St. Xavier's College (Autonomous), Kolkata

Department of Computer Science

CUSTOMER SEGMENTATION

Customer Analysis using
Unsupervised Machine Learning

Domain: Data Analysis and Machine Learning
Sector: Business

Prepared By:
Shria Banerjee (532)

In collaboration with:
Arkapriya Ghosh (523)
&
Anindita Mukherjee (504)

Supervised By:
Prof. Debabrata Datta

Submitted to the Department of Computer Science, in partial fulfillment of the requirements for the degree of Bachelor of Science (BSc.) Honours

Date of Submission: 10th May, 2021

Declaration

This is to declare that the project report entitled as “CUSTOMER SEGMENTATION: Customer Analysis Using Unsupervised Machine Learning” , contributed by myself, **Shria Banerjee [A01-2112-0847-18]** in collaboration with Arkapriya Ghosh and Anindita Mukherjee, under the guidance of Prof. Debabrata Datta, has all resources identified and that no part of this dissertation paper uses unacknowledged materials.

Name: Shria Banerjee

Address:
4, Deshbandhu Road (East),
Baranagar,
Kolkata - 700035

Registration No: A01-2112-0847-18

Roles and Responsibilities

Name of Project: CUSTOMER SEGMENTATION: Customer Analysis Using Unsupervised Machine Learning.

Date of Submission: 10.05.2021

Name	Roles and Responsibility
Shria Banerjee	ML Developer - Clustering Model, Analysis, Documentation, Background Study.
Anindita Mukherjee	Documentation, Literature Review, Background Study, Synopsis.
Arkapriya Ghosh	ML Developer - Clustering Model, Literature Review, Documentation, Background Study.

Acknowledgement

We would like to extend our heartfelt gratitude to our Project Mentor Prof. Debabrata Datta for guiding us throughout the project. We thank all the people who have given us valuable advice and ideas regarding the platform and resources which have been indispensable to the project.

We would like to thank the faculty of the Computer Science Department of St. Xavier's College (Autonomous), Kolkata for providing us this opportunity to work in this project.

1. INTRODUCTION

The evolution of computer technologies over the past few decades has led the whole world to experience the power of the internet to a great extent. The world wide web has allowed vigorous growth of various businesses all over the world. More and more customers tend to prefer online transactions to traditional offline systems as the physical location of the user is not a problem anymore. The continuously flourishing electronic commerce industry necessitates the storage and management of an enormous amount of data that are being generated by millions of online customers everyday. In order to thrive in this competitive market scenario, the available data needs to be readily analysed by the organisations to have a better understanding of customer demands. Data analysis is a significant step towards devising effective marketing and promotional strategies, enhancing budget efficiency, recognising customer interests and providing prompt services and solving marketing problems for a profitable business. Instead of mass marketing, that is assuming the whole population with internet access as potential customers, one should consider target marketing where proper grouping among the customers can give an idea of the percentage of potential customers for their promoted services or products. Therefore the analysis is done by classifying the customer base into groups of similar individuals called segments. This process of dividing customers into segments based on similar characteristics is known as Customer Segmentation.

Customer segmentation, also known as market segmentation, is a marketing approach that aims at mapping different groups of customers with appropriate products and services, in order to encourage them to buy more. Understanding and capability of providing their further needs creates brand dependence, so forth a loyal customer base. The viability and popularity of any company in today's competitive market majorly stands on customer demands, which are determined by particular geo-demographic and socio-economic backgrounds. Market varies from area to area with the characteristics of the population and so do the promotional approaches. It's evident that companies should be versatile in their ways of satisfying separate segments of customers rather than trying to meet every customer's needs in a single product or service. Extensive study on shopping behavior of different customers or a proper feedback system is essential to provide a customised experience to the buyers as well as enhance customer relationships.

Over the years, the analysis of e-commerce data has resulted in the observation that the sale of a particular product not only depends on its price but also on several other factors like the product value and customer satisfaction. To improve customer satisfaction and gain their loyalty, companies are recommended to design market programs targeting each of its customer segments. Exploring customer data like demographic and geographic characteristics, purchase behavior, satisfaction with services, etc., help in understanding and characterising the customer segments in a more efficient manner.

Customer segmentation can be broadly classified into four basic types (depending on which characteristics are targeted for the segmentation): Demographic Segmentation, Psychographic Segmentation, Geographic Segmentation and Behavioral Segmentation. Demographic segmentation deals with statistical characteristics like age, gender, income and profession. Psychographic segmentation works with customers' personalities and interests like hobbies, values and lifestyles. Geographic segmentation focuses on the physical location of customers, their country, region, city and postal code. While behavioral segmentation uses traits like purchasing habits, browsing habits and previous product ratings. All the types of segmentations are combined to facilitate better personalisations for individual customers.

Conventional hard computing techniques of data processing follow mathematical and statistical methodologies which require homogeneous input data to produce a deterministic output. Since real time data is huge and heterogeneous, these algorithms have been quite inefficient for solving complex real world problems like market analysis. On the other hand, soft computing methods like machine learning and Big Data analytics have proven to be effective and practically applicable for achieving an approximate solution to these problems with optimal usage of resources. The soft computing approach aims at automating the process of humans' logical thinking and the capability of learning from mistakes in an imprecise situation. One of the most typical ways of approaching a segmentation problem is the cluster analysis. Therefore the most extensively used soft computing algorithms with respect to customer segmentation are the clustering algorithms. Clustering algorithms are unsupervised machine learning techniques that divide datasets into unlebelled groups or clusters consisting of similar data-points. Clusters are computed based on similar characteristics or attribute values. Algorithms like K-means Clustering (hard clustering), Fuzzy C-means Clustering (soft clustering) and Hierarchical Clustering are collectively implemented to perform customer analysis.

2. LITERATURE REVIEW

2.1. Historical Overview

The concept of segmentation emerged as a formal component of contemporary marketing practice in the 1950's [1]. Wedel and Kamakura (2001) clarify that "Since the concept emerged in the late 1950's, segmentation has been one of the most researched topics in marketing literature" (Wedel and Kamakura, 2001, p. xix). Likewise, contemporary marketers report that segmentation represents an integral part of contemporary marketing practice. It is currently recognized that "Market Segmentation is an essential element of marketing in industrialized countries. Goods can no longer be produced and sold without considering customer needs and recognizing the heterogeneity of those needs" (Wedel and Kamakura, 2001, p. 3).

This increased literary focus has been predicated on a shift in marketing practice aligned with the diversification of industrial production (Wedel and Kamakura 2001). During the early twentieth century as production efficiency became enhanced and product variation increased, the concept of market segmentation became a formal component of marketing practice "industrial development in various sectors of the economy induced strategies of mass production and marketing. Those strategies were manufacturing oriented, focusing on reduction of production costs rather than satisfaction of consumers. But as production processes became more flexible, and consumer affluence led to the diversification of demand, firms that identified the specific needs of groups of customers were able to develop the right offer for one or more sub-markets and thus obtained a competitive advantage" (Wedel and Kamakura, 2001, p. 3).

Chamberlin (1933) laid the foundation for the prioritization of the consumer over the producer by pointing to the significance of aligning products with the needs and wants of consumers. Later in this decade, Robinson (1938) expanded this concept and formalized the economic theory of imperfect competition (Robinson 1938). The work of these two scholars set the stage for Smith's influential work in the 1950's. In 1956, he recognized "the existence of heterogeneity in the demand of goods and services, based on the economic theory of imperfect competition" (Wedel and Kamakura, 2000, p. 3) developed by Robinson in the late 1930's (Robinson 1938). Smith asserted that "Market segmentation involves viewing a heterogeneous market as a number of smaller homogeneous markets in response to differing preferences, attributable to the desires of consumers for more precise satisfaction of their varying wants" (Smith 1956 p 6).

Smith's decisive article from 1956 asserted that "segments should be based on consumer/user wants and a company should be better able to serve these needs when it has defined some segments within a larger market" (Anna-Lena 2001). Consumer segments were defined in Wind and Cardozo's seminal article "Industrial Market Segmentation" (1974) as "a group of present and potential customers with some common characteristic (s) which is relevant in explaining (and *predicting*) their response to a supplier's marketing stimuli" (Wind and Cardozo, 1974). Therefore, consumer segments should clarify groups of current and previous consumers while serving as predictors identifying the most likely candidates for future consumers. The predictive value of consumer segments should therefore have a profound impact on strategic Internet Marketing plans taking cognizance of matching promotional content with potential consumers.

This process undertaken to clarify the most receptive groups within a population with the most relevant messaging has been formalized in the literature as consumer segmentation (Frank, Massy, Wind 1972, McDonald & Dunbar 2004, Anna-Lena 2001, Jiang and Tuzhilin 2006). In 1974, Wind and Cardozo recognized that consumer segmentation (referred to then as market segmentation) “involves appropriate grouping of individual customers into a manageable and efficient (in a cost/benefit sense) number of market segments, for each of which a different marketing strategy is feasible and likely profitable” (Wind and Cardozo 1974 p. 155). A host of marketers since have advocated establishing consumer segments as a means to more closely align products and services with targeted groups (2006 p. 307). Therefore, the appropriate number of segments for the sole target of business has always been particular promotion are contextual and are determined based on the variation within a given population of potential and return customers.

The process of consumer segmentation rests on three primary assumptions: 1) the population of potential and return consumers is heterogeneous 2) heterogeneous groups have distinct characteristics that can be identified and analyzed 3) unique promotional content can cater to the varying needs, wants of consumers and perceived benefits of specific products and services.

Contemporary segmentation approaches recognize six primary criteria applied toward the evaluation of segmentation effectiveness (Kotler and Armstrong 2007, Wedel and Kamakura 2000, Anna-Lena 2001). These criteria measure effectiveness by evaluating segment formation and profitability and assert that consumer segments should be: identifiable/measurable, substantial, accessible, stable, actionable, and differentiable (Anna-Lena 2001 p. 5).[1]

The sole target of business has always been the customer demands, starting from local stores to giant industries.Unlike local stores where the customer demands can be physically encountered, all country wide businesses and today's online market, the only alternative in the form of customers purchasing data. Customers from different social economic geography demographic and cultural contexts have their own form of demands which may not match everywhere. Even people of different ages have different tastes and choices . Hence diversifying the services of targeting different groups of customers can be immensely advantageous in their way of profit making.

Alternatives additional to the previous are, well formed customer service , and effective interactions with the customers, a well monitored feedback system can add up to the producer-consumer relationship. Again, interaction with customers is direct at any customer care desk of a supermarket or mall or customer care system of any company, via phone call. But remote interaction through electronic mails or messages and delivering proper promotional offers for a proper set of customers requires the knowledge of their purchasing pattern. No one likes to be troubled by unnecessary advertisements or product suggestions out of the scope of our interests. A proper knowledge of customer demands helps the company to take prompt steps to attract customers without unnecessary talks.

2.2. Previous Works on Customer Segmentation

Every customer is different by some perspective or other and so are their tastes and choices [3]. Grouping or segmentation among them would require consideration of some factors based on which we can differentiate or draw analogy. Some most common and popular factors into consideration are -

1. Geographic i.e. country,town or city of residence, even locality of town or city, people's lifestyle, distribution of market, availability of products etc.
2. Demographic which considers gender, age, occupation and income, marital status etc.
3. Psycographic segmentation includes customers' personal traits,choices, interests, values etc.

4. Behavioral segmentation encompasses consumers spending habits, frequency, mostly bought products etc.

Some software companies often use the term 'technographic segmentation' to characterise customers' technological preferences and abilities.

2.2.1. RFM Models And CRM:

With the growth of the e-commerce industry and advancements in internet technologies, the marketers have realised that just broadcasting the promotional advertisements is not enough as a strategy to increase the profit for next quarter. The maturity of marketers in online business has led to increasing adoption of customer relationship management (CRM)(Cheng and Chen 2009).CRM refers to the procedures companies use to gain lifetime customer loyalty and thereby increase competitive advantage and profits. The key idea for CRM is"the right time, the right channel, the right price, and the right customers."[4] A successful web-based CRM can essentially strengthen a company's virtual interactions with its customers and provide a way to generate more revenue by developing an e-commerce strategy and redesigning its web pages.

(In [3])Pareto's 80|20 rules can give a better understanding of this picture. In 1896, Vilfredo Pareto showed us that 80% of the land in Italy was owned by just 20% of the people. It was soon realized that this 80–20 imbalance comes up a lot in the world.

- The richest 20% of humans have 82.7% of the world's income.
- 20% of words account for 80% of all words spoken
- 20% of customers are responsible for 80% of profits
- 80% of the complaints received by a business come from 20% of the customers.

(In [3])Accordingly, a small percentage of customers contribute to the major portion of the revenue(Bult and Wansbeek 1995,Stone and Jacobs 2001). Thus it is better to retain the set of customers who have spent the most, or have stayed with the company for the longest, than to run for new customers.Now the question is how to retain them?This requires knowledge on customers buying behaviour and thus choosing proper strategies to attract their interest.The most commonly-used strategies are one-to-one marketing and the recommendation system (Jiang and Tuzhilin 2006)[4]. Most of the successful companies have relied on web pages thoroughly personalized according to the interest of customers and tailored to a particular segment of market.

The RFM model stands for recency, frequency and monetary investment of shopping. This is the most popular model in analysis of customer behaviour. The RFM model segments customers based on their recency of shopping or visiting their website, time of searching about a product, frequency of buying the same product or products bought together and the money spent by them in shopping.

RFM model defines a scoring system in which each customer is scored accordingly as the mentioned three factors, and these scores are used to assign them to segments.In empirical research by Hughes(1996), customer records in a database were generally divided into five equal quintiles for each of the three RFM characteristics.[4]

However, this quintile method creates potential problems, because the arbitrary cutoffs are applied to the three RFM characteristics rather than to the customers (Migkautsch 2000). Using the means of continuous distributions of customer scores rather than quintiles not only yields greater sensitivity at both the top and bottom of the distribution, but it also isolates single customers.[3]

Another shortcoming of the RFM model is that, because of its three-dimensional nature, its predictive capacity is inferior to that of more sophisticated methods such as automatic interaction detection and regression analysis (McCarty and Hastak 2007).[4]

2.2.2. Segmentation of Online Customers:

Applications of machine learning, like regression and classification, focus on predicting the outcome as value of an instance. These types of algorithms can’t identify the similarity between the instances. Thus segmentation turns the focus towards supervised and unsupervised learning.

The determination of whether the algorithm used falls under supervised or unsupervised learning is contingent upon the following-

The data used as training instances should be paired with the target value, which may be scalar or vector. In contrast unsupervised learning algorithms are fed with data instances that are not paired with the target values.

For example , if a retail store owner ever thinks to analyse the customer’s spending behavior form a year's data to predict how much will they spend on their next visit, then the required methodologies will of supervised learning, as each instance has the spending record associated with it and based on that analysis is done.[5]

Now if the same seller decides to analyse customer data to understand differences and similarities in their demands then this will not have any particular result associated with the instances. Rather his has no clear cut target value. This exemplifies Unsupervised learning.[5]

Computational capabilities for data storage and processing knew no bounds from the last century. Similar for the advancement of data mining techniques, which has enabled extraction of hidden and predictive information from large databases.

In supervised classification methods (Vellido et al. 1999) such as neural networks, linear discriminant analysis, and decision-tree induction, the available observations or samples have class labels. The aim is to construct a model that assigns one of these class labels to each new observation.(from [4])

In case of unsupervised learning the outcome is out of the scope of the available classes. Some applications of unsupervised learning are as follows.(From [4])

1. Partition-based clustering algorithms group observations that are close to one another in distance (Bose and Chen 2010).
2. Model-based clustering approaches estimate membership probabilities for the purpose of assigning observations to the appropriate clusters.
3. Latent class clustering, also referred to as finite mixture model clustering, was designed for the analysis of grouped categorical data (Magidson and Vermunt 2002). Latent classes are unobservable subgroups or segments.
4. Finite mixture models result in overlapping classes but each observation is considered to belong to a discrete class. Real time problems often result in non-exclusive memberships.
5. Fuzzy clustering enables the use of non-numerical attributes and has a single object belonging to several classes simultaneously, thus customer relation assessments get more refined.

2.2.3. E-Commerce Data:

Reference: Customer Segmentation | Kaggle

Customer Segmentation by F. Daniel(September 2016)

Aim: Analyzing the content of an E-commerce database that lists purchases made by around 4000 customers over a period of one year (from 2010/12/01 to 2011/12/09) and develop a model that allows to anticipate the purchases that will be made by a new customer, during the following year and this, from its first purchase based on this analysis.

The dataframe is of dimension (541908,8) having columns-

- **Invoice No:** Invoice number. A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
- **Stock Code:** Product code. A 5-digit integral number uniquely assigned to each distinct product.
- **Description:** Product name.
- **Quantity:** Quantity of each product per transaction.
- **Invoice Date:** Datetime of purchase.
- **Unit Price:** Price of a unit product.
- **Customer ID:** Unique number for each customer.
- **Country:** The country from which the customer belongs.

In the dataframe, products are uniquely identified through the StockCode variable. The Description variable is used to group the products into 5 different categories using the k-means method of sklearn that makes use of the Euclidean distance. In a second step, the customers were analysed by their consumption habits over a period of 10 months into 11 major categories based on the type of products they usually buy, the number of visits they make and the amount they spent during the 10 months. For this, the classifier is based on 5 variables which are:

mean : amount of the basket of the current purchase.

categ_N with N∈[0:4] : percentage spent in product category with index N.

Finally, the quality of the predictions of the different classifiers was tested over the last two months of the dataset. The data were then processed in two steps: first, all the data was considered (over the 2 months) to define the category to which each client belongs, and then, the classifier predictions were compared with this category assignment. 75% of clients are awarded the right classes.
Shortcomings of the current model concerns the seasonality of purchases and the fact that purchasing habits will potentially depend on the time of year (for example, Christmas). In order to correct such bias, it would be beneficial to have data that would cover a longer period of time.

2.2.4. Bank Customer Segmentation:

Reference: Bank customer segmentation | Kaggle

The dataframe contains columns as follows:

- Age
- Sex
- Job
- Housing
- Savings Accounts
- Checking Account
- Credit Amount
- Duration
- Purpose

Clustering with Affinity Propagation algorithm is used:
In this algorithm there are two relevant parameters: preference and dumping. It means that we don't define the upfront number of clusters, the algorithm itself chooses their number.Together with decreasing value of preference parameter number of clusters goes down as well and levels for very small preference values.

Result:

Cluster 0 – high mean of credit amount, long duration, younger customers.
Cluster 1 – low mean of credit amount, short duration, younger customers.
Cluster 2 - low mean of credit amount, short duration, older customers.
Cluster 3 - high mean of credit amount, middle-time duration, older customers.

2.2.5. Market Basket Analysis:

Link: https://www.kaggle.com/mgmarques/customer-segmentation-and-market-basket-analysis

The Online Retail a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail is used to explore customer segmentation through the interesting task of unsupervised learning method.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.Then association rule mining approach is used to find interesting rules and patterns in this transaction database.

Market basket analysis is a method to gain insights into granular behavior of customers. This is helpful in devising strategies which uncovers deeper understanding of purchase decisions taken by the customers.

The description of each column:

Invoice No: A unique identifier for the invoice. An invoice number shared across rows means that those transactions were performed in a single invoice (multiple purchases).

- **Stock Code:** Identifier for items contained in an invoice.
- **Description:** Textual description of each of the stock items.
- **Quantity:** The quantity of the item purchased.
- **Invoice Date:** Date of purchase.
- **Unit Price:** Value of each item.
- **Customer ID:** Identifier for customer making the purchase.
- **Country:** Country of customer.

Result with k=5:

- Cluster 0 presents the fourth best purchase and a reasonable frequency, but this is a long time without buying. This group should be sensible to promotions and activations, so that they do not get lost and make their next purchase.
- The cluster 1 appears more robust on the affirmation of those who shop often and with high amounts.
- The cluster 2 are those who have a decent spend but are not as frequent as the cluster 1
- The cluster 3 makes low-cost purchases, with a relatively low frequency, but above 1, and made their last purchase more recently. This group of customers probably respond to price discounts and can be subject to loyalty promotions to try to increase the medium-ticket strategy that can be better defined when analyzing the market basket.

- The cluster 4 purchases medium amounts, with a relatively low frequency and not very recent
- Cluster 5 is similar to 0, but has made its purchases more recently and has a slightly better periodicity. Then actions must be taken to raise their frequency and reduce the chances of them migrating to cluster 0 by staying longer without purchasing products.

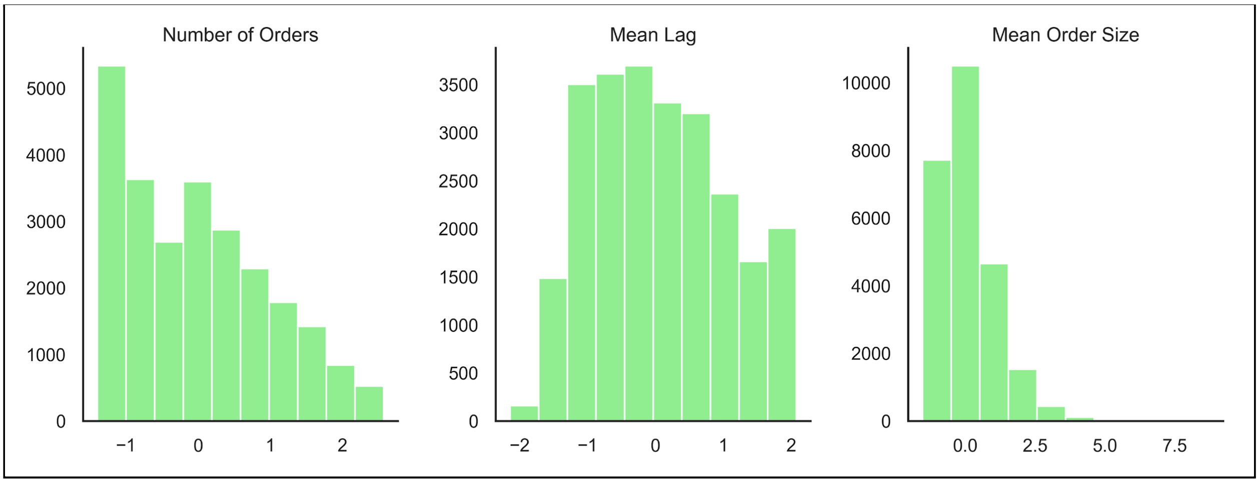
2.2.6. Customer segmentation using the Instacart dataset:

Link:https://towardsdatascience.com/customer-segmentation-using-the-instacart-dataset-17e24be9c0fe by JR Kreiger

Instacart is an grocery shopping service through an app.The aim was to try to predict which items a customer would order again in the future.The dataset contains the following types of information:

- A record for every order placed, including the day of week and hour of day (but no actual timestamp);
- A record of every product in every order, along with the sequence in which each item was added to a given order, and an indication of whether the item had been ordered previously by the same customer; and
- The name, aisle, and department of every product.

They made use of elbow plots using the silhouette score for various numbers of clusters and for producing snake plots to summarize the attributes of each cluster.



Here number of orders signify frequency,mean lag signifies recency and mean order size can refer to the monetary value of the commodities purchased.

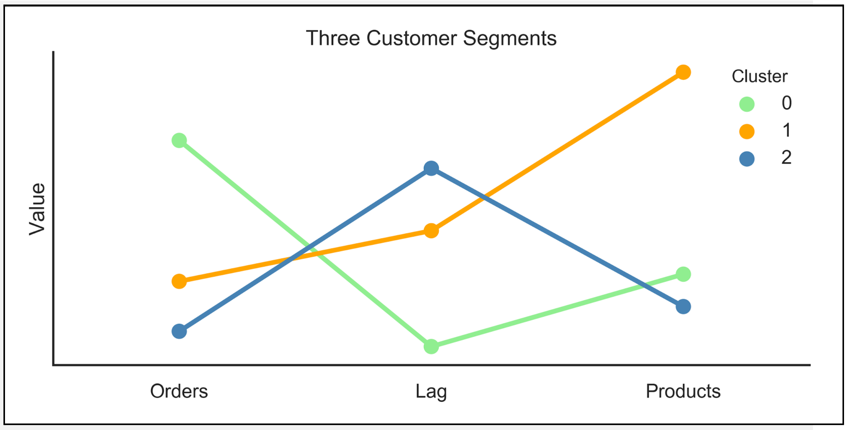
Since clustering is easily influenced by both non-normal distributions and outliers, log transformation and sklearn.preprocessing.StandardScaler were used. There are several metrics that can be used to evaluate how well *k* clusters fit a given dataset like: **distortion score** and **silhouette score**.

Distortion score is kind of like the residual sum of squares; it measures the error within a cluster, or the distance between each datapoint and the centroid of its assigned cluster. A lower distortion score means a tighter cluster, which means the customers in that cluster would have a lot in common.

Silhouette score compares the distance between any given datapoint and the center of its assigned cluster to the distance between that datapoint and the centers of other clusters. Basically, silhouette score is asking, “Is this point actually closer to the center of some other cluster?” A low value indicates our clusters are tighter and also farther from each other in the vector space. TSNE plots take everything we know about each customer and reduce that to just two dimensions so that we can easily see how clusters relate to one another.

When there are only 3 clusters, they look pretty easily separable (and also fairly evenly balanced — no one cluster is much bigger than the rest).

A “snake plot” (a Seaborn pointplot) was used to visualize the average value of each of the three features for each cluster.



- Cluster 0: These customers use Instacart a lot and make medium-sized orders. Marketing for these customers could focus on maintaining their loyalty while encouraging them to place orders that bring in more revenue for the company (whether that means more items, more expensive items, etc.).
- Cluster 1: These customers don’t use Instacart as often, but when they do, they place big orders. Of course we can focus on turning them into more frequent users, and depending on exactly how Instacart generates revenue from orders, we might nudge them to make more frequent, smaller orders, or keep making those big orders.
- Cluster 2: This is the segment where we have the most room for improvement. They have tried Instacart, but they don’t use it often, and they don’t purchase many items. A marketing strategy for these folks could focus on increasing order frequency, size, or both

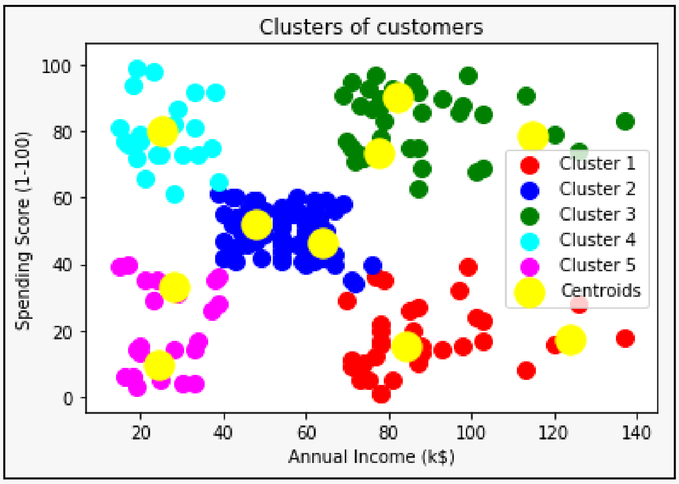
2.2.7. Segmentation of Mall customers:

Reference: https://www.kaggle.com/vjchoudhary7/kmeans-clustering-in-customer-segmentation

The dataset has the following columns:

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)

Considering only 2 features (Annual income and Spending Score).The Elbow method is used to find the optimal number of clusters in KMeans Clustering.k is taken as 5.



Model Interpretation :

Cluster 1 (Red Color) -> earning high but spending less
Cluster 2 (Blue Color) -> average in terms of earning and spending
Cluster 3 (Green Color) -> earning high and also spending high [TARGET SET]
Cluster 4 (cyan Color) -> earning less but spending more
Cluster 5 (magenta Color) -> Earning less , spending less

We can put Cluster 3 into some alerting systems where email can be sent to them on a daily basis whereas others we can set like once in a week or once in a month.

Clustering:

Is an unsupervised Machine Learning method which involves collecting as much data about the customers as possible in the form of features or attributes and then finding out the different clusters that can be obtained from that data. Finally, we can find traits of customer segments by analyzing the characteristics of the clusters.

Exploratory Data Analysis:
This is usually done by analysts who have a good knowledge about the domain relevant to both products and customers. It can be done flexibly to include the top decision points in an analysis.

RFM Model for Customer Value:
RFM (Recency, Frequency and Monetary Value) based model of customer value is used for finding our customer segments. 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.

The K-means clustering algorithm is used. 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 0 and variance of 1. 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. This is akin to feature scaling.The elbow method is used to find the optimal number of clusters.

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. This rule-based approach also generates new rules as it analyzes more data. The ultimate goal, assuming a large enough dataset, is to help a machine mimic the human brain's feature extraction and abstract association capabilities from new uncategorized data.Apriori, Eclat or FP-Growth algorithms are used.

3. K - MEANS CLUSTERING ALGORITHM

Clustering is one of the most prevalent methods for automated data analysis. To get insights on the underlying nature and structure of data the main purpose of clustering is to organise a set of data into clusters, such that the elements in each cluster are similar and different from those in other clusters. One of the most used clustering algorithms is K-means, because of its ease of implementation and interpretation [6].

K-means Clustering is an unsupervised machine learning algorithm that divides n observations into k groups or clusters based on similar characteristics or attributes. It is an exclusive (hard) clustering technique, hence a particular observation or data point can be assigned to only one cluster.

3.1. History

The concept of the k-means algorithm was first proposed by Hugo Steinhaus in 1956. In 1957, the standard algorithm was proposed by Stuart Lloyd of Bell Labs as a technique for pulse-code modulation. Later in 1967, the term "*k*-means" was coined by James MacQueen. In 1965, Edward W. Forgy published the same method, which is why it is sometimes referred to as the Lloyd–Forgy algorithm.[7]

3.2. Aim of K-means algorithm

Given a set of n observations (x_1, x_2, ..., x_n), where each observation is a d-dimensional
(d≥1) real vector, k-means clustering aims to partition the n observations into k ($2 \leq k \leq n$) sets
S = {S_1, S_2, ..., S_k}, such that the total sum of variance of k clusters is minimised.[7]
Variance is the degree of spread of data in a dataset with respect to the centroid (mean) of a dataset. Therefore greater the distance of the data points from the centroid greater is the variance. To minimise the variance, each data point must be assigned to a cluster whose centroid is at the nearest distance from the data point. Variance is calculated as the sum of squared Euclidean distance between the data points in a dataset and it's centroid. Any optimal solution of this algorithm will result in the generation of k number of means (centroids) , hence the method was named *k*-means algorithm.

In a two-dimensional space (i.e., a Cartesian plane), if two data points are
(x_1, y_1) and (x_2, y_2) then,

$$Euclidean\ Distance\ \delta_{1,2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad (1)$$

The general formula for calculating the distance between two objects
$(x_1, x_2, ..., x_n)$ and $(y_1, y_2, ..., y_n)$ with n attributes is,

$$\delta_{1,2}^2 = \sum_{i=1}^{n} (x_i - y_i)^2 \qquad (2)$$

3.3. Mathematical Model

In [6] the k-means algorithm is defined as an iterative method that consists of partitioning a set of n objects into k ($2 \leq k \leq n$) clusters, such that the objects in a cluster are similar to each other and are different from those in other clusters.

Let $N = \{x_1, x_2, ..., x_n\}$ be the set of n objects to be clustered by a similarity criterion, where $x_i \in \Re^d$ for $i = 1,\dots,n$ and $d \geq 1$ is the number of dimensions.

Let $k \geq 2$ be an integer. For a k-partition, $P = \{G(1), G(2),\dots, G(k)\}$ of N,let μ_j denote the centroid of cluster G(j), for $j \in k$, and let $M = \{\mu_1,\dots,\mu_k\}$ and $W=\{w_{11},\dots,w_{ij}\}$.

Therefore, the clustering problem can be formulated as an optimization problem whose objective function is described by Eq. (3):

$$P: \mathit{minimize}\ Z(M, W) = \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij}\, d(x_i, \mu_j)^2 \qquad (3)$$

$$\mathit{subject\ to}:$$

$$\sum_{j=1}^{k} w_{ij} = 1\ \mathit{for}\ i = 1,\dots\dots, n,$$

where $w_{ij}=1$ if object x_i belongs to cluster G(j) or 0 otherwise for $i = 1,...,n$ and $j = 1,...,k$ and $d(x_i,\mu_j)$ denotes the Euclidean distance between x_i and μ_j for $i=1,\dots,n$ and $j=1,\dots,k$.

3.4. Complexity Analysis

The classical k-means clustering problem for n observations in d dimensions is:

- NP-hard in general d-dimensional Euclidean space, even for k = 2 [9]
- NP-hard for a general number of clusters k in the plane [10][11]

Thus, a variety of heuristic algorithms are used to improve the hardness of the problem. These algorithms converge to a local optimum but do not guarantee a global optimum. Some examples of such algorithms are Lloyd's algorithm (Standard K-means algorithm) and Hartigan–Wong method.

The running time of Lloyd's algorithm (and most variants) is O(nkdi)[13] where:

- n is the number of d-dimensional vectors (to be clustered)
- k the number of clusters
- i the number of iterations needed until convergence.

In practice, i is usually O(n), therefore Lloyd's algorithm is often considered to be of polynomial time complexity. Although in the worst-case, Lloyd's algorithm needs $i=2^{\Omega(\sqrt{n})}$ iterations, so the worst-case time complexity of Lloyd's algorithm is exponential [12].

3.5. Lloyd’s Algorithm - Standard K-means Algorithm

The most common k-means algorithm considers the following steps [6][8]:

Step 1: The number of clusters k and the centroid of each cluster are initialised manually by the user.

Step 2: The Euclidean distances from x_i to the k centroids are calculated.

Step 3: x_i is assigned to the nearest centroid.

Step 4: The centroid of the cluster to which x_i is assigned, is recomputed.

Step 5: Steps 2, 3 and 4 are repeated for all $x_i \in N$ where $1 \leq i \leq n$.

Step 6: Convergence condition is checked:

- If the centroids remain unchanged in two consecutive iterations then stop the process.
- Stop the algorithm if a given number of iterations by the user is attained.

Else go to Step 2.

Step 7: End.

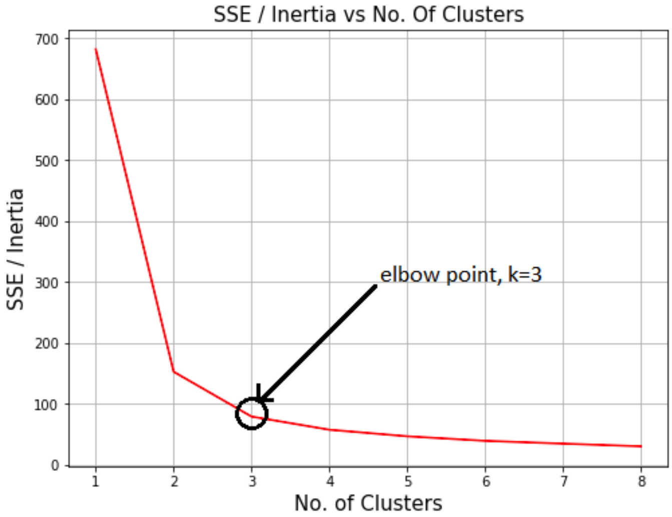
3.6. Determination of the Value of ‘K’

The optimal value of the number of clusters ‘k’ can be determined by a method called the elbow method.

In this method the sum of squared error (SSE) is computed for some values of k (eg: 2,3,4,5,etc.) The SSE or inertia is defined as the sum of the squared distances between each member of a cluster and it’s centroid.

$$SSE = \sum_{i=1}^{k} \sum_{x \in G(i)} d(x, \mu_i)^2$$

where μ_i is the centroid of the cluster G(i) and $d(x, \mu_i)^2$ is the euclidean distance between x and μ_i.



The SSE is plotted against the no. of clusters k to obtain the graph beside.

The graph[1] shows the elbow point at k=3 for which the change in SSE first starts to diminish, indicating the optimum value of ‘k’ to be 3.

[1] Image taken from the internet.

3.7. Advantages

The advantages of k-means clustering are as follows:-

- Easy to understand and implement, without the need of complex statistics.
- Interpretation of the clustering results is easy.
- Fast and efficient practically.
- Scalable.
- Convergence is guaranteed.

3.8. Disadvantages

The disadvantages of k-means clustering are as follows:-

- Depends on initial values and ‘k’ must be defined manually.
- Performs poorly in case of clusters of varying shapes, sizes and density.
- Unable to handle noisy data and outliers efficiently which may affect the overall result.

3.9. Variations of K-means Algorithm

In [7] the following variations of the k-means algorithm are mentioned:-

- *k*-medians clustering uses the median in each dimension instead of the mean, and this way minimizes L_1 norm (Taxicab geometry).
- *k*-medoids (also: Partitioning Around Medoids, PAM) uses the medoid instead of the mean, and this way minimizes the sum of distances for *arbitrary* distance functions.
- Fuzzy C-Means Clustering is a soft version of *k*-means, where each data point has a fuzzy degree of belonging to each cluster.
- Gaussian mixture models trained with expectation-maximization algorithms (EM algorithm) maintain probabilistic assignments to clusters, instead of deterministic assignments, and multivariate Gaussian distributions instead of means.
- *k*-means++ chooses initial centers in a way that gives a provable upper bound on the WCSS objective.
- The filtering algorithm uses kd-trees to speed up each *k*-means step.
- Some methods attempt to speed up each *k*-means step using the triangle inequality.
- Escape local optima by swapping points between clusters.
- The Spherical *k*-means clustering algorithm is suitable for textual data.
- Hierarchical variants such as Bisecting *k*-means, X-means clustering and G-means clustering repeatedly split clusters to build a hierarchy, and can also try to automatically determine the optimal number of clusters in a dataset.
- Internal cluster evaluation measures such as cluster silhouette can be helpful at determining the number of clusters.
- Minkowski weighted *k*-means automatically calculates cluster specific feature weights, supporting the intuitive idea that a feature may have different degrees of relevance at different features. These weights can also be used to re-scale a given data set, increasing the likelihood of a cluster validity index to be optimized at the expected number of clusters

4. METHODOLOGY[2]

As it is discussed before that a typical way of approaching customer segmentation is by RFM analysis, which is the analysis based on the answers to some simple questions like-

- What was the last purchase date ?
- How often does the customer purchase or pay a visit?
- Total bill amount or how much money is invested.

These questions pretty much explain the three factors -Recency, Frequency and Monetary investment respectively for each customer which is abbreviated as RFM. Clustering is done based on these three factors only and the nature of clusters are analysed to find the target customers.

4.1. Data Set Information

The dataset chosen for this project is available at UCI ML Repository as “Online Retail Data Set”.This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.[14]

Attribute Information [14]:

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

Size of the dataset: 541909 rows × 8 columns.

4.2. Cleaning the data

Real time data contains many missing value tuples, which are hard to work with. So these tuples need to be removed. This is the very first step in data cleaning. Again from the description of the dataset it is found that there are rows with ‘Unit Price’ and ‘Quantity’ with negative values, maybe indicating some cancelled orders, or return transactions. These were also removed. For convenience of handling ‘Customer ID’s are converted to int type and the columns ‘InvoiceNo’, ‘StockCode’, ‘Description’ and ‘Country’ are removed as they would not be used in this project. After cleaning the resulting dataframe has 397884 rows and 4 columns.

[2] All images in this section are taken from the internet.

4.3. Sampling the data

The resulting data frame after cleaning is still too large to be handled. Therefore, a randomised sample of 10,000 tuples is considered and a column named ‘TotalPrice’ is added to the dataframe where ‘TotalPrice’ = ‘Quantity’ x ‘UnitPrice’. ‘Invoice Date’ is modified from date-time type to date for further convenience.

4.4. Creating RFM value or score

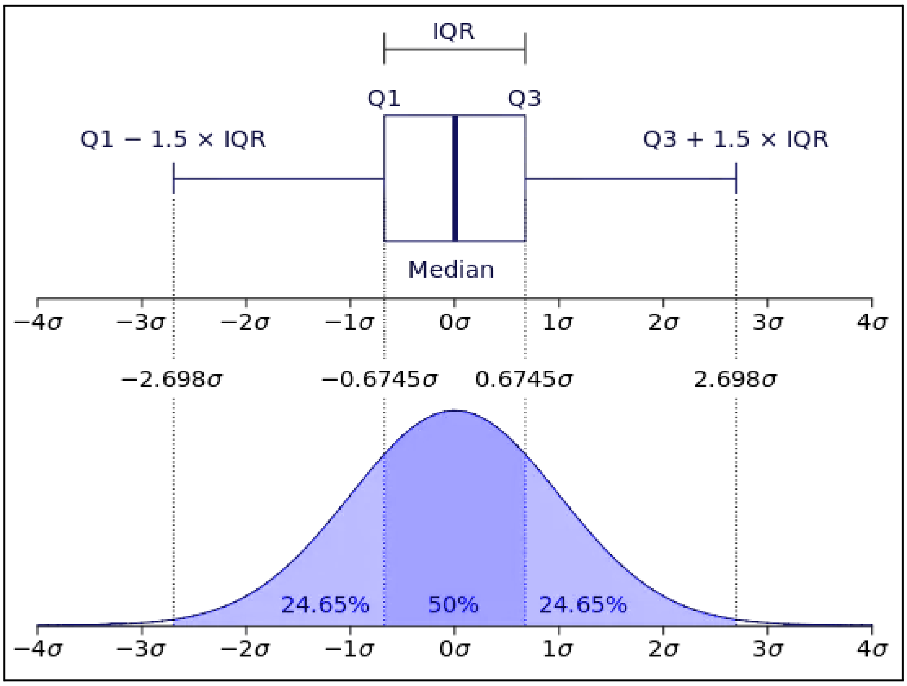
To implement the RFM model individual customers are alloted scores for each of the three variables taken as - ’Recency’,’Frequency’ and ‘Monetary’. ’Recency’ score of a customer is the difference of the most recent ‘InvoiceDate’ of the dataframe and the most recent ‘InvoiceDate’ of that customer, ‘Frequency’ is the number of times each ‘CustomerID’ is found in the dataframe and ‘Monetary’ is assigned the sum of ‘TotalPrice’ against each ‘CustomerID’. These values are stored in the “rfm” dataframe.

4.5. Analysis of the data extracted after RFM Modeling

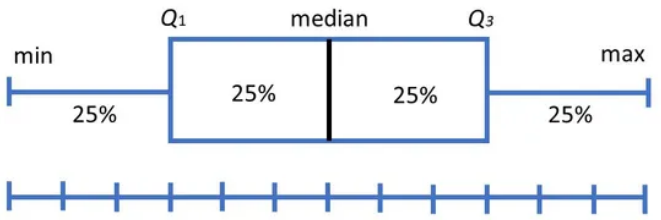
The description of the “rfm” dataframe is used to get an idea about the dataframe. After this different types of plots are used to analyse the nature and distribution of data.

4.5.1. Box plot:

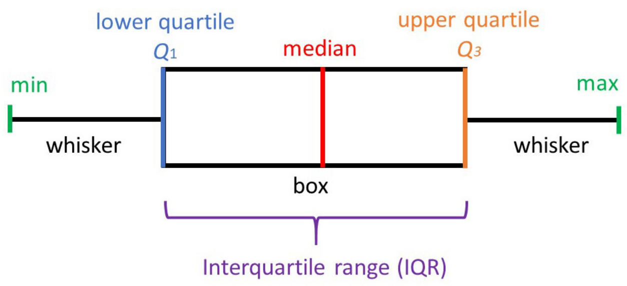
Box plots [15] visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages. Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median,third (upper) quartile, and maximum score.



Correspondence of Box Plot with Normal Probability Distribution

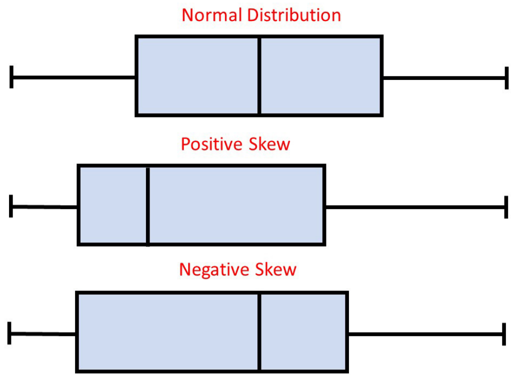


Data distribution in Box Plot

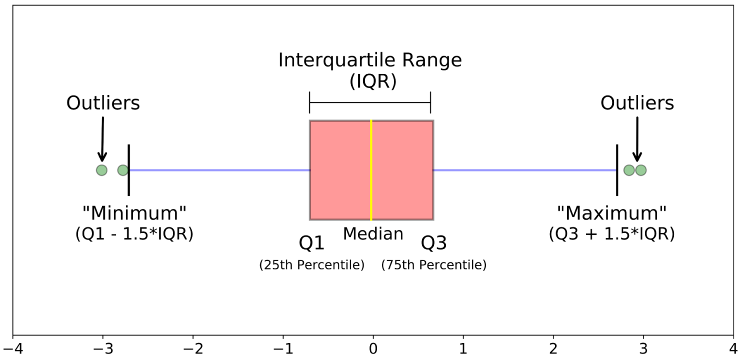


Structure of a Box Plot

Box plots divide the data into sections that each contain approximately 25% of the data in that set. Box plots show the average score of a data set by the median.The position of the median in the box plot, gives the idea of the skewness of a data set.The dispersion of a data set i.e.the extent to which a distribution is stretched or squeezed. The minimum and maximum value is found at the two ends of the whiskers. The interquartile range (IQR) is the box plot showing the middle 50% of scores and can be calculated by subtracting the lower quartile from the upper quartile. Outliers are the observations that are numerically distant from the rest of the data. In Box plots, the outliers within a data set are located outside of the whiskers. Box plot of the “rfm” dataframe shows the range of the values of ‘Recency’ ,’Frequency’ and ’Monetary’.



Normal and skewed distribution

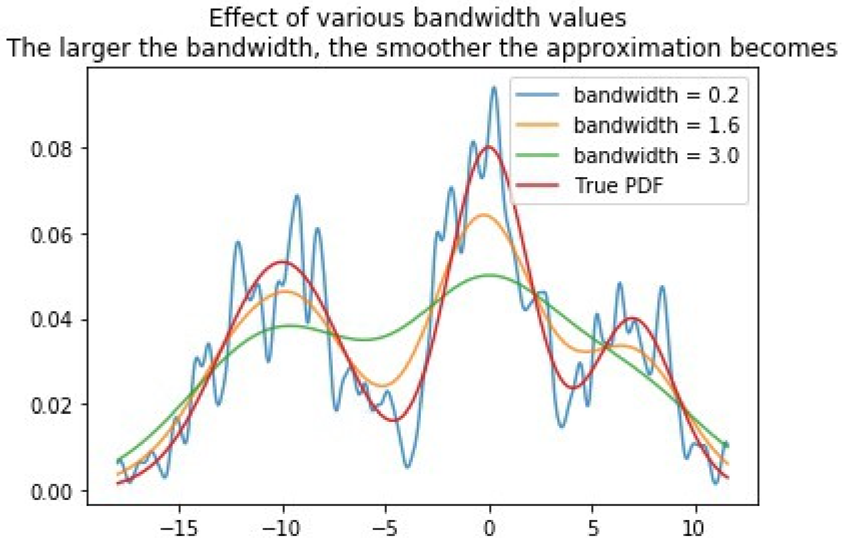


Outliers in a Box plot

4.5.2. Kernel Density Estimation[28]:

The Kernel Density Estimation (KDE) is a mathematical process of finding an estimated probability density function of a random variable. The estimation can also be used to generate points that only appear to have come from a specific sample set. While a histogram counts the number of data points in somewhat arbitrary regions, a **kernel density estimate** is a function **defined** as the sum of a **kernel** function on every data point.

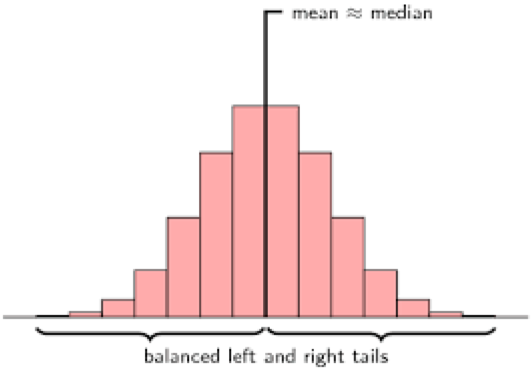
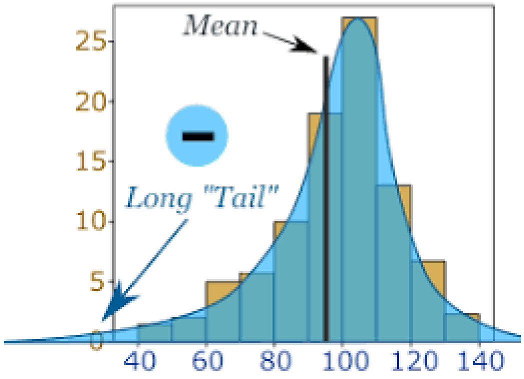
KDE works by plotting the data points and creating a probabilistic curve. The curve is calculated by weighing it’s distance from the data points.If there are more points grouped locally, the estimation is higher as the probability of seeing a point at that location increases. The shape of the curve depends on the bandwidth of the kernels. Lower bandwidth limits the scope of the function and the estimated curve becomes rough. Whereas, higher bandwidth leads to smoother curves.



Effects of different bandwidths in KDE

4.5.3. Histogram:

A histogram takes continuous (measured) data and displays its distribution.[16] A histogram is a relatively faster way to check the dispersion and nature of variation of a dataset without detailed statistical analysis or graphing.

 Balanced Histogram	 Skewed Histogram (Left)

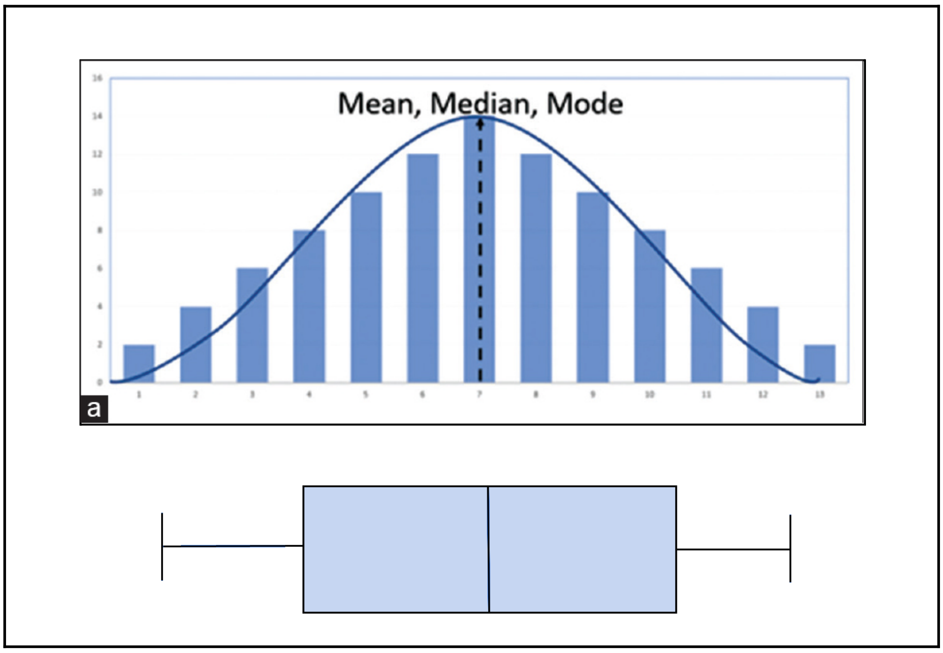
4.5.4. Skewness in Data:

Skewness is the measure of the asymmetry or degree of distortion of a histogram (frequency distribution). A histogram with normal distribution is symmetrical. In other words, the same amount of data falls on both sides of the mean. A normal distribution will have a skewness of 0. The direction of skewness is “to the tail.” The larger the number, the longer the tail. If skewness is positive, the tail on the right side of the distribution will be longer. If skewness is negative, the tail on the left side will be longer. [17] Function distplot from the seaborn library is used to create and compare the histogram of “rfm” dataframe to visualise the degree of skewness.

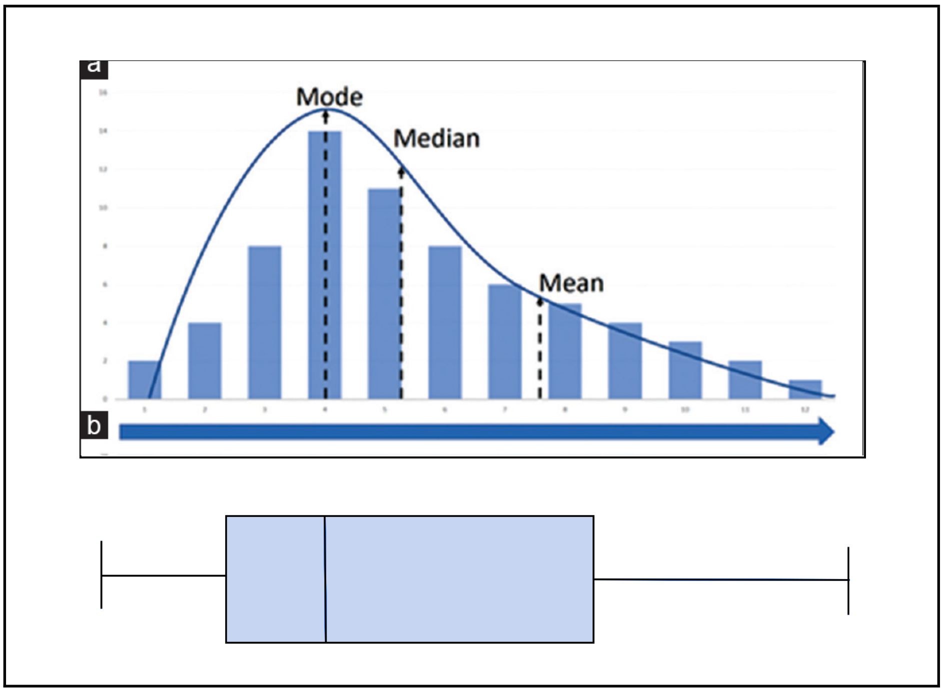
The formula given as,

Skew = 3 * (Mean – Median) / Standard Deviation
is known as an alternative Pearson Mode Skewness.

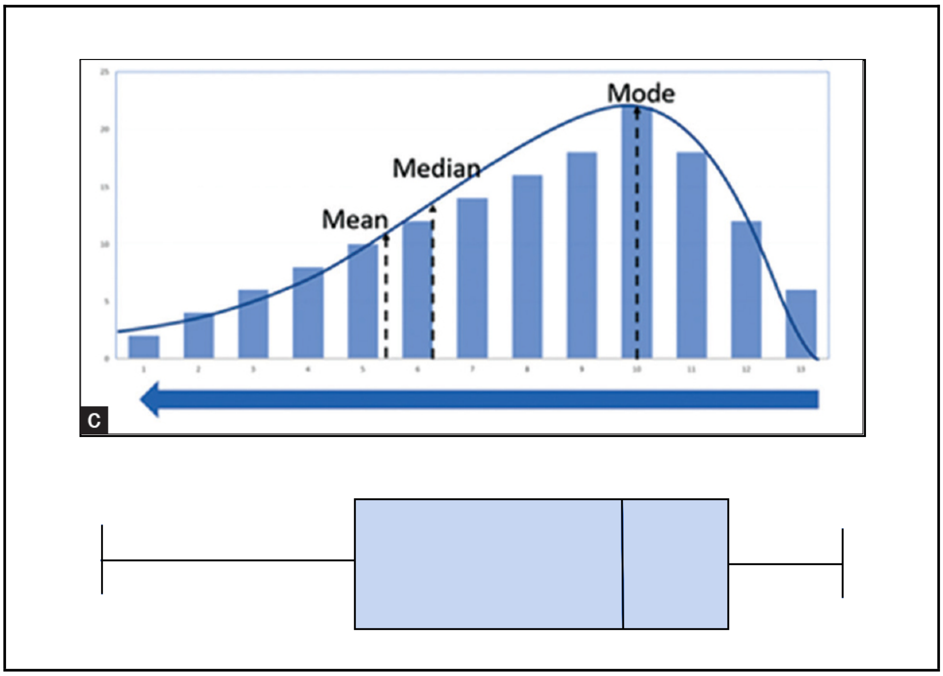
Three types of skewness are seen in the probability distribution of data:



Symmetric Data Distribution (0 Skew)



Right - Skewed Data Distribution (Positive Skew)



Left - Skewed Data Distribution (Negative Skew)

4.6. Transformation of Data

One of the assumptions that the k-means algorithm makes is that the input data is normally distributed (symmetric). Hence feeding asymmetric data to the algorithm may lead to inaccurate outcome.Therefore the data, if skewed, needs to be transformed into symmetric data. Three most common and useful methods of data transformation are:

- Log transformation.
- Square-root transformation.
- Box-cox transformation.

4.6.1. Log Transformation:

The **log transformation** is, arguably, the most popular among the different types of **transformations used to transform** skewed data to approximately conform to normality.**Log transformation** is a data **transformation method** in which it replaces each variable x with a **log**(x). The choice of the **logarithm** base is usually left up to the analyst and it would depend on the purposes of statistical modeling.[18]

4.6.2. Square-root Transformation:

The **square-root transformation** is a procedure for converting a set of data in which each value, x_i, is replaced by its **square root**, another number that when multiplied by itself yields x_i. Square-root transformations often result in homogeneity of variance for the different levels of the independent variable (x) under consideration.[19]It is weaker than the logarithm but it can be applied to zero values as an advantage.

4.6.3. Box-cox Transformation:[20]

Normality is an important assumption for many statistical techniques. A Box Cox transformation is a transformation of a non-normal dependent variable into a normal shape.

At the core of the Box Cox transformation is an exponent, lambda (λ), which varies from -5 to 5. All values of λ are considered and the optimal value for your data is selected; The "optimal value" is the one which results in the best approximation of a normal distribution curve. The transformation of Y has the form:

$$\boxed{y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}}$$

$$\boxed{y(\boldsymbol{\lambda}) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \text{if } \lambda_1 \neq 0; \\ \log(y+\lambda_2), & \text{if } \lambda_1 = 0. \end{cases}}$$

4.7. Feature Scaling

After transformation from the box plot the range and variance of 'Recency' is found to be very large, compared to 'Frequency' and 'Monetary'. K-Means being a distance based algorithm can be highly affected by range of the features, as it tends to get biased towards features of higher magnitude. Therefore, for k-means to produce appropriate output, the features need to be rescaled. The two methods used for feature rescaling are Standardisation and Normalisation. The sklearn.preprocessing package in python provides several common utility functions and transformer classes. The MinMaxScaler and StandardScaler objects are used for performing normalisation and standardisation respectively. Thereafter the results of the two methods are compared to analyse which one of them performed better on the input data.

4.7.1. Normalisation:[21]

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

1. When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0.
2. On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1.
3. If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1.

4.7.2. Standardisation:[21]

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

Feature rescaling brings all the features equally in picture, therefore distance between them becomes more comparable and chances of a biased result decreases.

4.8. Clustering Using K-Means Algorithm

K-means clustering algorithm makes certain assumptions about the nature of the input data. The assumptions are as follows [31]:

- Each variable in the data is symmetric
- The variance of the distribution of each attribute (variable) is spherical
- All variables have the same variance
- Each cluster has roughly equal number of observations

It is important for the data to satisfy the above assumptions for k-means to produce a reliable and unbiased result. This is the reason why the data needs to be preprocessed by transformation and scaling.

The Lloyd's K-Means approach, which is used in this project, considers the Euclidean distance between two data points to find a cluster and it's centroid. Therefore, scaling of data is necessary as this tends to be biased towards higher value data points.

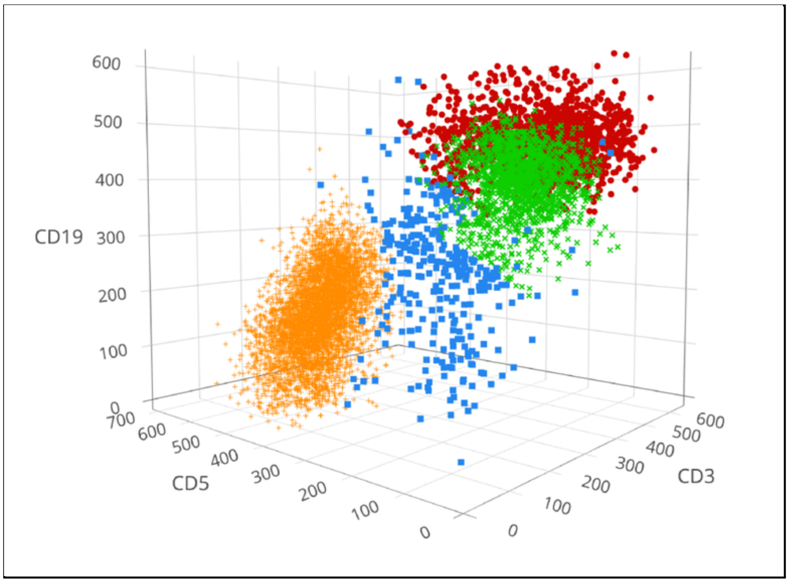
After the preprocessing stage, the data obtained is ready for clustering. Elbow method is applied on the dataset where k is varied from 1 to 10 and a rectangular hyperbolic curve is obtained, with the elbow containing k values 3,4 and 5. It is difficult to choose any one of them hence all of the three are considered for the cluster analysis.

While using the k-means method the random-variable is assigned a fixed value 42 so that the same centroids are chosen in every execution.

4.8.1. Data Visualisation and Choosing the Optimum value of "k":

Another popular way of data visualization is by using scatter plots. It displays the properties of data using the cartesian coordinate system, thereby revealing information about relationships between data attributes.

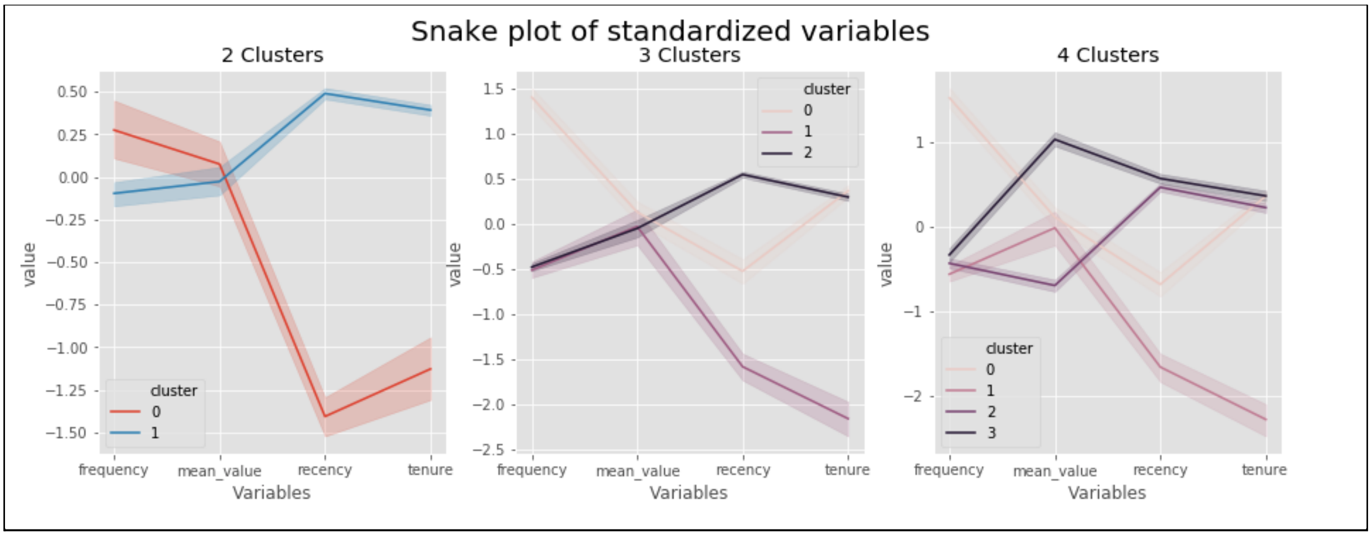
As mentioned before, the data that is input to the clustering stage has three variables, 'Recency', 'Frequency' and 'Monetary', hence the clusters are visualized using 3-Dimensional scatter plots.



3-D Scatter Plot

All the clusters and their centroids are plotted in 3-D scatter plots and Snake plots are used to plot the clusters statistics for analysing the result.

Snake plot is used to visualise and analyse the clusters. The x axis marks the three variables ‘Recency’, ‘Frequency’ and ‘Monetary’. The y-axis shows the range for each variable and since the variables are scaled they have quite uniform ranges. The dataset is melted using pandas.melt() and a mean for each value is calculated.Pandas.melt() unpivots a DataFrame from wide format to long format.
melt() function is useful to massage a DataFrame into a format where one or more columns are identifier variables, while all other columns, considered measured variables, are unpivoted to the row axis, leaving just two non-identifier columns, variable and value.



Structure of Snake Plot

4.8.2 Cluster Analysis:

The 3-D scatter plot uses different colours for different clusters and plots the data points with their respective colours. By visualising this plot the value of “k”, which creates well distinguished clusters, can be chosen.

From the snake plots the range of each of the three variables can be derived and using those properties of each cluster can be visualized. Here also, the value of “k” for which well distinguished clusters can be derived can be chosen.

5. IMPLEMENTATION

Programming Language: **Python**

Python is an object-oriented high level programming language along with the feature of dynamic typing and dynamic binding. It is a general purpose programming language with a wide variety of application domains. Python is also an interpreted language with dynamic semantics and is easy to read as well as easy to learn which makes it suitable for Rapid Application Development (RAD) .

Being an open-source project, Python has a huge collection of open-source libraries covering most of the fields of software application development. There are over 137,000 python libraries available [22] and as a result Python is one of the most dynamic and versatile programming languages present in today's world.

One of the most important applications of Python is in the field of data analysis. Some of the popular Python libraries that are extensively used for data analysis (which are also used in the following experiment) are as follows:

- Pandas - It uses the efficient Dataframe object for streamlining complex data manipulation. It also provides tools for reading from and writing to various file formats like text, CSV, excel, etc.
- Matplotlib and Seaborn - They are the most powerful data visualisation libraries used for plotting data. Various plots like histogram, box plots, scatter plots are very much useful while analysing the relationships between various components of data.
- Numpy and Scipy - They consist of various numeric and scientific functions like log, square root, box-cox, etc.
- Scikit-learn - A library that provides means to implement various machine learning algorithms like regression, classification, clustering, etc., efficiently and accurately.

Software Platform: **Google Colaboratory**

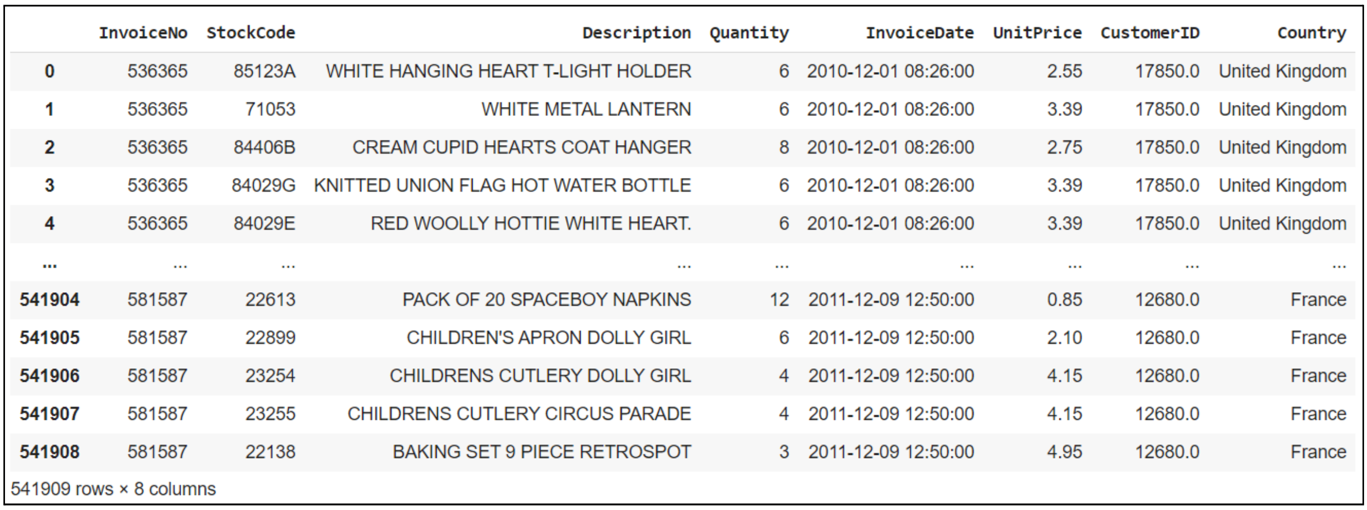
Google Colaboratory or Google Colab is a cloud-based Jupyter Notebook. It uses the Anaconda distribution of Python which comes with several pre-installed libraries like Pandas, Numpy, Matplotlib, which are extensively used in the field of Data Science and Machine Learning. Availability of the integrated GPU (Graphics Processing Unit) and TPU (Tensor Processing Unit) also makes it suitable for implementation of resource intensive tasks, for example, training deep learning neural networks.

5.1. Source Code:

```
### INCLUDING THE REQUIRED LIBRARIES

import pandas as pd
import numpy as np
from scipy import stats
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib import rcParams

rcParams['figure.figsize'] = (8,6)
```

```
### IMPORTING THE DATASET

df=pd.read_excel('https://github.com/BShria/Customer-Segmentation/blob/main/Online
%20Retail.xlsx?raw=true')
```

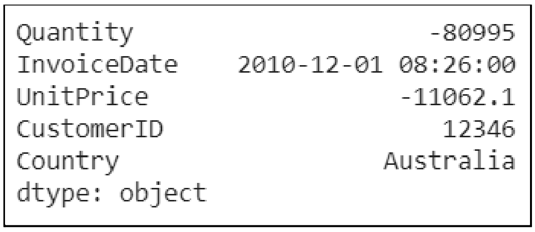
The dataset looks like this:



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

541909 rows × 8 columns

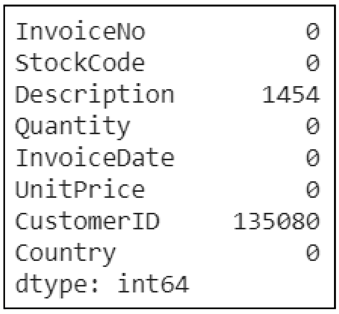
```
### CLEANING THE DATA

# Column minimum
df.min()
```



```
Quantity                    -80995
InvoiceDate    2010-12-01 08:26:00
UnitPrice                 -11062.1
CustomerID                   12346
Country                  Australia
dtype: object
```

Observation: The columns ‘Quantity’ and ‘UnitPrice’ contain negative values.

```
# Checking the number of missing values corresponding to each column
df.isnull().sum()
```



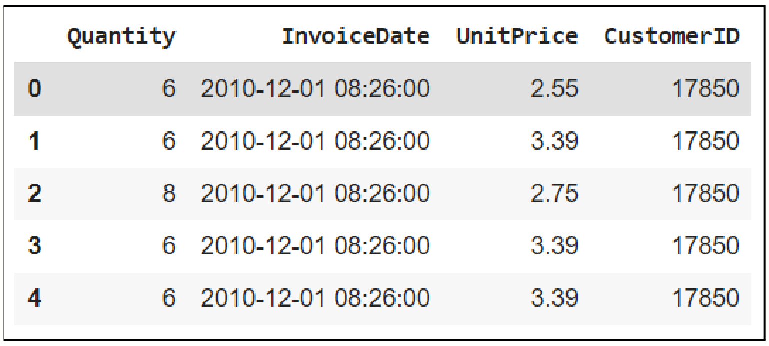
```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

Observation: A lot of rows have missing CustomerId and Description.

```
# Removing rows with missing values
df = df.dropna()

# Dropping InvoiceNo, StockCode, Description, Country as they are redundant
df = df.drop(['InvoiceNo','StockCode','Description' ,'Country'], axis=1)
```

```
# Removing rows with Quantity <= 0
negative_quantity = df[df['Quantity']<=0].index
df.drop(negative_quantity, inplace=True, axis=0)

# Removing rows with UnitPrice <= 0
negative_price = df[df['UnitPrice']<=0].index
df.drop(negative_price, inplace=True, axis=0)

# Converting CustomerID values from float type to int
df.CustomerID = df.CustomerID.astype(int)
df.head(5)
```

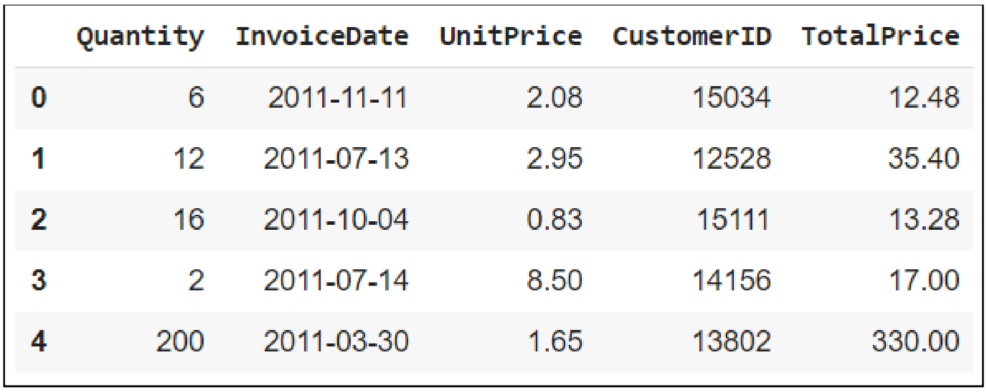
After Cleaning the data looks like this:



	Quantity	InvoiceDate	UnitPrice	CustomerID
0	6	2010-12-01 08:26:00	2.55	17850
1	6	2010-12-01 08:26:00	3.39	17850
2	8	2010-12-01 08:26:00	2.75	17850
3	6	2010-12-01 08:26:00	3.39	17850
4	6	2010-12-01 08:26:00	3.39	17850

```
### SAMPLING THE DATASET

sample = df.sample(10000, random_state = 42)
index = pd.Series(list(range(0,10000)))
sample.set_index(index, drop=False, inplace=True);

# Calculating Total Price against each purchase
sample["TotalPrice"] = sample["Quantity"] * sample["UnitPrice"]

# Extracting the dates from Invoice Date which is in datetime format
from datetime import datetime
sample["InvoiceDate"] = sample["InvoiceDate"].dt.date
sample.head(5)
```

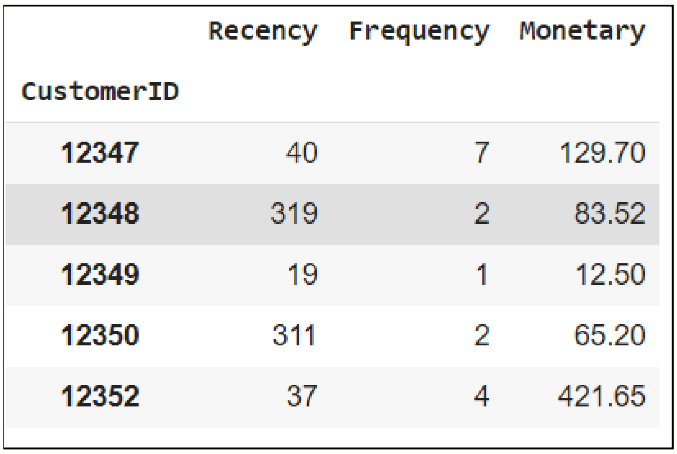
After Sampling the data looks like this:



	Quantity	InvoiceDate	UnitPrice	CustomerID	TotalPrice
0	6	2011-11-11	2.08	15034	12.48
1	12	2011-07-13	2.95	12528	35.40
2	16	2011-10-04	0.83	15111	13.28
3	2	2011-07-14	8.50	14156	17.00
4	200	2011-03-30	1.65	13802	330.00

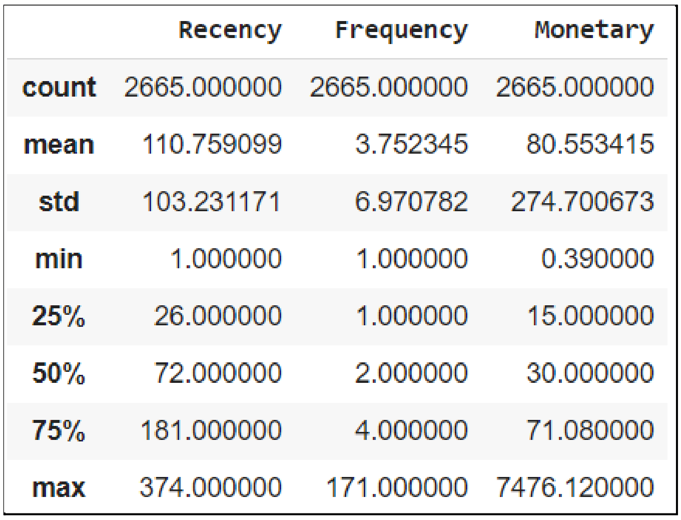
```
### CREATING THE RFM MODEL

max_date=max(sample.InvoiceDate)
rfm=sample.groupby('CustomerID').agg({'InvoiceDate':   lambda   date:(max_date   -
date.max()).days + 1,
                                      'CustomerID': lambda cust: len(cust),
                                       'TotalPrice': lambda price: price.sum()})
rfm.columns=['Recency','Frequency','Monetary']
rfm.head(5)
```

The RFM Model looks like this:



	Recency	Frequency	Monetary
CustomerID			
12347	40	7	129.70
12348	319	2	83.52
12349	19	1	12.50
12350	311	2	65.20
12352	37	4	421.65

```
### ANALYSIS OF THE RFM DATASET

## Statistics of the dataset
rfm.describe()
```



	Recency	Frequency	Monetary
count	2665.000000	2665.000000	2665.000000
mean	110.759099	3.752345	80.553415
std	103.231171	6.970782	274.700673
min	1.000000	1.000000	0.390000
25%	26.000000	1.000000	15.000000
50%	72.000000	2.000000	30.000000
75%	181.000000	4.000000	71.080000
max	374.000000	171.000000	7476.120000

Observation: K-means clustering algorithm assumes that all the variables in the dataset have the same mean and variance. But the 'mean' and 'std' information that is shown in the above table corresponding to each feature, indicates that there are huge differences between the means and standard deviations (variance) of the attributes.

