

Customer SEGMENTATION

**PROJECT REPORT-MACHINE LEARNING TECHNIQUES-A04**



*Submitted by*

**SRI MEGHA SARVANI IKKURTHY**

**AP23322130052**

**MASTER OF BUSINESS ADMINISTRATION**

**PAARI SCHOOL OF BUSINESS, SRM UNIVERSITY, AP**

A logo of a tree

Description automatically generated

Table of Contents

**Abstract 1**

**Related Works**2

**Introduction**3

**Algorithm & Architecture4**

**Code**5

**Interpretation and Findings**9



A **Special Thanks** to Shri ram sir and Sini mam for making complex things easy to understand and for inculcating a love for learning!

**Abstract:**

**Motivation:**

In today’s highly competitive market, businesses need to understand their customers better to deliver personalized services, improve customer satisfaction, and optimize marketing efforts. Customer segmentation allows companies to group their customers based on shared characteristics, enabling them to target specific segments with tailored offerings.

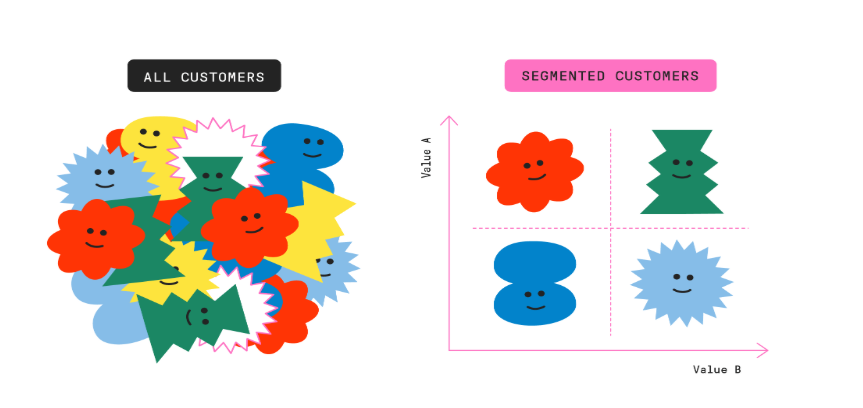
In this project, the motivation is to use data-driven techniques to segment customers effectively, providing actionable insights for targeted marketing campaigns.

**Objectives:**

The main objectives of this customer segmentation project are:

1. **To analyse customer data** and identify distinct customer segments based on purchasing behaviour and demographic information.
2. **To apply machine learning algorithms** (K-means clustering) to classify customers into groups with similar characteristics.
3. **To evaluate the performance** of the segmentation model using appropriate metrics (e.g., silhouette score, intra-cluster distance).
4. **To generate insights** that can guide marketing strategies, product development, and customer service improvement.
5. **To optimize marketing campaigns** by targeting specific customer segments with personalized offerings and promotions.

**Problem Statement:**

**The goal of this project is to segment a customer base into distinct groups using machine learning techniques.** **The dataset contains customer transaction data**

**1 | Page**

**Related works:** Let’s see how different companies leverage customer segmentation models to enhance the customer experience and drive personalized results.

1. **Netflix (Entertainment)**

* **Use Case**: Content Personalization
* **How**: Netflix segments viewers by viewing habits, genre preferences, and engagement levels to recommend shows and movies.
* **Impact**: Improved user retention and higher streaming hours.

2. **Uber (Ride-Sharing)**

* **Use Case**: Dynamic Pricing and Service Offerings
* **How**: Uber segments customers by usage patterns, location, and income levels, offering services like UberX for cost-conscious users and Uber Black for premium customers.
* **Impact**: Increased accessibility and revenue from diverse user segments.

3.**LinkedIn (Professional-Connectivity)**

* **Segment**: Job seekers based on their skills, industries, and experience levels.
* **How**: LinkedIn uses behavioural data, such as profile details and job application history, to recommend relevant job postings.
* **Impact**: Improved user engagement and higher job application success rates.

Every company needs to do customer segmentation to cater the needs of customers.

2**|**Page

**Introduction:**

Our objective is to **“Segment customers based on purchasing behaviour to identify marketing strategies for different groups”**

We have retail store customer transactional data, now we are going to use **K-Means Clustering**, a popular unsupervised machine learning algorithm, to segment customers based on their characteristics and behaviours.

By analysing customer data such as demographics, purchase history, and spending patterns, the project aims to group customers into distinct segments with similar traits.

**Dataset:** Transactional data from a retail store with 5,40,000 records

Data dictionary:

* Invoice No: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction.
* Stock Code: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
* Description: Product (item) name. Nominal.
* Quantity: The quantities of each product (item) per transaction. Numeric.
* Invoice Date: Invoice Date and time. Numeric, the day and time when each transaction was generated.
* Unit Price: Unit price. Numeric, Product price per unit in sterling.
* Customer ID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
* Country: Country name. Nominal, the name of the country where each customer resides

We have to do data cleaning, preprocessing before we step on to the clustering.

3|Page

**Algorithm:**

K-Means Clustering is an unsupervised machine learning algorithm used to group data into clusters based on their similarity. Here’s how it works:

1. **Initialization:** Choose the number of clusters (k) and randomly initialize (k) centroids (points representing the centre of each cluster).
2. **Assignment:** Assign each data point to the nearest centroid based on a distance metric (e.g., Euclidean distance).
3. **Update:** Recalculate the centroids as the mean of all points assigned to each cluster.
4. **Repeat:** Iterate between the assignment and update steps until centroids stabilize (no significant change) or a maximum number of iterations is reached.

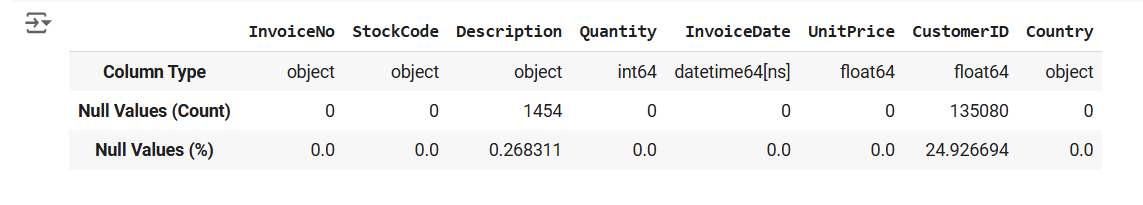
The result is k clusters, where data points within the same cluster are more similar to each other than to those in other clusters.

**Architecture:**

* Data Preparation
* Handling Missing Values
* Handling Duplicates
* Outliers Removal
* Feature Engineering
* Standard Scaling
* Elbow method for optimal clusters
* K-means Clustering
* Interpretations and Findings

4|Page

**Data Preparation:**

* Imported all the necessary libraries
* Loaded the data set into workbook with name ‘cdata’
* Visualized the data using basic functions like head(), tail(), dtypes, info, shape, describe.
* Created a Data Frame to tab\_info to visualize the column properties.

**Handling Missing Values:**

* Imported missingno module to handle the missing values

**A screenshot of a computer

Description automatically generatedA graph of numbers and text

Description automatically generated**

* Visualized the missing values and found out that there are missing values in **Customer ID** and **Description columns**
* Tried to find the correlation impact of both variables and removed the missing values from the data set. [Correlation 0.1 is very less]

**Handling Duplicate values:**

* Removed duplicate values from the data set

****

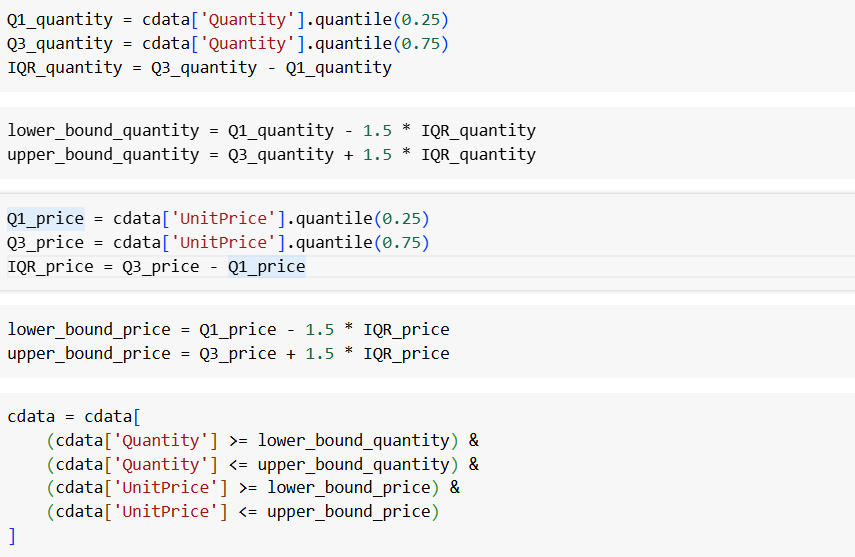
* A black and white image of a person's hand

  Description automatically generatedMade sure that there no duplicates in the dataset

5 | P a g e

**Outlier Detection:**

We will be working on **unit price and quantity,** the two numerical columns, so lets remove the outliers in those columns.



* Defined the IQR for both columns and filtered out the rows that falls in the limits.

**Feature Engineering:**

* Created New column total spending



* Created another data frame tab\_info2 to visualize products, transactions and customers.

A close up of a number

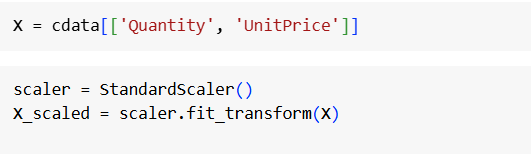
Description automatically generated

* A screenshot of a number of transactions

  Description automatically generatedTried to visualize the top customers based on number of transactions
* Created product matrix using pivot table
* Tried to cut extract the name of the product from description.

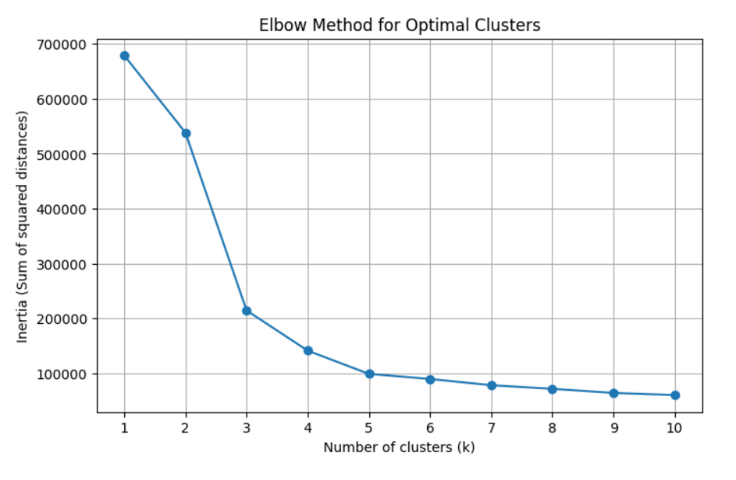
**6| Page**

**Standard Scaling:** The 2 numeric columns in the dataset are Quantity and Price. We standardize the values of both columns to visualize the clusters effectively.



**Elbow method:**

* To find the optimal number of clusters we try on iterative basis.
* The point where SSE stop decreasing further is called elbow point. That is the optimal k-value.

****

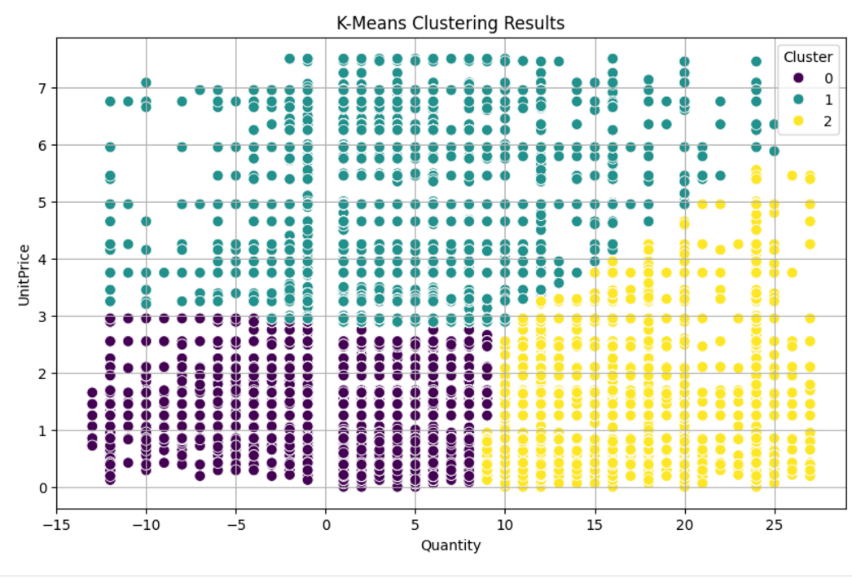
**K-Means Clustering:**

A screenshot of a computer code

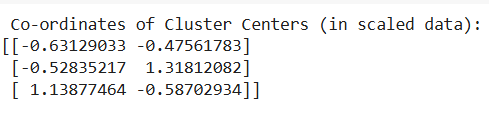
Description automatically generated

* Fitting the model and assigning the data to the dataset.

7|Page



**Cluster Centres and Coordinates:**

****

**A white background with black text

Description automatically generated**

**Results and Findings:**

* Cluster 0: Centred at (-0.631, -0.476), suggesting lower Quantity and Unit Price.
* Cluster 1: Centred at (1.138, -0.587), indicating higher Quantity but lower Unit Price.
* Cluster 2: Centred at (-0.528, 1.318), representing a group with a higher average Unit Price but moderate Quantity

8|Page

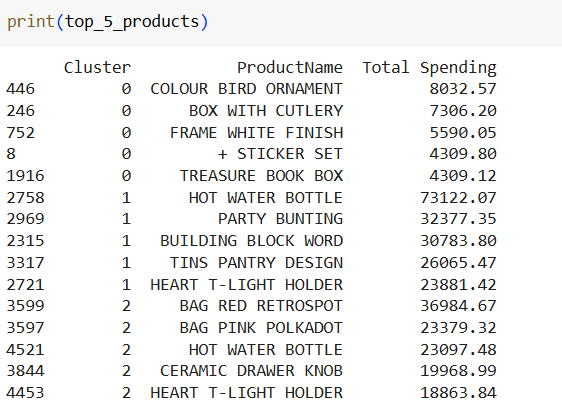
**Cluster Counts:**

* Cluster 0: Contains the largest number of data points (126,863), indicating a substantial portion of the data falls within the lower Quantity and Unit Price.
* Cluster 1: Contains 97,176 data points, representing items with higher Quantity but relatively lower Unit Price.
* Cluster 2: Contains 115,414 data points, showing items with higher Unit Price but moderate Quantity.

**Interpretation of Clusters**:

* Cluster 0: Represents low-quantity, low-price items, likely involving customers making smaller purchases of inexpensive items. The cluster might indicate **Casual or single-item buyer’s segment**
* Cluster 1: Represents large-quantity, low-price purchases, potentially reflecting bulk buying of low-cost items. **Bulk Buying Segment**
* Cluster 2: Represents moderate quantities of higher-priced items, possibly indicating purchases of higher-quality or premium products but in smaller quantities. **Premium customer segment**

