北京高档酒店价格因素分析

数据分析实战又来啦，今天我们进行的是北京高档酒店的价格因素分析，话不多说，直接上代码。

1. 导入所需要的包

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import warnings
warnings.filterwarnings("ignore")
```

2. 读取文件

```
hotel=pd.read_csv('hoteldata.csv')
#将四项评分的平均分作为总体评分
hotel['总体评分']=(hotel['卫生评分']+hotel['服务评分']+hotel['设施评分']+hotel['位置评分'])/4
#2015之前的旧装修，2015之后的为新装修
hotel['装修新旧']=pd.cut(hotel['装修时间'],[0,2015,2019],labels=['旧装修','新装修'])
hotel.head()
```

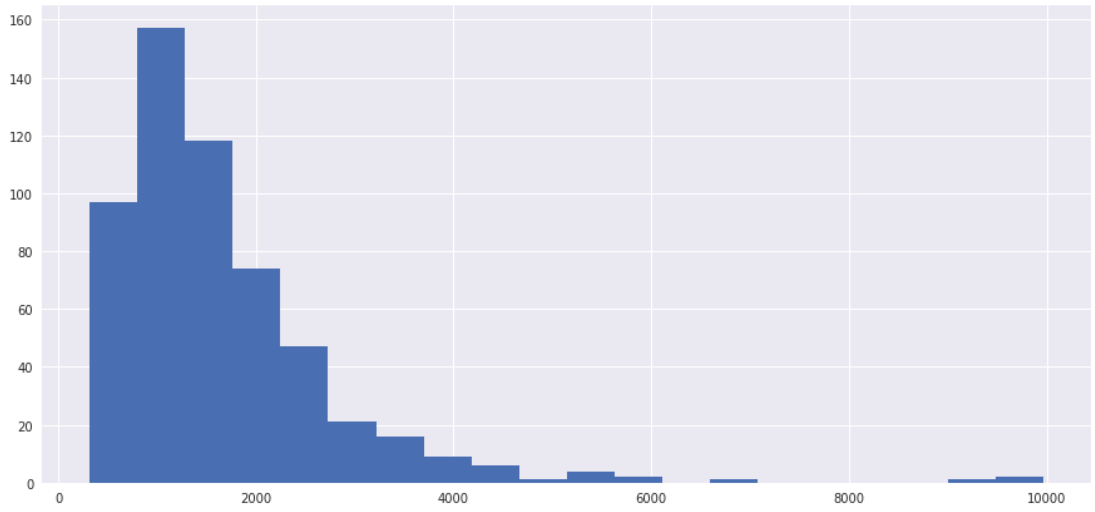
	酒店名称	地区	地址	卫生评分	服务评分	设施评分	位置评分	评价数	装修时间	房间类型	房价	经度	纬度	公司	出行住宿	校园生活	总体评分	装修新旧
0	北京朗丽兹西山花园酒店	海淀区	海淀永丰路与北清路十字路口往南600米路南	4.8	4.8	4.7	4.4	143	2014	豪华套间	9970	116.292419	40.095804	0	0	0	4.675	旧装修
1	北京钓鱼台国宾馆	海淀区	海淀阜成路2号	4.9	4.8	4.8	4.6	9	2013	豪华套间	9888	116.339444	39.928419	28	80	43	4.775	旧装修
2	北京颐和安缦酒店	海淀区	海淀颐和园宫门前街1号	4.7	4.6	4.4	4.4	104	2008	豪华套间	9269	116.288607	40.005692	2	18	5	4.525	旧装修
3	北京华尔道夫胡同四合院	东城区	东城金鱼胡同5-15号	5.0	5.0	4.5	4.5	7	2016	豪华套间	6777	116.420463	39.922276	33	185	35	4.750	新装修
4	北京颐和安缦酒店	海淀区	海淀颐和园宫门前街1号	4.7	4.6	4.4	4.4	104	2008	商务间	5813	116.288607	40.005692	2	18	5	4.525	旧装修

将各个酒店的情况进行评分，总体评分由卫生评分、服务评分、设施评分和位置评分构成，装修的新旧以装修时间来划分。

3 描述性统计分析

3.1 酒店房价分布直方图

```
price=hotel['房价']
plt.figure("hist",figsize=(15,7))
n, bins, patches = plt.hist(price, bins=20)
plt.show()
```



3.2 因变量数字特征

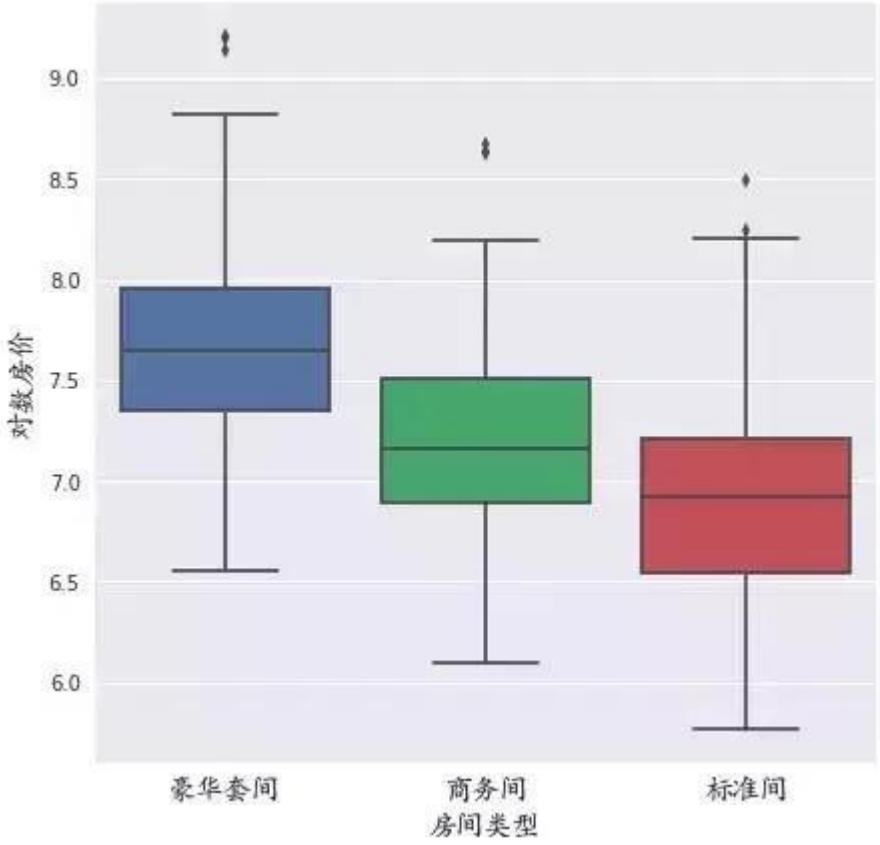
```
price=hotel['房价']
plt.figure("hist",figsize=(15,7))
n, bins, patches = plt.hist(price, bins=20)
plt.show()
```