```
# Box Plot to analyse the distribution of data
fig = rfm.plot.box()
fig.set_title('Fig.5.1: Range of the RFM features')
```

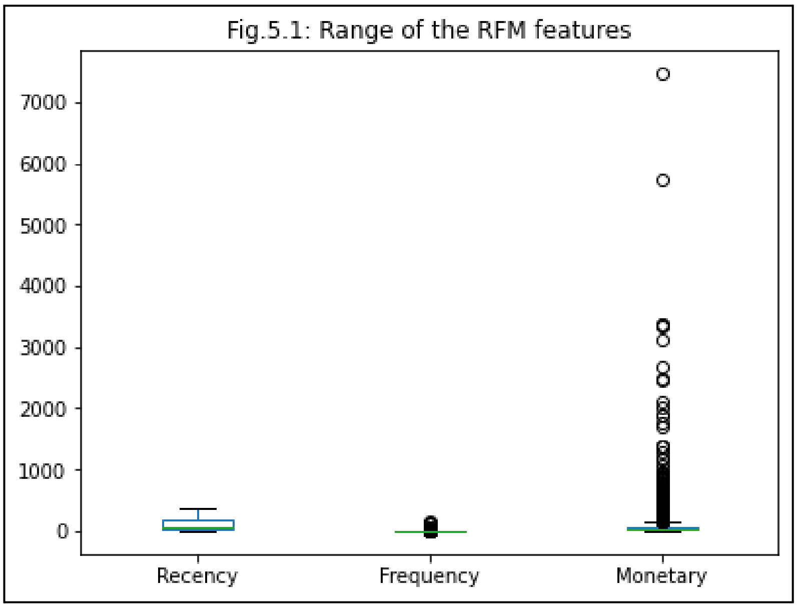
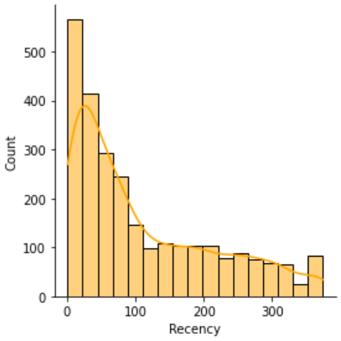
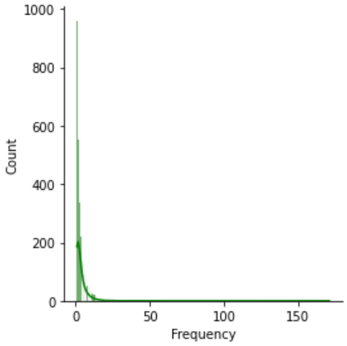
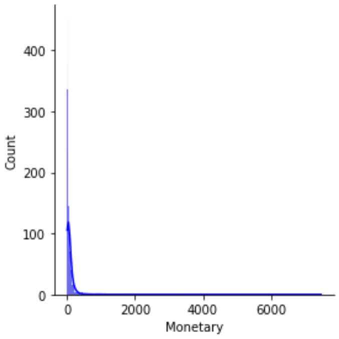


Fig.5.1: Range of the RFM features

Observation: Here the box plots are not clearly visible proving the difference in scales of the variables. The box plot also indicates signs of asymmetry (skewness) in the data distribution. K-means also expects the data to be normally distributed, that is symmetric. And applying k-means to asymmetric data may lead to faulty clustering. Therefore to get a better idea about the skewness in the data, the three variables are analysed individually.

```
## Exploring the Skewness in each of the features
# Recency
rfm['Recency'].skew()
sns.displot(rfm.Recency, kde=True, color='orange', height=4, aspect=1)

# Frequency
rfm['Frequency'].skew()
sns.displot(rfm.Frequency, kde=True, color='Green', height=4, aspect=1)

# Monetary
rfm['Monetary'].skew()
sns.displot(rfm.Monetary, kde=True, color='Blue', height=4, aspect=1)
```

Table 5.1: Skewness in the Data Distribution

Features	Recency	Frequency	Monetary
Skewness	0.9419160349502619	13.18802368680814	14.910987361225443
Histogram + **KDE Plot**			

Observation: The raw data are heavily right-skewed, especially Frequency and Monetary. Hence the data are required to be transformed such that the skewness of the data is as close as possible to 0. If the skewness is in between -0.5 and 0.5 then the data are said to be fairly symmetrical.

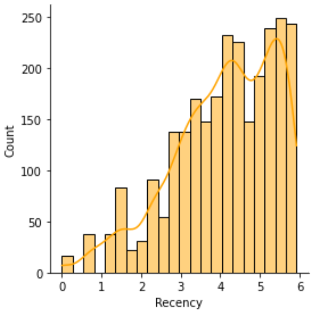
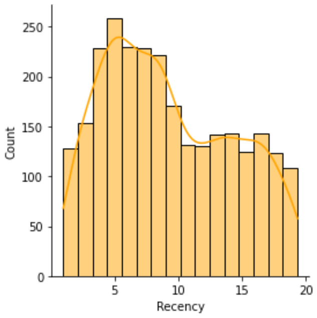
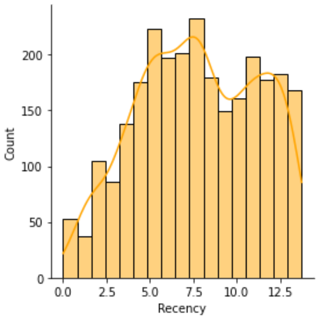
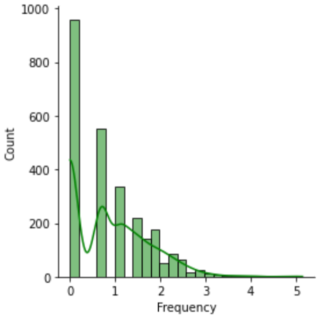
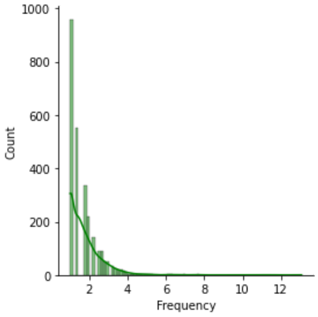
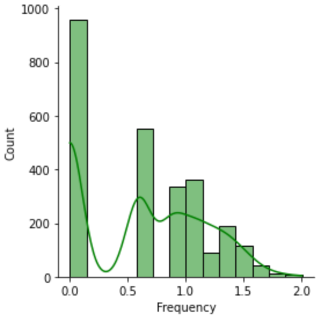
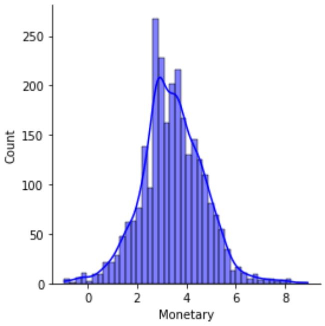
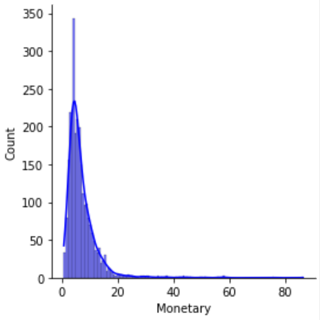
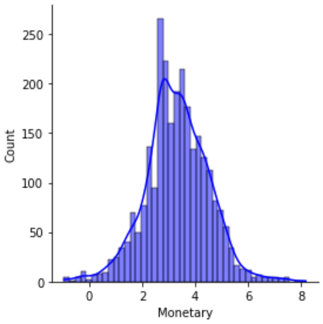
```
### TRANSFORMATION OF DATA

## Log Transformation
# Recency
recency_log = np.log(rfm['Recency'])
recency_log.skew()
sns.displot(recency_log, kde=True, color='orange', height=4, aspect=1)

# Frequency
frequency_log = np.log(rfm['Frequency'])
frequency_log.skew()
sns.displot(frequency_log, kde=True, color='green', height=4, aspect=1)

# Monetary
monetary_log = np.log(rfm['Monetary'])
monetary_log.skew()
sns.displot(monetary_log, kde=True, color='blue', height=4, aspect=1)
```

```
## Square Root Transformation
# Recency
recency_sqrt = np.sqrt(rfm['Recency'])
recency_sqrt.skew()
sns.displot(recency_sqrt, kde=True, color='orange', height=4, aspect=1)

# Frequency
frequency_sqrt = np.sqrt(rfm['Frequency'])
frequency_sqrt.skew()
sns.displot(frequency_sqrt, kde=True, color='green', height=4, aspect=1)

# Monetary
monetary_sqrt = np.sqrt(rfm['Monetary'])
monetary_sqrt.skew()
sns.displot(monetary_sqrt, kde=True, color='blue', height=4, aspect=1)
```

```
## Box-Cox Transformation
# Recency
recency_bcox = pd.Series(stats.boxcox(rfm['Recency'])[0])
recency_bcox = recency_bcox.rename('Recency')
recency_bcox.skew()
sns.displot(recency_bcox, kde=True, color='orange', height=4, aspect=1)

# Frequency
frequency_bcox = pd.Series(stats.boxcox(rfm['Frequency'])[0])
frequency_bcox = frequency_bcox.rename('Frequency')
frequency_bcox.skew()
```

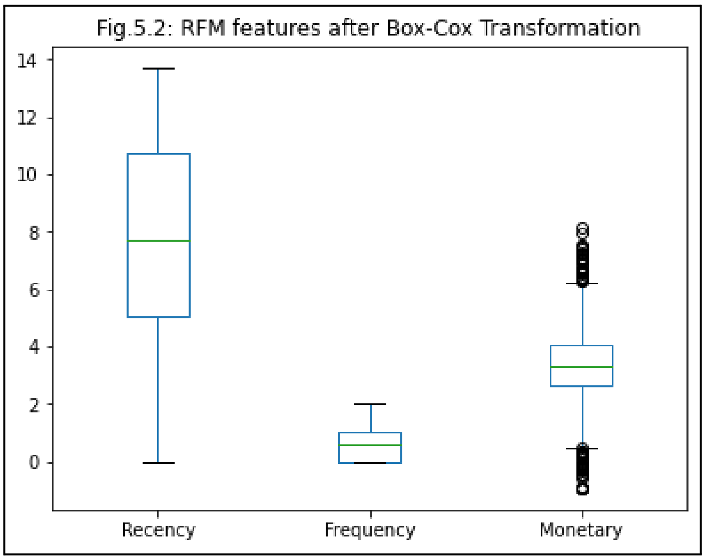
```
sns.displot(frequency_bcox, kde=True, color='green', height=4, aspect=1)
# Monetary
monetary_bcox = pd.Series(stats.boxcox(rfm['Monetary'])[0])
monetary_bcox = monetary_bcox.rename('Monetary')
monetary_bcox.skew()
sns.displot(monetary_bcox, kde=True, color='blue', height=4, aspect=1)
```

Table 5.2: Comparison of the Results corresponding to each Transformation

Features		**Log Transformation**	**Square Root Transformation**	**Box-Cox Transformation**
Recency	**Skewness**	-0.7198816256981649	0.3327355555263751	-0.10304787698346707
	Histogram + KDE Plot			
Frequency	**Skewness**	0.8218818004304456	3.519791158958472	0.1513177739575201
	Histogram + KDE Plot			
Monetary	**Skewness**	0.10209844098355485	4.712375019810757	-0.007362053674569914
	Histogram + KDE Plot			

Observation: With respect to the above comparison, it is clearly seen that the Box-Cox Transformation has performed better than the other transformations because the skewness corresponding to Box-Cox transformation is the closest to 0. Therefore the Box-Cox transformed data is considered for further analysis.

```
frame = {'Recency':pd.Series(recency_bcox),'Frequency':pd.Series(frequency_bcox),
'Monetary':(pd.Series(monetary_bcox))}
transformed_rfm=pd.DataFrame(frame)

# Box plot of transformed data
fig = transformed_rfm.plot.box()
fig.set_title('Fig.5.2: RFM features after Box-Cox Transformation')
```

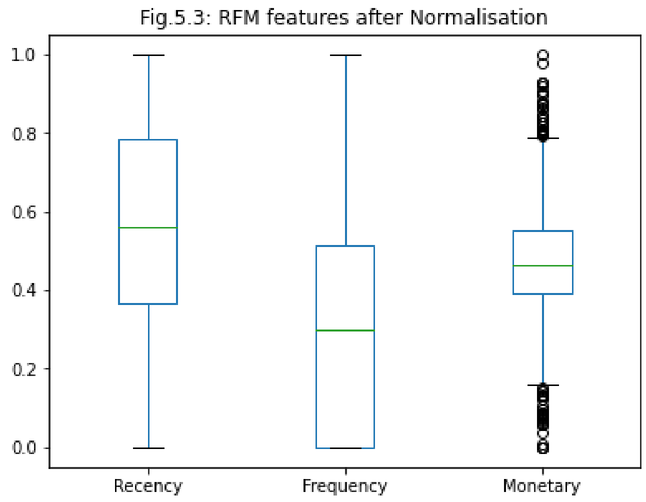
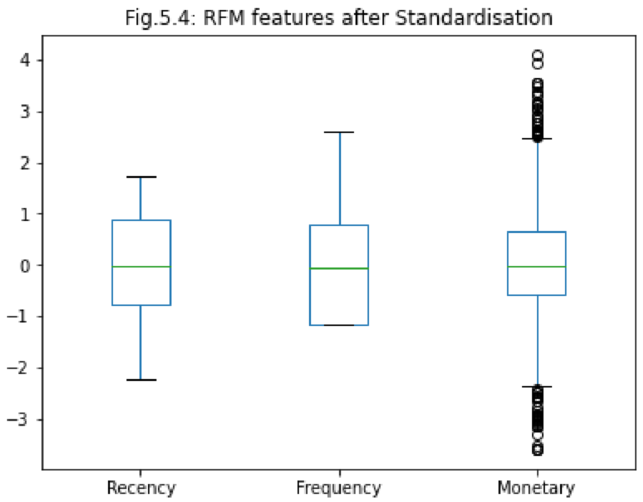


Observation: The process of transformation rendered the data reasonably symmetric. However, comparing the locations of the boxes and their medians, it is seen that there are differences in the magnitudes of the features. Recency has higher values whereas frequency has lower values. Distance based machine learning algorithms (like k-means) tend to get biased towards features of higher magnitudes. Therefore it is compulsory to scale the features such that they have similar ranges, for k-means to produce an unbiased output. Also, the monetary feature has a number of outliers, but as there are a lot of them they might reveal interesting patterns in the data. Hence they are not removed.

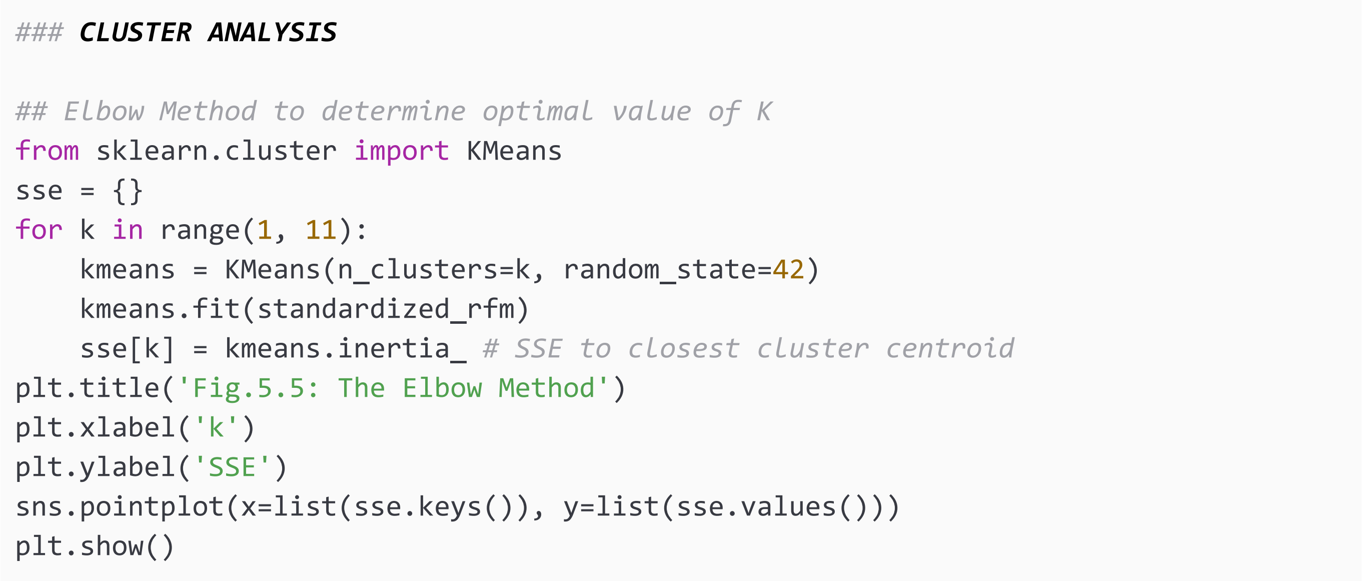
```
### FEATURE SCALING

# Normalisation using MinMaxScaler from sklearn.preprocessing package
from sklearn.preprocessing import MinMaxScaler

normalised_rfm = MinMaxScaler().fit_transform(transformed_rfm)
normalised_rfm = pd.DataFrame(normalised_rfm, columns=transformed_rfm.columns)
fig = normalised_rfm.plot.box()
fig.set_title('Fig.5.3: RFM features after Normalisation')
```

```
#Standardisation using StandardScaler from sklearn.preprocessing
from sklearn.preprocessing import StandardScaler

standardised_rfm = StandardScaler().fit_transform(transformed_rfm)
standardised_rfm = pd.DataFrame(standardised_rfm, columns=transformed_rfm.columns)
fig = standardised_rfm.plot.box()
fig.set_title('Fig.5.4: RFM features after Standardisation')
```

Table 5.3: Normalisation versus Standardisation

Normalised Data	Standardised Data
 Fig.5.3: RFM features after Normalisation 	 Fig.5.4: RFM features after Standardisation

Observation: It can be observed that Standardisation has produced better results than Normalisation. Hence the standardised data is subjected to the final cluster analysis using k-means clustering algorithm.



```
### CLUSTER ANALYSIS

## Elbow Method to determine optimal value of K
from sklearn.cluster import KMeans
sse = {}
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(standardized_rfm)
    sse[k] = kmeans.inertia_ # SSE to closest cluster centroid
plt.title('Fig.5.5: The Elbow Method')
plt.xlabel('k')
plt.ylabel('SSE')
sns.pointplot(x=list(sse.keys()), y=list(sse.values()))
plt.show()
```

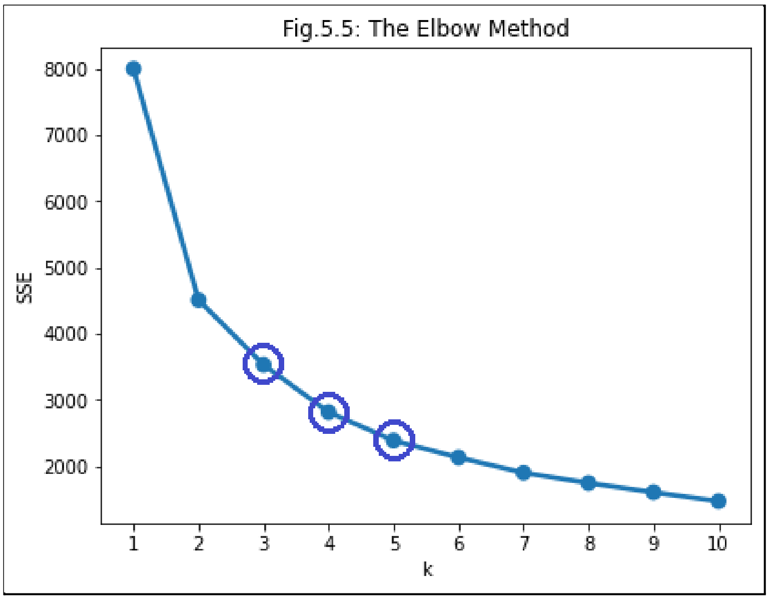


Fig.5.5: The Elbow Method

Observation: Here k = 3, 4 and 5 forms the elbow of the plot. All three of them are considered for performing clustering and then the most optimal one is chosen depending on the output.

```
## K-Means Clustering for K = 3
rcParams['figure.figsize'] = (12,9)
k = 3
rfm3 = standardized_rfm.copy()
kmeans3 = KMeans(n_clusters=k, random_state=42)
kmeans3.fit(rfm3)
labels = kmeans3.predict(rfm3)

# Scatter Plot
ax = plt.axes(projection='3d')
for i in range(0,k):
  labeled_rfm = rfm3[labels == i]
  x = labeled_rfm['Recency']
  y = labeled_rfm['Frequency']
  z = labeled_rfm['Monetary']
  ax.scatter(x, y, z, label = i)
ax.set_xlabel('Recency')
ax.set_ylabel('Frequency')
ax.set_zlabel('Monetary');
ax.legend(loc = 'upper left')
ax.set_title('3-D Clustering with K=3')

# Plotting the Centroids
centroids3 = kmeans3.cluster_centers_
ax.scatter(centroids3[:,0],  centroids3[:,1],  centroids3[:,2],  c='black',  s=100,
depthshade=False)

# Using a Snake Plot for differentiating between the clusters
rfm3['Cluster'] = labels
clusters3 = rfm3.groupby('Cluster').agg(['mean'])
clusters3['Cluster'] = clusters3.index
clusters3 = clusters3.melt(id_vars='Cluster', var_name = 'Attributes',
              value_name = 'Values')

rcParams['figure.figsize']=(8,6)
sns.lineplot(x='Attributes',  y='Values',  hue='Cluster',  data=clusters3,  palette=
['blue','orange','green'])
```

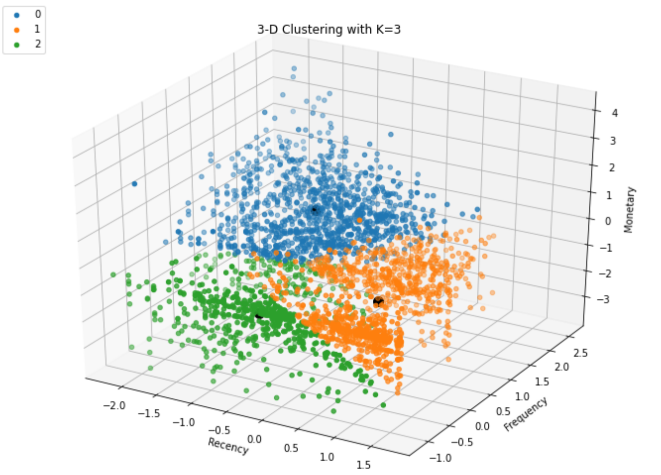
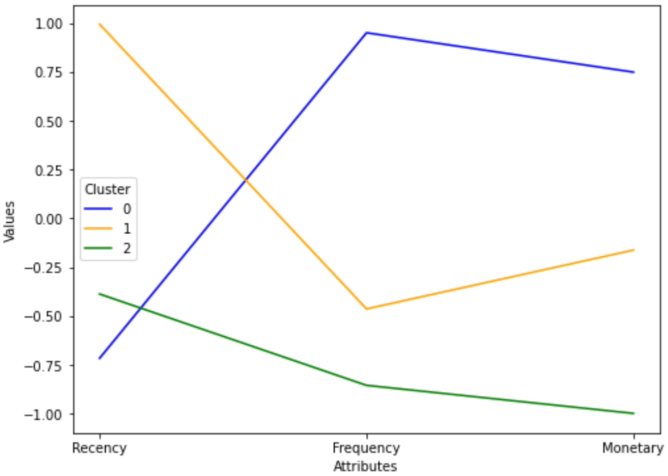
```
## K-Means Clustering for K = 4
rcParams['figure.figsize'] = (12,9)
k = 4
rfm4 = standardized_rfm.copy();
kmeans4 = KMeans(n_clusters=k, random_state=42)
kmeans4.fit(rfm4)
labels = kmeans4.predict(rfm4)
```

```
# Scatter Plot
ax = plt.axes(projection='3d')
for i in range(0,k):
  labeled_rfm = rfm4[labels == i]
  x = labeled_rfm['Recency']
  y = labeled_rfm['Frequency']
  z = labeled_rfm['Monetary']
  ax.scatter(x, y, z, label = i)
ax.set_xlabel('Recency')
ax.set_ylabel('Frequency')
ax.set_zlabel('Monetary');
ax.legend(loc = 'upper left')
ax.set_title('3-D Clustering with K=4')

# Plotting the Centroids
centroids4 = kmeans4.cluster_centers_
ax.scatter(centroids4[:,0], centroids4[:,1], centroids4[:,2], c='black', s=100,
depthshade=False)

# Using a Snake Plot for differentiating between the clusters
rfm4['Cluster'] = labels
clusters4 = rfm4.groupby('Cluster').agg(['mean'])
clusters4['Cluster'] = clusters4.index
clusters4 = clusters4.melt(id_vars='Cluster', var_name = 'Attributes',
              value_name = 'Values')

rcParams['figure.figsize']=(8,6)
sns.lineplot(x='Attributes', y='Values', hue='Cluster', data=clusters4, palette=
['blue','orange','green','red'])
```

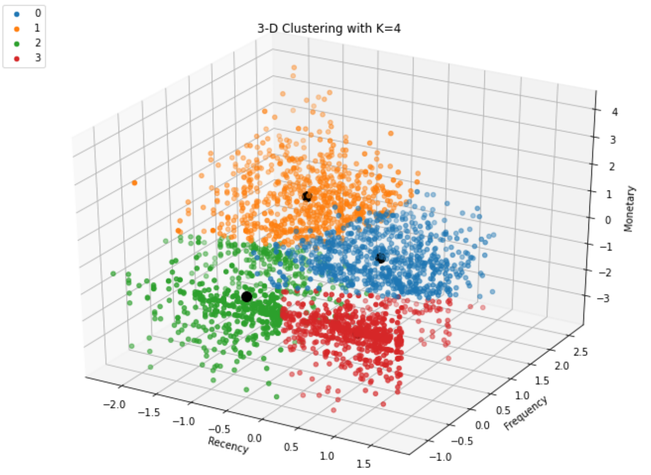
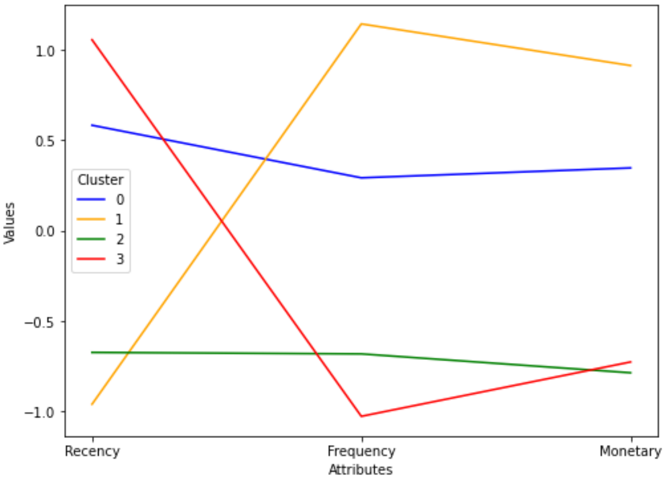
```
## K-Means Clustering for K = 5
rcParams['figure.figsize'] = (12,9)
k = 5
rfm5 = standardized_rfm.copy();
kmeans5 = KMeans(n_clusters=k, random_state=42)
kmeans5.fit(rfm5)
labels = kmeans5.predict(rfm5)

# Scatter Plot
ax = plt.axes(projection='3d')
for i in range(0,k):
  labeled_rfm = rfm5[labels == i]
  x = labeled_rfm['Recency']
  y = labeled_rfm['Frequency']
  z = labeled_rfm['Monetary']
  ax.scatter(x, y, z, label = i)
ax.set_xlabel('Recency')
ax.set_ylabel('Frequency')
ax.set_zlabel('Monetary');
ax.legend(loc = 'upper left')
ax.set_title('3-D Clustering with K=5')
```

```
# Plotting the Centroids
centroids5 = kmeans5.cluster_centers_
ax.scatter(centroids5[:,0],  centroids5[:,1],  centroids5[:,2],  c='black',  s=100,
depthshade=False)

# Using a Snake Plot for differentiating between the clusters
rfm5['Cluster'] = labels
clusters5 = rfm5.groupby('Cluster').agg(['mean'])

clusters5['Cluster'] = clusters5.index
clusters5 = clusters5.melt(id_vars='Cluster', var_name = 'Attributes',
              value_name = 'Values')

