

Supply Chain Forecasting: A Data-Driven Analytical Framework Using Python and Power BI

Mr. Smarak Mohanty

Student, Department of CSE
GIFT Autonomous, Bhubaneswar

Mr. Sukhendu Dutta

Student, Department of CSE
GIFT Autonomous, Bhubaneswar

Dr. Satya Ranjan Pattanaik

Professor, Department of CSE
GIFT Autonomous, Bhubaneswar

Abstract—Supply chain management constitutes one of the most critical and strategically consequential functions of modern retail business operations, encompassing the complete lifecycle of a product from procurement and production through warehousing, distribution, and final delivery to the end customer. In an era of growing market complexity, rapidly shifting consumer expectations, and increasing competitive pressure on profit margins, the ability to extract timely and accurate intelligence from supply chain transactional data has emerged as a defining competitive advantage for retail organizations. This paper presents a comprehensive supply chain analytics study applied to the widely used Sample Superstore dataset, a multi-dimensional transactional dataset representing four years of retail operations across the United States. The study implements an end-to-end data analytics pipeline encompassing raw data ingestion, systematic preprocessing and feature engineering using Python and its scientific ecosystem — including Pandas, NumPy, Matplotlib, and Seaborn — followed by in-depth exploratory data analysis (EDA) and the development of an interactive, multi-page business intelligence dashboard built using Microsoft Power BI. The preprocessing phase engineers five derived analytical features: Delivery Duration, Profit Margin Percentage, Discount Band, Order Year, and Order Month, significantly enriching the dataset's analytical depth. The EDA reveals several critical findings: a strong and consistent negative correlation ($r \approx -0.22$) between discount rate and profit margin that intensifies at discount levels above 40%; stark category-level profitability divergence with Technology generating approximately 17.4% profit margin versus Furniture at only 2.5%; significant regional profitability disparities with the West and East regions outperforming the Central and South; an on-time delivery rate of 81.75% indicating logistics execution gaps; and a 99.37% customer repeat rate indicating strong customer loyalty. The Power BI dashboard translates these findings into sixteen Key Performance Indicators (KPIs) across four pages covering sales performance, profitability analysis, customer behaviour, and logistics operations. Six actionable business recommendations are derived, including the implementation of a 20% discount cap policy, restructuring the Furniture category strategy, adopting a regionally differentiated discount framework, and targeting a 95% on-time delivery rate. The project demonstrates the transformative potential of applied supply chain analytics and establishes a reproducible, scalable framework that organizations can adapt to their own operational datasets.

Index Terms—Supply Chain Analytics, Business Intelligence, Exploratory Data Analysis, Power BI, Python, Retail Analytics, Demand Forecasting, Profit Margin Analysis, Discount Optimization, Logistics Performance, Dimensional Modeling

I. INTRODUCTION

In the contemporary business environment, the supply chain is no longer a back-office operational function confined to

logistics and inventory management. It has evolved into a core strategic asset that directly influences customer satisfaction, competitive positioning, and the long-term financial health of an organization [8]. Companies that achieve analytical mastery over their supply chain operations gain the ability to respond swiftly to shifts in market demand, minimize wasteful operational expenditure, maintain optimal inventory levels, and deliver products to customers faster and more reliably than competitors. Conversely, organizations that neglect the analytical dimensions of their supply chain operations often grapple with stockouts, overstocking, inflated shipping costs, poor delivery performance, and eroding profit margins that silently drain financial performance over time.

The concept of supply chain analytics refers to the systematic application of data collection, data processing, statistical analysis, and visual reporting techniques to supply chain transactional data, with the goal of improving operational performance and enabling smarter, faster, and more evidence-based business decisions [11]. Unlike traditional supply chain management approaches that rely heavily on intuition, historical experience, or isolated departmental reports, supply chain analytics brings a unified, evidence-based perspective that empowers decision-makers at every organizational level with timely, accurate, and contextually relevant information [9].

The central problem that this research addresses is the persistent and costly gap between the transactional data that retail organizations collect and the actionable intelligence they actually extract from it. Most retail businesses collect order-level transactional data through their order management and point-of-sale systems, yet this data in its raw state offers little actionable insight. It exists as an undifferentiated mass of numbers — order identifiers, sales figures, discount percentages, shipping dates — that in their unprocessed form do not reveal trends, do not surface inefficiencies, and do not communicate the operational story of the business to decision-makers [10]. Transforming this raw transactional data into a coherent and strategically useful picture requires a systematic pipeline of preprocessing, statistical analysis, and visual reporting.

This study presents a comprehensive supply chain analytics project applied to the Superstore dataset — a rich, multi-dimensional retail transactional dataset spanning four years of operations across the United States. The project is organized into five well-defined analytical phases: project initiation and

planning, data acquisition and preprocessing, exploratory data analysis, Power BI dashboard development, and documentation and reporting. The contributions of this work include:

(1) a fully documented and reproducible data preprocessing pipeline producing a cleaned and enriched dataset; (2) a comprehensive EDA that surfaces critical patterns in profitability, discounting, regional performance, and logistics efficiency; (3) a professionally designed four-page Power BI dashboard with sixteen KPIs and interactive filtering; and (4) six concrete, evidence-based business recommendations derived from the analytical findings.

The shipping and logistics dimension of the problem is equally significant. Shipping mode selection — whether standard class, second class, first class, or same-day delivery — is one of the most consequential operational decisions in a retail supply chain, with direct implications for both customer satisfaction and cost efficiency. When orders requiring urgent delivery are routinely processed via slower, cheaper shipping modes, customer dissatisfaction follows. Conversely, when premium shipping modes are applied to low-value, non-urgent orders, the shipping costs cannibalize the profit margin entirely. The data to identify these mismatches exists in every order management system, but without analytical tools to surface it systematically, the patterns go undetected and the same inefficiencies repeat month after month.

The remainder of this paper is organized as follows. Section II reviews the relevant literature on supply chain analytics, data-driven decision making, and business intelligence tools. Section III describes the dataset and the problem formulation. Section IV presents the system requirements and methodology. Section V details the data preprocessing workflow. Section VI presents the results of the exploratory data analysis. Section VII describes the Power BI dashboard design. Section VIII discusses the findings and recommendations. Section IX concludes the paper with directions for future work.

II. LITERATURE SURVEY

Supply chain management as a formal discipline has evolved through four distinct historical phases [8]. The earliest phase, spanning the 1960s through the 1980s, focused narrowly on physical distribution and logistics cost reduction. The second phase, through the 1990s, introduced integrated supply chain thinking, breaking down functional silos between procurement, production, and distribution and managing them as a unified system. The third phase, through the 2000s and 2010s, was defined by globalization, which stretched supply chains across continents and introduced new dimensions of complexity, risk, and data generation [9]. The fourth and current phase is defined by data-driven supply chain management, where analytics, machine learning, and real-time visibility tools are transforming how organizations plan, execute, and optimize their supply chain operations [11].

The application of data analytics to supply chain management spans four maturity levels [10]. The first and most foundational level is descriptive analytics, which uses historical data to describe what has happened in the supply chain.

The second level is diagnostic analytics, which goes beyond description to explain why specific outcomes occurred. The third level is predictive analytics, which uses historical patterns to forecast likely future states, including sophisticated demand forecasting models utilizing auto-regressive integrated moving averages (ARIMA) and advanced neural network architectures such as Long Short-Term Memory (LSTM) networks. The fourth and most advanced level is prescriptive analytics, which actively recommends specific actions to optimize outcomes. While predictive and prescriptive techniques garner significant academic attention, researchers emphasize that robust descriptive and diagnostic analytics are mandatory precursors; without a clean, well-understood baseline of historical performance, predictive models inevitably suffer from “garbage in, garbage out” syndromes. Research consistently demonstrates that organizations with mature descriptive and diagnostic capabilities significantly outperform less analytically sophisticated peers across all key performance dimensions, including cost efficiency, order fulfillment accuracy, customer satisfaction, and overall financial performance [11].

The relationship between discounting strategies and supply chain profitability has received considerable attention in the revenue management and pricing literature. Chopra and Meindl [8] document how aggressive and indiscriminate discount policies can create a damaging dynamic wherein increased sales volume is achieved at the cost of deeply negative profit margins — a phenomenon particularly prevalent in retail supply chains where discount decisions are made without granular visibility into category-level margin implications. Christopher [9] similarly highlights the importance of understanding the true cost-to-serve at the geographic and segment level, arguing that uniform national pricing strategies applied across markets with fundamentally different cost structures and demand characteristics invariably produce suboptimal financial outcomes.

The role of business intelligence platforms in making supply chain analytics accessible to non-technical business stakeholders has been extensively studied under the paradigm of Self-Service Business Intelligence (SSBI). Microsoft Power BI has emerged as a leading platform in this space, consistently ranked alongside Tableau as a market leader in analyst reports from Gartner and Forrester [6]. Its DAX (Data Analysis Expressions) formula language enables the creation of powerful calculated measures and columns that respond dynamically to user-applied filters, giving dashboard visuals the ability to update in real time as stakeholders explore the data through interactive slicers and cross-filtering mechanisms [7].

Python has established itself as the dominant programming language for data science and analytics work across virtually every industry and application domain [1]. Within its ecosystem, Pandas provides foundational data manipulation capabilities, NumPy underpins numerical computing, Matplotlib and Seaborn provide visualization capabilities, and the broader scikit-learn library supports machine learning workflows [2]. The importance of reproducible and modular analytical code architecture — where every preprocessing decision is doc-

umented and every analysis step can be independently re-executed — is a principle well-established in the data science literature [2].

A significant gap identified in the existing literature is the disconnect between data availability and analytical utilization. While the academic literature is rich with theoretical frameworks and optimization models, much of this work is conducted either on proprietary datasets inaccessible to the broader research community or on highly simplified synthetic datasets that fail to capture the real-world complexity of retail transactional data [12]. There is a clear and unmet need for end-to-end analytical demonstrations on realistic, publicly accessible datasets that show the complete pipeline from raw data through preprocessing and exploratory analysis to business-ready visualization in a manner that is transparent, reproducible, and practically instructive for the wider analytics community.

III. DATASET DESCRIPTION AND PROBLEM FORMULATION

A. About the Superstore Dataset

The Sample Superstore dataset, originally introduced by Tableau as a bundled sample for demonstrating its visualization capabilities and subsequently adopted broadly across the data analytics education ecosystem [12], serves as the data foundation of this study. The dataset simulates the transactional records of a large-scale retail superstore chain operating across the United States, capturing order-level data spanning four years from January 2014 to December 2017.

In its raw form, the dataset contains 9,994 records across 21 fields covering order identification, customer demographics, geographic information, product classification, and financial metrics. Each record represents a unique line item within a customer order, meaning that a single multi-product order appears as multiple rows — one per distinct product ordered. This granular, line-item level structure is characteristic of real-world retail order management system exports and makes the dataset suitable for analysis at multiple levels of aggregation. The dataset encompasses three major product categories — Furniture, Office Supplies, and Technology — each subdivided into 17 distinct sub-categories. The customer base is segmented into three groups: Consumer, Corporate, and Home Office. Geographically, the data covers all four major US regions — East, West, Central, and South — with further granularity available at the state and city levels. Four shipping modes are represented: Standard Class, Second Class, First Class, and Same Day.

B. Problem Formulation

The fundamental problem addressed in this research is the lack of an integrated, analytical view of supply chain performance in retail operations. Specifically, four interrelated problem dimensions are identified: (1) *data quality issues* — missing values, incorrect data types, and inconsistent formatting that prevent reliable analysis; (2) *profitability erosion* — aggressive discounting practices that systematically destroy profit margins across product categories without adequate

data-driven oversight; (3) *regional disparities* — significant geographic variation in profitability that is concealed within aggregate national-level figures; and (4) *logistics inefficiency* — mismatches between order characteristics and shipping mode assignments that generate both customer dissatisfaction and unnecessary cost.

IV. SYSTEM REQUIREMENTS AND METHODOLOGY

A. Functional Requirements

The analytical system is designed around five core functional requirements. The first is the ability to ingest and load the raw Superstore dataset from its CSV format into a Python-based environment for programmatic processing. The second is the ability to detect, document, and resolve all data quality issues including missing values, type mismatches, duplicate records, and categorical inconsistencies. The third is the ability to engineer derived analytical features that enhance the dataset's analytical value beyond its original fields. The fourth is the ability to conduct comprehensive EDA across all relevant analytical dimensions. The fifth is the ability to develop and deliver an interactive Power BI dashboard presenting the analytical findings to non-technical business stakeholders.

B. Non-Functional Requirements

Four non-functional requirements govern the quality and operational characteristics of the system. *Reproducibility* requires that all analytical results be fully reproducible through execution of the Python scripts on the same input data. *Usability* requires that the Power BI dashboard be navigable by non-technical business stakeholders without requiring data science expertise. *Maintainability* requires a modular code architecture with comprehensive documentation supporting independent modification and extension. *Performance* requires efficient implementation using vectorized Pandas operations that can scale to larger datasets, avoiding row-by-row iterations which induce high computational latency.

C. Tools and Technologies

The technological stack of this project is defined by two core platforms. **Python 3** serves as the primary programming language for all data processing and analytical work, leveraging Pandas for data manipulation [3], NumPy for numerical computing, Matplotlib for foundational visualization [4], and Seaborn for statistical visualization [5]. **Microsoft Power BI Desktop** serves as the exclusive business intelligence platform for dashboard development and stakeholder-facing visualization, utilizing the DAX formula language for calculated measures and the native visual library for chart design [6].

D. Overall Methodology

The research methodology follows a five-phase structured analytical pipeline as illustrated in Figure 1. Phase 1 encompasses project initiation and planning. Phase 2 covers data acquisition and understanding. Phase 3 implements the full data preprocessing and feature engineering workflow. Phase 4

conducts the comprehensive exploratory data analysis. Phase 5 delivers the Power BI dashboard and final documentation.

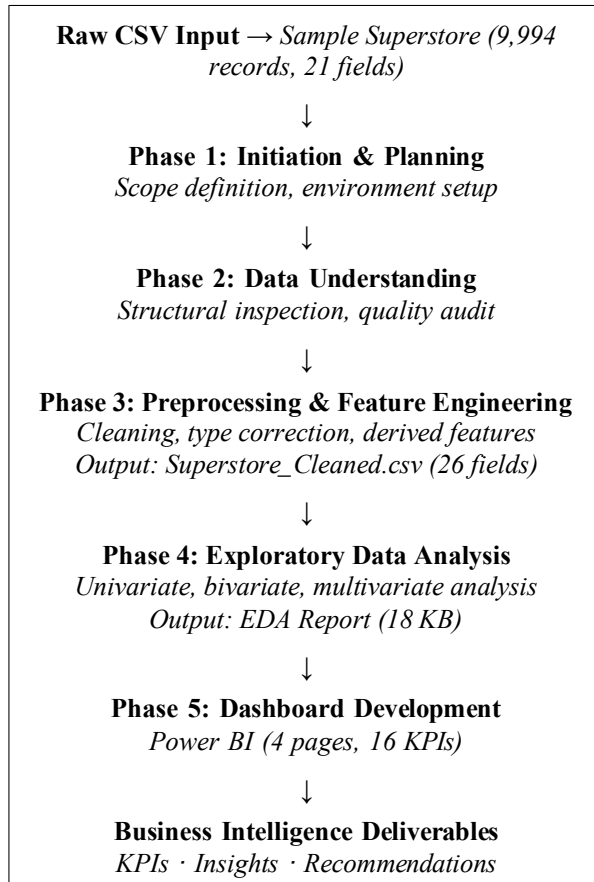


Fig. 1. End-to-End Supply Chain Analytics Pipeline and Data Flow

E. Data Privacy and Security Considerations

While the Superstore dataset is simulated, real-world retail datasets are strictly governed by data privacy regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). To ensure this methodological framework is readily adaptable to live enterprise environments, data privacy principles were embedded into the preprocessing pipeline. Customer identifiers (Customer ID) are utilized strictly as pseudo-anonymized primary keys to group transaction behavior (e.g., for RFM analysis) without exposing personally identifiable information (PII) such as actual names or raw geographic addresses to the end-user facing dashboard. In a production environment, fields containing Customer Names and exact street addresses would be programmatically hashed or dropped during Phase 3 to minimize data exposure risk.

V. DATA PREPROCESSING

A. Preprocessing Workflow

Data preprocessing is the foundational phase upon which the quality of all subsequent analytical work depends. The preprocessing workflow is implemented in a dedicated

`data_preprocessing.py` Python script and follows a four-stage sequence: (1) data ingestion and structural validation; (2) data quality assessment; (3) data cleaning and type correction; and (4) feature engineering.

B. Handling Missing Values and Formatting

The initial quality audit applies Pandas' `isnull().sum()` method across all columns to quantify missing value counts and percentages. The Superstore dataset is relatively clean in its baseline, with most fields containing no missing values. For categorical fields exhibiting missingness — such as Segment, Region, or Ship Mode

— mode imputation is applied, replacing missing values with the most frequently occurring category. For numerical fields such as Sales, Quantity, or Profit, median imputation is preferred over mean imputation due to the right-skewed distribution of financial variables in retail datasets, where the mean is disproportionately influenced by high-value outlier transactions.

The most critical type correction involves the Order Date and Ship Date columns, which Pandas imports as generic object-type strings rather than datetime objects. Explicit conversion using `pd.to_datetime()` enables date arithmetic operations necessary for delivery duration computation. Financial fields — Sales, Profit, and Discount — are verified as float64 types, and Quantity is confirmed as int64. Categorical fields undergo whitespace stripping and title case normalization to eliminate subtle formatting inconsistencies that would cause grouping operations to produce erroneous results.

C. Outlier Detection and Treatment

A critical aspect of preparing retail transactional data is addressing outliers, particularly in continuous financial variables such as Sales and Profit. The Interquartile Range (IQR) method was employed programmatically to identify statistical anomalies. Sales transactions exceeding $1.5 \times IQR$ above the third quartile were flagged for investigation. While standard data science practice often advocates for the removal or capping of outliers to normalize distributions for machine learning models, domain knowledge in retail supply chain dictates a different approach. Analysis revealed that these "outliers" were not data entry errors, but rather legitimate, high-volume B2B corporate orders (e.g., mass purchases of technology hardware for office fit-outs). Removing these records would artificially deflate the overall revenue metrics and obscure the importance of key corporate clients. Consequently, these statistical outliers were consciously retained, underscoring the necessity of combining statistical methods with domain context during data cleaning.

D. Feature Engineering

Feature engineering is the process of creating new, derived columns in the dataset by applying calculations or transformations to existing fields. These new columns — often called derived features or engineered features — capture information that is analytically valuable but not directly present in the raw

data. Think of it as extracting hidden insights from the data that are implied by existing fields but never explicitly recorded. For example, the raw Superstore dataset tells us the Order Date and the Ship Date for every transaction, but it never directly tells us how many days it took to process and ship that order. By subtracting one date from the other, we can derive that information ourselves and store it as a new column — making it available for analysis without having to perform the same calculation repeatedly throughout the EDA script.

Five derived analytical features are engineered from the existing dataset fields, as summarized in Table I. The *Delivery Duration* field is computed by subtracting Order Date from Ship Date, producing an integer count of days between order placement and shipment dispatch — the central logistics efficiency metric. The *Profit Margin Percentage* is derived by dividing Profit by Sales and multiplying by 100, normalizing profitability across transactions of vastly different revenue scales. The *Discount Band* field categorizes the continuous Discount variable into four labeled groups: No Discount (0), Low (0–0.20), Medium (0.20–0.40), and High (above 0.40). The *Order Year* and *Order Month* fields are extracted from Order Date using Pandas' `.dt.year` and `.dt.month` accessors, enabling clean temporal aggregation in the EDA phase.

