Data Science Project 1

Machine Learning Project:

Project 1: Customer Segmentation and Transactional Analysis

SUBMITTED BY:

Farjad Ahmad Mir

RIMT University

18010102033

B.Tech (CSE)

SUBMITTED TO:

SachTech Solutions Pvt.Ltd

TABLE OF CONTENTS

Chapter 1: Introduction

A very important skill set for data scientists is to match the technical aspects of analytics with its business value, i.e., its monetary value.

This can be done in a variety of ways and is very much dependent on the type of business and the data available. We can use our data science tools and techniques to tackle the framed business problems and link it directly with the business's revenue generation.

Customer Segmentation: Segmentation is the process of segregating any aggregated entity into separate parts or groups (segments). These parts may or may not share something common across them.

Objective: My main objective is to directly focus on two very important problems that can directly have a positive impact on the revenue streams of businesses and establishments particularly from the retail domain.

Why I did this project: My main objective from beginning of my data science career has been to gather valuable experience in tackling the various problems faced by many business's in our country especially the retail sector. I have already started collecting raw Unstructured Data from many retailers from various backgrounds and have started my work on improving their business.

This project is going to help me by expanding my tools which I can use in my future problems and applications.

I will learn to work on this project by following the standardized models and processes used by the Industry.

Chapter 2: The CRISP-DM Process Model

Introduction:

The CRISP-DM model stands for CRoss Industry Standard Process for Data Mining. CRISP-DM is a tried, tested, and robust industry standard process model followed for data mining and analytics projects. CRISP-DM clearly depicts necessary steps, processes, and workflows for executing any project right from formalizing business requirements to testing and deploying a solution to transform data into insights.

Data Science, Data Mining, and Machine Learning are all about trying to run multiple iterative processes to extract insights and information from data. Hence, we can say that analyzing data is truly both an art as well as a science, because it is not always about running algorithms without reason; a lot of the major effort involves in understanding the business, the actual value of the efforts being invested, and proper methods to articulate end results and insights.

The CRISP-DM model tells us that for building an end-to-end solution for any analytics project or system, there are a total of six major steps or phases, some of them being iterative.

Processes:

1. Business Understanding.

2. Data Understanding.
3. Data Preparation.
4. Modeling
5. Evaluation
6. Deployment

I will be following all the necessary steps until we achieve our goal.

I will provide all the information and steps I will be using in a detailed manner so that no-one will have any problem understanding the tools and techniques opted for.

Chapter 3: Business Understanding

This is the initial phase before kick starting any project in full flow. However, this is one of the most important phases in the lifecycle! The main objective here starts with understanding the business context and requirements for the problem to be solved at hand.

Benefits of the Project: One of the primary objectives of a customer segmentation process is a deeper understanding of a firm's customers

and their attributes and behavior. These insights into the customer base can be used in different ways

Targeted Marketing: The most visible reason for customer segmentation is the ability to focus marketing efforts effectively and efficiently. If a firm knows the different segments of its customer base, it can devise better marketing campaigns which are tailor made for the segment.

Optimal Product Placement: A good customer segmentation strategy can also help the firm with developing or offering new products. This benefit is highly dependent on the way the segmentation process is leveraged.

Higher Revenue: This is the most obvious requirement of any customer segmentation project. The reason being that customer segmentation can lead to higher revenue due to the combined effects of all the advantages identified by the model.

Chapter 4: Data Understanding

Exploratory Data Analysis:

Online Retail Transactions Dataset: The online retail transactions dataset is available from the UCI Machine Learning Repository. Based on its description on the UCI web site, it contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

We learn a valuable information about the company form the website itself, we learn that the business sells unique all-occasion gift items and a lot of customers of the organization are wholesalers.

The dataset is also available on Kaggle and can be download from the same too.

The dataset I have used in this project is also provided as a CSV format file.

Platform Used:

1. Programming Language – Python
2. Platform or IDE: Anaconda – Jupyter Notebook
3. Operating System: Kali Linux

Data Exploration with Jupyter Notebook:

I will be using Jupyter Notebook for my project. I have provided a complete notebook in this directory.

My first Step will be to import the dataset into Jupyter Notebook and Python Pandas library provides an easy and efficient way in achieving this task.

The Notebook provides the list of Dependencies.

Let’s view the dataset and gather information about its size:

Exploratory Data Analysis

- EDA is one of the first major analysis stages in the CRISP DM Process Model.
- The main objective is to explore and understand the data in detail.

Loading and Viewing the Dataset

In [2]:
```
# Reading the Customer Transaction dataset with file format csv via pandas
df = pd.read_csv("data.csv", encoding='latin1')
df.head()
```

Out[2]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Size of the Dataset =>

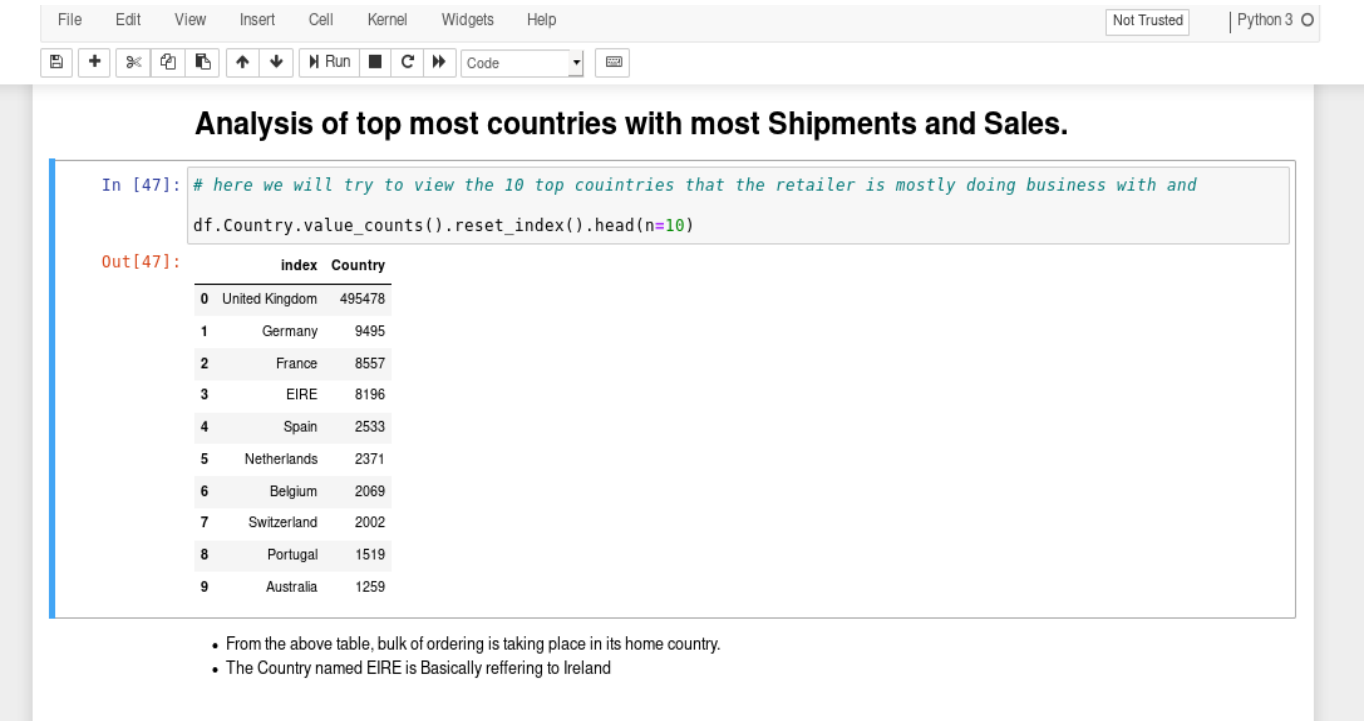
In [3]:
```
df.shape
```

Out[3]: (541909, 8)

Visualizing few lines of the dataset give information about the attributes of the dataset.

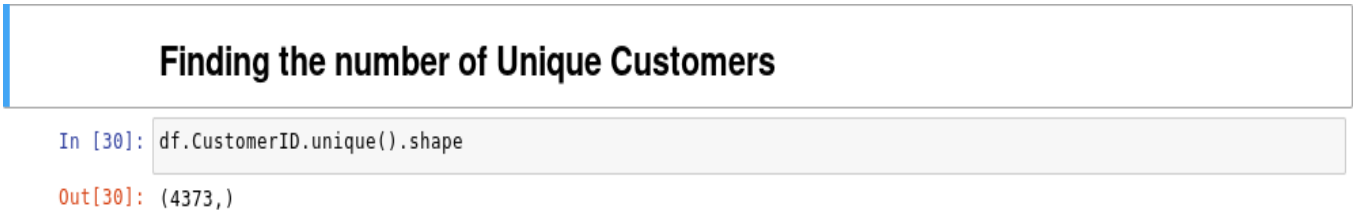
My next step will be to get a list of the countries that the retailer ships the most of its goods to.

For the sake of convenience, I will look for top 10 countries.



The data tells us that the most of the orders are from United Kingdom and are in bulk.

Let’s find the total count of unique customers:



Therefore, we now know that there are 4373 unique customers for the retailer.

I will now try to find out the number of customers that make up up to 10 % of the total revenue generated.

Percentage of orders made by top 10 customers:

- We must keep in mind that the company also has wholesalers as customers.
- Mathematically ==========>

```
|        DF of Unique no of customers      *  100 |
|  ------------------------------------           |
|    Revenue Generated by Uniqe customers         |
```

In [32]:
```
(df.CustomerID.value_counts()/sum(df.CustomerID.value_counts())*100).head(n=10).cumsum()
```

Out[32]:
```
17841.0    1.962249
14911.0    3.413228
14096.0    4.673708
12748.0    5.814728
14606.0    6.498553
15311.0    7.110850
14646.0    7.623350
13089.0    8.079807
13263.0    8.492020
14298.0    8.895138
Name: CustomerID, dtype: float64
```

Clearly top 10 customers make about 8 % of total sales, lets find the number of customers that generate 10% of the total sales.

In [46]:
```
(df.CustomerID.value_counts()/sum(df.CustomerID.value_counts())*100).head(n=13).cumsum()
```

Out[46]:
```
17841.0    1.962249
14911.0    3.413228
14096.0    4.673708
12748.0    5.814728
14606.0    6.498553
15311.0    7.110850
14646.0    7.623350
13089.0    8.079807
13263.0    8.492020
14298.0    8.895138
15039.0    9.265809
14156.0    9.614850
18118.0    9.930462
Name: CustomerID, dtype: float64
```

This shows that 10% of total sales are contributed by only 13 customers

As we notice, only 13 customers out of 4373 customers make up for 10 % of the Revenue. This goes in hand with the information that many of the customers are Wholesalers.

Now we have an idea of the Countries and the customers. I will now look into the data about the products sold by the retailer.

I want to get the count of the total number of unique products and if they have the matching Descriptions.

```
In [48]: df.StockCode.unique().shape
Out[48]: (4070,)
```

Total Number of Unique Product Descriptions

```
In [50]: df.Description.unique().shape
Out[50]: (4224,)
```

As we can see we have a mismatch between the two. The Descriptions exceed the total unique stock codes.

Data Quality Analysis:

Data quality analysis is the final stage in the data understanding phase where we analyze the quality of data in our datasets and document potential errors, shortcomings, and issues that need to be resolved before analyzing the data further or starting modeling efforts.

I will start by getting information about those StockCodes which have more than one Description and make list of it. I will create a new pandas dataframe and store the obtained list in it.

```
n_df = df.groupby(["StockCode","Description"]).count().reset_index()
```

StockCodes having more than one Description

```
n_df.StockCode.value_counts()[n_df.StockCode.value_counts()>1].reset_index().head()
```

	index	StockCode
0	20713	8
1	23084	7
2	85175	6
3	21830	6
4	21181	5

```
n_df.StockCode.value_counts()[n_df.StockCode.value_counts()>1].reset_index().shape
```

```
(650, 2)
```

This Implies that there are 650 Stock Codes with more than one Description.

Visualizing one of such Stock Codes

We get to see that we have 650 stockcodes which are having more than one Description.

Visualizing one such StockCode:

Visualizing one of such Stock Codes

```
In [72]: df[df['StockCode'] == n_df.StockCode.value_counts()[n_df.StockCode.value_counts()>1].
            reset_index()['index'][5]]['Description'].unique()
Out[72]: array(['SET/3 ROSE CANDLE IN JEWELLED BOX', 'wet pallet', 'damages',
                '???missing', 'AMAZON'], dtype=object)
```

After Visualizing the Stockcode we can clearly see that a Human error sometimes can lead to a possible loss of the data.

We also want to make sure our datatypes are all good to go before modeling.

```
In [52]: df.dtypes
Out[52]: InvoiceNo       object
         StockCode       object
         Description     object
         Quantity         int64
         InvoiceDate     object
         UnitPrice      float64
         CustomerID     float64
         Country         object
         dtype: object
```

We can see the dates in InvoiceDate Column are stored as object datatype, we will have to change it to valid date_time format before moving further.

```
df['invdatetime'] = pd.to_datetime(df.InvoiceDate)
```

```
df.head(3)
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	invdatetime
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom	2010-12-01 08:26:00
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	2010-12-01 08:26:00
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom	2010-12-01 08:26:00

```
df.dtypes
```

```
InvoiceNo              object
StockCode              object
Description            object
Quantity                int64
InvoiceDate            object
UnitPrice             float64
CustomerID            float64
Country                object
invdatetime    datetime64[ns]
dtype: object
```

Finally let's analyze our main two attributes, Quantity and UnitPrice.

```
df.Quantity.describe()
```

```
count    541909.000000
mean          9.552250
std         218.081158
min      -80995.000000
25%           1.000000
50%           3.000000
75%          10.000000
max       80995.000000
Name: Quantity, dtype: float64
```

```
df.UnitPrice.describe()
```

```
count    541909.000000
mean          4.611114
std          96.759853
min      -11062.060000
25%           1.250000
50%           2.080000
75%           4.130000
max       38970.000000
Name: UnitPrice, dtype: float64
```

- Both of these attributes are having negative values, which may mean that we may have some return transactions in our data also.
- We need to handle these before we proceed to our analysis

Chapter 5: Data Preparation

The third phase in the CRISP-DM process takes place after gaining enough knowledge on the business problem and relevant dataset. Data preparation is mainly a set of tasks that are performed to clean, wrangle, curate, and prepare the data before running any analytical or Machine Learning methods and building models.

Clustering and Clustering Strategy:

We are dealing with an unlabeled, unsupervised transactional dataset from which we want to find out customer segments. Thus, the most obvious method to perform customer segmentation is using unsupervised Machine Learning methods like clustering.

RMF Model:

The RFM model is a popular model in marketing and customer segmentation for determining a customer's value. The RFM model will take the transactions of a customer and calculate three important informational attributes about each customer:

• Recency: The value of how recently a customer purchased at the establishment

• Frequency: How frequent the customer's transactions are at the establishment

• Monetary value: The dollar (or pounds in our case) value of all the transactions that

the customer made at the establishment

A combination of these three values can be used to assign a value to the customer.

Data Cleaning:

As we have already seen, the bulk of our revenue comes from United Kingdom, my objective therefore for this project will be to work upon it only.

Data Cleaning

- Selecting the Data for A Particular Country - We will analyse for UK

```
cln_df = df[df.Country == 'United Kingdom']
```

We create a new dataframe which will have all the cleaned data in it.

Next thing we want to do is to remove the negative transactions or return transactions from our dataset.

