Spending, a type of behavioral data, is a common appropriate way to identify customer segments (Sagar, 2019). An examination of the customers dataset via clustering methods enables the food company to group customers into segments based on their annual spending on the six product categories, then plan resource allocations targeting each segment.

The purpose of this study is to help the company better understand its consumers' behaviorby identifying common behavior in spending within the customer segment. K-means will be used because of its emphasis on homogeneity within a group rather than the separation between groups (Hennig, 2015). It is simple while proving to outperform hierarchical clustering in the existence of noisy and outliers (Punj & Stewart, 1983).

Data Screening & Transformation

The distributions of all 6 features are extremely right-skewed (see Appendix A), which the noise and redundancy object will negatively affect the clustering method's performance (Virmani et al., 2015). Log transformation is performed on the whole dataset to achieve improved normality of variables distribution. Besides, Berget (2018) recommends to define the evaluation for each cluster size depending on the dataset size (e.g. 100-200 observations need 20-50 cluster size). Thus, each cluster size should be at least over 20 for the `customers` dataset.

Samples found to be univariate outliers in multiple variables do not necessarily mean they are actually multivariate outliers. Therefore, the multivariate outlier detection method based on a robust method is used (function `mvoutlier.CoDa` in `mvoutlier` package), recommended by Filzmoser et al. (2005). A total of 45 outliers are found, which occupies 10.23% of the total observation. Removing all 45 outliers would possibly cause excessive loss of information and generality for this research; hence they were all kept in the study.

Comparison of Unscaled Data & Scaled Data

After transforming data, since the range of mean and variance of the features is not vastly different (see Table 1), clustering is performed on both scaled and unscaled data to discover the existence of any dominant variable. Euclidean distances are commonly used for clustering consumer segmentation (Berget, 2018). The dominance of variables is considered in terms of scaling data in the next part. Therefore, Euclidean, which can greatly impact the difference in scale among dimensions, is used to compare the results between scaled and unscaled data.

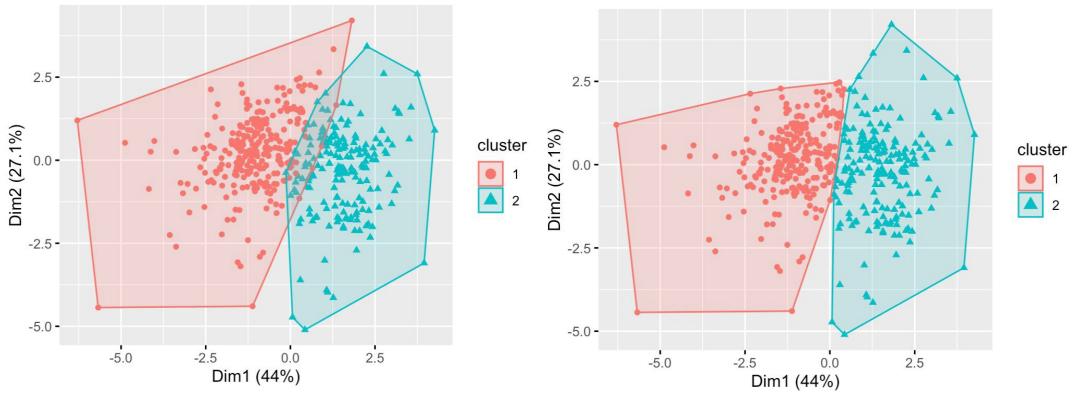
Two plots of within the sum of squares, silhouette and prediction strength are performed to choose the best K between the 1 to 10 range, where $K = 2$ is found to be the best value from all tests in both datasets (see Appendix D Figure 3). After meeting the minimum cluster size, cluster 1 and cluster 2 of unscaled data contains 261 and 179 observations respectively, while in scaled data there are 252 and 188.

Table 1 - Summary of mean and variance for each product category

	Fresh	Milk	Groceries	Frozen	DP	Deli
mean	8.7305	8.1210	8.4412	7.3014	6.7860	6.6651
variance	2.1906	1.1694	1.2458	1.6500	2.9619	1.7183

To compare clustering in both datasets, 96.14% of the number of observations are classified in the same cluster (only 17 observations are unmatched), which shows no significant difference in K-means performance on these datasets. Moreover, the Jaccard coefficient for clusters in both datasets exceeds 0.95, which indicates the high stability of clusters. However, from the clustering visualization (Figure 1), presented by the first two dimensions, clusters from scaled data are more visually separatable. Therefore, scaled data is considered as having better results and is chosen to perform customer analysis.

Figure 1 - Uscaled data (left) and Scaled data (right)



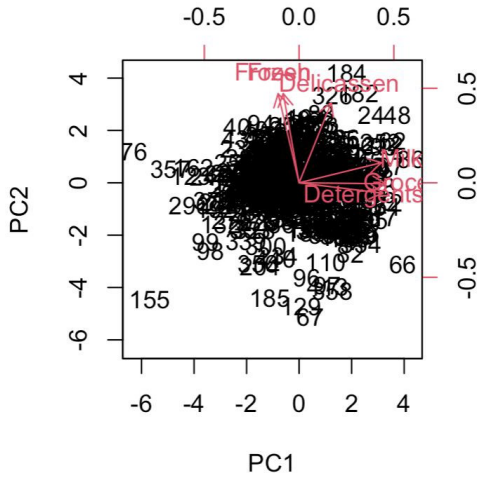
Customer Segment Analysis

Table 2 - Average total annual customer spending based on groups

Cluster	Fresh	Milk	Groceries	Frozen	DP	Deli	Total spending
1	13973	2402	2919	3706	492	1038	24529
2	9356	10346	14697	2222	6085	2177	44884

On average, cluster 2 customers' annual spending is 1.83 times higher than that of cluster 1 customers. Compared to cluster 1, cluster 2 spends over 4 times more on `Milk`, 5 times on `Groceries` and 12 times more on `DP` (69.35% of total spending). Meanwhile, cluster 1 spends much more on `Fresh` and `Frozen` food products (72.07% of total spending). Both clusters have similar spending on `Deli` products.

Figure 2 - Biplot, PC1 vs PC2



PCA is performed to further explain the clusters found. PC1 explains 43.99% of the total variance is dominated by `Milk`, `Groceries`, and `DP` (Figure 2). The narrow spread of the direction vectors for these three variables strongly indicates a high correlation among them. The correlation matrix (Table 3) also shows a high correlation between `Groceries` and `DP` (79.64%), `Milk` and `Groceries` (75.89%), and a moderate correlation between `Milk` and `DP` (67.79%). Thus, customers who spend a considerable amount on one of these three product types are predicted to also spend more on the other two. PC2 explains 27.13% oftotal variance which mainly presents the `Fresh`, `Frozen` and `Deli` variables.

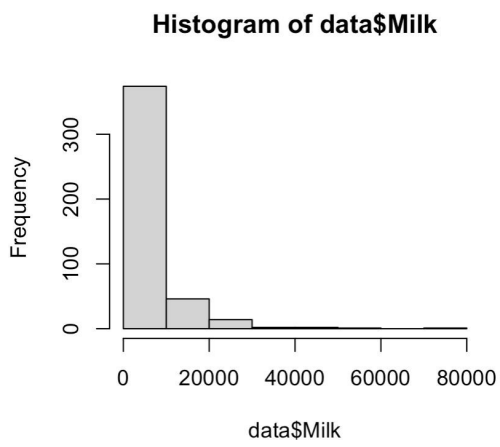
Table 3 - Correlation matrix of `customers` dataset

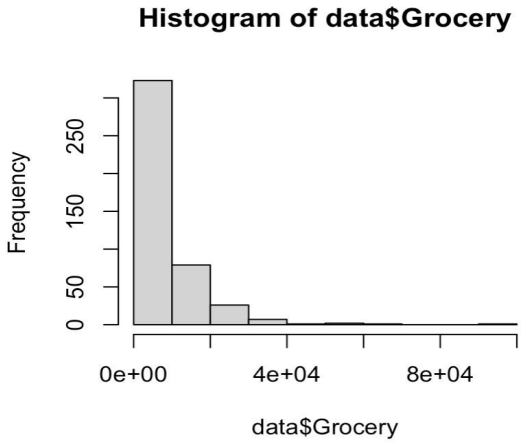
	Fresh	Milk	Groceries	Frozen	DP	Deli
Fresh	1.0000	-0.0198	-0.1327	0.3840	-0.1559	0.2252
Milk	-0.0198	1.0000	0.7589	-0.0553	0.6779	0.3378
Groceries	-0.1327	0.7589	1.0000	-0.1645	0.7964	0.2357
Frozen	0.3840	-0.0553	-0.1645	1.0000	-0.2116	0.2547
DP	-0.1559	0.6779	0.7964	-0.2116	1.0000	0.1667
Deli	0.2552	0.3378	0.2357	0.2547	0.1667	1.0000

