Summary of Findings – Superstore Sales Dataset

Dataset Overview

- The dataset consists of **9,994 sales records** from a U.S.-based Superstore covering **categories, regions, and customers** between 2014 and 2017.
- It includes **21 columns** containing both numerical and categorical features such as Sales, Profit, Discount, Quantity, Category, Segment, and Region.
- There were **no missing values** or duplicates found, and data types were appropriately formatted for analysis.

Univariate Analysis

- **Sales** and **Profit** distributions are **right-skewed**, indicating a few high-value transactions contribute significantly to total revenue.
- **Discount** is mostly concentrated around **0.2 (20%)**, showing a consistent discounting strategy across most orders.
- **Quantity** values are generally small (between 1 and 5 units per order), suggesting the majority of sales are retail-level, not bulk.
- **Shipping Mode:** *Standard Class* is the most used mode (~60% of all shipments).
- **Segment:** *Consumer* segment dominates (~52%), followed by *Corporate* and *Home Office.*
- **Category:** *Office Supplies* generates the highest order volume, while *Technology* contributes fewer but higher-value transactions.

Bivariate Analysis

- **Sales vs Profit:** Positive relationship — higher sales tend to yield higher profit, except in cases of large discounts.
- **Discount vs Profit:** Negative correlation — as discounts increase, profit margins decline sharply.
- **Category vs Profit:** *Technology* products produce the highest average profit; *Furniture* (especially *Tables*) often results in losses due to higher discounts and shipping costs.
- **Segment vs Sales:** *Corporate* customers have higher average order values compared to *Consumers*, who generate more total orders overall.

- **Region-wise Analysis:** *West* and *East* regions lead in total sales, while the *South* region underperforms.

Multivariate Analysis

- **Correlation Matrix:**
    - *Sales* and *Profit* show a moderate positive correlation (~0.48).
    - *Discount* has a negative correlation with *Profit (-0.22)*, reinforcing that high discounts hurt profitability.
    - *Quantity* shows very weak correlation with other numeric variables, suggesting order size alone doesn't drive sales or profit.
- **VIF Results:** All variables have **VIF < 2**, confirming **no multicollinearity** issues.
- **Pairplot:** Clusters of points in scatterplots indicate potential segmentation by region or customer type.

Outlier & Skewness Analysis

- **Outliers:** Detected primarily in Profit, Sales, and Discount — representing large or heavily discounted orders.
- **Skewness:** Both Sales and Profit are highly right-skewed; after **log transformation**, the distributions became more symmetrical, improving readiness for predictive modeling.

Business Insights

- Heavy **discounting** leads to significant profit drops — especially in *Furniture* products.
- *Technology* is the **most profitable** category; targeted promotions and stock optimization could enhance revenue further.
- *Corporate* customers bring higher average order values, suggesting potential for loyalty or premium programs.
- *West* and *East* regions outperform others — expansion efforts could focus on *South* to balance revenue streams.
- The *Standard Class* shipping mode dominates, implying limited adoption of faster (premium) delivery options.