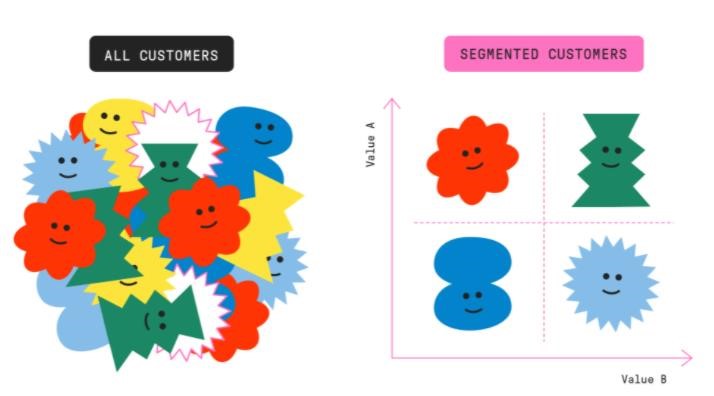
**Capstone Project - 4**

**Online Retail Customer Segmentation Unsupervised ML Model**

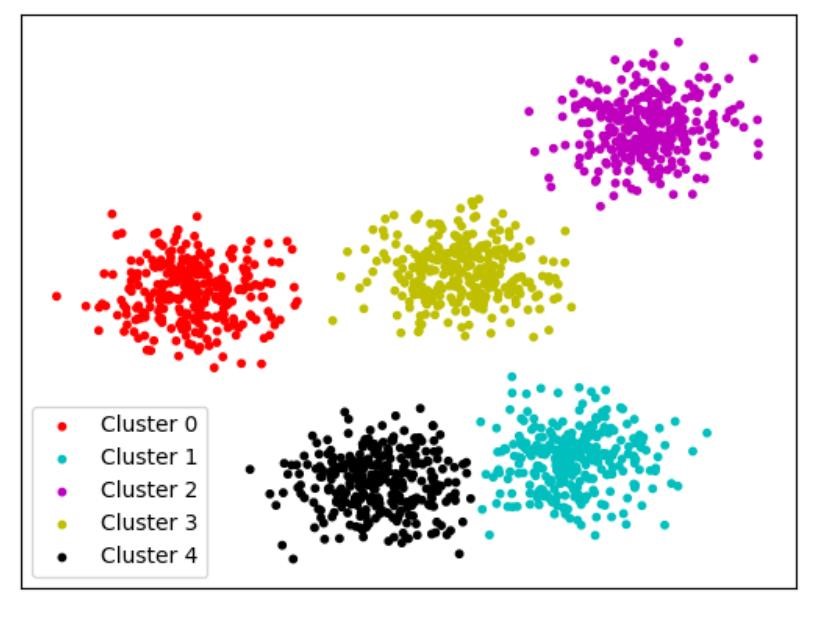
**Raavi sahu**

**Why Customer Segmentation?**

**“We needed a new way to understand our customers in a structured manner”**



# Introduction to Clustering

*Clustering can be considered the most important unsupervised learning problem. So, as every other problem of this kind, it deals with finding a structure in a collection of unlabelled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”.*

A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

# Content

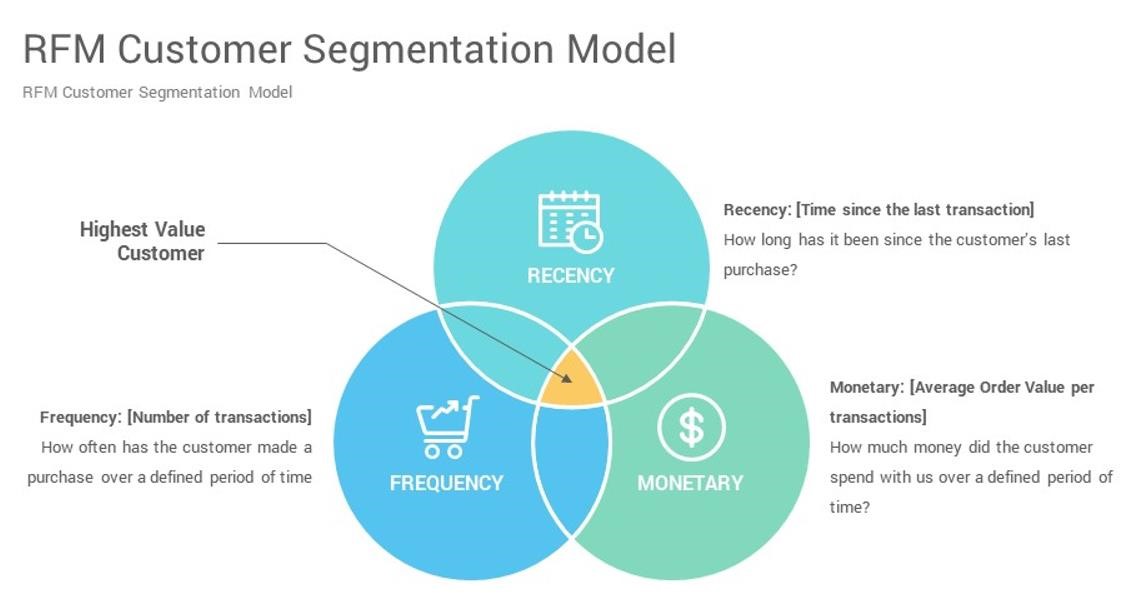
* Problem Statement
* Data Summary
* EDA / Feature analysis
* Data preparation
* Create RFM model
* Implementing various clustering

Models

* Challenges
* Conclusion

## Problem statement

* This project aims to identify major customer segments on a transnational data set for a UK-based online retail.
* Create RFM table
* We need to analyse and identify major customer segmentation using k means algorithm and also different algorithms to confirm our result.

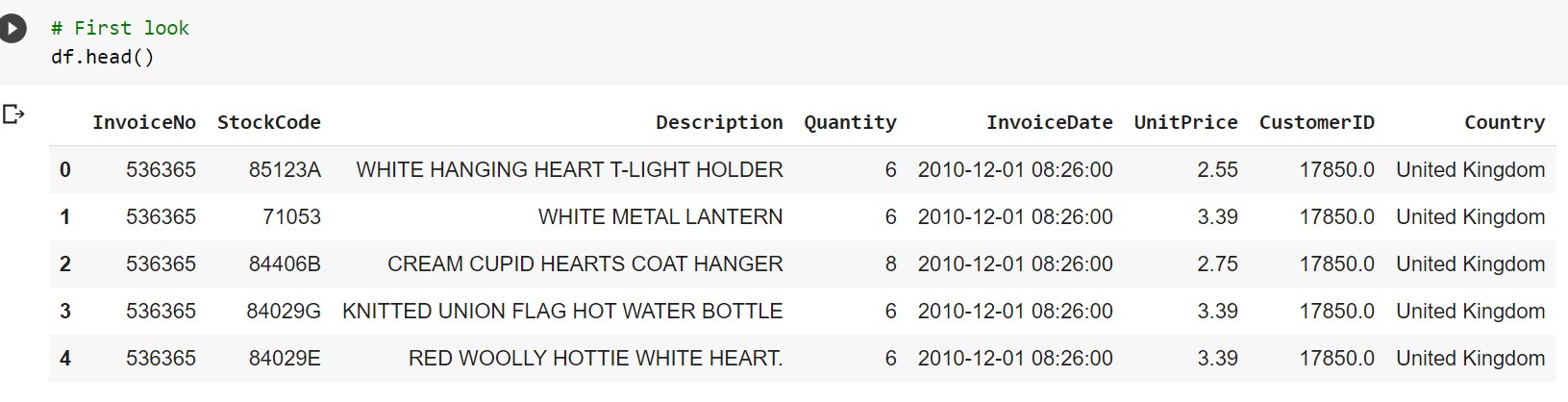


## Data Summary

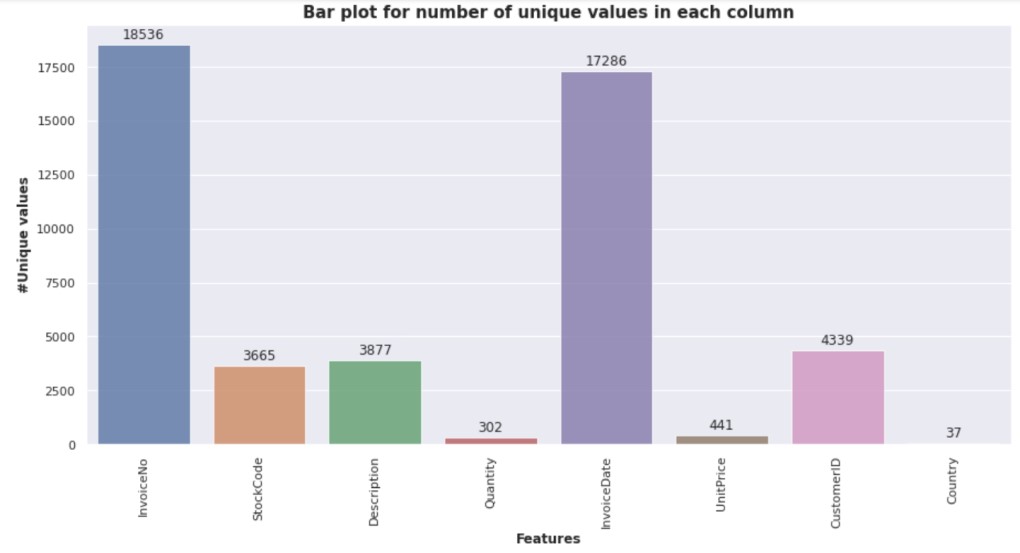
* InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction..
* StockCode: Product (item) code. 5-digit integral number uniquely assigned to each distinct product.
* Description: Product (item) name.
* Quantity: The quantities of each product (item) per transaction.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

## Basic Data Exploration

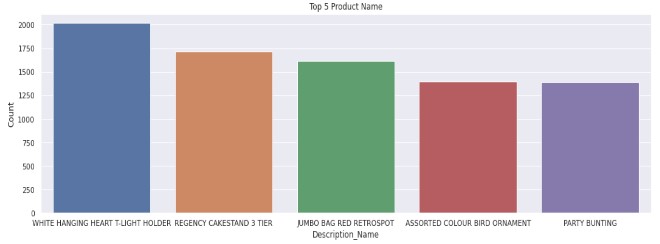
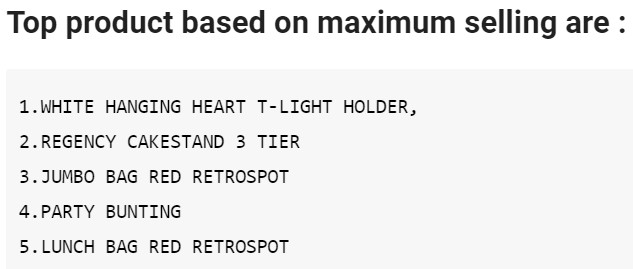
* A transnational data set with transactions occurring between 1st December 2010 and 9th December 2011 for a UK-based online retailer.
* Dataset has rows- 541909 & columns-8.
* The company mainly sells unique all-occasion gifts.
* Many customers of the company are wholesalers



**EDA - Which feature has the highest number of unique values?**

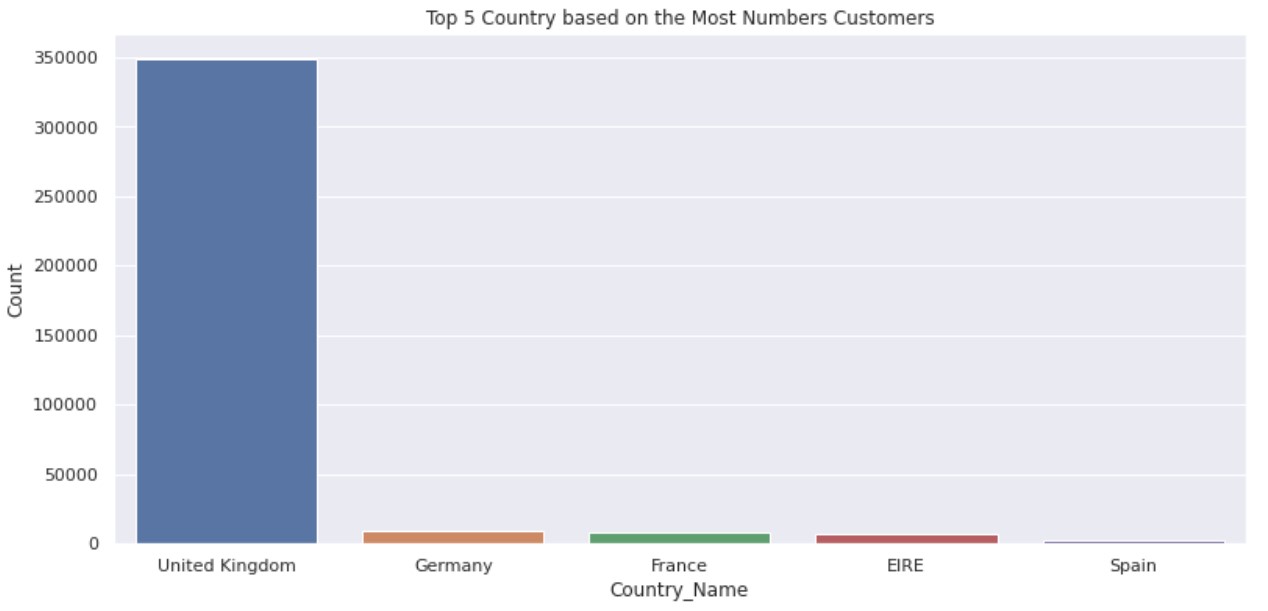
*The invoice number is unique for every transaction. Invoice Date has second highest count.*

### EDA – Finding Top product based on maximum selling

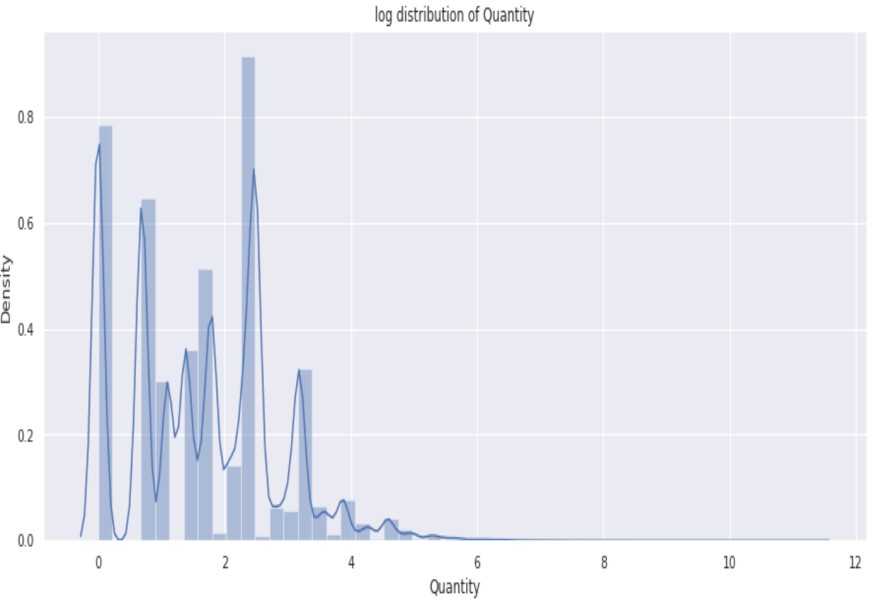
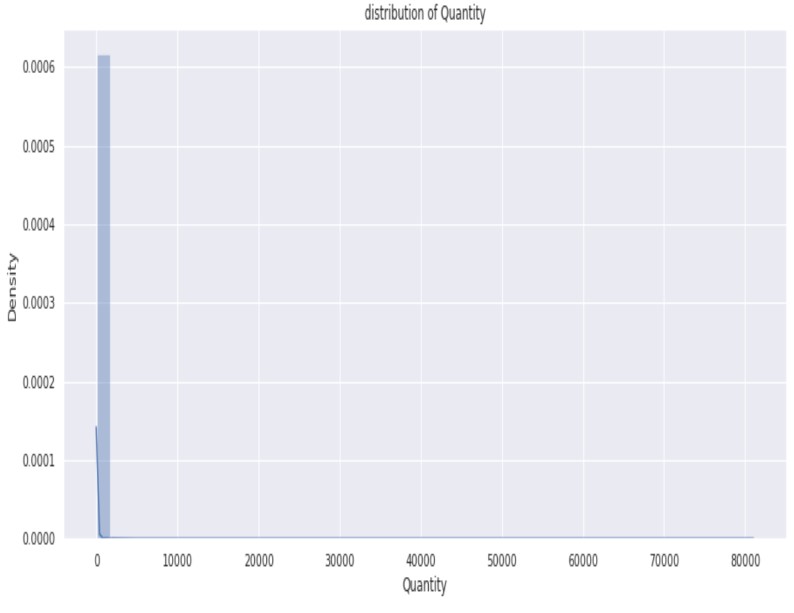


**White Hanging Heart T- Light Holder is the top product.**

**EDA - Top 5 Country based on the Most Numbers Customers?**

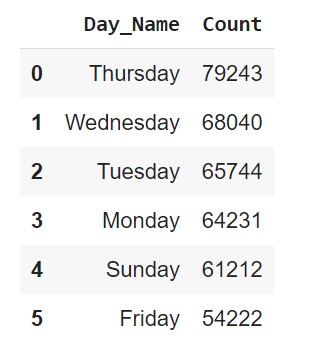
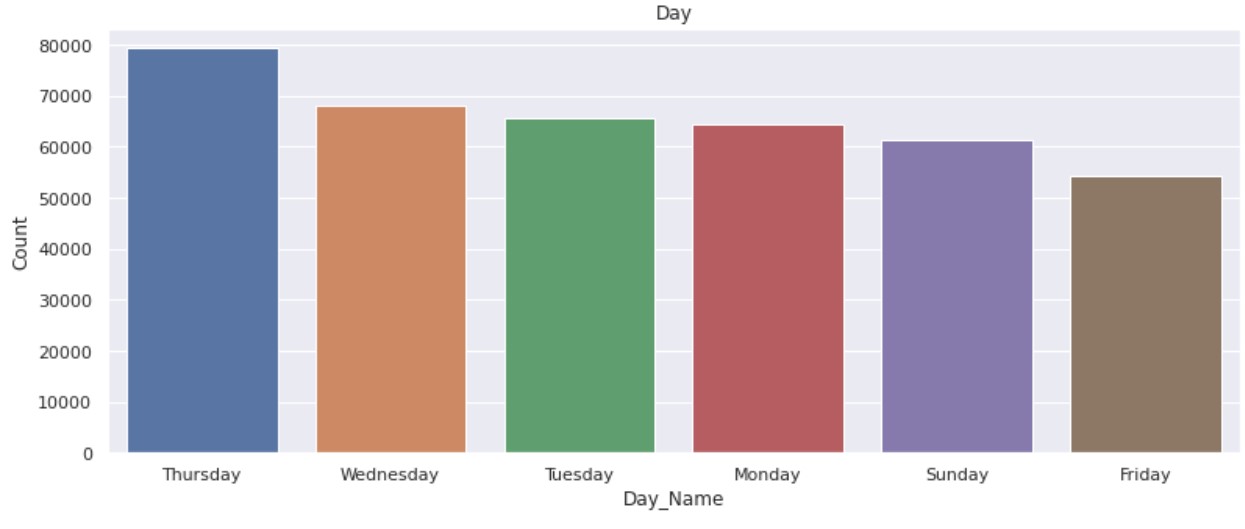
*In This graph, we can observe that most purchases are from the United Kingdom. It is justifiable also, as this is UK’s company.*

## Log transformation of quantity



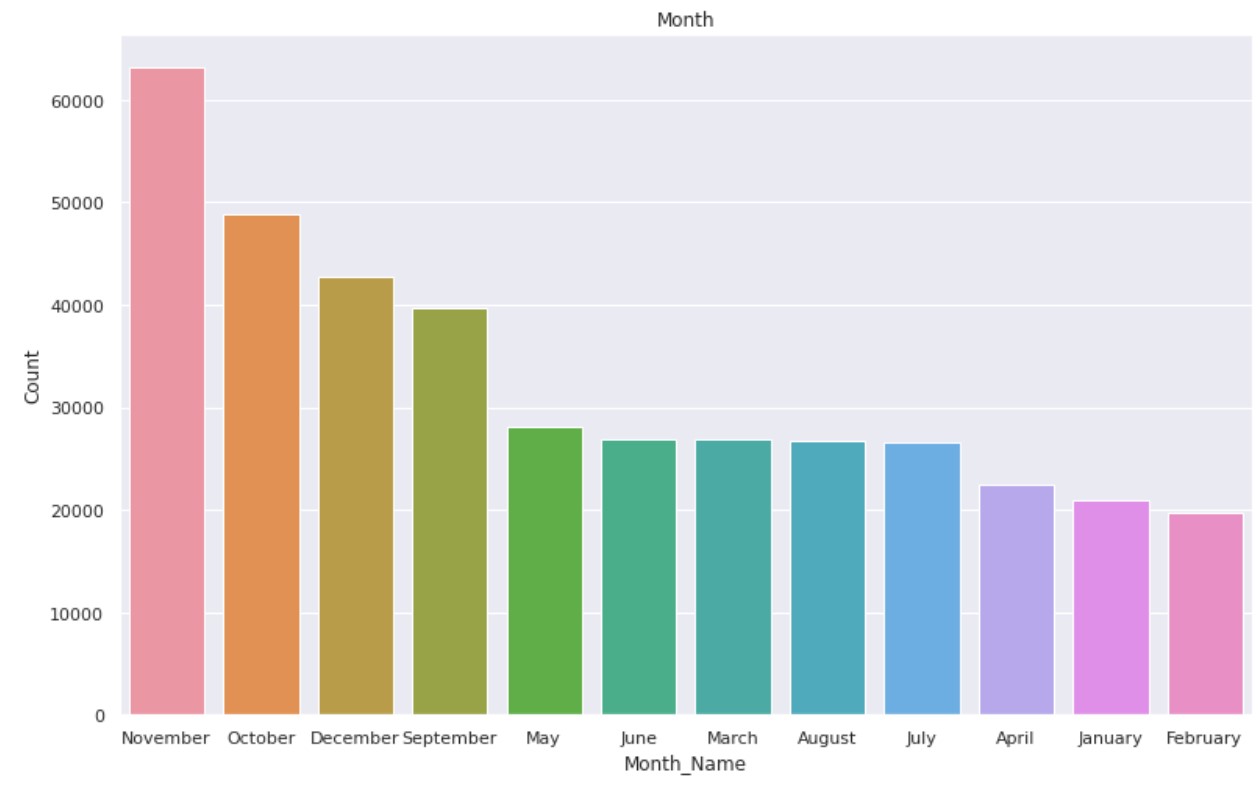
**For better accuracy, applied log distribution on quantity.**

**EDA- which day has highest count?**



**Most of the customers have purchased items on Thursday, Wednesday, Tuesday.**

**Which month has the highest count ?**

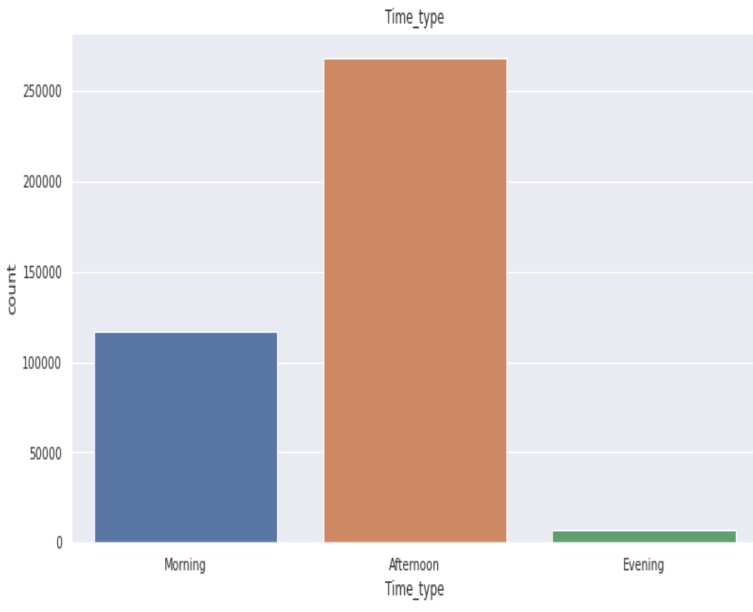


**Most of the customers have purchased items in November, October,**

**December, and the least number of purchases in April, January, February.**

|  |
| --- |
| We have divided hours of the day into 3-time types.   1. Morning 2. Afternoon 3. Evening |

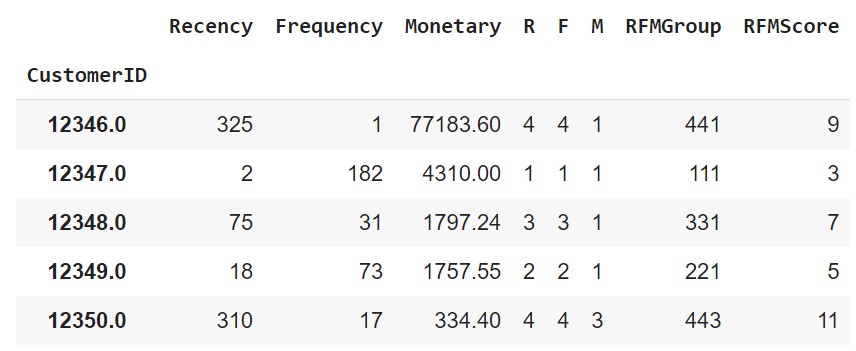
## Hour wise Analysis



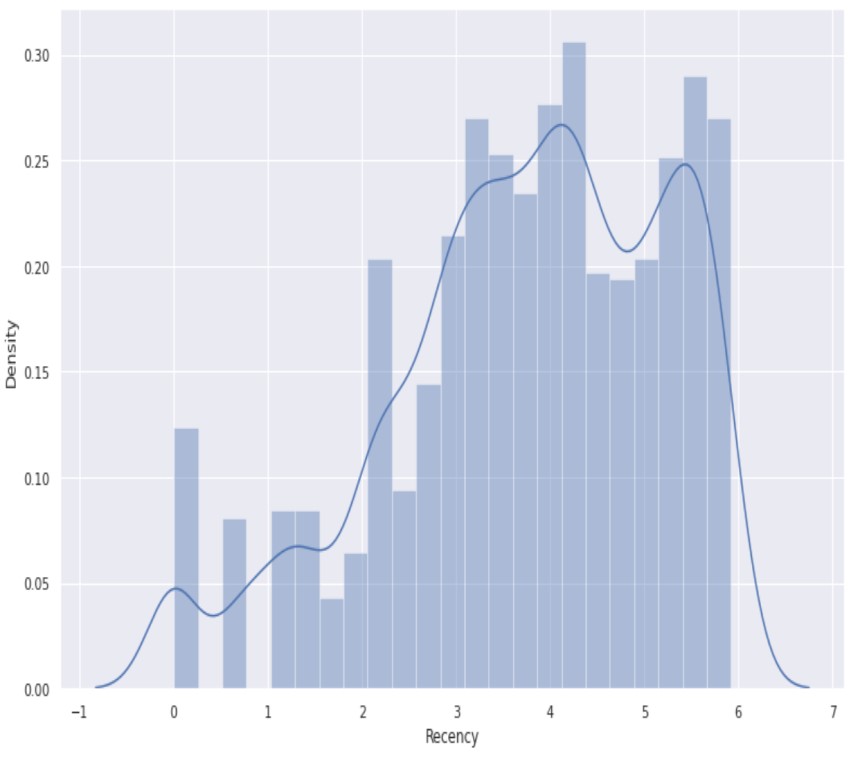
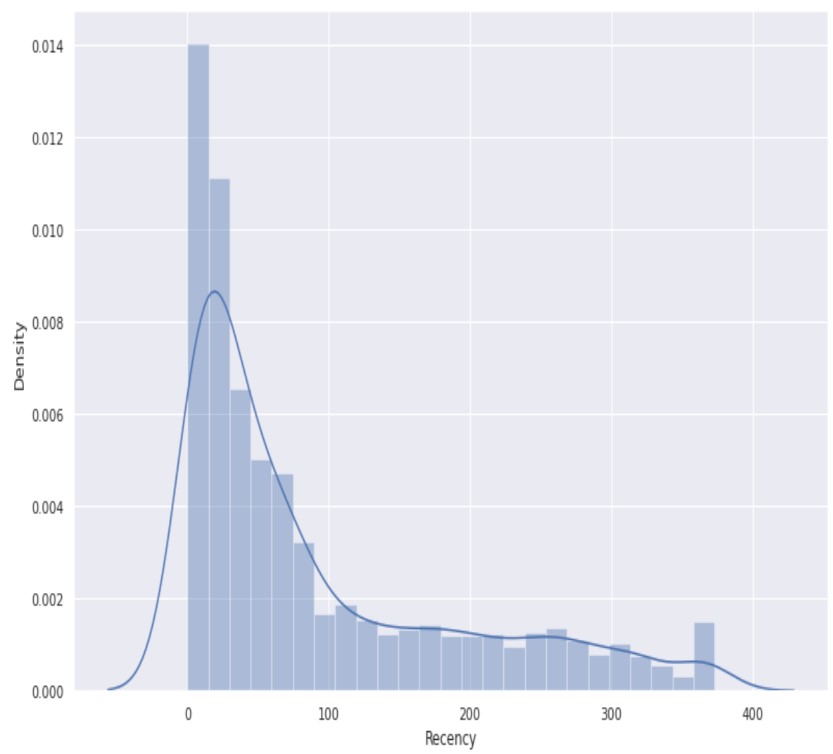
**Most of the customers purchase in the afternoon time. The 12th hour of the day is a peak for purchasing items. Moderate numbers of customers have purchased the items in the Morning and the least numbers of customers have purchased the items in the Evening.**

### Create the RFM model (Recency, Frequency,

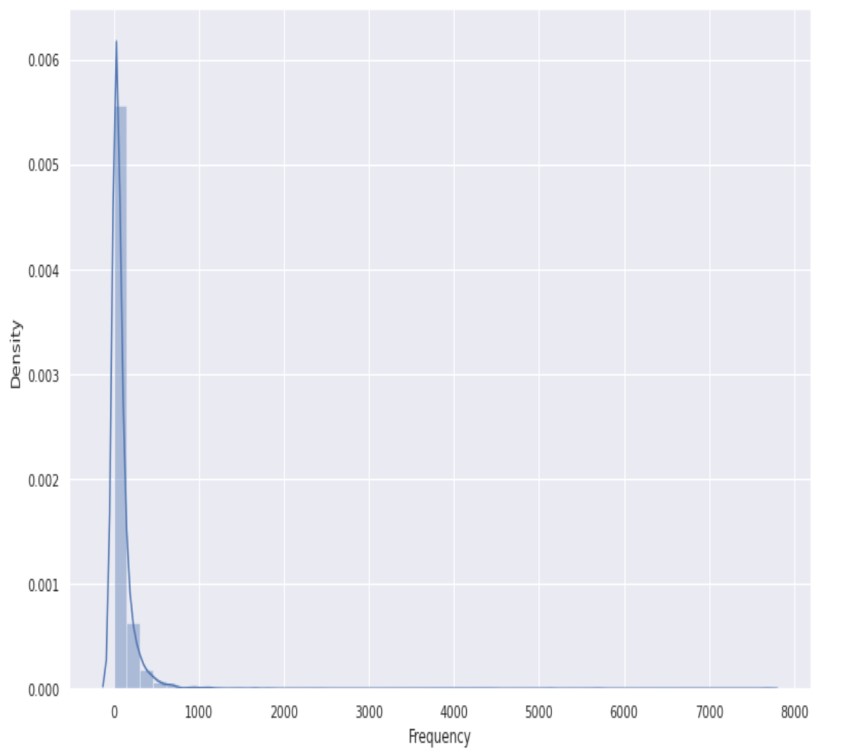
**Monetary value)**



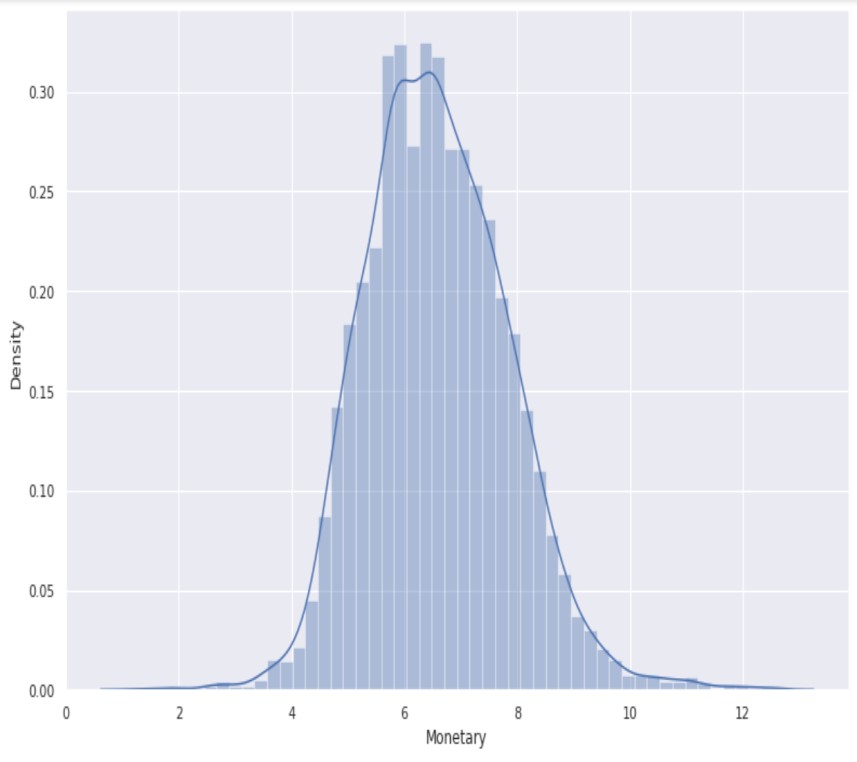
## Log Transformation of Recency



## Log Transformation of Frequency



### Log Transformation of Monetary



## Model Overview

**Let’s get some insight about Clustering models:**

**Silhouette score method :** It is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. It ranges from -1 to1 , where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters.

**Elbow method :** a point from where the value of clusters starts decreasing suddenly, indicates the optimal number of clusters.

**DBSCAN (Density Based Spatial Clustering of Application with Noise) :** Finds core samples of high density and expands clusters from them.

**Dendrogram :** It is representation of hierarchical clustering.

## Model Overview

* K-Means with silhouette score for RM
* K-Means with Elbow method FOR RM
* DBSCAN for RM
* K-Means with silhouette score for FM
* K-Means with Elbow method for FM
* DBSCAN for FM
* K-Means with silhouette score for RFM
* K-Means with Elbow method for RFM
* Hierarchical clustering for RFM
* DBSCAN for RFM

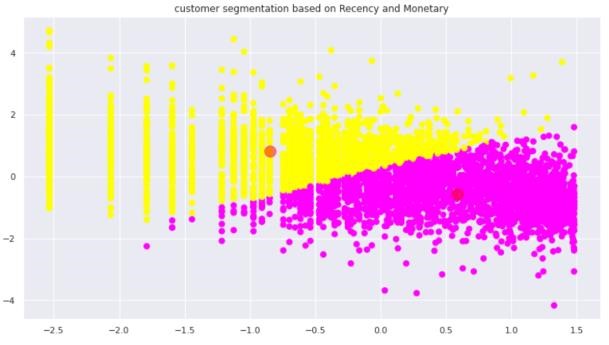
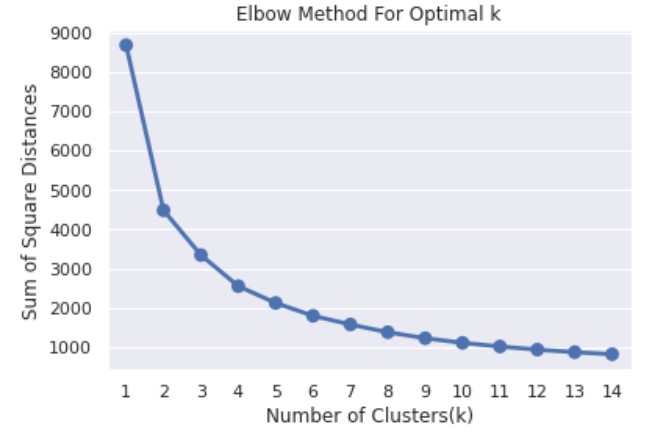
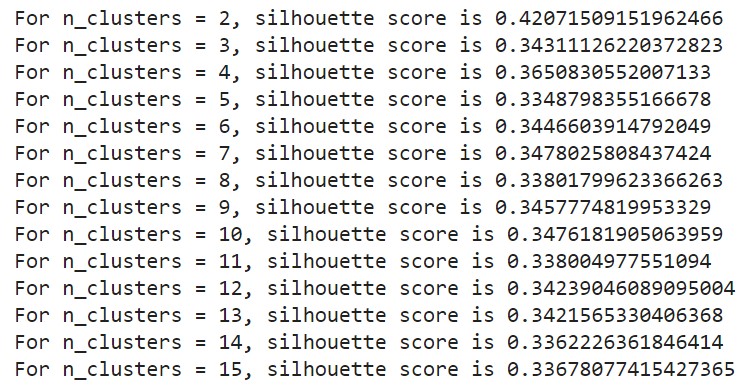
### Applying Silhouette Score and Elbow Method on Recency and Monetary

*We can see that, Customers*

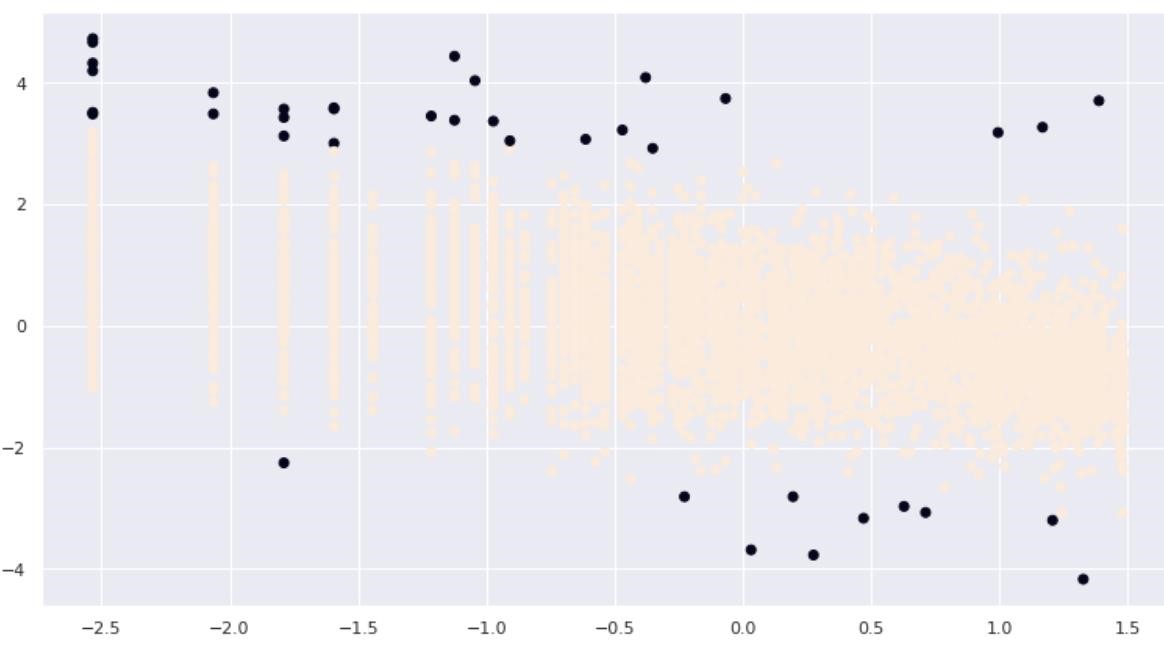
*are well separated when we*

*cluster them by Recency and*

*Monetary.*



### Applying DBSCAN on Recency and Monetary



**From above plot, we can observe that Customers are well separate when we cluster them by Recency and Monetary. We got 2 as optimal number of clusters.**

#### Applying silhouette Score and Elbow Method on Frequency and Monetary

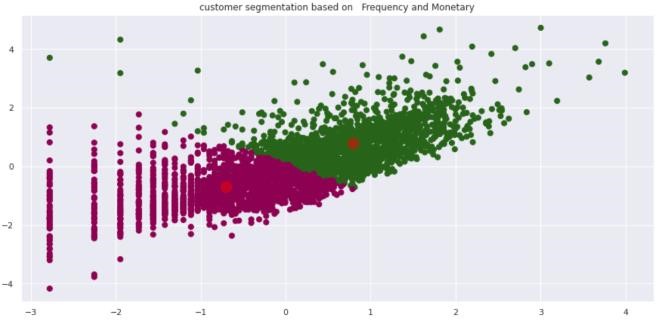
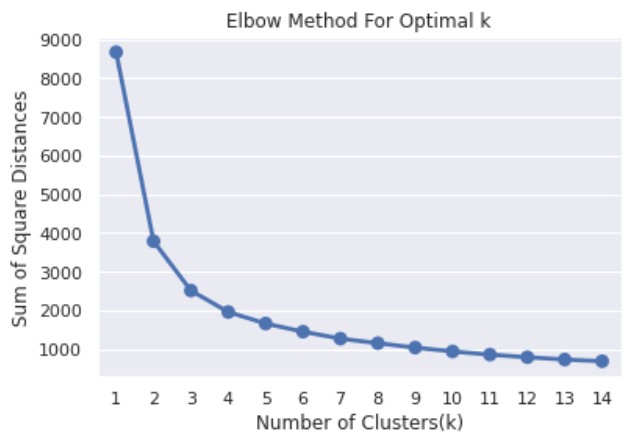
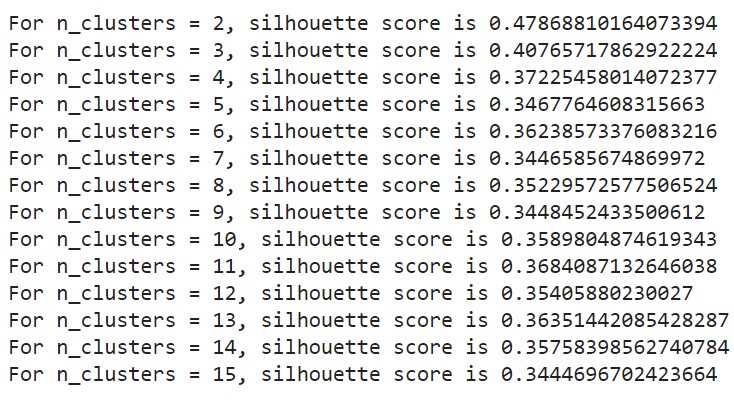
*From this plot, We found*

*Customers are well*

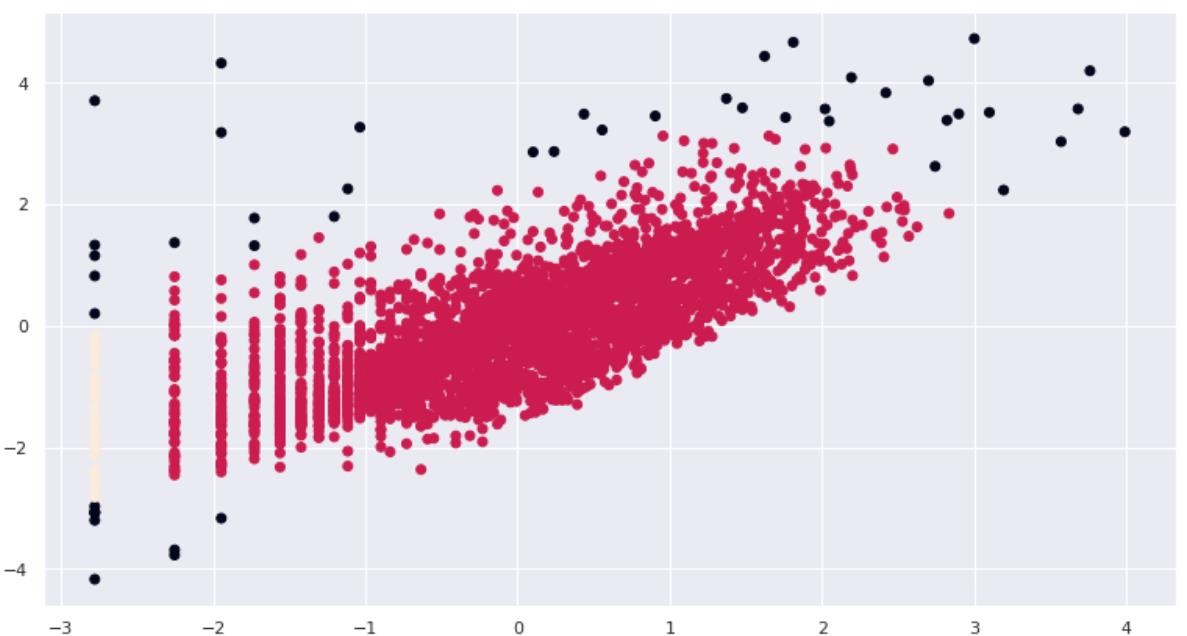
*separated when we cluster*

*them by Frequency and*

*Monetary.*

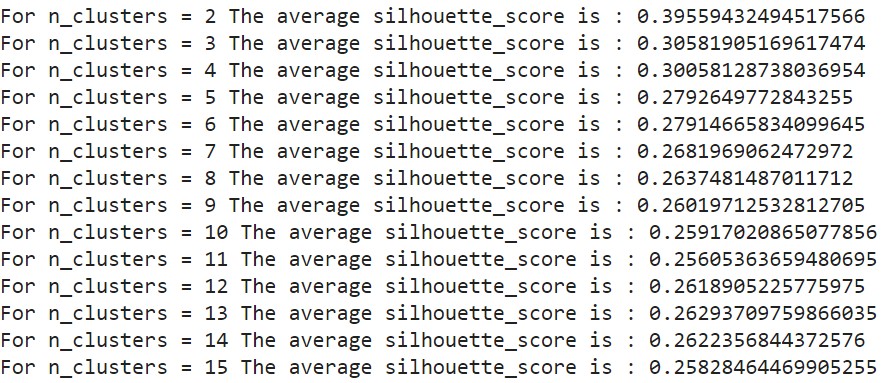


#### Applying DBSCAN on Frequency and Monetary



**We can see that Customers are well separated when we cluster them by Frequency and Monetary. We got 2 as optimal number of clusters.**

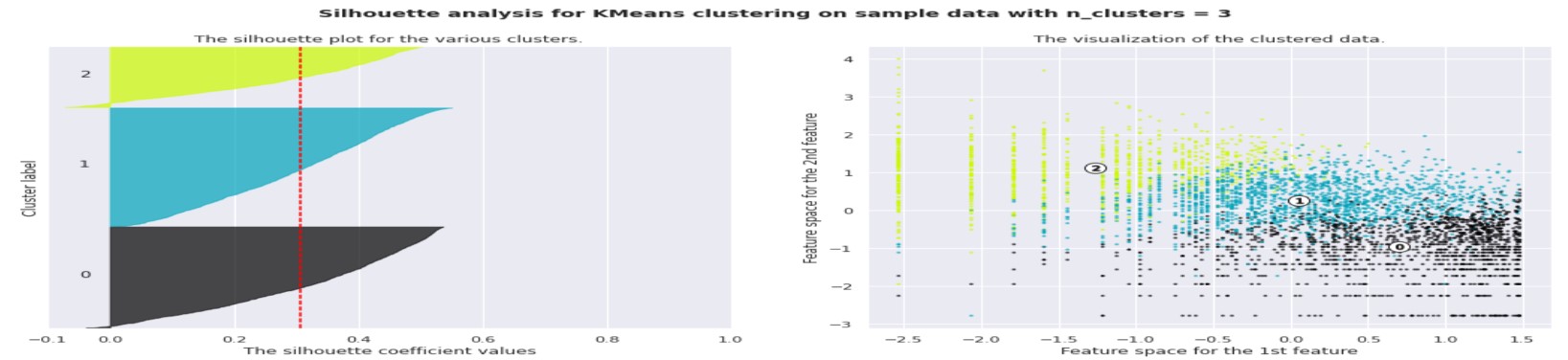
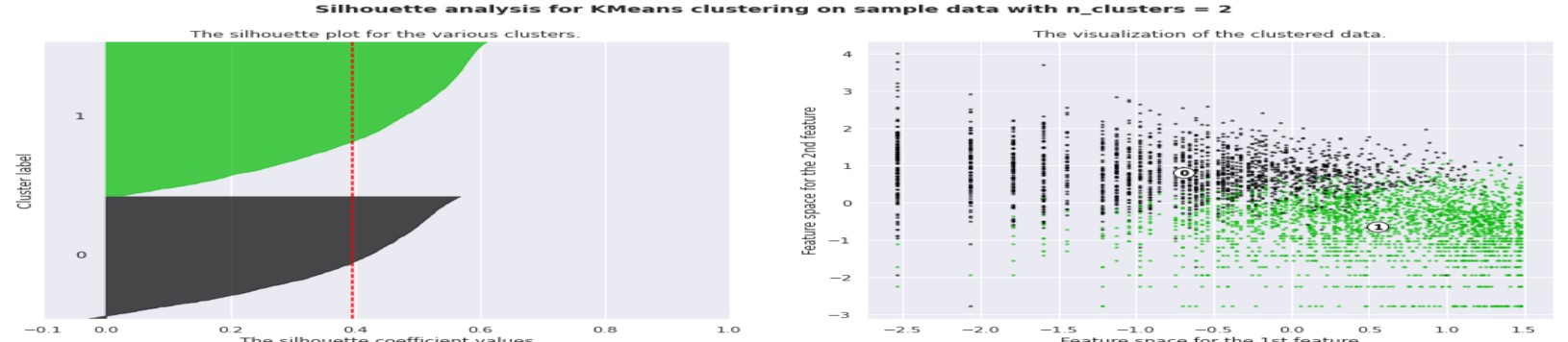
### Applying Silhouette Method on Recency, Frequency and Monetary



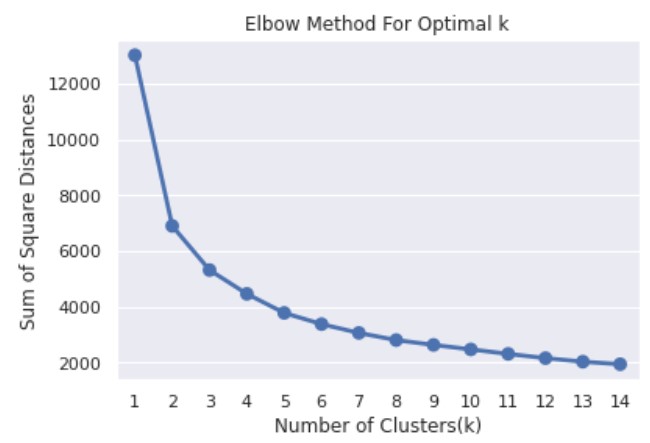
### Applying Silhouette Method on Recency, Frequency and Monetary

No. of cluster = 2

No. of cluster = 3



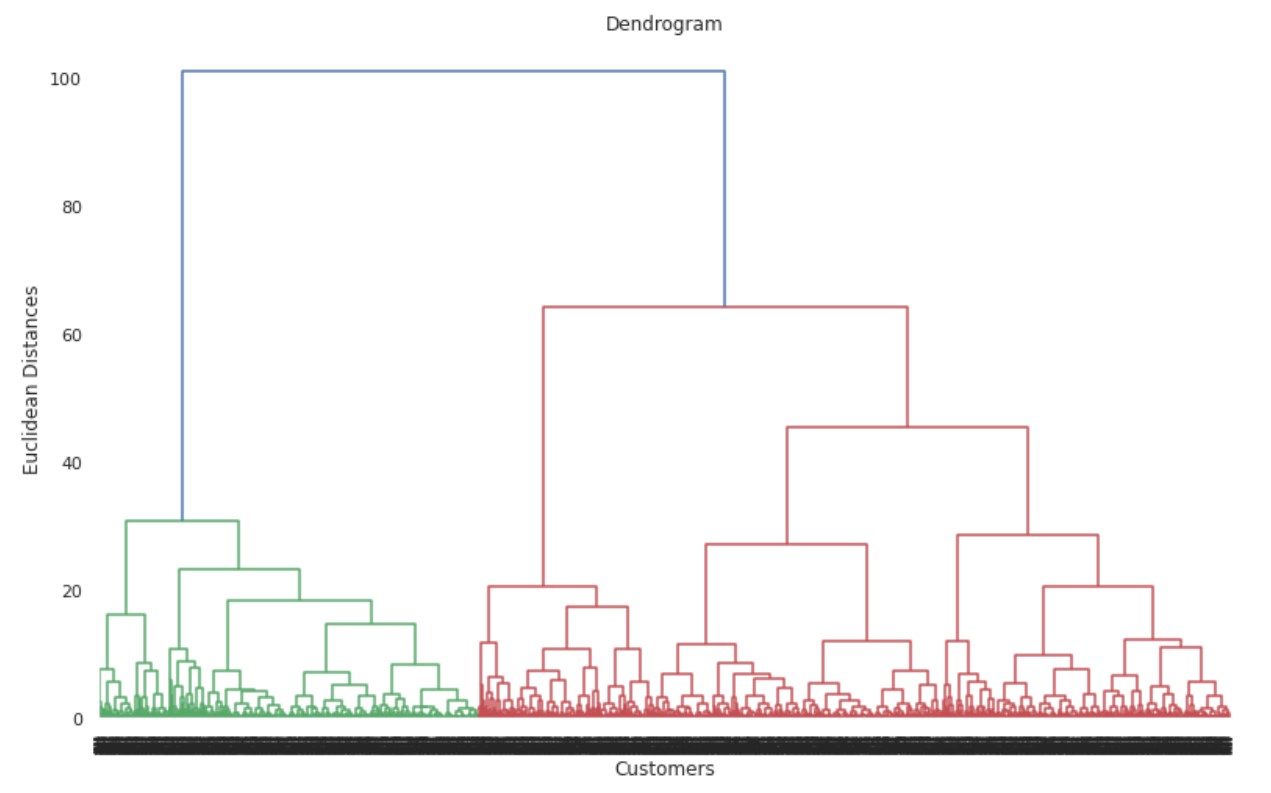
### Applying Elbow Method on Recency, Frequency and Monetary



