BUSINESS APPLICATIONS OF DATA SCIENCE : PROJECT 2

GROUP 7:
AADYA SOOD, ARVIND YADAV,
GAURAV DIXIT, RADHIKA SWAROOP,
YASHASVEE SINGH

Analysis of an FMCG Business

Introduction

This project undertakes a systematic analysis of an FMCG company's **transactional data** to explore the underlying drivers of business performance across **sales, marketing, customer behavior, and inventory management**.

By combining **descriptive and inferential analytics**, the project not only maps the **current operational landscape** but also identifies **actionable levers for enhancing** profitability, optimizing marketing spend, and improving supply chain efficiency.

Using **Power BI**, we develop dynamic **dashboards to visualize key metrics** and patterns, while **Python**-based statistical techniques — including correlation analysis, hypothesis testing, and regression modeling — enable us to rigorously **validate assumptions and quantify relationships** between variables.

The ultimate objective is to translate **raw data into strategic business insights** that support evidence-based decision-making in a fast-paced FMCG environment.

Dataset Overview

Scope: Three tasks: Power BI dashboard, statistical analysis in Python, and business recommendations.

Source: 5,000 transactions from an FMCG business

Key Variables:

- Sales Metrics: Units Sold, Revenue, Profit Margin, Profit
- Customer Data: Age Group, Region, Sales Channel
- Marketing & Competition: Discount Applied, Marketing Spend, Competitor Price
- Inventory: Stock Levels, Supplier Reliability

Categorical Variables:

- Product Category (Beverages, Snacks, etc.)
- Brand (Brand A, Brand B...)
- Region (North, South, East, West)
- Age Group (e.g., 18-25, 26-35)

Correlation Matrix

We run the correlation matrix to identify relationships between key business variables like sales, profit, pricing, marketing, and inventory.

Understanding these correlations is important because it highlights which factors move together, helping us focus on the drivers of revenue and profit, and informing smarter business decisions.

