K-Means clustering

In [34]:

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

In [35]:

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

Out[35]:

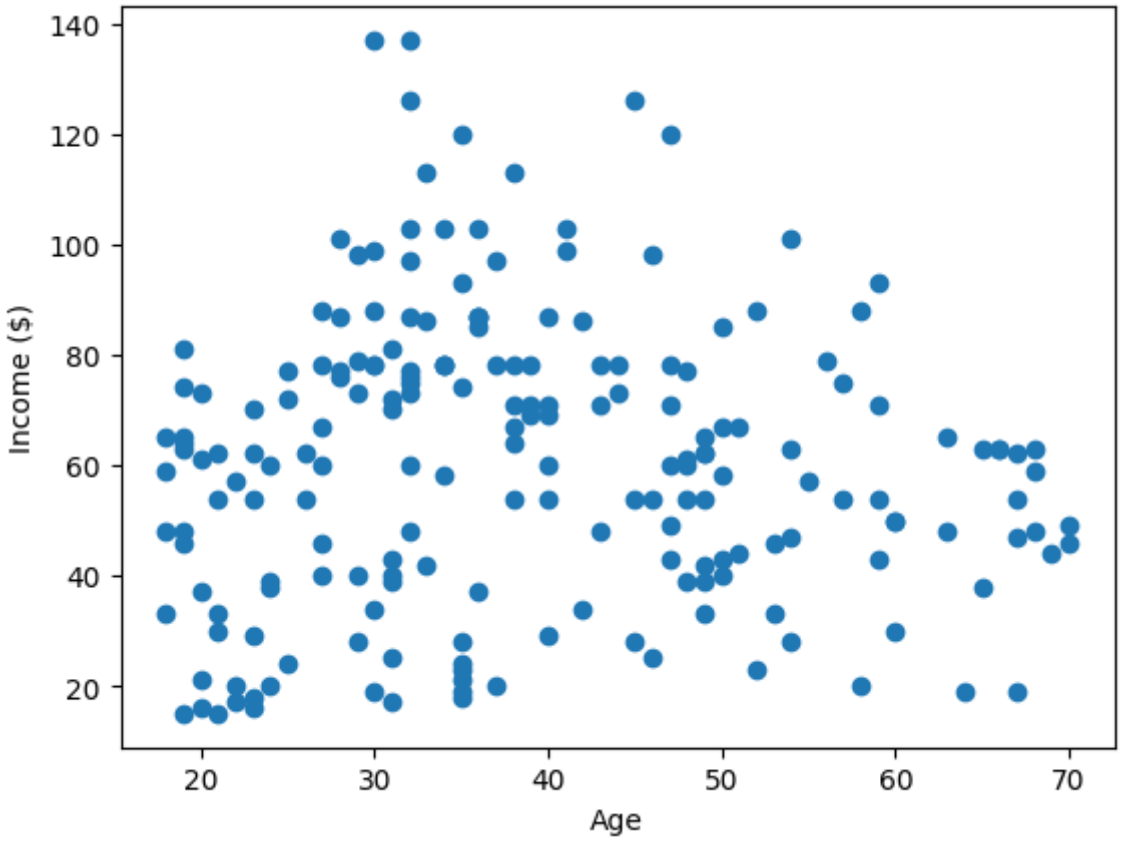
	Gender	Age	Income($)
0	Male	19	15
1	Male	21	15
2	Female	20	16
3	Female	23	16
4	Female	31	17

In [36]:

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

Out[36]:

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



In [37]:

```
from sklearn.cluster import KMeans
```

In [38]:

```
km=KMeans()
km
```

Out[38]:

```
KMeans()
```

In [39]:

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

```
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
```

Out[39]:

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

In [40]:

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

Out[40]:

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

In [41]:

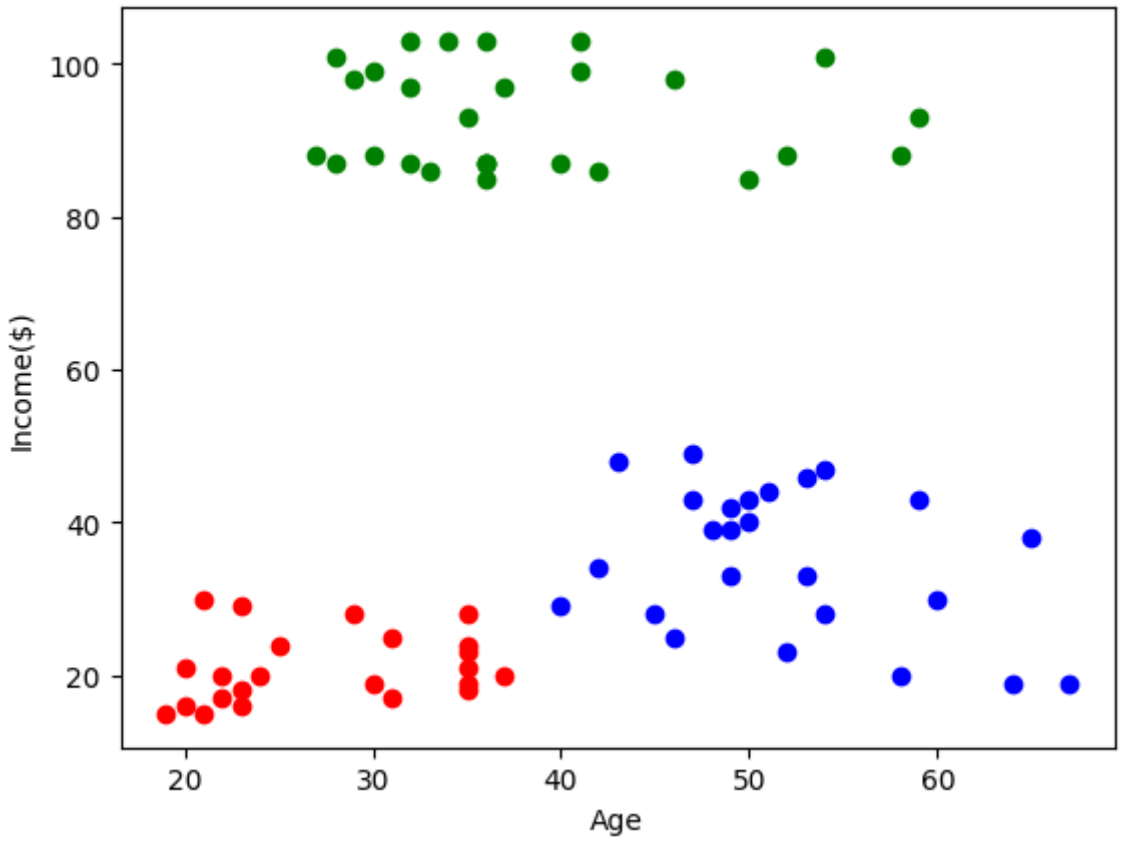
```
df1=df[df.Cluster==0]
df2=df[df.Cluster==2]
df3=df[df.Cluster==3]

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[41]:

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



In [42]:

```
from sklearn.preprocessing import MinMaxScaler
```

In [43]:

```
scaler=MinMaxScaler()
```

In [44]:

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

Out[44]:

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

In [45]:

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

Out[45]:

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

In [46]:

```
km=KMeans()
```

In [47]:

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

```
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
```

Out[47]:

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

In [48]:

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

Out[48]:

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

In [49]:

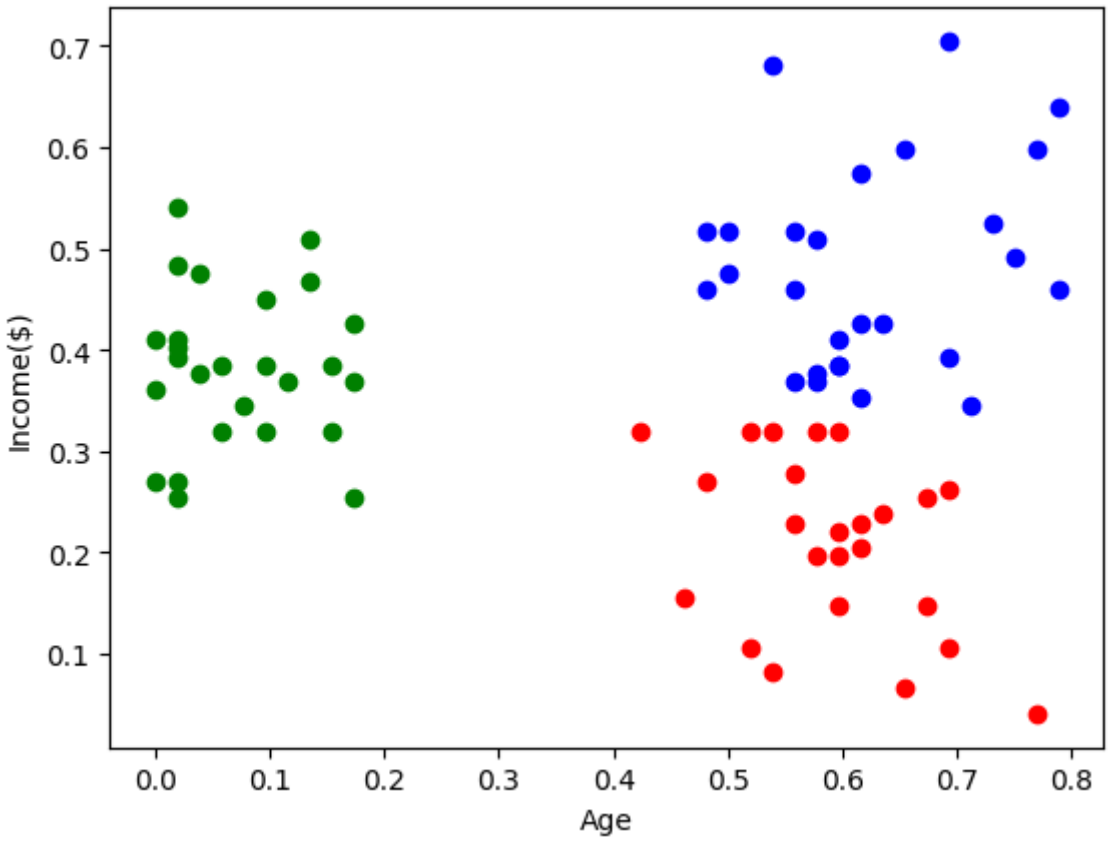
```
df1=df[df["New Cluster"]==0]
df2=df[df["New Cluster"]==1]
df3=df[df["New Cluster"]==2]

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[49]:

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



In [50]:

```
km.cluster_centers_
```

Out[50]:

```
array([[0.58974359, 0.20969945],
       [0.07322485, 0.38272383],
       [0.62037037, 0.47996357],
       [0.07239819, 0.08003857],
       [0.89799331, 0.28011404],
       [0.30944056, 0.50428465],
       [0.28388278, 0.1245121 ],
       [0.32905983, 0.78551913]])
```

In [51]:

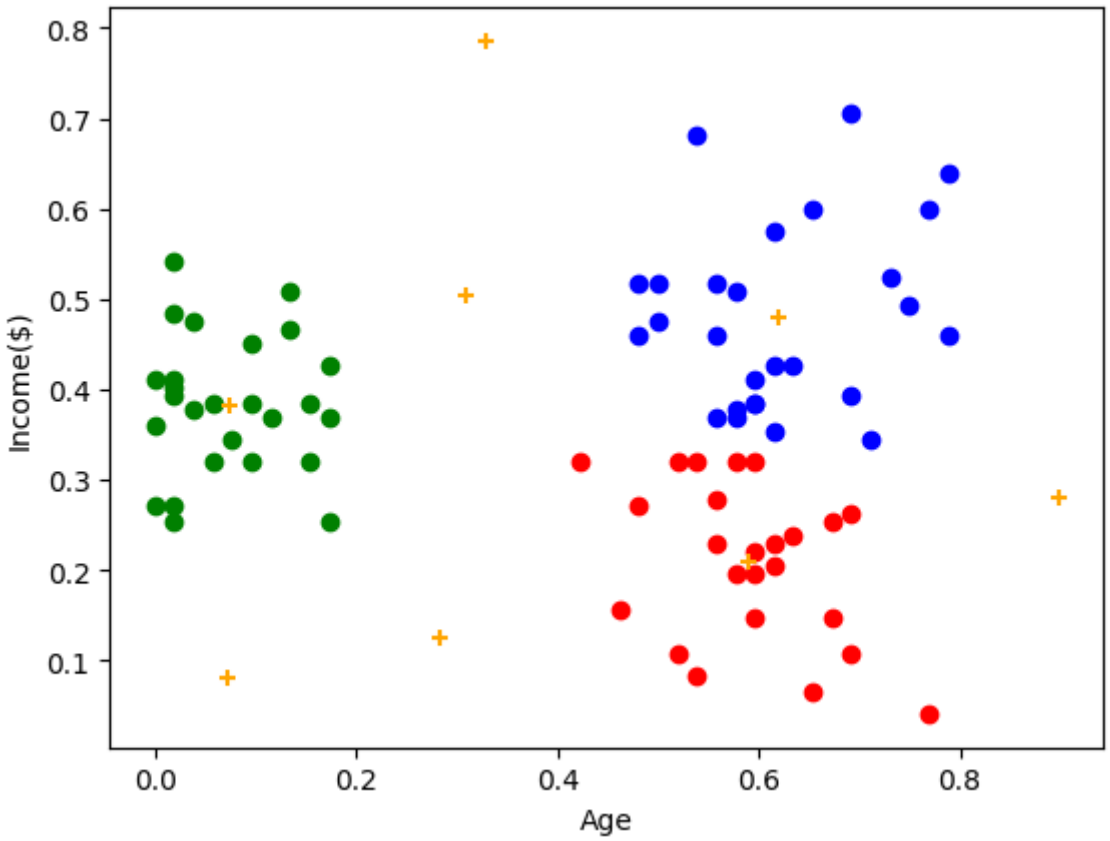
```
df1=df[df["New Cluster"]==0]
df2=df[df["New Cluster"]==1]
df3=df[df["New Cluster"]==2]

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[51]:

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



In [52]:

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

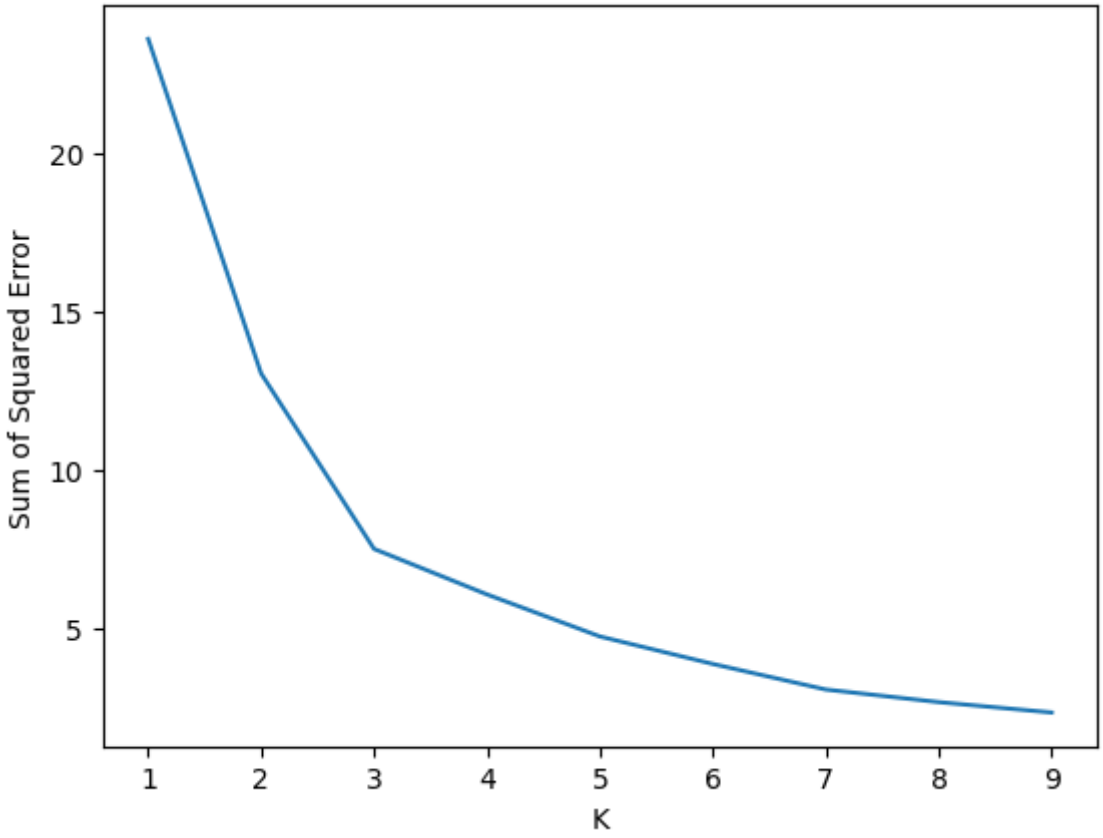
In [53]:

```
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 errorprint(sse)
plt.plot(k_rng,sse)
plt.xlabel("K")
plt.ylabel("Sum of Squared Error")
```

```
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
C:\Users\chila\AppData\Local\Programs\Python\Python310\lib\site-packages\s
klearn\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` explicit
ly to suppress the warning
  warnings.warn(
```

Out[53]:

```
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