Indian Institute of Management, Indore
Master of Science in Data Science & Management

DSM 404 – Business Analytics

Instructor: **Prof. Aditya Maheshwari**

सिद्धिमूलं प्रबन्धनम्

भा. प्र. सं. इन्दौर

IIM INDORE

Term Paper

Submission Date : 27th November 2024

Sales & Profit Optimisation of a Superstore using Business Analytics Techniques

Submitted by:

Aadar Pandita (2304107001)

Batch 3 MSDSM

1. Abstract

This study explores the methods to **optimize sales and profit** for a Superstore by using various business analytics techniques. Using a dataset of 9,944 rows, **we performed exploratory data analysis (EDA), hypothesis testing, regression analysis – logistic and linear, clustering, and decision tree technique** to analyse our data and draw meaningful insights and understand key factors that drive sales, profitability, and category-wise performance. The study identified high-performing product categories & sub-categories, regions, and marketing strategies to improve long-term planning and resource allocation. The study also investigates the relationships between critical variables like sales, profit, quantity, and discount using visualizations to uncover underlying patterns and correlations, potentially leading to improved results and helps provide **actionable recommendations for enhancing the process of decision-making in sales, giving out promotional offers, and discount for the Superstore** .

This holistic approach aims to guide not only Superstore toward data-driven decisions that maximize profitability, improve customer satisfaction, and strengthen its competitive position in the market but also other companies that are looking for similar results.

2. Introduction

In a highly competitive world driven on data and its insights for better decision making and business outcomes, optimizing sales and profitability has become of imperative importance for the sustainability of any business.

We applied various business analytics techniques to analyse the data in order to look for trends, patterns and identify factors that are driving sales and profit in the Superstore.

For the Superstore, effective data-driven decision-making has become pivotal in addressing challenges such as optimizing product offerings, managing inventory, and devising impactful promotional strategies.

Objectives

The study aims to guide the Superstore in identifying high-performing categories and underperforming areas, refining discount strategies, and improving customer segmentation.

Research Questions

The study aims to answer the following key questions:

1. Which product categories and sub-categories drive sales & profitability?
2. What is the sales and profit distribution across all US states?
3. What is the impact of discounts on sales and profit?
4. How can we optimize and improve the current sales & profit?

By trying to answering these questions, the study highlights the pragmatic approach of business analytics in effectively managing key metrics that are central to generating huge profits.

3. Literature Review

Business Analytics in Profit Maximising

Studies reveal that business analytics enhances operational efficiency, customer targeting, and decision-making. Regression and decision trees are frequently used for sales prediction, while clustering aids in customer segmentation (Davenport & Harris, 2017).

Analytics in Sales

The application of business analytics to aid data-driven insights and decisions has gained significant attention in recent years. Studies highlight the role of **predictive analytics** in sales forecasting, **clustering algorithms** in customer segmentation, and **hypothesis testing** for identifying key profit drivers. For instance, Kumar et al. (2022) demonstrated the effectiveness of ARIMA models in demand forecasting, while Johnson et al. (2021) used Markov Chains to study customer purchasing patterns.

However, many studies focus on large-scale platforms, leaving gaps in understanding profitability strategies for mid-sized businesses. Moreover, while clustering and hypothesis testing are widely used, their integration into a comprehensive optimization framework remains underexplored.

Research Gap

This paper aims to fill the gap by providing a holistic analysis of sales and profit data using diverse analytics techniques for Superstore.

4. Methodology

4.1 Data Collection

Dataset : We used the **Superstore Dataset (SampleSuperstore.csv)** from Kaggle.

The dataset has 9,994 rows and 13 columns.

It is suitable for deriving useful insights across multiple dimensions like profitability by region, category and region based performance of sales, and the impact of discounts on profit.

Key Features:

1. **Sales**: Revenue from transactions in US dollars.
2. **Quantity**: Number of units sold.
3. **Discount**: Discounts applied to each sales transactions in percentage.
4. **Profit**: Profit generated in US dollars.

4.2 Data Preprocessing

- Raw data is transformed into usable format – The data was in .csv format.
- We filter the data and check for missing values – They were dropped.
- We detected duplicate values – They were removed from the dataset.
- Noise in the data – in the form of impossible and extreme values were also checked and we found no such values in our data.
- Data Integration was not required for our dataset since there were no conflicting values
- Data transformation was also not used to scale any values as they fell in the acceptable range to be used for visualisation and analysis.
- Data Reduction was done by dropping 2 columns – "Country" and "Postal Code" as these were not relevant to our analysis.

Above steps were taken to ensure data quality.

4.3 Tools Used

Python libraries - pandas, NumPy, sklearn, matplotlib, seaborn, plotnine and SciPy were used for data preprocessing, modelling, and visualization.

4.4 Techniques Used

We used multiple statistical methods and analytic techniques to analyse our data effectively.

1. **Exploratory Data Analysis (EDA):**
   We visualized sales and profit trends across categories, sub-categories, segments and regions.

2. **Hypothesis Testing:**
   Conducted tests like **One-Way ANOVA** and **Kruskal-Wallis Test**, to compare profit margins across regions and the impact of discounts on sales revenue.

3. **Clustering Analysis:**
   We applied K-means clustering to group customers based on sales and profit distribution along with preferences, highlighting the different purchasing behaviours across segments.

4. **Linear & Logistic Regression:**
   Built a linear and logistic regression model that aided us to predict high-profit products based on key features such as sales, category, and discount levels.

5. **Decision Trees**: Helped in analysing factors that drive sales.

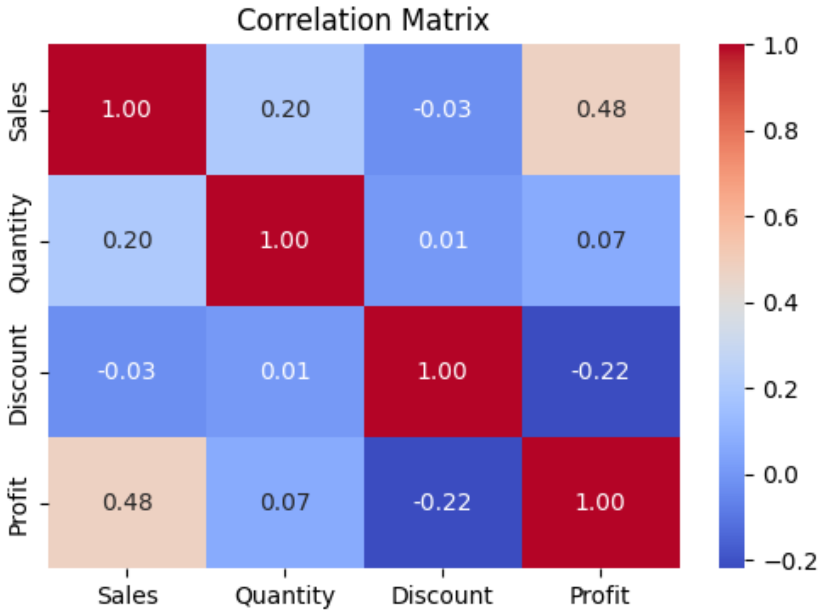
Limitations:

- Lack of sequential data didn't allow us to do Markov Chain analysis.
- No time-series column within data restricted us to perform ARIMA model analysis.

5. Results and Analysis

5.1 Descriptive Analysis

- Correlation Matrix Heatmap helped us identify that the data didn’t have any multicollinearity and therefore, we didn’t have to worry about any sort of over-fitting.



- Sales and Profit have a positive linear correlation of 0.48, meaning when sales increase by 1 dollar, profit is likely to go up by about 0.48 dollars. Superstore is popular amongst its consumers, has a good demand for its products, since it ensures high profit margins.

- Discount and Profit have a negative linear correlation of 0.22, meaning when discount is increased by 1%, the profit is likely going to decrease by 0.22 dollars. Discount decreases the overall revenue generated from sales.

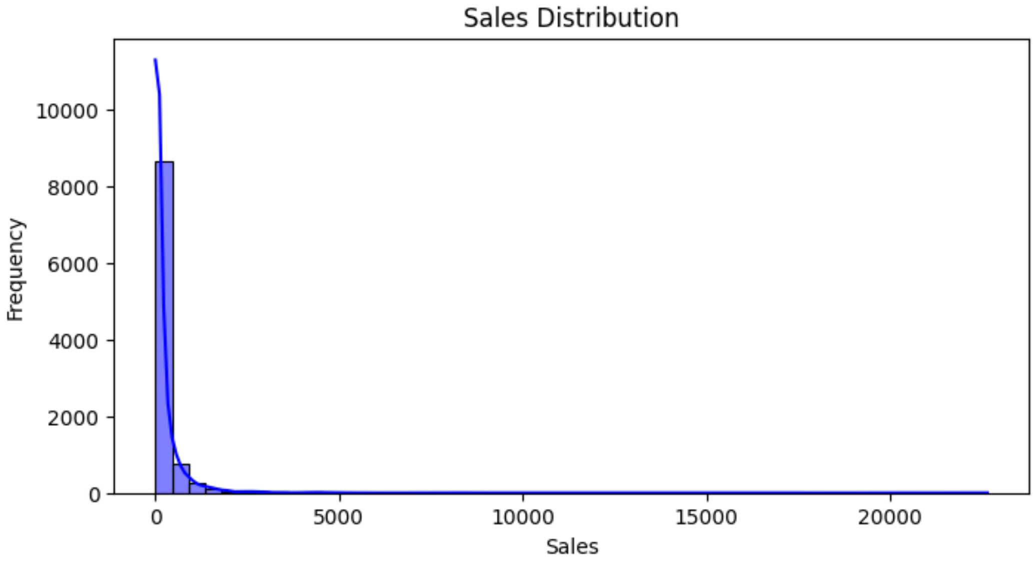
- Quantity and Profit show minimal linear correlation, suggesting that the Superstore sells large volumes at low profit margins, possibly due to bulk discounts or "Buy One Get One" type of promotions that they might be offering, reducing their average profit.

- Discount and Sales have little correlation, implying that discounts which were supposed to boost their sales, might be creating negative perceptions of quality, potentially harming the brand. Superstore should consider alternative promotional strategies.

- Discount and Quantity show minimal correlation, indicating that discounts are not an effective promotional strategy.

- Quantity and Sales have a weak positive correlation where sales increase by 0.20 dollars per unit.

5.2 Exploratory Data Analysis (EDA)

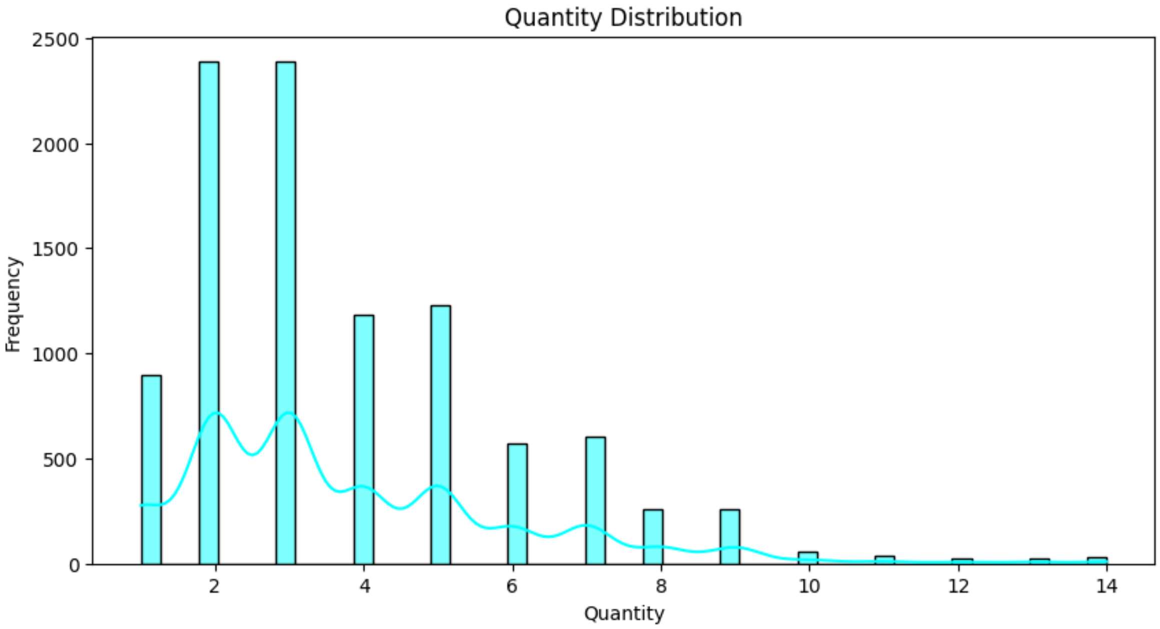


Sales Distribution:

- The distribution is highly **right-skewed**, indicating that most transactions have low sales values, while only a small number of transactions generate very high sales.

Implications:

- The business relies heavily on a few high-sales transactions to drive revenue.
- It might be beneficial to increase the volume of mid-range sales rather than relying on outliers.

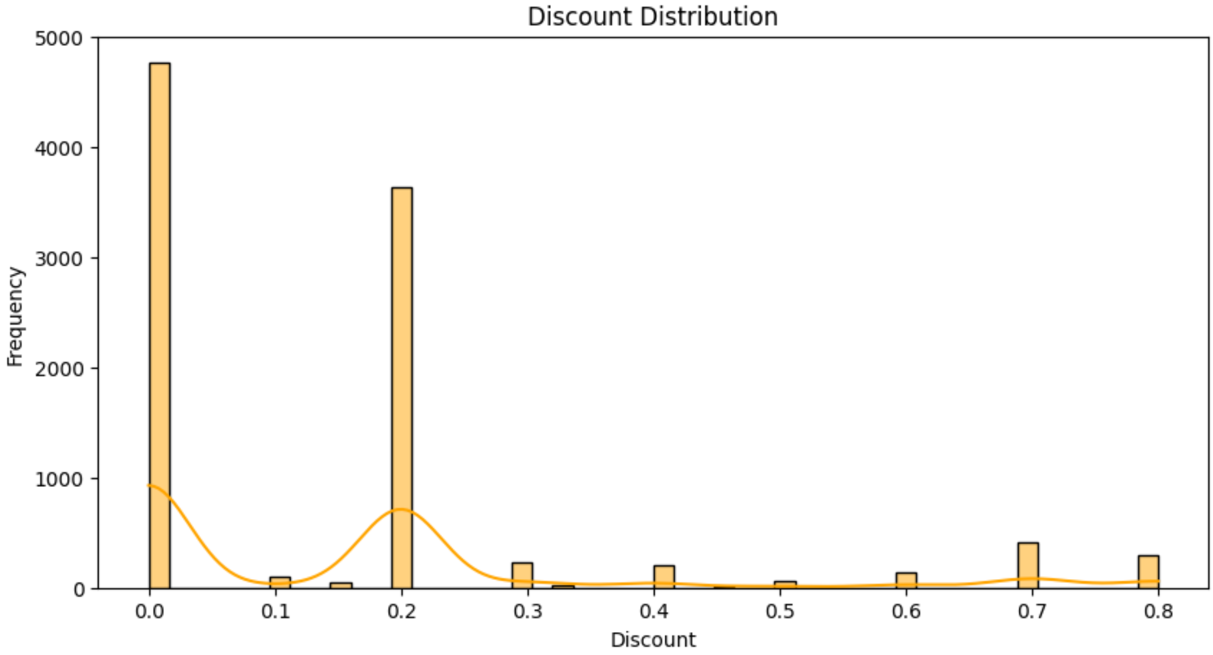


Quantity Distribution:

- The majority of transactions involve small quantities (2 or 3 units per transaction).
- Transactions with higher quantities (>10 units) are very rare.

Implications:

- Transactions involving smaller quantities have opportunity to upsell or be bundled.
- We need to understand which products are usually sold in larger quantities and then we can focus on creating a marketing strategy for those products.

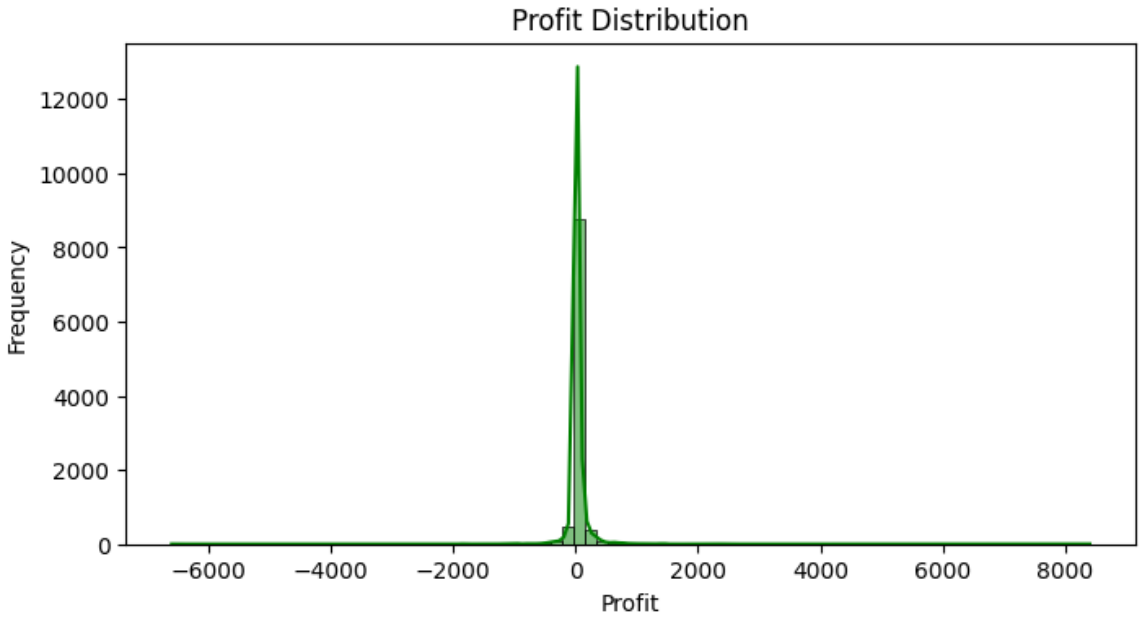


Discount Distribution:

- Discounts are mostly 0% (No discount) or 20%, but sometimes there are few higher discount rates like 70 and 80%.

Implications:

- 20% discount seems to be the standard promotional tactic.
- The high number of transactions have no discount and we can potentially introduce promotions on less selling items to increase their sales.

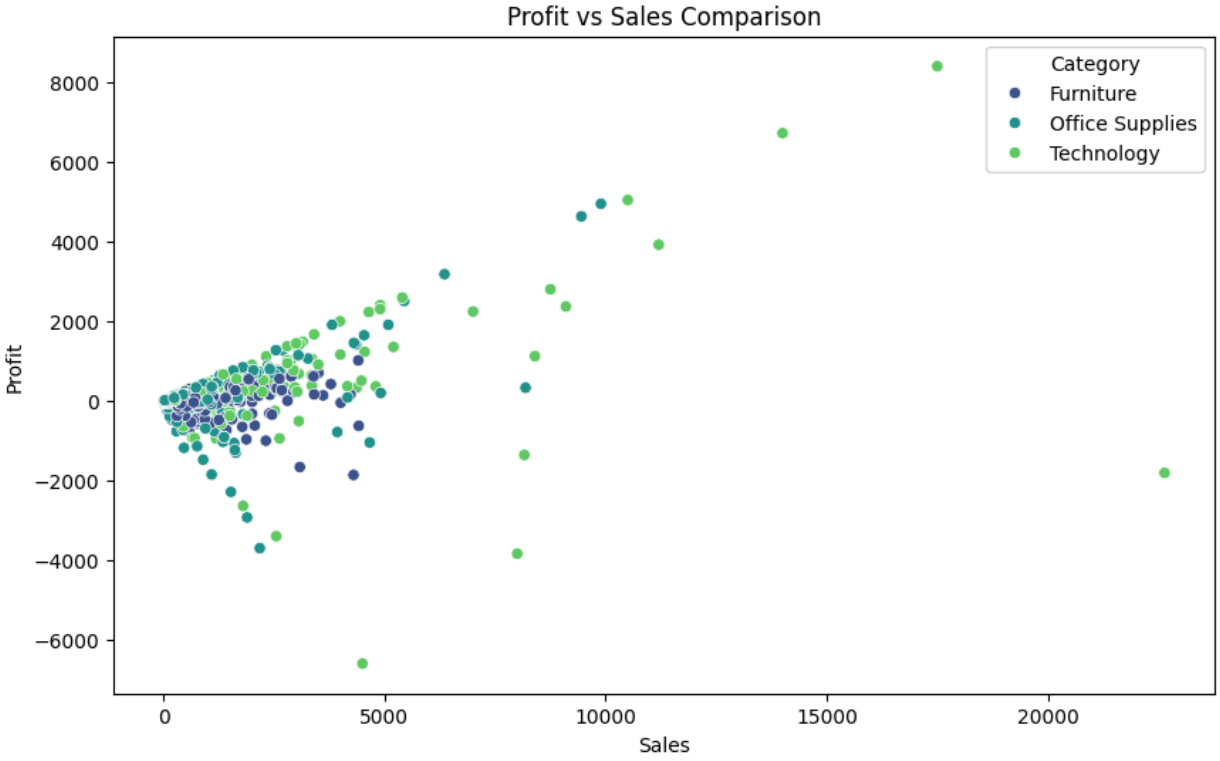


Key Observations from Profit Distribution:

1. **Skewness**:
   - The distribution is **right-skewed**, with most transactions clustered around low profit or losses.
   - A significant number of data points are in the negative range, indicating **unprofitable transactions**.
2. **Outliers**:
   - Some transactions have **very high profits** (highlighted by long extended tail of the distribution), thereby, some product categories or regions may perform exceptionally well.
3. **Concentrated Distribution**:
   - Majority of the transactions are close to break-even, with small profits or losses as they are cantered around zero.

Business Implications:

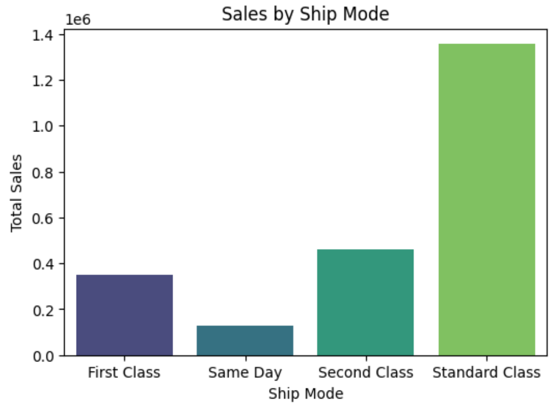
- **High Proportion of Losses**: There exist many loss-making transactions and need to be investigated further to identify potential reasons like:
  - Discounts offered
  - Product categories or subcategories
  - Regions or customer segments

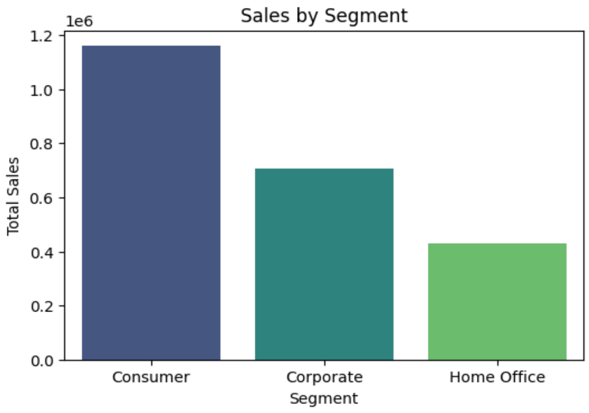


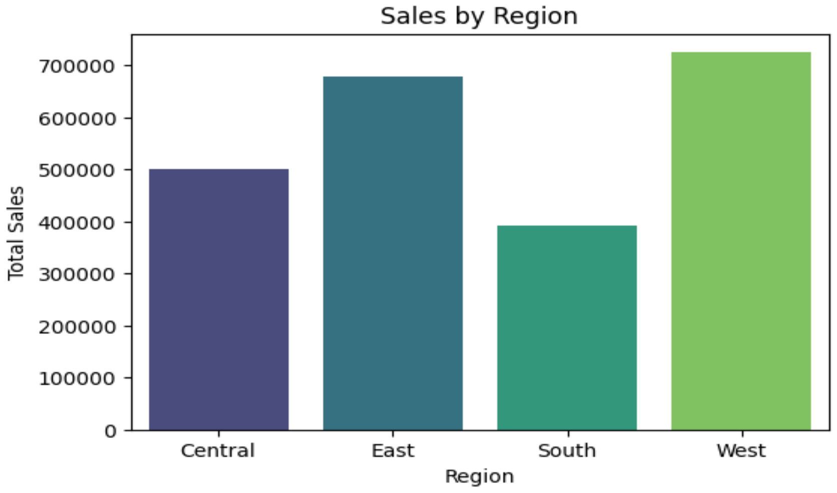
- There is a positive trend in general – higher sales leading to higher profits
- There are outliers to the general rule and some high sales have resulted in loss
- Technology have the most profit for high sales and best profit margin relatively
- For Furniture category, some high sales have resulted in significant losses meaning there seems to be ineffective discounts, shipping or promotions being offered
- Office Supplies have consistent but low profit margins

Implications :

1. Focus on technology should be high due to higher sales and profit margins
2. Loss management for Furniture category needs to be looked into
3. Office Supplies might benefit from targeted marketing and scaling

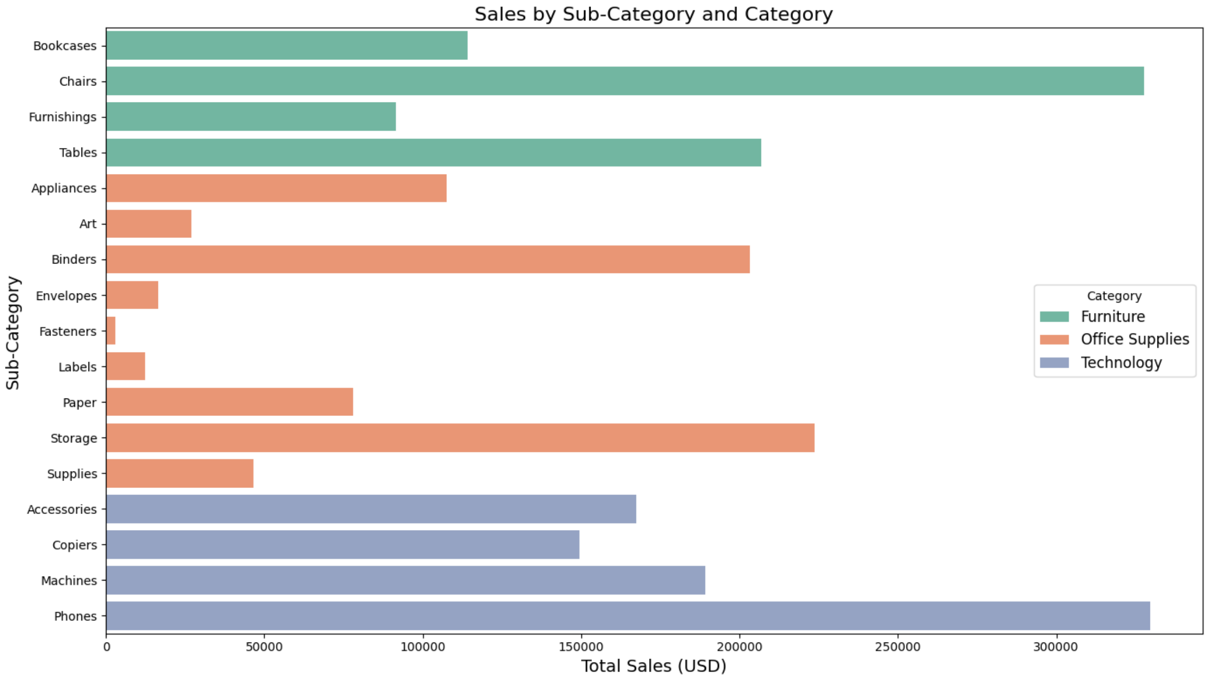






Insights from the graphs :

- **Standard Class** is the highest contributor to sales amongst the Ship Modes
- **Consumer Segment** has the most sales in all 3 segments
- **West and East region** contribute significantly more to sales than the other 2 regions

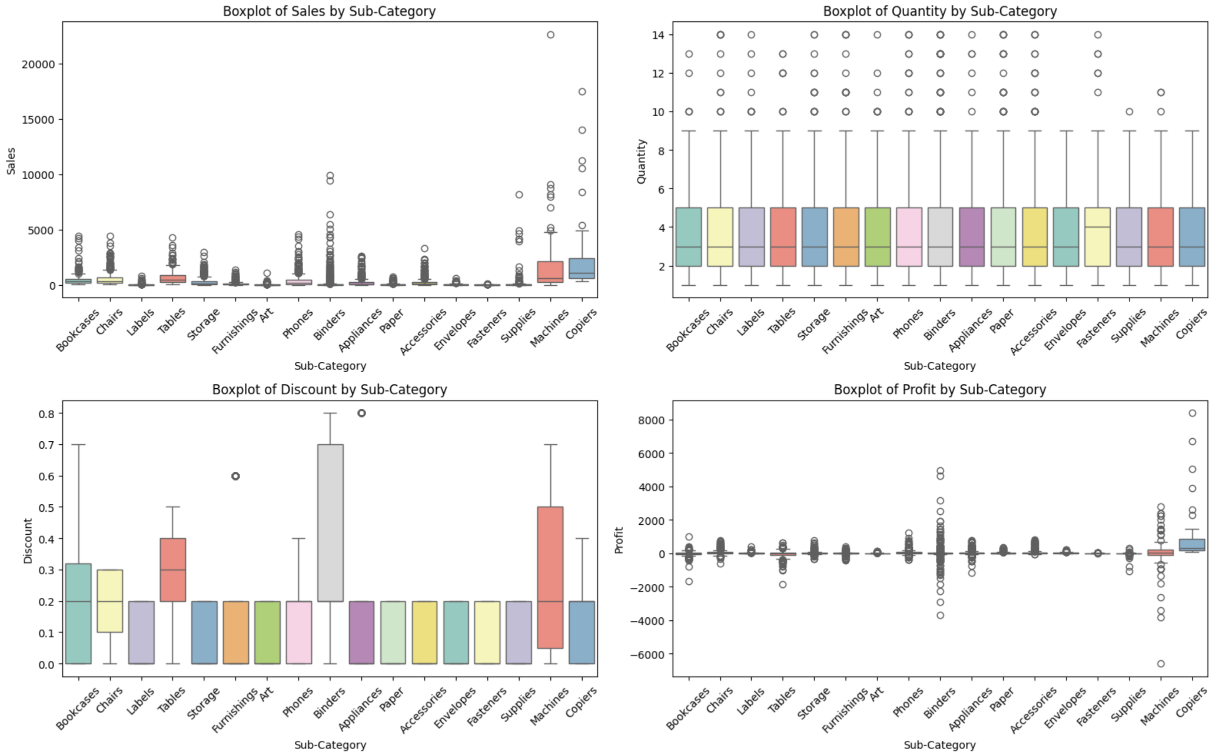


Sales by Sub-Category and Category Graph :

1. Furniture Category - **Chairs** and **Tables** generate the highest sales while **Furnishings** and **Bookcases** have comparatively lower sales

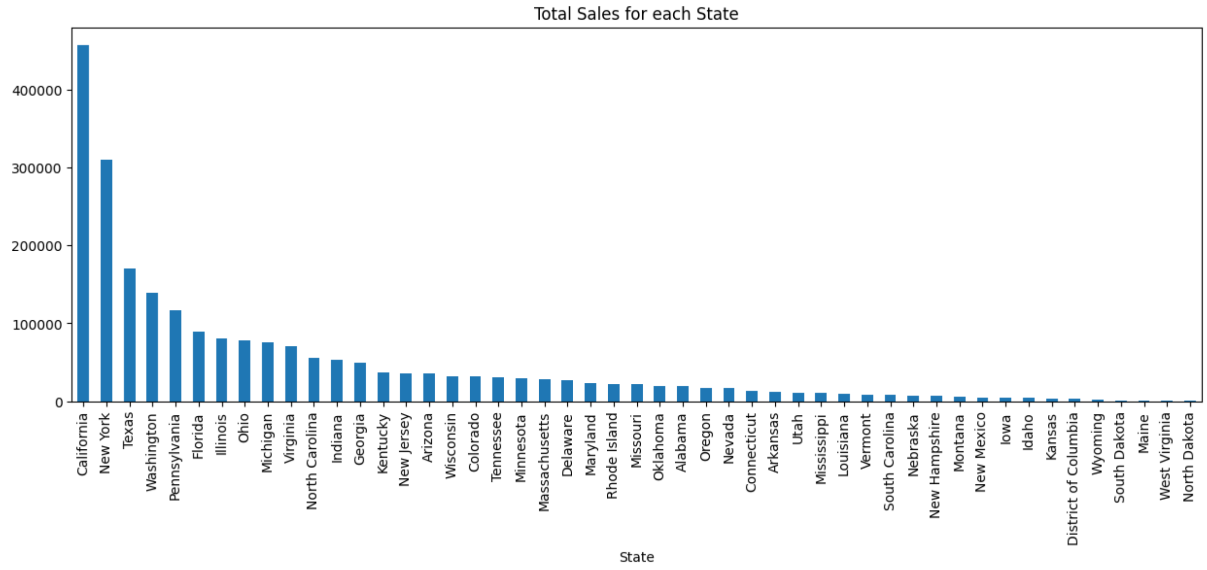
2. Office Supplies Category - **Binders** and **Storage** are the leading contributors to sales while **Fasteners**, **Labels**, and **Envelopes** show significantly lower volumes of sales.

3. Technology Category - **Phones** are the top-performing sub-category across categories and also perform better compared to the other 2 categories in terms of overall sales.



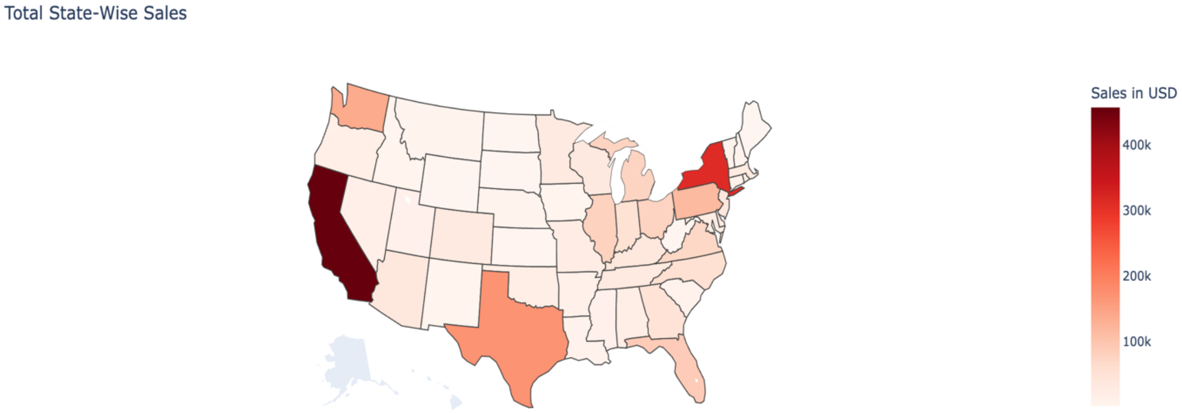
Insights :

- Copiers have very high variance in sales and has lots of extreme outliers
- Copiers also is a key driver of profitability with consistent profits despite variance
- Tables have some huge losses
- Phones and Binders have decent sales with relatively less outliers
- Most sub-categories have similar quantity with little variance
- There is consistent volume of sales across sub-categories
- Phones and Labels have low discounts indicating standard prices while Appliances have range of discounts thereby highlighting aggressive strategy in pricing



Insights :

1. **California and New York** have the highest sales amongst all states
2. **West Virginia and North Dakota** have the least sales.



Implications :

1. Superstore should **continue to focus on California,, New York and Texas** as they are the most revenue generating markets with highest demand for them
2. **Target** the markets of **Washington, Pennsylvania and Florida for growth** as these market have a lot of potential to grow further
3. There is a **risk of geographical concentration** as a majority of the revenue is generated from a few states – Superstore should start reducing dependency on top states by expanding in markets with rich consumers but that are still untapped

Recommendations :

1. **Tailor marketing strategy –** As regional preferences and demand vary, so customizing products and marketing strategies for different states can help in driving growth in local markets.
2. **Focused resource allocation –** Allocate more resources to high and medium performing states and reducing resources from areas that don't give much benefit.

5.3 Hypothesis Testing

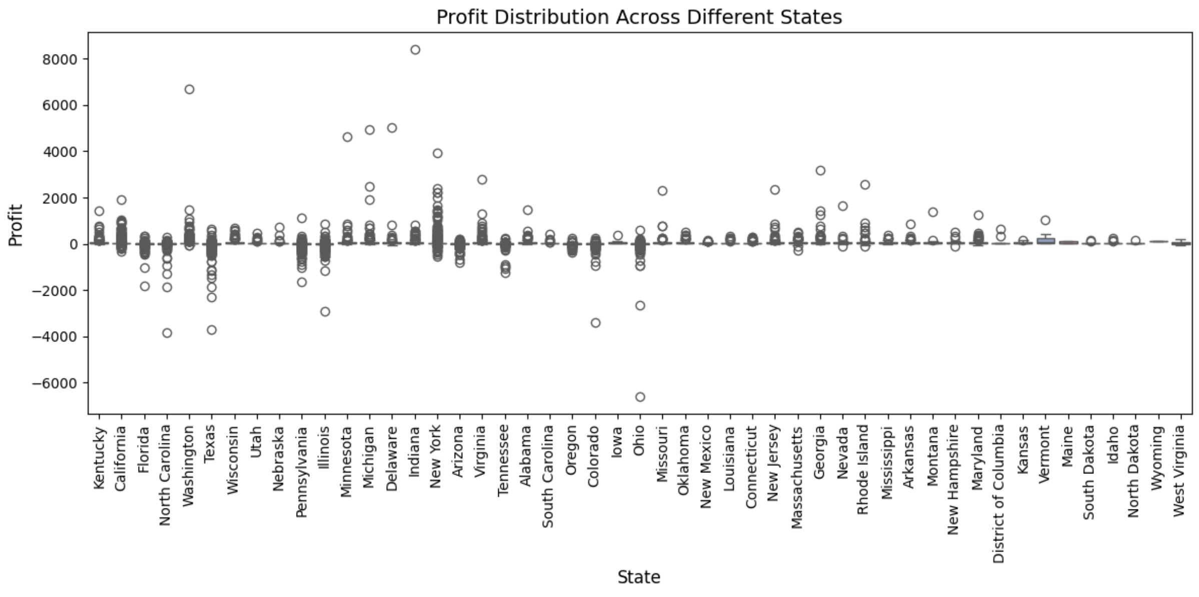
1. **Comparing Profit Across States**:
    - **Null Hypothesis (H_0)**: There is no significant difference in profit between different states.
    - **Alternative Hypothesis (H_1)**: There is a significant difference in profit between at least two states.

Test: One-way ANOVA or Kruskal-Wallis test (if data is not normally distributed).

<u>ANOVA Test Results:</u>

F-statistic: 8.821617512027952

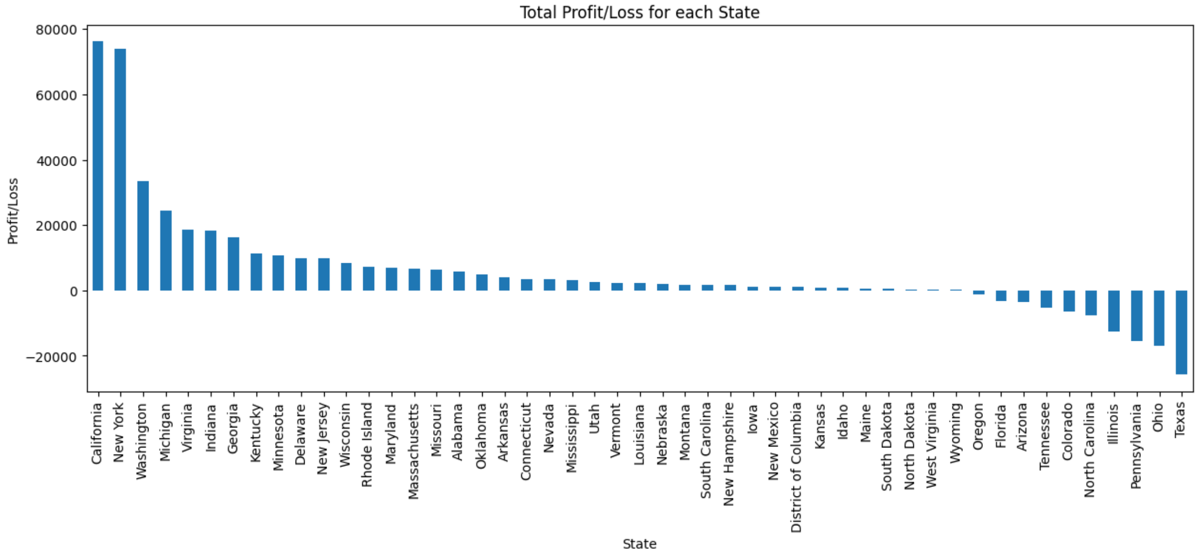
P-value: 5.011140011572623e-60 = 0.00



Kruskal-Wallis Test Results:

H-statistic: 2573.2102831302336

P-value: 0.0



We reject the Null Hypothesis (H₀): Hence, there is significant difference in Profit across states.

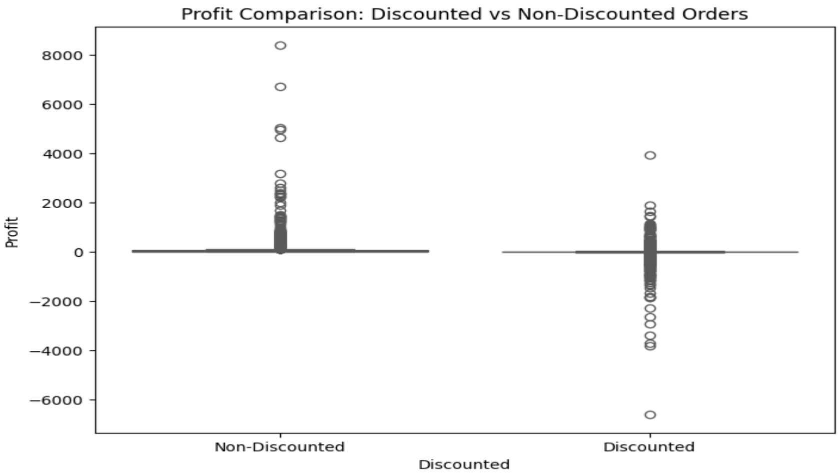
2. **Testing the Impact of Discounts on Profit**:
   - **Null Hypothesis (H₀)**: Discounts do not have an impact on profit.
   - **Alternative Hypothesis (H₁)**: Discounts do have an impact on profit.

Test: Independent t-test between discount and profit.

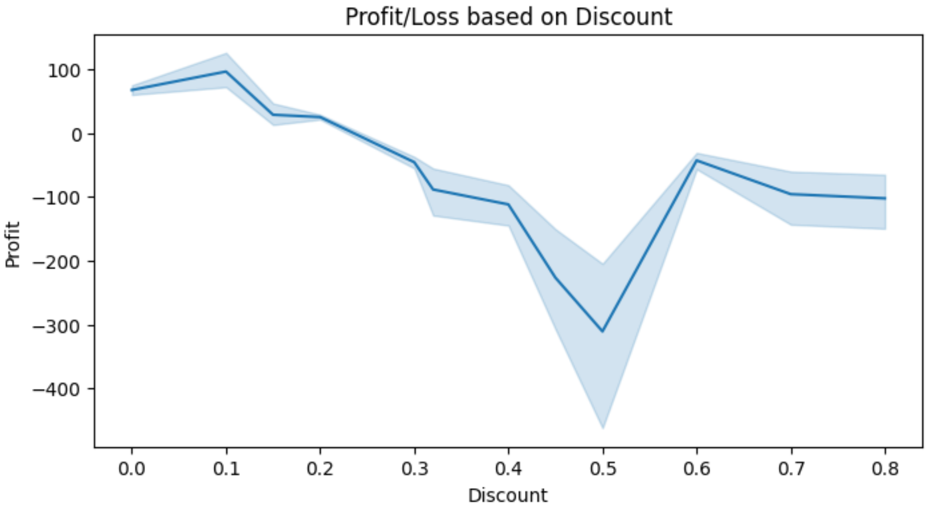
Independent T-test Results :

T-statistic: -15.88048735940772

P-value: 4.221733014595584e-56



We reject the Null Hypothesis: There is a significant difference in profit between discounted and non-discounted products.



- Profit is highest at **0% discount**, and starts decreasing as discount percentage increases.

- Between **20% to 30% discount**, the company moves closer to a break-even point, indicating that the profit margins are shrinking significantly

- After **30% discount**, the company starts getting losses

- **Maximum loss occurs around 50% discounts**

- After **50% discount**, the Superstore is unprofitable but there is some recovery and then becomes stable

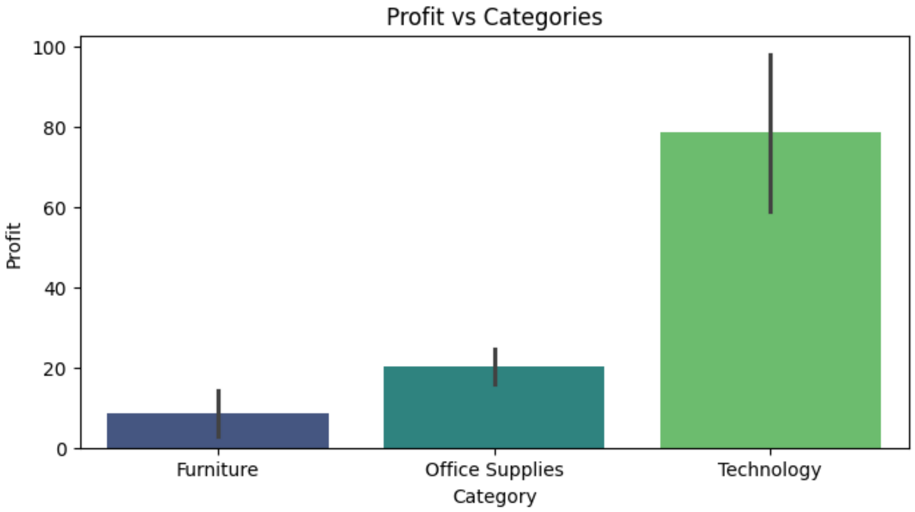
Business Insights :

1. To maintain profitability – It's better to stick to discounts below 20%
2. Higher discount ranges hamper the profit and would make it difficult for the Superstore to sustainability
3. Discounts above 50% can be used for clearing out unsold inventory and retaining customers once in a while
4. Need to better market the products to showcase their quality so that the sales are not affected by low discounts

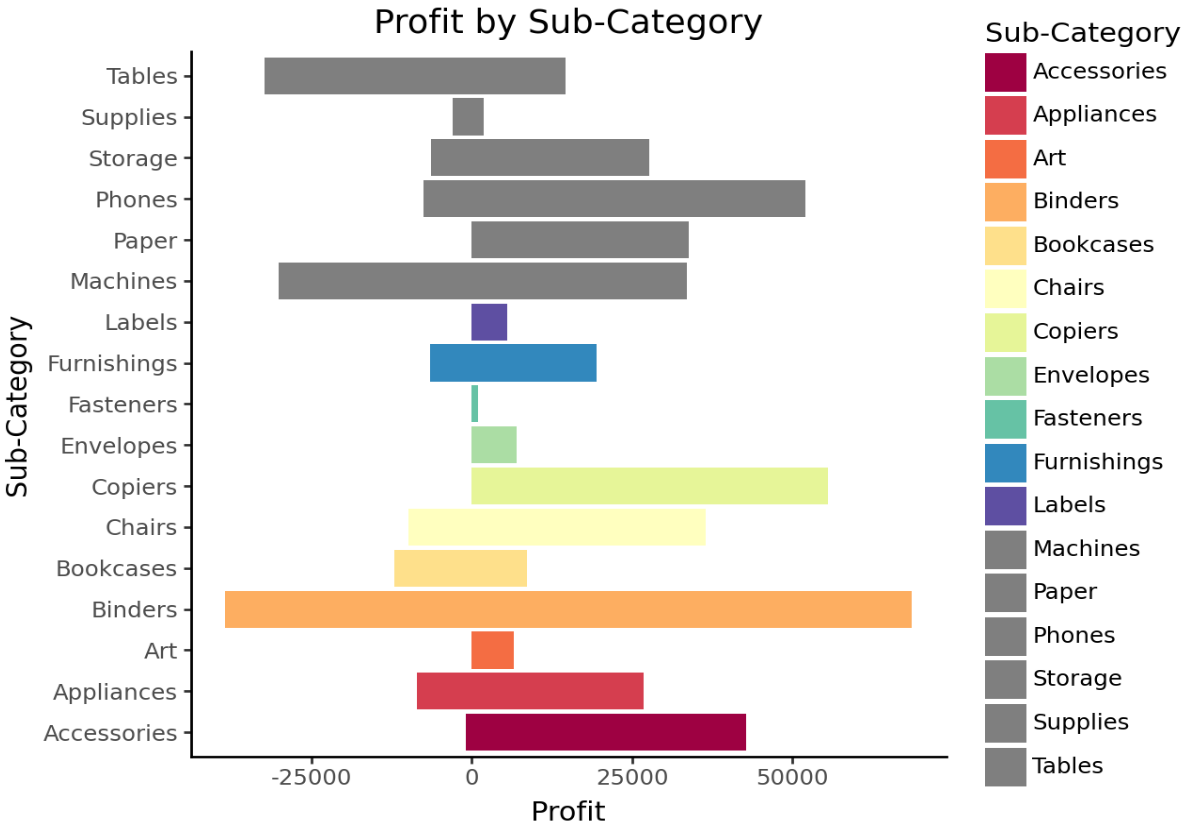
3. **Comparing Profit by Product Category**:

   - **Null Hypothesis (H_0)**: There is no significant difference in profit between different product categories.
   - **Alternative Hypothesis (H_1)**: There is a significant difference in profit between product categories.

Test: One-way ANOVA or Kruskal-Wallis test.



Hence, we reject the Null Hypothesis as there is significant difference in profit and different categories of products - Technology drives the highest profit amongst all categories



Implications :

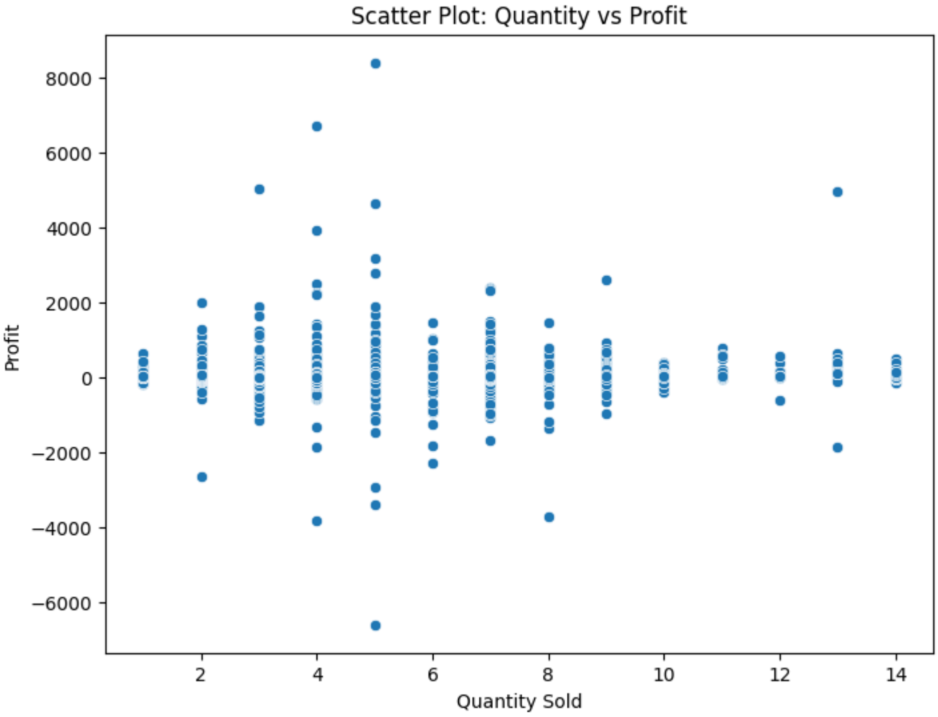
- Binders are critical to the profitability of Superstore
- Tables and Machines are drivers for loss
- Marketing team of the Superstore could focus on promoting the profitable sub-categories while addressing the operational and marketing issues causing low demand for the loss making sub-categories

4. **Testing the Relationship Between Quantity and Profit**:
    - **Null Hypothesis (H_0)**: There is no relationship between quantity sold and profit.
    - **Alternative Hypothesis (H_1)**: There is a relationship between quantity sold and profit.

Pearson correlation Test:

Pearson Correlation: 0.06608931737970732

p-value: 4.201697394619053e-11

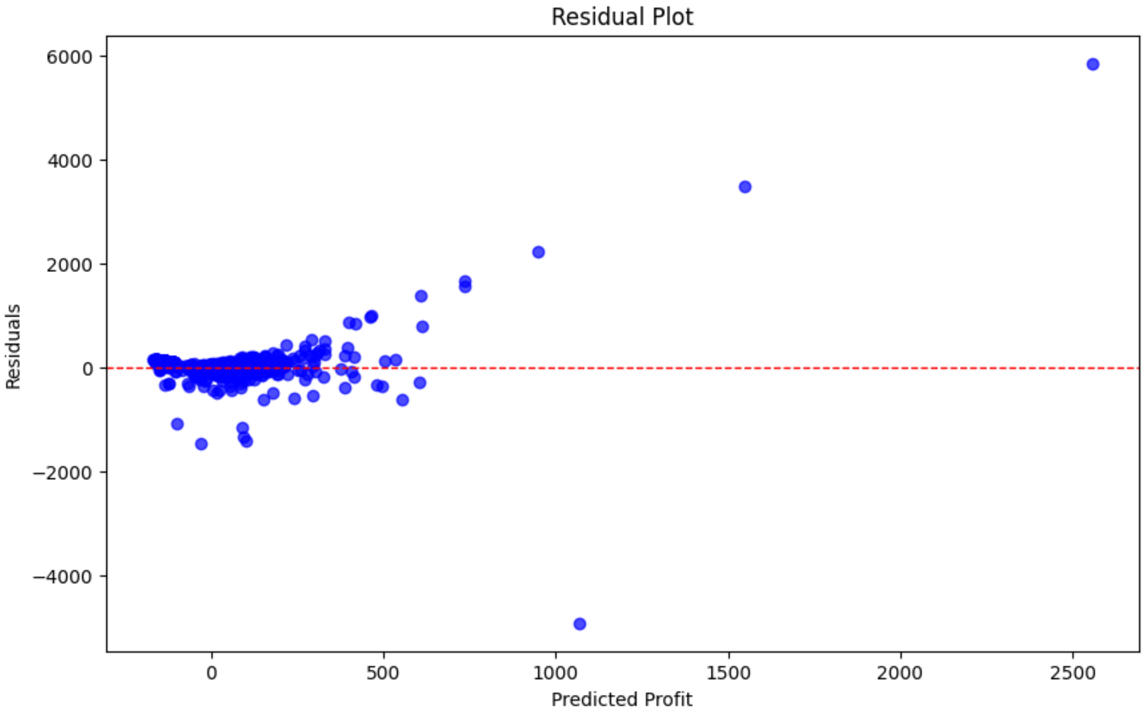


We reject the Null Hypothesis: There is a significant correlation between quantity and profit.

5.4 Regression Analysis

1. Linear Regression :

- **Coefficients**: [1.44326093e-01 -1.18761006e+00 -2.37639727e+02]
- **Intercept:** 35.89596583539688
- **Mean Squared Error (MSE):** 53968.30340230214
- **R-squared:** 0.3703106479096253

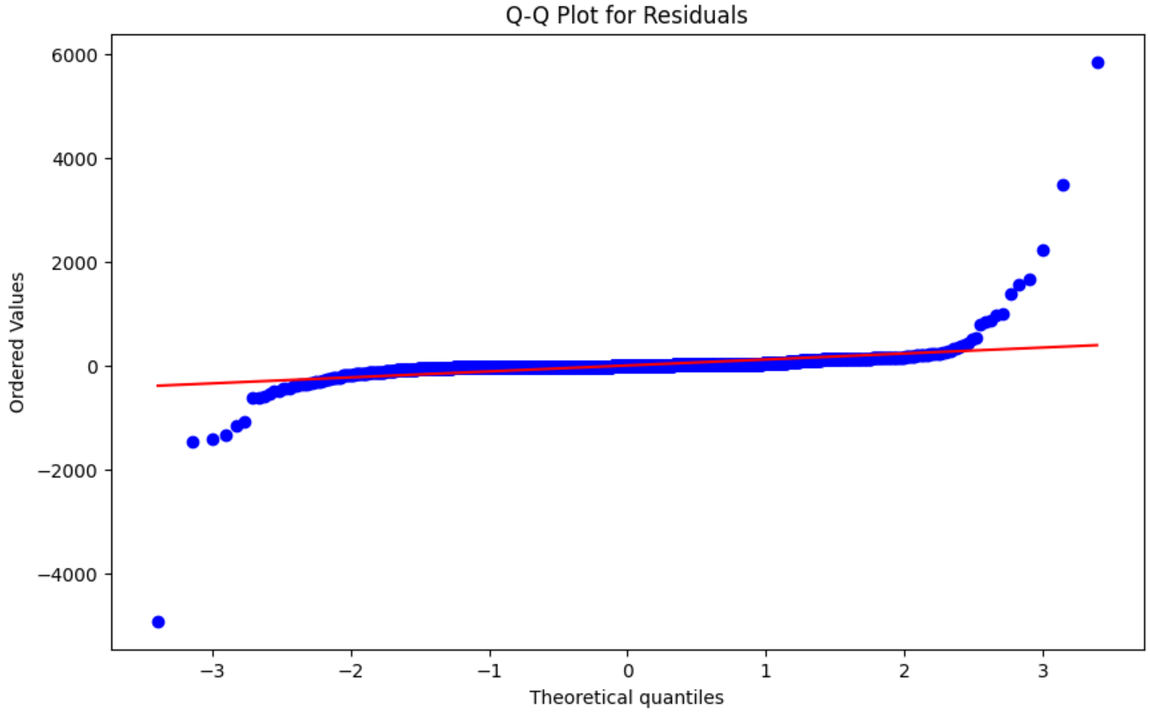


Residual Plot :

This plot visualizes the errors of the linear regression model, which are the differences between the actual and predicted values of the dependent variable (Profit).

Interpretation:

1. Ideally, the residuals should be randomly distributed around 0, indicating no clear pattern, which suggests that the model's assumptions are valid.
2. We can see that the residuals tend to spread out as the value of predicted profit increases, this indicates potential variance of errors that is not constant.



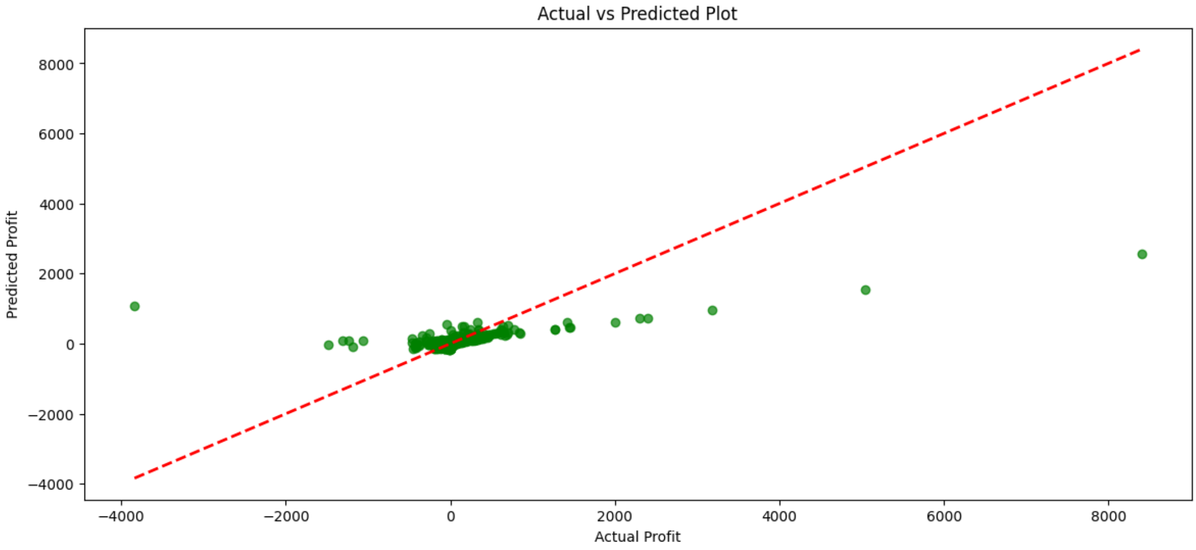
Q-Q Plot :

This quantile-quantile (Q-Q) plot checks if the residuals follow a normal distribution.

Insight from the graph:

Since the residuals follow a straight line, similar to the red line shown in the above graph, they are approximately normally distributed.

There are some deviations from the red line, at the ends, indicating that the residuals may not be properly normally distributed – especially due to outliers.



Actual vs. Predicted Plot :

The scatter plot compares the actual values of Profit to the predicted values.

Insight from the graph:

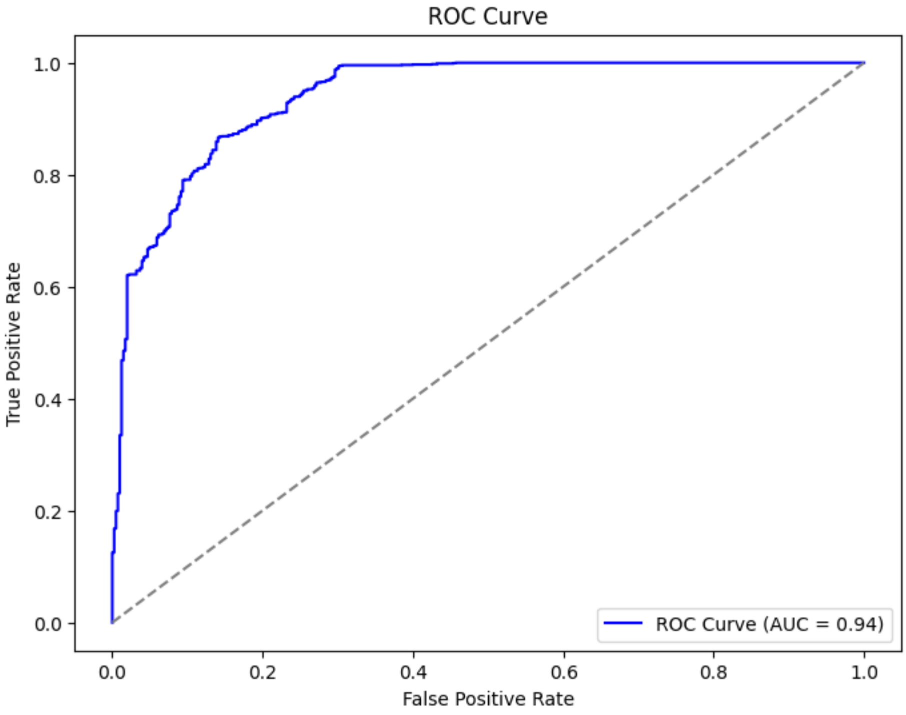
- The red dashed line represents the ideal case when the actual values are equal to the predicted values.
- The points clustered near the red line indicate good predictions.
- Points far from the line represent large prediction errors.
- Model faces some challenges with predicting some data points accurately, possibly due to outliers.

Recommendations:

- We need to look into the outliers that are affecting the model's performance.
- We can try to do certain transformations by taking log or square roots of the dependent variable or independent variables to help with the non-constant variance of errors.
- We can explore advanced models like Ridge, Lasso, or non-linear regression to further analyse the data.

2.Logistic Regression :

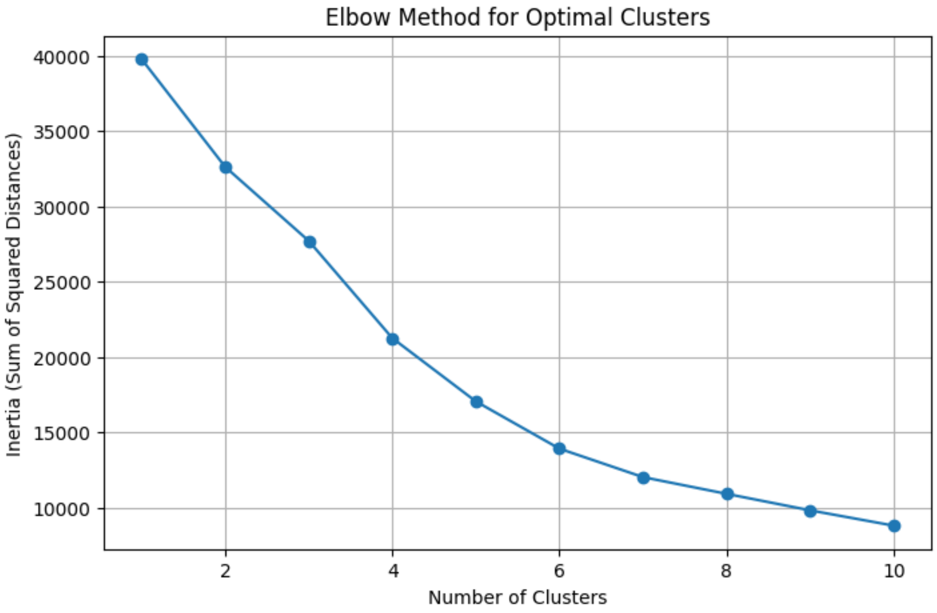
- The model achieved an **AUC of 0.94**
- It effectively classified high-profit sub-categories based on input features.



Insights from the ROC Curve :

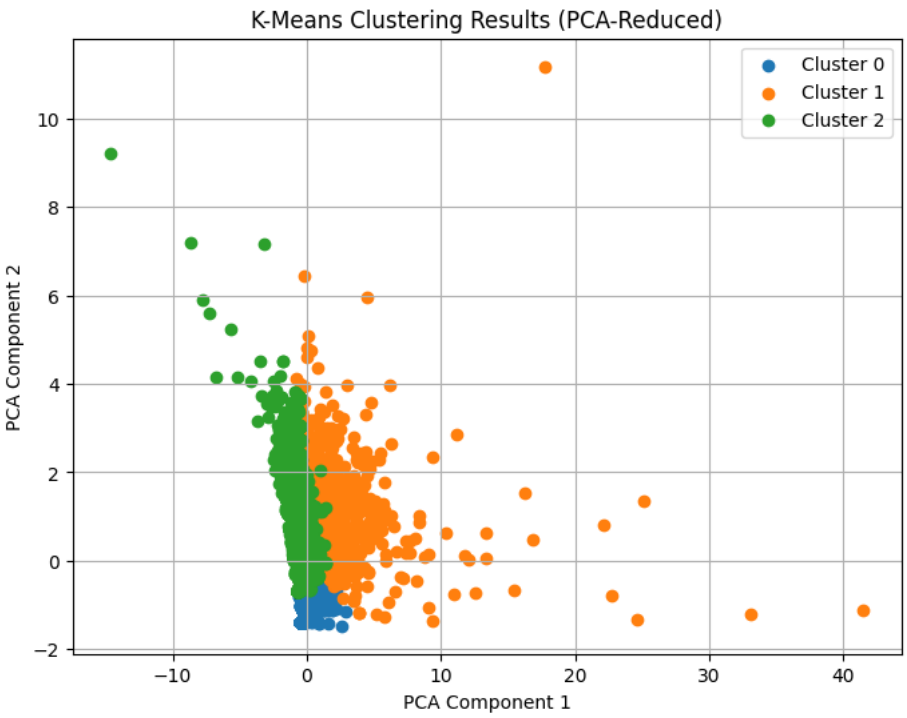
1. **Performance Evaluation**:
   - The **AUC (Area Under Curve)** value of 0.94 indicates an **excellent model performance**
   - It shows that the model is highly capable of distinguishing between the classes.
2. **False Positive vs. True Positive Trade-Off**:
   - The steep rise in the curve shows that the model achieves a high **true positive rate** and a low **false positive rate** for the majority of the threshold range.
   - It can be said that the model is effective at making correct predictions while minimizing incorrect ones.

5.5 Clustering



- We used Elbow method to find the optimal number of clusters (k) for our k-means clustering algorithm.
- The graph shows a steep decrease in inertia around **k=3** and slows significantly forming an elbow
- Therefore **k=3** is our optimal choice for the number of clusters.

We further run the PCA reduced K-means clustering and Principal Component Analysis (PCA) helps in reducing high-dimensional data into two components for visualization.



Insights from the graph :

The data points are grouped into three distinct clusters (Cluster 0, Cluster 1, and Cluster 2), similar to our analysis of elbow method.

Cluster Interpretation:

1. **Cluster 0 (Blue)**: The cluster is located close to the origin and is quite compact and dense, indicating similarity among the data points within this cluster
2. **Cluster 1 (Orange)**: Its spread out, indicating more variation in this cluster compared to others.
3. **Cluster 2 (Green)**: Its slightly overlapping with Cluster 1 and is a bit sparse, indicating that some data points are close in feature space but assigned to different clusters.

Cluster Separation:

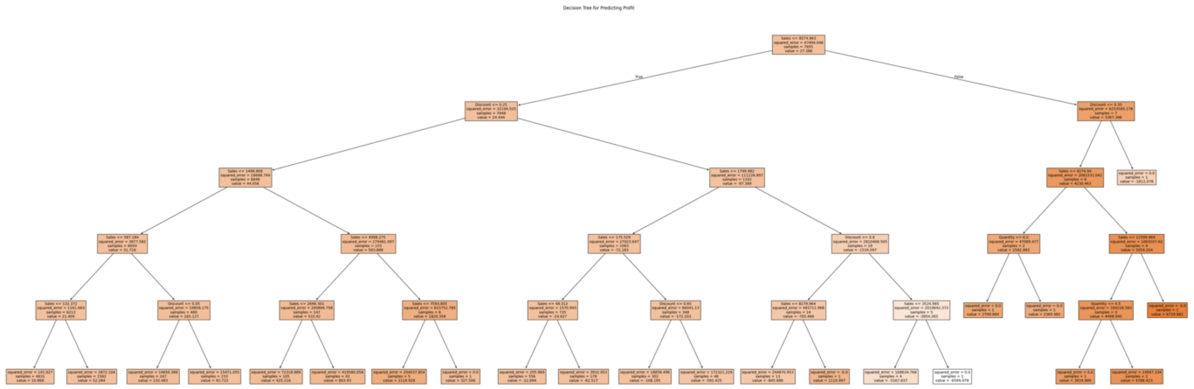
There is a decent amount of separation between the 3 clusters, although we see some overlap between Clusters 1 and 2. This could indicate that there are some shared features or similar data points between these clusters.

Analysis :

We can try to analyse the clusters and the type of segments they might represent :

1. **Cluster 0**: This is potentially Superstore's core group with consistent behaviour.
2. **Cluster 1**: This group has high variability and possibly consists of high-value customers or diverse products.
3. **Cluster 2**: These could be outliers or unique segments needing targeted strategies.

5.6 Decision Trees

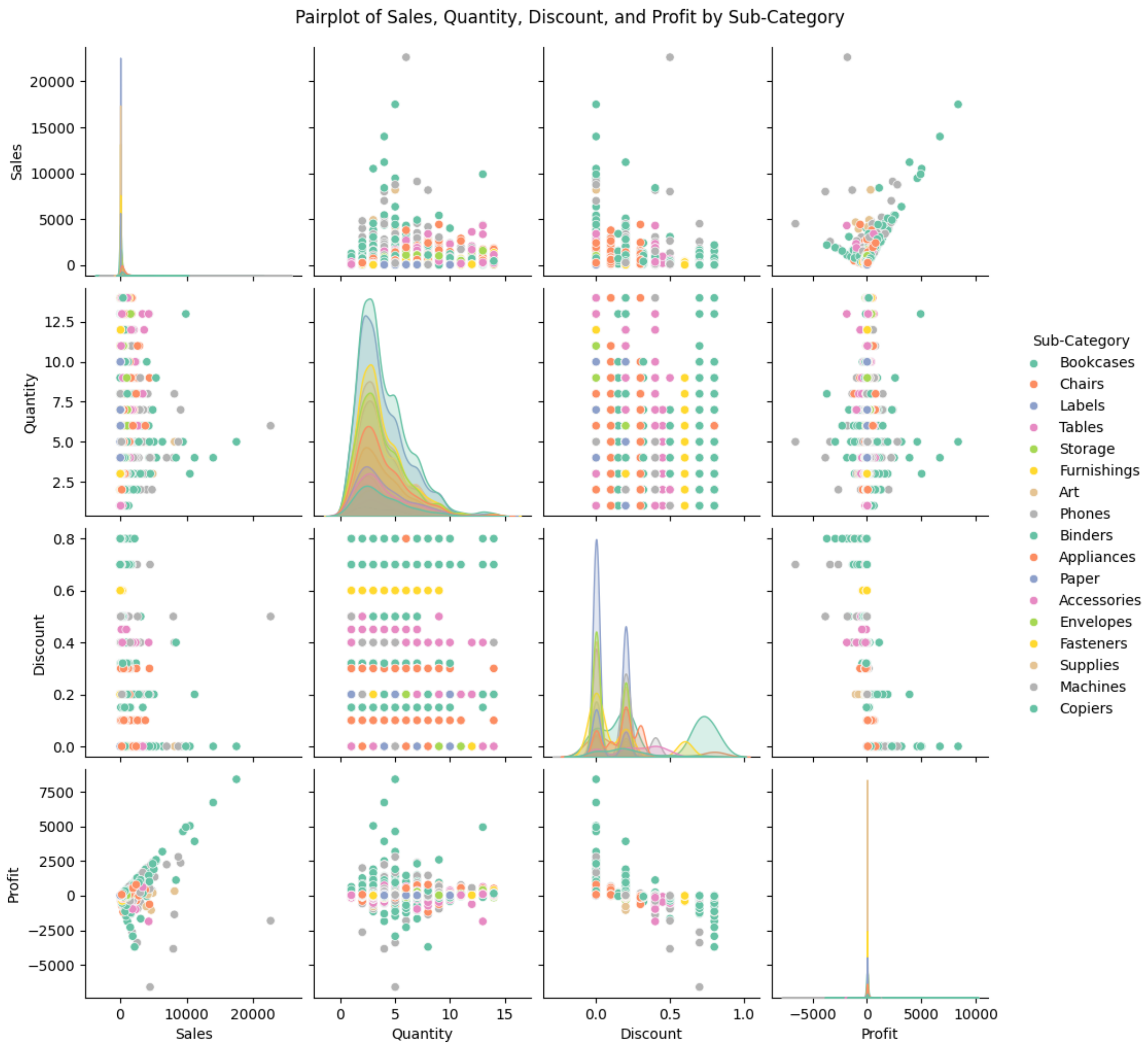


Insights from the Decision Tree:

1. **Top-level Split**:
   - The decision tree's root node splits on **Sales <= 1438.808** indicating that Sales is the most important variable influencing Profit at the highest level.
2. **Secondary Splits**:
   - For Sales <= 1438.808, further splits on **Discount** in this branch, showing the role of discounts in influencing profitability.
   - For Sales > 1438.808, Discount plays a significant role, followed by Sales and Quantity.
3. **Influence of Discount**:
   - Several splits are driven by the Discount feature. This indicates a strong relationship between discounts offered and profitability.
   - Higher discounts likely reduce profit margins, as seen in branches dominated by Discount > 0.5.
4. **Quantity Impact**:
   - The variable Quantity appears less frequently but contributes to splits where profitability depends on the number of units sold, particularly in the high Sales branches.
5. **Profitability Factors**:
   - Sales is a primary driver of profitability.
   - Offering high discounts can negatively impact profits.
   - In some cases, selling higher quantities offsets the impact of discount, leading to higher profits.

6. Discussion

6.1 Analysis of overall relationship between the metrics :



Insights from Pair plot :

1. **Sales vs. Profit**:
   - A positive correlation exists, as higher sales generally lead to higher profits.
2. **Discount vs. Profit**:
   - Higher discounts often result in reduced profits.
3. **Quantity vs. Sales**:
   - Higher quantities don't always translate to higher sales.
4. **Quantity vs. Profit**:
   - Selling larger quantities does not guarantee higher profitability. Superstore needs strategic focus on high-margin products.

6.2 Implications of Results on our Research Questions :

The analysis highlights actionable insights for **sales and profit optimization**:

1. **Targeted Marketing:**
    - Focus on high-value customers (Cluster 1) with personalized offers.
    - Offer discounts on frequently co-purchased items like **Phones** and **Accessories**.
2. **Regional Strategies:**
    - Investment in the **South region on marketing would benefit**, since profits are lagging in these areas.
    - Focus on catering to high-performing regions like the West to maximize our ROI.
3. **Category Optimization:**
    - Discounts on sub-categories like **Tables** that cause huge losses needs to be reduced or removed.
    - **Copiers** and **Phones** are the hero products and need more push to further maximise our profits.

6.3 Limitations

- Lack of detailed competitor analysis.
- Lack of time-series data restricted the use of ARIMA forecasting.
- Sequential data limitations didn't allow for a more robust Markov Chain model.

6.4 Recommendations for Future Research

- Incorporate time-series data for forecasting based analysis.
- Explore sentiment analysis to understand customer reviews and preferences by including rating given to these products.
- Expand analysis to include global markets instead of just the United States.

7. Conclusion

This study demonstrated how diverse business analytics techniques can help optimize sales and profitability for a Superstore. It provides a comprehensive framework for identifying profit drivers, effect of discount and promotional offers on sales and profit, and regional prosperity leading to higher purchasing power of consumers of that geographical location leading to higher profit margins in particular areas. These insights contribute to more informed decision-making for the companies trying to make effective marketing strategies, improving resource allocation, and thereby improving profitability and operational efficiency.

Objective Achieved :

We guided the Superstore in identifying high-performing categories and underperforming areas, refining discount strategies.

Research Questions Answered :

We got the answers to our research problems :

1. Which product categories and sub-categories drive sales & profitability?
    - **Technology Category**
    - **Phones Sub-Category**
2. What is the sales and profit distribution across all US states?
    - **California, and New York** have highest sales.
    - **West Virginia and North Dakota** have the least sales.
3. What is the impact of discounts on sales and profit?
    - **Discounts less than 20% help the business stay profitable**
4. How can we optimize and improve the current sales & profit?
    - **Focusing on profit generating regions, categories and sub-categories while giving minimum discounts (<20%)**

8. References

1. Davenport, T. H., & Harris, J. G. (2017). *Competing on Analytics: The New Science of Winning,* 84(1):98-107, 134. PMID: 16447373
2. Kumar, D., & George, D. (2021). *Business Analytics for Decision Making.* http://dx.doi.org/10.4018/IRMJ.291693
3. Kumar, R., & Smith, J. (2022). Demand forecasting in retail using ARIMA models. Journal of Business Analytics, 15(3), 234-245. http://dx.doi.org/10.2478/mmcks-2020-0012
4. Aman Sharma. (2020). Sample Superstore Dataset**.** Version 1. Retrieved Novemeber 20, 2024 from https://www.kaggle.com/datasets/bravehart101/sample-supermarket-dataset/data