Literature review

Decades of research in the grocery retail industry have been dedicated to optimizing market efficacy through customer segmentation. This approach aims to identify and target loyal and potential consumers across various sales channels. By employing customer segmentation, businesses not only gain the ability to identify profitable and loyal customer groups for targeted marketing efforts but also develop a deeper understanding of customer behavior and preferences. This understanding enables them to provide personalized services and implement tailored marketing strategies accordingly (Peker, Kocyigit, & Eren, 2017).

Previous research has employed customer segmentation to effectively group customers according to both general demographics variables, and product-specific variables, such as purchasing behaviors and intentions (Wedel and Kamakura, 2000).

In recent years, it has become evident that relying solely on general variables for customer segmentation may yield unreliable results in capturing purchase behaviors, as customers with similar demographics and lifestyles may exhibit varying patterns. As a result, integrating product-specific variables has gained importance in identifying customer groups with more homogeneous responses to marketing programs (Tsai and Chiu, 2004).

However, incorporating product-specific variables is more challenging, resulting in inconclusive findings and limitations in implementation. To address this gap, our study aims to utilize advanced clustering techniques to analyze essential product-specific variables such as customer consumption frequency, timing of purchases, and specific types of products purchased. This approach seeks to develop a robust and effective marketing segmentation strategy.

By refining our understanding of customer segmentation and purchase behaviors through the comprehensive examination and integration of these variables, we aim to provide valuable insights for marketing practitioners and decision-makers in the grocery retail industry. Ultimately, this research will contribute to enhancing marketing intelligence, leading to improved business performance and customer satisfaction.

Research Questions

Based on our literature review, our research questions are: 1) Does the frequency of purchasing specific types of products influence customer buying habits and decisions at Hunter's Supermarket?
2) Does the timing of purchases (hours of the day and days of the week) impact the types of products consumers choose to buy at Hunter's Supermarket?

We hypothesize that the frequency of purchasing specific types of products will significantly influence customer buying habits and decisions, the timing of purchases will significantly impact the types of products consumers choose to buy.

The anticipated effects of these investigations are multifold. If our hypotheses hold, this study will emphasize the importance of understanding customer buying patterns in terms of product type, purchasing frequency, and timing. Such insights could inform Hunter's supermarket's strategies for product placement, promotions, and stocking decisions, optimize operations and marketing efforts based on peak times, and guide in cultivating customer loyalty and enhancing targeted marketing strategies. Thus, the study's findings could ultimately contribute to improved business performance and customer satisfaction.

Data Description and Preliminary Analysis Results

We have used the Supermarket dataset for predictive marketing 2023 provided by Kaggle (Hunter, 2017). This dataset was created to predict market trends and contains consumer behavior data from supermarkets. There are 2,019,501 observations of 12 variables in this dataset. Among them, there are ten variables that are numerical, while the remaining two are character variables. The two-character variables are used to distinguish the names of different products sold in each order and the department they come from. The remaining ten numerical variables represent the order number, user number, number of the order, day of the week the order was made, time of the order, history of the order, ID of the product, number of items added to the cart, if the reorder took place, and unique number allocated to each department. We can easily observe that both the "ID of the product" and "unique number allocated to each department" can be used to differentiate between different products and departments. Therefore, in the subsequent analysis, we can organize a mapping table for the IDs and character variables. During the analysis process, we can use only the IDs to represent these products and departments, and then retrieve the corresponding products in the final report.

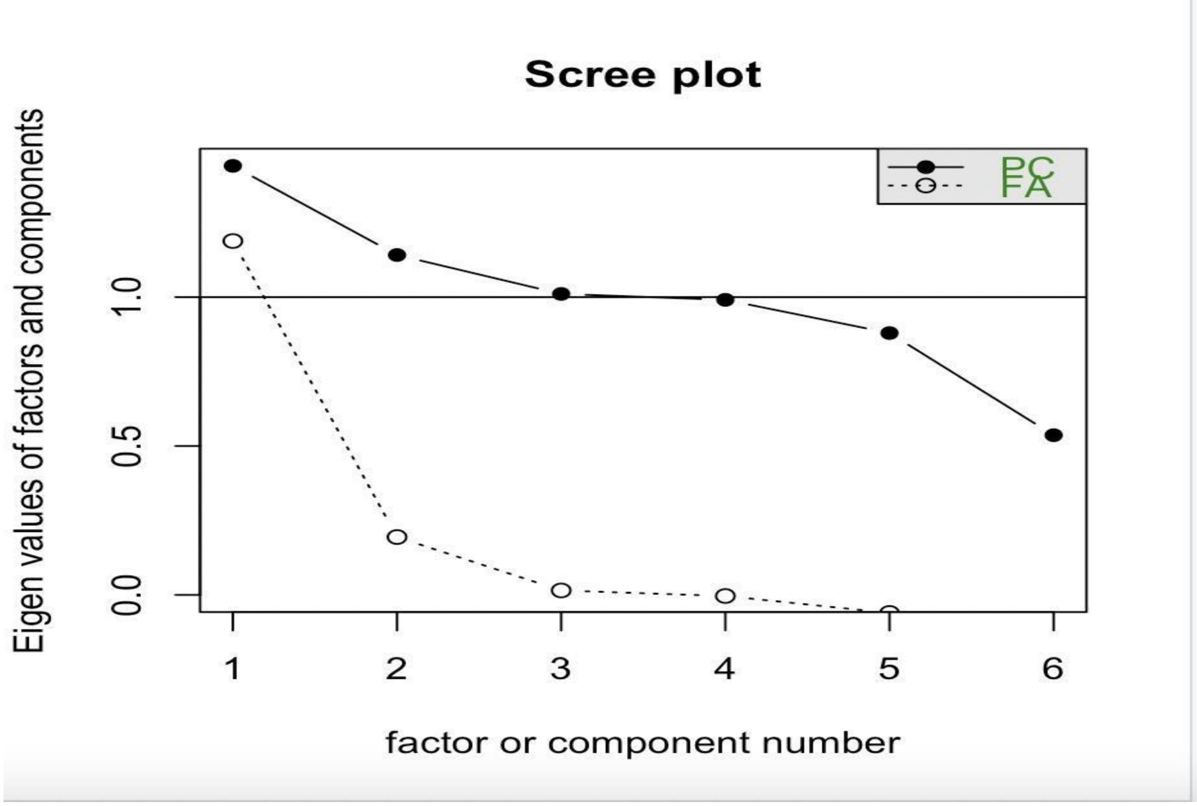
We first conducted correlation visualization. Through visualizing the graphs, we discovered that there is no strong correlation between variables, indicating that no multicollinearity exists in this dataset.

