Capstone Project – Online Retail Segmentation

Contents

Problem Statement

An online retail store is trying to identify Customer Segmentation or Market Segmentation. This will enable firm to understand the various customer purchase patterns and accordingly plan future strategies.

Project Objective

- Find useful insights about the customer purchasing history
- Segment the customers based on their purchasing behavior.

Data Description

This Online Retail data set contains all the transactions records. It contains 541909 reco rds and 8 features.

Below is brief insight to each feature:

- InvoiceNo: Invoice number. Nominal. A 6-digit integral number uniquely assigned to each transaction.
- StockCode: Product (item) code. Nominal. A 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice date and time. Numeric. The day and time when a transaction was generated.
- UnitPrice: Unit price. Numeric. Product price per unit.
- CustomerID: Customer number. Nominal. A 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal. The name of the country where a customer resides.

Snippet:

```
df.head()
```

	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

Data Pre-processing Steps and Inspiration

- Rearranged columns to enable better readability.

```
df.head()
```

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

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   CustomerID   406829 non-null  float64
 1   InvoiceNo    541909 non-null  object
 2   StockCode    541909 non-null  object
 3   Description  540455 non-null  object
 4   Quantity     541909 non-null  int64
 5   UnitPrice    541909 non-null  float64
 6   InvoiceDate  541909 non-null  object
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

Here, Invoice date is object type. We will convert this into datetime for calculating all the values. Further, we are addressing negative value of Quantity and Unitprice.

Null Values were addressed by dropping records containing null data as dataset is big enough. Post cleanup 530104 records are available in dataset.

Invoice data has total 18499 unique invoices.

```
ref.InvoiceDate.describe()
```

```
C:\Users\Admin\AppData\Local\Temp,
han numeric in `.describe` is depr
to silence this warning and adopt
  ref.InvoiceDate.describe()
```

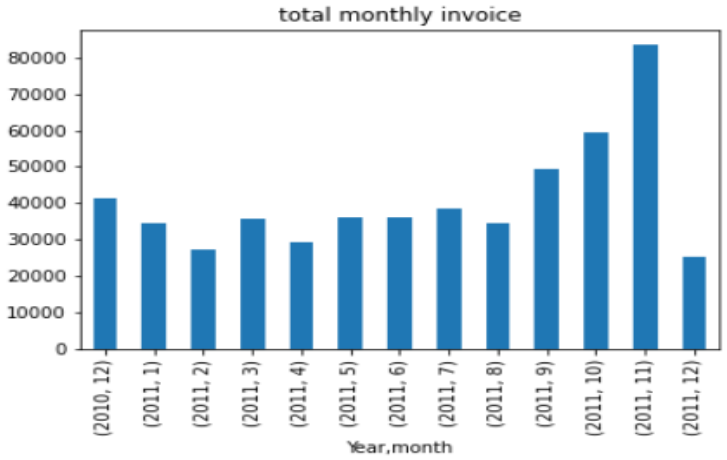
```
count                  530104
unique                  18499
top       2011-10-31 14:41:00
freq                     1114
first     2010-12-01 08:26:00
last      2011-12-09 12:50:00
Name: InvoiceDate, dtype: object
```

Analysis on “country’ feature shows 38 unique countries.

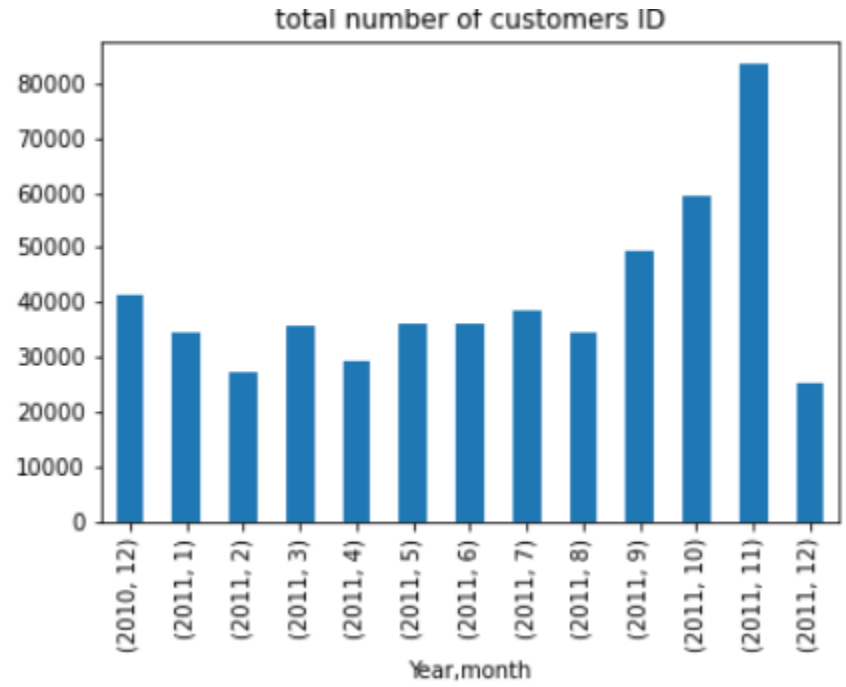
```
ref.Country.nunique()
```

```
38
```

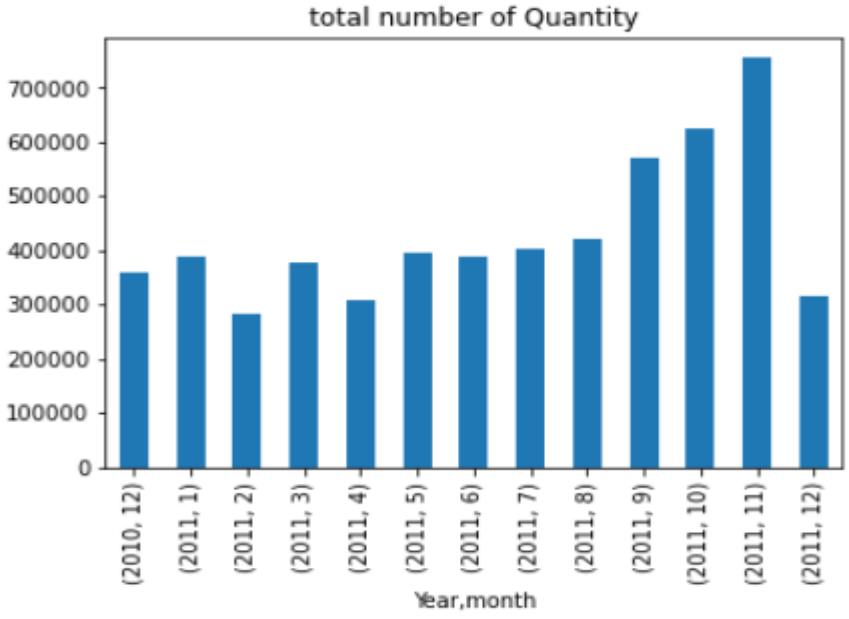
Below depicts Monetary Value (Total Price) per month. Here we are getting our monetary value by grouping customer with their customer id and total no. of sales.



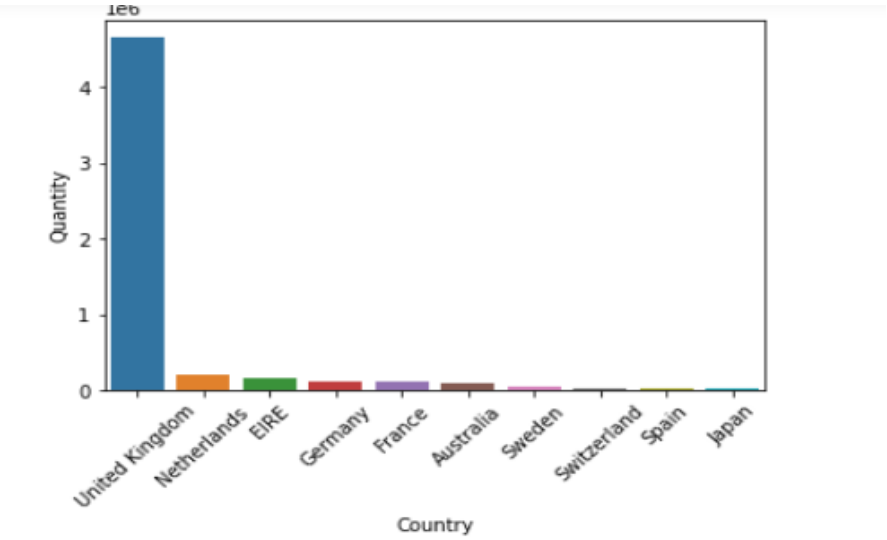
Below bar chart depicts total number of customers and their distribution.



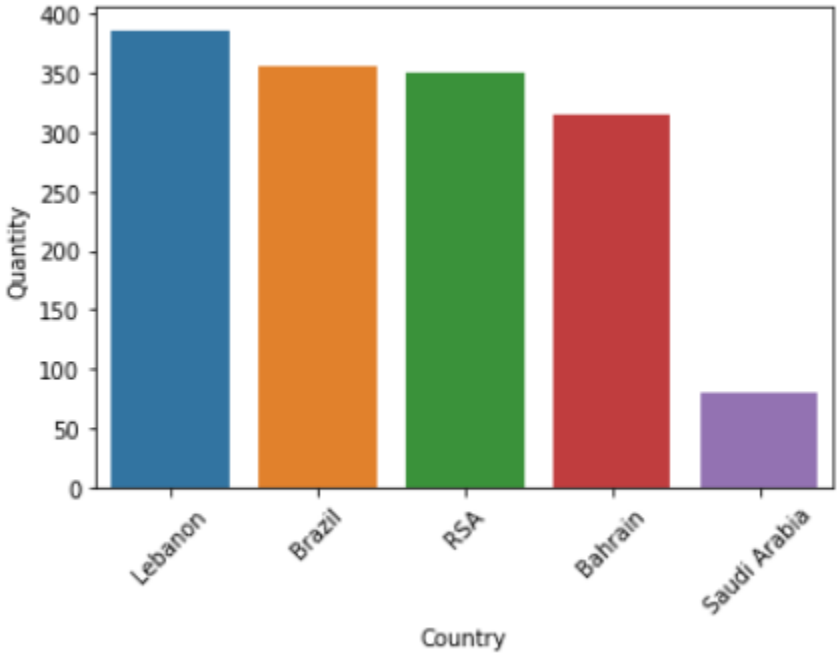
Total quantity sold per month can be seen thru below chart



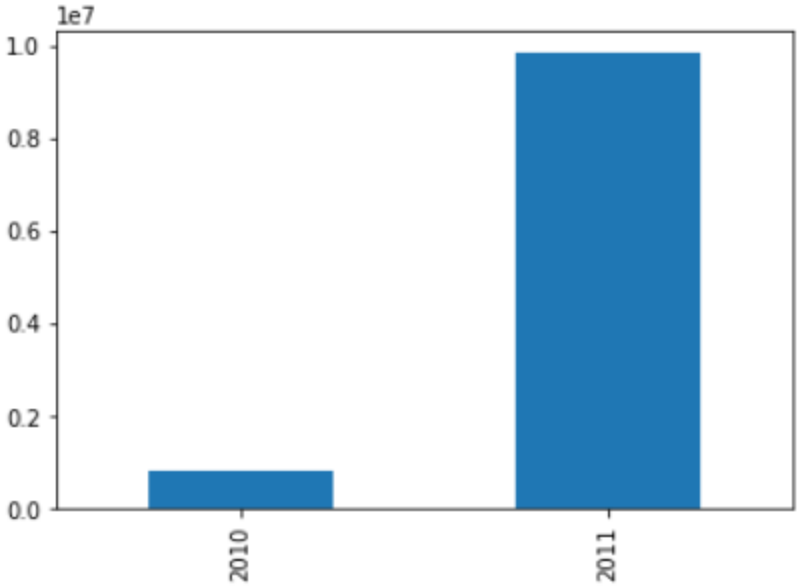
Top 10 countries based on Total quantity of Sales



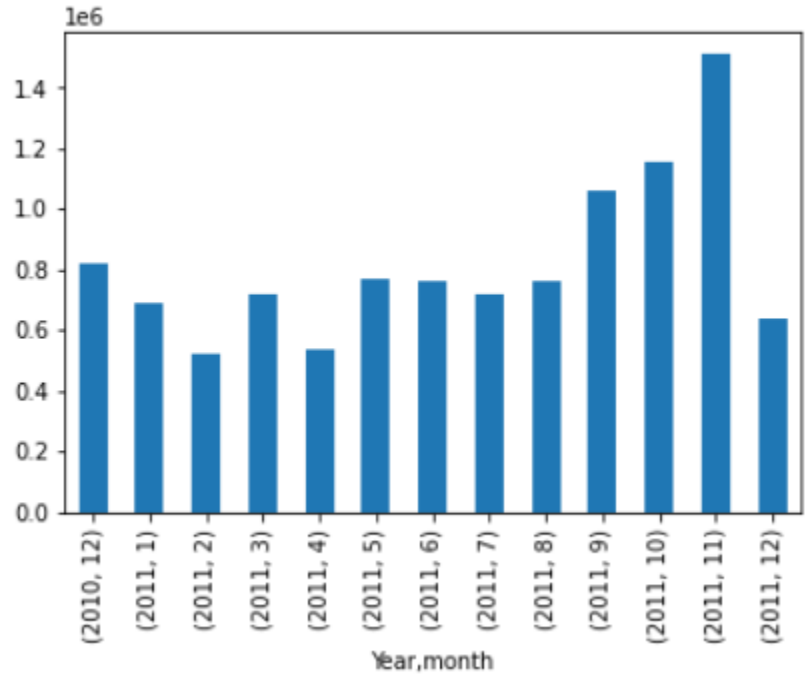
Lowest 5 countries based on Total Sales



Revenue Per year



Total Sales Per month



Choosing the Algorithm for the Project

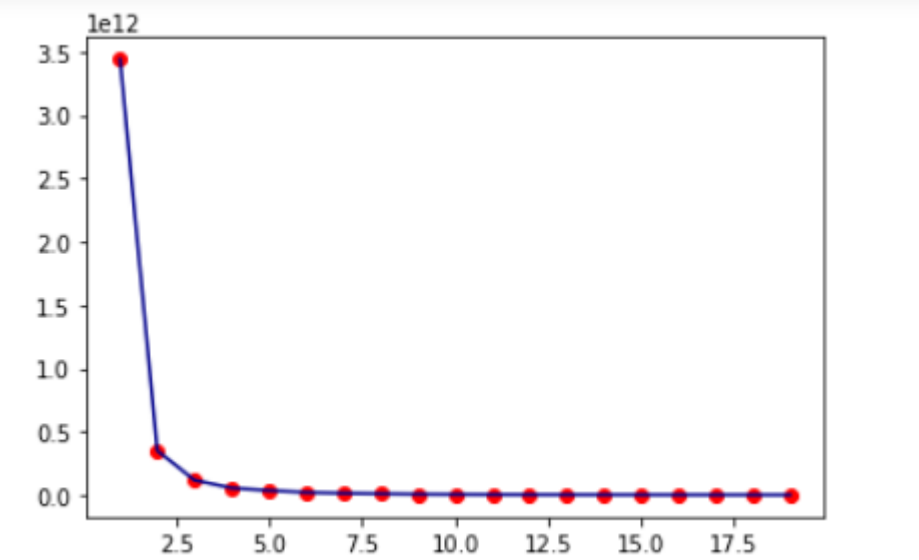
Here we will adopt K means clustering technique. The goal of K means is to group data points into distinct non-overlapping subgroups.

Motivation and Reasons for Choosing the Algorithm

The goal of K means is to group data points into distinct non-overlapping subgroups. We will divide the whole data of customers on the basis of RMF i.e. Recency, Monetary and Frequency and we will also visualize these groups on the basis of these 3 terms. This segmentation will help us to get a better understanding of customers which in turn could be used to increase the revenue of the company.

Model Evaluation and Techniques

In the KMean algo we are using elbow method to find the no. of clustering groups. From above 5 is ideal value for k



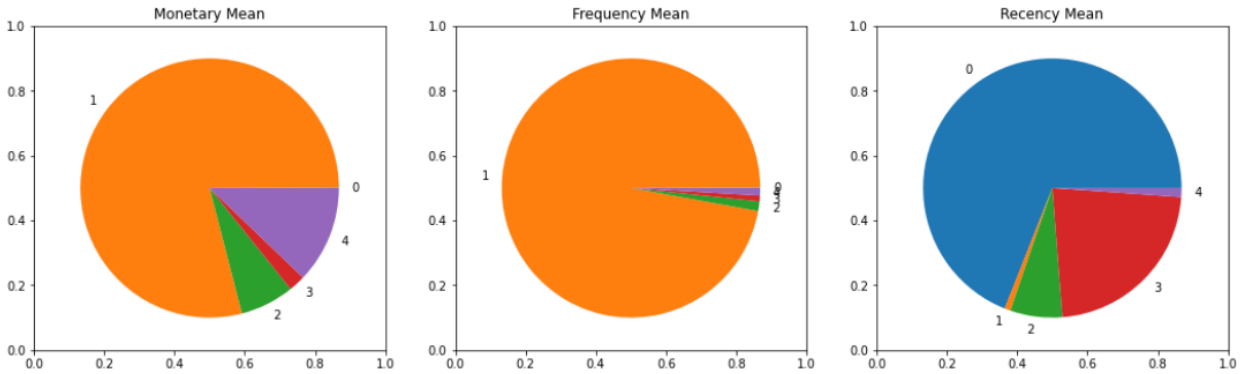
Inferences from the Same

On basis of Preprocessing Visualization

- This company is more into online retail across 38 countries.

- We should continue cooperation with EIRE, Germany, France, Australia and change relationships with Saudi Arabia
- The best sales month in November 2011
- We can concentrate on improving the sales for the other 8 months
- We see that September to December we have very high sales so we can concentrate on improving the sales for the other 8 months

On basis of Segmentation after adoption of Kmeans and below pie chart we can easily understand our 5 groups according to Recency mean,Frequency mean and Monetary mean.



- Group 1 is the group of customer who spends maximum amount of money and also has a good frequency and low recency rate.
- Group 4 are the customers whose frequency rate is maximum and monetary value is also good and recency rate is also quite good, whereas Group 0 is the group of customers who has a very high recency rate means they have not purchased anything from the past.

Future Possibilities of the Project

Next steps would be

- which one is the best weekday for sales
- which time is best for sales

Conclusion

With above inference captured we can conclude Project Online Retail Customer Segmentation is complete.

References

https://www.analyticsvidhya.com/

https://www.youtube.com/watch?v=EItlUEPCIzM

https://jovian.ai/