|  |  |
| --- | --- |
| Staples  Business Intelligence Report | Group 1 Section 2  Ayush Khandelwal  Hardeep Sahu  Varun Ashar  Ajinkya Deshmukh  Sai Teja Adimulam |

Table of Contents

[Chapter 1: Overview of Client and Problem Statement 2](#_Toc165656075)

[1.1 Client Background 2](#_Toc165656076)

[1.2 Problem Statement 2](#_Toc165656077)

[1.3 Justification for Choosing the Problem 2](#_Toc165656078)

[Chapter 2: Dataset Description. 3](#_Toc165656079)

[Chapter 3: Data Warehouse Design and Implementation. 6](#_Toc165656080)

[3.1 4-step dimension modeling process 6](#_Toc165656081)

[3.2 Star Schema 6](#_Toc165656082)

[3.3 Implementation 7](#_Toc165656083)

[3.4 Control Flow Diagram 8](#_Toc165656084)

[3.5 Data Flow Diagram (DFD) 8](#_Toc165656085)

[3.5.1 Customer DFD 8](#_Toc165656086)

[3.5.2 Order DFD 9](#_Toc165656087)

[3.5.3 Derived Column 9](#_Toc165656088)

[3.5.4 FACT DFD 9](#_Toc165656089)

[Chapter 4: Data Preparation 10](#_Toc165656090)

[Chapter 5: Data Exploration 15](#_Toc165656091)

[5.1 Exploratory Data Analysis (EDA) 15](#_Toc165656092)

[Descriptive Statistics Results 15](#_Toc165656093)

[5.2 Data Transformation 17](#_Toc165656094)

[Chapter 6: Data Analysis and Results 19](#_Toc165656095)

[Chapter 7: Business Implications 25](#_Toc165656096)

# Chapter 1: Overview of Client and Problem Statement

## 1.1 Client Background

Our client, Staples, was established in 1986 by Thomas G. Stemberg and Leo Kahn. It was one of the pioneers in the office supply superstore market, rapidly expanding across the United States and internationally. Staples offers a diverse range of products, including office supplies, technology products, and furniture. This broad product range caters to a wide variety of consumer needs, from small businesses to large corporations.

## 1.2 Problem Statement

In the pursuit of enhancing sales strategies, Staples faces a critical challenge centered on analyzing consumer behavior. Understanding and optimizing these behavioral patterns is paramount for maximizing profitability and maintaining a competitive edge in the market. By delving deep into consumer actions, preferences, and purchasing habits, Staples aims to develop targeted approaches that resonate with its customer base, ultimately driving increased sales and sustained business growth.

## 1.3 Justification for Choosing the Problem

This problem was chosen due to its direct impact on Staples’ strategic goals of increasing sales and market share. By leveraging detailed data analysis, we can provide actionable insights into how different factors affect sales outcomes. These insights will allow Staples to tailor their marketing and operational strategies to better meet consumer demands and improve financial performance. The analysis will focus on sales across different months and states, deviations from sales targets, and the effects of shipping options on delivery efficiency.

# Chapter 2: Dataset Description.

Our analysis leverages multiple datasets from Staples to understand consumer behavior, sales distribution, and the effectiveness of promotions. Below is the data dictionary for each dataset utilized in our study.

