***Online Retail Analysis***

**Submitted to**: Mr Rajeev Nath

**Submitted by**: Simarpreet Kaur

**Module Code**: B005

|  |
| --- |
| Table of Contents |
| [INTRODUCTION 3](#_Toc68103086)  [UNDERSTANDING CUSTOMER SEGMENTATION 3](#_Toc68103087)  *BUSINESS PROBLEM……………………………………………………………………...6*  *PROJECT FLOW……………………………………………………………………………7*  *EXPLORING THE DATA…………………………………………………………………..8*  *CONCLUSION………………………………………………………………………………..*  *SUGGESTIONS……………………………………………………………………...*  **BIBLIOGRAPHY……………………………………………………………………………** |

**INTRODUCTION**

Most customers mean a lot of business. Intuitively, it is safe to

assume that a company's ability to attract a growing number of customers has a direct relationship with its growing sales. However, just an increase in the number of customers may indicate that it is a trivial metric, which means it will not fulfil its potential to run the business. In addition, finding or retaining the right customers can be very effective in increasing sales and saving money instead of wasting unreasonable resources on lost causes.

# With the help of our data, we will try to generate information that helps the complete inventory station. The data we will be analyzing is that of a UK online retailer that specializes in gift products. The company's customer base is mainly the best sellers. These whole-sellers are bulk products, usually with low prices.

# To create a successful wholesale strategy, let’s take a step-by-step description of creating customer segregation in Online Retail Dataset using python in jupyter notebook, understanding purchase patterns using RFM (Recency, Frequency and Monetary) analysis and clustering.

**Consumer Segmentation: An Overview**

Cosumer division is the classification of consumers into classes based on common behaviours or other attributes. Groups should be homogeneous between themselves and against one another. The aim of this method is to find consumers that have higher growth opportunities or are more profitable.

**How do you set yourself apart from your competitors?**

When it comes to consumer segmentation, a company must have a clear strategy and goal in mind. The following steps should be taken to obtain components from a broader customer base.

1. Analyse the existing user base, taking into account considerations such as geographic delivery, consumer preferences and values, and empirical input from the web search page, among others.
2. Raise each client's awareness: Map each customer's preferences in order to anticipate and understand their behaviour: products, services, and material that might be of interest to them.
3. Describe the Segment's Opportunities: After the Components have been established, each segment's industry, as well as its challenges and opportunities, should be thoroughly examined. A wide spectrum of clients will be used in the firm's overall marketing campaign.
4. Analysis section: After assessing the definition and market value of different types, a business must understand how to adjust its products or services in order to better serve customers. In contrast to others, it can choose to give higher discounts to some customers in order to increase the base of active clients.
5. Strategy Development: Examine new approaches to understand the impact of customer buying time and savings at various stages. The method can be replicated as long as the procedure is optimized as far as possible.

**PROBLEM IN BUSINESS:**

An e-commerce business needs to distinguish its clients and develop marketing campaigns focused on these characteristics.

With this in mind, we will characterize consumer actions and divide them into classes based on their beliefs.

To put it another way, we can group people who use common sense together and attempt to come up with marketing and communications tactics for them.

**The below is the methodology:**

The store in question is a global corporation headquartered in the United Kingdom that provides customers with convenience shopping services. The business works in a variety of ways. Some of the stores are managed and managed by the company, while some are run by a distributor or a supplier.

The data for four separate European economies were examined in this article. The research was conducted over a one-year period. Transaction data, commodity data, shop data, loyalty data, and rival data are among the data types used.

• Attributes such as Invoice No., Invoice Year, Amount, and so on are collected in transaction details.

• The product dataset includes attributes such as unit price, product definition, category description, name, and so on.

• Store data includes information such as store size, location, and operating channel.

• The loyalty data captures information about consumers such as their ID, purchasing behaviour of customers who use the loyalty card, payment processes, rewards, point redemption, and so on.

