Segmentation and Trend Analysis in U.S. Retail - Geodemographics, Consumer Behavior, and Customer Lifetime Value

By

Claudia Dispinzeri

Instructor: Dr. Zhuojun Gu

BITM 603 (9863) - Business Analytics & Data Mining

University of Albany

Master of Science in Business Analytics

December 2024

ABSTRACT

This report represents the culmination of a data-driven analysis project aimed at addressing key business questions raised by the company's leadership. Utilizing advanced **classification** and **clustering methodologies**, we focused on delivering actionable insights related to customer segmentation, behavior prediction, and product affinity identification. By applying these techniques, we were able to connect theoretical concepts to real-world challenges, generating strategic recommendations to support informed decision-making. The findings presented in this report are designed to enhance targeted marketing initiatives, optimize resource allocation, and strengthen customer relationships, aligning with the company's strategic goals.

TABLE OF CONTENTS

INTRODUCTION

In this project, we will explore data analysis techniques to address four strategic problems related to customer behavior in the U.S. retail sector. The dataset used, "shopping_trends.csv", which is detailed below, was obtained from the Kaggle platform and contains behavioral information and consumer trends. Our goal is to provide insights to support strategic decision-making by the company's management, assisting in the implementation of targeted marketing campaigns and maximizing customer value.

The first business problem concerns the factors influencing a customer's decision to subscribe to the company's services. To identify the determinants of this decision, we will use a combination of techniques, such as Lasso CV, Backward Elimination, and Logistic Regression. The model will be evaluated using metrics including, but not limited to, Precision, Recall, F1-Score, Accuracy, and AUC-ROC. These metrics will allow us to analyze the model's ability to distinguish subscribers from non-subscribers, ensuring an understanding of the factors influencing subscriptions, such as gender, age, preferred payment methods, purchase frequency, among others.

Regarding customer segmentation for marketing campaigns, we will apply clustering methods, such as K-Means and Hierarchical Clustering, to create clusters based on the strategies labeled S1, S2, and S2H, which are based on behavioral, geographical, and demographic attributes. Cluster analysis will be optimized using the Elbow Method and Silhouette Score, ensuring that the segments are meaningful and well-differentiated. These groups will provide the foundation for personalized marketing campaigns, promoting greater efficiency in engagement efforts.

To identify customer affinities for product categories, we will explore their preferences for both general categories and specific items, using separate strategies: S3 for categories and S4 for items. For this purpose, we will use Hierarchical Clustering. Affinities will be evaluated through dendrograms, enabling a clear visual analysis of the proximity between groups. This approach will help us identify specific consumption patterns, such as categories or items with higher recurrence among certain clusters, enabling more accurate recommendations and optimized inventory management.

With respect to grouping customers by lifetime value (Customer Lifetime Value - CLV), we will implement K-Means to segment customers into clusters (S5 strategy), categorizing them as low, medium, or high value. This analysis will be essential to prioritize retention and loyalty efforts. We will evaluate metrics such as the mean and variance of CLV within each cluster, ensuring that strategies are aligned with the revenue generation potential of each group.

Data preparation will include cleaning, transformation, and the creation of new variables, such as Frequency of Purchases per Year, Subscription Status (binary), Region, Latitude, and Longitude, among others. Categorical variables will be converted into dummies, and numerical variables will be standardized using the StandardScaler, ensuring consistency across different scales.

Clusters will be statistically validated through centroid analysis and performance metrics, such as the within-cluster sum of squares (WCSS). Additionally, we will explore the relationship between clusters and key variables, such as region, purchase frequency, and product categories, to ensure that segmentations are useful for decision-making.

Finally, the final report will present a formal description of the methodology, with technical terms defined in a glossary, as well as the results, which will be the most relevant part for the management team. Graphs and strategic recommendations will be presented, guiding the adoption of campaigns in a combined, isolated, or staggered manner, always aiming to improve the company's results.

The following additional files are provided:

1. **Project_TeamCLAUDIADISPINZERI_V3:** contains 98% of the code used in the project.
2. **Project_TeamClaudiaDispinzeri_Coordinates.ipynb:** supplementary file that generates the coordinates using the *geopy.geocoders (Nominatim)* library.
3. **dataset_with_coordinates.csv:** processed dataset containing the geographic coordinate values for immediate use, due to the time required for processing.
4. **shopping_trends.csv:** the original dataset used as the project's foundation.

By using a data-driven approach, this project will enable the generation of meaningful insights, empowering the organization to make strategic decisions based on evidence and aligned with the challenges and opportunities of the U.S. retail market.

A. DATASET USED FOR THE FINAL PROJECT

The dataset chosen for the final project, named “shopping_trends.csv”, was obtained from the Kaggle platform, with the exact source being “https://www.kaggle.com/datasets/bhadramohit/customer-shopping-latest-trends-dataset/data”. It provides detailed information on consumer shopping behavior in the United States, reflecting typical trends in the country's retail sector. Comprising 3,900 rows and 19 original columns, the dataset combines categorical and numerical variables, enabling detailed analyses, such as classification and clustering, which were selected for our project. This diversified structure makes the dataset ideal for exploring shopping patterns, customer segmentation, and developing data-driven strategies.

Initially, the dataset contained many categorical variables, such as:

1. **Gender**: Customer gender (e.g., Male, Female).
2. **Item Purchased**: Product purchased (e.g., Blouse).
3. **Category**: Product category (e.g., Clothing, Electronics).
4. **Location**: Customer location (e.g., Montana).
5. **Size**: Product size (e.g., M, L).
6. **Color**: Product color (e.g., Green).
7. **Season**: Season of purchase (e.g., Spring).
8. **Subscription Status**: Customer subscription status (e.g., Yes, No).
9. **Payment Method**: Payment method used (e.g., Credit Card, PayPal).
10. **Shipping Type**: Type of shipping (e.g., Free Shipping).
11. **Discount Applied**: Indicates whether a discount was applied (e.g., Yes, No).
12. **Promo Code Used**: Indicates whether a promotional code was used (e.g., Yes, No).
13. **Preferred Payment Method**: Customer's preferred payment method (e.g., PayPal).
14. **Frequency of Purchases**: Frequency of purchases (e.g., Every 3 months).

There were also some relevant numerical variables, such as:

15. **Customer ID**: Unique identifier for each customer (e.g., 1-3900).
16. **Age**: Customer age (range: 18-70 years).
17. **Purchase Amount (USD)**: Purchase value in USD (e.g., 20-100).
18. **Review Rating**: Product rating (e.g., 2.5-5).
19. **Previous Purchases**: Number of previous purchases (e.g., 1-50).

After preprocessing and adjustments, the final version of the dataset includes the following columns:

Original Variables:

- Age, Gender, Item_Purchased, Category, Purchase_Amount_(USD), Location, Size, Color, Season, Review_Rating, Subscription_Status, Payment_Method, Shipping_Type, Discount_Applied, Previous_Purchases, Preferred_Payment_Method, Frequency_of_Purchases, Promo_Code_Used

New Variables Created:

- Frequency_Purch_per_year, Subscription_Num, Latitude, Longitude, Region, Cluster_S1, Cluster_S2, Cluster_S2H, Cluster_S3, Cluster_S4, Cluster_S5, CLV (Customer Lifetime Value), CLV_Normalized.

Removed Variable:

- **Customer ID** was removed since its values were identical to the dataset's index and not relevant for our analyses.

The interaction between categorical and numerical variables provides important insights into regional and demographic patterns in the United States. These variables enable behavioral and trend analyses, such as identifying the distribution of purchases across different genders, regions, or commonly used payment methods in the U.S. retail sector. Moreover, they allow for the identification of metrics such as average customer spending, the most profitable product categories, and purchasing trends over time.

In conclusion, this dataset reflects typical characteristics of the U.S. retail market, including seasonality, demographic preferences, and promotional strategies. Its organized structure facilitates the application of clustering and segmentation methods, making it ideal for the personalized strategies and educational analyses we aimed to achieve. For these reasons, we selected this dataset for our final project.

B. REPORT FOR DIRECTORS

1. Introduction

This report presents the results and insights derived from our comprehensive analysis, conducted to address the key business questions raised by leadership. Our primary goal was to explore customer behaviors, preferences, and value factors to support informed decision-making and improve strategic outcomes. Using data-driven methodologies and advanced segmentation techniques, we aimed to uncover actionable insights that enable the organization to enhance marketing strategies, optimize resource allocation, and strengthen customer relationships.

Through this project, we systematically analyzed customer data to identify patterns and trends directly aligned with the highlighted business challenges. The following sections outline our approach, detailed findings, and tailored recommendations to address the critical issues raised by the leadership team.

2. Business Problems

Our work was driven by critical questions raised by leadership, reflecting the strategic challenges the company currently faces. In a highly competitive market, understanding customer behavior and preferences is essential for delivering personalized experiences, increasing retention, and directing more effective marketing campaigns.

To address these challenges, our team structured the project around four key questions, each designed to provide actionable insights and targeted strategies:

1. What factors influence customers' decisions to subscribe to our services?
2. How can we segment our customers into distinct groups for targeted marketing campaigns?
3. How can we identify customers' affinity for certain product categories?
4. How can we group customers based on their lifetime value (CLV)?

These questions were crafted to address both immediate needs and long-term objectives, with a focus on enhancing the customer experience, optimizing marketing resources, and maximizing overall value. In the following sections, we outline how the project was structured to address these questions and present the key findings that underpin our recommendations.

Throughout this report, each Business Problem is presented with a focus on two key aspects: (1) Methodology, where we explain the technical approach used to address the problem and how we arrived at the results. This section is more technical in nature. (2) Results, Conclusions, and Recommendations, which provides the most relevant information, addressing the core Business Problem itself—the primary focus of leadership. If your interest lies in the business outcomes, you may wish to concentrate on this section. Technical definitions and terms are provided in the Glossary at the end of this report.

3. Describing Initial Work with the Dataset

To answer the proposed questions, we began by performing essential data preparation tasks, including verification, cleaning, and enrichment of the dataset.

Initially, we added Latitude and Longitude columns, as the original dataset only included state names. This addition allowed for more precise geospatial analysis, helping us identify regional patterns. We also introduced a categorical column called Region, which mapped each state to one of the four major U.S. regions: Northeast, Midwest, South, and West. This categorization aimed to determine whether region or numerical location significantly influenced subsequent analyses.

We retained the original Subscription_Status column but created a new binary version where “0” represented “Non-Subscriber” and “1” represented “Subscriber.” This transformation simplified the use of the variable in statistical and machine learning models.

To better quantify customer behavior, we transformed the Frequency_of_Purchases column, initially categorical (e.g., “Weekly,” “Monthly”), into a numerical variable representing annual purchase frequency. For example, a monthly purchase was converted to a value of 12 annual purchases. This change provided a more consistent and meaningful metric for analysis.

After these adjustments, we conducted statistical analyses to evaluate the number of unique values in each non-numerical column and calculated descriptive statistics, including mean, median, standard deviation, minimum, maximum, and mode for relevant variables. This preliminary analysis provided information about data distribution and potential inconsistencies.

We then plotted histograms to visualize variable distributions, overlaying lines representing the mean, median, and mode (for categorical data). These visualizations highlighted central tendencies and asymmetries in the data, helping us identify potential outliers. While we initially considered addressing these outliers, further analysis showed they were not representative enough to skew results. Additionally, location-based outliers (latitude and longitude) were retained as they provide critical geospatial insights.

We also created a correlation matrix to examine relationships between numerical variables. The results indicated low correlations, suggesting that the variables were largely independent—a favorable condition for modeling as it reduces the risk of multicollinearity.

To ensure robust analysis and visualization, we utilized Python libraries for graphical representation, and machine learning and statistical tools for deeper analysis. Throughout the process, we applied rigorous cleaning and transformation practices to maintain data integrity and reliability.

These initial steps provided a strong foundation for understanding the dataset’s structure and characteristics, enabling us to extract insights and proceed effectively to the subsequent stages of the project.

4. Answering the Business Problem 1: Factors Influencing Customer Subscription

In response to the question “**What factors influence a customer’s decision to subscribe to our services?**” we conducted a detailed analysis in structured steps, using statistical and machine learning techniques. The goal was to identify the most important factors that directly impact subscription decisions, providing a clearer understanding of customer behavior and supporting more effective decision-making strategies.

4.1. Methodology

Our team conducted an exploratory analysis to observe how subscription relates to categorical and quantitative variables such as Discount_Applied, Gender, Promo_Code_Used, Season, Category, Payment_Method, Item_Purchased, Shipping_Type, and Region. This step provided an initial overview and guided the subsequent work. Given the high number of non-numerical variables in the dataset, we converted these variables into binary format using a technique called one-hot encoding. This approach transformed categories into individual columns, making the analysis easier but significantly increasing the dataset's dimensionality.

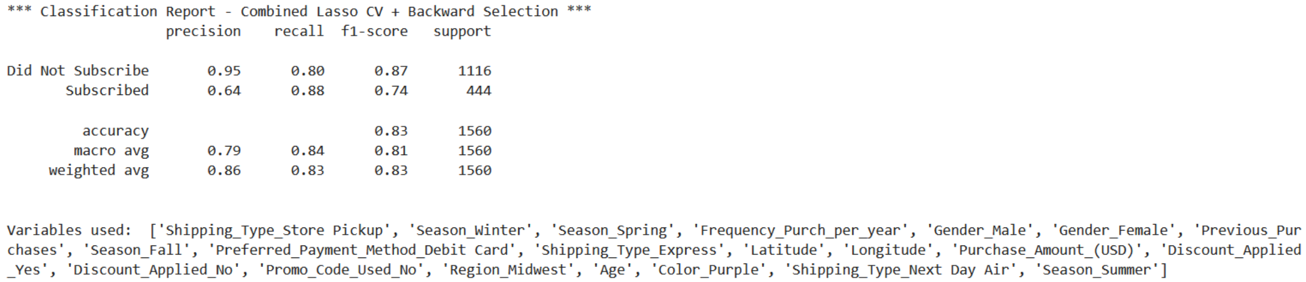
To address the challenges of high dimensionality, such as computational costs and risks of bias or overfitting, we applied variable selection techniques. Our primary goal was to identify the factors influencing customer subscriptions using variables selected through a combination of the Lasso-CV and Backward Selection models.

By applying Logistic Regression to the variables derived from the combined models, we achieved an overall accuracy of 83%, indicating the model's effectiveness in correctly classifying cases. Additionally, the AUC-ROC of 0.892 demonstrated the model's excellent ability to distinguish between subscribers and non-subscribers, even in complex scenarios. Metrics such as precision and recall were also crucial for understanding the results in each category. For instance, for subscribing customers, the recall (0.88) indicates that 88% of subscribers were correctly identified, while the precision (0.64) shows that 64% of the subscription predictions were accurate based on the applied variables. The balance between these metrics is reflected in the f1-score, which was 0.74 for subscribers, further reinforcing the model's reliability.

This combination of metrics validated the quality of the exploratory analysis and the final model, providing a comprehensive view of the factors that most influence subscription decisions.

For more details on the definitions and technical terms mentioned, please refer to the glossary at the end of this report.

The results show that the combined techniques delivered robust performance, successfully identifying the most significant variables related to subscription.



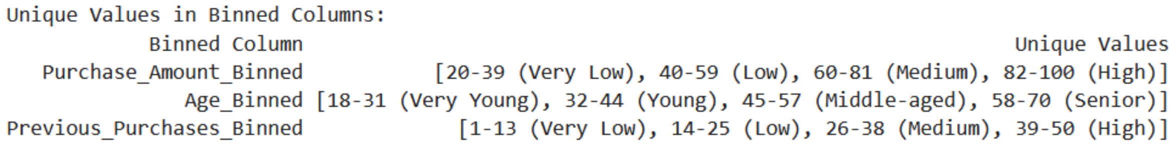
```
*** Classification Report - Combined Lasso CV + Backward Selection ***
                  precision    recall  f1-score   support

Did Not Subscribe       0.95      0.80      0.87      1116
       Subscribed       0.64      0.88      0.74       444

         accuracy                           0.83      1560
        macro avg       0.79      0.84      0.81      1560
     weighted avg       0.86      0.83      0.83      1560


Variables used:  ['Shipping_Type_Store Pickup', 'Season_Winter', 'Season_Spring', 'Frequency_Purch_per_year', 'Gender_Male', 'Gender_Female', 'Previous_Pur
chases', 'Season_Fall', 'Preferred_Payment_Method_Debit Card', 'Shipping_Type_Express', 'Latitude', 'Longitude', 'Purchase_Amount_(USD)', 'Discount_Applied
_Yes', 'Discount_Applied_No', 'Promo_Code_Used_No', 'Region_Midwest', 'Age', 'Color_Purple', 'Shipping_Type_Next Day Air', 'Season_Summer']
```

To enhance the analysis, we used the binning technique to group continuous data into strategic categories. This approach allowed us to assess impacts within more representative ranges. The table below summarizes these groupings.



```
Unique Values in Binned Columns:
            Binned Column                                              Unique Values
   Purchase_Amount_Binned          [20-39 (Very Low), 40-59 (Low), 60-81 (Medium), 82-100 (High)]
               Age_Binned [18-31 (Very Young), 32-44 (Young), 45-57 (Middle-aged), 58-70 (Senior)]
Previous_Purchases_Binned            [1-13 (Very Low), 14-25 (Low), 26-38 (Medium), 39-50 (High)]
```

The significant variables identified using the techniques were as follows:

Region_Midwest: Represents customers located in the Midwest. This variable was analyzed to understand regional differences that could influence subscription rates.

Shipping_Type: Refers to the delivery methods used by customers, such as Store Pickup, Next Day Air, and Express. Each method was evaluated for its influence on subscription rates.

Discount_Applied_Yes: Indicates the presence or absence of applied discounts. This variable was analyzed to understand how promotional offers directly impact subscription decisions.

Preferred_Payment_Method: Represents the preferred payment methods of customers, such as debit and credit cards. The relationship between these preferences and subscription decisions was explored.

Frequency_Purch_per_year: Refers to the annual purchase frequency of customers. It was analyzed to identify behavioral patterns and their impact on subscription rates.

Promo_Code_Used_No: Examines the absence of promotional code usage. This variable was considered to understand how behaviors related to promotions influence subscription decisions.

Season: Represents the seasons, such as Fall, Spring, Summer, and Winter. This variable was analyzed to identify seasonal variations impacting consumer behavior.

Purchase_Amount_Binned: Groups spending amounts into specific ranges, enabling a more targeted analysis of how average spending affects subscription decisions.

Age_Binned: Refers to the age ranges of customers. This variable was grouped to understand how different age groups behave regarding subscriptions.

Previous_Purchases_Binned: Groups customers' purchase histories into strategic intervals. This analysis assessed the impact of prior engagement on subscription decisions.

Gender: Refers to customers' genders, such as male and female. Including this variable helped identify how subscription behavior varies between genders. For example, male customers demonstrated higher subscription rates compared to those who did not subscribe.

Latitude and Longitude: Geographic indicators were included to analyze the influence of customers' physical locations on subscription decisions. These variables enabled mapping regional patterns and identifying geographic areas with higher or lower subscription propensities. For instance, some latitude ranges exhibited higher non-subscription rates, highlighting potential regional factors' impact.

4.2. Results, Conclusions, and Recommendations

The detailed analysis of the data shed light on several factors that directly influence customers' decisions to subscribe to the services. Significant trends were identified across demographic, behavioral, and geographic variables, providing a foundation for targeted marketing strategies and personalized offers.

For instance, the Midwest region exhibited lower subscription rates compared to other regions, with only 25.72% of customers subscribing. This highlights the need for a more focused approach to attract consumers from this area. Delivery method also proved to be an important factor, with the Store Pickup option showing the highest subscription rate at 29.23%. This finding suggests that convenient delivery options can serve as a compelling differentiator for this audience.

The11 application of discounts emerged as one of the most impactful factors in the analysis. Customers who received discounts had a subscription rate of 62.79%, underscoring that strategic promotions can be decisive in converting undecided customers. This price sensitivity was further observed in spending behavior: consumers in the 20-39 spending range (Very Low) showed a lower likelihood of subscribing, with only 26.92% opting for the service.

Payment method also played a significant role. Customers who chose to pay with debit cards were more likely to subscribe, with a subscription rate of 29.87%. This behavior highlights the importance of offering payment options that align with customer preferences. Similarly, annual shopping frequency revealed a clear pattern: consumers who made 52 purchases per year had a higher subscription rate (29.13%), indicating that regular shopping habits are associated with greater likelihood of subscribing.

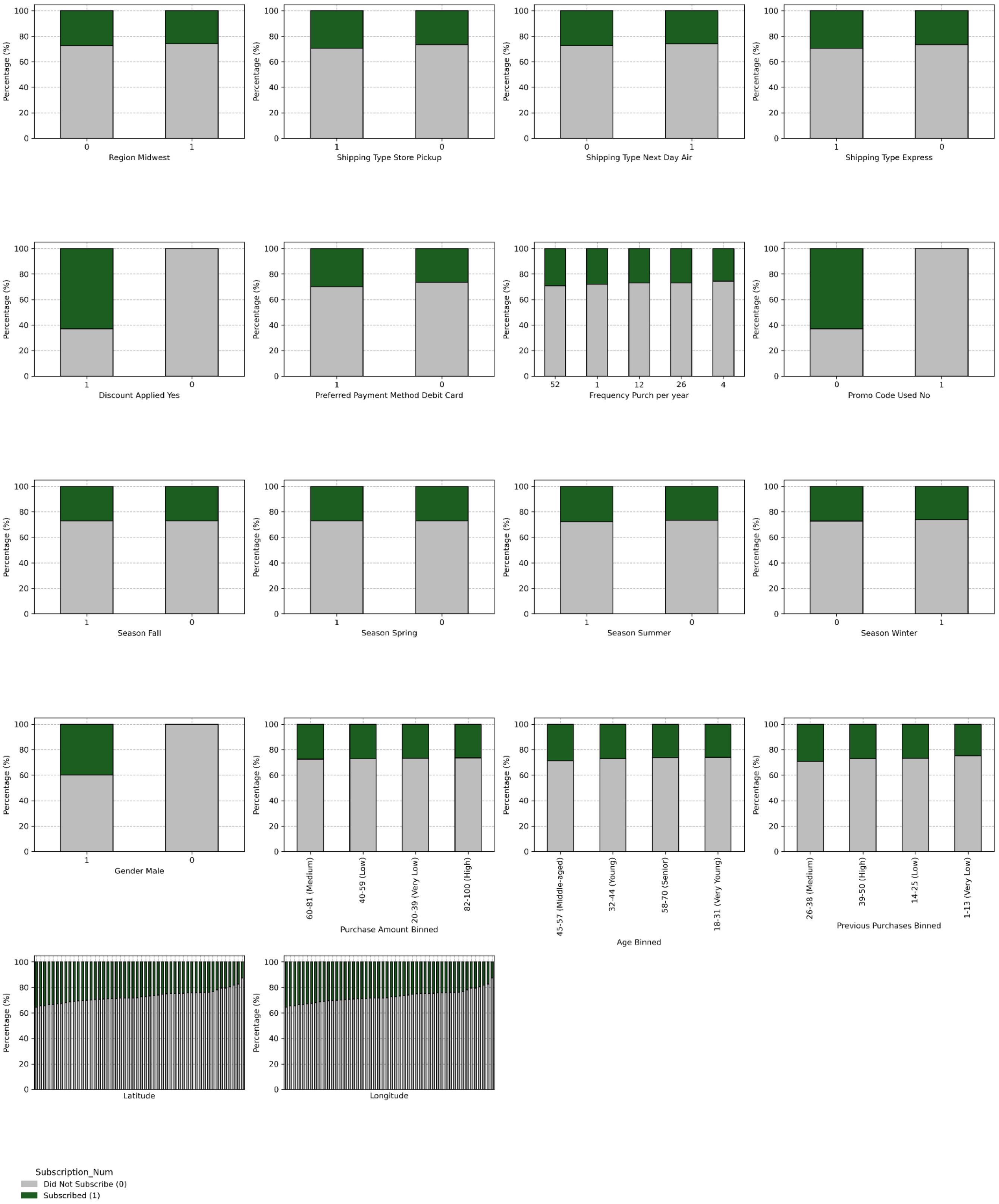
Gender analysis showed that men had a subscription rate of 39.71%, significantly higher than other groups. This suggests potential for more targeted campaigns aimed at women to balance subscription rates across genders. Age group analysis also revealed important insights: customers aged 18-31 (Very Young) recorded the lowest subscription rate at 25.97%, signaling the need for tailored strategies to engage younger audiences.

Geographic location emerged as another key factor. Certain latitude and longitude ranges showed striking differences in subscription behavior. For example, customers located near latitude 38.27 had a high non-subscription rate of 87.30%, while regions around longitude -89.68 showed lower non-subscription rates at 17.33%. These geographic patterns present opportunities for regionalized and personalized marketing campaigns.

Finally, historical customer behavior played a significant role. Customers with fewer previous purchases (1-13, Very Low) showed lower engagement, with a subscription rate of only 24.85%. This finding emphasizes the importance of encouraging active participation from less engaged customers.

These conclusions highlight the value of adopting a strategic and personalized approach, taking into account the most influential factors identified in the analysis. Campaigns leveraging discounts, convenient delivery methods, seasonal promotions, and incentives for specific groups, such as young people and women, are highly recommended. Additionally, using geographic and historical data can further enhance strategy personalization, ensuring a significant impact on subscription conversion rates. Aligning these variables with future initiatives has the potential to strengthen customer relationships and optimize outcomes.

Relationship Between Subscription_Num and Selected Variables



5. Answering Business Problem 2: Segmenting Customers for Targeted Marketing Campaigns

5.1. Methodology

To address the question of how to segment our customers into distinct groups for targeted marketing campaigns, we divided our analysis into three main strategies: Strategy 1 (S1), Strategy 2 (S2), and Strategy 2H (S2H). Each approach employed different clustering techniques to identify unique behavioral and regional patterns. Below, we outline how each strategy was implemented.

5.1.1. Strategy 1 (S1): Clustering Behavioral Patterns

In this strategy, we focused on identifying customer behavioral patterns based on demographic and behavioral data. First, we selected numerical variables and transformed categorical variables into binary representations (one-hot encoding), such as gender, applied discounts, payment methods, and shipping types. Next, we normalized the data to ensure that all variables had equal weight in the clustering process.

To determine the optimal number of clusters, we used two complementary approaches: the Elbow Method and the Silhouette Score, ultimately selecting 8 clusters. This decision was based on balancing model complexity with segmentation quality.

After applying K-means clustering, we identified eight distinct groups of customers, each with unique characteristics such as preferences for discounts, frequent shopping habits, or specific shipping methods. These results enable us to create personalized strategies to better engage each segment.

5.1.2. Strategy 2 (S2): Clustering Regional Consumption Patterns

This strategy was designed to explore customer patterns with a focus on regional differences. We aggregated data by state, consolidating metrics such as total purchase value, average purchase frequency, and the number of customers. Additionally, we included regional and behavioral variables, such as shipping preferences and geographic regions (Northeast, Midwest, South, and West).

After normalizing the aggregated data, we applied K-means clustering again. Using the same evaluation methods as in S1, we determined 4 clusters as the optimal grouping, with a Silhouette Score of 0.18. This approach revealed behavioral differences across regions, such as spending patterns, purchase frequency, and shipping preferences. These insights are essential for designing region-specific marketing campaigns tailored to the unique characteristics of each area.

5.1.3. Strategy 2H (S2H): Clustering Hierarchically Refined Regional Patterns

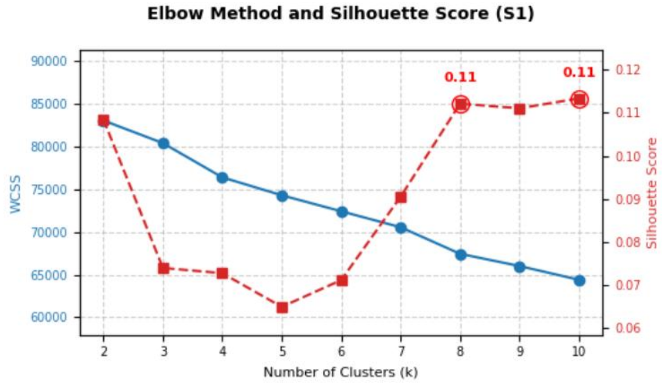
The third strategy extended Strategy 2 by employing Hierarchical Clustering to refine the regional segmentation. We retained the same regional and behavioral variables as in S2 but used the Average Linkage method with Euclidean distance to uncover deeper relationships between clusters.

To determine the optimal number of clusters, we combined visual analysis of the dendrogram with the Silhouette Score, leading to the identification of 6 clusters with a Silhouette Score of 0.15. This technique allowed for more detailed segmentation, exploring variations within the regions already identified in Strategy 2.

5.2. Results, Conclusions, and Recommendations

5.2.1. Strategy 1 (S1): Clustering Behavioral Patterns

When applying the S1 strategy, we tested different numbers of clusters to determine the optimal choice. Our evaluation relied on Silhouette Scores, which indicated the best results for 8 and 10 clusters. After careful analysis, we decided on 8 clusters, as they provided a balance between efficient segmentation and ease of strategic implementation.



Each cluster was named and described based on its key characteristics, allowing us to craft specific strategies tailored to the needs and preferences of each group. Below are the recommendations we developed for each cluster, based on variables and their central values:

Cluster 1: Discount-Sensitive Male Shoppers (409 clients): Cluster 1 consists primarily of male customers who are highly responsive to discounts (high *Discount_Applied_Yes*). These customers tend to focus on promotional offers and prefer using standard shipping as their main delivery method. While their purchasing frequency is moderate, their average transaction value remains relatively low. To effectively engage this group, the recommendation is to implement targeted promotional campaigns and seasonal discounts. Highlighting cost-effective products and offering value-driven promotions will appeal to their price sensitivity and drive greater purchasing activity.

Cluster 2: Transaction-Focused Female Shoppers (494 clients): Cluster 2 is dominated by female customers who rarely take advantage of discounts (high *Discount_Applied_No*). These shoppers prioritize fast shipping, with a clear preference for 2-Day Shipping options. They predominantly use bank transfers as their payment method while showing low interest in digital alternatives like PayPal or Venmo. Although their annual purchase frequency is low, their transaction value remains slightly positive. To engage this segment, focusing on speed and convenience is key. Offering expedited shipping options and ensuring a seamless shopping experience with reliable payment methods will enhance satisfaction and loyalty.

Cluster 3: Frequent Discounted Shoppers (659 clients): This cluster is characterized by high sensitivity to discounts and a male-dominated demographic. Their buying behavior is heavily influenced by promotions, and they show a strong preference for free shipping. While their purchase frequency is slightly below average, they engage more during promotional periods. They tend to use a mix of cash and credit cards infrequently, with minimal usage of these payment methods. The ideal strategy for this group is to emphasize attractive discounts and free shipping offers. Campaigns that encourage bulk purchases through targeted promotions can maximize engagement and transaction value.

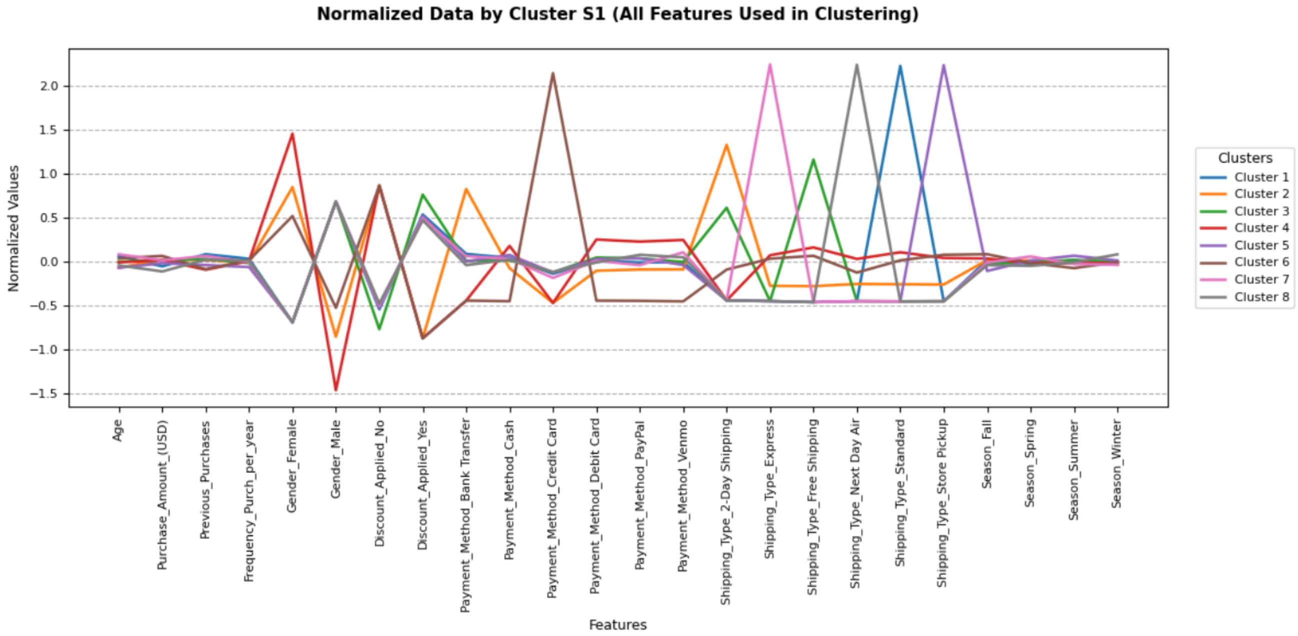
Cluster 4: High-Spending Female Shoppers (671 clients): Cluster 4 consists predominantly of female customers who exhibit slightly above-average transaction values and consistent purchase frequency. This group values quality and reliability, with a slight preference for express shipping options. They show a low inclination toward using credit cards as their payment method, preferring alternative methods like debit cards or PayPal. To engage this segment effectively, the recommendation is to promote premium products and offer exclusive deals that emphasize quality and reliability. Highlighting reliable shipping options and providing personalized shopping experiences will further enhance engagement and satisfaction.

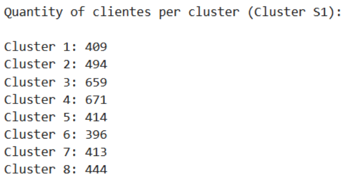
Cluster 5: Male Practical Shoppers (414 clients): Cluster 5 consists mainly of male customers who engage in infrequent purchases with slightly below average transaction values. This group values convenience, with a strong preference for in-store pickup over standard shipping. Their moderate engagement with discounts indicates a focus on practical and reliable shopping methods. To appeal to this segment, enhancing the in-store pickup experience is crucial. Offering localized promotions and emphasizing the ease and convenience of in-store pickup will strengthen their connection to the brand.

Cluster 6: High-Frequency Female Shoppers (396 clients): Cluster 6 consists predominantly of female customers who exhibit high purchase frequency and moderate transaction values. Contrary to the initial description, this group shows low responsiveness to discounts, indicating they do not heavily rely on promotional offers. They prefer credit cards as their main payment method and have minimal engagement with digital alternatives like Venmo. Their shipping preferences are slightly inclined towards express shipping, though not strongly. To engage this group, focus on loyalty programs that reward frequent purchases and offer exclusive benefits that do not solely rely on discounts. Emphasizing credit card rewards and enhancing the shopping experience with personalized offers can maximize their long-term value and strengthen loyalty.

Cluster 7: Delivery-Oriented Practical Shoppers (413 clients): Cluster 7 comprises male customers who exhibit moderate spending habits and a strong preference for express shipping. They demonstrate steady purchasing behavior throughout the year, with only slight responsiveness to discounts. This group values delivery speed and convenience, favoring express shipping over 2-Day Shipping options. To retain this segment, focus on enhancing express shipping services and ensuring a seamless delivery experience. Highlighting premium shipping options and maintaining reliable, fast delivery will resonate well with their preferences.

Cluster 8: High-Value Infrequent Male Buyers (444 clients): Cluster 8 consists predominantly of male customers who engage in infrequent purchases with slightly below average transaction values. This group shows a strong preference for premium delivery options, specifically Next Day Air. They have a low preference for credit cards and are only slightly responsive to discounts. To effectively engage this segment, the recommendation is to focus on personalized shopping experiences and premium delivery services. Offering exclusive products and ensuring fast, high-quality delivery can appeal to their preference for convenience and reliability, even though their transaction values are not as high as initially described.





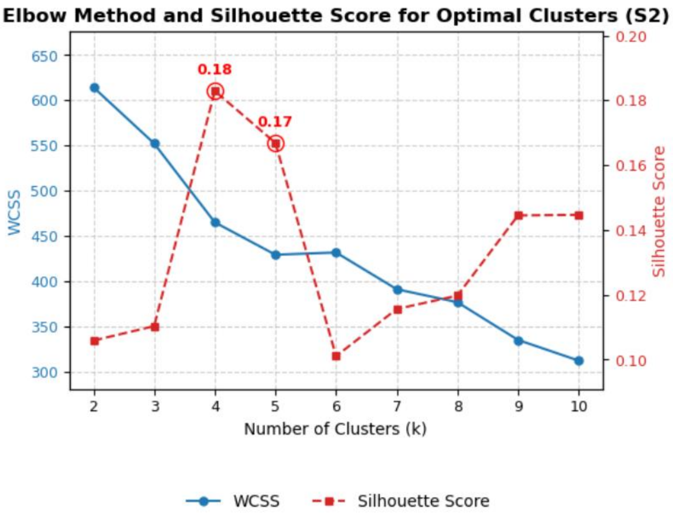
```
Quantity of clientes per cluster (Cluster S1):

Cluster 1: 409
Cluster 2: 494
Cluster 3: 659
Cluster 4: 671
Cluster 5: 414
Cluster 6: 396
Cluster 7: 413
Cluster 8: 444
```

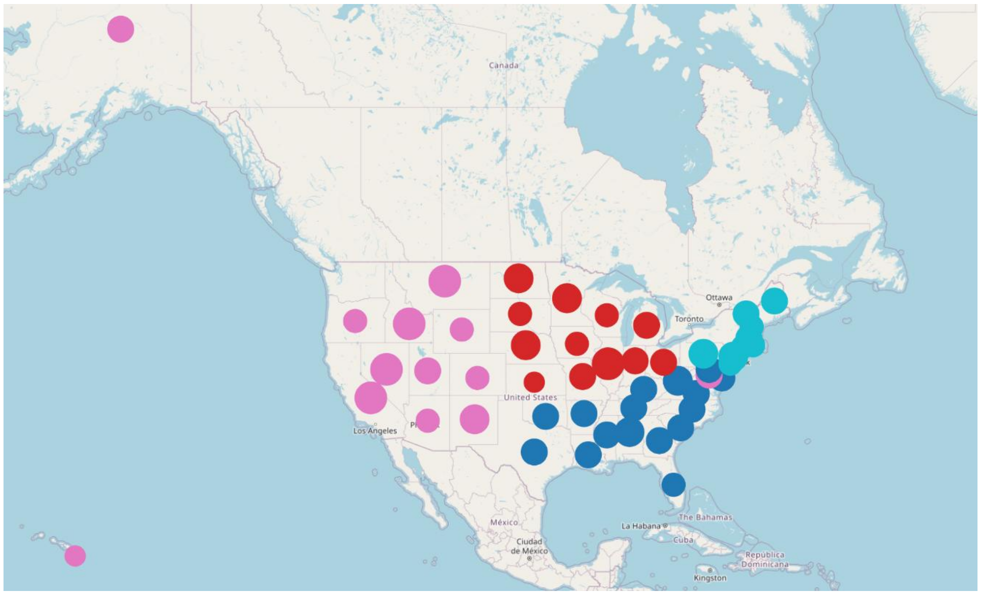
This approach allows for detailed segmentation, providing actionable insights for targeted and personalized marketing efforts. By aligning strategies with the specific interests of each group, we can maximize engagement and optimize campaign performance.

5.2.2. Strategy 2 (S2): Clustering Regional Consumption Patterns

In the application of the K-means Clustering method in the S2 strategy, which focused on regional analyses, we determined the optimal number of clusters using the Silhouette Score. The score reached 0.18, supporting the creation of 4 distinct clusters. This segmentation allowed us to identify unique purchasing patterns across different regions, enabling more targeted strategic actions.



Cluster S2 - Regional Consumption Patterns



```
Cluster Legend:
● Cluster 1 | Total Purchased: $75275.00 | Customers: 1271 | Avg: $59.23 | Median: $58.00 | Most Common Category: Clothing | Most Common Item: Hoodie
● Cluster 2 | Total Purchased: $55584.00 | Customers: 937 | Avg: $59.32 | Median: $60.00 | Most Common Category: Clothing | Most Common Item: Sandals
● Cluster 3 | Total Purchased: $62289.00 | Customers: 1018 | Avg: $61.19 | Median: $62.00 | Most Common Category: Clothing | Most Common Item: Backpack
● Cluster 4 | Total Purchased: $39933.00 | Customers: 674 | Avg: $59.25 | Median: $59.00 | Most Common Category: Clothing | Most Common Item: Blouse
```

The identified clusters were named and described based on their key regional and behavioral characteristics. Each cluster was also paired with specific recommendations:

Cluster 1: Southern Budget Shoppers (1271 clients): Cluster 1 includes customers primarily from the Southern region who demonstrate moderate total purchases and a relatively high transaction frequency compared to other clusters. These shoppers are particularly responsive to free shipping and moderately engage with express shipping options. Their purchasing behavior suggests that they seek value for money without compromising convenience. To better engage this group, the strategy should focus on localized promotions and free shipping offers tailored to the cultural and economic preferences of the Southern region. Highlighting cost-effective products and providing occasional express shipping incentives can drive greater engagement and loyalty.

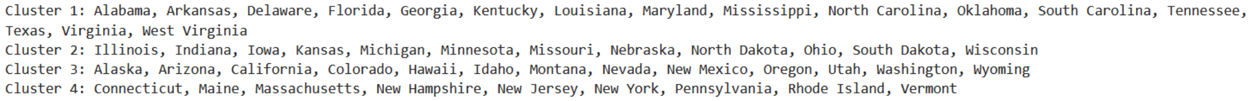
Cluster 2: Midwest Occasional Buyers (937 clients): Cluster 2 includes customers primarily from the Midwest region who exhibit low transaction frequency and below-average purchasing behavior. This group shows minimal engagement with premium shipping options and relies more on standard shipping or in-store pickup. Their below-average spending patterns suggest untapped growth potential. To encourage more frequent purchases, it is recommended to introduce personalized loyalty programs and exclusive rewards that align with their preferences. Enhancing the in-store pickup experience and offering incentives for regular transactions can strengthen their connection with the brand and increase purchase frequency.

Cluster 3: Western Practical Shoppers (1018 clients): Cluster 3 consists of customers primarily from the Western region who exhibit high average purchase values but only slightly below-average purchase frequency. While they demonstrate a preference for premium delivery options, such as Next Day Air, they also make use of free shipping to some extent. This indicates that their behavior balances quality and practicality, focusing on purchasing fewer, high-value items with a preference for convenience and premium delivery. The strategy for this group should emphasize premium products and personalized recommendations while also acknowledging their practical purchasing habits. Highlighting high-quality, value-driven offerings, combined with occasional free shipping promotions, can further enhance engagement and drive sustained value.

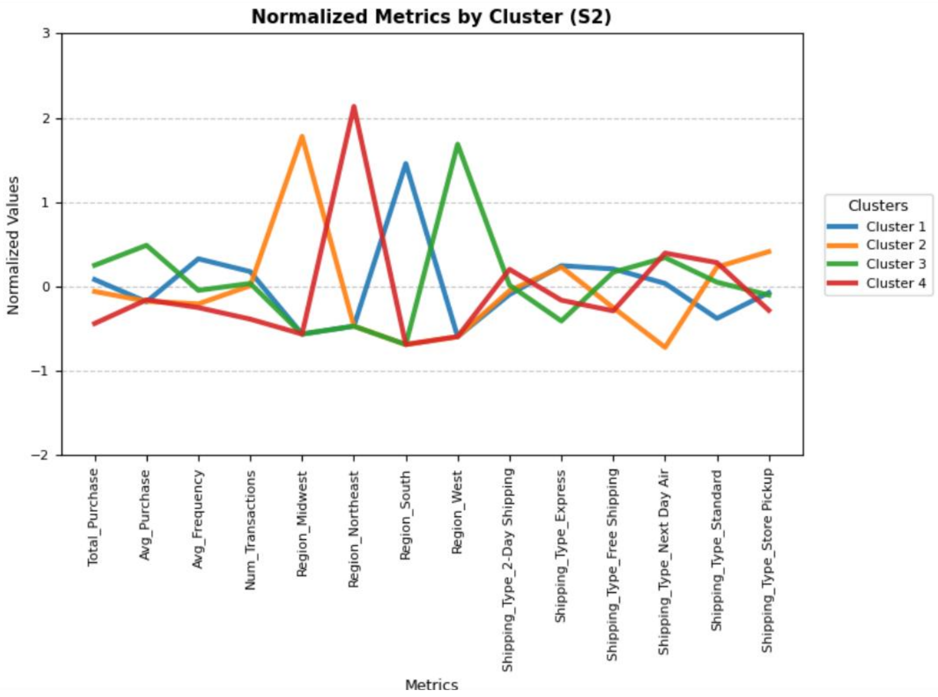
Cluster 4: Northeastern Value Shoppers (674): Cluster 4 represents customers primarily from the Northeast region who make infrequent purchases but with slightly below-average transaction values. They display a slight preference for standard shipping and Next Day Air options but are less responsive to free shipping and express methods. To capitalize on this group's purchasing habits, the recommendation is to promote high-value products with messaging that emphasizes exclusivity and added value. Targeted marketing campaigns that highlight unique or limited-time offers can increase engagement and encourage higher spending during their less frequent shopping trips.

This refined analysis integrates specific data-driven insights below with actionable strategies for each regional cluster, ensuring campaigns are better aligned with customer behaviors and preferences. Below we can show the customers trends at line chart per cluster.

Table with Cluster per State (S2):



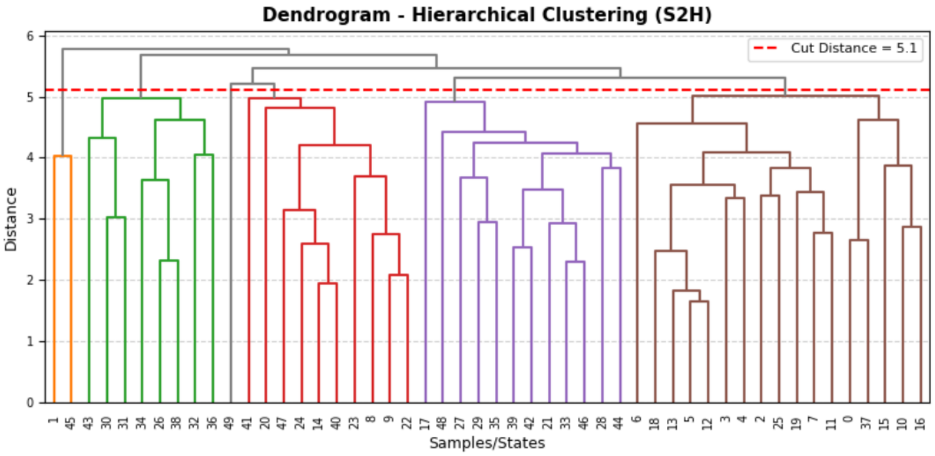
```
Cluster 1: Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee,
Texas, Virginia, West Virginia
Cluster 2: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin
Cluster 3: Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming
Cluster 4: Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont
```



The detailed regional segmentation provided a clearer understanding of purchasing patterns across different areas. It enables the company to design strategies tailored to the specific needs and preferences of customers in each region. By applying these recommendations, the effectiveness of marketing campaigns can be improved, and the brand's presence can be strengthened in each targeted market.

5.2.3. Strategy 2H (S2H): Clustering Hierarchically Refined Regional Patterns

By applying Hierarchical Clustering (S2H) using the Average Linkage method and Euclidean distance, we identified six distinct clusters with a Silhouette Score of 0.15. This segmentation grouped customers based on behavioral traits and purchasing patterns, providing insights for tailored strategies. Bellow at Dendrogram we can observe how this method distributed cluster and it works like a "tree" that connects similar groups and helps determine the number of clusters we should use to better understand customer patterns.



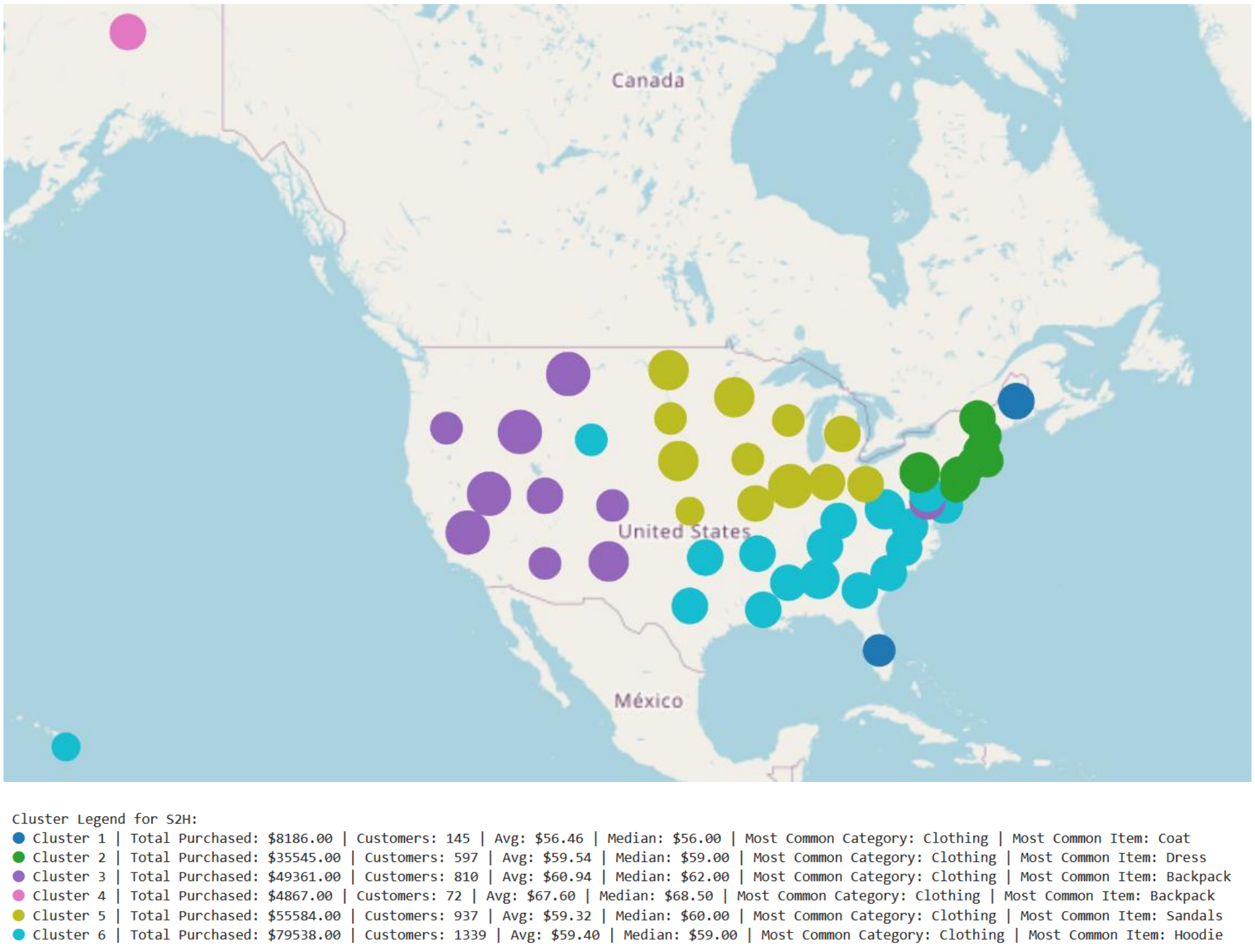
```
Silhouette Score for hierarchical clustering: 0.15
```

The identified clusters were named and paired with specific recommendations to address their unique characteristics:

Cluster 1: Price-conscious Buyers (145 clients): Cluster 1 includes customers who show the lowest total purchases, average purchase value, and frequency. These shoppers are highly price-sensitive, prioritizing low costs and budget-conscious decisions. The Northeastern region stands out, as this group is predominantly from there, with minimal representation from other areas. Additionally, they show a clear preference for express and free shipping options, likely indicating that they seek both affordability and some level of convenience. The recommended strategy for this group is to implement aggressive discount campaigns and regular promotions that emphasize affordability and value. Focusing on products with strong cost-benefit ratios, alongside messaging around savings, can effectively engage this group. Highlighting free shipping offers can also reinforce their interest in cost-effective purchases.

Cluster 2: Northeastern Occasional Buyers (597 clients): Cluster 2 includes customers primarily from the Northeast region who exhibit low transaction frequency and below-average purchasing behavior. This group shows moderate engagement with premium shipping options such as Next Day Air and 2-Day Shipping, while having minimal reliance on in-store pickup. Their below-average spending patterns suggest untapped growth potential. To encourage more frequent purchases, it is recommended to introduce personalized loyalty programs and exclusive rewards that align with their preferences. Enhancing the standard and premium shipping experiences and offering incentives for regular transactions can strengthen their connection with the brand and increase purchase frequency.

Cluster S2 - Hierarchically Refined Regional Patterns



```
Cluster Legend for S2H:
● Cluster 1 | Total Purchased: $8186.00 | Customers: 145 | Avg: $56.46 | Median: $56.00 | Most Common Category: Clothing | Most Common Item: Coat
● Cluster 2 | Total Purchased: $35545.00 | Customers: 597 | Avg: $59.54 | Median: $59.00 | Most Common Category: Clothing | Most Common Item: Dress
● Cluster 3 | Total Purchased: $49361.00 | Customers: 810 | Avg: $60.94 | Median: $62.00 | Most Common Category: Clothing | Most Common Item: Backpack
● Cluster 4 | Total Purchased: $4867.00 | Customers: 72 | Avg: $67.60 | Median: $68.50 | Most Common Category: Clothing | Most Common Item: Backpack
● Cluster 5 | Total Purchased: $55584.00 | Customers: 937 | Avg: $59.32 | Median: $60.00 | Most Common Category: Clothing | Most Common Item: Sandals
● Cluster 6 | Total Purchased: $79538.00 | Customers: 1339 | Avg: $59.40 | Median: $59.00 | Most Common Category: Clothing | Most Common Item: Hoodie
```

Cluster 3: Western Balanced Shoppers (810 clients): Cluster 3 consists of customers from the Western region who display moderate total and high average purchase values, indicating they are willing to spend on quality but with fewer transactions. While this group shows a strong engagement with premium delivery options such as Next Day Air, they are not entirely exclusive to them, also engaging with standard shipping methods to a moderate degree. This behavior suggests that while they prioritize convenience and value premium services, their purchasing habits reflect a balanced approach rather than a purely premium focus. To maximize engagement with this group, the recommendation is to highlight premium products and high-margin offerings while also ensuring flexibility in shipping options. Tailored promotions emphasizing value, such as bundled deals with premium services, can resonate well. Personalized recommendations that focus on high-quality, exclusive products will further encourage higher spending during their relatively infrequent purchases.

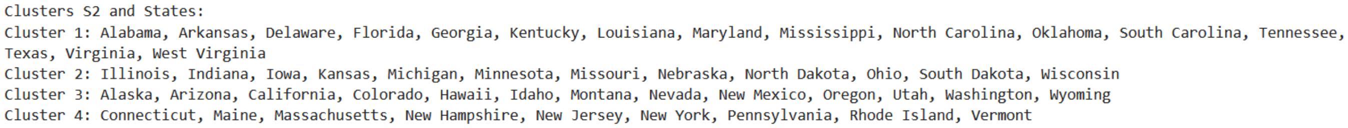
Cluster 4: Western Occasional High-Spenders (72 clients): Cluster 4 comprises customers from the Western region who stand out with high average purchase values but low total transactions. While their average spending per transaction is high, the cluster's overall low total purchase value reflects their infrequent buying behavior. These customers exhibit moderate engagement with Next Day Air and 2-Day Shipping, suggesting that while they value premium delivery options, they are not entirely committed to prioritizing speed over reliability. To better engage this segment, the focus should be on targeted campaigns for exclusive, high-ticket products and personalized shopping experiences. Highlighting premium bundles and time-sensitive offers that align with their preference for quality can encourage repeat transactions. Additionally, offering tailored incentives, such as exclusive deals or early access to premium launches, can capitalize on their willingness to spend and gradually increase transaction frequency.

Cluster 5: Seasonal Buyers (937 clients): Cluster 5 consists of customers who display low average purchase values and transaction frequency, but their purchases are concentrated during specific periods. These shoppers are particularly dominant in the Midwest and are less engaged with premium shipping options. Their behavior indicates that they are most active during seasonal peaks, such as holidays or major sales events. To maximize engagement, the strategy should focus on aligning marketing campaigns with seasonal trends. Creating targeted promotions during peak periods, such as holidays or back-to-school seasons, can encourage them to increase their spending and frequency. Highlighting bundled deals and cost-effective options during these times will also resonate with this group.

Cluster 6: Southern Discount-driven Loyalists (1339 clients): Cluster 6 is characterized by customers with slightly above-average total purchases and moderate to high purchase frequency. They are most prevalent in the Southern region and exhibit notable interest in discounts and promotions. This group shows minimal engagement with premium shipping options such as 2-Day and Express Shipping, indicating that while they value affordability through discounts, they may not prioritize premium shipping. The recommended strategy for this cluster is to focus on targeted discount campaigns and personalized offers. Frequent promotions, loyalty rewards, and volume-based discounts can effectively boost their engagement and transaction values. Highlighting savings opportunities with standard shipping options will encourage repeat purchases and enhance their loyalty to the brand.

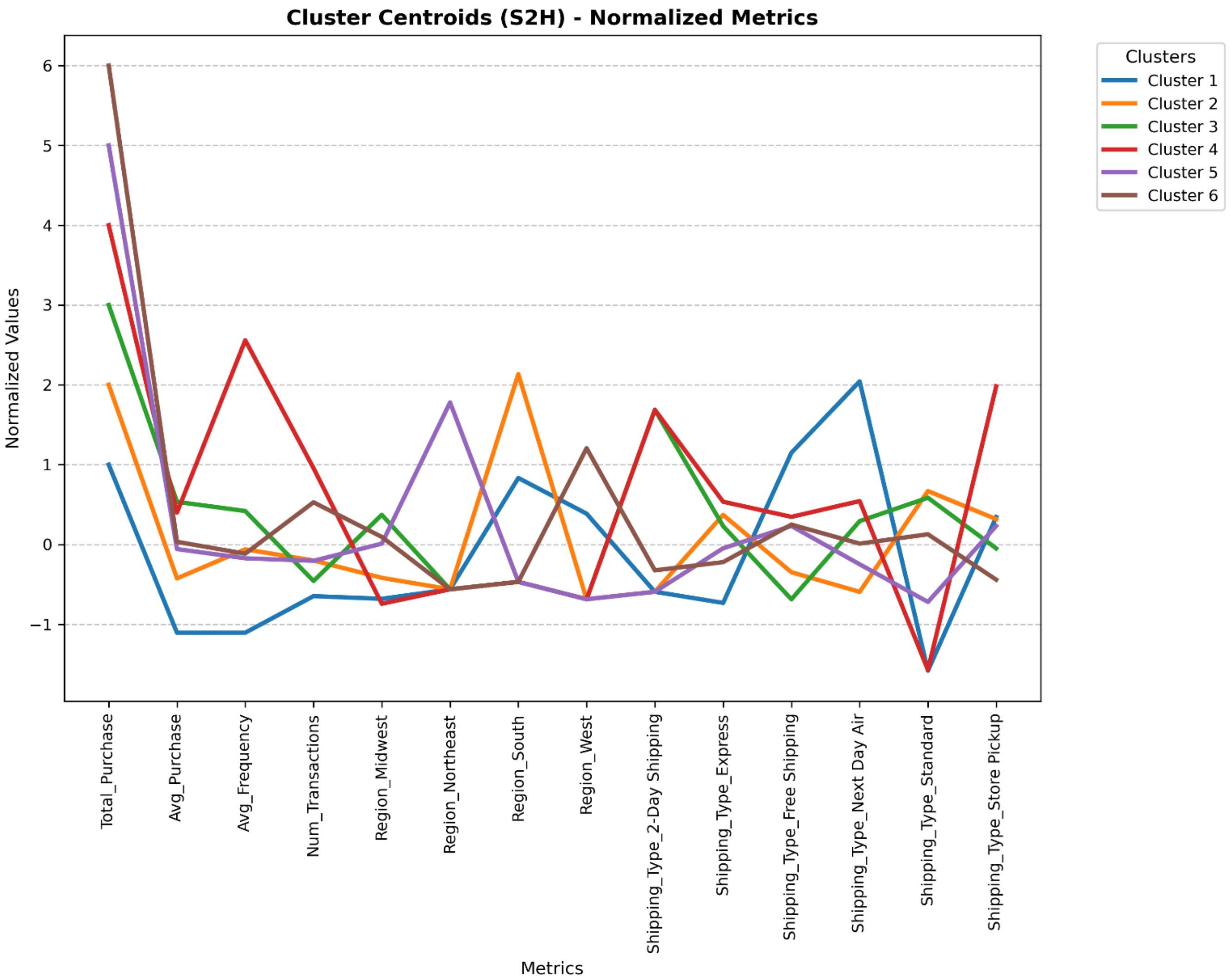
This enriched analysis for S2H clusters integrates key behavioral and regional insights, allowing for more precise, actionable strategies. By understanding the distinct characteristics of each cluster, marketing efforts can be effectively tailored to maximize engagement, satisfaction, and revenue.

Table with Cluster per State (S2H):



```
Clusters S2 and States:
Cluster 1: Alabama, Arkansas, Delaware, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee,
Texas, Virginia, West Virginia
Cluster 2: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin
Cluster 3: Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming
Cluster 4: Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont
```

The hierarchical segmentation provides a detailed understanding of customer purchasing behaviors, enabling the development of specific strategies to maximize engagement and retention for each profile. We can see below trend lines per cluster to analyze the partterns.



By aligning marketing actions with the needs and characteristics of each cluster, companies can enhance the effectiveness of their campaigns and optimize sales outcomes.

5.2.4. Recommendations and Strategic Actions for S1, S2, S2H

Based on the comprehensive analyses derived from the S1 (Behavioral Patterns), S2 (Regional Consumption Patterns), and S2H (Hierarchically Refined Regional Patterns) strategies, our team has formulated targeted strategic actions aimed at enhancing marketing campaigns and maximizing their impact across the identified customer segments.

For the S1 clusters, which encompass diverse behavioral profiles, we recommend implementing personalized promotions tailored to each group's unique purchasing behaviors. Discount-Sensitive Male Shoppers can be effectively engaged through aggressive discount campaigns and regular promotions that emphasize affordability and value, aligning with their high responsiveness to discounts and preference for standard shipping. Transaction-Focused Female Shoppers, who prioritize fast shipping and prefer bank transfers, would benefit from initiatives that enhance speed and convenience, such as promoting expedited shipping options and ensuring a seamless shopping experience with reliable payment methods. For Frequent Discounted Shoppers, emphasizing attractive discounts and free shipping offers is crucial, as their purchasing behavior is heavily influenced by promotions. High-Spending Female Shoppers should be targeted with campaigns that promote premium products and exclusive deals, highlighting quality and reliability to match their slightly above-average transaction values and consistent purchase frequency. Male Practical Shoppers, who favor in-store pickup, would respond well to enhancements in the in-store pickup experience, including improved efficiency and localized promotions that emphasize convenience. Lastly, High-Frequency Female Shoppers would benefit from robust loyalty programs that reward their frequent purchases with personalized rewards and exclusive benefits, fostering greater brand loyalty without relying solely on discounts.

Addressing the regional differences identified in the S2 and S2H strategies, we propose the development of regional campaigns that resonate with the specific behaviors and preferences of each area. Southern Budget Shoppers and Midwest Occasional Buyers demonstrate distinct purchasing patterns that can be leveraged through marketing messages tailored to local cultural and economic contexts, utilizing regional events or trends as focal points. Western Premium High Spenders, characterized by their high spending and preference for premium delivery options like Next Day Air, require campaigns that emphasize high-margin products and exclusive services, ensuring that their desire for quality and convenience is met through personalized offerings and tailored recommendations. Northeastern Occasional Buyers and Northeastern Value Shoppers exhibit below-average purchasing behaviors and infrequent transactions, respectively. For these groups, introducing personalized loyalty programs and enhancing both standard and premium shipping experiences can encourage more frequent purchases and strengthen their connection with the brand.

Additionally, for Western High-Value Frequent Shoppers and Southern Discount-driven Loyalists, the implementation of exclusive high-ticket product offerings and personalized shopping experiences is essential. These clusters benefit from targeted messaging around premium offerings and exclusive deals, which can help build a stronger brand connection and encourage continued high-value purchases. For Seasonal Buyers, aligning marketing campaigns with seasonal trends and peak periods is imperative. Creating targeted promotions during holidays and back-to-school seasons, along with highlighting bundled deals and cost-effective options, will resonate with their concentrated purchasing behavior during specific times of the year.

Moreover, enhancing convenience through improved in-store pickup services for clusters that prefer this method, such as Male Practical Shoppers and Seasonal Buyers, will increase engagement and satisfaction. Providing better conditions for pickup, such as increased efficiency and exclusive promotions, can make this delivery method more appealing. For clusters like Transaction-Focused Female Shoppers, who prefer conventional contact methods, utilizing traditional communication channels such as email newsletters and direct mail will maintain effective engagement and strengthen their connection to the brand.

In summary, these strategic actions are meticulously designed to align with the specific characteristics and preferences of each customer segment identified through the S1, S2, and S2H strategies. By tailoring marketing efforts to address the unique needs of each cluster, we can enhance customer engagement, foster loyalty, and optimize overall campaign performance, ultimately driving sustained business growth and improved customer satisfaction.

6. Answering Business Problem 3: Identify customers' affinity for certain product categories

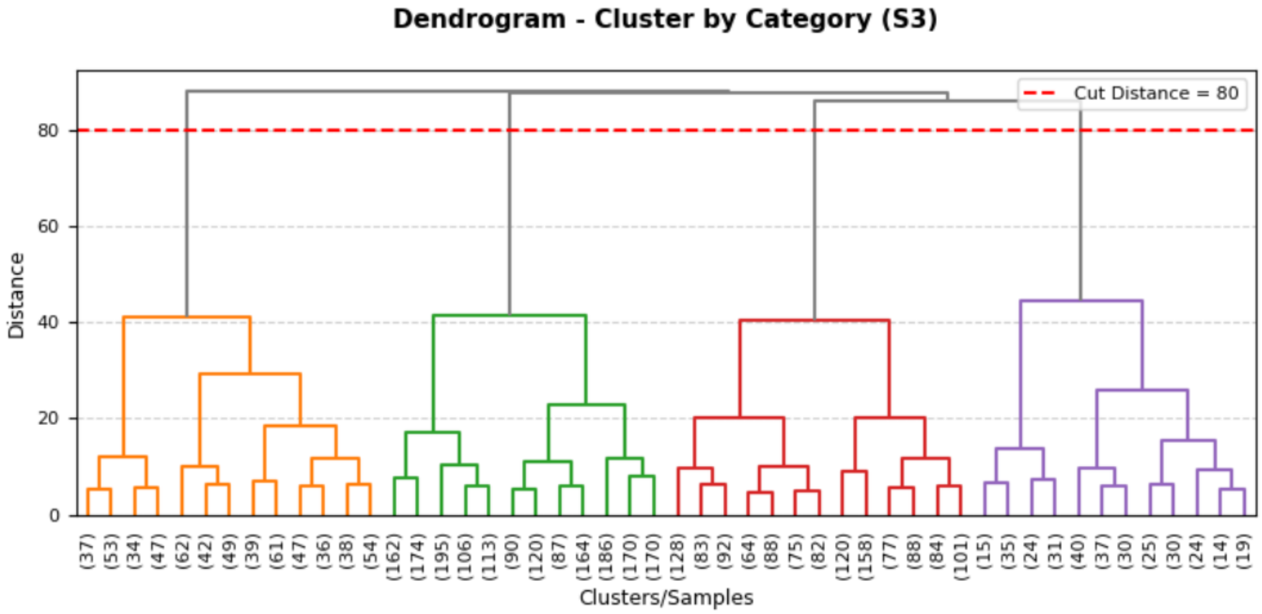
6.1. Methodology

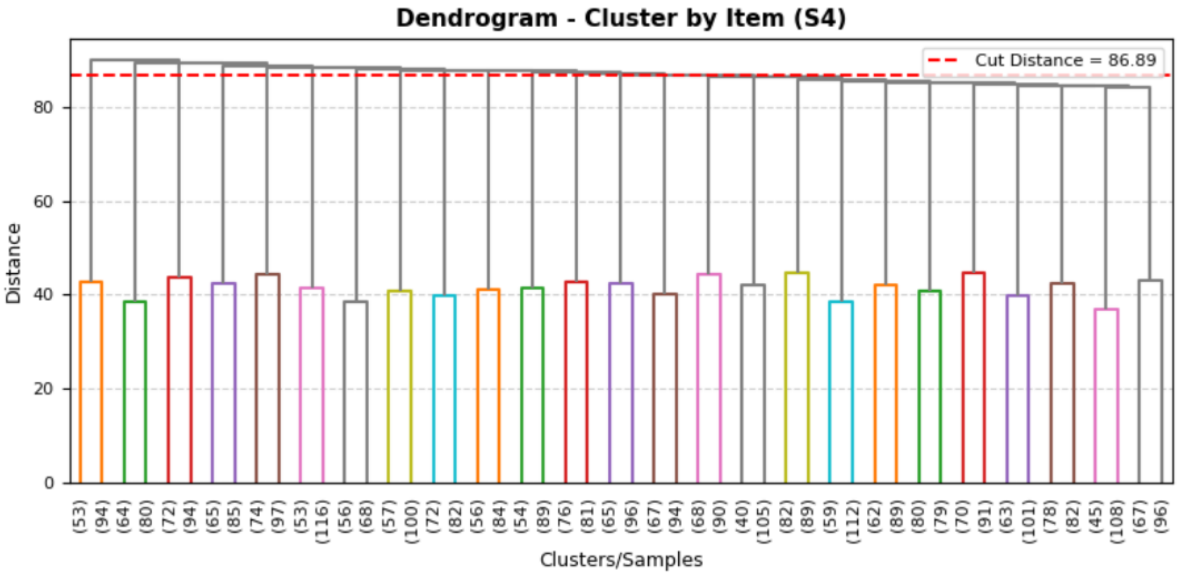
To identify customers' affinity for certain product categories and specific items, we applied two hierarchical clustering approaches: S3 (Clustering by Category) and S4 (Clustering by Item).

In the S3 strategy, we focused on clustering by product categories. We transformed the 'Category' column into dummy variables and aggregated the data by customer based on the purchase amount in each category. We utilized the average linkage method with Euclidean distance to perform the clustering, setting a distance cutoff of 80 to define the resulting four clusters. This approach allowed us to identify groups of customers with clear affinities for specific categories such as Footwear, Clothing, Accessories, and Outerwear.

In the S4 strategy, we directed the clustering towards individual items. By converting the 'Item_Purchased' column into dummy variables and aggregating the data by customer based on the purchase amount for each item, we again applied the average linkage method with Euclidean distance, adjusting the distance cutoff to 86.89. This segmentation resulted in the formation of fourteen clusters, each with notable affinities for specific items such as T-shirts, Boots, Dresses, Shoes, among others.

Below, we can observe the Dendrogram that shows clusters formed by hierarchical grouping in S3 and S4.

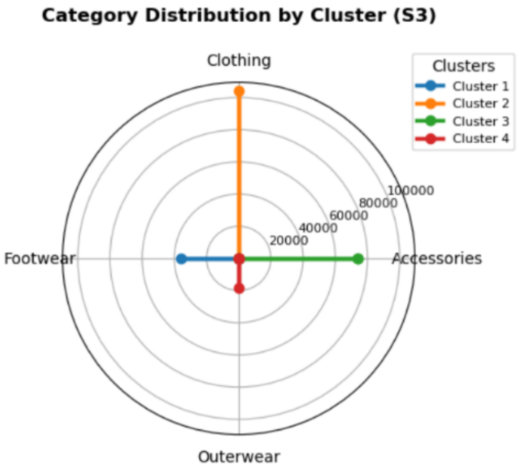




6.2. Results, Conclusions, and Recommendations

6.2.1. S3 – Clustering by Category

The results of the S3 strategy revealed four distinct clusters with clear affinities for specific product categories.



```
Quantity of clientes per cluster (Cluster S3):

Cluster 1: 599
Cluster 2: 1737
Cluster 3: 1240
Cluster 4: 324
```

Cluster 1: Footwear Enthusiasts (599 clientes) : These customers demonstrated a strong preference for Footwear, evidenced by the high purchase value in this category. This indicates that customers in this group are passionate about footwear and likely seek variety and quality within this segment. Our recommendation is to launch campaigns that highlight new releases and exclusive offers in the footwear category, promoting product diversity and quality to effectively engage this group.

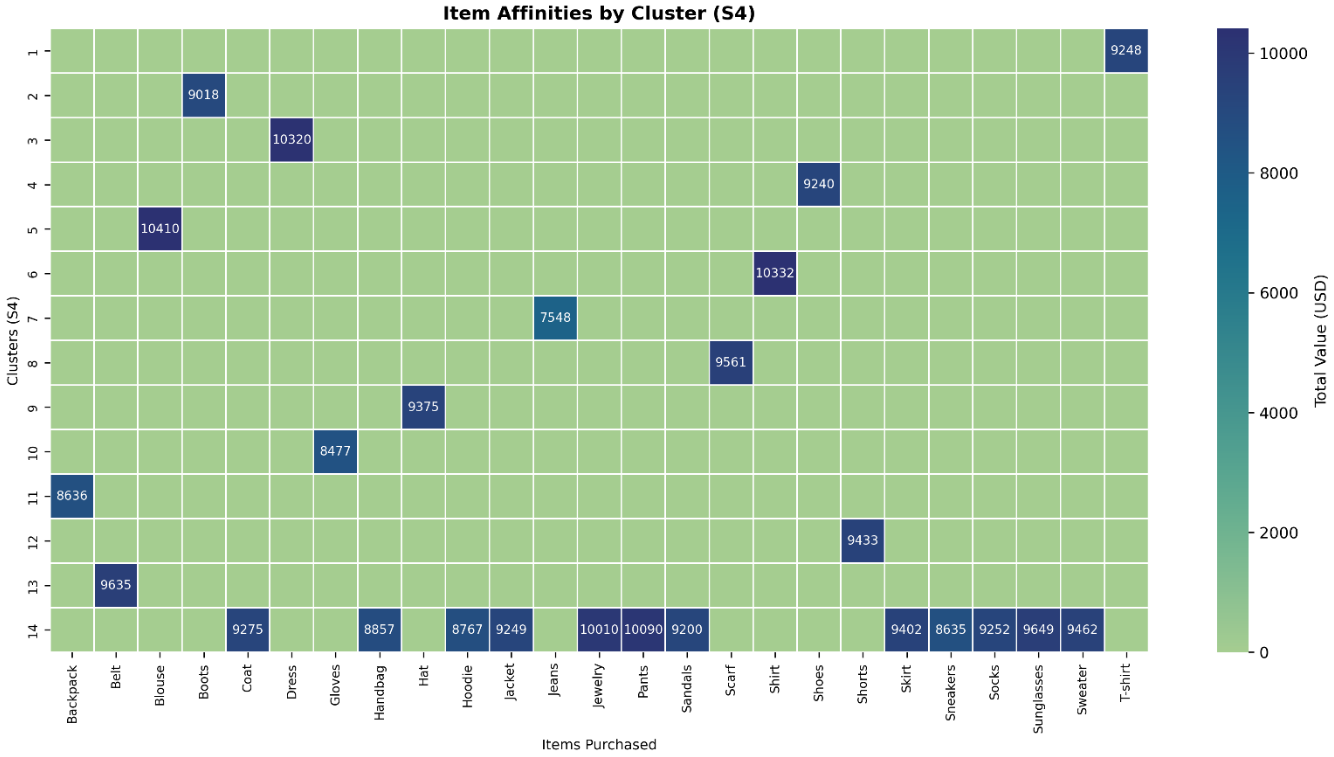
Cluster 2: Clothing Aficionados (1737 clients): This cluster stood out for their exclusive affinity for Clothing, suggesting that these customers prioritize apparel and keep up with the latest fashion trends. Campaigns that emphasize the latest fashion trends and the variety of available options can significantly increase engagement and sales for this segment, meeting their expectations for newness and updated styles.

Cluster 3: Accessories Lovers (1240 clients): Cluster 3 showed a notable preference for Accessories, indicating that these customers value complementary items such as bags, belts, and jewelry, which add value and style to their outfits. For these customers, it is recommended to promote collections of bags, belts, and jewelry, highlighting how these accessories complement the apparel purchased, thereby strengthening the perception of value and style.

Cluster 4: Outerwear Focused (324 clients): This cluster revealed a substantial preference for Outerwear, signaling that these customers specifically seek pieces that offer comfort and protection against the cold, such as coats and jackets. Seasonal campaigns that emphasize the quality and functionality of these pieces can attract these customers during peak demand periods, ensuring that their needs for protection and comfort are effectively met.

6.2.2. S4 – Clustering by Item

In the S4 analysis, the fourteen identified clusters exhibited differentiated affinities for specific items.



Cluster 1: T-shirt Enthusiasts (147 clients): Showed a high preference for T-shirts, indicating that these customers value basic and versatile items in their wardrobe. The recommendation is to offer promotions for new basic collections and discounts on multiple T-shirt purchases, encouraging loyalty through the provision of essential products.

Cluster 2: Boots Aficionados (144 clients): Stood out for their affinity for Boots, suggesting that these customers seek robust and highly durable footwear. Campaigns that highlight the durability and style of these boots, possibly accompanied by extended warranties or customization services, are recommended to meet the expectations of this group.

Cluster 3: Dresses Lovers (166 clients): Displayed a strong preference for Dresses, revealing a group of customers who prioritize elegant and formal pieces. It is essential to promote elegant and exclusive dresses, as well as offer personalized shopping experiences such as style consultations or launch events for new collections, to attract and retain these engaged customers.

Cluster 4: Shoes Focused (150 clients): Exhibited a significant preference for Shoes, reflecting a search for varied and quality footwear. It is recommended to promote a wide variety of high-quality shoes and offer personalized recommendations based on purchase history, thereby increasing customer satisfaction and loyalty.

Cluster 5: Blouse Aficionados (171 clients): Demonstrated a strong affinity for Blouses, indicating that these customers value specific pieces that combine comfort and style. Campaigns that emphasize the versatility and elegance of Blouses are recommended to meet the needs of this segment.

Cluster 6: Socks Enthusiasts (169 clients): Showed a high preference for Socks, suggesting that these customers seek comfort and functionality. Promoting high-quality sock sets and offering discounts on multiple unit purchases can increase engagement and sales within this group.

Cluster 7: Jeans Lovers (124 clients): Exhibited a strong affinity for Jeans, indicating that these customers value durable and versatile pieces in their wardrobe. Campaigns that highlight the durability and style of Jeans, along with exclusive offers, are recommended to attract this segment.

Cluster 8: Handbag Enthusiasts (157 clients): Presented a significant preference for Handbags, reflecting a search for bags that add value and style to customers' looks. It is recommended to promote exclusive handbag collections and offer discounts on multiple unit purchases to increase sales and customer satisfaction.

Cluster 9: Hat Enthusiasts (154 clients): showed a high preference for Hats, indicating that these customers value accessories that complement their outfits. Campaigns that highlight the variety and style of Hats are recommended to cater to this group.

Cluster 10: Handbag Lovers (140 clients): Demonstrated a strong affinity for Handbags, suggesting that these customers seek bags that add value and style to their looks. Promoting exclusive handbag collections and offering discounts on multiple unit purchases can increase sales and customer satisfaction for this group.

Cluster 11: Backpack Enthusiasts (143 clients): Exhibited a high preference for Backpacks, indicating that these customers value practicality and functionality. Campaigns that highlight the durability and versatility of Backpacks are recommended to meet the needs of this group.

Cluster 12: Skirt Lovers (157 clients): Showed a strong affinity for Skirts, reflecting a search for feminine and versatile pieces. Promoting different styles of Skirts and offering personalized recommendations can attract this group and increase engagement.

Cluster 13: Belt Aficionados (161 clients): Demonstrated a high preference for Belts, indicating that these customers value accessories that complement their outfits. Campaigns that highlight the variety and quality of Belts are recommended to cater to this segment.

Cluster 14: Multi-Item (1917 clients): Affinity stood out for affinities towards multiple items with moderate purchase values, indicating a diversity of preferences within this group. It is recommended to create campaigns that promote the complementarity of products, encouraging the purchase of multiple items that complement each other, thereby increasing the average transaction value.

6.2.3. Final Recommendation for S3 and S4

Based on the analyses conducted in strategies S3 and S4, it is evident that detailed segmentation by categories and items provides meaningful insights for personalizing marketing campaigns. It is recommended that strategic actions be continuously adjusted as new data is collected, ensuring that customer preferences are always effectively met. Investing in data analysis tools and dedicated segmentation teams will enable the company to respond agilely to changes in consumption patterns, promoting sustainable growth and strengthening customer loyalty. By aligning campaigns with the specific preferences of each cluster, we can increase the effectiveness of our initiatives, enhance customer satisfaction, and consequently drive solid business growth.

7. Answering Business Problem 4: Customers Based on Their Lifetime Value (CLV)

7.1. Methodology

To address the business question of grouping customers based on their lifetime value (CLV), we employed clustering techniques that focused on two critical metrics: total purchase amount and purchase frequency. The CLV for each customer was calculated using the formula:

$$CLV = \text{Purchase Amount (USD)} \times \text{Frequency of Purchases per Year}$$

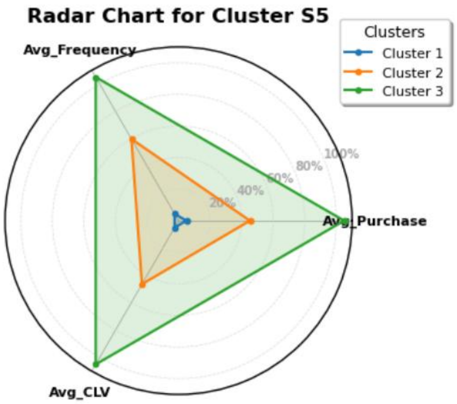
The calculated CLV values were normalized using the StandardScaler to ensure consistency and comparability during the clustering process. Leveraging the K-Means algorithm, we segmented customers into three distinct clusters. This number was chosen as the optimal solution based on the distribution of CLV values and its alignment with business priorities.

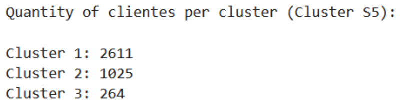
To enhance the interpretability of the clusters, they were ordered according to their mean CLV, making it easier to derive actionable insights. Visualization techniques, including radar charts, box plots, and bar plots, were utilized to provide a comprehensive view of each cluster's characteristics. These visualizations enabled us to identify distinct behavioral patterns and purchasing habits among the clusters.

Finally, we calculated the centroids for each cluster, summarizing key metrics such as average purchase amount, frequency, and lifetime value, which serve as a foundation for targeted business strategies.

7.2. Results, Conclusions, and Recommendations

The analysis revealed three distinct customer clusters based on their lifetime value, each with unique characteristics and strategic implications:





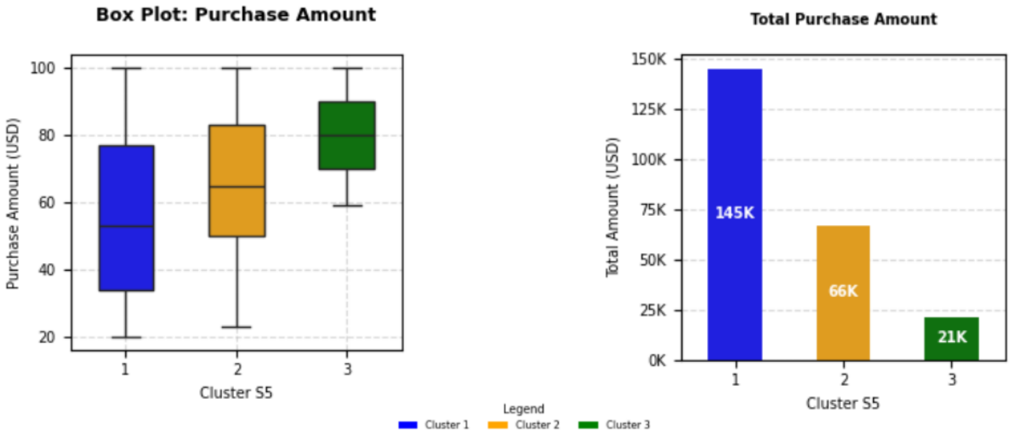
```
Quantity of clientes per cluster (Cluster S5):

Cluster 1: 2611
Cluster 2: 1025
Cluster 3: 264
```

Cluster 1: Low-Value Shoppers (2611 clients): This group is characterized by the lowest individual CLV ($373.75) and a lower average purchase value ($55.59); however, it accounts for the highest total purchase volume (145k) among all clusters. With 2,611 customers, this cluster represents a large customer base where each individual contributes through small, frequent transactions, resulting in significant cumulative volume. These customers are generally price-sensitive and focus on value-oriented purchases. Strategies to engage this segment include campaigns that highlight affordable products, bundle offers, and discounts on higher-value purchases. Additionally, implementing simple loyalty programs with rewards for recurring purchases can help increase the average transaction value and improve customer retention.

Cluster 2: Medium-Value Shoppers (1025 clients): With an average CLV of $1,934.01, this cluster represents customers with a moderate level of engagement. They purchase at a relatively high frequency (32.2 times per year) and have a higher average transaction value ($65.28) compared to Cluster 1. Despite a lower total purchase volume (66k), this group, which consists of 1,025 customers, demonstrates a positive affinity with the brand while still showing growth potential. Recommended strategies include enhanced loyalty programs with rewards for frequent purchases and incentives to increase transaction value, such as bundle promotions, personalized offers, and subscription-based models. Implementing cross-selling and upselling campaigns can also strengthen engagement and boost the CLV of this group.

Cluster 3: High-Value Shoppers (262 clients): Customers in this cluster are the most valuable on an individual level, with the highest average CLV ($4,142.47) among all clusters. They purchase with high frequency (52 times per year) and exhibit the highest average transaction value ($79.66). However, the total purchase volume (21k) is the lowest, reflecting that this segment consists of only 264 customers. These customers demonstrate strong loyalty and significant purchasing power, making them essential for revenue generation. To maximize the potential of this group, we recommend personalized campaigns that promote high-value products, exclusive VIP offers, and unique experiences, such as early access to product launches or premium services. Subscription models for premium products and tailored reward programs are also effective for reinforcing loyalty and expanding their long-term value.



Above the graph show clear this clustering S5 about value of each group have per purchase and value total. Each cluster's unique characteristics offer a foundation for strategic initiatives aimed at optimizing customer engagement and revenue generation. By tailoring campaigns to address the specific behaviors and preferences of each segment, the organization can enhance customer satisfaction and drive sustainable growth.

8. REPORT CONCLUSION

This report addressed four key business questions, using classification and clustering techniques adapted to customer behaviors and preferences. Below, we summarize the findings and recommendations for each question.

To understand the factors that drive customer subscriptions, we analyzed behavioral, demographic, and geographic aspects. The results revealed that regional differences, gender, promotional effectiveness, preferred payment methods, delivery preferences, and seasonal trends significantly impact subscription rates. Customers in certain regions showed lower engagement, highlighting the need for targeted marketing efforts. Detailed recommendations were presented to address these factors, and adapting strategies can help improve subscription rates and overall retention.

In segmenting customers into distinct groups, we used clustering techniques to identify behavioral and regional patterns. Through multiple approaches, including behavioral segmentation and regional analysis, we identified unique customer groups with specific habits, preferences, and priorities. These insights provide a foundation for personalized marketing strategies, depending on leadership's decisions, timing, and focus. Each strategy was discussed, recommendations were made, and the numerical clustering was saved in our customer database for future application.

When identifying customer affinity for product categories, our analysis highlighted clear preferences among different consumer groups. Segmenting by product categories and individual items revealed unique clusters and their specific preferences. These findings guide marketing strategies focused on promoting high-demand categories, offering personalized product recommendations, and aligning campaigns with customer preferences to enhance satisfaction and loyalty.

Finally, in classifying customers based on lifetime value (CLV), we identified three distinct segments: low, medium, and high-value customers. Each group demonstrated unique characteristics, such as transaction frequency and average purchase value, which guide targeted strategies. Recommendations include focusing on loyalty programs for medium-value customers and offering premium, personalized experiences for high-value customers to maximize engagement and revenue potential across all segments.

Overall, the results of this project provide valuable information and actionable recommendations to guide strategic decision-making. By integrating behavioral analysis, regional segmentation, product affinity studies, and customer lifetime value insights, we developed a framework to enhance customer engagement and optimize marketing campaigns. These multiple strategies enable the organization to allocate resources effectively, strengthen customer relationships, and achieve sustainable growth by carefully selecting and timing the application of each strategy.

9. Glossary of Technical Terms

This explanation was simplified for Non-Technical Readers:

1. **AUC-ROC (Area Under the Curve - Receiver Operating Characteristic)**: A performance measurement for models that predict categories, like subscribers vs. non-subscribers. A score close to 1.0 means the model is very accurate in distinguishing between the two categories.
2. **Behavioral Patterns**: Recurring habits or actions of customers, such as how often they shop, how much they spend, and what shipping options they prefer. Identifying these patterns helps businesses create personalized experiences.
3. **Binning**: A method of grouping continuous numbers (like ages or spending amounts) into ranges or categories. For example, customers aged 18–25 can be grouped as "Young Adults," and spending between $20–$50 can be grouped as "Low Spenders."
4. **Centroid**: The “center” or average position of all customers within a cluster. For instance, if Cluster 1 customers usually spend $50 and purchase 10 times a year, these values represent the centroid of the group.
5. **Clustering**: A technique used to group customers into smaller segments based on similar behaviors or characteristics. For example, customers who purchase frequently but spend small amounts can form one group, while those who make fewer but high-value purchases can form another. This helps businesses target each group with tailored strategies.
6. **Correlation**: A statistical relationship between two variables. A positive correlation means both variables increase together (e.g., higher spending correlates with higher purchase frequency).

7. **Correlation Matrix**: A table that shows the relationships between numerical variables. For example, it might reveal that higher spending is often related to frequent purchases. This helps businesses understand connections between behaviors.
8. **Customer Lifetime Value (CLV)**: The total amount of money a customer is expected to spend on a company's products or services during their relationship. For example, a loyal customer who buys regularly has a higher CLV than someone who only shops occasionally.
9. **Dendrogram**: A diagram that looks like a tree, used to show how clusters are formed in hierarchical clustering. It visually represents which customers are similar and how they are grouped step by step.
10. **Elbow Method**: A visual technique for choosing the optimal number of clusters. It involves plotting the number of clusters against how well the data fits. The "elbow point" on the graph shows the best number of clusters to use.
11. **Geospatial Analysis**: The use of geographic data (like latitude and longitude) to study customer behavior based on their location. For example, identifying that customers in the Northeast region shop less frequently but spend more per purchase.
12. **Hierarchical Clustering**: A clustering method that groups customers step by step, starting with smaller clusters and combining them into larger ones. The results are visualized as a "tree" (called a dendrogram) that shows how clusters are connected.
13. **K-Means Clustering**: A method for dividing customers into groups (or clusters) based on their similarities. The algorithm organizes customers into a chosen number of clusters, such as 3 or 4, depending on the business needs.
14. **Logistic Regression**: A model used to predict outcomes that have two possible results. For example, whether a customer subscribes to a service (Yes/No) based on their shopping behavior and preferences.
15. **Machine Learning**: A form of artificial intelligence that allows computers to analyze data and learn patterns without being explicitly programmed. In this report, it was used to group customers and predict their behaviors.
16. **Multicollinearity**: A situation in which two or more variables (e.g., purchase frequency and total spending) are very similar or strongly related. This can confuse models and make it difficult to identify which variable matters most.
17. **One-Hot Encoding**: A process of converting text-based categories (e.g., "Male" or "Female") into numerical data (0s and 1s) so they can be used in mathematical models. This ensures that computers can analyze all types of data accurately.
18. **Outliers**: Unusual data points that are very different from the rest. For example, a customer who spends 100 times more than anyone else may be considered an outlier. These points are studied separately to avoid misleading conclusions.
19. **Regional Patterns**: Trends and habits observed among customers based on their location. For instance, customers in the South may prefer free shipping, while those in the West prioritize premium shipping options.

20. **Segmentation**: Dividing customers into smaller, meaningful groups based on shared traits (e.g., spending habits, preferences). This allows businesses to create personalized strategies to better meet the needs of each group.

21. **Silhouette Score**: A number that measures how well customers are grouped within their clusters. A higher score means that the customers in a group are more similar to each other and different from customers in other groups.

22. **StandardScaler**: A method to adjust (or "scale") numerical data so it is easier to compare. For example, when purchase amounts range from $10 to $10,000, scaling ensures that all numbers are treated fairly in the analysis.

CONCLUSION

We concluded the final project by applying classification and clustering techniques covered during the course to analyze data, aiming to solve four strategic problems related to customer behavior in the U.S. retail sector. Using the *"shopping_trends.csv"* dataset obtained from the Kaggle platform, we analyzed behavioral patterns and consumption trends, generating relevant information to support strategic decision-making by management. The study enabled the creation of targeted marketing campaigns and the optimization of strategies to increase customer value.

In the first business problem, related to service subscription, we investigated the variables influencing customers' decisions to subscribe to the company's services. Through methods such as Lasso CV, Backward Elimination, and Logistic Regression, we identified the key factors contributing to subscriptions, including demographic and behavioral characteristics such as gender, age, preferred payment methods, and purchase frequency.

To segment customers and create more effective marketing campaigns, we applied clustering techniques such as K-Means and Hierarchical Clustering, defining distinct groups (S1, S2, and S2H) based on geographic, demographic, and behavioral attributes. The clusters were refined using the Elbow Method and Silhouette Score, ensuring clear and actionable segmentations. These groups provided the foundation for more personalized and efficient marketing strategies.

The identification of customer affinities for product categories and specific items was conducted through Hierarchical Clustering, resulting in strategies S3 (categories) and S4 (items). The dendrogram analysis revealed clear patterns of consumption preferences, enabling targeted recommendations and more efficient inventory management.

In the fourth problem, the segmentation of customers based on Customer Lifetime Value (CLV) was performed using K-Means, grouping customers into clusters (S5) of low, medium, and high value. This analysis guided retention and loyalty strategies, prioritizing efforts based on each group's revenue potential.

The data preparation process included appropriate cleaning, the creation of new variables such as *Frequency of Purchases per Year* and *Subscription Status*, and the inclusion of geographic coordinates. Categorical variables were converted into dummies, while numerical variables were standardized using StandardScaler to ensure consistency and better model performance.

We consolidated the information into the final report, which presented the methodology, results, and strategic recommendations in detail. Using visualizations such as graphs, maps, and dendrograms, we highlighted key findings and suggested approaches that can be implemented individually or in combination to improve the company's overall performance.

Finally, by adopting a data-driven approach, this project provided a solid foundation for informed and evidence-based decision-making, addressing challenges and leveraging opportunities in the U.S. retail market.