Development of an End-to-End Data Analytics System for Vendor Performance and Profitability Optimization

Vaibhav Bhagat

Advised by

Prof. Ausif Mahmood

SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIRMENTS FOR THE DEGREE OF MASTER OF SCIENCE
IN COMPUTER SCIENCE

THE SCHOOL OF ENGINEERING UNIVERSITY OF
BRIDGEPORT CONNECTICUT

Abstract:

The recent growth of data analytics capabilities provides businesses with the unique ability to automate and optimize complex operational processes, particularly in supply chain and sales management. For retail and wholesale businesses, effective inventory and sales management are critical. Many companies, however, struggle to integrate and analyse data from disparate sources (e.g., sales, inventory, purchasing) to effectively gauge vendor performance, identify underperforming brands, and manage inventory turnover. This lack of integrated analysis can lead to significant financial losses from inefficient pricing, poor inventory turnover, and over-dependency on key vendors.

To address these challenges, this project details the development of a complete, end-to-end data analytics system designed to analyse vendor performance and optimize profitability. The system is built as a multi-stage pipeline. First, an ETL (Extract, Transform, Load) script built with Python and SQL ingests, cleans, and transforms large volumes of raw data from multiple tables into a centralized SQLite database. A final, pre-aggregated summary table is then generated using optimized SQL queries to support efficient analysis. This clean data is analysed in a Python (Pandas) environment to engineer key features such as Gross Profit, Profit Margin, and Stock Turnover. The analysis answers critical business questions, such as identifying brands needing promotional adjustment and the impact of bulk purchasing on unit cost.

The final findings are delivered through an interactive Power BI dashboard and a formal business report. These tools allow stakeholders to visualize key performance indicators (KPIs) like total unsold capital, top vendor contributions, and low-performing brands, enabling data-driven decision-making and enhancing operational efficiency.

Table of Contents

Chapter 1: Introduction

1.1 Introduction

In the 21st century, data has emerged as one of the most valuable assets for the modern enterprise. This valuation is particularly evident in the retail and wholesale industries, which are predicated on complex supply chains, the management of vast inventories, and engagement with diverse customer bases. The proliferation of digital systems—spanning from point-of-sale (POS) terminals and e-commerce platforms to sophisticated inventory management and vendor-relation systems—has precipitated an exponential increase in the volume of data available to businesses.

This "data explosion," however, presents a significant challenge. An abundance of data does not automatically equate to superior decision-making. Within many organizations, this data remains fragmented, residing in "data silos"—disparate, non-communicating systems. For example, sales data may be archived in one set of files, purchase orders in another, and vendor invoices in a third. Such fragmentation renders it nearly impossible for decision-makers to obtain a holistic, accurate, and timely view of business operations. Consequently, answering seemingly fundamental questions, such as, "Which products are the most *profitable*, as opposed to merely the highest-selling?" or "Are bulk purchasing strategies *actually* yielding cost reductions?" becomes a time-consuming, manual, and error-prone endeavor.

The solution to this problem lies within the field of Data Analytics and Business Intelligence (BI). The core objective of data analytics is the construction of systems and processes that transform raw, chaotic data into clean, structured, and actionable information. A modern, end-to-end data pipeline accomplishes this by creating an automated workflow: ingesting data from its various sources, centralizing and transforming it into a "single source of truth," enabling in-depth analysis, and ultimately presenting the findings in an accessible format for stakeholders.

This project presents the design, development, and implementation of such an end-to-end data analytics pipeline. Utilizing a real-world dataset that represents a company's vendor and sales operations, this project constructs a complete system. The process commences with raw, large-scale CSV files, which are ingested into a SQL database for robust management. It proceeds with complex transformations to create a unified analysis table, conducts in-depth statistical analysis in Python to uncover profound insights, and culminates in an interactive Power BI dashboard designed for business-facing communication and strategic decision-making.

1.2 Statement of the Problem

The project addresses two distinct yet interrelated problems: a primary business problem and a secondary technical-academic problem.

The Business Problem:

The central business challenge is the inability to perform effective vendor performance and profitability analysis stemming from the existence of data silos. The company's data is fragmented across multiple large CSV files (e.g., Purchases, Sales, VendorInvoice, Inventory), each containing millions of records. This fragmentation introduces several critical issues:

1. **Inaccessibility:** The sheer volume of the data, totaling over 2GB, renders it impossible to analyze using standard spreadsheet software such as Microsoft Excel, which possesses row limitations and exhibits poor performance with large datasets.
2. **Lack of Integration:** Owing to the data's separation, cross-functional analysis is not feasible. A manager cannot easily compare the PurchasePrice from the Purchases table against the SalesPrice from the Sales table for a specific brand to ascertain profitability.
3. **Operational Inefficiency:** Business leaders are consequently forced to make decisions predicated on intuition or incomplete estimations rather than on empirical evidence. This can precipitate significant financial losses arising from:
    - **Inefficient Pricing:** A failure to identify brands characterized by high profit margins but low sales volumes, which represent missed promotional opportunities.
    - **Poor Inventory Management:** The unnecessary sequestration of capital in "excess stock" from vendors exhibiting low inventory turnover.
    - **Supply Chain Risk:** An over-dependency on a few key vendors without a clear, quantitative understanding of their contribution to overall profitability.

The Technical-Academic Problem:

Many academic and portfolio projects within the data analytics domain tend to be simplistic "toy" projects that focus on a single tool or technique in isolation. One project might demonstrate Exploratory Data Analysis (EDA) within a Python notebook, another might

illustrate the construction of a Power BI dashboard from a clean, pre-processed file, and a third might solve a standalone SQL challenge.

This single-tool approach is unrealistic and inadequately prepares analysts for real-world business challenges. The most significant challenges in a corporate environment arise not from the use of one tool, but from the *integration* of multiple tools to build a cohesive and automated system. There is a discernible lack of comprehensive, "company-standard" project examples that demonstrate how to connect the entire pipeline: from messy, large-scale raw data (requiring Python for ingestion), to database management and transformation (requiring SQL), to deep statistical analysis (requiring Python and its libraries), and finally to executive-level visualization (requiring Power BI).

