洛杉矶酒店市场研究与定价指导

致理-数理 1　刘苏青* 2021013371

机械 00　姚圣泰† 2019011497

致理-数理 2　沈宇捷‡ 2022012290

2024 / 01 / 14

摘　要

本文基于洛杉矶地区 1388 家酒店的信息，包括设施、服务、地理位置、价格等因素，深入分析洛杉矶酒店市场，并使用统计学方法，探索建立一个合理的定价模型，够为酒店业者提供一种科学而有效的定价指导。最后基于 Rshiny 制作交互界面，为用户提供一个直观、灵活的工具，能够根据具体情况调整参数，快速获取并应用我们的定价模型。

关键词：**酒店数据**；**爬虫**；**主成分分析**；**逻辑回归**；**支持向量机**；**随机森林**；Rshiny

*liu-sq21@mails.tsinghua.edu.cn

†yaost19@mails.tsinghua.edu.cn

‡shenyj22@mails.tsinghua.edu.cn

目录

1 引言

洛杉矶作为美国最大的城市之一，以其独特的文化、娱乐和商业氛围吸引着数以百万计的游客。在这个充满活力的城市中，酒店行业扮演着至关重要的角色，为游客提供舒适的住宿环境和多样化的服务。随着旅游业的不断发展和游客对住宿期望的提高，洛杉矶的酒店市场也面临着日益激烈的竞争。

本研究旨在深入分析洛杉矶的酒店市场，着重关注酒店提供的设施、服务、地理位置等方面的信息，以及住客对这些因素的评价。通过对这些客观数据的综合分析，我们将尝试揭示洛杉矶酒店市场的现状和趋势，为酒店业者提供有价值的市场洞察。

在研究的第二阶段，我们将基于酒店提供的设施和服务等客观信息，探索建立一个合理的定价模型。通过运用先进的数据分析和机器学习技术，我们希望能够为酒店业者提供一种科学而有效的定价指导，使他们能够更好地满足市场需求，提高竞争力。

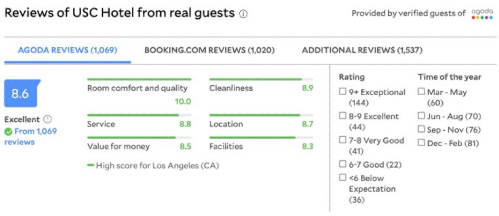
通过对洛杉矶酒店市场的全面研究，本研究旨在为酒店业者、投资者和市场从业者提供深刻的理解和有针对性的建议，以促进酒店业的可持续发展。同时，通过利用先进的分析工具，我们将为酒店业者提供一种更精准、智能的定价策略，帮助他们在激烈的市场竞争中脱颖而出。

2 数据获取

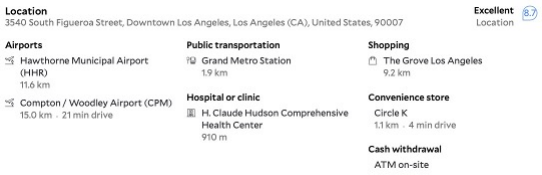
2.1 网站选择

Agoda 是一个全球在线酒店预订网站，提供全球范围内的酒店、度假村、公寓和其他类型的住宿预订服务。Agoda 通过其网站和移动应用程序为用户提供广泛的住宿选择，覆盖了世界各地的目的地。

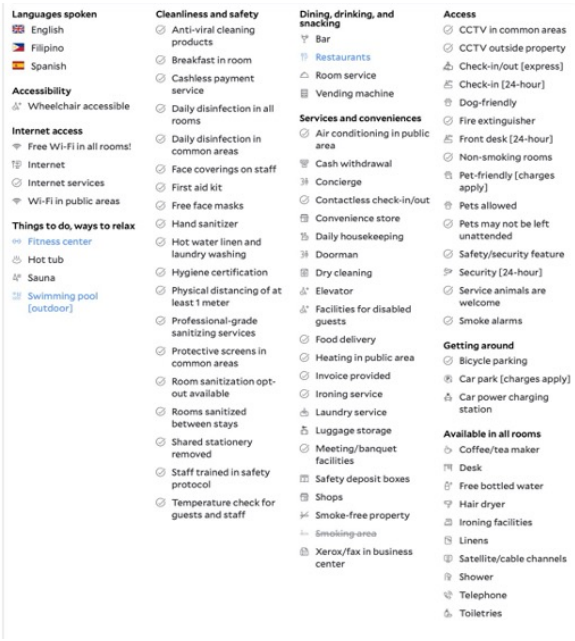
更重要的是，Agoda 网站中提供了很多酒店的标准化参数，包括设施、服务、位置等，为用户提供了全面的了解酒店的途径。以Figure 1为例，是由 Agoda 提供选项，各酒店选择其具有的信息。标准化的信息框架，使得我们能够系统性地收集和比较各酒店的特征，为后续进行数据分析提供了可能性及便利性。



(a) 酒店各项评分参数



(b) 酒店周边信息



(c) 酒店设施信息

Figure 1: Agoda 中提供的酒店标准化参数

2.2 网络爬虫

打开 Agoda 网页，我们尝试使用 rvest 包进行爬取，结果发现爬取信息不全，很多节点缺失。调查发现，Agoda 网页使用了动态加载技术，它可以在网页加载时延迟加载不必要的资源, 以提高页面的加载速度和性能，因此使用 rvest 爬取时出现节点缺失的问题。

我们采用 Rselenium 包进行网页爬取工作。Rselenuim 包可以通过模拟浏览器行为，实现滚动、点击、输入等功能。并且通过时间延迟，等待网页全部加载完毕再进行爬取，实现对动态网页的爬取。

爬取工作分为两个部分。首先，在 Agoda 中选择洛杉矶地区，网页如Figure 2所示。对此网页，爬取获得洛杉矶所有酒店的链接，爬取流程图如Figure 3所示。去重后，得到了 1388 个酒店链接。接着对每个酒店页面进行爬取，酒店页面如Figure 4所示。一次打开上述保存的链接，爬取网页，流程如Figure 5所示。

