**ABSTRACT**

“Customer is the king.” In today’s world, customer is indeed the king and hence it’s not only very important to retain our loyal customers but also the ones who are on the verge of churning out. Every day there is a transaction process performed by Customer. The process generates a lot of data where there are 82,648 transactions from the month of January-December 2017. Many small online retailers and new entrants to the online retail sector are keen to practice data mining and consumer-centric marketing in their businesses yet technically lack the necessary knowledge and expertise to do so. Companies need to understand the customers’ data better in all aspects. Detecting similarities and differences among customers, predicting their behaviors, proposing better options and opportunities to customers became very important for customer-company engagement. Segmenting the customers according to their data became vital in this context. The objective of this study is to apply business intelligence in identifying potential customers by providing relevant and timely data to business entities in the Retail Industry. The data furnished is based on systematic study and scientific applications in analyzing sales history and purchasing behavior of the consumers. The curated and organized data as an outcome of this scientific study not only enhances business sales and profit, but also equips with intelligent insights in predicting consumer purchasing behavior and related patterns. In order to execute and apply the scientific approach using **K-Means algorithm**, the real time transactional and retail dataset are analyzed. Spread over a specific duration of business transactions, the dataset values and parameters provide an organized understanding of the customer buying patterns and behavior across various regions. This study is based on the **RFM** (Recency, Frequency and Monetary) model and deploys dataset segmentation principles using K-Means Algorithm**.** A variety of dataset clusters are validated based on the calculation of Silhouette Coefficient. The results thus obtained with regard to sales transactions are compared with various parameters like Sales Recency, Sales Frequency and Sales Volume. RFM (Recency, Frequency and Monetary) values have been used for many years to identify which customers valuable for the company, which customers need promotional activities, etc. **Data-mining** tools and techniques widely have been used by organizations and individuals to analysis their stored data. Clustering, which one of the tasks of data mining has been used to group people, objects, etc.  In this article a case study of using data mining techniques in customer-centric business intelligence for an online retailer is presented. The main purpose of this analysis is to help the business better understand its customers and therefore conduct customer-centric marketing more effectively. On the basis of the Recency, Frequency, and Monetary model, customers of the business have been segmented into various meaningful groups using the *k*-means clustering algorithm and decision tree induction, and the main characteristics of the consumers in each segment have been clearly identified. We detected that the current customer segmentation which built by just considering customers’ expense is not sufficient. Hence, models that recommended in this research are expected to provide better customer understanding, well-designed strategies, and more efficient decisions.

Much of the advertising seen on both desktop and mobile Web sites is not the result of savvy ad buyers picking up the phone and calling publishers, but instead is procured through programmatic advertising, a process by which ads are automatically served up based on a bidding system driven by algorithms. Marketers can create algorithms that automatically purchase ads across the Web based on these criteria, instead of having to pay for a specific number of ads, as was the practice with traditional ads.  Furthermore, while a campaign is ongoing, these algorithms can evaluate what is working best, in terms of geographic segmentation, daytimes, audience segments, and publishers, to help marketers narrow their target so they are paying only for highly effective ads.

**KEYWORDS**

* Customer segmentation
* RFM model
* Clustering
* K- means clustering
* Recency
* Frequency
* Monetary
* Elbow Criterion
* Business intelligence
* Customer Churn.

Chapter

|  |  |
| --- | --- |
| 1 | INTRODUCTION |

In today's business competition, customers are the main focus of the company to maintain its excellence. Companies must plan and use clear strategies in serving customers. The company's primary focus is not on how to get new potential customers but how to sell more products to the existing customers because the cost that companies must incur to acquire new customers is much more expensive than to retain existing customers. In the credit business, the data can be obtained based on historical data, so the data will increase continuously such as the transaction data from each agent. The transaction process of agents in a credit server generates abundant data in the form of profiles of transactions that the agent performs. This will happen repeatedly to the credit business. Agent transaction data cumulation will slow down the search for information on that data. This data can be called as **data mining**. Data mining is a part of knowledge discovery data which is an information extraction process that is useful, not known before, and hidden from data. Based on the number of available agent transaction data, the unknown or hidden information can be known by processing the data so that it is useful for the credit business agent, for example in which information on the grouping of agent data has the potential to give the most profit to the company which will help companies to make decisions in product marketing. The model used by the researcher is **RFM (Recency, Frequency, Monetary)** commonly used to perform the last visit time grouping, visit frequency, and revenue obtained by the company. The reason why continuing to use the RFM model is that it is easy to use and quickly implemented in companies, and in addition RFM is easily understood by managers and marketing decision makers. In the light of data **segmentation**, customers are divided into set of individuals with distinct similarities. Some of the attributes relevant to customer segmentation are gender, age, lifestyle, location, purchase and income behavior. Such attributes are mainly categorized based on the historical purchasing behavior that can lead to a specific outcome, for example, an increase in sales and the profit for the company. As it is well known by marketers, customers have various kinds of needs and wants. Companies have used several segmentation criteria and techniques to better identify and understand customer groups and provide preferable products and services to them in order to satisfy these different needs and wants. Also, segmentation is important that the company can create profitable segments and react to the selected segment based on its competitive advantages. However, many marketers have difficulty in identifying the right customer segments to organize marketing campaigns (Mohammadian & Makrani, 2016). This causes unsuccessful loyalty programs and promotions conjunction with waste of marketing resources. **Customer churn** is the tendency of customers to stop purchasing with a company over a time period. Customer churn is also called **customer attrition** or **customer defection**. Churning reduces growth. Therefore, companies should have a proper defined method to compute customer churn rate for a given time. By keeping track of churn rate, organizations are often equipped to succeed in terms of customer retention. Retailers need a good strategy to manage customer churn. Measuring the churn rate is kind of crucial for retail businesses because the metric reflects customer response towards the service, quality, price and competition. Churn prediction envision the likelihood of customers to churn. It pares the investment on gaining new customers and helps to retain the existing customer. The marketing efforts and amount spent on attracting a new customer is higher and more difficult than clinging to existing customers. Customers who are unlikely to make a purchase or willing to shift the shopping site because of cautiousness with money, expecting standard and assortment in products can be convinced and clutched. The customers who are ending the relationship due to valuable and unavoidable reasons are free to leave. Compared with traditional shopping in retail stores, online shopping has some unique characteristics: each customer's shopping process and activities can be tracked instantaneously and accurately, each customer's order is usually associated with a delivery address and a billing address, and each customer has an online store account with essential contact and payment information. These desirable, special online shopping characteristics have enabled online retailers to treat each customer as an individual with personalized understanding of each customer and to build upon customer-centric business intelligence.

In relation to customer-centric business intelligence, online retailers are usually concerned with the following common business concerns:

* Which items/products’ web pages has a customer visited? How long has a customer stayed with each web page, and in which sequence has a customer visited a set of products’ web pages?
* Who are the most/least valuable customers to the business? What are the distinct characteristics of them?
* Who are the most/least loyal customers, and how are they characterized?
* What are customers’ purchase behavior patterns? Which products/items have customers purchased together often? In which sequence the products have been purchased?
* Which types of customers are more likely to respond to a certain promotion mailing? and
* What are the sales patterns in terms of various perspectives such as products/items, regions and time (weekly, monthly, quarterly, yearly and seasonally), and so on?

In order to address these business concerns, data mining techniques have been widely adopted across the online retail sector, coupled with a set of well-known business metrics about customers’ profitability and values, for instance, the recency, frequency and monetary **(RFM) model**,  and the **customer life value model**. For many online retailers in the United Kingdom and internationally alike, especially the leading companies including Amazon, Walmart, Tesco, Sainsbury's, Argos, Marks and Spencer, John Lewis, and EasyJet, data mining has now become a common practice and an integral part of the business processes in creating customer-centric business intelligence and supporting customer-centric marketing.

"In order to build strong targeting, you need to build a really strong attribution model because otherwise, you’re assigning incorrectly credit to stuff that really didn’t make a difference," says Jay Friedman, COO of Goodway Group, a nationwide firm that has worked on digital campaigns with Fortune 500 companies such as Subaru, McDonalds, and General Motors. For example, if a marketer had a set of data consisting of females who purchased makeup and also frequented salons to get their hair done, perhaps a close, automated scan of these users would uncover that this segment over-represents in the purchase of, say, Snickers candy bars. By developing an algorithm that can quickly identify new data patterns, and then targeting new ads based on this data, marketers can make the most of any data they encounter.

Algorithms are being used to do more than just serve up relevant ads online.

Ad frequency is another issue that should be addressed by marketers, since there is technology available to manage the process.

"At what point is showing another ad to a user not worth it anymore?" Friedman asks. "Our algorithm just cuts off users once they’ve hit a certain amount, because we’ve deemed it not worth it anymore. But I think that it annoys consumers; ‘you’ve shown me the ad 10 times, and I’m not going to buy.’  It’s beneficial for both marketers and consumers to mind the frequency."

In this project we have created an approach where advertisements will be delivered through e-mails and text messages to customers according to the k-means value while segmenting those customers behaviorally. The segment with higher k-mean score will receive advertisements on the basis of their previous purchasing time, say during Diwali or Eid or Christmas, they will get special offers, whereas segments with low k-mean scores will get discounts or special offers no matter what occasion, because as they are on the verge of churning out, marketer will have to focus on the strategy on how they can get back those customers as well.

Chapter

|  |  |
| --- | --- |
| 2 | LITERATURE REVIEW |

## 2.1. Customer Segmentation

Customers can have different types of characteristics and can be of different importance to a company. For companies to know which customers are of significance, a segmentation of customers need to be done (**McDonald & Dunbar, 2012**)[7]. The theory of segmentation is the process where identifying characteristics of different customers and dividing them into groups. What companies often do when segmenting their customers is to divide them based on how much revenue they contribute to the company based on their purchase volumes (**Batt, 2000**)[1].

According to **Bottcher[2], Spott, Nauck& Kruse (2009)** identifying and classifying customers leads to a better understanding of who the customers are and what type of demand the customers require. Some customer groups can have a high degree of innovation where changes within the customer group over time often occur. For these type of customers, you need to be aware of the requirement changes to meet the customer demand in the best way which also fulfils the customer needs.

According to **Sandström (2003)[10]** segmentation has its basis in the concept that consumers who take part of the company's products and services are not proportionally valuable. For companies, customers are of different significance and to be able to stay in the market, companies need to distribute their attention unevenly, meaning that they need to move attention from the non-profit consumers to the ones with higher profit. For a company to continue gaining profit, they need to target a lot of attention to the customers that consume their products or services frequently or in greater volumes to create groups that are fruitful.

**Buttle (2009)[3]** states that when a company has the right knowledge about the customer requirements it will give them the ability to easier divide the customers into segmentation groups. Furthermore, the company can easier find out what satisfies their customers and even surprise them. This kind of information can be used for further improvements into their services or products. These days, customer service is as important for the customers as the actual product or service, and it is important that the companies have this part set. Finally, segmenting customers can simplify the choices of how much and what the company should put emphasis on when it comes to the degree of services that the different groups should get.