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Type** | **Description** |
| **Category** | VARCHAR(50) | Represents the classification of products or services |
| **City** | VARCHAR(50) | The city in which the customer or transaction is located |
| **Color** | VARCHAR(50) | The color specification of the product |
| **Cost** | VARCHAR(50) | Cost price of the product or service |
| **Country** | VARCHAR(50) | The country where the transaction takes place or where the customer resides. |
| **Customer ID** | VARCHAR(50) | Unique identifier for each customer |
| **Customer Name** | VARCHAR(50) | Name of the customer |
| **CustomerSK** | VARCHAR(50) | Surrogate key for the customer dimension in a data warehouse |
| **Date** | DATE | Specific date when an event occurs |
| **DateSK** | INT | Surrogate key for the date dimension in a data warehouse |
| **DateString** | VARCHAR(50) | Date in string format |
| **Day** | INT | Day of the month (1-31) |
| **DayofWeek** | INT | Numeric representation of the day of the week (1-7) |
| **DayofWeekName** | VARCHAR(10) | The name of the day of the week |
| **DayofYear** | INT | The day number of the year |
| **Discount** | NUMERIC(18,3) | The discount amount applied to a sale |
| **IsLeapYear** | BIT | A Boolean indicating whether the year is a leap year |
| **IsWeekend** | BIT | A Boolean indicating whether the day is part of the weekend |
| **Month** | INT | Numeric representation of the month |
| **MonthName** | VARCHAR(10) | The name of the month |
| **On Time** | INT | Indicates whether the delivery or shipment was on time |
| **Order Date** | DATETIME | The date on which the order was placed |
| **Order ID** | VARCHAR(50) | A unique identifier for each order |
| **OrderSK** | VARCHAR(50) | A surrogate key for the order |
| **Payment Type** | VARCHAR(50) | The method of payment used |
| **Payment\_modeSK** | VARCHAR(50) | A surrogate key for the payment mode |
| **Postal Code** | VARCHAR(50) | The postal code where the order is delivered |
| **Product ID** | VARCHAR(50) | A unique identifier for each product |
| **ProductSK** | VARCHAR(50) | A surrogate key for the product |
| **Profit** | NUMERIC(18,3) | The profit made from the sale of a product |
| **Promotion Description** | VARCHAR(50) | Description of any promotions applied to the order |
| **Promotion ID** | VARCHAR(50) | A unique identifier for each promotion |
| **PromotionSK** | VARCHAR(50) | A surrogate key for the promotion |
| **Quantity** | NUMERIC(18,3) | The number of units of the product sold |
| **Quarter** | INT | The quarter of the year in which the transaction occurred |
| **Region** | VARCHAR(50) | The region in which the order was placed or delivered |
| **Retail Cost** | NUMERIC(18,3) | The cost price of the product |
| **Sales** | NUMERIC(18,3) | The total sales amount for the order |
| **Segment** | VARCHAR(50) | The market segment to which the customer belongs |
| **Ship Date** | DATETIMED | The date on which the order was shipped |
| **Ship Mode** | VARCHAR(50) | The mode of shipment used |
| **State** | VARCHAR(50) | The state where the order is delivered |
| **Sub-Category** | VARCHAR(50) | A further classification of the product category |
| **Turnaround Time** | INT | The time taken from order placement to shipment |
| **Unit Price** | NUMERIC(18,3) | The price per unit of the product |
| **Week** | INT | The week number of the year |
| **Year** | INT | The year in which the transaction occurred |

# Chapter 3: Data Warehouse Design and Implementation.

## 3.1 4-step dimension modeling process

A screenshot of a computer

Description automatically generated

## 3.2 Star Schema



## 3.3 Implementation

We used python to make connection with our database “MIS587\_project\_grp1” and import out dataset which is flat file (CSV) to our OLTP Database “MIS587\_dataset\_preprocessed”.

**Snapshot of Our Dataset:**

A screenshot of a computer

Description automatically generated

**Snapshot of Python code to connect and data import:**

A screenshot of a computer

Description automatically generated

**Snapshot of imported data in OLTP form:**

A screenshot of a computer

Description automatically generated

## 3.4 Control Flow Diagram

A screenshot of a computer

Description automatically generated

We used SSIS for generating a star schema.

## 3.5 Data Flow Diagram (DFD)

### 3.5.1 Customer DFD

A screenshot of a computer

Description automatically generated

### 3.5.2 Order DFD

A screenshot of a computer

Description automatically generated

### 3.5.3 Derived Column

#### On Time

(Ship\_Mode== "First Class" && (DATEDIFF("day",(DT\_DBTIMESTAMP)Order\_Date,(DT\_DBTIMESTAMP)Ship\_Date) <= 2)) ? 1 : (Ship\_Mode == "Second Class" && (DATEDIFF("day",(DT\_DBTIMESTAMP)Order\_Date,(DT\_DBTIMESTAMP)Ship\_Date) <= 3)) ? 1 : (Ship\_Mode == "Standard Class" && (DATEDIFF("day",(DT\_DBTIMESTAMP)Order\_Date,(DT\_DBTIMESTAMP)Ship\_Date) <= 5)) ? 1 : (Ship\_Mode == "Same Day" && DATEDIFF("day",(DT\_DBTIMESTAMP)Order\_Date,(DT\_DBTIMESTAMP)Ship\_Date) <= 1) ? 1 : 0

#### Turnaround Time

DATEDIFF("DAY",(DT\_DBTIMESTAMP)Order\_Date,(DT\_DBTIMESTAMP)Ship\_Date)

### 3.5.4 FACT DFD

A screenshot of a computer

Description automatically generated

# Chapter 4: Data Preparation

The dataset was successfully loaded into a panda DataFrame, which contains 10,194 entries and 20 columns, detailing order information, customer data, product categories, sales, quantities, discounts, and profits.

A screenshot of a computer

Description automatically generated

We removed a few uncertainties in the data for clarity and uniformity.

A screenshot of a computer

Description automatically generated

Duplicate Removal: We checked for duplicate rows in the data, identifying and removing a total of 5 duplicate entries.

A screen shot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

Missing Values: We identified missing values in columns 'Segment', 'Sales', and 'Quantity', with 50 missing entries each.

A screen shot of a computer

Description automatically generated

Data Visualization: Several visualizations were created, including histogram for Quantity sold based on Sub-Category, and a scatter plot comparing sales and profit, providing visual insights into the data distributions and relationships between key variables. The visualizations which are generated for the Data preprocessing can be seen below.

A histogram for ‘Quantity’ sold based on ‘Sub-Category’ distribution showed that the Sub-Category ‘Binders’ sold the most quantity among other categories.  
  
A black background with white text

Description automatically generated  
  
A graph with different colored bars

Description automatically generated

A scatter plot between 'Sales' and 'Profit' revealed the relationship between these two variables, highlighting the variance in profit for given sales figures and potential outliers or high-leverage points that could influence further analysis.

A black screen with white text

Description automatically generated

A graph showing a graph of sales

Description automatically generated with medium confidence

Chapter 5: Data Exploration

## 5.1 Exploratory Data Analysis (EDA)

Descriptive Statistics: We calculated descriptive statistics, providing a summary of the central tendency, dispersion, and shape of the dataset's distribution. Descriptive statistics for 'Sales' and 'Profit' were included, offering insights into average sales and profit margins, as well as understanding the variability and distribution of these key metrics.



### Descriptive Statistics Results

The average (mean) 'Sales' were approximately $228, but the median (50th percentile) was much lower at $53.91, indicating the presence of high-value sales transactions that skew the average.

The 'Profit' had a mean of approximately $28.67 and a median of $8.69, suggesting that most transactions yielded a modest profit, with a few transactions resulting in significantly higher profits.

The standard deviation for 'Profit' was relatively high, pointing to large variations in profitability across transactions.A screenshot of a computer

Description automatically generated  
A screen shot of a computer

Description automatically generated  
A screen shot of a computer

Description automatically generated

### 5.2 Data Transformation

One Hot Encoding: This technique converts categorical variables into a form that could be provided to machine learning algorithms to better predict the outcome. However, an error with the OneHotEncoder suggests there might have been a problem with this step.

A screenshot of a computer

Description automatically generated

Standardization: We included a computation of z-scores for the 'Sales' column, which is a form of standardization. This process adjusts the data to have a mean of 0 and a standard deviation of 1, which is particularly useful for comparison between different datasets and for models that assume normally distributed data.

A screenshot of a computer

Description automatically generated

# Chapter 6: Data Analysis and Results

A graph with green and blue squares

Description automatically generated

The visualization is a waterfall chart that displays the sales contributions of various sub-categories within a company. It aims to show how each sub-category contributes to the total sales, highlighting both positive and negative impacts on the overall financial performance.

A map of the united states

Description automatically generated

The visualization is a geographical map of the United States displaying the sum of sales by state, differentiated by payment type—credit card, debit card, and digital wallet. Each state has pie charts overlaid on the map, where the size of each pie segment represents the proportion of sales from each payment method in that state. This map helps in identifying payment preferences across different regions and can be used to tailor marketing strategies, like targeted promotions and discounts, based on the preferred payment methods in each area.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

The visualization consists of two parts:

*Top Chart (Line Graph)*: Shows the sum of sales and last month's sales by month over a period from January 2019 to around mid-2023. This graph is useful for observing the trends in sales over time, identifying any seasonal patterns, and noting any significant peaks or drops in sales.

*Bottom Chart (Bar Graph):* Displays the month-over-month (MoM) percentage change in sales during the same period. Positive values (shown in blue) indicate months where sales increased compared to the previous month, while negative values (shown in red) indicates a decrease.

Overall, this visualization helps in understanding how sales fluctuated month to month over several years, providing insights that could be used to make strategic business decisions, optimize sales efforts, and forecast future trends.

A screenshot of a computer

Description automatically generated

The visualization is an interactive hierarchy chart, also known as a drill-down chart, used to display sales data over multiple years, broken down further into quarters and months for detailed analysis. The left side of the chart shows a timeline of annual sales totals, which can be expanded to reveal sales figures for each quarter within a year, and further into each month.

On the right side, the chart offers a geographical breakdown of sales by state and even more detailed data by city within each state. This allows users to click through and see precisely how much was sold each year, quarter, and month, and understand the distribution of sales geographically.

The purpose of this visualization is to provide a comprehensive view of sales trends over time and across different regions. This insight is crucial for identifying patterns, planning marketing strategies, and allocating resources efficiently based on where and when products are selling best.

A screenshot of a computer

Description automatically generated

The visualization displays a comprehensive analysis of sales data over several years. It consists of three parts:

*Data Table:* This table shows monthly sales data for each year, including actual sales, target sales, and the deviation between the two. This allows for a detailed month-by-month comparison of expected versus actual financial performance.

*Bar Chart:* Represents the count of sales transactions each month.

*Line Chart:* Plots the sum of sale deviations (difference between actual and target sales) over the same period.

The combination of these elements in the visualization provides a clear overview of sales performance, highlighting where and when sales fell short of or exceeded targets. This kind of analysis is critical for businesses to understand sales trends, identify potential issues, and make informed decisions to improve future performance.

A screenshot of a computer

Description automatically generated

The visualization is focused on analyzing the impact of different shipping modes on the on-time delivery rate of orders. It clearly shows that 'Same Day' shipping significantly increases the likelihood of on-time delivery compared to other shipping options like 'Standard Class,' 'First Class,' and 'Second Class.'

The bar chart illustrates this by displaying a much higher average on-time delivery rate for 'Same Day' shipments. This suggests that if timeliness is critical, 'Same Day' shipping is the most effective mode. This insight can help businesses prioritize shipping options and improve customer satisfaction by ensuring timely deliveries.

# 

# Chapter 7: Business Implications

Linking back to the initial problem statement of increasing competition, customer retention, and cost optimization, our proposed solutions provide a comprehensive strategy to leverage data-driven insights for tactical and strategic decision-making. Implementing these recommendations will enable Staples to not only defend but also expand its market share by enhancing the customer experience and optimizing operational efficiencies.

Targeted Investment in High-Impact Categories: Our analysis identifies categories that significantly contribute to sales, aligning with the goal of strengthening market position against rising e-commerce competition. Resources can be allocated more effectively to ensure inventory meets demand, particularly for high-performing categories identified in our analysis.

Optimization of Shipping for Enhanced Customer Retention: With 'Same Day' shipping showing a positive impact on on-time delivery rates, optimizing shipping options can directly enhance customer satisfaction. This aligns with the need for personalized customer experiences, as reliable and fast delivery is a critical component of such strategies. Implementing a tiered shipping strategy that prioritizes 'Same Day' delivery for premium customers can significantly enhance customer retention and satisfaction.

Promotional Strategies Tailored to Payment Preferences: By leveraging insights from payment mode preferences, promotional and discount strategies can be more effectively aligned with customer behavior, driving sales and enhancing customer engagement. Introducing promotions based on observed payment trends and seasonal buying patterns can increase sales volume and improve inventory turnover, aligning with cost optimization strategies.

Performance Tracking and Adjustment: Integrating real-time tracking tools to monitor sales performance against targets enables quick strategic adjustments, helping to mitigate risks associated with inventory excess or shortages.

Utilization of Sales Transaction Frequencies: Insights into when sales transactions peak provides invaluable data for managing inventory more effectively, crucial for responding to increased competition and optimizing costs. Using predictive analytics to understand customer purchasing patterns can facilitate more personalized marketing and sales strategies, thereby increasing customer engagement and retention.