FORE SCHOOL OF MANAGEMENT
NEW DELHI

Academic Year 2023-25

Machine Learning for Managers

TOPIC: To find out the user segmentation of the e-commerce retail store

---

Submitted to: Prof. Amarnath Mitra

Submitted by:

Name- Arindam Chakraborty

Roll No.-321128 PGDM 32 Section C

Table of contents

Contents	Page no.
Objective of the project	3
Data descriptions	3
Analysis	5
Observation	8
Managerial Insights	10

1. OBJECTIVES

- **To find out the segmentation of the marketing data**
- **To find the no of cluster**
- **K = 2- 3- 4- 5- and compare the score by silhouette score (PERFORMANCE MATRIX)**
- **Characteristics of each**

2. DESCRIPTION OF DATA

2.1. Data description

2.1.1 Total no of variables = 8

2.1.2 no of observations = 61539

2.1.3 data Size = 10 MB

2.2. Characteristics of variables-

2.2.1 index variable = invoice_id

2.2.2 categorical Variable = Stock_code, Description, Customer_ID, Country, invoice date

2.2.2.1 nominal Stock_code, Description, Customer_ID, Country, invoice date

2.2.2.2ordinal = No ordinal Data found

2.2.3 Non-cat= Quantity, Unit_Price

2.3. Source of data- https://www.kaggle.com/datasets/yasserh/customer-segmentation-dataset?rvi=1

2.4 Descriptive Statistics

2.4.1 Count, frequency statistics

Count of country Data

Row ID	count
Australia	243
Austria	30
Bahrain	2
Belgium	263
Channel Islands	59
Cyprus	256
Denmark	40
EIRE	1026
Finland	28
France	1456
Germany	1680
Greece	32
Hong Kong	57
Iceland	60
Israel	32
Italy	169
Japan	154
Lebanon	45
Lithuania	35
Netherlands	492
Norway	179
Poland	61
Portugal	285
Singapore	56
Spain	560
Sweden	74
Switzerland	237
United Arab E...	30
United Kingdom	92358

Statistics

Rows: 2 | Columns: 5

Name	Minimum	Maximum	Mean	Standard Deviation
Quantity	-74,215	74,215	8.893	336.564
UnitPrice	0	16,888.02	5.397	120.892

For NON CATEGORICAL

Statistics

Rows: 5 | Columns: 4

Name	Type	# Missing values	10 most common values ↓
Description	String	299	WHITE HANGING HEART T-LIGHT HOLDER (543; 0.54%), REGENCY CAKESTAND 3 TIER (472; 0.47%), HEART OF WICKER SMALL (385; 0.39%), JUMBO BAG RED RETROSPOT (350; 0.35%), SET OF 3 C...
Country	String	0	United Kingdom (92358; 92.36%), Germany (1680; 1.68%), France (1456; 1.46%), EIRE (1026; 1.03%), Spain (560; 0.56%), Netherlands (492; 0.49%), Portugal (285; 0.29%), Belgium (263; 0.26%), Cyprus ...
ï»¿InvoiceNo	String	0	537434 (675; 0.68%), 538071 (652; 0.65%), 538349 (620; 0.62%), 537638 (601; 0.6%), 537237 (597; 0.6%), 536876 (593; 0.59%), 536592 (592; 0.59%), 537823 (591; 0.59%), 537240 (568; 0.57%), 5376...
StockCode	String	0	85123A (533; 0.53%), 22423 (473; 0.47%), 22469 (385; 0.39%), 85099B (350; 0.35%), 22720 (345; 0.35%), 22961 (322; 0.32%), 22470 (300; 0.3%), 22960 (299; 0.3%), 22457 (298; 0.3%), 22197 (271; 0.2...
CustomerID	Number (i...	34915	15,311 (718; 1.1%), 12,748 (713; 1.1%), 17,841 (709; 1.09%), 14,606 (684; 1.05%), 14,911 (549; 0.84%), 14,646 (468; 0.72%), 13,089 (456; 0.7%), 15,039 (370; 0.57%), 18,118 (325; 0.5%), 14,298 (324; 0....

3. Analysis

3.1 Preprocessing

3.1.1 Data bifurcation

3.1.1.1 Categorical

Categorical Stock_code, Description, Customer_ID, Country, invoice data

3.1.1.2 Non Categorical = Quantity, Unit Price

3.1.2 Missing Data treatment

Statistics

Rows: 5 | Columns: 4

Name	Type	# Missing values	Mean
ï»¿InvoiceNo	String	0	
StockCode	String	0	
Description	String	299	
CustomerID	Number (integer)	34915	15,350.988
Country	String	0	

0

Cat(treatment done by mode)

Statistics

Rows: 3 | Columns: 4

Name	Type	# Missing values	Mean
ï»¿InvoiceNo	String	0	
Quantity	Number (integer)	0	8.893
UnitPrice	Number (double)	0	5.397

And Non cat (is done by mean)

3.1.3 Cat data = Numerical Encoding

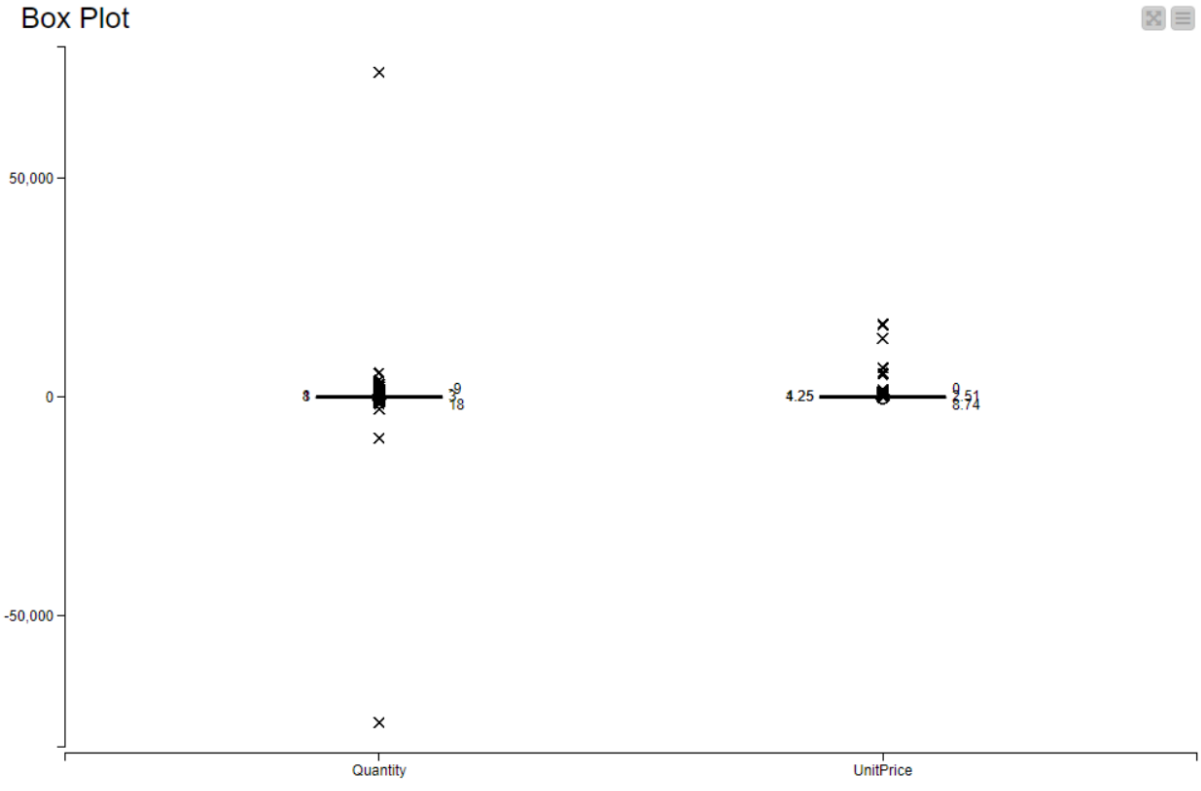
- **Variable which are already Numerical = Customer_ID**
- **Variables which are not numerical= Description, StockCode, Country, Invoice**

Row ID	ï»¿Invoi...	StockC...	Description	Custom...	Country	ï»¿Invoi...	StockC...	Countr...
Row0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	17850	United Kingdom	0	0	0
Row1	536365	71053	WHITE METAL LANTERN	17850	United Kingdom	0	1	0
Row2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	17850	United Kingdom	0	2	0
Row3	536365	84029G	KNITTED UNION FLAG HOT WATER BOT...	17850	United Kingdom	0	3	0
Row4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	17850	United Kingdom	0	4	0
Row5	536365	22752	SET 7 BABUSHKA NESTING BOXES	17850	United Kingdom	0	5	0
Row6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	17850	United Kingdom	0	6	0
Row7	536366	22633	HAND WARMER UNION JACK	17850	United Kingdom	1	7	0
Row8	536366	22632	HAND WARMER RED POLKA DOT	17850	United Kingdom	1	8	0
Row9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	13047	United Kingdom	2	9	0
Row10	536367	22745	POPPY'S PLAYHOUSE BEDROOM	13047	United Kingdom	2	10	0
Row11	536367	22748	POPPY'S PLAYHOUSE KITCHEN	13047	United Kingdom	2	11	0
Row12	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	13047	United Kingdom	2	12	0
Row13	536367	22310	IVORY KNITTED MUG COSY	13047	United Kingdom	2	13	0
Row14	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPO...	13047	United Kingdom	2	14	0
Row15	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	13047	United Kingdom	2	15	0
Row16	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	13047	United Kingdom	2	16	0
Row17	536367	21754	HOME BUILDING BLOCK WORD	13047	United Kingdom	2	17	0
Row18	536367	21755	LOVE BUILDING BLOCK WORD	13047	United Kingdom	2	18	0
Row19	536367	21777	RECIPE BOX WITH METAL HEART	13047	United Kingdom	2	19	0
Row20	536367	48187	DOORMAT NEW ENGLAND	13047	United Kingdom	2	20	0
Row21	536368	22960	JAM MAKING SET WITH JARS	13047	United Kingdom	3	21	0
Row22	536368	22913	RED COAT RACK PARIS FASHION	13047	United Kingdom	3	22	0

