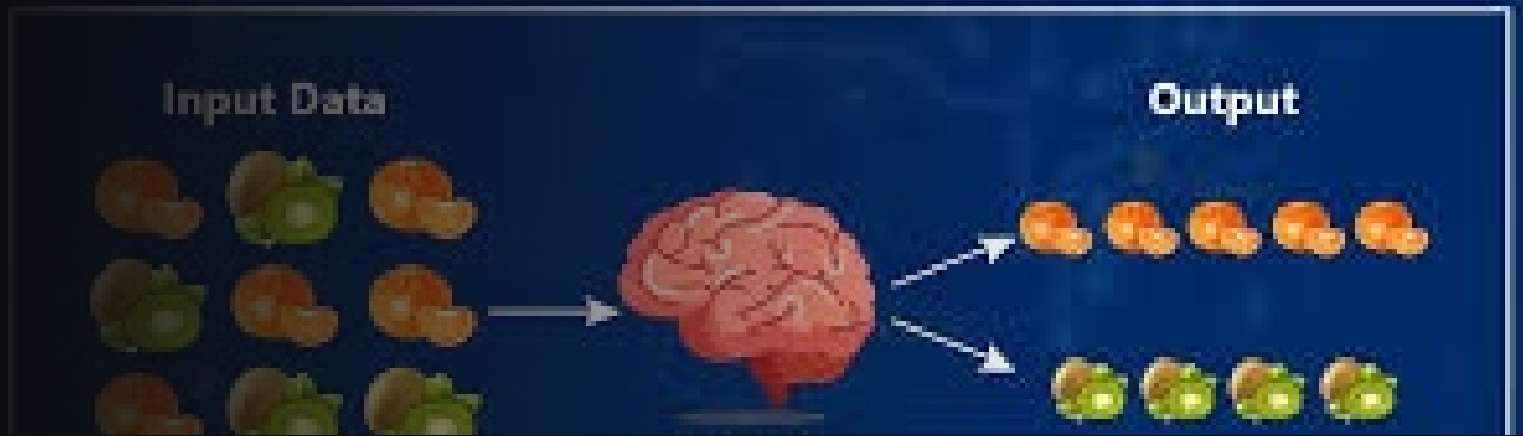
Unsupervised Learning Project



By:
Rana Al-suffi

For this project:

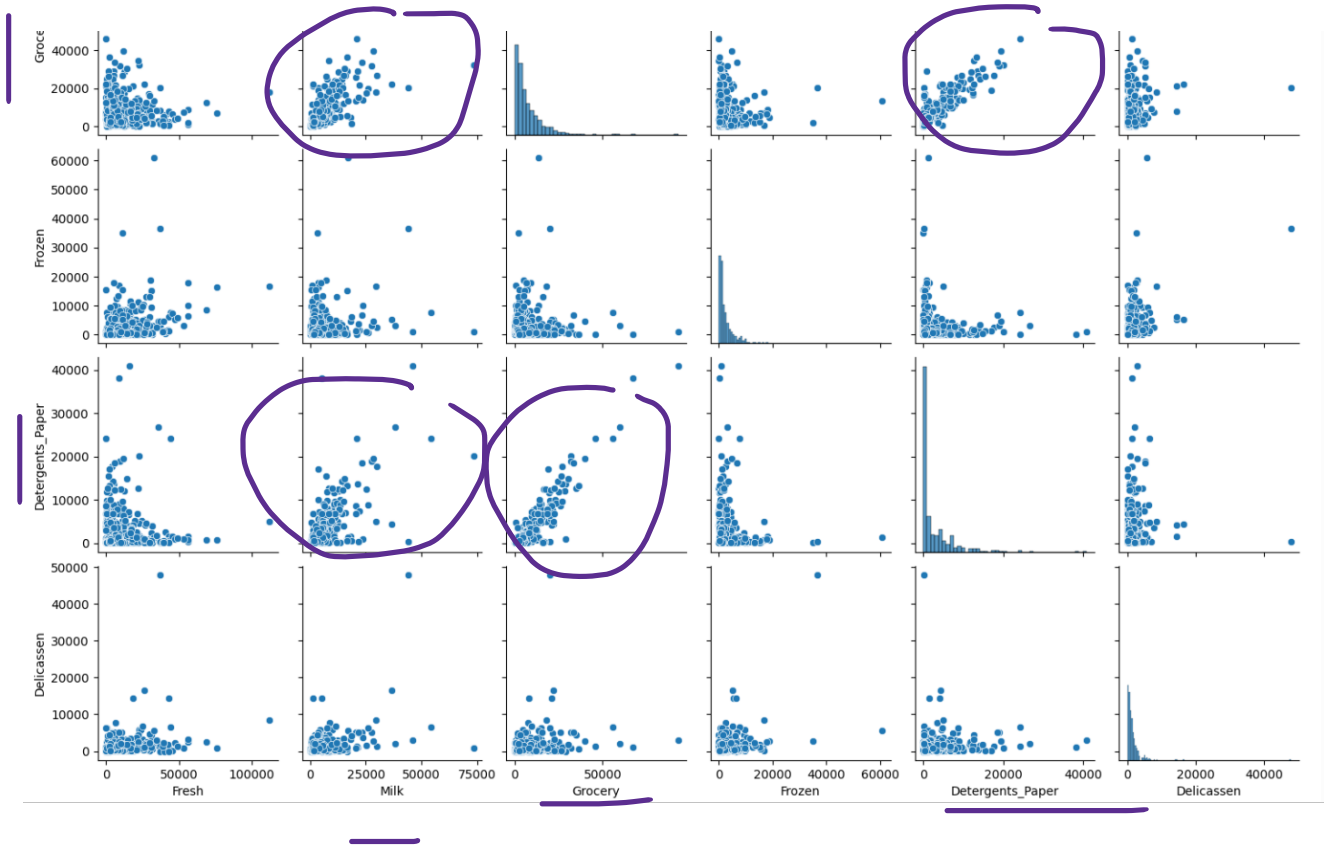
The dataset that I used was wholesale_data.csv.

I used Jupyter notebook

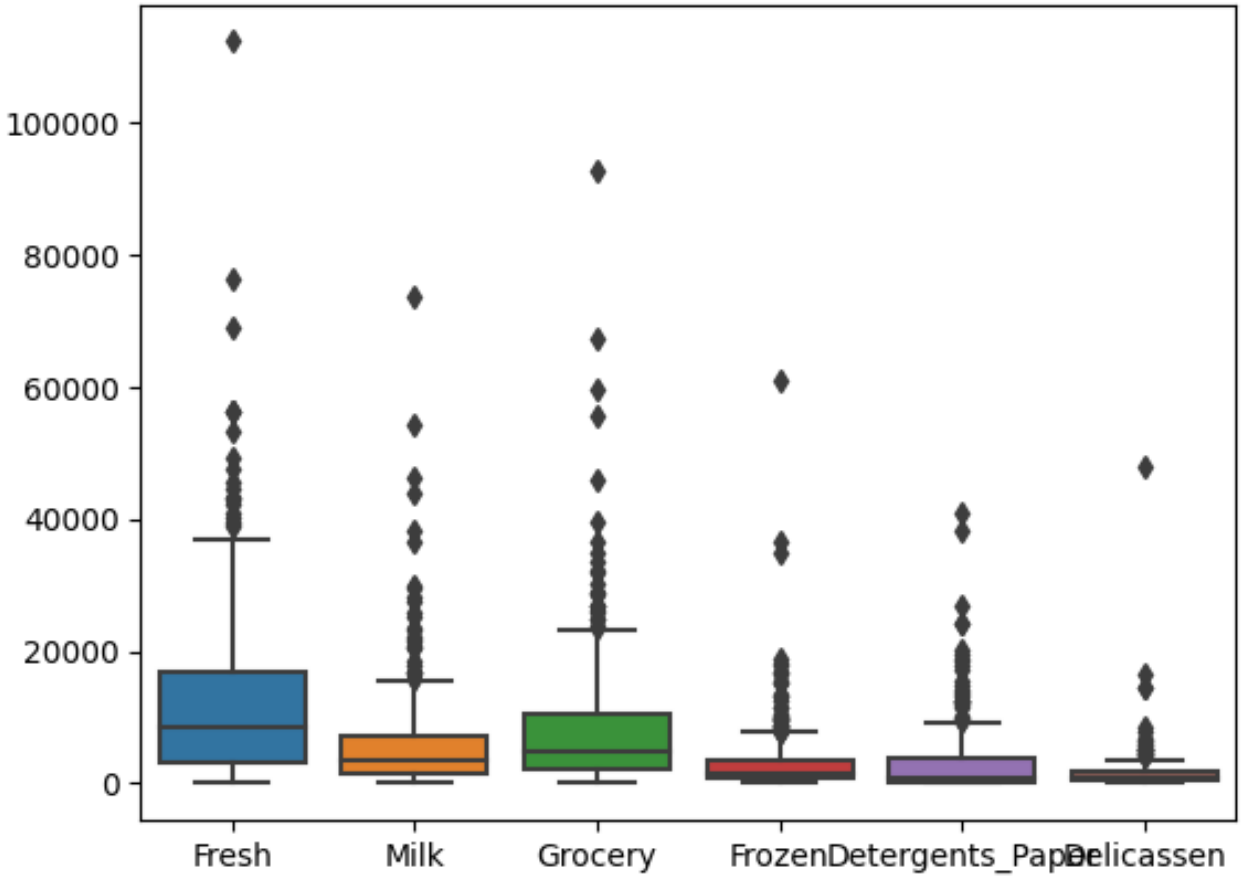
Libraries include pandas, NumPy, matplotlib, scipy, sklearn, and seaborn

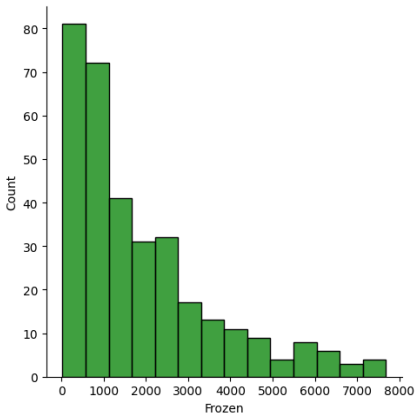
EDA - Exploratory Data Analysis & Pre-processing

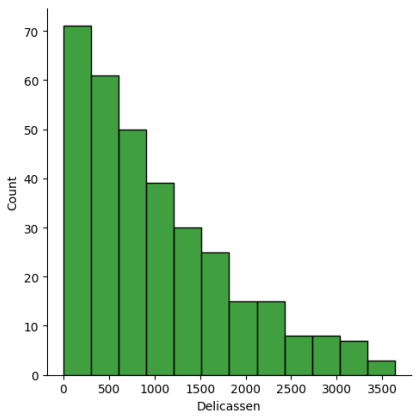
- It looks as if all columns has a positive correlation with each other.
- However, the relationship between Detergents, Milk and grocery are higher correlation then other columns.
- Customers intend to purchase these products more frequently than others.

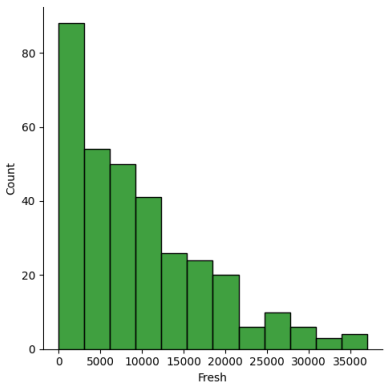


- From the data, it is obvious that there are outliers and in multiple columns.
- I used boxplot to take another look.
- And it looks like there are too many outliers.

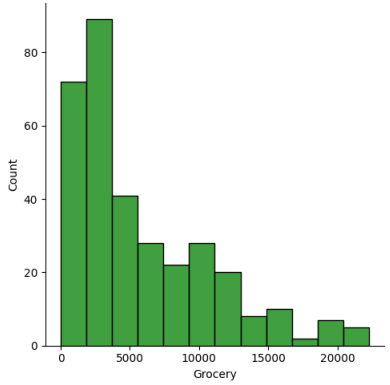


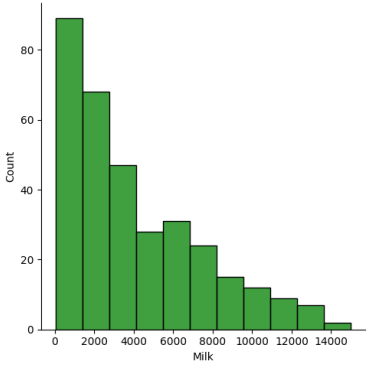


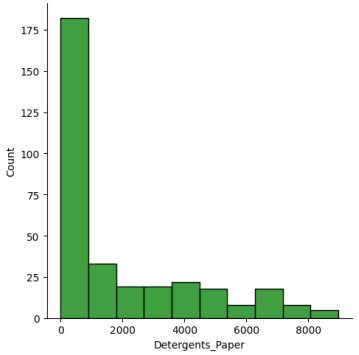




- I used IQR to remove the outliers from the dataset
- I checked the data after removing the outliers and it looked better.







Processing the data:

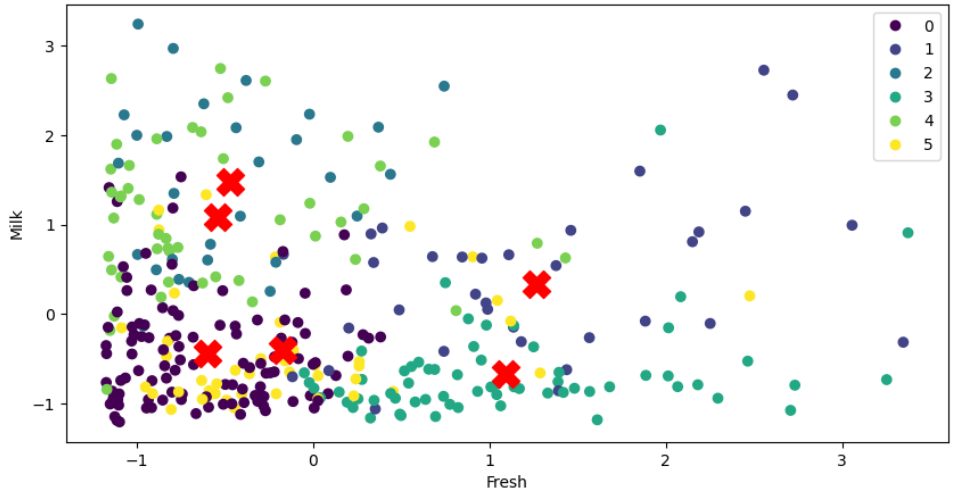
KMeans Clustering:

- Using KMeans, I started with three clustering groups however using the inertia score showed that it is best to make the clustering into 6 groups.

- To confirm the findings, I used the silhouette score which gave the same results.

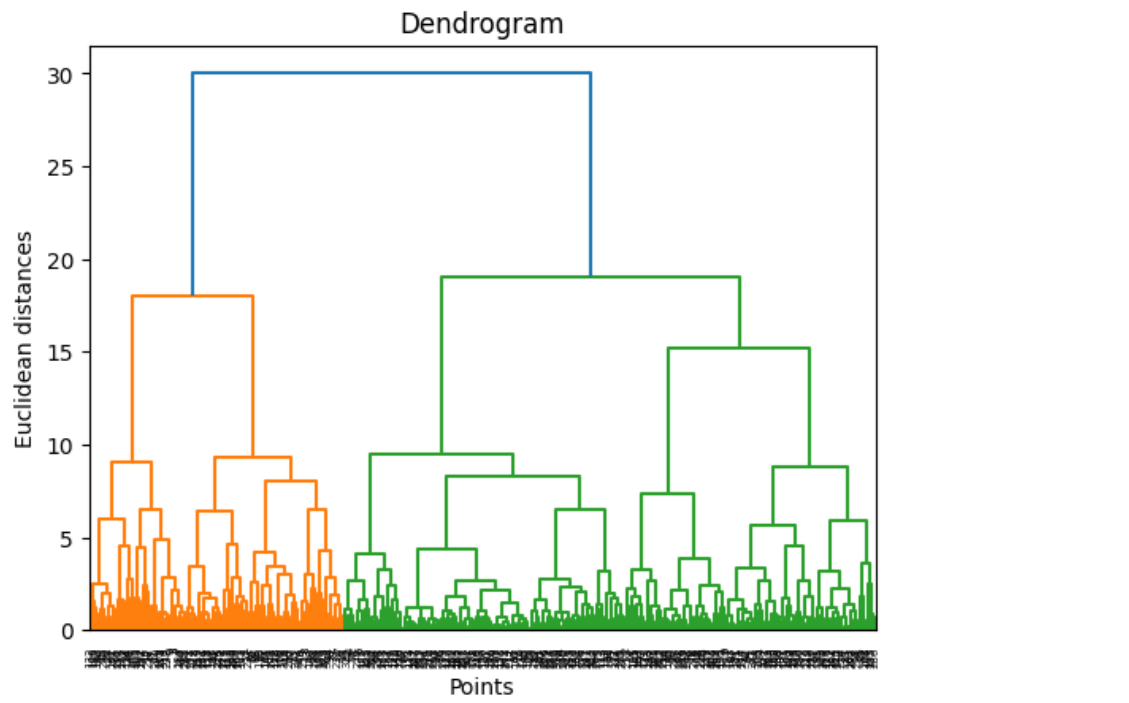
- From the result, it means that all the shoppers can be divided into six groups and that depends on the type of items they purchase for each shopping trip.

```
  warnings.warn(
4]: <matplotlib.legend.Legend at 0x24ab29b27d0>
```



Hierarchical Clustering:

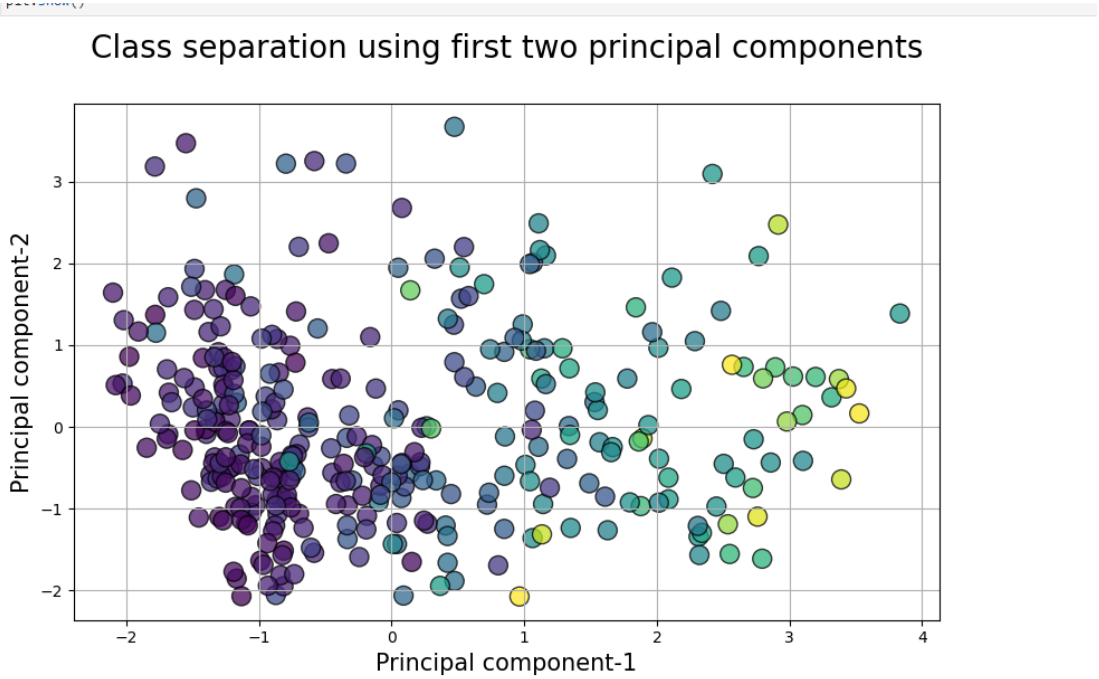
- After multiple tries, using different numbers of clusters, Dendrogram showed the best number of clusters is three



PCA:

- I used five products, which are 'Fresh', 'Milk', 'Frozen', 'Detergents_Paper', and 'Delicassen. From processing the data, it is apparent that the first two components combined make 64% of the variance ratio.

- Using the previous information, we used only the first two components in our training.



- From the data, we can see that there are certain products are sold together.
- The company can use this data to leverage their sales and see which product should make promotions on for better margins

Thank you