Customer Segmentation Using ML

**Soham Sawant**

**Electrical & Electronics Engineering, Mumbai University**

**Abstract**—The emergence of many business competitors has engendered severe rivalries among competing businesses in gaining new customers and retaining old ones. Due to the preceding, the need for exceptional customer services becomes pertinent, notwithstanding the size of the business. Furthermore, the ability of any business to understand each of its customers’ needs will earn it greater leverage in providing targeted customer services and developing customised marketing programs for the customers. This understanding can be possible through systematic customer segmentation. Each segment comprises customers who share similar market characteristics. The ideas of Big data and machine learning have fuelled a terrific adoption of an automated approach to customer segmentation in preference to traditional market analyses that are often inefficient especially when the number of customers is too large. In this paper, the kMeans clustering algorithm is applied for this purpose.

Keywords: Clustering; Customer Segmentation; K-Means Algorithm, Elbow method.

# I. INTRODUCTION

Over the years, the competition amongst businesses is increased and the large historical data that is available has resulted in the widespread use of data mining techniques in extracting the meaningful and strategic information from the database of the organisation. Data mining is the process where methods are applied to extract data patterns in order to present it in the human readable format which can be used for the purpose of decision support. According to, Clustering techniques consider data tuples as objects. They partition the data objects into groups or clusters, so that objects within a cluster are similar to one another and dissimilar to objects in other clusters.

Customer Segmentation is the process of division of customer base into several groups called as customer segments such that each customer segment consists of customers who have similar characteristics. The segmentation is based on the similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decision; identification of products associated with each customer segment and to manage the demand and supply of that product; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions.

The thrust of this paper is to identify customer segments using the data mining approach, using the partitioning algorithm called as K-means clustering algorithm.

# II. LITERATURE REVIEW

## A. Customer Segmentation

Over the years, the commercial world is becoming more competitive, as such organizations have to satisfy the needs and wants of their customers, attract new customers, and hence enhance their businesses. The task of identifying and satisfying the needs and wants of each customer in a business is a very complex task. This is because customers may be different in their needs, wants, demography, geography, tastes and preferences, behaviours and so on. As such, it is a wrong practice to treat all the customers equally in business. This challenge has motivated the adoption of the idea of customer segmentation or market segmentation, in which the customers are subdivided into smaller groups or segments wherein members of each segment show similar market behaviours or characteristics. Customer segmentation is a strategy of dividing the market into homogenous groups. The purpose of segmentation is the concentration of marketing energy and force on subdivision (or market segment) to gain a competitive advantage within the segment.

It’s analogous to the military principle of concentration of force to overwhelm energy. Customer or Market segmentation includes geographic segmentation, demographic segmentation, media segmentation, price segmentation, psychographic or lifestyle segmentation, distribution segmentation and time segmentation .

## B. Clustering and k-Means Algorithm

Clustering algorithms generates clusters such that within the clusters are similar based on some characteristics. Similarity is defined in terms of how close the objects are in space.

K-means algorithm in one of the most popular centroid based algorithm.

**Algorithm:** The k-means algorithm for partitioning, where each cluster’s center is represented by the mean value of the objects in the cluster.

*Input:* k: the number of clusters, D: a data set containing n objects.

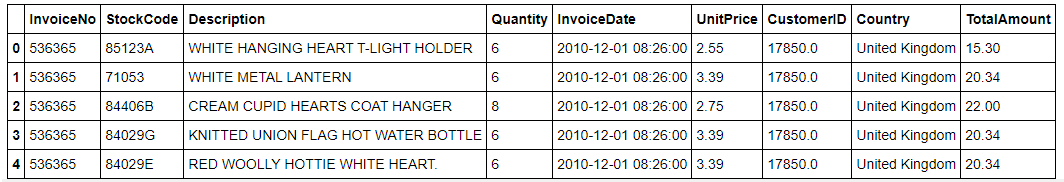
*Output:* A set of k clusters.

*Method:* (1) arbitrarily choose k objects from D as the initial cluster centers; (2) repeat (3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster; (4) update the cluster means, that is, calculate the mean value of the objects for each cluster; (5) until no change.

# III.METHODOLOGY

The data used in this paper was collected from UCI ML Repository. The Data set namely Online Retail Dataset contains 54909 instances and 8 attributes. The attributes include Customer ID , Stock Code, Invoice ID, Product description, quantity, Invoice Date ,Country , Unit Price.

