UNIVERSITÀ DEGLI STUDI DELL'AQUILA

Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica

Tesi di Laurea Magistrale in Ingegneria Matematica

Clustering Customers in a Retail Market: Data Driven Approaches

Relatore
Prof. Fabrizio Rossi

Correlatori
Andrea Manno
Prof. Stefano Smriglio

Laureando
Vipul Chalotra
280269

Anno Accademico 2022-2023

"It always seems impossible until it's done."

- Nelson Mandela

Abstract

The retail industry, as one of the largest customer-driven sectors, faces significant challenges in comprehending consumer habits. This thesis delves into the utilization of customer clustering techniques to optimize marketing strategies within the ever-changing dynamics of the retail market. Acknowledging the growing intricacies of consumer behavior, retailers are increasingly embracing data-driven approaches to gain insights into customer preferences and purchasing patterns. The central aim of this study is to pinpoint distinct customer segments within a retail context, utilizing an array of data-driven methods, and analyze their characteristics to tailor marketing efforts effectively.

The research employs a blend of statistical methods and machine learning algorithms to unravel nuanced understandings of customers based on their historical transaction data, demographic information, and behavioral attributes. The dataset, encompassing a diverse range of customers from two locations of a leading retail chain, provides a rich source for analysis. The study zeroes in on extracting meaningful patterns that unveil variations in shopping behavior, preferences, and responsiveness to marketing stimuli.

Results spotlight the existence of discernible customer segments, each showcasing unique preferences and characteristics. These identified clusters illuminate the heterogeneity in customer needs and expectations, providing valuable insights for targeted marketing strategies. Furthermore, the study delves into the potential impact of these findings on key performance indicators.

The implications of customer clustering transcend marketing, influencing inventory management, product assortment, and overall business strategy. By comprehending the diverse needs of distinct customer segments, retailers can optimize resource allocation and enhance the overall customer experience.

This research significantly contributes to the evolving field of retail analytics by presenting a comprehensive framework for customer segmentation. The findings empower retailers to transcend generic marketing approaches and embrace personalized strategies resonating with specific customer groups. As the retail industry evolves, the insights derived from this study offer a valuable foundation for enhancing competitiveness and fostering customer-centric retail practices.

Keywords: Customer Clustering, Data Analysis, Data Driven Methods, Machine Learning Algorithms, Retail Analytics, Retail Industry, Statistical Methods, Targeted Marketing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The retail landscape has undergone a profound transformation, propelled by technological advancements and changing consumer behaviors. In this era of digitization, retailers are inundated with vast amounts of data, presenting both opportunities and challenges. Understanding and effectively utilizing this data is imperative for retailers aiming to stay competitive and relevant.

Traditionally, customer segmentation in the retail sector has relied on broad categorizations, often overlooking the nuanced and dynamic nature of consumer preferences. As consumers demand personalized experiences, the need for more sophisticated and data-driven approaches to customer clustering becomes evident. The amalgamation of advanced analytics, machine learning, and statistical techniques provides a promising avenue for retailers to unravel intricate patterns within their customer base.

This research delves into the realm of customer clustering in the retail market, seeking to bridge the gap between traditional segmentation methods and the evolving expectations of modern consumers. By exploring data-driven approaches, this study aims to unearth actionable insights that can empower retailers to tailor their strategies, enhance customer satisfaction, and ultimately thrive in an increasingly competitive marketplace.

1.2 Research Problem

Customer segmentation is the process of dividing a large and diverse customer base into smaller homogeneous groups based on certain characteristics such as demographics, behaviors, and preferences. This process helps the service providers better understand their customers' needs, and thus, provide targeted customer service. Overall, customer segmentation is a valuable approach that helps businesses understand their customers better, tailor their marketing efforts, and improve their sales.

With the increasing digitization, the challenges and the opportunities associated with customer clustering are increasing as well. Some of the major challenges associated with customer segmentation are as follows:

1. Data Quality: The process of clustering depends heavily on validity of the data. Incorrect or missing data can potentially lead to incorrect and/or useless clusters.

2. Costs: Overall customer segmentation can be expensive if the data collection channels are extensive surveys. The benefits of customer segmentation for smaller businesses might not be a profitable investment.

3. Over-Segmentation: In general, understanding that clustering can be a useful resource is important. But deciding when further clustering might be pointless in terms of business profits and advancement is necessary.

4. Inadequate Testing: Segmentation strategies need regular testing and refining to ensure that they are effective. Inadequate testing and optimizing can lead to campaigns that do not resonate with customers and/or do not generate the desired results [18].

In this research, our primary focus revolves around leveraging data-driven approaches for customer segmentation and addressing the inherent challenges to transform them into opportunities. The advent of advanced analytical techniques has significantly eased the task of managing data quality, offering efficient quality control mechanisms. The overall costs associated with data collection have witnessed a decline, given the widespread adoption of digital channels, thereby facilitating easier data analysis. Moreover, modern data-driven approaches for segmenting and analyzing trends are equipped with robust safeguards against pitfalls, preventing the risk of over-segmentation. Notably, advanced machine learning techniques are incorporating real-time testing and calibration, reinforcing the argument that, with technological advancements and increased digitization, data-driven approaches to customer segmentation are gaining substantial ground.

1.3 Objectives of the Study

The primary objectives of this research are to systematically investigate, analyze, and derive actionable insights into the realm of customer clustering in the retail market. The study aims to address the following key objectives:

1.3.1 Conduct Descriptive Analysis to Uncover Customer Characteristics

The first fundamental objective is to conduct comprehensive descriptive analysis to uncover baseline customer characteristics, providing a foundational understanding of the customer base. This involves exploring key demographic information, transactional patterns, and engagement metrics across the entire customer dataset. By establishing this baseline, the research aims to set the context for subsequent clustering analyses, ensuring that the identified segments are not only algorithmically derived but also align with and enrich the inherent characteristics of the customer population.

1.3.2 Identify Effective Data-Driven Customer Clustering Methods

The second objective is to assess and compare various data-driven customer clustering methods, including but not limited to machine learning algorithms and statistical techniques. By evaluating the efficacy of these methods, the research aims to provide insights into the most suitable approaches for segmenting customers in the retail context. This involves a comprehensive examination of clustering algorithms such as K-Means and other advanced machine learning models.

1.3.3 Assess the Impact of Customer Clustering on Retail Performance Metrics

The third objective is to measure the impact of employing data-driven customer clustering on key retail performance metrics. These metrics include but are not limited to customer retention, average transaction value, conversion rates, and overall sales. By quantifying the influence of customer clustering strategies, the research seeks to provide empirical evidence of the benefits that retailers can derive from more targeted and personalized approaches.

1.3.4 Understand Customer Segmentation Patterns and Characteristics

The fourth objective is to gain a deep understanding of the patterns and characteristics inherent in the identified customer segments. This involves conducting a thorough analysis of the clusters generated through data-driven methods, exploring commonalities, differences, and potential behavioral trends within each segment. By elucidating these patterns, the research aims to offer retailers actionable insights into the diverse needs and preferences of their customer base.

1.3.5 Provide Practical Recommendations for Retail Strategy Enhancement

Building on the findings from the data analysis, the final objective is to formulate practical recommendations for retailers seeking to enhance their strategies based on customer clustering insights. These recommendations will be grounded in the identified patterns and performance impacts, offering a road-map for implementing data-driven approaches in a manner that aligns with industry best practices and ethical considerations.

1.4 Research Questions

This study aims to address the following research questions, which collectively guide the investigation into customer clustering in the retail market using data-driven approaches:

What Baseline Customer Characteristics Emerge from Descriptive Analysis, and How Do They Inform Subsequent Clustering Strategies?

This fundamental research question delves into the insights gained from conducting descriptive analysis to uncover baseline customer characteristics. The study aims to identify key demographic information, transactional patterns, and engagement metrics across the entire customer dataset. By understanding these foundational aspects, the research seeks to elucidate how the identified baseline characteristics can inform and enrich subsequent data-driven clustering strategies within the retail sector.

What Data-Driven Methods are Most Effective for Customer Clustering in the Retail Sector?

This primary research question seeks to evaluate and compare the effectiveness of various data-driven methods for customer clustering within the retail sector. The study will explore clustering algorithms such as k-means, hierarchical clustering, and machine learning models to identify which methods yield the most meaningful and actionable customer segments.

How Does Customer Clustering Impact Key Retail Performance Metrics?

This research question focuses on assessing the tangible impact of data-driven customer clustering on key performance metrics in the retail industry. By analyzing customer behavior and purchasing patterns, the study aims to quantify the influence of clustering strategies on metrics such as customer retention, average transaction value, conversion rates, and overall sales.

What Patterns and Characteristics Define Identified Customer Segments?

To deepen our understanding of the customer segments identified through data-driven clustering, this research question aims to uncover the inherent patterns and characteristics within each segment. By exploring commonalities and differences, the study seeks to reveal insights into the diverse needs, preferences, and behaviors of distinct customer groups.

How Can Retailers Enhance Strategies Based on Data-Driven Customer Clustering Insights?

This research question focuses on providing practical recommendations for retailers based on the analysis of data-driven customer clustering. By synthesizing the findings, the study aims to offer actionable insights that guide retailers in optimizing their strategies, tailoring marketing efforts, and improving overall customer satisfaction.

These objectives align with the research questions to collectively form the foundation of this investigation into customer clustering in the retail market. Through empirical analysis and interpretation of results, the study aims to contribute valuable insights that can inform both academic discourse and practical applications within the dynamic landscape of retail. Subsequent chapters will delve into the methodology employed to achieve the stated objectives while answering these questions, presenting a robust framework for the empirical research that follows.

Chapter 2

Literature Review

2.1 Retail Industry Trends

The retail sector has experienced significant metamorphosis, driven primarily by technological strides and shifts in consumer preferences. Amid these driving forces, one undeniable reality is the intensified intra-industry competition in retail, a stark contrast to the landscape of just a few years ago. Navigating through this heightened competition and addressing associated challenges, companies in the retail sphere are increasingly relying on customer data. This wealth of data serves as a cornerstone for deriving insightful strategies and innovative campaigns. The overarching goal is to not only retain existing customers but also to allure potential ones in a dynamic and fiercely competitive market [13].

In the given ever-evolving landscape of the retail industry, a nuanced understanding of customers and their behaviors is paramount. This was addressed in the National Conference on Foreign Direct Investment in India in 2013. Specifically, the importance of leveraging fact-based data on customer trends across different regions and regular updates to this data as a foundational resource for retailers was established [12]. The dynamic nature of consumer preferences demands a data-driven approach in retail marketing. By harnessing the power of up-to-date and region-specific customer insights, retailers can tailor their strategies, enhance personalized experiences, and stay agile in meeting the evolving demands of their diverse customer base.

Another crucial trend enhancing retailers' comprehension of customers revolves around item association. This phenomenon holds particular significance in the grocery retail market, offering substantial advantages to retailers when executed effectively. By delving into the data-driven intricacies of item associations, retailers can unlock valuable insights into customer preferences and purchasing patterns. This strategic understanding empowers retailers to optimize product placement, tailor marketing strategies, and ultimately elevate the overall shopping experience for their customers [19].

Amidst the backdrop of globalization and the unprecedented expansion of the retail sector, the imperative to provide customer-focused services has become more pronounced than ever. In this dynamic landscape, the application of data-driven techniques, particularly through robust data-mining frameworks and data-based decision-making, takes center stage. The retail sector's ability to harness the power of data analytics becomes a strategic cornerstone for delivering personalized services, understanding global consumer trends, and staying agile in the face of industry shifts [7].

The adoption of data-driven approaches becomes not just a competitive advantage but a necessity to navigate the complexities of the modern retail environment and ensure sustained customer satisfaction. In our research, we tackle the formidable challenges faced by the industry through two primary data-driven methodologies: Descriptive Analytics, aimed at comprehending customers' shopping behaviors, and Customer Clustering, designed to group together customers with similar behavioral patterns.

2.2 Customer Segmentation in Retail

In recent times, customer segmentation in the retail industry has garnered significant attention, fueled by frequent technological advancements. A thorough exploration of existing literature revealed that K-Means Clustering stands out as the predominant approach for customer clustering in retail [2]. This popularity can be attributed to the method's apparent advantages. Notably, the simplicity of K-Means is particularly advantageous, especially when compared to other highly complex methods. Additionally, its computational efficiency becomes crucial as other methods tend to exhibit exponential increases in run-time and memory requirements, particularly when dealing with large datasets [10].

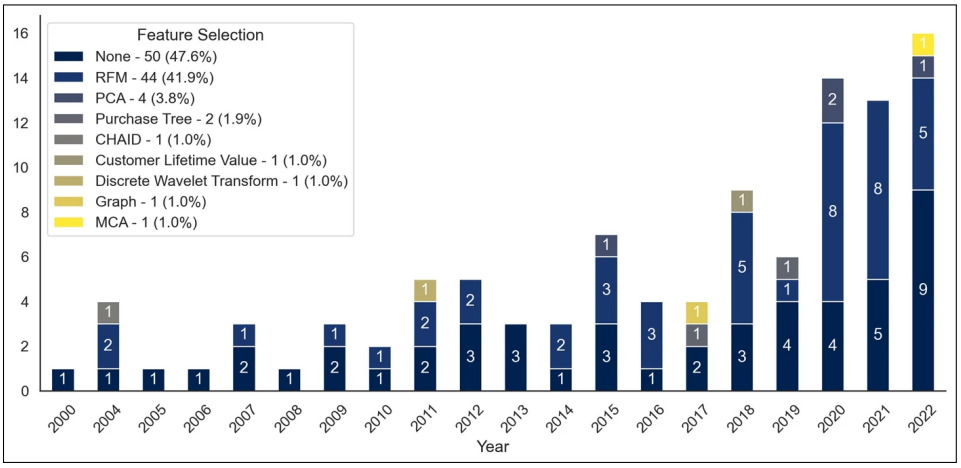


Figure 2.1: Distribution of Surveyed Feature Selection Methods over Years of Publications [2]

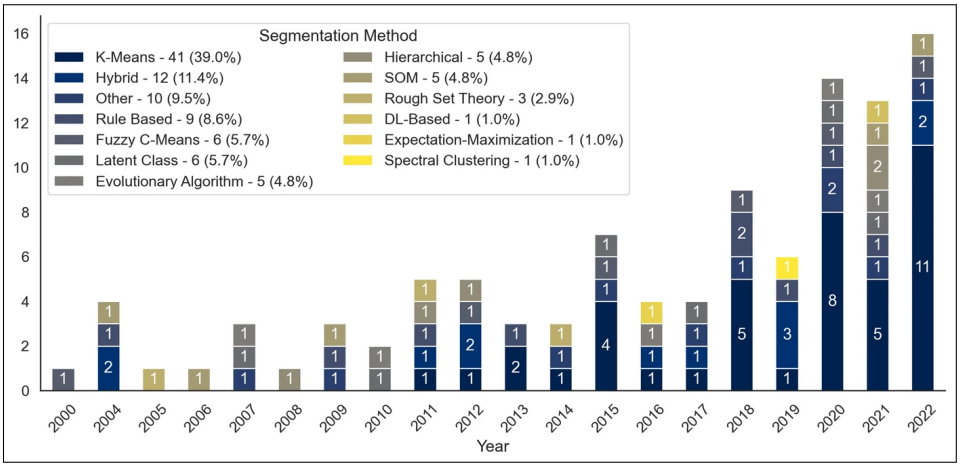


Figure 2.2: Distribution of Surveyed Clustering Methods over Years of Publications [2]

After a thorough examination of the extensive research on Clustering Algorithms and Complexity Analysis, along with a detailed exploration of key findings presented in Figures 2.1 and 2.2, the inclusion of RFM Analysis and K-Means Clustering was deemed essential. This decision was motivated by their strong alignment with the objectives and focus of our research.

In tandem with the fundamental procedures rooted in historical patterns of Customer Segmentation in the retail industry, we also employed additional data-driven approaches, as outlined in the following section, to delineate the characteristics of the datasets and segment them appropriately.

2.3 Data-Driven Approaches

2.3.1 Describing the Statistics

In this section, we discuss the rationale behind opting for multiple statistical approaches employed to achieve comprehensible results in Chapter 4: Exploratory Analysis and Insights.

Chi-Square Test

The chi-square test is a statistical method used to determine if there is a significant association between two categorical variables. It assesses whether the observed distribution of categorical data differs from the distribution that would be expected under the assumption of independence between the variables [9].

Cramér's V

Cramér's V is a statistical measure of association between two categorical variables. It is derived from the chi-square statistic and provides a normalized indication of the strength and direction of the relationship between the variables. Cramér's V ranges from 0 to 1, where 0 signifies no association, and 1 indicates a perfect association. This measure is particularly useful in assessing the degree of dependence between categorical variables, with higher values implying a stronger association [4, 9].

Apriori Algorithm

The Apriori algorithm is a classic and widely used algorithm in data mining and association rule learning. It is specifically designed for discovering frequent itemsets in transactional databases to establish associations between different items. The algorithm was introduced by Rakesh Agrawal and Ramakrishnan Srikant in 1994 and is fundamental for association rule mining [1]. The working methodology of the Apriori algorithm is delineated into two primary steps: identifying frequently occurring itemsets and establishing associations between them. Frequent itemsets are determined based on their support, indicating their frequency within the dataset. Subsequently, association rules are generated using IF-THEN criteria. For instance, a rule $A \rightarrow B$ with 0.70 confidence value implies that if A occurs, then there is a 70% likelihood of B occurring.

Spearman Correlation Coefficient

Considering the nature of our data, exploring potential relationships among the features became imperative. In this context, the Spearman Correlation Coefficient took precedence over the Pearson

Correlation Coefficient. This choice was informed by the recognition that the Pearson Correlation Coefficient assesses linear correlations, whereas our focus was on capturing monotonic relationships, linear and otherwise [4, 9].

2.3.2 Clustering Methods

In this section, we discuss the rationale behind opting for multiple clustering methods employed in Section 3.3 to achieve actionable results in Chapter 5: Data Driven Customer Segmentation.

Recency Frequency Monetary (RFM) Predefined Segmentation

Given the extensive research into RFM Analysis and its demonstrated success in multiple case studies within the retail industry [2, 6, 8], it was deemed an essential component for our research. RFM Analysis centers on three key criteria for characterizing customers: recency, frequency, and monetary value. Recency pertains to how recently a customer made their last visit, frequency indicates the number of visits over a specified period, and monetary value reflects the amount spent during that period. RFM Analysis, along with any subsequent RFM-based clustering, relies solely on these three features to inform decision-making. A particular approach we followed was using the three RFM features in a predefined way to achieve a customer RFM Score and then segment customers based on that score.

K-Means Clustering

The choice to incorporate K-Means Clustering was evident, supported by thorough research in the methodology. It is recognized as one of the most efficient and effective clustering method in the retail industry as backed by Figure 2.2. K-Means Algorithm is an unsupervised machine learning algorithm that works by iteratively assigning data points to clusters based on their similarity to the cluster centroids and then refining those centroids [4, 9]. The "K" in K-Means refers to the number of clusters, which is a value that needs to be provided ahead of time for this algorithm to work. The k-Value decision is consistently based on the Elbow Method and Silhouette Score results throughout our research.

The Elbow Method involved running the K-means clustering algorithm on the dataset for a range of values of k. The sum of squared distances from each point to its assigned cluster centroid is called inertia. This value is calculated for multiple k-values. As k increases, the inertia typically decreases because the data points are closer to their centroids. However, beyond a certain point, the rate of decrease slows down. The "elbow" in the plot of inertia versus k signifies the point where the rate of decrease sharply changes. The optimal k is often considered to be at this elbow, as it indicates a balance between clustering accuracy and simplicity [4, 9].

Silhouette Scores provide a measure of how well-separated the clusters are and ranges from -1 to 1, with close to 1 indicating well-defined clusters and close to -1 indicating highly overlapping clusters. A high silhouette score indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. The optimal number of clusters is often associated with the maximum average silhouette score across different values of k. Throughout our research, we chose clustering based on the first or the second highest silhouette score [4].

Density-Based Spatial Clustering of Applications with Noise

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a widely used clustering algorithm in machine learning and data analysis. Unlike traditional clustering methods such as K-Means, DBSCAN doesn't require the specification of the number of clusters beforehand. Instead, it identifies dense regions in the data, separating them from less dense areas, and can even classify data points as noise if they do not belong to any cluster. This flexibility makes DBSCAN particularly suitable for datasets with irregular shapes and varying cluster densities. It operates based on the concept of density, where clusters are formed by regions of high data point concentration, separated by areas of lower point density [5].

Chapter 3

Methodology

3.1 Data Collection

The foundation of this research rests upon a comprehensive dataset generously provided by a prominent entity in the realm of Computer Science and Research within the retail industry. This dataset encompasses four primary sections, each offering distinctive insights into the dynamics of transactions at a specific grocery store's two locations during the month of March 2023.

3.1.1 Sector Information

The first segment of the dataset revolves around sector details, furnishing essential information for different business divisions. Comprising 21 rows and 2 columns, this dataset includes sector ID numbers and corresponding sector names. These categorical identifiers delineate the diverse areas in which items are retailed, forming a fundamental basis for the subsequent analysis.

3.1.2 Detailed Item Information

The second dataset delves into a granular portrayal of the items vended at the grocery store. With 360 rows and 4 columns, this dataset provides intricate details, including categorical item ID numbers, scan-codes, item names, and their associated sectors. This detailed itemization lays the groundwork for a nuanced examination of consumer purchasing patterns and preferences.

3.1.3 Card-Holders' Information

The third dataset encompasses crucial information about cardholders, presenting a holistic view of the consumers engaged in transactions at locations 1 and 2. Boasting 7,950 rows, 4 columns and 2,600 rows, 4 columns respectively for locations 1 and 2, these datasets offer a comprehensive set of details, including card ID numbers, birth years, CAP codes, and gender information of the cardholders. This demographic information forms a crucial aspect of the analysis, allowing for a deeper understanding of consumer behavior.

3.1.4 Transaction Records

The cornerstone of the dataset is the comprehensive transaction record, capturing a wealth of information related to all transactions conducted in the distinct locations of the specified grocery store

during March 2023. These pivotal datasets comprise about 374,200 rows, 8 columns and 143,600 rows, 8 columns respectively for locations 1 and 2. The columns encapsulate vital transaction details, including date and time stamps, cash register identification, transaction counts at specific registers on particular days, categorical item IDs, quantities of items purchased, promotional item counts, total expenditure, and cardholder ID numbers. These datasets serve as the bedrock for the subsequent analyses, providing a detailed perspective on the transactional landscape within the grocery stores.

The selection of this multifaceted dataset was driven by the imperative to conduct a comprehensive analysis of customer clustering in the retail market, with a particular focus on a grocery store's transactions during March 2023. The inclusion of sector information, detailed item specifics, card-holders' demographics, and transaction records serves a strategic purpose. The sector data provides context by categorizing items into distinct business divisions, allowing for a more nuanced examination of customer preferences across various sectors. Detailed item information enriches the analysis by offering insights into specific products, their categorization, and potential correlations with consumer behavior. Card-holders' information contributes a demographic layer, enabling a more profound understanding of purchasing patterns based on age, location, and gender. The transaction records, being the backbone of the dataset, empower the study with a granular view of each transaction, encompassing essential details such as timing, frequency, and expenditure. This judicious selection ensures a holistic exploration of customer clustering dynamics, providing a robust foundation for the ensuing analytical processes.

3.2 Data Preprocessing

The multifaceted nature of the dataset, comprising distinct data frames related to sectors, items, cardholders, and transactions, necessitated a comprehensive approach to data preprocessing. Recognizing the diversity of information and the unique characteristics inherent in each dataset, multiple analyses were conducted, each accompanied by tailored preprocessing steps to ensure data quality and coherence. As each dataset played a crucial role in shaping the insights derived from the study, this section outlines the meticulous preprocessing procedures undertaken to cleanse, integrate, and transform the raw data. The ensuing discussion delves into the intricacies of handling missing values, merging disparate datasets, engineering features, addressing imbalances, and ensuring temporal alignment, providing a transparent account of the steps taken to fortify the datasets for robust analyses.

3.2.1 Data Cleaning

Among the five datasets scrutinized, only the dataset containing cardholders' information exhibited missing values, which were distributed across all columns except the cardholder ID number. Of about 7,950 distinct registered cardholders, who visited location 1 at least once in March 2023, approximately 1,600 lacked birth year data, 300 lacked gender information, and 340 were without CAP (Postal Code) information. Notably, about 300 cardholders were missing all three of these demographic values, while a total of 1,700 cardholders had at least one missing value. Similarly for location 2, out of about 2,600 registered cardholders, 250 lacked birth year data, 130 lacked gender information, and 160 were without CAP information. Overall, 125 cardholders were missing all

three of these demographic values, while a total of 280 cardholders had at least one missing value. Throughout the research process, instances with missing values were judiciously managed, particularly in demographical analyses. IDs associated with missing data were systematically excluded from relevant analyses. Moreover, erroneous data entries, such as ages recorded as 3 years or 123 years for registered cardholders, were meticulously addressed to prevent any undue influence on the research outcomes. While NULL values and illogical entries were effectively handled, the datasets did not encounter issues related to duplicates.

3.2.2 Data Integration

The meticulous process of merging the four distinct data frames derived from the comprehensive dataset was executed with precision to facilitate nuanced analyses. The linchpin of this integration was the association of the main transactional data frame with the cardholders' information using card ID numbers shared between the two datasets. In instances where a customer was not a cardholder, the card ID number in the transactional data was marked as "0". Subsequently, the transactional data frame was seamlessly merged with the detailed item information and sector information data frames, leveraging categorical item ID numbers and sector ID numbers, respectively. Notably, the individual data frames were stored separately for logistical convenience and were selectively merged only when necessitated by specific analyses. This judicious approach to data integration ensured the preservation of dataset integrity and the facilitation of targeted analytical investigations.

3.2.3 Data Transformation

This section encapsulates the transformative steps undertaken to enhance the richness and analysis-ready quality of the integrated datasets. Two pivotal processes, namely feature engineering and normalization/scaling, were meticulously applied to distill meaningful insights from the multidimensional data. Feature engineering involved crafting new variables to augment the existing dataset, fostering a more nuanced understanding of the retail dynamics under scrutiny. Simultaneously, normalization and scaling procedures were implemented to standardize the variable ranges, ensuring equitable contributions to subsequent analyses. The ensuing discussion delves into the specific strategies employed for feature engineering and the rationale behind normalization and scaling, underscoring their collective role in fortifying the datasets for robust and meaningful analyses.

1. **Feature Engineering** emerged as a strategic imperative, introducing nuanced dimensions to the datasets to unveil potential trends within the retail landscape. Leveraging the date-time column in the transactional dataset, multiple new features were artfully crafted throughout the process of analysis. These features encapsulate the temporal context, specifically capturing the information regarding day of the week, and the hour of operation during which each transaction occurred at the grocery store. An additional layer of sophistication was added through the aggregation of transactional data at the transaction level, offering a comprehensive view per transaction instead of per item sold. This aggregation involved summation of values corresponding to items bought in offer, total items purchased, and total expenditure. However, to mitigate potential data loss linked to categorical item ID numbers, supplementary features were engineered using one-hot encoding. A parallel process was executed for sector ID

numbers, documenting the count of items bought within each sector. Notably, a binary feature was introduced to distinguish cardholders from non-cardholders, taking the value "1" for cardholders and "0" for non-cardholders, facilitating streamlined analyses. These orchestrated feature engineering endeavors form a pivotal foundation for subsequent clustering algorithms, detailed comprehensively in the relevant clustering processes.

2. In the pursuit of robust clustering analyses, a judicious application of **Standardization, Normalization, and Scaling techniques** was executed throughout the analytical journey. The imperative to address potential clustering challenges arising from disparate data scales led to the implementation of these techniques. Specifically, when engaging in k-means clustering, renowned for its sensitivity to variable magnitudes, Robust Scaling was meticulously applied. This process rectified disparities in feature scales, accommodating diverse data types such as costs, counts, and binary indicators.

It is worth noting that, in alignment with the principle of preserving the original dataset integrity, both feature engineering and scaling techniques were conducted selectively, responding directly to the analytical demands of the moment. This ensured that each transformation was purposeful and aligned with the specific needs of the analyses conducted, underscoring a commitment to maintaining the authenticity of the data wherever feasible.

3.2.4 Data Reduction

In our approach, aiming to mitigate information loss and considering the manageable size of most datasets, we selectively applied data reduction techniques for the purposes of visualizing our high-dimensional results and deriving results from high-dimensional datasets. Principal Component Analysis (PCA) was strategically employed as needed to achieve our objectives of both comprehension and effective representation. Following the application of various algorithms, PCA played a crucial role in post-algorithm visualization, aiding in the exploration and communication of our results in a more accessible and interpretable manner. PCA also played a crucial role in Section 5.5, Sector Clustering. Through intelligent dimensionality reduction, PCA effectively reduced the features by approximately 70%, while retaining 90% of the variance in the data.

3.2.5 Handling Categorical Variables

This research undertook several critical analyses that necessitated the inclusion of categorical variables, including days of the week, specific hours of operation, item categories, item sectors, and more. To incorporate these categorical variables into the datasets, feature engineering techniques, specifically one-hot encoding, were selectively employed. Additionally, a meticulous approach was taken when dealing with categorical data to ensure the application of statistical techniques most suitable for this type of data.

3.2.6 Temporal Alignment

A crucial aspect of data preprocessing was the careful treatment of timestamps in the transactional data, particularly emphasized in Section 4.5 dealing with daily and hourly analyses. Temporal consistency played a pivotal role in several analyses, necessitating meticulous handling of time-related

information. To maintain uniformity and facilitate precise analysis, timestamps were standardized to the nearest hour, effectively rounding down the transaction time. For instance, a transaction occurring between 11:00:00 and 12:00:00 hours was consistently timestamped as 11:00:00. This meticulous adjustment ensured coherence and consistency across the datasets, laying the foundation for accurate temporal analyses.

3.2.7 Data Quality Assurance

Ensuring data quality is a foundational aspect of any analysis, as the reliability of findings hinges on the integrity of the data. Addressing the question of trustworthiness begins with a robust approach to data quality assurance. In our demographic analyses, meticulous measures were taken to handle outliers, particularly in age-oriented analyses. To enhance accuracy, customers below the age of 5 years and above the age of 100 years were excluded. Notably, in the transactional data, outliers in features such as total items bought and total amount spent were retained, acknowledging their real-life relevance. Instead of outright removal, we employed techniques such as standardization, normalization, and scaling on a case-by-case basis before proceeding with clustering, ensuring a balanced consideration of the data's authenticity and complexity.

3.2.8 Ethical Considerations

Given the sensitivity of the data involved, stringent measures were implemented to safeguard against any inadvertent disclosure of critical identification information. Particularly, in the analysis of transactional data and cardholders' information, the customer identification numbers were meticulously excluded from our analyses. Instead, a binary indicator was employed to represent cardholding status while maintaining privacy. Moreover, during demographic analyses, the inclusion of CAP codes data was strictly avoided in adherence to both data protection and ethical guidelines, ensuring a responsible and compliant approach to handling sensitive information. In the research, most numerical values are also rounded to prevent the possibility of making strict associations with our data.

3.3 Data Driven Methods

This section delves into the intricacies of the approaches undertaken, shedding light on the diverse techniques utilized to discern patterns and groupings within our data. Through a comprehensive examination of these clustering methodologies, we aim to unravel valuable insights that contribute to a nuanced understanding of customer behaviors and preferences in the retail domain.

3.3.1 Pre-Clustering: Leveraging Association Rules for Initial Data Insights

Prior to undertaking segmentation, it is essential to grasp the underlying trends within the data. This preliminary understanding was achieved by applying statistical approaches outlined in Section 2.3: Data Driven Approaches. In this section, we delve into the specific methodologies associated with these approaches.

A glimpse into the application of these approaches in the Python environment is presented to enhance comprehension of the methodology. The Chi-Square Test necessitated the input of a

contingency table for categorical data, while the Apriori algorithm required a Boolean dataset.

Chi-Square Test & Cramér's V:

```
>>> # Importing necessary libraries
>>> from scipy.stats import chi2_contingency
>>> # Performing a chi-square test
>>> chi2, _, _, _ = chi2_contingency(contingency_table)
>>> # Calculating Cramer's V
>>> n = np.sum(contingency_table)
>>> k = contingency_table.shape[1]
>>> r = contingency_table.shape[0]
>>> cramers_v = np.sqrt(chi2 / (n * min(k-1, r-1)))
```

Association Rules using Apriori Algorithm:

```
>>> # Importing necessary libraries
>>> from mlxtend.frequent_patterns import apriori, association_rules
>>> # Choosing Minimum Support (num1) & Minimum Confidence (num2) on a case-basis
>>> # Using the Apriori algorithm to find frequent itemsets
>>> frequent_itemsets = apriori(data, min_support=num1, use_colnames=True)
>>> # Generating association rules
>>> rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=
    num2)
>>> # num1 and num2 were values between 0 and 1
```

3.3.2 K-Means Clustering

One of the preeminent methods employed in customer clustering, valued for its simplicity, efficiency, and efficacy, is K-Means Clustering. Given its widespread use and alignment with previous extensive research in the retail industry, it was deemed crucial to incorporate it into our investigation.

A central challenge associated with K-Means Clustering is determining the optimal number of well-defined and distinct clusters within the dataset under scrutiny. To address this challenge, a combination of the Elbow Method and Silhouette Scores, utilizing Python's Data Science Libraries, was employed. This allowed for a robust evaluation and selection of the most suitable number of clusters for our analysis. Basic implementations of these approaches are provided for better understanding.

Elbow Method:

```
>>> # Importing necessary libraries
>>> from sklearn.cluster import KMeans
>>> # Performing the Elbow Method to figure out optimal value for clusters
>>> inertias = []
>>> for k in range(1, 11):
    >>> kmeans_model = KMeans(n_clusters=k, init='k-means++', random_state=42,
        n_init='auto')
    >>> kmeans_model.fit(data)
    >>> inertias.append(kmeans_model.inertia_)
>>> # Plotting the elbow curve
>>> plt.plot(range(1, 11), inertias, marker='o')
```

Silhouette Score:

```
>>> # Importing necessary libraries
```

```
>>> from scipy.cluster.vq import kmeans, vq
>>> from sklearn.metrics import silhouette_score
>>> # Performing K–Mean Clustering with a generic k value
>>> centroids, distortion = kmeans(data, k)
>>> clusters_k, _ = vq(data, centroids)
>>> # Computing silhouette score
>>> silhouette_avg = silhouette_score(data, clusters_k)
```

The investigation into the clusters persisted following the execution of K-Means Clustering. This phase aimed to discern noteworthy properties and articulate defining characteristics associated with the clusters obtained to validate the results obtained from K-Means.

3.3.3 Density-Based Spatial Clustering of Applications with Noise

We have already established that DBSCAN is a prominent clustering algorithm in machine learning. Here we provide a Python environment snippet of the implementation of DBSCAN and an explanation into the required inputs. The minimum distance between two points for them to be considered in neighborhood of each other is given by "eps" and the minimum number of points to be considered a cluster are determined by "min_samples". These two inputs of "eps" and "min_samples" were consistently 0.5 and 1000 to guarantee the consistency of the algorithm for our datasets. This decision was rooted in our goal of achieving well-categorized clusters with at least 1000 transactions.

DBSCAN:

```
>>> # Importing necessary libraries
>>> from sklearn.cluster import DBSCAN
>>> # Applying DBSCAN
>>> dbscan = DBSCAN(eps=0.5, min_samples=1000)
>>> labels = dbscan.fit_predict(data)
```

Chapter 4

Exploratory Data Analysis and Insights

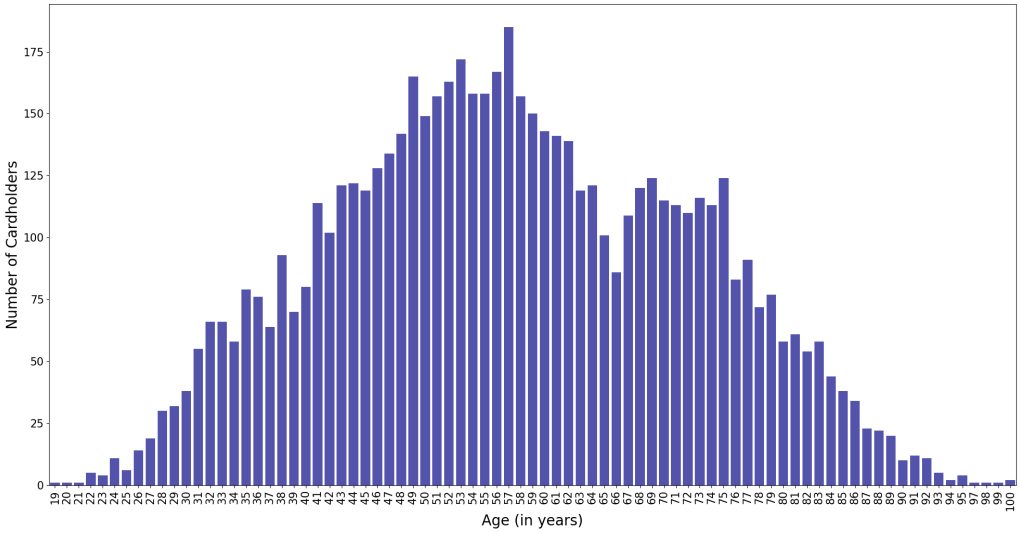
In this pivotal chapter, we unveil the culmination of extensive research and analytical endeavors, presenting the results and insights gleaned from the multifaceted datasets. The focus here is on translating raw data into actionable knowledge, unraveling patterns, and distilling meaningful findings. Through the lens of diverse analytical approaches, including statistical analyses and association techniques, we navigate the intricate landscape of the retail market. Each section within this chapter serves as a gateway to the revelations discovered, offering a comprehensive overview of customer behavior, transactional dynamics, and the interplay of external factors. As we delve into the results, the narrative seeks not only to illuminate the outcomes but also to provide a foundation for informed decision-making within the retail sector. This chapter stands as a testament to the thesis's overarching goal — to empower stakeholders with a deep understanding of the intricate tapestry that defines customer behaviors in the retail market.

We embark on an in-depth exploration of the diverse analytical approaches undertaken, each yielding unique findings. Our journey delves into the intricate groundwork laid to prepare the stage for multifaceted customer clustering. By scrutinizing each approach, we unravel the insights derived from statistical analyses and advanced methodologies, illuminating the nuanced dynamics of the retail landscape. This comprehensive discussion not only showcases the diverse tools in our analytical toolkit but also provides a coherent narrative of the strategic steps taken to uncover patterns and trends at various levels.

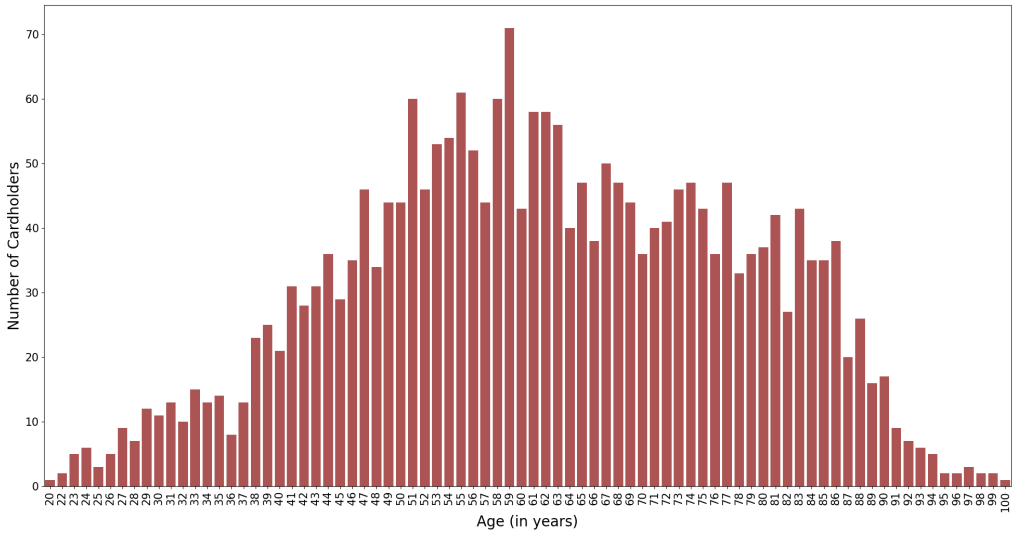
4.1 Demographic Analysis: Unveiling Customer Insights

Our analytical journey commenced with a meticulous examination of the demographic landscape of our customer base. Focused on discerning trends and insights, this initial approach involved a comparative analysis against existing research. Leveraging the cardholders' information dataset—chosen due to the absence of comparable data for non-cardholders—we delved into three key columns: birth year, gender, and CAP. A fundamental aspect of this exploration was the engineering of age within the dataset, accomplished by subtracting the birth year from the current year (2023). Building upon the data cleaning procedures detailed in Section 3.2.1, our analyses unfolded, revealing valuable patterns that were subsequently visualized for a comprehensive understanding of the demographic nuances within our customer base.

The primary demographic among registered cardholders, those who visited location 1 at least



(a) Location 1

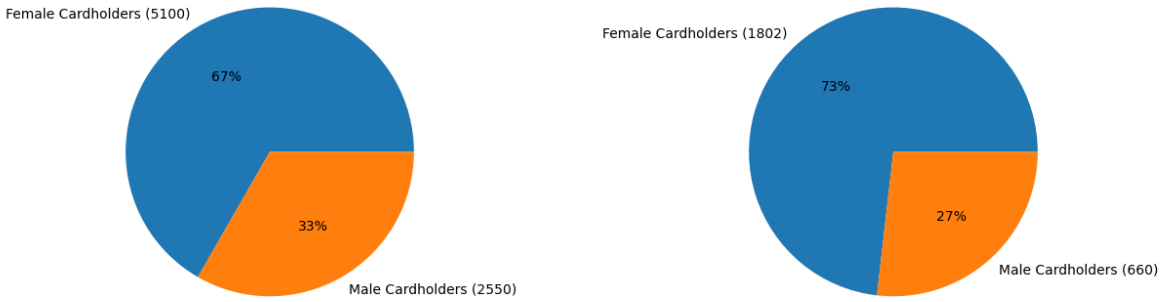


(b) Location 2

Figure 4.1: Age Distribution among the registered Cardholders

once in March 2023, predominantly fell within the age range of 40 to 75 years, constituting approximately 74% of the total. Remarkably, the apex age group was 57 years, representing the highest cardholder count at 187. Further exploration into gender distribution uncovered that the count of female cardholders surpassed that of males by approximately twofold. Transitioning to the analysis of location 2 revealed a distinct demographic pattern. The predominant age range for cardholders who visited location 2 at least once spanned longer from 45 to 85 years, with the peak age group centered around 59 years and a corresponding cardholder count of 71. Intriguingly, the gender distribution at location 2 depicted a notable contrast, with three times as many female cardholders as their male counterparts.

Upon scrutiny of the CAP data, a remarkable revelation unfolded— the majority of registered cardholders at both the locations came from very localized areas. Although excluded from the report



(a) Location 1 **(b)** Location 2

Figure 4.2: Gender Split among the registered Cardholders

for privacy concerns, the analyses and the visualizations of the CAP information underscored the profound clustering of cardholders within specific zones, shedding light on the localized nature of customer engagement and the gradual decrease in density of cardholders on moving away from the locations.

4.2 Transactional Analysis

In this segment, we aim to analyze disparities in spending habits among diverse customers, shedding light on distinctions between cardholders and non-cardholders, as well as their utilization of promotional offers. The dataset under examination pertains to our primary dataset encompassing transactional data of about 45,700 and 19,800 distinct transactions from the locations 1 and 2 of the grocery store respectively throughout March 2023.

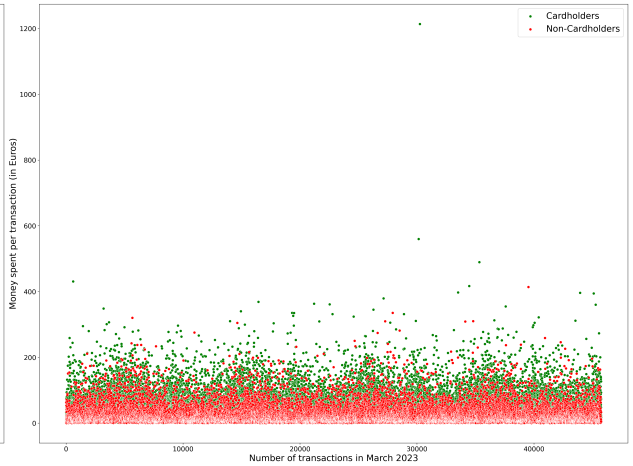
	Location 1		**Location 2**	
Category:	Cardholders	Non-Cardholders	Cardholders	Non-Cardholders
Items bought on Offer	4.3	2.1	3.8	1.6
Total Items bought	19	11	13	8
Total Amount Spent	45.00	26.00	28.00	17.80
% Items on Offer	23%	19%	30%	19%

Table 4.1: Per-Transaction Average Values Comparison Chart for Location 1 & 2

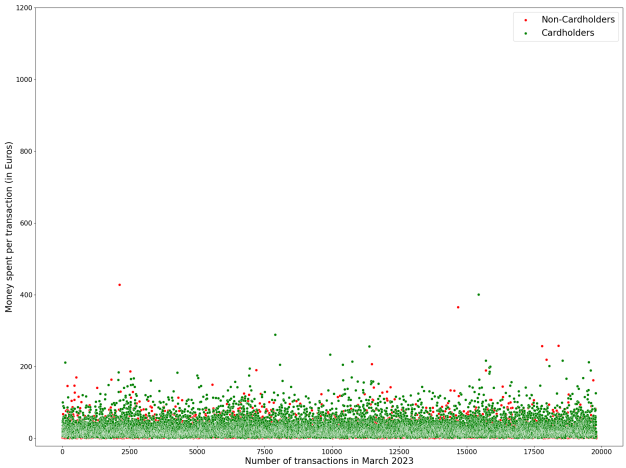
The analysis of Table 4.1 reveals a conspicuous contrast in engagement levels between cardholders and non-cardholders at both locations. At location 1, cardholders demonstrated a higher level of involvement, purchasing an average of approximately 2 more items on offer and 8 more total items per shopping trip compared to their non-cardholding counterparts. Moreover, when examining the percentage of items bought on offer per shopping trip separately for cardholders and non-cardholders, cardholders exhibited a 4% higher rate. The most notable difference emerged in spending behavior, where cardholders consistently outspent non-cardholders by an average of about 19.00 Euros per shopping trip.

Similarly, at location 2, cardholders showcased a heightened level of engagement, acquiring approximately 2 more items on offer and 5 more total items per shopping trip compared to non-cardholders. Additionally, the percentage of items bought on offer per shopping trip, when averaged separately for cardholders and non-cardholders, reflected a notably higher rate for cardholders at 11%, a significant increase compared to location 1. Cardholders at this location also outspent their non-cardholding counterparts by an average of about 10.20 Euros per shopping trip.

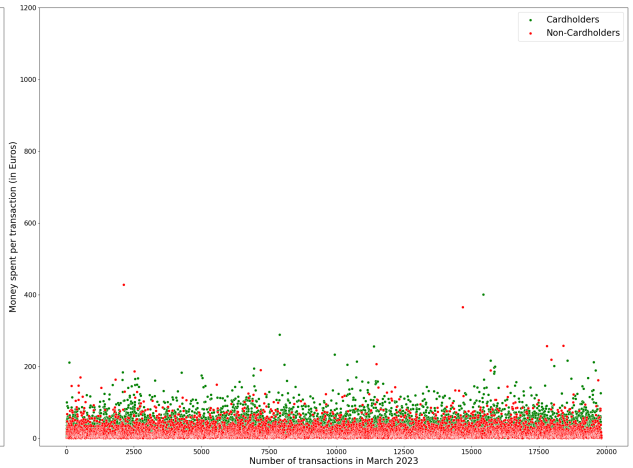
(a) Location 1: Cardholders over Non-Cardholders



(b) Location 1: Non-Cardholders over Cardholders



(c) Location 2: Cardholders over Non-Cardholders



(d) Location 2: Non-Cardholders over Cardholders

Figure 4.3: Transactional Expenditure of Customers

The assertion that cardholders tend to spend more than non-cardholders finds robust support in Figure 4.3. A clear pattern emerges in the plots in Figure 4.3 (a) and (c) representing cardholders' spending plotted following those of non-cardholders. The plots are uniformly green, vividly illustrating that the expenditures of cardholders consistently surpass those of their non-cardholding counterparts. Notably, this pattern is reiterated by the green top linings in the plots in Figure 4.3 (b) and (d), emphasizing the dominance of cardholders' spending when they are plotted before those of non-cardholders. These compelling visualizations and the analysis of Table 4.1 enforce the notion that loyalty card ownership significantly influences both the quantity and financial dimensions of shopping behavior.

Moreover, our analytical findings were substantiated through correlation analysis, where we

explored the relationship between a binary cardholding indicator and three key features: items bought on offer, total items bought, and total amount spent. The associated p-values for all six correlations (three for each location) were uniformly 0.00, decisively rejecting the null hypothesis that posits no correlation. This statistical evidence underscores the robustness of our findings. For location 1, Spearman correlation coefficients of 0.28, 0.31, and 0.31 were observed, while for location 2, the coefficients were 0.36, 0.31, and 0.31, respectively. These coefficients illuminate the strength and directionality of the relationships, affirming the significant impact of cardholding status on the quantities and financial aspects of shopping behavior at both locations, specifically that cardholders tend to buy more items on offer at location 2.

4.3 Item Sector Analysis

This section delves into the analyses conducted regarding the classification of items within various sectors. As established in Section 3.1.1, our dataset comprises 21 distinct sectors, and further exploration in Section 3.1.2 reveals the existence of 360 unique item categories. Notably, a detailed examination of the item information dataset exposes an interesting finding: none of the items fall under Sector 14, titled "Insert Manually." The complete breakdown of item classifications across sectors is presented in Table 4.2.

Sector ID	Sector Name	Number of Items
1	Beverages	37
2	Self-Service Counter	25
3	Non-Food Items	32
4	Food Items	56
5	Confectionery - 1st Breakfast	39
6	Butcher Shop	10
7	Bakery	1
8	House Cleaning	58
9	Personal Hygiene	48
10	Christmas Celebration	11
11	Celebrations	1
12	Easter Celebration	4
13	Fruits & Vegetables	4
14	Insert Manually	0
15	Delicatessen Cutting Counter	6
16	Frozen Foods	7
17	Fish Shop	3
18	Childhood	7
19	Petfood	6
20	Cured Meat Cutting Counter	2
21	Cheese Cutting Counter	3

Table 4.2: Count of distinct items classified under different Sectors

One pivotal aspect of our analysis revolved around investigating potential correlations among the total number of items purchased in various sectors at both locations. This exploration extended to examining relationships with other major influencing factors, all scrutinized on a per-transaction

basis. Before delving into the intricate details, we begin by discussing fundamental sectoral statistics and drawing inferences to foster a deeper understanding of expectations.

		Items Bought			Amount Spent	
ID	**Sector Name**	**On Offer**	**Total**	**% Offer**	**Total**	**% Total**
1	Beverages	20,400	182,000	11.22	442,600	33.88
2	Self-Service Counter	29,300	90,600	32.41	198,600	15.20
4	Food Items	19,100	83,300	23.00	174,000	13.32
5	Confectionery - 1st Breakfast	14,000	63,000	22.26	149,900	11.48
13	Fruits & Vegetables	8,500	38,300	22.28	86,800	6.65
16	Frozen Foods	5,800	14,200	41.32	49,500	3.79
8	House Cleaning	4,500	17,400	26.10	45,000	3.44
9	Personal Hygiene	3,300	14,000	23.63	40,300	3.09
12	Easter Celebration	480	4,200	11.32	34,600	2.65
6	Butcher Shop	2,200	7,200	30.51	29,500	2.26

Table 4.3: Location 1: Cardholders Top 10 Sectoral Expenditure

		Items Bought			Amount Spent	
ID	**Sector Name**	**On Offer**	**Total**	**% Offer**	**Total**	**% Total**
1	Beverages	5,900	63,900	9.26	147,900	33.81
4	Food Items	4,500	27,000	16.63	60,200	13.76
2	Self-Service Counter	6,600	25,500	26.16	58,600	13.41
5	Confectionery - 1st Breakfast	3,900	22,800	17.43	55,200	12.63
13	Fruits & Vegetables	2,300	12,300	19.34	28,200	6.46
9	Personal Hygiene	800	4,900	16.45	15,200	3.48
8	House Cleaning	800	5,100	15.83	14,400	3.29
16	Frozen Foods	1,000	3,600	29.56	13,600	3.13
12	Easter Celebration	180	1,500	11.57	12,400	2.84
6	Butcher Shop	690	2,500	26.62	11,200	2.56

Table 4.4: Location 1: Non-Cardholders Top 10 Sectoral Expenditure

Tables 4.3 and 4.4 delineate the top ten sectors where both cardholders and non-cardholders made the most significant expenditures at location 1. Although the rankings exhibit some variations, these differences are not substantial enough to warrant further analysis. Notably, Table 4.3 consistently reveals higher figures for cardholders compared to non-cardholders in Table 4.4. This aligns with the overarching trend observed, indicating that cardholders, on both total and per-transactional bases, tended to purchase and spend more than their non-cardholding counterparts.

In conducting a comparative analysis of sectoral spending between the two locations, a discernible trend emerged. At location 2, cardholders exhibited an increased expenditure in Sector 20, Cured Meat Cutting Counter, and Sector 21, Cheese Cutting Counter, as opposed to their preferences at location 1, which included Sector 9, Personal Hygiene, and Sector 12, Easter Celebration. Conversely, non-cardholders exhibited only one notable variance; at location 2, there was a higher sales volume in Sector 20, Cheese Cutting Counter, compared to their preference for Sector 12, Easter Celebration, at location 1.

As we transition completely to location 2, the prevailing pattern observed in Tables 4.5 and 4.6

reveals that the numbers associated with cardholders surpass those of non-cardholders, consistent with the trend observed at location 1. This reaffirms the inclination of cardholders to engage in higher levels of purchasing and expenditure compared to their non-cardholding counterparts, even at location 2. Upon closer scrutiny, a noteworthy shift emerges: among the top 10 sectors for cardholders in Table 4.5, Sector 21, Cheese Cutting Counter, supplants Sector 9, Personal Hygiene, from the top 10 sectors for non-cardholders in Table 4.6, marking a distinct divergence in consumption preferences.

		Items Bought			Amount Spent	
ID	**Sector Name**	**On Offer**	**Total**	**% Offer**	**Total**	**% Total**
1	Beverages	5,600	31,900	17.55	67,100	19.41
4	Food Items	8,100	27,700	29.38	55,700	16.12
2	Self-Service Counter	9,700	25,500	38.26	48,400	13.98
5	Confectionery - 1st Breakfast	6,200	21,800	28.34	47,800	13.81
13	Fruits & Vegetables	3,300	9,100	36.49	17,800	5.16
20	Cured Meat Cutting Counter	3,500	5,100	68.86	17,100	4.96
16	Frozen Foods	3,100	5,400	57.35	16,400	4.76
8	House Cleaning	2,200	6,400	34.01	15,400	4.45
21	Cheese Cutting Counter	1,500	2,900	50.66	14,900	4.30
6	Butcher Shop	1,100	2,800	40.17	11,000	3.18

Table 4.5: Location 2: Cardholders Top 10 Sectoral Expenditure

		Items Bought			Amount Spent	
ID	**Sector Name**	**On Offer**	**Total**	**% Offer**	**Total**	**% Total**
1	Beverages	1,600	14,000	11.94	28,300	20.90
4	Food Items	1,700	10,200	17.41	22,200	16.41
5	Confectionery - 1st Breakfast	1,600	8,800	18.75	20,800	15.37
2	Self-Service Counter	2,600	8,600	30.50	17,400	12.91
13	Fruits & Vegetables	1,100	3,800	28.19	7,800	5.77
8	House Cleaning	340	2,200	15.57	5,700	4.19
20	Cured Meat Cutting Counter	940	1,600	58.39	5,500	4.04
16	Frozen Foods	680	1,500	43.13	5,300	3.92
9	Personal Hygiene	410	1,800	22.75	4,700	3.52
6	Butcher Shop	370	1,000	36.32	4,200	3.09

Table 4.6: Location 2: Non-Cardholders Top 10 Sectoral Expenditure

Given the industrial imperative to optimize promotions and offers, particularly tailored to the diverse customer base of the grocery store under examination, a crucial facet of our analysis involved sectoral categorization of offers. This categorization was performed based on cardholding statuses and overall purchasing behavior. The outcomes of this categorization are meticulously presented in Tables 4.7 and 4.8 for locations 1 and 2, respectively.

Notably, across both locations, cardholders consistently exhibited a higher average percentage of items bought on offer per sector compared to non-cardholders and the overall customer base (including both cardholders and non-cardholders). However, a notable anomaly surfaced in Sector 3, Non-Food Items, at the first location. Here, non-cardholders displayed an exceptionally high

percentage of items bought on offer, reaching 50.7%, while cardholders lagged behind at 32.4%. This divergence in behavior underscores the importance of sectoral analysis in tailoring promotional strategies to different customer segments.

ID	Sector Name	Cardholders	Non-Cardholders	All Customers
1	Beverages	11.22	9.26	10.71
2	Self-Service Counter	32.41	26.16	31.03
3	Non-Food Items	35.93	50.71	40.20
4	Food Items	23.00	16.63	21.44
5	Confectionery - 1st Breakfast	22.26	17.43	20.97
6	Butcher Shop	30.51	26.62	29.48
8	House Cleaning	26.10	15.83	23.76
9	Personal Hygiene	23.63	16.45	21.75
10	Christmas Celebration	0.00	0.00	0.00
11	Celebrations	0.00	0.00	0.00
12	Easter Celebration	11.32	11.57	11.38
13	Fruits & Vegetables	22.28	19.34	21.56
15	Delicatessen Cutting Counter	0.00	0.00	0.00
16	Frozen Foods	41.32	29.56	38.92
17	Fish Shop	0.00	0.00	0.00
18	Childhood	0.51	0.00	0.36
19	Petfood	13.66	11.81	13.17
20	Cured Meat Cutting Counter	5.41	5.49	5.43
21	Cheese Cutting Counter	23.18	18.49	22.14

Table 4.7: Percentage of items bought on offer in different sectors at Location 1

ID	Sector Name	Cardholders	Non-Cardholders	All Customers
1	Beverages	17.55	11.94	15.83
2	Self-Service Counter	38.26	30.50	36.30
3	Non-Food Items	0.03	0.00	0.02
4	Food Items	29.38	17.41	26.14
5	Confectionery - 1st Breakfast	28.34	18.75	25.57
6	Butcher Shop	40.17	36.32	39.15
7	Bakery	0.00	0.00	0.00
8	House Cleaning	34.01	15.57	29.30
9	Personal Hygiene	37.58	22.75	33.40
11	Celebrations	0.00	0.00	0.00
12	Easter Celebration	23.16	11.71	20.04
13	Fruits & Vegetables	36.49	28.19	34.04
15	Delicatessen Cutting Counter	15.83	10.00	14.36
16	Frozen Foods	57.35	43.13	54.14
18	Childhood	0.00	0.00	0.00
19	Petfood	14.71	10.10	13.47
20	Cured Meat Cutting Counter	68.86	58.39	66.36
21	Cheese Cutting Counter	50.66	42.31	49.06

Table 4.8: Percentage of items bought on offer in different sectors at Location 2

An intriguing observation is the absence of data in Sectors 7 and 14 for Location 1, and Sectors

10, 14, and 17 for Location 2 in Tables 4.7 and 4.8. This anomaly can be attributed to the absence of shopping activity in these specific sectors or the absence of the sectors altogether at their respective locations, a fact further substantiated by the corresponding heat maps in Figure 4.4.

Our existing datasets already incorporated crucial columns detailing items bought on offer, the overall count of items purchased, the total amount spent, a binary indicator denoting cardholders, and twenty additional feature-engineered columns as outlined in Section 3.2.3. These additional columns elucidated the distribution of items across various sectors. Furthermore, two more columns were introduced to capture categorical and sectoral variety, indicating the count of distinct item categories and the number of sectors visited during each transaction, respectively.

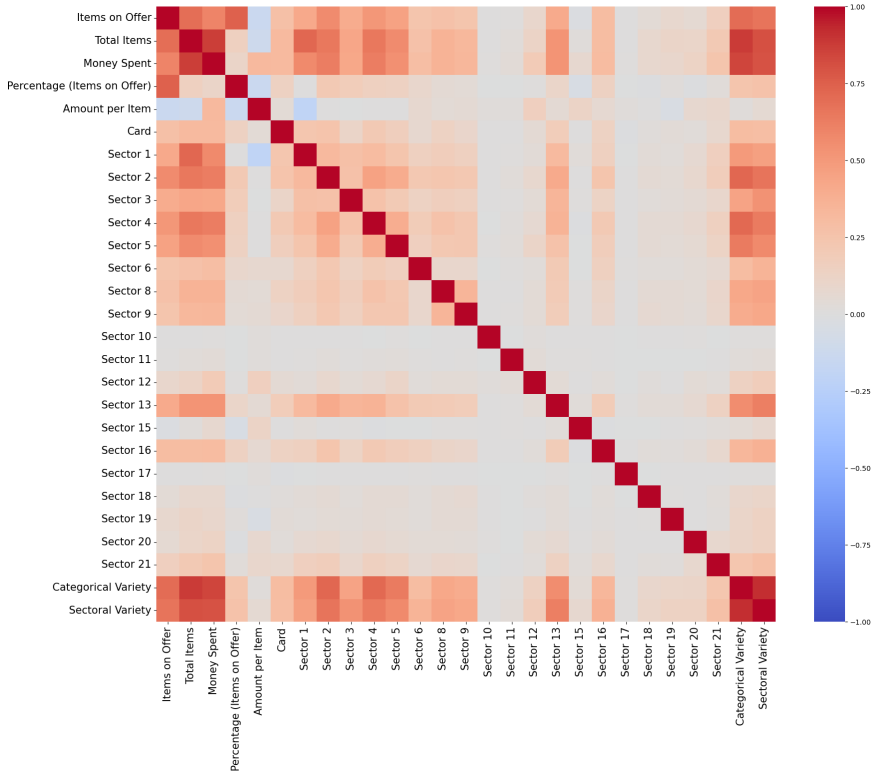
Conducting an inter-feature correlation analysis allowed us to uncover trends within sectors and compare them against essential transactional features. The heat maps, providing basic overview of such analyses, are presented in Figure 4.4 for both the locations and significant correlation coefficients are discussed further in detail.

Expanding on the insights derived from Figure 4.4, we further delve into significant Spearman correlation coefficient values, specifically those equal to or greater than 0.60 at location 1. This exploration is encapsulated in Table 4.9. This table aims to illuminate the sectors that wield substantial influence in driving sales, promoting items, and encapsulating diverse features within a customer's shopping basket. The emphasis is on providing a nuanced understanding of the pivotal sectors contributing to the dynamics of customer engagement at each location.

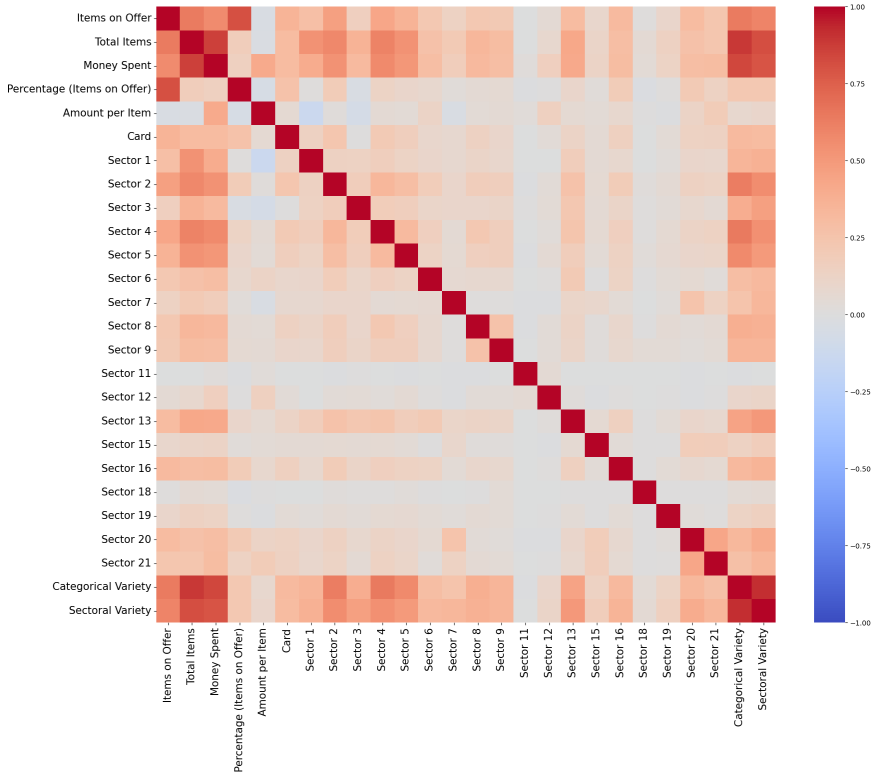
Feature 1	**Feature 2**	**Location 1**	**Location 2**
Sectoral Variety	Categorical Variety	0.93	0.93
Categorical Variety	Total Items	0.88	0.89
Money Spent	Total Items	0.87	0.87
Categorical Variety	Money Spent	0.86	0.86
Sectoral Variety	Total Items	0.81	0.82
Sectoral Variety	Money Spent	0.80	0.79
Percentage (Items on Offer)	Items on Offer	0.75	0.81
Sector 1	Total Items	0.73	0.53
Categorical Variety	Sector 2	0.73	0.63
Categorical Variety	Sector 4	0.71	0.64
Categorical Variety	Items on Offer	0.70	0.64
Total Items	Items on Offer	0.69	0.64
Sectoral Variety	Sector 2	0.67	0.56
Sectoral Variety	Items on Offer	0.67	0.60
Sector 2	Total Items	0.66	0.58
Sector 4	Total Items	0.65	0.61
Sectoral Variety	Sector 4	0.63	0.55
Sector 2	Money Spent	0.63	0.54
Sector 4	Money Spent	0.63	0.58
Categorical Variety	Sector 5	0.63	0.57
Sectoral Variety	Sector 13	0.62	0.50
Money Spent	Items on Offer	0.60	0.57

Table 4.9: Significant Spearman Correlation Coefficient Values from Sectoral Analysis

Table 4.9, displaying the coefficients ranging from 0.60 to 0.93, unveils substantial relationships between various features, offering key insights into the interconnected dynamics of the retail



(a) Location 1



(b) Location 2

Figure 4.4: Heat Maps depicting Sectoral Correlation Coefficient Values

landscape.

Notably, the exceptionally high coefficients of 0.93 for Sectoral Variety with both Categorical

Variety and Total Items at both locations signify a robust positive correlation, indicating that as the variety within a sector expands, so does the diversity in categorical items and the total number of items purchased. The consistent positive correlations across multiple features, such as Categorical Variety with Total Items and Money Spent, underscore the intricate relationships shaping customer behavior.

Specifically, validating our Sector-based analysis, the correlations involving specific sectors (e.g., Sector 1, Sector 2, Sector 4) shed light on the sectors that play pivotal roles in influencing total items purchased, money spent, and the prevalence of items on offer. For instance, the positive correlations between Categorical Variety in Sector 2 and Sector 4 with various features suggest that these sectors significantly contribute to the overall composition of customers' shopping baskets.

Building further on results from Figure 4.4, an analysis regarding association and correlation between sectors and cardholding status was performed. Association results with minimum support of 0.4 and minimum confidence of 0.7 (Location 1) and 0.65 (Location 2) are presented in Tables 4.10 and 4.11 respectively for locations 1 and 2.

Antecedent	**Consequent**	**Confidence**	**Lift**	**Support**
Card	Sector 1	0.88	1.05	0.56
Sector 2	Card	0.73	1.15	0.42
Sector 4	Card	0.71	1.12	0.43
Sector 2	Sector 1	0.89	1.07	0.52
Sector 3	Sector 1	0.88	1.05	0.50
Sector 4	Sector 1	0.88	1.05	0.54
Sector 5	Sector 1	0.87	1.03	0.49
Sector 2	Sector 4	0.75	1.23	0.44
Sector 4	Sector 2	0.71	1.23	0.44
Sector 5	Sector 2	0.70	1.21	0.40
Sector 5	Sector 4	0.73	1.20	0.42

Table 4.10: Association Rules: Key Relationships between Cardholders and Sectors at Location 1

Antecedent	**Consequent**	**Confidence**	**Lift**	**Support**
Card	Sector 1	0.69	1.08	0.42
Sector 1	Card	0.66	1.08	0.42
Card	Sector 2	0.66	1.15	0.41
Sector 2	Card	0.71	1.15	0.41
Card	Sector 4	0.68	1.12	0.42
Sector 4	Card	0.69	1.12	0.42
Sector 4	Sector 1	0.68	1.06	0.41
Sector 4	Sector 2	0.67	1.17	0.41
Sector 2	Sector 4	0.71	1.17	0.41

Table 4.11: Association Rules: Key Relationships between Cardholders and Sectors at Location 2

Tables 4.10 and 4.11 outline key association rules revealing meaningful relationships between cardholders and specific sectors at Locations 1 and 2 respectively. These rules, generated through analysis, provide valuable insights into the patterns and dependencies within customer behavior:

1. **Card and Sector Associations**: The high confidence values of 0.88 at location 1 and

0.69 at location 2 indicate a robust association between cardholders and Sector 1, suggesting that customers holding loyalty cards are 88% and 69% likely to engage with products from Sector 1 respectively at locations 1 and 2. Similarly, the confidence values for Sector 2 and Sector 4 further underscore the influence of cardholders in these sectors. The associations between cardholding status and specific sectors, notably Sector 1, Sector 2, and Sector 4, gain additional support through examination of their Spearman correlation coefficient values. This support is presented in Tables 4.12 and 4.13 for locations 1 and 2, respectively. While the correlation coefficient values indicate a relatively weak correlation individually, their alignment with the association rules offers substantial support. The combined insights from both analyses suggest that cardholders exhibit significant engagement with specific sectors. This synergy between correlation coefficients and association rules strengthens the inference of meaningful associations between cardholding status and targeted sectors, providing valuable insights for strategic decision-making in marketing and customer engagement strategies.

	Correlation Coefficient
Sector 2	0.25
Sector 1	0.24
Sector 4	0.21
Sector 13	0.18
Sector 5	0.17

Table 4.12: Top 5 Spearman Correlation Coefficient Values of Sectors against Cardholding status at Location 1

	Correlation Coefficient
Sector 2	0.24
Sector 4	0.21
Sector 5	0.17
Sector 16	0.15
Sector 1	0.14

Table 4.13: Top 5 Spearman Correlation Coefficient Values of Sectors against Cardholding status at Location 2

2. **Inter-Sector Relationships**: The association rules unveil cross-sector dependencies, exemplified by the high confidence values between Sector 2 and Sector 1, Sector 3 and Sector 1, Sector 4 and Sector 1, and Sector 5 and Sector 1 at location 1, and Sector 4 and Sector 1, Sector 4 and Sector 2, and Sector 2 and Sector 4 at location 2. These relationships shed light on the interconnected preferences of customers, emphasizing the potential influence of products in one sector on those in another.

3. **Lift and Support Metrics**: The lift values greater than 1 (e.g., 1.05 and 1.08 for Card and Sector 1 respectively at locations 1 and 2) signify that the antecedent (Card) has a positive impact on the consequent (Sector 1) compared to random chance. The support values reflect the proportion of transactions supporting these associations, offering an understanding of the prevalence of these patterns within the dataset.

These association rules provide actionable insights for targeted marketing strategies, allowing the grocery store to tailor promotions, optimize inventory placement, and enhance customer engagement by leveraging the identified dependencies between cardholders and specific sectors at different locations.