```
1389.5
```

```
# 酒店房价平均值
hotel['房价'].mean()
1655.5125899280577
```

3.3 酒店因素箱型图

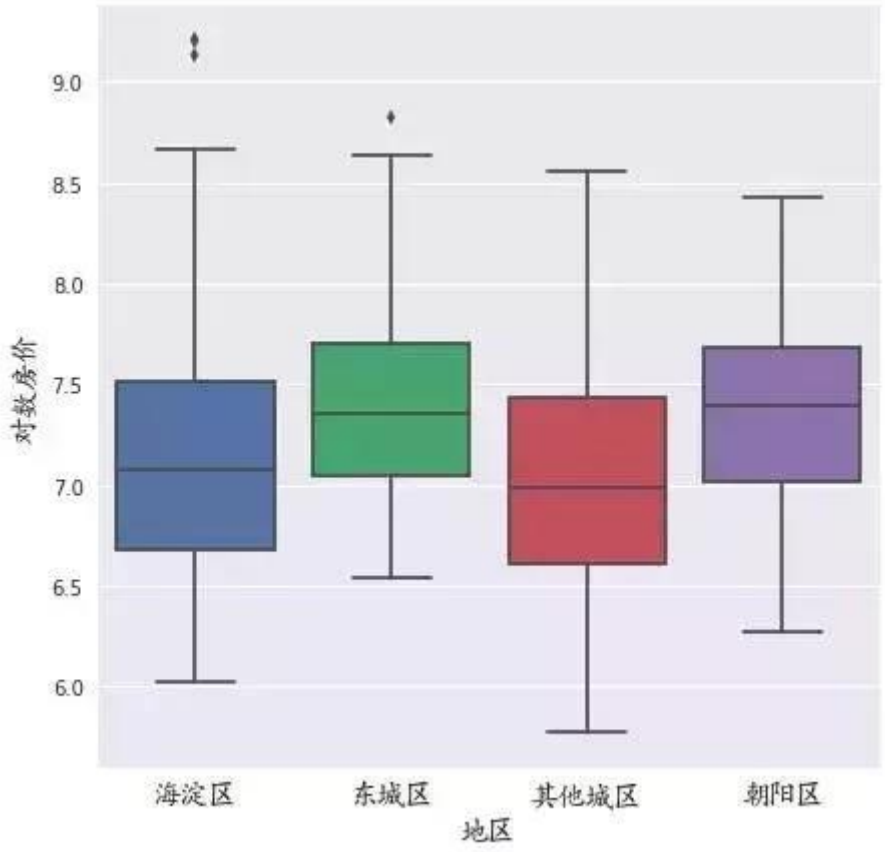
```
#酒店房间类型
hotel['对数房价']=np.log(hotel['房价'])
plt.figure(figsize=(7,7))
sns.boxplot(x='房间类型',y='对数房价',data=hotel)
```



符合一般的房价标准，按照标准间、商务间、豪华套间价格依次递增。

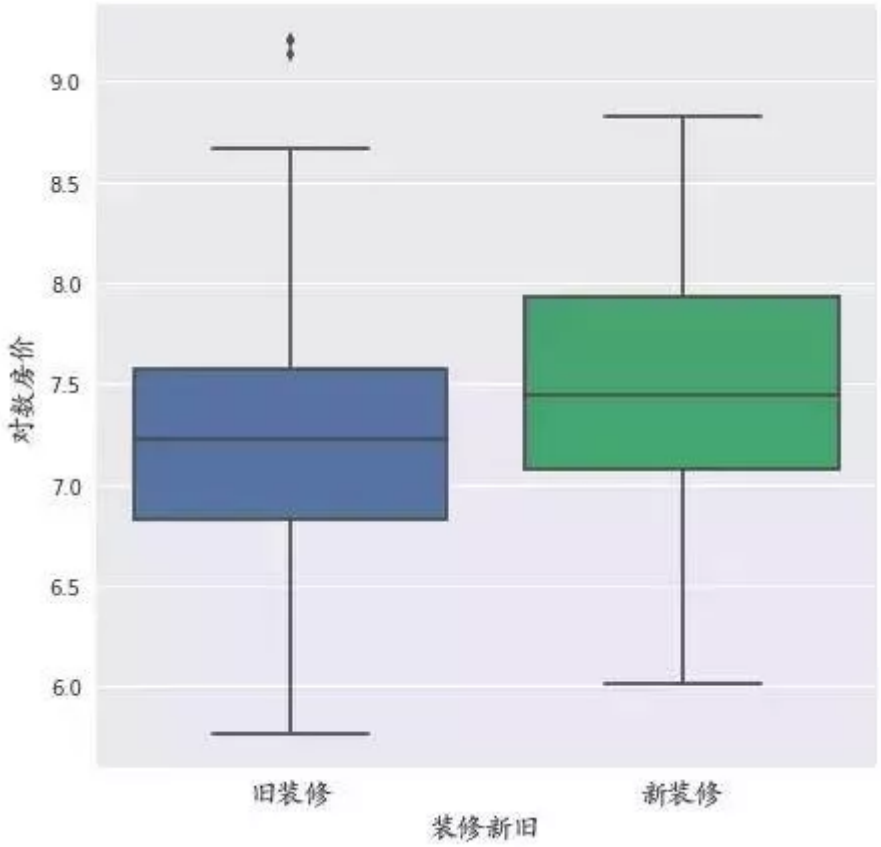
```
#酒店区域因素分析
```

```
plt.figure(figsize=(7,7))
sns.boxplot(x='地区',y='对数房价',data=hotel)
```



根据地区划分的箱型图展示，其中，东城区和朝阳区的房价最高，海淀区紧随其后。

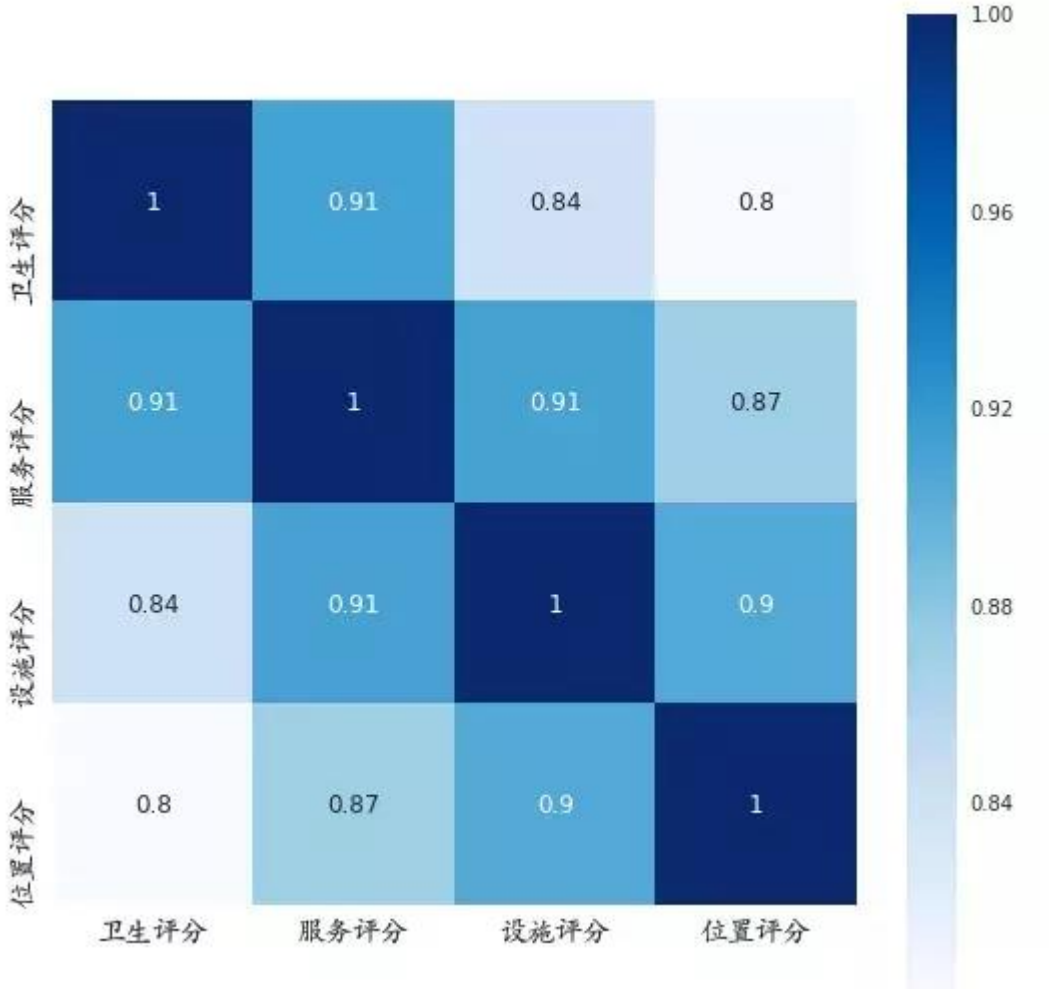
```
#酒店装修时间
hotel['对数房价']=np.log(hotel['房价'])
plt.figure(figsize=(7,7))
sns.boxplot(x='装修新旧',y='对数房价',data=hotel)
```



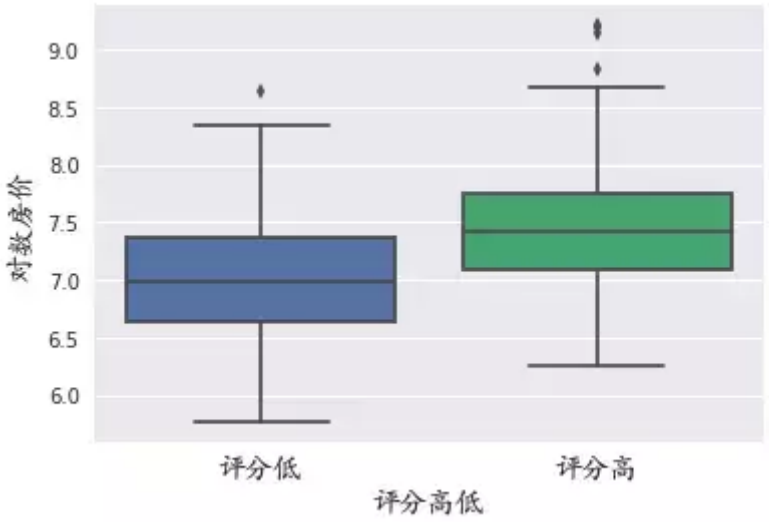
新装修的价格高于就旧装修，并且价格差异明显。

3.4 评分因素相关系数

```
grade=pd.DataFrame([hotel['卫生评分'],hotel['服务评分'],hotel['设施评分
'],hotel['位置评分']]).transpose()
correlation=grade.corr()
plt.subplots(figsize=(9, 9))  # 设置画面大小
sns.heatmap(correlation, annot=True, vmax=1, square=True, cmap="Blues")
```




```
#评分因素箱型图
hotel['评分分组']=pd.cut(hotel['总体评分'],[0,4.5,5.0],labels=['评分低','评分高'])
sns.boxplot(x='评分分组',y='对数房价',data=hotel)
```



评分高的房价高于评分低的房价。

4 对数线性回归模型

4.1 特征处理

```
#特征选择与处理
features=['地区','房间类型','装修新旧','总体评分','校园生活','公司','出行住宿']
X=hotel[features]
X['地区']=pd.get_dummies(X['地区'])
X['房间类型']=pd.get_dummies(X['房间类型'])
X['装修新旧']=pd.get_dummies(X['装修新旧'])

# 对特征进行归一化处理
from sklearn import preprocessing
X['总体评分']=preprocessing.scale(X['总体评分'])
X['校园生活']=preprocessing.scale(X['校园生活'])
X['公司']=preprocessing.scale(X['公司'])
X['出行住宿']=preprocessing.scale(X['出行住宿'])
```

4.2 模型拟合

```
from  sklearn  import  linear_model
model=linear_model.LinearRegression()
model.fit(X,y)
```

4.3 计算残差

```
np.mean(abs(model.predict(X)-y))
0.375942
```

4.4 查看模型拟合情况

```
import statsmodels.api  as  sm
est=sm.OLS(y,X).fit()
print(est.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   对数房价   R-squared:                       0.901
Model:                            OLS   Adj. R-squared:                  0.900
Method:                 Least Squares   F-statistic:                     712.4
Date:                Sat, 29 Dec 2018   Prob (F-statistic):          9.44e-271
Time:                        10:25:05   Log-Likelihood:                -1249.0
No. Observations:                 556   AIC:                             2512.
Df Residuals:                     549   BIC:                             2542.
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
地区             0.8242      0.307      2.683      0.008       0.221       1.428
房间类型         0.9906      0.201      4.930      0.000       0.596       1.385
装修新旧         6.7286      0.137     49.179      0.000       6.460       6.997
总体评分         0.3948      0.104      3.801      0.000       0.191       0.599
校园生活        -0.1958      0.106     -1.839      0.066      -0.405       0.013
公司            -0.0832      0.200     -0.417      0.677      -0.475       0.309
出行住宿        -0.0156      0.227     -0.069      0.945      -0.462       0.431
==============================================================================
Omnibus:                      257.645   Durbin-Watson:                   1.694
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              899.966
Skew:                           2.270   Prob(JB):                     3.76e-196
Kurtosis:                       7.271   Cond. No.                         5.73
==============================================================================
```