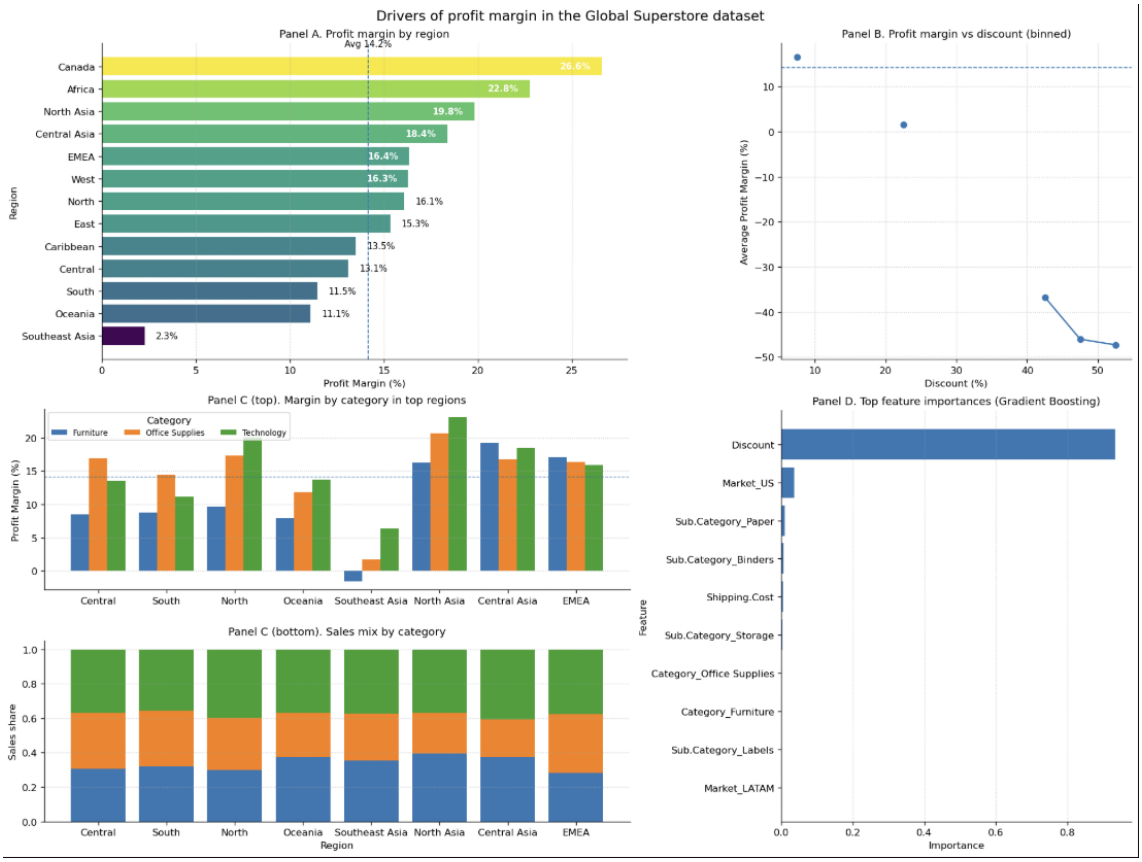
Understanding Regional Profitability in the Global Superstore Dataset

Legend:



Panel A – Profit margin by region. Horizontal bars show average profit margin (Profit ÷ Sales) by region, sorted high to low. Bar colors (viridis) encode margin level. A dashed vertical line marks the overall average margin (≈14.2%), and labels show percentages.

Panel B – Profit margin vs. discount. Points represent discount bins (0–10%, 10–20%, etc.). The x-axis is the bin midpoint; the y-axis is the average profit margin in that bin. A dashed horizontal line shows the overall average margin; points below it correspond to discount ranges with below-average or negative margins.

Panel C – Margin and sales mix by category in top regions. The top subplot shows average profit margin for Furniture, Office Supplies, and Technology in the top-selling regions, with a dashed line for the global average. The bottom subplot shows stacked bars of sales share by category in the same regions; each bar sums to 1.0.

Panel D – Feature importances. Horizontal bars show the relative importance of features in a gradient boosting regression model predicting order-level profit margin, including numeric variables (Discount, Shipping Cost, Quantity) and one-hot encoded categorical features. Longer bars indicate more influential predictors.

Findings

Regional differences

Canada, Africa, and North Asia have the highest margins (≈20–27%), clearly above the global average.

Central Asia, EMEA, West, North, and East cluster around the average, while Oceania, South, and especially Southeast Asia perform poorly (Southeast Asia ≈2%).

Discount effects

Profit margin decreases as discount increases: low discounts keep margins positive; high discounts push margins down.

At discounts above roughly 40–50%, average margins turn strongly negative, so products are often sold at a loss.

A global Spearman correlation (computed separately) confirms a strong negative relationship between discount and margin.

Product mix across regions

Technology usually has the highest margins, especially in North Asia and Central Asia; Furniture and Office Supplies are lower and more volatile.

Some region–category pairs (e.g., Furniture in Southeast Asia) have negative margins, indicating unprofitable niches.

Several regions rely heavily on lower-margin categories, which keeps their overall margins near or below the global average.

Model-based drivers

A gradient boosting model achieves good out-of-sample performance ($R^2 \approx 0.70$, MAE ≈ 0.12 in margin), so the chosen features explain meaningful variation.

Discount is by far the most important feature, dominating all others. Market indicators and specific subcategories (such as Paper and Binders) are secondary but still contribute.

Data and Methods

The Global Superstore dataset contains order-level records with fields such as region, market, product category and subcategory, order date, sales, profit, discount, shipping cost, quantity, and shipping mode. I loaded the data in Python using pandas, removed records with non-positive sales, converted key numeric columns to numeric types, and

defined profit margin as Profit ÷ Sales at the order level. To limit the impact of clear outliers and data errors, I restricted attention to orders with margins between –100% and +100%.

For Panels A–C, I used group-by aggregations. Panel A groups by region and computes total sales, total profit, and average margin per region; the global average margin is total profit divided by total sales across all orders. Panel B filters orders with discounts between 0 and 60%, bins discounts into fixed-width intervals, and computes average margin within each bin; Spearman correlations (computed separately) summarize the negative relationship between discount and margin. Panel C focuses on the top-selling regions (ranked by sales), computing region × category margins for the grouped bars and a row-normalized cross-tab of sales for the stacked sales-share bars.

For Panel D, I trained a gradient boosting regression model using scikit-learn. The target is clipped order-level profit margin. Numeric predictors are Discount, Shipping Cost, and Quantity; categorical predictors (Region, Market, Category, Subcategory, Ship Mode, Year) are one-hot encoded. I split the data into 80% training and 20% test sets, standardized numeric features, fit a GradientBoostingRegressor, evaluated performance using R^2 and mean absolute error, and visualized feature importances as a horizontal bar chart.

Significance

The integrated figure provides a compact view of how geography, pricing, product mix, and modeled drivers jointly shape profitability. It shows that only a few regions generate very high margins, that aggressive discounting is a primary cause of low or negative margins, and that category mix can make some region–category niches systematically unprofitable. The model-based panel confirms that discount is the dominant driver of margin, while market and product features refine the picture. These results suggest concrete actions such as reducing extreme discounts in weak regions, shifting the product mix toward higher-margin categories like Technology, and using the model to explore "what-if" scenarios for pricing and assortment decisions.

GitHub Repository

WSQB666/Global-Superstore-Margin-Analysis