最终，我们得到了 1388 个酒店的名称、价格、设施、服务、位置等信息，具体获取的信息内容将在2.3小节中详细描述。最终的 csv 文件，共有 1388 行，每行有 52 个特征，可在文件夹中 HOTELS.csv 中查看。

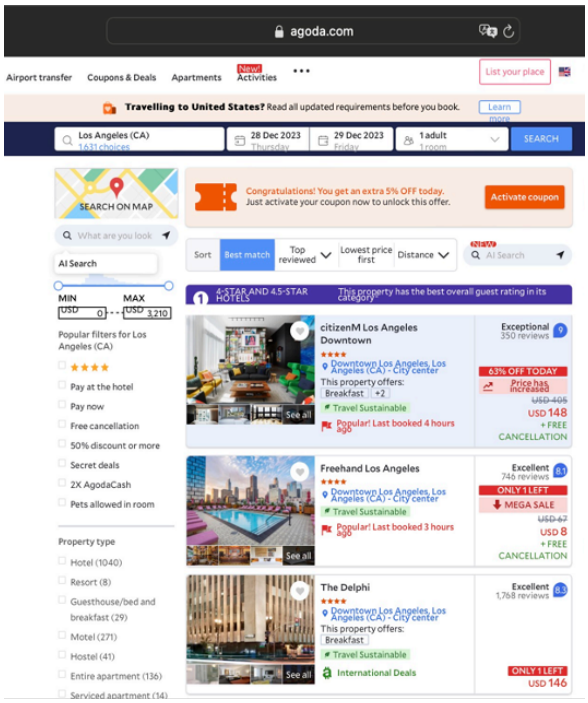


Figure 2: Agoda 上洛杉矶地区的所有酒店

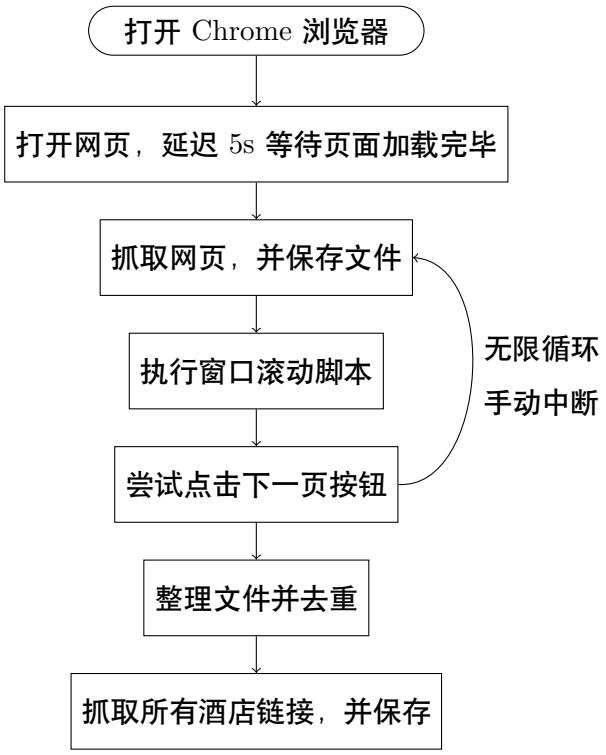


Figure 3: Rselenium 爬取酒店链接流程

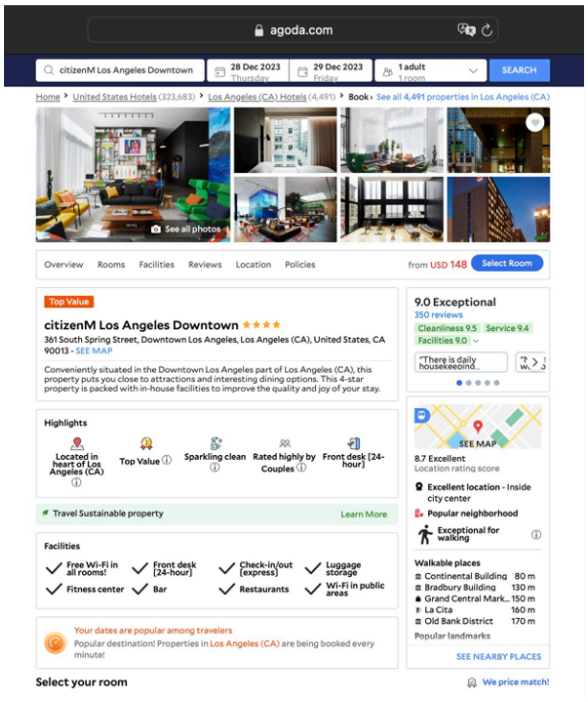


Figure 4: 每个酒店具体信息的页面

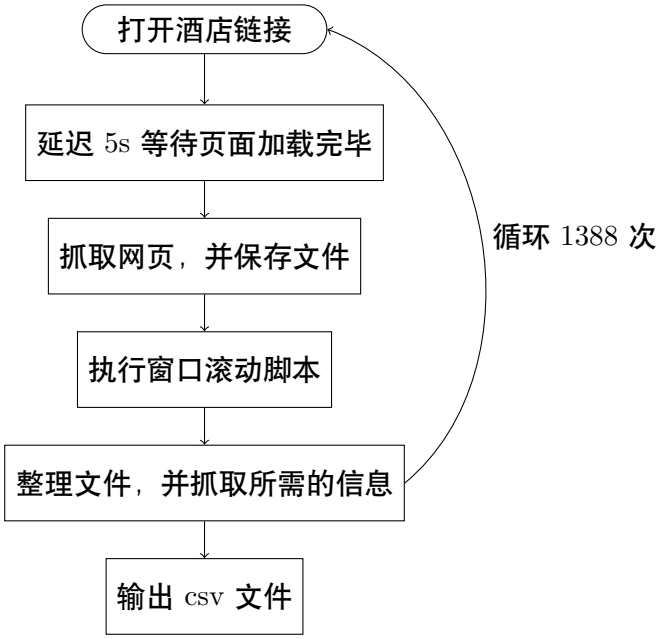


Figure 5: Rselenium 爬取酒店链接流程

2.3 获取信息

如Table 1所示，从 Agoda 的酒店页面中，爬取了酒店基本信息、住客评分、设施和服务、位置等信息。其中，住客评分是一个 0f0 的小数，越接近 10 代表评分越高。设施和服务中，Table 1中所列是项目名称，每一项下是该酒店具有的该项的所有设施服务。以上网服务一项为例，该酒店包含公用区 WiFi；上网服务；所有客房免费 WiFi；网络服务。爬取时，每一项中的不同内容使用“@@”分隔，例如“Free Wi-Fi in all rooms! @@ Internet @@ Internet services @@ Wi-Fi in public areas”。位置信息中，每一项都是距离该项的距离，并且列举出一项或多项。其中地名与距离之间使用“##”分隔，不同地名之间使用“@@”分隔，例如“Staples Center ## 3.8 km @@ La Brea Tar Pits and Museum ## 5.0 km @@ Hollywood Walk Of Fame ## 5.0 km @@ Madame Tussauds Hollywood ## 5.8 km @@ Griffith Observatory ## 6.3 km @@ Griffith Park ## 7.3 km @@ Hollywood Sign ## 8.0 km @@ Universal Studios Hollywood ## 9.8 km @@ The Getty Center ## 16.1 km @@ Venice Beach ## 17.8 km”

基本信息	住客评分	设施和服务	位置
• 酒店名称 • 酒店价格 • 酒店星级	• 总评分 • 客房舒适度 • 服务 • 性价比 • 环境和清洁度 • 位置 • 设施	• 可用语言 • 无障碍设施 • 上网服务 • 休闲娱乐 • 清洁度和安全 • 餐饮服务设施 • 便利服务设施 • 其他设施服务 • 交通服务 • 所有客房均提供	• 附近机场 • 公交/地铁站 • 医疗机构 • 逛街购物 • 便利店 • 取款服务

Table 1: 从 Agoda 获取的酒店的信息

3 数据预处理

获取的四类信息中，星级与评分为数字类型数据，而设施服务与位置为字符串类型，需要进行预处理。

3.1 设施和服务数据处理

设施和服务每一项例如可用语言下分多个小类，只可能为有或没有，为离散数据，因此采用独热编码生成 1388 行 360 列的 0-1 矩阵。列数为特征数量，维度较高，因此采用 PCA 主成分分析方法进行降维处理，把可能具有相关性的高维变量合成线性无关的低维变量，降低数据复杂度，识别最重要的多个特征。最终获得 Facilities 矩阵。

V	W
Languages spoken	Accessibility
English @@ Filipino @@ Spanish	Wheelchair accessible
English @@ Korean @@ Spanish	Wheelchair accessible
English @@ Spanish	Wheelchair accessible
English @@ Arabic @@ Chinese [Cantonese] @@ Ch	NA
English @@ Spanish	Wheelchair accessible
English @@ Arabic @@ Filipino @@ Hindi @@ Japa	Wheelchair accessible
English	Wheelchair accessible
English @@ Hindi @@ Spanish	Wheelchair accessible
English @@ Spanish	Wheelchair accessible
English	NA
English @@ Chinese [Mandarin] @@ French @@ Spa	NA
English @@ Spanish	Wheelchair accessible
English @@ Chinese [Mandarin] @@ Filipino @@ Fre	Wheelchair accessible
English @@ Spanish	NA
English @@ Spanish	Wheelchair accessible
English @@ Chinese [Mandarin] @@ Filipino @@ Sp	Wheelchair accessible
English @@ Chinese [Mandarin] @@ Dutch @@ Filip	NA
English @@ Spanish	Access all room by interic
English	Wheelchair accessible
English @@ Spanish	Wheelchair accessible

(a) 原数据

	PC1	PC2	PC3	PC4	PC5	PC6
1	-6.6782530	-3.72435260	-0.9817806641	0.046531744	2.230721349	-0.7332653
2	-9.5815055	-2.85200399	-1.0057042607	-0.366894324	-1.942497592	1.5640168
3	-6.7658094	-5.48328456	-0.5883686326	3.868434406	3.380328259	-1.1662749
4	-4.6446171	1.24526806	29.2375324077	-9.065285532	0.252707264	4.3387362
5	-4.2808033	2.53130167	-2.2927452604	-5.752053977	-2.206124575	-6.4608037
6	-6.3893260	-0.59258967	-0.1970932999	-4.206503090	1.254579729	0.8971488
7	-3.2317524	1.03704808	-1.7941927913	-5.173952486	0.293147116	-0.1958897
8	1.8860699	-2.11520093	-0.2398341032	0.858566660	4.580473887	0.4601399
9	-5.4362160	-3.52777531	-0.0004950138	2.913190275	-3.668030632	4.1501219
10	3.2247871	-1.58324219	-0.4221852872	0.901058939	3.683717172	0.3105751
11	0.2908885	0.18084303	-0.4579100727	-1.972528304	0.012954124	0.8253338
12	-10.5818166	4.46997638	-1.0314327556	-1.421510625	2.825991659	1.4980244
13	-9.5073852	-4.54230037	2.0137580044	2.909922483	-1.748891462	1.9087812

(b) Facilities 矩阵

Figure 6: 设施服务信息处理

3.2 位置数据处理

位置数据为连续变量，每一项中包含若干地点以及相应距离。考虑到距离便利店、提款服务等地点的最近距离往往体现了便利程度以及用户倾向，因此选择每项中的最近距离作为该项的数据。若为 NA，则认为该酒店附近没有较近的设施，代替以该列的最远距离；若为”on site”，则取距离 0。处理时注意保持单位一致。最终获得 Locations 矩阵。

AH	AI
Convenience store	Hospital or clinic
Circle K ## 1.1 km 4 min drive	H. Claude Hudson Comprehensive Health Center ## 910 m
7 Eleven ## 370 m	First Family Dental ## 110 m
On-site	MedMen Los Angeles - Downtown (DTLA) ## 220 m
7 Eleven ## 650 m	The Spine Clinic Of Los Angeles ## 530 m
7 Eleven ## 240 m 2 min drive	Coast Dental Group ## 170 m
On-site	Meaningfuluseexpertscom ## 530 m 2 min drive
7 Eleven ## 1.5 km 4 min drive	Valley Village Dental ## 1.2 km
7 Eleven ## 780 m	Famiy Medical ## 1.6 km
7 Eleven ## 1.3 km 5 min drive	East Los Angeles Health Center ## 2.9 km
7 Eleven ## 970 m	Children's Dental Group ## 510 m
7 Eleven ## 1.2 km	Reasons Eating Disorder Center ## 850 m
7 Eleven ## 740 m	Reliant Urgent Care ## 1.3 km
On-site	Reliant Urgent Care ## 230 m
7 Eleven ## 880 m 2 min drive	Garden Grove Hospital And Medical Center ## 1.6 km
7 Eleven ## 1.5 km	Memorial Hospital Of Gardena ## 2.6 km
7 Eleven ## 550 m	Daniel E. Cronk, DDS ## 7.1 km
7 Eleven ## 320 m	MedMen Los Angeles - Downtown (DTLA) ## 590 m
On-site	Us Healthworks Medical Group ## 310 m 10 min walk
7 Eleven ## 310 m	T. Yamashita, M.D., Diplomate of Ophthalmogy ## 720 m

(a) 原数据

	V1	V2	V3	V4	V5	V6	V7	V8
1	11600	1900	910	9200	1100	2000	2900	60
2	15700	80	110	5300	370	2000	3800	80
3	17400	400	220	9800	3000	10	1000	50
4	15400	620	530	8200	650	2000	1200	350
5	10800	4800	170	10000	240	180	34200	170
6	8400	9800	530	10000	3000	280	19300	170
7	6500	3100	1200	10000	1500	160	5400	1200
8	3900	740	1600	10000	780	2000	33800	250
9	14300	820	2900	10000	1300	2000	12300	40
10	12800	3300	510	10000	970	2000	4500	510
11	3900	4100	850	10000	1200	280	18800	520
12	3300	1800	1300	10000	740	70	8000	1000
13	1900	2000	230	10000	3000	2000	6600	230
14	10800	2300	1600	10000	880	30	43000	310
15	3300	1400	2600	10000	1500	2000	14300	1400

(b) Locations 矩阵

Figure 7: 距离信息处理

4 探索性数据分析

4.1 酒店星级分布

Figure 8是洛杉矶酒店的星级分布，其中众数为 2 星酒店，中位数是 2.5 星，洛杉矶地区酒店平均星级是 2.5 星。

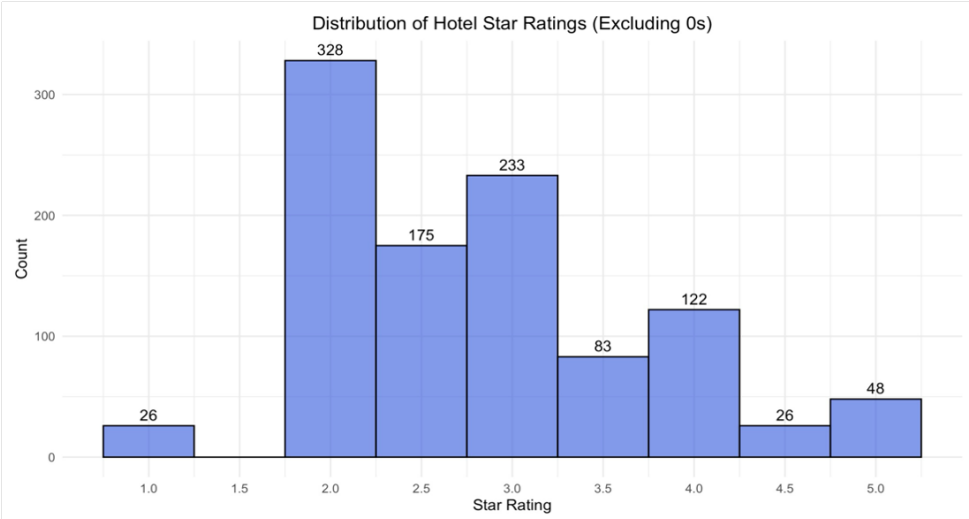
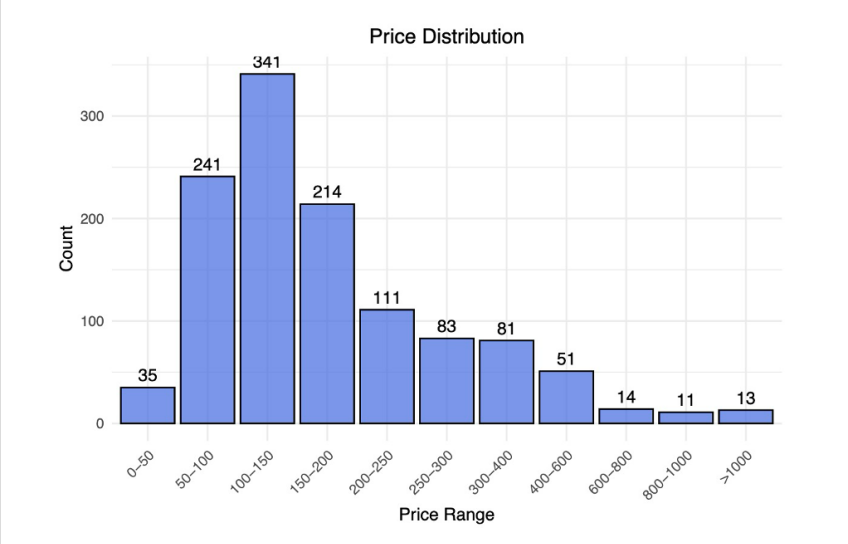


Figure 8: 洛杉矶酒店星级分布

4.2 价格分布

Figure 9是洛杉矶酒店价格分布，中位数是 146$, 平均数是 202.04$，最大值是 3642$。受限于图幅，价格的区间如Figure 9中横坐标所示。在交互界面中，我们制作了价格分布的可交互式界面，用户可以调整价格区间的间隔大小，已获得更细节、更具体的信息，见6.1。




```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
14.17  102.41  146.00  202.04  228.91 3642.00
```

Figure 9: 洛杉矶酒店价格分布

4.3 用户评分分布

Figure 10为用户评分分布，包括一个总分 “Score”，与 6 项小分，具体信息见Table 1。使用不同颜色的曲线，标明不同类型的分数的分布密度曲线。

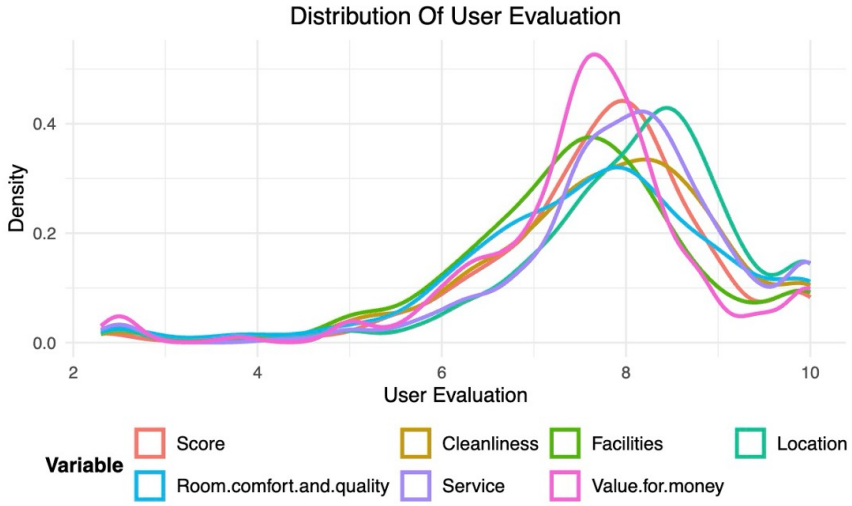


Figure 10: 洛杉矶酒店评分分布

4.4 相关性分析（评分、星级与价格之间）

Figure 11为酒店价格、星级与上述各项评分之间的相关性图。从图中可以看出，酒店价格与位置评分的相关性最高，符合我们对酒店价格的认知，越靠近市中心，位置越好的酒店价格相对越贵；酒店星级与环境和清洁度评分以及设施评分的相关程度最高。需要指出的是，酒店价格和星级之间具有较强的相关性，这点也符合酒店价格的定价规律。

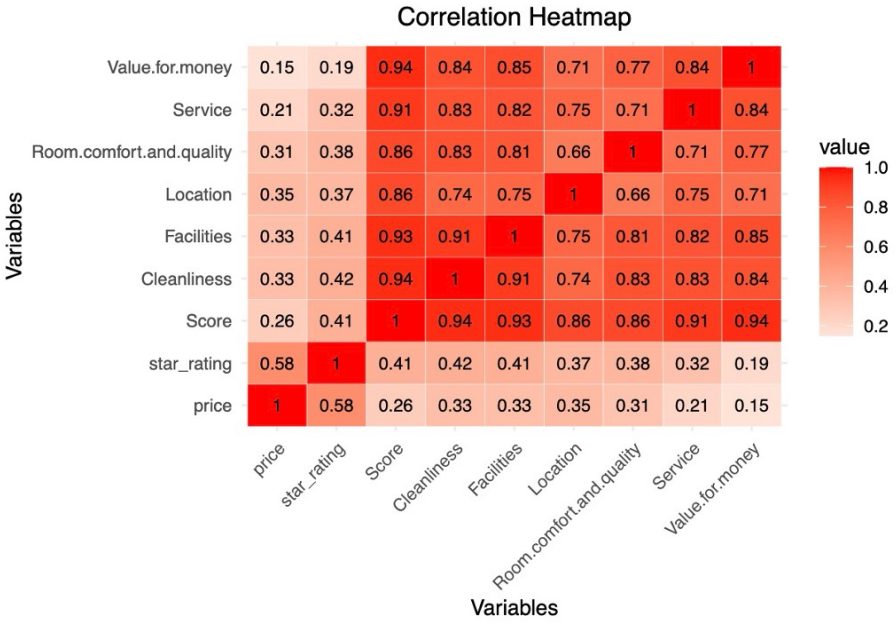


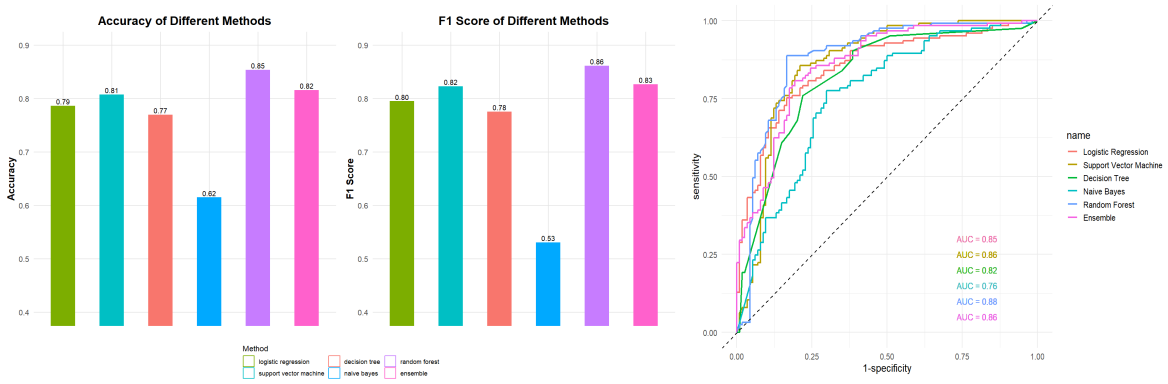
Figure 11: 评分、星级与价格之间相关性

5 价格预测

在section 3中，我们通过独热编码和 PCA 降维的方法获得了酒店的 Facilities 矩阵，它包括了酒店的可用语言、休闲娱乐、清洁度和安全性、餐饮、便利服务、其他设施以及客房提供的物品等设施信息；并通过正则表达式匹配每个单元格中的距离最小值来获取 Locations 矩阵，它包括了酒店到附近机场、公交/地铁站、医疗机构、逛街购物、便利店以及取款服务等距离信息。这些信息都是酒店的客观条件，即住客和酒店管理者可以显式观测到的信息，所以接下来我们将 Facilities 矩阵和 Locations 矩阵合并起来作为自变量，并将酒店定价作为因变量，进行价格预测任务。

5.1 分类问题

我们考虑从两个角度来进行这个预测问题，首先是尝试进行一个较为简单的分类问题。我们取所有酒店价格的中位数 ($140~¥1000) 作为分界线，将数据集分为酒店价格小于等于 $140 和大于 $140 的两类。其中，前者我们可以认为是较为便宜的酒店；而后者我们可以认为是较为昂贵的酒店。这里选择中位数作为分界点的一个原因是这种分法可以获得较为平衡的数据集，便于后续的训练及测试；同时中位数相较均值的一大优点是它不容易被个别酒店价格极高的样本拉偏。在给数据集贴完标签以后，我们就得到了一个很常见的二分类问题，因此我们选择了 5 种常见的二分类方法来处理这个问题，包括逻辑回归 (Logistic Regression)、支持向量机 (Support Vector Regression)、决策树 (Decision Tree)、朴素贝叶斯 (Naive Bayes) 以及随机森林 (Random Forest)，最后再将训练集上所得的分类器通过最大投票的方法进行集成，获得第六个分类器。我们将数据集按照 4:1 的比例划分成训练集和测试集，这六个分类器在测试集上的结果如下，下面三个图从左到右分别是测试集上各分类器的准确率、F1-score 和 ROC 曲线，可以看到三张图表表现出一致性，其中左侧的准确率和中间的 F1-score 都显示随机森林方法获得的分类器效果最好，而朴素贝叶斯方法获得的分类器效果最差。通过计算右侧的 ROC 曲线下方的面积，我们也可以发现随机森林方法的 AUC 能达到 0.88，因此是处理该分类问题的最好办法。同时，我们可以发现之前尝试通过最大投票获得的集成分类器效果反而略逊于随机森林，这可能是由于朴素贝叶斯分类器的分类效果偏弱，在最大投票的过程中产生了一定的干扰。

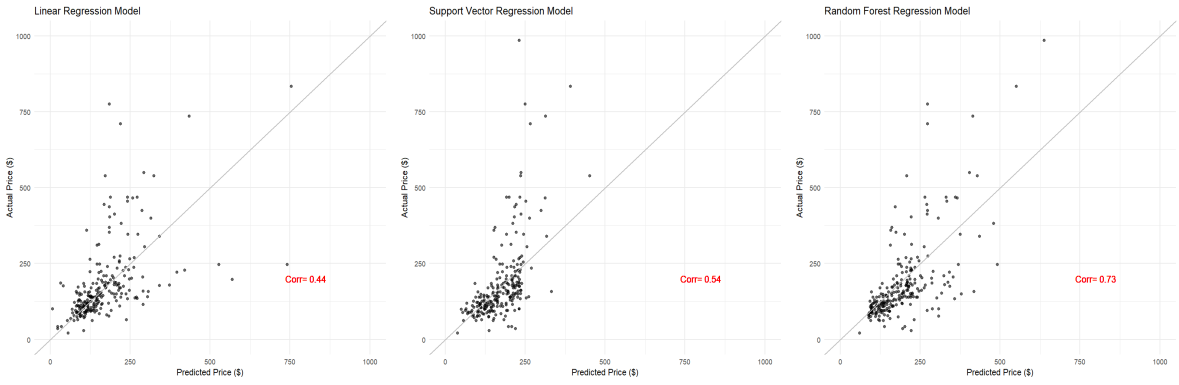


(a) Accuracy&F1-Score (b) ROC Curve

Figure 12: 分类结果

5.2 回归问题

然后我们考虑更进一步地给出酒店价格的估计值，以帮助酒店管理者更好地定价。因此我们直接取酒店价格作为因变量，关于和分类问题相同的自变量做回归，包括线性回归 (Linear Regression)、支持向量回归 (Support Vector Regression) 和随机森林回归 (Random Forest Regreesion)。从前面的探索性数据分析部分和经验知识都可以知道，酒店的价格一定是非负值，因此它的分布应该是一个右偏分布，即酒店价格较高的一侧会出现拖尾，所以我们在做回归之前还需要对酒店价格做一步 log 变换。我们将三种回归所得的预测价格关于真实价格作图，并且在图上标注 y=x 的直线。我们可以发现这些数据点基本分布在 y=x 附近，即预测价格和真实价格相差较小，并且三种方法预测值和真实值的相关系数依次增大，其中随机森林回归的相关系数可以达到 0.73。



(a) LR (b) SVR (c) RF

Figure 13: 回归结果

6 交互界面

交互界面使用 Rshiny 制作，分为三个部分，价格分布图、词云与价格预测。如Figure 14所示，在每一部分标题下，均写明了该部分的功能与使用方法。

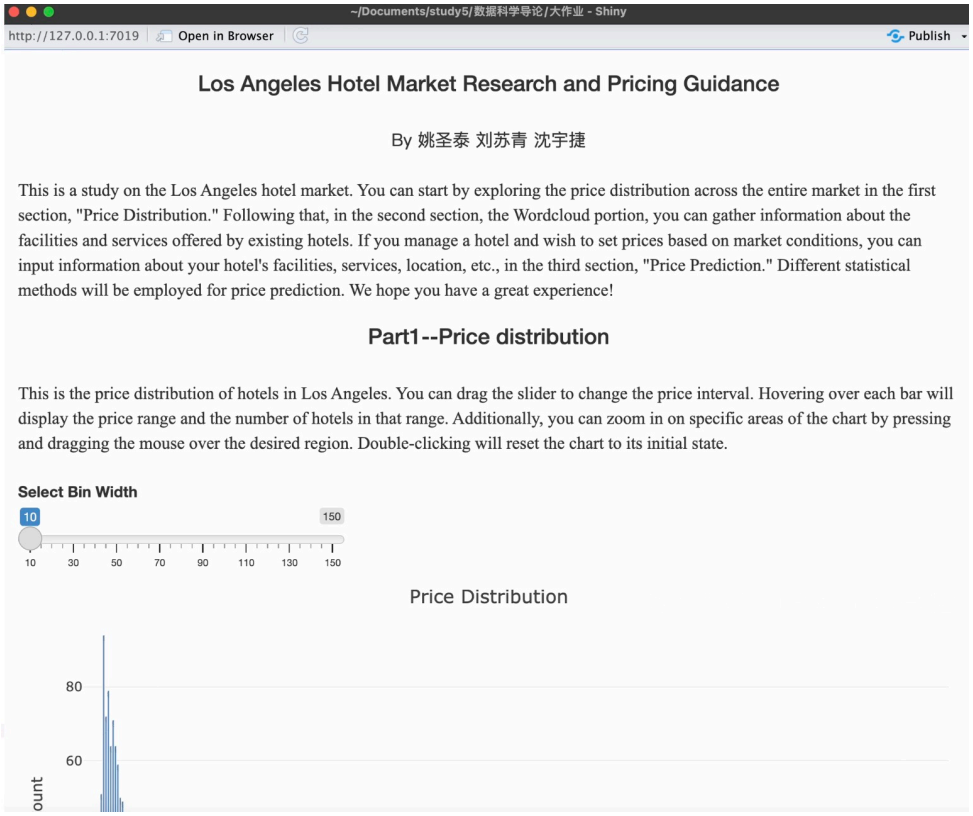


Figure 14: 交互界面

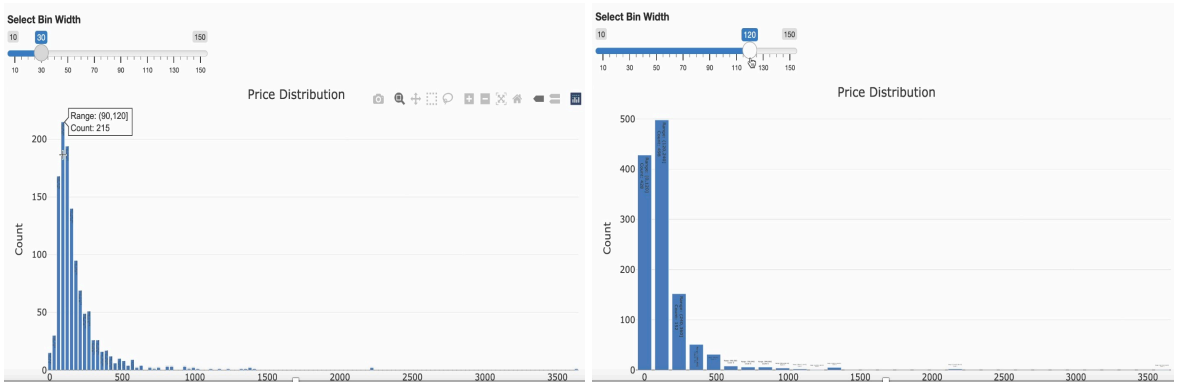
6.1 价格分布图

第一部分是洛杉矶现有酒店的价格分布，可以通过拖拽滑动条调整柱的宽度，如Figure 15所示。将鼠标悬浮在每一个柱上，可以显示所在的价格区间以及这个价格区间酒店的频数，如Figure 15a所示。

拖拽鼠标可以进行局部放大，观察细节的信息，如Figure 16所示。

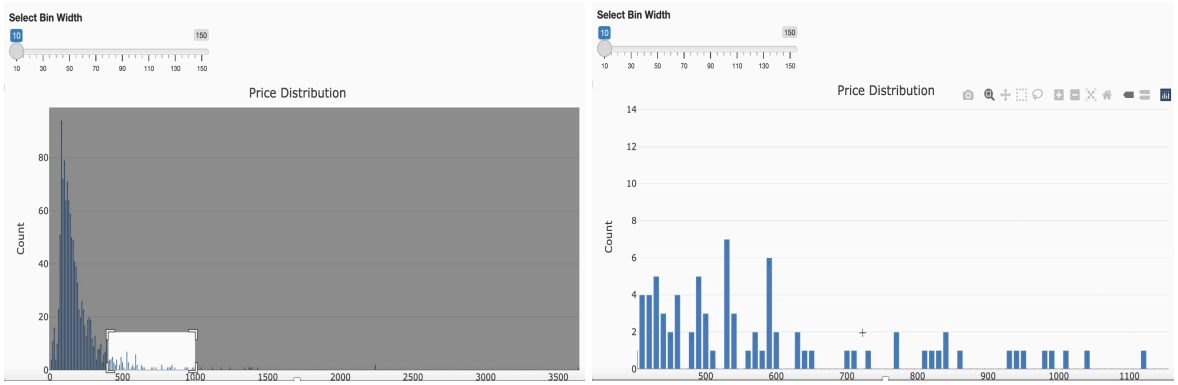
6.2 词云

第二部分是词云，汇总了现有酒店提供的清洁度、安全性、餐饮、便利等服务和设施，以帮助使用者了解酒店中最常用、最为必要的一些设施和服务，如Figure 17所示。



(a) 宽 30$ (b) 宽 120$

Figure 15: 价格分布图



(a) 原图 (b) 价格分布图局部放大

Figure 16: 局部放大

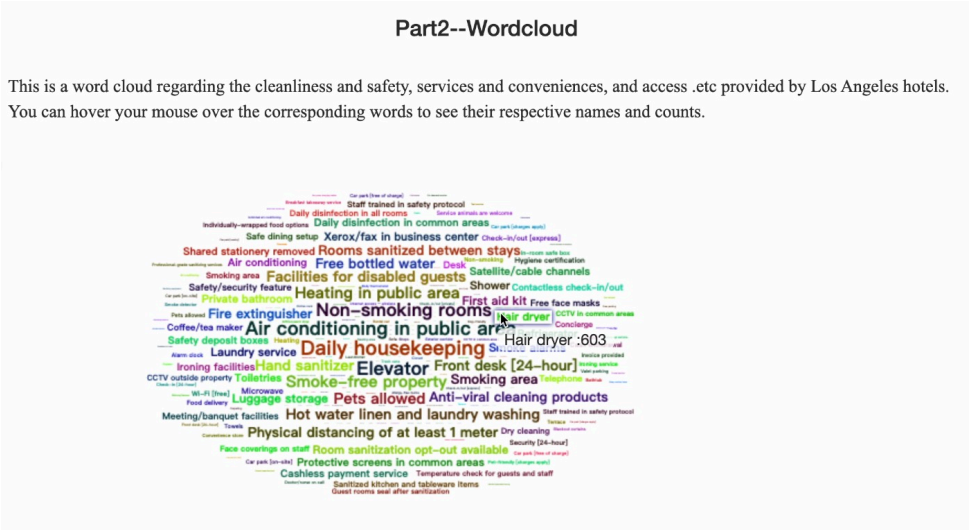


Figure 17: 设施服务词云

6.3 价格预测

第三部分是价格预测。使用者可以再左侧的界面中选择酒店具有的设施服务信息，我们设置了多选框。并且输入酒店的位置信息，离最近的机场、购物中心、医院等重要位置的距离，如Figure 18所示。在左侧选择模型，提供了线性回归，决策树和随机森林三种算法，得到预测的价格。从前面的介绍，决策树和随机森林预测效果较。

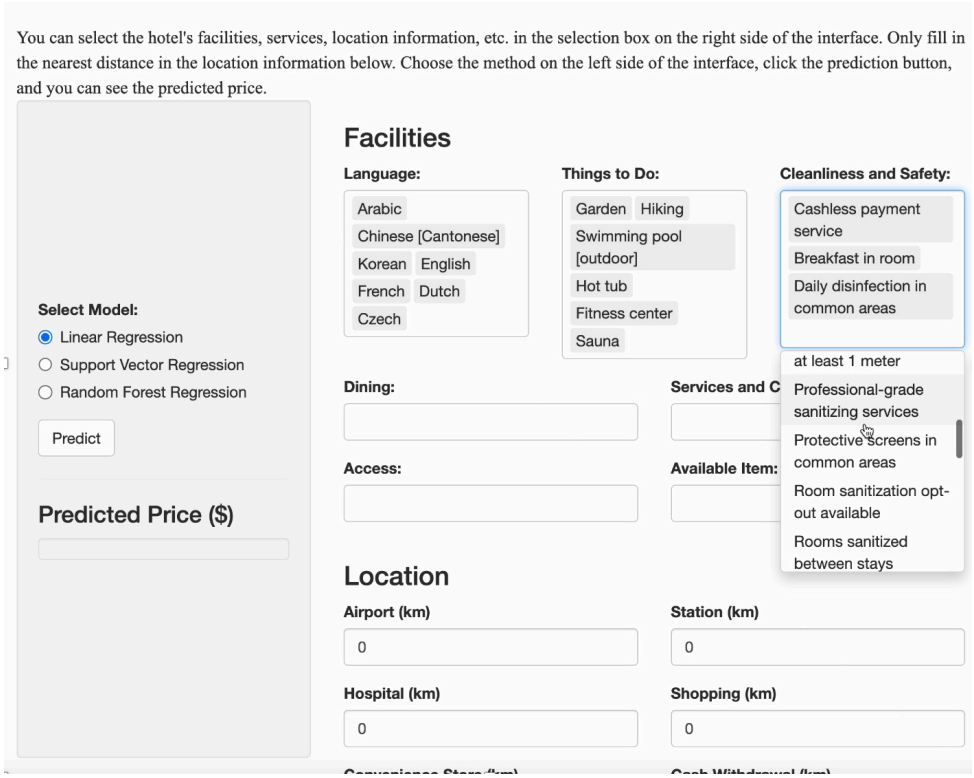


Figure 18: 价格预测

7 总结与反思

在本研究中，我们依托洛杉矶 1388 家酒店的综合数据——包括设施、服务和地理位置等因素，开发了一个酒店价格预测模型。此模型不仅深化了我们对洛杉矶酒店行业现状的认识，还旨在提供一个科学的定价参考，以支持业内的决策制定。我们综合考量了酒店的多项关键属性，并采用了包括逻辑回归、支持向量机和随机森林在内的多种统计分析和机器学习手段来构建预测模型。通过这些方法，我们成功地映射了酒店特征与其价格之间的关系，并尝试为酒店业提供一个基础但高效的价格设定工具。

我们的研究表明，随机森林在价格预测问题上的准确性最高，其在二分类问题上可以达到 85% 的准确率，在回归问题上可以达到 0.73 的相关系数，可以视为当前问题下

的最优预测模型。此外，我们还利用 Rshiny 开发了一个交互界面，使用户可以根据具体情况调整参数，快速获取并应用我们的定价模型，从而更好地理解酒店市场并进行价格预测。这个工具的直观性和灵活性对于那些希望深入洛杉矶酒店市场的业者、投资者和市场分析师来说是非常有价值的。

尽管我们的模型在价格预测方面表现较好，但我们也认识到了的一些本研究的一些局限性。首先，我们没有考虑时间变量，如节假日和旅游旺季等对酒店价格的影响。其次，在从 Agoda 网站收集数据时，我们发现部分酒店存在信息不全的问题，这些限制可能会影响模型的泛化能力和预测的精确度。

未来的工作将致力于解决这些问题，并进一步完善我们的定价模型。我们可以考虑引入时间序列分析来捕捉价格随时间变化的动态特性，并尝试采用更多样的数据源来丰富我们的数据集，以提高模型的全面性和准确性。通过这些努力，我们希望为洛杉矶酒店市场提供更加科学、合理的定价策略，并帮助酒店业者在竞争激烈的市场中取得成功。

8 小组分工

- 刘苏青：网络爬虫；数据预处理；价格预测；交互界面
- 姚圣泰：网络爬虫；探索性数据分析；交互界面
- 沈宇捷：网络爬虫；数据预处理

9 文件说明

- Capstone_report.pdf 为大作业报告；
- Webscrape.Rmd 为爬虫代码；
- HOTELS.csv 为爬取到的数据文件；
- EDA.Rmd 为探索性数据分析代码；
- Main.R 为价格预测及交互界面代码；