

**PROJECT NAME – CUSTOMER SEGMENTATION USING K MEANS CLUSTERING**

**Submitted by: Submitted to:**

**Naman Agarawal Prof. Amarnath Mitra**

**(064030)**

**Project Contents**

1. **Project Objectives | Problem Statements**
2. **Description of Data**
3. **Analysis of Data**
4. **Results | Observations**
5. **Managerial Insights**

**Project 1: Report (Sample)**

**Project Title: Segmentation of Consumer Data**

**1. Project Objectives | Problem Statements**

1.1. PO1 | PS1: Segmentation of Consumer Data using Unsupervised Machine Learning Clustering Algorithms

1.2. PO2 | PS2: Identification of Appropriate Number of Segments or Clusters

1.3. PO3 | PS3: Determination of Segment or Cluster Characteristics

**2. Description of Data**

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

2.1.1. Data Source - [https://www.kaggle.com](https://www.kaggle.com/datasets/mervemenekse/ecommerce-dataset/data)

2.1.2. Data Size – 5.75 MB

2.1.3.1 Data Shape – 51289\*16

2.1.3.2 Dimension: Number of Variables – 16 | Number of Records - 51289

**2.2. Description of Variables**

2.2.1. Index Variable(s): I1 - Order Date

I2 – Time

I3 – Customer Id

2.2.2. Variables or Features having Categories | Categorical Variables or Features (CV)

2.2.2.1. Variables or Features having Nominal Categories | Categorical Variables or Features - Nominal Type: CNV1 – Gender - Gender of customer.

CNV2 – Device Type - The device the customer uses to actualize the transaction (Web/Mobile).

CNV3 – Product Category - Product category

CNV4 – Product – Product

CNV5 – Payment Method - Payment method

CNV6 – Customer login type - The type the customer logged in. Such as Member, Guest etc.

2.2.2.2. Variables or Features having Ordinal Categories | Categorical Variables or Features - Ordinal Type: COV1 – Order Priority - Order priority. Such as critical, high etc.

2.2.3. Non-Categorical Variables or Features:

NCV1 – Aging - The time from the day the product is ordered to the day it is delivered.

NCV2 – Sales - Total sales amount

NCV3 – Quantity - Unit amount of product

NCV4 – Discount - Percent discount rate

NCV5 – Profit - Profit

NCV6 – Shipping Cost - Shipping cost

**2.3. Descriptive Statistics**

2.3.1. Descriptive Statistics: Categorical Variables or Features

2.3.1.1. Count – 51290 | Frequency Statistics

2.3.1.2. Proportion (Relative Frequency) Statistics – 43.75 %

2.3.2. Descriptive Statistics: Non-Categorical Variables or Features

2.3.1.1. Count – 51290 | Frequency Statistics

2.3.1.2. Proportion (Relative Frequency) Statistics – 37.5 %

2.3.2.1. Measures of Central Tendency

NCV1 – Aging | mean - 5.255 | mode - | median - 5 |

NCV2 – Sales | mean - 152.341 | mode - | median – 133 |

NCV3 – Quantity | mean - 2.503 | mode - | median - 2 |

NCV4 – Discount | mean - 0.304 | mode - | median - 0.3 |

NCV5 – Profit | mean - 70.407 | mode - | median - 59.9 |

NCV6 – Shipping Cost | mean - 7.042 | mode - | median - 6 |

2.3.2.2. Measures of Dispersion

NCV1 – Aging | std dev - 2.9599 | variance - 8.761 | skewness - 0.0656 |

NCV2 – Sales | std dev - 66.4954 | variance - 4421.641| skewness - -0.0878 |

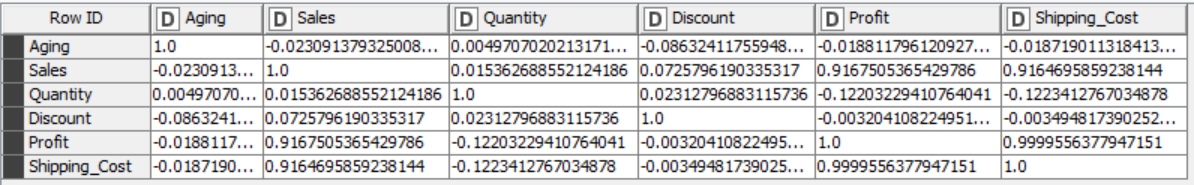
NCV3 – Quantity | std dev - 1.5119 | variance - 2.286 | skewness - 0.4642 |

NCV4 – Discount | std dev - 0.131 | variance - 0.017 | skewness - 0.0332 |

NCV5 – Profit | std dev - 48.7295 | variance - 2374.563 | skewness - 0.261 |

NCV6 – Shipping Cost | std dev - 4.8717 | variance - 23.734 | skewness - 0.2625 |

2.3.2.3. Correlation Statistics (with Test of Correlation)



**3. Analysis of Data**

**3.1. Data Pre-Processing**

**3.1.1. Missing Data Statistics and Treatment**

3.1.1.1.1. Missing Data Statistics: Records = 6

3.1.1.1.2. Missing Data Treatment: Records

3.1.1.1.2.1. Removal of Records with More Than 50% Missing Data: None |

3.1.1.2.1. Missing Data Statistics: Categorical Variables or Features

3.1.1.2.2. Missing Data Treatment: Categorical Variables or Features

3.1.1.2.2.1. Removal of Variables or Features with More Than 50% Missing Data: None |

3.1.1.2.2.2. Imputation of Missing Data using Descriptive Statistics: Mode

3.1.1.3.1. Missing Data Statistics: Non-Categorical Variables or Features

NCV1 – Aging – 1

NCV2 – Sales - 1

NCV3 – Quantity - 2

NCV4 – Discount - 1

NCV5 – Profit - 0

NCV6 – Shipping Cost - 1

3.1.1.3.2. Missing Data Treatment: Non-Categorical Variables or Features

3.1.1.3.2.1. Removal of Variables or Features with More Than 50% Missing Data: None

3.1.1.3.2.2. Imputation of Missing Data using Descriptive Statistics: Mean

**3.1.2. Numerical Encoding of Categorical Variables or Features** (Encoding Schema - Alphanumeric Order)

CV1 – Gender | Female – 0 | Male – 1 |

CV2 – Device Type | web – 0 | Mobile – 1 |

CV3 – Product Category | Auto & Accessories – 0 | Fashion – 1 | Electronic – 2 | Home & Furniture – 3 |

CV4 – Product | Car Media Players - 0 | Car Speakers - 1 | Car Body Covers - 2 | Car & Bike Care - 3 | Tyre - 4 | Bike Tyres - 5 | Car Mat - 6 | Car Seat Covers - 7 | Car Pillow & Neck Rests - 8 | Shirts - 9 | jeans - 10 | Suits - 11 | Sports Wear - 12 | Casula shoes - 13 | Running Shoes - 14 | Formal Shoes - 15 | Sneakers - 16 | Tilak Watch - 17 | Fossil Watch - 18 | T-shirt - 19 | Samsung Mobile - 20 | Watch - 21 | Fans - 22 | Iron - 23 | Tablet - 24 | Mouse - 25 | Keyboard - 26 | Apple Laptop - 27 | Mixer/Juicer - 28 | LED - 29 | LCD - 30 | Speakers - 31 - | Sofa Covers - 32 | Bed Sheets - 33 | Curtains - 34 | Towels - 35 | Sofas - 36 | Beds - 37 | Dining Tables - 38 | Shoe rack - 39 | Umbrellas - 40 | Dining Crockery - 41 |

CV5 – Payment Method | Credit card – 0 | Money Order – 1 | E-Wallet – 2 | Debit Card – 3 | Not Defined – 4 |

CV6 – Customer login type | Member – 0 | Guest – 1 | New – 2 | First Signup – 3 |

CV7 – Order Priority | Medium – 0 | Critical – 1 | High – 2 | Low – 3 |

**3.1.3. Outlier Statistics and Treatment** (Scaling | Transformation)

3.1.3.1.1. Outlier Statistics: Non-Categorical Variables or Features | There is no Outlier present in any non – Categorical variables |

3.1.3.1.2. Outlier Treatment: Non-Categorical Variables or Features

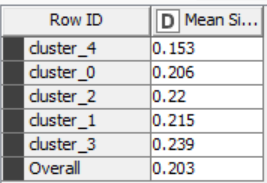
3.1.3.1.2.1. Normalization using Min-Max Scaler:

**3.1.4. Data Bifurcation: Training & Testing Sets** | Not Required |

**3.2. Data Analysis**

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

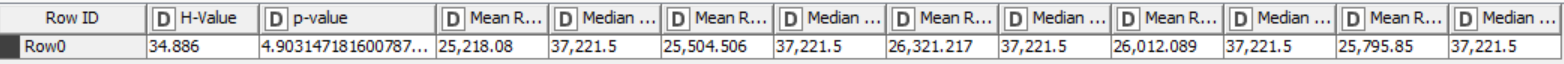
3.2.2.1.1. PO2 | PS2:: Clustering Model Performance Evaluation: Silhouette Score | (Base Model: K-Mean)



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 | Kruskal wallis test

Gender :



DEVICE TYPE



Customer Login Type



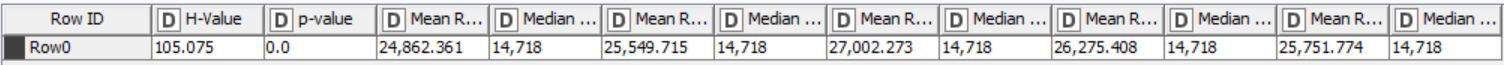
Product Category



Product



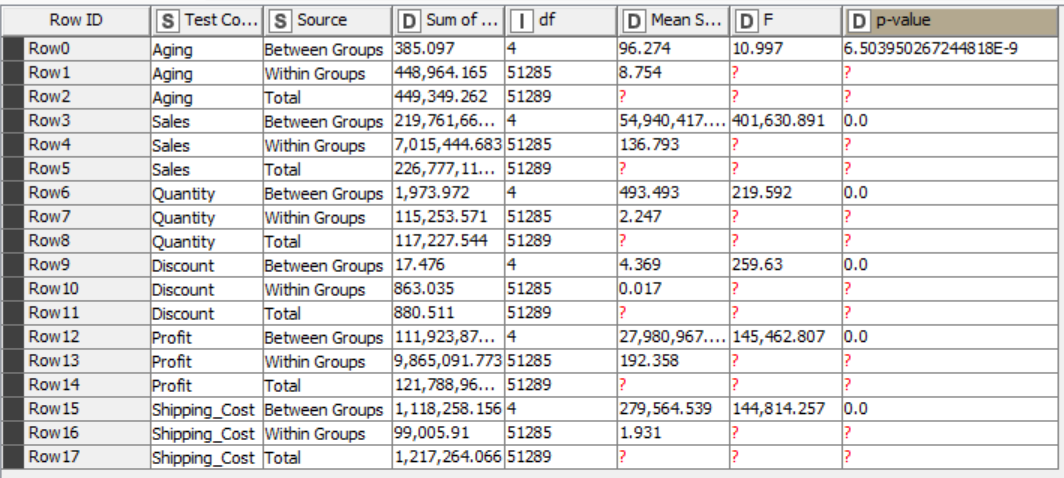
Order Priority



Payment Method



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



**4. Results | Observations**

4.1. Appropriate Number of Segments | Clusters: Base Model (K-Means) | 5 |

4.2. Cluster Analysis: Base Model (K-Means) | Comparison Models (DBSCAN | BIRCH | OPTICS: At Least One)

4.2.1. Categorical Variables or Features: **Contributing or Significant** |

Gender | p value = 4.903147181600787E^-7 | null hypothesis rejected | there is significant difference between the clusters in terms of Gender |

Device type | p value = 3.281441848667077E^-4 | null hypothesis rejected | there is significant difference between the clusters in terms of Device Type |

Product Category | p value = 0.00 | null hypothesis is rejected | there is significant difference between the clusters in terms of Product Category |

Product | p value = 0.00 | null hypothesis is rejected | there is significant difference between the clusters in terms of Product |

Order Priority | p value = 0.00 | null hypothesis is rejected | there is significant difference between the clusters in terms of Order Priority |

Non-Contributing or Non-Significant

Customer login type | p value = 0.414 | null hypothesis is not rejected |there is no difference between the clusters in terms of customer login type |

Payment Method | p value = 0.747 | null hypothesis is not rejected |there is no difference between the clusters in terms of Payment Method |

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

Aging | p value = 6.503950267244818E^-9 | null hypothesis is rejected | at least one cluster differs significantly from the overall mean of the dependent variable

Quantity | p value = 0.0 | null hypothesis is rejected | at least one cluster differs significantly from the overall mean of the dependent variable

Sales | p value = 0.0 | null hypothesis is rejected | at least one cluster differs significantly from the overall mean of the dependent variable

Discount | p value = 0.0 | null hypothesis is rejected | at least one cluster differs significantly from the overall mean of the dependent variable

Profit | p value = 0.0 | null hypothesis is rejected | at least one cluster differs significantly from the overall mean of the dependent variable

Shipping cost | p value = 0.0 | null hypothesis is rejected | at least one cluster differs significantly from the overall mean of the dependent variable

Non-Contributing or Non-Significant | None |

**5. Managerial Insights**

5.1. Appropriate Model | K-Means |

**Efficient Online Shoppers:**

Insight: These customers value efficiency and prompt delivery.

