CAPSTONE PROJECT ON UNSUPERVISED MACHINE LEARNING

Customer Segmentation

KARAN K. KARLE
PCAI, FEB 2021 BATCH

Contents

Introduction

- **Machine learning** is a subfield of Data Science,
- Machine learning is an approach to data analysis that involves building and adapting models, which allow programs to "learn" through experience
- Machine Learning has two types i.e. Supervised and Unsupervised Learning,
- **RFM** stands for Recency - Frequency - Monetary Value. As the methodology, we need to calculate Recency, Frequency and Monetary Value and apply unsupervised machine learning to identify different groups (clusters) for each.
- **K-means clustering** is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*.
- **Association rule** mining is a technique to identify underlying relations between different items.
- **Apriori** algorithms is one of the algorithm which has been developed to implement association rule mining.

Objectives

- To create Clusters Based on the characteristics of users to categorize users based on their transactions And Association Rules for Products which are “Frequently bought together”.

- Benefits:-
  - Understand the customers relations with company
  - Avoid stockouts
  - Identify opportunities to boost revenue
  - Make better merchandising decisions
  - Understand and plan marketing efforts

Python Packages

- Pandas, Numpy, Seaborn, Matplotlib
- StandardScaler
- Algorithm
  - Clustering
    - K-Means
  - Association Rule Mining
    - Apriori
- All other needed dependencies

The Data

The Data Attributes

I. InvoiceNo
II. StockCode
III. Description
IV. Quantity
V. InvoiceDate
VI. UnitPrice
VII. Country

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

“Amount” column is created by multiplying Quantity and Unit Price

:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Amount
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	20.34

Exploratory Data Analysis

Data Cleaning:

- Duplicate values are removed
- Missing values are imputed
- Canceled Invoices are removed

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France

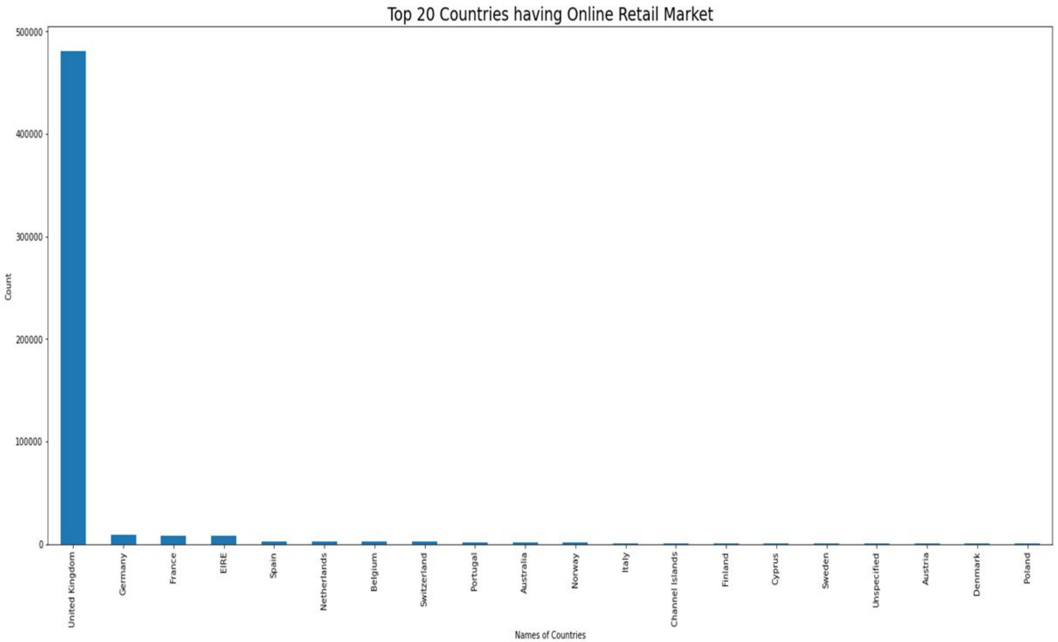
525923 rows × 8 columns

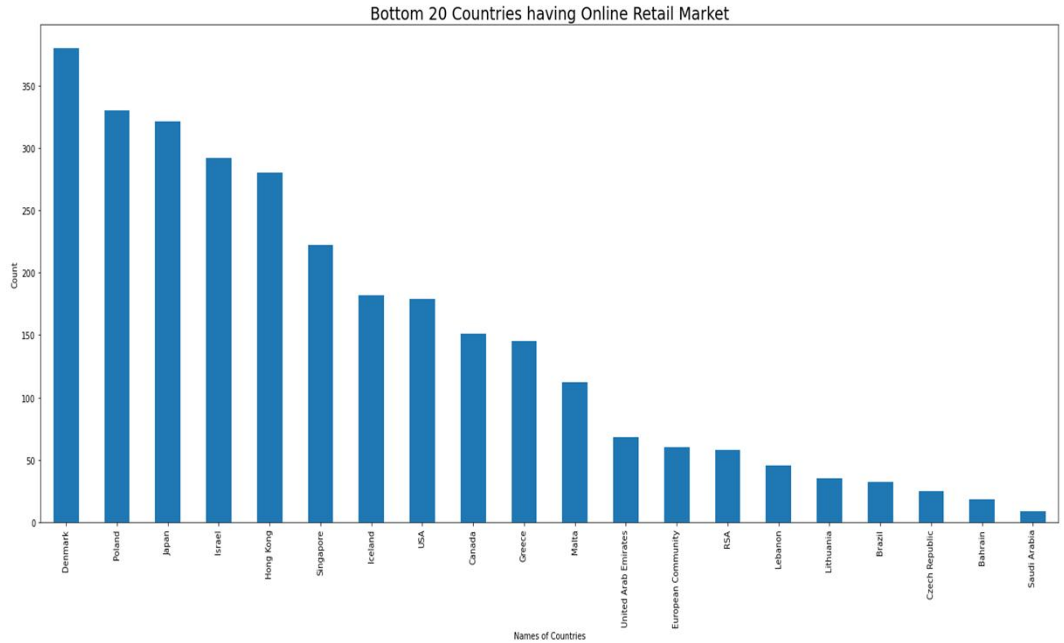
There are 38 countries in this data, UK has highest number of Customers

United Kingdom	481012
Germany	9027
France	8393
EIRE	7883
Spain	2480
Netherlands	2363
Belgium	2031
Switzerland	1959
Portugal	1492
Australia	1184
Norway	1072
Italy	758
Channel Islands	747
Finland	685
Cyprus	603
Sweden	450
Unspecified	442
Austria	398
Denmark	380
Poland	330
Japan	321
Israel	292
Hong Kong	280
Singapore	222
Iceland	182
USA	179
Canada	151
Greece	145
Malta	112
United Arab Emirates	68
European Community	60
RSA	58
Lebanon	45
Lithuania	35
Brazil	32
Czech Republic	25
Bahrain	18
Saudi Arabia	9

