Data Storm 5.0

KJ Marketing

Presented By
DataDragons
(DataStorm549)

https://github.com/anush47/datastorm5.git

Team Members

- **Kahapola K. V.** - kaushalvirajk@gmail.com
- **Galappaththi A. S.** - anushangasharada@gmail.com
- **Wickramarathna H. K. G. V. L.** - gayani.20@cse.mrt.ac.lk

Powered By OCTAVE

Enhancing Customer Engagement through Personalized Marketing: A Data-Driven Segmentation Approach for KJ Marketing

---

Project Overview:

In Sri Lanka, KJ Marketing is a well-known chain of retail supermarkets with 22 outlets in both urban and suburban areas. The company sells a variety of commodities, such as luxury goods, fresh goods, and dry goods. The business has observed that traditional and standard marketing techniques are no longer as successful in attracting and retaining their present clientele in recent years. KJ Marketing wants to solve this problem by implementing a customized marketing plan based on the interests of each individual customer.

Objective:

This project's main goal is to provide an advanced analytical technique for classifying customers into different categories according to how they make purchases. The objective is to improve KJ Marketing's marketing strategy by using past sales data to create customized and targeted promotions.

1. Methodologies for Data Preprocessing:

Handling missing values:

- **Null Replacement:** Initially, 'null' and 'nul' strings in the dataset were converted to NaN values using the replace method.
- **Dropping Rows:** Rows with missing values in critical columns like 'outlet_city' and 'cluster_category' were dropped using dropna. This was done because these columns are crucial for our model, and missing values would significantly impact the predictions. Imputation could introduce significant bias or errors, especially with high variability in these categories. Moreover, maintaining data integrity is essential; with only 0.000258% missing values in 'outlet_city' and 0.000129% in 'cluster_category', the impact of dropping these rows is minimal and ensures robust and trustworthy model predictions.
- **Numeric Conversion and Text Mapping:** Non-numeric values in the sales columns ('luxury_sales', 'fresh_sales', 'dry_sales') were identified. A mapping function text_to_number was used to convert known text representations of numbers to numeric values. Any remaining non-numeric values were coerced to NaN and imputed using the mean sales values grouped by 'cluster_category' and 'outlet_city'.

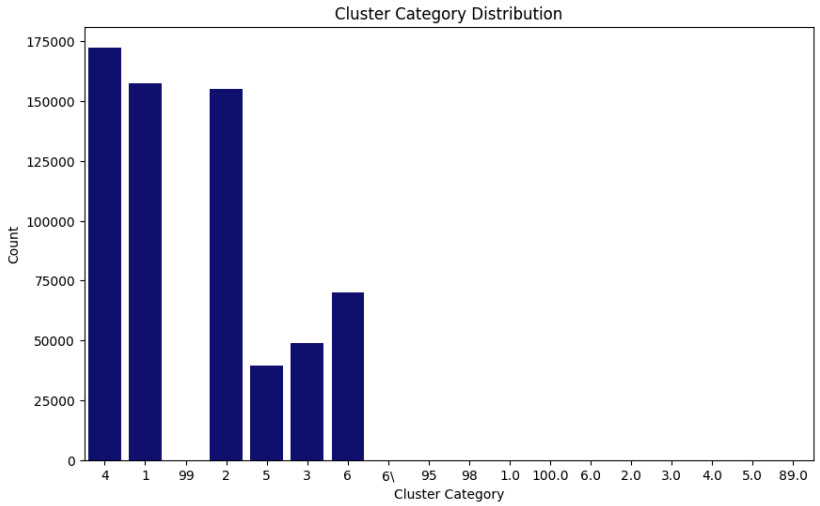
- **Imputation:** For missing values in 'luxury_sales', 'fresh_sales', and 'dry_sales' after the text conversion, imputation was performed using the mean of each combination of 'cluster_category' and 'outlet_city'.
  This method minimizes the risk of introducing bias that could occur with more generic imputation methods, such as global mean imputation, which might not account for differences across various clusters and cities. Additionally, using the mean gave more accuracy than the median or mode because the mean is a better representative measure for continuous data like sales figures, as it takes into account all values, providing a balanced approach that captures the average trend within each group.

Handling Duplicates:

- **Typo Correction:** Typos in the 'outlet_city' column were corrected by converting all text to lowercase and standardizing known incorrect entries (e.g., 'trincomale' to 'trincomalee').
- **Lowercase Conversion:** To identify and ignore duplicates effectively, the text data in relevant columns (e.g., 'outlet_city') was converted to lowercase. This ensures that entries like 'Peliyagoda' and 'PeliyagodA' are treated as duplicates and only one unique entry is retained.
- We verified that there were no duplicates in the Customer_ID column, ensuring data integrity and accurate analysis, as it serves as the primary key in our analysis.

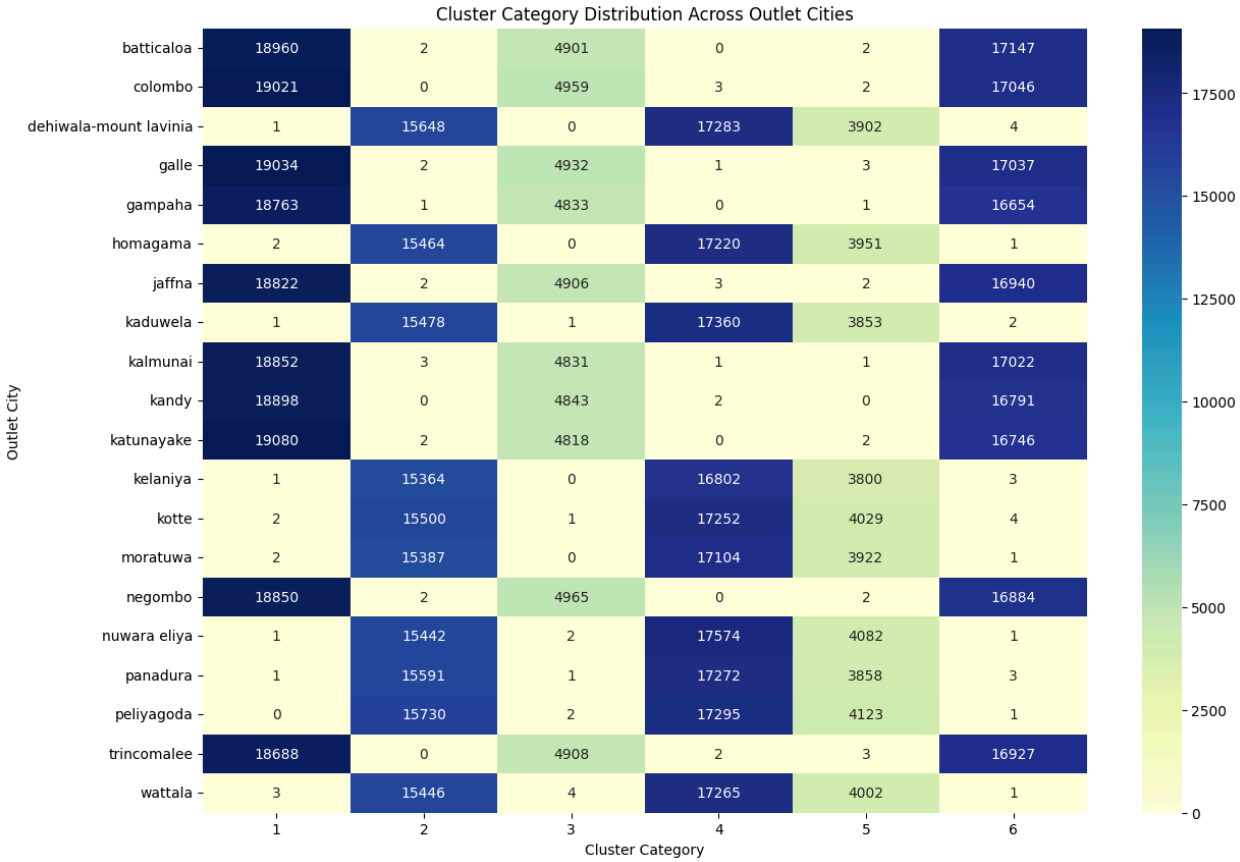
Handling Outliers:

- **Data Validation and Conversion:** Outliers in the 'cluster_category' column were removed by replacing invalid values (e.g., 89.0, 98, 100.0, '99') with NaN and then dropping those rows, as the question clearly mentions that there were 6 clusters.



Categorization of Outlet Cities:

To enhance the analysis and segmentation process, outlet cities were categorized into two distinct categories based on their sales patterns and customer demographics. This categorization aims to provide deeper insights into regional variations in sales behavior and facilitate more targeted marketing strategies.

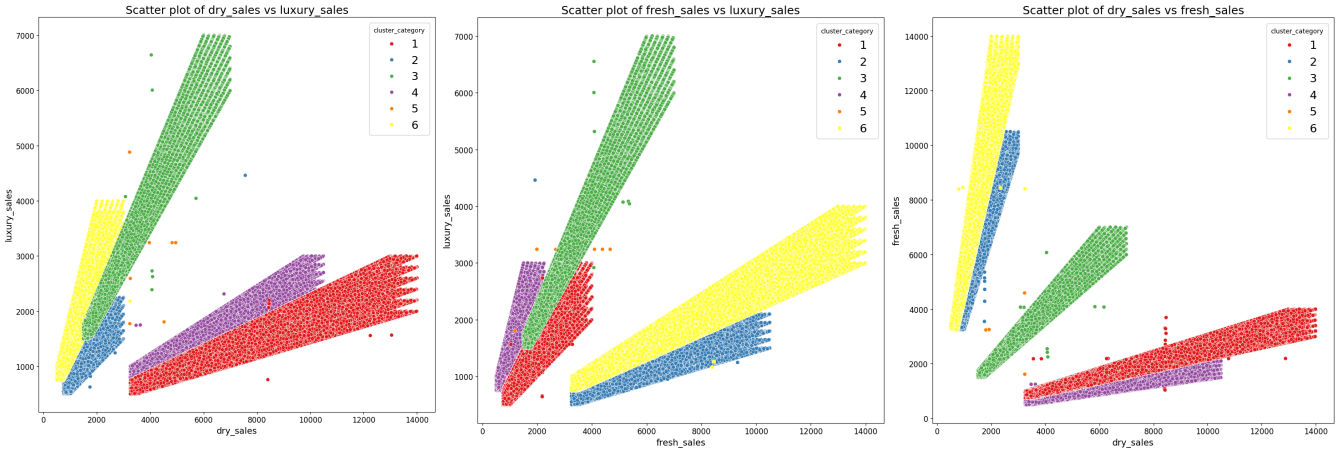


The categorization process involved:

- Identifying patterns in sales data across different outlet cities.
- Grouping outlet cities into two categories based on similarities in sales trends and customer preferences.
- Implementing the categorization within the preprocessing steps to ensure alignment with subsequent analytical techniques.

In the test dataset, two new outlet cities, Anuradhapura and Madawachiya, were discovered, which were not present in the training dataset. To categorize these cities, their sales metrics' mean values were compared to those of existing cities in the training data. By utilizing the Euclidean distance to measure the similarity between the sales metrics of the new and existing cities, we identified the most similar category for each new city.

2. Feature Correlation and Inter-feature relationships:



The scatter plots provide a visual representation of the relationships between different types of sales data (luxury_sales, fresh_sales, and dry_sales) across different clusters. These plots help in understanding the purchasing behavior of customers grouped into various clusters based on their sales patterns.

1. **Fresh_sales vs luxury_sales:**
   The analysis of fresh_sales versus luxury_sales reveals distinct consumer preferences across six clusters. Cluster 1 indicates a balanced inclination towards fresh and luxury items, with moderate sales in both categories. Cluster 2 prioritizes fresh items over luxury, showing significantly higher fresh sales. In Cluster 3, customers favor luxury items alongside moderate fresh sales. Cluster 4 represents budget-conscious consumers with low sales in both categories. Cluster 5 exhibits diverse spending habits, with moderate luxury sales and a broad range of fresh sales. Finally, Cluster 6 reflects a strong preference for fresh items with moderate interest in luxury products. These findings provide valuable insights into consumer behavior, guiding strategic marketing decisions.