rcParams['figure.figsize']=(8,6)
sns.lineplot(x='Attributes',  y='Values',  hue='Cluster',  data=clusters5,  palette=
['blue','orange','green','red','purple'])
```

Table 5.4: Comparison of Results after Clustering with K = 3, 4 and 5

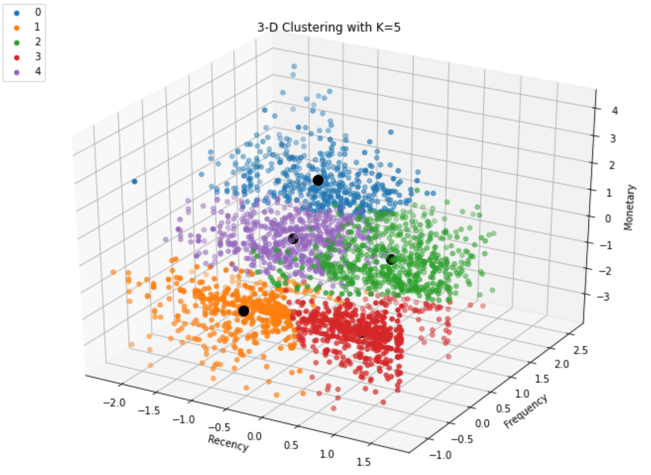
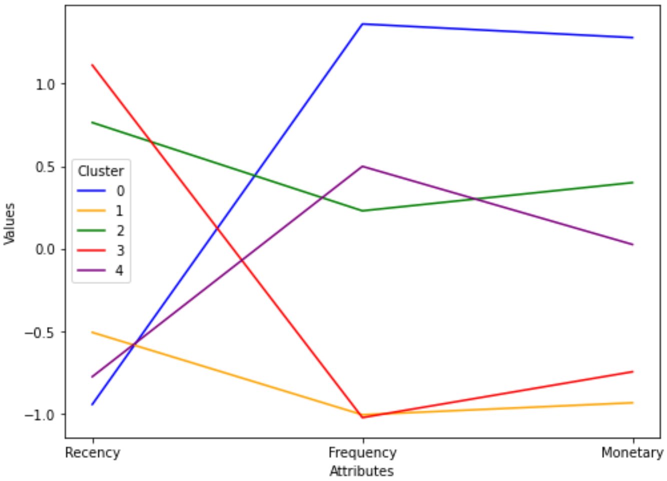
K = 3	
Scatter Plot	**Snake plot**
	
Observation: For k = 3, it can be said that - ● Cluster-0 (blue) is the most recent with highest frequency and monetary value. Hence this is the loyal customer segment. ● Cluster-1 (orange) has the highest recency, lower frequency and monetary values. Therefore this could be the cluster of the former customers. ● However for Cluster-2 (green) the recency value is low to medium (> -0.5) and it is difficult to conclude whether it is the cluster of the new customers or the former customers whose interaction with the business was infrequent. Therefore this information is not enough to draw appropriate business inferences.	

K = 4

Scatter Plot	Snake plot
	

Observation: Clustering with k = 4 results in 4 distinct clusters each of which provides us with specific information to reach a conclusion about each segment.

K = 5

Scatter Plot	Snake plot
	

Observation: The primary aim of k-means clustering is to divide the dataset in such a way that the data points are similar to the ones in its own cluster and are different from the ones in the other clusters. Here, both Cluster-0 (blue) and Cluster-4 (purple) have low recency, medium to high frequency and medium to high monetary value. As a result they appear to be similar, making Cluster-4 redundant.

5.2. Results:

From the above observations it is clear that the optimal number of Clusters is 4. Therefore analysing the resultant Clusters for k = 4:

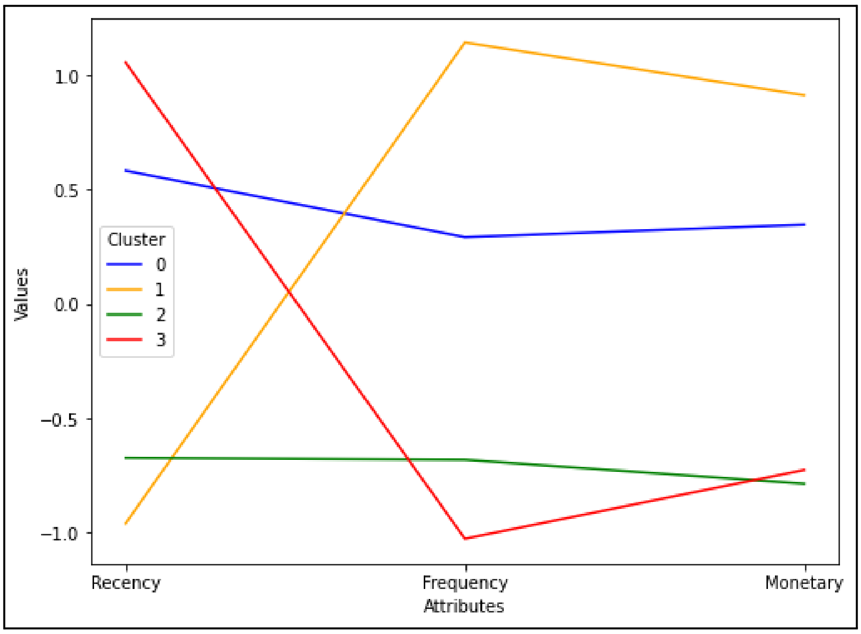
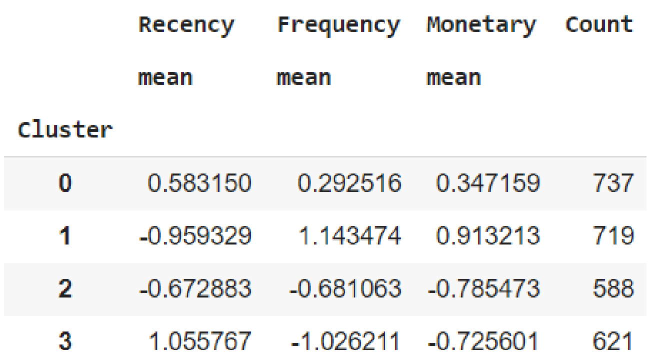


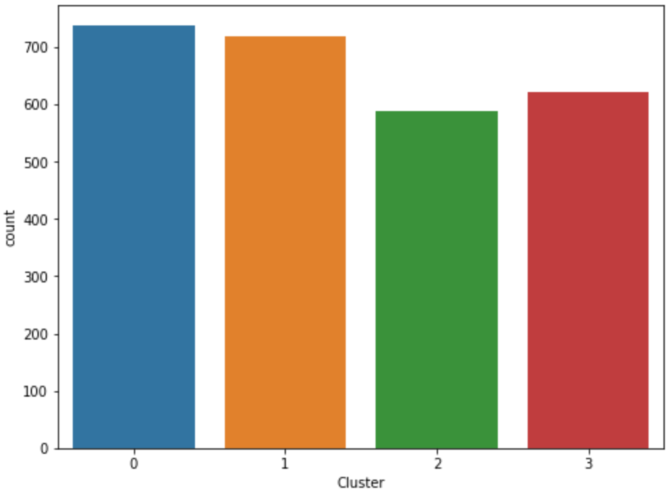
Table 5.5: Interpretation of the Clusters

Cluster - 0 (Blue)	● More frequent and spent more but bought the products long ago. Therefore this could be the cluster of former but potential customers. The concerned company should try to re-engage these customers by target marketing and advertising as former buyers are usually better prospects than non buyers because they have already shown interest and ability to buy.
Cluster - 1 (Orange)	● Bought the products recently and are the most frequent customers with the highest expenditure. Hence this could be the cluster of regular and loyal customers and therefore of highest priority. Marketing policies should be devised for retaining this group of potential buyers because it is observed that increasing customer retention rates by 5% has increased the profit by 25% to 95% for an average business [23].
Cluster - 2 (Green)	● Bought the products recently, but are less frequent with lesser expenditure. Therefore this could be the cluster of new customers. Coherent strategies should be developed for building their brand trust.
Cluster - 3 (Red)	● Least frequent with less expenditure and bought the products long ago. Hence this could be the cluster of churned customers. This category of customers are the least profitable but can be extremely helpful for churn analysis. Churn rate is the rate at which customers stop engaging with a business over a specified amount of time and churn analysis is the process of understanding the reason behind it based on statistical data. Therefore analysing these fluctuating churn rates can provide insightful observations which may help to actively reduce churn rates and eventually resolve customer retention problems [24].

Table 5.6: Cluster Statistics

<table>
<tr><th>Mean of the Features and the number of data points (Count) in each Cluster</th><th>Bar Graph for Count in each Cluster</th></tr>
<tr><td>


	Recency mean	Frequency mean	Monetary mean	Count
Cluster				
0	0.583150	0.292516	0.347159	737
1	-0.959329	1.143474	0.913213	719
2	-0.672883	-0.681063	-0.785473	588
3	1.055767	-1.026211	-0.725601	621

</td><td>


</td></tr>
<tr><td colspan="2"><b>Observation:</b> The sample that has been used in our experiment contains more former customers than loyal customers. Hence the concerned organisation should try to improve their customer retention.</td></tr>
</table>

5.3. Discussion:

Our experiment clearly demonstrates how to prepare a RFM model from raw data, how to preprocess the RFM data, how to perform k-means clustering on the preprocessed data using python and finally how to derive marketing insights from the resultant clusters. As clustering is an unsupervised machine learning model that deals with real time data, the resultant clusters may gradually change over time as more customers begin to interact with the organisation. This fact ensures that the clusters are reflecting the current state of data, and hence this method can provide the most precise and practical insights on the market segments.

Customer segmentation can also be achieved by a deterministic rule-based approach. This approach considers segmentation of customers based on specific heuristic rules. However, there are several disadvantages of the rule based segmentation process as compared to the cluster analysis approach. Segmentation based on rules defined by the organisation may satisfy the initial requirements but might not yield accurate results eventually because it is slow to adapt to changes. Moreover, it is practically infeasible to create appropriate rules that effectively predict user interests over more than two dimensions. On the contrary, clustering allows the available data to automatically reveal patterns and trends that are inherently present in the customer dataset, without the need of human intervention. Furthermore, unlike clustering, rule-based segmentation might not be able to preserve the homogeneity in the segments because the intra-segment variances may be large [25]. These limitations of the rule-based method makes machine learning models more suitable for practical problems like market research.

Despite the merits of clustering technique, the k-means clustering algorithm that has been used for cluster analysis in this project, has certain drawbacks. Firstly, the number of groups is to be predefined manually which may be difficult. Secondly, k-means starts with a random selection of cluster centers and therefore a faulty initialization of centroids can result in poor clustering. K-means also cannot handle clusters of varying shape, size and density. To overcome these problems, alternate algorithms like k-means++, gaussian mixture model (distribution based model), hierarchical clustering algorithm, etc., can be taken into account depending on the complexity of the problem.

6. CONCLUSION

Customer segmentation is a highly effective approach to comprehend the target market for maximising business turnovers and profits. It is not feasible for companies to keep the entire customer base satisfied all the time. Customer segmentation allows for recognising intra-segmental similarities as well as inter-segmental differences. As a result identification of the most profitable and the least profitable segments becomes easier which facilitates formulation of consistent marketing propositions to satisfy the needs of the potential customers. Determination of the most attractive customer segments in the market makes target marketing possible which in turn leads to economic resource allocation, increased sales and improved customer retention.

Though the needs of individual consumers differ, customers belonging to the same segment are most likely to share similar needs and interests, and thus are attracted to similar kinds of products and services. This fact allows a business to understand consumer interests efficiently and pursue focus strategy. Focus strategy can be defined as a marketing strategy that focuses on delivering a product or a service to a specific market segment [26]. It helps to improve product value as well as maximize market penetration and promotion with higher customer satisfaction.

However, along with several benefits, there are also a few limitations of market segmentation. Segmentation only gives an approximation about the potential customers of an organisation, but communicating the brand message to these customers using effective marketing techniques depends on the expertise and experience of the entrepreneur. In the worst case, incorrect judgement of the market may lead to product failure as well as poor sales. It is also observed that sometimes manufacturing a variety of custom goods, according to the demands of the different target segments, becomes costlier than mass producing standard products. Implementation of various marketing mixes for multiple segments increases administrative, promotional and storage expenses [27]. This counters the fundamental objective of profitability and therefore in such cases a segmented approach can prove to be inefficient as compared to non-segmented approach of marketing.

7. FUTURE SCOPE OF STUDY

The present study is limited to a very rudimentary idea about customer segmentation and its approaches. It has explored the application of an Euclidean distance based algorithm on the available numerical data for a business to consumer segmentation. However in practice, the segmentation process for large organisations can be extremely complex. In that case it may be required to consider categorical data like customer satisfaction, product reviews, etc. These inevitable complexities in the problem open up a scope for further research. Few of the recommended areas for future studies are:

- Customer Segmentation using Hierarchical Clustering Algorithm - For complex data it is difficult to meet the assumptions and predetermine the optimal value of k for k-means algorithm. Hierarchical clustering is an improvement over k-means where the user does not have to specify the number of clusters, rather the algorithm itself builds a hierarchical tree of nested clusters revealing insightful patterns in the data.
- Categorical Data Clustering - Categorical (qualitative) attributes may be assigned certain scores or ranks to make them numerical (quantitative) but these scores do not hold any numeric significance, nor do their means or the Euclidean distance between them. That is why k-means

fails to produce a reliable output in this case. To resolve this issue it is proposed to compute the mode, that is the most occurring categorical value of a feature in order to find the centroids of a cluster. This proposal has facilitated the formulation of mode and median based algorithms. Some of them that may be considered for exploration are k-modes, k-medians and k-medoids algorithms [32].

- Clustering of Heterogeneous Data - As heterogeneity in data increases it becomes more difficult to apply standard algorithms to analyse them. A number of alternate clustering frameworks have been proposed to approach such mixed data. Some of them are multiview clustering, ensemble clustering, collaborative clustering and semi-supervised clustering [29].
- Customer Feedback Analysis - Customer feedback (surveys and reviews) hold insights that can drive retention, increase conversion rates, and improve customer lifetime value. Analysis of customer feedback encompasses sentiment analysis which deals with unstructured text data. If the data is huge then the process can be automated using machine learning based natural language processing (NLP) [30].

There are a plethora of clustering algorithms that can be used for segmenting data. Different algorithms handle different types of data using different types of techniques. Therefore to solve complex problems all of these techniques can be subject to future studies.

References:

1. https://kcossin.com/2011/02/16/consumer-segmentation-a-contemporary-historical-perspective/
   Consumer Segmentation: A Contemporary-Historical Perspective by: David S.B. Butler, PhD.
2. https://clevertap.com/blog/customer-segmentation-examples-for-better-mobile-marketing/
   6 Customer Segmentation Examples for Better Mobile Marketing by Emily Bonnie.
3. Customer Segmentation with RFM analysis — Part 1 | by Aman Prasad | Merino Services Analytics Blog | Medium.
4. Customer segmentation of multiple category data in e-commerce using a soft-clustering approach by Roung-Shiunn Wu , Po-Hsuan Chou.
5. 288175101.pdf (core.ac.uk) CUSTOMER SEGMENTATION ANALYSIS OF CANNABIS RETAIL DATA: A MACHINE LEARNING APPROACH By Ryan Henry Papett.
6. Ortega, Joaquín & Almanza-Ortega, Nelva & Vega-Villalobos, Andrea & Pazos-Rangel, Rodolfo & Zavala-Diaz, José Crispin & Martínez-Rebollar, Alicia. (2019). The K-Means Algorithm Evolution. 10.5772/intechopen.85447.
   https://www.researchgate.net/publication/332340076_The_K-Means_Algorithm_Evolution.
7. K-means clustering, https://en.wikipedia.org/wiki/K-means_clustering#cite_note-3.
8. Kalra, Monika & Lal, Niranjan & Qamar, Shamimul. (2018). K-Mean Clustering Algorithm Approach for Data Mining of Heterogeneous Data. 10.1007/978-981-10-3920-1_7.
9. Aloise, D., Deshpande, A., Hansen, P. *et al.* NP-hardness of Euclidean sum-of-squares clustering. *Mach Learn* 75, 245–248 (2009). https://doi.org/10.1007/s10994-009-5103-0
10. Andrea Vattani. The hardness of k-means clustering in the plane. manuscript, 2009.
11. Mahajan, Meena & Nimbhorkar, Prajakta & Varadarajan, Kasturi. (2009). The Planar k-Means Problem is NP-Hard. Theoretical Computer Science. 442. 274-285. 10.1007/978-3-642-00202-1_24.
12. Arthur, David; Vassilvitskii, Sergei (2006-01-01). *How Slow is the* k*-means Method?*. *Proceedings of the Twenty-second Annual Symposium on Computational Geometry*. SCG '06. New York, NY, USA: ACM. pp. 144–153. doi:10.1145/1137856.1137880. ISBN 978-1595933409. S2CID 3084311.
13. Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A *k*-Means Clustering Algorithm". *Journal of the Royal Statistical Society, Series C*. 28 (1): 100–108. JSTOR 2346830.
14. UCI data repository-Online Retail data https://archive.ics.uci.edu/ml/datasets/online+retail

15. Simply psychology-What does a box plot tell you?-https://www.simplypsychology.org/boxplots.html
16. What is the importance of histogram?-https://socratic.org/questions/what-is-the-importance-of-a-histogram
17. Quality Advisor-histogram:Calculate descriptive Statistics-Skewness https://www.pqsystems.com/qualityadvisor/DataAnalysisTools/interpretation/histogram_stats.php
18. Log Transformation: Purpose and Interpretation by Kyaw Saw Htoon https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9
19. APA dictionary of psychology-square-root transformation- https://dictionary.apa.org/square-root-transformation
20. Statistics how to- Box Cox Transformation- https://www.statisticshowto.com/box-cox-transformation/
21. Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization by Aniruddha Bhandari- https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/
22. Vaishali Advani. 34 Open-Source Python Libraries You Should Know About https://www.mygreatlearning.com/blog/open-source-python-libraries/
23. Customer Retention Should Outweigh Customer Acquisition https://www.retentionscience.com/blog/customer-retention-should-outweigh-customer-acquisition.
24. Customer Churn Analysis: One of SaaS's Most Important Processes. https://www.profitwell.com/customer-churn/analysis
25. Varda Tirosh; Behavior-based Customer Segmentation for More Effective Retail Marketing; https://www.optimove.com/blog/customer-segmentation-for-more-effective-marketing
26. Abdul Haseeb Ahmad. Focus Strategy – Definition, Types & Examples. https://www.marketingtutor.net/focus-strategy/
27. Advantages and Disadvantages of Market Segmentation https://accountlearning.com/advantages-and-disadvantages-of-market-segmentation/
28. Kernel Density Estimation https://deepai.org/machine-learning-glossary-and-terms/kernel-density-estimation
29. Abdullin, Artur & Nasraoui, Olfa. (2012). Clustering Heterogeneous Data Sets. Proceedings - 2012 8th Latin American Web Congress, LA-WEB 2012. 1-8. 10.1109/LA-WEB.2012.27.
30. Customer sentiment analysis guide: Improve satisfaction https://www.sentisum.com/customer-sentiment-analysis
31. David Robinson. K-means clustering is not a free lunch http://varianceexplained.org/r/kmeans-free-lunch/
32. https://www.quora.com/Why-does-K-means-clustering-perform-poorly-on-categorical-data-The-weakness-of-the-K-means-method-is-that-it-is-applicable-only-when-the-mean-is-defined-one-needs-to-specify-K-in-advance-and-it-is-unable-to-handle-noisy-data-and-outliers
33. Geeks for geeks-Python-Pandas.melt()-https://www.geeksforgeeks.org/python-pandas-melt/