**This method also gives information about the optimal number of clusters.**

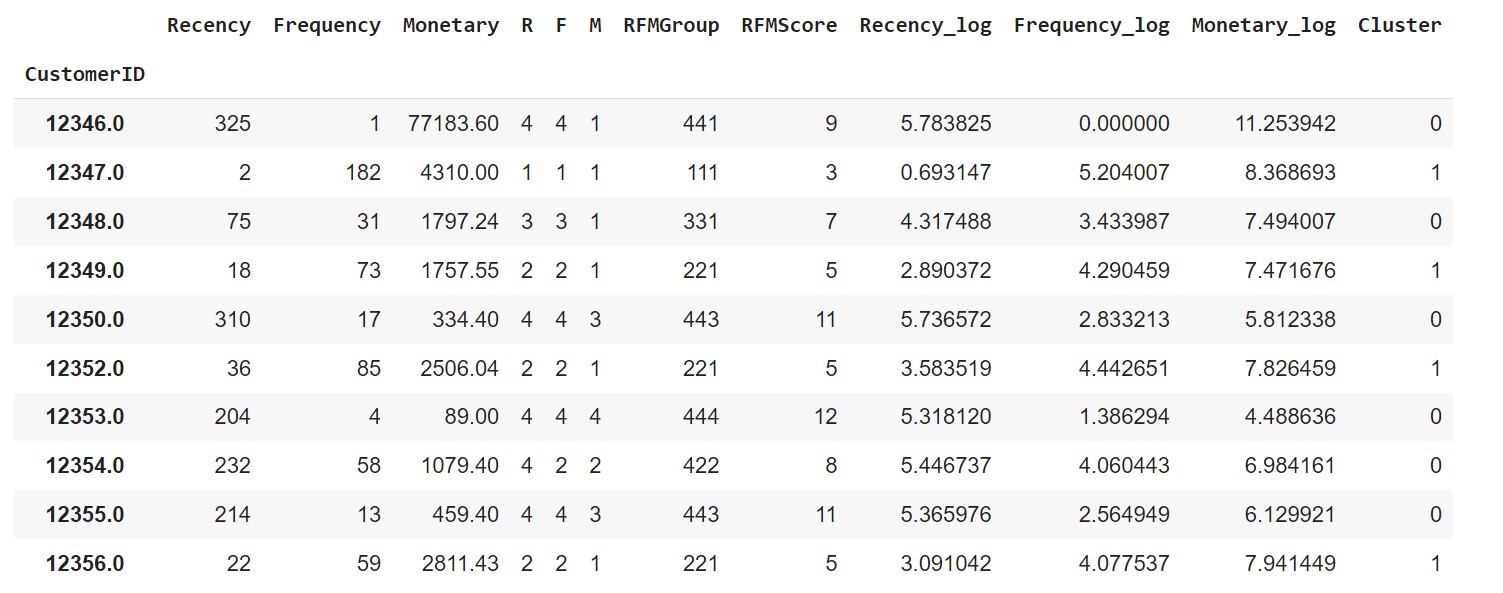
**According to it, 2 is the optimal number of clusters.**

#### Using the Dendrogram to find the optimal number of clusters

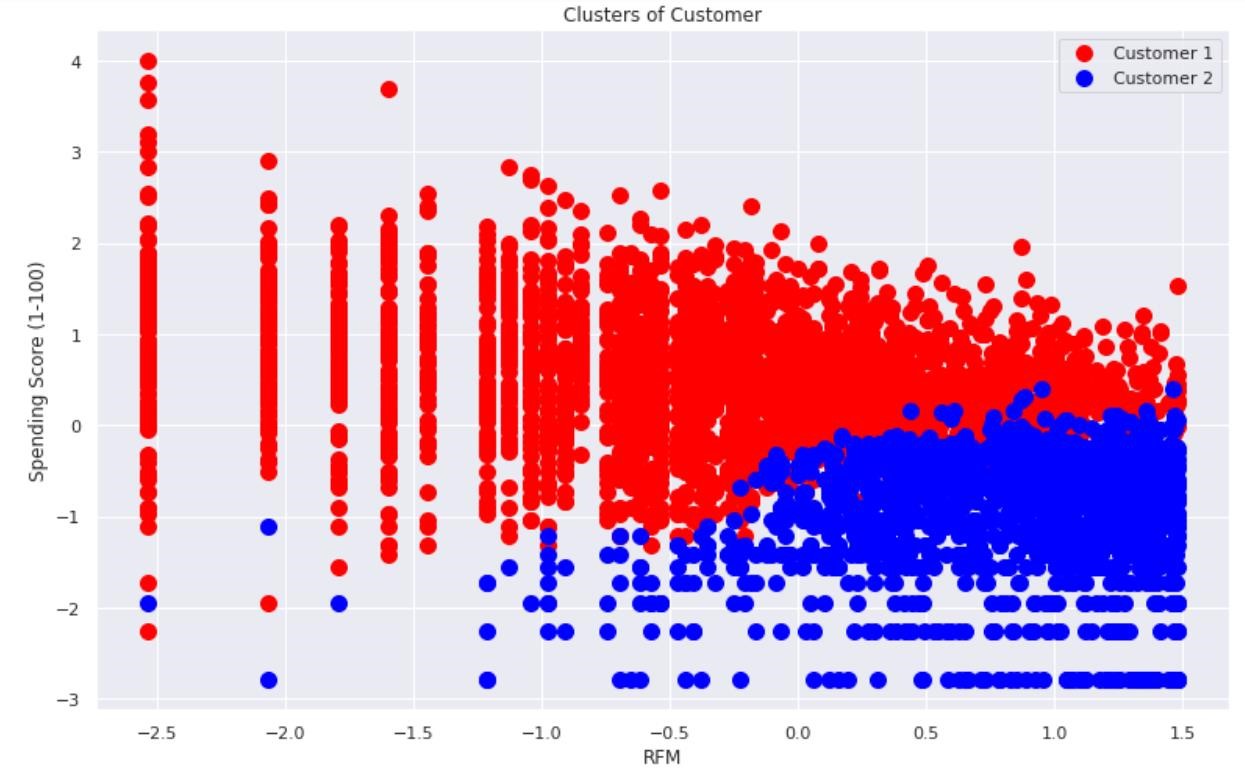


*The number of clusters will be the number of vertical lines which are being intersected by the line drawn using the threshold=90. No. of Cluster = 2*

### Find the clusters on the basis on RFM table



### Visualizing the clusters (two dimensions only)



## Challenges

* Understanding the problem statement. ● Figuring Out right Approach
* Dealing with Null And duplicate values
* Treatment of cancelled orders
* Extracting Datetime Column Properly and creating RFM variables.
* Designing multiple visualizations to summarize the Data points in the dataset and effectively communicating the results and insights to the reader.
* Finding optimal number of clusters 

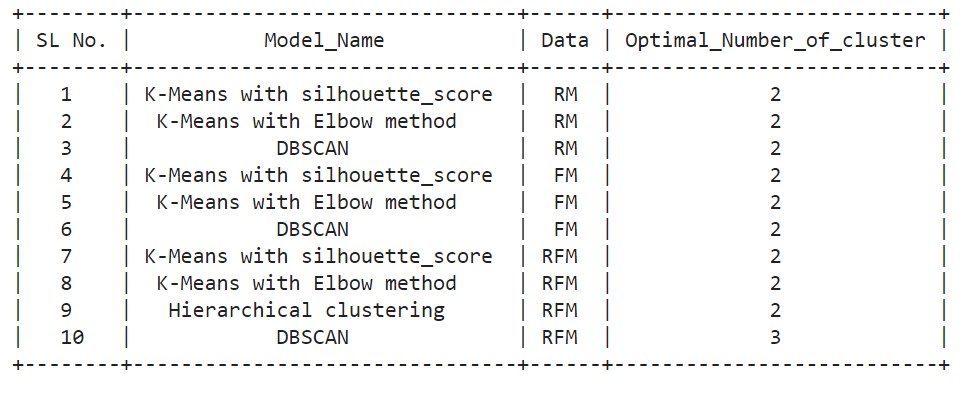
# Conclusion

## Descriptive Analytics

In conclusion, the data exploration of Online customer segmentation dataset shows :

* Missing and duplicate values were found.
* Most of the purchases are from the United Kingdom.
* Most of the customers have purchased items on Thursday, Wednesday, Tuesday.
* Most of the customers have purchased items in November, October, December, and the least number of purchases in April, January, February.
* Most of the customers purchase in the afternoon time. The 12th hour of the day is a peak for purchasing items.

## Conclusion



**By applying different clustering algorithm to our dataset, we get the optimal number of cluster is equal to 2.**

## Final Thought

Customer segmentation is an important marketing approach that businesses should employ in order to gain a better understanding of the market and make more informed decisions in order to increase sales.

K-Means clustering is a basic but effective machine learning algorithm that businesses can use. Finally, in order to optimise our marketing success, we must keep the RFM client segmentation up to date.

## Further analysis

* New variables have been added, such as tenure, which is the number of days since each customer's first transaction. This will reveal how long each customer has been a member of the system.
* Customers are being segmented more deeply based on their physical location, as well as demographic and psychographic factors.
* Incorporating data from the company's Google Analytics account. Google Analytics is an excellent tool for tracking a variety of essential business data, including Customer Lifetime Value, Traffic Source/Medium, Pageviews per Visit, and Bounce Rate of a company's website, among others.