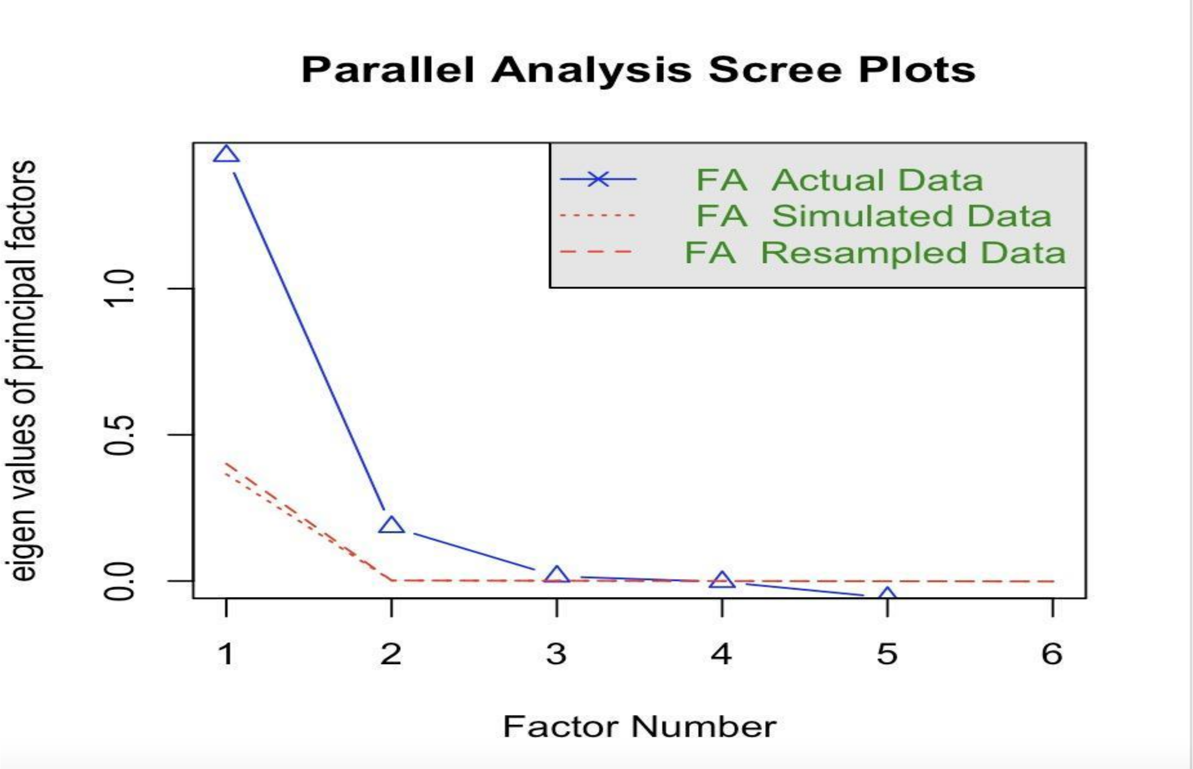
Then we conducted an investigation into anomalies within this dataset. We performed a visualization of quantiles for these variables to identify outliers. Based on our analysis, outliers were found in "order_number" and "add_to_cart_order". This indicates that some customers have very high order volumes, and some orders have a large number of items. In our final analysis, we have addressed and analyzed these outliers separately.

Dimension reduction

To efficiently analyze the data and reduce the dimensionality of variables, we initially opted for factor analysis as our primary tool, but encountered several challenges.

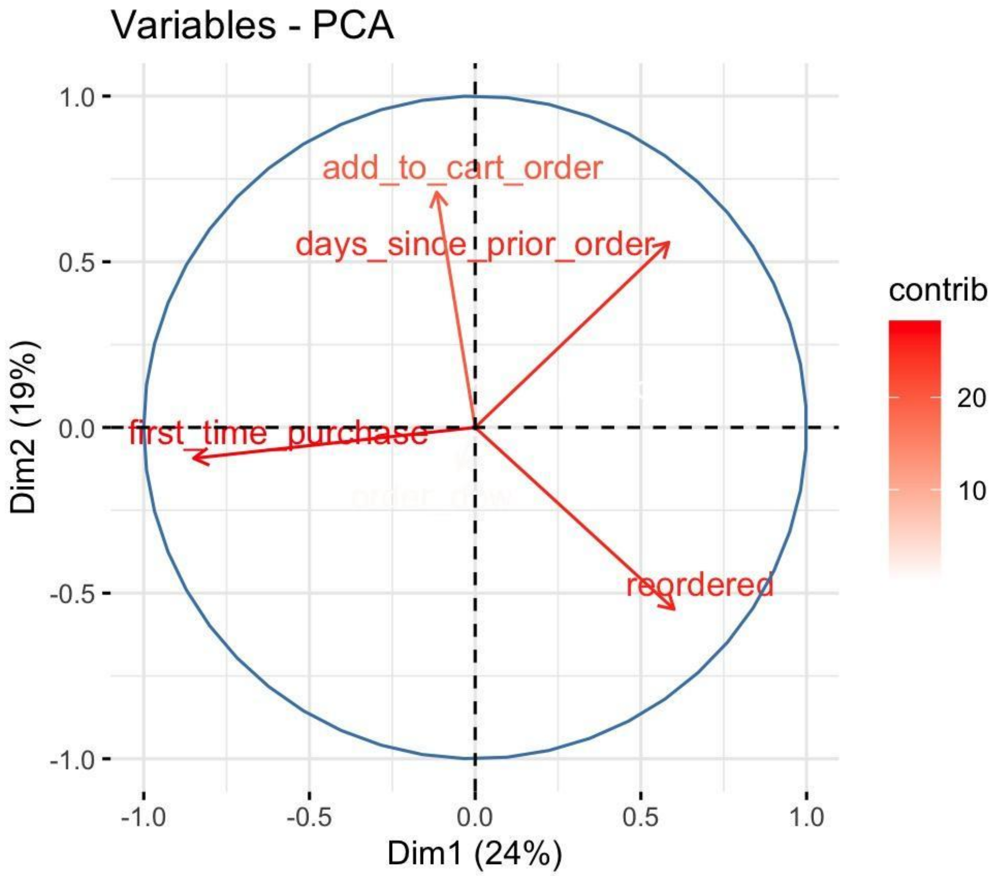
Initially, we evaluated the suitability of the data for factor analysis through correlation checks, KMO, and the Bartlett test. Observations from the scree plot suggested two components, yet other methodologies, such as fa.parallel, hinted towards three. At this juncture,





would offer us tangible evidence. However, with a KMO value of 0.45, falling below the commonly accepted threshold of 0.5, it hinted that our data might not be highly suitable for factor analysis. Even though we opted for three factors, they explained only 25% of the data variance, which was far from our expectations.

Seeking an alternative approach, we turned to Principal Component Analysis (PCA). While RStudio hinted at the selection of three principal components, they accounted for just 59% of the data variance. Here,



would provide a clearer perspective on this outcome. Although this percentage is higher, we felt that it's still insufficient for our study, given that we didn't want to lose significant details about customer buying habits.

Reflecting upon the entire process, lessons to be drawn include recognizing that a mere reduction in data dimensions when handling data related to consumer purchasing habits might lead to loss of crucial insights, potentially skewing subsequent findings and conclusions. Also, ensuring the data's aptness for factor analysis or PCA is pivotal.

Final Cluster Analysis

We choose to use k - means clustering. We first performed the Total within sum of squares Plot and ratio Analysis of plots. As shown in these two figures, the position of the black line is the "elbow" we observed, and the position of the elbow we observed in both figures is 4. Therefore, we finally decided: choose 4 clusters for further analysis.

After determining the number of clusters, we start to perform the k - means clustering. The parameters we choose, in addition to the previously determined " centers = 4 ", also have " iter.max = 100, nstart = 25 ". Then we list the amount of data in each cluster for subsequent analysis:

1	2	3	4
498040	508522	502657	505533

And we also store the cluster index to which each data belongs in a long integer, which is also for the convenience of our subsequent analysis.

It is also very important to obtain the insight of each cluster. By obtaining the average value of each column in different clusters, we can roughly grasp the situation and characteristics of each cluster. The specific results are as follows:

k_cluster	**order_id**	**user_id**	**order_number**	**order_dow**	**order_hour_of_day**
1	2993775	102994.3	-0.0005	0.003	0.0006
2	431226.1	102956.1	-0.0071	-0.0051	-0.0014
3	2142826	103112.1	0.0107	0.0018	-0.0006
4	1289366	103211.7	-0.003	0.0005	0.0014

days_since_prior_order	**product_id**	**add_to_cart_order**	**reordered**	**department_id**	**first_time_purchase**

0.0064	71.14	-0.0047	0.0025	9.9075	-0.0064
0.0052	71.1388	0.0014	-0.0035	9.8936	-0.0004
-0.0135	71.1778	-0.0013	0.0009	9.9114	0.0051
0.0019	71.0959	0.0046	0.0002	9.8967	0.0016

The above is the overall situation of the entire data set. Next we start our specific analysis on our research questions.

For question one, we set add_to_cart_order as our outcome. Then we perform Total within sum of squares Plot and ratio Plot. After analyzing them, we decided to set the number of clusters to 2. Then we start to apply a clustering solution from ‘Train’ to ‘Test’. We use ‘as. kcca’ in the package ‘flexclust’ to complete this process. Then we use ‘table()’ to confirm that we have divided the data into two clusters. After confirmation, we split train and test based on cluster membership, that is, divide the train into train 1 and train 2, and divide the test into test 1 and test 2 according to the cluster to which it belongs.

Then we predict test 1 and test 2 through regression according to train 1 and train 2 and calculate ‘sse ' respectively. Then we compare the obtained ‘sse’ with the previous ‘sseLinear’. The result is that the ‘sse’ of the overall in question one is slightly smaller. This can strengthen the credibility of our conclusion on question one. But in question 2, sse is slightly greater than " sseLinear ", which may indicate that the construction of our cluster model has some drawbacks, and we may have not been able to use ‘sse’ to strengthen our conclusion on question 2.

Conclusion

In conclusion, our analysis confirms that Hypothesis 1 is indeed supported, as it demonstrates that the frequency of purchasing specific types of products does influence customer buying habits and decisions at Hunter's Supermarket. Besides, our analysis also provides support for Hypothesis 2 that the timing of purchases does have an impact on the types of products consumers select at Hunter's Supermarket. Variables such as *order_dow* (day of the week), *order_hour_of_day* (hour of the day), and *days_since_prior_order* (number of days since the previous order) have coefficients that are statistically significant, indicating that the timing of purchases does indeed influence the types of products consumers choose to buy. Based on our cluster analysis, we have concluded that different customer clusters exhibit variations in average values for certain variables, highlighting the existence of diverse buying preferences among different groups.

The following table compares the average purchasing behaviors of the customers in each cluster with the overall average level.

Cluster	**Order Number (Number of times a user purchased at Hunter)**	**Order Hour of each Day**	**Days Since Prior Order**	**Add to Cart Order (Number of items added to cart)**	**Reordered**
1	Less	Earlier	Higher	More	No difference
2	Less	Later	Higher	More	Less
3	Much More	No difference	Much Lower	Less	No difference
4	Less	No difference	Higher	Less	More

Based on the results, Cluster 3 stands out as the most prominent and noteworthy cluster. Users belonging to this particular cluster demonstrate a higher amount and frequency of purchases and a greater tendency to engage in repeat purchases, suggesting their potential loyalty to Hunter Supermarket. To further capitalize on this user segment, it is recommended to implement targeted strategies such as personalized recommendations and promotional campaigns aimed at enhancing their purchase volume and their engagement in repeat purchases.

For research question 2, our result indicates that there might not be significant variations or distinct patterns within different customer groups identified by the clustering algorithm. However, we can still provide effective strategies for decision makers if they wish to fulfill specific customer needs at different times of the day. For example, for Cluster 1 that tends to place orders earlier in the day, we could consider highlighting breakfast and morning-related products or promotions based on their past preferences. This could include breakfast cereals, fresh bakery items, coffee, tea, or morning-specific discounts. Adjusting marketing messages to emphasize convenience and quick solutions for busy mornings can also be effective. For Cluster 2 that tends to place orders later in the day, we could focus on promoting dinner or evening-related products based on their past preferences. This could include ready-to-cook meal kits, ingredients for quick and easy dinners, snacks for late-night cravings, or offers for evening-specific discounts. Emphasize convenience, relaxation, and indulgence in the marketing messages to cater to their preferences.

By considering the timing preferences of different customer clusters and tailoring promotions, product offerings, and overall marketing strategies accordingly, Hunter's Supermarket can effectively capture the attention and meet the needs of its customers at different times of the day, ultimately enhancing customer satisfaction and purchase behavior.

Bibliography

Hunter. (2017). “Supermarket dataset for predictive marketing 2023”. https://www.kaggle.com/datasets/hunter0007/ecommerce-dataset-for-predictive-marketing-2023

Peker, S., Kocyigit, A. and Eren, P.E. (2017), "LRFMP model for customer segmentation in the grocery retail industry: a case study", Marketing Intelligence & Planning, Vol. 35 No. 4, pp. 544-559. https://doi.org/10.1108/MIP-11-2016-0210

Wedel M, Kamakura WA (2000) Market segmentation. Conceptual and methodological foundations. Kluwer Academic Publishers, Boston

Tsai, C.-Y. and Chiu, C.-C., (2004), "A purchase-based market segmentation methodology", Expert Systems with Applications, Vol. 27 No. no. 2, pp. 265-276