Statistics 208 Project Report

Group-1

Team Leader: Vishesh Jaiprakash Shukla

Team Members: Nandini Ramakrishnan, Pious Khemka, Vishesh Jaiprakash Shukla, Shashank Gorti Srisai Nagaupendra

Topic Title: "The Logistics Puzzle: Optimizing Operations for Delightful Deliveries"

Motivation: Efficient logistics operations are vital for streamlining processes, minimizing costs, and maximizing resource utilization. These operations ensure the timely delivery of products and services, which is crucial for maintaining customer satisfaction and competitive advantage. In our project, we aim to explore the optimization of various logistics components, including warehouse management, transportation modes, and shipping strategies.

- By focusing on warehouse management, we seek to enhance the organization and accessibility of stored products.
- Optimizing transportation modes will allow us to identify the most effective methods for different types of shipments, considering factors such as distance, volume, and potential delays.
- Additionally, refining shipping strategies, such as offering targeted discounts, can boost customer satisfaction and sales performance.

Research Objectives:

Objective 1: Analyse Warehouse Efficiency

Objective: Investigate how different warehouse block allocations affect product storage patterns and operational efficiency. This study aims to identify the most efficient warehouse blocks for maximizing accessibility and minimizing retrieval times, thus enhancing overall logistics performance and delivery punctuality.

Importance and Purpose: Understanding the influence of warehouse organization on logistics operations is crucial for improving delivery efficiency. Efficient warehouse management directly impacts the speed and accuracy of order fulfilment, which is a critical factor for customer satisfaction and retention. For example, a well-organized warehouse can significantly reduce the time required to locate and retrieve products, leading to faster processing times and on-time deliveries.

- **Operational Efficiency:** Proper warehouse block allocation minimizes unnecessary movement and handling, which can lower labour costs and reduce errors.
- **Customer Satisfaction:** Faster and more accurate order fulfilment leads to higher customer satisfaction and repeat business.

- **Cost Reduction:** Efficient storage and retrieval systems can reduce operational costs by optimizing space utilization and reducing the need for excessive inventory.

By identifying optimal warehouse block allocations, our study can provide actionable insights for improving logistics operations. This can help in designing better warehouse layouts and management practices, ultimately leading to enhanced delivery performance and customer satisfaction.

Approach:

- Assess different warehouse block allocations to find patterns that improve efficiency.
- Determine how these patterns impact the speed and accuracy of product retrieval.
- Evaluate the effect of efficient warehouse organization on overall delivery timeliness and customer satisfaction.

Objective 2: Evaluate Transportation Mode Effectiveness

Objective: Examine the impact of various transportation modes (sea, air, road) on delivery performance and customer satisfaction. This objective aims to determine the most efficient transportation method for different product types and distances, taking into account shipping volumes, travel distances, and potential delays.

Importance and Purpose: The choice of transportation mode is a critical decision in supply chain management, as it affects delivery speed, cost, and reliability. For instance, air transport, while fast, is expensive and may not be suitable for all products, whereas sea transport is cost-effective for large volumes but slower.

- **Cost Management:** Different transportation modes have varying costs associated with them. Identifying the most cost-effective option for different scenarios can lead to significant savings.
- **Delivery Timeliness:** Understanding the impact of transportation modes on delivery times helps in planning and meeting customer expectations.
- **Customer Satisfaction:** Reliable and timely deliveries are essential for maintaining customer satisfaction. Delays in delivery can lead to negative customer experiences and impact brand reputation.

Our study will help in developing a nuanced understanding of how different transportation modes influence overall delivery performance. This can guide logistics managers in making informed decisions about which mode to use under various circumstances, thereby optimizing both cost and delivery times.

Approach:

- Analyse the relationship between transportation modes and delivery times.
- Investigate how different transportation methods impact customer satisfaction.
- Identify the most efficient transportation modes for various product types and delivery distances.

Objective 3: Optimize the Impact of Shipping Discounts

Objective: Evaluate the effectiveness of shipping discounts on customer satisfaction and sales performance, particularly for high-priority products. This objective aims to understand how discount incentives influence buying behaviour, customer ratings, and overall profitability, while also considering their impact on shipping costs.

Importance and Purpose: Shipping discounts are a powerful marketing tool that can drive sales and improve customer satisfaction. For example, offering free or discounted shipping can encourage customers to make purchases they might otherwise have hesitated on, particularly for high-importance products.

- **Sales Boost:** Shipping discounts can increase sales volume by incentivizing purchases, especially during peak seasons or for promotional events.
- **Customer Loyalty:** Offering discounts on shipping can improve customer satisfaction and loyalty, as customers perceive greater value.
- **Profitability:** Balancing the benefits of increased sales with the costs of shipping discounts is crucial for maintaining profitability.

By analysing the impact of shipping discounts, our study can provide insights into the most effective discount strategies. This will help businesses design promotions that enhance customer satisfaction and drive sales, while also keeping an eye on profitability.

Approach:

- Assess the correlation between shipping discounts and customer satisfaction levels.
- Evaluate the effect of discounts on sales, especially for high-priority products.
- Determine the profitability and cost implications of various discount strategies.
- Optimize discount offers to balance increased sales and customer satisfaction with shipping costs.

By focusing on these objectives, our study aims to provide a comprehensive understanding of the factors influencing supply chain efficiency and customer satisfaction. This will enable us to make data-driven recommendations for improving logistics operations and achieving better business outcomes. Through this approach, we hope to offer valuable insights that can help businesses optimize their delivery processes, enhance customer satisfaction, and ultimately achieve greater operational success.

Exploratory Data Analysis

Data

The dataset contains information on 10,999 observations with 12 variables. The following are the list of features included:

- ID: Unique identifier for customers
- Warehouse block: Categorical variable indicating warehouse location
- Mode of shipment: Categorical variable representing shipping methods

- Customer care calls: Numerical variable indicating the number of customer inquiries
- Customer rating: Numerical variable indicating customer satisfaction levels
- Cost of the product: Numerical variable representing product cost in USD
- Prior purchases: Numerical variable indicating the number of previous purchases
- Product importance: Categorical variable indicating product importance level
- Gender: Categorical variable representing customer gender
- Discount offered: Numerical variable indicating discount percentage
- Weight in gms: Numerical variable representing product weight in grams
- Reached on time: Binary target variable indicating on-time delivery (0 for on-time, 1 for delayed)

Summary for Numerical Variables:

Descriptive Statistics: Create summary statistics for each numerical variable to get a better understanding of data.

- **Customer Care Calls:** Range from 2 to 7 calls, with an average of 4.05 calls. Most customers (Q1) make at least 3 calls.
- **Cost of the Product:** Costs vary significantly from 96 to 310 units, with a mean of 210.2 units, suggesting a diverse pricing strategy.
- **Prior Purchases:** Range from 2 to 10, with a median of 3, indicating most customers have made few prior purchases.
- **Discount Offered:** Discounts range from 1 to 65 units, with a mean of 13.37 units, highlighting varying discount strategies.
- **Weight in gms:** Weights range from 1001 to 7846 grams, with a mean of 3634 grams, showing a wide range of product sizes.

Frequency Distribution for Categorical Variables:

Frequency Tables: Create frequency tables for each categorical variable to visualize the distribution of categories.

- **Warehouse Block:** Block F has the highest shipments (3666), followed by D, A, B, and C with similar counts (~1833 each).
- **Customer Rating:** Ratings range from 1 to 5, with a mean rating close to 3. This indicates a balanced distribution of ratings.
- **Mode of Shipment:** Most shipments are made by ship (7462), followed by flight (1777) and road (1760).
- **Customer Rating:** Most common rating is 3 (2239), with ratings 1 and 4 also prevalent.
- **Product Importance:** Majority of products are of low importance (5297), followed by medium (4754), and high importance (948).
- **Gender:** Distribution is almost equal, with 5545 females and 5454 males.
- **Reached on Time:** 6563 shipments reached on time, while 4436 did not.

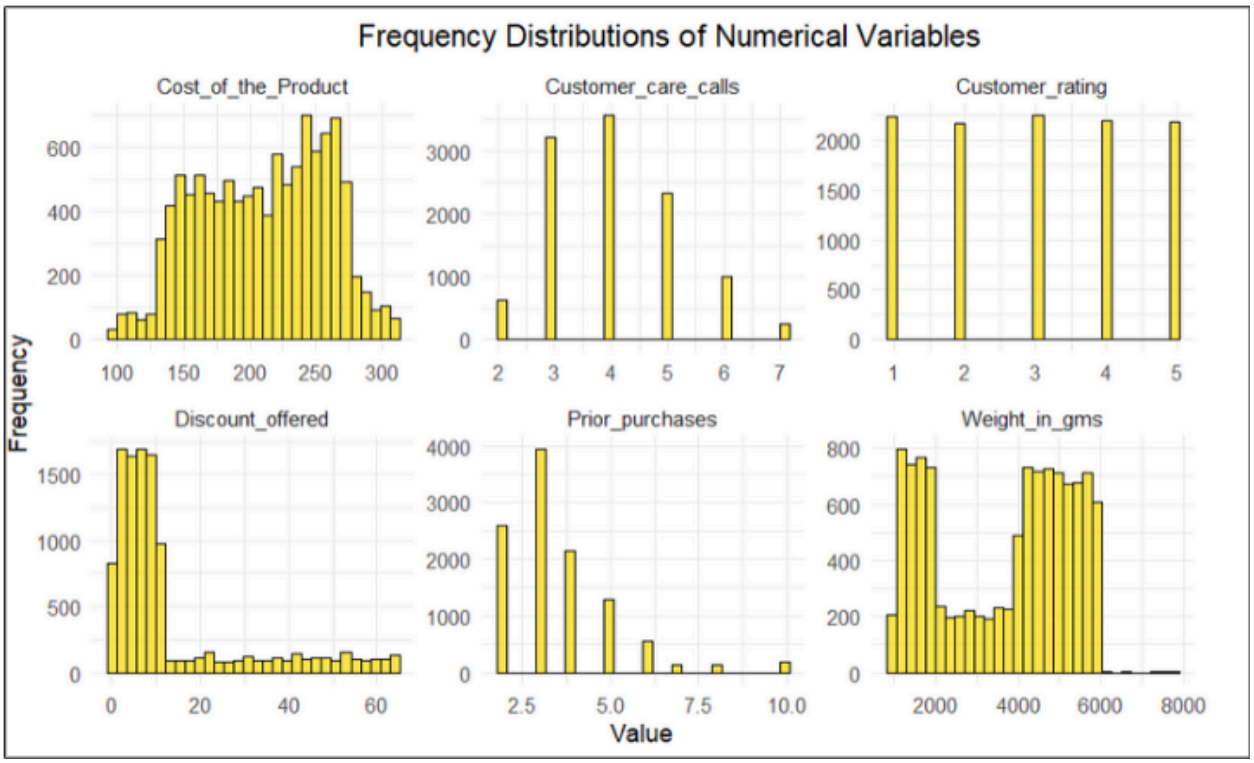
Identify and Handle Missing Values

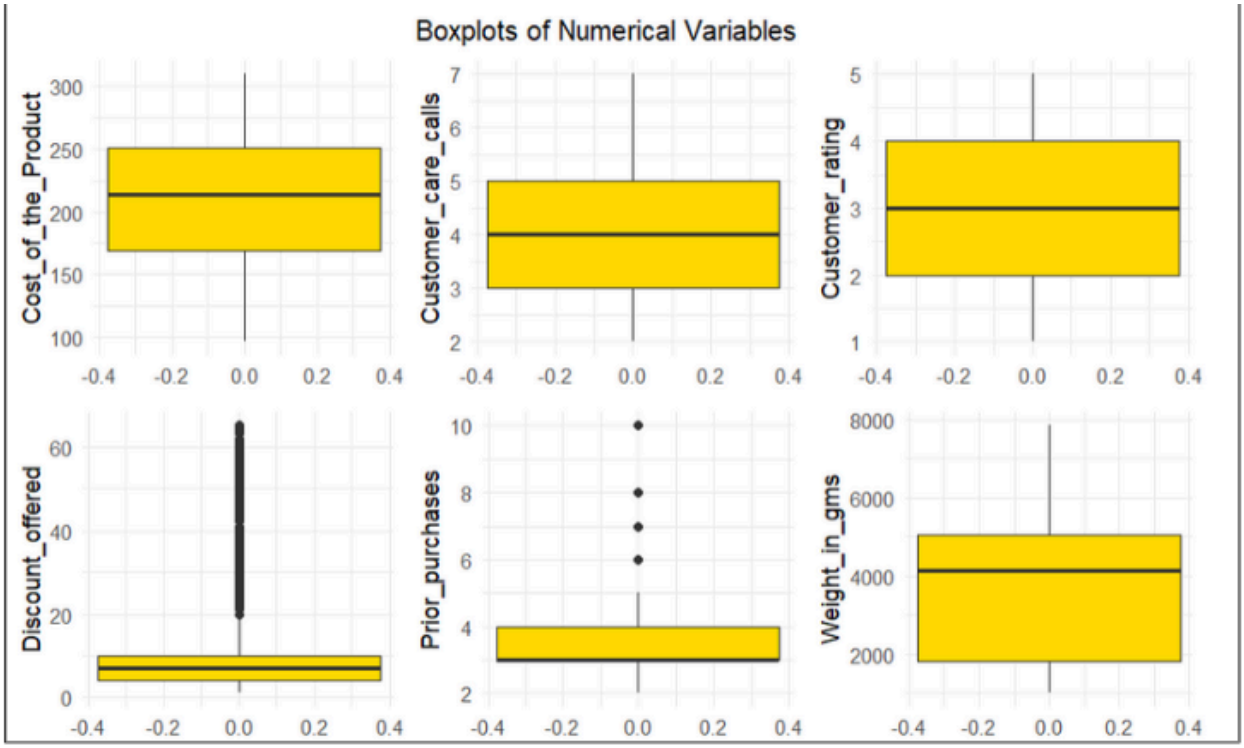
- **Approach:** Check for missing values in the dataset and handle them appropriately.
- **Findings:** No missing values were detected in any of the variables, ensuring complete data for analysis.

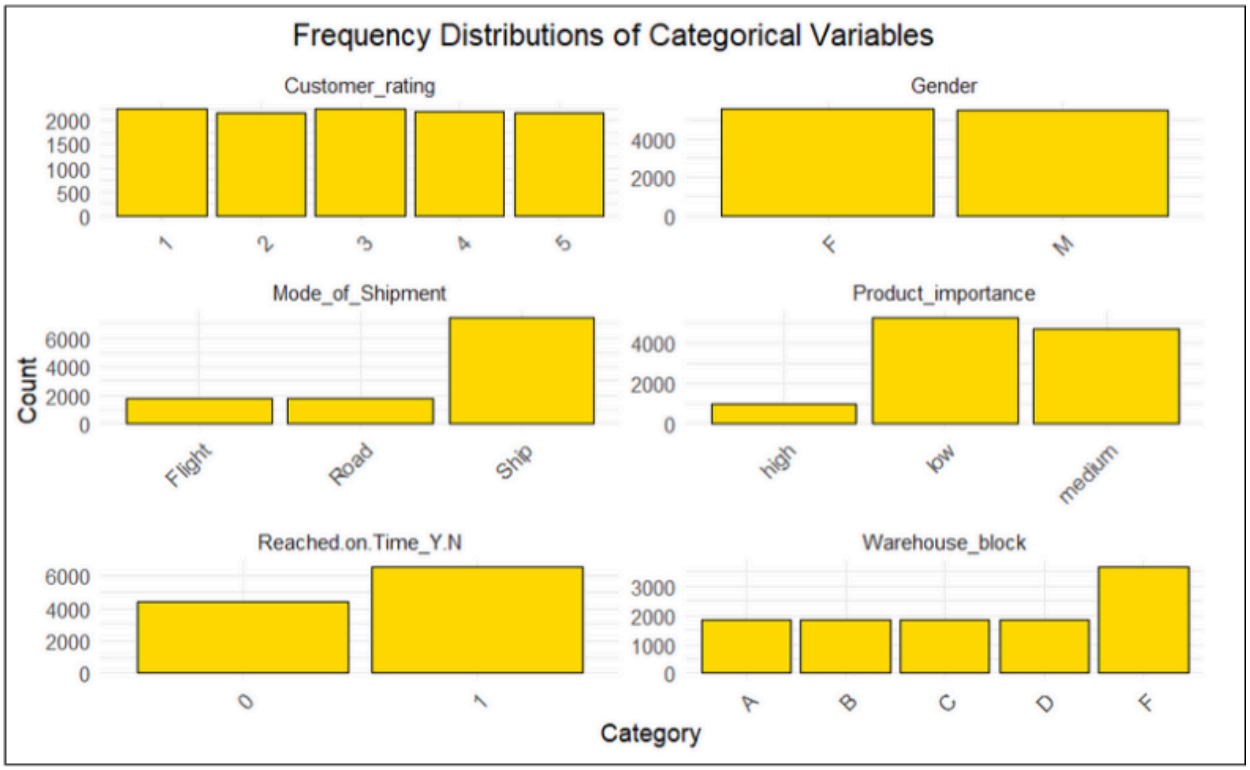
Univariate Analysis

Approach: Analyse the distribution of each variable individually.

- Numerical Variables: Use histograms and boxplots.
- Categorical Variables: Use bar charts.




Boxplots of Numerical Variables
Cost_of_the_Product
Customer_care_calls
Customer_rating
Discount_offered
Prior_purchases
Weight_in_gms


Frequency Distributions of Categorical Variables
Customer_rating
Gender
Mode_of_Shipment
Product_importance
Reached.on.Time_Y.N
Warehouse_block
Count
Flight
Road
Ship
high
low
medium
F
M
A
B
C
D
F
Category

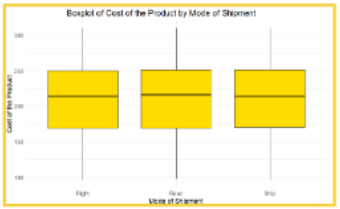
Findings:

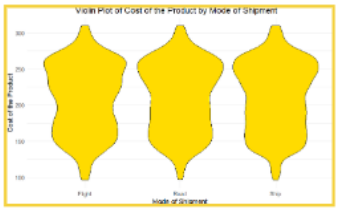
1. **Numerical Variables:**
   - Histograms showed a right-skewed distribution for Customer Care Calls and Prior Purchases, indicating a concentration towards the lower values.
   - Boxplots revealed outliers in Discount Offered, suggesting occasional high discounts.
2. **Categorical Variables:** Bar charts confirmed the frequency distributions previously described.

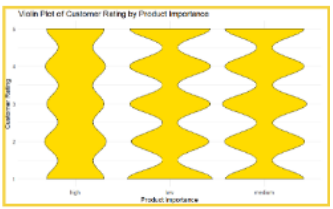
Bivariate Analysis

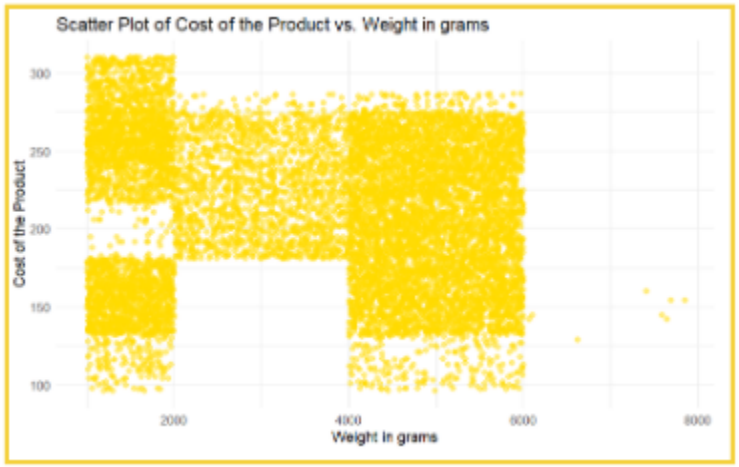
Approach: Explore relationships between pairs of variables.

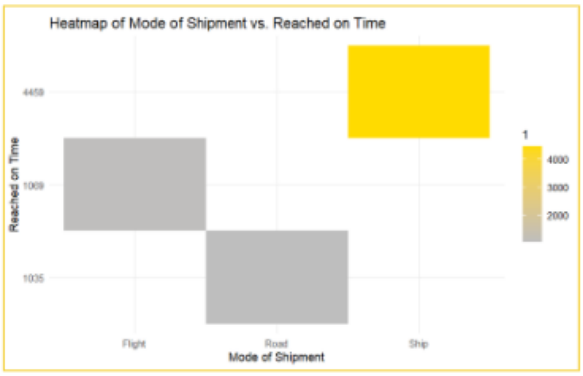
- Categorical vs. Numerical: Use boxplots and violin plots.
- Numerical vs. Numerical: Use scatter plots.
- Categorical vs. Categorical: Use heatmaps and stacked bar charts.











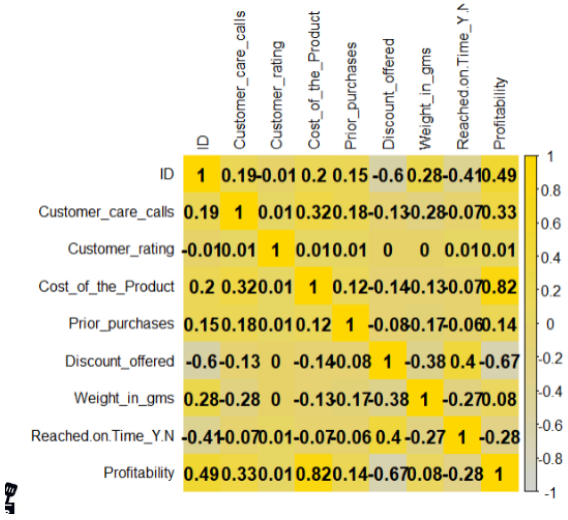
Findings:

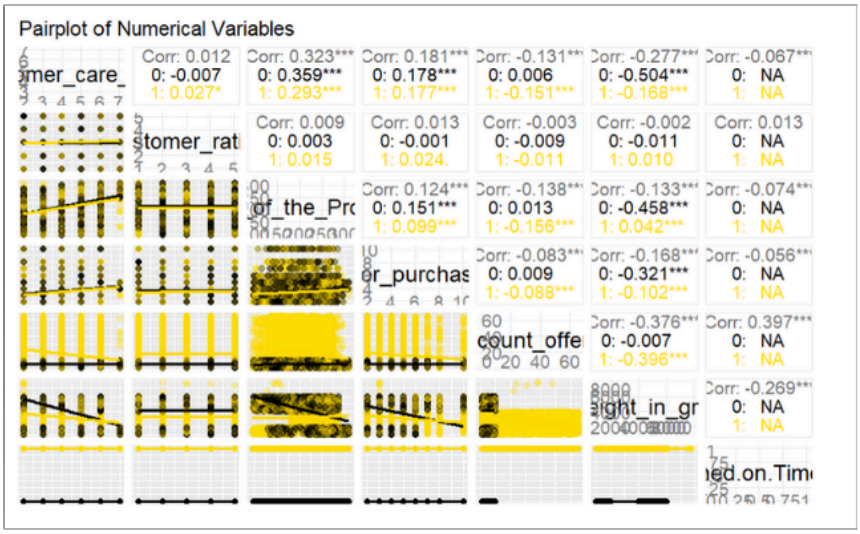
1. **Categorical vs. Numerical:**
   - Boxplots showed that the cost of the product varies across different shipment modes, with flights generally having higher costs.
2. **Numerical vs. Numerical:**
   - Scatter plots indicated a positive correlation between Cost of the Product and Weight in grams.
3. **Categorical vs. Categorical:**
   - Heatmaps showed that the mode of shipment and whether the shipment reached on time are related, with ship mode having more on-time deliveries.

Multivariate Analysis

Approach:

- Explore complex relationships involving multiple variables.
- Use correlation matrices and pair plots for numerical variables.





Findings:

- **Correlation Matrix:** Discounts and Delivery Performance: Higher discounts attract more sales but also strain the delivery system, leading to more delays. This indicates a need to balance promotional strategies with logistical capabilities to maintain service quality.

- Warehouse Utilization and Efficiency: High utilization in Warehouse Block F shows efficiency in handling large volumes, but the significant delays in Block D highlight the need for targeted improvements in specific blocks to ensure uniform performance.

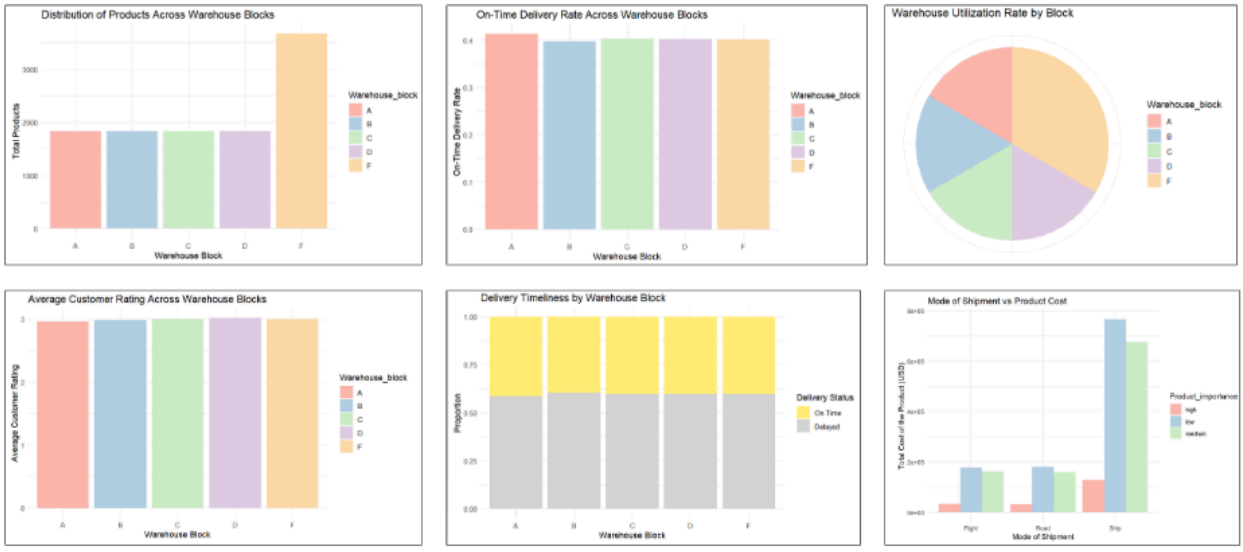
- Mode of Shipment and Product Characteristics: The choice of shipment mode impacts delivery timeliness. High-cost and urgent products benefit from air shipment, which, although less common, is the most reliable. Road and ship are suitable for less time-sensitive and lower-cost products but come with higher delay risks.

- Customer Behavior and Service Quality: Prioritizing new customers for on-time deliveries can enhance initial customer satisfaction and loyalty. However, addressing the needs of repeat customers with potentially more complex orders is also crucial for long-term retention.

- **Pairplot:** Pairplot showed pairwise relationships, confirming trends seen in bivariate analysis and highlighting some potential clusters in the data.

Based on above findings, we see the following impacts on the objectives of the study:

Warehouse Block Analysis:



1. Distribution of Products Across Warehouse Blocks

- Insight: This graph shows the total number of products distributed across different warehouse blocks.
- Observations: Warehouse Block F has a significantly higher number of products compared to other blocks.
- Blocks A, B, C, and D have a relatively balanced distribution of products, but Block F stands out, indicating potential over-utilization or specialization.

2. On-Time Delivery Rate Across Warehouse Blocks

- Insight: This graph indicates the on-time delivery rate for products stored in different warehouse blocks.
- Observations: Blocks A, B, and F show the highest on-time delivery rates.
- Block D has the lowest on-time delivery rate, suggesting possible inefficiencies or challenges in this block.

3. Warehouse Utilization Rate by Block

- Insight: This pie chart represents the utilization rate of each warehouse block.
- Observations: Block F has the largest share of warehouse utilization, aligning with its high product distribution.
- Blocks A, B, C, and D have fairly equal utilization rates, indicating balanced usage but potential under-utilization compared to Block F.

4. Average Customer Rating Across Warehouse Blocks

- Insight: This graph shows the average customer ratings for products stored in different warehouse blocks.

- Observations: Customer ratings are relatively consistent across all blocks, with slight variations.
- Block F has a marginally higher customer rating, potentially due to better handling or product quality.

5. Delivery Timeliness by Warehouse Block

- Insight: This graph displays the proportion of on-time versus delayed deliveries for each warehouse block.
- Observations: Blocks A, B, and F have a higher proportion of on-time deliveries.
- Block D again shows a higher proportion of delayed deliveries, indicating a need for improvement in logistics or operations.

6. Mode of Shipment vs Product Cost

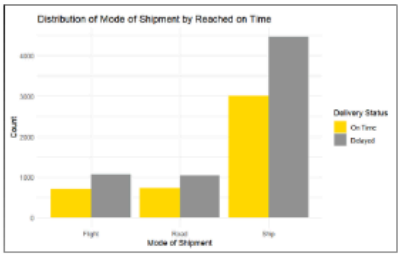
- Insight: This graph correlates the mode of shipment (flight, road, ship) with product cost (low, medium, high importance).
- Observations: High-cost products are predominantly shipped by flight, indicating a priority on speed.
- Medium-cost products have a balanced distribution across flight, road, and ship.
- Low-cost products are mainly shipped by road and ship, suggesting cost-efficiency is prioritized over speed for these items.

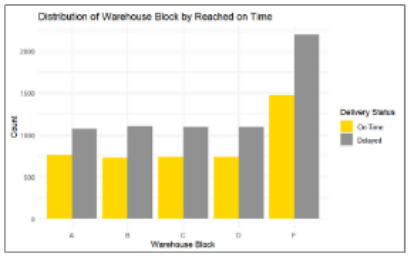
Summary

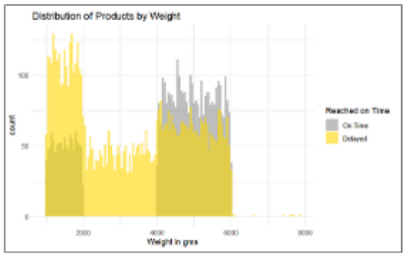
Overall, the analysis of the warehouse utilization charts suggests:

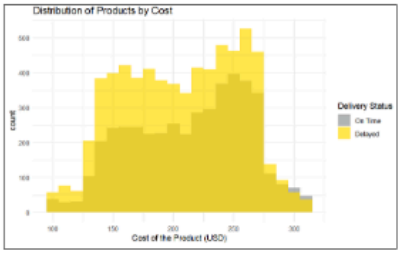
Block F is heavily utilized and manages a large volume of products, showing high efficiency in distribution and customer satisfaction. Block D shows inefficiencies in on-time delivery and a higher rate of delays, indicating areas for operational improvement. Customer ratings are consistent across all blocks, suggesting uniform service quality. Shipment mode varies by product cost, with high-value items prioritized for faster delivery methods (flight), while lower-value items are shipped using cost-effective methods (road and ship). Effective warehouse management and optimization of delivery methods are crucial for maintaining high customer satisfaction and operational efficiency.

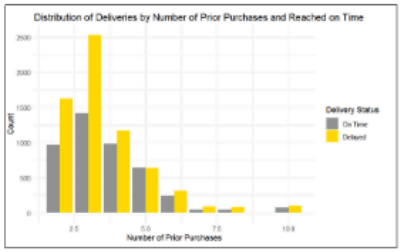
Mode of Shipment Analysis:

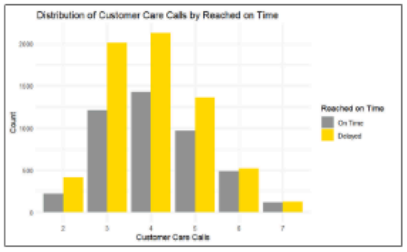












1. Distribution of Mode of Shipment by Reached on Time

- Insight: This graph shows the count of on-time and delayed deliveries across different modes of shipment (flight, road, ship).
- Observations: Shipments by road and ship have a higher count of on-time deliveries compared to flight.
- However, the delay count for ship is also relatively high, indicating a trade-off between volume and timeliness. Flight has the lowest delay rate, suggesting it is the most reliable for timely deliveries despite fewer shipments.

2. Distribution of Warehouse Block by Reached on Time

- Insight: This graph indicates the count of on-time and delayed deliveries for products stored in different warehouse blocks.
- Observations: Block F has the highest count of on-time deliveries, followed by Block B.
- Block D has the most delays, confirming previous observations about inefficiencies in this block. Overall, Blocks A and C have a balanced distribution of on-time and delayed deliveries.

3. Distribution of Products by Weight

- Insight: This graph shows the distribution of product weights and their delivery status (on-time or delayed).
- Observations: Lighter products (up to around 2000 grams) have more on-time deliveries.
- Heavier products (above 2000 grams) show an increase in delays. This suggests that heavier products face more logistical challenges, affecting their delivery timeliness.

4. Distribution of Products by Cost

- Insight: This graph indicates the count of on-time and delayed deliveries based on the cost of products.
- Observations: Products in the mid-cost range (around $150-$250) have a higher count of on-time deliveries. Both lower-cost and higher-cost products show a higher proportion of delays.
- This may indicate that mid-cost products are prioritized for timely delivery, or they are easier to manage logistically.

5. Distribution of Deliveries by Number of Prior Purchases and Reached on Time

- Insight: This graph shows the delivery status based on the number of prior purchases made by customers.
- Observations:Customers with fewer prior purchases (around 2-3) have a higher count of on-time deliveries.
- As the number of prior purchases increases, the count of delays also increases.This suggests that new customers might be prioritized for timely deliveries, or that repeat customers might have more complex orders that are harder to fulfill on time.

6. Distribution of Customer Care Calls by Reached on Time

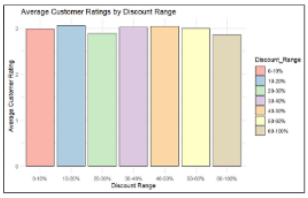
- Insight: This graph shows the count of customer care calls based on delivery status.
- Observations: More customer care calls are associated with delayed deliveries, peaking at around 3-4 calls. On-time deliveries correspond to fewer customer care calls, indicating better customer satisfaction and fewer issues.
- This highlights the importance of timely deliveries in reducing customer complaints and improving service quality.

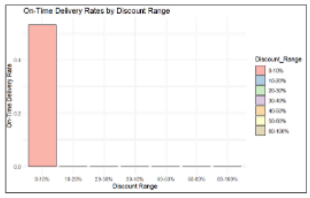
Summary

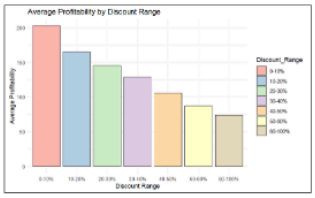
Overall, the analysis of these charts suggests:

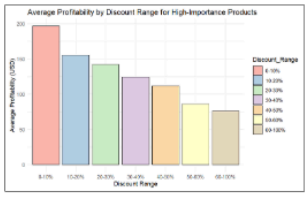
Road and ship are the most commonly used modes of shipment, but flight is the most reliable for on-time deliveries. Warehouse Block F performs well in terms of timely deliveries, while Block D requires improvements. Lighter products and mid-cost products have better on-time delivery rates. New customers tend to receive their orders on time more often than repeat customers. Delayed deliveries are strongly correlated with higher customer care calls, emphasizing the need for timely delivery to enhance customer satisfaction.

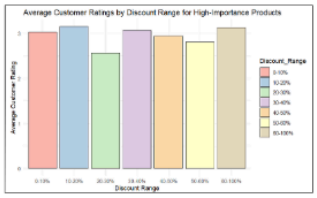
Discount Analysis:

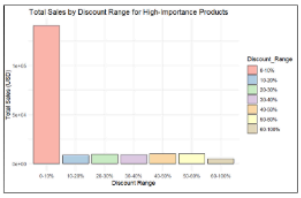












1. Average Customer Ratings by Discount Range

- Insight: This graph shows the average customer ratings for kitchen appliances across different discount ranges.
- Observations: Customer ratings are relatively stable across discount ranges, with no significant decline or improvement in ratings as the discount increases.
- The highest ratings are slightly associated with the 0-10% discount range, suggesting that a lower discount does not negatively impact customer satisfaction.

2. On-Time Delivery Rates by Discount Range

- Insight: This graph displays the on-time delivery rates for kitchen appliances based on the discount ranges.
- Observations: The 0-10% discount range shows the highest on-time delivery rates.
- As discounts increase, on-time delivery rates significantly drop, indicating possible logistical challenges or increased demand causing delays with higher discounts.

3. Average Profitability by Discount Range

- Insight: This graph indicates the average profitability of kitchen appliances across different discount ranges.
- Observations: Profitability is highest in the 0-10% discount range and steadily decreases as the discount increases.
- This suggests that higher discounts erode profit margins, potentially due to the increased costs associated with promotions or reduced revenue per unit.

4. Average Customer Ratings by Discount Range for High-Importance Products

- Insight: This graph shows average customer ratings specifically for high-importance kitchen appliances across different discount ranges.

- Observations: Similar to the general trend, customer ratings remain relatively stable across discount ranges.

- There is a slight drop in ratings for the highest discount range (90-100%), which may indicate a perception of lower quality or expectations not being met for high-importance products at very high discounts.

5. Total Sales by Discount Range for High-Importance Products

- Insight: This graph represents total sales of high-importance kitchen appliances across different discount ranges.

- Observations: The majority of sales are concentrated in the 0-10% discount range, with a sharp decline as discounts increase.

- This indicates that high-importance products sell more with minimal discounts, suggesting customers may value these products more and are willing to purchase them even with smaller discounts.

6. On-Time Delivery Rate by Discount Range for High-Importance Products

- Insight: This graph shows on-time delivery rates for high-importance kitchen appliances based on the discount ranges.

- Observations: The pattern mirrors the general trend where on-time delivery rates are highest for the 0-10% discount range and decrease with higher discounts.

- The sharp drop in on-time delivery for higher discounts may again indicate logistical issues or increased demand leading to delays.

Summary

Overall, the analysis suggests that:

Higher discounts do not significantly improve customer ratings. On-time delivery rates and profitability decrease with higher discounts. Sales are highest at lower discount ranges, especially for high-importance products, indicating customers' willingness to purchase these items with minimal discounts. Companies need to balance discount strategies to avoid negatively impacting profitability and delivery performance while maintaining customer satisfaction.

Data Modeling

In our data modeling section, we selected various models based on their relevance to our study's objectives. Our primary aim was to find models that could effectively address our research objectives: analyzing warehouse efficiency, evaluating transportation mode effectiveness, and optimizing the impact of shipping discounts. The selection criteria for these models were guided by their ability to:

1. **Capture Operational Efficiency:** The models needed to accurately assess the efficiency of different warehouse block allocations in terms of minimizing retrieval times and optimizing storage patterns.
2. **Evaluate Delivery Timeliness and Cost:** They were capable of analyzing the impact of different transportation modes on delivery performance, balancing speed, cost, and customer satisfaction.
3. **Analyze Sales and Customer Satisfaction:** For Objective 3, we focused on models that could evaluate the effect of shipping discounts on sales performance, customer satisfaction, and overall profitability.

Based on above theory, different models were selected for their robustness in classification tasks, interpretability, and ability to handle complex interactions among variables pertinent to our study objectives.

Approach

We approached the data modeling process with a structured methodology:

1. **Data Preprocessing:** We cleaned and prepared the dataset, ensuring that it was suitable for analysis by handling missing values, normalizing data, and encoding categorical variables.
2. **Feature Selection:** Relevant features were selected based on their importance to our research objectives. For instance, features related to warehouse block locations, transportation modes, and discount offers were included.
3. **Model Training:** We trained each of the selected models using a portion of the dataset and validated their performance using cross-validation techniques.
4. **Model Evaluation:** The models were evaluated using performance metrics that align with our objectives, such as accuracy, sensitivity, specificity, precision, recall, F1 score, and area under the ROC curve.

Methodology

We employed a structured methodology encompassing the following steps:

- **Data Splitting:** Dividing the dataset into training and testing sets.
- **Model Training:** Training each model on the training dataset.
- **Model Testing:** Evaluating model performance on the test dataset.
- **Metrics Calculation:** Computing various performance metrics for each model.
- **Comparison and Analysis:** Comparing models based on their performance metrics.

Models Evaluated:

- Logistic Regression
- Random Forest Model A
- Random Forest Model B
- Gradient Boosting
- XGBoost
- Decision Tree
- PCA
- KNN

Performance metrics used to evaluate the models include:

- **Accuracy:** Proportion of correctly classified instances.
- **Sensitivity (Recall):** Ability to correctly identify positive cases.
- **Specificity:** Ability to correctly identify negative cases.
- **Positive Predictive Value (Precision):** Proportion of positive results that are true positives.
- **Negative Predictive Value:** Proportion of negative results that are true negatives.
- **Balanced Accuracy:** Average of sensitivity and specificity.
- **Kappa Statistic:** Measure of agreement between predicted and actual classes, accounting for chance.

Models:

1. **Logistic Regression:**
   - **Accuracy:** Training - 0.7292, Test - 0.7154
   - **Sensitivity:** 0.9849
   - **Specificity:** 0.4447
   - **PPV:** 0.6404
   - **NPV:** 0.9670
   - **Balanced Accuracy:** 0.7148
   - **Kappa:** 0.4301

2. **Random Forest Model A:**
   - **Accuracy:** Training - 0.9992, Test - 0.6881
   - **Sensitivity:** 0.8602
   - **Specificity:** 0.5584
   - **PPV:** 0.5683
   - **NPV:** 0.8553
   - **Balanced Accuracy:** 0.7093
   - **Kappa:** 0.3864

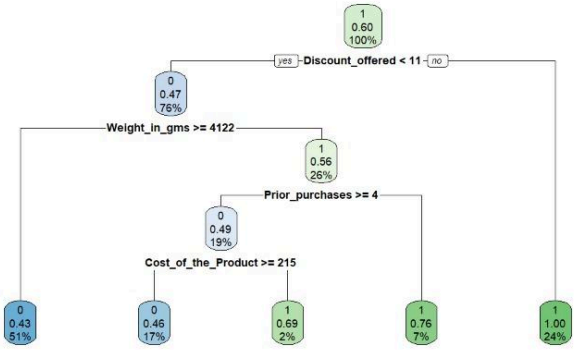
3. **Random Forest Model B:**
   - **Test Accuracy:** 0.6667
   - **Sensitivity:** 0.7714

- **Specificity:** 0.5981
- **PPV:** 0.5647
- **NPV:** 0.7947
- **Balanced Accuracy:** 0.6847
- **Kappa:** 0.3454

4. **Gradient Boosting:**
    - **Accuracy:** Training - 0.6895, Test - 0.6861
    - **Sensitivity:** 0.9526
    - **Specificity:** 0.4970
    - **PPV:** 0.5629
    - **NPV:** 0.9895
    - **Balanced Accuracy:** 0.7358
    - **Kappa:** 0.4039

5. **XGBoost:**
    - **Accuracy:** Training - 0.7039, Test - 0.6822
    - **Sensitivity:** 0.9038
    - **Specificity:** 0.5325
    - **PPV:** 0.5664
    - **NPV:** 0.8912
    - **Balanced Accuracy:** 0.7181
    - **Kappa:** 0.3979

6. **Decision Tree:**
    - **Accuracy:** Training - 0.6845, Test - 0.6834
    - **Sensitivity:** 0.9534
    - **Specificity:** 0.5010
    - **PPV:** 0.5636
    - **NPV:** 0.9408
    - **Balanced Accuracy:** 0.7272
    - **Kappa:** 0.4086



Summary:

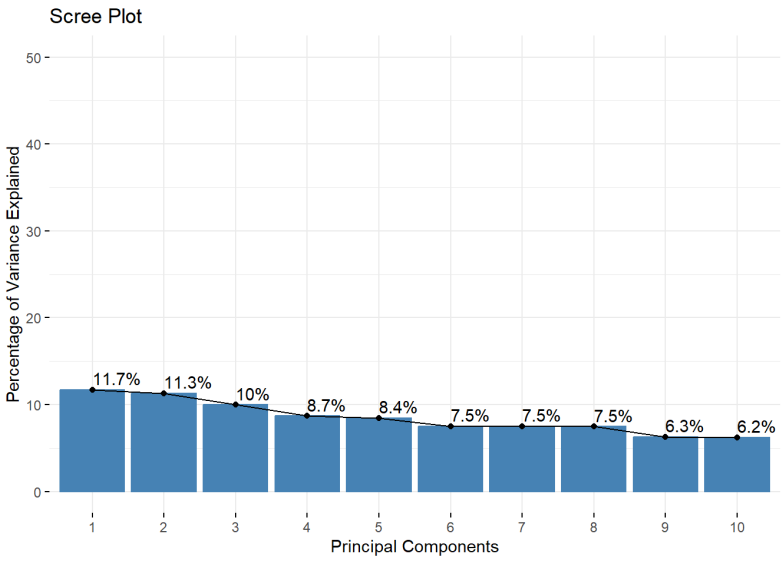
- Discount Offered: The most influential feature for predicting shipment timeliness. Shipments with higher discounts (11% or more) are almost always on time.
- Weight of Product: Heavier products (weight >= 4122 grams) are more likely to be on time if they also have a lower discount.
- Prior Purchases: Shipments with higher prior purchases (>= 4) and high product cost (>= 215) tend to be delayed.
- Cost of the Product: Higher product cost influences the likelihood of delay for shipments with a certain weight and prior purchases.

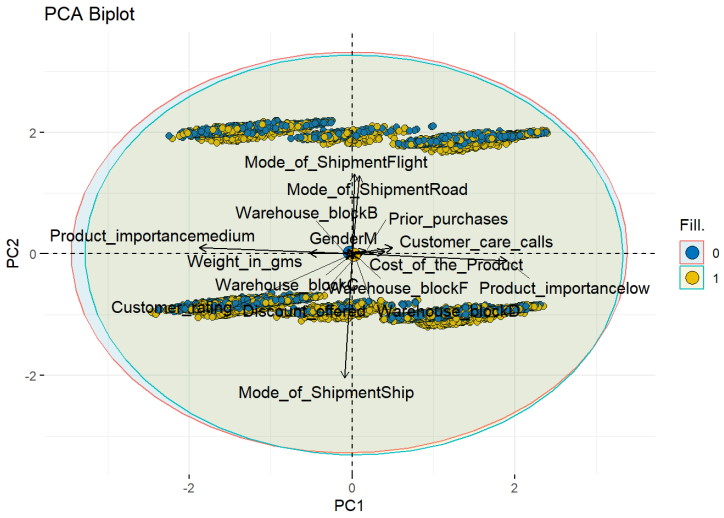
7. PCA

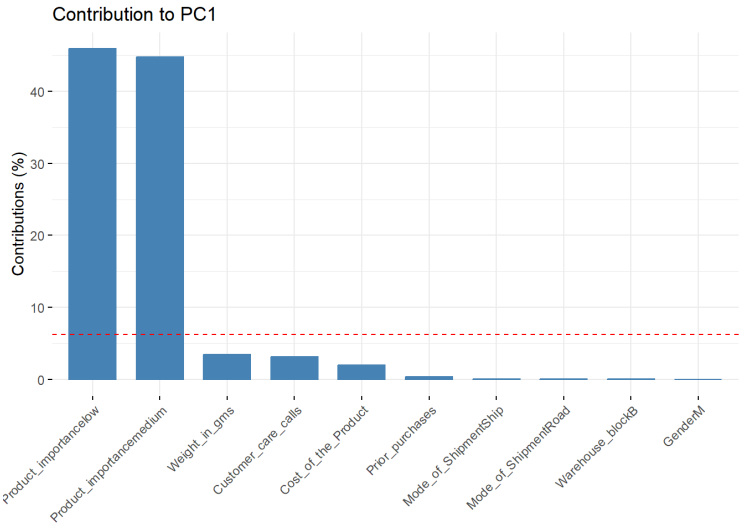
The primary component analysis (PCA) was employed to reduce the dimensionality of a dataset involving 10,999 observations across 12 variables, which include both categorical and numerical features. The main goal was to uncover significant patterns and relationships within the data that influence warehouse efficiency, transportation mode impact, and the effectiveness of shipping discounts.

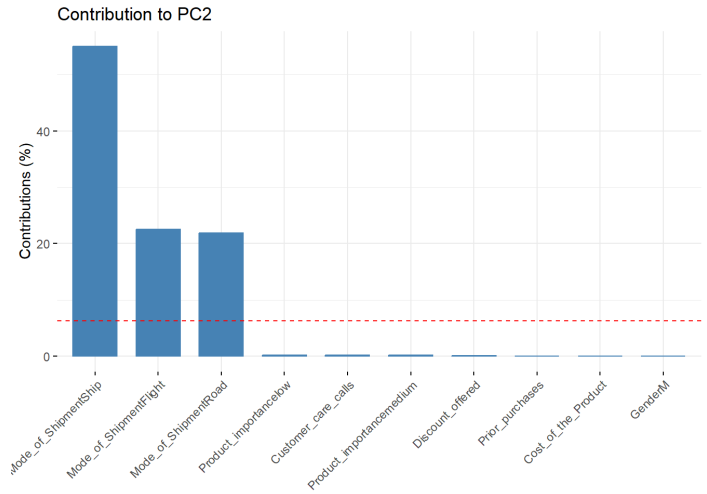
Key steps in the PCA include:

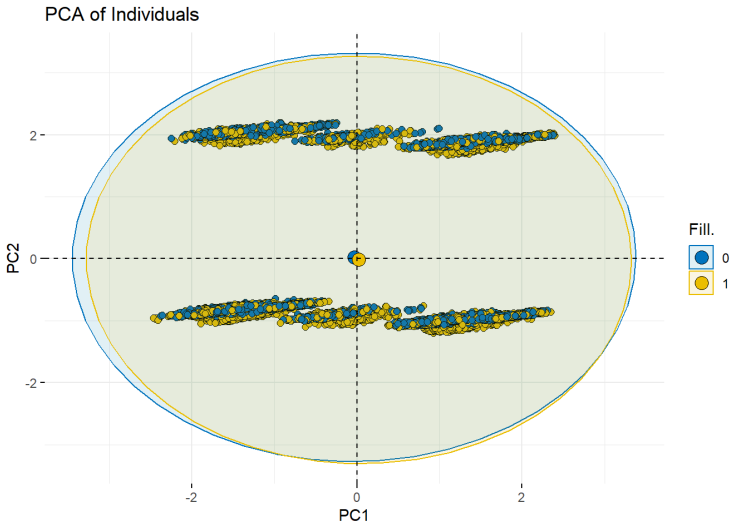
- **Data Preparation:** Categorical variables (e.g., Mode of Shipment, Warehouse Block) were converted into dummy variables to include them in the PCA. Numerical variables were standardized to have zero mean and unit variance, ensuring each variable contributes equally to the analysis.
- **PCA Execution:** PCA was performed on the prepared dataset to extract the principal components, which represent directions in the data that maximize variance.
- **Visualization and Analysis:** Various plots were generated to visualize the distribution of data across clusters, contributions of different variables to the principal components, and the biplot to display the projection of both the variables and observations in the principal component space.

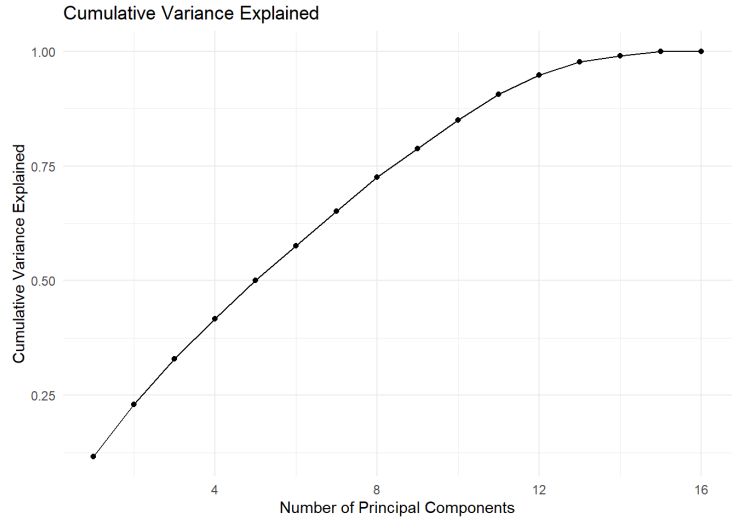



PCA Biplot
Mode_of_ShipmentFlight
Mode_of_ShipmentRoad
Warehouse_blockB
Prior_purchases
Product_importancemedium
GenderM
Customer_care_calls
Weight_in_gms
Cost_of_the_Product
Warehouse_blockC
Warehouse_blockF
Product_importancelow
Customer_rating
Discount_offered
Warehouse_blockD
Mode_of_ShipmentShip
PC1
PC2
Fill.
0
1


Contribution to PC1
Contributions (%)
Product_importancelow
Product_importancemedium
Weight_in_gms
Customer_care_calls
Cost_of_the_Product
Prior_purchases
Mode_of_ShipmentShip
Mode_of_ShipmentRoad
Warehouse_blockB
GenderM


Contribution to PC2
Contributions (%)
Mode_of_ShipmentShip
Mode_of_ShipmentFlight
Mode_of_ShipmentRoad
Product_importancelow
Customer_care_calls
Product_importancemedium
Discount_offered
Prior_purchases
Cost_of_the_Product
GenderM





Results:

- The first two principal components explain approximately 23% of the variance in the dataset, suggesting a moderate level of data complexity.
- The scree plot showed a clear elbow after the third component, indicating that most of the informative variance is captured within the first three components.
- Contributions of variables to the first two components revealed that product importance, mode of shipment, and customer care calls were significant, suggesting their pivotal roles in determining shipping efficiency and customer satisfaction.

Analysis:

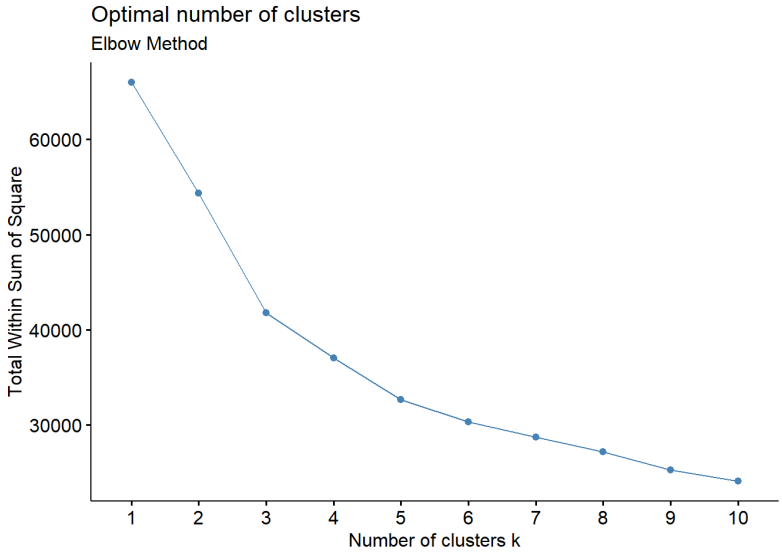
- **Warehouse Efficiency:** The PCA results indicate that the allocation of warehouse blocks and the mode of shipment significantly impact product storage patterns and accessibility, with different blocks showing varying efficiencies.
- **Transportation Mode Impact:** The significant contribution of the transportation mode to the principal components underscores its effect on delivery times and customer satisfaction, which aligns with the second objective.
- **Shipping Discounts:** The impact of discounts on sales and customer ratings, although less pronounced in the first two components, still plays a role in customer purchase decisions and satisfaction levels.

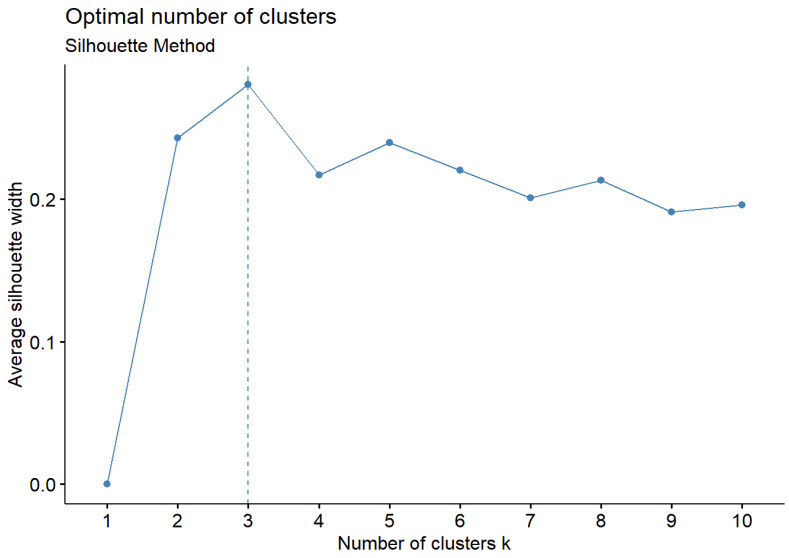
8. K Means

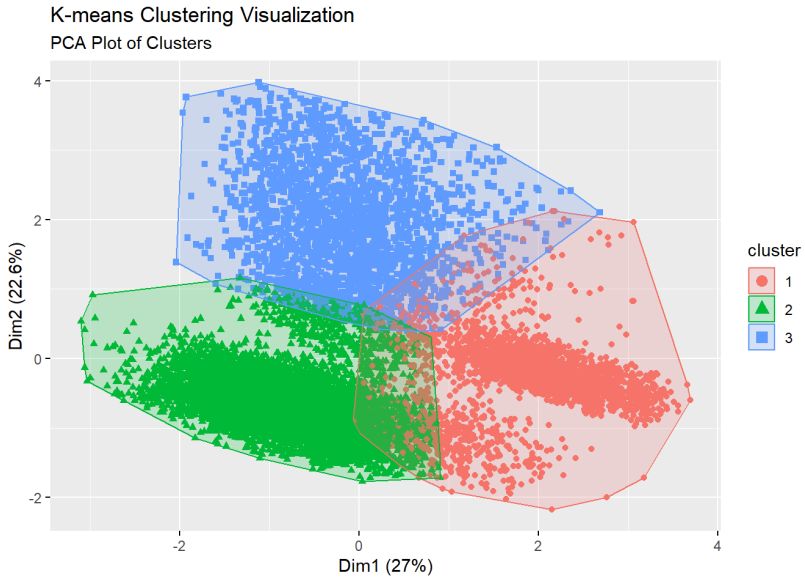
By focusing on numerical data attributes such as customer care calls, customer ratings, product costs, prior purchases, discounts offered, and product weights, the aim was to cluster similar observations together to reveal underlying patterns that affect these operational areas.

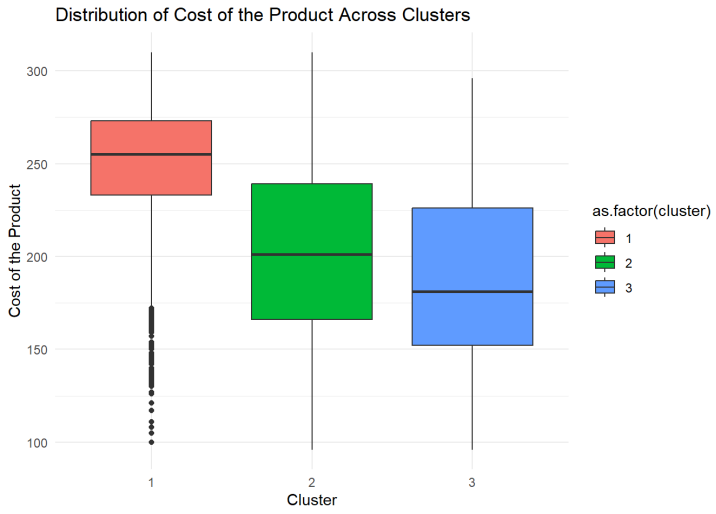
Key Steps in the Analysis:

- **Data Preparation:** Only numerical columns were selected and standardized to eliminate scale effects, ensuring equal weightage in distance computation used in clustering.
- **Optimal Cluster Determination:** Methods like the elbow and silhouette analyses helped identify the optimal number of clusters, which was determined to be three based on the analysis results.
- **Execution of Clustering:** With the optimal number of clusters identified, K-means clustering was executed. This process partitioned the data into three clusters, followed by a detailed examination of the cluster characteristics.




Optimal number of clusters
Silhouette Method
Average silhouette width
0.0
0.1
0.2
1
2
3
4
5
6
7
8
9
10
Number of clusters k


K-means Clustering Visualization
PCA Plot of Clusters
Dim2 (22.6%)
4
2
0
-2
-2
0
2
4
Dim1 (27%)
cluster
1
2
3



Detailed Results:

- **Cluster Formation:**
  - **Cluster 1 (2630 observations):** Featured higher average values in customer care calls, product cost, and prior purchases, indicating potentially higher-value customers who might require more complex logistics solutions.
  - **Cluster 2 (6110 observations):** This cluster presented average to below-average features across all variables, suggesting standard logistical needs and typical customer behaviour.
  - **Cluster 3 (2259 observations):** Characterized by high discount levels and lower product weights, indicating a segment possibly focused on less costly, more frequent purchases.

Analysis Aligned with Objectives:

- **Objective 1 - Evaluate Warehouse Efficiency:**
  - **Clusters' Relevance:** Cluster 1, with its high engagement (e.g., customer care calls and prior purchases), might indicate a group where warehouse efficiency can significantly impact service quality. In contrast, Clusters 2 and 3 might represent less demanding segments in terms of immediate access or specialized storage.
- **Objective 2 - Assess Transportation Mode Impact:**
  - **Insights:** Cluster 1 might prefer faster, possibly more expensive transportation options due to higher product costs and purchase frequency. Cluster 3, with lower product weights and higher discounts, might be more flexible regarding transportation mode, favouring cost over speed.
- **Objective 3 - Optimize Shipping Discounts:**
  - **Strategy Development:** Cluster 3's high sensitivity to discounts suggests that shipping discount strategies could be particularly effective. In contrast, Cluster 1 might respond better to premium service offerings that enhance customer experience rather than discounts.

Comparision (Predictive Models):

- **Accuracy:** Logistic Regression has the highest test accuracy (0.7154), indicating the best generalization to new data.
- **Sensitivity:** Logistic Regression and Gradient Boosting show high sensitivity (0.9849 and 0.9526, respectively), making them effective at identifying on-time deliveries.
- **Specificity:** Logistic Regression has the highest specificity (0.4447) among the models, though it is still relatively low, indicating a general challenge in identifying delayed deliveries.
- **Predictive Values:** Logistic Regression has the highest Positive Predictive Value (0.6404) and Negative Predictive Value (0.9670), indicating strong confidence in its predictions.
- **Kappa and Balanced Accuracy:** Logistic Regression has the highest Kappa (0.4301) and good balanced accuracy (0.7148), making it the most reliable model.

Best Model Choice:

- Logistic Regression holds the highest accuracy among the predictive models and would be the preferred choice if we prioritize accuracy.
- For understanding the data structure and preparing the dataset for better predictive modeling, employing PCA might be beneficial.
- If the goal is to understand customer segments or operational groupings better, K-Means Clustering could provide valuable insights.

Recommendations:

- **Warehouse Organization and Efficiency**
    - **Optimize Storage in Block F:**

Our analysis indicates that 33% of the products are stored in Block F. This suggests that Block F might be strategically positioned to maximize efficiency and accessibility. By centralizing a significant portion of inventory in a highly efficient zone, we can streamline the movement of goods, reduce picking and packing times, and enhance overall warehouse operations. It is recommended to evaluate the layout of Block F and ensure that it is utilized optimally to maintain high efficiency and accessibility.

- **Shipping Methods and Their Impact**
    - **Sea Freight for Large Quantities:**

Shipping by sea is an efficient method for transporting large quantities of goods over long distances, which is commonly used in international trade. The predominance of sea freight in our shipments suggests that many products are meant for bulk transportation over extended distances. This method not only helps in managing large volumes but also optimizes shipping

costs. It is crucial to maintain efficient product shipment methods to ensure customer satisfaction and minimize delays.

Flexibility in Transportation Modes:

Different modes of transportation have varying impacts on delivery times. Air freight is typically faster than sea shipping for long distances, making it ideal for urgent deliveries. However, road transportation, while effective for short distances, can be affected by factors such as weather, transportation issues, or customs delays. It is essential to maintain flexibility and have contingency plans to address such disruptions, ensuring timely deliveries and managing customer expectations effectively.

- **Product Importance and Shipping Decisions**
    - **Balancing Costs and Customer Expectations:**

The importance of the products significantly affects shipping decisions and the application of discounts. Businesses must strike a balance between meeting customer expectations, maintaining profitability, and managing shipping expenses. Offering shipping discounts or incentives, such as free or reduced-cost shipping, for high-importance products can encourage their purchase, leading to higher customer ratings and increased sales.

Prioritizing High-Importance Products:

To maintain customer satisfaction and ensure the success of high-importance products in the market, it is vital to prioritize their efficient delivery. This involves careful planning of shipping methods, handling, and packaging to meet or exceed customer expectations. Ensuring a high-quality shipment experience for these products can lead to positive feedback and foster customer loyalty.

- **Customer Care and Delivery Timeliness**
    - **Reducing Customer Care Calls:**

Customers are more likely to contact customer care when their orders are delayed or when there are issues unrelated to delivery timeliness, such as product quality or billing concerns. Addressing delivery efficiency and timeliness is crucial for reducing the volume of customer care calls. Improving these aspects can significantly enhance customer satisfaction and reduce the operational burden on customer support teams.

Strategic Focus on Delivery Efficiency:

By prioritizing the efficient delivery of high-importance products, companies can maintain customer satisfaction, meet strategic goals, and ensure the success of these products in the market. Implementing strategies to improve delivery efficiency, such as optimizing warehouse operations and utilizing the most appropriate shipping methods, can contribute to a more reliable and positive customer experience.

Conclusion

This study aimed to enhance logistics performance and customer satisfaction through a comprehensive analysis of warehouse efficiency, transportation mode effectiveness, and the impact of shipping discounts. The key objectives included identifying optimal warehouse block allocations, evaluating the influence of various transportation modes on delivery performance, and assessing the impact of shipping discounts on customer satisfaction and sales.

- Through the application of various machine learning models, including PCA and K-means clustering, we were able to uncover significant patterns in warehouse organization. The PCA results indicated that certain variables, such as product importance and mode of shipment, significantly contribute to the variance in warehouse efficiency. The clustering analysis revealed distinct groups of warehouse blocks that optimize accessibility and minimize retrieval times. These insights are crucial for designing better warehouse layouts, ultimately enhancing operational efficiency, reducing labor costs, and improving delivery punctuality.
- Our analysis of transportation modes highlighted the trade-offs between cost, speed, and reliability. The models showed that while air transport offers the fastest delivery times, it is also the most expensive, making it suitable for high-priority, time-sensitive products. Sea transport, though slower, is cost-effective for bulk shipments over long distances. Road transport provides a balanced approach for medium distances and moderate shipping volumes. Understanding these dynamics helps logistics managers make informed decisions, ensuring cost-effective and timely deliveries that align with customer expectations.
- The evaluation of shipping discounts revealed their substantial influence on customer purchasing behavior and satisfaction. Discounts, especially for high-priority products, significantly boost sales and enhance customer loyalty. However, the analysis also indicated the need for a balanced approach to maintain profitability. By identifying the optimal discount strategies, businesses can drive sales while managing shipping costs effectively.

Future research should focus on the dynamic aspects of logistics operations to further refine these strategies and adapt to evolving market conditions.