Team-S2O\_Group5

Gummaraj Shivakumar Sridevi, Trivikram - (gsstrivikram18)

Mandapati, Puneeth Varma - (varmapuneeth07)

Ramesh, Kirthana - (kirthanaramesh)

Thondepu, Manish Kumar - (manisht)

Tumma, Teja - (krishnateja)

Dataco supply chain analysis

MIS 587 Project Report

Table of Contents

[Introduction and Client Overview 2](#_Toc165557861)

[Dataset Description 3](#_Toc165557862)

[Data Warehouse Design 4](#_Toc165557863)

[Star Schema Design 4](#_Toc165557864)

[Data Preparation 7](#_Toc165557865)

[Remove Columns 7](#_Toc165557866)

[Rename Columns 7](#_Toc165557867)

[Handling Null Values 8](#_Toc165557868)

[Transformations 9](#_Toc165557869)

[Data Exploration 10](#_Toc165557870)

[Descriptive statistics 10](#_Toc165557871)

[Outlier Analysis 10](#_Toc165557872)

[Data Visualization 11](#_Toc165557873)

[Data Analysis & Results 13](#_Toc165557874)

[Visualization Description 13](#_Toc165557875)

[Insights from Visuals 14](#_Toc165557876)

[Business Implications 17](#_Toc165557877)

# Introduction and Client Overview

DataCo Supply Chain is a leading player in the retail industry, specializing in clothing, sports, and electronic supplies. With a global presence and a diverse product range, DataCo Supply Chain has established itself as a key player in the market, focusing on efficient supply chain management and inventory optimization to drive sales and customer satisfaction. DataCo Supply Chain caters to a wide range of customers, from individual consumers to large corporations. Its customer base spans across different demographics and geographic regions, reflecting the company's commitment to providing quality products and services to diverse markets.

DataCo Supply Chain is engaged in various activities across the supply chain, including provisioning, production, sales, and commercial distribution. These core activities are strategically managed to ensure seamless operations and timely delivery of products to customers worldwide. It leverages advanced technologies to enhance its operations and drive innovation. By integrating structured and unstructured data sources, the company gains valuable insights into customer behavior, market trends, and operational efficiency.

The primary objectives of the project include optimizing inventory management, reducing stockouts, and enhancing customer satisfaction with delivery status. These objectives are aligned with the company's mission to deliver quality products and services while maintaining a competitive edge in the retail industry.

# Dataset Description

The dataset under analysis comprises key operational metrics pertaining to a supply chain management system. Below is the data dictionary that details each column present in the dataset, providing insights into its structure and significance.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Datatype** |
| Type | Type of transaction made. | String |
| Days for shipping (real) | Actual shipping days of the purchased product. | Double |
| Days for shipment (scheduled) | Days of scheduled delivery of the purchased product. | Double |
| Benefit per order | Earnings per order placed. | Double |
| Sales per customer | Total sales per customer. | Double |
| Delivery Status | Status of the order's delivery (e.g., Late, On time). | String |
| Late\_delivery\_risk | Indicates risk level of late delivery (0 for low risk, 1 for high risk). | Double |
| Category Id | Identifier for product category. | Double |
| Category Name | Name of the product category. | String |
| Customer City | City of the customer. | String |
| Customer Country | Country of the customer. | String |
| Customer Email | Email address of the customer. | String |
| Customer Fname | First name of the customer. | String |
| Customer Id | Unique identifier for each customer. | Double |
| Customer Lname | Last name of the customer (surnames). | String |
| Customer Password | Encrypted customer password. | String |
| Customer Segment | Segment of the customer (e.g., Consumer, Corporate). | String |
| Customer State | State of the customer's location. | String |
| Customer Street | Street address of the customer. | String |
| Customer Zipcode | Zip code of the customer's location. | Double |
| Department Id | Identifier for department within the company. | Double |
| Department Name | Name of the department. | String |
| Latitude | Latitude of the store location. | Double |
| Longitude | Longitude of the store location. | Double |
| Market | Market region to which the order was delivered. | String |
| Order City | City where the order was placed. | String |
| Order Country | Country where the order was placed. | String |
| Order Customer Id | Identifier linking the order to a specific customer. | Double |
| order date (DateOrders) | Date when the order was placed. | Date |
| Order Id | Unique identifier for the order. | Double |
| Order Item Cardprod Id | Product identifier. | Double |
| Order Item Discount | Discount amount applied to the order item. | Double |
| Order Item Discount Rate | Discount rate applied to the order item. | Double |
| Order Item Id | Identifier for each item within an order. | Double |
| Order Item Product Price | Price of the product as per the order item. | Double |
| Order Item Profit Ratio | Profit ratio per order item. | Double |
| Order Item Quantity | Quantity of each item ordered. | Double |
| Sales | Total sales amount for the order. | Double |
| Order Item Total | Total amount for the order item after discounts. | Double |
| Order Profit Per Order | Total profit made from the order. | Double |
| Order Region | Region to which the order was delivered. | String |
| Order State | State to which the order was delivered. | String |
| Order Status | Current status of the order (e.g., Complete, Pending, Closed). | String |
| Product Card Id | Identifier for the product card. | Double |
| Product Category Id | Category identifier for the product. | Double |
| Product Description | Description of the product. | String |
| Product Image | URL of the product image. | String |
| Product Name | Name of the product. | String |
| Product Price | Price of the product. | Double |
| Product Status | Availability status of the product. | String |
| Shipping date (DateOrders) | Exact date and time of shipping. | Date |
| Shipping Mode | Mode of shipping used for the delivery of the order. | String |

Business Process Representation

The dataset provides a comprehensive overview of the business processes involved in supply chain management and sales. By analyzing real and scheduled shipment days, the dataset helps in optimizing logistics and improving delivery efficiency. The sales-related fields assist in understanding customer purchase patterns and profitability per transaction, crucial for strategic decision-making in marketing and finance. This data is instrumental in driving improvements in operational efficiency, customer satisfaction, and overall business performance.

The dataset provided for the DataCo Supply Chain Analysis Inventory Management system comprises a comprehensive array of structured data fields essential for understanding and optimizing various aspects of the supply chain and inventory management processes. These fields encompass transactional details, customer information, product specifications, geographical data, and order statuses, among others.

The dataset encompasses a wide range of transactional metrics, including "Days for shipping (real)" and "Days for shipment (scheduled)," which provide insights into the actual and expected durations for product delivery, respectively. These metrics are crucial for assessing the efficiency and reliability of the shipping process. Additionally, the dataset includes financial indicators such as "Benefit per order" and "Sales per customer," offering insights into revenue generation and customer spending patterns.

Customer-related information is also captured comprehensively within the dataset, with fields such as "Customer City," "Customer Country," "Customer Segment," and "Customer Zipcode." These details enable the analysis of customer demographics, geographic distribution, and segmentation, facilitating targeted marketing strategies and customer relationship management initiatives.

Product-related attributes play a pivotal role in inventory management and sales analysis. The dataset includes fields such as "Category Name," "Product Description," "Product Price," and "Product Status," providing insights into product categories, descriptions, pricing, and stock availability. These details are instrumental in identifying popular product categories, optimizing pricing strategies, and ensuring adequate stock levels to meet demand.

Geographical data, including "Market," "Order City," "Order Country," and "Order Region," offer valuable insights into the global distribution of orders and sales. Analysis of this data enables the identification of regional trends, market penetration strategies, and opportunities for expansion or optimization in specific geographic areas.

Order-related metrics, such as "Order Date," "Order Status," "Shipping Mode," and "Delivery Status," provide a comprehensive overview of the order fulfillment process. These details allow for the assessment of order processing times, delivery efficiency, and overall customer service performance. Additionally, they facilitate the identification of potential bottlenecks or issues in the supply chain workflow.

Furthermore, the dataset includes unique identifiers such as "Customer Id," "Order Id," and "Product Card Id," enabling the linkage and aggregation of data across different tables or databases. These identifiers facilitate data integration, analysis, and reporting, ensuring a cohesive and holistic view of the supply chain operations.

In summary, the dataset for the DataCo Supply Chain Analysis Inventory Management system encompasses a rich array of structured data fields covering transactional, customer, product, geographical, and order-related aspects. This comprehensive dataset serves as a valuable resource for conducting in-depth analysis, generating actionable insights, and optimizing inventory management strategies to enhance overall operational efficiency and business performance.

# Data Warehouse Design

A comprehensive data warehouse to streamline and enhance DataCo’s data management is created along with reporting capabilities. This project implements a star schema design, integrating OLTP tables and OLAP dimension and fact tables.

## Star Schema Design

The star schema comprises a central Orders Fact Table, surrounded by several dimension tables: Product, Customer, Orders, and Date Dimensions. Surrogate keys are created for each of the dimension tables to uniquely identify each row and are used to map with the fact table.

Fact Table

The Orders Fact Table serves as the central hub for the star schema, capturing key metrics related to transactions. It includes identifiers such as OrderSkey, which uniquely identifies each order, and links to dimension tables, including ProductSkey, CustomerSkey, and OrderDateSkey. Additionally, it holds details such as OrderID (for tracking individual transactions), ProductID (the identifier for each product), ProductPrice (price per unit), Quantity (number of items purchased), ProfitAmount (profit earned from each sale), ItemDiscount (discount applied per item), and PriceAfterDiscount (the discounted price of the product).

Dimensions

The Date Dimension provides a comprehensive structure for tracking transaction dates. It includes a unique identifier for each date, DateSkey, along with a variety of ways to view the date, such as DateString, Day, DayOfYear, DayOfWeek, and DayOfWeekName. It also provides aggregated periods like Week, Month, MonthName, Quarter, and Year. Additionally, it includes indicators for special cases, such as IsWeekend (indicating if the date falls on a weekend) and IsLeapYear (indicating if the year is a leap year).

The Customer Dimension offers insights into customer details, including unique identifiers such as CustomerSkey and CustomerID for tracking. It also includes personal details like CustomerFname and CustomerLname, with a concatenated CustomerFullname for completeness. Additionally, CustomerSegment helps categorize customers into various segments, aiding in customer profiling and targeted analysis.

The Orders Dimension provides detailed tracking of individual transactions, including unique identifiers such as OrderSkey and OrderID. It tracks order-related details such as OrderDate, ShippingDate, and both ActualShippingDays and ExpectedShippingDays, allowing for comparisons. It also includes DeliveryStatus and IsDeliveryLate to indicate the state of each order. Additionally, it captures geographic information such as OrderCity, Country, Region, and State, along with the WarehouseCity, WarehouseCountry, WarehouseStreet, and WarehouseState. This dimension concludes with ShippingMode, outlining how the order was shipped.

The Product Dimension provides comprehensive product details, including unique identifiers such as ProductSkey and ProductID. It categorizes products into CategoryName and DepartmentName, aiding in tracking and analysis. It also includes the ProductName, providing a complete overview of each product.

A diagram of a data flow

Description automatically generated with medium confidence

SSIS Process for Data Warehouse Population

Data Warehouse is created using SSIS (SQL Server Integration Services) process. This process starts by truncating all data warehouse tables and then populating the dimension tables followed by the fact tables. Below is a detailed explanation of each stage:

**Stage 1: Truncate DW Tables**

The first step clears existing data from all dimensions and fact tables in the data warehouse, ensuring a clean slate for the new data to be loaded.

**Stage 2: Populate Dimension Tables**

1. Customers Dimension: This table stores information about customers. The process extracts data from the OLTP (Online Transaction Processing) tables, sorts it to remove duplicate records based on IDs, and then processes it through a derived column stage, converting attributes to their respective data types before loading into the data warehouse.
2. Similarly, Products, Orders, Customers, and Date dimension tables are populated.

**Stage 3: Populate Orders Fact Table**

The fact table's population process involves a series of steps to ensure accurate and comprehensive data storage. It begins by performing lookups to fetch surrogate keys from the dimension tables, including Product, Customer, Orders, and Date Dimensions. These lookups provide references for attributes such as ProductSkey, CustomerSkey, OrderDateSkey, and others, linking the fact table to relevant descriptive data. Additionally, the process includes a Derived Column stage, which generates new columns by applying expressions to existing attributes. In this stage, two new attributes are derived:

* CostPrice: Calculated as product\_price \* order\_product\_quantity, representing the total cost based on product price and quantity.
* PriceAfterDiscount: Calculated as (product\_price \* order\_product\_quantity) - order\_item\_discount, representing the total price after discounts are applied.
* These derived attributes provide key metrics for tracking each transaction, enhancing the table's data granularity.

A screenshot of a computer

Description automatically generated

# Data Preparation

In this chapter, the initial steps taken to prepare the data for the ETL (Extract, Transform, Load) process are discussed, executed using Python. This encompasses data cleaning, preprocessing, and transformations applied during the ETL process. Leveraging Python's extensive libraries such as NumPy, pandas, seaborn, and matplotlib for data manipulation and analysis, a robust framework was provided for efficiently handling and transforming the dataset.

## Remove Columns

After review, it was decided to remove columns that were not relevant to the analysis. These included 'Type', 'Order Item Discount Rate', 'Order Item Profit Ratio', 'Benefit per order', 'Sales per customer', 'Category Id', 'Customer Email', 'Customer Password', 'Department Id', 'Order Item Cardprod Id', 'Order Zip code', 'Order Status', 'Product Card Id', 'Product Description', 'Product Image', 'Product Status', 'Customer Zip code', 'Order Customer Id', and 'Product Category Id'. The decision to omit these fields was driven by their limited impact on the analysis objectives.

## Rename Columns

Next, adjustments were made to the column names to improve clarity and consistency across the dataset. The changes included:

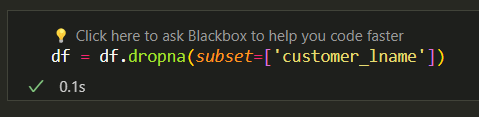
|  |  |
| --- | --- |
| **Original Column Name** | **New Column Name** |
| Days for shipping (real) | Actual Shipping Days |
| Days for shipment (scheduled) | Expected Shipping Days |
| Late\_delivery\_risk | Is Late Delivery |
| Customer City | Warehouse City |
| Customer Country | Warehouse Country |
| Customer State | Warehouse State |
| Customer Street | Warehouse Street |
| Customer zipcode | Warehouse Zipcode |
| Order Item Id | Product ID |
| Order Item Quantity | Order Product Quantity |
| Sales | Cost Price |
| Order Item Total | Price after Discount |
| Order Profit Per Order | Profit Amount |
| Shipping date (DateOrders) | Delivery Date |
| Order date (DateOrders) | Order Date |

Additionally, necessary adjustments were made to the data types within the dataset to ensure compatibility with analytical processes.

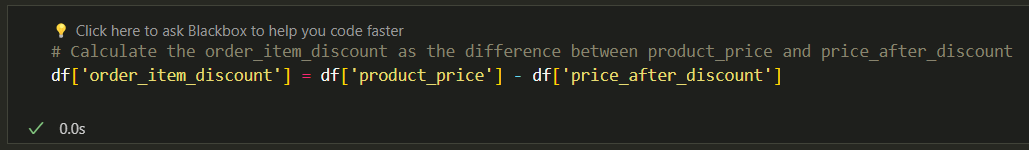
## Handling Null Values

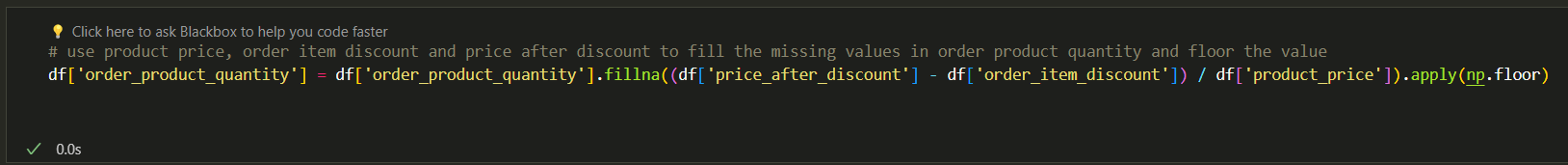
The presence of null values in the dataset was addressed, notably found in the customer\_lname, order\_item\_discount, and order\_product\_quantity columns.

* Customer Last Name (customer\_lname): Missing values in the customer\_lname column were removed due to their importance in customer identification and analysis.



* Order Item Discount (order\_item\_discount): Missing values were recalculated using the formula df['product\_price'] - df['price\_after\_discount'] to maintain accuracy and consistency with financial data.

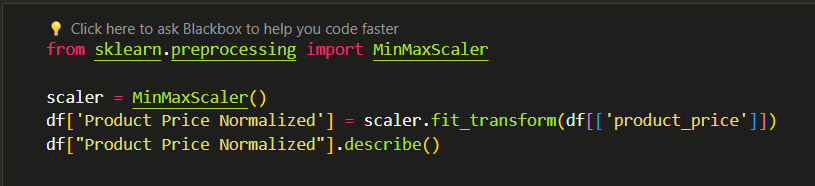


* Order Product Quantity (order\_product\_quantity): Missing values were imputed using a calculated approach based on price and discount relationships. The formula (df['price\_after\_discount'] - df['order\_item\_discount']) / df['product\_price'] was utilized, and the results were floored for whole numbers, ensuring logical consistency in transactional data. 

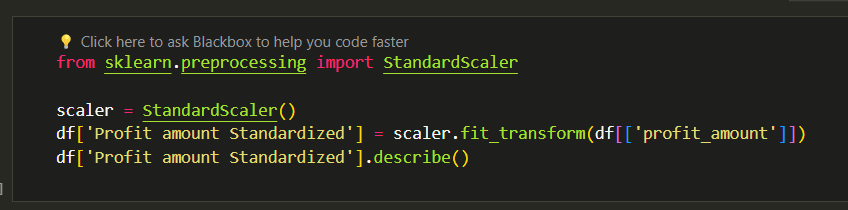
## Transformations

Transformations were performed to normalize and standardize key financial metrics in the dataset.

* Normalization of product\_price: Min-Max scaling was applied to the product\_price column to normalize these values between 0 and 1.



* Standardization of profit\_amount: To standardize the profit\_amount, the StandardScaler from Scikit-learn was employed. This transformation adjusted the data to have a mean of zero and a standard deviation of one.



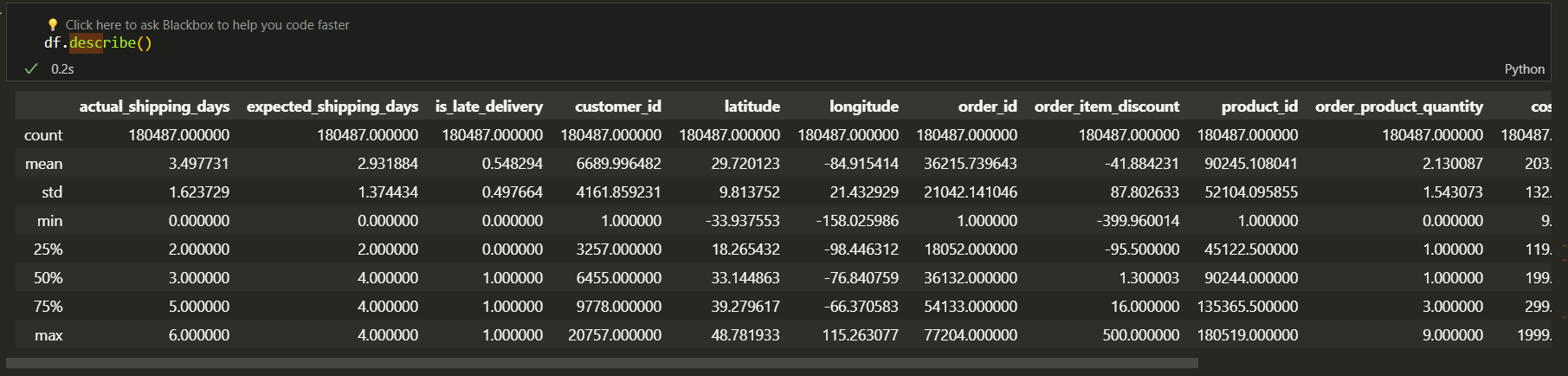
The dataset is now reliable, organized, and ready for the next step.

# Data Exploration

The initial focus was the dataset's underlying structure and distributions through Descriptive Statistics and a targeted Outlier Analysis.

## Descriptive statistics

Exploratory data analysis began with descriptive statistics, revealing central tendencies, variability, and distribution patterns essential for accurate interpretations.



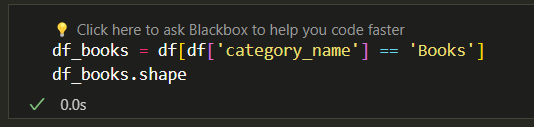
## Outlier Analysis

Identifying and understanding outliers is vital, as they can significantly impact the outcomes of the analyses and could potentially indicate data issues or special scenarios that require separate investigation.

In the dataset, detecting outliers required a nuanced approach, particularly within the category\_name column. It was observed that anomalies predominantly surfaced in the entries categorized under 'Books'. This revelation was critical, as it highlighted specific areas within the data that might skew analysis results or reflect unique cases.

To address these outliers effectively, the following steps were taken:

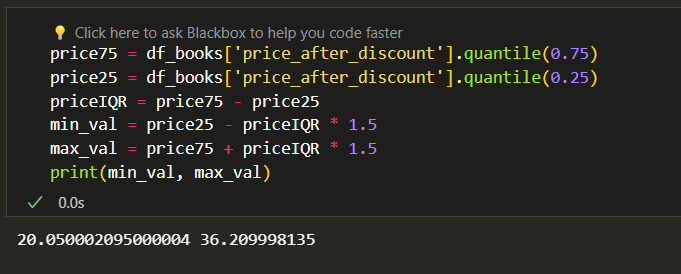
* Isolation of the Category: To closely examine the characteristics and distributions, the data for the 'Books' category was isolated.



* Boxplot for Outlier Detection: To identify outliers in the order\_item\_discount, boxplots were utilized. This visual helped in assessing the extremities in discount values which could be potential outliers or errors.



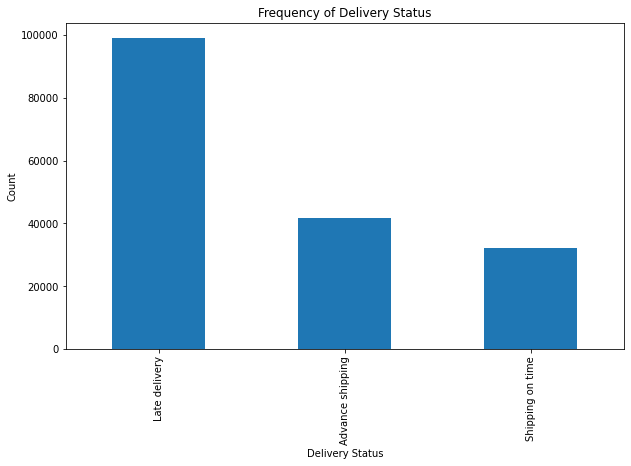
* Statistical Analysis: Using statistical methods, the data for any values that significantly deviated from the norm was assessed. This involved looking at metrics such as z-scores or IQR (Interquartile Range) to identify and confirm these outliers.



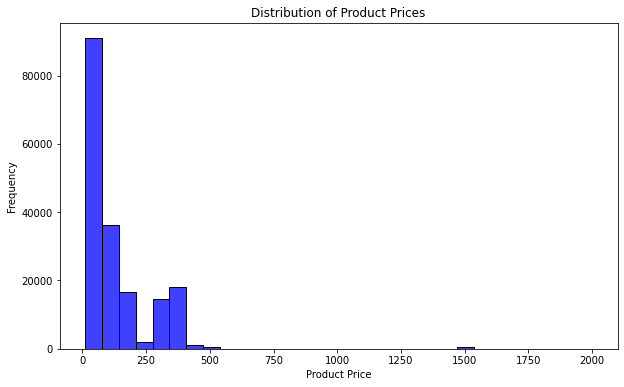
## Data Visualization

The visualizations help illustrate the behavior and characteristics of different data aspects within the dataset, facilitating better understanding and decision-making.

For the categorical variable delivery\_status, a bar chart was used to visualize its frequency distribution. This visualization helps in understanding the most common delivery outcomes and identifying any potential issues with order fulfillment.



For the numerical variable product price, a histogram examines its distribution.

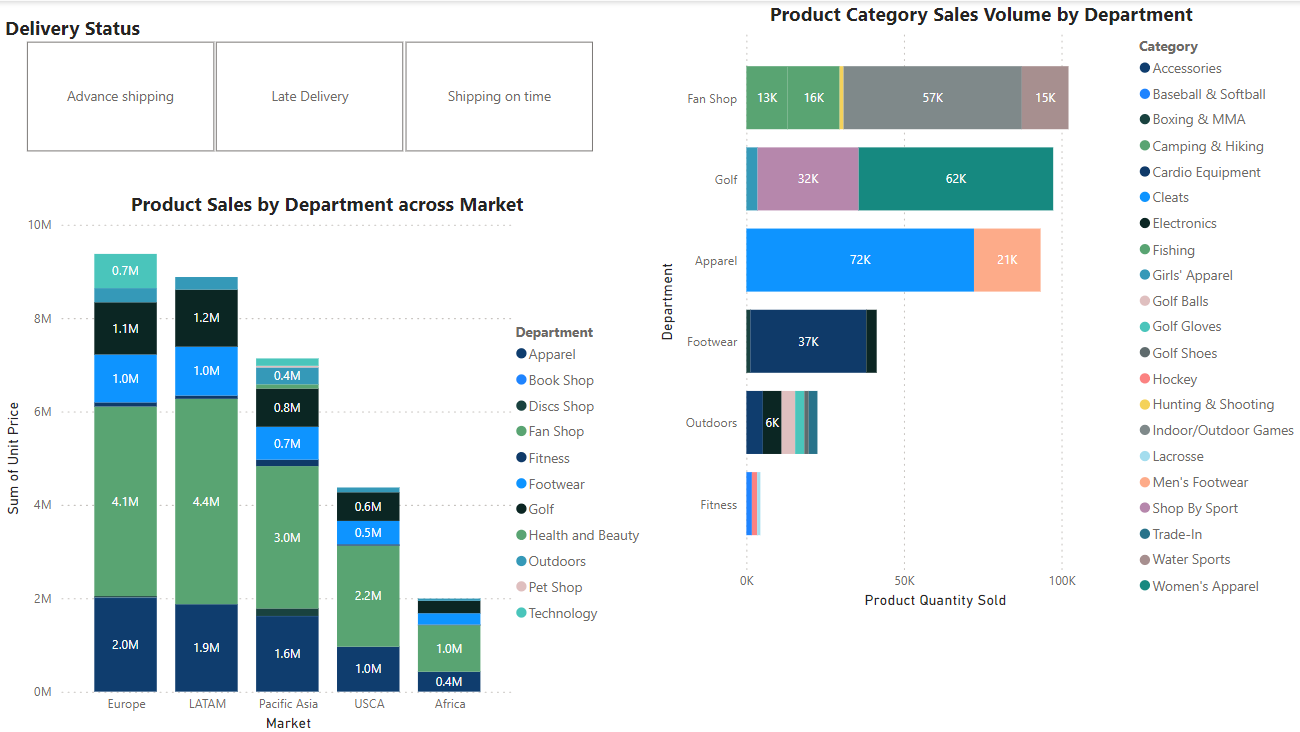


The cleaned data has been exported to a CSV file for integration into the data warehouse, setting the stage for deeper analysis in subsequent phases.

# Data Analysis & Results

For data analysis, Microsoft Power BI for creating visualizations is chosen. This tool enables to effectively display and interpret complex datasets, facilitating a clearer understanding of sales and inventory metrics across different dimensions.

This section highlights two main visualizations designed to analyze product sales across various departments and markets, as well as sales volume across different categories within each department. These insights are key to understanding sales performance and consumer preferences, aiding in effective inventory management.

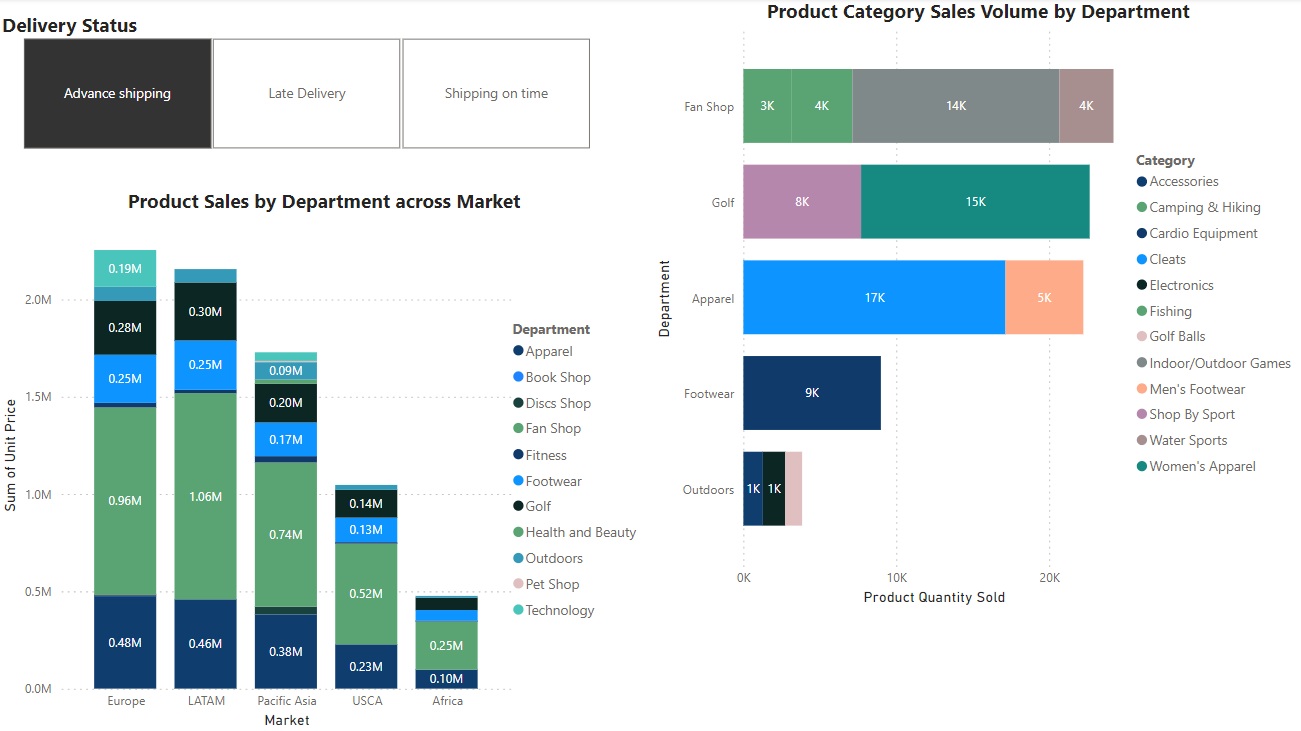


## Visualization Description

1. ***Product Sales by Department Across Various Markets:***This stacked column chart visualizes product sales by department across different global markets. The X-axis lists the markets—Europe, LATAM, Pacific Asia, USCA, and Africa—while the Y-axis represents the sum of the unit price. Each color segment in the stack indicates a department, such as Apparel, Book Shop, Discs Shop, etc.
2. ***Product Category Sales Volume by Department:***This visualization represents the sales volume of various product categories within each department. It employs a horizontal stacked chart with the X-axis showing departments like Fan Shop, Golf, Apparel, Footwear, and Outdoors, while the Y-axis quantifies the product quantity sold. Each stack segment is color-coded by product category such as Accessories, Cardio Equipment, and Cleats etc.
3. ***Delivery Status Slicer*:** A slicer for 'Delivery Status' is applied across the visualizations to allow dynamic filtering of data based on three categories: Advanced Shipping, Late Delivery, and On Time. This interactive feature enables users to focus on specific segments of the data, facilitating deeper analysis of shipping efficiencies and their impact on overall sales and inventory management.

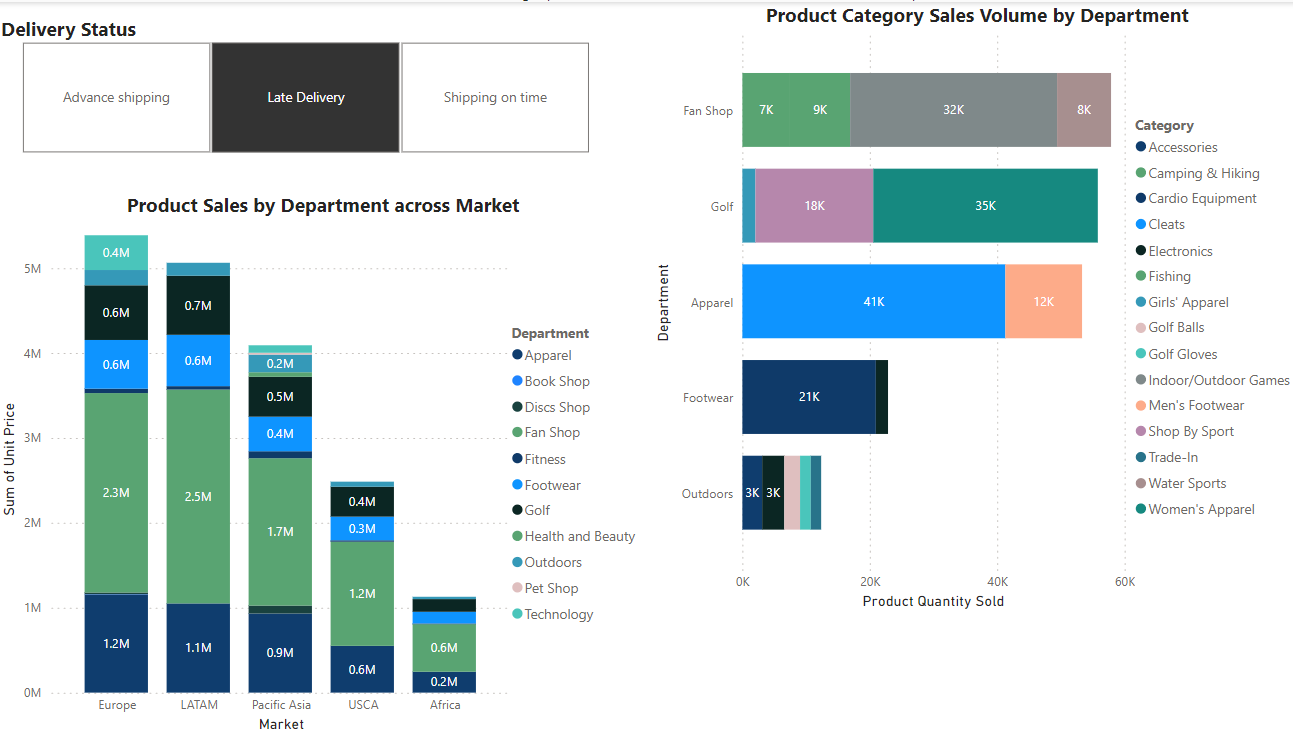
## Insights from Visuals

1. ***No slicer selected:***
   * *Product Sales by Department Across Markets:* The Fan Shop department shows the highest sales in Europe and LATAM, suggesting it is a dominant player in these markets. The Fitness department is not much affected by any slicer shipping condition
   * *Product Category Sales Volume by Department:* The Apparel department has a high sales volume in "Cleats", indicating consistent popularity. "Women's Apparel" in the Golf department shows significant sales volume.
2. ***Advanced Shipping:***



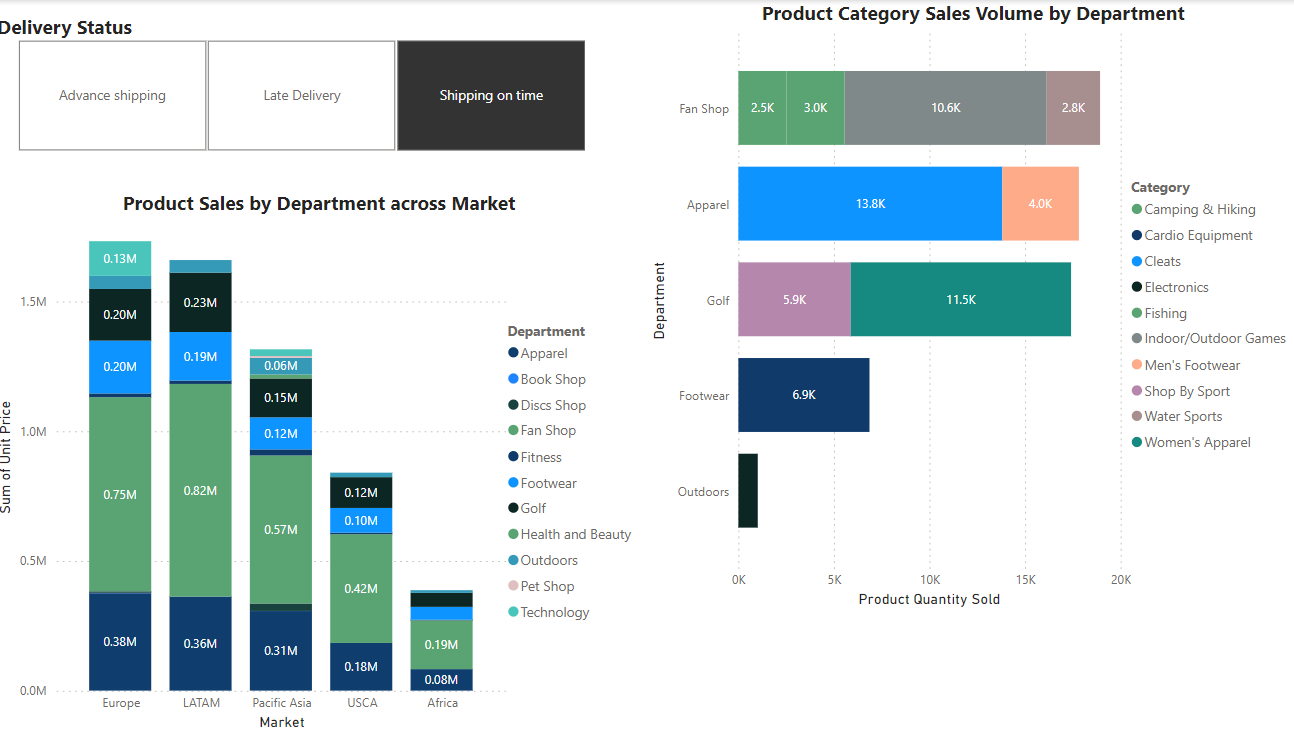
* + *Product Sales by Department Across Markets:* Sales in the Apparel department are high in Europe under Advanced Shipping conditions, which could suggest efficient supply chain operations in this region.
  + *Product Category Sales Volume by Department:* High sales volumes of "Women Apparel" in the Golf department under Advanced Shipping conditions suggest successful pre-season stocking or promotions. The "Cardio equipment" category in the Footwear department shows robust sales, which could be driven by seasonal sports events.

1. ***Late Delivery:***



* + *Product Sales by Department Across Markets:* Notably, the Fan Shop and Golf departments show resilience in sales despite Late Delivery conditions, especially in Europe, LATAM and Pacific Asia.
  + *Product Category Sales Volume by Department:* "Cardio equipment" in the Footwear department still performs well in terms of sales volume despite late delivery scenarios, suggesting strong demand overrides delivery timing issues.

1. ***Shipping On-Time:***



* + *Product Sales by Department Across Markets:* On-time shipping boosts sales across all departments with notable spikes in Apparel in LATAM and Europe, reflecting the critical impact of timely delivery on sales.
  + *Product Category Sales Volume by Department:* "Electronics" in the Outdoors department shows increased sales volume with on-time shipping. This also signifies that customers ordering other product categories from the Outdoors department are not tolerant of late deliveries.

# Business Implications

**1. Strategic Overstocking in Key Departments and Markets**

The consistent high sales volumes in the 'Apparel' and 'Footwear' departments across Europe and LATAM markets suggest a strong consumer demand. Overstocking these items may prevent stockouts during peak shopping periods, ensuring customer satisfaction and capturing maximum sales.

*Recommendation:* Develop a strategic overstocking plan for these departments in key markets and monitor sales trends to dynamically adjust inventory levels as needed.

**2. Addressing Late Deliveries in Pacific Asia**

The data indicates issues with late deliveries, particularly in the 'Apparel' and 'Footwear' departments in the Pacific Asia market. This could affect customer satisfaction and potentially lead to lost sales if consumers seek alternatives due to availability issues.

*Recommendation:* Improve logistic operations by collaborating with reliable local carriers and consider setting up a local distribution hub to enhance delivery efficiency and reduce lead times.

**3. High Demand Product Categories Need Consistent Restocking**

Categories like 'Indoor/Outdoor Games' in the 'Fan Shop' and 'Women's Apparel' in the 'Golf' department show significant sales, indicating strong market demand. Consistent availability is crucial to maintain sales momentum.

*Recommendation:* Prioritize these categories in inventory restocking strategies to avoid potential stockouts. Employ predictive analytics to accurately forecast demand and schedule timely restocking.

**4. Stable Demand Products Require Effective Inventory Management**

Products such as 'Cleats' in the 'Golf' department demonstrate consistent sales, suggesting steady demand. This predictability provides an opportunity to optimize inventory management and reduce the risk of unsold stock.

*Recommendation:* Maintain optimal inventory levels for these products to ensure continuous availability without overstocking, leveraging data analytics for precise restocking.