**Lambert (1990)[5]** state the importance of segmenting markets in emerging production industries as well as in service industries. All kinds of organisations need to find a method that fits to categorize the market into different segments to meet the customer demand in best way and increase the revenue for the company. According to Lambert (1990) when an organisation starts the process of segmenting it is recommended to look at the customers from a need based point of view where the customer with high demand should be prioritised.

A company provides the market with either a service or a product. Because of this it is vital according to **Fang, Palmatier[9] & Steenkamp (2008)** for a company to reach the customer service elements to please their customers. Well established service companies have the right skillset and right knowledge to fulfil the demands, expectations and needs of their customers (**Mattson, 2004**)[10]. The concept of customer service can be defined as what a company does to include the purchasers, sellers and other groups that can boost their product or service. A successful customer segmentation within services benefits the company to enhance their relationship with their purchasers and sellers which also contributes to an enhanced competitiveness (**Pauline, 2009**)[10].

**Anderson, Naru & Narayandas (2009)[10]** Out in the market there are a huge range of service providing companies with many different types of customers. To be competitive and meet the customer demand in best way, service companies have different strategies of how to target and segment their customers.

## 2.2. Clustering and K-Means Algorithm

**Tajunisha et al.[12]**

Performance Analysis of k-Means with different initialization methods for high dimensional data uses Principal Component Analysis (PCA) for dimension reduction and to find initial cluster centers. The variable with the highest Eigen value calculated using PCA is taken as first principal component along which partitioning is done, on the basis of which k subsets are formed and k median values are taken as initial k centers.

**Bouhmala et al.[8]**

Combined Genetic Algorithm and K-Means to improve the quality of clusters formed and speed up their search process. The performance of GAKM is tested over the datasets such as iris, glass, etc., and that has been taken from Machine learning repository. The experimental results have proved that GAKM converges faster while comparing to standard Genetic Algorithm. Though this algorithm failed to capture the best quality of clusters, it is unsuitable for the maximizing both homogeneity and heterogeneity within same clusters and with different clusters respectively.

Bara’a Ali Attea et al. [12]. Discovered that performance of

**Chetna Sethi et al.[4]**

Proposed a Linear PCA based hybrid K-Means clustering and PSO algorithm (PCA-K-PSO). In (PCA-K-PSO) algorithm the fast convergence of K-Means algorithm and the global searching ability of Particle Swarm Optimization (PSO) are combined for clustering large data sets using Linear PCA. Better clustering results can be obtained with PCA-K-PSO as compared to ordinary PSO. This was effectively developed in order to make its use for efficient clustering of high- dimensional data sets.

**Tapas Kanungo et al.[13]**

An efficient k-means clustering algorithm: analysis and implementation .In k-means clustering, they are given a set of n data points in d-dimensional space Rd and an integer k and the problem is to determine a set of k points in Rd, called centers, so as to minimize the mean squared distance from each data point to its nearest center. A popular heuristic for k-means clustering is Lloyd's algorithm. In this paper, present a simple and efficient implementation of Lloyd's k-means clustering algorithm, which they call the filtering algorithm. This algorithm is easy to implement, requiring a kd-tree as the only major data structure. A establish the practical efficiency of the filtering algorithm in two ways. First, present a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, present a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization, data compression, and image segmentation.

**2.3. RFM model**

**Wei et al.[14]**

This paper provides a comprehensive review on the application of RFM model. First, this paper depicts the definition and the scoring scheme of RFM. Later, this paper summarizes how RFM model has been applied in various areas. Next, the advantages and disadvantages of RFM model are presented and discussed. Moreover, this paper also elaborates on the relative advantages and disadvantages of RFM and other models. Finally, this paper reviews the articles about the extended RFM to show how RFM can be combined with other variables and other models. The review on RFM model is essential and can provide fruitful insight to researchers and decision makers. In fact, RFM model has been proven to be very successful in a variety of practical areas. Therefore, RFM can help identify valuable customers and develop effective marketing strategy for not only profit organizations (including marketing industry, banking and insurance industries, telecommunication industry, travelling industry and on-line industry), but also non-profit organizations and government agencies. For researchers, they can get a full understanding on the overview of RFM model so that they can have more ideas on the refined application of RFM. On the other hand, decision makers can identify valuable customers and develop important strategy by adopting RFM. As a matter of fact, RFM facilitates decision makers to observe customer behavior (Buckinx and Poel, 2005), segment customers (Hughes, 1996; Kahan, 1998), estimate the response probability for each offer type (Spring et al., 1999), calcuate customer value and customer lifetime value (Liu and Shih 2005a; Sohrabi and Khanlari, 2007) and evaluate on-line reviewers (Li et al., 2010). Particularly, direct marketing has a long history in using RFM model (Tsai and Chiu, 2004). Therefore, through the Wei et al. 4205 review of the application of RFM model, decision makers would gain insights on RFM and would be able to apply RFM more effectively to resolve the problems encountered in daily activities and develop effective strategy to satisfy a wide variety of customer needs.

**2.4. Elbow Criterion**

**M A Syakur et al.[6]**

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. The same method can be used to choose the number of parameters in other data-driven models, such as the number of principal components to describe a data set.

Illustration of K value on Elbow combination with K-Means was graph of cluster relationship with error decreasing, increasing value of K then graph will decrease slowly until result of value of K is stable.

For example, the value of the cluster K = 2 to K = 3, then from K = 3 to K = 4, it shows a drastic decrease to form the elbow at point K = 3 then the ideal cluster k is K = 3. The combined Elbow and K-Means Methods can determine the value of K at the best cluster.

1. Find k as the number of clusters formed. This study will use the elbow criterion method to select the number of k clusters to be used for grouping data on the K-Means algorithm. The elbow method is expressed by Sum of Squared Error. With k = many clusters formed = the i-th cluster, x = the data present in each cluster.

2. Determine the cluster's center point at the beginning at random. Early centroid determination is done randomly from the available objects as much as cluster k, then to calculate the next i-cluster centroid, by the following formula: ∑ ;

3. Calculate the distance of each object to each centroid using the Euclidian Distance. With xi : Variable on Object x to-i and yi : Variable output y n : The number of objects

4. Allocate each object into the nearest centroid.

5. Allocation of objects into each cluster at iteration with k-means. Where each cluster member object has been measured the proximity distance to the cluster's center point.

6. Perform iteration, then process determine the position of new centroid.

7. Repeat step 3 if the new centroid position with the old centroid is not the same.

Chapter

|  |  |
| --- | --- |
| 3 | PRELIMINARIES |

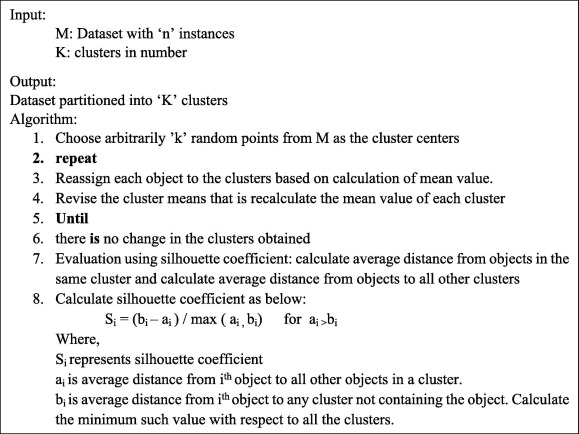
**3.1. What is RFM Model ?**

• ***Recency****:* It refers to the number of days before the reference date when a customer made the last purchase. Lesser the value of recency, higher is the customer visit to a store.

• ***Frequency****:* It is the period between two subsequent purchases of a customer. Higher the value of frequency more is the customer visit to the company.

• ***Monetary****:* This refers to the amount of money spent by a customer during a specific period of time. Higher the value, more is the profit generated to the company.

**3.2. Algorithm for Customer Segmentation**



**3.3. Flowchart and broader steps for k means**

NO OF CLUSTER K

CENTEROID

CALCULATE OBJEC DESTENCE TO CENTROID

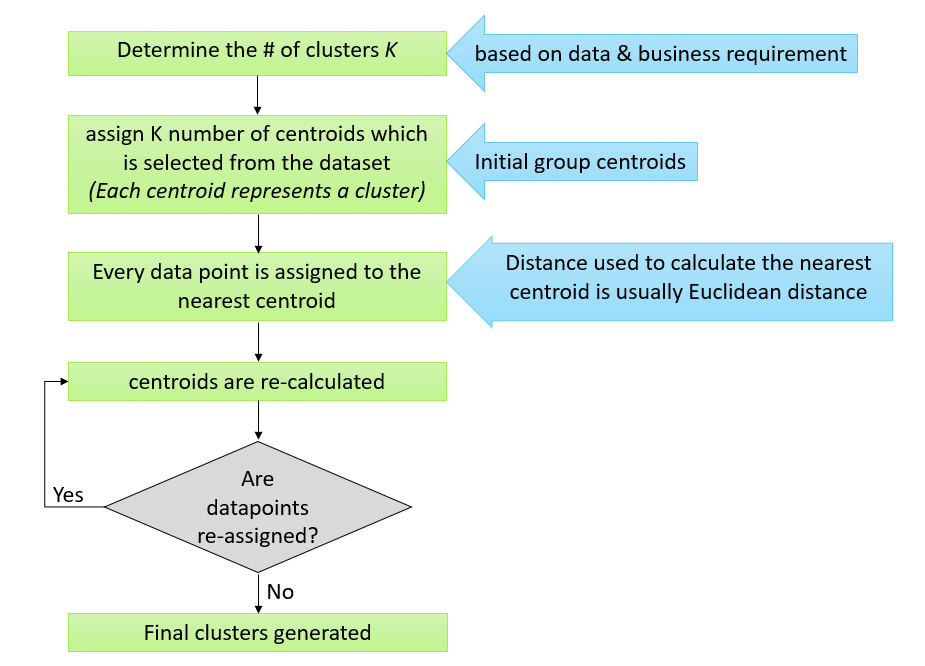
GROUPING ON MINIMUM DISTENCE

NO OBJECT MOVE ?

**NO**

**YES**

Fig. 3.3.1. K-means Flowchart



**Fig. 3.3.2. Broad steps of the k-means algorithm**

**3.4. Algorithm for Advertising on the basis of k-means values**

**3.5. Algorithm for Elbow Method**

1. Find k as the number of clusters formed. This study will use the elbow criterion method to select the number of k clusters to be used for grouping data on the K-Means algorithm. The elbow method is expressed by Sum of Squared Error. With k = many clusters formed = the i-th cluster, x = the data present in each cluster.

2. Determine the cluster's center point at the beginning at random. Early centroid determination is done randomly from the available objects as much as cluster k, then to calculate the next i-cluster centroid, by the following formula: ∑ ;

3. Calculate the distance of each object to each centroid using the Euclidian Distance. ( ) ‖ ‖ √∑ ( ) with xi : Variable on Object x to-i and yi : Variable output y n : The number of objects

4. Allocate each object into the nearest centroid.

5. Allocation of objects into each cluster at iteration with k-means. Where each cluster member object has been measured the proximity distance to the cluster's center point.

6. Perform iteration, then process determine the position of new centroid by using equation (2.2).

7. Repeat step 3 if the new centroid position with the old centroid is not the same.

Chapter

|  |  |
| --- | --- |
| 4 | PROPOSED SCHEME |

**4.1. Research Methodology flowchart**

Define business object

Collecting data

Data preparation

Analyzing variable & looking for the responsive amongst the varible

Date processing with selected method

Performance evolution

Fig. 3.2 Research Methodology flowchart

**4.2.1. DFD DIAGRAM**

**LEVEL ZERO**

CUSTOMER

ADMIN

**Q A Q A**

**4.2.2. DFD DIAGRAM**

**LEVEL ONE**

|  |
| --- |
|  |

**4.3. ER DIAGRAM**

FILTERED CUSTOMER DATA SET

RFM MODEL

CUSTOMER DATA SET

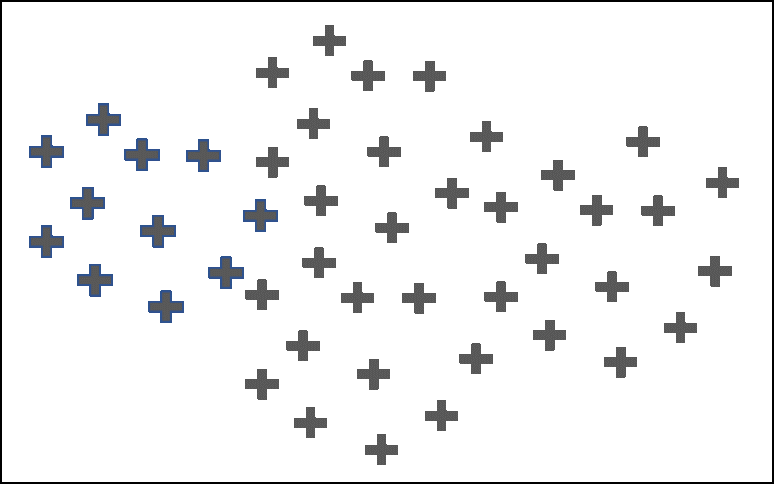
K MEANS CLUSTER

GGROUP OF CUSTOMER

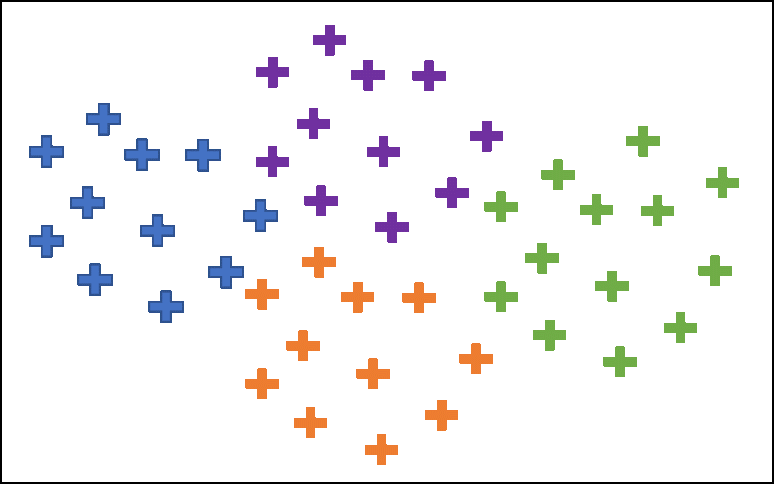
Fig. 4.3. E.R. Diagram

* 1. **Selection of clustering algorithm :**

Clustering is an unsupervised machine learning technique, where there are no defined dependent and independent variables. The patterns in the data are used to identify / group similar observations.

****

**Fig 4.4.1** Original Dataset



**Fig. 4.4.2** After Clustering

The objective of any clustering algorithm is to ensure that the distance between datapoints in a cluster is very low compared to the distance between 2 clusters. In other words, members of a group are very similar, and members of different groups are extremely dissimilar. A cluster is understood as a conceptually meaningful group of objects that have common characteristics. Clustering can be used for customer segmentation for additional analysis. The literature survey that one of the applications of K-means is customer segmentation. K-Means clustering algorithm is a prototype based partition clustering technique that finds the user specified number of clusters, which are represented by their centroids. K-Means is computationally faster and performs well on large datasets compared to other clustering methods. Another advantage of using K-Means is that the algorithm requires only one input parameter ‘K‘ than other algorithms. Also it decreases the rate of misclassification of data. One of the major applications of K-Means is customer segmentation. The present work uses K-Means algorithm.

* 1. **Proposed methodology :**

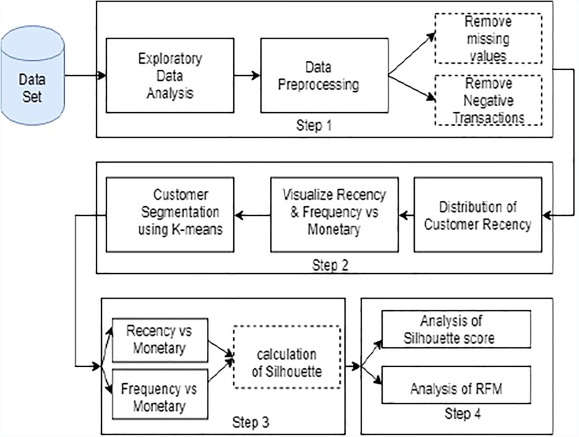
****

Fig. 4.5. Steps of Proposed Methodology

The proposed methodology can be broadly divided into 4 steps as shown in the Fig.1. The corresponding details are explained as below:

* **Step 1:** Exploratory Analysis and Data Preprocessing.

Exploratory data analysis (EDA) refers to initial exploration of data in order to extract or discover the patterns with the help of statistics or graphical representations. In this activity, EDA helps in identifying unique customers, percentage of orders by top 10 or more, information about the data, mismatch in description, stock code and to check null values. Further, data preprocessing is applied to identify and remove missing customer identification number, negative transactions and so on.

* **Step 2**:

a) Execution of RFM Analysis

After data is preprocessed, check for recent transactions, frequency and the amount spent by the customers. In order to create recency variable, decide the reference date - that is one day prior to the last transaction. RFM analysis is a very popular customer segmentation and identifiable technique in database marketing. It is significant especially in Retail Industry. Each customer under RFM is scored based on three factors.

b) K-Means algorithm is applied using Euclidean distance metric to partition the customers for RFM values. K-Means is used twice to analyze the amount obtained for Recent and Frequent transactions as mentioned below:

1. To partition the customers based on the amount generated with recent transactions.
2. To group the customers on the amount generated with frequent transactions

* Step 3: Calculation of Silhouette Score

Clusters obtained in step 2.b) are evaluated using silhouette score, which analyzes how well the resulting clusters are separated. It lies in a range of [−1, +1]. If the value is near to +1, then objects (customers) are grouped far away from neighboring clusters, whereas if it is −1, then objects (customers) might have been assigned to a wrong cluster or preprocessing of data is not correct.

* Step 4: Evaluation of clusters

Let K = number of clusters. Silhouette values are compared for K = 3 and K = 5 to identify the optimal clusters based on the value. After the analysis, compare the sales recency with sales amount and sales frequency with sales amount from one cluster to another cluster respectively. This helps in identifying the group of customers having highest sales recency, sales frequency and the sales amount.

### Mathematical model :

Clustering using K-means algorithm is a method of unsupervised learning used for data analysis. This algorithm identifies ‘K’ centroids from the dataset ‘D’ and assigns the non- overlapping data points to each of the nearest clusters. The intra-cluster distance is maximum compared to inter-cluster distance in K-means algorithm. Since it is an iterative approach, data points are moved to different clusters, based on the centroids calculation.

As per the pseudo algorithm shown in fig.2, the mathematical model for the manual calculation of silhouette for an object is given below. Consider K clusters of which each cluster contains variable objects. Since K-Means is applied twice in the present experiment, objects are clustered based on a customer transaction data for recency vs monetary and frequency vs monetary values.

K=p1,q1,p2,q2…px,qx,p1,q1,p2,q2…py, qy,…p1,q1,p2,q2…pz,qz where,

K = number of clusters, (p,q) = object in a cluster.

The mathematics behind clustering, in very simple terms involves minimizing the sum of square of distances between the cluster centroid and its associated data points:

Minimize2…………………….……………..(eq. 1)

* *K* = number of clustersij
* *N*= number of data points
* *C*=centroid of cluster j
* (*x`ij — cj*)– Distance between data point and centroid to which it is assigned

### Deciding on the optimum number of clusters ‘K’ :

### The main input for k-means clustering is the number of clusters. This is derived using the concept of *minimizing within cluster sum of square (WCSS)*. A screen plot is created which plots the number of clusters in the X axis and the WCSS for each cluster number in the y-axis.

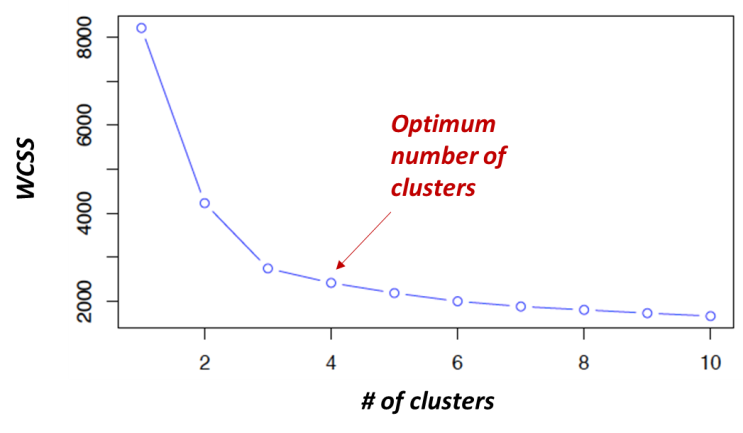


Fig. 4.7. Screen plot / Elbow method to determine optimum number of clusters

As the number of clusters increase, the WCSS keeps decreasing. The decrease of WCSS is initially steep and then the rate of decrease slows down resulting in an elbow plot. The number of clusters at the elbow formation usually gives an indication on the optimum number of clusters. This combined with specific knowledge of the business requirement should be used to decide on the optimum number of clusters.

* 1. **Demonstrating with a sample dataset :**

The data set contains the **annual income** of ~300 customers and their **annual spend** on an e-commerce site. We will use the k-means clustering algorithm to derive the optimum number of clusters and understand the underlying customer segments based on the data provided.

* + 1. **About the data set :**

The dataset consists of Annual income (in $000) of 303 customers and their total spend (in $000) on an e-commerce site for a period of one year. Let us explore the data using numpy and pandas libraries in python.

