K_MEANS_CLUSTERING

June 14, 2023

#K-MEANS CLUSTERING

---

[1]:
```
import pandas as pd
from matplotlib import pyplot as plt
%matplotlib inline
```

[2]:
```
df=pd.read_csv(r"/content/Income.csv")
```

[23]:
```
df
```

[23]:

	Gender	Age	Income($)	cluster	New Cluster
0	Male	0.019231	0.000000	4	5
1	Male	0.057692	0.000000	4	5
2	Female	0.038462	0.008197	4	5
3	Female	0.096154	0.008197	4	5
4	Female	0.250000	0.016393	4	0
..	…	…	…	…	…
195	Female	0.326923	0.860656	3	6
196	Female	0.519231	0.909836	3	6
197	Male	0.269231	0.909836	3	6
198	Male	0.269231	1.000000	3	6
199	Male	0.230769	1.000000	3	6

[200 rows x 5 columns]

[24]:
```
df.head()
```

[24]:

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

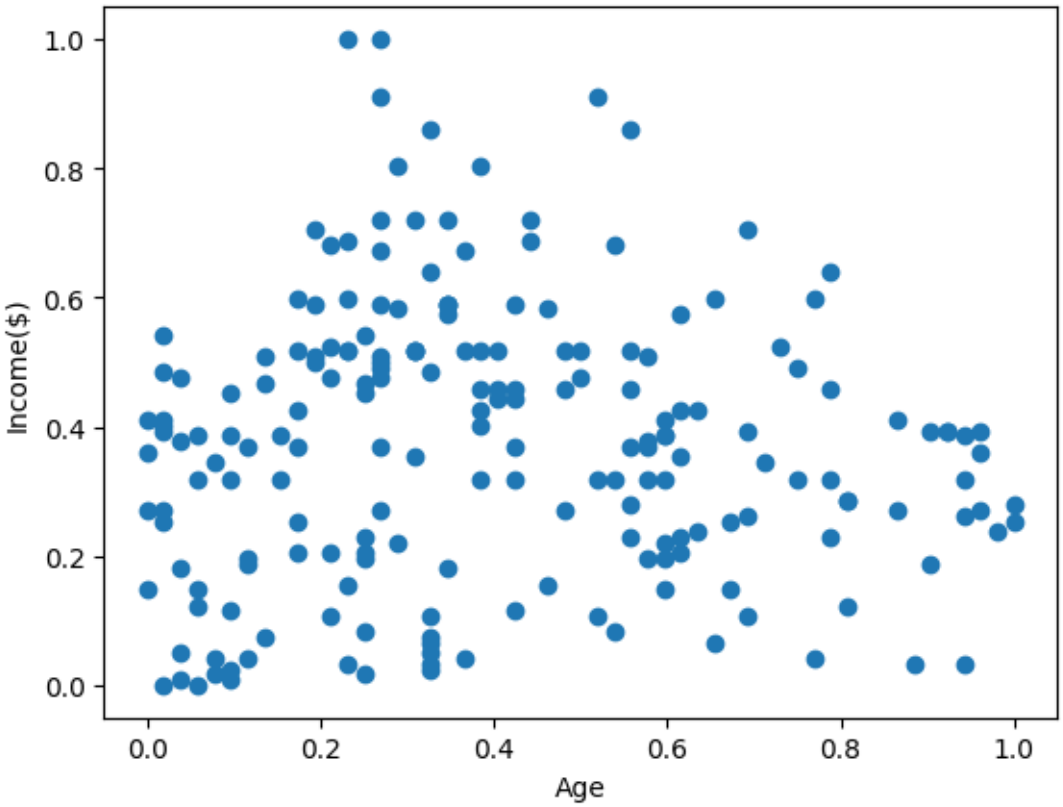
[25]:
```
df.tail()
```

[25]:

	Gender	Age	Income($)	cluster	New Cluster
195	Female	0.326923	0.860656	3	6
196	Female	0.519231	0.909836	3	6
197	Male	0.269231	0.909836	3	6
198	Male	0.269231	1.000000	3	6
199	Male	0.230769	1.000000	3	6

[26]:
```
plt.scatter(df["Age"],df["Income($)"])
plt.xlabel("Age")
plt.ylabel("Income($)")
```

[26]:
```
Text(0, 0.5, 'Income($)')
```



[27]:
```
from sklearn.cluster import KMeans
km=KMeans()
km
```

[27]:
```
KMeans()
```

[28]:
```
y_predicted=km.fit_predict(df[["Age","Income($)"]])
y_predicted
```

```
/usr/local/lib/python3.10/dist-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(
```

[28]:
```
array([5, 5, 5, 5, 1, 5, 1, 5, 2, 1, 2, 1, 4, 5, 1, 5, 1, 5, 4, 1, 1, 5,
       4, 1, 4, 1, 4, 1, 1, 5, 2, 5, 4, 5, 4, 5, 4, 1, 1, 5, 2, 5, 4, 1,
       4, 5, 4, 1, 1, 1, 4, 1, 1, 2, 4, 4, 4, 2, 7, 4, 2, 7, 2, 4, 2, 7,
       4, 2, 7, 1, 2, 4, 2, 2, 2, 7, 4, 4, 7, 4, 2, 3, 2, 4, 7, 4, 0, 7,
       3, 0, 2, 7, 0, 3, 3, 7, 0, 7, 0, 7, 7, 0, 2, 7, 0, 7, 2, 0, 2, 2,
       2, 7, 3, 7, 7, 7, 2, 0, 0, 0, 7, 3, 3, 3, 7, 3, 0, 3, 0, 3, 0, 3,
       7, 3, 7, 3, 0, 3, 7, 3, 0, 3, 3, 3, 7, 3, 0, 3, 3, 3, 0, 3, 0, 3,
       0, 3, 3, 3, 3, 3, 0, 3, 7, 3, 0, 3, 3, 3, 3, 3, 3, 3, 3, 3, 0, 3,
       0, 3, 0, 3, 6, 6, 0, 6, 6, 6, 0, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6,
       6, 6], dtype=int32)
```

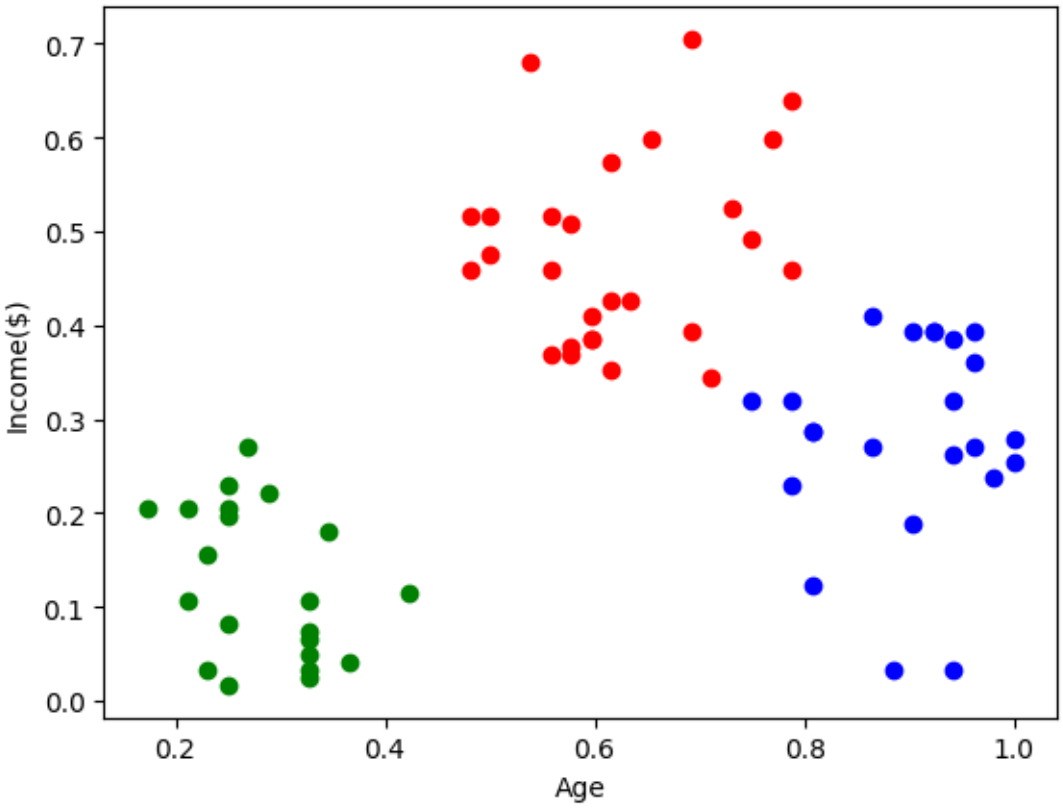
[29]:
```
df["cluster"]=y_predicted
df.head()
```

[29]:

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

[30]:
```
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($)"],color="blue")
plt.xlabel("Age")
plt.ylabel("Income($)")
```

[30]:
```
Text(0, 0.5, 'Income($)')
```



[31]:
```
from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
scaler.fit(df[["Income($)"]])
df["Income($)"]=scaler.transform(df[["Income($)"]])
df.head()
```

[31]:

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

[32]:
```
scaler.fit(df[["Age"]])
df["Age"]=scaler.transform(df[["Age"]])
df.head()
```

[32]:

	Gender	Age	Income($)	cluster	New Cluster
0	Male	0.019231	0.000000	5	5
1	Male	0.057692	0.000000	5	5
2	Female	0.038462	0.008197	5	5

```
3  Female  0.096154   0.008197        5            5
4  Female  0.250000   0.016393        1            0
```

[33]:
```
km=KMeans()
```

[34]:
```
y_predicted=km.fit_predict(df[["Age","Income($)"]])
y_predicted
```

```
/usr/local/lib/python3.10/dist-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(
```

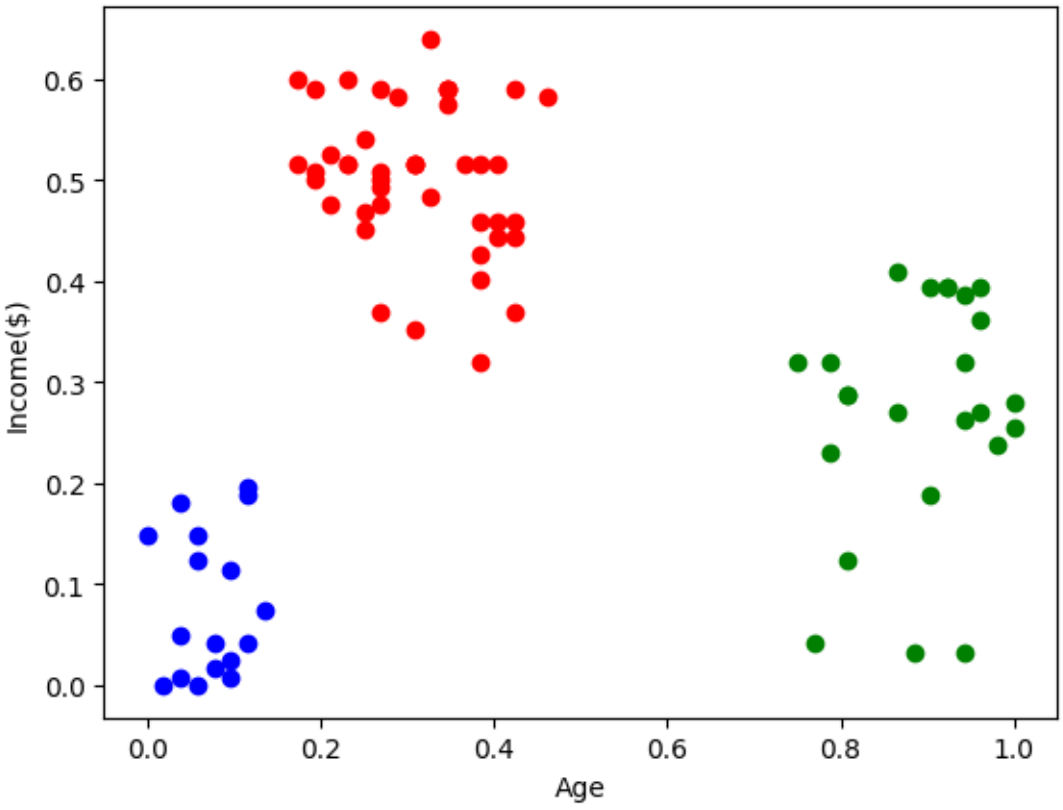
[34]:
```
array([2, 2, 2, 2, 7, 2, 7, 2, 1, 7, 1, 7, 1, 2, 7, 2, 7, 2, 3, 7, 7, 2,
       3, 7, 3, 7, 3, 7, 7, 2, 1, 2, 3, 2, 3, 2, 3, 7, 7, 2, 1, 2, 3, 7,
       3, 2, 3, 7, 7, 7, 3, 7, 7, 1, 3, 3, 3, 1, 4, 3, 1, 4, 1, 3, 1, 4,
       3, 1, 4, 7, 1, 3, 1, 1, 1, 4, 3, 3, 4, 3, 1, 0, 1, 3, 4, 3, 3, 4,
       0, 3, 1, 4, 3, 0, 0, 4, 3, 4, 3, 4, 4, 3, 1, 4, 3, 4, 1, 5, 1, 1,
       1, 4, 0, 4, 4, 4, 1, 5, 5, 5, 4, 0, 0, 0, 4, 0, 5, 0, 5, 0, 5, 0,
       4, 0, 4, 0, 5, 0, 4, 0, 5, 0, 0, 0, 4, 0, 5, 0, 0, 0, 5, 0, 5, 0,
       5, 0, 0, 0, 0, 0, 5, 0, 4, 0, 5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0,
       5, 0, 5, 0, 6, 6, 5, 6, 6, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6,
       6, 6], dtype=int32)
```

[35]:
```
df["New Cluster"]=y_predicted
df.head()
```

[35]:
```
   Gender       Age  Income($)  cluster  New Cluster
0    Male  0.019231   0.000000        5            2
1    Male  0.057692   0.000000        5            2
2  Female  0.038462   0.008197        5            2
3  Female  0.096154   0.008197        5            2
4  Female  0.250000   0.016393        1            7
```

[36]:
```
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($)"],color="blue")
plt.xlabel("Age")
plt.ylabel("Income($)")
```

[36]:
```
Text(0, 0.5, 'Income($)')
```

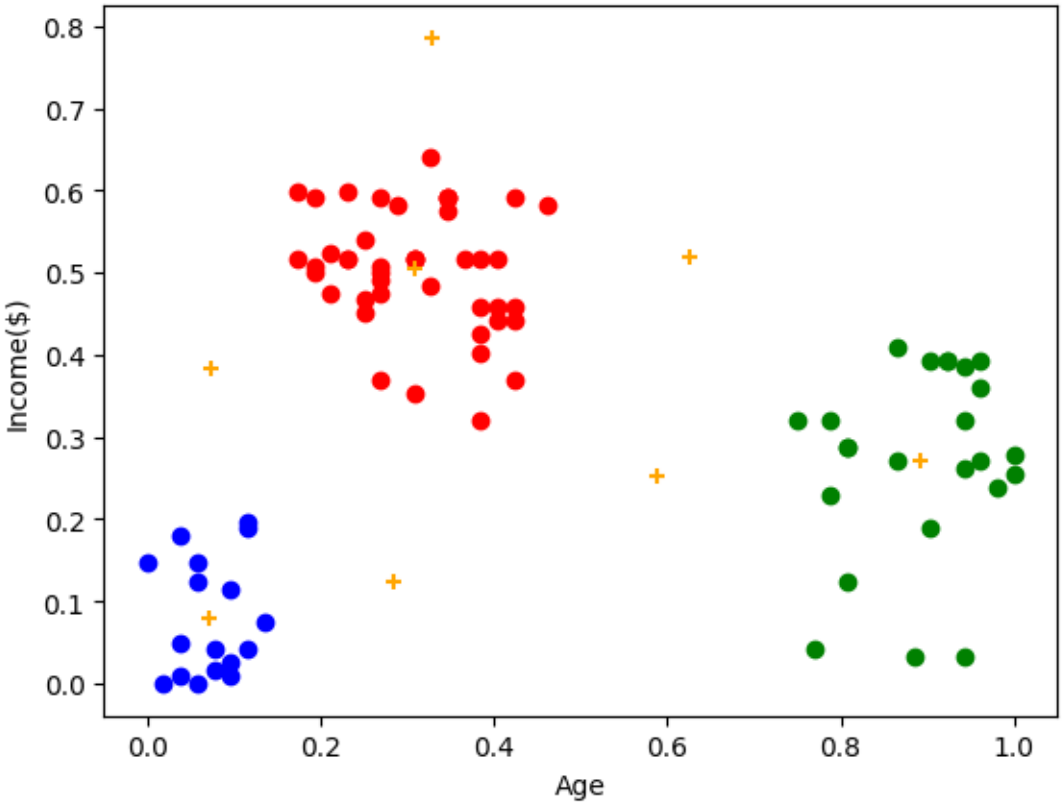


[37]:
```
km.cluster_centers_
```

[37]:
```
array([[0.30944056, 0.50428465],
       [0.89262821, 0.27015027],
       [0.07239819, 0.08003857],
       [0.58717949, 0.25245902],
       [0.07322485, 0.38272383],
       [0.62596154, 0.51885246],
       [0.32905983, 0.78551913],
       [0.28388278, 0.1245121 ]])
```

[38]:
```
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($)"],color="blue")
plt.scatter(km.cluster_centers_[:,0],km.cluster_centers_[:
 ↪,1],color="orange",marker="+")
plt.xlabel("Age")
plt.ylabel("Income($)")
```

[38]: Text(0, 0.5, 'Income($)')



[39]:
```
k_rng=range(1,10)
sse=[]
```

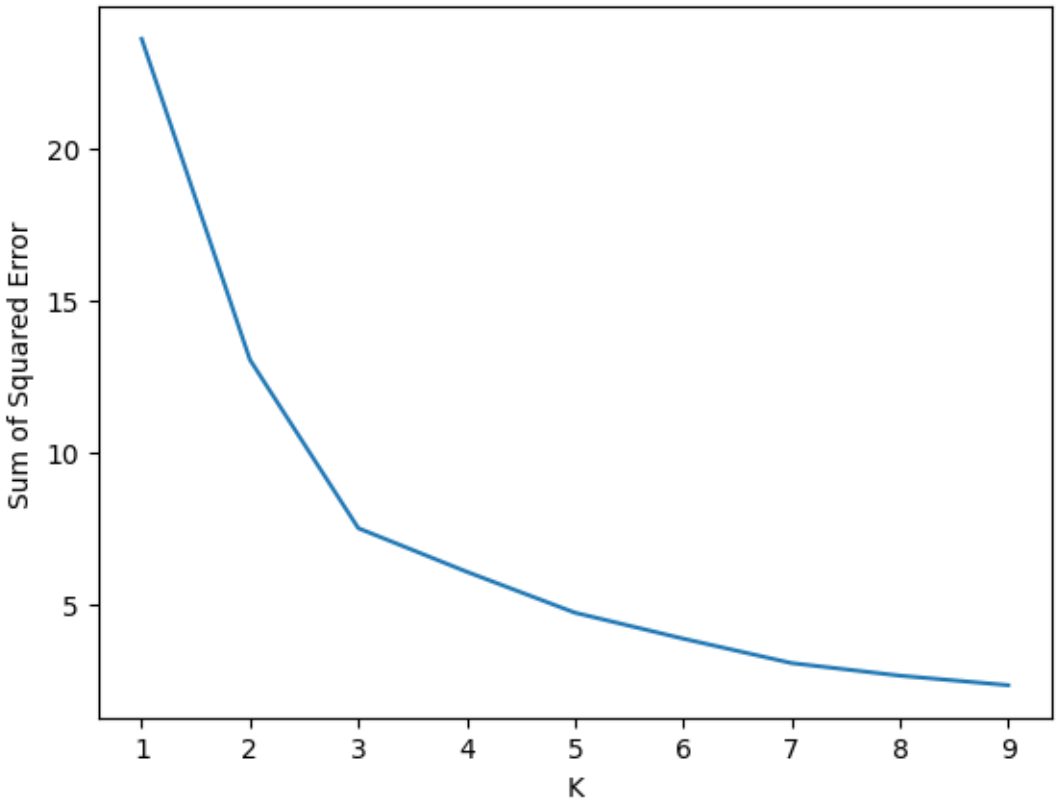
[40]:
```
for k in k_rng:
 km=KMeans(n_clusters=k)
 km.fit(df[["Age","Income($)"]])
 sse.append(km.inertia_)
#km.inertia_ will give you the value of sum of square error
print(sse)
plt.plot(k_rng,sse)
plt.xlabel("K")
plt.ylabel("Sum of Squared Error")
```

```
/usr/local/lib/python3.10/dist-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(
/usr/local/lib/python3.10/dist-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(
/usr/local/lib/python3.10/dist-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(
/usr/local/lib/python3.10/dist-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(
/usr/local/lib/python3.10/dist-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(
/usr/local/lib/python3.10/dist-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(
/usr/local/lib/python3.10/dist-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(
/usr/local/lib/python3.10/dist-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(
/usr/local/lib/python3.10/dist-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(

[23.58390615036361, 13.028938428018284, 7.492107868586011, 6.06068162047012,
4.713025598595382, 3.859055754701024, 3.0559862119202013, 2.642693946921809,
2.3291765544665175]
```

```
[40]: Text(0, 0.5, 'Sum of Squared Error')
```


Sum of Squared Error
20
15
10
5
1
2
3
4
5
6
7
8
9
K