This project aims to resolve both issues concurrently by constructing a realistic, multi-tool pipeline to solve a complex, data-siloed business problem.

1.3 Significance of the Study

The development of this end-to-end analytics system holds significant value for both business and technical stakeholders.

For Business Stakeholders:

The primary significance of this work is its provision of a direct blueprint for data-driven decision-making. By transforming a collection of unusable raw files into a dynamic and interactive dashboard, this system empowers a company to:

- **Increase Profitability:** Identify and promote high-margin brands, while re-evaluating pricing strategies for low-margin, high-volume goods.
- **Optimize Cash Flow:** Pinpoint with precision the amount of capital locked in unsold inventory and identify the vendors responsible, thereby allowing for targeted reductions in stock.
- **Strengthen the Supply Chain:** Quantify the financial contribution and reliability of each vendor, which enables stronger negotiations and the development of strategies to mitigate vendor dependency.
- **Improve Marketing ROI:** Target marketing efforts precisely at underperforming brands that demonstrate high potential, rather than utilizing a less-discriminating "shotgun" approach.

For Data Analysts and Academics:

This project serves as a high-fidelity, "company-standard" case study. Its significance lies in its holistic methodology. It demonstrates a complete workflow that is highly sought after by

employers:

- **ETL Pipeline Construction:** It illustrates the use of Python (Pandas, SQLAlchemy) to perform the "Extract" and "Load" portions of an ETL process on large files.
- **Advanced SQL Transformation:** It leverages the power of SQL (Common Table Expressions, Joins, and Aggregations) to execute the "Transform" step, creating a clean, aggregated, and query-ready data mart.
- **Integrated Analysis:** It substantiates that SQL and Python are not competing tools, but are instead complementary. SQL is employed for pre-aggregation and efficient data retrieval, while Python (Pandas, SciPy) is utilized for complex statistical analysis (such as hypothesis testing) that is difficult or impossible to perform in SQL or Power BI alone.
- **Effective Communication:** It completes the analytics lifecycle by channeling all complex findings into a simple, intuitive, and interactive Power BI dashboard, thereby demonstrating the final and most crucial step: communicating insights to non-technical stakeholders.

1.4 Purpose of the Study

The primary purpose of this project is to design, implement, and document a robust, end-to-end data analytics pipeline. This pipeline is engineered to ingest raw, disparate, and large-volume data files and transform them into a unified, clean, and interactive set of business insights.

To guide the development of this pipeline and demonstrate its value, the study will focus on answering a set of specific, high-impact business questions (termed "research questions"). These questions define the analytical goals of the project:

1. **Q1 (Brand Performance):** Which brands exhibit low sales performance but maintain high profit margins, thereby indicating a need for promotional or pricing adjustments?
2. **Q2 (Vendor Contribution):** Who are the top 10 vendors contributing to overall sales revenue and, more significantly, to gross profit?
3. **Q3 (Bulk Purchasing):** Does purchasing in larger volumes (bulk) demonstrably reduce the unit price for a product, and what constitutes the optimal purchase volume for cost savings?
4. **Q4 (Inventory Efficiency):** Which vendors are associated with the lowest inventory turnover, indicating a surplus of excess stock and slow-moving products?
5. **Q5 (Locked Capital):** How much capital is currently sequestered in unsold inventory, and which vendors are the largest contributors to this locked capital?
6. **Q6 (Statistical Validation):** Does a *statistically significant* difference exist between the profit margins of top-performing vendors and low-performing vendors?

By answering these questions, the project will serve as a proof-of-concept for resolving

complex business problems through the systematic integration of multiple data technologies.

1.5 Research Hypothesis

This project is predictive in nature, hypothesizing that a structured analytics process will reveal specific, actionable insights that are not discernible from the raw data.

Hypothesis 1: It is hypothesized that by centralizing the siloed CSV data into a unified SQLite database, complex, multi-table join-and-aggregation queries can be performed. It is predicted these queries will reveal a holistic "vendor summary" that effectively links purchasing costs, sales revenue, and profit margins at the brand and vendor level—a perspective that is impossible to achieve from the separate files.

Hypothesis 2: It is hypothesized that in-depth Python analysis will uncover non-obvious, statistically significant relationships. Specifically, it is predicted that a cohort of vendors/brands will be identified that possess *low sales* but *high profit margins*. It is also hypothesized that a T-test will confirm a significant statistical difference in the profit margins of high-performing versus low-performing vendors, suggesting they operate on different business models (e.g., volume-based vs. premium-based pricing).

Hypothesis 3: It is hypothesized that the final Power BI dashboard, by aggregating all key metrics (e.g., "Total Unsold Capital," "Top 10 Vendors by Profit") into one interactive view, will provide a superior tool for business decision-making compared to static analysis reports. It is predicted this will enable stakeholders to identify optimization opportunities (e.g., which vendor to de-prioritize) "at-a-glance."

Collectively, the central hypothesis is that an *integrated multi-tool pipeline* (SQL + Python + Power BI) will provide exponentially more business value and deeper analytical insights than any project constructed with a single-tool approach.

1.6 Definitions of The Keywords

- **End-to-End Data Pipeline:** A complete, multi-stage process that handles all steps of a data-driven task, commencing from original raw data collection (ingestion) through transformation, analysis, and final presentation of insights (visualization).
- **ETL (Extract, Transform, Load):** A standard data integration framework.
    - **Extract:** The process of reading and retrieving data from its original source systems (e.g., CSV files, APIs, databases).
    - **Transform:** The process of cleaning, validating, aggregating, joining, and structuring the data to prepare it for analysis.
    - **Load:** The process of writing the newly transformed data into a target destination,

such as a database, data warehouse, or data mart.

- **Vendor Performance Analysis:** The process of measuring, analyzing, and managing vendor performance with the objective of reducing costs, mitigating risks, and driving continuous improvement.
- **Inventory Turnover:** A financial ratio indicating how many times a company has sold and replaced its inventory during a specified period. A low turnover rate is indicative of weak sales or excess inventory.
- **Gross Profit Margin:** A profitability metric, expressed as a percentage, that denotes the proportion of revenue remaining after accounting for the Cost of Goods Sold (COGS). It is calculated as: (Total Sales - Total Purchase Cost) / Total Sales.
- **Business Intelligence (BI):** The application of software and services to convert data into actionable insights that inform an organization's strategic and tactical business decisions. BI tools, such as Power BI, are used to create reports and dashboards.
- **Key Performance Indicator (KPI):** A quantifiable measure of performance over time for a specific objective. In this project, "Total Sales," "Total Unsold Capital," and "Profit Margin" function as key KPIs.
- **SQLite:** A C-language library that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine. It is particularly suitable for prototyping and projects, as the entire database is contained within a single file.

