Customer Segmentation using K-Means Clustering

G.Mitra Datta-AP19110010552,V.Mahesh-AP19110010388,
A.Harshith-AP19110010515, M.Roop Sagar-AP19110010376

1.Abstract

Customer segmentation is one of the key aspects in business driven companies . Identifying potential customers is important as to know where to put the company's efforts.Majority of the companies run on the ability to retain customers ,due to the abundance of products there is a high chance of risk customers may get confused and couldn't find the product desired. Customer segmentation focusing on the relationships not only between customers to customer,but also the interactions between customer and products can achieve high promising results in the outcome. In this paper, we are proposing to use K-means Clustering algorithm for solving the customer segmentation.The algorithm is applied on a dataset, which is first preprocesed.Data preprocessing is an essential stage in the Data mining process. Preprocessing the data improves the quality of data thereby improving the outcome of a data mining task.First we will apply data integration methods such as normalization (Min-Max,Z-score) which ensures that all the attributes in the dataset are given equal importance. Later we will remove redundancies in the dataset using correlation analysis. Now the curated dataset is ready , which can be used for customer segmentation. We will use the K-Means clustering algorithm to segment the data set into K dissimilar clusters depending upon the similarities of the attributes.Choosing the 'K' value in the algorithm plays an important role in segmenting clusters. Inorder to choose the optimal value of K, different methods such as Silhouette or Elbow method. We aim to show the results after applying the above mentioned techniques.

2.Introduction

In the current era of Business industry especially in the e-commerce and retail sectors, the competition between companies is cumulatively increasing along with a huge base of customers.Accordingly the data of customers has also been increasing which lead to ideas to use this data for the benefit of the company. To stay in line with the heavy competition,obtaining useful insights from the data is one of the essential needs for a company.The effective use of this data directly or indirectly leads to effective decision making. Due to the high change, volume and variety of data ,the manual checking of the data is exhausting due to the era of Big Data, and expensive also requires a lot of human

effort. The aim here is to automate the process of getting possible insights from data without human intervention.This is where Data mining techniques come into rescue , automating the entire process resulting in required information with no human bias. Data mining is the process of extracting knowledge from an existing database. Data Mining Techniques are implemented as programs are used to find the underlying or hidden patterns in the data. Data mining applications include Fraud Detection ,Product Recommendations, Customer segmentation etc. Customer segmentation is the process of forming groups of customers with similar characteristics.It can be applied in a company is to understand the customers who have similar and distinct characteristics.Customers are grouped into clusters or segments, where each customer in a cluster in turn have similar characteristics with others within that cluster but rather have different characteristics of those from other segments. The importance of customer segmentation is the ability of a business to customize marketing strategies that would be appropriate for each segment of its customers.This identification and segmentation of customers happens using attributes such as demographics, number of items purchased , time spent on website, items purchased in past etc from the customer related database within the company.The promising and possible outcomes of customer segmentation are better understanding of customers, increase in sales, retain customers, improve customer experience , provide targeted ads and hence incrementing the overall marketing.

3.Literature Review:

In [1] Tushar Kansal et al have worked on 3 different clustering algorithms (K-Means, Agglomerative, and Mean Shift) which have been implemented to segment the customers and finally compare the results of the clusters obtained from the algorithms. A python code has been developed and trained by applying scalar onto the dataset having 2 features of 200 training samples taken from a local retail shop. Through clustering , 5 segments of clusters have been formed labeled as Careless, Careful, Standard, Target, and Sensible customers. However , two new clusters have emerged on applying mean shift clustering labeled as high buyers and occasional buyers.

The results have 2 internal clustering measures, Silhouette score and Calinski- Harabasz index. In conclusion, As the data set was unlabeled they have opted for internal clustering validation rather than external clustering validation which depends on some external data like labels. In this case internal clustering validation which best suits the dataset and can correctly cluster data into opposite clusters.

In [2] Aman Banduni et al have identified customer segments into a commercial business using the data

mining method such as customer segmentation. The data used in the paper were collected from the UCI Machine learning repository which contains 8 attributes. In this paper several steps were taken to obtain an accurate result. It includes a feature with Centro's first stage, allocation phase and update phase , which are the most common phase K-Means algorithms.

The results showed that orange clusters have the highest value customers, green as the lowest value customers, blue and red as the high opportunity customers. Overall, as the dataset is unbalanced they have opted for internal clustering validation techniques rather than external clustering validation techniques which require extra data such as labels. Internal clustering validation techniques showed the best data cluster.

4.Methodology:

4.1.Dataset Description:

The data set used to implement the Customer segmentation using K-Means clustering algorithm is the Online Retail II data set. This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011.The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.The data set contains 8 attributes and has 525461 tuples, representing the data of 4384 customers.

4.2.Feature Description:

The attributes in the data set are InvoiceNo, StockCode, Description, Quantity,InvoiceDate, UnitPrice(Â£), CustomerID, Country.

- **InvoiceNo**:It consists of Invoice number. It is of Nominal data type. A 6-digit unique integral number assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.

- **StockCode**: It consists of Product (item) code. It is of Nominal data type. A 5-digit integral number uniquely assigned to each distinct product.

- **Description**: it consists of Product (item) name.it is of Nominal data type.

- **Quantity**: It specifies each product (item) quantity per transaction. It is of Numeric datatype.

- **InvoiceDate**: It consists of Invoice date and time when a transaction was generated. It is of Numeric datatype.

- **UnitPrice**: It specifies Unit price. It is of Numeric datatype. Product price per unit in sterling (Â£).(1 Pound sterling equals 96.22 Indian Rupee)

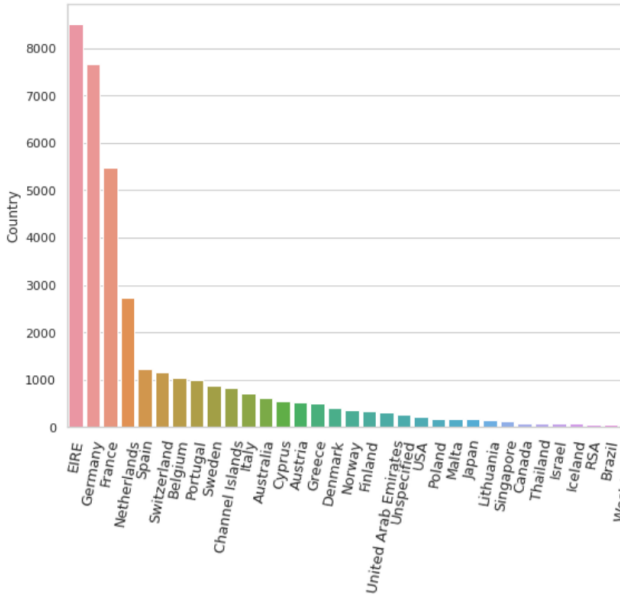
- **CustomerID**:It consists of Customer number. It is a Nominal data type. A 5-digit integral number uniquely assigned to each customer.

- **Country**: It consists of Country names. It is a Nominal data type. The name of the country where a customer resides.

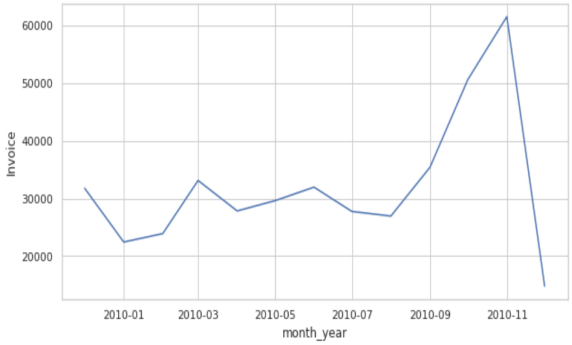
4.3 Exploratory Data Analysis

Before beginning the modeling work, EDA is used to see what the data can tell us. Deriving insights from raw data can be tiresome, uninteresting, and/or overpowering. In this case, exploratory data analysis approaches have been developed as a help.

Transactions count in each country



Number of transactions per year



4.4. Data Preprocessing

Data preprocessing is the process by which raw data is prepared and suitable for machine learning models. This is the first important step in creating a machine learning model. Raw data can typically contain noise, missing values, and unusable formats that aren't directly available for machine learning models. Data preprocessing is required to clean up the data and make it suitable for machine learning models. This also improves the accuracy and efficiency of machine learning models.

Methods Applied :

1.Handling Missing Values: Null Values from the dataset are removed, missing values from Description and CustomerID's rows are removed. Size of Dataset (541909, 8) before Removing null values is reduced to (406829, 8).

```
print(dataset.isnull().sum())
```

```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID     135080
Country            0
dtype: int64
```

Before Removing Null Values

After removing null values

```
dataset = dataset.dropna()
print(dataset.isnull().sum())
print(dataset.shape)
```

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
(406829, 8)
```

2. The Quantity attribute has values less than 0 so, these negative values in the dataset will be removed so that it will not affect while forming clusters along with which duplicates records in the data are also removed .

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Coun
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/01/10 8:26	2.55	17850.0	United Kingd
1	536365	71053	WHITE METAL LANTERN	6	12/01/10 8:26	3.39	17850.0	United Kingd
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/01/10 8:26	2.75	17850.0	United Kingd
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/01/10 8:26	3.39	17850.0	United Kingd
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/01/10 8:26	3.39	17850.0	United Kingd
...	...	...	...	...	...	...	...	
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/09/11 12:50	0.85	12680.0	Fran
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/09/11 12:50	2.10	12680.0	Fran
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/09/11 12:50	4.15	12680.0	Fran
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/09/11 12:50	4.15	12680.0	Fran
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/09/11 12:50	4.95	12680.0	Fran

526054 rows × 8 columns

3. Min-Max normalization: Min-Max normalization transforms each feature individually so that it fits within the specified range of the training set.

This technique is used on the dataset prepared from the RFM Analysis in the Section.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

4.Features extraction: Revenue calculation, here we multiplied Price and Quantity inorder find and add revenue generated column[Price * Quantity = Revenue]. And we have extracted data and month from InvoiceDate.

```
dataset["revenue"] = dataset["Price"]*dataset["Quantity"]
dataset["revenue"]
```