```
cln_df['TotalAmount'] = df.Quantity*df.UnitPrice
```

Removing Negative Transactions or Return Transactions.

```
cln_df = cln_df[~(cln_df.TotalAmount<0)]
```

```
cln_df = cln_df[~(cln_df.TotalAmount<0)]
```

```
cln_df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	invdatetime	TotalAmount
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom	2010-12-01 08:26:00	15.30
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	2010-12-01 08:26:00	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom	2010-12-01 08:26:00	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	2010-12-01 08:26:00	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom	2010-12-01 08:26:00	20.34

Remove all the transactions that have a missing value for the CustomerID field as all our subsequent transactions will be based on the customer entities

```
cln_df = cln_df[~(cln_df.CustomerID.isnull())]
```

```
cln_df.shape
```

```
(354345, 10)
```

Hence, we have a new dataframe which is cleaned of the negative transactions.

```
cln_df.describe()
```

	Quantity	UnitPrice	CustomerID	TotalAmount
count	354345.000000	354345.000000	354345.000000	354345.000000
mean	12.048913	2.963793	15552.436219	20.625073
std	190.428127	17.862067	1594.546025	326.033014
min	1.000000	0.000000	12346.000000	0.000000
25%	2.000000	1.250000	14194.000000	4.160000
50%	4.000000	1.950000	15522.000000	10.200000
75%	12.000000	3.750000	16931.000000	17.700000
max	80995.000000	8142.750000	18287.000000	168469.600000

We will now build RMF variables to build customer value:

Note that I have taken the reference date to be the one day after the last transaction date.

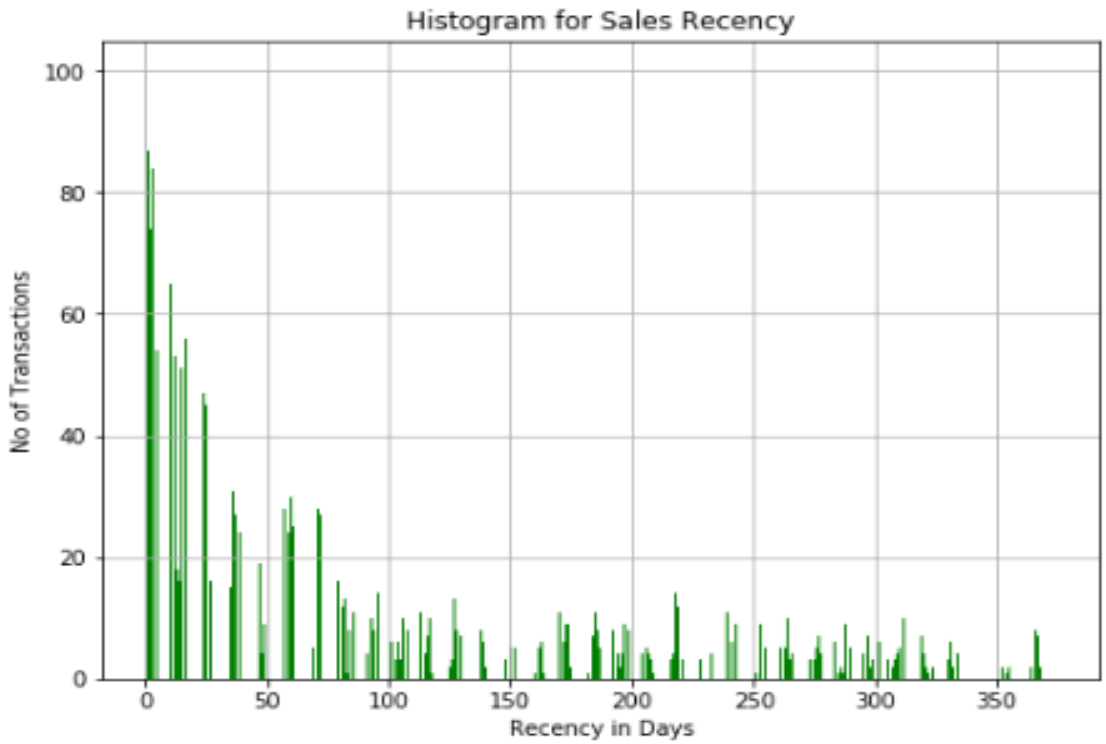
Calculating Recency

```
cust_his_df = cln_df.groupby("CustomerID").min().reset_index()[['CustomerID', 'days_since_last_purchase_num']]
cust_his_df.head()
```

	CustomerID	days_since_last_purchase_num
0	12346.0	326.0
1	12747.0	2.0
2	12748.0	1.0
3	12749.0	4.0
4	12820.0	3.0

```
cust_his_df.rename(columns={'days_since_last_purchase_num':'recency'}, inplace=True)
cust_his_df.head()
```

	CustomerID	recency
0	12346.0	326.0
1	12747.0	2.0
2	12748.0	1.0
3	12749.0	4.0
4	12820.0	3.0



Same way i will create the Monetary and Frequency values. For detailed code go to the Notebook Section.

```
cust_his_df = cust_his_df.merge(cust_freq, how='outer')
cust_his_df.head(3)
```

	CustomerID	recency	TotalAmount	frequency
0	12346.0	326.0	77183.601	1
1	12747.0	2.0	4196.011	103
2	12748.0	1.0	33719.731	4596

Now we have all the three RMF values:

We will now move to the last step in Data preparation that is Data Pre-Processing.

Data Pre-Processing:

Now we have our customer value dataframe, we will perform preprocessing on the data. For our clustering, we will be using the K-means clustering algorithm. One of the requirements for proper functioning of the algorithm is the mean centering of the variable values.

Mean centering of a variable value means that we will replace the actual value of the variable with a standardized value, so that the variable has a mean of 1 and variance of 0. This ensures that all the variables are in the same range and the difference in ranges of values doesn't cause the algorithm to not perform well.

Data Pre-Processing

```
from sklearn import preprocessing
import math
```

Changing Variables to log scale

```
cust_his_df['recency_log'] = cust_his_df['recency'].apply(math.log)
cust_his_df['frequency_log'] = cust_his_df['frequency'].apply(math.log)
cust_his_df['amount_log'] = cust_his_df['TotalAmount'].apply(math.log)
```

```
feature_vector = ['amount_log', 'recency_log','frequency_log']
```

```
X_subset = cust_his_df[feature_vector].as_matrix() # Note: If An error has occured please update your Package
```

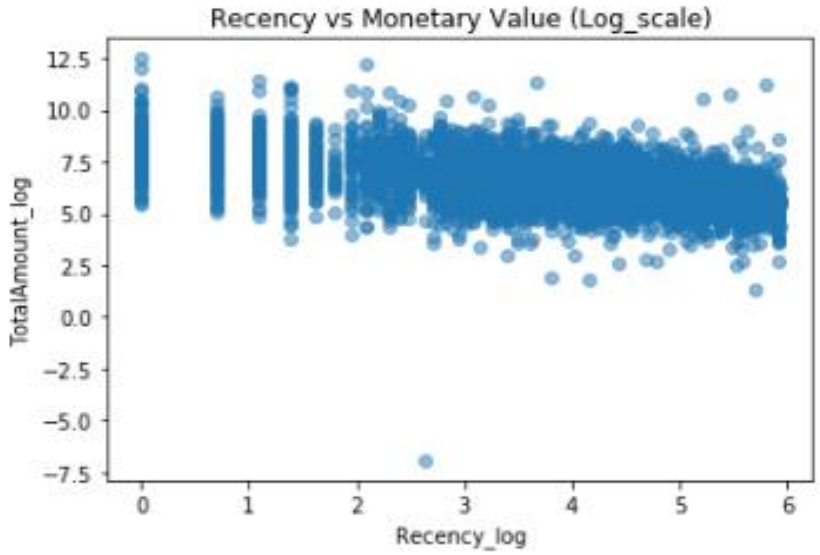
```
C:\Users\User\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: FutureWarning: Method .as_matrix will be
removed in a future version. Use .values instead.
  """Entry point for launching an IPython kernel.
```

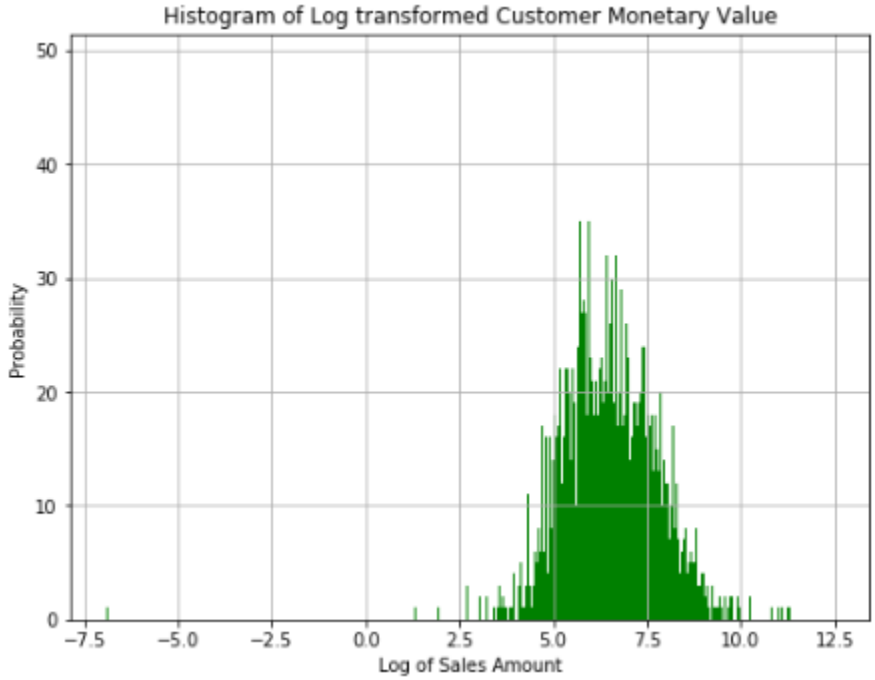
```
scaler = preprocessing.StandardScaler().fit(X_subset)
X_scaled = scaler.transform(X_subset)
```

Data Visualization:

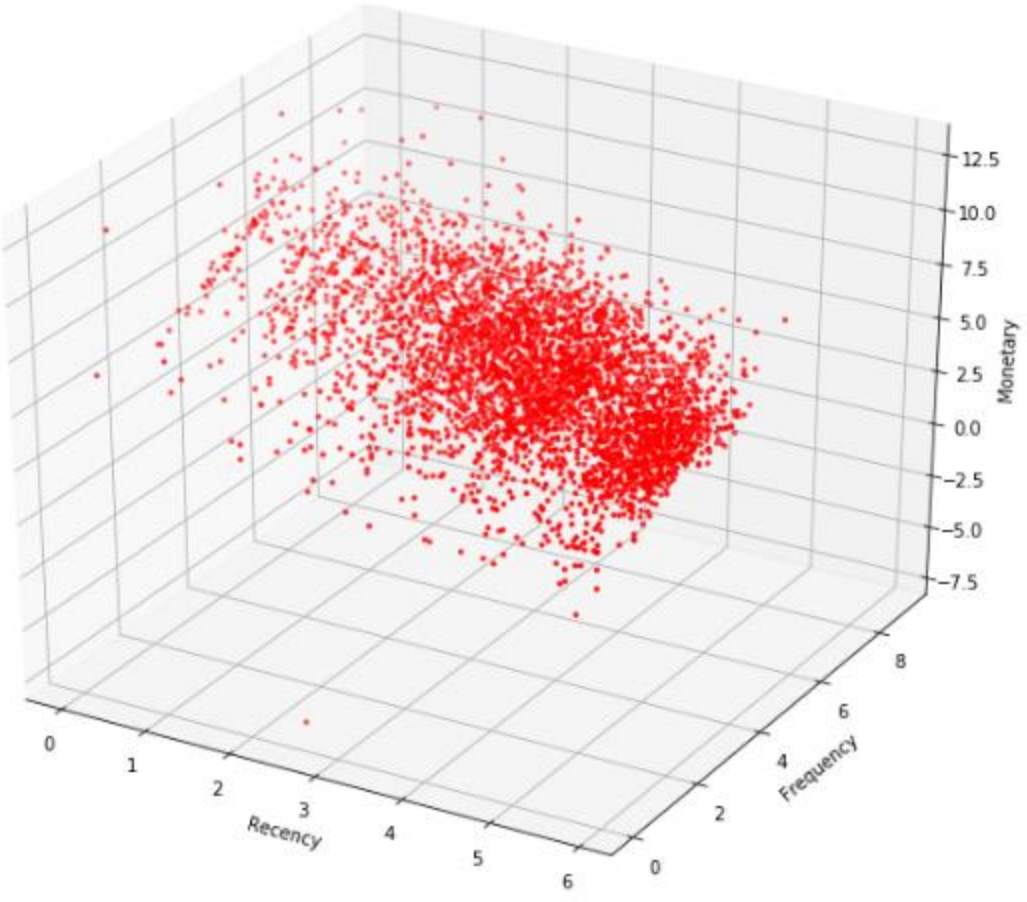
Visualization is an important step in every data science project.

I will be using both 2D and 3D plots for visualizing our data.





Visualize our three main Features on a 3D – plot



The patterns we can see from our 3D plot shows us that people who buy with a higher frequency and more recency tend to spend more based on the increasing trend in Monetary value with a corresponding increasing and decreasing trend for Frequency and Recency, respectively.

This is an example of how we can generate the insights from visualization tools using data science. all the information and insights are valuable for the organizations or a business.

Chapter 6: Modeling

The fourth phase in the CRISP-DM process is the core phase in the process where most of the analysis takes place with regard to using clean, formatted data and its attributes to build models to solve business problems.

Clustering for Segments:

We will be using the K-means clustering algorithm for finding out clusters (or segments in our data). It is one of the simplest clustering algorithms that we can employ and hence it is widely used in practice.

K-means clustering:

k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The problem with every clustering algorithm we face is to choose the value for K which means the number of clusters we want to create. We can do that manually but it takes a lot of time, the way I am going to tackle this problem is by using Sk-learn's Silhouette Analysis.