1.7 Research Limitations

While this project is designed as a realistic, "company-standard" model, it operates under several limitations that must be acknowledged.

1. **Static Data Source:** The project utilizes a static set of CSV files as its raw data source. The ingestion process is consequently a one-time "batch" operation. In a genuine corporate environment, this data would likely be "streaming" or updated on a daily basis. A production-grade pipeline would necessitate scheduling (e.g., via Airflow or a similar orchestrator) to run incrementally, processing only new or modified data.
2. **Database Technology:** The project employs SQLite due to its simplicity, portability, and ease of setup (the entire database being a single file). A large-scale enterprise solution would typically utilize a more robust, server-based Relational Database Management System (RDBMS) such as PostgreSQL, MySQL, or Microsoft SQL Server, or a cloud data warehouse (e.g., Google BigQuery, Amazon Redshift, Snowflake) designed to manage petabyte-scale data and high-concurrency queries.
3. **Scope of Analysis:** The analysis is intentionally focused on vendor performance and profitability. A comprehensive business analysis would also incorporate other dimensions, such as customer segmentation, time-series forecasting of sales, and geographical (spatial) analysis, all of which are considered outside the defined scope of this specific project.

4. **Generalizability of Findings:** The analytical findings and specific thresholds (e.g., using the 15% quantile to define "low sales" or the 75% quantile for "top performers") are specific to this dataset. These "magic numbers" are not universally applicable and would require re-calibration and validation against the specific business context of any other company.

Chapter 2: Background and Literature Review

2.1 Introduction

This chapter provides a review of the established literature and foundational concepts that underpin this project. The objective of this project is not to invent a new algorithm, but rather to synthesize and apply industry-standard best practices from several key domains in a novel, integrated manner. This review will encompass the evolution of business intelligence, the core technologies in a modern data pipeline (SQL, Python, and Power BI), and the ETL (Extract, Transform, Load) framework that binds them. Finally, it will identify the gap in existing literature that this project endeavors to address.

2.2 The Evolution of Business Intelligence

Business Intelligence (BI) refers to the processes and technologies employed by enterprises to analyze data and present actionable information, thereby assisting executives, managers, and other end-users in making more informed business decisions.

The concept of BI has undergone a significant evolution. In the late 20th century, BI was a highly centralized, IT-led function. Business users submitted requests to the IT department, which subsequently executed manual database queries and utilized tools like Crystal Reports to generate static, text-heavy reports. This process was notoriously slow, often requiring weeks or months, by which time the information was frequently outdated (Gartner, 2013). The primary tools were spreadsheets, notably Microsoft Excel, which, while powerful for small-scale analysis, suffer from data size limitations, a lack of version control, and a high risk of manual error.

The 2000s witnessed the rise of the "data warehouse," championed by thought leaders such as Bill Inmon and Ralph Kimball. A data warehouse is a central repository of integrated data derived from one or more disparate sources. The objective was to establish a "single source of truth" for the entire organization (Kimball & Ross, 2013). This project, in effect, constructs a miniature "data mart" (a subset of a data warehouse) by consolidating the vendor and sales

data into one central SQLite database.

The modern era of BI, which this project embodies, is defined by "self-service analytics." Propelled by tools like Microsoft's Power BI and Salesforce's Tableau, the power of data analysis has been democratized. These platforms permit business users, not exclusively IT specialists, to connect to data sources, perform analyses, and construct their own interactive visualizations. This project utilizes Power BI as the final "self-service" layer, where a stakeholder can interactively filter and explore the data to answer their own questions.

2.3 Core Technologies in the Data Pipeline

This project integrates three distinct but complementary technology stacks.

2.3.1 Data Storage and Management (SQL)

Relational Database Management Systems (RDBMS) have constituted the backbone of business operations for decades. Systems like MySQL, PostgreSQL, and Microsoft SQL Server provide a structured, reliable, and scalable method for storing and retrieving transactional data.

The language used to communicate with these databases is **SQL (Structured Query Language)**. SQL remains the industry standard for managing data. While often perceived as a tool for simple data *retrieval* (e.g., SELECT * FROM table), its true power resides in data *transformation* (Codd, 1970). This project demonstrates the transformative power of SQL by utilizing:

- **Joins (INNER JOIN, LEFT JOIN):** To combine data from multiple tables (e.g., linking a vendor's name from the VendorInvoice table to a purchase in the Purchases table).
- **Aggregations (GROUP BY, SUM, AVG):** To "roll up" millions of individual transaction records into meaningful summaries (e.g., calculating Total_Sales_Dollar for each brand).
- **Common Table Expressions (CTEs):** To decompose highly complex queries into logical, readable, and modular steps, as demonstrated in the creation of the final vendor_sales_summary table.

2.3.2 Data Analysis and Statistics (Python)

While SQL excels at data transformation and aggregation, its analytical and statistical capabilities are limited. It is in this domain that Python has become the "lingua franca" of data science and analysis (McKinney, 2012). The maturity of its open-source libraries makes it the tool of choice for in-depth analysis. This project leverages several key libraries:

- **Pandas:** Built atop NumPy, Pandas provides high-performance, easy-to-use data

structures, most notably the "DataFrame." This project employs Pandas to load the aggregated SQL table into memory, perform final data cleaning (e.g., imputing null values), and conduct feature engineering.
- **Feature Engineering:** This is the process of applying domain knowledge to create new features (columns) from existing data. The creation of Gross_Profit, Profit_Margin, and Stock_Turnover are all examples of feature engineering performed within Pandas.
- **Matplotlib and Seaborn:** These are visualization libraries utilized for Exploratory Data Analysis (EDA). They permit the analyst to gain familiarity with the data by plotting distributions, identifying outliers, and visualizing relationships *prior* to formal analysis.
- **SciPy / Statsmodels:** These libraries elevate the analysis from simple description to statistical inference. Whereas a bar chart can *illustrate* a difference, these libraries can *substantiate* it. This project uses SciPy.stats to perform an independent two-sample T-test, quantifying whether the observed difference in profit margins between vendor groups is statistically significant or merely an artifact of random chance.

2.3.3 Data Visualization (Power BI)

The final stage of the analytics lifecycle is communication. An insight is without value if it is not comprehended by the individuals who can act upon it. Data visualization is the practice of translating complex information into a visual context that is readily understood.

Modern BI platforms like **Power BI** are engineered for this purpose. They transcend the static charts of Excel or Python by facilitating the creation of *interactive dashboards*. A dashboard consolidates multiple KPIs and visualizations into a single screen, furnishing a high-level overview of the business (Few, 2006). This project uses Power BI to:

- **Display KPIs:** Present top-level metrics such as "Total Sales" and "Total Unsold Capital."
- **Interactive Filtering:** Allow a user to select a specific vendor and observe all other charts on the dashboard update instantaneously to reflect that vendor's data.
- **Narrate a Story:** Guide the user's attention from a high-level summary to specific, actionable insights, such as the list of low-sales, high-profit brands.

2.4 The ETL (Extract, Transform, Load) Framework

The process that connects all the aforementioned technologies is ETL (Extract, Transform, Load). ETL constitutes the foundational process of any data warehouse or analytics pipeline.

- **Extract:** This stage involves the retrieval of data from its myriad sources. In this project, the "Extract" step is the Python script responsible for reading the large, raw CSV files from the local disk.
- **Transform:** This represents the most complex and value-adding stage. It encompasses all the work required to convert raw data into clean, analysis-ready data. This project

demonstrates a powerful "two-stage" transformation:

1. **SQL-based Transformation:** Performing large-scale joins and aggregations within the database. This method is highly efficient as it leverages the database's optimized query engine and avoids loading billions of raw rows into memory.
2. **Python-based Transformation:** Performing fine-grained cleaning and feature engineering (e.g., calculating Profit_Margin) in Pandas, which offers greater flexibility and expression than SQL for complex calculations.

- **Load:** This stage involves loading the newly transformed data into its final destination. This project features a "multi-load" process:
  1. The raw data is *loaded* into staging tables within the SQLite database.
  2. The transformed, aggregated data is *loaded* into a new, permanent vendor_sales_summary table in the same database.
  3. This final, clean table is subsequently *loaded* into Power BI, which serves as the final destination for the end-user.

2.5 Gap in Existing Literature and Portfolio Projects

A review of academic literature and online data analytics tutorials reveals a significant gap, which this project directly addresses. The existing literature is heavily segmented by tool. There are innumerable books and courses on "Mastering SQL," "Data Analysis with Python," or "Power BI for Beginners." However, there is a distinct lack of practical, end-to-end case studies that demonstrate how to architect a system *utilizing all of them in conjunction*.

Students and aspiring analysts are often left with the difficult task of determining "what to use when." They might pose questions such as: "Should aggregations be performed in SQL or in Pandas?" or "How is data transferred from a database into Power BI?"

This project provides a clear, opinionated, and realistic answer to these questions. It presents a "company-standard" architecture:

1. Utilize **Python** for ingestion and orchestration.
2. Employ **SQL** for large-scale transformations, joins, and aggregations *as close to the data source as possible*.
3. Utilize **Python (Pandas)** for complex, row-by-row feature engineering and statistical analysis.
4. Employ **Power BI** as the final presentation layer for interactive, self-service insights.

By documenting this complete, integrated pipeline, this project fills a critical gap, moving beyond single-tool tutorials to provide a holistic blueprint for solving real-world data analytics challenges.

Chapter 3: Methodology and Exploratory Data Analysis

3.1 Introduction to Analysis Methodology

The methodology for this project involved a rigorous, multi-stage analytical process applied to the consolidated vendor and sales data. Following the establishment of the unified data mart (as detailed in Chapter 2), the analytical process focused on two primary phases: Exploratory Data Analysis (EDA) to establish data quality and distribution, and the execution of specific research queries and statistical tests to address the core business problems.

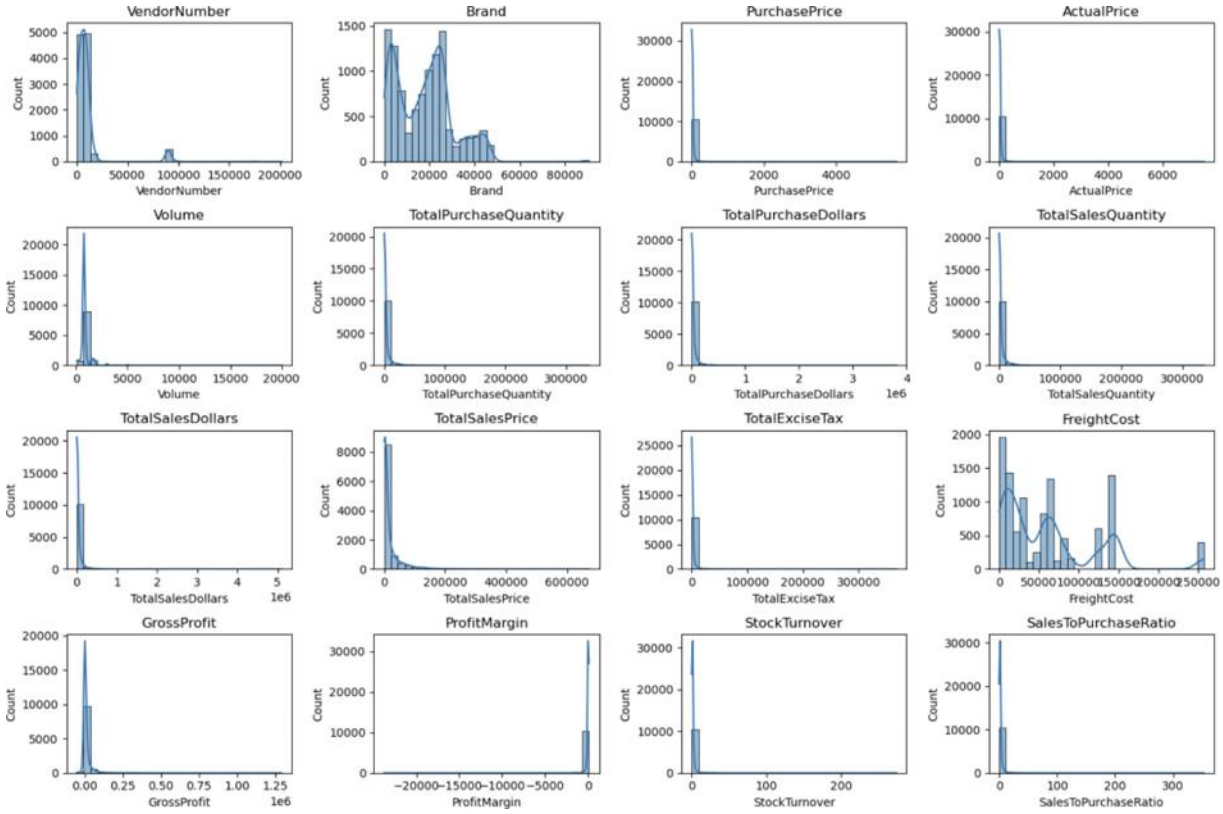
3.2 Exploratory Data Analysis (EDA) Insights

