Sales & Customer Analysis

Project Phase - 1

Group Name: Data Vizards
Members:
Muskan Garg
Vatsal Gohel

1. Description of the dataset: Provide an overview of the data, including its content and limitations?

Description of the Dataset:

This dataset contains information about orders placed by customers on an e-commerce website. It includes details about the orders such as the order ID, order date, order status, item ID, quantity ordered, price, value, discount amount, total, customer ID, year, reference number, age, zip code, and discount percentage. The dataset has a total of 286,392 entries with 36 columns. There are several categorical and numerical variables such as quantity ordered, price, value, discount amount, total are numerical data, while the order date, and order status and so on are categorical data.

order_id	Unique identifier for each sales order
order_date	Date when the order was placed
status	Current status of the order (received, completed, order_refunded, etc.)
item_id	Unique identifier for each item in the order
sku	SKU (Stock keeping unit) code for each item
qty_ordered	Quantity of the item ordered
price	Price of one unit of the item
value	Total value of the item ordered
discount_amount	Amount of discount applied to the item
total	Total amount paid for the item after discount
category	Category of the item
payment_method	Method used for payment (COD, Payaxis, Easypay, etc.)
bi_st	Billing Status represents the status or type of the sales transaction, indicating whether it is valid, net or gross
cust_id	Unique identifier for each customer
year	Year in which the order was placed
month	Month in which the order was placed
ref_num	Reference number for the order
Name Prefix	Prefix of the customer's name (Drs., Prof., etc.)
First Name	First Name of the customer
Middle Initial	Middle Initial of the customer's name
Last Name	Last Name of the customer
Gender	Gender of the customer
age	Age of the customer
full_name	Full Name of the customer
E Mail	E-mail of the customer
Customer Since	Data when the customer started their relationship with the company
SSN	Social Security number of the customer
Phone No.	Contact Detail of the customer
Place Name	Location of the customer
County	County where the customer resides
City	City where the customer resides
State	State where the customer resides
Zip	ZIP code of the customer's Location

Region	Region where the customer resides
User Name	User name associated with the customer
Discount_Percent	Percentage of discount applied to the order

Limitations:

- The dataset provided is a sample, so it might not cover the entire data range. A major limitation of our dataset is that, although a year is a good time for observing and understanding market trends and patterns. The year 2020-21 was the peak time of the global fight against Corona virus, and everything was different as a result. There was a section of people shopping for necessities, whereas there was another section who were just out shopping for leisure, fun moments within the boundaries of their houses. Therefore, we might get skewed insights and patterns.
- Dataset does not include any information on marketing campaigns, customer behavior, or external factors that may affect sales.
- The importance of data privacy should not be underestimated when dealing with some columns, such as SSN (Social Security Number), that contain sensitive personal information. Therefore, when working with such columns, appropriate steps should be taken to ensure confidentiality.
- Datasets do not contain contextual information about their businesses. It is important to understand industry dynamics, market conditions, or specific business objectives to interpret findings accurately.

2. Reason for selecting the dataset: Explain the rationale behind choosing this particular dataset?

Reason for Selecting the Dataset:

Our choice of this dataset was based on our emphasis on understanding the market patterns on a bi-directional basis. To understand consumer behavior, how purchasing power works? Today's business and customers are both interested in the Market Basket concept. The market basket (a basket of things) is an assortment of goods or resources set up to track a market segment's performance. There are a lot of baskets of things that people like to buy together when they go out shopping, such as butter and bread, bread and milk.

Few things which we can do using this dataset efficiently are:

- We can evaluate pricing strategies after analyzing sales. Inventory management, product pricing, marketing campaigns, and customer service can all be improved by understanding product consumption and sales.
- A sales dataset provides information on order status, payment methods and billing status, which can be used to measure key performance indicators, identify areas for improvement, and access sales performance.
- Demographic information, email addresses, and customer since dates are included in this dataset. Using this data, marketers can segment customers, analyze preferences, identify loyal customers, and customize marketing strategies.

- Dataset contains information about total order value, discounts, and payment methods. Among other things, it can be used to analyze revenue trends, evaluate discounts, compare payment methods, and calculate financial metrics.

3. Ambiguity in the data: Identify any unclear or ambiguous aspects of the data that may affect the analysis?

Ambiguity in the Data:

In the Sales dataset, there are few ambiguous or unclear aspects. Here are some examples:

- **Status Terminology:** The status column contains values like "received", "complete", "order_refunded", and "canceled". The specific definitions and criteria for this status are not stated explicitly. What factors trigger an order to be considered "complete" or "canceled", and what circumstances lead to an order being refunded, is not clearly defined. Analysis would be improved by clarifying these status values.
- **Bi_st Column:** In the "bi_st" column, we can find values like "Valid", "Net", and "Gross". While we previously named it as billing status in the data dictionary through potential interpretations, however, without further information or a clear definition provided, the exact meaning of these values remains uncertain. Understanding the intended definition or context of "bi_st" would be necessary for accurate analysis.
- **Data Privacy:** Dataset contains personal information like email addresses, Social Security Numbers (SSNs), and contact details. In order to ensure compliance with data protection regulations and maintain the privacy of individuals, appropriate measures must be taken to protect the handling of sensitive personal data.

4. Intended audience: Specify the target audience for the analysis, who will be presented with the results.

Intended Audience:

The groups we are looking to target are distinct groups with distinct goals such as Sales and Marketing people to enhance customer targeting, Operations and Inventory management teams, Business Executives and Decision-Makers, Finance and Accounting Departments, Online Retailers, Market Researchers, Sentiment Analysts, Marketing Professionals and Business Owners.

5. Objectives of the analysis: Outline the questions you hope to answer through the final analysis. At this stage, the report should highlight your research questions, rather than providing answers.

The objectives of the analysis of the sales dataset are to:

1. How much is the average discount percentage applied to orders?
2. Which of the following categories or products have the highest sales?
3. Is there an impact of discounts on the overall revenue of a company?
4. What is the relationship between the age of the customer and their purchase behavior?
5. How is the distribution of statuses of different orders determined (received, complete, refunded, canceled)?

Some of the specific research questions that will guide the analysis include:

A. What are the overall sales trends over time? Are there any specific items or stock keeping units (SKUs) that are generating the highest level of sales volume or revenue within your company? Finding out which categories of products are most popular than others in order to prepare the inventory of products so that it will be able to satisfy the needs of the customers when the time comes.
B. Understanding the characteristics of the company's highest-value customers as well as the average length of time the customer has been associated with the organization in order to figure out the loyal customer segments and strategize pricing.
C. What is the preferred method of payment for customers? Are there any differences in the preferences of payment methods between individuals based on the demographics such as gender or age?
D. Does the performance of sales in each region differ from the performance in other regions? Are there particular cities, counties, states, or regions that play a significant role in contributing to sales?
E. Analyzing the trends of which products are more likely to be sold or recommended in more attractive ways, so that they can be stacked or recommended in a way that makes them more readily accessible to consumers.

Reference:

→ **Dataset:** https://www.kaggle.com/datasets/nhiyen/sales-data-fy-2020-2021