CUSTOMER SEGMENTATION USING K-MEANS CLUSTERING

Presented by: Ipsita Kundu

Abstract

01 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.

02

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.

03

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.

04

. 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.

Contents

01

Introduction

02

Existing method

03

Proposed method with architecture

04

Methodology

05

Implementation

06

Conclusion

Introduction

STRATIGY

- 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

- Data mining helpful to extract data from the database in a human readable format. But, we may not known the actual beneficiaries in the whole dataset.
- .Data mining helpful to extract data from the database in a human readable format. But, we may not known the actual beneficiaries in the whole dataset.

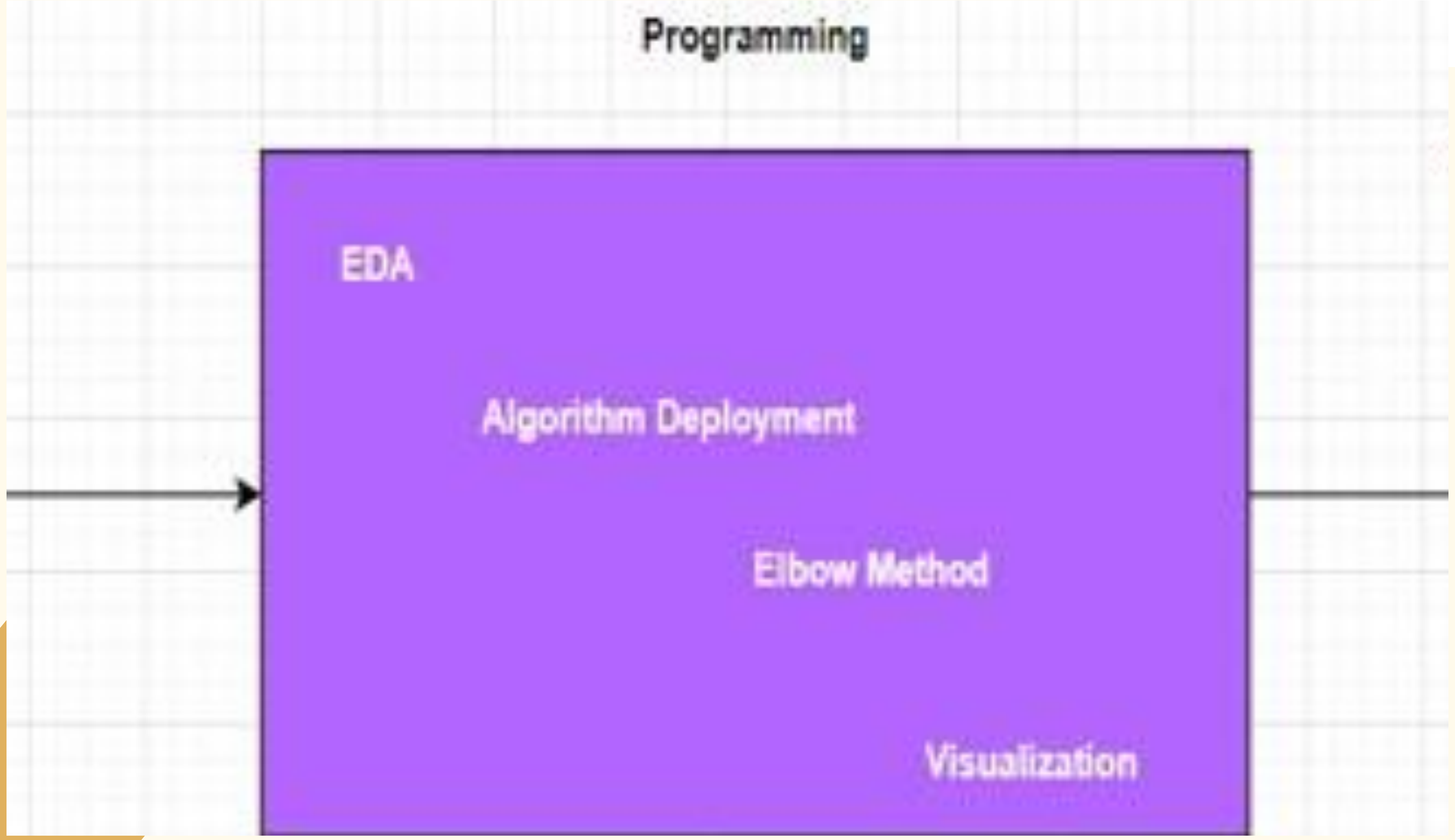
TARGET

- 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.
- 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.

Existing method

- 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.

Proposed method with architecture



01

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 po innovation which is used to predict the final outcome whic 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.

02

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.

Methodology

1.

First of all we will import all the necessary libraries or modules (pandas, numpy, seaborn).

2.

Then we will read dataset and anyalse 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.

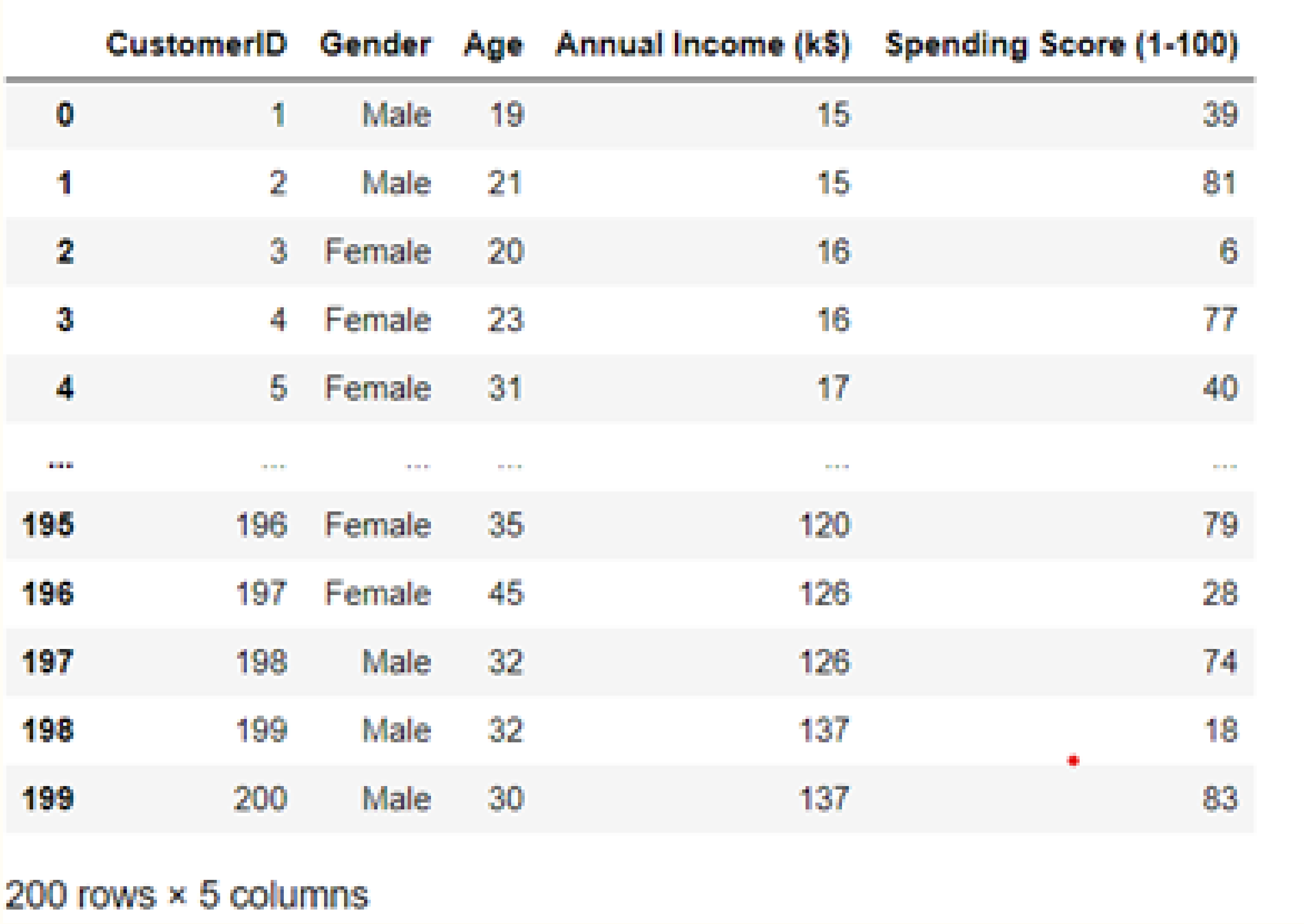
3.

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.

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.

Overview of a Dataset



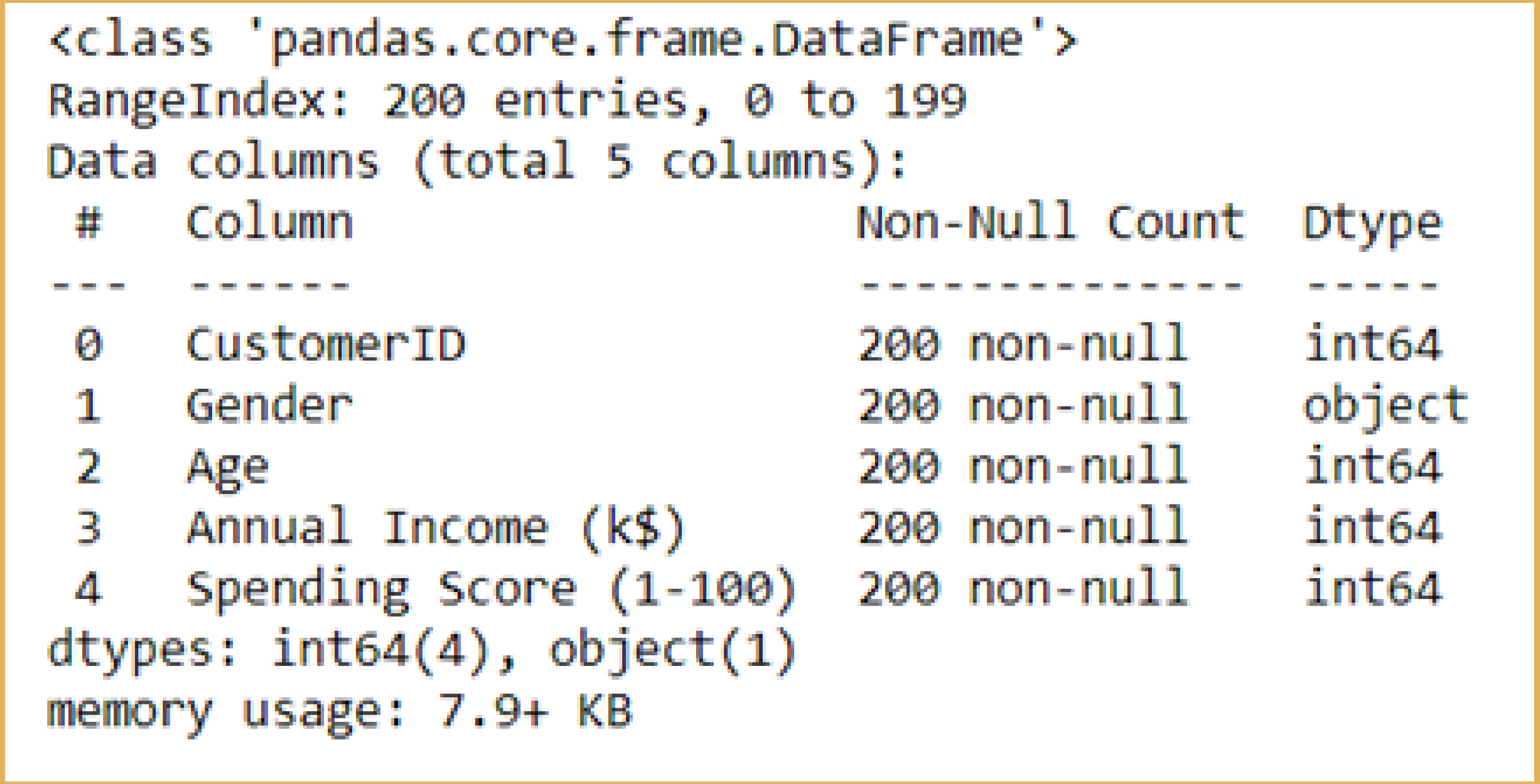
	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

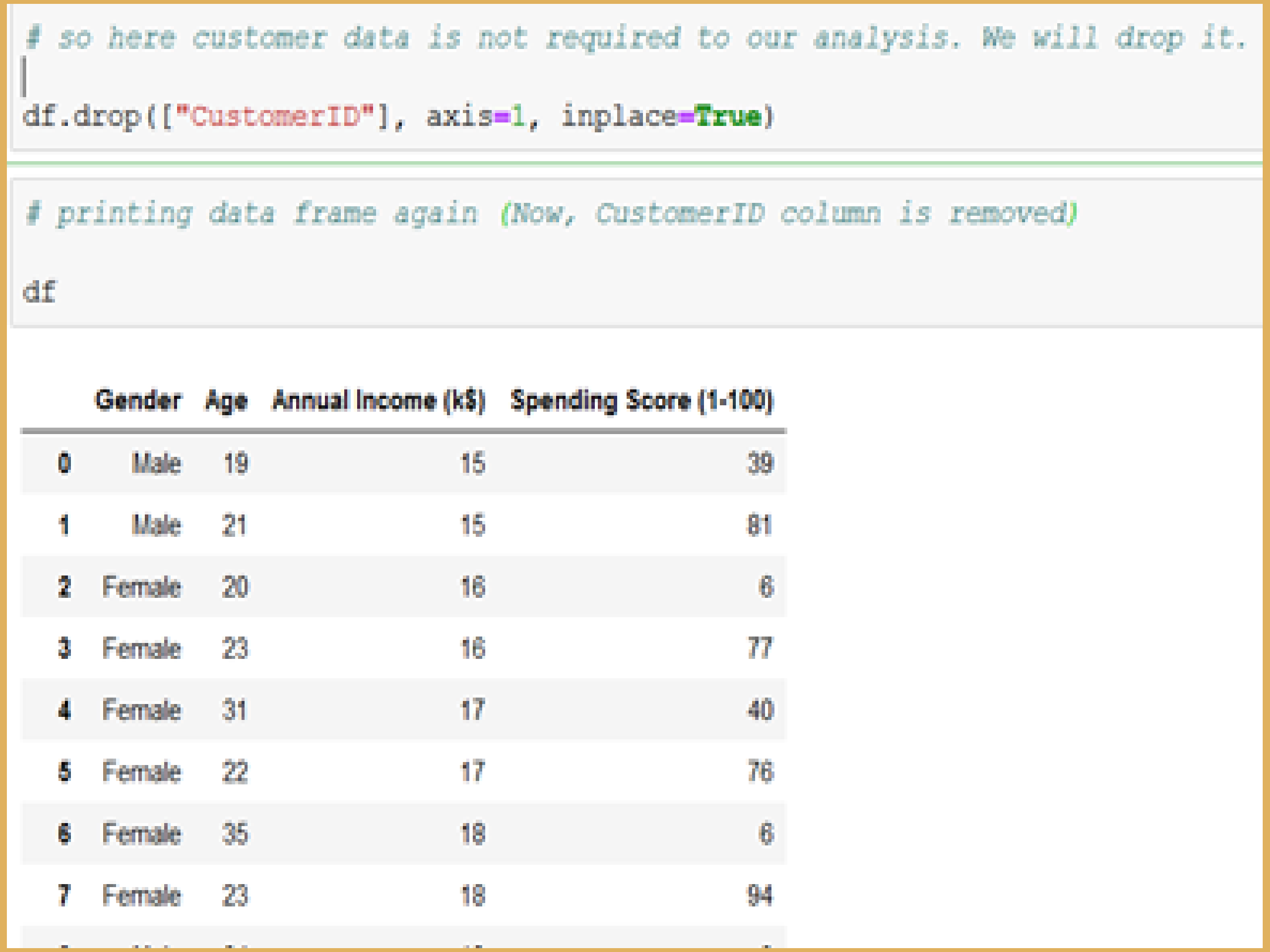
1. This is a mall customer segmentation data which contains 5 columns and 200 rows.

Information of the dataset

- #df.info()
- 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)



```
<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
```



```
# so here customer data is not required to our analysis. We will drop it.

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

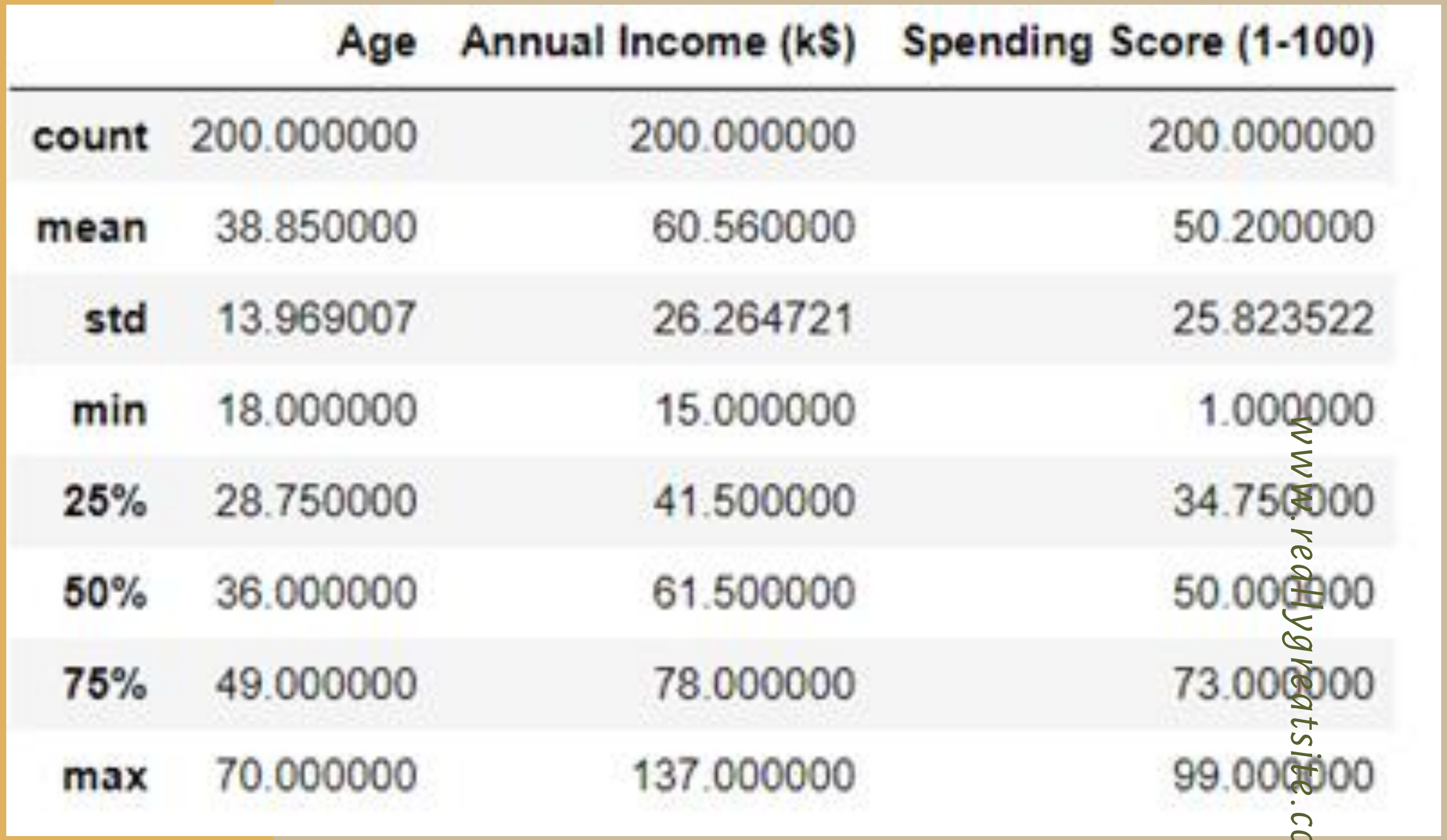
```
# printing data frame again (Now, CustomerID column is removed)

df
```

	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
5	Female	22	17	76
6	Female	35	18	6
7	Female	23	18	94

Description of the data

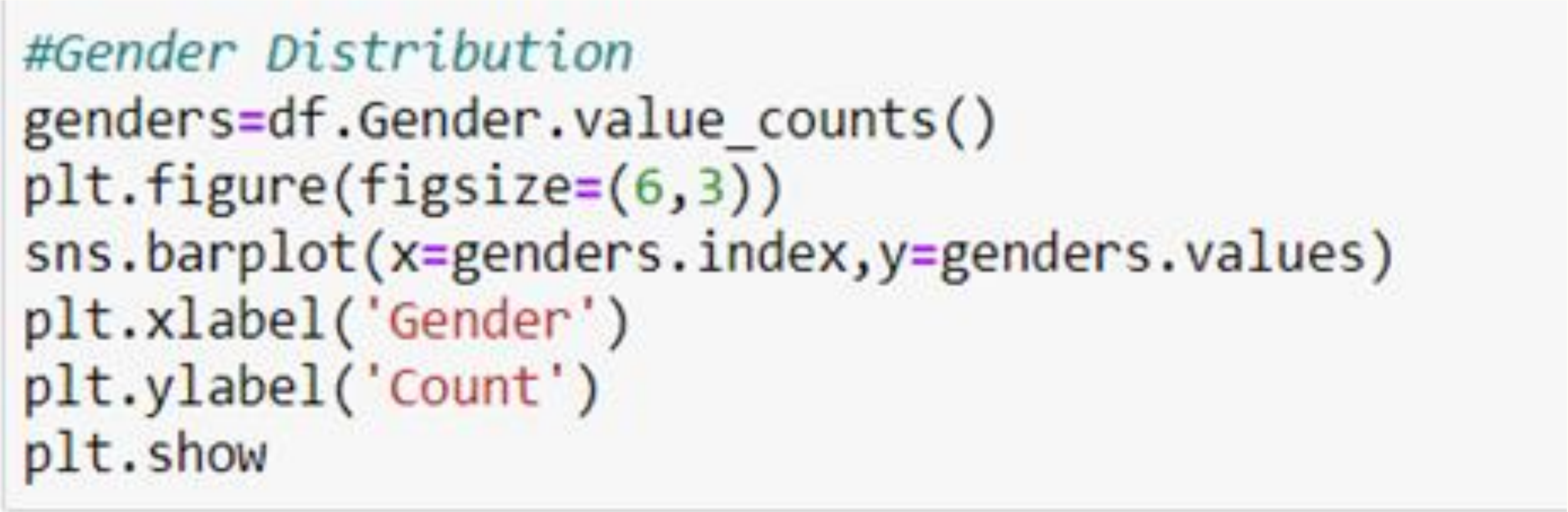
- #df.describe()
- It describes about the count which counts the no of rows in it, mean of the columns, standard deviations, maximum and minimum and percentiles etc.



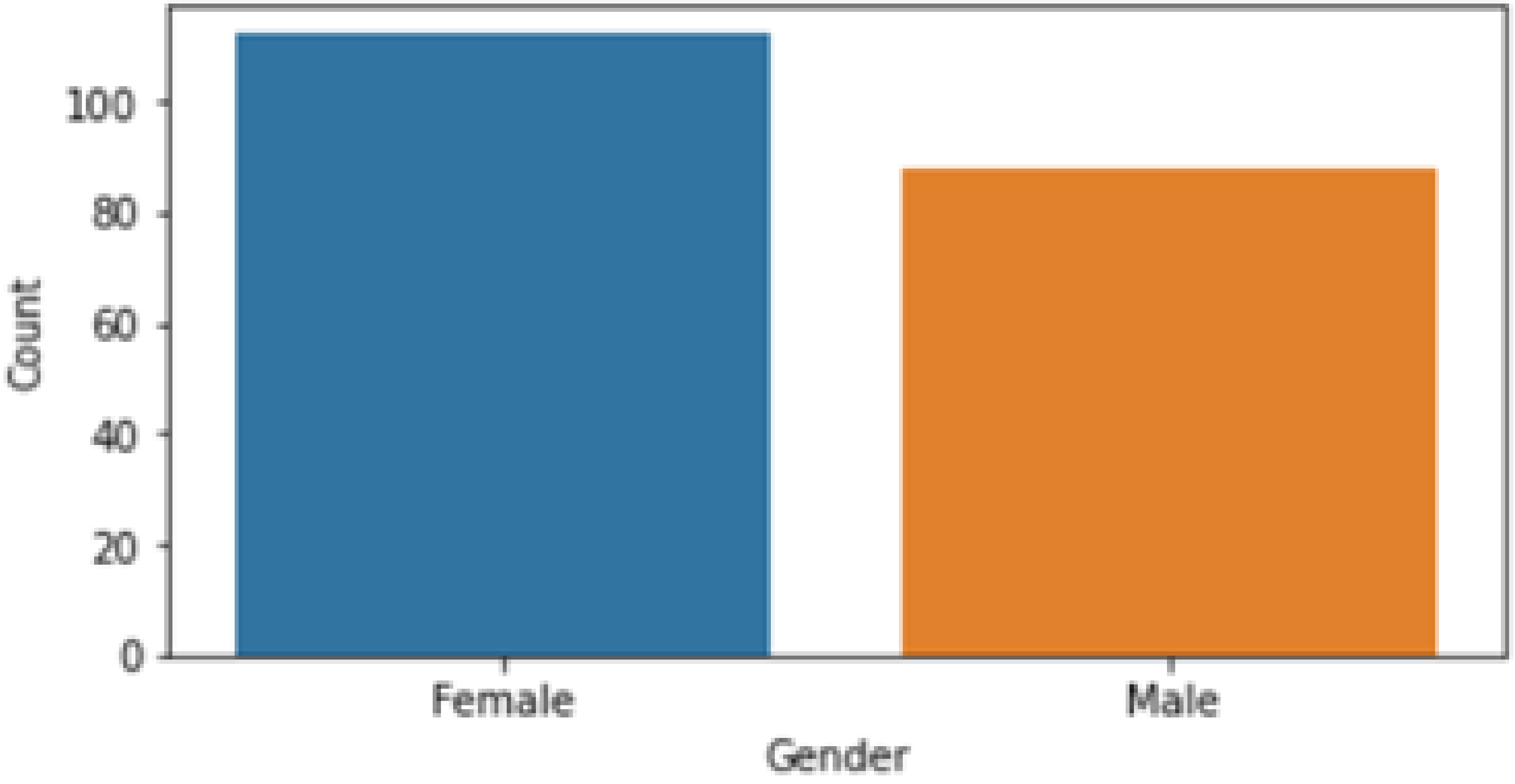
	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000
mean	38.850000	60.560000	50.200000
std	13.969007	26.264721	25.823522
min	18.000000	15.000000	1.000000
25%	28.750000	41.500000	34.750000
50%	36.000000	61.500000	50.000000
75%	49.000000	78.000000	73.000000
max	70.000000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

Gender plot Analysis

Here it overview the gender analysis



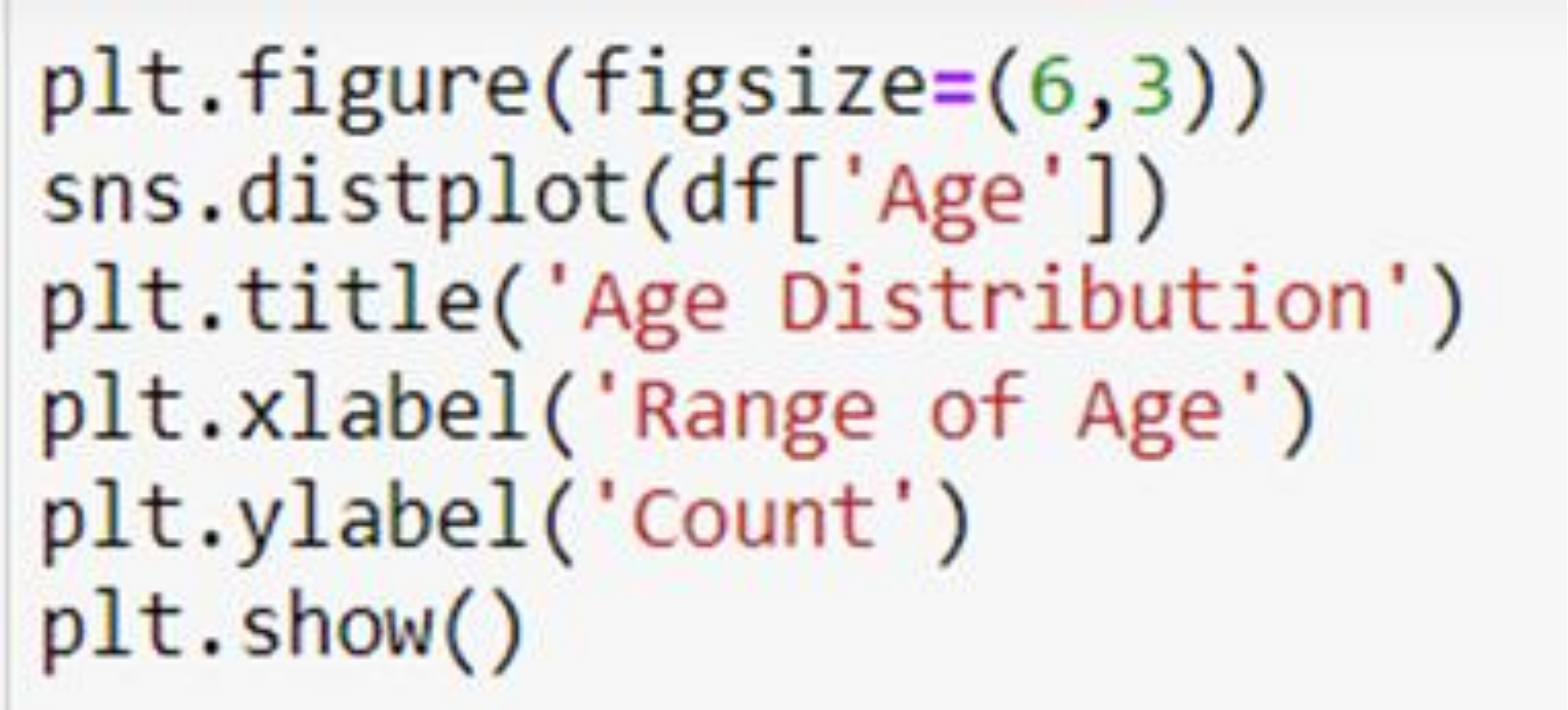
```
#Gender Distribution
genders=df.Gender.value_counts()
plt.figure(figsize=(6,3))
sns.barplot(x=genders.index,y=genders.values)
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show
```



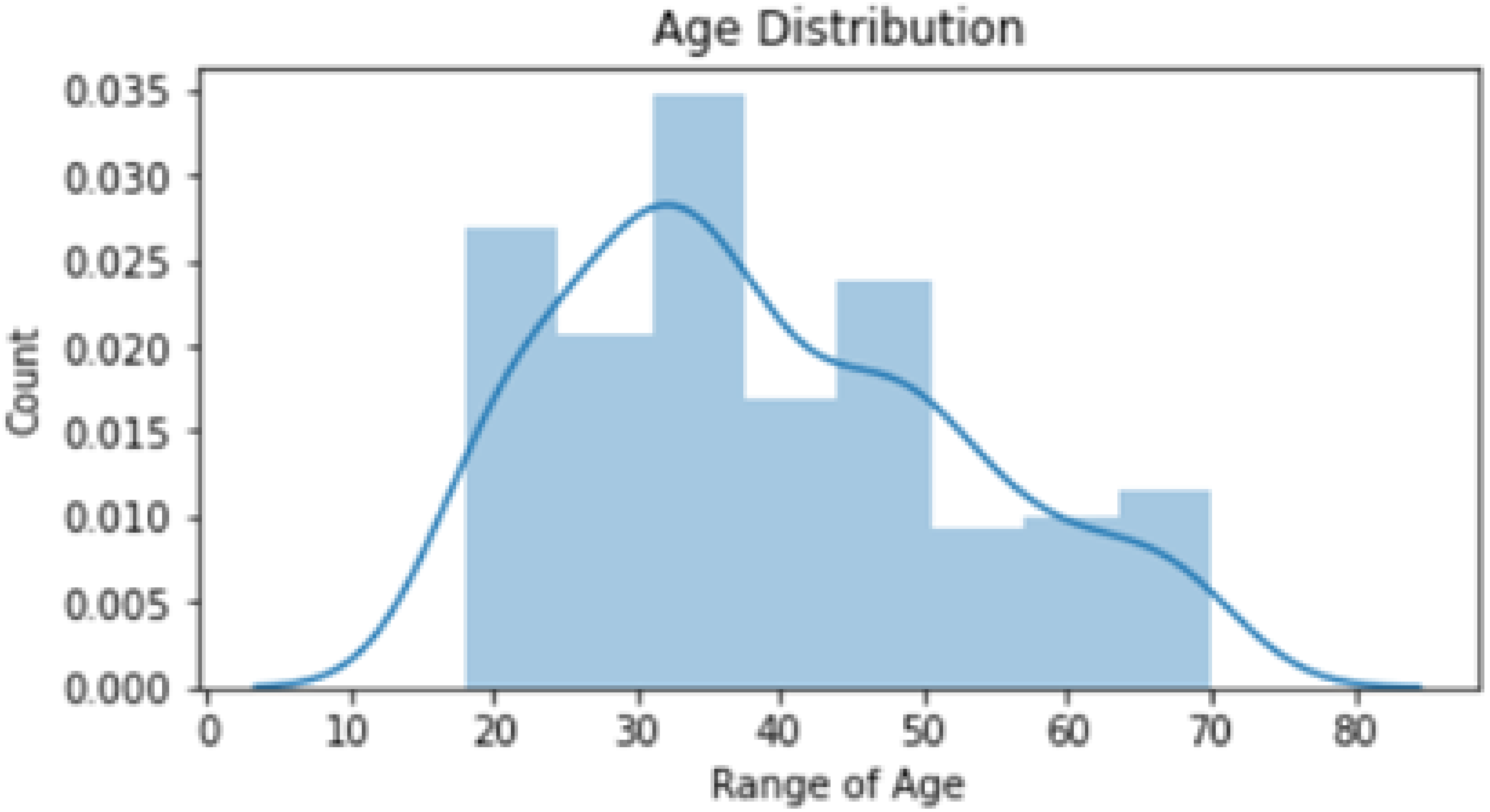
So we label the x-axis as Gender and y-axix as Count and we plot it by using barplot.
From the plot we will conclued that the there are more female customers than the male customers i.e female customers are more than 100 whereas male customers are nearly 80.

Age plot Analysis

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()
```

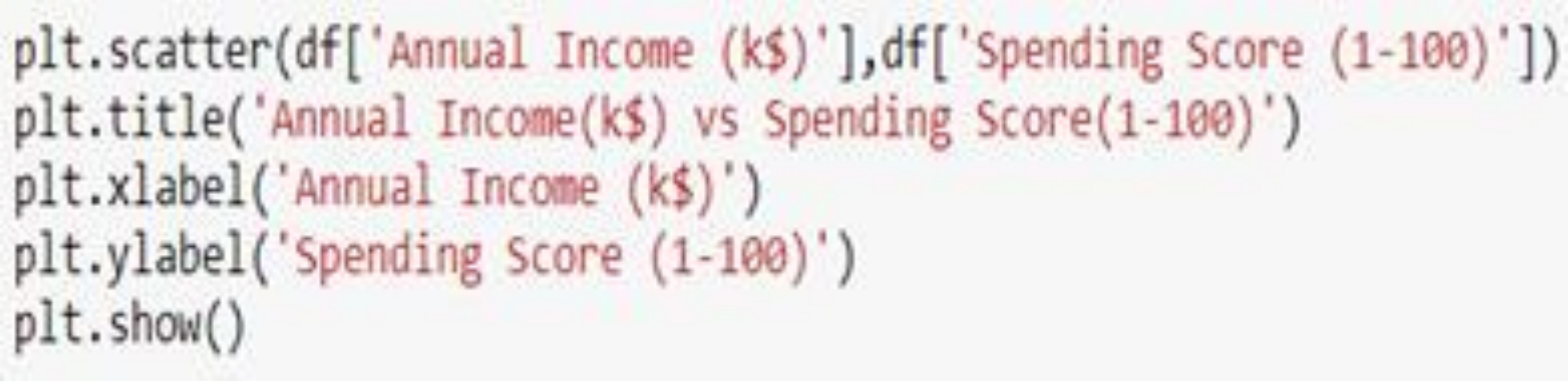


So we label X-axis as range of age and y-axis as count.

