Assignment submission

TABLE OF CONTENT

Introduction and data Dictionary
Importing libraries
Data Wrangling
- Missing values
- Structuring

Analysis
- Exploratory Data Analysis
- RFM Analysis

Final Conclusions
- Findings
- Recommendation

Data

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
525456	538171	22271	FELTCRAFT DOLL ROSIE	2	2010-12-09 20:01:00	2.95	17530.0	United Kingdom
525457	538171	22750	FELTCRAFT PRINCESS LOLA DOLL	1	2010-12-09 20:01:00	3.75	17530.0	United Kingdom
525458	538171	22751	FELTCRAFT PRINCESS OLIVIA DOLL	1	2010-12-09 20:01:00	3.75	17530.0	United Kingdom
525459	538171	20970	PINK FLORAL FELTCRAFT SHOULDER BAG	2	2010-12-09 20:01:00	3.75	17530.0	United Kingdom
525460	538171	21931	JUMBO STORAGE BAG SUKI	2	2010-12-09 20:01:00	1.95	17530.0	United Kingdom

Explore the data — validation and new variables

- Missing values in Customer's ID column
- Customers' distribution in each country
- Unit price and Quantity should > 0
- Invoice date should < today.

Missing values

- There are `107927` missing values in "CustomerID "column
- I`ll make a new dataframe without null values for RFM Analysis

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

Our data is clean now :)

Price

After observing we see that there are values under zero. These values are debt and we don't need them.
Let's filter out the data

```
#lets remove the negative quantity
df =df[df['Quantity']>0]
df.shape
```

```
(407695, 8)
```

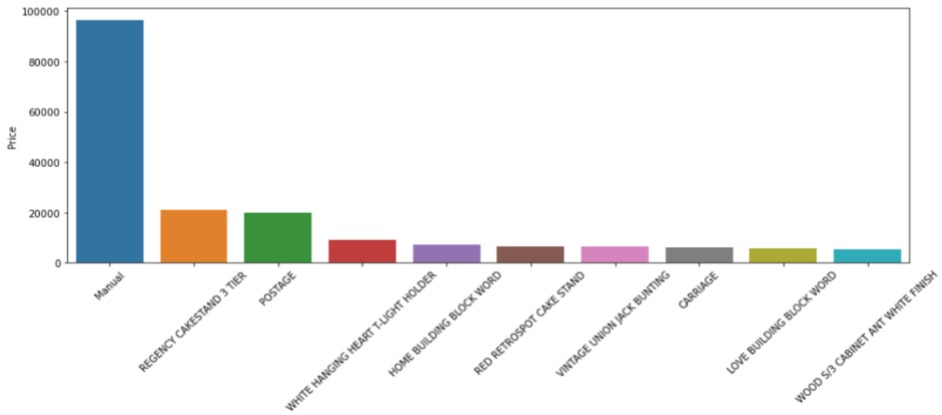
Adding Total Price column

I will add new column which will be the Total price column

```
#Lets add a column for total price
df['TotalPrice']= df['Quantity'] * df['Price']
```

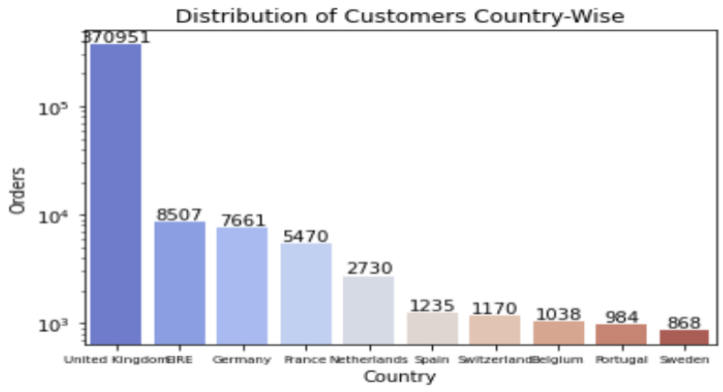
EXPLORATORY DATA ANALYSIS

Q1: Which Products has the highest sales?



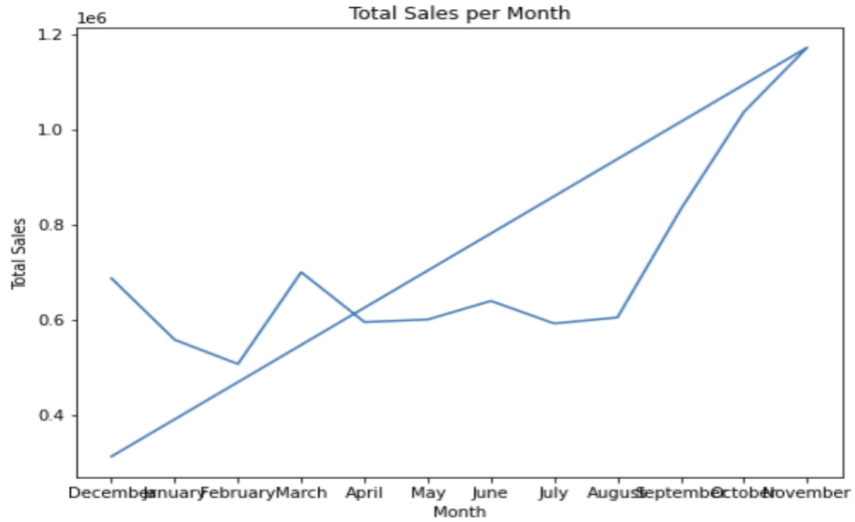
Some descriptions are just Manual and not items

Q2: In which country we have more customers?



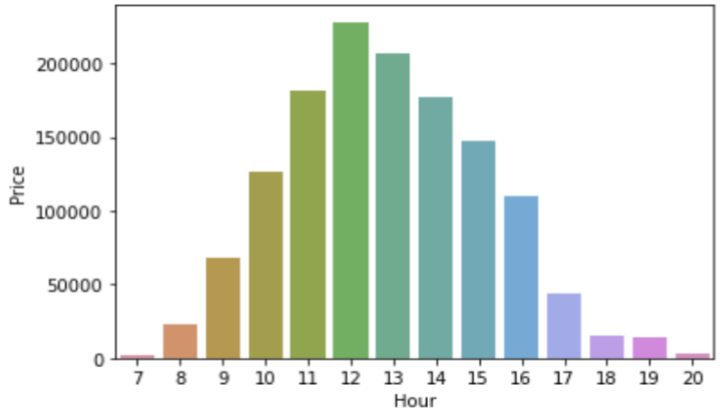
We can see that United kingdom contributes to the maximum number of orders hence more sales. Hence, the countries with high orders have potential or growth and we must improve customer acquisition in countries with less orders.

Q3: In which month we gained the highest sales?



We can see that the best sale is in November 2011. We can identify the months in which the demand is high and the months in which it is low. Based on this analysis, we can plan company inventory levels, adjust pricing strategies, and optimize marketing efforts.

Q4:Which hour is the busiest?



We can see that 12pm is the busiest hour in a day.This information can be used to optimize the store operations, such as scheduling staff, replenishing inventory at 12pm.

RMF Analysis

1. RMF Metrics for each customer

RFM (Recency, Frequency, Monetary) analysis is a customer segmentation technique that uses past purchase behaviour to divide customers into groups. RFM helps divide customers into various categories or clusters to identify customers who are more likely to respond to promotions and also for future personalization services.

RECENCY (R): Days since last purchase

FREQUENCY (F): Total number of purchases

MONETARY VALUE (M): Total money this customer spent.

We will create these 3 customer attributes for each customer.

RMF TABLE

Customer ID	recency	frequency	monetary_value
12346.0	529	33	372.86
12347.0	367	71	1323.32
12348.0	438	20	222.16
12349.0	407	102	2671.14
12351.0	375	21	300.93

---

2. Calculating RMF scores

Let's Split the metrics

The easiest way to split metrics into segments is by using quartiles.

- This gives us a starting point for the detailed analysis.

- 4 segments are easy to understand and explain.

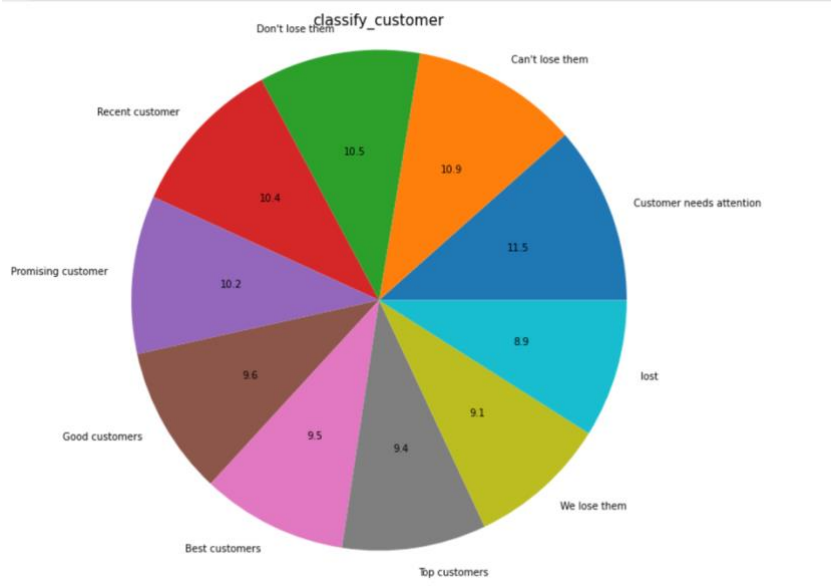
Customer ID	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore
12346.0	529	33	372.86	4	3	3	433
12347.0	367	71	1323.32	1	2	2	122
12348.0	438	20	222.16	3	3	4	334
12349.0	407	102	2671.14	2	2	1	221
12351.0	375	21	300.93	1	3	4	134

3. Creating & Analysing RFM Segments

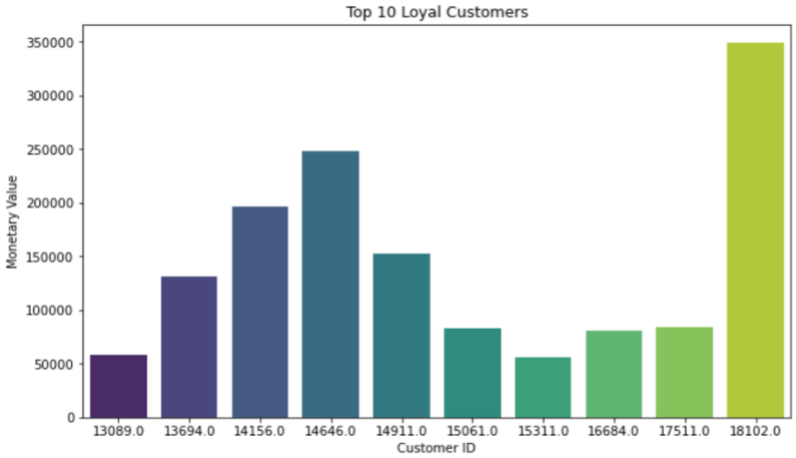
Customer ID	recency	frequency	monetary_value	r_quartile	f_quartile	m_quartile	RFMScore	RFM_Sum	classify_customer
12346.0	529	33	372.86	4	3	3	433	10	Don't lose them
12347.0	367	71	1323.32	1	2	2	122	5	Good customers
12348.0	438	20	222.16	3	3	4	334	10	Don't lose them
12349.0	407	102	2671.14	2	2	1	221	5	Good customers
12351.0	375	21	300.93	1	3	4	134	8	Customer needs attention

Segmentation work has been done. We can do various analyses of the data according to our wishes. I will show a few ones now.

Q1:What are the percentage of each customer segmentation?

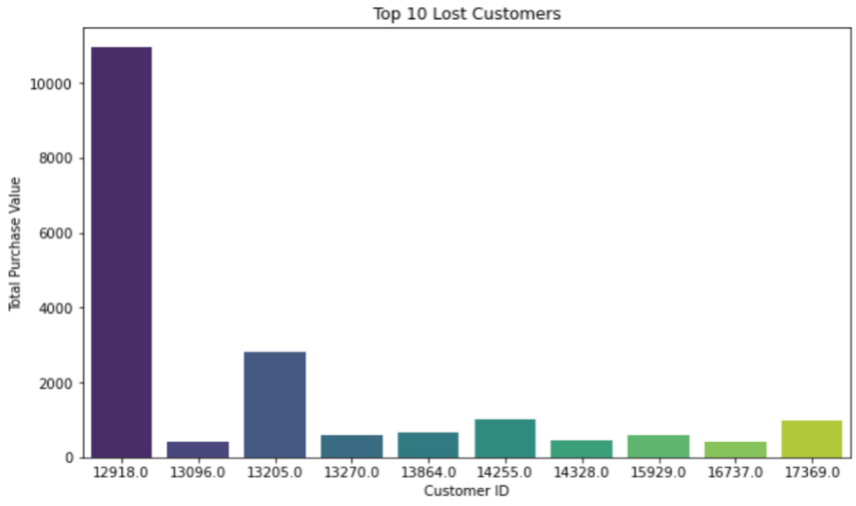


Q2:Who are our loyal customers?



Loyalty of the customers depends on Frequency and Monetary value. Costumers with **high frequency** and **high monetary** value and assign the value high to the loyalty.
In above we see the top 10 loyal customers.

Q3: Who are our lost customers?



In the above analysis, we are identifying lost customers as those who have a recency value of more than 180 days (i.e. 6 months) and a frequency value of 1 (i.e. they have made only one purchase). The resulting plot shows the total purchase value for each of the top 10 lost customers, which can help the retail business to identify the customers who have stopped purchasing and take actions to win them back.

RECOMMENDATIONS

Recommendations

- We should focus more on united kingdom
- We see that we have 8.9% of our customers are lost, so we must attract them again

Thank you : -)
Aishwarya