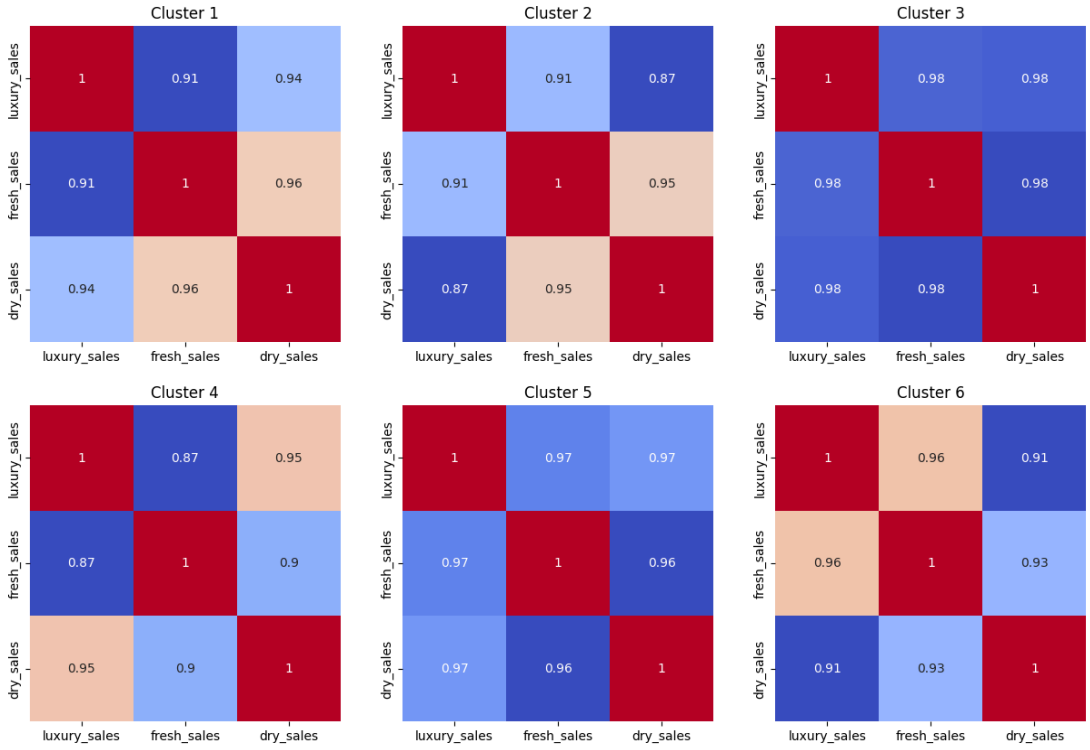
2. **Dry_sales vs luxury_sales:**
   The analysis of dry_sales versus luxury_sales across six clusters reveals distinct consumer preferences and spending behaviors. Cluster 1 exhibits moderate dry sales alongside low luxury sales, suggesting a focus on dry goods over luxury items. In Cluster 2, customers prioritize dry goods over luxury items, with low luxury sales and moderate dry sales. Cluster 3 showcases high luxury sales alongside moderate dry sales, indicating a preference for luxury products. Budget-conscious consumers are represented in Cluster 4, characterized by low luxury sales and very high dry sales. Cluster 5 displays moderate luxury sales alongside a wide range of dry sales, reflecting diverse spending habits.

Finally, Cluster 6 demonstrates a strong preference for dry goods with moderate interest in luxury products, showing high dry sales and moderate luxury sales. These insights inform strategic decision-making for targeted marketing and product offerings.

3. **Dry_sales vs fresh_sales:**

The comparison between dry_sales and fresh_sales across six clusters reveals diverse consumer spending behaviors. Cluster 1 demonstrates moderate sales in both fresh and dry categories, suggesting a balanced approach to spending. In Cluster 2, customers prioritize fresh items over dry goods, with high fresh sales and moderate dry sales. Cluster 3 exhibits balanced spending habits, characterized by moderate sales in both fresh and dry categories. Consumers in Cluster 4 show a preference for dry goods over fresh items, with very high dry sales and low fresh sales. Cluster 5 reflects diverse spending habits, displaying a wide range of sales in both fresh and dry categories. Finally, Cluster 6 showcases a preference for fresh items with some interest in dry goods, evidenced by high fresh sales and moderate dry sales. These insights provide valuable guidance for marketing strategies and product offerings tailored to different consumer preferences.

Inter-feature relationships



Insights

The scatter plots and correlation matrices provide several key insights:

- There is a consistent positive relationship between all types of sales (luxury, fresh, and dry), indicating that higher spending in one category often correlates with higher spending in others.
- Clusters show distinct purchasing behaviors, with some clusters focusing more on fresh or dry goods, while others show a balanced spending pattern across different categories.
- Understanding these relationships and patterns is crucial for tailoring marketing strategies to target specific customer segments more effectively.

These insights can be leveraged to develop targeted marketing campaigns, optimize product offerings, and improve customer satisfaction by addressing the specific needs and preferences of different customer segments.

3. Feature Selection:

Chosen Features:

1. **Sales Data:**
    - 'luxury_sales'
    - 'fresh_sales'
    - 'Dry_sales'

    These features were selected because they **directly relate to sales performance and are likely to influence the cluster categorization** according to the scatter plots drawn. Sales data provides crucial insights into the purchasing behavior and preferences of customers, which are essential for accurate clustering.

2. **Categorical Data:**
    - 'Outlet_city' (Categorized)

    The inclusion of the 'Outlet_city' feature, now categorized into two distinct groups, provides crucial geographical context that can influence sales patterns. By categorizing outlet cities into two categories based on their sales patterns and customer demographics, we enhance the granularity of our analysis and improve the effectiveness of our clustering algorithm. This enrichment allows us to capture regional nuances and tailor marketing strategies more effectively to the unique characteristics of each category.

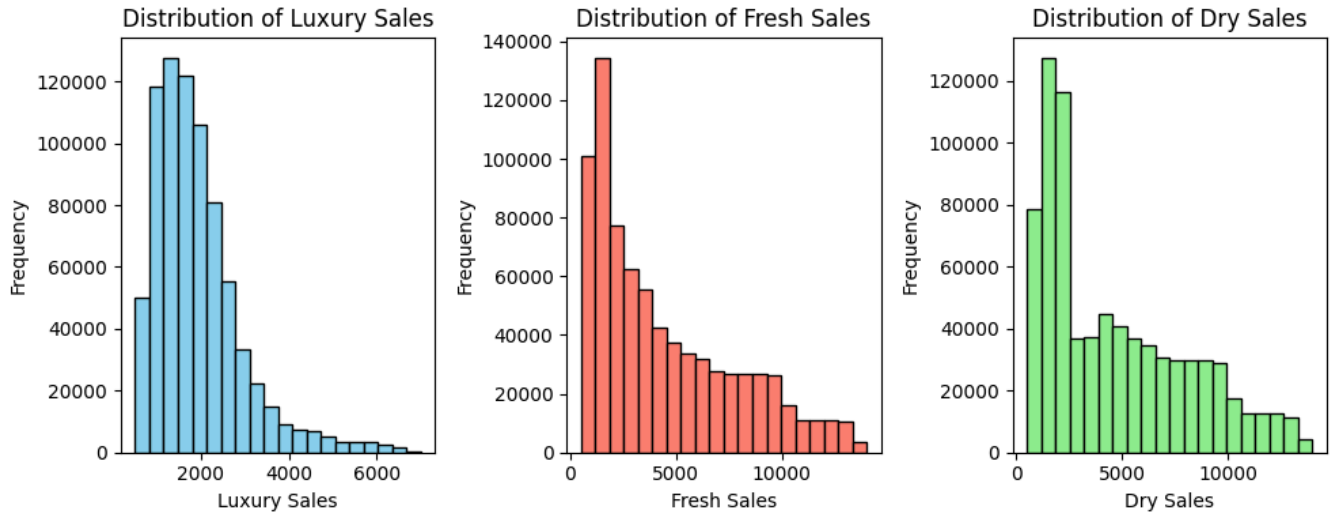
Relevance Determination:

- **Sales Data:** The chosen sales features are critical indicators of customer behavior and product performance. By analyzing the sales figures, we can better understand which products are popular in different regions and among various customer segments.
- **Categorical Data:** Including geographical information helps capture regional variations in sales, which is essential for creating accurate and meaningful customer segments.

Exclusion of Irrelevant Features:

- **Customer ID:** The Customer_ID feature was not considered because it is a unique identifier and does not provide any additional information relevant to the clustering task. Including it could introduce noise and negatively impact the model's performance. This decision helps ensure that the model focuses on features that contribute meaningfully to the clustering process.

4. Feature Scaling:



Scaling Method:

- **StandardScaler:**
  The numerical features ('luxury_sales', 'fresh_sales', 'dry_sales') were scaled using StandardScaler to standardize the features by removing the mean and scaling to unit variance. This ensures that each feature contributes equally to the model's performance, preventing features with larger scales from dominating the model.

5. Encoding Strategies:

Encoding Method:

- **One-Hot Encoding:** The 'outlet_city' column was encoded using one-hot encoding, converting the categorical text data into binary vectors. This prevents the model from assuming any ordinal relationship between the categories and makes the categorical data suitable for machine learning algorithms.

6. Target Variable and Interpretation:

- **Target Variable:** Cluster Category

  The "cluster_category" serves as the central element of our analysis, categorizing observations into six distinct groups labeled 1 through 6. This variable is crucial for understanding patterns within the data and guiding strategic decisions. It plays a key role in predictive modeling, serving as the target variable for model training and evaluation. Accurate interpretation of the cluster categories is essential for informed decision-making. The characteristics that define the different customer segments are detailed in cluster interpretation (section 9).

7. Model Selection:

Considered Algorithms:

- **Random Forest Classifier:** Chosen for its robustness, ability to handle large datasets, and ability to provide good performance without extensive parameter tuning. It also handles both numerical and categorical data well.
- Other algorithms like SVM, K-Nearest Neighbors, and Gradient Boosting could be considered, but Random Forest was chosen for its balance between performance and ease of implementation.

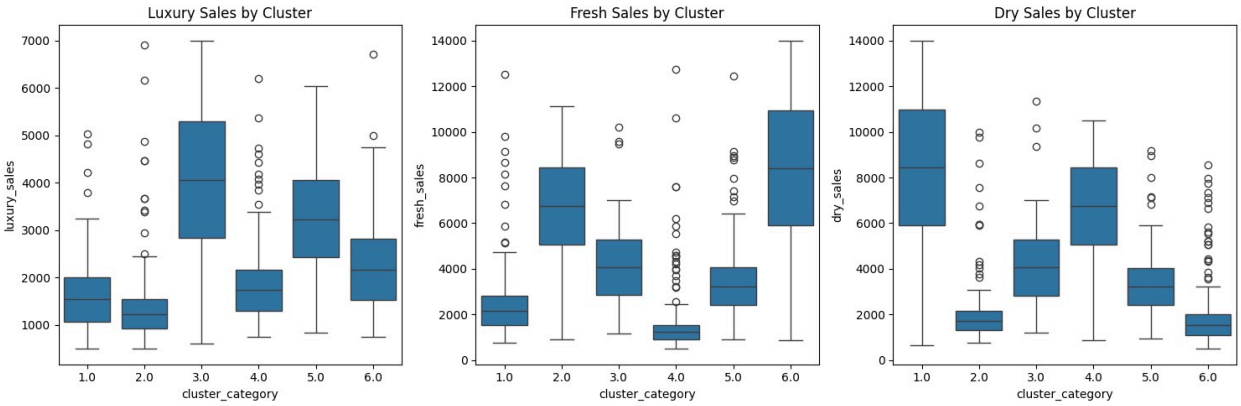
Evaluation Matrices:

In assessing the performance of our model, we utilized standard evaluation metrics such as **accuracy, recall, F1-score, and support**. These metrics provide valuable insights into the model's ability to accurately classify data points into their respective clusters, accounting for both correct and incorrect classifications. By considering these metrics, we gain a comprehensive understanding of the model's performance, enabling us to make informed decisions regarding its deployment and suitability for various applications.

8. Challenges and Solutions:

- Converting sales data from text to numbers was challenging; developed a custom function. Also manually checked and corrected typos.
- The main challenge in outlet city categorization was defining precise criteria to distinguish between the two categories based on sales patterns and customer demographics. This required careful analysis and interpretation of ambiguous sales trends, as well as robust validation processes to ensure the accuracy and reliability of the categorization outcomes.
- Confusion arose from consistently high accuracy. Employed validation sets and cross-validation; accuracy remained unchanged.
- Model runtime was considerable but justified by the output value it provided.
- There was a class imbalance within the train dataset. As a next step, we can implement solutions such as oversampling the minority class, undersampling the majority class, or using techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the class distribution and improve the model's performance.

9. Cluster Interpretation and Naming:



- Cluster 1: Dry Goods Shoppers
    - low spending on luxury items (median around 1500 units, with some outliers reaching up to around 5000 units).
    - low spending on fresh items (median around 2000 units, with some outliers reaching up to around 14000 units).
    - High spending on dry goods (median around 8000 units).
    - These customers are likely to be shoppers who purchase with a focus on dry goods.
    - These customers are primarily found in outlet category 1.

