# Statistical Analysis For Super Store Sales

Canonical Correlation Analysis

**By Omar Wahby**

**5/2025**

# Overview

This analysis is based on the Superstore Sales dataset, which includes various sales transactions for a retail store. The dataset contains key features like product categories, sales amounts, profits, and shipping modes across different regions. The goal of this analysis is to extract meaningful insights to help improve sales and profits for the business.

The company generates sales across various categories and customer segments, but profitability patterns appear inconsistent. **This raises a key question:**

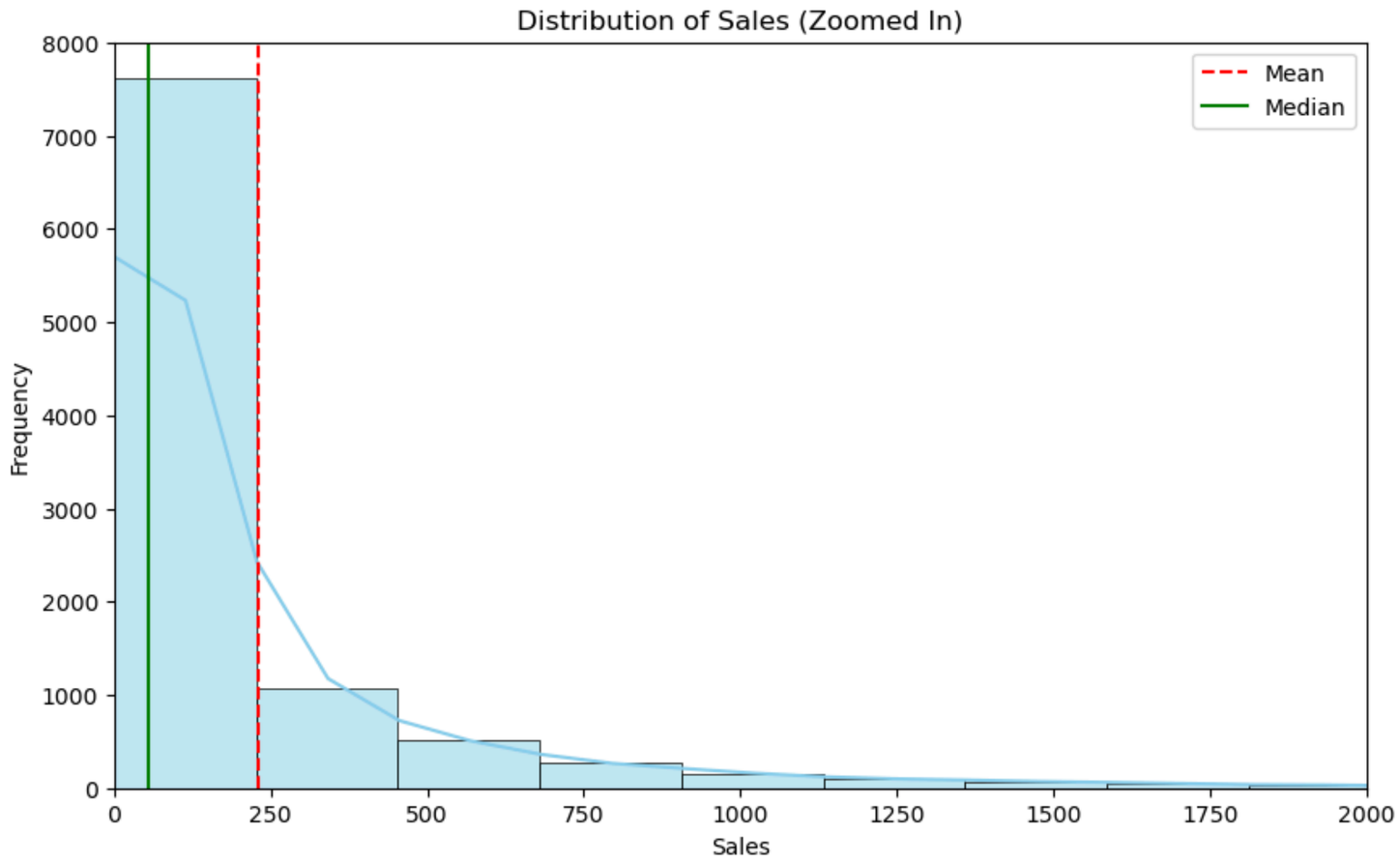
1. To what extent do sales behaviors impact profit ?
2. Does offering discounts lead to an increase in profit, or does it result in losses ?
3. Is there always a direct correlation between higher sales and increased revenue, or does this relationship vary ?
4. How does profitability differ across various product categories ?
5. What is the variance and stability of key metrics like Sales and Profit ?
6. Are there significant differences in these metrics across different categories ?
7. Are there any hidden relationships or inefficiencies influencing overall business performance ?

# Descriptives Statistics

## Measures of Central Tendency

| | Mean | Median | Mode |
|---|---|---|---|
| **Sales** | 229.86 | 54.49 | 12.96 |
| **Profit** | 28.66 | 8.67 | 0 |
| **Quantity** | 3.79 | 3 | 3 |
| **Discount** | 0.16 | 0.2 | 0 |

The mean sales per order is 229.86, while the median is significantly lower at 54.49, indicating a right-skewed distribution due to high-value outliers



# Descriptives Statistics

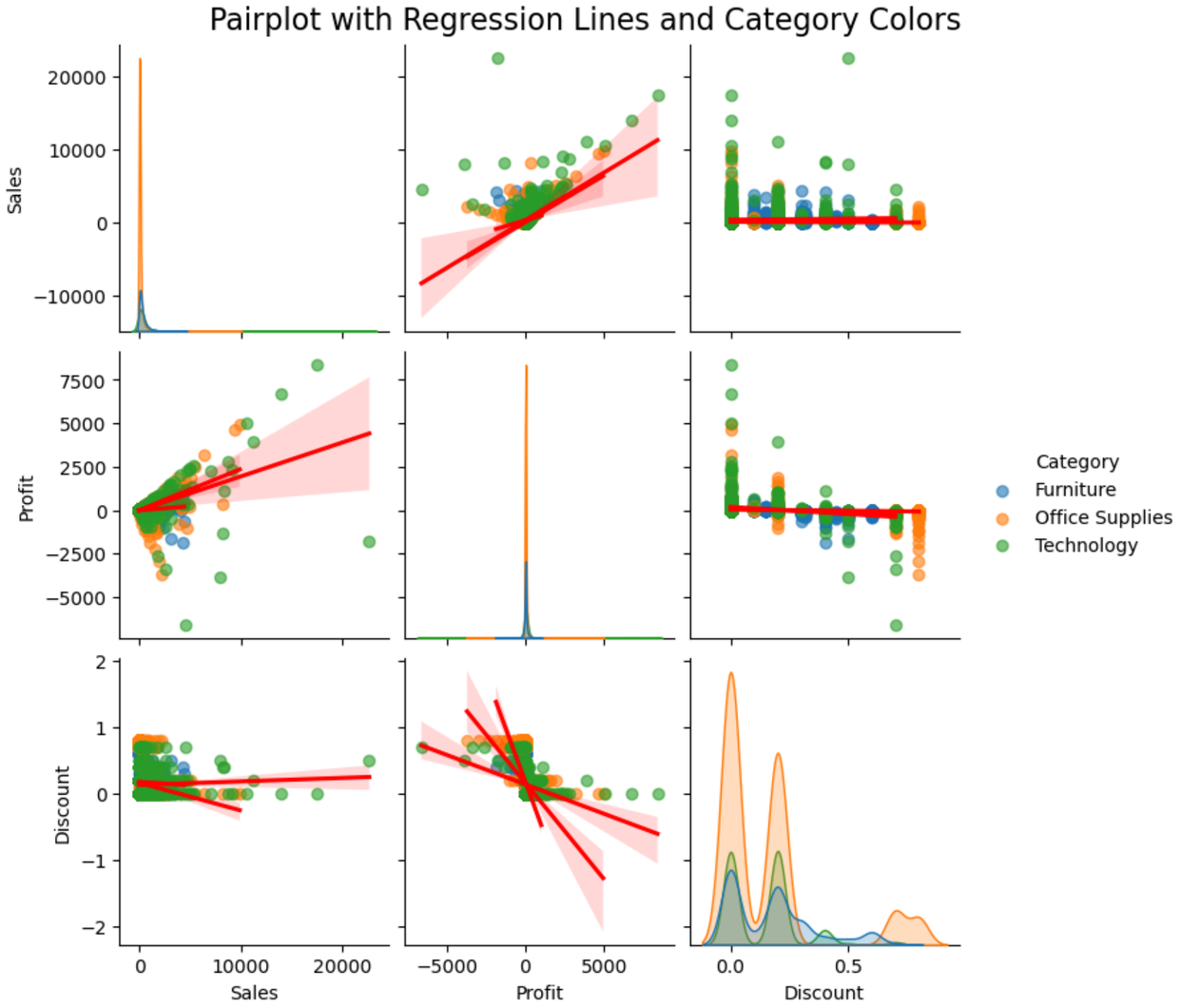
Covariance Matrix

| | Sales | Profit | Quantity | Discount |
|---|---|---|---|---|
| **Sales** | 388.43K | 69.94K | 278.46 | -3.63 |
| **Profit** | 69.94K | 54.88K | 34.53 | -10.62 |
| **Quantity** | 278.46 | 34.53 | 4.95 | 0 |
| **Discount** | -3.63 | -10.62 | 0 | 0.04 |

Higher sales often lead to higher profits, and this is evident in all categories, especially **technology**

High discounts often result in a loss or very low profit, especially in some product categories

Discounts don't necessarily increase sales in an obvious way



# Descriptives Statistics

Covariance Matrix

| | Sales | Profit | Quantity | Discount |
|---|---|---|---|---|
| **Sales** | 388.43K | 69.94K | 278.46 | -3.63 |
| **Profit** | 69.94K | 54.88K | 34.53 | -10.62 |
| **Quantity** | 278.46 | 34.53 | 4.95 | 0 |
| **Discount** | -3.63 | -10.62 | 0 | 0.04 |

Discounts Tend to Reduce Profit

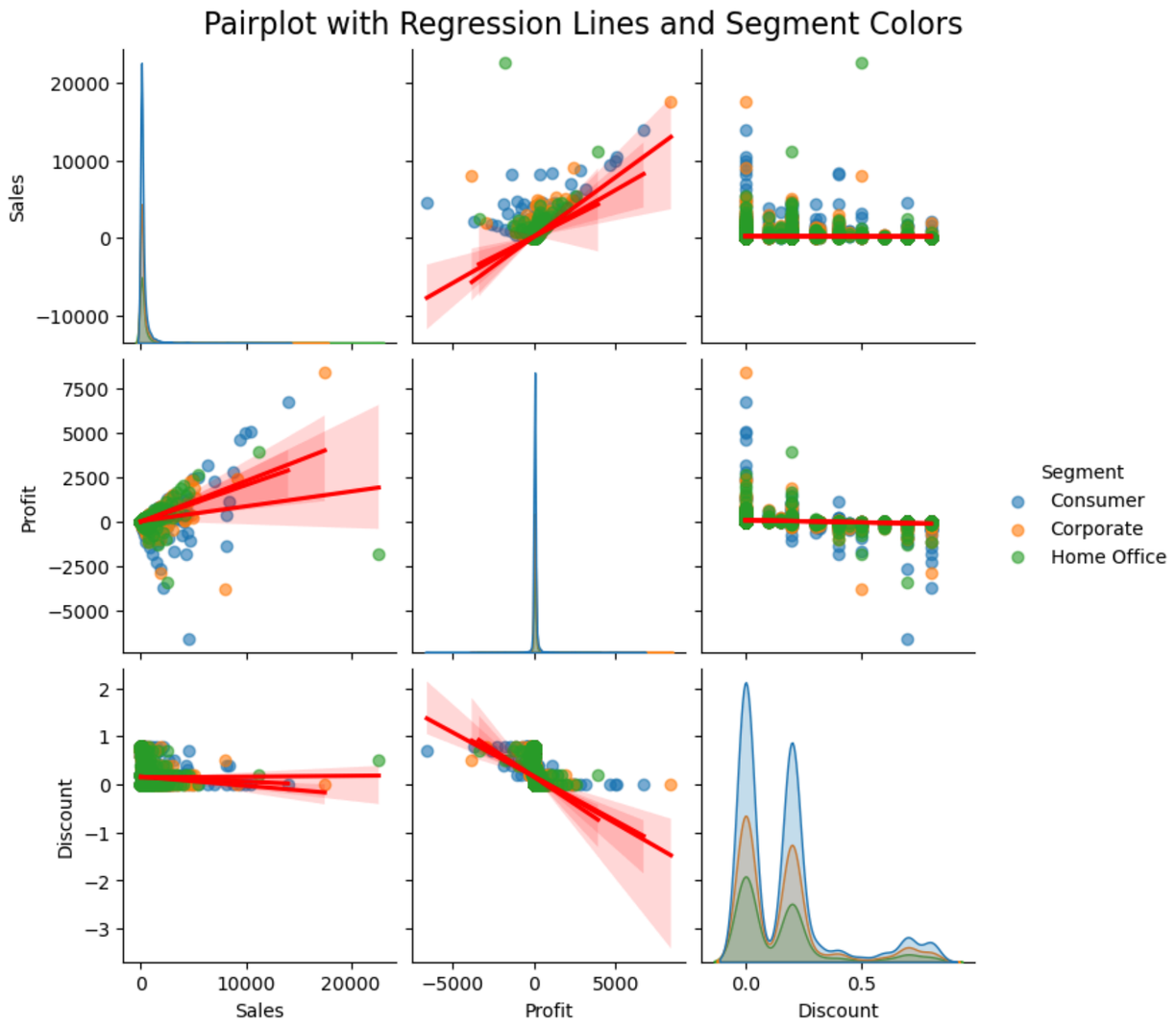
Sales Do Not Always Increase with Discounts

**Consumer** Segment is the Most Volatile

**Corporate** Segment Appears More Stable and Profitable

**Home Office** Segment Has Moderate Activity

There Are Some **Unprofitable Sales**



# Descriptives Statistics

## Correlation Matrix

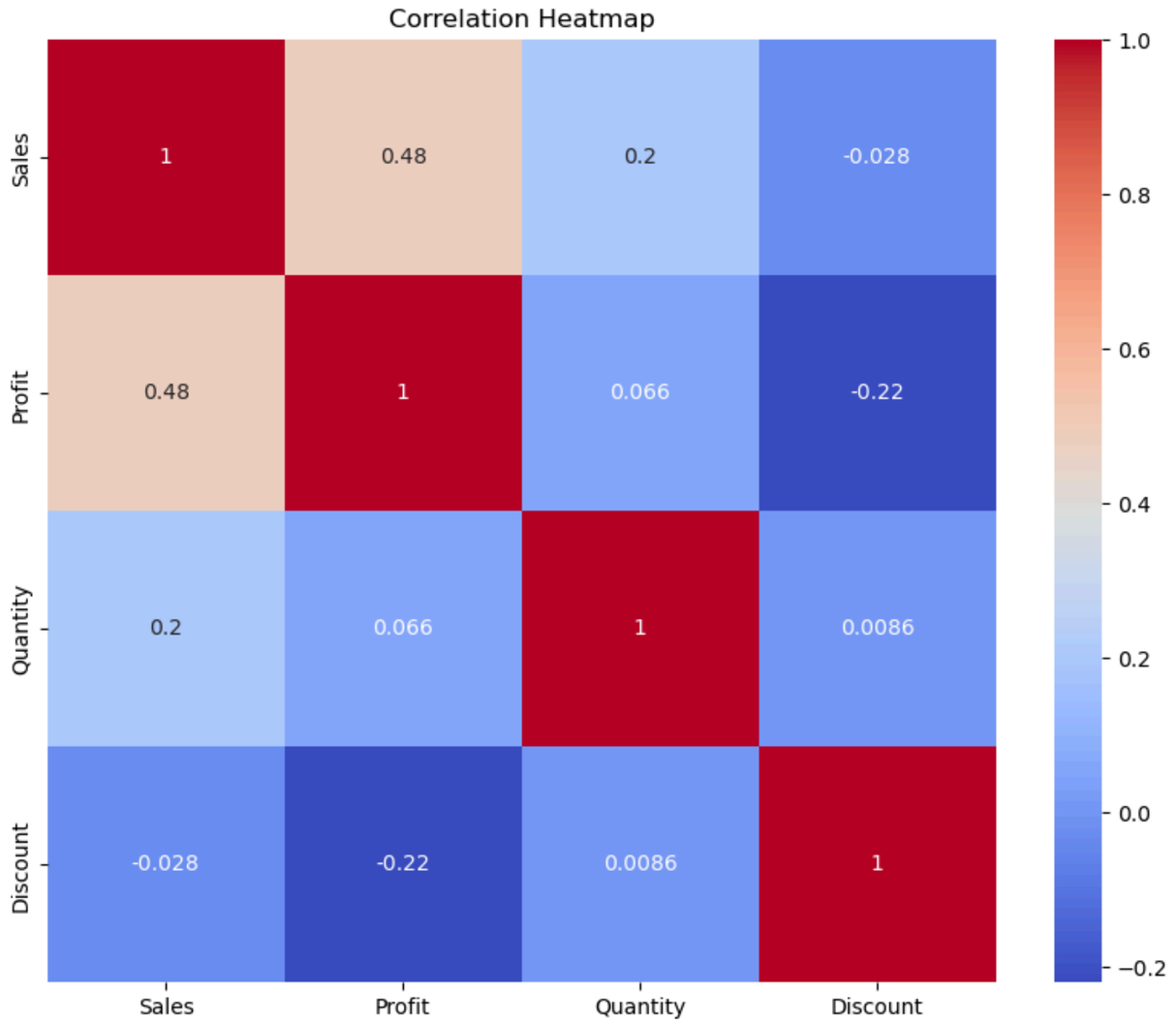
| | Sales | Profit | Quantity | Discount |
|---|---|---|---|---|
| **Sales** | 1 | 0.48 | 0.2 | -0.03 |
| **Profit** | 0.48 | 1 | 0.07 | -0.22 |
| **Quantity** | 0.2 | 0.07 | 1 | 0.01 |
| **Discount** | -0.03 | -0.22 | 0.01 | 1 |

Sales vs Profit (0.48): Moderate positive correlation , as sales increase, profits tend to increase too

Sales vs Discount (-0.03): No meaningful relationship, which could imply that giving discounts doesn't lead to higher sales

Profit vs Discount (-0.22): Negative correlation, indicating that higher discounts might reduce profit, as expected

Quantity correlations are weak, suggesting it's not a strong predictor for sales or profit



# Multivariate Normality Test

To evaluate whether our dataset follows a multivariate normal distribution, we performed a normality test

**Null Hypothesis ($H_0$)**:
The data follows a multivariate normal distribution

**Alternative Hypothesis ($H_1$):**
The data does not follow a multivariate normal distribution

**Test Result**
p-value < 0.05

**Conclusion :**
Since the p-value is less than 0.05, we reject the null hypothesis.
This means our data does not follow a multivariate normal distribution

