Customer Segmentation Using Data Science
Team Members

VANDNA KUMARI REG NO:-310521104123

What is customer segmentation and why does it matter?

Also known as market segmentation, customer segmentation is the division of potential customers in a given market into discrete groups. That division is based on customers having similar:

1. Needs
2. Buying characteristics

There are three main approaches to market segmentation:

1.A priori segmentation, the simplest approach, uses a classification scheme based on publicly available characteristics—such as industry and company size—to create distinct groups of customers within a market. However, a priori market segmentation may not always be valid since companies in the same industry and of the same size may have very different needs.

2.Needs-based segmentation is based on differentiated, validated drivers (needs) that customers express for a specific product or service being offered. The needs are discovered and verified through primary market research, and segments are demarcated based on those different needs rather than characteristics such as industry or company size.

3.Value-based segmentation differentiates customers by their economic value, grouping customers with the same value level into individual segments that can be distinctly targeted.

This guide will focus on the value-based approach, which allows expansion-stage companies to clearly define and target their best prospects (based on its current knowledge of the market) and satisfy most of their needs for segmentation in the expansion stage—without consuming the time and resources of a traditional, descriptive segmentation research process.

PART 1 : Analyze and Clean the dataset

- Cleaning the data
- Exploratory analysis
- Feature engineering

PART 2 : Creating customer categories

- Intermediate dataset grouped by invoices
- Final dataset grouped by customers
- K-means clustering

- Interpreting the clusters

PART 3 : Testing models for predicition

- Classic sklearn models
- Random Forest
- XGBoost

Customer Segmentation and Analysis :

Steps to solve the problem :

1. Importing Libraries.
2. Exploration of data.
3. Data Visualization.
4. Clustering using K-Means.
5. Selection of Clusters.
6. Ploting the Cluster Boundry and Clusters.
7. 3D Plot of Clusters.

Importing Libraries.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns  import
plotly as py import
plotly.graph_objs as go from
sklearn.cluster import KMeans
import warnings import os
warnings.filterwarnings("ignore")
py.offline.init_notebook_mode(connected = True)
```

Data Exploration

```
df = pd.read_csv(r'../input/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

```
df.shape
```

```
(200, 5)
```

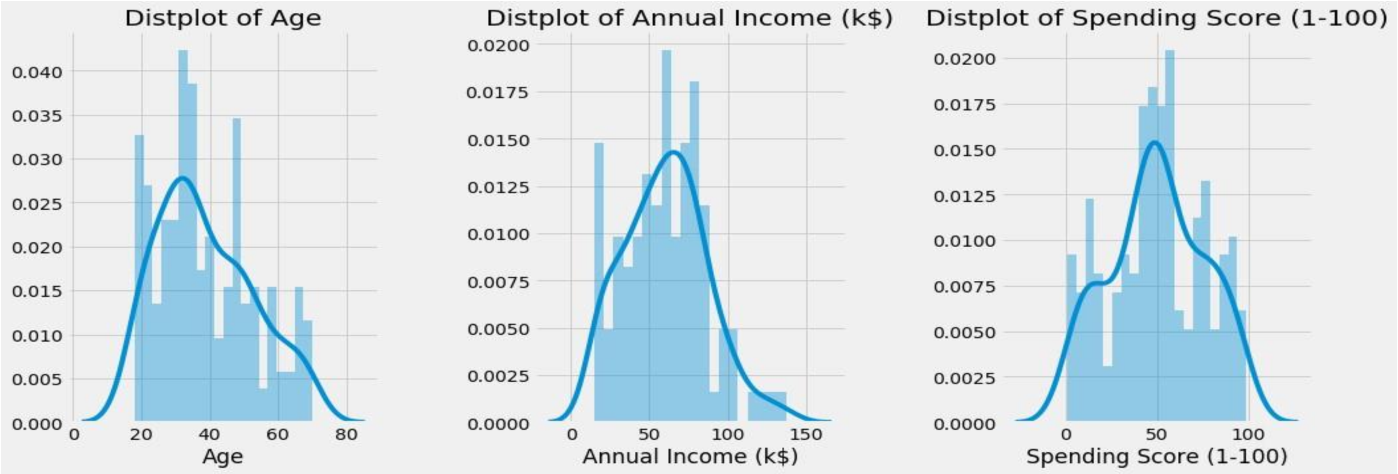
```
df.describe()
```

	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

```
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
```

Data Visualization

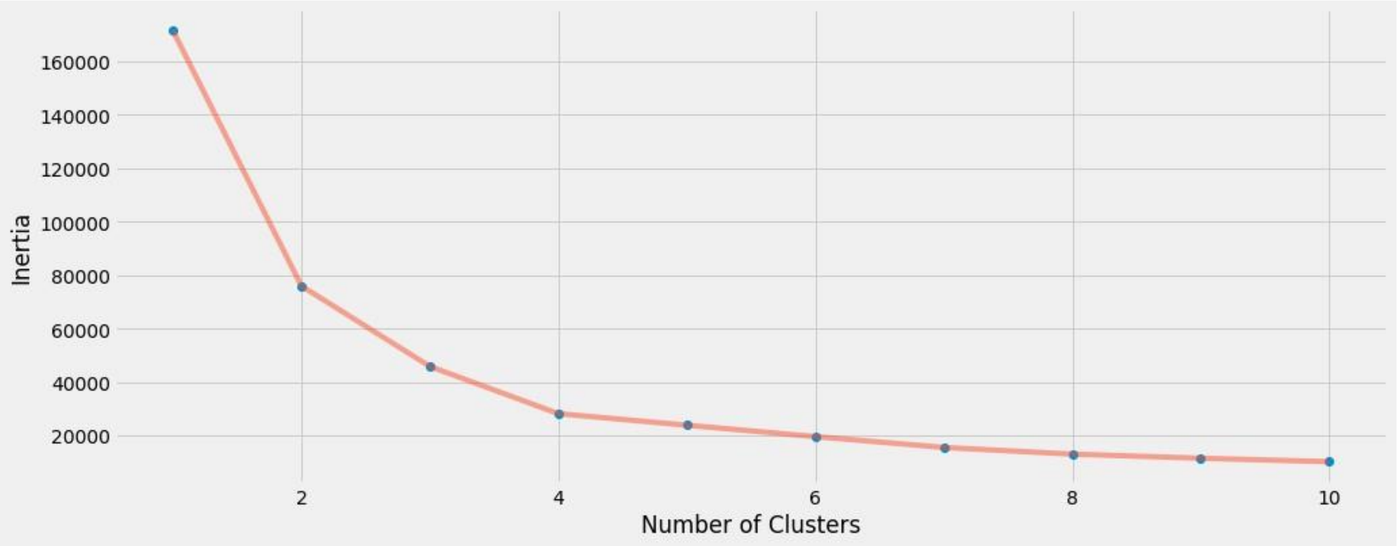


Distribution of values in Age , Annual Income and Spending Score according to Gender

Clustering using K- means

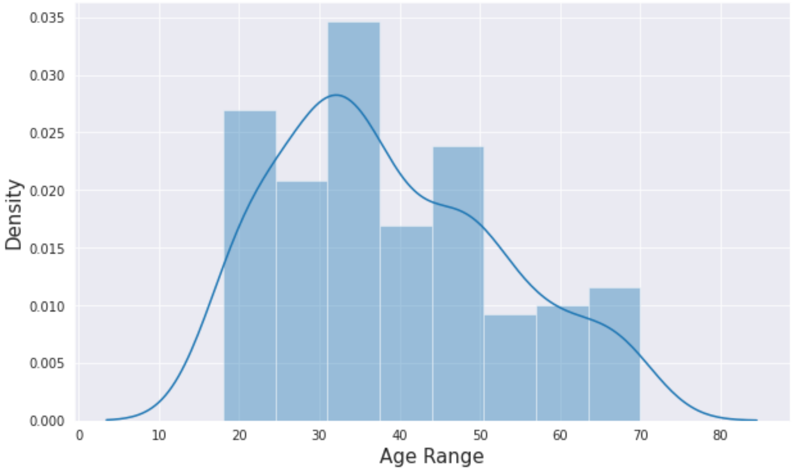
1.Segmentation using Age and

Spending Score :



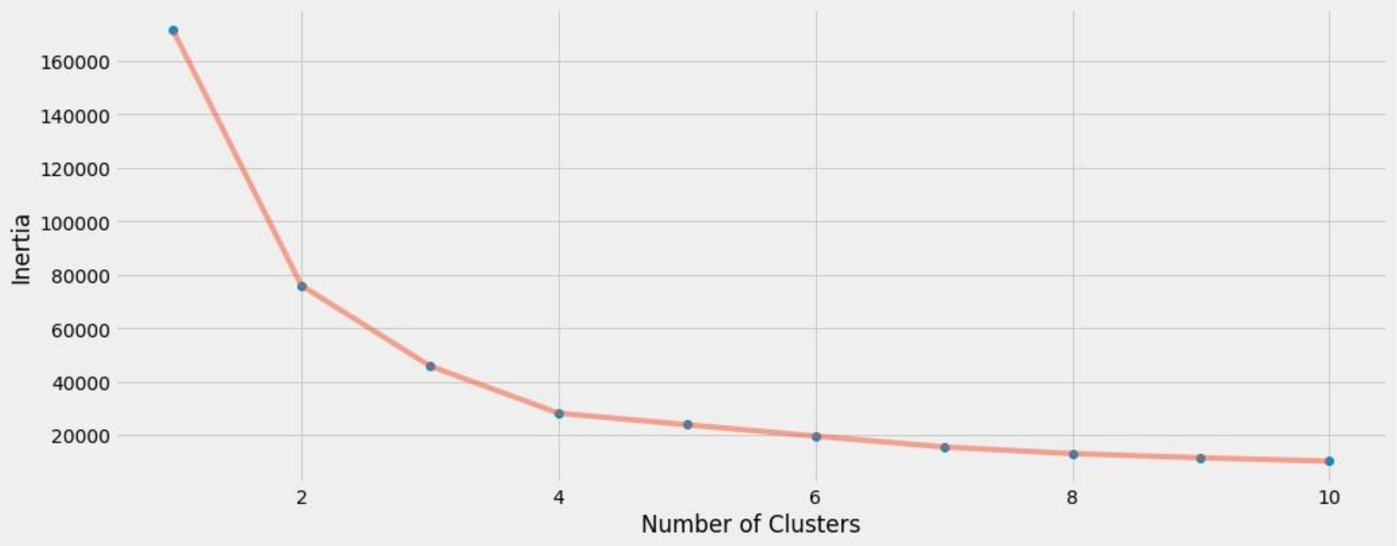
```
'''Age and spending Score'''
X1 = df[['Age' , 'Spending Score (1-100)']].iloc[: , :].values
inertia = [] for n in range(1 , 11):
    algorithm = (KMeans(n_clusters = n ,init='k-means++', n_init = 10 ,max_iter=300,
tol=0.0001,  random_state= 111  , algorithm='elkan') )     algorithm.fit(X1)
    inertia.append(algorithm.inertia_)
```

Selecting N Clusters based in Inertia (Squared Distance between Centroids and data points, should be less)



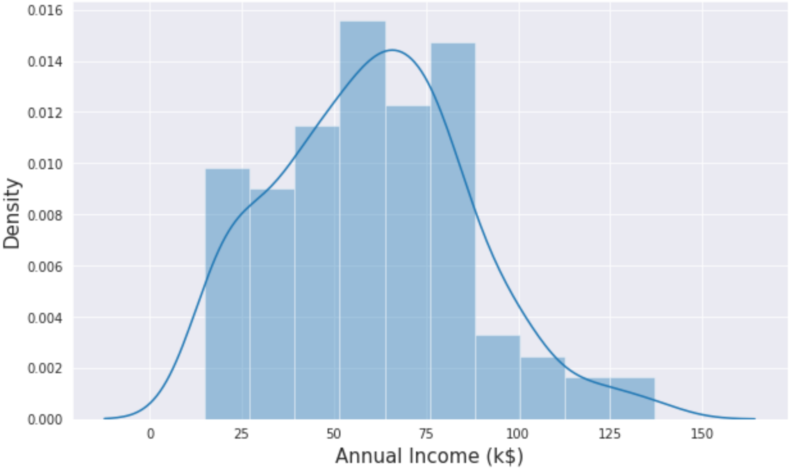
```
algorithm = (KMeans(n_clusters = 4 ,init='k-means++', n_init = 10 ,max_iter=300,
tol=0.0001,  random_state= 111  , algorithm='elkan') ) algorithm.fit(X1) labels1
= algorithm.labels_
centroids1 = algorithm.cluster_centers_
```

2.Segmentation using Annual Income and Spending Score :



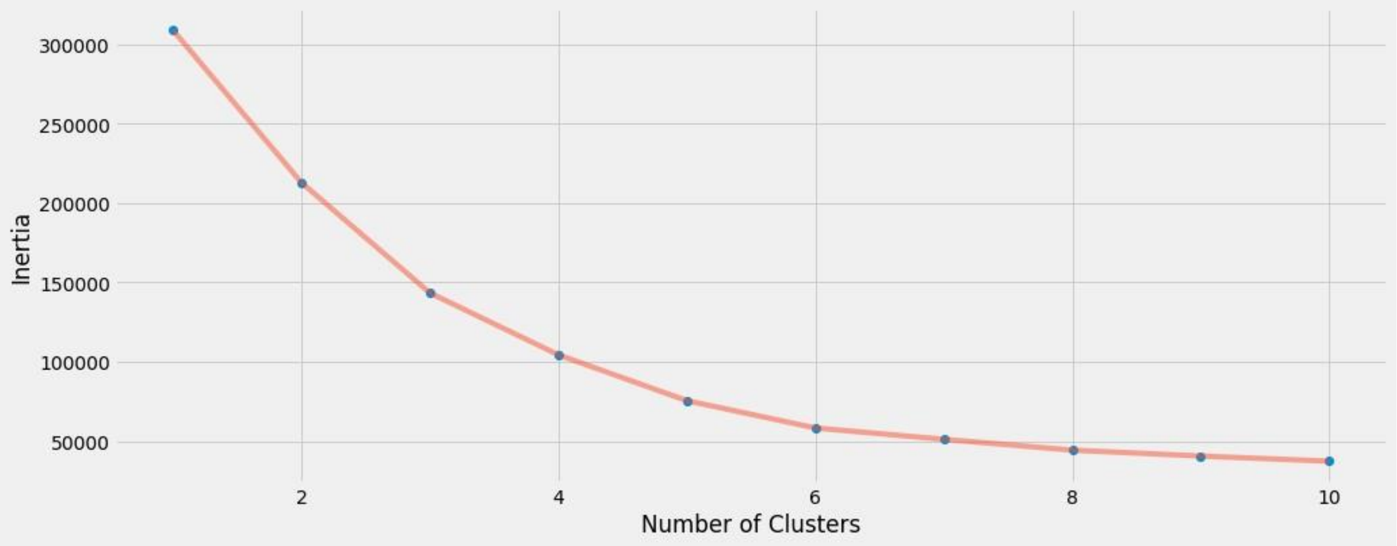
```
'''Annual Income and spending Score'''
X2 = df[['Annual Income (k$)' , 'Spending Score (1-100)']].iloc[: , :].values
inertia = [] for n in range(1 , 11):     algorithm = (KMeans(n_clusters = n
,init='k-means++', n_init = 10 ,max_iter=300,                         tol=0.0001,
random_state= 111  , algorithm='elkan') )     algorithm.fit(X2)
    inertia.append(algorithm.inertia_) In
[21]:
```

```
algorithm = (KMeans(n_clusters = 5 ,init='k-means++', n_init = 10 ,max_iter=300,
tol=0.0001,  random_state= 111  , algorithm='elkan') ) algorithm.fit(X2) labels2
= algorithm.labels_
centroids2 = algorithm.cluster_centers_
```

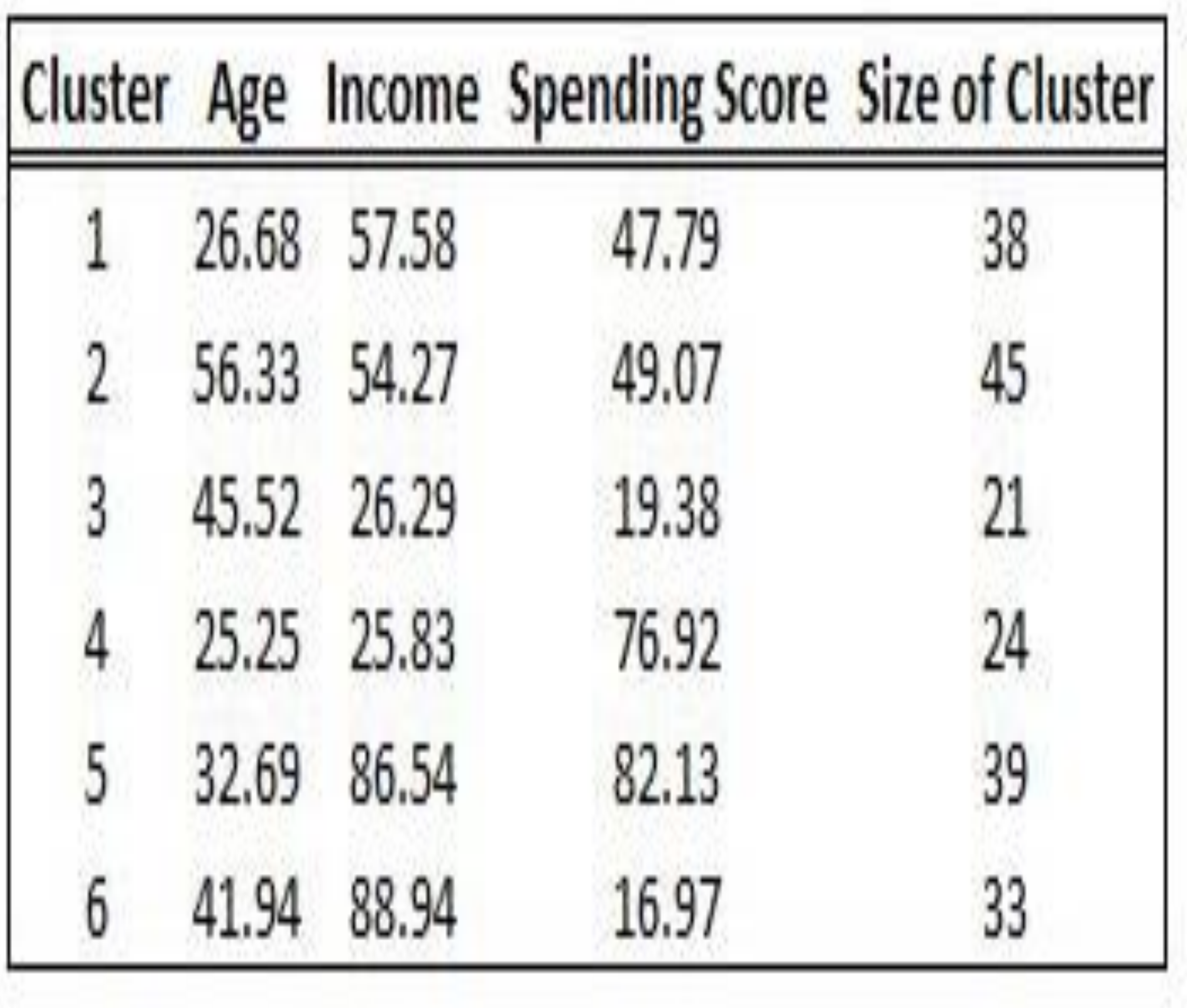


3.Segmentation using Age , Annual Income and Spending Score :

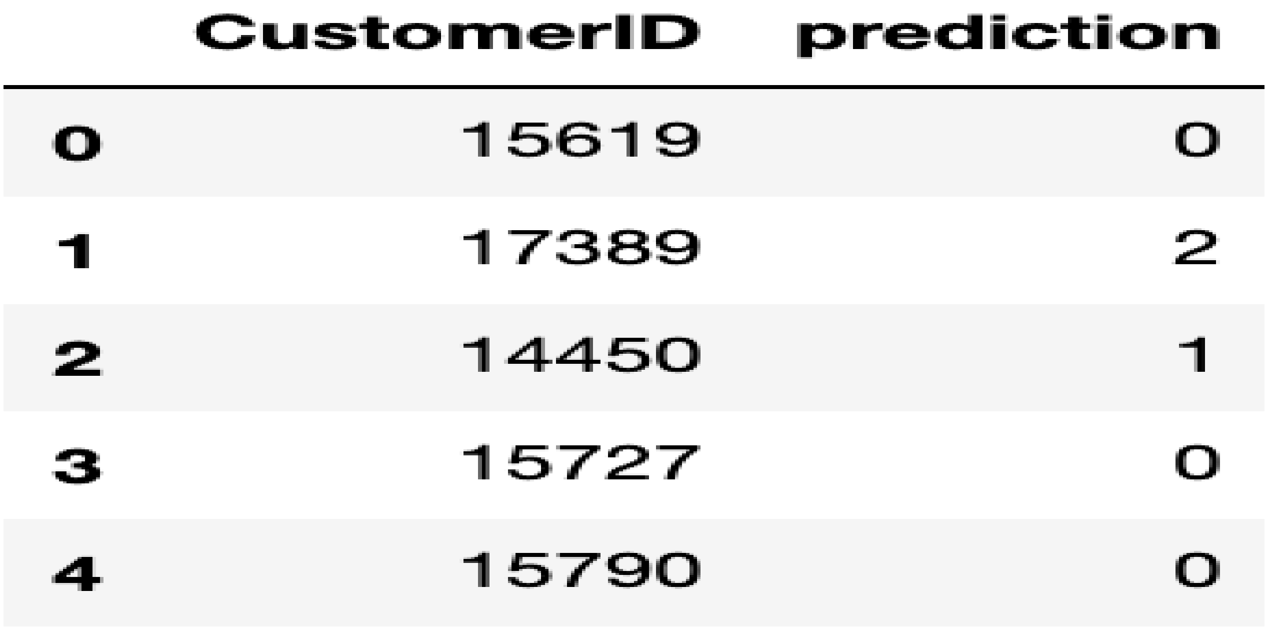
```
X3 = df[['Age' , 'Annual Income (k$)' ,'Spending Score (1-100)']].iloc[: , :].values
inertia = [] for n in range(1 , 11):
    algorithm = (KMeans(n_clusters = n ,init='k-means++', n_init = 10 ,max_iter=300,
tol=0.0001,  random_state= 111  , algorithm='elkan') )     algorithm.fit(X3)
    inertia.append(algorithm.inertia_)
```



