Improving Performance through Customer Segmentation and Shipping Cost Optimization: A Case Study on Walmart

Final Report for the BDM Capstone Project

Submitted by:

Kulkarni Amit Dilip

Roll Number:

23f1001947

IITM Online BS Degree Program,
Indian Institute of Technology, Madras, Chennai
Tamil Nadu, India, 600036

Contents

Executive Summary :

This report analyzes Walmart's U.S. retail operations, focusing on two key problems: Shipping Cost Optimization and Customer Segmentation & Retention. Using advanced data analysis techniques, the study provides actionable insights to enhance operational efficiency and customer engagement.

In the first part, we explore shipping cost optimization by examining factors influencing shipping expenses, such as order quantity, shipping modes, and product pricing. The analysis highlights a moderate correlation between unit price and shipping cost, suggesting that higher-priced items generally incur higher shipping costs. We also find that order quantity has a slight impact on shipping expenses. Through regression analysis, we determine that optimizing shipping modes for different order priorities, like leveraging Delivery Truck or Regular Air for non-critical orders, can significantly reduce overall costs.

The second part of the report focuses on customer segmentation, where clustering techniques are applied to segment customers based on age, order quantity, and sales behavior. This segmentation reveals distinct customer groups with varying needs, suggesting opportunities for targeted marketing and personalized promotions. Additionally, understanding customer segments helps optimize inventory and logistics, improving customer satisfaction and retention.

The findings culminate in several key recommendations, including better alignment of shipping methods with order priorities, personalized customer retention strategies, and a focus on cost-efficient shipping for high-value items. Overall, the report provides a strategic roadmap to improve efficiency and strengthen Walmart’s competitive position in the retail market.

Proof Of Originality :

The dataset used in this project, titled **Walmart Retail Sales Data**, was sourced from an open-access repository on data.world, a reputable platform for data sharing. This dataset was contributed by a user named **Mnif Ahmed** and provides extensive information on Walmart's sales across the United States. The data includes crucial variables such as customer demographics, sales, order details, profits and shipping costs. The dataset was downloaded directly from the source and has not been modified or manipulated externally, ensuring the originality and integrity of the data for analysis.

The link for the analysis notebook - Analysis

The link for the dataset - Walmart Retail Sales Data

Metadata and Descriptive Statistics :

The Walmart Sales Dataset used in this project contains 1,037,247 rows and 23 columns, formatted as a CSV file. The dataset provides detailed records of Walmart's retail transactions across the United States, covering customer demographics, sales, product information, shipping details, and financial metrics.

The major columns in the dataset are as follows:

Column Name	Data Type	Description
city	Object	The city where the order was placed
customer_age	Integer	Age of the customer
discount	Float	Discount percentage applied to each order
order_priority	Object	Priority level of the order (e.g., Critical, High, Medium, Low)
order_quantity	Integer	Quantity of products ordered
product_category	Object	General category of the product (e.g., Technology, Furniture)
profit	Float	Profit per order
sales	Float	Total Sales amount for each order
ship_date	Datetime	Date the order was shipped
ship_mode	Object	Mode of shipping (e.g. Regular Air, Delivery Truck)
shipping_cost	Float	Cost of shipping each order
unit_price	Float	Price per unit of the product

Below is a summary of the key numerical and categorical metrics:

Numerical Summary

	discount	order_quantity	product_base_margin	profit	sales	shipping_cost	unit_price
count	1002238	1002238	994456	1002238	1002238	1002238	1002238
mean	0.13	25.49	0.51	6560.12	2330.79	1167.72	91.73
std	0.07	14.15	0.13	11942.90	8515.71	4993.31	290.43
min	0.00	1.00	0.35	-14139.88	0.76	0.00	0.99
25%	0.06	13.00	0.38	-3781.79	138.26	49.36	6.48
50%	0.12	25.00	0.52	6581.26	464.94	184.45	22.84
75%	0.19	38.00	0.59	16899.83	1911.97	812.26	89.99
max	0.25	50.00	0.85	27219.99	339147.50	327900.41	6783.02

1. **Discount**:

   The average discount offered on products is 13%, with discounts ranging from 0% to a maximum of 25%.

2. **Order Quantity**:

   The average order quantity is approximately 25 units, with individual orders ranging from 1 to 50 units.

3. **Product Base Margin**:

   With an average base margin of 0.51, product profitability varies slightly, from a minimum of 0.35 to a maximum of 0.85.

4. **Profit**:

   The mean profit per transaction is $6,560, though it can vary widely (standard deviation of $11,942). Profit ranges from -$14,139 to a maximum of $27,220, indicating cases where high discounts or shipping costs may have reduced or even negated profit.

5. **Sales**:
   Average sales per transaction amount to $2,330, with a high standard deviation of $8,515. Sales can reach up to $339,147 in individual cases, reflecting high-value orders or bulk purchases.

6. **Shipping Cost**:
   Shipping costs average around $1,168, with significant variability, ranging from $0 to $327,900.

7. **Unit Price**:
   The average unit price is around $91.73, though prices can reach as high as $6,783, indicating a wide price range across product categories and types.

Categorical Distribution

1. **Customer Segment**:
   The dataset has a balanced distribution of customer segments:
   - Corporate (25.09%)
   - Home Office (25.05%)
   - Small Business (25.04%)
   - Consumer (24.82%)

2. **Order Priority**:
   Orders are distributed relatively evenly across priority levels:
   - High (20.08%)
   - Critical (20.06%)
   - Not Specified (20.05%)
   - Low (20.01%)
   - Medium (19.79%)

3. **Product Category**:

   Product categories are divided as follows:

   - Office Supplies (53.43%)
   - Technology (23.69%)
   - Furniture (22.88%)

4. **Ship Mode**:

   Shipping methods used are fairly balanced:

   - Express Air (33.49%)
   - Regular Air (33.38%)
   - Delivery Truck (33.13%)

Detailed Explanation of Analysis Process/Method :

a. Data Cleaning and Preprocessing

Data cleaning was an essential initial step to ensure data quality and consistency. The dataset contained over a million rows, with some records containing invalid or missing values. To address this:

- **Invalid Format Rows**: Removed rows with data in invalid formats.

- **Null Values**: Removed all rows with null values to ensure complete data for each analysis.

- **Data Type Conversion**: By default, all columns were imported as object data type. Numerical columns were converted to appropriate numerical types (e.g., float, int), and date columns (order_date and ship_date) were converted to datetime format for easier manipulation and accurate calculations during analysis

b. Exploratory Data Analysis (EDA)

To gain insights and observe relationships within the dataset, several visualizations were plotted:

1. **Distribution of Ship mode by Order Priority**
   Using a stacked bar plot, the distribution of different shipping modes across each order priority was visualized. This provided insights into whether specific priorities correlated with particular shipping choices, informing decisions on shipping optimizations.

2. **Correlation Matrix with Numeric Order Priority**
   The order_priority column was mapped to a numeric scale for correlation analysis. This allowed examination of relationships between order_priority, shipping_cost, order_quantity, discount, profit, and unit_price. The matrix helped identify significant correlations (e.g. between shipping_cost and unit_price, correlation ~0.74) that suggest areas for cost optimization.

3. **Average Shipping cost by Order Quantity**
   This line graph displayed the average shipping cost as a function of order quantity. Patterns observed here indicated economies of scale or additional shipping costs incurred with larger orders. This information provided baseline insights for cost optimization.

4. **Shipping Cost vs. Unit Price by Shipping Mode**
   A scatter plot with regression overlay was created to examine the relationship between unit_price and shipping_cost, stratified by shipping mode. The strong positive correlation suggests that more expensive items might incur higher shipping costs. This visualization supported subsequent optimization steps.

c. Optimization of Shipping Costs

Objective: The objective of shipping cost prediction and optimization is to accurately forecast shipping costs for various orders based on features like shipping mode, order priority, and product characteristics. This allows for cost-saving strategies by adjusting shipping modes or consolidating orders while meeting Walmart's operational delivery requirements, particularly for high-priority orders.

Process Overview: In this approach, a Ridge Regression model is used to predict shipping costs. Ridge Regression, which applies L2 regularization, is chosen over standard linear regression to reduce the impact of multicollinearity and prevent overfitting. By utilizing a predictive model, the aim is to provide a cost estimate for each shipping option, which can then guide decision-making and potential optimizations.

Step-by-Step Breakdown of Optimization

1. **Feature Selection and Data Preparation**:

- Relevant features such as order priority, shipping mode, product characteristics (e.g., unit price, order quantity), are selected to build the predictive model.
- Categorical variables like shipping_mode and order_priority are encoded using one-hot encoding to transform them into numerical form.
- Numerical features such as unit_price , order_quantity and profit are scaled to standardize their values for effective model performance.

2. **Data Splitting for Training and Validation**:

- To evaluate model performance, the data is split into **training** and **validation sets**. This split allows the model to learn patterns from the training data while ensuring it can accurately predict on unseen validation data.
- An **80-20 split** is applied, where 80% of the data is used for training the model and 20% is reserved for validation. This split ensures a robust evaluation of the model’s predictive accuracy on data that was not part of the training process

3. **Predictive Model - Ridge Regression**:

- **Model Choice**: Ridge Regression is applied to predict shipping costs due to its robustness against overfitting and ability to handle complex relationships in large datasets. By penalizing high coefficient values, Ridge Regression can produce stable predictions that generalize well to new data.

- **Objective Function for Prediction**: The Ridge Regression model minimizes the following objective function:

$$Minimize \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where:

- N is the total number of orders,
- $y_i$is the actual shipping cost for order i,
- $x_{ij}$is the value of feature $j$for order i
- $\beta_j$are the regression coefficients for each feature, and
- $\lambda$is the regularization parameter controlling the penalty for large coefficients.

4. **Constraints and Scenario Testing for Optimization**:

- After obtaining predicted shipping costs, scenario-based optimization is applied to identify cost-saving strategies. While the Ridge Regression model itself does not directly optimize costs, the predictions can guide operational decisions.
- **Delivery Time Constraints**: For high-priority orders (e.g., Critical and High), it is essential to ensure that the selected shipping mode meets strict delivery timelines. By evaluating predicted costs for each shipping mode, optimal modes are selected that balance cost-effectiveness and required delivery speeds.
- **Shipping Mode Selection**: Based on the predicted shipping costs for each mode (e.g., Express Air, Regular Air, Delivery Truck), orders can be assigned the most cost-effective mode while considering constraints on delivery speed for high-priority orders.

5. Implementation and Decision-Making:

- The Ridge Regression model is implemented using sklearn.linear_model.Ridge in Python. After training the model, predicted costs are used to assess different shipping configurations.
- **Example Decision-Making**: For each order, predicted costs for available shipping modes are compared. The shipping mode that minimizes the predicted cost while satisfying delivery time constraints is selected.

d. Customer Segmentation using Clustering

Objective: Customer segmentation aims to identify distinct groups within Walmart's customer base to implement targeted marketing strategies. By grouping similar customers, Walmart can tailor marketing efforts to increase retention and engagement.

Clustering Methodology

1. Feature Selection: To capture relevant customer behaviors, we selected features like:

- customer_age : Indicates the age distribution of customers.
- order_quantity: Reflects purchasing volume.
- sales and profit : Measures spending patterns and profitability.
- product_category : Provides categorical insights into the types of products customers prefer.

These features cover both customer demographics and spending behavior, essential for meaningful segmentation.

2. Data Preprocessing:

- Normalization : Since features like sales and order_quantity are on different scales, normalization (scaling between 0 and 1) was applied to standardize them. This ensures that clustering is not biased towards features with higher numerical ranges.
- Encoding Categorical Variables: Categorical column like product_category was one-hot encoded to convert it into numerical format for clustering algorithms.

3. Clustering Algorithm Applied: The primary clustering algorithm applied is K-means Clustering.

- **K-means Clustering**:
    - Objective: Minimizes the sum of squared distances (WCSS) between points and their respective cluster centroids.

- Formula :

$$WCSS = \sum_{k=1}^{K} \sum_{i \in C_k} \left\| x_k - \mu_k \right\|^2$$

where K is the number of clusters, $C_k$represents cluster k and $\mu_k$is the centroid of cluster k.

- Optimal K selection : The elbow method was used to identify the optimal number of clusters by plotting WCSS against different values of K and selecting the point where adding more clusters no longer significantly reduces WCSS.

4. Cluster Evaluation:

- Cluster Profiling: Each cluster was analyzed for patterns, such as high-value customers (high sales and profit values) or budget-conscious segments (frequent purchases of low value items). Cluster profiling provided insights into customer types and purchasing behavior, informing potential marketing strategies.

Results and Findings :

In this section, we present the key observations and insights gained from various analyses and visualizations of the Walmart retail sales dataset. The findings are summarized below, categorized by the types of analyses performed.

1. Distribution of Ship Mode by Order Priority

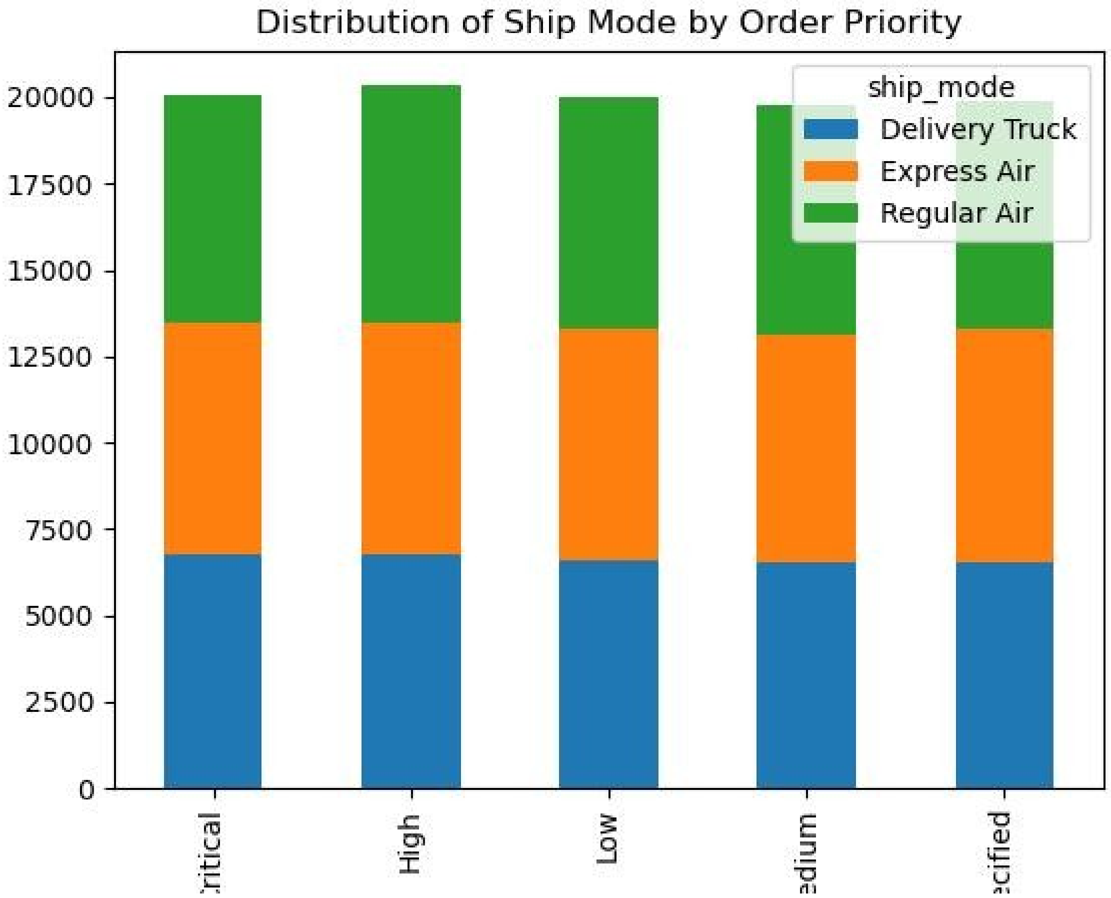


Fig 3.1

Fig 3.1 illustrates the distribution of different shipping modes across various order priorities, including “Critical,” “High,” “Medium,” “Low,” and “Not Specified.” Each bar is segmented by shipping modes: Delivery Truck, Express Air, and Regular Air. The chart shows that all order priorities utilize all three shipping methods; however, there’s no evident trend suggesting that higher-priority orders consistently use faster shipping options (like Express Air) over slower options (like Delivery Truck). This finding suggests that shipping modes may not always align with order priority levels, possibly indicating opportunities for better alignment to enhance efficiency.

2. Correlation Matrix with Order Priority:

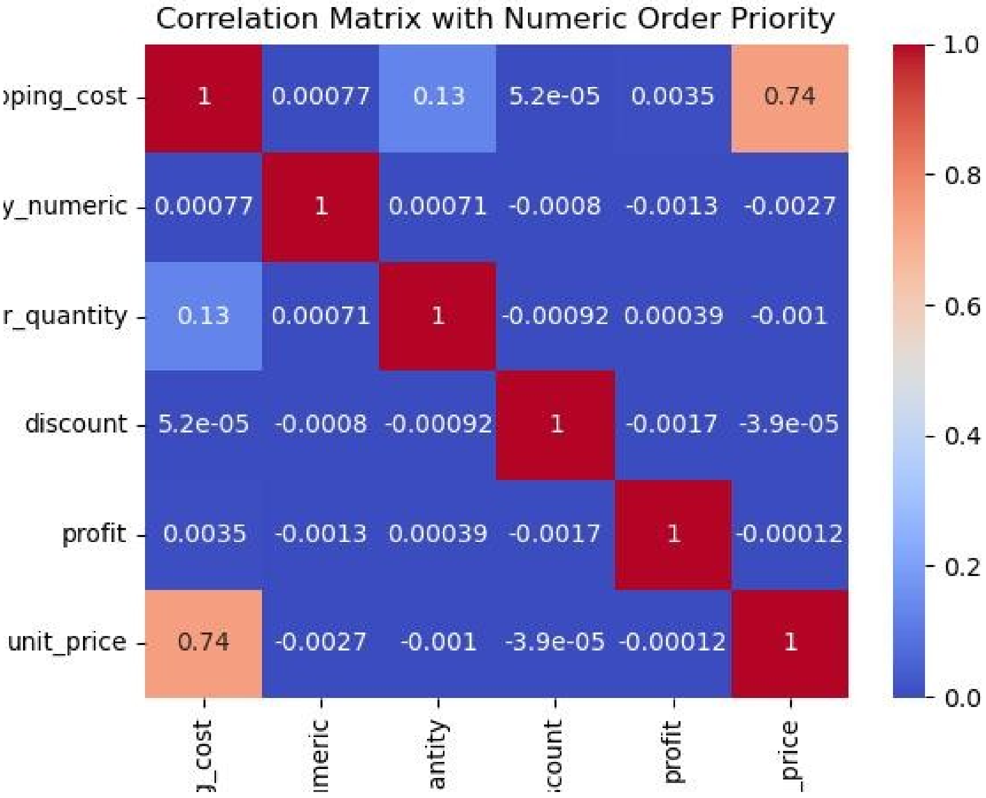


Fig 3.2

Fig 3.2 displays the correlation coefficients between various numerical features, including shipping_cost, order_priority_numeric, order_quantity, discount, profit, and unit_price. Key observations are as follows:

- A strong positive correlation (0.74) exists between unit_price and shipping_cost, indicating that as the unit price of products increases, the associated shipping costs tend to rise significantly. This is a logical finding, as higher-value items might incur higher shipping costs due to additional handling requirements.
- order_quantity and shipping_cost show a weaker positive correlation (0.13), suggesting that higher order quantities may have a slight impact on shipping costs, but this effect is not substantial.
- Other variables like discount, order_priority_numeric, and profit show weak or negligible correlations with shipping_cost, which implies these factors do not directly influence the shipping expenses.

3. Average Shipping Cost by Order Quantity

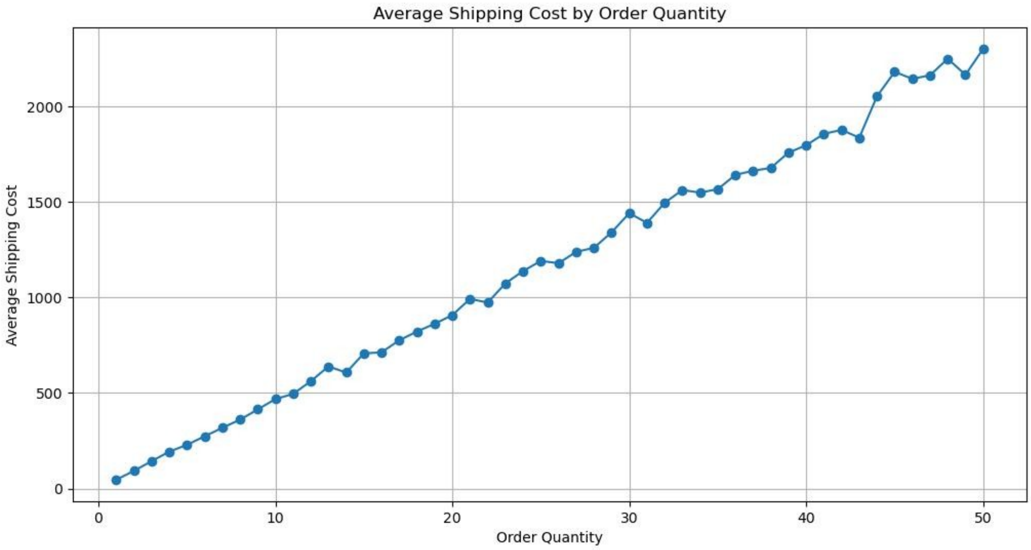


Fig 3.3

Fig 3.3 demonstrates the trend of average shipping cost as a function of order quantity. The curve exhibits a gradual upward trend, indicating that as the number of items ordered increases, the shipping cost also tends to rise. This trend suggests a near-linear relationship between order quantity and shipping cost, reflecting economies of scale for smaller orders and consistent increases for larger ones. However, the growth rate appears to be consistent, implying there may be room for optimization in shipping costs for larger orders.

4. Shipping Cost Vs Unit Price by Shipping mode

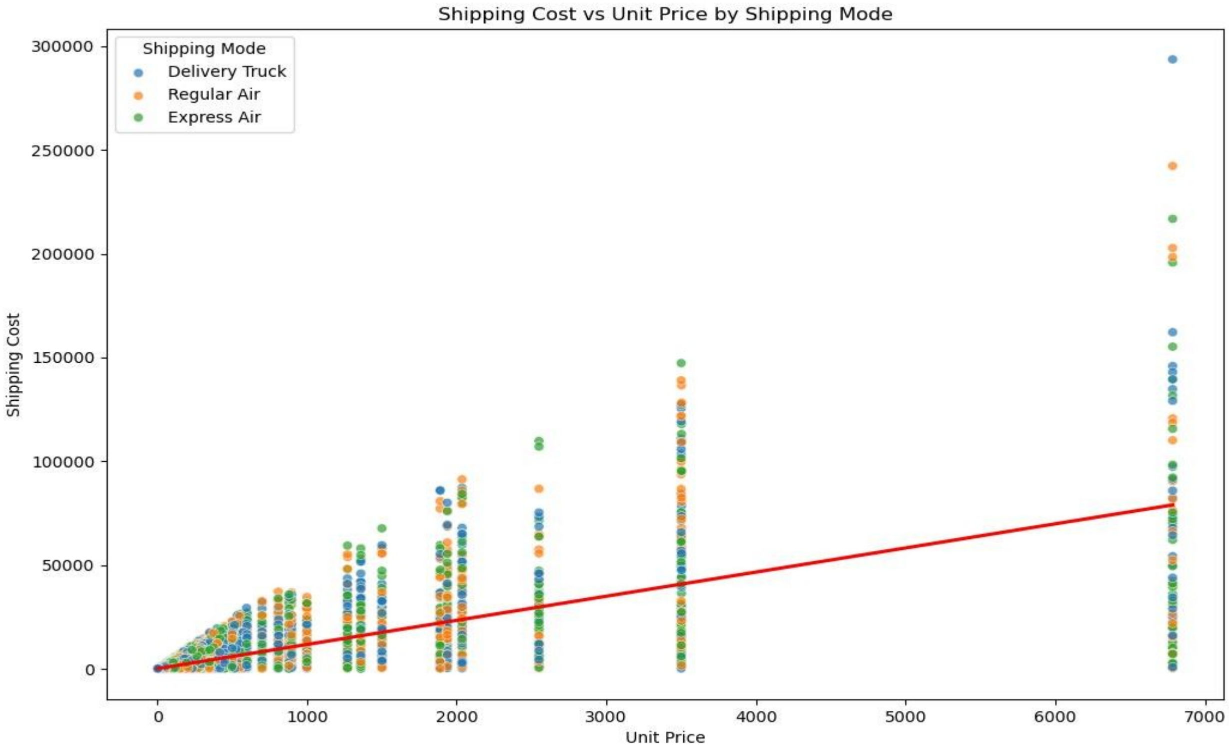


Fig 3.4

Fig 3.4 depicts the relationship between unit price and shipping cost, with different colors representing various shipping modes (Express Air, Regular Air, and Delivery Truck). Additionally, a regression line has been plotted to indicate the general trend.

Key findings from the scatter plot:

- **Concentration at Low Unit Prices**: The graph reveals a high concentration of points at lower unit prices (0–1,000). Within this range, shipping costs vary widely, indicating that even for low-priced products, shipping costs can be substantial depending on the shipping mode.
- **Relationship Between Shipping Mode and Shipping Cost**: For products priced above 1,000 units, shipping costs increase considerably, with Delivery Truck and Regular Air modes appearing frequently at higher cost levels.
- **Sharp Increase at 7,000 Unit Price**: At a unit price of around 7,000, the shipping costs peak dramatically, with most points at the top corresponding to Delivery Truck and Regular Air modes. This could indicate that for high-value items, Walmart prefers more secure or traditional shipping methods over Express Air, possibly due to cost or logistical considerations.

5. Shipping Cost Optimization

1. Model Performance

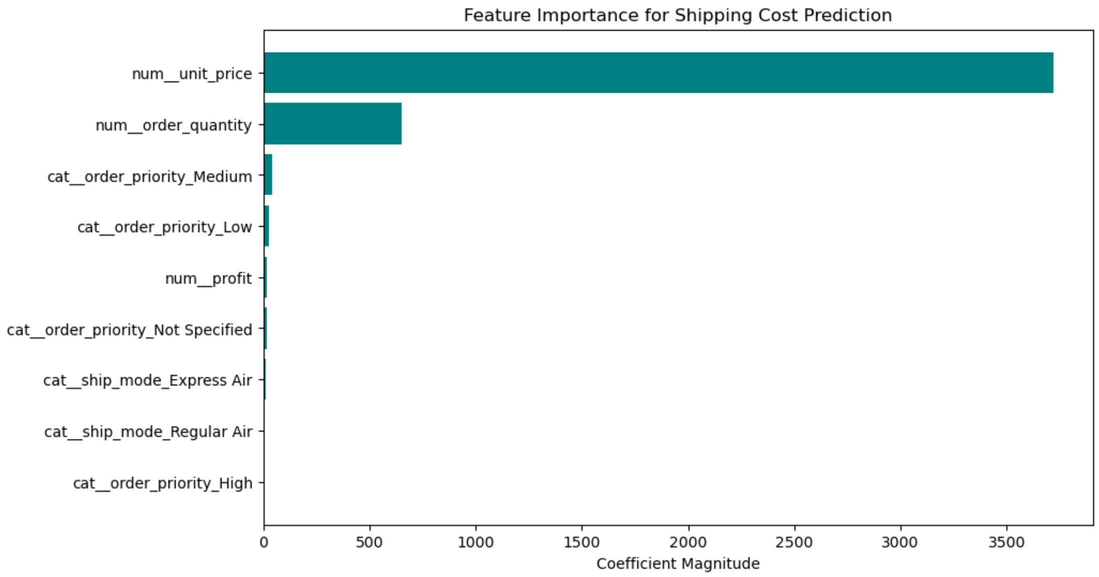
To evaluate the performance of the Ridge Regression model, we used two key metrics: Mean Absolute Error (MAE) and R-squared (R^2). The results for the validation and training sets are:

- **Validation Set Metrics:**
    - **Mean Absolute Error (MAE):** This metric represents the average absolute difference between predicted and actual shipping costs. A lower value indicates better prediction accuracy. The model achieved a MAE of **968.8029** on the validation set.
    - **R-squared (R^2):** The R^2 score indicates the proportion of the variance in the target variable (shipping costs) that is explained by the model. The model achieved an R^2 of **0.5623** on the validation set.
- **Training Set Metrics:**
    - **R-squared (R^2):** The R^2 on the training set was **0.5594**, which confirms that the model is capable of explaining a significant proportion of the variance in the training data, validating its generalization potential.

For real world business problems, It is difficult to get a very high score due to the complexity of data. A R^2 score between 0.5-0.7 may not predict accurate values but is enough to analyze trends and give a rough estimate of shipping costs. The lower MAE indicates that the predictions are reasonably close to actual shipping costs.

2. Optimization Insights

a. Feature Importance Analysis



- From the feature importance analysis, it was found that *unit_price* had the highest coefficient, indicating it is the most significant factor affecting shipping costs. This suggests that higher-priced products contribute significantly to higher shipping costs, likely because they may require different handling or higher shipping tiers.

- *order_quantity* also had a notable impact. This aligns with the intuition that larger orders, with more units, typically result in higher shipping expenses due to volume-related costs.
- Order priorities such as medium and low had much smaller coefficients. This implies that, while order priority is a consideration, it does not significantly impact shipping costs when compared to product-related features like price and quantity.

b. Heatmap Analysis of Shipping Costs by Mode and Priority

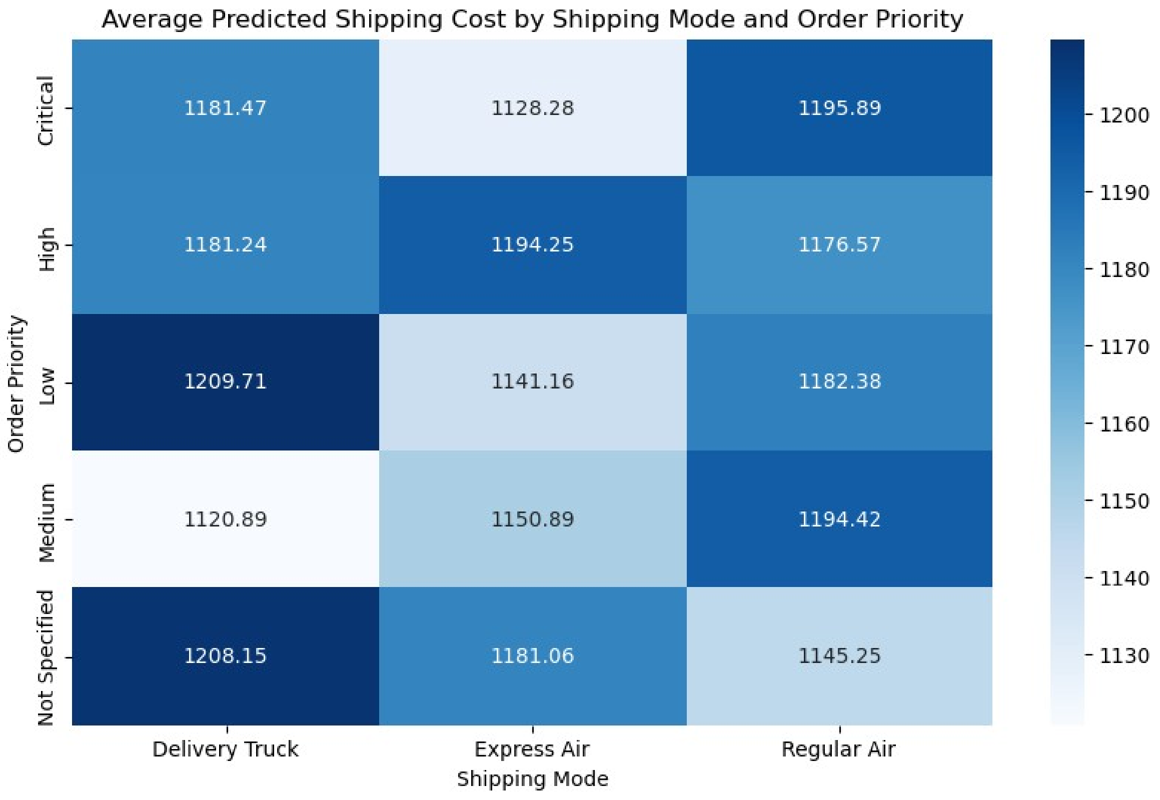


Fig 3.6

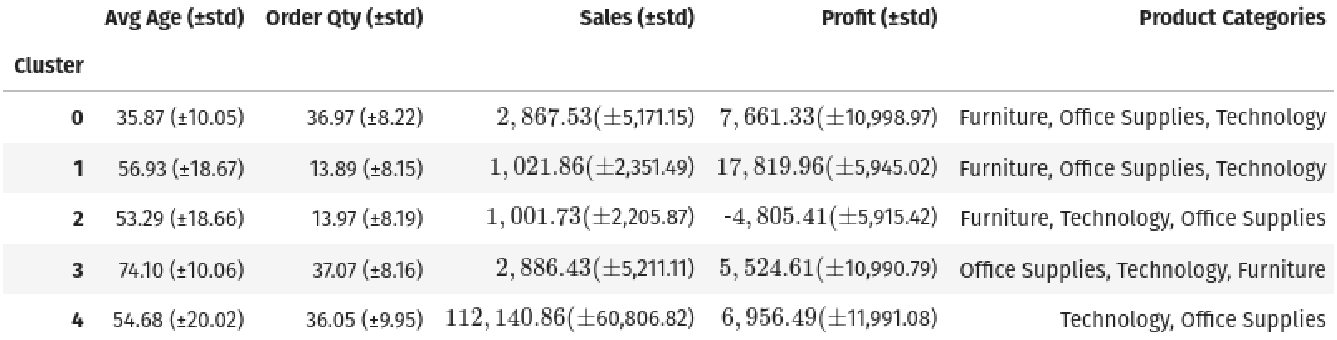
- **For Critical Orders**: The predicted shipping costs for all three modes (Delivery Truck, Express Air, and Regular Air) were similar, with **Express Air** showing the lowest costs (around **$1128**) compared to **Regular Air** at **$1195.89**. This suggests that critical orders may need to rely on **Express Air** if cost minimization is a priority.

- **For High Priority Orders**: Shipping costs across the modes are relatively close, with **Delivery Truck** (around **$1181**) and **Regular Air** (around **$1176.57**) being the more cost-effective choices compared to **Express Air** at **$1194.25**. This suggests that, for high-priority orders, **switching from Express Air to Regular Air** or **Delivery Truck** could offer savings without a substantial impact on delivery speed, especially for orders that are not time-critical.

- **For Medium Priority Orders**: **Delivery Truck** appears to be the most cost-effective shipping option across in medium priority orders. For **Medium Priority Orders**, the predicted shipping costs were lowest with **Delivery Truck** at **$1120.89**, compared to **Express Air** and **Regular Air** at higher costs.

- **For Non-Specified Priority Orders**: The predicted shipping costs for non-specified priority orders were slightly more variable, with **Delivery Truck** at **$1208.15**, **Regular Air** at **$1145.25**, and **Express Air** showing a slightly higher cost. This suggests that for orders with unspecified priorities, **Regular Air** is the best mode to minimize shipping costs.

6. Customer Segmentation.



Cluster	Avg Age (±std)	Order Qty (±std)	Sales (±std)	Profit (±std)	Product Categories
0	35.87 (±10.05)	36.97 (±8.22)	2,867.53(±5,171.15)	7,661.33(±10,998.97)	Furniture, Office Supplies, Technology
1	56.93 (±18.67)	13.89 (±8.15)	1,021.86(±2,351.49)	17,819.96(±5,945.02)	Furniture, Office Supplies, Technology
2	53.29 (±18.66)	13.97 (±8.19)	1,001.73(±2,205.87)	-4,805.41(±5,915.42)	Furniture, Technology, Office Supplies
3	74.10 (±10.06)	37.07 (±8.16)	2,886.43(±5,211.11)	5,524.61(±10,990.79)	Office Supplies, Technology, Furniture
4	54.68 (±20.02)	36.05 (±9.95)	112,140.86(±60,806.82)	6,956.49(±11,991.08)	Technology, Office Supplies

Cluster Characteristics

- **Cluster 0**:
  - Average customer age is around 35.87, indicating a younger segment.
  - This cluster has a high order quantity (36.97) and moderate average sales ($2,867.53) and profit ($7,661.33).
  - Common product categories include Furniture, Office Supplies, and Technology.
  - Customer segments include Small Business, Consumer, Home Office, and Corporate.

- **Cluster 1**:
  - Older customers with an average age of 56.93 and a lower order quantity (13.89).
  - Sales and profit are on the lower end ($1,021.86 and $17,819.96 respectively).
  - Includes diverse product categories, similar to Cluster 0.
- **Cluster 2**:
  - Similar to Cluster 1 in age (53.29) and order quantity (13.97), but has a negative profit (-$4,805.41).
- **Cluster 3**:
  - Oldest age group with an average of 74.10, and high order quantity (37.07).
  - Moderate sales ($2,886.43) and profit ($5,524.61), showing a potential for frequent purchases with moderate returns.
- **Cluster 4**:
  - Mid-age (54.68), high order quantity (36.05), extremely high sales ($112,140.86) and moderate profit ($6,956.49).
  - Primarily interested in Technology and Office Supplies.

Visualization Insights

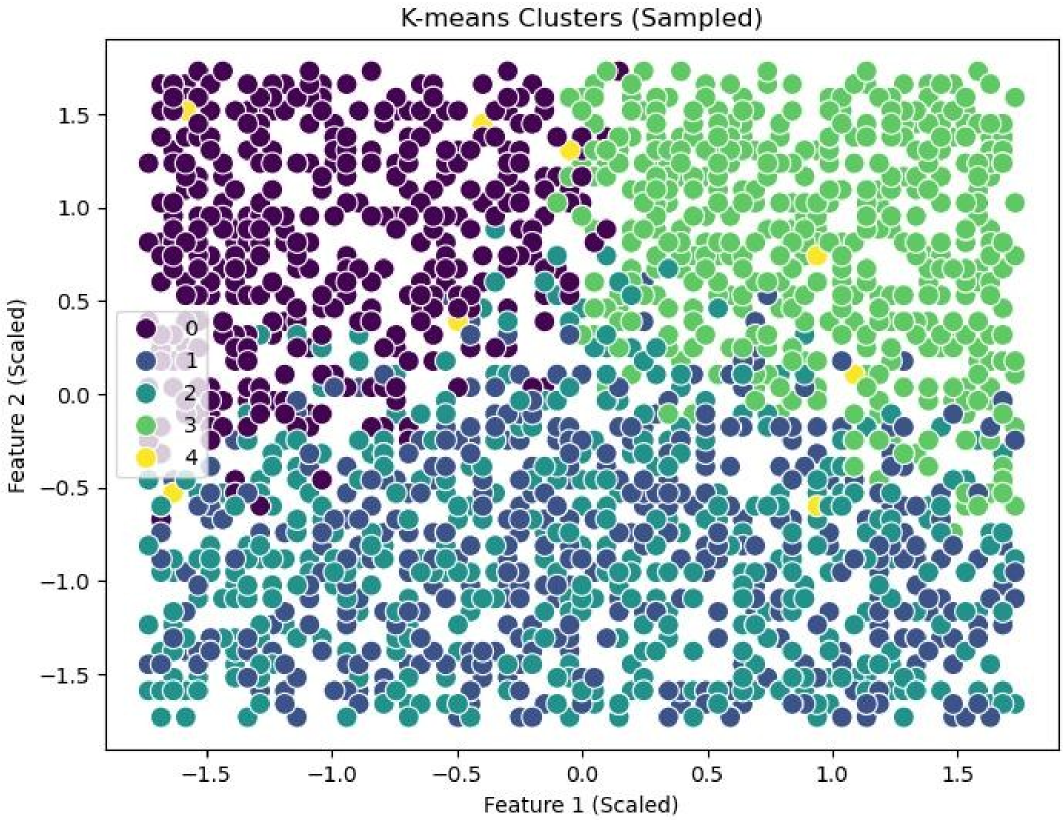


Fig 4.1

- Fig 4.1 shows distinct separation between clusters, indicating good segmentation.
- Some clusters (e.g., Cluster 4, in yellow) have clear distinctions in terms of feature space, suggesting that they represent unique customer groups.
- Clusters 0, 2, and 3 are spread widely, showing diversity in customer characteristics within these segments.

Interpretation of Results and Recommendations :

1. Shipping Modes and Order Priority Alignment

Fig 3.1 suggests that Walmart does not consistently align higher-priority orders with faster, more expensive shipping methods like Express Air. This lack of alignment may lead to inefficiencies and unnecessary costs, especially for critical and high-priority orders.

Recommendation: Walmart should consider implementing a priority-based shipping strategy where higher-priority orders, especially critical ones, are allocated faster shipping options to meet customer expectations. For low and medium-priority orders, more cost-effective options like Delivery Truck can be favored, balancing speed and expense without sacrificing customer satisfaction.

2. Shipping Cost Drivers and Optimization Potential

Fig 3.2 and Fig 3.5 indicate that shipping cost is most strongly influenced by unit price and order quantity. Higher-priced items tend to incur higher shipping costs, likely due to additional handling and insurance requirements. Order quantity also has an impact, albeit smaller, suggesting economies of scale that could be further optimized. Notably, the R^2 score of 0.56 implies that while the model captures significant trends, additional factors could still be considered to refine predictions.

Recommendation: Walmart can leverage this insight by exploring bulk shipping options or negotiated rates for high-value items, potentially lowering costs. Additionally, integrating more predictive factors (e.g., weight, distance, or warehousing costs) into the shipping cost model could improve the accuracy of predictions and lead to more targeted cost-saving strategies.

3. Cluster-Based Customer Segmentation and Targeted Marketing

Fig 4.1 reveals five distinct customer segments, each with unique characteristics. Cluster 0, for instance, consists of younger customers with high order quantities and moderate sales, suggesting frequent but lower-value purchases. In contrast, Cluster 4 includes customers with high sales volumes, primarily focused on Technology and Office Supplies. This diversity presents opportunities for Walmart to tailor marketing and customer retention strategies based on the specific needs and behaviors of each segment.

Recommendation: Walmart should consider a targeted marketing approach based on cluster profiles:

- **For Cluster 0**: Develop loyalty programs or small-business-targeted discounts to encourage repeat purchases.
- **For Cluster 1** and **Cluster 3**: Since these clusters are older and more consistent in ordering smaller quantities, a focus on personalized customer service or senior-friendly promotions may increase engagement.
- **For Cluster 4**: Given the high sales and interest in Technology, Walmart could target this segment with exclusive product releases, bulk purchase discounts, or specialized service offerings.

4. Shipping Mode Recommendations by Order Priority

Fig 3.6 reveals opportunities for cost optimization. For critical and high-priority orders, the use of Express Air can reduce costs slightly while meeting delivery expectations. In cases of medium and non-specified priority orders, opting for Delivery Truck or Regular Air minimizes expenses without significant impact on service quality.

Recommendation: Walmart can introduce a tiered shipping policy based on order priority:

- **Critical Orders**: Prioritize Express Air for timely delivery at minimal cost.
- **High Priority Orders**: Shift from Express Air to Regular Air or Delivery Truck for slight savings.
- **Medium and Low Priority Orders**: Favor Delivery Truck to maintain low costs, particularly for bulk items or low-margin products.