**Online Retail K-Means Clustering**

## **Overview :**

## Online retail is a transnational data set which contains all the transactions occurring between 04/01/2011 and 09/12/2011 for a online retail.

## **Business Goal**

We aim to segment the Customers based on RFM so that the company can target its customers efficiently.

#### **The steps are broadly divided into:**

1. [Step 1: Reading and Understanding the Data](https://www.kaggle.com/hellbuoy/online-retail-k-means-hierarchical-clustering#1)
2. [Step 2: Data Cleansing](https://www.kaggle.com/hellbuoy/online-retail-k-means-hierarchical-clustering#2)
3. [Step 3: Data Preparation](https://www.kaggle.com/hellbuoy/online-retail-k-means-hierarchical-clustering#3)
4. [Step 4: Model Building](https://www.kaggle.com/hellbuoy/online-retail-k-means-hierarchical-clustering#4)
5. [Step 5: Final Analysis](https://www.kaggle.com/hellbuoy/online-retail-k-means-hierarchical-clustering#5)

## **Step 1 : Reading and Understanding Data**

Reading the data set from local drive using pandas and using appropriate libraries for the calculation, analysis and plotting.

## **Step 2 : Data Cleansing**

It mainly involves in removing the null values ,Duplicate values and outliers , I excluded all these types of data from the data set.

## **Step 3 : Data Preparation**

This step includes calculating the revenue as “invoice value” ,changing the data type of each column into the appropriate data type for usage.

#### **We are going to analysis the Customers based on below 3 factors:**

* R (Recency): Number of days since last purchase
* F (Frequency): Number of tracsactions
* M (Monetary): Total amount of transactions (revenue contributed)

## **Step 4 : Building the Model**

### **K-Means Clustering**

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

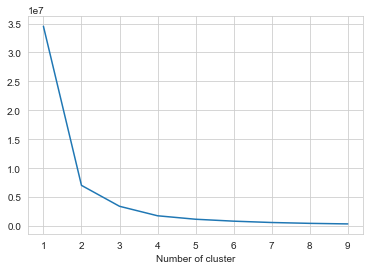
The algorithm works as follows:

* First we initialize k points, called means, randomly.
* We categorize each item to its closest mean and we update the mean’s coordinates, which are the averages of the items categorized in that mean so far.
* We repeat the process for a given number of iterations and at the end, we have our clusters.

### **Finding the Optimal Number of Clusters**

#### **Elbow Curve to get the right number of Clusters**

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.



## **Step 5 : Final Analysis**

### **Inference:**

K-Means Clustering with 3 Cluster segments worst, medium, and best

* Customers with Cluster segments best are the customers with high amount of transactions as compared to other customers.
* Customers with Cluster segments medium are frequent buyers.
* Customers with Cluster segments worst are not recent buyers and hence least of importance from business point of view.

Choosing the values of k as 3 we can get the following data for frequency recency and monetary value