**Techniques for Clustering:**

Clustering methods are divided into two categories:

• Hard clustering: A data point refers to exactly one cluster in hard clustering. The hard clustering description involves K-means and hierarchical clustering.

• Soft clustering: A data point has a chance of belonging to all clusters by using soft clustering. Weak clustering grouping includes self-organizing maps (SOM), Gaussian Mixture Models (GMM), and Fuzzy C-means (FCM).

However, the clustering algorithm is best for you is determined by your company's needs. Let us take a look at each of them and then determine which one is best for us.

**1**.**Partition based Clustering**

* • A partition-based clustering algorithm is an unsupervised method for grouping data points around a central point known as the centroid. Partition-based clustering divides groups into groups based on the distance between data points and assigns them to one of a collection of clusters. It's important to remember that each data point can only be assigned to one cluster at a time. A partitioning algorithm creates an initial grouping of data points before doing iterations/reallocations depending on the set's mean or median data points before it meets the final groupings.
* • Since it is simple, reliable, and versatile, K-means clustering is a common partitioning algorithm. You must choose the number of clusters (k) that you want to use while using this algorithm.
* • You can use the ELBOW METHOD or industry-relevant experience for the specific product segment to decide the optimum number of clusters for your market.

# • After the k-means algorithm runs, each data point is assigned to the nearest centroid, ensuring that data points from various clusters are separated by a wide enough radius. The algorithm repeats iterations until the data points are allocated to their final clusters and an average (mean) is reached.

# • When you have categorical data, this sort of algorithm is ideal (e.g., grouping based on category, subcategory and brand).

* •
* The k-means algorithm is modular, easy to use, and easier to run than hierarchical clustering. It can handle massive data sets like those used in retail POS and loyalty data, and it generates well-separated clusters that are lightweight and well-defined.

**2. Hierarchical clustering:**

# • A hierarchical clustering algorithm is an unsupervised method for grouping data points into a dendrogram, which is a hierarchical tree. You have the luxury of deciding how many clusters you want to create. This determines the interval between clusters as well as the grouping distinctions. The accuracy of the hierarchical clustering algorithm improves as the value of k increases.

• When working with a random dataset, it's better to use a hierarchical algorithm. This algorithm can be used in classification and research and development. A hierarchical algorithm that determines the department, division, subcategory, section, and so on can be used to classify products into a category hierarchy. However, since all types of customers are important to our business, we don't use this clustering algorithm.

**3. Fuzzy clustering**

# Soft clustering, also known as fuzzy clustering or soft k-means clustering, is an unsupervised approach. Each data point can be assigned to several clusters using this algorithm. Since you must select the number of clusters you want to build, this algorithm is close to the k-means algorithm. The algorithm repeats itself before a final classification is found.

# Fuzzy clustering can be used in marketing applications. Consumers may be divided into groups based on their likes, desires, shopping habits, LSM, and psychographic profiles.

# Data points can be assigned to several clusters using fuzzy clustering. This is a more natural reflection of customer behaviour, and knowing complex consumer behaviour will give retailers who use this strategy a competitive edge.

# However, this algorithm has the disadvantage of requiring you to choose the number of clusters you want to build as well as a cut-off value for membership in the groupings. The initial positioning of the centroids is also taken into account by this algorithm. Since we don't know ahead of time what kinds of consumers, we'll be able to get from a given dataset, this approach could be unsuccessful. Furthermore, the dataset being fed to the model is extremely complex, so our model must be adaptable enough to cope with it.

# We may infer that there is a plethora of other clustering algorithms available, each with its own set of advantages and capabilities, that you should investigate. Each algorithm is ideally suited to a specific business challenge and set of data. It could be beneficial to test the algorithms on your datasets before choosing which is best for you. During my research, I discovered that K-Mean clustering is the perfect strategy for our current needs, since we need to classify related classes and there is no clear estimation or classification result.

**Project Flow**

In this demonstration, we will be processing and analyze the database of online retail stores. Full details of each step are provided within the notebook. The summary of this project is as follows:

* The first step is data cleaning and data understanding.
* The second step is EDA and data visualization where we will inspect the data for useful information.
* In the last step we will use clustering for customer segmentation.

• Once you've finished cluster research, you'll be able to classify market groups. You can do this by clustering the customers for a certain product segment using geographical, psychographic, and behavioural data, as well as output data.

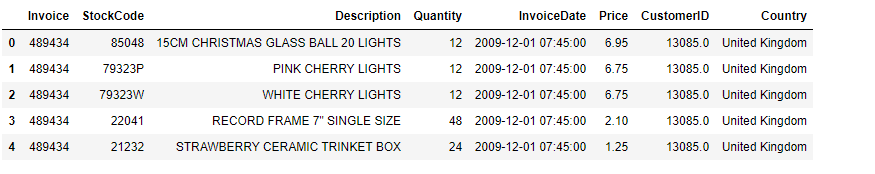
•

Following that, we'll profile our clusters in order to better understand our customers and identify them using the cluster analysis variables. We will use this data to customise marketing communications, product assortments, and the overall shopping experience to satisfy the demands of your customers and maximise your return on investment.

• We'll use clustering to analyze and group customer turnover in order to profile and identify customers based on retention. To analyze and forecast retention rates in specific customer groups, we can use variables such as frequency of sales, how recently the buyer visited the store, average spend per ride, and basket structure.

**EXPLORING THE DATA**

The Online Retail Data Set, which was discovered in the final area of UCI Machine Learning, will be used in the following study. The data provide details on international web marketing operations carried out by UK-based companies. The online retail dataset can be found [here](file:///C:\Users\Hp\Desktop\online_retail_II.xlsx).



**Data Attributes**

**Dataset Description:**

• It includes content procurement invoices. There are 541909 invoices from 4372 individual consumers in 38 different countries.

• The transfers were completed between December 2011 and December 2012.

**Characteristics of the Dataset:**

• **Invoice**- Invoice number, a 6-digit integral number uniquely allocated to each account, is one of the dataset's eight attributes. This code implies a cancellation if it begins with the letter "c."

**• Stock Code**- A 5-digit integral number uniquely assigned to each distinct product; Stock Code is a product (item) code.

• **Description-** Product (item) name.

• **Quantity**—the number of units of each commodity (item) sold in a sale.

• **Invoice Date** is the date of the invoice. Time and date. The date and time of each transaction's formation.

• **Price-per-unit** cost of the product

• **Customer ID**- This is a five-digit integral number that is unique to each customer.

• **Region**- The country in which each consumer lives.

**Data Collection:**

Retail consumer data obtained at the point of sale is not only a fantastic solution for inventory control, but it also provides some great advantages that cannot be overlooked. A point-of-sale device illuminates metrics like profit margins, buyer numbers, pricing trends, and more, allowing you to see which goods and vendors are generating revenue. As a result, it is easier to make more reliable choices. You may do the following with data collected at the point of sale:

• Sort consumers according to their brand preferences, spending levels, purchase categories, and so on.

• Forecast orders and monitor inventory accordingly.

• Streamline the cash flow and other store processes by further identifying shopper tastes and desires of products.

Taking Online Surveys – Asking your customers what they think of your goods and services is the best way to find out! Online polls allow you to get a better understanding of your consumers' viewpoints and think like them. This enables the retailing departments to make strategic choices that are error-free.

While consumers generally avoid taking surveys to provide input, you can use an online survey platform to make your surveys and survey designs exciting to pique their interest. Online survey software not only allows you to produce intelligent and innovative surveys, but it also allows you to create dynamic analyses that give you a bird's eye view and encourage you to do meaningful research.

**Here are few things to keep in mind when creating an online survey:**

**• Make sure the survey has a definite goal.**

**• Make a good questionnaire by starting with standardised questions and ending with personal demographic questions.**

**• Inform the audience of the survey's intent and the time it will take to complete it.**

**• Describe how it would help the audience.**

**Foot Traffic and Web Analysis – If you don't keep track of your numbers, you may be losing out on a chance to succeed. Analysing foot traffic in brick-and-mortar shops is the safest and fastest way to go. Shopper traffic monitoring allows you to keep track of the overall number of customers to your shop as well as their events. You may also calculate other details based on this count, such as:**

**• Translation rate.**

**• The most famous sections of the shop.**

**• The most popular days and hours for customers to visit the shop.**

**• Transactions made on the basis of sales and discounts.**

**• Patterns in shopping for special events.**

**Market analysis helps you to gather data using a market research tool to see how your goods are meeting the demands of your clients and shoppers. Market analysis will help you develop actionable information on your clients, rivals, and target market, offering you a strategic advantage. It entails identifying the right audience, asking the right questions, collecting useful information, and taking corrective action.**

**Market analysis aids retailers in gauging demand for their offerings, identifying target markets, assessing the market and competitiveness, gauging consumer and purchasing habits, and introducing creativity.**

Social Media – Social media has evolved into one of the most credible data providers. What you have to do now is find out which platforms your target audience is most active on. Then, using surveys and social media groups, you will learn what your target audience thinks about your goods and services. The advantages include:

• Knowing your consumers' most recent requirements

• Knowing what your consumer thinks about your rivals

• Knowing your audience better and then anticipating demand, and so on.

**Data Storage**:

Suppose state of affairs inquiry could be an employment designing, moreover modelling system want to give varied predictions for a few consequences supported by selection dynamical information. Now, we are using WISE for the data storage because our retail company is global brand and we are collecting millions of GBs of data every day from around the globe. But we need to have contingency plan as in today’s world data is proving more valuable than gold, and this data for the consumer-based company which solely depends on the behaviour of their customers, so it cannot afford to lose the its customers data at any cost.

Cloud Storage seems very feasible option to keep the company safe from above scenario. But why cloud storage is one of the major questions that comes in mind. The answer is crystal clear and is as follows:

Nowadays, controlling, saving plus studying the facts and figures has become an intensive piece of work. Businesses are rapidly embracing cloud computing in order to efficiently handle and process data in order to provide real-time insights. Retailers must concentrate on expanding their access to consumers through many stores and controlling supply chains and have a unique consumer service to stay afloat in a fiercely competitive market. Incorporating cloud computing services into the retail market not only lowers IT costs, but it also streamlines processes, improves performance, and improves the end-user experience.

We can take benefit of Cloud-Enabled Disaster Management System.

We all know that data is the most important resource of the modern era, so we must still take the necessary precautions to secure it. Previously, data was recovered manually, which was time-consuming and necessitated ongoing servicing and assistance. You can store data in the cloud disaster recovery model, which can be accessed from any location and at any time. A cloud storage and disaster recovery can be started from anywhere, with each phase of the process being automated. Cloud storage disaster recovery has been the most convenient choice for all companies thanks to its scalability, cost-effectiveness, and durability. Cloud disaster management employs remote cloud data centers to assist company processes that are disrupted by a variety of incidents.

**Data Manipulation:**

The data must be interpreted at the store level in order to achieve store clustering. Following the collation of the datasets, all variables are rolled at the store stage. Aggregation methods such as number, count, cap, and min are used depending on the type of the element. In the case of sales, for example, an amount number is computed; but, in the case of purchases, a separate count is computed. Many derived variables are generated, such as spending per segment, total price, revenue for various months, weeks of the year, and times of the day. These factors offer a comprehensive view of stores, capturing dimensions such as profiles, firmographics, transaction patterns, and purchasing patterns, among others.

A cleansing process is used to ensure that only high-quality data is used for clustering. The phases are as follows.

• The percentage ranges (0.01, 0.05, 0.1,0.25, 0.5, 0.75, 0,9,0.95,0.99), as well as the count of missed values, was calculated using a univariate study.

• Missing value imputation methods such as mean/median/mode substitution are used depending on the type of the attribute.

• Factors with a substantial number of missed values are not used in the study.

• Factors with limited uncertainty are also omitted.

• The outlier therapy is the last move. The procedure is carried out based on the distribution of the element. For certain factors, the outlier at the upper end is replaced by the 95th percentile value, and for others, a different threshold is used.

The review does not contain any of the stores. Only stores that are owned by the company and are open for more than 80% of the time are accepted.

**Consider the following:**

We'll divide the clients into groups based on their behaviour. We'll use three factors to do so: the most recent shop visit, the monthly buying volume, and monetary values.

I'll compare the behavior of two cycles over the course of 13 months of buying activity. The first is to establish a baseline of behavior, and the second is to assess it. We'll use 8 months for the first and 5 months for the second. It would take up 60 percent and 40 percent of the day, respectively.

In terms of technicalities, K-Means is a good algorithm for finding data similarities. Separating the data into potential classes would be large enough, but not so large that minor differences would be noticeable.

Since the store's aim is to maximize either the frequency of purchases or the amount invested, the categories can be categorized as follows:

**• Dormant** — There have been very little trips and very little money invested. To reactivate them, they would require a lot of stimulus.

• **Capacity** — Similar to the previous category, with the exception that the stimuli to enhance their customer experience would be lower.

• **Medium** — A typical consumer.

• **Irregular** — They only come once or twice a year, but when they do, they spend a lot of money.

• **Loyal** — They frequent the store but do not spend a lot of money.

Depending on the company's plan, the names and group numbers can change. It's vital to provide clear coordination and understanding between data scientists and stakeholders in order to better convert market criteria into models.

# Data Visualization

For any online merchandiser achieving customer’s fulfilment is the means for staying prior the competition. At the side of providing distinctive looking experiences for the web client, knowledge analytics has been used in online retail in order to switch from a vast retailing scheme to the practical approach.

Data visualization is a basic method of viewing information. It assists you in interpreting for clients, allowing you to enhance your marketing plan and interactions with them, as well as improve their shopping experience and meet their needs.

For any Online retail firm, data analytics gives many advantages. Some of them are:

* Enhance the sale’s income.
* Enhance the availability of series administration.
* Execute higher production valuation ways.
* Targeting the proper client for the promotions and further suggestions of the product.
* Knowing the merchandise line for the interests of the client towards their products in which they are interested to buy.

From the dataset, analysis is done and found that **85123A** is the **Most Popular Stock Code.**

The company receives the maximum number of orders from the consumers of the country United Kingdom (UK).

Henceforth, Top 6 countries according to the Quantity Sold Online are:

* United Kingdom
* Denamrk
* Netherlands
* Eirie
* France
* Germany

# RFM Segmentation-

Recency, Frequency, and Monetary are acronyms for Recency, Frequency, and Monetary. It's a brand research technique that uses those metrics to determine a company's or organization's best consumers. Three quantitative considerations underpin the RFM model:

• Recency: The time before a consumer made a transaction.

• Frequency: The number of times a consumer buys something.

• Monetary Value: The amount of money spent by a buyer on sales.

Marketers should use RFM research to find answers to the following questions:

• Who are the most valuable customers?

• Which of your clients could be a factor in your high churn rate?

• Who are the people who have the ability to become valuable customers?

• Which of your clients are you able to keep?

• Which of your clients are the most open to engagement campaigns?

Creating ‘quartiles' on each of the parameters and assigning the appropriate order is a standard procedure after having the RFM values.

We will sort the customers according to their RMF ratings for our own convenience:

• Champions: they are the most loyal clients who buy often and invest a lot of money. Reward these clients. They will help market the brand by being early adopters of new products.

• Potential Loyalists: These are recent consumers who have purchased from you on a regular basis and have invested a significant amount of money. To upsell them and make them become your Loyalists or Champions, provide membership or loyalty plans or suggest similar items.

• New Customers: These are customers who have a high RFM score overall but aren't regular shoppers. To maximise their visits, start establishing relationships with these customers by offering onboarding assistance and special deals.

• At-Risk Consumers: These are the customers who used to buy regularly and spend a lot of money, but haven't bought in a while. Send them customised reactivation campaigns to reconcile with them, as well as renewals and useful items to entice them to make another buy.

• Can't Lose Them: They are people who used to come in and shop often but haven't done so lately. Bring them back with relevant incentives and conduct polls to figure out what went wrong to prevent them from going to a rival.

**K-Means Algorithm Preprocessing:**

K-means is a common clustering algorithm that is commonly used in unsupervised learning.

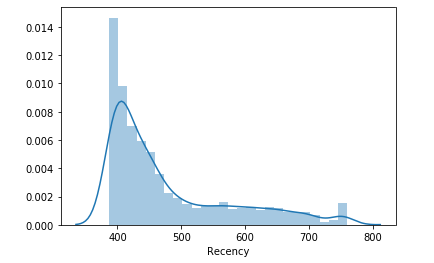
We need to realise that the algorithm makes some predictions about the data for our purposes. As a result, we must process the data in order to satisfy the algorithm's key principles, namely:

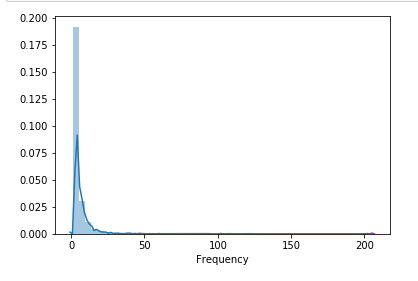
• The variables should be assigned in a logical order

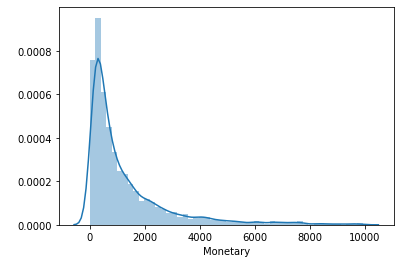
• The variable's median values must be the same

• The variable's standard deviation values must be the same.

The assumption is tested by using the seaborn library to build histograms of Recency, Frequency, and Monetary variables:



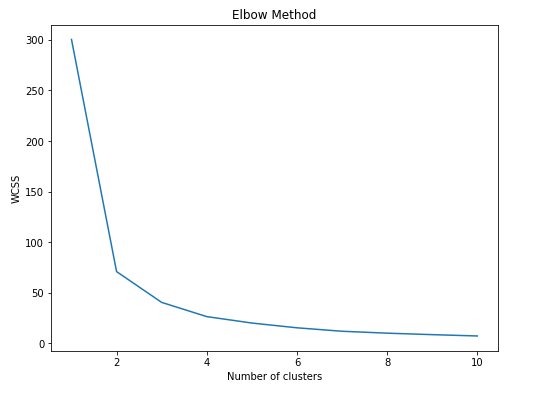




**Clustering with K-Means:**

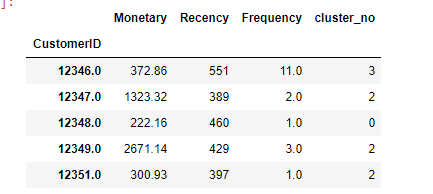
In this part, we'll use the elbow method to create clusters from our normalized RFM data and try to find the optimum number of clusters in our data.

The ideal number of clusters is 3 or 4, as seen in the graph below.



Some methods can assist users in estimating the number of clusters that better suit their needs. The elbow system is what it's called. The elbow process curve is represented graphically above.

As a final step in this study, we can now provide each customer with the following facts, which can be used to map a customer's perceptions of company values:



**FINAL REMARKS**

We can see from the above review that we have combined Customer IDs based on three variables over RFM metrics.

Using k processes, this data collection was reduced to four sets based on market variations.

About the fact that we couldn't find two distinct categories, we were able to develop a model that could categorise potential consumers into "low cost" and "high value" groups.

**K-Means Clustering's Limitations-**

* There is no guarantee that it will lead to the best solution possible.
* Where the clusters have complex geometric forms, k-means fails to cluster the data effectively.
* You can't compete with a particular form (one that isn't round) and search for possibilities for one point to appear in several sets.
* Because it tries to reduce within-cluster variation, the K-means algorithm does not allow data points that are far apart to share the same cluster, even though they clearly belong to the same cluster.
* Because it tries to minimize within-cluster variation, it gives greater weight to larger clusters than smaller clusters. To put it another way, data points in smaller clusters can be moved away from the centroid so that more attention may be paid to the larger cluster.
* The negative k-means indicate that for many databases (especially low-level data sets), it may not work as well up to the expectations.

# SUGGESTIONS

How we continue with this study is decided by how the company wants to use the conclusions and the degree of hypocrisy that business members hope to see in phases.

**Identifying New Clients –**

• The shop should be on the lookout for potential vendors who are interested in a certain category of gift items. A strategic research should be performed in some way to determine who their clients are.

• New consumers learn about your goods and services in a variety of formats and from a variety of outlets, which leads to sales. When they are about to shop, the more they know from you, the more likely they are to accept what you have to sell.

* **Building Trust** –
* The more a client has a relationship with you, the more important you get. This is contingent on how the shop establishes credibility in all sets of customers (new and repetitive).
* It is important to have dedicated support personnel and to set high expectations for the speed and consistency of the operation.
* Reliability can be increased by resources such as prompt distribution, quick recovery, and special discounts for current clients, as well as ensuring product quality, etc.
* Regardless of the specific strategy you use, hearing from other customers will make potential customers feel more at ease about giving your company a shot — which will help them trust your brand right away.
* One of the most critical aspects in achieving consumer satisfaction is honesty. This entails being open and frank about what they should expect from your company. Ensure that all of the clients are happy with their service is much more crucial than creating as much profits as possible by being as honest as possible about what you have to sell and setting accurate consumer preferences from the outset.
* **Increasing Purchase Frequency** :
  + Provide recommendations to potential retailers and buyers who have been very successful and do research on their market needs.
  + You should always encourage repeat purchase in the name of festivity, yearly occasions like birthday and anniversary so that your store and worthwhile buying could be remembered by your customers in the long run.
  + Instead of just waiting for the next customer’s call or query, or simply getting new sign-ups for the program where you supply your product or service automatically at predefined intervals.
  + You can offer lower costs for pre-instalment or for the long haul or “till further notice” duties or you can offer contract costs to high volume clients. You can give value motivating forces, rewards, premiums, redesigns, discounts, or credits to clients who assure you to maintain a frequency of purchase of a base volume in a year.

# BIBLIOGRAPHY:

1. [**https://learn.datacamp.com/courses/customer-segmentation-in-python**](https://learn.datacamp.com/courses/customer-segmentation-in-python)
2. [**https://www.shopify.in/encyclopedia/customer-segmentation**](https://www.shopify.in/encyclopedia/customer-segmentation)
3. [**https://www.intercom.com/blog/customer-segmentation/**](https://www.intercom.com/blog/customer-segmentation/)
4. [**https://looker.com/blog/creating-actionable-customer-segmentation-models**](https://looker.com/blog/creating-actionable-customer-segmentation-models)
5. [**(368) Customer Segmentation using RFM K-Means & Python | Who are your Loyal Customers ? - YouTube**](https://www.youtube.com/watch?v=fdUofaT8gUw)
6. [**https://www.business2community.com/customer-experience/4-types-of-customer-segmentation-all-marketers-should-know-02120397**](https://www.business2community.com/customer-experience/4-types-of-customer-segmentation-all-marketers-should-know-02120397)
7. [**https://www.youtube.com/watch?v=IpDPFOLsn0k**](https://www.youtube.com/watch?v=IpDPFOLsn0k)