#**Load the required packages**  
import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt#**Plot styling**  
import seaborn as sns; sns.set() # for plot styling  
%matplotlib inlineplt.rcParams['figure.figsize'] = (16, 9)  
plt.style.use('ggplot')#**Read the csv file**  
dataset=pd.read\_csv('CLV.csv')#**Explore the dataset**  
dataset.head()#top 5 columns  
len(dataset) # of rows#**descriptive statistics of the dataset**  
dataset.describe().transpose()

Table 4.8.1.1. dataset.head()

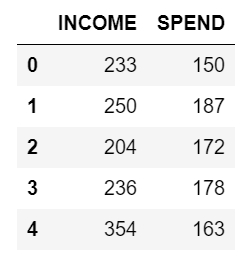
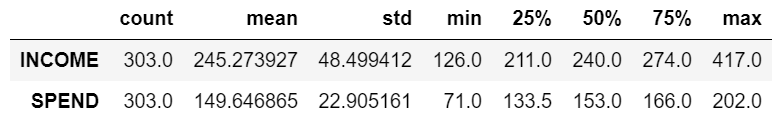


Table 4.8.1.2. dataset.describe().transpose()



The dataset consists of 303 rows. The mean annual income is 245000 and the mean annual spend is 149000. The distribution of the annual income and annual spend has been illustrated with a distplot and violinplot.

* + 1. **Visualizing the data :**

The displot and violinplot give an indication of the data distribution of Income and Spend.

**#Visualizing the data – displot**

plot\_income = sns.distplot(dataset["INCOME"])  
plot\_spend = sns.distplot(dataset["SPEND"])  
plt.xlabel('Income / spend')

For our dataset, we will arrive at the optimum number of clusters using the elbow method:

**#Using the elbow method to find the optimum number of clusters**

from sklearn.cluster import KMeans  
wcss = []  
for i in range(1,11):  
 km=KMeans(n\_clusters=i,init='k-means++', max\_iter=300, n\_init=10, random\_state=0)  
 km.fit(X)  
 wcss.append(km.inertia\_)  
plt.plot(range(1,11),wcss)  
plt.title('Elbow Method')  
plt.xlabel('Number of clusters')  
plt.ylabel('wcss')  
plt.show()

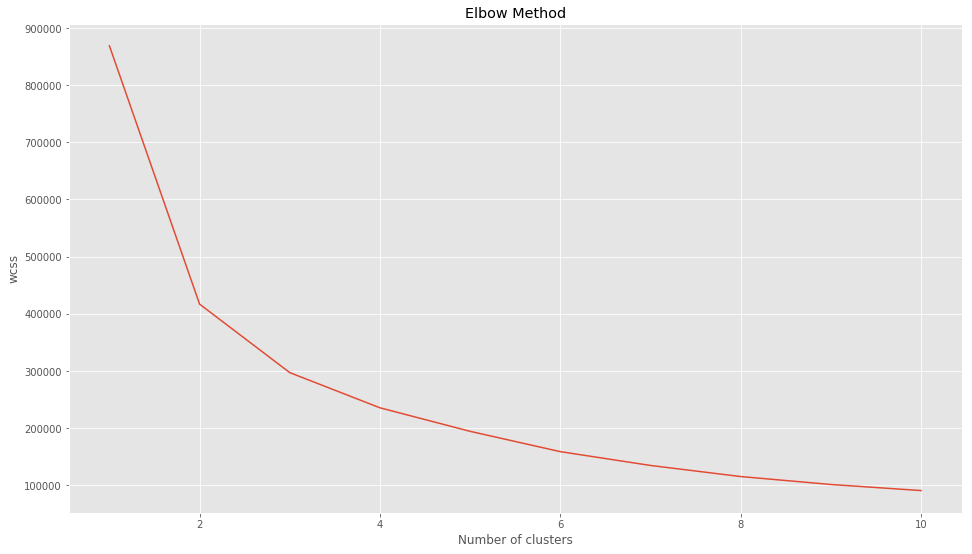


Fig. 4.8.2. Screen plot of given dataset on customer Income & Spend

Based on the elbow plot, we could choose 4,5 or 6 clusters. Let us try both the number of clusters and visualize the clusters to decide on the final number of clusters.

* + 1. **Fitting the k-means to the dataset with k=4 :**

**##Fitting kmeans to the dataset with k=4**

km4=KMeans(n\_clusters=4,init='k-means++', max\_iter=300, n\_init=10, random\_state=0)  
y\_means = km4.fit\_predict(X)**#Visualizing the clusters for k=4**  
plt.scatter(X[y\_means==0,0],X[y\_means==0,1],s=50, c='purple',label='Cluster1')  
plt.scatter(X[y\_means==1,0],X[y\_means==1,1],s=50, c='blue',label='Cluster2')  
plt.scatter(X[y\_means==2,0],X[y\_means==2,1],s=50, c='green',label='Cluster3')  
plt.scatter(X[y\_means==3,0],X[y\_means==3,1],s=50, c='cyan',label='Cluster4')plt.scatter(km4.cluster\_centers\_[:,0], km4.cluster\_centers\_[:,1],s=200,marker='s', c='red', alpha=0.7, label='Centroids')  
plt.title('Customer segments')  
plt.xlabel('Annual income of customer')  
plt.ylabel('Annual spend from customer on site')  
plt.legend()  
plt.show()

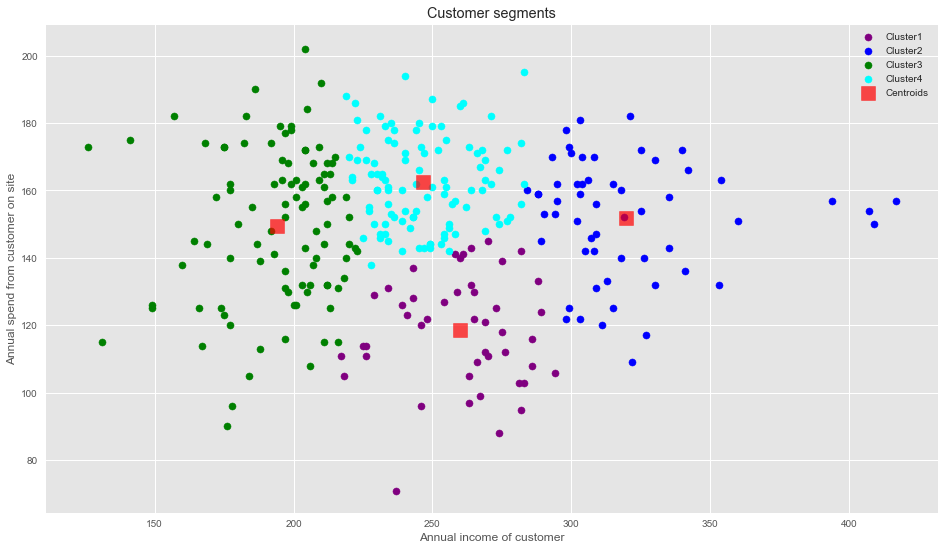


Fig. 4.8.3. Cluster plot : k=4

The plot shows the distribution of the 4 clusters. We could interpret them as the following customer segments:

1. Cluster 1: Customers with medium annual income and low annual spend
2. Cluster 2: Customers with high annual income and medium to high annual spend
3. Cluster 3: Customers with low annual income
4. Cluster 4: Customers with medium annual income but high annual spend
   * 1. **Fitting the k-means to the dataset with k=6 :**

Cluster 4 straight away is one potential customer segment. However, Cluster 2 and 3 can be segmented further to arrive at a more specific target customer group. Let us now look at how the clusters are created **when k=6:**

**##Fitting kmeans to the dataset - k=6**

km4=KMeans(n\_clusters=6,init='k-means++', max\_iter=300, n\_init=10, random\_state=0)  
y\_means = km4.fit\_predict(X)**#Visualizing the clusters**  
plt.scatter(X[y\_means==0,0],X[y\_means==0,1],s=50, c='purple',label='Cluster1')  
plt.scatter(X[y\_means==1,0],X[y\_means==1,1],s=50, c='blue',label='Cluster2')  
plt.scatter(X[y\_means==2,0],X[y\_means==2,1],s=50, c='green',label='Cluster3')  
plt.scatter(X[y\_means==3,0],X[y\_means==3,1],s=50, c='cyan',label='Cluster4')  
plt.scatter(X[y\_means==4,0],X[y\_means==4,1],s=50, c='magenta',label='Cluster5')  
plt.scatter(X[y\_means==5,0],X[y\_means==5,1],s=50, c='orange',label='Cluster6')plt.scatter(km.cluster\_centers\_[:,0], km.cluster\_centers\_[:,1],s=200,marker='s', c='red', alpha=0.7, label='Centroids')  
plt.title('Customer segments')  
plt.xlabel('Annual income of customer')  
plt.ylabel('Annual spend from customer on site')  
plt.legend()  
plt.show()

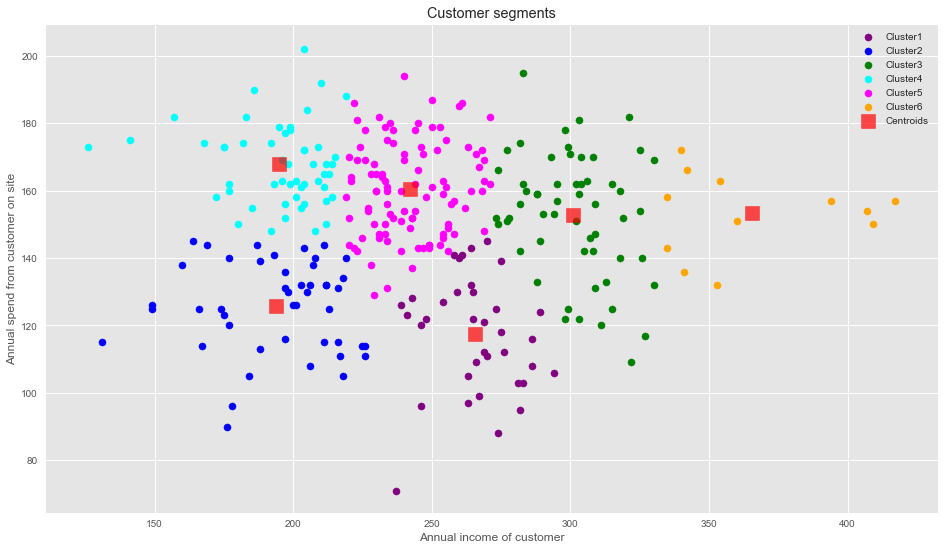


Fig. 4.8.4. Cluster plot : k=6

Setting the number of clusters to 6 seems to provide a more meaningful customer segmentation.

1. Cluster 1: Medium income, low annual spend
2. Cluster 2: Low income, low annual spend
3. Cluster 3: High income, high annual spend
4. Cluster 4: Low income, high annual spend
5. Cluster 5: Medium income, low annual spend
6. Cluster 6: Very high income, high annual spend

Thus it is evident that 6 clusters provides a more meaningful segmentation of the customers.

