摘　要

近几年来，北京的房价涨势迅猛，同时也带动了租房价格的不断上涨。本文以北京市为例，采集自如网的租房数据，累计 76319 条房源数据。通过对数据的统计分析和可视化，揭示了自如的房源的空间地理分布情况，各个行政区域合租和整租的价格分布。最后，本文根据六个维度综合比较各个行政区域房源的性价比。分析结果可为租房者在选房时提供宏观上的参考。

关键字：　租房；数据采集；数据可视化；统计分析

Abstract

In recent years, Beijing's housing prices have been rising rapidly, and at the same time, it has also led to the rising rental prices. Taking Beijing as an example, this paper collects the rental data from Free Net, totaling 76319 housing source data. Through statistical analysis and visualization of the data, this paper reveals the spatial and geographical distribution of comfortable housing resources, and the price distribution of combined rent and integrated rent in each administrative region. Finally, according to six dimensions, we comprehensively compares the performance-price ratio of housing in each administrative region. The analysis results can provide a macro reference for renters when choosing a house.

Key Words: Rental; Data Acquisition; Data Visualization; Statistical Analysis

目录

第一章 绪论

(一)选题背景及意义

近几年来，北京的房价涨势迅猛，同时也带动了租房价格的不断上涨。租房对于绝大多数北漂都是必须要面对的一个问题，每月过高的租房支出，增加了北漂一族的压力，同时也在降低他们的生活质量。在北京，哪里的租房质量好、价格比较便宜、离地铁近、离公司近、上下班方便等方面都是人们在选择租房时的一些重要考察指标。通常人们都是通过自己的主观综合分析来判断房间是否合适，价格是否合理，自己是否能够承受，然后做出一个选择。但是最后真正租到一个满意的房间的人是很少的。通过分析其中原因发现以下几个原因：1、缺乏市场调研，对市场的租房价格没有一个整体的了解；2、市场不透明，不规范，导致中介泛滥，人们容易被中介误导，做出不合理的判断；3、人们的精力有限，在有限时间内能够参考的房间太少，选择过于仓促。这些因素综合起来，导致人们在选择租房过程中的犹豫和痛苦。

按照经验，一般距离地铁越近，租房价格也会越高；距离北京市的热点区域(如国贸)越近, 租房价格也会越高，每年 6，7 月份，大学生毕业季找房时，房价也会上涨很多。影响租房价格的因素有很多，比如区域、地铁（线路，距离）、房屋年龄、居室、面积、朝向、是否供暖、是否合租、时间等。本文以北京为例，通过对国内知名的租房中介自如网的租房数据进行采集和整理，统计分析房源的分布规律，各个区域的价格分布情况，并综合六个维度评价各个区域的房源特点。最后，本文的分析结果可为租房者在选房时提供宏观上的参考。

第二章 数据采集和预处理

(一)数据采集

自如网是链家旗下的租房网站。自如网的房源数据规整，便于前期的数据爬取，除噪，去重，量化统一，图形化展示。本文采用 Scrapy 开源爬虫框架，采集了自如网的友家合租和自如整租的房源数据。

图 2.1 自如网截图

通过持续一个周的采集，对数据去重后的得到 92927 条租房数据，将采集到的数据存入到 mongo 数据库中，然后对数据进行清洗，去重，去燥，提取出有效的数据。

(二)数据预处理

由于信息抽取不全或错误，导致采集到数据中存在一部分异常值，主要包括数据缺失，数据错误，数据重复等。数据缺失包括缺少重要的地铁信息，地理位置信息。数据错误包括租房价格异常，房屋大小异常，超出取值范围等问题，本文根据具体的问题采取不同的措施：

1、 数据缺失：采集到的数据中不包含房源的所处位置的经纬度，利用百度地图 API 获取其经纬度。

2、 数据错误或非预期，如采集的每个房源的房价都应该是按月来收费，但是会出现某些按天收费的异常数据，因此采取剔除或者乘以一个月的天数，算出房价来避免数据的浪费。

3、 数据重复，数据采集出现重复采集的情况，因此需要在采集完毕之后，根据采集到的房屋的 ID 进行去重。

4、 数据格式化，统一对数据保留两位小数。

通过对数据的预处理，最终得到 76319 条可用的房源数据。

	id	出租类型	房间大小(m2)	区域	朝向	户型	经度	纬度	临近地铁	离地铁最近距离(米)	所在楼层	出租价格(元)
0	BJZRGY081769284_01	合	10.00	海淀	南	3室1厅	40.024767	116.289076	16号线	347	1/11层	2660
1	BJZRGZ111784464	整	54.97	朝阳	南北	2室1厅	39.965058	116.469961	10号线	485	3/6层	7260
3	BJZRHD68044884	整	57.00	海淀	东西	2室1厅	40.037549	116.336905	13号线	912	6/6层	5660
4	BJZRCY16000582	整	46.12	朝阳	东	2室1厅	39.933755	116.621507	6号线	332	11/17层	5780
5	BJZRXC95874513	整	43.10	西城	南	1室1厅	39.940291	116.353434	6号线	264	6/6层	5490

图 2.2 清理后的数据

第三章 统计分析

(一)房屋数量及空间分布

统计各个行政区域，合租和整租的房间数量，如图 3.1 所示。自如大部分的房源集中在朝阳、海淀和丰台地区，其中朝阳区出租的房屋数量，比第二名海淀区和第三名丰台区加起来还要多。朝阳、海淀为北京城市功能核心区，人口密度最大，对租房的需求更大。

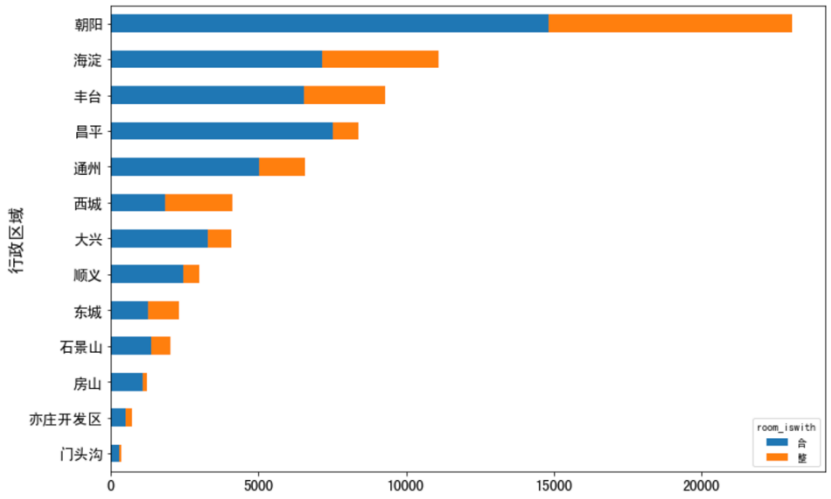

图 3.1 各个行政区域，合租和整租的房间数量

同时，我们分析了各个行政区域，合租和整租的占比，如图 3.2 所示。整体上来看，自如

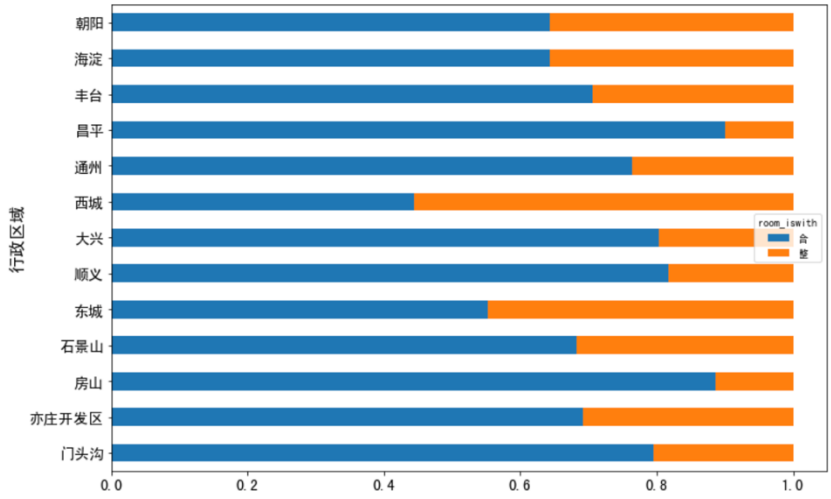

图 3.2 各个行政区域，合租和整租的占比

合租的供给量大于整租，从租房市场来看，租房者更加倾向于合租，减少租金压力，因此合租的需求量大于整租。但在北京首都功能核心区的东城区和西城区，整租的房源数占比较大，租金偏高，一般不是普通的租房者居住。

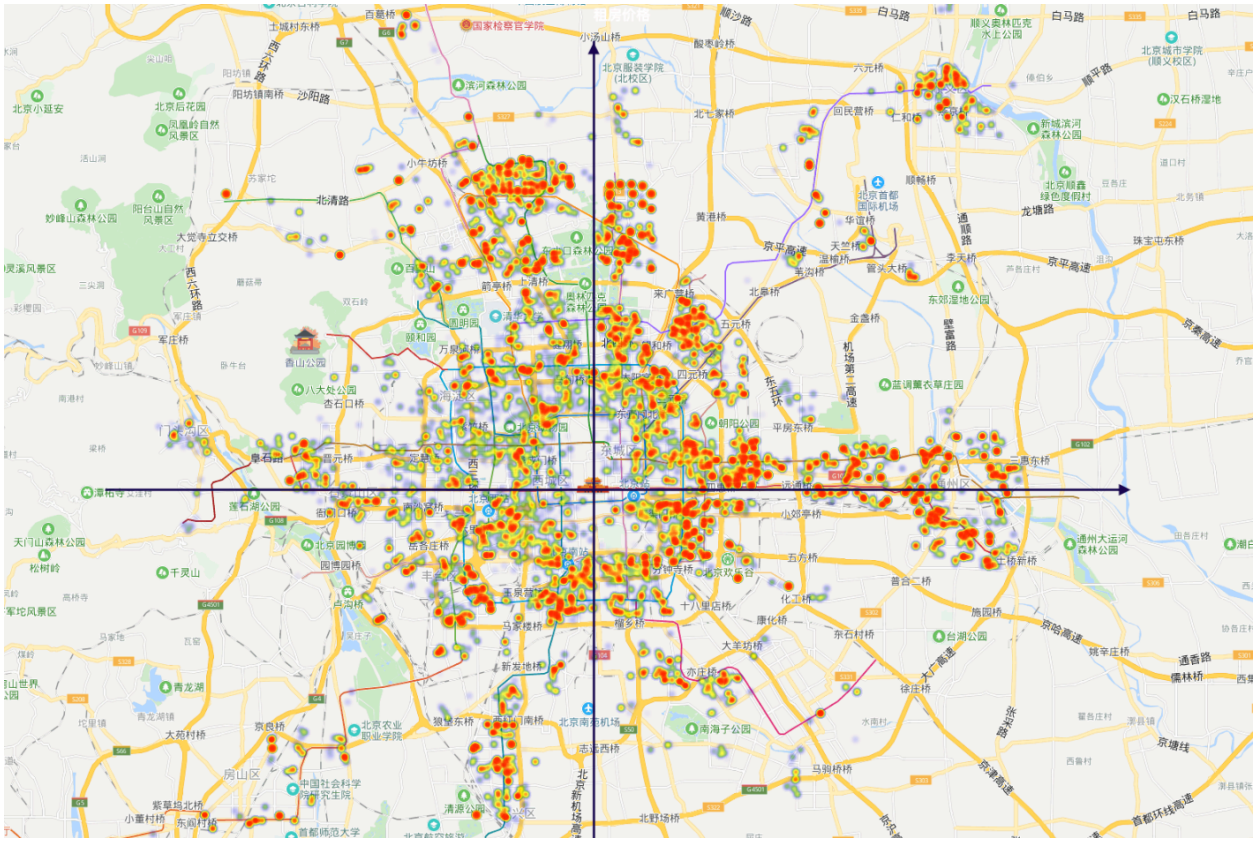

图 3.3 自如房源分布

进一步，我们对根据房源的地理位置绘制地图热力图，如图 3.3 所示。房屋的分布呈现以下几个特点：

1、沿地铁线分布：主要分布在 10 号线的东南段，1 号线的东西段和八通线沿线。以及四号线南部末段。和 5 号线北段，以及 15 号末端。这些位置交通较为便利，地铁末端距离城市中心较远，租金也相对便宜。

2、大聚居，小散居：在西二旗，望京，国贸，房源较为集中。这几个地方公司较多，适合想离公司较近的工作者居住。在海淀区二环到四环之间的区域，房源较为稀疏，这里聚集了大量的高校，高校占地面积较大，将租房区域切分成零散的区域。还有一些更为偏远的行政区，如门头沟区，房源稀少，分布零散。

3、空间聚集：以一号线和北京中轴线划分，房源主要分布在东北，东南，南北和北部位置。在西部区域分布较少。

(二)房屋价格分布

通过对房租价格的分析，如图 3.4 所示，按合租价格降序排序。不考虑房间大小，行政区域的热点地区等因素的影响。合租和整租的房间，可以分为三档，第一档为西城、东城、朝阳和海淀,合租平均价为 2969 元，整租平均价为 6275 元；第二档为丰台、亦庄开发区、昌平和石景山，合租平均价为 2278 元左右，整租平均价为 4822 元；第三档为大兴、通州、顺义、门头沟和房山，合租平均价为 1740 元左右，整租平均价为 3728 元。综合来看，整租的平均价为合租的平均价 2 倍以上。

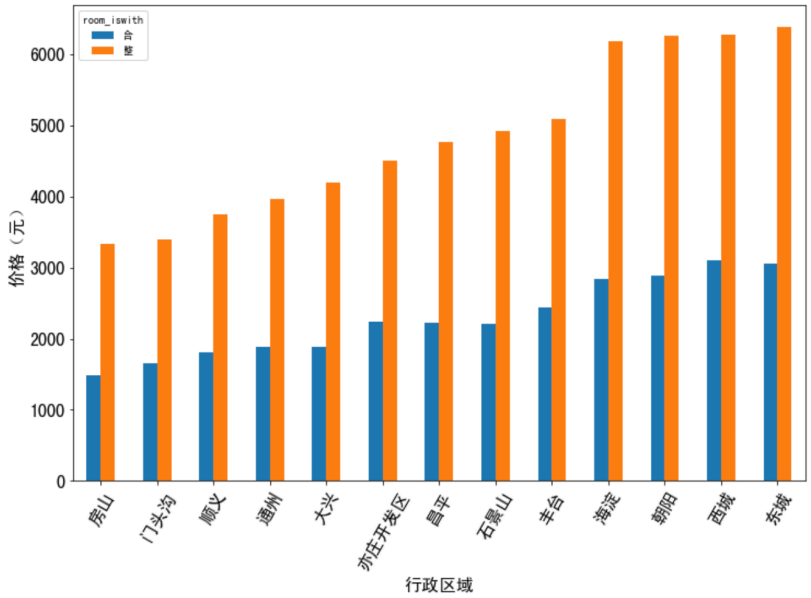

图 3.4 各行政区域，合租和整租的平均价格分布

一般来说。房间越大，价格会越高，因此排除房间大小对价格的影响，我们通过计算每个行政区域每平米平均价，如图 3.5 所示。与平均价的分布一致，可以分为三档。东城、西城、海淀和朝阳为第一档，平均价 239 元/m^2；丰台、石景山、昌平和亦庄开发区为第二档，平均价 170 元/m^2； 通州、大兴、门头沟、顺义和房山为第三档，平均价 134 元/m^2。

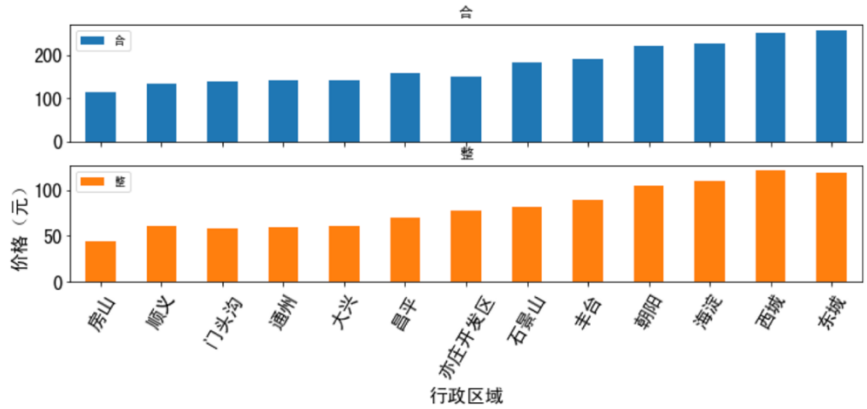

图 3.5 各行政区域，合租和整租的每平米价格分布

进一步，我们对合租和整租分别绘制了箱线图，如图 3.6 所示。整体来说，北京中部、西北部的租房价格相比于北京南部和最东部区域，租房的价格相对较高。四环以内的租房价格比四环外，价格更高。北京经济活力较大，人群密集的几个区域如中关村，望京，国贸等附近，其房价也相对偏高。