4.4 Item Category Analysis

This section explores the analyses conducted on the classification of items within different categories. As outlined in Section 3.1.2, our dataset encompasses 360 unique item categories. However, upon a detailed examination of the transactional dataset, a noteworthy observation emerges: at location 1, only 303 distinct item categories were shopped, while at location 2, only 286 distinct item categories were shopped. This finding underscores the variability in customer shopping patterns between the two locations and sets the stage for a more nuanced exploration of category-specific behaviors.

A crucial facet of our analysis centered on exploring potential correlations among the total number of items purchased in various categories at both locations. This investigation extended to scrutinizing relationships with other influential factors, all analyzed on a per-transaction basis. Prior to delving into intricate details, we initiate the discussion by examining fundamental categorical statistics and drawing inferences to cultivate a deeper understanding of expectations.

		Items Bought			Amount Spent	
Sector ID	**Category Name**	**On Offer**	**Total**	**% Offer**	**Total**	**% Total**
Unknown	Uncensored	10,700	73,800	14.47	329,700	25.24
13	Vegetables	8,400	33,400	25.11	72,400	5.54
5	Biscuits	4,800	17,900	26.83	39,900	3.06
1	Wine	970	7,300	13.28	32,200	2.47
1	Water	2,000	71,400	2.88	31,500	2.41
2	Pre-packaged Cheeses	3,800	10,300	37.08	24,500	1.88
12	Easter Egg	260	2,300	11.39	24,500	1.88
2	Fresh Milk & Cream	20	10,900	0.18	22,200	1.70
2	Mozzarella Cheeses	4,900	8,200	59.74	18,700	1.43
16	Frozen Fish	1,700	3,700	47.92	18,500	1.42

Table 4.14: Location 1: Cardholders Top 10 Categorical Expenditure

		Items Bought			Amount Spent	
Sector ID	**Category Name**	**On Offer**	**Total**	**% Offer**	**Total**	**% Total**
Unknown	Uncensored	2,600	23,600	11.28	99,000	22.63
13	Vegetables	2,400	10,900	21.74	23,900	5.46
5	Biscuits	1,200	6,500	19.10	14,700	3.35
1	Wine	280	2,600	10.37	14,000	3.20
1	Water	650	25,800	2.50	11,600	2.65
12	Easter Egg	80	800	9.27	8,500	1.93
2	Pre-packaged Cheeses	920	2,900	31.47	7,000	1.61
6	Packaged Meat	230	1,200	18.04	5,800	1.34
1	Normal Beer	840	3,500	23.62	5,800	1.33
2	Fresh Milk & Cream	1	2,800	0.04	5,800	1.32

Table 4.15: Location 1: Non-Cardholders Top 10 Categorical Expenditure

Tables 4.14 and 4.15 reveal a clear pattern in the spending habits of cardholders and non-cardholders at location 1. The majority of their expenditures align in the same eight categories, with two distinct categories setting them apart: Mozzarella Cheeses and Frozen Fish for card-

holders, and Packaged Meat and Normal Beer for non-cardholders. Notably, Table 4.14 exhibits consistently higher figures, indicative of cardholders consistently spending more than their non-cardholding counterparts. This aligns with our overarching observation that, on both an overall and per-transaction basis, cardholders tend to spend more. The key inference drawn is that both cardholders and non-cardholders exhibit a consistent preference for specific categories, emphasizing the sustained significance of certain items in their shopping baskets.

Upon comparing the top categorical expenditure tables for both locations, a notable distinction emerges: the category of "Easter Egg" consistently appearing at location 1 is replaced by the categories of "Raw & Cut Cured Meats" and "Cooked & Cut Cured Meats" at location 2. However, in general, overall categorical differences between the two locations are not markedly evident, which may limit the validation of further analyses based on these distinctions. This observation was reinforced by conducting a Spearman correlation analysis on categorical data between the two locations. The correlation coefficients consistently surpassed 0.9 across the board, indicating a high degree of correlation. This suggests that the categories that exhibited high shopping frequencies at location 1 were also the same ones that showed high shopping frequencies at location 2.

		Items Bought			Amount Spent	
Sector ID	**Category Name**	**On Offer**	**Total**	**% Offer**	**Total**	**% Total**
Unknown	Uncensored	2,200	9,200	23.53	41,600	12.03
13	Vegetables	3,100	8,100	38.79	15,300	4.43
5	Biscuits	2,000	5,700	35.67	11,500	3.32
20	Raw & Cut Cured Meats	1,900	2,800	67.06	10,700	3.07
21	Seasoned Cheeses Counter	650	1,400	45.70	9,800	2.82
7	Scaled Bread Items	0	5,000	0.00	8,100	2.34
2	Fresh Milk & Cream	20	4,100	0.41	8,000	2.33
20	Cooked & Cut Cured Meats	1,600	2,300	71.06	6,500	1.88
1	Wine	380	1,800	21.29	6,500	1.88
16	Frozen Fish	890	1,400	63.20	6,500	1.88

Table 4.16: Location 2: Cardholders Top 10 Categorical Expenditure

		Items Bought			Amount Spent	
Sector ID	**Category Name**	**On Offer**	**Total**	**% Offer**	**Total**	**% Total**
Unknown	Uncensored	530	3,400	15.56	14,500	10.69
13	Vegetables	1,000	3,400	29.66	6,800	5.05
5	Biscuits	510	2,300	22.50	5,100	3.75
1	Wine	150	800	19.25	3,600	2.67
7	Scaled Bread Items	0	2,200	0.00	3,500	2.61
20	Raw & Cut Cured Meats	480	830	56.90	3,200	2.34
1	Water	250	6,300	3.86	3,000	2.22
2	Fresh Milk & Cream	3	1,300	0.23	2,600	1.92
20	Cooked & Cut Cured Meats	470	780	59.97	2,300	1.70
2	Pre-packaged Cheeses	290	880	32.65	2,100	1.59

Table 4.17: Location 2: Non-Cardholders Top 10 Categorical Expenditure

Tables 4.16 and 4.17 uncover a discernible spending pattern among cardholders and non- card-

holders at location 2. The bulk of their expenditures overlap across the same eight categories, with two notable exceptions: Seasoned Cheeses Counter and Frozen Fish for cardholders, and Prepackaged Cheeses and Water for non-cardholders. This pattern suggests a consistent preference for specific categories among both cardholders and non-cardholders, underscoring the enduring significance of particular items in their shopping selections at location 2, mirroring the trends observed at location 1.

In addressing the imperative to optimize promotions for the diverse customer base of the examined grocery store, our analysis focused on categorizing offers based on cardholding statuses and overall purchasing behavior. This categorization specifically targeted item categories highly frequented by customers. Tables 4.18 and 4.19 meticulously detail the outcomes of this categorization for locations 1 and 2.

Across both locations, cardholders consistently demonstrated a higher average percentage of items bought on offer per category compared to non-cardholders and the overall customer base. An interesting exception arose at the first location for "Plastic Bags," where non-cardholders exhibited an exceptionally high percentage of items bought on offer, surpassing cardholders. This anomaly, unrelated to a significant category, does not warrant further analysis.

Moreover, a notable disparity in percentages between the two tables is evident, with consistently higher figures in Table 4.19 than in Table 4.18. This highlights distinct shopping behaviors between the two locations, emphasizing that customers at location 2 consistently purchase a higher percentage of items on offer compared to those at location 1.

Category Name	**Cardholders**	**Non-Cardholders**	**All Customers**
Mozzarella Cheeses	59.74	47.97	57.26
Pre-packaged Cured Meat	53.63	44.48	51.69
Semolina Pasta	54.25	41.51	51.43
Butter & Margarine	49.77	39.78	47.97
Pre-packaged Cheese Spreads	47.82	38.50	45.87
Packaged Fresh Pasta	47.52	37.61	45.37
Frozen Fish	47.92	31.92	44.78
Poultry	43.99	44.04	44.00
Plastic Bags	38.41	53.93	42.92
Napkins	42.15	42.42	42.20

Table 4.18: Top 10 Percentage of Items bought on Offer in different Categories at Location 1

The Spearman correlation coefficient table 4.20 reveals significant insights into the relationships among various features in the dataset for locations 1 and 2 on a categorical basis. The strong positive correlation coefficients between "Money Spent" and both "Total Items" and "Items on Offer" at both locations suggest a robust relationship. This indicates that as the total number of items or items on offer increases, the overall spending also tends to rise consistently.

Distinctive patterns emerge in category-specific correlations. Particularly noteworthy are categories such as "Water" and "Biscuits," which, although displaying weak correlations, are notably significant for our analysis. These categories exhibit a discernible influence on the total number of items purchased, especially evident at location 1. In contrast, "Vegetables" showcase robust correlations with both "Items on Offer" and "Total Items," underscoring their pronounced impact on shopping behavior.

Category Name	Cardholders	Non-Cardholders	All Customers
Cooked & Cut Cured Meats	71.06	59.97	68.27
Raw & Cut Cured Meats	67.06	56.90	64.74
Frozen Fish	63.20	50.47	60.22
Mozzarella Cheeses	61.35	47.36	58.06
Frozen Legumes & Vegetables	60.60	43.74	56.92
Ready Meals - Pizzas & Bases	59.89	44.21	55.98
Fresh Soft Cheeses	55.19	46.86	53.72
Pre-packaged Cheese Spreads	56.15	42.34	52.86
Tuna Oil	53.97	43.27	51.48
Fresh Sauces & Ready Meals	55.27	41.03	51.47

Table 4.19: Top 10 Percentage of Items bought on Offer in different Categories at Location 2

Delving deeper into the dynamics of two intriguing categories, "Plastic Bags" and "Uncensored," offers insights into the distinct shopping and operational dynamics of the grocery store's two locations. The prevalence of "Plastic Bags" is more pronounced at location 1 than at location 2. Similarly, the significance of "Uncensored" is more prominent at location 1. These trends can be rationalized within the context of location 1 being the larger of the two, implying a higher likelihood of having more items categorized as "Uncensored" than location 2. In further categorical analyses, the category "Uncensored" was handled properly as it doesn't provide essential descriptive feedback.

Feature 1	Feature 2	Location 1	Location 2
Total Items	Items on Offer	0.69	0.64
Money Spent	Items on Offer	0.60	0.57
Money Spent	Total Items	0.87	0.87
Percentage (Items on Offer)	Items on Offer	0.75	0.81
"Water"	Total Items	0.42	<0.40
"Biscuits"	Total Items	0.40	<0.40
"Plastic Bags"	Items on Offer	0.42	<0.40
"Plastic Bags"	Total Items	0.44	<0.40
"Plastic Bags"	Money Spent	0.42	<0.40
"Vegetables"	Items on Offer	0.40	<0.40
"Vegetables"	Total Items	0.51	0.41
"Vegetables"	Money Spent	0.51	0.40
"Uncensored"	Total Items	0.52	<0.40
"Uncensored"	Money Spent	0.56	<0.40
"Uncensored"	"Plastic Bags"	0.41	<0.40
Amount per Item	Money Spent	<0.40	0.41
"Insecticide Tablets"	"Insecticide Sprays"	<0.40	0.58

Table 4.20: Significant Spearman Correlation Coefficient Values from Categorical Analysis

In parallel with sectoral association and correlation analyses related to cardholding status, a comprehensive exploration was conducted for categorical association and correlation with cardholding status. An overview of these analyses is presented below.

Table 4.21 and Table 4.22 present key association rules elucidating the relationships between cardholders and specific item categories at locations 1 and 2, respectively. These rules were estab-

Antecedent	Consequent	Confidence	Lift	Support
"Water"	Card	0.70	1.10	0.16
"Biscuits"	Card	0.72	1.14	0.21
"Fresh Milk & Cream"	Card	0.78	1.23	0.16
"Plastic Bags"	Card	0.66	1.05	0.35
"Vegetables"	Card	0.73	1.15	0.29
"Vegetables"	"Plastic Bags"	0.67	1.27	0.27
"Plastic Bags", "Vegetables"	Card	0.72	1.14	0.19
"Vegetables", Card	"Plastic Bags"	0.67	1.26	0.19

Table 4.21: Association Rules: Key Relationships between Cardholders and Categories at Location 1

lished for Location 1 using a minimum support threshold of 0.15 and a minimum confidence level of 0.65. In contrast, for Location 2, association rules were generated with a minimum support of 0.15 and a minimum confidence set at 0.5. These criteria were employed to extract meaningful and relevant patterns from the data, tailoring the parameters to the characteristics of each location's dataset.

Table 4.21 - Location 1:

- Notably, categories such as "Water," "Biscuits," and "Fresh Milk & Cream" exhibit strong associations with cardholders, as indicated by confidence values ranging from 0.70 to 0.78.
- "Plastic Bags" and "Vegetables" also feature prominently, showcasing significant associations with cardholders. Interestingly, an association rule involving both "Plastic Bags" and "Vegetables" suggests a synergistic relationship, reinforcing their combined impact on cardholder behavior.

Antecedent	Consequent	Confidence	Lift	Support
"Biscuits"	Card	0.71	1.15	0.17
"Scaled Bread Items"	Card	0.68	1.10	0.18
"Fresh Milk & Cream"	Card	0.76	1.23	0.16
"Plastic Bags"	Card	0.61	0.99	0.27
"Vegetables"	Card	0.70	1.13	0.22
"Vegetables"	"Plastic Bags"	0.56	1.25	0.18

Table 4.22: Association Rules: Key Relationships between Cardholders and Categories at Location 2

Table 4.22 - Location 2:

- In the context of Location 2, "Scaled Bread Items" replace "Water" from Location 1 in displaying high association with cardholding status. In addition, "Fresh Milk & Cream" and "Vegetables" display robust associations with cardholders, as reflected in high confidence values of 0.76 and 0.70, respectively.
- Unlike Location 1, the association between "Plastic Bags" and cardholders at Location 2 is comparatively weaker, with a confidence value of 0.61. And the association rule involving both "Vegetables" and "Plastic Bags" at Location 2 suggests a distinctive interaction, albeit with a slightly lower confidence value than observed at Location 1.

	Value
"Vegetables"	0.17
"Fresh Milk & Cream"	0.16
"Mozzarella Cheeses"	0.13
"Pre-packaged Cheese Spreads"	0.12
"Pre-packaged Cheeses"	0.12

Table 4.23: Top 5 Spearman Correlation Coefficient Values of Categories against Cardholding status at Location 1

	Value
"Fresh Milk & Cream"	0.15
"Raw & Cut Cured Meats"	0.13
"Vegetables"	0.12
"Semolina Pasta"	0.11
"Mozzarella Cheeses"	0.11

Table 4.24: Top 5 Spearman Correlation Coefficient Values of Categories against Cardholding status at Location 2

These association rules shed light on specific item categories that significantly influence the purchasing behavior of cardholders at each location, providing valuable insights for targeted promotional strategies and inventory management. The impact of specific categories on cardholding status is corroborated through Spearman correlation analysis. Key correlation values for each location are outlined in Tables 4.23 and 4.24. Notably, categories such as "Vegetables" and "Fresh Milk & Cream" consistently emerge as influential factors in relation to cardholding status, a trend that had persisted across various stages of our categorical analysis.

4.5 Daily & Hourly Analyses

In this section, we employ visualization techniques to discern variations in shopping trends at the two locations throughout March 2023, focusing on daily and hourly bases separately. The objective is to categorize and identify specific disparities, providing a foundation for more targeted and nuanced analyses in subsequent sections. Location 1, being the larger of the two, boasted 8 cash registers, while Location 2 had a more modest setup with 3 cash registers. The operating hours differed between the two locations, with the store at Location 1 open from 8 AM to 9 PM throughout the week. In contrast, the store at Location 2 operated in the same timeframe on all days except Sunday. These variations in size and operating hours contribute to the distinct characteristics of each location, influencing shopping patterns and behaviors visualized below.

4.5.1 Daily Analyses

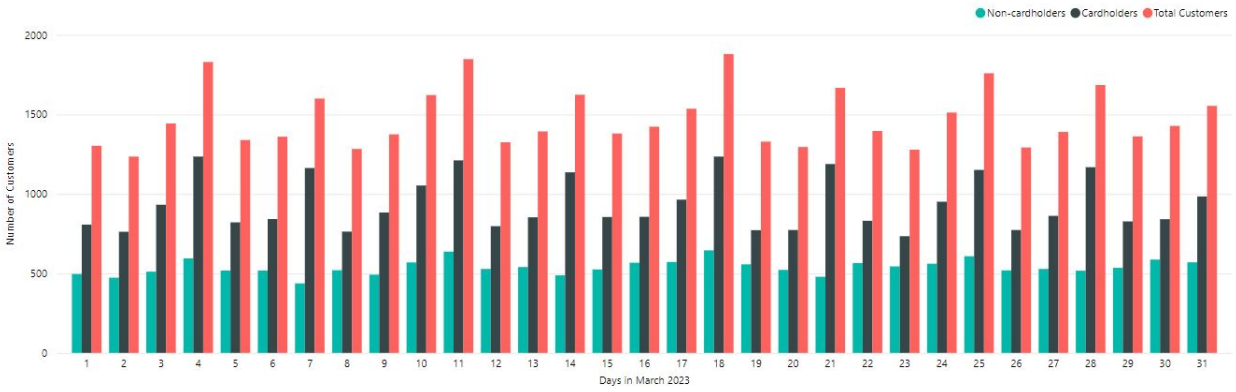


Figure 4.5: Total Customers across days of March 2023 at Location 1

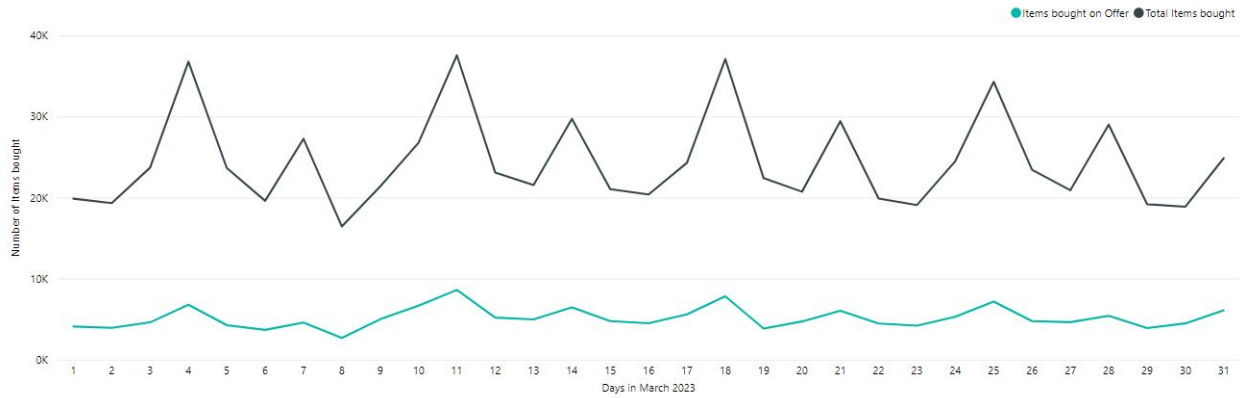


Figure 4.6: Items bought across days of March 2023 at Location 1

Figure 4.5 reveals a noteworthy trend in customer activity at Location 1. While the highest customer density aligns with weekends, a common pattern in grocery shopping, an intriguing deviation is observed on Tuesdays. This trend persists prominently in both total items purchased and total amount spent, as depicted in Figures 4.6 and 4.7 respectively.

Figure 4.7: Amount Spent across days of March 2023 at Location 1

At Location 2, depicted in Figure 4.8, the Tuesday anomaly took an intriguing turn by surpassing customer density compared to weekends. However, it's essential to note that this location was not operational on Sundays. The observed difference between Saturday's and Tuesday's statistics remains insubstantial throughout Figures 4.9 and 4.10, mirroring the pattern seen at Location 1. Interestingly, akin to Location 1, the surge in customer activity on Tuesdays at Location 2 doesn't translate proportionally to an increase in items bought on offer. This emphasizes a commonality between the two locations in the unique Tuesday shopping trend, suggesting potential shared factors influencing customer behavior on this specific day.

However, this Tuesday surge is not equally reflected in the items bought on offer. The noticeable difference in magnitude suggests that the heightened customer engagement on Tuesdays may not be directly correlated with variations in items being offered on discount. This observation hints at a distinct factor or influencing variable that contributes to the significant Tuesday activity, possibly unrelated to promotional offers. Further investigation into the nature of this unique Tuesday pattern

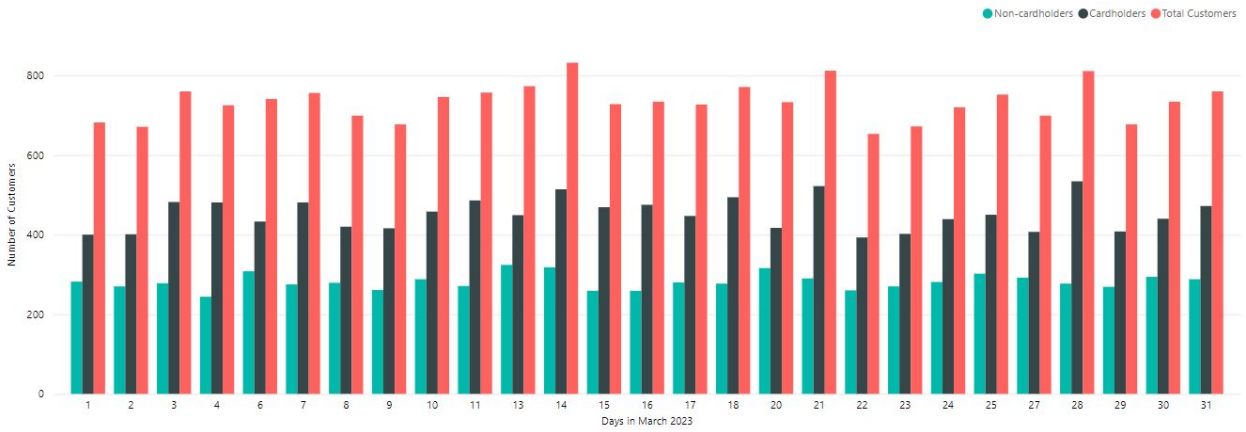


Figure 4.8: Total Customers across days of March 2023 at Location 2

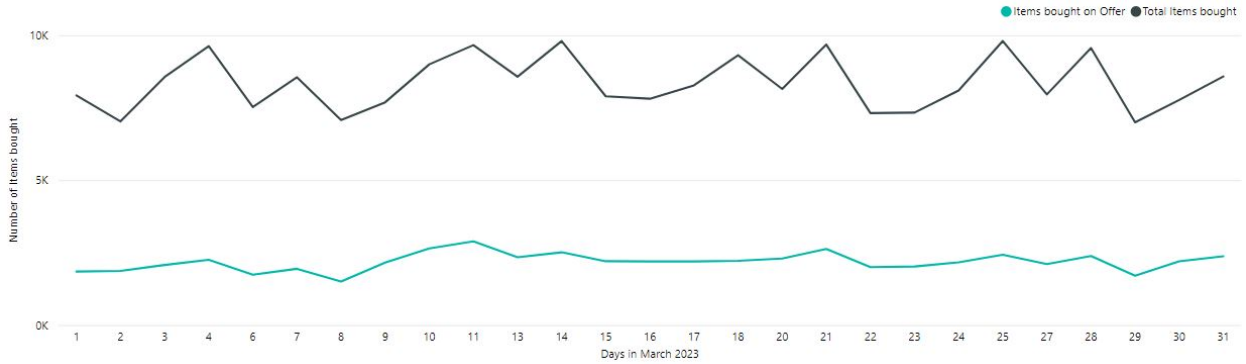


Figure 4.9: Items bought across days of March 2023 at Location 2

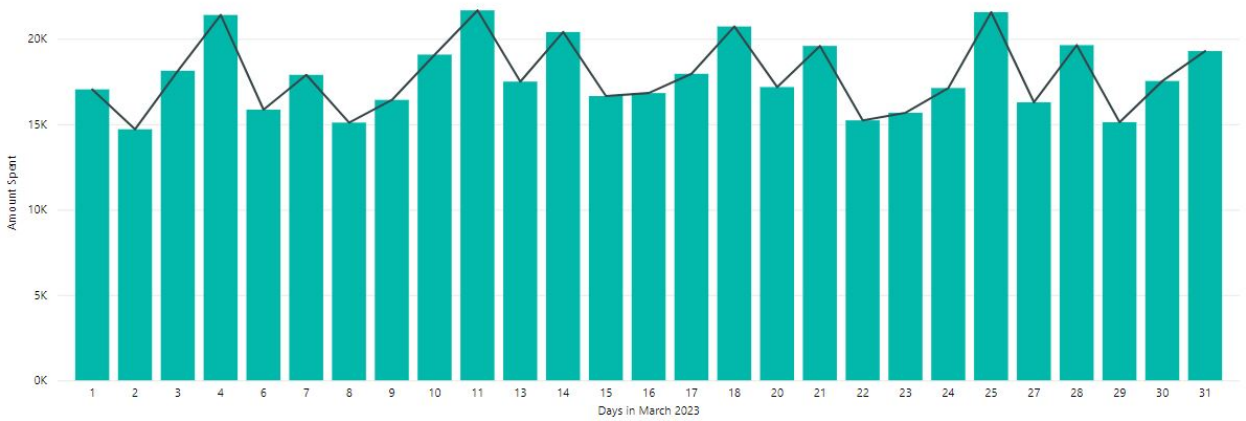


Figure 4.10: Amount Spent across days of March 2023 at Location 2

was continued using statistical techniques.

The initial approach to assess the association involved conducting a Chi-Square Test and computing Cramér's V values to explore the relationship between cardholding status and various days. The findings, as summarized in Table 4.25, indicate that the differences in values are not particularly significant across individual days but reveal noticeable variations between the two locations. In the case of location 1, Monday, Sunday, and Tuesday exhibit the highest Cramér's V values, while in location 2, Monday, Saturday, and Tuesday display the strongest associations. However, it is important to note that these associations lack the strength required for making definitive

Day	Location 1	Location 2
Monday	0.220	0.102
Tuesday	0.200	0.097
Wednesday	0.197	0.094
Thursday	0.197	0.093
Friday	0.185	0.091
Saturday	0.189	0.101
Sunday	0.223	NULL

Table 4.25: Cramér's V Values: Associating Days & Cardholding Status

conclusions.

The analysis proceeds with association testing utilizing the Apriori algorithm. The outcomes for locations 1 and 2 are presented in Tables 4.26 and 4.27, respectively. The selection of minimum confidence was based on the observation that 63.4% and 61.5% of transactions were made by cardholders at location 1 and 2, respectively. Additionally, the minimum support was chosen to ensure that all associations between days and cardholding status had the opportunity to be reflected in the considered item sets before the formation of the final association rules. The association rules tables for both locations reveal interesting insights into the relationships between cardholders and specific days.

Antecedent	Consequent	Confidence	Lift	Support
Tuesday	Card	0.71	1.12	0.10
Friday	Card	0.64	1.00	0.11
Saturday	Card	0.66	1.04	0.11

Table 4.26: Association Rules: Key Relationships between Cardholders and Days at Location 1

At Location 1, the rules indicate the following key associations:

- On Tuesday, cardholders show a confidence of 71%, suggesting a moderately strong association. The lift value of 1.12 indicates that the presence of a cardholder on Tuesday is 12% more likely compared to other days.
- On Friday, cardholders display a confidence of 64%, with a lift value of 1.00, suggesting a moderate association without a significant increase in likelihood.
- On Saturday, cardholders exhibit a confidence of 66%, indicating a moderate association, and a lift value of 1.04, implying a slight increase in the likelihood of cardholders on Saturdays.

Antecedent	Consequent	Confidence	Lift	Support
Tuesday	Card	0.64	1.04	0.10
Friday	Card	0.62	1.01	0.12
Saturday	Card	0.64	1.03	0.10

Table 4.27: Association Rules: Key Relationships between Cardholders and Days at Location 2

For Location 2, the association rules are as follows:

- On Tuesday, cardholders have a confidence of 64%, indicating a moderately strong association. The lift value of 1.04 suggests a slight increase in the likelihood of cardholders on Tuesdays.
- On Friday, cardholders show a confidence of 62%, with a lift value of 1.01, implying a moderate association without a significant increase in likelihood.
- On Saturday, cardholders exhibit a confidence of 64%, suggesting a moderately strong association, with a lift value of 1.03, indicating a slight increase in the likelihood of cardholders on Saturdays.

In summary, both locations demonstrate a notable association between cardholders and specific days, with Tuesday being consistently significant for cardholders at both locations. These findings strongly indicate a correlation between heightened customer traffic, particularly driven by cardholders, on Tuesdays.

4.5.2 Hourly Analyses

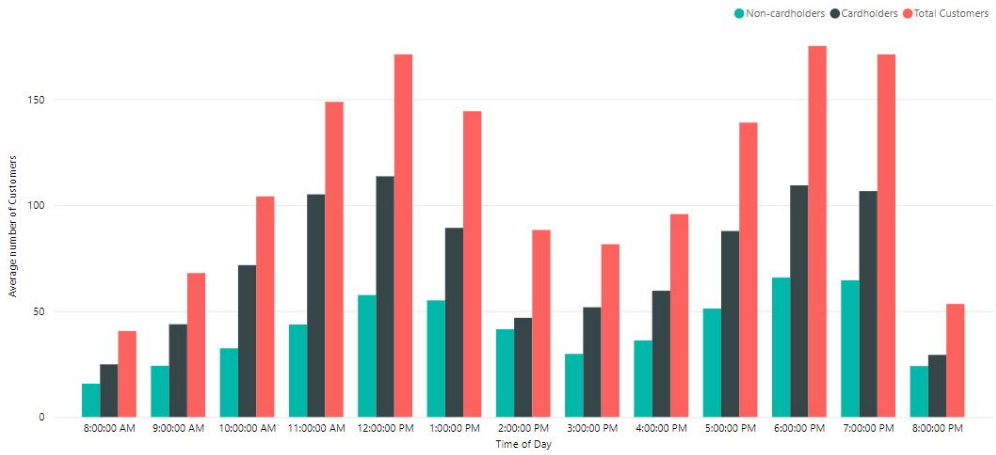


Figure 4.11: Average Number of Customers across operational timeframes at Location 1

During this phase of analysis, the transactional data was aggregated based on hourly intervals, rounding each transaction's timestamp to the nearest hour. Subsequently, the features related to cardholding status, items bought on offer, total items bought, and amount spent were averaged within each hourly timeframe. The resulting dataset was then visualized to discern trends and patterns, contributing to a more comprehensive understanding of the temporal dynamics of customer behavior. This approach allowed for a finer-grained exploration of variations and tendencies within specific hours, facilitating the identification of temporal patterns that may have implications for promotional strategies and operational decisions for both the locations.

The findings for location 1 are succinctly summarized using data visualizations presented in Figures 4.11, 4.12, and 4.13. Similarly, the insights for location 2 are derived from the corresponding Figures 4.14, 4.15, and 4.16.

For location 1, our analysis revealed that customers exhibited a consistent pattern of frequenting the store during the timeframes of 10:00 AM to 01:00 PM and 05:00 PM to 07:00 PM. This trend was particularly prominent in metrics such as total items bought, items bought on offer, and total amount spent during those specified hours. Interestingly, this temporal pattern was not confined

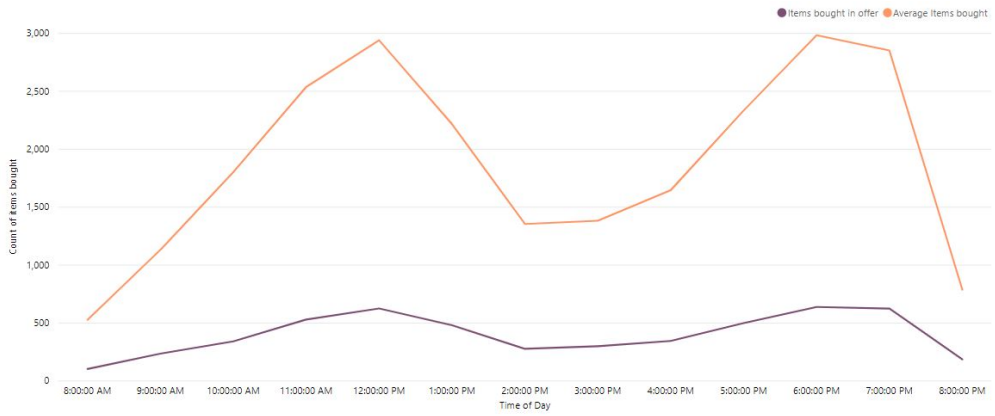


Figure 4.12: Average Number of Items bought across operational timeframes at Location 1

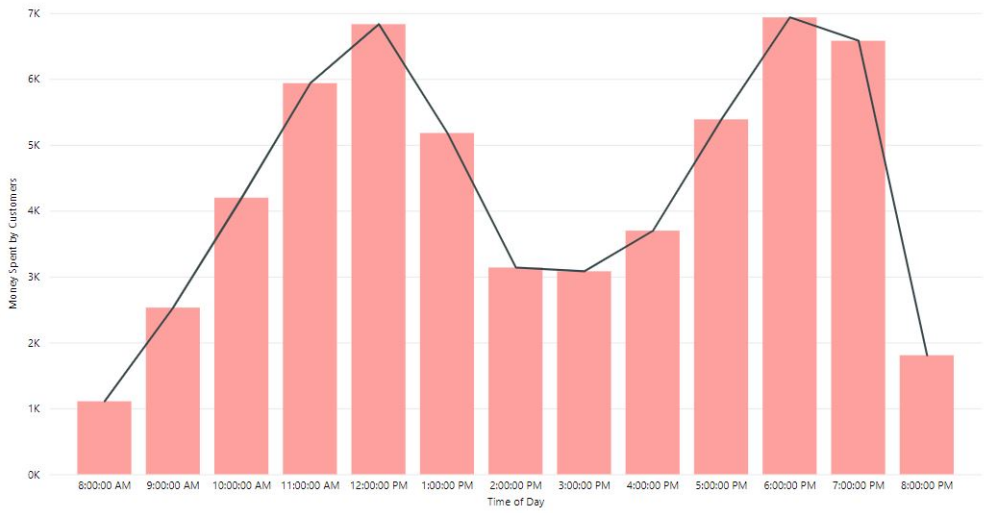


Figure 4.13: Average Amount Spent across operational timeframes at Location 1

to location 1 alone; location 2 mirrored the same peak hours of customer activity. However, a notable disparity between the two locations emerged during the 08:00 PM timeframe. In location 2, a sharp decline was observed across all categories, setting it apart from the more sustained customer activity observed in location 1 during the same period. This discrepancy underscores the importance of considering local nuances in customer behavior when devising location-specific strategies.

Continuing the analysis, we investigated the association of different time frames with cardholding status, considering the significant majority of transactions made by cardholders at both locations. This analysis followed a similar approach to the daily analyses, incorporating the Chi-Square Test, calculation of Cramér's V values, and the Apriori Algorithm to identify association rules.

In Table 4.28, the Cramér's V values reveal a consistent trend, indicating that the association between hours of operation and cardholding status is noteworthy. Specifically, these values are consistently high during the morning and afternoon hours, with the most significant associations found in the early opening hours. This suggests a moderate association between the time of day and the cardholding status of customers.

The Apriori analysis provides additional support to the association findings by identifying significant pairs of time frames associated with cardholding status. Tables 4.29 and 4.30 highlight key associations between time frames and cardholding status at locations 1 and 2, respectively. Similar

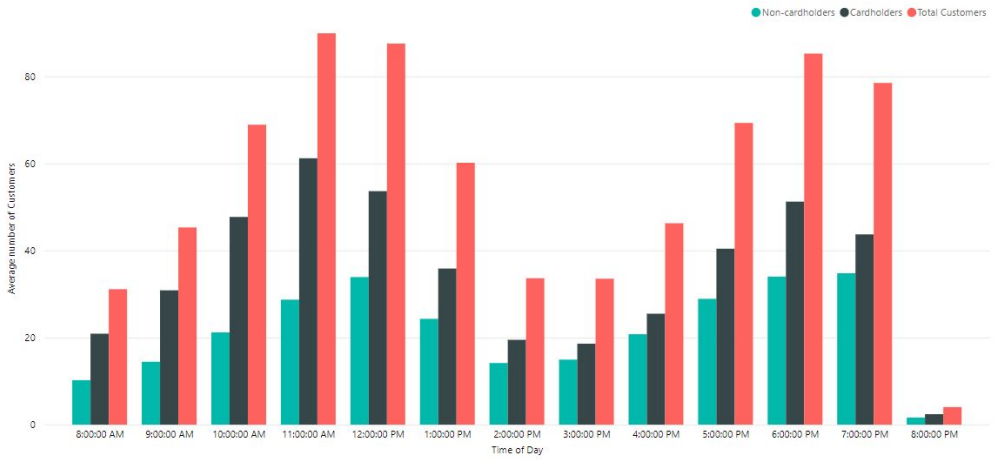


Figure 4.14: Average Number of Customers across operational timeframes at Location 2

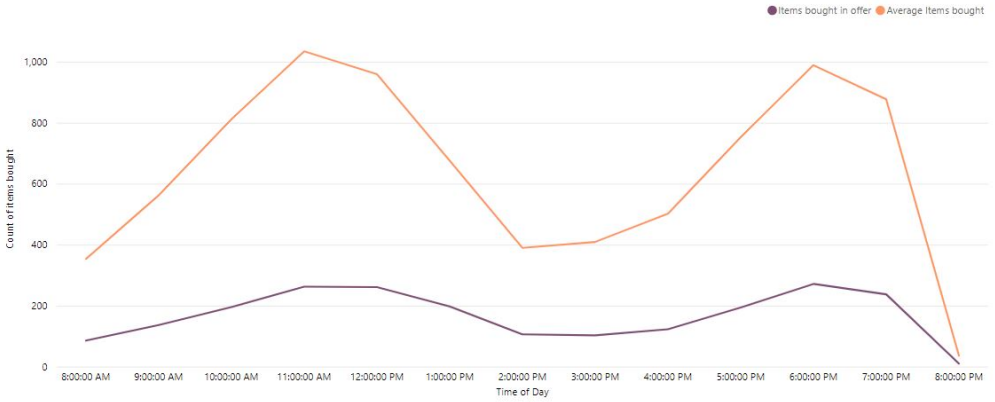


Figure 4.15: Average Number of Items bought across operational timeframes at Location 2

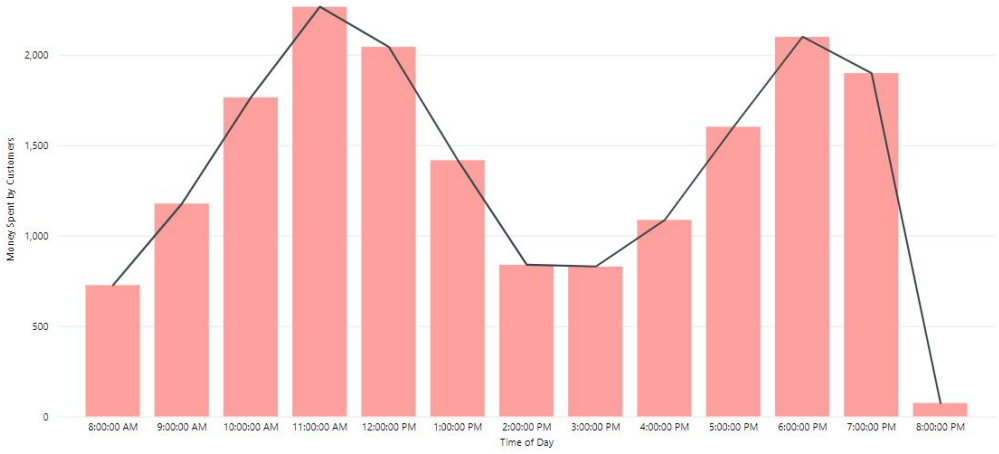


Figure 4.16: Average Amount Spent across operational timeframes at Location 2

to the daily analyses, the confidence cutoffs were determined based on the percentage of cardholders in the transactional data, while the support cutoffs were set to give all time frames an opportunity to be represented in the final association rules. The association rules between cardholders and time frames reveal insightful patterns.

Location 1:

Time Frame	Location 1	Location 2
08:00:00	0.572	0.498
09:00:00	0.412	0.413
10:00:00	0.333	0.335
11:00:00	0.279	0.293
12:00:00	0.260	0.297
13:00:00	0.283	0.358
14:00:00	0.362	0.479
15:00:00	0.376	0.480
16:00:00	0.347	0.409
17:00:00	0.288	0.334
18:00:00	0.257	0.301
19:00:00	0.260	0.314
20:00:00	0.465	NULL

Table 4.28: Cramér's V Values: Associating Time Frames & Cardholding Status

Antecedent	Consequent	Confidence	Lift	Support
09:00:00	Card	0.64	1.02	0.03
10:00:00	Card	0.69	1.09	0.05
11:00:00	Card	0.71	1.11	0.07
12:00:00	Card	0.66	1.05	0.08

Table 4.29: Association Rules: Key Relationships between Cardholders and Time Frames at Location 1

- At 09:00:00, cardholders exhibit a confidence of 64%, suggesting that there's a moderate association with this time frame.
- The association strengthens at 10:00:00, with a confidence of 69%, and further increases at 11:00:00 (71%) and decreases at 12:00:00 (66%).
- Generally, the lift values are above 1, indicating a positive association between cardholders and these specific time frames.

Antecedent	Consequent	Confidence	Lift	Support
08:00:00	Card	0.67	1.09	0.03
09:00:00	Card	0.68	1.11	0.04
10:00:00	Card	0.69	1.12	0.07
11:00:00	Card	0.68	1.11	0.08

Table 4.30: Association Rules: Key Relationships between Cardholders and Time Frames at Location 2

Location 2:

- Similar to location 1, there's a consistent increase in confidence from 08:00:00 to 11:00:00, ranging from 67% to 69%.
- The lift values are again above 1, indicating a positive association, with the strongest association observed at 11:00:00.

In both locations, the morning hours consistently show a positive association with the cardholders. These results can guide strategic decisions related to staffing, promotions, or other initiatives during these hours to maximize engagement with cardholding customers.

4.6 Big Basket Analysis

In this section, we explore a fundamental analysis widely employed in the retail industry known as Basket Analysis, with a specific focus on identifying significant customer segments—referred to as "Big Baskets." This analysis is particularly prevalent in the retail sector as it centers on discerning heavy shoppers and high spenders, who often contribute substantially to a company's profits. Our approach involves categorizing transactions as Big Baskets based on two criteria: having a higher-than-average number of items bought and surpassing the average amount spent across the entire transactional dataset. Notably, we conducted this categorization separately for each location, recognizing the distinct characteristics and customer behaviors observed at the two locations, necessitating tailored approaches to customer segmentation.

Upon conducting a preliminary analysis of Big Baskets and discerning the trends exhibited by these substantial shoppers, notable patterns emerged in their time and day preferences across both locations. The observed distinctions are systematically presented in Tables 4.31 and 4.32, corresponding to days and hours, respectively.

Day	Location 1	Location 2
Monday	26.81	28.21
Tuesday	33.65	31.86
Wednesday	24.88	28.23
Thursday	24.81	28.35
Friday	29.80	30.95
Saturday	38.98	35.94
Sunday	31.71	NULL

Table 4.31: Daily Heavy Shopper Percentage

It was evident that heavy shoppers demonstrated a preference for weekends at both locations, with the highest percentage of shoppers concentrating their spending on Saturdays—38.98% and 35.94% for locations 1 and 2, respectively. The Tuesday anomaly identified earlier persisted, with weekdays exhibiting lower percentages except for Tuesdays at both locations. Remarkably, Tuesday emerged as the second-heaviest day for Big Baskets, following closely behind Saturday.

Further examination in Table 4.32 revealed that heavy shoppers tended to favor morning hours from 09:00 AM to noon and, in the evening, from 04:00 PM to 07:00 PM at location 1. On the contrary, at location 2, they displayed a stronger inclination towards early mornings, particularly from 09:00 AM to 11:00 AM, with a notable spike at 06:00 PM in the evening. This data furnishes crucial insights for tailoring targeted promotional offers, taking into account variations in timeframes, days, and locations.

In our preliminary analysis, we conducted a detailed comparison between big basket statistics and their counterparts within the overall customer base. The key findings are synthesized in Table 4.33. Across both locations, heavy spenders exhibited a substantial increase in items bought on offer,

Time Frame	Location 1	Location 2
08:00:00	20.44	28.57
09:00:00	30.20	32.95
10:00:00	32.04	32.90
11:00:00	31.21	32.88
12:00:00	31.84	28.90
13:00:00	27.41	30.67
14:00:00	28.55	30.84
15:00:00	30.46	29.83
16:00:00	30.73	28.10
17:00:00	31.60	27.37
18:00:00	31.54	32.20
19:00:00	30.80	29.97
20:00:00	24.85	16.51

Table 4.32: Hourly Heavy Shopper Percentage

	Location 1			Location 2		
Category:	Small	Big	All	Small	Big	All
Items bought on Offer	2.0	6.9	3.5	1.7	5.9	3.0
Total Items bought	9	34	16	6	22	11
Total Amount Spent	19.00	82.00	38.00	13.00	49.00	24.00
Amount Spent per Item	2.70	2.50	2.60	2.31	2.27	2.30
% Items bought on Offer	21.9%	20.5%	21.5%	25.4%	26.5%	25.7%
% Cardholders	56.8%	78.9%	63.4%	54.9%	76.7%	61.6%

Table 4.33: Per-Transaction Average Values Comparison Chart for Location 1 & 2
(Small: Light Spenders | Big: Heavy Spenders | All: All Customers)

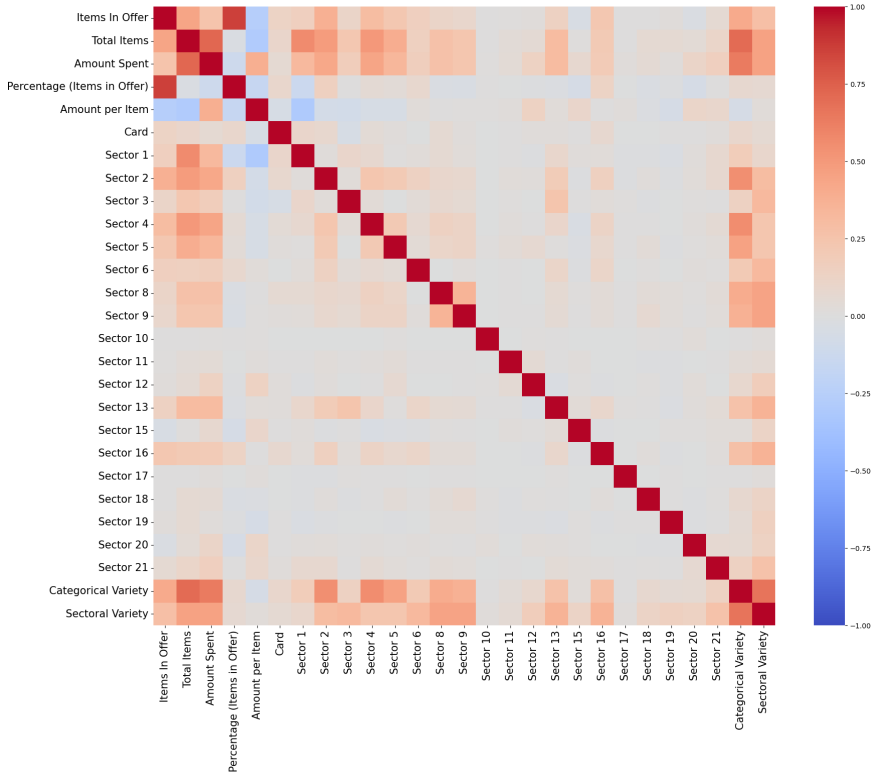
total items bought, and total amount spent, nearly doubling the figures compared to the overall customer base and more than tripling those of their counterpart shoppers who spent less. The percentage of cardholders among big basket customers surpassed that of the overall customer base and small basket customers at both locations. However, there was a divergence in the percentage of items bought on offer, with heavy spenders showing a higher percentage at location 2 but a lower one at location 1 compared to their counterparts. Intriguingly, the average amount spent per item decreased for heavy spenders at both locations. The variation was more pronounced at location 1, with a decrease of 10 cents, while location 2 saw a more modest 3-cent reduction.

Additionally, we conducted a comprehensive exploration of sectoral and categorical patterns specific to big basket transactions. The objective was to discern the preferences and associations of heavy spenders with particular sectors and item categories, guiding our subsequent analytical steps.

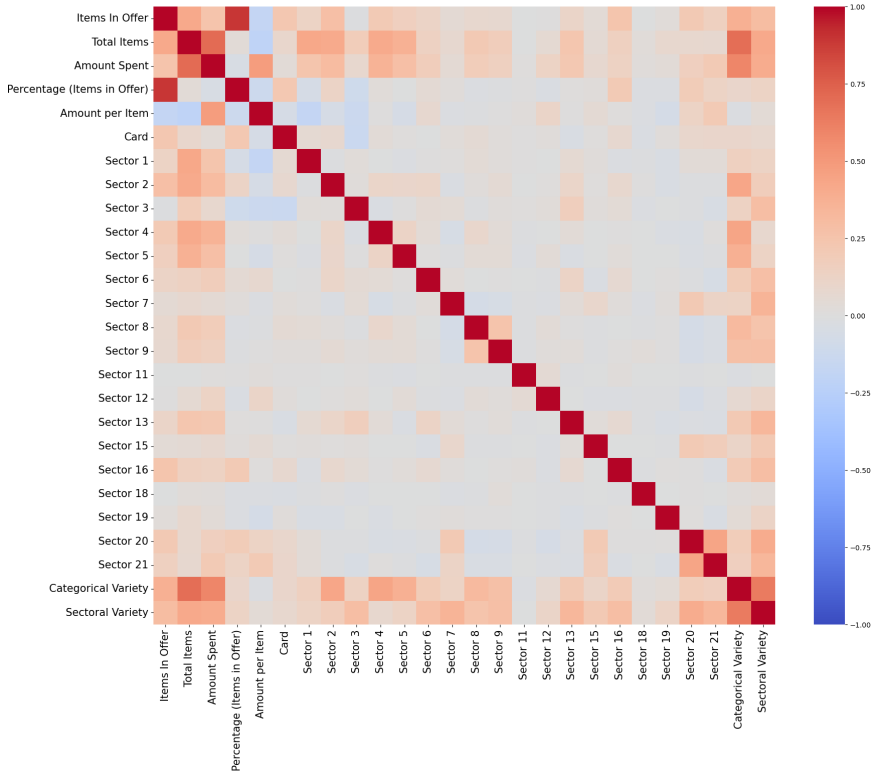
4.6.1 Sectoral Analysis

The sectoral analysis unfolded in two strategic phases: Spearman Correlation Analysis and Sector Association Rules. The heat maps illustrating the Spearman correlation coefficient values are depicted in Figure 4.17 for both locations, providing an initial overview of the results. The more salient correlation values are extracted and detailed in the subsequent analysis for a more in-depth

understanding.



(a) Location 1



(b) Location 2

Figure 4.17: Heat Maps depicting Sectoral Correlation Coefficient Values for Big Basket Transactions

A notable divergence emerged when comparing the findings of Table 4.34 with the results from

Section 4.3, particularly in the correlation coefficients for Percentage (Items on Offer) against Items on Offer. The coefficients for these variables exhibited an increase for both locations when considering big spenders, emphasizing a robust positive monotonic relationship between the total number of items bought on offer and the percentage of items bought on offer. Another significant difference surfaced in the reduced correlation coefficients between Categorical Variety and Sectoral Variety for both locations among big spenders. This indicates that heavy spenders tend to concentrate their purchases within specific sectors, as opposed to buying a few items across various sectors. This insight, coupled with high correlation coefficients between sectors such as "Beverages," "Food Items," and "Self-Service Counter" with Total Items, highlights these sectors as focal points for heavy shoppers. While overall patterns remained consistent for both locations, heavy shoppers at the second location displayed a distinctive correlation between the "Cheese Cutting Counter" and "Cured Meat Cutting Counter" sectors, underscoring the particular significance of these sectors for shoppers at location 2.

Feature 1	**Feature 2**	**Location 1**	**Location 2**
Percentage (Items on Offer)	Items on Offer	0.86	0.90
Amount Spent	Total Items	0.73	0.71
Categorical Variety	Total Items	0.71	0.69
Sectoral Variety	Categorical Variety	0.67	0.64
Categorical Variety	Amount Spent	0.64	0.59
"Beverages"	Total Items	0.57	0.42
Categorical Variety	"Food Items"	0.56	0.44
Categorical Variety	"Self-Service Counter"	0.55	0.43
"Food Items"	Total Items	0.50	0.41
"Self-Service Counter"	Total Items	0.49	0.41
Categorical Variety	"Confectionery - 1st Breakfast"	0.46	<0.40
Sectoral Variety	Total Items	0.45	0.42
Sectoral Variety	Amount Spent	0.45	0.40
Sectoral Variety	"Personal Hygiene"	0.45	<0.40
Sectoral Variety	"House Cleaning"	0.45	<0.40
Total Items	Items on Offer	0.44	0.42
"Food Items"	Amount Spent	0.44	<0.40
"Self-Service Counter"	Amount Spent	0.42	<0.40
Categorical Variety	Items on Offer	0.41	<0.40
Amount per Item	Amount Spent	<0.40	0.48
"Cheese Cutting Counter"	"Cured Meat Cutting Counter"	<0.40	0.44
Sectoral Variety	"Cured Meat Cutting Counter"	<0.40	0.40

Table 4.34: Significant Spearman Correlation Coefficient Values from Sectoral Analysis of Big Basket Transactions

Our observation that heavy shoppers exhibit a focus on the "Beverages," "Food Items," and "Self-Service Counter" sectors gains additional support through our association analysis. The top association rules, detailed in Tables 4.35 and 4.36 for locations 1 and 2, respectively, contribute valuable insights into the connections between cardholders and specific sectors in big basket transactions.

To extract meaningful associations, a support cutoff of 0.70 and a confidence cutoff of 0.95 were applied for location 1. For location 2, considering its smaller scale, the established cutoffs

were set at 0.65 for support and 0.85 for confidence. These association rules provide a deeper understanding of the relationships between cardholders and particular sectors, highlighting the sectors that significantly contribute to big basket transactions at both locations. The findings underscore the strategic importance of sectors like "Beverages," "Food Items," and "Self-Service Counter" in tailoring promotions and offers to cater specifically to heavy shoppers.

Antecedent	Consequent	Confidence	Lift	Support
Card	"Beverages"	0.98	1.00	0.77
"Self-Service Counter"	"Beverages"	0.98	1.00	0.87
"Non-Food Items"	"Beverages"	0.98	1.00	0.71
"Food Items"	"Beverages"	0.97	1.00	0.89
"Confectionery - 1st Breakfast"	"Beverages"	0.97	1.00	0.83
"Fruits & Vegetables"	"Beverages"	0.98	1.01	0.73
Card, "Self-Service Counter"	"Beverages"	0.98	1.01	0.70
Card, "Food Items"	"Beverages"	0.98	1.01	0.71
"Self-Service Counter", "Food Items"	"Beverages"	0.98	1.00	0.81
"Confectionery...", "Self-Service Counter"	"Beverages"	0.97	1.00	0.75
"Confectionery...", "Food Items"	"Beverages"	0.97	1.00	0.76
"Confectionery...", "Self-Service...", "Food..."	"Beverages"	0.97	1.00	0.70

Table 4.35: Big Basket Association Rules: Key Connections between Cardholders & Sectors at Location 1

At Location 1, cardholders demonstrated a strong association with the "Beverages" sector, with various combinations such as "Self-Service Counter," "Food Items," and "Confectionery - 1st Breakfast." The confidence values ranged from 0.97 to 0.98, indicating a high likelihood of cardholders purchasing items from the "Beverages" sector when also buying from these associated sectors. These associations underscore the distinct preferences and purchasing patterns of cardholders within specific sectors.

Antecedent	Consequent	Confidence	Lift	Support
Card	"Beverages"	0.85	1.01	0.65
Card	"Self-Service Counter"	0.86	1.02	0.66
Card	"Food Items"	0.90	1.00	0.69
"Beverages"	"Food Items"	0.90	1.00	0.75
"Food Items"	"Self-Service Counter"	0.86	1.01	0.77
"Self-Service Counter"	"Food Items"	0.90	1.01	0.77
"Confectionery - 1st Breakfast"	"Self-Service Counter"	0.86	1.01	0.71
"Confectionery - 1st Breakfast"	"Food Items"	0.91	1.01	0.75

Table 4.36: Big Basket Association Rules: Key Connections between Cardholders & Sectors at Location 2

Similarly, at Location 2, cardholders exhibited significant associations with the "Beverages" sector, along with notable connections to "Self-Service Counter" and "Food Items." Confidence values ranged from 0.85 to 0.91, emphasizing a strong correlation between cardholders and their preference for items from these sectors. The associations highlight the distinct shopping behavior of high spending cardholders, particularly in relation to specific sectors.

When considering both locations collectively, the commonality lies in the strong affiliation between cardholders and the "Beverages" sector. While the specific combinations of associated sectors

may differ between locations, the overall trend suggests that cardholders, in big basket transactions, consistently show a preference for items from the "Beverages" sector. This unified observation underscores the potential for targeted promotional strategies within this sector for big basket customers across both locations.

4.6.2 Categorical Analysis

Similar to the sectoral analysis discussed earlier, this section delves into two primary strategies for big basket categorical analysis: Spearman Correlation Analysis and Category Association Rules. Examining the Spearman Correlation Coefficients presented in Table 4.37, a trend reminiscent of the Big Basket Sectoral Analysis emerges. Notably, the correlation coefficient values related to the relationship between the Percentage of Items on Offer and Items on Offer exhibit a significant increase compared to the results from Section 4.4. While other trends show some fluctuations across locations and basket categorization, the overall patterns remain consistent.

Feature 1	**Feature 2**	**Location 1**	**Location 2**
Percentage (Items in Offer)	Items In Offer	0.86	0.90
Amount Spent	Total Items	0.73	0.71
Categorical Variety	Total Items	0.71	0.69
Sectoral Variety	Categorical Variety	0.67	0.64
Categorical Variety	Amount Spent	0.64	0.59
Sectoral Variety	Total Items	0.45	0.42
Sectoral Variety	Amount Spent	0.45	0.40
Total Items	Items In Offer	0.44	0.42
Categorical Variety	Items In Offer	0.41	0.38
"Water"	Total Items	0.40	<0.35
Amount per Item	Amount Spent	0.39	0.48
"Water"	Amount per Item	-0.43	<0.35
"Hand Cream"	"Dark Beer"	<0.35	0.35
"Cooked & Cut Cured Meats"	"Raw & Cut Cured Meats"	<0.35	0.39
Sectoral Variety	"Scaled Bread Items"	<0.35	0.36

Table 4.37: Significant Spearman Correlation Coefficient Values from Categorical Analysis of Big Basket Transactions

Zooming into location-specific results, at location 1, the category "Water" demonstrates a positive correlation with Total Items and a negative correlation with Amount per Item, suggesting a logical connection between these variables. Location 2 presents intriguing findings, indicating a noteworthy correlation of 0.35 between "Hand Cream" and "Dark Beer" and 0.39 between "Cooked & Cut Cured Meats" and "Raw & Cut Cured Meats." These correlations signify that these items either consistently occur together or do not occur together. A closer look at the association rules in Table 4.39 reveals that these categories most commonly do not occur together, as evidenced by their combined support and confidence levels not placing them in the Table 4.39.

To enhance the granularity of the analysis and capture the nuanced relationships between different categories within big basket transactions, a crucial preprocessing step involved the exclusion of the cardholding status feature. This adjustment aimed to mitigate the dominant influence of cardholders in the dataset, allowing for a more detailed exploration of interconnected categories

within big basket transactions.

The association rules presented in Tables 4.38 and 4.39 were derived under specific criteria: a minimum support of 0.20 for location 1 and 0.10 for location 2, and a minimum confidence of 0.7 for location 1 and 0.6 for location 2. These thresholds were set to ensure that the established associations were both statistically significant and provided meaningful insights into the relationships between different categories for heavy spenders at each location.

Antecedent	Consequent	Confidence	Lift	Support
"Fresh Fruits"	"Plastic Bag"	0.80	1.16	0.20
"Fresh Fruits"	"Vegetables"	0.84	1.20	0.21
"Fresh Milk & Cream"	"Vegetables"	0.73	1.04	0.27
"Plastic Bag"	"Vegetables"	0.74	1.06	0.51
"Vegetables"	"Plastic Bag"	0.73	1.06	0.51
"Pre-packaged Cheese Spreads"	"Vegetables"	0.76	1.09	0.23
"Pre-packaged Cheeses"	"Vegetables"	0.77	1.09	0.28
"Mozzarella Cheeses"	"Vegetables"	0.76	1.09	0.24
"Biscuits", "Plastic Bag"	"Vegetables"	0.74	1.06	0.26
"Biscuits", "Vegetables"	"Plastic Bag"	0.72	1.05	0.26
"Pre-packaged Cheeses", "Plastic Bag"	"Vegetables"	0.81	1.15	0.20
"Pre-packaged Cheeses", "Vegetables"	"Plastic Bag"	0.73	1.06	0.20

Table 4.38: Big Basket Association Rules: Key Connections between Categories at Location 1

The association rules presented in Tables 4.38 and 4.39 unveil key connections between categories within big basket transactions at locations 1 and 2, respectively. At location 1, noteworthy associations include a strong connection between "Fresh Fruits" and "Vegetables," with a confidence of 0.84, indicating that when customers purchase fresh fruits, there is an 84% likelihood they will also buy vegetables. Additionally, "Pre-packaged Cheeses" and "Vegetables" exhibit a strong connection, emphasizing a consistent shopping pattern among these categories.

Antecedent	Consequent	Confidence	Lift	Support
"Rough Bread"	"Vegetables"	0.62	1.10	0.10
"Fresh Fruits"	"Vegetables"	0.75	1.33	0.10
"Poultry"	"Vegetables"	0.67	1.19	0.10
"Vegetables"	"Plastic Bag"	0.61	1.08	0.35
"Pre-packaged Cheese Spreads"	"Vegetables"	0.61	1.09	0.16
"Pre-packaged Cheeses"	"Vegetables"	0.63	1.12	0.15
"Mozzarella Cheeses"	"Vegetables"	0.61	1.09	0.14
"Biscuits", "Scaled Bread Items"	"Vegetables"	0.61	1.09	0.10
"Vegetables", "Scaled Bread Items"	"Plastic Bag"	0.62	1.09	0.13
"Scaled Bread Items", "Plastic Bag"	"Vegetables"	0.62	1.10	0.13

Table 4.39: Big Basket Association Rules: Key Connections between Categories at Location 2

On the other hand, at location 2, associations like "Fresh Fruits" and "Vegetables" (confidence of 0.75) and "Poultry" and "Vegetables" (confidence of 0.67) highlight specific item combinations favored by customers. Interestingly, the rule "Vegetables" and "Plastic Bag" indicates that when customers buy vegetables, there is a 61% likelihood they will also purchase plastic bags, suggesting

a consistent pairing of these items in shopping baskets.

Combining the insights from both locations, it's evident that certain categories tend to co-occur more frequently in big basket transactions. These associations provide valuable information for inventory management and promotional strategies, enabling the store to enhance the shopping experience for heavy spenders.

4.7 Cardholder Demographic Associations: Exploring Day & Time Preferences

In this segment of the analysis, we delve into demographical trends among cardholding customers. Due to the sensitive nature of the data, details regarding geographical information, particularly CAP codes, are omitted. Instead, the focus is on revealing insights into the day and time preferences of cardholding customers, categorized by age and gender groups.

The analysis of the age distribution within the cardholding customer base was conducted in a straightforward manner, focusing on average data across various days of the week and time frames throughout the day. This approach was chosen based on the insights derived from Figure 4.1, which indicated a concentrated density of cardholding customers in the 45-80 years age range. Attempts to group ages further proved impractical, as the data from these groups introduced significant distortions in the overall analysis results.

	Location 1		**Location 2**	
Days	Birth Year	Age	Birth Year	Age
Monday	1967	56	1962	61
Tuesday	1961	62	1958	65
Wednesday	1966	57	1961	62
Thursday	1965	58	1960	63
Friday	1965	58	1960	63
Saturday	1966	57	1960	63
Sunday	1968	55	NULL	NULL

Table 4.40: Average Birth Year and Age of Cardholders by Days

Table 4.40 presents valuable insights into the age distribution of cardholders at both locations. Further examination revealed that the average age of cardholders at location 1 was 58 years, while those at location 2 averaged 63 years. This suggests a trend where an older customer base tends to prefer the smaller location. The table also sheds light on the Tuesday anomaly: on this day, the average age was 62 years at location 1 and 65 years at location 2, surpassing the respective overall averages. This indicates an association between high customer density on Tuesdays and their age.

Upon analyzing the data presented in Table 4.41, a clear trend emerges, indicating that older cardholders exhibit a preference for morning grocery shopping at both locations. In location 1, the highest average age, 63 years, is observed during the 09:00:00 and 10:00:00 time slots. Similarly, at location 2, significantly higher average ages of 69 and 68 years are registered during the 10:00:00 and 11:00:00 time slots, surpassing the averages observed at location 1. These findings, when considered alongside the results from Tables 4.29 and 4.30, suggest a consistent preference among older cardholders for morning time frames when engaging in grocery shopping.

	Location 1		Location 2	
Time Frames	Birth Year	Age	Birth Year	Age
08:00:00	1966	57	1961	62
09:00:00	1963	60	1957	66
10:00:00	1960	63	1954	69
11:00:00	1960	63	1955	68
12:00:00	1962	61	1957	66
13:00:00	1966	57	1961	62
14:00:00	1967	56	1963	60
15:00:00	1966	57	1963	60
16:00:00	1967	56	1962	61
17:00:00	1967	56	1963	60
18:00:00	1968	55	1965	58
19:00:00	1970	53	1967	56
20:00:00	1972	51	1969	54

Table 4.41: Average Birth Year and Age of Cardholders by Time Frames

In the following part of this section, we delve into an analysis of cardholding customers, categorizing them based on their self-classified genders. The analysis focuses on the association of days of the week and time frames of the day with two gender categories: men and women. Tables 4.42 and 4.43 present essential numerical data and percentages pertaining to the time frames of the day and days of the week, respectively, for both locations.

	Location 1				Location 2			
Time Frames	Men	Women	% Men	% Women	Men	Women	% Men	% Women
08:00:00	8	16	35	65	5	15	27	73
09:00:00	15	27	36	64	8	21	29	71
10:00:00	26	44	37	63	13	33	29	71
11:00:00	36	67	35	65	16	43	26	74
12:00:00	39	72	35	65	14	38	27	73
13:00:00	30	57	34	66	9	26	26	74
14:00:00	14	32	30	70	5	13	29	71
15:00:00	16	34	32	68	6	12	33	67
16:00:00	19	39	33	67	8	17	32	68
17:00:00	30	55	35	65	11	28	29	71
18:00:00	37	69	35	65	13	37	26	74
19:00:00	41	64	39	61	13	30	30	70
20:00:00	11	18	36	64	1	2	44	56

Table 4.42: Average Numbers and Percentages of Men & Women Shoppers by Time Frames

Upon a cursory examination of Tables 4.42 and 4.43, no glaring patterns or anomalies are immediately evident. The data appears to be relatively consistent, with minor variations in certain instances. To gain deeper insights and draw meaningful conclusions regarding potential relationships between the days of the week, time frames, and the genders of shoppers, we employ statistical methods for calculating associations among variables.

Assessing the Cramér's V values presented in Tables 4.44 and 4.45, it is evident that the as-

	Location 1				Location 2			
Days	Men	Women	% Men	% Women	Men	Women	% Men	% Women
Monday	285	526	35	65	123	291	30	70
Tuesday	400	733	35	65	135	360	27	73
Wednesday	270	526	34	66	99	305	25	75
Thursday	275	518	35	65	119	294	29	71
Friday	331	626	35	65	127	317	29	71
Saturday	429	746	37	63	138	321	30	70
Sunday	284	486	37	63	NULL	NULL	NULL	NULL

Table 4.43: Average Numbers and Percentages of Men & Women Shoppers by Days

Time Frames	**Location 1**	**Location 2**
08:00:00	0.278	0.194
09:00:00	0.197	0.161
10:00:00	0.154	0.129
11:00:00	0.127	0.114
12:00:00	0.122	0.121
13:00:00	0.138	0.149
14:00:00	0.191	0.202
15:00:00	0.182	0.207
16:00:00	0.169	0.176
17:00:00	0.140	0.140
18:00:00	0.125	0.124
19:00:00	0.126	0.134
20:00:00	0.242	NULL

Table 4.44: Cramér's V Values: Associating Time Frames & Gender of Customers

Days	**Location 1**	**Location 2**
Monday	0.059	0.109
Tuesday	0.050	0.100
Wednesday	0.053	0.099
Thursday	0.053	0.098
Friday	0.049	0.094
Saturday	0.049	0.104
Sunday	0.060	NULL

Table 4.45: Cramér's V Values: Associating Days & Gender of Customers

sociation between genders and both the days of the week and time frames of the day is generally weak. However, subtle patterns emerge, indicating some association with early morning and mid-afternoon time frames. It's noteworthy that in Table 4.44, the Cramér's V values for time frames 08:00:00 and 20:00:00 are notably high for both locations. However, caution is required in interpreting these values, as opening and closing hours may lack sufficient data compared to the rest of the day, potentially leading to statistical skews based on insufficient data.

To further delve into associations, we employed the Apriori algorithm, generating association rules for both locations categorized by days of the week and time frames of the day. Key results are

summarized in Tables 4.46, 4.47, 4.48, and 4.49 shedding light on potential patterns in the data. We set confidence cutoff rules for association rules based on the average ratios of gender categories. At location 1, the confidence cutoffs were 0.35 for men and 0.65 for women. At location 2, the corresponding cutoffs were 0.28 for men and 0.72 for women. Support cutoffs were carefully chosen to capture a substantial number of frequently occurring itemsets.

Antecedent	Consequent	Confidence	Lift	Support
10:00:00	Man	0.37	1.06	0.03
13:00:00	Woman	0.66	1.02	0.06
14:00:00	Woman	0.70	1.07	0.03
15:00:00	Woman	0.67	1.04	0.04
16:00:00	Woman	0.67	1.03	0.04
19:00:00	Man	0.39	1.10	0.04

Table 4.46: Association Rules: Key Connections between Time Frames & Gender at Location 1

Antecedent	Consequent	Confidence	Lift	Support
8:00:00	Woman	0.74	1.02	0.03
10:00:00	Man	0.29	1.03	0.03
11:00:00	Woman	0.73	1.02	0.10
12:00:00	Woman	0.73	1.02	0.09
13:00:00	Woman	0.74	1.03	0.06
17:00:00	Man	0.29	1.02	0.03
18:00:00	Woman	0.74	1.03	0.08
19:00:00	Man	0.31	1.08	0.03

Table 4.47: Association Rules: Key Connections between Time Frames & Gender at Location 2

The association rules in Tables 4.46 and 4.47 reveal interesting connections between time frames and gender at Locations 1 and 2, respectively. At Location 1, the rules suggest that men are more likely to shop at 10:00:00 and 19:00:00 time frames, while women show a higher likelihood of shopping during afternoon hours, particularly at 13:00:00, 14:00:00, 15:00:00, and 16:00:00 time frames. At Location 2, women exhibit a higher likelihood of shopping during morning and midday hours (8:00:00, 11:00:00, 12:00:00, and 13:00:00), whereas men are associated with evening hours, especially at 17:00:00 and 19:00:00. These findings provide valuable insights into the temporal shopping preferences based on gender at each location.

Antecedent	Consequent	Confidence	Lift	Support
Wednesday	Woman	0.66	1.02	0.09
Thursday	Woman	0.65	1.01	0.09
Friday	Woman	0.65	1.01	0.11
Saturday	Man	0.37	1.04	0.06
Sunday	Man	0.37	1.05	0.04

Table 4.48: Association Rules: Key Connections between Days & Gender at Location 1

The association rules in Tables 4.48 and 4.49 provide insights into the connections between days of the week and gender at Locations 1 and 2, respectively. At Location 1, the rules suggest that

Antecedent	Consequent	Confidence	Lift	Support
Monday	Man	0.30	1.06	0.04
Tuesday	Woman	0.73	1.01	0.12
Wednesday	Woman	0.75	1.05	0.13
Thursday	Man	0.29	1.03	0.05
Friday	Man	0.29	1.02	0.05
Saturday	Man	0.30	1.07	0.05

Table 4.49: Association Rules: Key Connections between Days & Gender at Location 2

women are more likely to shop on Wednesday, Thursday, and Friday, while at Location 2, Tuesday and Wednesday are associated with women shoppers. On the other hand, men at Location 1 are linked to Saturday and Sunday, and at Location 2, Monday, Thursday, Friday, and Saturday are associated with male shoppers. These findings indicate variations in shopping preferences based on both gender and specific days of the week at each location.

In conclusion, the comprehensive exploration undertaken in this chapter has provided valuable insights into various aspects of customer behavior and preferences. The demographic analysis unveiled intriguing patterns, shedding light on the diverse characteristics of our customer base. Transactional analyses, item sector and category breakdowns, and daily-hourly examinations added depth to our understanding of customer engagement. The specific focus on Big Basket transactions and the subsequent cardholder demographic analysis deepened our insights into customer preferences regarding days and time slots.

Now as we transition from our thorough exploratory analyses to the formulation of customer segments, it is crucial to underscore the significance of our descriptive investigations. Throughout this chapter, we meticulously unearthed numerous trends within our retail market datasets, offering valuable insights into customer behavior and market dynamics. This wealth of understanding serves as a robust foundation for the upcoming chapter, where we will delve into the application of various data-driven techniques to perform segmentation. The insights gained through exploratory analyses not only inform our understanding of the data but also guide the strategic application of segmentation methods to uncover meaningful patterns and clusters within the customer landscape.

Chapter 5

Data Driven Customer Segmentation

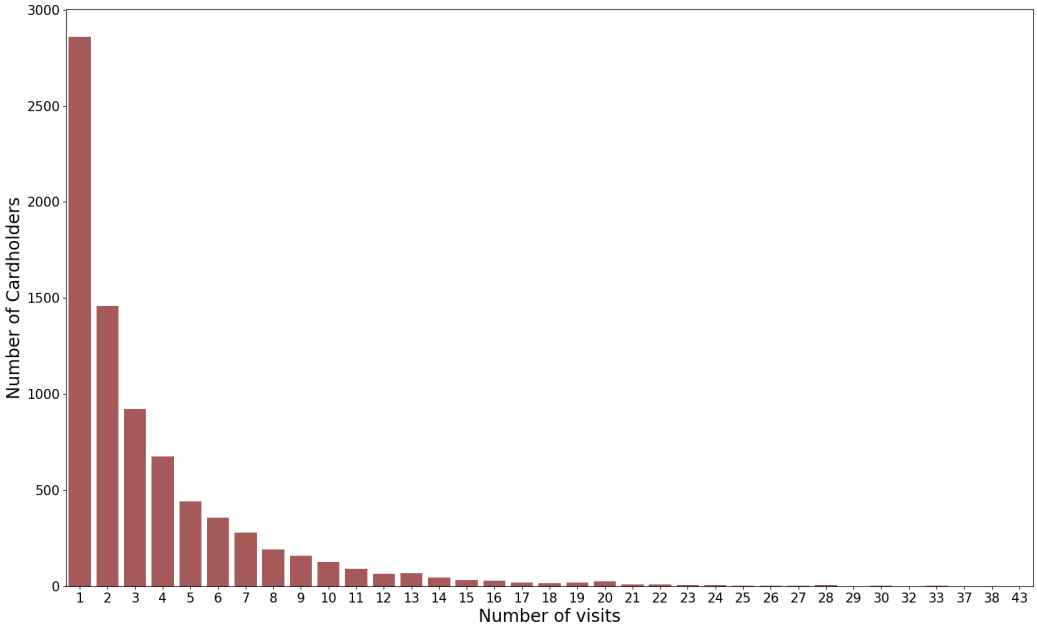
In this chapter, we thoroughly investigate the results obtained from employing various clustering techniques to identify patterns and groupings within the customer base. By harnessing advanced analytical methods, our goal is to uncover hidden structures that offer valuable insights into customer behavior, preferences, and engagement. This chapter delves into the diverse clusters formed based on multiple factors, illuminating the unique characteristics of each group. Furthermore, it employs a range of data-driven techniques for cluster formulation, spanning from predefined segmentation rules to sophisticated machine learning clustering algorithms. Through a comprehensive analysis of these clusters and methods, along with cross-examination, our objective is to derive actionable conclusions that can inform strategic decision-making and enrich our comprehension of the customer landscape.

5.1 Predefined Customer Segmentation through RFM: Analyzing Cardholders

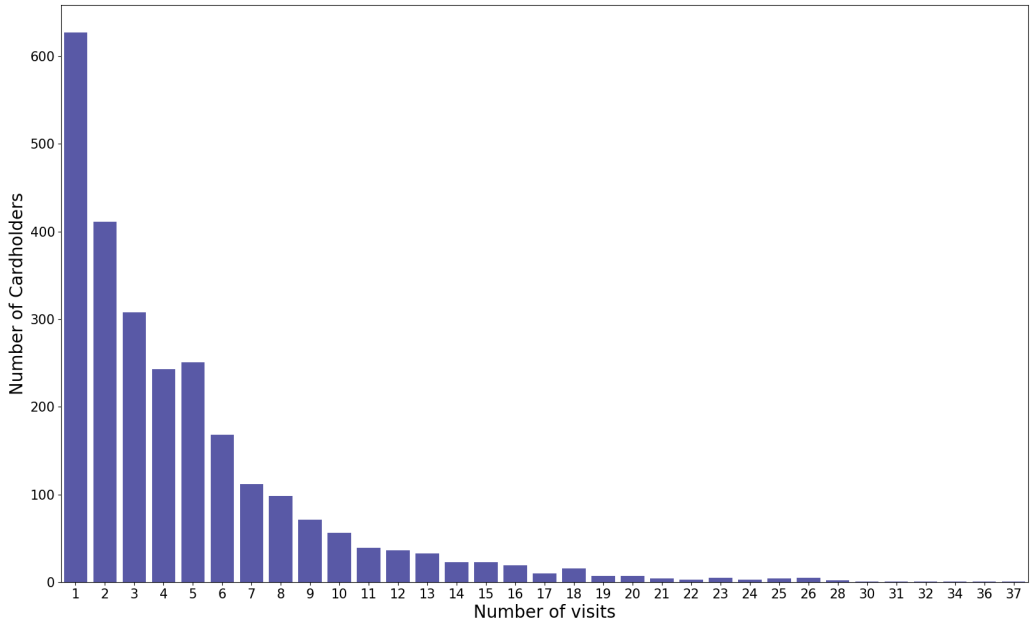
In this section, we delve into the dynamic realm of RFM (Recency, Frequency, and Monetary) analysis, a powerful method for customer segmentation and behavior analysis. RFM provides a comprehensive lens through which we examine customer transactions, focusing on the recency of their purchases, the frequency of interactions, and the monetary value of their expenditures. By dissecting these key dimensions, we aim to uncover distinctive patterns and gain valuable insights into customer segments, enabling us to tailor strategies that resonate with their unique preferences and behaviors. This nuanced approach promises to unveil hidden trends and foster a deeper understanding of our cardholding customer base.

This analysis focuses on scrutinizing the spending and visiting patterns of cardholding customers. Unfortunately, a parallel examination for non-cardholding customers was unfeasible due to the absence of identification linking multiple transactions by the same customer. To maintain data integrity and consistency, we opted to exclude non-cardholder transactional data, representing approximately 37.5% at location 1 and 38.4% at location 2. Given that cardholders accounted for over 60% of transactions at both locations, this strategic exclusion ensures the reliability and accuracy of subsequent analyses.

$$RFMScore = 0.1 * (RecencyRank) + 0.3 * (FrequencyRank) + 0.6 * (MonetaryRank) \quad (5.1)$$



(a) Location 1



(b) Location 2

Figure 5.1: Cardholders Frequency of Visits

In the analysis of cardholders' transactional data, we aggregate information from three key columns: the total amount spent, the frequency of store visits, and the recency of the most recent visit. To mitigate the impact of outliers, each of these columns is transformed into their respective ranks. Subsequently, we calculate the RFM (Recency, Frequency, Monetary) score for each customer

using Equation 5.1. This RFM score serves as the basis for further analytical segmentation, dividing cardholders into five distinct categories.

Equation 5.1 was meticulously designed to align with the specific characteristics of our dataset. The allocation of 10% for recency rank reflects the limited timeframe of our data, where recency of a visit was deemed less influential compared to other factors. Frequency emerged as a crucial factor, evident from Figure 5.1, showcasing a substantial number of cardholders making multiple visits. Monetary value was accorded the highest weight, aligning with our earlier analyses in Section 4.6 that highlighted the significance of heavy spenders among cardholders. This weighting aimed to provide further granularity in distinguishing cardholders based on their spending behavior, offering valuable insights.

Customer Type	**RFM Score Range**
Top Customer	$4.5 < RFMScore \leq 5.0$
High Value Customer	$4.0 < RFMScore \leq 4.5$
Medium Value Customer	$3.0 < RFMScore \leq 4.0$
Low Value Customer	$1.6 < RFMScore \leq 3.0$
Lost Customer	$0.0 \leq RFMScore \leq 1.6$

Table 5.1: Segmentation Criteria based on RFM Score on a 5-Point Scale

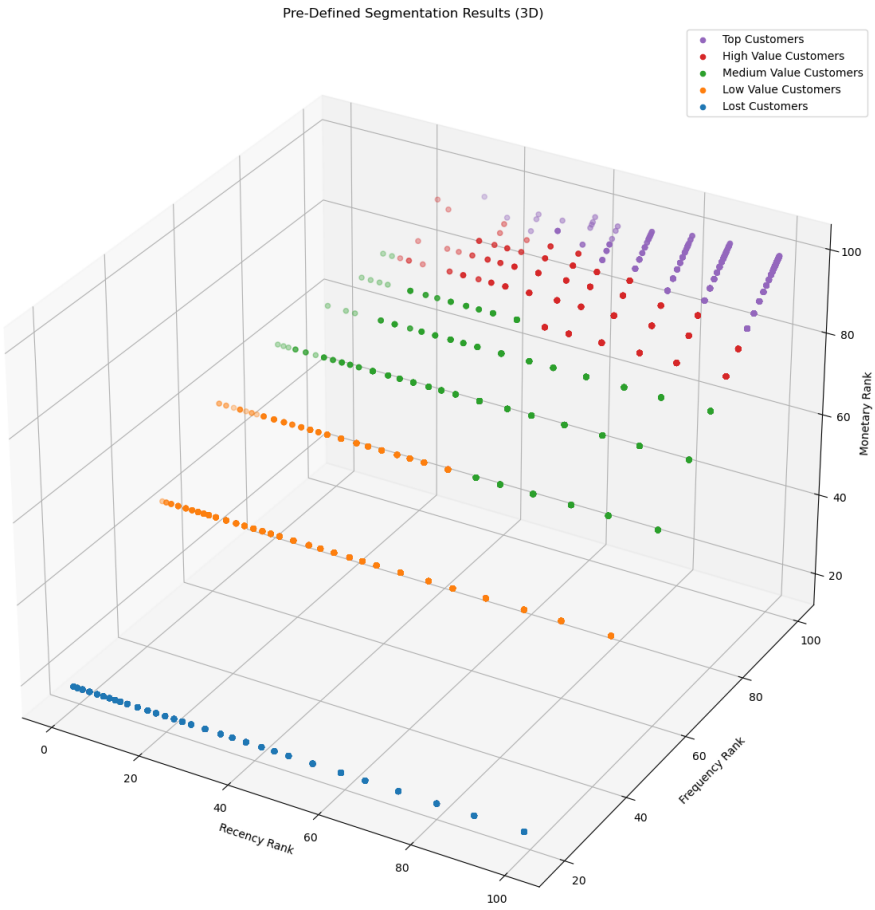
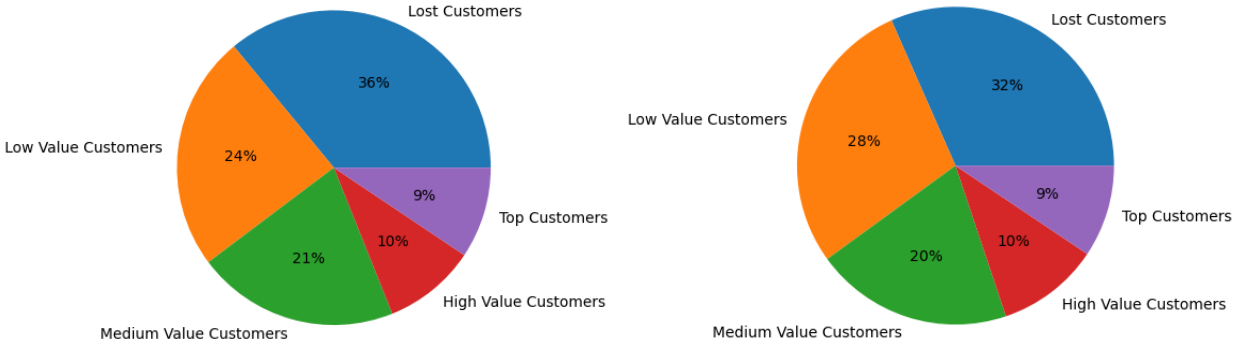


Figure 5.2: Location 1: RFM Score-Based Predefined Segmentation Results

A predefined segmentation of the customer base was carried out utilizing the RFM Score scaled down to a maximum value of 5 instead of 100. Detailed classification criteria can be found in Table 5.1. The objective behind this segmentation was to gain a comprehensive overview of the cardholding customer base and to glean preliminary insights that could guide subsequent clustering techniques.



(a) Location 1 **(b)** Location 2

Figure 5.3: RFM-Based Predefined Customer Segmentation Overview

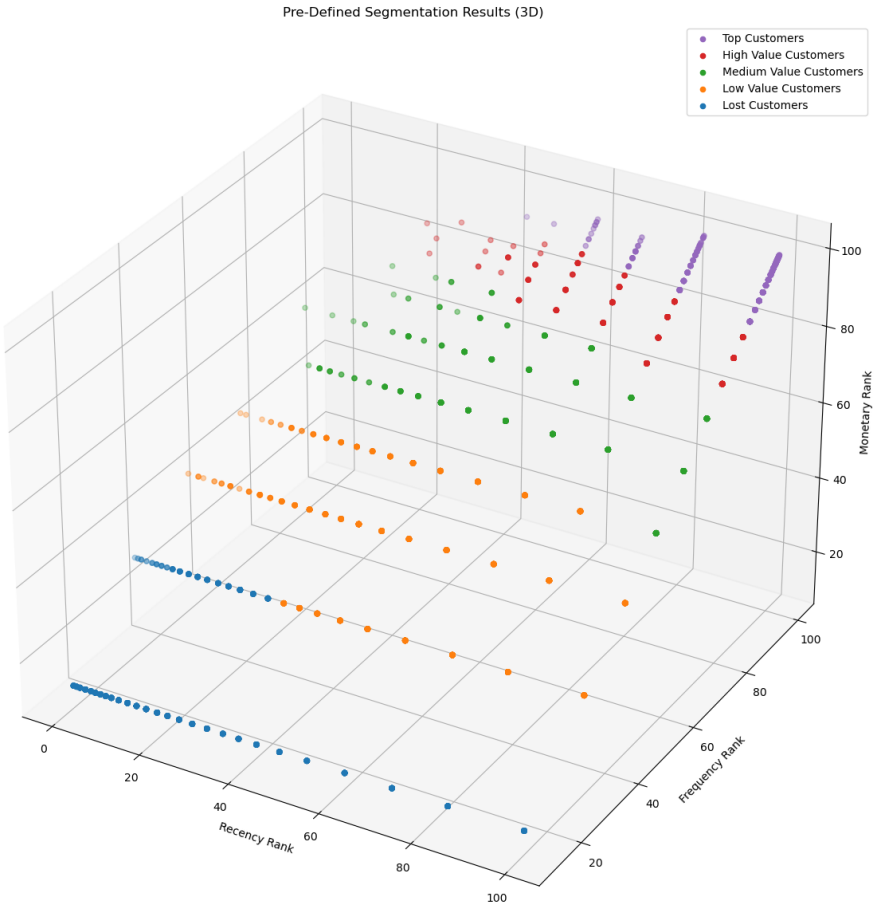


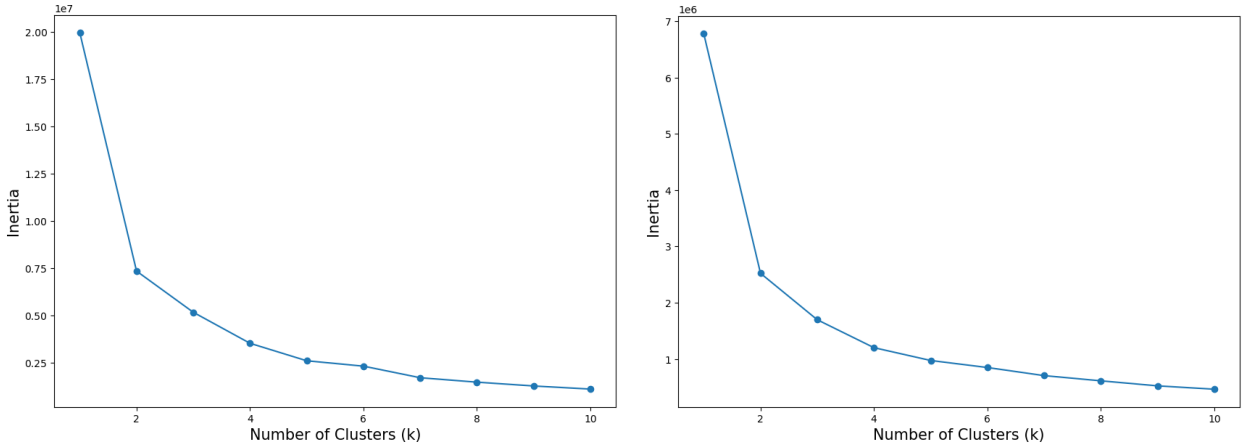
Figure 5.4: Location 2: RFM Score-Based Predefined Segmentation Results

The outcomes of the predetermined segmentation are visualized through pie charts in Figure 5.3 for both locations. To enhance the understanding of the outcomes, additional visual aids are presented by plotting customers in three dimensions across Recency rank, Frequency rank, and Monetary rank. Refer to Figures 5.2 and 5.4 for Locations 1 and 2, respectively.

Location 1 comprises approximately 7,950 cardholders, while Location 2 has around 2,600 cardholders. The distribution of customers across the five categories exhibits minimal variation between the two locations. The notable distinction lies in the 4% increase in Low-Value Customers at Location 2 compared to the 24% of Low-Value Customers at Location 1. This shift is counterbalanced by a decrease in Lost Customers, dropping from 36% at Location 2 to 32% at Location 1.

5.2 RFM into K-Means Clustering

In this section, we build upon the insights gained in Section 5.1. Our primary objective is to compare the results obtained from the RFM predefined segmentation approach with those derived from K-Means Clustering. Instead of predefining customer segments based on the RFM Score from Equation 5.1 and segmentation criteria outlined in Table 5.1, we opt for a K-Means clustering method. This involves utilizing the Recency Rank, Frequency Rank, and Monetary Rank as the input features for the clustering algorithm. The aim is to assess how well the K-Means method aligns with or diverges from the predefined RFM segmentation, providing a comprehensive perspective on customer groupings.



(a) Location 1 **(b)** Location 2

Figure 5.5: Elbow Method Results to find Optimal k-Value

To comprehensively determine the optimal number of clusters for the K-Means Clustering approach, we employed both the Elbow Method and Silhouette Scores calculations. The Elbow Method results, depicted in Figure 5.5, clearly indicate that the optimal number of clusters is 2 for each location. This determination is based on the observation that the elbow in the graph is formed at that specific point, suggesting a distinct transition in the clustering quality.

Examining the Silhouette Scores calculations for both locations, we refer to Table 5.2 for values ranging from 2 to 6 clusters. Notably, the highest Silhouette Score values for both locations are consistently observed at 2 clusters. Based on the combined results, we decided to perform clustering into 2 and 4 clusters, the two highest Silhouette Scores.

	k=2	k=3	k=4	k=5	k=6
Location 1	0.528	0.445	0.505	0.485	0.502
Location 2	0.516	0.435	0.469	0.437	0.438

Table 5.2: Silhouette Scores for K-Means Clustering with Varying Cluster Numbers

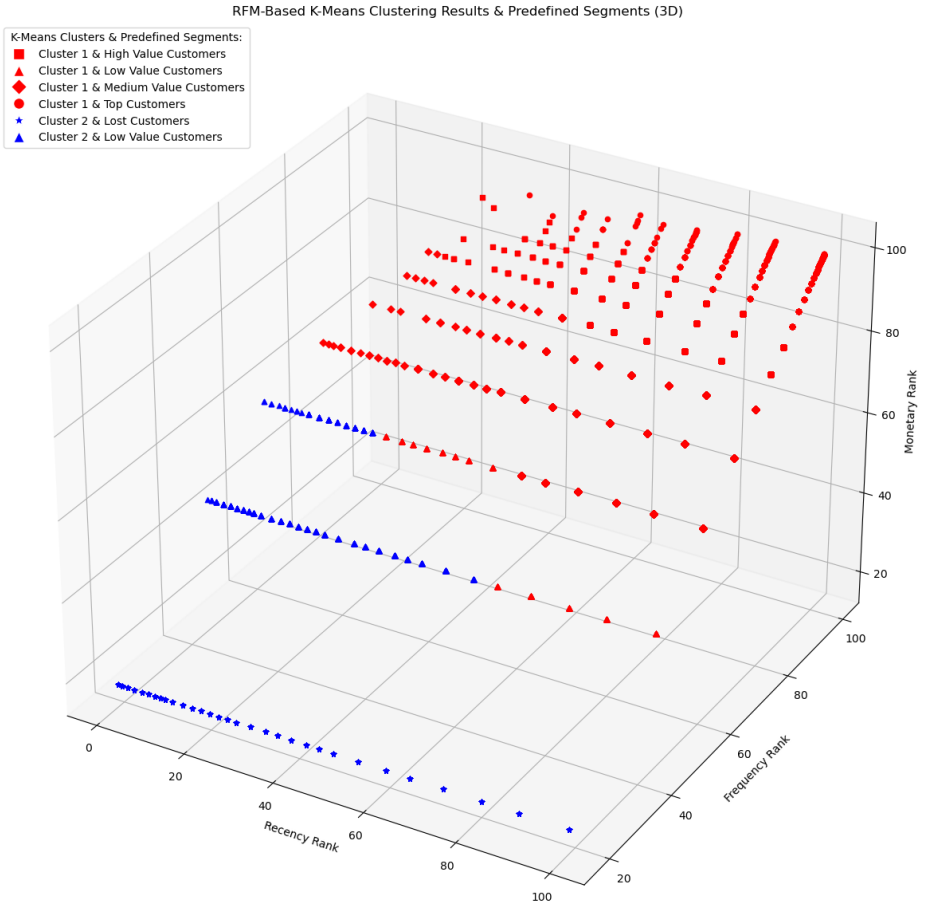


Figure 5.6: RFM-Based K-Means Clustering with k=2 Results for Location 1

Examining the details from Table 5.3 and referencing Figure 5.6, the outcomes of the K-Means clustering into 2 clusters at Location 1 demonstrate meaningful insights. The initial cluster comprises customers who visit frequently, exhibit high spending behavior, and have a low recency score, implying recent visits within the last week of the month. In contrast, the second cluster identifies moderate spenders with lower visit frequencies, distributing their visits more evenly throughout the month, as indicated by the quartile calculations.

Similarly, upon examining the outcomes presented in Table 5.4 and analyzing Figure 5.7, four distinctive clusters emerge at Location 1, each characterized by unique properties. The first and second segments share the characteristic of a single visit, but differ in monetary and recency aspects. The first segment visited in the last week of the month and spent moderately, while the second segment visited in the first two weeks and had the lowest spending among all segments. The third cluster identifies customers who exhibit very high spending and frequent visits, with their last visit occurring in the last two days. The fourth cluster comprises customers with moderate visit

Cluster 1	**Freq.**	**Monetary**	**Recency**
Count	3,950	3,950	3,950
Mean	6.0	264.6	3.5
Std	4.3	214.5	3.6
Min	2.0	4.9	0.0
25%	3.0	112.5	1.0
50%	5.0	202.6	3.0
75%	7.0	357.1	5.0
Max	43.0	2,551.7	24.0

Cluster 2	**Freq.**	**Monetary**	**Recency**
Count	4,000	4,000	4,000
Mean	1.3	65.5	14.5
Std	0.5	69.0	8.1
Min	1.0	0.8	0.0
25%	1.0	22.8	8.0
50%	1.0	45.5	14.0
75%	2.0	84.5	21.0
Max	3.0	1,773.1	30.0

Table 5.3: Descriptive Statistics of RFM-Based K-Means Clustering into 2 Clusters at Location 1

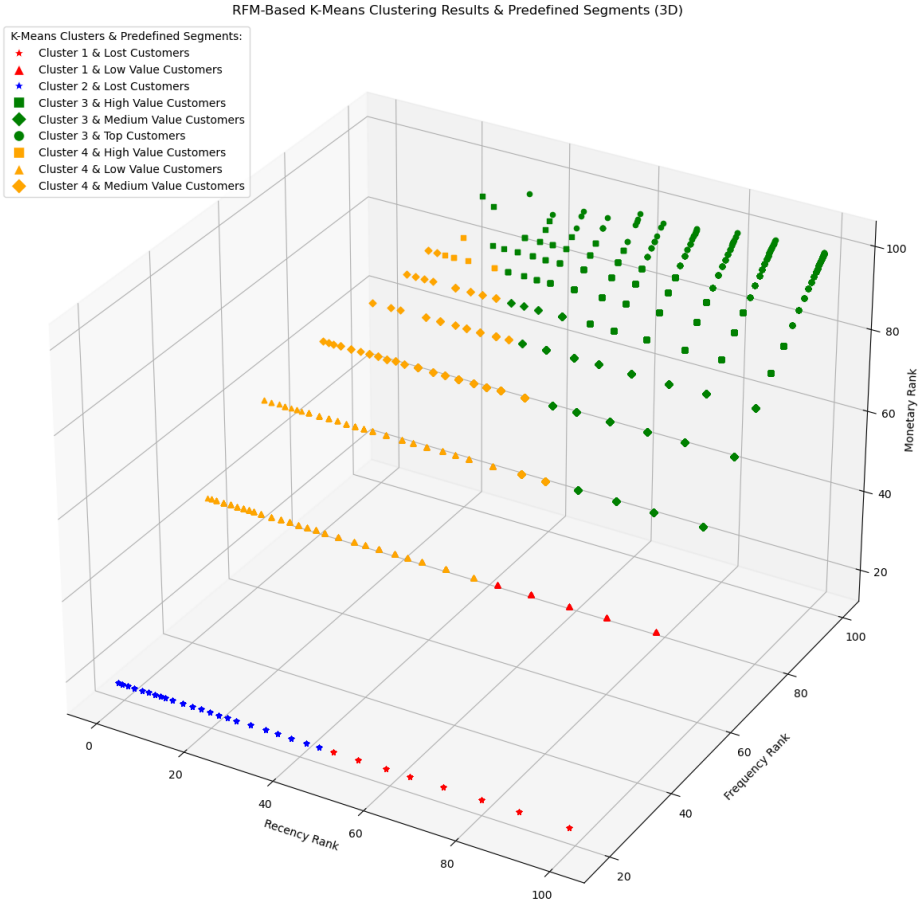


Figure 5.7: RFM-Based K-Means Clustering with k=4 Results for Location 1

frequencies, high spending behavior, and the last visit recorded during the middle two weeks of the month. These nuanced segmentations allow for a granular understanding of customer behaviors within each cluster.

Implementing K-Means Clustering for customers at Location 2 into two clusters yields two segments of equal size with distinct properties, as evident from Table 5.5 and Figure 5.8. The first group of customers is characterized by high frequency and high spending, with each customer making at least two trips during the month. Their recency values average at 2, indicating that most of them visited the store within the last two days. The second cluster comprises customers with low

Cluster 1	**Freq.**	**Monetary**	**Recency**
Count	1,150	1,150	1,150
Mean	1.4	70.5	3.0
Std	0.5	67.3	2.1
Min	1.0	2.0	0.0
25%	1.0	25.4	1.0
50%	1.0	50.5	3.0
75%	2.0	92.7	5.0
Max	2.0	708.9	7.0

Cluster 2	**Freq.**	**Monetary**	**Recency**
Count	2,150	2,150	2,150
Mean	1.0	50.4	19.0
Std	0.0	48.2	6.3
Min	1.0	0.8	8.0
25%	1.0	18.0	13.0
50%	1.0	34.7	19.0
75%	1.0	67.4	25.0
Max	1.0	363.1	30.0

Cluster 3	**Freq.**	**Monetary**	**Recency**
Count	2,650	2,650	2,650
Mean	7.5	313.8	2.1
Std	4.6	229.8	2.1
Min	3.0	16.6	0.0
25%	4.0	147.7	0.0
50%	6.0	256.0	2.0
75%	9.0	417.3	3.0
Max	43.0	2,551.7	14.0

Cluster 4	**Freq.**	**Monetary**	**Recency**
Count	2,000	2,000	2,000
Mean	2.7	141.7	11.3
Std	0.9	123.9	5.6
Min	2.0	3.7	4.0
25%	2.0	58.7	7.0
50%	2.0	106.2	10.0
75%	3.0	187.3	14.0
Max	8.0	1,773.1	30.0

Table 5.4: Descriptive Statistics of RFM-Based K-Means Clustering into 4 Clusters at Location 1

frequency and low spending habits, exhibiting visits spread almost evenly throughout the month, as reflected in the quartile results for recency values. This segmentation provides valuable insights into the varying behaviors of customers within each cluster at Location 2.

Cluster 1	**Freq.**	**Monetary**	**Recency**
Count	1,300	1,300	1,300
Mean	7.6	211.8	2.1
Std	4.7	158.3	2.6
Min	2.0	14.1	0.0
25%	5.0	99.6	0.0
50%	6.0	173.1	1.0
75%	9.0	276.0	3.0
Max	37.0	1,328.0	20.0

Cluster 2	**Freq.**	**Monetary**	**Recency**
Count	1,300	1,300	1,300
Mean	1.8	55.3	12.3
Std	0.9	55.1	7.9
Min	1.0	0.8	0.0
25%	1.0	20.5	6.0
50%	2.0	38.2	11.0
75%	2.0	72.0	18.0
Max	5.0	583.9	30.0

Table 5.5: Descriptive Statistics of RFM-Based K-Means Clustering into 2 Clusters at Location 2

The implementation of four-way clustering at Location 2 revealed intriguing results, summarized in Table 5.6 and illustrated in Figure 5.9. The four distinct customer segments are delineated as follows:

1. The first segment comprises customers with medium frequency and high spending habits, with most of them visiting the store at least once in the last week.

2. The second cluster consists of customers who made, on average, two grocery trips during the month, spending approximately 70 Euros. Notably, all of these customers visited the store once in the last four days.

3. The third cluster encompasses the lowest frequency and lowest spending group, with their

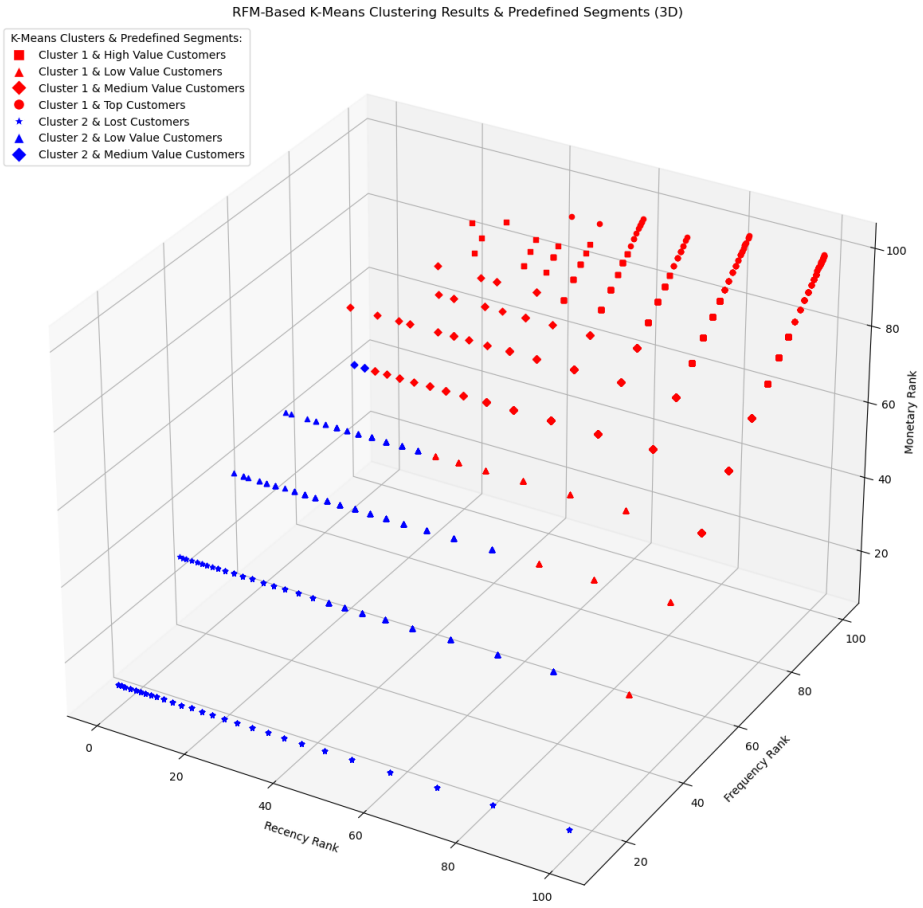


Figure 5.8: RFM-Based K-Means Clustering with k=2 Results for Location 2

Cluster 1	**Freq.**	**Monetary**	**Recency**
Count	660	660	660
Mean	4.4	133.9	6.6
Std	1.3	91.2	4.0
Min	3.0	9.0	2.0
25%	3.0	68.8	3.0
50%	4.0	110.6	6.0
75%	5.0	176.2	9.0
Max	12.0	710.3	22.0

Cluster 2	**Freq.**	**Monetary**	**Recency**
Count	380	380	380
Mean	2.3	70.7	1.3
Std	1.0	65.4	1.2
Min	1.0	4.0	0.0
25%	2.0	26.9	0.0
50%	2.0	51.8	1.0
75%	3.0	92.6	2.0
Max	4.0	551.9	4.0

Cluster 3	**Freq.**	**Monetary**	**Recency**
Count	820	820	820
Mean	1.4	41.6	15.9
Std	0.5	41.4	7.0
Min	1.0	0.8	4.0
25%	1.0	15.9	10.0
50%	1.0	30.2	16.0
75%	2.0	52.0	21.0
Max	3.0	583.9	30.0

Cluster 4	**Freq.**	**Monetary**	**Recency**
Count	740	740	740
Mean	10.0	269.4	1.0
Std	5.0	173.0	1.2
Min	5.0	23.6	0.0
25%	7.0	152.9	0.0
50%	9.0	227.1	1.0
75%	12.0	342.4	2.0
Max	37.0	1,328.0	7.0

Table 5.6: Descriptive Statistics of RFM-Based K-Means Clustering into 4 Clusters at Location 2

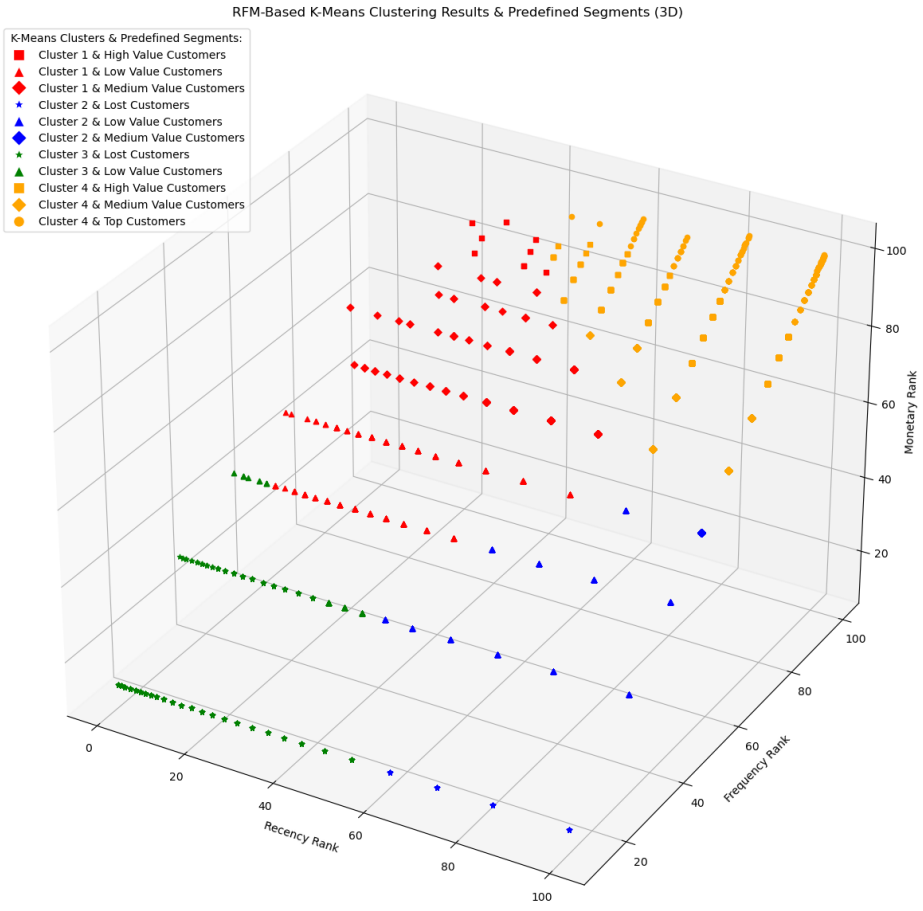


Figure 5.9: RFM-Based K-Means Clustering with k=4 Results for Location 2

average single visit spread evenly across the month, as indicated by the quartile results.

4. The fourth cluster represents the highest frequency and highest spending customer base, with the majority visiting the store on the final day of the month.

These findings offer valuable insights into the diverse behaviors of customers within each of the four clusters at Location 2.

The comparison between K-Means Clustering results derived here and predetermined segmentation using RFM Scores performed in Section 5.1 is conducted in three steps. First insight is provided by the legend labels generated in Figures 5.6, 5.7, 5.8, and 5.9. Second input is provided by the calculations of the average RFM Scores for the different clusters obtained from K-Means Clustering, as outlined in Equation 5.1. The resulting RFM Scores are presented in Table 5.7. Finally, the most vital information is provided by Silhouette Scores for the predefined RFM segmentation cluster.

In the Two-Way Clustering analysis, the K-Means Algorithm efficiently segmented the customer base into two distinctive groups: Low Value Customers and High Value Customers. Notably at Location 1, customers were grouped from Low Value to Top Customers into Cluster 1, while Cluster 2 comprised only Lost and Low Value Customers. At Location 2, Cluster 1 exhibited a similar range, while Cluster 2 included some Medium Value Customers in addition to Lost and Low Value Customers.

Likewise, in the Four-Way Clustering analysis, the algorithm partitioned the customer base

into Lost Customers, Low Value Customers, Medium Value Customers, and High Value Customers. Examining the outcomes for Location 1 in Figure 5.7, it is evident that Clusters 1 and 2 encompassed Low Value and Lost Customers, respectively, while Clusters 3 and 4 captured High Value and Medium Value Customers, respectively. Similarly, for Location 2, Clusters 2 and 3 encompassed Low Value and Lost Customers, respectively, while Clusters 4 and 1 captured High Value and Medium Value Customers, respectively.

A distinct contrast between the two locations becomes evident in the observation that our segmented Medium Value Customers consistently found placement in the upper half of the clusters through K-Means Clustering at Location 1. Specifically, for k=2, they consistently belonged to the high-value cluster, and for k=4, they exclusively occupied the top two high-value clusters. Notably, this pattern did not manifest in the K-Means Clustering results for Location 2. This suggests the presence of a noteworthy threshold between Low Value and Medium Value Customers at Location 1, one that was consistently identified by the K-Means Algorithm.

	Two Clusters		**Four Clusters**			
	First	Second	First	Second	Third	Fourth
Location 1	3.71	1.34	1.63	0.91	4.08	2.65
Location 2	3.71	1.34	2.89	1.98	0.97	4.24

Table 5.7: Average RFM Scores for Clusters from K-Means Clustering

An analysis presented in Table 5.7 uncovered significant variations in the average RFM Score values across different clusters, emphasizing the efficacy of K-Means Clustering in discerning distinct customer segments based on their RFM characteristics. This assertion is further reinforced by the Silhouette Score calculations for the predefined segmentation using RFM Scores.

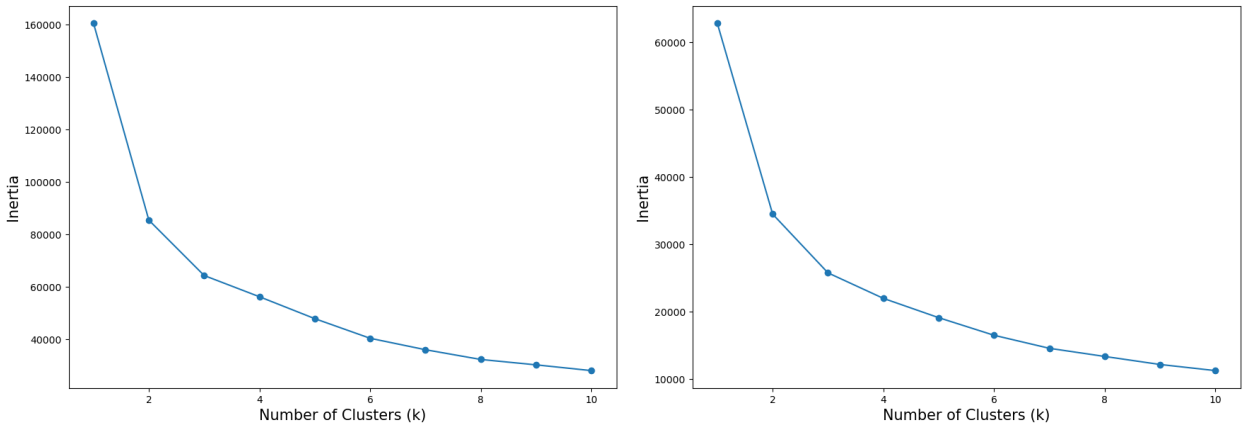
It is essential to highlight that none of the silhouette scores (refer to Table 5.2) for the K-Means Algorithm, spanning k-values up to 6, dip below 0.430, indicating a robust separation between clusters. In contrast, the silhouette scores for the predefined segmentation clusters stood at 0.330 and 0.291 for Location 1 and 2, respectively.

Subsequent to this observation, an intriguing test was conducted to assess the comparative strength and validity of our K-Means Clustering results for k=2 and k=4 in both locations against our predefined segmentation approach. A new balanced RFM Score was computed, assigning equal one-third weight to all three features: Recency Rank, Frequency Rank, and Monetary Rank. Subsequently, 2 and 4 clusters were formed based on equal one-half and one-fourth divisions of the balanced RFM Score, respectively. Interestingly, the silhouette scores for the 2 predefined clusters were 0.528 and 0.516 for Locations 1 and 2, respectively. Remarkably, these values aligned precisely with the silhouette scores for K-Means Clustering with k=2 for the respective locations.

However, the superiority of the K-Means Clustering algorithm became more evident when we examined the silhouette scores for 4 predefined clusters. These predefined clusters only achieved silhouette scores of 0.397 and 0.372 for Locations 1 and 2, respectively, significantly trailing behind the performance of K-Means. This starkly underscores the clear dominance of the K-Means Algorithm over our predefined segmentation approach in adeptly capturing and distinguishing inherent patterns within the data, particularly with the increase in the number of clusters.

5.3 Transactional Data Focused K-Means Clustering

In this section, our attention turns to the application of K-Means Clustering, concentrating on four key features extracted from the transactional dataset: Items bought on Offer, Total Items bought, Total Amount Spent, and Cardholding Status. Due to the presence of outliers in the Total Items bought and Total Amount Spent columns, scaling becomes imperative, considering the sensitivity of the K-Means Clustering Algorithm to outliers. To address this, we opted for Robust Scaling, a technique designed to mitigate the impact of outliers on the clustering process.



(a) Location 1 **(b)** Location 2

Figure 5.10: Elbow Method Results to find Optimal k-Value

Before delving into cluster formulation and categorization, a crucial step involves determining the optimal number of clusters. Utilizing the Elbow Method and Silhouette Scores, we aim to define well-distinguished and distinct clusters. The Elbow Method results are showcased in Figure 5.10 for both locations. These results suggest that our customer base can be most effectively categorized into two or three clusters. Silhouette Scores further substantiate this, with Location 1 yielding scores of 0.521 and 0.290 for two and three clusters, and Location 2 providing scores of 0.495 and 0.297 for two and four clusters, respectively.

Conducting K-Means Clustering for the entire transactional dataset proved to be challenging, given the substantial volume of distinct transactions at both locations. Location 1 boasted 45,700 distinct transactions, while Location 2 featured 19,800 distinct transactions, highlighting the complexity of the analysis. Despite the challenges, the effort yielded valuable and insightful results.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	1.87	9.87	22.48	20.21	2.67	0.58
Std	1.90	7.00	16.85	21.31	2.40	0.49
Min	0.00	1.00	0.15	0.00	0.15	0.00
25%	0.00	4.00	9.28	0.00	1.62	0.00
50%	1.00	8.00	18.50	15.38	2.19	1.00
75%	3.00	14.00	31.88	30.00	2.98	1.00
Max	12.00	53.00	128.34	100.00	112.19	1.00

Table 5.8: Descriptive Statistics of First Cluster of K-Means Clustering with k=2 at Location 1

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	8.66	38.03	88.72	25.54	2.41	0.82
Std	6.08	17.76	48.34	17.47	1.15	0.38
Min	0.00	4.00	6.83	0.00	0.19	0.00
25%	5.00	26.00	56.51	12.50	1.86	1.00
50%	7.00	34.00	78.03	21.74	2.28	1.00
75%	11.00	46.00	109.12	34.78	2.77	1.00
Max	62.00	425.00	1,213.42	100.00	43.80	1.00

Table 5.9: Descriptive Statistics of Second Cluster of K-Means Clustering with k=2 at Location 1

In Tables 5.8 and 5.9, we present descriptive statistical values for the two clusters resulting from K-Means Clustering at Location 1. The algorithm grouped 34,900 transactions into Cluster 1 and 10,800 into Cluster 2. The two clusters exhibit notable distinctions across various parameters. Transactions in Cluster 2 are characterized by high monetary values, with an average amount of 88 Euros per transaction, compared to a minimal average amount of 22 Euros per transaction in Cluster 1. In Cluster 2, 82% of transactions were performed by cardholders, contrasting with the 58% cardholder involvement in Cluster 1. Despite higher monetary values in Cluster 2, the amount per item is 25 cents less than in Cluster 1. These well-defined clusters showcase distinct properties and contribute valuable insights.

Applying K-Means Clustering with k=2 at Location 2 resulted in clusters outlined in Tables 5.10 and 5.11. The clustering outcomes closely resemble those observed at Location 1. Cluster 1 comprises 5,050 transactions characterized by substantial shopping and high spending, with approximately 82% of them conducted by cardholders. In contrast, Cluster 2 classifies 14,750 transactions associated with lighter shopping and lower spending, with around 45% performed by non-cardholders. Notably, Cluster 1 exhibits a higher proportion of items bought on offer (32%) compared to Cluster 2, where only about 23.5% of items were purchased on offer.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	7.06	24.03	52.50	32.02	2.28	0.82
Std	5.09	11.62	28.51	20.71	1.54	0.39
Min	0.00	1.00	9.19	0.00	0.30	0.00
25%	4.00	17.00	34.87	16.00	1.74	1.00
50%	6.00	21.00	45.69	28.81	2.14	1.00
75%	9.00	28.00	62.10	44.44	2.61	1.00
Max	47.00	165.00	428.08	100.00	90.40	1.00

Table 5.10: Descriptive Statistics of First Cluster of K-Means Clustering with k=2 at Location 2

Transitioning to the next phase, K-Means Clustering is applied to the same dataset, this time forming three distinct clusters instead of two. Descriptive properties of the three clusters formed at Location 1 are outlined in Tables 5.12, 5.13, and 5.14. The algorithm segmented approximately 14,700, 21,600, and 9,400 transactions into Clusters 1, 2, and 3, respectively. In this scenario, the clusters are intriguingly delineated into three categories: Low Spending Non-Cardholder Transactions, Low Spending Cardholder Transactions, and High Monetary Transactions.

Cluster 1, outlined by the statistics in Table 5.12, characterizes the lowest monetary transactions,

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	1.57	7.06	14.68	23.54	2.31	0.55
Std	1.72	4.26	9.89	26.05	1.85	0.50
Min	0.00	1.00	0.09	0.00	0.09	0.00
25%	0.00	4.00	6.89	0.00	1.49	0.00
50%	1.00	6.00	12.76	16.67	1.97	1.00
75%	2.00	10.00	20.62	37.50	2.59	1.00
Max	10.00	30.00	77.79	100.00	64.57	1.00

Table 5.11: Descriptive Statistics of Second Cluster of K-Means Clustering with k=2 at Location 2

all made by non-cardholders. This cluster also exhibits the lowest average percentage of items bought on offer, approximately 18.3%. Cluster 2, detailed in Table 5.13, features low monetary transactions, with 99% of them made by cardholders. These transactions consist of an average of 12 total items, costing a total of 27.70 Euros on average, with about 21.9% of those items bought on offer.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	1.41	7.88	17.81	18.26	2.64	0.00
Std	1.64	6.56	15.82	21.93	2.62	0.00
Min	0.00	1.00	0.15	0.00	0.15	0.00
25%	0.00	3.00	5.79	0.00	1.53	0.00
50%	1.00	6.00	12.86	12.50	2.14	0.00
75%	2.00	11.00	25.26	27.27	2.97	0.00
Max	11.00	50.00	110.68	100.00	99.20	0.00

Table 5.12: Descriptive Statistics of First Cluster of K-Means Clustering with k=3 at Location 1

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	2.40	12.11	27.66	21.87	2.67	0.99
Std	2.18	7.57	18.28	20.64	2.18	0.11
Min	0.00	1.00	0.35	0.00	0.16	0.00
25%	1.00	6.00	13.44	6.25	1.69	1.00
50%	2.00	11.00	23.73	16.67	2.24	1.00
75%	4.00	17.00	38.16	33.33	2.98	1.00
Max	13.00	62.00	128.34	100.00	112.19	1.00

Table 5.13: Descriptive Statistics of Second Cluster of K-Means Clustering with k=3 at Location 1

Lastly, Cluster 3 at Location 1 consists of high monetary transactions, with an average of 93.75 Euros spent per transaction. This cluster comprises approximately 81% of the transactions made by cardholders, with the highest overall percentage (about 25.6%) of items being bought on offer. As a result, these transactions have the lowest per-item cost, approximately 25 cents less than the averages from the other two clusters.

Moving on to the results of K-Means Clustering at Location 2 with k=3 as detailed by the Tables 5.15, 5.16, and 5.17, a similar trend as Location 1 is observed. The algorithm segmented approximately 6,200, 3,800, and 9,800 transactions into three categories: Low Spending Non-Cardholder

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	9.17	40.09	93.76	25.56	2.42	0.81
Std	6.27	17.95	49.23	17.30	1.17	0.39
Min	0.00	4.00	11.46	0.00	0.19	0.00
25%	5.00	28.00	60.92	12.50	1.87	1.00
50%	8.00	36.00	83.26	21.88	2.28	1.00
75%	12.00	48.00	114.94	34.78	2.77	1.00
Max	62.00	425.00	1,213.42	100.00	43.80	1.00

Table 5.14: Descriptive Statistics of Third Cluster of K-Means Clustering with k=3 at Location 1

Transactions, High Monetary Transactions, and Low Spending Cardholder Transactions respectively.

Clusters 1 and 3 consist of low monetary transactions, with the key distinction being that Cluster 3 comprises about 92% transactions made by Cardholders, while Cluster 1 includes almost 99% transactions made by Non-Cardholders. Cluster 3 exhibits a higher average number of total items bought, average amount spent, and percentage of items bought on offer when compared to Cluster 1. This indicates that Cardholders, even when making small purchases, tend to spend more than Non-Cardholders.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	0.83	5.28	10.73	17.35	2.25	0.01
Std	1.12	3.61	8.48	25.20	1.79	0.11
Min	0.00	1.00	0.09	0.00	0.09	0.00
25%	0.00	2.00	4.05	0.00	1.38	0.00
50%	0.00	5.00	8.49	0.00	1.90	0.00
75%	1.00	7.00	15.19	25.00	2.57	0.00
Max	7.00	25.00	58.33	100.00	33.26	1.00

Table 5.15: Descriptive Statistics of First Cluster of K-Means Clustering with k=3 at Location 2

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	7.86	26.55	58.27	32.22	2.27	0.80
Std	5.42	12.10	29.92	20.61	0.87	0.40
Min	0.00	6.00	10.86	0.00	0.30	0.00
25%	4.00	19.00	39.52	16.06	1.76	1.00
50%	7.00	23.00	51.48	29.17	2.17	1.00
75%	10.00	30.00	69.04	44.44	2.62	1.00
Max	47.00	165.00	428.08	100.00	14.70	1.00

Table 5.16: Descriptive Statistics of Second Cluster of K-Means Clustering with k=3 at Location 2

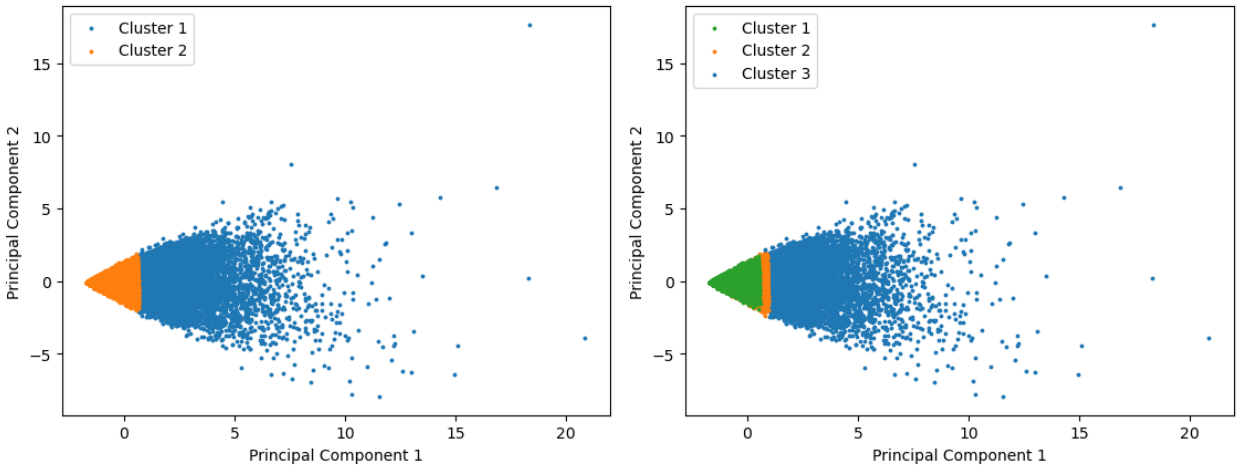
Examining the high-spending transactions cluster at Location 2, it is evident that approximately 80% of these transactions are made by cardholders, mirroring the trend observed at Location 1. These transactions are distinguished by their high average numbers, including around 26 total items bought per transaction, €58 spent per transaction, and 32.2% of items bought on offer.

To facilitate a more concise and insightful representation of the data, Principal Component

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	2.40	9.28	19.53	28.43	2.34	0.92
Std	2.10	4.83	11.23	25.24	2.02	0.27
Min	0.00	1.00	0.84	0.00	0.30	0.00
25%	1.00	5.00	10.73	8.33	1.58	1.00
50%	2.00	9.00	17.51	25.00	2.02	1.00
75%	4.00	13.00	26.74	42.86	2.59	1.00
Max	11.00	34.00	90.40	100.00	90.40	1.00

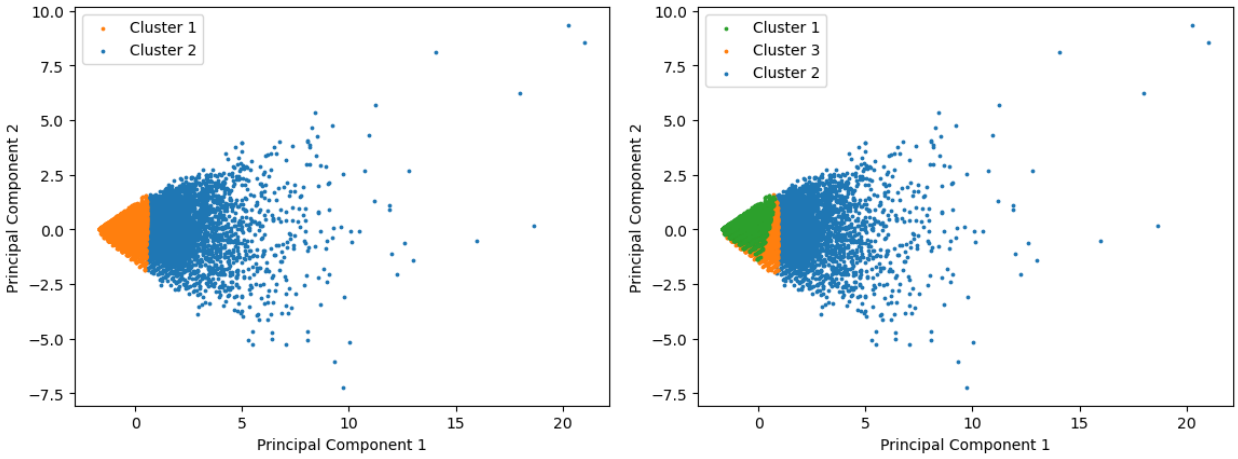
Table 5.17: Descriptive Statistics of Third Cluster of K-Means Clustering with k=3 at Location 2

Analysis (PCA) was employed on the four selected features, reducing the dimensionality to two principal components. The objective was to provide visualization and enhance comprehension of the K-Means Clustering outcomes.



(a) 2 Clusters **(b)** 3 Clusters

Figure 5.11: Location 1: Transactional Data K-Means Clustering results plotted on Principal Components



(a) 2 Clusters **(b)** 3 Clusters

Figure 5.12: Location 2: Transactional Data K-Means Clustering results plotted on Principal Components

In Location 1, the variance was distributed as follows: Principal Component 1 explained 73.8%, and Principal Component 2 contributed 16.3%, resulting in a cumulative variance of 90.1%. Similarly, at Location 2, Principal Components 1 and 2 accounted for 74.8% and 14.6% of the explained variances, respectively, yielding a cumulative explanatory power of 89.4%.

The clustering outcomes for Location 1 are visually represented in Figure 5.11, juxtaposed with the corresponding Principal Components. Figure 5.11 (a) exhibits a distinct separation along the Principal Component 1 axis. Furthermore, in Figure 5.11 (b), a pronounced separation is once again evident on the Principal Component 1 axis.

The clustering results for Location 2 are visually depicted in Figure 5.12, juxtaposed with the corresponding Principal Components. K-Means Clustering into 2 clusters exhibits similar outcomes to Location 1, where the primary distinction is driven by Principal Component 1, as illustrated in Figure 5.12 (a). A dissimilarity is observed when examining the results for 3 clusters, as depicted in Figure 5.12 (b), especially when compared to the first location. While distinct clustering is heavily influenced by Principal Component 1, the separation between Clusters 1 and 3 is also contributed to by Principal Component 2.

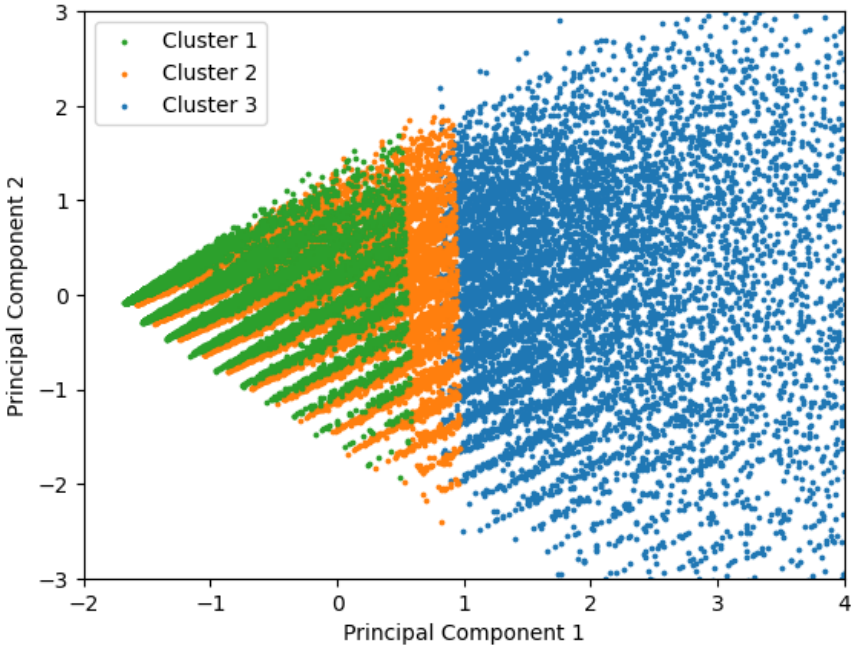


Figure 5.13: Location 1: Deeper Look into Cluster Separation from Figure 5.11(b)

Figure 5.11 illustrates that the K-Means Clustering results with k=3 at Location 1 are primarily influenced by Principal Component 1. However, a more nuanced insight is revealed when examining Clusters in Figure 5.13. The initial separation of Cluster 3 from Clusters 1 and 2 is clear, but Clusters 1 and 2, representing the low-expenditure clusters at Location 1, exhibit substantial overlap. It is evident that the distinction between them is shaped by a combination of factors beyond just the first and second principal components, specifically influenced by a principal component capturing the most variance in cardholding status.

Continuing into the similar investigation of Clusters 1 and 3 at Location 2, Figure 5.13 presents a similar insight into the clustering trend as observed at Location 1. Clusters 1 and 3 observe a distinction between them that cannot be explained solely by the first and the second principal

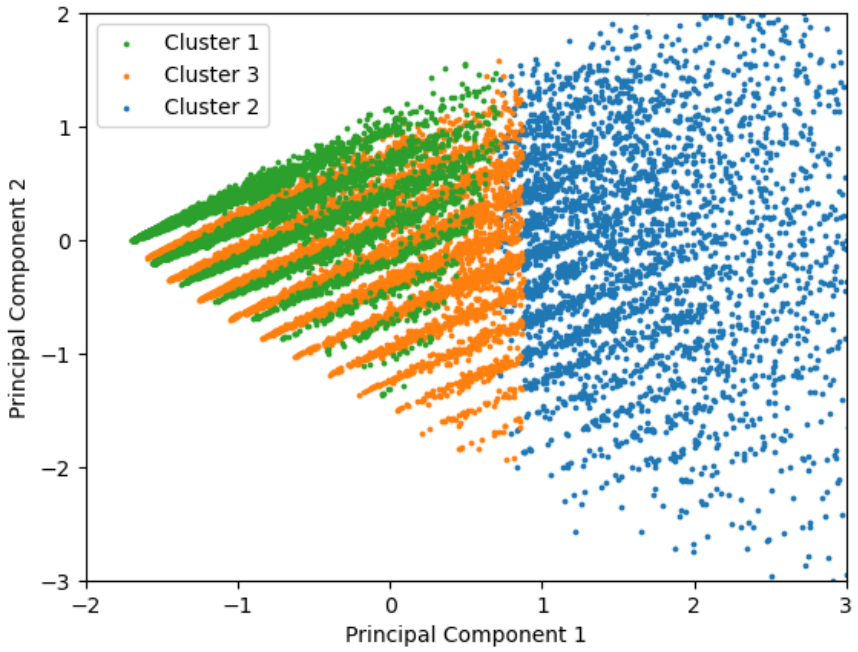


Figure 5.14: Location 2: Deeper Look into Cluster Separation from Figure 5.12(b)

components.

In summary, the K-Means clustering analysis on transactional data and PCA-reduced visualizations of the obtained clusters at both locations have provided valuable insights into customer behavior. The clusters reveal distinct patterns in spending, frequency, and preferences among cardholders and non-cardholders. The identified clusters offer a nuanced understanding of customer segments, aiding in targeted strategies for marketing, promotions, and customer engagement.

5.4 Transactional Data Focused DBSCAN

In this section, we aim to compare the outcomes of our K-Means Clustering results with k=3, as discussed in Section 5.3, against a more sophisticated clustering algorithm commonly employed in machine learning and data analysis: Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The primary objective of this comparison is to assess the effectiveness of our clusters and explore potential improvements.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	1.35	7.29	16.70	18.25	2.61	0.00
Std	1.53	5.70	14.19	21.82	2.21	0.00
Min	0.00	1.00	0.15	0.00	0.15	0.00
25%	0.00	3.00	5.56	0.00	1.55	0.00
50%	1.00	6.00	12.28	12.50	2.15	0.00
75%	2.00	11.00	24.02	27.27	2.96	0.00
Max	7.00	27.00	69.66	100.00	53.18	0.00

Table 5.18: Descriptive Statistics of First Cluster of DBSCAN at Location 1

Tables 5.18, 5.19, and 5.20 present a foundational set of descriptive statistics for the clusters

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	2.61	12.48	28.76	22.27	2.63	1.00
Std	2.32	7.63	18.61	20.07	1.87	0.00
Min	0.00	1.00	0.35	0.00	0.20	1.00
25%	1.00	6.00	13.78	7.69	1.73	1.00
50%	2.00	11.00	24.72	18.18	2.24	1.00
75%	4.00	18.00	40.51	33.33	2.94	1.00
Max	10.00	35.00	87.67	100.00	49.60	1.00

Table 5.19: Descriptive Statistics of Second Cluster of DBSCAN at Location 1

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	8.69	39.82	92.14	24.41	2.55	0.72
Std	6.65	18.47	51.33	19.26	2.71	0.45
Min	0.00	1.00	5.09	0.00	0.15	0.00
25%	4.00	28.00	58.62	9.83	1.77	0.00
50%	8.00	37.00	84.82	20.00	2.27	1.00
75%	12.00	48.00	115.05	34.62	2.83	1.00
Max	62.00	425.00	1,213.42	100.00	112.19	1.00

Table 5.20: Descriptive Statistics of "Noise" Transactions of DBSCAN at Location 1

resulting from the DBSCAN analysis at Location 1. The algorithm segmented approximately 14,100, 22,200, and 9,400 transactions into Cluster 1, Cluster 2, and Noise, respectively. In comparison, K-Means Clustering had allocated 14,700, 21,600, and 9,400 transactions to the corresponding clusters replacing Noise with Cluster 3. Notably, both methods delineated the transactions into the same three categories: Low Spending Non-Cardholder Transactions, Low Spending Cardholder Transactions, and High Monetary Transactions. DBSCAN classified High Monetary Transactions as Noise.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	0.72	4.99	9.98	16.26	2.21	0.00
Std	0.90	3.30	7.46	24.25	1.63	0.00
Min	0.00	1.00	0.09	0.00	0.09	0.00
25%	0.00	2.00	3.99	0.00	1.38	0.00
50%	0.00	4.00	7.98	0.00	1.89	0.00
75%	1.00	7.00	14.39	25.00	2.55	0.00
Max	3.00	16.00	34.79	100.00	24.90	0.00

Table 5.21: Descriptive Statistics of First Cluster of DBSCAN at Location 2

Similar to the DBSCAN results for Location 1, Tables 5.21, 5.22, and 5.23 present a fundamental set of descriptive statistics for the clusters at Location 2. Continuing alike trends, both methods at Location 2 characterized the transactions into the same three categories: Low Spending Non-Cardholder Transactions, High Monetary Transactions, and Low Spending Cardholder Transactions.

The DBSCAN algorithm, in this case as well, classified transactions into three categories: Cluster 1, Cluster 2, and Noise. Here's a breakdown of the results compared against the K-Means Clustering results:

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	1.81	7.64	15.93	26.14	2.29	1.00
Std	1.54	3.87	8.67	24.37	1.41	0.00
Min	0.00	1.00	0.15	0.00	0.15	1.00
25%	1.00	5.00	9.08	6.25	1.59	1.00
50%	2.00	7.00	14.57	22.22	2.02	1.00
75%	3.00	10.00	21.91	40.00	2.58	1.00
Max	5.00	18.00	40.69	100.00	29.80	1.00

Table 5.22: Descriptive Statistics of Second Cluster of DBSCAN at Location 2

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	6.19	21.04	45.90	33.26	2.39	0.71
Std	4.86	11.61	28.06	23.87	2.21	0.45
Min	0.00	1.00	3.31	0.00	0.30	0.00
25%	3.00	14.00	28.85	13.64	1.68	0.00
50%	5.00	19.00	40.54	29.17	2.11	1.00
75%	8.00	25.00	55.70	50.00	2.63	1.00
Max	47.00	165.00	428.08	100.00	90.40	1.00

Table 5.23: Descriptive Statistics of "Noise" Transactions of DBSCAN at Location 2

Cluster 1 (Low Spending Non-Cardholder Transactions):

- DBSCAN: Approximately 5,700 transactions
- K-Means: Approximately 6,200 transactions

Cluster 2 (Low Spending Cardholder Transactions):

- DBSCAN: Approximately 7,400 transactions
- K-Means (Cluster 3 - Table 5.17): Approximately 9,800 transactions

Noise (Potentially High Monetary Transactions):

- DBSCAN: Approximately 6,700 transactions
- K-Means (Cluster 2 - Table 5.16): Approximately 3,800 transactions

It's worth noting that DBSCAN classified transactions that it couldn't assign to either Cluster 1 or Cluster 2 as "Noise." In contrast, K-Means had a dedicated cluster for what seems to be the equivalent of High Monetary Transactions for both the locations.

The general characteristics of the clusters identified by DBSCAN closely resemble those produced by K-Means Clustering, with one notable distinction. DBSCAN exhibited a more stringent classification of low-expenditure transactions into distinct categories of cardholders and non-cardholders, surpassing the approximately 99% and 92% threshold achieved by K-Means at Locations 1 and 2 respectively. However, this enhanced separation came with less cardholder transactions being classified into the High Monetary Transactions segments, which now comprise 72% and 71% transactions by cardholders compared to the earlier 81% and 80% when analyzed with K-Means respectively for Locations 1 and 2.

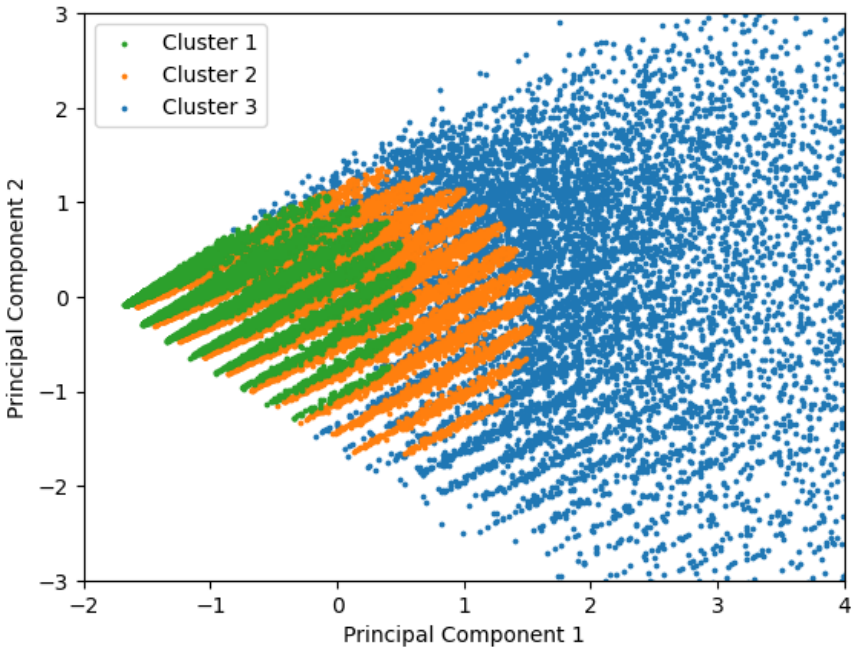


Figure 5.15: Location 1: Transactional Data DBSCAN Results plotted on Principal Components

For a more detailed comparison between the results of DBSCAN and K-Means Clustering with k=3, we examine the sizes of each cluster. Table 5.24 provides a side-by-side comparison of the cluster sizes for both locations. It is important to note that the category of Potentially High Monetary Transactions, which corresponded to a cluster from K-Means Clustering results emerged as Noise from DBSCAN. At Location 1, the number of transactions classified as High Monetary Transactions remains consistent across both algorithms. However, the count for Low Spending Non-Cardholder Transactions decreases by 600 in DBSCAN compared to K-Means Clustering, while the count for Low Spending Cardholder Transactions increases by the same amount in DBSCAN.

	Location 1		**Location 2**	
Category	**K-Means**	**DBSCAN**	**K-Means**	**DBSCAN**
Low Spending Non-Cardholder Transac...	14,700	14,100	6,200	5,700
Low Spending Cardholder Transactions	21,600	22,200	9,800	7,400
Potentially High Monetary Transactions	9,400	9,400	3,800	6,700

Table 5.24: Cluster Size Comparison: K-Means Clustering with k=3 & DBSCAN

The disparities between the two algorithms are more pronounced for Location 2 than for Location 1. In Location 2, there is a substantial increase in the count of High Monetary Transactions when transitioning from K-Means Clustering to DBSCAN. After DBSCAN, about one-thirds of the total data is classified as Noise which corresponded to an additional 2,900 transactions being classified as High Monetary Transactions. However, the average amount spent by this cluster decreases from 58 Euros (K-Means) to 46 Euros, which is still relatively high for this location. The other two clusters experience a reduction in size, with Low Spending Non-Cardholder Transactions following a similar trend as in Location 1, dropping by 500, and Low Spending Cardholder Transactions exhibiting an opposite trend and decreasing by 2,400 transactions.

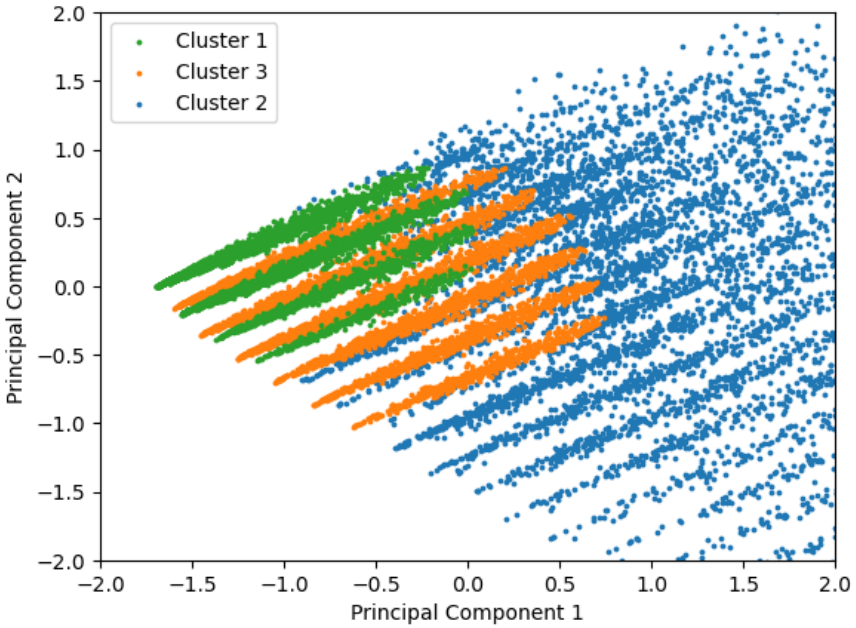


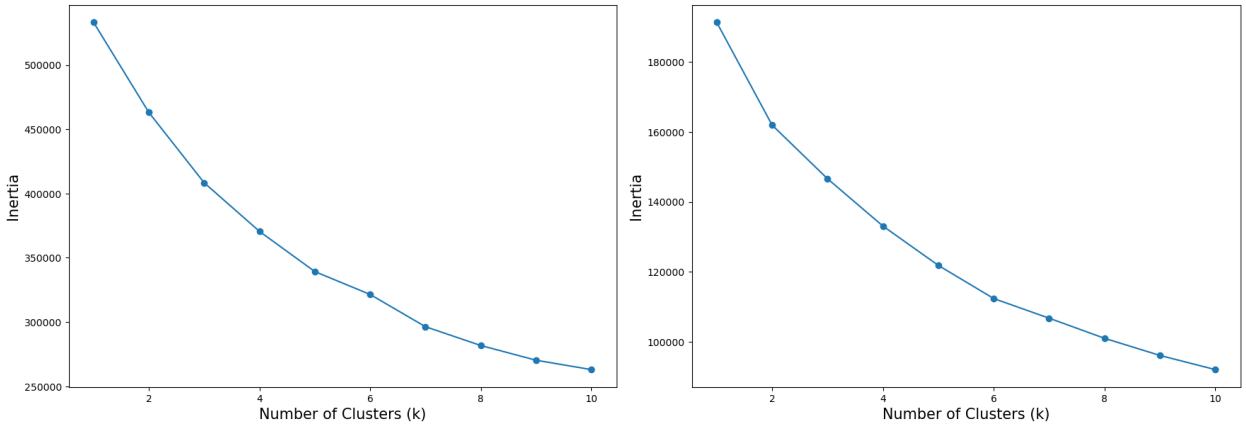
Figure 5.16: Location 2: Transactional Data DBSCAN Results plotted on Principal Components

A comparison of DBSCAN and K-Means Clustering for our transactional dataset is evident when examining Figures 5.15 and 5.16, and contrasting them with Figures 5.13 and 5.14. K-Means Clustering, being a distance-based technique, exhibits a straight line cut-off criterion, primarily visible along the first principal component. In contrast, DBSCAN, a density-based clustering algorithm, forms curved cuts.

The difference in results between DBSCAN and K-Means suggests that the two algorithms have different interpretations of the data. DBSCAN, being density-based, may be more conservative in forming clusters and may be sensitive to variations in data density. It is observed that DBSCAN algorithm was more stringent in separating Low Spending Transactions into Cardholders and Non-Cardholders. In the case of DBSCAN, it also categorized a portion of transactions, potentially High Monetary, as Noise. The usual association of noise in data as a bad thing doesn't hold true for our specific goals as high monetary transactions are always the most important transactions for a retailer. Therefore, understanding the characteristics of the data and the behavior of each algorithm is essential in interpreting and validating clustering results because if the DBSCAN "Noise" Transactions are classified as a cluster, the obtained results from both the clustering techniques are highly comparable.

5.5 Sectoral Shopping Clustering using K-Means

In this section, the primary objective was to categorize shopping habits according to various sectors in which customers made purchases. Before diving into the details of the clustering process, certain steps of feature engineering were undertaken. Specifically, upon examining Table 4.2 and drawing insights from our previous analyses, a decision was made to merge Sectors 6 and 17 (Butcher Shop and Fish Shop), Sectors 10, 11, and 12 (Christmas Celebration, Celebrations, and Easter Celebrations), as well as Sectors 15, 20, and 21 (Delicatessen Cutting Counter, Cured Meat Cutting Counter, and Cheese Cutting Counter) due to their close association with each other.



(a) Location 1 **(b)** Location 2

Figure 5.17: Elbow Method Results for K-Means Sector Clustering to find Optimal k-Value

In this instance, K-Means Clustering was applied to a set of seven features. Utilizing the transactional data from the preceding Sections 5.3 and 5.4, two principal components were derived, explaining approximately 90% of the total variance. Additionally, from the reduced set of fifteen features representing the total items bought in a sector on a per-transactional basis, as previously employed in Section 4.3, five principal components were extracted, accounting for about 90% of the total variance. This resulted in a set of seven features that were utilized for the K-Means Clustering analysis in this context.

The determination of the optimal number of clusters relied on the Elbow Method, Silhouette Scores, and actual K-Means Clustering to assess the distinctiveness between clusters and finalize the k-value. The results of the Elbow Method, depicted in Figure 5.17 for both locations, did not exhibit a clear elbow formation. Upon calculating Silhouette scores for various cluster numbers at both locations, it was noted that the most suitable number of clusters was either 2 or 3. The choice to proceed with three clusters was informed by the distinctive characteristics observed in the descriptive statistics of the three clusters formed at both locations. As these clusters demonstrated sufficient uniqueness, the decision was made to proceed with k=3.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	2.04	9.60	21.74	21.24	2.69	0.58
Std	2.33	7.42	17.33	22.00	2.49	0.49
Min	0.00	1.00	0.15	0.00	0.15	0.00
25%	0.00	4.00	8.78	0.00	1.60	0.00
50%	1.00	8.00	17.36	16.67	2.19	1.00
75%	3.00	13.00	30.17	33.33	3.02	1.00
Max	25.00	77.00	208.69	100.00	112.19	1.00

Table 5.25: Transactional Descriptive Statistics of First Cluster of K-Means Sector Clustering with k=3 at Location 1

At Location 1, K-Means Sector Clustering resulted in Cluster 1 encompassing approximately 32,700 transactions characterized as low-expenditure transactions described analytically in Table 5.25. On average, these transactions involve around nine items purchased for an average total of

ID	Sector Name	Mean	Std	Min	25%	50%	75%	Max
1	Beverages	3.84	5.47	0.00	1.00	2.00	5.00	74.00
2	Self-Service Counter	1.09	1.48	0.00	0.00	0.00	2.00	8.00
4	Food Items	1.06	1.45	0.00	0.00	0.00	2.00	10.00
5	Confectionery - 1st Breakfast	0.99	1.45	0.00	0.00	0.00	2.00	13.00
3	Non-Food Items	0.82	1.07	0.00	0.00	1.00	1.00	13.00
13	Fruits & Vegetables	0.61	1.13	0.00	0.00	0.00	1.00	11.00
8	House Cleaning	0.26	0.81	0.00	0.00	0.00	0.00	17.00
9	Personal Hygiene	0.23	0.73	0.00	0.00	0.00	0.00	12.00

Table 5.26: Sectoral Descriptive Statistics of First Cluster of K-Means Sector Clustering with k=3 at Location 1

about 22 Euros. Among these transactions, approximately 21.7% of items were on offer, and around 58% were made by cardholders. Moreover, these transactions typically include about one to two items from the top sectors (Sectors 1, 2, 4, 5, and 3) outlined in Table 5.26, with a higher preference for the Beverages sector.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	6.17	33.60	80.12	19.06	2.46	0.76
Std	5.72	18.64	50.91	15.32	1.14	0.43
Min	0.00	7.00	6.00	0.00	0.15	0.00
25%	2.00	21.00	46.51	8.33	1.87	1.00
50%	5.00	30.00	68.33	15.38	2.32	1.00
75%	8.00	42.00	100.65	26.02	2.84	1.00
Max	62.00	425.00	1,213.42	100.00	43.80	1.00

Table 5.27: Transactional Descriptive Statistics of Second Cluster of K-Means Sector Clustering with k=3 at Location 1

ID	Sector Name	Mean	Std	Min	25%	50%	75%	Max
1	Beverages	9.48	10.79	0.00	3.00	7.00	13.00	409.00
4	Food Items	7.67	5.28	0.00	4.00	7.00	10.00	79.00
5	Confectionery - 1st Breakfast	4.81	4.35	0.00	2.00	4.00	7.00	59.00
2	Self-Service Counter	3.24	3.53	0.00	0.00	2.00	5.00	29.00
3	Non-Food Items	2.17	2.48	0.00	0.00	2.00	3.00	53.00
13	Fruits & Vegetables	2.15	2.55	0.00	0.00	1.00	3.00	21.00
8	House Cleaning	1.19	2.29	0.00	0.00	0.00	2.00	42.00
9	Personal Hygiene	0.93	1.80	0.00	0.00	0.00	1.00	40.00

Table 5.28: Sectoral Descriptive Statistics of Second Cluster of K-Means Sector Clustering with k=3 at Location 1

The second cluster at Location 1 (refer Table 5.27) comprises approximately 7,000 high-monetary transactions, with 76% of them made by cardholders. Within this group, customers purchase an average of about 34 items per transaction, including approximately 6 items on offer, accounting for an average of 19% of items bought on offer. Their shopping preferences, as detailed in Table 5.28, are notable in Sectors 1, 4, and 5, with moderate engagement in Sectors 2, 3, and 13.

Similarly, the third cluster at Location 1 consists of around 6,000 high-monetary transactions, with approximately 81% made by cardholders. This group also averages about 34 items per transaction, with about 8 items on offer, contributing to an average of 25.5% of items bought on offer. Their sector shopping pattern differs from the second cluster, as they heavily engage in Sectors 1 and 2, with moderate involvement in Sectors 3, 4, 5, and 13. The overview provided here is outlined in more details in Tables 5.29 and 5.30.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	8.15	34.23	78.28	25.47	2.32	0.81
Std	6.42	17.90	45.68	17.20	0.76	0.39
Min	0.00	6.00	4.83	0.00	0.27	0.00
25%	4.00	22.00	46.00	12.50	1.85	1.00
50%	7.00	30.00	67.88	22.22	2.22	1.00
75%	11.00	43.00	98.83	34.78	2.66	1.00
Max	57.00	189.00	431.00	100.00	13.34	1.00

Table 5.29: Transactional Descriptive Statistics of Third Cluster of K-Means Sector Clustering with k=3 at Location 1

ID	Sector Name	Mean	Std	Min	25%	50%	75%	Max
2	Self-Service Counter	9.52	5.01	0.00	6.00	8.00	11.00	58.00
1	Beverages	8.97	8.55	0.00	3.00	7.00	12.00	150.00
4	Food Items	3.67	3.60	0.00	1.00	3.00	5.00	35.00
5	Confectionery - 1st Breakfast	3.30	3.23	0.00	1.00	2.00	5.00	28.00
13	Fruits & Vegetables	2.62	2.75	0.00	1.00	2.00	4.00	22.00
3	Non-Food Items	2.12	2.35	0.00	0.00	2.00	3.00	28.00
8	House Cleaning	0.93	1.78	0.00	0.00	0.00	1.00	18.00
16	Frozen Foods	0.92	1.73	0.00	0.00	0.00	1.00	21.00

Table 5.30: Sectoral Descriptive Statistics of Third Cluster of K-Means Sector Clustering with k=3 at Location 1

Transitioning to Location 2, basic statistics for the first cluster are outlined in Tables 5.31 and 5.32. Analogous to Location 1, this cluster encompasses approximately 13,900 transactions characterized by low expenditure. These transactions reflect an average spending of around 15 Euros per visit, with about 25% of items purchased on offer out of a total of nearly 7 items per transaction. Approximately 55% of these transactions were conducted by cardholders. These transactions prominently corresponded to the top sectors provided in Table 5.32.

The second cluster at Location 2 represents the first of the two high monetary transactions clusters. It consists of approximately 3,200 transactions, with cardholders making about 70% of these transactions as explained in Table 5.33. This cluster is characterized by an average spending of around 50.5 Euros per transaction on a total of 24 items, out of which 17% (about 4 items) were purchased on offer. The primary sectors frequented by this group are detailed in Table 5.34, with a strong emphasis on Sectors 1 and 4, while Sectors 5, 2, 3, and 13 are moderately explored.

The third and final cluster at Location 2 represents a high monetary transactions cluster, comprising approximately 2,700 distinct transactions, as detailed in Table 5.35. A significant portion, 83%, of these transactions were conducted by cardholders, spending around 42 Euros per trip on

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	1.72	6.70	14.83	24.96	2.39	0.55
Std	1.96	4.00	10.61	26.47	2.04	0.50
Min	0.00	1.00	0.09	0.00	0.09	0.00
25%	0.00	4.00	6.69	0.00	1.54	0.00
50%	1.00	6.00	12.53	20.00	2.01	1.00
75%	3.00	9.00	20.64	40.00	2.66	1.00
Max	16.00	26.00	90.40	100.00	90.40	1.00

Table 5.31: Transactional Descriptive Statistics of First Cluster of K-Means Sector Clustering with k=3 at Location 2

ID	Sector Name	Mean	Std	Min	25%	50%	75%	Max
1	Beverages	1.16	1.57	0.00	0.00	1.00	2.00	10.00
4	Food Items	1.10	1.48	0.00	0.00	1.00	2.00	14.00
5	Confectionery - 1st Breakfast	1.00	1.47	0.00	0.00	0.00	2.00	18.00
2	Self-Service Counter	0.84	1.10	0.00	0.00	0.00	1.00	6.00
3	Non-Food Items	0.52	0.71	0.00	0.00	0.00	1.00	10.00
15,20,21	Cutting Counters	0.39	0.87	0.00	0.00	0.00	0.00	9.00
13	Fruits & Vegetables	0.39	0.78	0.00	0.00	0.00	1.00	7.00
7	Bakery	0.31	0.65	0.00	0.00	0.00	0.00	6.00

Table 5.32: Sectoral Descriptive Statistics of First Cluster of K-Means Sector Clustering with k=3 at Location 2

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	4.29	24.03	50.54	17.27	2.13	0.70
Std	4.29	12.60	32.91	13.95	1.01	0.46
Min	0.00	5.00	3.60	0.00	0.30	0.00
25%	1.00	16.00	30.70	6.90	1.57	0.00
50%	3.00	21.00	44.03	14.29	2.09	1.00
75%	6.00	28.00	63.35	25.00	2.60	1.00
Max	40.00	165.00	428.08	100.00	15.08	1.00

Table 5.33: Transactional Descriptive Statistics of Second Cluster of K-Means Sector Clustering with k=3 at Location 2

ID	Sector Name	Mean	Std	Min	25%	50%	75%	Max
1	Beverages	6.92	8.08	0.00	2.00	5.00	9.00	121.00
4	Food Items	4.79	4.46	0.00	2.00	4.00	7.00	50.00
5	Confectionery - 1st Breakfast	2.71	3.14	0.00	0.00	2.00	4.00	41.00
2	Self-Service Counter	2.06	2.54	0.00	0.00	1.00	3.00	43.00
3	Non-Food Items	1.46	2.05	0.00	0.00	1.00	2.00	37.00
13	Fruits & Vegetables	1.45	1.83	0.00	0.00	1.00	2.00	13.00
8	House Cleaning	0.99	1.93	0.00	0.00	0.00	1.00	30.00
15,20,21	Cutting Counters	0.90	1.46	0.00	0.00	0.00	2.00	11.00

Table 5.34: Sectoral Descriptive Statistics of Second Cluster of K-Means Sector Clustering with k=3 at Location 2

an average of about 20 items. Notably, nearly 40% of the items (almost 8 items) were purchased on offer. This cluster exhibited a strong preference for Sector 2, with moderate associations with Sectors 5, 1, and 4, as outlined in Table 5.36.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	7.86	20.53	42.06	39.52	2.03	0.83
Std	5.54	9.77	24.26	22.27	0.58	0.37
Min	0.00	4.00	3.53	0.00	0.59	0.00
25%	4.00	14.00	25.78	23.08	1.62	1.00
50%	7.00	19.00	37.03	37.50	1.96	1.00
75%	10.00	24.00	52.01	54.55	2.36	1.00
Max	47.00	124.00	219.40	100.00	5.11	1.00

Table 5.35: Transactional Descriptive Statistics of Third Cluster of K-Means Sector Clustering with k=3 at Location 2

ID	Sector Name	Mean	Std	Min	25%	50%	75%	Max
2	Self-Service Counter	5.84	3.13	0.00	4.00	5.00	7.00	28.00
5	Confectionery - 1st Breakfast	2.99	3.00	0.00	1.00	2.00	4.00	36.00
1	Beverages	2.89	3.33	0.00	1.00	2.00	4.00	27.00
4	Food Items	2.69	2.65	0.00	1.00	2.00	4.00	23.00
13	Fruits & Vegetables	1.06	1.39	0.00	0.00	1.00	2.00	8.00
3	Non-Food Items	0.98	1.28	0.00	0.00	1.00	2.00	14.00
15,20,21	Cutting Counters	0.94	1.45	0.00	0.00	0.00	2.00	8.00
16	Frozen Foods	0.78	1.49	0.00	0.00	0.00	1.00	22.00

Table 5.36: Sectoral Descriptive Statistics of Third Cluster of K-Means Sector Clustering with k=3 at Location 2

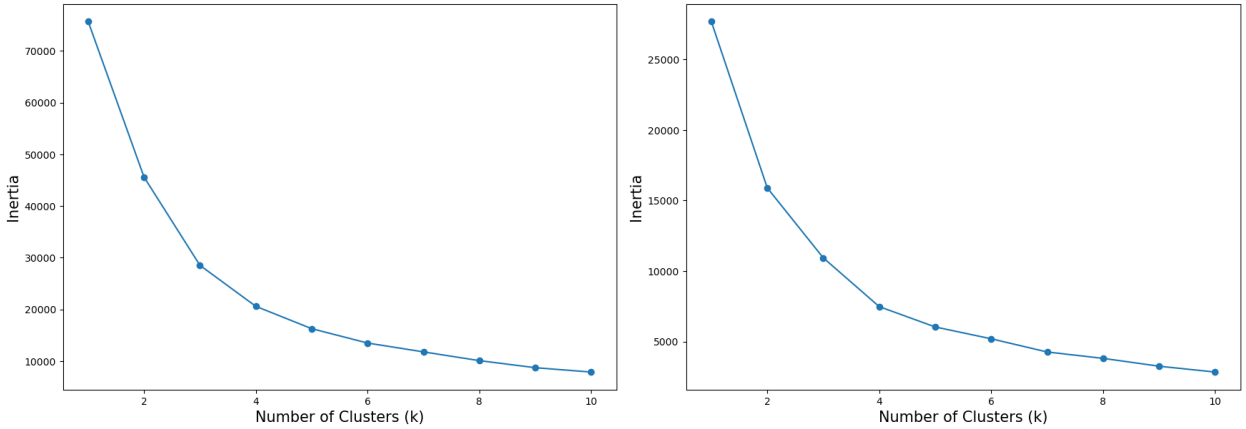
A notable trend observed in the Sector Clustering analysis is the higher percentage of items bought on offer during high monetary transactions, particularly when customers heavily shopped in Sector 2, the "Self-Service Counter," as opposed to Sector 1, "Beverages," and Sector 4, "Food Items." This trend was more pronounced in Location 2 compared to Location 1.

Specifically, at Location 2, there was a consistent presence of our combined Sectors 15, 20, and 21, corresponding to "Delicatessen Cutting Counter," "Cured Meat Cutting Counter," and "Cheese Cutting Counter." This suggests that customers at Location 2 demonstrated a significantly higher preference for these specific sectors compared to those at Location 1.

This observation can have several implications and could guide retailers in tailoring their marketing strategies, promotions, and product placements to maximize the impact on high-value transactions. Understanding the sectors that contribute most to substantial purchases, especially those with enticing offers, allows for targeted efforts to enhance customer engagement and drive sales in specific areas of the store. This prepares us seamlessly for the upcoming section, where we will delve into clustering customers based on their preferences for purchasing items on offer.

5.6 Promotional Shopping Clustering using K-Means

In this segment, our primary objective is to classify customers according to their buying behavior in response to promotions, discounts, and special offers. To achieve the specified goals, we focused on a limited feature set, including the total amount spent and the percentage of items bought on offer per transaction.



(a) Location 1 **(b)** Location 2

Figure 5.18: Elbow Method Results for K-Means Promotional Clustering to find Optimal k-Value

The clustering process was initiated using K-Means, and determining the optimal number of clusters involved a combination of the Elbow Method, Silhouette Scores, and the actual implementation of K-Means to assess the distinguishability of the clusters. The results of the Elbow Method are illustrated in Figure 5.18, revealing a slight discrepancy between Location 1 and Location 2. In Location 1, the most significant reduction in slope occurs at k=3, while in Location 2, it happens at k=4. Additionally, after calculating Silhouette Scores for both locations with k=2, 3, and 4, it was observed that k=2 had the highest values for both locations. However, in line with the Elbow Method results, the Silhouette Scores for k=3 in Location 1 and k=4 in Location 2 were the second-highest values, at 0.453 and 0.427, respectively. Considering this and the distinctiveness observed in the actual K-Means results, the decision was made to perform K-Means Clustering with k=3 and k=4 at Locations 1 and 2, respectively. K-Means Clustering with k=2 was disregarded due to its specific focus on separating based on the value of transactions giving us clusters of high and low monetary transactions.

5.6.1 Location 1: K-Means Promotional Shopping Clustering with k=3

In Location 1 analysis, the K-means algorithm assigned 8,000, 26,400, and 11,300 transactions to Clusters 1, 2, and 3, respectively. Basic statistics for these clusters are presented in Tables 5.37, 5.38, and 5.39. The first and smallest cluster mainly comprises high monetary transactions with a below-average percentage of items bought on offer. Cardholders conducted 80% of these transactions, suggesting a group of brand-loyal customers who predominantly fulfill their grocery needs at this location, regardless of available offers.

Moving on to clusters categorizing below-average spenders, both Cluster 2 and Cluster 3 involve transactions with approximately 11 total items costing around 24 Euros. The key distinction lies

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	7.44	40.55	106.02	17.95	2.99	0.80
Std	6.54	19.35	45.46	11.80	2.70	0.40
Min	0.00	1.00	59.89	0.00	0.23	0.00
25%	3.00	28.00	75.22	9.09	2.18	1.00
50%	6.00	37.00	92.24	15.87	2.60	1.00
75%	10.00	50.00	122.36	25.00	3.20	1.00
Max	62.00	425.00	1,213.42	81.54	112.19	1.00

Table 5.37: Descriptive Statistics of First Cluster of K-Means Promotional Clustering at Location 1

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	1.42	11.34	23.61	10.52	2.64	0.57
Std	1.67	9.30	16.48	9.65	2.27	0.50
Min	0.00	1.00	0.16	0.00	0.15	0.00
25%	0.00	4.00	9.75	0.00	1.60	0.00
50%	1.00	9.00	20.23	10.00	2.18	1.00
75%	2.00	16.00	35.30	19.05	2.94	1.00
Max	15.00	150.00	65.37	29.63	59.00	1.00

Table 5.38: Descriptive Statistics of Second Cluster of K-Means Promotional Clustering at Location 1

in the percentage of items bought on offer. Cluster 2 comprises about 57% transactions made by cardholders, with an average of about 1-2 items on offer, resulting in an average of around 10% of items bought on offer. In contrast, Cluster 3 consists of 67% transactions made by cardholders, averaging about 5-6 items on offer, leading to approximately 50% of items bought on offer. This percentage is significantly higher than the location average of 21.5%.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	5.48	11.73	24.32	49.45	2.26	0.67
Std	4.31	8.58	17.13	18.01	1.30	0.47
Min	1.00	1.00	0.15	29.03	0.15	0.00
25%	2.00	5.00	10.47	35.29	1.57	0.00
50%	4.00	10.00	20.25	44.44	2.00	1.00
75%	7.00	16.00	34.91	57.14	2.57	1.00
Max	46.00	64.00	106.84	100.00	31.27	1.00

Table 5.39: Descriptive Statistics of Third Cluster of K-Means Promotional Clustering at Location 1

Drawing additional insights from Figure 5.19, which classifies clusters on a two-axes plot between total amount spent and percentage of items bought on offer, and conducting calculations based on the results of the descriptive cluster analyses, it becomes evident that Cluster 3 constitutes approximately 24.7% of the total transactions recorded at Location 1. This suggests a substantial portion of consumers who are particularly responsive to promotions.

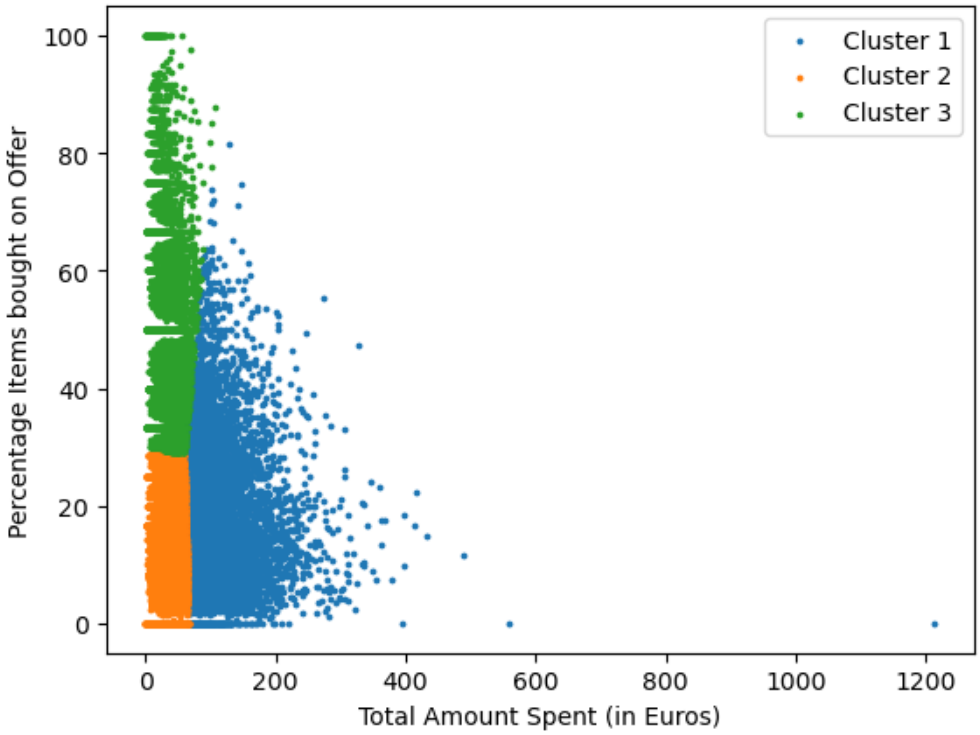


Figure 5.19: Location 1: K-Means Promotional Clustering Results

5.6.2 Location 2: K-Means Promotional Shopping Clustering with k=4

Moving on to Location 2, where K-Means Clustering with k=4 was performed, the algorithm grouped 1,750, 7,850, 4,400, and 5,800 transactions into Clusters 1, 2, 3, and 4 respectively. Figure 5.20 provides insights into the distribution of transactions in each cluster based on the features: total amount spent and percentage of items bought on offer. Additionally, basic descriptive statistics for each cluster are presented in Tables 5.40, 5.41, 5.42, and 5.43 from Cluster 1 to 4.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	8.16	32.22	80.82	25.65	2.91	0.79
Std	6.70	14.98	30.94	18.07	3.13	0.41
Min	0.00	1.00	52.68	0.00	0.54	0.00
25%	3.00	23.00	61.42	11.76	2.14	1.00
50%	6.00	30.00	71.93	22.22	2.54	1.00
75%	11.00	37.00	88.46	35.71	3.01	1.00
Max	47.00	165.00	428.08	96.30	90.40	1.00

Table 5.40: Descriptive Statistics of First Cluster of K-Means Promotional Clustering at Location 2

Similar to Location 1, the first cluster at Location 2 represented brand-loyal customers who primarily chose this location for the majority of their grocery needs, spending approximately 81 Euros per trip. In this cluster, 79% of the transactions were made by cardholders, and about 25.7% of the items purchased were on offer. Notably, this percentage aligns with the location average for the percentage of items bought on offer, indicating that even high-spending loyal customers at this location are inclined to seek out offers more than those at the first location.

Clusters 2 and 3 share similarities in categorizing low monetary transactions but differ signifi-

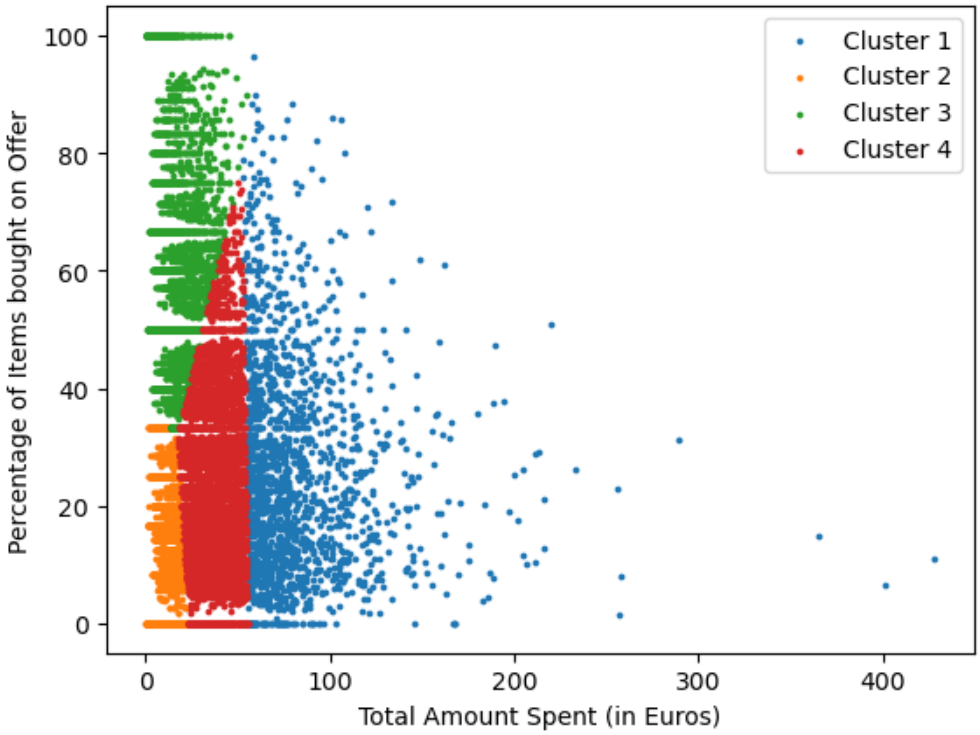


Figure 5.20: Location 2: K-Means Promotional Clustering Results

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	0.59	5.85	10.01	8.73	2.07	0.45
Std	0.82	4.32	5.66	11.45	1.45	0.50
Min	0.00	1.00	0.09	0.00	0.09	0.00
25%	0.00	3.00	5.29	0.00	1.35	0.00
50%	0.00	5.00	9.51	0.00	1.82	0.00
75%	1.00	8.00	14.29	16.67	2.37	1.00
Max	7.00	66.00	23.62	33.33	22.40	1.00

Table 5.41: Descriptive Statistics of Second Cluster of K-Means Promotional Clustering at Location 2

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	4.47	7.61	14.44	61.16	2.16	0.71
Std	3.29	4.89	8.83	19.66	1.64	0.45
Min	1.00	1.00	0.46	33.33	0.35	0.00
25%	2.00	4.00	7.65	50.00	1.43	0.00
50%	4.00	7.00	13.48	55.56	1.80	1.00
75%	6.00	10.00	19.59	72.73	2.37	1.00
Max	32.00	36.00	54.34	100.00	34.65	1.00

Table 5.42: Descriptive Statistics of Third Cluster of K-Means Promotional Clustering at Location 2

cantly in alignment with our promotional clustering goals. Cluster 2 is characterized by quick-grab transactions, averaging about 6 total items purchased at an average cost of 10 Euros per transaction. These transactions exhibit an average of 9% of items bought on offer, with approximately 55% of them conducted by non-cardholders. Unlike Location 1, the majority of cardholders is not

a characteristic of this cluster at Location 2.

Cluster 3, on the other hand, represents a heavily promotions-oriented cluster, with all transactions featuring at least one item on offer. These transactions averaged about 14 Euros spent on a total of approximately 7 to 8 items, with 61% of them bought on offer (equivalent to 4 to 5 items). This percentage is more than double the location average for items bought on offer. Notably, about 71% of these transactions were made by cardholders, slightly higher than that observed for Location 1.

	Offer	Total Items	Amount Spent	% Offer	Amount per Item	Card
Mean	3.49	15.50	34.19	21.70	2.54	0.71
Std	3.03	6.30	9.23	14.65	1.65	0.45
Min	0.00	1.00	17.90	0.00	0.40	0.00
25%	1.00	11.00	26.44	10.00	1.86	0.00
50%	3.00	15.00	32.92	20.00	2.24	1.00
75%	5.00	19.00	41.16	31.58	2.74	1.00
Max	26.00	78.00	55.49	75.00	36.00	1.00

Table 5.43: Descriptive Statistics of Fourth Cluster of K-Means Promotional Clustering at Location 2

The fourth cluster formed at Location 2 grouped together transactions that were moderately above-average in terms of money spent but slightly below-average in the number of items bought on offer. This cluster represents customers who exhibit typical shopping behavior, addressing their basic needs while incorporating a mix of regular and promotional items. Occasionally taking advantage of offers or discounts, these customers maintain a spending pattern that strikes a balance between budget-conscious choices and occasional indulgences, categorizing them as moderate spenders compared to the overall customer base.

Ultimately, akin to Location 1, approximately 23% of the transactions at Location 2 were primarily focused on promotions. However, a notable disparity exists between the locations. Even the more brand-loyal heavy spenders at Location 2 demonstrated a significant inclination towards purchasing items on offer, highlighting its importance for the retailer. The absence of a proportionately distinctive fourth cluster at Location 1, as indicated by the results from the Elbow Method and Silhouette Scores, is intriguing and open to various interpretations. One possible conclusion is that typical shoppers at Location 2 maintained their typical habits, forming a distinct cluster, while at Location 1, external factors might have influenced them to deviate from their usual shopping patterns.

This underscores a significant distinction between the locations, highlighting that customers engaged with promotions differently at each site. Although location-targeted promotions may offer a straightforward solution, any decision in this regard should be informed by the various factors discussed throughout the research.

Chapter 6

Discussion and Conclusions

The preceding chapters have unfolded an in-depth exploration into the realms of retail customer behavior, employing a multifaceted approach that amalgamates traditional descriptive analysis with advanced data-driven methodologies. The journey embarked upon sought to address pivotal questions concerning customer characteristics, clustering techniques, and the consequential impact on retail performance metrics. As we navigate through the culminating chapter of this thesis, we distill the myriad findings and insights, paving the way for a comprehensive understanding of the intricate dynamics shaping contemporary retail landscapes, notably at two distinct outlets of the same grocery store in a city in Southern Europe during March 2023. This chapter is a synthesis of analytical rigor, methodological insights, and actionable recommendations, which collectively contribute to the broader discourse on enhancing retail strategies in an era defined by data-driven decision-making.

6.1 Key Findings

In this section, we revisit the primary discoveries and offer a succinct overview of the results derived from the data-intensive segments of the research. Additionally, our objective is to delineate persistent trends that hold significant value throughout the study. Specifically, we delve into the results stemming from the two focal points of our study: Exploratory Data Analysis and Customer Clustering.

6.1.1 Exploratory Data Analysis

We conducted an exhaustive exploratory analysis of our dataset, meticulously examining all its prominent features. The thoroughness of our Exploratory Data Analysis (EDA) was paramount, given its significance and efficacy in revealing trends and patterns. This comprehensive exploration serves as the cornerstone for employing a diverse set of data-driven techniques to attain actionable insights.

Demographic Analysis

We conducted a demographic analysis focused on cardholders due to missing associated information for other customers. At both locations, age distributions exhibited nearly Gaussian patterns, if not precisely Gaussian. The modal ages were 57 and 59 years for Locations 1 and 2, respectively. The gender distribution skewed significantly toward women, constituting approximately 67% and 73% of

cardholders at Locations 1 and 2, respectively. It's worth noting that while CAP data was available, we adhered to ethical guidelines and refrained from utilizing it in our study to ensure the protection of sensitive information related to associated individuals.

Transactional Analysis

The findings from Transactional Analysis revealed valuable insights. Location 1 outpaced Location 2 in transaction volume, recording 45,700 and 19,800 distinct transactions, respectively. An intriguing pattern emerged as Cardholders consistently exhibited higher spending per transaction compared to Non-Cardholders at both locations. Notably, a significant divergence in shopping behavior surfaced, with Cardholders at Location 2 purchasing approximately 30% of items on offer, contrasting with the 23% recorded at Location 1. This distinction was further underscored by Spearman correlation coefficients, revealing that, between the two locations, the correlation between items bought on offer and cardholding status intensified specifically for Location 2.

Item Sector Analysis

These findings prominently emphasized the sectors with the highest footfall at each location, providing valuable insights into distinct shopping behaviors. The analysis delved into trends categorized by cardholding status, revealing that cardholders exhibited a propensity to purchase more items and spend higher amounts at both locations. The exploration identified a consistent set of top five sectors for all customers across locations. Additionally, associations between sectors and cardholding status were uncovered, with "Beverages," "Self-Service Counter," and "Food Items" showing significant ties to cardholders at both locations. A notable divergence between the locations emerged as at Location 2, "Cured Meat Cutting Counter" and "Cheese Cutting Counter" sectors consistently replaced the "Easter Celebrations" sector's prominence at Location 1, contributing to a nuanced understanding of shopping preferences.

Item Category Analysis

This analysis aimed to delve into more specific insights than those obtained through the Item Sector Analysis. An unexpected finding in this analysis was the emergence of an item category labeled "Uncensored," likely representing daily over-the-counter items that defied conventional categorization. Although there were no significant distinctions between the most purchased item categories by cardholders and non-cardholders at both locations, variations were observed between the two locations.

One notable difference aligned with our Item Sector Analysis: the highly purchased "Easter Egg" item in the "Easter Celebrations" sector at Location 1 was replaced by meat and cheese items in the "Cured Meat Cutting Counter" and "Cheese Cutting Counter" sectors at Location 2.

Another noteworthy contrast between the locations was evident when meat and cheese items in the "Cured Meat Cutting Counter" and "Cheese Cutting Counter" sectors at Location 2 displayed a high percentage of items bought on offer. An interesting and expected association emerged, indicating that customers buying vegetables at both locations were 25% more likely to purchase plastic bags than usual.

Furthermore, specific item-category associations with cardholders were identified. At both locations, items under the categories of biscuits, fresh milk and cream, and vegetables were highly associated with cardholders. However, at Location 1, water exhibited a high association with cardholders, while at Location 2, scaled bread items were prominently associated with cardholders.

Daily and Hourly Analyses

In this segment of the Exploratory Analysis, we commenced our inquiry with an examination of daily shopping behaviors. The first location operated throughout the week, whereas the second location remained closed on Sundays. Overall, the weekends and Tuesdays exhibited the highest density of footfall for both locations. At Location 1, the highest and second-highest values for total customers, total items purchased, and total amount spent consistently occurred on Saturdays and Tuesdays, respectively. In contrast, at Location 2, while this trend persisted for total amount spent, variations were observed in the data capturing total customers and total items bought. This discrepancy suggested the presence of some promotion, not adequately reflected in the data, leading to a reduced amount spent.

An additional analysis regarding the preference for certain days among cardholders was conducted for both locations. Cardholders consistently favored Tuesdays, Fridays, and Saturdays. Specifically for Location 1, cardholders were found to be 12% more likely to shop on Tuesdays, indicating a potential association with a promotion linked to their cardholding status.

A comparable approach was undertaken to capture the hourly preferences of the customer base. A consistent trend was observed at both locations, where customers showed a preference for shopping in the morning hours before lunch and then again in the evening hours after lunch. To gain a deeper understanding of these trends, associations between specific time frames and cardholding status were explored to identify if cardholders exhibited preferences for certain shopping times.

At Location 1, cardholders demonstrated a preference for the morning hours from 9 AM to noon, with approximately a 10% higher likelihood of a transaction involving a cardholder at 10 AM and 11 AM. Similarly, at Location 2, cardholders preferred even earlier hours, from 8 AM to 11 AM, with about an 11% higher likelihood of them shopping during these hours.

Big Basket Analysis

In the context of the Big Basket Analysis, the primary focus is on discerning the behavior of high-spending customers, often characterized by their brand loyalty and frequent visits for most of their grocery needs. The criteria for identifying these heavy spenders were based on two key factors: purchasing above the location-average quantity of items and spending beyond the location-average monetary amount.

The initial step involved providing an overview of the day and time preferences exhibited by these heavy spenders. The analysis revealed a preference for Tuesdays and weekends among bulk shoppers at both locations. Regarding timing, it was observed that at Location 1, these shoppers frequented the store in the morning before lunch and in the evening after lunch, while at Location 2, the morning hours were preferred.

Delving into specifics, at Location 1, the majority (79%) of bulk shoppers were cardholders. They, on average, purchased 34 items per transaction, with an expenditure of approximately 82

Euros. Meanwhile, at Location 2, constituting around 77% cardholders, these shoppers bought about 22 items per transaction, spending an average of 49 Euros.

Following the general analysis, we conducted specific item sectoral and item categorical analyses tailored to the unique shopping behaviors of bulk buyers—customers of substantial significance to retailers due to their substantial purchasing patterns. These analyses aimed to provide insights into the item-oriented preferences of this particular customer segment.

In the sectoral analysis, we consistently observed high Spearman Correlation coefficients linking the total items bought with the "Beverages," "Food Items," and "Self-Service Counter" sectors. This underscored the noteworthy importance of these sectors to bulk shoppers at both locations. An intriguing trend emerged concerning the correlation between categorical variety and total items, revealing that bulk shoppers tended to explore a variety of items within specific prominent sectors, rather than across a broad range of sectors. Exploration into sectoral association rules emphasized the significance of the "Beverages" sector further. High cross associations were identified between the "Self-Service Counter" and "Food Items" sectors at Location 1, with similar prevalence at Location 2. Additionally, a consistent sector related to breakfast items was noted.

In the categorical analysis, only "Water" emerged as a notable item category at Location 1 based on Spearman Correlation coefficients. To delve deeper into this trend, further investigation was conducted using association rules. At Location 1, it was revealed that bulk shoppers frequently purchased vegetables in combination with fresh fruits, mozzarella cheeses, fresh milk and cream, biscuits, pre-packaged cheeses, and plastic bags. On the other hand, at Location 2, the association among vegetables, fruits, mozzarella cheeses, biscuits, pre-packaged cheeses, and plastic bags persisted, but new associations with rough bread, poultry items, and scaled bread items were identified. This highlighted distinct shopping behaviors of bulk consumers between the two locations.

Cardholder Demographic Associations

In this segment of the analysis, we specifically targeted the cardholding customer base, aiming to discern any prevailing trends among cardholders of distinct age groups and gender categories. By calculating the average age of cardholders visiting on different days of the week and during various time frames for both locations, we deduced that the older cardholding customer base exhibited a preference for Location 2 over Location 1, as evidenced by consistently higher average ages. Additionally, within each location, it was observed that older cardholders favored shopping on Tuesdays and during the early morning hours.

We delved deeper into the analysis, exploring the preferred days and time frames for both male and female customer bases. Initial calculations based on the percentages of men and women shopping at different times of the day and days of the week did not reveal significant trends at either location. However, to ensure a comprehensive understanding and uncover potential hidden trends, association rules were applied.

Upon closer examination, intricate observations emerged, such as men being nearly 10% more likely to make a transaction at 7 PM at both locations. Additionally, women exhibited a higher tendency to shop during daytime hours compared to men at both locations. Location 1 showed that men preferred weekends, while women were inclined to shop during weekdays. At Location 2, men were more likely to transact on Mondays and Saturdays, while women showed a preference for Tuesdays and Wednesdays.

These findings underscore the nuanced differences in shopping behaviors based on gender, age groups, and geographical locations.

6.1.2 Customer Segmentation

Following an extensive Exploratory Data Analysis, the focus shifted to the pivotal phase of our study: Customer Segmentation. Leveraging the insights revealed during the previous analyses, this stage aimed to utilize these trends to establish customer clusters, thereby laying the foundation for achieving actionable objectives.

Predefined Customer Segmentation through RFM

This constituted our initial approach towards establishing actionable customer clusters. The primary objective of this section was to concentrate on three fundamental aspects of any customer base: the recency of their last visit, the frequency of their visits, and the net amount spent during those visits. The computation of these three features was exclusively feasible for the cardholding customer base, given our ability to associate each transaction they made with their unique identifier. Through the calculation of RFM scores and subsequent categorization using Equation 5.1 and Table 5.1, the cardholding customer base was segmented into five categories, ranging from Lost Customers to Top Customers. The outcomes derived from this predefined segmentation mirrored consistency across the two locations, with 60% of customers falling within the Low Value to Lost Customers category and 40% falling within the Medium Value to Top Customers category.

RFM into K-Means Clustering

This phase of the study focused primarily on generating clusters from RFM Ranks using K-Means Clustering and conducting a detailed comparison between predefined segmentation and the K-Means algorithm. K-Means Clustering was executed for k=2 and k=4 for both locations, guided by insights obtained from the Elbow Method and Silhouette Scores.

The clustering outcomes with k=2 remained uniform across both locations, with the algorithm generating two distinct clusters. One cluster aligned with high-frequency and heavy-spending customers who had visited in the last couple of days, while the other corresponded to low-frequency and low-spending customers who visited once a month. The primary disparity observed was in the juxtaposition of these clusters against our predefined clusters. At Location 1, the low-value cluster from K-Means exclusively comprised Lost and Low-Value Customers from the predefined segmentation, while the high-value K-Means cluster grouped the other three predefined segments, including some Low-Value Customers. At Location 2, the low-value cluster from K-Means included some Medium-Value Customers as well, while the high-value K-Means cluster showed no differences when compared to Location 1.

With a clustering configuration of k=4 for Location 1, distinct and identifiable clusters emerged. The first and second segments were marked by low spending and a single visit, differing in the timing of the visit—either in the last week of the month or the first two weeks of the month. The third group of customers exhibited high spending and frequent visits, with their last visit recorded within the last two days. The fourth set of customers displayed high spending and moderate visit frequencies, with their last visit occurring during the middle two weeks of the month.

Similarly, applying K-Means Clustering with k=4 to Location 2 revealed distinctive segments. The first segment comprised customers with medium frequency and high spending, mainly visiting the store at least once in the last week. The second group of customers made, on average, two grocery trips during the month, with one occurring in the last four days, spending approximately 70 Euros. The third group was characterized as the least frequenting and lowest spending, with their average single visit spread across the month. The fourth cluster represented the highest frequency and highest spending customer base, with the majority visiting the store on the final day of the month.

Comparing the K-Means Clustering results for both locations with the predefined segmentation, Location 1 exhibited a noticeable K-Means algorithm threshold between Low Value Customers and Medium Value Customers from the predefined segmentation. The two low-value K-Means clusters at Location 1 did not include Medium Value Customers from the predefined segmentation, while the same two low-value K-Means clusters at Location 2 included Medium Value Customers. Further examination of the RFM Scores for the clusters derived from K-Means Clustering led to the conclusion that the K-Means algorithm generated sufficiently distinct clusters, even aligning with our interpretation of RFM Score.

This section ultimately underscored the superiority of the K-Means algorithm over our predefined segmentation by consistently providing high Silhouette scores. These scores indicate robust levels of intra-cluster similarities and inter-cluster dissimilarities for K-Means clusters compared to the predefined segments.

Transactional Data Focused K-Means Clustering

This segment of the research centered on clustering customers according to their primary shopping habits and cardholding status. Clustering was executed with k=2 and k=3, guided by the Elbow Method and Silhouette scores. The clustering outcomes for k=2 aligned with our exploratory analyses focused on transactional data. The algorithm delineated transactions into high and low monetary categories for both locations. The high monetary transactions cluster closely resembled the findings from the Big Basket Analysis for the respective locations, with approximately 80% of transactions made by cardholders.

Expanding our clustering analysis to k=3, the algorithm generated two clusters associated with low monetary transactions and one with high monetary transactions. The high monetary transaction cluster, consistent with k=2, comprised approximately 80% cardholder transactions for both locations. A more detailed examination of the new clusters formed with k=3 revealed that the algorithm segregated the low expenditure clusters based on cardholding status, achieving around 95% accuracy. The significant distinction observed was that even within the group of smaller transactions, cardholders were spending approximately 10 Euros more per transaction compared to non-cardholders. This underlines the significance of cardholders for the retailer.

Additional intriguing insights emerged when Principal Component Analysis (PCA) was applied to the four features to visualize our clustering outcomes. Utilizing two principal components, we were able to capture 90% of the variance in the data. Plotting our clustering results against these principal components underscored their significance, as evident distinctions were clearly visible in these plots.

Transactional Data Focused DBSCAN

This research segment delved into a more advanced clustering algorithm: Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The outcomes of applying DBSCAN to the transactional data shared some similarities with K-Means Clustering with k=3. Notably, DBSCAN labeled high monetary transactions at both locations as Noise. Additionally, it distinctly separated low monetary transactions into two clusters, rigorously distinguishing cardholders and non-cardholders, in contrast to the more lenient approach of K-Means. Despite some variations in the descriptive statistics of the clusters, the overall differences between DBSCAN and K-Means with k=3 were not significant. Plotting our clustering results on principal components and conducting an overall comparison highlighted that the two algorithms offer distinct interpretations of the data, and their outcomes are aligned with their individual methodologies. This suggests that the choice between the two algorithms depends on the specific goals and interpretation needs of the analysis.

Sectoral Shopping Clustering using K-Means

This research branch emerged directly from the insights gained during the exploratory analysis of item sectors and item categories (Sections 4.3 and 4.4). Combining the findings from these exploratory analyses, we recognized the sparse nature of categorical item data. Consequently, we opted to aggregate transactions into sectoral shopping data and performed clustering based on this refined approach. This strategy aimed to gain a more nuanced understanding of shoppers' spending patterns within different grocery store sectors.

The results from this segment revealed intriguing differences compared to previous clustering outcomes. Notably, the K-Means algorithm here produced one low monetary cluster and two high monetary clusters for both locations. The low monetary cluster comprised predominantly non-cardholder transactions (55-60%) involving small transactions, primarily focused on sectors such as Beverages, Food Items, Confectionery - 1st Breakfast, and Self-Service Counter. In contrast, the other two clusters featured approximately 70-80% cardholder transactions, distinguished by their substantial shopping activities, with variations centered around key sectors.

One cluster exhibited transactions concentrated in the "Beverages" and "Food Items" sectors, with only 17-19% of items bought on offer. The second cluster focused on purchases within the "Self-Service Counter" sector, with a higher range of 25-39% items bought on offer. This discernment in heavy-spending transactions based on sectors, leading to differences in the percentage of items bought on offer, indicates that the retailer is offering a significantly higher number of items on offer in the "Self-Service Counter" sector at both locations.

Promotional Shopping Clustering using K-Means

In this segment of our study, our attention was primarily directed towards two key transactional features: the total amount spent and the percentage of items bought on offer. Leveraging insights from the Elbow Method and Silhouette Scores, we applied K-Means Clustering with k=3 at Location 1 and k=4 at Location 2. The clustering outcomes yielded three comparable groups at both locations, with the fourth cluster at Location 2 encompassing transactions from typical shoppers exhibiting a mix of regular and promotional item preferences.

The first cluster encapsulated high monetary transactions, representing brand-loyal customers who conducted the majority of their grocery shopping at the investigated locations. Noteworthy distinctions between the two locations arose as these brand-loyal customers, at Location 1, purchased items with an average percentage bought on offer lower than the location average, while at Location 2, the percentage aligned with the location average. This underscored a higher involvement of brand-loyal customers in purchasing promotional items at Location 2 compared to Location 1.

The second cluster comprised quick-grab transactions, characterized by low amounts spent and a low percentage of items bought on offer. Lastly, a group heavily focused on promotional items emerged at both locations, featuring low spenders who purchased approximately 50% and 61% of items on offer at Locations 1 and 2, respectively. A consistent trend surfaced, indicating that Location 2 exhibited a higher percentage of items bought on offer across various criteria. This led us to the conclusion that the retailer could derive benefits from targeted promotions, leveraging the distinct shopping behaviors of customers at the two locations.

6.2 Interpretations and Expectations

In this section, we categorize and highlight significant observed trends, examining their implications and assessing their alignment or deviation from established expectations within the retail industry.

Cardholders spend more than Non-Cardholders

In our research, a detailed examination revealed that cardholders consistently outspent non-cardholders at both locations. Notably, even in scenarios where cardholders engaged in smaller transactions, their expenditures surpassed those of non-cardholders in similar transactions. This aligns with broader industry trends, as evidenced by the spending patterns of members in leading retail loyalty programs. For instance, Amazon Prime members tend to spend twice as much as non-members, while Walmart+ members show an average spending increase of approximately $15 compared to non-members [17].

Women are more Brand-Loyal than Men

Our analysis revealed a notable gender disparity among cardholders, with women outnumbering men at both locations. Particularly, at Location 1, women constituted twice the number of men, while at Location 2, they were three times more numerous. This trend aligns with findings from a prior quantitative study on Swedish customers' experiences with loyalty programs conducted by students at Linnaeus University, which emphasized the greater involvement of women in loyalty programs [11]. While our research suggests that women may exhibit higher loyalty, an existing study suggests women are more loyal toward individual service providers and men toward firms [15], it underscores the importance of tailoring strategies to meet the distinct needs of both gender groups while not alienating one from another.

Foot Traffic is the highest on the Weekends

Our analysis uncovered that, while the highest customer frequency occurred on weekends, Tuesdays also witnessed a considerable influx of customers. The well-established trend of in-person grocery

shopping during weekends was evident, but the notable spike on Tuesdays across both locations suggested the success of a promotion in drawing customers. Based on our analyses, it was inferred that this particular promotion was effective in attracting older cardholders.

50-to-65 Year Olds are the Most-Loyal Generation

Our investigation revealed that customers in the age group of 50 to 65 years old exhibited the highest level of loyalty for both locations, with Location 2 showing a greater concentration in the older age range. This finding resonates with a YouGov study, emphasizing that individuals aged 55 to 64 years old constitute the most brand-loyal generation [3].

Men are more likely to shop in the evenings and during the weekends

Our examination revealed that men tend to prefer shopping during evening hours and on weekends, potentially influenced by full-time employment commitments throughout the week. This pattern aligns with findings from prior research in the retail industry [14].

6.3 Implications and Applications

In this segment, we aim to thoroughly explore the ramifications of our research and identify actionable applications that can effectively enhance critical retail performance metrics.

6.3.1 Understanding Cardholding Customer Base

In our research, we placed a significant emphasis on understanding the customer base, particularly those holding cards. This strategic focus was motivated by various factors, such as the observation that a majority of transactions are carried out by cardholders and that high spenders predominantly fall within this category. The comprehensive exploratory analysis of cardholders, followed by clustering based on RFM features, establishes a robust foundation for making informed decisions related to cardholding customers for the retailer. This process enables the retailer to categorize the shopping habits of cardholders, facilitating the implementation of customer-focused promotions and an enhanced shopping experience. Achieving an in-depth understanding of the customer base is crucial for effectively tailoring strategies to meet the unique needs of cardholders.

6.3.2 Increasing Cardholding Customer Base

Through extensive research across various studies within the retail industry, a consistent trend has emerged—cardholders consistently exhibit higher spending patterns compared to non-cardholders. This underscores the significant value associated with the cardholding customer base for retailers. In practical terms, we propose implementing an "At Counter Loyalty Card Sign-Up" program, leveraging a straightforward criterion based on a minimum spending amount. To enhance the efficacy of this initiative, an immediate promotional offer can be extended during the transaction. This recommendation stems from our analysis, which revealed that approximately 20% of the high monetary transactions at each location were conducted by non-cardholders during a thorough big basket analysis. This signifies a substantial pool of potential high-value customers for grocery stores. Moreover, our strategy aligns with findings from a Loyalty Programs study conducted by

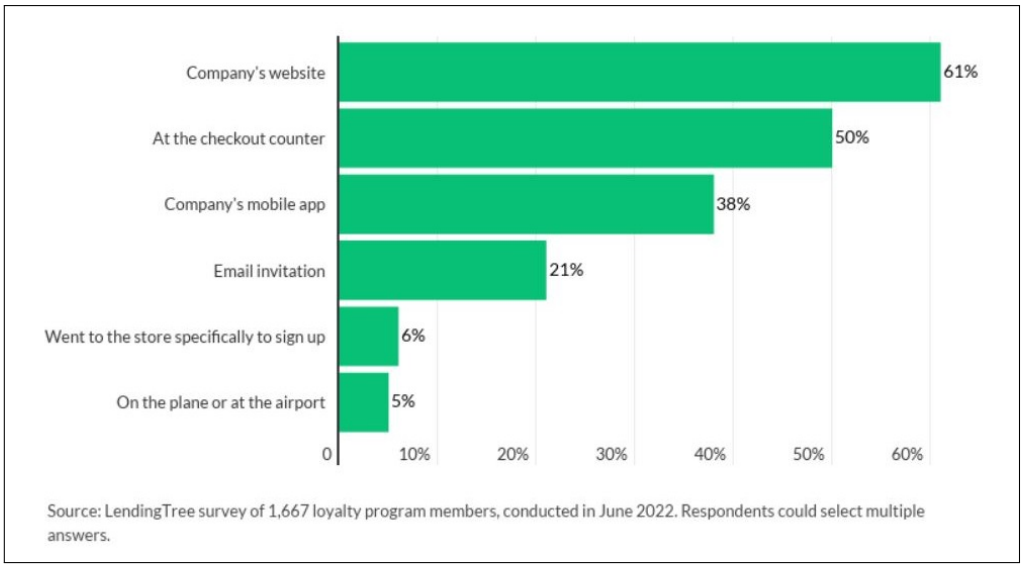


Figure 6.1: Signing-Up Methods of Loyalty Program Members

LendingTree, indicating that 50% of loyalty card owners enrolled in the program at the counter [16], as illustrated in Figure 6.1.

6.3.3 Targeted Promotions

This section elaborates on the clustering outcomes, emphasizing the sectors and promotions discussed in Sections 5.5 and 5.6. Our findings underscore notable distinctions in customer shopping behaviors between the two locations. This presents an opportunity for the retailer to tailor strategies independently for each location, establishing a more resilient and customer-centric framework. By optimizing the customer experience based on location-specific insights, the retailer stands to enhance profits and create a more targeted approach to engaging with their customer base.

6.3.4 Associating Item Sectors and Item Categories

Our research focused on establishing fundamental associations between item sectors, item categories, and cardholding status. The objective was to gain a nuanced understanding of frequently co-occurring elements. This exploration yielded intriguing insights that the retailer can leverage in grocery stores to enhance the overall shopping experience for customers. Specifically, this knowledge can be applied to organize stores in a sequential sector fashion, optimizing the layout for smoother navigation and improved customer satisfaction. By strategically placing items associated with each other, the retailer can create a more intuitive shopping environment, facilitating a seamless and enjoyable experience for their customer base.

6.4 Answering the Research Questions

As we conclude our research, it is pertinent to revisit the initial research questions that guided our study. These questions served as a compass throughout our investigation, shaping the flow of our research. In brief, we will now elucidate how we have addressed these questions and elucidate their role in formulating the trajectory of our inquiry.

What Baseline Customer Characteristics Emerge from Descriptive Analysis, and How Do They Inform Subsequent Clustering Strategies?

This question played a pivotal role in our research, serving as a catalyst for a thorough exploratory analysis. The outcomes of this analysis brought to the forefront crucial insights into key demographic details, transactional behaviors, and engagement metrics. This foundational phase of the research then seamlessly directed subsequent clustering techniques, elucidating influential features such as expenditure amounts, cardholding status, purchases during promotions, and other pertinent factors in a clear and concise manner.

What Data-Driven Methods are Most Effective for Customer Clustering in the Retail Sector?

The inquiry regarding customer clustering in the retail sector was addressed through a comprehensive comparison of multiple clustering approaches. A detailed study in the field, conducted by Alves Gomes and Tobias in 2023 [2], identified K-Means Clustering based on RFM features as the most prominent method in the retail industry. Our study extended this exploration by focusing on four clustering approaches: Predefined Segmentation, K-Means Clustering, Hierarchical Clustering, and DBSCAN.

The initial comparison, utilizing RFM features, revealed that K-Means Clustering surpassed Predefined Segmentation as a superior data-driven method for cluster formulation. Subsequently, a broader evaluation pitted three algorithms — K-Means, Hierarchical, and DBSCAN — against each other using transactional data features. The results excluded hierarchical clustering due to its failure to generate outcomes within a reasonable timeframe. K-Means and DBSCAN demonstrated highly comparable results, with DBSCAN exhibiting a slightly more sophisticated outcome. Ultimately, the K-Means algorithm was deemed the most effective, owing to its ease of implementation and providing a clearer understanding of the data, when compared to DBSCAN.

How Does Customer Clustering Impact Key Retail Performance Metrics? What Patterns and Characteristics Define Identified Customer Segments?

These inquiries found comprehensive and intertwined responses throughout Chapter 5: Data Driven Customer Segmentation. The identification of patterns and characteristics within the established customer segments contributed to a nuanced comprehension of both the commonalities and distinctions within the customer base. These clusters exert a substantial and applicable influence on key retail performance metrics. Furthermore, the insights derived from understanding these patterns and characteristics empower targeted customer engagement strategies, focused on enhancing the identified key retail performance metrics.

How Can Retailers Enhance Strategies Based on Data-Driven Customer Clustering Insights?

This chapter delved into a comprehensive response to this pivotal question. An important subquestion was posed by the retailer regarding the necessity for location-based targeted promotions. Despite the associated costs of implementing targeted promotions, we have concluded that the retailer stands to gain substantial benefits from heightened customer involvement, particularly as

our analyses revealed disparities in customer engagement and offer redemption between the two locations. Considering these disparities and a wealth of research highlighting the advantages of personalizing the customer experience — such as the statistic that 80% of consumers express a greater likelihood to engage with a company offering personalized experiences [17] — it is strongly recommended that the retailer places high importance on the implementation of targeted promotions.

6.5 Limitations and Recommendations

In acknowledging the scope and boundaries of our research, it is imperative to address the inherent limitations encountered during the study. This section delineates these limitations, providing valuable insights into the constraints that may impact the interpretation and generalization of our findings. Subsequently, we offer thoughtful recommendations aimed at mitigating these limitations in future studies, thus fostering a more robust and comprehensive exploration of the subject matter.

Limitations:

- One significant constraint in our demographic analyses revolved around missing data concerning cardholders. This absence of information has the potential to introduce biases and skew the outcomes of our analysis. The missing data was managed by exclusion.
- The presence of the "Uncensored" item category posed a challenge as it contributed to transactions without providing any informative data. Consequently, a strategic decision was made to exclude this category from most analyses related to item categories and clustering, while retaining its inclusion in other aspects of the study. This selective exclusion has the potential to introduce non-coherent trends in certain analyses, impacting the overall interpretability of the results.
- The RFM-focused clustering methods employed in our study may encounter potential limitations due to the restricted timeframe of the data, limited to March 2023. This temporal constraint poses a challenge to the accurate assessment of the recency feature, despite its deemed significance in K-Means Clustering. The reduced value of the recency feature within the constrained timeframe could impact the overall effectiveness of the clustering methods.

Recommendations:

- Addressing missing data through exclusion, particularly when alternative methods for gauging the missing data are nearly impossible, is considered the appropriate approach. This method was implemented in our study. Opting to exclude the associated data throughout the entire dataset is deemed the most optimal strategy to uphold the highest level of data integrity, even though it comes at the expense of losing some data.
- Implementing RFM-focused analyses with a strategically calculated weightage for retail strategy not only aligns with specific retailer needs but also mitigates the risk of inaccurately classifying a customer who visited in March 2023 as lost. Classifying lost customers necessitates a more comprehensive RFM dataset, beyond the limited timeframe explored in this study. Legg, Mancini, and Webb offer a detailed method for identifying when a customer

is lost in their article titled "Identifying When a Customer is Lost?" [20]. Leveraging this method could serve as a foundation for enhancing the precision of results obtained in our study.

6.6 Conclusion

Our study not only reaffirms established trends in the retail industry, such as the higher spending patterns of cardholders compared to non-cardholders, underscoring the significance of well-crafted loyalty programs for retailers but also delves into the nuances of customer shopping habits by analyzing extensively and clustering appropriately. By highlighting both similarities and disparities between the two locations of the same grocery store brand within the city, our research emphasizes the importance of tailoring promotions based on specific locations and engaging customer bases.

In evaluating clustering algorithms, K-Means Clustering, boasting efficiency and efficacy, emerged as notably superior for our dataset. We posit that our study contributes positively to the retail industry by uncovering nuanced patterns and employing clustering based on these uncovered insights to offer actionable goals for real-world application. While measuring key retail performance metrics post-implementation would provide detailed insights, we are confident in the positive contribution of our data-driven methods and results to the field of retail marketing.

Bibliography

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, 1994.

[2] M. Alves Gomes and T. Meisen. A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, 21(3):527–570, Sep 2023.

[3] D. Berthiaume. Study: The most loyal generation is 55-to-64-year-olds, 04 2019.

[4] P. Bruce, A. Bruce, and P. Gedeck. *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. O'Reilly Media, 2020.

[5] ChatGPT. Introduction to dbscan. Online, 2023. Generated by ChatGPT.

[6] D. Chen, S. L. Sain, and K. Guo. Data mining for the online retail industry: A case study of rfm model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19:197–208, 2012.

[7] N. Dawar and J. Singh. Usage of data mining for customers profiling in supermarkets and grocery stores. *International Journal of Applied Research*, 8(8):85–88, 2022.

[8] O. Dogan, E. Ayçin, and Z. Bulut. Customer segmentation by using rfm model and clustering methods: a case study in retail industry. *International Journal of Contemporary Economics and Administrative Sciences*, 8, 2018.

[9] A. Downey. *Think Stats*. O'Reilly Media, 2015.

[10] S. Firdaus and M. A. Uddin. A survey on clustering algorithms and complexity analysis. *International Journal of Computer Science Issues (IJCSI)*, 12(2):62, 2015.

[11] S. Follin and V. Fransson. The impact of gender and age on customer loyalty: A quantitative study of swedish customers' experiences of a loyalty program, 2013.

[12] R. Hariharan and S. Mahapatra. Retailing in india: issues and challenges. In *National conference on FDI*, 2013.

[13] J. Kaur, V. Arora, and S. Bali. Influence of technological advances and change in marketing strategies using analytics in retail industry. *International Journal of System Assurance Engineering and Management*, 11(5):953–961, Oct 2020.

[14] L. Majoor. Predicting the type of shopper (weekend or weekday) from online grocery data., 2018.

[15] V. Melnyk, S. van Osselaer, and T. Bijmolt. Are women more loyal customers than men? gender differences in loyalty to firms and individual service providers. *Journal of Marketing*, 07 2009.

[16] D. Papandrea. Survey: Loyalty programs more important than ever?, Jul 2022.

[17] Queue-it. Loyalty program statistics: 2022 market growth & trends, 2023.

[18] T. Raitaluoto. The role of customer segmentation in customer segmentation research, May 2023.

[19] S. Shariff, Z. Bakri, and P. Hamzah. Association rules for purchase dependency of grocery items. *Social and Management Research Journal*, 13:61, 12 2016.

[20] T. Webb, M. Legg, and M. Mancini. Identifying when a customer is lost?, Nov 2021.

Acknowledgements

I want to express my heartfelt gratitude to my parents, whose unwavering love and support have been a constant throughout my life. My sister has been an inspiring role model, and I am grateful for her guidance. A special acknowledgment goes to my esteemed educators and advisors, Prof. Fabrizio Rossi, Prof. Stefano Smriglio, and Prof. Andrea Manno, whose consistent support and guidance have been invaluable on this academic journey.

I extend my thanks to all the teachers who have played a significant role in shaping and motivating me to achieve this academic milestone. I am also deeply appreciative of my colleagues and friends who accompanied me on this journey, sharing laughter, tears, love, and mutual support.

Last but not least, I want to express gratitude to myself for believing in my abilities and continually pushing my limits, embracing a path of self-improvement every day.

Vipul Chalotra, L'Aquila, December 2023