# 问题求解与实践大作业报告

——纽约市 2019 年 Airbnb 开放数据统计与房租预测

## 运行说明：

通过运行命令行进入解压后的文件的 data_statistics_and_visualization 文件夹中，运行名字为“NY.exe”的文件（exe 不太清楚需不需要，在自己电脑上运行时没加 exe 能正常运行），运行详情可看 demo 视频。

## 一．数据输入

在 kaggle 上获得的数据为 csv 格式，文件共 16 列，均为线性结构，并且考虑到我后续预测与可视化部分均不需要排序搜索等操作，因此直接将所有数据储存于类成员结构体数组中。

## 二．脏数据的发现和处理

**id、host_id 项：**为 airbnb 系统生成的随机数值，无分析意义。数据读入时跳过读取。

**name、host_name、last_review 项：**数据与其余数据关联性较小，也不具有可视化效果，同样不具有分析意义。数据读入时跳过读取。

**neighbourhood 项**：该数据与 neighbourhood_group 属性重复，两者相比选择分类性较好的后者，能够更好的进行预测分析。数据读入时跳过读取。

**number_of_reviews 0 项：**在数据筛选过程中，而 airbnb 的统计规则中，没有 reviews 即代表无入住，考虑到其中复杂的原因，无法将其放入价格预测中来，但这部分数据数据量较大，仍可在数据可视化中体现。

**Price 异常项：**存在部分个体价格过于高或直接为 0，这部分数据所占比例低于 2%，同样不具有预测意义且对预测结果造成不小影响，因此在数据读入时判断后跳过，后选择价格预测中舍弃该部分数据。

## 三．数据的统计与可视化

### 1.纽约市 Airbnb 民宿分布

得到的数据包含民宿经纬度信息，因此我找到纽约市城市地图，将所有租房标记于地图之中。将数据按照所属区块分类，以不同的颜色标记。因此能够从图

像中得到纽约市五个区的民宿分布图，并从中了解其大致分布，对于纽约市airbnb 的租房分布有个大致印象。

**2.五大区域价格分布**

分析每个住房数据的信息，最具有分析价值的便是 price。按照每个租房的区块分化，分别统计各个区域价格分布。为了取得良好的可视化效果，将步长设置为 10，最后绘制各个步长点的价格连线图。从图中能够得到各个区域价格分布随价格的变化，以及各个区域之间的横向对比。为了体现真实分布效果，价格统计可视化这一项中我并未清除价格异常数据。

**3.Price-Number_of_reviews 散点图**

这一项可视化，我决定服务于后续预测部分,画出 Price-Number_of_reviews 的散点图，能够的到两者间的大致分布关系，以便后续预测模型的选择。

## 四．趋势预测

之前通过 Price-Number_of_reviews 的散点图能够大致看出两者之间有一定的相关关系，考虑到目前学习的技术较少，就采用线性回归。后续几周又学习机器学习中的多元线性拟合，调用 c++能使用的 mlpack 库完整多元线性回归的实现。最终比对两者的结果，实现对价格的预测。我将 csv 顺序读取的前 30000 个数据作为训练集，剩余的数据作为验证集，判断模型的准确性。

一元线性回归：除去脏数据之后，需要在剩下的数据中选择与 Price 线性性较好的数据列。筛选之后（筛选过程不做展示，均为计算相关系数），得到相关系数最大的为 Number_of_reviews，但系数仍低于 0.75，初步判定两者线性相关性较差。

多元线性回归：模型形式为

$$y = a_1x_1 + a_2x_2 + a_3x_3 + a_4\chi_4 + a_5\chi_5 + a_6\chi_6 + a_7x_7 + b$$

调用 mlpack 库，使用多元线性回归对剩余的非脏数据列进行模型计算，得到各个参数对应系数和常数项，详见附页。在程序中直接使用计算出的模型。

模型准确度判断，一元线性回归相关系数 R 小于 0.75 已经可以判断其不具有线性相关性。但仍将其放到最后的预测比对中去。多元线性回归决定使用校正

决定系数 $R^2\_adjusted$ 判断其准确性，消除了样本数量和特征数量的影响,最终计算得 0.324247<0.4，预测效果较差，但相对于一元线性回归准确性大大提高。计算公式如下：

$$R^2\_adjusted = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

**总结感悟：**

本次大作业自己的数据集虽然方便处理，但是数据效果并不是很好，目前学习的知识并不能满足对于该数据的准确处理和预测，最后还是自学机器学习中的多元线性回归，通过调用 malpack 库进行预测，但最后结果在自己的预测方式来看，仍旧差强人意，若是以后能系统的进行学习，相信能训练出更好的模型。本次大作业通过对于数据集的处理，提升了自己的代码能力，并且主动学习未知的东西，对于我自身也是很大的提高。也非常感谢两位助教和两位老师对我的帮助，感谢，祝好。

**附页：**

多元线性回归参数表：

| | |
|---|---|
| Latitude/a1 | 160.345588 |
| Longitude/a2 | 684.218346 |
| minimum_nights/a3 | 0.150842297 |
| number_of_reviews/a4 | 0.143023417 |
| reviews_per_month/a5 | 0.339119434 |
| calculated_host_listings_count/a6 | 0.167945419 |
| availability_365/a7 | 0.134275895 |
| b | -57002.2532356945 |