Capstone Project

Unsupervised ML Clustering

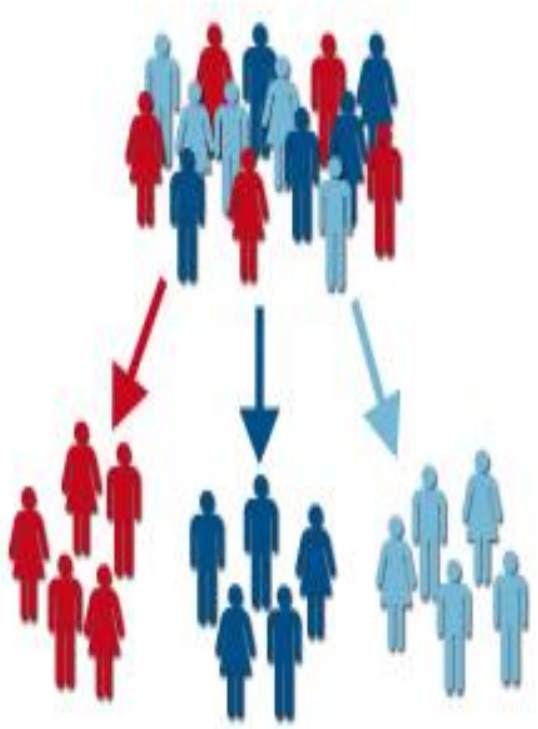
Online retail Customer Segmentation

Team member:

Mohd. Navaid Ansari
Mohd Atif Ansari
Mohammad Anas Ansari

Customer segmentation

- 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.
- A key consideration for a company would be whether or not to segment its customers and how to do the process of segmentation. This would depend upon the company philosophy and the type of product or services it offers. The type of segmentation criterion followed would create a big difference in the way the business operates and formulates its strategy. This is elucidated below.



1. **Zero segments:** <*Undifferentiated approach*> This means that the company is treating all of its customers in a similar manner. In other words, there is no differentiated strategy and all of the customer base is being reached out by a single mass marketing campaign.
2. **One segment:** <*Focused approach*> This means that the company is targeting a particular group or niche of customers in a tightly defined target market.
3. **Two or more segments:** <*Differentiated approach*> This means that the company is targeting 2 or more groups within its customer base and is making specific marketing strategies for each segment.
4. **Thousands of segments:** <*Hyper segmentation approach*> This means that the company is treating each customer as unique and is making a customized offer for each one of them.

Once the company has identified its customer base and the number of segments it aims to focus upon, it needs to decide the factors on whose basis it will decide to segment its customers.

How to segment your customers?

- To start with customer segmentation, a company needs to have a clear vision and a goal in mind. The following steps can be undertaken to find segments in the customer base on a broad level.
- Analyze the existing customer pool: Understanding the geographical distribution, customer preferences/beliefs, reviewing website search page analytics, etc.
- Develop an understanding of each customer: Mapping each customer to a set of preferences to understand and predict their behavior: the products, services, and content they would be interested in.
- Define segment opportunities: Once the segments have been defined, there should a proper business understanding of each segment and its challenges and opportunities. The entire company's marketing strategy can be branched out to cater to different niches of customers.
- Research the segment: After cementing the definition and business relevance of different customer segments, a company needs to understand how to modify its products or services to better cater to them. For example, it may decide to provide higher discounts to some customers compared to others to expand its active customer base.
- Tweak strategy: Experiment with new strategies and understand the impact of time and economy on the purchasing behavior of customers in different segments. And then the process should be repeated for refining the strategy as much as possible.

Dataset Information:

In [ ]:

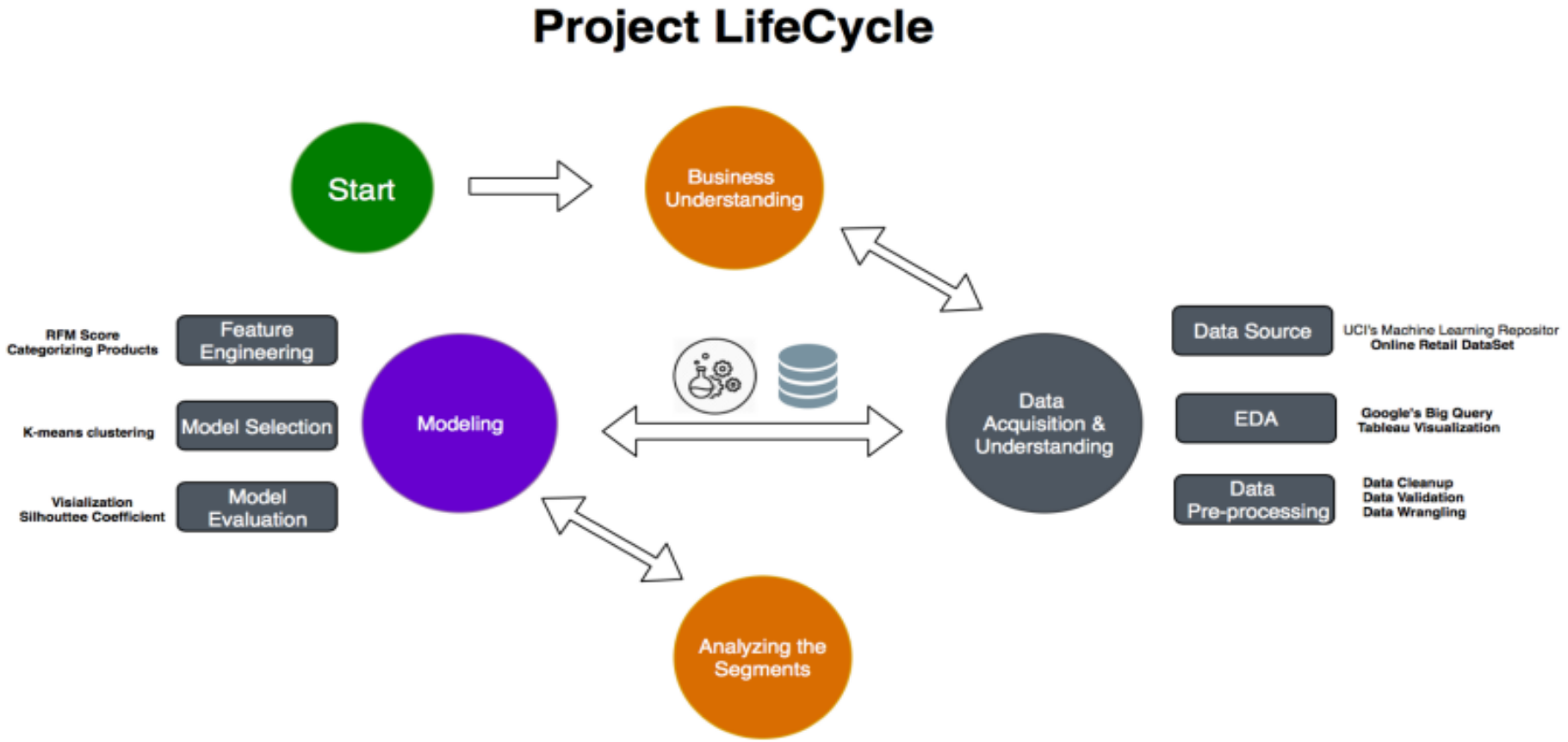
```
# Here is a list of all the columns and rows, how many columns and rows are present in this data set
orcs
```

Out[ ]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680.0	France

541909 rows × 8 columns

Flow Chart


Project LifeCycle
Start
Business Understanding
Data Acquisition & Understanding
Data Source
UCI's Machine Learning Repositor
Online Retail DataSet
EDA
Google's Big Query
Tableau Visualization
Data Pre-processing
Data Cleanup
Data Validation
Data Wrangling
Modeling
Feature Engineering
RFM Score
Categorizing Products
Model Selection
K-means clustering
Model Evaluation
Visialization
Silhouttee Coefficient
Analyzing the Segments

Data Attributes

1.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.
2.StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
3.Description: Product (item) name. Nominal.
4.Quantity: The quantities of each product (item) per transaction. Numeric.
5.InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
6.UnitPrice: Unit price. Numeric, Product price per unit in sterling.
7.CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
8.Country: Country name. Nominal, the name of the country where each customer resides.

Sample Data

```
# check information all about the data
orcs.info()
```

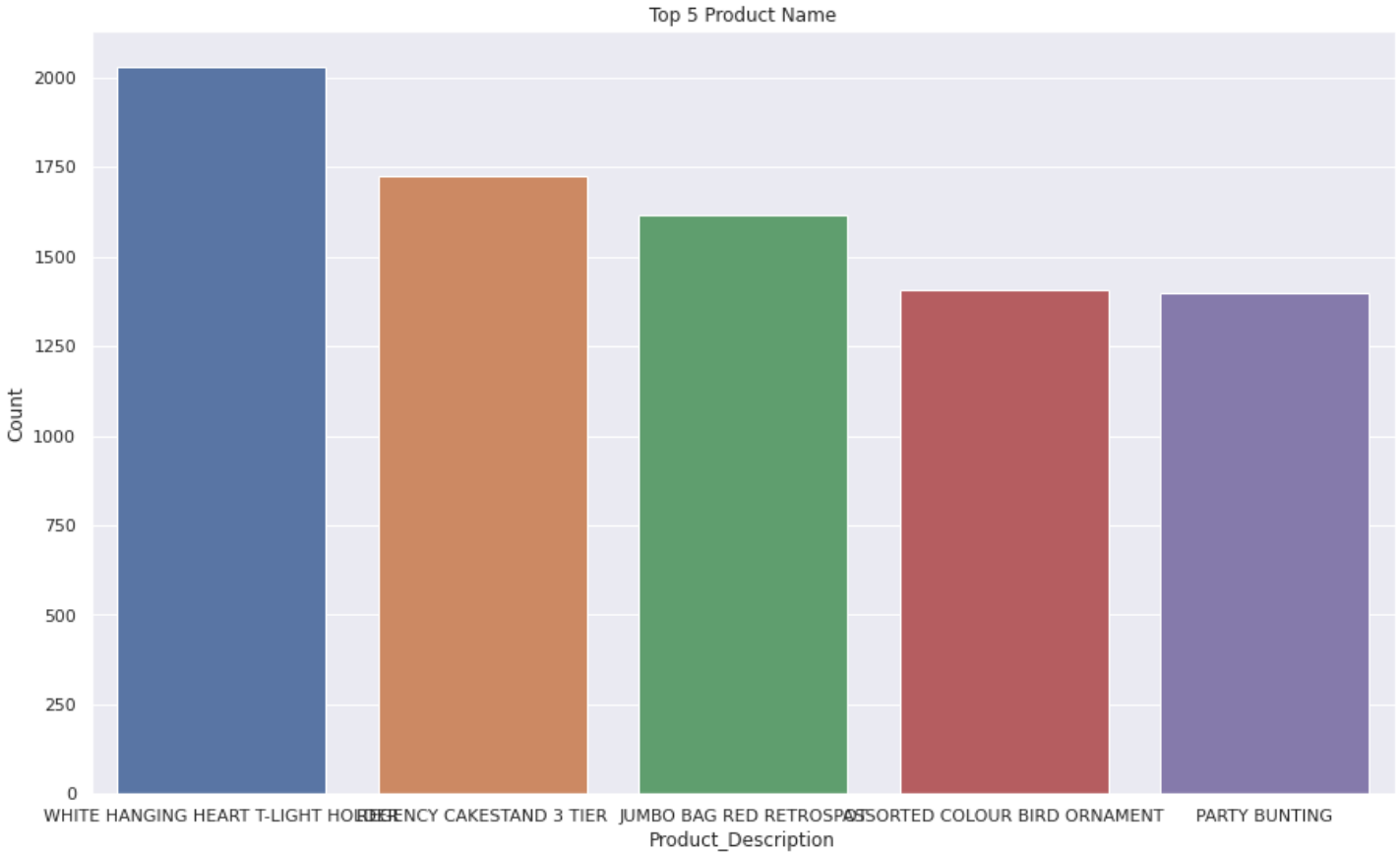
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 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
```

There are column in this sample data as you can see are invoice no., stock code , Description, Quantity, invoice Date , unit price , CustomerID and country

Exploring the data

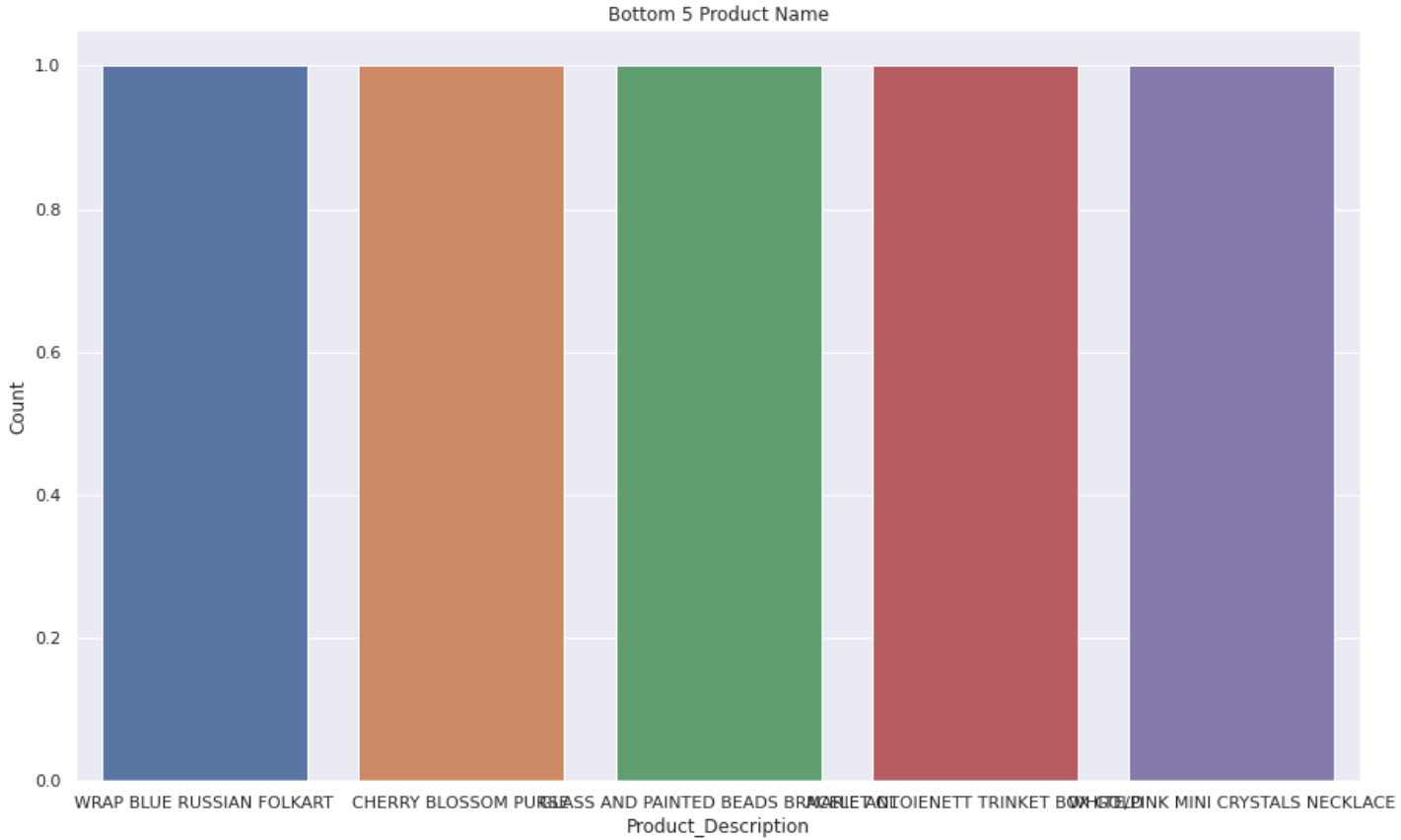
- The data set contains 541909 rows and 8 columns
- Before diving into insights from the data, duplicate entries were removed from the data. The data contained 5268 duplicate entries (about ~1%).
- In our data one column named InvoiceNo has some entries starting with letter "c" stands for cancellation of invoice
- Hence we are dropping those rows which has InvoiceNo start with "c“. Now the data reduced to 397924 rows and 8 columns.
- Let us now look at the total number of products, transactions, and customers in the data, which correspond to the total unique stock codes, invoice number, and customer IDs present in the data.
- Thus, for 4070 products, there are 25900 transactions in the data. This means that each product is likely to have multiple transactions in the data. There are almost as many products as customers in the data as well.
- Since the data, taken from the UCI Machine Learning repository describes the data to based on transactions for a UK-based and registered non-store online retail, let us check the percentage of orders from each country in the data.

Describe Top 5 Product Name:



This is the graph of top 5 product name that's on the top of the list are :
1: White hanging heart t-light holder
2: Regency cakestand 3 tier
3: Jumbo bag red retrospot
4: Assorted colour bird ornament
5: Party bunting

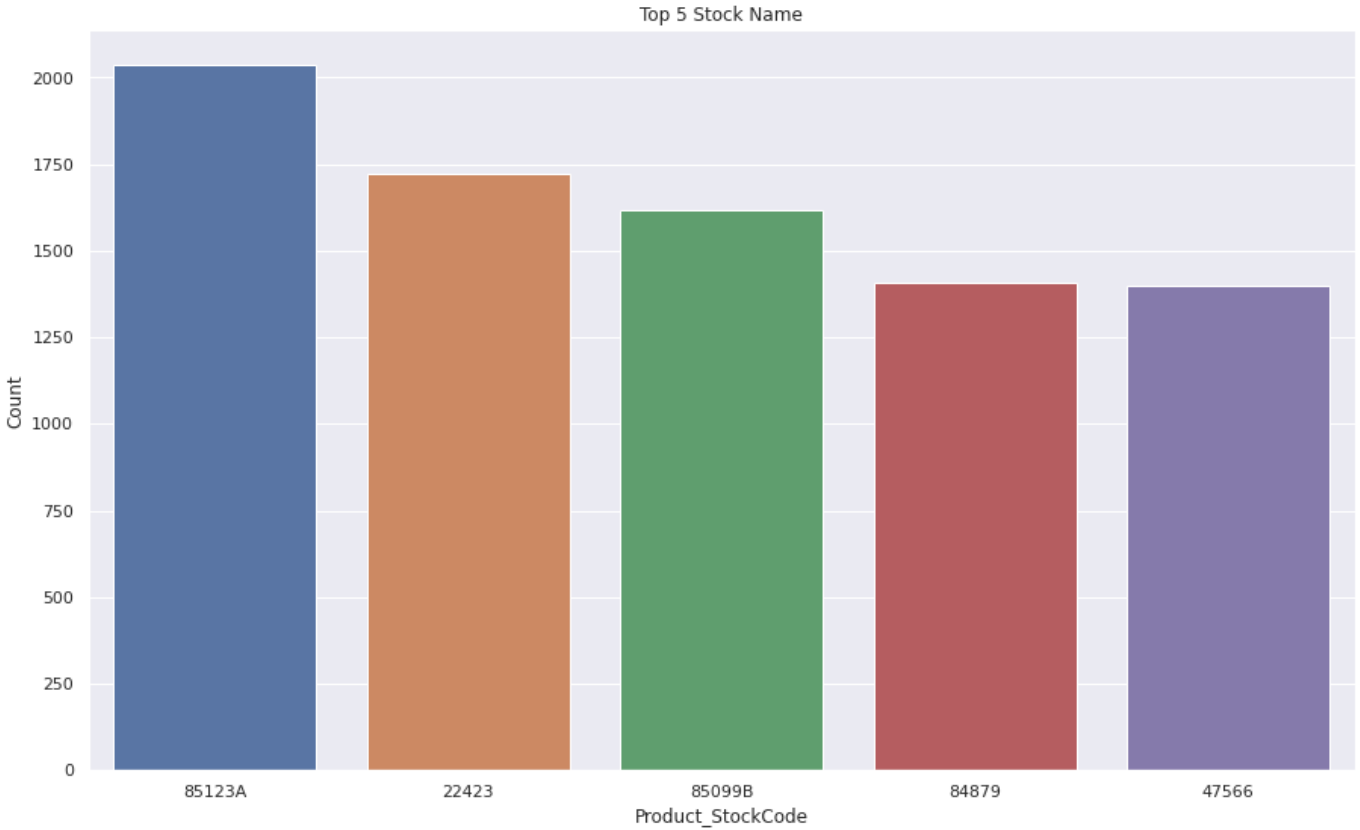
Top 5 Bottom product are:



As you can top 5 bottom product from our analyzation are:

1. Light decoration battery operated
2. Water damaged
3. Throw away
4. Re dotcom quick fix.
5. Birthday banner tape.

Top 5 Stock Name



Top 5 stock name are:
1.85123A
2.22423
3.85099B
4.47566
5.20725

Customers belong to country

We analyses the customer which country they belong in ascending order are:
From first Germany , France , Eire , Spain

Then from last :
Lithuania , Brazil , Czech republic , Bahrain and Saudi Arabia

In [ ]:
```
# Number of customer belogns to country in descending order
country_df=orcs['Country'].value_counts().reset_index()
country_df.rename(columns={'index': 'Country_Name'}, inplace=True)
country_df.rename(columns={'Country': 'Count'}, inplace=True)
country_df.head()
```

Out[ ]:

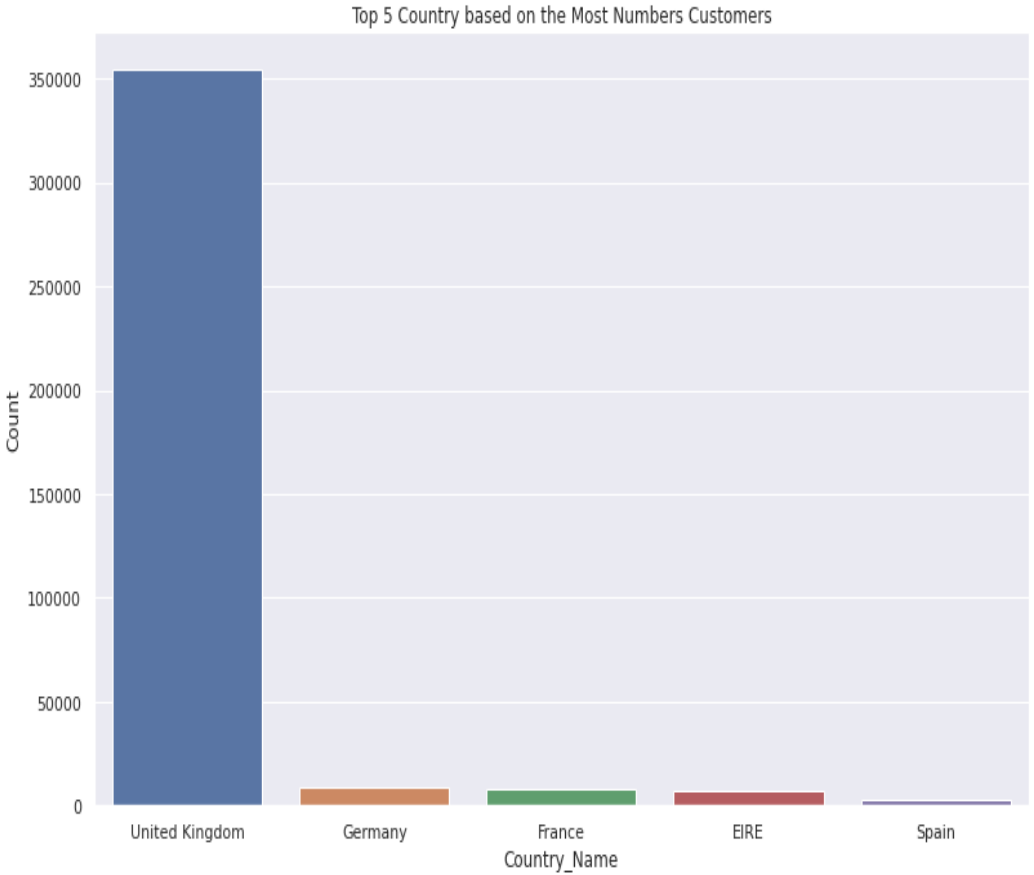
	Country_Name	Count
