python读取CSV文件方法：

方法1.可先用以下代码查看当前工作路径，然后将CSV文件放在该路径下。

import os

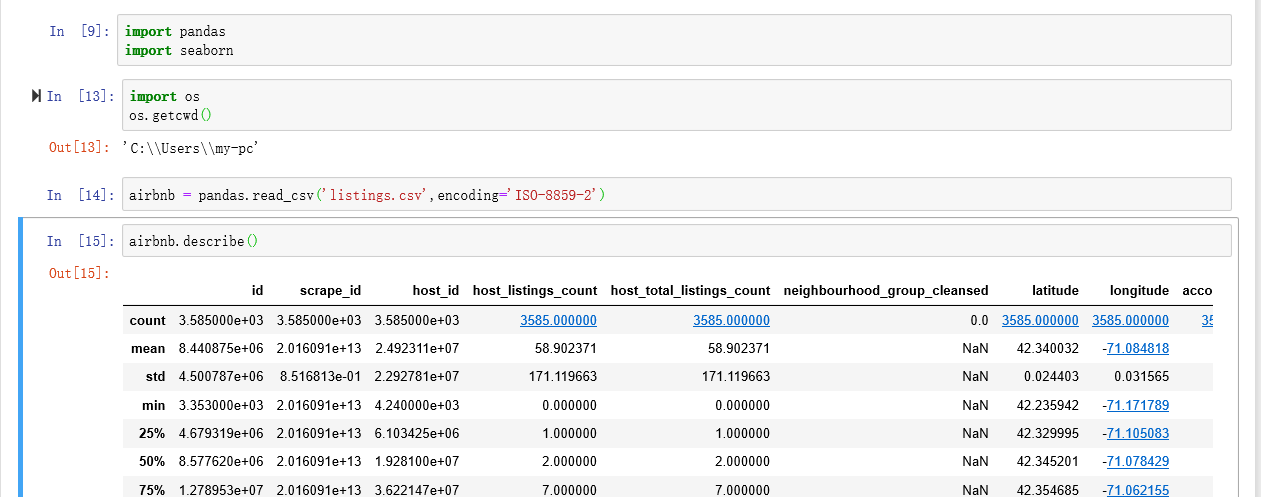
os.getcwd()

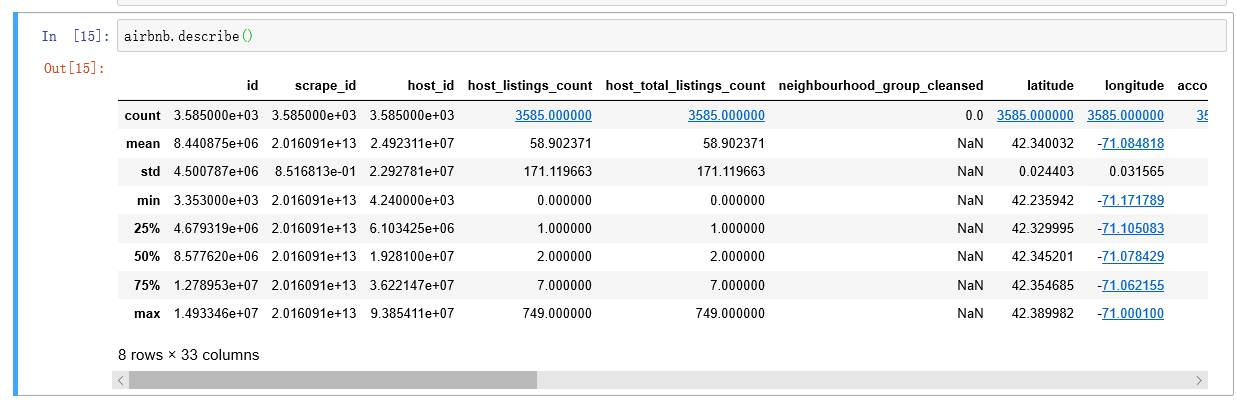
方法2.（绝对路径）

import pandas as pd

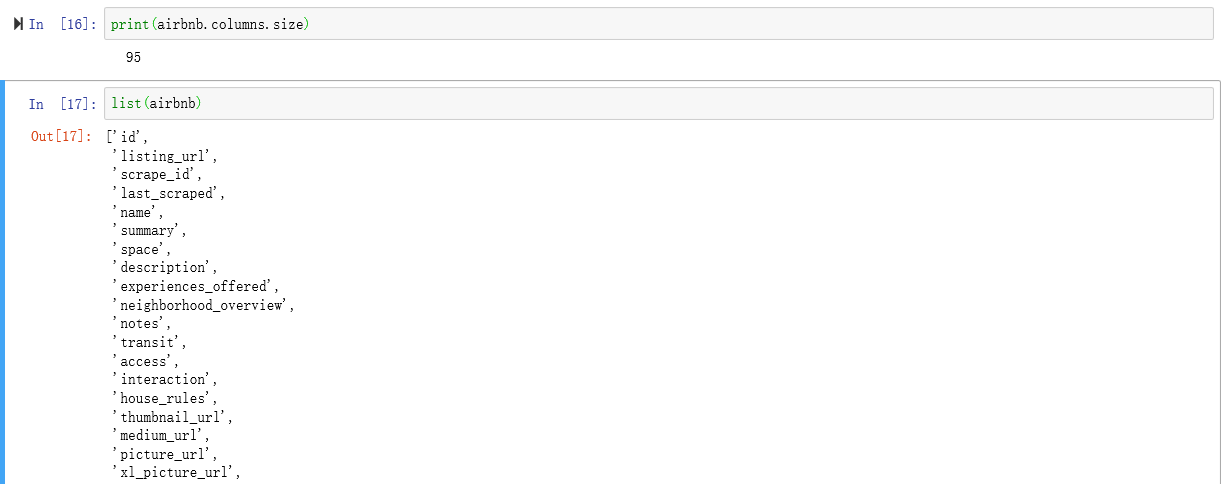
iris\_train=pd.read\_csv('E:\Study\DataSets\iris\_train.csv')

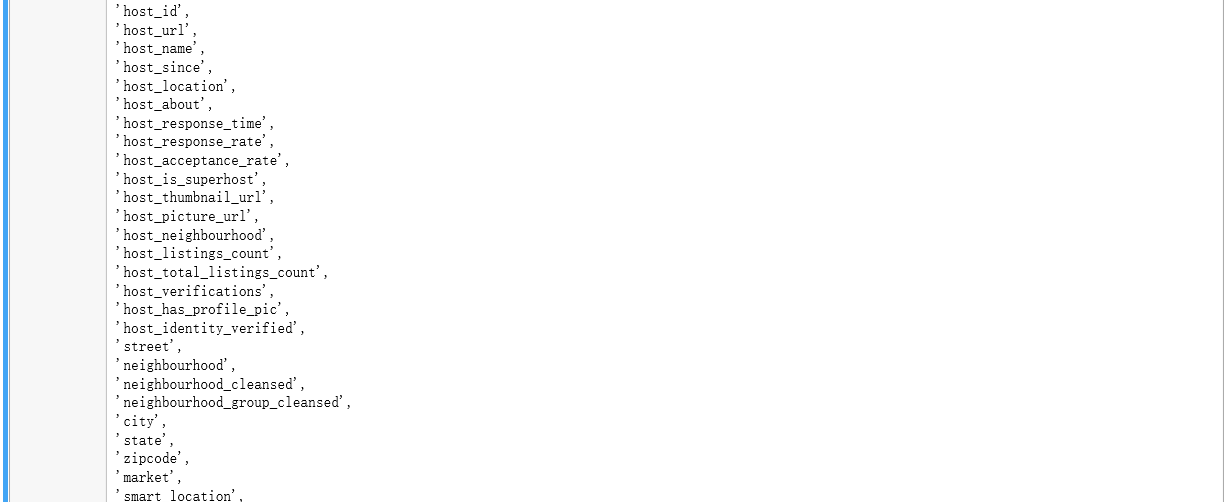
1.安装模块--读取路径--读取csv文件并编码--查看文件内容

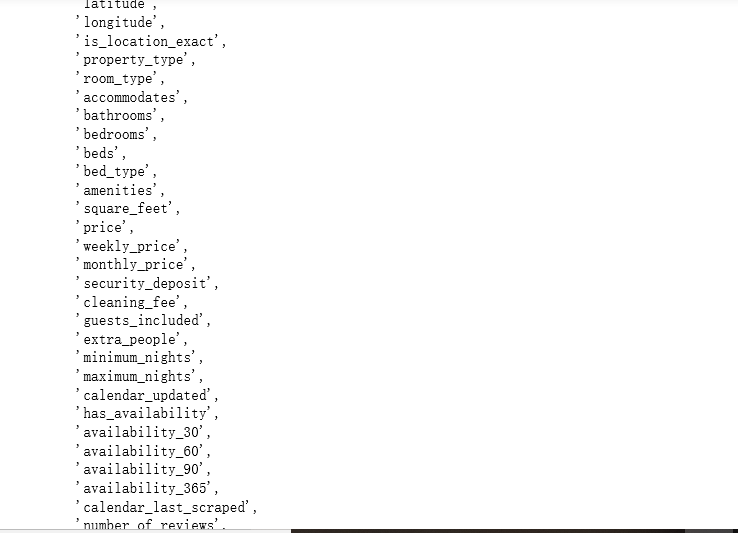




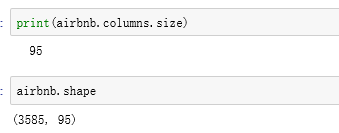
2.①查询有多少行，95行





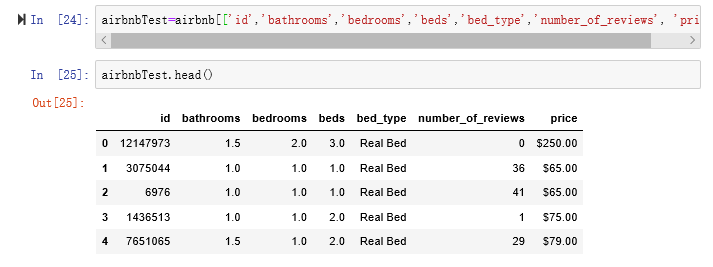


②也可以使用shape方法，查看到数据总共有3585行90列：

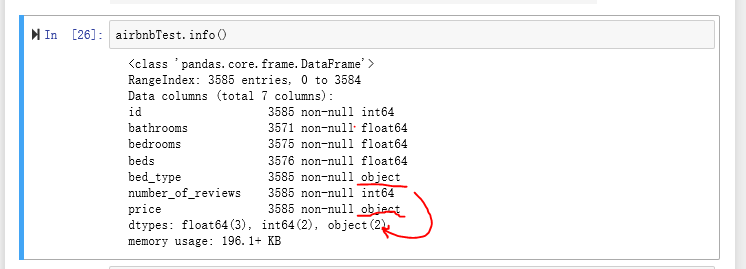


3.重新建立一个数据集airbnbTest，创建的数据库包含只需要用到的列：

airbnbTest=airbnb[['id','bathrooms','bedrooms','beds','bed\_type','number\_of\_reviews', 'price']]



4.使用info（）方法查看数据，发现其他列有空数据，且发现price也是object类型。



到了这一步就知道哪些需要进行清洗：

4.清洗方法：

（1）**填充NaN值**

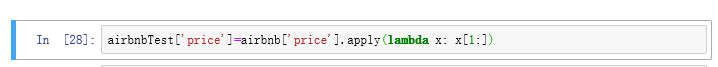
思路：处理列中NaN值，通过分析每列的数据都应该是数值类型，既然是NaN,那使用0来填充：airbnbTest=airbnbTest.fillna('0')



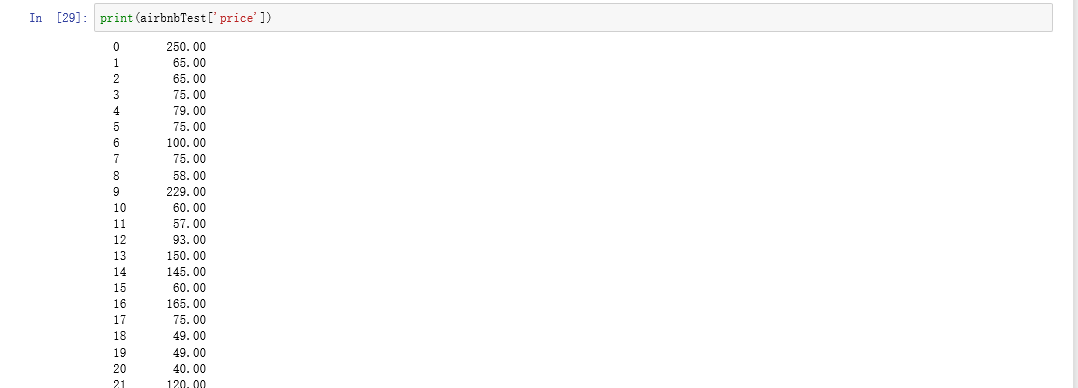
（2） **去掉price列中的$符号：**

思路：分列去掉$符号，取$之后的数值，但object是不能进行分列操作，所以先转换为str类型；再分列操作，取从第1这个位置开始后面所有的数据。





打印看看price这一列清洗的结果：

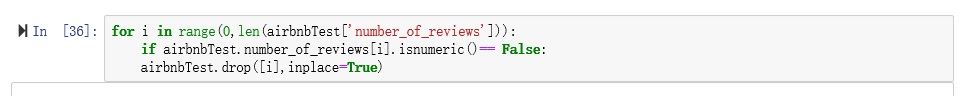


（3）**异常值处理：**

①在评论数（number\_of\_reviews）这一列，只应该有数字：（如果describe查询结果为object就进行处理，这里举例number\_of\_reviews这一列）

思路：判断是否为数字，不是就删除。（直接删除异常值）

注意：这里drop方法里面一定要加上inplace=True，才是表示在数据中将这一行数据完整删除（如果不加上inpalce结果中反而会多出数据）。然后再将这列type转化为int类型。

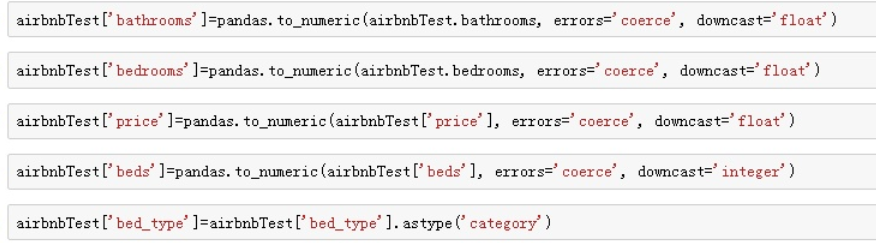




②数据中有异常值则替换为：NaN，且将数据转换格式为：float（异常值替换并转换类型）



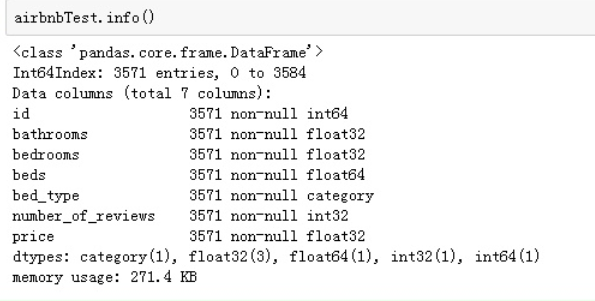
其他列处理如下，



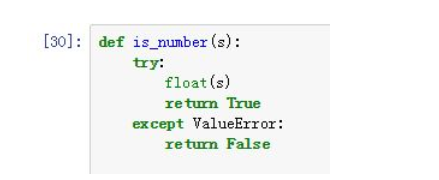
然后用0 填充NaN.



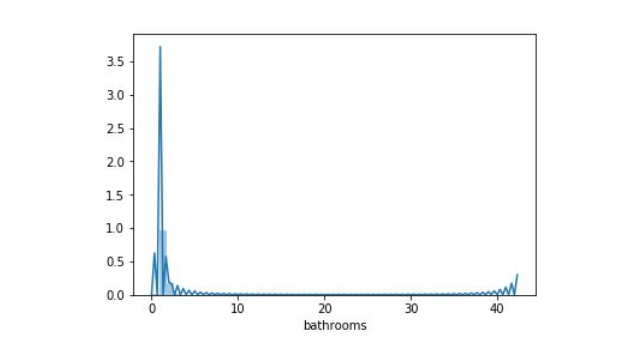
最后查看所有数据格式结果：

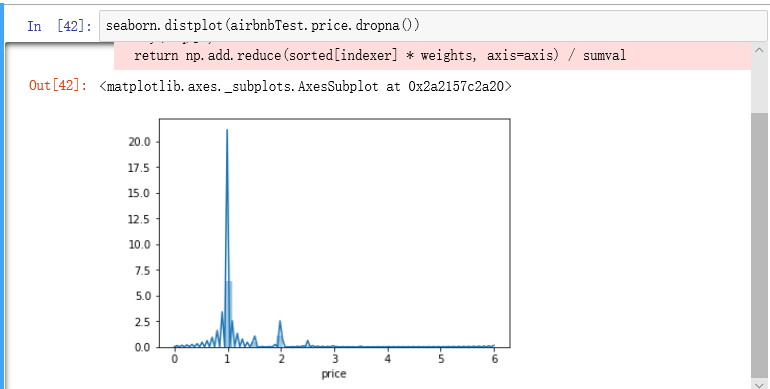


③还可以使用自定义函数的方式筛选异常数据，如果数据有异常筛选出来后则处理掉。



5.**画图查看异常值：**





6.**箱型查看异常值：**

