MINOR PROJECT REPORT

ON

CUSTOMER SEGMENTATION

BY

AYUSHI GUPTA (00916405320)

Submitted in partial fulfillment of the requirements
For the award of the degree of

Master of Technology
(INFORMATION TECHNOLOGY)

UNDER THE GUIDANCE OF:
DR. Udayan Ghose
SIGNATURE__________________

SUBMITTED TO:
(UNIVERSITY SCHOOL OF INFORMATION, COMMUNICATION AND TECHNOLOGY
GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY
(2020-2022))

TABLE OF CONTENTS

CERTIFICATE

This is to certify that the Term Paper titled "**CUSTOMER SEGMENTATION**" submitted to GGSIPU, Dwarka Delhi by **AYUSHI GUPTA**, in partial fulfillment of the requirement for the award of degree of masters of computer application during the academic year 2020-2022, is a bonafide work carried out by the student under my supervision and guidance. This work is the original one and has not been submitted anywhere else for any other degree.

____

Dr. Udayan Ghose

(Professor)

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude and respect to my supervisor **Prof. Udayan Ghose** for his excellent guidance, suggestions and support. I consider myself extremely lucky to be able to work under the guidance of such a dynamic personality. I would like to render heartiest thanks to my friend who's ever helping nature and suggestion has helped us to complete this present work.

I would like to thank all faculty members and staff of the IT Engineering, U.S.I.C.T, GGSIPU, Dwarka for their extreme help throughout the course. An assemblage of this nature could never have been attempted without reference to and inspiration from the works of others whose details are mentioned in the reference section. We acknowledge our indebtedness to all of them.

AYUSHI GUPTA

(00916405320)

ABSTRACT

In customer relationship management, consumer segmentation is crucial. It enables businesses to devise and implement various strategies aimed at maximising customer value. Customer segmentation is the process of categorising consumers based on shared factors such as age, geography, and purchasing patterns. Similarly, clustering entails grouping items together in such a manner that comparable categories of items are kept together. In this work, a machine learning (ML) hierarchical agglomerative clustering (HAC) method is built in Python to conduct consumer segmentation on credit card data sets and suggest effective marketing tactics. Customer segmentation splits consumers into groups based on shared traits, which helps banks, corporations, and companies enhance their goods and services.The study looks at how K-means, Hierarchical clustering, and Principal Component Analysis (PCA) may be used to determine a company's client categories based on their credit card transaction history. The information utilised in the research summarises the usage patterns of 8950 active credit card holders over the last six months, and our goal is to apply clustering algorithms to achieve the most accurate consumer segmentation

possible. For client segmentation, the project takes two approaches: first, it employs Hierarchical clustering and K-means to consider all factors in clustering methods. Second, by using Principal Component Analysis (PCA) to reduce the dataset's dimensionality, then determining the ideal number of clusters and performing the clustering analysis with the updated number of clusters. The results suggest that the PCA may be used as a check tool for K-means and hierarchical clustering in the clustering process.

INTRODUCTION

Companies that use customer segmentation believe that each client has unique needs that require a targeted marketing strategy to satisfy. Companies want to obtain a better understanding of the customers they're after. As a result, their goal must be particular and designed to meet the needs of each and every individual consumer. Furthermore, by analysing the data acquired, businesses may gain a better grasp of client preferences as well as the needs for identifying profitable categories. This allows them to more effectively design their marketing strategies while reducing the chance of their investment being harmed. Customer segmentation is based on a number of significant differentiators that split customers into groups that may be targeted. Demographics, geography, economic position, and behavioural tendencies all have a part in shaping the company's approach to solving various issues.

What is Customer Segmentation and its analysis?

Client segmentation is the practice of dividing a customer base into many groups of people who are similar in various aspects significant to marketing, such as gender, age, hobbies, and other spending patterns.

Understanding the components of a retailer's client base is critical to maximise their market potential; ceteris paribus, the merchant that attracts the most consumers will obtain the largest market share. Indeed, the high expenses of acquiring a new client or regaining an existing one encourage merchants to critically evaluate how to spend resources to optimise not only customer numbers, but also customer retention. Furthermore, it is widely accepted in the retail business that the Pareto Principle most likely applies to the company: The profits from 20% of the clients account for 80% of the total. The fact that retail enterprises depend on recurring purchases is one of the most important reasons why this theory stays true. As a result, a single customer's net change might have a large influence on a company's long-term earnings. As a result, it is often in the retailer's best advantage to focus on client retention.A priori analysis entails constructing the segments ahead of time and then placing each consumer into them after analysing the data. While the purpose of customer segmentation research has been consistent throughout retailers for many years, previous methodologies depended on far less sophisticated analytical tools than those accessible today. It's pointless to criticise corporations in the past for not effectively utilising their data; technology and data infrastructure just weren't available or affordable enough to allow firms to collect vast volumes of data in the way they do today. Despite this, many businesses continue to use primitive approaches to try to understand their clients, the most common of which is solely demographic analysis.

Challenges of Performing Analysis

Customer segmentation study has a lot of advantages. Retailers can better deploy resources to acquire and mine relevant data to enhance profitability by having a better grasp of their customer base. For many merchants, however, getting to the stage of undertaking high-level consumer segmentation analysis is more challenging than they anticipated. Many businesses may have access to the data needed to do the study, but they lack the capacity to access it in a user-friendly way or an employee with the essential capabilities.Smaller businesses' ability to do such analysis is hampered by a lack of appropriate employees or technology to manage the required volume of data. Although the availability of open source programming software like R or Python has made this sort of research more accessible, it still requires merchants to have someone on staff who can write in one of those languages. Furthermore, some shops are either unaware of the scope of their data collecting or lack the motivation to investigate it. However, businesses who haven't fully embraced consumer segmentation analysis are likely doing so because they can't afford to spend the time, money, or manpower required to do so. As a result, one of the goals of this work is to demonstrate that this type of in-depth analysis may be done inexpensively and effectively.However, there is a less obvious but equally important reason why businesses do not use consumer segmentation analysis: it is too difficult to comprehend. High-level consumer segmentation analysis necessitates significantly more accurate understanding of machine learning and the mathematics that define how the algorithms function when compared to standard demographic segmentation or RFM analysis. Furthermore, conventional marketing analysts lack the arithmetic and programming abilities required to properly deploy machine learning methods for consumer segmentation analysis; similarly, programmers and data analysts are not well-suited to undertake marketing responsibilities. This creates a new problem because it requires converting a typical marketing task—segmenting customers based on purchasing behaviors—into a purely programming task, which means the marketing team lacks the skills to code it themselves but the programming team lacks the marketing skills to interpret the results. As a result, a hybrid system is required.Traditional analysts have had some success with demographic or RFM research, but these models simply lack the technology capacity to give extensive insight into more precise client characteristics. Customer segmentation research paired with machine learning approaches, on the other hand, has the potential to change the way a store thinks about their data. As a result, merchants are looking for low-cost, simple solutions to apply clustering and describe how it may be used to segment their customers. After a thorough introduction to customer segmentation analysis, it's time to delve into the inner workings of several clustering methods before moving on to a discussion of the findings.

Clustering Using Machine Learning Methods

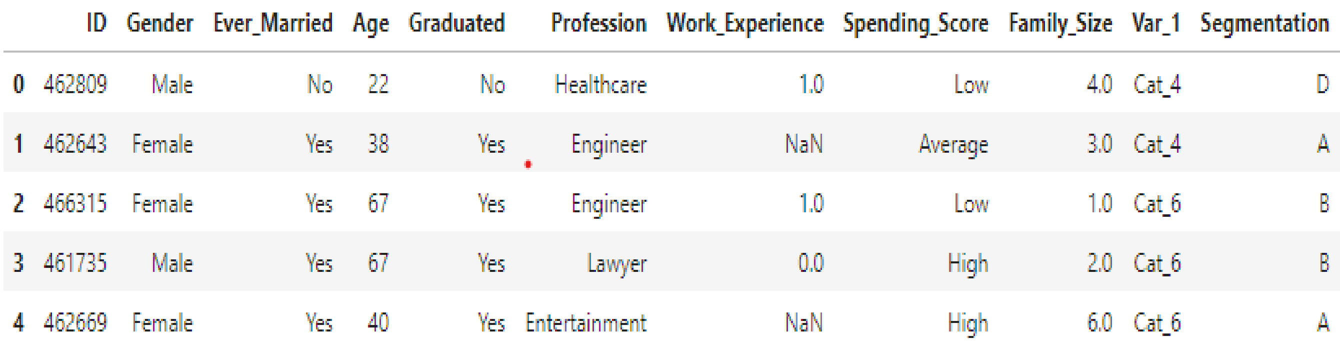
While many machine learning applications, such as regression and classification, aim to anticipate an instance's result or value, they don't try to analyse similarities across examples; instead, they focus on the link between instances and their respective outputs. As a result, the focus must shift from supervised machine learning to unsupervised machine learning when looking for algorithms or approaches that seek similarities between characteristics of instances. The presence of the goal value in the instances used to train the model in the training data determines whether the method is supervised or unsupervised machine learning.In all supervised machine learning training scenarios, instances are coupled with a goal value, which can be a scalar or a vector depending on the situation. Unsupervised machine learning, on the other hand, works with input that isn't matched with a goal value. It's probably better to look at these differences — and some similarities — through the lens of an example. Consider a retail business owner who has been open for more than a year and is interested in analysing their data to better understand their consumers while also projecting how much they will spend on their next visit. To forecast their next ticket, the owner examines their prior purchases and devises a method for estimating the worth of the future purchase based on the previous tickets.This is an example of supervised machine learning since it incorporates prediction and the results of past data and their outcomes (the tickets themselves). To be more exact, this form of technique is known as regression since the owner is most likely attempting to forecast a financial amount that the consumer will spend. To improve their understanding of their clients, the owner, on the other hand, chooses to examine the acquired customer data and see if there are any bigger patterns or similarities among the customers. This is a sort of unsupervised machine learning since there is no obvious outcome or target value connected with the data or the procedure. This is a perfect example of clustering. Clustering, in technical terms, is an unsupervised machine learning approach for grouping instances into clusters based on their commonalities. This simply indicates that clustering is a method of seeing or assessing data by examining it in groups.