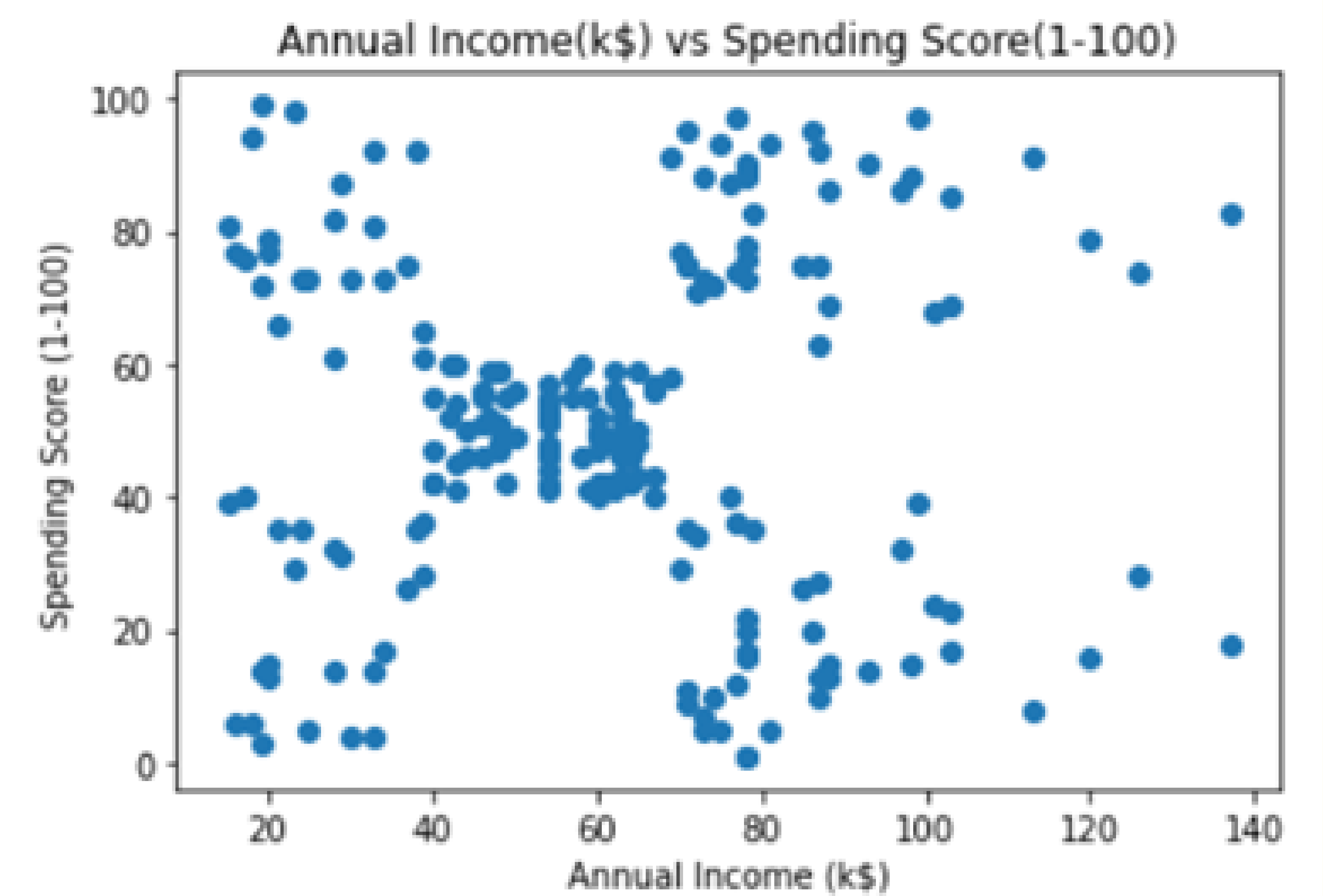
From the plot, it varies the age from nearly 20 to 70. it is evident that the age of the customers between 30 - 40 are more, then after 20-30 etc.

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.

Elbow Method

Objective 01

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.

Objective 02

This is because having more clusters allows one to capture finer groups of data objects that are more similar to each other.

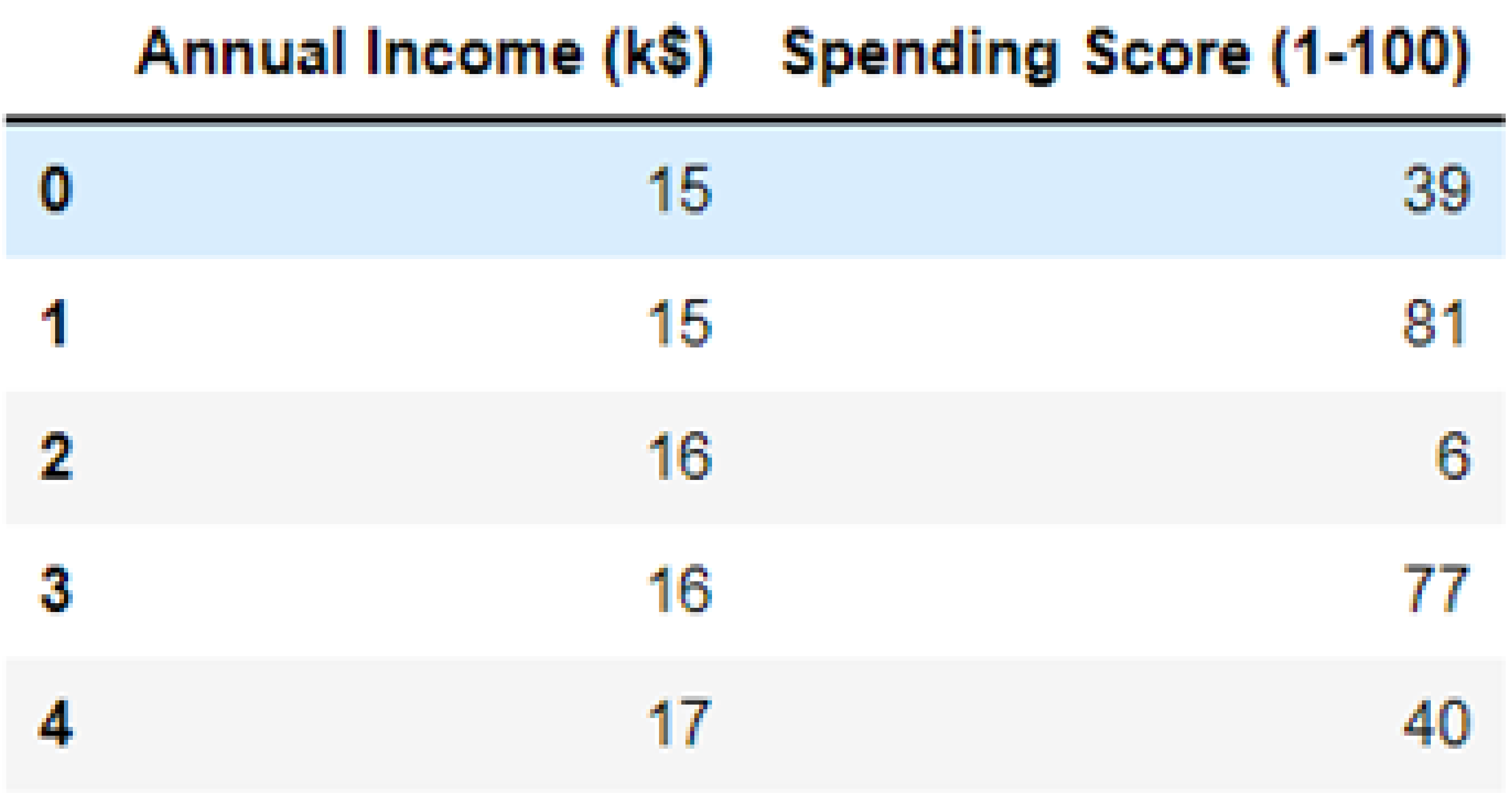
Objective 03

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.

Objective 04

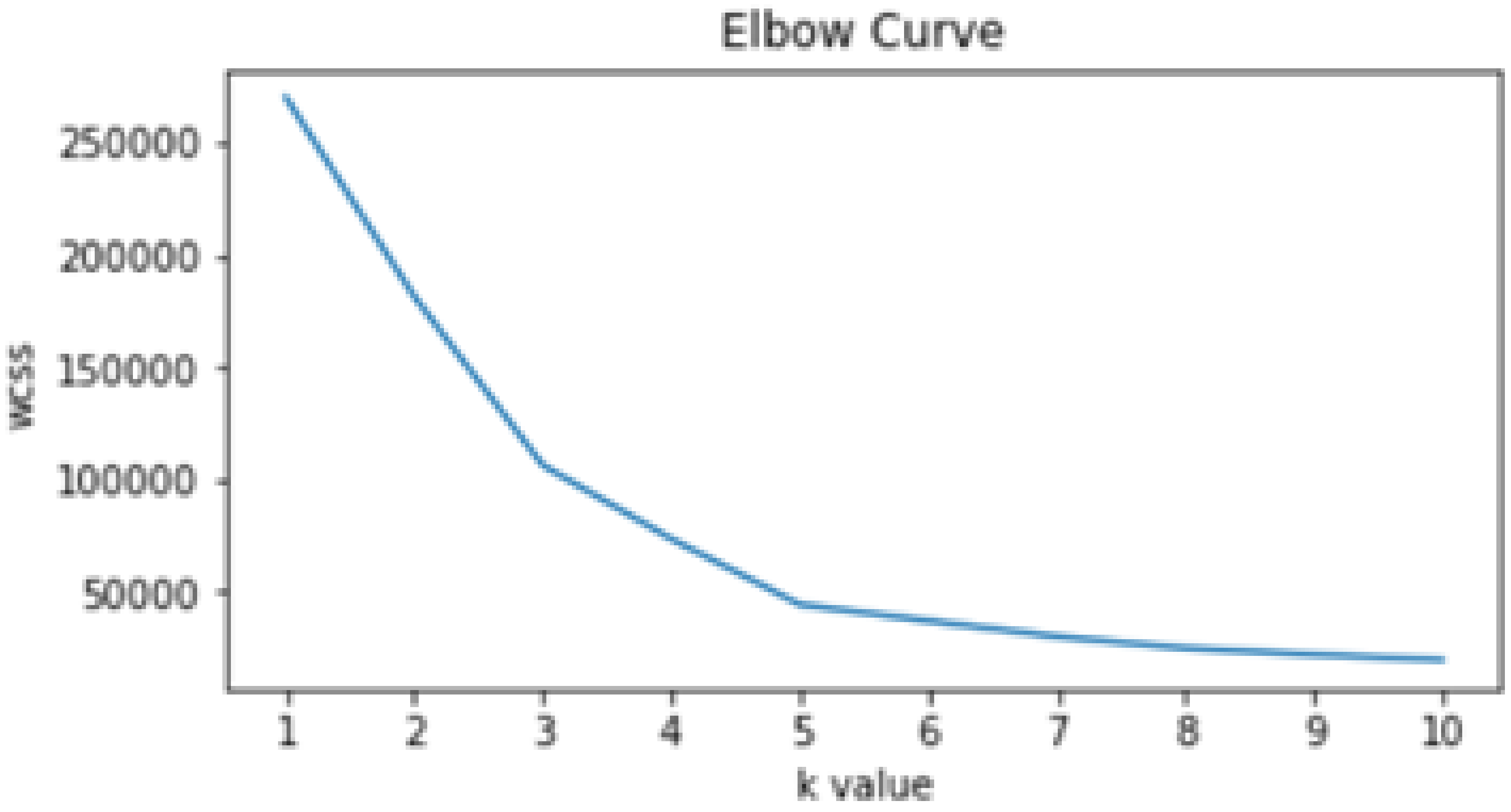
Then we calculate the total intra-cluster sum of square. Then, we proceed to plot intra-cluster sum of square based on the number of clusters. The plot denotes the approximate number of clusters required in our model. The optimum clusters can be found from the graph where there is a bend in the graph

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



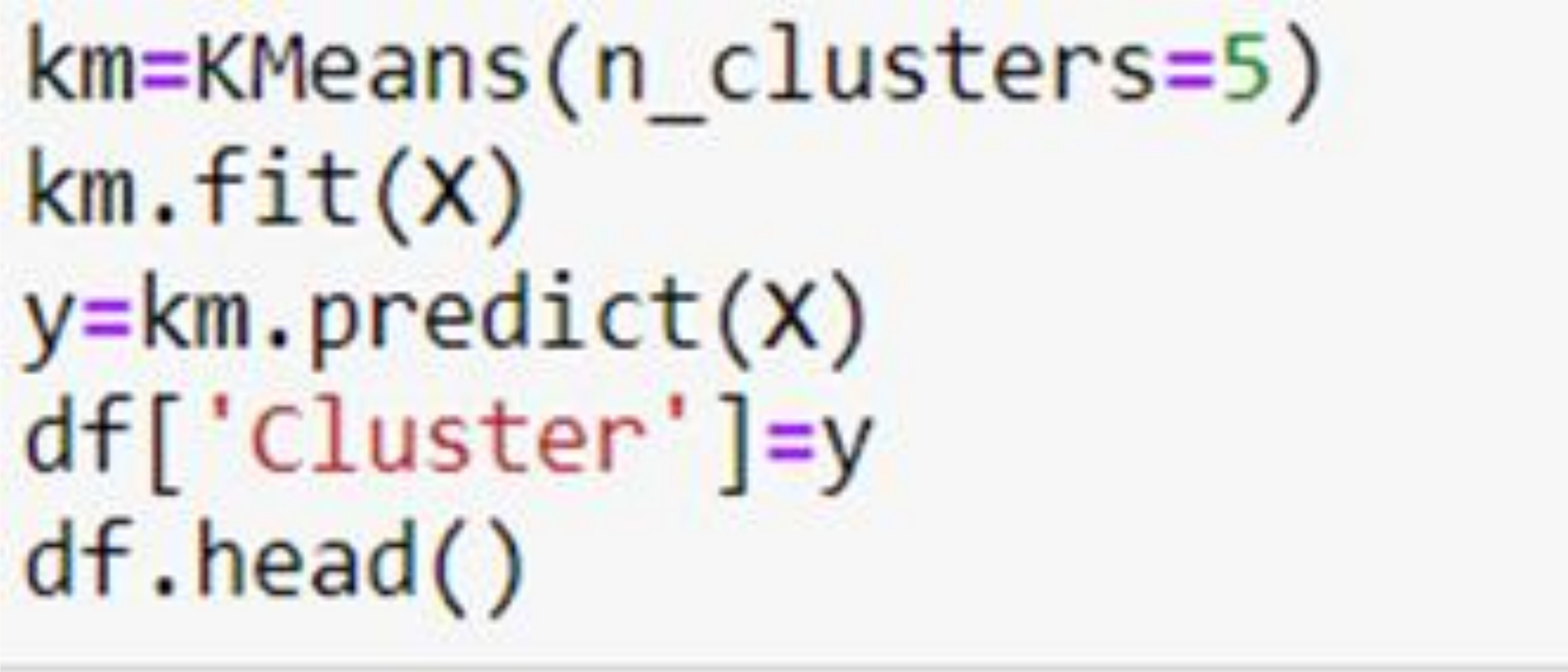
	Annual Income (k$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40

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

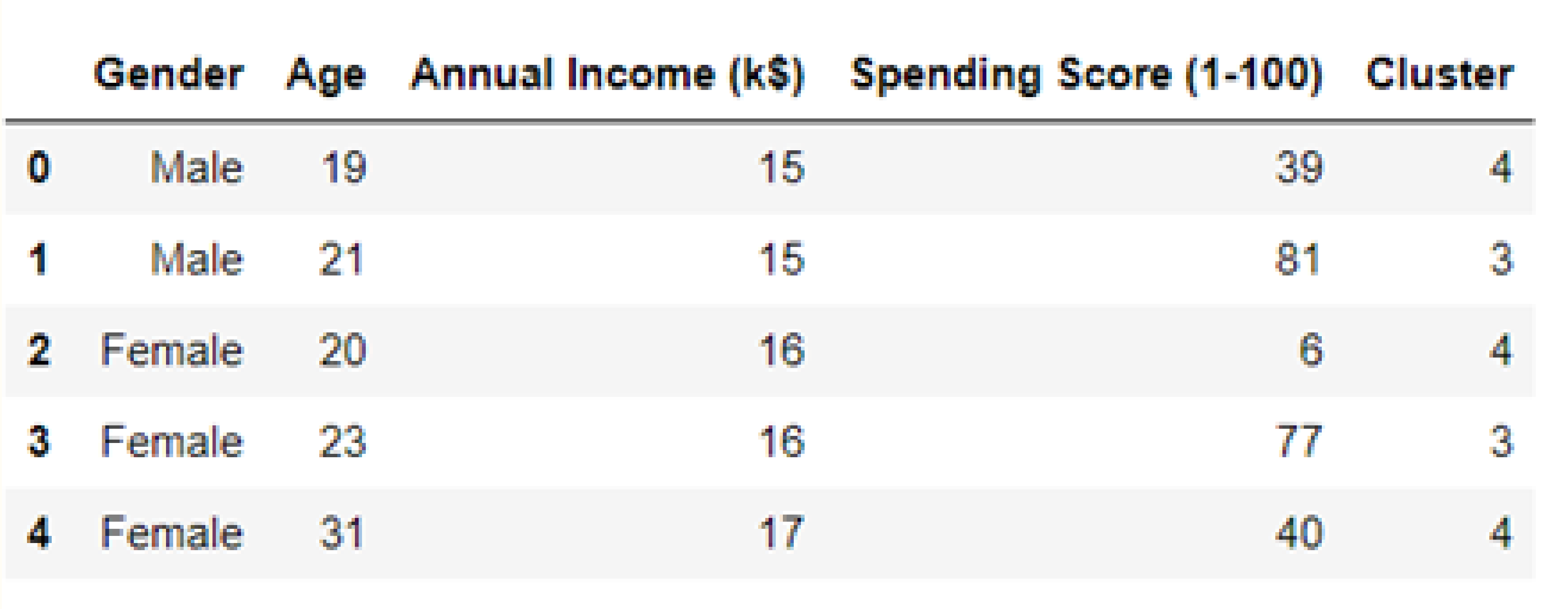


Fitting the Algorithm

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.



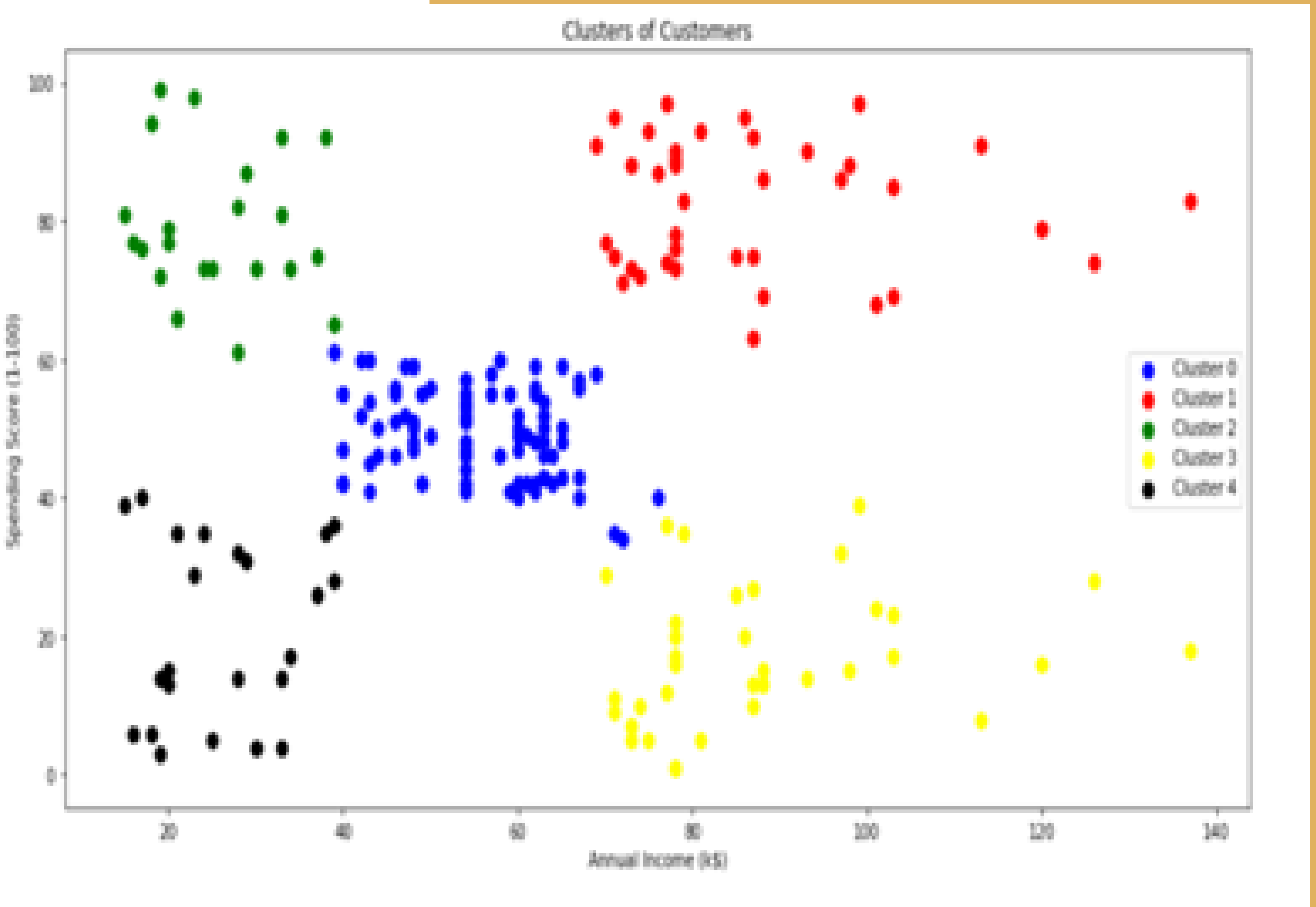
```
km=KMeans(n_clusters=5)
km.fit(X)
y=km.predict(X)
df['Cluster']=y
df.head()
```



	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	16	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.

Visualization the clusters



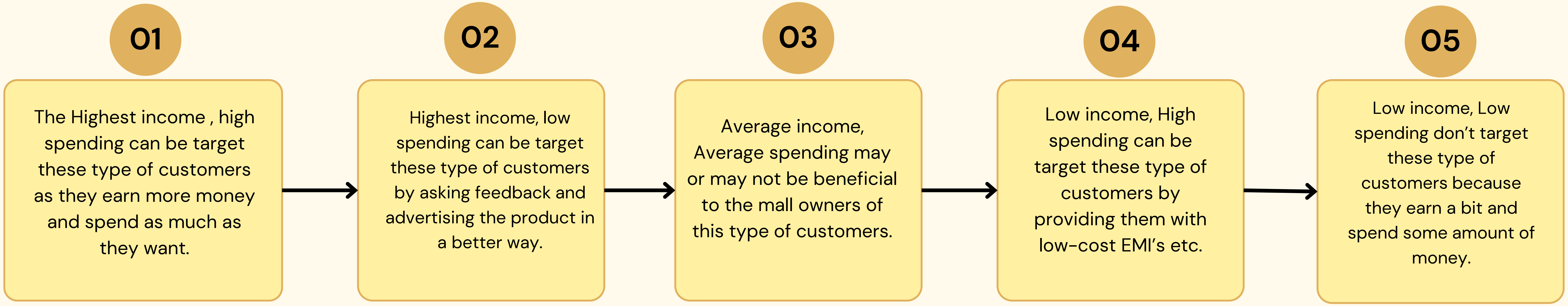
- Visualizing the clusters based on Annual Income and Spending Score of the customers. As here we plot a graph named as Clusters of Customers to visualize the data in terms of groups or cluster.

- So from the above one we observed that the 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.

Conclusion

So we concluded that the ,



01

The Highest income , high spending can be target these type of customers as they earn more money and spend as much as they want.

02

Highest income, low spending can be target these type of customers by asking feedback and advertising the product in a better way.

03

Average income, Average spending may or may not be beneficial to the mall owners of this type of customers.

04

Low income, High spending can be target these type of customers by providing them with low-cost EMI's etc.

05

Low income, Low spending don't target these type of customers because they earn a bit and spend some amount of money.

So high income, high spending are the most beneficial ones to the mall owners which increases the owner's business. (Cluster 1)