黔南民族职业技术学院

QIANNAN POLYTECHNIC FOR NATIONALITIES

毕业设计说明书

题目：《贝壳平台租房数据分析》

系　　部：	大数据与电子商务系
专　　业：	大数据技术与应用
班　　级：	2020 大数据技术与应用（5）班
学　　号：	202003170536
姓　　名：	窦友俊
指导教师：	黎龙珍

二〇 22 年 9 月 6 日

承 诺 诚 信

我将在此保证：窦友俊的毕业设计《贝壳平台租房数据分析》为自己独自编写，没有作弊行为，源代码中大部分代码写法有独特风格，均作了特别处理，如果说谎，后果由本人承担。

承诺人（签名）：

年　　月　　日

论文题目：《贝壳平台租房数据分析》

专　　业：大数据技术与应用

姓　　名：窦友俊　　（签名）_____

指导教师：黎龙珍　　（签名）_____

摘要

在现在的个人生活中，租房是不可缺少的一部分。如何利用技术更多掌握租房信息已变得重要,数据分析正是为租客和自己提供了这样一种有效的分析方法。

未来租赁市场发展将走向成熟化。住房租赁市场机构化比例将继续提升，专业化、精细化运营的租赁企业将为不同群体提供多样化的租赁产品和服务，未来的租房市场只会越来越大，太多的社会青年更倾向于高品质的生活，更愿意超前消费，加上国家政策的支持，国家在宏观调控下经济越来越好。

本章主要从地点、优势、价格三个方面进行分析，收集官网提供的数据集与 Python 爬取的数据信息，利用 MYsql 对数据进行进一步处理与存储，使用 pyecharts 进行分析与可视化实现，最后对书局结果进行表明并提供相应处理。

关 键 词：Python、mysql、pyecharts、贝壳租房数据、数据技术分析

1 目录

1.1 简介与目的

1.1.1 简介

随着国家政策支持力度逐渐加大，流动人口租赁需求持续支撑，住房租赁市场将迎来蓬勃发展期，重点城市群、核心城市的人口吸引力更强，这些城市的住房租赁市场拥有更广阔的发展前景，随着家庭户规模逐渐小型化，小户型住房需求将有所增加，利用平台的房主发布的房源信息，快速找到相关房源，从而达到精准营销的快速商场理念；利用平台的信息，以及其他房主发布的信息后有效的进行手里的资源整合，达到多方面投放。

1.1.2 目的

目标：

本文从租房平台网站同步数据访问及个别房源两个方面的多维度来进行资源分析，通过大量数据的真实情况和线下情况可以把租客对房源的大致了解，对现有的房源有所了解，在某个时间段的范围内销售情况了解，达到了解整体是进还是退。为了进一步了解租房平台房源的整体销售情况，从时间空间多维度分析房销量趋势、拐点、异常，以及用户的整体活跃情况，留存情况等，并通过各种查看数据对地方房源运营状态，发现经营问题，找到提高销量的可行性方法，为接下来的运营工作提供建议。

意义：

（1） 提示房源的问题所在。

（2） 提示某个地方的房源为什么销量差。

（3） 指导我们在将来如何有效提高销量的方法和建议。

1.2 应用领域

python 在很多领域有着很重要的作用，在各个领域中发挥价值，比如云平台，人工智能，金融分析

1) python 有·相对较小的关键字，结构简单，语法定义明确

2) 易于阅读，代码定义清晰，

3) 易于维护，它的源代码相对于其他语言要简单得多

4) 具有丰富的库，这是他最大的优势之一

5) 可与移植性，python 可以应用到多平台

6) 免费，开源

7) 简单开源强大多平台的

1.3 python Scrapy 框架

（1）Item 对象、容器

框架编写，定义 item 类，对数据进行预选择，预存储到容器进行处理，便于后续开发

（2）数据打印

数据运用 css 选择器选择出 item，打印出来后通过爬虫方法，获取到原数据达到数据获取目的

（3）数据持久化存储

得到经过处理后的数据进行数据写入数据库操作，进行持久化存储，便以分析

（4）分析数据

分析数据以便后续便以开发可视化，找到数据中有关系的数据

（5）数据可视化处理

主要是把数据的关系进行可视化处理，让数据的显示方式以图表的方式展示出来

1.4.1 框架构建

（1）数据获取：本次项目数据的获取方式是利用 Scrapy 、对 https://bj.ke.com/ 网站进行爬取。

（2）数据清洗：本文数据清洗使用的是 python 函数，通过 python 语言的 Scrapy 框架爬取数据，写入库操作。在写入数据库之前，去除无用字段、整理数据入库。

（3）数据分析：在数据库中使用 sql 语言进行数据分析。

（4）分析结果展示与说明：将分析结果通过 pyecharts 绘制图表可视化展现。

1.4.2 路由分析

本文结合研究内容和研究方法，制定具体技术路线图如图 1.1 所示。

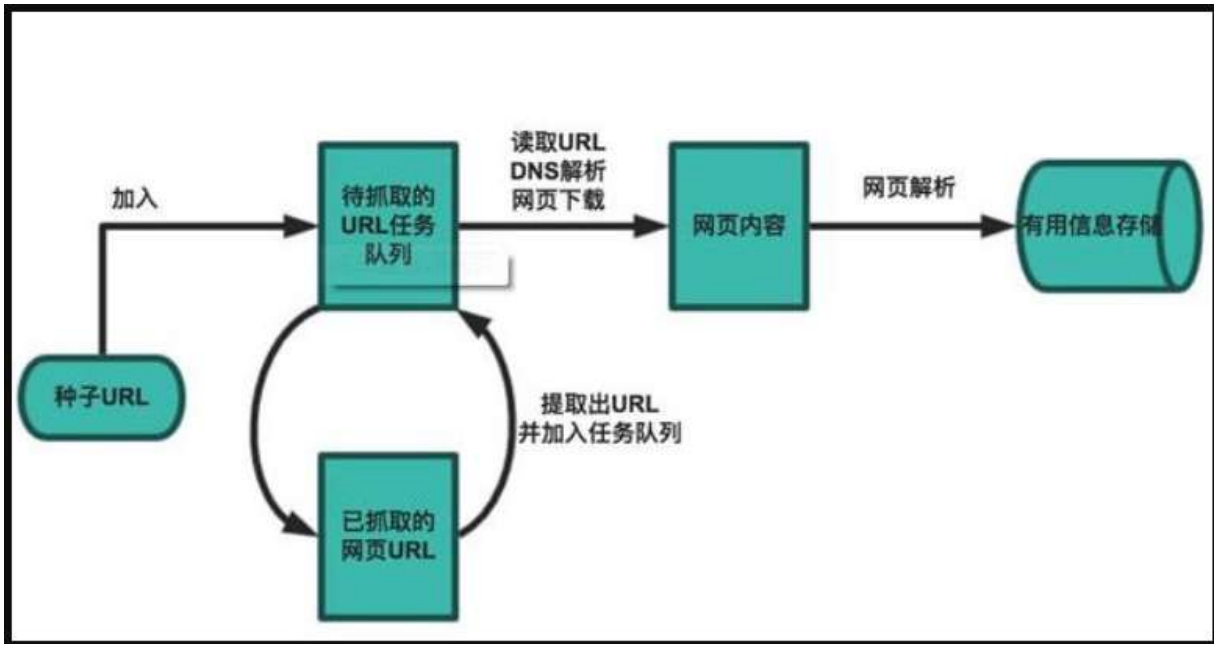


图 1.1 路由示意图

2 开发环境及软硬件

2.1 设备措施

2.1.1 操作系统及 CPU

liunx、windows10、max 等均可

CPU：酷睿 i5

其他：网络、交换机、防火墙、输入输出设备等

2.2IDE 与软件

pyCharm 专业版，专业版具有相对于社区版更好的稳定性和更多的专业支持，相比于社区版，专业版在编写代码时明显比社区版要支持得多。

2.2.2 IDE

这次使用 PyCharm 专业版开发环境，IDE 得使用会增加代码得流畅度以及编写速度，PyCharm 是编写 Python 程序的大多数人的选择，它是专门写 Python 的一款 IDE。

2.2.1 MySql 数据库

MySQL 是现代技术中小型关系型数据库管理系统，Mysql 关系型数据库是现代数据库最常用的数据库类型，它的访问速度稳定、数据表示更便捷。MySQL 通过结构化查询语言 SQL，数据查询简便快捷。MySQL 软件使用了 GPL，即 GNU 通用公共许可证，保证了我们开源共享有网络协议支持。即使与其他的大型数据库如 SQLite、大型数据库 等相比，MySQL 规模相对比较小，功能没有大型数据库强大，甚至还有大量缺点，但是对于中小型系统来说，MySQL 的性能和稳定的 SQL 于体积小、速度快、跨平台，特别是开放源码这一特点，可以获得大量免费资源，GitHub 仓库有许多开源网站搭建系统都采用 MYSQL 来搭建网站，以减低网站在服务器上部署带来大额账单。

Mysql 的主要核心部分采用完全的多线程，服务灵活，支持多线程，充分发挥出 CPU 资源。MySQL 支持多种接口，JDBC、ODBC、PHP、C 语言、Python、Perl、Ruby、VB 都可以与 MySQL 数据库直接连接提出数据。MySQL 服务器提供了连接池，也就是多线程，可以同时为多个客户端提供服务。 常用的 MySQL 管理工具有 MySQL Workbench、MYSQL Front 和 PHPMy Admin 等，其中最方便的是 Web 的 PHPMyAdmin 工具。PHPMyAdmin 是用 PHP 编写的，免费地数据库 web 端管理工具，可以通过 web 前端和控制 Mysql 数据库。它支持常用的数据库操作，包括管理数据库、表格、字段、联系、索引、用户、许可等增、删、改、查各种操作。PHPMyAdmin 的缺点是必须安装在 Web 服务器上，所以访问权限如果设置的不正确，就有可能遭到黑客攻击使数据受损。