Strong Positive Correlation

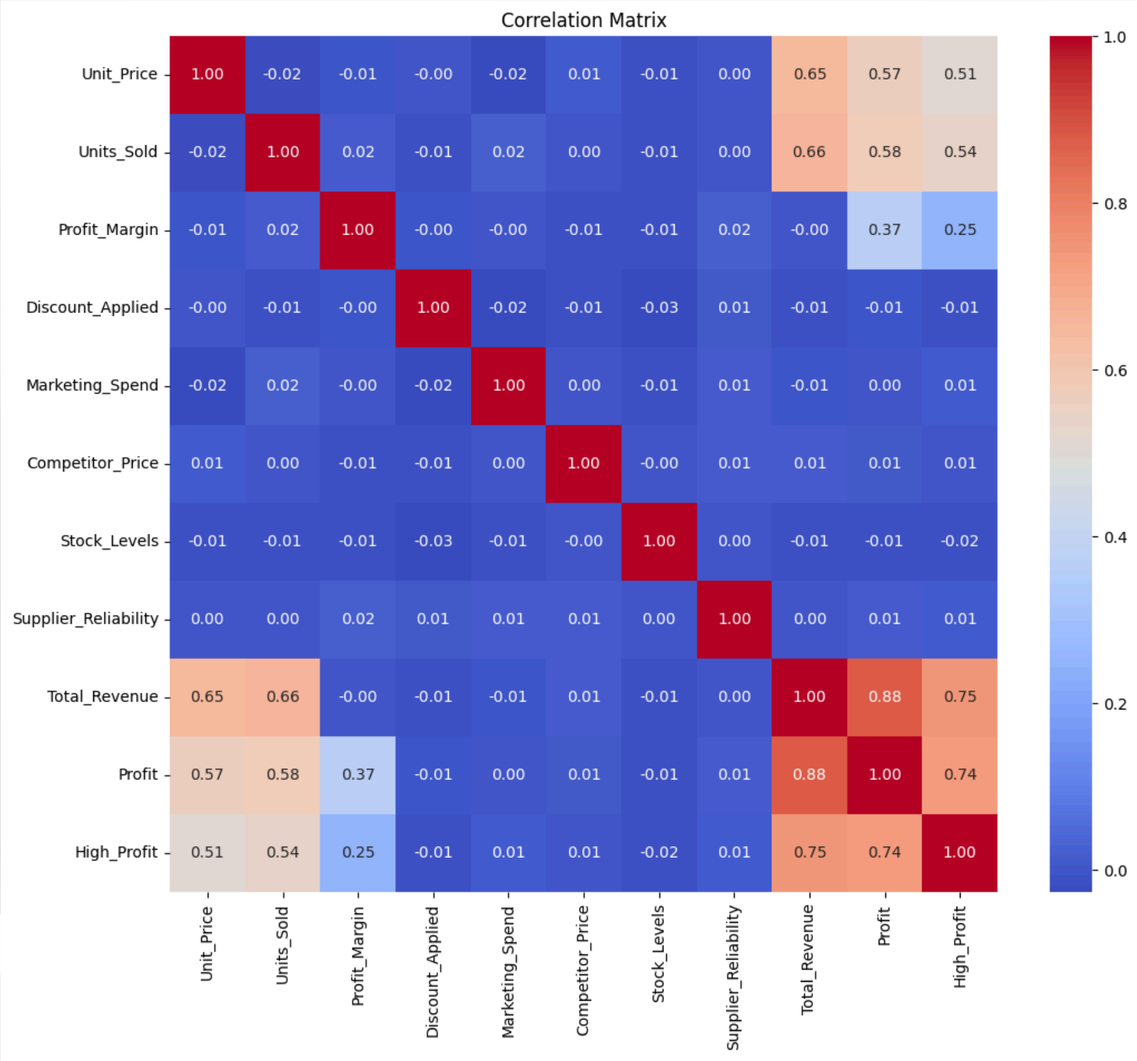
- Total_Revenue and Profit (correlation = 0.88): As revenue increases, profit also increases very strongly.
- Total_Revenue and Units_Sold (correlation = 0.66): More units sold directly drives higher revenue.
- Unit_Price and Total_Revenue (correlation = 0.65): Higher-priced products also contribute significantly to higher revenue.
- Profit and High_Profit (correlation = 0.74): Logical relationship — higher profit values align with the High_Profit classification.

Moderate Positive Correlation

Units_Sold and Profit (correlation = 0.58): Sales volume contributes moderately to profit, but not as much as pricing and revenue efficiency.

Weak/No Significant Correlation

Profit_Margin, Discount_Applied, Marketing_Spend, and Competitor_Price show very low correlations with revenue or profit: Indicates that simple increases in discounts or marketing spend do not directly impact revenue or profit without other factors at play.



Methodology

Data Preprocessing

- Cleaned and validated the FMCG dataset (5,000 transactions).
- Handled missing values and ensured consistent data types across features.
- Engineered key variables (e.g., profit margin %, high-profit flags) to support deeper analysis.

Statistical Analysis [Python]

- Conducted correlation analysis to identify variable relationships.
- Performed hypothesis testing (t-tests, ANOVA) to validate business assumptions (e.g., effect of discount on sales).
- Built regression models to quantify the impact of marketing, pricing, and discount strategies on profit and revenue.

