

| **Title: Implementation of Clustering algorithm.** |
| --- |

**Aim:** To understand & implement clustering algorithms using python.

**Expected Outcome of Experiment:**

CO3: Analyse unsupervised learning methods.

**Books/ Journals/ Websites referred:**

1. https://wikipedia.com/kmean-clustering
2. **3rd Edition**

**Theory of K-Mean Clustering Algorithm:**

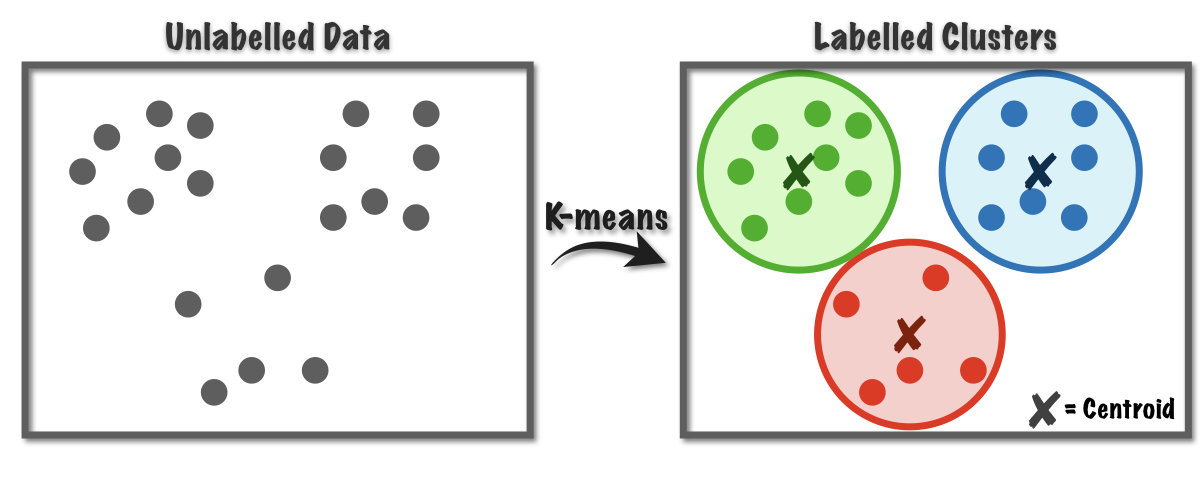
K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



**Problem Definition:**

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

**What Is Recency, Frequency, Monetary Value (RFM)?**

Recency, frequency, monetary value is a [marketing](https://www.investopedia.com/terms/m/marketing.asp) analysis tool used to identify a company's or an organization's best customers by using certain measures. The RFM model is based on three quantitative factors:

**Recency**: How recently a [customer](https://www.investopedia.com/terms/c/customer.asp) has made a purchase

**Frequency**: How often a customer makes a purchase

**Monetary Value**: How much money a customer spends on purchases

The RFM (Recency, Frequency, Monetary value) based model of customer value for finding the customer segments.

**Dataset: https://archive.ics.uci.edu/ml/datasets/online+retail**

**Perform the following tasks:**

1. Exploratory Data Analysis & Visualizations
2. Data Preprocessing if required
3. Customer segmentation for Recency vs Monetary Value
4. Cluster Segmentation for Frequency vs Monetary value
5. Cluster Analysis

**API/Function used:**

1. **sklearn.cluster.KMeans**:

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.

1. **sklearn.preprocessing.StandardScaler**

Standardize features by removing the mean and scaling to unit variance.

The standard score of a sample x is calculated as:

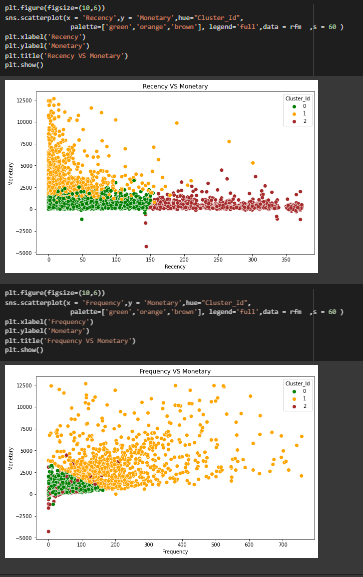
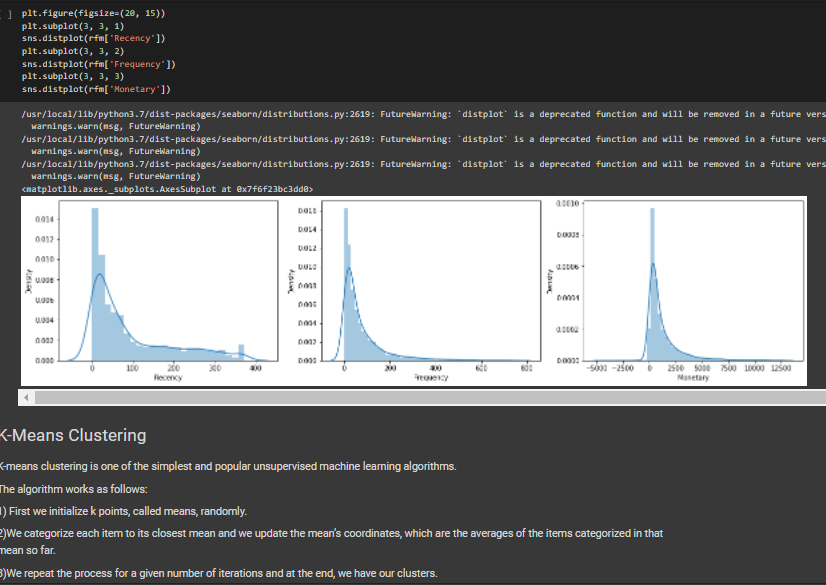
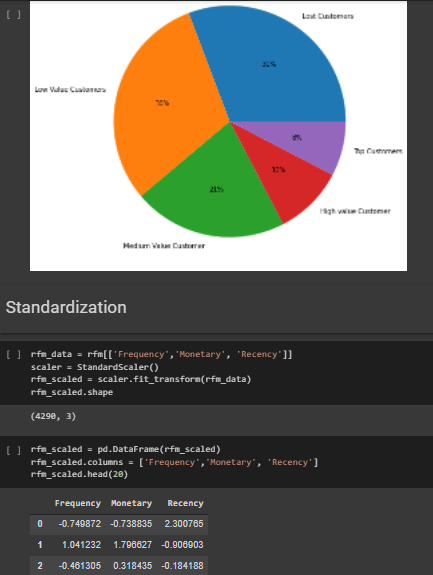
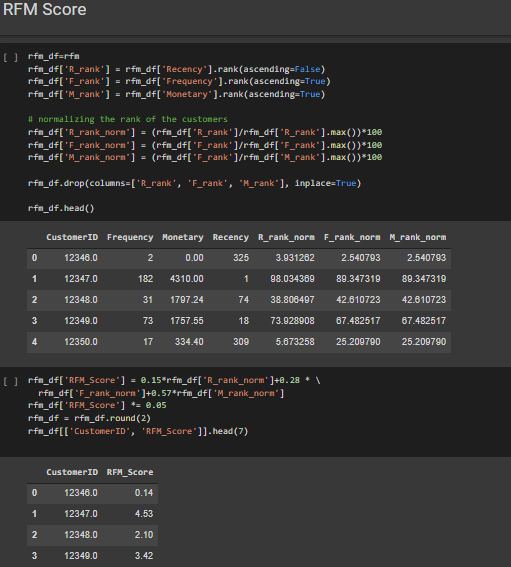
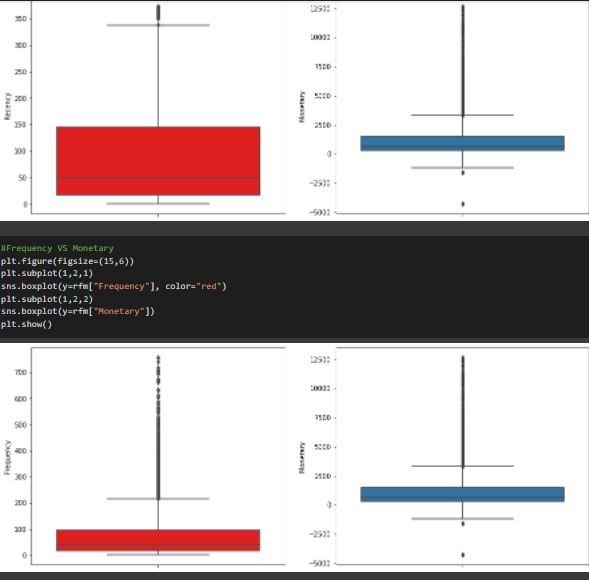
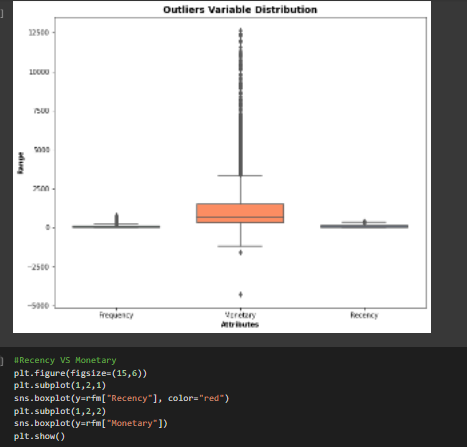
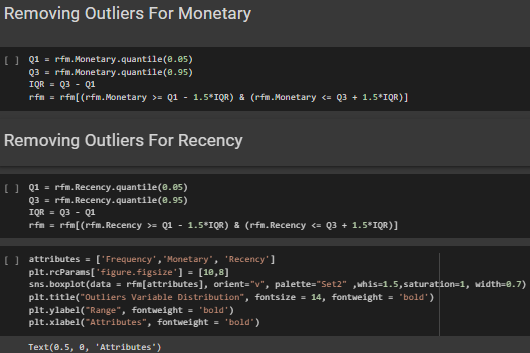
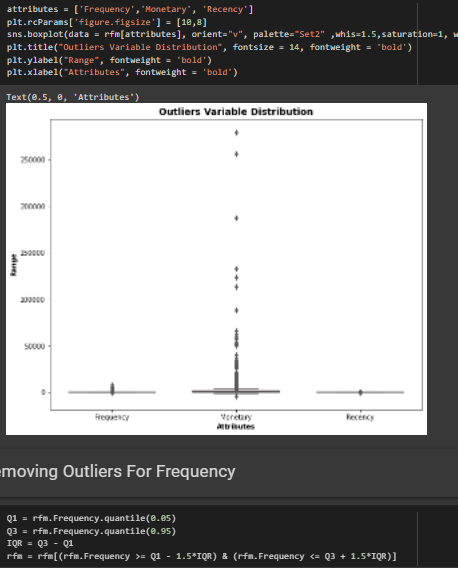
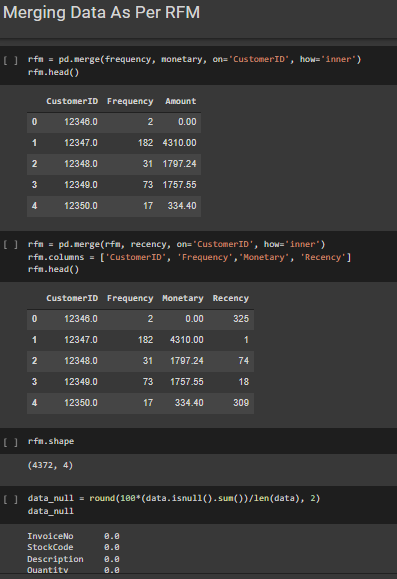
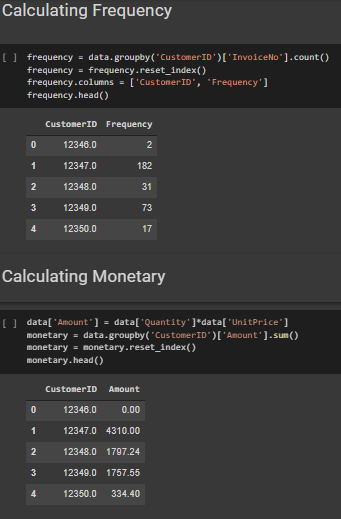
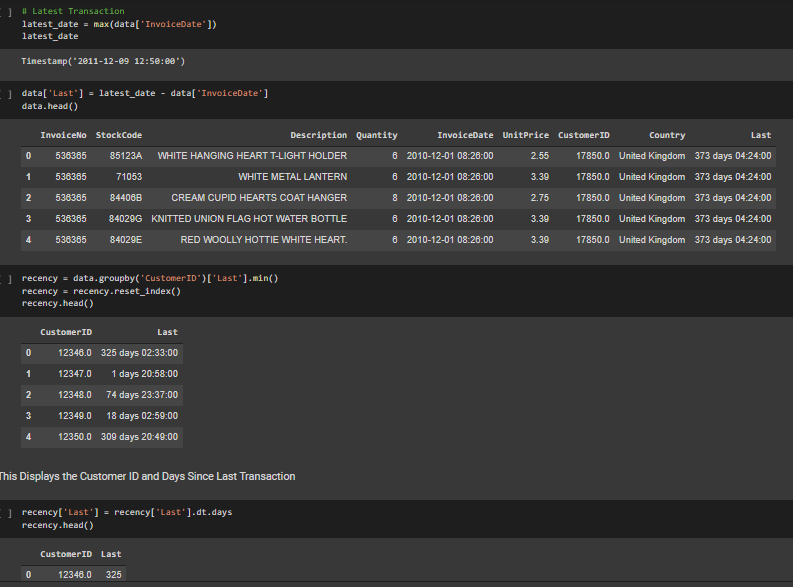
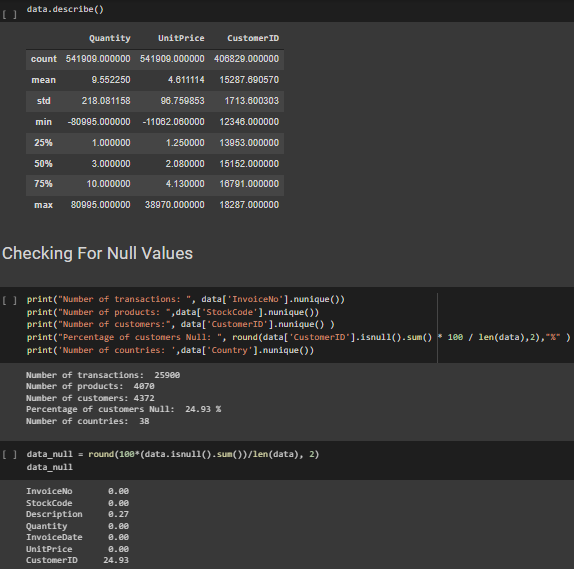
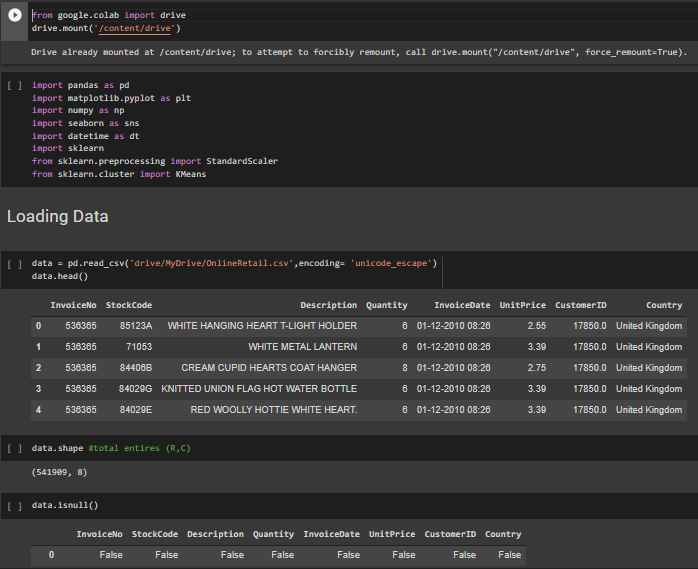
z = (x - u) / s

where u is the mean of the training samples or zero if with\_mean=False, and s is the standard deviation of the training samples or one if with\_std=False.

**Notebook Link:**

https://colab.research.google.com/drive/12Cwyb3wnNyEHx4aPZ3mdp0oPBNbgUfYN

**Code & Output:**



**Conclusion:**

We successfully understood & implemented the concept of clustering algorithm over a given dataset in Colab notebook using python.