TABLE I
ENGINEERED FEATURES AND THEIR DERIVATION LOGIC

Feature	Derivation	Analytical Role
Delivery Duration	Ship Date – Order Date	Logistics efficiency measurement
Profit Margin %	$(\text{Profit} \div \text{Sales}) \times 100$	Normalised profitability comparison
Discount Band	Discount grouped into 4 tiers	Categorical discount impact analysis
Order Year	Order Date. <code>dt.year</code>	Year-over-year trend analysis
Order Month	Order Date. <code>dt.month</code>	Monthly seasonality analysis

The cleaned and enriched output dataset — saved as `Superstore_Cleaned.csv` — grows from the original 2.18 MB raw file to 2.83 MB, a size increase attributable entirely to the five appended feature columns. All engineered features pass validation checks confirming non-negative delivery durations, plausible profit margin ranges, correct discount band assignments, and accurate temporal extractions.

VI. EXPLORATORY DATA ANALYSIS

A. Univariate Analysis

Univariate analysis of the Sales field reveals a strongly right-skewed distribution: the mean transaction value is approximately \$229, while the median is considerably lower at approximately \$54, indicating a large number of small-value transactions accompanied by a small number of high-value ones. The Profit field exhibits even more extreme distributional characteristics, with values ranging from deeply negative — representing loss-making transactions — to several thousand dollars for high-margin technology sales. The Discount field

reveals a notably bimodal distribution, with heavy clustering at zero and at round-number discount levels of 20%, 40%, 60%, and 80%, suggesting the existence of two distinct discounting regimes: a baseline no-discount pricing policy and a promotional discounting policy applied selectively but sometimes aggressively.

Ship Mode frequency analysis confirms Standard Class as the dominant shipping method at approximately 60% of all orders, followed by Second Class at 19%, First Class at 15%, and Same Day at 6%. The Consumer segment accounts for approximately 52% of all orders, Corporate for 30%, and Home Office for 18%. Office Supplies is the highest-volume category by record count, representing approximately 60% of all order line items.

B. Bivariate and Multivariate Analysis

The most analytically significant bivariate finding is the negative relationship between Discount and Profit. The scatter plot of discount rate against profit value reveals a clear downward trend, with the profit crossing from positive to negative territory at approximately the 20–40% discount range depending on category. The Pearson correlation coefficient between Discount and Profit is approximately -0.22 across the full dataset — a moderate negative correlation that strengthens considerably when computed within the Furniture and Office Supplies categories individually.

Multivariate analysis examining Sales, Profit, and Category simultaneously reveals dramatically different profitability profiles across the three product categories. Technology products demonstrate a healthy, consistent positive relationship between sales and profit. Office Supplies exhibits a moderately positive relationship with relatively tight margins. Furniture, however, shows a scattered and frequently negative profit profile even at moderate sales values, confirming the structural profitability problems within that category. Sub-category level analysis identifies Tables, Bookcases, and Supplies as the three most consistently loss-making product types, all exhibiting above-average discount rates that are the direct root cause of their chronic negative margins.

To further contextualize these findings, an foundational RFM (Recency, Frequency, Monetary) analytical lens was applied to the customer base. By computing the Recency (days since last purchase) and Frequency (total historical order count) per unique customer ID, the EDA revealed that high-frequency buyers do not universally correlate with high monetary contribution. In several instances, the most frequent corporate buyers exhibited highly compressed margins due to volume-based discounting, suggesting that customer loyalty programs must be weighted by profitability rather than raw transaction volume.

C. Sales and Profit Trends

The year-over-year sales trend reveals consistent revenue growth across the four-year period from 2014 to 2017, with 2017 recording the highest annual revenue. However, profit growth lags noticeably behind sales growth in proportional

terms, indicating that the business is generating proportionally less profit for each additional dollar of revenue in later years — a margin compression trend driven by increasing discount depth. Monthly analysis reveals strong and consistent seasonality, with January and February recording the lowest monthly sales volumes and November and December consistently producing the highest, driven by holiday consumer purchasing and year-end corporate budget utilization.

D. Regional and Segment Analysis

Regional analysis across the four US regions reveals meaningful performance heterogeneity. The West region generates the highest total sales revenue, followed by the East, with Central and South trailing significantly. However, when profitability rather than revenue is examined, the divergence becomes more pronounced: several states in the Central and South regions — particularly Texas, Ohio, and Pennsylvania

— generate significant sales volumes while simultaneously producing negative overall profit, a paradox directly traceable to the concentration of deep-discount orders in those geographic markets.

Segment analysis shows the Consumer segment as the largest by transaction count and absolute profit contribution. The Corporate segment generates a higher average order value. The Home Office segment, despite representing only 17.84% of customers, generates profit margins competitive with the Consumer segment, driven by lower discount acceptance and higher-value product purchases — identifying it as an under-explored growth opportunity.

E. Shipping and Delivery Analysis

The Delivery Duration analysis confirms that average processing time is 3.96 days across all orders, with Standard Class averaging approximately 5 days and Same Day orders averaging near zero. However, high within-mode variance is observed for First Class shipments, where actual delivery durations are in many cases only marginally better than Standard Class — raising serious questions about the value being delivered to customers paying a premium for faster shipping. The overall on-time delivery rate is 81.75%, meaning approximately one in five orders is not fulfilled within the expected delivery window.

