CUSTOMER SEGMENTATION IN E-COMMERCE

Students: Pavithra V(312217205065)
Pratheeksha V(312217205067)
Priyadharshini P(312217205068)

Project guide: Dr.S.Mohanavalli
Associate Professor

CUSTOMER SEGMENTATION

- Customer segmentation is used for dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

- This is done in various fields to meet the growing population and economy. Some of it’s applications are e-commerce ,telecom , online streaming services, etc.

- We have discussed it’s application in e-commerce briefly in our project.

RESEARCH MOTIVATION

- Customers base is increasing day by day exponentially.
- Lack of customer segmentation in e-commerce leads to an unhealthy relationship with customers.
- Key benefits of Customer segmentation
  - Improved Focus
  - Increased Competitiveness
  - Ability to Expand
  - Price Optimization

PROBLEM STATEMENT

To group large number of customers on an e-commerce website based on behavioural data, geographical data , purchase and transaction history using machine learning algorithms. This system would help in cutting down unnecessary costs and concentrate on most valuable customers.

METHODOLOGY

- We have used technique called RFM analysis and clustering algorithm, where:
- RFM Analysis:
  - R – Recency
  - F – Frequency
  - M – Monetary
- Text Mining
- Clustering Technique: K-means

IMPLEMENTATION

DATASET

- The data for our analysis is a transnational data which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based online retail company.
- Attributes:

-InvoiceNo
-StockCode
-Description
-Quantity
-InvoiceDate
-UnitPrice
-CustomerID
-Country

INITIAL DATASET

Total initial records in dataset: (541909,8).

Out[2]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047.0	United Kingdom

DATA PREPROCESSING

- Dropping records containing missing values:

Out[3]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
column type	object	object	object	int64	datetime64[ns]	float64	float64	object
null values (nb)	0	0	1454	0	0	0	135080	0
null values (%)	0	0	0.268311	0	0	0	24.9267	0

After dropping missing records:(406829, 8)

- Dropping duplicate entries(5225):

Out[5]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
565	536412	21448	12 DAISY PEGS IN WOOD BOX	2	2010-12-01 11:49:00	1.65	17920.0	United Kingdom
601	536412	21448	12 DAISY PEGS IN WOOD BOX	2	2010-12-01 11:49:00	1.65	17920.0	United Kingdom
604	536412	21448	12 DAISY PEGS IN WOOD BOX	2	2010-12-01 11:49:00	1.65	17920.0	United Kingdom

After dropping duplicate entry records: (401604, 8)

EXPLORING INDIVIDUAL FEATURES

- **Country**: There are a 37 unique countries in the dataset.

  90%- UK. So we consider only the transactions of UK.

- Number of Unique Customers and products and Transactions:

	products	transactions	customers
quantity	3661	19857	3950

- **In** Out[10]:

	CustomerID	InvoiceNo	Number of products
0	12346.0	541431	1
1	12346.0	C541433	1
2	12747.0	537215	7
3	12747.0	538537	8

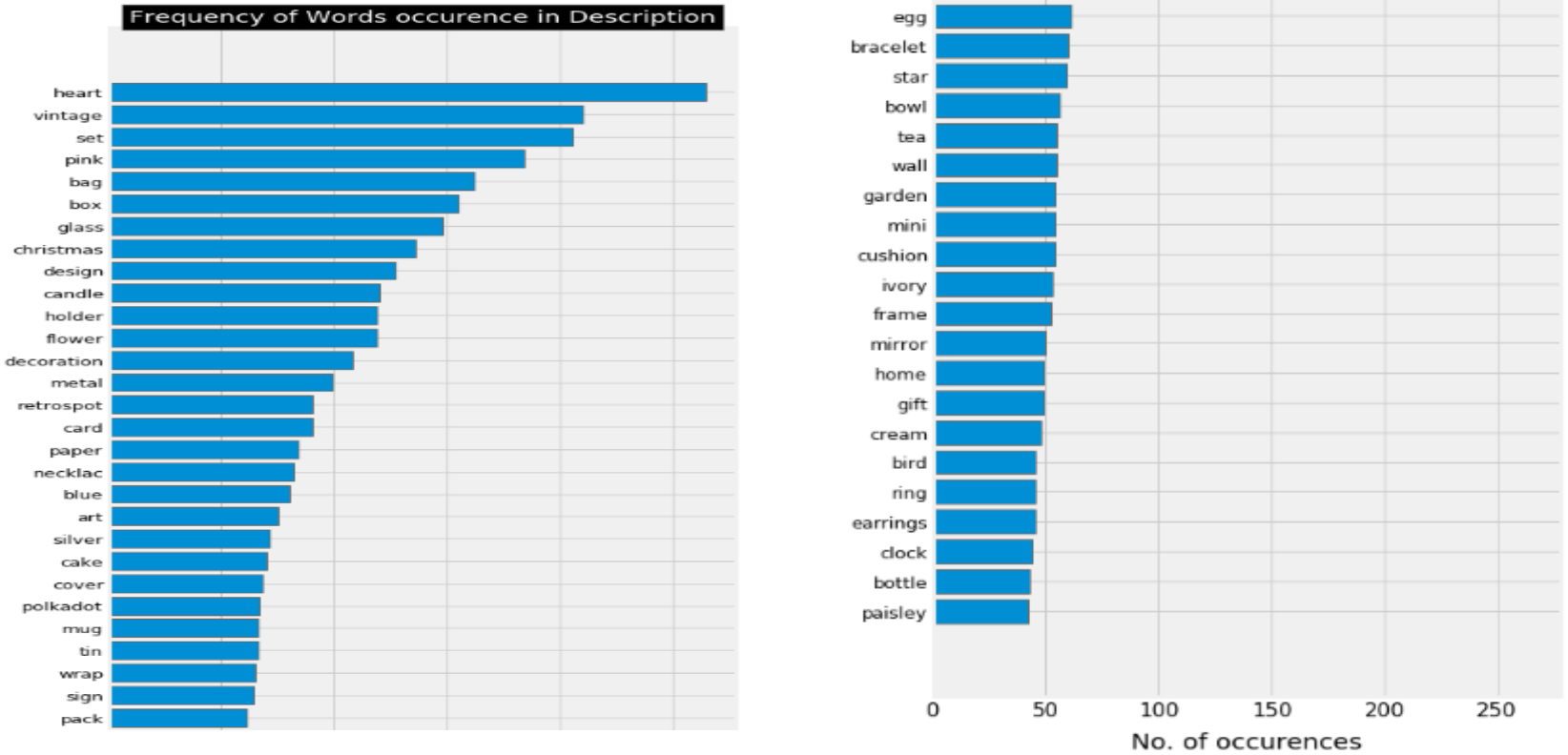
Here C refers to cancelled orders.

EXPLORING INDIVIDUAL FEATURES

- **Description:**

No. of keywords in variable 'Description': 1480

Listing the top-50 words appearing in most of the descriptions:



EXPLORING INDIVIDUAL FEATURES

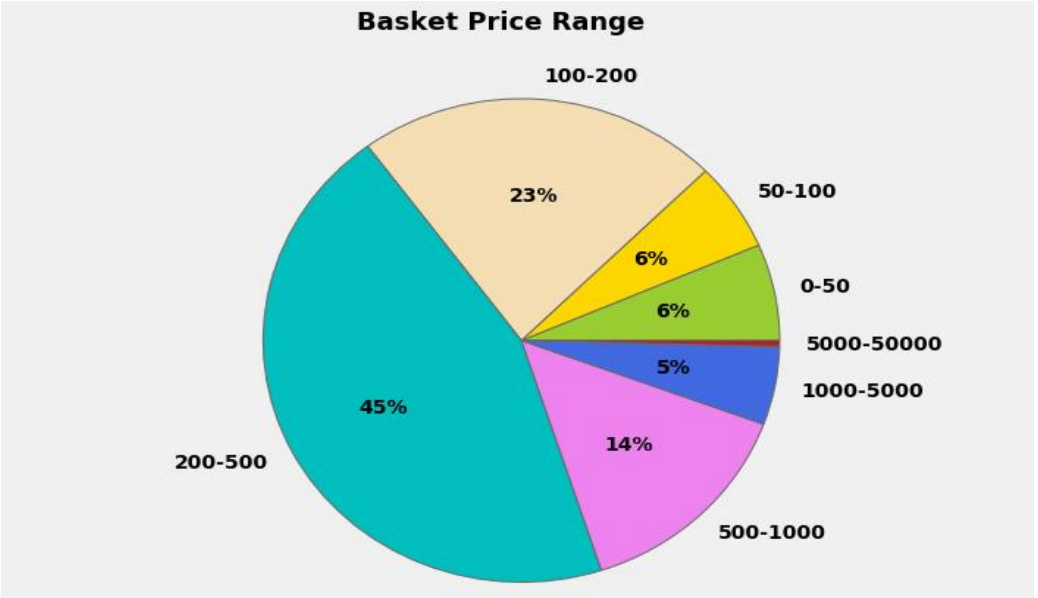
- **Basket Price:**

Formula: (Quantity purchased - Quantity canceled) * Unit Price

(16538, 4)

Out[22]:

	CustomerID	InvoiceNo	Basket_Price	InvoiceDate
10	12747.0	577104	312.73	2011-11-17 17:13:00
9	12747.0	569397	675.38	2011-10-04 08:26:00
8	12747.0	563949	301.70	2011-08-22 10:38:00
7	12747.0	558265	376.30	2011-06-28 10:06:00



TEXT MINING

- **Generating Categories for Products**
  - One-hot data encoding technique is used here.

Out[24]:

	heart	vintage	set	bag	box	glass	christmas	design	candle	holder	...	tidy	plant	house	square	diner	lace	pan
0	1	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0

- **Creating clusters of products**

For n_clusters = 3 The average silhouette_score is : 0.09689439106249706
For n_clusters = 4 The average silhouette_score is : 0.12752536288729166
For n_clusters = 5 The average silhouette_score is : 0.1240144559057432
For n_clusters = 6 The average silhouette_score is : 0.15102728685983333
For n_clusters = 7 The average silhouette_score is : 0.130999063434301
For n_clusters = 8 The average silhouette_score is : 0.15470465876902081
For n_clusters = 9 The average silhouette_score is : 0.12484232443063882

For n_clusters = 5 The average silhouette_score is : 0.1452148389646187

TEXT MINING

- **Word Cloud**

TEXT MINING

- Different products were grouped in five clusters.

```
Out[35]:
  InvoiceNo                          Description  categ_product
0    536365   WHITE HANGING HEART T-LIGHT HOLDER              4
1    536365                WHITE METAL LANTERN              2
2    536365     CREAM CUPID HEARTS COAT HANGER              2
3    536365  KNITTED UNION FLAG HOT WATER BOTTLE              2
4    536365     RED WOOLLY HOTTIE WHITE HEART.              2
5    536365       SET 7 BABUSHKA NESTING BOXES              0
6    536365   GLASS STAR FROSTED T-LIGHT HOLDER              2
7    536366              HAND WARMER UNION JACK              4
8    536366           HAND WARMER RED POLKA DOT              1
9    536367      ASSORTED COLOUR BIRD ORNAMENT               1
```

- The amount spent in each product category:

InvoiceNo	Description	categ_product	categ_0	categ_1	categ_2	categ_3	categ_4
536365	WHITE HANGING HEART T-LIGHT HOLDER	4	0	0	0	0	15.3
536365	WHITE METAL LANTERN	2	0	0	20.34	0	0
536365	CREAM CUPID HEARTS COAT HANGER	2	0	0	22	0	0
536365	KNITTED UNION FLAG HOT WATER BOTTLE	2	0	0	20.34	0	0
536365	RED WOOLLY HOTTIE WHITE HEART.	2	0	0	20.34	0	0

MODELLING

- **K-means Clustering:**

  As our feature variables are numerical and our goal is unsupervised to find out some sort of structure/grouping in the customers, we used k-means clustering.

- **Feature selection:**

  We are applying k-means clustering

  i. 1st using Recency, Frequency, Monetary.

  ii. 2nd using the product categories.

CLUSTERING

- **Clustering based on Recency , Frequency , Monetary**

	Recency	Frequency	Monetary
count	3912.000000	3912.000000	3912.000000
mean	91.876534	4.227505	1747.337910
std	99.695909	7.134453	6730.162064
min	0.000000	1.000000	2.900000
25%	17.000000	1.000000	292.007500
50%	50.000000	2.000000	635.070000
75%	143.000000	5.000000	1536.315000
max	373.000000	205.000000	259657.300000

- **Square Root Transformation:**

The RFM feature selected for our analysis has different scales:

Recency : 0 – 373 Frequency : 1 – 205 Monetary : 2.9 : 259,657

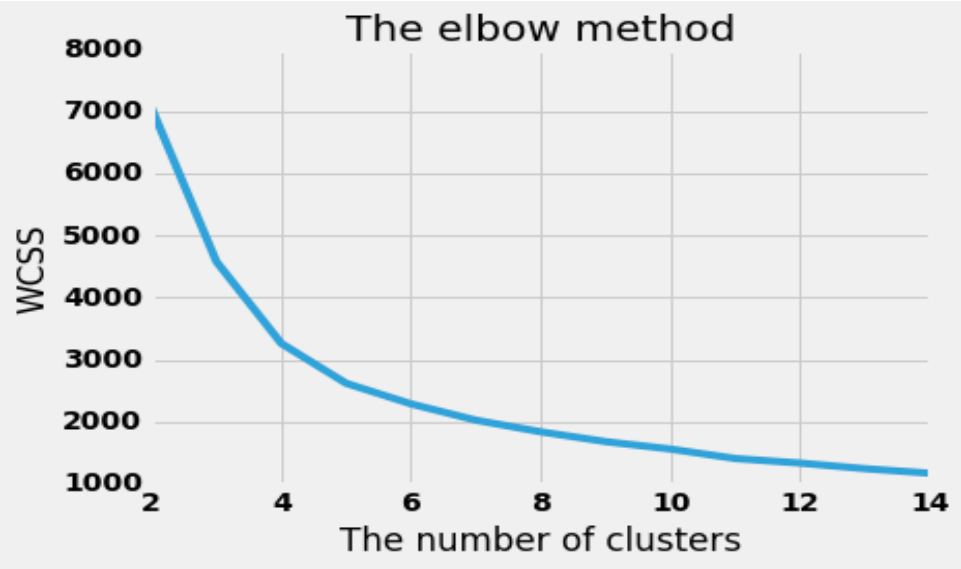
Since, we are using k-means clustering which basically finds the eucledian distance between the data points and the cluster mean, its important to scale or transform the data before analysis.

Variables mean values:

[8.15011469 1.79362304 32.07254545]

CLUSTERING

- **ELBOW METHOD**



SILHOUETTE SCORES

```
For n_clusters = 3 The average silhouette_score is : 0.6005040987032123
For n_clusters = 4 The average silhouette_score is : 0.5297904992840151
For n_clusters = 5 The average silhouette_score is : 0.523596172966738
For n_clusters = 6 The average silhouette_score is : 0.45166919105390346
For n_clusters = 7 The average silhouette_score is : 0.4164862040731145
For n_clusters = 8 The average silhouette_score is : 0.3967995933097067
For n_clusters = 9 The average silhouette_score is : 0.4064028398846738
```

N_CLUSTERS=5

CLUSTERING

- **No of customers in each clusters:**

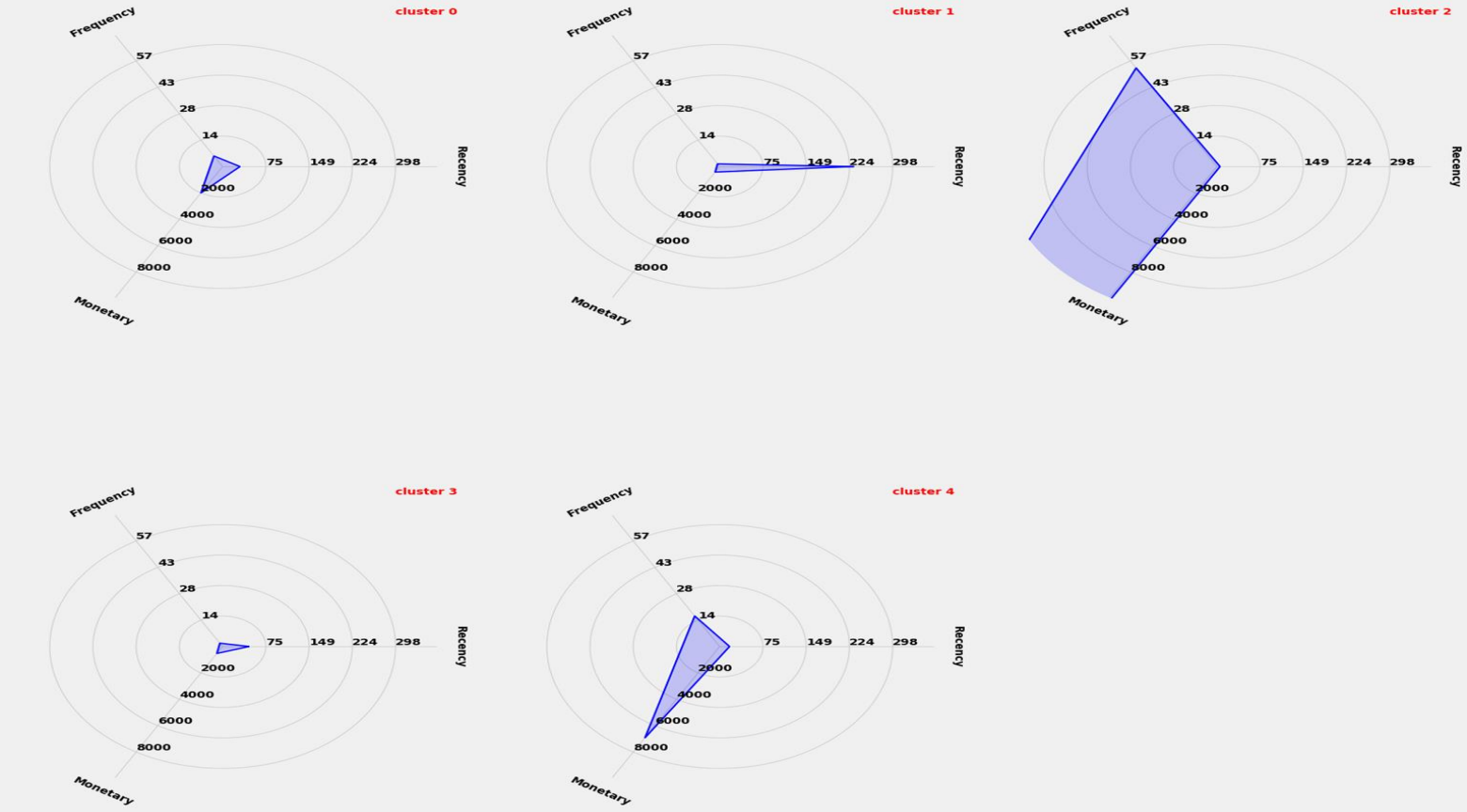
	3	1	0	4	2
nb. of customers	1499	1101	1001	281	30

- **Final dataset for RFM clustering:**

total number of customers: 3912

CustomerID	Recency	Frequency	Monetary	First Purchase	size	cluster
15593.87284	231.8065395	1.544959128	422.3600917	262.9981835	1101	1
15594.43229	46.03802535	1.893929286	526.9339179	137.5757171	1499	3
15482.16983	30.18181818	5.768231768	1996.501329	271.4655345	1001	0
15583.5694	17.113879	16.48398577	6915.607117	337.3131673	281	4
15306.66667	5.666666667	53.06666667	54630.33567	353.7666667	30	2