Data Visualization

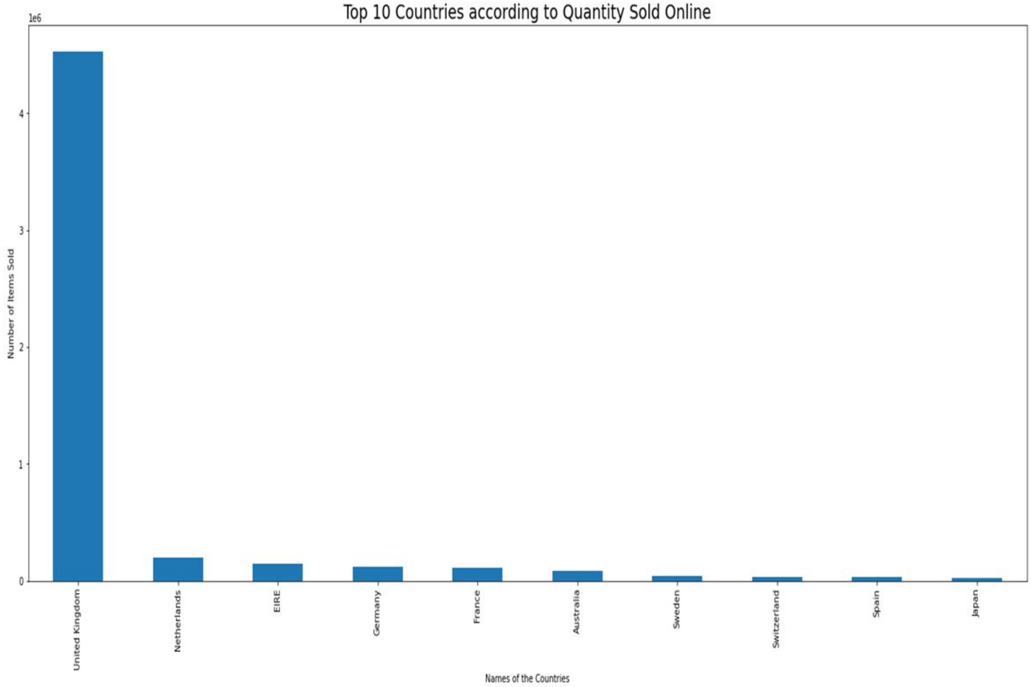
Countries having Online Retail Market



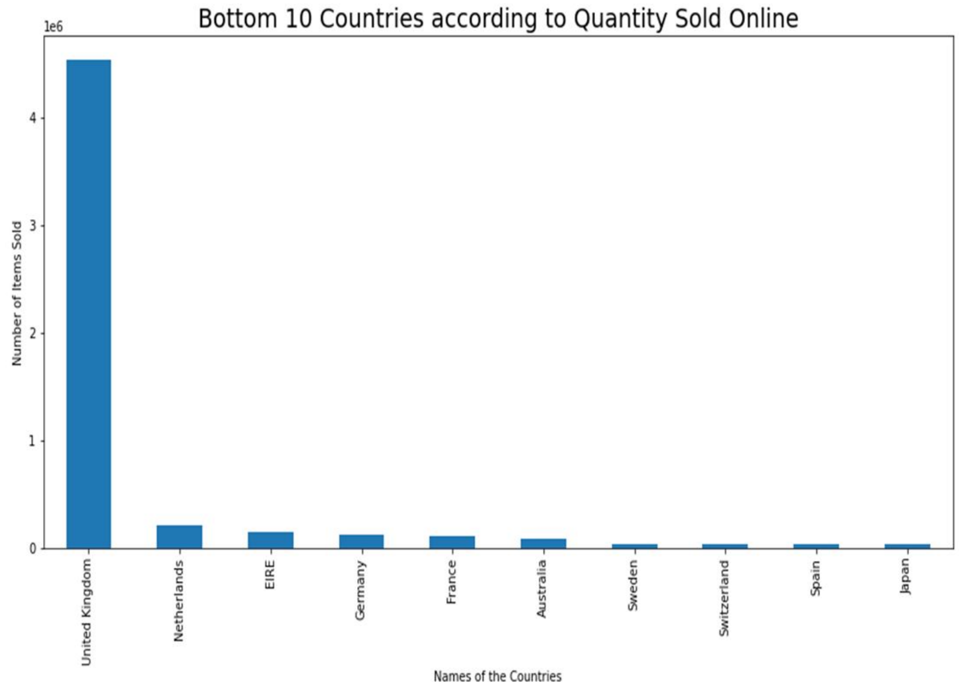


