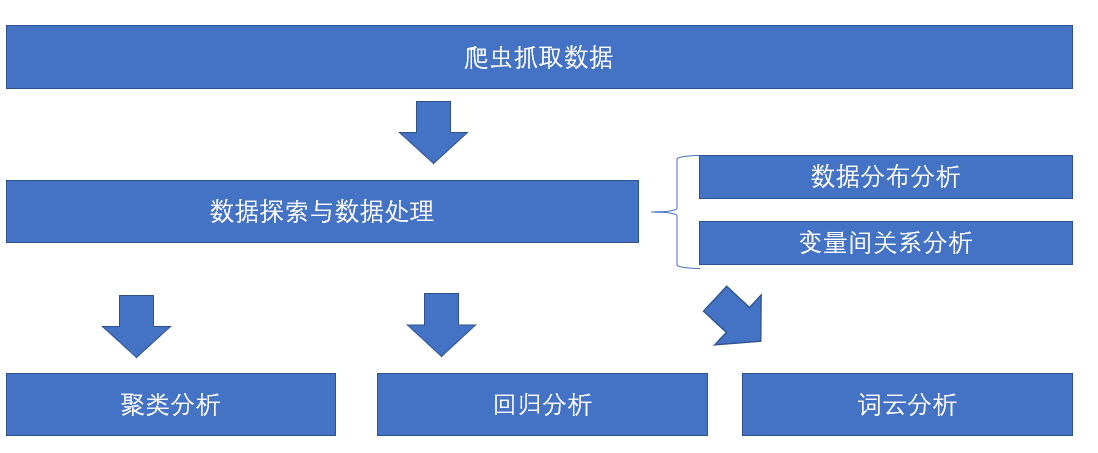
上海链家网租房数据分析报告

# 1分析简介

上海链家网租房数据分析框架如下图所示

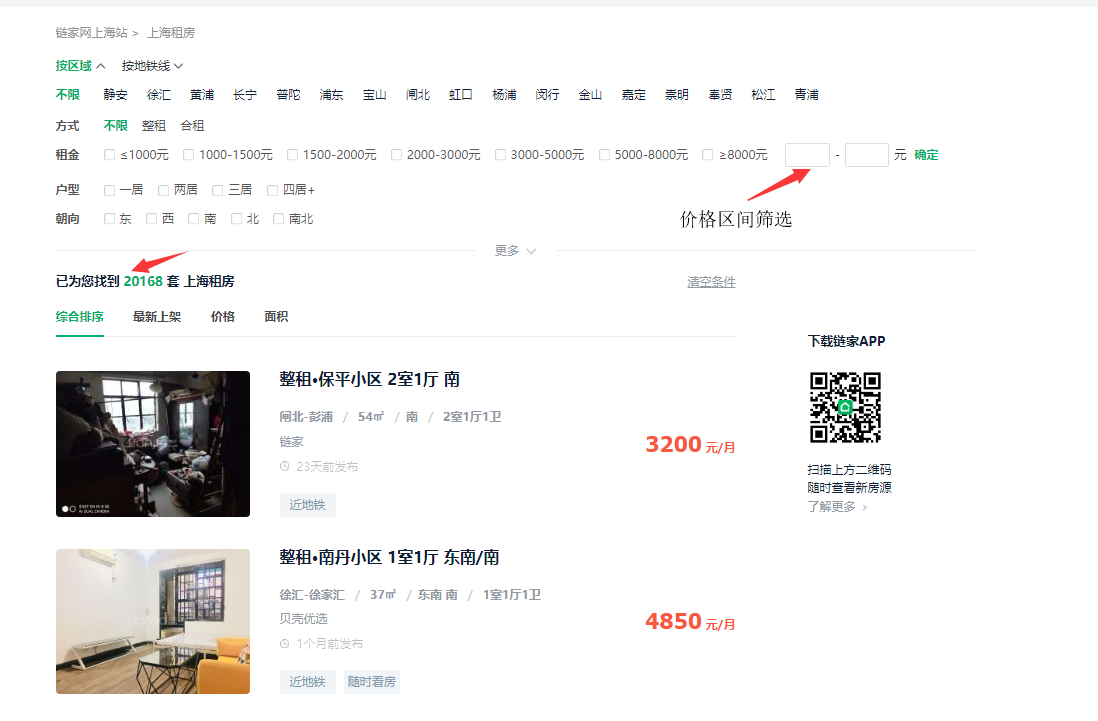


本文首先使用爬虫程序抓取房源数据，之后对数据进行了探索与处理，最后对进行了聚类分析、回归分析以及词云分析。

# 2数据获取

本文使用python scrapy框架对上海链家网租房(https://sh.lianjia.com/zufang)数据进行抓取。

网站页面如下图所示，



租房网站上，已知的租房房源大约在20000套，而该站点一次只显示3000条，单一页面下，无法一次性抓取，因此需要采取别的措施。本文通过价格区间筛选，构造了初始链接，对20000余套房源进行了抓取。



根据网站上的信息，本文抓取了每套房源的id(房源编号)、area(区域)、adress（地址）、community（小区）、shelf\_time（上架时间）、rent（租金）、house\_type（户型）、square（面积）、towards（朝向）、floor（楼层）、describe（介绍）、subway（地铁信息）、img\_url（图片链接）、url（租房房源链接）、download\_time（采集时间）。

