## **Data Description**

### **Attribute Information:**

### InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

### 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: Invice Date and time. Numeric, the day and time when each transaction was generated.

### UnitPrice: Unit price. Numeric, Product price per unit in sterling.

### Customer-ID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

### Country: Country name. Nominal, the name of the country where each customer resides.

# ****Create the RFM model (Recency, Frequency,Monetary value)****

### **RFM is a method used for analyzing customer value. It is commonly used in database marketing and direct marketing and has received particular attention in retail and professional services industries.**

****RFM stands for the three dimensions:****

### **Recency – How recently did the customer purchase?**

### **Frequency – How often do they purchase?**

### **Monetary – How much do they spend?**

**Recency** - In order to find the recency value of each customer, we need to determine the last invoice date as the current date and subtract the last purchasing date of each customer from this date.

**Frequency** - In order to find the frequency value of each customer, we need to determine how many times the customers make purchases.

**Monetary** - In order to find the monetary value of each customer, we need to determine how much do the customers spend on purchases.

# ****Clustering:****

### ****K-Means Clustering****

**Clustering is an unsupervised classification techinque to understand the groups of classes in the data. We use the K-means clustering algorithm to determine the ideal segments of customers.**

**KMeans requires the number of clusters to be specified during the model building process. To know the right number of clusters, methods such as silhouette analysis and elbow method can be used. These methods will help in selection of the optimum number of clusters.**

### ****Silhouette score method:****

### **Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters.**

### ****Elbow Method:****

**Elbow is one of the most famous methods by which you can select the right value of k and boost your model performance. We also perform the hyperparameter tuning to chose the best value of k. It is an empirical method to find out the best value of k. it picks up the range of values and takes the best among them. It calculates the sum of the square of the points and calculates the average distance.**

### ****DBSCAN Clustering:****

**DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. DBSCAN is a density-based clustering algorithm that works on the assumption that clusters are dense regions in space separated by regions of lower density.It groups ‘densely grouped’ data points into a single cluster.**

# ****Conclusion****

* ***This project mainly focused on developing customer segments for a UK based online store, selling unique all occasion gifts.***
* ***Using a recency, frequency and monetary(RFM) analysis, the customers have been segmented into various clusters and got a silhoutte score of 0.39 for two clusters***
* ***By applying different clustering algorithm to our dataset, we get the optimal number of cluster is equal to 2.***
* ***The business can focus on these different clusters and provide customer with services of each sector in a different way, which would not only benefit the customers but also the business at large.***