Customer market segmentation
Jayaprakash (prakashjz121@gmail.com)

Problem statement : To understand the various customer purchase patterns for a online firm from a given dataset

Data: https://www.kaggle.com/code/anushasuresh348/recommendation-system-product/notebook

What is customer segmentation?

- *Customer segmentation is an effective tool for businesses to closely align their strategy and tactics with, and better target, their customers. Every customer is different, and every customer journey is different so a single approach often isn't going to work for all. This is where customer segmentation becomes a valuable process.*
- *The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business. Let's do deep dive on the given business problem….*

Problem Statement

- *An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence based insights to provide the same.*

Project Objective

- Using the above data, find useful insights about the customer purchasing history that can be an added advantage for the online retailer.
- Segment the customers based on their purchasing behavior.

Data Description

- Given dataset contains 541909 rows and 8 columns,
- Primary key is CustomerID, and
- Data volume is very HIGH , processing time took more than expected.

Data Pre-processing Steps and Inspiration

Data Cleanup

- During data analysis, few rows/columns had null values which were cleaned up to improve the data quality,
- Columns details in the given dataset,

Feature Name	Description
Invoice	Invoice number
StockCode	Product ID
Description	Product Description
Quantity	Quantity of the product
InvoiceDate	Date of the invoice
Price	Price of the product per unit
CustomerID	Customer ID
Country	Region of Purchase

- Validate the null values and drop them if exists, and

```
df.isnull().sum()

InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID    135080
Country            0
dtype: int64
```

- There are no missing values in the given dataset
    - Converted date column to appropriate format and also created few more features (labels) for better insights using date column.

```
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])

from datetime import datetime

df['Invoice Year'] = (df["InvoiceDate"].dt.year).astype(int)
df['Invoice Month'] = (df["InvoiceDate"].dt.month).astype(int)
df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Invoice Year	Invoice Month
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	2010	12
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010	12
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	2010	12
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010	12
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010	12

- Changed datatype for CustomerId for better data analysis, and
- Created few user defined / helper functions for reusability.

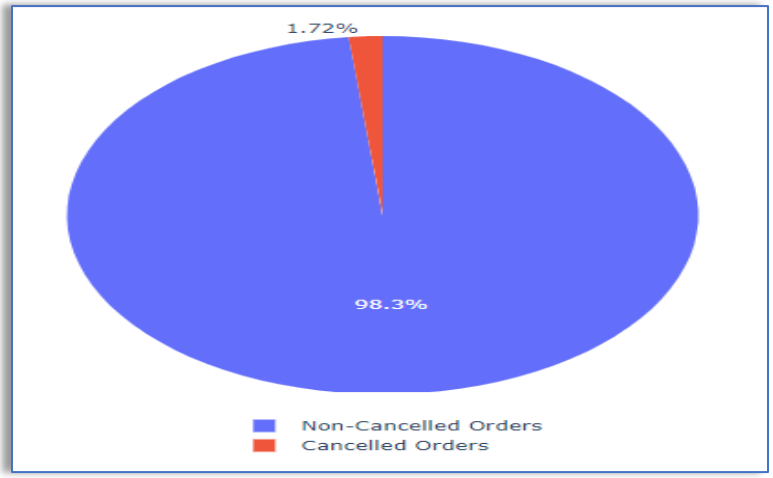
Exploratory Data Analysis (EDA)

Top 5 Sales Insights Dashboard

- Above analysis shows that,
  - Sales were high in UK when compared to other countries,
  - From Year wise sales, it is evident that in year 2011, there were high sales than the 2010, and
  - From monthly distribution chart, sales were high in October, November and December (mostly in year end).

Proportion of Cancelled orders

- Canceller orders where the quantity is under zero from the given data,



- From the cancelled order distribution , sounds like cancellation distribution was very less % (2%) so data seems to be good for further analysis ,
- Looking deeper into why these orders were cancelled may prevent future cancellations. Now let's find out what a negative Unit Price or Quantity means ,

```
cprint("Cancelled Orders",'blue')
df[df['Quantity'] <= 0][['InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate', 'UnitPrice', 'CustomerID', 'Country']]
```

Cancelled Orders

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
...	...	...	...	...	...	...	...	...
540449	C581490	23144	ZINC T-LIGHT HOLDER STARS SMALL	-11	2011-12-09 09:57:00	0.83	14397.0	United Kingdom
541541	C581499	M	Manual	-1	2011-12-09 10:28:00	224.69	15498.0	United Kingdom
541715	C581568	21258	VICTORIAN SEWING BOX LARGE	-5	2011-12-09 11:57:00	10.95	15311.0	United Kingdom
541716	C581569	84978	HANGING HEART JAR T-LIGHT HOLDER	-1	2011-12-09 11:58:00	1.25	17315.0	United Kingdom
541717	C581569	20979	36 PENCILS TUBE RED RETROSPOT	-5	2011-12-09 11:58:00	1.25	17315.0	United Kingdom

9762 rows × 8 columns

  - Invoice number starts w/ 'C', most likely it could be the order which was cancelled or customers who abandon their order,
  - Descriptions showed that some of the transactions were discounted , probably it could be a promotion to make better sales,
  - Another interesting observations was that most of the cancelled (discounted) orders were from UK where sales were HIGH , this could be a reason for better sales than other countries so online store can use this findings for other countries for better sales rates, and

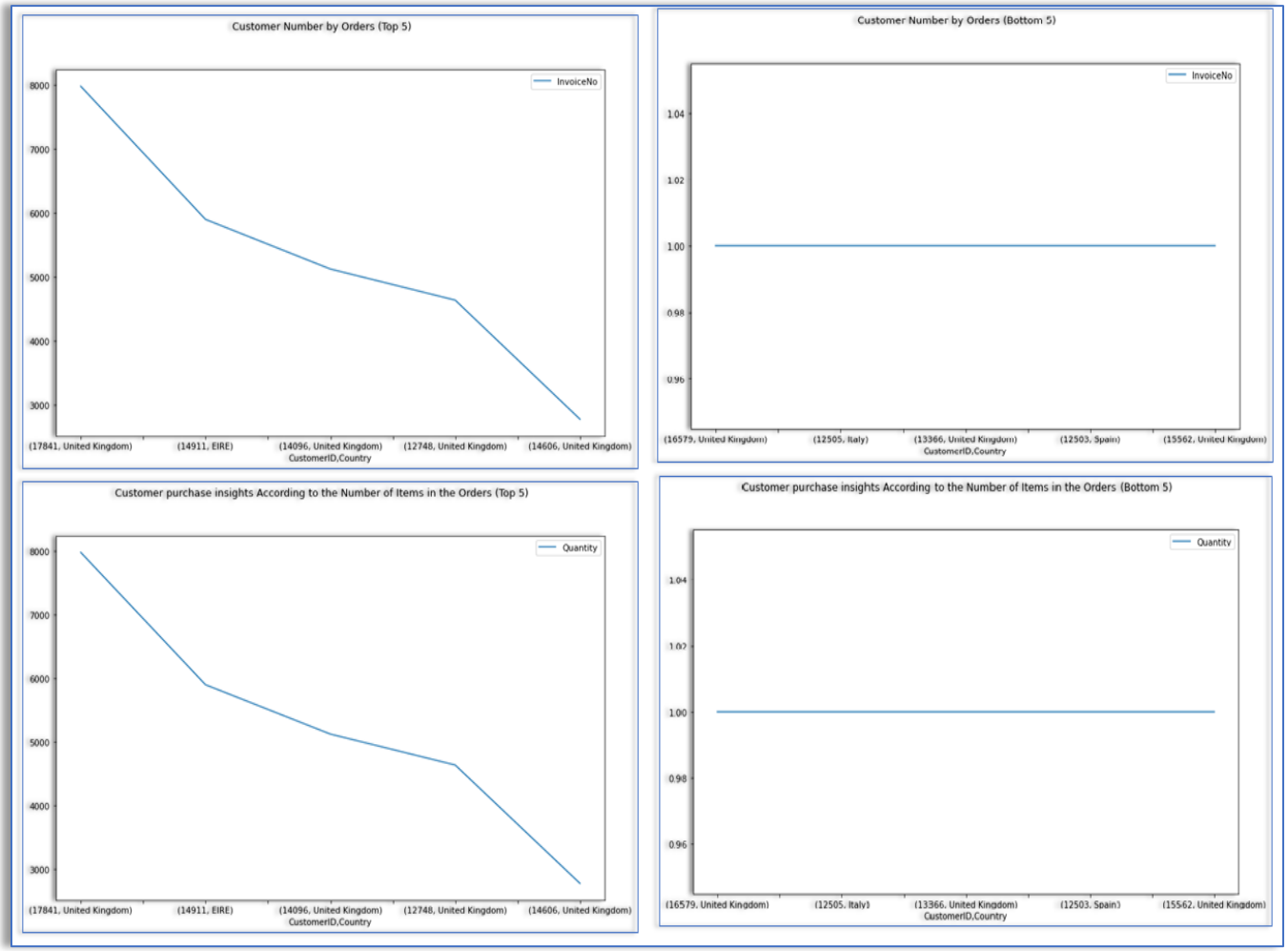
```
round(df[df["InvoiceNo"].str.startswith('C')]['CustomerID'].nunique() / df['CustomerID'].nunique()*100,2)
```

```
36.36
```

  - 36% --> cancelled w/ 'customerid"].nunique() ratio , so cancelled orders were distributed from more one customers .

Top 5 Customer Insights dashboard

Exploring the Orders or spend

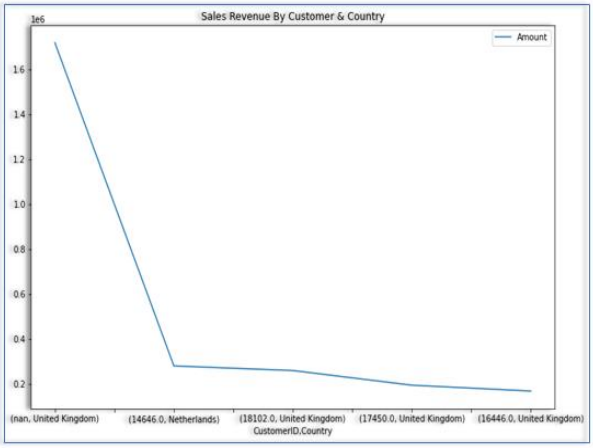


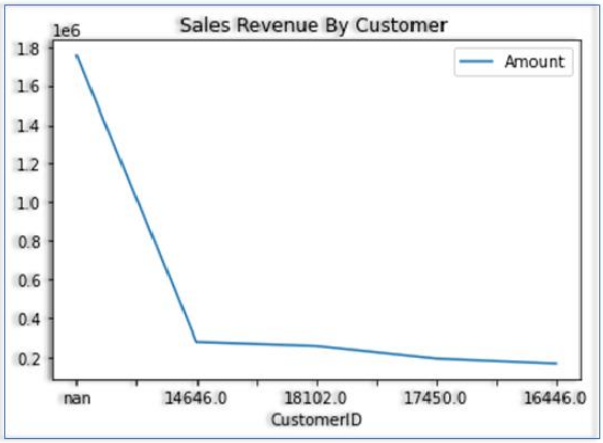
- Observations from the above Orders dashboard is below , and
  - In the given dataset, for some of the rows where customerID were empty or null, since it is reference number so I was unable to assign default values,
  - Interestingly , from bottom 5 order chart lowest order was almost 1 for all 5 countries ,
  - It was evident that sales were high in year end , so each country can have more products or choice to improve sales opportunities
- There was another observation that some of the customers were purchased from more one countries, Ex: CustomerID :12347.

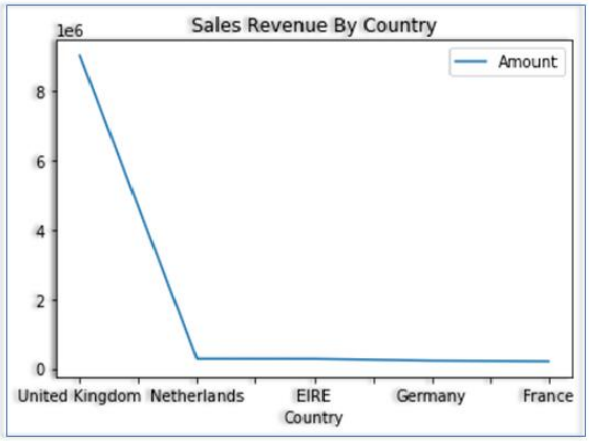
```
df.groupby('CustomerID')['Country'].nunique().value_counts()

1    4364
2       8
9       1
```

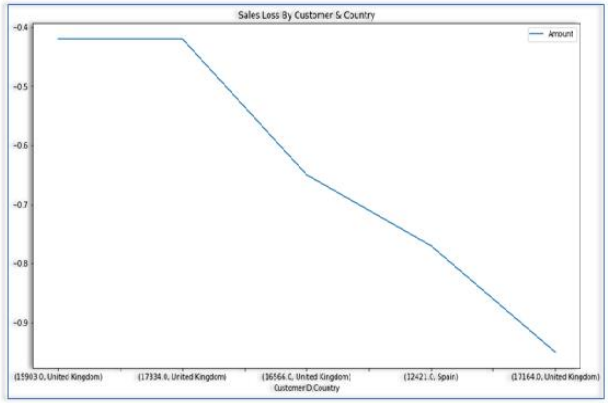
Top 5 Revenue & Loss Insights

- For revenue analysis, rows which had zero or under zero amount were NOT considered , and for loss , it was in reverse where rows w/ -ve amount were considered ,









- Above analysis shows that,
    - Revenue was high in UK when compared to other countries but for some of the sales where there were reference to customers, and other countries like Netherlands, EIRE, Germany & France were also in the top 5 revenue list ,
    - Top 5 Customers which contributed towards revenue were nan [No customerid reference],14646,18102,17450 & 16446 ,
    - Top 5 Loss revenue or cancelled orders were from UK [Customer: 15903,17334,16566] & Spain [Customer :17164] countries, and

Most popular products from UK

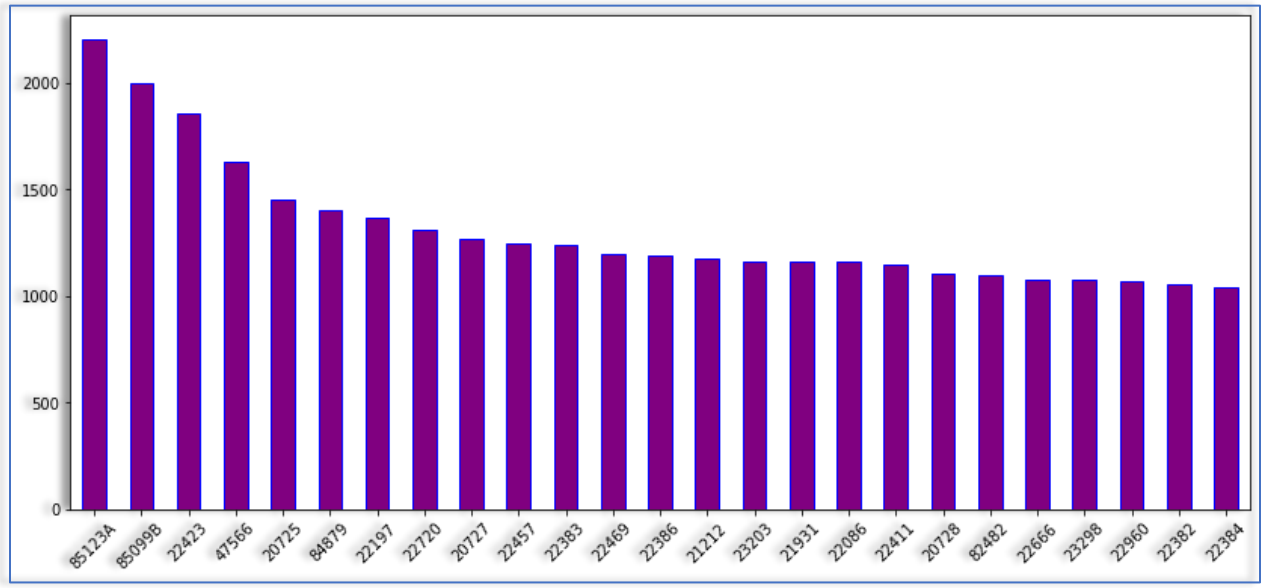
- The UK not only has the most sales revenue, but also the most customers. Since the majority of this dataset contains orders from the UK, we can explore the UK market further analysis by finding out what products the customers buy together and any other buying behaviors to improve our sales and targeting strategy.

Most Popular Products from UK (Top 5)

StockCode	Description	Quantity
84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	48326
85099B	JUMBO BAG RED RETROSPOT	43167
22197	POPCORN HOLDER	34365
84879	ASSORTED COLOUR BIRD ORNAMENT	33679
85123A	WHITE HANGING HEART T-LIGHT HOLDER	32901

Top 25 most purchased products

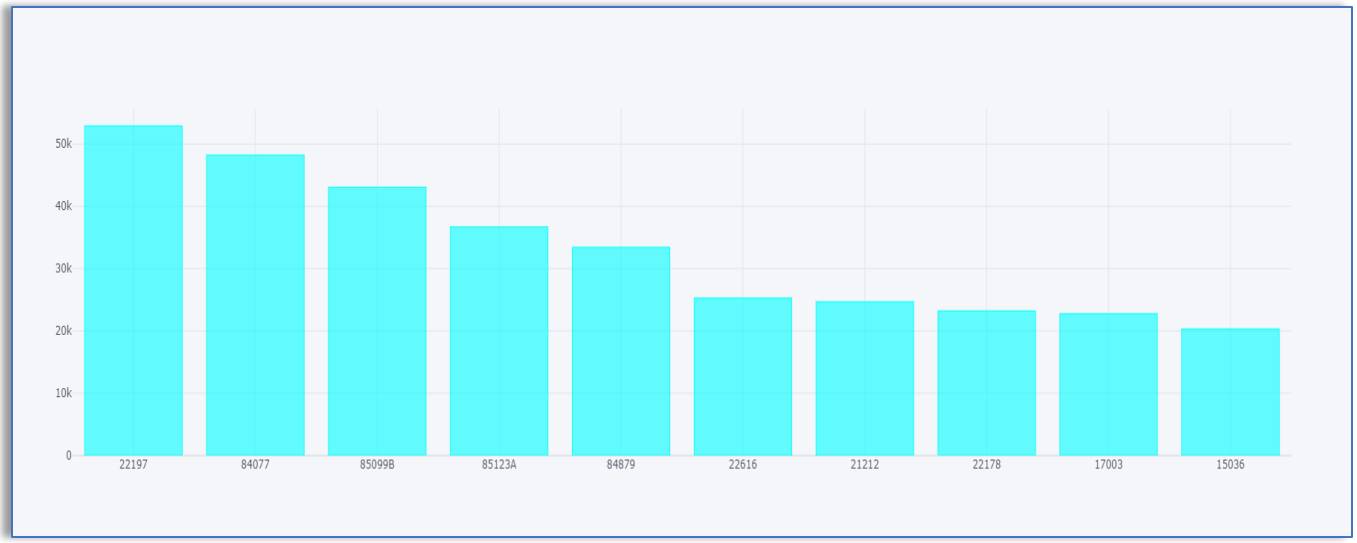
- Below chart shoes top 25 products which were purchased by UK customers.



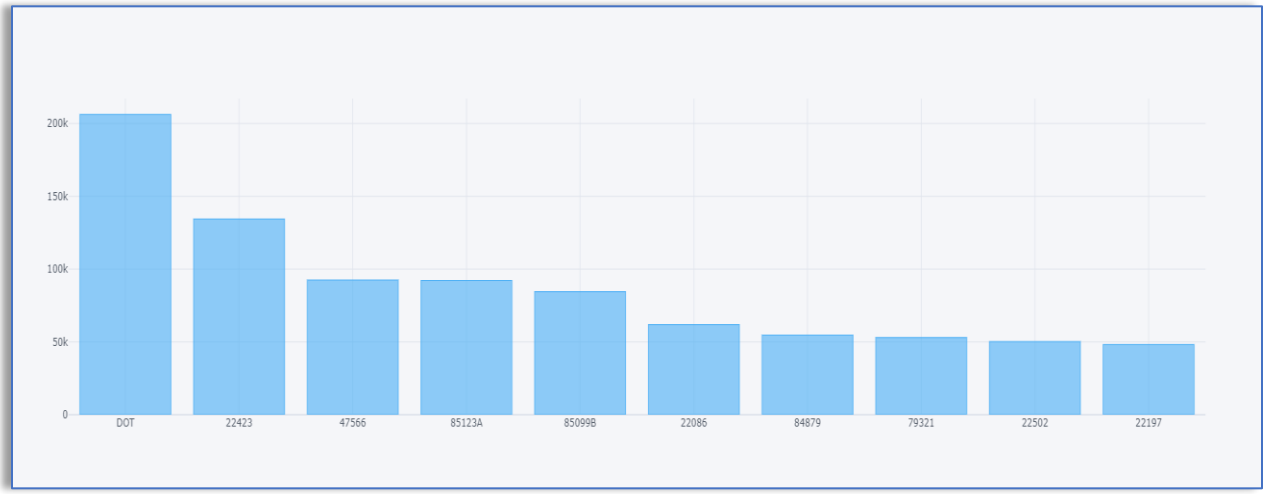
Top 10 Demanded Products by Quantity

Top 10 Demanded Products [StockCode] By Quantity were,

- 22197 - SMALL POPCORN HOLDER
- 84077 - WORLD WAR 2 GLIDERS ASSTD DESIGNS
- 85099B - JUMBO BAG RED RETROSPOT
- 85123A - WHITE HANGING HEART T-LIGHT HOLDER
- 84879 -ASSORTED COLOUR BIRD ORNAMENT and so on…



Top 10 Demanded Products by Total Price



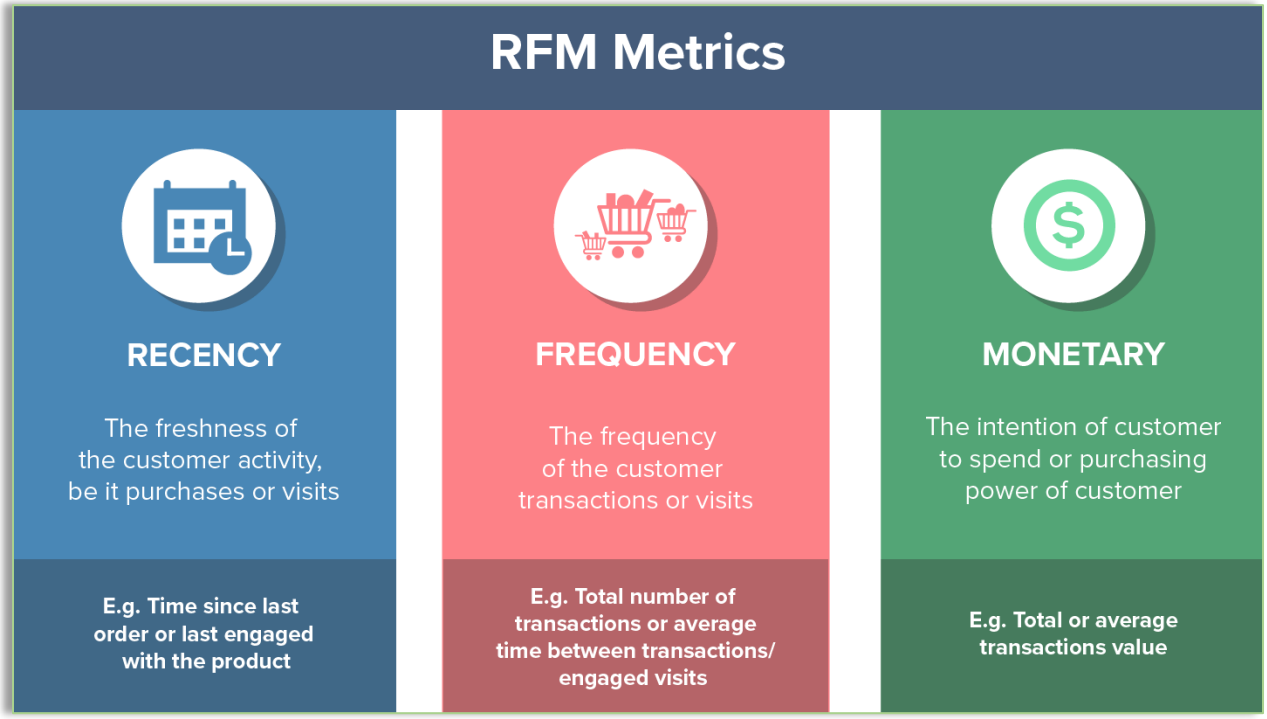
- From the above 2 charts, there was an interesting observation that three product were part of highly demanded by quantity and price , this can be looked further for better sales across countries.

```
filtered_df = df_uk.query( "stockcode == ['85123A' ,'22197','85099B']" )

filtered_df.head()
```

	invoiceno	stockcode	description	quantity	invoicedate	unitprice	customerid	country	invoice_year	invoice_month	amount
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	2010	12	15.3
49	536373	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 09:02:00	2.55	17850.0	United Kingdom	2010	12	15.3
66	536375	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 09:32:00	2.55	17850.0	United Kingdom	2010	12	15.3
177	536386	85099B	JUMBO BAG RED RETROSPOT	100	2010-12-01 09:57:00	1.65	16029.0	United Kingdom	2010	12	165.0
220	536390	85123A	WHITE HANGING HEART T-LIGHT HOLDER	64	2010-12-01 10:19:00	2.55	17511.0	United Kingdom	2010	12	163.2

Customer Segmentation using RFM ANALYSIS in UK market

- *RFM (Recency, Frequency, Monetary) Analysis is a customer segmentation technique for analyzing customer value based on past buying behavior. RFM analysis was first used by the direct mail industry more than four decades ago, yet it is still an effective way to optimize your marketing,*
    - ***RECENCY (R):** Time since last purchase*
    - ***FREQUENCY (F):** Total number of purchases*
    - ***MONETARY VALUE (M):** Total monetary value*

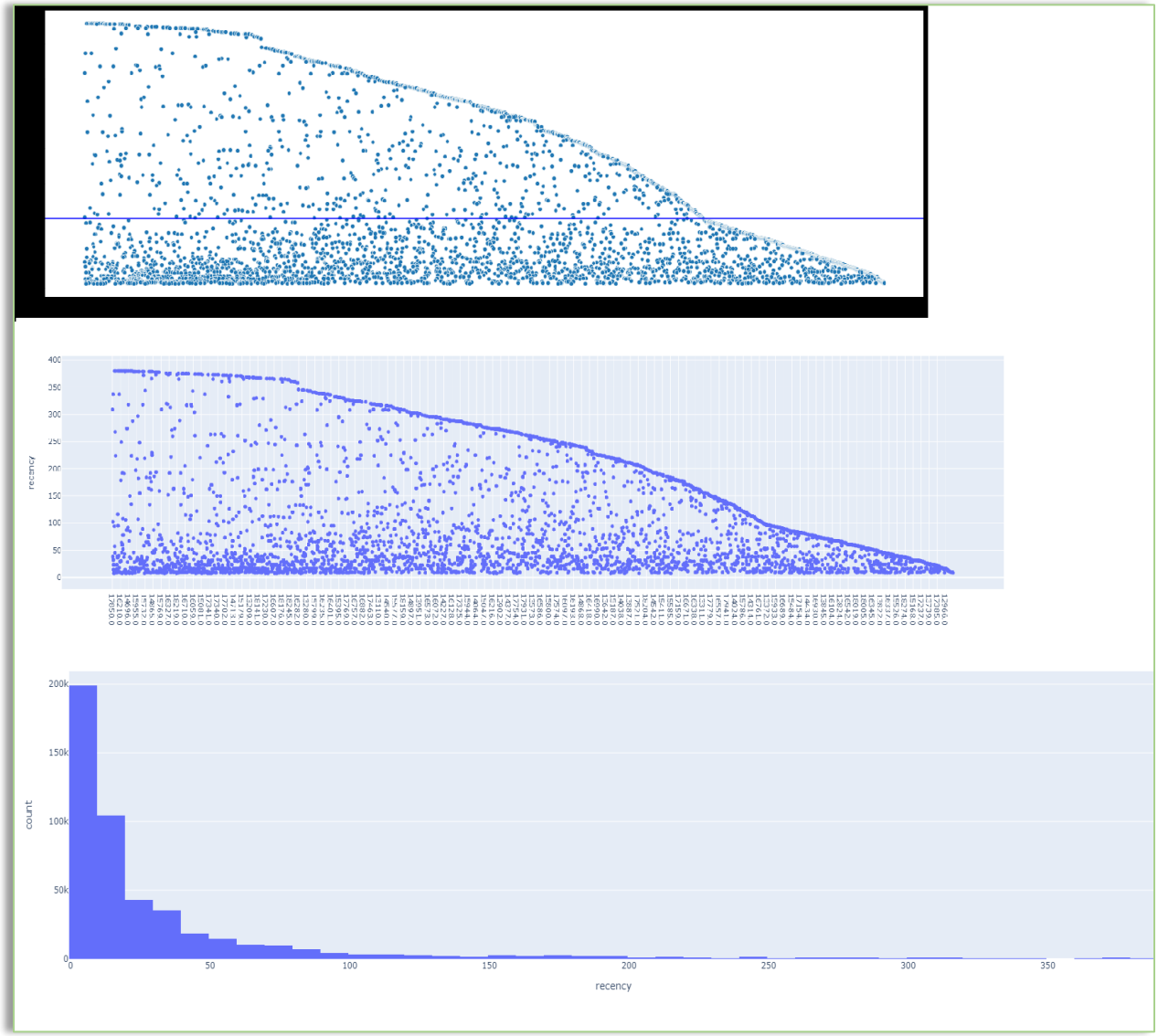


Calculating Recency

- To calculate the recency values, follow these steps in order:
    1. To calculate recency, we need to choose a date as a point of reference to evaluate how many days ago was the customer's last purchase.
    2. Create a new column called Date which contains the invoice date without the timestamp.
    3. Group by CustomerID and check the last date of purchase.
    4. Calculate the days since last purchase.
    5. Drop Last_Purchase_Date since we don't need it anymore.
    6. Plot RFM distributions.
- *Below was Recency table for a given dataset after following the above steps,*

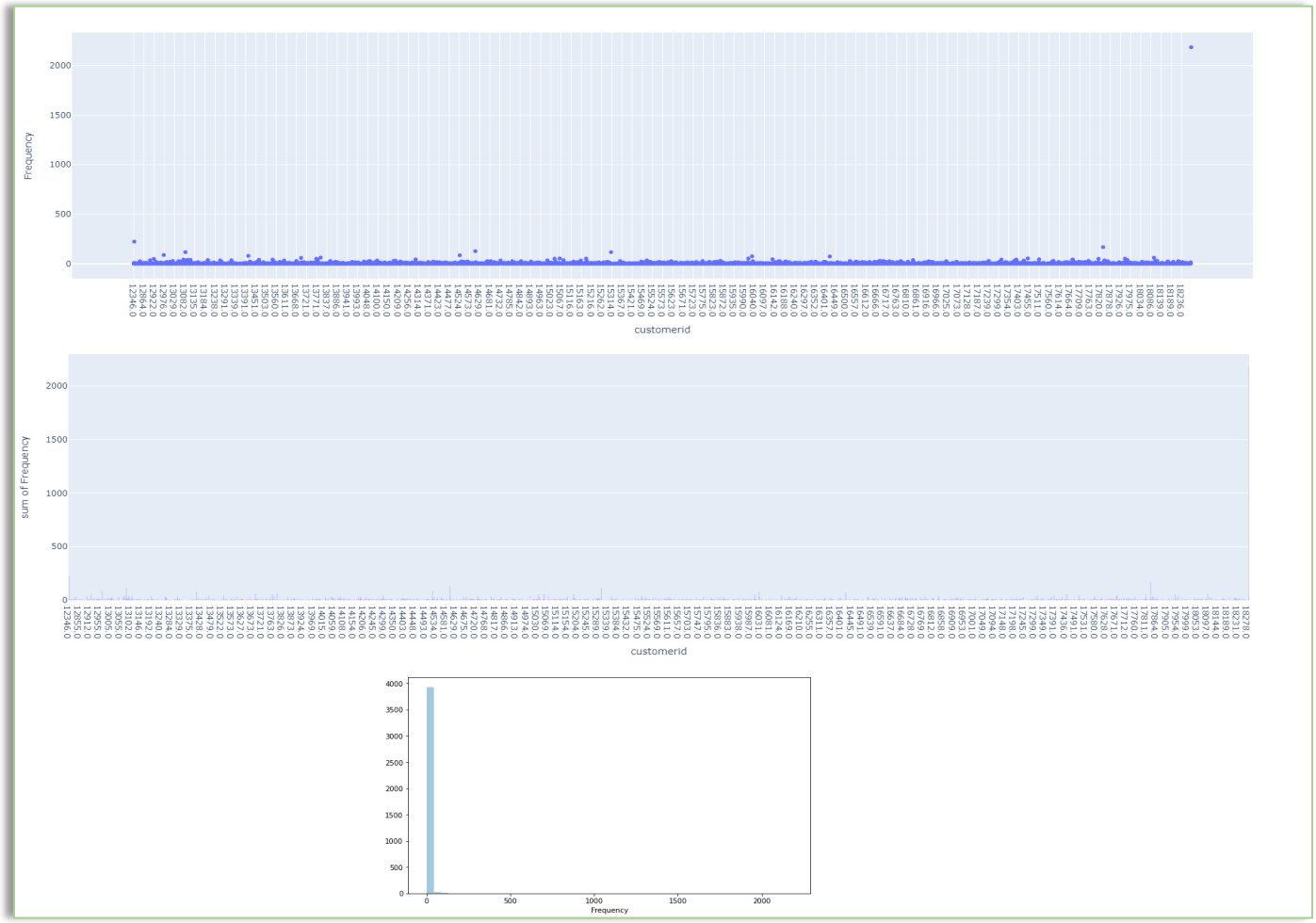
	customerid	Recency
1013	14237.0	380.0
3748	18011.0	380.0
2495	16274.0	380.0
3718	17968.0	380.0
177	13065.0	380.0

Plot RFM distributions



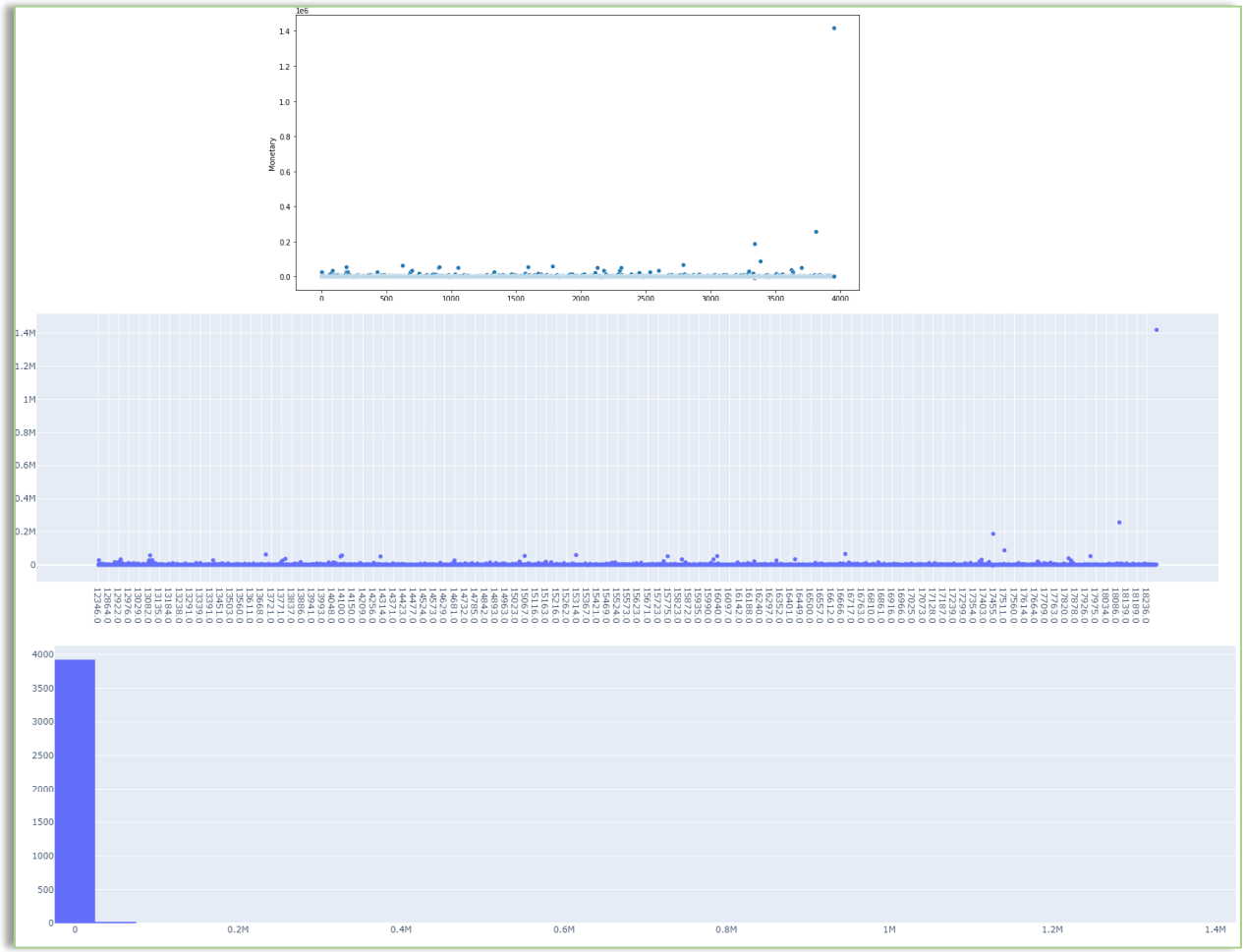
- *Here, we can note that the histogram is biased towards the left side and hence this is a sign of distribution which is a right-skewed distribution and also we can see that the rug plot is crowded between 0 and 50. Based on that we can see that we have a high concentration of customers in the last 50 days, i.e. the last 2 months.*

Calculating Frequency



- Frequency chart did not provide much insight to analyze further but sounds like , It is common that people do purchasing less frequently.

Calculating Monetary Values



- From the above chart ,it was marked that not many customers spend more than € 25,000.

Creating RFM Table

- RFM Table created by merging the recency, frequency and monetary dataframes.

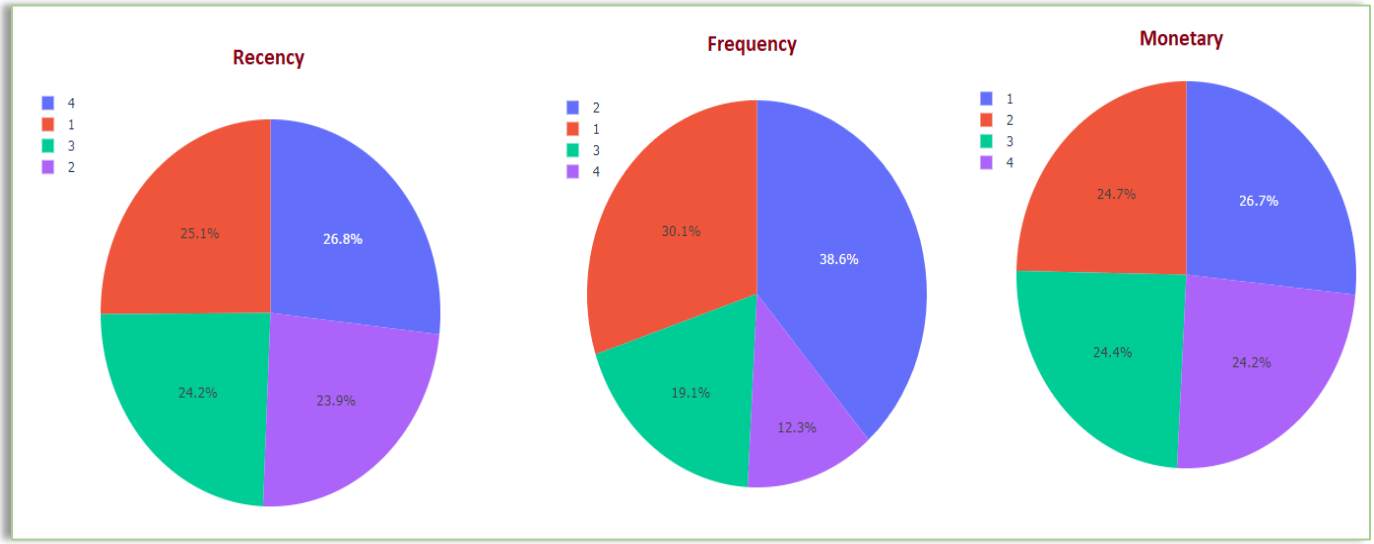
	customerid	Recency	Frequency	Monetary
0	12346.0	332.0	2	0.00
1	12747.0	9.0	11	4196.01
2	12748.0	7.0	224	28405.56
3	12749.0	10.0	8	3868.20
4	12820.0	10.0	4	942.34

RFM w/ Quantile

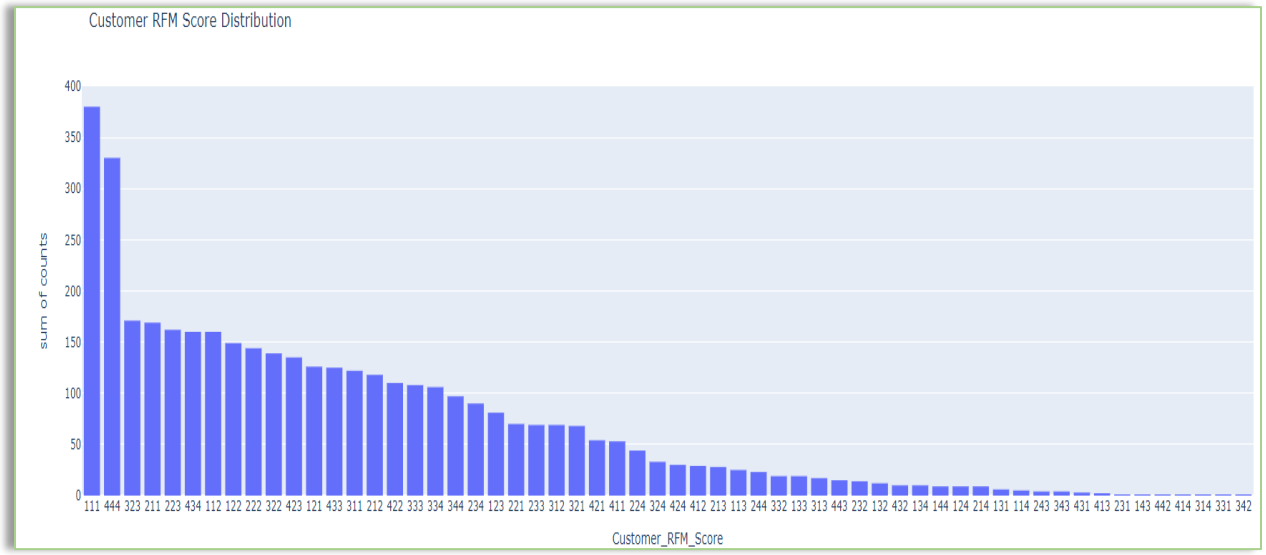
- Using RFM score , created below table w/ quantiles,

	Recency	Frequency	Monetary
0.25	23.0	1.0	281.160
0.50	57.0	3.0	623.390
0.75	150.0	5.0	1518.925

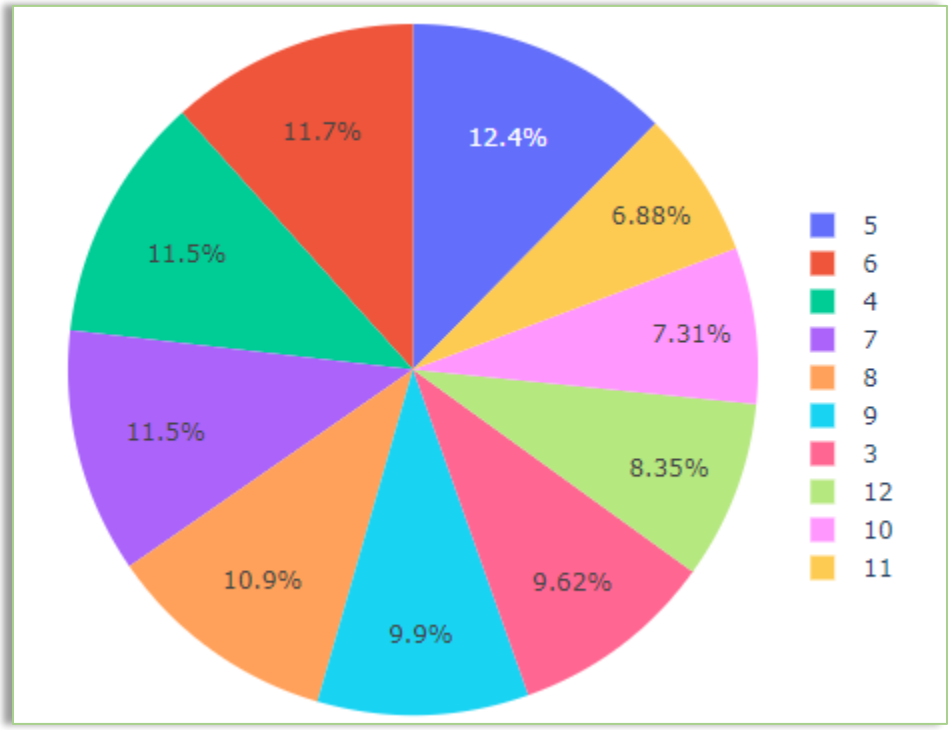
RFM Distribution



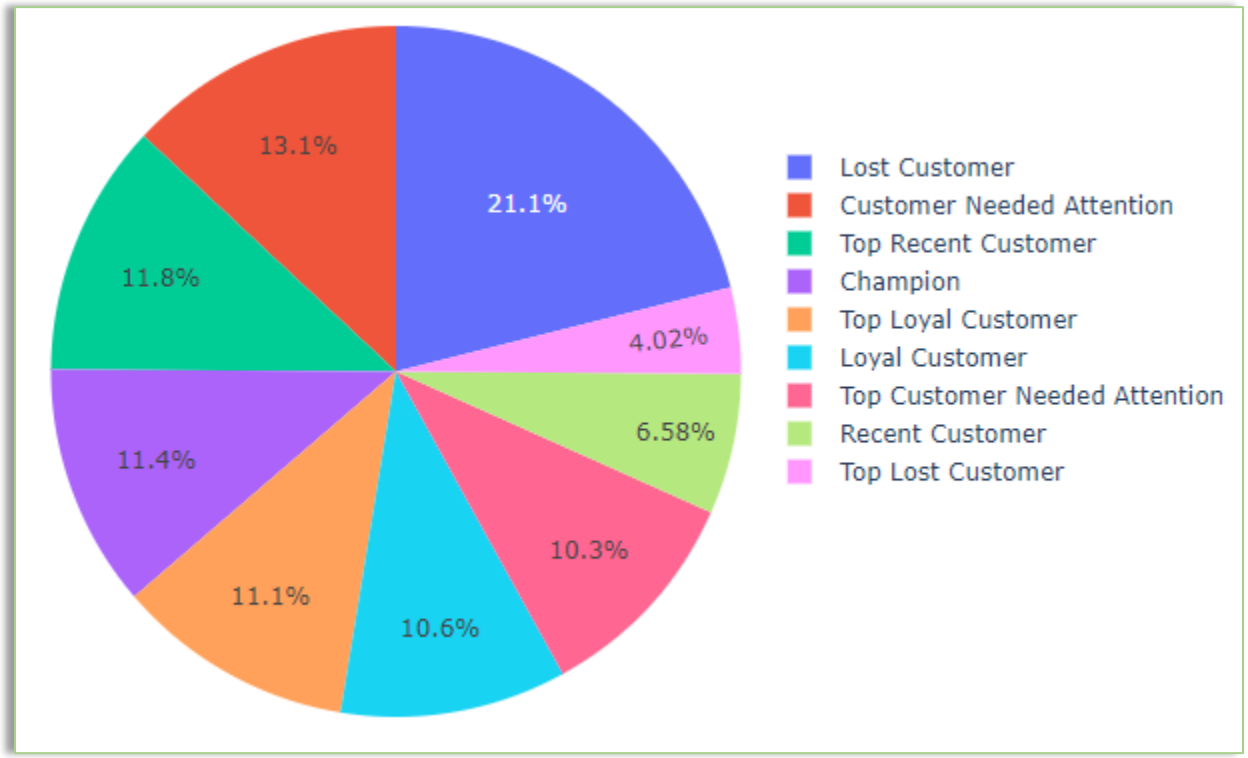
Customer RFM Score Distribution



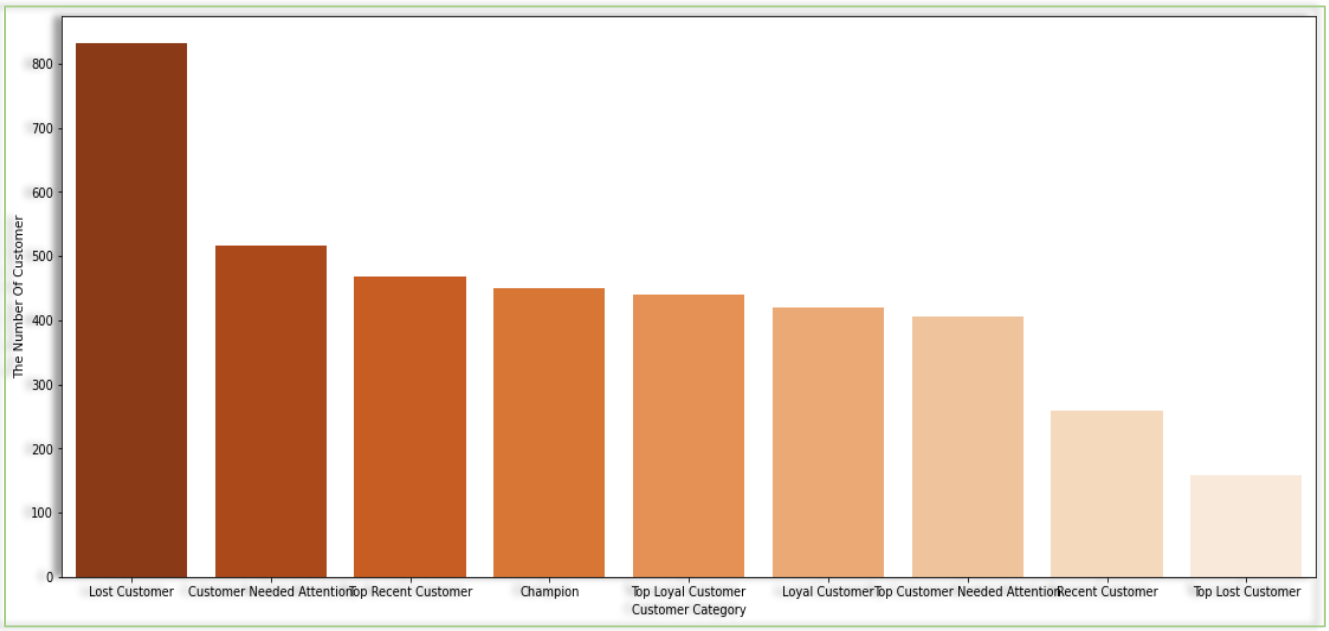
RFM Label distribution



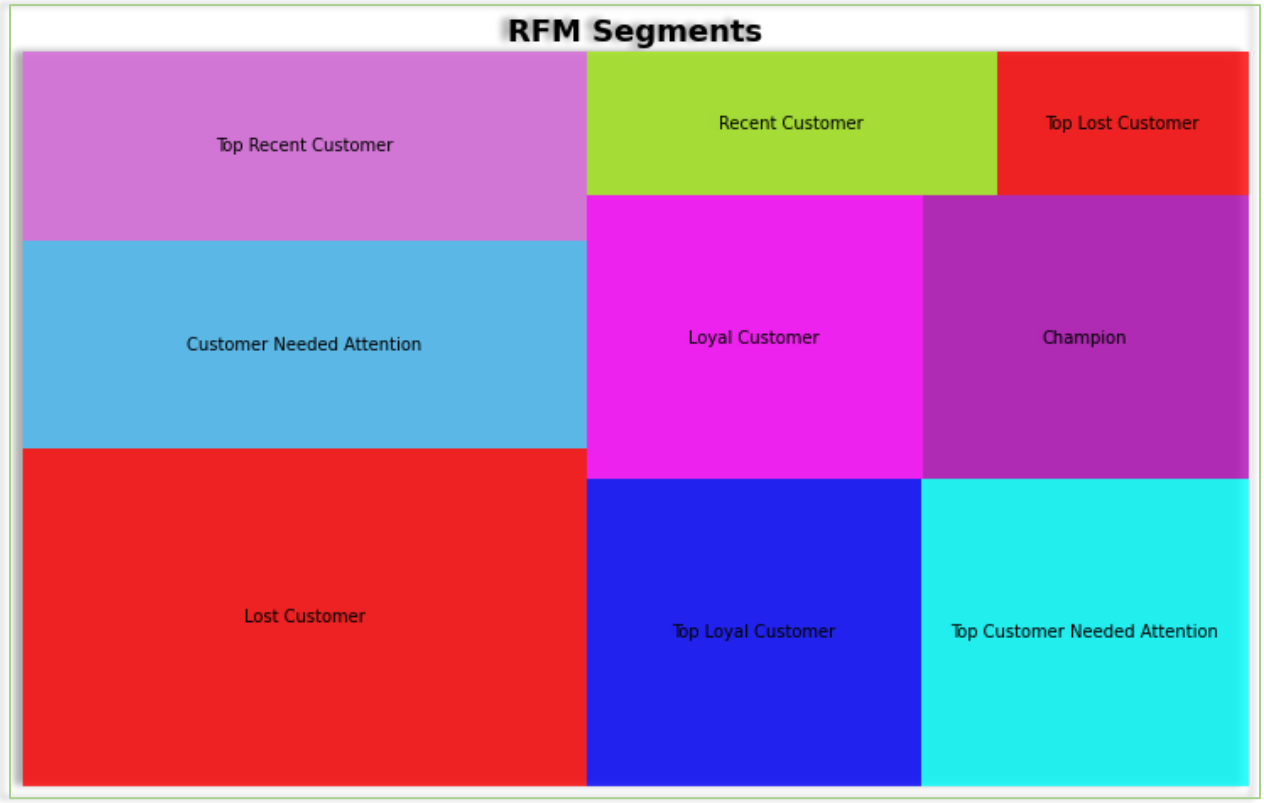
Customer Category Distribution



Plotting RFM Segments



Customers in each segment



Segment Visualization with WorldCloud

Conclusion

- In this project, three different types of analyses were applied independently, to get better insights from the given dataset which are listed below,
    - It was evident that sales were high in year end, so each country can have more products or choice to improve sales opportunities .
    - In the given dataset, some of the rows (25 %) did not customer references so it will have an impact on customer segmentation insights
    - From top demanded products by quantity and price, there was an interesting observation that three product were part of highly demanded by quantity and price, this can be looked further for better sales across countries. Except UK, other countries can have promotion or discounted price for better sales.
- For Customer Segmentation, RFM technique was performed, which is the technique of diving customers into groups based on their purchase patterns to identify who are the most profitable groups.
- In segmenting customers, various criteria can also be used depending on the market such as geographic, demographic characteristics or behavior bases. This technique assumes that groups with different features require different approaches to marketing and wants to figure out the groups who can boost their profitability the most.
- In this sense, the customers in our dataset were divided into insightful clusters. What makes RFM analysis attractive is the flexibility it offers so the segmentation could be specified in terms of business needs and an analyst could create customer groups based on their purchase history – how recently, with what Frequency, and what value they bought. On the other hand, it is susceptible to user-induced biases.
- We can start taking actions with this segmentation.
- The main strategies are quite clear:
    - High Value: Improve Retention
    - Mid Value: Improve Retention + Increase Frequency and
    - Low Value: Increase Frequency

Appendix

- We can develop different types of customer segments with RFM modeling, but here are 11 segments commonly used to make better decision.

Champions	Bought recently, buy often and spend the most!	Reward them. Can be early adopters for new products. Will promote your brand.
Loyal Customers	Spend good money with us often. Responsive to promotions.	Upsell higher value products. Ask for reviews. Engage them.
Potential Loyalist	Recent customers, but spent a good amount and bought more than once.	Offer membership / loyalty program, recommend other products.
Recent Customers	Bought most recently, but not often.	Provide on-boarding support, give them early success, start building relationship.
Promising	Recent shoppers, but haven't spent much.	Create brand awareness, offer free trials
Customers Needing Attention	Above average recency, frequency and monetary values. May not have bought very recently though.	Make limited time offers, Recommend based on past purchases. Reactivate them.
About To Sleep	Below average recency, frequency and monetary values. Will lose them if not reactivated.	Share valuable resources, recommend popular products / renewals at discount, reconnect with them.
At Risk	Spent big money and purchased often. But long time ago. Need to bring them back!	Send personalized emails to reconnect, offer renewals, provide helpful resources.
Can't Lose Them	Made biggest purchases, and often. But haven't returned for a long time.	Win them back via renewals or newer products, don't lose them to competition, talk to them.
Hibernating	Last purchase was long back, low spenders and low number of orders.	Offer other relevant products and special discounts. Recreate brand value.
Lost	Lowest recency, frequency and monetary scores.	Revive interest with reach out campaign, ignore otherwise.