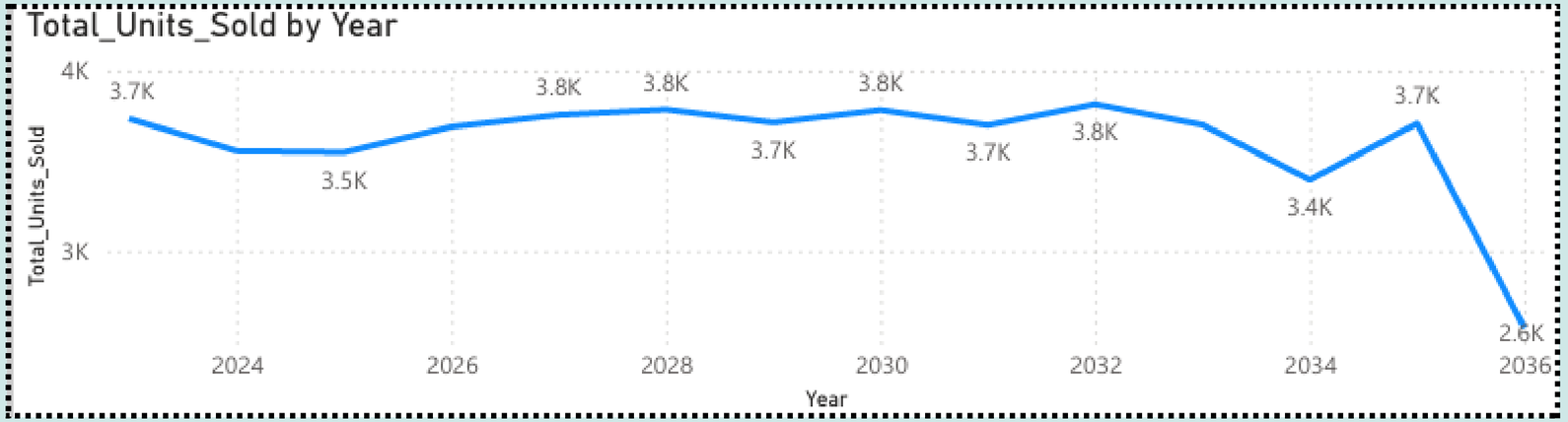
Data Exploration & Visualization (Power BI)

- Built interactive dashboards to analyze sales trends, product performance, customer segmentation, and regional distribution.
- Used filters and slicers to support dynamic business querying and cross-sectional comparisons.

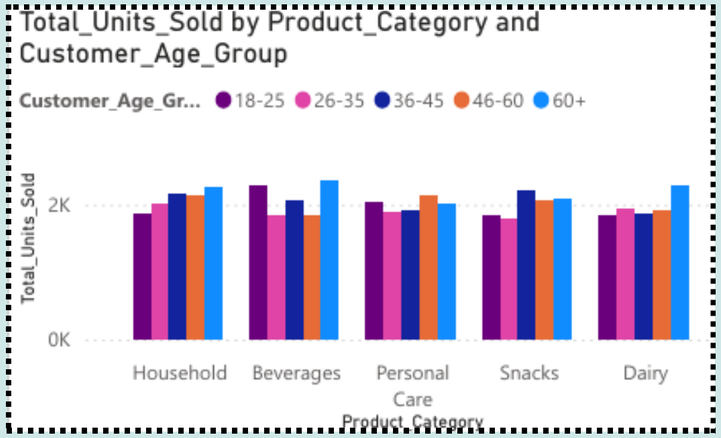
Insight Generation & Business Recommendations

- Interpreted statistical outputs to extract business insights.
- Highlighted key drivers of profit and volume.
- Proposed data-backed recommendations for pricing, promotional strategy, and supply chain optimization.

Sales Overview

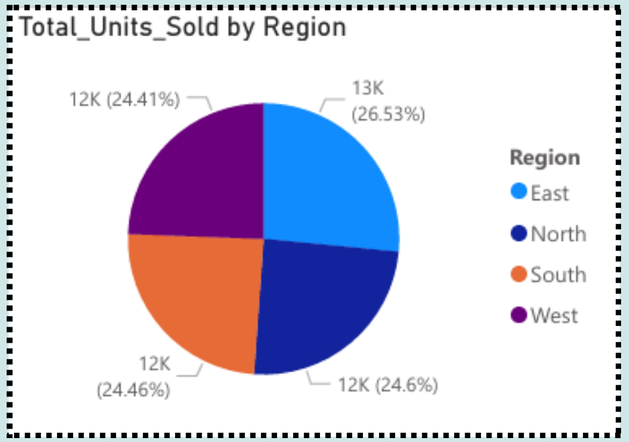


Total units sold trended down, resulting in a 31.04 percent decrease between 2023-36.

