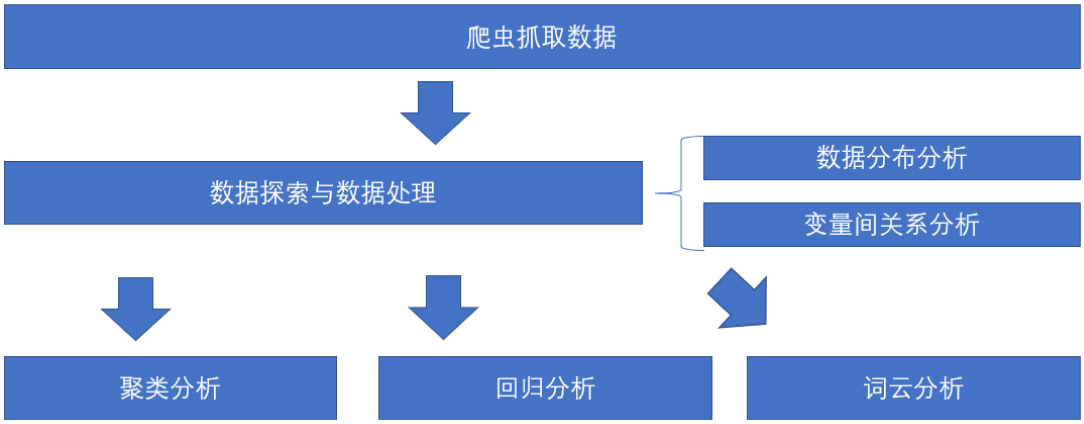
上海链家网租房数据分析报告

1 分析简介

上海链家网租房数据分析框架如下图所示

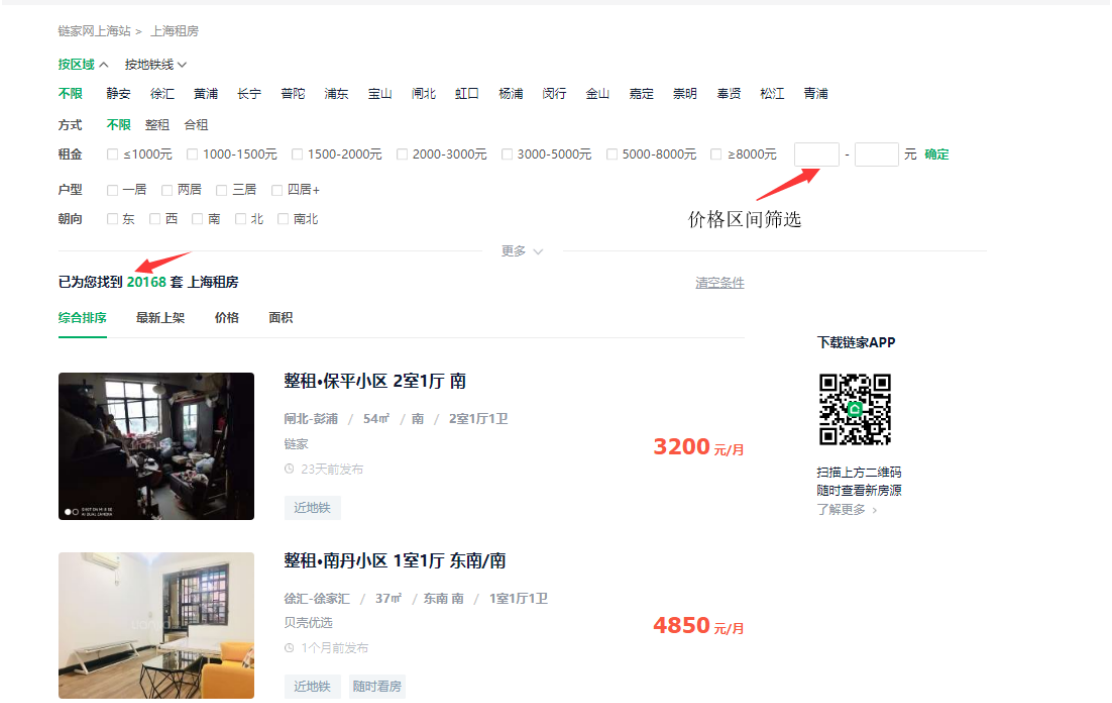


本文首先使用爬虫程序抓取房源数据，之后对数据进行了探索与处理，最后对进行了聚类分析、回归分析以及词云分析。

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）。

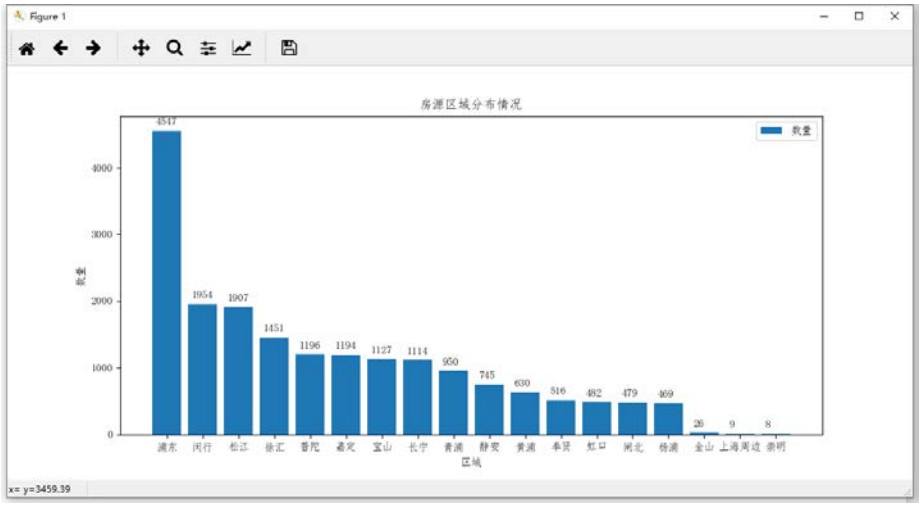
最终爬取 18805 条房源数据，如下图所示，大致接近所有租房房源数据，可以进行整体分析。爬虫代码见 lianjiaSpider 文件夹。

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

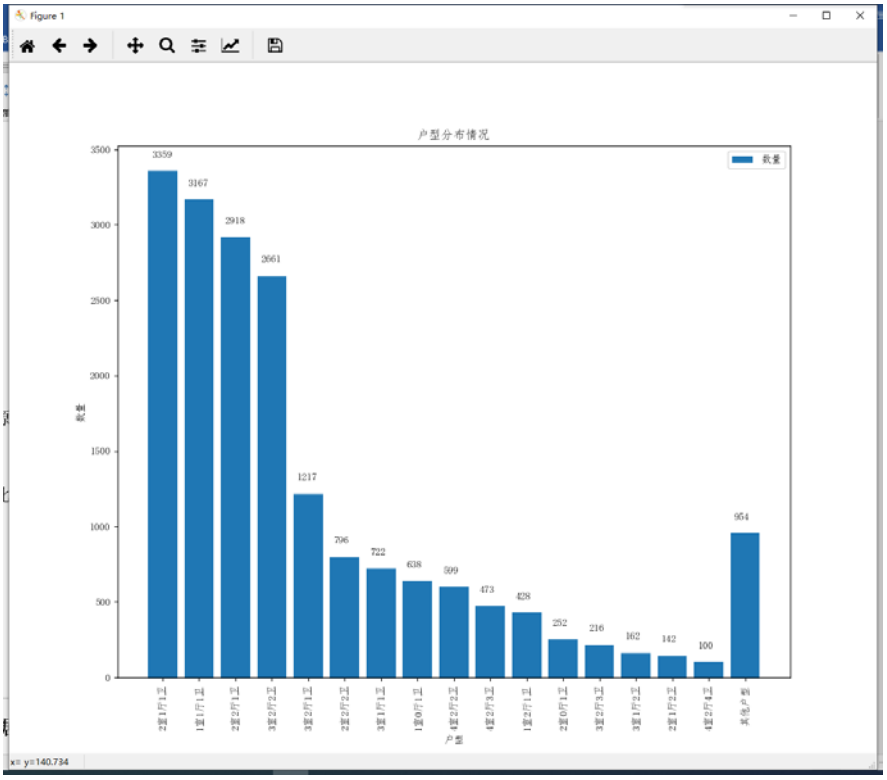
3.1.1 房源区域分布



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

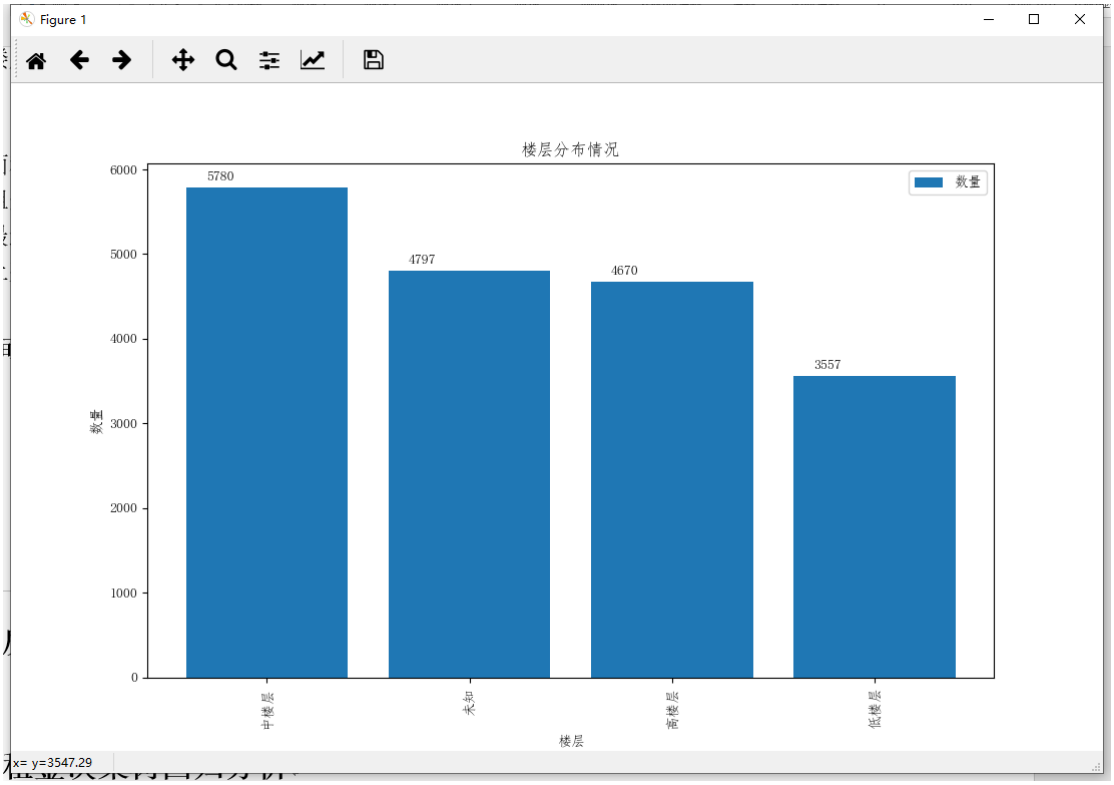
统计发现户型的种类多达 122 种，而大部分户型市场上的数量很少甚至只有 1 个，因此考虑将户型数量小于 100 的归为其他户型，之后绘制分布图如下。



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

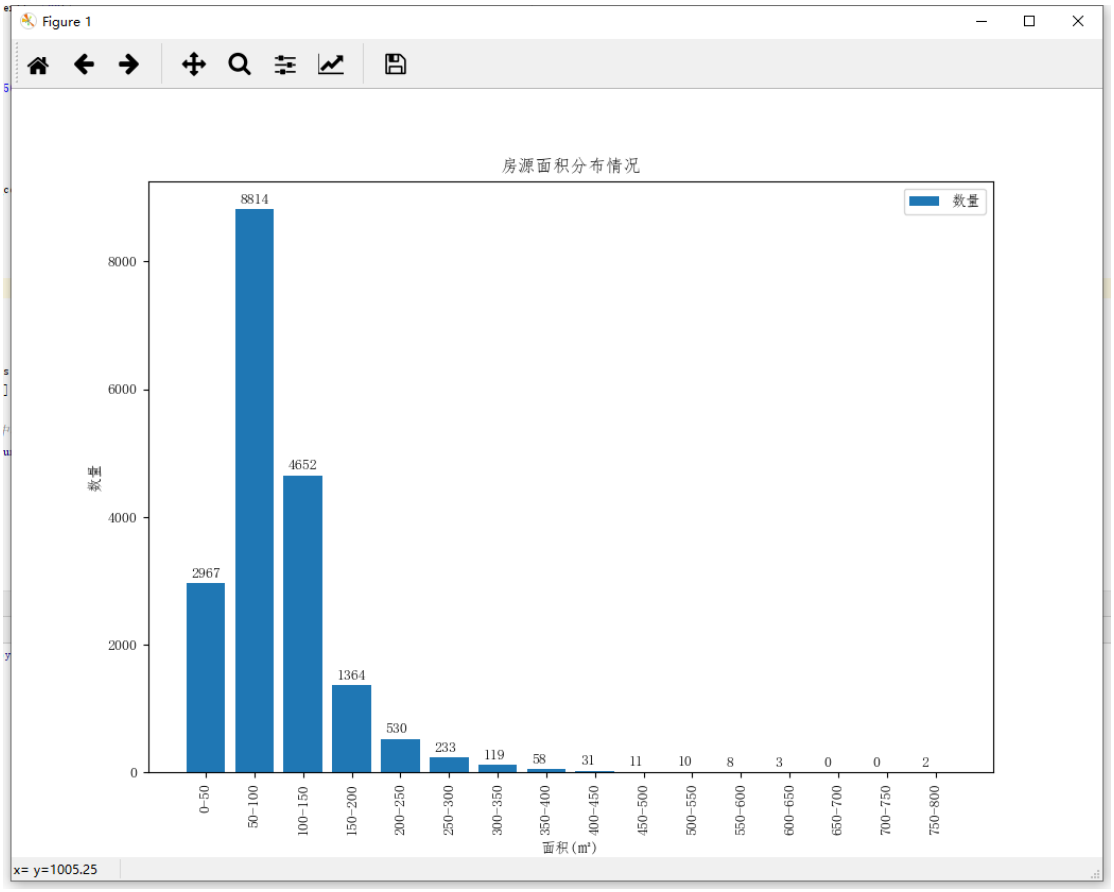
统计发现市场上的楼层描述为楼层类型/总楼层，因而楼层的水平多达 171 种，为了便于分析，统一将楼层划分为低楼层、中楼层、高楼层三类，剩余的归为未知楼层，分布如下图所示。



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

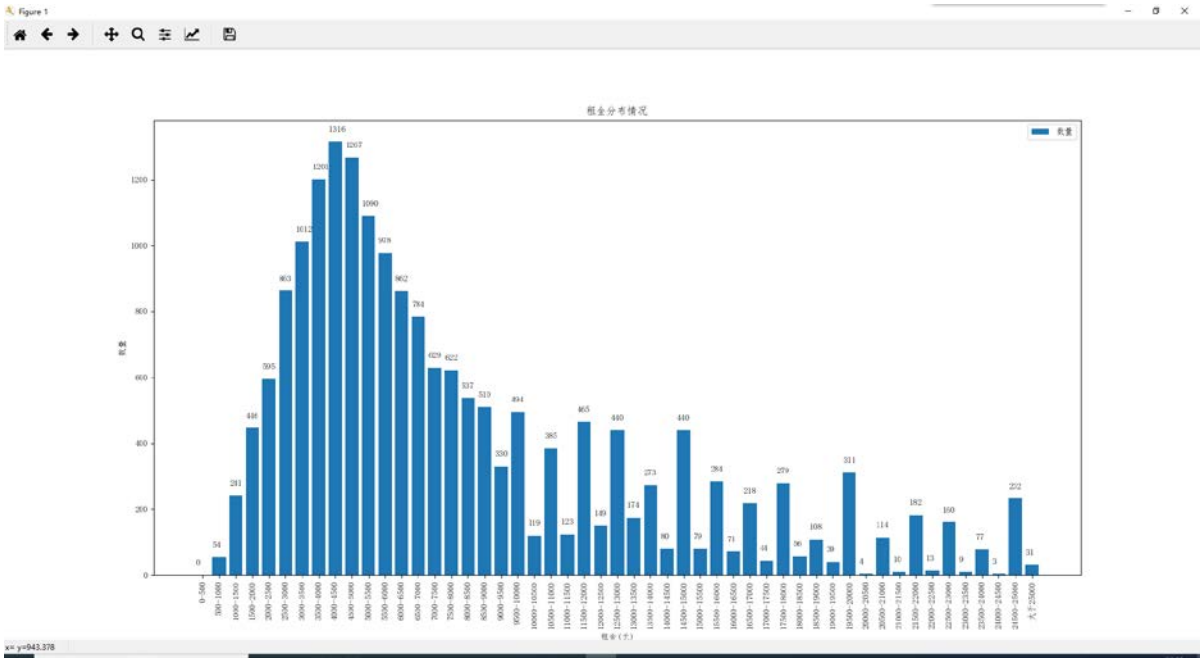
面积数据属于连续数据，需要进行离散化处理才能绘制分布图，本文将面积数据按照一定的区间进行划分，之后绘制分布图，如下所示。