**Now we will try to find top 5 products in each cluster:**

****

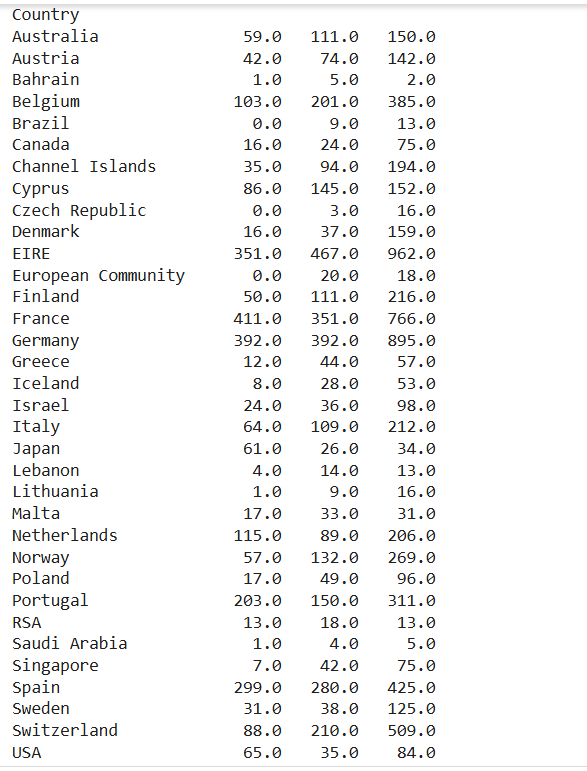
In the first cluster **casual buying segment** the products were cutlery, sticker set and book box

In the 2nd cluster **Bulk buying segment,** the products wereParty bunting, building blocks word and pantry design

In the 3rd cluster **Premium buying segment** Bags, Ceramic knobs, T-light holders

9|Page

**Country wise distribution of clusters:**

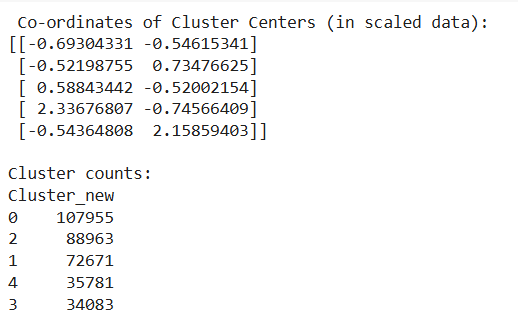
* I tried to create pivot table to visualize the country wise distribution of 3 clusters.
* Tried to visualize that using heatmap.
* It helps us to know which segment is doing good in which country.
* This helps to customize the marketing strategies based on the purchasing behaviour distributed across 3 categories.
* We can visualize the pivot table using the heatmap.

**Conclusion: We can devise the marketing strategies based on the data generated.**

**10 | Page**

**Trying with another K-value-5:**

* Tried deploying another K-means clustering model with different k-value
* My findings are



**Cluster 0:** Quantity: 2.58 (low) Unit Price: 1.36 (low)

Interpretation: These customers buy small quantities of low-priced items. Likely budget-conscious or infrequent buyers.

**Cluster 1:** Quantity: 3.75 (moderate) Unit Price: 3.34 (moderate)

**11 | Page**

Interpretation: These customers purchase moderate quantities of moderately priced items. Could be occasional buyers of mid-range products.

**Cluster 2:** Quantity: 11.41 (high) Unit Price: 1.39 (low)

Interpretation: These customers buy high quantities of low-priced items. Likely wholesale buyers or price-sensitive bulk purchasers.

**Cluster 3:** Quantity: 23.45 (very high) Unit Price: 1.05 (very low)

Interpretation: These are heavy bulk buyers of very low-priced items. Could represent resellers or highly price-conscious bulk customers.

**Cluster 4:** Quantity: 3.60 (moderate) Unit Price: 5.55 (high)

Interpretation: These customers buy moderate quantities of high-priced items. Likely premium buyers with specific preferences.

**Marketing Strategy:**

**Cluster 0 & 1:** Target with discounts or promotions for upselling to increase purchase quantities.

**Cluster 2 & 3:** Offer bulk deals and loyalty rewards, as these customers are likely to respond well to volume-based incentives.

**Cluster 4:** Focus on premium branding and exclusive offers to retain these high-value customers.

**12 | Page**