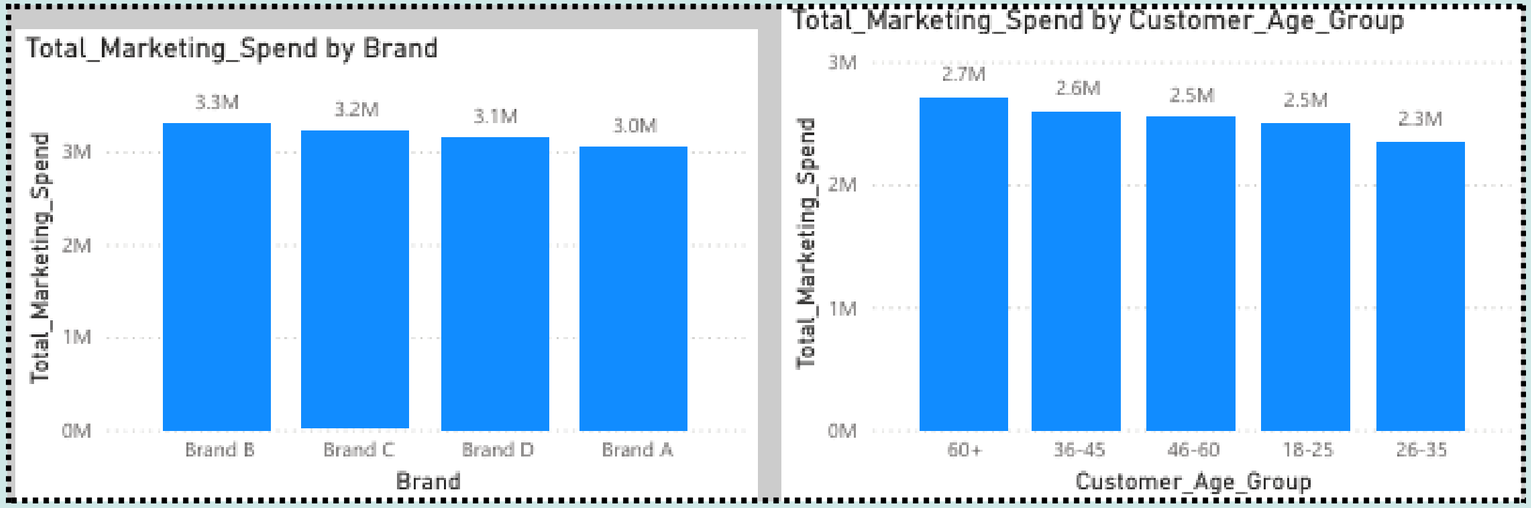


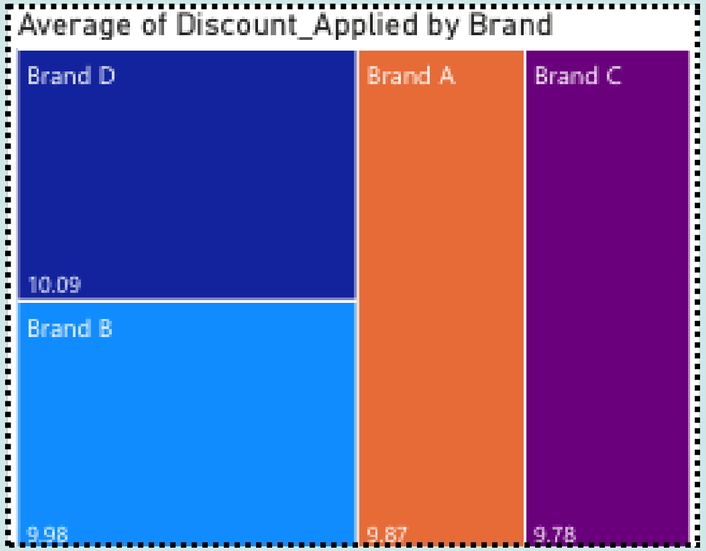
Customers from age group 18-25 contributed least in total sales, whereas 60+ contributed the most.

