A blue logo with a arrow

Description automatically generated

**FORE SCHOOL OF MANAGEMENT**

**Academic Year 2023-2025**

MACHINE LEARNING FOR MANAGERS - II

**TITLE:** **Customer Segmentation on Online Sales Data**

(UNSUPERVISED LEARNING – K-MEANS CLUSTER)

Submitted to: Prof. Amarnath Mitra

Submitted by: Kshitij Tiwari

Roll No: 321146

PDGM 32 Section C

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S. No** | **Project Contents** | **Page No.** |
| **1** | **Project Objectives** | **3** |
| **2** | **Description of Data** | **4** |
| **3** | **Analysis of Data** | **6** |
| **4** | **Results | Observations** | **9** |
| **5** | **Managerial Insights** | **11** |

**Project Title: User segmentation on Online sales.**

**1. PROJECT OBJECTIVES**

1.1 Segmentation of Consumer Data about online sales.

1.2 Identification of the Appropriate Number of Customer Segments, with K = 2,3,4 and 5 and comparison of Silhouette Scores.

1.3 Determination of the Characteristics of each Customer Segment.

**2. DESCRIPTION OF DATA**

**2.1. Data Source, Size, Shape**

2.1.1 Data Source: <https://www.kaggle.com/datasets/hellbuoy/online-retail-customer-clustering>

2.1.2. Data Size: The size of the data is 3 MB.

2.1.3 Data Shape: Number of Variables = 8, Number of Observations: The dataset has 54,099 records (based on the number of rows).

**2.2. Description of Variables**

**2.2.1 Index Variables:** CustomerID: gives a unique identification to customers.

**2.2.2 Categorical Variables:**

* InvoiceNo: a number that appears on an invoice and uniquely identifies it for easy organisation and identification later.
* StockCode: A stock code identifies an item of stock on a stock quotation scheme.
* Description: Describes the product
* InvoiceDate: Date of transaction
* Country: Country at which transaction took place.

**2.2.2.1 Categorical Variables - Nominal Type:**

InvoiceNo, StockCode, Description, CustomerID, Country

**2.2.2.2 Categorical Variables - Ordinal Type:** No ordinal data

**2.2.3 Non-Categorical Variables:**

* Quantity: Quantity of items sold
* UnitPrice: Unit price of each item sold.

**2.3 Descriptive Statistics**

2.3.1 Descriptive Statistics: Categorical Variables or Features

2.3.1.1. Count | Frequency Statistics

A screenshot of a table

Description automatically generated

**2.3.2. Descriptive Statistics: Non-Categorical Variables or Features**

2.3.2.1. Measures of Central Tendency

2.3.2.2. Measures of Dispersion

For Categorical Variables

A screenshot of a calendar

Description automatically generated

For Non-Categorical Variables

A screenshot of a calendar

Description automatically generated

**3. ANALYSIS OF DATA**

**3.1 Data Pre-Processing**

**3.1.1 Missing Data Statistics and Treatment**

3.1.1.1.1. Missing Data Treatment: For Categorical Variables, it is done my Mode of the variables.

A screenshot of a calendar

Description automatically generated

3.1.1.1.2 For Non – categorical variables, missing data treatment is done my Mean.

A screenshot of a computer

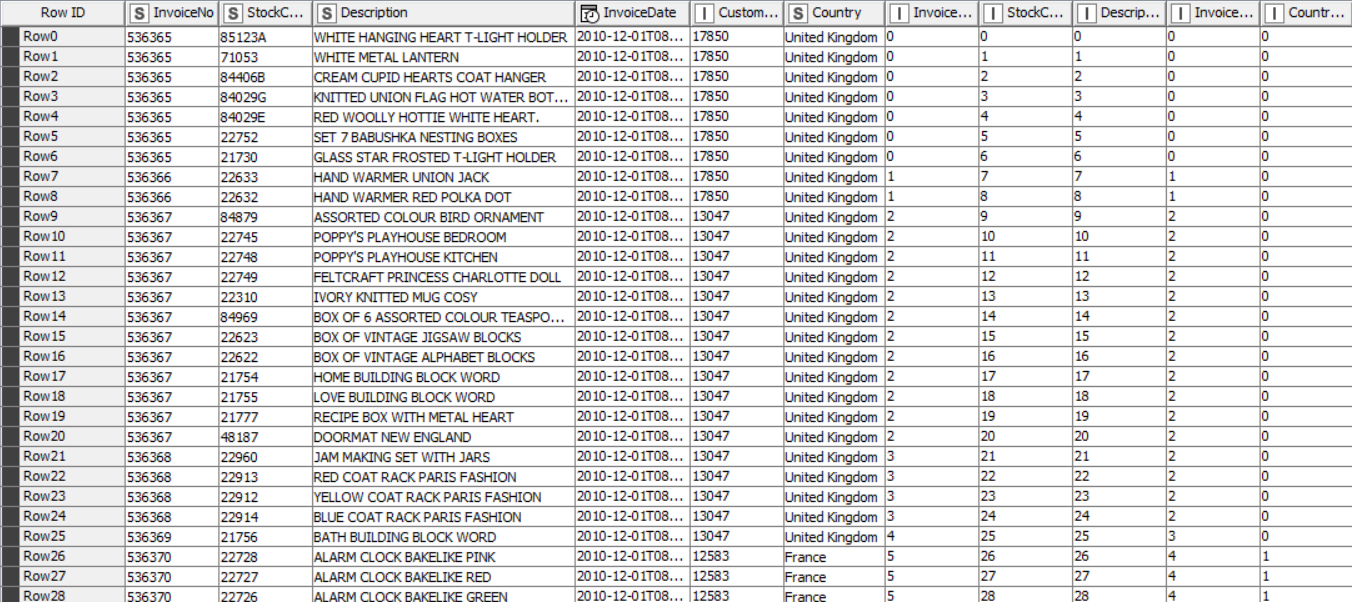
Description automatically generated

3.1.1.1.2.1. Removal of Records with More Than 50% Missing Data: None since no record has more than 50% missing data.

**3.1.2. Numerical Encoding of Categorical Variables or Features**

Encoding Schema - Alphanumeric Order

Variables that are converter into numerical form: Invoice, StockCode, Description, Country

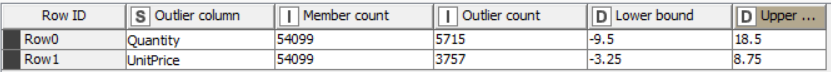
****

**3.1.3. Outlier Statistics and Treatment**

A black and white image of a cross

Description automatically generated

For the non-categorical data, we do outlier treatment by using box plot to analyse the outliers. The numeric outliers are shown below.



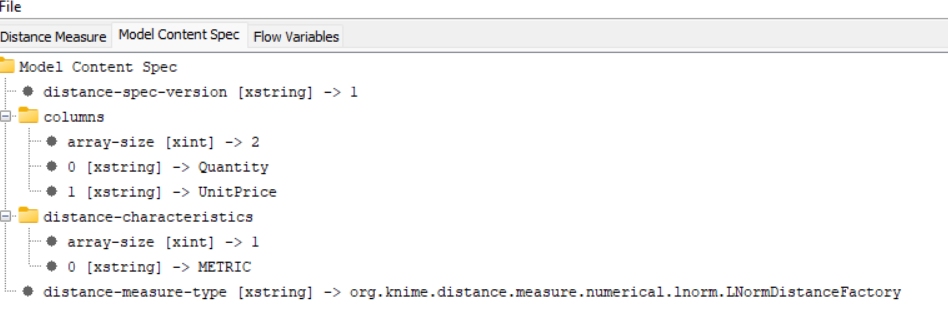
For Normalization, we have used Mean-max Normalization, also referred to as feature scaling, which is a method utilized in data preprocessing to rescale numerical attributes within a defined range, commonly between 0 and 1. The min-max normalization formula is expressed as follows:

A black text on a white background

Description automatically generated

**3.2. Data Analysis**

3.2.1.1. PO1 | PS1:: Unsupervised Machine Learning Clustering Algorithm: K-Means | Metrics Used - Euclidean Distance



3.2.1.2. PO1 | PS1:: Unsupervised Machine Learning Clustering Algorithms:

{DBSCAN} | Metrics Used - Euclidean Distance

A screenshot of a computer

Description automatically generated

3.2.3.1. PO3 | PS3:: Cluster Analysis: Base Model (K-Means)

3.2.3.1.1. Cluster Analysis with Categorical Variables or Features: Chi-Square Test of Independence

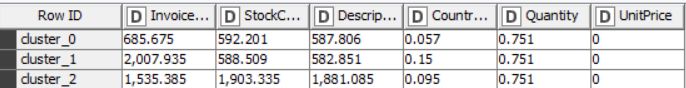


3.2.3.2.2. Cluster Analysis with Non-Categorical Variables or Features: Analysis of Variance (ANOVA)

**4. RESULTS**

4.1. Appropriate Number of Segments | Clusters: Base Model (K-Means)

Through analysis, we found that the appropriate number of clusters is 3.



4.2. Non-Categorical Variables or Features: Contributing or Significant | Non-Contributing or Non-Significant





We conducted an ANOVA test for the two non-categorical variables, UnitPrice and Quantity. The test yielded significant differences between groups.

For Description, the Chi-Square test of independence through Kruskal Walis test yielded significant difference.



For Country, the Chi-Square test of independence through Kruskal Walis test yielded significant difference.

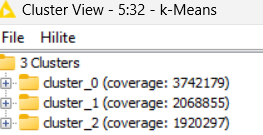
****

For CustomerID, the Chi-Square test of independence through Kruskal Walis test yielded significant difference.

****

**OBSERVATIONS**

It is observed that the sales data is segmented and the appropriate number of segments in the data is 3.

****

These categorical variables Description, Country and CustomerID can be individually distinguished as they exhibit a significant p-value among groups.







Description



Country

****

CustomerID

****

**5. MANAGERIAL INSIGHTS**

5.1. Appropriate Model: K-means

5.2. Appropriate Number of Segments: 3

5.3. Segment | Cluster - (Heterogeneous) Identity

Various groups exhibit distinct average quantities and unit prices: A statistically significant difference is observed in both quantity and unit price across the three groups (referred to as cluster 0, cluster 1, and cluster 2). This indicates that the average quantity and unit price vary among the groups.

1. One group will consistently purchase larger quantities at lower unit prices, possibly benefiting from bulk discounts.
2. One other group will procure smaller quantities at higher unit prices, suggesting purchases akin to those made at convenience stores.
3. A mix of scenarios is likely present across the different groups, indicating diverse purchasing behaviours and preferences within each cluster.