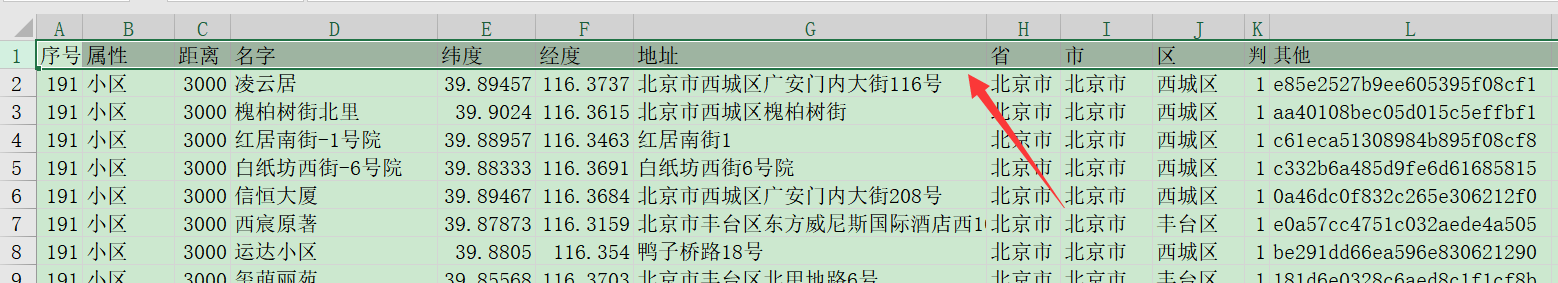
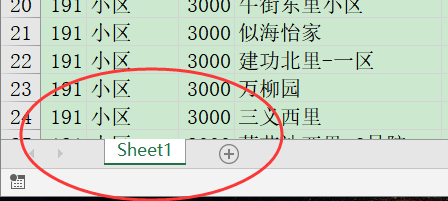
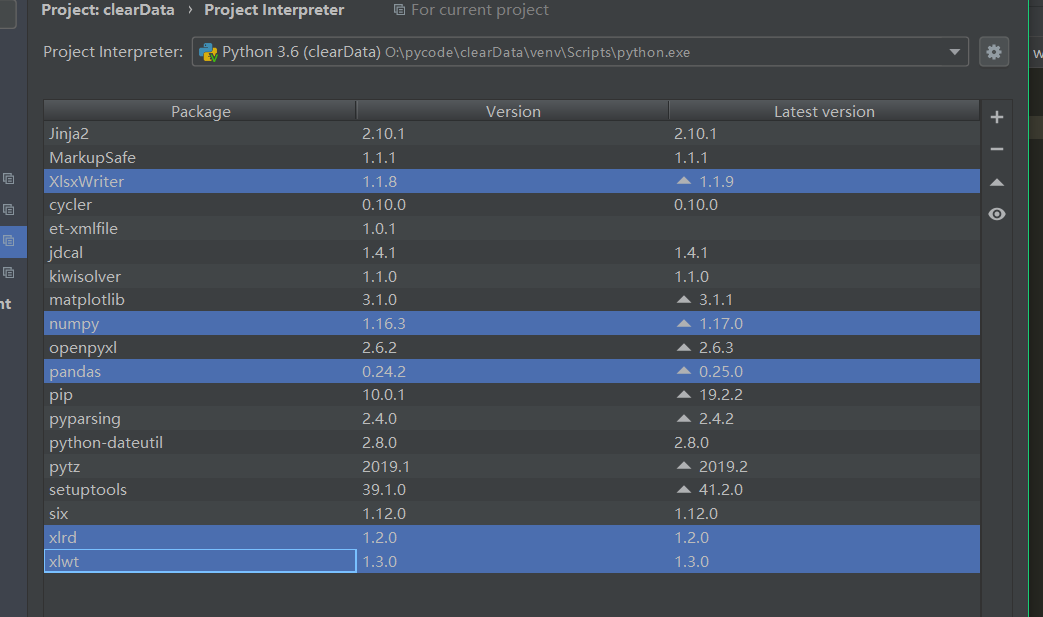
### 程序运行与excel清洗注意事项：

#### Excel表格预处理：

需要处理第一行(表头)，如图  
  
A列到L列的值需如图一致！！！  
属性值处理，属性列（B列）的数据值不能有空格，清除所有空格，在做属性统计时，有空格的和没空的是不一样的属性如：‘商场’和‘商场 ’。  
sheet的名字需要为‘Sheet1’，‘序号’列需要升序排列！  


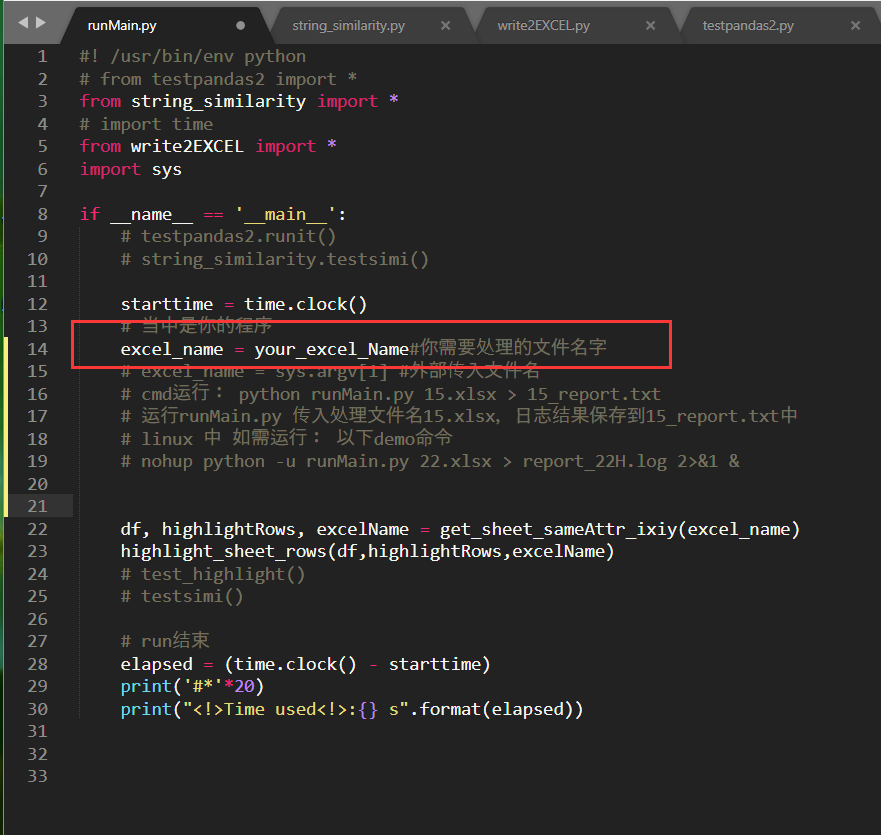
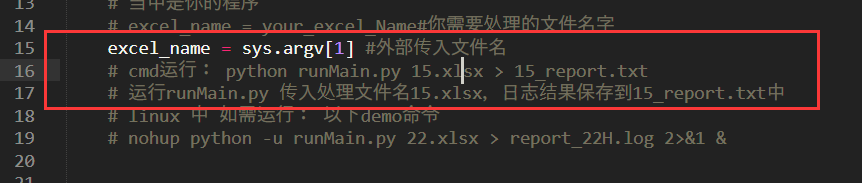
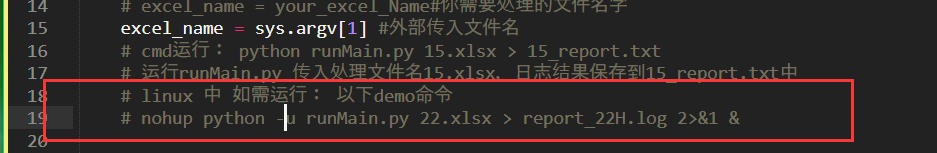
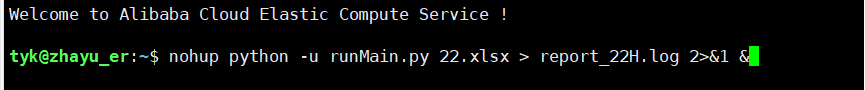