```
0         83.40
1         81.00
2         81.00
3        100.80
4         30.00
          ...
525456     5.90
525457     3.75
525458     3.75
525459     7.50
525460     3.90
Name: revenue, Length: 407695, dtype: float64
```

RFM (Recency, Frequency, Monetary) Analysis

Recency, frequency, monetary value is a marketing analysis tool used to identify a company`s or an organization's best customers by measuring and analyzing spending habits. RFM analysis reasonably predicts how much revenue a company will make from customers who

are likely to buy the product again, new customers (compared to frequent customers), and how to convert casual buyers to frequent customers.

The RFM model is based on three quantitative factors:

- Recency- How recently a customer has made a purchase.
- Frequency- How often a customer makes a purchase.
- Monetary- How much money a customer spends on purchases.

Recency: Recency factor indicates the most recent purchase of all the customers.In the dataset , we will consider the most recent purchase date as a reference date. Then we will find the recency of each customer by subtracting the transaction date from the reference date.Finally we will get one integer value that will be the recency of the customer.

	Customer ID	Diff
0	12346.0	164
1	12347.0	2
2	12348.0	73
3	12349.0	42
4	12351.0	10
...	...	...
4309	18283.0	17
4310	18284.0	66
4311	18285.0	295
4312	18286.0	111
4313	18287.0	17

[4314 rows x 2 columns]

Frequency: Customer transaction frequency can be influenced by factors such as product type, purchase price, and the need for replenishment or replacement. If you can predict the purchase cycle, marketing activities can be directed towards encouraging them to go to the store when they are low on staples. Here, we are grouping the Customer ID by the total count of the invoice register on the customer, in order to find the number of transactions done by the customer(frequency).

	Customer ID	Frequency
0	12346.0	11
1	12347.0	2
2	12348.0	1
3	12349.0	3
4	12351.0	1
...	...	...
4309	18283.0	6
4310	18284.0	1
4311	18285.0	1
4312	18286.0	2
4313	18287.0	4

[4314 rows x 2 columns]

Monetary: Monetary value is how much a customer spends on purchase in a period. Overall amount spent by the customers in all transactions. This is to put more emphasis on encouraging customers who spend the most money to continue to do so. It produces a better return on investment in marketing and customer service, and also runs the risk of isolating customers who have been consistent but may not spend as much with each transaction.

Monetary value is computed by considering customer id and summing up the amount he/she spends in each

transaction. The output displays each customer id and Monetary value of that specific customer id.

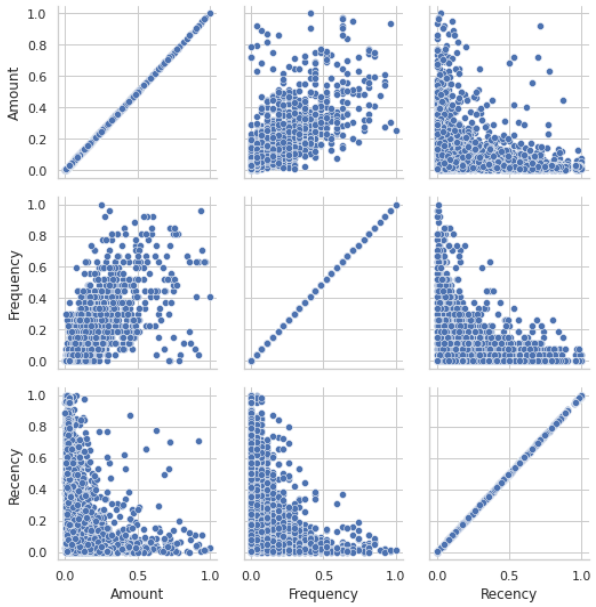
```
Customer ID
12346.0     372.86
12347.0    1323.32
12348.0     222.16
12349.0    2671.14
12351.0     300.93
             ...
18283.0     641.77
18284.0     461.68
18285.0     427.00
18286.0    1296.43
18287.0    2345.71
Name: Amount, Length: 4314, dtype: float64
```

From the features extracted by RFM Analysis , the customized dataset is prepared with Amount ,Frequency and Recency as features on which K-means Clustering is applied.

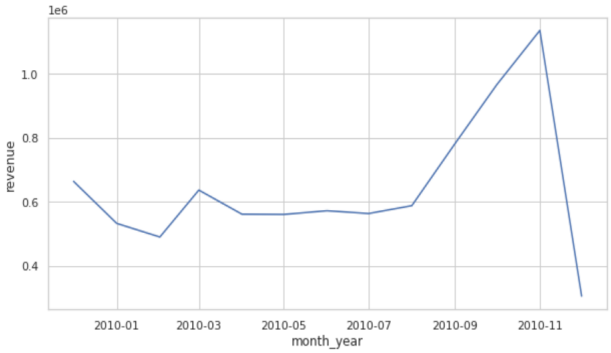
Data transformation is applied on the dataset to improve the performance of the algorithm , Min-max normalization technique is used to normalize all the features.

	Amount	Frequency	Recency
0	0.024629	0.370370	0.439678
1	0.087409	0.037037	0.005362
2	0.014674	0.000000	0.195710
3	0.176437	0.074074	0.112601
4	0.019877	0.000000	0.026810

Dataset derived from RFM Analysis



Pair plot between all features



Revenue per each month

5.Proposed Method

We used K-means clustering technique to segment the customers, followed the below systematic approach for optimizing the performance of K-means clustering.

K-means++ method for the initialisation of cluster centroids.Choosing the ‘K’

value in the algorithm plays an important role in segmenting clusters.

Inorder to choose the optimal value of K, Silhouette Method is used, there are many other methods are also available like Elbow method but Silhouette Method has shown the optimal value of K.

K-means clustering algorithm is one of the most popular centroid based algorithms. It is an unsupervised learning algorithm and used for clustering tasks which works really well with complex datasets. It is an iterative algorithm that partitions the dataset into “k” pre-defined non overlapping subgroups (clusters) where each data point belongs to only one group. This clustering algorithm relies on centro, where each data point is placed in one of the overlapping ones, which is pre-sorted in the K-algorithm.

