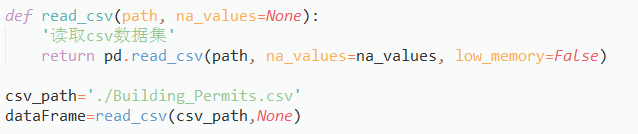
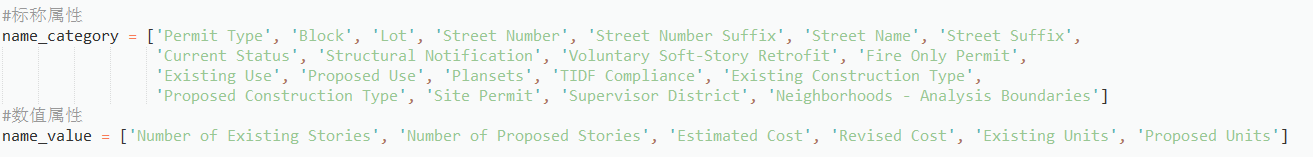
**Building\_Permits数据分析与预处理**

一、数据读取和属性分类

利用Python中的pandas库进行csv数据文件的读取：



对属性进行分类，分为标称属性和数值属性：



二、数据可视化和摘要

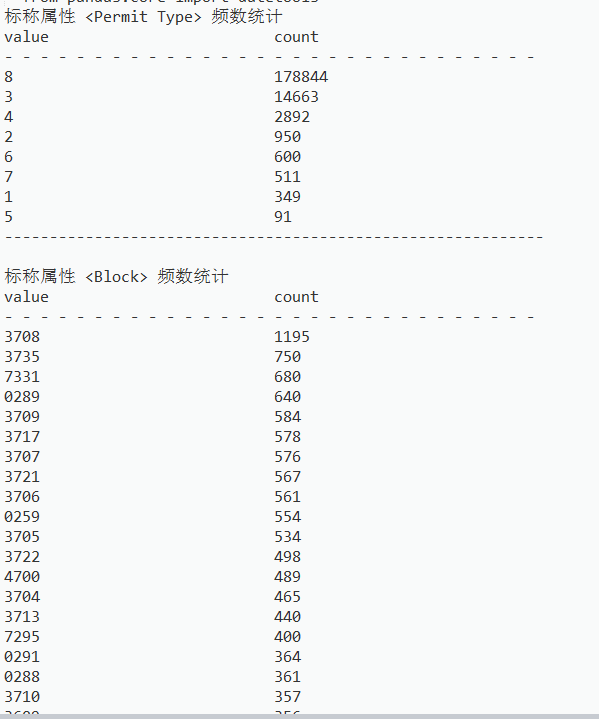
2.1 数据摘要

**标称数据：**

对于标称属性，使用pandas中的value\_counts函数统计每个标称属性的所有可能取值的频数。

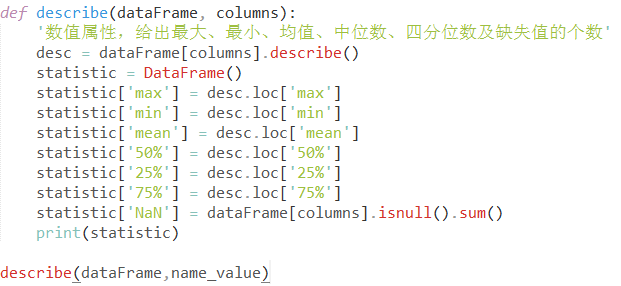


部分结果如下：

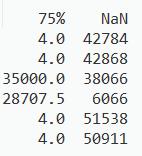
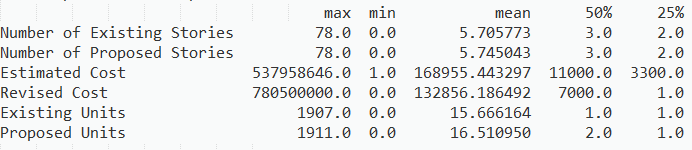