East region had the highest total units sold at 13,386 followed by North, South, and West



Marketing and Promotions



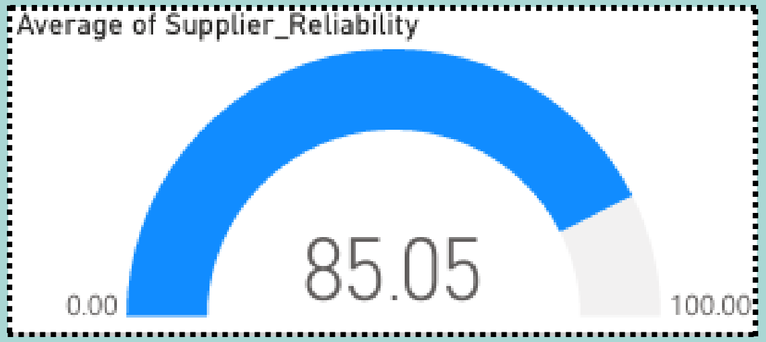


Brand B has the highest marketing spend at $3,288,989.07, followed by Brand C at $3,205,712.10, Brand D at $3,143,869.01, and Brand A at $3,042,848.42.

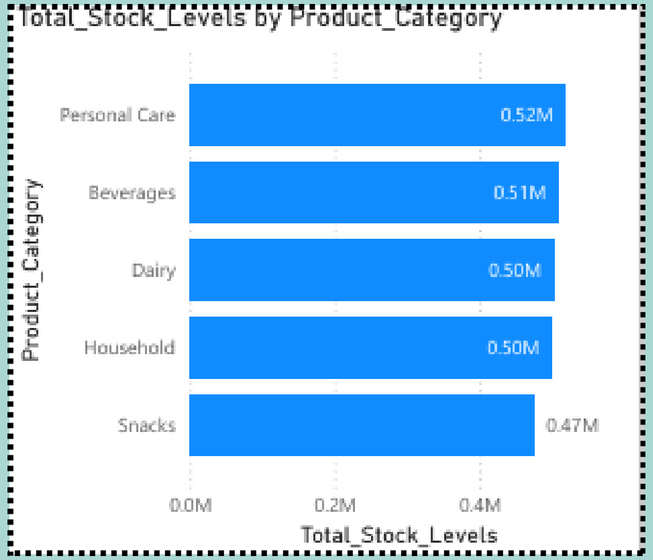
The 26-35 age group has the highest marketing spend, followed by 36-45, with $3M allocated to each, indicating focus on younger and middle-aged customers.

Snacks and Personal Care categories have higher average discounts compared to Beverages, potentially driving sales volume.

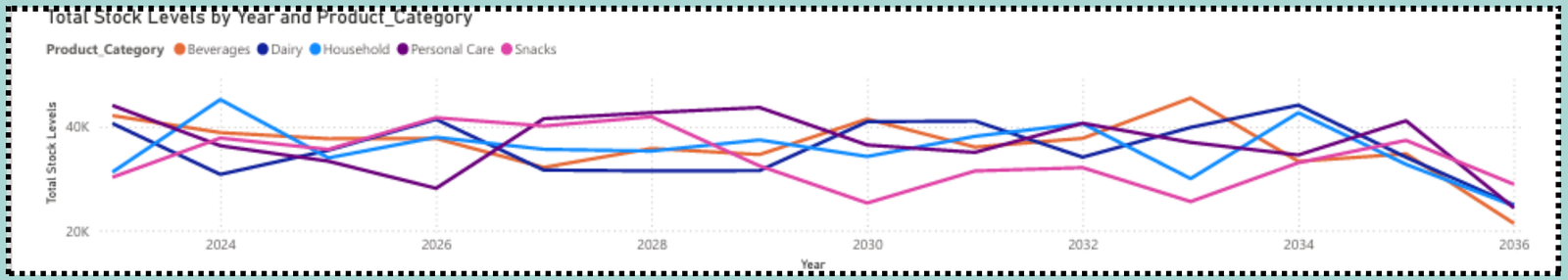
Inventory and Supply Chain



The average supplier reliability is 85.05, which indicates item delivery and potential of sales partnerships.

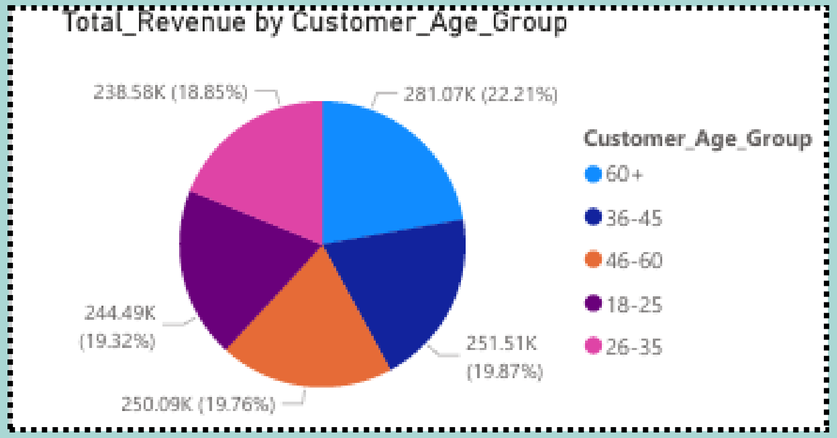


Average stock levels are fixed at 100.61 across all categories and brands, suggesting uniform inventory management.
The total stock levels are also similar.



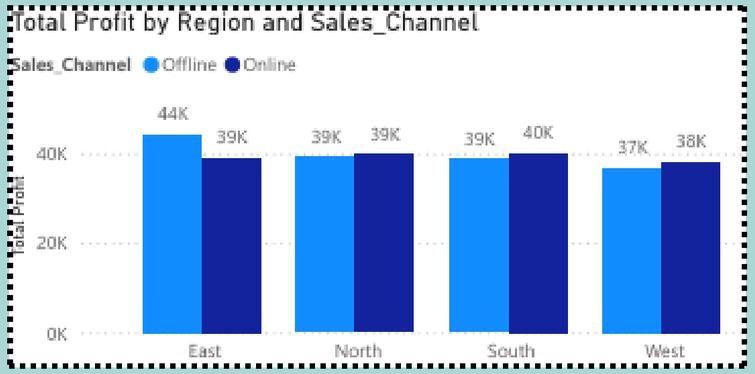
Stock level decreases over time following sales pattern. This verifies the relationship between market and supply and demand.