- **Schema used- Alphabetical**

i. Non-cat data – outlier treatment

- **Doing box plot to identify outlier**



- **Normalisation procedure done on Unit_price and Quantity using Mean max scalar.**

Min-max normalization, also known as feature scaling, is a technique used in data preprocessing to scale numerical features to a specific range, typically between 0 and 1. The formula for min-max normalization is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

3.2 Clustering

Objective 1-
We are using Unsupervised learning ie Clustering K-means

Obj2-
To find out the Appropriate no. of clusters

4 Characteristics

We found out that the appropriate no of clusters is 3

Row ID	Custom...	ï»¿Invoi...	Countr...	Quantity
cluster_0	15,376.338	733.575	0.053	0.5
cluster_1	15,196.774	2,368.005	0.228	0.5
cluster_2	15,670.202	3,921.963	0.22	0.5

Anova test for Non-Categorical variable

We did the anova test for the 2 non Categorical Variable ie Price And Quantity
And the Test is significant between groups

	Source	Sum of Squares	df	Mean Square	F	p-value
Quantity	Between Groups	0.0003	2	0.0001	3,269.4606	0.0
UnitPrice	Between Groups	0.0155	2	0.0077	3,023.7432	0.0

Categorical variable() – chi sq test of Independence: Kruskal walis test
For Customers the value is significant

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median ...	Mean R...	Median ...	Mean R...	Median ...
Row0	435,350.622	0.0	7,121,327.174	6,887,915	6,527,009.54	6,887,915	7,981,397.765	6,887,915

For Country also there is a significant difference among the clusters

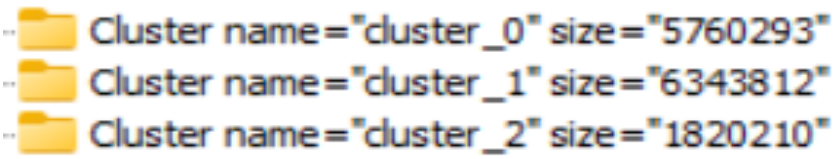
Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2
Row0	89,750.227	0.0	6,861,010.499	6,797,425.5	7,033,713.61	6,797,425.5	7,032,871.391	6,797,425.5

5. Observation

5.1 Marketing is segmented. The appropriate no. of segment is 3

5.2 Proof-

Characteristics of cluster 3



These cat variables(Country and customerID) can be distinguished distinguished individually because they have a significant p-value between groups.

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median ...	Mean R...	Median ...	Mean R...	Median ...
Row0	435,350.622	0.0	7,121,327.174	6,887,915	6,527,009.54	6,887,915	7,981,397.765	6,887,915

Row ID	H-Value	p-value	Mean Rank of Group cluster_0	Median Rank of Group cluster_0	Mean Rank of Group cluster_1	Median Rank of Group cluster_1	Mean Rank of Group cluster_2	Median Rank of Group cluster_2
Row0	89,750.227	0.0	6,861,010.499	6,797,425.5	7,033,713.61	6,797,425.5	7,032,871.391	6,797,425.5

Quantity and price (non categorical variables can be distinguished individually because they have a significant p-value between groups.

	Source	Sum of Squares	df	Mean Square	F	p-value
Quantity	Between Groups	0.0003	2	0.0001	3,269.4606	0.0
UnitPrice	Between Groups	0.0155	2	0.0077	3,023.7432	0.0

6. Managerial insights

Different groups have different average quantities and unit prices: There's a statistically significant difference in both quantity and unit price across the three groups (cluster 0, cluster 1, and cluster 2). This means the average quantity and average unit price are not the same for all groups.

6.1 Cluster (Heterogenous) identity

- One group might consistently buy a higher quantity at a lower unit price (bulk discount).
- Another group might buy a lower quantity at a higher unit price (convenience store purchase).
- There could be a mix of scenarios across the groups.