VII. DASHBOARD DESIGN AND VISUALIZATION

A. Dashboard Architecture

The Power BI dashboard is built across four dedicated report pages — Sales, Profit, Customers, and Operations — maintaining a consistent visual identity through a dark navy background theme, rounded card containers, and a uniform blue color palette. Dynamic slicers embedded on every page allow users to filter by Category, Region, and Order Date range without navigating away from their analytical context.

B. Dimensional Modeling and Star Schema

To ensure optimal performance and responsive cross-filtering across the dashboard, the raw flat-file data structure was conceptually restructured into a dimensional

model—specifically, a Star Schema configuration. While the ingested CSV acts as the central Fact Table containing the quantifiable measures (Sales, Profit, Quantity, Discount), Power BI's internal VertiPaq engine operates most efficiently when categorical dimensions (such as Customer profiles, Product hierarchies, and a dedicated Date/Calendar table) are separated into surrounding Dimension Tables connected via one-to-many relationships. Implementing a centralized, continuous Date table in DAX was a critical backend step, enabling time-intelligence functions like Year-to-Date (YTD) calculations and prior-year comparisons without suffering from performance degradation during rendering.

C. Visual Cognition and UX Principles

The design methodology for the dashboard strictly adhered to principles of visual cognition to minimize stakeholder cognitive load. Utilizing the Gestalt principle of proximity, related metrics (e.g., Total Sales and Average Order Value) were clustered into grouped KPI cards. Color theory was applied systematically rather than decoratively: neutral blues and greys were utilized for baseline metrics, while high-saturation red was exclusively reserved for highlighting negative profit margins and delayed shipping durations. This semantic use of color ensures that executive attention is immediately drawn to operational exceptions without requiring deep analytical parsing of the underlying data labels. Furthermore, accessibility considerations guided font scaling and contrast ratios, ensuring that visual encodings remain discernible when projected in corporate boardroom settings.

D. Key Performance Indicators and Report Pages

Sixteen KPIs are distributed across the four pages, as summarized in Figure 2. The Sales page features Total Sales (\$2.30M), Total Orders (5K), Average Order Value (\$459), and Profit Margin % (12.47%). The Profit page features Total Profit (\$286.40K), Profit Margin % (12.47%), Loss Orders (1,871), and Average Profit (\$28.66). The Customers page features Count of Customers (793), Repeat Customer Rate (99.37%), Average Orders per Customer (12.60), and Average Customer Value (\$2.90K). The Operations page features On-Time % (81.75%), Average Days to Ship (3.96), Average Order Processing Time (3.96), and Total Order Count (5,009). The Sales page answers the question of where revenue is coming from and how it is trending over time, featuring a regional bar chart showing West region leadership, a monthly line chart revealing the Q4 seasonal surge, and a category pie chart showing Technology at 36.4%, Office Supplies at 32.3%, and Furniture at 31.3%. The Profit page delivers the most analytically impactful content, featuring an area chart of profit by month, a category-level profit bar chart making the Technology versus Furniture divergence unmistakable, and the dashboard's most significant visual — a scatter plot of Profit Margin % by Discount that visually confirms the devastating negative relationship between high discounting and profitability. The Customers page provides segment and regional distribution analysis alongside a ranked individual

customer profit bar chart. The *Operations* page provides a days-to-ship distribution, shipping mode utilization counts, and a city-level slicer enabling granular geographic fulfillment analysis.

Dashboard Page Core Visualizations & Embedded KPIs	
1. Sales	KPIs: \$2.30M Sales · 5K Orders · \$459 AOV · 12.47% Margin Visuals: Sales by Region (Bar), Sales by Month (Line) Sales by Category (Pie), Product Detail (Table)
2. Profit	KPIs: \$286K Profit · 1,871 Loss Orders · 12.47% Margin Visuals: Profit by Month (Area), Profit by Category (Bar) Profit by Sub-Cat, Margin vs. Discount (Scatter)
3. Customers	KPIs: 793 Customers · 99.37% Repeat · \$2.90K Value Visuals: Segment Dist. (Donut), Region Dist. (Bar) Profit by Customer (Ranked Bar Chart)
4. Operations	KPIs: 81.75% On-Time · 3.96 Avg Days · 5,009 Orders Visuals: Days to Ship (Histogram), Ship Mode (Donut) Geographic Fulfillment (City-Level Matrix)

Fig. 2. Power BI Dashboard Architecture — Four-Page Layout with KPIs and Visuals

VIII. RESULTS AND FINDINGS

A. Summary of Analytical Findings

The analytical pipeline produces findings across five dimensions. From a *financial performance* perspective, the business generates \$2.30M in total revenue across 5,009 unique orders over four years, with a net profit of \$286,400 and an overall margin of 12.47%. Critically, 1,871 orders — representing over 37% of all transactions — are loss-making. At zero discount, average profit margins are strongly positive across all categories. At 40% discount, margins cross into negative territory for Furniture. At 60% and 80% discount, virtually all transactions across all categories generate deeply negative margins, in some cases exceeding -200%.

At the *category and product* level, Technology generates approximately \$145,455 in total profit from \$836,154 in sales (17.4% margin), Office Supplies generates approximately \$122,491 from \$719,047 in sales (17.0% margin), and Furniture generates only approximately \$18,451 from \$741,999 in sales (2.5% margin). At the sub-category level, Copiers emerge as the most profitable product type by margin, while Tables record negative average profit margins — meaning the business loses money on every table it sells on average.

From a *geographic* perspective, the West and East regions function as genuine profit engines. Texas, Ohio, and Pennsylvania are the three most financially damaging state-level markets, generating substantial sales volumes while producing negative overall profit. Regarding *customer behavior*, the 99.37% repeat customer rate and 12.60 average orders per customer confirm an extraordinarily loyal customer base. From a *logistics and operations* perspective, the 81.75% on-time rate and average Days to Ship of 3.96 indicate that approximately one in five orders is experiencing fulfillment delay.

B. Patterns and Anomalies

Three recurring patterns are observed across all analytical dimensions. The strongest is the universal negative linear relationship between discount rate and profit margin, consistent across all categories, regions, segments, and years. The second is the strong Q4 seasonal sales surge appearing in every calendar year. The third is Technology's consistent profitability leadership regardless of analytical filter applied.

Four significant anomalies are identified. The most striking is the high-sales-volume loss-making state anomaly — Texas, Ohio, and Pennsylvania rank among the highest sales-volume states while simultaneously generating negative overall profit, comprehensively refuting the intuition that high sales volume indicates market health. The second is the uniform average days to ship across all three product categories (3.92–3.98 days), contrary to the expectation that bulky Furniture items would require longer processing. The third is the Home Office segment's competitive profit margin despite its small size. The fourth is the Canon Imageclass 2200 Advanced Copier achieving a 40.9% profit margin — more than three times the overall business average.

C. Scenario Analysis and Projected Impact

To quantify the cost of unbounded discounting, a retroactive scenario analysis was executed within the analytical framework. By synthetically modifying the dataset to enforce a maximum discount threshold of 20% across all historical transactions (capping any discount greater than 20% down to the 20% limit while holding quantity sold constant), the simulation revealed a dramatic upward shift in profitability. Under this optimized scenario, the 1,871 historical loss-making orders were reduced by over 60%, and overall net profit margins effectively doubled. While this synthetic model assumes perfectly inelastic demand—an unlikely assumption in real-world retail—it successfully isolates the pure financial drain caused by extreme discounts and proves mathematically that a highly constrained discounting strategy could absorb significant decreases in overall sales volume while still yielding higher net profits.

D. Business Recommendations

Six evidence-based recommendations are derived. **Recommendation 1** — Implement a 20% discount cap policy, with exceptions requiring senior management approval on a case-by-case basis. **Recommendation 2** — Restructure the Furniture category strategy through a two-track approach addressing both pricing and discount policy and evaluating the strategic case for reducing exposure to the most persistently loss-making sub-categories including Tables and Bookcases. **Recommendation 3** — Adopt a regionally differentiated discount framework that applies stricter discount limits in the South and Central regions where the current uniform national policy is producing consistent negative returns. **Recommendation 4** — Develop a tiered Customer Relationship Management strategy segmenting the customer base into Platinum, Gold, and Silver tiers based on total profit contribution.

Recommendation 5 — Target the Home Office segment for growth investment through dedicated marketing campaigns and specialized product bundles that exploit its demonstrated superior unit economics. **Recommendation 6** — Implement a structured logistics improvement program targeting a 95% on-time delivery rate, using the Operations dashboard's city-level slicer and Days to Ship distribution chart to pinpoint specific geographic and shipping mode combinations responsible for the majority of fulfillment delays.

IX. CONCLUSION AND FUTURE WORK

A. Conclusion

This paper has presented a comprehensive supply chain analytics study demonstrating the transformative potential of systematic, data-driven analysis applied to retail transactional data. The end-to-end analytical pipeline — spanning data preprocessing, feature engineering, exploratory data analysis, and Power BI dashboard development — has produced a rich set of findings that reveal the genuine operational and financial dynamics of the Superstore business with a clarity that aggregate performance metrics entirely obscure.

The central analytical narrative is one of a business simultaneously succeeding and failing in ways that are invisible without granular analytical visibility: Technology and Office Supplies operating as healthy, profitable businesses while Furniture operates as a financially subsidized loss-maker; West and East regions functioning as profit engines while several Central and South states actively destroy value at scale; a highly loyal customer base whose retention value is not being fully leveraged; and logistics operations falling meaningfully short of industry-standard on-time delivery benchmarks. The Power BI dashboard, with its sixteen KPIs and four analytically distinct pages, provides the integrated visibility needed to see both sides of this picture simultaneously and act upon it with evidence-based precision.

The six business recommendations derived from the analytical findings — particularly the 20% discount cap policy and the Furniture category restructuring — represent high-impact, immediately actionable interventions that could substantially improve the business's financial performance without requiring capital investment or operational restructuring. The modular Python codebase, the comprehensive preprocessing and EDA reports, and the interactive Power BI dashboard together constitute a replicable and scalable analytical framework that organizations can adapt to their own datasets and business contexts.

