```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

Import all necessary libraries such as numpy, pandas, matplotlib and seaborn.

```
df=pd.read_csv("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

Import data set of Mall customer and name as df.

df.head() gives a top 5 data set. Here data set consist of CustomerID, Gender, Age, Annual Income, and Spending Score.

```
df.shape
```

```
(200, 5)
```

Data set consist of 200 row and 5 coloumn.

```
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.describe() give the sumarray of data set with mean, standard deviation, min, 25 percentile, 50 percentile, 75 percentile and max value. Here mean value of age,annual income and spending score is 38.85, 60.56 and 50.20 respectively.

From the above, the average income of 200 members is 60.56k dollars, with a maximum income of 137k dollars and a minimum income of 15k dollars. In addition, the customer's majority income is around 78k dollars.

```
df.dtypes
```

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

df.dtypes gave an datatypes of data set. Customer ID, Age, Annual Income and Spending Score are integer types. The Gender is of object data types consists of male and female data.

```
# to check data set has any null value.
df.isnull().sum()
```

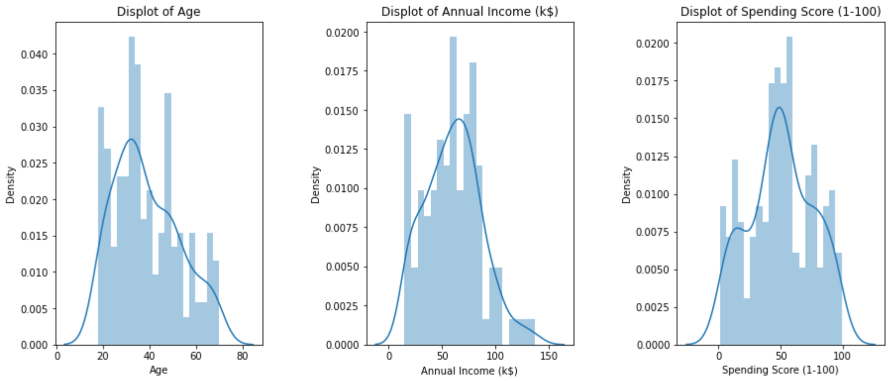
```
CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

There is no null data in dataset.

```
# to drop cusotmer ID coloumn as it not required.
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

```
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('Displot of {}'.format(x))
plt.show()
```

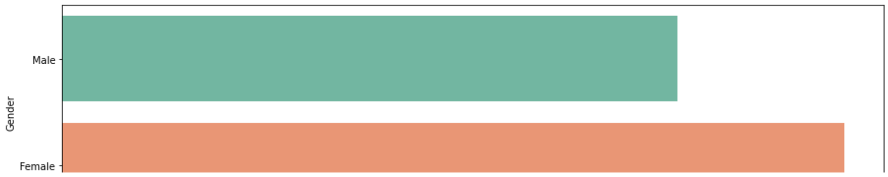


Here draw the displot of age, annual income and spending score. Age group around 20 to 40 has maximum density means data set of age consists of more that between 20 to 40 and peak at 30.

Similary Annual income has maximum density between 50 to 100 k. The large amount of cusotmer has annual income between 50 to 70 k.

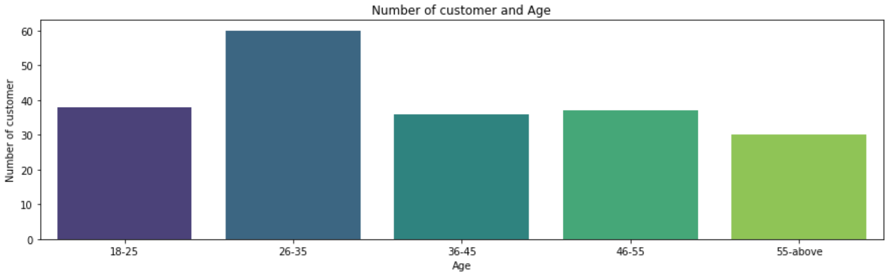
The most people has spending score of 50.

```
plt.figure(figsize=(15,4))
sns.countplot(y='Gender',data=df,palette='Set2')
plt.show()
```



The female customer has more membership card and purchase more with compare to male cusotmer.

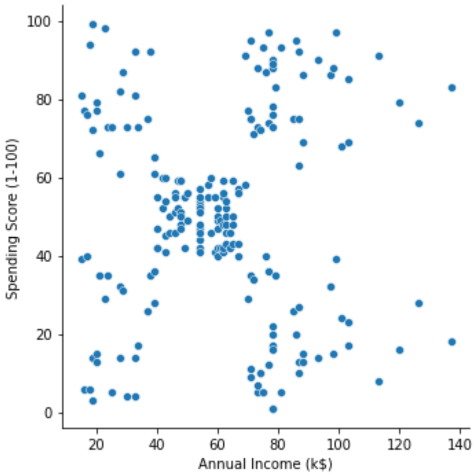
```
age_18_25=df.Age[(df.Age>=18)& (df.Age<=25)]
age_26_35=df.Age[(df.Age>=26)& (df.Age<=35)]
age_36_45=df.Age[(df.Age>=36)& (df.Age<=45)]
age_46_55=df.Age[(df.Age>=46)& (df.Age<=55)]
age_55_above=df.Age[(df.Age>=55)]

agex=["18-25","26-35","36-45","46-55","55-above"]
agey=[len(age_18_25.values),len(age_26_35.values),len(age_36_45.values),len(age_46_55.values),len(age_55_above.values)]

plt.figure(figsize=(15,4))
sns.barplot(x=agex,y=agey,palette="viridis")
plt.title("Number of customer and Age")
plt.xlabel("Age")
plt.ylabel("Number of customer")
plt.show()
```



From the above bar graph it is clear that the customer of age between 26-35 has more number of customer followed by age group of 46 to 55 and then age group between 36 to 45.

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

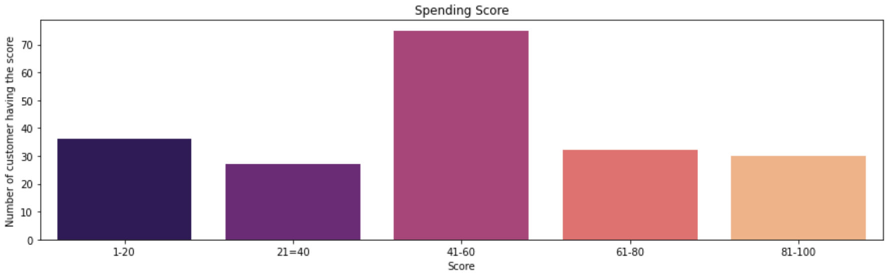
```
<seaborn.axisgrid.FacetGrid at 0x7fdfcabe3e20>
```



There is not much relation between Annual income and spending score but their is some relation i.e, more concentration of data point between annual income in between 40 to 60 K$ which lies between spending score of 40 to 60.

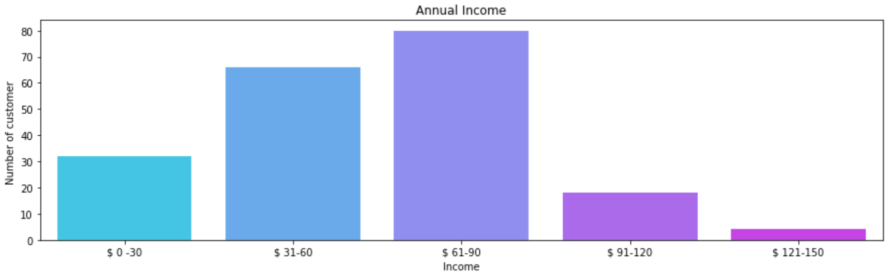
```
ss_1_20=df["Spending Score (1-100)"][(df["Spending Score (1-100)"]>=1)& (df["Spending Score (1-100)"]<=20)]
ss_21_40=df["Spending Score (1-100)"][(df["Spending Score (1-100)"]>=21)& (df["Spending Score (1-100)"]<=40)]
ss_41_60=df["Spending Score (1-100)"][(df["Spending Score (1-100)"]>=41)& (df["Spending 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=(15,4))
sns.barplot(x=ssx,y=ssy,palette="magma")
plt.title("Spending Score")
plt.xlabel("Score")
plt.ylabel("Number of customer having the score")
plt.show()
```



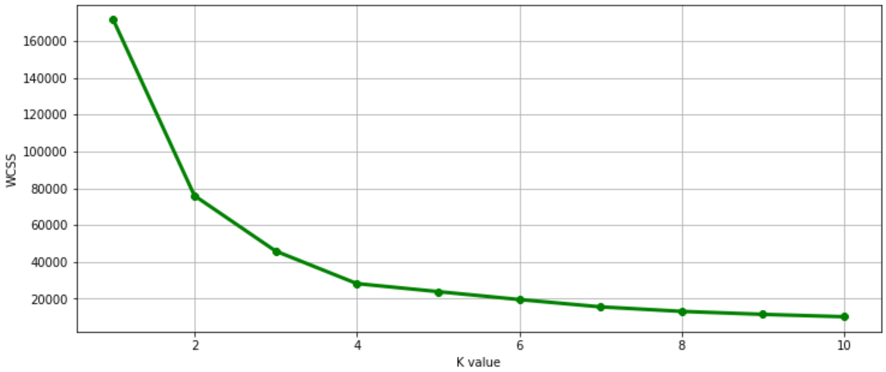
Majority of customer spending scores ranging from 41-60.

```
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","$ 31-60","$ 61-90","$ 91-120","$ 121-150"]
aiy=[len(ai0_30.values),len(ai31_60.values),len(ai61_90.values),len(ai91_120.values),len(ai121_150.values)]

plt.figure(figsize=(15,4))
sns.barplot(x=aix,y=aiy,palette="cool")
plt.title("Annual Income")
plt.xlabel("Income")
plt.ylabel("Number of customer ")
plt.show()
```



The most of customer has annual income between $61 to 90k.

```
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=(12,5))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=3,color="green",marker="8")
plt.xlabel("K value")
plt.ylabel("WCSS")
plt.show()
```



As we seen in the graph the bend is at k=4 afterward it is straight. So the optimum value of cluster is 4, hence we will proceed further with k=4.

```
kmeans=KMeans(n_clusters=4)
#predict the label of cluster. Here we are finding cluster based on age and spending score.
label=kmeans.fit_predict(x1)
print(label)
```

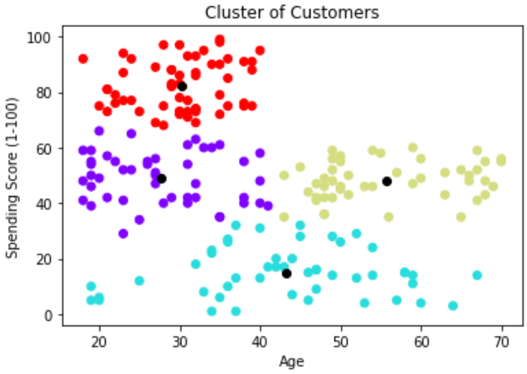
```
[0 3 1 3 0 3 1 3 1 3 1 3 1 3 1 3 0 0 1 3 0 3 1 3 1 3 1 0 1 3 1 3 1 3 1 3 1
 3 1 3 2 3 2 0 1 0 2 0 0 0 2 0 0 2 2 2 2 2 0 2 2 0 2 2 2 0 2 2 0 0 2 2 2 2
 2 0 2 0 0 2 2 0 2 2 0 2 2 0 0 2 2 0 2 0 0 0 2 0 2 0 0 2 2 0 2 0 2 2 2 2 2
 0 0 0 0 0 2 2 2 2 0 0 0 3 0 3 2 3 1 3 1 3 0 3 1 3 1 3 1 3 1 3 0 3 1 3 2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

Here we see that data is divided into 0,1,2,3 as cluster is four.

```
# to find out centroid with x and y cordinate
print(kmeans.cluster_centers_)
```

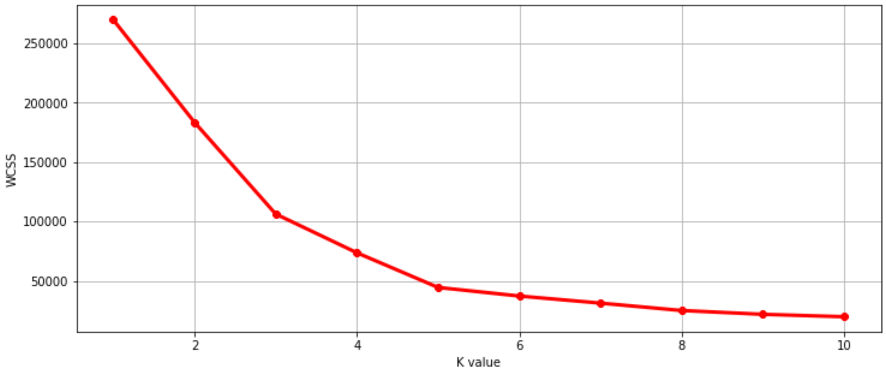
```
[[27.61702128 49.14893617]
 [43.29166667 15.02083333]
 [55.70833333 48.22916667]
 [30.1754386  82.35087719]]
```

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



Here we see the 4 cluster where whole data is divided, and black dot represent centroid of each of those cluster.

```
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=(12,5))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=3,color="red",marker="8")
plt.xlabel("K value")
plt.ylabel("WCSS")
plt.show()
```



As we seen in the graph the bend is at k=5 afterward it is straight. So the optimum value of cluster is 5, hence we will proceed further with k=5.

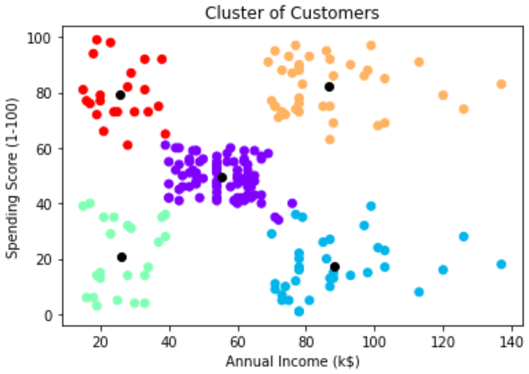
```
kmeans=KMeans(n_clusters=5)
#predict the label of cluster. Here we are finding cluster based on age and spending score.
label=kmeans.fit_predict(x2)
print(label)
```

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

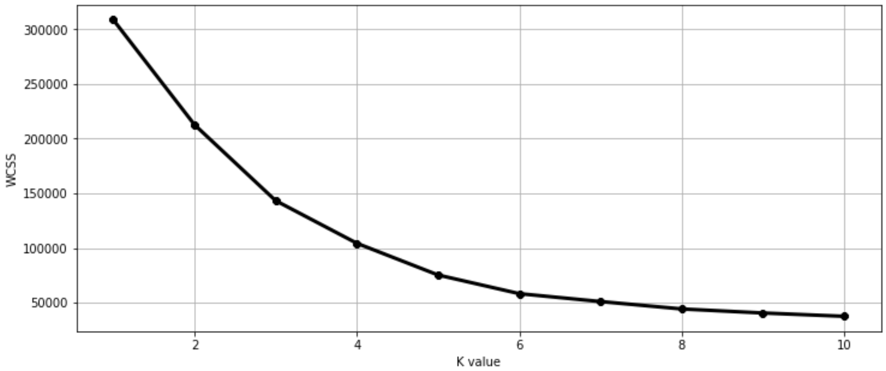
```
# to find out centroid with x and y cordinate
print(kmeans.cluster_centers_)
```

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

```
plt.scatter(x2[:,0],x2[:,1],c=kmeans.labels_,cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],color='black')
plt.title('Cluster 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=(12,5))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=3,color="black",marker="8")
plt.xlabel("K value")
plt.ylabel("WCSS")
plt.show()
```



```
kmeans=KMeans(n_clusters=5)
#predict the label of cluster. Here we are finding cluster based on age and spending score.
label=kmeans.fit_predict(x3)
print(label)
```

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

```
# to find out centroid with x, y and z cordinate
print(kmeans.cluster_centers_)
```

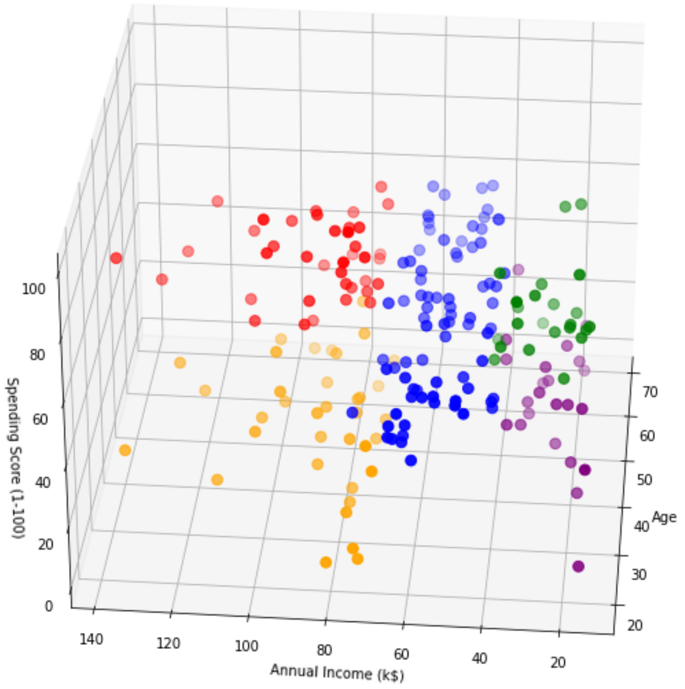
```
[[32.69230769 86.53846154 82.12820513]
 [43.08860759 55.29113924 49.56962025]
 [25.52173913 26.30434783 78.56521739]
 [45.2173913  26.30434783 20.91304348]
 [40.66666667 87.75       17.58333333]]
```

```
df=pd.read_csv("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

```
clusters=kmeans.fit_predict(x3)
df["label"]=clusters

from mpl_toolkits.mplot3d import Axes3D

fig=plt.figure(figsize=(21,11))
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()
```



```
cust1=df[df["label"]==1]
print('Number of customer in 1st group=', len(cust1))
print('They are -', cust1["CustomerID"].values)
print("                                                    ")
cust2=df[df["label"]==2]
print('Number of customer in 2nd group=', len(cust2))
print('They are -', cust2["CustomerID"].values)
print("                                                    ")
cust3=df[df["label"]==0]
print('Number of customer in 3rd group=', len(cust3))
print('They are -', cust3["CustomerID"].values)
print("                                                    ")
cust4=df[df["label"]==3]
print('Number of customer in 4th group=', len(cust4))
print('They are -', cust4["CustomerID"].values)
print("                                                    ")
cust5=df[df["label"]==4]
print('Number of customer in 5th group=', len(cust5))
print('They are -', cust5["CustomerID"].values)
print("                                                    ")
```

```
Number of customer in 1st group= 39
They are - [124 126 128 130 132 134 136 138 140 142 144 146 148 150 152 154 156 158
 160 162 164 166 168 170 172 174 176 178 180 182 184 186 188 190 192 194
 196 198 200]

Number of customer in 2nd group= 23
They are - [ 2  4  6  8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46]

Number of customer in 3rd group= 79
They are - [ 47  48  49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64
  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82
  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100
 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118
 119 120 121 122 123 127 143]

Number of customer in 4th group= 36
They are - [125 129 131 133 135 137 139 141 145 147 149 151 153 155 157 159 161 163
 165 167 169 171 173 175 177 179 181 183 185 187 189 191 193 195 197 199]

Number of customer in 5th group= 23
They are - [ 1  3  5  7  9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45]
```

There are total five cluster are formed between three dimension i.e,Age, Annual Income and Spending Score. The 3d plot are fromed with five color of red, green, orange, blue and purple.

The cusotmer in different cluster in blue, red, green, orange and purple are 39, 23, 79, 36 and 23 respectively.

The owner of mall need to focus on orange cluster which has high annual income and low spending score, to those cusotmer need to give some offer so that they incline to increase more average order value.

The cluster with purple color which as 23 cusotmer for which owner need to give more discount as they have less annual income but for spending scoreso that they can increase in overall revenue.

✓ 0s completed at 8:12 PM ● ×