* + 1. **Marketing strategies for the customer segments :**

Based on the 6 clusters, we could formulate marketing strategies relevant to each cluster:

* A typical strategy would focus certain promotional efforts for the high value customers of Cluster 6 & Cluster 3.
* Cluster 4 is a unique customer segment, where in spite of their relatively lower annual income, these customers tend to spend more on the site, indicating their loyalty. There could be some discounted pricing based promotional campaigns for this group so as to retain them.
* For Cluster 2 where both the income and annual spend are low, further analysis could be needed to find the reasons for the lower spend and price-sensitive strategies could be introduced to increase the spend from this segment.
* Customers in clusters 1 and 5 are not spending enough on the site in spite of a good annual income — further analysis of these segments could lead to insights on the satisfaction / dissatisfaction of these customers or lesser visibility of the e-commerce site to these customers. Strategies could be evolved accordingly.
  1. **Portions of Code for our proposed methodology :**

#**RFM\_Modelling**

In [14]:

#Recency = Latest Date - Last Inovice Data, Frequency = count of invoice no. of transaction(s), Monetary = Sum of Total

#Amount for each customer

import datetime as dt

#Set Latest date 2011-12-10 as last invoice date was 2011-12-09. This is to calculate the number of days from recent purchase

Latest\_Date = dt.datetime(2011,12,10)

#Create RFM Modelling scores for each customer

RFMScores = Rtl\_data.groupby('CustomerID').agg({'InvoiceDate': lambda x: (Latest\_Date - x.max()).days, 'InvoiceNo': lambda x: len(x), 'TotalAmount': lambda x: x.sum()})

#Convert Invoice Date into type int

RFMScores['InvoiceDate'] = RFMScores['InvoiceDate'].astype(int)

#Rename column names to Recency, Frequency and Monetary

RFMScores.rename(columns={'InvoiceDate': 'Recency',

'InvoiceNo': 'Frequency',

'TotalAmount': 'Monetary'}, inplace=True)

RFMScores.reset\_index().head()

Out [14] : Table 5.6.5.

In [15]:

#Descriptive Statistics (Recency)

RFMScores.Recency.describe()

Out [15]: Table 5.6.6.

In [16]:

#Recency distribution plot

import seaborn as sns

x = RFMScores['Recency']

ax = sns.distplot(x)

Out [16]: Fig. 5.6.1.

In [17]:

#Descriptive Statistics (Frequency)

RFMScores.Frequency.describe()

Out [17]: Table 5.6.7.

In [18]:

#Frequency distribution plot, taking observations which have frequency less than 1000

import seaborn as sns

x = RFMScores.query('Frequency < 1000')['Frequency']

ax = sns.distplot(x)

Out [18]: Fig. 5.6.2.

In [19]:

#Descriptive Statistics (Monetary)

RFMScores.Monetary.describe()

Out [19]: Table 5.6.8

In [20]:

#Monateray distribution plot, taking observations which have monetary value less than 10000

import seaborn as sns

x = RFMScores.query('Monetary < 10000')['Monetary']

ax = sns.distplot(x)

Out [20]: Fig. 5.6.3.

In [21]:

#Split into four segments using quantiles

quantiles = RFMScores.quantile(q=[0.25,0.5,0.75])

quantiles = quantiles.to\_dict()

In [22]:

quantiles

#**K\_Means\_Clustering**

In [29]:

#Handle negative and zero values so as to handle infinite numbers during log transformation

def handle\_neg\_n\_zero(num):

if num <= 0:

return 1

else:

return num

#Apply handle\_neg\_n\_zero function to Recency and Monetary columns

RFMScores['Recency'] = [handle\_neg\_n\_zero(x) for x in RFMScores.Recency]

RFMScores['Monetary'] = [handle\_neg\_n\_zero(x) for x in RFMScores.Monetary]

#Perform Log transformation to bring data into normal or near normal distribution

Log\_Tfd\_Data = RFMScores[['Recency', 'Frequency', 'Monetary']].apply(np.log, axis = 1).round(3)

In [30]:

#Data distribution after data normalization for Recency

Recency\_Plot = Log\_Tfd\_Data['Recency']

ax = sns.distplot(Recency\_Plot)

Out [30]: Fig. 5.6.4.

In [31]:

#Data distribution after data normalization for Frequency

Frequency\_Plot = Log\_Tfd\_Data.query('Frequency < 1000')['Frequency']

ax = sns.distplot(Frequency\_Plot)

Out[31]: Fig. 5.6.5.

In [32]:

#Data distribution after data normalization for Monetary

Monetary\_Plot = Log\_Tfd\_Data.query('Monetary < 10000')['Monetary']

ax = sns.distplot(Monetary\_Plot)

Out [32]: Fig. 5.6.6.

In [33]:

from sklearn.preprocessing import StandardScaler

#Bring the data on same scale

scaleobj = StandardScaler()

Scaled\_Data = scaleobj.fit\_transform(Log\_Tfd\_Data)

#Transform it back to dataframe

Scaled\_Data = pd.DataFrame(Scaled\_Data, index = RFMScores.index, columns = Log\_Tfd\_Data.columns)

In [36]:

from sklearn.cluster import KMeans

sum\_of\_sq\_dist = {}

for k in range(1,15):

km = KMeans(n\_clusters= k, init= 'k-means++', max\_iter= 1000)

km = km.fit(Scaled\_Data)

sum\_of\_sq\_dist[k] = km.inertia\_

#Plot the graph for the sum of square distance values and Number of Clusters

sns.pointplot(x = list(sum\_of\_sq\_dist.keys()), y = list(sum\_of\_sq\_dist.values()))

plt.xlabel('Number of Clusters(k)')

plt.ylabel('Sum of Square Distances')

plt.title('Elbow Method For Optimal k')

plt.show()

Out [36]: Fig. 5.6.7.

In [37]:

#Perform K-Mean Clustering or build the K-Means clustering model

KMean\_clust = KMeans(n\_clusters= 3, init= 'k-means++', max\_iter= 1000)

KMean\_clust.fit(Scaled\_Data)

#Find the clusters for the observation given in the dataset

RFMScores['Cluster'] = KMean\_clust.labels\_

RFMScores.head()

Out [37]: Table 5.6.12.

In [38]:

from matplotlib import pyplot as plt

plt.figure(figsize=(7,7))

##Scatter Plot Frequency Vs Recency

Colors = ["red", "green", "blue"]

RFMScores['Color'] = RFMScores['Cluster'].map(lambda p: Colors[p])

ax = RFMScores.plot(

kind="scatter",

x="Recency", y="Frequency",

figsize=(10,8),

c = RFMScores['Color']

)

Out [38]: Fig. 5.6.8.

In [39]:

RFMScores.head()

Out [39]: Table 5.6.13.

Chapter

|  |  |
| --- | --- |
| 5 | **EXPERIMENTAL RESULTS** |

Cluster Analysis:

The following clusters are created by the model

1. Cluster Orange

2. Cluster Blue

3. Cluster Purple

4. Cluster Red

5. Cluster Green

Fig. 5.1. Cluster analysis

* 1. **Cluster Orange - Balanced Customers:**

They earn less and spend less. We can see people have low annual income and low spending scores, this is quite reasonable as people having low salaries prefer to buy less , in fact, these are the wise people who know how to spend and save money. The shops/mall will be least interested in people belonging to this cluster.

* 1. **Cluster Blue - Pinch Penny Customers:**

Earning high and spending less. We see that people have high income but low spending scores, this is interesting. Maybe these are the people who are unsatisfied or unhappy by the mall’s services. These can be the prime targets of the mall, as they have the potential to spend money. So, the mall authorities will try to add new facilities so that they can attract these people and can meet their needs.

* 1. **Cluster Purple - Normal Customer:**

Customers are average in terms of earning and spending An Average consumer in terms of spending and Annual Income we see that people have average income and an average spending score, these people again will not be the prime targets of the shops or mall, but again they will be considered and other data analysis techniques may be used to increase their spending score

* 1. **Cluster Red - Spenders:**

This type of customers earns less but spends more Annual Income is less but spending high, so can also be treated as potential target customer we can see that people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. Maybe it’s because these people are more than satisfied with the mall services. The shops/malls might not target these people that effectively but still will not lose them.

* 1. **Cluster Green - Target Customers:**

Earning high and also spending high Target Customers. Annual Income High as well as Spending Score is high, so a target consumer. We see that people have high income and high spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. These people might be the regular customers of the mall and are convinced by the mall’s facilities.

* 1. **Outputs:**

Table 5.6.1. Few Entries from our Dataset

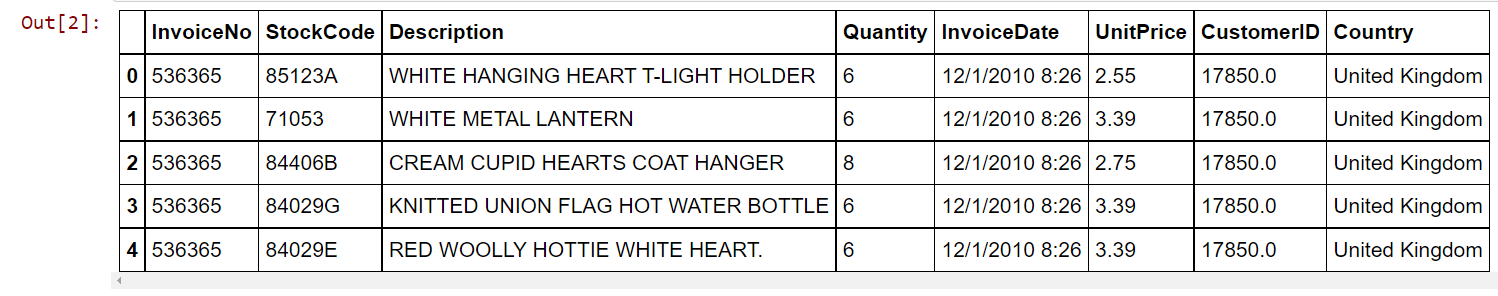


Table 5.6.2. Selecting the country with maximum no. of customers

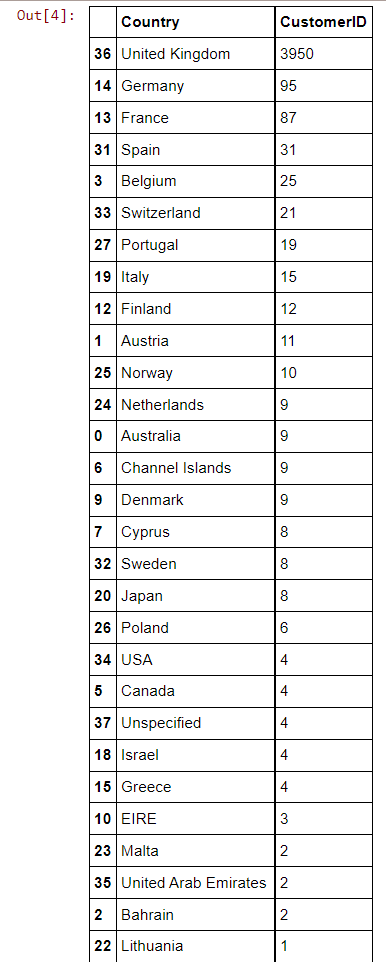


Table 5.6.3. Figuring out which attributes have how many null entries

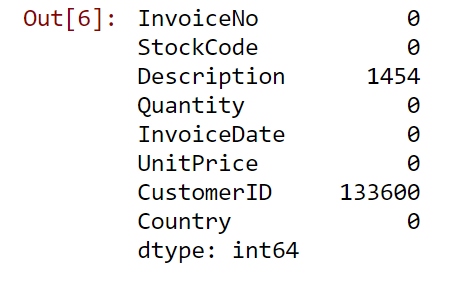


Table 5.6.4. Adding an extra column for storing total amount spent

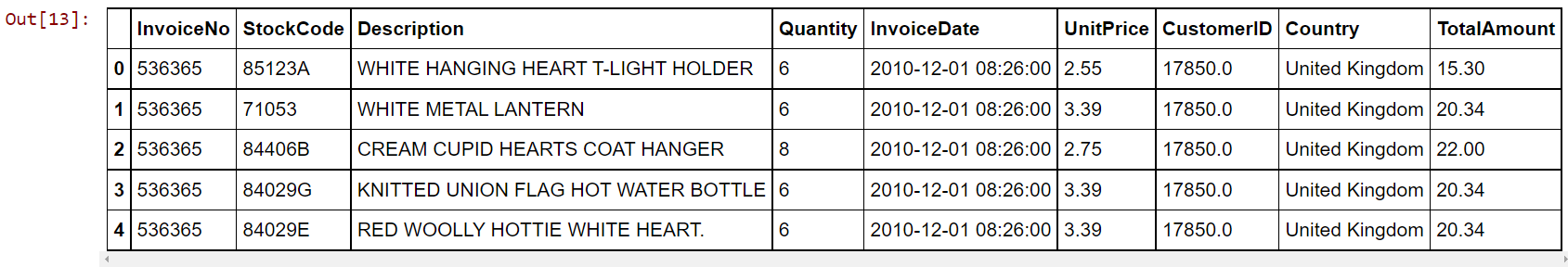


Table 5.6.5. RFM Scores

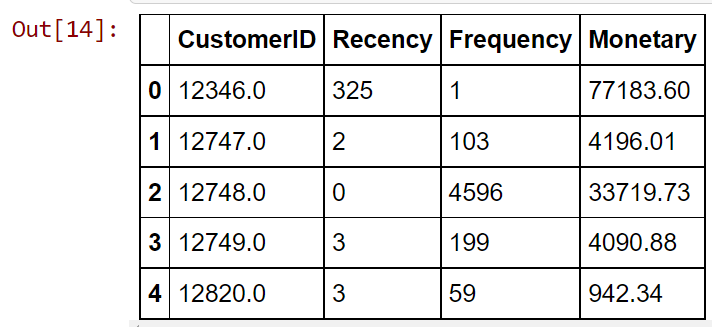
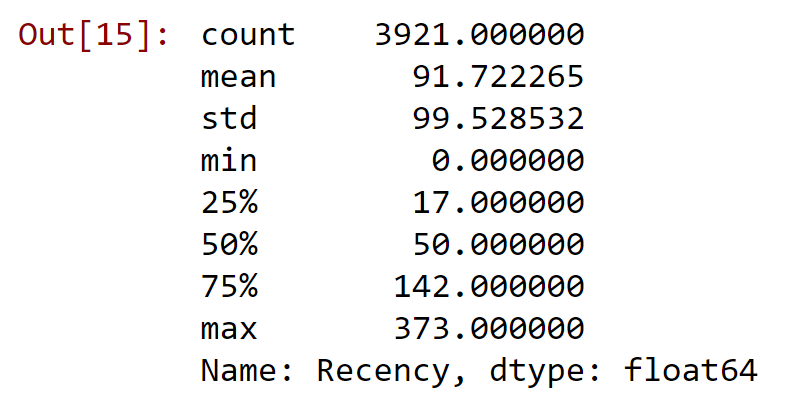
****

Table 5.6.6. Descriptive Statistics of Recency

****

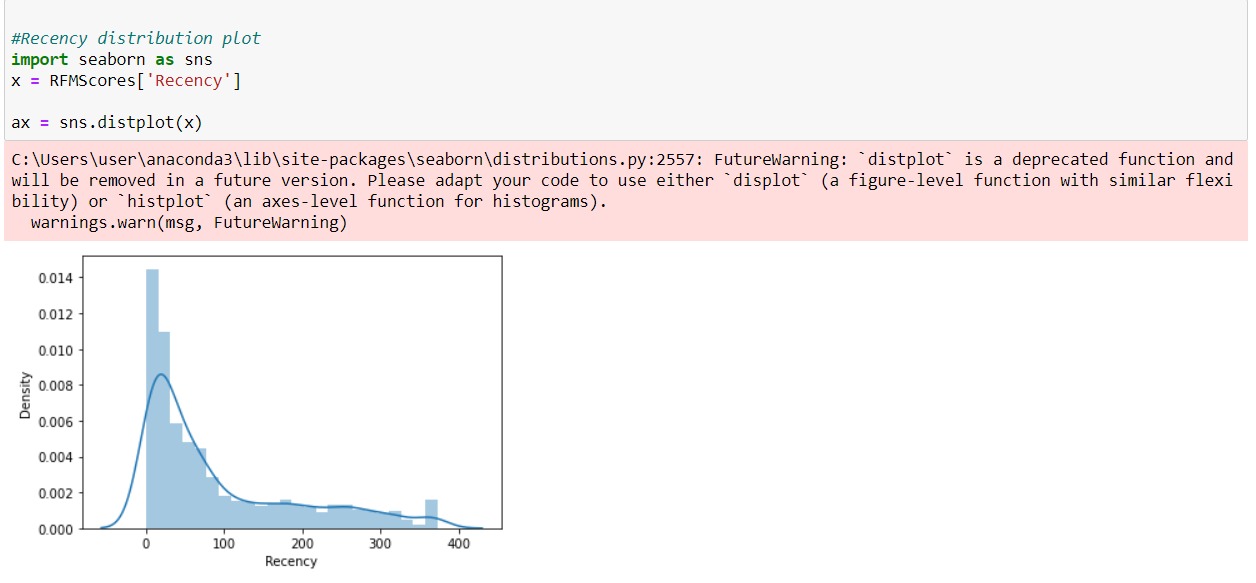
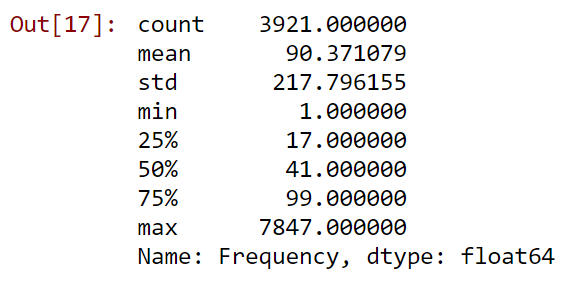
****

Fig. 5.6.1. Recency Distribution Plot

Table 5.6.7. Descriptive Statistics of Frequency

****

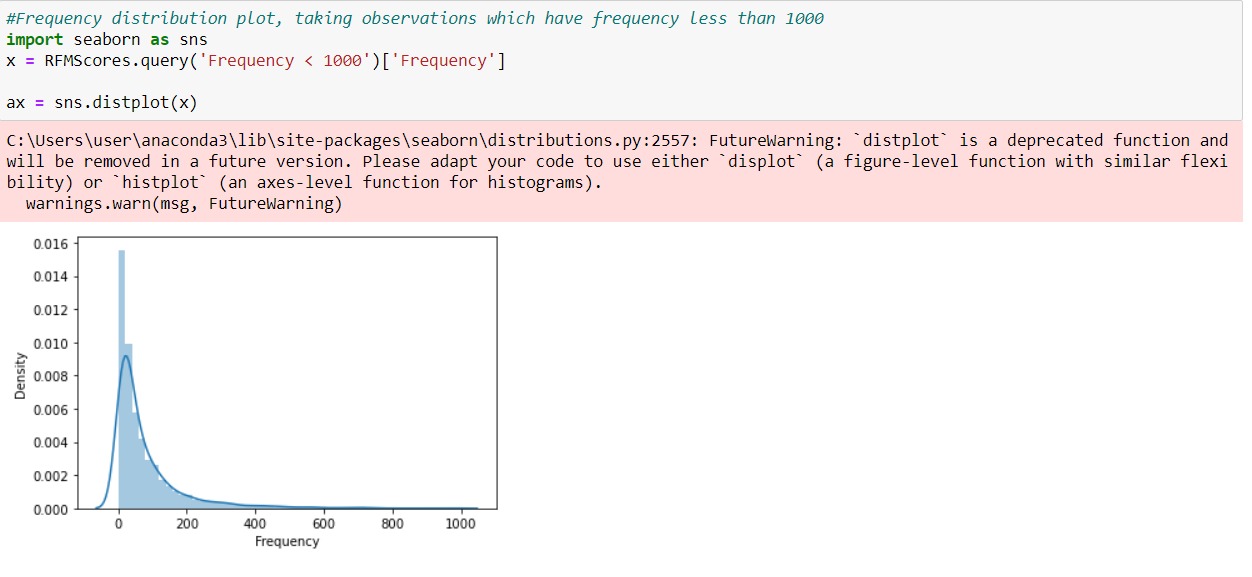
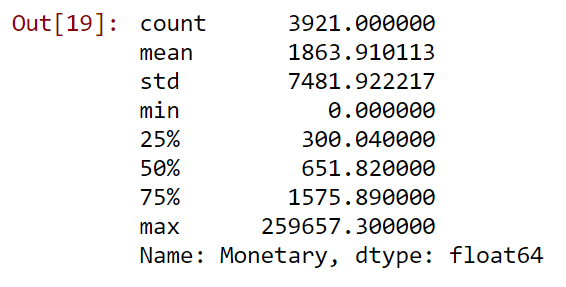
****