**数值属性：**

对于数值属性，使用pandas中describe()函数给出其最小、最大、均值、中位数、四分位数及缺失值个数：



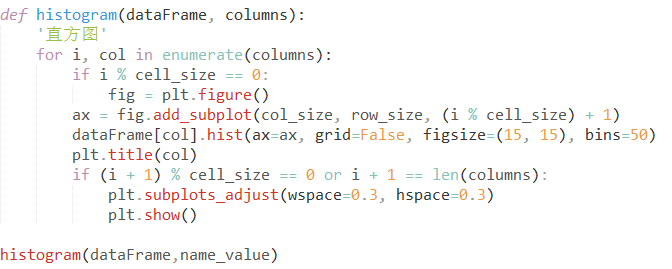
部分结果如下：



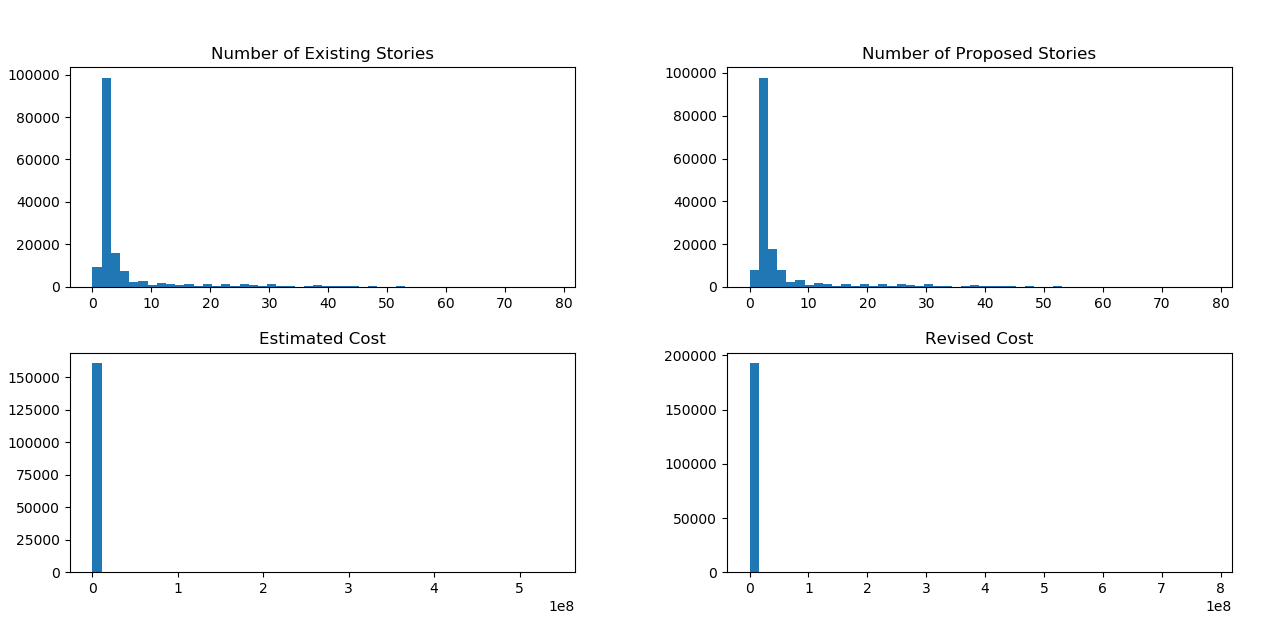
2.2 数据可视化

**直方图**

使用matplotlib绘制直方图

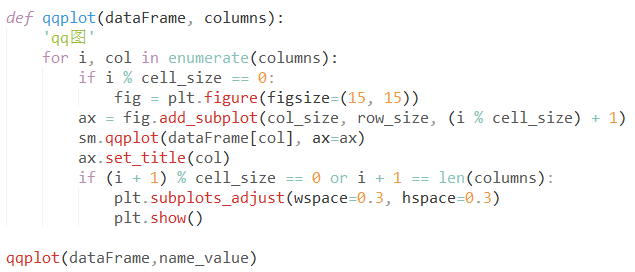


部分结果如下：

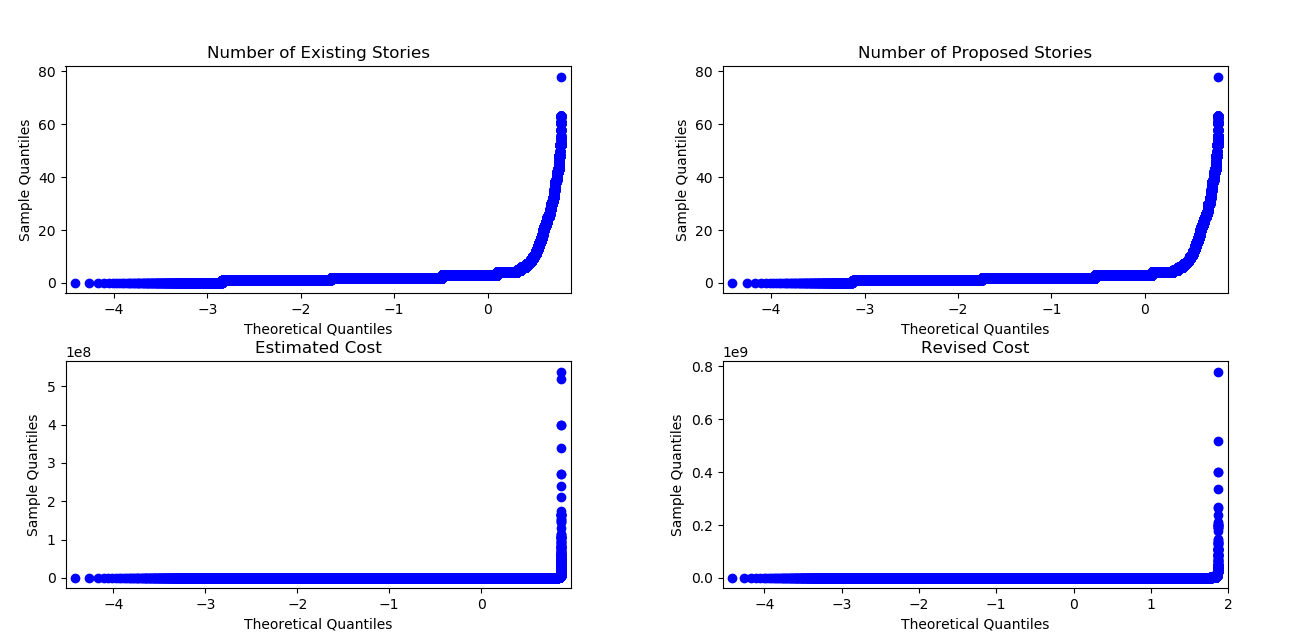


**qq图：**

使用matplotlib绘制qq图



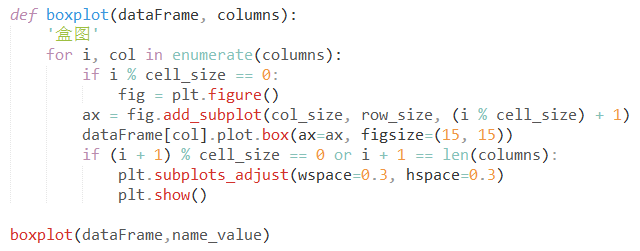
部分结果如下：



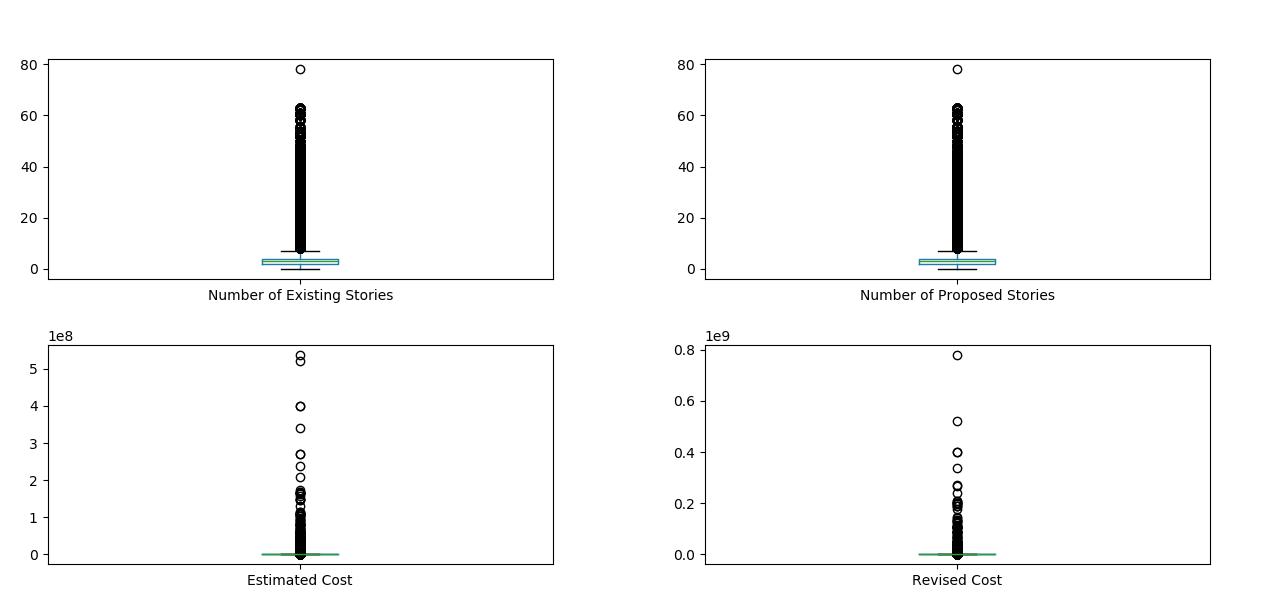
根据qq图可知图像若是近似直线的，其对应属性为正态分布态。

**盒图：**

使用matplotlib绘制盒图，对离群值进行识别：



部分结果如下：



三、数据缺失处理

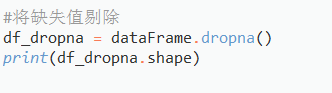
3.1 将缺失部分剔除

无缺失的字段：Permit Number，Permit Type，Permit Type Definition，Permit Creation Date，Block，Lot，Street Number，Street Name，Current Status，Current Status Date，Filed Date，Record ID 。

无填充意义的字段：Unit，Unit suffix，Description，Issued Date，Completed Date，First Construction Document Date，Permit Expiration Date，Existing Construction Type Description，Proposed Construction Type Description，Zipcode，Location，Street Number Suffix，Street Name Suffix，Existing Use，Existing Units，Proposed Use，Proposed Units，Plansets，Existing Construction Type，Proposed Construction Type，Supervisor District，Neighborhoods - Analysis Boundaries，Number of Existing Stories，Number of Proposed Stories，Estimated Cost，Revised Cost。

可填充的属性字段：Structural Notification，Voluntary Soft-Story Retrofit，Fire Only Permit，TIDF Compliance，Site Permit

可填充的属性字段中除了TIDF Compliance均为布尔型，空表示否，可用N填充；TIDF Compliance字段只有两条记录不为空，空表示否，可用N填充，对其进行剔除。



输出：



说明剔除缺失部分数据后，数据集为空。

3.2 用最高频率值来填补缺失值

无

3.3 通过属性的相关关系来填补缺失值



部分结果如下：



3.4 通过数据对象之间的相似性来填补缺失值

无