

George Brown College

Course: Foundation of Data Management

Professor: Daniel Vitaver

Assignment 1 & 2

Submitted by Group 5:

John Allan Ellingson: 1007711881

Pegah Farazmand: 101611990

Jorge Martinez: 100853469

Jeffrey Okoduwa: 101477700

Ryan Tran: 101460443

Gordon Yeung: 101624785

Business Context

1. Industry and business entities

Industry:

Big-box retailer of office supplies, furniture and business technology

Entities impacting the reporting process:

Customers: purchasing behavior affects profit and satisfaction.

Sales Teams: apply for discounts and track regional performance.

Shipping/Logistics: responsible for delivery times, which affect satisfaction.

IT/BI Analysts: manage data pipelines and reporting tools.

Store and warehouse team: responsible for giving information about products and their inventory

Executives: decision-makers who interpret summary-level reports.

Orders: Core source for KPIs like profit margin, discount, and shipping time.

Geography: Enables regional analysis and strategy based on area-level KPIs.

2. Target Audience

Operational Report:

Sales Supervisors: Monitor discount usage and sales trends.

Logistics Coordinators: Track shipping durations and fulfillment times.

Store and Warehouse Managers: Act on inventory and regional performance insights.

Executive Report:

VP of Sales: Evaluate regional sales performance and adjust strategy.

CFO (Finance Executive): Monitor profit margins and assess discount impact.

VP of Marketing: Allocate campaign resources to high-performing regions.

Company Stakeholders / Investors: Review high-level performance indicators.

3. Purposes of reports

Operational Report

Monitor discount effectiveness by category

Track shipping delays and fulfillment time

Support daily decisions on pricing and logistics

Improve store-level performance and efficiency

Executive Report:

Evaluate regional sales and profitability trends

Adjust pricing and discount strategies quarterly

Plan campaigns and allocate budgets by region

Analyze impact of shipping on customer satisfaction

4. Publishing and Distribution Frequency, Maintenance, and Use

Operational Report:

Frequency: Weekly (or monthly summaries)

Distribution: PDF or dashboard (e.g., Power BI or Excel)

Maintenance: Data is regularly updated from MySQL and reviewed each week

Use: Monitor discounts, shipping times, and regional sales in real-time

Executive Report:

Frequency: Monthly or Quarterly

Distribution: PDF or presentation with key insights

Maintenance: Reviewed by BI analysts before release to ensure accuracy

Use: Strategic planning, budget allocation, and pricing decisions

5. Assumptions and metrics

Based on the provided data, the assumptions and hypotheses are as follows:

- Returned orders will be for all the items in 1 order
- Each individual product has 1 unique ID as identification and will be override to ensure the accuracy of products and performance reports

Hypothesis 1: High discounts are linked to low profitability in certain categories or regions.

This hypothesis investigates: It evaluates whether larger discounts reduce profit margins, especially in specific product categories or geographic areas, and whether discount policies need to be revised.

Operational Report: Store Managers, Sales Supervisors

Executive Report: CFO

Managers adjust daily discount policies, and the CFO monitors overall financial impact.

Hypothesis 2: Certain regions consistently outperform others in sales and profit.

This hypothesis investigates: It explores whether specific regions generate higher revenue and profitability, guiding where to focus marketing and sales efforts.

Operational Report: Regional Sales Managers, Store Managers

Executive Report: VP of Sales, VP of Marketing

Helps identify high-performing areas and underperforming regions for better resource allocation.

Hypothesis 3: Longer shipping times negatively affect customer satisfaction and operational performance.

This hypothesis investigates: It assesses whether delayed shipping correlates with lower profit margins or customer dissatisfaction, highlighting a need for improved logistics.

Operational Report: Warehouse Coordinators, Logistics Managers

Executive Report: VP of Operations

Supports decisions on optimizing shipping methods and reducing delivery delays.

6. Metrics for Operational and Executive Reports

Feature	Operational Report	Executive Report
Sales, Quantity, Profit	daily tracking by region and category	aggregated by region and product line
Discounts	per order and category	monthly average by region/category
Shipping Duration	calculated as Ship Date, Order Date	average per region/product
Region/Province	for shipment routing, local trends	to analyze regional investment strategies
KPIs	Profit Margin, Discount Amount, Shipping Time	Region Contribution %, Avg Profit, Profit Margin

KPIs derived from hypotheses:

- Profit Margin = Profit / Sales (H1, H3)
- Discount Amount = Sales × Discount (H1)
- Region Contribution % = (Region Sales / Total Sales) × 100 (H2)
- Shipping Time = Ship Date – Order Date (H3)

B. Technical Data Analysis

Data cleaning

Initially, the data inspection was conducted on excel using filter and power query for summary.

Problems & Solutions

1. Table: Sample-Superstore

- Duplicates: Order Id, multiple transaction happened for different items on the same order. Rename the column “Row ID” into “transaction_id” so it becomes the unique identifier for each row of data.
- Redundancies: country_region, all values “United States”, the column was removed
- Inconsistencies:
 - o product_id should be unique identifier to determine the product, however, there are multiple products having different id numbers (1849 unique product name but 1852 unique product id). The product_id data of the products having multiple id was overridden by the first value appeared in the spreadsheet to ensure consistency using power query on excel sheet.
 - o Sales data were collected with different decimal rounding. All sales numbers were rounded to 2 decimal places.

2. Table: Returns

Redundancies: Returned columns because all of them are “Yes”. So this column was removed.

Adding returned_id for the index of each return.

Recommended to add information about the amount of return

3. Table: People

Information about the manager of each region, no cleaning needed.

C. Data Normalization

Renamed all the column with lower letters and replace the space with “_”

The Sample_superstore sheet was UNF, in order to maintain efficient data storage and management the data was normalized into smaller tables as reference data with the following tables:

Customers: customer_id, customer_name, region, city, state, segment, postal_code

Products: product_id, product_name, category, sub_category

Orders: transaction_id, order_id, order_date, ship_date, customer_id, product_id, sales, quantity, discount

Shipments: ship_id, order_id, order_date, ship_date, ship_mode

People: region, manager

Returns: return_id, order_id

Orginal data: includes all the original data and added columns for metrics and reports

ERD and DFD are included.