Countries according to Quantity Sold Online

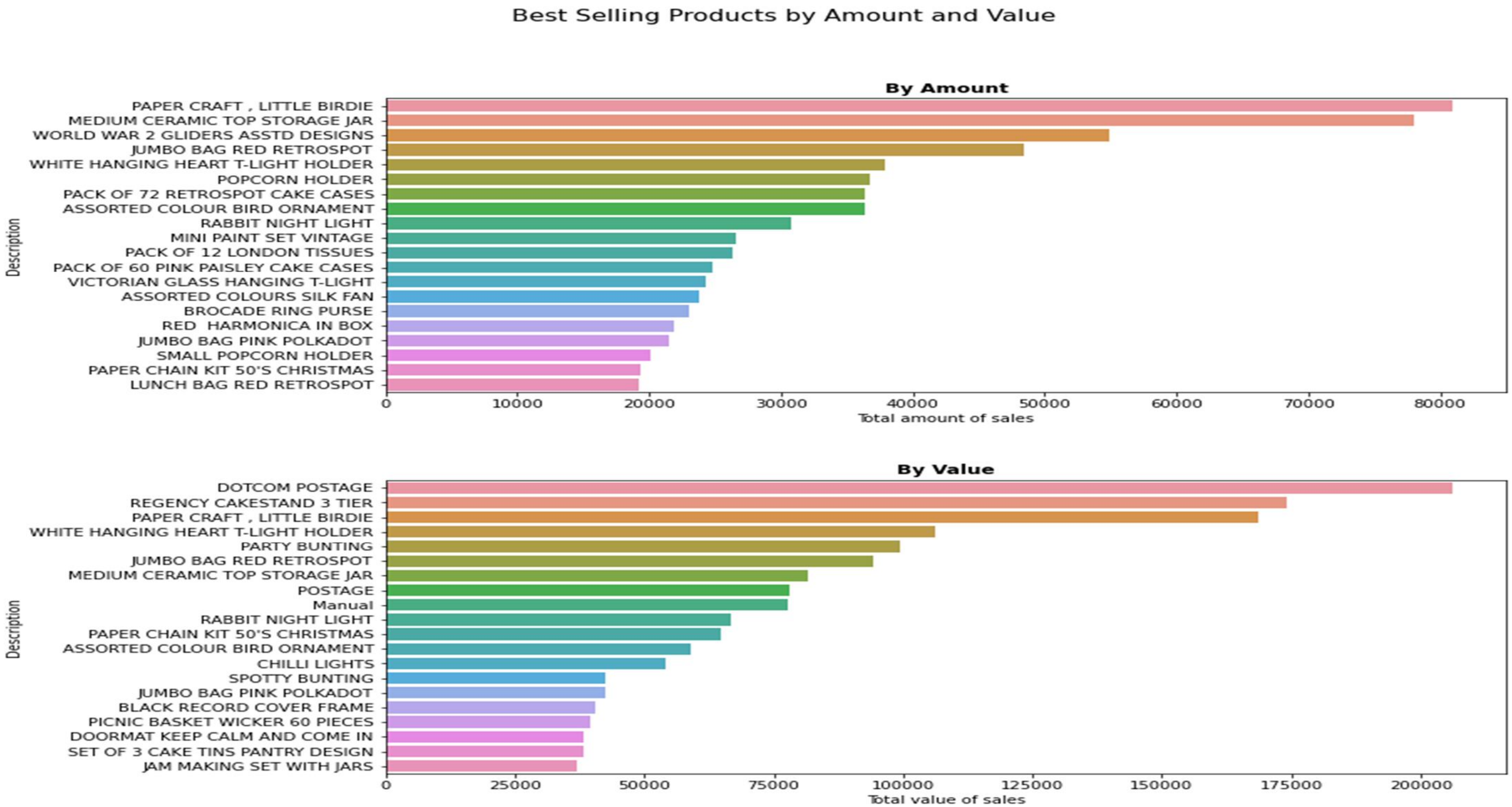
Top 10 countries



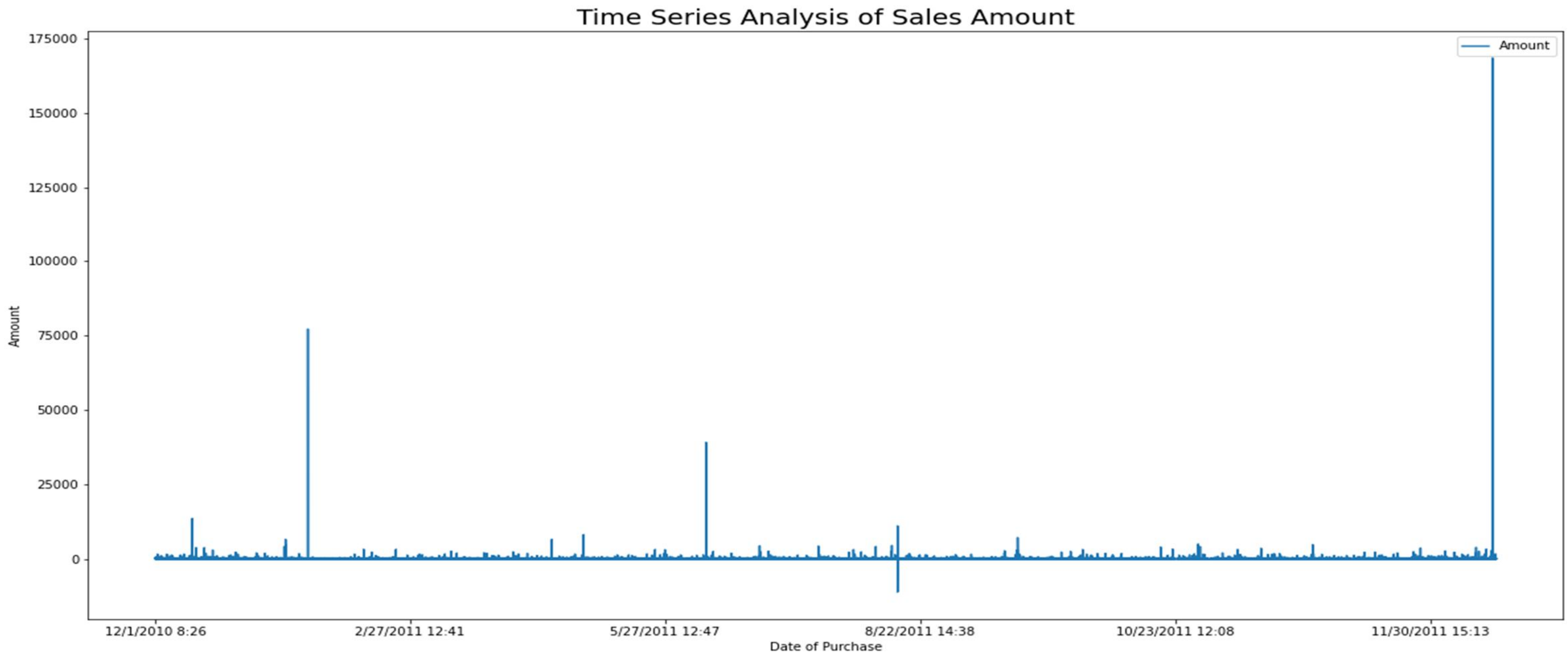
Bottom 10 countries



Best Selling Products



Time Series analysis for Sales Amount



Clustering for Customer segmentation

❖ **RMF Analysis:**

Customer segmentation by 3 important features:

•Recency — Number of days since the last purchase

•Frequency — Number of transactions made over a given period

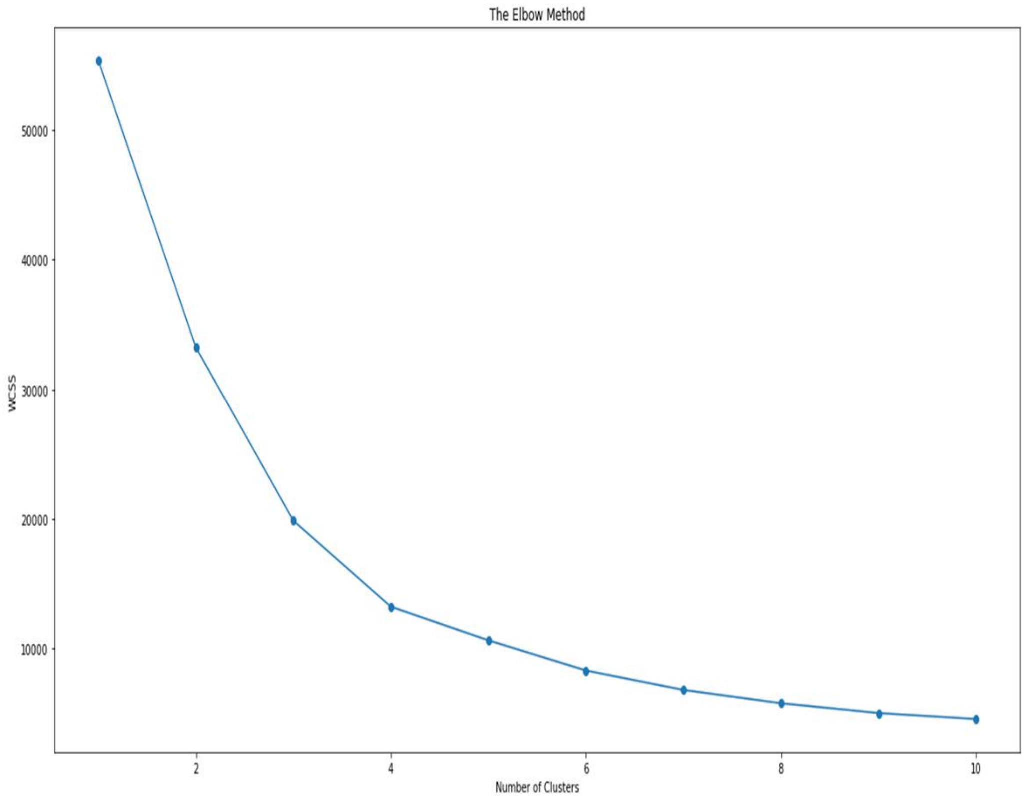
•Monetary — Amount spent over a given period of time

```
RFM.head()
```

	Recency	Monetary	Frequency
0	326	77183.6	1
1	2	4310.0	182
2	40	4310.0	182
3	130	4310.0	182
4	183	4310.0	182

Choosing the number of clusters

Elbow Curve method:- WCSS -> Within Clusters Sum of Squares



Elbow Curve method :- Optimal Value of K is 4

Silhouette analysis

```
For n_clusters=2, the silhouette score is 0.751827695095344
For n_clusters=3, the silhouette score is 0.534149404203003
For n_clusters=4, the silhouette score is 0.5426450015703492
For n_clusters=5, the silhouette score is 0.5610983019710827
For n_clusters=6, the silhouette score is 0.4935289444936344
For n_clusters=7, the silhouette score is 0.49584088405519006
For n_clusters=8, the silhouette score is 0.5050579426293625
For n_clusters=9, the silhouette score is 0.4420910212585138
For n_clusters=10, the silhouette score is 0.44614657355320564
```