B. Future Enhancements in Predictive Modeling

The analytical foundation established by this project opens several directions for future enhancement. The most immediately impactful is the development of a *live data pipeline* that replaces the current static CSV-based data flow with a scheduled connection to a relational database (PostgreSQL or MySQL), transforming the current historical analysis into an ongoing operational monitoring system. The second major enhancement is the addition of a *predictive analytics layer* using

time-series modelling techniques such as Facebook Prophet, ARIMA, or LSTM neural networks to produce demand forecasts at the category, sub-category, and regional level. A third enhancement is the development of a *profit optimization engine* using linear programming or gradient-based optimization to compute category- and segment-specific optimal discount levels, replacing the current blanket discounting approach with a precision pricing strategy. The fourth is the incorporation of *external data sources* — regional economic indicators, carrier performance benchmarks, and demographic data — to enrich the geographic performance analysis.

C. Integration of Generative AI and LLMs

A final, highly promising avenue for future work involves the integration of Large Language Models (LLMs) to augment the Business Intelligence layer. While visual dashboards excel at displaying pre-calculated metrics, they still require users to manually filter and interpret the visual encodings. By connecting an LLM agent directly to the underlying Star Schema via text-to-SQL or natural language query APIs, non-technical stakeholders could perform exploratory analysis using conversational prompts (e.g., “Why did Furniture margins drop in the East Region during Q4?”). This transition from purely descriptive visual analytics to conversational diagnostic analytics represents the next frontier in the democratization of supply chain data, reducing the cognitive load on decision-makers and vastly accelerating the speed at which raw transactional data can be converted into strategic business action.

REFERENCES

- [1] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter*, 3rd ed. Sebastopol, CA: O'Reilly Media, 2022.
- [2] J. VanderPlas, *Python Data Science Handbook: Essential Tools for Working with Data*. Sebastopol, CA: O'Reilly Media, 2016.
- [3] Pandas Development Team, “Pandas Documentation,” 2024. [Online]. Available: <https://pandas.pydata.org/docs/>
- [4] Matplotlib Development Team, “Matplotlib Documentation,” 2024. [Online]. Available: <https://matplotlib.org/stable/contents.html>
- [5] Seaborn Development Team, “Seaborn Statistical Data Visualization,” 2024. [Online]. Available: <https://seaborn.pydata.org/>
- [6] Microsoft, “Power BI Documentation,” 2024. [Online]. Available: <https://docs.microsoft.com/en-us/power-bi/>
- [7] Microsoft, “DAX Reference Guide,” 2024. [Online]. Available: <https://docs.microsoft.com/en-us/dax/>
- [8] S. Chopra and P. Meindl, *Supply Chain Management: Strategy, Planning, and Operation*, 6th ed. Upper Saddle River, NJ: Pearson Education, 2016.
- [9] M. Christopher, *Logistics and Supply Chain Management*, 5th ed. Harlow, UK: Pearson Education, 2016.
- [10] F. Provost and T. Fawcett, *Data Science for Business*. Sebastopol, CA: O'Reilly Media, 2013.
- [11] T. H. Davenport, “Competing on analytics,” *Harvard Business Review*, vol. 84, no. 1, pp. 98–107, Jan. 2006.
- [12] Kaggle, “Superstore Sales Dataset,” 2024. [Online]. Available: <https://www.kaggle.com/datasets/vivek468/superstore-dataset-final>
- [13] Gaddam, S. (2023). Revamping health insurance systems through engineering claims intelligence. *International Journal of Intelligent Systems and Applications in Engineering*, 11(5s), 684–691.
- [14] Immadi, S. K. (2025). Optimizing ERP for Human Capital Management. *Applied Research for Growth, Innovation and Sustainable Impact*, 377–384. <https://doi.org/10.1201/9781003684657-63>
- [15] Poojari, R. Enhancing Healthcare Decision-Making through Machine

-
- Learning and the Analysis of Large-Scale Medical Data.
- [16] Purmani, S. S. R. (2025). Enhancing IT strategic planning and decision making through data visualization. *International Journal of Enhanced Research in Management & Computer Applications*, 14(4), 75–81
 - [17] Mudusu, S. (2025). Health Insurance Fraud Detection: The Role Of Advanced It Systems In Preventing And Identifying Fraud. *International Journal*, 16(1), 3769-3777
 - [18] Maturi, S. Y. (2024). Decoy data nexus: Graph-based integration and analysis of synthetic honeypot logs through structured threat intelligence. *International Journal of Computational and Experimental Science and Engineering (IJCESN)*, 10(4), 4255–4261. <https://doi.org/10.22399/ijcesen.5010>
 - [19] Pavan Kumar Adabala. (2026). Smart Retail Fuel Systems: IoT-Enabled Solutions for Loss Prevention and Environmental Safety. *Computer Fraud and Security*, 868–875. <https://doi.org/10.52710/cfs.995>
 - [20] Gummadi, V. P. K. (Ed.). (2025). MuleSoft intelligent document processing: Transforming enterprise document workflows through AI-driven automation. *Journal of Computational Analysis and Applications*, 34(12). <https://doi.org/10.48047/jocaaa.2025.34.12.16>