# One Variance Test (Sales)

| Describe | Value |
|---|---|
| count | 9.99K |
| mean | 229.86 |
| Variance | 388.44K |
| std | 623.25 |
| Q1(25%) | 17.28 |
| Q2(50%) | 54.49 |
| Q3(75%) | 209.94 |
| min | 0.44 |
| max | 22.64K |

## Statistical Summary for Sales

- **Standard Deviation and Variance**

The standard deviation is 623.25 and variance is 388.44K
confirming the presence of large **fluctuations and outliers**

- **Minimum and Maximum Sales**

The lowest sales value is just **$0.44**, while the maximum reaches **$22,640.00**
This extreme range reinforces the idea of **outliers and right-skewed distribution**

- **Interquartile Range (IQR)**

25% of sales are below **$17.28**

50% (Median) are below **$54.49**

75% are below **$209.94**

These values show that most transactions are in the low-sales range, confirming that high-value orders are rare but impactful.

# Two Variances Test (Regions)

The null hypothesis ($H_0$) assumes the variances are equal

A p-value > 0.05 means we do not have enough evidence to say the variances are different, so we accept equal variances.

Test Results Summary

Sales Variance: East vs. West

**p-value = 0.33**

Fail to reject $H_0$ → The variances in sales between East and West are statistically equal

Sales Variance: Central vs. South

**p-value = 0.31**

Fail to reject $H_0$ → The variances in sales between Central and South are statistically equal

Conclusion: p-values (0.33 and 0.31) are well above the common threshold of 0.05, showing no significant difference in variance.

# Canonical Correspondence Analysis

Explore the relationship between two sets of variables

The goal of this CCA analysis is to understand how **categorical features** (Sub_Categories, Categories, Regions, Ship Mode) influence **numerical features** (Sales , Profit)

The Canonical Component 1 showed a moderate correlation of **0.454 > 0.3 ,** indicating a meaningful relationship worth exploring

