3.3_7

2021 年 12 月 20 日

1 西安酒店聚类分析

[15]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
from sklearn.cluster import KMeans # 导入 K 均值聚类算法
import pylab as mpl  # 导入中文字体，避免显示乱码
mpl.rcParams['font.sans-serif']=['SimHei']  # 设置为黑体字

poi_gpd=pd.read_pickle('../data/poiAll_gpd.pkl') # 读取已经存储为.pkl 格式的 POI 数据，其中包括 geometry 字段，为 GeoDataFrame 地理信息数据，可以通过 poi_gpd.plot() 迅速查看数据。

df = poi_gpd.reset_index()
df = df[df.level_0 == 'poi_1_hotel']
df = df.dropna(subset = ['detail_info_price','detail_info_overall_rating'],axis=0) # 删除缺省值
df.head()
```

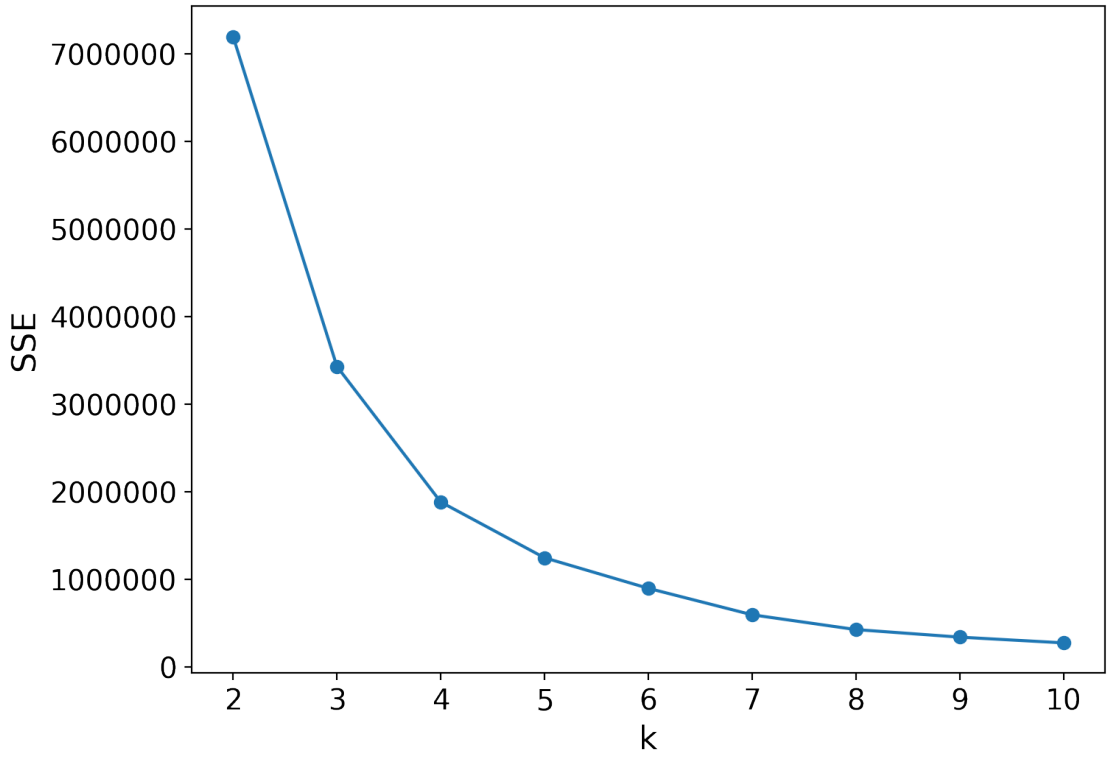
[15]:

	level_0	level_1	name	location_lat	location_lng
11579	poi_1_hotel	1191	志诚丽柏酒店	34.240030	108.912124
11580	poi_1_hotel	1194	水晶岛酒店	34.213837	108.893900
11581	poi_1_hotel	1195	西安高新希尔顿酒店	34.226686	108.894191
11582	poi_1_hotel	1196	西安海升酒店	34.218452	108.891532
11583	poi_1_hotel	1197	西安天骊君廷大酒店	34.224739	108.919048

	detail_info_tag	detail_info_overall_rating	detail_info_price	\

```
11579          酒店;星级酒店                    4.6                376
11580            酒店;其他                  4.4                299
11581          酒店;星级酒店                    4.6                614
11582            酒店;其他                  4.6                264
11583          酒店;快捷酒店                    4.8                655

                          geometry
11579  POINT (108.91212 34.24003)
11580  POINT (108.89390 34.21384)
11581  POINT (108.89419 34.22669)
11582  POINT (108.89153 34.21845)
11583  POINT (108.91905 34.22474)
```

[2]:
```
# 手肘法看 k 值
d=[]
for i in range(2,11):    #k 取值 1~10, 做 kmeans 聚类, 看不同 k 值对应的簇内误差平方和
    km=KMeans(n_clusters=i)
    km.fit(df[['detail_info_price','detail_info_overall_rating']])
    d.append(km.inertia_)  #inertia 簇内误差平方和

# 生成 figure 对象
plt.figure(figsize = (8,6), dpi = 200)
plt.plot(range(2,11),d,marker='o')
plt.xlabel('k',fontsize = 16)
plt.ylabel('SSE',fontsize = 16)
plt.xticks(fontsize = 14)
plt.yticks(fontsize = 14)
plt.show()
```



[3]:
```
# K-means 聚类
k = 6
km=KMeans(n_clusters=k)
km.fit(df[['detail_info_price','detail_info_overall_rating']])
df['k_clusters'] = km.labels_
df.head()
```

[3]:

	level_0	level_1	name	location_lat	location_lng	\
11579	poi_1_hotel	1191	志诚丽柏酒店	34.240030	108.912124	
11580	poi_1_hotel	1194	水晶岛酒店	34.213837	108.893900	
11581	poi_1_hotel	1195	西安高新希尔顿酒店	34.226686	108.894191	
11582	poi_1_hotel	1196	西安海升酒店	34.218452	108.891532	
11583	poi_1_hotel	1197	西安天骊君廷大酒店	34.224739	108.919048	

	detail_info_tag	detail_info_overall_rating	detail_info_price	\
11579	酒店;星级酒店	4.6	376	
11580	酒店;其他	4.4	299	

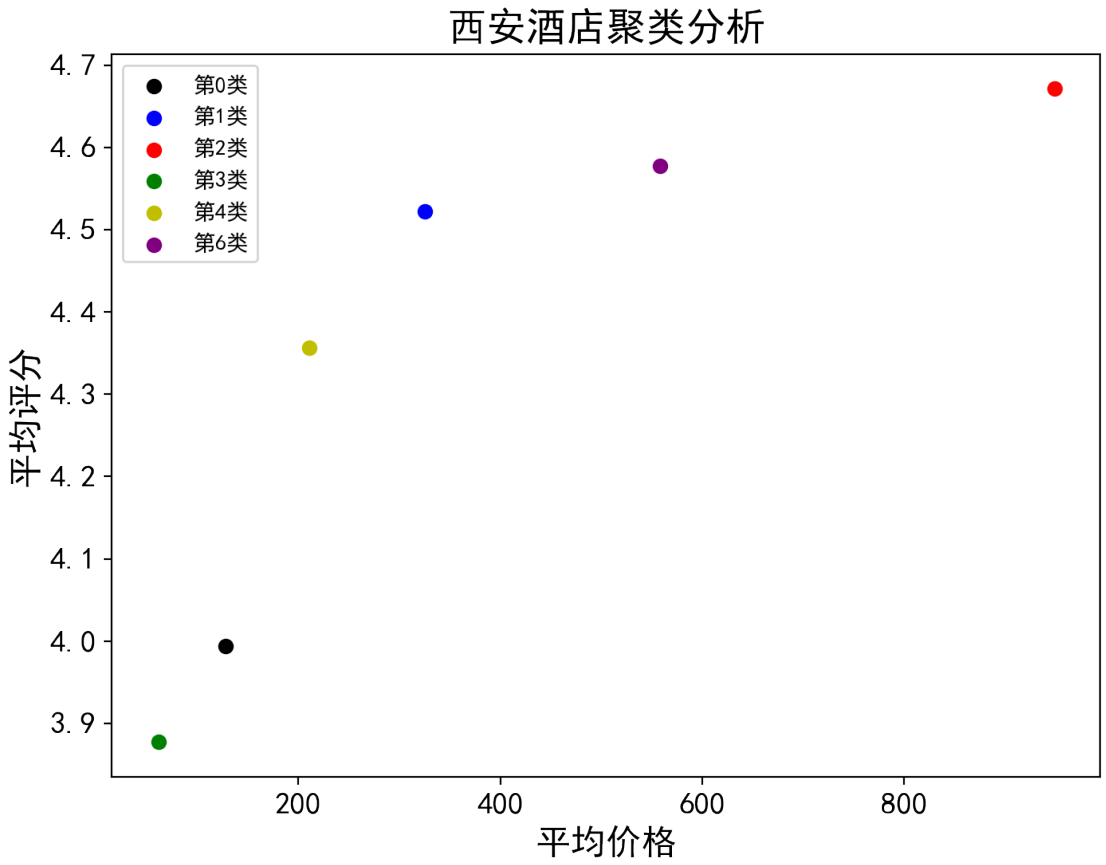
```
11581          酒店;星级酒店                          4.6                614
11582            酒店;其他                        4.6                264
11583          酒店;快捷酒店                          4.8                655

                          geometry  k_clusters
11579  POINT (108.91212 34.24003)           1
11580  POINT (108.89390 34.21384)           1
11581  POINT (108.89419 34.22669)           5
11582  POINT (108.89153 34.21845)           4
11583  POINT (108.91905 34.22474)           5
```

[10]:
```
price = []
rating = []
for i in range(0,k):
    price_mean = df[df.k_clusters == i]['detail_info_price'].mean()
    rating_mean = df[df.k_clusters == i]['detail_info_overall_rating'].mean()
    price.append(price_mean)
    rating.append(rating_mean)
    print('第{}类: 平均价格为 {},  平均评分为 {}'.
 ↪format(i,round(price_mean,2),round(rating_mean,2)))
```

```
第 0 类: 平均价格为 128.04,  平均评分为 3.99
第 1 类: 平均价格为 325.52,  平均评分为 4.52
第 2 类: 平均价格为 949.57,  平均评分为 4.67
第 3 类: 平均价格为 61.52,  平均评分为 3.88
第 4 类: 平均价格为 210.91,  平均评分为 4.36
第 5 类: 平均价格为 558.08,  平均评分为 4.58
```

[21]:
```
# 生成 figure 对象
labels = ['第 0 类','第 1 类','第 2 类','第 3 类','第 4 类','第 6 类']
colors = ['black','blue','red','green','y','purple']
plt.figure(figsize = (8,6), dpi = 200)
for i in range(0,k):
    plt.scatter(price[i], rating[i], marker='o',c=colors[i],label = labels[i])
plt.xlabel('平均价格',fontsize = 16)
plt.ylabel('平均评分',fontsize = 16)
plt.title('西安酒店聚类分析',fontsize = 18)
```

```
plt.legend()
plt.xticks(fontsize = 14)
plt.yticks(fontsize = 14)
plt.show()
```



[ ]: