K means clustering

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:

```
df=pd.read_csv(r"C:\Users\Teju\Downloads\Income.csv")
df
```

Out[2]:

	Gender	Age	Income($)
0	Male	19	15
1	Male	21	15
2	Female	20	16
3	Female	23	16
4	Female	31	17
...	...	...	...
195	Female	35	120
196	Female	45	126
197	Male	32	126
198	Male	32	137
199	Male	30	137

200 rows × 3 columns

In [3]:

```
df.head()
```

Out[3]:

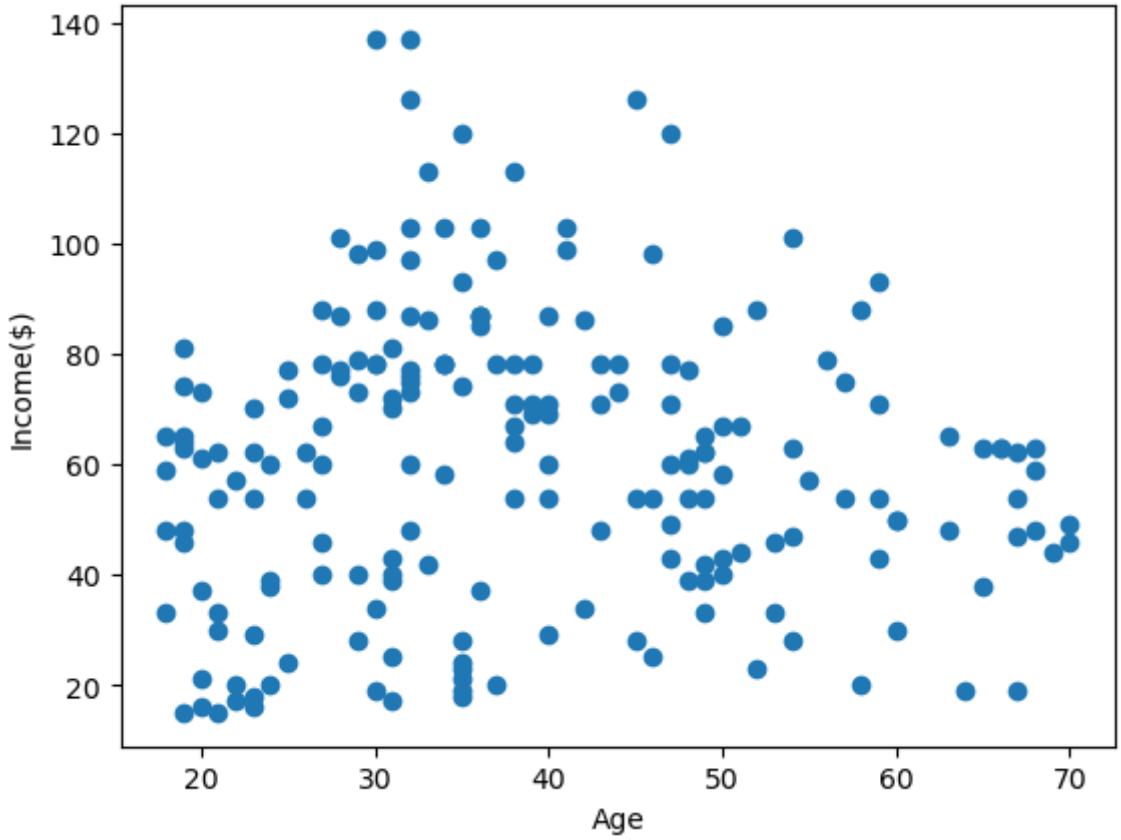
	Gender	Age	Income($)
0	Male	19	15
1	Male	21	15
2	Female	20	16
3	Female	23	16
4	Female	31	17

In [4]:

```
plt.scatter(df["Age"],df["Income($)"])
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[4]:

```
Text(0, 0.5, 'Income($)')
```



In [5]:

```
from sklearn.cluster import KMeans
km=KMeans()
km
```

Out[5]:

```
KMeans()
```

In [6]:

```
y_predicted=km.fit_predict(df[["Age","Income($)"]])
y_predicted
```

```
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
```

Out[6]:

```
array([6, 6, 6, 6, 6, 6, 6, 6, 4, 6, 4, 6, 4, 6, 6, 6, 6, 6, 4, 6, 6, 6,
       4, 6, 4, 6, 4, 6, 4, 6, 4, 6, 4, 2, 4, 2, 4, 2, 2, 2, 4, 2, 4, 2,
       4, 2, 4, 2, 2, 2, 4, 2, 2, 4, 4, 4, 4, 0, 2, 0, 0, 2, 0, 0, 0, 2,
       0, 0, 2, 2, 0, 0, 0, 0, 0, 5, 0, 5, 5, 0, 0, 5, 0, 0, 5, 0, 0, 5,
       5, 0, 0, 5, 0, 5, 5, 5, 0, 5, 0, 5, 5, 0, 0, 5, 0, 5, 0, 0, 0, 0,
       0, 5, 1, 5, 5, 5, 0, 0, 0, 0, 5, 1, 1, 1, 5, 1, 1, 1, 0, 1, 1, 1,
       5, 1, 5, 1, 1, 1, 5, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7,
       7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 3, 3, 3, 3, 3, 3,
       3, 3])
```

In [7]:

```
df["cluster"]=y_predicted
df.head()
```

Out[7]:

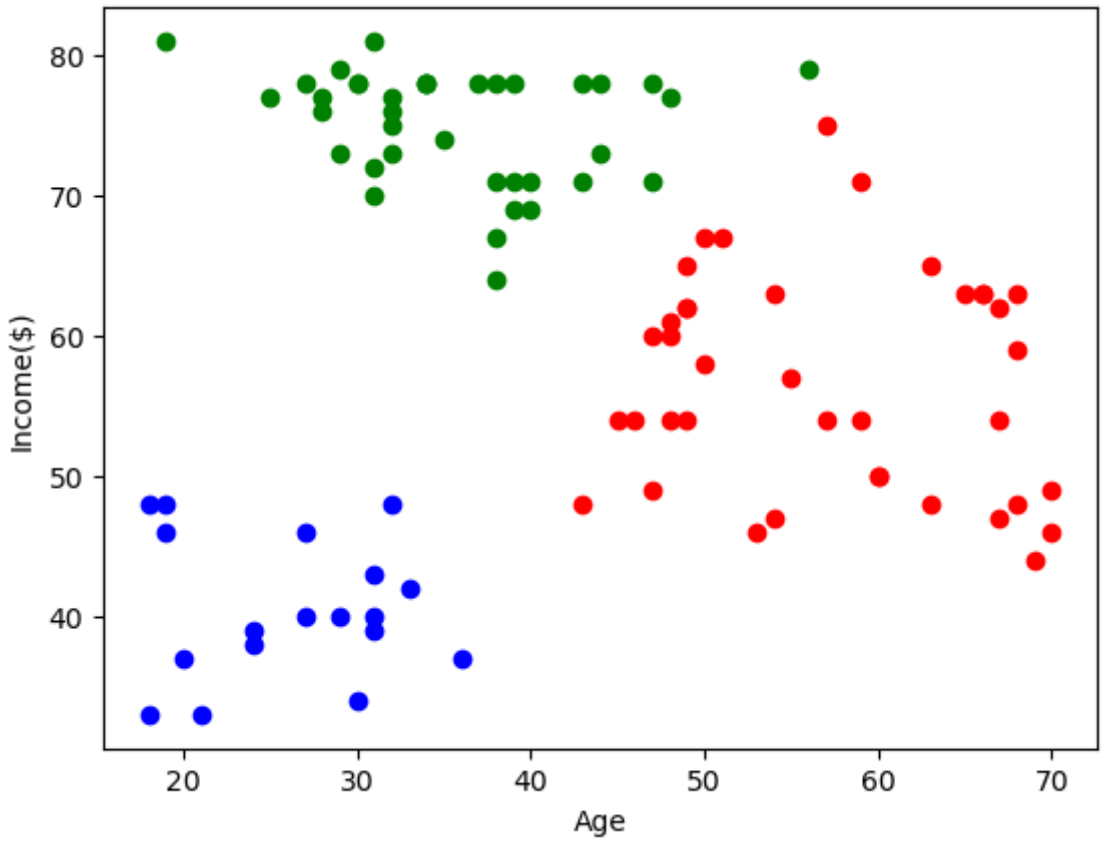
	Gender	Age	Income($)	cluster
0	Male	19	15	6
1	Male	21	15	6
2	Female	20	16	6
3	Female	23	16	6
4	Female	31	17	6

In [8]:

```
df1=df[df.cluster==0]
df2=df[df.cluster==1]
df3=df[df.cluster==2]
plt.scatter(df1["Age"],df1["Income($)"],color="red")
plt.scatter(df2["Age"],df2["Income($)"],color="green")
plt.scatter(df3["Age"],df3["Income($)"],color="blue")
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[8]:

```
Text(0, 0.5, 'Income($)')
```



In [9]:

```
from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
scaler.fit(df[["Income($)"]])
df["Income($)"]=scaler.transform(df[["Income($)"]])
df.head()
```

Out[9]:

	Gender	Age	Income($)	cluster
0	Male	19	0.000000	6
1	Male	21	0.000000	6
2	Female	20	0.008197	6
3	Female	23	0.008197	6
4	Female	31	0.016393	6

In [10]:

```
scaler.fit(df[["Age"]])
df["Age"]=scaler.transform(df[["Age"]])
df.head()
```

Out[10]:

	Gender	Age	Income($)	cluster
0	Male	0.019231	0.000000	6
1	Male	0.057692	0.000000	6
2	Female	0.038462	0.008197	6
3	Female	0.096154	0.008197	6
4	Female	0.250000	0.016393	6

In [11]:

```
km=KMeans()
```

In [12]:

```
y_predicted=km.fit_predict(df[["Age","Income($)"]])
y_predicted
```

```
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
```

Out[12]:

```
array([2, 2, 2, 2, 5, 2, 5, 2, 6, 5, 6, 5, 7, 2, 5, 2, 5, 2, 7, 5, 5, 2,
       7, 5, 7, 5, 7, 5, 5, 2, 6, 2, 7, 2, 7, 2, 7, 5, 5, 2, 6, 2, 7, 5,
       7, 2, 7, 5, 5, 5, 7, 5, 5, 6, 7, 7, 7, 6, 4, 7, 6, 4, 6, 7, 6, 4,
       7, 6, 4, 5, 6, 7, 6, 6, 6, 4, 7, 7, 4, 7, 6, 0, 6, 7, 4, 7, 1, 4,
       0, 1, 6, 4, 1, 0, 0, 4, 1, 4, 1, 4, 4, 1, 6, 4, 1, 4, 6, 1, 6, 6,
       6, 4, 0, 4, 4, 4, 6, 1, 1, 1, 4, 0, 0, 0, 4, 0, 1, 0, 1, 0, 1, 0,
       4, 0, 4, 0, 1, 0, 4, 0, 1, 0, 0, 0, 4, 0, 1, 0, 0, 0, 1, 0, 1, 0,
       1, 0, 0, 0, 0, 0, 1, 0, 4, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       1, 0, 1, 0, 3, 3, 3, 3, 3, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
       3, 3])
```

In [13]:

```
df["New Cluster"]=y_predicted
df.head()
```

Out[13]:

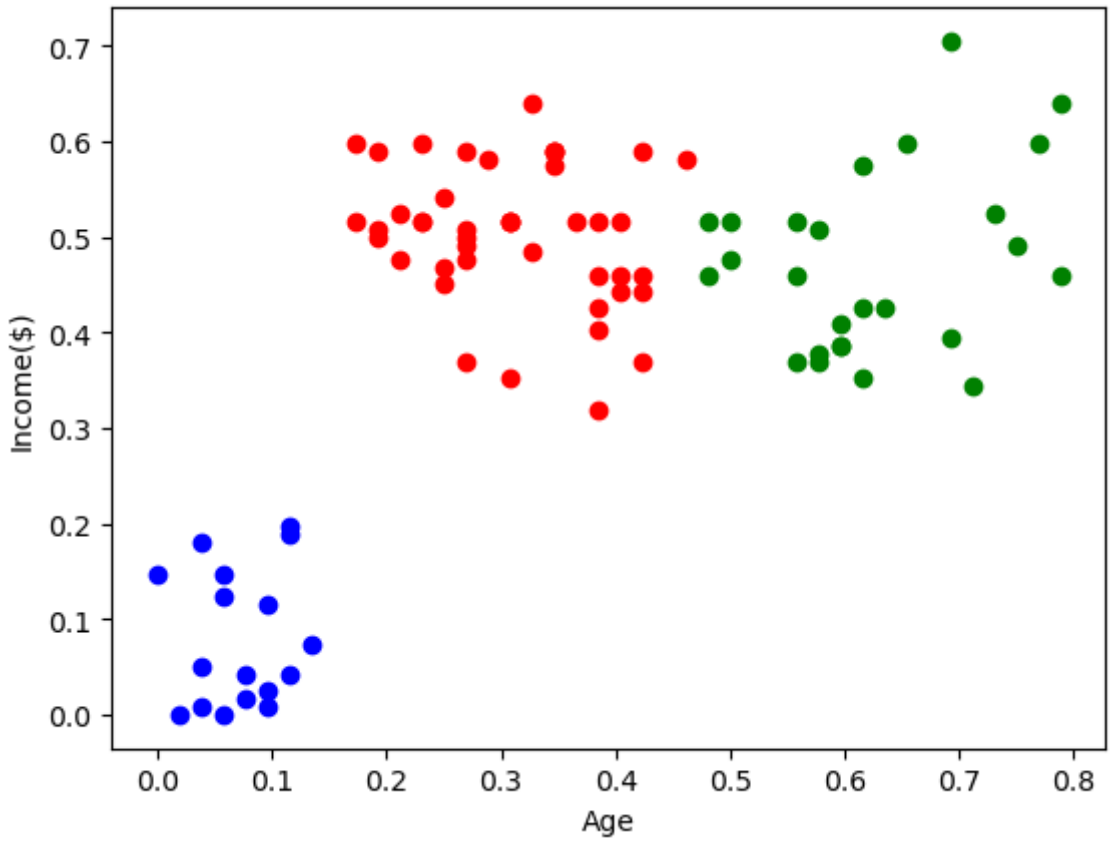
	Gender	Age	Income($)	cluster	New Cluster
0	Male	0.019231	0.000000	6	2
1	Male	0.057692	0.000000	6	2
2	Female	0.038462	0.008197	6	2
3	Female	0.096154	0.008197	6	2
4	Female	0.250000	0.016393	6	5

In [14]:

```
df1=df[df["New Cluster"]==0]
df2=df[df["New Cluster"]==1]
df3=df[df["New Cluster"]==2]
plt.scatter(df1["Age"],df1["Income($)"],color="red")
plt.scatter(df2["Age"],df2["Income($)"],color="green")
plt.scatter(df3["Age"],df3["Income($)"],color="blue")
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[14]:

```
Text(0, 0.5, 'Income($)')
```



In [15]:

```
km.cluster_centers_
```

Out[15]:

```
array([[0.30944056, 0.50428465],
       [0.62352071, 0.47225725],
       [0.07239819, 0.08003857],
       [0.34008097, 0.77998274],
       [0.07322485, 0.38272383],
       [0.28388278, 0.1245121 ],
       [0.89799331, 0.28011404],
       [0.58974359, 0.20969945]])
```

In [16]:

```
km.cluster_centers_
```

Out[16]:

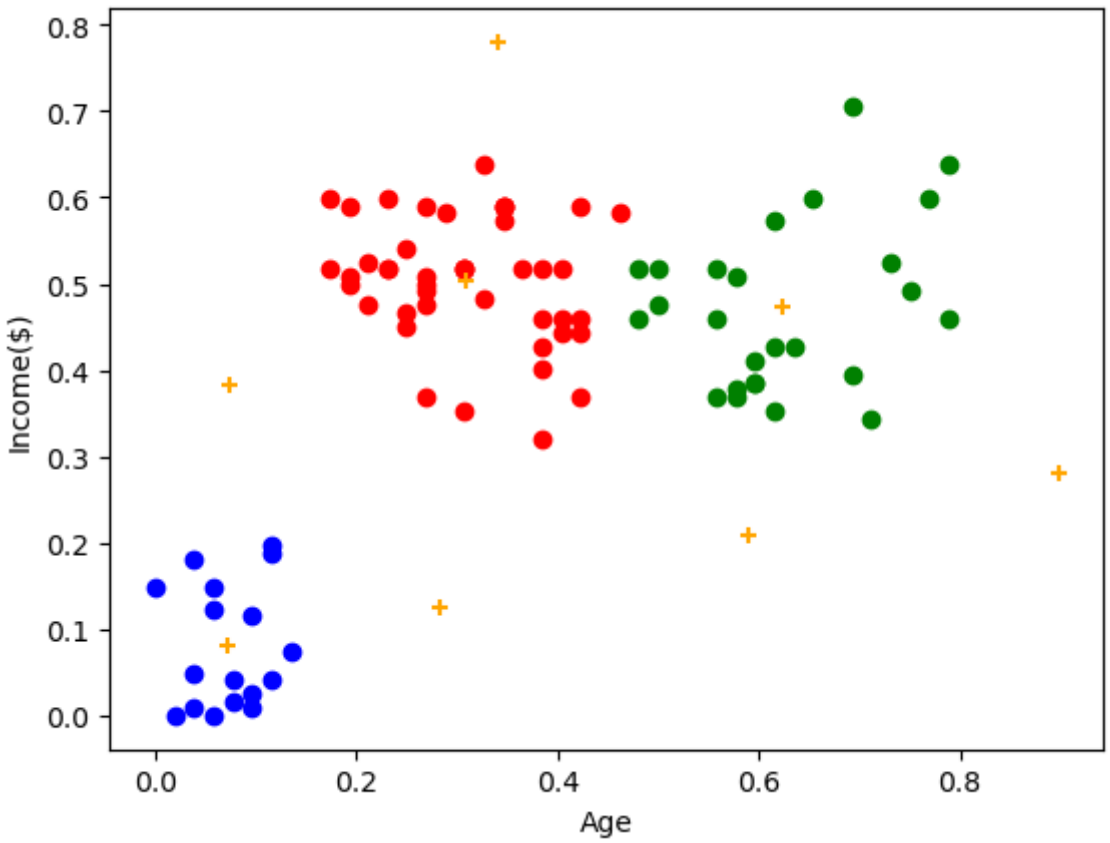
```
array([[0.30944056, 0.50428465],
       [0.62352071, 0.47225725],
       [0.07239819, 0.08003857],
       [0.34008097, 0.77998274],
       [0.07322485, 0.38272383],
       [0.28388278, 0.1245121 ],
       [0.89799331, 0.28011404],
       [0.58974359, 0.20969945]])
```

In [17]:

```
df1=df[df["New Cluster"]==0]
df2=df[df["New Cluster"]==1]
df3=df[df["New Cluster"]==2]
plt.scatter(df1["Age"],df1["Income($)"],color="red")
plt.scatter(df2["Age"],df2["Income($)"],color="green")
plt.scatter(df3["Age"],df3["Income($)"],color="blue")
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:,1],color="orange",marker="+")
plt.xlabel("Age")
plt.ylabel("Income($)")
```

Out[17]:

```
Text(0, 0.5, 'Income($)')
```



In [18]:

```
k_rng=range(1,10)
sse=[]
```

In [19]:

```
for k in k_rng:
    km=KMeans(n_clusters=k)
    km.fit(df[["Age","Income($)"]])
    sse.append(km.inertia_)
#km.inertia_ will give you the value of sum of square error
print(sse)
plt.plot(k_rng,sse)
plt.xlabel("K")
plt.ylabel("Sum of Squared Error")
```

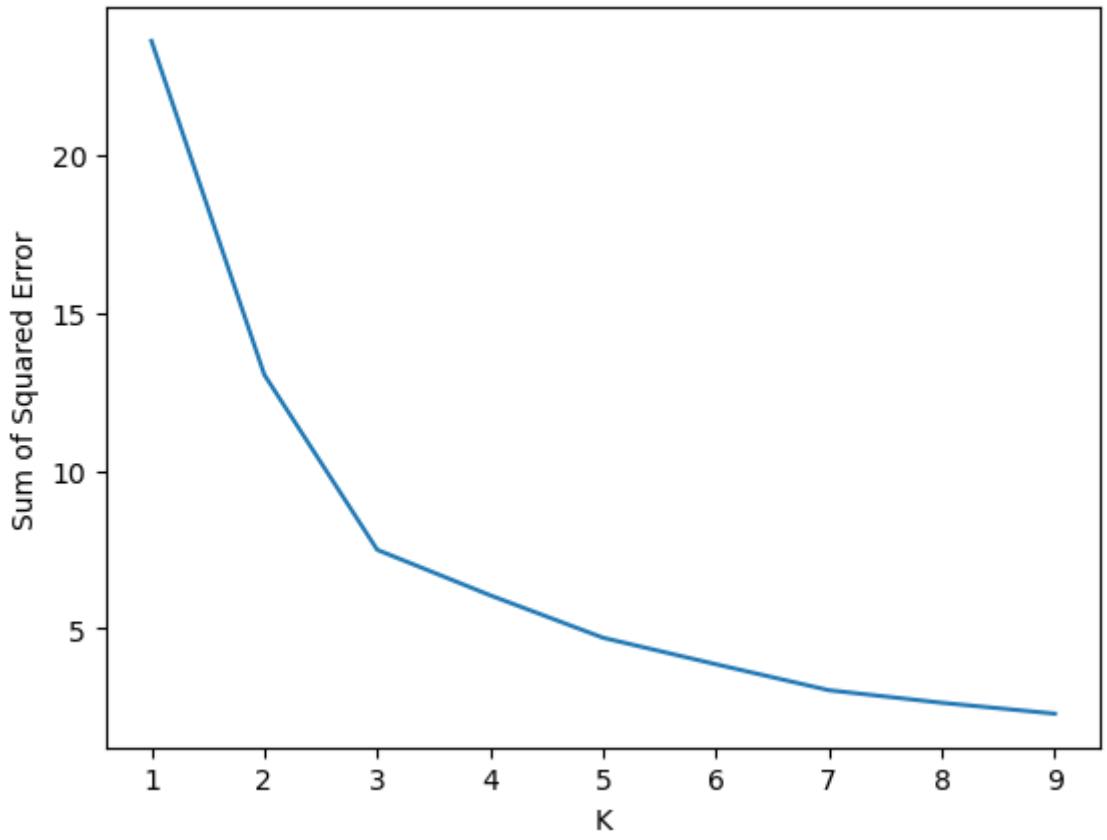
```
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
```

```
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:870:
FutureWarning: The default value of `n_init` will change from 10 to 'auto'
in 1.4. Set the value of `n_init` explicitly to suppress the warning
  warnings.warn(
C:\Users\Teju\anaconda3\lib\site-packages\sklearn\cluster\_kmeans.py:1382:
UserWarning: KMeans is known to have a memory leak on Windows with MKL, wh
en there are less chunks than available threads. You can avoid it by setti
ng the environment variable OMP_NUM_THREADS=1.
  warnings.warn(

[23.583906150363603, 13.02893842801829, 7.49302484330499, 6.05585864481254
7, 4.713025598595381, 3.8746475334239223, 3.054717436369359, 2.65738659219
7303, 2.3135720353543285]
```

Out[19]:

```
Text(0, 0.5, 'Sum of Squared Error')
```



In [ ]: