# 数据清洗案例

数据清洗是指把输入中的问题数据清洗掉，包括检查数据一致性，处理异常值、重复值、缺失值等，从而提高数据质量，便于后续的数据处理、可视化、模型构建等。

## 1. 导入相关包

In [65]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
#jupyter notebook运行这一行代码后在cell中显示图形
```

## 2.导入数据集

In [66]:
```
df=pd.read_csv('qunar_freetrip.csv',index_col=0)
```

In [67]:
```
df.head(3)
```

Out[67]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 哈尔滨 | 北海 | 2208.0 | 650.0 | 哈尔滨-北海3天2晚 \| 入住北海祥丰嘉年华大酒店 + 春秋航空往返机票 | 北海祥丰嘉年华大酒店 舒适型 4.7分/5分 | 标准双人间(双床) 双床 不含早 1间2晚 | 春秋航空 9C8741 | 直飞 | 17:10-21:50 | 春秋航空 9C8742 | 直飞 | 10:20-15:05 |
| **1** | 成都 | 泸沽湖 | 1145.0 | 376.0 | 成都-泸沽湖3天2晚 \| 入住7天酒店丽江古城中心店 + 成都航空往返机票 | 7天酒店丽江古城中心店 经济型 4.0分/5分 | 经济房-不含早-限时特... 其他 不含早 1间2晚 | 成都航空 EU2237 | 直飞 | 19:45-21:20 | 成都航空 EU2738 | 直飞 | 23:30-01:05 |
| **2** | 广州 | 沈阳 | 2702.0 | 618.0 | 广州-沈阳3天2晚 \| 入住沈阳中煤宾馆 + 南方航空/深圳航空往返机票 | 沈阳中煤宾馆 舒适型 4.5分/5分 | 大床间(内宾) 大床 双早 1间2晚 | 南方航空 CZ6384 | 直飞 | 08:05-11:45 | 深圳航空 ZH9652 | 经停 | 08:20-13:05 |

## 3. 初步探索数据

In [68]:
```
#查看数据形状
df.shape
```

Out[68]: (5100, 13)

In [69]:
```
#了解数据的结构
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5100 entries, 0 to 5099
Data columns (total 13 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   出发地     5098 non-null   object 
 1    目的地    5099 non-null   object 
 2   价格      5072 non-null   float64
 3   节省      5083 non-null   float64
 4   路线名     5100 non-null   object 
 5   酒店      5100 non-null   object 
 6   房间      5100 non-null   object 
 7   去程航司    5100 non-null   object 
 8   去程方式    5100 non-null   object 
 9   去程时间    5100 non-null   object 
 10  回程航司    5100 non-null   object 
 11  回程方式    5100 non-null   object 
 12  回程时间    5100 non-null   object 
dtypes: float64(2), object(11)
memory usage: 557.8+ KB
```

In [70]:
```
#快速查看数据的描述性统计信息
df.describe() #显示数值型数据的描述统计
```

Out[70]:

| | 价格 | 节省 |
|---|---|---|
| **count** | 5072.000000 | 5083.000000 |
| **mean** | 1765.714905 | 474.139878 |
| **std** | 2580.129644 | 168.893780 |
| **min** | 578.000000 | 306.000000 |
| **25%** | 1253.000000 | 358.000000 |
| **50%** | 1632.000000 | 436.000000 |
| **75%** | 2028.250000 | 530.000000 |
| **max** | 179500.000000 | 3500.000000 |

## 4.简单数据处理

In [71]:
```
df.columns
```

Out[71]:
```
Index(['出发地 ', ' 目的地', '价格 ', '节省', '路线名', '酒店', '房间', '去程航司', '去程方式', '去程时间',
       '回程航司', '回程方式 ', '回程时间'],
      dtype='object')
```

In [72]:
```
df.head(2)
```

Out[72]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 哈尔滨 | 北海 | 2208.0 | 650.0 | 哈尔滨-北海3天2晚 \| 入住北海祥丰嘉年华大酒店 + 春秋航空往返机票 | 北海祥丰嘉年华大酒店 舒适型 4.7分/5分 | 标准双人间(双床) 双床 不含早 1间2晚 | 春秋航空 9C8741 | 直飞 | 17:10-21:50 | 春秋航空 9C8742 | 直飞 | 10:20-15:05 |
| **1** | 成都 | 泸沽湖 | 1145.0 | 376.0 | 成都-泸沽湖3天2晚 \| 入住7天酒店丽江古城中心店 + 成都航空往返机票 | 7天酒店丽江古城中心店 经济型 4.0分/5分 | 经济房-不含早-限时特... 其他 不含早 1间2晚 | 成都航空 EU2237 | 直飞 | 19:45-21:20 | 成都航空 EU2738 | 直飞 | 23:30-01:05 |

In [73]:
```
col = df.columns.values
```

In [74]:
```
col
```

Out[74]:
```
array(['出发地 ', ' 目的地', '价格 ', '节省', '路线名', '酒店', '房间', '去程航司', '去程方式',
       '去程时间', '回程航司', '回程方式 ', '回程时间'], dtype=object)
```

In [75]:
```
col[0].strip() #strip函数一次只能处理一个数据，用于删除字符串头尾指定的字符（默认为空格或换行符）
```

Out[75]:
```
'出发地'
```

In [76]:
```
[x.strip() for x in col]
#strip去除前后空格
```

Out[76]:
```
['出发地',
 '目的地',
 '价格',
 '节省',
 '路线名',
 '酒店',
 '房间',
 '去程航司',
 '去程方式',
 '去程时间',
 '回程航司',
 '回程方式',
 '回程时间']
```

In [77]:
```
df.columns = [x.strip() for x in col]
```

In [78]:
```
df.columns
```

Out[78]:
```
Index(['出发地', '目的地', '价格', '节省', '路线名', '酒店', '房间', '去程航司', '去程方式', '去程时间',
       '回程航司', '回程方式', '回程时间'],
      dtype='object')
```

## 5.重复值的处理

### 检查重复值duplicated()

Duplicated函数功能：查找并显示数据表中的重复值
这里需要注意的是：

- 当两条记录中所有的数据都相等时duplicated函数才会判断为重复值
- duplicated支持从前向后(first)，和从后向前(last)两种重复值查找模式
- 默认是从前向后进行重复值的查找和判断，也就是后面的条目在重复值判断中显示为True

In [79]:
```
#第一行数据
#第二行数据和第一行一样
#从前向后就把第二行数据判断为重复值
#从后向前就把第一行数据判断为重复值
```

In [80]:
```
df.duplicated() #返回布尔型数据，告诉重复值的位置
```

Out[80]:
```
0       False
1       False
2       False
3       False
4       False
        ...
