摘 要

居住是国民安居乐业和社会稳定的关键所在，也是社会发展水平的综合表现，历来都受到国家的重视。当一个国家或地区的经济水平发展到一定程度以后，房地产业就会迎来一段快速增长时期，并进一步刺激和带动整个国家的经济增长，促进社会稳定繁荣。然而，随着中国房地产市场的发展，各种矛盾也逐渐暴露，特别是高房价问题，在中大型城市最为突出，已经远远超出了大多数居民的购买能力，为了更直观的反应如今的高房价问题，我们以成都为例，大数据为基础，开发了这个展示房价的网站。

开发环境：Windows、MacOS,Python3。

技术：mongodb、scrapy、beautifulsoup、django、pyecharts、numpy、pandas、matplotlib、wordcloud **and so on**...

开发人员：张鹏，龙栉，刘国洋，金先华

时间：2019.6.3 -- 2019.6.14

目录

[摘 要 1](#_Toc11606253)

[一、需求分析 3](#_Toc11606254)

[二、数据采集及数据处理 3](#_Toc11606255)

[1、数据采集 3](#_Toc11606256)

[2、数据清洗 3](#_Toc11606257)

[3、数据分析及可视化 4](#_Toc11606258)

一、需求分析

成都作为西南地区的门户，开发时间较早，人口较多，住房需求量大，让用户可以更直观的了解成都各个区域的房源及房价；但如今的新房房源并不多，所以为了有足够的数据，我们从新房入手，以成都市主城区为主进行分析。

二、数据采集及数据处理

## 1、数据采集

明确了需求之后，开始采集数据，数据采集即从网页上采集我们需要的指定信息，我们使用爬虫实现。当前开源的爬虫非常多，处于简便及学习的目的，在此使用python的scrapy框架，并BeautifulSoup解析网页获取指定的字段信息。获取的贝壳网上的新房数据，在URL链家网的房产数据都是以列表的方式存在，比较容易获取，其中包含的信息有楼盘名称、地址、价格等信息。采集的数据我们存放在mongodb中。



## 2、数据清洗

数据清洗，顾名思义就是将不合规的数据清理掉，留下可供我们能够正确分析的数据，至于哪些数据需要清理掉，则和我们最终的分析目标有一定的关系，可谓仁者见仁智者见智了。在这里，由于是基于地理位置做的一个统计分析，显然爬取的地理位置必须是准确的才行。但由于售房者填写的地址和楼盘名称可能有误，如何将这些有误的识别出来成为这里数据清洗成败的关键。经过清洗后的数据格式为：



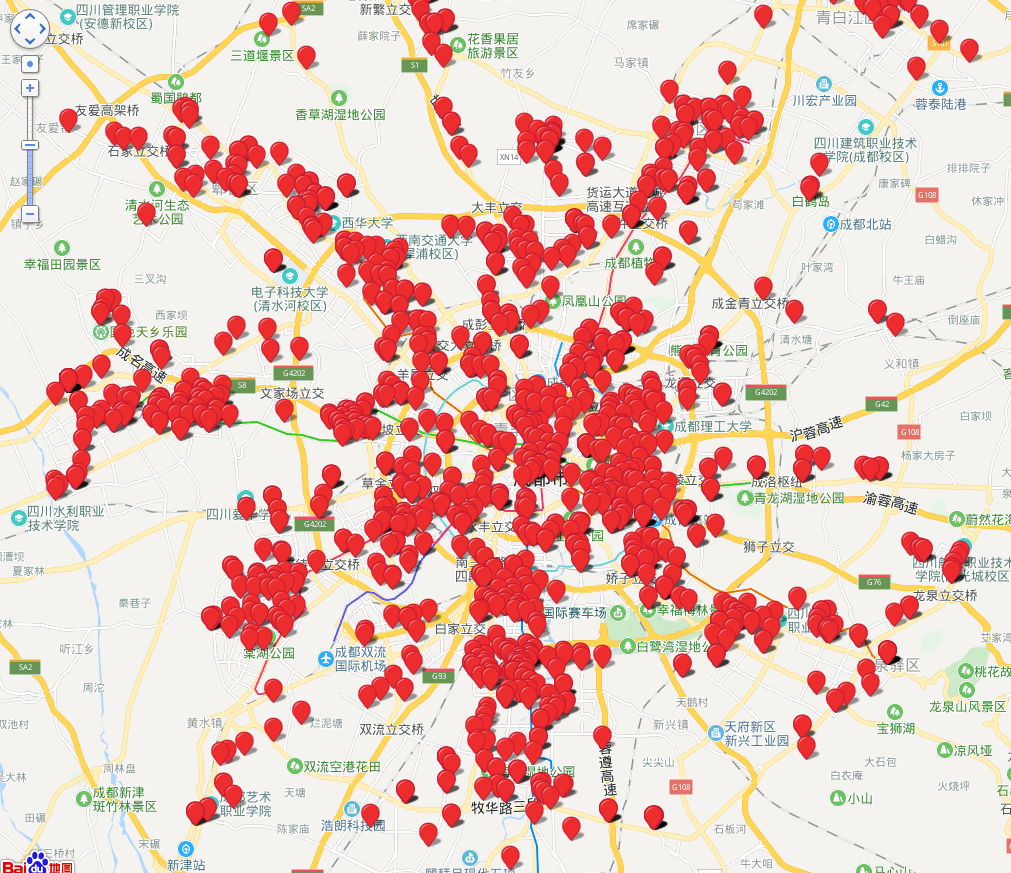
包括区、楼盘/房屋名称、地址、价格和新房类型五个维度。

## 3、数据分析及可视化

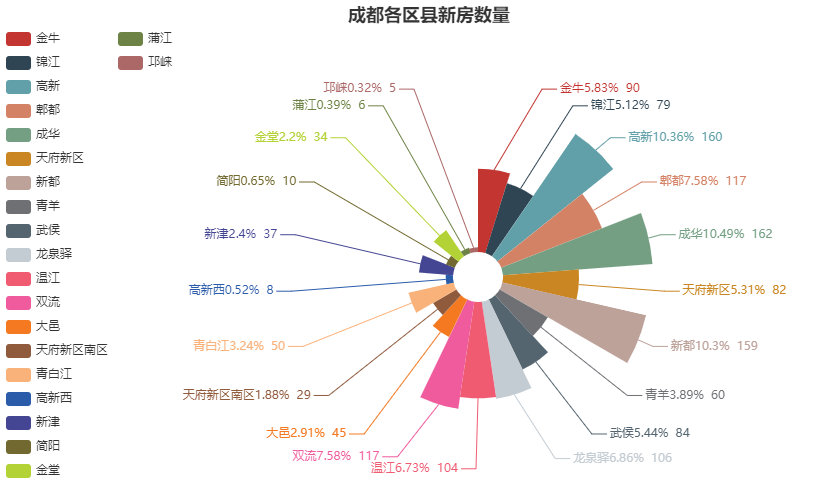
通过分析这些数据，再确定以怎样形式的图展示出来，才能更直观的看出各个区域的价格及房源。如：根据分析不同区域的房源及价格画出相应的统计图，并可得出结论：市中心依然遵循了寸独存金的原则，销售价格远远高于郊县，一方面原因是位置地段、配套的独特性，一方面也是由于可供销售的土地面积、楼盘数量极为有限。

为了有更好的用户体验，我们最终将得出的结论及相应的统计图用Django框架简单展示在网页上，并且将新房房源的具体位置展示在百度地图上，使用户在查看时更方便快捷。

房屋销售的位置分布可视化：



各区县新房数量比重可视化：



成都各种类新房各个价位分段及数量可视化：

