# Report on

# Customer Clustering in E-Commerce Platforms Using K-Means

**Team Members:**

B. Ganga Reddy (23265A0503)

K. Samba (23265A0507)

**Date:** 13-11-2024

# Abstract

The **e-commerce industry** is rapidly evolving, driven by advancements in technology and changing consumer expectations. As competition intensifies, businesses face significant challenges in retaining customers and delivering personalized experiences that resonate with their unique needs. Customers now expect highly tailored interactions, which requires a deep understanding of their behavior, preferences, and purchasing patterns. To address this challenge, this project applies **K-Means clustering**, a powerful unsupervised machine learning algorithm, to analyze customer behavior and optimize marketing strategies. By using a dataset of 500 customers sourced from **Kaggle**, we aim to identify meaningful patterns and group customers based on their shared characteristics, such as **spending habits**, **membership duration**, and **engagement levels.**

This customer segmentation process provides actionable insights that can enhance **product recommendations**, ensuring that the right products are shown to the right customers. Moreover, by understanding distinct customer segments, businesses can implement more effective **retention strategies**, such as personalized offers and loyalty programs, to keep high-value customers engaged. The clustering model successfully identified **five distinct customer segments**, including **high-value long-term customers**, **occasional shoppers**, **at-risk customers**, and others. These segments enable businesses to tailor their marketing and customer service efforts in a way that resonates with each group’s specific needs, ultimately improving **customer satisfaction**, boosting **sales**, and increasing overall **customer loyalty**.

# Introduction

## Context and Background

The e-commerce industry is one of the fastest-growing sectors, characterized by intense competition. To stay ahead in such a dynamic environment, companies must gain a deep understanding of **customer behavior**. This knowledge is essential for developing **effective marketing strategies** that can not only **attract new customers** but also **retain existing ones**. Given that customers today expect personalized experiences, it is critical for businesses to leverage data-driven insights to meet these expectations.

The **e-commerce industry** is one of the fastest-growing sectors in the global economy. It has revolutionized the way businesses operate, offering consumers the convenience of shopping online from anywhere at any time. With the rise of platforms like Amazon, eBay, Alibaba, and various others, e-commerce has become a key driver of economic activity worldwide. According to recent statistics, e-commerce sales are expected to continue growing rapidly in the coming years, driven by advances in mobile technology, internet accessibility, and changes in consumer behavior.

## Purpose and Significance

The purpose of this case study is to apply **machine learning techniques**, specifically **K-Means clustering**, to analyze and segment customers within an e-commerce platform. By grouping customers based on shared characteristics, this study aims to provide businesses with valuable insights into their customer base, enabling them to create more **personalized experiences** and develop highly targeted marketing strategies.

In the highly competitive world of e-commerce, understanding customer behavior is crucial for **enhancing customer retention** and **optimizing marketing efforts**. Companies that fail to identify the differing needs of their customers may struggle to keep loyal buyers engaged or fail to convert new visitors into long-term clients. This case study aims to bridge that gap by analyzing patterns in customer data—such as spending habits, frequency of visits, and membership duration—and clustering them into distinct groups based on these behaviors.

#### ****Significance of the Study:****

1. **Enhancing Customer Retention:** By identifying the characteristics of high-value customers, businesses can create retention strategies tailored to specific segments. For example, offering personalized discounts or loyalty programs to long-term customers could increase their engagement and prevent churn.
2. **Optimizing Marketing Strategies:** Targeted marketing is more effective than one-size-fits-all approaches. Understanding the unique needs of each customer segment allows businesses to design personalized marketing campaigns, such as customized email promotions, advertisements, or product recommendations.
3. **Improving Product Recommendations:** Customer segmentation allows for the development of tailored product recommendations based on the purchasing behaviors of different segments. This can increase the likelihood of cross-selling or upselling, improving overall sales and customer satisfaction.
4. **Cost Efficiency:** By focusing on high-value customers and identifying which customer segments are less engaged, businesses can allocate resources more efficiently. This can reduce marketing costs by targeting the right customers with relevant offers.
5. **Strategic Business Decisions:** Segmenting customers enables business leaders to make data-driven decisions about product offerings, pricing strategies, and customer service initiatives. It helps in identifying profitable customer segments, optimizing product assortments, and improving customer experience.

## Problem Statement

In today’s highly competitive e-commerce landscape**, customers demand personalized experiences** that cater to their individual preferences and needs. However, many businesses struggle to engage or retain their most loyal customers because they lack a deep understanding of customer behavior and preferences. As a result, these businesses often miss out on valuable opportunities for growth and profitability.

The inability to analyze and act on customer data leads to inefficient marketing strategies, generic promotions, and a one-size-fits-all approach, all of which fail to resonate with customers who are seeking more relevant, tailored interactions. This disconnect can result in **customer churn**, where satisfied customers leave for competitors who offer more personalized experiences.

Moreover, businesses often overlook valuable customer segments, such as long-term loyal customers who may require different engagement strategies than newer customers. Without the ability to segment customers effectively and understand their distinct behaviors, companies are at a disadvantage in building lasting relationships and fostering customer loyalty.

This problem is particularly critical in the **e-commerce industry**, where the volume of customer data is vast, and the competition is fierce. Businesses that fail to leverage this data to create personalized experiences risk falling behind in a market that increasingly values customer-centric approaches.

# Methodology

**Techniques Applied:**

**K-means Clustering:**

**K-Means Clustering** is crucial in the case study for segmenting customers based on their behaviors, such as spending habits, membership duration, and time on the website. By grouping customers into distinct clusters, K-Means helps identify patterns and differences between customer segments, allowing for targeted marketing strategies and personalized recommendations.

**Customer Segmentation:**

* + K-Means divides customers into clusters based on similarity in behavior, such as high-value long-term customers versus low-spending customers, which helps in targeting specific segments more effectively.

**Improved Marketing Strategies:**

* + By identifying distinct customer groups, businesses can create personalized marketing campaigns, discounts, or promotions for each cluster, increasing engagement and retention.

**Product Recommendations:**

* + K-Means helps tailor product recommendations by understanding which products appeal to which customer clusters, enhancing cross-selling and upselling opportunities.

**Identifying Retention Opportunities:**

* + The algorithm highlights loyal customers who may need retention strategies and identifies those at risk of churn, helping businesses take proactive measures.

**Steps of Analysis:**

**Data Exploration and Pre-processing:**

Data Exploration and Data Preprocessing play a crucial role in preparing the customer dataset for clustering and ensuring that the results are meaningful.

**Data Preprocessing:**

* + **Handling Missing Data:** The dataset from Kaggle may contain missing values. Preprocessing ensures these are handled (either by imputation or removal) before clustering is performed, which is essential because K-Means is sensitive to missing data.
  + **Removing Duplicates:** Inconsistent or duplicate records can distort the clustering results. By removing duplicates, the model focuses on unique customer behaviors, which leads to more accurate segmentation.
  + **Scaling Features:** K-Means is sensitive to the scale of data, so numerical features like Yearly Amount Spent and Avg. Session Length should be standardized or normalized. This ensures that features with larger ranges do not dominate the clustering process.
  + **Encoding Categorical Data:** If the dataset contains categorical features (e.g., customer segments based on membership type), these need to be converted into numerical data through techniques like one-hot encoding.
  + **Outlier Detection:** Outliers, such as unusually high spenders or extremely low-engagement customers, might skew the clustering process. Identifying and handling these outliers ensures that the clustering algorithm doesn't produce misleading results.

### ****Data Exploration****

* **Objective:** Understand the structure, distribution, and potential issues in the data.
* **Action:**
  + Summarize the dataset to understand its shape (rows and columns), data types, and missing values.
  + Visualize the data using techniques such as **scatter plots, histograms**, and **box plots** to identify trends, outliers, and potential relationships between features.
  + Check for missing values or duplicates using functions like is.na() and duplicated(). Decide whether to remove or impute missing data.
  + Explore the basic statistics (mean, median, range) of numerical features to get a sense of their distribution.

### ****Apply K-Means Clustering:****

Once the number of clusters (K) is decided, apply the **K-Means algorithm** to the dataset:

* **Initialize Centroids:** Randomly select K centroids.
* **Assign Data Points to Clusters:** Each data point is assigned to the nearest centroid based on a distance metric.
* **Recomputed Centroids:** Update the centroids by computing the mean of all points in each cluster.
* **Repeat:** The process of assigning points and updating centroids is repeated iteratively until the centroids no longer move significantly or the maximum number of iterations is reached.

### ****Visualize the Clusters:****

Visualization is key to understanding and interpreting the clustering results, especially when dealing with multi-dimensional data. Some visualization techniques include:

* **Scatter Plots:**
  + Visualize the clusters by plotting two or three relevant features on the axes, such as **Annual Spend** and **Membership Length**, and use colors to distinguish between clusters.

# Analysis and Results

# Data collection and preprocessing:

* **is.na()**:
* **Purpose:** Checks for any missing values in the dataset.
* **Explanation:** This function identifies NA (Not Available) values, which represent missing or incomplete data entries. It allows for the identification of fields that require imputation or removal during data cleaning.
* **duplicated()**:
* **Purpose:** Identifies and removes any duplicate records from the dataset.
* **Explanation:** This function checks for rows that are exact copies of others and marks them as duplicates. By removing these duplicates, the dataset becomes more accurate, reducing bias that may result from repeating data.

**Result**

A number and percentages on a white background

Description automatically generated

## Exploratory Data Analysis



**- Key Insights:**

 **High-Spending, Long-Term Customers:**

* **Insight:** Customers with 5+ years of membership who spend significantly more annually are a valuable segment for the business. They are likely to be loyal and more engaged.
* **Action:** Focus on retaining this group through loyalty programs, exclusive offers, and VIP treatment. Personalized communications such as special discounts or exclusive product recommendations can increase their engagement and lifetime value.

 **Low-Spending, Long-Term Customers:**

* **Insight:** Customers who have been with the company for a long time but spend very little annually represent a potential retention challenge.
* **Action:** Offer re-engagement strategies, such as special offers, discounts, or new product recommendations that may incentivize them to spend more. Implement personalized emails or loyalty programs to reactivate this segment.

 **High-Spending, Short-Term Customers:**

* **Insight:** These customers have recently joined the platform but are already spending significantly, suggesting they may be new, high-potential customers.
* **Action:** Nurture this segment with targeted campaigns to deepen their relationship with the brand. Provide exclusive deals and early access to products to encourage long-term loyalty.

 **Low-Spending, Short-Term Customers:**

* **Insight:** These customers have only been on the platform for a short time and are not spending much. They may need more engagement to see the value of the platform.
* **Action:** Target this group with welcome offers, educational content, and initial discounts to encourage more interaction and increase their likelihood of becoming high-value customers.

# Discussion

## Implications of Findings

* **Improved Customer Segmentation and Targeting:**
  + Customers are segmented into distinct groups based on spending behavior, engagement, and membership duration.
  + Enables businesses to tailor marketing strategies for each segment (e.g., high-value customers receive exclusive offers, low-spending customers get discounts).
  + Leads to more efficient marketing and better resource allocation.
* **Personalized Marketing Strategies:**
  + - Allows businesses to develop targeted marketing campaigns based on the characteristics of each customer group.
  + Marketing messages and product recommendations become more relevant to each segment, increasing engagement and conversion rates.
* **Enhanced Customer Retention Efforts:**
  + Identifies at-risk customers who may be spending less or showing signs of disengagement.
  + Proactive churn prevention strategies, like personalized outreach, exclusive promotions, or loyalty rewards, can be implemented to retain these customers.
* **Optimized Product Recommendations:**
  + Informs **product recommendation engines** by identifying the types of products preferred by each customer segment.
  + **Targeted recommendations** based on customer preferences (e.g., premium products for high-value customers and budget-friendly products for low-spending customers).
* **Focused Customer Service and Engagement:**
  + Segmentation helps identify which customers may need more attention or support, such as **new customers** who may require more guidance or long-term customers who need re-engagement.
* **Strategic Decision-Making:**
* Provides actionable insights that can inform key business decisions, such as pricing strategies, product offerings, and resource allocation.

## Summary

In this case study, **K-Means clustering** was applied to customer data from an e-commerce platform to segment customers based on their behaviors, such as spending habits, time spent on the platform, and membership duration. The objective was to identify distinct customer groups and derive actionable insights to optimize marketing strategies, product recommendations, and retention efforts.

Key steps included **data exploration**, where the dataset was analyzed for patterns and cleaned for missing values and duplicates, followed by **data preprocessing** to standardize the features for clustering. The k-means algorithm was then applied to group customers into clusters, with the optimal number of clusters determined using methods like the **Elbow method** and **Silhouette score**.

### ****Key Findings:****

* **High-value, long-term customers** were identified as a critical group for retention, while **low-spending, long-term customers** represented a segment needing re-engagement strategies.
* **Short-term, high-spending customers** showed significant potential for growth, requiring targeted nurturing to increase loyalty.
* The analysis revealed opportunities for **personalized marketing, tailored product recommendations**, and **improved customer retention strategies** based on customer segmentation.

## Lessons Learned

Through the process of applying K-Means clustering to the customer data, several key lessons were learned. First, **data preprocessing** is crucial for accurate clustering, as cleaning the data (handling missing values, removing duplicates, and scaling features) ensures meaningful results. The importance of **selecting the right number of clusters (K)** was also highlighted, as methods like the **Elbow Method** and **Silhouette Score** helped ensure the clusters were distinct and relevant. **Outlier sensitivity** was another important consideration, as outliers can skew results, making it essential to address them during preprocessing. The power of **customer segmentation** became evident, enabling businesses to identify valuable customer groups and tailor marketing efforts effectively. Additionally, the process emphasized the need for **iteration and model refinement**, as adjusting parameters and continuously evaluating the model led to more precise and actionable insights. Finally, the analysis reinforced the value of **personalization**, showing how targeted strategies based on customer segments can drive higher engagement, satisfaction, and sales. These lessons underscore the importance of thorough data handling, careful modeling, and customer-centric strategies for business growth.

# References

* 1. [Ecommerce Customers](https://www.kaggle.com/datasets/srolka/ecommerce-customers)
  2. [K means Clustering - Introduction - GeeksforGeeks](https://www.geeksforgeeks.org/k-means-clustering-introduction/)
  3. [Visualizing K-Means Clustering](https://www.naftaliharris.com/blog/visualizing-k-means-clustering/)