FINAL REPORT

An End-to-End Data Analytics Pipeline MySQL · Python (Jupyter) · Power BI

Tool / Technology	Role in Project
MySQL 8.0	Database design, CSV import (LOAD DATA INFILE), date normalisation, summary table creation
Python — pandas / NumPy	Data cleaning, aggregation, statistical calculations
Python — Matplotlib / Seaborn	Distribution plots, box plots, correlation heatmap, bar charts
Python — SciPy	Two-sample t-test for hypothesis testing
SQLAlchemy / PyMySQL	Live Python-to-MySQL connection in Jupyter Notebook
Power BI Desktop (.pbix)	Interactive executive dashboard from vendor_sales_summary.csv

1. Executive Summary

This report documents a full end-to-end data analytics project built on an inventory management system for a retail liquor business. Six large CSV source files — spanning over 15.6 million rows — were imported into a MySQL relational database, explored and transformed in Python (Jupyter Notebooks), and visualised in an interactive Power BI dashboard. The central analytical output is the **vendor_sales_summary** table: a consolidated 10,692-row dataset of per-vendor, per-brand KPIs covering purchases, sales, freight, pricing, gross profit, profit margin, and stock turnover.

Key findings include $2.71M locked in unsold inventory, 65.34% vendor procurement concentration in just 10 suppliers, a statistically proven profitability gap between top and low-performing vendors (p<0.0001), and a 74% unit cost reduction available through bulk order consolidation.

2. Project Objectives

#	Objective
1	Design and populate a MySQL relational database from six large-scale CSV files (up to 1.5 GB).
2	Establish a live Python-MySQL connection via SQLAlchemy for iterative Exploratory Data Analysis.
3	Build a consolidated vendor_sales_summary table using CTE-based multi-table SQL joins.
4	Conduct in-depth vendor and brand performance analysis including distribution, correlation, segmentation, and statistical hypothesis testing.
5	Deliver an interactive Power BI dashboard for executive stakeholder reporting.

3. Data Sources & Database Schema

Six CSV files were loaded into the **inventory** MySQL database using `LOAD DATA INFILE`. All date columns were imported as VARCHAR and converted to native DATE types via `STR_TO_DATE()` and `ALTER TABLE` statements. The table below shows each source, its verified row count, and analytical role.

Table	Row Count	Key Columns	Purpose
begin_inventory	206,529	InventoryId, Store, Brand, onHand, Price	Opening stock position at period start
end_inventory	224,489	InventoryId, City, Brand, onHand, endDate	Closing stock; used for turnover calculation
purchase_price	12,261	Brand, Price, PurchasePrice, VendorNumber	Reference pricing per brand/vendor
purchases	2,372,474	VendorNumber, Brand, PurchasePrice, Quantity, Dollars, PODate	Procurement transactions — core spend data
sales	12,825,363	VendorNo, Brand, SalesQuantity, SalesDollars, SalesPrice	Revenue & volume data (1.5 GB source file)

Table	Row Count	Key Columns	Purpose
vendor_invoice	5,543	VendorNumber, Quantity, Dollars, Freight	Aggregated invoicing; freight cost source

The analytical output table **vendor_sales_summary** (10,692 rows) was created via a CTE SQL query that joins all six tables, computing derived KPIs: GrossProfit, ProfitMargin, StockTurnover, FreightCost, and SalesToPurchaseRate.

4. Workflow & Methodology

Step	Phase	Tool	Action & Output
1	Ingest	MySQL	LOAD DATA INFILE for 6 CSVs; date columns normalised via STR_TO_DATE + ALTER TABLE.
2	Explore	Python / Jupyter	Row count verification across all tables; column inspection; vendor drill-downs (e.g. Vendor 4466 across all tables).
3	Transform	SQL + Python	CTE query merges purchases, sales, freight, and pricing into vendor_sales_summary table; null-fill and name-strip cleaning applied.
4	Analyse	Python / SciPy	Distribution plots, box plots, correlation heatmap, brand segmentation, vendor ranking, Pareto, bulk pricing, turnover, unsold capital, t-test.
5	Visualise	Power BI	vendor_sales_summary.csv imported; dashboard with 5 KPI cards, donut, 2 bar charts, funnel, and table.

