TEAM ARES

Kaggle Username - DataStorm019

Data Storm 4.0

Semi-Final Round

No of clusters : 3

Lowest inertia : 3037.43

Highest Silhouette Coefficient : 0.55

Lowest Davies-Bouldin Index : 0.688

Highest Calinski-Harabasz Index : 1455.28

GitHub link : https://github.com/sandani98/DataStorm_4_semi_finals

Chalani Ekanayake
Sachini Chandanayake
Sandani Jayawardena

Project Description

1.1. Business Problem

When allocating assets, businesses are in need of optimally allocating them which would improve their sales against the cost invested in the asset. When the number of outlets of a company increases, it becomes hard to accomplish that manually. Thus, the AI analytic solutions come in handy where outlets of the similar characteristics could be segmented. Then the most suitable asset allocation that would benefit the company could be recommended for a particular cluster.

1.2. Use Case Definition

In the challenge, we are required to design and develop an advanced analytics solution that can be used by Beverages Company XYZ to perform store segmentation (identify the stores with similar characteristics). The purpose of the segmentation is to recommend a suitable freezer for each identified outlet segment to maximize the **Return of Investment (RoI)** (ratio of ice cream sales to the freezer maintenance and power consumption cost) and **Item Sales Ratio** (ratio of ice cream sales volume to the freezer capacity).

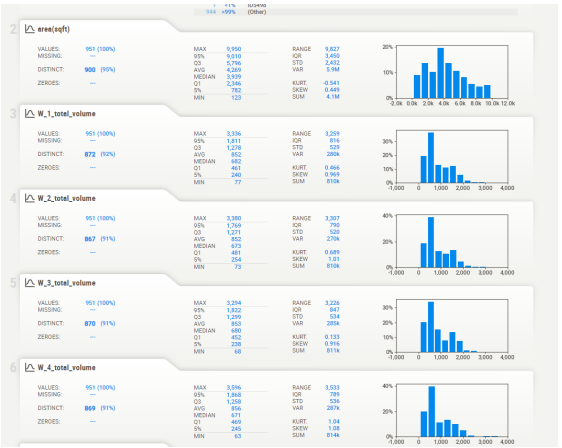
1.3. Data sources

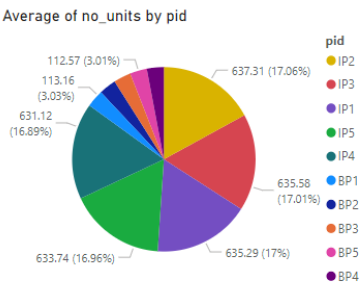
For training purposes, we are given five distinct datasets.

- A historical product sales data set consisting of 81000 records which are collected through 12 weeks (from 2nd January 2023 to 26th March 2023).
- An outlet dataset consisting 988 records with each outlet's area.
- A product data set consisting of the product details.
- A week dataset consisting of the week details.
- A freezer dataset consisting of details about different types of freezers.

Tools that were utilized to crack the case

- **SQL workbench** to make relations between the two datasets
- **Sweetviz** generates beautiful, high-density visualizations to kickstart EDA
- **Excel** for calculations
- **Power BI** to draw plots and analyze the datasets
- **Matplotlib** for creating static, animated, and interactive visualizations





Approach

1. Exploratory Data Analysis (EDA)

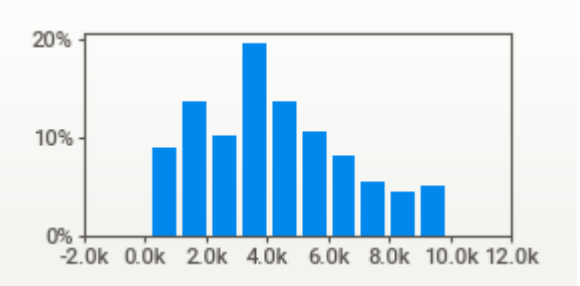
We used Power Bi, matplotlib and Sweetviz to generate some plots and calculated value counts manually and used Excel to get a clearer idea of the data available.

Datasets : We have been provided with five datasets;

1. Sales Data Set : A historical product sales data set consisting of 81000 records which are collected through 12 weeks (from 2nd January 2023 to 26th March 2023).

2. Outlet Data Set : The outlet data set consists of 988 records. Each record gives the shop space of each outlet. Among them 37 outlets have been duplicated. We took the 'last' recorded value with the assumption that the outlet underwent some modification during the time so the last recorded value gives the current shop space. So altogether XYZ company has unique 951 outlets.

   The figure illustrates the distribution of the area of outlets (in sqft).



3. Product Data Set : The XYZ company manufactures 10 unique ice-cream products.
4. Week Data Set : The sales data are recorded during a 12 week period. (from 2nd January 2023 to 26th March 2023)
5. Freezer Data Set : There are 10 unique types of freezers.

2. Dataset Preparation

1.1. Feature Engineering

We mapped information from other datasets into the outlet dataset using python codes. We came up with the following features for each store,

1. *shop_area* : already given
2. *total_units_sold of each item(10)* : There are 10 unique products and we calculated the total units sold of each of those products at each outlet.
3. *total_volume* : total volume of items sold during all the 12 weeks
4. *total_earning* : total earning within the 12 weeks

It gave altogether 13 eventual features for each store.

2.2 Checking for missing values & categorical variables

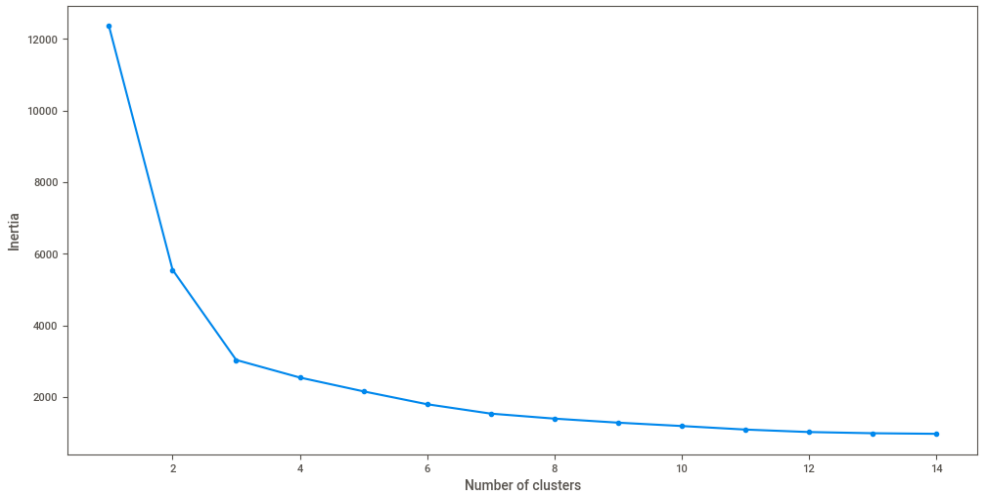
There were no missing values or categorical variables.

2.3 Standardization

We applied standard normalization to the selected features (mentioned in section 2.2) before sending through the segmentation model.

3. Segmentation Technique

We used the K-Means clustering model to segment data into non-overlapping sub-groups. To determine the number of clusters, we created a loop and ran the K-Means algorithm from 1 to 15 clusters and plotted the inertia at each instance. The graph is given below. The elbow (the point of inflection) of the curve was at the 3-cluster mark and we selected the number of clusters to be used as 3.

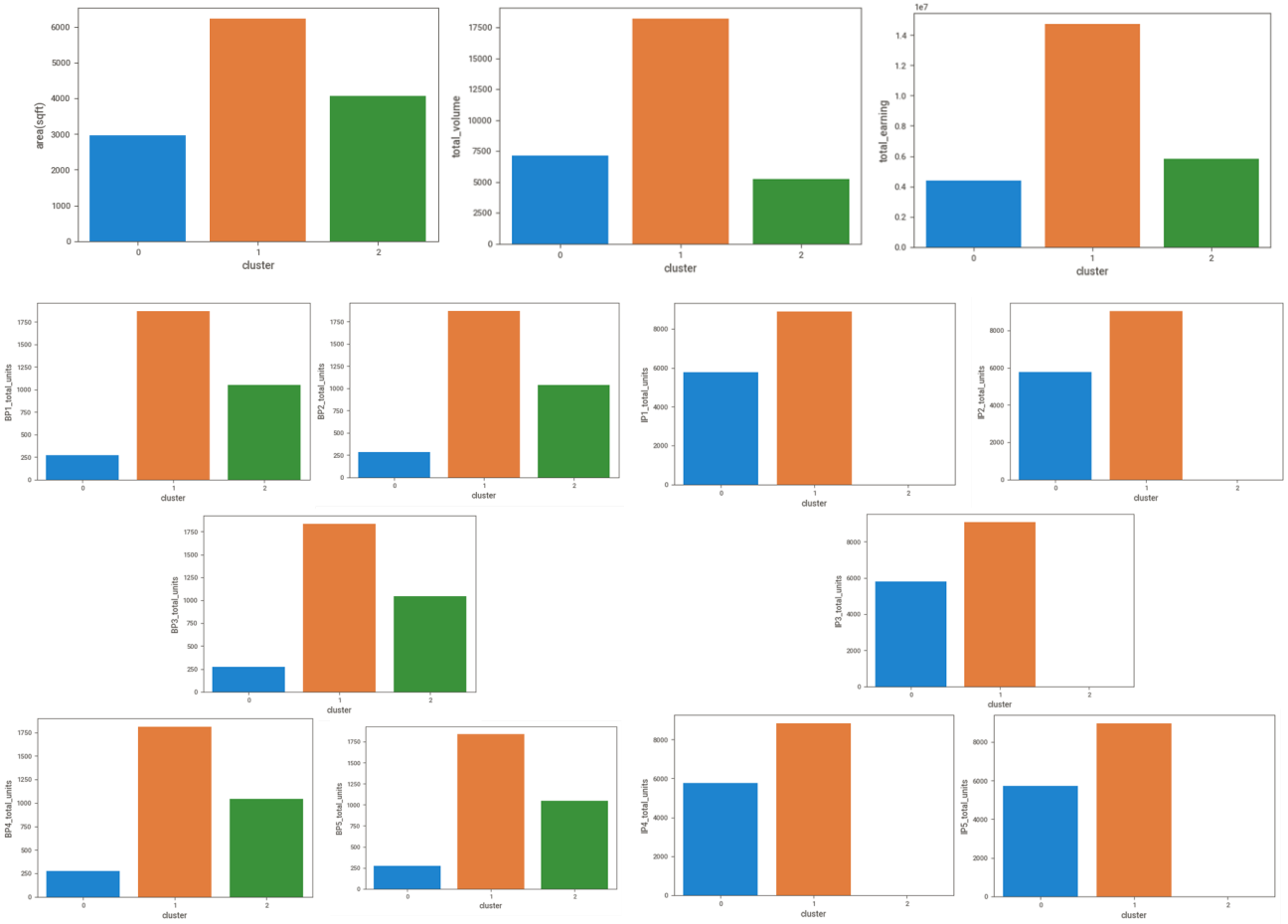


The clustering results obtained are as follows. The best value at each evaluation metric is colored in light orange.

	No of clusters				
Evaluation metric	**2**	**3**	**6**	**7**	**8**
Inertia	5554.658	3037.43	1801.33	1539.47	1399.65
Silhouette Coefficient	0.528	0.55	0.506	0.547	0.435
Davies-Bouldin Index	0.738	0.688	0.873	0.744	1.026
Calinski-Harabasz Index	1163.19	1455.28	1108.16	1106.16	1055.2

We can see that selecting 3 as the elbow gives relatively better results.

4. Characteristics of each segment



Based on the above graphs we can arrive at the following conclusions about each cluster.

Cluster	Characteristics
0	Have the lowest shop area and the lowest earning. Outlets in this cluster sell all the products but in a lesser quantity.
1	Have the highest shop area, highest volume of sold items and the highest earnings. Outlets in this cluster sell all the products in large quantities.
2	All the outlets that sell high-price and high-volume products (BP1, BP2, BP3, BP4, BP5) belong to this category. As a total they have the lowest volume of sold items but have a higher earning than cluster 0.

5. Allocation of freezers to each cluster

Both the ***ROI value*** and the **Sales Ratio** descends in the order,
M008,M004,M001,M006,M002,M009,M007,M003,M005,M010
But when considering a freezer it is important paying attention to the average volume an outlet would need during a day.

Cluster	Average Volume per day (L)	Freezer Type	ROI	Sales Ratio
0	**85** *(7142/(12*7))*	**M007**	3675.94	5.95
1	**216** *(18188/(12*7))*	**M010**	6136.108	7.578
2	**62** *(5224/(12*7))*	**M006**	7435.128	6.698

6. Conclusion and intervention strategies

Using the above segmentation we could effectively segment 951 outlets into 3 clusters and allocate freezers that most suits their requirement.

Some recommendations for the XYZ company to maximize their sales against the cost invested in freezers are as follows.

- Introducing new flavors.
- Price the ice cream competitively to attract customers while still making a profit. Consider offering discounts for bulk purchases or promotions to encourage sales.
- Maintenance of freezers to make sure that they are in good working condition.
- Regularly monitor sales data to identify popular flavors and adjust inventory levels accordingly.
- Keep the freezers well stocked in outlets that use more than one freezer.