- Cluster 2: Fresh & Healthy Buyers
    - Low spending on luxury items (median around 1000 units).
    - Relatively high spending on fresh items (median above 6000 units).
    - Low spending on dry goods (median below 2000 units, with some outliers reaching up to around 10000 units).
    - These customers prioritize purchasing fresh items, indicating a focus on health or culinary preferences.
    - These customers are primarily found in outlet category 2.

- Cluster 3: Luxury Aficionados
    - High spending on luxury items (median around 4000 units).
    - Moderate spending on fresh items (median around 4000 units).
    - Moderate spending on dry goods (median around 4000 units)
    - These customers prioritize purchasing luxury items.
    - These customers are primarily found in outlet category 1.

- Cluster 4: Budget-conscious Consumers
    - Low spending on luxury items
    - Fresh sales are the lowest among all clusters.
    - High spending on dry items (median above 6000 units).
    - These customers may prioritize budget-friendly options and purchase fewer fresh items compared to other clusters.
    - These customers are primarily found in outlet category 2.

- Cluster 5: Diverse Spenders
    - This cluster has moderate luxury sales but low fresh and dry item sales.
    - The spread for luxury and dry goods is also quite significant indicating a varied customer base within this cluster.
    - Similar to Cluster 3, these customers have a high spending capacity and purchase a wide range of products across all categories, but lower than cluster 3 customers.
    - These customers are primarily found in outlet category 2.

- Cluster 6: Fresh & Luxury Enthusiasts
    - This cluster shows the highest sale of fresh items.
    - Luxury sales are moderate, higher than cluster 2.
    - Low spending on dry goods (median below 2000 units, with some outliers reaching up to around 8000 units).

- These customers prioritize purchasing fresh items, indicating a focus on health or culinary preferences, while still having a moderate interest in luxury items.
- These customers are primarily found in outlet category 1.

10. Marketing Strategy Enhancement

The company can enhance its marketing strategies by implementing targeted campaigns, providing personalized recommendations, and allocating resources effectively based on the classified clusters.

Enhanced Marketing Strategies:

- **Targeted Campaigns:**
  For targeted campaigns, the company can launch campaigns that highlight the variety and quality of dry goods for Cluster 1, promote the freshness and health benefits of fresh items for Cluster 2, showcase the exclusivity and premium nature of luxury items for Cluster 3, emphasize the affordability and value of products for Cluster 4, conduct segmented campaigns within Cluster 5 to cater to the varied customer base, and initiate campaigns that highlight the quality of fresh items and the moderate range of luxury items for Cluster 6.

- **Personalized Recommendations:**
  In terms of personalized recommendations, the company can recommend new arrivals or top-selling dry goods to Cluster 1, suggest fresh items based on previous purchases or popular trends to Cluster 2, recommend premium products or exclusive offers to Cluster 3 based on their buying behavior, suggest budget-friendly options or value deals to Cluster 4, provide personalized recommendations to Cluster 5 based on further segmentation within this cluster, and recommend fresh items and luxury items to Cluster 6 based on their preferences.

- **Resource Allocation:**
  Regarding resource allocation, the company can allocate more resources to ensure the availability and variety of dry goods for Cluster 1, invest in maintaining the freshness and quality of fresh items for Cluster 2, allocate resources to procure and maintain a diverse range of luxury items for Cluster 3, invest in sourcing and offering budget-friendly options for Cluster 4, allocate resources for further segmentation and personalized marketing within Cluster 5, and ensure the availability of fresh items and a moderate range of luxury items for Cluster 6.

- **Outlet category specific strategies:**
  The company can tailor advertisements and inventory based on the cluster composition of each outlet category. For Outlet Category 1, focusing on Clusters 1, 3, and 6, the emphasis would be on dry goods, luxury items, and fresh products. Outlet Category 2, comprising Clusters 2, 4, and 5, would highlight fresh items, budget-friendly options, and a varied range of products. This strategy ensures marketing efforts and inventory levels align with customer preferences, enhancing satisfaction and resource utilization.

By implementing these strategies, the company can effectively cater to the specific needs and preferences of different customer segments, thereby enhancing the effectiveness of its marketing strategies. It's also important to continuously monitor and adjust these strategies based on changing customer behaviors.

Conclusion:

By implementing this analytical solution, KJ Marketing can significantly improve its marketing effectiveness through tailored strategies that align with the unique preferences of its customer segments. The accurate classification of customers into distinct clusters will enable more focused and efficient marketing efforts, leading to increased customer satisfaction and loyalty.