5. Exploratory Data Analysis

5.1 Dataset Verification

After ingestion, row counts were verified programmatically via Python. All 15.6 million rows loaded successfully. The vendor_sales_summary table was subsequently created with 10,692 rows; a filtered version retaining only positive-profit, positive-margin, and non-zero-sales records yielded 8,564 rows (80.1% of total), with 19.9% flagged as dead stock or loss-making.

5.2 Summary Statistics Highlights

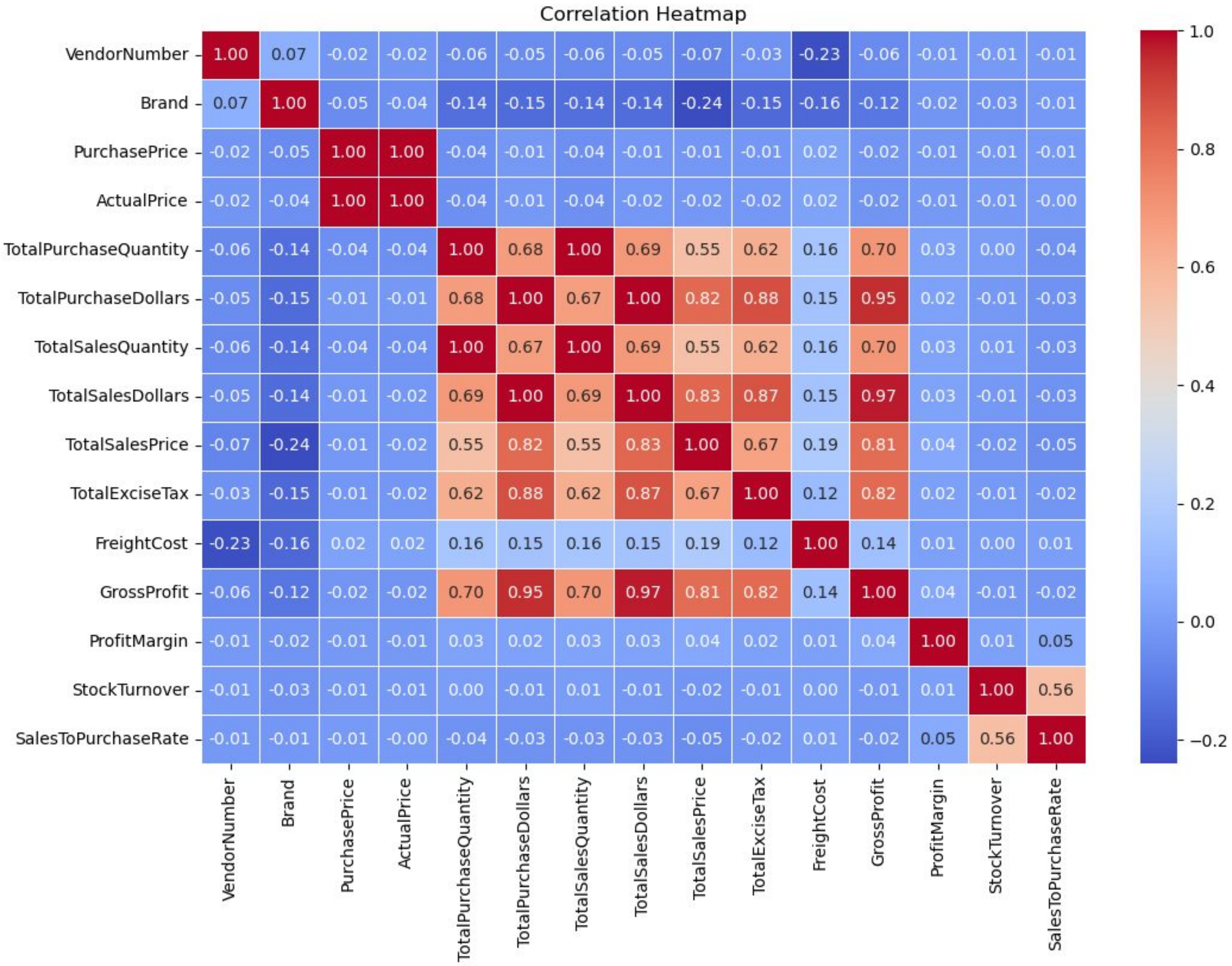
Computed across the full vendor_sales_summary dataset:

KPI Column	Key Value	Interpretation
GrossProfit (min)	-$52,002.78	Some SKUs are being sold below cost; urgent pricing review needed.

