In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```
df=pd.read_csv("Mall_Customers.csv")
df
```

Out[2]:

	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
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   CustomerID              200 non-null    int64 
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64 
 3   Annual Income (k$)      200 non-null    int64 
 4   Spending Score (1-100)  200 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

In [4]:
```
df.isnull().sum()
```

Out[4]:
```
CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

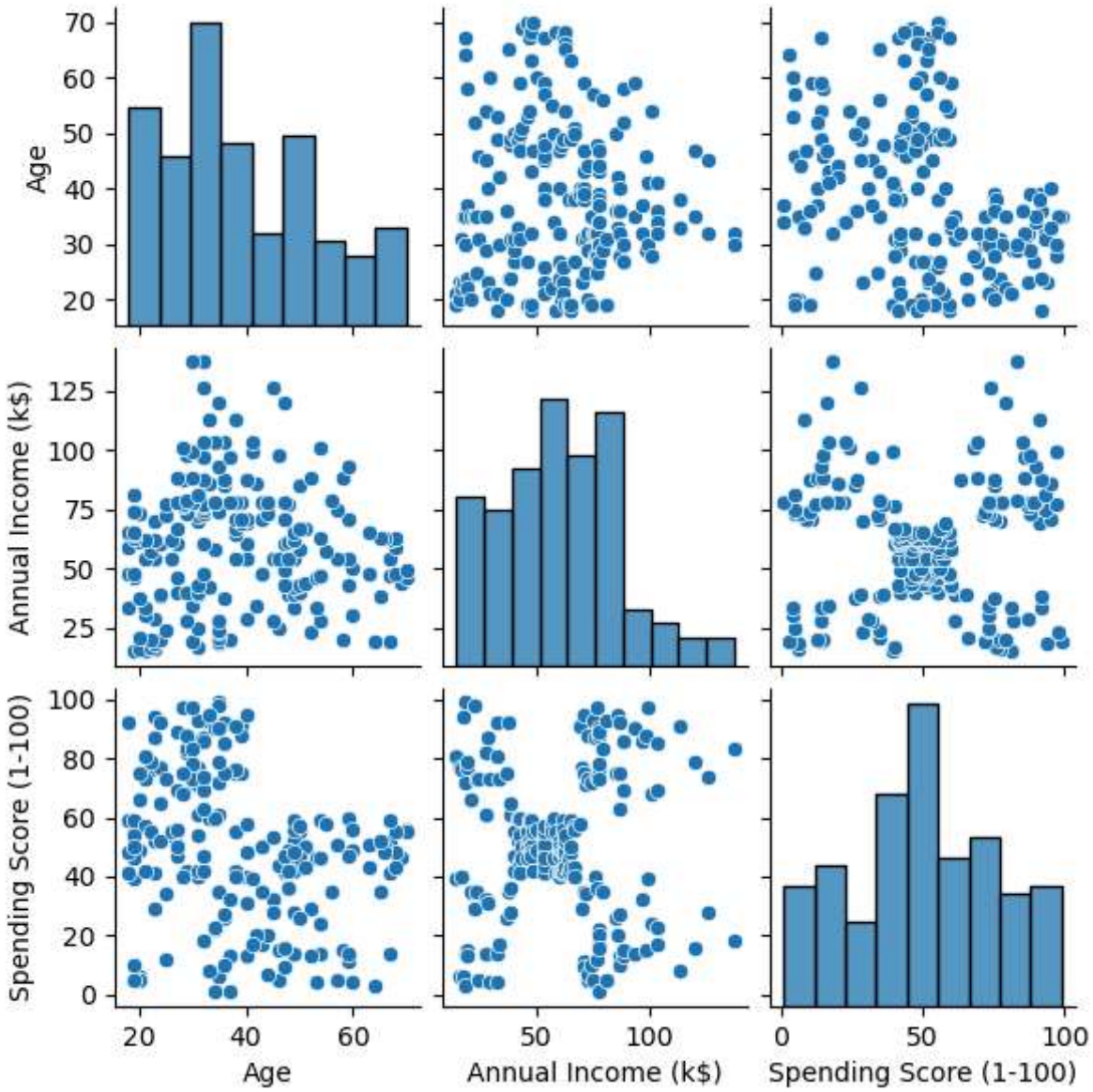
In [5]:
```
df.describe()
```

Out[5]:

	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

In [6]:
```
sns.pairplot(df[["Age", "Annual Income (k$)",   "Spending Score (1-100)"]],height=2
```

Out[6]:
```
<seaborn.axisgrid.PairGrid at 0x29d9ec71a10>
```



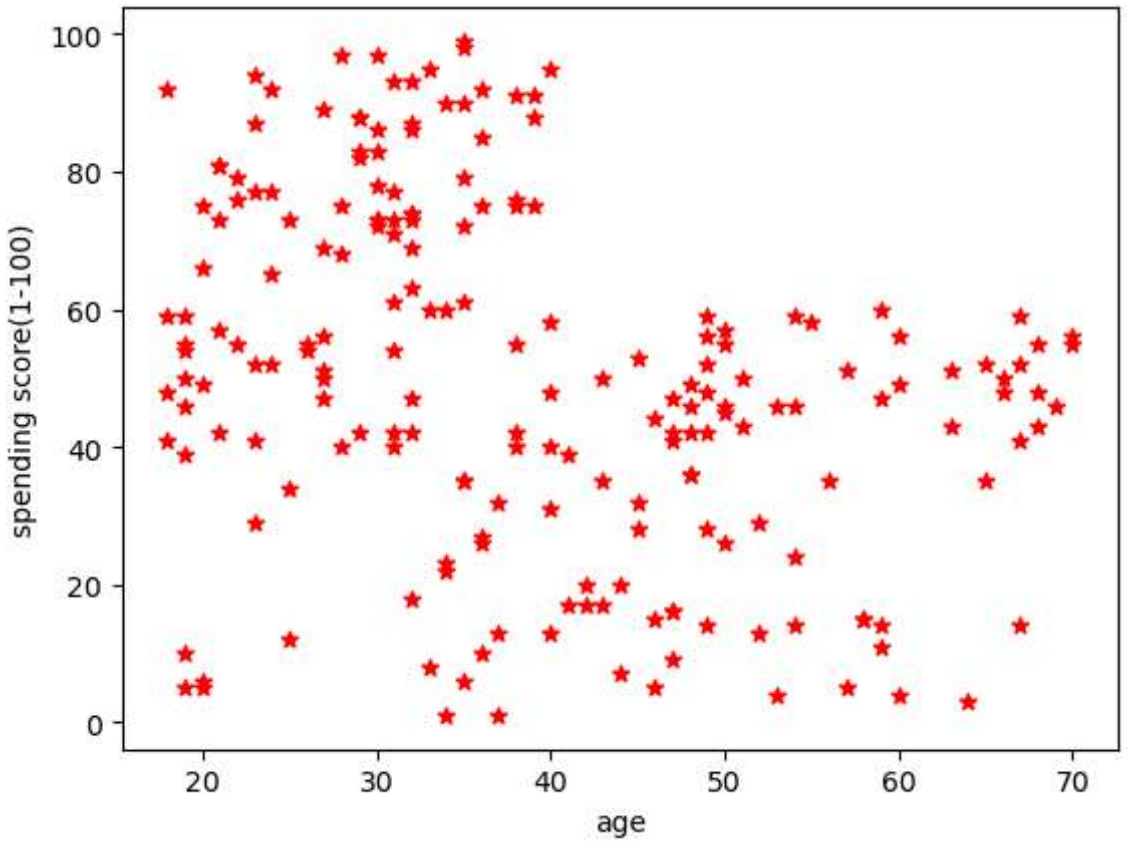
In [7]:
```
df1=df[["Age","Spending Score (1-100)"]]
df1
```

Out[7]:

	Age	Spending Score (1-100)
0	19	39
1	21	81
2	20	6
3	23	77
4	31	40
...	...	...
195	35	79
196	45	28
197	32	74
198	32	18
199	30	83

200 rows × 2 columns

In [8]:
```
plt.scatter(df1["Age"],df1["Spending Score (1-100)"],marker="*",color="r")
plt.xlabel("age")
plt.ylabel("spending score(1-100)")
plt.show()
```



In [9]:
```
from sklearn.cluster import KMeans
km=KMeans(n_clusters=3)
y_pred=km.fit_predict(df[["Age","Spending Score (1-100)"]])
y_pred
```

Out[9]:
```
array([2, 1, 0, 1, 2, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 2, 1, 0, 1, 2, 1,
       0, 1, 0, 1, 2, 2, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 2, 1, 2, 2,
       0, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1,
       2, 2, 1, 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, 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, 1, 0, 1, 2, 1, 0, 1, 0, 1,
       2, 1, 0, 1, 0, 1, 0, 1, 0, 1, 2, 1, 0, 1, 2, 1, 0, 1, 0, 1, 0, 1,
       0, 1, 0, 1, 0, 1, 2, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1,
       0, 1, 0, 1, 2, 1, 0, 1, 2, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1,
       0, 1])
```

In [10]:
```
df1["cluster"]=y_pred
df1
```

Out[10]:

	Age	Spending Score (1-100)	cluster
0	19	39	2
1	21	81	1
2	20	6	0
3	23	77	1
4	31	40	2
...	...	...	...
195	35	79	1
196	45	28	0
197	32	74	1
198	32	18	0
199	30	83	1

200 rows × 3 columns

In [11]:
```
df1[df1.cluster==0]
```

Out[11]:

	Age	Spending Score (1-100)	cluster
2	20	6	0
6	35	6	0
8	64	3	0
10	67	14	0
12	58	15	0
14	37	13	0
18	52	29	0
22	46	5	0
24	54	14	0
28	40	31	0
30	60	4	0
32	53	4	0
34	49	14	0
36	42	17	0
38	36	26	0
44	49	28	0
124	23	29	0
128	59	11	0
130	47	9	0
134	20	5	0
136	44	7	0
138	19	10	0
140	57	5	0
144	25	12	0
148	34	22	0
150	43	17	0
152	44	20	0
154	47	16	0
156	37	1	0
158	34	1	0
162	19	5	0
164	50	26	0
166	42	20	0
168	36	27	0
170	40	13	0
172	36	10	0

	Age	Spending Score (1-100)	cluster
174	52	13	0
176	58	15	0
178	59	14	0
182	46	15	0
186	54	24	0
188	41	17	0
190	34	23	0
192	33	8	0
194	47	16	0
196	45	28	0
198	32	18	0

In [12]:
```
df1[df1.cluster==1]
```

Out[12]:

	Age	Spending Score (1-100)	cluster
1	21	81	1
3	23	77	1
5	22	76	1
7	23	94	1
9	30	72	1
...	...	...	...
191	32	69	1
193	38	91	1
195	35	79	1
197	32	74	1
199	30	83	1

62 rows × 3 columns

In [13]:
```
df1[df1.cluster==2]
```

Out[13]:

	Age	Spending Score (1-100)	cluster
0	19	39	2
4	31	40	2
16	35	35	2
20	35	35	2
26	45	32	2
...	...	...	...
142	28	40	2
146	48	36	2
160	56	35	2
180	37	32	2
184	41	39	2

91 rows × 3 columns

In [14]:
```
df1[df1.cluster==3]
```

Out[14]:

Age	Spending Score (1-100)	cluster

In [15]:
```
df1[df1.cluster==4]
```

Out[15]:

Age	Spending Score (1-100)	cluster

In [16]:
```
kc=km.cluster_centers_
kc
```

Out[16]:
```
array([[42.95744681, 14.59574468],
       [29.56451613, 80.74193548],
       [43.05494505, 47.78021978]])
```

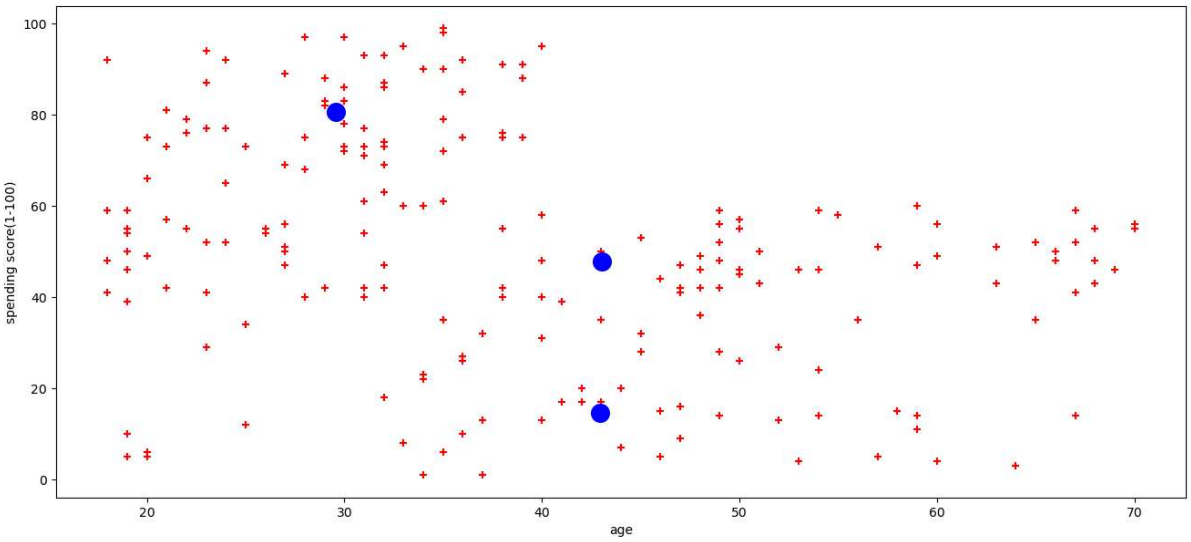
In [17]:
```
kc[:,0]
```

Out[17]:
```
array([42.95744681, 29.56451613, 43.05494505])
```

In [18]:
```
kc[:,1]
```

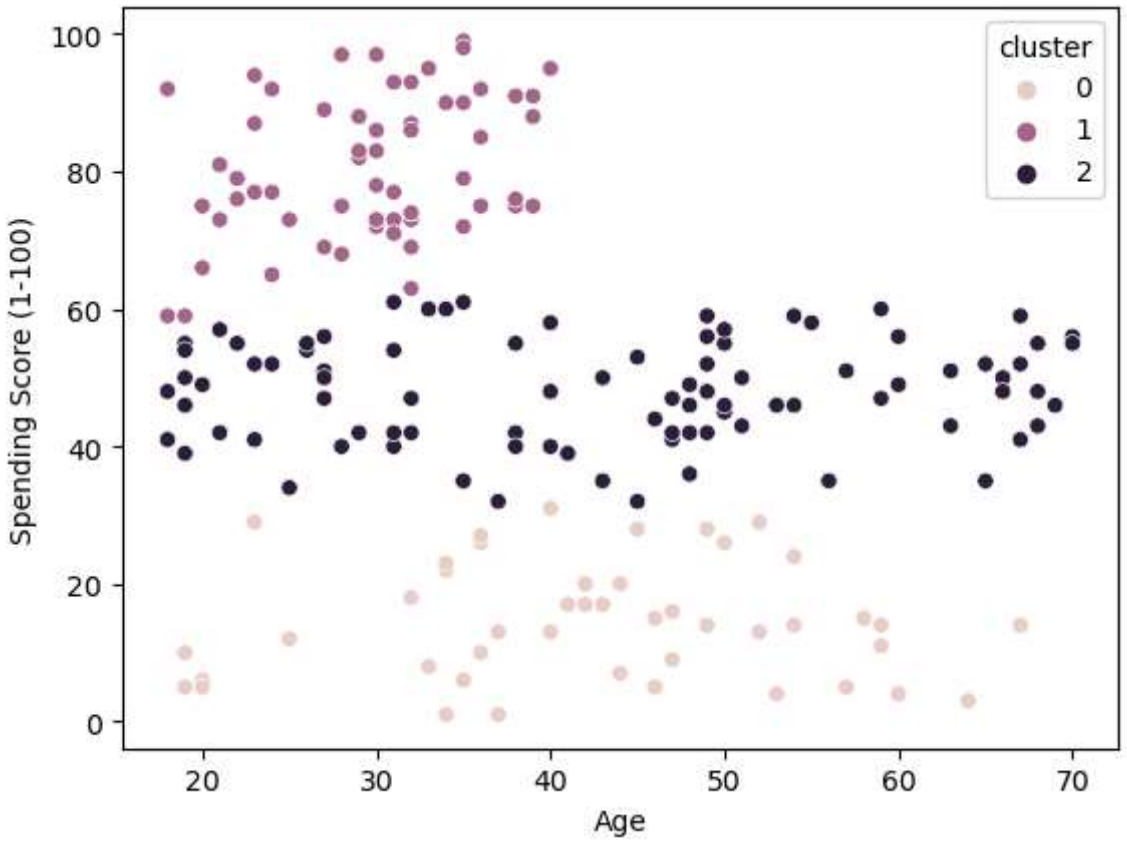
Out[18]:
```
array([14.59574468, 80.74193548, 47.78021978])
```

In [19]:
```
plt.figure(figsize=(16,7))
plt.scatter(df1["Age"],df1["Spending Score (1-100)"],marker="+",color="r")
plt.scatter(kc[:,0],kc[:,1],marker="o",color="b",s=200)
plt.xlabel("age")
plt.ylabel("spending score(1-100)")
plt.show()
```



In [20]:
```
sns.scatterplot(x="Age",y="Spending Score (1-100)",data=df1,hue="cluster")
```

Out[20]:
```
<Axes: xlabel='Age', ylabel='Spending Score (1-100)'>
```

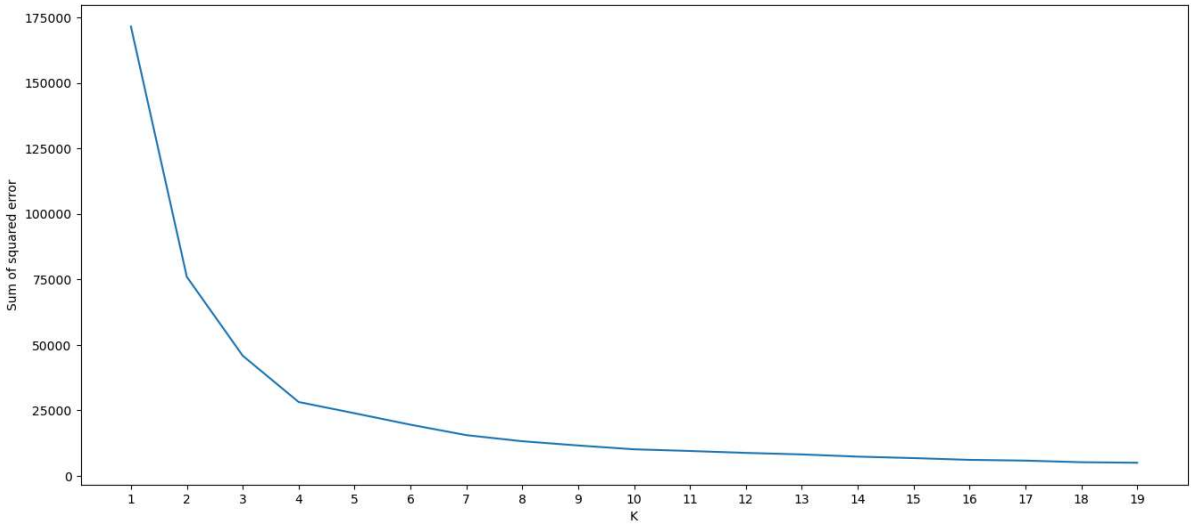


In [21]:
```
sse = []
k_rng = range(1,20)
for k in k_rng:
    km = KMeans(n_clusters=k)
    km.fit(df[['Age','Spending Score (1-100)']])
    sse.append(km.inertia_)
```

In [22]:
```
sse
```

Out[22]:
```
[171535.50000000003,
 75949.15601023019,
 45840.67661610866,
 28165.583566629342,
 23872.69755069491,
 19498.41264031118,
 15514.19313435103,
 13180.954546618257,
 11562.064417577105,
 10104.02786557042,
 9465.75771173271,
 8734.883814333814,
 8150.583848640466,
 7310.53087244999,
 6746.278380361469,
 6075.033123537111,
 5785.191693274044,
 5177.55992063492,
 4970.068596681097]
```

In [23]:
```
plt.figure(figsize=(16,7))
plt.xlabel('K')
plt.ylabel('Sum of squared error')
plt.plot(k_rng,sse)
plt.xticks(range(1,20))
plt.show()
```



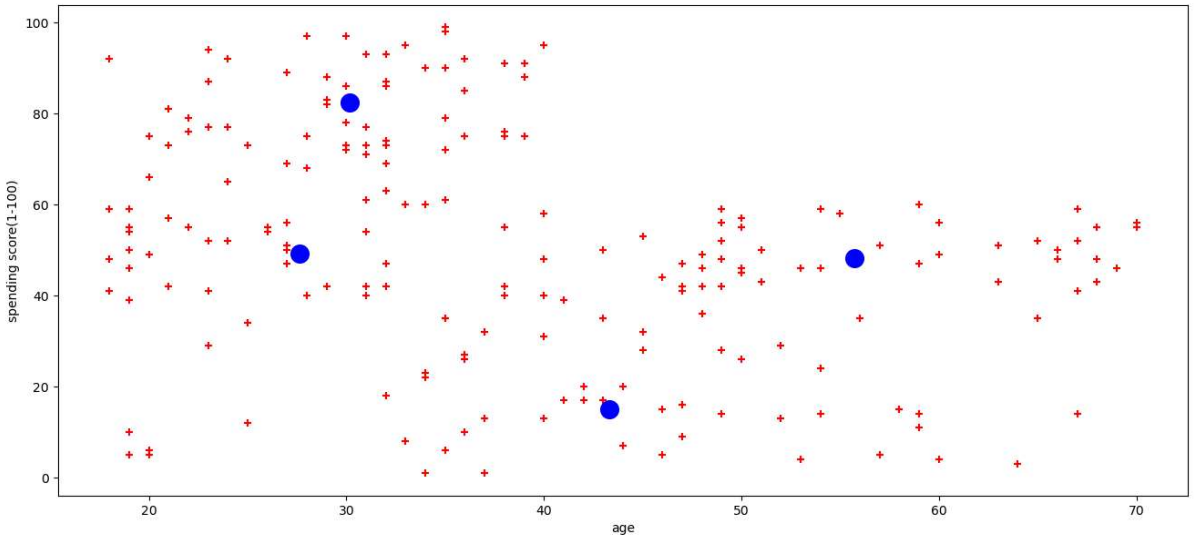
In [24]:
```
from sklearn.cluster import KMeans
km=KMeans(n_clusters=4,random_state=42)
y_pred=km.fit_predict(df[["Age","Spending Score (1-100)"]])
y_pred
```

Out[24]:
```
array([3, 1, 2, 1, 3, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 3, 3, 2, 1, 3, 1,
       2, 1, 2, 1, 2, 3, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 0, 1, 0, 3,
       2, 3, 0, 3, 3, 3, 0, 3, 3, 0, 0, 0, 0, 0, 3, 0, 0, 3, 0, 0, 0, 3,
       0, 0, 3, 3, 0, 0, 0, 0, 0, 3, 0, 3, 3, 0, 0, 3, 0, 0, 3, 0, 0, 3,
       3, 0, 0, 3, 0, 3, 3, 3, 0, 3, 0, 3, 3, 0, 0, 3, 0, 3, 0, 0, 0, 0,
       0, 3, 3, 3, 3, 3, 0, 0, 0, 0, 3, 3, 3, 1, 3, 1, 0, 1, 2, 1, 2, 1,
       3, 1, 2, 1, 2, 1, 2, 1, 2, 1, 3, 1, 2, 1, 0, 1, 2, 1, 2, 1, 2, 1,
       2, 1, 2, 1, 2, 1, 0, 1, 2, 1, 2, 1, 2, 1, 2, 3, 2, 1, 2, 1, 2, 1,
       2, 1, 2, 1, 2, 1, 2, 1, 3, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, 1,
       2, 1])
```

In [25]:
```
kc=km.cluster_centers_
kc
```

Out[25]:
```
array([[55.70833333, 48.22916667],
       [30.1754386 , 82.35087719],
       [43.29166667, 15.02083333],
       [27.61702128, 49.14893617]])
```

In [26]:
```
plt.figure(figsize=(16,7))
plt.scatter(df1["Age"],df1["Spending Score (1-100)"],marker="+",color="r")
plt.scatter(kc[:,0],kc[:,1],marker="o",color="b",s=200)
plt.xlabel("age")
plt.ylabel("spending score(1-100)")
plt.show()
```



In [27]:
```
km.predict([[39,40]])
```

Out[27]:
```
array([3])
```

In [ ]: