**Capstone Project –**

**Segmentation and Predictive Modeling**

BIA 690 Capstone

Professor Nabeel

Tristin Burdick

December 14, 2023

**Table of Contents**

[Introduction 2](#_Toc153439997)

[Methodology 3](#_Toc153439998)

[Results and Analysis 4](#_Toc153439999)

[Solution 10](#_Toc153440000)

[Conclusion 12](#_Toc153440001)

[Appendix 13](#_Toc153440002)

[Data Source 17](#_Toc153440003)

[References 17](#_Toc153440004)

## Introduction

Convenient Food Mart (CFM), founded in 1958, is a nationwide chain of convenience stores in the United States and Mexico. Their initial vision was to provide a small grocery store experience that was easy and quick for customers. Primarily located in neighborhoods, these stores offered a practical alternative to larger supermarkets. Their business model was simple but effective: bring the store to the customer, not the customer to the store. As reported by CSP Daily News in 2017, CFM's strategic placement within walking distance for many residents gave them an advantage (CSP Daily News, 2017). For decades, these stores were the perfect solution for anyone needing to grab a few items without the hassle of a long drive.

In the past, CFM’s stores were often the first choice for people looking to pick up daily essentials. They stocked a variety of items, from snacks and beverages to basic household supplies. This range of products, combined with the convenience of their locations, made them a staple in the daily lives of many consumers. People knew they could rely on CFM for a quick shopping trip, and this reliability turned many first-time visitors into regular customers.

However, the retail world has been changing rapidly with the advancement of digital technology and the growth of e-commerce. The convenience that once set CFM apart is now being challenged by the convenience of online shopping. With just a few clicks, consumers can now have a wide range of products delivered to their doorsteps, sometimes at prices lower than those at physical stores. This shift has led CFM to reevaluate their approach to business. The company recognized that it needed to attract new customers who were becoming accustomed to the digital marketplace. To stay competitive, CFM began to look beyond their immediate neighborhoods and consider broader community outreach.

To adapt to these shifts, CFM Management has launched targeted marketing campaigns in various U.S. regions and select locations in Mexico. The company has been collecting detailed data from these campaigns, including who visits their stores, what items they are buying, and which types of advertising are most effective in driving them to CFM locations. They are seeking help from a data professional to provide insights on next steps for CFM Management and how they can effectively reach out and retain new customers.

To assist with CFM Management, they’ve engaged Tristin, a data analyst, to develop a comprehensive data solution. Tristin's initial task is to provide CFM with a detailed analysis of customer profiles derived from the recent marketing campaigns. This will clarify which consumer segments are engaging with CFM's campaigns. Following this, Tristin is responsible for developing a predictive model that will forecast the customer acquisition cost for future marketing efforts. This model will enable CFM to strategically plan their budget by estimating the investment required to gain each new customer.

## Methodology

For this business case, CFM will supply the necessary data in a CSV file extracted from their internal database. This file will be made accessible to Tristin, allowing him to download it to his computer for further analysis. The dataset used for this project is sourced from Kaggle.com, which is known for its comprehensive repository of open-source data.

The data provided by CFM, sourced directly from their internal systems, will already have undergone initial cleaning and preprocessing. Tristin's task will be to further refine this data to ensure it is fully prepared for analysis and modeling. He will apply data cleaning techniques including standardizing column names, adjusting data types, encoding categorical variables, and discarding any irrelevant columns. These steps are essential for the accuracy and efficiency of his analysis and predictive modeling.

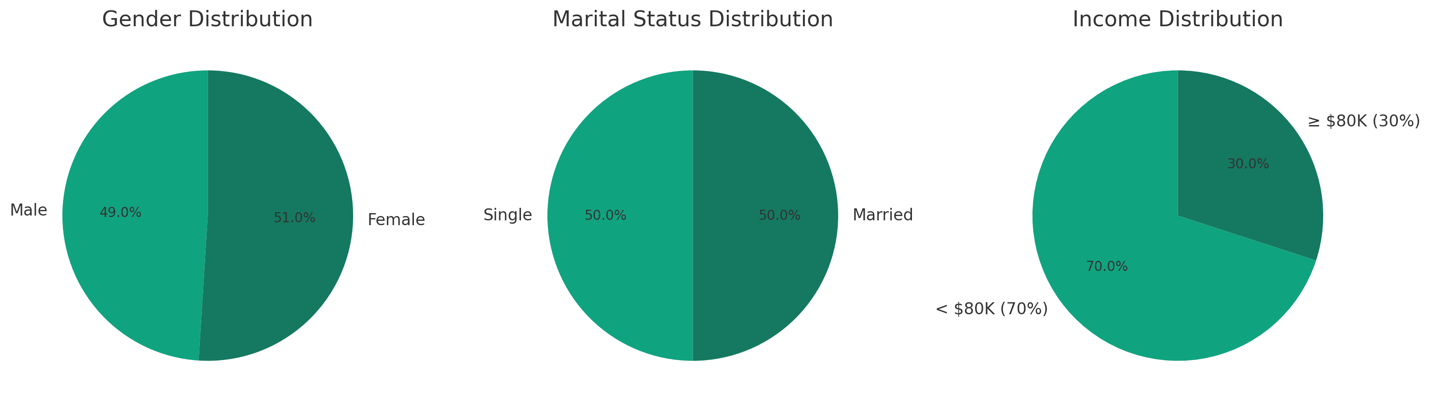
Tristin will utilize a Python environment to process the data and develop his scripts for analysis and model construction. His initial step will involve examining the dataset in Excel to implement any immediate modifications. Following this, he will transfer the data into a Google Colab environment to carry out more sophisticated analysis and modeling.

Tristin's initial step will be to perform exploratory data analysis (EDA) to gain a comprehensive understanding of the dataset. He will examine the size of the dataset, review the number of rows and columns, and compile descriptive statistics. Additionally, he will create visual representations including frequency charts, boxplots, and a correlation matrix to identify patterns and insights within the data.

Tristin will begin his modeling efforts with consumer segmentation to understand the various customer groups within CFM's marketing campaigns, thereby enhancing future personalized campaign strategies. He plans to use k-means clustering for this segmentation and will determine the optimal number of clusters by applying the elbow method. An under/over index analysis will be used to determine the key characteristics of each segment to understand what makes them unique.

Finally, Tristin will develop a predictive model incorporating the segmentation labels and other relevant features to estimate customer acquisition costs. This model will serve as a strategic tool for CFM management, enabling them to forecast the expenses associated with acquiring new customers and facilitating the allocation of marketing resources.

## Results and Analysis

 **Figure 1: Key Demographic Variables**

The dataset presents a comprehensive view of consumer, product, and marketing-related variables for Convenient Food Mart. After preprocessing the data, the entire file contained over 50,000 rows of data with 312 features.

#### Key demographic variables:

* Male: 49%, Female: 51%
* Single: 50%, Married 50%
* 70% make less than $80k per year
* 2.5 kids on average
* 2.2 cars at home

#### Target variable:

A graph and diagram of a graph

Description automatically generated with medium confidence

**Figure 2: Target Variable**

The frequency distribution and boxplot of our target variable (customer acquisition cost) reveals a few key things. There appears to be a bimodal distribution in the target variable, with multiple high peaks. This could indicate multiple consumer segments, each with different CAC values. The average customer acquisition cost for CFM is $99.31. We can expect this number to increase or decrease depending on the input variables used in the predictive model.

#### Correlation matrix:

A screenshot of a computer

Description automatically generated

**Figure 3: Correlation Matrix**

The correlation matrix shows the correlated values between all variables. This is helpful for identifying multicollinearity in our model and for determining which variables influence each other. For instance, store\_sq\_ft has a correlation of 0.78 to frozen sq\_ft, telling us that the larger a store is, the larger frozen food section each store will have.

K-means segmentation was used for cluster analysis because it effectively groups customers into distinct categories based on similarities within their data, which is essential for understanding and targeting different market segments (Ong, 2020). This method is ideal for this problem to recognize patterns within the data and tailor marketing strategies to reduce customer acquisition costs (CAC) and optimize resource allocation.

#### Identifying Clusters (Elbow Method):

A screen shot of a computer

Description automatically generated

**Figure 4: Elbow Method for Cluster Identification**

The elbow method was used to determine the number of clusters found within the dataset. According to the chart, there should be approximately 4 unique clusters given the point where the graph starts to level out.

#### How characteristics in each cluster were identified

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
| total\_children | 101 | 98 | 100 | 100 |
| num\_children\_at\_home | 105 | 92 | 100 | 103 |
| avg\_cars\_at\_home | 96 | 101 | 101 | 101 |
| marital\_status\_S | 99 | 104 | 99 | 98 |
| gender\_M | 83 | 106 | 105 | 107 |
| avg\_yearly\_income\_40000.0 | 96 | 101 | 102 | 101 |
| avg\_yearly\_income\_60000.0 | 117 | 94 | 94 | 95 |
| avg\_yearly\_income\_80000.0 | 96 | 100 | 106 | 98 |
| avg\_yearly\_income\_100000.0 | 82 | 117 | 103 | 97 |
| avg\_yearly\_income\_120000.0 | 107 | 100 | 102 | 91 |
| education\_Graduate Degree | 76 | 104 | 104 | 116 |
| education\_High School Degree | 96 | 102 | 104 | 98 |
| education\_Partial College | 108 | 96 | 94 | 102 |
| education\_Partial High School | 105 | 99 | 98 | 98 |
| occupation\_Management | 114 | 95 | 100 | 91 |
| occupation\_Manual | 85 | 104 | 102 | 109 |
| occupation\_Professional | 93 | 103 | 101 | 104 |
| occupation\_Skilled Manual | 110 | 98 | 98 | 93 |
| houseowner\_Y | 95 | 101 | 101 | 103 |
| member\_card\_Golden | 111 | 92 | 97 | 99 |
| member\_card\_Normal | 99 | 101 | 99 | 101 |
| member\_card\_Silver | 85 | 108 | 103 | 104 |
| product\_food\_family\_Food | 133 | 133 | 0 | 134 |
| product\_food\_family\_Non-Consumable | 31 | 31 | 308 | 30 |
| sales\_country\_Mexico | 251 | 41 | 72 | 36 |
| sales\_country\_USA | 0 | 126 | 117 | 157 |
| store\_type\_Gourmet Supermarket | 195 | 123 | 83 | 0 |
| store\_type\_Mid-Size Grocery | 163 | 39 | 88 | 111 |
| store\_type\_Small Grocery | 0 | 0 | 110 | 290 |
| store\_type\_Supermarket | 64 | 30 | 102 | 204 |

**Table 1: Indices of Each Persona**

To identify key characteristics within each cluster, an index was developed using the average values of each feature across all clusters (as shown in Table 1 above). This method involved calculating the centroids, or the central point, for each feature within each cluster. By comparing these centroids to the overall averages, an index was created for every feature in each cluster. Features with indices below 90 or above 110 were particularly noteworthy, as they highlighted the most distinct characteristics of each customer segment, helping to understand what makes each group unique.



**Figure 5: Persona 1**

Cluster 1 at Convenient Food Mart (CFM) predominantly consists of women who generally earn around $60k per year, with a notable preference for management-related positions. They exhibit a high preference for shopping at stores offering a variety of amenities and large grocery sections, indicative of their health-conscious nature and propensity for low-fat products. This group is also more likely to be golden cardholders and prefers to shop at gourmet supermarkets and mid-size grocery stores, primarily in Mexican locations. They are most receptive to marketing through street handouts, television and radio advertisements, and Sunday newspapers, showing a lesser tendency to use in-store coupons.



**Figure 6: Persona 2**

Cluster 2 customers share characteristics with the average consumer but have a higher average income of around $100k. They display a strong preference for store amenities such as coffee bars, video stores, salad bars, prepared food, and florists, with a notable inclination towards larger frozen and meat sections. These consumers are more likely to hold silver loyalty cards and show a preference for gourmet supermarkets within the United States. In terms of marketing, they are most likely to be swayed by Sunday newspapers and in-store coupons, while being less influenced by street handouts and radio ads.



**Figure 7: Persona 3**

Cluster 3 represents the most ‘average-looking’ consumers demographically and do not prioritize low-fat products, suggesting less health consciousness than most shoppers. They are average card members who typically shop for non-consumables at small grocery stores and supermarkets across the United States. Their conversion into customers is most likely through Sunday newspapers, with street handouts being the least effective. This group has a customer acquisition cost that is close to the average for CFM (around $99).



**Figure 8: Persona 4**

Cluster 4 customers, while resembling the average consumer, tend to have college graduate degrees. They are less likely to frequent stores with additional amenities and have an average likelihood of holding loyalty cards. Their shopping is food-oriented, with a preference for small to mid-size grocery stores in the United States. They are most likely to be converted into customers by daily papers, in-store coupons, and cash register handouts, and are least likely to be influenced by TV and radio ads, or product attachment ads. Like Clusters 2 and 3, they also represent an average acquisition cost for CFM.

The silhouette score is one metric used for evaluating the performance of a k-means segmentation model. The silhouette score is used to calculate the goodness of the clustering technique, ranging from -1 to 1 (Bhardwaj, 2020). In this case, 1 would indicate the clusters are far apart from each other and are truly distinct, 0 means the clusters are indifferent and may even overlap. The silhouette score of the k-means segmentation model was 0.25, which indicate the clusters are somewhat distinct but not highly separated.

The predictive model chosen for estimating customer acquisition cost was XGBoost, a type of ensemble learning method. XGBoost stands out due to its gradient-boosting framework which constructs decision trees in a sequential manner, with each tree correcting the errors of the previous one (Analytics Vidhya, 2023). This process enhances both accuracy and robustness with each iteration. Due to its scalability and computational efficiency, XGBoost is well-suited for large datasets and was particularly useful for this business case, which required the identification of complex patterns and the modeling of non-linear relationships in the data. The model utilized 30 significant features for training, with five of the most important features being shown below.

#### Top 5 Features:

1. Grocery stores frozen section sqft (0.06)
2. Grocery stores grocery section sqft (0.04)
3. Marketing campaign “Price Slashers” (0.03)
4. Marketing campaign “Big Time Discounts” (0.03)
5. Advertising strategy – Cash Register Handout (0.03)

A graph of blue dots

Description automatically generated

**Figure 9: XGBoost Residual Plot**

MSE: 25.64

MAE: 3.72

R2: 0.97

The XGBoost model performed adequately with the dataset provided. The residual plot showed that residuals are scattered around the horizontal line, which typically indicates a model that fits the data well. However, a pattern was observed where residuals increased with higher actual values, pointing to inaccuracies in predictions for larger numbers. The R-squared value of 0.97 suggests that the model can explain most of the variability of the target variable. The mean squared error (MSE) and mean absolute error (MAE) are relatively low.

The model showed good performance on the training dataset. However, there is a potential risk of overfitting as the model is applied to predict new, real-world data. For future applications, implementing cross-validation would help verify the model's performance on test data and any new data collected subsequently. Ongoing efforts in feature engineering and hyperparameter tuning are necessary to prevent overfitting and to preserve the model's predictive accuracy in practical use.

## Solution

The core business challenge faced by Convenient Food Mart (CFM) lies in enhancing the effectiveness of its marketing strategies to boost customer acquisition and retention. This challenge is paramount in ensuring CFM’s sustained growth and market competitiveness.

The analytical approach in this report, consisting of customer segmentation and predictive modeling, has yielded valuable insights. The segmentation analysis revealed distinct customer groups, each characterized by unique purchasing behaviors and preferences. This deep understanding of CFM’s customer base is crucial for tailoring marketing efforts. For instance, one segment identified was particularly receptive to health-conscious products, while another showed a preference for gourmet and specialty items. Detailed segmentation allows for marketing strategies that resonate deeply with each group, thereby increasing engagement and conversion rates.

Furthermore, the predictive modeling used (XGBoost) provided a clear understanding of customer acquisition costs (CAC). This model has the ability to estimate a predicted acquisition value for each customer based on a given set of input variables. For instance, we can predict how much it would cost us to acquire a customer in a certain region, using specific marketing mediums, or even how much it will cost to acquire a customer from a particular store. This model helps CFM in allocating marketing resources to the proper channels to maximize return on Investment (ROI).

These findings resemble a multifaceted solution to CFM’s data-driven marketing strategy. By utilizing the insights gained from the segmentation, CFM can develop targeted marketing initiatives, such as personalized email campaigns, customized in-store promotions, and targeted online advertisements, that cater to the specific needs and preferences of each segment.

Budget allocation will be guided by the insights from the predictive modeling. Marketing funds will be strategically directed towards segments that not only show higher potential for conversion but also align with lower CAC. This approach ensures that marketing spend is not just effective but also efficient.

Additionally, CFM could implement a dynamic pricing model to adapt to market changes. A dynamic pricing model is a product pricing model used to adjust prices of goods in real-time using external factors such as market demand, season, supply changes, and more (Dublino, 2023). This model not only drives sales but also enhances customer satisfaction by offering value when it matters most to the consumer.

Since part of CFM’s goal was to find a data-driven solution for customer retention, customer relationship management (CRM) software could also be beneficial to implement in conjunction with the segmentation and predictive models. According to Sarah Williams from Big Contacts, CRMs can help improve customer retention by keeping your customer data organized, personalizing the content they receive, and automating crucial processes, among other benefits (Williams, 2023). The CRM will help improve customer retention and can also be used as a data tool to feed into the segmentation and predictive models so they can perform better over time.

Overall, this solution is particularly suited to CFM’s business case for several reasons. First, the solution proposed in grounded in data insights, ensuring that the suggestions for solving the business problem is data-driven, and not determined by hunches or guesses. Second, the solution offered provides a high degree of personalization, which is key in today’s marketing landscape. Customers are more likely to engage with brands that understand their unique needs and preferences. Finally, this approach is scalable. As CFM grows and the market evolves, the strategies can be adapted to meet new challenges and opportunities, ensuring long-term sustainability and success for CFM.

## Conclusion

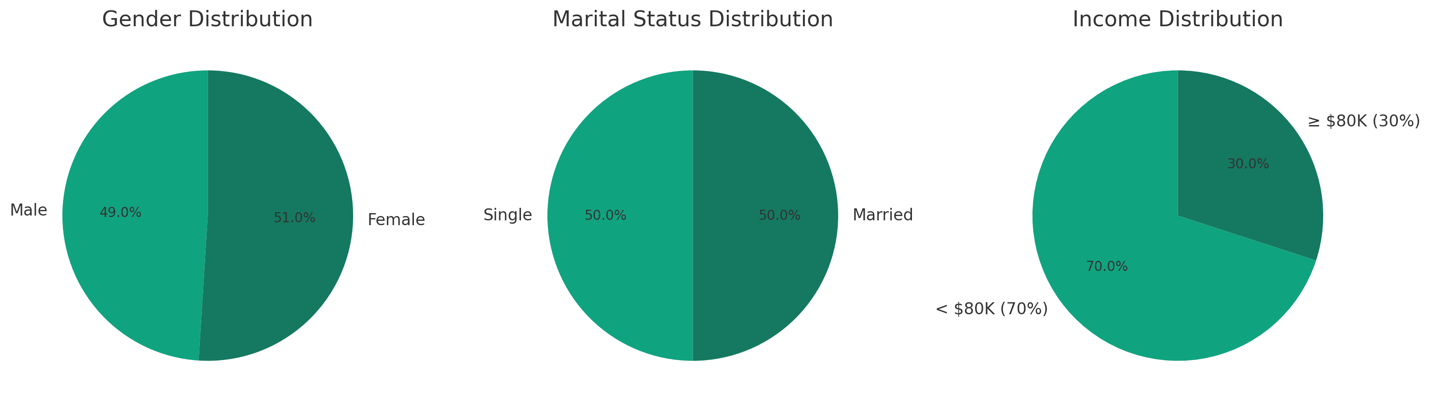
The data-driven approach, centered around customer segmentation and predictive modeling, provides CFM with a clear path to optimize their marketing campaigns. By focusing on specific customer segments identified through detailed analysis, CFM can tailor their marketing efforts more effectively, ensuring that resources are directed towards the most promising and cost-effective areas. The implementation of a dynamic pricing strategy further enhances CFM's ability to respond to market changes and customer preferences in real-time, potentially increasing both sales and customer satisfaction.

The predictive model, developed using XGBoost, offers a strategic tool for forecasting customer acquisition costs. This enables CFM to make informed decisions on where and how to allocate marketing resources, maximizing the return on investment. The model's findings, particularly the identification of top features influencing customer acquisition costs, guide CFM in refining their marketing strategies.

Additionally, the recommendation to implement a Customer Relationship Management (CRM) system is aimed at improving customer retention. A well-integrated CRM can help CFM in organizing customer data, personalizing interactions, and automating crucial processes. This not only aids in retaining existing customers but also provides valuable data that feeds back into the segmentation and predictive models, enhancing their accuracy and effectiveness.

The conclusion emphasizes that this data-driven solution is not a one-off strategy but a dynamic, evolving approach. As CFM continues to grow and the market landscape changes, the strategies can be adapted to meet new challenges and seize new opportunities for CFM.

## Appendix



**Figure 1: Key Demographic Variables**

#### Target variable:

A graph and diagram of a graph

Description automatically generated with medium confidence

**Figure 2: Target Variable**

A screenshot of a computer

Description automatically generated

**Figure 3: Correlation Matrix**

A screen shot of a computer

Description automatically generated

**Figure 4: Elbow Method for Cluster Identification**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
| total\_children | 101 | 98 | 100 | 100 |
| num\_children\_at\_home | 105 | 92 | 100 | 103 |
| avg\_cars\_at\_home | 96 | 101 | 101 | 101 |
| marital\_status\_S | 99 | 104 | 99 | 98 |
| gender\_M | 83 | 106 | 105 | 107 |
| avg\_yearly\_income\_40000.0 | 96 | 101 | 102 | 101 |
| avg\_yearly\_income\_60000.0 | 117 | 94 | 94 | 95 |
| avg\_yearly\_income\_80000.0 | 96 | 100 | 106 | 98 |
| avg\_yearly\_income\_100000.0 | 82 | 117 | 103 | 97 |
| avg\_yearly\_income\_120000.0 | 107 | 100 | 102 | 91 |
| education\_Graduate Degree | 76 | 104 | 104 | 116 |
| education\_High School Degree | 96 | 102 | 104 | 98 |
| education\_Partial College | 108 | 96 | 94 | 102 |
| education\_Partial High School | 105 | 99 | 98 | 98 |
| occupation\_Management | 114 | 95 | 100 | 91 |
| occupation\_Manual | 85 | 104 | 102 | 109 |
| occupation\_Professional | 93 | 103 | 101 | 104 |
| occupation\_Skilled Manual | 110 | 98 | 98 | 93 |
| houseowner\_Y | 95 | 101 | 101 | 103 |
| member\_card\_Golden | 111 | 92 | 97 | 99 |
| member\_card\_Normal | 99 | 101 | 99 | 101 |
| member\_card\_Silver | 85 | 108 | 103 | 104 |
| product\_food\_family\_Food | 133 | 133 | 0 | 134 |
| product\_food\_family\_Non-Consumable | 31 | 31 | 308 | 30 |
| sales\_country\_Mexico | 251 | 41 | 72 | 36 |
| sales\_country\_USA | 0 | 126 | 117 | 157 |
| store\_type\_Gourmet Supermarket | 195 | 123 | 83 | 0 |
| store\_type\_Mid-Size Grocery | 163 | 39 | 88 | 111 |
| store\_type\_Small Grocery | 0 | 0 | 110 | 290 |
| store\_type\_Supermarket | 64 | 30 | 102 | 204 |

**Table 1: Indices of Each Persona**



**Figure 5: Persona 1**



**Figure 6: Persona 2**



**Figure 7: Persona 3**



**Figure 8: Persona 4**

A graph of blue dots

Description automatically generated

**Figure 9: XGBoost Residual Plot**

## Data Source

<https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart>

## Python Script

<https://colab.research.google.com/drive/1UPjeWuc8RZ4Ey5RWA8imirX3cmImu3gg?usp=sharing>

## References

Analytics Vidhya. (2023, March 30). *Introduction to XGBoost Algorithm in Machine Learning*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/#:~:text=XGBoost%20is%20a%20machine%20learning,won%20several%20machine%20learning%20competitions.

Bhardwaj, A. (2020, May 26). *Silhouette Coefficient*. Retrieved from Towards Data Science: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c#:~:text=Silhouette%20Coefficient%20or%20silhouette%20score%20is%20a%20metric%20used%20to,each%20other%20and%20clearly%20distinguished.

CSP Daily News. (2017). *CONVENIENT FOOD MART*. Retrieved from CSP Daily News: https://www.cspdailynews.com/top-202-convenience-stores-2017/convenient-food-mart

Dublino, J. (2023, November 6). *What Is Dynamic Pricing, and How Does It Affect E-commerce?* Retrieved from Business: https://www.business.com/articles/what-is-dynamic-pricing-and-how-does-it-affect-ecommerce/

Ong, A. (2020, March 16). *Segmenting Customers using K-Means, RFM and Transaction Records*. Retrieved from Towards Data Science: https://towardsdatascience.com/segmenting-customers-using-k-means-and-transaction-records-76f4055d856a

Raj, A. (2020, June 18). *A Quick and Dirty Guide to Random Forest Regression*. Retrieved from Towards Data Science: https://towardsdatascience.com/a-quick-and-dirty-guide-to-random-forest-regression-52ca0af157f8#:~:text=ADVANTAGES%20OF%20RANDOM%20FOREST,of%20the%20data%20are%20missing.

Williams, S. (2023, June 14). *CRM Customer Retention: 7 Best Ways to Increase Customer Retention With CRM*. Retrieved from Big Contacts: https://www.bigcontacts.com/blog/crm-customer-retention/