A Project Report
On

CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING

Submitted in partial fulfillment of the requirements for the award of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

By

SREERAM MEGHANA (19BQ1A12F3)

SHAIK SUPHIYA NAWAZ BANU (19BQ1A12E9)

POOJITHA MANNE (19BQ1A12D2)

SINGAVARAPU RISHITHA (19BQ1A12F1)

Under the esteemed guidance of

Sk. MULLA ALMAS ***MTech (Ph.D.)***

ASSISTANT PROFESSOR

VVIT
VASIREDDY VENKATADRI
INSTITUTE OF TECHNOLOGY

Department Of Information Technology

Vasireddy Venkatadri Institute of Technology
Jawaharlal Nehru Technological University, Kakinada, AP, India VASIREDDY
VENKATADRI INSTITUTE OF TECHNOLOGY: NAMBUR

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY
NAMBUR

BONAFIDE CERTIFICATE

This is to certify that this project report is the bonafide work of **SREERAM MEGHANA, SHAIK SUPHIYA AWAZ BANU, POOJITHA MANNE, SINGAVARAPU RISHTHA** Reg No: **19BQ1A12F3,19BQ1A12E9,19BQ1A12D2,19BQ1A12F1** who carried out the project entitled **"CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING"** under our supervision during the year 2022.

SIGNATURE OF SUPERVISIOR

Sk. MULLA ALMAS , ***MTech (Ph.D.)***

Assistant Professor

HEAD OF THE DEPARTMENT

Dr. KALAVATHI ALLA

Professor & HOD

External Viva voice conducted on ___________________

Internal Examiner

External Examiner

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY: NAMBUR

CERTIFICATE OF AUTHENTICATION

I Solemnly declare that this project report **“CUSTOMER SEGMENTATION USNG K-MEANS CLUSTERING”** is bonafide work done purely by us, carried out under the supervision of Ms. **Sk. Mulla Almas** towards partial fulfillment of the requirements of the Degree of **Bachelor of Technology** in Information Technology from Jawaharlal Nehru Technological University, Kakinada during the year 2022.

It is further certified that this work has not been submitted, either in part or in full, to any department of the Jawaharlal Nehru Technological University, Institution or elsewhere, or for publication in any form

Signature of the Student

DECLARATION

We **SREERAM MEGHANA, SHAIK SUPHIYA NAWAZ BANU, POOJITHA MANNE, SINGAVARAPU RISHITHA** hereby declare that the project report entitled "**CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING**" done by us under the guidance of Sk. MULLA ALMAS, ASSISTANT PROFESSOR at **VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY** is submitted in fulfillment of their acquirement for the award of a degree in **INFORMATION TECHNOLOGY**. The results embodied in this project have not been submitted to any other university or college for the award of any degree or diploma.

DATE:

PLACE: NAMBUR

SIGNATURE OF CANDIDATES:

(SREERAM MEGHANA-19BQ1A12F3)

(SHAIK SUPHIYA NAWAZ BANU 19BQ1A12E9)

(POOJITHA MANNE –19BQ1A12D2)

(SINGAVARAPU RISHITHA-19BQ1A12F1)

ACKNOWLEDGEMENT

First, we would like to express our sincere gratitude to our beloved Chairman, **Sri V. VIDYASAGAR**. We would be grateful to our beloved Principal, **Dr. Y.**

MALLIKARJUNA REDDY, and we would be grateful to our beloved Director, **Dr. NAVEEN RAVELA**. We would be grateful to our beloved Dean of Academics**, Dr. N. Kumara Swamy**. And people who made this project work easier with words of encouragement, motivation, discipline, and faith by offering different places to look forward in expanding our ideas and helping us towards the successful completion of this project work.

We take this opportunity to express our deepest gratitude and appreciation to our HOD**, Dr. A. KALAVATHI**, Professor, Guide, **SK. MULLA ALMAS, Assistant professor** and Project Coordinator **MR. KRANTHI KUMAR, Assistant Professor,** Department of Information Technology, for their motivating suggestions, insightful advice, invaluable guidance, help, and support in the successful completion of this project, and also for constant encouragement and advice throughout our B. Tech program.

We would like to take this opportunity to express our thanks to the **teaching and non-teaching staff** of the Department of Information Technology, VVIT for their valuable help and support in these four years of our study. We would like to thank the Department of Information Technology, VVIT for providing all facilities to complete this project work. We are also grateful to all our classmates for their help, encouragement, and invaluable suggestions. We would be grateful to our parents for their incessant support and cooperation.

Finally, we would like to thank all those whose direct and indirect support helped us in completing our project work on time

ABSTRACT

Customer segmentation is one of the key aspects of business-driven companies . Identifying potential customers is important to know where to put the company's efforts. The majority of companies run on the ability to retain customers , due to the abundance of products there is a high chance of risk customers may get confused and couldn't find the product desired. Customer segmentation focusing on the relationships not only between customers to customers, but also the interactions between customers and products can achieve highly promising results in the outcome. In this project deals with propose using the K-means Clustering algorithm to solve customer segmentation. The algorithm is applied to a dataset, which is first pre-processed. Data pre-processing is an essential stage in the Data mining process. Pre-processing the data improves the data quality, thereby improving the outcome of a data mining task. First the system will apply data integration methods such as normalization (Min-Max, Z-score) , ensuring that all the dataset attributes are given equal importance. Later the model will remove redundancies in the dataset using correlation analysis. Now the curated dataset is ready , which can be used for customer segmentation. This project will use the K-Means clustering algorithm to segment the data set into K dissimilar clusters depending upon the similarities of the attributes. Choosing the ‘K’ value in the algorithm plays an important role in segmenting clusters. In order to choose the optimal value of K, different methods such as Silhouette or Elbow method. The project aim to show the results after applying the above mentioned techniques.

LIST OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction:

In the current era of Business industry especially in the e-commerce and retail sectors, the competition between companies is cumulatively increasing along with a huge base of customers. Accordingly, the data of customers has also been increasing which lead to ideas to use this data for the benefit of the company. To stay in line with the heavy competition, obtaining useful insights from the data is one of the essential needs for a company. The effective use of this data directly or indirectly leads to effective decision making. Due to the high change, volume and variety of data ,the manual checking of the data is exhausting due to the era of Big Data, and expensive also requires a lot of human effort. The aim here is to automate the process of getting possible insights from data without human intervention. This is where Data mining techniques come into rescue, automating the entire process resulting in required information with no human bias. Data mining is the process of extracting knowledge from an existing database. Data Mining Techniques are implemented as programs are used to find the underlying or hidden patterns in the data. Data mining applications include Fraud Detection , Product Recommendations, Customer segmentation etc. Customer segmentation is the process of forming groups of customers with similar characteristics. It can be applied in a company is to understand the customers who have similar and distinct characteristics. Customers are grouped into clusters or segments, where each customer in a cluster in turn have similar characteristics with others within that cluster but rather have different characteristics of those from other segments. The importance of customer segmentation is the ability of a business to customize marketing strategies that would be appropriate for each segment of its customers. This identification and segmentation of customers happens using attributes such as demographics, number of items purchased , time spent on website, items purchased in past etc from the customer related database within the company. The promising and possible outcomes of customer segmentation are better understanding of customers, increase in sales, retain customers, improve customer experience , provide targeted ads and hence incrementing the overall marketing.

1.2 Literature Overview:

In [1] Tushar Kansal et al have worked on 3 different clustering algorithms (K-Means, Agglomerative, and Mean Shift) which have been implemented to segment the customers and finally compare the results of the clusters obtained from the algorithms. A python code has been developed and trained by applying scalar onto the dataset having 2 features of 200 training samples taken from a local retail shop. Through clustering , 5 segments of clusters have been formed labelled as Careless, Careful, Standard, Target, and Sensible customers.

However, two new clusters have emerged on applying mean shift clustering labeled as high buyers and occasional buyers.

The results have 2 internal clustering measures, the Silhouette score and Calinski- Harabasz index. In conclusion, As the data set was unlabeled they have opted for internal clustering validation rather than external clustering validation which depends on some external data like labels. In this case internal clustering validation which best suits the dataset and can correctly cluster data into opposite clusters.

In [2] Aman Banduni et al have identified customer segments into a commercial business using the data mining method such as customer segmentation. The data used in the paper were collected from the UCI Machine learning repository which contains 8 attributes. In this paper

several steps were taken to obtain an accurate result. It includes a feature with Centro’s first stage, allocation phase and update phase , which are the most common phase
K-Means algorithms.

