## 数据探索性分析与数据预处理

数据集一：Melbourne Airbnb Open Data

**1.数据摘要**

1.标称属性：以“room\_type”属性为例，列举出了所有可能的取值，以及对应的频数：

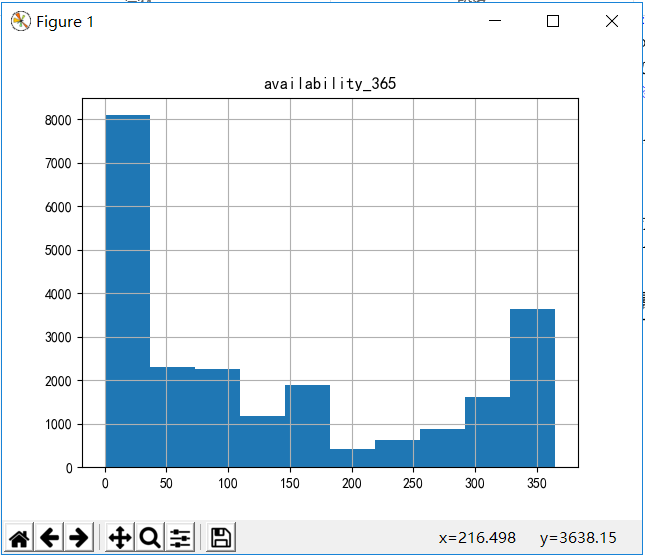


2.数值属性：以“availability\_365”属性为例，分别给出了空值数据的个数、平均值、最小值、四分位数、最大值。

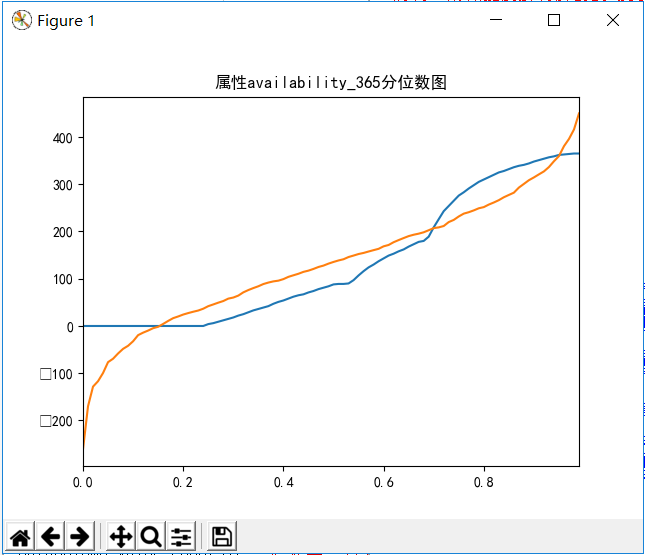


**2.数据可视化**

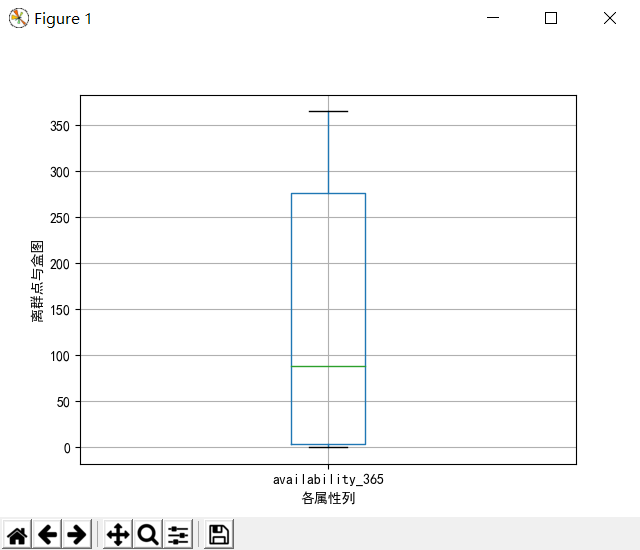
1.直方图：



2.Q-Q图：



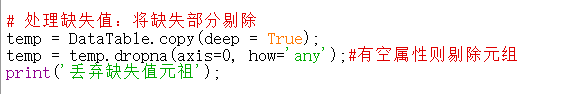
3.盒图：



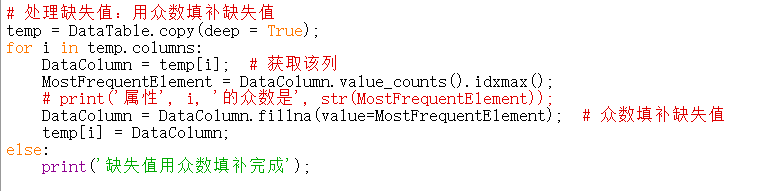
**3.处理缺失数据**

主要使用了三种方法处理缺失值：

1.丢弃有缺失值的元组；



2.用属性列上的众数填充缺失值；



3.用最相关的元组来填充有缺失值的元祖；

主要思路是：对于任意一个有空值的元组a，遍历整个数据表，按照依据：标称属性相似度最高、数值属性欧氏距离最小来寻找最相似的元祖b，把a的缺失属性值用b的相应属性填补。该做法的时间复杂度为O(N^2)。

4.用属性之间的相关性填补缺失值；

主要思路是：对于任意一个有空值的属性列A，计算所有属性列与A的相关系数，选取最相关的属性列B，统计A与B的线性关系，把A中的缺失值用B中的数值计算出来。但这个思路要求属性列B本身不能有太多空值，还要求数据表里的数值属性列的数量要足够多，所以该思路还没有实现。

数据集二：MLB Pitch Data 2015-2018

**1.数据摘要**

1.标称属性：以“stand”属性为例，列举出了所有可能的取值，以及对应的频数：

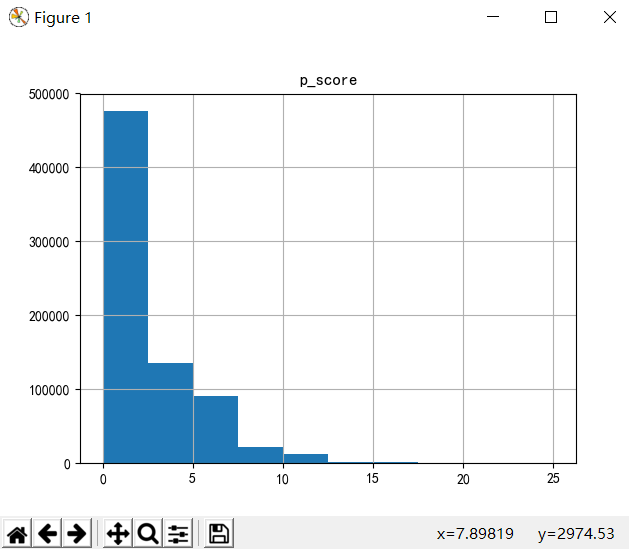


2.数值属性：以“p\_score”属性为例，分别给出了空值数据的个数、平均值、最小值、四分位数、最大值。

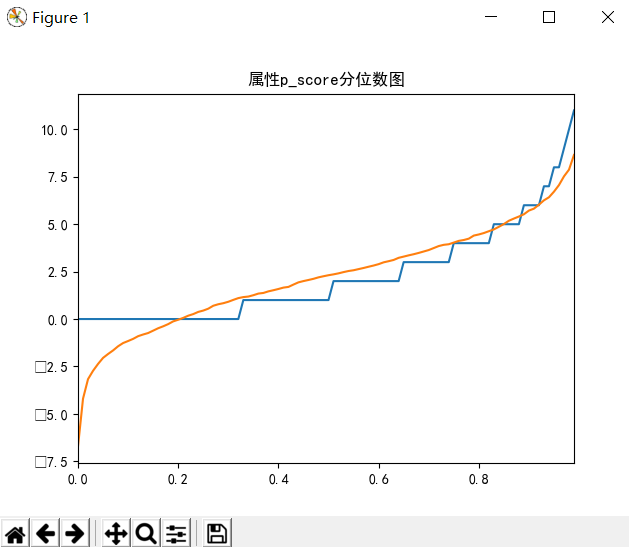


1. **数据可视化**

1.直方图



2.Q-Q图：



3.盒图:

