数据简介

**数据组一**

这是一组以小区为单位的北京二手房数据，包含了东城、西城、海淀区全部有二手房的小区（时间截止至2017年4月21日）。数据通过脚本抓取形式，在房天下官方网站(fang.com)进行抓取、导入Mogo dB数据库、并导出进行数据基础清理。基础清理中包括去除不相关字符、合并同一小区的不同名字、找出空值并标记、清理小区地址格式、清理经度纬度等。

|  |  |  |
| --- | --- | --- |
| **数据文件名** | **数据数量** | **简介** |
| dongcheng\_xiaoqu\_cleaned.xlsx | 487条 | 东城区-有二手房房源的全部小区及其基本信息 |
| xicheng\_xiaoqu\_cleaned.xlsx | 705条 | 西城区-有二手房房源的全部小区及其基本信息 |
| haidian\_xiaoqu\_cleaned.xlsx | 1139条 | 海淀区-有二手房房源的全部小区及其基本信息 |

截止录入数据日，其中东城区487条数据，西城区705条数据，海淀区1139条数据；每一个小区所包含信息包括小区均价、二手房房源数量和交易量、小区名称、地址等基本信息，以及小区地理位置的经度和纬度（见下表《表格内容》）。

通过已有的基础数据信息，可以对上述三个区的二手房小区的交易平均价格进行分析，并通过经度纬度的地理位置信息，在地图中标注不同区域内二手房交易小区的价格比较，辅以小区基础信息例如建筑年份等作为参照。

表格内容

|  |  |
| --- | --- |
| **标题行名称** | **对应内容** |
| Id | 数据库id |
| district | 北京所属行政区名 |
| average-price | 小区均价 |
| second-hand-resource | 二手房房源数量 |
| second-hand-trading | 二手房交易量 |
| court | 小区名称 |
| address | 小区地址 |
| house-type | 住宅类型 |
| build-time | 房屋建造年份 |
| url | 房天下网页链接 |
| latitude | 小区地理位置纬度 |
| longitude | 小区地理位置经度 |

**数据组二**

这是一组以单个二手房房源为单位的北京二手房数据，包含东城、西城、海淀区全部在列房天下官方网站的二手房数据（时间截止至2017年4月21日）。数据通过脚本抓取形式，在房天下官方网站(fang.com)进行抓取、导入Mogo dB数据库、并导出进行数据基础清理。基础清理中包括数据条目重新分类、去除不相关字符、找出空值并标记、合并重复地址等。

|  |  |  |
| --- | --- | --- |
| **数据文件名** | **数据数量** | **简介** |
| dongcheng-all-cleaned.xlsx | 22029条 | 东城区-全部二手房房源及其基本信息 |
| xicheng-all-cleaned.xlsx | 27325条 | 西城区-全部二手房房源及其基本信息 |
| haidian-all-cleaned.xlsx | 55417条 | 海淀区-全部二手房房源及其基本信息 |

截止录入数据日，其中东城区22029条数据，西城区27325条数据，海淀区55417条数据；每一个小区所包含信息包括单套二手房挂牌价、单价，以及房屋面积、户型、朝向、建造年份、楼层、是否装修、产权信息等基础房源信息（见下表《表格内容》）。

通过已有的基础数据信息，可以对上述三个区的所有二手房房源的基本信息作了解，通过挂牌价格、房屋面积等信息比较分析不同区域内二手房存量的情况，可以找出房屋面积特别小，但同时挂牌单价特别高的房源进行单个研究。

表格内容

|  |  |
| --- | --- |
| **标题行名称** | **对应内容** |
| Id | 数据库id |
| district | 北京所属行政区名 |
| sub\_district | 所属商圈名字 |
| build\_form | 建筑类别 |
| build\_time | 房屋建造年份 |
| unit\_price | 单价 （/平方米） |
| update\_time | 发布时间 |
| house\_form | 住宅类别 |
| floor\_plan | 户型 |
| total\_price | 总价 （万） |
| court | 小区名字 |
| direction | 朝向 |
| area | 面积 （平方米） |
| title | 房天下关于此房源的标题 |
| add | 房源地址 |
| which\_floor | 楼层 |
| structure | 楼层结构 |
| down\_payment | 参考首付 |
| decoration | 装修 |
| url | 房源房天下网址 |
| property\_right | 产权性质 |
| author | 发布房源者 |

**数据组三**

这是一组西城、东城和海淀区义务教育入学学区划分数据，数据内容包含小学全称、所属学区、学校分级、对应划片小区。数据通过网页脚本抓取，人工输入两种方式录入数据库，数据来源于上述三区的教委、招生中心官方网站。时间跨度为2013年至2016年，其中缺少西城、海淀的2013年数据和东城2015、2016年数据。

|  |  |  |
| --- | --- | --- |
| **数据文件名** | **数据数量** | **简介** |
| ♥数据准备-JIACHUAN-0418.xlsx | 2016：368  2015：332  2014：407  2013：34 | 学区、学校名称、划片小区整理 |

从学区数量的整理上看，上述三个区在2014、2015年分别做过学区划片调整，从原有的大划片，细化出多个学区划片。