01检查原始数据是否具有较强的相关性

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
```

In [2]:

```
origin= pd.read_excel("城市投资潜力.xlsx",encoding="ANSI")
```

In [3]:

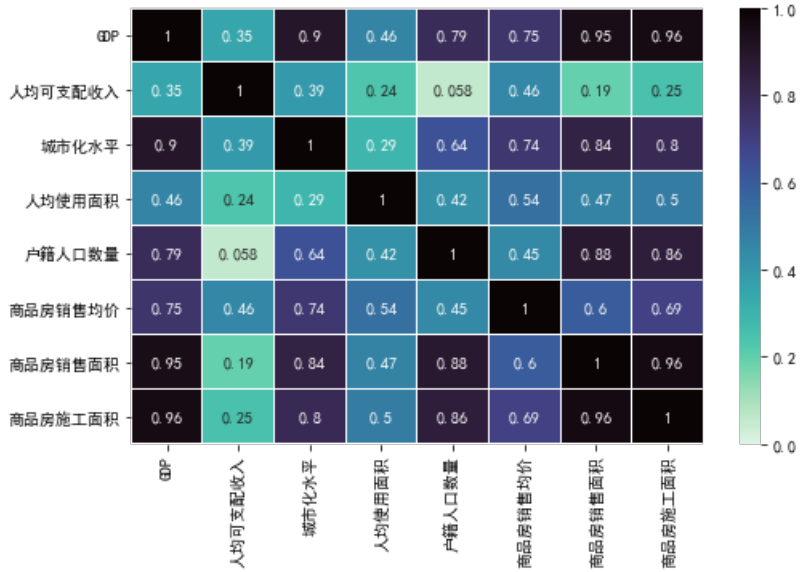
```
origin.head(3)
```

Out[3]:

	城市	GDP	人均可支配收入	城市化水平	人均使用面积	户籍人口数量	商品房销售均价	商品房销售面积	商品房施工面积
0	深圳	5684.39	22567	83.7	19.7	196.83	9230	784.63	3122.10
1	北京	9006.20	21989	84.3	20.1	1213.30	8792	2176.60	10438.60
2	杭州	3440.99	19027	62.1	21.0	666.31	7751	762.50	4545.33

In [15]:

```
#相关系数矩阵
plt.figure(figsize=(8,5))
import seaborn as sns
correlations = origin.iloc[:,1:].corr()
correction=abs(correlations)# 取绝对值，只看相关程度 ， 不关心正相关还是负相关
# plot correlation matrix
ax = sns.heatmap(correction,cmap='mako_r', linewidths=0.05,vmax=1, vmin=0 ,annot=True,annot_kws={'size'
:10,'weight':'bold'})
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus']=False
```



小结：从上图的相关系数矩阵可看出各变量之间存在相关性，适合进行主成分降维

02对相关系数矩阵进行特征值分解得到主成分方差和主成分系数

In [5]:

```
#标准化
st = StandardScaler().fit_transform(origin.iloc[:,1:])
```

```
B:\anaconda\lib\site-packages\sklearn\preprocessing\data.py:625: DataConversionWarning: Data with inpu
t dtype int64, float64 were all converted to float64 by StandardScaler.
  return self.partial_fit(X, y)
B:\anaconda\lib\site-packages\sklearn\base.py:462: DataConversionWarning: Data with input dtype int64,
float64 were all converted to float64 by StandardScaler.
  return self.fit(X, **fit_params).transform(X)
```

In [6]:

```
df = pd.DataFrame(st)
```

In [7]:

```
df.columns = origin.columns[1:]
```

In [8]:

```
#变量方差占比
pca = PCA(n_components=6)
principalComponents = pca.fit_transform(df)
pca.explained_variance_ratio_
```

Out[8]:

```
array([0.66509037, 0.17130862, 0.09869093, 0.03202314, 0.02046742,
       0.00842093])
```

In [9]:

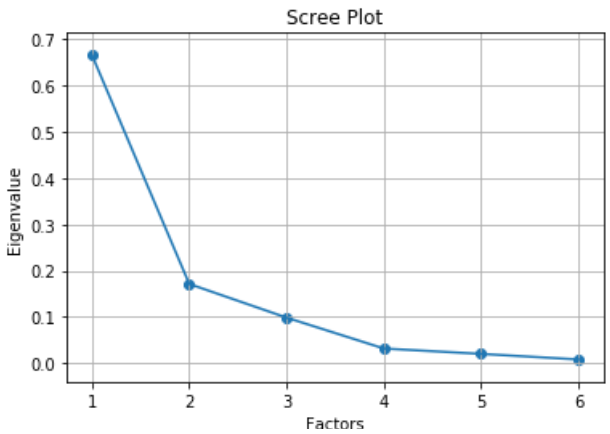
```
#前两个主成分的方差累计贡献率
0.66509037+0.17130862
```

Out[9]:

```
0.8363989900000001
```

In [10]:

```
# 进行可视化
importance = pca.explained_variance_ratio_
plt.scatter(range(1,7),importance)
plt.plot(range(1,7),importance)
plt.title('Scree Plot')
plt.xlabel('Factors')
plt.ylabel('Eigenvalue')
plt.grid()
plt.show()
```



小结：前两个主成分的方差累计贡献率为0.836，遂取两个主成分

03降维

In [11]:

```
# 进行降维,降到2维
pca = PCA(n_components=2)
principalComponents = pca.fit_transform(df)
# 查看降维后的数据
principalDf = pd.DataFrame(data=principalComponents, columns=['principal component 1', 'principal compo
nent 2'])
finalDf = pd.concat([principalDf, origin[['城市']]], axis = 1)
finalDf.head(5)
```

Out[11]:

	principal component 1	principal component 2	城市
0	1.903036	1.949203	深圳
1	5.294752	0.700785	北京
2	0.900859	0.043277	杭州
3	6.072103	-0.263522	上海
4	1.979812	0.025369	广州

In [12]:

```
#查看转换系数
pca.components_
```

Out[12]:

```
array([[ 0.42568554,  0.13179163,  0.38628529, -0.2355548 ,  0.35927873,
         0.34508358,  0.41497454,  0.41994362],
       [ 0.05991076,  0.77482091,  0.19584228,  0.47750208, -0.28389729,
         0.19738284, -0.08851747, -0.0480397 ]])
```

得到的主成分表达式为：

F1 = 0.426*GDP*+*0.132*人均可支配收入+0.386*城市化水平*-*0.236*人均使用面积+0.359*户籍人口数量*+ *0.345* 商品房销售均价+0.415*商品房销售面积*+*0.420*商品房施工面积

F2= 0.060*GDP*+*0.775*人均可支配收入+0.196*城市化水平*+*0.478*人均使用面积-0.284*户籍人口数量*+ *0.197* 商品房销售均价-0.089*商品房销售面积*-*0.048*商品房施工面积

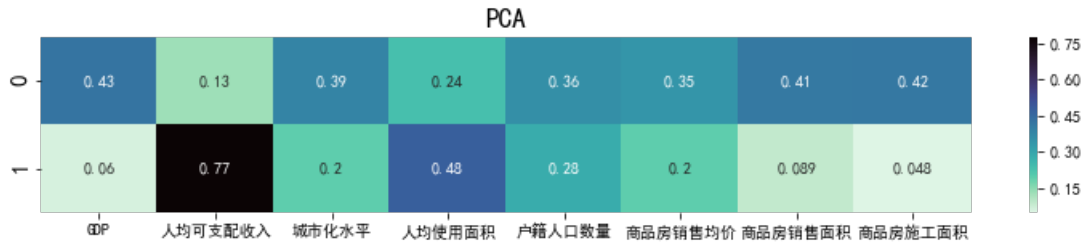
04分析主成分系数

In [13]:

```
# 对系数进行可视化
import seaborn as sns
df_cm = pd.DataFrame(np.abs(pca.components_), columns=origin.columns[1:])
plt.figure(figsize = (13,2))
ax = sns.heatmap(df_cm, annot=True,cmap="mako_r")
# 设置y轴的字体的大小
ax.yaxis.set_tick_params(labelsize=15)
ax.xaxis.set_tick_params(labelsize=10)
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus']=False
plt.title('PCA', fontsize='xx-large')
# Set y-axis label
#plt.savefig('factorAnalysis.png', dpi=200)
```

Out[13]:

```
Text(0.5, 1.0, 'PCA')
```



小结：

上图可看出在第一主成分的组成中，系数较大的指标为：GDP、城市化水平、商品房销售面积和商品房施工面积，这四个指标都与城市发展水平紧密相关，因此在这里将第一主成分命名为“发展指数”

同样地，在第二主成分中，系数较大的指标为：人均可支配收入、人均使用面积，这两个指标都与人民生活水平紧密相关，因此在这里将第二主成分命名为“生活指数”

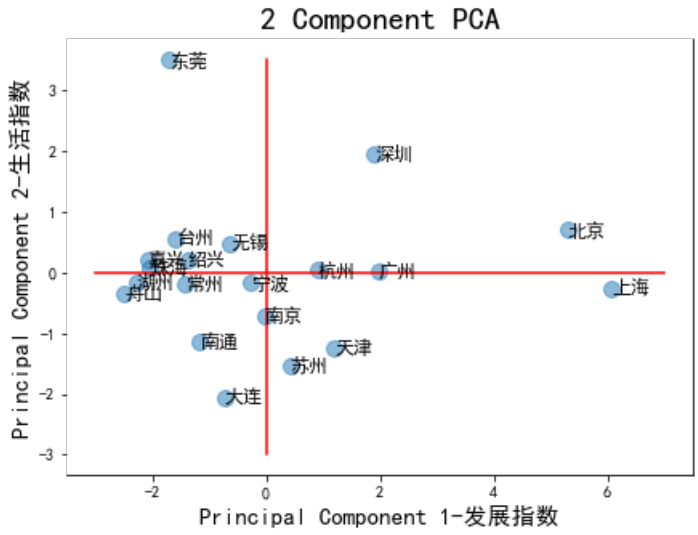
05观察各个城市在两个新建指数上的分布，寻找投资潜力

In [14]:

```
a = finalDf.iloc[:,0]
b = finalDf.iloc[:,1]
fig = plt.figure(figsize = (7,5))
ax = fig.add_subplot(1,1,1)
ax.set_xlabel('Principal Component 1-发展指数', fontsize = 15)
ax.set_ylabel('Principal Component 2-生活指数', fontsize = 15)
ax.set_title('2 Component PCA', fontsize = 20)
ax.scatter(a,b,s=100,alpha=0.5)
for i,j,z in zip(a,b,finalDf.iloc[:,2]):
    ax.text(x=i,y=j-0.1,s=z,fontsize=12)
ax.spines["top"].set_visible(False)
ax.spines["left"].set_visible(False)
ax.plot((0,0), (-3,3.5),'r')
ax.plot((-3,7), (0,0),'r')
```

Out[14]:

```
[<matplotlib.lines.Line2D at 0x17f57c8e780>]
```



小结：

1.综合来看，最有投资潜力的时深圳和北京，二者的生活指数与发展指数均较高，说明两座城市建设得较为繁荣，居民也较为富裕，适合投资

2.其次，若投资更看重居民生活水平得服务业、餐饮业等行业，可在发展指数方面做一些让步，转而选择东莞、台州、无锡、嘉兴、绍

兴这样的城市；反之，若投资大型建设型产业，与居民日常生活相对不甚紧密的行业可选择上海、广州、杭州、天津这样的城市

In [ ]: