

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

MAAC Accredited with A++ Grade



Project Report

on

Customer Segmentation

Submitted By:

Piyush Rathore(0901AM211037)

Anshika Tripathi(0901AM211012)

Faculty Mentor:

Prof. Nitya Thagele

Assistant Professor

CENTRE FOR ARTIFICIAL INTELLIGENCE

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE

GWALIOR - 474005 (MP) est. 1957

JULY-DEC. 2023

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

CERTIFICATE

This is certified that **Piyush Rathore**(0901AM211037) and **Anshika Tripathi**(0901AM211012) has submitted the project report titled **Customer Segmentation** under the mentorship of **Prof. Nitya Thagele** in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in AIML from Madhav Institute of Technology and Science, Gwalior.



Prof. Nitya Thagele
Faculty Mentor
Assistant Professor
Centre for Artificial Intelligence



Dr. R. R. Singh
Coordinator
Centre for Artificial Intelligence

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in AIML at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of **Prof. Nitya Thagele**, Assistant Professor, Centre for Artificial Intelligence.

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Piyush Rathore
0901AM211037

Piyush

Anshika Tripathi
0901AM211012

Anshika

3rd Year,
Centre for Artificial Intelligence

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Centre for Artificial Intelligence**, for allowing me to explore this project. I humbly thank **Dr. R. R. Singh**, Coordinator, Centre for Artificial Intelligence, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Prof. Nitya Thagele**, Assistant Professor, Centre for Artificial Intelligence, for her continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Piyush Rathore
0901AM211037

Anshika Tripathi
0901AM211012

3rd Year,
Centre for Artificial Intelligence

ABSTRACT

Customer segmentation is simply grouping customers with similar characteristics. These characteristics include geography, demography, behavioural, purchasing power, situational factors, personality, lifestyle, psychographic, etc. The goals of customer segmentation are customer acquisition, customer retention, increasing customer profitability, customer satisfaction, resource allocation by designing marketing measures or programs and improving target marketing measures.

Clustering is an efficient technique used for customer segmentation. Clustering places homogenous data points in a given dataset. Each of these groups is called a cluster. While the objects in each cluster are similar between themselves, they are dissimilar to the objects of other groups. Clustering is a type of data mining approach in machine learning classified under unsupervised learning. This is because it is able to discover patterns and information from unlabelled data. It is used extensively in machine learning, classification, and pattern recognition.

Clustering algorithms include the K-means algorithm, hierarchical clustering, DBSCAN. In this project, the k-means clustering algorithm has been applied in customer segmentation. K-means is a clustering algorithm based on the principle of partition. The letter k represents the number of clusters chosen. It is the most common centroid-based algorithm.

Table Of Contents

Page No

1-7

Certificate.....

Declaration.....

Acknowledgement.....

Title/Abstract.....

List Of figures.....

Chapter 1 :Project Overview 7-9

1. INTRODUCTION.....

2. DATASET.....

3. Scope of Analysis.....

Chapter 2: Micro Level Analysis 9-12

1. Customer Segmentation Analysis

1. Challenges of Performing Analysis.....

2. Algorithm.....

Chapter 3: Macro Level Analysis 12-19

1. Clustering

2. Centroid Based : K means.....

3.43Solution.....

Chapter 4: Mini Level Analysis (Final Analysis and Design) 19-28

1. Preparing the Data.....

2. Criteria for Clustering.....

3. Scaling and Reformatting the Data.....

4. Brief Overview of Code.....

5. Clustering Results.....

Chapter 5: Conclusion..... 28-32

References.....

List Of Figures

Figure No.	Name of Figure	Page No.
Figure 2.4.1	Structure of Department	4
Figure 4.6.1	Gender Distribution	12
Figure 4.6.2	Age Distribution	13
Figure 4.6.3	Annual Income Distribution	14
Figure 4.6.4	Spending Score Distribution	15
Figure 4.6.5	Annual Income Vs Spending Score	16
Figure 4.6.6	Elbow Graph	15
Figure 4.6.7	Graph after applying K-means	17

Chapter 1: Project Overview

Effective decisions are mandatory for any company to generate good revenue. In these days competition is huge and all companies are moving forward with their own different strategies. We should use data and take a proper decision. Every person is different from one another and we don't know what he/she buys or what their likes are. But, with the help of machine learning technique one can sort out the data and can find the target group by applying several algorithms to the dataset. Without this, it will be very difficult and no better techniques are available to find the group of people with similar character and interests in a large dataset. Here, The customer segmentation using K-Means clustering helps to group the data with same attributes which exactly helps to business the best. We are going to use elbow method to find the number of clusters and at last we visualize the data.

2. Keywords

Clustering, Elbow Method, K-Means Algorithm, Customer Segmentation, Visualization.

Introduction

1.1 Introduction

Nowadays the competition is vast and lot of technologies came into account for effective growth and revenue generation. For every business the most important component is data. With the help of grouped or ungrouped data, we can perform some operations to find customer interests.

Data mining helpful to extract data from the database in a human readable format. But, we may not know the actual beneficiaries in the whole dataset. Customer Segmentation is useful to divide the large data from dataset into several groups based on their age, demographics, spent, income, gender, etc. These groups are also known as clusters. By this, we can get to know that, which product got huge number of sales and which age group are purchasing etc. And, we can supply that product much for better revenue generation.

Initially we are going to take the old data. As we know that old is gold so, by using the old data we are going to apply K-means clustering algorithm and we have to find the number of clusters first. So, at lastly, we have to visualize the data. One can easily find the potential group of data while observing that visualization.

The goal of this paper is to identify customer segments using the data mining approach, using the

partitioning algorithm called as K-means clustering algorithm. The elbow method determines the optimal clusters.

3.2 Problem Statement

Customer Segmentation is the best application of unsupervised learning. Using clustering, identify segments of customers in the dataset to target the potential user base. They divide customers into various groups according to common characteristics like gender, age, interest, and spending habits so they can market to each group effectively. Use K-Means Clustering and also visualize the gender and age distributions. Then analyze their annual income and spending scores. As it describes about how we can divide the customers based on their similar characteristics according to their needs by using k-means clustering which is a classification of unsupervised machine learning.

4. Existing System

The existing method is storing customer data through paperwork and computer software (digital data) is increasing day by day. At end of the day they will analyse their data as how many things are sold or actual customer count etc. By analysing the collected data they got to know who is beneficial to their business and increase their sales. It requires more time and more paperwork. Also, it is not much effective solution to find the desired customers data.

5. Proposed System

1. Proposed Method

To overcome the traditional method i.e paper work and computerized digital data this new method will play vital role. As we collect a vast data day by day which requires more paperwork and time to do. As new technologies were emerging in today's world. Machine Learning which is powerful innovation which is used to predict the final outcome which has many algorithms. So for our problem statement we will use K-Means Clustering which groups the data into different clusters based on their similar characteristics. And then we will visualize the data.

2. System Architecture

Initially we will see the dataset and then we will perform exploratory data analysis which deals with the missing data, duplicates values and null values. And then we will deploy our algorithm k-means clustering which is unsupervised learning in machine learning.



As in order to find the no of clusters we use elbow method where distance will be calculate through randomly chosen centres and repeat it until there is no change in cluster centres. Thereafter we will analyse the data through data visualization. Finally we will get the outcome.

3. Algorithm

1. K-Means Clustering

- ⦿ K Means algorithm in an iterative algorithm that tries to partition the dataset into K predefined distinct non overlapping sub groups which are called as cluster.
- ⦿ Here K is the total no of clusters.
- ⦿ Every point belongs to only one cluster.
- ⦿ Clusters cannot overlap.

5.3.2 Steps of Algorithm

- ⦿ Arbitrarily choose k objects from D as the initial cluster centers.
- ⦿ Repeat.
- ⦿ Assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster.

- Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
- Until no change.

6. Methodology

1. First of all we will import all the necessary libraries or modules (pandas, numpy, seaborn).
2. Then we will read dataset and analyse whether it contains any null values, missing values and duplicate values. So we will fix them by dropping or fixing the value with their means, medians etc which is technically named as Data Preprocessing.
3. We will deploy our model algorithm K-Means Clustering, which divides the data into group of clusters based on similar characteristics. To find no. of clusters we will use elbow method.
4. Finally, we will visualize our data using matplotlib, which concludes the customers divided into groups who are similar to each other on their group.

7. Implementation And Analysis

1. Overview of a Dataset

This is a mall customer segmentation data which contains 5 columns and 200 rows. For this Project we have used Mall Customer Dataset taken from "kaggle", here our main objective is to divide customers into groups according to common characteristics.

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
...					
196	196	Female	36	120	70
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	63

200 rows × 5 columns

2. Exploratory Data Analysis

It deals with the data preprocessing, whether it contains any missing values or null values. There after we will see the information and description of the dataset.

1. Information of the dataset

```
#df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
#   Column                Non-Null Count  Dtype
---  --
0   CustomerID            200 non-null    int64
1   Gender                200 non-null    object
2   Age                   200 non-null    int64
3   Annual Income (k$)    200 non-null    int64
4   Spending Score (1-100) 200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

As here it overview the information of the data. And it gives it doesn't contain any null values.

As we will remove the irrelevant data which is customer id.

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

```
In [10]: df
Out[10]:
   Gender  Age  Annual Income (k$)  Spending Score (1-100)
0   Male   19                15               40
1   Male   21                17               51
2   Female  21                16               3
3   Female  21                16               51
4   Female  21                17               40
5   Female  22                17               70
6   Female  23                16               7
7   Female  23                16               56
```

7.2.2 Description of the data

```
#df.describe()
```

	Age	Annual Income (k\$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000
mean	38.850000	60.560000	50.200000
std	13.969007	26.264721	25.823622
min	18.000000	15.000000	1.000000
25%	28.750000	41.500000	34.750000
50%	36.000000	61.500000	50.000000
75%	49.000000	78.000000	73.000000
max	70.000000	137.000000	99.000000

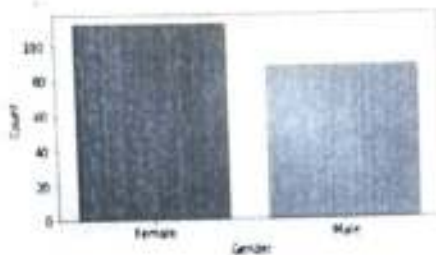
It describes about the count which counts the no of rows in it, mean of the columns, standard deviations, maximum and minimum and percentiles etc.

7.3 Gender plot Analysis

Here it overview the gender analysis

```
#Gender Distribution
genders=df.Gender.value_counts()
plt.figure(figsize=(6,1))
sns.barplot(x=genders.index,y=genders.values)
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show
```

So we label the x-axis as Gender and y-axis as Count and we plot it by using barplot.



From the plot we will conclude that there are more female customers than the male customers i.e. female customers are more than 100 whereas male customers are nearly 80.

7.4 Age plot

We will use distplot for the distribution of age of the customers.

```
plt.figure(figsize=(6,3))
sns.distplot(df['Age'])
plt.title('Age Distribution')
plt.xlabel('Range of Age')
plt.ylabel('Count')
plt.show()
```

So we label X-axis as range of age and y-axis as count.

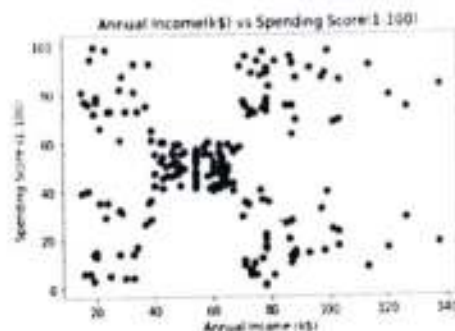


From the plot, it varies the age from nearly 20 to 70. It is evident that the age of the customers between 30 - 40 are more, then after 20-30 etc.

7.5 Annual Income vs Spending Score

As we will use scatterplot and labelled x-axis as Annual Income(k\$) and y-axis as Spending Score(1-100)

```
plt.scatter(df['Annual Income (k$)',df['Spending Score (1-100)'])  
plt.title('Annual Income(k$) vs Spending Score(1-100)')  
plt.xlabel('Annual Income (k$)')  
plt.ylabel('Spending Score (1-100)')  
plt.show()
```



From the plot we observed that it varies from low annual income with low expenditure or spending money to high annual income with high expenditure.

7.6 Elbow Method

The elbow method is based on the observation that increasing the number of clusters can help to reduce

the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture

finer groups of data objects that are more similar to each other.

To define the optimal clusters, Firstly, we use the clustering algorithm for various values of k. This is

done by ranging k from 1 to 10 clusters. Then we calculate the total intra-cluster sum of square. Then,

we proceed to plot intra-cluster sum of square based on the number of clusters. The plot denotes the

approximate number of clusters required in our model. The optimum clusters can be found from the graph

where there is a bend in the graph.

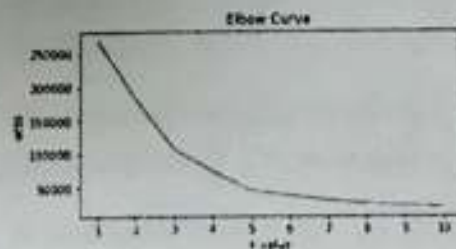
First we will consider the data X which as only two columns they are annual income and spending score.

```
X=df[['Annual Income (k$)','Spending Score (1-100)']]
```

X.head()

	Annual Income (k\$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

```
wcss=[]  
for i in range(1,11):  
    km=KMeans(n_clusters=i)  
    km.fit(X)  
    wcss.append(km.inertia_)  
  
plt.figure(figsize=(6,3))  
plt.plot(range(1,11),wcss)  
plt.title('Elbow Curve')  
plt.xlabel('k value')  
plt.xticks(np.arange(1,11,1))  
plt.ylabel('wcss')  
  
plt.show()
```



So from the graph we observed that the at 5 there is bend and it can be considered as k which is no of clusters.

Therefore, k=5 i.e no of clusters are equal to 5.

7.7 Fitting the Algorithm

```
km=KMeans(n_clusters=5)
km.fit(X)
y=km.predict(X)
df['cluster']=y
df.head()
```

As here we initialized the kmeans as km with 5 clusters and we will fit it. There after we will predict the data and store it in y. And then we will add new column named as Cluster and data as y.

	Gender	Age	Annual Income (k\$)	Spending Score (1-100)	Cluster
0	Male	19	15	39	4
1	Male	21	15	81	3
2	Female	20	18	6	4
3	Female	23	16	77	3
4	Female	31	17	40	4

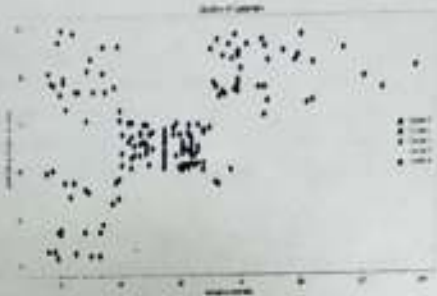
So from the figure we observed that each customer is labelled with cluster which is based on their characteristics.

7.8 Visualization the clusters

Visualizing the clusters based on Annual Income and Spending Score of the customers. As here we plot a graph named as Clusters of Customers to visualize the data in terms of groups or cluster.

```
plt.figure(figsize=(15,7))
plt.scatter(df["Annual Income (k$)"][df.Cluster == 0], df["Spending Score (1-100)"]
[df.Cluster == 0], c='blue', s=60, label='Cluster 0')
plt.scatter(df["Annual Income (k$)"][df.Cluster == 1], df["Spending Score (1-100)"]
[df.Cluster == 1], c='red', s=60, label='Cluster 1')
plt.scatter(df["Annual Income (k$)"][df.Cluster == 2], df["Spending Score (1-100)"]
[df.Cluster == 2], c='green', s=60, label='Cluster 2')
plt.scatter(df["Annual Income (k$)"][df.Cluster == 3], df["Spending Score (1-100)"]
[df.Cluster == 3], c='yellow', s=60, label='Cluster 3')
plt.scatter(df["Annual Income (k$)"][df.Cluster == 4], df["Spending Score (1-100)"]
[df.Cluster == 4], c='black', s=60, label='Cluster 4')
```

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



So from the above one we observed that there are 5 clusters which are named as 0, 1, 2, 3, 4.

- Cluster 0 which is at centre, average annual income with average spending score.
- Cluster 1 which is at top right, highest annual income with highest spending score.
- Cluster 2 which is at top left, lowest annual income with highest spending score.
- Cluster 3 which is at bottom right, high annual income with low spending score.
- Cluster 4 which is at bottom left, lowest annual income with lowest spending score.

Customer Segmentation - Using k-means

About: Customer Segmentation is a popular application of unsupervised learning. Using clustering, identify segments of customers to target the potential user base. They divide customers into groups according to common characteristics like gender, age, interests, and spending habits so they can market to each group effectively.

Use K-means clustering and also visualize the gender and age distributions. Then analyze their annual incomes and spending scores.

CODE:

In [1]:

```
#importing required libraries
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```

```
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

In [2]:

```
#reading data using pandas to a dataframe and printing its head values
customer = pd.read_csv('Customer-
Segmentation.csv')
customer.head()
```

Out[2]:

	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

There are 5 columns CustomerID, Gender, Age, Annual Income and Spending Score in our dataframe 'customer'

In [3]: #checking size of data
customer.shape

Out[3]: (200, 5)

We have a data set with 200 rows and 5 columns.

In [4]: #checking dataframe for any NULL values
customer.isnull().sum()

Out[4]: CustomerID 0
Gender 0
Age 0
Annual Income (k\$) 0
Spending Score (1-100) 0
dtype: int64

It clearly shows that there is no NULL value present in our dataframe.

```
customer.describe()
```

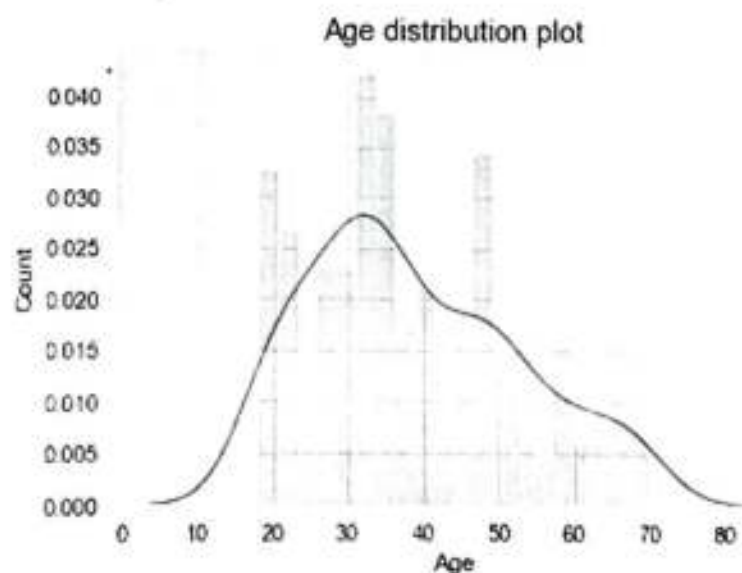
	CustomerID	Age	Annual Income (k\$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000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

We got values like mean, std deviation, min, max, Q1, Q2 and Q3 for all attributes.

```
In [6]: # applying grid to all our plots for better visuals
sns.set(style="whitegrid")
```

Visualizing various Distributions

```
In [7]: # distribution plot for 'Age'
sns.distplot(customer['Age'], color='green', bins=20)
plt.title('Age distribution plot', fontsize=15)
plt.xlabel('Age', fontsize=12)
plt.ylabel('Count', fontsize=12)
plt.show()
```



This shows that our data has customer ranges from 10 years to 80 years.

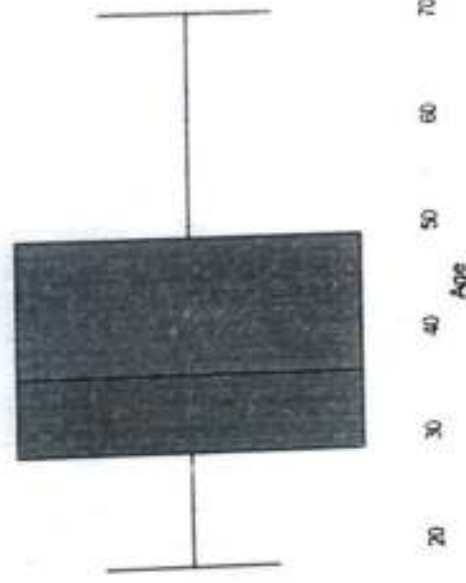
```
plt.figure(figsize=(10, 10))
sns.countplot(customer['Age'])
plt.title('Age count plot', fontsize = 12)
plt.xlabel('Age', fontsize = 12)
plt.ylabel('Count', fontsize = 12)
plt.show()
```



This plot is more clear view on counting customer based on their Age. Also we can see that 11 customers are 32 years old which is the most value count.

```
In [9]: # box plot for 'Age'
sns.boxplot(customer['Age'])
plt.title('Age box plot', fontsize = 12)
plt.xlabel('Age', fontsize = 12)
plt.show()
```

Age box plot

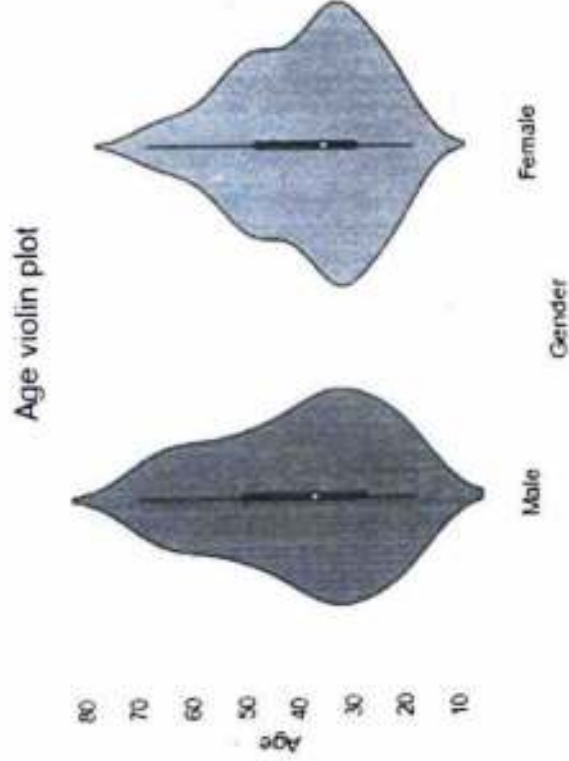


Based on 5 point summary we can get a clear picture of various aspect of customer based on their age.

```

sns.violinplot(y='Age', x='Gender', data=customer)
plt.title('Age Violin Plot', fontsize=12)
plt.xlabel('Gender', fontsize=12)
plt.ylabel('Age', fontsize=12)
plt.show()

```

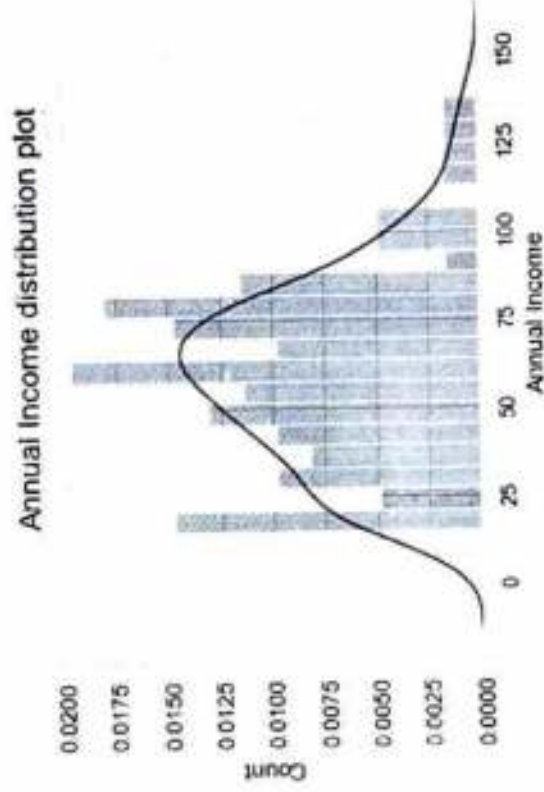


This violin plot shows that we have higher number of female customer who belongs to age group of 30 years.

```

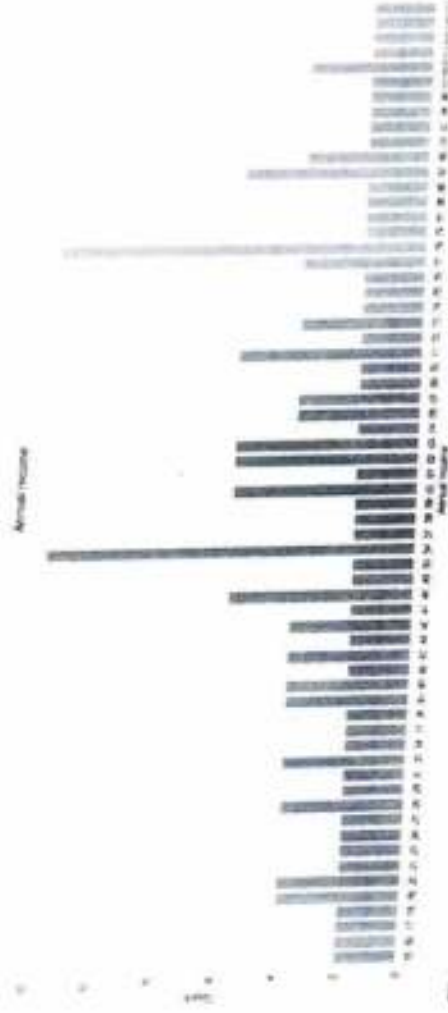
In [11]: # distribution plot for 'Annual Income'
sns.distplot(customer['Annual Income (k$)'], color='blue', bins=10)
plt.title('Annual Income distribution plot', fontsize=12)
plt.xlabel('Annual Income', fontsize=12)
plt.ylabel('Count', fontsize=12)
plt.show()

```



This shows that our data has customer ranges from income of 0k to 150k.

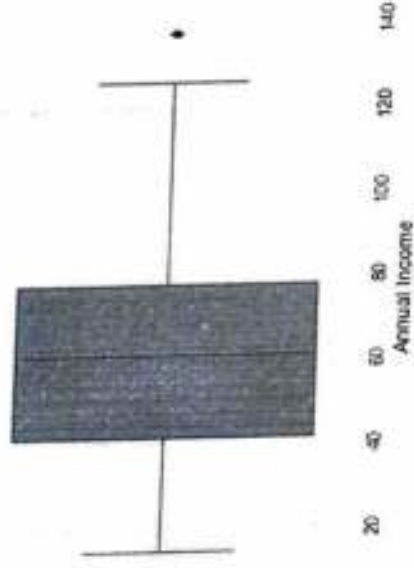
```
plt.figure(figsize=(10, 10))
sns.countplot(customer, hue='Annual Income (k$)',
               plt.title='Annual Income', font_size=12,
               plt.xlabel='Age (years)', font_size=12,
               plt.ylabel='Count', font_size=12,
               plt.show())
```



This plot is more clear view on counting customer based on their Income. Also we can see that 12-12 customers are 54 years and 78 years old which is the most value count.

```
In [13]: #sns plot for Annual Income
sns.boxplot(customer, 'Annual Income (k$)',
             plt.title='Annual Income box plot', font_size=12,
             plt.xlabel='Annual Income', font_size=12,
             plt.show())
```

Annual Income box plot

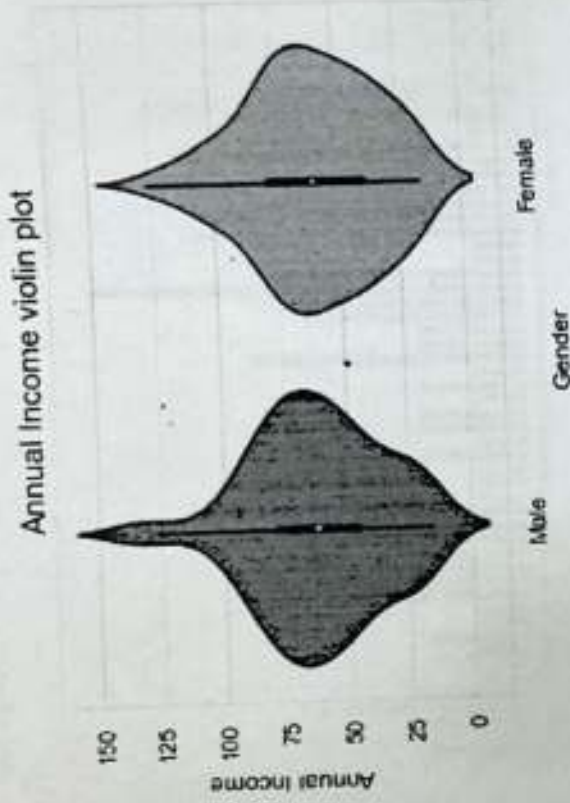


Based on 5 point summary we can get a clear picture of various aspect of customer based on their Income.

```

In [14]: sns.violinplot(x='Annual Income ($)', y='Gender', data=customer)
plt.title('Annual Income violin plot', fontsize=12)
plt.xlabel('Gender', fontsize=12)
plt.ylabel('Annual Income', fontsize=12)
plt.show()

```

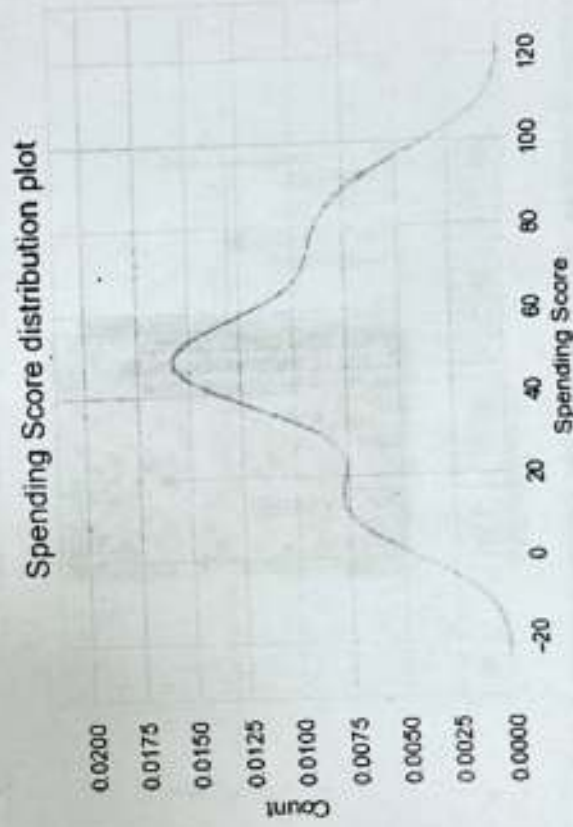


This violin plot shows that we have higher number of male customer who have more income.

```

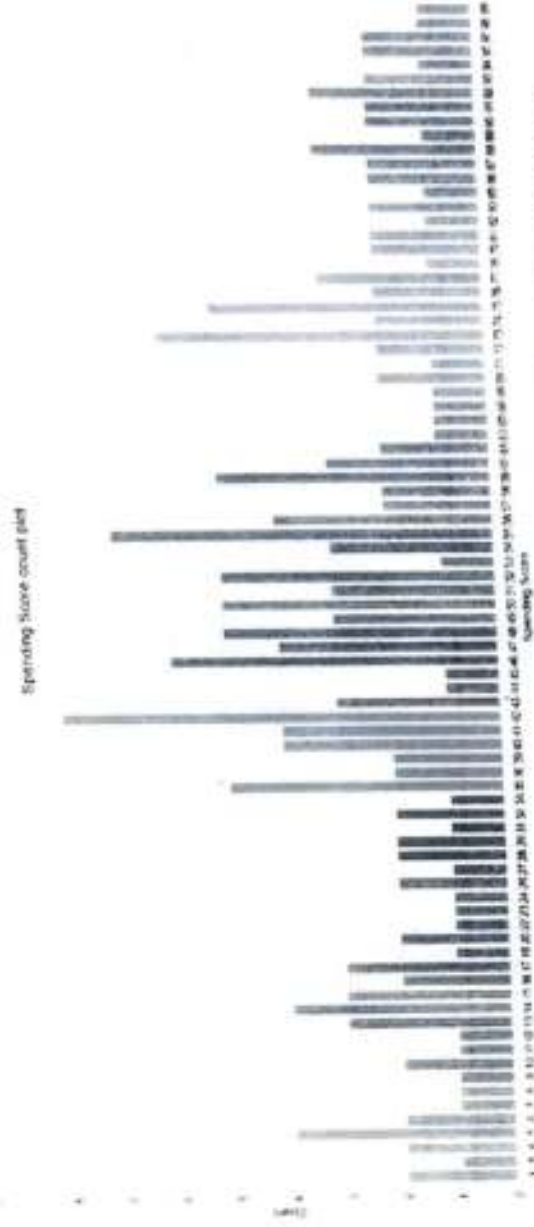
In [15]: # distribution plot for 'Spending Score'
sns.distplot(customer['Spending Score (1-100)'], color='orange', bins=20)
plt.title('Spending Score distribution plot', fontsize=12)
plt.xlabel('Spending Score', fontsize=12)
plt.ylabel('Count', fontsize=12)
plt.show()

```



This shows that our data has customer ranges from with -20 to 120 spending score.

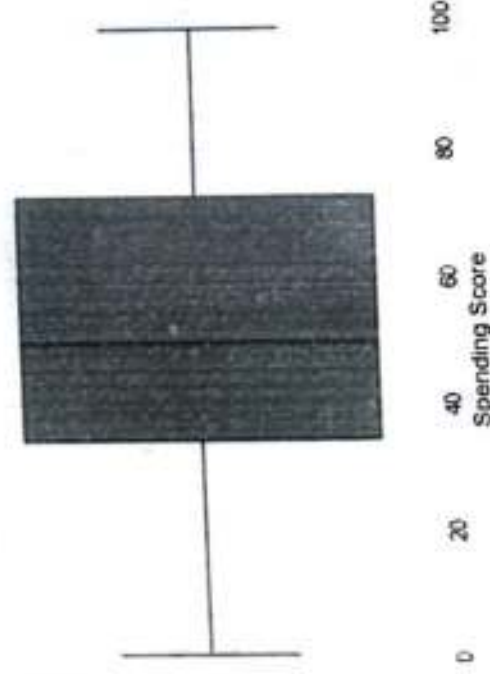
```
plt.figure(figsize=(10,10))
sns.countplot(customer['Spending Score (1-100)'])
plt.title('Spending Score count plot', fontsize = 15)
plt.xlabel('Spending Score', fontsize = 15)
plt.ylabel('count', fontsize = 15)
plt.show()
```



This plot is more clear view on counting customer based on their Age. Also we can see that 8 customers are 42 years old which is most value count.

```
In [17]: # plt plot for 'Spending Score'
sns.boxplot(customer['Spending Score (1-100)'])
plt.title('Spending Score', fontsize = 15)
plt.xlabel('Spending Score', fontsize = 12)
plt.show()
```

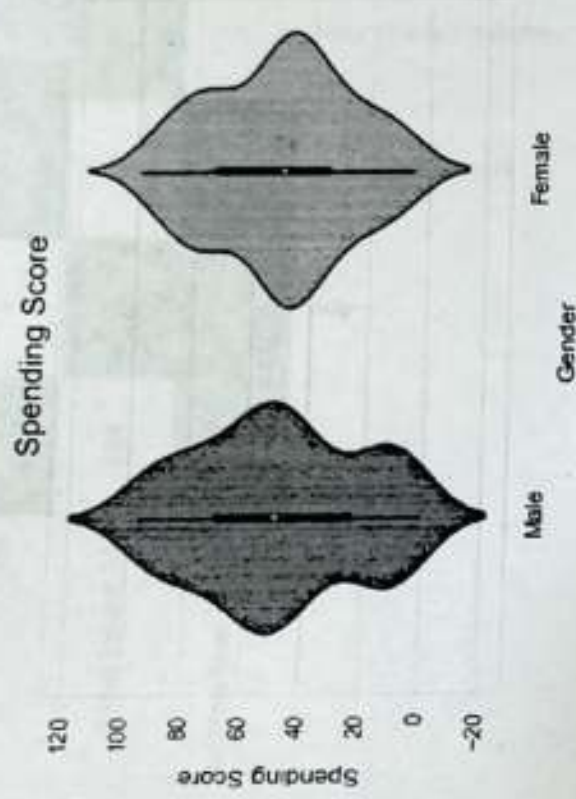
Spending Score



Based on 5 point summary we can get a clear picture of various aspect of customer based on their spending score.

```
In [18]:
```

```
# Violin Plot for 'Spending Score'  
sns.violinplot(y = 'Spending Score (1-100)', x = 'Gender', data = customer  
plt.title('Spending Score', fontsize = 12)  
plt.xlabel('Gender', fontsize = 12)  
plt.ylabel('Spending Score', fontsize = 12)  
plt.show()
```

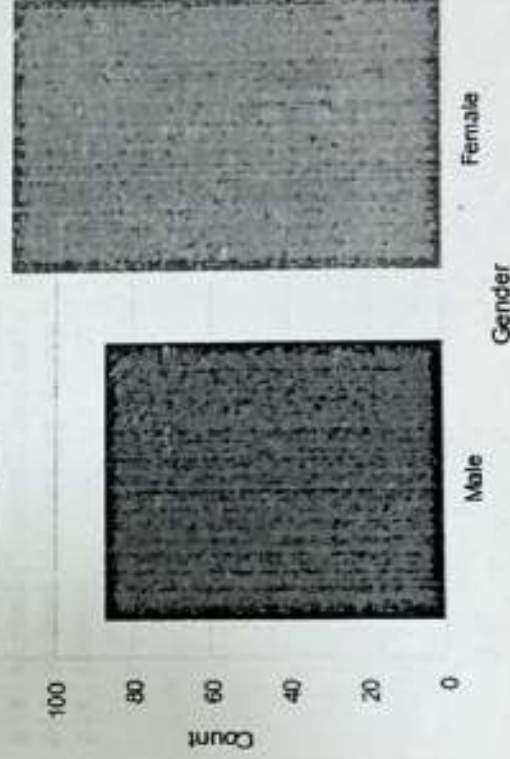


This violin plot shows that we have higher number of female customer who have mostly spending score around 50 .

```
In [19]:
```

```
# count plot for 'Gender'  
sns.countplot(x='Gender', data=customer)  
plt.title('Spending Score count plot', fontsize = 15)  
plt.xlabel('Gender', fontsize = 12)  
plt.ylabel('Count', fontsize = 12)  
plt.show()
```

Spending Score count plot

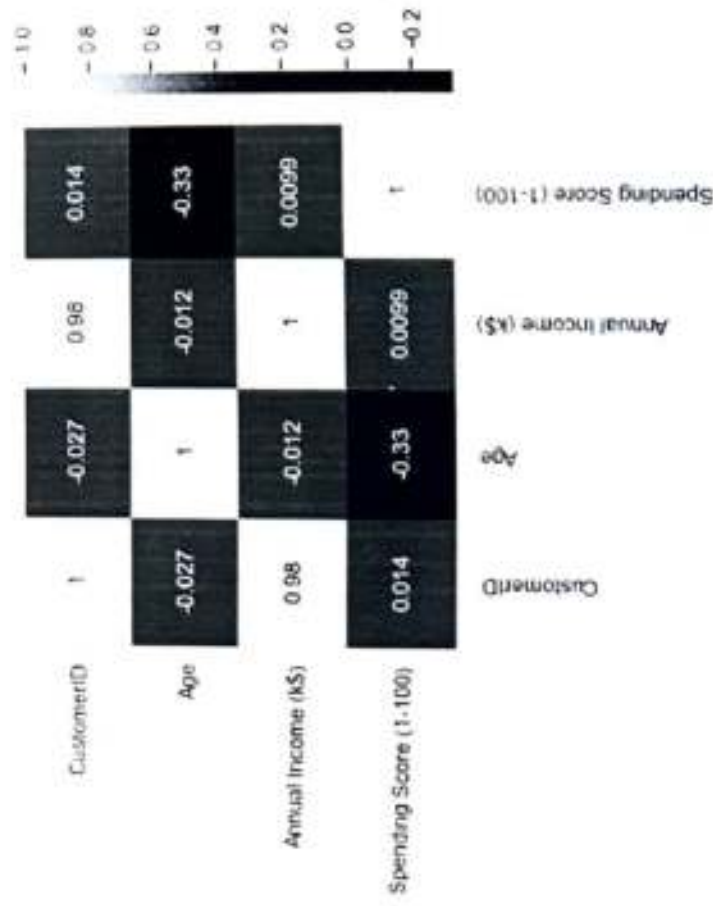


This plot clearly shows that we have more female customer compare to male customers.

```
In [20]:
```

```
# heatmap to show correlation of various features  
sns.heatmap(customer.corr(), annot = True)
```

```
<AxesSubplot: >
```



From this plot we got that income and spending score correlates to each other with a good score. But age and spending score does not correlates efficiently.

Cluster based on Annual Income and Spending Score

```
In [21]: plt.figure(figsize=(15,8))  
sns.scatterplot(customer['Annual Income (k$)'], customer['Spending Score  
(1-100)'], palette='red', 'green', alpha=0.6)  
plt.title('Distribution of Gender based on Annual Income and Spending Score')  
plt.xlabel('Annual Income', fontsize=12)  
plt.ylabel('Spending Score', fontsize=12)  
plt.show()
```

Distribution of Gender based on Annual Income and Spending Score

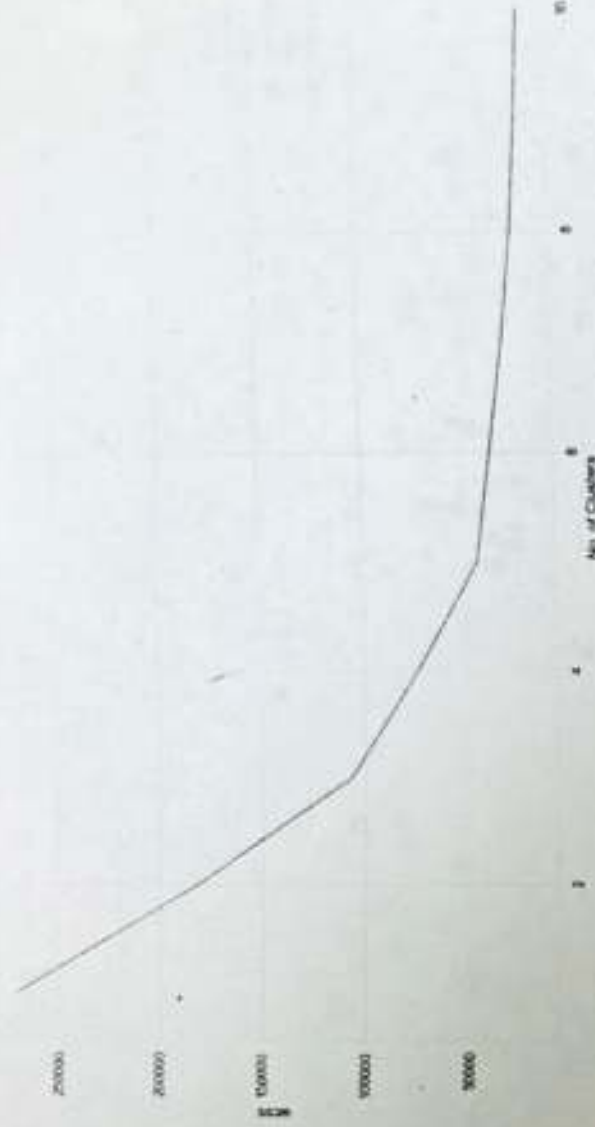


This scatter plot show the distribution of customers based on their income, spending score and gender. And we can see customer cluster clearly in this plot.

```
In [22]: Income_Spend = customer[['Annual Income (k$)', 'Spending Score (1-100)']]
from sklearn.cluster import KMeans

wcss = []
for i in range(1, 11):
    km = KMeans(n_clusters = i, init = 'k-means++', max_iter = 300, n_init
    km.fit(Income_Spend)
    wcss.append(km.inertia_)

plt.figure(figsize=(15,8))
plt.plot(range(1, 11), wcss)
plt.title('The Elbow Method', fontsize = 15)
plt.xlabel('No. of Clusters', fontsize = 12)
plt.ylabel('wcss', fontsize = 12)
plt.show()
```



This elbow method show a low slope line after 5 number of cluster so we can take 5 as optimum number of cluster.

```
In [23]: km = KMeans(n_clusters = 5, init = 'k-means++', max_iter = 300, n_init = 1
y_means = km.fit_predict(Income_Spend)

plt.figure(figsize=(15,8))
plt.scatter(Income_Spend[y_means == 0, 0], Income_Spend[y_means == 0, 1],
plt.scatter(Income_Spend[y_means == 1, 0], Income_Spend[y_means == 1, 1],
plt.scatter(Income_Spend[y_means == 2, 0], Income_Spend[y_means == 2, 1],
plt.scatter(Income_Spend[y_means == 3, 0], Income_Spend[y_means == 3, 1],
plt.scatter(Income_Spend[y_means == 4, 0], Income_Spend[y_means == 4, 1],
plt.scatter(km.cluster_centers_[0,0], km.cluster_centers_[0, 1], s = 50, c
plt.legend()

plt.title('Customer Segmentation using Annual Income and Spending Score',
plt.xlabel('Annual Income', fontsize = 12)
plt.ylabel('Spending Score', fontsize = 12)
plt.show()
```



Based on the above clustering we can clearly say that there are five cluster segments present based on customers' Annual Income and Spending Score. We named them as Low budget, Spenders, Average, Savers, and Best.

Cluster based on Age and Spending Score

```
In [24]: plt.figure(figsize=(15,8))
sns.scatterplot(customer['Age'], customer['Spending Score (1-100)'], hue=customer['Gender'])
plt.title('Distribution of Gender based on Age and Spending Score', fontsize=12)
plt.xlabel('Age', fontsize=12)
plt.ylabel('Spending Score', fontsize=12)
plt.show()
```

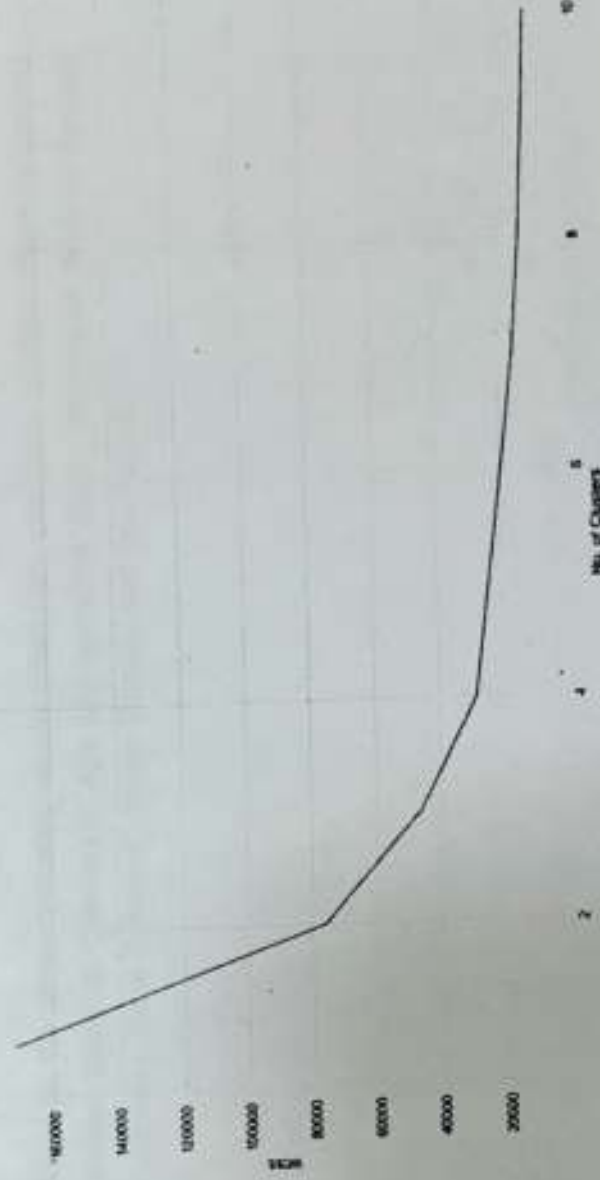


This scatter plot show the distribution of customers based on their age, spending score and gender. And we can clearly observe that aged people don't have higher spending score.

```
In [25]: Age_Spend = customer[['Age', 'Spending Score (1-100)']].iloc[:, :].value
wcss = []
for i in range(1, 11):
    km = KMeans(n_clusters=i, init='k-means++', max_iter=100, n_init=
    km.fit(Age_Spend)
    wcss.append(km.inertia_)

plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), wcss)
plt.title('The Elbow Method', fontsize=15)
plt.xlabel('No. of Clusters', fontsize=12)
plt.ylabel('wcss', fontsize=12)
plt.show()
```

The Elbow Method



This elbow method show a low slope line after 4 number of cluster so we can take 4 as optimum number of cluster.

```
In [26]: km = KMeans(n_clusters=4, init='k-means++', max_iter=300, n_init=1
y_means = km.fit_predict(Age_Spend)

plt.figure(figsize=(15, 8))
plt.scatter(Age_Spend[y_means == 0, 0], Age_Spend[y_means == 0, 1], s=100,
plt.scatter(Age_Spend[y_means == 1, 0], Age_Spend[y_means == 1, 1], s=100,
plt.scatter(Age_Spend[y_means == 2, 0], Age_Spend[y_means == 2, 1], s=100,
plt.scatter(Age_Spend[y_means == 3, 0], Age_Spend[y_means == 3, 1], s=100,
plt.scatter(km.cluster_centers_[0, 0], km.cluster_centers_[0, 1], s=50,
plt.legend()
plt.title('Customer Segmentation using Annual Income and Spending Score',
plt.xlabel('Age', fontsize=12)
plt.ylabel('Spending Score', fontsize=12)
plt.show()
```

Customer Segmentation using Annual Income and Spending Score



Based on the above clustering we can clearly say that there are four cluster segments present based on customers' Age and Spending Score. We named them as Regular Customers, Usual Customer, Young Targets, and Old Targets.

Conclusion

The goal of K means is to group data points into distinct non-overlapping subgroups.

Cluster 3: high spending scores and high-income; alert them with new arrivals as they are potential customer for increase in revenue.

```
##      ID Gender Age AnnualIncome SpendingScore
## 124 124 Male 39      59          91
## 125 125 Female 31      70          88
## 126 126 Male 40      71          95
## 127 127 Male 38      71          79
## 128 128 Male 39      71          75
## 129 129 Female 31      72          71
```

Cluster 1: high income and low spending score; ask them for feedback and advertise them with new produces that might attracts them, they have the potential to convert into cluster 4.

```
##      ID Gender Age AnnualIncome SpendingScore
## 127 127 Male 39      71          91
## 128 128 Male 39      71          91
## 129 129 Male 47      71          88
## 130 130 Male 20      73          85
## 131 131 Female 44      73          71
## 132 132 Male 19      74          70
```

Cluster 2: low income and high spending scores; can help them by providing new deals and sales offers so that despite low income they still remain loyal.

```
##      ID Gender Age AnnualIncome SpendingScore
## 2 2 Male 21      15          81
## 4 4 Female 23      16          77
## 6 6 Female 22      17          72
## 8 8 Female 23      18          64
## 10 10 Female 30      19          73
## 12 12 Female 14      19          59
```

Cluster 3: low income and low spending score; it won't be beneficial to both the parties to target these customers.

```
##      ID Gender Age AnnualIncome SpendingScore
## 1 1 Male 19      15          39
## 3 3 Female 20      16          6
## 5 5 Female 31      17          60
## 7 7 Female 25      18          6
## 9 9 Male 54      19          3
## 11 11 Male 47      19          14
```

References

- [1] Cooil, B., Aksoy, L., & Keiningham, T. L. (2008), 'Approaches to customer segmentation', *Journal of Relationship Marketing* 6(3-4), 9-39.
- [2] D. A l o i s e , A . D e s h p a n d e , P. H a n s e n , a n d P. P o p a t . "The Basis Of Market Segmentation" Euclidean sum-of-squares clustering." *Machine Learning*, vol. 75, pp. 245-249, 2009.
- 3 T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 881-892, 2002.
- 4Bhatnagar, Amit,Ghose, S. (2004), 'A latent class segmentation analysis of e-shoppers', *Journal of Business Research* 57, 758-767.
- 5 Marcus, C. (1998), 'A practical yet meaningful approach to customer segmentation approach to customer segmentation', *Journal of Consumer Marketing* 15, 494-504.