Capstone Project - 4

Customer Segmentation

Project By :

Sanjay Yadav

Contents

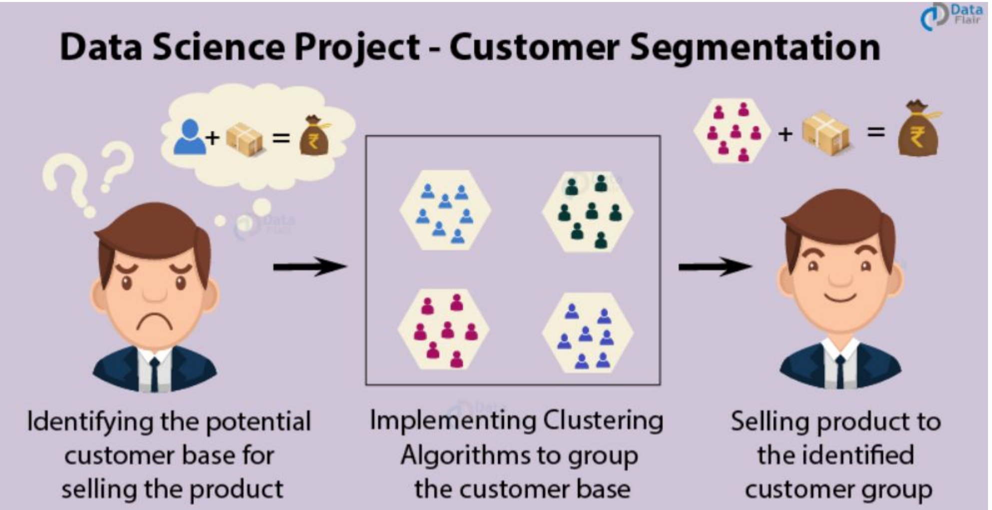
Introduction

Customer segmentation is the process of separating customers into groups on the basis of their shared behavior or other attributes. The groups should be homogeneous within themselves and should also be heterogeneous to each other. The overall aim of this process is to identify high-value customer base i.e. customers that have the highest growth potential or are the most profitable.

Insights from customer segmentation are used to develop tailor-made marketing campaigns and for designing overall marketing strategy and planning.

Problem Statement

In this project, our task is to identify major customer segments on 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.



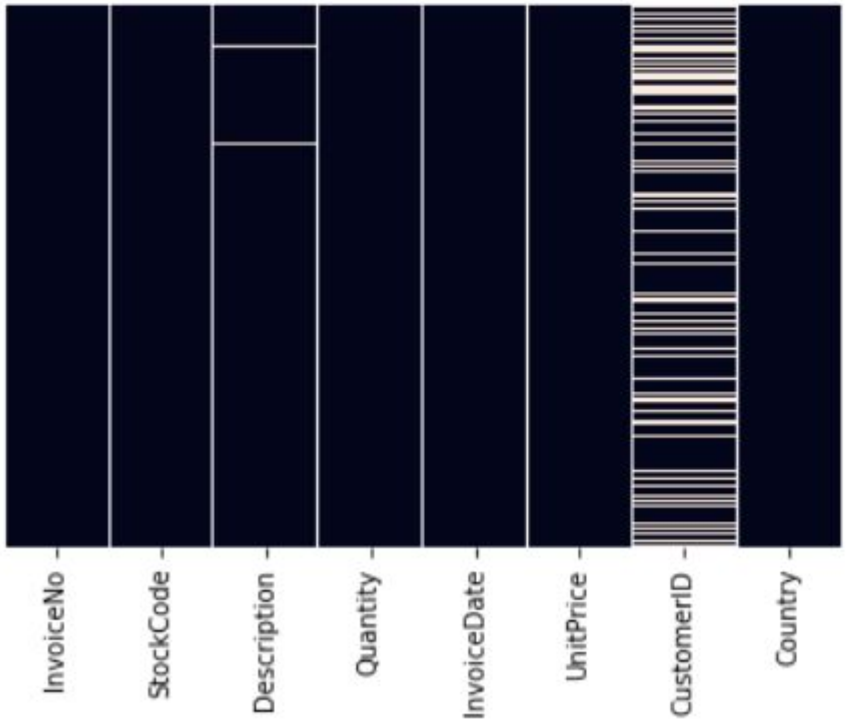
Data Summary

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

Missing Values

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



Exploratory Data Analysis

```
array(['United Kingdom', 'France', 'Australia', 'Netherlands', 'Germany',
       'Norway', 'EIRE', 'Switzerland', 'Spain', 'Poland', 'Portugal',
       'Italy', 'Belgium', 'Lithuania', 'Japan', 'Iceland',
       'Channel Islands', 'Denmark', 'Cyprus', 'Sweden', 'Finland',
       'Austria', 'Greece', 'Singapore', 'Lebanon',
       'United Arab Emirates', 'Israel', 'Saudi Arabia', 'Czech Republic',
       'Canada', 'Unspecified', 'Brazil', 'USA', 'European Community',
       'Bahrain', 'Malta', 'RSA'], dtype=object)
```

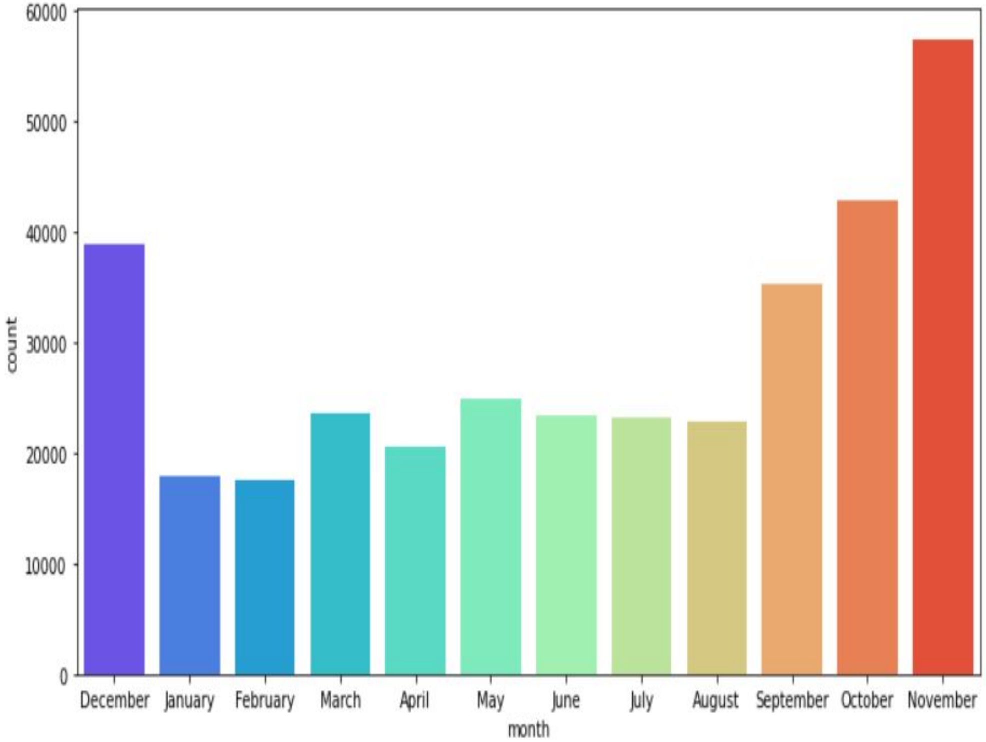
```
United Kingdom    349203
Germany             9025
France              8326
EIRE                7226
Spain               2479
Netherlands         2359
Belgium             2031
Switzerland         1841
Portugal            1453
Australia           1181
Name: Country, dtype: int64
```

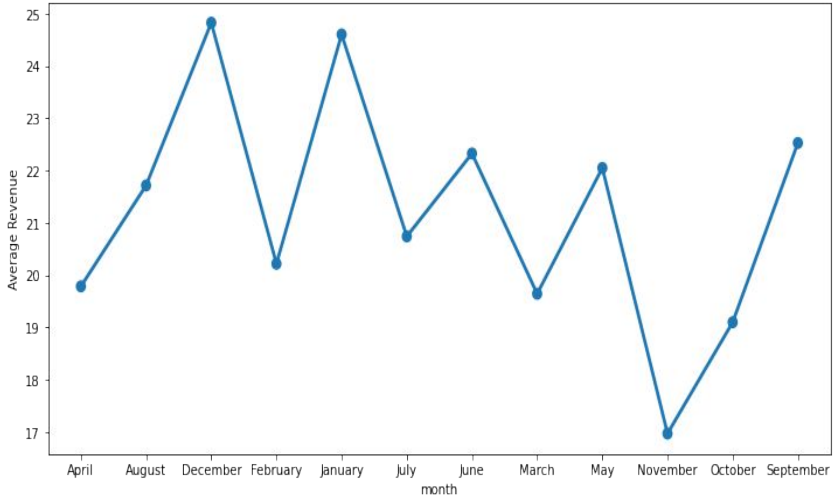
Most of the data is from UK, since different region can affect clustering we will only consider data from UK.

```
#total unique cunstomers
df_uk['CustomerID'].nunique()
```

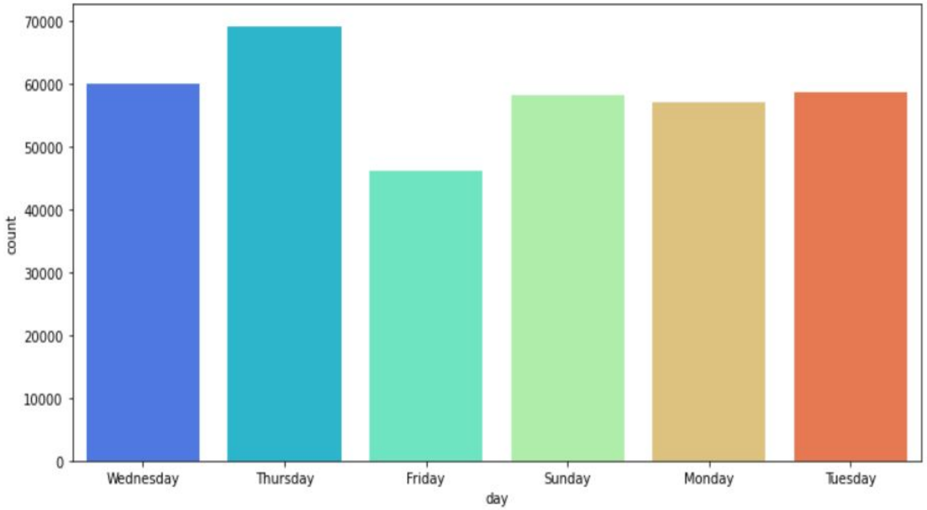
```
3920
```

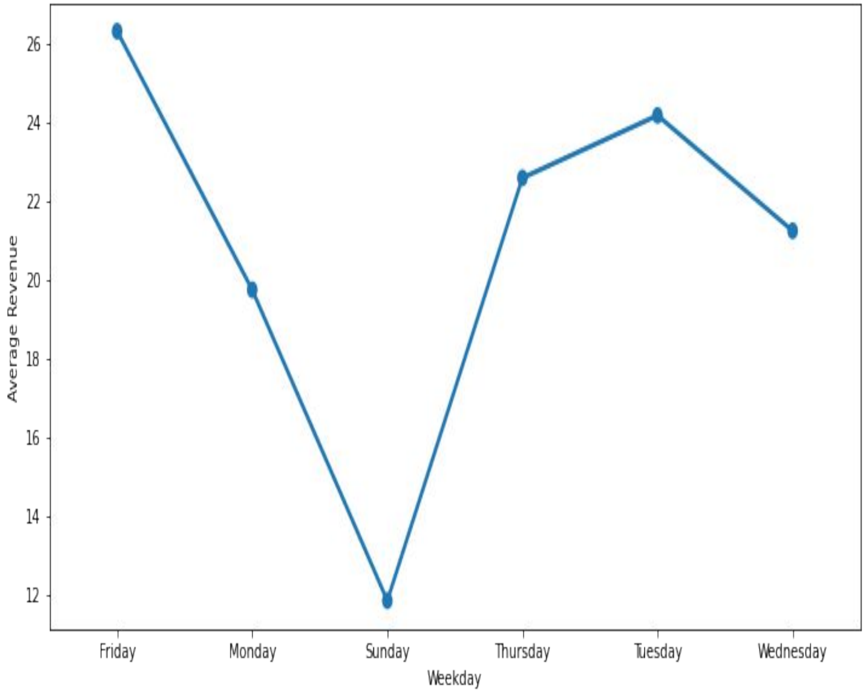
Monthly Analysis



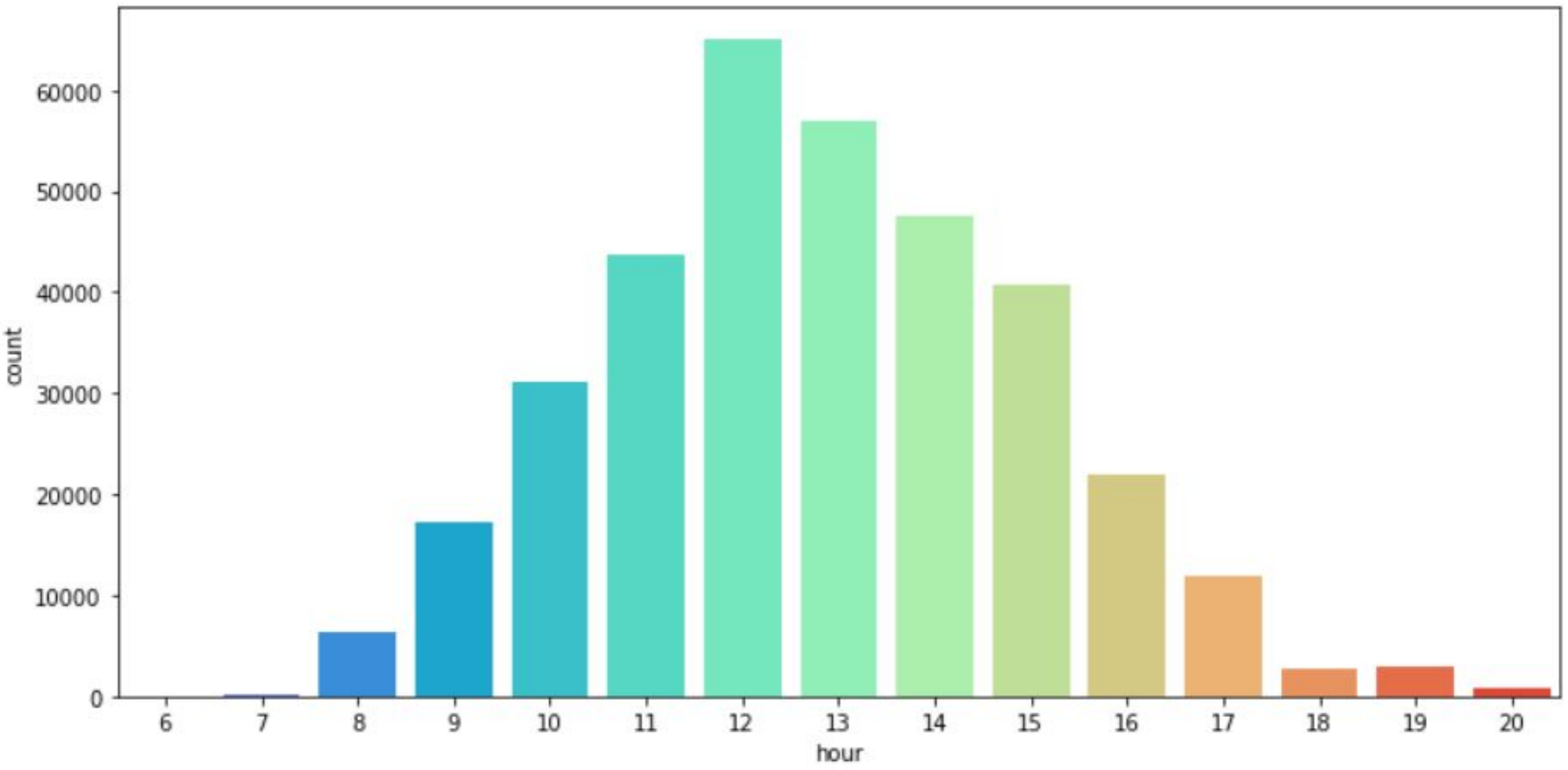


Weekly Analysis





Hourly Analysis



RFM Segmentation

RFM analysis is a customer behavior segmentation technique. Based on customers' historical transactions, RFM analysis focuses on 3 main aspects of customers' transactions: recency, frequency and purchase amount. Understanding these behaviors will allow businesses to cluster different customers into groups.

Recency : How recently did the customer visit our website or how recently did a customer purchase

Frequency : How often do they visit or how often do they purchase

Monetary : How much revenue we get from their visit or how much do they spend when they purchase

Individual RFM Values

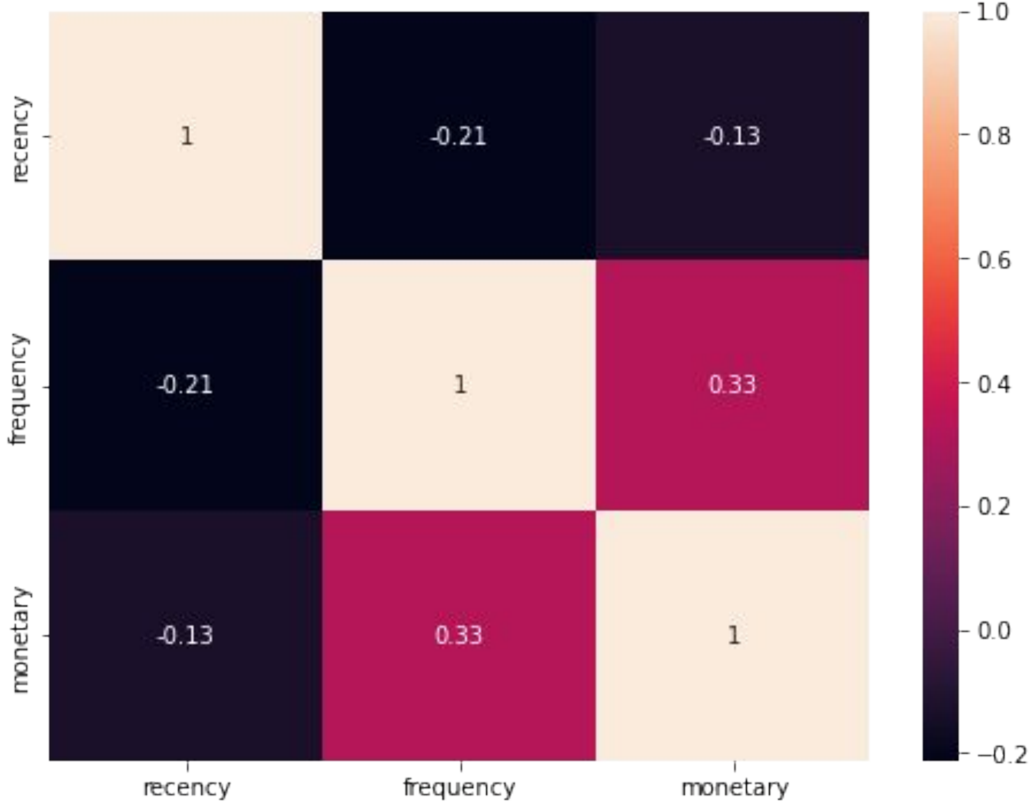
	CustomerID	recency	frequency	monetary
0	12346	326	1	77183.60
1	12747	2	103	4196.01
2	12748	1	4412	33053.19
3	12749	4	199	4090.88
4	12820	3	59	942.34

Quartile RFM values

	CustomerID	recency	frequency	monetary
0.25	14208.75	18.0	17.0	298.185
0.50	15569.50	51.0	40.0	644.975
0.75	16913.25	143.0	98.0	1571.285

	CustomerID	recency	frequency	monetary	r_quartile	f_quartile	m_quartile	RFM_Segment	RFM_Score
0	12346	326	1	77183.60	4	4	1	441	9
1	12747	2	103	4196.01	1	1	1	111	3
2	12748	1	4412	33053.19	1	1	1	111	3
3	12749	4	199	4090.88	1	1	1	111	3
4	12820	3	59	942.34	1	2	2	122	5

Correlation Between RFM



Observation :

Frequency and monetary value are positively correlated with each other implying an increase in frequency implies increase in monetary value

Frequency and Recency are negatively correlated with each other implying an increase in frequency implies decrease in monetary value

Best customers :

RFM Score: 111

Who They Are: Highly engaged customers who have bought the most recent, the most often, and generated the most revenue.

Marketing Strategies: Focus on loyalty programs and new product introductions. These customers have proven to have a higher willingness to pay, so don't use discount pricing to generate incremental sales. Instead, focus on value added offers through product recommendations based on previous purchases.

Big Spenders

RFM Score: XX1

Who They Are: Customers who have generated the most revenue for your store.

Marketing Strategies: These customers have demonstrated a high willingness to pay. Consider premium offers, subscription tiers, luxury products, or value add cross/up-sells to increase AOV. Don't waste margin on discounts.

Loyal customers

RFM Score: X1X

Who They Are: Customers who buy the most often from your store.

Marketing Strategies: Loyalty programs are effective for these repeat visitors. Advocacy programs and reviews are also common X1X strategies. Lastly, consider rewarding these customers with Free Shipping or other like benefits.

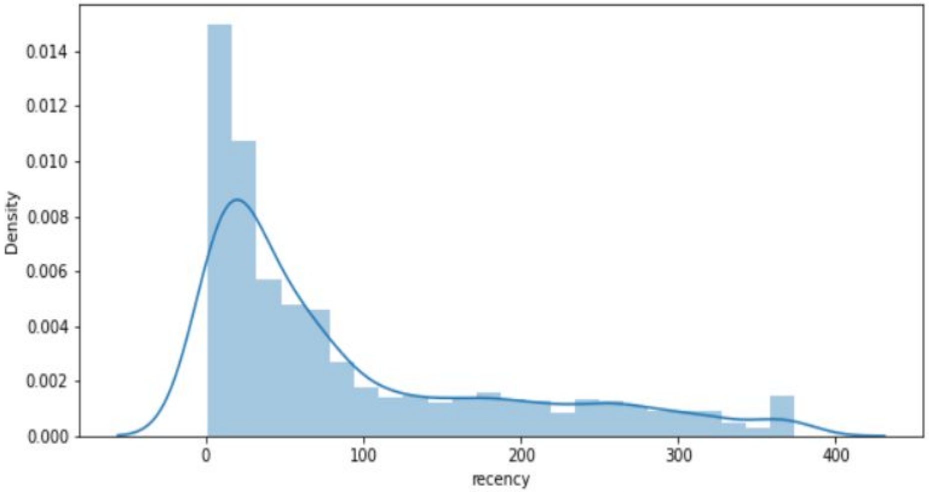
Newest Customers

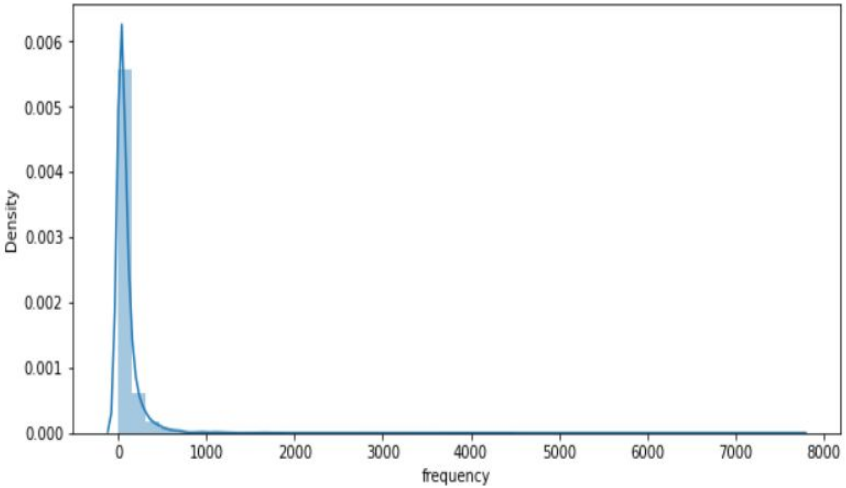
RFM Score: 14X

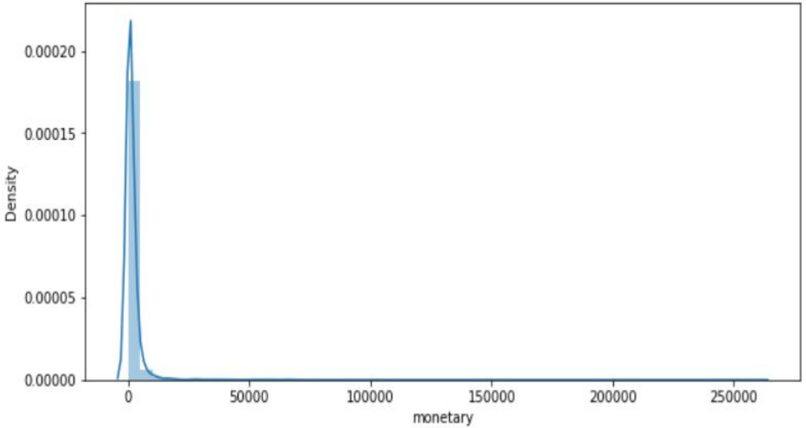
Who They Are: First time buyers on your site.

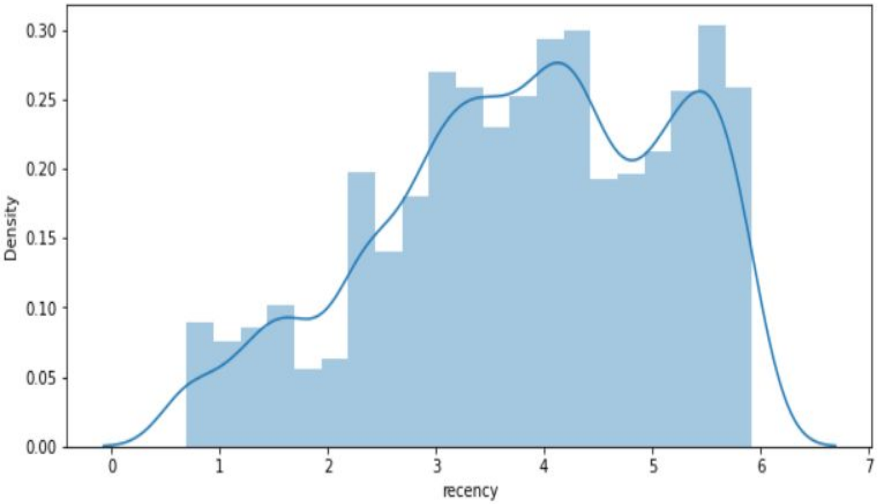
Marketing Strategies: Most customers never graduate to loyal. Having clear strategies in place for first time buyers such as triggered welcome emails will pay dividends.

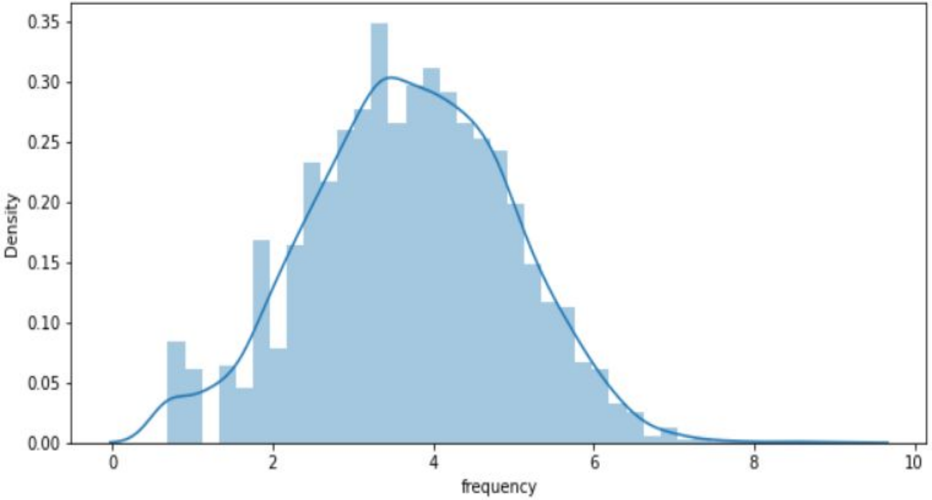
RFM Distribution

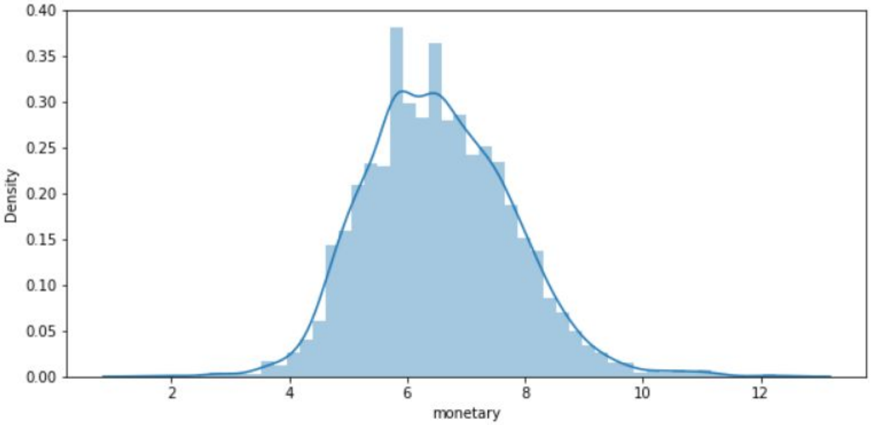







Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
1
2
3
4
5
6
7
recency


Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0
2
4
6
8
10
frequency

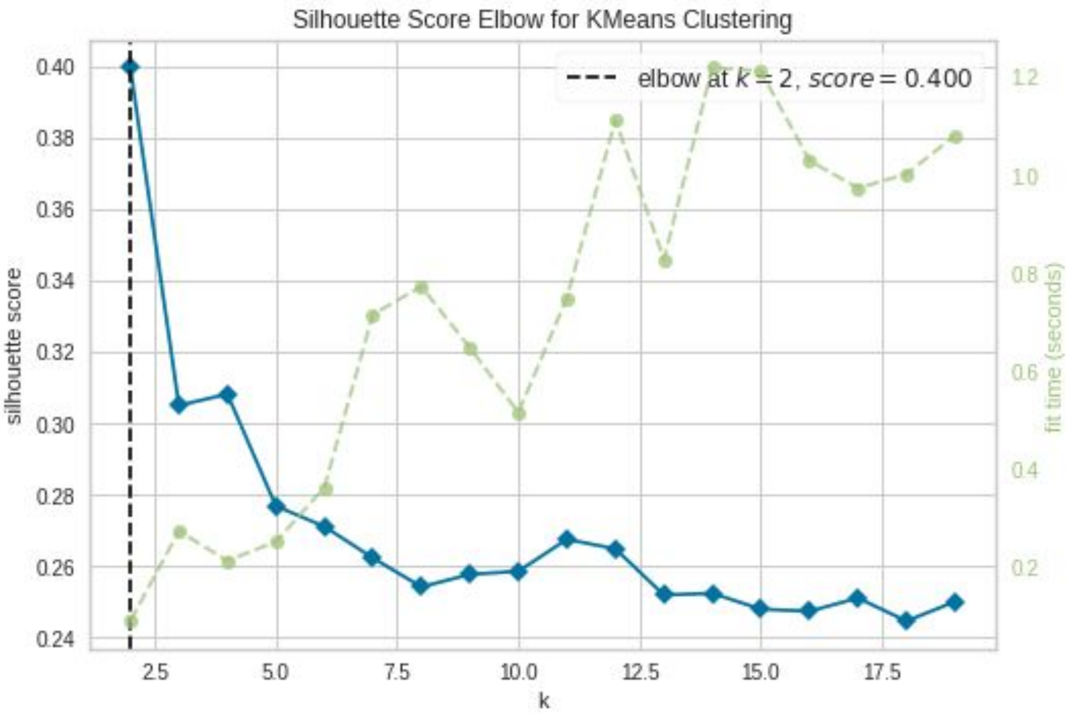

Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
2
4
6
8
10
12
monetary

Optimal number of clusters

Silhouette score for a set of sample data points is used to measure how dense and well-separated the clusters are.

Silhouette score takes into consideration the intra-cluster distance between the sample and other data points within the same cluster and inter-cluster distance between the sample and the next nearest cluster.

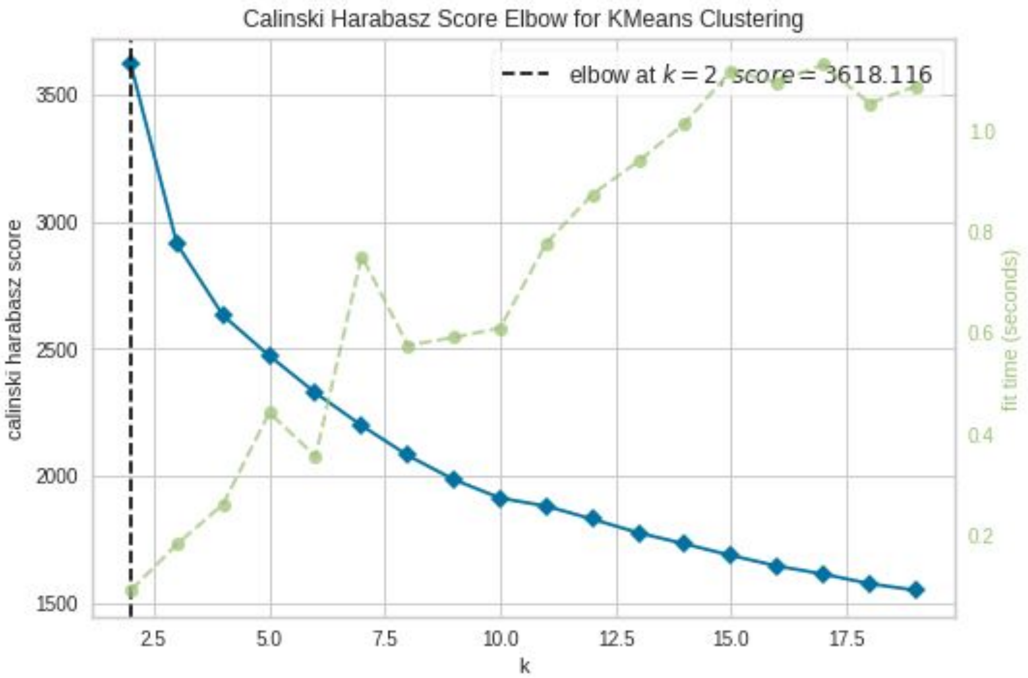
The silhouette score falls within the range [-1, 1].



The Calinski-Harabasz index also known as the Variance Ratio Criterion, is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters, the higher the score , the better the performances.

Advantages :

- The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster.
- The score is fast to compute.

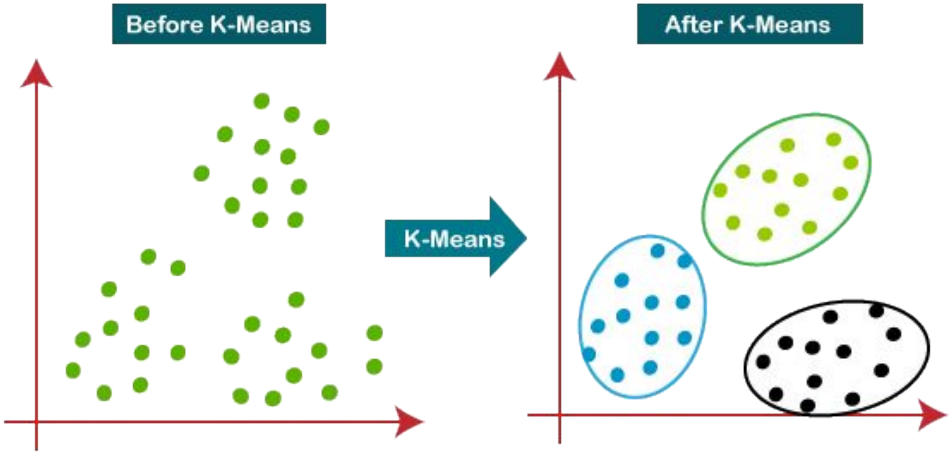


KMEANS

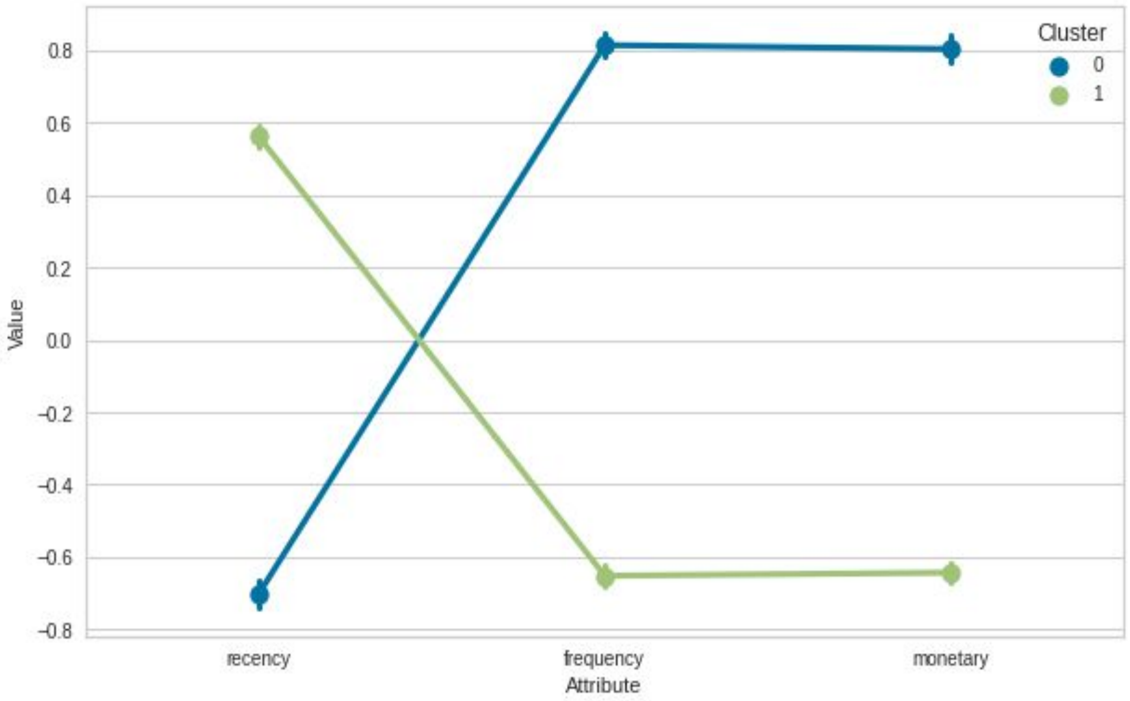
K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid.

The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

K-Means has the advantage that it's pretty fast, as all we're really doing is computing the distances between points and group centers; very few computations! It thus has a linear complexity O(n).



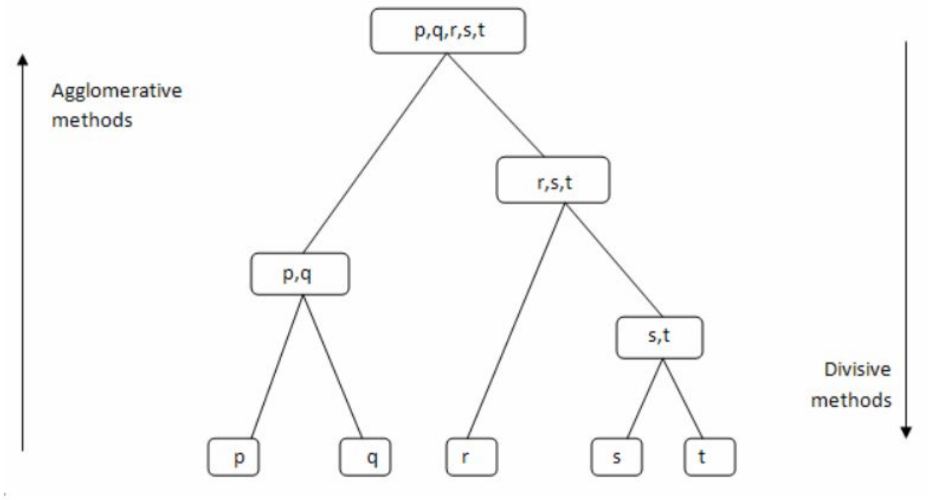
	recency	frequency	monetary	
	mean	mean	mean	count
Cluster				
0	31.0	170.0	3614.6	1743
1	141.2	24.3	452.3	2177

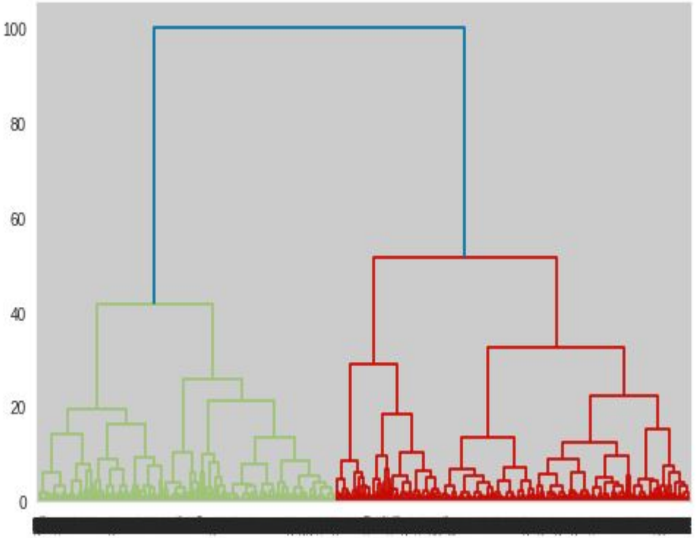


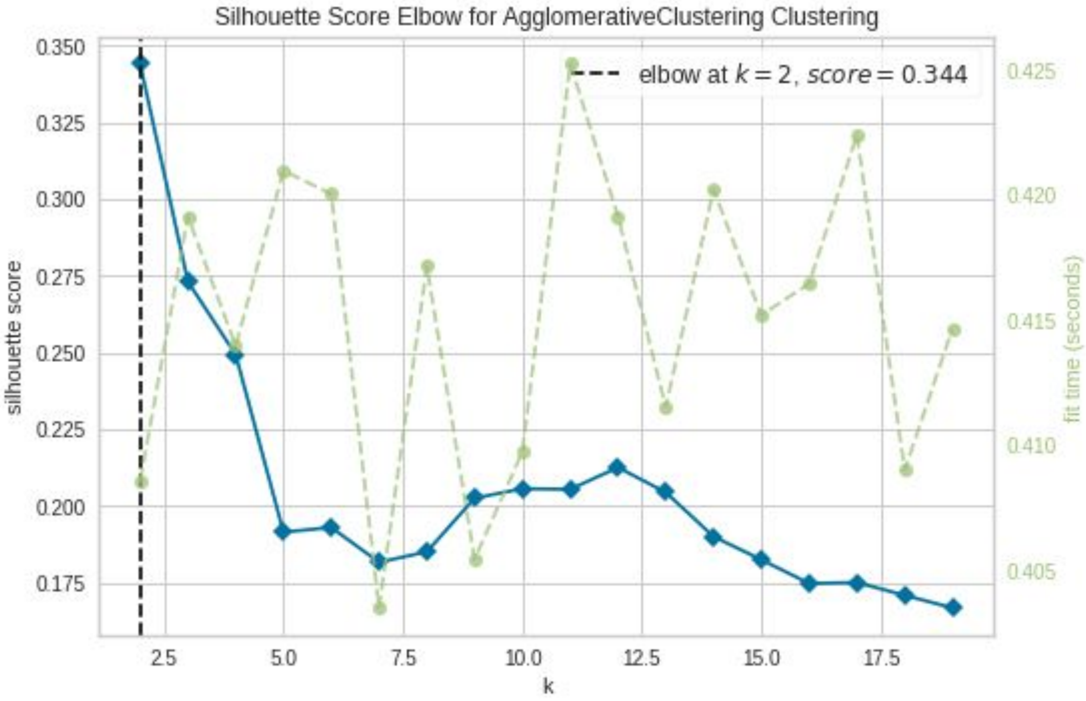
Hierarchical Agglomerative Clustering

Agglomerative clustering works in a "bottom-up" manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are member of just one single big cluster (root)

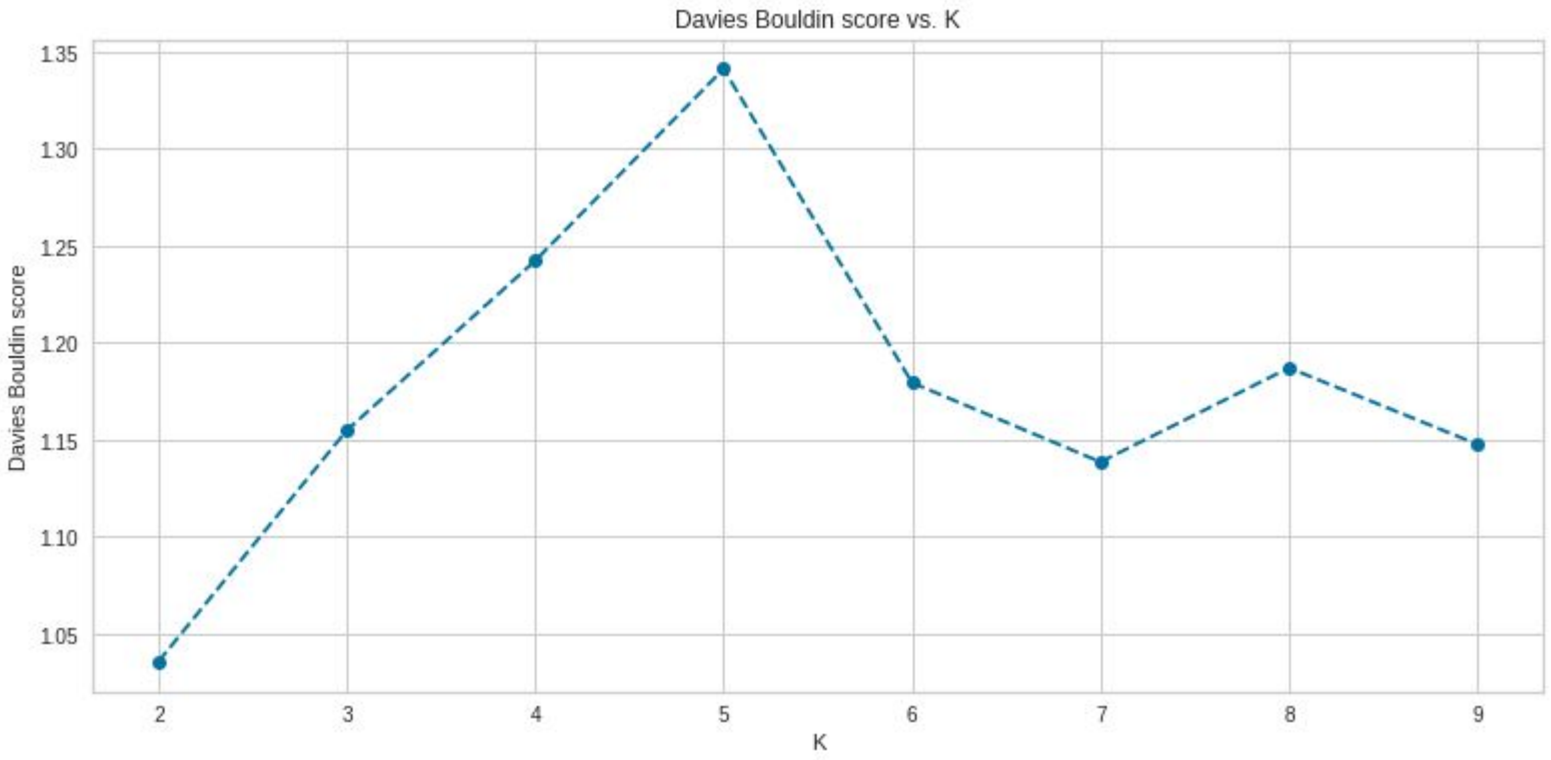
The result is a tree-based representation of the objects, named dendrogram.



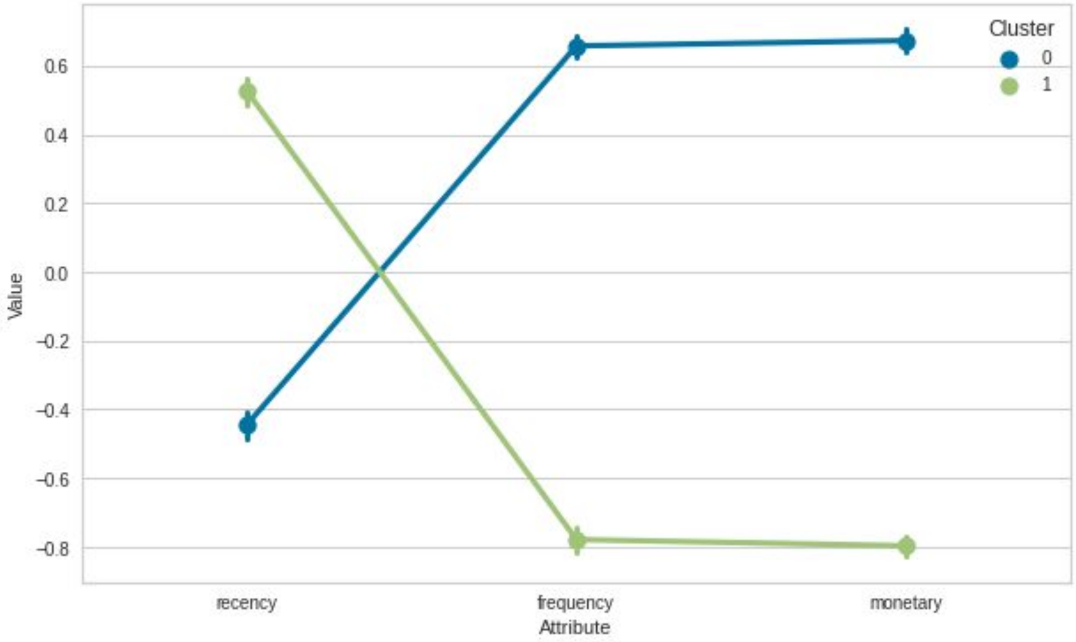

100
80
60
40
20
0


Silhouette Score Elbow for AgglomerativeClustering Clustering
elbow at $k = 2$, $score = 0.344$
silhouette score
0.350
0.325
0.300
0.275
0.250
0.225
0.200
0.175
fit time (seconds)
0.425
0.420
0.415
0.410
0.405
2.5
5.0
7.5
10.0
12.5
15.0
17.5
k

The Davies-Bouldin Index is defined as the average similarity measure of each cluster with its most similar cluster. Similarity is the ratio of within-cluster distances to between-cluster distances. In this way, clusters which are farther apart and less dispersed will lead to a better score



	recency	frequency	monetary	
	mean	mean	mean	count
Cluster				
0	47.3	146.4	3066.2	2125
1	145.4	21.2	428.6	1795



CONCLUSION

The Dataset was large enough summing around 5.4 lakh samples with most of the samples from UK.

On Thursday highest sales can be seen but on Friday highest revenue is generated.

High sales volume can be observed for November but most revenue generating months are December and January.

For the segmentation we used RMF Technique to create working table as it is most common segmentation technique.

Using various metrics such as silhouette score , calinski harabasz index and Davies Bouldin we have generated optimal number of clusters.

By using KMeans and Hierarchical clustering we formed appropriate clusters for the customer data.

Challenges

- Large dataset.
- Handling missing values.
- Proper RMF segmentation.
- Choosing the optimal number of clusters.
- Inferences from the cluster using clustering algorithms.

THANK YOU !!