5095     True
5096    False
5097    False
5098    False
5099    False
Length: 5100, dtype: bool
```

In [81]:
```
df.duplicated().sum()#说明有100个重复值
```

Out[81]:
```
100
```

In [82]:
```
#查看重复的记录
df[df.duplicated()]
```

Out[82]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **454** | 广州 | 黄山 | 1871.0 | 492.0 | 广州-黄山3天2晚 \| 入住黄山汤口醉享主题酒店 + 南方航空往返机票 | 黄山汤口醉享主题酒店 舒适型 4.8分/5分 | 睫毛弯弯(大床)大床 不含早 1间2晚 | 南方航空 CZ3627 | 直飞 | 19:20-21:15 | 南方航空 CZ3628 | 直飞 | 22:05-23:50 |
| **649** | 济南 | 长沙 | 1134.0 | 360.0 | 济南-长沙3天2晚 \| 入住长沙喜迎宾华天大酒店 + 山东航空往返机票 | 长沙喜迎宾华天大酒店 高档型 3.7分/5分 | 特惠双间(特惠抢购)(... 双床 不含早 1间2晚 | 山东航空 SC1185 | 直飞 | 18:40-20:50 | 山东航空 SC1186 | 直飞 | 10:20-12:15 |
| **685** | 青岛 | 重庆 | 1474.0 | 420.0 | 青岛-重庆3天2晚 \| 入住怡家丽景酒店重庆垫江店 + 山东航空/华夏航空往返机票 | 怡家丽景酒店重庆垫江店 舒适型 4.3分/5分 | 法式房(内宾)(无窗)... 大床 不含早 1间2晚 | 山东航空 SC4709 | 经停 | 19:30-00:05 | 华夏航空 G54710 | 经停 | 18:00-22:25 |
| **852** | 北京 | 哈尔滨 | 1450.0 | 368.0 | 北京-哈尔滨3天2晚 \| 入住哈尔滨水逸城市酒店 + 南方航空/大新华航空往返机票 | 哈尔滨水逸城市酒店 舒适型 4.6分/5分 | 标准间-【预付特惠】独... 双床 双早 1间2晚 | 南方航空 CZ6202 | 直飞 | 22:20-00:20 | 大新华航空 CN7150 | 直飞 | 22:50-00:55 |
| **922** | 北京 | 长沙 | 1289.0 | 334.0 | 北京-长沙3天2晚 \| 入住浏阳市华尔宫大酒店 + 海南航空/南方航空往返机票 | 浏阳市华尔宫大酒店 3.8分/5分 | 豪华双人间(双床)双床 不含早 1间2晚 | 海南航空 HU7135 | 直飞 | 17:45-20:30 | 南方航空 CZ3855 | 直飞 | 22:55-01:10 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **5045** | 杭州 | 丽江 | 2872.0 | 718.0 | 杭州-丽江3天2晚 \| 入住丽江松竹居客栈玉观音店 + 长龙航空/首都航空往返机票 | 丽江松竹居客栈玉观音店 高档型 4.6分/5分 | 特惠房(大床) 大床 不含早 1间2晚 | 长龙航空 GJ8869 | 经停 | 08:50-13:35 | 首都航空 JD5192 | 直飞 | 13:00-16:15 |
| **5066** | 哈尔滨 | 西安 | 1843.0 | 450.0 | 哈尔滨-西安3天2晚 \| 入住西安铁通商务酒店贵宾楼 + 天津航空/东方航空往返... | 西安铁通商务酒店贵宾楼 舒适型 4.2分/5分 | 标准间(持房卡尊享清凉... 双床 不含早 1间2晚 | 天津航空 GS7584 | 经停 | 12:30-17:40 | 东方航空 MU2211 | 经停 | 08:45-13:25 |
| **5068** | 南京 | 成都 | 1922.0 | 552.0 | 南京-成都3天2晚 \| 入住成都伊丽特酒店 + 东方航空/西藏航空往返机票 | 成都伊丽特酒店 高档型 4.5分/5分 | 行政标准间-含早立即确... 双床 双早 1间2晚 | 东方航空 MU2880 | 直飞 | 21:55-00:35 | 西藏航空 TV9839 | 直飞 | 06:30-08:55 |
| **5081** | 上海 | 青岛 | 769.0 | 354.0 | 上海-青岛3天2晚 \| 入住青岛金中太大酒店 + 春秋航空/吉祥航空往返机票 | 青岛金中太大酒店 舒适型 4.8分/5分 | 特惠大床房[无早]大床 不含早 1间2晚 | 春秋航空 9C8853 | 直飞 | 19:50-21:35 | 吉祥航空 HO1242 | 直飞 | 23:05-00:35 |
| **5095** | 宁波 | 九寨沟 | 2085.0 | 562.0 | 宁波-九寨沟3天2晚 \| 入住黑水县达古冰山国际大酒店 + 成都航空/四川航空往... | 黑水县达古冰山国际大酒店 豪华型 3.9分/5分 | B区豪华标间(双床) 双床 不含早 1间2晚 | 成都航空 EU2730 | 经停 | 19:45-00:45 | 四川航空 3U8927 | 经停 | 07:55-12:15 |

100 rows × 13 columns

**删除重复值drop_duplicates()**

drop_duplicates函数功能是：删除数据表中的重复值，判断标准和逻辑与duplicated函数一样

In [83]:
```
df.drop_duplicates(inplace=True)
#inplace=True表示直接在源数据上进行操作
```

In [84]:
```
df.head()
```

Out[84]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 哈尔滨 | 北海 | 2208.0 | 650.0 | 哈尔滨-北海3天2晚 \| 入住北海祥丰嘉年华大酒店 + 春秋航空往返机票 | 北海祥丰嘉年华大酒店 舒适型 4.7分/5分 | 标准双人间(双床) 双床 不含早 1间2晚 | 春秋航空 9C8741 | 直飞 | 17:10-21:50 | 春秋航空 9C8742 | 直飞 | 10:20-15:05 |
| **1** | 成都 | 泸沽湖 | 1145.0 | 376.0 | 成都-泸沽湖3天2晚 \| 入住7天酒店丽江古城中心店 + 成都航空往返机票 | 7天酒店丽江古城中心店 经济型 4.0分/5分 | 经济房-不含早-限时特... 其他 不含早 1间2晚 | 成都航空 EU2237 | 直飞 | 19:45-21:20 | 成都航空 EU2738 | 直飞 | 23:30-01:05 |
| **2** | 广州 | 沈阳 | 2702.0 | 618.0 | 广州-沈阳3天2晚 \| 入住沈阳中煤宾馆 + 南方航空/深圳航空往返机票 | 沈阳中煤宾馆 舒适型 4.5分/5分 | 大床间(内宾) 大床 双早 1间2晚 | 南方航空 CZ6384 | 直飞 | 08:05-11:45 | 深圳航空 ZH9652 | 经停 | 08:20-13:05 |
| **3** | 上海 | 九寨沟 | 1954.0 | 484.0 | 上海-九寨沟3天2晚 \| 入住红原芸谊大酒店 + 成都航空往返机票 | 红原芸谊大酒店 舒适型 4.6分/5分 | 豪华双床房[双早] 双床 双早 1间2晚 | 成都航空 EU6678 | 直飞 | 21:55-01:15 | 成都航空 EU6677 | 直飞 | 17:45-20:35 |
| **4** | 广州 | 天津 | 1608.0 | 422.0 | 广州-天津3天2晚 \| 入住天津逸海明珠大酒店 + 奥凯航空/海南航空往返机票 | 天津逸海明珠大酒店 高档型 4.1分/5分 | 豪华双床房(预付) 双床 不含早 1间2晚 | 奥凯航空 BK2787 | 直飞 | 06:55-10:00 | 海南航空 HU7201 | 直飞 | 20:15-23:25 |

In [85]:
```
df.shape
```

Out[85]: (5000, 13)

In [86]:
```
df
```

Out[86]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 哈尔滨 | 北海 | 2208.0 | 650.0 | 哈尔滨-北海3天2晚 \| 入住北海祥丰嘉年华大酒店 + 春秋航空往返机票 | 北海祥丰嘉年华大酒店 舒适型 4.7分/5分 | 标准双人间(双床) 双床 不含早 1间2晚 | 春秋航空 9C8741 | 直飞 | 17:10-21:50 | 春秋航空 9C8742 | 直飞 | 10:20-15:05 |
| **1** | 成都 | 泸沽湖 | 1145.0 | 376.0 | 成都-泸沽湖3天2晚 \| 入住7天酒店丽江古城中心店 + 成都航空往返机票 | 7天酒店丽江古城中心店 经济型 4.0分/5分 | 经济房-不含早-限时特... 其他 不含早 1间2晚 | 成都航空 EU2237 | 直飞 | 19:45-21:20 | 成都航空 EU2738 | 直飞 | 23:30-01:05 |
| **2** | 广州 | 沈阳 | 2702.0 | 618.0 | 广州-沈阳3天2晚 \| 入住沈阳中煤宾馆 + 南方航空/深圳航空往返机票 | 沈阳中煤宾馆 舒适型 4.5分/5分 | 大床间(内宾) 大床 双早 1间2晚 | 南方航空 CZ6384 | 直飞 | 08:05-11:45 | 深圳航空 ZH9652 | 经停 | 08:20-13:05 |
| **3** | 上海 | 九寨沟 | 1954.0 | 484.0 | 上海-九寨沟3天2晚 \| 入住红原芸谊大酒店 + 成都航空往返机票 | 红原芸谊大酒店 舒适型 4.6分/5分 | 豪华双床房[双早] 双床 双早 1间2晚 | 成都航空 EU6678 | 直飞 | 21:55-01:15 | 成都航空 EU6677 | 直飞 | 17:45-20:35 |