The results showed that orange clusters have the highest value customers, green as the lowest value customers, blue and red as the high opportunity customers. Overall, as the dataset is unbalanced they have opted for internal clustering validation techniques rather than external clustering validation techniques which require extra data such as labels. Internal clustering validation techniques showed the best data cluster.

1.3 Software and Hardware Requirements:

Hardware necessities:

Hardware choice is essential to the standard and potency of any software package.
In Hardware choice, size and power necessities are necessary.
Customer isolation will be with success run on the system with AN i3 processor with a minimum of four GB RAM and disc drive with 500GB and fifteen.6 inches to observe system performance.

- Pentium processor -------- two GHz or on top of
- RAM capability -------- four GB
- Hard Disk -------- five hundred GB

Software necessities:

- Operating System: Windows seven or ten
- Software: Jupyter Notebook
- Databases: Excel sheets
- Python Libraries
- Packages

CHAPTER 2

SYSTEM ANALYSIS

2.1 Definition:

It is a process of collecting and interpreting facts, identifying the problems, and decomposition of a system into its components. System analysis is conducted for the purpose of studying a system or its parts in order to identify its objectives. It is a problem-solving technique that improves the system and ensures that all the components of the system work efficiently to accomplish their purpose.

2.2 Existing System:

In the business industry's current era, especially in the e-commerce and retail sectors, the competition between companies is cumulatively increasing along with a huge base of customers. Accordingly, the data of customers has also been increasing which leads to ideas to use this data for the benefit of the company. To stay in line with the heavy competition, obtaining useful insights from the data is one of the essential needs for a company. Currently, we are manually checking the data to obtain data insights. We depend on the tools like Excel, a tableau in order to get data insights. This requires a huge time and human effort.

Limitations of the Existing System:

More time-consuming:

Manual customer segmentation is time-consuming. It takes months, even years to analyze piles of data and find patterns manually. Also if done heuristically, it may not have the accuracy to be useful as expected. Customer segmentation used to be done manually and wasn't too precise. You'd manually create and populate different data tables, and analyze the data like a detective with a looking glass. Now, it's much better (and relatively easy thanks to rapid progress in ML) to just use machine learning, which can free up your time to focus on more demanding problems that require creativity to solve.

Hard to retrain:

Customer Segmentation is not a “develop once and use forever” type of project. Data is ever-changing, trends oscillate, and everything keeps changing after your model is deployed. Usually, more labeled data becomes available after development, and it’s a great resource for improving the overall performance of your model. Therefore manual segmentation is not at all suitable as there is a continuous change in the data.

Poor scaling:

This manual segmentation is not quite flexible for future changes and feedback. For example, consider a company that has 10000 customers today, and they’ve implemented a customer segmentation model. After a year, if the company has 1 million customers, then ideally we need to create a separate team to handle this increased data. Manual segmentation does not have the inherent capability to handle more data and scale in production.

2.3 Proposed System:

As we know, the effective use of customer data directly or indirectly leads to effective decision-making. Due to the high change, volume, and variety of data, the manual checking of the data is exhausting since it is very expensive and also requires a lot of human effort. The aim here is to automate the process of getting possible insights from data without human intervention. This automates the entire process resulting in required information with no human bias. The promising and possible outcomes of customer segmentation are a better understanding of customers, an increase in sales, retaining customers, an improving customer experience, providing targeted ads , and hence incrementing the overall marketing.

Advantages of the Proposed System:

High Scalability:

Scalability refers to the ability to scale up or down with demand. Employing a cloud infrastructure increases your ability to scale up your resources as you acquire and retain customers. Let's say an e-commerce platform is starting up and presently has ten thousand customers. After two years from now, it will have two million customers. Given this increase in customers, a team of data scientists without cloud infrastructure at their disposal would have to ask for additional expensive infrastructure to handle their work. But with cloud computing, this would not happen as they can just keep on working as usual and all the additional needs for infrastructure like servers, databases, and containers would scale up with demand.

Less time-consuming:

It is easier to apply machine learning with the support of the tools and software packages available to use and get insights that can be used to make decisions. Since Customer segmentation was previously done manually which took months or years. An analyst would acquire reams of data and populate different tables and then apply statistics to find patterns like a detective with a magnifying glass. The insights obtained with this approach would have become obsolete as new data reflecting changes in customer behavior and business conditions pours in

Better Insights on Customers for Acquisition and Retention:

Applying and optimizing the right machine learning clustering algorithms on your customer data could yield insights into underserved customer segments and could indicate potential strategies and actions to serve them better. Customer segmentation can also be used on CRM data to acquire new customers. For example, customer segmentation could indicate that men living in urban areas in their 20s tend to avail of a loan to buy their first or second vehicle. Sending the right marketing messages targeting the right customers of your competitors who fit in this segment with lesser interest rates can make a sizable percentage of the shift to your company's loan product.

Better Customer Experience and Loyalty

The intelligence derived from customer segmentation can target customers with appropriate content and offer them customized products and services specific to their needs and financial status to retain and grow your organization's relationship with them.

Benefits of Customer Segmentation in Retail

The landscape of the retail industry and the customer expectations associated with it has undergone a dramatic change and upheaval in the last few recent years. Instead, customers look and prioritize shopping with retailers who cater to their needs and ever-changing expectations in line with the changing trends in popular culture. In this new environment, customer segmentation has become the need of the hour for retailers to identify what their customers need and serve them effectively. The benefits for retailers including customer segmentation as a tool in the arsenal of their AI tech stack are listed here.

Better serve the customers

Customer segmentation is all about segmenting your customers into groups with similar needs. Defining your segments better will allow you to employ specialized marketing strategies and target those segments better by creating personalized incentives to shop for products of different ranges. For example, suppose that one of the segments is women under the age of thirty-five shopping for clothes at the lower end of the spectrum. This will allow you to create personalized marketing messages with discounts for clothes that are fashionable and are also priced appropriately for their respective budgets. This will incentivize the customers of that segment to shop with your brand as you create a personal connection by addressing their necessities.

CHAPTER 3

SYSTEM DESIGN

3.1 Definition:

The most creative and challenging face of the system development is System Design. It provides the understanding and procedural details necessary for the logical and physical stages of development.

In designing a new system, the system analyst must have a clear understanding of the objectives, which the design is aiming to full fill. The first step is to determine how the output is to be produced and in what format. Second, input data and master files have to be designed to meet the requirements of the proposed output. The operational phases are handled through program construction and testing.

Design of the system can be defined as a process of applying various techniques and principles for the purpose of defining a device, a process or a system in sufficient detail to permit its physical realization. Thus system design is a solution to ―how to approach to the creation of a new system.

This important phase provides the understanding and procedural details necessary for implementing the system recommended in the feasibility study. The design step provides a data design, an architectural design, and a procedural design.

3.2 System Architecture:

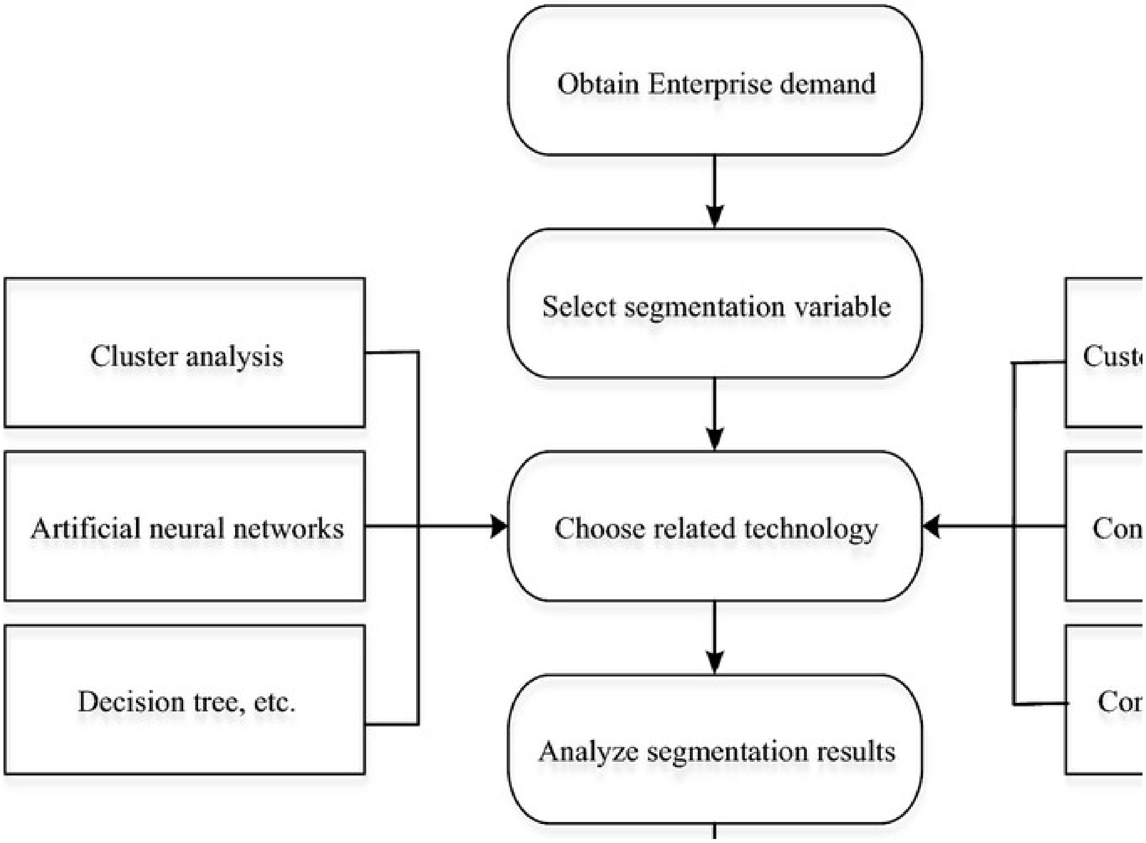

Obtain Enterprise demand
Select segmentation variable
Cluster analysis
Artificial neural networks
Decision tree, etc.
Choose related technology
Analyze segmentation results
Cust
Con
Cor

Figure: 3.1

3.3 Use Case Diagram:

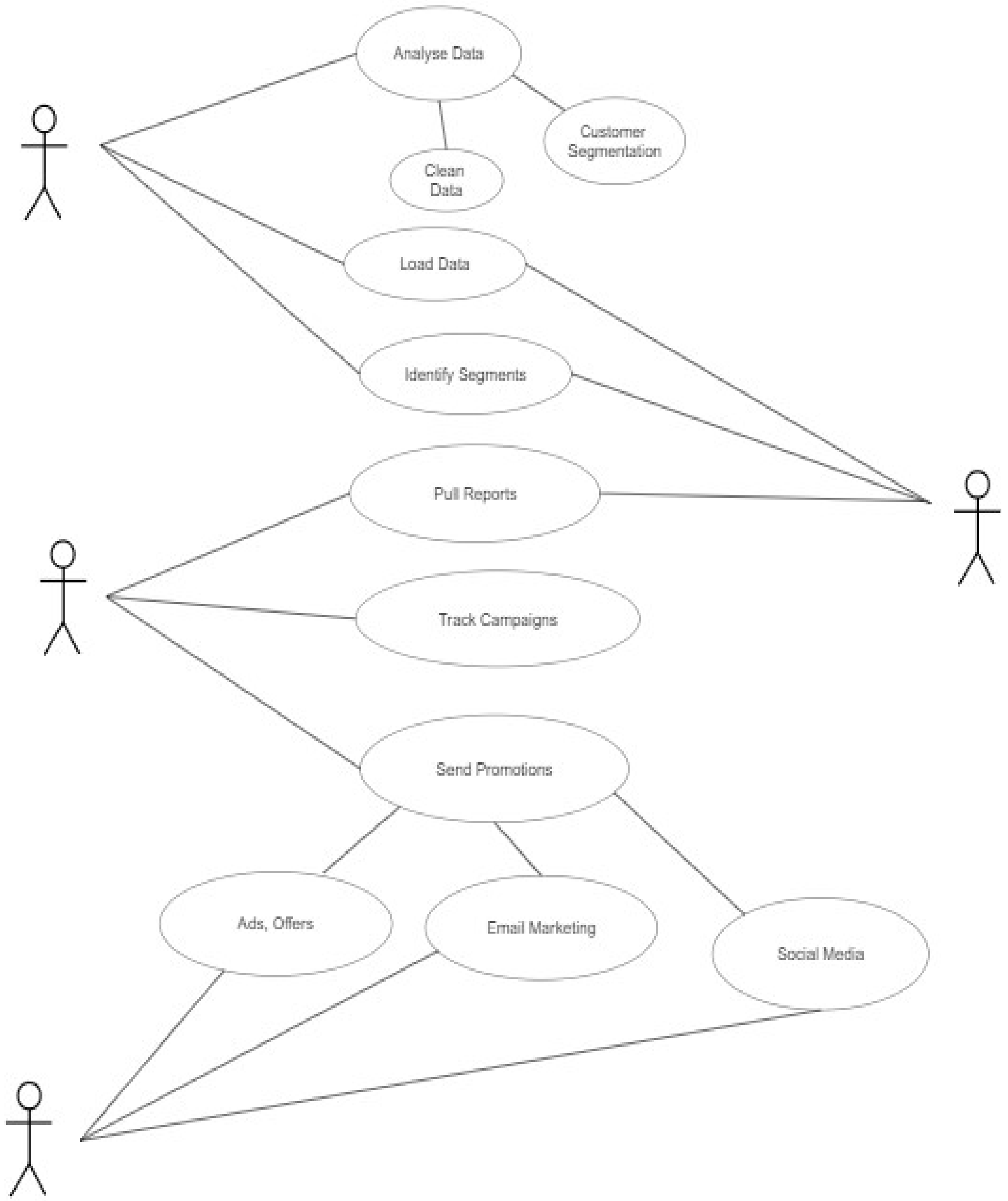


Figure: 3.2

3.4 Data Flow Diagram:

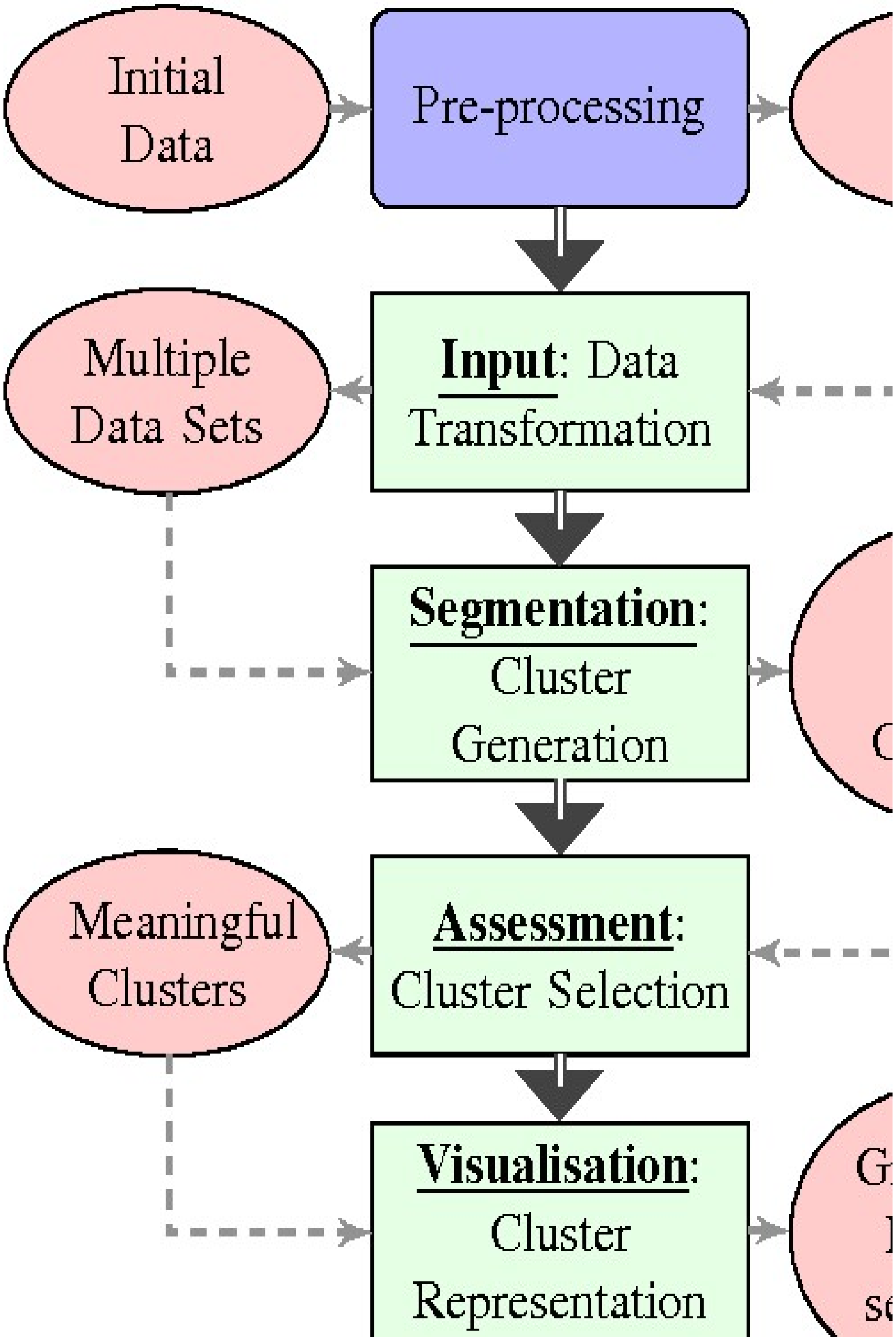

Initial Data
Pre-processing
Multiple Data Sets
Input: Data Transformation
Segmentation: Cluster Generation
C
Meaningful Clusters
Assessment: Cluster Selection
Visualisation: Cluster Representation
G
I
s

Figure: 3.3

CHAPTER 4

METHODOLOGY

4.1 Dataset Description:

The data set used to implement the Customer segmentation using K-Means clustering algorithm is the Online Retail II data set. This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers. The data set contains 8 attributes and has 525461 tuples, representing the data of 4384 customers.

4.2 Feature Description:

The attributes in the data set are Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price(Â£), Customer ID, Country.

- **Invoice No** : It consists of Invoice number. It is of Nominal data type. A 6-digit unique integral number assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.

- **Stock Code** : It consists of Product (item) code. It is of Nominal data type. A 5-digit integral number uniquely assigned to each distinct product.

- **Description** : It consists of Product (item) name.it is of Nominal data type.

- **Quantity** : It specifies each product (item) quantity per transaction. It is of Numeric datatype.

- **Invoice Date** : It consists of Invoice date and time when a transaction was generated. It is of Numeric datatype.

- **Unit Price** : It specifies Unit price. It is of Numeric datatype. Product price per unit in sterling (Â£).(1 Pound sterling equals 96.22 Indian Rupee)

- **Customer ID** : It consists of Customer number. It is a Nominal data type. A 5-digit integral number uniquely assigned to each customer.

- **Country** : It consists of country names. It is a Nominal data type. The name of the country where a customer resides.

4.3 Exploratory Data Analysis:

Before beginning the modelling work, EDA is used to see what the data can tell us. Deriving insights from raw data can be tiresome, uninteresting, and/or overpowering. In this case, exploratory data analysis approaches have been developed as a help.

Transactions count in each country

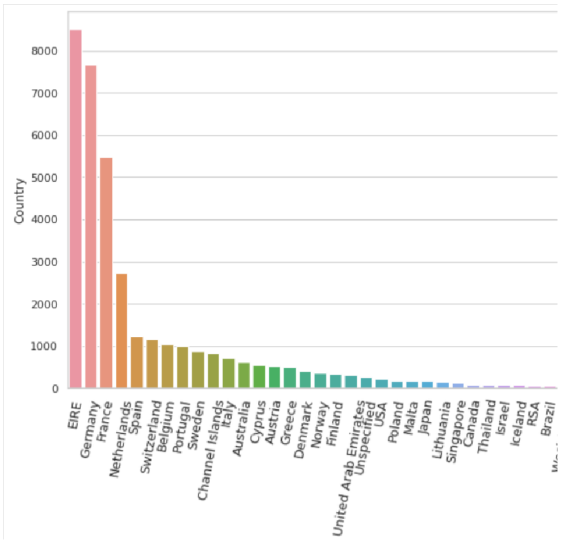


Figure: 4.1

Number of transactions per year

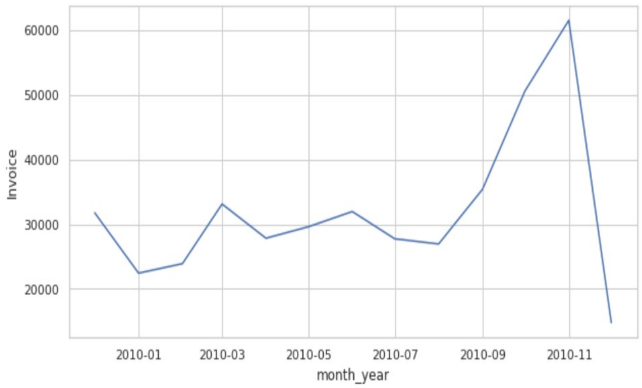


Figure: 4.2

4.4 Data Preprocessing:

Data pre-processing is the process by which raw data is prepared and suitable for machine learning models. This is the first important step in creating a machine learning model. Raw data can typically contain noise, missing values, and unusable formats that aren't directly available for machine learning models. Data pre processing is required to clean up the data and make it suitable for machine learning models. This also improves the accuracy and efficiency of machine learning models.

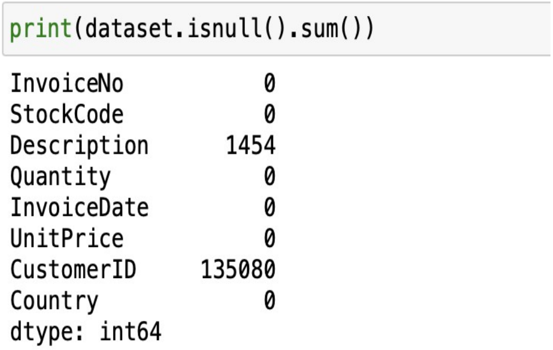
4.5 Methods Applied :

1. Handling Missing Values: Null Values from the dataset are removed, missing values from Description and Customer ID's rows are removed. Size of Dataset (541909, 8) before Removing null values is reduced to (406829, 8).

Result of null values reduction

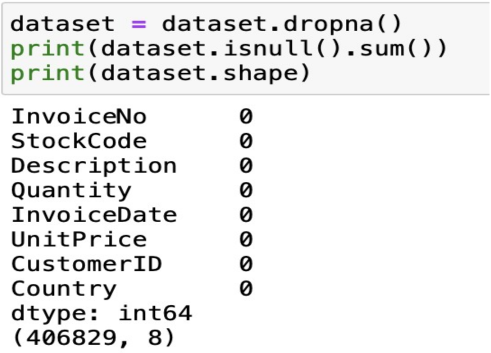
Before removing null values

After removing null values



```
print(dataset.isnull().sum())
```

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



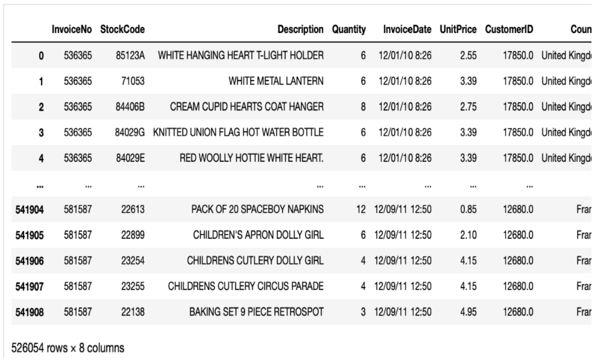
```
dataset = dataset.dropna()
print(dataset.isnull().sum())
print(dataset.shape)
```

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
(406829, 8)
```

Figure: 4.3

2. The Quantity attribute has values less than 0 so, these negative values in the dataset will be removed so that it will not affect while forming clusters along with which duplicate records in the data are also removed.

Data pre-processing result



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Coun
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/01/10 8:26	2.55	17850.0	United Kingd
1	536365	71053	WHITE METAL LANTERN	6	12/01/10 8:26	3.39	17850.0	United Kingd
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/01/10 8:26	2.75	17850.0	United Kingd
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/01/10 8:26	3.39	17850.0	United Kingd
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/01/10 8:26	3.39	17850.0	United Kingd
...	...	...	...	...	...	...	...	
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/09/11 12:50	0.85	12680.0	Frar
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/09/11 12:50	2.10	12680.0	Frar
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/09/11 12:50	4.15	12680.0	Frar
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/09/11 12:50	4.15	12680.0	Frar
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/09/11 12:50	4.95	12680.0	Frar

526054 rows × 8 columns

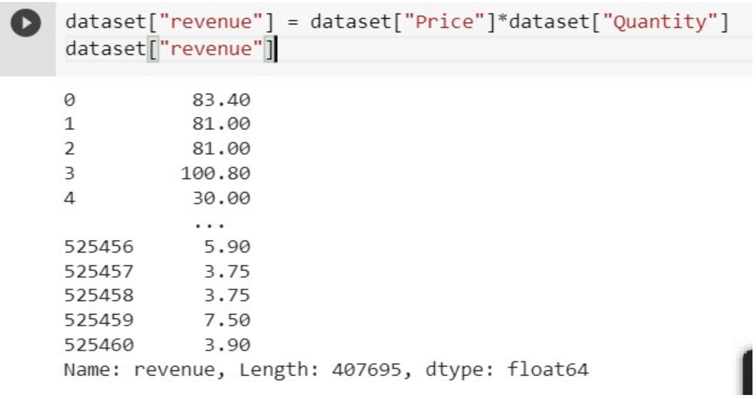
Figure: 4.4

3. ***Min-Max normalization***: Min-Max normalization transforms each feature individually so that it fits within the specified range of the training set. This technique is used on the dataset prepared from the RFM Analysis in the Section.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

4. ***Features extraction***: Revenue calculation, here we multiplied Price and Quantity in order find and add revenue generated column[Price * Quantity = Revenue]. And we have extracted data and month from Invoice Date.

Feature extraction



```
dataset["revenue"] = dataset["Price"]*dataset["Quantity"]
dataset["revenue"]
```

```
0          83.40
1          81.00
2          81.00
3         100.80
4          30.00
           ...
525456      5.90
525457      3.75
525458      3.75
525459      7.50
525460      3.90
Name: revenue, Length: 407695, dtype: float64
```

Figure: 4.5

4.6 RFM (Recency, Frequency, Monetary):

Analysis

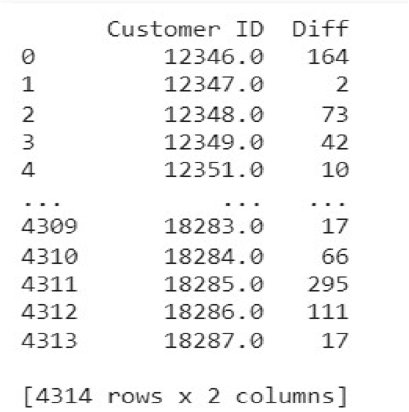
Recency, frequency, monetary value is a marketing analysis tool used to identify a company`s or an organization's best customers by measuring and analysing spending habits. RFM analysis reasonably predicts how much revenue a company will make from customers who are likely to buy the product again, new customers (compared to frequent customers), and how to convert casual buyers to frequent customers.

The RFM model is based on three quantitative factors:

- Recency- How recently a customer has made a purchase.
- Frequency- How often a customer makes a purchase.
- Monetary- How much money a customer spends on purchases.

Recency: Recency factor indicates the most recent purchase of all the customers In the dataset , we will consider the most recent purchase date as a reference date. Then we will find the recency of each customer by subtracting the transaction date from the reference date Finally we will get one integer value that will be the recency of the customer.

Calculated Recency Table

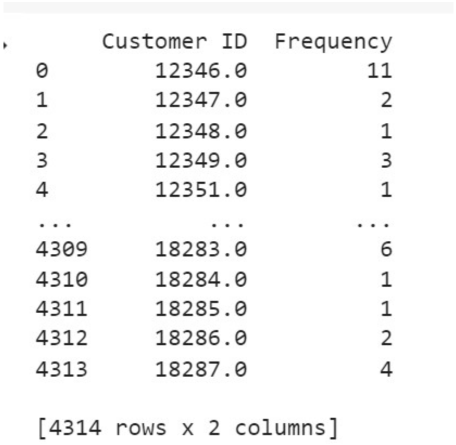


```
      Customer ID  Diff
0         12346.0   164
1         12347.0     2
2         12348.0    73
3         12349.0    42
4         12351.0    10
...           ...   ...
4309      18283.0    17
4310      18284.0    66
4311      18285.0   295
4312      18286.0   111
4313      18287.0    17

[4314 rows x 2 columns]
```

Figure: 4.6

Frequency: Customer transaction frequency can be influenced by factors such as product type, purchase price, and the need for replenishment or replacement. If you can predict the purchase cycle, marketing activities can be directed towards encouraging them to go to the store when they are low on staples. Here, we are grouping the Customer ID by the total count of the invoice register on the customer, in order to find the number of transactions done by the customer(frequency).

Calculated Frequency Table

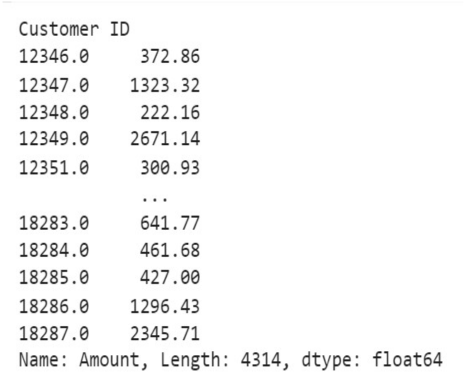


```
      Customer ID  Frequency
0         12346.0         11
1         12347.0          2
2         12348.0          1
3         12349.0          3
4         12351.0          1
...           ...        ...
4309      18283.0          6
4310      18284.0          1
4311      18285.0          1
4312      18286.0          2
4313      18287.0          4

[4314 rows x 2 columns]
```

Figure: 4.7

Monetary: Monetary value is how much a customer spends on purchase in a period. Overall amount spent by the customers in all transactions. This is to put more emphasis on encouraging customers who spend the most money to continue to do so. It produces a better return on investment in marketing and customer service, and also runs the risk of isolating customers who have been consistent but may not spend as much with each transaction. Monetary value is computed by considering customer id and summing up the amount he/she spends in each transaction. The output displays each customer id and Monetary value of that specific customer id.

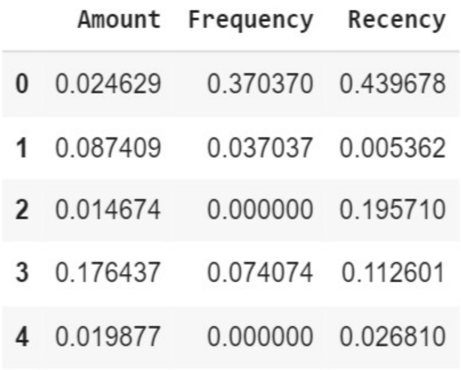
Calculated Monetary table



```
Customer ID
12346.0     372.86
12347.0    1323.32
12348.0     222.16
12349.0    2671.14
12351.0     300.93
             ...
18283.0     641.77
18284.0     461.68
18285.0     427.00
18286.0    1296.43
18287.0    2345.71
Name: Amount, Length: 4314, dtype: float64
```

Figure: 4.8

From the features extracted by RFM Analysis , the customized dataset is prepared with Amount ,Frequency and Recency as features on which K-means Clustering is applied. Data transformation is applied on the dataset to improve the performance of the algorithm , Min-max normalization technique is used to normalize all the features.



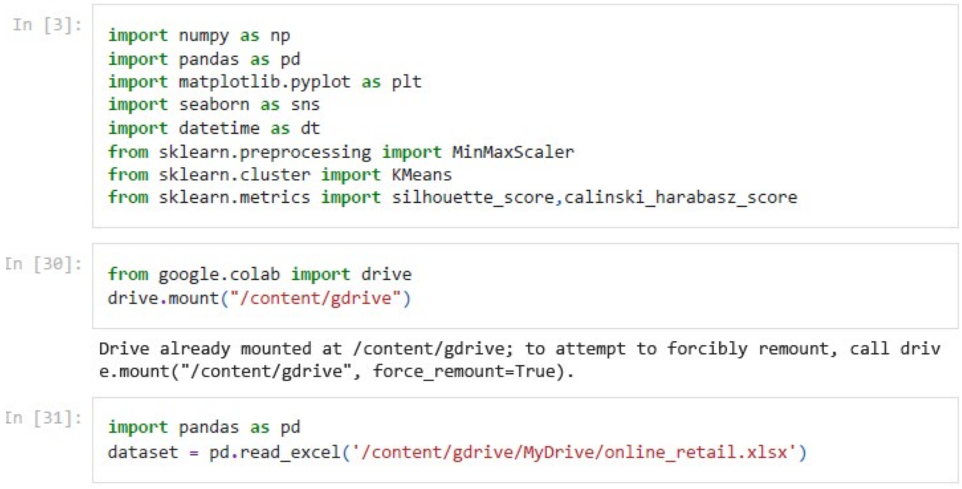
	Amount	Frequency	Recency
0	0.024629	0.370370	0.439678
1	0.087409	0.037037	0.005362
2	0.014674	0.000000	0.195710
3	0.176437	0.074074	0.112601
4	0.019877	0.000000	0.026810

Dataset derived from RFM Analysis

Figure: 4.9

CHAPTER 5

IMPLEMENTATION

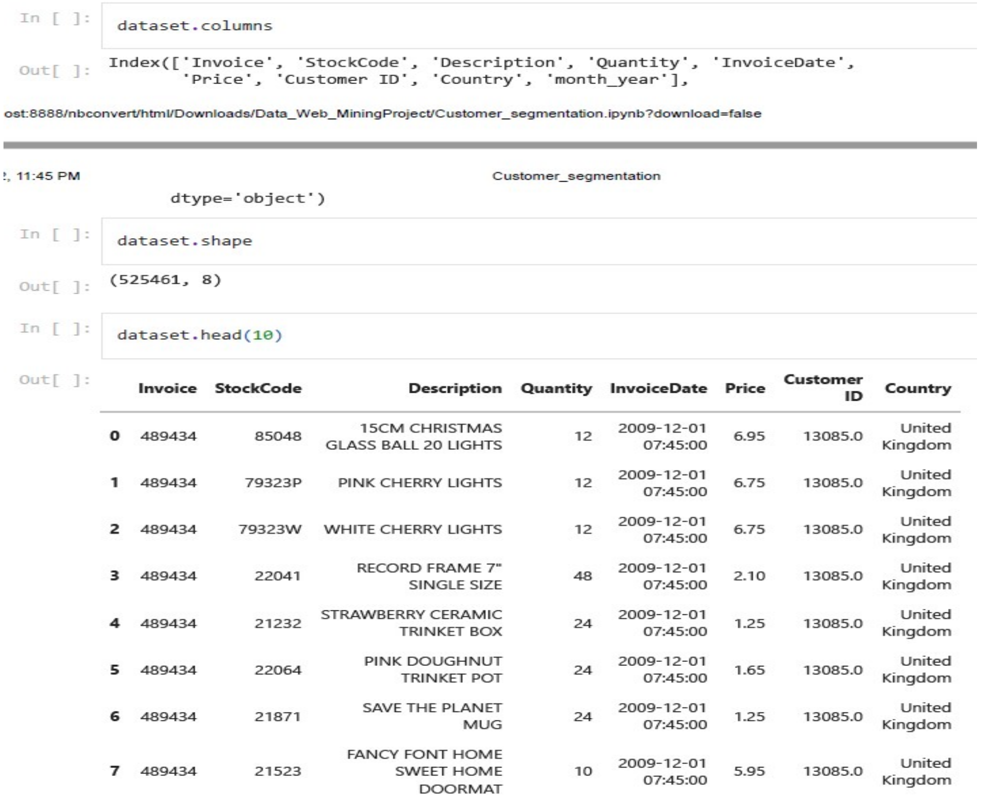


```
In [3]:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
from sklearn.preprocessing import MinMaxScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score,calinski_harabasz_score
```

```
In [30]:
from google.colab import drive
drive.mount("/content/gdrive")
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call driv
e.mount("/content/gdrive", force_remount=True).
```

```
In [31]:
import pandas as pd
dataset = pd.read_excel('/content/gdrive/MyDrive/online_retail.xlsx')
```



```
In [ ]:
dataset.columns
```

```
Out[ ]:
Index(['Invoice', 'StockCode', 'Description', 'Quantity', 'InvoiceDate',
       'Price', 'Customer ID', 'Country', 'month_year'],
```

ost:8888/nbconvert/html/Downloads/Data_Web_MiningProject/Customer_segmentation.ipynb?download=false

!, 11:45 PM Customer_segmentation

```
      dtype='object')
```

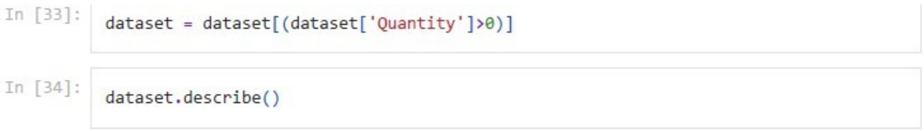
```
In [ ]:
dataset.shape
```

```
Out[ ]:
(525461, 8)
```

```
In [ ]:
dataset.head(10)
```

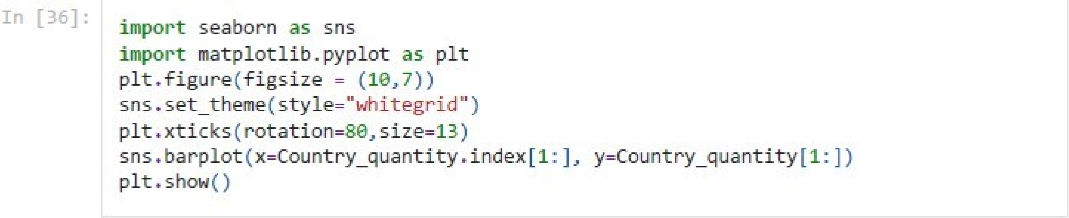
Out[ ]:

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom
5	489434	22064	PINK DOUGHNUT TRINKET POT	24	2009-12-01 07:45:00	1.65	13085.0	United Kingdom
6	489434	21871	SAVE THE PLANET MUG	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom
7	489434	21523	FANCY FONT HOME SWEET HOME DOORMAT	10	2009-12-01 07:45:00	5.95	13085.0	United Kingdom



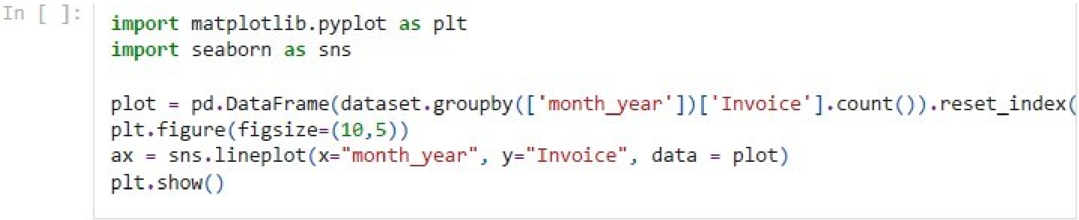
```
In [33]:
dataset = dataset[(dataset['Quantity']>0)]
```

```
In [34]:
dataset.describe()
```



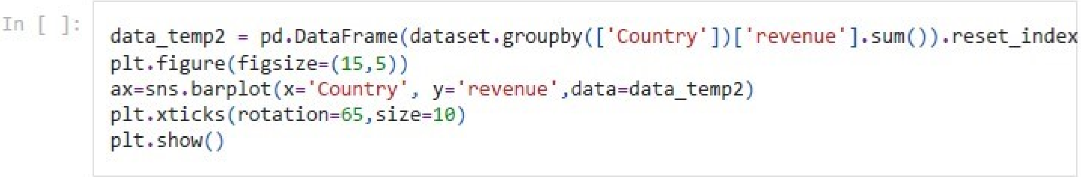
In [36]:

```
import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize = (10,7))
sns.set_theme(style="whitegrid")
plt.xticks(rotation=80,size=13)
sns.barplot(x=Country_quantity.index[1:], y=Country_quantity[1:])
plt.show()
```



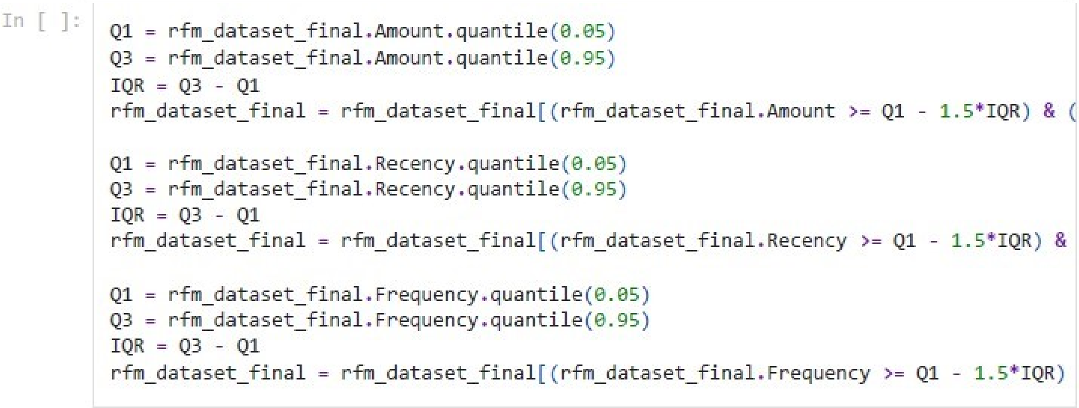
In [ ]:

```
import matplotlib.pyplot as plt
import seaborn as sns

plot = pd.DataFrame(dataset.groupby(['month_year'])['Invoice'].count()).reset_index(
plt.figure(figsize=(10,5))
ax = sns.lineplot(x="month_year", y="Invoice", data = plot)
plt.show()
```



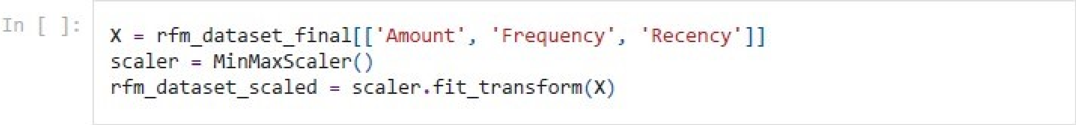
In [ ]:

```
data_temp2 = pd.DataFrame(dataset.groupby(['Country'])['revenue'].sum()).reset_index
plt.figure(figsize=(15,5))
ax=sns.barplot(x='Country', y='revenue',data=data_temp2)
plt.xticks(rotation=65,size=10)
plt.show()
```



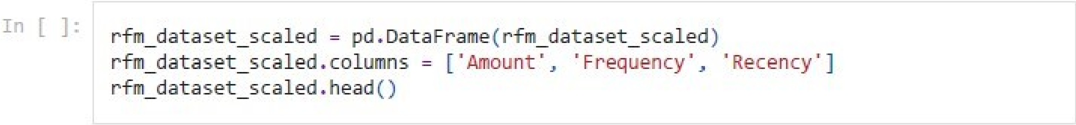
In [ ]:

```
Q1 = rfm_dataset_final.Amount.quantile(0.05)
Q3 = rfm_dataset_final.Amount.quantile(0.95)
IQR = Q3 - Q1
rfm_dataset_final = rfm_dataset_final[(rfm_dataset_final.Amount >= Q1 - 1.5*IQR) & (

Q1 = rfm_dataset_final.Recency.quantile(0.05)
Q3 = rfm_dataset_final.Recency.quantile(0.95)
IQR = Q3 - Q1
rfm_dataset_final = rfm_dataset_final[(rfm_dataset_final.Recency >= Q1 - 1.5*IQR) &

Q1 = rfm_dataset_final.Frequency.quantile(0.05)
Q3 = rfm_dataset_final.Frequency.quantile(0.95)
IQR = Q3 - Q1
rfm_dataset_final = rfm_dataset_final[(rfm_dataset_final.Frequency >= Q1 - 1.5*IQR)
```



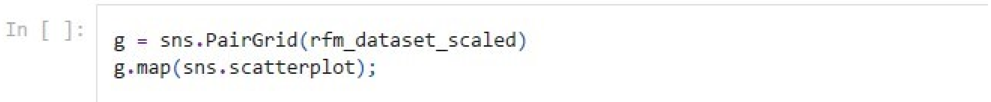
In [ ]:

```
X = rfm_dataset_final[['Amount', 'Frequency', 'Recency']]
scaler = MinMaxScaler()
rfm_dataset_scaled = scaler.fit_transform(X)
```

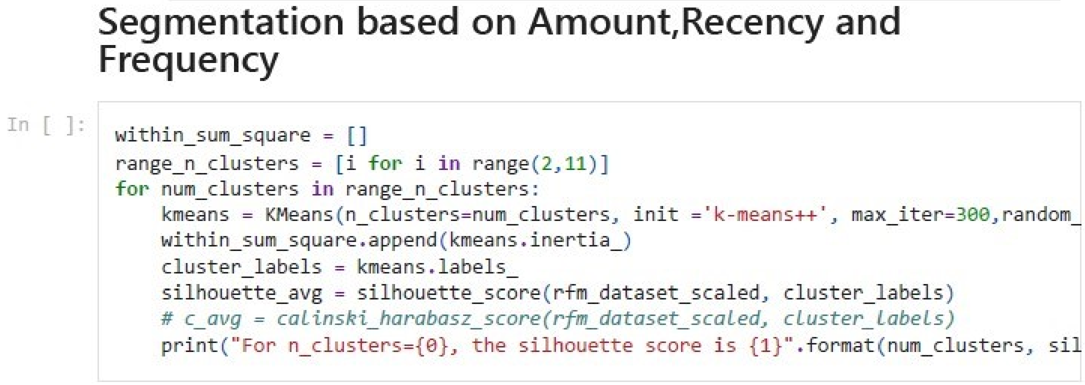


In [ ]:

```
rfm_dataset_scaled = pd.DataFrame(rfm_dataset_scaled)
rfm_dataset_scaled.columns = ['Amount', 'Frequency', 'Recency']
rfm_dataset_scaled.head()
```

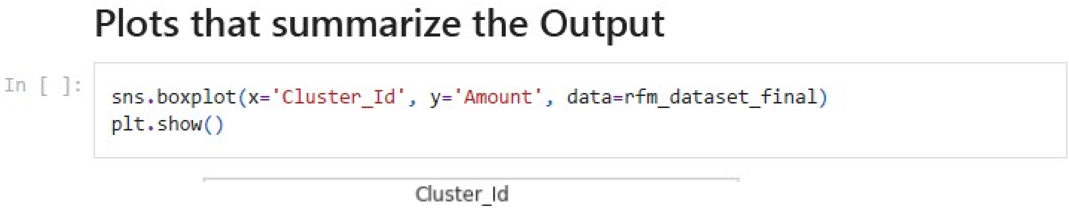


```
In [ ]:
g = sns.PairGrid(rfm_dataset_scaled)
g.map(sns.scatterplot);
```



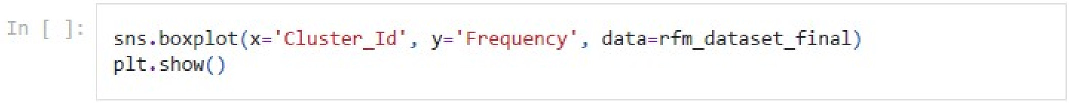
Segmentation based on Amount,Recency and Frequency

```
In [ ]:
within_sum_square = []
range_n_clusters = [i for i in range(2,11)]
for num_clusters in range_n_clusters:
    kmeans = KMeans(n_clusters=num_clusters, init ='k-means++', max_iter=300,random_
    within_sum_square.append(kmeans.inertia_)
    cluster_labels = kmeans.labels_
    silhouette_avg = silhouette_score(rfm_dataset_scaled, cluster_labels)
    # c_avg = calinski_harabasz_score(rfm_dataset_scaled, cluster_labels)
    print("For n_clusters={0}, the silhouette score is {1}".format(num_clusters, sil
```

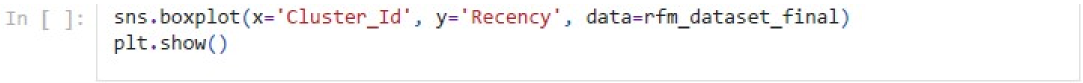


Plots that summarize the Output

```
In [ ]:
sns.boxplot(x='Cluster_Id', y='Amount', data=rfm_dataset_final)
plt.show()
```



```
In [ ]:
sns.boxplot(x='Cluster_Id', y='Frequency', data=rfm_dataset_final)
plt.show()
```



```
In [ ]:
sns.boxplot(x='Cluster_Id', y='Recency', data=rfm_dataset_final)
plt.show()
```

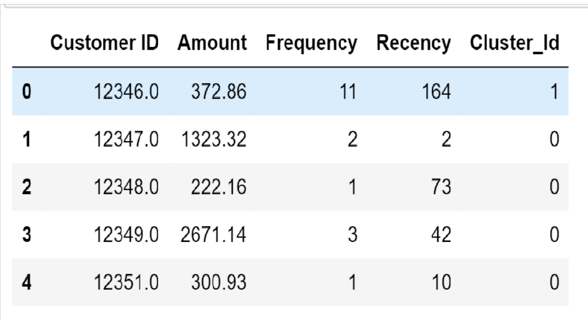
CHAPTER 6

RESULT

After using the Silhouette approach and interpretation of Data, K=3 is chosen. Hence, customers are divided into three clusters. For more insights the K-means is applied first on all features and then on 2 selected features. The graphs and results displayed below.

The cluster labels (0,1,2) predicted by the K-means clustering algorithm are assigned to corresponding customers.

Assigned cluster labels to respective customers



	Customer ID	Amount	Frequency	Recency	Cluster_Id
0	12346.0	372.86	11	164	1
1	12347.0	1323.32	2	2	0
2	12348.0	222.16	1	73	0
3	12349.0	2671.14	3	42	0
4	12351.0	300.93	1	10	0

Figure-6.1

Considered Amount, Frequency, and Recency as the features ,K-means clustering is applied using all these features. A 3D visualization of the resultant clusters obtained is shown above. Customers have been segmented into 3 groups. Red ,blue and green indicate clusters 0,1 and 2 respectively.

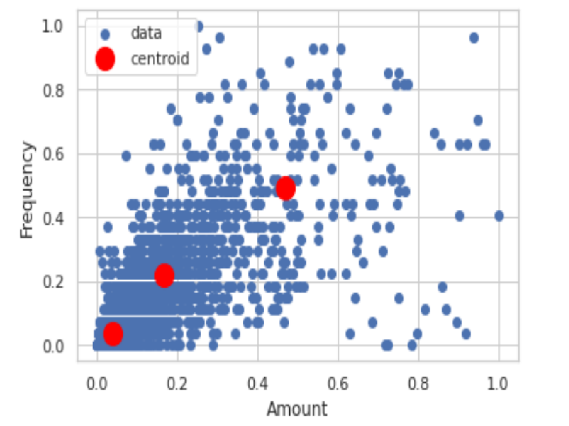


Figure: 6.2

In the shown graph only 2 features are considered(Amount and Frequency). We performed K-means clustering on these 2 features. Cluster centroids obtained from the k-means++ method are represented in red colour. Horizontal axis represents the amount and vertical axis represents Frequency.

Centroid and cluster represented graph

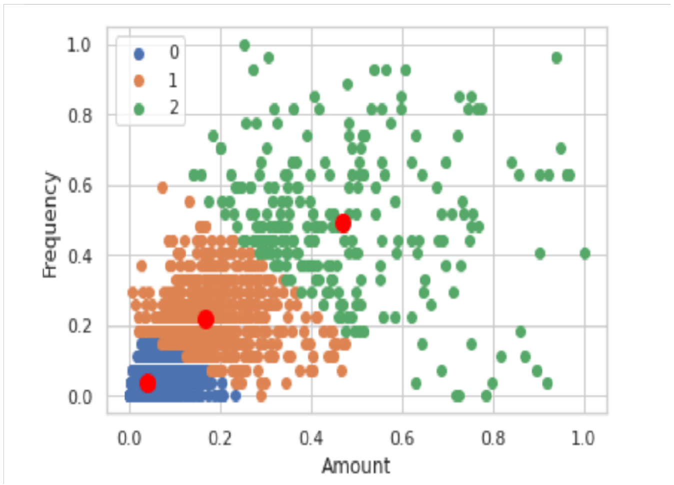
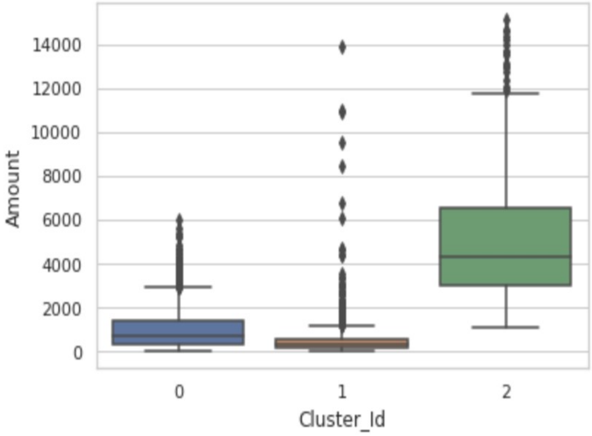


Figure: 6.3

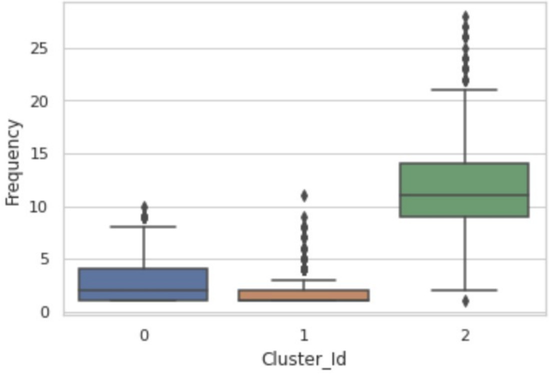
Now after applying K means clustering, Data is divided into 3 clusters and clearly seen on the graph. 0 represents blue data points, 1 represents orange data points, 2 represents green data points. To summarize the results of the proposed model, the below information along with the boxplots are described.

Here, we can see the distribution when Cluster Id attribute as x and Amount attribute as y. Cluster 1 has min 0 to max 3000,Cluster 2 has min 0 to max 1000 and Cluster 3 has min 1000 to max 12000



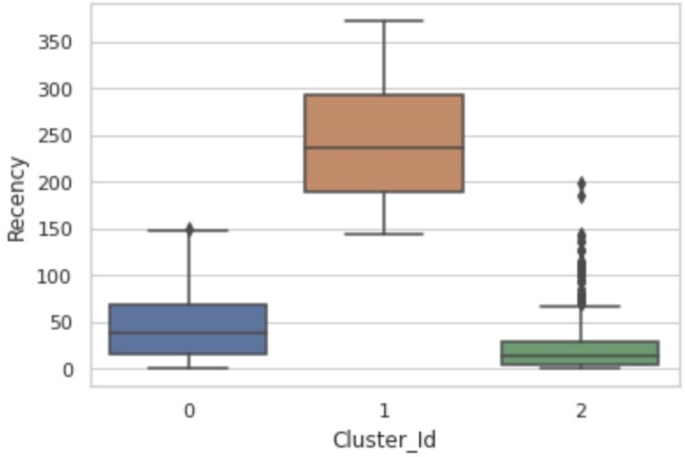
Clusters vs Amount (figure:6.4)

Here, we can see the distribution when Cluster Id attribute as x and Frequency attribute as y. Cluster 1 has min 0 to max 8,Cluster 2 has min 0 to max 2 and Cluster 3 has min 0 to max 21.



Clusters vs Frequency (figure: 6.5)

Here, we can see the distribution when Cluster Id attribute as x and Recency attribute as y. Cluster 1 has min 0 to max 150, Cluster 2 has min 150 to max 370 and Cluster 3 has min 0 to max 70.



Clusters vs Recency (figure: 6.6)

CHAPTER 7

CONCLUSION

This project introduced the RFM Model implementation of a K-Means clustering algorithm for customer segmentation using data collected from online retailers. Our model divides customers into mutually exclusive groups (in this case, three clusters).As shown in the above plots , cluster 1 represents the customers with highest recency , cluster 2 represents customers with lowest recency and accordingly cluster 1 is in between. This insight is obtained by segmenting the customers and hence leading to better understanding them, which can be used to increase the company's sales.This will help to apply further data mining strategies, and the more insights you gain the more it will help the business decisions.

CHAPTER 8

REFERENCES

[1] T. Kansal, S. Bahuguna, V. Singh and T. Choudhury, "Customer Segmentation using K-means Clustering," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS),2018,pp.135-139,doi: 10.1109/CTEMS.2018.8769171.

[2] Aman Banduni, A. Ilavendhan, "Customer Segmentation using Machine Learning," 2021 International Journal of Innovative Research in Technology(IJIRT), 2021,Volume 7, Issue 2, pp.116-122,doi: http://10.10.11.6/handle/1/1850.

[3] Abhinav Sagar, A. (n.d.). *Customer Segmentation Using K Means Clustering - KDnuggets*.KDnuggets;www.kdnuggets .com. Retrieved May 1, 2022, from https://www.kdnuggets.com/2019/11/cust omer-segmentation-using-k-means-clustering.html

[4] Dhiraj Kumar. (2021, June 18). *Implementing Customer Segmentation Using Machine Learning [Beginners Guide] neptune.ai*. Neptune.Ai; neptune.ai. https://neptune.ai/blog/customer-segmentati on-using-machine-learning

[5] Muhal, H. (n.d.). *(PDF) Two-Stage Customer Segmentation using K-Means Clustering And Artificial Neural Network | Harshit Muhal - Academia.edu*. (PDF) Two-Stage Customer Segmentation Using K-Means Clustering And Artificial Neural Network | Harshit Muhal - Academia.Edu; www.academia.edu. Retrieved May 1, 2022, from https://www.academia.edu/45443368/Two_Stage_Customer_Segmentation_using K_Means_Clustering_And_Artificial_Neural_Network

[6] Li, Y., Qi, J., Chu, X., & Mu, W. (2022, January 9). *Customer Segmentation Using K-Means Clustering and the Hybrid Particle Swarm Optimization Algorithm | The Computer Journal | Oxford Academic*. OUP Academic; academic.oup.com. https://academic.oup.com/comjnl/advance-article-abstract/doi/10.1093/comjnl/bxab206/6501352?redirectedFro