Data Mining & Data Warehousing

BSCS - 633

Clustering Model Report

Submitted By:

Mohammad Taha	**B18101055**

Course Incharge:

Miss Maryam Feroze

Online Retail Dataset

Data Description:

Transactions Count	:	541909
No# of Features	:	8
Feature Description	:	

- **Invoice Number (object)**
- **StockCode (string)**
- **Decsription (string)**
- **Quantity (int64)**
- **InvoiceDate (datetime)**
- **UnitPrize (float64)**
- **CustomerID (float64)**
- **Country (string)**

Data Pre-Processing:

- Data Transformations (If needed):

Duplicate Records	:	No Duplicate Records
Null or NA Values	:	Dropped 135080 Records
Handling Typos or Strange Names	:	Converted Them

For Example:

- For the last purchase date converted into date and take the days diff b/w the latest transaction and that transaction for customer recency.

- CustomerID is converted for our ease.

- Data Preparation:

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

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

- Column Outliers:

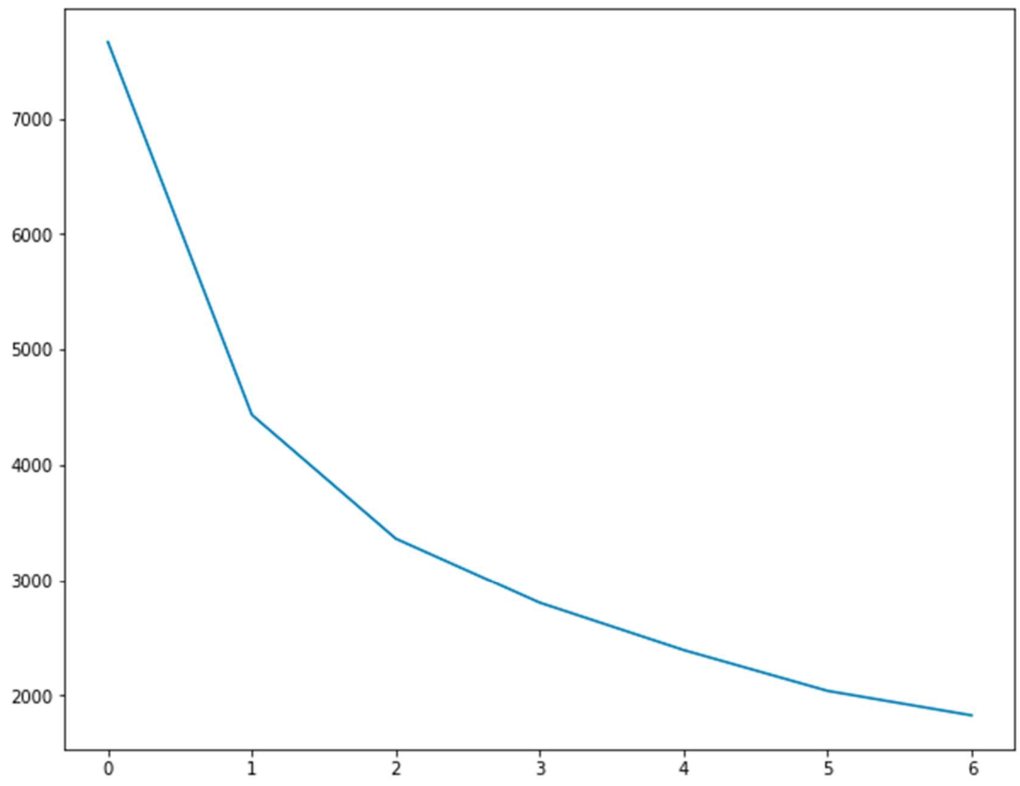
Removing Statistical Outliers (Using Q1 = 0.05 and Q3 = 0.95)

Scaling Values, using Standarization (mean = 0, sigma = 1)

Modeling:

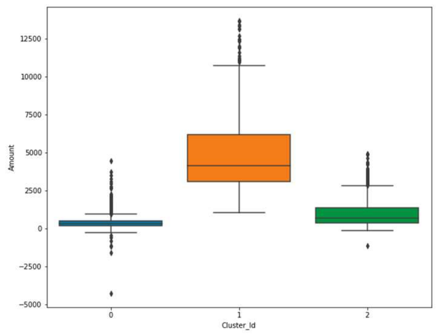
Under the process of modelling, I tried k Means Clustering and also implemented Hierarchical Clustering.

- K Means Clustering:

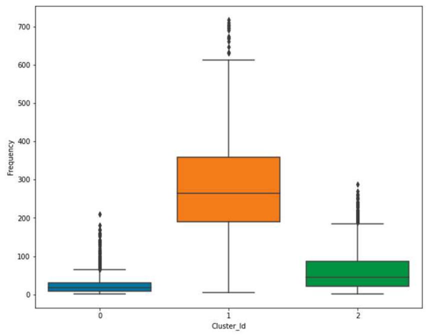


Note: Impact of 3 clusters is very on the data Clustering with numOfClusters = 3

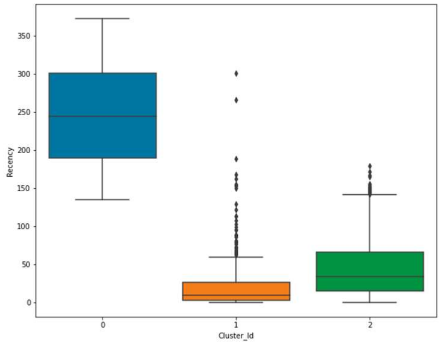
Box Plot to Visualize Amount



Box Plot to Visualize Frequency

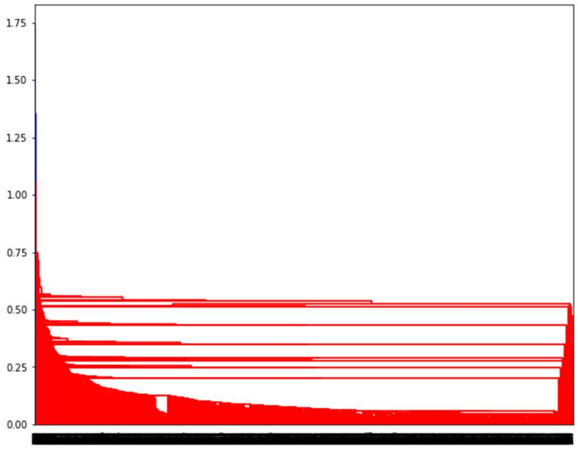


Box Plot to Visualize Recency

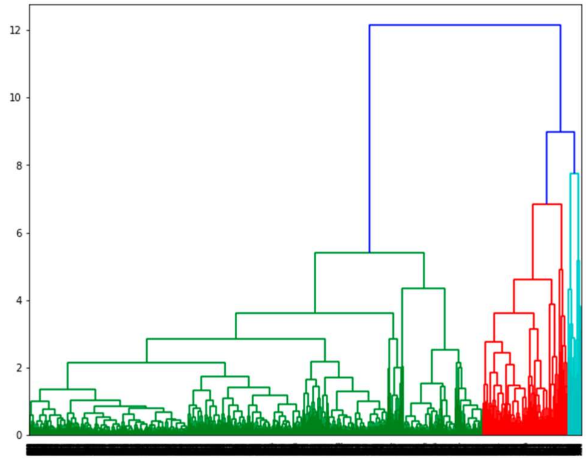


- Hierarchical Clustering:

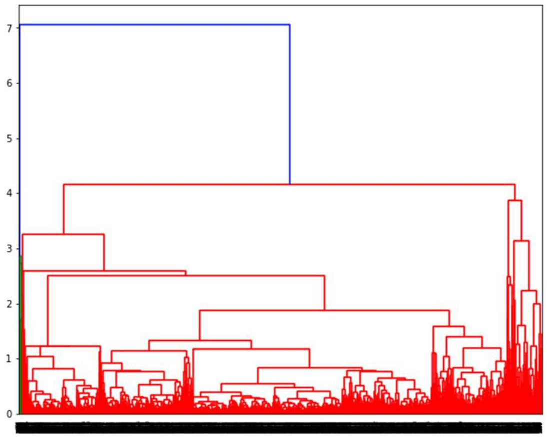
Single Linking:



Complete Linking:

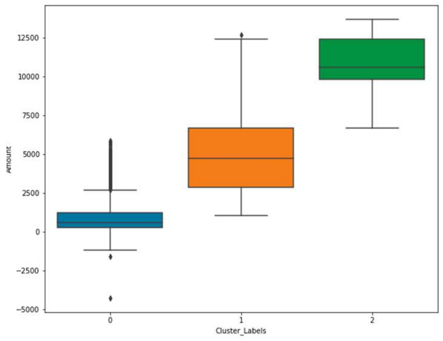


Average Linking:

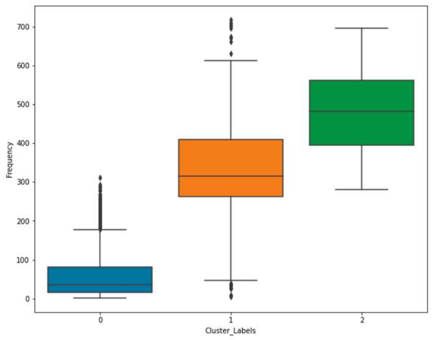


Cutting the Dendrogram based on K (3 Clusters):

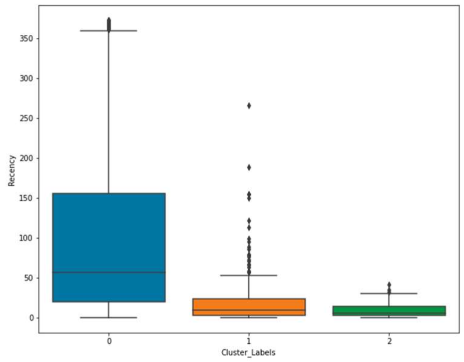
Box Plot to Visualize Amount



Box Plot to Visualize Frequency



Box Plot to Visualize Recency



Inference:

K-Means Clustering with 3 Cluster Ids

- Customers with Cluster Id 1 are the customers with high number of transactions as compared to other customers.
- Customers with Cluster Id 1 are frequent buyers.
- Customers with Cluster Id 2 are not recent buyers and hence least of importance from business point of view.

Hierarchical Clustering with 3 Cluster Labels

- Customers with Cluster_Labels 2 are the customers with high number of transactions as compared to other customers.
- Customers with Cluster_Labels 2 are frequent buyers.
- Customers with Cluster_Labels 0 are not recent buyers and hence least of importance from business point of view