0	United Kingdom	354345
1	Germany	9042
2	France	8342
3	EIRE	7238
4	Spain	2485

In [ ]:
```
country_df.tail()
```

Out[ ]:

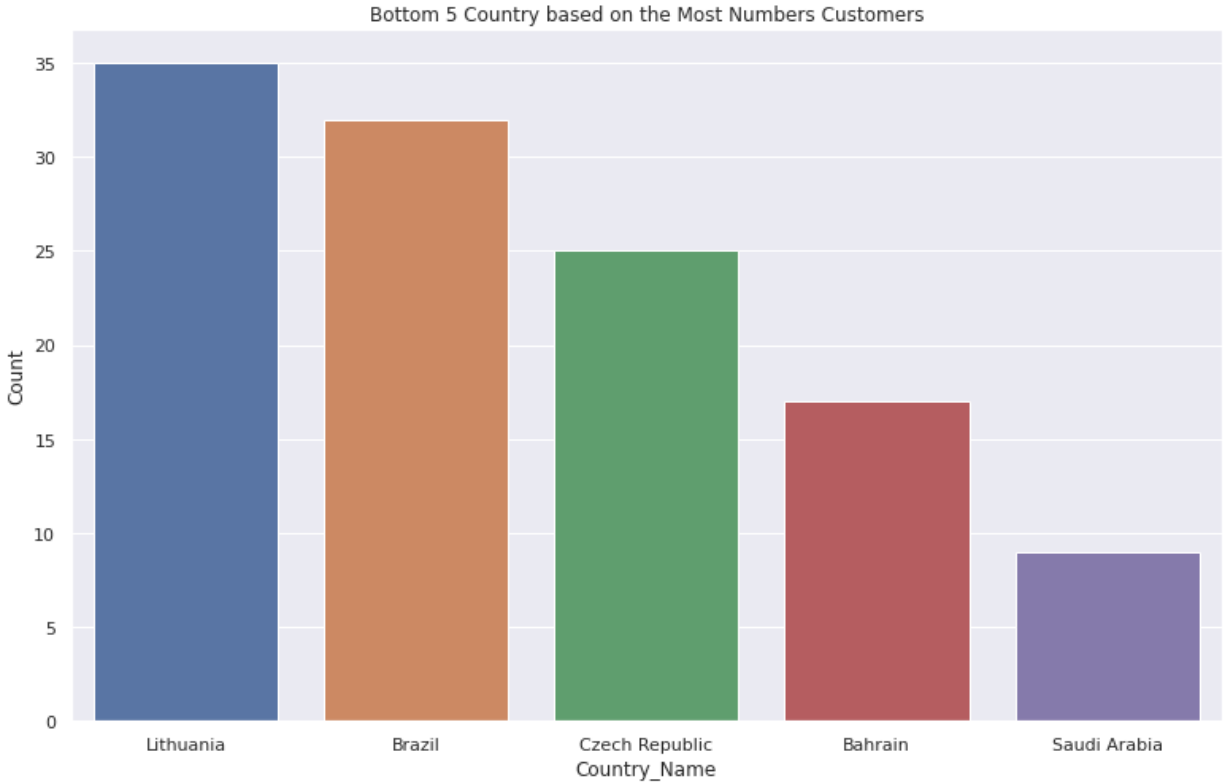
	Country_Name	Count
32	Lithuania	35
33	Brazil	32
34	Czech Republic	25
35	Bahrain	17
36	Saudi Arabia	9

Top countries most customers:



From this graph we can see that most of the customers are from United Kingdom ,Germany ,France ,EIRE and Spain

Bottom 5 countries the most customers come



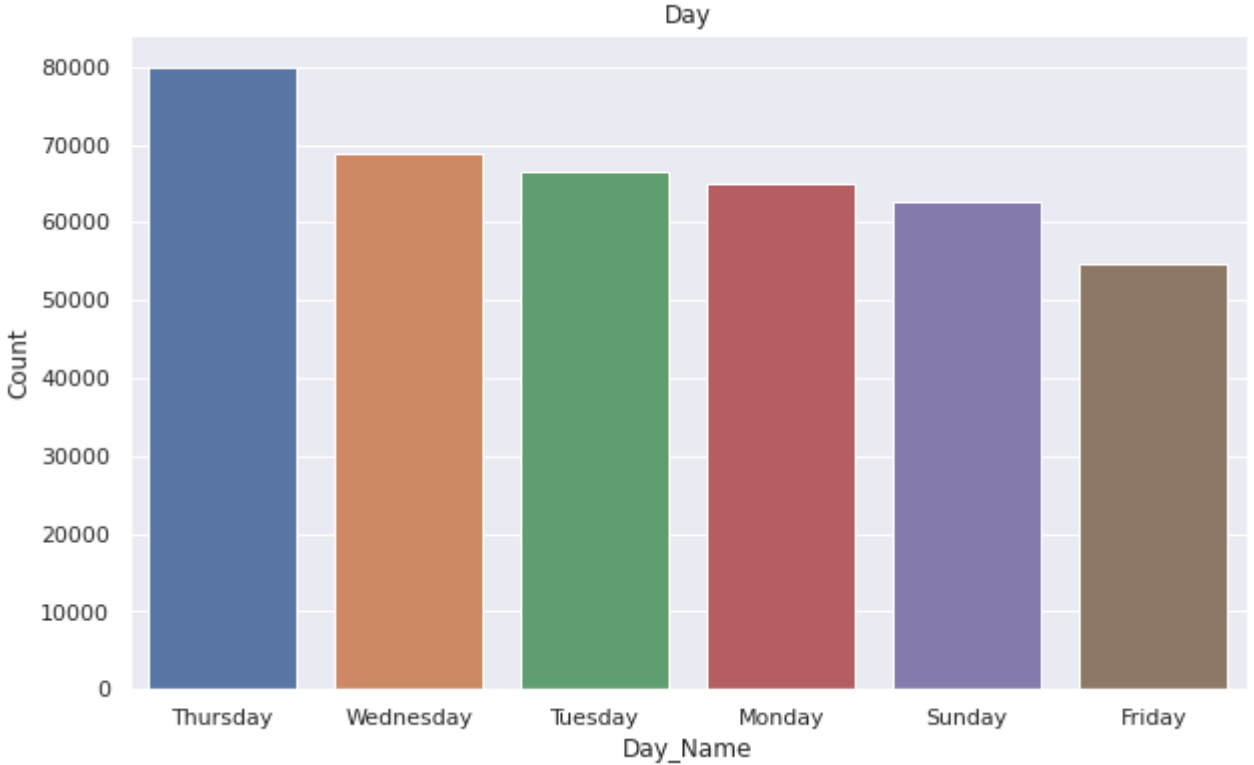
These are the 5 most countries where customers come are Lithuania , Brazil , Czech republic, Bahrain, Saudi Arabia

Feature Engineering

- Convert InvoiceDate column into date time format
- Creating new features from invoice date
- Creating new columns by extracting year, month, date, hour, minute and second from invoice date
- Preparing data to run on RFM model by creating new column “total amount = quantity * unitprice”.

Days when customers shopped

Most customer shopped on the Thursday, Wednesday and Tuesday

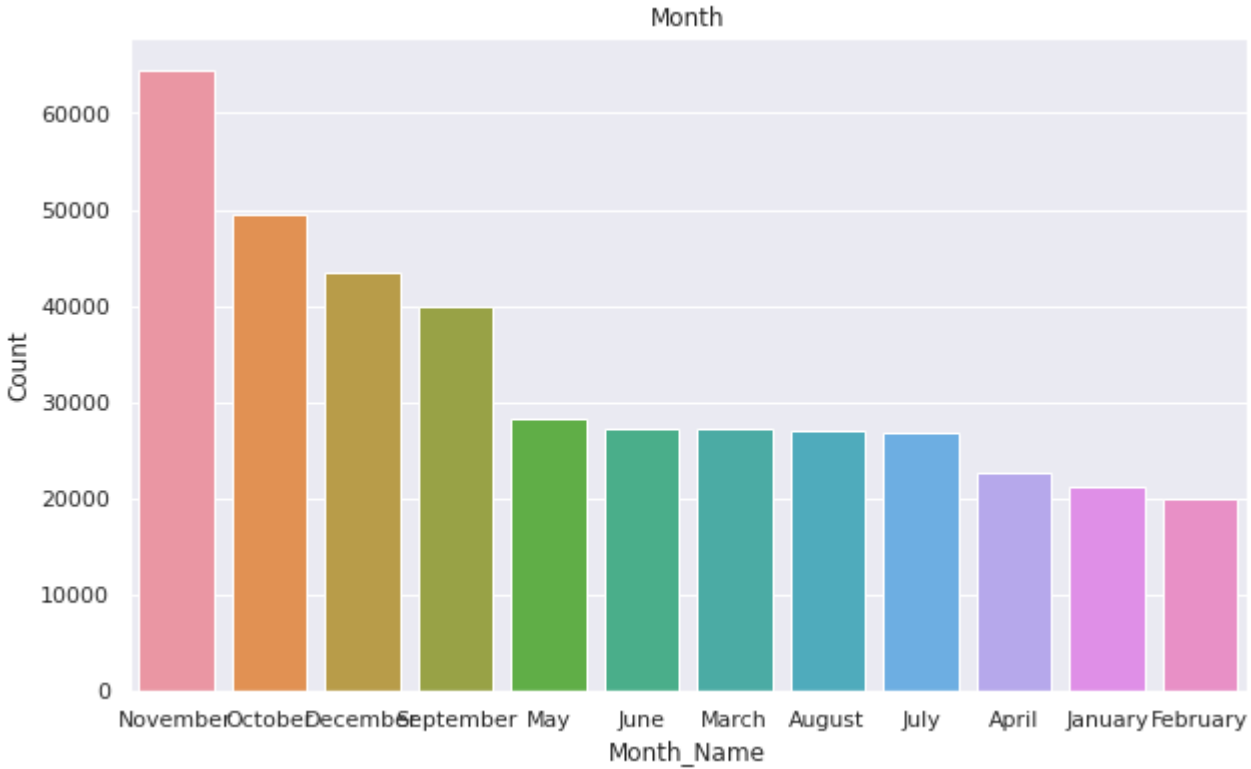


Maximum sale on the month

	Month_Name	Count
0	November	64531
1	October	49554
2	December	43461
3	September	40028
4	May	28320
5	June	27185
6	March	27175
7	August	27007
8	July	26825
9	April	22642
10	January	21229
11	February	19927

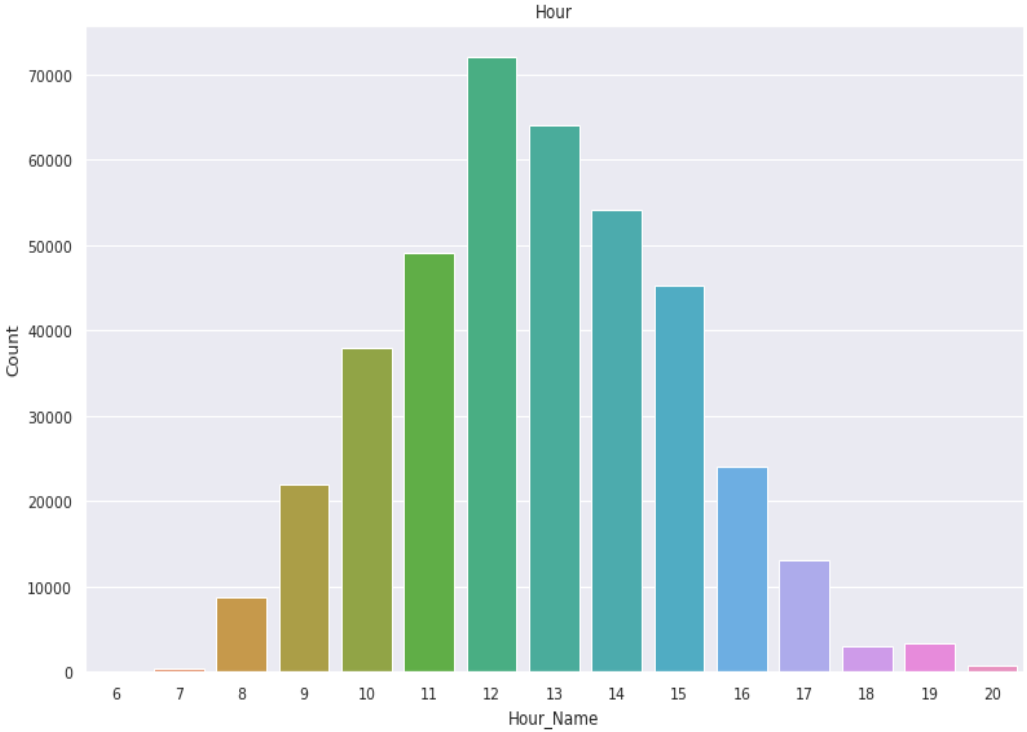
Most numbers of customers have parches the gifts in the month of November ,October and December September
And the values are :
November: 64531
October : 49554
December : 43461
September: 40028
May: 28320
So on as you see in the graph

Month Wise Sale

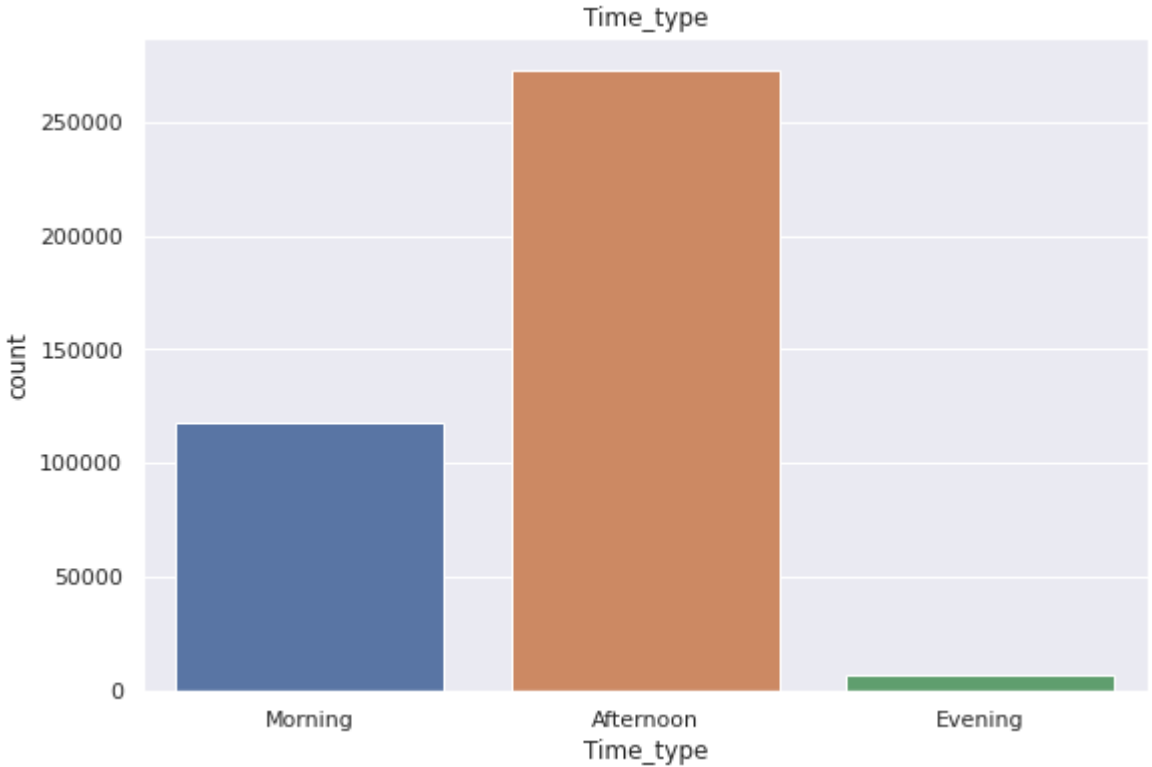


Hours

From this graph we can see that in Afternoon Time most of the customers have purchase the item



dividing in the three group of morning, afternoon and evening time



Create the RFM model (Recency, Frequency and Monetary value)

- RFM is a method used to analyze customer value. RFM stands for RECENCY, Frequency, and Monetary.
- 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?

- Performing RFM model step by step
  - The first step in building an RFM model is to assign Recency, Frequency and Monetary values to each customer
  - The second step is to divide the customer list into tiered groups for each of the three dimensions (R,F and M)
- Calculating RFM score
  - The number is typically 3 or 5. If you decide to code each RFM attribute into 3 categories, you'll end up with 27 different coding combinations ranging from a high of 333 to a low of 111. Generally speaking, the higher the RFM score, the more valuable the customer.

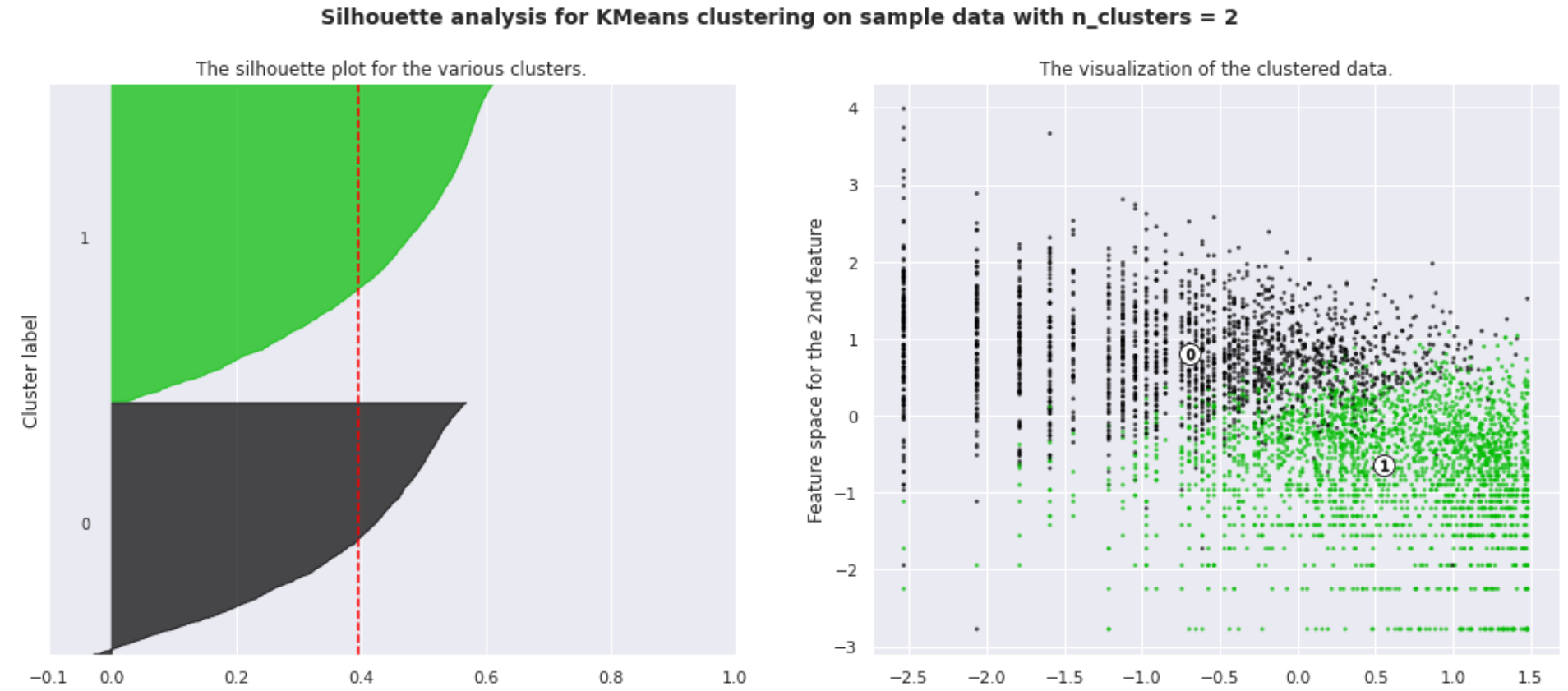
Create the RFM model (Recency, Frequency and Monetary value)

