飘向北方

——Airbnb 北京民宿数据分析

第 5 小组：张川，王琦、邓文、黄义捷

一、 背景介绍与任务目标

我们身处在共享时代，卷着一股共享的潮流，从街头的共享单车，再到共享书店、共享篮球，共享经济已经成为我们这个时代闪闪放光的标志。但你知道，开启了一个辉煌时代的共享经济“老祖宗”是谁吗?

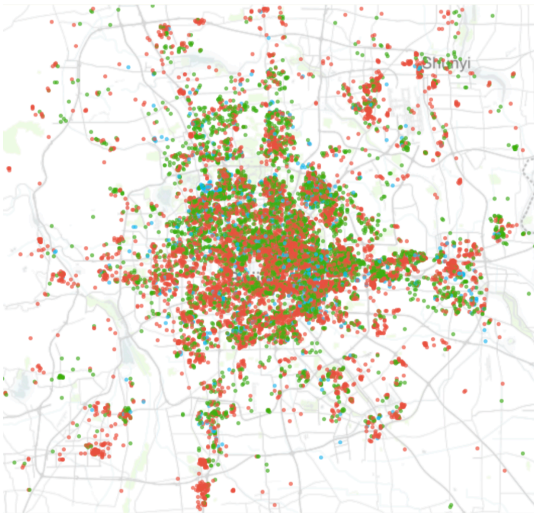
2008 年，以为个人房主和租客牵线搭桥为服务宗旨的 Airbnb（中文名：爱彼迎）网站在美国旧金山成立，拉开了互联网加持下共享经济新时代的大幕。Airbnb 重塑了酒店行业，因为有了它，你可以从个人的手中租住一间房屋，而不是从一家酒店中租住。此后，民宿行业急速发展：2017 年，中国在线民宿市场规模达到 126 亿元人民币，预计到 2020 年，这一数字将超过 300 亿。根据 Airbnb 官方统计，到 2020 年，中国市场将成长为其全球最大的客源国。

民宿给了市场一个惊喜，也给了我们旅途中便捷、多样的选择。从青旅、公寓、别墅、城堡、帐篷、房车到树屋，Airbnb 为旅行者们提供数以百万计的独特入住选择，你可以在这里找到几乎所有你想要的住房体验。因此，民宿已成为广大旅行者的优先选项。

而我们的主人公，一位计划在暑期前往北京短住几天的穷学生，正在为此苦恼：怎样才能在北京挑选到便宜又便捷的民宿呢？于是我们决定，让数据亲自说话，替他答疑解惑。相信我们的研究，对于未来将要出行的各位，也都具有普遍的参考价值。

二、 数据说明

我们使用的数据集爬取自 Airbnb，其中包括 31457 条遍布北京市全境的 Airbnb 民宿信息，数据截止至 2019 年 2 月 11 日。每条数据中的标题字段包含大量中、英文文本信息，我们对其做了切词处理（如图 1），并根据高频词汇人工归纳出了“是否临近地铁”、“是否临近学校”、“是否临近景点”、“是否临近商圈”共 4 个哑变量和“房源特色“这个分类变量。



(a) 样本点地理分布

图 1 (b) 标题文本词云

变量类型	字段	类型	示例	说明
因变量	价格	连续变量	500	[35,19999] 元
原始自变量	最少入住天数	类别变量	非 1 天	1 天、非一天
	一年可租天数	连续变量	329	[1,365] 天
	评论数、月均评论数	连续变量	90，0.85	[0,333] 条，[0,55] 条/月
	房型	类别变量	整个房源	整房、单间、共享单间
	经度、纬度	连续变量	39.89503	
	区域	类别变量	朝阳区	覆盖北京全部 16 个区
	临近地铁、临近学校、	类别变量	1	1 代表是，0 代表不是

提取自变量	临近景点、临近商圈			
	房源特色	类别变量	装修风格	装修风格、设计布局、价格优惠、娱乐设施、可做饭、其他

三、 描述性分析

1、 民宿价格：

首先我们对因变量——民宿的价格进行频数分布的统计。结果如下：

•有两间民宿价格为 0，有十余套超过两万高价房源，经过查看后发现为**测试房源或早已下架并未出售**，因此我们将其**删除**；

• 绝大多数民宿的价格在 200~700 元之间，分类时也以 700 元为界；

• 其中有少数几套**小于 50** 元的民宿，我们发现它们都为**青年旅舍**之类，仅提供床位；价格最高的是一处名为“古北水镇首排观景小叠墅”的整套民宿，高达 19999 元一晚，是名副其实的“土豪房”。

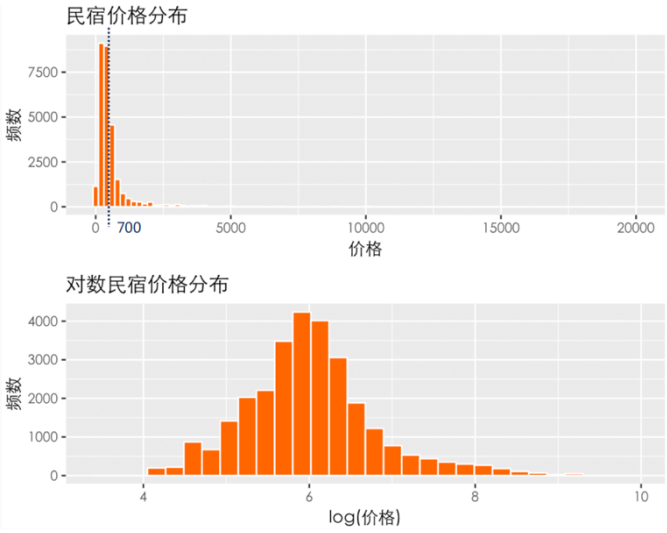


图 2. 民宿价格分布

2、 民宿区域对价格的影响

• 朝阳区的民宿众多，有 10183 处，远超其他区域——朝阳区有很多 CBD 以及酒吧街等，如三里屯、国贸等；

• **东城区和西城区的民宿价格的平均水平明显高于其他区**——这两处是北京二环以内的中心地带；

•海淀区民宿价格平均较低——海淀主要是大学城，消费水平不高；

• 每个区都存在价格很高的离群点——别墅，四合院等

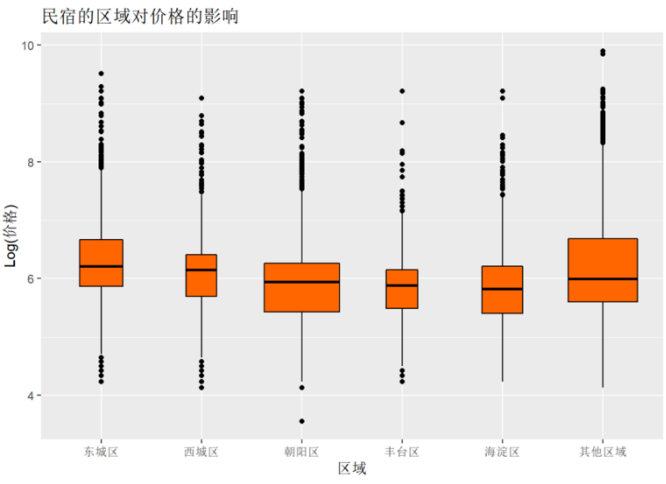


图 3. 区域对价格影响箱线图

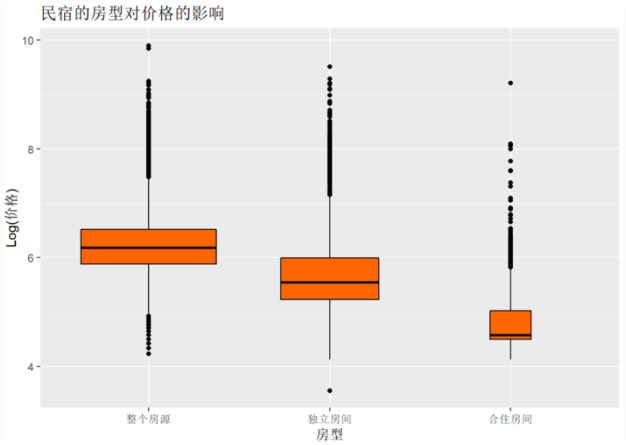


图 4.　房型对价格影响箱线图

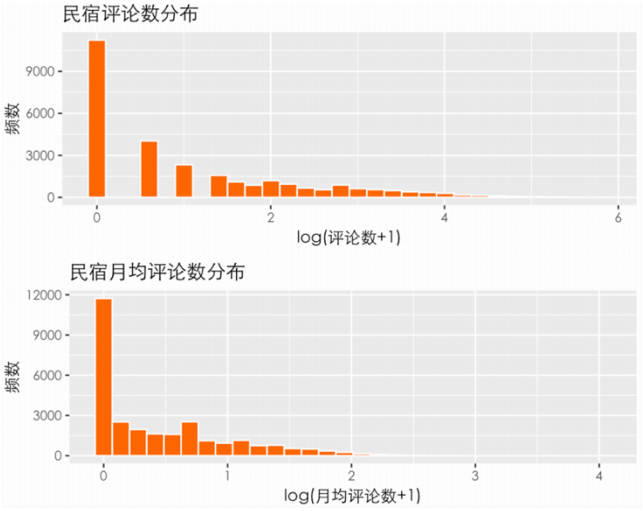


图 5.　民宿评论数分布

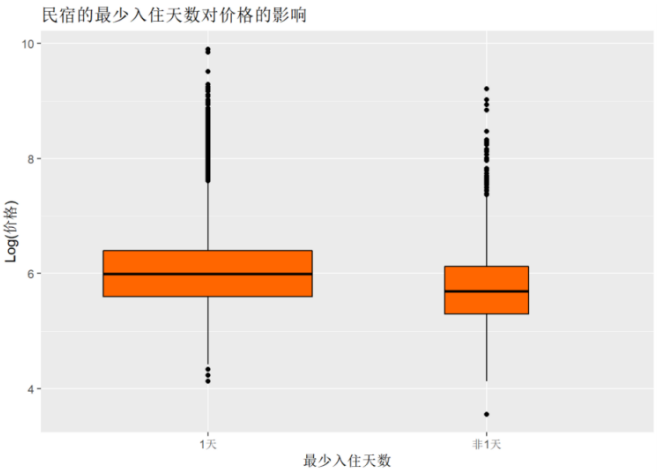


图 6.　最少入住天数对价格影响箱线图

3、民宿房型对价格的影响：

•整个房源的类型最多，有 17757 套，独立房间次之，合住房间最少；

•三种出租方式的均价差异明显：整房出租均价约 450 元，独立房间均价约 244 元，合住房间均价约 110 元——整房出租主要以民居为主，合住房间主要以青旅为主，二者目标群体不同；

4、民宿评论数分布：

• 尽管民宿众多，但**评论数和月均评论数却很少**，这与“大部分人都倾向于住完民宿后就不会再去评论”的实际情况符合；

• 评论数最多的是名为“梵溪故宫旁日式花园庭院禅意卧房”的独立房间，有 333 条评论，其价格为 581 元一晚；

• 月均评论数最多的是一套紫禁城附近的，名为“Hutong Studio”的整间民宿，月均 55 条评论，价格为 1051 元一晚。

5、民宿的最少入住天数对价格的影响：

•最少入住天数为 1 天的房源占总数的 86%，因此将超过 1 天的都归为一类；

• **最少入住天数为 1 天的民宿价格要明显高于非 1 天的民宿**——这可能是因为当住的天数多了，房东会相应的优惠一点价格。

6、民宿一年可租天数的分布：

•可租天数主要集中在三个时间：3 个月左右，6 个月左右，1 年左右

•绝大部分的民宿都是长年出租的，可租天数在 350 天以上，这种可能是有闲置房产的房东；

•一部分在 80~95 天，应该一季一季地出租；一部分在 170 天左右，应该是半年半年地出租。这两种可能是房东暂时外出或季节性地使用该处房产。

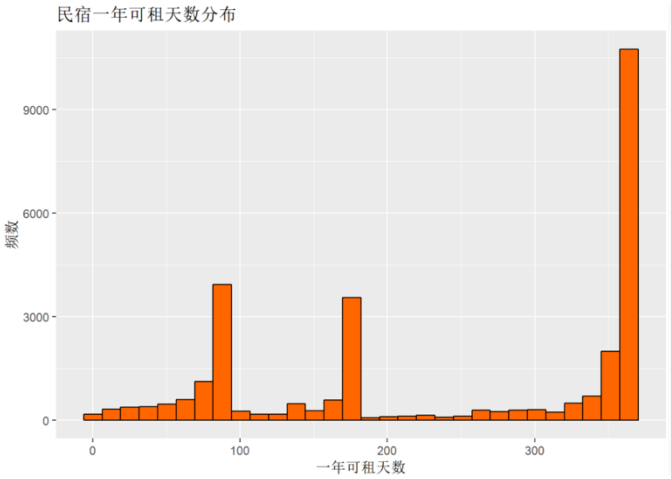


图 7.　一年可租天数分布

7、民宿临近地对价格的影响：

•临近地铁的民宿反而平均价格较低，查看数据发现，这主要是因为实惠的民宿更倾向于宣传自己的便捷性，同时最贵的四合院、别墅、景区房等基本也都不在地铁沿线；

•临近学校的民宿平均价格较低，我们发现学区房单间出租比例较高，而单间出租的民宿价格普遍较低；

•临近景点的民宿平均价格较高，这与实际情况相符合，景区消费水平较高，同时也是民宿的主要市场；

• 是否临近商圈对民宿价格的影响并不明显

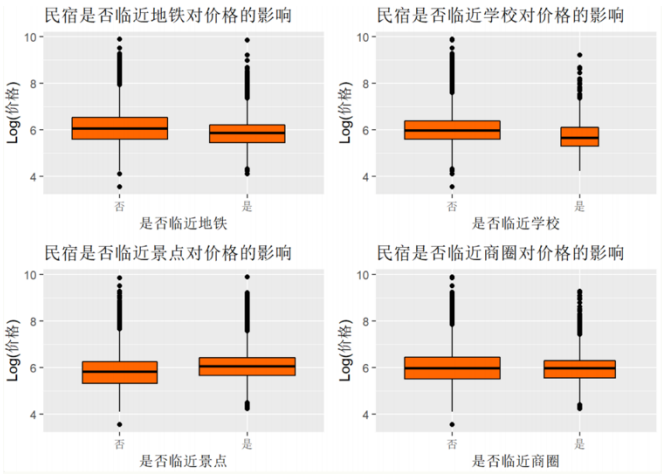


图 8.　是否临近地铁、学校、景点、商圈对价格的影响

8、民宿的房源特色对价格的影响：

•大部分民宿的特色以装修设计为主，其次是以设施为主，将其他数量较小的民宿特色归为“其他”；

• 设施类均价稍高一点，但总体来说差别不大

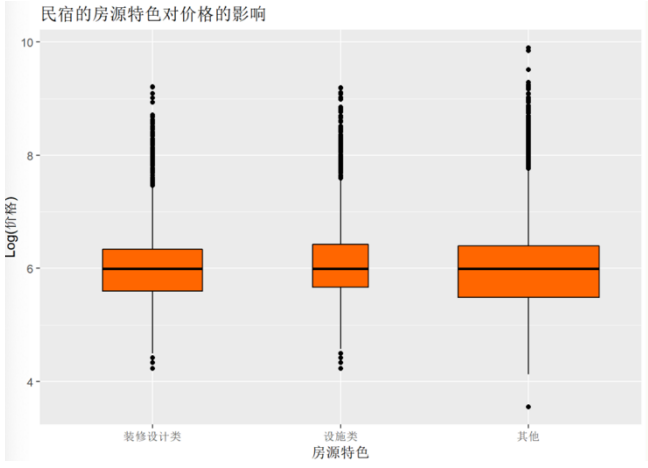


图 9.　房源特色对价格的影响

四、 回归分析与模型选择

A. 回归模型

1. 简单线性回归

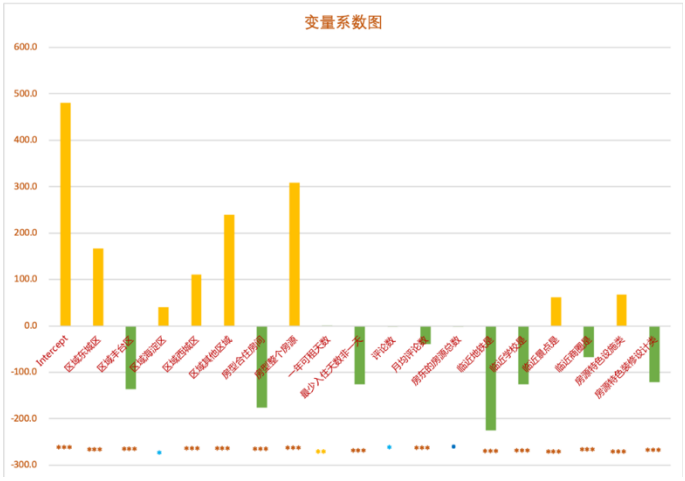


图 10. 简单线性回归变量系数图

1. 朝阳区作为基准，只有丰台区均价更低；
2. 独立房间作为基准，整个房源更贵，而合住房间更便宜；
3. 最少入住一天为基准，非 1 天的更便宜；
4. 评论越多价格越低；
5. 临近景点对价格正加成，地铁、学校、商圈都是负加成；
6. 房源特色其他类为基准，设施类更贵，装修设计类更便宜。

采用 10 折交叉验证，得到 MSE=788.67。从 full model 的结果中发现房东的房源总数不太显著，去除后重新建模，得到 MSE=788.70，发现结果并没有得到改善。

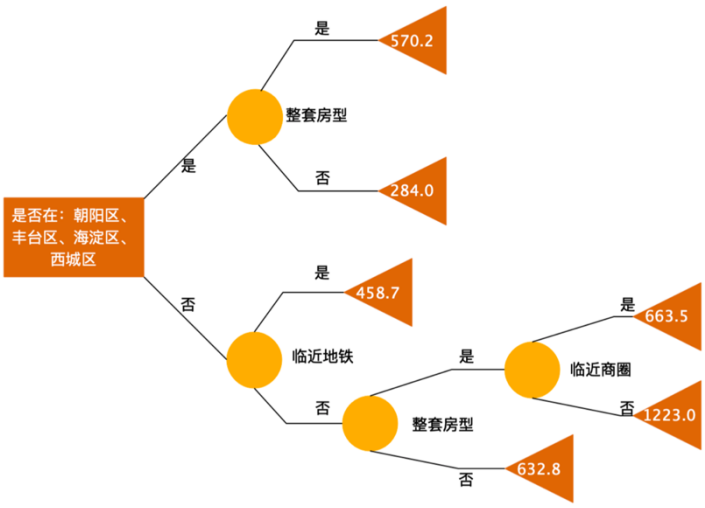


图 11. 回归树模型

2. 回归树

经剪枝操作后得到如下回归树：

1. 区域、房型、是否临近地铁在回归树模型中的作用很大。
2. 最贵的房子的特点：不在朝阳、海淀、丰台、西城+无地铁+整套房+无商圈。
3. 最便宜房子的特点：在朝阳、海淀、丰台、西城+不是整套房出租。
4. 使价格下降的因素：不是整套房+靠近地铁+靠近商圈。

3. **随机森林**

重要变量依次是一年可组天数、房东的房源总数、月评论数、区域、房型等。

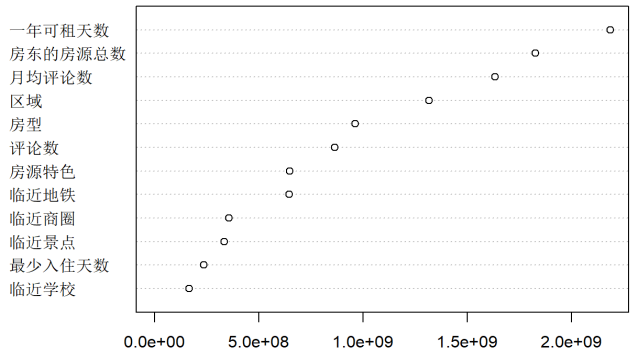


图 12. 随机森林模型变量重要性图

4. **回归模型结果比较**

随机森林在三个回归模型中的拟合效果最好，但效果不够理想。

模型	MSE
线性回归	788.7
回归树	790.3
随机森林	744.5

表 1. 回归模型结果比较

B. 分类模型

1. 分类依据

1. 选择价格时，游客往往会在一个大致范围内选择，而不是选择精确价格。
2. 价格大于 700 元时，数量骤减；
3. 查询发现，北京四星级酒店一般在 600+，绝大多数游客出行首选三星级舒适型酒店，部分能接受四星级酒店的价格，极少数人会选择住五星级酒店，因此认为若民宿价格超过四星级酒店将会对游客的选择产生不小的影响；

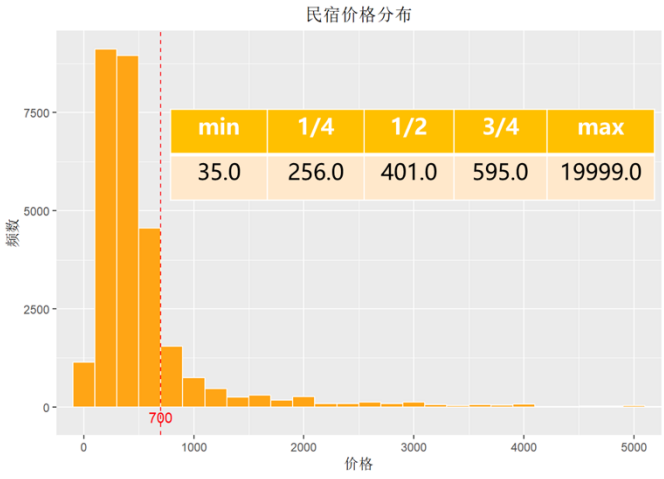


图 13. 民宿价格分布

综合考量，价格高于 700 元的民宿设为“高”，这样设置价格高、低的民宿比例大约为 1:5，比较合理。

2. 分类模型结果比较

从结果上看，随机森林的正确率和 AUC 值最大，预测效果最佳，但其在“高价格”上的预测效果不佳；在“高价格”预测上表现最好的是 QDA，但它在“低价格”预测上表现

比较差；分类树模型虽然正确率比较高，尤其是预测“低价格”的正确率极高，但这是它将所有预测为“低”导致的。

模型	accuracy	AUC	TPR	TNR
逻辑回归	0.830	0.777	0.083	0.983
LDA	0.827	0.773	0.023	0.992
QDA	0.655	0.781	**0.792**	0.627
朴素贝叶斯	0.815	0.758	0.427	0.895
kNN	0.842	0.730	0.311	0.950
分类树	0.830	0.737	**0.000**	**1.000**
随机森林	**0.862**	**0.827**	0.338	0.969
Adaboost	0.845	0.816	0.298	0.957

表 2. 分类模型结果比较

五、 聚类分析

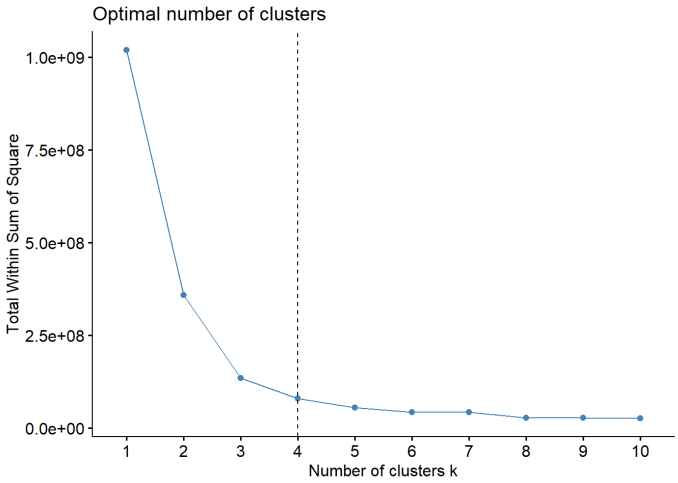


图 14. 崖底碎石图

1. 碎石图分析

首先，我们利用碎石图进行类别数量的确定。从图中可以看出，碎石图的斜率在**类别数为 3** 时变化最为明显。因此，我们选择将所有的房源聚为 3 类。

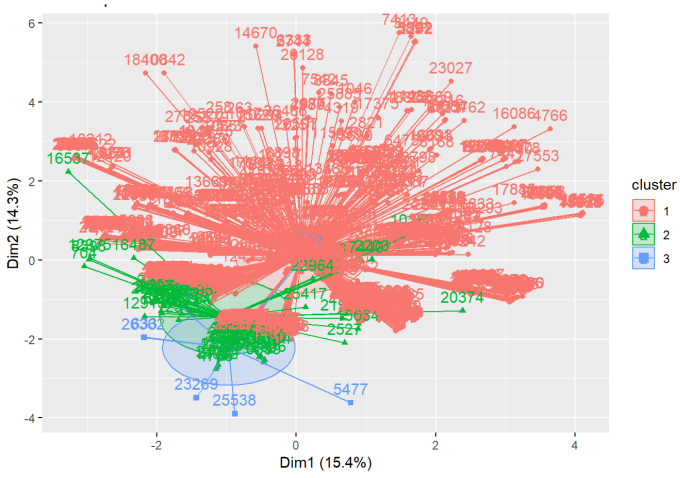


图 15. K-means 聚类结果

2. k-means 聚类分析

利用 k-means 聚类法，获得的聚类结果。从图中可以看出，k-means 法将所有房源聚为三类，**绝大部分的房源都被归于第一类。**同时，第一类与第二类、第三类的距离**都较远，**证明第一类与其它两类的**差异较大。**

3. 层次聚类分析

我们继续来看看层次聚类的结果。

从层次聚类的结果可以更明显地看出，第一类的房源**占据了绝大部分**，而且第二类与第三类的距离**显著近于**第一类。

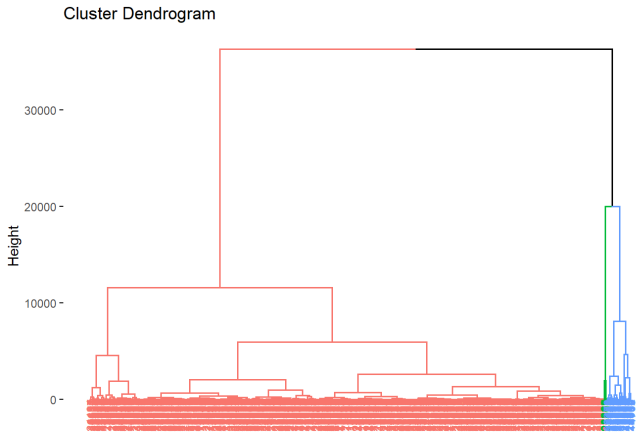


图 16.　层次聚类结果

4. 聚类数据解读

1）从对聚类数据特征的观察表明，聚类特征间的差异最主要体现于**价格**，第一类、第二类、第三类房源的平均价格分别**大幅递增**。

2）第一类房源**临近地铁**的比例**显著更高**。

3）第一类房源与第三类房源在**临近学校**与**临近商圈**的比例相近，而第二类房源在这两点的表现都**显著更差**。

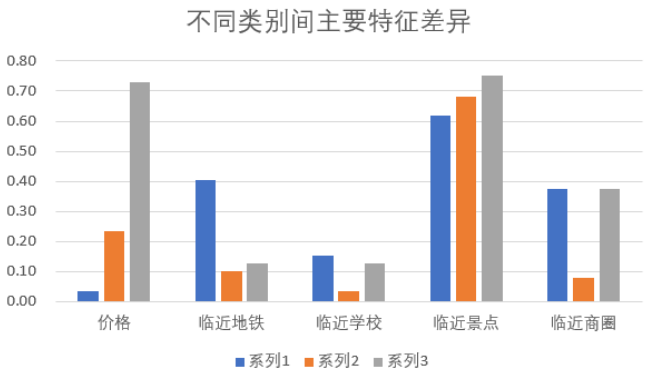


图 17.　聚类特征

4）根据以上现象，我们将第一类房源归为“经济大众房”，在价格上显著更低，平均仅为 431 元，而且在诸多便利设施方面都有优势；第二类房源归为“坑爹宰客房”，不仅价格大幅高于第一类房，达到了 2864 元，而且在各方面都无特别优势，仅离景区比第一类房更近；第三类房则归为“土豪奢华房”，在临近地铁以外的各个方面都具有优势，然而价格也是显著飙升，高达 8918 元。

5. 地理位置分析

在地图上对三类房分别进行位置分析，可以发现第一类房占绝大多数，从市中心二环内到郊区均有广泛分布；第二类房主要集中于主干道以及景区位置，分布同样分散；第三类房则主要集中于景区核心位置与二环内的城市核心。这一结果与我们的三类房差异分析与总结相符。

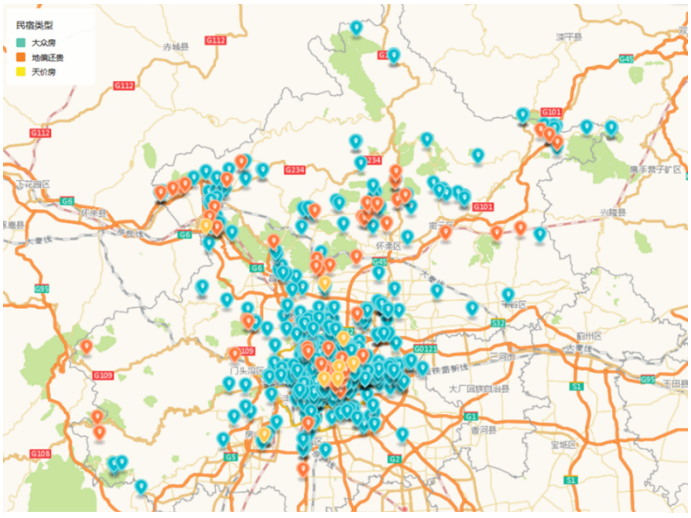


图 18.　聚类结果地理分布

六、 建议与总结

1、 首先我们发现，北京的民宿数量众多、分布广泛、各有特色，无论你是想要体验京城文化还是像我们的主人公一样想要搜寻到便宜的房源，民宿都是合适的选择；

2、 考虑民宿的价格，绝大部分民宿价格在 200～700 元之间，最便宜的有 50 元的青旅床位，最贵的也有 19999 元的整栋别墅；

3、 我们发现影响民宿价格的主要因素包括区域（东城区、西城区明显偏高）、房型（整房最贵，合住房间最便宜）、是否临近地铁（临近的反而普遍便宜）、是否临近学校（学校附近的普遍便宜）、是否临近景区（临近景区的明显偏高）、房东拥有的房源数、评论数、一年可租天数、最少入住天数等，其中最重要的变量是一年可租天数、房东房源数和评论数，最不重要的变量是是否临近学校；

4、 应用回归模型预测民宿价格，在线性回归、回归树和随机森林三个模型中，随机森林的效果最好，但 MSE 依然高达 700 以上；

5、 结合实际逻辑，以 700 元为界，将价格分为“低”“高”两档。尝试多种分类模型：整体上，随机森林的 AUC 最低，但在预测“高”价的准确率上，QDA 远优于其他模型；

6、 对房源信息做聚类，我们发现北京的民宿房源明显可归为三类：第一类“经济大众房”，便宜又方便，特点是大多位于学校和地铁旁，单间房源比例较高；第二类“宰客坑爹房”，价格又高，交通又不方便，离景区也不是很近；第三类“土豪奢华房”，除了价格高、离地铁稍远，其他方面都很完美，适合享受宝贵假期的多金人士。

7、 综上，推荐咱们的主人公大胆选择地铁旁、学校旁的单间房源，这样的房源既实惠又方便，而且往往评论数多，可供参考的信息也比较丰富，相信是学生党的不二选择～