CUSTOMER SEGMENTATION

MINOR PROJECT REPORT

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR

THE AWARD OF THE DEGREE OF

BACHELOR OF TECHNOLOGY

Information Technology

Submitted By:

Anant Shree (1805490)

Arshpreet Singh (1805494)

Mehak (1805528)

Submitted To:

Prof. Ranjodh Kaur

Assistant Professor

Minor Project Coordinator

Department of Information Technology

Guru Nanak Dev Engineering College,

Ludhiana-141006

Abstract

Customer segmentation and pattern extraction is one of the key aspects of business decision support system. In order to grow the business intelligently in competitive market, identification of potential customer should be done timely. This paper proposes an integrated novel approach for determining target customers using predictive model and also discover their associative buying patterns using algorithm. After identification of targeted customers and their associative buying pattern, the business managers take the strategic profitable decisions accordingly.

ACKNOWLEDGEMENT

We are highly grateful to Dr. Sehijpal Singh, Principal, Guru Nanak Dev Engineering College (GNDEC), Ludhiana, for providing this opportunity to carry out the minor project work at making 'Customer Segmentation Using Machine LEARNING' using unity. The constant guidance and encouragement received from Dr. K.S. Mann, H.O.D., IT Department, GNDEC Ludhiana has been of great help in carrying out the project work and is acknowledged with reverential thanks. We would like to express a deep sense of gratitude and thanks profusely to DR. Kamaljit Kaur Dhillon , without his wise counsel and able guidance, it would have been impossible to complete the project in this manner. We express gratitude to other faculty member of computer science and engineering department of GNDEC for their intellectual support throughout the course of this work Finally, we are indebted to all whosoever have contributed in this report work.

Anant Shree (URN – 1805490)

Arshpreet Singh (URN – 1805494)

Mehak (URN – 11805528)

List of Figures

Title page

Contents

1 Introduction

1.1 Introduction to Project Page

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.

Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers (and potential customers), a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioral and other indicators.

It is always easier to make assumptions and use "gut feelings" to define rules which will segment customers into logical groupings, e.g., customers who came from a particular source, who live in a particular location or who bought a particular product/service. However, these high-level categorizations will seldom lead to the desired results.

It is obvious that some customers will spend more than others during their relationship with a company. The best customers will spend a lot for many years. Good customers will spend modestly over a long period of time, or will spend a lot over a short period of time. Others won't spend too much and/or won't stick around too long.

The right approach to segmentation analysis is to segment customers into groups based on predictions regarding their total future value to the company, with the goal of addressing each group (or individual) in the way most likely to maximize that future, or lifetime, value.

1.2 Project Category (Internet based, Application or System Development, Research based, Industry Automation, Network or System Administration)

It is a System Based and Industry Based project as due to utilization of customer segmentation in industry it requires a lot of clustering and segmentation on the basis of several features for ex- income, salary, type of employee etc. so that they can increase their productivity and since it is basically a proposed system and hence developed for future use for any industry so it is a system-based project which could be utilized anywhere and everywhere.

1.3 Objectives

1. **Determine the Pricing -** A key objective for market segmentation is determining what price different groups of consumers are willing to pay for your product.

2. **Improvisation of product -** Market segmentation in terms of promotion lets you target members of each group in terms of what is important to them.

3. **Offering a product/service for maximum convenience -** Market segmentation lets you decide, so you can tailor your sales channels to the preferences of the members of each market segment .

4. **Giving promotional initiatives -** Your product could be improved in various ways, but you don't want to spend money on extra features if they don't result in additional sales.

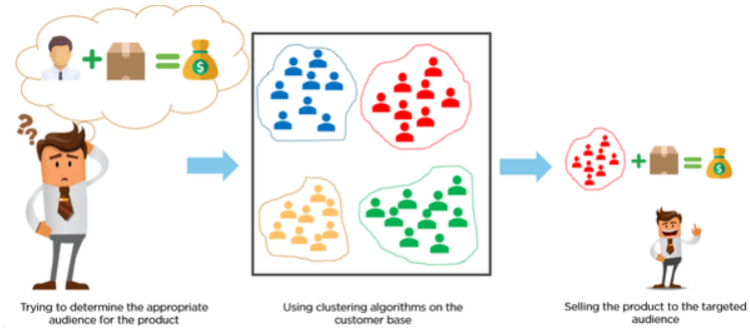


Figure 1: Implementation of Customer Segmentation

1.4 Problem Formulation

Algorithm used in this project is K-Means

K-MEANS ALGORITHM -

1. Specify number of clusters K.

2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.

3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

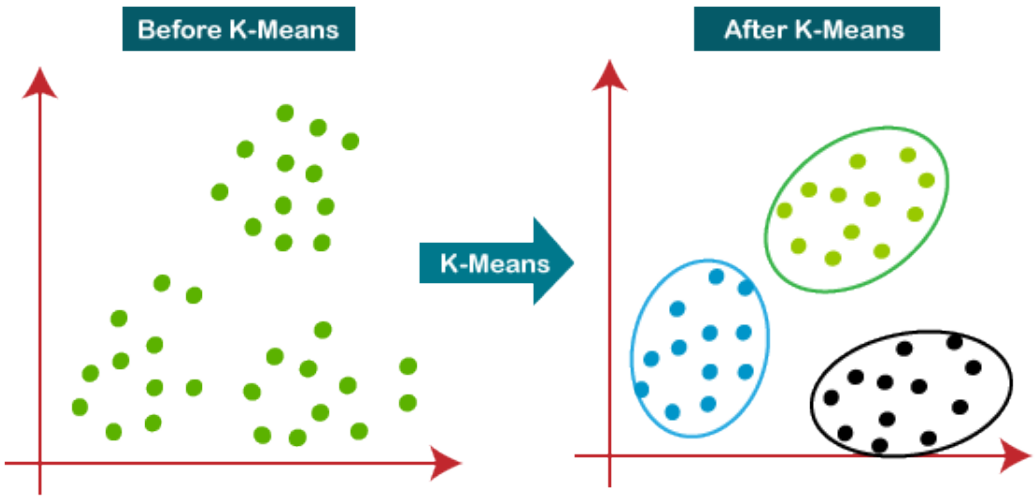


Figure 2: K-Means Algorithm

Why K-Means algorithm used for this project ?

1. K means clustering is one of the most popular clustering algorithms and used for partitioning and clusterings in most of the projects developed for segmentation
2. The first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset.
3. The goal of K means is to group data points into distinct non-overlapping subgroups.
4. One of the major application of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

Concept Used -

Customer Segmentation with Machine Learning

An additional approach to customer segmentation is leveraging machine learning algorithms to discover new segments. Different to marketer-designed segmentation models, as the ones described above, machine learning customer segmentation allows advanced algorithms to surface insights and groupings that marketers might find difficulty discovering on their own. Furthermore, marketers that create a feedback loop between the segmentation model and campaign results will have ever improving customer segments. In these cases, the machine learning model will be not only able to refine its definition of segments, but also be able to identify if a specific subset of the segment is outperforming the rest, optimizing marketing performance.

1.5 Identification/Reorganization of Need

Segmentation allows businesses to make better use of their marketing budgets, gain a competitive edge over rival companies and, importantly, demonstrate a better knowledge of your customers' needs and wants. It can also help :

- **Marketing efficiency** – Breaking down a large customer base into more manageable pieces, making it easier to identify your target audience and launch campaigns to the most relevant people, using the most relevant channel.
- **Determine new market opportunities** – During the process of grouping your customers into clusters, you may find that you have identified a new market segment, which could in turn alter your marketing focus and strategy to fit.
- **Better brand strategy** – Once you have identified the key motivators for your customer, such as design or price or practical needs, you can brand your products appropriately.
- **Improve distribution strategies** – Identifying where customers shop and when can informatively shape product distributions strategies, such as what type of products are sold at particular outlets.
- **Customer retention** – Using segmentation, marketers can identify groups that require extra attention and those that churn quick, along with customers with the highest potential value. It can also help with creating targeted strategies that capture your customers' attention and create positive, high-value experiences with your brands.

1.6 Existing System

Customer segmentation is a existing project in the business world where it needs the segmentation of customer as the maximum productivity must be increased by the segmentation so that organization must know how and where to invest the products so that it must reached to the potential customers.

Current Customer Segmentation is done right, however, the business benefits are numerous. For example, a best current customer segmentation exercise can tangibly impact your operating results by:

1. Improving your whole product
2. Focusing your marketing message
3. Allowing your sales organization to pursue higher percentage opportunities
4. Getting higher quality revenues

1.7 Proposed System

It is a proposed system developed for industry purposed to solve the problem of higher productivity. Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers (and potential customers), a customer segmentation analysis allows marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioral and other indicators.

1.8 Unique Features of the System

1. **Identifiable -** You should be able to identify customers in each segment and measure their characteristics, like demographics or usage behavior.
2. **Substantial -** It's usually not cost-effective to target small segments — a segment, therefore, must be large enough to be potentially profitable.
3. **Accessible -** It sounds obvious, but your company should be able to reach its segments via communication and distribution channels. When it comes to young people, for example, your company should have access to Twitter and Tumblr and know how to use them authentically — or, as Clearblue smartly did, reach out to celebrities with active Twitter presences to do some of your marketing for you. .
4. **Stable -** In order for a marketing effort to be successful, a segment should be stable enough for a long enough period of time to be marketed to strategically. For example, lifestyle is often used as a way to segment.
5. **Differentiable -** The people (or organizations, in B2B marketing) in a segment should have similar needs that are clearly different from the needs of other people in other segments.

2 Requirement Analysis and System Specification

2.1 Feasibility Study (Technical, Economical, Operational)

The project needs to implement only the analysis of how the behaviour of customer is done on the basis of price xed thus , this project implements the feasibility by determining the price by customers and then clustering to other personal to see the stock market profit.

So , Further feasibility is studied by the following factors :

1. It will help in identifying the most potential customers.
2. It will help managers to easily communicate with a targeted group of the audience.

3. Also, help in selecting the best medium for communicating with the targeted segment.
4. It improves the quality of service, loyalty, and retention.
5. Improve customer relationship via better understanding needs of segments.
6. It provides opportunities for upselling and cross-selling.
7. It will help managers to design special offers for targeted customers, to encourage them to buy more products. It helps companies to stay a step ahead of competitors.
8. It also helps in identifying new products that customers could be interested in.

2.2 Software Requirement Specification Document which must include the following: (Data Requirement, Functional Requirement, Performance Requirement, Dependability Requirement, Maintainability requirement, Security Requirement, Look and feel requirement)

SRS Documentation for it includes –

2.2.1 Introduction

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.

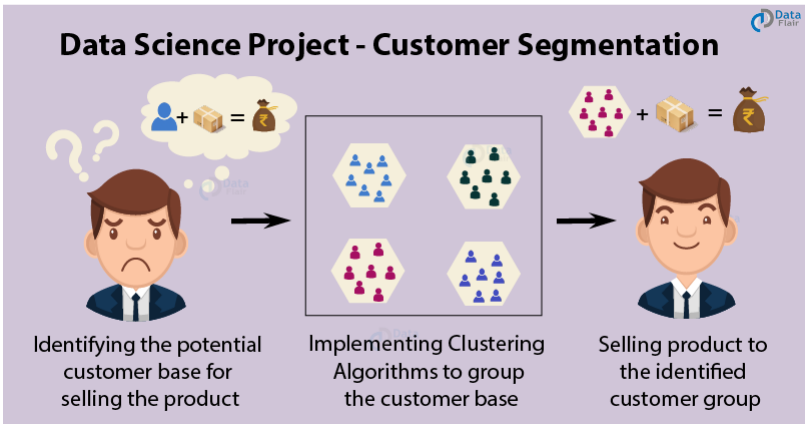


Figure 3: Overview of Customer Segmentation

2.2.2 User Needs

- Marketing efficiency
- Better brand strategy
- Customer retention

2.2.3 Intended Audience

1. Any small or large private business organization which requires the status of segmentation on the basis of type of customers. (For example - investing in different areas of stock market).
2. Any government organization which requires the segmentation status (For example - Covid vaccination status).

2.2.4 Intended Use

With a range of competitors available in the market, customer segmentation gives you a perfect opportunity to stand out from the crowd. You can easily find the right messaging that will make your clients knock at your door time and time again.

2.2.5 Scope of development

This project is developed Since the marketer's goal is usually to maximize the value (revenue and/or profit) from each customer, it is critical to know in advance how any particular marketing action will influence the customer. Ideally, such "action-centric" customer segmentation will not focus on the short-term value of a marketing action, but rather the long- term customer lifetime value (CLV) impact that such a marketing action will have. Thus, it is necessary to group, or segment, customers according to their CLV.

2.2.6 Overall Description

Customer segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately.

In business-to-business marketing, a company might segment customers according to a wide range of factors, including :

- Industry
- Number of employees
- Products previously purchased from the company
- Location

In business-to-consumer marketing, companies often segment customers according to demographics that include:

- Age
- Gender
- Marital status
- Location (urban, suburban, rural)
- Life stage (single, married, divorced, empty-nester, retired, etc.)

2.2.7 Purpose

Market segmentation and targeting refer to the process of identifying a company's potential customers, choosing the customers to pursue, and creating value for the targeted customers. It is achieved through the segmentation, targeting, and positioning (STP) process.

2.2.8 System Features and Requirements

1. Functional Requirements

1. Manage and store all its data in a single solution (customer and historical knowledge)
2. Track dashboards to facilitate reporting and decision making.
3. Manage billing and refillable hours to third parties (external services).

2. NON – Functional Requirements

1. Process requirement to validate customer details.
2. Process requirement to retrieve customer details.
3. Maintain compliance with financial regulations

2.3 Expected hurdles

1. Segmentation increases costs.
2. Characteristics of a customer segment change, investment made already might become useless.

2.4 SDLC Model to be used

SDLC model used is **Spiral model**.

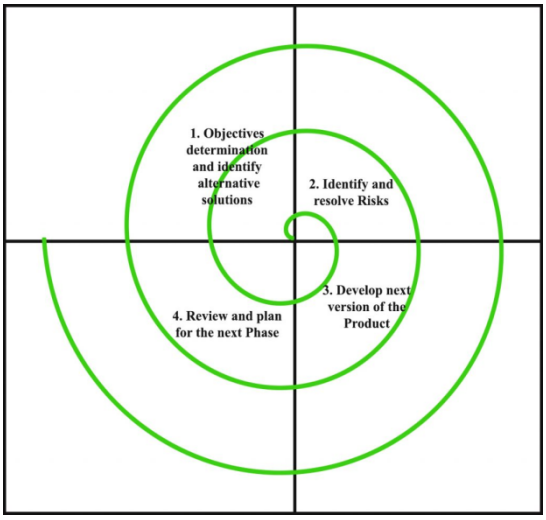


Figure 4: Spiral Model

The **spiral model** is a risk-driven software development process **model**. Based on the unique risk patterns of a given project, the spiral model guides a team to adopt elements of one or more process models, such as incremental, waterfall, or evolutionary prototyping.

3 System Design

3.1 Detail Design

Activity Diagram is shown below for detailed design concept -

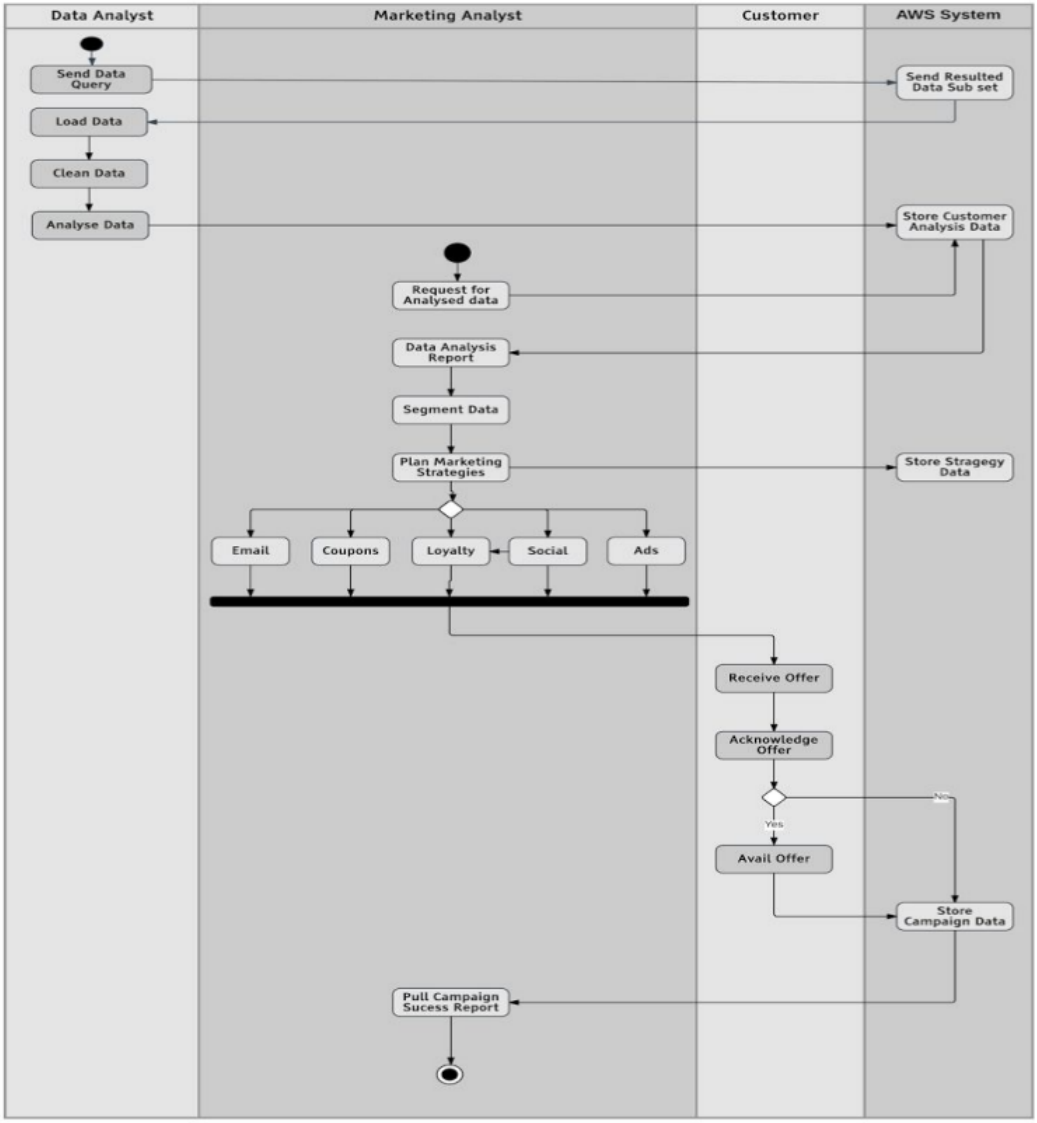


Figure 5: Activity Diagram

3.2 System Design using various structured analysis and design tools such as: DFD's, Data Dictionary, Structured charts, Flowcharts or UML

1. Data Flow Diagram -

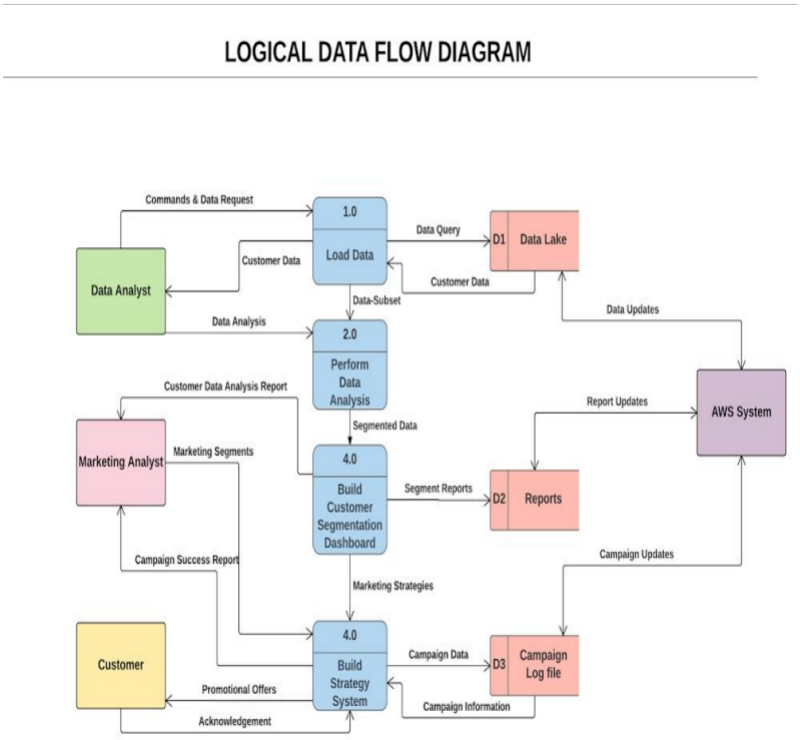


Figure 6: Data Flow Diagram

2. UML Diagram -

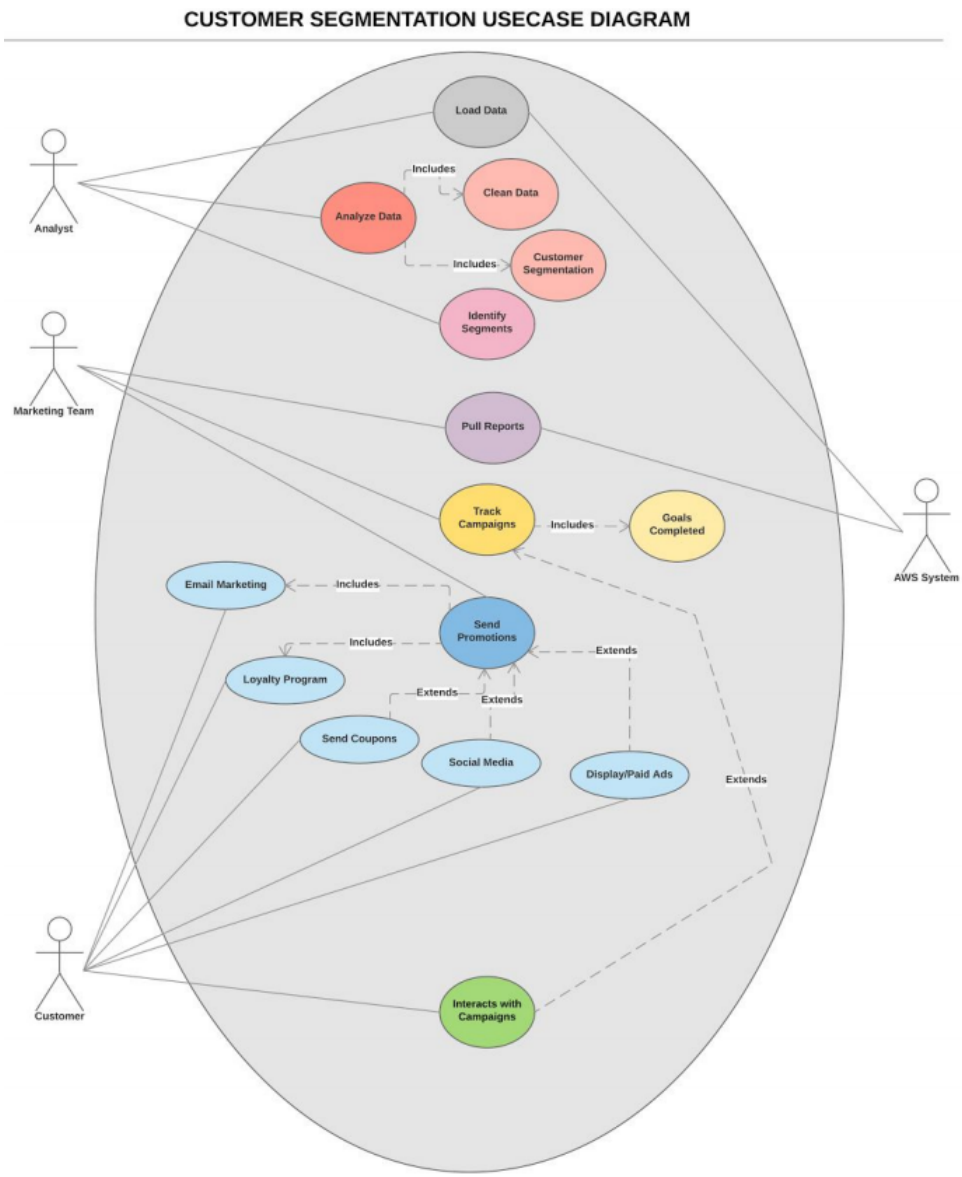


Figure 7: UML Diagram

3. Flow Chart -

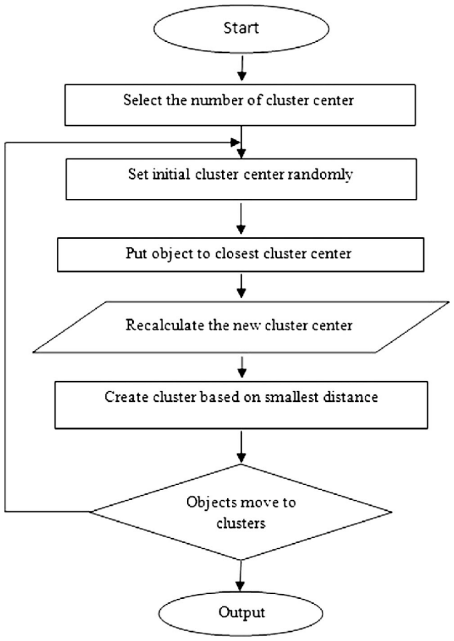


Figure 8: Flow Chart

3.3 Database Design

3.3.1 ER Diagrams

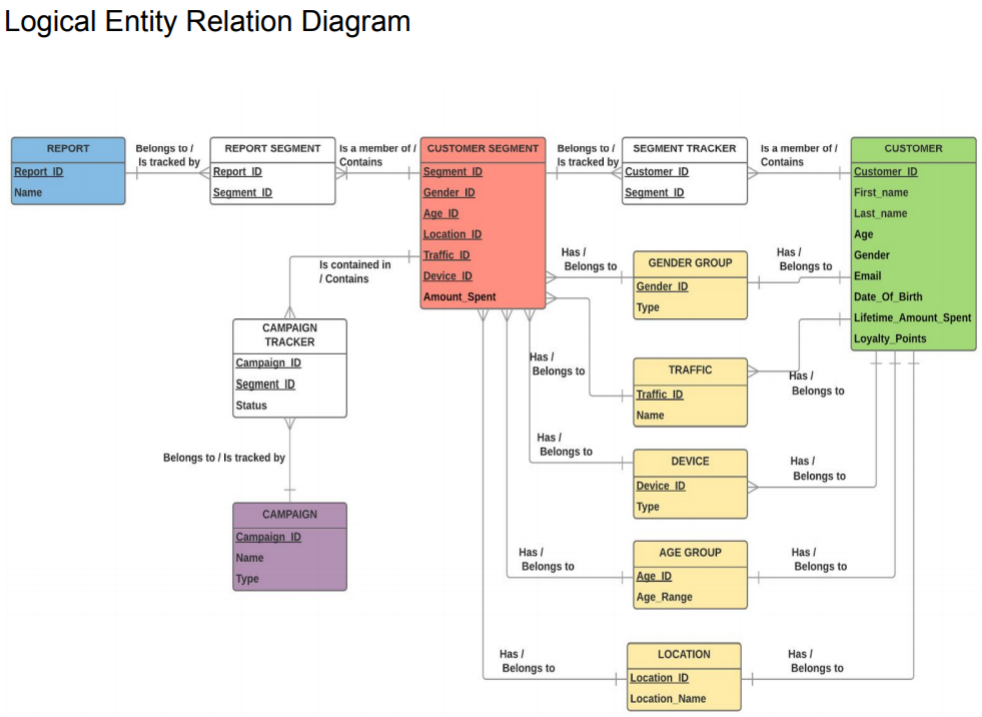
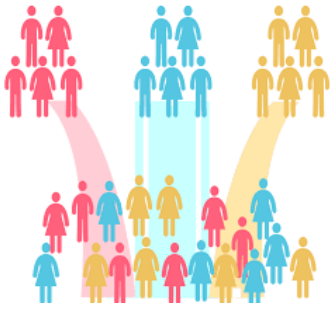


Figure 9: ER Diagram

3.3.2 Database Sets



Since , at first we had mixed group of customers in datasets but after applying algorithm to datasets we got the classified view and hence potential customers are tracked according to situation and criteria of segmentation .

	Genre	Age	Annual Inc	Spending Score (1-100)
1	Genre	Age	Annual Inc	Spending Score (1-100)
2	Male	19	15	39
3	Male	21	15	81
4	Female	20	16	6
5	Female	23	16	77
6	Female	31	17	40
7	Female	22	17	76
8	Female	35	18	6
9	Female	23	18	94
10	Male	64	19	3
11	Female	30	19	72
12	Male	67	19	14
13	Female	35	19	99
14	Female	58	20	15
15	Female	24	20	77
16	Male	37	20	13
17	Male	22	20	79
18	Female	35	21	35
19	Male	20	21	66
20	Male	52	23	29
21	Female	35	23	98
22	Male	35	24	35
23	Male	25	24	73
24	Female	46	25	5
25	Male	31	25	73

Figure 10: Dataset Used in Customer Segmentation

3.3.3 Database Connection Controls and Strings

3.4 Methodology

Identifying right customer and providing right service at right time and treating different types of customers differently is the key to success in business.

In the first step of this data science project, we will perform data exploration. We will import the essential packages required for this role and then read our data. Finally, we will go through the input data to gain necessary insights about it.

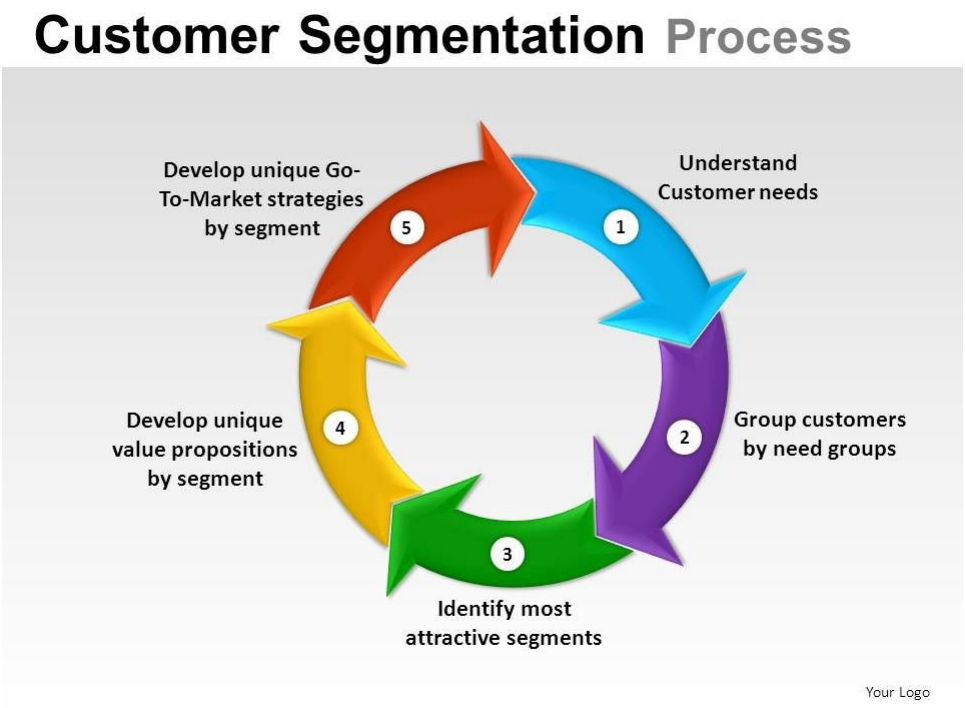


Figure 11: Process of Customer Segmentation

Understand Customer Needs

By understanding the needs of users we can observe what actually consumer needs for the business strategy.

- **Group customers by need** - Separate one type of customers at one place so that product or services could easily reach to them without specially classifying .
- **Identify Most Attractive Segments** - Identify the most attractive potential customers so that maximum profit could be gained by delivering services there .
- **Develop Unique Value Propositions By Segment** - Unique value propositions must be set so that every customers must be able to satisfy according to their value of needs of organization.
- **Develop Unique Go-to-Market strategies** - If we develop unique market go-to strategy then it will be easy for us to observe which type of customers were reached and who are left also which type of services and where it will be needed.

4 Implementation, Testing, and Maintenance

4.1 Introduction to Languages, IDE's, Tools and Technologies used for Implementation

LANGUAGE USED - The entirety of the code written for this project was in ***Python***.

Being a full-fledged programming language, Python is a great tool to implement algorithms for production use. There are several Python packages for basic data analysis and machine learning.

IDE USED - Jupyter notebooks basically provides an interactive computational environment for developing Python based Data Science applications. They are formerly known as ipython notebooks. The following are some of the features of Jupyter notebooks that makes it one of the best components of Python ML ecosystem :

- Jupyter notebooks can illustrate the analysis process step by step by arranging the stuff like code, images, text, output etc. in a step by step manner.
- It helps a data scientist to document the thought process while developing the analysis process.
- One can also capture the result as the part of the notebook.
- With the help of jupyter notebooks, we can share our work with a peer also.

ENVIRONMENTS & TOOLS USED

The following Python packages played pivotal roles in the execution and development of this project:

1. **scikit-learn** - Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. The library is built upon the SciPy (Scientific Python). The library is focused on modeling data. Clustering is one of the model provided by scikit-learn and here in our project we have done it using K-Means Clustering.
2. **Seaborn:** - Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
3. **Numpy** - NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.
4. **Pandas** - Pandas is a Python library used for working with data sets.It has functions for analyzing, cleaning, exploring, and manipulating data. Pandas allows us to analyze big data and make conclusions based on statistical theories.

5. **Matplotlib -** Matplotlib is one of the most popular Python packages used for data visualization. It is a cross-platform library for making 2D plots from data in arrays. It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits.

4.2 Coding standards of Language used

For Python, PEP 8 has emerged as the style guide that most projects adhere to; it promotes a very readable and eye-pleasing coding style. Major coding standards include:

1. Use 4-space indentation and no tabs.
2. Use docstrings.
3. Wrap lines so that they don't exceed 79 characters.
4. Use of regular and updated comments are valuable to both the coders and users.
5. Use Python's default UTF-8 or ASCII encodings and not any fancy encodings.
6. Use spaces around operators and after commas, but not directly inside bracketing constructs.
7. While naming of function of methods always use self for the first argument.
8. Characters such as 'l' (lowercase letter el), 'O' (uppercase letter oh), or 'I' (uppercase letter eye) should not be used for identifiers.
9. Name your classes and functions consistently.
10. Follow proper naming conventions throughout the program

4.3 Testing Techniques and Test Plans

4.3.1 Purpose

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a project. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software wit the intent of ensuring that the software system meets its requirements and user expectations and does not fail in an unacceptable manner.

4.3.2 Testing Techniques

A typical software testing suite will include -

1. **unit tests** which operate on atomic pieces of the codebase and can be run quickly during development

2. **regression tests** replicate bugs that we've previously encountered and fixed

3. **integration tests** which are typically longer-running tests that observe higher-level behaviors that leverage multiple components in the codebase, and follow conventions such as:

- don't merge code unless all tests are passing
- always write tests for newly introduced logic when contributing code
- when contributing a bug fix, be sure to write a test to capture the bug and prevent future regressions.

4.3.3 Test Plans

ML testing has a couple of peculiarities: it demands that you test the quality of data, not just the model, and go through a couple of iterations adjusting the hyperparameters to get the best results. So, this testing basically consistes of two parts:

1. Data Debugging - First of all, we started with data debugging because the accuracy of predictions made by the model depends not only on the algorithm but on the quality of data itself.

The engineered data was checked separately. While raw data was alright, engineered data went through some changes and looked totally different. We wrote tests to ensure that the outliers were handled or that the missing values were replaced by mean or default values.

2. Model Debugging - We controlled the model performance by manual testing for a random couple of data points.

We checked the general logic of the model. After we got reasonable results, we jumped to unit tests to check the model performance on the real data.

4.3.4 Test Results

All the test cases were passed successfully on the real data. No defects were encountered.

5 Results and Discussions

5.1 Snapshots of system with brief detail of each

 # We started with loading all the libraries and dependencies. The columns in the dataset are customer id, gender, age, income and spending score.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df=pd.read_csv("D:\Minor Project\Mall_Customers.csv")
df.head()
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

 # We described the whole dataset and calculated the values such as: No. of rows and columns, count, mean, min, max, etc.

```
df.shape
```

```
(200, 5)
```

```
df.describe()
```

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

 # We checked the data types of various attributes and checked if no column has null value.

```
df.dtypes
```

```
CustomerID                int64
Gender                   object
Age                       int64
Annual Income (k$)        int64
Spending Score (1-100)    int64
dtype: object
```

```
df.isnull().sum()
```

```
CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

 # We dropped the "CustomerID" column as that does not seem relevant to the context.

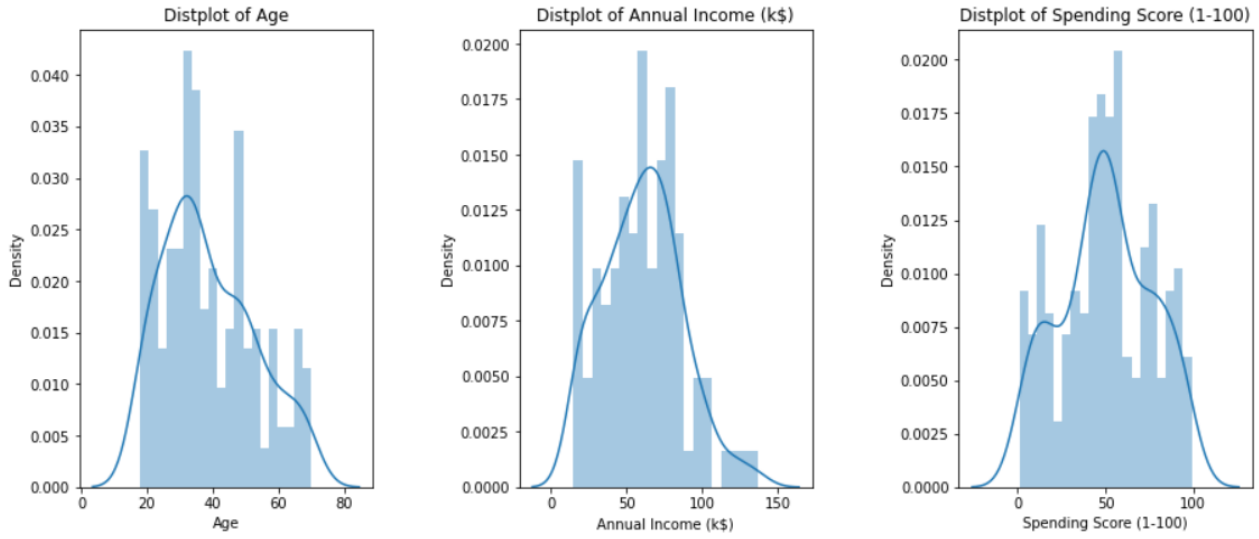
```
df.drop(['CustomerID'],axis=1, inplace=True)
```

```
df.head()
```

	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	Male	19	15	39
1	Male	21	15	81
2	Female	20	16	6
3	Female	23	16	77
4	Female	31	17	40

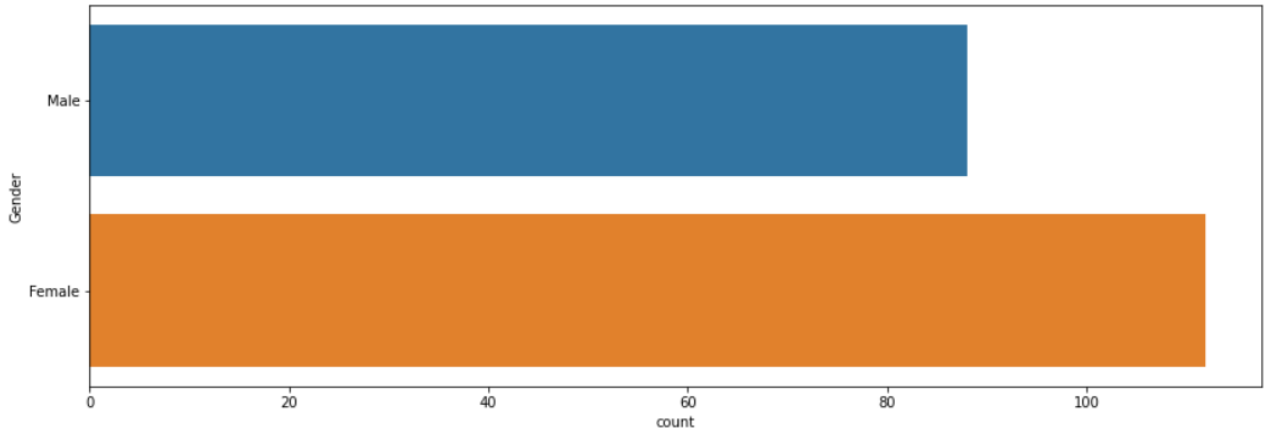
 # Also we plotted the age frequency of customers.

```
plt.figure(1, figsize=(15,6))
n=0
for x in ['Age' , 'Annual Income (k$)' , 'Spending Score (1-100)']:
    n += 1
    plt.subplot(1 , 3 , n)
    plt.subplots_adjust(hspace = 0.5 , wspace = 0.5)
    sns.distplot(df[x] , bins = 20)
    plt.title('Distplot of {}'.format(x))
plt.show()
```

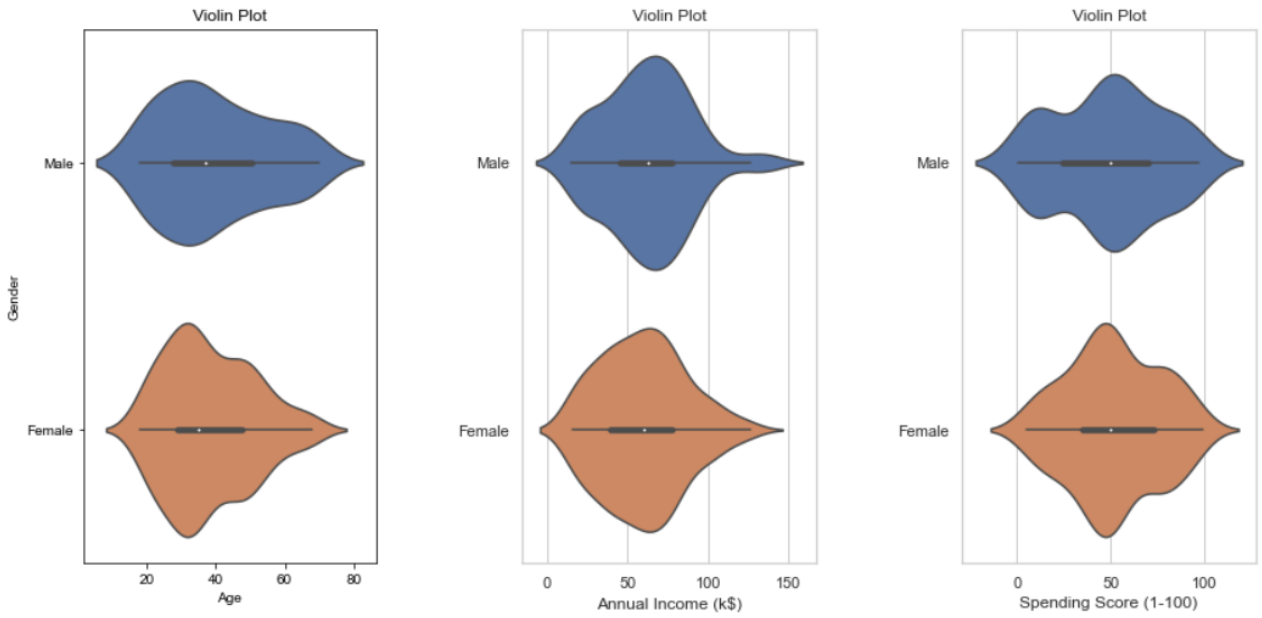


 # We made a bar plot to check the distribution of male and female population in the dataset. The female population clearly outweighs the male counterpart.

```
plt.figure(figsize =(15,5))
sns.countplot(y='Gender' ,data=df)
plt.show()
```

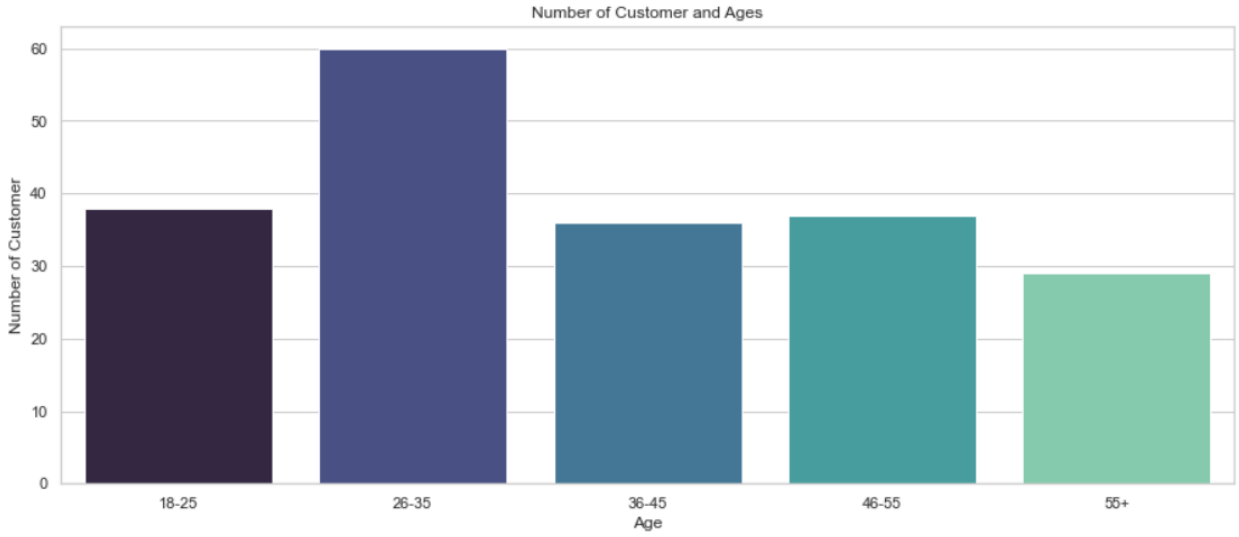



```
plt.figure(1,figsize=(15,7))
n=0
for cols in ['Age' , 'Annual Income (k$)' , 'Spending Score (1-100)']:
    n += 1
    plt.subplot(1 , 3 , n)
    sns.set(style="whitegrid")
    plt.subplots_adjust(hspace = 0.5 , wspace = 0.5)
    sns.violinplot(x = cols , y = 'Gender' , data = df)
    plt.ylabel('Gender' if n == 1 else '')
    plt.title('Violin Plot')
plt.show()
```



 # Next we made a bar plot to check the distribution of number of customers in each age group. Clearly the 26–35 age group outweighs every other age group.

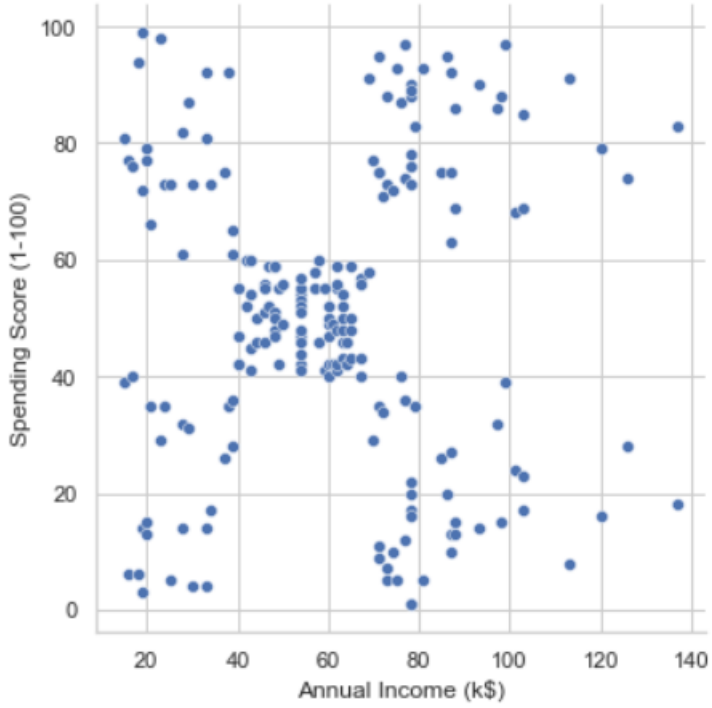
```
age_18_25 = df.Age[(df.Age >= 18) & (df.Age <= 25)]
age_26_35 = df.Age[(df.Age >= 26) & (df.Age <= 35)]
age_36_45 = df.Age[(df.Age >= 36) & (df.Age <= 45)]
age_46_55 = df.Age[(df.Age >= 46) & (df.Age <= 55)]
age_55above = df.Age[df.Age >= 56]

agex = ["18-25","26-35","36-45","46-55","55+"]
agey = [len(age_18_25.values),len(age_26_35.values),len(age_36_45.values),len(age_46_55.values),len(age_55above.values)]

plt.figure(figsize=(15,6))
sns.barplot(x=agex, y=agey, palette="mako")
plt.title("Number of Customer and Ages")
plt.xlabel("Age")
plt.ylabel("Number of Customer")
plt.show()
```

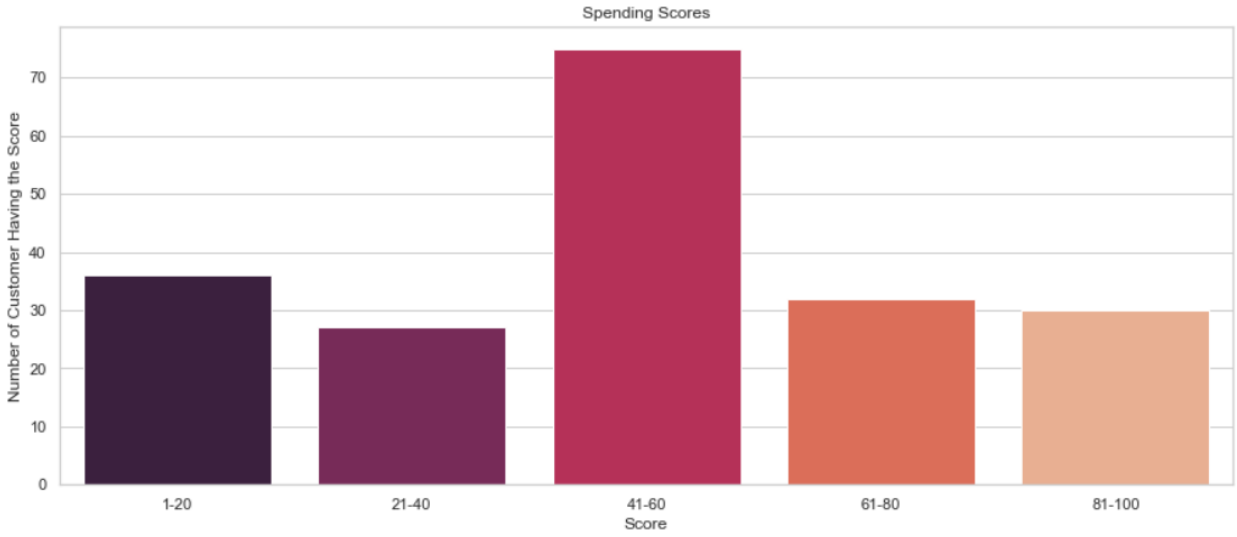


 # We continued with making a bar plot to visualize the number of customers according to their spending scores. The majority of the customers have spending score in the range 41–60.

 # Along with this, we made a scatter plot between Spending Score and Annual Income.

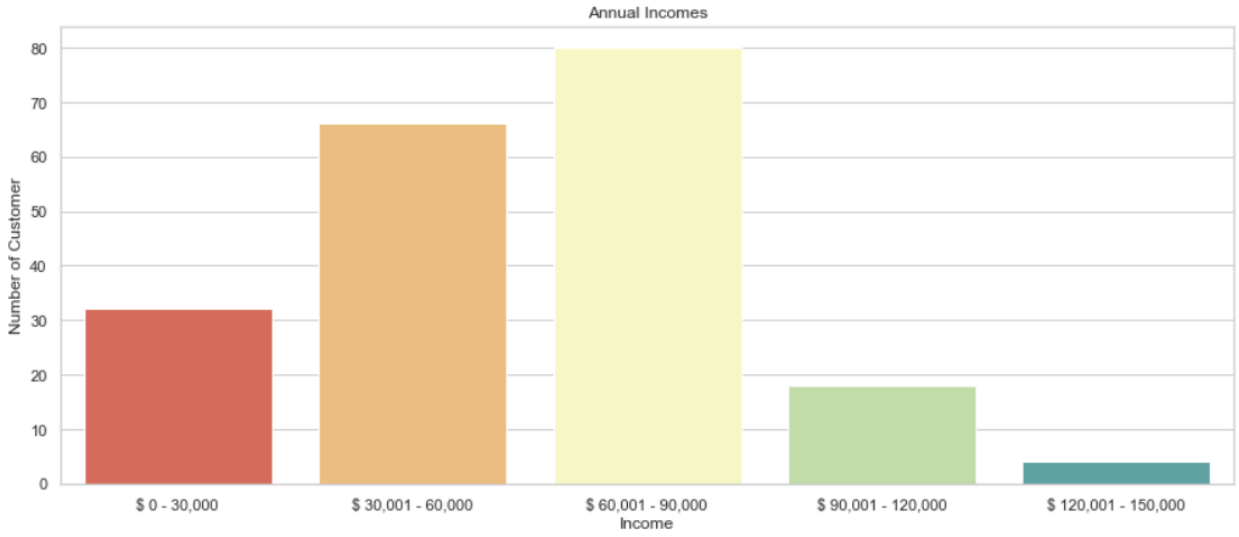
```
sns.relplot(x="Annual Income (k$)", y="Spending Score (1-100)", data=df)

ss_1_20 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 1) & (df["Spending Score (1-100)"] <= 20)]
ss_21_40 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 21) & (df["Spending Score (1-100)"] <= 40)]
ss_41_60 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 41) & (df["Spending Score (1-100)"] <= 60)]
ss_61_80 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 61) & (df["Spending Score (1-100)"] <= 80)]
ss_81_100 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 81) & (df["Spending Score (1-100)"] <= 100)]

ssx = ["1-20", "21-40", "41-60", "61-80", "81-100"]
ssy = [len(ss_1_20.values), len(ss_21_40.values), len(ss_41_60.values), len(ss_61_80.values), len(ss_81_100.values)]

plt.figure(figsize=(15,6))
sns.barplot(x=ssx, y=ssy, palette="rocket")
plt.title("Spending Scores")
plt.xlabel("Score")
plt.ylabel("Number of Customer Having the Score")
plt.show()
```





 # Also we made a bar plot to visualize the number of customers according to their annual income. The majority of the customers have annual income in the range 60000 and 90000.

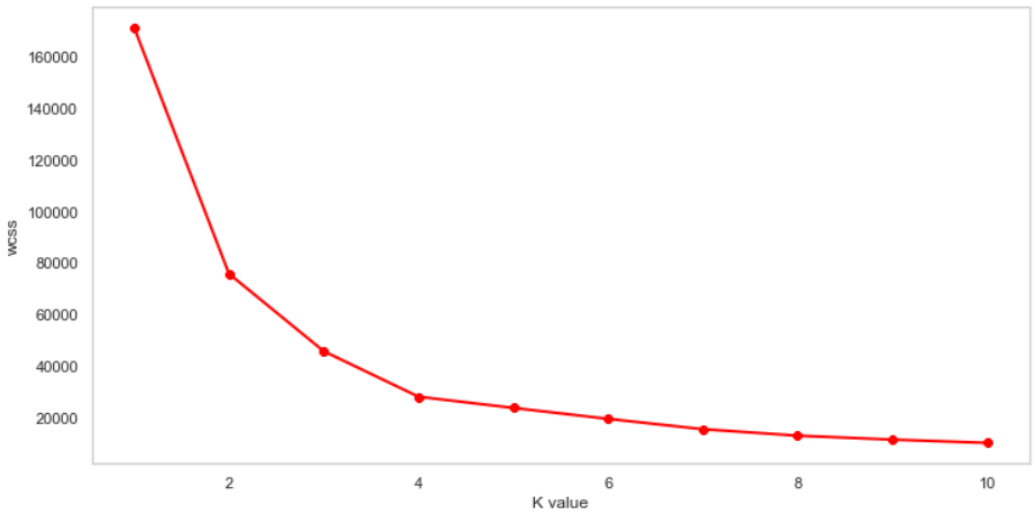
```
ai0_30 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 0) & (df["Annual Income (k$)"] <= 30)]
ai31_60 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 31) & (df["Annual Income (k$)"] <= 60)]
ai61_90 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 61) & (df["Annual Income (k$)"] <= 90)]
ai91_120 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 91) & (df["Annual Income (k$)"] <= 120)]
ai121_150 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 121) & (df["Annual Income (k$)"] <= 150)]

aix = ["$ 0 - 30,000", "$ 30,001 - 60,000", "$ 60,001 - 90,000", "$ 90,001 - 120,000","$ 120,001 - 150,000"]
aiy = [len(ai0_30.values), len(ai31_60.values), len(ai61_90.values), len(ai91_120.values), len(ai121_150.values)]


plt.figure(figsize=(15,6))
sns.barplot(x=aix , y=aiy, palette="Spectral")
plt.title("Annual Incomes")
plt.xlabel("Income")
plt.ylabel("Number of Customer")
plt.show()
```



 # Next we plotted Within Cluster Sum Of Squares (WCSS) against the the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids.

 # Calculate the Within Cluster Sum of Squared Errors (WCSS) for different values of k, and choose the k for which WCSS first starts to diminish. In the plot of WCSS-versus k, this is visible as an elbow.

```
X1 =df.loc[:,["Age", "Spending Score (1-100)"]].values
from sklearn.cluster import KMeans
wcss = []
for k in range (1,11):
  kmeans = KMeans(n_clusters=k,init="k-means++")
  kmeans.fit(X1)
  wcss.append(kmeans.inertia_)

plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker="8")
plt.xlabel("K value")
plt.ylabel("wcss")
plt.show()
```



```
kmeans = KMeans(n_clusters=4)
label = kmeans.fit_predict(X1)
print(label)
```

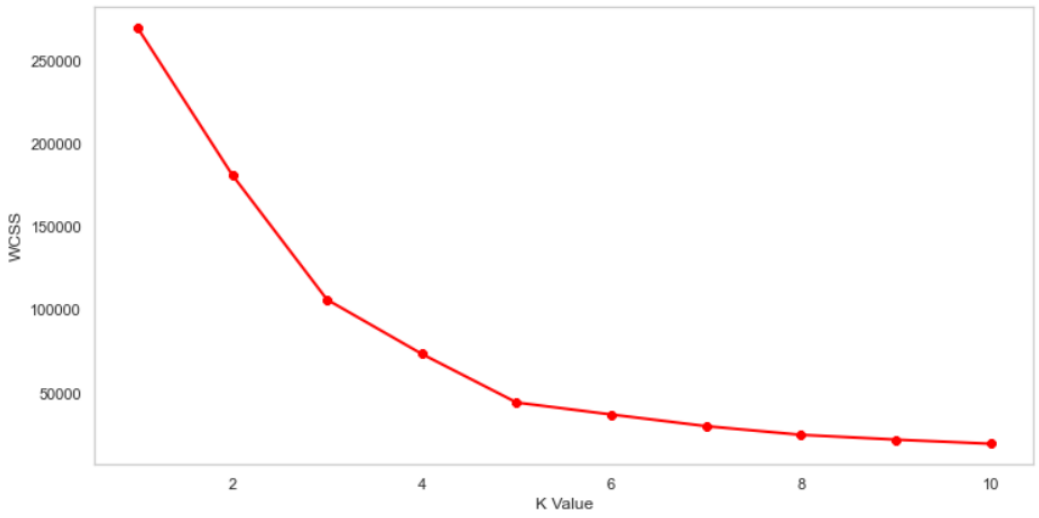
```
[0 3 1 3 0 3 1 3 1 3 1 3 1 3 1 3 0 0 1 3 0 3 1 3 1 3 1 0 1 3 1 3 1 3 1 3 1
 3 1 3 2 3 2 0 1 0 2 0 0 0 2 0 0 2 2 2 2 2 0 2 2 0 2 2 2 0 2 2 0 0 2 2 2 2
 2 0 2 0 0 2 2 0 2 2 0 2 2 0 0 2 2 0 2 0 0 0 2 0 2 0 0 2 2 0 2 0 2 2 2 2 2
 0 0 0 0 0 2 2 2 2 0 0 0 3 0 3 2 3 1 3 1 3 0 3 1 3 1 3 1 3 1 3 0 3 1 3 2 3
 1 3 1 3 1 3 1 3 1 3 1 3 2 3 1 3 1 3 1 3 1 0 1 3 1 3 1 3 1 3 1 3 1 3 1 3 0
 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3]
```

```
print(kmeans.cluster_centers_)
```

```
[[27.61702128 49.14893617]
 [43.29166667 15.02083333]
 [55.70833333 48.22916667]
 [30.1754386  82.35087719]]
```

```
plt.scatter(X1[:,0],X1[:,1],c=kmeans.labels_,cmap='rainbow')

plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],color='black')

plt.title('Clusters of Customers')
plt.xlabel('Age')
plt.ylabel('Spending Score (1-100)')
plt.show()
```

```
X2 =df.loc[:,["Annual Income (k$)","Spending Score (1-100)"]].values
from sklearn.cluster import KMeans
wcss=[]
for k in range (1,11):
  kmeans = KMeans(n_clusters=k,init="k-means++")
  kmeans.fit(X2)
  wcss.append(kmeans.inertia_)

plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=2,color='red',marker="8")
plt.xlabel("K Value")
plt.ylabel("WCSS")
plt.show()
```



```python
kmeans = KMeans(n_clusters=5)
label = kmeans.fit_predict(X2)
print(label)
```

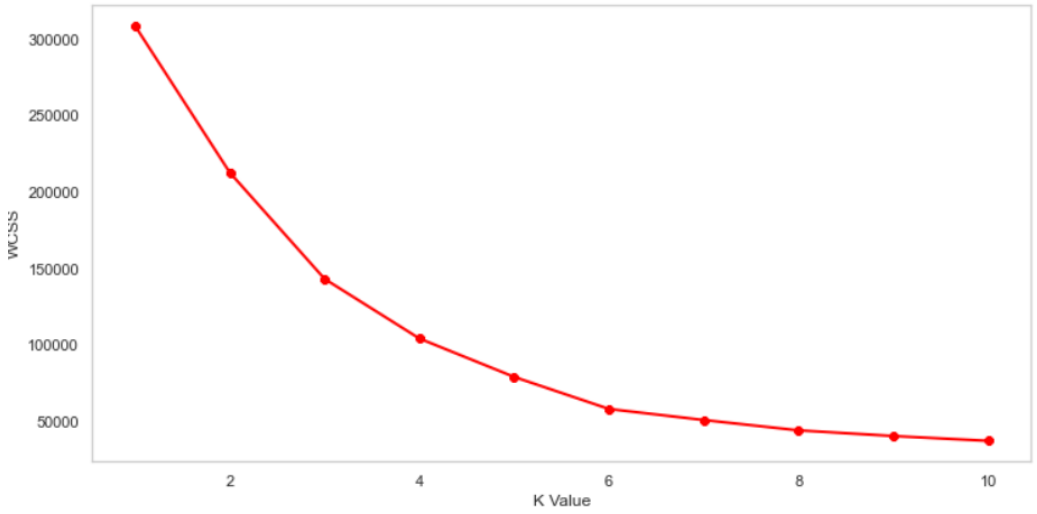
```
[3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3 1 3
 1 3 1 3 1 3 4 3 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
 4 4 4 4 4 4 4 4 4 4 4 4 0 2 0 4 0 2 0 2 0 4 0 2 0 2 0 2 0 2 0 4 0 2 0 2 0
 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0 2
 0 2 0 2 0 2 0 2 0 2 0 2 0 2 0]
```

```python
print(kmeans.cluster_centers_)
```

```
[[86.53846154 82.12820513]
 [25.72727273 79.36363636]
 [88.2        17.11428571]
 [26.30434783 20.91304348]
 [55.2962963  49.51851852]]
```

```python
plt.scatter(X2[:,0],X1[:,1],c=kmeans.labels_,cmap='rainbow')

plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],color='black')

plt.title('Clusters of Customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score (1-100)')
plt.show()
```

```
X3=df.iloc[:,1:]

wcss=[]
for k in range (1,11):
  kmeans = KMeans(n_clusters=k,init="k-means++")
  kmeans.fit(X3)
  wcss.append(kmeans.inertia_)

plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=2,color='red',marker="8")
plt.xlabel("K Value")
plt.ylabel("WCSS")
plt.show()
```




```
kmeans = KMeans(n_clusters=5)
label = kmeans.fit_predict(X3)
print(label)
```

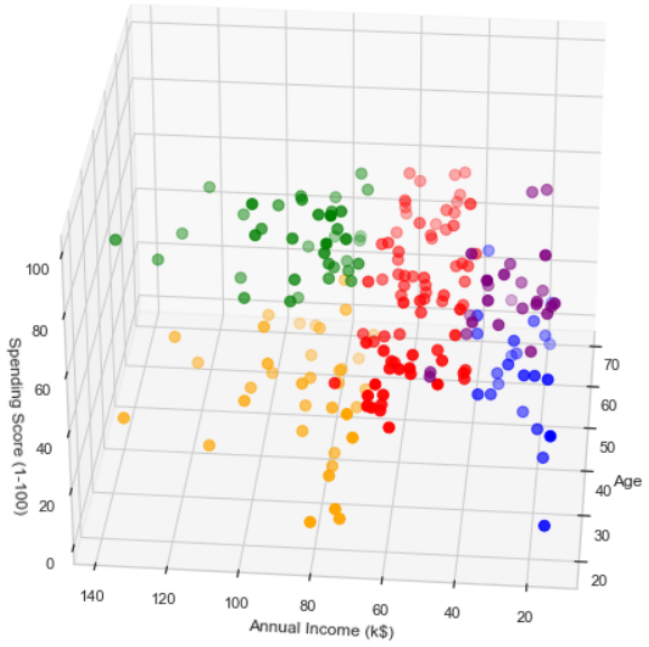
```
[3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3
 4 3 4 3 4 3 4 3 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 1 2 1 0 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 0 1 2 1 2 1
 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1]
```

```
print(kmeans.cluster_centers_)
```

```
[[43.08860759 55.29113924 49.56962025]
 [32.69230769 86.53846154 82.12820513]
 [40.66666667 87.75       17.58333333]
 [45.2173913  26.30434783 20.91304348]
 [25.52173913 26.30434783 78.56521739]]
```

 # Finally we made a 3D plot to visualize the spending score of the customers with their annual income. The data points are separated into 5 classes which are represented in different colours as shown in the 3D plot.

```
clusters=kmeans.fit_predict(X3)
df["label"] = clusters

import matplotlib.pyplot as plt
import mpl_toolkits
from mpl_toolkits.mplot3d import Axes3D

fig = plt.figure(figsize=(20,10))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(df.Age[df.label == 0], df["Annual Income (k$)"][df.label == 0], df["Spending Score (1-100)"][df.label == 0],
           c='blue', s=60)
ax.scatter(df.Age[df.label == 1], df["Annual Income (k$)"][df.label == 1], df["Spending Score (1-100)"][df.label == 1],
           c='red', s=60)
ax.scatter(df.Age[df.label == 2], df["Annual Income (k$)"][df.label == 2], df["Spending Score (1-100)"][df.label == 2],
           c='green', s=60)
ax.scatter(df.Age[df.label == 3], df["Annual Income (k$)"][df.label == 3], df["Spending Score (1-100)"][df.label == 3],
           c='orange', s=60)
ax.scatter(df.Age[df.label == 4], df["Annual Income (k$)"][df.label == 4], df["Spending Score (1-100)"][df.label == 4],
           c='purple', s=60)
ax.view_init(30, 185)
plt.xlabel("Age")
plt.ylabel("Annual Income (k$)")
ax.set_zlabel('Spending Score (1-100)')
plt.show()
```



6 Conclusion and Future Scope

6.0.1 Conclusion

In this project, we made a 3D plot to visualize the spending score of the customers with their annual income. The data points are separated into 5 classes which are represented in different colours in the 3D plot. For this purpose, we used K-Means clustering algorithm. The goal of K means is to group data points into distinct non-overlapping subgroups. We implemented K means clustering in segmentation of customers to get a better understanding of them and their behaviours which in turn could be used to increase the revenue of the company as with this we are able to understand the customers like who are the target customers so that the sense can be given to marketing team and plan the strategy accordingly.

6.0.2 Future Scope

Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

7 References

[1] Bruce Cooil , Vanderbilt University , "Journal Of Relationship Marketing" , Approaches to Customer Segmentation , Available at : https://www.researchgate.net/publication/230557972_Customer_Segmentation

[2] Sulekha Goyat Department of humanities &social sciences, National Institute of Technology, Kurukshetra, "Journal of Business and Management" , "The basis of market segmentation" , Available at : https://core.ac.uk/download/pdf/234624114.pdf

[3] Mark Anthony Camilleri , PhD (Edinburgh)", Customer Segmentation , Targeting and Positioning", "Market Segmentation with Targeting" , Available at : https://www.um.edu.mt/123456789/Market%20Segmentation