Fig. 5.6.2. Frequency Distribution Plot

Table 5.6.8. Descriptive Statistics of Monetary

****

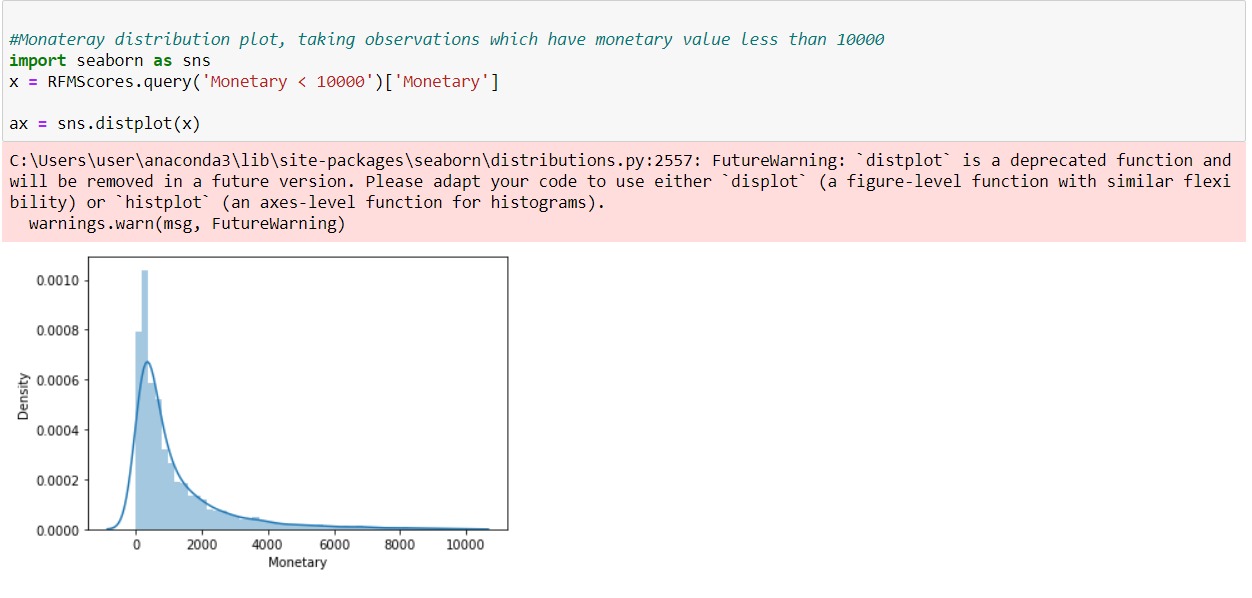
****

Fig. 5.6.3. Monetary Distribution Plot

Table 5.6.9. RFM Segment Values

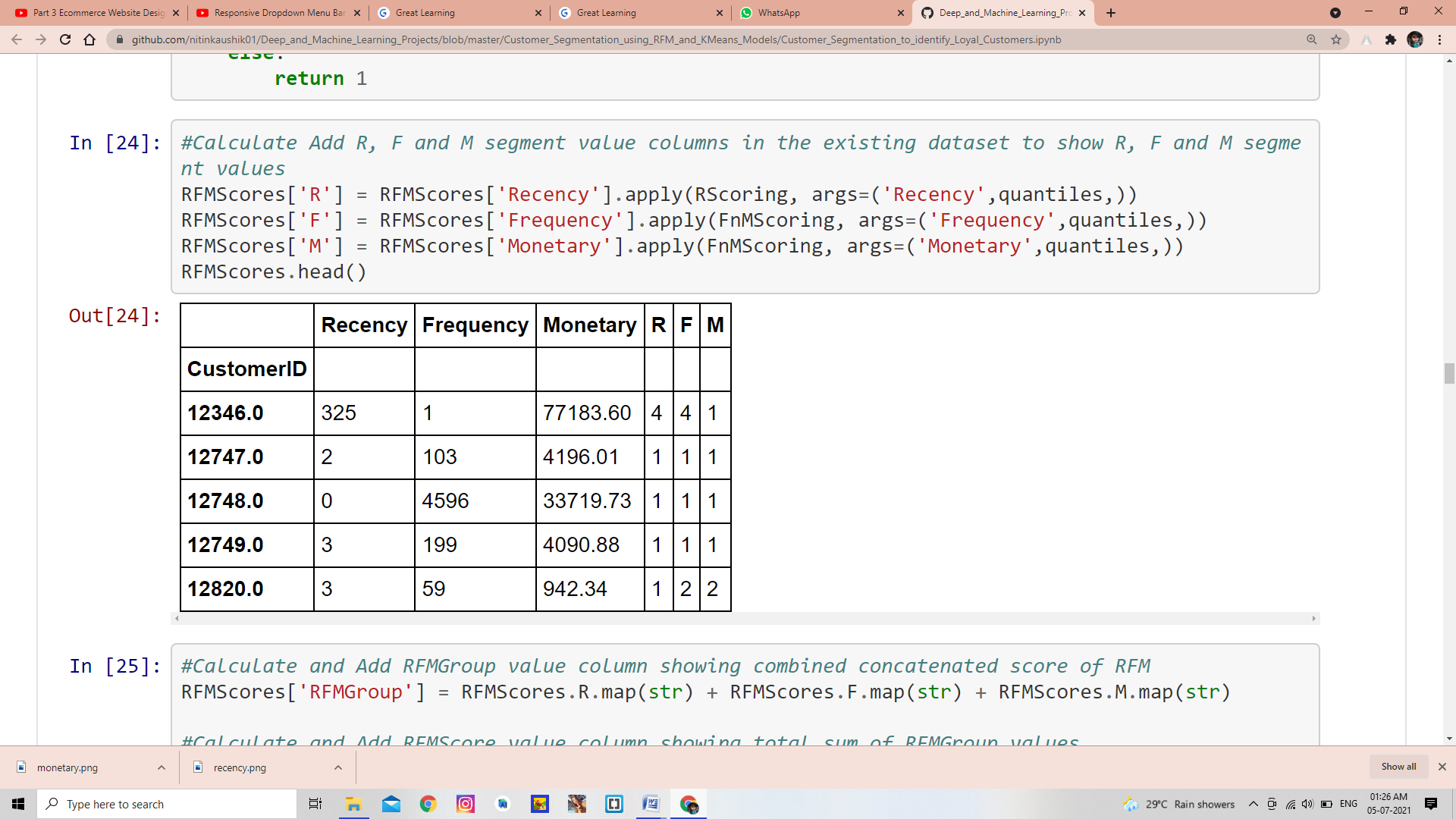
****

Table 5.6.10. R+F+M=RFM Scores

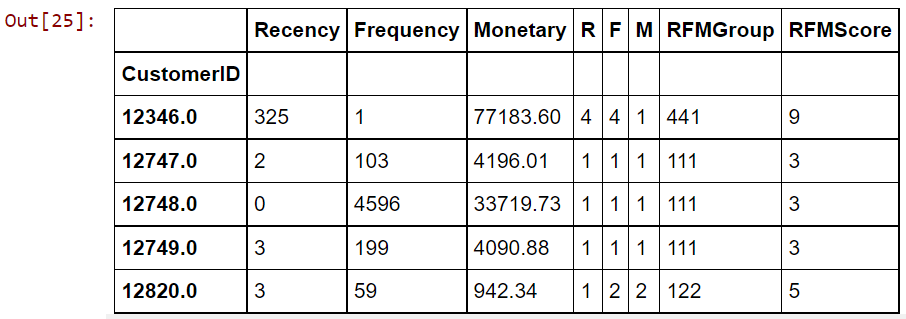
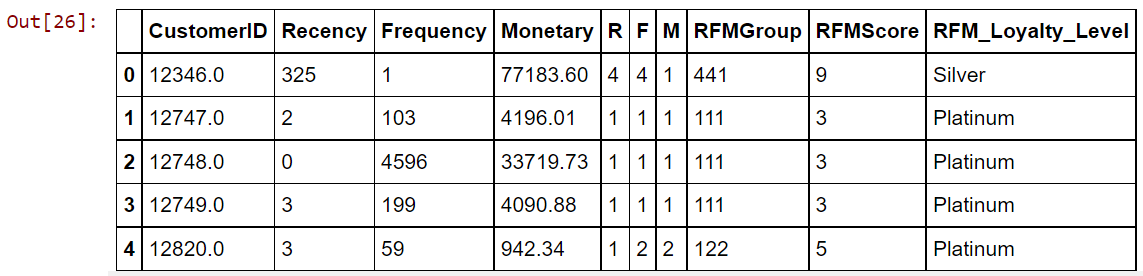
****

