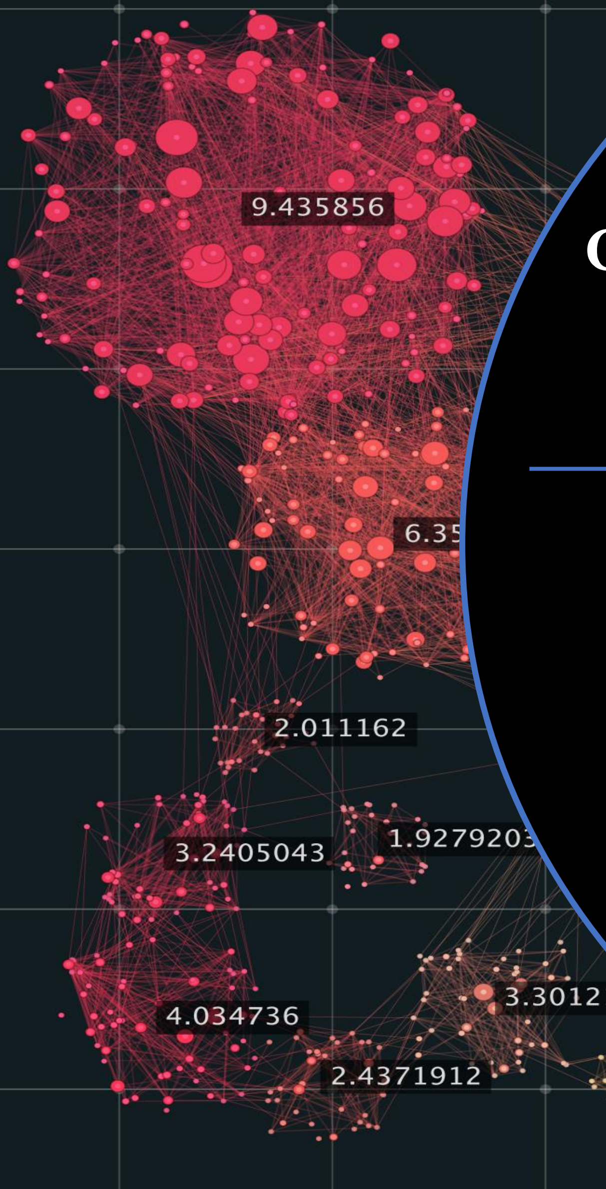


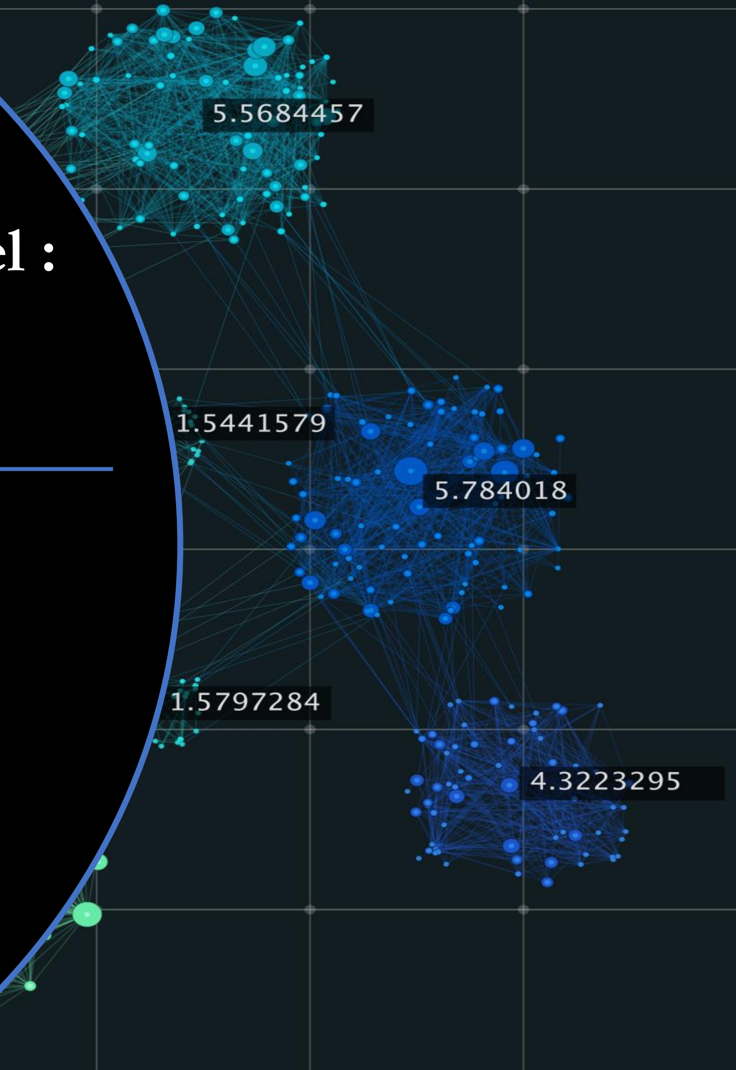
Customer Segmentation using RFM Model :

An Application of k-means++ clustering algorithm

NAME : Amrita Bhattacherjee

DEPARTMENT : Statistics (STSA)
SUPERVISOR – Dr. Ayan Chandra

St. Xavier's College (Autonomous), Kolkata



CONTENTS :

INTRODUCTION

What is Customer Segmentation?

Finding discernible patterns or clusters from
customer data to obtain valuable marketing insights

- **Obtain Customer Transaction Data from a business enterprise**
- **Data cleaning, data pruning, feature engineering**
- **Perform Clustering**
- **Present valuable insights**

Objective :

- To cluster the customers based on their value to this company.
- This will enable us to obtain important insights on the behaviour of customers
- Formulate targeted strategies to retain valuable customers and attract new customers

Data Description

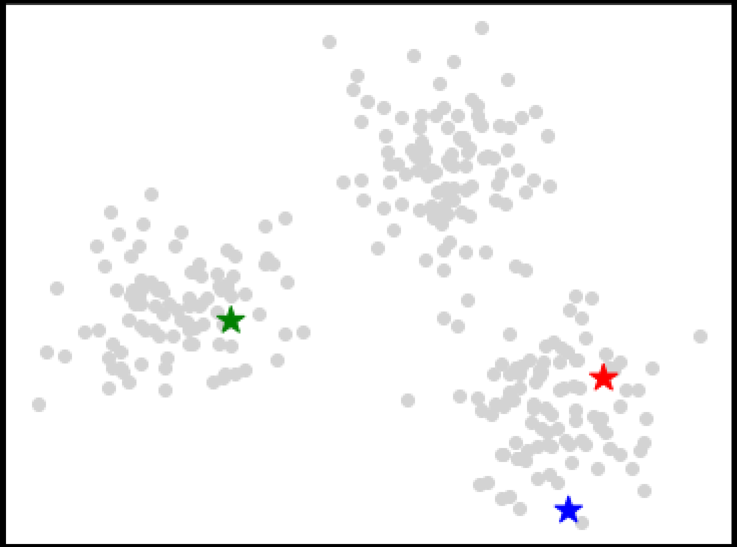
The first 3 rows are shown below to understand the format of the data -

	INVOICE_DATE	CUST_ACCOUNT	Amount
0	2019-06-15 00:00:00	[redacted]	400.0
1	2019-06-15 00:00:00	[redacted]	221.0
2	2019-09-08 00:00:00	[redacted]	0.0

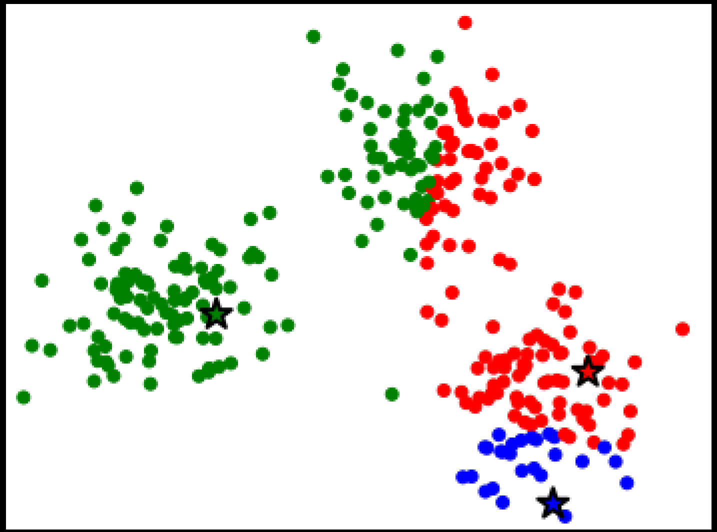
- Transaction Details of 5000 customers of the enterprise for the year 2019-2020
- Amount column also contains non-positive values.
- The date is followed by a time stamp, which is information that is not needed for this study.
- The Customer ID’s are anonymised later in the project.

K-Means Clustering

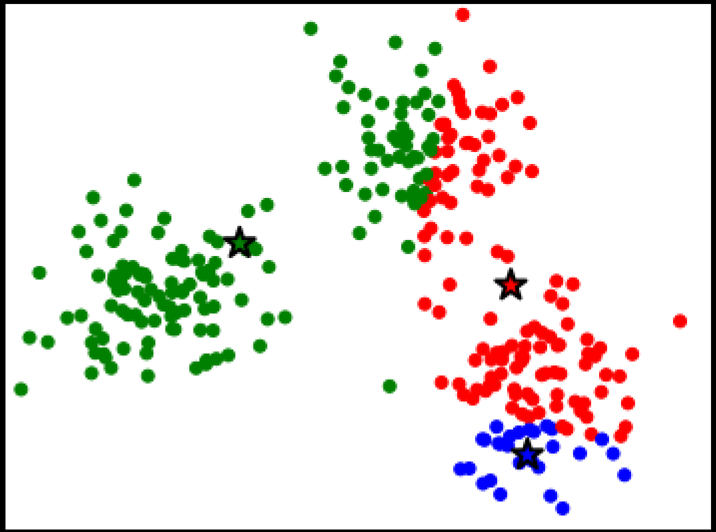
- Unsupervised machine learning algorithm
- The term 'k-means' was first coined by James MacQueen in 1967
- **Tries to find *k* number of representative centers or prototypes from a given set of data, such that each prototype is used to identify a cluster and any given datapoint is assigned to one of the *k* prototypes obtained from the algorithm**
- We will refer to these prototypes as 'cluster centers' for future reference



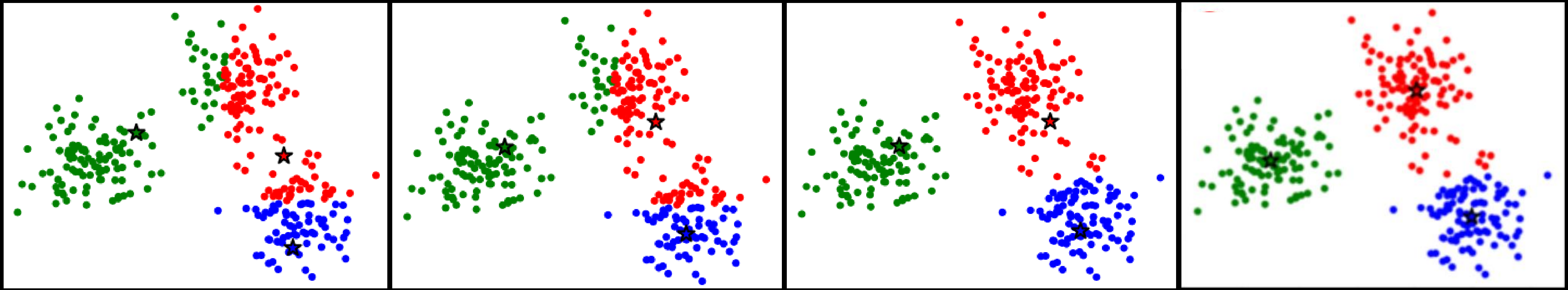
Step 1 : Randomly select k points from the dataset and treat them as initial cluster centers.



Step 2 : Iterate through the entire dataset and assign each datapoint to its nearest cluster center.



Step 3 : Recompute the cluster centers by calculating the mean of all observations assigned to that particular cluster.



Step 4 : Repeat Steps 2 and 3 iteratively until cluster assignments stabilise, that is, they do not change anymore.

Mathematically,

Dataset : $\{x_1, x_2, x_3, \dots, x_n\}$ where each x_i is a p-dimensional point on the Euclidean Space.

Define –

- $\{\boldsymbol{\mu_1}, \boldsymbol{\mu_2}, \boldsymbol{\mu_3}, \dots, \boldsymbol{\mu_k}\}$: k cluster centers, each having p components. These centers are initially randomly selected from the dataset
- $\boldsymbol{I_{ij}}$: an indicator variable $\in \{0, 1\}$ such that –

$$I_{ij} = \begin{cases} 1 & \text{, if the } i^{th} \text{ datapoint is assigned to the } j^{th} \text{ cluster} \\ 0 & \text{, otherwise} \end{cases}$$

$$i = 1\ to\ n \text{ and } j = 1\ to\ k$$

PROBLEM : **Minimise** $\boldsymbol{J = \sum_{i=1}^{n} \sum_{j=1}^{k} I_{ij} \left\| x_i - \mu_j \right\|^2}$

STEP - 1

Minimise J with respect to $\{I_{ij}\}_{\substack{i=1\ to\ n\\ j=1\ to\ k}}$, keeping $\{\mu_j\}_{j=1\ to\ k}$ fixed. This is done by assigning 0s and 1s in the following way –

$$I_{ij} = \begin{cases} 1, & if\ j = arg\ min_{m=1\ to\ k}\left(\|x_i - \mu_m\|^2\right) \\ 0, & otherwise \end{cases}$$

STEP - 2

Minimise J with respect to $\{\mu_j\}_{j=1\ to\ k}$ keeping $\{I_{ij}\}_{\substack{i=1\ to\ n\\ j=1\ to\ k}}$ fixed. Since J is quadratic in μ_j, we can minimise J by simply by equating $\frac{\partial J}{\partial \mu_j}$ to zero and solving for μ_j. By doing this, we obtain –

$$\mu_j = \frac{\sum_{i=1}^{n} I_{ij} x_i}{\sum_{i=1}^{n} I_{ij}}$$

K-Means ++

- The initial choice of centers heavily influences the results
- If the k initial centers are chosen very close to each other, not only will there be a higher possibility of being stuck at a bad optimum value, but it will also take longer (a greater number of iterations) to reach the desired result.

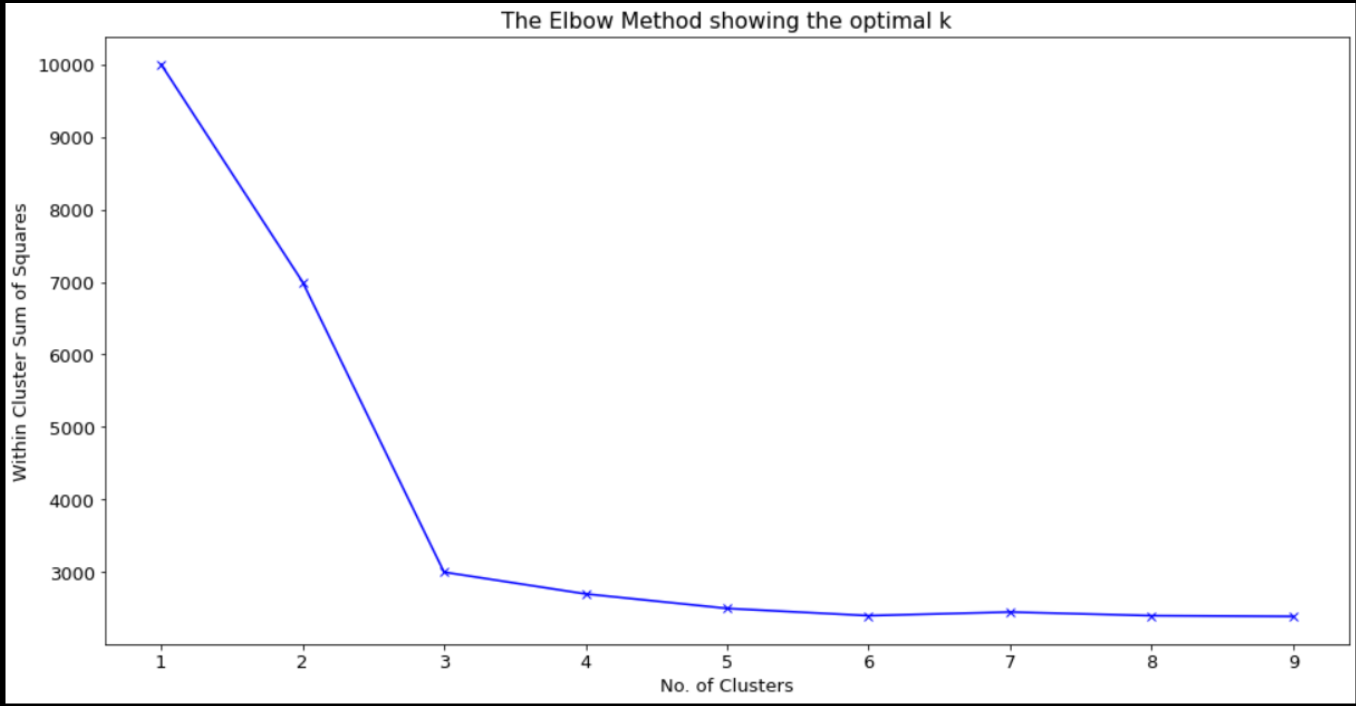
Step 1 : Calculate distances of every point from this randomly chosen cluster

Step 2 : Assign a probability of choosing each point as the next center, where this probability is proportional to the square of the distance of the point from the nearest center

Step 3 : Select the next center by sampling it based on the probabilities calculated in Step 2

Elbow Method

- Used for choosing k
- Plot the final value of J (sum of squares within clusters) for a range of choices of k
- Select that value of k where the curve faces a sharp bend



From the adjacent graph,
k = 3 is a feasible choice

Feature Engineering

RFM MODEL

We cluster the data based on **RECENCY, FREQUENCY** and **MONETARY VALUE** of each customer

Step 1 : From the data, the unique customer ID's are extracted and stored in a list.
Step 2 : We run a loop through these unique ID's and extract information for each customer.

RECENCY

For 'Recency', we first find the most recent date in the entire dataset, say D_0. We then calculate the difference in days between D_0 and D_i, where D_i is the last date of transaction for the i-th customer ($i = 1, 2, \ldots, 5000$)

FREQUENCY

For each customer, we calculate the number of transactions, and store them in feature vector 'Frequency'.

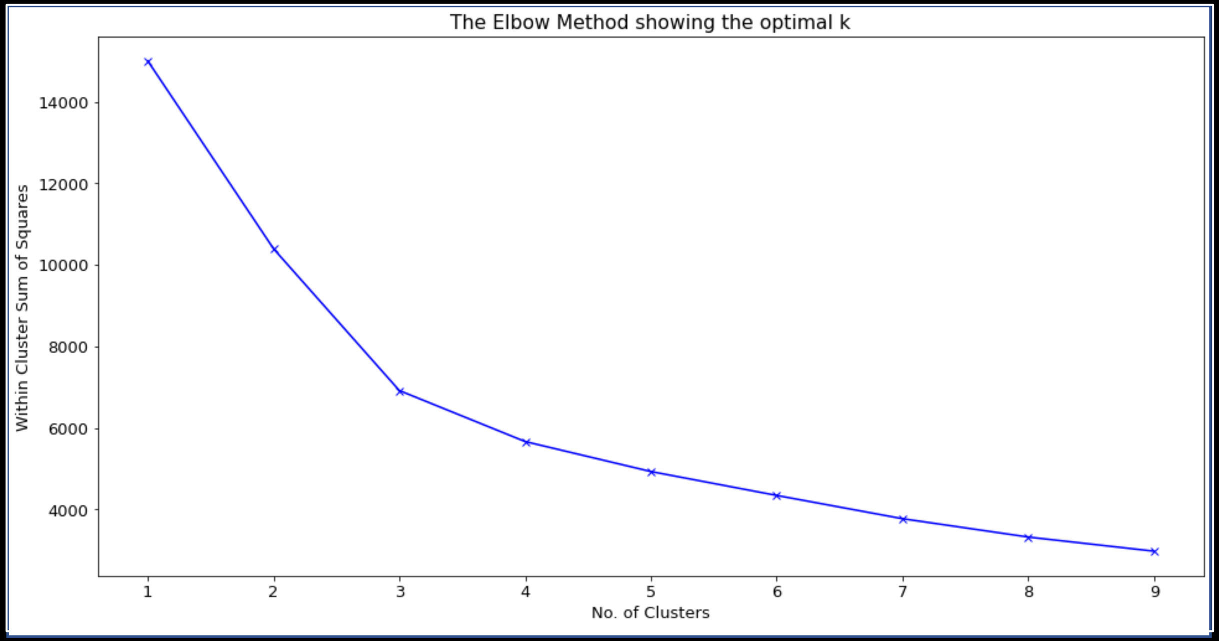
MONETARY VALUE

For each customer, we calculate the total amount transacted and store them in feature vector 'Monetary Value'.

Step 3 : So, after step 2, we have three feature vectors – Recency, Frequency and Monetary Value. We now create 3-dimensional data points for the i-th customer by taking the i-th observation in each of the feature vectors.

Step 4 : Then, the dataset is scaled using StandardScaler( ) method in Scikit-learn. The final dataset, therefore, has 5000 points, where each point is a 3-dimensional vector. This dataset is now ready for clustering.

Step 5 : We implement the k-means clustering algorithm (standard implementation from Scikit-learn, with k-means++ initialisation) to the dataset obtained in Step 3.



We take,

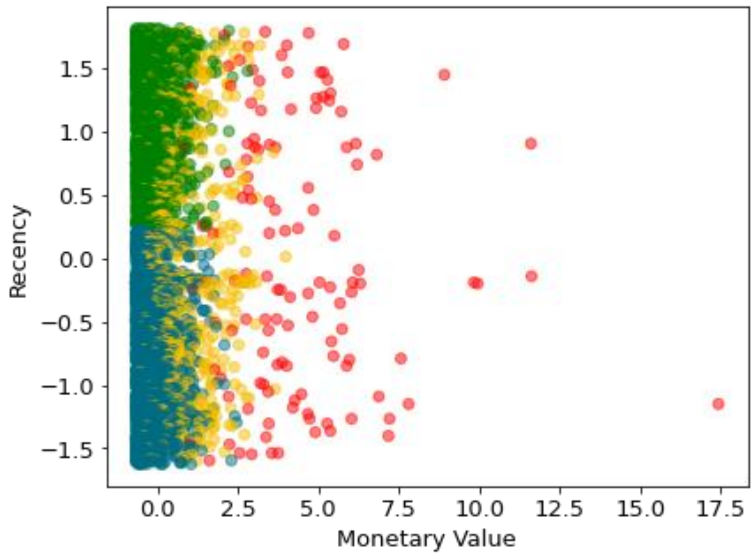
k = 4

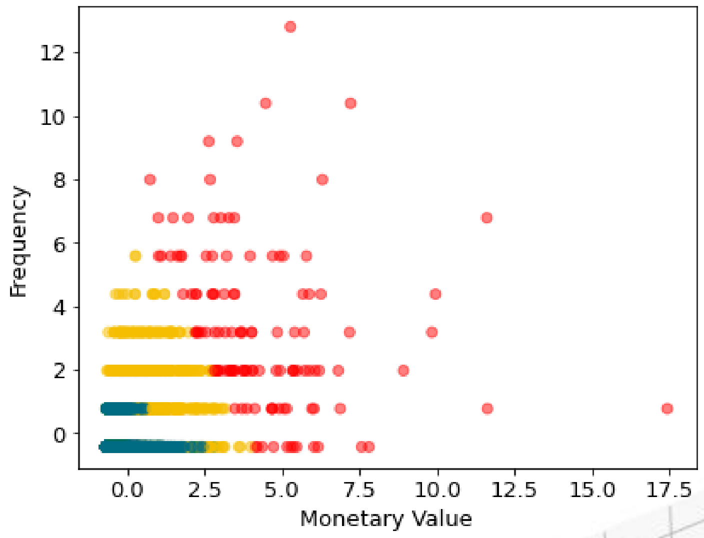
The algorithm specifications are then given below –

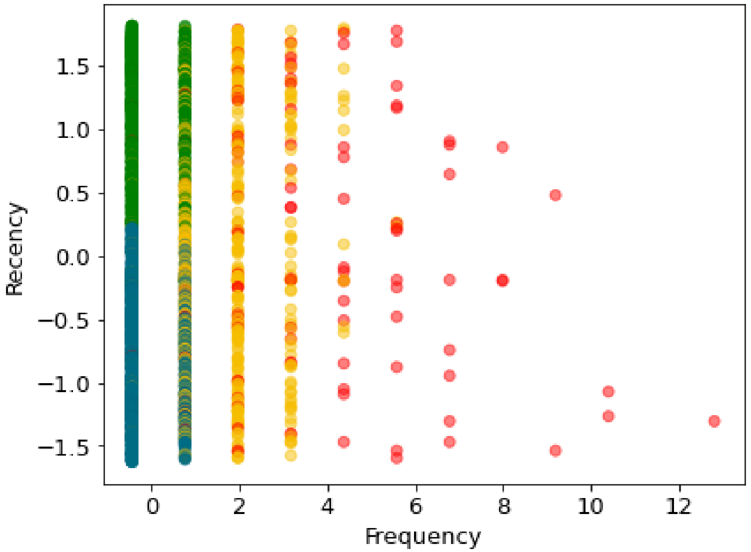
n_clusters = 4

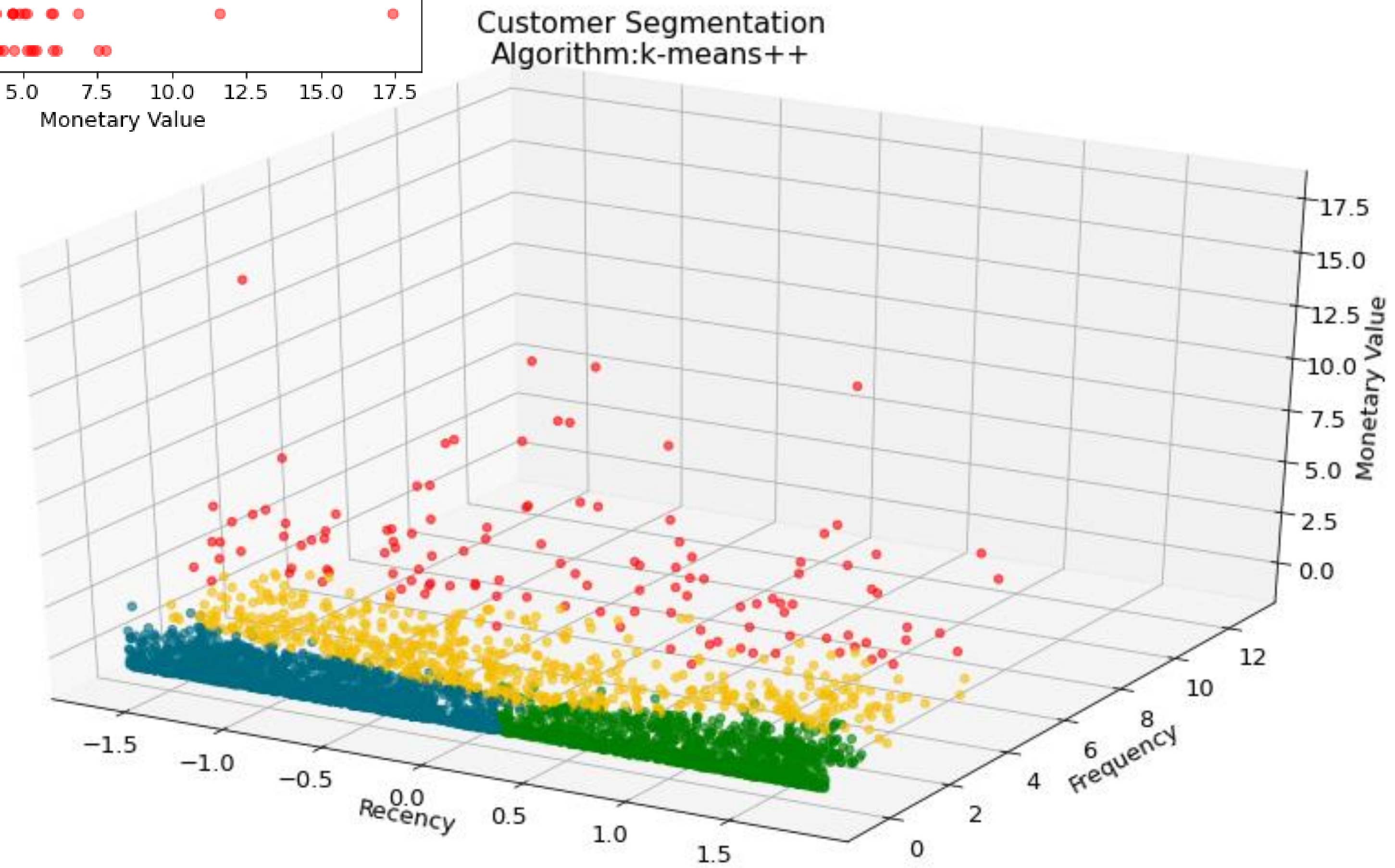
random_state = 0

Cluster results

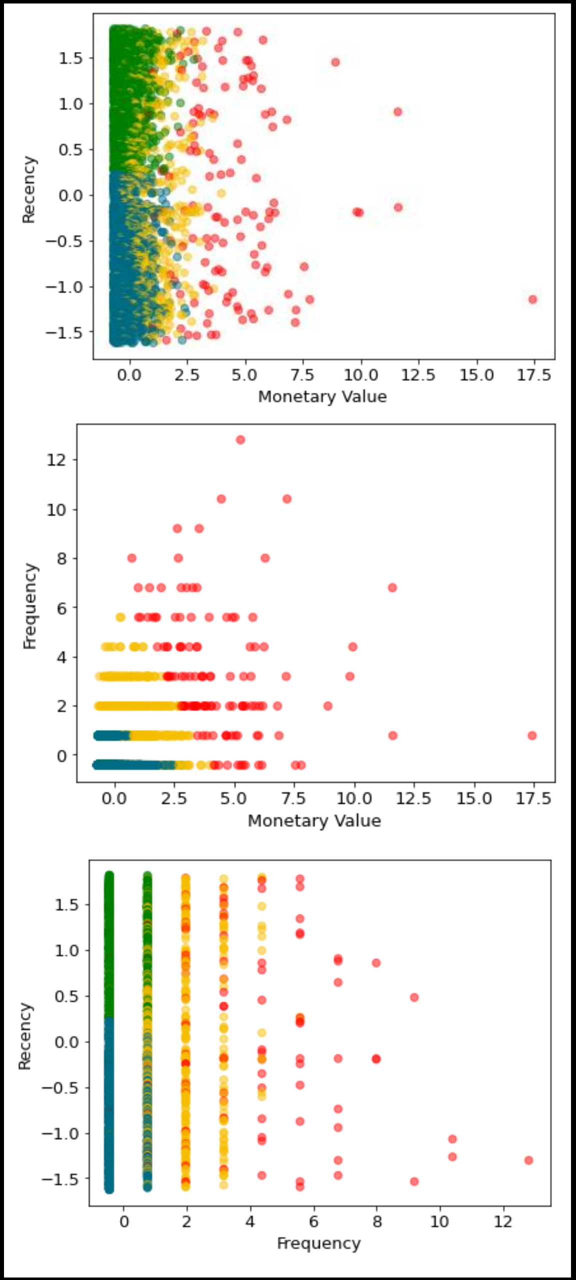








Cluster Analysis



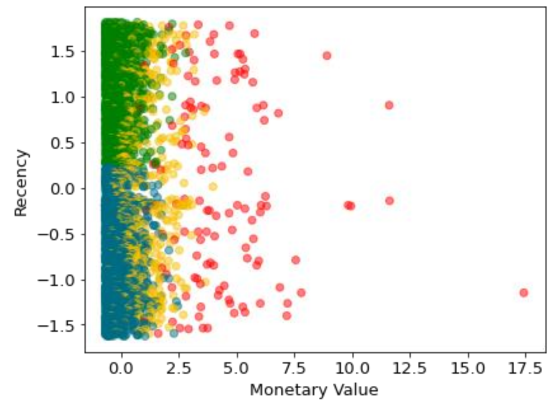
The 4 clusters are distinguished by 4 different colours in the plots. For further reference, we denote the Blue, Green, Yellow and Red clusters by 1, 2, 3 and 4 respectively.