**For CCA - Component 1 :**

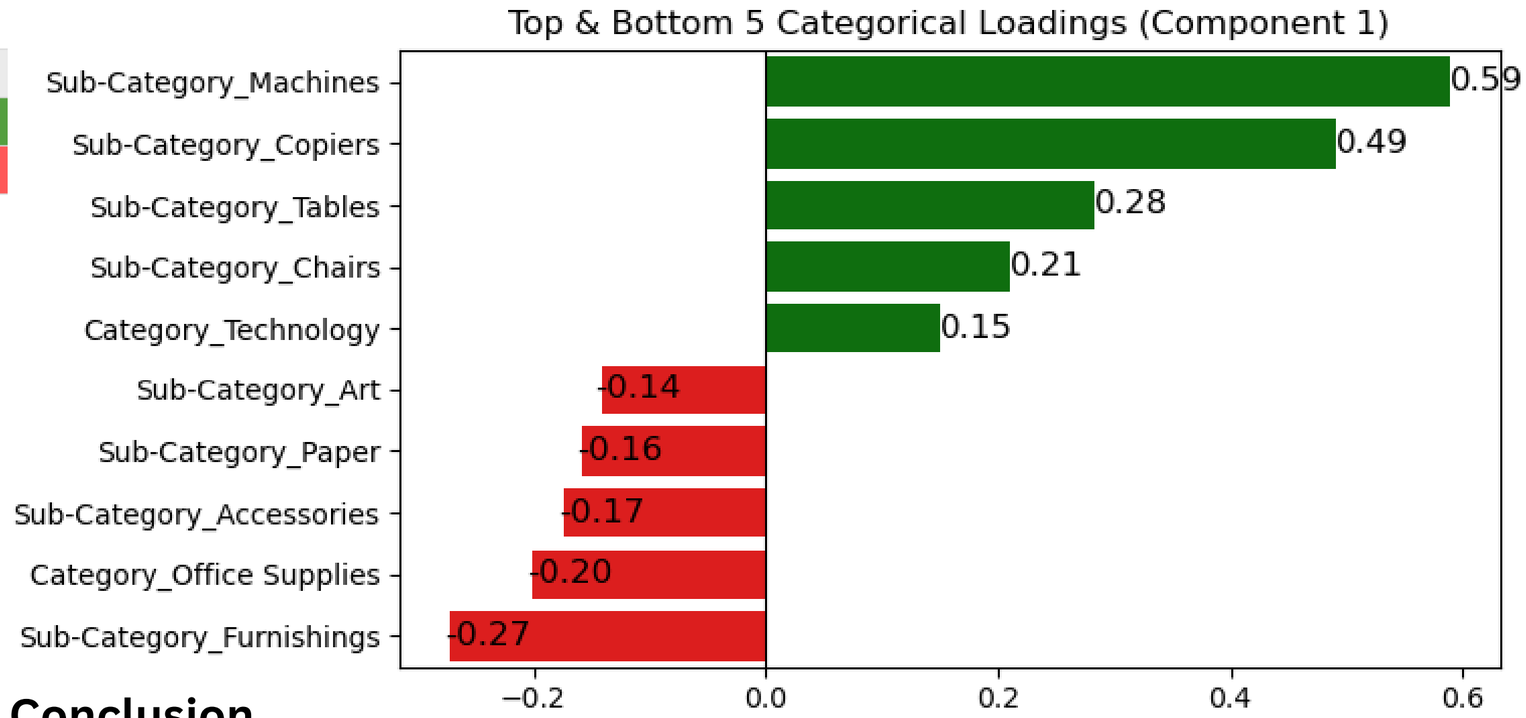
| Numerical Feature | CCA - Component 1 | |
|---|---|---|
| Sales | 0.977 | **Very strong positive** |
| Profit | -0.210485 | **Weak Negative** |

**Strong Positive Correlation**

| Categorical Feature | CCA - Component 1 |
|---|---|
| Sub-Category_Machines | 0.588 |
| Sub-Category_Copiers | 0.49 |
| Sub-Category_Tables | 0.283 |
| Sub-Category_Chairs | 0.21 |
| Category_Technology | 0.15 |

**Strong Negative Correlation**

| Categorical Feature | CCA - Component 1 |
|---|---|
| Sub-Category_Furnishings | -0.272 |
| Category_Office Supplies | -0.201 |
| Sub-Category_Accessories | -0.173 |
| Sub-Category_Paper | -0.158 |
| Sub-Category_Art | -0.141 |




## Conclusion

- **SubCategory (Machines , Copiers , Tables , Chairs )** is the Most **Positive** influential feature for Sales, There are other factors that influence but not significantly
- **SubCategory (Furnishings )** and **Category(Office Supplies)** is the Most **Negative** influential feature for Sales

# Canonical Correspondence Analysis

The **Canonical Component 2** has a **lower correlation 0.272 < 0.3**

Although multiple components were extracted during the Canonical Correlation Analysis (CCA), the canonical correlation value for Component 2 was only 0.277, which is considered relatively weak and indicates a limited association between the two sets of variables in this component. As such, Component 2 lacks sufficient explanatory power and does not contribute meaningfully to the overall interpretation, leading to its exclusion from the analysis in favor of focusing on the more informative Component 1

# Data-Driven Recommendations

We show you the best recommendations based on user reviews and your needs

**IF you want Increase Sales ?**

**Machines & Copiers & Tables** : These categories have a strong correlation with **high sales**, so focusing efforts here would help in boosting overall sales

---

**Why is the company losing ?**

**Furnishings & Office Supplies(Category):** These categories **reduce sales performance**, so efforts to minimize their sales or focus on other higher-performing categories could be beneficial for boosting overall sales

**Please feel free to ask**
**any questions about this presentation.**

# THANK YOU

*For Listening*

Omar Wahby