```
#Descriptive Statistics (Recency)
rfm_df.Recency.describe()
```

```
count    4338.000000
mean       92.059474
std       100.012264
min         0.000000
25%        17.000000
50%        50.000000
75%       141.750000
max       373.000000
Name: Recency, dtype: float64
```

- This gives us an idea of how consumer spending is distributed in our data. We can see that the mean value is 92.05 and the standard deviation is 100.01. But the maximum value is 373. This is a very large value. Therefore, the Total Sum values in the Top 25% of our data increase very rapidly from 141.75 to 373.
- Now, for RFM analysis, we need to define a ‘snapshot date’, which is the day on which we are conducting this analysis. Here, I have taken the snapshot date as the highest date in the data + 1 (The next day after the date till which the data was updated). This is equal to the date 2011–12–10. (YYYY-MM-DD)
- Next, we confine the data to a period of one year to limit the recency value to a maximum of 365 and aggregate the data on a customer level and calculate the RFM metrics for each customer.

KMeans Clustering With 2 clusters

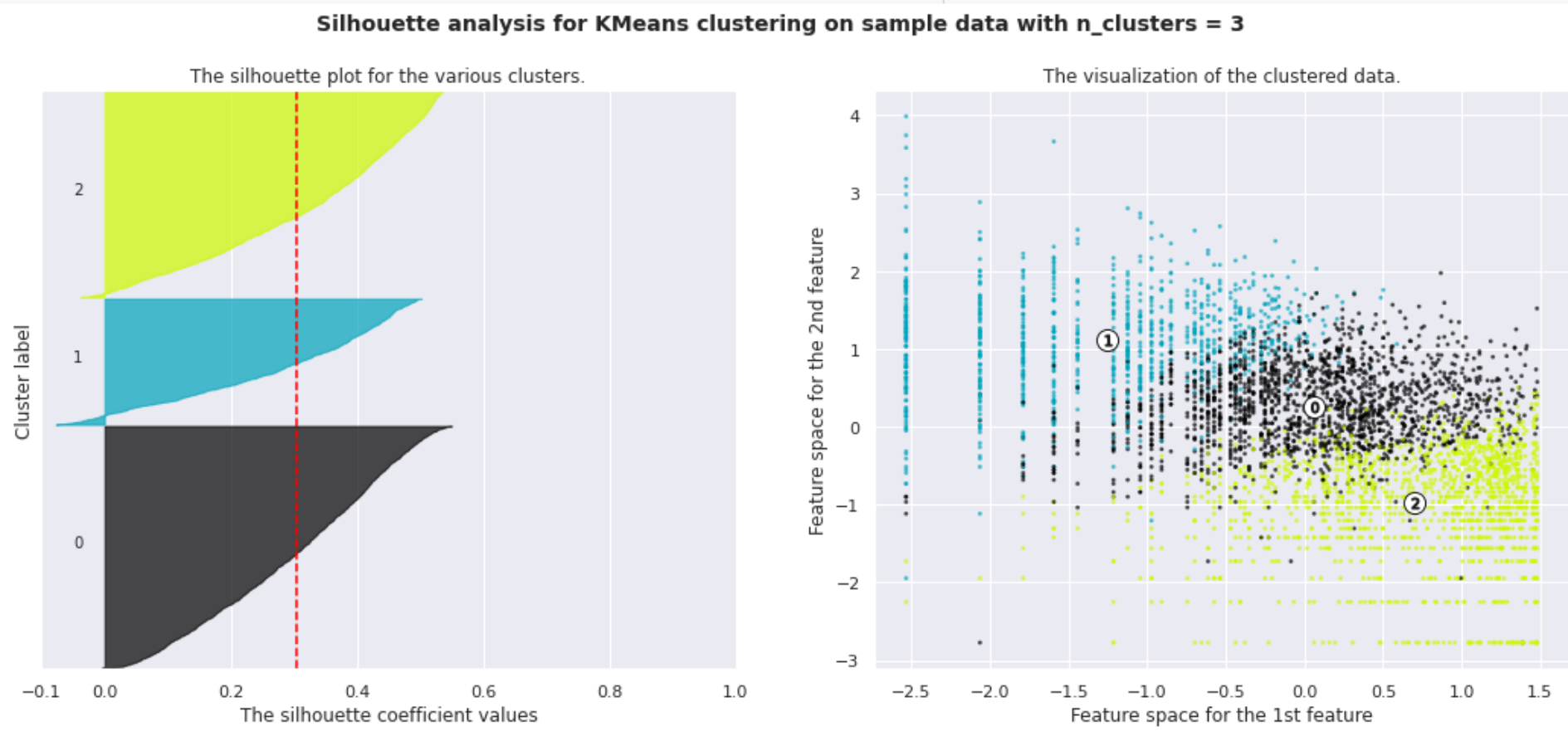


Silhouette Coefficient or Silhouette Score

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

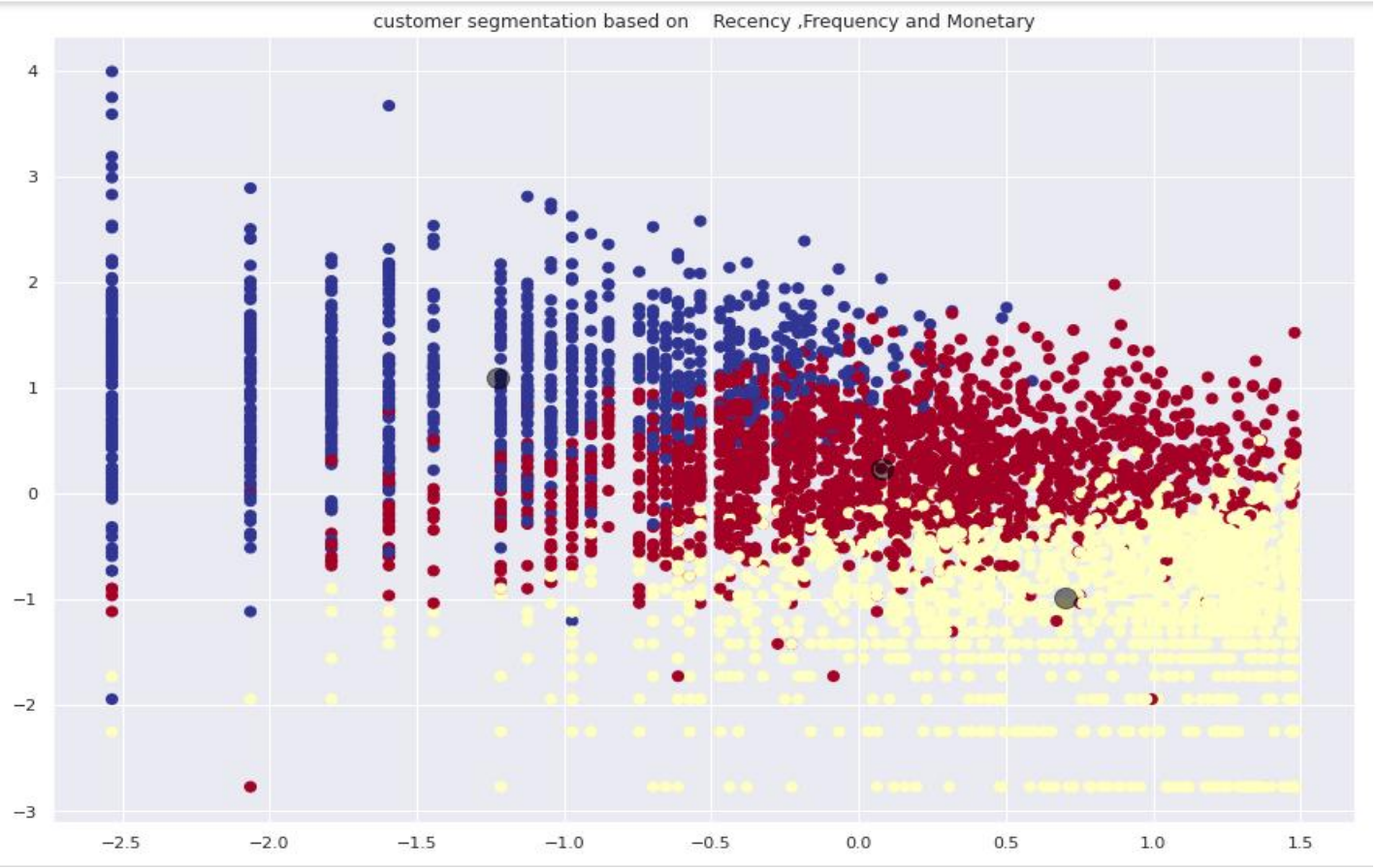
- 1 : Means clusters are well apart from each other and clearly distinguished.
- 0 : Means clusters are indifferent, or we can say that the distance between clusters is not significant.
- -1 : Means clusters are assigned in the wrong way.

KMeans Clustering With 3 clusters

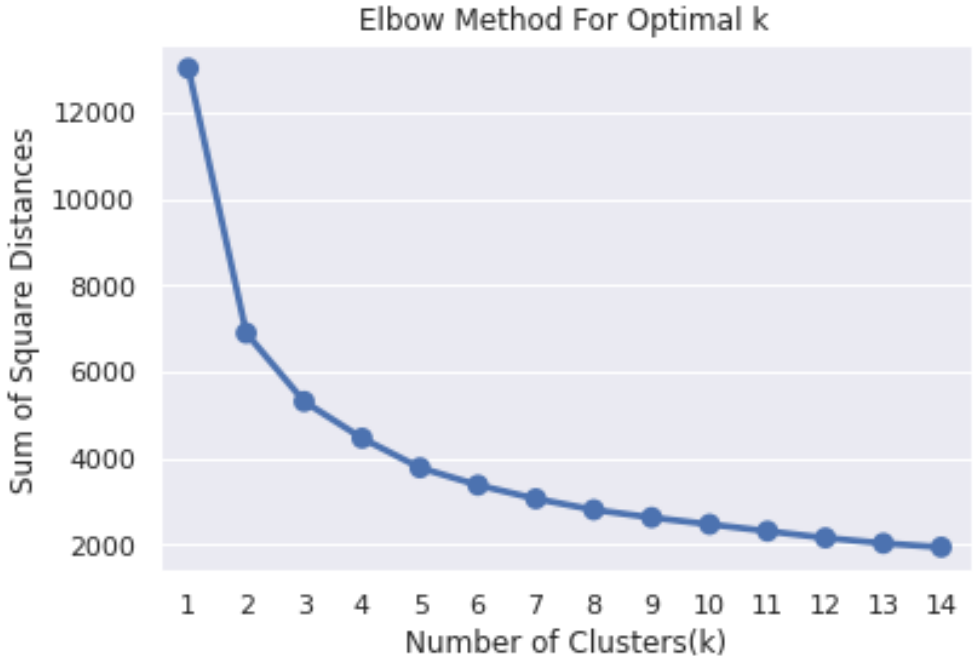


Based on Recency, Frequency And Monetary With 2

Based on Recency, Frequency And Monetary With 3



Elbow method for clustering for K = 1 -14



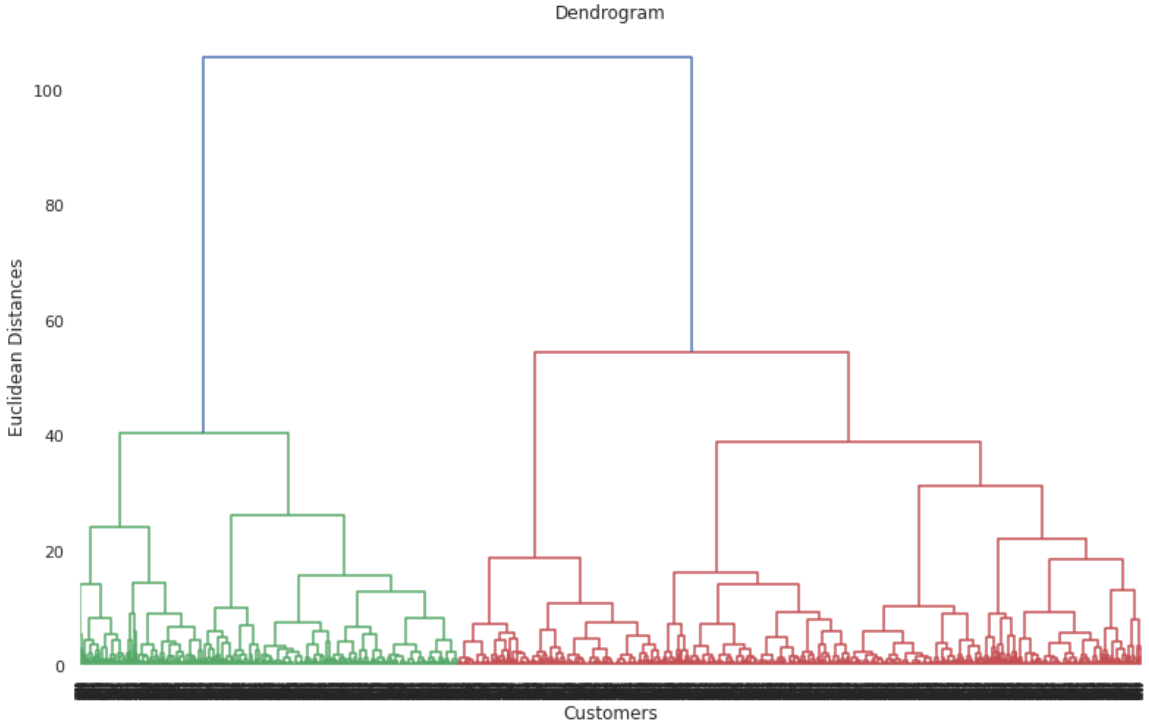
The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

Perform K-Mean Clustering

Out[123...

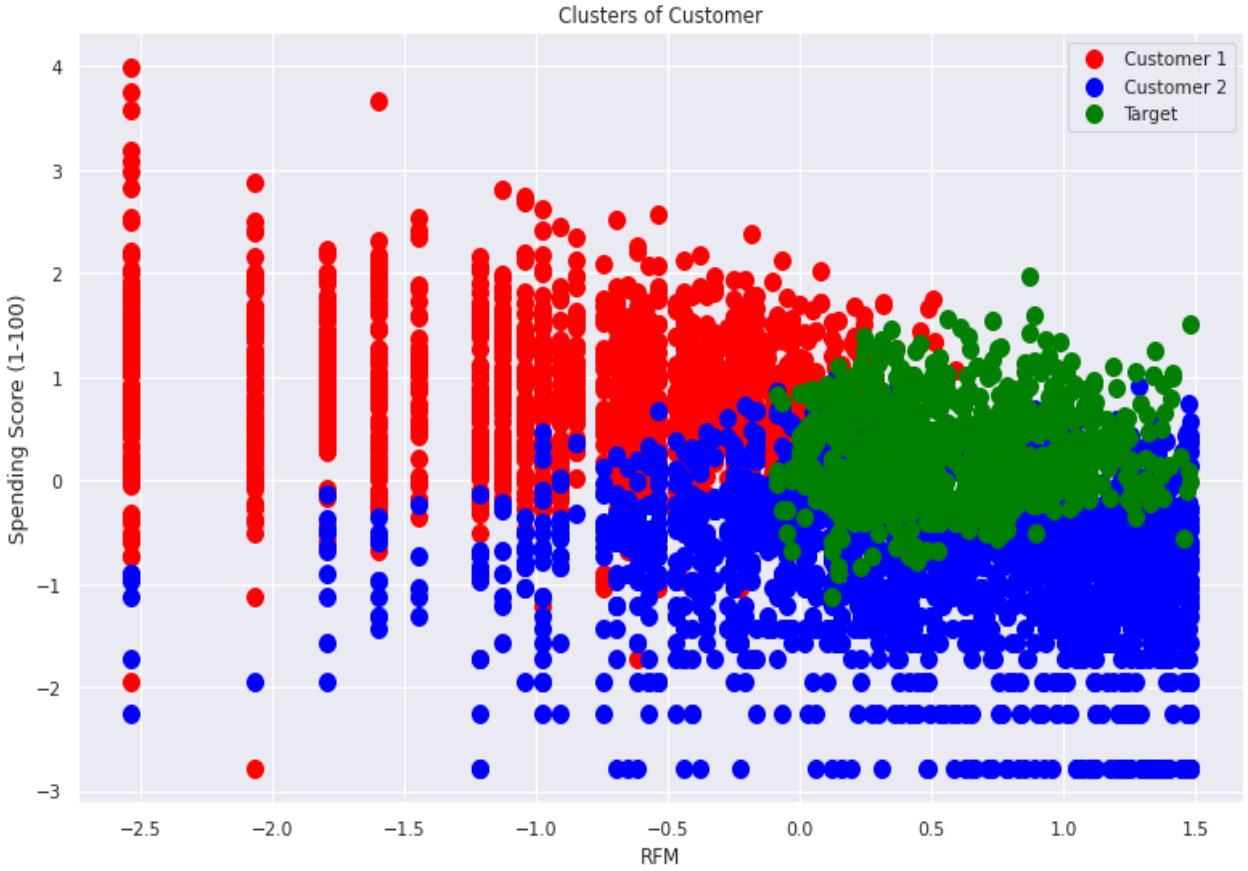
CustomerID	Recency	Frequency	Monetary	R	F	M	RFMGroup	RFMScore	Recency_log	Frequency_log	Monetary_log	Cluster
12346.0	325	1	77183.60	4	4	1	441	9	5.783825	0.000000	11.253942	2
12347.0	2	182	4310.00	1	1	1	111	3	0.693147	5.204007	8.368693	1
12348.0	75	31	1797.24	3	3	1	331	7	4.317488	3.433987	7.494007	2
12349.0	18	73	1757.55	2	2	1	221	5	2.890372	4.290459	7.471676	2
12350.0	310	17	334.40	4	4	3	443	11	5.736572	2.833213	5.812338	0
12352.0	36	85	2506.04	2	2	1	221	5	3.583519	4.442651	7.826459	2
12353.0	204	4	89.00	4	4	4	444	12	5.318120	1.386294	4.488636	0
12354.0	232	58	1079.40	4	2	2	422	8	5.446737	4.060443	6.984161	2
12355.0	214	13	459.40	4	4	3	443	11	5.365976	2.564949	6.129921	0
12356.0	22	59	2811.43	2	2	1	221	5	3.091042	4.077537	7.941449	2

Elbow method for clustering for K = 1 -14



A **dendrogram** is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering.

Fitting hierarchical clustering to the all dataset



These illustrates us the all hierarchy clustering

By applying different clustering algorithm to our dataset .we get the optimal number of cluster is equal to 3

Summary and result