CENTROID - BASED K-MEANS :

K-Means Although there are many other types of clustering algorithms, each of which deserves its own discussion and explanation, this paper will focus on two of them: centroid-based and hierarchical-based clustering algorithms. Before diving into the details of centroid-based clustering, it's important to first understand what a centroid is and how it fits into clustering. A centroid is the centre of a data cluster in the context of centroid-based clustering. Although the centre of a cluster may be defined in a variety of ways, in a k-means cluster, the centre is the arithmetic mean of each feature in the space where the data reside. To put it another way, the centroid is the average of the attributes of the instances assigned to that cluster. However, it may not be immediately apparent why centroids are required in the first place or how to define them. After all, there hasn't been anything.

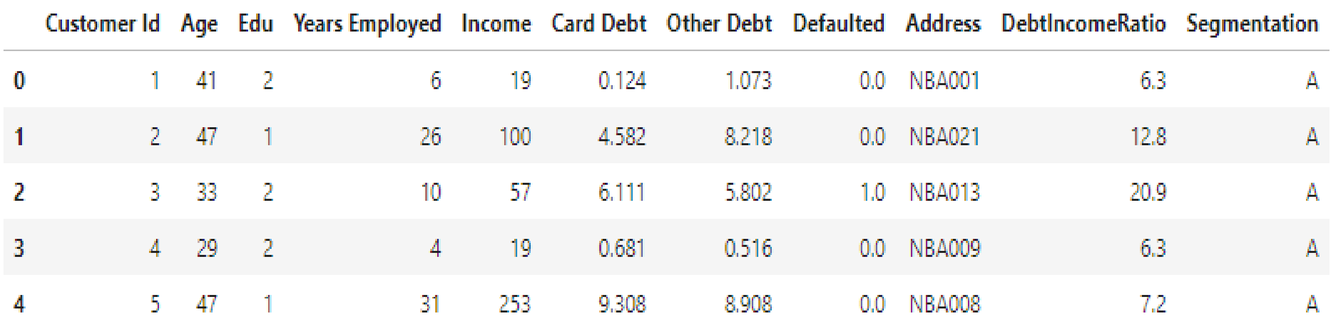
DATASET COLLECTION

➔DATASET 1



	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
0	462809	Male	No	22	No	Healthcare	1.0	Low	4.0	Cat_4	D
1	462643	Female	Yes	38	Yes	Engineer	NaN	Average	3.0	Cat_4	A
2	466315	Female	Yes	67	Yes	Engineer	1.0	Low	1.0	Cat_6	B
3	461735	Male	Yes	67	Yes	Lawyer	0.0	High	2.0	Cat_6	B
4	462669	Female	Yes	40	Yes	Entertainment	NaN	High	6.0	Cat_6	A

➔DATASET 2



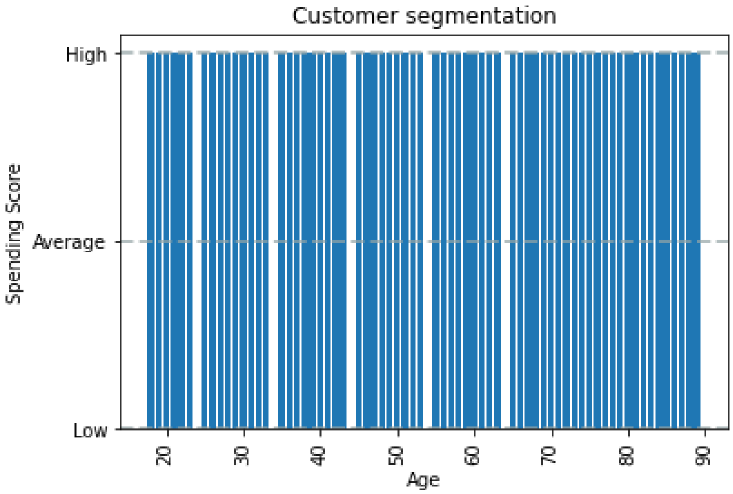
	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Defaulted	Address	DebtIncomeRatio	Segmentation
0	1	41	2	6	19	0.124	1.073	0.0	NBA001	6.3	A
1	2	47	1	26	100	4.582	8.218	0.0	NBA021	12.8	A
2	3	33	2	10	57	6.111	5.802	1.0	NBA013	20.9	A
3	4	29	2	4	19	0.681	0.516	0.0	NBA009	6.3	A
4	5	47	1	31	253	9.308	8.908	0.0	NBA008	7.2	A

BARPLOT BETWEEN THE FEATURES

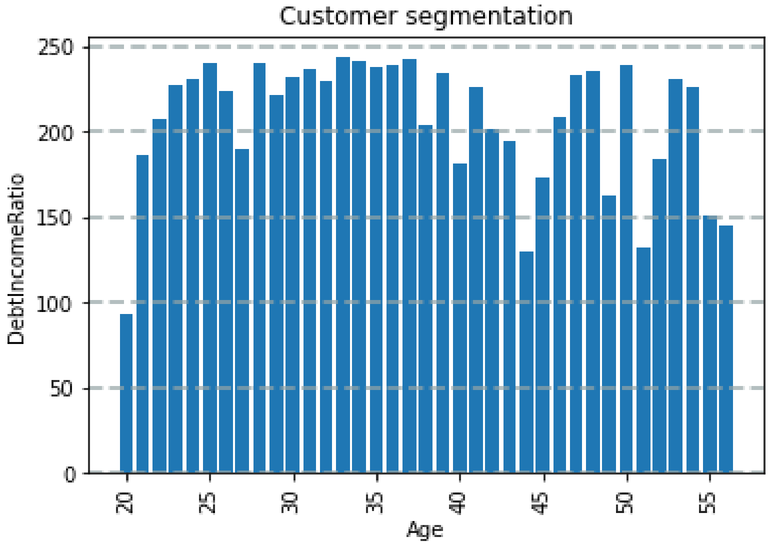
One of the most prevalent forms of graphics is a barplot (or barchart). It shows how a numeric and a categorical variable are related. A barplot is a tool for aggregating categorical data using many methods, the most common of which is the mean. It can also be seen as a communal vision through activity.

We pick a category column for the x-axis and a numerical column for the y-axis to apply this plot, and we notice that it makes a plot with a mean per categorical column.

➔ DATASET 1



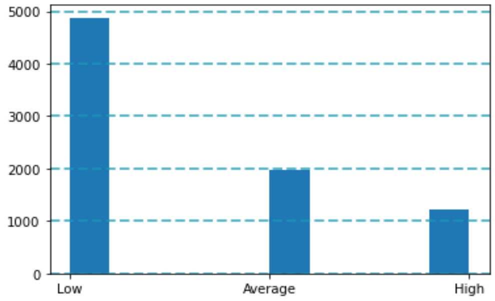
➔ DATASET 2

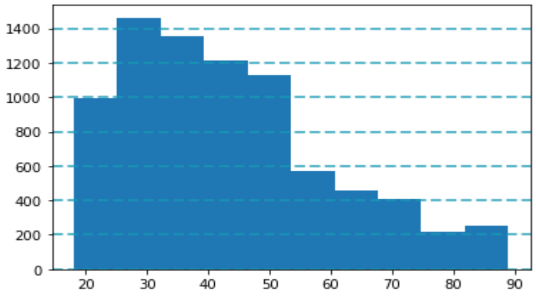


HISTOGRAM

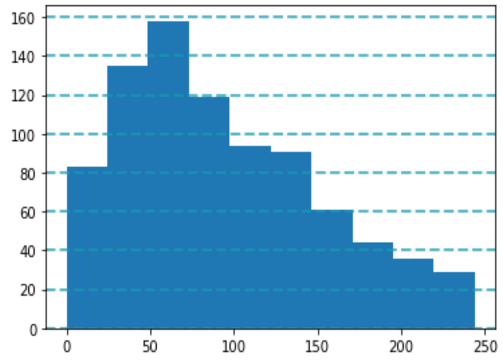
A histogram is a visual representation of data presented in the form of groupings. It's a precise way for displaying numerical data distribution graphically. It's a sort of bar plot in which the X-axis shows bin ranges and the Y-axis represents frequency.

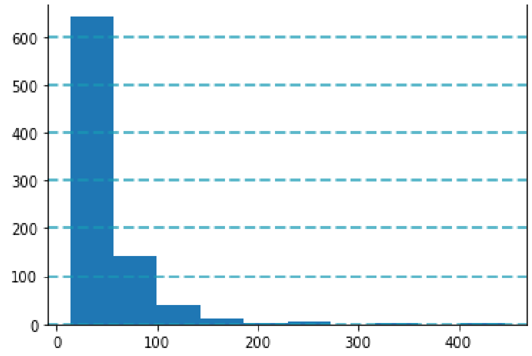
➔ DATASET 1 (AGE, SPENDING SCORE)





➔DATASET 2 (DebtIncomeRatio , Income)

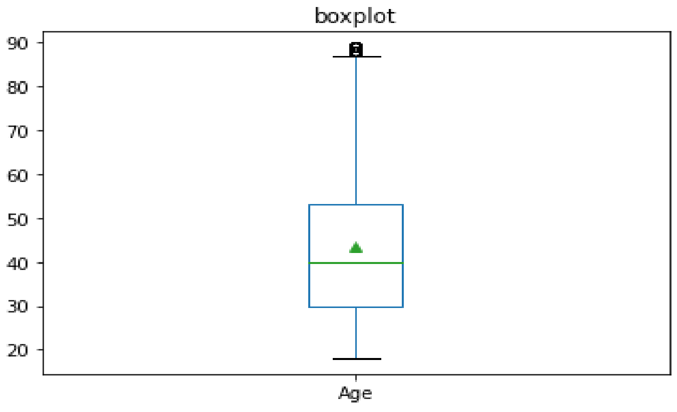




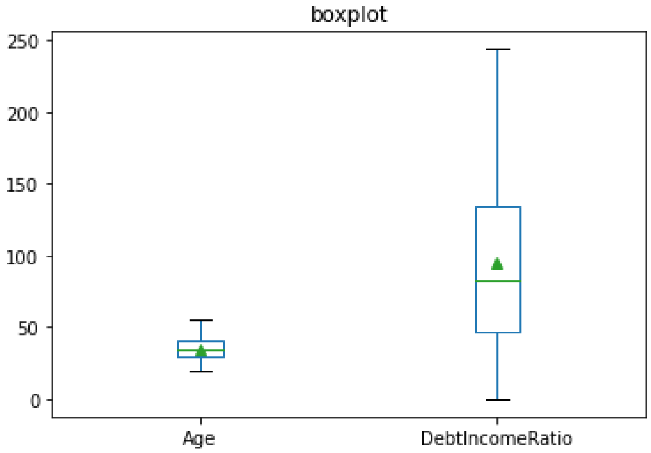
BOXPLOT

Boxplots are a way to see how well a data set's data is distributed. The data set is divided into three quartiles. This graph depicts the data set's lowest, maximum, median, first quartile, and third quartile.

➔DATASET 1



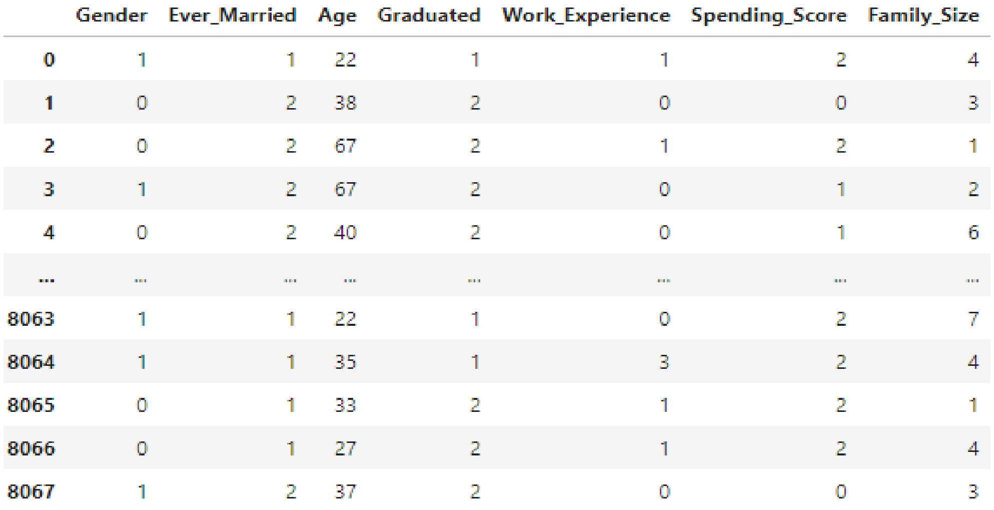
➔DATASET 2



LABEL ENCODING

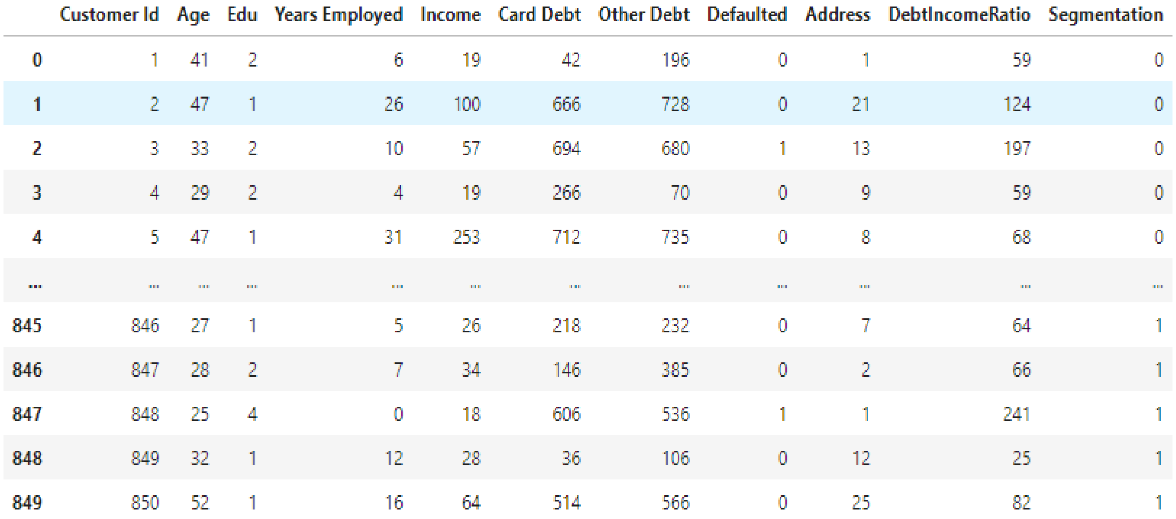
For categorical data, label encoding is a popular encoding approach. Each label is given a unique number based on alphabetical order in this approach. Let's look at how to use the scikit-learn module to create label encoding in Python, as well as the issues that come with it.

➔ DATASET 1



	Gender	Ever_Married	Age	Graduated	Work_Experience	Spending_Score	Family_Size
0	1	1	22	1	1	2	4
1	0	2	38	2	0	0	3
2	0	2	67	2	1	2	1
3	1	2	67	2	0	1	2
4	0	2	40	2	0	1	6
...	...	...	...	...	...	...	...
8063	1	1	22	1	0	2	7
8064	1	1	35	1	3	2	4
8065	0	1	33	2	1	2	1
8066	0	1	27	2	1	2	4
8067	1	2	37	2	0	0	3

➔ DATASET 2

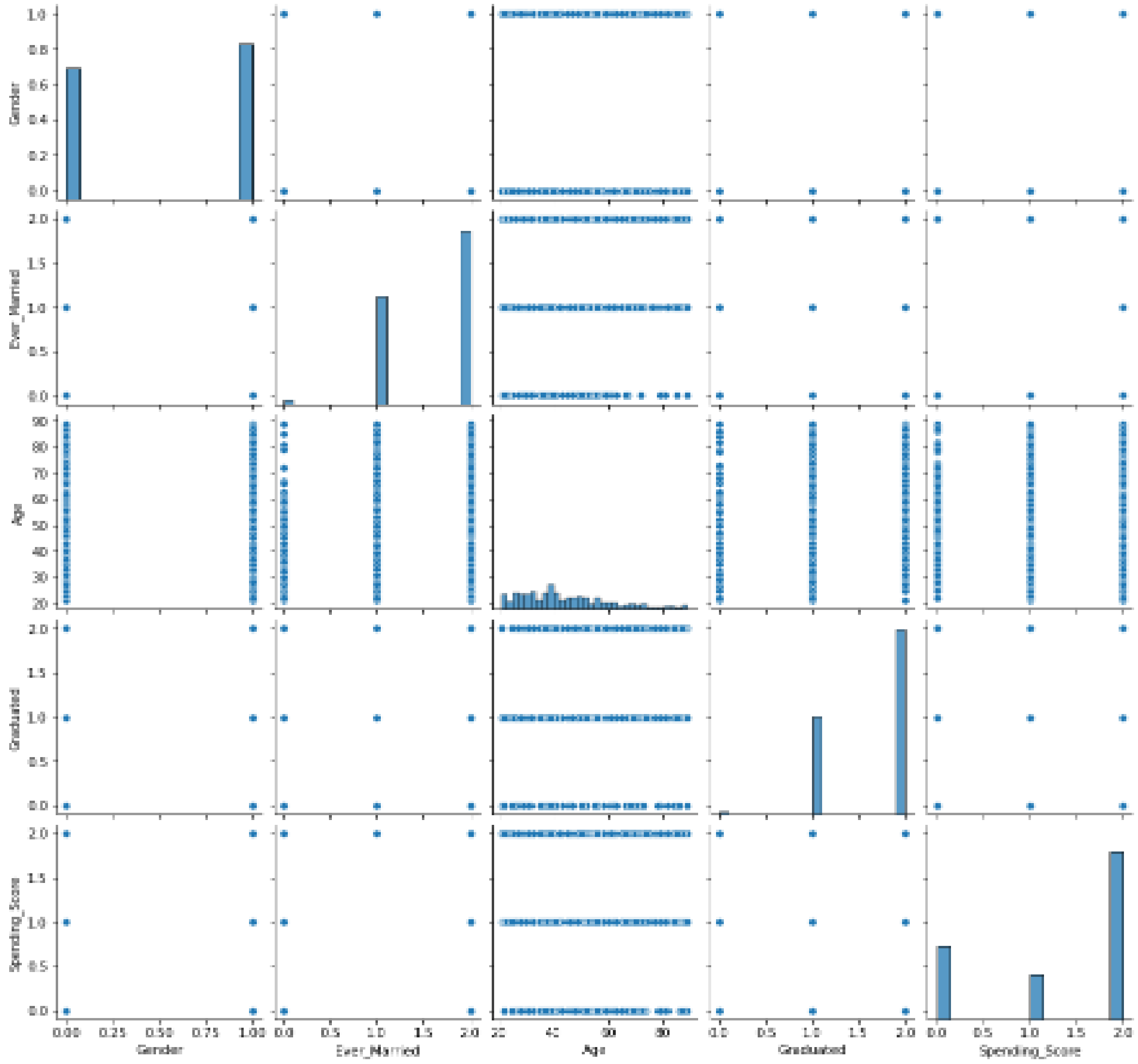


	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Defaulted	Address	DebtIncomeRatio	Segmentation
0	1	41	2	6	19	42	196	0	1	59	0
1	2	47	1	26	100	666	728	0	21	124	0
2	3	33	2	10	57	694	680	1	13	197	0
3	4	29	2	4	19	266	70	0	9	59	0
4	5	47	1	31	253	712	735	0	8	68	0
...	...	...	...	...	...	...	...	...	...	...	...
845	846	27	1	5	26	218	232	0	7	64	1
846	847	28	2	7	34	146	385	0	2	66	1
847	848	25	4	0	18	606	536	1	1	241	1
848	849	32	1	12	28	36	106	0	12	25	1
849	850	52	1	16	64	514	566	0	25	82	1

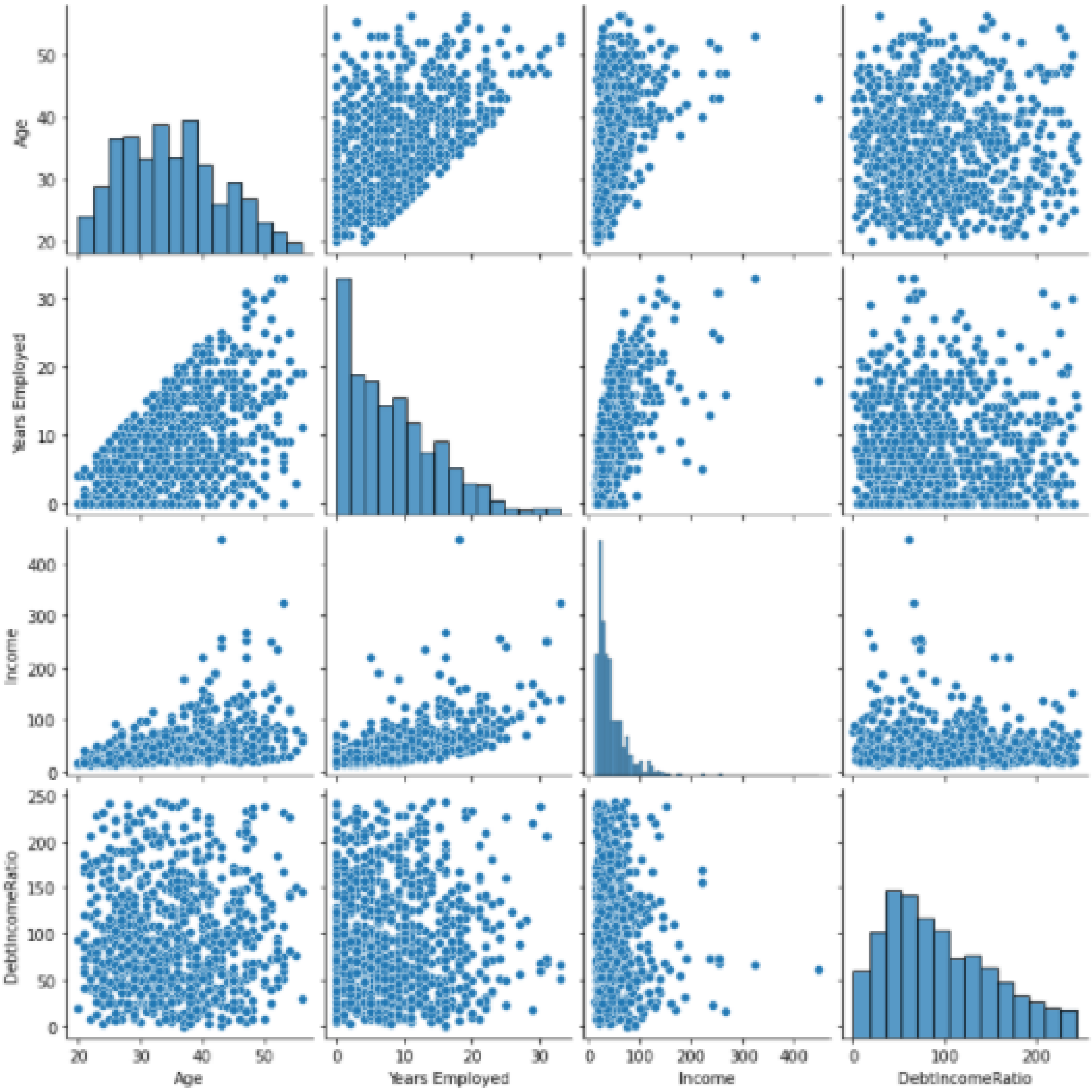
PAIRPLOT

In a dataset, a pairplot depicts a paired connection. The pairplot function generates a grid of Axes in which each variable in the data is spread over a single row and a single column on the y-axis.

➔ DATASET 1



➔DATASET 2


Age
Years Employed
Income
DebtIncomeRatio
20
30
40
50
0
10
20
30
0
100
200
300
400
0
50
100
150
200
250
Age
Years Employed
Income
DebtIncomeRatio

K-means Algorithm For Cluster

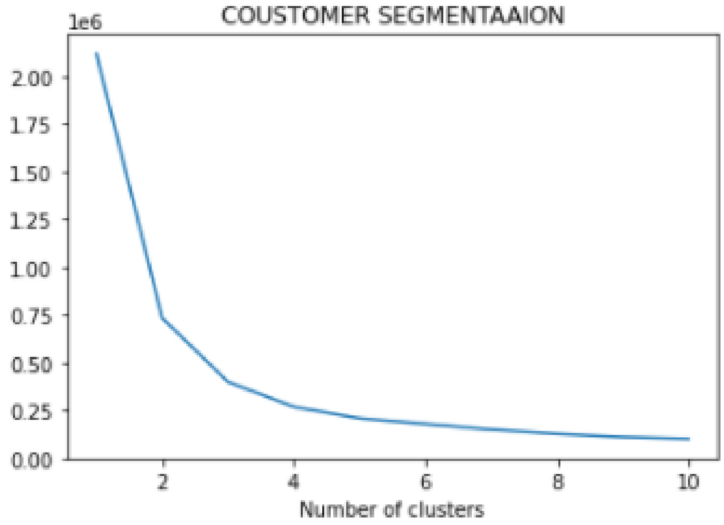
The first step in applying the k-means clustering method is to specify the number of clusters (k) that we want in the final result. The procedure begins by randomly picking k items from the dataset to act as the cluster's initial centres. The cluster means, also known as centroids, are the chosen objects. The closest centroid is then assigned to the remaining items. The object's centroid is determined by the Euclidean Distance between it and the cluster mean. This is referred to as "cluster assignment." After the assignment is finished, the algorithm calculates a new mean value for each cluster in the data. The observations are reviewed to see if they are closer to a different cluster once the centres have been recalculated. The objects are reassigned using the updated cluster mean. This is performed multiple times until the cluster allocations are finalised.

VISUALIZING THE OPTIMAL NUMBER OF CLUSTERS

➔ELBOW GRAPH

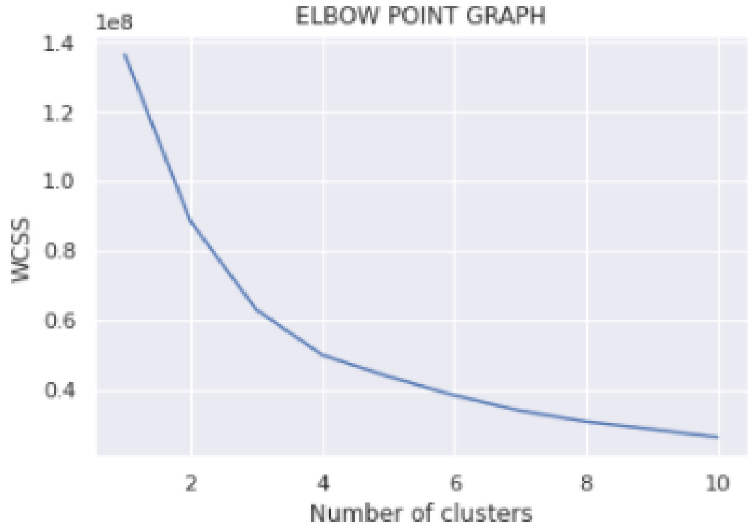
K-means is a basic unsupervised machine learning technique that divides data into a set of clusters (k).... The elbow approach performs k-means clustering on the dataset for a range of k values (say, 1-10), and then computes an average score for all clusters for each value of k.

- **DATASET 1**



OPTIMUM NUMBER OF CLUSTER = 3

- **DATASET 2**



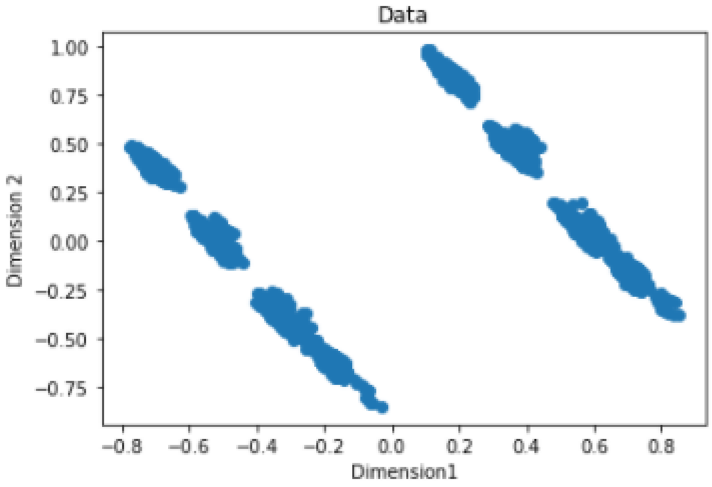
OPTIMUM NUMBER OF CLUSTERS = 4

Visualizing the Clustering Results using the First Two Principal Components

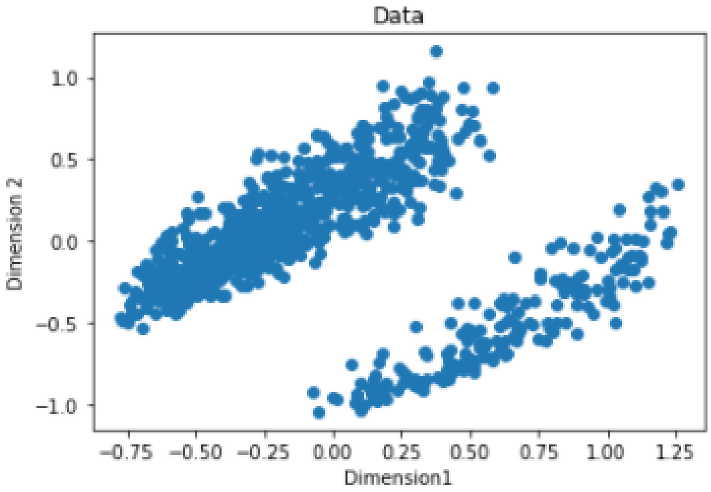
➔ PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis is an unsupervised learning approach used in machine learning to reduce dimensionality. With the aid of orthogonal transformation, it is a statistical technique that turns observations of correlated characteristics into a collection of linearly uncorrelated data. The Principal Components are the newly altered characteristics. It's one of the most widely used programmes for exploratory data analysis and predictive modelling. It's a method for extracting strong patterns from a dataset by lowering variances. PCA seeks for the lowest-dimensional surface on which to project the high-dimensional data.The variance of each characteristic is taken into account by PCA since the high attribute indicates a good separation between the classes and so minimises dimensionality. Image processing, movie recommendation systems, and optimising power allocation in multiple communication channels are some of the real-world uses of PCA. Because it is a feature extraction approach, it keeps the significant variables while discarding the less important ones.

➔DATASET 1



➔ DATASET 2

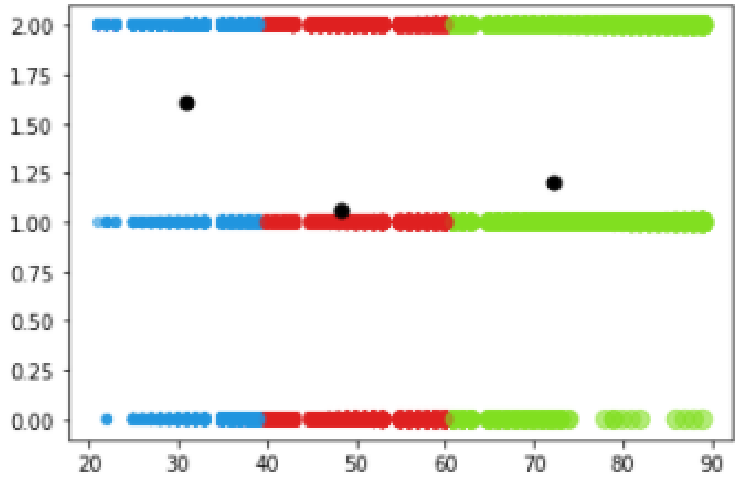


VISUALISING THE CLUSTERS BY K-MEANS

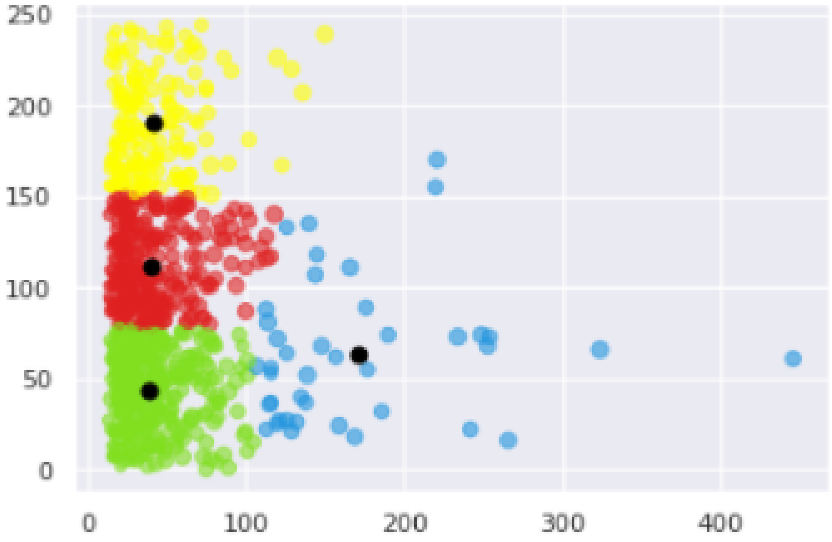
➔ SCATTER PLOT

Scatter plots are used to visualise correlations between variables, with dots representing the link. A scatter plot is created using the matplotlib library's scatter() function. Scatter plots are commonly used to depict relationships between variables and how one affects the other.

➔ DATASET 1



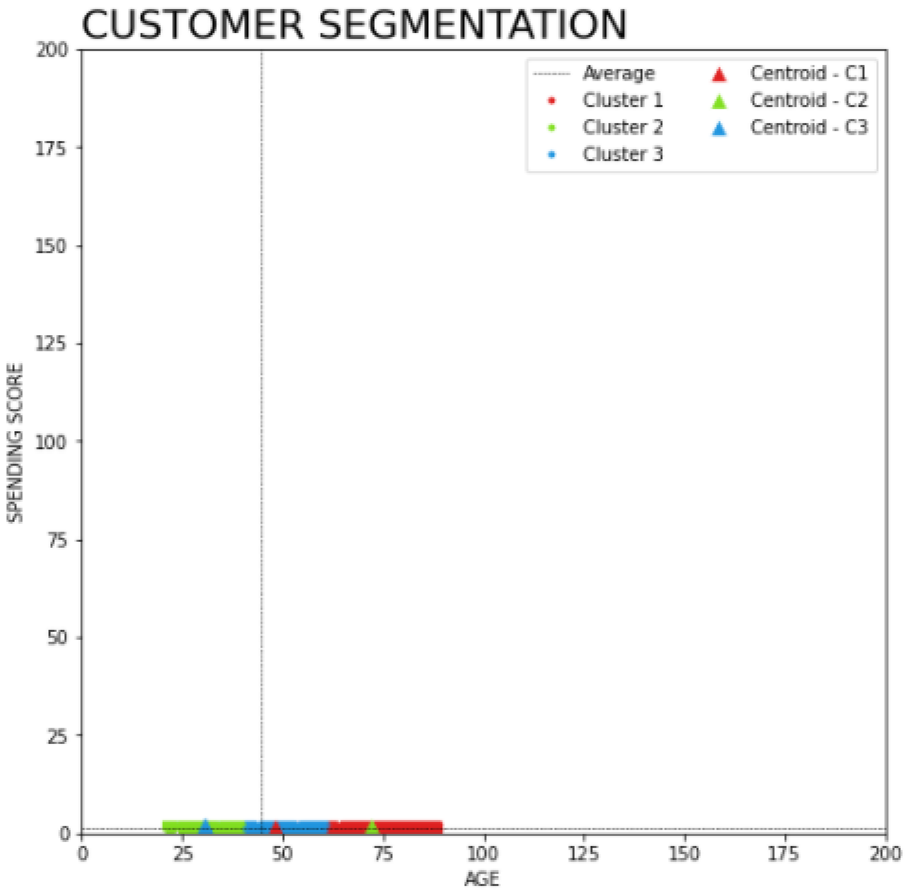
➔ DATASET 2



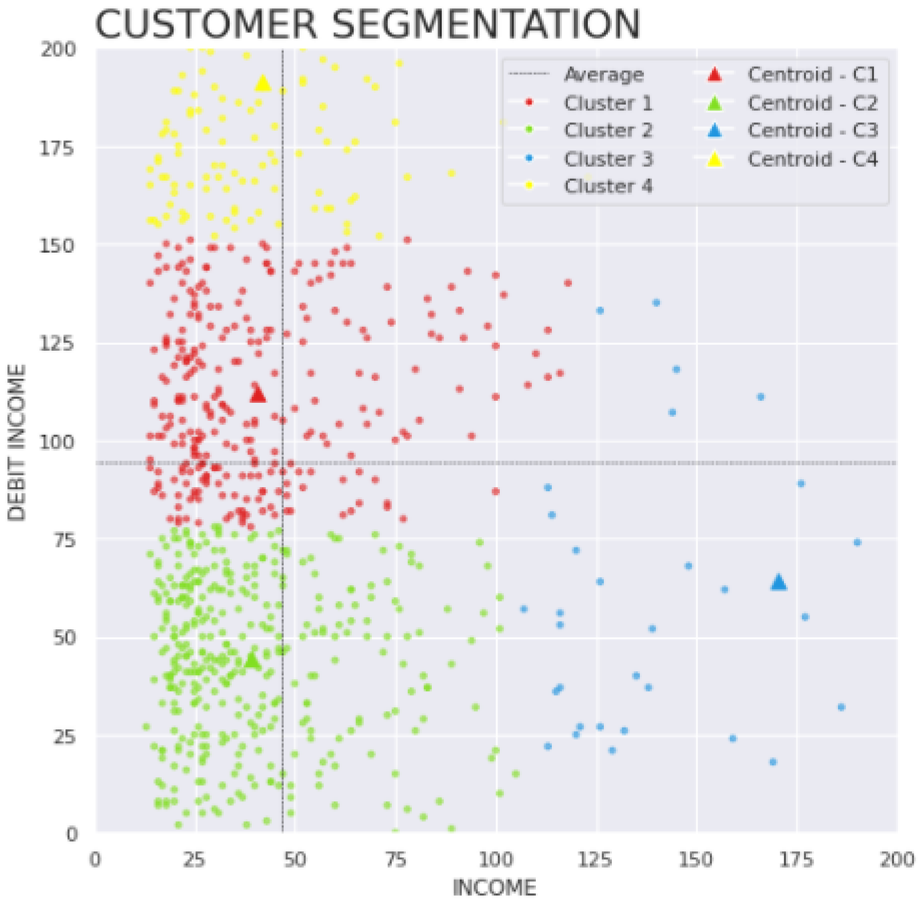
POKEMON STATS

To put it another way, the goal of clustering is to separate groups with similar qualities and assign them to clusters. It may assist you in defining actions for a set of items rather than dealing with each one individually, and elements might include clients, tasks, students, projects, goods, and Pokemons.

➔DATASET 1

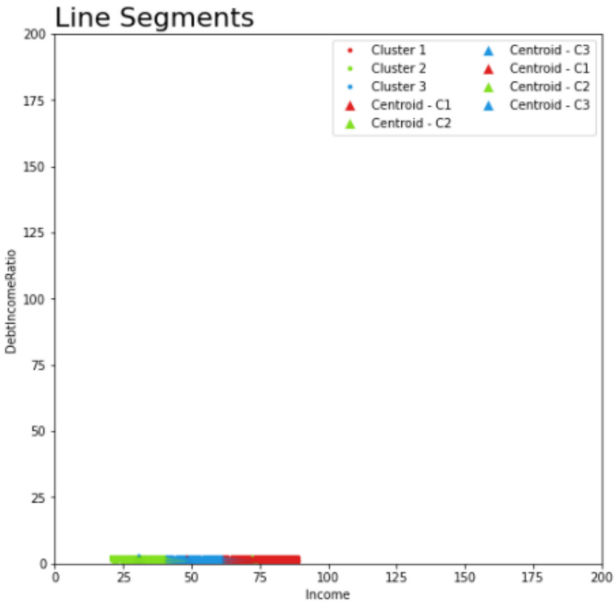


➔ DATASET 2

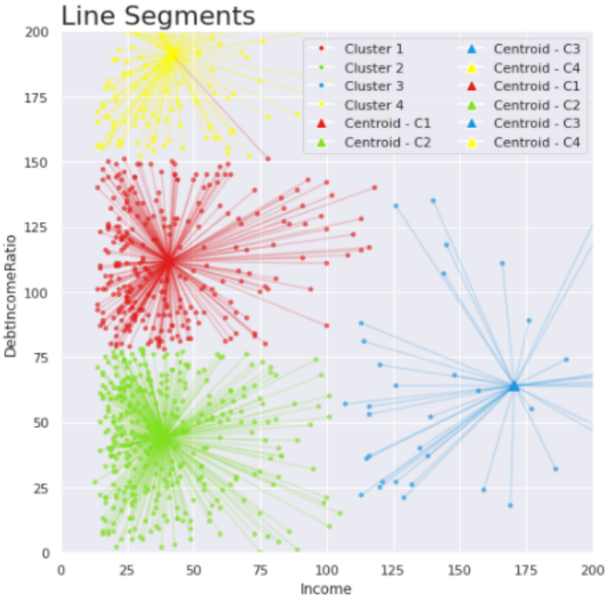

CUSTOMER SEGMENTATION
DEBIT INCOME
INCOME
0
25
50
75
100
125
150
175
200
Average
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Centroid - C1
Centroid - C2
Centroid - C3
Centroid - C4

LINE SEGMENT

➔DATASET 1



➔DATASET 2



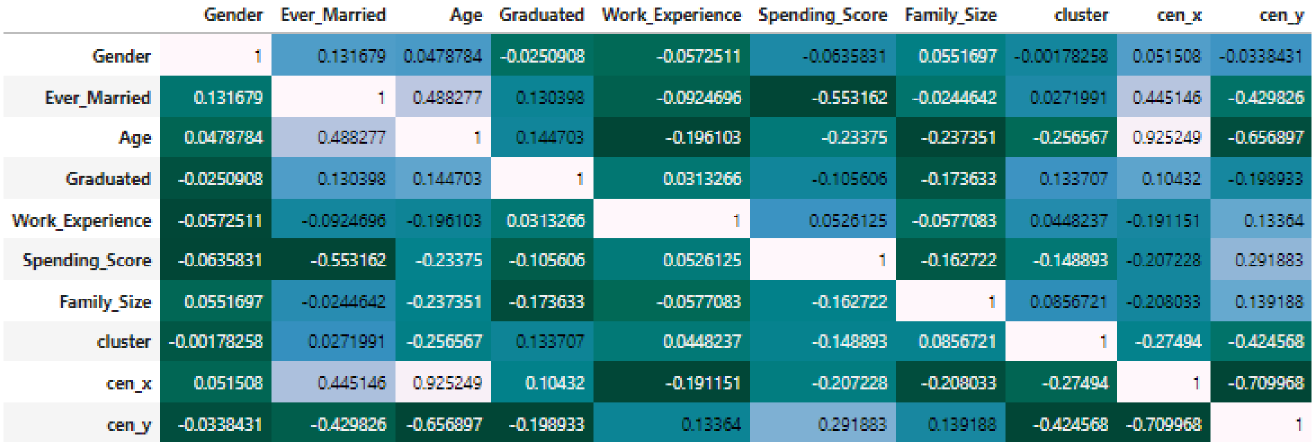
CORRELATION MATRIX

Correlation means an association, It is a measure of the extent to which two variables are related.

1. Positive Correlation: When two variables grow and drop at the same time. They have a nice relationship. The number '1' denotes a complete positive correlation. For example, demand and profit are positively connected, meaning that the higher the demand for a product, the higher the profit, and vice versa.

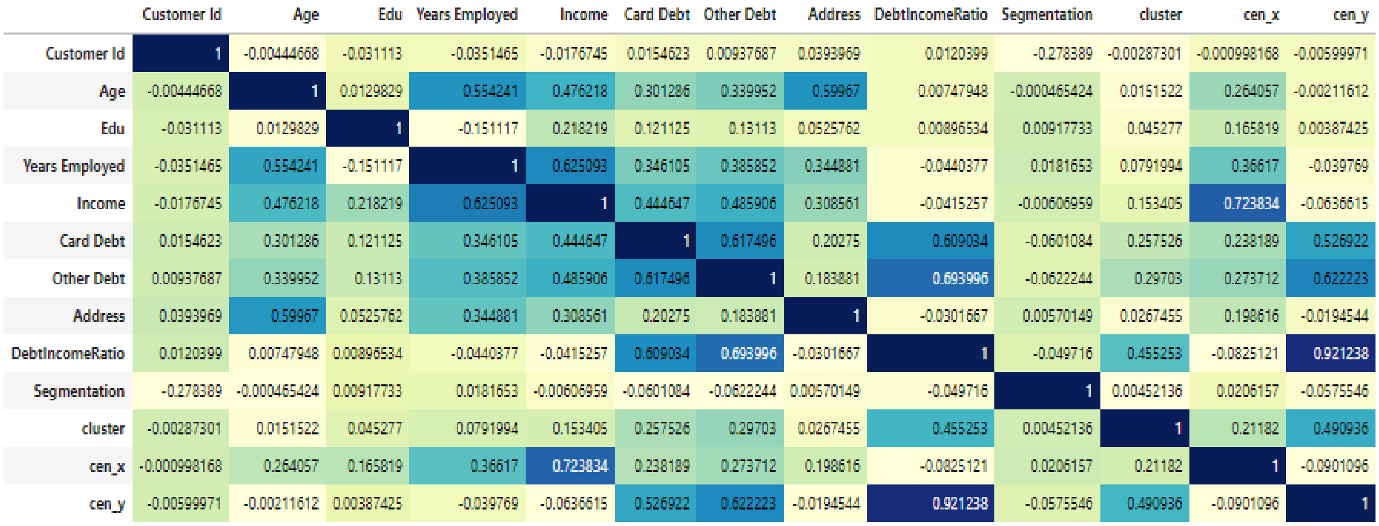
2. Negative Correlation: When one variable rises and the other falls at the same time, and vice versa. They have an unfavourable relationship. When the distance between magnets is increased, for example, their attraction diminishes, and vice versa. As a result, there is a negative association. '-1' indicates that there is no relationship.

➔DATASET1



	Gender	Ever_Married	Age	Graduated	Work_Experience	Spending_Score	Family_Size	cluster	cen_x	cen_y
Gender	1	0.131679	0.0478784	-0.0250908	-0.0572511	-0.0635831	0.0551697	-0.00178258	0.051508	-0.0338431
Ever_Married	0.131679	1	0.488277	0.130398	-0.0924696	-0.553162	-0.0244642	0.0271991	0.445146	-0.429826
Age	0.0478784	0.488277	1	0.144703	-0.196103	-0.23375	-0.237351	-0.256567	0.925249	-0.656897
Graduated	-0.0250908	0.130398	0.144703	1	0.0313266	-0.105606	-0.173633	0.133707	0.10432	-0.198933
Work_Experience	-0.0572511	-0.0924696	-0.196103	0.0313266	1	0.0526125	-0.0577083	0.0448237	-0.191151	0.13364
Spending_Score	-0.0635831	-0.553162	-0.23375	-0.105606	0.0526125	1	-0.162722	-0.148893	-0.207228	0.291883
Family_Size	0.0551697	-0.0244642	-0.237351	-0.173633	-0.0577083	-0.162722	1	0.0856721	-0.208033	0.139188
cluster	-0.00178258	0.0271991	-0.256567	0.133707	0.0448237	-0.148893	0.0856721	1	-0.27494	-0.424568
cen_x	0.051508	0.445146	0.925249	0.10432	-0.191151	-0.207228	-0.208033	-0.27494	1	-0.709968
cen_y	-0.0338431	-0.429826	-0.656897	-0.198933	0.13364	0.291883	0.139188	-0.424568	-0.709968	1

➔DATASET 2

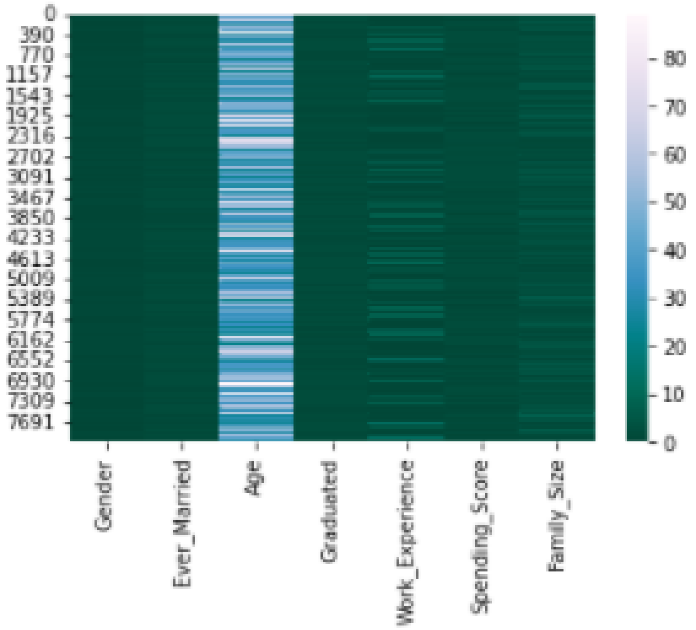


	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Segmentation	cluster	cen_x	cen_y
Customer Id	1	-0.00444668	-0.031113	-0.0351465	-0.0176745	0.0154623	0.00937687	0.0393969	0.0120399	-0.278389	-0.00287301	-0.000998168	-0.00599971
Age	-0.00444668	1	0.0129829	0.554241	0.476218	0.301286	0.339952	0.59967	0.00747948	-0.000465424	0.0151522	0.264057	-0.00211612
Edu	-0.031113	0.0129829	1	-0.151117	0.218219	0.121125	0.13113	0.0525762	0.00896534	0.00917733	0.045277	0.165819	0.00387425
Years Employed	-0.0351465	0.554241	-0.151117	1	0.625093	0.346105	0.385852	0.344881	-0.0440377	0.0181653	0.0791994	0.36617	-0.039769
Income	-0.0176745	0.476218	0.218219	0.625093	1	0.444647	0.485906	0.308561	-0.0415257	-0.00606959	0.153405	0.723834	-0.0636615
Card Debt	0.0154623	0.301286	0.121125	0.346105	0.444647	1	0.617496	0.20275	0.609034	-0.0601084	0.257526	0.238189	0.526922
Other Debt	0.00937687	0.339952	0.13113	0.385852	0.485906	0.617496	1	0.183881	0.693996	-0.0622244	0.29703	0.273712	0.622223
Address	0.0393969	0.59967	0.0525762	0.344881	0.308561	0.20275	0.183881	1	-0.0301667	0.00570149	0.0267455	0.198616	-0.0194544
DebtIncomeRatio	0.0120399	0.00747948	0.00896534	-0.0440377	-0.0415257	0.609034	0.693996	-0.0301667	1	-0.049716	0.455253	-0.0825121	0.921238
Segmentation	-0.278389	-0.000465424	0.00917733	0.0181653	-0.00606959	-0.0601084	-0.0622244	0.00570149	-0.049716	1	0.00452136	0.0206157	-0.0575546
cluster	-0.00287301	0.0151522	0.045277	0.0791994	0.153405	0.257526	0.29703	0.0267455	0.455253	0.00452136	1	0.21182	0.490936
cen_x	-0.000998168	0.264057	0.165819	0.36617	0.723834	0.238189	0.273712	0.198616	-0.0825121	0.0206157	0.21182	1	-0.0901096
cen_y	-0.00599971	-0.00211612	0.00387425	-0.039769	-0.0636615	0.526922	0.622223	-0.0194544	0.921238	-0.0575546	0.490936	-0.0901096	1

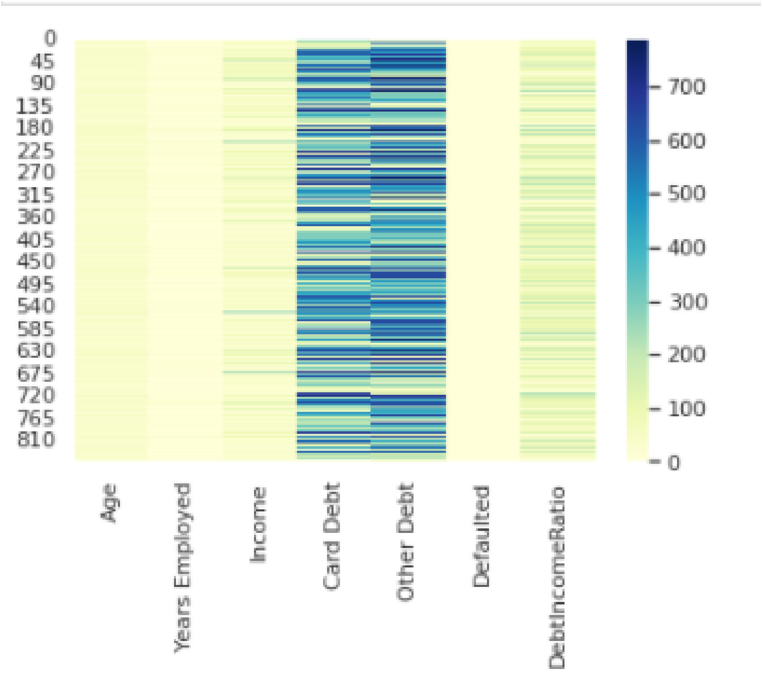
HEATMAP

A heatmap is a graphical representation of data that uses colours to display the matrix's value. To symbolise more common values or greater activities, brighter colours, primarily reddish hues, are utilised, whereas darker colours are favoured to represent less common or activity values. The name of the shading matrix also defines the heatmap. The seaborn.heatmap() method may be used to create heatmaps in Seaborn.

- ➔ **DATASET 1**



➔ DATASET 2

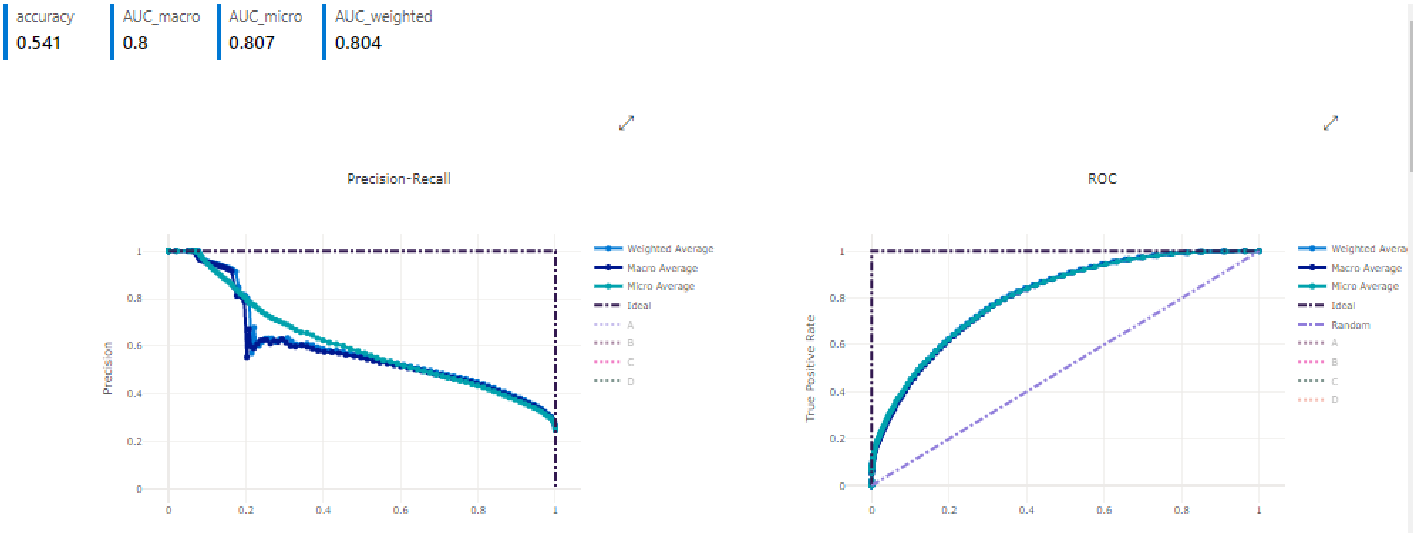


0
45
90
135
180
225
270
315
360
405
450
495
540
585
630
675
720
765
810
Age
Years Employed
Income
Card Debt
Other Debt
Defaulted
DebtIncomeRatio
700
600
500
400
300
200
100
0

COMPARATIVE STUDY

PARAMETERS	DATASET 1	DATASET 2
accuracy	0.5411514071509841	0.985882352941176 5
precision_score_weighted	0.540228615767184	0.986824670532924 9
f1_score_macro	0.5290983530013584	0.982375814880148 2
recall_score_weighted	0.5411514071509841	0.985882352941176 5

- **DATASET 1**



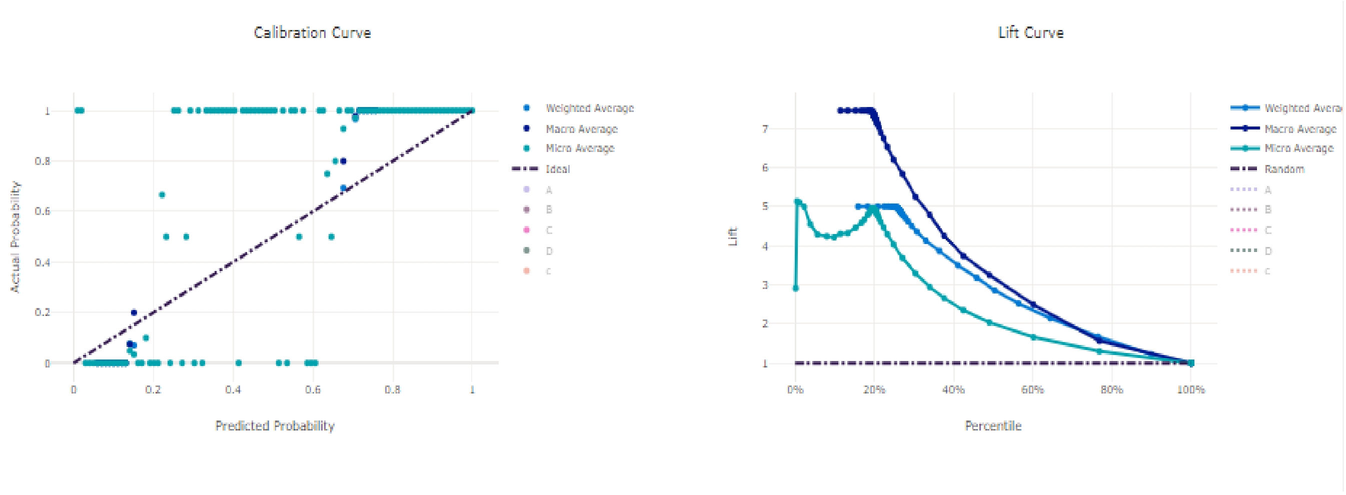
accuracy 0.541

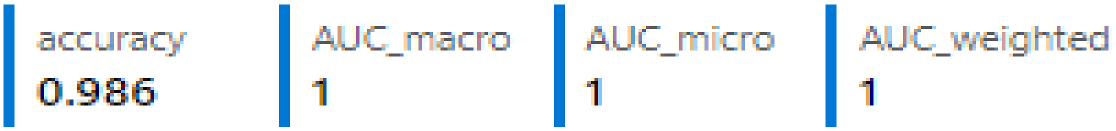
AUC_macro 0.8

AUC_micro 0.807

AUC_weighted 0.804

- **DATASET 2**





accuracy
0.986

AUC_macro
1

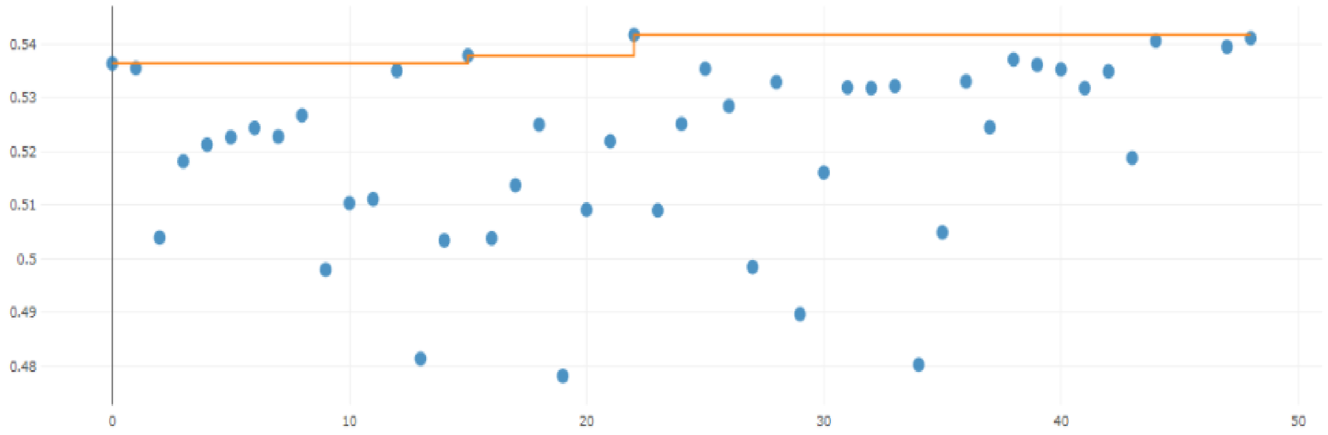
AUC_micro
1

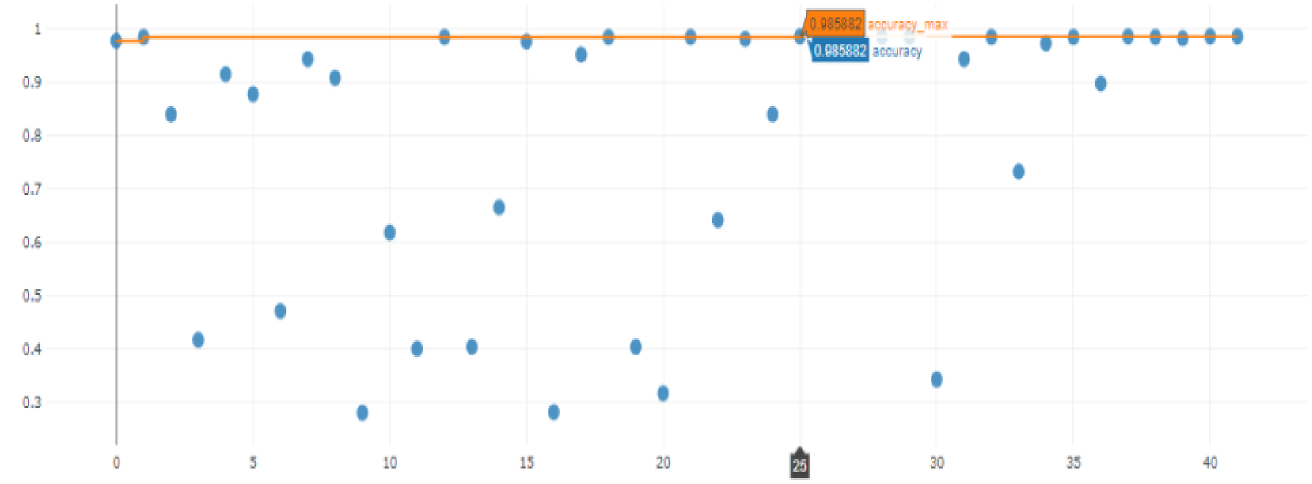
AUC_weighted
1

RESULTS

There are important ramifications that come with these findings, particularly in terms of cluster organisation. To begin, the ideal number of clusters to pick is somewhere between five and six, according to both algorithms. It's simple to separate the clusters to find patterns when there are just five or six of them. Although separating fewer clusters may be easier, the clusters will be significantly less useful and too generic to make accurate predictions. Traditional consumer segmentation study frequently has this issue. Traditional customer segmentation study typically leads to oversimplification of the clusters and thus hampers management action since there is a motive to keep analysis basic and not add attributes unless absolutely essential. Regardless, throughout the customer segmentation analysis study, the decision to cluster with five or six parts is evident. In some ways, this means that clustering cannabis retail data, even using cannabis-specific factors, is a waste of time.

```
0.5411514071509841
```





CONCLUSION AND FUTURE WORK

The implications of these discoveries are significant, particularly in terms of cluster organisation. To begin, both algorithms agree that the best number of clusters to choose is between five and six. When there are just five or six clusters, separating them to identify patterns is straightforward. Although it may be easier to separate fewer clusters, the clusters will be far less valuable and too general to make accurate forecasts. This is a common problem in traditional customer segmentation studies. Traditional customer segmentation studies frequently result in oversimplification of clusters, which impedes management action since there is a desire to keep analysis simple and not add features unless absolutely necessary. Regardless, the decision to cluster with five or six pieces is visible throughout the consumer segmentation research study. This implies that clustering cannabis retail data, even using cannabis-specific characteristics, is a waste of effort in several aspects.