Data Storm 4.0
Semi Final - Case Crack
Team Milkman

Team Members-DataStorm120

Minidu Thiranjaya
Themira Chathumina
Navinda Perera

Project Description

Problem Statement
Company XYZ sells beverages and ice cream items, increasing its reach to just under 1000 outlets. It gives different kinds of freezers with varying sizes and power usages to each outlet based on the evaluation done by the Area Distributor Managers (ADM) who look at outlet dimensions, space availability, sales, and location, However, as the reach grows, it wants to assign freezers to outlets in a way to boost sales while reducing freezer costs. So, the problem is to group the outlets based on similar features and suggest an appropriate freezer for each group, maximizing the sales and ROI of freezers.

Objectives
To perform a store segmentation to identify the stores with similar characteristics.
To evaluate the performance of the segmentation model against the metrics - Inertia, Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index.
To recommend a suitable freezer type for each identified outlet segment, maximizing the ROI and sales of freezers.

EDA and Feature Engineering

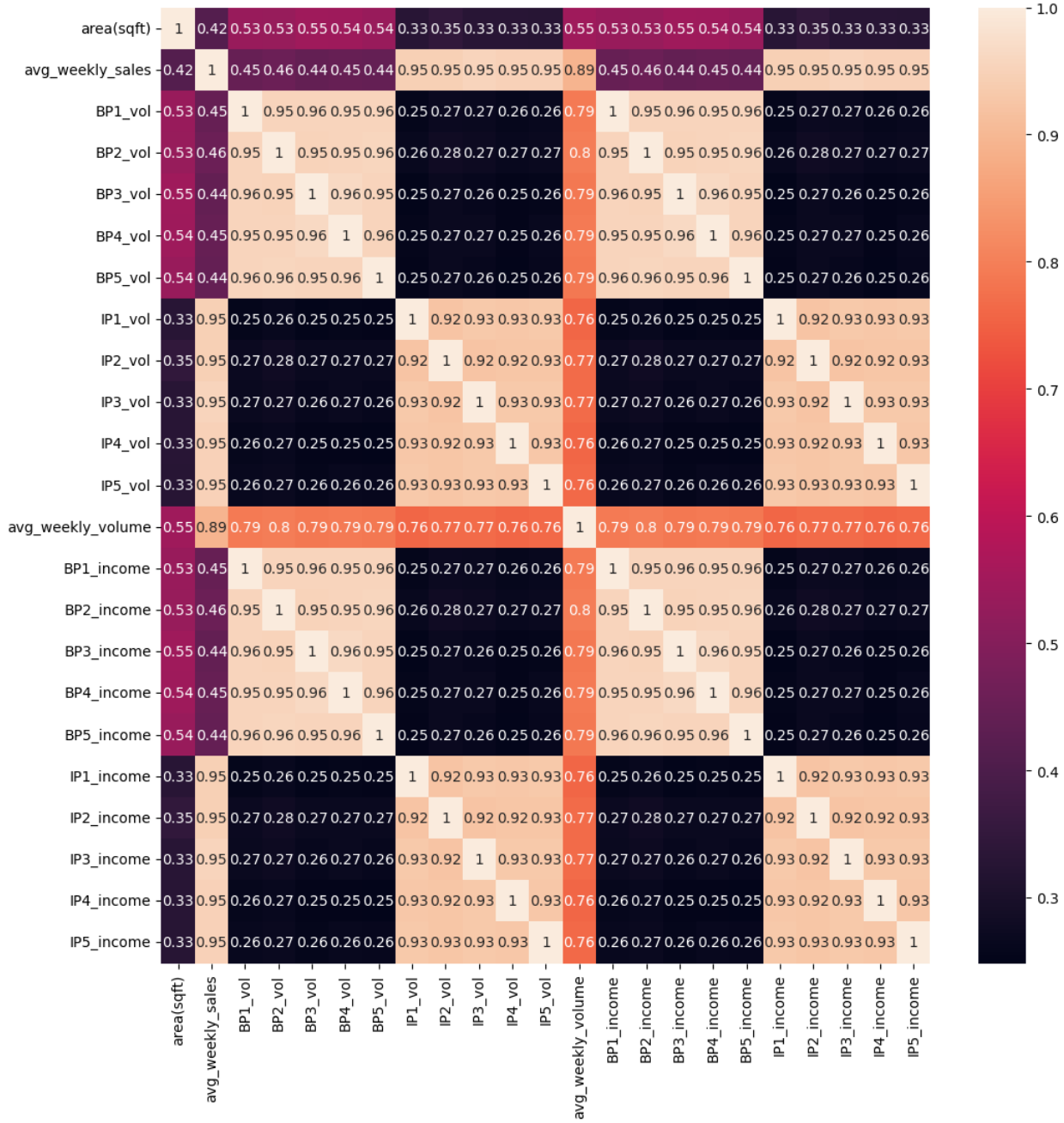
When we first analyzed the outlets_data table we found that even though there are 981 records there are only 951 unique outlets and some outlets have multiple areas recorded. To resolve that issue we took the mean of the areas and replace them with the area.

We decided not to scale the data since when scaled it would significantly reduce performance measured by the metrics given.

To run a model and cluster outlets together we need to create features to help us achieve that. To do that first, we created 20 separate features. They are constructed using the average

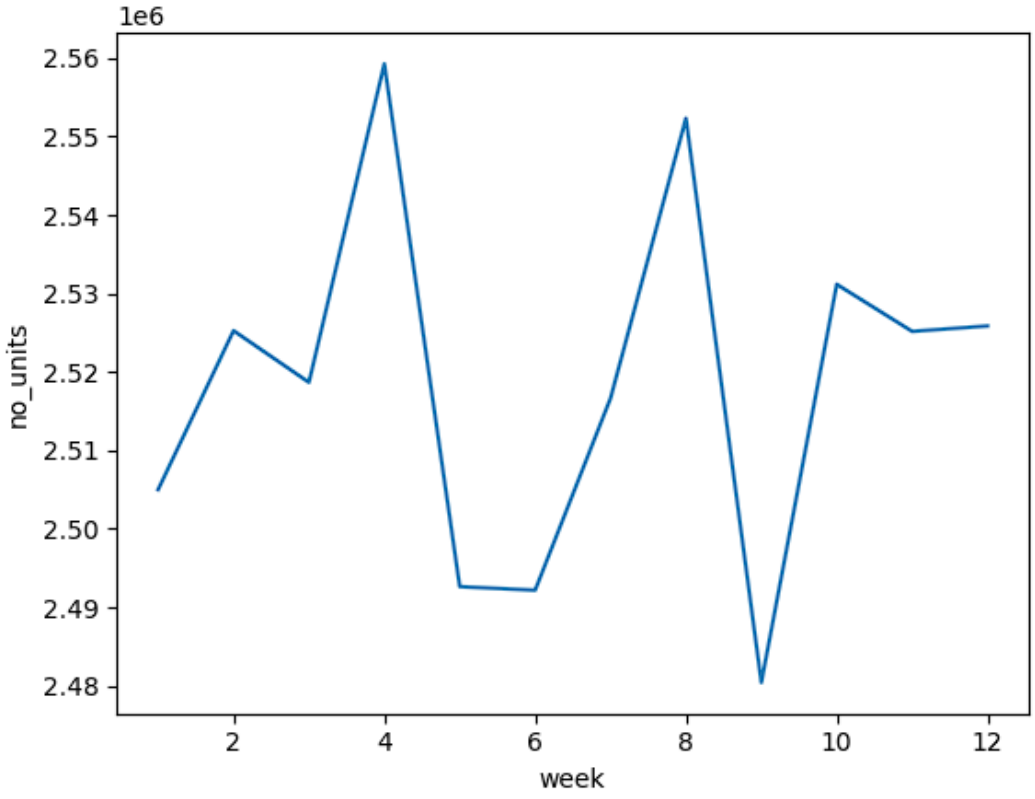
volume for each item taken in that outlet and the average income each item has generated for that outlet.

Then visualizing the correlation



We find that there is a high correlation between features with the same volume. For example, with items volume 0.2 features are highly correlated to each other. It's the same for items with 1. So we group with volume and create 4 features.

We then created the new features avg_weekly_sales, avg_weekly_volume, average_weekly_income, most_popular_item, and unique_products. Here most_popular_item means the item that has been sold most in the given time period and the unique_item is several unique items that are available in a given outlet.



Line plot of weekly sales: This plot shows the total number of units sold per week. This graph shows that the number of units sold has stayed consistent throughout the given timeline.

Clustering and Segmentation Technique

We tried out multiple algorithms to do the clustering and segment the outlets into different clusters. DBSCAN, K-Means, and OPTICS are the three algorithms we tried initially. Without tuning hyperparameters we ran the models and evaluated them based on inertia, Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index. (Note: inertia can't be used as a measurement in cases of density-based algorithms like DBSCAN and OPTICS). As we run our initial test the algorithms gave the following results

DBSCAN

After running for multiple values. For the eps = 100000 and min_samples = 50, we got the best results

Silhouette Score: 0.7627810374032912
Davies-Bouldin Index: 1.0927360503892698
Calinski-Harabasz Index: 1968.3791087890029

We got 6 clusters and outliers.

OPTICS

After initial guess and check methods, we observed that the OPTICS algorithm performs well. So we decided to tune the hyperparameters for the OPTICS algorithm. The hyperparameters we choose are min_samples, xi. After running three tests we found that 50 for min_samples and 0.05 for xi values give the optimal results. And after hyperparameter tuning our final model for segmentation gives the following results for the metrics mentioned above

Silhouette Score: 0.7933686855159273
Davies-Bouldin Index: 1.0995746842912488
Calinski-Harabasz Index: 1787.8276423304949

We got 6 clusters and outliers. The number of outliers is 11.

K-Means

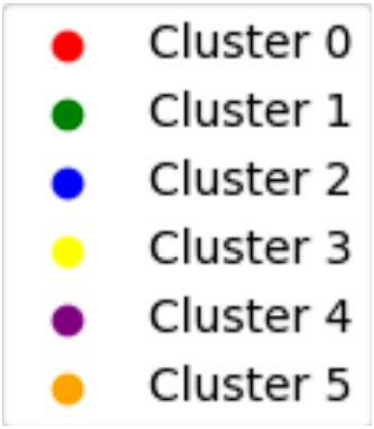
K-Means also performed well. We tested it on several k values and a different set of features. Finally, it gives the best performance when k = 6. Also, the distance between two different points seems to be large. So naturally, the value for inertia is very high.

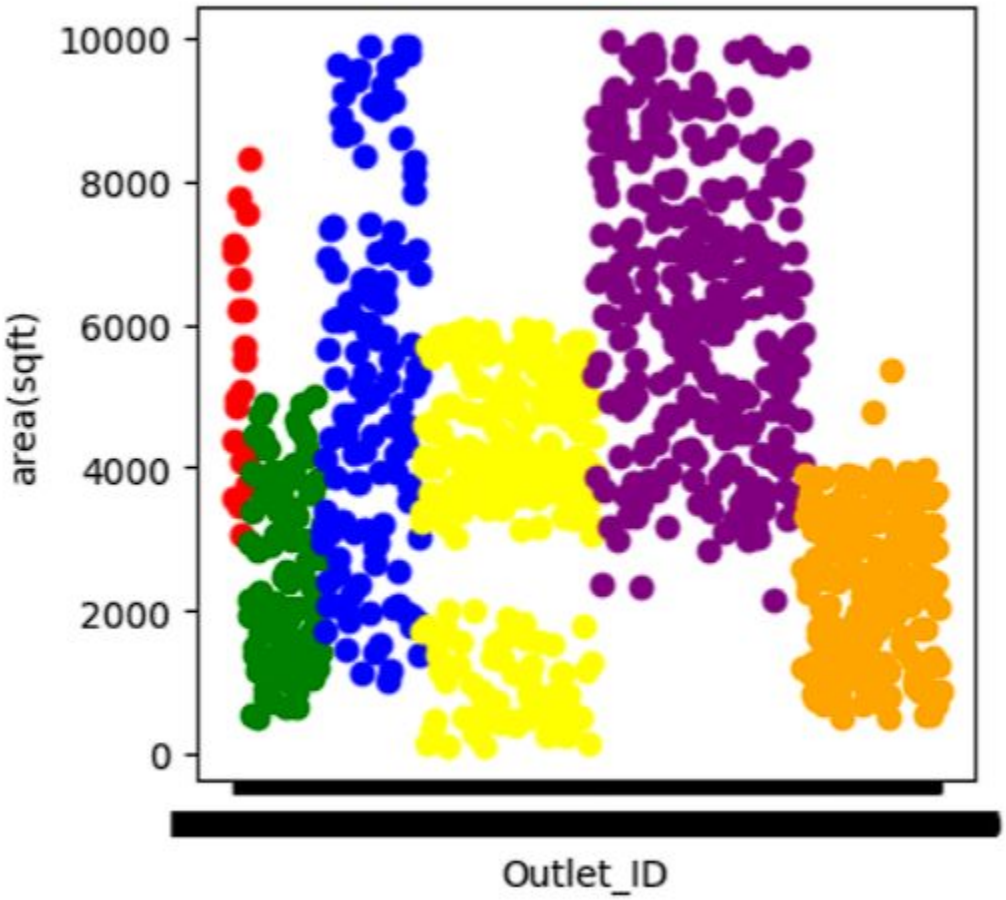
Inertia: 12652268300834.0
Silhouette 0.7801138884573174
Davies-Bouldin Index: 0a .40976698193830835
Calinski-Haraba,sz Index: 4358.241219531024

After observing, we found that we can't use Inertia to compare density based algorithms like DBSCAN and OPTICS. So we used the other three metrics to compare the three models we created. We concluded that K-Means with k = 6 seem to give the best overall performance to the given metrics.

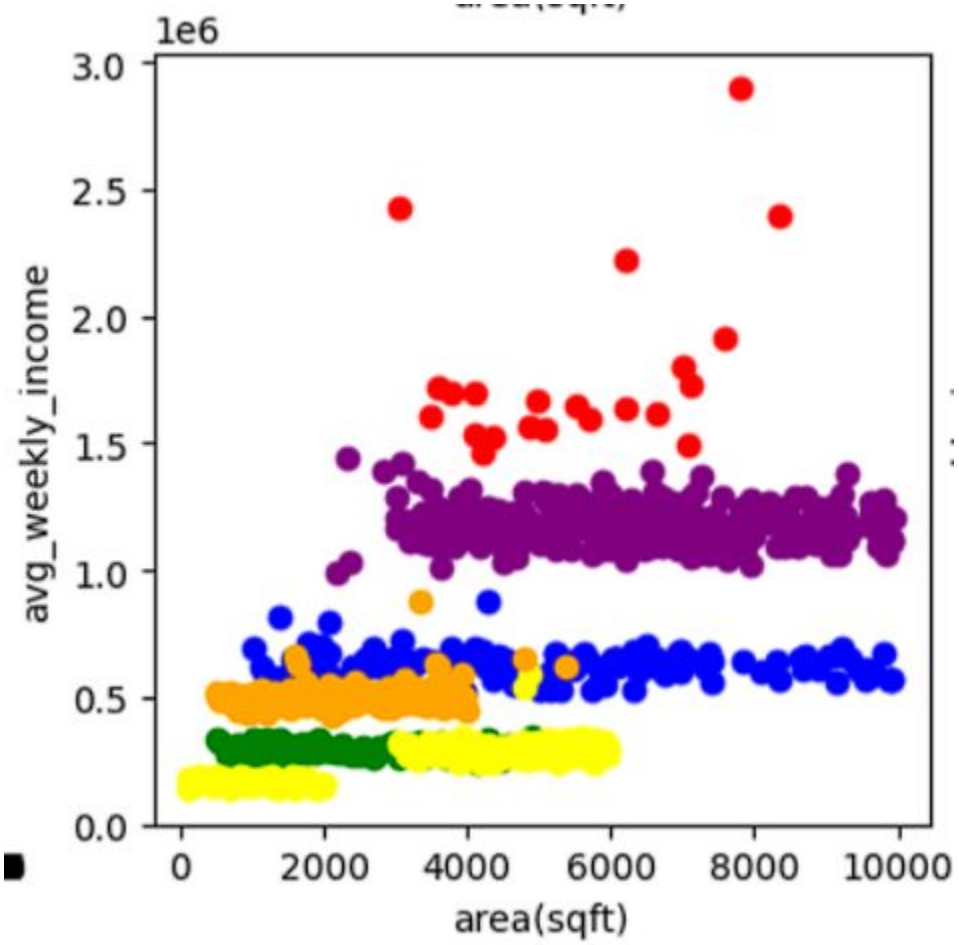
Characteristics of Segments

For simplicity let's name the the clusters by color.

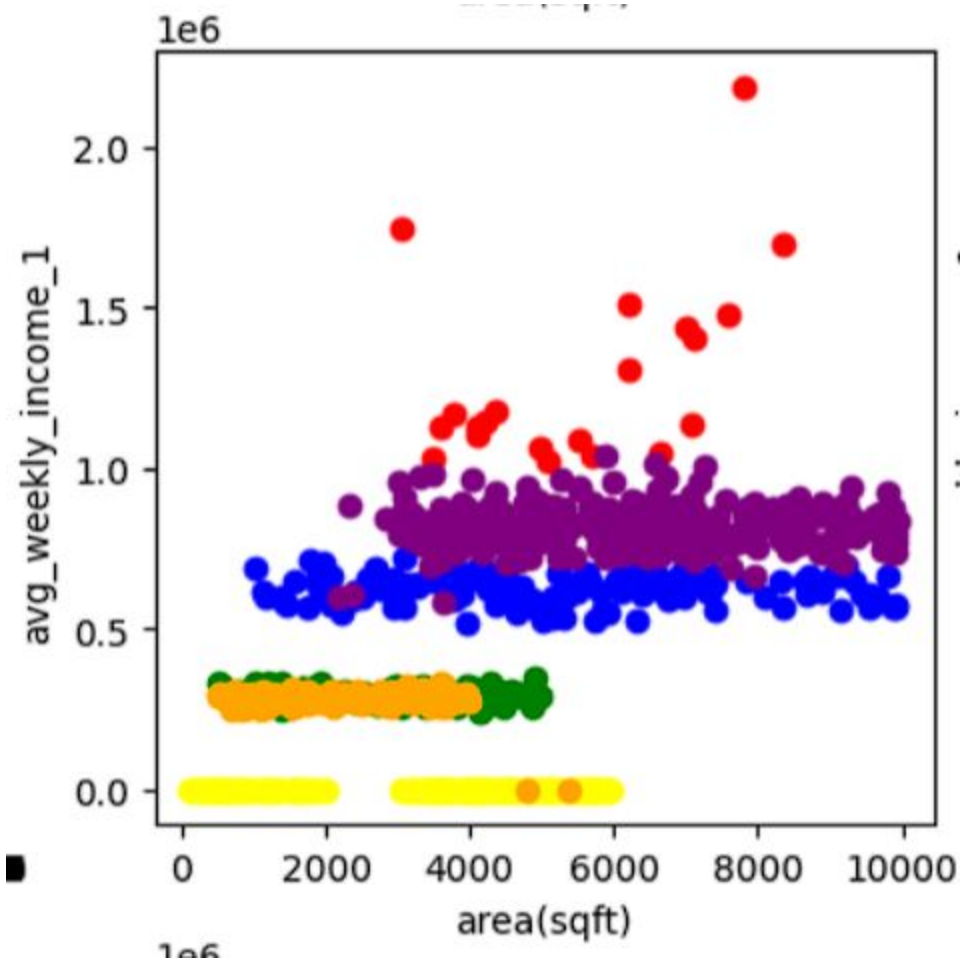




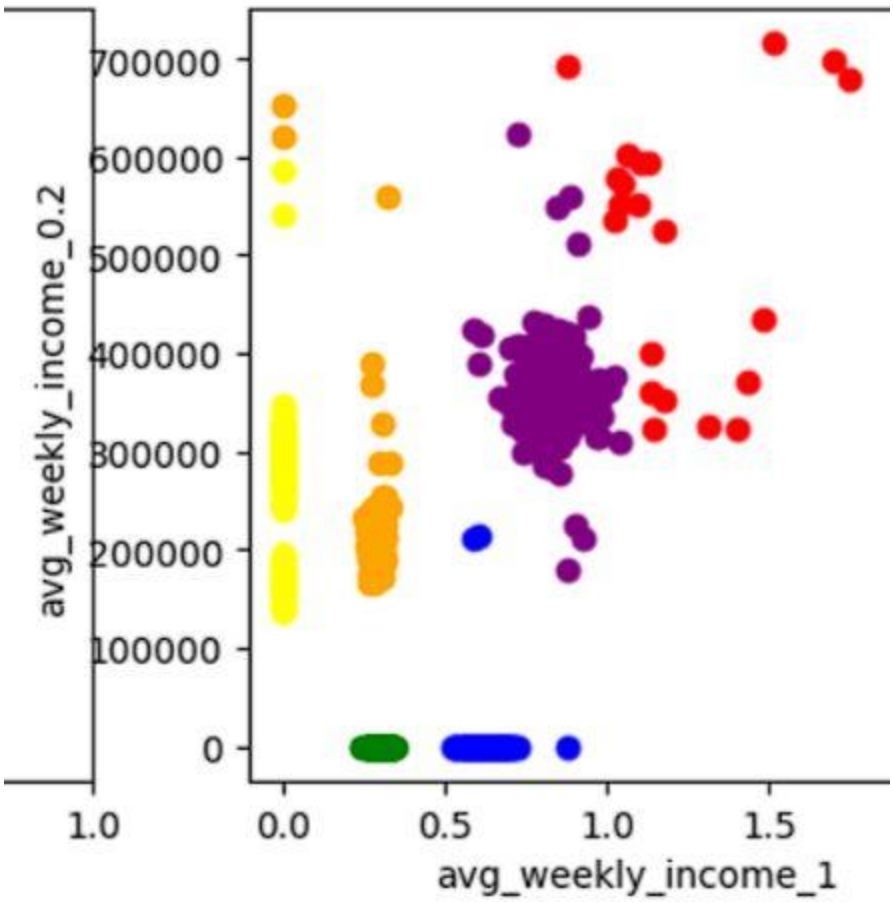
Observing this graph, we can have estimates of the ranges of clusters.
Blue clusters have the outlets with the widest range of areas. While green and orange clusters seem to have the narrowest range of areas.



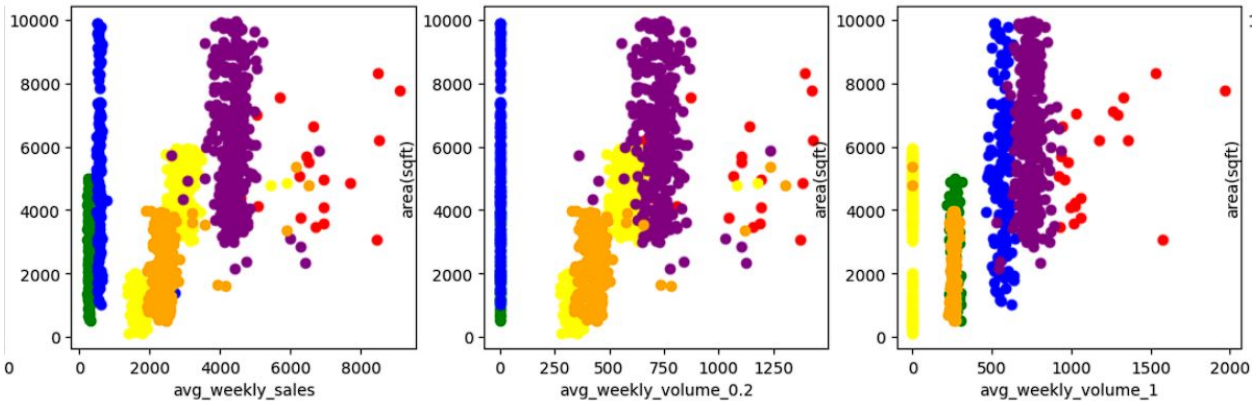
By area and avg_weekly_income we can differentiate blue, orange, purple, and red clusters. The red cluster has the highest average income. Then followed by purple, blue, and orange respectively. We can't differentiate between blue and green clusters here.



By observing this graph, we can differentiate between green and yellow clusters. Cluster yellow consists of outlets that do not sell items of volume 1L.



This graph shows that the outlets in the green cluster do not sell items that have a volume of 0.2L.



In all three graphs, we can see at least 3 clusters.

Analytical approach to allocating freezers to stores

We first calculated the average volume needed weekly for an outlet and calculated the number of freezers that are needed to store the needed amount. Then we calculate the Item sales ratio

by dividing the weekly needed volume by the volume given by the total number of freezers to store that weekly volume. Then we calculate the Return on Investment(ROI) by dividing weekly income by the number of freezers calculated above * maintenance & power consumption cost (since freezer cost is not given in the dataset, it was not included.)

We calculated the ROI and Item Sales Ratio for each cluster and each freezer. Then we multiplied both values for each cluster and each freezer. After that for each cluster, we selected the freezer
with the highest final value.
Here is the final freezer for each cluster.

Cluster: 0.0, Max Column: IceTech, Max Value: 21.429414292575665
Cluster: 1.0, Max Column: IceBlast Pro, Max Value: 22.451075416198705
Cluster: 2.0, Max Column: IceTech, Max Value: 26.29492090327781
Cluster: 3.0, Max Column: IceTech, Max Value: 9.708001995220052
Cluster: 4.0, Max Column: IceBlast Pro, Max Value: 19.64464127973494
Cluster: 5.0, Max Column: IceBlast Pro, Max Value: 15.975459905777722

Since the freezer cost was not given the correctness may be slightly different. Above freezers are optimized for weekly volume and weekly cost.

The Conclusion

Since the cost of freezers has not been given the effectiveness of modeling the freezers to clusters can be low.
But outlets with higher sales and volumes always use high-capacity freezers like IceTech while other stores use more balanced freezers like IceBlast. These values were optimized for total weekly income and volume sold.