The initial phase of analysis involved a detailed examination of the key financial metrics within the aggregated dataset. This process, conducted using Python, focused on summary statistics, value distribution, and correlation analysis to identify anomalies and establish baseline understanding.

3.2.1 Summary Statistics and Anomaly Identification

A descriptive statistical review of the data revealed several critical anomalies that required mitigation to ensure the reliability of subsequent analyses:

	count	mean	std	min	25%	50%	75%	max
VendorNumber	10692.0	1.065065e+04	18753.519148	2.00	3951.000000	7153.000000	9552.000000	2.013590e+05
Brand	10692.0	1.803923e+04	12662.187074	58.00	5793.500000	18761.500000	25514.250000	9.063100e+04
PurchasePrice	10692.0	2.438530e+01	109.269375	0.36	6.840000	10.455000	19.482500	5.681810e+03
ActualPrice	10692.0	3.564367e+01	148.246016	0.49	10.990000	15.990000	28.990000	7.499990e+03
Volume	10692.0	8.473605e+02	664.309212	50.00	750.000000	750.000000	750.000000	2.000000e+04
TotalPurchaseQuantity	10692.0	3.140887e+03	11095.086769	1.00	36.000000	262.000000	1975.750000	3.376600e+05
TotalPurchaseDollars	10692.0	3.010669e+04	123067.799627	0.71	453.457500	3655.465000	20738.245000	3.811252e+06
TotalSalesQuantity	10692.0	3.077482e+03	10952.851391	0.00	33.000000	261.000000	1929.250000	3.349390e+05
TotalSalesDollars	10692.0	4.223907e+04	167655.265984	0.00	729.220000	5298.045000	28396.915000	5.101920e+06
TotalSalesPrice	10692.0	1.879378e+04	44952.773386	0.00	289.710000	2857.800000	16059.562500	6.728193e+05
TotalExciseTax	10692.0	1.774226e+03	10975.582240	0.00	4.800000	46.570000	418.650000	3.682428e+05
FreightCost	10692.0	6.143376e+04	60938.458032	0.09	14069.870000	50293.620000	79528.990000	2.570321e+05
GrossProfit	10692.0	1.213238e+04	46224.337964	-52002.78	52.920000	1399.640000	8660.200000	1.290668e+06
ProfitMargin	10692.0	-inf	NaN	-inf	13.324515	30.405457	39.956135	9.971666e+01
StockTurnover	10692.0	1.706793e+00	6.020460	0.00	0.807229	0.981529	1.039342	2.745000e+02
SalesToPurchaseRatio	10692.0	2.504390e+00	8.459067	0.00	1.153729	1.436894	1.665449	3.529286e+02



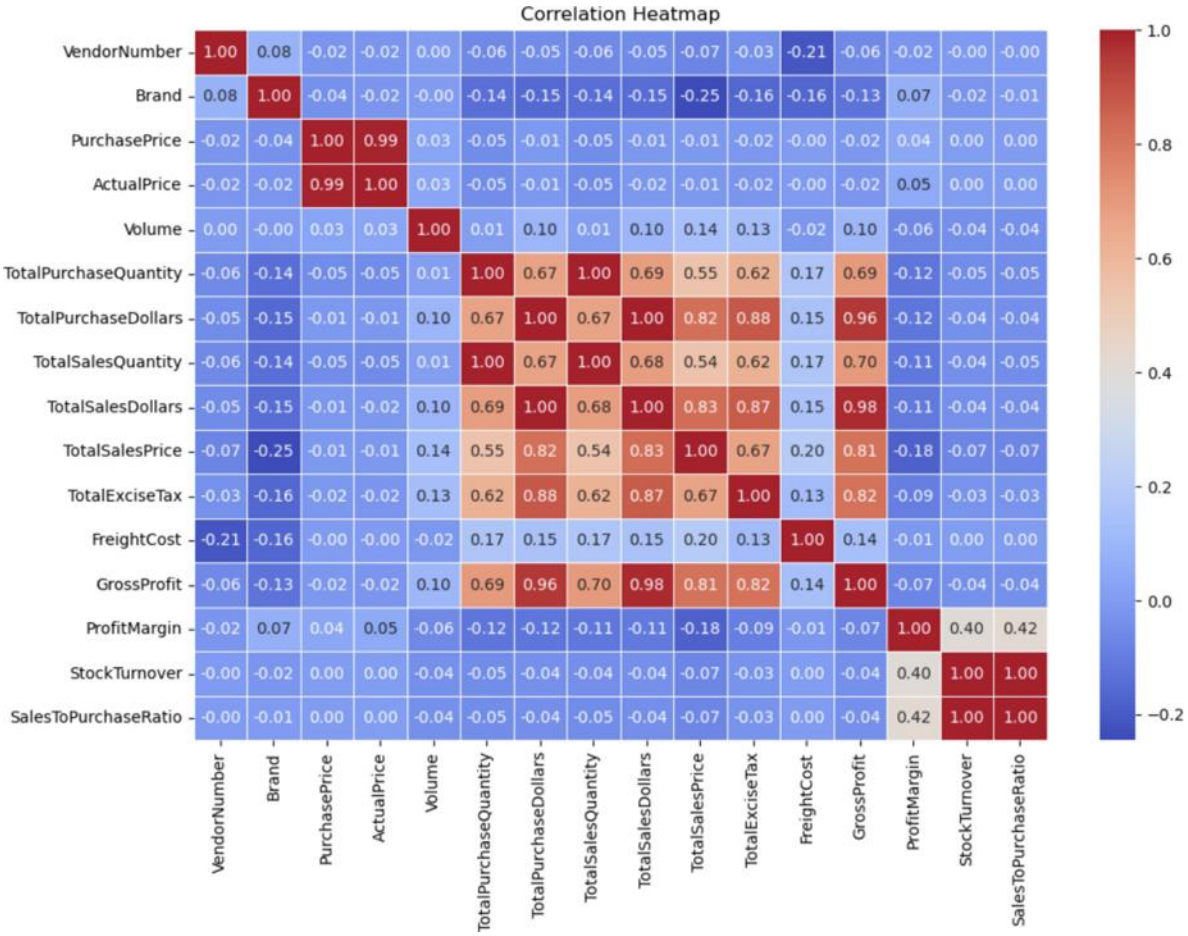
- Negative and Zero Values:
  - **Gross Profit:** A minimum value of -52,002.78 was observed, indicating substantial potential losses due to costs exceeding sales revenue, likely resulting from heavy discounting or errors in pricing/costing models.
  - **Profit Margin:** The presence of a minimum value of negative infinity suggested instances where revenue was zero or less than total cost, leading to extreme negative profit margins due to division by zero or near-zero total sales.
  - **Total Sales Quantity & Sales Dollars:** Numerous products exhibited zero sales, confirming that they were purchased but never sold. This cohort represents slow-moving or obsolete stock and is a primary driver of inventory inefficiency.
- Outliers Detected by High Standard Deviations:
  - **Purchase & Actual Prices:** Maximum values (e.g., 5,681.81 and 7,499.99) were observed to be significantly higher than the respective mean values (24.39 and 35.64). This extreme deviation confirms the presence of premium product offerings within the dataset, justifying a cautious approach to outlier removal.
  - **Freight Cost:** Extreme variation (ranging from 0.09 to 257,032.07) was noted. This wide range suggests complex logistics scenarios, including significant bulk shipments or highly erratic shipping costs across various product types.
  - **Stock Turnover:** The range extended from 0 to 274.5. Values near zero indicate stagnant inventory, while values significantly greater than 1 suggest that current sales are being fulfilled by older stock, affirming the need for inventory assessment.

3.2.2 Data Filtering and Refinement

To enhance the objectivity and reliability of the analytical findings, inconsistent data points were systematically removed or filtered prior to in-depth analysis:

- **Gross Profit < 0:** Transactions resulting in a gross loss were excluded to focus the primary profitability analysis solely on transactions that generated a return.
- **Profit Margin < 0:** Records with non-positive profit margins were eliminated to ensure the calculation of average profitability metrics was based exclusively on gainful transactions.
- **Total Sales Quantity = 0:** Inventory records that were purchased but never successfully sold were filtered out to eliminate non-performing stock from vendor efficiency metrics.

3.2.3 Correlation Insights



A correlation matrix was generated to assess the linear relationships between the critical financial metrics, providing predictive context for the subsequent research questions:

Metric 1	Metric 2	Correlation Coefficient	Insight
Purchase Price	Total Sales Dollars	-0.012 (Weak Negative)	Price variations do not exhibit a statistically significant impact on sales revenue.
Purchase Price	Gross Profit	-0.016 (Weak Negative)	Price variations do not significantly influence gross profitability.
Total Purchase Qty	Total Sales Qty	0.999 (Strong Positive)	Confirms a highly efficient relationship between purchased volume and sales volume (high inventory turnover).
Profit Margin	Total Sales Price	-0.179 (Negative)	Increasing sales prices may be associated with reduced profit margins, possibly due to pressure from competitive market pricing.
Stock Turnover	Gross Profit	-0.038 (Weak	Faster stock turnover does not

		Negative)	necessarily correlate with achieving higher gross profitability.

3.3 Research Questions and Analytical Execution

The project proceeded by executing targeted queries and statistical tests to address the six established research questions.

3.3.1 Q1: Identifying Brands for Promotional or Pricing Adjustments

Brands with Low Sales but High Profit Margins:

	Description	TotalSalesDollars	ProfitMargin
6199	Santa Rita Organic Svgn Bl	9.99	66.466466
2369	Debauchery Pnt Nr	11.58	65.975820
2070	Concannon Glen Ellen Wh Zin	15.95	83.448276
2188	Crown Royal Apple	27.86	89.806174
6237	Sauza Sprklg Wild Berry Marg	27.96	82.153076
...	...	...	...
5074	Nanbu Bijin Southern Beauty	535.68	76.747312
2271	Dad's Hat Rye Whiskey	538.89	81.851584
57	A Bichot Clos Marechaudes	539.94	67.740860
6245	Sbragia Home Ranch Merlot	549.75	66.444748
3326	Goulee Cos d'Estournel 10	558.87	69.434752

198 rows × 3 columns

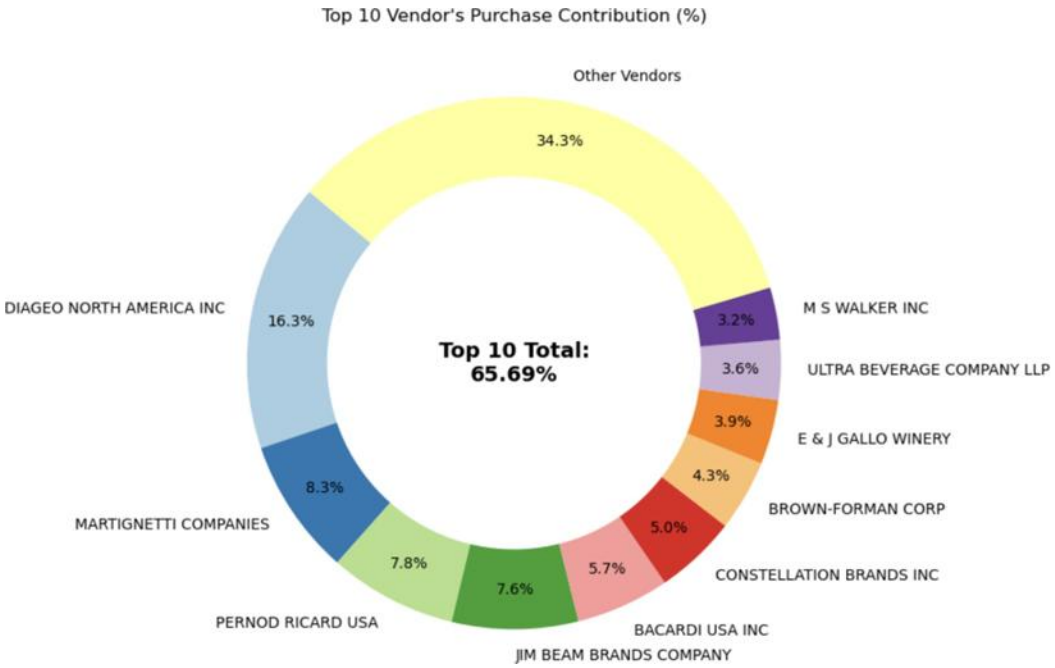
The objective was to identify brands with high-profit potential but suppressed market volume. This was achieved by segmenting brands based on two criteria:

- **Low Sales:** Brands falling below the 15th percentile of Total Sales Dollars.
- **High Profit Margin:** Brands ranking above the 85th percentile of Profit Margin.

Key Finding: 198 brands were identified that exhibit lower sales volume but maintain significantly higher profit margins. These brands present a clear opportunity for targeted marketing, promotional campaigns, or price optimizations aimed at increasing sales volume without compromising underlying profitability.

3.3.2 Q2: Determining Top Vendor Contribution and Q4: Assessing Inventory Efficiency

Analysis was performed to quantify the business reliance on key suppliers and to assess which vendors contributed most to total unsold inventory capital.



Key Finding on Contribution: The top 10 vendors collectively contribute **65.69%** of total purchases, while the remaining vendors account for only **34.31%**. This significant reliance on a small number of suppliers introduces considerable risk concerning potential supply chain disruptions and necessitates a strategy for vendor diversification.

Key Finding on Inventory: The total capital sequestered in slow-moving or unsold inventory (calculated by multiplying unsold quantity by purchase price) amounts to **$2.71M**. Identification of vendors contributing to this low inventory turnover is crucial for optimizing cash flow and reducing holding costs.

3.3.3 Q3: Analyzing the Impact of Bulk Purchasing on Unit Costs

This question investigated whether volume-based purchasing yielded unit cost advantages. Purchasing transactions were segmented into Small, Medium, and Large quantities based on quantile analysis.

OrderSize	UnitPurchasePrice
Small	39.057543
Medium	15.486414
Large	10.777625

Key Finding: Vendors who purchase in large quantities realize a **72% lower unit cost** ($10.78 per unit) compared to the unit costs associated with smaller orders. This finding validates the efficacy of the current bulk pricing strategies in encouraging larger order volumes, thereby increasing total sales while successfully maintaining or enhancing overall profitability.

3.3.4 Q5 & Q6: Investigating Profitability Variance and Statistical Validation

Vendors were categorized into Top-Performing (75th percentile of sales) and Low-Performing (25th percentile of sales) groups to investigate their respective profit margin characteristics.

VendorName	StockTurnover
ALISA CARR BEVERAGES	0.615385
HIGHLAND WINE MERCHANTS LLC	0.708333
PARK STREET IMPORTS LLC	0.751306
Circa Wines	0.755676
Dunn Wine Brokers	0.766022
CENTEUR IMPORTS LLC	0.773953
SMOKY QUARTZ DISTILLERY LLC	0.783835
TAMWORTH DISTILLING	0.797078
THE IMPORTED GRAPE LLC	0.807569
WALPOLE MTN VIEW WINERY	0.820548

VendorName	UnsoldInventoryValue
DIAGEO NORTH AMERICA INC	722.21K
JIM BEAM BRANDS COMPANY	554.67K
PERNOD RICARD USA	470.63K
WILLIAM GRANT & SONS INC	401.96K
E & J GALLO WINERY	228.28K
SAZERAC CO INC	198.44K
BROWN-FORMAN CORP	177.73K
CONSTELLATION BRANDS INC	133.62K
MOET HENNESSY USA INC	126.48K
REMY COINTREAU USA INC	118.60K

Profit Margin Comparison (95% Confidence Interval, CI):

Vendor Group	Mean Profit Margin	95% Confidence Interval (CI)
Top-Performing Vendors	31.17%	(30.74%, 31.61%)
Low-Performing Vendors	41.55%	(40.48%, 42.62%)

