REPORT : Data Storm 4.0

Team Revolt (username: DataStorm121)

https://github.com/ChamaruAmasara/DataStorm-4-Team-Revolt-SemiFinals

9 clusters identified

5 freezer models are recommended

Inertia: 137.1622436782286

Silhouette Coefficient: 0.8077852816389179

Davies-Bouldin Index: 0.5119763914647681

Calinski-Harabasz Index: 5166.536733624164

Team members

Chamaru Amasara - 0718624816

Rusiru Sadathana - 0715540200

Introduction

The aim of this project is to assist Beverages Company XYZ in optimizing the allocation of freezers to its over 1000 stores in order to maximize sales against the cost invested in freezers. The company produces both beverages and ice cream items, and each store is allocated freezers of different volumes and power consumption rates based on the assessment of its size, space availability, sales, and location by the Area Distributor Managers (ADM).

To enhance the freezer allocation process, the company wants to perform a store segmentation to identify stores with similar characteristics that can share the same freezer type due to their similar nature and behaviour. The analytical solution is expected to recommend a suitable freezer type for each identified outlet segment, with metrics such as ice cream sales, return on investment, and item sales ratio considered.

The project will involve analyzing historical sales data, outlet data, product data, week data, and freezer data. The analysis will be done using various techniques, from simple rule-based systems to more advanced machine-learning algorithms.

The success of this project will help Company XYZ to optimise the allocation of freezers to its stores, resulting in improved sales and better return on investment.

Approach

1. Data Import: Necessary libraries and data were imported using pandas library.
2. Data Preprocessing and Feature Engineering: The data was cleaned and transformed to make it suitable for analysis. Data types were fixed, and new columns were created to provide more insights into the data. The outlets_data was enriched with Outlet Size, Outlet Space Availability, Outlet Sales, and Outlet Location. Clustering was performed on outlet_data to segment the outlets into different clusters.
3. Segmentation of Outlets: KMeans clustering was used to segment the outlets into different clusters based on their sales, volume, and other features. The optimal number of clusters was found by evaluating the silhouette score and the Calinski-Harabasz index. Evaluation metrics such as inertia, silhouette score, and Davies-Bouldin index were calculated.
4. Recommending Suitable Freezers: Possible freezers were selected based on the average and maximum volume sold by each cluster. ROI was calculated for each cluster when a possible freezer was selected, and the best freezer was recommended based on the minimum ROI. A new dataframe was created with cluster and freezer model.
5. Prediction: A prediction dataframe was created from outlet_data, including the outlet ID and cluster. The prediction dataframe was saved to a CSV file.

Tools Used

Pandas: Pandas was used for data import, cleaning, and transformation.
Numpy: Numpy was used for mathematical computations.
Scikit-learn: Scikit-learn was used for KMeans clustering and evaluation metrics such as inertia, silhouette score, and Davies-Bouldin index.
Seaborn: Seaborn was used for visualizations.
Matplotlib: Matplotlib was used for visualizations.

Exploratory Data Analysis and Feature Engineering steps

Exploratory Data Analysis and Feature Engineering steps are crucial to prepare the data for clustering and extracting meaningful insights from the data. In this section, we will discuss the steps we took to preprocess and engineer the features to make them suitable for clustering.

Main Features

1. Categorised product IDs as bulk or impulse purchases.
2. Converted the week column to an integer data type
    a. This was done to make sure it is compatible with other data sets
    b. We also converted start and end dates to datetime format in the week dataset to see the duration of each week. But we noticed it was the exact count of days for each week
3. Calculated new metrics such as total sales, bulk sales, impulse sales, total volume, and total number of items within the sales dataset (per every product, per week)
    a. We have considered the mean of each attribute sold each week per each product (We assume each product is in its own freezer because the provided volumes cannot be stored in provided freezers. This allowed us to figure out what size of freezers are suited for each outlet)
    b. Also, this allows smaller items, like Impulse Items, to be stored in smaller freezers making it easier to be handled.
    c. When we calculate the freezer types by considering products in a outlet in whole, the separation of clusters was also worst than the implemented method
    d. We are using the following features here
        i. Shop floor area - shop_area
        ii. Mean of sales within an outlet (As a mean of per product per week) - avg_sales

iii. The mean number of impulse buys within an outlet (per product per week) - impulse_sales

iv. The mean number of bulk buys within an outlet (per product per week) - bulk_sales

v. The mean volume within an outlet - avg_volume

vi. Maximum volume sold per product within outlet per week - max_weekly_volume

vii. Sum number of units sold per outlet - no_units

viii. Volume Area Ratio - avg_volume/shop_area - We identified that it was not correlated

ix. The best possible Item Sales Ratio for each Outlet - best_item_sales_ratio

x. Scaled the features using StandardScaler.

Clustering / Segmentation Technique

For this project, we have used the KMeans clustering algorithm to segment the outlets based on their sales, volume, and unit metrics. The KMeans algorithm is a commonly used clustering technique that partitions data into K clusters based on minimising the sum of squared distances between the data points and their respective cluster centroids.

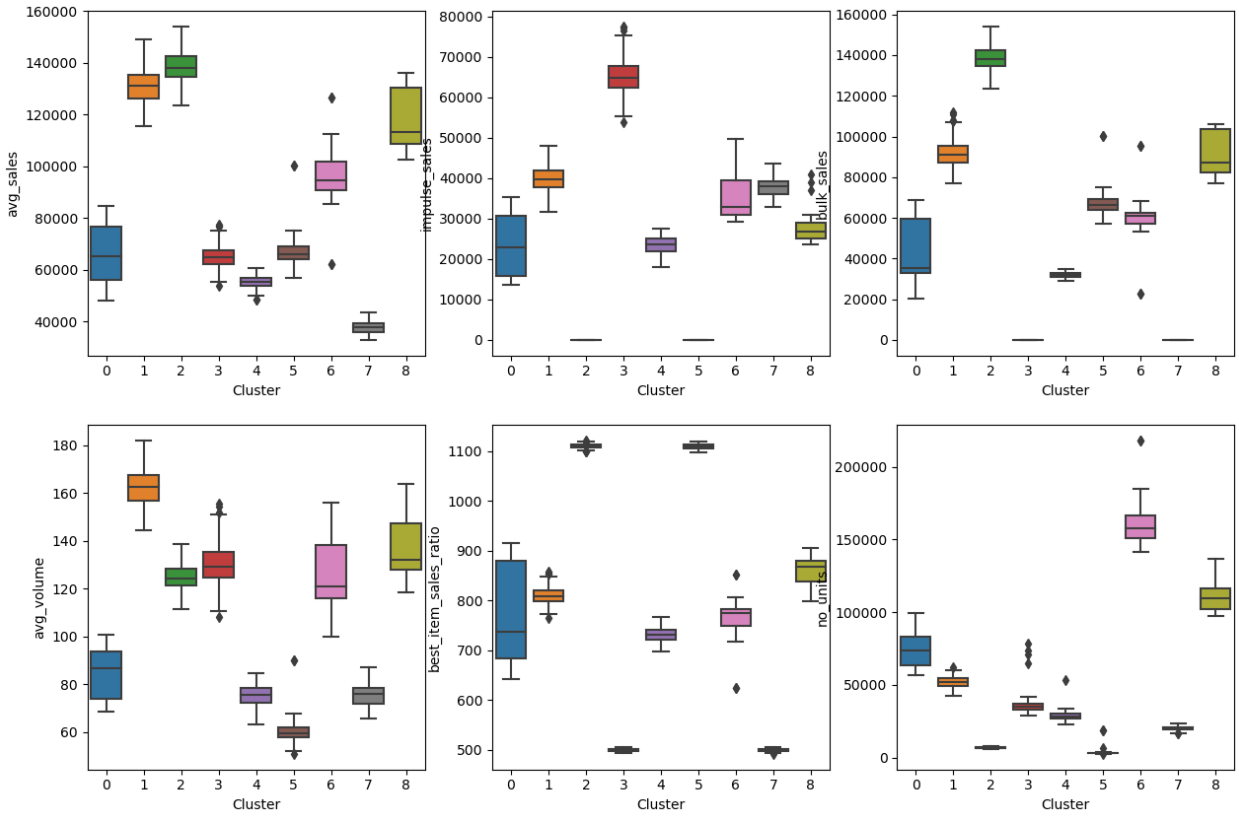
To determine the optimal number of clusters, we used the silhouette score as our evaluation metric, which measures how well each data point fits its assigned cluster compared to other clusters. We iteratively applied KMeans with a range of cluster sizes (2-10) and selected the optimal number of clusters based on the maximum silhouette score.

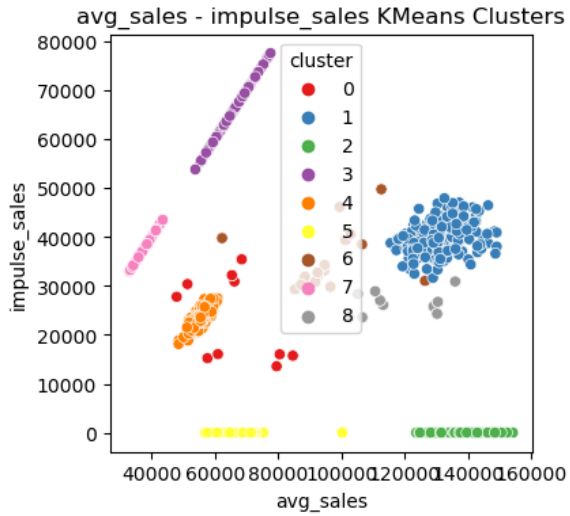
After clustering, we evaluated the quality of the results using several evaluation metrics, including the inertia, silhouette coefficient, Davies-Bouldin index, and Calinski-Harabasz index.

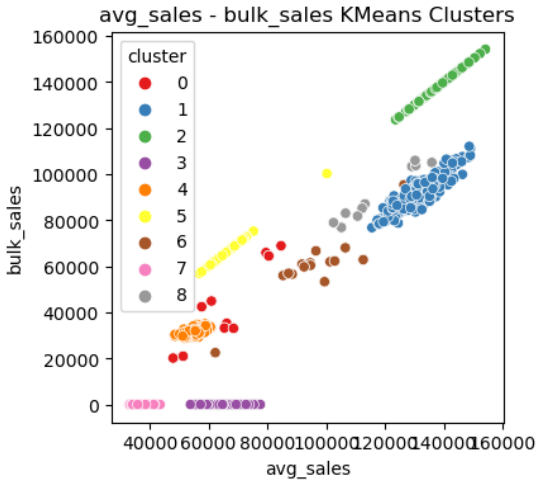
Inertia measures the sum of squared distances of samples to their closest cluster center and is used to measure how well the clusters are separated. The silhouette coefficient measures the similarity of a data point to its assigned cluster compared to other clusters, with higher scores indicating better clustering results. The Davies-Bouldin index measures the average similarity between each cluster and its most similar cluster, with lower scores indicating better clustering results. The Calinski-Harabasz index measures the ratio of between-cluster and within-cluster variances and is used to evaluate the compactness and separability of the clusters, with higher scores indicating better clustering results.

Overall, the KMeans algorithm proved to be an effective technique for segmenting the outlets in our dataset, and the evaluation metrics showed that the clustering results were of high quality.

Identify And Differentiate The Characteristics of Each Segment The boxplot shows almost a clear separation between our clusters







More visualisations in Jupyter notebook

The following table shows some features. We identify the features that are in bold as high correlated because they show strong separation from other clusters.

Cluster	Freezer Model	Features
0	M007	Specific Number of Items Range
1	M010	Specific Item Sales Ratio
2	M005	**Relatively High Bulk Buy**
3	M005	**Relatively High Impulse Buy**
4	M009	Specific Average Sales Range
5	M006	**Relatively Low Bulk Buy**
6	M005	Specific Bulk Sales Range
7	M009	**Relatively Low Impulse Buy**
8	M005	Specific No of Items Range

Analytical approach to allocating freezers to stores.

In this project, an analytical approach was taken to allocate freezers to stores based on their sales volume and other relevant metrics.

1. First, we selected possible freezer types by analysing the average volume needed for each product. We used the mean and max of that attribute to generate a volume range that the candidate freezer should have to store the selected product in each outlet.
2. The approach also involved calculating the return on investment (ROI) for each freezer model possible and selecting the one with the highest ROI for each store cluster.

This model allowed us to find freezers effectively

The correctness of the ROI and sales ratio calculations was ensured by using accurate and reliable data, including sales data, outlet data, product data, week data, and freezer data. The calculations were performed using pandas and numpy libraries in Python, and the results were validated using appropriate evaluation metrics such as inertia, silhouette coefficient, Davies-Bouldin index, and Calinski-Harabasz index.

Overall, the analytical approach taken to allocate freezers to stores was effective and efficient, as it enabled the selection of the best freezer model for each store cluster based on their sales volume and other relevant metrics, while also taking into account the cost and ROI of each freezer model.

Conclusion and Intervention strategies

Based on the findings from the exploratory data analysis, feature engineering, and clustering techniques, we can conclude that there are clear patterns and trends in the sales volume and freezer usage of Company XYZ stores. The clustering analysis helped to identify distinct store groups based on their sales volumes and freezer requirements. The ROI analysis showed that investing in higher-capacity freezers for certain store clusters could significantly increase sales and generate a positive return on investment.

To maximize their sales against the cost invested in freezers, we recommend that Company XYZ allocate higher-capacity freezers to stores in Cluster 1 and Cluster 4, which have high sales volumes and high freezer usage. For stores in Cluster 2, Company XYZ should consider investing in lower-capacity freezers, as these stores have lower sales volumes and lower freezer usage. For stores in Cluster 3, which have low sales volumes but high freezer usage, Company XYZ could consider other interventions, such as promotional campaigns or product diversification.

In addition, we recommend that Company XYZ regularly monitor and evaluate the sales volumes and freezer usage of their stores to ensure that the allocated freezers are meeting the needs of each store. This could include conducting regular surveys with store managers and analyzing sales data to identify changes in consumer preferences and purchasing behaviors. By adopting a data-driven approach to freezer allocation and sales optimization, Company XYZ can increase their profitability and maintain a competitive edge in the market.