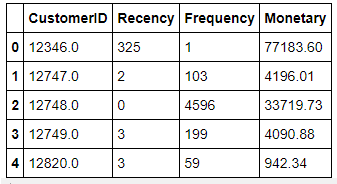
## A. Preparing the Data

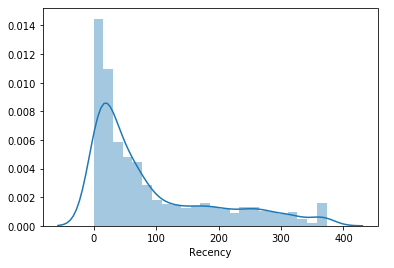
This is a data preparation stage. Dataset was imported and checked for shape. The Customers here were distributed by Country and the Country with the highest count was kept which was United Kingdom and rest all were removed. Here the dataset was checked for any negative, missing or duplicate value and cleaned accordingly. A new Column was added namely Total amount derived from unit price and quantity.

## B. RFM Modelling

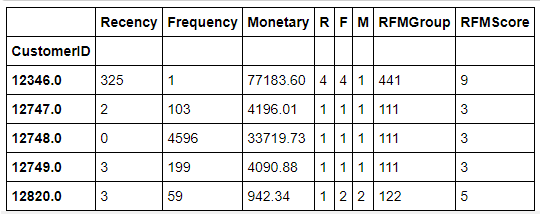
Recency, frequency, monetary value is a marketing analysis tool used to identify a company's or an organization's best customers by using certain measures. The RFM Model is based on three quantitative factors: Recency: How recently a customer has made a purchase. Frequency: How often a customer makes a purchase Monetary: How much do they spend?

Here the renaming of columns was done, Invoice Date, Invoice No and Total Amount were replaced By Recency , Frequency and Monetary. After Descriptive statistics were applied and Distributions were plotted using seaborn library. Here are some previews of table and plot.

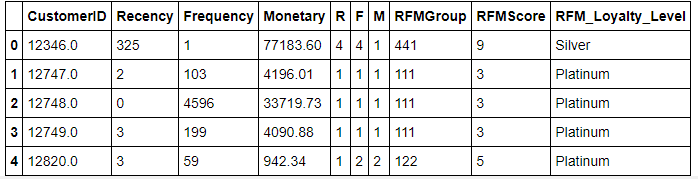




The RFM model were Splitted into 4 segments using quantiles. Then the functions were defined to assign the individual score to each R, F and M , thereby making it easier to further visualization. Now added the RFM Group column to display the Concatenated score of RFM and RFM Score Column to show total sum of RFM Group.

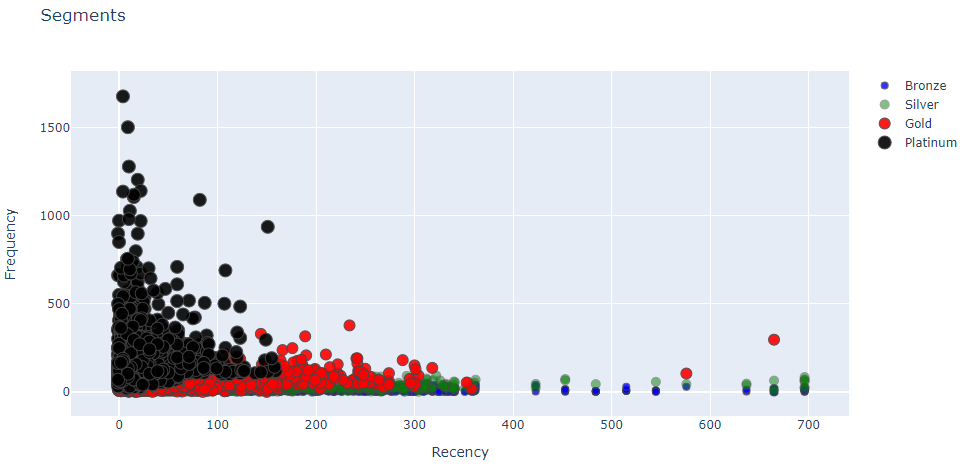


According to the score the customers then were assigned Loyalty Level as Platinum, Gold, Silver and Bronze.

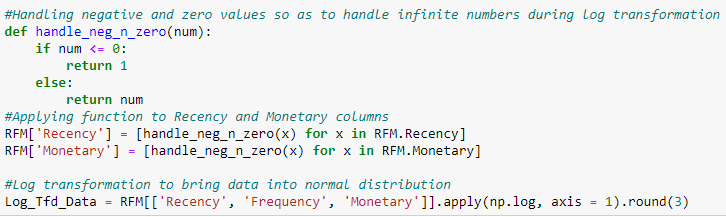


## C. Data Standardization & Normalization

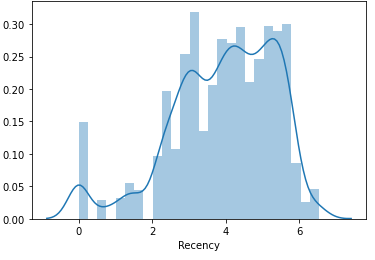
In this stage, the visualizations was done, with the help of plotly library the graphs of Recency vs Frequency, Frequency vs Monetary, Recency vs Monetary were plotted considering the Loyalty Levels such as Platinum, Gold, Silver and Bronze were considered as the data points.



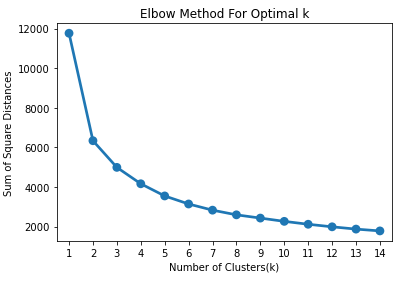
Defining Function to handle negative and 0 values so as to handle infinite numbers in log transformation for normal distributed form.



After the normalisation the Data Distribution for model would be as follows, For Ex. Recency.



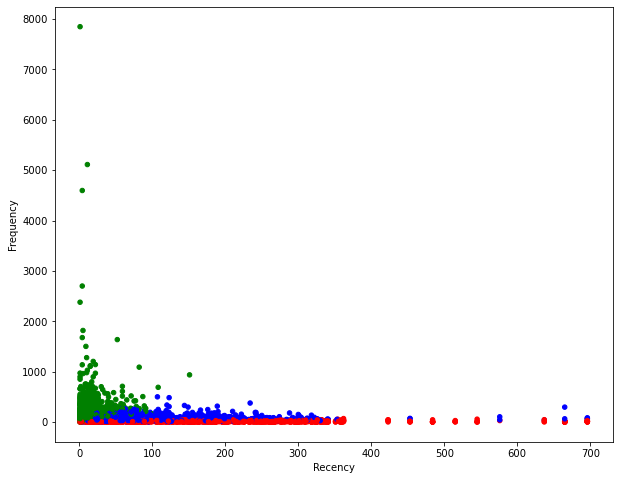
## D. Elbow Method

The elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture finer groups of data objects that are more similar to each other. After the standardization of Data on the same scale, this method was applied.

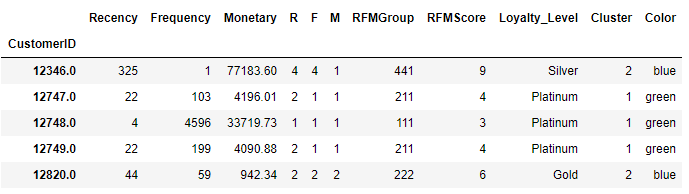
Since, the value is dramatically decreasing at k =3, it is considered as optimal value.

## D. K-Means Clustering

After the Elbow method , we got the value of clusters as 3, then the model creation was started with the defined clusters values and also a new column Clusters was created so as to assign a specific cluster value to each Customer. Now using the matplotlib library function plotting the scatter plot for Frequency vs Recency, assigning the colors Red, Green and Blue to particular clusters values. The plot is as follows,



The Color column was added to specify the values and colours assigned to the plots thereby making the visualization easier to understand. The final dataset preview is as follows,



# IV.CONCLUSION

From the above visualization it can be observed that Cluster 1 which was assigned to platinum group in green color, it represents the Customers which are very recent and have very high frequency of visit and most of the revenue is generated from that group. The Cluster 2 represent both the Gold and Silver in Blue Color because of their Common factors, This group also contribute considerable amount to the income of the company but are not frequent compared to the Cluster 1. The Cluster 0 represents the Bronze level section , the data points are almost flat representing very less frequency and recency and not much revenue is generated here ,this is a concern for the company to look at reason for this section of the Customer.

# V.SWOT ANALYSIS

**Strength:** A good amount of Customer base is on Platinum Level as Compared to the overall Customer base.

**Weakness:** There are also considerable amount of the Customer base at Bronze Level which cannot be ignored as they may cause the impact on the Sustainable Business Model.

**Opportunities:** The data shows there are many Customers visiting but are not buying the products due to some reason, so deep analysis could be done to identify the problem and thereby making more Opportunity to increase Revenue.

**Threat:** As there are always many competitors in the market, The company should be always be aware of the Low Prices and Maintain Good Service they are offering Compared to their Peers.

# REFERENCES

[1] I. S. Dhillon and D. M. Modha, “Concept decompositions for large

sparse text data using clustering,” Machine Learning, vol. 42, issue 1,

pp. 143-175, 2001.

[2] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R.

Silverman, and A. Y. Wu, “An efficient K-means clustering

algorithm,” IEEE Trans. Pattern Analysis and Machine Intelligence,

vol. 24, pp. 881-892, 2002.

[3] MacKay and David, “An Example Inference Task: Clustering,”

Information Theory, Inference and Learning Algorithms, Cambridge

University Press, pp. 284-292, 2003.

[4] Jiawei Han, Micheline Kamber, Jian Pei “Data Mining Concepts and Techniques”, Third Edition.

[5] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, “The Basis Of Market Segmentation”

Euclidean sum-of-squares clustering,” Machine Learning, vol. 75, pp.

# 