Table 5.6.11. Determining the Loyalty Level according to RFM Score

****

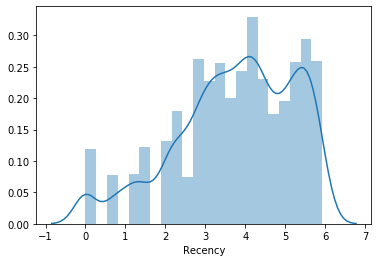
****

Fig. 5.6.4. Data Distribution after Data Normalization for Recency

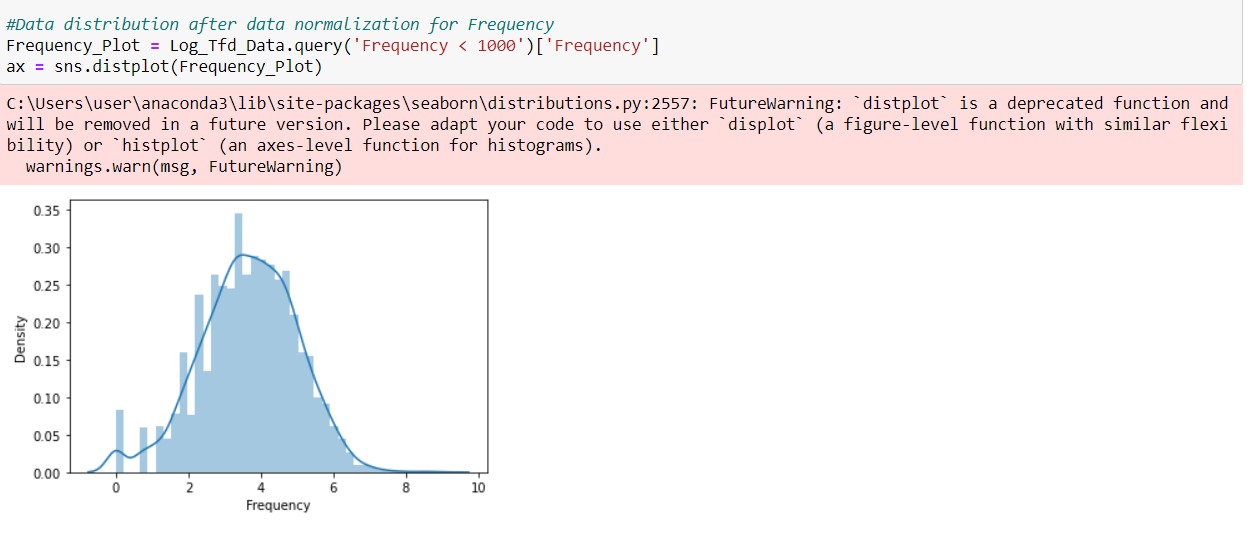
****

Fig. 5.6.5. Data Distribution after Data Normalization for Frequency

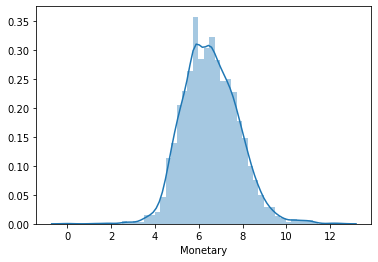


Fig. 5.6.6. Data Distribution after Data Normalization for Monetary

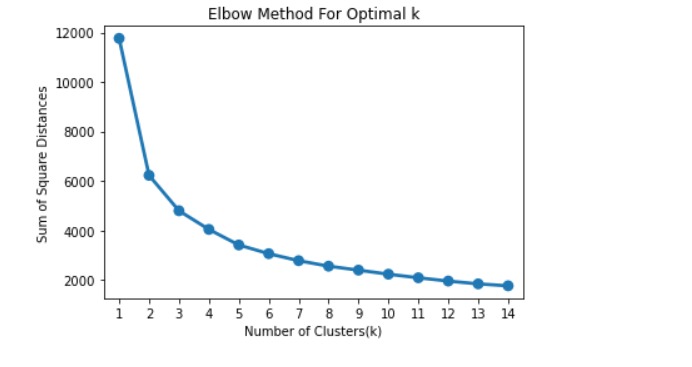
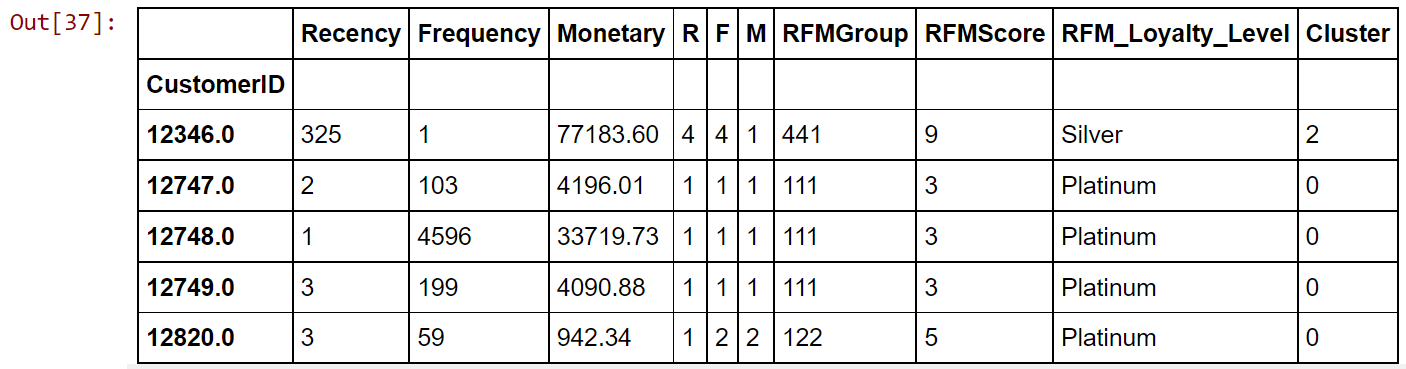
****

Fig. 5.6.7. Elbow Method

Table 5.6.12. K-means Clustering Model



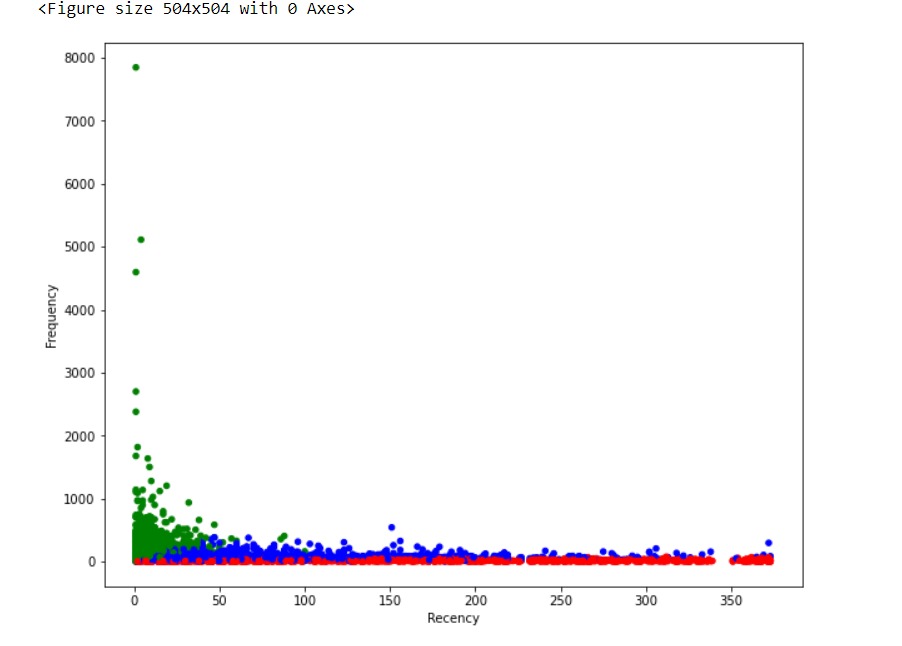
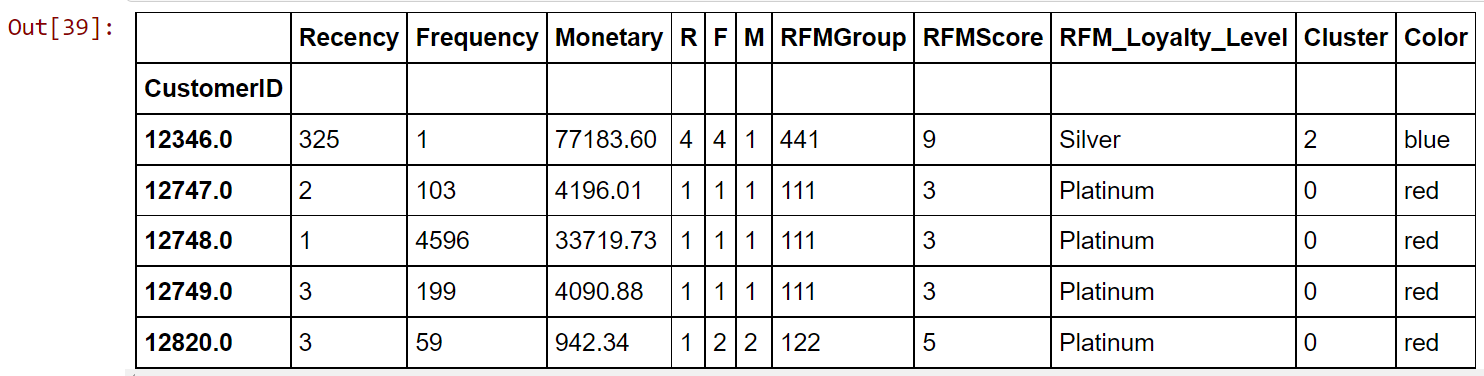
****

Fig. 5.6.8. Frequency vs. Recency

Table 5.6.13. Final Grouping of customers according to their Loyalty Level



Chapter

|  |  |
| --- | --- |
| 6 | **SYSTEM REQUIREMENTS** |

**6.1. HARDWARES :**

1. P4 2.8 GB Processor and Above
2. RAM 2GB and Above
3. HDD (Hard Disk Drive) 500 Gb And Above
   1. **SOFTWARES :**
4. Jupyter Notebook
5. My SQL DataBase
6. Anaconda Command Prompt

Chapter

|  |  |
| --- | --- |
| 7 | **FUTURE SCOPE** |

The scope of future work in this area lies in the study and analysis of specific categories of products, for example, Mobile Phones and Accessories. Various other business parameters such as the most preferred product or the most effective sales technique during as specific event, or some threshold parameters in different regions can be studied for designing effective business enhancement. Such advancements and deliberations in this area will help the enterprises to improve businesses by offering promotions and designing innovative strategies that can prove cutting edge against the competitors.

Chapter

|  |  |
| --- | --- |
| 8 | **CONCLUSION** |

Customer segmentation based on the buying pattern of customers though strategically important, is an equally challenging task. Customer retention is another major concern for both online and physical enterprises. In the present work, the RFM model is implemented for synthetic and real datasets, to analyze customer segmentation. Also, clusters are evaluated using Silhouette Analysis for K-Means clustering algorithm with different number of clusters. Based on the Silhouette Score, the Sales Recency, Sales Frequency and Sales Monetary can be analyzed and an optimal solution is found.

We have thus seen, how we could arrive at meaningful insights and recommendations by using clustering algorithms to generate customer segments. For the sake of simplicity, the dataset used only 2 variables — income and spend. In a typical business scenario, there could be several variables which could possibly generate much more realistic and business-specific insights.

Chapter

|  |  |
| --- | --- |
| 9 | **REFERENCES** |

[1] Batt, R. (2000). *Strategic segmentation in front line services: matching customers, employees and human resource systems. International Journal of Human Resource Management*, 11(3), 540-561.

[2] Bottcher, M., Spott, M.D. Nauck, & Kruse, R. (2009). *Mining changing customer segments in dynamic markets.* Expert Systems with Applications, 36, 155–164.

[3] Buttle, F. 2009. *Customer Relationship management: Concepts and technologies*, Second Edition. Burlington: Elsevier Ltd.

[4] Chetna Sethi and Garima Mishra*, “A Linear PCA based hybrid K-Means PSO algorithm for clustering large dataset,”* International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013, pp.1559-1566.

[5] Lambert, D. M. 2010. “Customer relationship management as a business process.” The Journal of Business & Industrial Marketing 25(1): 4-17.

[6] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, *“Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster,”* in *IOP* Conference Series: Materials Science and Engineering, p. 12017, Surabaya, Indonesia, 2018, IOP Publishing.

[7] McDonald M, Dunbar I (2004*). Market Segmentation: How to Do It, How to Profit from it* ; Elsevier Butterworth Heinemann. Oxford.

[8] N. Bouhmala, A. Viken, J. B. Lonnum, “*Enhanced Genetic Algorithm with K-Means for the Clustering Problem*”, International Journal of Modeling and Optimization, pp. 150-154, 2015.

[9] Palmatier, Robert W. , Dant, Rajiv P. , and Evans, Kenneth R. (2006), “*Factors Influencing the Effectiveness of Relationship Marketing: A Meta-Analysis*,” Journal of Marketing, 70 (October), 136–53.

[10] "Scikit-learn", [online] Available: <https://scikit-learn.org>.

[11] Sharma , Lambert 1990 , *Segmentation of Markets based on Customer Service* , International Journal of Physics Distribution & Logistics Management.

[12] Tajunisha and Saravanan, “*Performance Analysis of kMeans with different initialization methods for high dimensional data*” International Journal of Artificial Intelligence & Applications (IJAIA), Vol.1, No.4, October 2010 .

[13] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, “*An Efficient k-Means Clustering Algorithm: Analysis and Implementation*”, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, No. 7, July 2002.

[14] Wei et al. ***,*** *A review of the application of RFM model,*African Journal of Business Management Vol. 4(19), pp. 4199-4206 (2010)