```
from prettytable import PrettyTable

# Specify the Column Names while initializing the Table
myTable = PrettyTable(['SL No.',"Model_Name",'Data', "Optimal_Number_of_cluster"])

# Add rows
myTable.add_row(['1',"K-Means with silhouette_score ", "RFM", "3"])
myTable.add_row(['2',"K-Means with Elbow methos  ", "RFM", "3"])
myTable.add_row(['3',"Hierarchical clustering  ", "RFM", "3"])
print(myTable)
```

```
+--------+---------------------------------+------+---------------------------+
| SL No. |           Model_Name            | Data | Optimal_Number_of_cluster |
+--------+---------------------------------+------+---------------------------+
|   1    | K-Means with silhouette_score   | RFM  |             3             |
|   2    |  K-Means with Elbow methos      | RFM  |             3             |
|   3    |   Hierarchical clustering       | RFM  |             3             |
+--------+---------------------------------+------+---------------------------+
```

Conclusion:

It is critical requirement for business to understand the value derived from a customer. RFM and cohort analysis is a method used for analyzing customer value. Business optimization can be achieved with the above RFM customer segmentation with having segregated the customer base into groups of individuals based on well defined characteristics and traits. Visualization is added to implement the user story with relevant charts. Necessary promotion campaigns with aggressive price incentives and discounts can help monitor customer attrition.

Reference:

- 1) https://www.almabetter.com/
- 2) https://www.wikipedia.org
- 3) https://www.kaggle.com/
- 4) https://github.com/