链家二手房数据初步分析

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。

2. 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。

3. 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

     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，将各个城区做了简单划分。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？
  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，从三个月以前房屋数量陡然下降来看，优质房源一般会在三个月内完成交易。

4. 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
5. 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，请教老师。

- 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，请教老师。
- 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

6. 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。
7. 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