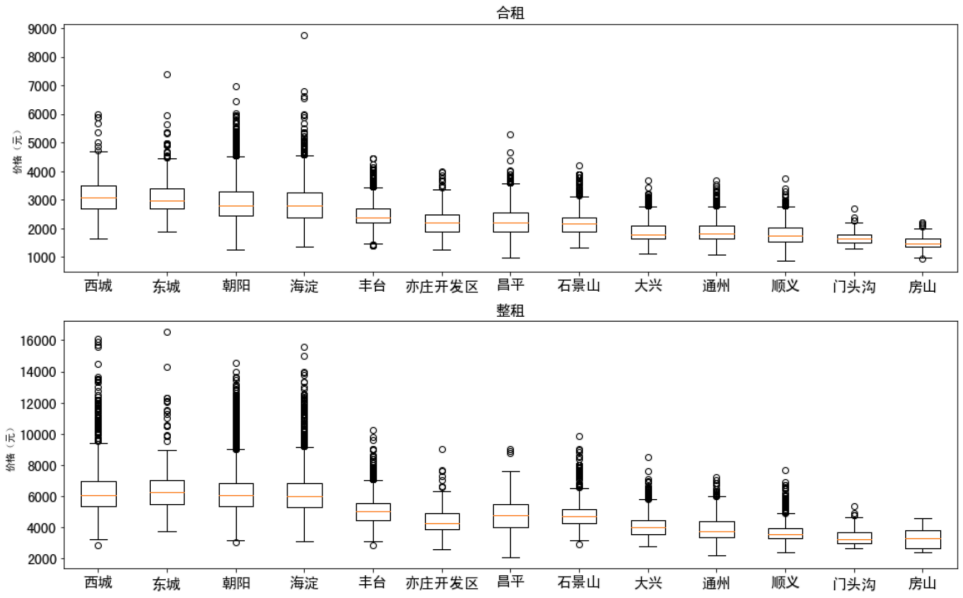

图 3.6 各行政区域，合租和整租的价格统计分布

(三)房屋性价比

租房者在选房时，一般会考虑离地铁距离、公司距离、房屋大小、价格、楼层高度等因素。因此为了综合比较各个行政区域的租房性价比，我们选择了每个行政区域的六个维度：

1、平均地铁站距离

2、合租和整租每平米价格

3、房间的平均大小

4、楼层平均高度

5、房源到北京市中心（以天安门为测量点）的平均距离

6、距离热门区域（国贸、望京、中关村、西二旗）的最短平均距离。

然后对这六个维度进行打分（0-100 分），计算方式如下：

$$score = (1 - \frac{\mathrm{x} - \mathrm{x}_{min}}{x_{max} - x_{min}}) * 100$$

特殊的，对于房间的平均大小的分数的计算方式为：

$$score = \frac{\mathrm{x} - \mathrm{x}_{min}}{x_{max} - x_{min}} * 100$$

一般距离越短，得分越高；价格越贵，得分越低；房间越大，得分越高。对于合租和整租的房屋分别打分，根据打分结果，我们绘制了各个区域的性价比雷达图，如图 3.7，3.8 所示:

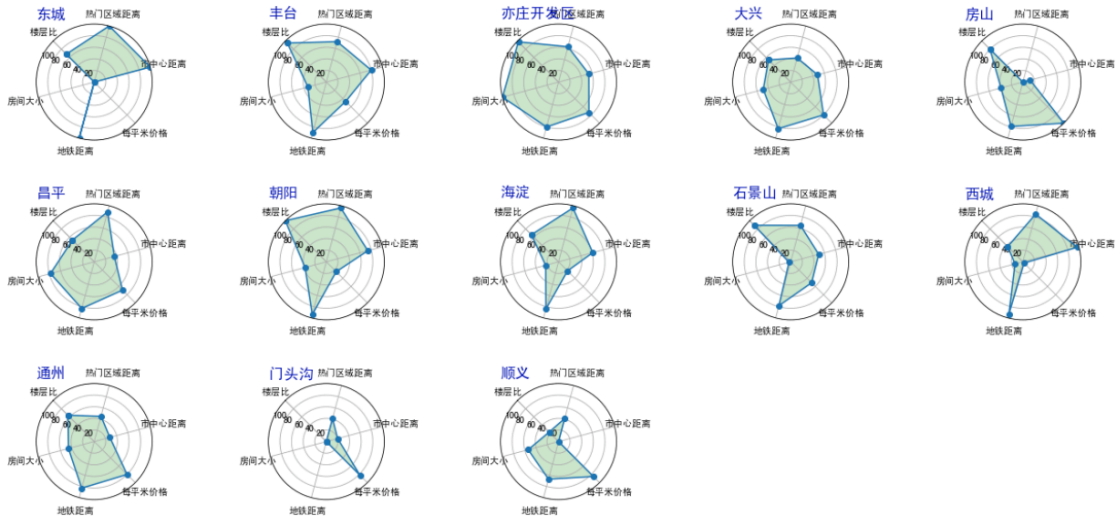

图 3.7 合租性价比雷达图

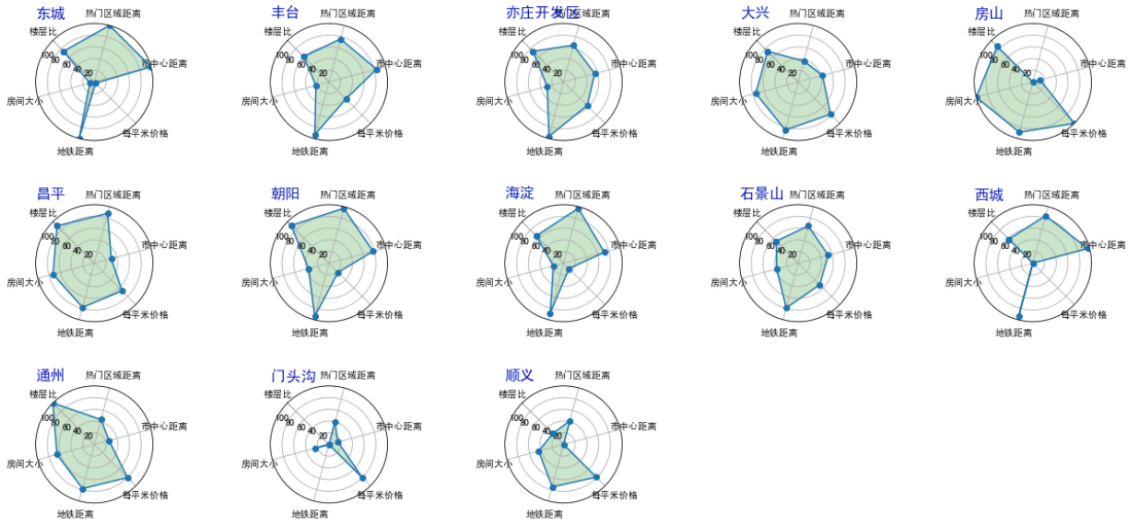

图 3.8 整租性价比雷达图

通过雷达图，我们可以发现，对于东城和西城等城市中心的区域，其房间大小和每平米价格得分都很低，但是其距离地铁近，位置好。而对于门头沟这些偏远地区，除了价格较低，其它的都不行。雷达图很好的反映了各个区域房源的优点和缺点，可为租房者提供一个参考。

第四章 总结和展望

本文通过对自如租房数据的采集，处理，分析了自如的房源的数量和空间分布，并分析了每个区域的价格，最后用六个维度评价每个区域的优点和缺点。当然，本文对数据的分析不够深入，未对房源的朝向，房间数量进行分析。另外，房间的价格跟时间存在一点的关系，如每年 5-7 月毕业季，必然存在房价上涨的情况，每个区域的平均价格会有变化，由于采集数据的时间较短，无法充分的使用时间这个特征。本文也未深入分析各个热点地区和地铁线价格分布，后期可继续探索的空间很大。

附录

代码和数据：https://github.com/xuxping/ziru