Customer Segmentation/Clustering:

Objective:

The goal of this analysis was to better understand our customer base by dividing them into distinct groups, or segments, based on both their purchasing habits and demographic information. By doing so, we can offer more targeted marketing, improve customer experiences, and ultimately drive business growth

Datasets Overview:

We used two datasets to perform this analysis:

1. **Customers.csv:**
    - This dataset provides demographic details about each customer, like:
        - Region: The geographic location of the customer.
        - Signup Date: When the customer first registered**.**
2. **Transactions.csv:**
    - This dataset includes transaction details, such as:
        - Quantity: The total number of items purchased in each transaction.
        - Product ID: The unique IDs for each product purchased.

Methodology

1. Data Preprocessing:

Region Encoding:

Since the Region was stored as text (like "North", "South", etc.), we needed to convert it into numerical form. This was done using Label Encoding, which assigns a unique number to each region. This step was crucial because clustering algorithms, like K-Means, work best with numeric data.

Signup Date Conversion:

The Signup Date, which was originally in a date format, was converted into a Unix timestamp. A Unix timestamp is a numerical representation of a date, which made it easier for the algorithm to process and analyze the timing of customer signups.

2. Data Merging:

To get a more complete understanding of each customer's behavior, we merged the customer data with their transaction data. This was done by matching on the Customer ID, so we could analyze both the demographic and transactional aspects of each customer together.

3. Feature Aggregation:

For each customer, we calculated two important metrics:

- **Total Quantity Purchased**: This is the total number of products each customer bought. It tells us how much they are buying overall.
- **Number of Unique Products Purchased**: This metric shows how diverse their purchases are — whether they tend to buy a wide range of different products or focus on a few specific ones.

4. Feature Selection:

We chose the following features to help segment our customers into meaningful groups:

- **Region (encoded)**: This represents where each customer is located.
- **Signup Date (timestamp)**: This indicates when each customer first signed up.
- **Total Quantity Purchased**: This reflects how much a customer has bought in total.
- **Number of Unique Products Purchased**: This shows how varied the customer's purchases are across different products.

5. Feature Scaling:

To ensure fairness in how each feature contributes to the clustering process, we scaled the features. This step was important because without scaling, features with larger numeric ranges (like total quantity) could disproportionately influence the clustering, potentially overshadowing features like region or signup date.

6. Clustering (K-Means):

Finally, we applied **K-Means clustering**, a method that groups customers based on similarities in their purchasing behaviour and demographics. We decided to divide customers into **4 clusters**. This number is a starting point and can be adjusted as needed based on the data and specific business objectives.

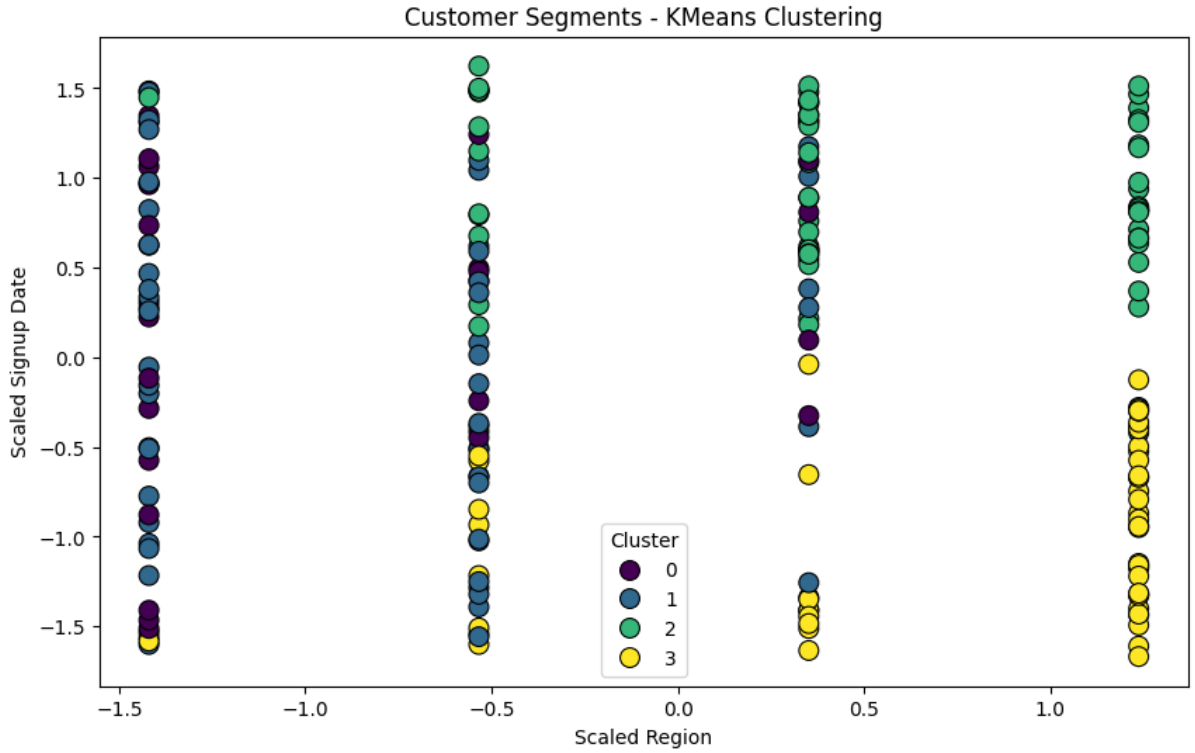
Results:

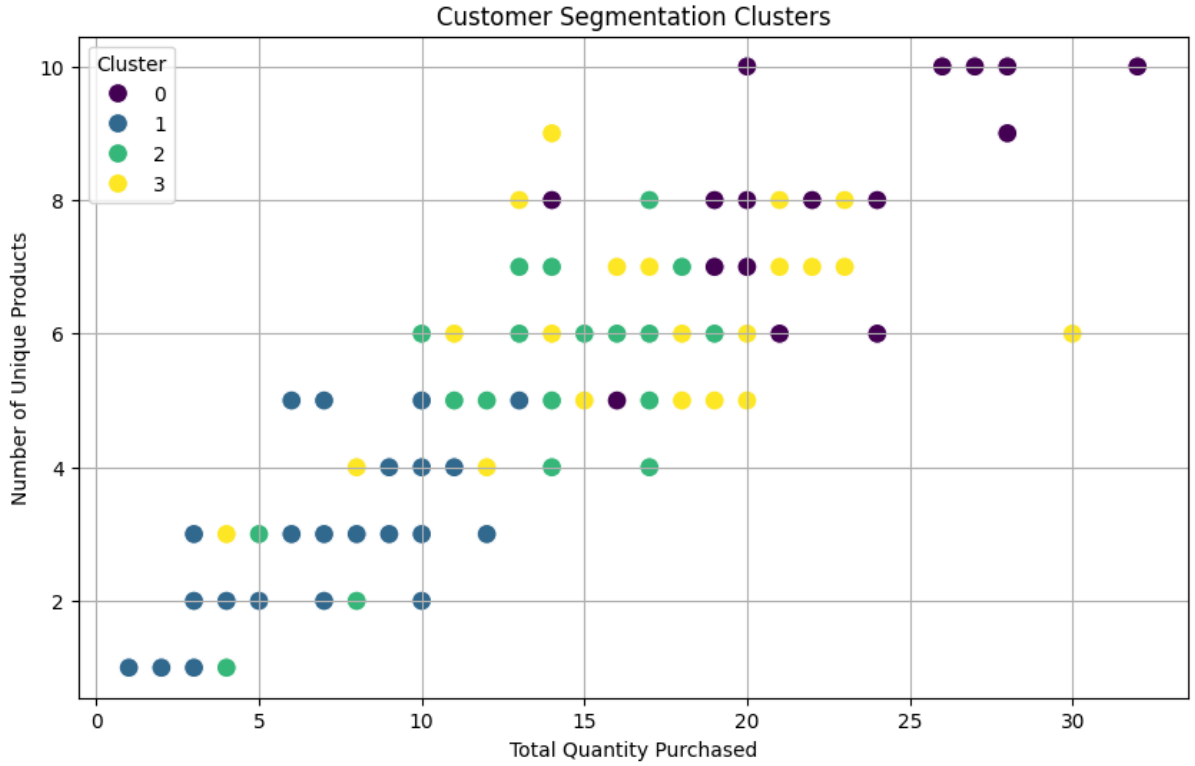
• Number of Clusters Formed: 4

• Davies-Bouldin Index: 1.1725

• Average Silhouette Score: 0.3736

Snapshots


Customer Segments - KMeans Clustering
Scaled Signup Date
Scaled Region
Cluster
0
1
2
3
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5


Customer Segmentation Clusters
Number of Unique Products
Total Quantity Purchased
Cluster
0
1
2
3
10
8
6
4
2
0
5
10
15
20
25
30