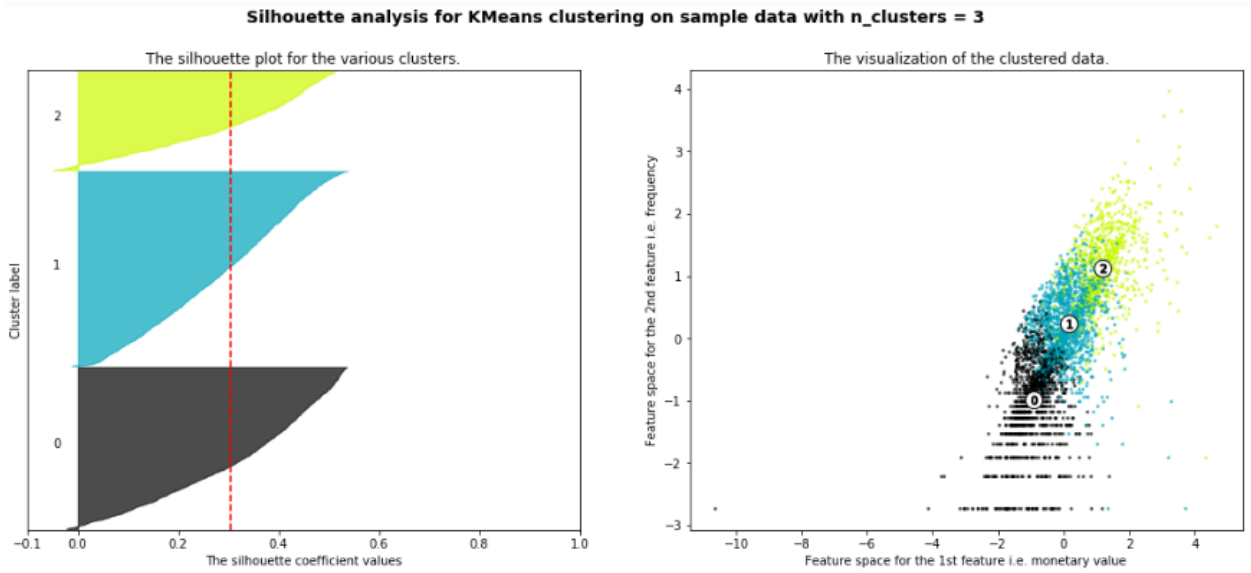
Silhouette Analysis on K-means Clustering:

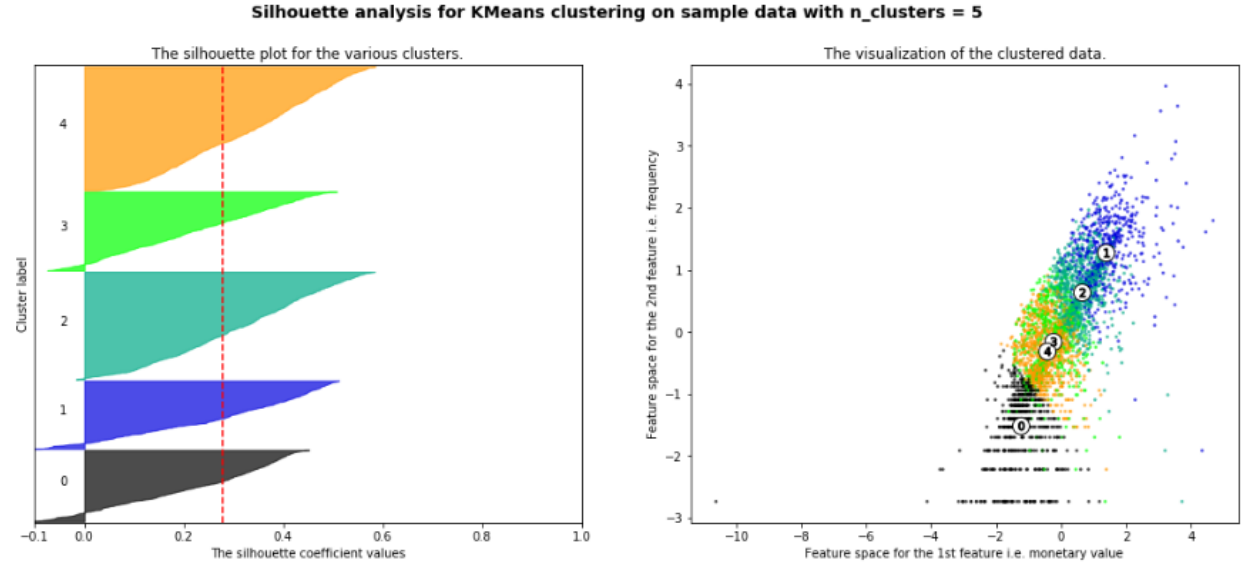
Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1].

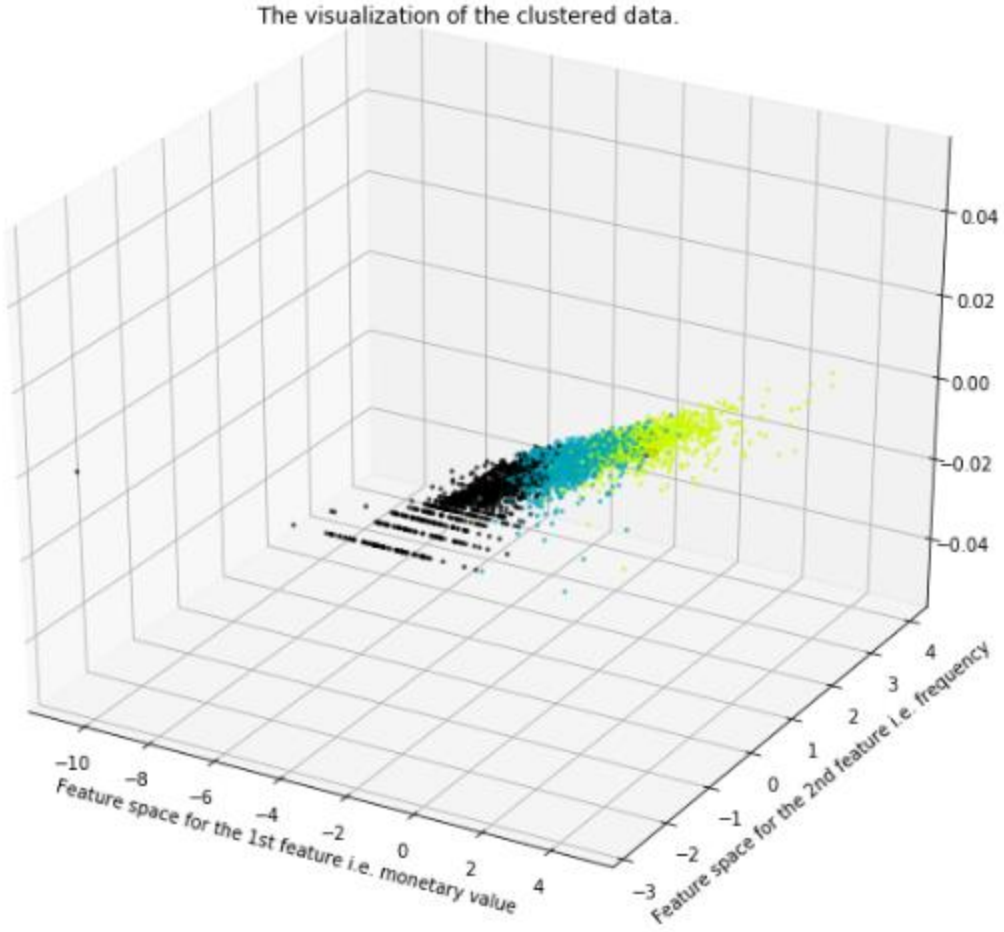
I will show Silhouette Analysis with 2D as well as 3D plot.

