The dataset

Orders	Slaes orders information_ bao gồm thông tin tài chính quan trọng liên quan đến từng đơn hàng, chẳng hạn như số tiền bán hàng và giá vốn hàng bán - còn được gọi là giá vốn hàng bán. Biến này cho biết chi phí trực tiếp liên quan đến việc sản xuất quần áo mà công ty bán.
Returns	Orders that have been returned to the company _ chi tiết bất kỳ đơn đặt hàng nào được trả lại sau khi mua do lỗi hoặc sai sót
Products	Details about products sold_ các mặt hàng quần áo do CT sản xuất
Retailers	Information related to the customers_ chứa tất cả dữ liệu liên quan đến các nhà bán lẻ mua mặt hàng đó - về cơ bản là khách hàng của công ty.

➔ Đối tượng ban đầu là Phòng Kinh doanh, những thành viên đã hiểu rõ về các số liệu liên quan đến hiệu quả hoạt động của công ty và các sản phẩm được bán. Tuy nhiên, phòng Kinh doanh muốn điều tra chi tiết hơn về loại đơn đặt hàng mà công ty đang xử lý và những sản phẩm nào đang được đặt hàng.

1.Six steps to EDA

There are six tasks of :

Understanding data structure.

Indentifying missing data.

Describing data with descriptive statistics & distributions.

Indentifying outliers, examining and quantifying relationships between variables.

1.1 Understading data structure

Trước khi đi sâu vào bất kỳ tập dữ liệu nào, điều quan trọng là phải biết cấu trúc dữ liệu - số hàng, cột và kiểu dữ liệu. Có hai loại biến cơ bản.

- Các biến liên tục thường là số, có tập giá trị vô hạn.

Ví dụ như số lượng đơn hàng, tỷ lệ nhấp và khoảng cách giữa các thành phố.

- Các biến phân loại hoặc phi số có thể có hai hoặc nhiều nhóm.

Một số ví dụ ở đây có thể là loại nhà, quốc gia và công ty. Biết cấu trúc này giúp ảnh hưởng đến các bước tiếp theo của EDA.

Bộ Dataset có 4 data con, đi sâu hơn vào các data

Tập 1 : Return.

Có 549 hàng và 8 cột lần lượt là các cột.

Order_ID, Return_year, Return_Month , Return_day, Hour , Minute, Second , Return_Date

Order_ID	Chứa các giá trị số ID của retailer
Return_Year	Chứa các giá trị năm trả lại hàng
Return_Month	Chứa các giá trị tháng trả lại hàng
Return_day	Chứa các giá trị ngày trả lại hàng
Return_Date	Chứa đầy đủ ngày tháng năm trả lại hàng, thuộc kiểu dữ liệu chuỗi

Dataset 2: Orders.

Gồm 8392 dòng và 13 dòng gồm các cột tương ứng với:

Order_ID, Order_Date, Order_YearMonth, Retailer_ID, Product_SKU Product_Price , Product_Cost , Order_Quantity , Sales_Amount ,Cost_of_Goods_Sold , Product_Discount, Profit, Profit_Margin.

Oder_ID	Số lượng nhận đơn
Order_Date	Ngày đặt hàng
Order_YearMonth	Năm và Tháng đặt hàng
Retailer_ID	Nhà bán lẻ
Product_SKU	Mã nhận dạng sản phẩm
Product_Price	Giá của sản phẩm
Product_Cost	Số thành sản phẩm
Order_Quantity	Số lượng sản phẩm được đặt hàng
Sales_Amount	Tổng số tiền bán hàng
Cost_of_Goods_Sold	Giá vốn bán hàng
Product_Discoun	Mức giảm gái sản phẩm ở định dạng %
Profit	Số tiền lợi nhuận
Profit_Margin	Tỷ suất lợi nhuận

Dataset 3: Product.

Có 1211 hàng và 7 hàng tương ứng với:

Product_SKU, Product_Full_Description, Product_Gender, Product_Category, Product_Name, Product_Size, Product_Color

Product_SKU	Đơn vị lưu kho(SKU) cho mỗi sản phẩm, là mã nhận dạng duy nhất cho mỗi biến thể sản phẩm
Product_Full_Description	Mô tả đầy đủ về sản phẩm
Product_Gender	Danh mục giới tính mà sản phẩm được nhắm mục tiêu

Product_Category	Phân loại sản phẩm thành 1 danh mục cục thể
Product_Name	Chứa tên sản phẩm
Product_Size	Kích thước của sản phẩm
Product_color	Thể hiện màu của sản phẩm

Dataset 4: Retailer.

Chứa thông tin các nhà bán lẻ.

Retailer_ID	Nhận dạng duy nhất cho các nhà bán lẻ
Retailer_Channel	Kênh phân phối
Retailer_Name	Tên các nhà bán lẻ sử dụng để nhận dạng
City	Địa điểm nơi đặt trụ sở của nhà bán lẻ
Region	Khu vực đặt trụ sở của nhà bán lẻ
Area	Khu vực cục thể của region
Country	Quốc gia
Distance from Warehouse	Khoảng cách từ vị triscuar nhà bán lẻ đến nhà kho hoặc trung tâm phân phối

Trước khi đi sâu vào bất kỳ tập dữ liệu nào, điều quan trọng là phải biết cấu trúc dữ liệu - số hàng, cột và kiểu dữ liệu. Phân tích kiểu dữ liệu trong data sau

1.2: Identifying missing data.

1.2.1: Khi bị missing data sẽ ảnh hưởng như thế nào?

1.2.2: Các xử lí.

Thiếu dữ liệu là trường hợp thường xảy ra khi thực hiện Phân tích dữ liệu khám phá trên tập dữ liệu mới. Thay vì xóa các bản ghi bị thiếu, đôi khi việc sử dụng kỹ thuật quy nạp để điền vào các giá trị với ước tính về giá trị của chúng sẽ hữu ích hơn. Điều này đặc biệt đúng khi số lượng giá trị bị thiếu là nhỏ

Xác định bằng column quality để xác định missing data.

Dựa vào column quality để xác định missing data, chừng nào không có missing data.

Và có dùng thêm datavizalition

1.3: Describing data with descriptive statistics & distributions

1.3.1: Statistics

Một phần quan trọng của quy trình Phân tích dữ liệu khám phá (EDA) chỉ đơn giản là mô tả dữ liệu. Bắt đầu là chọn một số biến quan tâm và sử dụng số liệu thống kê mô tả cơ bản - tối thiểu, trung bình, trung bình, tối đa, độ lệch chuẩn - để tìm hiểu thêm về phạm vi giá trị của các biến đó (ví dụ: phân phối của chúng).

Power Query sẽ không đầy đủ bằng khi khái quái ra biểu đồ.

Thì dựa vào các chỉ số em có thể lọc ra được các sản phẩm đang được bán chạy và ưu chuộng ở hiệu tại và từ đó thông qua điều chỉnh giá của sản phẩm.

Từ đó đưa ra các chiến lược marketing phù hợp cho từng chuyên mục riêng.

1.3.2. Distributions

Distributions của một biến đề cập đến tập hợp tất cả các giá trị có thể có của biến và tần số liên quan. Nói cách khác, đó là số lần mỗi giá trị của biến xuất hiện trong các quan sát.

Được sử dụng cho cả 2 loại dữ liệu

Biểu đồ được sử dụng để hình dung một phân phối. Chúng trông giống như biểu đồ thanh và hiển thị các giá trị thay đổi trên trục x và tần số hoặc số lượng quan sát với các giá trị đó trên trục y.

Thông qua biểu đồ distrubutions ta có thể biết được về xu hướng trung tâm – trung bình- cũng như sự lệch của biểu đồ (cân đối, lệch phải, lệch trái). Mức độ phân bố rộng như thế nào cho thấy sự phân tán của dữ liệu xung quanh các điểm trung tâm này. Nếu phân phối có mức chênh lệch rộng thì sẽ có độ biến thiên lớn hơn so với mức trung bình, tức là độ lệch chuẩn lớn. Nếu nó hẹp thì độ biến thiên và độ lệch chuẩn sẽ nhỏ hơn

Phần trăm là giá trị mà tại đó phần trăm quan sát rơi vào bằng hoặc thấp hơn. Trung vị là phân vị thứ 50, tức là 50% số quan sát ở dưới giá trị trung bình, 50% ở trên. Thường xem xét sự phân bổ theo các phần tư, cụ thể là các phần tư thứ 1, 2 và 3 hoặc các phân vị thứ 25, 50 và 75.

1.4: Indentifying outliers, examining and quantifying relationships between variables.

1.4.1: Indentifying outliers.

Các ngoại lệ là các điểm dữ liệu nằm ngoài mẫu tổng thể trong một phân phối.

Có hai phương pháp phổ biến để tìm kiếm các giá trị ngoại lệ trong tập dữ liệu một cách định lượng.

Using standard devistion

Lower = -3*SD

Upper = 3*SD

Outlier when: value < lower OR upper < value.

Interquartile Range(IQR)

Lower = 25 percentile – (1.5* IQR)

Upper = 25 percentile – (1.5* IQR)

Outlier when: value < lower OR upper < value.

Có nhiều con đường để giải quyết các yếu tố ngoại lại nhưng hai cách remove observations và imputation giống khi giải quyết missing data. Bên cạnh đó Winsorzing là phương pháp khá hiệu quả, tương tự như áp đặt, cập nhật các điểm ngoại lệ bằng một số khác.

Nếu value < 5^{th} percentile thì value = 5^{th} percentile

Nếu 95th percentile > value thì value =95^{th} percentile

1.4.2: EDA with categorical variables and continuous variables

Categorical variables:

Dựa vào mối quan hệ giữa các biến ta có thể rút ra nhiều insight để đưa ra các quyết đoán cho chiến lược trong tương lai. Cũng có thể xác định xu hướng về một biến liên tục bằng cách khám phá sự khác biệt của thống kê mô tả giữa các giá trị của biến phân loại.

Biểu đồ hình hộp rất tuyệt vời trong việc so sánh các phân bố trên một biến phân loại. Đặt mỗi cạnh nhau, bạn có thể rút ra kết luận nhanh chóng về sự khác biệt.

Ví dụ: Nam và Nữ thị phần nào chiếm nhiều hơn khi mua quần áo, và tại địa điểm nào thì người nữ sẽ bỏ tiền ra nhiều hơn. Từ đó đưa ra ý kiến cho chiến lược quảng cáo đúng cách.

Continuous variables:

Nếu Categorical variables dùng biểu đồ hình hộp để khái quát hóa mối quan hệ giữa các biến thì đối với Continuous variables ta sẽ dùng biểu đồ phân tán.