Camarines Sur Polytechnic Colleges
College of Computer Studies
Nabua, Camarines Sur

NAME: SEAN XANDER B. AQUINO, BSCS - 2A

The UCI Machine Learning Repository provides the Online Retail Dataset for this program. The dataset includes transaction data of a UK-based online retailer between December 2010 and December 2011. The dataset tracks purchase data of customers, mostly from wholesale businesses, on their product range.

Key Features:

- ***InvoiceDate***: Timestamp of purchase.
- ***Quantity***: Number of units purchased per product,
- ***UnitPrice***: Price of each unit, which is measured in GBP.
- ***CustomerID***: Unique customer identifier.
- ***Description***: Product name.
- ***Country***: Country of the customer.

A set of data-preparation procedures was needed before conducting clustering analysis. The data cleaning started by eliminating everything that lacked *CustomerID* values, because this step maintained a clear link between all transactions and their respective customers. Transactions containing negative values in *Quantity* and *UnitPrice* fields were removed, because they represented either returns or processing errors. Additional filtering of the dataset maintained only transactions involving *United Kingdom* customers, in order to conduct a consistent market analysis. *TotalPrice* served as a new calculation that computed the total transaction costs by multiplying *Quantity* with *UnitPrice*. Aggregation of customer-level features completed to produce three summarized behavior metrics that included *TotalQuantity* (items purchased), *TotalSpent* (transaction sum), and *ProductVariety* (unique items).

The analyzed data underwent *KMeans* Clustering analysis, as an unsupervised learning method to organize similar purchasing behavior customer groups.

- The application of StandardScaler accomplished feature scaling.
- Visual inspection will be possible through 2D representation, after applying Principal Component Analysis (PCA) to the data.
- The Silhouette Score identified the right cluster number as k = 3.

The customer segments, formed according to *TotalSpent*, *ProductVariety*, and *PurchaseFrequency* data, received these labels:

- **High-Value Customers**
  - Highest spending, most frequent purchases, and high product diversity.
- **Moderate Customers**
  - Average spending and frequency, loyal but not premium buyers.
- **Low-Value Customers**
  - Infrequent, low-spending customers with limited product engagement.

Using PCA, the 2D scatter plot displays visible clusters, which have been named after customer segments. This visualization helps businesses:

- Administrators should select VIP customers for membership in loyalty schemes.
- Target moderate buyers with promotions.
- Companies should understand the threat of customer turnover in their low-value market segments.