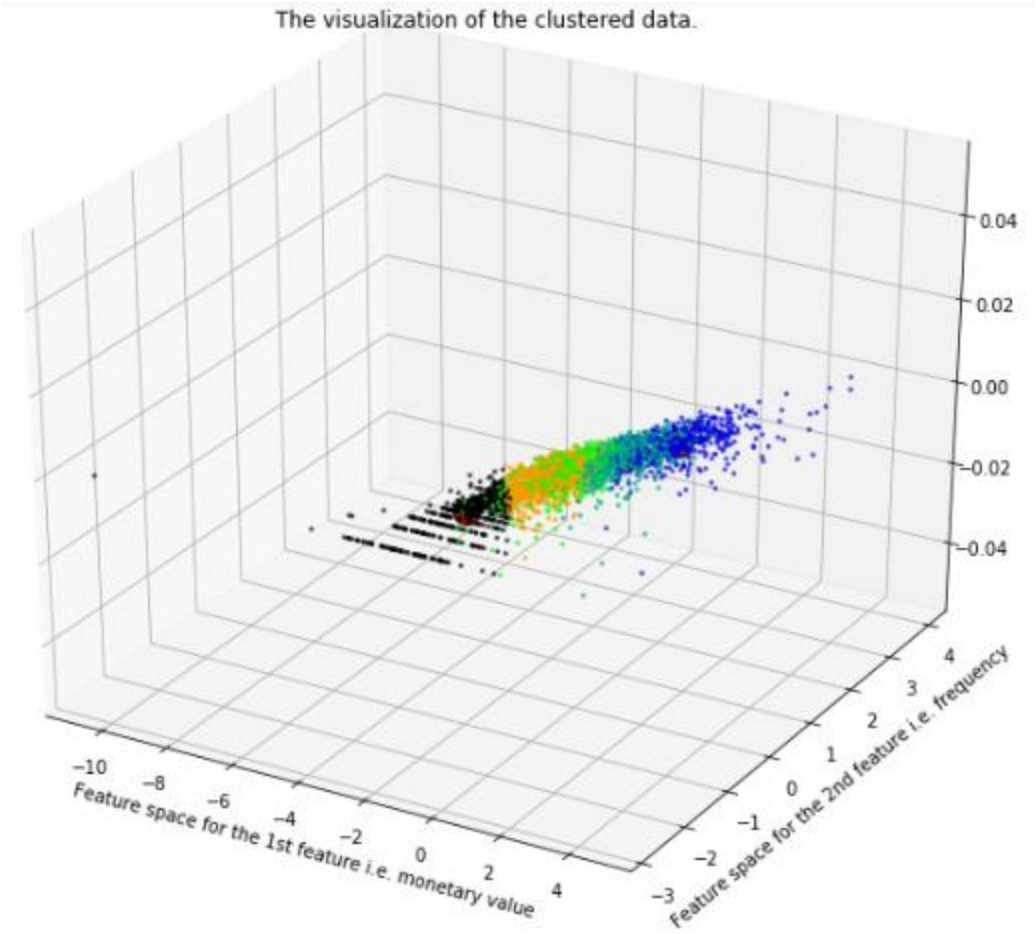
NOTE: The centers on 3D plot are not visible because the centers are lying inside the data clusters.

I will iterate by starting with k = 3 and then k = 5






The visualization of the clustered data.
0.04
0.02
0.00
−0.02
−0.04
−10
−8
−6
−4
−2
0
2
4
Feature space for the 1st feature i.e. monetary value
−3
−2
−1
0
1
2
3
4
Feature space for the 2nd feature i.e. frequency



Pre-Segmentation Analysis:

As we had transformed all the values to a log scale, now we want to transform the values back to the normal scale. We also want to get the values of the centers thus obtained by our K-Means clustering algorithm.

```
for i in range(3,6,2):
    print("for {} number of clusters".format(i))
    cent_transformed = scaler.inverse_transform(cluster_centers[i]['cluster_center'])
    print(pd.DataFrame(np.exp(cent_transformed),columns=feature_vector))
    print("Silhouette score for cluster {} is {}". format(i, cluster_centers[i]['silhouette_score']))
    print()
```

```
for 3 number of clusters
    amount_log  recency_log  frequency_log
0   220.198311   122.258999      10.545038
1   828.741744    44.343353      53.275214
2  3158.748880     7.396805     177.087308
Silhouette score for cluster 3 is 0.303449108662755

for 5 number of clusters
    amount_log  recency_log  frequency_log
0   144.581761   121.494119       5.203414
1  3983.000207     5.669752     216.919083
2  1529.103414    48.475555      93.359920
3   491.225157    13.227883      31.937363
4   407.421058   138.857191      25.576757
Silhouette score for cluster 5 is 0.2782580724931221
```

We simply perform mathematical inverse transformation on the center values and change the numbers from the log scale to normal values.

The numbers provided are valuable and many insights can be drawn from it.

Valuable Insights:

Considering Three Clusters:

- We can see a clear distinction in the Monetary value of the customer.
- Cluster 2 is the cluster of high value customer who shops frequently and is certainly an important segment for each business.
- We also get information about the customers with low spend and medium spends.

Considering Five Clusters:

The results are surprising in the five-cluster analysis.

- Our High value customer is further made of two subgroups.
- Customers who shop frequently and with high amounts.
- Customers who spend high amounts but do not shop frequently.

Note: The silhouette analysis showed that the five-cluster analysis wat less optimal, this is because there is a higher overlap in the five-cluster analysis as compared to the 3-cluster analysis.

Also, we need to sometimes overlook mathematical insights and look at the business aspects of the data.

Cluster Description:

Once we have the labels assigned to each of the customers, our task is simple. We can find out how the summary of customer in each group is varying.

If we can visualize that information, we will able to find out the differences in the clusters of customers and we can modify our strategy on the basis of those differences.

	CustomerID	recency	TotalAmount	frequency	recency_log	frequency_log	amount_log	num_cluster5_labels	num_cluster3_labels
0	12346.0	326.0	77183.601	1	5.786897	0.000000	11.253942	1	0
1	12747.0	2.0	4196.011	103	0.693147	4.634729	8.341890	0	2
2	12748.0	1.0	33719.731	4596	0.000000	8.432942	10.425838	0	2
3	12749.0	4.0	4090.881	199	1.386294	5.293305	8.316516	0	2
4	12820.0	3.0	942.341	59	1.098612	4.077537	6.848367	4	2

Now, all customer in each cluster have been labeled and now we are ready for visualization.

Instead of using normal visualization tools like matplotlib and seaborn as I have used them earlier, I will be using another powerful visualization tool Plotly.

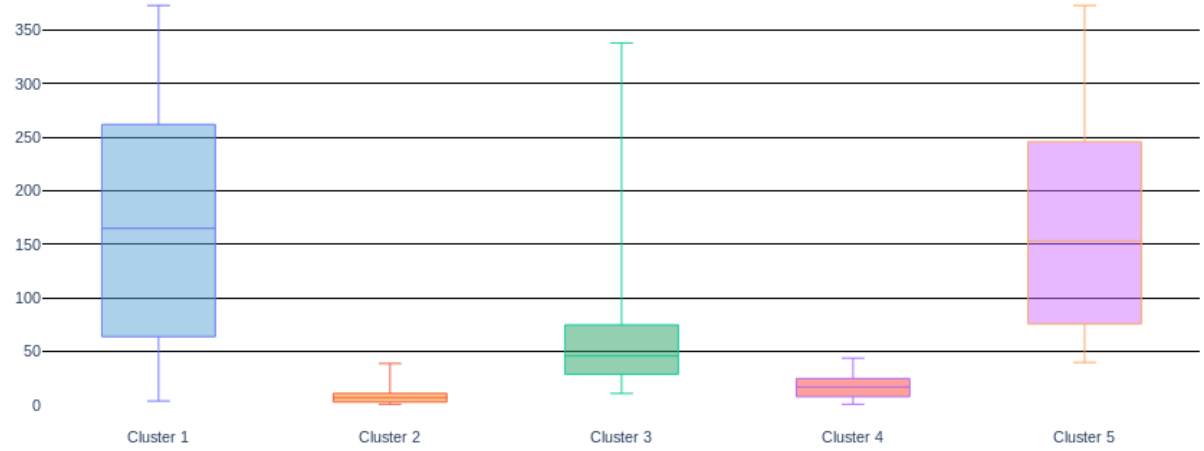
Note: We will restrict the data such that only data points which are less than 0.8th percentile of the cluster is used. This will give us good information about the majority of the users in that cluster segment.

We want to do so in order to avoid the extreme values.

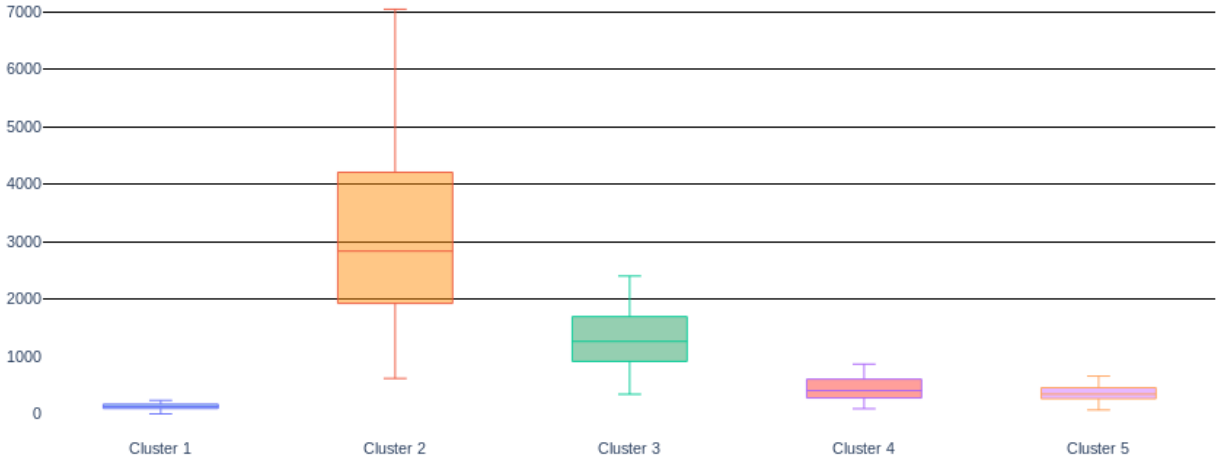
Visualizing Segments:

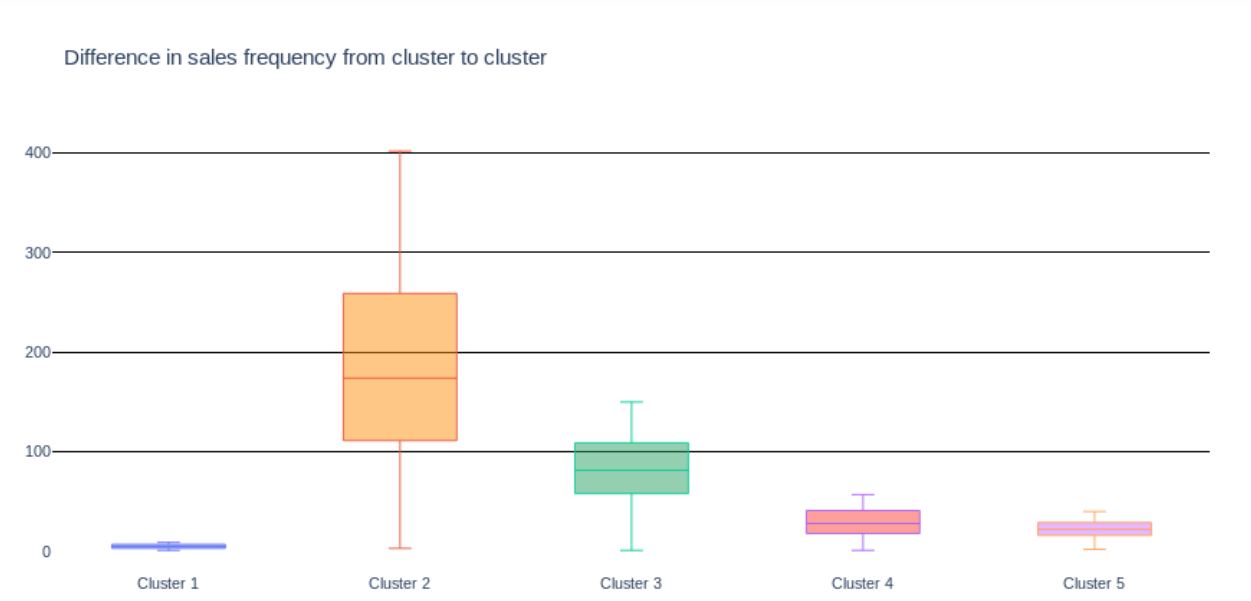
Five Cluster Visualization:

Difference in sales recency from cluster to cluster



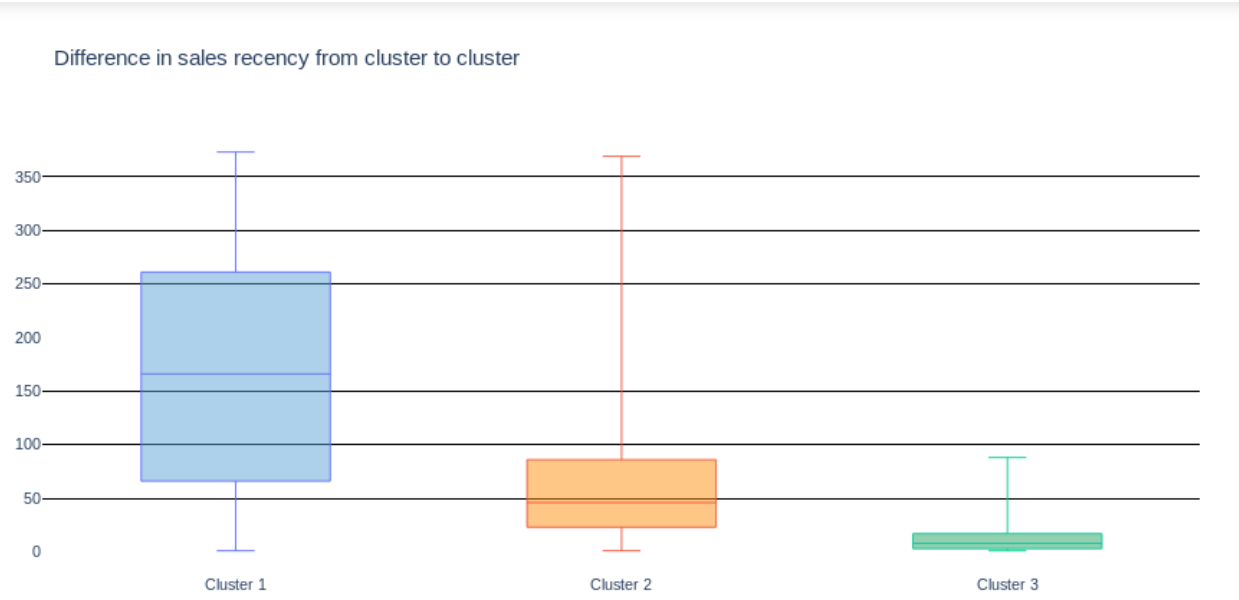
Difference in sales TotalAmount from cluster to cluster



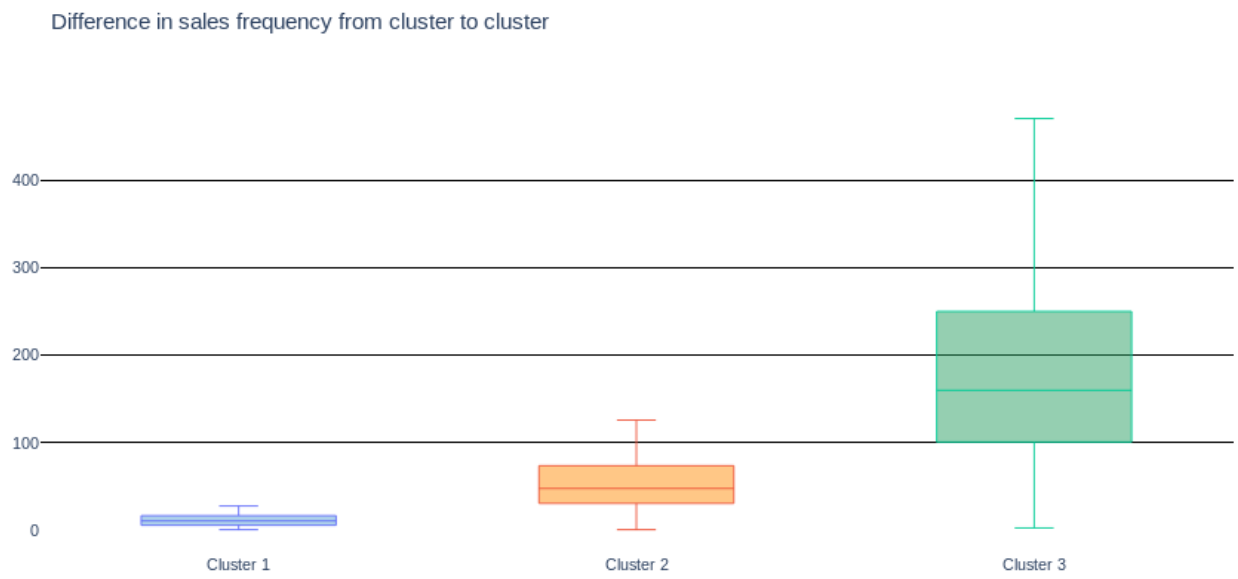


The above give us the complete visualization of the RMF clusters for five segments.

Coming to three cluster analysis:


Difference in sales recency from cluster to cluster
350
300
250
200
150
100
50
0
Cluster 1
Cluster 2
Cluster 3


Difference in sales TotalAmount from cluster to cluster
5000
4000
3000
2000
1000
0
Cluster 1
Cluster 2
Cluster 3



Finally, I will analyze the segments based upon the Customer Segmentation process we have come up with using K-Means Clustering Algorithm.

Segmentation Analysis:

Five Segment Analysis:

Valuable Insights:

1. The customers in segment 2 are of high value, their Recency is very low, which means that it hasn't been long since they have visited or shopped at the retailer.

   The customers in segment 2 spend more money than any other customer segment.

   Which can also mean that the customers in segment two will be more likely to buy products in large quantities and also products which cost more.

   Another insight which can be generated is that the customers of segment 2 have a very high frequency of shopping.

   Thus, we can say the customers of the second cluster must be the top priority for the business, and to ensure that they keep on a healthy relationship with the business, a target marketed campaign must be carried out for them.

   A detailed analysis can be further carried upon the customers in cluster segment 2, this can be done via using the Market Basket Analysis.

2. The cluster segment 3 is also very interesting for the business, as this cluster spends medium amount, their frequency and recency are relatively medium ranged.

   Our main aim as a business would be to increase the revenue generated by this cluster segment.

   This can be achieved by lowering the recency bar and increasing the frequency bar.

   Our strategy for this cluster segment is not to get them buying costlier products, but to only increase their frequency of shopping and lowering the recency.

   By doing so the bar for total amount spend will rise automatically and maybe would even be relatively equal with the amount bar of the cluster 2 segment.

   Our marketing strategy would be to give out discounts and other important customer attractive offers to ensure we reach our target.

3. The other clusters form a low value customer segment. A general marketing campaign is a useful tool through which we can somewhat make those low value customers into our medium and even high value customer.

Three Segment Analysis:

1. My own understanding tells me that the three-cluster analysis was a general outlook of the segments. It was the five-cluster analysis which helped us look even deeper into the data.
2. We were able to analyze that the third cluster of our three-segment analysis was actually composed of two further clusters.
3. The three-cluster analysis goes in hand with our targeted strategy for the High value customers we proposed earlier.

Chapter 6: Conclusion

The one of the most important take-away from this project is that we must not always restrict our analysis to the mathematical aspects and drawings of the data. We learn that when it comes to business analysis, we must look beyond mathematical insights of the data and look from the business aspect as well.

Once we have developed such informative and useful insights, we can come in touch with the marketing team for developing necessary and appropriate practices for the particular cluster-segments.

This will be useful for the revenue model of the business as this will directly help improve the business for the organization.

In this project we used a dataset for our analysis, but further more now have the tools to develop such insights. We can use our knowledge of data science to explore many businesses around us and help improve their revenues.

Future Aspect:

I had mentioned in the beginning of this project, that the reason behind selecting this particular dataset was that I am focused to improve the retail sector of our country.

The data generated by businesses around us is enormous. But that data is totally unstructured. The businesses around us have no idea how useful this data can be

for them. This is because they lack appropriate tools and do not have the proper knowledge of the uses of data.

The data is like oil in this century, it is us who are equipped with the tools to perform useful data analysis and predictive analysis. This is a huge opportunity for us.

I have already started gathering data from interested businesses on my own.

This way I will be able to gather valuable and necessary skills to be a better data scientist.

What's Next?

The machine learning algorithms are mathematical models and I showcased their use in this project.

My next goal and project will be a deep learning project.

I have already decided which project for deep learning I am going to work upon and it related to the Computer Vision. This time I am getting more creative and will try to bring out the beauty of my Kashmir Valley using computer vision and deep learning.

It is of course going to take a bit but I am sure the end product will be a creative one.

This project did not make it necessary for me to create a deploying mechanism but I will be creating a deployment mechanism if required by my next project.

Bibliography:

https://www.kaggle.com/carrie1/ecommerce-data

http://archive.ics.uci.edu/ml/datasets/Online+Retail/

https://en.wikipedia.org/wiki/K-means_clustering

https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

https://www.shopify.in/encyclopedia/customer-segmentation

http://www.datascience-pm.com/crisp-dm-2/

https://www.sololearn.com/Play/data-science

https://www.youtube.com/channel/UCh9nVJoWXmFb7sLApWGcLPQ

https://www.youtube.com/watch?v=NPznsxeL3FM&list=PLH6mU1kedUy9HTC1n9QYtVHmJRHQ97DBa

https://www.youtube.com/user/Simplilearn

https://www.youtube.com/watch?v=OGxgnH8y2NM

https://scikit-learn.org/stable/modules/clustering.html#clustering

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html#sklearn.metrics.silhouette_score

https://www.kaggle.com/hendraherviawan/customer-segmentation-using-rfm-analysis-r

In addition of the resources provided above I have gained valuable Data Science experience in SachTech Solutions.

I have followed my mentors online and offline and I am daily working upon increasing my knowledge of Industry standard Data science tools.

Thank You