**Case Study 2** (Data-driven Approach)

When it comes to the Retail Industry, meeting the customer needs is what plays a key role for the business to run successfully. So, in this case, I would be focusing on Customer Behavior to study and analyze what the customer needs. Customer Behavior mainly focuses on finding out what the customer purchases. These purchases depend on various factors such as Geographical Location, Community, Season/Weather and Preferences/Interests and coming up with offers/suggestions to meet customer satisfaction

The dataset which I would be using is Online Retail (OnlineRetail.csv) from Kaggle [URL: https://www.kaggle.com/datasets/tunguz/online-retail].

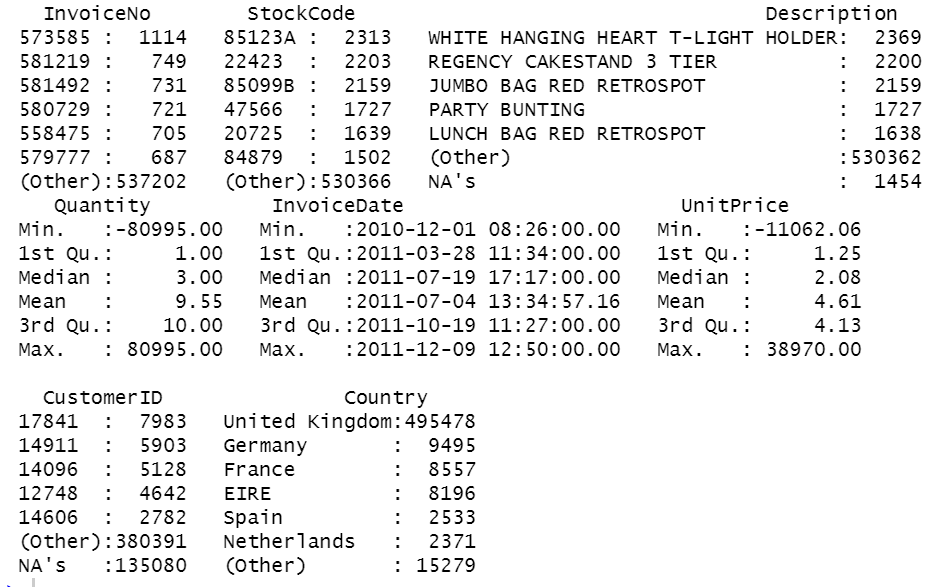
The data is in an Excel Workbook (XLSX) File with 8 Attributes and 541909 Observations.

The 8 attributes are as follows:

1. InvoiceNo - Invoice Number for the transaction/purchase - ***Numeric***
2. StockCode - Unique Code given to the Stock - ***Character / Alphanumeric***
3. Description - Description of the product - ***Character String***
4. Quantity - Quantity of the Products purchased - ***Numeric/Decimal***
5. InvoiceDate - Date on which the Invoice was generated (Purchase Date) - ***Date and Time - YYYY-MM-DD HH:MM: SS***
6. UnitPrice - Price per 1 unit of the Product - ***Numeric/Decimal***
7. CustomerID - ID of the Customer who did the transaction - ***Numeric (With NAs)***
8. Country - Country in which the transaction was made - ***Character String***

I'm planning to divide/filter the data based on the Country and then analyze the data. Instead of working on the whole data, I'll take the countries I'm interested in analyzing and look for insights. I'm planning to use R (4.2.1) and RStudio to work on the dataset. I'll change the data type (for convenience) and then plot graphs with various combinations of attributes.

Following is the Summary of the Dataset – OnlineRetail:



For cleaning and handling the data, *dplyr/tidyr* would be used, and for plotting, *ggplot2* would be helpful (*libraries and functions in R*).

We have NAs (Null Values) in the data for Customer ID. Customer ID is one of the important attributes that should not be empty. Since null values exist, working with them would be less efficient. We won’t be a replace the data for the CustomerID since it is a unique attribute to identify who purchased the data and cannot be replaced with mean or random values.

After working on the data and plotting the graphs, we will be able to study the pattern and trends of the purchases made by the customer. With this, we would be able to analyze the customer behavior and come up with new business strategies to satisfy customer.