各城市投资潜力

罗亭轩

资料介绍

某房地产开发公司拟在长江、珠江三角洲及环渤海地区中选择具有投资潜力的目标城市进行投资。 此数据报含了下列各种指标：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。

资料探索

叙述统计量

由下报表可以看出 20 个不同城市 8 项城市指标的分布情形。

```
##       城市          GDP           人均可支配收入     城市化水平
##  上海    : 1   Min.   :  333.2   Min.   :13350   Min.   :52.00
##  大连    : 1   1st Qu.: 1536.9   1st Qu.:17290   1st Qu.:57.42
##  天津    : 1   Median : 2699.3   Median :18009   Median :62.10
##  北京    : 1   Mean   : 3371.8   Mean   :18333   Mean   :64.73
##  台州    : 1   3rd Qu.: 4458.4   3rd Qu.:19718   3rd Qu.:69.62
##  舟山    : 1   Max.   :10297.0   Max.   :25320   Max.   :88.70
##  (Other):14
##    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
##  Min.   :16.50   Min.   :  92.63   Min.   :4134   Min.   :  91.8
##  1st Qu.:20.00   1st Qu.: 316.10   1st Qu.:4454   1st Qu.: 364.4
##  Median :23.60   Median : 562.53   Median :4788   Median : 613.5
##  Mean   :24.93   Mean   : 552.14   Mean   :5440   Mean   : 801.4
##  3rd Qu.:28.65   3rd Qu.: 689.91   3rd Qu.:5714   3rd Qu.: 932.4
##  Max.   :45.20   Max.   :1368.10   Max.   :9230   Max.   :3025.4
##
##  商品房施工面积
##  Min.   :  300
##  1st Qu.: 1326
##  Median : 2198
##  Mean   : 3198
##  3rd Qu.: 4340
##  Max.   :10939
##
```

盒形图

将资料进行可视化以便判读。由下图 1 可以更清楚看出各个指标的分布，像是商品房的销售均价及施工面积，都呈现右偏，代表有少数城市的商品房售价及施共面积远大于其他城市，并且可以发现每个变量都存在少数离群值。

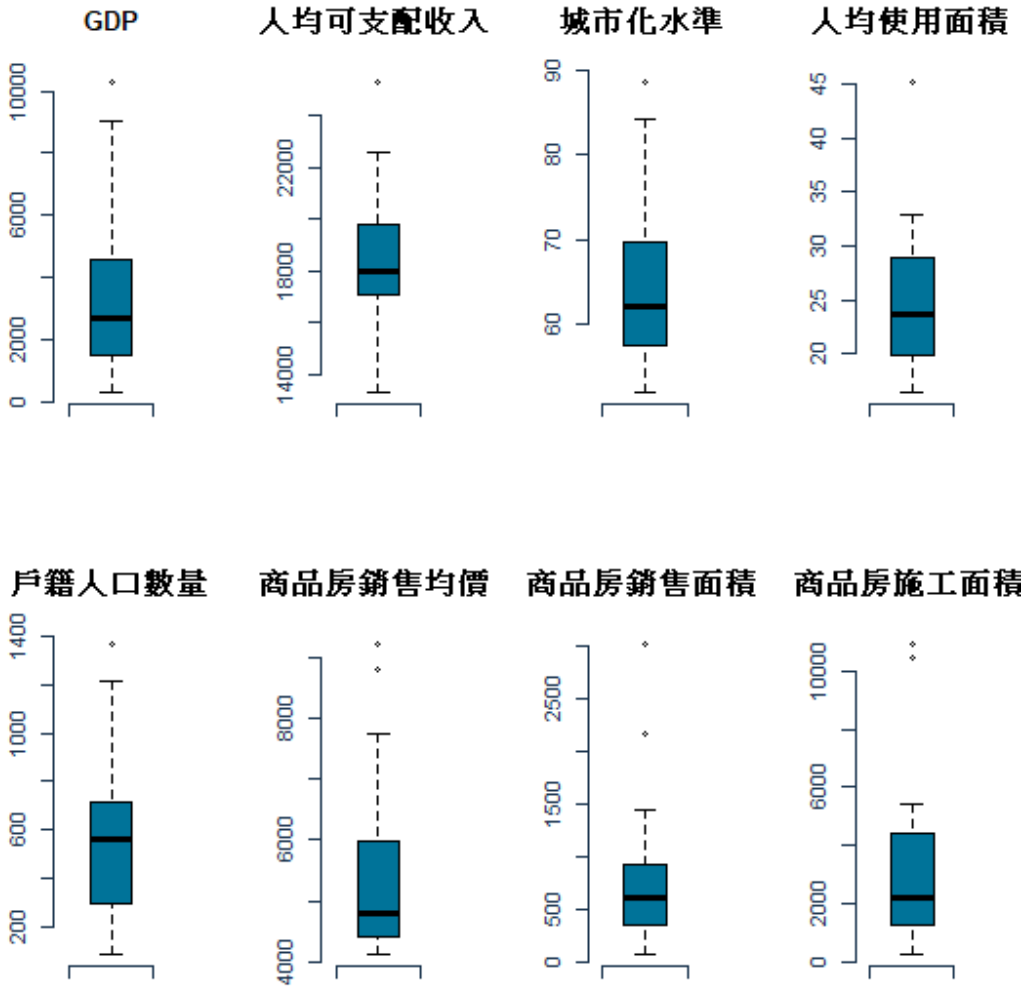

图 1 城市发展指标盒型图

散点图矩阵

接着探讨变量间的关系，由图 2 可以看出各种资指标间大部分呈现正相关，其中上三角为变量间相关性的强弱，不考虑正负。

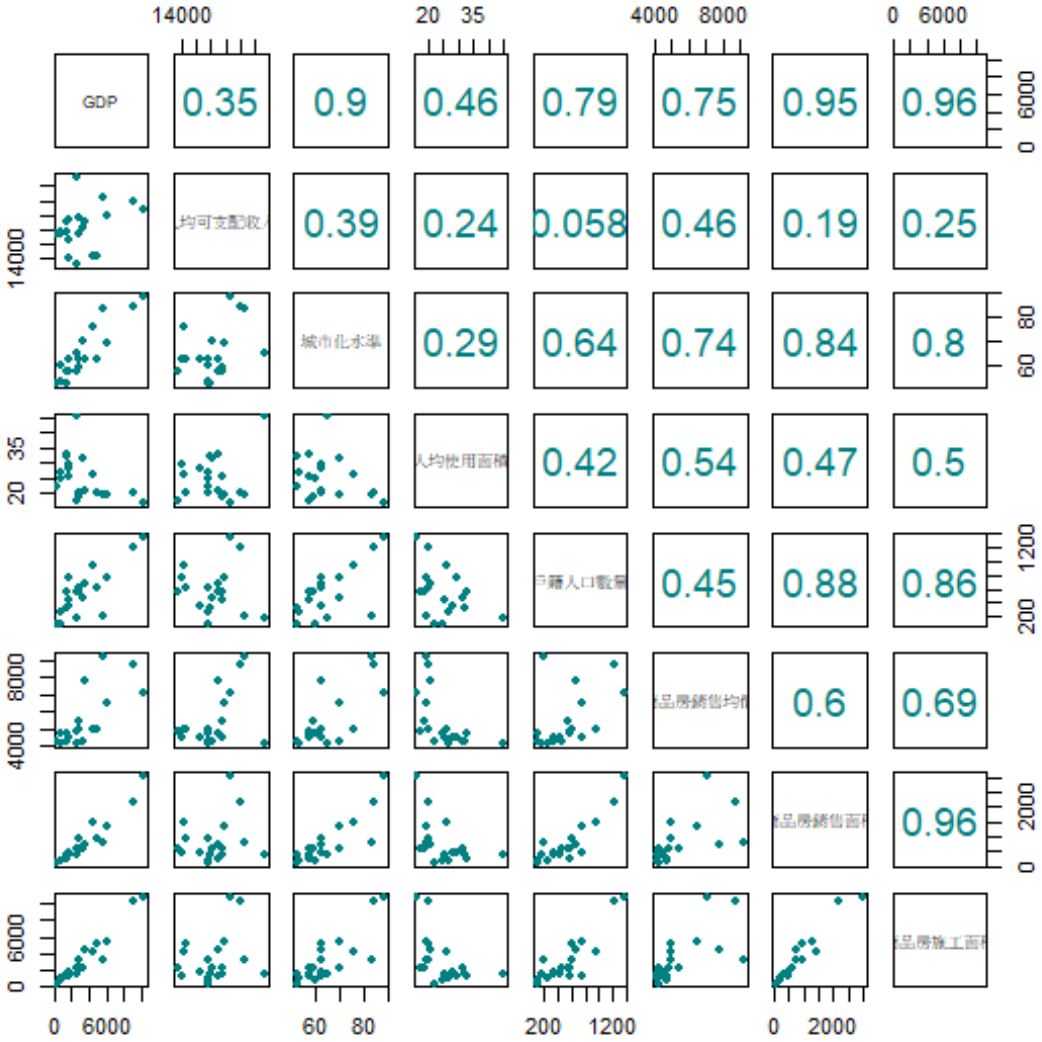

图 2 城市发展指标散点图矩阵

相关系数矩阵

再由相关系数矩阵，如下图 3 更可以清楚发现，人均使用面积与其他众指标呈现负相关。

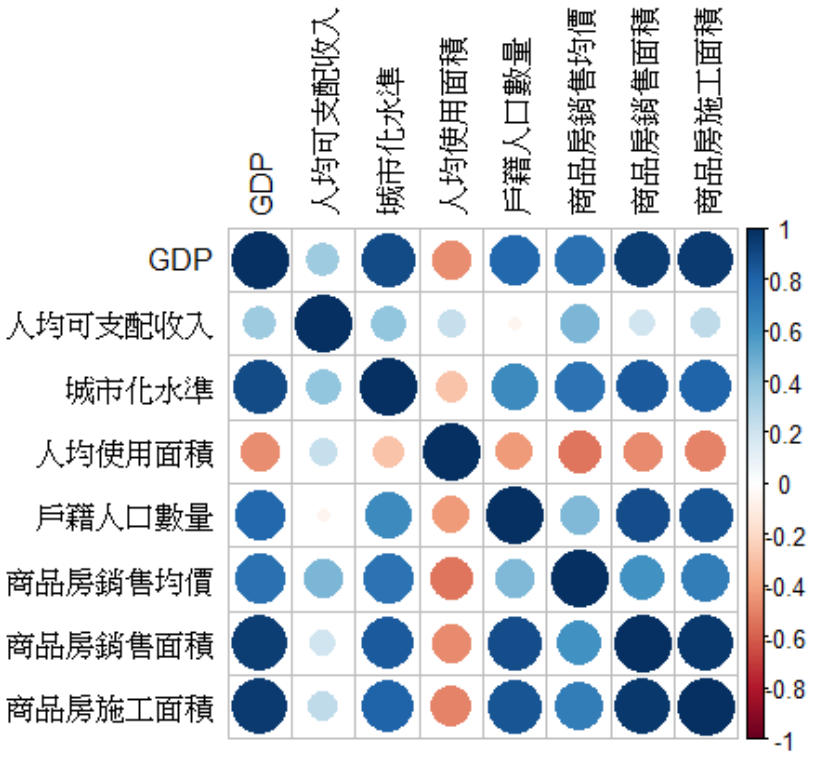

图 3 城市发展指标相关系数矩阵

主成分分析

先利用主成分分析将数据维度缩减至二维，以方便观察原始数据的分布情形。 又由探索性分析中的盒形图知数据中存在离群值，故将每个变量的离群值数值取出来。

在 8 个离群值中，有三笔数据同为上海、两笔为北京及东莞、一笔为深圳。

```
##
## 上海 北京 东莞 深圳
##    3    2    2    1
```

维度缩减后的第一主成分与第二组成分的累积贡献率为百分之 83.64，并将离群值在图中标出，数据分布情形结果如下图 4。

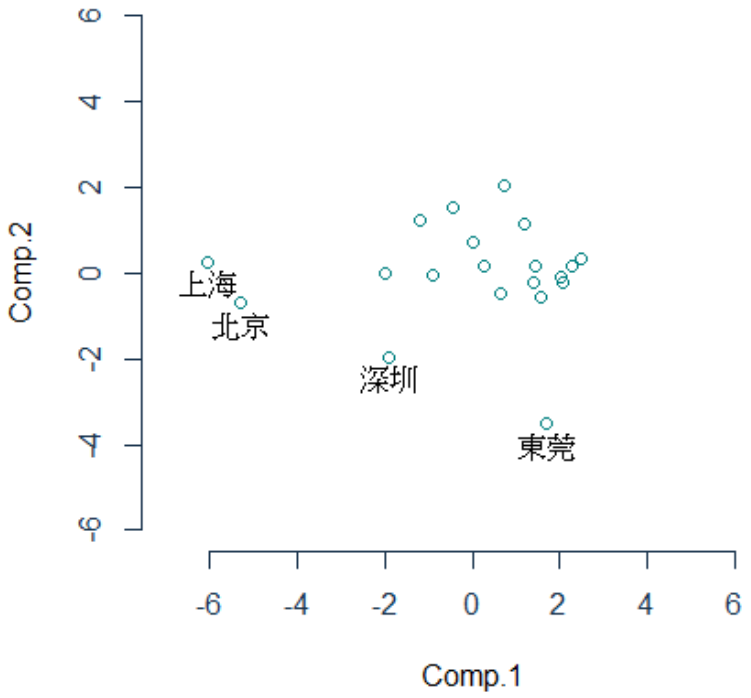

图 4 个城市分布情

选择主成分

为避免离群值影响主成分的选择，先占时将离群值移除，再进行主成分分析。然而，在第一主成分的系数中，仍然有负。

因此将人均使用面积与户籍人口数量相乘，命名为新的变量：面积。

```
##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## GDP             0.437                -0.242 -0.318  0.233  0.612  0.466
## 人均可支配收入         -0.653 -0.595 -0.233 -0.240 -0.303         -0.101
## 城市化水平      0.363  0.284 -0.326 -0.361  0.651 -0.234  0.165 -0.211
## 人均使用面积   -0.215  0.397 -0.724  0.269         0.416         0.131
## 户籍人口数量    0.384  0.240         0.578 -0.335 -0.533  0.130 -0.195
## 商品房销售均价  0.312 -0.502         0.555  0.492  0.205         0.238
## 商品房销售面积  0.438  0.131        -0.196 -0.127        -0.742  0.422
## 商品房施工面积  0.436                       -0.211  0.540 -0.149 -0.662
##
```

经过转换后的数据其第一主成分系数均为正，可当作综合指标，第一主成分的贡献率为百分之 61.44，其中将在选择主成分时，占时移除的离群点加回来，数据分布状况便如下图 5 所示。

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4    Comp.5
## Standard deviation     2.073722 1.1857338 0.77190247 0.63370176 0.4557141
## Proportion of Variance 0.614332 0.2008521 0.08511906 0.05736827 0.0296679
## Cumulative Proportion  0.614332 0.8151841 0.90030319 0.95767146 0.9873394

##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## GDP             0.459         0.153  0.217 -0.350 -0.546  0.539
## 人均可支配收入        -0.727 -0.577  0.291 -0.200
## 城市化水平       0.411  0.177 -0.301  0.433  0.688 -0.172 -0.139
## 商品房销售均价  0.307 -0.524  0.147 -0.619  0.444         0.169
## 商品房销售面积  0.464                0.193 -0.121  0.809  0.251
## 商品房施工面积  0.453 -0.150  0.301        -0.283        -0.771
## 面积            0.320  0.362 -0.657 -0.509 -0.269
##
```

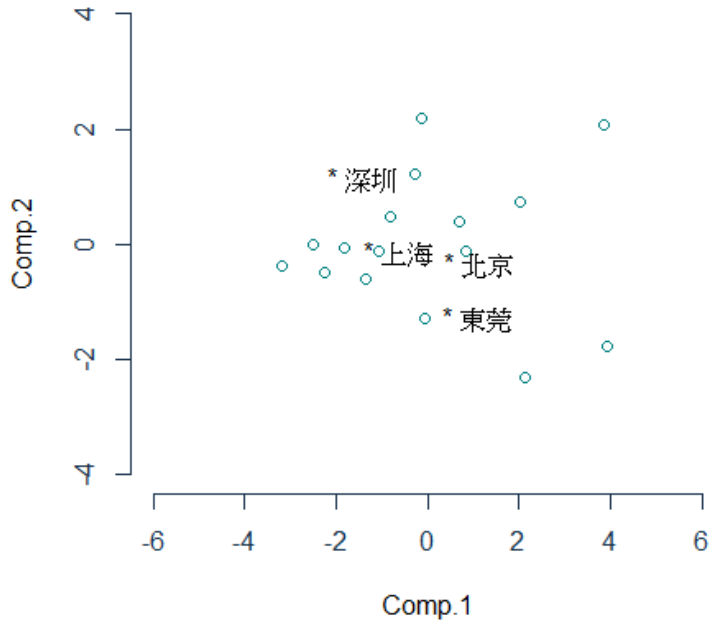

图 5 主成分得分

结论

经过移除离群值即便数转换后选择的第一主成份为：

$$F1 = (0.459 \quad 0.411 \quad 0.307 \quad 0.464 \quad 0.453 \quad 0.32) \begin{pmatrix} \text{GDP} \\ \text{城市化水准} \\ \text{商品房销售均价} \\ \text{人均使用面积} \\ \text{商品房施工面积} \\ \text{人均使用面积} * \text{户籍人口数量} \end{pmatrix}$$

最终投资的目标城市首选应为广州，其次为天津，较有机会获利。

```
invest <- sort(c(city.pca3$scores[, 1], out.comp1), decreasing = T)
city[as.numeric(names(invest)), 1]
```

```
##  [1] 广州 天津 杭州 苏州 北京 上海 南京 无锡 宁波 南通 大连 深圳 常州 台州
## [15] 绍兴 东莞 嘉兴 珠海 湖州 舟山
## 20 Levels: 上海 大连 天津 北京 台州 舟山 杭州 东莞 南京 南通 珠海 ... 苏州
```