Customer Purchase Behavior Analysis

Abstract- Understanding consumer purchasing behavior has become crucial for businesses to create successful marketing strategies and raise customer satisfaction in the cutthroat business world of today. Customer Purchase Behavior Analysis is the main emphasis of this research, which uses data analytics methods to pinpoint important variables that affect consumer spending trends. The study makes use of a publicly accessible marketing dataset that includes demographic data, household composition, socioeconomic characteristics, and specific purchasing behavior for 2,240 clients from various nations.

To deal with missing values, eliminate outliers, and guarantee data consistency, preprocessing and data cleaning procedures were first carried out. After that, connections between variables like income, education, age, family structure, and total spending were examined using exploration data analysis (EDA). According to the analysis, customers' purchasing behavior is strongly influenced by their wealth and level of education, with higher-income and better-educated consumers spending more, particularly on high-end goods.

Additionally, web visits and online sales were examined to understand client involvement patterns. The findings demonstrated that customers' browsing and comparison behavior does not always translate into more sales. K-Means clustering was used to divide clients into discrete groups according to their purchasing and spending habits in order to improve their study even more. Three significant client segments—low-value, medium-value, and high-value customers—were produced by applying the elbow technique to calculate the ideal number of clusters.

Businesses can better understand client wants, develop marketing efforts, and adopt customized strategies for various customer categories with the help of the information gleaned from this analysis. All things considered, this project shows how data analytics can be utilized successfully to assist data-driven decision-making and convert unprocessed customer data into useful business insights.

Keywords: Python, Customer Segmentation, Descriptive Statistics, Data Analytics, Customer Behavior Analysis, K- means clustering, Data-Driven Decision making

I. INTRODUCTION

In today's data-driven and highly competitive business environment, understanding customer purchasing behavior has become essential for organizations aiming to design effective marketing strategies and improve decision-making. With the rapid growth of digital platforms and the availability of large volumes of customer data, businesses are increasingly relying on data analytics to gain insights into customer preferences, spending patterns, and engagement behavior. Marketing analytics enables firms to transform raw data into actionable insights that support customer segmentation, personalized marketing, and improved customer retention.

Customer purchasing behavior is influenced by a wide range of factors, including demographic characteristics, socioeconomic conditions, household composition, and preferred purchasing channels. Variables such as income, education level, age, marital status, and family structure play a crucial role in shaping consumption decisions across different product categories. Analyzing these factors allows businesses to better understand which customer groups contribute the most value and how purchasing behavior varies across regions and sales channels.

The primary objective of this study is to analyze customer purchasing behavior using a marketing dataset that includes demographic, socioeconomic, and behavioral attributes. The analysis focuses on identifying key factors that influence customer spending behavior across multiple product categories such as wine, meat, fish, sweets, and gold products, as well as across different purchasing channels including web, store, catalog, and promotional deals. Special attention is given to understanding the impact of income, education level, age, household composition, and country of residence on total customer spending.

To achieve these objectives, the study follows a systematic data analytics approach involving data preprocessing, descriptive statistical analysis, and exploratory data analysis (EDA). Python is used as the primary analytical tool due to its efficiency in handling large datasets and its wide range of libraries for data cleaning, visualization, and analysis. In addition, customer segmentation is performed using K-means clustering to group customers with similar purchasing behaviors and spending patterns.

The findings of this research provide valuable insights for businesses seeking to optimize marketing strategies, improve customer targeting, and enhance customer engagement. By understanding how different customer segments behave, organizations can develop more effective marketing campaigns, allocate resources efficiently, and strengthen long-term customer relationships.

II. DATA DESCRIPTION & METHODOLOGY

A. Data Source

The dataset is being taken from Kaggle. The dataset has the customers' information about a marketing store operating across multiple countries. Spain (SP), Canada (CA), United States (US), Australia (AUS), Germany (GER), India (IND), Saudi Arabia (SA), and Montenegro (ME)).

B. Data Overview

This dataset contains 2,240 customers, that includes 28 variables. The household information (teenagers, number of children), demographic characteristics (age, country, educational and marital status), income and spending behavior across multiple product categories. Data set also includes customers engagement metrics such as website visits, and the purchase frequency across promotional, catalog, online and retail channels.

C. VariableDescription

The key variables used in this analysis are defined as follows.

Variables	Description	Type / Unit
Income	Annual household income of the customer	Numeric (Currency)
Education	Highest education level attained by the customer	Categorical
Country	Country of residence of the customer	Categorical
Age	Age of the customer	Numeric (Years)

	(derived from year of birth)	
Kidhome	Number of young children living in the household	Numeric (Count)
Teenhome	Number of teenagers living in the household	Numeric (Count)
Recency	Number of days since the customer's last purchase	Numeric (Days)
mnt_wines	Amount spent on wine products	Numeric (Currency)
mnt_fruits	Amount spent on fruit products	Numeric (Currency)
mnt_meat_products	Amount spent on meat products	Numeric (Currency)
mnt_fish_products	Amount spent on fish products	Numeric (Currency)
mnt_sweet_products	Amount spent on sweet products	Numeric (Currency)
mnt_gold_products	Amount spent on gold products	Numeric (Currency)
num_web_visits_month	Number of website visits per month	Numeric (Count)
num_web_purchases	Number of purchases made through the website	Numeric (Count)
num_store_purchases	Number of purchases made in physical stores	Numeric (Count)
num_catalog_purchases	Number of purchases made through catalogs	Numeric (Count)
num_deals_purchases	Number of purchases made using promotional deals	Numeric (Count)
total_spending	Total amount spent across all product categories (derived)	Numeric (Currency)

III. DATA PREPROCESSING

Several preprocessing steps were performed to ensure data quality. Income values originally stored as string containing currency symbols were converted into numerical formats. Null values were found only in the income columns which were replaced by median due to the right-skewed income distribution. Duplicate records were not found in the data. Date variables were converted into an appropriate date of time format.

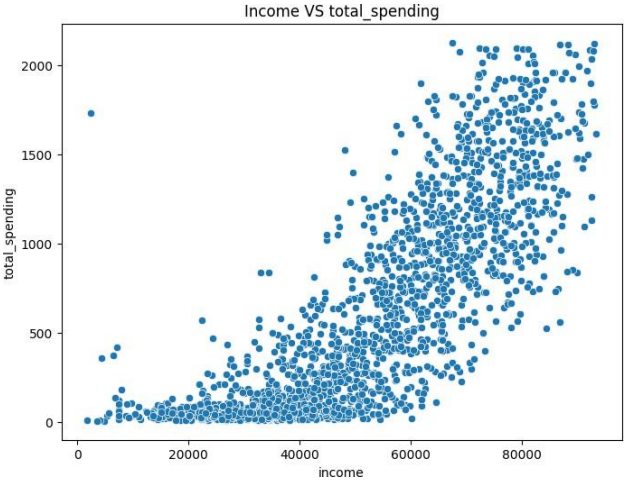
Outlier Treatment & Feature Engineering

A new variable (total_spending) was developed to indicate the total value (monetary) of all purchases made by customers and was created by adding together all the amounts spent in each product category. The outlier values found in the income and total_spending columns were removed with a percentile-based method so as not to bias the analysis results. The outliers were identified using a boxplot. Any observations greater than the 99th percentile of the income or total_spending were removed from the dataset. All column names were converted to SNAKE_CASE to maintain consistency among the names and improve the readability of the dataset for easier manipulation while during the analysis.

IV. EXPLORATORY DATA ANALYSIS (EDA)

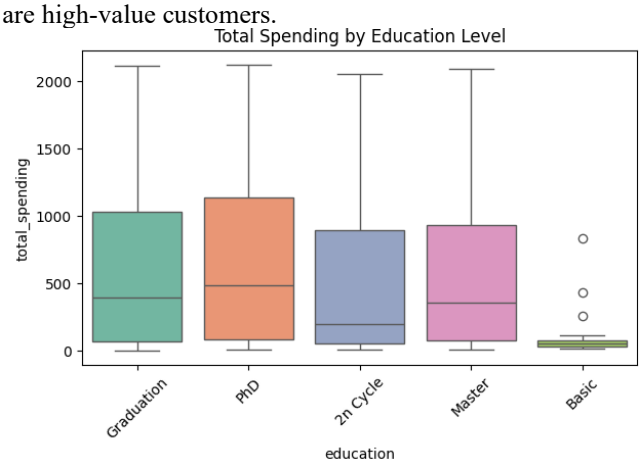
A. Income vs Purchasing Behavior

A strong positive relation is observed between income and total spending across all product categories. The spread of points widens at higher income levels, suggesting heteroscedasticity (more variation in spending among higher- income customers) which indicate different lifestyles or access to premium product categories. Low-income customers cluster at low spending, indicating similar purchasing behavior and reflecting budget constraints and more necessity-driven consumption.



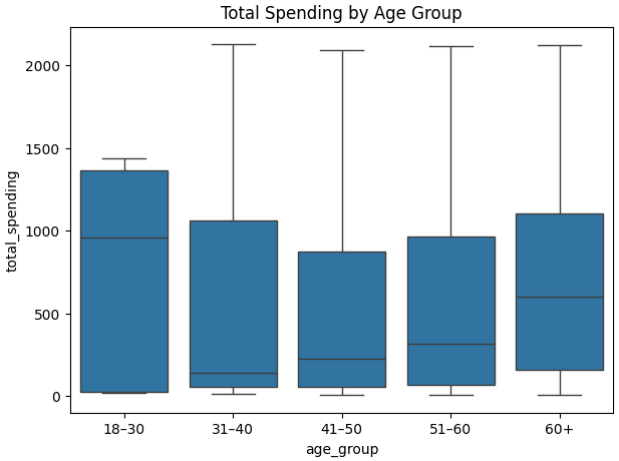
B. Education vs Purchasing Behavior

Median total spending is high for the customers with PhD, master's & graduation and moderate for the people with 2n cycle & lower for the basic group customers. This indicates that people with basic education spend less comparatively with other groups of people. Higher-educated groups have high spending, which shows they are high-value customers.



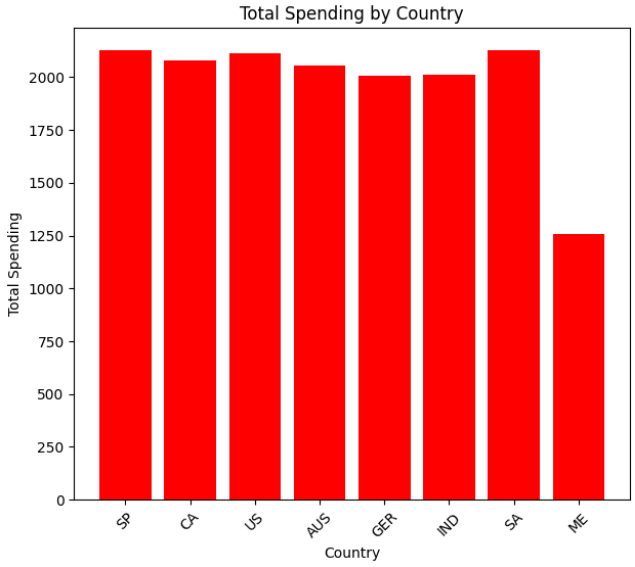
C. Age vs Purchasing Behavior

Total purchasing behavior varies across all age groups. The youngest group (18-30) and oldest (60+) have the highest median compared to middle age group people (31-50). The possible reason that middle-aged people spend less could be high family responsibilities, such as childcare and household expenses. Due to which they might be spending only on the essential categories.



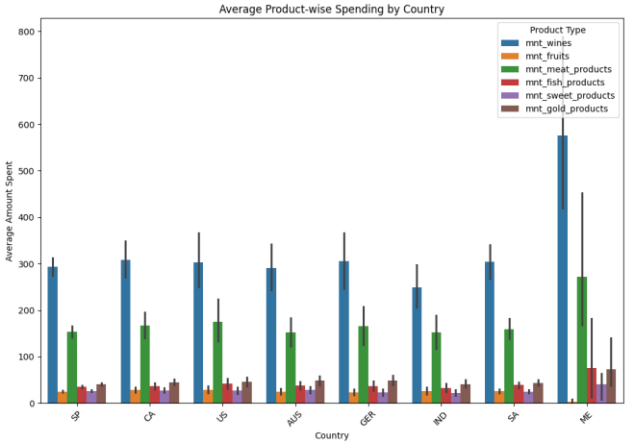
D. Countries vs Purchasing Behavior

Countries' spending is similar, with the US slightly ahead and Montenegro at the bottom.



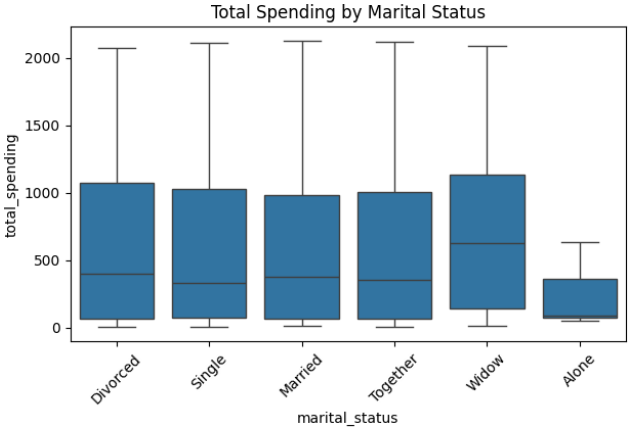
E. Average Product-wise Spending Across Countries

For each country on the x-axis, there are multiple bars that represent the mean spent on the product type (wines, fruits, meat, fish, sweets, gold products). Most countries have dominated spending on wine and meat products. ME (Montenegro) noticeably stands out with high demand for wine, meat and gold products, indicating high demand for these categories.



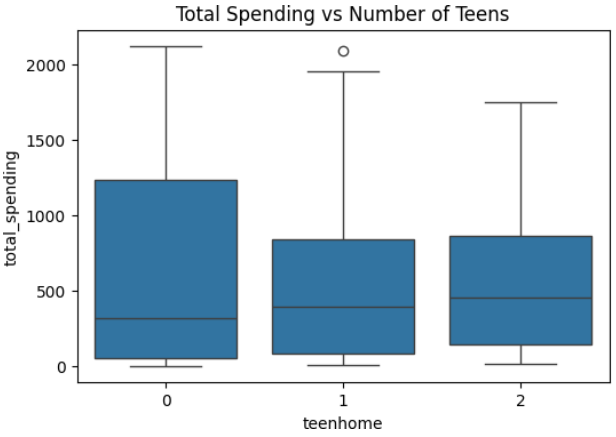
F. Marital Status vs Purchasing Behavior

Boxplot shows that the marital-status groups: Married, Together, Divorced, and Widow Customers have higher medians than Single, Alone, YOLO, and Absurd. It suggests that marital status is a meaningful driver of consumption levels.



G. Number of Teens vs Purchasing Behavior

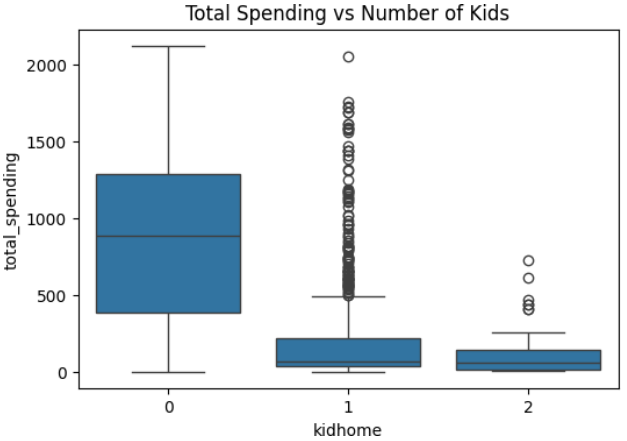
The total spending increases with the number of teens at home. This means that adolescents have higher consumption levels.



H. Number of Kids vs Purchasing Behavior

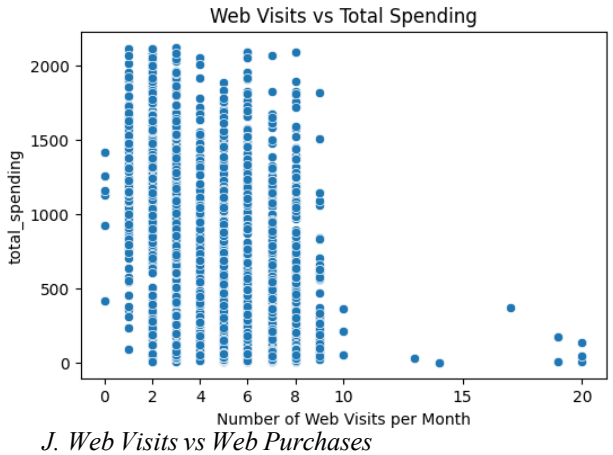
Households with more young kids have lower medians spending compared to households with no kids. This means

that families with children might have tight budget, which leads to less expenditure on the products in the dataset.



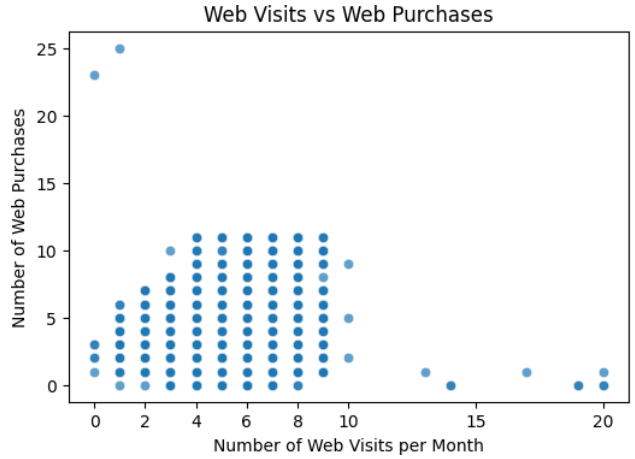
I. Web Visits vs Purchasing Behavior

There is no strong relation between web visits and total spending, which suggests that customers visit the web site just to compare the prices, rather than having the purchase intent.



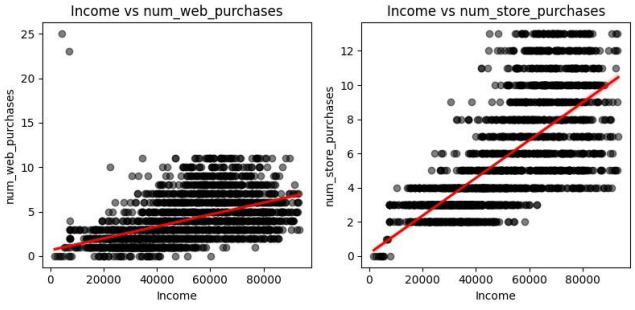
J. Web Visits vs Web Purchases

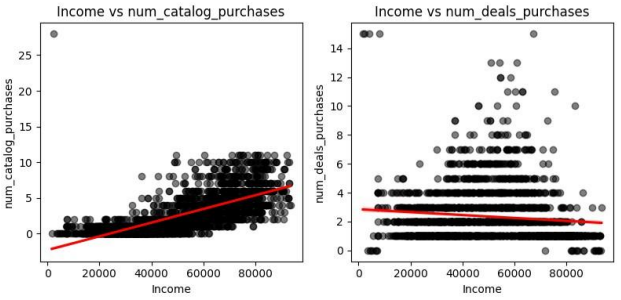
Web Visits vs Web Purchases have a non-linear relationship. Web Purchases increase initially with the number of monthly web visits up to a moderate level but then flatten.



K. Income vs Purchases by Sales Channel

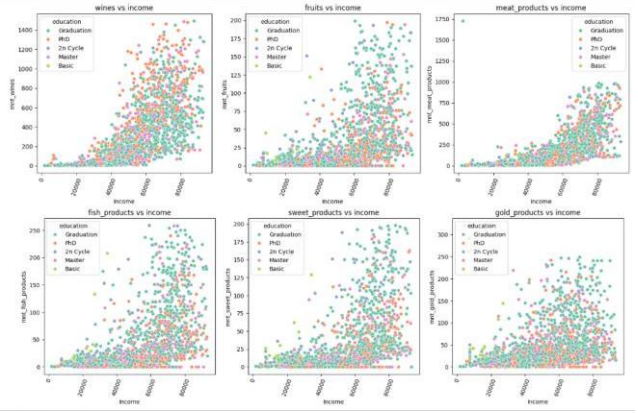
This measures customer purchase frequency through online, offline, catalog, and promotional mediums. There is a weak to moderate, positive relationship between income and web purchases, indicating that people with higher incomes are slightly more prone to web purchasing. Store purchases display a very positive correlation with income, indicating that high-income customers have continued to actively participate in physical shopping channels. There is very little to no positive correlation between deal purchases and income, implying that price sensitivity is lower for those who earn more money. This indicates that the number of catalog purchases is directly proportional to the consumer's income, with higher-income groups being more receptive to conventional methods of advertising.





L. Income, Product Spending, And Education Level

High-income consumers spend more on meat, wine, and gold products across all product categories; within the cloud, higher-education groups (Graduation, Master, PhD) are more prevalent at high spending, indicating that education amplifies the positive income-spending relationship for luxury goods.



V. CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING

A. Feature selection and data preparation

To understand more about the similarity between the customer purchasing behavior, K-means clustering was applied to segment the customers. Using numerical variables that represented customer income, recency, product-level spending, and frequency of purchases across web and store channels, K-means clustering was carried out .It measures the Euclidean distance between each data point and its cluster center and chooses the number of clusters based on where change in "within cluster sum of squares" (WCSS) levels off. This value represents the total variance within each cluster that gets plotted against the number of clusters.

B. Determination of the optimal number of clusters

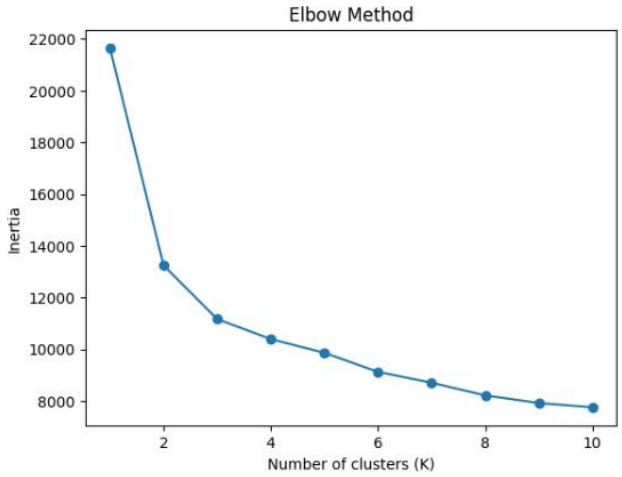
The elbow method was used to determine the appropriate number of customer segments by evaluating the within-cluster sum of squares (WCSS) for different values of k. In this method the algorithm partitions data into k clusters by minimizing the distances between points and their cluster centroids. As more clusters are added, the improvement in clustering quality gradually decreases. The "elbow" point represents the value of a point where the improvement slows down. Based on this analysis, the optimal number of clusters for the final customer segmentation was determined to be k = 3

C. Interpretation and cluster characteristics

Cluster0: Low-Value and Low-Engagement Customers: Customers in this segment have the lowest average income and spending across all the product categories. This segment of customers is price sensitive with low purchasing power of premium products (wine, meat & gold) in comparison to other clusters.

This group has low purchasing frequency across both web and store channels, shows limited engagement with the

marketing department. To increase engagement, marketing strategies may focus on promotions, discounts, or basic product offerings.



Cluster1: High-Value and High-Engagement Customers: Customers in this segment have the highest average income and are spending across all the product categories, particularly on meat, fish, wine, sweets, and gold products, which shows a strong inclination for luxury and premium items. They are the high value customers, and ideal for loyalty programs, personalized marketing and premium product promotion.

Cluster 2: Medium-Value and Medium-Engagement Customers: Customers in this segment have a moderate level of income and are spending across all the product categories. These customers fall somewhere between cluster0 and cluster1. These customers have balanced spending across all the product categories. By giving personalized suggestions, reward incentives, these people can be motivated to contribute to overall higher spending.

VI. CONCLUSION

This study focused on the analysis of the customer purchasing behavior analysis, through descriptive statistics analysis, exploratory data analysis techniques and K-means cluster analysis. From this analysis outcome, it can be concluded that income and education level are key drivers of customer expenditure, particularly for premium product categories such as wine and gold. Household composition also influences the purchasing patterns, especially the presence of teenagers while age & geographic location has a weaker relationship with purchasing behavior. With the help of k- means clustering, we segmented the customers with similar spending behavior, income levels, and purchasing frequency. Customers were segmented into low-value, medium-value, and high-value customers according to their purchasing behavior. This provided actionable insights for targeted marketing strategies and customer retention initiative

VII. LIMITATIONS AND FUTURE WORK

This research is limited to descriptive and unsupervised analytical techniques. In future work predictive modeling approaches, such as classification or regression, could be used to forecast the customer's purchasing behavior. Also, promotions or economic conditions could be incorporated to give more insights into customer segmentation and marketing. An open question that remains is to what extent can customer segments identified from a single dataset be generalized or transferred to different markets, time periods, or customer populations?

ACKNOWLEDGMENT

The authors would like to thank Kaggle for providing the publicly available data set used in this study.

REFERENCES

1. Scribd,"Customer Purchase Behavior Analysis". [Online]. Available:https://www.scribd.com/document/752771036/Int roduction
2. GeeksforGeeks, "K means Clustering".[Online]. Available: https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/
3. H. Badrnezhad, "Customer segmentation clustering," Kaggle, 2023. [Online]. Available:https://www.kaggle.com/code/hosseinbadrnezhad/ customer-segmentation-clustering
4. Coursera Staff, "Marketing analytics: What it is, why it's important, and more," Coursera, Mar. 15, 2025. [Online]. Available: https://www.coursera.org/articles/marketing-analytics

M. R. Sadeghi, "Marketinganalysis," GitHub repository.