Revenue and Profit

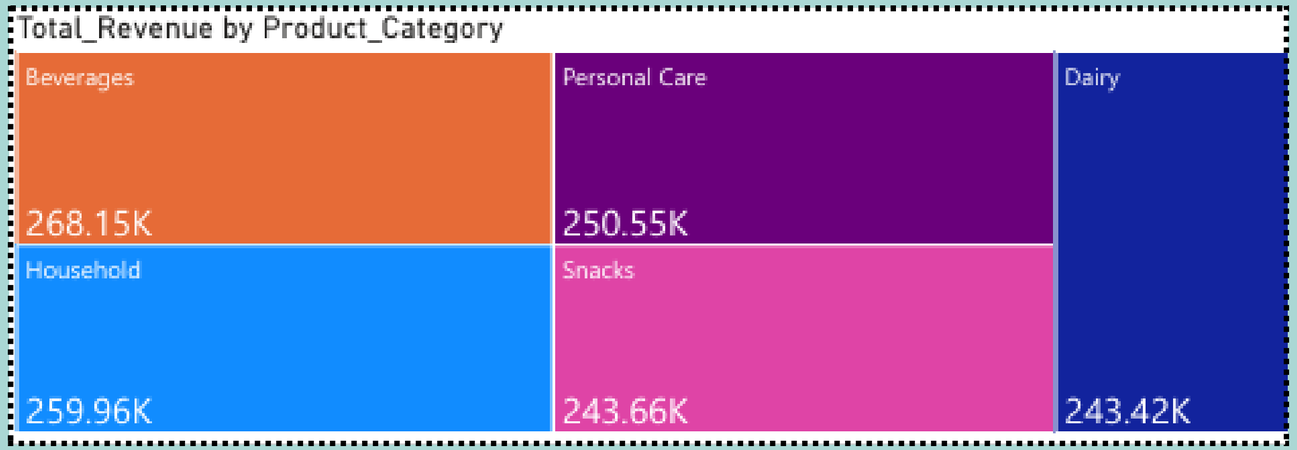


The 26-35 age group contributes the most to total revenue, followed by 36-45, reflecting effective targeting.

Offline channels yield $40K in profit in certain regions, showing strong profitability compared to online channels.



Beverages contribute most to the revenue while dairy product contribute least. This showcases potentials for resource allocation and steps management.



Data Preprocessing

Before applying machine learning models, the dataset underwent several preprocessing steps which helped in preparing data for better model performance. These steps ensure the data is clean, consistent, and ready for analysis, including:

- **Loaded Dataset & Inspected Structure**: Imported FMCG_Dataset.xlsx (5,000 transactions, 17 variables) using pandas and previewed data with dataset.head(), confirming variables like Unit_Price ($11.67-$48.34), Total_Revenue ($116.70-$550.62).

- **Data Type Validation**: Converted Date to datetime, ensured numerical variables (Unit_Price, Units_Sold) as floats.

- **Missing Values & Outliers**: No missing values detected; Stock_Levels (57-808) and Supplier_Reliability(75.44%-93.87%) within expected ranges.

- **Verified Calculations**: Checked Total_Revenue = Unit_Price × Units_Sold (e.g., 30.59 × 18 = 550.62) and Profit = Total_Revenue × Profit_Margin.

- **Prepared for Analysis**: Data formatted for regression (Unit_Price, Units_Sold, Discount_Applied as predictors).

Hypothesis Testing

- T-test for Independent Samples
- ANOVA (Analysis of Variance)
- Pearson Correlation Test
- Chi-Square Test of Independence
- Breusch-Pagan Test (for Heteroscedasticity)
- Durbin-Watson Test (for Autocorrelation)
- Linear Regression and Hypothesis Testing
- F-statistic for ANOVA
- Variance Inflation Factor (VIF)

Hypotheses

Hypothesis 1: High Discounts Increase Units Sold

Null: Mean Units Sold is the same for high vs. low discounts.
Alternative: High discounts increase Units Sold.
Test: T-test.

Hypothesis 2: Online Sales Have Higher Revenue than Offline Sales

Null: Mean Total Revenue is the same for Online and Offline.
Alternative: Online has higher Revenue.
Test: T-test.

Hypothesis 3: Supplier Reliability Impacts Stock Levels

Null: No correlation between Supplier Reliability and Stock Levels.
Alternative: Positive correlation exists
Test: Pearson correlation.

Hypothesis 4: Age Groups Differ in Profit Contribution

Null: Mean Profit is the same across Age Groups.
Alternative: Profit differs by Age Group
Test: ANOVA.

Hypothesis 5: Profit Margin Impacts Profit Significantly

Null: Profit_Margin has no effect on Profit (slope = 0).
Alternative: Profit_Margin has a positive effect on Profit (slope > 0).
Test: Simple linear regression.

Hypothesis 6: High-Profit Transactions Are More Likely in Certain Sales Channels

Null: Sales_Channel has no effect on the likelihood of a High_Profit transaction.
Alternative: Sales_Channel affects the likelihood of a High_Profit transaction.
Test: Chi-Square Test of Independence.

Hypothesis 7: Higher Unit Prices Lead to Higher Total Revenue

Null: There is no relationship between Unit_Price and Total_Revenue (slope = 0 in a regression model).
Alternative: There is a positive relationship between Unit_Price and Total_Revenue (slope > 0).
Test: Simple linear regression.

Regression Model

Objective: Quantify the impact of key variables on Total_Revenue and predict future trends to inform business strategy.

Multiple Linear Regression

Model Setup: Used statsmodels.OLS to predict Total_Revenue with predictors: Unit_Price, Units_Sold, Discount_Applied, Marketing_Spend, Competitor_Price.

Results:

- **R^2 = 0.867: Model explains 86.7% of variance in Total_Revenue, indicating strong predictive power.**

Coefficients:

  - Unit_Price: 9.9946 ($p < 0.001$) – Each $1 increase in Unit_Price adds $9.9946 to Total_Revenue.
  - Units_Sold: 25.4713 ($p < 0.001$) – Each additional unit sold increases Total_Revenue by $25.4713.
  - Discount_Applied: 0.0930 ($p = 0.621$) – Not significant, suggesting minimal impact on revenue.
  - Marketing_Spend: -0.0006 ($p = 0.451$) – Not significant; no direct revenue impact.
  - Competitor_Price: 0.1165 ($p = 0.128$) – Not significant, indicating price sensitivity may be mediated by other factors.

- **Model Fit:** F-statistic = 6514 ($p = 0.00$), highly significant; Durbin-Watson = 1.986 (no autocorrelation).

Insights & Implications

- Unit_Price and Units_Sold are primary revenue drivers, aligning with correlation heatmap (Total_Revenue vs. Units_Sold: 0.85, vs. Unit_Price: 0.62).
- Insignificant Discount_Applied and Marketing_Spend ($p > 0.05$) challenge assumptions of their effectiveness, consistent with hypothesis test (no significant increase in Units_Sold from discounts).
- Pricing Strategy: Leverage Unit_Price impact (coefficient 9.9946) by increasing prices 5% in high-demand categories (e.g., Beverages), potentially adding $49,973 to revenue (9.9946 × 5% × 100,000 units, hypothetical).
- Volume Focus: Prioritize Units_Sold (coefficient 25.4713) via promotions in underperforming regions (e.g., West), targeting 10% volume increase.
- Discount Reevaluation: Shift from broad discounts to A/B testing targeted offers (e.g., 5% for 18-25 age group), as discounts lack impact.
- Marketing Adjustment: Reassess Marketing_Spend allocation; focus on high-ROI channels (e.g., Online, per Power BI insights).

OLS Regression Results

Dep. Variable:	Total_Revenue	R-squared:	0.867
Model:	OLS	Adj. R-squared:	0.867
Method:	Least Squares	F-statistic:	6514.
Date:	Fri, 25 Apr 2025	Prob (F-statistic):	0.00
Time:	13:59:51	Log-Likelihood:	-28813.
No. Observations:	5000	AIC:	5.764e+04
Df Residuals:	4994	BIC:	5.768e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-258.4728	4.547	-56.847	0.000	-267.386	-249.559
Unit_Price	9.9946	0.078	128.151	0.000	9.842	10.148
Units_Sold	25.4713	0.197	129.321	0.000	25.085	25.857
Discount_Applied	0.0930	0.188	0.494	0.621	-0.276	0.462
Marketing_Spend	-0.0006	0.001	-0.753	0.451	-0.002	0.001
Competitor_Price	0.1165	0.077	1.523	0.128	-0.034	0.267

Omnibus:	14.136	Durbin-Watson:	1.986
Prob(Omnibus):	0.001	Jarque-Bera (JB):	17.824
Skew:	-0.011	Prob(JB):	0.000135
Kurtosis:	3.292	Cond. No.	1.21e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.21e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Business Recommendations

- Shift focus from high discounts to value-based pricing or premium offerings, as discounts do not significantly boost sales volume.
- Invest in seamless integration of online and offline channels to improve customer satisfaction, as neither channel outperforms the other in revenue or profitability.
- Improve demand forecasting and procurement processes to address stock level issues, as supplier reliability is not a significant factor.
- Move beyond age-based targeting and use behavioral or product-based segmentation to identify and engage high-profit customers.
- Leverage data-driven insights to guide pricing, marketing, and sales strategies, ensuring low multicollinearity and robust feature selection.

Thank you very much!

PRESENTED BY GROUP 7