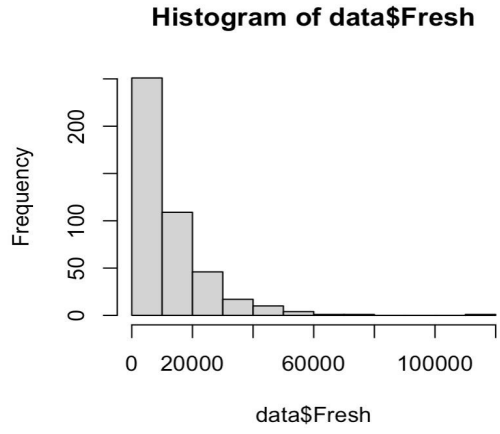
In conclusion, based on the purpose of the study and the characteristics of the dataset, the food company can divide the customer segment into 2 main groups. However, it is noted that the outliers detected from individual features are not removed which may impact the algorithms' performance. Besides, more customer information and further clustering methods are also recommended to analyze customer behavior better and validate clusters.

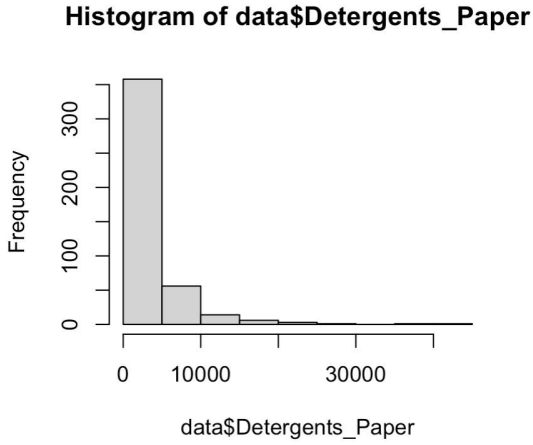
Appendix A

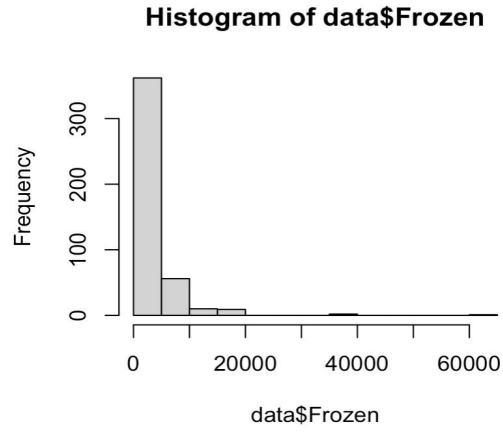
Histograms - Six Product Categories

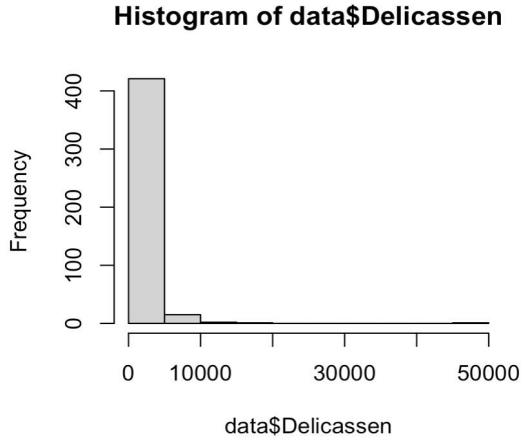












References

- Berget (2018). *Statistical Approaches to Consumer Segmentation*. Methods in Consumer Research, 1, 353-382 [Online]. Available at:https://doi.org/10.1016/B978-0-08-102089-0.00014-5 [Accessed 11 March 2021].

- Filzmoser, P., Garrett, R. G. and Reimann, C. (2005). *Multivariate outlier detection inexploration geochemistry*. Computers & Geosciences, 31(5). 579-587 [Online]. Available at: https://doi.org/10.1016/j.cageo.2004.11.013 [Accessed 16 March 2021].

- Hennig, C. (2015). Clustering strategy and method selection. arXiv [Online]. Available at:https://arxiv.org/abs/1503.02059 [Accessed 16 March 2021].

- Punj, G. and Stewart, D. W. (1983). *Cluster analysis in marketing research: Review and suggestions for application*. Journal of Marketing Research, 20(2), 134-148 [Online].Available at: https://doi.org/10.2307/3151680 [Accessed 12 March 2021].

- Sagar, A. (2019). *Customer Segmentation Using K Means Clustering*. Towards Data Science[Online]. Available at:
- https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3 [Accessed 10 March 2021].

- Virmani, D., Taneja, S. and Malhotra, G. (2015). *Normalization based K means Clustering Algorithm*. arXiv [Online]. Available at: https://arxiv.org/abs/1503.00900 [Accessed 13 March 2021].