Data storm 5.0

DataStorm539

MAY 20, 2024

DATA STORM ROTARACT

Table of Contents

GitHub Links

We have uploaded our notebooks to GitHub repository https://github.com/rithakith/Datastorm5.0 These are the notebook links:

- https://github.com/rithakith/Datastorm5.0/blob/main/feature-correlation.ipynb
- https://github.com/rithakith/Datastorm5.0/blob/main/preprocessing.ipynb
- https://github.com/rithakith/Datastorm5.0/blob/main/training.ipynb

1). Elaborate on the methodologies implemented to address missing values, duplicates and outliers within the dataset? Please describe any specific techniques used for imputation or exclusion, and the rationale behind these choices.

In the process of preparing the retail supermarket chain dataset for machine learning modeling was carefully cleaned, with special emphasis paid to addressing missing values, duplicates, and outliers, in order to prepare it for machine learning modeling. In addition to providing the reasoning behind the decisions made, this section describes the precise methods used for imputation or exclusion.

1. Handling Missing Values

Methodologies Implemented

I. Data Cleaning: Initial data cleaning involved examining the dry_sales, fresh_sales, and luxury_sales columns to identify and rectify any non-numeric values. This included handling null values (NaN) and any textual entries mistakenly included in numeric columns.

- Identified and isolated rows containing non-numeric values.
- Converted textual entries to appropriate numeric values or excluded them if conversion was not possible.
- Ensured all values in these columns were numeric before proceeding with further analysis.

II. Handling Missing and Incorrect Values in cluster_category: The cluster_category column was scrutinized for any missing or incorrect values that could affect the segmentation model's performance.

   - Removed a row containing NaN values to ensure that all entries in the cluster_category column were valid numerical values.
   - Corrected 6\ value to the correct category value of 6.
   - Ensured that the cluster_category column only contained values from the six identified categories, removing any entries with categories higher than 6 to maintain the integrity of the classification.

III. Imputation of Missing Values: To address missing values in key numerical columns (dry_sales, fresh_sales, and luxury_sales) imputation techniques were applied.

   - Replace missing values in the dry_sales, fresh_sales, and luxury_sales columns with the mean values of their respective columns. This approach assumes that the missing data points are likely to be similar to the average of the observed data.

IV. Handling Missing Values in customer_id: The customer_id column, a crucial identifier, had two missing entries which needed to be addressed to maintain data integrity.

   - Generated new unique customer IDs by identifying the current maximum value in the customer_id column and incrementing it by 1 for each missing entry. This ensured no duplicate IDs and maintained the uniqueness of each customer entry.

V. Addressing Missing Values in outlet_city: The outlet_city column contained some missing values which were critical for geographic analysis and segmentation.

   - Replaced missing values in the outlet_city column with the mode (most frequently occurring value) of the column. This method presumes that the most common city is a reasonable substitute for the missing entries, maintaining the distribution of data.

Techniques for Imputation/Exclusion:

- Mean Imputation: For numerical columns (luxury_sales, fresh_sales, dry_sales) where data is missing, the mean value of the column was used to fill in the missing entries. Mean imputation is appropriate when the data is normally distributed as it maintains the central tendency of the data.
- Mode Imputation: For categorical data in categorical columns (outlet_city) missing values were replaced with the mode (most frequent value) of the column. This method is chosen to preserve the most common category, thereby minimizing disruption to the distribution of the categorical variable.
- Data cleaning and outlier exclusion: To ensure data integrity in the cluster_category column, we employed outlier detection and removal techniques. We removed a row containing NaN values to ensure that all entries were valid numerical values. Additionally, we corrected a misentry of "6\" to the correct category value of 6. Finally, we ensured that the cluster_category column only contained values from the six identified categories, removing any entries with categories higher than 6 to maintain the integrity of the classification.

Rationale:

- Mean Imputation: This method is effective for maintaining the overall distribution and mean of the data. It is particularly useful when the dataset is large and the proportion of missing values is small.
- Mode Imputation: Preserving the most frequent category ensures that the overall distribution of the categorical variable remains consistent, which is crucial for maintaining the integrity of the data.

2. Handling Duplicates

Duplicate entries will be identified and removed to ensure the model doesn't learn from redundant data. Duplicates were identified using the duplicated() method to detect any repeated rows in the dataset.

Exclusion Technique**:**

- Dropping Duplicates: The drop_duplicates() function was used to eliminate identified duplicate rows from the dataset. In order to prevent bias and overfitting in the model, it is imperative that every element in the dataset be verified as unique.

Rationale:

- Data Integrity: Removing duplicates is vital to ensure that the model learns from unique instances only. Duplicate data can lead to overfitting, where the model performs well on training data but poorly on unseen data due to repetitive information.

3. Handling Outliers

Descriptive statistics (mean, standard deviation) and visual tools (box plots) were employed to identify potential outliers in the numerical columns (luxury_sales, fresh_sales, dry_sales). Techniques like Interquartile Range (IQR) used to identify potential outliers. Depending on the distribution and business context, outliers might be winsorized (capped to a specific percentile) or removed entirely.

Techniques for Handling Outliers:

- Data Cleansing and Outlier Detection: To ensure data integrity in the cluster_category column, we employed outlier detection and removal techniques. We removed a row containing NaN values to ensure that all entries were valid numerical values. Additionally, we corrected a misentry of "6" to the correct category value of 6. Finally, we ensured that the cluster_category column only contained values from the six identified categories, removing any entries with categories higher than 6 to maintain the integrity of the classification.

Rationale:

- Ensuring that all entries in the cluster_category column are accurate and within the valid range prevents errors in the segmentation process, maintains dataset integrity, and enhances the model's performance by avoiding skewed results from anomalous data points.

By using these techniques, we made sure the dataset is dependable, clean, and prepared for strong machine learning modeling, which will produce predictions that are more precise and broadly applicable.

2). Explain the features you chose for the above task. How did you determine their relevance to the problem?

Choosing criteria that effectively capture the key elements of consumer purchase behavior was a carefully considered process in order to classify clients into appropriate segments for individualized marketing tactics. Based on their direct relevance to the issue at hand and their capacity to yield insightful information, the following features were selected:

I. Outlet City (outlet_city) - The city where the retail outlet is located.
    - Geographical Influence: A customer's purchase decisions might differ greatly depending on where they are in the world because of things like area events, socioeconomic status, product availability, and local preferences.
    - Market Segmentation: Helps in understanding regional market segmentation, which is crucial for tailoring marketing strategies.

II. Luxury Sales (luxury_sales) - The average monthly sales per customer for luxury goods.
    - Purchasing Power: Spending on luxury goods is a strong indicator of a customer's purchasing power and lifestyle preferences.
    - Customer Segmentation: Customers who spend more on luxury items are likely to belong to different segments compared to those who spend less or none.

III. Fresh Sales (fresh_sales) - The average monthly sales per customer for fresh goods.

- Health and Freshness Preference: Reflects customer preference for perishable and potentially healthier food options, which can be an important segment differentiator.
- Consumption Patterns: Helps in understanding consumption patterns that are more immediate and frequent.

IV. Dry Sales (dry_sales) - The average monthly sales per customer for dry goods.

- Staple Goods: Spending on dry goods indicates a customer's purchasing habits for non-perishable items, which are typically staple goods with a longer shelf life.
- Frequency and Bulk Buying: Can help identify customers who buy in bulk or those with more regular, smaller purchases.

V. Cluster Category (cluster_category) - The category or segment assigned to each customer in the training dataset.

- Target Variable: This is the target variable for the classification task. It represents the segment to which each customer belongs based on historical data.

3). Has feature scaling or normalization been applied to the data? If so, which methods were utilized and explain how these techniques improve the performance of the model?

Yes, feature scaling and normalization were applied to the data to improve the performance of the model. Specifically, the StandardScaler from the sklearn.preprocessing module was used to standardize the sales-related features.

<u>Methods Utilized</u>

- The StandardScaler method:
  - This method standardizes features by removing the mean and scaling to unit variance, ensuring each feature has a mean of 0 and a standard deviation of 1.
  - The columns luxury_sales, fresh_sales, dry_sales, and total_sales were standardized in both the training and test datasets.

Impact on Model Performance

- Improved Convergence and Speed: Scaling features ensures uniform weight updates and faster convergence for algorithms like Random Forests and Logistic Regression, leading to quicker training times.

- Enhanced Model Accuracy: Scaling features improves the performance of distance-based algorithms, such as k-nearest neighbors and support vector machines, by preventing disproportionate influence from features with larger scales. In our Random Forest model, scaling ensured balanced feature influence, resulting in more accurate predictions.

- Reduced Model Bias: Standardization prevents features with larger ranges from dominating the model, ensuring that all features contribute equally. This reduces bias and improves the model's ability to generalize to new data.

- Improved Interpretability: Standardizing features makes the coefficients and feature importance scores more comparable, aiding in the interpretation of which features are most influential in predicting the target variable.

Feature scaling and normalization are crucial preprocessing steps that enhance the performance and interpretability of machine learning models. By standardizing the features, we ensure fairer comparisons, faster convergence, improved accuracy, reduced bias, and better interpretability of the model's results.

4). Have you used any encoding strategies? Provide a comprehensive explanation of the chosen encoding methods and their impact on the model’s input requirements and performance.

Yes, we have used the "one-hot" encoding method in building this model.The main category characteristic in this project, "outlet_city," was one-hot encoded. For machine learning models that are not naturally proficient at handling categorical data, this encoding technique is essential.

One-Hot Encoding: Application of this approach is performed on the categorical feature "outlet_city." Every distinct city is converted into an independent binary feature, where a value of 1 signifies the customer's affiliation with that city, and 0 denotes the opposite of that. This enables the algorithm to independently understand how each city affects consumer behavior.This expands the feature set, which may enhance the model's capacity to identify minute variations in consumer behavior depending on location.

Example: If the 'outlet_city' column has values ['Colombo', 'Kandy', 'Galle'], one-hot encoding will transform this into three binary columns: 'outlet_city_Colombo', 'outlet_city_Kandy', 'outlet_city_Galle'. A row with 'Colombo' will have values [1, 0, 0] in these columns, respectively.

<u>Impact on Model’s Input Requirements and Performance:</u>

- Input Requirements:
    - Enhanced Dimensionality: In this one-hot encoding method, the unique categories in the 'outlet_city' column are used to replace the 'outlet_city' column while increasing the number of input features.
    - No Ordinal Assumptions: Since there is no inbuilt ordering between the cities, one-hot encoding ensures that the model does not assume any ordinal relationship between them.

- Performance:

- Enhanced Interpretability of the Model: Each and every city's unique impact on the target variable can be individually identified by the model. This level of precision assists in providing the unique features and impacts of each city.
- Management of new Categories: When new categories are added to the data, one-hot encoding makes it relatively easy to expand it. These new categories won't have been observed by the existing model, but they can be easily included by retraining the model using an updated one-hot encoding strategy.

5). How do the features correlate with the target variable, and are there any notable inter-feature relationships?

Mainly two important techniques were used to analyze the relationship between features and the target variable (cluster_catgeory). Initially a Random Forest model was trained to extract the feature importance, and then a correlation analysis was performed using a correlation matrix. When combined, these techniques offer a thorough understanding of how distinct features relate to and affect the target variable.

Feature Importance from Random Forest Model

Feature importance scores that indicate how much each individual feature contributes to the predictive power of the model were obtained by training a Random Forest model. Fresh sales, dry sales, and luxury sales have a high importance, indicating that these sales categories are important in identifying customer clusters.

Correlation Analysis

The correlation matrix explains the linear relationships between features and the target variable, which supports the feature importance analysis. The results of the analysis show:

- High Correlation with Sales Features: There is a significant relationship between luxury sales, fresh sales, and dry sales with the cluster_catgeory. The Random Forest model's findings are supported by the high correlation that shows changes in sales are highly correlated with variations in customer clusters.

- Weak Correlation with One-Hot Encoded City Features: There are weak correlations between the target variable and the one-hot encoded outlet_city features. This implies that, although location does play a role in identifying customert groups, sales have a greater influence than location. The relationship between cluster_catgeory and city-specific effects is comparatively weak, suggesting that the goods that customers purchase have a greater influence on the target variable than the location of their purchases.

Combined Insights

An in-depth understanding of the factors affecting customer segmentation is obtained by integrating the correlation matrix analysis with the feature importance derived from the Random Forest model. The luxury sales, fresh sales, and dry sales features have a strong association with the target variable, indicating how important a role they play in distinguishing between different customer categories. Conversely, a limited link between city features and customer clusters shows that the customer clusters are more behaviorally oriented based on purchasing behaviors and less geographically driven.

In the context of the provided dataset, examining the relationships between features helps uncover patterns and dependencies that might influence customer purchasing behavior. Notable inter-feature relationships can reveal how different types of sales (luxury, fresh, and dry goods) correlate with each other and with total sales.

Inter-feature Relationships

- Sales Features: The features 'luxury_sales', 'fresh_sales', and 'dry_sales' were found to be positively correlated with 'total_sales'. This is expected as 'total_sales' is the sum of these three features.
- City Features: One-hot encoded city features have minimal correlation with one another, suggesting that they each contribute to the model independently and with minimal overlap.

https://www.kaggle.com/code/datastorm539/feature-correlation illustrates the correlation of features with the target variable.

6). Describe the target variable and interpret each category within it, detailing the characteristics that define the different customer segments.

The target variable ‘cluster_catgeory’ which consists of categorical data, classifies customers into six distinct segments based on their purchasing behavior. Each cluster is a collection of consumers with comparable purchasing habits, which enables the business to effectively target these groups with product offerings and marketing campaigns. A thorough explanation of each category is given below:

<u>Cluster 1</u>

- Expenditure on dry goods is significantly higher in this cluster than on luxury and fresh goods.
- They may put a greater focus on non-perishable goods than on premium or perishable ones, suggesting a preference for long-term stocking or maybe a focus on necessities.

<u>Cluster 2</u>

- These consumers spend less on luxury and dry products and more on fresh goods.
- They may value fresh food above non-perishable or luxuries, or they may be health-conscious, which could point to a lifestyle that is more focused on health.

<u>Cluster 3</u>

- This group of people puts a higher value on luxury products than on fresh and dry goods.
- These customers probably appreciate high-end goods and might be members of a wealthier population that places a higher priority on luxury items.

<u>Cluster 4</u>

- Consumers in this cluster spend a greater amount on dry goods but much less on luxury and fresh goods.
- This could indicate a preference for products with a longer shelf life as well as less expensive or careful spending habits.

Cluster 5

- These consumers spend less on fresh and dry goods and more on luxury items.
- Similar to cluster 3, but with less expenditure on luxury items, these consumers prefer high-end goods on a smaller scale or budget.

Cluster 6

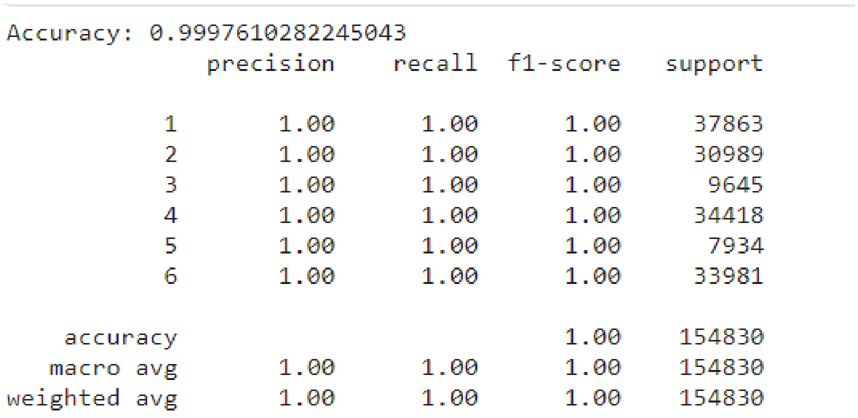
- This cluster of customers spends significantly more on fresh goods, a moderate amount on luxury goods, and a small amount on dry goods.
- This indicates a focus on fresh, healthful diet, much like cluster 2, but with a greater tendency to also spend money on luxuries, possibly signifying a budget that finds a balance between luxury and health.

7). What are the algorithms you considered for this problem, and why did you choose the final algorithm?

Initially, we examined our target variable, `cluster_category`, which is a categorical nominal variable. Since this is a supervised learning problem, we decided to use classification algorithms. We experimented with several popular classification algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Naive Bayes.

Details on how we implemented these algorithms using our processed training dataset can be found https://www.kaggle.com/code/datastorm539/training-and-clustering Based on the results, we observed that several algorithms outperformed the others in terms of accuracy. Consequently, we chose Random Forest as our final model due to its superior performance compared to the other algorithms.

Figure 1: Random forest Classification report



```
Accuracy: 0.9997610282245043
              precision    recall  f1-score   support

           1       1.00      1.00      1.00     37863
           2       1.00      1.00      1.00     30989
           3       1.00      1.00      1.00      9645
           4       1.00      1.00      1.00     34418
           5       1.00      1.00      1.00      7934
           6       1.00      1.00      1.00     33981

    accuracy                           1.00    154830
   macro avg       1.00      1.00      1.00    154830
weighted avg       1.00      1.00      1.00    154830
```

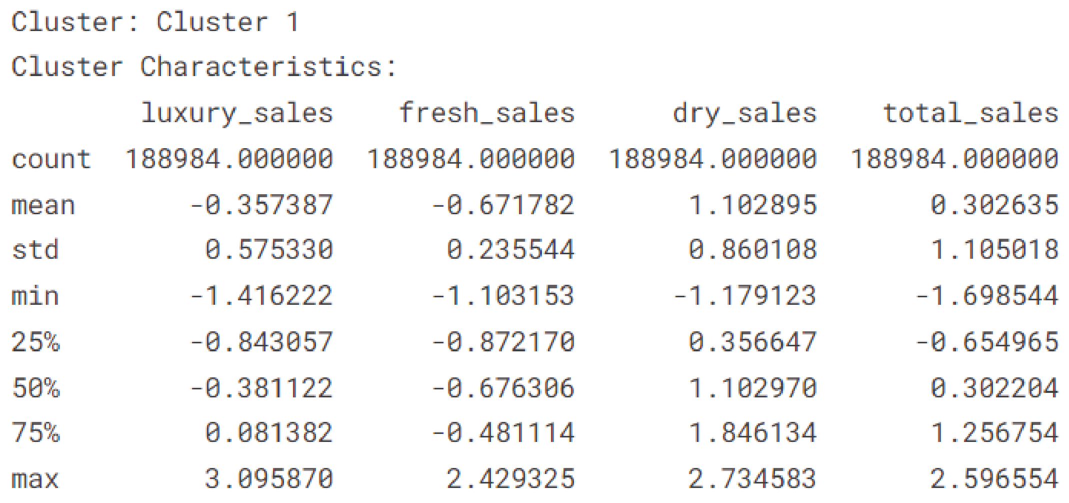
9). Briefly define and explain all the classified clusters while providing appropriate names.

In this analysis, we utilized **Hierarchical Clustering** to classify the data into distinct categories. The steps involved are as follows:

1. Data Preparation: The dataset was cleaned and pre-processed to ensure all features were in suitable formats.

2. Grouping by Cluster Category: We grouped the dataset based on the pre-defined `cluster_catgeory` column.

3. Descriptive Statistics: For each cluster, we calculated and analyzed key statistics (mean, standard deviation, minimum, maximum, etc.) for the relevant features, including `luxury_sales`, `fresh_sales`, `dry_sales`, and `total_sales`.

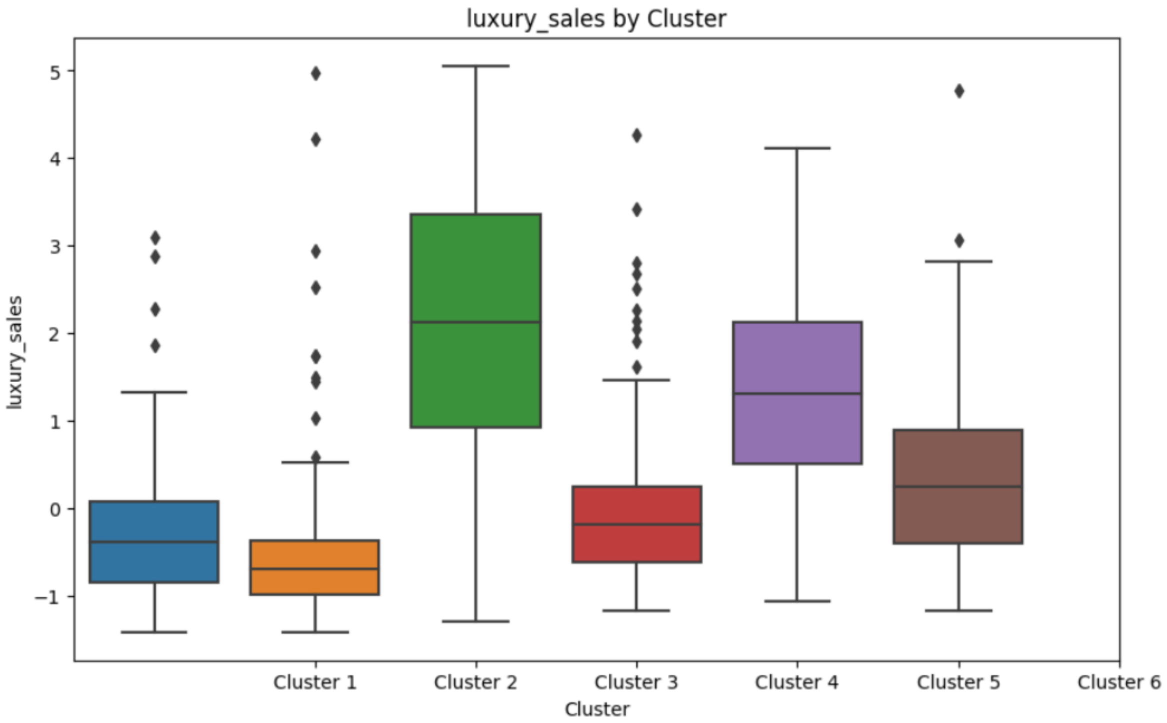


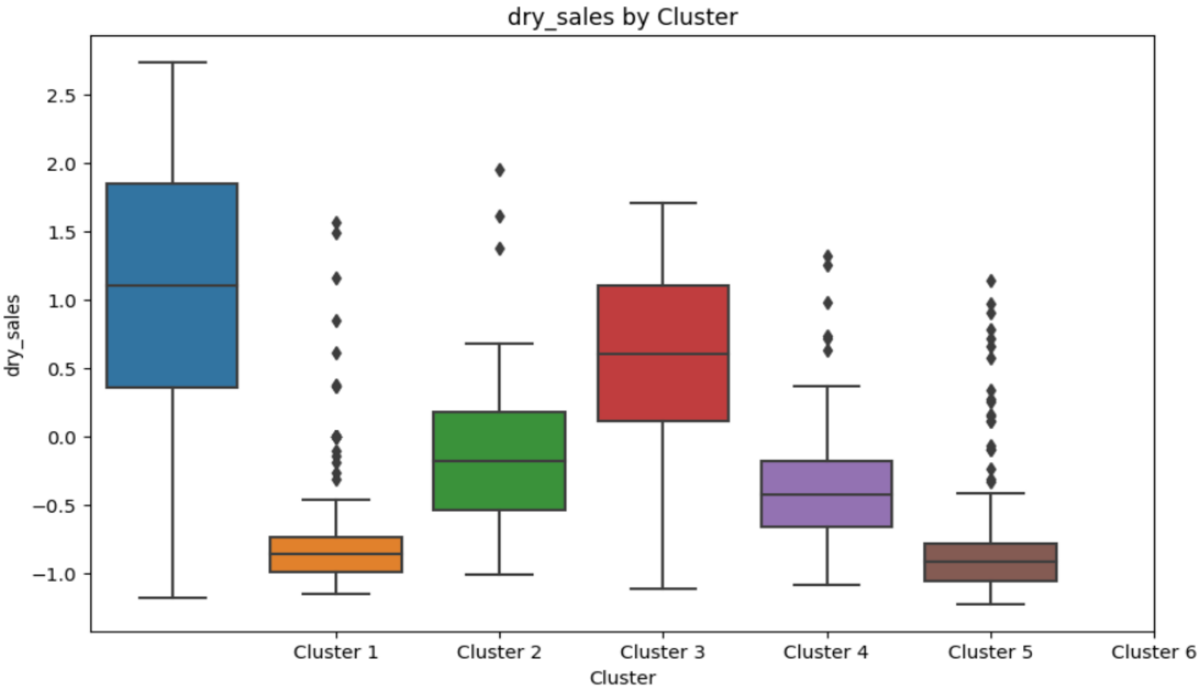
```
Cluster: Cluster 1
Cluster Characteristics:
        luxury_sales    fresh_sales      dry_sales    total_sales
count  188984.000000  188984.000000  188984.000000  188984.000000
mean       -0.357387      -0.671782       1.102895       0.302635
std         0.575330       0.235544       0.860108       1.105018
min        -1.416222      -1.103153      -1.179123      -1.698544
25%        -0.843057      -0.872170       0.356647      -0.654965
50%        -0.381122      -0.676306       1.102970       0.302204
75%         0.081382      -0.481114       1.846134       1.256754
max         3.095870       2.429325       2.734583       2.596554
```

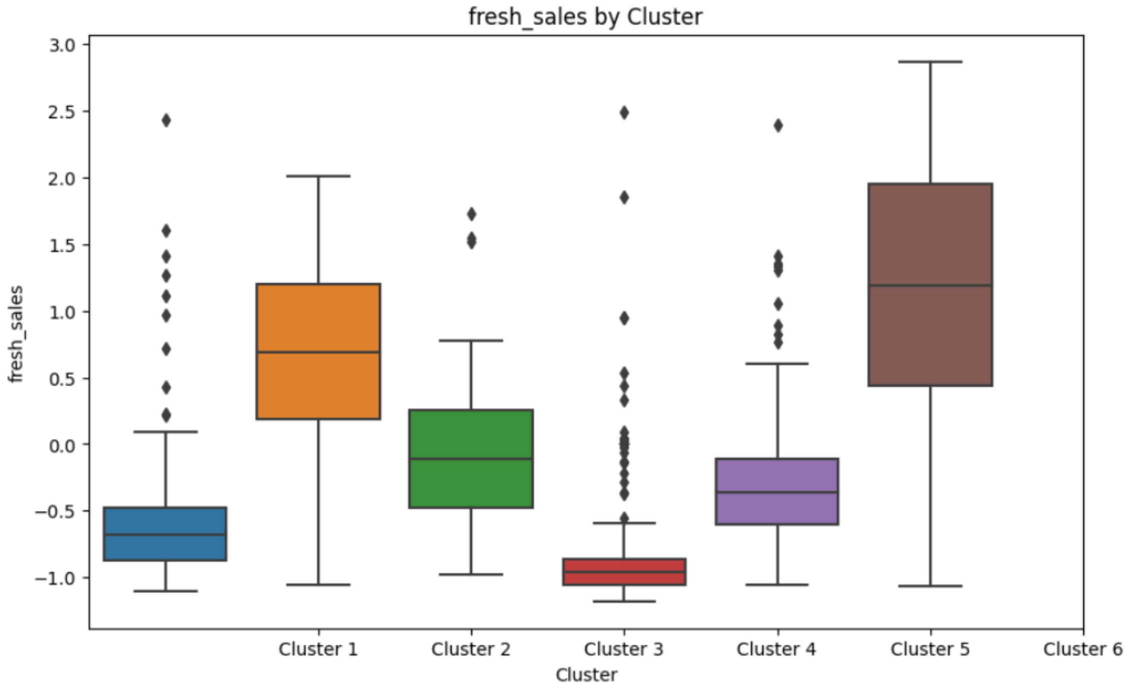
https://www.kaggle.com/code/datastorm539/training-and-clustering further illustrates the method used to group the data and compute the statistics for each cluster:

4. Visualization

To further analyze and visualize these clusters, we plotted the distribution of key features such as `luxury_sales`, `fresh_sales`, `dry_sales`, and `total_sales` for each cluster using box plots. This helped in identifying trends and patterns unique to each cluster, which can significantly enhance the effectiveness of targeted marketing strategies.


luxury_sales by Cluster
luxury_sales
5
4
3
2
1
0
−1
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster


dry_sales by Cluster
dry_sales
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster



Cluster 1: The Necessity Shoppers (blue box plot)

Characteristics:

High Expenditure on Dry Goods: This cluster significantly prioritizes spending on non-perishable items.

Lower Spending on Luxury and Fresh Goods: There is a noticeable lower expenditure on premium and perishable goods.

Interpretation: These customers likely focus on long-term stocking and essential items, suggesting practical or budget-conscious shopping behaviors.

Cluster 2: The Health Enthusiasts (orange box plot)

Characteristics:

Higher Spending on Fresh Goods: Customers in this cluster prioritize fresh food items.

Lower Spending on Luxury and Dry Goods: There is less expenditure on non-perishable and luxury items.

Interpretation: This group may be health-conscious, placing a higher value on fresh, perishable goods, possibly indicating a lifestyle centered around health and wellness.

Cluster 3: The Luxury Seekers (green box plot)

Characteristics:

High Expenditure on Luxury Goods: This cluster focuses heavily on premium, high-end products.

Lower Spending on Fresh and Dry Goods: There is a reduced emphasis on both perishable and non-perishable necessities.

Interpretation: These customers likely come from a wealthier demographic that prioritizes luxury items, indicating a preference for high-end goods and experiences.

Cluster 4: The Practical Buyers (red box plot)

Characteristics:

High Expenditure on Dry Goods: Similar to Cluster 1, this group prioritizes spending on non-perishable items.

Lower Spending on Luxury and Fresh Goods: There is minimal expenditure on premium and perishable goods.

Interpretation: These consumers may prefer long shelf-life products and demonstrate careful spending habits, focusing on practicality and cost-efficiency.

Cluster 5: The Selective Spenders (purple box plot)

Characteristics:

Moderate Expenditure on Luxury Goods: There is a preference for high-end items, though not as pronounced as in Cluster 3.

Lower Spending on Fresh and Dry Goods: Less focus on both perishable and non-perishable essentials.

Interpretation: These consumers prefer luxury goods but within a constrained budget, indicating selective, value-conscious spending on premium items.

Cluster 6: The Balanced Shoppers (brown box plot)

Characteristics:

High Expenditure on Fresh Goods: This cluster prioritizes spending on fresh, perishable items.

Moderate Spending on Luxury Goods: There is a balanced expenditure on premium items.

Low Spending on Dry Goods: Minimal focus on non-perishable essentials.

Interpretation: Customers in this cluster balance between a healthful diet and luxury spending, suggesting a lifestyle that values both health and occasional indulgence in premium goods.

10). How can your solution enhance the effectiveness of the company's marketing strategies based on the classified clusters?

By leveraging the insights gained from the classified clusters, the company can tailor its marketing strategies to better meet the needs and preferences of each customer segment. Here's how our solution can enhance marketing effectiveness:

1. Targeted Promotions and Discounts:

Cluster 1 - The Necessity Shoppers:

- Strategy: Focus on promotions for bulk purchases and discounts on essential non-perishable items.

- Benefit: Encourages frequent purchases and builds customer loyalty among budget-conscious shoppers.

Cluster 2 - The Health Enthusiasts:

- Strategy: Promote fresh and organic food products, with special deals on health-related items.

- Benefit: Aligns with their health-conscious lifestyle, increasing their likelihood of repeat purchases.

Cluster 3 - The Luxury Seekers:

- Strategy: Highlight premium and exclusive products through targeted ads and personalized offers.

- Benefit: Appeals to their preference for high-end items, driving higher-margin sales.

2. Customized Product Recommendations:

Cluster 4 - The Practical Buyers:

- Strategy: Recommend cost-effective products with a long shelf life, focusing on value-for-money offerings.

- Benefit: Enhances the shopping experience by meeting their practical needs, increasing customer satisfaction.

Cluster 5 - The Selective Spenders:

- Strategy: Suggest a mix of moderately priced luxury items and essential goods, balancing quality and cost.

- Benefit: Matches their selective spending habits, encouraging them to explore a broader range of products.

3. Personalized Marketing Campaigns:

Cluster 6 - The Balanced Shoppers:

- Strategy: Develop campaigns that emphasize both fresh, healthful foods and occasional luxury treats.

- Benefit: Resonates with their balanced lifestyle, fostering a deeper emotional connection with the brand.

4. Optimized Inventory Management:

- Insight: Understanding the purchasing patterns of each cluster allows the company to optimize inventory levels for different product categories.

- Benefit: Reduces stockouts and overstock situations, ensuring that popular items are always available for the right customer segments.

5. Enhanced Customer Engagement:

- Cluster-Specific Engagement: Use cluster insights to tailor communication strategies.

- Example: For Cluster 2 - The Health Enthusiasts, send newsletters with tips on healthy living and recipes featuring fresh produce.

- Benefit: Increases engagement and builds a community around shared interests and values.

6. Geographic-Specific Strategies:

- Insight: Leverage outlet city information to customize marketing efforts based on regional preferences.

- Strategy: For cities with a higher concentration of **Cluster 3 - The Luxury Seekers**, prioritize the availability and promotion of luxury items.

- Benefit: Ensures marketing efforts are regionally relevant and effective.

7. Loyalty Programs and Membership Benefits:

- Insight: Develop loyalty programs tailored to the spending habits and preferences of each cluster.

- Example: For Cluster 1 - The Necessity Shoppers, offer points and rewards for frequent purchases of essential items.

- Benefit: Encourages repeat business and strengthens customer loyalty.

8. Cross-Selling and Upselling Opportunities:

- Insight: Use cluster characteristics to identify cross-selling and upselling opportunities.

- Strategy: For Cluster 5 - The Selective Spenders, suggest complementary luxury items when they purchase their regular products.

- Benefit: Increases the average order value and enhances the customer shopping experience.

By implementing these targeted strategies, the company can more effectively meet the diverse needs of its customer base, leading to increased customer satisfaction, loyalty, and ultimately, higher sales and profitability.