| **4** | 广州 | 天津 | 1608.0 | 422.0 | 广州-天津3天2晚 \| 入住天津逸海明珠大酒店 + 奥凯航空/海南航空往返机票 | 天津逸海明珠大酒店 高档型 4.1分/5分 | 豪华双床房(预付) 双床 不含早 1间2晚 | 奥凯航空 BK2787 | 直飞 | 06:55-10:00 | 海南航空 HU7201 | 直飞 | 20:15-23:25 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **5094** | 大连 | 武汉 | 1473.0 | 368.0 | 大连-武汉3天2晚 \| 入住武汉保利大酒店 + 南方航空/东方航空往返机票 | 武汉保利大酒店 豪华型 4.3分/5分 | 附楼标准间（双床） 双床 双早 1间2晚 | 南方航空 CZ8306 | 直飞 | 18:35-21:10 | 东方航空 MU2517 | 经停 | 07:40-12:05 |
| **5096** | 成都 | 泸沽湖 | 1158.0 | 376.0 | 成都-泸沽湖3天2晚 \| 入住丽江望月阁客栈 + 成都航空往返机票 | 丽江望月阁客栈 经济型 4.6分/5分 | 标准双人间-不含早（预... 双床 不含早 1间2晚 | 成都航空 EU2237 | 直飞 | 19:45-21:20 | 成都航空 EU2738 | 直飞 | 23:30-01:05 |
| **5097** | 天津 | 丽江 | 1616.0 | 426.0 | 天津-丽江3天2晚 \| 入住丽江凡间度假连锁客栈青旅店 + 天津航空/首都航空往... | 丽江凡间度假连锁客栈青旅店 经济型 4.5分/5分 | 大床房-预付 大床 不含早 1间2晚 | 天津航空 GS7861 | 直飞 | 16:25-19:45 | 首都航空 JD5739 | 直飞 | 07:50-10:50 |
| **5098** | 大连 | 重庆 | 1703.0 | 446.0 | 大连-重庆3天2晚 \| 入住重庆酉阳锦宏大酒店 + 华夏航空/山东航空往返机票 | 重庆酉阳锦宏大酒店 舒适型 4.0分/5分 | 特惠房(大床) 大床 不含早 1间2晚 | 华夏航空 G52762 | 经停 | 18:25-23:30 | 山东航空 SC4837 | 经停 | 07:00-11:30 |
| **5099** | 天津 | 哈尔滨 | 1192.0 | 356.0 | 天津-哈尔滨3天2晚 \| 入住哈尔滨钰轩酒店中央大街店 + 奥凯航空/福州往返机票 | 哈尔滨钰轩酒店中央大街店 舒适型 4.0分/5分 | 标准大床房 大床 大床 双早 1间2晚 | 奥凯航空 BK2918 | 直飞 | 21:10-23:25 | 福州 FU6556 | 直飞 | 11:35-13:45 |

5000 rows × 13 columns

In [87]:
```
df.shape[0]
```

Out[87]: 5000

In [88]:
```
range(df.shape[0])
```

Out[88]:
```
range(0, 5000)
```

In [89]:
```
df.index = range(df.shape[0])
```

In [90]:
```
df.index
```

Out[90]:
```
RangeIndex(start=0, stop=5000, step=1)
```

## 6.异常值的处理

In [91]:
```
df.describe().T
```

Out[91]:

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **价格** | 4972.0 | 1767.782381 | 2604.329780 | 578.0 | 1253.0 | 1633.0 | 2031.0 | 179500.0 |
| **节省** | 4983.0 | 474.490869 | 169.148391 | 306.0 | 358.0 | 436.0 | 532.0 | 3500.0 |

In [92]:
```
#计算标准分数，找出'价格'异常值
sta=(df['价格']-df['价格'].mean())/df['价格'].std()
```

In [93]:
```
sta[:10]
```

Out[93]:
```
0     0.169033
1    -0.239133
2     0.358717
3     0.071503
4    -0.061353
5    -0.072104
6    -0.054825
7    -0.036394
8    -0.229534
9    -0.175393
Name: 价格, dtype: float64
```

In [94]:
```
sta.abs()[:10]
```

Out[94]:
```
0     0.169033
1     0.239133
2     0.358717
3     0.071503
4     0.061353
5     0.072104
6     0.054825
7     0.036394
8     0.229534
9     0.175393
Name: 价格, dtype: float64
```

In [95]:
```
sta.abs()>3
```

Out[95]:
```
0       False
1       False
2       False
3       False
4       False
        ...
4995    False
4996    False
4997    False
4998    False
4999    False
Name: 价格, Length: 5000, dtype: bool
```

In [96]:
```
df[sta.abs()>3]
```

Out[96]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2763** | 杭州 | 九寨沟 | 179500.0 | 538.0 | 杭州-九寨沟3天2晚 \| 入住九寨沟九乡宾馆 + 成都航空/长龙航空往返机票 | 九寨沟九乡宾馆 舒适型 4.3分/5分 | 特惠房(双床)双床 不含早 1间2晚 | 成都航空EU2206 | 经停 | 20:30-01:00 | 长龙航空GJ8680 | 经停 | 20:25-00:50 |

In [97]:
```
df.head()
```

Out[97]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 哈尔滨 | 北海 | 2208.0 | 650.0 | 哈尔滨-北海3天2晚 \| 入住北海祥丰嘉年华大酒店 + 春秋航空往返机票 | 北海祥丰嘉年华大酒店 舒适型 4.7分/5分 | 标准双人间(双床)双床 不含早 1间2晚 | 春秋航空9C8741 | 直飞 | 17:10-21:50 | 春秋航空9C8742 | 直飞 | 10:20-15:05 |
| **1** | 成都 | 泸沽湖 | 1145.0 | 376.0 | 成都-泸沽湖3天2晚 \| 入住7天酒店丽江古城中心店 + 成都航空往返机票 | 7天酒店丽江古城中心店 经济型 4.0分/5分 | 经济房-不含早-限时特... 其他 不含早 1间2晚 | 成都航空EU2237 | 直飞 | 19:45-21:20 | 成都航空EU2738 | 直飞 | 23:30-01:05 |
| **2** | 广州 | 沈阳 | 2702.0 | 618.0 | 广州-沈阳3天2晚 \| 入住沈阳中煤宾馆 + 南方航空/深圳航空往返机票 | 沈阳中煤宾馆 舒适型 4.5分/5分 | 大床间(内宾) 大床 双早 1间2晚 | 南方航空CZ6384 | 直飞 | 08:05-11:45 | 深圳航空ZH9652 | 经停 | 08:20-13:05 |
| **3** | 上海 | 九寨沟 | 1954.0 | 484.0 | 上海-九寨沟3天2晚 \| 入住红原芸谊大酒店 + 成都航空往返机票 | 红原芸谊大酒店 舒适型 4.6分/5分 | 豪华双床房[双早]双床 双早 1间2晚 | 成都航空EU6678 | 直飞 | 21:55-01:15 | 成都航空EU6677 | 直飞 | 17:45-20:35 |
| **4** | 广州 | 天津 | 1608.0 | 422.0 | 广州-天津3天2晚 \| 入住天津逸海明珠大酒店 + 奥凯航空/海南航空往返机票 | 天津逸海明珠大酒店 高档型 4.1分/5分 | 豪华双床房(预付)双床 不含早 1间2晚 | 奥凯航空BK2787 | 直飞 | 06:55-10:00 | 海南航空HU7201 | 直飞 | 20:15-23:25 |

In [98]:
```
sum(df.价格>df.节省)
```

Out[98]: 4952

In [99]:
```
#找出'节省'异常值
df[df.节省>df.价格]
```

Out[99]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2904** | 武汉 | 西安 | 949.0 | 3500.0 | 武汉-西安3天2晚 \| 入住西安西稍门大酒店 + 东方航空往返机票 | 西安西稍门大酒店 舒适型 3.3分/5分 | 标准间B(丝路之旅)(... 双床 不含早 1间2晚 | 东方航空MU2194 | 直飞 | 21:50-23:30 | 东方航空MU2462 | 直飞 | 19:35-21:20 |
| **3108** | 济南 | 大连 | 911.0 | 3180.0 | 济南-大连3天2晚 \| 入住普兰店科洋大酒店 + 山东航空/厦门航空往返机票 | 普兰店科洋大酒店 舒适型 4.4分/5分 | 大床房(限量促销) 大床 不含早 1间2晚 | 山东航空SC4916 | 直飞 | 19:45-20:50 | 厦门航空MF8042 | 直飞 | 13:10-14:20 |
| **3660** | 沈阳 | 青岛 | 924.0 | 3200.0 | 沈阳-青岛3天2晚 \| 入住星程酒店青岛台东步行街店 + 青岛航/南方航空往返机票 | 星程酒店青岛台东步行街店 舒适型 4.2分/5分 | 大床房(内宾)(提前1... 大床 不含早 1间2晚 | 青岛航QW9780 | 直飞 | 22:35-00:10 | 南方航空CZ6568 | 直飞 | 20:55-22:35 |

- 对于建模来说，通常会删掉异常值
- 但是对于业务来说，异常值中可能包含有更多的价值

In [100]:
```
pd.concat([df[df.节省>df.价格],df[sta.abs()>3]])
```

Out[100]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2904** | 武汉 | 西安 | 949.0 | 3500.0 | 武汉-西安3天2晚 \| 入住西安西稍门大酒店 + 东方航空往返机票 | 西安西稍门大酒店 舒适型 3.3分/5分 | 标准间B(丝路之旅)(... 双床 不含早 1间2晚 | 东方航空 MU2194 | 直飞 | 21:50-23:30 | 东方航空 MU2462 | 直飞 | 19:35-21:20 |