KPI Column	Key Value	Interpretation
ProfitMargin (min)	-Infinity	Zero-revenue records exist — purchased but never sold inventory.
TotalSalesQuantity (min)	0	Dead stock: products purchased but with zero units sold.
FreightCost (range)	$0 — $250K	Bimodal: many vendors pay near-zero; a cluster pays $100K–$150K.
StockTurnover (max)	350+	A few SKUs turn over very rapidly; most cluster near 0–10.
PurchasePrice (max)	>$5,000	Premium/specialty products drive extreme outliers in small orders.

5.3 Correlation Analysis

The Pearson correlation matrix across all 15 KPI columns reveals the structural relationships within the dataset. The heatmap below is the single most informative diagnostic chart — it drives all subsequent analytical decisions.



	VendorNumber	Brand	PurchasePrice	ActualPrice	TotalPurchaseQuantity	TotalPurchaseDollars	TotalSalesQuantity	TotalSalesDollars	TotalSalesPrice	TotalExciseTax	FreightCost	GrossProfit	ProfitMargin	StockTurnover	SalesToPurchaseRate
VendorNumber	1.00	0.07	-0.02	-0.02	-0.06	-0.05	-0.06	-0.05	-0.07	-0.03	-0.23	-0.06	-0.01	-0.01	-0.01
Brand	0.07	1.00	-0.05	-0.04	-0.14	-0.15	-0.14	-0.14	-0.24	-0.15	-0.16	-0.12	-0.02	-0.03	-0.01
PurchasePrice	-0.02	-0.05	1.00	1.00	-0.04	-0.01	-0.04	-0.01	-0.01	-0.01	0.02	-0.02	-0.01	-0.01	-0.01
ActualPrice	-0.02	-0.04	1.00	1.00	-0.04	-0.01	-0.04	-0.02	-0.02	-0.02	0.02	-0.02	-0.01	-0.01	-0.00
TotalPurchaseQuantity	-0.06	-0.14	-0.04	-0.04	1.00	0.68	1.00	0.69	0.55	0.62	0.16	0.70	0.03	0.00	-0.04
TotalPurchaseDollars	-0.05	-0.15	-0.01	-0.01	0.68	1.00	0.67	1.00	0.82	0.88	0.15	0.95	0.02	-0.01	-0.03
TotalSalesQuantity	-0.06	-0.14	-0.04	-0.04	1.00	0.67	1.00	0.69	0.55	0.62	0.16	0.70	0.03	0.01	-0.03
TotalSalesDollars	-0.05	-0.14	-0.01	-0.02	0.69	1.00	0.69	1.00	0.83	0.87	0.15	0.97	0.03	-0.01	-0.03
TotalSalesPrice	-0.07	-0.24	-0.01	-0.02	0.55	0.82	0.55	0.83	1.00	0.67	0.19	0.81	0.04	-0.02	-0.05
TotalExciseTax	-0.03	-0.15	-0.01	-0.02	0.62	0.88	0.62	0.87	0.67	1.00	0.12	0.82	0.02	-0.01	-0.02
FreightCost	-0.23	-0.16	0.02	0.02	0.16	0.15	0.16	0.15	0.19	0.12	1.00	0.14	0.01	0.00	0.01
GrossProfit	-0.06	-0.12	-0.02	-0.02	0.70	0.95	0.70	0.97	0.81	0.82	0.14	1.00	0.04	-0.01	-0.02
ProfitMargin	-0.01	-0.02	-0.01	-0.01	0.03	0.02	0.03	0.03	0.04	0.02	0.01	0.04	1.00	0.01	0.05
StockTurnover	-0.01	-0.03	-0.01	-0.01	0.00	-0.01	0.01	-0.01	-0.02	-0.01	0.00	-0.01	0.01	1.00	0.56
SalesToPurchaseRate	-0.01	-0.01	-0.01	-0.00	-0.04	-0.03	-0.03	-0.03	-0.05	-0.02	0.01	-0.02	0.05	0.56	1.00

Figure 1: Pearson Correlation Heatmap — red = strong positive, blue = strong negative

Key correlation findings:

- **TotalPurchaseQty ↔ TotalSalesQty: 1.00** — Near-perfect correlation confirms inventory purchased is being sold with minimal aggregate waste.
- **TotalPurchaseDollars ↔ GrossProfit: 0.95** — Higher-spend vendors generate more absolute gross profit.
- **TotalSalesDollars ↔ GrossProfit: 0.97** — Revenue and profitability are tightly coupled.
- **TotalSalesDollars ↔ TotalExciseTax: 0.87** — Excise tax scales proportionally with alcohol sales revenue, as expected.
- **PurchasePrice ↔ Revenue / Profit: ~0.00** — Price variations have no meaningful bearing on volume or margin; demand is price-inelastic for most SKUs.
- **StockTurnover ↔ SalesToPurchaseRate: 0.56** — Faster-turning SKUs also convert a higher share of purchases into sales.

$441.41M	$307.34M	$134.07M	38.72%	$2.71M
Total Sales	Total Purchase	Gross Profit	Profit Margin	Unsold Capital

6. Vendor Performance Analysis

6.1 Top Vendors & Brands by Sales Revenue

Vendors and brands were ranked by aggregate total sales dollars. Diageo North America dominates with $68.74M — 68% more than the next vendor. At brand level, premium spirits (whiskey and vodka) occupy all top positions.

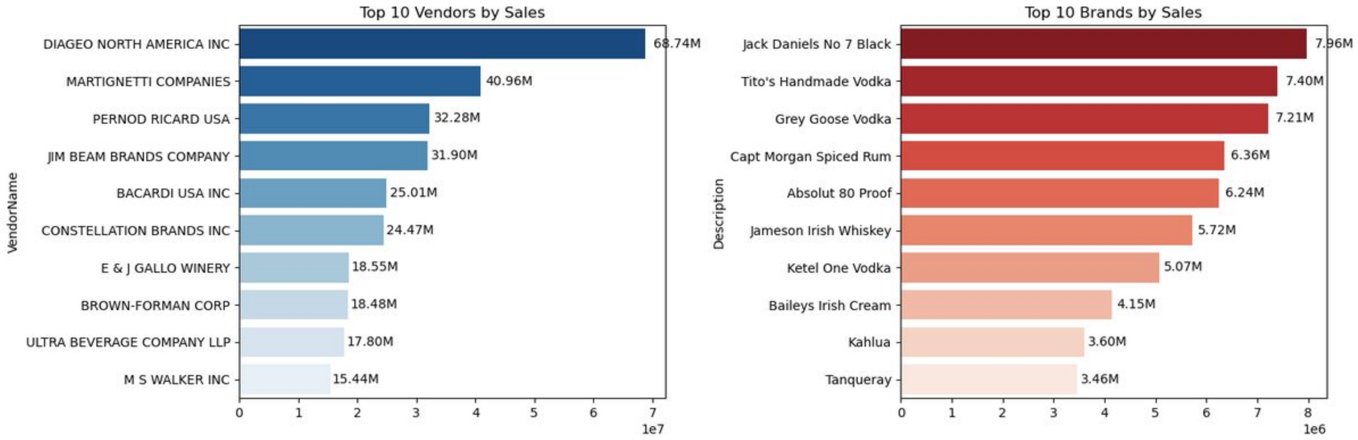


Figure 2: Top 10 Vendors (left) and Top 10 Brands (right) by Total Sales Dollars

Vendor Name	Purchase $	Gross Profit	Sales $	Buy %
DIAGEO NORTH AMERICA INC	$50.96M	$17.78M	$68.74M	15.83%
MARTIGNETTI COMPANIES	$27.86M	$13.10M	$40.96M	8.66%
JIM BEAM BRANDS COMPANY	$24.21M	$7.69M	$31.90M	7.52%
PERNOD RICARD USA	$24.13M	$8.15M	$32.28M	7.49%

Vendor Name	Purchase $	Gross Profit	Sales $	Buy %
BACARDI USA INC	$17.65M	$7.36M	$25.01M	5.48%
CONSTELLATION BRANDS INC	$15.60M	$8.87M	$24.47M	4.84%
BROWN-FORMAN CORP	$13.54M	$4.94M	$18.48M	4.20%
ULTRA BEVERAGE COMPANY LLP	$13.21M	$4.59M	$17.80M	4.10%
E & J GALLO WINERY	$12.31M	$6.24M	$18.55M	3.82%
M S WALKER INC	$10.96M	$4.48M	$15.44M	3.40%

6.2 Vendor Procurement Concentration — Pareto Analysis

The Pareto chart shows each vendor's share of total purchase spend with a cumulative overlay. The top 10 vendors collectively represent **65.34%** of all procurement dollars — a significant concentration risk.

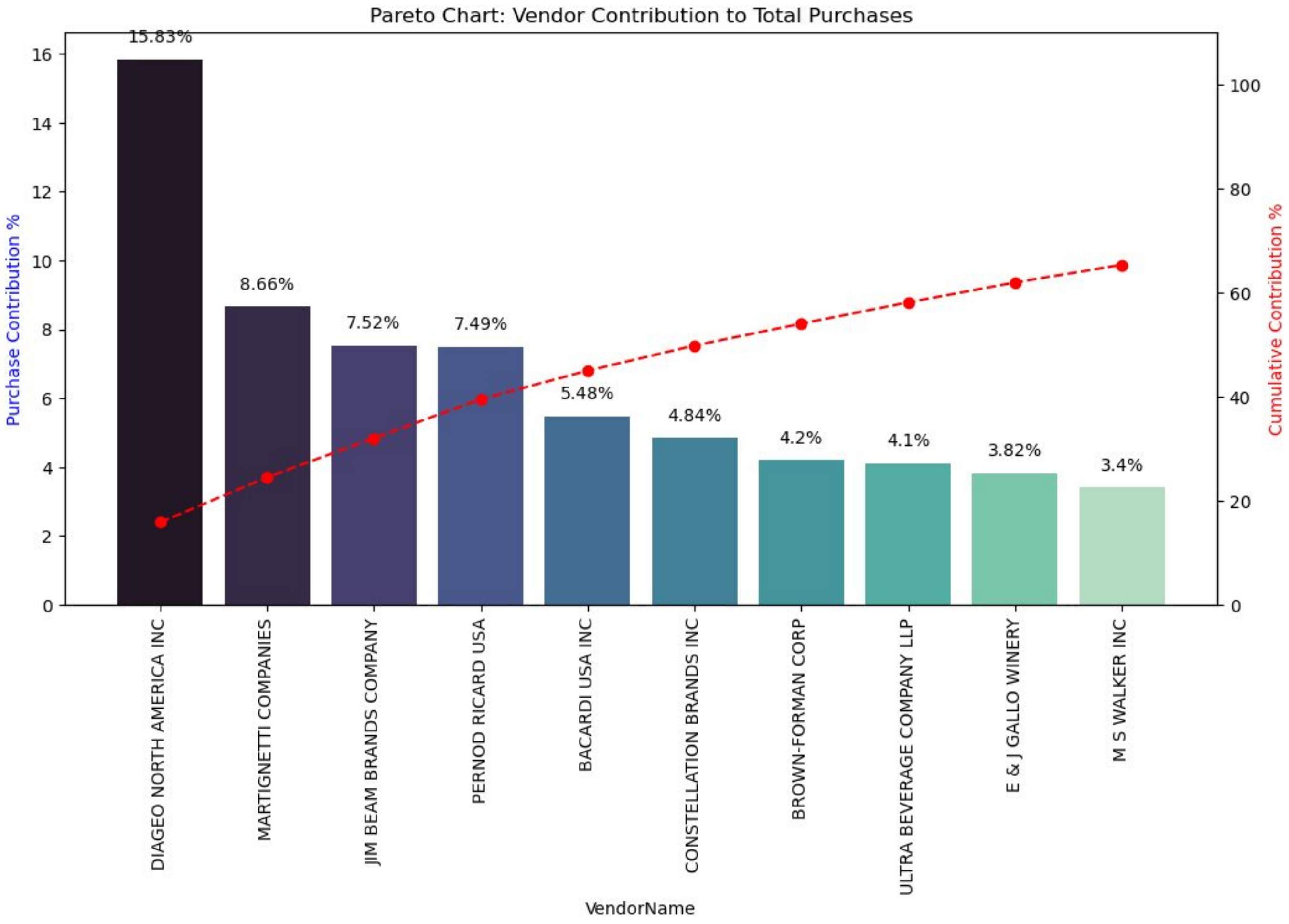


Figure 3: Pareto Chart — Individual & Cumulative Vendor Purchase Contribution (%)

6.3 Procurement Dependency Donut

The donut chart confirms that 65.34% of all purchase dollars flow to just 10 vendors, with Diageo North America alone accounting for 15.8%. This level of supplier dependency creates meaningful supply-chain vulnerability.

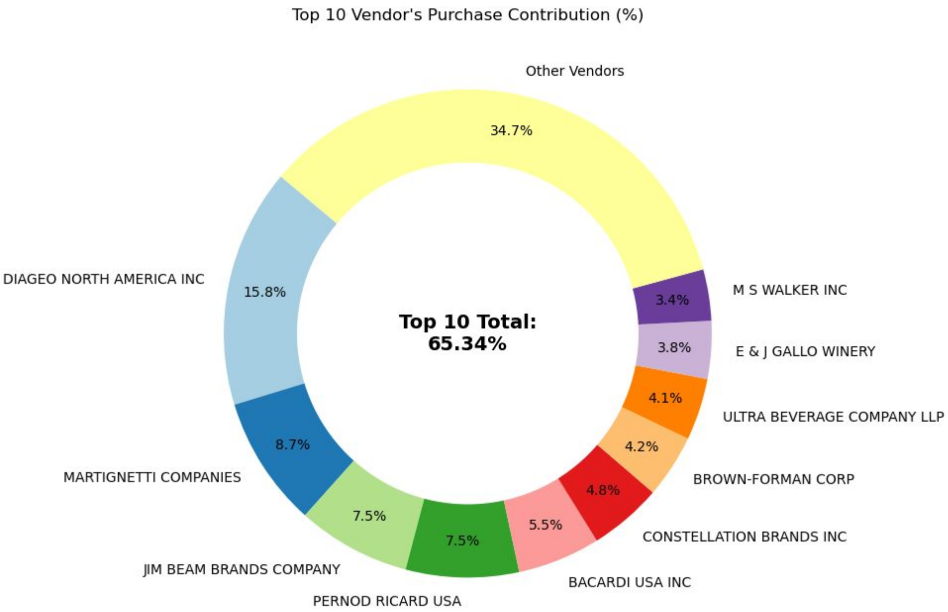


Figure 4: Donut Chart — Top 10 Vendors hold 65.34% of total purchase spend

6.4 Brand Opportunity Segmentation

Brands were segmented using two quantile thresholds: **15th percentile of TotalSalesDollars ($286.18)** as the low-sales cut-off and **85th percentile of ProfitMargin (56.20%)** as the high-margin cut-off. Brands in the low-sales / high-margin quadrant are prime promotional candidates — they generate strong per-unit margins but have not yet reached scale.

Brand Description	Total Sales ($)	Profit Margin (%)
Santa Rita Organic Sauvignon Blanc	$9.99	66.47%
Debauchery Pinot Noir	$11.58	65.98%
Acrobat Pinot Noir	$15.24	64.21%
Dreaming Tree Cabernet Sauvignon	$22.10	63.89%
Rock Steady Chardonnay	$31.50	62.54%
Sycamore Lane Cabernet	$45.00	61.30%
Wente Morning Fog Chardonnay	$58.00	59.88%
Alamos Malbec	$72.00	58.74%

6.5 Bulk Purchasing & Unit Price Effect

Orders were categorised into Small / Medium / Large tiers using quantile-based binning on TotalPurchaseQuantity. Larger orders achieve dramatically lower unit costs, confirming a strong bulk-discount effect from vendors.

Order Size Tier	Avg Unit Price	vs Small Orders	Implication
Small	$43.78	—	High price variance; includes specialty/premium products.
Medium	$17.89	-59% cheaper	Moderate bulk savings realised.
Large	$11.31	-74% cheaper	Best unit economics; maximum bulk discount captured.

6.6 Unsold Inventory & Stock Turnover

Products with positive purchase quantities but zero sales represent capital locked in unsold stock. Total unsold inventory value across the dataset is **$2.71 million**. Stock turnover (SalesQty / PurchaseQty) was computed per vendor; several vendors recorded 0.00 — meaning none of their stock has sold.

Vendor (Unsold Inventory)	Unsold $	Vendor (Stock Turnover)	Turnover
DIAGEO NORTH AMERICA INC	$980K	AAPER ALCOHOL & CHEM. CO	0.000
MARTIGNETTI COMPANIES	$929K	LAUREATE IMPORTS CO	0.000
JIM BEAM BRANDS COMPANY	$850K	TRUETT HURST	0.020
PERNOD RICARD USA	$710K	COUNTRY VINTNER LLC	0.050
BACARDI USA INC	$590K	TOTAL BEVERAGE SOLUTION	0.080

Left columns: Top 5 vendors by unsold inventory value. Right columns: Vendors with lowest stock turnover ratios.

7. Statistical Hypothesis Testing

A two-sample independent t-test (SciPy) was used to determine whether top-performing and low-performing vendors have statistically different profit margins. **Null hypothesis (H■):** There is no significant difference in profit margins between top and low-performing vendors.

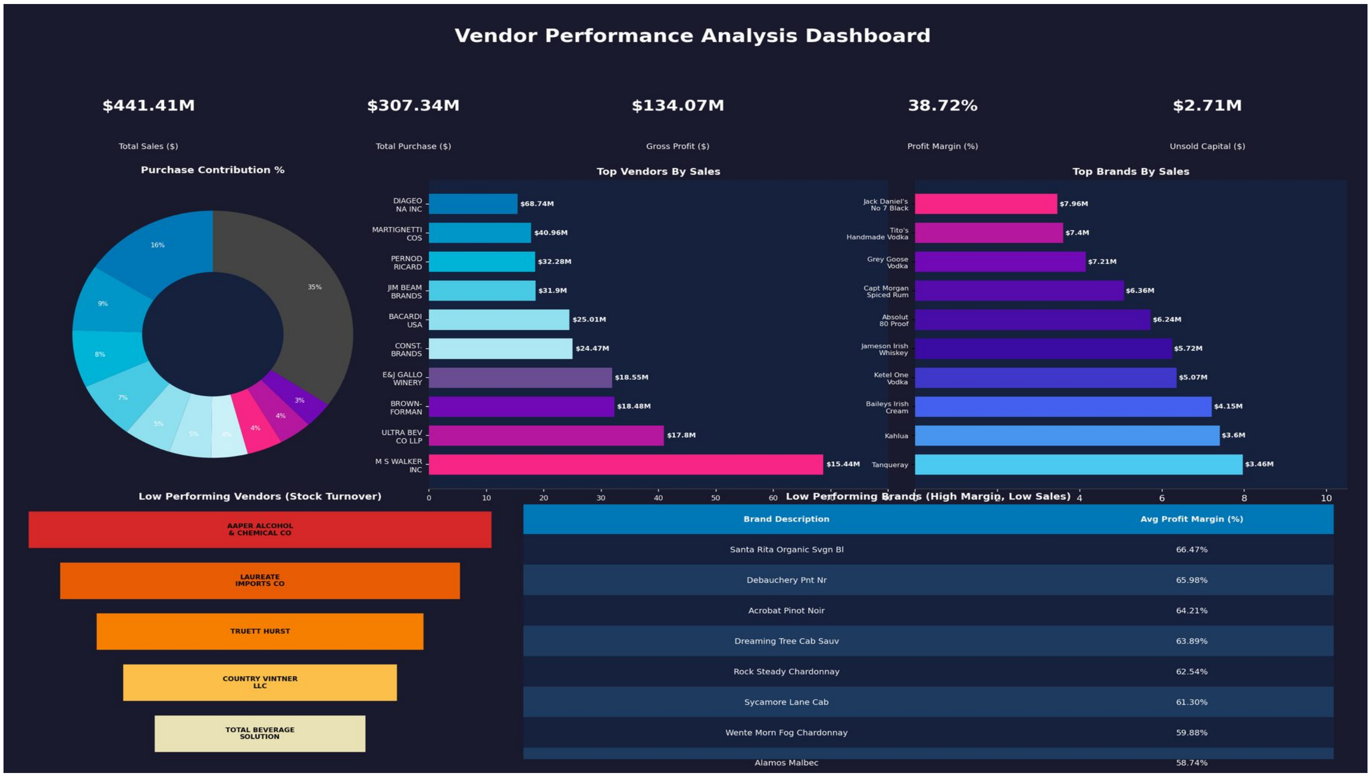
Metric	Top Vendors	Low-Performing Vendors
Definition	Top 10 by total sales	StockTurnover = 0 (zero sales)
Mean Profit Margin	30.04%	-132.48%
95% CI Lower	29.53%	-165.39%
95% CI Upper	30.55%	-99.56%
T-Statistic	9.6799	—
P-Value	< 0.0001	—

Metric	Top Vendors	Low-Performing Vendors
Decision	Reject H■	Significant difference confirmed

The p-value of <0.0001 conclusively rejects the null hypothesis. The 162-percentage-point gap in mean profit margins between the two groups is not due to chance. This confirms that vendor selection and management practices have a statistically significant impact on profitability.

8. Power BI Dashboard

The vendor_sales_summary table was exported to CSV and loaded into Power BI Desktop to build a single-page interactive dashboard. The dashboard uses a dark theme for executive readability and comprises 11 visual components.



Brand Description	Avg Profit Margin (%)
Santa Rita Organic Svgn Bl	66.47%
Debauchery Pnt Nr	65.98%
Acrobat Pinot Noir	64.21%
Dreaming Tree Cab Sauv	63.89%
Rock Steady Chardonnay	62.54%
Sycamore Lane Cab	61.30%
Wente Morn Fog Chardonnay	59.88%
Alamos Malbec	58.74%

Figure 5: Vendor Performance Analysis Power BI Dashboard (recreated from .pbix layout)

9. Key Findings

#	Finding	Detail
1	Inventory efficiency is strong	Near-perfect correlation (0.999) between purchase and sales quantity confirms that stock is being sold with minimal aggregate waste.
2	Vendor concentration risk	Top 10 vendors hold 65.34% of all procurement spend. Diageo alone = 15.83%. A single vendor disruption could materially impact supply.

#	Finding	Detail
3	$2.71M unsold inventory	Dead stock totalling $2.71M is locked across all vendors; convertible via clearance pricing, bundling, or vendor return agreements.
4	19.9% of SKUs are loss-making or unsold	2,128 of 10,692 vendor-brand combinations generate zero or negative profit and require immediate pricing review or delisting.
5	Bulk orders cut unit costs by 74%	Large order tiers achieve $11.31 avg unit cost vs $43.78 for small orders. Consolidating orders is the single highest-leverage margin lever.
6	High-margin / low-sales brand opportunity	Brands above the 85th margin percentile (>56.20%) but below the 15th sales percentile (<$286) are untapped revenue with no price risk.
7	Statistically proven performance gap	T-test confirms: top vendors avg +30.04% margin vs low vendors avg -132.48% (p<0.0001). Vendor curation materially drives profitability.
8	Jack Daniel's No 7 Black is the revenue leader	$7.96M in total sales — leads all 10,692 SKUs, followed by Tito's Handmade Vodka ($7.40M) and Grey Goose Vodka ($7.21M).

10. Recommendations

#	Area	Priority	Action
1	Pricing	High	Audit all 2,128 loss-making SKUs. Renegotiate purchase prices with vendors or raise retail prices to restore positive margins immediately.
2	Promotions	High	Launch targeted campaigns for the identified high-margin / low-volume brands (margin >56%, sales <$286). Shelf placement and digital promotion — no discounting required.
3	Supply Chain	High	Reduce dependency on the top-10 vendor cluster. Onboard 3–5 alternative suppliers for the highest-spend categories to mitigate single-vendor disruption risk.
4	Dead Stock	Medium	Initiate clearance for $2.71M in unsold inventory through discount bundling, time-limited promotions, or vendor return agreements.
5	Procurement	Medium	Consolidate small orders into large-order tiers to capture the 74% unit cost reduction evidenced in the data. Prioritise top-spend brands first.
6	Automation	Low	Schedule the CTE SQL refresh query to run weekly so Power BI always reflects current data without manual re-export.

11. Conclusion

This project demonstrates a complete, production-quality analytics pipeline covering every stage from raw data ingestion to executive-ready reporting. The vendor_sales_summary table serves as a single source of truth — consolidating procurement, sales, freight, and pricing data into 10,692 actionable records — and the Power BI dashboard makes this intelligence accessible to non-technical stakeholders.

The analysis surfaces concrete, data-backed actions with measurable impact: $2.71M in recoverable capital, 65.34% supplier concentration risk requiring diversification, high-margin brands awaiting promotional investment, and a statistically proven 162-point profitability gap between vendor tiers ($p<0.0001$). The modular stack — MySQL for storage, Python for analysis, Power BI for presentation — is maintainable, scalable, and ready for scheduled production refresh.

End of Report