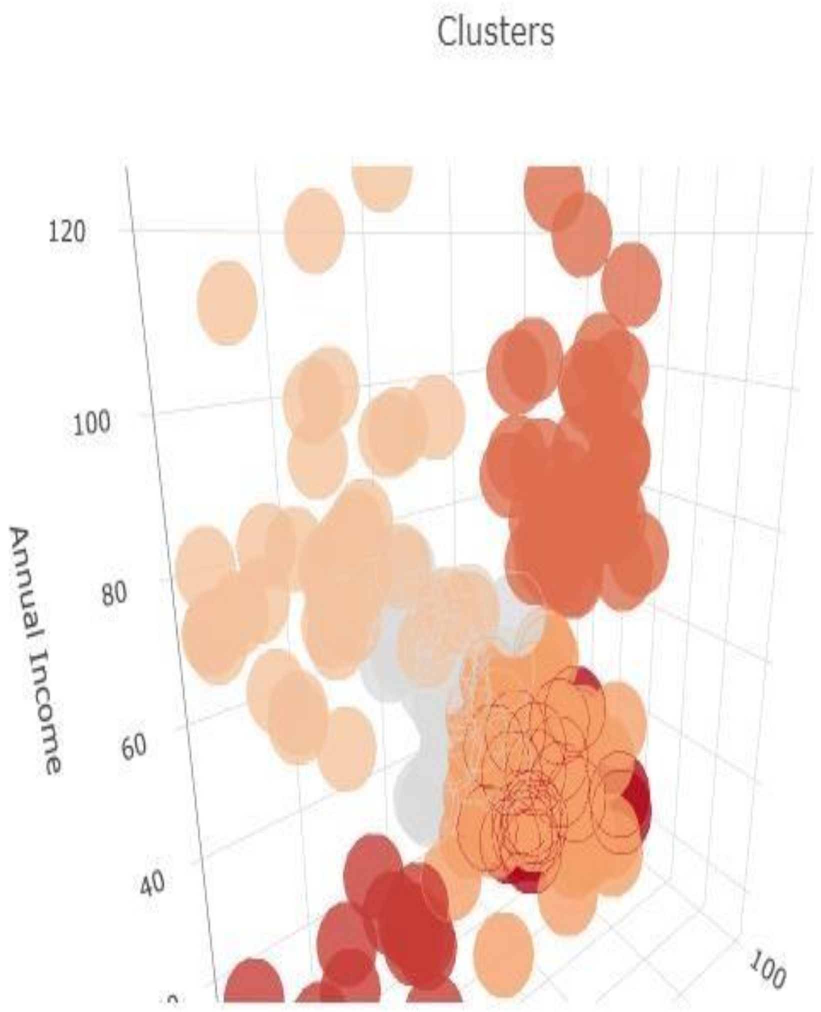
```
algorithm = (KMeans(n_clusters = 6 ,init='k-means++', n_init = 10 ,max_iter=300,
tol=0.0001,  random_state= 111  , algorithm='elkan') ) algorithm.fit(X3) labels3
```

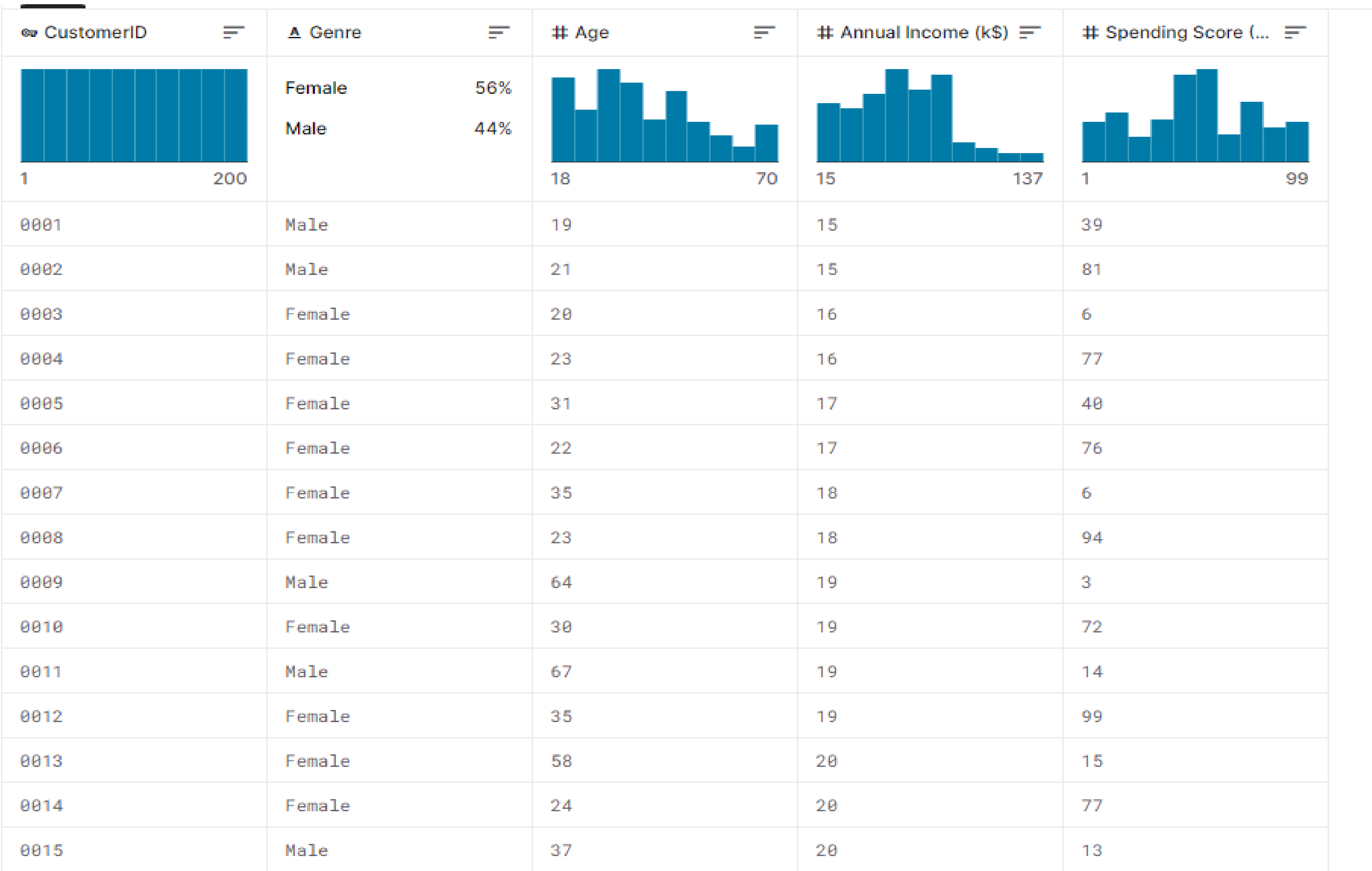


Cluster	Age	Income	Spending Score	Size of Cluster
1	26.68	57.58	47.79	38
2	56.33	54.27	49.07	45
3	45.52	26.29	19.38	21
4	25.25	25.83	76.92	24
5	32.69	86.54	82.13	39
6	41.94	88.94	16.97	33

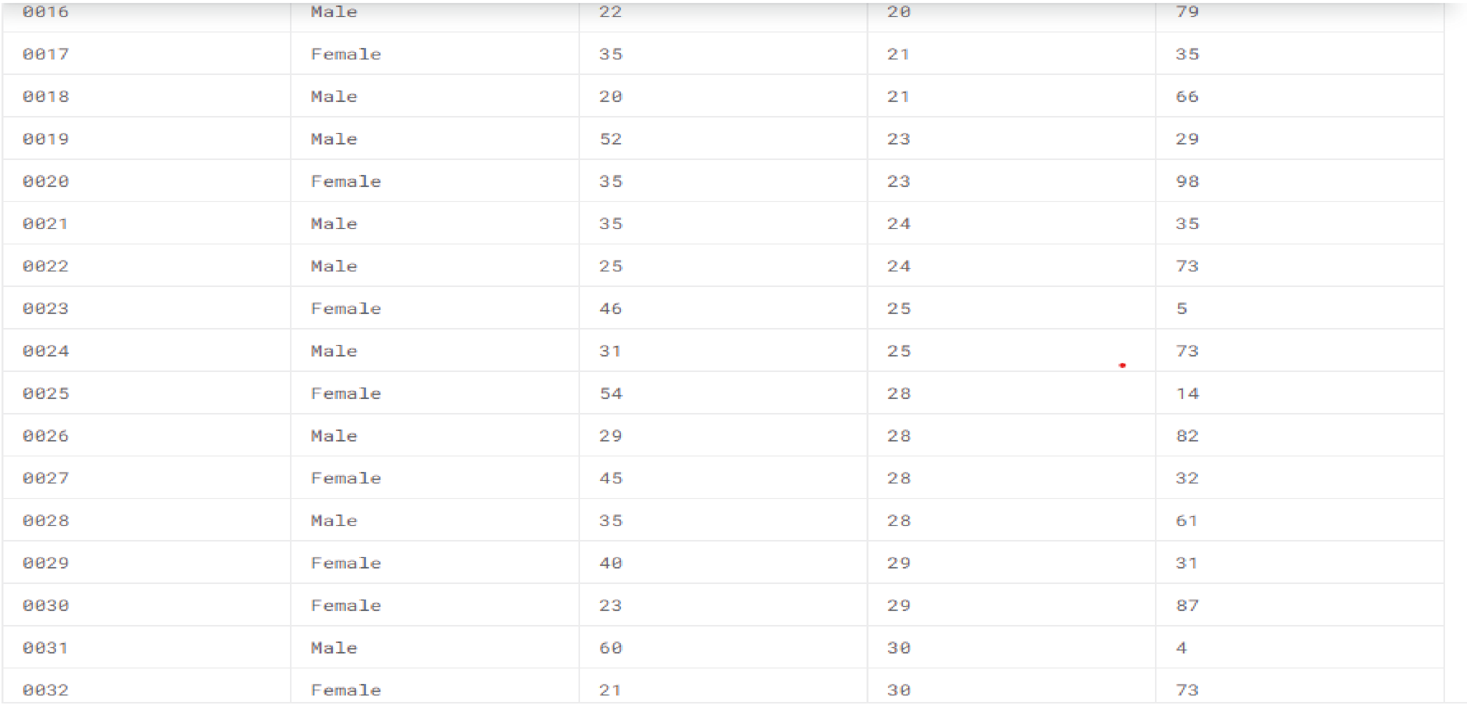


	CustomerID	prediction
0	15619	0
1	17389	2
2	14450	1
3	15727	0
4	15790	0

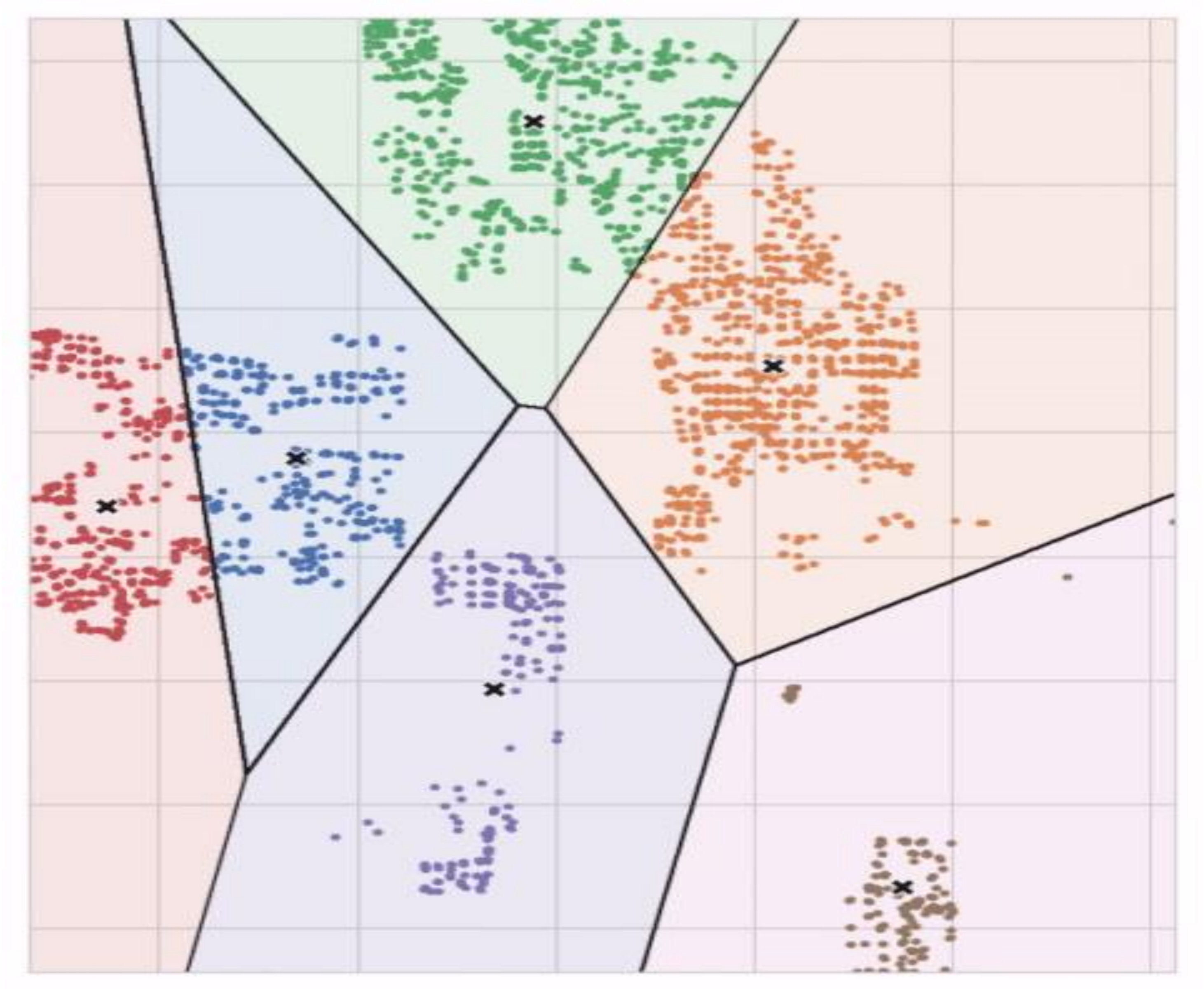




CustomerID	Genre	Age	Annual Income (k$)	Spending Score (...
1 200	Female 56% Male 44%	18 70	15 137	1 99
0001	Male	19	15	39
0002	Male	21	15	81
0003	Female	20	16	6
0004	Female	23	16	77
0005	Female	31	17	40
0006	Female	22	17	76
0007	Female	35	18	6
0008	Female	23	18	94
0009	Male	64	19	3
0010	Female	30	19	72
0011	Male	67	19	14
0012	Female	35	19	99
0013	Female	58	20	15
0014	Female	24	20	77
0015	Male	37	20	13



0016	Male	22	20	79
0017	Female	35	21	35
0018	Male	20	21	66
0019	Male	52	23	29
0020	Female	35	23	98
0021	Male	35	24	35
0022	Male	25	24	73
0023	Female	46	25	5
0024	Male	31	25	73
0025	Female	54	28	14
0026	Male	29	28	82
0027	Female	45	28	32
0028	Male	35	28	61
0029	Female	40	29	31
0030	Female	23	29	87
0031	Male	60	30	4
0032	Female	21	30	73



Conclusion

Customer segmentation is essential. Machine learning can get control over the complete process. Discovering all of the different groups that build up a more meaningful customer base permits you to get into customers’ brains and give them precisely what they crave, enhancing their participation and expanding profits.