Elbow and Silhouette analysis

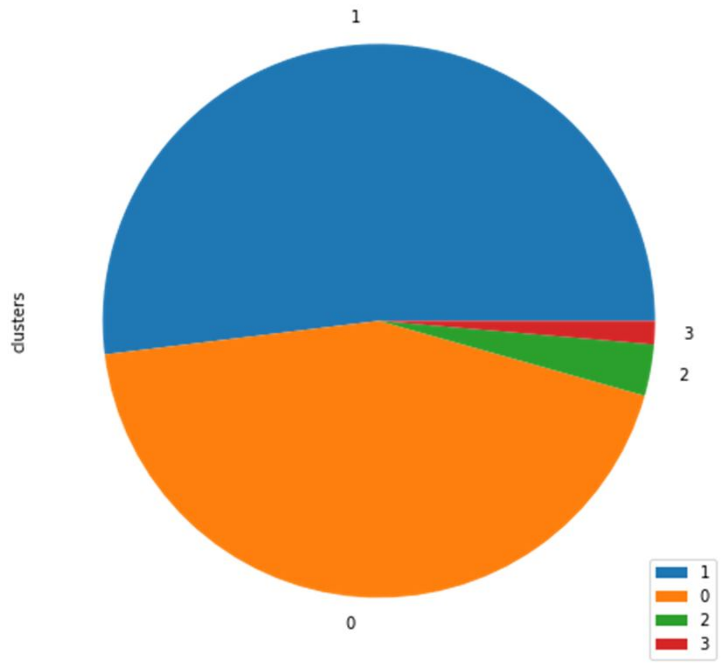
- From the elbow curve we observe the elbow at cluster 3 and cluster 4.
- Also from Silhouette analysis we see the value is better when number of cluster will be 4 rather than 3.
- So we now categorize the data into 4 clusters and check their RFM values and its distribution.

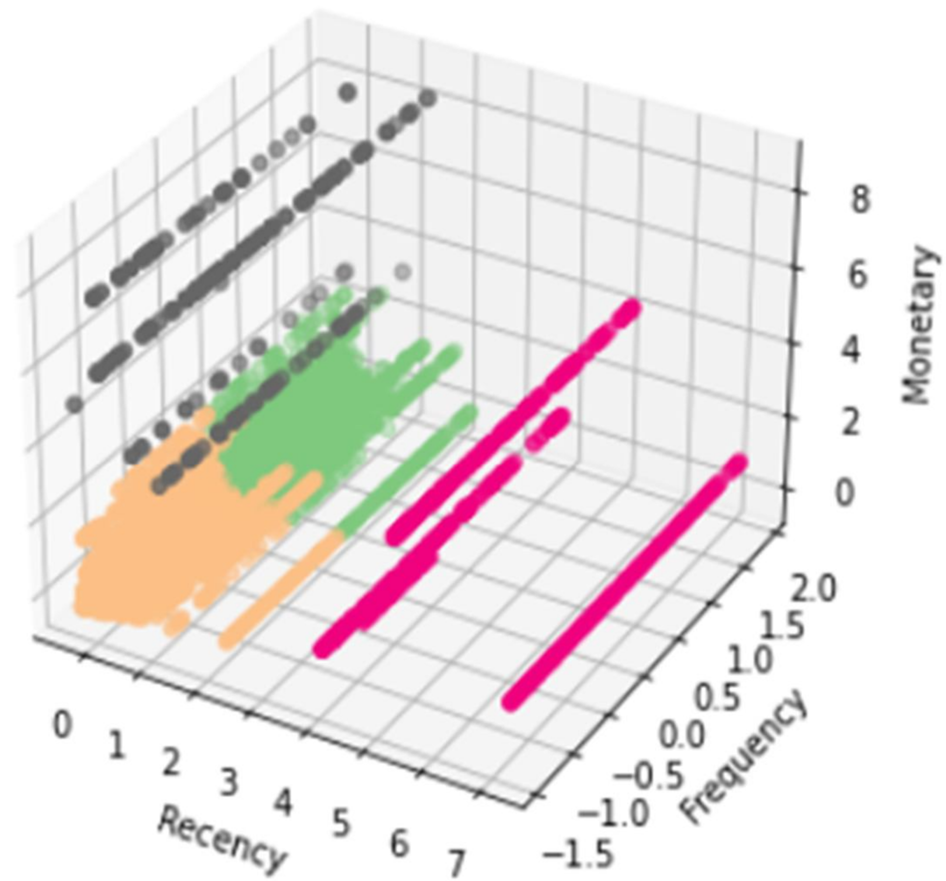
Training the K-Means Clustering Model

```
RFM_Scaled.head()
```

	Frequency	Monetary	Recency	clusters
0	1.458653	2.124858	-0.393769	0
1	-1.415066	-0.211840	-0.214893	1
2	-1.078025	-0.211840	-0.214893	1
3	-0.279769	-0.211840	-0.214893	1
4	0.190314	-0.211840	-0.214893	0

Visualizing the K-Means Clustering Model





Association Rule Mining by using Apriori Algorithm

Basket : Created Market basket Data by grouping “Description”,”InvoiceNo” and "Quantity“ column from Given Data

```
basket.head()
```

Description	*Boombox Ipod Classic	*USB Office Mirror Ball	10 COLOUR SPACEBOY PEN	12 COLOURED PARTY BALLOONS	12 DAISY PEGS IN WOOD BOX	12 EGG HOUSE PAINTED WOOD	12 HANGING EGGS HAND PAINTED	12 IVORY ROSE PEG PLACE SETTINGS	12 MESSAGE CARDS WITH ENVELOPES	12 PENCIL SMALL TUBE WOODLAND	...	wet boxes	wet damaged	wet pallet	we rusty
InvoiceNo															
536365	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
536366	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
536367	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
536368	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
536369	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0

5 rows × 4183 columns

Association rules

List of Frequently Bought Together Items

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.051027	0.047580	0.031072	0.608944	12.798384	0.028645	2.435508
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.047580	0.051027	0.031072	0.653061	12.798384	0.028645	2.735276
2	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.037190	0.049279	0.030733	0.826371	16.769220	0.028900	5.475581
3	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.049279	0.037190	0.030733	0.623645	16.769220	0.028900	2.558252
4	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.049279	0.051755	0.037287	0.756650	14.619817	0.034737	3.896634
5	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.051755	0.049279	0.037287	0.720450	14.619817	0.034737	3.400901
6	(JUMBO BAG PINK POLKADOT)	(JUMBO BAG RED RETROSPOT)	0.059135	0.101568	0.040054	0.677340	6.668819	0.034048	2.784453
7	(JUMBO BAG RED RETROSPOT)	(JUMBO BAG PINK POLKADOT)	0.101568	0.059135	0.040054	0.394359	6.668819	0.034048	1.553504
8	(JUMBO BAG RED RETROSPOT)	(JUMBO SHOPPER VINTAGE RED PAISLEY)	0.101568	0.057047	0.033015	0.325048	5.697880	0.027220	1.397066
9	(JUMBO SHOPPER VINTAGE RED PAISLEY)	(JUMBO BAG RED RETROSPOT)	0.057047	0.101568	0.033015	0.578723	5.697880	0.027220	2.132641
10	(JUMBO STORAGE BAG SUKI)	(JUMBO BAG RED RETROSPOT)	0.057484	0.101568	0.035151	0.611486	6.020453	0.029312	2.312485
11	(JUMBO BAG RED RETROSPOT)	(JUMBO STORAGE BAG SUKI)	0.101568	0.057484	0.035151	0.346080	6.020453	0.029312	1.441333
12	(LUNCH BAG RED RETROSPOT)	(LUNCH BAG BLACK SKULL.)	0.075933	0.061805	0.031121	0.409847	6.631272	0.026428	1.589747
13	(LUNCH BAG BLACK SKULL.)	(LUNCH BAG RED RETROSPOT)	0.061805	0.075933	0.031121	0.503535	6.631272	0.026428	1.861292

we can see the items that were most often bought together in the above table:

- antecedent=purchased,
- consequents= going purchase
- confidence= chances of buying toghether

Conclusion

- **Cluster** (for customer segmentation and to find customer groups with similar behaviors for further analysis and business strategy planning)
  - For clustering we have used RFM Analysis and then by Elbow curve method we can see 4 is optimal value, for validation we used Silhouette analysis score which is confirming 4 cluster can be made
- **Association Rules** (to see which set of products were Frequently Bought Together)
  - We used Apriori Algorithm and then association rules for products which were Frequently Bought Together

Future Improvements

- Scope for Improvements

  - By doing more Feature Engineering , we can do it without RFM too.
  - With Gathering more data.
  - Different ways of clustering like DBScan algorithm etc.

Thank You