```
for k = 1 to K do
    μ_k ← some random location          // randomly initialize mean for kth
end for
repeat
    for n = 1 to N do
        z_n ← argmin_k ||μ_k − x_n||    // assign example n to closest
    end for
    for k = 1 to K do
        μ_k ← MEAN({ x_n : z_n = k })   // re-estimate mean of cl
    end for
until converged
return z                                // return cluster assign
```

Silhouette method is considered as a better metric than elbow is used to determine the optimum number of clusters. The silhouette score of a point measures how close that point lies to its nearest neighbor points, across all clusters. It provides information about clustering quality which can be used to determine whether further refinement by clustering should be performed on the current clustering.

Silhouette score for each sample is calculated using the formula: Calculate the silhouette s(i) as follows , ratio of the difference between cluster cohesion and separation to the greater of the two :
s(i) = b(i) – a(i) / [max{a(i),b(i)}]

K-means++ is an algorithm for choosing the initial values fthe K-means clustering algorithm.This algorithm ensures a smarter initialization of the centroids and improves the quality of the clustering. Apart from initialization, the rest of the algorithm is the same as the standard K-means algorithm. That is, K-means++ is the standard K-means algorithm coupled with a smarter initialization of the centroids.

Initially , we considered clusters as n=2 to n=10 , with 300 iterations

```
For n_clusters=2, the silhouette score is 0.5661945492047095
For n_clusters=3, the silhouette score is 0.5477575520199691
For n_clusters=4, the silhouette score is 0.4743336103129437
For n_clusters=5, the silhouette score is 0.4377793045333027
For n_clusters=6, the silhouette score is 0.3880489018428877
For n_clusters=7, the silhouette score is 0.381351139040909
For n_clusters=8, the silhouette score is 0.387409124390399
For n_clusters=9, the silhouette score is 0.37027024047295526
For n_clusters=10, the silhouette score is 0.35262500025185517
```

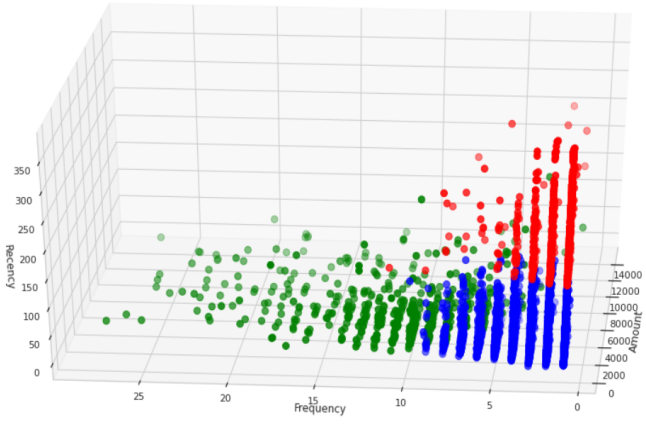
6.Results

After using the Silhouette approach and interpretation of Data, K=3 is chosen. Hence, customers are divided into three clusters.For more insights the K-means is applied first on all features and then on 2 selected features.The graphs and results displayed below.

The cluster labels (0,1,2) predicted by the K-means clustering algorithm are assigned to corresponding customers.

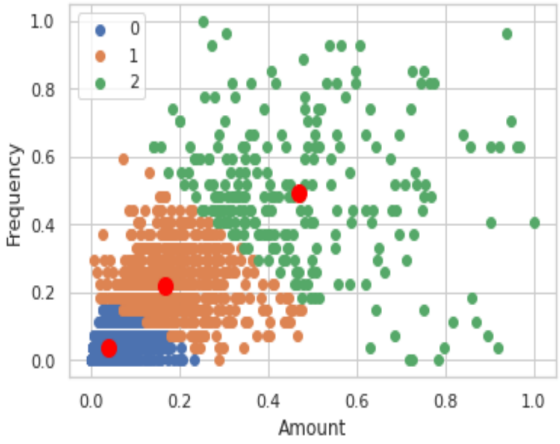
	Customer ID	Amount	Frequency	Recency	Cluster_Id
0	12346.0	372.86	11	164	1
1	12347.0	1323.32	2	2	0
2	12348.0	222.16	1	73	0
3	12349.0	2671.14	3	42	0
4	12351.0	300.93	1	10	0

Assigned cluster labels to respective customers



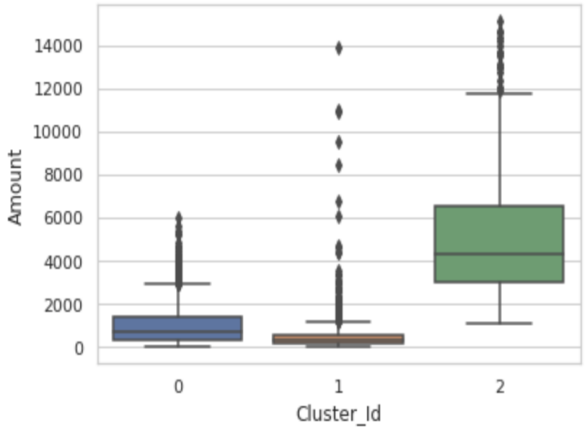
Considered Amount, Frequency, and Recency as the features ,K-means clustering is applied using all these features. A 3D visualization of the resultant clusters obtained is shown above. Customers have been segmented into 3 groups. Red ,blue and green indicate clusters 0,1 and 2 respectively.

In the shown graph only 2 features are considered(Amount and Frequency). We performed K-means clustering on these 2 features.Cluster centroids obtained from the k-means++ method are represented in red colour.Horizontal axis represents the amount and vertical axis represents Frequency.



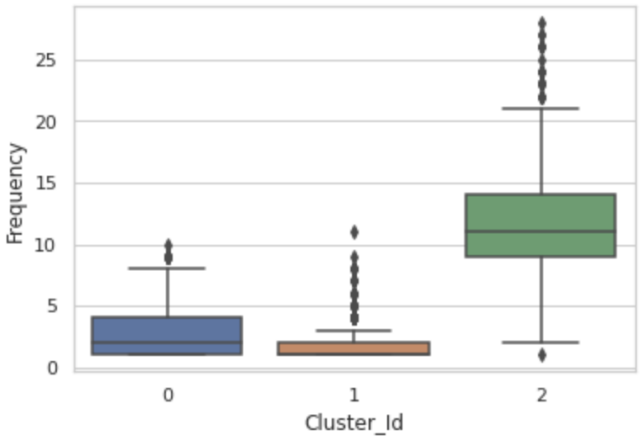
Now after applying K means clustering, Data is divided into 3 clusters and clearly seen on the graph. 0 represents blue data points, 1 represents orange data points, 2 represents green data points.
To summarize the results of the proposed model, the below information along with the boxplots are described.

Here, we can see the distribution when Cluster_Id attribute as x and Amount attribute as y. Cluster 1 has min 0 to max 3000,Cluster 2 has min 0 to max 1000 and Cluster 3 has min 1000 to max 12000



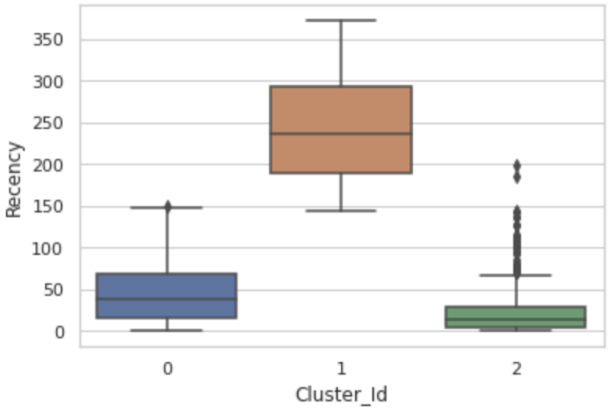
Clusters vs Amount

Here, we can see the distribution when Cluster_Id attribute as x and Frequency attribute as y. Cluster 1 has min 0 to max 8,Cluster 2 has min 0 to max 2 and Cluster 3 has min 0 to max 21.



Clusters vs Frequency

Here, we can see the distribution when Cluster_Id attribute as x and Recency attribute as y. Cluster 1 has min 0 to max 150, Cluster 2 has min 150 to max 370 and Cluster 3 has min 0 to max 70.



Clusters vs Recency

Conclusion:

This paper introduced the RFM Model implementation of a K-Means clustering algorithm for customer segmentation using data collected from the online retailers. Our model divides customers into mutually exclusive groups (in this

case, three clusters).As shown in the above plots , cluster 1 represents the customers with highest recency , cluster 2 represents customers with lowest recency and accordingly cluster 1 is in between. This insight is obtained by segmenting the customers and hence leading to better understanding them, which can be used to increase the company's sales.This will help to apply further data mining strategies, and the more insights you gain the more it will help the business decisions.

References :

[1] T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K-means Clustering," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 135-139, doi: 10.1109/CTEMS.2018.8769171.

[2] Aman Banduni, A. Ilavendhan, “Customer Segmentation using Machine Learning,” 2021 International Journal of Innovative Research in Technology(IJIRT), 2021, Volume 7, Issue 2, pp. 116-122, doi: http://10.10.11.6/handle/1/1850.

[3] Abhinav Sagar, A. (n.d.). *Customer Segmentation Using K Means Clustering -KDnuggets*.KDnuggets;www.kdnuggets.com. Retrieved May 1, 2022, from https://www.kdnuggets.com/2019/11/customer-segmentation-using-k-means-clustering.html

[4] Dhiraj Kumar. (2021, June 18). *Implementing Customer Segmentation Using Machine Learning [Beginners Guide] - neptune.ai*. Neptune.Ai; neptune.ai. https://neptune.ai/blog/customer-segmentation-using-machine-learning

[5] Muhal, H. (n.d.). *(PDF) Two-Stage Customer Segmentation using K-Means Clustering And Artificial Neural Network | Harshit Muhal - Academia.edu*. (PDF) Two-Stage Customer Segmentation Using K-Means Clustering And Artificial Neural Network | Harshit Muhal - Academia.Edu; www.academia.edu. Retrieved May 1, 2022, from https://www.academia.edu/45443368/Two_Stage_Customer_Segmentation_using_K_Means_Clustering_And_Artificial_Neural_Network

[6] Li, Y., Qi, J., Chu, X., & Mu, W. (2022, January 9). *Customer Segmentation Using K-Means Clustering and the Hybrid Particle Swarm Optimization Algorithm | The Computer Journal | Oxford Academic*. OUP Academic; academic.oup.com. https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxab206/6501352?redirectedFrom=fulltext