# 基于链家网数据分析上海市租房报告

## 一 项目背景

根据国家统计局显示，2018年上海市在岗职工平均工资142983元，居民人均可支配收入64182.65元，2019年人均可支配收入69442元，全国居民恩格尔系数28.2%。以2019年数据为例，每人每月平均可支配收入为5786.83元。本项目通过分析链家网上海市租房数据来分析房租在可支配收入中所占比例。

## 二 数据获取方法

本项目使用Python编写网络爬虫，通过爬取链家网上海地区租房数据进行分析，涉及上海全部16个区县、204个区域，获得有效数据13605条。

所获得的数据包括房源ID、标题、预览图、房源所在区县、区域、小区、房源面积、户型、朝向、楼层、租金、标签、房源信息维护时间、装修类型、出租方式、支付方式、押金、服务费、中介费、房源与地铁的距离、配套设施、基本信息、房源描述以及轮播图等。

所获得的数据中，房租单位为“元/月”，房源面积单位为“m²”，配套设施（电视、冰箱等）均以“TRUE/FALSE”来判断是否配备，标签、房源到地铁的距离、轮播图均将多个数据存储于一个单元格内，多条数据间以英文分号“;”进行分割。

因为爬虫存在被链家网封禁的情况，在重连过程中可能造成部分数据存在问题，故以单条数据中存在至少35个非空子数据为标准来判断该数据是否有效。

## 三 数据概览

上海市平均房租为11320元，平均每平米租金为111.99元，平均租房面积为103.73m²。

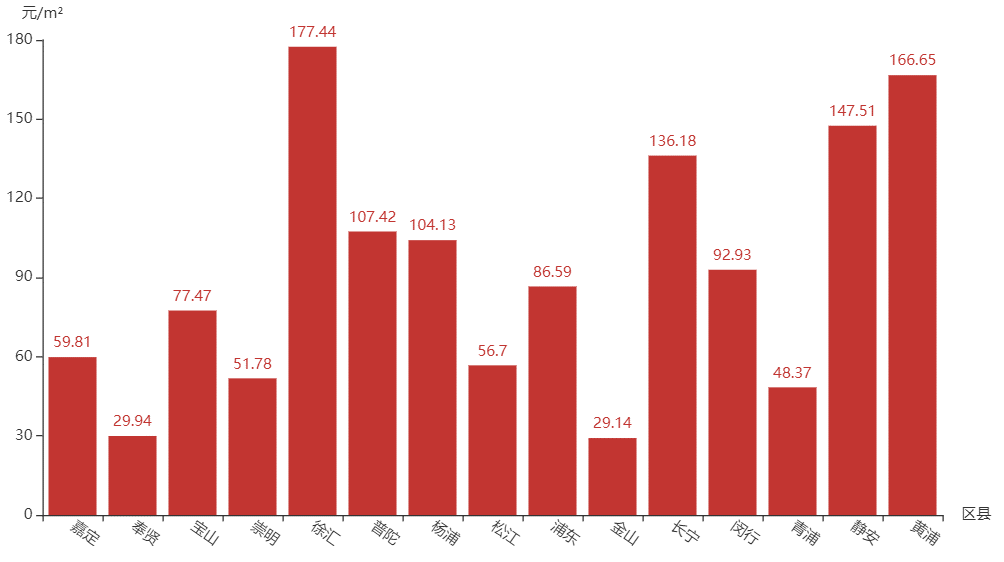
从图1可以看到，靠近市中心的徐汇区每平米租金最高，为177.44元，高出平均租金58.44%。次之是最繁华的黄浦区，每平米租金为166.65元，高出平均租金48.81%。而奉贤区和金山区平均租金最低，仅不到30元/m²。

图 1 上海市租房均价

各区房源数量占比如图2所示。从图中可知，浦东房源占比最多，有2551条数据。松江次之，有1400条数据。浦东与松江的房源数量多，租金水平中等，适宜人们租住。

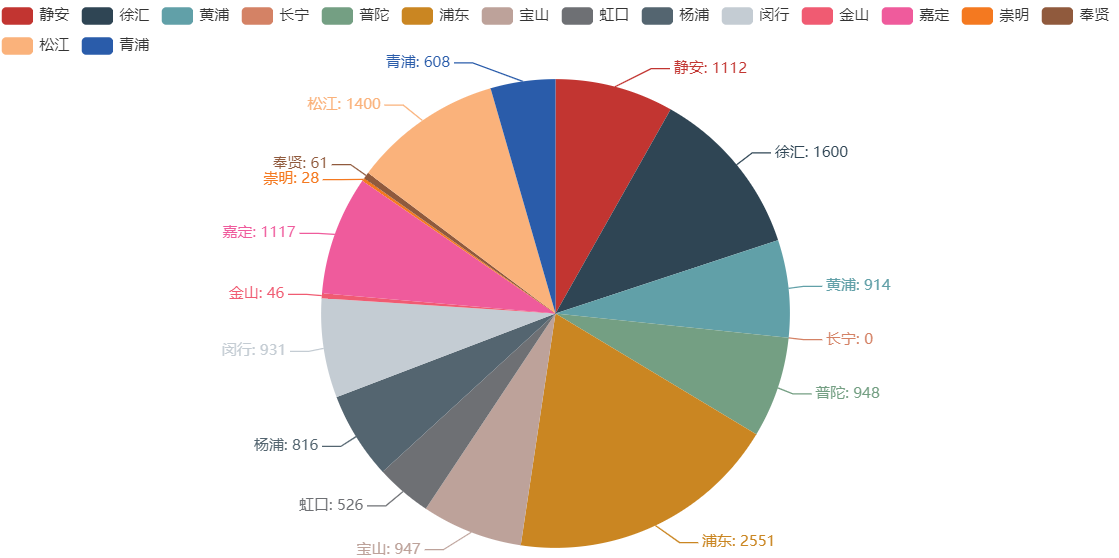


图 2 上海市房源数量

通过对所有有效数据进行分析，获得了热点词云，如图3所示。从图中可以看出房源的宣传重点主要在于小区周边的交通设施以及附近的公园超时等便民店。这说明不仅房源本身影响租金，其周边各种配套设施也会影响到租金。



图 3 房源描述热点词

图片包含 屏幕截图

描述已自动生成通过对面积、距离、楼层、租金进行统计，获取到数据直方图，如图4所示。从图中我们可以看出，绝大部分房源面积小于400m²，楼层低于40楼，租金低于40000元，链家网上所统计地铁到房源的距离只到1200m。故对所获到的数据进行筛选，最终保留了8728条所需的数据。

图 4 部分信息统计直方图

对经过处理后的数据进行统计分析，所获得的数据图如图5所示。经过处理后的范围有所缩小，数据更集中，便于进行后续统计。处理后的数据中，租金的平均值为9371.28元，最低为1000元，最高为39500元。租金偏度为1.8277，为右偏态，即租金小于房租均值的房源数量多于大于均值的房源数量，上海租户图片包含 屏幕截图

描述已自动生成的住房成本能控制在一定的范围内。

图 5 部分数据经过处理后的统计直方图

通过对房租及多种因素进行相关性分析，可以获得房租、平均房租受到面积、楼层、距地铁站的距离以及房源配套设施的相关性统计图，如图6所示。需要注意的是，因水平有限，区县、区域等因素没能考虑进去，并且链家网上部分房源的数据录入不全，所以数据结论可能存在一定的误差。从图中我们可以看到，房租受面积的影响最大，相关性为0.68；次之为楼层，相关性0.45。由此可以简单推断出，随着面积增加，楼层增高，房租也会相应增长。但从每平米平均房租上来看，与面积的相关性为-0.16，可以简单理解为随着面积增大，每平米平均房租会有所下降。房租与房源内配套设施的相关性大体在0.2-0.3之间，每平米平均房租与配套设施的相关性大多在0.25-0.29之间。虽然相关性不强，但是完善的配套设施还是可能会增加一定的租金。同时，从图中也能看出，床、电视、冰箱、空调、洗衣机、热水器有着强相关性，即这些设施大多是一齐配备的。楼层与是否有电梯的相关性为0.74，即楼层越高，房源配备电梯的可能性就越大。

图片包含 屏幕截图

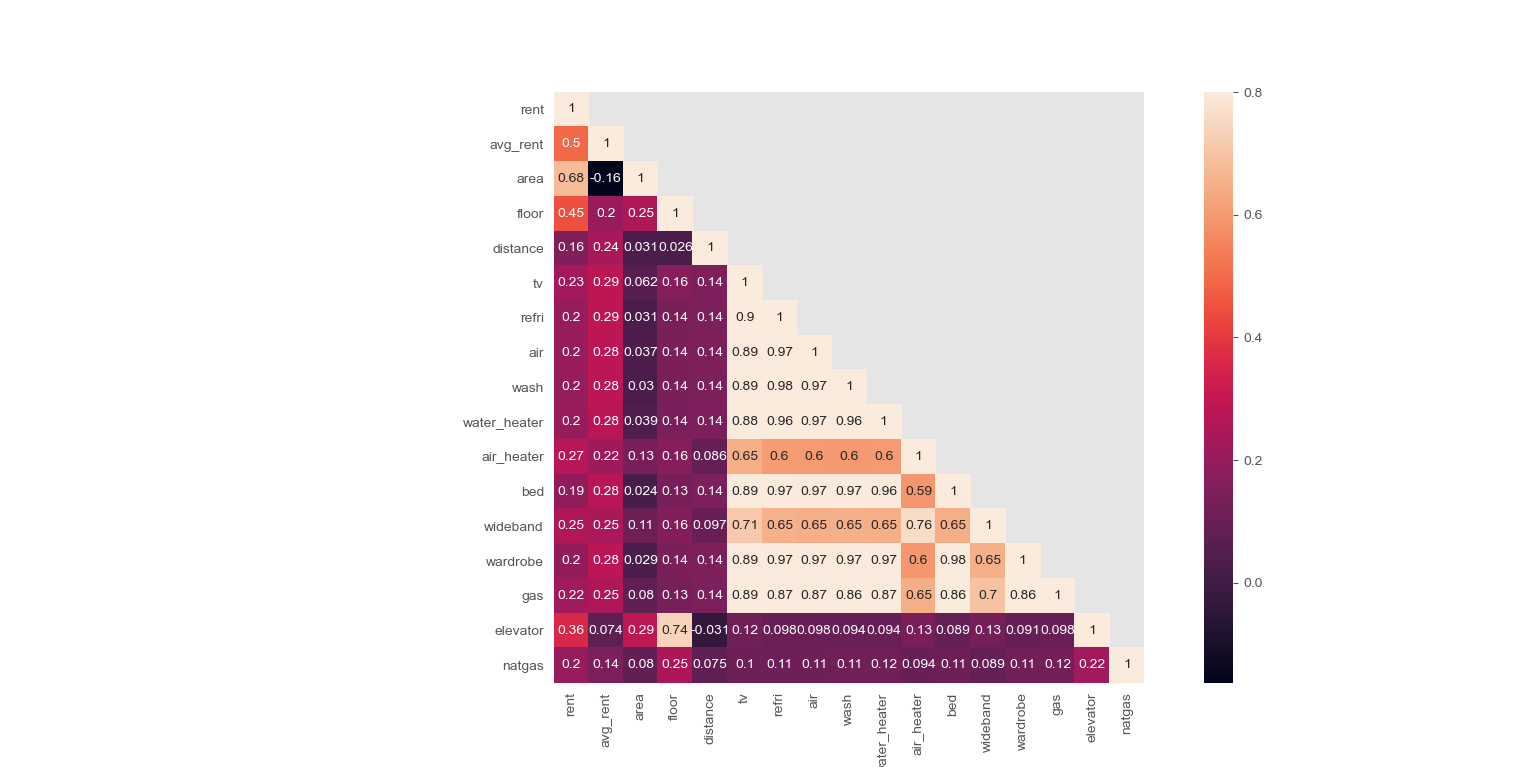
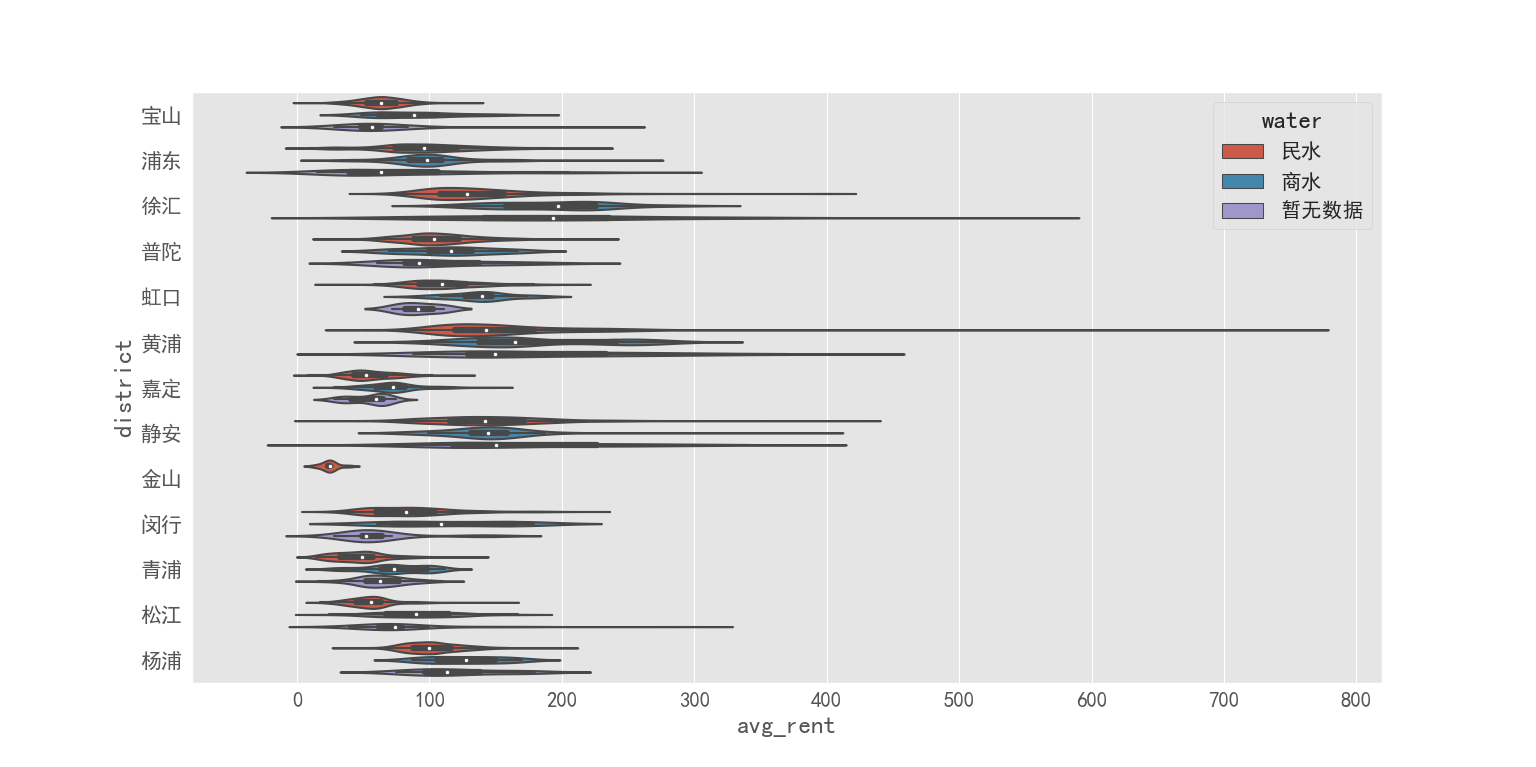
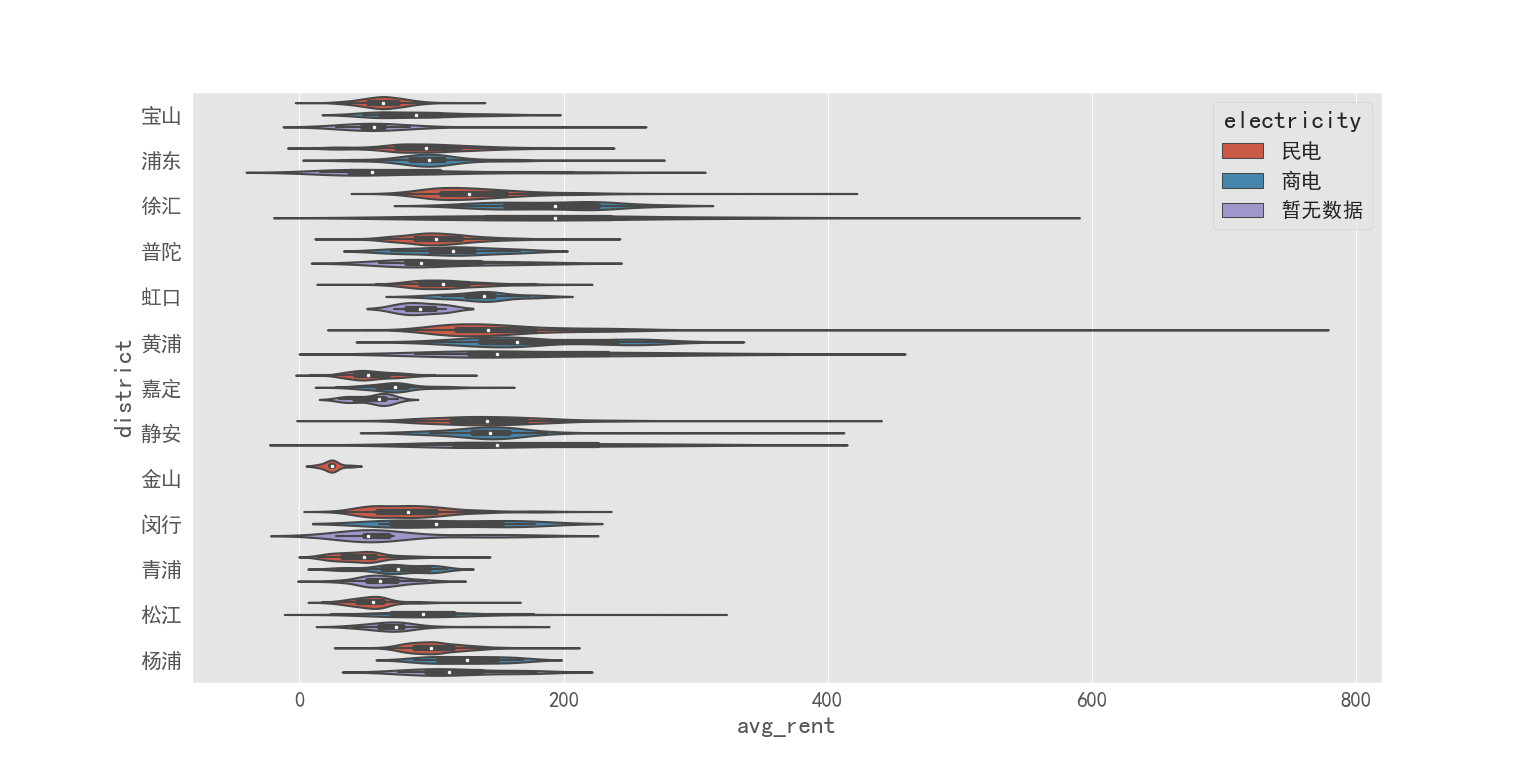
描述已自动生成为了验证配套设施对房租的影响，以床为例，按区县进行分组，我们通过小提琴图来进行统计，可得图7来表示是否配备床造成的房租范围差异。但因为部分房源信息中，配套设施数据没有录入，所以默认为无，可能造成一定程度上的数据误差。从图中可以看出，配备了床的房源的最低租金要高于没有配备床的，同时大部分区县（如：宝山、普陀、嘉定）等，配备了床的房源的房租中位数要略高于没有配备床的。据此，我们可以得出结论，设施对房租是有一定的影响。

图 7 房租受是否配备床的影响

图 6 房租与部分因素的相关性统计

图 9 房租受用电类型的影响

图 8 房租受用水类型的影响

根据经验可以提出猜想：房源的用水用电类型也会影响房租。按区县进行分组，以房租作为横轴，使用小提琴图进行统计分析。可以看出，使用民水、民电的房源具有更广的房租范围，使用商水、商电的房源的最低房租要高于使用民水、民电的，且中位数也高于使用民水、民电的。经查阅可知，使用商水、商电的多为公寓或者写字楼，因此租金相对来说会有所偏高。

预测模型可以在一定程度上揭示事物间的内在规律性，可以把它作为计算预测值的直接依据，故选择一个好的适当的预测模型是非常重要的。本项目分别建立了GBDT回归模型和随机森林回归模型，用于进行比较。因为类型限制，故横坐标由面积、楼层、房源到地铁站的距离以及配套设施组成，纵坐标为每平米平均房租。将本项目中爬取到的数据分成两批，一批作为训练集，占75%；一批作为测试集用于评估模型的质量，占25%。经查阅可知，GBDT回归模型在参数调整合适的情况下要优于随机森林回归模型。但因硬件设施不足，本次项目没能自适应调整其相关参数。在模型训练结束后，通过公式(1)进行质量评估：

(1)

由该公式可得，GBDT回归模型获得的评估分为0.5195，所获得的对比图如图10所示；随机森林回归模型获得的评估分为0.5279，要略优于GBDT回归模型，所获得的对比图如图11所示。两种模型的评估分数都不高，说明仍然存在影响房租的因素未能考虑在内，但拟合度较高，模型有效。

图片包含 物体

描述已自动生成

图 10 GBDT回归模型图

图片包含 物体

描述已自动生成

图 7 随机森林回归模型图

## 四 结论汇总

在网上查询资料可得，房租在可支配收入中占比应为1/5到3/1之间。结合上海房租每平米均价可知，居民在上海所能承受的租房面积平均约为10.33 m²-17.22 m²。但目前上海平均房租为11320元， 平均面积为103.73m²，远超人均可支配收入，大大提高了居民在上海的生活成本。

影响房租的因素有很多，不同的区县、不同的楼层，甚至不同的配套设施都会影响到房租的高低。在预算有限的情况下，租户可以选择面积较小、楼层较低、配套设施不那么完善的房源来租住，从而降低租房成本。

## 五 后续改进

因时间有限，之前也很少接触数据挖掘与分析，所以在这四天左右的时间中大部分时间都用在查资料和改错误上，导致参考的数据、分析图表不多，分析不够精确，未能发现明显影响房租的因素。

在后续的学习中，需要解决以下问题：

1. 因为爬虫程序会受到目标网站的封禁，所以需要考虑以下几点：
   1. 增加断点续爬功能，减小因为中断导致的时间浪费；
   2. 增加代理池，随机更换代理伪装，减小被封禁的可能性
2. 学习数据分析相关内容，学会数据建模。在额外增加的一天中，主要精力都投入在了分析上。但是随着我了解的越多，越发的感觉我学要学习的更多。