2.3python 程序

Python 是一个高层次的结合了解释性、编译性、互动性和面向对象的脚本语言。

2.3.1 Scrapy 框架部署

python 、pyecharts 、Mysql 、json 、xpath、css 选择器等。

1. Python：主要是使用 Scrapy 框架
2. Pyecharts 是可视化的一把利器
3. Mysql 是小型关系型数据库
4. Json 是数据集
5. Xpath 是 python 中选择数据的表达式
6. Css 是 python 中选择数据的表达式

2.3.2 Python

python 是我们在校所学的程序，主要是用来做爬虫的，我们学了很久，python 语言是非常强大的，而且很简单，在很多大公司都在使用 python 语言比如 Google 爬虫，Google 投放的广告大量都用 python 来进行开发，就连世界上最大的视频 YouTube 也在使用 python，如今 python 已经是世界上最受欢迎的语言之一了，python 能走到今天也是有很大的原因的，他有非常强大的各种库支持，就算没有经理过严格训练的普通人也能感受到 python 的强大，python 的应用不止应用在爬虫，还有人工智能，python 可以运用于自动化办公，比如 **SeaTable** 一款新型的协同表格和信息管理工具 python 运用的领域不止于以上这些，这些都是凤毛麟角，python 的强大是证明它能在编程语言里面脱颖而出的重要原因，Python 语言是跨平台的，他的语法很简单，很短的代码也能做很多事，同时他也是脚本语言，随便一段语言就可以处理数据，十分方便。同时它也是面向对象语言，对初学者十分友好。Python 在处理各个领域的类库也十分丰富，爬虫、机器学习、数据处理、图像处理等等满足了大部分领域的需要。

2.3.3 SQL

SQL 是在处理数据库时使用的标准语言，具有查询复杂数据以及多数据的功能

在这次论文中，具体来使用 SQL 和处理数据系统中的数据，这类数据库包括：MySQL、SQL Server、Access、Oracle、Sybase、SQLite 等等；

3 程序流程控制

Scrapy 是基于 Python 编写的一个为爬虫、为了提取网站数据的应用框架，Scrapy 经常被应用在一系列的程序中，包括数据挖掘、信息处理或存储历史数据我们可以使用 Scrapy 的 python 框架爬虫爬取一个网站的数据或者图片

3.1 部署 Scrapy 项目

scrapy startproject deins

cd deins

scrapy genspider junwd junwd.xyz

3.1.1 部署项目管道

配置 settings.py 文件，开通请求头配置项目管道开启项目等，本人掌握 python、sql 等数据处理技术，并且熟练使用 mySql 数据库、pycharm 集成开发环境，并且在 YouTube 上比较容易获取找到相关技术文档，因此技术思路清晰。

3.1.2 项目爬虫数据分析

在 items.py 文件中定义数据类型，编写主程序 bk.py，使用 css 选择器选择数据，并存储本此分析数据由爬取网站信息的并进行整理，数据来源虽有时效性，但真实可靠而且可以轻松获取。

3.2 程序编写

基于 python 的 scrapy 框架爬虫速度非常快，很让人省心

3.2.1 分析网站数据

数据集源自 https://bj.ke.com/网站的爬取，是贝壳的实时更新数据，包括字段：房名、价格、租赁方式、房屋类型、房屋朝向、面积、楼层、用水、用电、租期、燃气、电梯、车位。

3.2.2 数据打印查看

先打印出数据分析需要的字段以及对数据进行关联、清洗等预处理内容。将收集到的信息表导入 Mysql 数据库中进行存储与预处理，因分析内容涉及较多的分析，所以首先需关联房名信息表和价格信息表以及其他表，并在表 1 中添加不同字段，接下来只需要对表 1 进行清洗。

（1）添加排序

根据数据信息写入数据库自动添加排序，方便最后查看数据总量。

（2）列重命名

为更好的方便区分数据，使用中文来为表命名

（3）删除重复值

数据在写入之前已经做个数据清洗

（4）缺失值处理

对数据精准抓取，所以暂不存在数据缺失

（5）一致化处理

数据面面积是两条数据，为整理区分数据，故作处理为一条

（6）数据排序

按写入时间排序。

（7）异常值处理

a. 房名有大量空格，但不影响数据整体使用，所以故不作处理。

b. 有无电梯字段大量相同，但是有用数据，故不作处理。

c. 用电字段大量相同，但是有用数据，故不作处理。

d. 用水字段大量相同，但是有用数据，故不作处理。

3.2.3 数据分析思路

租房平台总体分析可以从网站爬取的数据和观察其他网站两个大的维度来进行综合分析，通过整体情况可以把控平台房源的现状，对房源以及数据有所了解，在一定的时间范围内房源情况如何，整体是增加还是减少。为了解租房平台房源的整体情况，可以从时间维度分析房源发布趋势、拐点、异常，以及用户的整体活跃情况，留存情况等。

4 系统的设计与实现

先确定好需要面向的数据，使用 python scrapy 框架进行程序编写

1.1 各文件实现

beike －目录

spiders—目录

bk.py—主程序

items.py—定义 item 类

middlewares.py—其他

pipelines.py—项目管道

settings.py—配置项

scrapy.cfg--配置文件

4.1.1 请求数据

请求网站 https://bj.ke.com/，在网站搜索栏中搜索需要的数据集，选择需要的数据集并使用 css 选择器抓取。

4.1.2 数据存储

（截图部分数据，）

数据表与少部分内容如 3.1 所示，目前共有 590 条数据。

1	花果园M区 4居室 南卧	1160	合租	4室1厅2卫	南	18.00㎡	8/41层	民水	民电	3~12个月	有	有
2	高速路政管理所宿舍 5	1200	合租	5室1厅2卫	南	22.00㎡	12/31层	民水	民电	6~12个月	有	无
3	花香上海城 1室1厅 南	1080	整租	1室1厅1卫	南	50.00㎡	16/20层	民水	民电	1个月以上	有	无
4	世纪城龙祺苑 3室2厅	1700	整租	3室2厅2卫	东南	111.00㎡	中楼层/33层	民水	民电	暂无数据	有	有
5	未来方舟G1组团 4居室	580	合租	4室1厅2卫	南	25.00㎡	27/31层	民水	民电	1个月以上	有	有
6	中铁逸都国际C区 4室2	2200	整租	4室2厅2卫	南	157.00㎡	低楼层/17层	民水	民电	暂无数据	有	有
7	中大国际广场 1室1厅	1030	整租	1室1厅1卫	南	50.00㎡	29/42层	民水	民电	1个月以上	有	有
8	碧桂园西南上城 5室2厅	3500	整租	5室2厅2卫	南	204.41㎡	低楼层/28层	暂无数据	暂无数据	1年以内	暂无数据	无
9	一鸣	580	合租	4室1厅2卫	南	25.00㎡	29/33层	民水	民电	1个月以上	有	有
10	龙凯苑 3室2厅 东南	2200	整租	3室2厅2卫	东南	118.07㎡	低楼层/25层	民水	民电	暂无数据	有	有
11	贵阳宇虹万花城 5居室	880	合租	5室1厅2卫	南	13.00㎡	30/32层	民水	民电	6~12个月	有	无
12	中天会展城B区 2室2厅	9600	整租	2室2厅1卫	南	89.00㎡	低楼层/32层	民水	民电	1年以内	有	有
13	凤凰栖一期 2室2厅 东	2000	整租	2室2厅1卫	东南	80.00㎡	中楼层/32层	暂无数据	暂无数据	1年以内	暂无数据	无
14	经典时代花园广场 2室	3200	整租	2室1厅1卫	南	90.00㎡	中楼层/27层	民水	暂无数据	1年以内	有	有
15	贵阳金融城 1室1厅 南	1400	整租	1室1厅1卫	南	47.00㎡	13/18层	民水	商电	1~2年	有	有
16	花果园A南区 3室2厅	2300	整租	3室2厅1卫	东南	85.00㎡	低楼层/35层	民水	民电	1年以内	有	有
17	盛世花城 3室2厅 南	4000	整租	3室2厅2卫	南	136.00㎡	中楼层/32层	民水	民电	暂无数据	有	有
18	优品道	1800	整租	2室2厅1卫	西南 西	65.00㎡	高楼层/33层	民水	民电	暂无数据	有	有
19	睿力上城 4居室 南卧	730	合租	4室1厅2卫	南	9.00㎡	12/30层	民水	民电	6~12个月	有	有
20	龙盛苑 3室2厅 南	2300	整租	3室2厅2卫	南	110.00㎡	低楼层/30层	民水	民电	1年以内	有	有
21	华润国际社区 5居室 南	1550	合租	5室1厅2卫	南	20.00㎡	6/33层	民水	民电	6~12个月	有	有

图 4.1 数据截图部分数据集

4.1.3 分析模块实现

（截图展示通过 mysql 实现的数据关联查询，多字段排序，获取随机值，取数据统计等，并使用文字大概说明）

4.1.4 可视化展示

详细描述进行可视化展示的平台如何处理得到的直观数据展示的结果，并对展示图的含义进行简要解释[3]。

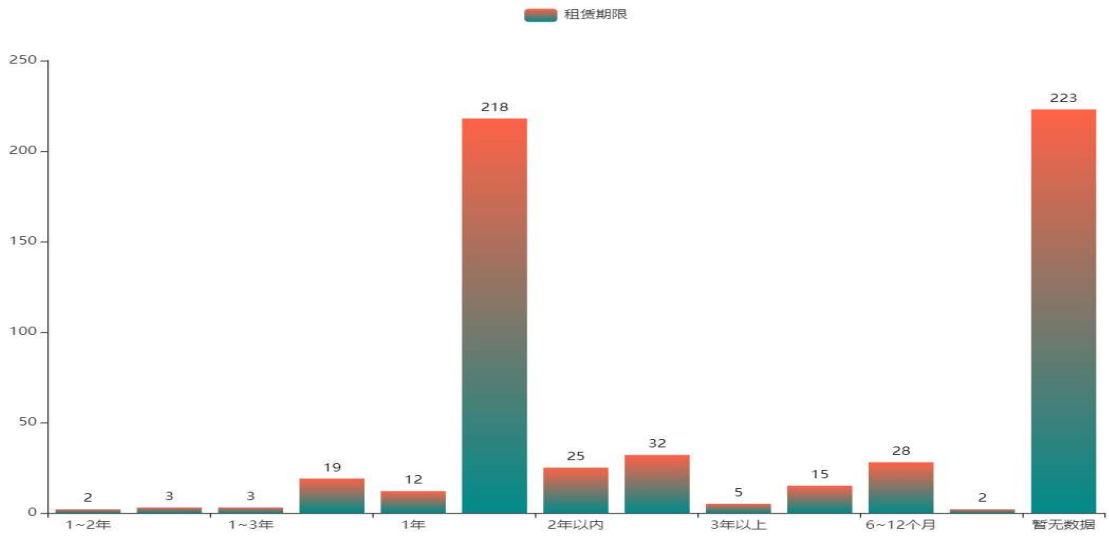
例如：

1. 店铺销量趋势

贝壳平台的房源和租期长度有销售关系。统计了现有房源对租期的要求，并对现有房源统计各房源需要租多少天进行统计，二者是不同的概念，所以需要分别进行分析与展示，首先从租期角度分析。

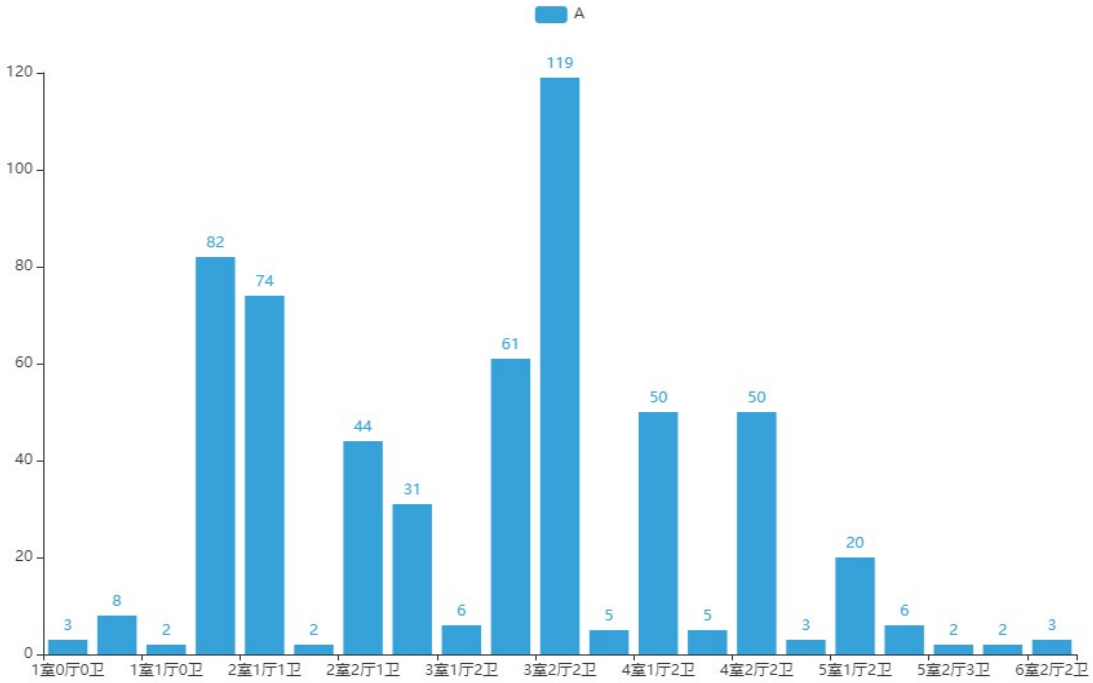
1 可视化租期图

1. 房源租期统计



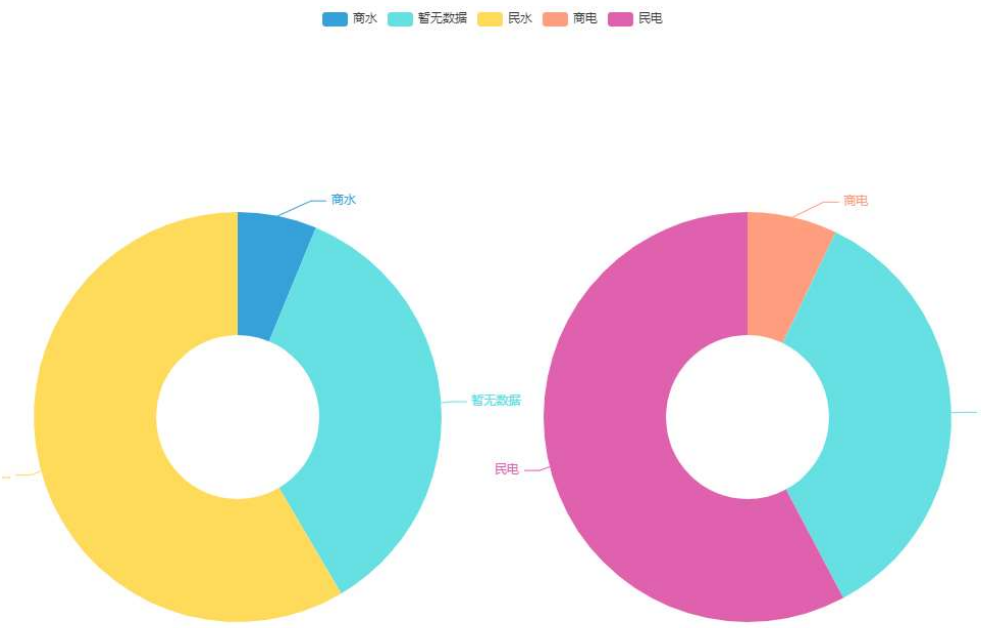
在随机抓取的数据中统计出期限时间数量，统计出租期的租房房源，其中一年以下最多，可以看出对租房的选择只是暂时性的，其他较多的一部风还是有待商议，并未做出租期选择。

2. 源类型趋势图



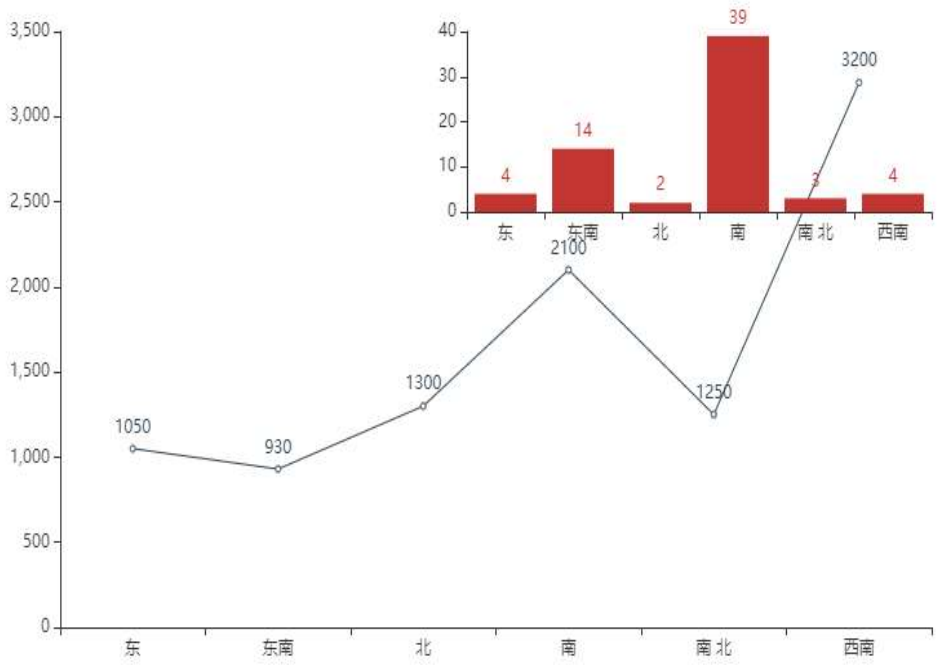
从图里可以看出，每种房源各有占比，其中三室二厅的房子最多，五室二厅三卫只有两套，可以看出需求最少，以及各种类型在同一时间类的占比。

3. 可视化水电图



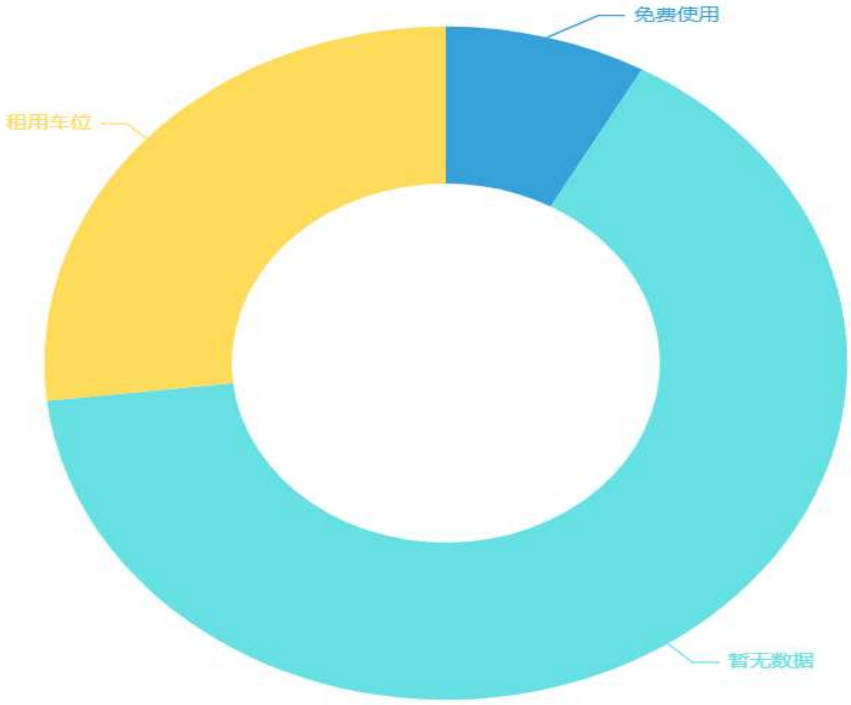
在数据有限的情况下统计出随机数据的统计结果，其中在贝壳租房平台的房源数据里，属于民房的占比明显比商用房的占比多，民水、民电数据统计出来结果大致相同，其中一部分并未表示出属于哪一类型。

4 可视化朝向图



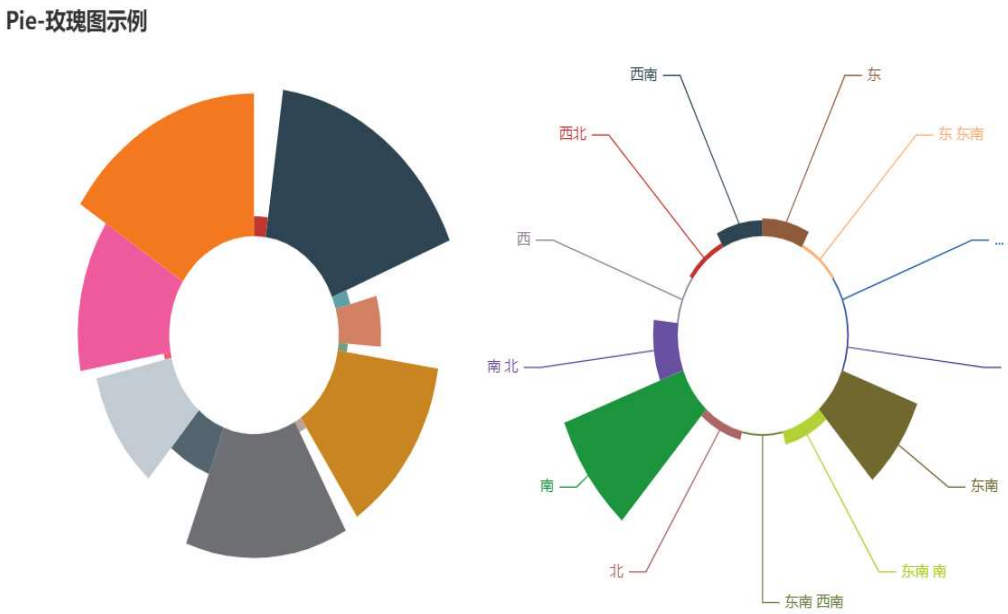
从图中可以看得出来，在所知数据库中房屋朝向朝南的最多，价格中等，是房源类型中最有竞争力的房源，其他朝向只有东南仅次于南向，大致原因可能是采光好，其他的可能是背对光源，所以房源较小。

5 . 车位占比图



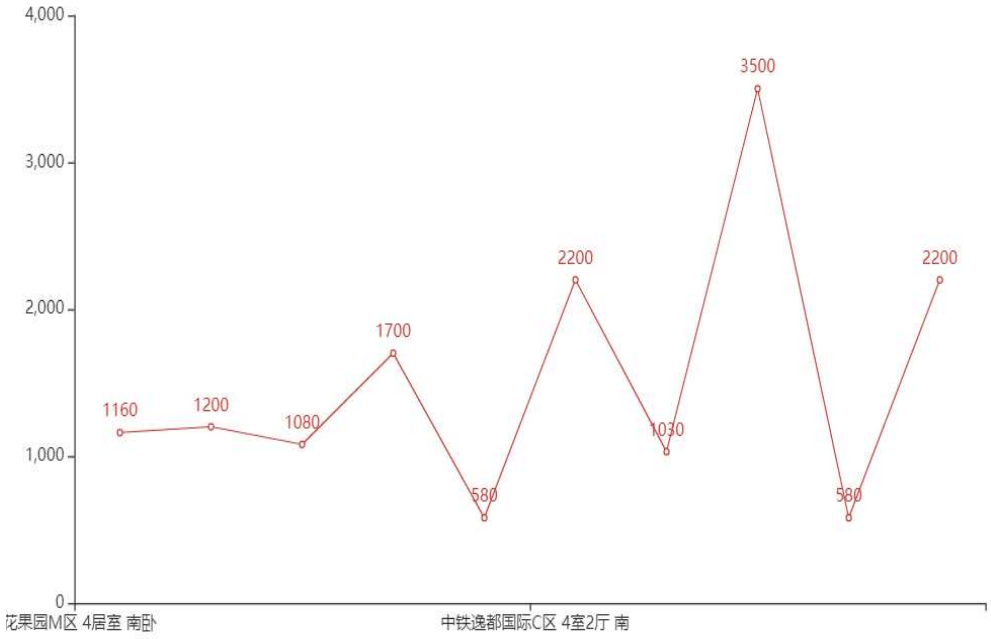
从图中可以看得出来，在所知数据库中大多数都没有配备车位，有车位的大多要钱，免费的车位很少。

6. 面积和朝向占比



从图中可以看得出来，在所知数据库中 150 平的房源最多，朝南的房型是最有竞争力的房源，其他朝向只有东南仅次于南向，大致原因可能是采光好，其他的可能是背对光源，所以房源较小。

7. 房价格趋势图



从图中可以看得出来，3500 元每月的房型最多，1160 的房型较少，1700 元的属于中等的房型

4.2 结论

通过数据分析得到的结果，针对问题总结说明数据中体现的规律与现象。

例如：

a. 根据数据展示可以得出对于租房的关键字为近地铁，交通便利、押一付一、租房保证等，分析得出租客 20 到 30 岁之间的社会青年，这部分群体更注重于超前消费和享受，对于租房有一点的要求。多是为了更容易的上下班。

b. 通过数据分析租房关键字我得出租客对于租房更多是整租，在数据库里统计的所有数据可以得出整租有 502 套，合租只有 82 套，可以看得出来消费群体更注重个人隐私，对于合租还是有一定心里芥蒂。

针对（1）结论中的现象中提出一定的建议。

例如：

a.开展一系列优惠活动，促进用户群体对于性价比的追求，重点主推近地铁，交通便利的房源，房源展示的图片是本房真实照片真实在租：房源在租的状态是真实的真实价格：

b.针对这一问题可以很好理解，客户对于隐私的保护、不轻易分析用户的数据，对外加强数据的管控，对内严格要求，更多注重客户隐私。让客户住的心安。加强对房源的来源进行真实性检查，在贝壳找房平台发布的房源价格是真实价格居住面积合规（群租）：房屋的人均居住面积符合当地有关部门的规定

5 总结

1.将论文全文内容进行总结性简单陈述。

2.叙述毕业设计的初心点：

（1）python 框架挖掘数据

（2）数据入库整理

（3）数据分析及数据可视化。

3.利用 scrapy 框架编写爬虫程序，定义各种数据 item，对数据进行分析并对数据进行 css 选择器选择，整理数据后对数据进行写入数据库处理，并做出分析，再利用数据进行可视化。简单做出数据分析与论文编写过程中发现未解决的毛病或可以改进的地方。

致谢

这次论文是在班级导师黎龙珍线上线下辅导和监督下完成的。从选题到论文，黎龙珍老师一直关心着班上的进度和完成情况，并始终督促同学要以认真顽强、实事求是和不断创新的态度来对待这次论文和代码工作，对学习过程中的文献以及书籍要认真整理和总结。导师在知识上严谨的治学态度、渊博的专业学识、宽厚的为人及忘我的奉献拼搏精神都时刻影响着我。从论文的立意、选题、撰写和修改，每一处都有老师的极大的心血。值此论文完成之际，谨向我的导师黎龙珍致以最诚挚的谢意。

论文主要工作是对数据获取、存储、分析与可视化呈现进行介绍与总结，在程序及论文的开发与编写中，同学和老师也给我了及时的帮助。在此，衷心感谢各位老师及同学论文实验过程中的悉心指导和耐心帮助！感谢在校期间，大数据与电子商务系所有老师对我的关怀和教导！感谢班主任付老师、黎龙珍老师、以及同学们等在我大学期间成长过程中给予的宝贵建议！另外，还要感谢大学期间跟我朝夕相处的好友、同学们在生活和学习中对我的无私帮助！

最后，还要感谢我的家人给予我的鼎力关怀和支持，为我完成学业付出了辛劳！

介于我文笔有限，论文中难免有大量错误，望各位评委专家、教师在百忙之中抽出时间对论文进行建议，将感激不尽！

参考文献

[1] 范金程、梅长林.数据分析(第二版) [M]. 北京：科学出版社，2010.
[2] 赵蔷. (2022). *基于Python爬虫的旅游网站数据分析与可视化*.

附录

大学期间获得的奖项

[1]知进畏奖