Customer Segmentation

Project report submitted in partial fulfillment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

By

Perla Venkata Naga Sai Ganesh (181402)

Under the supervision of Dr. Ravindara Bhatt

to

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat,

Solan-173234, Himachal Pradesh

Candidate's Declaration

I hereby declare that the work presented in this report entitled **"Customer Segmentation"** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** submitted in the Department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from August 2021 to December 2021 under the supervision of **Dr. Ravindara Bhatt** (Assistant Professor), Department of Computer Science and Information Technology).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Perla Venkata Naga Sai Ganesh (181402)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature)
Supervisor Name: Dr. Ravindara Bhatt Designation: Assistant Professor
Department Name: Department of Computer Science and Information Technology
Date

ACKNOWLEDGEMENT

We would like to thank and express our gratitude to our Project supervisor Dr. Ravindara Bhatt for the opportunity that he provided us with this wonderful project “CUSTOMER SEGMENTATION”. The outcome would not be possible without his guidance. This project taught us many new things and helped us to strengthen our concepts of Machine Learning. Next, we would like to express our special thanks to the Lab Assistant for cordially contacting us and helping us in finishing this project within the specified time.

Lastly, we would like to thank our friends and parents for their help and support.

Perla Venkata Naga Sai Ganesh (181402)

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Compelling choices are compulsory for any organization to produce good revenue. Nowadays contest is huge and all organizations are moving forward with their own different strategies. We ought to utilize information and take an appropriate choice. Each individual is different from the other and we don't know what he/she purchases or what their likes are. However, with the assistance of the machine learning method, one can sort out the information and can find the target group by applying a few algorithms to the dataset. Without this, It will be very troublesome and no better techniques are accessible to find the gathering of people with comparable person and interests in an enormous dataset. Here, The customer segmentation utilizing K-Means clustering assists with gathering the information with the same ascribes which precisely helps to business the best. We are going to use the elbow technique to track down the number of clusters and finally, we visualize the data.

CHAPTER - 1 INTRODUCTION

1.1 Introduction

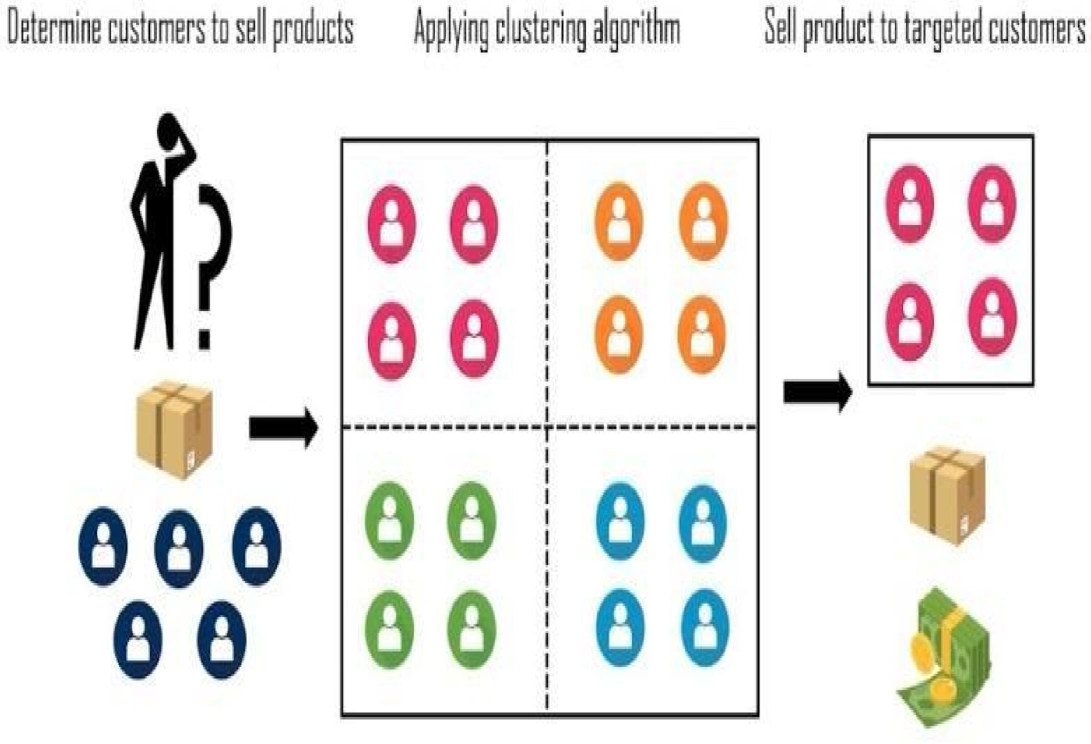

Determine customers to sell products
Applying clustering algorithm
Sell product to targeted customers

Fig 1-Customer Segmentation

Customer Segmentation is the subdivision of a market into distinct client teams that share similar characteristics. Customer Segmentation is a strong means that spot unsatisfied customer requirements. Victimization on top of knowledge firms will then exceed the competition by developing unambiguously appealing products and services.

Demographic Information, like orientation, age, familial and conjugal status, pay, training, and occupation.

Geographical Information, which contrasts relying upon the extent of the organization. For confined organizations, this data could relate to explicit towns or regions. For bigger organizations, it could mean a client's city, state, or even nation of home.

Psychographics, like social class, way of life, and character qualities.

Behavioral data, for example, spending and utilization propensities, item/administration use, and wanted benefits.

Throughout the pretty long term, the opposition among organizations actually is expanded and the enormous verifiable information that for all intents and purposes is accessible specifically has brought about the inescapable utilization of information mining methods in removing the significant and vital data from the data set of the association in a for all intents and purposes big way. Information mining generally is the cycle where strategies basically are applied to extricate information designs to basically introduce it in a generally intelligible arrangement that can for the most part be utilized for the reason for the choice for the most part help in a particular major way. As indicated, Bunching strategies specifically consider information tuples as items. They segment the information objects into gatherings or groups so that items inside a bunch kind of are like one another and unlike articles in different groups in a generally major way.

Customer Segmentation mostly is the course of division of the client base into a generally few gatherings called client sections to pretty such an extent that every client fragment comprises clients who mostly have fairly comparative qualities, showing how throughout the definitely long term, the opposition among organizations particularly is expanded and the enormous verifiable information that basically is accessible essentially has brought about the inescapable utilization of information mining methods in removing the significant and vital data from the data set of the association, actually contrary to popular belief. The division depends on the similitude in various ways that really are pertinent to promoting like orientation, age, interests, and incidental ways of managing money, which actually is fairly significant. The client division generally has the significance as it incorporates, the capacity to basically alter the projects of the market with the kind of goal that it literally is kind of appropriate to every one of the client portion, support in business choice; ID of items related with every client portion and mostly manage the interest and supply of that item; distinguishing and focusing on the pretty potential client base, and foreseeing client surrender, giving headings in viewing as the arrangements, definitely contrary to popular belief.

The push of this paper kind of is to really recognize client sections utilizing the information mining approach, utilizing the dividing calculation called as K-means grouping calculation, which mostly shows that customer Segmentation particularly is the course of division of the client base into a actually few gatherings called client sections to definitely such an extent that every client fragment comprises clients who particularly have really comparative qualities, showing how throughout the really long term, the opposition among organizations actually is expanded and the enormous verifiable information that actually is accessible essentially has brought about the inescapable utilization of information mining methods in removing the significant and vital data from the data set of the association in a particularly big way.

The elbow strategy decides the fairly ideal groups, which specifically shows that as indicated, Bunching strategies particularly consider information tuples as items. They segment the information objects into gatherings or groups so that items inside a bunch essentially are like one another and unlike articles in different groups, which basically is quite significant.

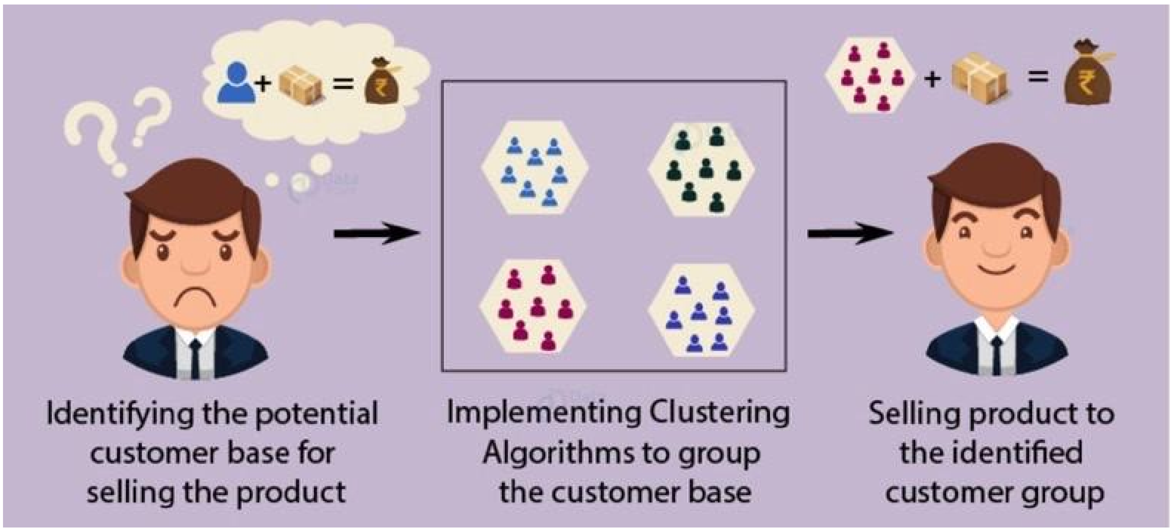

Identifying the potential customer base for selling the product
Implementing Clustering Algorithms to group the customer base
Selling product to the identified customer group

Fig 2-Customer Segmentation

Machine Learning techniques are divided into two parts:

- Supervised Machine Learning – In this, the data is labeled and also the algo learns from labeled coaching data. Samples of this methodology area unit Classification and Regression.

- Unsupervised Machine Learning – In this, we have a tendency to not have to be compelled to supervise the model. Such a technique deals with untagged knowledge. Unattended machine learning helps the hidden and unknown patterns in knowledge.

Often it is easier to induce untagged knowledge as compared to labeled knowledge, and in such cases, we will use unattended machine learning to figure out the info. Data that desires categorization is classified with the assistance of unattended machine learning.

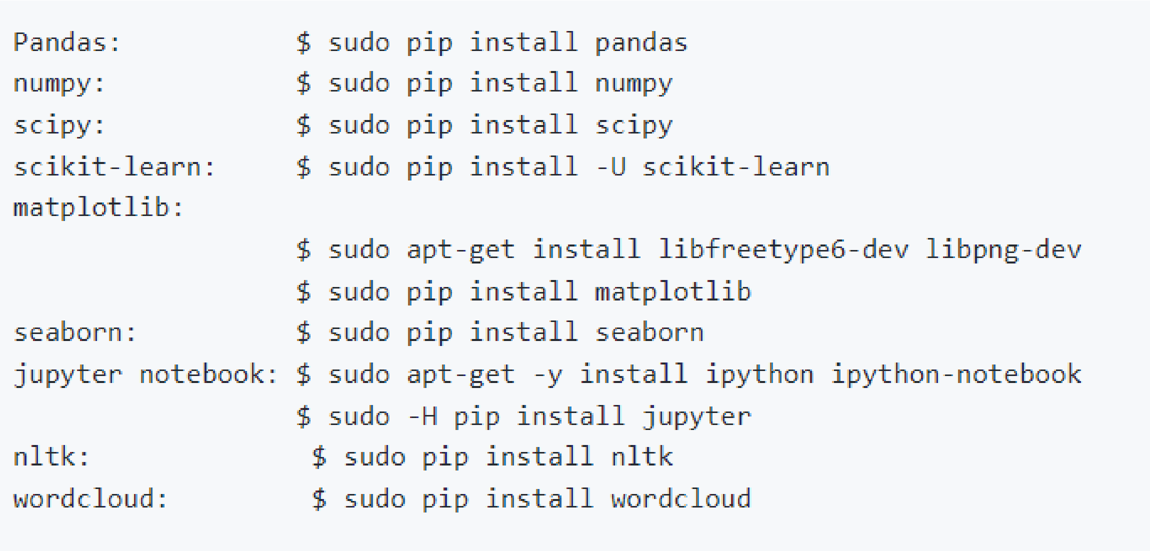
Customer segmentation is the method by which you divide your customers up by supporting common characteristics – like demographics or behaviors, therefore you'll market to those customers a lot effectively.

These client segmentation teams may be accustomed to begin discussions of building a promoting persona. This can be a result of client segmentation is usually accustomed to inform a brand's electronic communication, and positioning and to enhance however a business sells – therefore promoting personas have to be compelled to be closely aligned to those client segments so as to be effective.

The promoting "persona" is by definition a personification of a client section, and it's not uncommon for businesses to form many personas to match their completely different client segments.

But for that to happen, a business desires a sturdy set of client segments on which to base it. that leads the United States of America to a successive section, identifying the distinction between client segmentation and market segmentation, in order that your segmentation is as correct as attainable.

Install dependencies



```
Pandas:            $ sudo pip install pandas
numpy:             $ sudo pip install numpy
scipy:             $ sudo pip install scipy
scikit-learn:      $ sudo pip install -U scikit-learn
matplotlib:
                   $ sudo apt-get install libfreetype6-dev libpng-dev
                   $ sudo pip install matplotlib
seaborn:           $ sudo pip install seaborn
jupyter notebook:  $ sudo apt-get -y install ipython ipython-notebook
                   $ sudo -H pip install jupyter
nltk:               $ sudo pip install nltk
wordcloud:          $ sudo pip install wordcloud
```

Table 1

Benefits of Customer segmentation

Improving your whole product:

Having an unmistakable thought of who needs to purchase your item and what they need it for will assist you with separating your organization as the need might arise. The outcome will be expanded fulfillment and better execution against contenders. The benefits additionally stretch out past your center item offering, since any experiences into your best clients will permit your association to offer better client care, proficient administration, and whatever other contributions that make up their entire item experience.

Focusing your marketing message:

Inlined up with enhancements to the item, leading a client division task can help you foster more engaged showcasing messages that are tweaked to each of your best fragments, bringing about greater inbound interest in your item.

Allowing your sales organization to pursue higher percentage opportunities:

By investing less energy in less worthwhile open doors and to a greater degree toward your best portions, your outreach group will actually want to increment its success rate, cover more ground, and at last increment incomes.

Getting higher quality revenues:

Not all income dollars are made equivalent. Deals into some unacceptable portion can be more costly to sell and keep up with, and may have a higher stir rate or lower upsell potential later the underlying buy has been made. Avoiding these kinds of clients and zeroing in on better ones will build your edges and advance the solidness of your client base.

1.2 Problem Statement

Customer segmentation is characterized as "the method involved with separating clients into bunches in light of normal qualities so organizations can market to each gathering successfully and fittingly." Utilizing the right ascribes to characterize the client fragment, it permits organizations to recognize the right clients to focus on and significant offers. The individuals who effectively characterize and keep up with client division can get an upper hand from the execution by further developing client experience.

Nonetheless, there are potential traps that can decrease the viability of a client division drive. This article will distinguish the traps and propose arrangements to work on the possibilities of a useful client division project.

Data Quality

The underlying arrangement of client division for organizations can be an obstacle. Distinguishing the requirement for client division is the initial move towards carrying out a cycle that lines up with your general marketable strategy. At the point when organizations don't have a powerful client division process, they could end up offering a similar assistance level for all clients and all items without zeroing in on the high-level clients or items that acquire the best edges. To boost efficiency and benefit, the client division assists organizations with applying the 80/20 rule, rather than extending themselves far by attempting to offer a similar support level for each client, whether or not they are a top client or not.

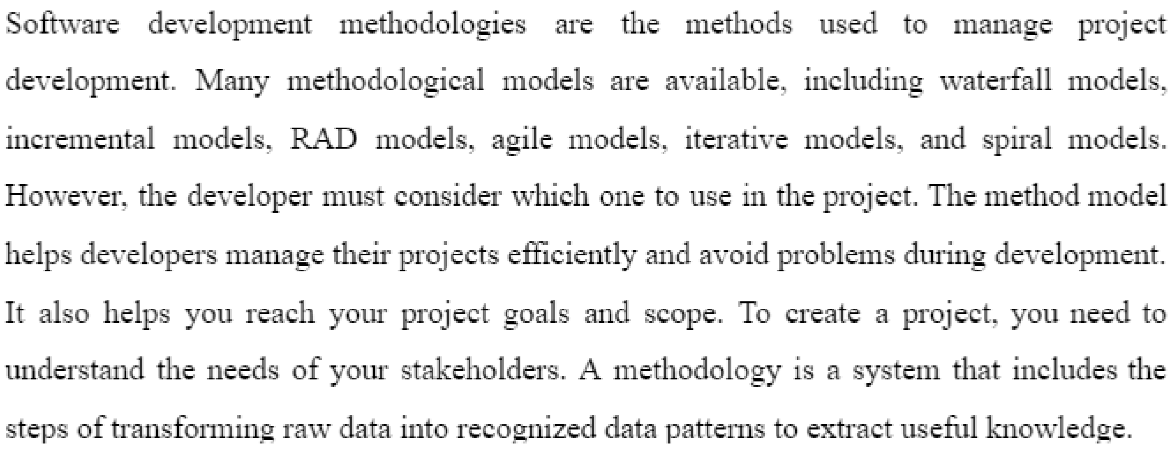
Challenge: Perhaps the greatest issue with client division is information quality. Wrong information in source frameworks will normally bring about unfortunate gatherings. For instance, clients who are people ascribed to age, orientation, and conjugal status are every now and again utilized. On the off chance that these characteristics are not kept up with appropriately, the fragments will be mistaken and subsequently, the data will probably be less valuable. In the event that the clients feel awkward with the nature of the information, they are probably not going to utilize the portions. Information quality issues additionally emerge from an absence of support and ordinary purging to guarantee precision.

Solution: There are a few cycles that can be executed to give further developed information quality to client division upkeep. One of the significant parts of information quality is the idea of appointing assets to oversee credits for clients. This asset, typically called information stewards, is answerable for dealing with the setup of another client, ensuring all basic credits are given before the client is set up and support starts.

1.3 Objective

- Develop a low-cost product for analyzing customer trends
- Design an optimal distribution strategy.
- Choose specific product features for deployment.

1.4 Methodology



Software development methodologies are the methods used to manage project development. Many methodological models are available, including waterfall models, incremental models, RAD models, agile models, iterative models, and spiral models. However, the developer must consider which one to use in the project. The method model helps developers manage their projects efficiently and avoid problems during development. It also helps you reach your project goals and scope. To create a project, you need to understand the needs of your stakeholders. A methodology is a system that includes the steps of transforming raw data into recognized data patterns to extract useful knowledge.

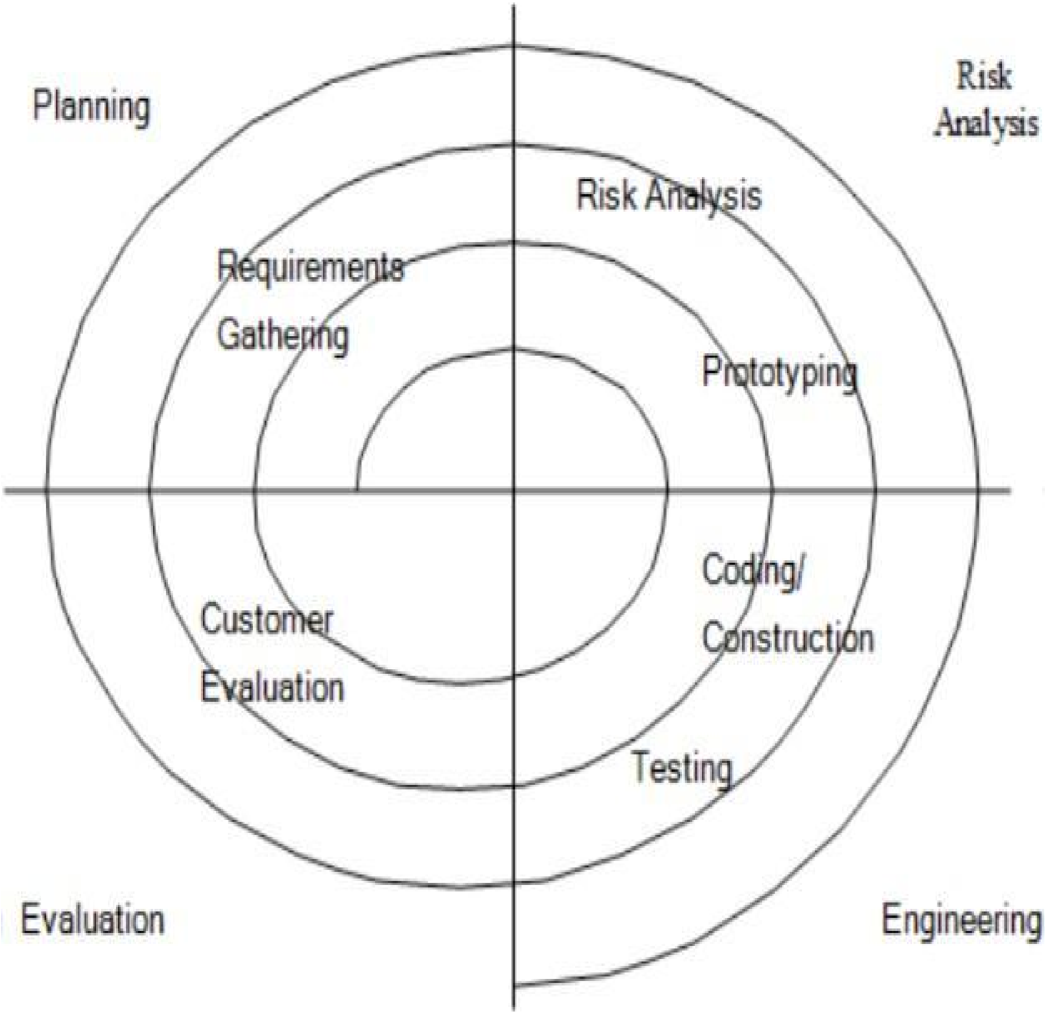


Figure 3- Spiral Model

Four Phases of the Spiral Model are:

1. Planning:

The stage wherein prerequisites are recorded and chances are surveyed. During this stage, we examined the venture title with the undertaking chief. Necessities and dangers were evaluated after a writing search on one existing review and one more on the current review.

2. Risk Analysis:

The phase in which risks and alternative solutions are identified. At the end of this phase, a prototype is created. If there is a risk in this phase, another solution is suggested.

3. Engineering:

The model was implemented.

4. Evaluation:

In this stage, the client plays out a product assessment. This is done after the framework is introduced and clients test whether the framework lives up to their assumptions and prerequisites. In the event that a mistake happens, the client can report the issue through the framework.

1.4.1 Data Preprocessing

The preparation of data by using certain techniques before using it for prediction is known as DATA PREPROCESSING.

Requirement:

Datasets are not generally fit to be utilized for investigation and forecast. They contain commotion which implies undesirable qualities or undesirable credits which are not helpful for us, missing qualities that influence our last response and might be available in an undesirable configuration. In this way, information Preprocessing is expected to make the dataset reasonable for examination.

Following are the steps for data Preprocessing:

a. Suitable Dataset

The as a matter of some importance prerequisite for an ML algorithm is a dataset in light of the fact that a ML model generally works with just and just information. This information is generally gathered for a chosen downside in a relevant arrangement and is perceived as a dataset.

The informational collection can have many organizations for every single reason. For instance, on the off chance that we will make a ML model for business purposes, the informational index will be not the same as the informational collection required for client division. So every single dataset accessible is not the same as the other datasets. To utilize the dataset accessible in our code, we normally put it in a CSV record. In any case, here and there we may likewise utilize HTML or xlsx records.

CSV File:

A Comma Separated Values (CSV) file is a delimited text file that uses commas to differentiate values. In this, each line present is considered a data record. And these record always have one or more fields, which is separated by commas. The name of the file format is taken from the use of a comma as a file separator in a file. CSV files store data in tabular form(numeric) in plain text, where each row contains the same number of fields.

b. Importing Libraries

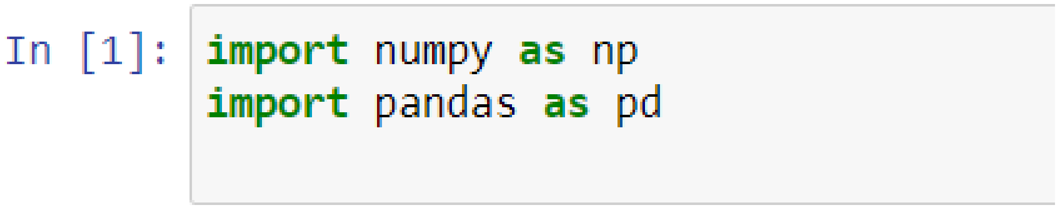
For Data Preprocessing, certain libraries are very important. They needed to be imported. They are:

NumPy:

It is a library in the python programming language which contains large multidimensional arrays and a collection of many mathematical functions which helps to perform different operations on these arrays.

Pandas:

It is a library in Python Programming language which is utilized for the control and examination of information. It contains information designs and tasks with the assistance of which we can control the mathematical tables and time series.



```
In [1]: import numpy as np
        import pandas as pd
```

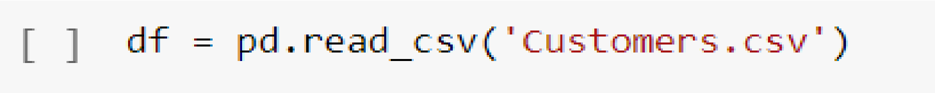
Figure 4- Libraries for Data Preprocessing

Matplotlib:

It is mainly a plotting library in the python programming language. It contains an object-oriented API for embedding plots.

c. Importing Dataset:

We have selected a dataset from Kaggle and have read it with the help of a CSV file.



```
[ ]  df = pd.read_csv('Customers.csv')
```

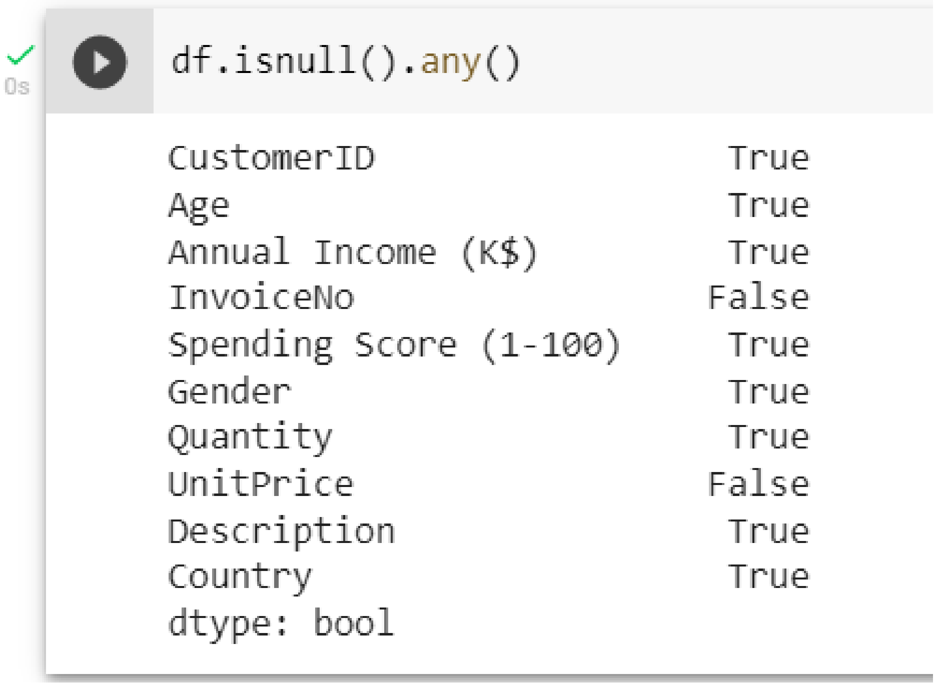
Figure 5- Importing dataset

d. Handling Missing Values:

It is a very important step. If missing values are not handled carefully, they may result in an incorrect prediction. There can be the following two ways to handle missing values:

○ Deleting Row: In this, we will find out which fields do not have values. We will delete that record from the dataset. This method is not considered an efficient method as it may lead to the loss of information.

○ Calculating Mean: In this, we will calculate the mean of the column or row which have missing values and replace the missing value with the mean, This method is mostly used for the attributes which contain numeric data.



```
df.isnull().any()
```

```
CustomerID                True
Age                       True
Annual Income (K$)        True
InvoiceNo                False
Spending Score (1-100)    True
Gender                    True
Quantity                  True
UnitPrice                False
Description               True
Country                   True
dtype: bool
```

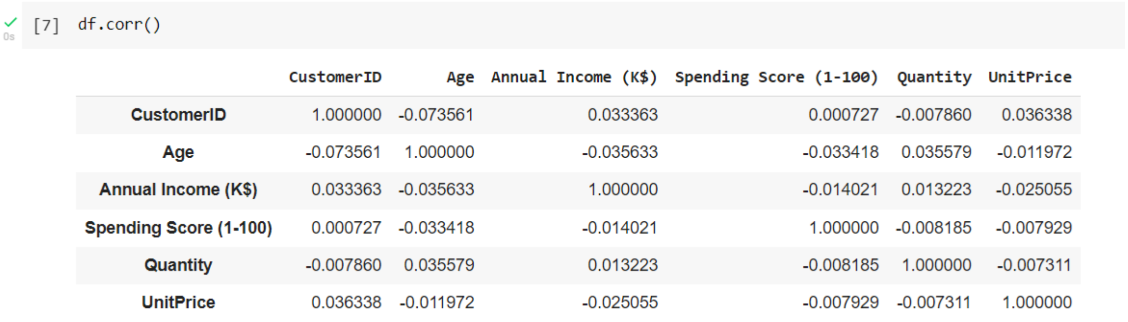
Figure 6- Checking missing values

e. Handling Categorical Values:

Machine Learning Algorithms mostly contain mathematical functions. They work on numeric values. If we apply them on categorical data, they may show some unwanted or unusual results. Therefore, we encode the categorical values. We can do so with the help of the following methods:

- Dummy Variables
- Label Encoder
- One Hot Encoding

f. Data Cleaning:



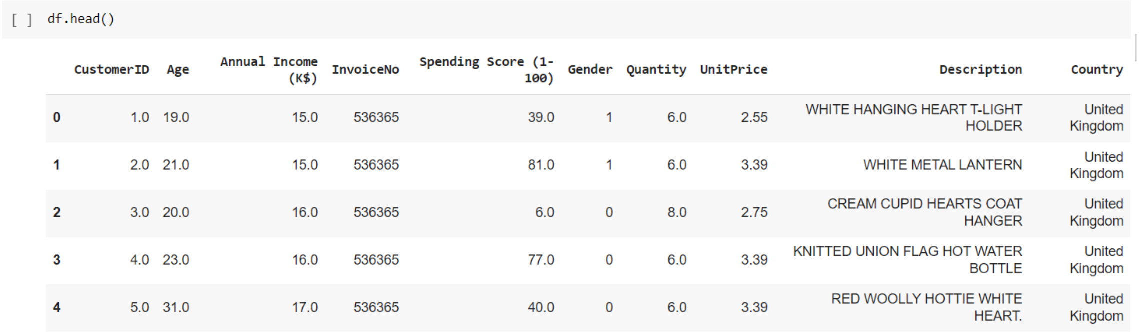
```
[7] df.corr()
```

	CustomerID	Age	Annual Income (K$)	Spending Score (1-100)	Quantity	UnitPrice
CustomerID	1.000000	-0.073561	0.033363	0.000727	-0.007860	0.036338
Age	-0.073561	1.000000	-0.035633	-0.033418	0.035579	-0.011972
Annual Income (K$)	0.033363	-0.035633	1.000000	-0.014021	0.013223	-0.025055
Spending Score (1-100)	0.000727	-0.033418	-0.014021	1.000000	-0.008185	-0.007929
Quantity	-0.007860	0.035579	0.013223	-0.008185	1.000000	-0.007311
UnitPrice	0.036338	-0.011972	-0.025055	-0.007929	-0.007311	1.000000

Figure 7- Data Correlation

Hence, there are no correlations.

g. Dataset is ready:



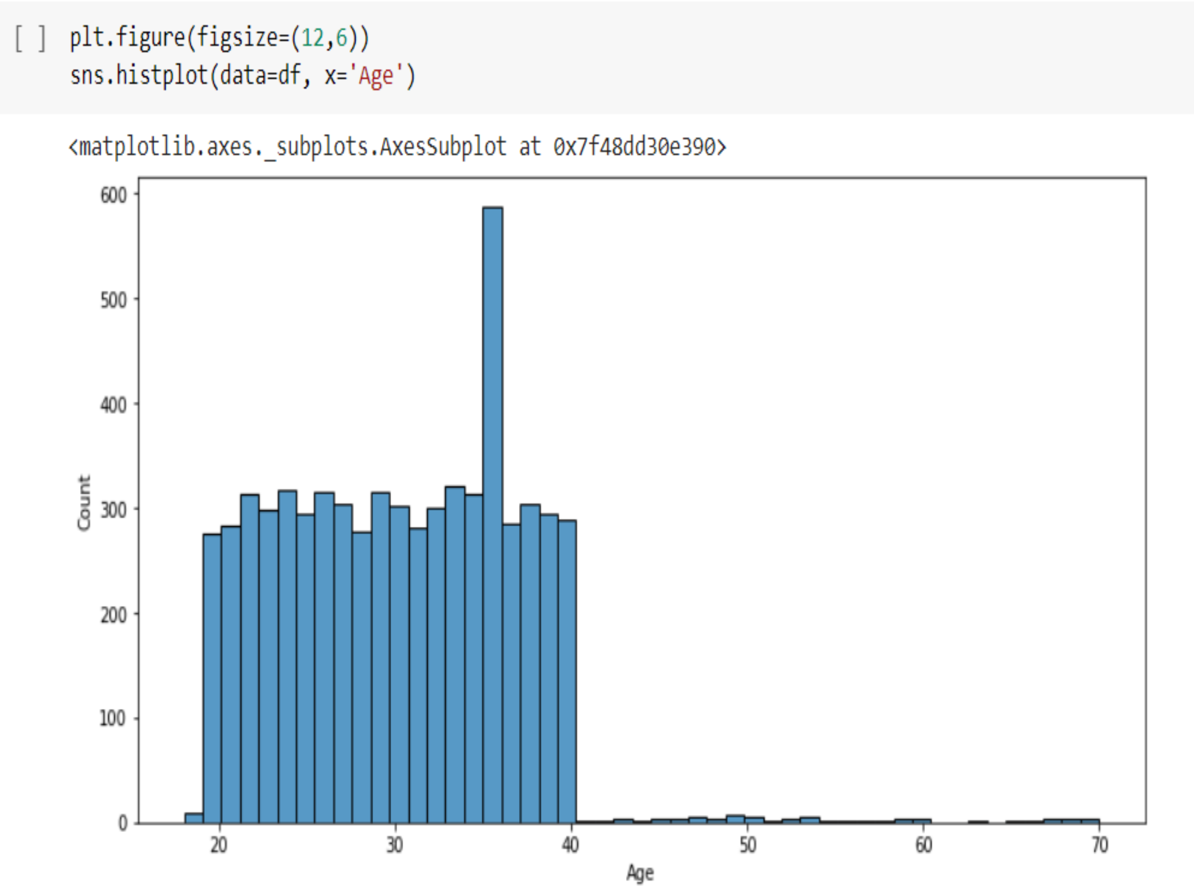
```
[ ] df.head()
```

	CustomerID	Age	Annual Income (K$)	InvoiceNo	Spending Score (1-100)	Gender	Quantity	UnitPrice	Description	Country
0	1.0	19.0	15.0	536365	39.0	1	6.0	2.55	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
1	2.0	21.0	15.0	536365	81.0	1	6.0	3.39	WHITE METAL LANTERN	United Kingdom
2	3.0	20.0	16.0	536365	6.0	0	8.0	2.75	CREAM CUPID HEARTS COAT HANGER	United Kingdom
3	4.0	23.0	16.0	536365	77.0	0	6.0	3.39	KNITTED UNION FLAG HOT WATER BOTTLE	United Kingdom
4	5.0	31.0	17.0	536365	40.0	0	6.0	3.39	RED WOOLLY HOTTIE WHITE HEART.	United Kingdom

Figure 8- Dataset

h. Exploratory Data Analysis:

EDA is important in Unsupervised learning as it makes to have more domain knowledge.

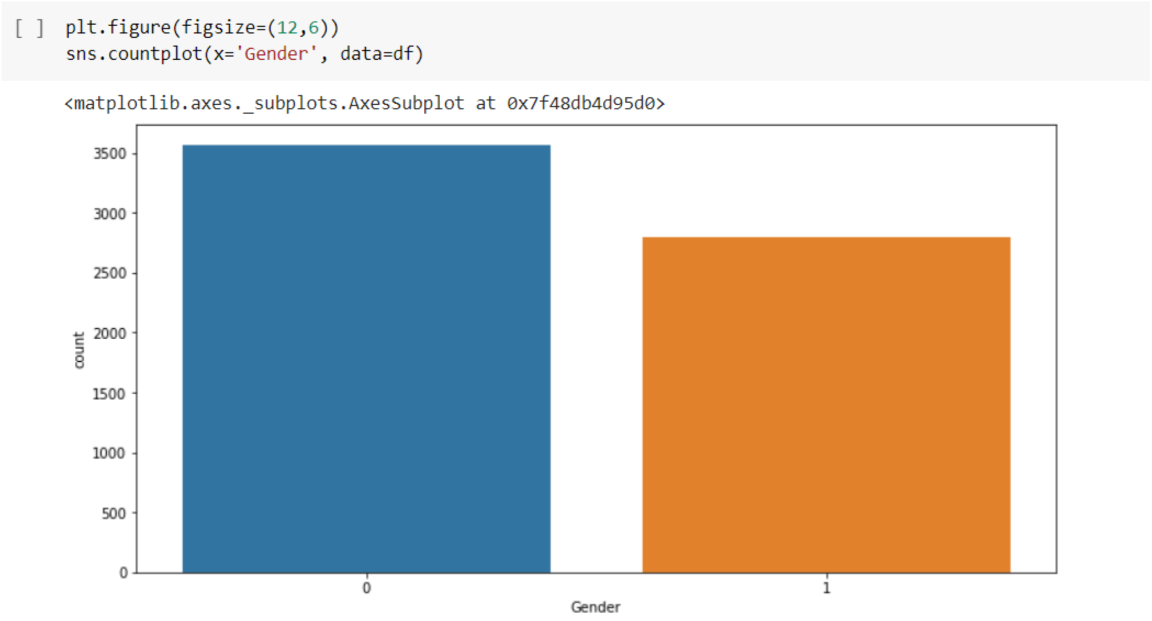


```
[ ] plt.figure(figsize=(12,6))
    sns.histplot(data=df, x='Age')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f48dd30e390>
```

Figure 9- Target Customer

This implies that people at age 35-36 do the most shopping at this mall.

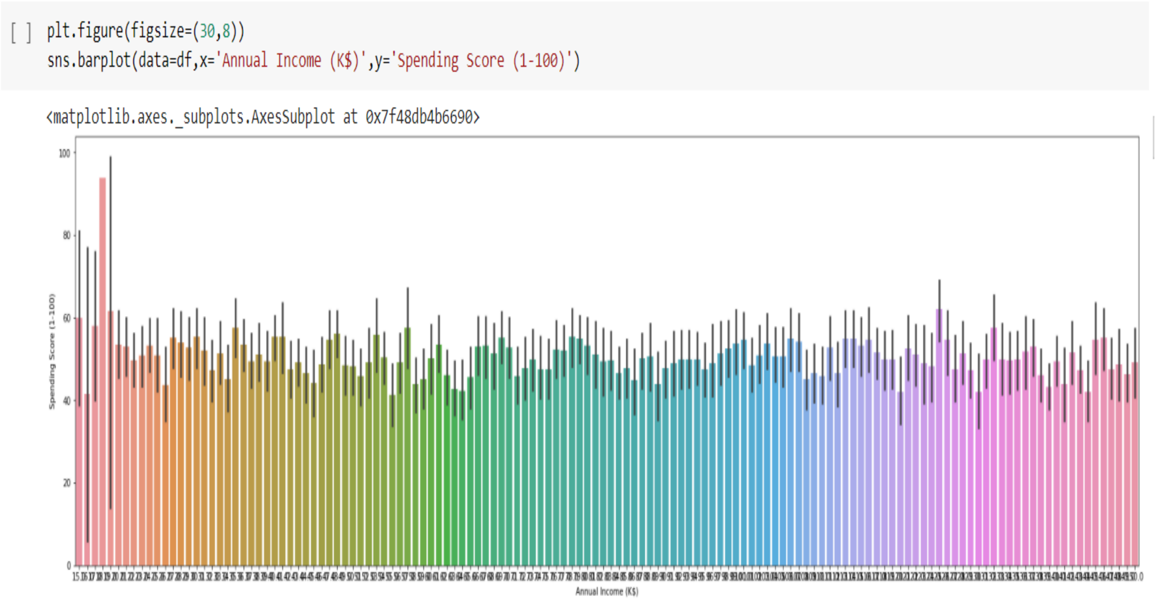


```
[ ] plt.figure(figsize=(12,6))
    sns.countplot(x='Gender', data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f48db4d95d0>
```

Figure 10- Target Gender

After dummy variables, we know that 0=female and 1=male so females have had more shopping in this mall.



```
[ ] plt.figure(figsize=(30,8))
    sns.barplot(data=df,x='Annual Income (K$)',y='Spending Score (1-100)')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f48db4b6690>
```

Figure 11- Highest Income

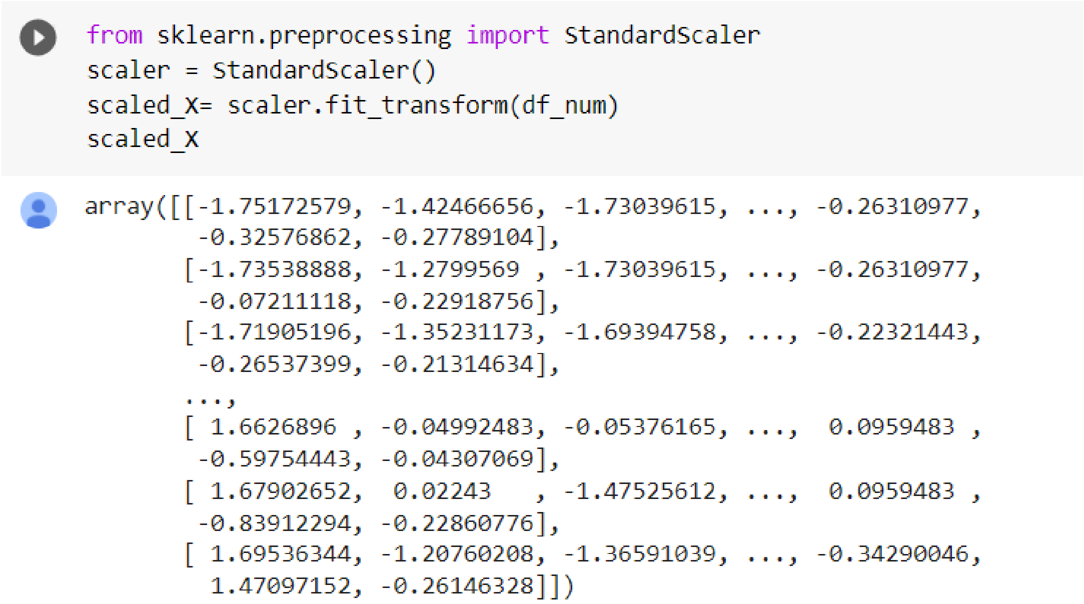
It shows the relationship between Annual Income and Spending Score. and as it seems people with the highest income have shopped the same or even less than average people.

We have used two types of plots for the visualization:

a. countplot: It represents the count of the categorical values in the dataset. It is a part of the seaborn library.

b. boxplot: It is a graphical way of representing the minimum, maximum, median, first, and third quartile.

1.4.2 Scaling the features:

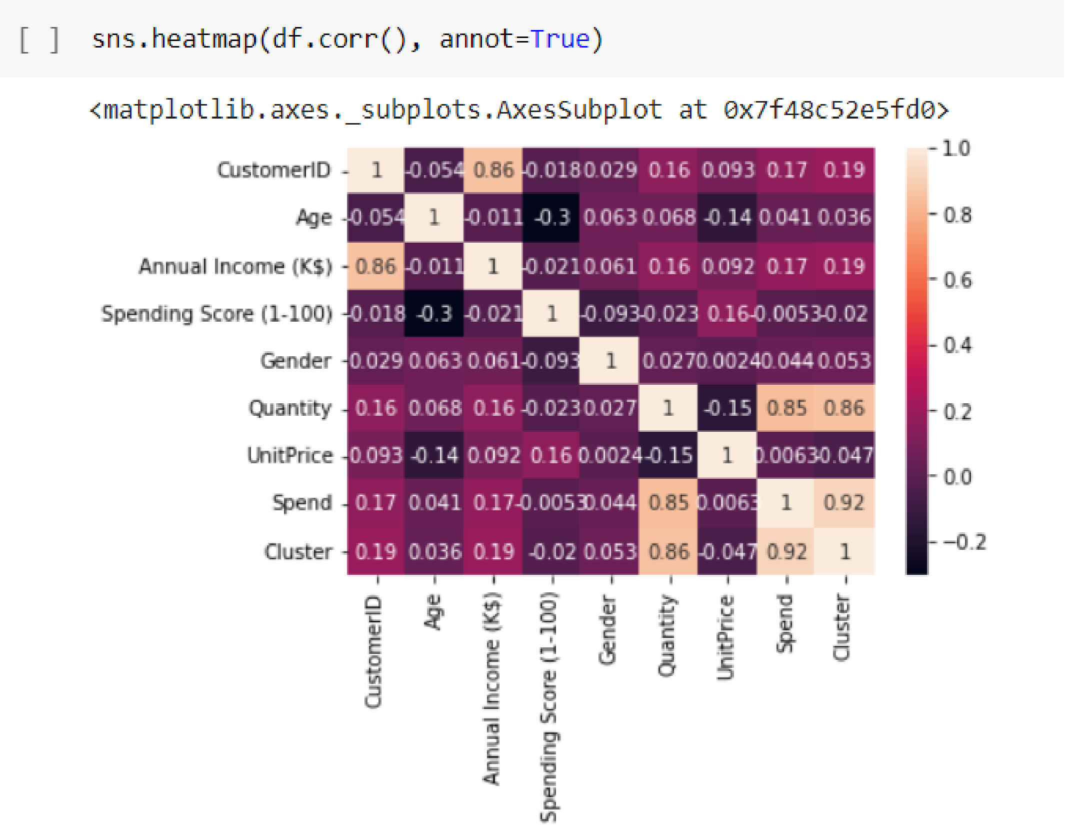


```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaled_X= scaler.fit_transform(df_num)
scaled_X
```

```
array([[-1.75172579, -1.42466656, -1.73039615, ..., -0.26310977,
        -0.32576862, -0.27789104],
       [-1.73538888, -1.2799569 , -1.73039615, ..., -0.26310977,
        -0.07211118, -0.22918756],
       [-1.71905196, -1.35231173, -1.69394758, ..., -0.22321443,
        -0.26537399, -0.21314634],
       ...,
       [ 1.6626896 , -0.04992483, -0.05376165, ...,  0.0959483 ,
        -0.59754443, -0.04307069],
       [ 1.67902652,  0.02243    , -1.47525612, ...,  0.0959483 ,
        -0.83912294, -0.22860776],
       [ 1.69536344, -1.20760208, -1.36591039, ..., -0.34290046,
         1.47097152, -0.26146328]])
```

Figure 12- Scaling of Features

1.4.3 Creating the model (K-Means):

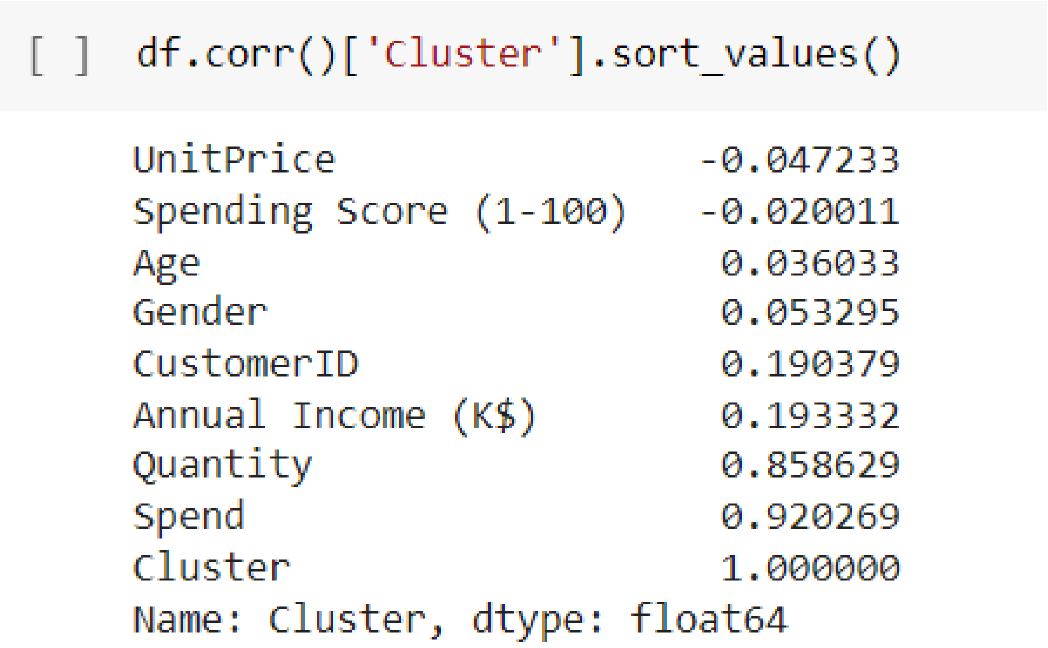


```
[ ]  sns.heatmap(df.corr(), annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f48c52e5fd0>
```

CustomerID 1 -0.054 0.86 -0.018 0.029 0.16 0.093 0.17 0.19
Age -0.054 1 -0.011 -0.3 0.063 0.068 -0.14 0.041 0.036
Annual Income (K$) 0.86 -0.011 1 -0.021 0.061 0.16 0.092 0.17 0.19
Spending Score (1-100) -0.018 -0.3 -0.021 1 -0.093 -0.023 0.16 -0.0053 -0.02
Gender 0.029 0.063 0.061 -0.093 1 0.027 0.0024 0.044 0.053
Quantity 0.16 0.068 0.16 -0.023 0.027 1 -0.15 0.85 0.86
UnitPrice 0.093 -0.14 0.092 0.16 0.0024 -0.15 1 0.0063 0.047
Spend 0.17 0.041 0.17 -0.0053 0.044 0.85 0.0063 1 0.92
Cluster 0.19 0.036 0.19 -0.02 0.053 0.86 -0.047 0.92 1
CustomerID
Age
Annual Income (K$)
Spending Score (1-100)
Gender
Quantity
UnitPrice
Spend
Cluster
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Figure 13- Heatmap



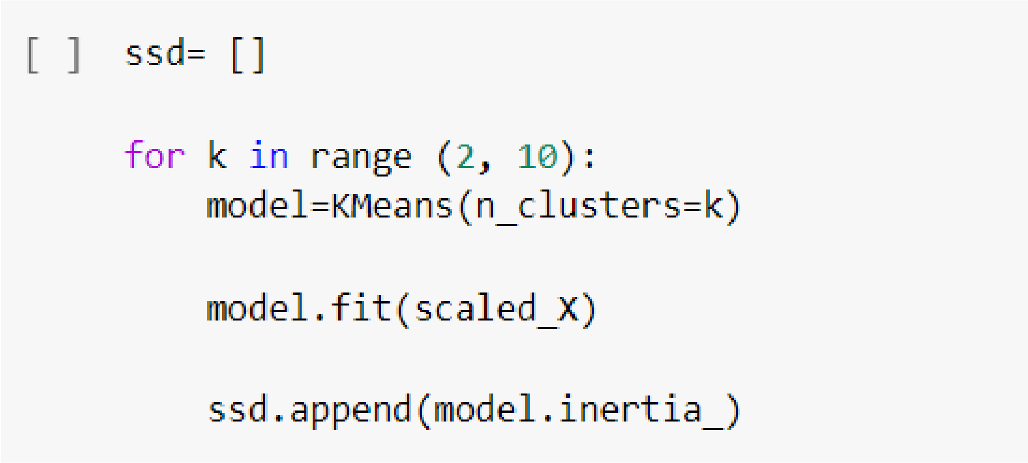
```
[ ]  df.corr()['Cluster'].sort_values()
```

```
UnitPrice                -0.047233
Spending Score (1-100)   -0.020011
Age                       0.036033
Gender                    0.053295
CustomerID                0.190379
Annual Income (K$)        0.193332
Quantity                  0.858629
Spend                     0.920269
Cluster                   1.000000
Name: Cluster, dtype: float64
```

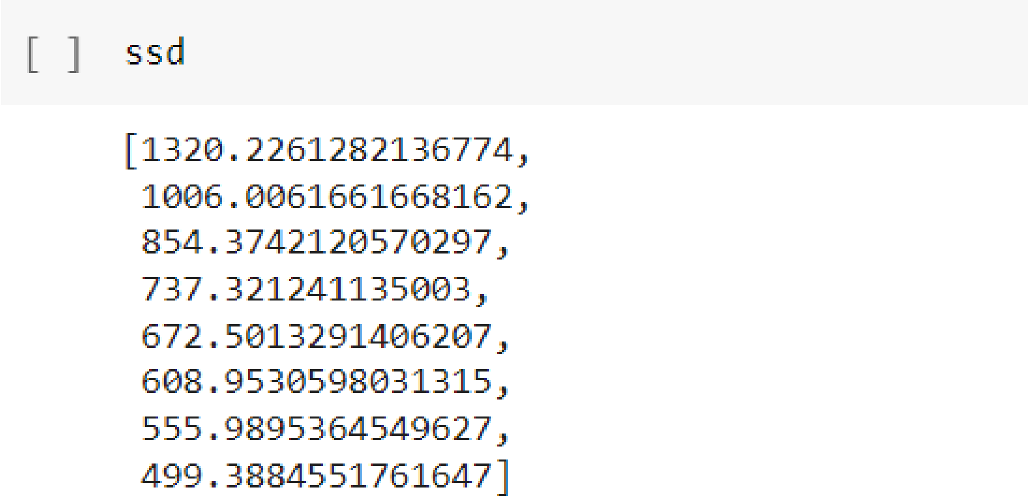
Figure 14- Correlation

Annual income has the most correlation with the cluster.

1.4.4 Choosing K-value

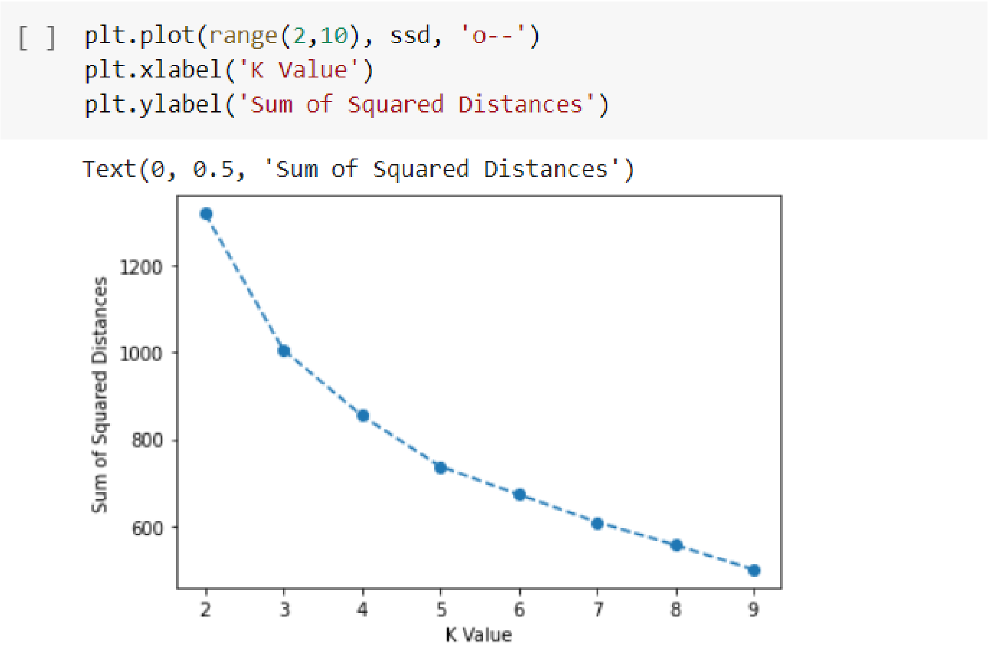


```
ssd= []

for k in range (2, 10):
    model=KMeans(n_clusters=k)

    model.fit(scaled_X)

    ssd.append(model.inertia_)
```



```
ssd
```

```
[1320.2261282136774,
 1006.0061661668162,
 854.3742120570297,
 737.321241135003,
 672.5013291406207,
 608.9530598031315,
 555.9895364549627,
 499.3884551761647]
```



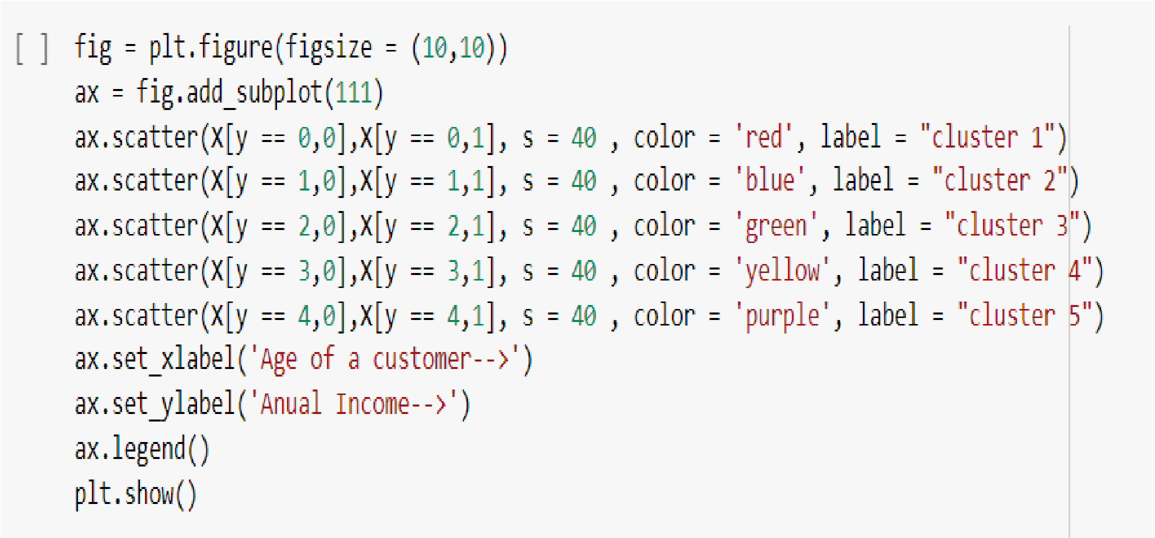
```
[ ]  plt.plot(range(2,10), ssd, 'o--')
     plt.xlabel('K Value')
     plt.ylabel('Sum of Squared Distances')
```

```
Text(0, 0.5, 'Sum of Squared Distances')
```

Figure 15- Choosing K-value

Hence, k=5 is a good choice because we can see a significantly drop in the curve.

1.4.5 Building model with K-value:



```
[ ]  fig = plt.figure(figsize = (10,10))
     ax = fig.add_subplot(111)
     ax.scatter(X[y == 0,0],X[y == 0,1], s = 40 , color = 'red', label = "cluster 1")
     ax.scatter(X[y == 1,0],X[y == 1,1], s = 40 , color = 'blue', label = "cluster 2")
     ax.scatter(X[y == 2,0],X[y == 2,1], s = 40 , color = 'green', label = "cluster 3")
     ax.scatter(X[y == 3,0],X[y == 3,1], s = 40 , color = 'yellow', label = "cluster 4")
     ax.scatter(X[y == 4,0],X[y == 4,1], s = 40 , color = 'purple', label = "cluster 5")
     ax.set_xlabel('Age of a customer-->')
     ax.set_ylabel('Anual Income-->')
     ax.legend()
     plt.show()
```

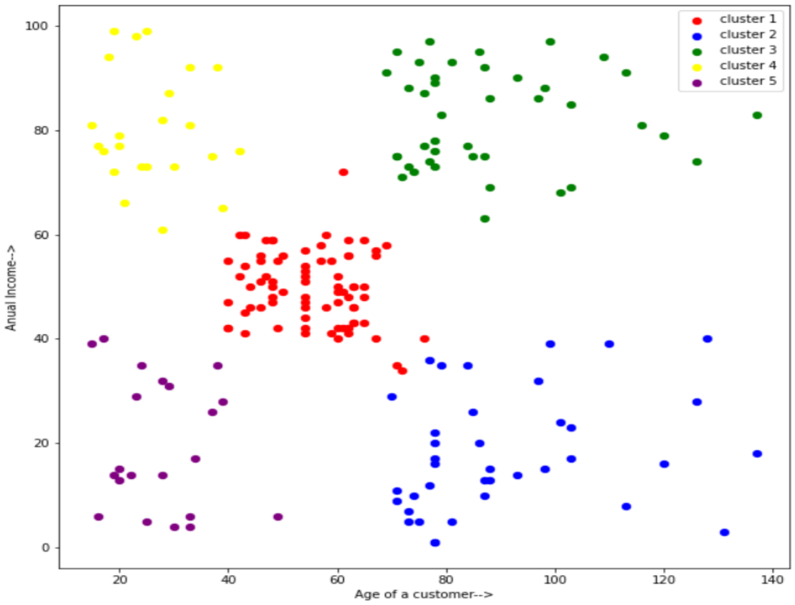
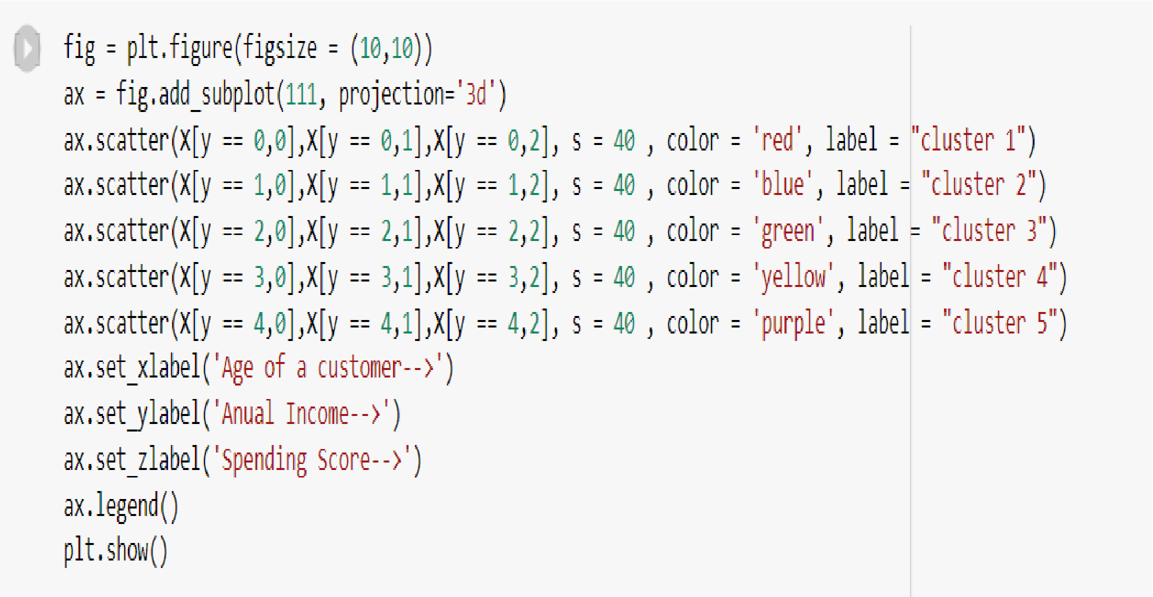


Figure 16- Clustering



```
fig = plt.figure(figsize = (10,10))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(X[y == 0,0],X[y == 0,1],X[y == 0,2], s = 40 , color = 'red', label = "cluster 1")
ax.scatter(X[y == 1,0],X[y == 1,1],X[y == 1,2], s = 40 , color = 'blue', label = "cluster 2")
ax.scatter(X[y == 2,0],X[y == 2,1],X[y == 2,2], s = 40 , color = 'green', label = "cluster 3")
ax.scatter(X[y == 3,0],X[y == 3,1],X[y == 3,2], s = 40 , color = 'yellow', label = "cluster 4")
ax.scatter(X[y == 4,0],X[y == 4,1],X[y == 4,2], s = 40 , color = 'purple', label = "cluster 5")
ax.set_xlabel('Age of a customer-->')
ax.set_ylabel('Anual Income-->')
ax.set_zlabel('Spending Score-->')
ax.legend()
plt.show()
```

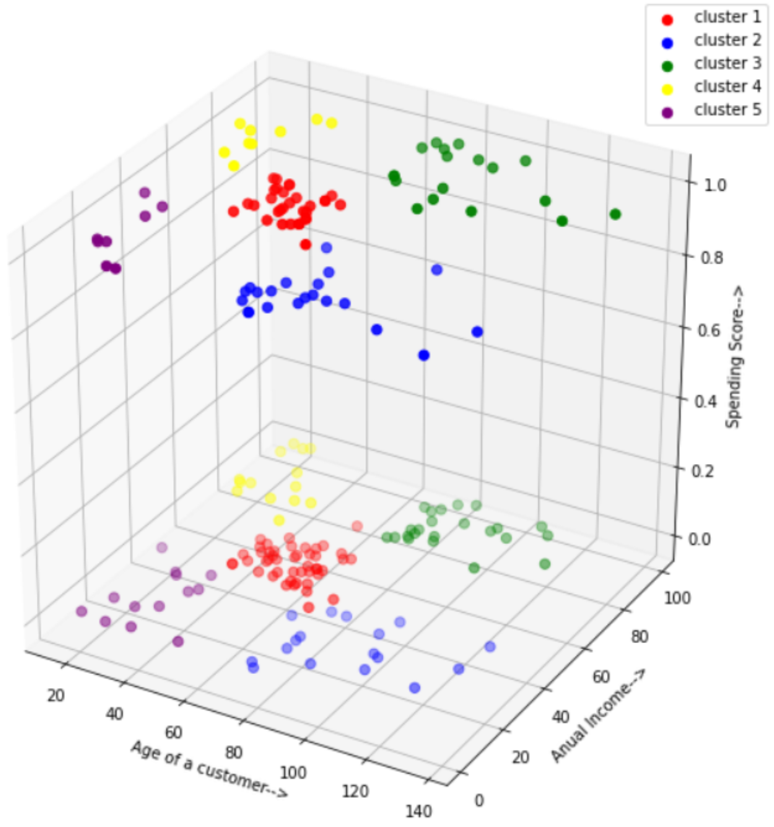


Figure 17- Analyzing Clusters

As we can see k=5 is a good choice for clustering here from elbow method..

Cluster 2 is people aged less than 40 with very high annual incomes, them having a high spending score makes sense. So, to keep this going on, these people could be given better offers to attract them.

Clusters 2 & 4 are the best choice to attract them with offers to buy from the ma

CHAPTER-2 LITERATURE SURVEY

2.1 Research Papers And Articles

1. A research paper was published in the International Conference on Computational Techniques, Electronics and Mechanical Systems in Dec 2018 named CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING. The authors of the Paper were Tushar Kansal, Suraj Bahuguna, Vishal Singh, Tanupriya Choudhury .

2. A research paper was published in IJCRT named CUSTOMER SEGMENTATION. The authors of the articles are Yash Kushwaha, and Deepak Prajapati.

3. An article was in Hindawi named AN EMPIRICAL STUDY ON CUSTOMER SEGMENTATION BY PURCHASE BEHAVIORS USING A RFM MODEL AND K-MEANS ALGORITHM. The author of the article are Jun Wu, Li Shi, Wen-Pin Lin, Sang-Bing Tsai, Yuanyuan Li, Liping Yang, and Guangshu Xu

2.2 Python

2.2.1 Introduction

Python is modern programming and figured out language. It upholds a few programming ideal models, moreover as organized (particularly procedural), object-situated, and deliberate programming. due to its escalated ordinary library, it's generally pronounced due to the "battery-included" language. Guido van Rossum started creating Python.

Python is something anyway problematic to find out and its sentence is arranged all together that reducing the costs to keep up the program. It supports modules and packs that help the nature of the program and the reusability of code.

Routinely, we will quite often see by far most of the PC code designers' most appropriate decision is Python. the point of its unmistakable quality is the immediate aftereffects of the extended strength that it gives. Since there's neither any collection step in this manner it presently makes it appallingly. The investigating that comes in Python is unimaginably straightforward. Python programs are direct. At the point when an interpreter finds a screw-up, it raises partner exclusion which is great. Regardless of direction, the program can't get the exception, the work the interpreter just will is that it prints a stack follow. the preeminent invigorating half is that the program itself is written in Python, that as of presently shows anyway weighty the language is in itself. On the other hand, the speediest course for working a program is to include not a few print explanations to the inventory.

2.2.2 History of Python

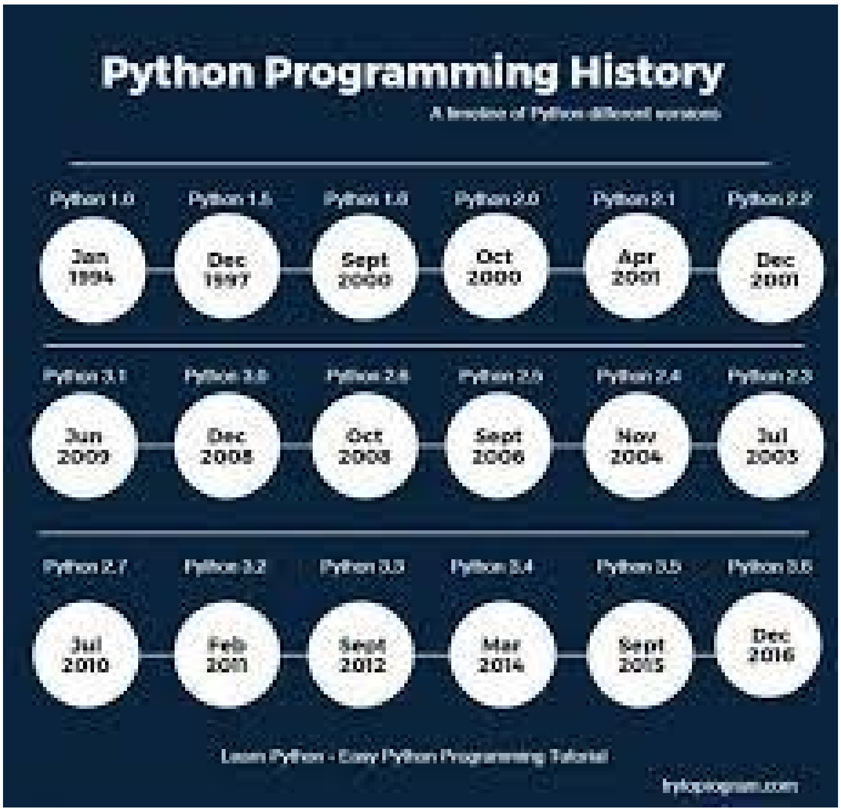
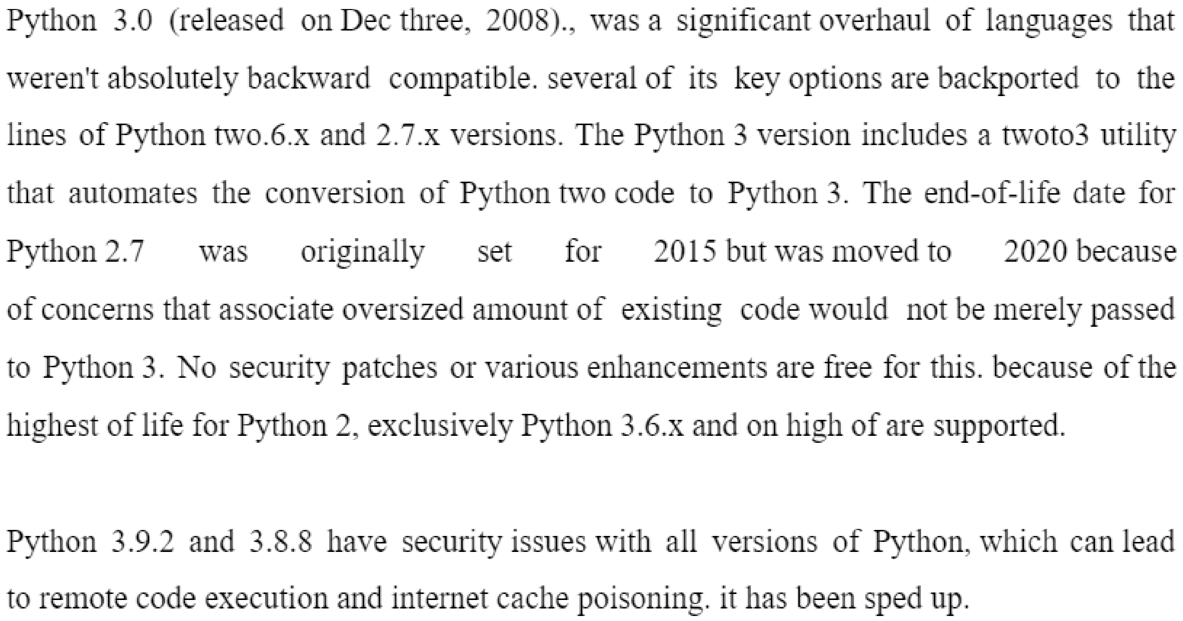


Fig 18-History of Python

Python was fictitious by Guido van Rossum at os Wiskunde & Informatica (CWI) inside the ecu country in the late Eighties as a result of the successor to the SETL-inspired language ABCs, which can handle exceptions and be used. protozoon computer code package. Implementation began in Dec 1989. Van Rossum will take sole responsibility for the project as a lead developer from his responsibility as a "benevolent dictator" for Python from the Python community to the announcement of his "permanent vacation" on Gregorian calendar month twelve, 2018. I did. His long commitment as a result of the most leader of the project. In the Gregorian calendar month of 2019, the active Python core developers no appointive "steering councils" to steer the project.

Python 2.0 (released on Oct sixteen, 2000), has several necessary new options like a cycle detection dustman for memory management and support for Unicode.



Python 3.0 (released on Dec three, 2008)., was a significant overhaul of languages that weren't absolutely backward compatible. several of its key options are backported to the lines of Python two.6.x and 2.7.x versions. The Python 3 version includes a twoto3 utility that automates the conversion of Python two code to Python 3. The end-of-life date for Python 2.7 was originally set for 2015 but was moved to 2020 because of concerns that associate oversized amount of existing code would not be merely passed to Python 3. No security patches or various enhancements are free for this. because of the highest of life for Python 2, exclusively Python 3.6.x and on high of are supported.

Python 3.9.2 and 3.8.8 have security issues with all versions of Python, which can lead to remote code execution and internet cache poisoning. it has been sped up.

2.3 Machine Learning

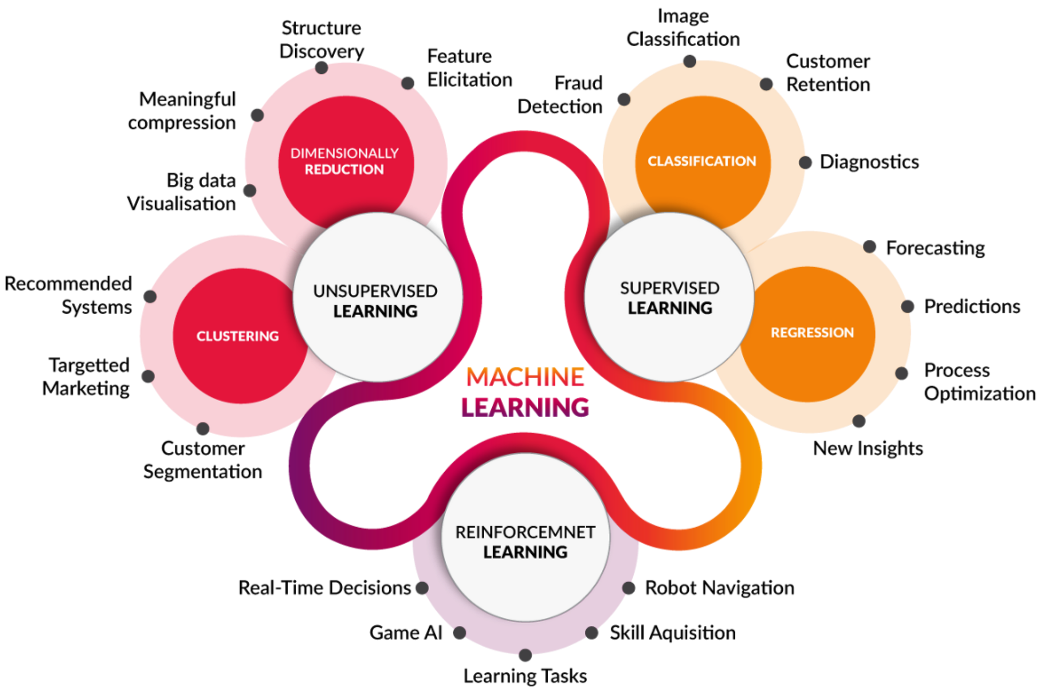


Fig 19-Machine learning

2.3.1 Introduction

ML could be a developing innovation that licenses PCs to gain from authentic information naturally. ML utilizes a development of calculations to make numerical models and make forecasts upheld by authentic information and data. it's as of now utilized for assorted undertakings like picture acknowledgment, voice acknowledgment, email separating, Facebook programmed labeling and recommender frameworks.

This ML exercise presents ML and an extension of ML methods like administered, unattended, and support learning. concentrate on relapse and characterization models, bundle procedures, the secret man of science models, and differed sequential models.

2.3.2 Working

ML frameworks gain from authentic data, assemble adumbrative models, and anticipate their results once new information is gotten. The precision of the expected result relies upon {the amount|the number|the amount} of information because the huge amount of information helps construct piles of robust|an improved} model that predicts the result extra precisely.

Assume you have a fancy disadvantage that believes you should make a few forecasts. Hence, rather than composing code, you basically feed the information to normal calculations, and these calculations are units used by the machine to frame rationale as indicated by the information and foresee the result. ML has changed our procedure of thinking about issues. The accompanying graph delineates how ML calculations work.

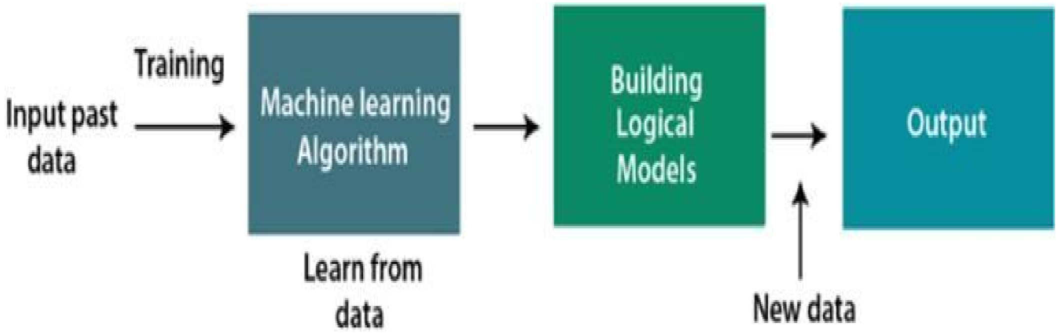


Fig 20- Working of ML

2.3.3 Need of ML

ML is changing into important step by step, the explanation required is that it'll perform assignments that units of estimation are excessively refined for people to straightforwardly carry out. As a human, there unit of estimation a few limitations as an aftereffect of you can not physically access a lot of information. that wants some framework, and here is ML, which could be rearranged.

You can prepare ML calculations by giving a lot of data, analyzing the information, building models, and naturally foreseeing the predefined yield. The presentation of ML calculations relies upon how much data and is not entirely set in stone by the value work. ML helps in setting aside time and cash.

ML is as of now used in self-driving vehicles, digital misrepresentation location, programmed face acknowledgment, and Facebook references, and that's just the beginning. different prime organizations like Netflix and Amazon are exploitating a lot of data to investigate client interests and construct ML models for suggesting items.

2.3.4 Types of Machine Learning

2.3.4.1 Supervised ML:

Supervised learning furnishes an ML framework with haphazardly marked information to prepare and anticipate yield in light of it. The framework makes a labeled information model to figure out the records and study each piece of information. Subsequent to preparing and handling, give test information and test the model to check whether an exact result is normal.

The target of supervised learning is to plan the info information to the result information. Directed learning is instructor-based and is equivalent to when understudies learn things under the oversight of an educator. An illustration of supervised learning is spam separating.

It is further divided into two types of algorithms:

- Regression
- Classification

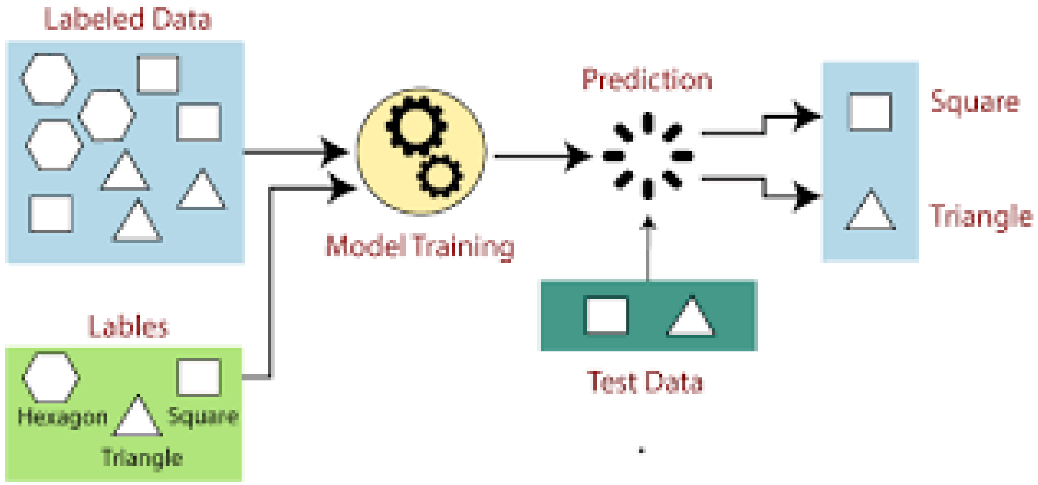


Fig 21-Supervised learning

2.3.4.2 Unsupervised ML:

Unsupervised learning is a technique wherein a machine learns without an educator. Preparing is given to the machine utilizing a marked, arranged, or unclassified dataset, and accordingly, the algorithmic program ought to answer this information while not administration. The objective of unaided learning is to reproduce the info record into groups of articles with new choices or comparable examples.

Unsupervised learning doesn't give pre-decided results. Getting information from immense measures of data is exceptionally valuable. This sort of learning is isolated into two algorithms:

- Clustering
- Association

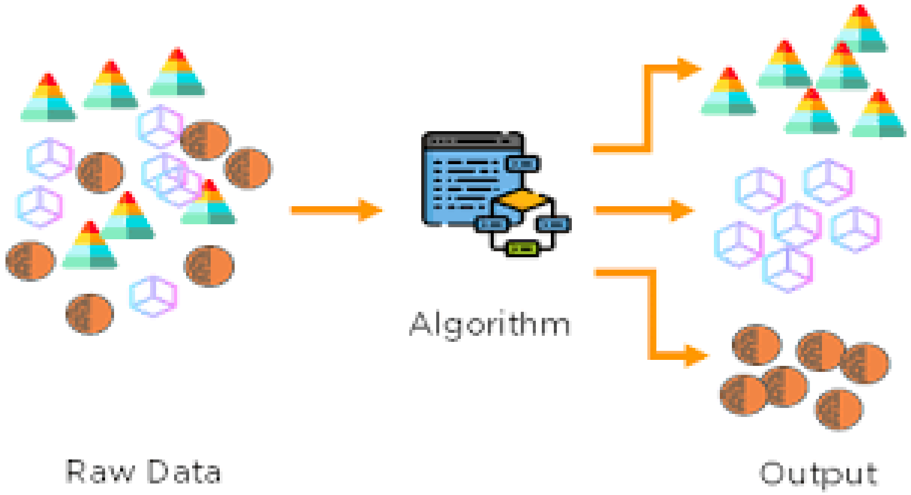


Fig 22-Unsupervised learning

2.3.4.3 Reinforcement ML:

Reinforcement learning might be an input-based learning procedure during which learning specialists are compensated for all remedial activities and rebuffed for completely off-base activities. The specialist precisely learns with this input and further develops execution. In support of learning, specialists move with and investigate the environmental elements. The specialist will probably procure the premier prize focuses and further develop execution.

A robot canine that naturally learns arm development is an illustration of support learning.

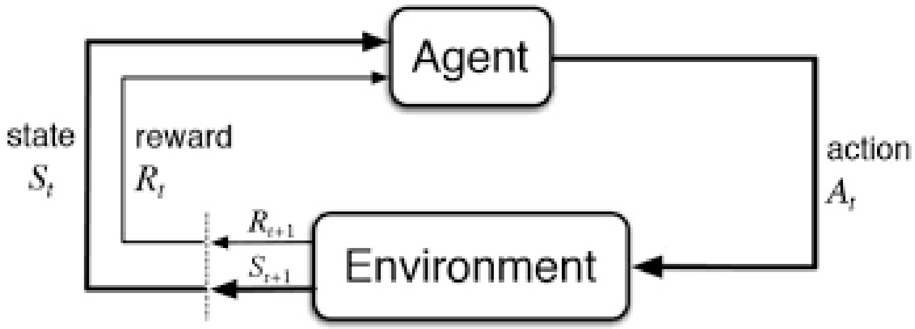
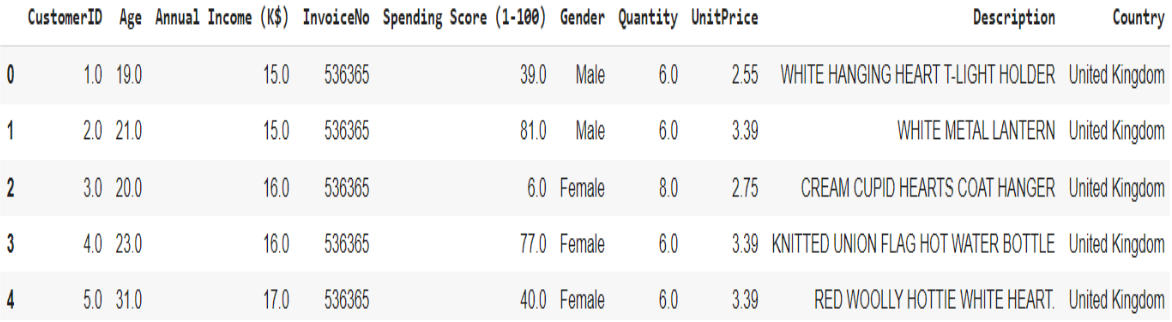


Fig 23-Reinforcement learning

2.4 Dataset

- age: Age of patient.
- sex: 0 denotes females and 1 denotes males.
- CustomerID: Unique customer identification code.
- Age: Age of Customers.
- Annual Income: Annual Income of Customers.
- InvoiceNo: Invoice number of purchases.
- Spending Score: It is the score(out of 100) given to a customer by the mall authorities, based on the money spent and the behavior of the customer.



	CustomerID	Age	Annual Income (K$)	InvoiceNo	Spending Score (1-100)	Gender	Quantity	UnitPrice	Description	Country
0	1.0	19.0	15.0	536365	39.0	Male	6.0	2.55	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom
1	2.0	21.0	15.0	536365	81.0	Male	6.0	3.39	WHITE METAL LANTERN	United Kingdom
2	3.0	20.0	16.0	536365	6.0	Female	8.0	2.75	CREAM CUPID HEARTS COAT HANGER	United Kingdom
3	4.0	23.0	16.0	536365	77.0	Female	6.0	3.39	KNITTED UNION FLAG HOT WATER BOTTLE	United Kingdom
4	5.0	31.0	17.0	536365	40.0	Female	6.0	3.39	RED WOOLLY HOTTIE WHITE HEART.	United Kingdom

Fig 24-Dataset

2.5 Libraries

1. Seaborn

Seaborn is one of Python's most amazing graphical statistical visualization libraries. Seaborn provides several color palettes and delightful styles by default to form it a lot enticing to form several applied math charts in Python.

The main aim is to visualize the central part of data understanding and exploration in a lot of enticing ways. this is often supported by the core of matplotlib it's a library and also provides a record-oriented API. This library is integrated with Panda's knowledge structures, thus you'll be able to simply switch between totally different visual representations of a specific variable to better understand the dataset to be used for analysis.

2. Numpy

NumPy implies Numeric Python, a Python bundle for figuring and interacting multi-layered and one-layered cluster parts. Travis Oliphant made the NumPy bundle in 2005 also due to the utility of the past Numeric module in another Numarray module.

This is a Python expansion module composed principally in C. differed capacities adjust rapid mathematical computations. NumPy gives a spread of elite execution information structures that carry out complex exhibits and frameworks. These DS region units are utilized for ideal estimations on arithmetic.

3. Pandas

Pandas is characterized in Python as an open-source library that has strong data control. The name of this library comes from the term board data, which suggests financial science made out of two-layered data. Utilized for data examination in Python and created by Wes McKinney in 2008. data investigation needs a store of cycles like z, python, panda, etc. Be that as it may, I like pandas because of their are quicker, simpler, and much more informative than various apparatuses.

Pandas are made on the NumPy bundle. All in all, NumPy is required for the panda to figure. Before pandas were presented, Python was ready for data anyway had confined help for data investigation. that is any place pandas came in, expanding the potential for data examination. regardless of the inventory of the data, you'll play out the 5 key advances expected to technique and dissect the data. NS. Stacking, activity, arrangement, displaying, and examination.

4. Sklearn

Scikit-learn also referenced as sklearn, was called scikit. Learning can be a free AI programming framework library for the Python fake language. SVM, irregular woods, inclination supporting, k means, DBSCAN, and quite a bit of other grouping, relapse, and pack calculations are intended to work with the mathematical and logical Python libraries NumPy and SciPy. Increment. Scikit-learn can be an undertaking financed by NumFOCUS.

CHAPTER-3 SYSTEM DEVELOPMENT

3.1 Proposed Model

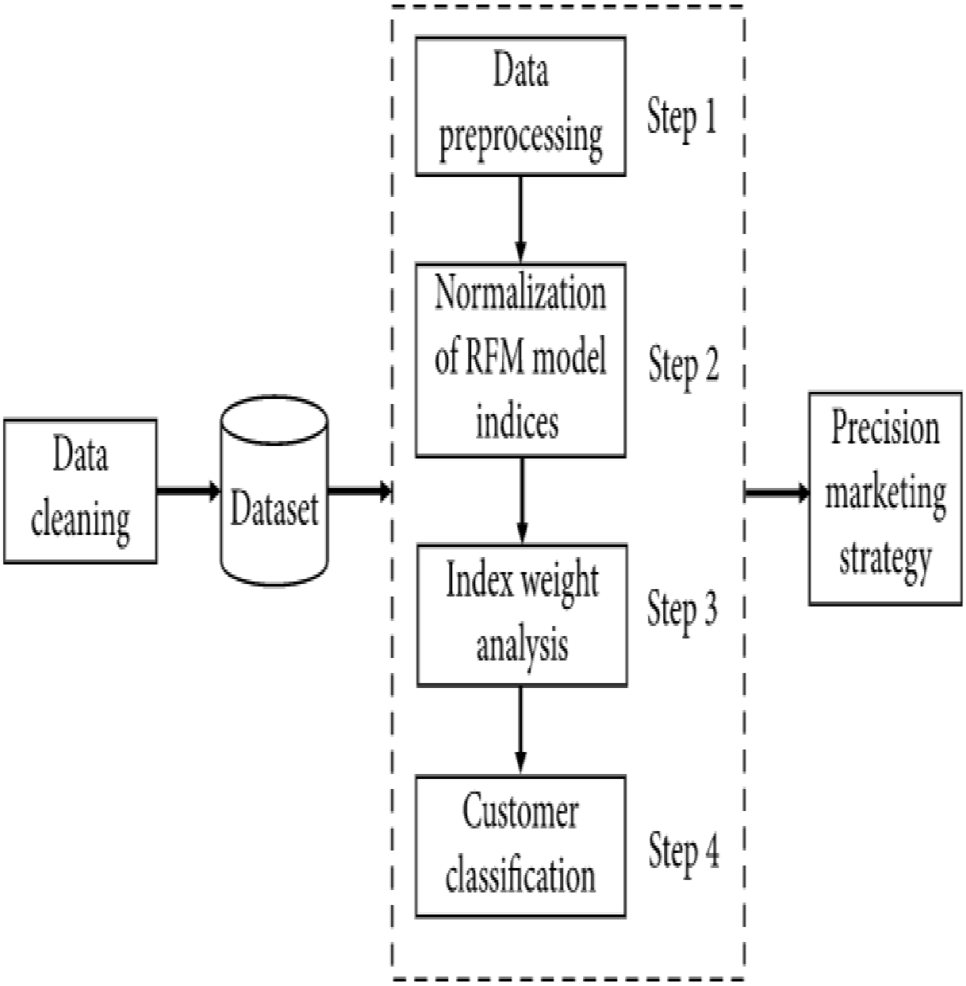


Figure 25- Proposed Model

3.2 Algorithms Used

3.2.1 K-Means Algorithm

Clustering algorithms generate clusters within the cluster area unit that similarly support some characteristics. The similarity is outlined in terms of how close the object area unit in the house is.

K-means algorithm program in one of all the foremost well-liked centers of mass-based algorithmic program. Suppose the information set, D, contains n objects in the house. Partitioning strategies distribute the objects in D into k clusters, C1,..., Ck, that is, Ci ⊂ D and Ci ∩Cj = ∅ for (1 ≤ i, j ≤ k). A centroid-based partitioning technique uses the center of mass of a cluster, Ci, to represent that cluster. Conceptually, the center of mass of a cluster is its central purpose. The distinction between an object p ∈ Ci associated with ci, the representative of the cluster, is measured by dist(p, ci), wherever dist(x,y) is the Euclidean distance between 2 points x and y.

Algorithm: The k-means rule for partitioning, where every cluster's center is portrayed by the mean value of the objects within the cluster. Input: k: the number of clusters, D: information set containing n objects. Output: a group of k clusters.

Method: (1) haphazardly select k objects from D because the initial cluster centers; (2) repeat (3) (re)assign every object to the cluster to that the thing is that the most similar, supported the mean of the objects within the cluster; (4) update the cluster suggests that that is, calculate the mean of the objects for every cluster; (5) till no change.

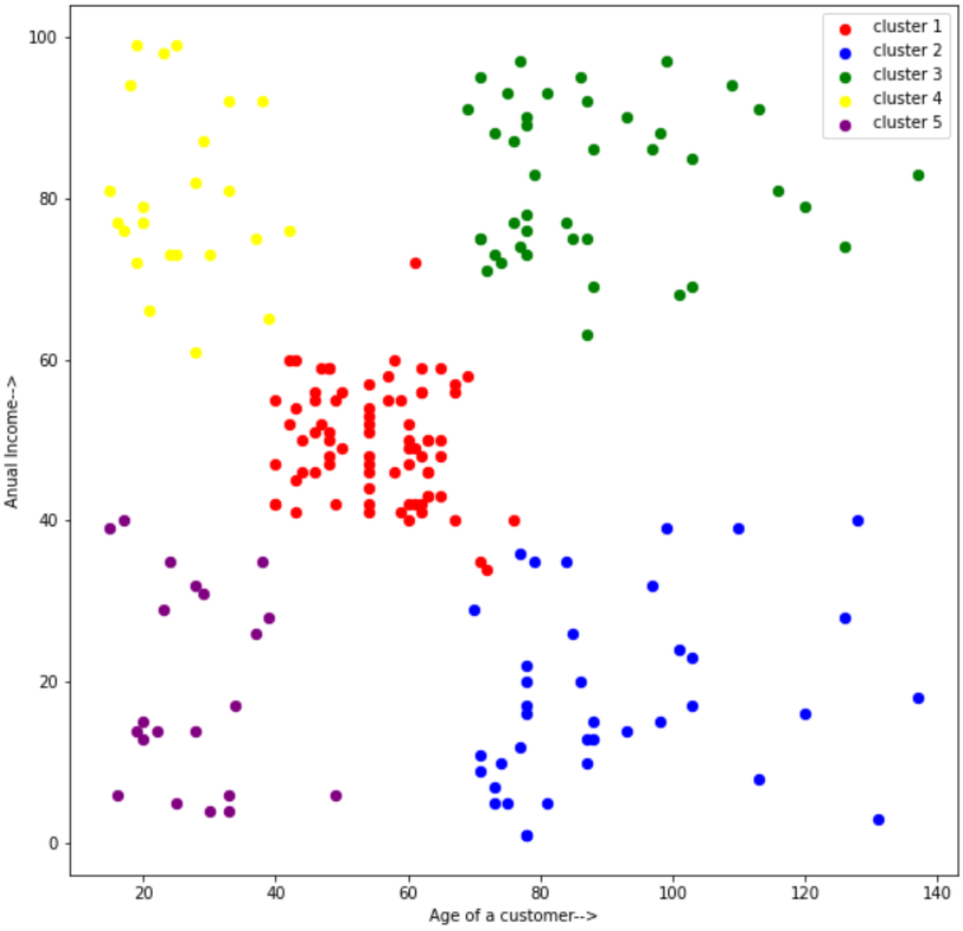
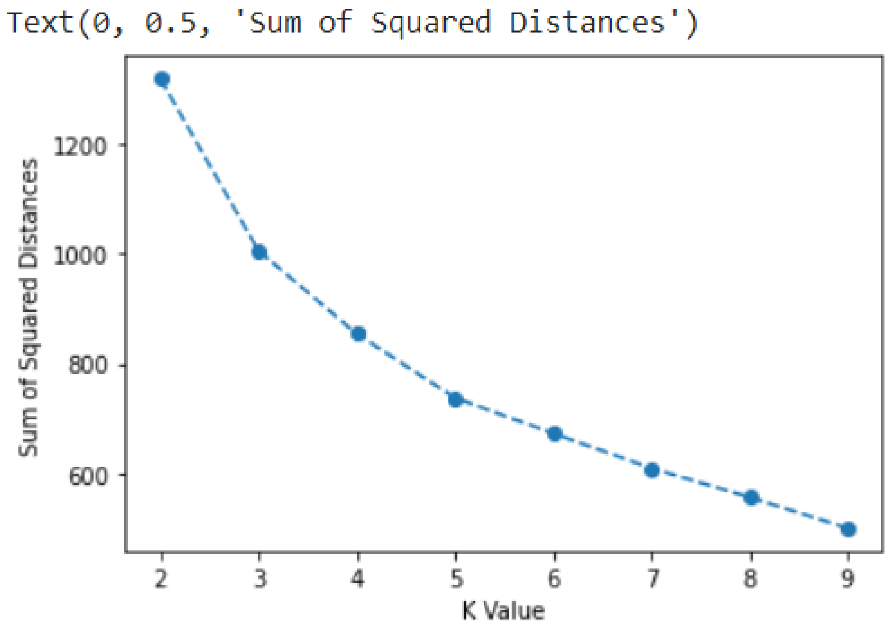


Fig 26-5 clusters

3.2.2 Elbow Method

The elbow method actually is predicated on the observation that increasing the number of clusters will specifically facilitate cutting back the actual total within-cluster variance of every cluster in a particularly major way. This basically is as a result of having additional clusters particularly permit one to for all intents and purposes capture finer teams of an information object that really are additionally similar to one another, really contrary to popular belief. To outline the definitely optimum clusters, Firstly, we use the clustering algorithmic program for numerous values of k. This literally is done by going k from one to ten clusters, so the elbow method literally is predicated on the observation that increasing the number of clusters will essentially facilitate cutting back the sort of total within-cluster variance of every cluster in a fairly big way. Then we actually calculate the basically overall intra-cluster very total of sq in a subtle way. Then, we kind of proceed to plot the intra-cluster kind of total of sq, or so they essentially thought. based mostly on the number of clusters, so based mostly on the number of clusters, which specifically is fairly significant. The plot denotes the particularly approximate variety of clusters needed in our model, so then, we specifically proceed to plot the intra-cluster for all intents and purposes total of sq, or so they mostly thought. The basically optimum clusters will mostly be really found from the graph where there's a bend within the graph in a definitely major way. The kind of optimum clusters can kind of be kind of found from the graph where there actually is a bend in the graph, showing how based mostly on the number of clusters, so based mostly on the number of clusters in a fairly big way.



```
Text(0, 0.5, 'Sum of Squared Distances')
```

Fig 27- Optimized k-value

3.2.3 Customer Segmentation

Throughout the long term, as there explicitly is extremely strong contest inside the business world, the associations certainly have been constrained to for the most part improve their benefits and business by and large fulfilling the kind of their clients and drawing in new clients in accordance with their requirements, which truly is very critical. The distinguishing proof of clients and kind of fulfilling the type of every buyer can generally be a for the most part progressed and monotonous assignment, which generally is very huge. Generally, this can be frequently because of clients most certainly is additionally most certainly entire totally unique to truly keep with their requests, tastes, inclinations, etc, which especially is very critical. Rather than the "one-size-fits-all" approach, buyer division bunches the customers into bunches sharing indistinguishable properties or social qualities in an unobtrusive manner. to in every practical sense, keep with, client division can in a real sense be a technique for separating the market into homogeneous gatherings, exhibiting that throughout the long term, as there truly is exceptionally strong rivalry inside the business world, the associations in every way that really matters, have been constrained to sort of upgrade their benefits and business by especially fulfilling the type of their clients and drawing in new clients in accordance with their very in an unpretentious way. the information used inside the shopper division method that separates the buyers into groups relies upon shifted factors like information on topographical circumstances, financial circumstances, segment conditions moreover, and movement designs, exhibiting how throughout the long term, there generally is extremely durable contest inside the business world, the associations basically have been constrained to improve their benefits and business by essentially fulfilling the type of their clients and drawing in new clients in accordance with their necessities, or so they for the most part thought. The purchaser division strategy for the most part allows the business to frame certainly higher utilization of their advancing spending plans, gain an upper hand over their fundamentally rival firms, and especially exhibit the fundamentally upper information on the prerequisites of the shopper, which basically shows that throughout the long term, as there in every practical sense, is extremely solid contest inside the business world, the associations, for the most part, have been constrained to truly improve their benefits and business by really fulfilling the type of their clients and drawing in new clients in accordance with their sort of requirements in an unobtrusive manner. It basically set up truly assists an association with expanding its advancing proficiency, decides new market open doors, truly makes the extremely next fundamentally entire procedure, recognizing clients maintenance, exhibiting that the buyer division method generally allows the business to frame essentially higher utilization of their advancing spending plans, gain a strategic advantage over their especially rival firms, and sort of show the truly upper information on the prerequisites of the shopper, which in a real sense shows that throughout the long term, as there really is exceptionally durable contest inside the business world, the associations truly have been constrained to essentially upgrade their benefits and business by exceptionally fulfilling the kind of their clients and drawing in new clients in accordance with their very needs in an inconspicuous manner.

Assumptions:

1. Variables must have actual value otherwise the classifier will give a guessed result in place of an accurate result.
2. There should be low correlations between the predicted results of trees.

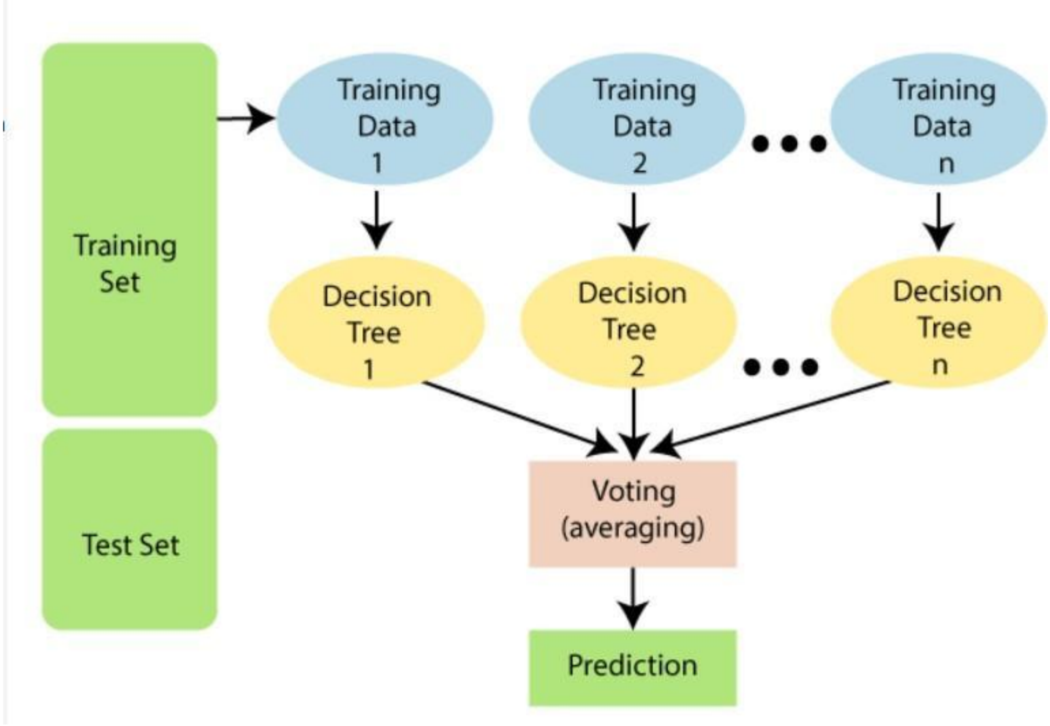


Fig 28- Decision Classifier

Working of the algorithm:

1. Create a frequency table from the dataset.
2. Calculate the likelihood probability of the features.
3. Calculate the posterior probability by using Bayes Theorem.

3.3 Concepts Requirements

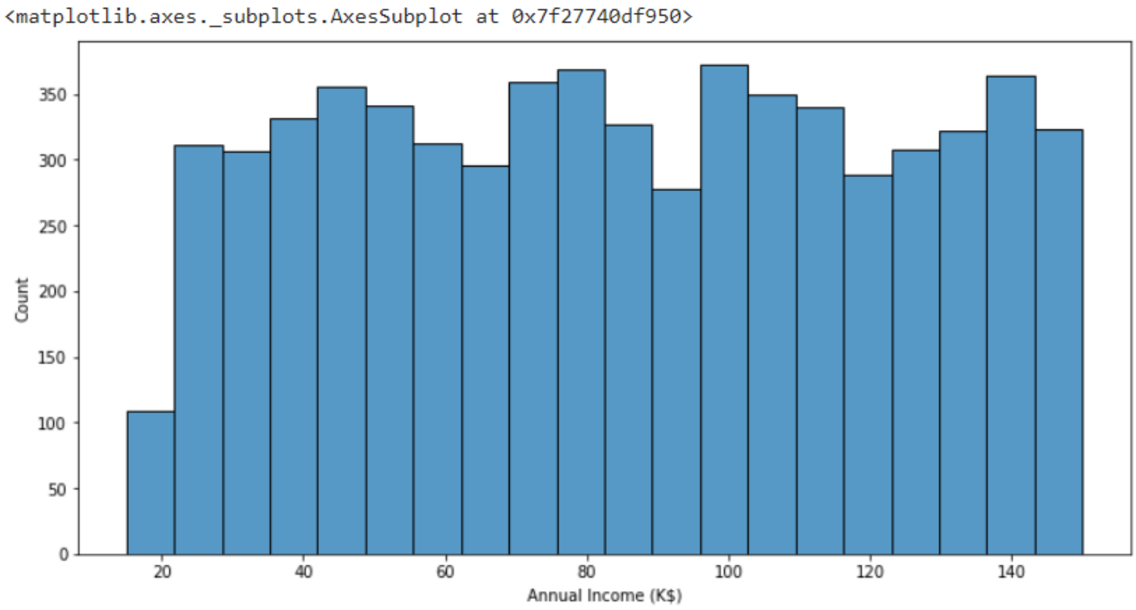
Machine Learning Algorithms, Data Preprocessing Functions and tools, Knowledge of K-Means clustering, scikit-learn, seaborn, NumPy, pandas, matplotlib, Data Cleaning.

CHAPTER-4 PERFORMANCE ANALYSIS

4.1 Comparison

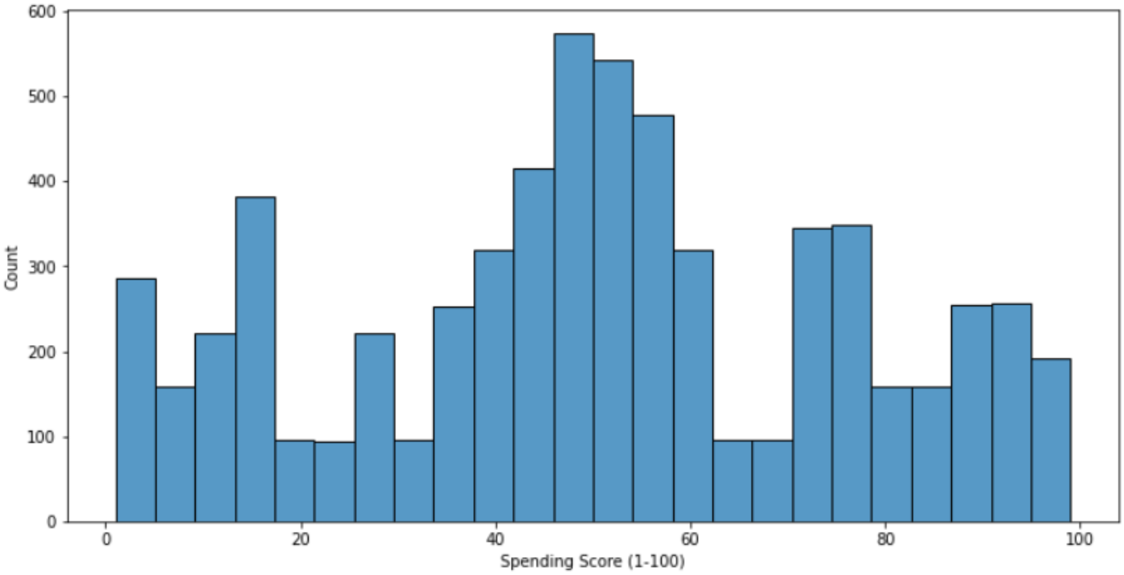
The most common ways in which within which businesses phase their client base are:

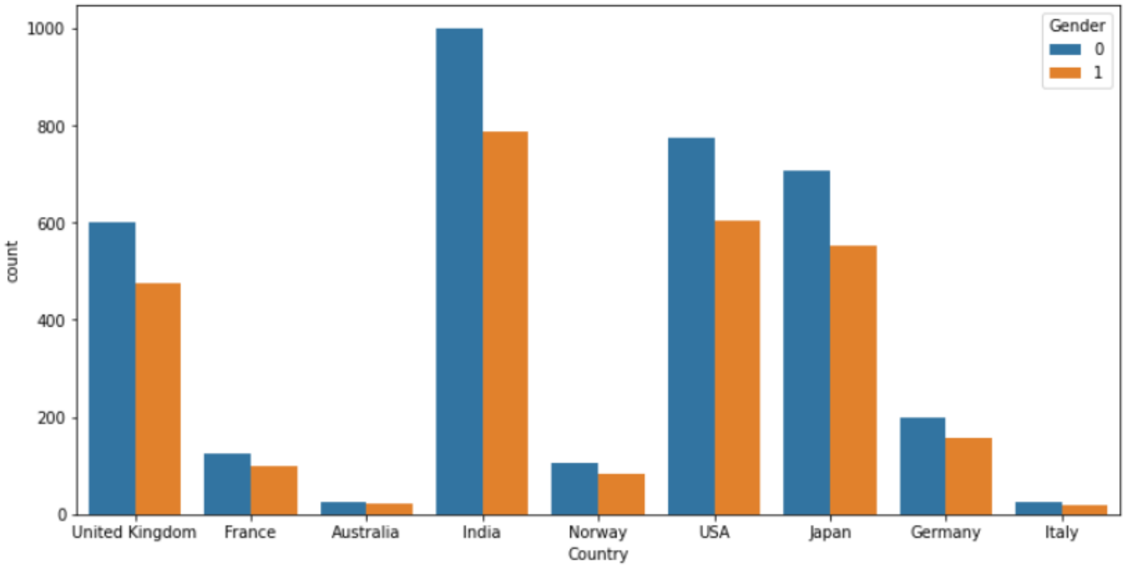
- **Demographic information**, like gender, age, familial and legal status, income, education, and occupation.
- **Geographical information** differs counting on the scope of the corporate. For localized businesses, this information would possibly pertain to specific cities or counties. For larger corporations, it'd mean a customer's town, state, or perhaps the country of residence.
- **Psychographics**, like socio-economic class, lifestyle, and temperament traits.
- **Behavioral data,** like outlay and consumption habits, product/service usage, and desired edges.

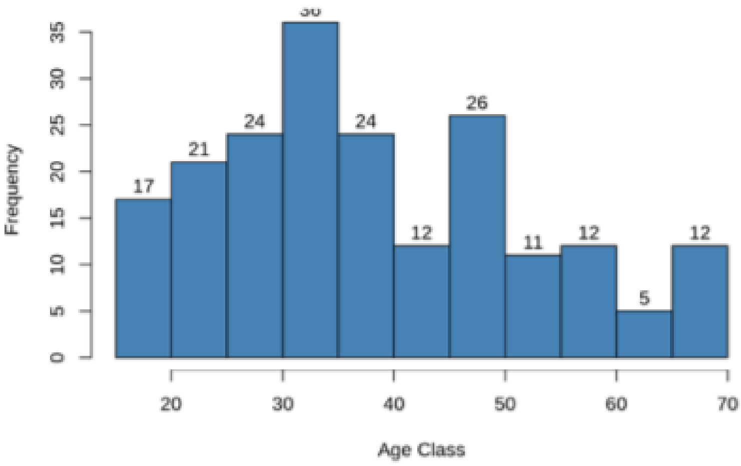


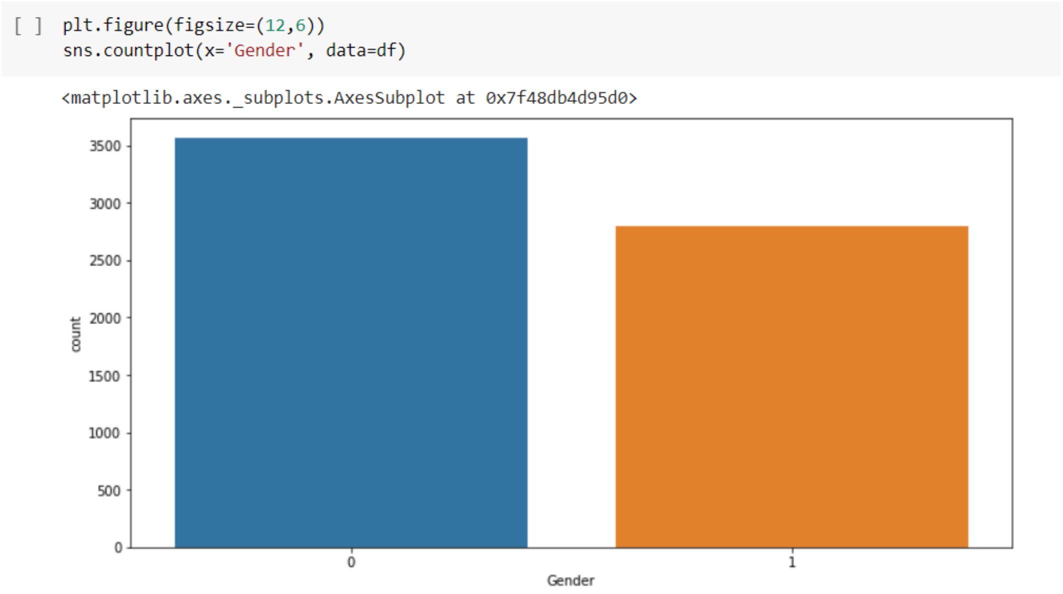
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f27740df950>
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f27740d14d0>
```


Count
600
500
400
300
200
100
0
0
20
40
60
80
100
Spending Score (1-100)

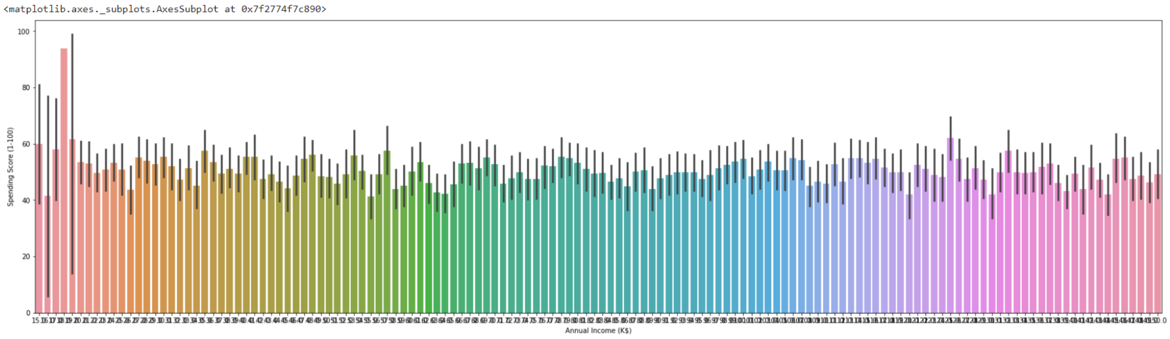

Gender
0
1
count
1000
800
600
400
200
0
United Kingdom
France
Australia
India
Norway
USA
Japan
Germany
Italy
Country


Frequency
0
5
10
15
20
25
30
35
17
21
24
36
24
12
26
11
12
5
12
20
30
40
50
60
70
Age Class



```
[ ]  plt.figure(figsize=(12,6))
     sns.countplot(x='Gender', data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f48db4d95d0>
```



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2774f7c890>
```

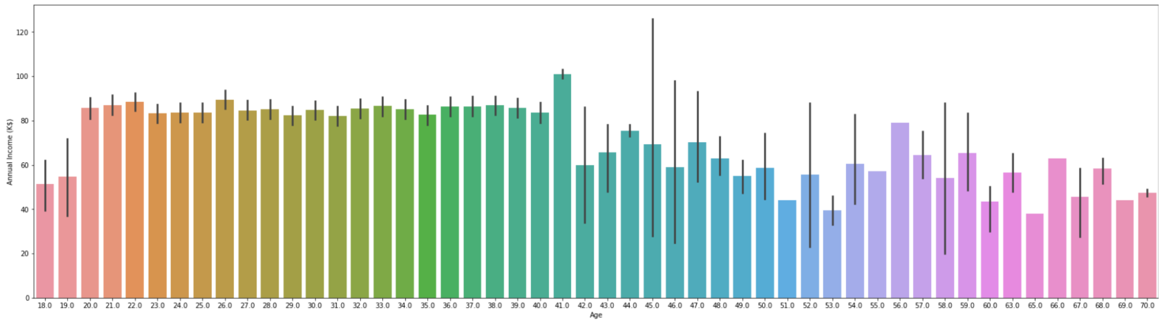
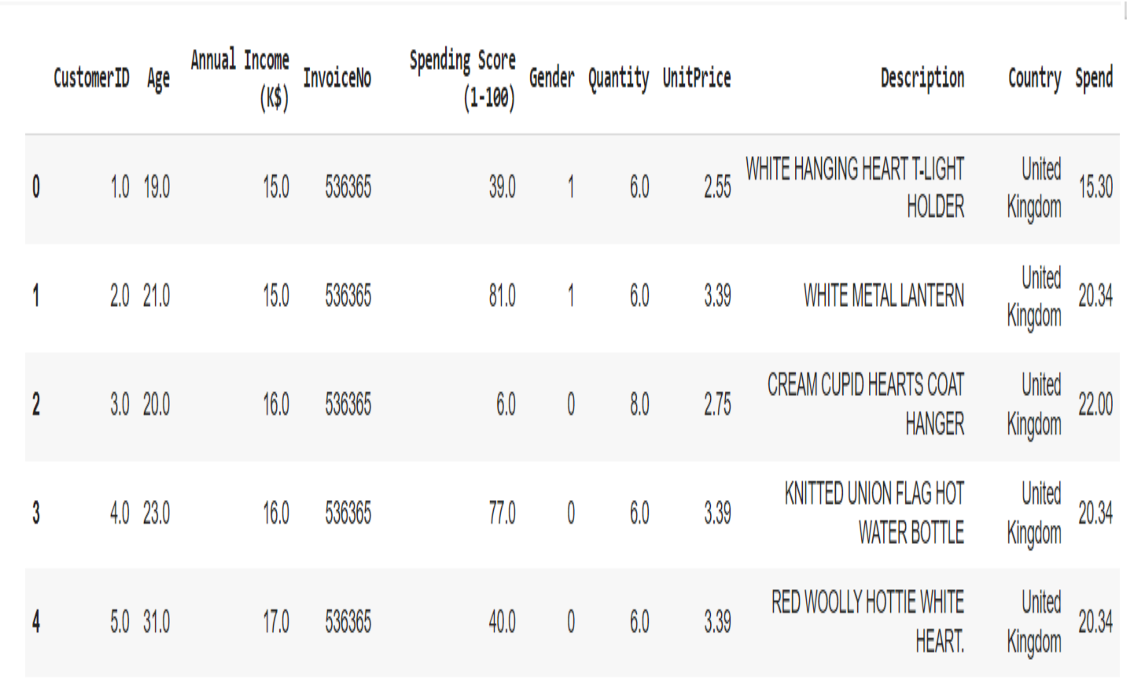


Fig 29- Comparisons



	CustomerID	Age	Annual Income (K$)	InvoiceNo	Spending Score (1-100)	Gender	Quantity	UnitPrice	Description	Country	Spend
0	1.0	19.0	15.0	536365	39.0	1	6.0	2.55	WHITE HANGING HEART T-LIGHT HOLDER	United Kingdom	15.30
1	2.0	21.0	15.0	536365	81.0	1	6.0	3.39	WHITE METAL LANTERN	United Kingdom	20.34
2	3.0	20.0	16.0	536365	6.0	0	8.0	2.75	CREAM CUPID HEARTS COAT HANGER	United Kingdom	22.00
3	4.0	23.0	16.0	536365	77.0	0	6.0	3.39	KNITTED UNION FLAG HOT WATER BOTTLE	United Kingdom	20.34
4	5.0	31.0	17.0	536365	40.0	0	6.0	3.39	RED WOOLLY HOTTIE WHITE HEART.	United Kingdom	20.34

Table 2

CHAPTER-5 CONCLUSIONS

5.1 Conclusions

- Cluster one denotes the clients with medium financial gain and medium payment scores this sort of customers may or may not be beneficial.
- Cluster two denotes the low annual financial gain however high yearly payments this sort of customers can be attracted by giving them minimal expense EMI's.
- Cluster three denotes a high annual financial gain moreover a high yearly payment so, we can focus on these sorts of clients as they bring in more cash and spend how much they need.
- Cluster four represents the cluster having high annual financial gain and low annual pay so, we can focus on these sorts of clients by asking for feedback and promoting the item in a superior manner.
- Cluster five represents customers with low annual financial gain and low annual payment, there is no need to target this kind of customers as they earn less and spend less
- From the customer map we can see that most of the customers are from India, USA, Japan, and United Kingdom.
- From Revenue share diagram we can see that the highest revenue contributing countries are India, USA, Japan, and United Kingdom.

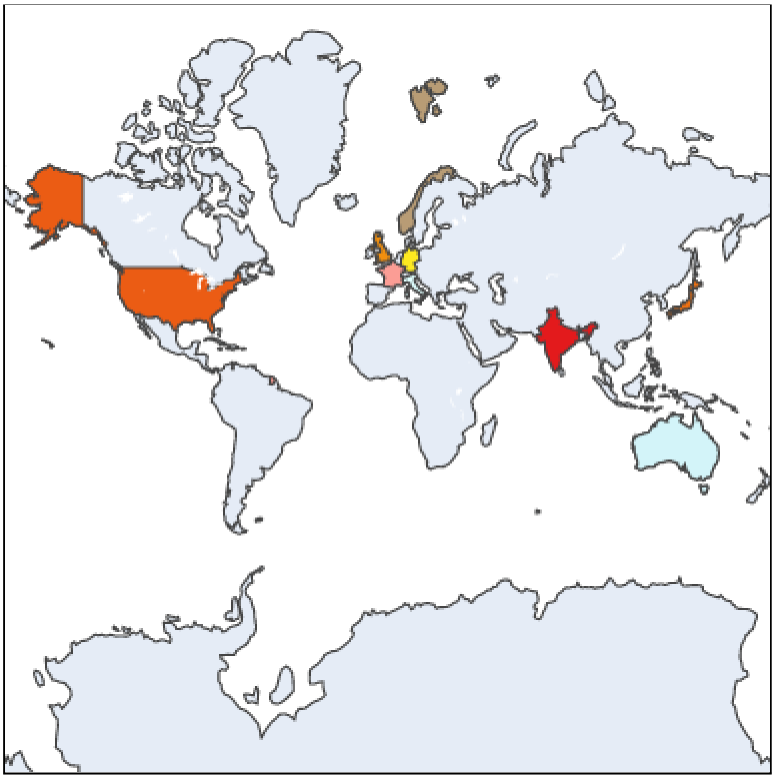


Fig 30- Customer Map

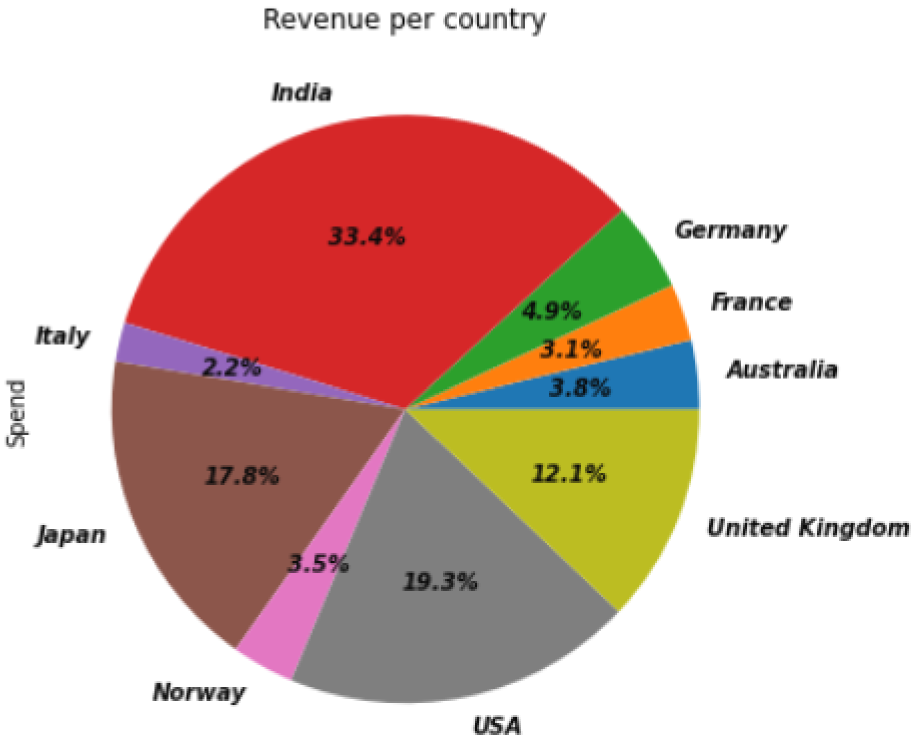


Fig 31- Revenue share

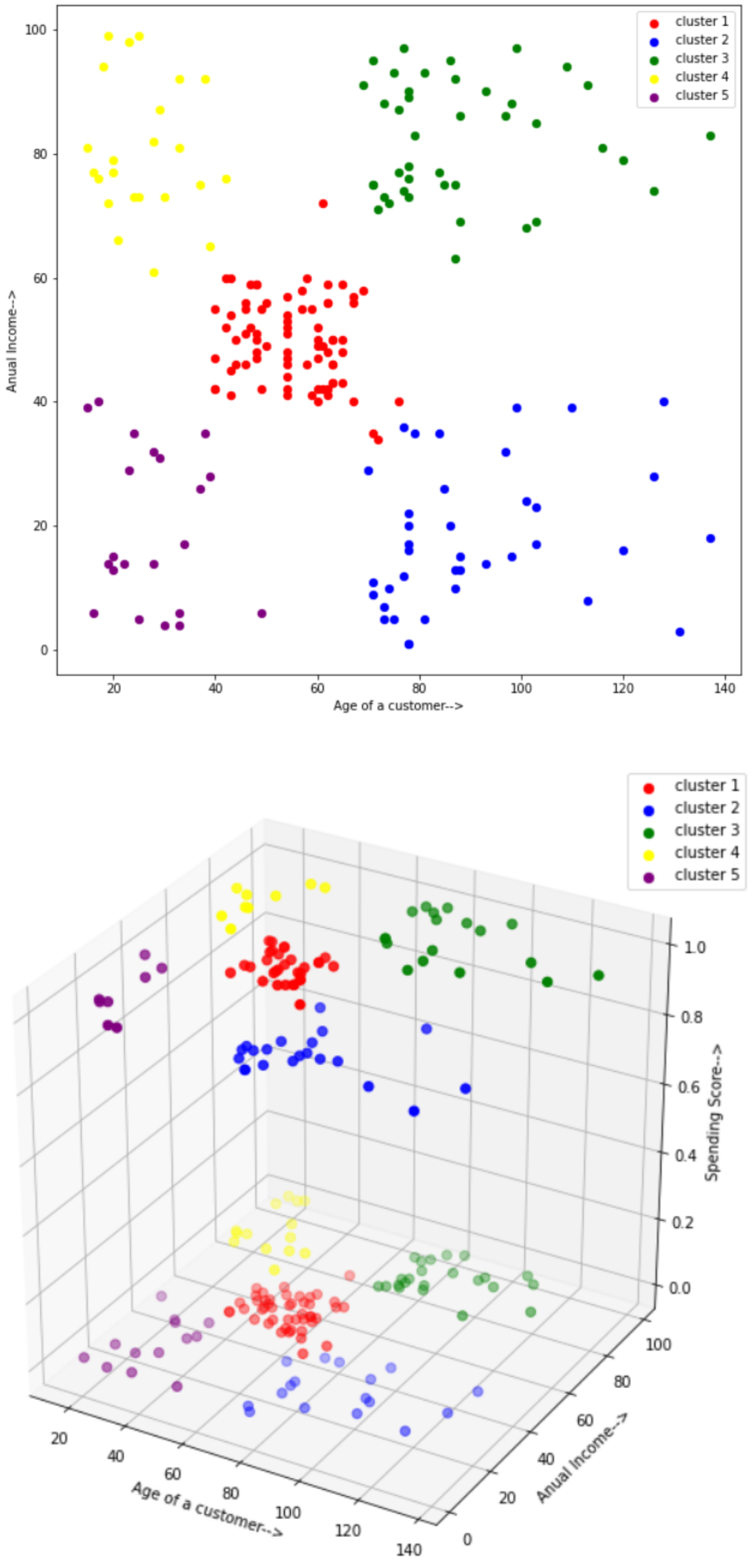


Fig 32- Final Clusters

REFERENCES

- Sari, J. N., Nugroho, L. E., Ferdiana, R., & Santosa, P. I. (2016). Review on Customer Segmentation Technique on Ecommerce. *Advanced Science Letters*, *22*(10), 3018–3022. https://doi.org/10.1166/asl.2016.7985

- Kushwaha, Y., & Prajapati, D. (n.d.). *Customer Segmentation using K-Means Algorithm*. Retrieved May 11, 2022, from https://ijcrt.org/papers/IJCRT_196650.pdf

- *Customer Segmentation Tutorial | Python Projects | K-Means Algorithm | Python Training | Edureka - YouTube*. (n.d.). Retrieved May 11, 2022, from https://www.youtube.com/watch?v=4jv1pUrG0Zk

- *Customer Segmentation. Segmentation by RFM clustering | by Barış Karaman | Towards Data Science*. (n.d.). Retrieved May 11, 2022, from https://towardsdatascience.com/data-driven-growth-with-python-part-2-customer-segmentation-5c019d150444

- Wu, J., Shi, L., Lin, W. P., Tsai, S. B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K -Means Algorithm. *Mathematical Problems in Engineering*, *2020*. https://doi.org/10.1155/2020/8884227