3.3_6

2021 年 12 月 20 日

1 西安餐饮聚类分析

[21]:
```
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_0_delicacy']
df = df.dropna(subset = ['detail_info_price','detail_info_overall_rating'],axis=0) # 删除缺省值
df.head()
```

[21]:

	level_0	level_1	name	location_lat	location_lng	\
23	poi_0_delicacy	2787	百姓厨房（高新店）	34.239950	108.908171	
24	poi_0_delicacy	2788	陕西巷子老菜馆（高新店）	34.241850	108.911848	
25	poi_0_delicacy	2789	莲花餐饮（高新店）	34.224543	108.903540	
26	poi_0_delicacy	2790	苏福记（紫薇臻品店）	34.243518	108.886904	
27	poi_0_delicacy	2791	大龙燚火锅（高新店）	34.241661	108.912056	

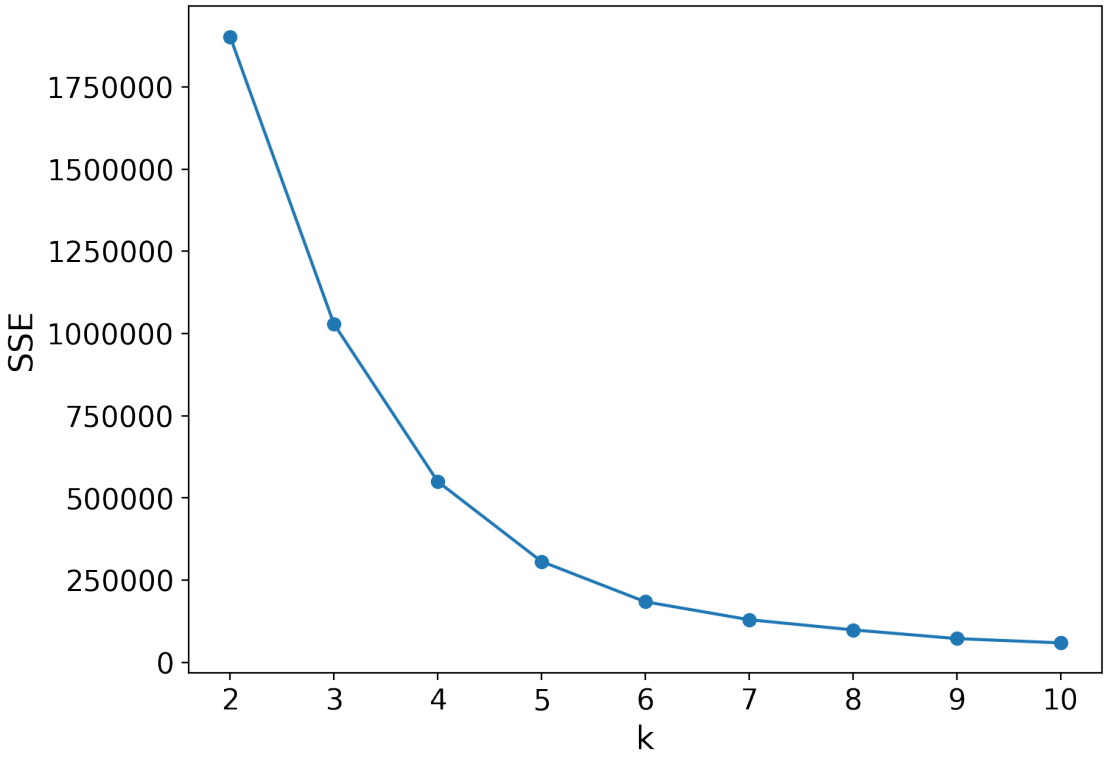
	detail_info_tag	detail_info_overall_rating	detail_info_price	\

23	美食;中餐厅	4.5	59
24	美食;中餐厅	4.7	64
25	美食;中餐厅	4.3	76
26	美食;中餐厅	4.6	44.5
27	美食;中餐厅	4.4	106

	geometry
23	POINT (108.90817 34.23995)
24	POINT (108.91185 34.24185)
25	POINT (108.90354 34.22454)
26	POINT (108.88690 34.24352)
27	POINT (108.91206 34.24166)

[9]:
```
# 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()
```



[17]:
```
# K-means 聚类
k = 4
km=KMeans(n_clusters=k)
km.fit(df[['detail_info_price','detail_info_overall_rating']])
df['k_clusters'] = km.labels_
df.head()
```

[17]:

	level_0	level_1	name	location_lat	location_lng	\
23	poi_0_delicacy	2787	百姓厨房（高新店）	34.239950	108.908171	
24	poi_0_delicacy	2788	陕西巷子老菜馆（高新店）	34.241850	108.911848	
25	poi_0_delicacy	2789	莲花餐饮（高新店）	34.224543	108.903540	
26	poi_0_delicacy	2790	苏福记（紫薇臻品店）	34.243518	108.886904	
27	poi_0_delicacy	2791	大龙燚火锅（高新店）	34.241661	108.912056	

	detail_info_tag	detail_info_overall_rating	detail_info_price	\
23	美食;中餐厅	4.5	59	
24	美食;中餐厅	4.7	64	

25	美食;中餐厅	4.3	76
26	美食;中餐厅	4.6	44.5
27	美食;中餐厅	4.4	106

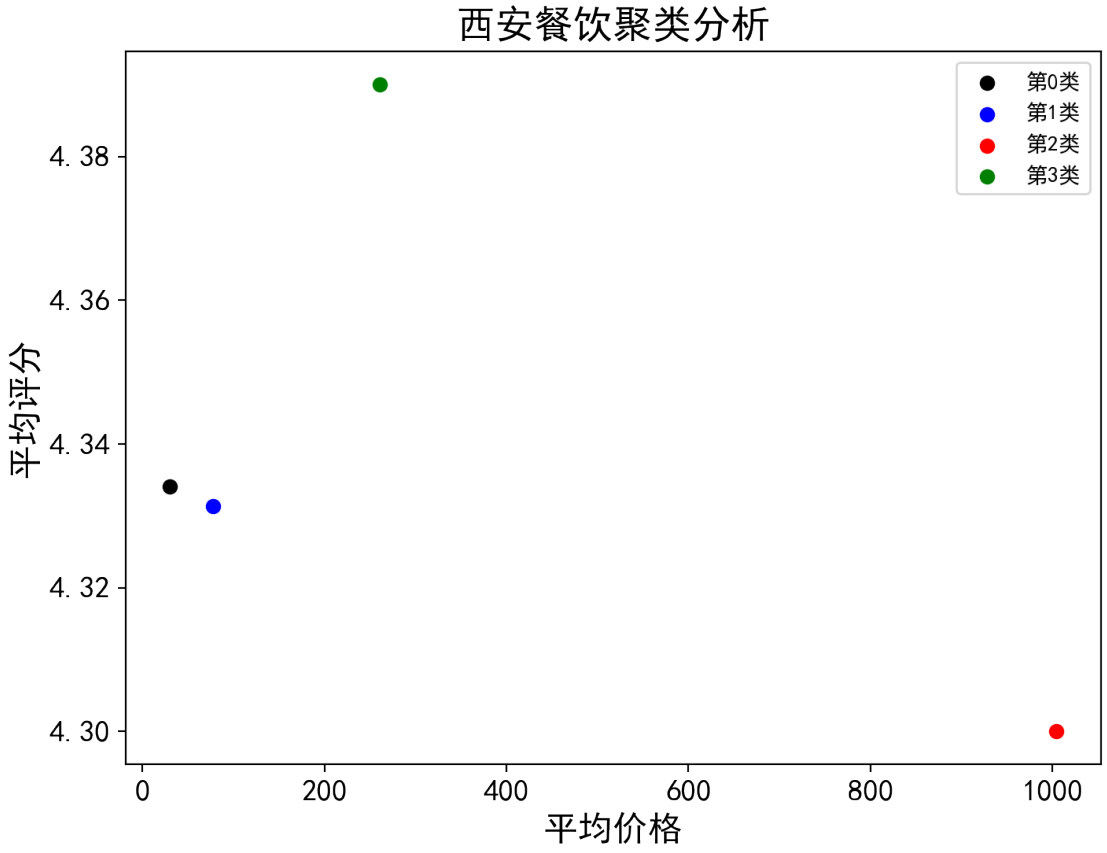
	geometry	k_clusters
23	POINT (108.90817 34.23995)	1
24	POINT (108.91185 34.24185)	1
25	POINT (108.90354 34.22454)	1
26	POINT (108.88690 34.24352)	0
27	POINT (108.91206 34.24166)	1

[19]:
```
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 30.08, 平均评分为 4.33
第 1 类: 平均价格为 77.91, 平均评分为 4.33
第 2 类: 平均价格为 1004.0, 平均评分为 4.3
第 3 类: 平均价格为 261.0, 平均评分为 4.39
```

[22]:
```
# 生成 figure 对象
labels = ['第 0 类','第 1 类','第 2 类','第 3 类','第 4 类','第 5 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('西安餐饮聚类分析',fontsize = 18)
plt.legend()
plt.xticks(fontsize = 14)
```

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



[ ]: