高开轩  
CS201-第二期数据分析项目  
gkx516830031@gmail.com

项目报告

# 写在前面

首先，很抱歉我的项目完成的非常粗糙和潦草，大概只完成了基础项目（结业证书）的大部分的内容，另外勉强完成了拓展项目（优秀证书）的pool策略的爬虫，而且大部分都是copy老师的代码，自己只修改了很少的东西。我想我应该为这个潦草的项目做一个解释，下面是我的几点原因。

1. 我基本上是零基础参加这个课程的，我现在是学材料工程的大四学生，之前能够和数据分析搭上边的经历就是参加过几次数学建模比赛，和C语言的基础课，另外我之前自学过10天左右的JavaScript。这导致我在上课的过程中很难跟得上老师的节奏，只好课后看视频和Google，再加上问问TA来解决一些问题，光解决各种运行环境的问题就几乎让我心力交瘁。。。。。。。。。
2. 我现在正处在申请季，在上课的这一个月期间又刷了托福（虽然刷的很烂），我实在没有能拿出太多精力来做这个项目，其实我做这个项目的目的也是为了在我的文书中能够有相应的project经历。
3. 我同时在bittiger上了三门课程，我发现我太贪心了，我并不能在繁忙的申请季handle这三门课程，虽然难度都是入门级的，但我还是需要很多时间来消化。

基于上述原因，我的项目完成的很差。但是我觉得如果能再给我三个周左右的时间的话，我大概能很好的补充完相应的基础知识，很好的完成这个项目和相应的拓展方向，我的现在R和sql还基本上处于只知道老师上课讲的东西上，只知其然而不知其所以然，稍微变一点东西就要Google，还不一定能解决，唯一我自己能做一点拓展的地方就是数据爬取的部分了，我修改了老师的pool的策略，并且同时爬取了两个页面的数据。

# 项目报告

### 数据爬取

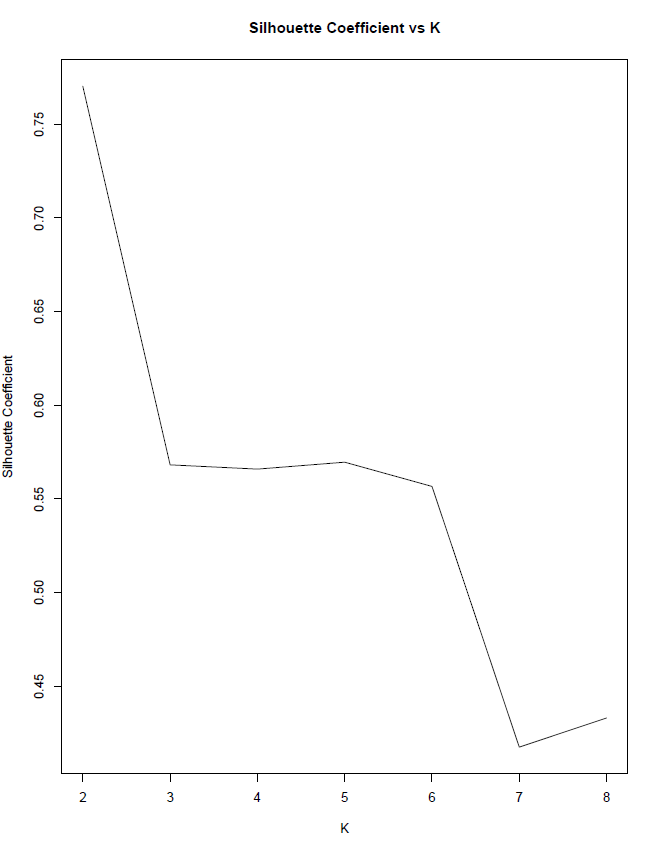
爬取上海链家网的小区数据，包括小区名字，经纬度，物业费，物业公司，开发商，均价，房龄，房屋类型。

我是在老师的v9基础上增加了一个parser2.js，用来爬取小区详细信息页面上的数据，并且把爬取的速度调整了一下。最后爬下来20000条左右的数据。存入到mongodb里面，然后导出为output.csv文件.

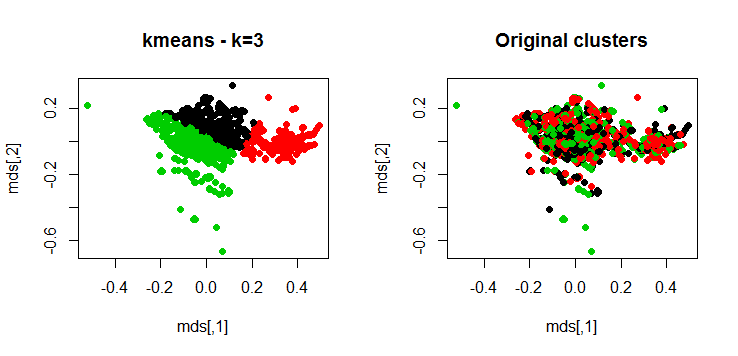
文件放在homework\_spider文件夹里面。

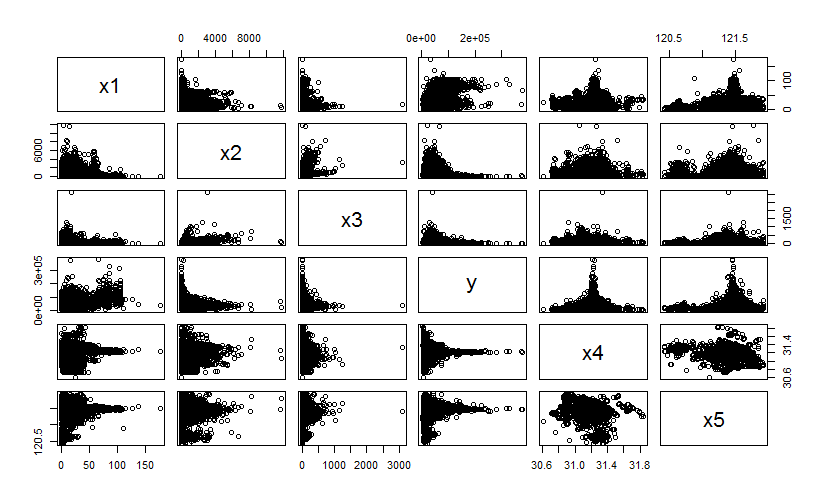
### 数据分析

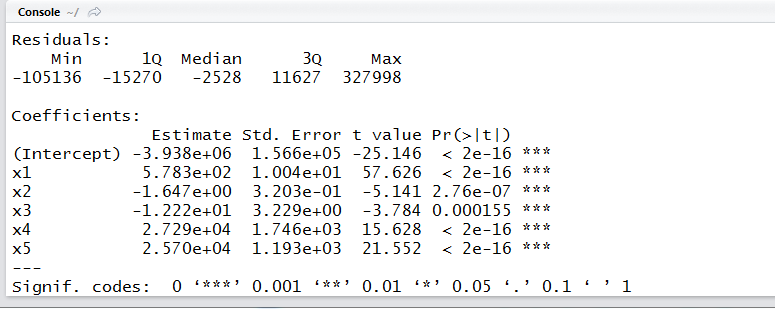
对爬取到的小区的经纬度进行K-means聚类分析，我先用excel剔除了lng和lat为0的数据和异常值，然后进行K-means聚类，结果因为数据太多，20000多条数据，我的电脑处理不了，于是选取了10000条数据进行计算。取k为2到8，round为1，进行计算。计算结果为k = 3时最佳。



选取k=3，下面用plot输出



房价预测方面，我将csv文件中的经度、纬度、住户数、楼盘数、房龄抽取出来，并且用excel将Null值替换为NA，然后在RStudio中剔除NA值，画出房价与这5个元素的散点图，从而建立房价与这5个元素的多元回归方程。



P值均小于0.05，所以方程可以建立。预测结果分别为41793和39023.

### 数据可视化

说实话，这部分是照搬的老师的代码，目前我还没有能力做出像样的修改。

用leaflet做出小区在上海地图上的分布。

把链家小区的坐标标在上海市的地图上，并且可以根据房价来调整坐标点的大小和颜色。坐标点的大小和范围可以自己设定。

文件放在homework\_visualization文件夹里面。

## 项目体验

虽然完成的不是很好，但是我感觉这仍然是一个很nice的项目，特别是对我这样的零基础的，想要转专业的同学来说。这个项目让我了解了data analyst的工作流程，更重要的是，让我对数据分析有了更强烈的兴趣。我知道了应该重点掌握什么样的技能，同时这也是一个很重要的奠基课程，我至少在短时间内掌握了很多以前很想知道的知识和技能。Anyway，我很开心参加这个课程。