该任务主要是清洗爬取的数据和对几个有趣的属性进行标记化。

清洗数据：

我爬取到了很多数据，但是，“细节”一栏存在问题，有如下问题：

1.“细节”是字符串，包含了大量信息而且尚未分类

2.“细节”存在部分未知属性，需要清除无效数据

直接利用csv转换成xlsx文件即可，利用xlsx文件的功能能快速实现细节的分类，无需使用python编写程序进行清洗。

无效数据包含以下几种情况：

1.装修为“其他”

2.版型为“暂无数据”

3.楼层为“空”

4.单价为“空”

变量数值化;

有些变量已经是数值了，比如面积，单价，楼层。

有些变量需要进行数值化，方便统计。比如版型，装修，朝向等。

可视化：

上述讨论可以知道，有些变量可以数值化，但是地址无法进行数值化，因此需要进行进一步的判断，首先先统计各个二手房的地址

我对房子的发布非常感兴趣，因此想做出房子地址的词云图。