**链家二手房数据初步分析**

修改内容以红色标出

1. 背景

一直想抓取链家的数据对北京的暴涨式房价做个初步分析，这次本来想把整个流程走完，其中提到的流程包括：数据爬取，数据清洗，数据分析，数据可视化。但数据爬取部分由于时间原因未由自己完成，使用了一个已知数据集，<https://zhuanlan.zhihu.com/p/25132058>，数据收集于大约2017/02/28，在此谢谢这位大牛的数据供给。该数据集的特征数目不算多，但可以从中看出一些信息。下次希望用自己爬取的数据来优化分析，其中最重要的就是增多爬取特征，增多数据样本，现在延庆、怀柔以及密云数据量较少，不清楚是本身交易量较少还是数据爬取不足。

1. 简要介绍

链家是一个房屋交易平台，上面记录了每个在售或者在租房屋的具体信息。该数据集中包含了以下几个特征：房屋所在城区，房屋所在小区，房屋类型，总房价，每平米单价，房屋面积，房屋关注度，房屋线下带看人数，房屋上架时间以及房屋经纬度。这几个特征非常直观，在此不做详细介绍。

1. 具体执行

* 数据清洗

数据清洗用python完成，具体见lianjia-data.ipynb。主要做几点工作：

* 将部分特征从object类型转换为number类型
* 将部分离散变量特征转换为category类型
* 将上架时间统一单位
* 将车位信息移除
* 将房屋类型为别墅的房屋面积信息进行修正
* Tableau数据可视化

作品链接：

V01: <https://public.tableau.com/profile/menglin4237#!/vizhome/lianjia_v01/LianjiaStory?publish=yes>

V03:

<https://public.tableau.com/profile/menglin4237#!/vizhome/lianjia_v03/Insight7?publish=yes>

V04:

<https://public.tableau.com/profile/menglin4237#!/vizhome/lianjia_v04/LianjiaStory?publish=yes>

拿到数据后，开始思考，我想知道什么？我能从数据里面得到这些答案吗？这些问题在数据可视化的过程中不断迭代。

* 这些在售房屋在地里位置上的分布情况是？

通过对城区着色，我们能很清楚的看出各个城区的地理位置分布情况。可以看出，中心地带房屋密集，周边地带房屋稀疏。

* 每平米房价在地理位置上是如果分布的？

因为有经纬度信息，所以很容易实现这点可视化。从可视化结果来看，符合我们预期，每平米房屋单价中心最高，周边最低，呈圆环形辐射开来。其中，西城，东城，海淀，朝阳以及丰台簇拥形成中心。

* 城区的平均每平米房价排序情况？

应用了Tableau的群集功能，将各个城区根据每平米单价做了简单划分。Tableau的群集划分功能使用的是K-means算法。从可视化结果来看，西城，东城以及海淀稳居第一梯队。符合预期，西城、东城是老北京城区，大量学区房聚集于此。海淀也是学区房聚集地。

* 各个城区的房源数如何？

一二梯队中海淀，朝阳和丰台房源较多。第三梯队中昌平房源较多。从中可以看出，昌平较其它的周边城区更具有居住属性。昌平与海淀和朝阳接壤，有相对的地理优势，是外地人买房的不错选择。

* 每个城区的TopN小区是哪些？有什么特点？

我们重点关注面积小的老城区，如果他们的单价较高，可能预示的周边有学区。由于这个数据集里面没有房屋年份，所以部分信息需要通过搜索获取。数据分析一定需对业务有相当充分的了解，这样才能对数据作出很好的解释。

* 整个北京的TopN小区是哪些？有什么特点？

我们按单价筛选出了全北京的前150名小区，不出意料，全部集中在了西城，东城，海淀以及朝阳这几个中心城区。其中，西城点较为密集，不愧是北京的老城区，配备大量学区；东城的点相对分散；海淀房源呈现明显的聚集性，说明周边配备学区；朝阳点较少，但同样呈现出一定的聚集性，说明红玺台附近配有学区。

* 各个城区的户型分布情况？

从整个可视化结果来看，各个城区的房型配比没有特别显著的差异。最为常见的房型为两室一厅，因为该房型可以满足基本的生活需求。如果综合考虑，两室占比最多，三室次之，继而一室，其它占比较少。密云、怀柔和延庆房源采集数较少，所采集的数据大多为大居室。

* 房屋单价会随房屋面积增长而降低吗？

我们知道，房屋单价受地理位置的影响很大，由于这种强特征存在，我们必须确定地理位置和房屋面积没有强相关性。除了做计算之外，我们从可视化结果来看，似乎没有明显的相关性。我们重点研究北京几大城区，西城、东城、海淀、朝阳以及丰台，可以看出，100平左右是单价凹点。为了彻底剔除地理位置影响，我们选取西城荣丰小区来进一步分析，发现100平左右确实是单价凹点。

* 哪些城区最受关注？

关于关注这一点，有两个指标：一个是关注度，表征有多少人关注此房源；一个是线下带看数，表征有多少人实际看过此房源。房屋上架时间越长，下线带看数可能会越多。线下带看的购房者的买房欲望更明显更迫切，所以我们主要参考线下带看数。从可视化结果可以看出，周边城区中门头沟，石景山以及昌平较为受欢迎，这几个区域的共同特点是紧邻海淀。从房价的中位数来看，昌平房屋价格较高，石景山次只，门头沟最低，这样给不同预算的购房者提供了相对广泛的选择空间。然后我们再重点观察关注度，发现房山和密云的关注度相对较高，这是什么原因造成的呢？

* 北京目前的购买力如何？是否仍是供不应求的状态？

如何衡量购买力？在这里我们用简单的房屋总价来衡量。从可视化结果来看，房屋总价呈现长尾特征。我们将600万作为划分节点（至于为什么选600万，目前是自己的主观态度，因为发现大量的刚需购房者将房买在了门头沟、石景山以及昌平，这几个城区的中位数价格在500万以下），我们发现，600万以下的房屋记录数仅仅是600万以上的两倍左右，也就是说，仍然存在大量的高价房源。然后我们来看这些房源的竞争程度如何。如何衡量竞争程度，简单来说，就是衡量目前出售房屋的数量与购房者人数的比值。比值越小，说明供不应求的程度越强。我们做一个简单的推导，这里取线下带看数并不是很准确。

房屋记录数＝总线下带看数／平均线下带看数

购房者人数＝总线下带看数／平均每人线下看房次数

房屋记录数／购房者人数 ＝ 平均每人线下看房次数／平均线下带看数

为什么做此计算，因为每个人在购房时不只看一套房。我们举个例子，小区中有五套房，每套房被线下带看过五次，问有多少个购房者？这个问题怎么算呢？可以是25个人，每个人看了一套房；可以是5个人，每个人看了所有的5套房。。。所以，到底有多少个购房者，取决于平均每人线下看房次数。这个值目前不是很清楚，但可以从lianjia网站上提取，但根据目前的实际经验来看，应该在15以内。

我们从可视化结果来看，所有房屋平均线下带看数在15以上，说明目前仍是卖方市场。而且，600万以上房源的平均线下带看数相对较高，说明可能学区房市场仍旧竞争激烈。北京高价位房源的线下带看数同样很高，说明北京聚集了大量了有钱人，太牛了。

* 房屋一般多久被出售？未被及时出售的房屋有什么特点？

从可视化结果来看，半年前上架的房屋占比较少，说明大部分房源在半年内完成交易。并且，从两个月以前房屋数量陡然下降来看，优质房源一般会在两个月内完成交易。未及时被出售的房屋的平均面积相对较大。

1. 设计准则

* 单价从中心向周边自高到低呈圆环形辐射开来

为了体现整个北京的房屋单价的分布情况，我们采用heatmap的形式。而且我们知道，北京的价格分布与城区资源丰富与否密切相关，所以将北京的城区分布图置于左上角，方便观察单价与城区的关系。

* 西城、东城及海淀稳居单价第一梯队。昌平房源相对较多

我们采用直方图对城区单价进行排序，这样可以直观看出各个城区平均单价情况。但同样用BoxPlot对图进行内容补充，说明整个城区单价的浮动情况。我们用Tableau的聚类功能根据单价对城区进行聚类。加入平均线对评价单价进行说明。我们知道，地理位置是房价的一个强相关特征，所以，如果可以，城区分布图应该放在每张展示图上，随时提醒客户。

* 北京的TopN小区位于四大城区

通过交互式方式，可以动态筛选出北京的TopN小区来观察最高房价区域。

* 各城区房型差别不大，二居室配备最为丰富

通过用百分比来表现各城区的房型分布情况。

* 100平左右可能为四大城区的房屋单价局部凹点

由于地理位置是单价的强相关特征，所以通过可视化来判断地理位置与房屋大小是否相关。用直方图来绘制每平米单价波动情况。加入交互式面积筛选器，从而筛选出重点关注的房屋范围内的单价波动情况。

* 周边地区中门头沟、石景山和昌平最受关注

线下带看数和关注度通过直方图表示，可以看出购房者重点关注城区。购房者在关注时，除了地理位置，另一个重要因素应该就是预算了，所以将总房价中值用双轴的形式绘制于右侧，这样可以很好的表征出各城区总房价中值情况。

* 北京房屋仍处于供不应求的状态，富人聚集

对房屋总价绘制直方图。用600万房价作分界来评估购房者的经济实力。采用标注对重点表达对象进行说明。

* 大部分房源在半年内完成交易，优质房源缩短为两个月

采用扇形图说明半年内和其它的比例关系。采用直方图表现各个月份房源数和平均房屋面积的差异。

1. 反馈

* 第一位审阅人
* 城区颜色表述不清晰
* 城区是片，为什么是点呢？
* 记录数改为房源数
* 怀柔和密云为什么这么贵？解释异常值
* 横纵坐标单位
* 第二位审阅人（Reviewer）
* 仪表版尺寸较小
* 只保留故事，将其它工作表和仪表板隐藏
* 4个群集未说明划分标准
* 地理位置和房屋大小没有明显相关性无法看出
* 单个样本说服力度不强
* 房价中位数表示不清晰，容易给客户带来困惑
* 建议剔除数据量较少的城区
* 最后一个图表过于单一

1. 修改历史

* V01 –> V02
* Story中的说明框直接写出结论，简单直白。
* 添加说明，简要概括结论
* V02 -> V03 (基于第一个审阅人的修改)
* 城区颜色表述不清晰

城区颜色这个问题，当时就很纠结，都列出来感觉很冗余，目前的解决方案是以多列的形式列出来。其实，最好是在城区分布图上直接显示出来，但是发现显示不出来。

* 城区是片，为什么是点呢？

这个问题，最好是把用点绘制的外轮廓勾勒出来，但是不是太会。

* 记录数改为房源数

已改

* 怀柔和密云为什么这么贵？解释异常值

通过加入点标注解释了异常值

* 横纵坐标单位

对坐标轴进行了重新命名和单位标注

* 修改了房屋百分比构图
* V03 -> V04 (基于reviewer的修改)
* 仪表版尺寸较小

已经改成自动。

* 只保留故事，将其它工作表和仪表板隐藏

已修改。

* 4个群集未说明划分标准

已经添加说明。

* 地理位置和房屋大小没有明显相关性无法看出

当时同样觉得不是很明显，但老师提的意见似乎也不能够完全表达我的意思。如果仅对城区做条形图，还是无法说明问题。我是想排除同一个城区内地理位置与房屋大小有强相关性，所以只对城区绘图无法说明城区内建房情况。

1. 放大绘图，这次提交采取的措施
2. 采用别的可视化方式来说明房屋大小与地理位置关系不大

* 单个样本说服力度不强

确实从统计上没有意义，将其删除。

* 房价中位数表示不清晰，容易给客户带来困惑

换成了双轴的形势，如果不修改，确实会给客户带来困惑。以后多注意可视化图表的制作。

* 建议剔除数据量较少的城区

我的理解可能稍微有点不一致，由于数据来源于第三方，所以不太确定是二手房出售数量在延庆、怀柔以及密云本身就少，还是抓取者抓取不充分导致。数据样本不充分必然会导致对总体的分析不准确。但如果二手房数量在这几个区本来就有限，说明总体本身有限，那即使样本数量少，也可以得出一些信息。将来自己会亲自采集数据，会对这部分有更强的概念。

* 最后一个图表过于单一

加入扇形图说明。同时加入另一个insight，发现滞留房屋的平均面积会相对比较大。

1. 计划

由于数据集采用的是别人的，很多想要的信息没有爬下来。时间有限，下次希望有机会自己多爬取一些信息。毕竟项目只有更好，没有最好。

1. 致谢

* 数据集发布者：<https://zhuanlan.zhihu.com/p/25132058>
* Tableau官方资源：<https://www.tableau.com/zh-cn/solutions/gallery>