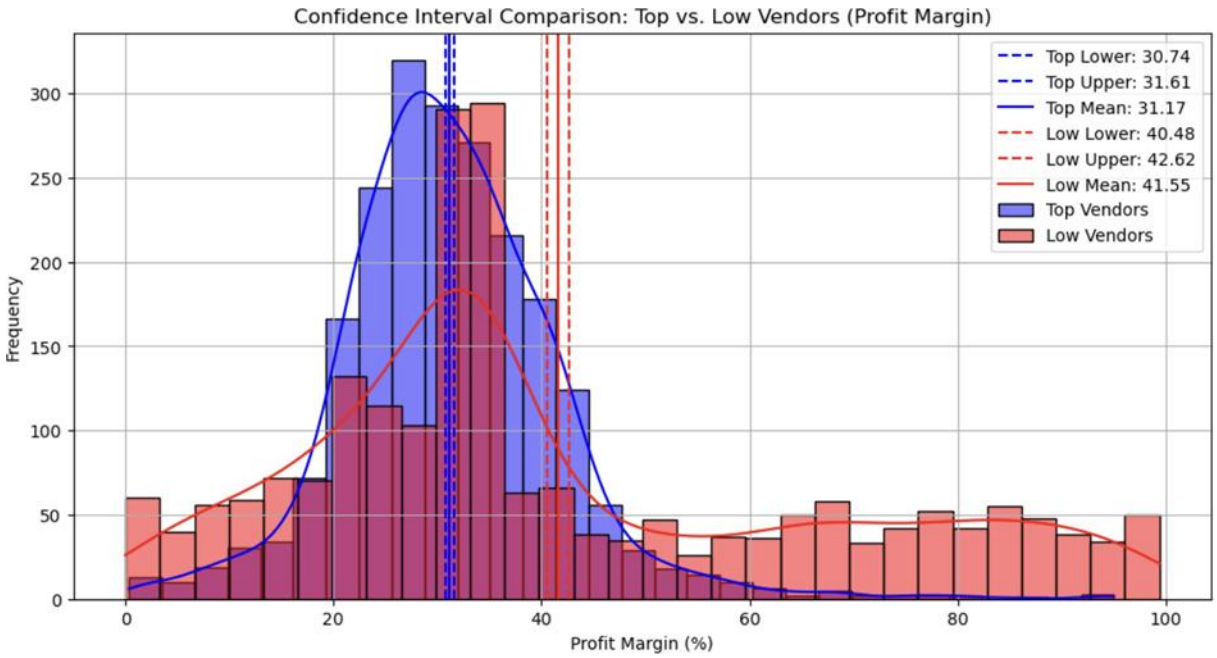
Initial Observation: Low-performing vendors maintain a substantially higher mean profit margin (41.55%) compared to the top-performing, high-volume vendors (31.17%). This

suggests the two groups operate on fundamentally different models, with low performers potentially focusing on premium, low-volume sales, and high performers focusing on high-volume, lower-margin sales.

Hypothesis Testing (T-Test):

To statistically validate this observation, a T-test was performed:

- **Null Hypothesis (H_0):** No significant difference exists in profit margins between the top-performing and low-performing vendors.
- **Alternative Hypothesis (H_1):** A significant difference exists in profit margins between the two vendor groups.



Result: The **Null Hypothesis (H_0) is rejected**. The statistical analysis confirms that the two vendor groups operate under distinctly different profitability models. This mandates separate strategic approaches for each cohort.

Chapter 4: Results and Final Recommendations

4.1 Executive Summary of Key Findings

The comprehensive end-to-end data pipeline successfully transformed fragmented, large-

scale data into actionable business intelligence. The analysis confirmed the presence of significant opportunities for optimization across pricing, inventory, and supply chain management.

Business Metric	Key Finding	Strategic Implication
Vendor Reliance	Top 10 vendors contribute **65.69%** of total purchases.	**Risk Mitigation:** Excessive reliance on a small number of suppliers introduces substantial supply chain risk and limits leverage in future negotiations.
Bulk Purchasing	Bulk orders yield a **72% reduction** in unit cost.	**Optimization:** The current volume discount structure is highly effective and should be maintained or further leveraged to drive additional bulk purchases.
Unsold Capital	Total capital locked in unsold inventory is **$2.71M**.	**Cash Flow:** This capital must be unlocked through targeted sales and reductions in future purchasing from identified high-contributor vendors.
Profitability Variance	Low-Performing Vendors have **+10% higher** profit margins (41.55%) than Top-Performing Vendors (31.17%).	**Strategy Differentiation:** Two distinct vendor strategies are required: one for high-volume efficiency (Top Vendors) and one for premium volume generation (Low Vendors).

Promotional Brands	**198 brands** were identified with low sales but high margins.	**Revenue Opportunity:** These brands are immediate candidates for targeted marketing efforts to boost volume without significant price reduction, converting high-margin potential into realized revenue.

4.2 Actionable Insights and Strategic Recommendations

Based on the quantitative and statistically validated findings, the following strategic and tactical recommendations are proposed to enhance operational efficiency, mitigate risk, and achieve sustainable profitability.

4.2.1 Pricing and Promotion Strategy

- **Re-evaluate Pricing for High-Margin, Low-Sales Brands:** An immediate strategic review should be initiated for the 198 identified brands. The goal is to implement promotional campaigns, targeted marketing, or minimal price adjustments to boost sales volume without sacrificing the inherent high profitability of these products.
- **Top-Performing Vendors: Focus on Cost Efficiency:** Given the statistical validation that these high-volume vendors operate on tighter profit margins (approx. 31.17%), the focus must shift from volume generation to operational cost optimization. Strategies may include optimizing logistics costs, offering bundled promotions, or reducing operational overhead associated with handling high-volume inventory.
- **Low-Performing Vendors: Enhance Market Reach:** For vendors with demonstrably high profit margins (approx. 41.55%) but low sales, the primary challenge is not profitability but market penetration. The recommendation is to enhance marketing efforts, optimize pricing strategies to be more competitive, and improve distribution networks to drive higher sales volumes.

4.2.2 Inventory and Supply Chain Management

- **Diversify Vendor Partnerships:** The current dependency structure, where the Top 10 vendors account for 65.69% of purchases, constitutes a supply chain risk. Management should prioritize building relationships and increasing purchase volume with reliable

secondary suppliers to reduce this dependency and gain greater negotiation leverage.

- **Optimize Slow-Moving Inventory:** Immediate action is required to address the **$2.71M** in locked capital. This includes:
  - **Inventory Reduction:** Adjusting future purchase quantities from the high-contributing vendors.
  - **Clearance:** Launching immediate clearance sales or liquidation strategies for the oldest and slowest-moving stock.
  - **Storage Revision:** Revising storage strategies to minimize holding costs associated with these slow-moving products.
- **Leverage Bulk Purchasing Advantage:** Maintain and reinforce the current bulk pricing strategy. Given the confirmed 72% unit cost reduction achieved via bulk orders, management should explore opportunities to push more vendors and products into the "Large" order size category to maximize savings.

4.2.3 Conclusion

By implementing these data-driven recommendations, the company is strategically positioned to achieve sustainable profitability, effectively mitigate supply chain risks inherent in vendor dependency, and significantly enhance overall operational efficiency and cash flow management.