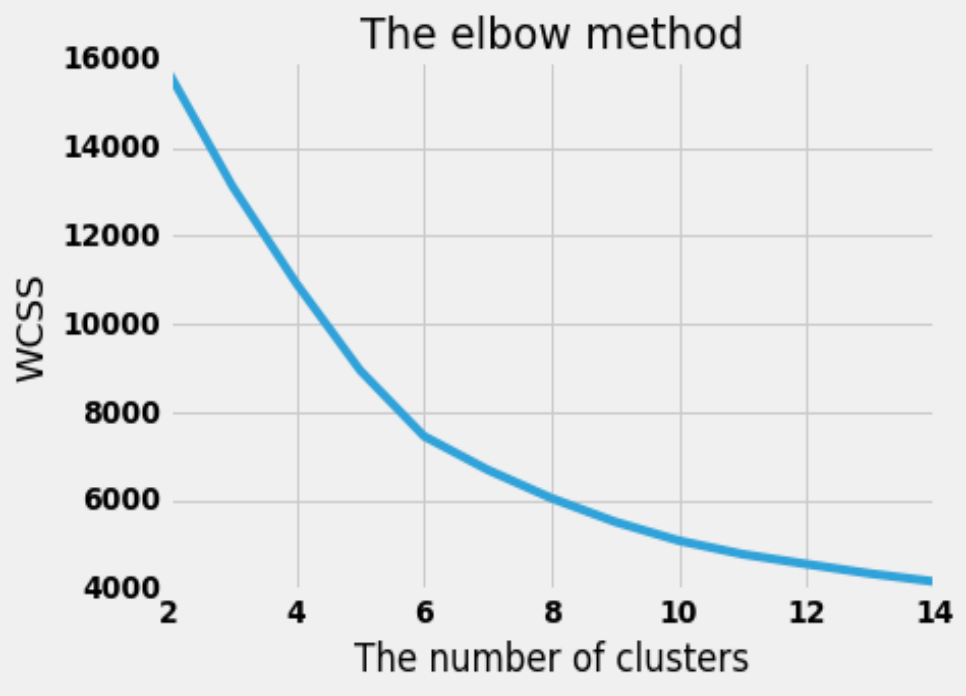
CLUSTERING



RADAR CHART (RFM)

CLUSTERING BASED ON CATEGORY

- **ELBOW METHOD**



SILHOUETTE SCORES

n_cluster=6
silhouette score:0.269

CLUSTERING BASED ON CATEGORY

- **No of customers in each clusters**

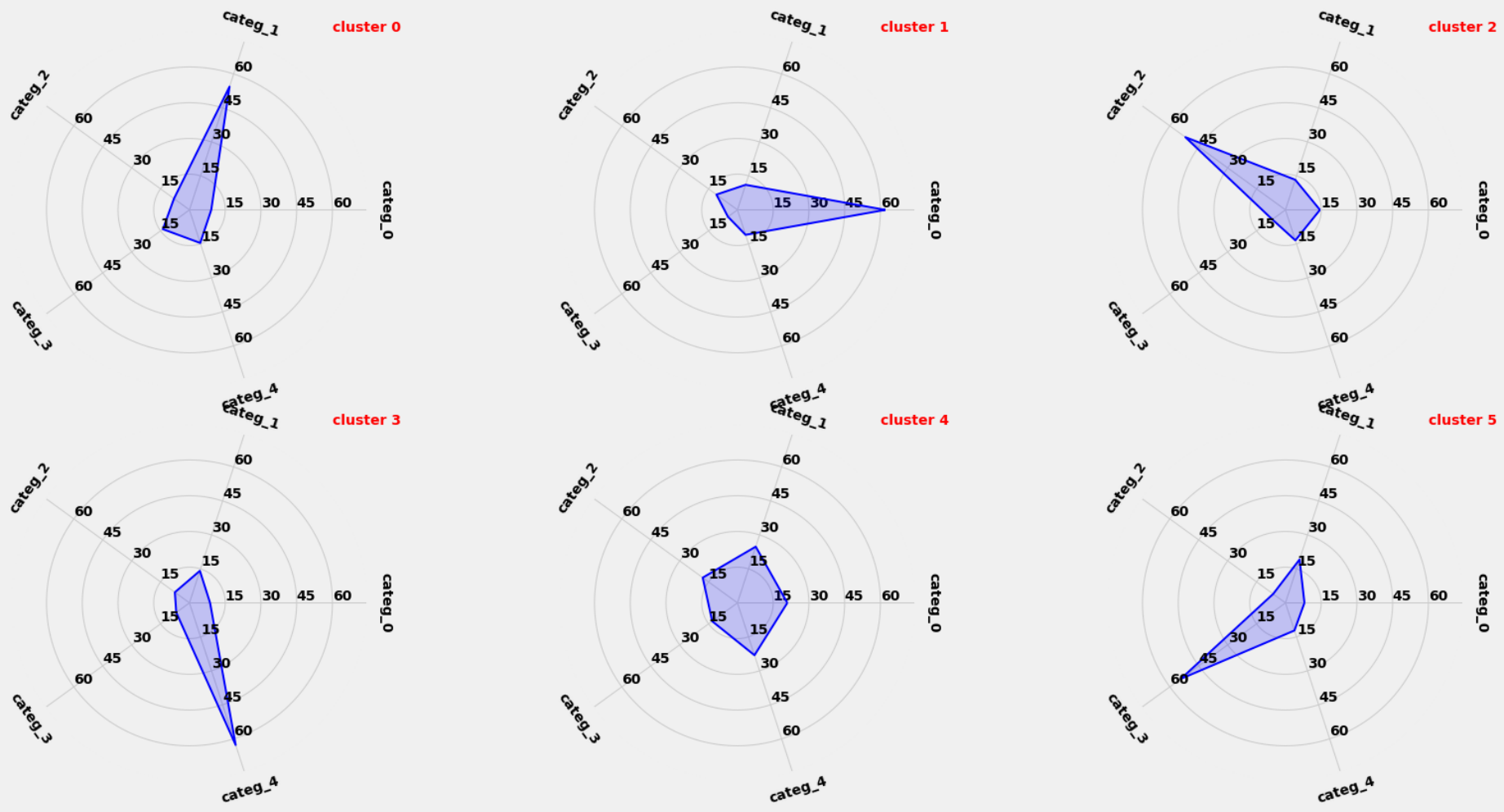
4 0 1 2 3 5

nb. of customers: 1873 584 519 319 317 300

- **Final dataset for clustering wrt category:**

Monetary	categ_0	categ_1	categ_2	categ_3	categ_4	size	cluster
736.6413333	7.97880423	19.1590191	6.387146791	54.27057143	12.20445845	300	5
1060.556149	9.14547082	54.40220734	8.043448735	13.78727268	14.62476855	584	0
1179.740408	14.4872365	13.29084741	51.94345288	6.841440148	13.46145262	319	2
1911.71006	62.0513239	11.06749929	10.92179006	4.943189064	11.0272634	519	1
1918.332145	8.65434907	14.15137911	7.583325845	6.868098667	62.7428473	317	3
2145.543359	20.9309811	24.82161663	18.02497005	13.11884334	23.11008377	1873	4

CLUSTERING BASED ON CATEGORY



CLUSTER ANALYSIS

- **Cluster 0:**
  - ✔ Average Monetary score
  - ✔ Have high tendency of purchasing products from category_1
- **Cluster 1:**
  - ✔ Good Monetary score
  - ✔ Have high tendency of purchasing products from category_0
- **Cluster 2:**
  - ✔ Average Monetary score
  - ✔ Have high tendency of purchasing products from category_2

CLUSTER ANALYSIS

- **Cluster 3:**
  - ✔ Good Monetary score
  - ✔ Have high tendency of purchasing products from category_4
- **Cluster 4:**
  - ✔ Highest Monetary score
  - ✔ Hare general buyers, can be targeted for multiple products
- **Cluster 5:**
  - ✔ Below Average Monetary score
  - ✔ Have high tendency of purchasing products from category_3

CONCLUSION

- Customer segmentation is a way to improve communication with the customer.
- It is an essential process to get potential customers and increase profits.
- This can be extended in various sectors.

THANK YOU