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

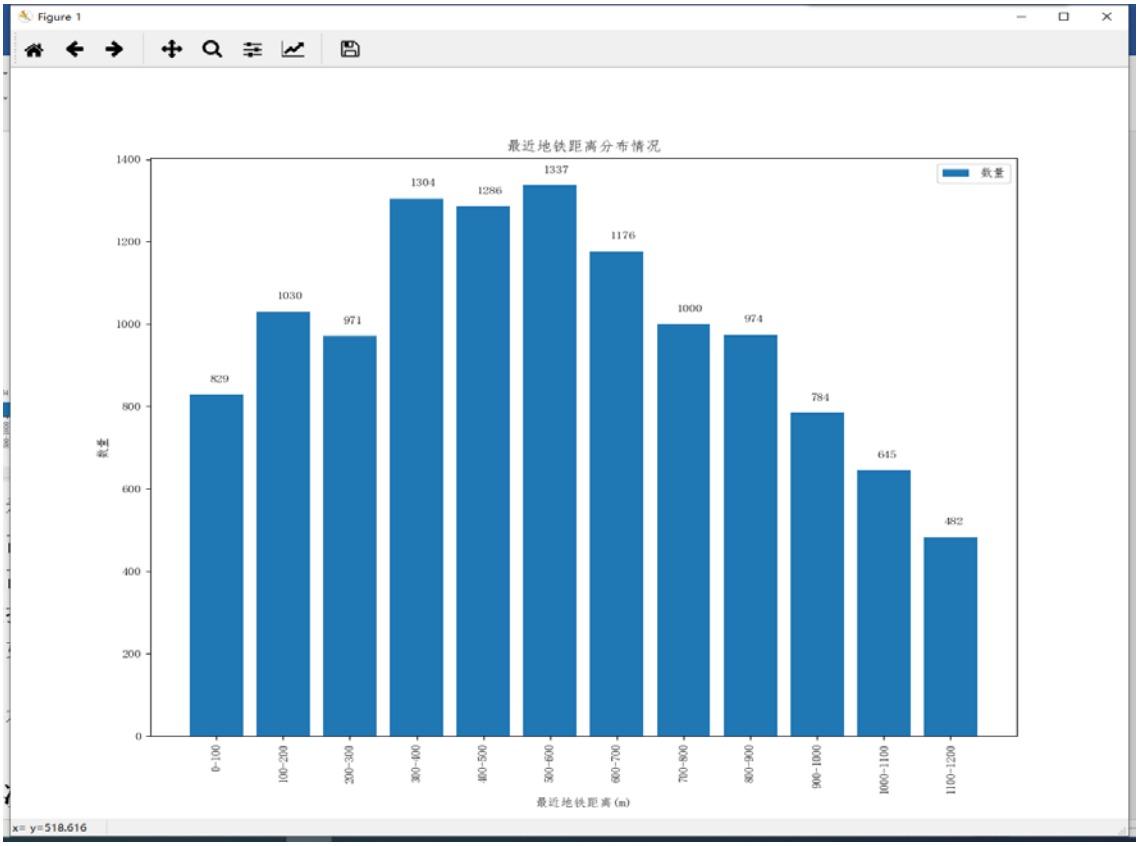
租金数据也为连续数据，因此也需要离散化处理，分布图如下。



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

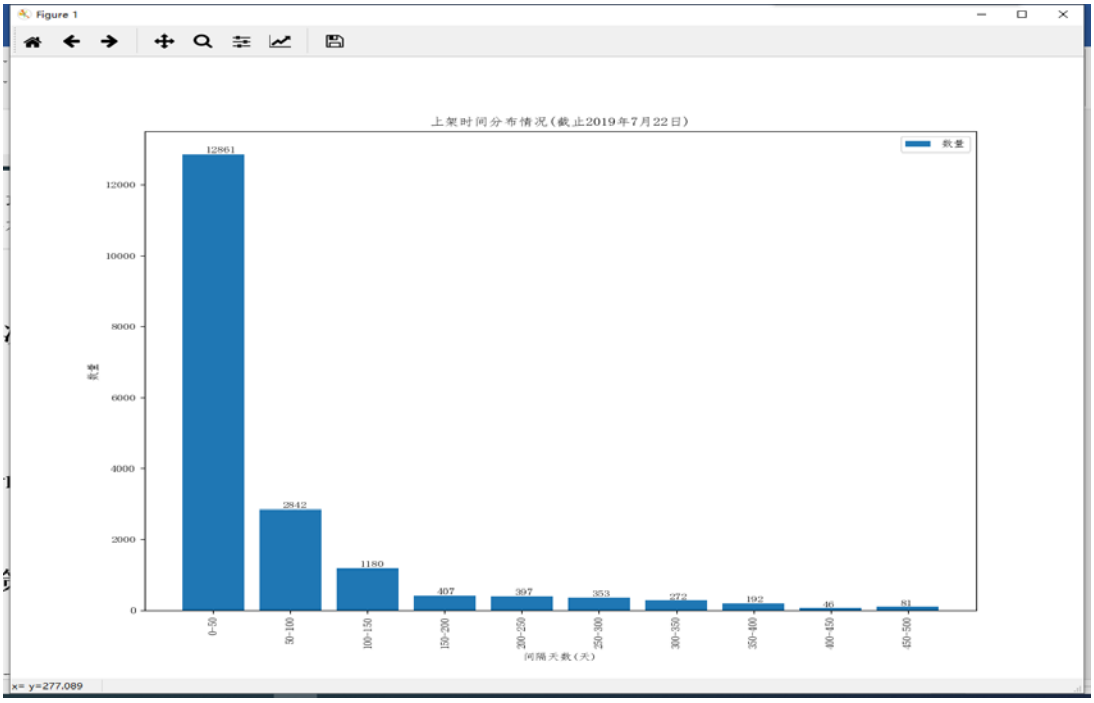
部分房源数据没有地铁相应数据，因此本文只对有地铁数据的房源进行了最近地铁距离分布分析。本文首先对抓取得地铁数据进行了最近距离提取，随后绘制了分布图如下。



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。

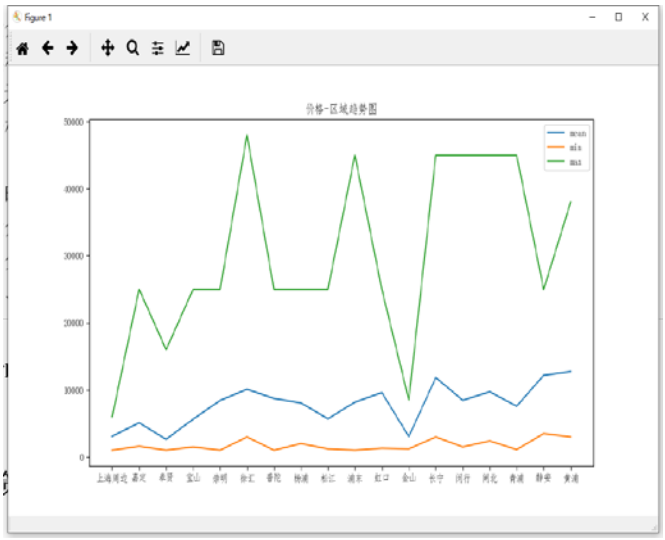


可以看到，上架时长呈分布呈递减趋势，时间越短数量越多。

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)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。

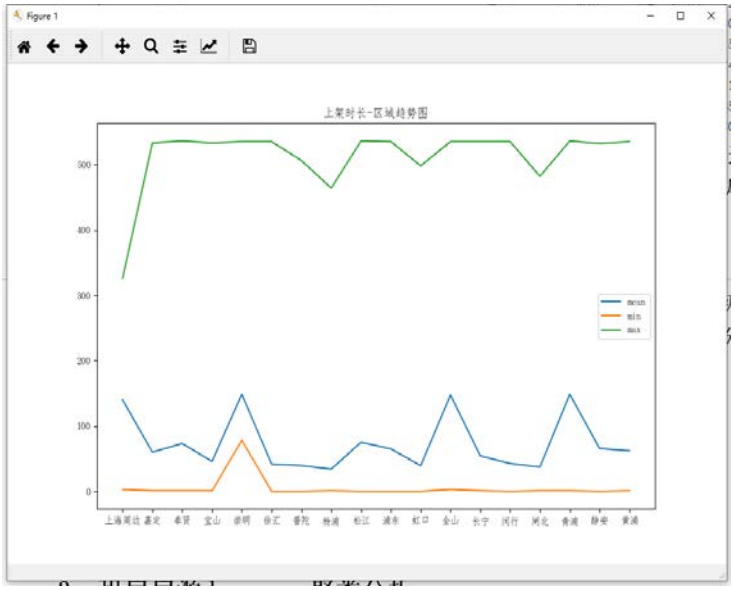


area	mean	min	max
上海周边	3100.000000	1000	6000
嘉定	5096.221943	1600	25000
奉贤	2640.821705	1000	16000
宝山	5634.454303	1500	25000
崇明	8437.500000	1000	25000
徐汇	10106.048276	3000	48000
普陀	8733.326087	1000	25000
杨浦	8074.520256	2000	25000
松江	5692.579444	1200	25000
浦东	8181.848558	1000	45000
虹口	9615.726141	1300	25000
金山	3042.307692	1200	8500
长宁	11848.360288	3000	45000
闵行	8461.511270	1500	45000
闸北	9730.688935	2400	45000
青浦	7586.246316	1100	45000
静安	12188.818792	3500	25000
黄浦	12734.904762	3000	38000

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

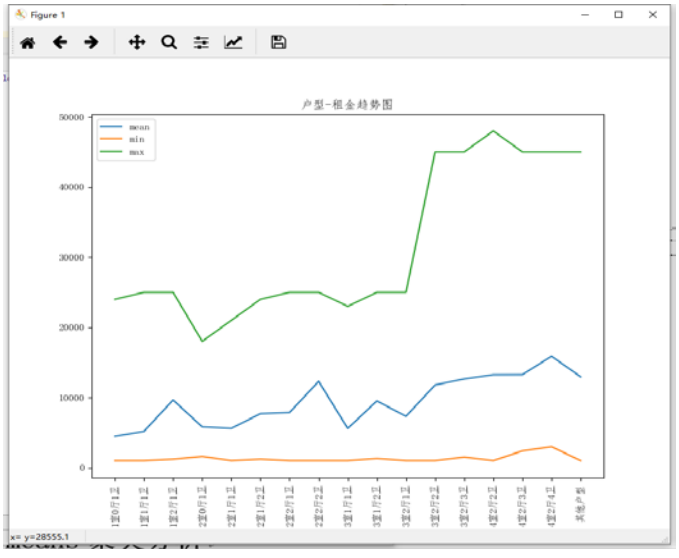
分析上架时长时发现，有套房源(链接)的发布时间竟然为 1970 年，初步断定为录入错误数据，因此本文将其进行了剔除。随后对各区域的上架时长进行了分析，如下所示。



area	mean	min	max
上海周边	140.000000	3	326
嘉定	59.868509	1	533
奉贤	73.250000	1	536
宝山	45.678793	1	533
崇明	148.500000	78	535
徐汇	41.564828	0	535
普陀	39.378763	0	506
杨浦	33.944563	1	464
松江	75.194546	0	536
浦东	64.903368	0	535
虹口	39.257261	0	498
金山	147.384615	3	535
长宁	54.113208	1	535
闵行	42.531762	0	535
闸北	37.123173	1	482
青浦	148.509474	1	536
静安	65.590604	0	532
黄浦	61.919048	1	535

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。

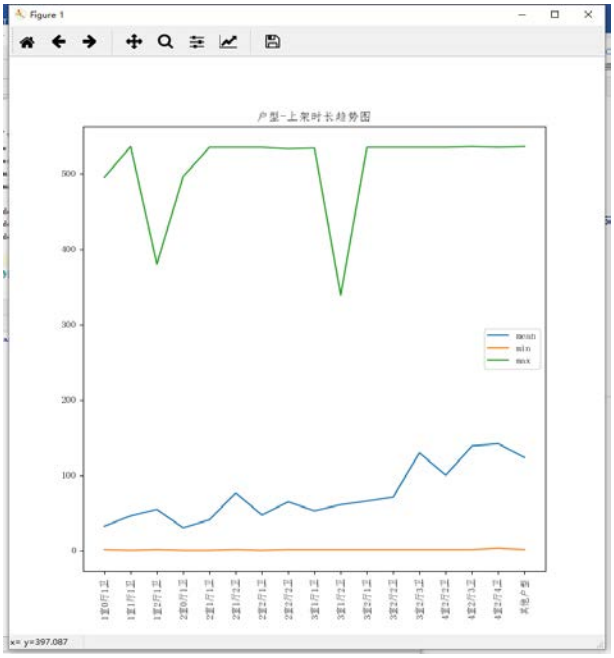
3.2.3 户型与租金分析



house_type	mean	min	max
1室0厅1卫	4483.164577	1000	24000
1室1厅1卫	5178.858812	1000	25000
1室2厅1卫	9655.808411	1200	25000
2室0厅1卫	5813.015873	1600	18000
2室1厅1卫	5602.377493	1000	21000
2室1厅2卫	7700.704225	1200	24000
2室2厅1卫	7864.013023	1000	25000
2室2厅2卫	12316.884422	1000	25000
3室1厅1卫	5590.872576	1000	23000
3室1厅2卫	9520.370370	1300	25000
3室2厅1卫	7325.002465	1000	25000
3室2厅2卫	11796.781661	1000	45000
3室2厅3卫	12659.203704	1500	45000
4室2厅2卫	13250.964942	1000	48000
4室2厅3卫	13263.057325	2400	45000
4室2厅4卫	15879.371134	3000	45000
其他户型	12955.063025	1000	45000

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。

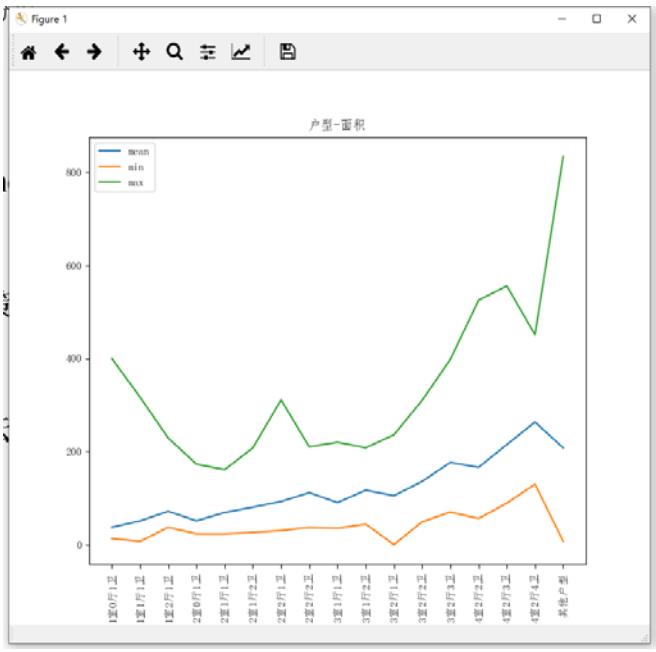
3.2.4 户型与上架时长分析



house_type	mean	min	max
1室0厅1卫	32.415361	1	495
1室1厅1卫	46.183512	0	536
1室2厅1卫	54.600467	1	380
2室0厅1卫	30.297619	0	496
2室1厅1卫	40.842810	0	535
2室1厅2卫	76.626761	1	535
2室2厅1卫	47.273132	0	535
2室2厅2卫	64.755025	1	533
3室1厅1卫	52.825485	1	534
3室1厅2卫	60.969136	1	339
3室2厅1卫	65.856204	1	535
3室2厅2卫	71.145434	1	535
3室2厅3卫	129.819444	1	535
4室2厅2卫	99.762938	1	535
4室2厅3卫	138.838641	1	536
4室2厅4卫	141.824742	3	535
其他户型	123.526261	1	536

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



house_type	mean	min	max
1室0厅1卫	37.131661	13	400
1室1厅1卫	50.997789	7	317
1室2厅1卫	71.626168	37	229
2室0厅1卫	51.170635	23	173
2室1厅1卫	68.963680	23	161
2室1厅2卫	80.436620	26	208
2室2厅1卫	92.703907	30	311
2室2厅2卫	111.765075	37	210
3室1厅1卫	90.371191	35	220
3室1厅2卫	116.975309	44	208
3室2厅1卫	105.230896	0	236
3室2厅2卫	135.757234	49	310
3室2厅3卫	176.625000	70	398
4室2厅2卫	166.126878	56	525
4室2厅3卫	214.819533	89	556
4室2厅4卫	263.371134	130	451
其他户型	207.924370	7	834

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

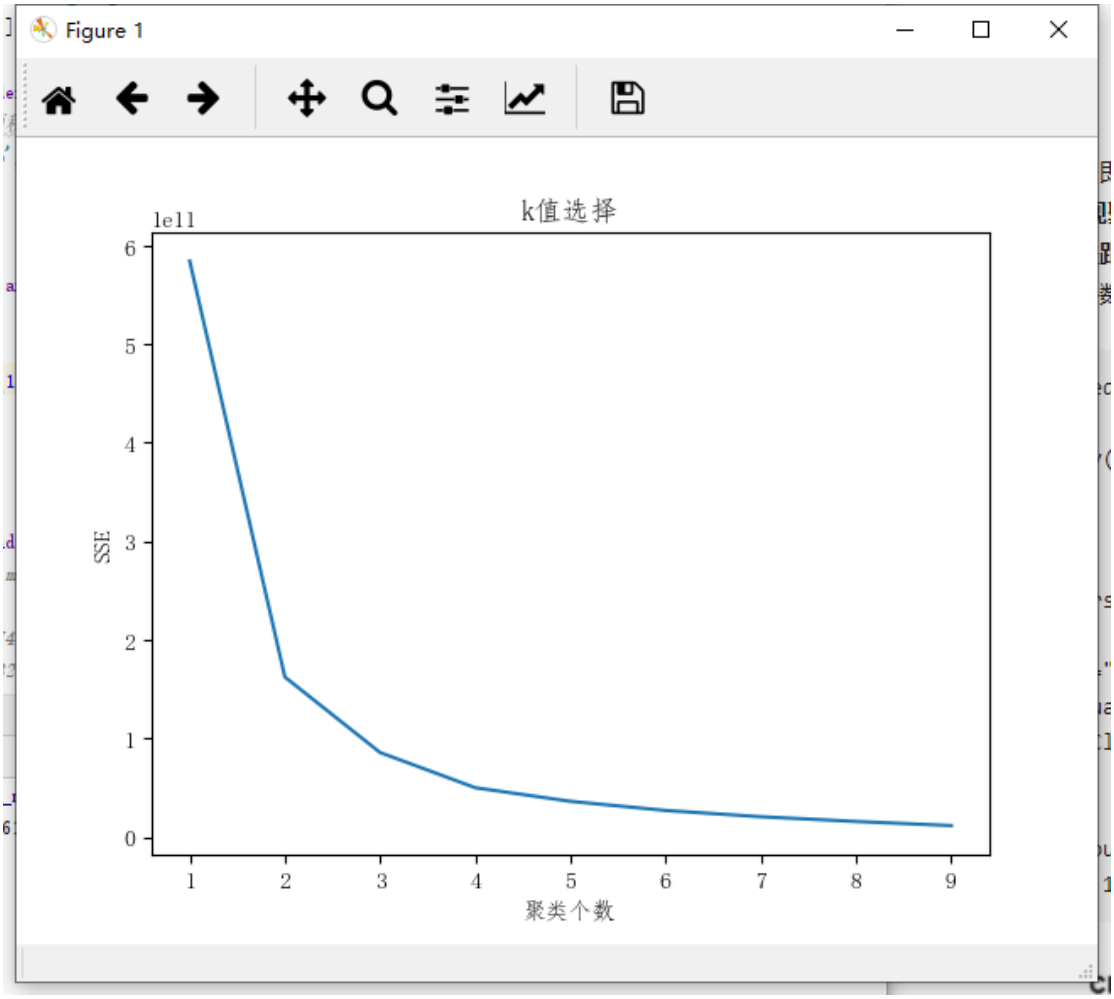
从表中可以看到，square（面积）与 rent（租金）呈正相关，square（面积）也与 time_delta（上架时长）呈弱正相关，而最近地铁距离与其余三个变量几乎不相关。因此为了便于后续分析，将地铁信息这一变量进行二分类化，代表有无地铁信息。

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

cls_result	0	1	2	3
area				
上海周边	8	0	0	1
嘉定	922	38	21	213
奉贤	497	5	0	14
宝山	720	48	9	350
崇明	5	0	2	1
徐汇	420	388	178	464
普陀	423	274	77	422
杨浦	192	73	30	174
松江	1308	106	54	439
浦东	2035	822	404	1282
虹口	161	103	59	159
金山	23	0	0	3
长宁	188	348	218	359
闵行	833	297	186	636
闸北	143	108	53	175
青浦	511	98	87	254
静安	69	256	147	273
黄浦	70	219	139	202

有无地铁（0 表示无地铁信息）与聚类结果数量统计

cls_result	0	1	2	3
subway				
0	4554	562	349	1519
1	3974	2621	1315	3902

4.3 聚类结果解释

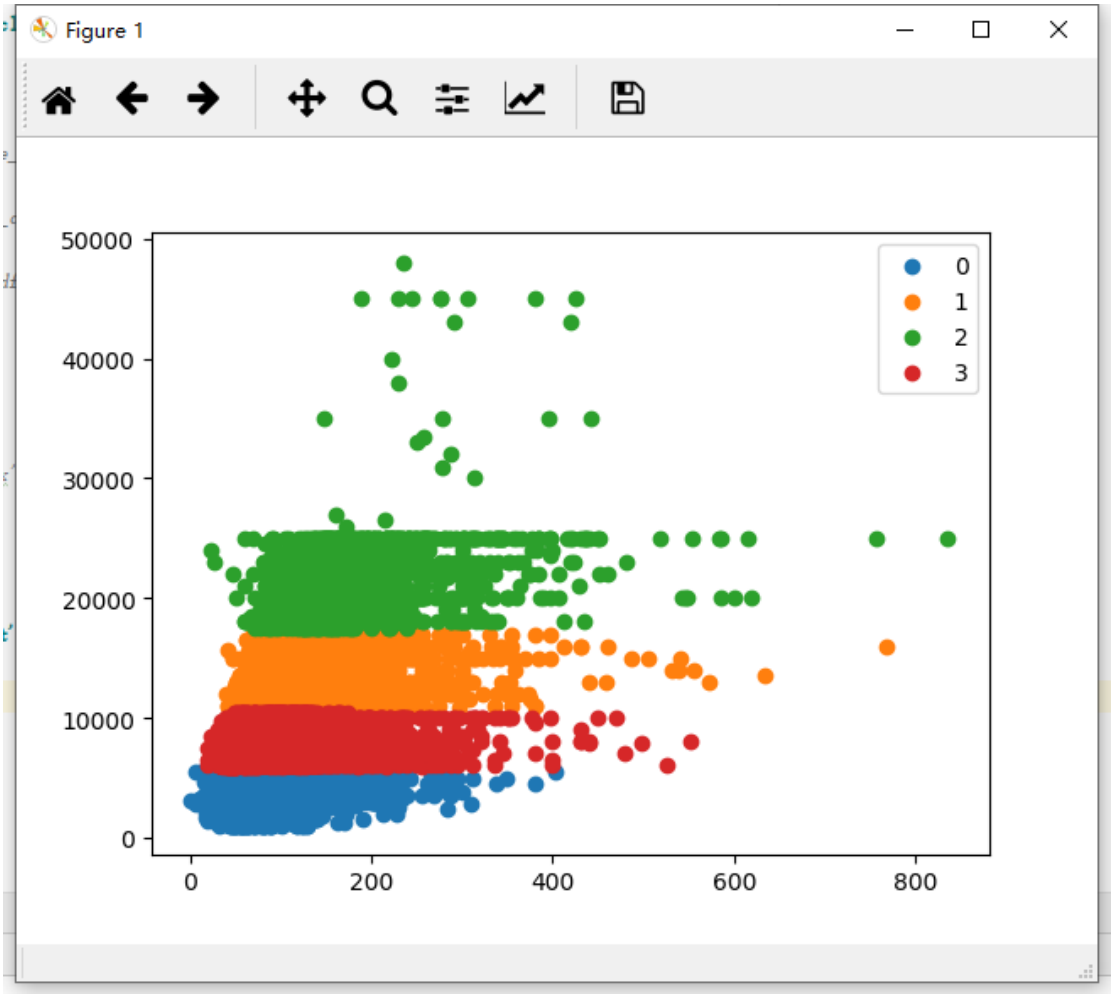
整体来看，可以将聚类结果合成如下：

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

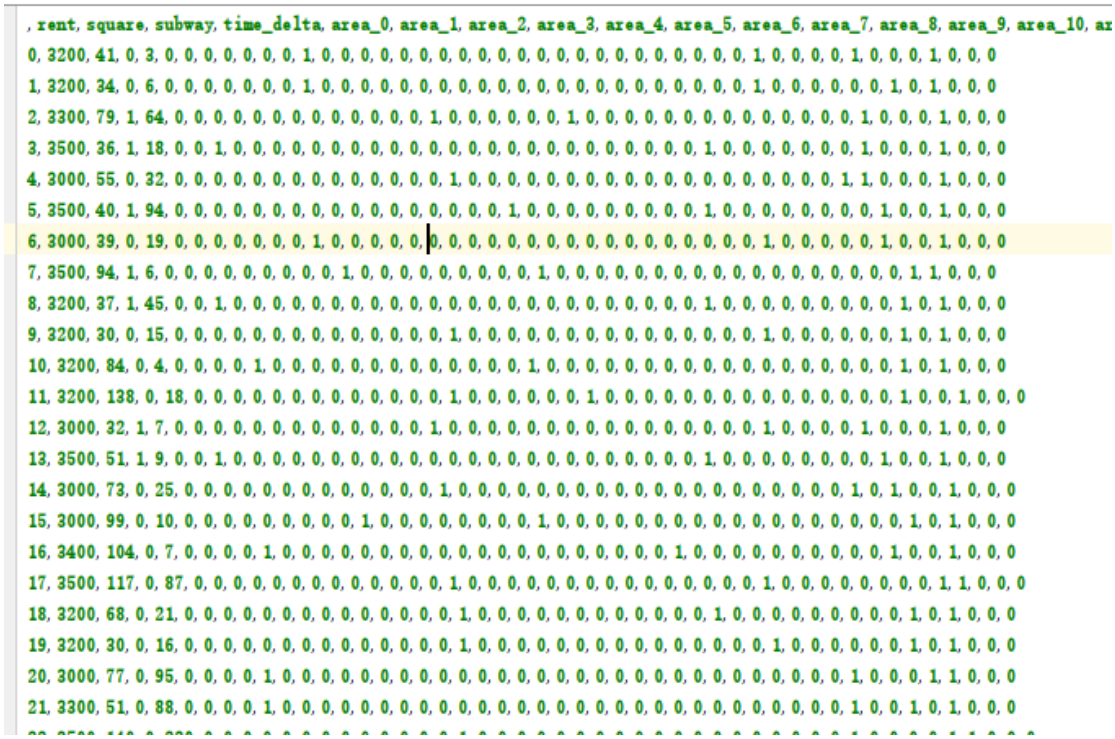
对于租金价格预测，本文选择了（面积、地铁、区域、户型、楼层、聚类水平）这些变量作为特征，构建了决策树回归模型。

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。



,rent,square,subway,time_delta,area_0,area_1,area_2,area_3,area_4,area_5,area_6,area_7,area_8,area_9,area_10,ar
0,3200,41,0,3,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0
1,3200,34,0,6,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,1,0,0,0
2,3300,79,1,64,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,0,0,0
3,3500,36,1,18,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0
4,3000,55,0,32,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,1,0,0,0
5,3500,40,1,94,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0
6,3000,39,0,19,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,1,0,0,0
7,3500,94,1,6,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0
8,3200,37,1,45,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,1,0,0,0
9,3200,30,0,15,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,1,0,0,0
10,3200,84,0,4,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0
11,3200,138,0,18,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0
12,3000,32,1,7,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0
13,3500,51,1,9,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,1,0,0,0
14,3000,73,0,25,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,1,0,0,0
15,3000,99,0,10,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,0,0
16,3400,104,0,7,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,1,0,0,0
17,3500,117,0,87,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1,0,0,0
18,3200,68,0,21,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,1,0,0,0
19,3200,30,0,16,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,1,0,0,0
20,3000,77,0,95,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,1,1,0,0,0
21,3300,51,0,88,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,1,0,0,0

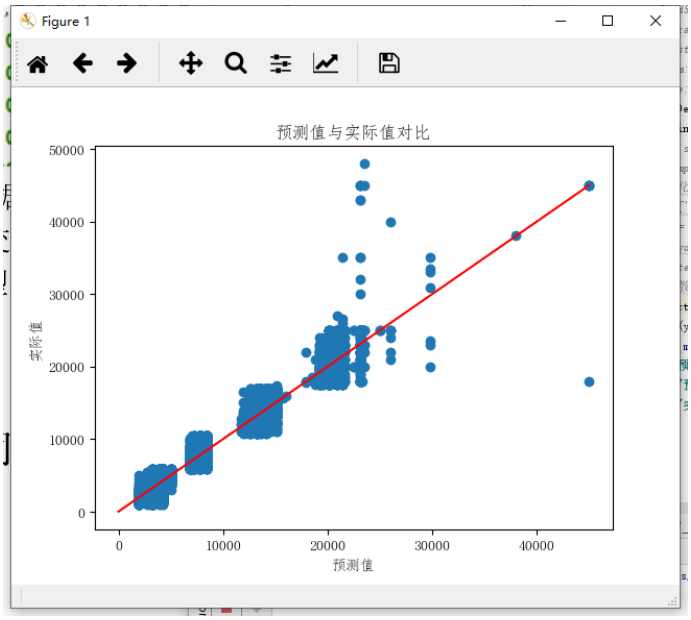
5.2 模型训练

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。

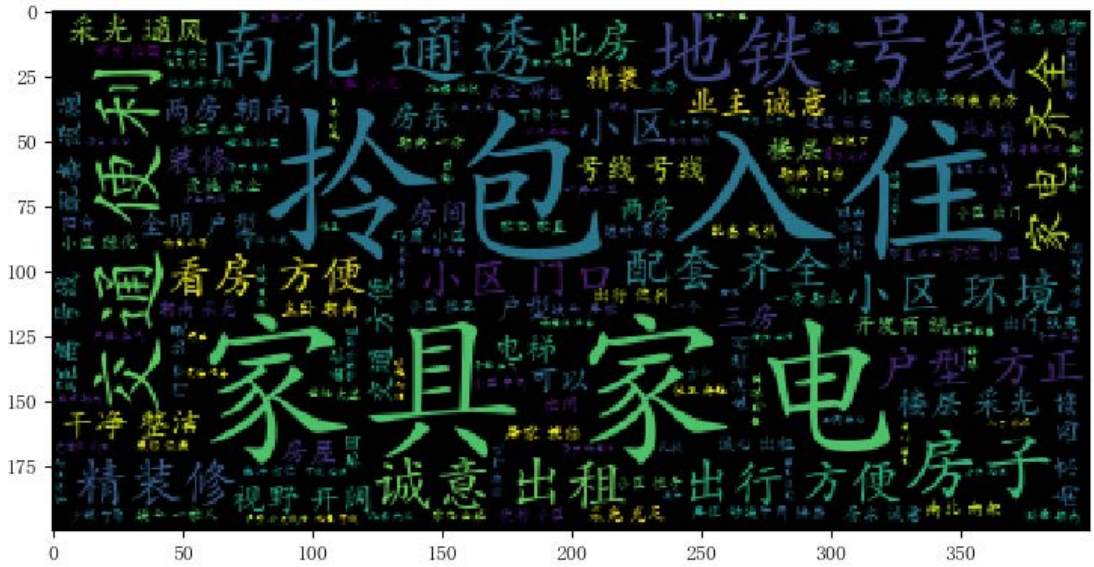


可以看到，模型的输出拟合效果也较好，模型可以初步用于上海租房房源市场租金预测。

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。