu theo thời gian.

# Xây dựng báo cáo

## Dashboard và report

**5 CÁCH TỐI ƯU HÓA REPORTS VÀ DASHBOARDS**

**❑ Áp dụng các bộ lọc để giới hạn số lượng dữ liệu**

❖Ít dữ liệu hơn sẽ giúp cải thiện tốc độ tải hình ảnh.

❖Khi áp dụng bộ lọc, hãy đặt câu hỏi: Người xem có cần thấy tất cả những dữ liệu này không?

❖Còn cách nào để giảm thiểu số lượng dữ liệu được hiển thị, và hạn chế tối đa tác động xấu đến trải nghiệm của người dùng không?

**❑ Kiểm soát số lượng hình ảnh trên trang báo cáo**

❖Nên giới hạn số lượng hình ảnh trên một trang báo cáo ở mức cần thiết và hợp lý. ❖Điều này sẽ giúp người xem dễ hiểu hơn khi đọc báo cáo, và giúp báo cáo vận hành ổn định hơn.

**❑ Đảm bảo hiệu suất cao của các hình ảnh tùy chỉnh**

❖Lưu ý là hình ảnh trực quan có tác động rất lớn, ảnh hưởng đến toàn bộ báo cáo. ❖Chúng ta nên đánh giá chất lượng và số lượng của các hình ảnh một cách chính xác.

**❑ Hiểu rõ ý nghĩa về hiệu suất của dashboard**

❖Khi chúng ta truy xuất dữ liệu từ bộ nhớ cache, dashboard sẽ hoạt động tốt hơn và nhất quán hơn so với khi chỉ dựa vào nguồn dữ liệu.

**❑ Tối ưu hóa môi trường vận hành**

❖Có thể tối ưu hóa môi trường Power BI bằng cách định cấu hình cài đặt dung lượng, định cỡ cổng dữ liệu và giảm độ trễ mạng

## Xây dựng báo cáo

### Dashboard vs Report

* ***So sánh Dashboard và Report***

|  |  |  |
| --- | --- | --- |
| **Item** | **Dashboard** | **Report** |
| Pages | Chỉ một trang. | Có thể tạo một hoặc nhiều trang. |
| Data sources | Tạo từ nhiều datasets hoặc reports. | Tạo từ một dataset. |
| Visualization | Xây dựng thông tin chi tiết về dữ liệu bằng cách sử dụng biểu đồ, hình ảnh sinh động, lưu đồ. | Xây dựng trang tổng hợp thông tin, cung cấp góc nhìn tổng quan. |
| Available in Power BI Desktop | Không thể tạo. | Có thể tạo và xem. |
| Filters and Slicers | Không thể dung, bị giới hạn trên trang đơn. | Có thể dùng các loại filter, highlight và slice khác nhau. |
| User Interactivity | Cho phép pin visuals từ report và datasets trên canvas đơn, làm cho nó đơn giản để nhóm những gì cần thiết cho người dùng. | Tập trung hơn vào khả năng trực quan hóa, áp dụng các phép biến đổi trên một tập dữ liệu. |
| Favourite | Có thể gán nhìu dashboards | Có thể gán nhìu dashboards |
| Q&A feature | Có | Có thể phân quyền chỉnh sửa report. |
| Alerts | Gửi email cảnh báo khi đáp ứng điều kiện, tiêu chí cụ thể hoặc vượt qua giới hạn. | Không thể |
| Subscribe | Có | Có |
| See underlying dataset tables and fields | Không thể xem dataset nhưng có thể xuất dữ liệu này. | Có thể xem dữ liệu trong tab dữ liệu. |
| Purpose | Được sử dụng để giám sát cấp cao, thường theo thời gian thực hoặc gần thời gian thực, cung cấp cái nhìn tổng hợp về hiệu quả kinh doanh. | Được sử dụng để phân tích chuyên sâu và khám phá dữ liệu nhằm trả lời các câu hỏi kinh doanh phức tạp. |

**Lý do cụ thể giải thích vì sao chúng ta cần làm dashboard và report cho dự án này là :**

**1. Theo dõi tiến độ dự án:**

* **Hiển thị rõ ràng các mục tiêu:** Dashboard và report giúp trực quan hóa các mục tiêu của dự án, so sánh tiến độ thực tế với kế hoạch đã đề ra.
* **Phát hiện rủi ro sớm:** Bằng cách theo dõi các chỉ số quan trọng, chúng ta có thể phát hiện sớm những vấn đề tiềm ẩn và đưa ra các biện pháp khắc phục kịp thời.
* **Đánh giá hiệu quả công việc:** Các báo cáo giúp đánh giá hiệu suất làm việc của từng cá nhân và nhóm, từ đó đưa ra các điều chỉnh cần thiết.

**2. Hỗ trợ ra quyết định:**

* **Cung cấp thông tin đầy đủ:** Dashboard và report cung cấp một nguồn thông tin đáng tin cậy để hỗ trợ quá trình ra quyết định.
* **Phân tích xu hướng:** Bằng cách phân tích dữ liệu lịch sử, chúng ta có thể dự đoán xu hướng tương lai và đưa ra các quyết định chiến lược.
* **So sánh các phương án:** Các báo cáo giúp so sánh hiệu quả của các phương án khác nhau, từ đó lựa chọn phương án tối ưu nhất.

**3. Cải thiện giao tiếp:**

* **Truyền đạt thông tin rõ ràng:** Dashboard và report giúp truyền đạt thông tin một cách trực quan và dễ hiểu, ngay cả đối với những người không có chuyên môn về dữ liệu.
* **Tăng tính minh bạch:** Các báo cáo giúp tăng tính minh bạch trong quá trình làm việc, tạo sự tin tưởng giữa các thành viên trong dự án.
* **Thúc đẩy hợp tác:** Bằng cách chia sẻ thông tin chung, dashboard và report giúp các thành viên trong dự án làm việc hiệu quả hơn.

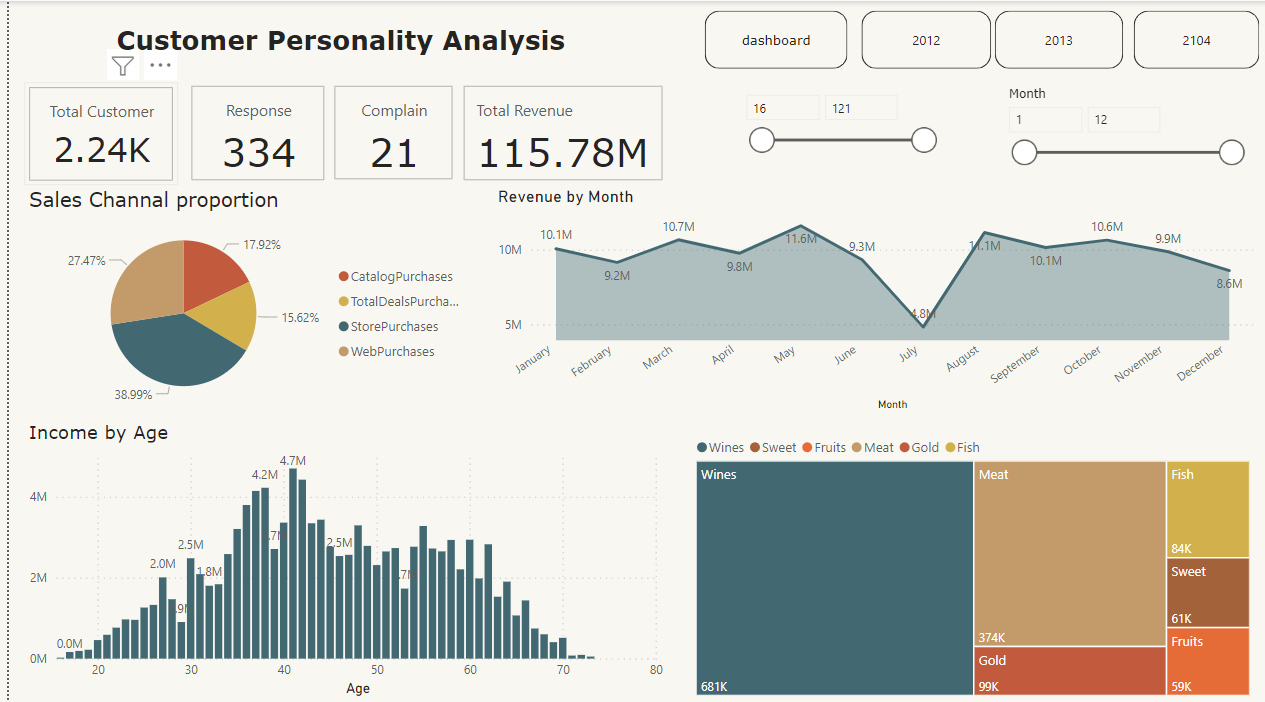
**4. Lưu trữ và báo cáo:**

* **Lưu trữ dữ liệu:** Dashboard và report giúp lưu trữ dữ liệu một cách có hệ thống, dễ dàng tìm kiếm và truy xuất khi cần.
* **Báo cáo định kỳ:** Các báo cáo định kỳ giúp theo dõi tiến độ dự án và cung cấp thông tin cho các cấp quản lý.
* **Hỗ trợ kiểm toán:** Các báo cáo chi tiết giúp kiểm toán viên dễ dàng kiểm tra và đánh giá quá trình thực hiện dự án.

**Tóm lại,** dashboard và report là những công cụ không thể thiếu trong quản lý dự án. Chúng giúp chúng ta theo dõi tiến độ, đưa ra quyết định, cải thiện giao tiếp và lưu trữ dữ liệu.

### Dashboard

* ***Dashboard có trong dự án.***

****

**Dashboard phân tích hành vi khách hàng:** Đây là loại dashboard phổ biến, tập trung vào việc trực quan hóa và phân tích các dữ liệu liên quan đến hành vi mua sắm, sở thích và các đặc điểm nhân khẩu học của khách hàng. Dashboard bạn mô tả có nhiều đặc điểm của loại dashboard này, như:

* **Phân tích theo nhóm khách hàng:** Chia khách hàng theo độ tuổi, giới tính, khu vực... để tìm hiểu hành vi mua sắm của từng nhóm.
* **Phân tích theo sản phẩm:** So sánh doanh thu của các sản phẩm khác nhau, xác định sản phẩm bán chạy và sản phẩm kém hiệu quả.
* **Phân tích theo kênh bán hàng:** Đánh giá hiệu quả của các kênh bán hàng khác nhau.

**Phân tích tổng quan:**

Dashboard này cung cấp một cái nhìn tổng quan về hoạt động kinh doanh trong giai đoạn 2012-2014, bao gồm các thông tin sau:

* **Thông tin chung về khách hàng:** Tổng số khách hàng, số lượng phản hồi, khiếu nại và tổng doanh thu.
* **Kênh bán hàng:** Tỷ lệ đóng góp doanh thu của từng kênh bán hàng (catalog, tổng đại lý, cửa hàng, mua hàng trực tuyến).
* **Doanh thu theo tháng:** Biểu đồ đường thể hiện sự biến động của doanh thu theo từng tháng trong giai đoạn 3 năm.
* **Doanh thu theo sản phẩm:** Biểu đồ TreeMap thể hiện doanh thu của từng loại sản phẩm (rượu vang, đồ ngọt, trái cây, thịt, cá vàng).
* **Doanh thu theo độ tuổi:** Biểu đồ cột thể hiện doanh thu được phân bổ theo các nhóm tuổi khác nhau.

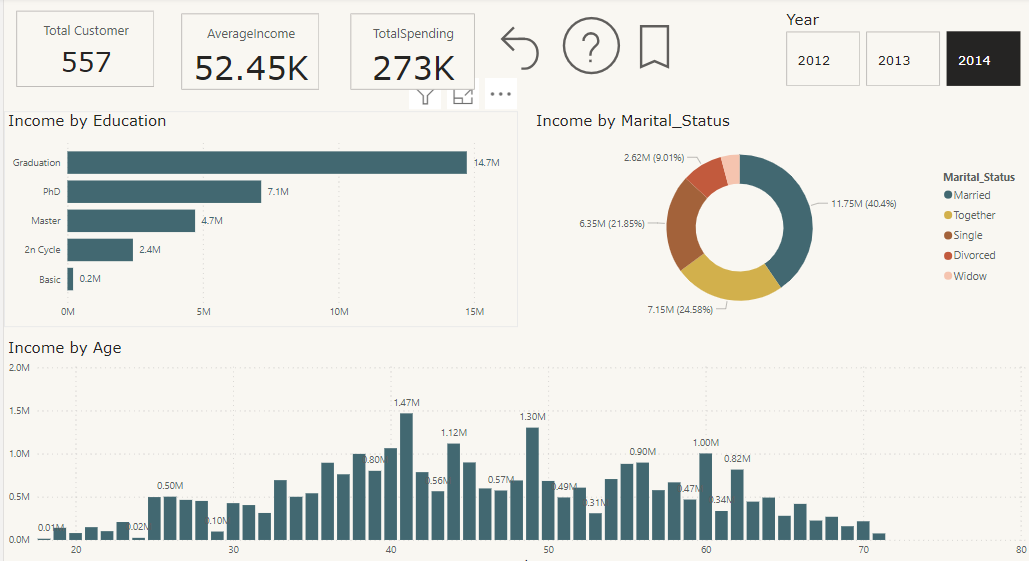
**Các điểm đáng chú ý:**

* **Doanh thu:** Doanh thu có xu hướng tăng trưởng từ năm 2012 đến năm 2013, sau đó có sự giảm nhẹ vào năm 2014.
* **Kênh bán hàng:** Kênh mua hàng trực tuyến đóng góp một tỷ lệ đáng kể vào tổng doanh thu.
* **Sản phẩm:** Rượu vang và đồ ngọt là hai loại sản phẩm mang lại doanh thu cao nhất.
* **Độ tuổi:** Khách hàng trong độ tuổi từ 30 đến 50 là nhóm khách hàng mang lại doanh thu lớn nhất.

Dashboard này cung cấp một cái nhìn tổng quan hữu ích về tình hình kinh doanh. Tuy nhiên, để đưa ra những quyết định kinh doanh hiệu quả, cần phải tiến hành phân tích sâu hơn và kết hợp với các nguồn dữ liệu khác.

### Report

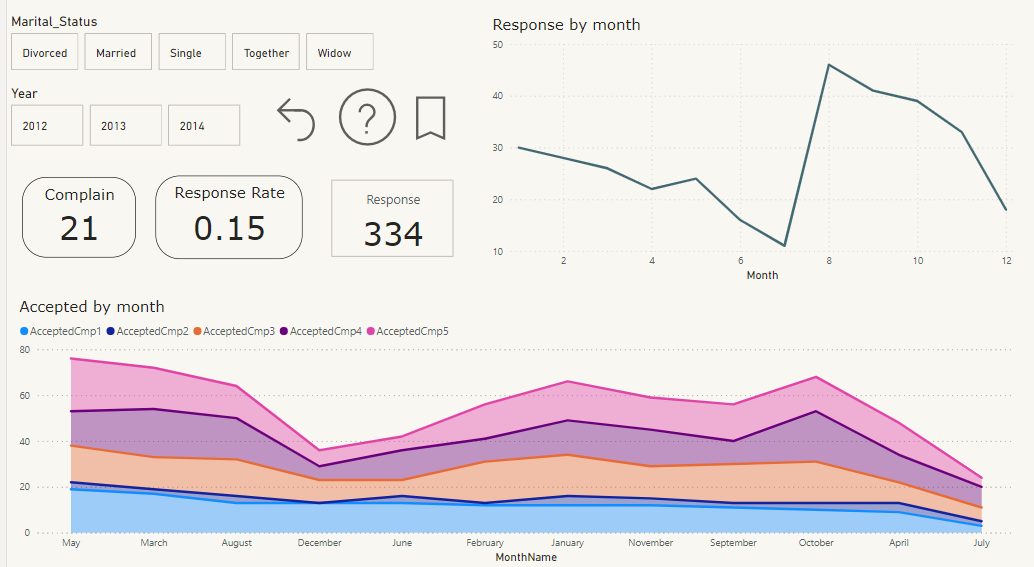
1. **Khách hàng:**



Giải thích:

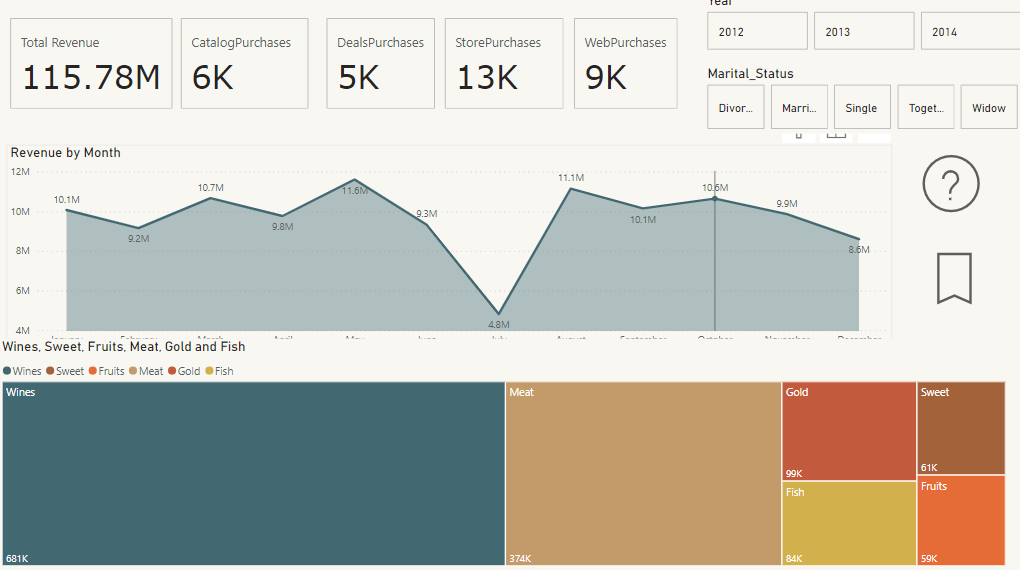
* Số lượng khách hàng tổng thể:
* Total Customer: 2,24K (2.240 khách hàng).
* averageINcome: Thu nhập trung bình của khách hàng.
* Thu nhập theo trình độ học vấn (Income by Education):
* Những người có trình độ Graduation (tốt nghiệp đại học) có tổng thu nhập cao nhất: 59 triệu.
* Những người có trình độ PhD (tiến sĩ) đạt 27 triệu.
* Các trình độ thấp hơn như Master, 2n Cycle, và Basic lần lượt có mức thu nhập giảm dần.
* Thu nhập theo tình trạng hôn nhân (Income by Marital Status):
* Married (đã kết hôn) chiếm 38,41% tổng thu nhập (44,33 triệu).
* Các nhóm khác như Together, Single, Divorced, và Widow có tỷ lệ đóng góp nhỏ hơn.
* Thu nhập theo độ tuổi (Income by Age):
* Thu nhập cao nhất nằm ở nhóm tuổi từ 35-45, với mức cao nhất đạt 4,7 triệu.
* Các nhóm tuổi lớn hơn hoặc nhỏ hơn có mức thu nhập giảm dần.

1. **Phản hồi:**

Giải thích:

* Bộ lọc dữ liệu:
* Marital\_Status (Tình trạng hôn nhân): Có thể lọc theo các trạng thái như Divorced, Married, Single, Together, Widow.
* Year (Năm): Dữ liệu được hiển thị từ năm 2012 đến 2014.
* Thông tin chính:
* Complain: Số lượng khiếu nại là 21.
* Response Rate (Tỷ lệ phản hồi): 0.15 (15%), cho thấy số khách hàng phản hồi là khá thấp so với tổng số.
* Response: Tổng số phản hồi là 334.
* Biểu đồ "Response by Month" (Phản hồi theo tháng):
* Dữ liệu này thể hiện số lượng phản hồi của khách hàng theo từng tháng.
* Có sự gia tăng đáng kể vào khoảng tháng 8 (tháng 7-9 đạt đỉnh), sau đó giảm dần.
* Biểu đồ "Accepted by Month" (Chấp nhận theo tháng):
* Biểu đồ khu vực này hiển thị số lượng khách hàng chấp nhận các chiến dịch (Campaigns) theo từng tháng, phân chia theo các chiến dịch cụ thể:
* AcceptedCmp1 đến AcceptedCmp5: Thể hiện các chiến dịch từ 1 đến 5.
* Các chiến dịch có xu hướng cao hơn trong các tháng như tháng 5, tháng 10, nhưng giảm mạnh vào các tháng mùa đông và đầu năm.
* Nhận xét:
* Tỷ lệ phản hồi thấp: Với Response Rate chỉ 15%, cần đánh giá lại cách tiếp cận khách hàng trong chiến dịch.
* Thời gian cao điểm: Các tháng như tháng 8 và tháng 10 có thể là cơ hội tốt để triển khai thêm chiến dịch mới.
* Chiến dịch hiệu quả: Cần phân tích chiến dịch nào (AcceptedCmp1 - 5) mang lại hiệu quả cao nhất và tập trung vào đó.

1. **Chi tiêu:**



Giải thích:

Tổng quan doanh thu và kênh mua sắm:

* Tổng doanh thu (Total Revenue): 115,78 triệu. Đây là tổng số tiền thu được từ khách hàng trong giai đoạn từ 2012 đến 2014.
* Các kênh mua sắm:
* Store Purchases (Mua tại cửa hàng): 13.000 giao dịch, là kênh phổ biến nhất.
* Web Purchases (Mua qua web): 9.000 giao dịch, đứng thứ hai.
* Catalog Purchases (Mua qua danh mục): 6.000 giao dịch.
* Deals Purchases (Mua qua khuyến mãi): 5.000 giao dịch.

Doanh thu theo tháng (Revenue by Month):

* Các tháng doanh thu cao nhất:
* Tháng 5: 11,6 triệu (cao nhất trong năm).
* Tháng 10: 10,6 triệu.
* Tháng 1: 10,1 triệu.
* Tháng có doanh thu thấp nhất:
* Tháng 7: 4,8 triệu (giảm mạnh so với các tháng khác).
* Nhận xét:
* Doanh thu ổn định ở mức cao trong nửa đầu năm, ngoại trừ sự sụt giảm vào tháng 7.
* Từ tháng 8 trở đi, doanh thu tăng trở lại và duy trì ở mức khá cao.

Doanh thu theo sản phẩm (Wines, Sweet, Fruits, Meat, Gold, and Fish):

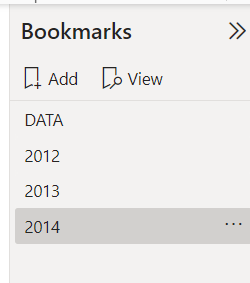
* Rượu vang (Wines):
* Mang lại doanh thu cao nhất: 681K, chiếm phần lớn tổng doanh thu từ các danh mục.
* Thịt (Meat):
* Xếp thứ hai với doanh thu 374K.
* Các sản phẩm khác:
* Vàng (Gold): 99K.
* Cá (Fish): 84K.
* Đồ ngọt (Sweet): 61K.
* Trái cây (Fruits): 59K.

Nhận xét:

* Rượu vang chiếm ưu thế lớn: Đây là sản phẩm đóng góp cao nhất vào doanh thu. Các chiến lược quảng bá và phân phối nên tập trung mạnh vào sản phẩm này.
* Doanh thu giảm rõ rệt vào tháng 7. Có thể do yếu tố mùa vụ hoặc nhu cầu giảm, cần phân tích thêm nguyên nhân.
* Ngoài rượu vang và thịt, các sản phẩm khác như vàng và cá cũng có tiềm năng. Việc tăng cường khuyến mãi hoặc đa dạng hóa nguồn cung cấp có thể thúc đẩy doanh số.

### Bookmarks

Tạo 3 Bookmarks qua từng năm để dễ nhìn ra dữ liệu để phân tích



# KẾT LUẬN

## Báo cáo

### Các bước viết báo cáo

**Các bước viết báo cáo phân tích dữ liệu**

* **Consider Your Audience (Xem xét đối tượng)**

❖Ai sẽ đọc báo cáo này? Họ hiểu chủ đề như thế nào?

- Doanh nghiệp và các nhà đầu tư.

❖Ngôn ngữ và từ vựng bạn đang sử dụng.

- Tiếng việt, có kèm một số thuật ngữ tiếng Anh

❖Chữ viết tắt và mức độ kỹ thuật.

- Mức độ hiểu cơ bản trong các cửa hàng mua sắm

❖Bạn sẽ đi để giải thích điều gì? Ở mức độ nào?

- Tôi đang muốn giải thích các giao dịch luôn mới theo ngày, việc khách hàng mua sắm theo các loại mặt hàng khác nhau

- Mục tiêu cụ thể là tách và phân nhóm dữ liệu thành các đối tượng cụ thể.

❖Loại hình ảnh nào sẽ dung trong báo cáo?

- Sau đó thể hiện sự thay đổi trên từng đối tượng này thông qua hình ảnh là các dạng biểu đồ

* **Draft Out the Sections (Soạn thảo nội dung)**

❖Giới thiệu. - Giới thiệu tổng quan về mua sắm của khách hàng

❖Nội dung - Dữ liệu, Phương pháp, Phân tích, Kết quả.

❖Kết luận - Trình bày các báo cáo tổng hợp dữ liệu rõ ràng và chính xác.

* **Develop The Body (Phát triển nội dung)**

❖ Giới thiệu về Dữ liệu

- Hãy cụ thể và giải thích cách bạn thu thập dữ liệu, mẫu của bạn là gì, công cụ và tài nguyên nào bạn đã sử dụng cũng như cách bạn tổ chức dữ liệu của mình. Điều này sẽ giúp người đọc hiểu sâu hơn về mẫu dữ liệu của bạn và làm cho báo cáo của bạn chắc chắn hơn.. => Thu thập dữ liệu thủ công từ file Excel.

❖ Phương pháp - Cần giải thích những phương pháp bạn đã sử dụng để phân tích dữ liệu. Điều này đơn giản có nghĩa là bạn cần giải thích tại sao và cách bạn chọn các phương pháp cụ thể. Bạn cũng cần giải thích lý do tại sao những phương pháp này phù hợp nhất với mục tiêu đã đặt ra và kết quả bạn đang cố gắng đạt được. => Phương pháp mô hình hóa dữ liệu thành các đối tượng. => Sử dụng các phép biến đổi để thêm các cột dữ liệu tính toán còn thiếu.

* **Develop The Body (Phát triển nội dung)**

❖ Phân tích - Sau khi bạn giải thích dữ liệu và phương pháp bạn đã sử dụng, phần tiếp theo này sẽ kết hợp hai phương pháp đó lại với nhau. Thêm mô tả và giải thích từng bước. Cố gắng làm cho nó càng đơn giản càng tốt để ngay cả những độc giả thiếu kinh nghiệm nhất cũng có thể hiểu. => Mục tiêu cụ thể là tách và phân nhóm dữ liệu thành các đối tượng cụ thể.

❖ Kết quả - Là phần quan trọng nhất trong báo cáo. => Thể hiện sự thay đổi trên từng đối tượng này thông qua hình ảnh là các dạng biểu đồ. => Trình bày các báo cáo tổng hợp dữ liệu rõ ràng và chính xác về khách hàng đã mua các sản phẩm.

* **Write the Introduction (Viết phần giới thiệu)**

❖ Có vẻ lạ khi viết phần giới thiệu ở cuối nhưng đó lại là cách thông minh nhất. Phần này giải thích ngắn gọn những gì báo cáo sẽ đề cập. Đó là lý do tại sao bạn nên viết nó sau khi viết xong phần phát triển Nội dung.

❖ Câu hỏi bạn đã nêu ra và trả lời bằng phân tích - Làm thế nào truyền đạt thông điệp dữ liệu rõ ràng, hấp dẫn giúp người nghe hiểu được vấn đề nhanh chính xác, từ bộ dữ liệu phi cấu trúc.

❖ Bối cảnh của phân tích và thông tin cơ bản - Nhiều doanh nghiệp hoặc cá nhân ứng dụng phân tích dữ liệu vào quá trình quản lý, kinh doanh, đầu tư - Dữ liệu sau khi xử lý, kết quả là hàng loạt con số khô khan.

❖ Dàn ý ngắn của báo cáo

* **Add a Short Conclusion (Thêm một kết luận ngắn)**

❖ Ngắn gọn và sử dụng ngôn ngữ đơn giản để cung cấp những phát hiện quan trọng nhất.

❖ Nên bao gồm một đoạn văn về ý nghĩa và tầm quan trọng của những phát hiện đó đối với khách hàng, doanh nghiệp hoặc công ty đã thuê bạn. =>Trong phạm vi dự án này, chỉ cung cấp kiến thức, khái niệm cơ bản về hành vi mua sắm của khách hàng => Hiểu được một số đối tượng và mặt hàng cơ bản => Biết cách nhìn biểu đồ trực quan.

* **Include Data Visualization Elements (Thêm các đối tượng trực quan hóa dữ liệu)**

❖ Hình ảnh là cách hiệu quả để thu hút người xem => Khi người xem không phải là người có chuyên môn sâu về dữ liệu hoặc lĩnh vực đang đề cập.

❖ Trình bày đơn giản, ngắn gọn, làm nổi bật dữ liệu quan trọng, đáng được lưu ý.

❖ Sử dụng các dạng biểu đồ, bộ lọc hiệu quả về hành vi mua sắm của khách hàng => Card => Line chart => Table => Matrix => Treemap => Slicer => Stacked column chart => Donut chart => Clustered bar chart => Bie chart.

* **Proofread & Edit Before Submission (Chỉnh sửa trước khi gửi)**

❖Để viết một báo cáo phân tích dữ liệu tốt có thể trở nên lãng phí nếu không đọc lại và chỉnh sửa.

❖Việc này sẽ giúp loại bỏ những sai sót tiềm ẩn và còn có cái nhìn khách quan khác về báo cáo của bạn.

- Đọc toàn bộ báo cáo một cách khách quan, giống như bạn đang nhìn thấy nó lần đầu tiên.

- Để lại một tâm trí cởi mở cho những thay đổi.

- Thêm hoặc bớt thông tin.

- Sắp xếp lại các phần.

- Tìm những từ tốt hơn để nói điều gì đó.

- Tìm và loại bỏ các lỗi ngữ pháp và chính tả

### Tổng hợp

**Bối cảnh và vấn đề cần giải quyết:**  
Một chuỗi siêu thị lớn tại Mỹ muốn hiểu sâu hơn về hành vi khách hàng từ dữ liệu mua sắm (2012-2014). Mục tiêu là tối ưu hóa chiến lược tiếp thị, cải thiện trải nghiệm khách hàng và tăng doanh thu.

**Mục tiêu chính của phân tích:**

* Xây dựng chân dung khách hàng lý tưởng dựa trên các đặc điểm như độ tuổi, thu nhập, tình trạng hôn nhân, và quy mô gia đình.
* Hiểu rõ hành vi khách hàng để đáp ứng nhu cầu cung ứng.
* Phân tích từ dữ liệu đưa ra chiến lược marketing hiệu quả dể tăng doanh thu cho công ty
* Hiểu rõ xu hướng chi tiêu, xác định danh mục sản phẩm được ưa chuộng nhất.
* Đánh giá hiệu quả các kênh bán hàng (trực tuyến, tại cửa hàng, hoặc qua catalog).
* Phân tích tỷ lệ phản hồi và mức độ hài lòng qua các chiến dịch tiếp thị.

**Phân tích dữ liệu khách hàng:**

* **Phân khúc khách hàng:** Sử dụng thông tin nhân khẩu học (tuổi, trình độ học vấn, thu nhập, số con...) để phân nhóm khách hàng.
* **Hành vi chi tiêu:**
* Danh mục sản phẩm có chi tiêu cao nhất là rượuvang và thịt.
* Khách hàng trong độ tuổi từ 30-40 chiếm phần lớn số lượng giao dịch.
* **Hiệu quả kênh bán hàng:** Cửa hàng trực tiếp mang lại doanh thu cao nhất 38.96%, trong khi giao dịch qua web và catalog cũng đáng chú ý.
* **Phản hồi khách hàng:** 334 phản hồi tích cực cho các chiến dịch tiếp thị, nhưng có 21 khiếu nại, cho thấy cần cải thiện ở một số khía cạnh.

**Các phương pháp và công cụ được sử dụng:**

* **Làm sạch dữ liệu:** Xử lý các giá trị ngoại lệ, dữ liệu bị thiếu bằng Excel và Power Query.
* **Mô hình hóa dữ liệu:** Xây dựng các bảng Fact và Dimension, kết nối qua Power BI.
* **Trực quan hóa:** Sử dụng biểu đồ cột, tròn, treemap, slicer trong Power BI để trình bày thông tin.
* **Tính toán DAX:** Dùng DAX để tạo các chỉ số như tuổi khách hàng, quy mô gia đình, phân loại thu nhập, và tính toán chi tiêu trung bình.

**Kết quả phân tích chính:**

* **Chân dung khách hàng:**
  + Khách hàng lý tưởng: 30-40 tuổi, thu nhập trung bình đến cao, gia đình nhỏ (1-2 con).
  + Hành vi: Thích mua rượu vang và thịt, giao dịch thường xuyên qua cửa hàng trực tiếp.
* **Đánh giá hiệu quả tiếp thị:** Các chiến dịch trước đó có tỷ lệ phản hồi khá tốt, cần cải thiện ở các kênh ít phổ biến như catalog.
* **Đề xuất:** Tăng cường tiếp cận nhóm khách hàng trẻ qua kênh trực tuyến và tập trung vào các sản phẩm chủ lực như rượu vang.

## Khó khăn

**Ngành công nghiệm thực phẩm là một lĩnh vực đầy biến động và cạnh tranh cao. Luôn đối mặt với nhiều thách thức và cơ hội khác nhau, dữ liệu cho thấy:**

1. **Biến động nguyên liệu đầu vào**: Giá cả và nguồn cung nguyên liệu nông sản thay đổi thất thường do yếu tố thời tiết, thiên tai, hoặc biến động kinh tế toàn cầu. Điều này gây ảnh hưởng lớn đến chi phí sản xuất và khả năng duy trì mức giá ổn định cho sản phẩm.
2. **Sự cạnh tranh cao**: Nhiều doanh nghiệp cùng tham gia thị trường, cả trong nước lẫn quốc tế, với các sản phẩm tương tự. Điều này buộc các doanh nghiệp phải không ngừng cải tiến sản phẩm, dịch vụ để giữ chân khách hàng.
3. **Yêu cầu về chất lượng và an toàn thực phẩm**: Ngày càng có nhiều quy định nghiêm ngặt hơn về tiêu chuẩn an toàn thực phẩm từ các cơ quan chức năng. Doanh nghiệp phải đầu tư nhiều hơn vào công nghệ, quản lý và kiểm tra chất lượng để đáp ứng các yêu cầu này.
4. **Xu hướng tiêu dùng thay đổi**: Người tiêu dùng ngày càng quan tâm đến các sản phẩm lành mạnh, hữu cơ và thân thiện với môi trường. Điều này đòi hỏi các doanh nghiệp phải nghiên cứu và phát triển các sản phẩm phù hợp, đôi khi với chi phí cao.
5. **Hội nhập kinh tế quốc tế**: Việc gia nhập các hiệp định thương mại quốc tế mang lại cơ hội nhưng cũng đi kèm với những thách thức lớn. Các doanh nghiệp phải cạnh tranh với các sản phẩm nhập khẩu, đồng thời đảm bảo đáp ứng các tiêu chuẩn quốc tế khi xuất khẩu.
6. **Đầu tư công nghệ và đổi mới sáng tạo**: Để nâng cao năng suất và chất lượng sản phẩm, các doanh nghiệp cần đầu tư vào công nghệ mới. Tuy nhiên, chi phí cao và rủi ro từ việc ứng dụng công nghệ là những rào cản lớn.
7. **Nguồn nhân lực**: Việc thiếu hụt nguồn nhân lực chất lượng cao trong ngành công nghiệp thực phẩm là một vấn đề cần được quan tâm. Đào tạo và giữ chân nhân viên có năng lực là bài toán khó đối với nhiều doanh nghiệp.

Các thách thức trên yêu cầu các doanh nghiệp ngành công nghiệp thực phẩm không ngừng đổi mới, nâng cao năng lực cạnh tranh, và tận dụng hiệu quả các cơ hội để phát triển bền vững.

**Bên cạnh đó, còn có các vấn đề như:**

**Kỹ thuật:** Thiếu kinh nghiệm trong sử dụng Power BI và DAX, dẫn đến mất thời gian trong việc xử lý và trực quan hóa dữ liệu.

**Dữ liệu:** Dữ liệu ban đầu có nhiều trường không cần thiết và thiếu tính nhất quán, đòi hỏi nhiều công đoạn làm sạch và chuyển đổi.

**Quản lý thời gian:** Phân bổ thời gian giữa các giai đoạn của dự án chưa tối ưu, gây áp lực trong việc hoàn thành báo cáo đúng hạn.

## Thuận lợi

Mặc dù đối mặt với nhiều khó khăn, ngành công nghiệp thực phẩm cũng có những thuận lợi lớn giúp các doanh nghiệp trong ngành phát triển:

1. **Nhu cầu tiêu thụ cao**: Thực phẩm là một nhu cầu thiết yếu và không thể thiếu trong cuộc sống hàng ngày. Với sự gia tăng dân số và mức sống ngày càng cao, nhu cầu tiêu thụ thực phẩm, đặc biệt là các sản phẩm chế biến và chất lượng cao, luôn tăng trưởng mạnh.
2. **Tiềm năng thị trường nội địa**: Các doanh nghiệp thực phẩm có thể tận dụng cơ hội để phát triển sản phẩm phù hợp với khẩu vị và thói quen tiêu dùng của người dân.
3. **Hỗ trợ từ chính phủ**: Ngành thực phẩm luôn được chính phủ quan tâm và hỗ trợ thông qua các chính sách phát triển nông nghiệp, công nghiệp chế biến, xúc tiến thương mại và các hiệp định thương mại tự do .
4. **Xu hướng xuất khẩu tích cực:** Các sản phẩm như cà phê, hạt điều, trái cây sấy, và đồ uống chế biến đang có vị thế cao trên thị trường quốc tế.
5. **Cơ hội từ hội nhập quốc tế**: Việc tham gia các hiệp định thương mại tự do (CPTPP, EVFTA, RCEP) mang lại cơ hội tiếp cận thị trường nước ngoài với thuế quan ưu đãi, mở rộng thị trường cho doanh nghiệp thực phẩm.
6. **Phát triển công nghệ và xu hướng tiêu dùng mới**: Công nghệ chế biến thực phẩm ngày càng tiên tiến, giúp các doanh nghiệp tạo ra sản phẩm đa dạng, đáp ứng xu hướng tiêu dùng như thực phẩm tiện lợi, hữu cơ, và chức năng.
7. **Phát triển hạ tầng logistics**: Với sự cải thiện hệ thống giao thông, kho bãi, và các dịch vụ logistics, các doanh nghiệp thực phẩm dễ dàng vận chuyển và phân phối sản phẩm đến các thị trường trong nước và quốc tế.
8. **Thói quen tiêu dùng hiện đại**: Xu hướng tiêu dùng ngày càng ưu tiên thực phẩm chế biến sẵn, tiện lợi, và an toàn, mở ra cơ hội lớn cho các doanh nghiệp đầu tư vào sản phẩm phù hợp.

Những thuận lợi trên là động lực lớn giúp ngành công nghiệp thực phẩm phát triển bền vững, đặc biệt trong bối cảnh hội nhập và toàn cầu hóa.

Về mặt dữ liệu :

**Dữ liệu rõ ràng:** Bộ dữ liệu đầy đủ thông tin về khách hàng, hỗ trợ tốt trong phân tích hành vi và chiến lược tiếp thị.

**Công cụ hiệu quả:** Sử dụng Power BI và các kỹ thuật trực quan hóa mạnh mẽ giúp trình bày dữ liệu dễ hiểu và chuyên nghiệp.

**Hợp tác nhóm:** Các thành viên trong nhóm phối hợp hiệu quả, chia sẻ công việc và hỗ trợ lẫn nhau trong quá trình thực hiện.

## Hướng phát triển

Để công ty ngày càng phát triển, chúng ta cần tập trung vào 2 hướng sau:

1. **Định hình sản phẩm cốt lõi và đa dạng hóa danh mục sản phẩm**

* Rượu vang (Wines): Chiếm doanh thu lớn nhất (681K), cho thấy đây là sản phẩm cốt lõi.
* Hướng phát triển:
* Đẩy mạnh chiến lược marketing:
* Xây dựng các chương trình quảng bá mạnh vào các mùa cao điểm (tháng 5, tháng 10).
* Kết hợp rượu vang với các sản phẩm liên quan như thịt (Meat) trong các gói quà tặng.
* Đầu tư vào phân khúc cao cấp:
* Phát triển dòng sản phẩm rượu vang cao cấp để tăng giá trị đơn hàng.
* Hợp tác với nhà hàng, khách sạn để đẩy mạnh tiêu thụ.
* Tăng cường bán hàng online:
* Triển khai website chuyên biệt hoặc ứng dụng bán hàng cho rượu vang.
* Tạo các ưu đãi độc quyền cho khách hàng đặt qua kênh trực tuyến.
* Thịt (Meat) và Cá (Fish): Đây là hai sản phẩm có tiềm năng tăng trưởng nhưng chưa được khai thác mạnh.
* Hướng phát triển:
* Thịt (Meat):
* Tăng cường các gói combo với rượu vang.
* Marketing nhấn mạnh vào chất lượng và nguồn gốc (organic, sạch, cao cấp).
* Cá (Fish):
* Tập trung vào khách hàng quan tâm đến sức khỏe và dinh dưỡng.
* Xây dựng dòng sản phẩm đặc sản (như cá hồi, cá nhập khẩu).

**2. Phân tích theo mùa vụ và điều chỉnh chiến lược kinh doanh**

* Doanh thu theo tháng:
* Tháng 7 là giai đoạn doanh thu thấp nhất (4,8M).
* Tháng 5 và tháng 10 là giai đoạn doanh thu cao nhất.
* Hướng phát triển:
* Tháng 7:
* Tổ chức các chương trình khuyến mãi, giảm giá để kích cầu.
* Kết hợp với lễ hội hoặc sự kiện để thu hút khách hàng (như ngày hội mua sắm giữa năm).
* Tháng cao điểm (tháng 5 và tháng 10):
* Gia tăng sản lượng hàng hóa, đảm bảo nguồn cung không bị gián đoạn.
* Tập trung quảng bá các gói quà tặng theo mùa lễ hội hoặc ngày đặc biệt.

**3. Tăng cường hiệu quả của các kênh bán hàng**

* Phân phối doanh số:
* Kênh cửa hàng truyền thống (Store Purchases) chiếm ưu thế (13K giao dịch).
* Kênh mua qua web (Web Purchases) có tiềm năng tăng trưởng nhưng chỉ đạt 9K giao dịch.
* Hướng phát triển:
* Mua tại cửa hàng:
* Cải thiện trải nghiệm khách hàng tại điểm bán, như cung cấp dịch vụ thử sản phẩm, tăng tính tương tác.
* Đưa ra chương trình thẻ thành viên với ưu đãi cho khách hàng trung thành.
* Mua qua web:
* Xây dựng nền tảng website dễ sử dụng, tối ưu hóa thanh toán và giao hàng.
* Kích thích mua sắm online qua khuyến mãi đặc biệt (miễn phí vận chuyển, giảm giá).
* Hợp tác với các sàn thương mại điện tử để mở rộng phạm vi khách hàng.

**4. Tập trung vào phân khúc khách hàng trọng tâm**

* Theo tình trạng hôn nhân (Marital\_Status):
* Nhóm khách hàng Married (Đã kết hôn) chiếm 38,41% doanh thu.
* Hướng phát triển:
* Phát triển các sản phẩm phù hợp cho gia đình, như:
* Gói sản phẩm ăn uống dành cho gia đình (kết hợp rượu, thịt, cá).
* Khuyến mãi đặc biệt vào các dịp lễ gia đình (ngày lễ Tạ ơn, Giáng sinh).
* Đẩy mạnh marketing cá nhân hóa, nhấn mạnh đến sự tiện lợi và giá trị khi mua gói combo.
* Theo độ tuổi:
* Khách hàng trong độ tuổi từ 30-50 chiếm doanh thu lớn nhất.
* Hướng phát triển:
* Phát triển sản phẩm nhắm đến đối tượng trung niên:
* Sản phẩm cao cấp hơn để phục vụ nhu cầu của nhóm khách hàng có thu nhập cao.
* Tăng cường sản phẩm chăm sóc sức khỏe (như cá, thực phẩm chức năng).
* Marketing qua kênh yêu thích của nhóm này như mạng xã hội (Facebook, YouTube).

**5. Nâng cao tỷ lệ phản hồi và giữ chân khách hàng**

* Tỷ lệ phản hồi (Response Rate) hiện chỉ đạt 15%.
* Hướng phát triển:
* Tăng tính cá nhân hóa:
* Sử dụng dữ liệu khách hàng để tạo ra các chương trình quảng cáo phù hợp với từng cá nhân.
* Gửi email hoặc thông báo ưu đãi đặc biệt dựa trên hành vi mua hàng trước đây.
* Triển khai khảo sát khách hàng:
* Hỏi khách hàng về trải nghiệm mua sắm và cải thiện dịch vụ dựa trên phản hồi của họ.
* Chương trình khách hàng thân thiết:
* Tặng điểm thưởng, phiếu giảm giá cho khách hàng mua nhiều.
* Tổ chức các sự kiện tri ân khách hàng (online hoặc offline).