最终爬取18805条房源数据，如下图所示，大致接近所有租房房源数据，可以进行整体分析。爬虫代码见lianjiaSpider文件夹。

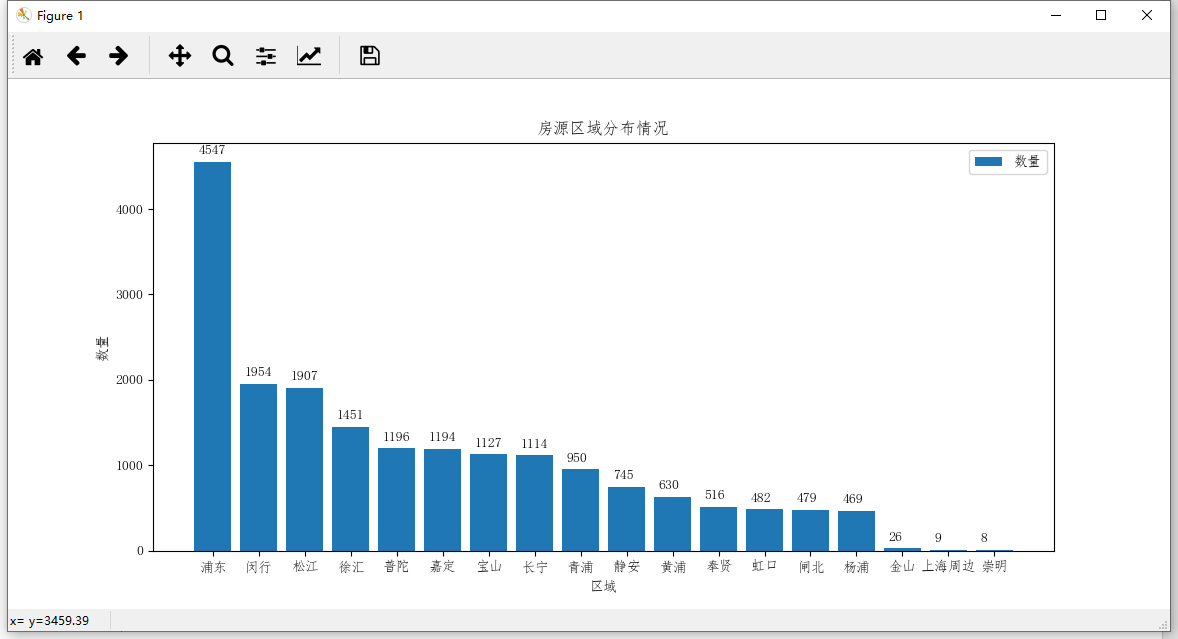


# 3数据探索与处理

本章对租房房源的数据进行了分布分析以及数据间相关情况分析，并针对数据的情况，进行了数据处理。

## 3.1数据分布分析

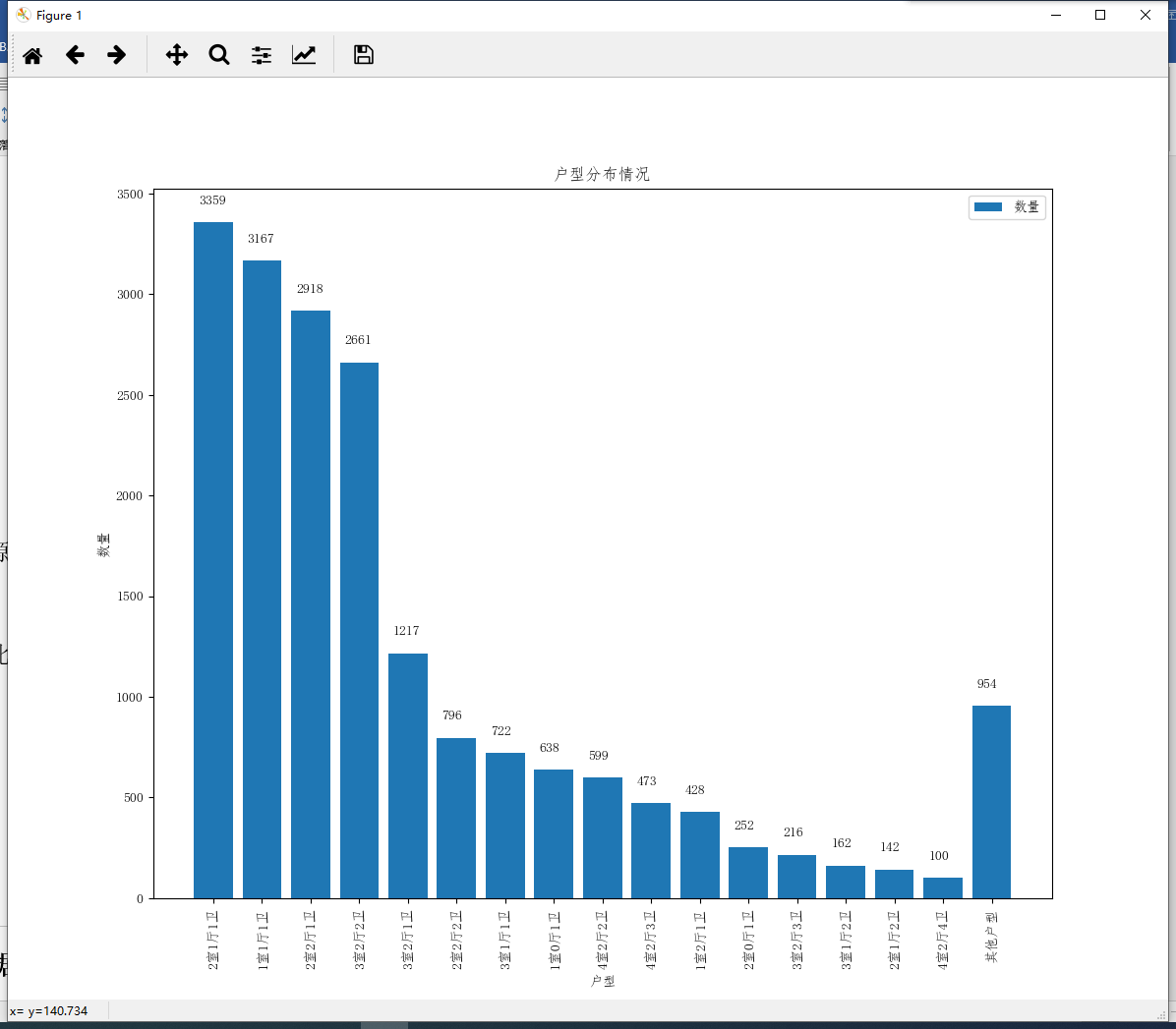
### 3.1.1房源区域分布



从图中可以看到，浦东的租房房源数量最多，而金山、上海周边、崇明则几乎没有房源，其余地区房源均有一定的数量。

### 3.1.2户型分布

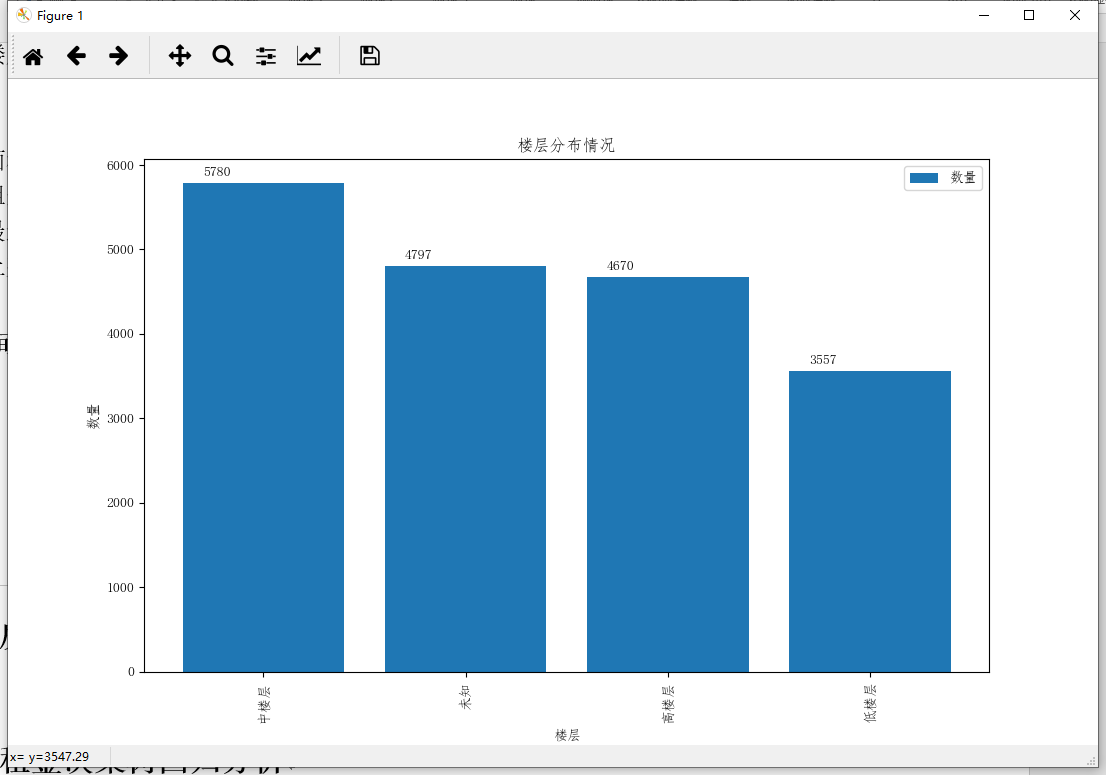
统计发现户型的种类多达122种，而大部分户型市场上的数量很少甚至只有1个，因此考虑将户型数量小于100的归为其他户型，之后绘制分布图如下。



从图中可以看到，市场上数量较多的户型为2室1厅1卫、1室1厅1卫、2室2厅1卫、3室2厅2卫，可见这些户型较受市场欢迎。

### 3.1.3楼层分布

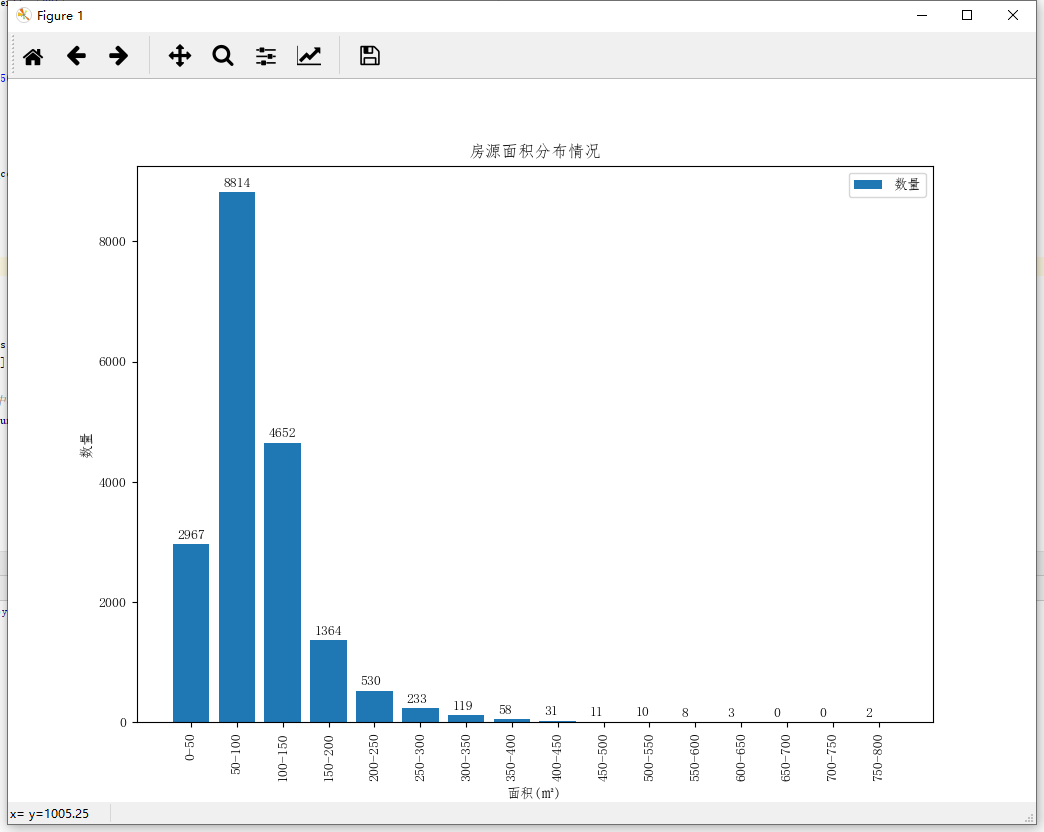
统计发现市场上的楼层描述为楼层类型/总楼层，因而楼层的水平多达171种，为了便于分析，统一将楼层划分为低楼层、中楼层、高楼层三类，剩余的归为未知楼层，分布如下图所示。



从图中可以看出，市场上各个楼层均有一定的数量分布，能够满足多样化需求。

### 3.1.4面积分布

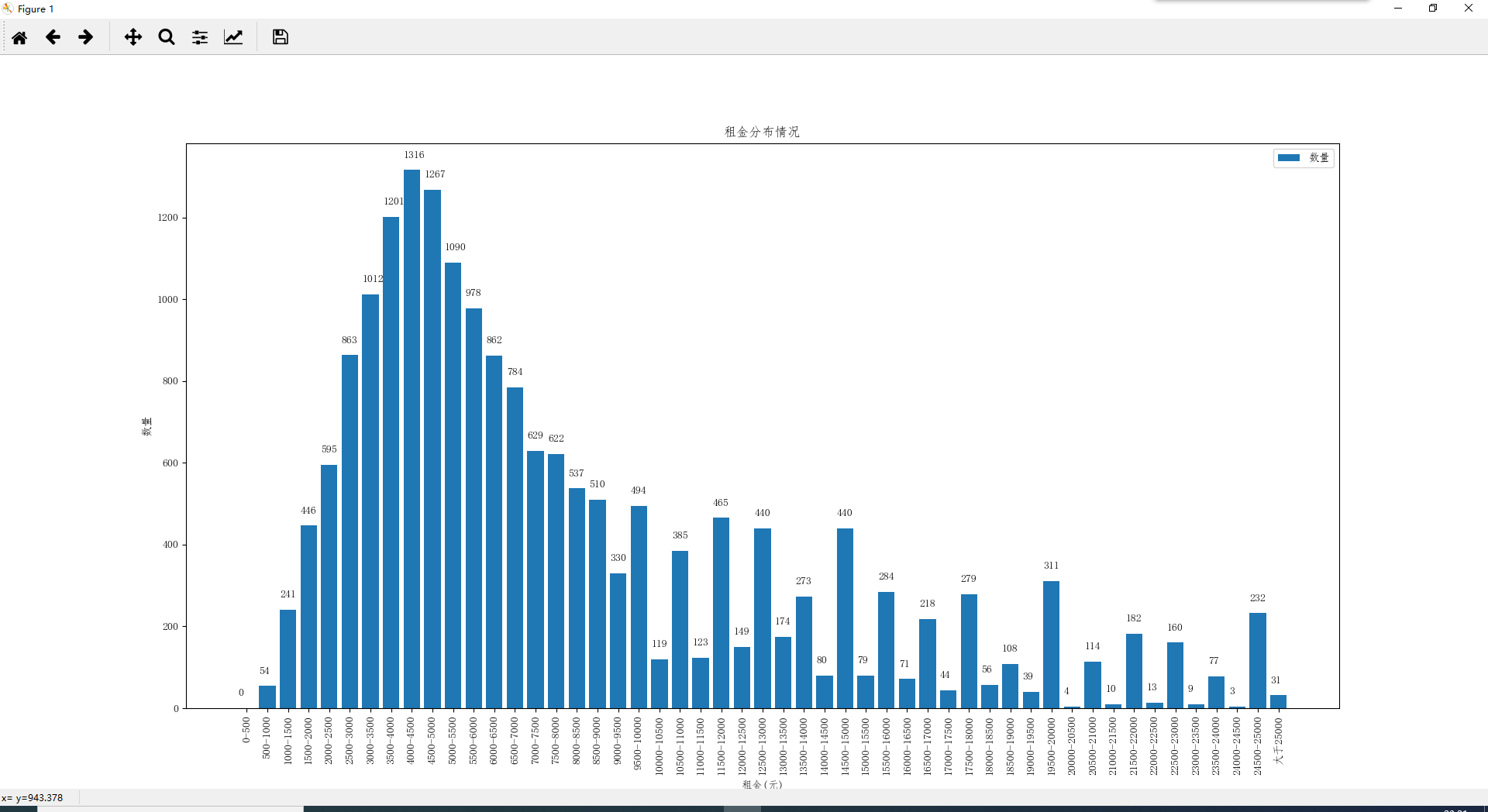
面积数据属于连续数据，需要进行离散化处理才能绘制分布图，本文将面积数据按照一定的区间进行划分，之后绘制分布图，如下所示。



可以看到，租房房源面积集中分布于0-150㎡，而超过300㎡的房源数量较少，房源面积分布整体呈右偏分布。

### 3.1.5租金分布

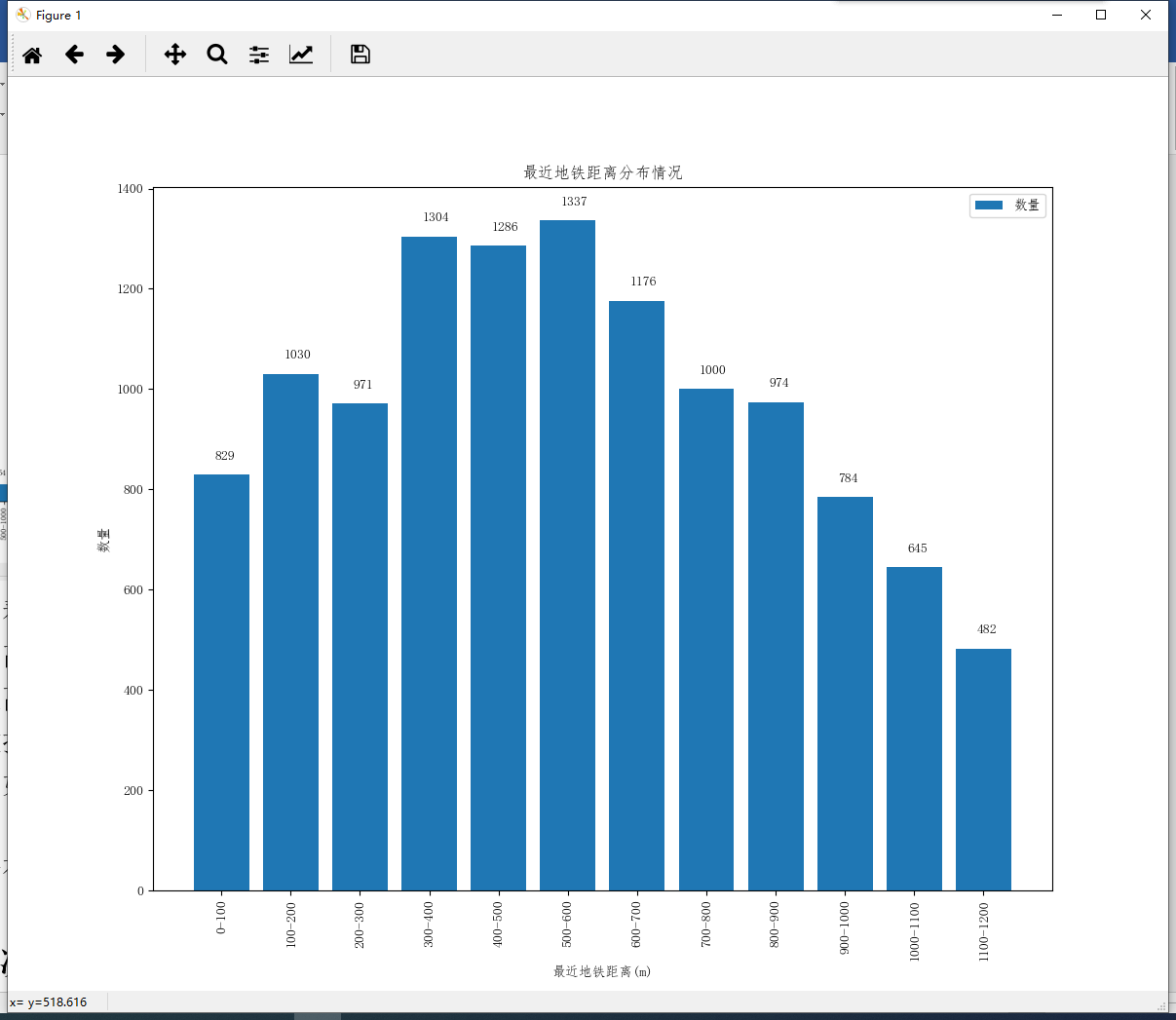
租金数据也为连续数据，因此也需要离散化处理，分布图如下。



从图中可以看到，租金价格主要集中于2500-6500元/月。0-500元/月这一区间没有相应租房房源，而在高价位仍有房源分布，可见上海租房市场的繁荣。

### 3.1.6最近地铁距离分布

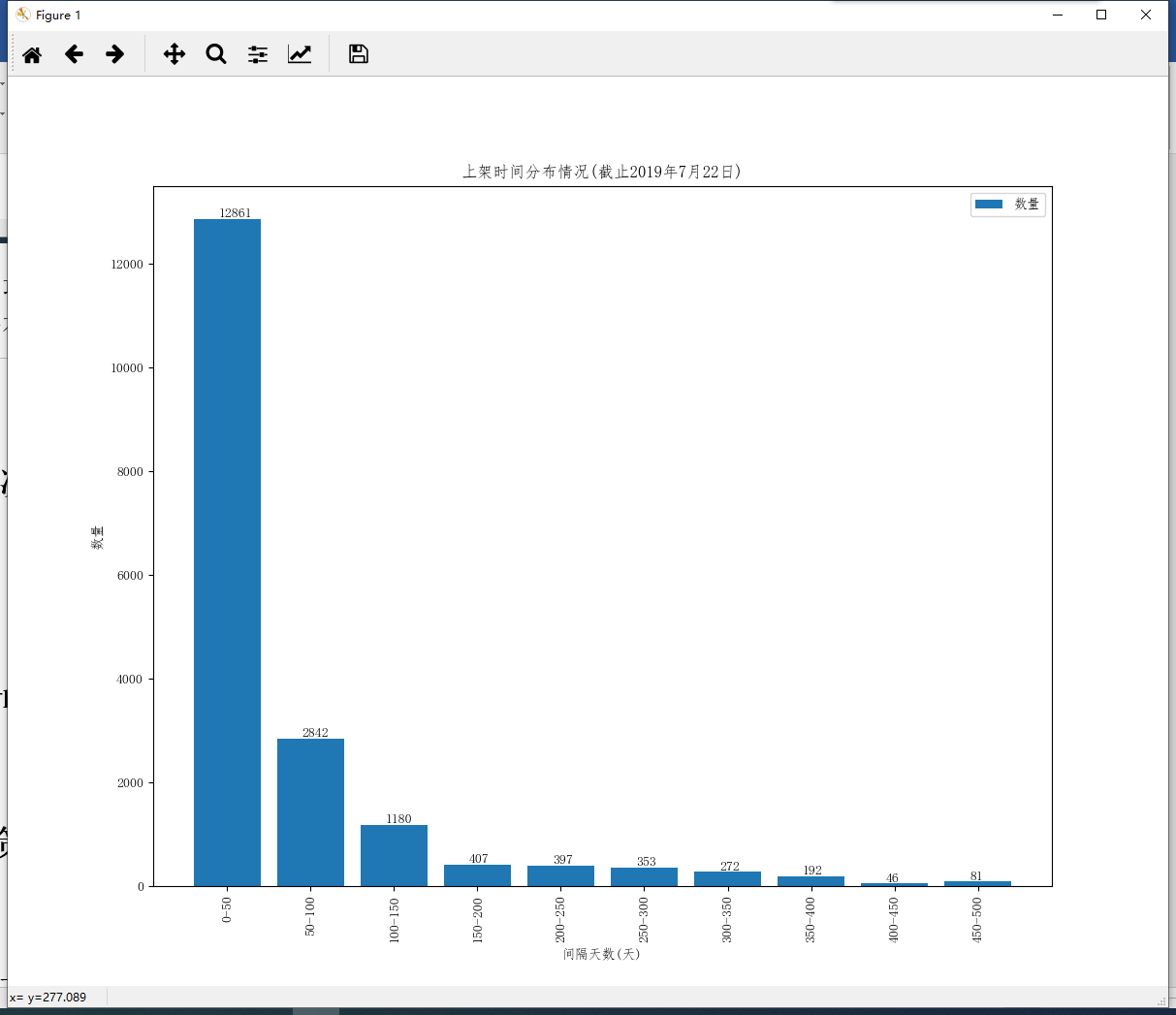
部分房源数据没有地铁相应数据，因此本文只对有地铁数据的房源进行了最近地铁距离分布分析。本文首先对抓取得地铁数据进行了最近距离提取，随后绘制了分布图如下。



可以看到，最近地铁距离在各个区间均有分布，无明显特征。

### 3.1.7上架时长分布

本文根据采集时间以及房源发布时间进行了上架时长计算，之后绘制了分布图如下。

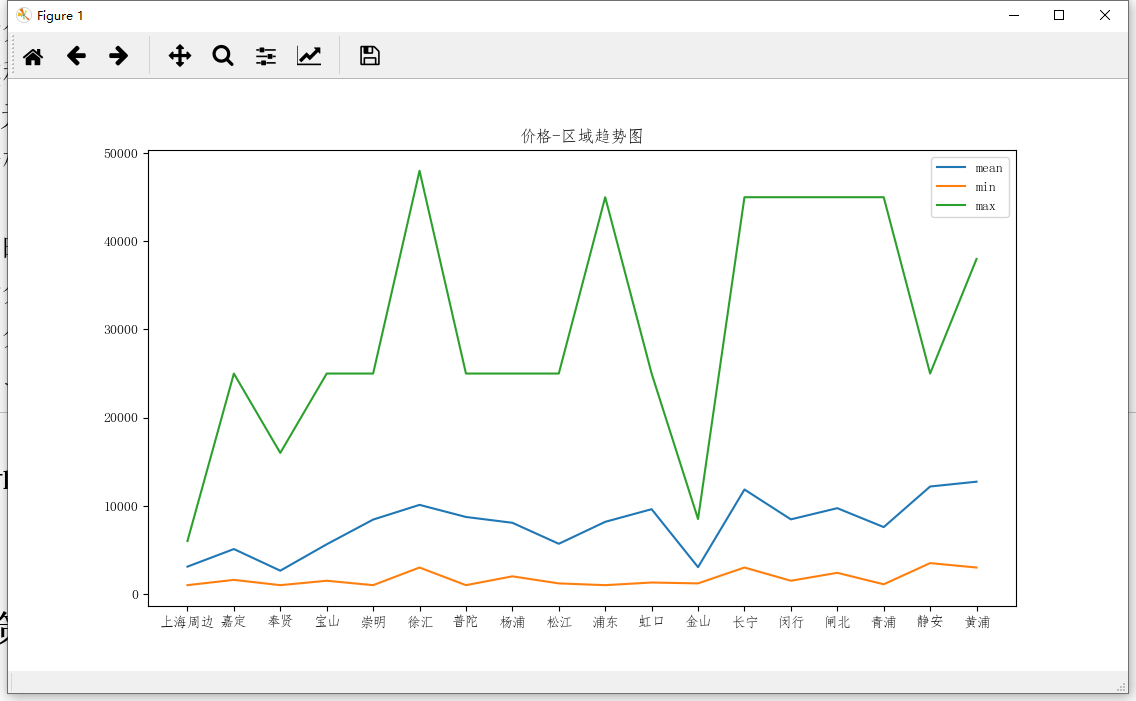


可以看到，上架时长呈分布呈递减趋势，时间越短数量越多。

## 3.2数据间相关情况分析

### 3.2.1区域与租金分析

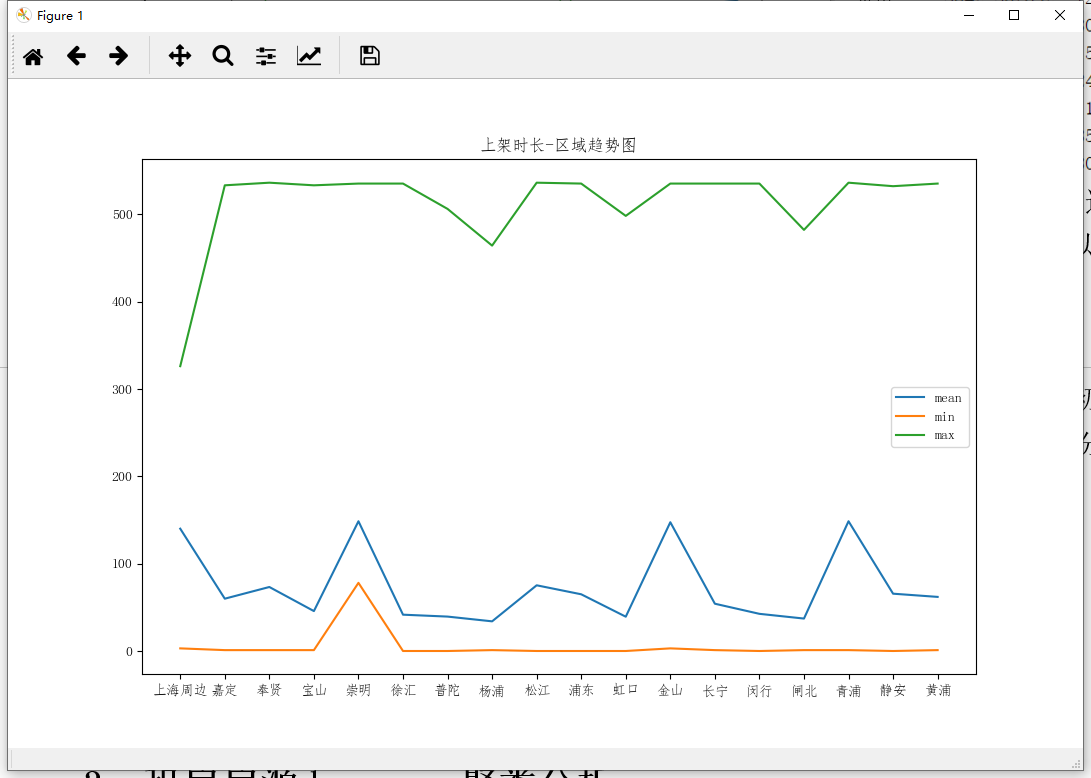
分析各区域租金时发现，某些地区的部分房源租金非常之高，例如长宁区有套房源租金高达42万元/月([链接](https://sh.lianjia.com/zufang/SH2249769180654485504.html))属于极端数据，为了便于分析，本文只保留了租金小于50000元/月的数据进行分析，下图是各区域的租金情况趋势图以及相对应表格。

可以看到徐汇、长宁、静安、黄浦的租房价格相对较高，平均价格超过了10000元/月，而房源数量最多的浦东地区的租房均价则位于中游。从上表也可以看到，上海地区的租房最低为1000元/月。

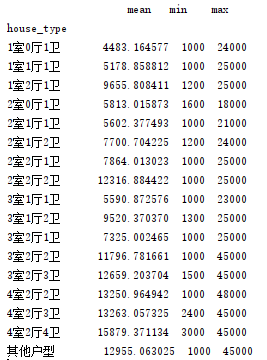
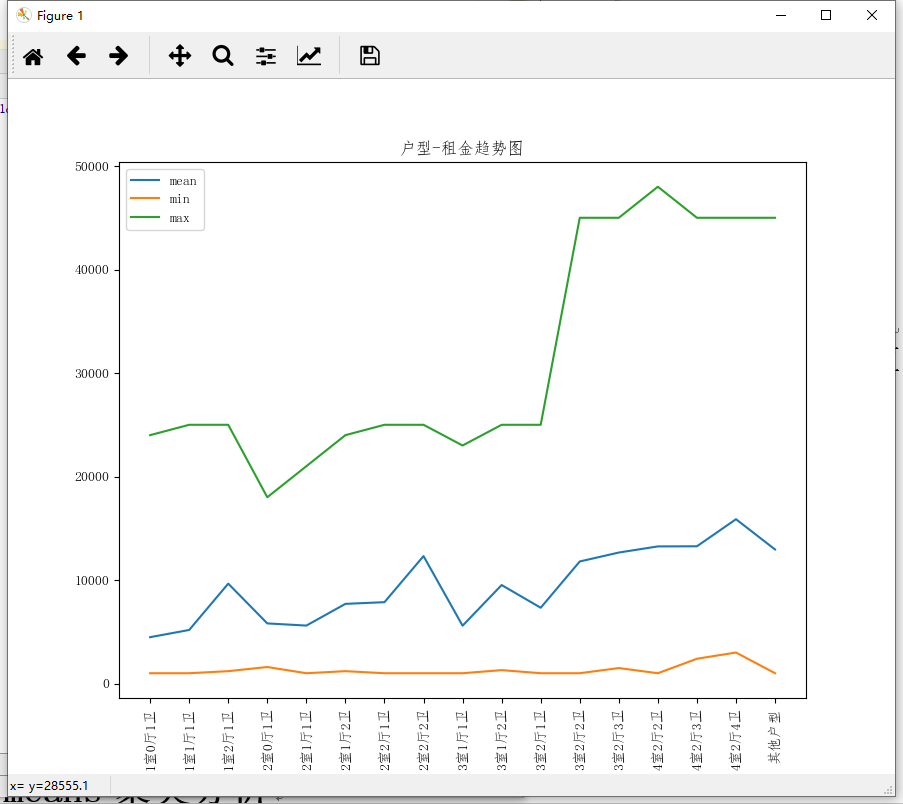
### 3.2.2区域与上架时长分析

分析上架时长时发现，有套房源([链接](https://sh.lianjia.com/zufang/SH2303655123387695104.html))的发布时间竟然为1970年，初步断定为录入错误数据，因此本文将其进行了剔除。随后对各区域的上架时长进行了分析，如下所示。



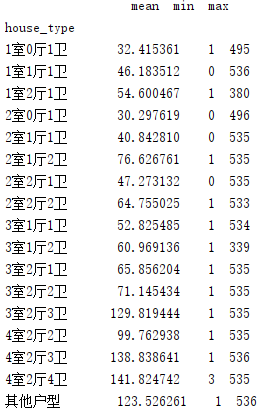
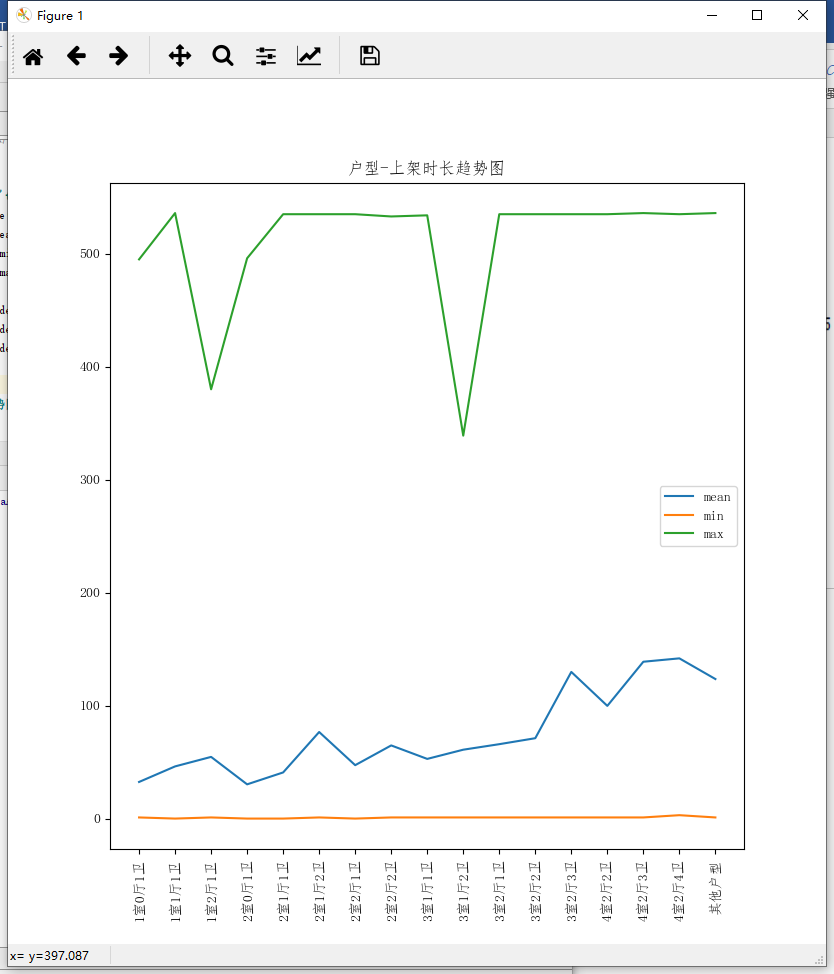
从表中可以看到，崇明这一区域的房源上架时长最久，可能是过于偏僻。其余地区的房源上架时长差异相对不明显。

### 3.2.3户型与租金分析



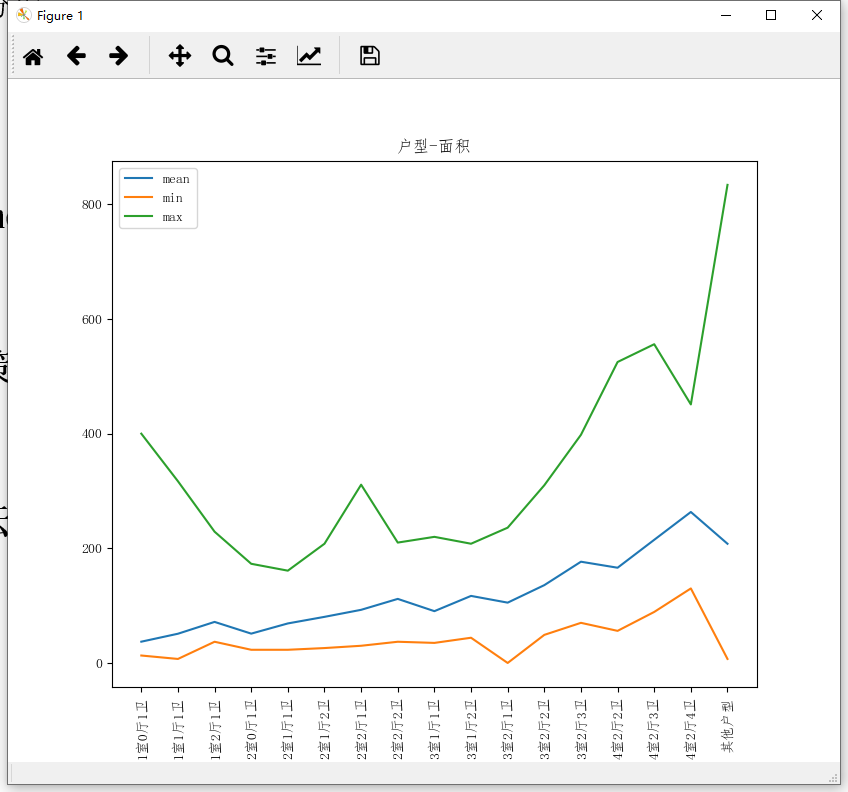
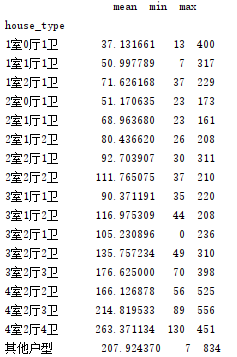
从中可以看出，随着房间数量的增加，租金有上升趋势。而常见户型（2室1厅1卫、1室1厅1卫）平均租金也超过了5000元/月。

### 3.2.4户型与上架时长分析



从中可以看到，大户型的平均上架时长相对较长，而像常见户型（2室1厅1卫、1室1厅1卫）以及其他小户型则平均上架时长较短。

### 3.2.5户型与面积分析

显然，随着房间数量的增加，面积也增加。然而对于户型与面积存在对应关系，例如2室1厅1卫的平均面积为68.96㎡。

### 3.2.6租金、面积、上架时长、最近地铁距离相关分析

对于这四种连续变量，本文首先计算了它们的相关系数矩阵，如下表所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | rent | square | time\_delta | subway |
| rent | 1.000000 | 0.649142 | 0.178392 | -0.088873 |
| square | 0.649142 | 1.000000 | 0.279384 | 0.034736 |
| time\_delta | 0.178392 | 0.279384 | 1.000000 | -0.006909 |
| subway | -0.088873 | 0.034736 | -0.006909 | 1.000000 |

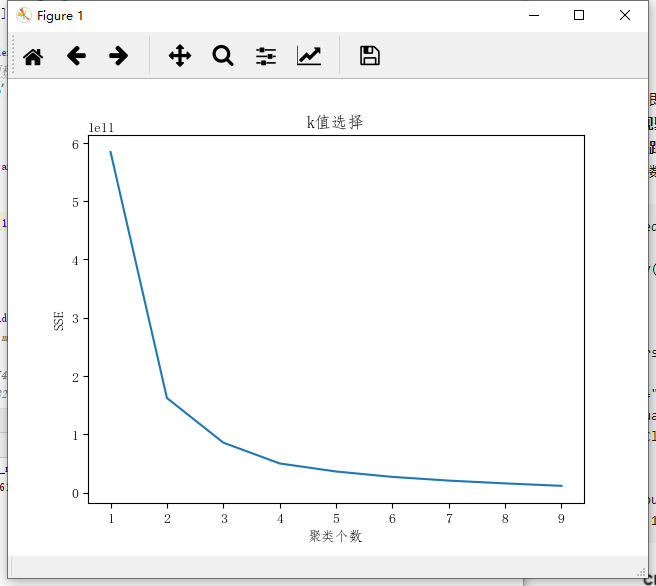
从表中可以看到，square（面积）与rent（租金）呈正相关，square（面积）也与time\_delta（上架时长）呈弱正相关，而最近地铁距离与其余三个变量几乎不相关。因此为了便于后续分析，将地铁信息这一变量进行二分类化，代表有无地铁信息。

# 4租房房源k-means聚类分析

本文选取了数值型变量租金(rent)、面积(square)、上架时间间隔(time\_delta)进行了k-means聚类分析。

## 4.1k值选择

在聚类之前需要确定**聚为几类**，即k值。根据聚类原则：组内差距要小，组间差距要大，本文绘制了不同类簇下的组内离差平方和图。



从图中可以看到，聚类个数达4个时，SSE取值接近平缓，因此k值选择为4。

## 4.2聚类结果统计

聚类后**各个类别的数量**如下表。

|  |  |  |  |
| --- | --- | --- | --- |
| 0 | 1 | 2 | 3 |
| 8528 | 3183 | 1664 | 5421 |

**聚类中心**的结果如下表所示。

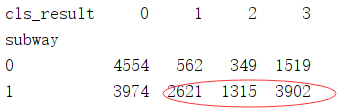
|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 租金 | 面积 | 上架时长 |
| 0 | 3918.60096154 | 72.89774859 | 54.47678236 |
| 1 | 13626.01916431 | 132.40182218 | 69.71190701 |
| 2 | 21248.48617788 | 179.64963942 | 97.20552885 |
| 3 | 7756.37114923 | 96.8544549 | 53.36081904 |

结合**前文3.2.5户型与面积**的关系推测，这四个中心可以分别代表1室2厅1卫、3室2厅2卫、3室2厅3卫、2室2厅1卫这4种户型。

**区域与聚类结果**数量统计



**有无地铁**（0表示无地铁信息）与聚类结果数量统计



## 4.3聚类结果解释

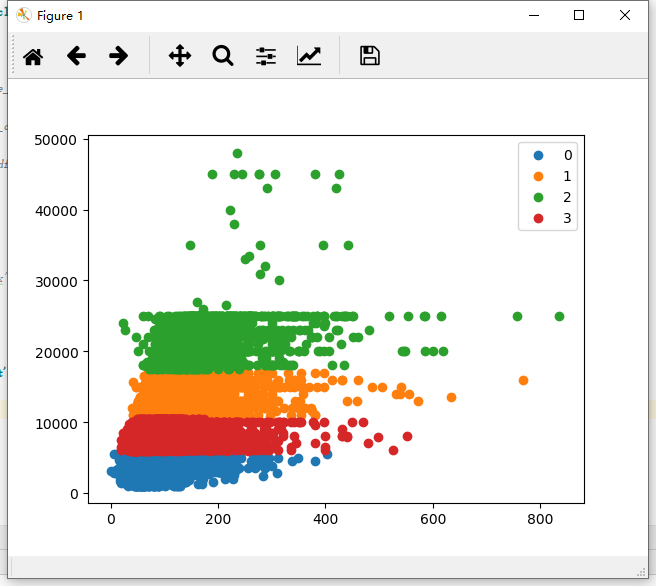
整体来看，可以将聚类结果合成如下：

**大户型**（3室2厅2卫、3室2厅3卫），属于第1、2类，平均面积超过130㎡，平均租金也较高，且上架时长较久，分布于浦东、长宁等地。

**经济型**（1室2厅1卫），属于第0类，平均租金较低，且周边大多无地铁，分布于浦东、松江等地。

**中间型**（2室2厅1卫），属于第3类，平均租金介于大户型与经济型之间，房源较多，平均上架时间相对较短，分布于浦东、闵行等地。

最后绘制聚类结果同**租金与面积**散点图如下



# 5租房租金决策树回归分析

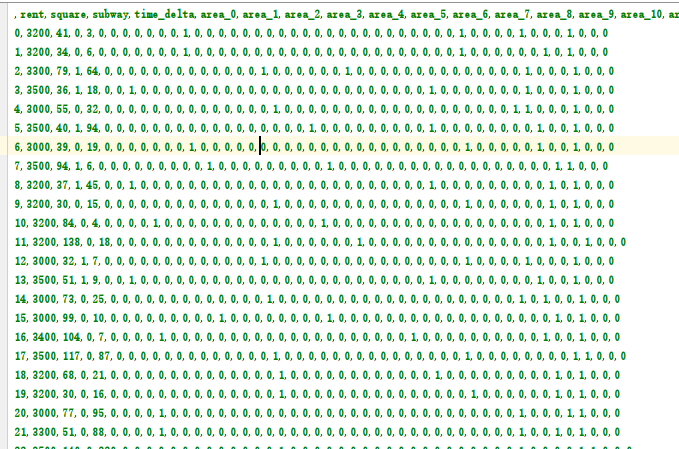
对于租金价格预测，本文选择了（面积、地铁、区域、户型、楼层、聚类水平）这些变量作为特征，构建了决策树回归模型。

## 5.1预处理编码

首先需要对非数值变量（区域、楼层、户型）进行编码，编码字典如下表所示。

|  |
| --- |
| {'静安': 0, '黄浦': 1, '闸北': 2, '嘉定': 3, '宝山': 4, '上海周边': 5, '普陀': 6, '徐汇': 7, '长宁': 8, '松江': 9, '崇明': 10, '奉贤': 11, '杨浦': 12, '浦东': 13, '闵行': 14, '青浦': 15, '金山': 16, '虹口': 17} |
| {'中楼层': 0, '未知': 1, '高楼层': 2, '低楼层': 3} |
| {'3室2厅1卫': 0, '3室1厅1卫': 1, '2室2厅1卫': 2, '3室1厅2卫': 3, '2室1厅2卫': 4, '其他户型': 5, '4室2厅3卫': 6, '2室2厅2卫': 7, '3室2厅3卫': 8, '1室1厅1卫': 9, '4室2厅4卫': 10, '3室2厅2卫': 11, '1室0厅1卫': 12, '2室0厅1卫': 13, '1室2厅1卫': 14, '4室2厅2卫': 15, '2室1厅1卫': 16} |

之后对上述变量采取one-hot编码措施，作为训练模型输入。部分模型输入数据如下图所示。



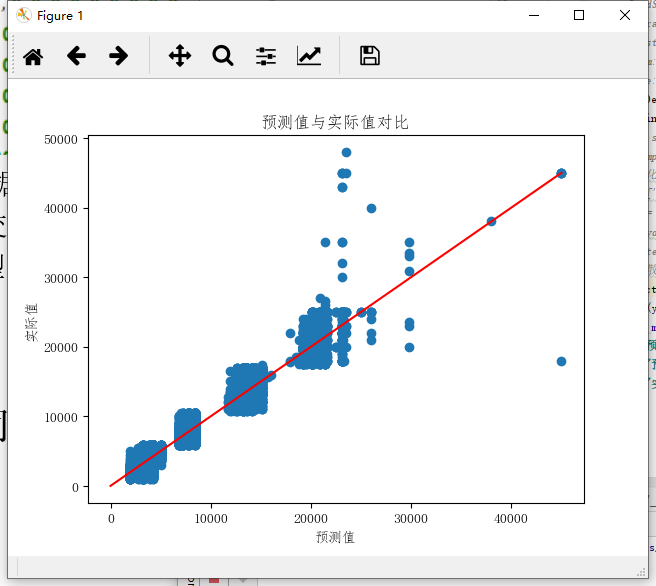
## 5.2模型训练

将模型输入数据按0.8:0.2划分训练集与测试集。决策树模型树的最大深度确定，使用网格搜索法，5折交叉验证，最终确定树的最大深度为6，训练集测试得分为0.923，测试集得分为0.919，模型效果较好。

## 5.3模型结果分析

决策树可视化结果保存于tree.png（图片过大，文档不宜展示）中。对图片中的结果观察可知，决策树共127个结点，其中聚类水平变量、面积变量位于树的上层，对预测结果影响较大，而楼层变量几乎不影响预测结果。根据决策树可视化结果可以得出，租金主要与面积有关，与区域、有无地铁、户型也有一定关系，而与楼层几乎没有关系。

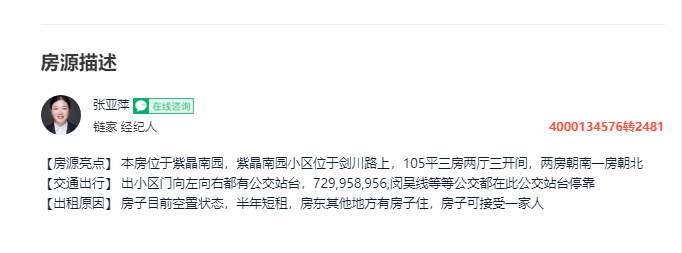
最后绘制模型预测值与实际值得散点图。



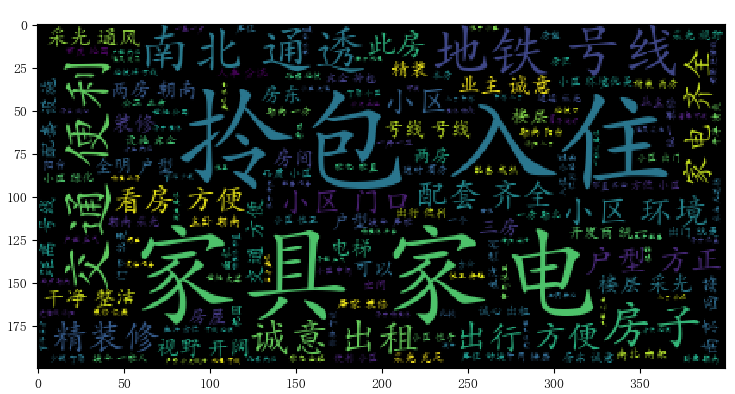
可以看到，模型的输出拟合效果也较好，模型可以初步用于上海租房房源市场租金预测。

# 6租房介绍词云分析

链家网上的租房介绍（房源描述）如下图所示。



本文对上述文本信息进行了词云分析。首先对房源亮点、装修描述、出租原因、交通出行、户型介绍这些无效频繁词进行了剔除，并且对文本中的标点、空格等符号也进行了剔除，之后利用jieba对文本进行了分词，最后生成词云，如下图所示。



从图中可以看到，家具、家电、拎包入住成为词云的亮点高频词，可以推测租房房源市场可能对快速入住有较大需求，因而在租房介绍上重点描述了这方面的信息。

# 7总结

本文基于上述分析，初步得出上海租房市场如下结论：

1. 房源集中分布于浦东、闵行、松江等地；
2. 房源户型、楼层种类很多，能够满足市场多样化需求；
3. 经济型房源大多无周边地铁信息；
4. 上架时长久的房源普遍为大户型房源；
5. 租金主要与面积有关，与区域、有无地铁、户型也有一定关系，而与楼层几乎没有关系；
6. 租房市场对快速入住有较高的需求。