| **3108** | 济南 | 大连 | 911.0 | 3180.0 | 济南-大连3天2晚 \| 入住普兰店科洋大酒店 + 山东航空/厦门航空往返机票 | 普兰店科洋大酒店 舒适型 4.4分/5分 | 大床房(限量促销) 大床 不含早 1间2晚 | 山东航空 SC4916 | 直飞 | 19:45-20:50 | 厦门航空 MF8042 | 直飞 | 13:10-14:20 |
| **3660** | 沈阳 | 青岛 | 924.0 | 3200.0 | 沈阳-青岛3天2晚 \| 入住星程酒店青岛台东步行街店 + 青岛航/南方航空往返机票 | 星程酒店青岛台东步行街店 舒适型 4.2分/5分 | 大床房(内宾)(提前1... 大床 不含早 1间2晚 | 青岛航 QW9780 | 直飞 | 22:35-00:10 | 南方航空 CZ6568 | 直飞 | 20:55-22:35 |
| **2763** | 杭州 | 九寨沟 | 179500.0 | 538.0 | 杭州-九寨沟3天2晚 \| 入住九寨沟九乡宾馆 + 成都航空/长龙航空往返机票 | 九寨沟九乡宾馆 舒适型 4.3分/5分 | 特惠房(双床) 双床 不含早 1间2晚 | 成都航空 EU2206 | 经停 | 20:30-01:00 | 长龙航空 GJ8680 | 经停 | 20:25-00:50 |

In [101]:
```
pd.concat([df[df.节省>df.价格],df[sta.abs()>3]]).index
```

Out[101]:
```
Int64Index([2904, 3108, 3660, 2763], dtype='int64')
```

In [102]:
```
delindex = pd.concat([df[df.节省>df.价格],df[sta.abs()>3]]).index
```

In [103]:
```
df.drop(delindex,inplace=True)   #删除异常数据
```

In [104]:
```
df.shape
```

Out[104]:
```
(4996, 13)
```

## 7.缺失值的处理

- df.isnull() #查看缺失值
- df.notnull() #查看不是缺失值的数据
- df.dropna() #删除缺失值
- df.fillna() #填补缺失值

In [105]:
```
df.isnull().sum()
```

Out[105]:
```
出发地       2
目的地       1
价格       28
节省       17
路线名       0
酒店        0
房间        0
去程航司      0
去程方式      0
去程时间      0
回程航司      0
回程方式      0
回程时间      0
dtype: int64
```

In [106]:
```
df[df.出发地.isnull()]
```

Out[106]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1850** | NaN | 烟台 | 647.0 | 348.0 | 大连-烟台3天2晚 \| 入住烟台海阳黄金海岸大酒店 + 幸福航空/天津航空往返机票 | 烟台海阳黄金海岸大酒店 3.7分/5分 | 海景标准间(内宾)[双... 双床 双早 1间2晚 | 幸福航空 JR1582 | 直飞 | 10:05-11:05 | 天津航空 GS6402 | 直飞 | 16:30-17:25 |
| **1915** | NaN | 西安 | 1030.0 | 326.0 | 济南-西安3天2晚 \| 入住西安丝路秦国际青年旅舍钟楼回民街店 + 华夏航空往返... | 西安丝路秦国际青年旅舍钟楼回民街店 经济型 4.4分/5分 | 标准间（独卫）- 吃货天... 双床 不含早 1间2晚 | 华夏航空 G54963 | 直飞 | 07:10-08:55 | 华夏航空 G58858 | 直飞 | 23:10-00:55 |

In [107]:
```
[str(x)[:2] for x in df.loc[df.出发地.isnull(),'路线名']]   #利用路线名中的信息补全出发地
```

Out[107]:
```
['大连', '济南']
```

In [108]:
```
df.loc[df.出发地.isnull(),'出发地'] = [str(x)[:2] for x in df.loc[df.出发地.isnull(),'路线名']]
```

In [109]:
```
df[df.出发地.isnull()]
```

Out[109]:

| 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

In [110]:
```
df.出发地.isnull().sum()
```

Out[110]:
```
0
```

In [111]:
```
df[df.目的地.isnull()]
```

Out[111]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1860** | 深圳 | NaN | 2149.0 | 494.0 | 深圳-大连3天2晚 \| 入住大连黄金山大酒店 + 南方航空/东海往返机票 | 大连黄金山大酒店 舒适型 3.4分/5分 | 标准间 大/双床 不含早 1间2晚 | 南方航空 CZ6833 | 直飞 | 09:10-12:40 | 东海 DZ6242 | 经停 | 12:40-18:00 |

In [112]:
```
str(df.loc[df.目的地.isnull(),'路线名'].values)[5:7]
```

Out[112]:
```
'大连'
```

In [113]:
```
df.loc[df.目的地.isnull(),'目的地'] = str(df.loc[df.目的地.isnull(),'路线名'].values)[5:7]
```

In [114]:
```
round(df['价格'].mean(),0)
```

Out[114]:
```
1733.0
```

In [115]:
```
#处理价格缺失值
df['价格'].fillna(round(df['价格'].mean(),0),inplace=True)
```

In [116]:
```
#处理节省缺失值
df['节省'].fillna(round(df['节省'].mean(),0),inplace=True)
```

In [117]:
```
df.isnull().sum()
```

Out[117]:
```
出发地      0
目的地      0
价格       0
节省       0
路线名      0
酒店       0
房间       0
去程航司     0
去程方式     0
去程时间     0
回程航司     0
回程方式     0
回程时间     0
dtype: int64
```

## 8. 处理文本型数据

In [118]:
```
# 如果我们想要在一系列文本提取信息,可以使用正则表达式
# 正则表达式通常被用来检索某个规则的文本
#正则表达式包含在引号里面，括号内表示要提取的部分
```

In [119]:
```
df.head()
```

Out[119]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 哈尔滨 | 北海 | 2208.0 | 650.0 | 哈尔滨-北海3天2晚 \| 入住北海祥丰嘉年华大酒店 + 春秋航空往返机票 | 北海祥丰嘉年华大酒店 舒适型 4.7分/5分 | 标准双人间(双床)双床 不含早 1间2晚 | 春秋航空 9C8741 | 直飞 | 17:10-21:50 | 春秋航空 9C8742 | 直飞 | 10:20-15:05 |
| 1 | 成都 | 泸沽湖 | 1145.0 | 376.0 | 成都-泸沽湖3天2晚 \| 入住7天酒店丽江古城中心店 + 成都航空往返机票 | 7天酒店丽江古城中心店 经济型 4.0分/5分 | 经济房-不含早-限时特... 其他 不含早 1间2晚 | 成都航空 EU2237 | 直飞 | 19:45-21:20 | 成都航空 EU2738 | 直飞 | 23:30-01:05 |
| 2 | 广州 | 沈阳 | 2702.0 | 618.0 | 广州-沈阳3天2晚 \| 入住沈阳中煤宾馆 + 南方航空/深圳航空往返机票 | 沈阳中煤宾馆 舒适型 4.5分/5分 | 大床间(内宾) 大床 双早 1间2晚 | 南方航空 CZ6384 | 直飞 | 08:05-11:45 | 深圳航空 ZH9652 | 经停 | 08:20-13:05 |
| 3 | 上海 | 九寨沟 | 1954.0 | 484.0 | 上海-九寨沟3天2晚 \| 入住红原芸谊大酒店 + 成都航空往返机票 | 红原芸谊大酒店 舒适型 4.6分/5分 | 豪华双床房[双早] 双床 双早 1间2晚 | 成都航空 EU6678 | 直飞 | 21:55-01:15 | 成都航空 EU6677 | 直飞 | 17:45-20:35 |
| 4 | 广州 | 天津 | 1608.0 | 422.0 | 广州-天津3天2晚 \| 入住天津逸海明珠大酒店 + 奥凯航空/海南航空往返机票 | 天津逸海明珠大酒店 高档型 4.1分/5分 | 豪华双床房(预付)双床 不含早 1间2晚 | 奥凯航空 BK2787 | 直飞 | 06:55-10:00 | 海南航空 HU7201 | 直飞 | 20:15-23:25 |

In [120]:
```
df.酒店[:10]
```

Out[120]:
```
0          北海祥丰嘉年华大酒店 舒适型 4.7分/5分
1        7天酒店丽江古城中心店 经济型 4.0分/5分
2             沈阳中煤宾馆 舒适型 4.5分/5分
3           红原芸谊大酒店 舒适型 4.6分/5分
4         天津逸海明珠大酒店 高档型 4.1分/5分
5         青岛康大豪生大酒店 豪华型 4.6分/5分
6          阿坝松潘松林酒店 舒适型 4.3分/5分
7     重庆运通假日酒店汽博中心店 舒适型 4.3分/5分
8      维也纳酒店南京南站汇景店 舒适型 4.4分/5分
9     厦门泊旅时尚酒店会展中心店 舒适型 4.0分/5分
Name: 酒店, dtype: object
```

In [121]:
```
df.酒店.str.extract('(\d\.\d)分/5分',expand=True)[:10]      #
```

Out[121]:

| | 0 |
|---|---|
| **0** | 4.7 |
| **1** | 4.0 |
| **2** | 4.5 |
| **3** | 4.6 |
| **4** | 4.1 |
| **5** | 4.6 |
| **6** | 4.3 |
| **7** | 4.3 |
| **8** | 4.4 |
| **9** | 4.0 |

In [122]:
```
#提取酒店评分
df['酒店评分'] = df.酒店.str.extract('(\d\.\d)分/5分',expand=False)      #\d表示匹配一个整数

#expand=False (return Index/Series)
#expand=True (return DataFrame)
```

In [123]:
```
df.head(2)
```

Out[123]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 | 酒店评分 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 哈尔滨 | 北海 | 2208.0 | 650.0 | 哈尔滨-北海3天2晚 \| 入住北海祥丰嘉年华大酒店 + 春秋航空往返机票 | 北海祥丰嘉年华大酒店 舒适型 4.7分/5分 | 标准双人间(双床) 双床 不含早 1间2晚 | 春秋航空 9C8741 | 直飞 | 17:10-21:50 | 春秋航空 9C8742 | 直飞 | 10:20-15:05 | 4.7 |
| **1** | 成都 | 泸沽湖 | 1145.0 | 376.0 | 成都-泸沽湖3天2晚 \| 入住7天酒店丽江古城中心店 + 成都航空往返机票 | 7天酒店丽江古城中心店 经济型 4.0分/5分 | 经济房-不含早-限时特... 其他 不含早 1间2晚 | 成都航空 EU2237 | 直飞 | 19:45-21:20 | 成都航空 EU2738 | 直飞 | 23:30-01:05 | 4.0 |

In [124]:
```
df.酒店[:10]
```

Out[124]:
```
0          北海祥丰嘉年华大酒店 舒适型 4.7分/5分
1        7天酒店丽江古城中心店 经济型 4.0分/5分
2            沈阳中煤宾馆 舒适型 4.5分/5分
3          红原芸谊大酒店 舒适型 4.6分/5分
4        天津逸海明珠大酒店 高档型 4.1分/5分
5          青岛康大豪生大酒店 豪华型 4.6分/5分
6          阿坝松潘松林酒店 舒适型 4.3分/5分
7    重庆运通假日酒店汽博中心店 舒适型 4.3分/5分
8     维也纳酒店南京南站汇景店 舒适型 4.4分/5分
9    厦门泊旅时尚酒店会展中心店 舒适型 4.0分/5分
Name: 酒店, dtype: object
```

In [125]:
```
df.酒店.str.extract('(.+)(店|宾馆)',expand=False)[:5]     # .表示匹配任意单个字符，+表示匹配前面的子表达式一次或多次
```

Out[125]:

| | 0 | 1 |
|---|---|---|
| **0** | 北海祥丰嘉年华大酒 | 店 |
| **1** | 7天酒店丽江古城中心 | 店 |
| **2** | 沈阳中煤 | 宾馆 |
| **3** | 红原芸谊大酒 | 店 |
| **4** | 天津逸海明珠大酒 | 店 |

In [126]:
```
#提取酒店等级
df['酒店等级'] = df.酒店.str.extract(' (.+) ',expand=False)
```

In [127]:
```
#提取天数信息
df['天数']=df.路线名.str.extract('(\d+)天\d晚',expand=False)
df.去程航司.str.extract('航空.(.+)',expand=False)[:5]
```

Out[127]:
```
0    9C8741
1    EU2237
2    CZ6384
3    EU6678
4    BK2787
Name: 去程航司, dtype: object
```

In [128]:
```
df.head()
```

Out[128]:

| | 出发地 | 目的地 | 价格 | 节省 | 路线名 | 酒店 | 房间 | 去程航司 | 去程方式 | 去程时间 | 回程航司 | 回程方式 | 回程时间 | 酒店评分 | 酒店等级 | 天数 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 哈尔滨 | 北海 | 2208.0 | 650.0 | 哈尔滨-北海3天2晚 \| 入住北海祥丰嘉年华大酒店 + 春秋航空往返机票 | 北海祥丰嘉年华大酒店 舒适型 4.7分/5分 | 标准双人间(双床) 双床 不含早 1间2晚 | 春秋航空 9C8741 | 直飞 | 17:10-21:50 | 春秋航空 9C8742 | 直飞 | 10:20-15:05 | 4.7 | 舒适型 | 3 |
| **1** | 成都 | 泸沽湖 | 1145.0 | 376.0 | 成都-泸沽湖3天2晚 \| 入住7天酒店丽江古城中心店 + 成都航空往返机票 | 7天酒店丽江古城中心店 经济型 4.0分/5分 | 经济房-不含早-限时特... 其他 不含早 1间2晚 | 成都航空 EU2237 | 直飞 | 19:45-21:20 | 成都航空 EU2738 | 直飞 | 23:30-01:05 | 4.0 | 经济型 | 3 |
| **2** | 广州 | 沈阳 | 2702.0 | 618.0 | 广州-沈阳3天2晚 \| 入住沈阳中煤宾馆 + 南方航空/深圳航空往返机票 | 沈阳中煤宾馆 舒适型 4.5分/5分 | 大床间(内宾) 大床 双早 1间2晚 | 南方航空 CZ6384 | 直飞 | 08:05-11:45 | 深圳航空 ZH9652 | 经停 | 08:20-13:05 | 4.5 | 舒适型 | 3 |
| **3** | 上海 | 九寨沟 | 1954.0 | 484.0 | 上海-九寨沟3天2晚 \| 入住红原芸谊大酒店 + 成都航空往返机票 | 红原芸谊大酒店 舒适型 4.6分/5分 | 豪华双床房[双早] 双床 双早 1间2晚 | 成都航空 EU6678 | 直飞 | 21:55-01:15 | 成都航空 EU6677 | 直飞 | 17:45-20:35 | 4.6 | 舒适型 | 3 |
| **4** | 广州 | 天津 | 1608.0 | 422.0 | 广州-天津3天2晚 \| 入住天津逸海明珠大酒店 + 奥凯航空/海南航空往返机票 | 天津逸海明珠大酒店 高档型 4.1分/5分 | 豪华双床房(预付) 双床 不含早 1间2晚 | 奥凯航空 BK2787 | 直飞 | 06:55-10:00 | 海南航空 HU7201 | 直飞 | 20:15-23:25 | 4.1 | 高档型 | 3 |

In [ ]:
```

```