Report and Analysis

The project implementation has been divided into 4 parts as follows:

1. *Exploratory Data Analysis(EDA)*
2. *Sales per country*
3. *Case study on UK*
4. *Customer Segmentation using RFM*

They have been discussed in detail as shown below.

Part 1:

For the initial step we perform Exploratory Data Analysis(EDA). The EDA is beneficial in understanding the data, discovering patterns, spotting anomalies or null values in the dataset. Following steps have been executed:

- Checking the shape, columns and top 5 head, tail and sample values of each column from the dataset.
- In the next step, we checked if the dataset contained null values and duplicate values or not. It was observed that 'Description' and 'CustomerID' had some null values.
- Then I checked for duplicate values and removed them using '.dropna()' function and reset the indexing of dataset.
- Later on to verify if the dataset was free of all duplicate and values. I again used basic EDA methods and confirmed the data was free of discrepancies.
- To better take insight into the data set, I did correlation for 'Quantity', 'Price' and 'CustomerID'. From the table it could be understood that 'Quantity' had negative correlation with both 'Price' and 'CustomerID' and similar case was observed individually with each of them.

Part 2:

For part 2, we will be visualizing the data to understand the target market for future sales. To understand sales, a new column - 'Total_sales' has been created which is computed with following formula:

```
Total sales = Price x Quantity
```

The graphs visualized are as below:

1. *Total sales v/s Country*

From here, it can be deciphered that United Kingdom(UK) has the maximum sales among the 41 countries. Next being followed by Norway and Poland.

2. *Quantity v/s Country*

From this graph, we could understand that again United Kingdom(UK) had been the country with most quantity of products sold. That was followed by Denmark and France.

Part 3:

Since it can be visualized from the above graphs that United Kingdom(UK) has been the country with most sales and products sold. This led me on to performing a case study on the data for the UK region. As it could be the future target region to increase sales and products sold.

Here, we store the data related to UK region in a new dataframe for it to accessed frequently. We compute that UK has solely contributed to nearly 716115 counts of the 797885 values in the data.

Following are unique statistical values observed for UK region:

- Number of transactions: 40505
- Number of products bought: 4631
- Number of customers: 5410

In the next step, check for the top10 most bought products. This would let the company know which products are being sold the most by analyzing from the 'StockCode' and 'Quantity'. Thus helping in replenishing and stocking up the inventory for those particular products on timely basis in UK.

Next we use Apriopri Algorithm to gain insight into the structed relationships of different item sets of the region. Before we begin with building the model, we do 'one-hot' encoding of the grouped data of UK region as it makes it more suitable and accessible for implementation. Taking forward, we build the Apriopri model to check for most frequent item sets and set minimum support count as 0.01.

When the rules for British transactions are examined more closely, it is discovered that the British buy different coloured tea-plates together. One reason for this could be that the British love tea and frequently collect different coloured tea-plates for different occasions.

Part 4:

In this part we implement Customer Segmentation using RFM(Recency, Frequency, Monetary). It is performed by combining these 3 metrics after they have been calculated. This is helpful in order to analyze the current status of customers and segmenting them according to their scores obtained.

Recency:

It is the information of how long the customer has been receiving the service from a company and its term of membership. Computed as follows:

Last subscription date/ last order date from today

Frequency:

This metric is often helpful in letting a company/organization know for how often the purchases are made by a customer. It results as mentioned below:

Order number/order code

Monetary:

This metric tells about the sum of customer's expenses. It is beneficial in letting an organization know the revenue collected after the services once receives from it. It can be calculated by summing up the expenses made by a customer throughout their life.

Following interpretation can be made from the above RFM implementation. They have been discussed below: Taking an example, let us interpret segment 111:

- This segment has 42 people.
- On average their last purchase was nearly 33 days ago
- Their shopping frequency is 1, so they have 1 purchases.
- A total of approximately $165.9 were spent.