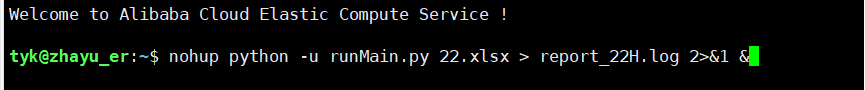
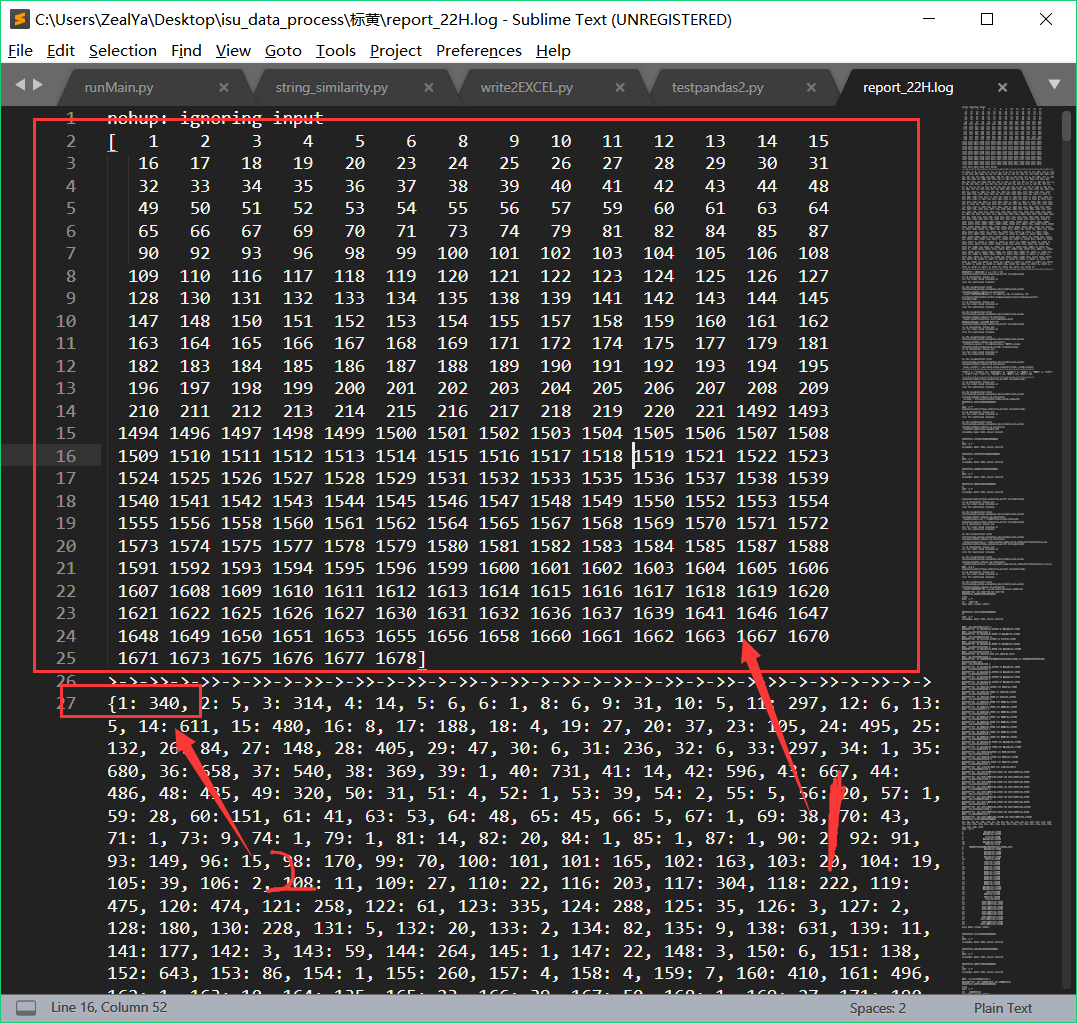
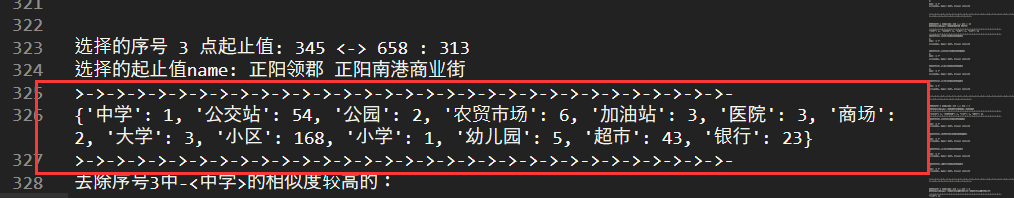
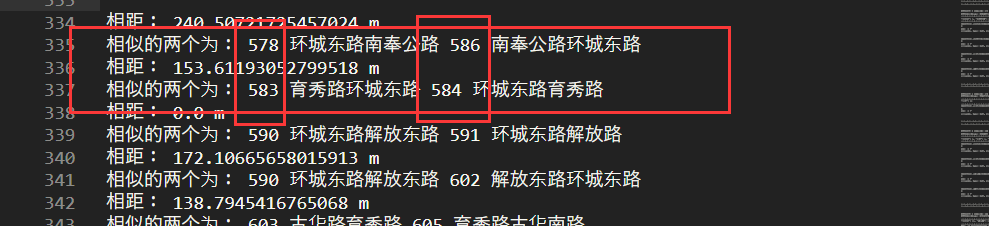
#### Python所需库：

Python3.6，蓝色高亮库



需要安装这些包才能正常运行程序。

#### 运行程序

1. 直接运行  
   
2. Cmd运行  
   
3. Linux运行  
     
     
   cmd和Linux运行需要文件名参数传入，所以excel\_name = sys.argv[1]不能注释掉
4. 运行结果展示：Linux为例  
     
   处理22.xlsx文件，输出报告于report\_22H.log（名字自己设置）中，如下：  
     
   箭头1所示为该excel的sheet1中所‘序号’列中包含的不同序号点，箭头2所示为1序号有340条记录，后面的也是如此。  
   下图是序号3中，各个属性对应的记录数目，比如 中学：1个  
     
   在检查到一个序号中同一属性下的两个名字比较相似且距离很近的，将标黄：  
     
   处理结果输出在中  
     
   这两个名字为相似，且有很近，标黄，但是实际情况为两个公交站，在后续清理时不能删除！  
     
   这两个名字正确的标出，但是都是xx小区x门，都要删除，不属于小区的范畴。

#### Excel清理（手动）

结合输出报告（report\_x.log）来手动清洗表格：

以22.excel和report\_22.log为例：

标黄的基本上都可以删除，不确定时参考一下log文件



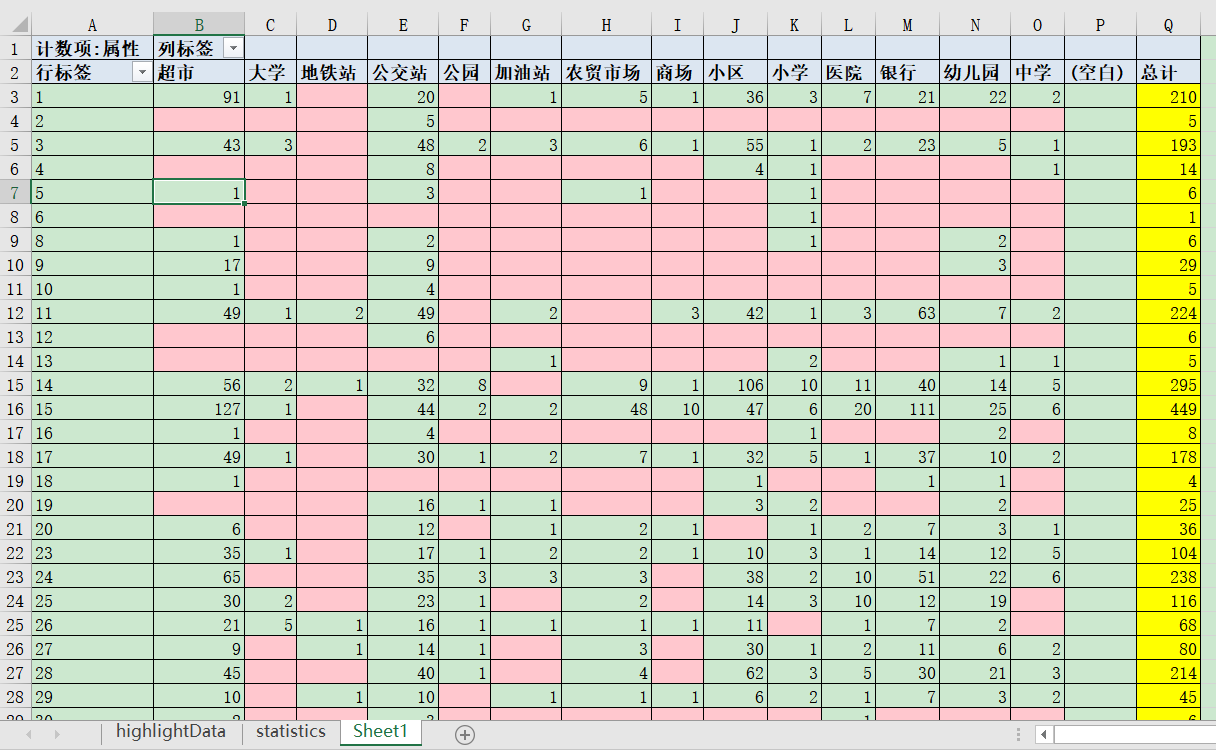
比如上图红框中是两个值在表中对应的序号



***删除规则：***

|  |  |  |
| --- | --- | --- |
| **小区** | xx小区xx号（楼）、xx小区xx门、xx小区xx停车场(停车场)、xx小区管理处（物业）、xx店（xx小区店）、标黄的 | 删除 |
| **医院** | xx医院xx楼、xx医院xx中心、标黄的 | 删除 |
| **商场** | 标黄的 | 删除 |
| **超市** | 标黄的 | 考虑删除 |
| **加油站** | 标黄的 | 考虑删除 |
| **幼儿园** | 标黄的 | 考虑删除 |
| **小学** | 标黄的 | 删除 |
| **中学** | 标黄的 | 删除 |
| **大学** | xx大学、xx大学-xx楼（校区），xx儿童大学，看起来不像大学的培训机构，标黄的 | 删除 |
| **公交站** | 标黄的 | 考虑删除 |
| **银行** | 标黄的 | 考虑删除 |
| **农贸市场** | 标黄的 | 删除 |
| **地铁站** | 基本上没有删除的 |  |

处理完表格后，新建一个sheet，做一个数据透视：



**詹御**

**2019.08.22**