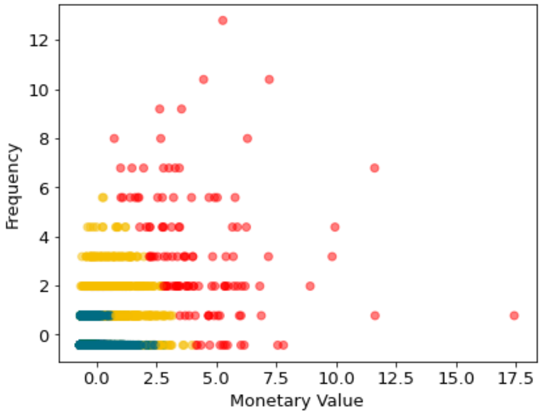
1. Clusters 1 and 2 correspond to **low Monetary Value** and **low Frequency** customers. There is also a logical explanation to both of these features existing in low value together : if a customer does not purchase often, then their monetary transaction is bound to be low. Only exceptions are one-time customers who purchased heavy orders. These exceptions fall in Cluster 4, as indicated by the lower-right points in the second figure.

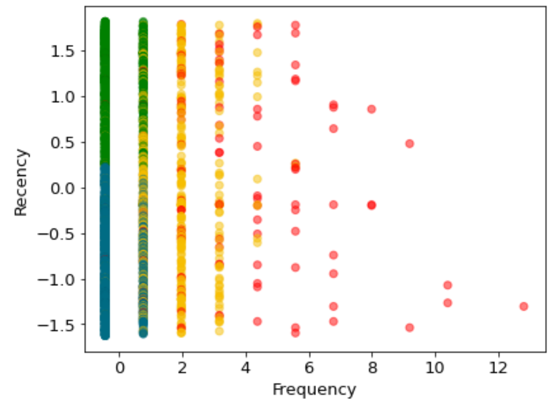
2. **Recency** does not seem to act as a major factor in distinguishing Clusters 1 and 2 from 3 and 4. However, Clusters 1 and 2 are <u>themselves</u> different in terms of Recency alone. This is further verified by the adjacent figures, where the top figure clearly indicates a rough margin between Clusters 1 and 2, and this differentiable margin vanishes in Figure 2, when Recency is not plotted. In Figure 2, Clusters 1 and 2 simply overlap each other.

3. In terms of customer value, Cluster 1 falls above Cluster 2. Cluster 2 contains customers who are not frequent, not recent and not heavy spenders. Cluster 1 contains customers who are not frequent, and their few transactions have not been very heavy in terms of monetary value. However, they are recent, indicating that regular targeted ads and other marketing strategies might consolidate their loyalty.

Cluster Analysis (Contd.)







4. Cluster 3 contains low to medium frequency customers with comparatively higher amounts of transactions than 1 and 2, but lower than 4. This cluster could be treated as potential high-value customers, i.e., customers who could become highly loyal with the correct strategies.

5. Cluster 4 only contains customers with very high monetary transactions. The most valuable customers within this cluster are the ones lying on the bottom-right of figures 1 and 3, and top-right of figure 2. The other points within this cluster represent less frequent, less recent customers with heavy transaction amount. A prospective way of treating these points is by capitalising on the fact that they spent a heavy amount on their transaction (however less recent or less frequent), thereby indicating that they must have some loyalty, for example, they might trust the brand value. Therefore, targeted differentiated strategies can turn them into frequent customers.

Further Scope

- **Performance measures :** Gap-statistic methods, silhouette scores.
- **More information :** With more high-dimensional data, containing information on the purchased products, demography and other aspects, one can study product preferences within each segment. This will help in making decisions that relate to promotional offers and targeted advertisements. Therefore, these studies can reveal information that helps maximising company revenue.
- **Effectiveness of the study :** An interesting avenue of further research could be the study of how the targeted strategies helped in increasing customer influx and retaining high-value customers. This could be done by collecting transaction data over multiple economic years and analysing how the customer segments changed over the years, as new strategies were employed.

THANK YOU

I would like to acknowledge my deepest gratitude to the entire Faculty of the Department of Statistics, my supervisor Prof. Ayan Chandra and the administration of St. Xavier's College (Autonomous), Kolkata.