Recommendation: Invest in optimizing the shipping and delivery processes to maintain customer satisfaction. Implement features like same-day or next-day delivery options. Additionally, focus on enhancing the mobile shopping experience to cater to their preferences.

**Bulk Buyers:**

Insight: Customers in this cluster prefer purchasing products in large quantities.

Recommendation: Offer bulk discounts or incentives to encourage them to buy even larger quantities. Consider implementing subscription models or bulk purchase deals to incentivize repeat purchases. Additionally, analyze their preferences to offer personalized product recommendations based on their buying patterns.

**High-Spending, High-Profit Customers:**

Insight: These customers contribute significantly to revenue and profit.

Recommendation: Provide personalized services and exclusive offers to retain their loyalty. Offer premium membership programs with added benefits such as early access to sales, dedicated customer support, or exclusive products. Implement targeted marketing campaigns to upsell or cross-sell to these high-value customers.

**Price-Sensitive Shoppers:**

Insight: Customers in this cluster are motivated by discounts and promotions.

Recommendation: Continue offering attractive discounts and promotions to incentivize purchases. Experiment with dynamic pricing strategies to offer personalized discounts based on individual customer behavior and preferences. Focus on building customer loyalty through rewards programs or referral incentives.

**Occasional Shoppers:**

Insight: These customers make infrequent purchases.

Recommendation: Implement strategies to increase customer engagement and encourage repeat purchases. Send personalized recommendations or reminders to re-engage these customers. Offer incentives such as discounts or free shipping on their next purchase to encourage them to return. Analyze their browsing and purchase history to tailor marketing efforts effectively.

5.2. Appropriate Number of Segments | Clusters (Given the Appropriate Model) | K means | 5 |

5.3. Segment | Cluster - (Heterogeneous) Identity | Nomenclature

Cluster 1: Efficient Online Shoppers

**Characteristics:**

Short Aging time: Customers in this cluster have short durations between ordering and delivery, indicating efficient shipping and delivery processes.

Majority of customers might be frequent online shoppers.

Balanced gender distribution.

Higher proportion of mobile users compared to web users.

Mostly members rather than guests.

Recommendations: Focus on retaining these customers by offering loyalty programs or personalized recommendations based on their purchase history.

Cluster 2: Bulk Buyers

**Characteristics:**

High Quantity: Customers in this cluster tend to purchase products in large quantities.

Moderate to high sales amounts.

Moderate aging time.

Balanced gender distribution.

Likely to use both web and mobile devices for transactions.

May include a mix of member and guest logins.

Recommendations: Offer bulk discounts or promotions to encourage repeat purchases.

Cluster 3: High-Spending, High-Profit Customers

**Characteristics:**

High Sales and Profit: Customers in this cluster generate significant revenue and profit for the business.

Moderate aging time.

Balanced gender distribution.

Predominantly use web for transactions.

Mostly members.

Recommendations: Provide personalized offers or premium services to enhance customer loyalty and increase sales further.

Cluster 4: Price-Sensitive Shoppers

**Characteristics:**

High Discount Rates: Customers in this cluster tend to make purchases when offered significant discounts.

Moderate to low sales amounts.

Longer aging time, possibly due to waiting for discounts or promotions.

Balanced gender distribution.

Similar usage of web and mobile devices.

May include more guest logins compared to members.

Recommendations: Focus on targeted discount campaigns to attract these customers and encourage repeat purchases.

Cluster 5: Occasional Shoppers

**Characteristics:**

Low Frequency: Customers in this cluster make infrequent purchases.

Low to moderate sales amounts.

Longer aging time, possibly due to sporadic shopping behavior.

Balanced gender distribution.

Similar usage of web and mobile devices.

May include a mix of member and guest logins.