Overview

* **Objective:** The objective to be achieved from the data set is to perform the customer segmentation based on the Recency, Frequency and Monetary values of the customers in the given dataset.
* **Methodology:** CRISP-DM

We create a Machine Learning pipeline, which consists of the following steps:

* 1. Problem Understanding: Online retail market is huge sector and having customers across the globe. We need to analyze customer purchase patterns using RFM scores and prioritize/focus on the customers where more revenue is generated.
  2. Data Understanding: This includes the collection, description, and exploration of data. The dataset *Online Retail.xlsx* is used for the Analytics. The data generated is from different regions. As we are performing clustering techniques, we need not divide it into training/testing data. There are 8 features in the dataset.
  3. Data Preparation: This includes the cleaning of data. Additionally, feature construction and feature selection are performed.
  4. Modelling: This involves training a model using specific algorithms. In this experiment, we train two models - a K-Means clustering model and Expectation Maximization models.
  5. Evaluation: Contrary to supervised learning where we have the ground truth to evaluate the model’s performance, clustering analysis does not have a solid evaluation metric that we can use to evaluate the outcome of different clustering algorithms. Moreover, since kmeans requires k as an input and doesn’t learn it from data, there is no right answer in terms of number of clusters that we should have in any problem. Sometimes domain knowledge and intuition may help but usually that is not the case. In the cluster-predict methodology, we can evaluate how well the models are performing based on different K Clusters since cluster are used in the downstream modeling.

We have used elbow method in determining which k value should be taken.

# Methodology

* We use two classifiers in our experiment:

**K-Means Clustering method**:

The [K-Means](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans) algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the *inertia* or

within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to large number of samples and has been used across a large range of application

areas in many different fields.

**Expectation Maximization Model:**

The [GaussianMixture](https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html" \l "sklearn.mixture.GaussianMixture" \o "sklearn.mixture.GaussianMixture) object implements the [expectation-maximization](https://scikit-learn.org/stable/modules/mixture.html#expectation-maximization) (EM) algorithm for fitting mixture-of-Gaussian models. It can also draw confidence ellipsoids for multivariate models, and

compute the Bayesian Information Criterion to assess the number of clusters in the data. A [GaussianMixture.fit](https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html" \l "sklearn.mixture.GaussianMixture.fit" \o "sklearn.mixture.GaussianMixture.fit) method is provided that learns a Gaussian Mixture Model from train data. Given

test data, it can assign to each sample the Gaussian it mostly probably belong to using the [GaussianMixture.predict](https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html" \l "sklearn.mixture.GaussianMixture.predict" \o "sklearn.mixture.GaussianMixture.predict) method.

# Feature Engineering Techniques

* **Feature creation:** We have created a new feature called amount\_spent which is derived from quantity and no. of units in the data set. It gives us the insights about the customer spending on the orders. Also, created Recency, Frequency and Monetary features from the data set.

## Features removed: We removed no features from out data set as there are minimal set of features and we cannot remove any features as correlation coefficient is not greater than .1 between any two features except for country and customer id which is 0.39.

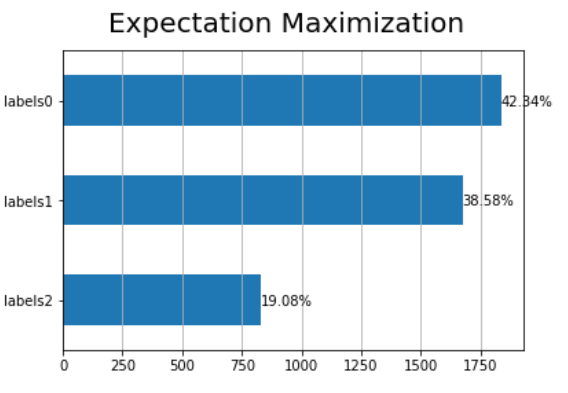
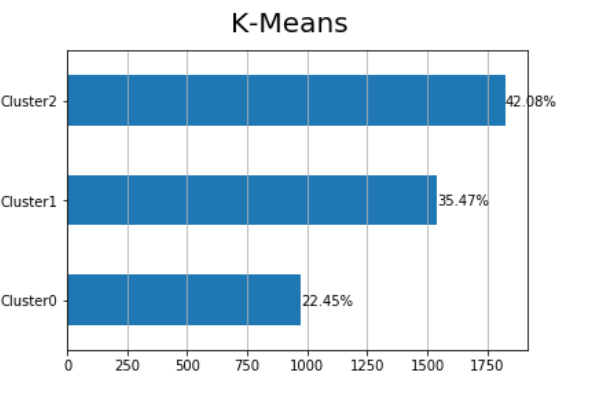
* **Feature ranking:** NA
* **Class imbalance treatment:** We haveapplied log transformation on RFM graphs to make the right skewed distribution to normal.

Dataset

* **How many features**: There are a total of 8 features.
* **Size of the dataset**- There are 541909 tuples in total.
* **Multiple files** - One File
* **What kind of data** – Combination of numerical and nominal data.
* **Balanced or imbalanced** – The distribution of RFM attributes which are extracted from the given features are right skewed.
* **Missing data and preprocessing challenges:** There are few NA customer IDs in the dataset which cannot be replaced with any technique, hence we have eliminated such entries. Few negative quantity numbers that are associated with invoice number which starts with C are considered as cancelled invoices and these records are removed as well.

Results

* **Plot of the graphs:**



* **Conclusion:**

From elbow technique, at K=3, the sum of squares has decreased drastically. K = 3 acts as an optimal value, hence we can observe that there is similarity in the results of K-means and EM models.

* **Our Observation:**

It is interesting to know that the InvoiceDate is an important feature in the entire dataset as the derived attributes recency & frequency are more dependent on it.