北京林业大学

“统计软件”课程论文

（普通高等教育）

题　　目	房价分析与建模
学　　院	理学院
专业名称	数学与应用数学
班　　级	数学 191
学　　号	191102124
姓　　名	

（2021-2022 学年第一学期）

摘要

随着大数据和互联网技术的高速发展,房地产行业的数据越来越多,如何从复杂多样的特征中得到房价的预测模型是非常重要的。本文通过探索性数据分析，得到了影响房价的六类主要特征，包括社区组，纬度，经度，房屋类型，最短租房时间因素。结合国内外的相关的房价预测模型，选择出了线性回归、CART 决策树、GBDT 算法进行预测。将回归模型进行一元到多元回归分析，CART 决策树进行剪纸，GBDT 算法进行交叉验证，得到了三种不同的效果达最优的模型。

其中 GBDT 算法能够更好的适应不平衡的数据集,同时也更不容易过拟合,泛化能力较好,在很多非线性的回归问题上也有良好的表现,适用范围很广泛。结果表明，在这三种模型中，GBDT 的效果最优。

关键词：EDA；线性回归；决策树；boosting

Abstract

With the rapid development of big data and Internet technology, there are more and more data in the real estate industry. It is very important how to obtain the prediction model of housing price from the complex and diverse characteristics. Through exploratory data analysis, this paper obtains six main characteristics that affect housing prices, including community group, latitude, longitude, housing type, and the shortest rental time factor. Combined with relevant domestic and foreign housing price prediction models, linear regression, CART decision tree, and GBDT algorithm are selected for prediction. The regression model was subjected to unary to multiple regression analysis, the CART decision tree was paper-cut, and the GBDT algorithm was cross-validated, and three different models with optimal effects were obtained.

Among them, GBDT algorithm can better adapt to unbalanced data sets, and it is also less prone to overfitting. It has good generalization ability. It also has good performance on many nonlinear regression problems and has a wide range of applications. The results show that among these three models, GBDT has the best effect.

Key Words: EDA ;linear regression; decision tree; boosting

目录

摘要 ..... **I**

Abstract ..... **II**

1 绪论 **1**

2 数据处理 **2**

2.1 变量提取与分析 2

2.2 变量详情 3

2.2.1 定性变量 3

2.2.2 定量变量 6

2.3 探索性数据分析 10

2.3.1 其他变量与价格变量 10

2.3.2 数据标准化 15

2.3.3 地理位置分析 17

3 模型构建 **19**

3.1 数据分割 19

3.2 线性回归模型 19

3.2.1 一元线性回归 20

3.2.2 多元线性回归 21

3.3 决策树 23

3.4 梯度提升树（boosting） 26

3.5 模型比较 27

参考文献 **28**

1 绪论

近年来随着共享经济的迅猛发展，在线短租成为人们出行的最佳选择之一[1]。而对于业主来说，如何确定合适的房价以吸引更多租户，成为了一个重要的问题。Airbnb 即爱彼迎，它是一个旅行房屋租赁社区，用户可通过网络或手机应用程序发布、搜索度假房屋租赁信息并完成在线预定程序。据官网显示以及媒体报，其社区平台在 191 个国家、65,000 个城市为旅行者们提供数以百万计的独特入住选择，不管是公寓、别墅、城堡还是树屋[2]。Airbnb 业主面临的一个挑战 是确定最优租金。在许多地区，租房者可以通过价格、卧室数量、房间类型等标准进行筛选。由于 Airbnb 是一个市场，房东的收费最终取决于市场价格。本文的数据集主要来源于 Inside Airbnb（IA）于 2021 年 9 月 1 号在它网站上公开的纽约市在 2010 到 2021 年间的上市活动和指标，网页链接：http://insideairbnb.com/get-the-data.html。 线性回归是数理统计中的一种统计分析方法，需要给出训练数据的分类标识，是机器学习系统的典型构成。有着建模速度快、可根据系数给出每个变量的解释、对异常值敏感三个有点优点。多元线性回归分析是指包括两个及以上自变量，且因变量和自变量满足线性关系[3]。决策树因其形状像树且又能用于决策故被称为决策树,是通过机器学习,从一系列无秩序、无规则的逻辑关系中推理出一套分层规则,将结局按照概率分布的树形图表达,从而进行精确预测或正确分类[4]。GBDT(Gradient Boosting Decision Tree) 又叫 MART（Multiple Additive Regression Tree)，是一种迭代的决策树算法，该算法由多棵决策树组成，所有树的结论累加起来做最终答案[5]。本文应用上述三个模型进行房价预测，并进行模型比较，得到效果最好的模型。

2 数据处理

2.1 变量提取与分析

在原始数据集中，共有：id（编号）, name（租户名称）, host_id（户主编号）, host_name（户主名称）, neighbourhood_group（社区组）, neighbourhood（社区）, latitude（纬度）, longitude（经度）, room_type（房屋类型）, price（价格）, minimum_nights（最短租房时间）, number_of_reviews（评论的数量）, last_review（最新评论时间）, reviews_per_month（平均一个月评论数）, calculated_host_listings_count（同一房东拥有的公寓总数）, availability_365（365 天内该房屋可以被预订的天数）, number_of_reviews_ltm（过去 12 个月的评论数量），license（许可证）这 18 个变量。本文从这 18 各变量中选择了“社区组”、“社区”、“纬度”、“经度”、“房屋类型”、“价格”、“最短租房时间”进行分析，并将它们单独存入 CleanData.csv 中。共有 7 个变量，37713 组观察值。数据概要见表 2.1。

表 2.1 数据概要

	社区组	社区	纬度	经度	房屋类型	价格	最短租房时间
1	Manhattan	Midtown	40.75	-73.99	Entire home/apt	150	30
2	Brooklyn	Bedford-Stuyvesant	40.68	-73.96	Entire home/apt	73	1
3	Brooklyn	Bedford-Stuyvesant	40.69	-73.96	Private room	60	30
4	Brooklyn	Sunset Park	40.66	-73.99	Entire home/apt	275	5
5	Manhattan	Midtown	40.76	-73.98	Private room	68	2
6	Manhattan	Upper West Side	40.80	-73.97	Private room	75	2

表 2.2 显示了一些基本数据类型。

表 2.2 基本数据类型展示

'data.frame'	37713 obs. of 7 variables
$ 社区组	Factor w/ 5 levels "Bronx","Brooklyn",..: 3 2 2 2 3 3 2 2 3 3 ...
$ 社区	Factor w/ 222 levels "Allerton","Arden Heights",..: 130 13 13 193 130 205 185 217 63 63 ...
$ 纬度	num 40.8 40.7 40.7 40.7 40.8 ...
$ 经度	num -74 -74 -74 -74 -74 ...
$ 房屋类型	Factor w/ 4 levels "Entire home/apt",..: 1 1 3 1 3 3 3 1 3 3 ...
$ 价格	int 150 73 60 275 68 75 98 89 65 62 ...
$ 最短租房时间	int 30 1 30 5 2 2 4 30 30 30 ...

由表 2.2，数据集共有 37713 组观察值，7 个变量。因子型变量社区组共有 5 个水平，因子型变量房屋类型共有 4 个水平，社区型变量也为因子型，共 222 个水平。纬度、经度均为数值型变量，价格、

最短租房时间则均为整数型变量。

对于数据的描述性统计见表 2.3 和表 2.4

表 2.3 定性变量描述性统计

社区组		社区		房屋类型	
名称	数量	名称	数量	名称	数量
Manhattan	16625	Bedford-Stuyvesant	2724	Entire home/apt	20062
Brooklyn	14510	Williamsburg	2596	Hotel room	246
Queens	5178	Harlem	1960	Private room	16829
Bronx	1059	Hell's Kitchen	1678	Shared room	576
Staten Island	339	(Other)	28754		

根据表 2.3，社区组共有 5 组，Manhattan 占比最多，Staten Island 占比最少。房屋类型共有 Entire home/apt（套房家庭酒店），Hotel room（酒店），Private room（私人房屋），Shared room（共享房屋）这四种，套房家庭酒店最多，酒店最少。在此图中无法比较各个社区的占比大小。

表 2.4　定量变量描述性统计

	纬度	经度	价格	最短租房时间
最小值	40.505	-74.250	0．00	1.000
25%中位数	40.689	-73.984	69．00	3.000
中位数	40.725	-73.955	110．00	30.000
均值	40.729	-73.949	165．43	22.131
75%中位数	40.763	-73.931	180．00	30.000
最大值	40.914	-73.711	10000.00	1250.000
方差	0.003098	0.002544	85477.959	949.224339
标准差	0.055662	0.050441	292.36614	30.809485
偏度	0.179671	1.266703	18.913585	12.402241
峰度	0.183836	3.949107	541.33082	298.280321

由表 2.4，价格的均值（165．43）比中位数（110．00）高出一些，说明数据具有右偏趋势，偏度为 18.913585。峰度较大为 541.33082，判定数据相对于正态分布而言陡峭一些。方差（85477.959）较大，说明数据波动较大。最短租房时间的均值（22.131）比中位数（30.000）低，偏度为 12.402241 大于 0，说明数据右偏，但偏移程度较价格比小。峰度为 298.280321，判定数据相对于正态分布而言陡峭。方差为 949.224339，数据有波动，但比价格的波动小。对于经度和纬度，它们最小值、中位数、均值、最大值近似，且方差、标准差、偏度、峰度的数值都比其他变量的数值小，故它们分布情况均匀。

2.2 变量详情

2.2.1 定性变量

（1）社区

频率分布，是指在统计分组的基础上，将总体中各单位按组归类整理，按一定顺序排列，形成的总体中各单位在各组间的分布[7]。表 2.5 给出了占比前十的社区的频数和频率。从中可知占比最多的

Bedford-Stuyvesant 为 2724 个，但是仅仅只占所有社区中的 7.22%。利用 describe 函数查看社区，发现它共有 222 组。

表 2.5 排名前十社区

	频数	频率
East Village	1194	3.1660170
Crown Heights	1200	3.1819267
Upper East Side	1358	3.6008803
Upper West Side	1451	3.8474796
Midtown	1604	4,2531753
Bushwick	1674	4.4387877
Hell’s Kitchen	1678	4.4493941
Harlem	1960	5.1971469
Williamsburg	2596	6.8835680
Bedford-Stuyvesant	2724	7.2229735

进行可视化处理。

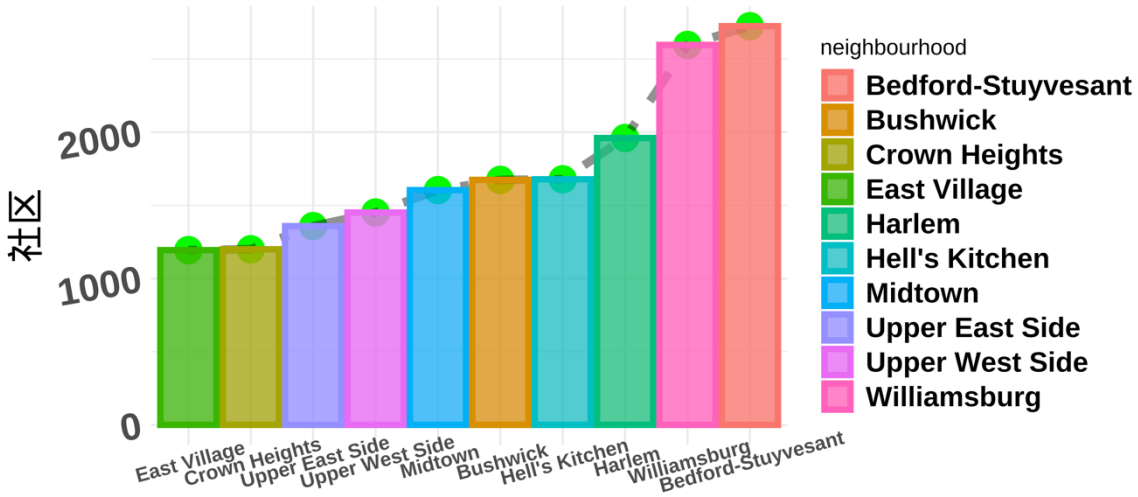

图 2.1 占比最多的前十个社区可视化

显然，仅仅是在这最多的十个社区中，最大和最小的频数还是有明显差别。

（2）社区组

表 2.6 社区组频率频数分布

Value	Manhattan	Brooklyn	Queens	Bronx	Staten Island
Frequency	16625	14512	5178	1059	339
Proportion	0.441	0.385	0.137	0.028	0.009

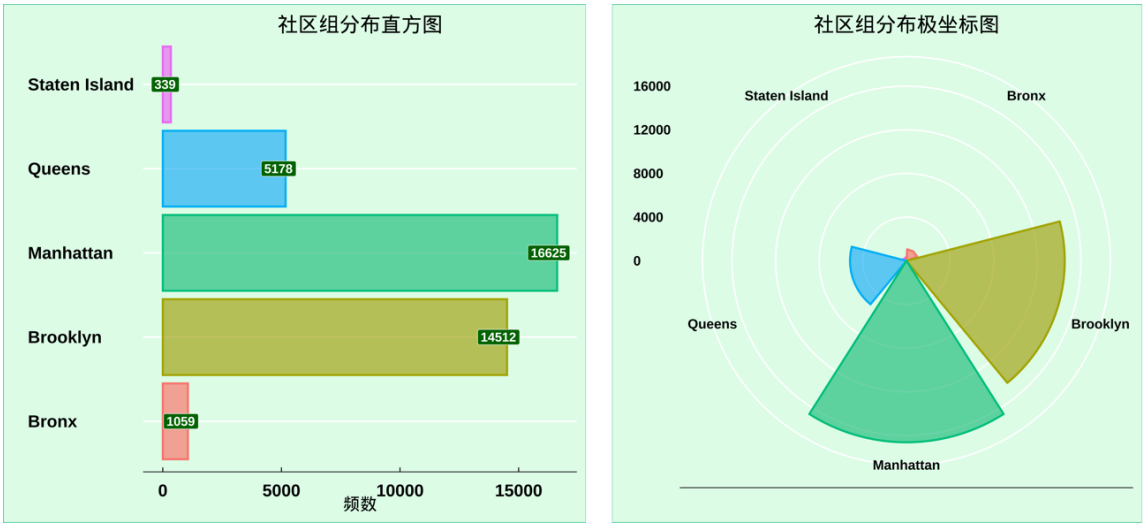

图 2.2 社区组频数分布直方图与极坐标图

由表 2.6 和图 2.2 得到了不同社区组的频数和频率。极坐标图将它们的占比可视化。观察发现，对于 Staten Island 和 Bronx 社区组来说，数据已经很少，如果再往下细分到具体社区，可能会出现数据量过少的情况，在数据分割时出现数据不平衡，从而影响建模效果。所以，不会将具体社区纳入建模的变量。

（3）房屋类型

表 2.7 房屋类型频率频数分布

Value	Entire home/apt	Hotel room	Shared room	Private room
Frequency	20062	16829	576	246
Proportion	0.532	0.446	0.015	0.007

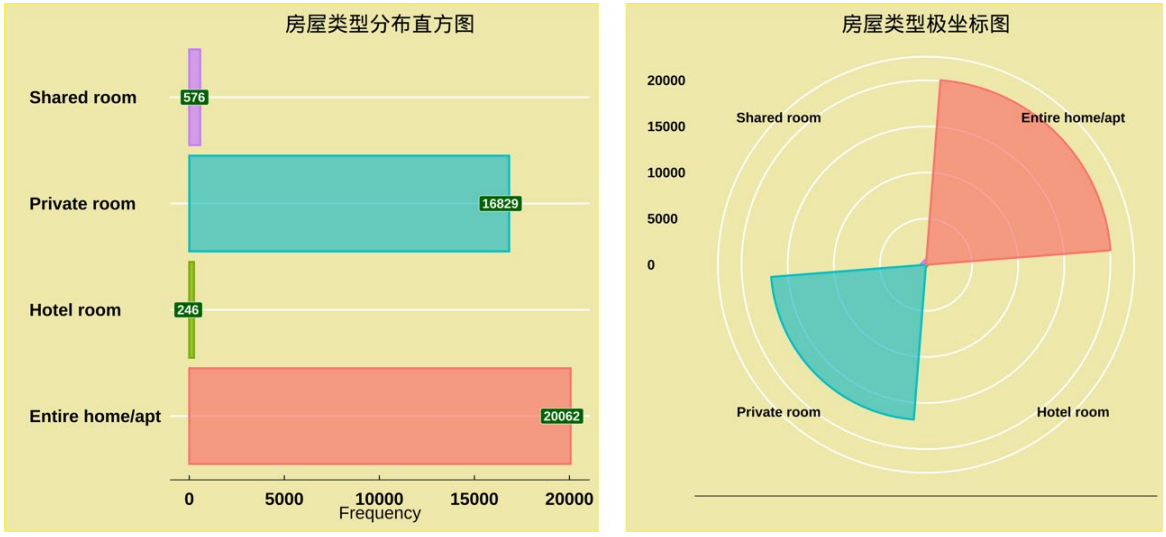

图 2.3 房屋类型频数分布直方图与极坐标图

由表 2.7 和图 2.3 得到了不同房屋类型的频数和频率。极坐标图将它们的占比可视化。发现部分房屋类型之间有明显差别。

2.2.2 定量变量

（1）最短租房时间

Sturges 在 1926 年在直方图制作方法上作出了开创性的工作，得到了分组数k关于样本量n的粗略关系。其主要思想是用对称的二项分布（p = 0.5）来近似正态分布[6]：

①考虑理想化的直方图，设其分组为k。

②假设：每一样本观测值落在直方图第i个区间中的概率近似服从概率$p=0.5$的二项分布，则第i组的样本频数平均为：

$$n = C_{k-1}^{0} + C_{k-1}^{1} + \cdots + C_{k-1}^{k-1} = 2^{k-1}$$

③将上式两边取对数即得到分组数k关于样本量n的 Sturges 公式：

$$k = 1 + \log_2 n$$

在本文中：

$$n = 37713$$

$$k = 1 + \log_2 37713 \approx 16$$

表 2.8 最短租房时间频率频数分布表

	Frequency	Percent
1 —| 78.125	36948	97.9715217564
78.125 —| 156.250	522	1.3841380956
156.250 —| 234.375	95	0.2519025270
234.375 —| 312.500	36	0.0954577997
312.500 —| 390.625	97	0.2572057381
390.625 —| 468.750	2	0.0053032111
468.750 —| 546.875	6	0.0159096333
546.875 —| 625.000	0	0.0000000000
625.000 —| 703.125	0	0.0000000000
703.125 —| 781.250	0	0.0000000000
781.250 —| 859.375	0	0.0000000000
859.375 —| 937.500	0	0.0000000000
937.500 —| 1015.625	5	0.0132580277
1015.625 —| 1093.750	0	0.0000000000
1093.750 —| 1171.875	1	0.0026516055
1171.875 —| 1250	1	0.0026516055

由表 2.8 发现最短租房时间在 1～78.125 的分布比较集中，约占总数的 97.97%，因此特别画出 0-40 内的最短租房时间频数分布直方图。

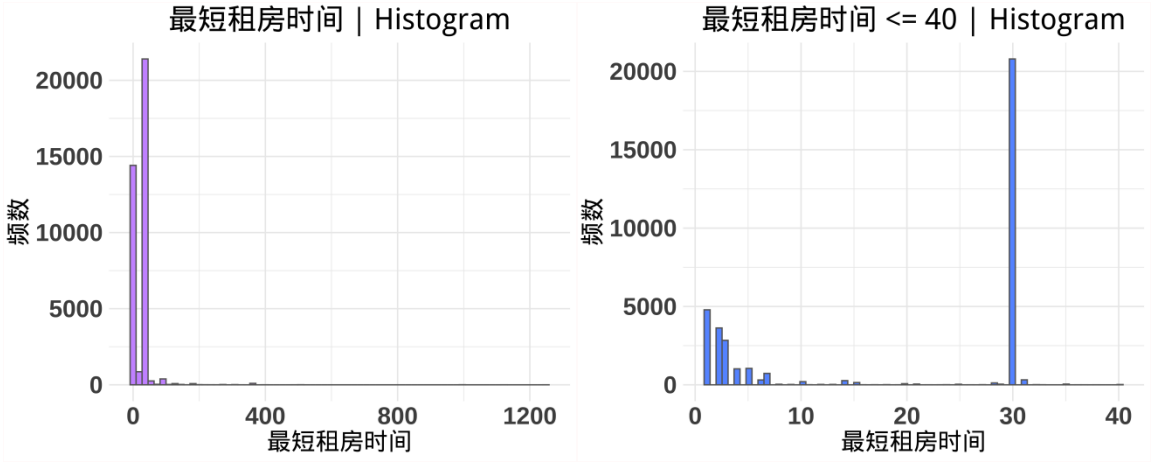

图 2.4 最短租房时间频数分布直方图

观察图 2.4 发现，最短租房时间并不是标准的偏态分布，它们中一大部分集中在 30 这个时间上。

（2）价格

表 2.9 是利用 Sturges 公式得到的对价格的分组：

表 2.9 价格频率频数分布表

	Frequency	Percent
0 —| 625.000	36814	97.6162066131
625.000 —| 1250.00	680	1.8030917721
1250.00 —| 1875.00	109	0.2890250046
1875.00 —| 2500.00	49	0.1299286718
2500.00 —| 3125.00	21	0.0556837165
3125.00 —| 3750.00	9	0.0238644499
3750.00 —| 4375.00	4	0.0106064222
4375.00 —| 5000.00	3	0.0079548166
5000.00 —| 5625.00	1	0.0026516055
5625.00 —| 6250.00	3	0.0079548166
6250.00 —| 6875.00	3	0.0079548166
6875.00 —| 7500.00	3	0.0079548166
7500.00 —| 8125.00	0	0.0000000000
8125.00 —| 8750.00	0	0.0000000000
8750.00 —| 9375.00	1	0.0026516055
9375.00 —| 10000.0	13	0.0344708721

由上面的分组画出频数直方图，发现价格在 0-1000 的分布比较集中，因此特别画出 0-1000 间的价格分布频数直方图。

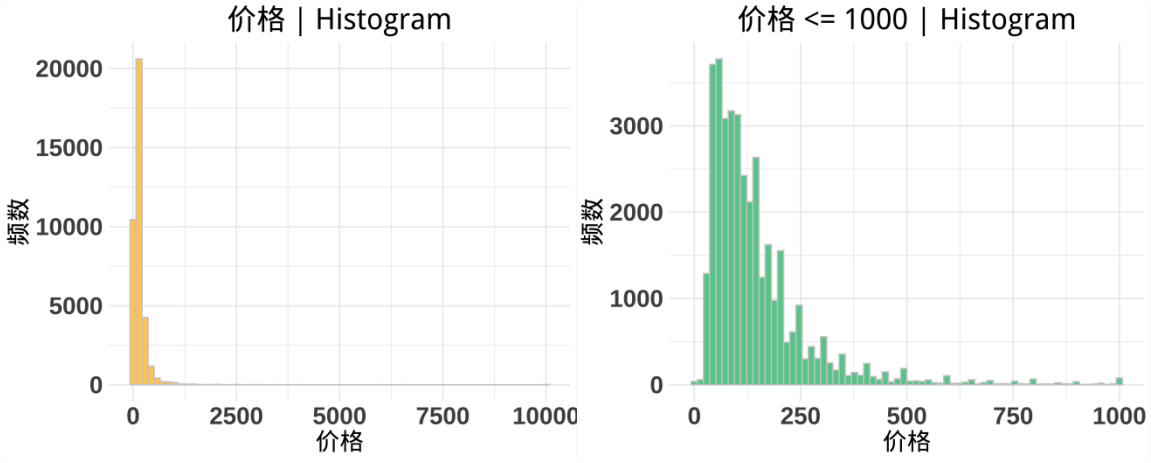

图 2.5 价格频数分布直方图

由图 2.5，价格呈偏右态分布。

（3）价格与最短租房时间

图 2.6 用众数，中位数，平均数来表现价格和最短租房时间的数据的集中趋势。

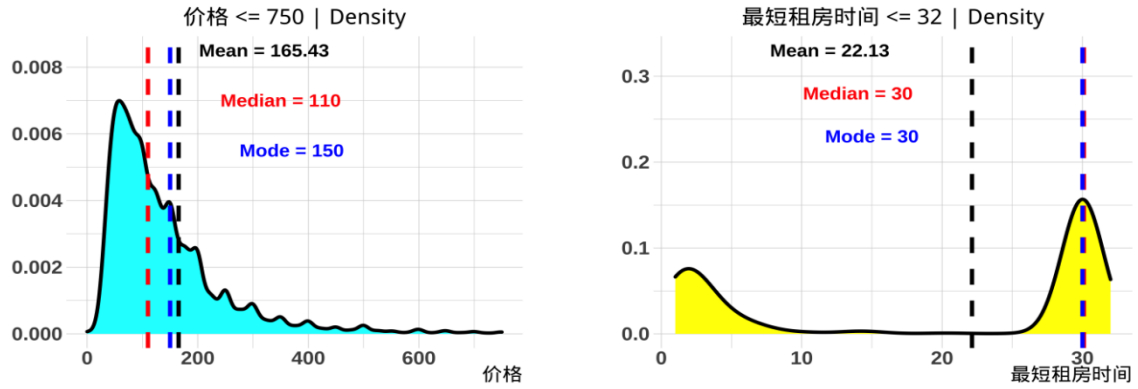

图 2.6 价格和最短租房时间的密度曲线

由上图可看出在小于等于 750 的范围内，价格的平均数是 165.43，中位数是 110，众数是 150；而最短租房时间受在 30 的人群的影响，中位数和众数都是 30，平均值为 22.13。

对于连续型变量，分离度量是将有序的数据序列划分为包含序列中相同数量元素的部分的实数。在描述性统计中，四分位数是将一组有序的数据分成四个相等部分的三个值中的任何一个，因此每一部分代表样本或总体的 1/4：

$$P_k = \frac{N_k}{100}$$

其中P_k为第 k 个百分比，N_k为第 k 个区间的样本量。

表 2.10 四分位与百分位混合结果。

	价格	最短租房时间
25.00%	69	3
50.00%	110	30
75.00%	180	30
99.00%	999	91

在对单独测量的分析中，我们获取了一些有用的信息：

（1）25.00%在 airbnb 上的预订价值等于或少于 69 美元，且最少三个晚上；

（2）在 airbnb 上，50.00%的预订是价值等于或少于 110 美元，最少 30 晚；

（3）75.00%在 airbnb 上的预订价值等于或少于 180 美元，且至少 30 晚；

（4）在 airbnb 上，99.00%的预订价值等于或少于 999 美元和 91 晚，这意味着只有 1.0%的预订价值超过 999 美元，而价钱的最大值为为 10000 美元、最短租房时间为 1250 晚。

可以看出最短租房时间和价格在 99.00%以上有非常奇怪的值，为了获得关于这些值的更多信息并做出决策，需要进行更精确的调查，以验证这些上限为 10000 美元和最低 1250 个夜晚的值是否确实是离群值。

小提琴图是一种绘制数值数据的方法。它类似于箱线图，只是在每一边都增加了一个旋转的核密度图，图 2.7 为我们展示了价格和最短租房时间的小提琴图。

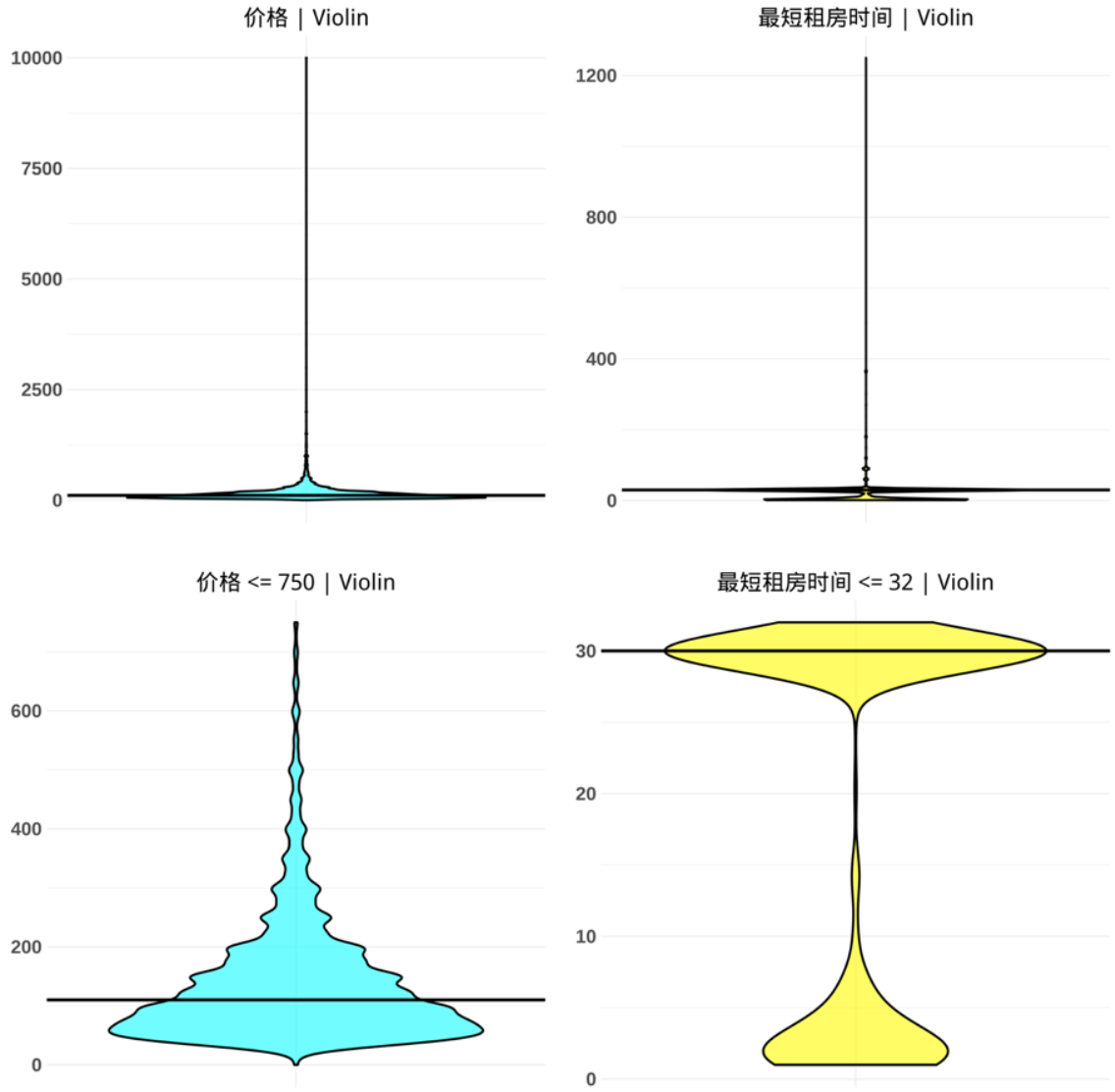

图 2.7 价格和最短租房时间的小提琴图

通过小提琴图并结合图 2.7，可以明显地看出，最短租房时间属于双模分布，因为它具有两个峰

值。而价格的小提琴图表现出它比正态分布更加偏斜。由小提琴图可以判断出，10000 美元的价格和 1250 的最短租房时间确实偏离群体。

（4）经度和纬度

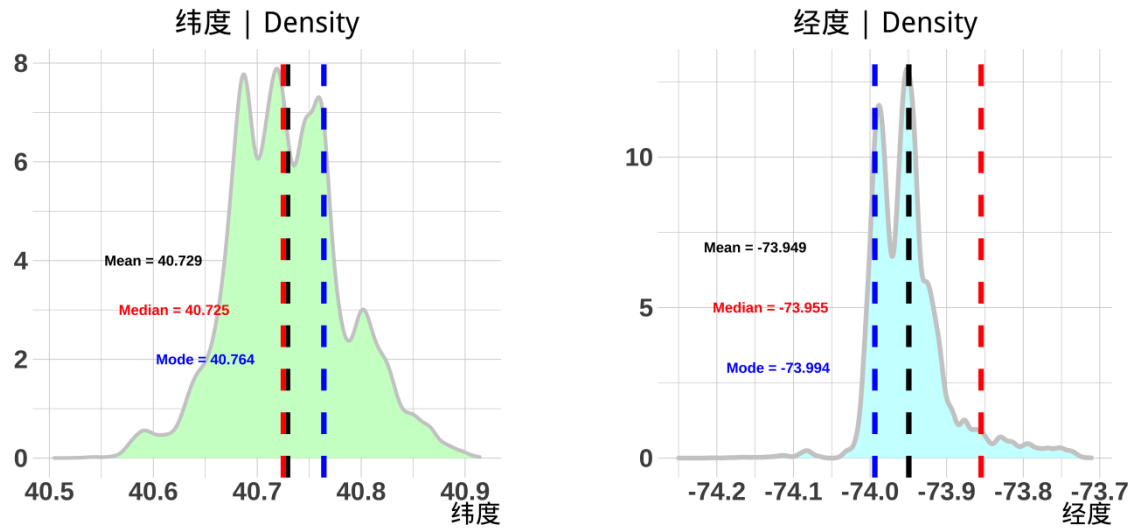

图 2.8 经度和纬度的密度曲线

由图 2.8，经度纬度的分布近似正态分布，比较均匀。

2.3 探索性数据分析

2.3.1 其他变量与价格变量

（1）最短租房时间与价格

为了研究最短租房时间与价格的关系，画出散点图，进行一元回归分析。

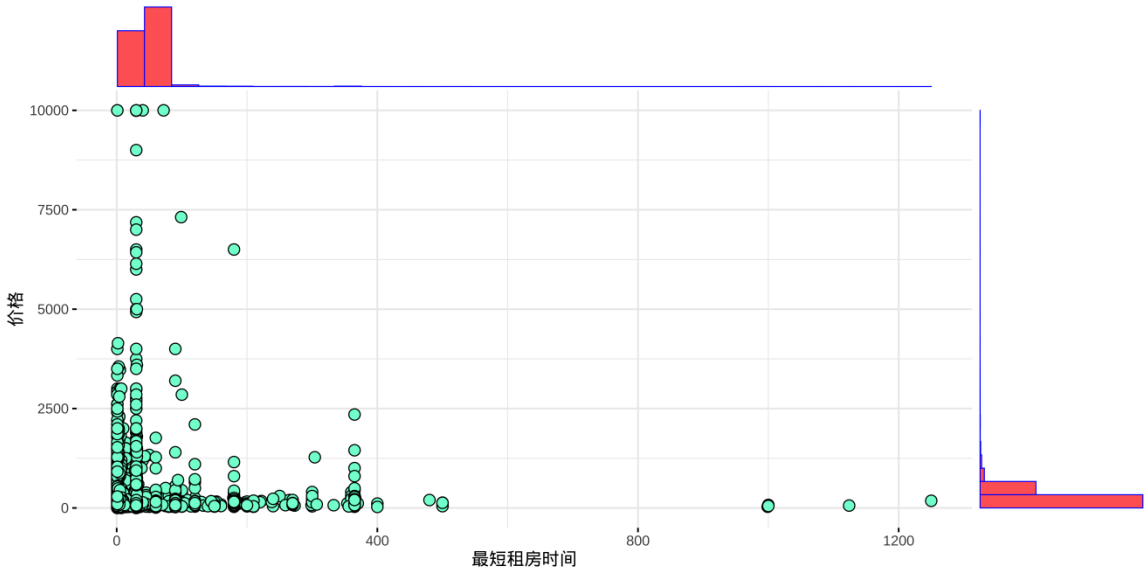

图 2.9 价格和最短租房时间二维散点图

显然，最短租房时间与价格呈现较明显的负相关关系。设定价格为因变量，最短租房时间为自变

量，进行一元线性回归。

表 2.11 价格与最短租房时间一元线性回归结果

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	169.768441	1.853297	< 2.2e-16 ***
最短租房时间	-0.196182	0.048856	5.943e-05 ***
Multiple R-squared: 0.0004274,		Adjusted R-squared: 0.00040089	
F-statistic: 16.124 on 1 and 37711 DF		p-value: 5.9428e-05	

观察发现，p 值小于 0.05，说明回归系数显著，但 R 方较小，说明拟合效果不好。最短租房时间的系数小于 0，也说明与价格呈现负相关关系。

（2）经度、纬度与价格

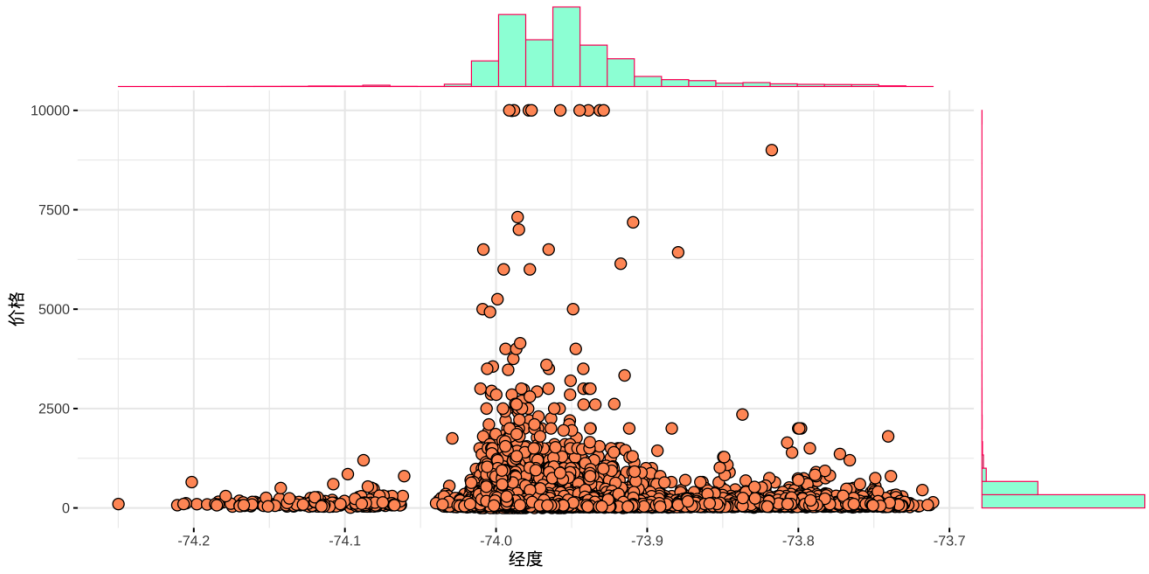

图 2.10 经度与价格

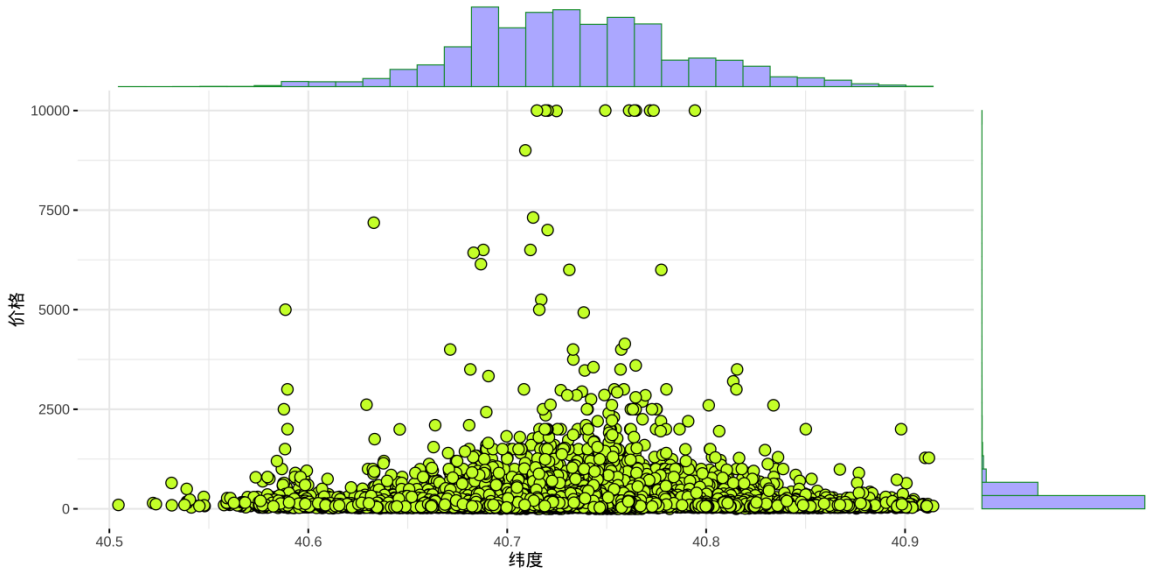

图 2.11 纬度与价格

可以看出，经度和纬度是呈现正态分布的，而没有做处理的价格则为偏态分布。

(3) 房屋类型与价格

表 2.12 房间类型的均价与占比

房间类型	平均价格	百分比
Hotel room	312.8862	41.03310
Entire home/apt	217.0543	28.46535
Shared room	129.6302	17.00021
Private room	102.9506	13.50134

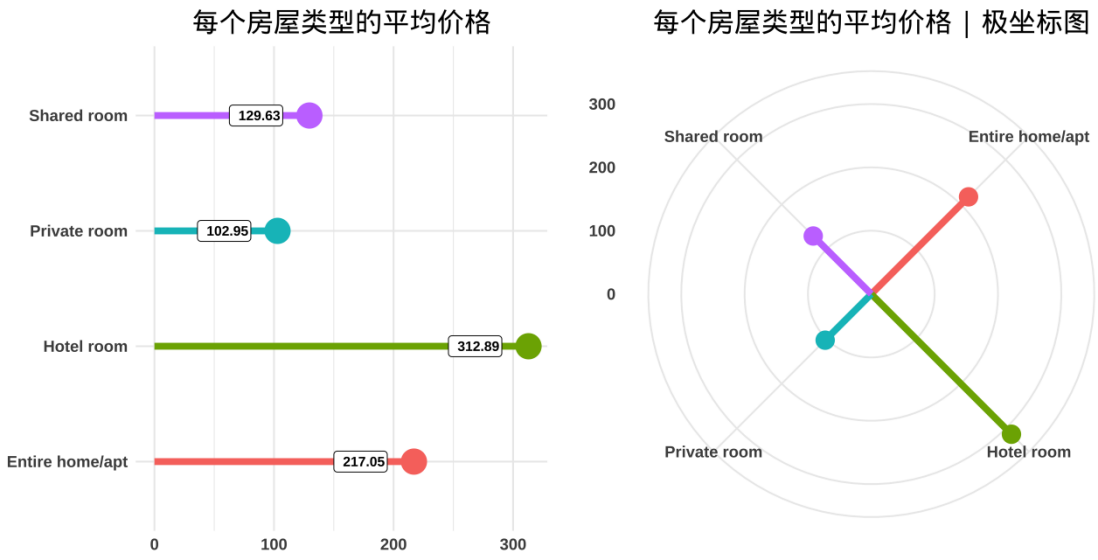

图 2.12 房屋类型的均价占比可视化

由表 2.12 和图 2.12，Hotel room 的平均预订价格约为 312.9 美元，占所有类型房间的 41.03%，是最贵的。Hotel room 的价格比 Entire home/apt 贵 12.56%，比 Shared room 贵 24.03%，比 Private room 贵 27.53%。其次是 Entire home/apt，平均预订价格约为 217.1 美元，占总价格的 28.47%。比 Hotel room 低 12.56%，分别比 Shared room 和 Private room 贵 11.47%和 14.97%。然后是 Shared room，平均预订价格在 103.0 美元左右，占所有类型房间的 13.50%。它的平均价格比 Private room 高 3.5%。最后是 Private room，平均预订价格在 103.0 美元左右，占所有类型房间的 13.50%，它的价格是最低的。

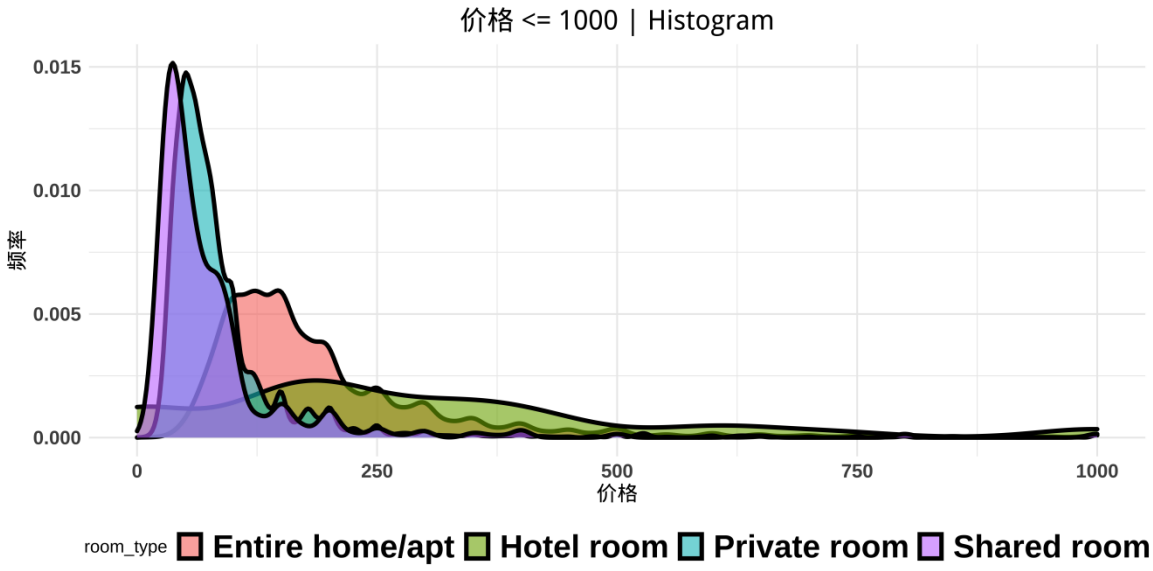

图 2.13 不同房屋类型的密度曲线

图 2.13 展示了不同房屋类型的密度曲线，可以看出这四个房屋类型都具有右偏趋势。

（4）社区与价格

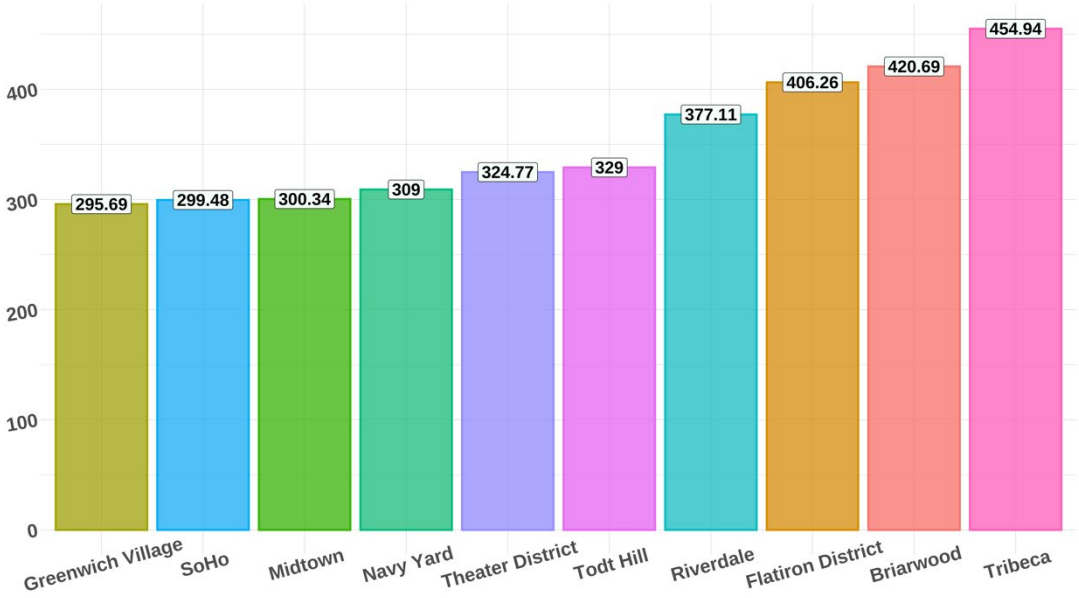

图 2.14 平均价格最高的前十组社区直方图

图 2.14 展示了价格最高的前十组社区的平均价格。

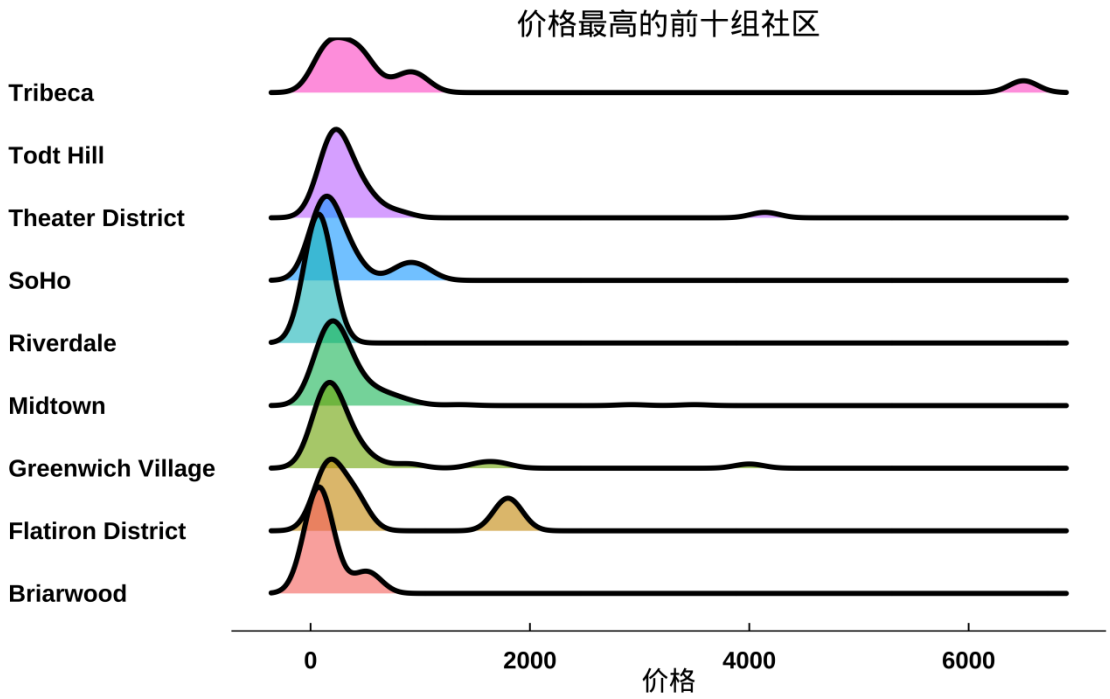

图 2.15 平均价格最高的前十组社区的价格分布图

由图 2.15，发现在 Todt Hill 上面没有密度曲线，因为它在整组数据集中可能只出现几次。

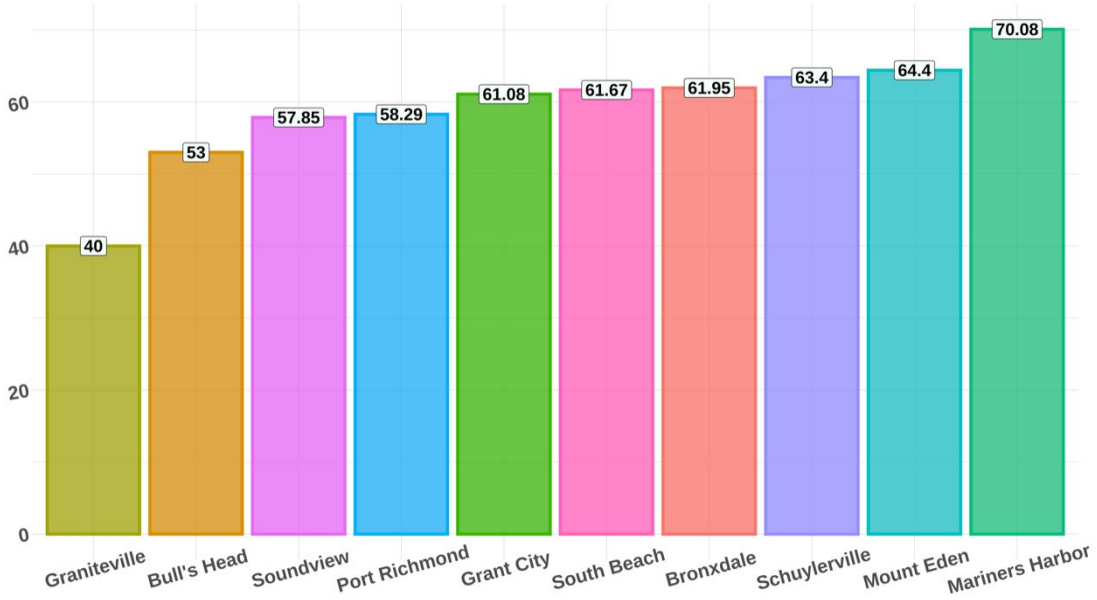

图 15　平均价格最低的后十组社区直方图

通过上面的分析，我们可以得到一些隐藏信息，比如在 airbnb 上预订最贵的 10 个社区和最便宜的 10 个社区。这些信息对于消费者来说是十分有利的。

（5）社区组与价格

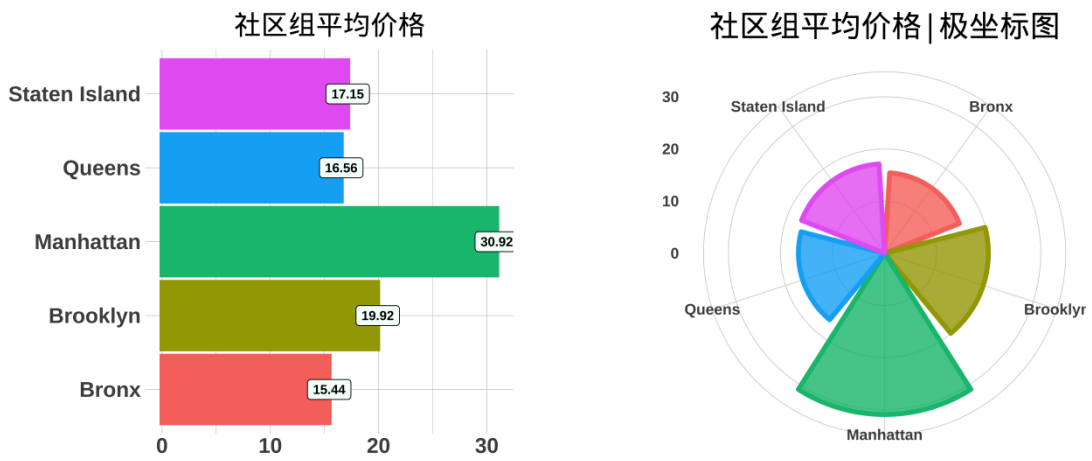

图 2.16 社区组平均价格

图 2.16 展示了不同社区组的平均价格。其中 Manhattan 的平均价格最高，为 30.92，Bronx 的平均价格最低，为 15.44。

2.3.2 数据标准化

对数变换是数据变换的一种常用方式，数据变换的目的在于使数据的呈现方式接近我们所希望的前提假设，从而更好的进行统计推断。

对价格进行对数转换：

$$P_2 = \log_2 P_1$$

其中，P2 为转换后的对数价格，P1 为数据集原本的价格。

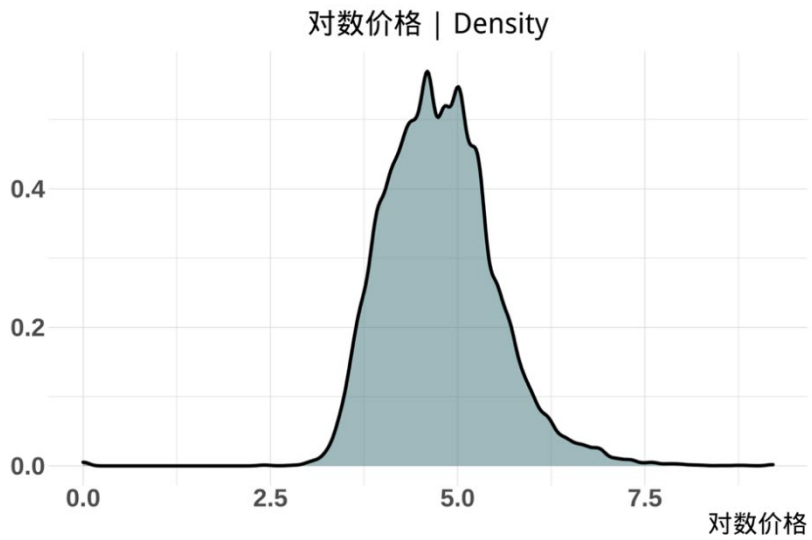

图 2.17 对数价格密度曲线

因为此数据集的样本量过大，无法使用 Shapiro-Wilk 正态检验方法，所以为了进一步确定对数价

格是否是正态分布，我们利用 QQ 图。据中一串数目的每个点都是该数据的某分位点，把这些点的（称为样本分位数点）和相应的理论上的分位数配对做出散点图，如果该数据服从正态分布，那么该图看上去应该像一条直线，否则就不服从正态分布。

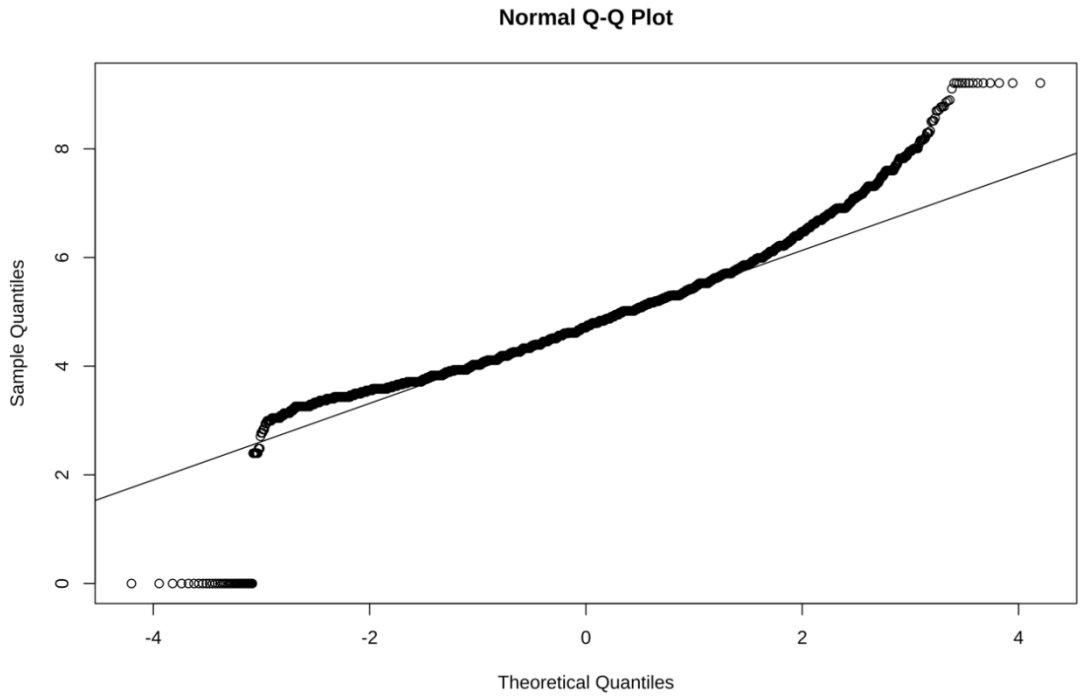

图 2.18 对数价格 QQ 图

观察 QQ 图，分布上的点近似落在直线上，所以显然，经过对数变换后的对数价格服从正态分布。

观察不同房屋类型的价格分布，会发现转换后的对数价格相比原价格更接近正态分布。

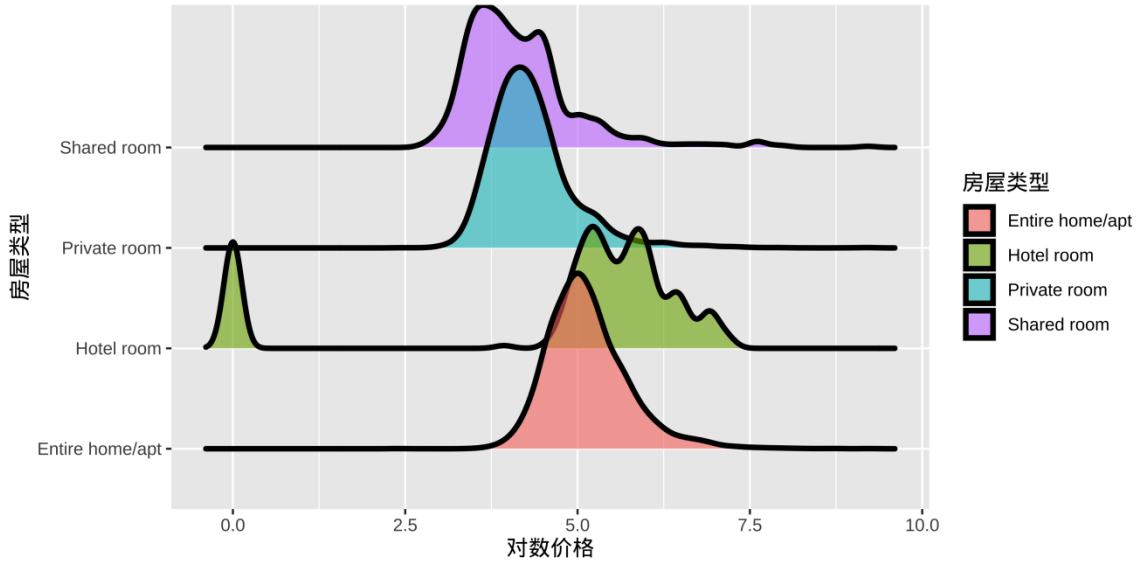

图 2.19 标准化后不同房屋类型的价格分布

2.3.3 地理位置分析

从地理分布上看，高房价区域主要集中在 纬度 40.65~40.75 区间范围内。

图 2.20 房价分布

接下来分别从房屋所在社区组和方形来看高房价的分布：

图 2.21 不同房屋类型的房价分布

从房屋类型上看各类房屋的高房价分布纬度区间相似，特别是样本最大的 Entire home/apt 分布最明显.。Entire home/apt 的分布与 Private room 的分布最相近。

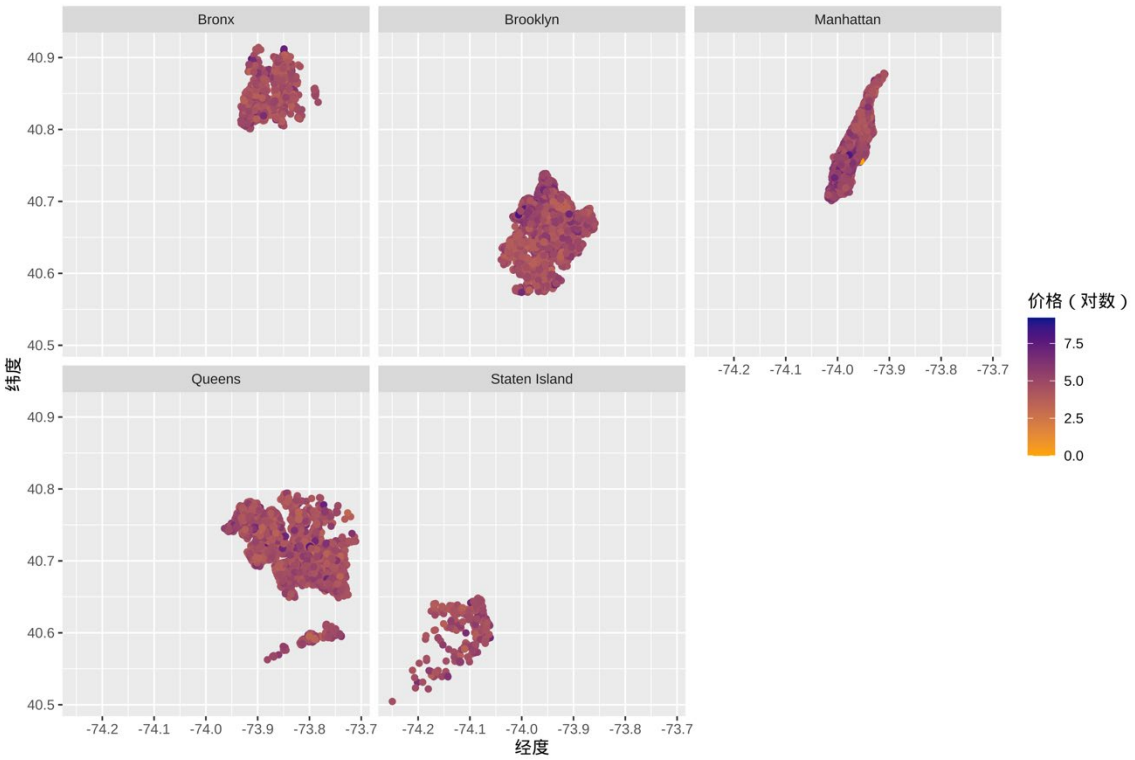

图 2.22 不同社区组的房价分布

从社区组上看，只要是横跨了 40.7 纬度的社区组，高房价多数集中在 40.65～47.5 区间内。

3 模型构建

3.1 数据分割

为了提高模型的效果，现将 70% 数据设为训练集，30%为验证集。得到训练集 26401 条，验证集 11312 条。

3.2 线性回归模型

首先，分别用房屋类型和社区组进行划分，考察纬度、经度、对数价格、最短租房时间的相关性。

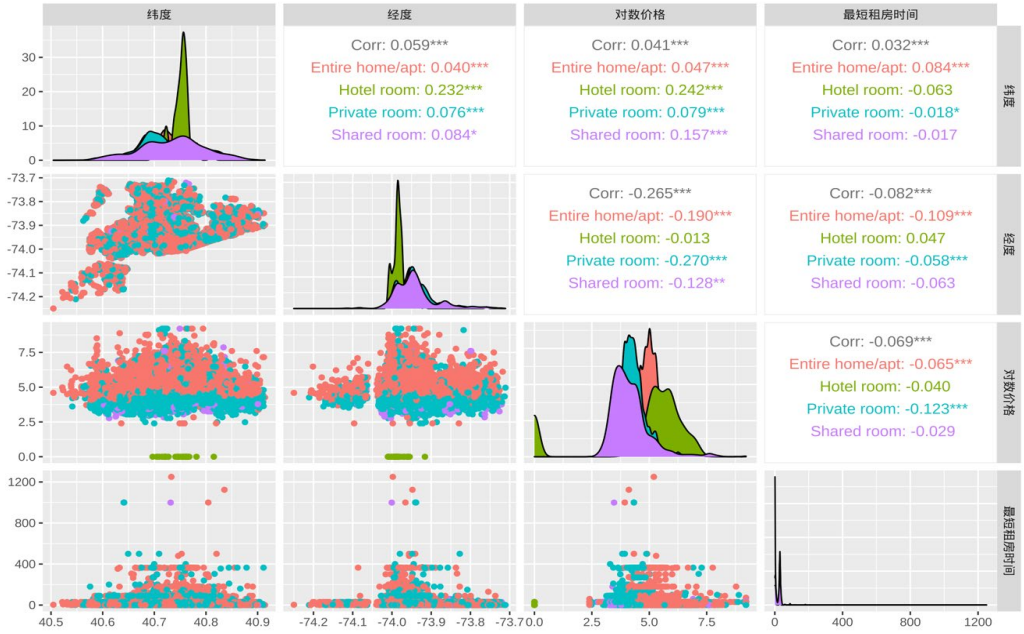

图 3.1 房屋类型分组得到相关性图

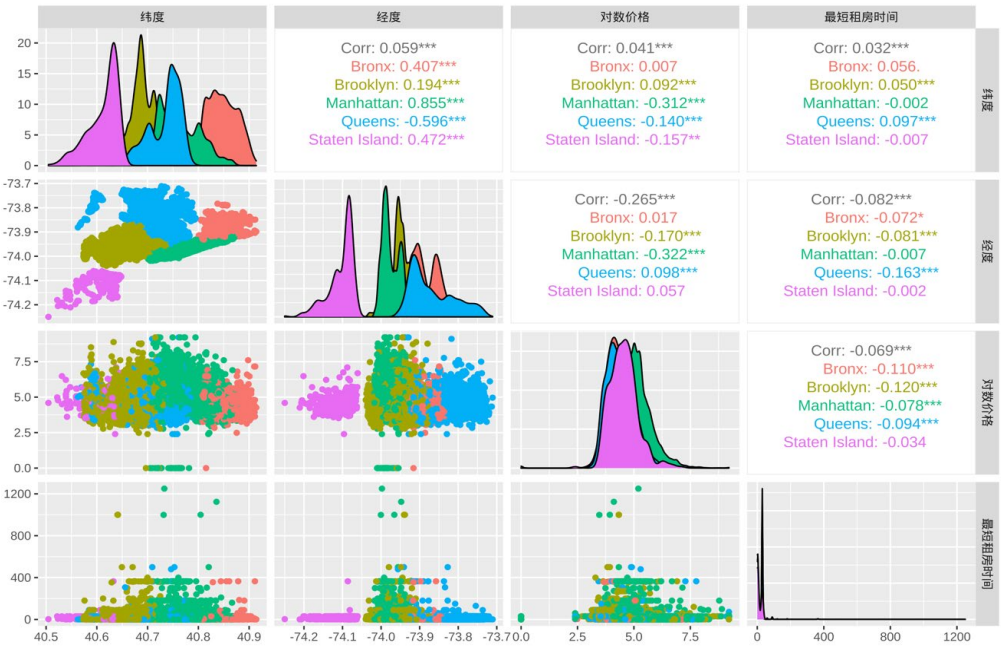

图 3.2 社区组分组得到相关性图

在建模中，除了线性回归模型外，同时使用两种树模型做拟合比较，一种是单模型 CART 决策树，一种是集成模型 梯度提升树（boosting），最后通过 RMSE 指标来比较预测结果和真实值的差异大小。

3.2.1 一元线性回归

对基本符合线性关系的数据，一元线性回归所使用的最小二乘法是：使回归的直线与散列的点在 Y 方向的距离最小为条件求出回归直线的系数 a 和 b 的。即对给定的 n 个点列：

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

设回归的直线方程为：

$$y = bx + a$$

点在 y 方向到直线的距离总的远近程度用$\sum_{i=1}^{n}[y_i - (a + bx_i)]^2$来进行定量的描述，所以可以把其看成是一个二元函数：

$$Q(a, b) = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2$$

从而把寻找一条直线，使其最接近 n 个点的问题，转化为找出两个数$\widehat{a}$，$\widehat{b}$，使二元函数$Q(a, b)$在$a = \widehat{a}; b = \widehat{b}$处达到最小的问题[8]。最后可得：

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})}$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$a = \overline{y} - b\overline{x}$$

由于各房型的价格分布峰值差异较大，因此首先考虑分布以房屋类型和社区组为控制变量的一元回归，因变量始终为对数价格。

表 3.1 定性变量线性回归

自变量	拟合曲线方程	Adjusted R-squared	p-value
房屋类型	Y= （-0.971）*Shared room + （-0.816）*Private room + （-0.337）*Hotel room + 5.147	0.29712	< 2.22e-16
社区组	Y= 0.21*Brooklyn + 0.599* Manhattan +0.002*Queens + 0.125* Staten Island + 4.42	0.098352	< 2.22e-16
房屋类型 + 社区组	Y=（-0.935）*Shared room + （-0.761）*Private room + （-0.475）*Hotel room + 0.128*Brooklyn + 0.441* Manhattan +0.02*Queens + 0.02* Staten Island + 4.877	0.35128	< 2.22e-16

通过方差分析比较控制变量贡献度。

表 3.2 方差分析

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	37709	14895.6				
2	37708	19107.6	1	-4211.93		
3	37705	13746.4	3	5361.14	4901.68	< 2.22e-16

由表 3.2，得到了 3 组拟合方程。它们的 p 值都很小，说明自变量的系数显著，但调整后的 R 方都很小，说明拟合效果都不好。但以社区组为控制变量的一元回归和以房屋类型为控制变量的一元回归相比，R 方下降，而包含两种控制变量的多元回归 R 方最高，ANOVA 分析中 p-value < 0.05,表示两种自变量都应该作为控制变量。

3.2.2 多元线性回归

接下来将最短租房时间和经度纬度分别做单自变量和所有变量组合，包括上述两种控制变量的回归，同时，由于前面的观察显示纬度在特定区间房价较高，因此再增加一个纬度二次项的模型做比较。

设随机变量 y 随着 m 个自变量 x1,x2,…,xm 变化,并有如下线性关系式:

$$y = \beta_0 + \beta_1 x_1 \cdots \beta_m x_m + \epsilon$$

根据最小二乘法可以得到结果：

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

表 3.3 对数价格 ~ 房屋类型 + 社区组 + 最短租房时间线性回归

对数价格 ~ 房屋类型 + 社区组 + 最短租房时间,		
	Estimate	Pr(>|t|)
(Intercept)	4.911	< 2.2e-16 ***
房屋类型 Hetel room	-0.508	< 2.2e-16 ***
房屋类型 Private room	-0.761	< 2.2e-16 ***
房屋类型 Shared room	-0.937	< 2.2e-16 ***
社区组 Brooklyh	0.142	1.352e-13 ***
社区组 Manhattan	0.461	< 2.2e-16 ***
社区组 Queens	0.027	0.1886
社区组 Staten lsland	0.014	0.7172
最短租房时间	0.002	< 2.2e-16 ***
Adjusted R-squared: 0.35932	F-statistic: 2644.8 on 8 and 37704 DF	p-value: < 2.22e-16

表 3.4 对数价格 ~ 房屋类型 + 社区组 + 经度线性回归

对数价格 ~ 房屋类型 + 社区组 + 经度		
	Estimate	Pr(>|t|)
(Intercept)	-157.63	< 2.2e-16 ***
房屋类型 Hetel room	-0.499	< 2.2e-16 ***
房屋类型 Private room	-0.751	< 2.2e-16 ***

房屋类型 Shared room	-0.928	< 2.2e-16 ***
社区组 Brooklyh	-0.018	0.373378
社区组 Manhattan	0.242	< 2.2e-16 ***
社区组 Queens	0.057	0.005312 **
社区组 Staten lsland	-0.463	< 2.2e-16 ***
经度	-2.199	< 2.2e-16 ***
Adjusted R-squared: 0.35967	F-statistic: 2648.9 on 8 and 37704 DF	p-value: < 2.22e-16

表 3.5 对数价格 ~ 房屋类型 + 社区组 + 纬度线性回归

对数价格 ~ 房屋类型 + 社区组 + 纬度

	Estimate	Pr(>|t|)
(Intercept)	85.623	< 2.2e-16 ***
房屋类型 Hetel room	-0.488	< 2.2e-16 ***
房屋类型 Private room	-0.746	< 2.2e-16 ***
房屋类型 Shared room	-0.924	< 2.2e-16 ***
社区组 Brooklyh	-0.200	< 2.2e-16 ***
社区组 Manhattan	0.276	< 2.2e-16 ***
社区组 Queens	-0.219	< 2.2e-16 ***
社区组 Staten lsland	-0.447	< 2.2e-16 ***
纬度	-1.976	< 2.2e-16 ***
Adjusted R-squared: 0.35952	F-statistic: 2647.1 on 8 and 37704 DF	p-value: < 2.22e-16

表 3.6 对数价格 ~ 房屋类型 + 社区组 + 最短租房时间 + 经度 + 纬度线性回归

对数价格 ~ 房屋类型 + 社区组 + 最短租房时间 + 经度 + 纬度

	Estimate	Pr(>|t|)
(Intercept)	-81.658	< 2.2e-16 ***
房屋类型 Hetel room	-0.542	< 2.2e-16 ***
房屋类型 Private room	-0.739	< 2.2e-16 ***
房屋类型 Shared room	-0.922	< 2.2e-16 ***
社区组 Brooklyh	-0.265	< 2.2e-16 ***
社区组 Manhattan	0.138	1.927e-10 ***
社区组 Queens	-0.136	2.980e-09 ***
社区组 Staten lsland	-0.827	< 2.2e-16 ***
最短租房时间	-0.002	< 2.2e-16 ***
经度	-2.073	< 2.2e-16 ***
纬度	-1.630	< 2.2e-16 ***
Adjusted R-squared: 0.37435	F-statistic: 2257.4 on 10 and 37702 DF	p-value: < 2.22e-16

表 3.7 对数价格 ~ 房屋类型 + 社区组 + 最短租房时间 + 经度 + 纬度 + I(纬度^2)线性回归

对数价格 ~ 房屋类型 + 社区组 + 最短租房时间 + 经度 + 纬度 + I(纬度^2)		
	Estimate	Pr(>|t|)
(Intercept)	-32325	< 2e-16 ***
房屋类型 Hetel room	-0.570	< 2e-16 ***
房屋类型 Private room	-0.733	< 2e-16 ***
房屋类型 Shared room	-0.909	< 2e-16 ***
社区组 Brooklyh	-0.461	< 2e-16 ***
社区组 Manhattan	-0.047	0.03966 *
社区组 Queens	-0.407	< 2e-16 ***
社区组 Staten lsland	-0.700	< 2e-16 ***
最短租房时间	-0.002	< 2e-16 ***
经度	-1.445	< 2e-16 ***
纬度	1584	< 2e-16 ***
I(纬度^2)	-19.465	< 2e-16 ***
Adjusted R-squared: 0.38392	F-statistic: 2137.5 on 11 and 37701 DF	p-value: < 2.22e-16

将上述模型进行方差分析：

表 3.8 模型方差分析结果

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	37704	13575.8				
2	37704	13568.2	0	7.5571		
3	37704	13571.5	0	-3.3095		
4	37702	13256.6	2	314.8809	454.724	< 2.22e-16 ***
5	37701	13053.3	1	203.3054	587.192	< 2.22e-16 ***

方差分析的结果表示加入了纬度二次项的模型 P-value < 0.05 拒绝原假设,效果最好.因此线性回归最终决定使用包含纬度二次项的模型，即最优模型为：

Y=(-0.57)*Hetel room -0.733*Private room-0.909*Shared room-0.461*Brooklyh-0.047*Manhattan-0.407*Queens-0.7*Staten Island -0.002*最短租房时间 -1.445*经度+1584*纬度-32325

3.3 决策树

CART（Classification And Regression Trees，分类回归树）算法是一种树构建算法，既可以用于分类任务，又可以用于回归，在回归任务中则以均方误差作为特征选择的依据

CART 回归树的度量目标是，对于任意划分特征 A，对应的任意划分点 s，可切分成数据集 D1 和 D2，求出使 D1 和 D2 各自集合的均方差最小，同时 D1 和 D2 的均方差之和最小所对应的特征和特征值划分点。

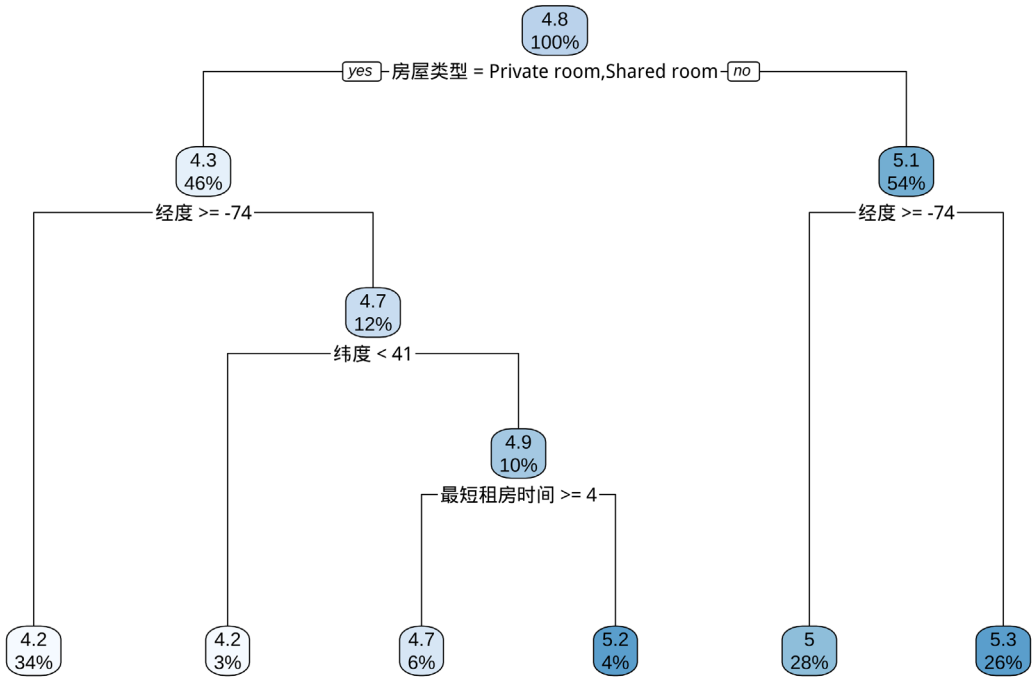

图 3.2 决策树初始划分预测

一棵树如果节点过多，则表明该模型可能对数据进行了“过拟合”。我们可通过降低决策树的复杂度来避免过拟合，最有效的手段是进行剪枝处理（pruning）。

cp 全称为 complexity parameter，指某个点的复杂度，对每一步拆分,模型的拟合优度必须提高的程度。

接下来通过筛选最小误差寻找最优剪枝点。

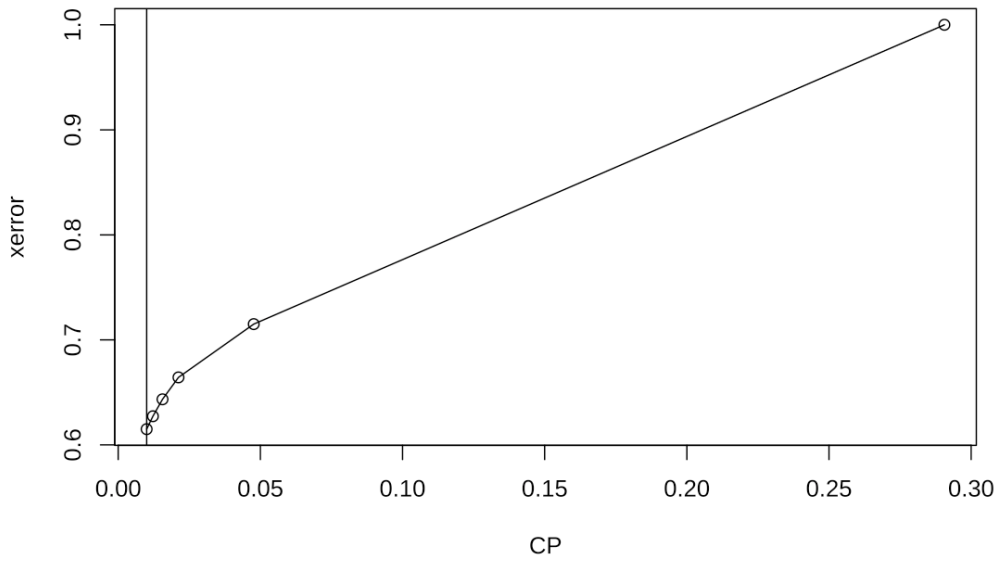

图 3.3 决策树初始划分预测

训练剪枝模型并画出剪枝后的树。

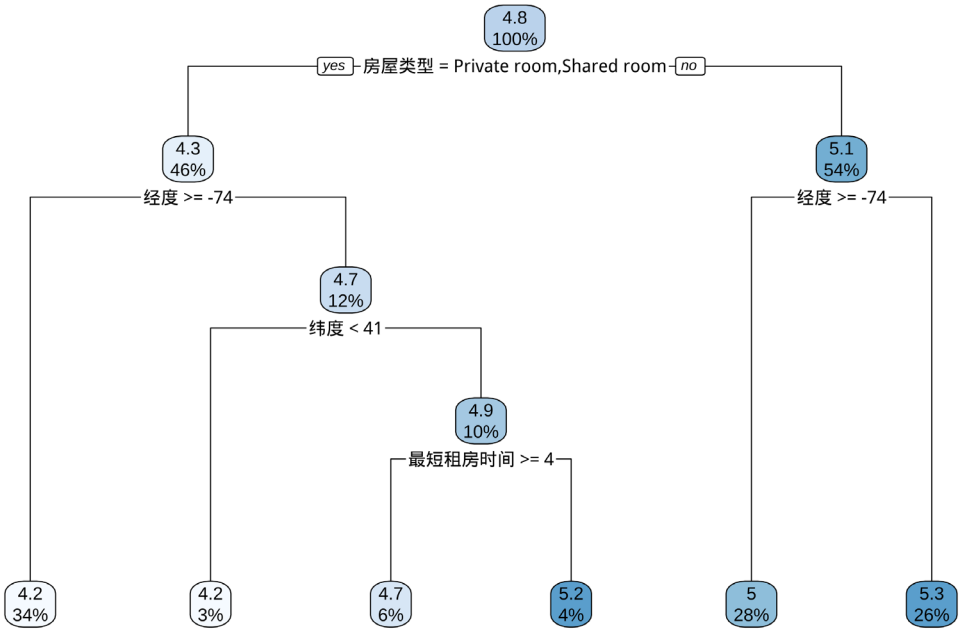

图 3.4 决策树初始划分预测

剪枝后的树和剪枝前的树几乎无变化，表示树模型已经足够简练。

最后通过总贡献度指标排序查看树模型的变量重要性。

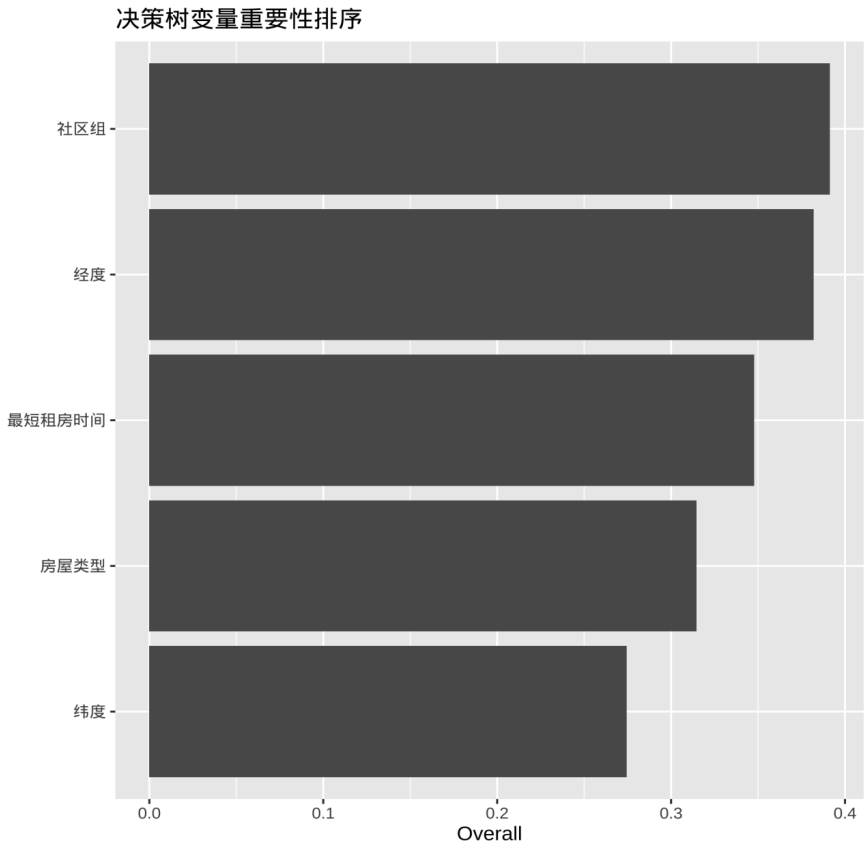

图 3.5 决策树变量重要性排序

上图显示影响价格最重要的变量是社区组，其次是经度，最后是纬度，此顺序与线性回归中似乎不太一样，最后我们来看 boosting 算法。

3.4 梯度提升树（boosting）

GBDT(Gradient Boosting Decision Tree)模型是一种迭代的决策树算法。该算法由多棵决策树组成，通常都是上百棵树，而且每棵树规模都较小，即树的深度较浅。模型预测时，对于输入的一个样本实例，首先会赋予一个初值，然后会遍历每一棵决策树，每棵树都会对预测值进行调整修正，最终的结果是将每一棵决策树的结果进行累加得到的最后得到预测的结果。

首先训练模型，使用 5 折交叉验证，然后使用最小误差模型，得到结果：

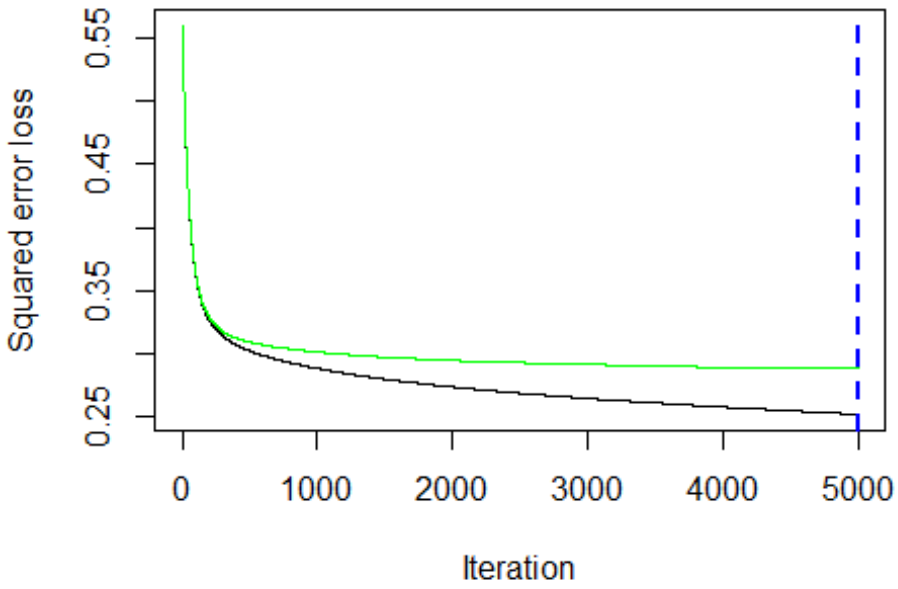

图 3.6 均方误差与迭代次数关系图

上图显示最小误差模型迭代次数次数为 4999 次，接下来看重要性排序。

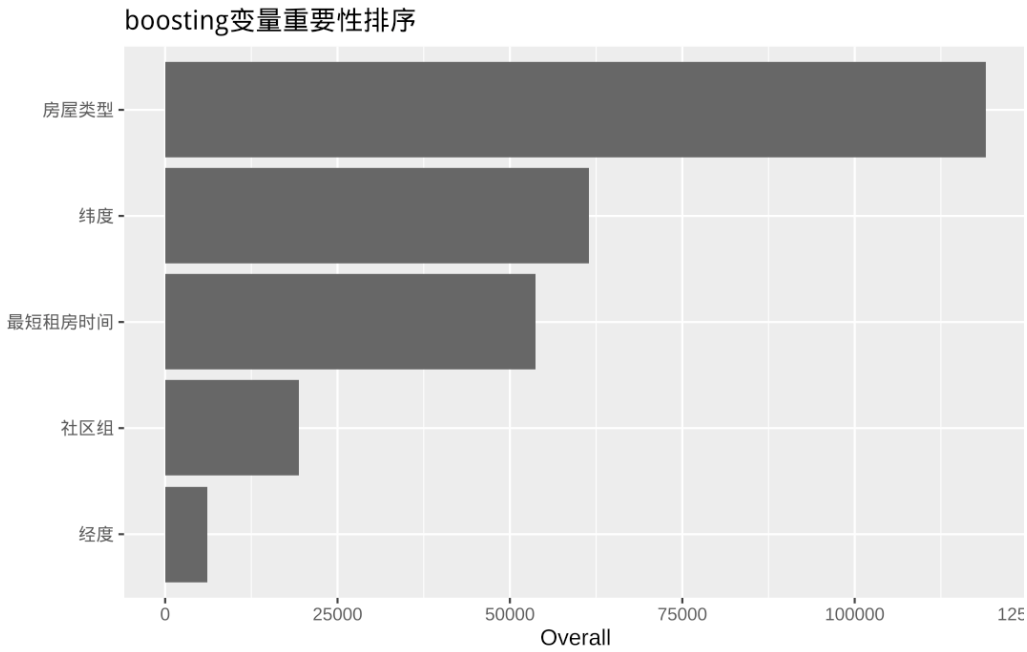

图 3.7 boosting 重要性排序

在 boosting 模型中最重要的变量是房屋类型，其次是纬度和最短租房时间，最后是社区组和经度。

在 boosting 模型中最重要的变量是房屋类型，其次是纬度和最短租房时间，最后是社区组和经度。

3.5 模型比较

表 3.9 三种模型

model	RMSE_log	rmse_prob
Linear Regression	0.58252757	0.79055848
Desicion Tree	0.57628478	0.77941522
Boosting	0.70119916	1.01616896

Rmse_log 指对数价格的预测值与真实值的 rmse，而 rmse_prob 是指将 Rmse_log 从对数转换为常数后得到的比例。

通过上表比较看出 boosting 模型的对房价对数值的预测误差最小，但换算过来仍然有接近 70%的误差区间，综合前述分析，决定价格的因素首先要看具体房屋类型，然后纬度在 40.7 左右的区域价格会偏高，租房时间越短房价越高，在此基础上不同的地段和经度都会不同程度影响房价。

参考文献

[1] 陈琳,陈涛.基于 LDA 模型和信任维度的在线短租用户信任感知空间分布研究——基于 Airbnb 北京地区数据[J].
[2] 中国发展,2021,21(05):53-61.DOI:10.15885/j.cnki.cn11-4683/z.2021.05.008.
[3] 百度百科.[DB/OL]. https://baike.baidu.com/item/Airbnb/5204658#reference-[1]-5898285-wrap
[4] 李盛达.基于多变量线性回归的房价预测模型[J].科学技术创新,2021(06):91-92.
[5] 马红丽,徐长英,杨新鸣.决策树模型在中医药领域的应用现状[J].世界中医药,2021,16(17):2648-2651+2656.
[6] SiyueLin, GBDT：梯度提升决策树, https://www.jianshu.com/p/005a4e6ac775
[7] 张建方, 王秀祥. 直方图理论与最优直方图制作[J]. 应用概率统计, 2009, 25(2):201-214. DOI:10.3969/j.issn.1001-4268.2009.02.010.
[8] 周健民．土壤学大辞典：科学出版社，2013.10
[9] 刘晓叙. 灰色预测与一元线性回归预测的比较[J]. 四川理工学院学报：自然科学版, 2009, 22(1):3.