西安电子科技大学网络与信息安全学院
大数据分析与安全 实验报告

班　　级：2118021

学　　号：

姓　　名：

电子邮箱：

指导教师：李 玥

2024 年 10 月 28 日

实验题目：基于纽约 Airbnb 数据的区域租房流量的探究

实验摘要：

本次实验根据数据分析流程及探索性数据分析（EDA）方法，分析了各个区域的流量，并通过可视化展示了实验结果，最终得出了以下结论：曼哈顿和布鲁克林是流量较高的区域，尤其是低价的私人房间和整套公寓较受欢迎。同时，价格较高的房源在各区域流量相对较低。我们还发现布朗克斯和斯塔滕岛的房源尽管数量较少，但流量较为集中，可能与房源类型和价格定位有关。

题目描述：

采用数据分析流程及探索性数据分析（EDA）方法，探究哪些区域的流量比其他区域大，分析原因，并使用可视化进行成果展示。

实验内容：

一 **实验目的（包含自己提出的研究假设或洞见）**

探究哪些区域的流量比其他区域大，，分析原因并使用可视化进行成果展示

二 **实验步骤（可依据讲授的数据分析流程与方法）**

1. 数据加载与初步查看

- 加载数据并检查数据结构，包括缺失值、数据类型、基本统计描述等
- 观察数据的地理分布，初步了解不同区域的出租信息分布情况

2. 数据清洗与处理

- 检查数据中的缺失值，并决定是否填补或删除
- 处理与研究问题相关的字段

3. 各区域房源数与评论数分析

- 分析不同 **neighbourhood_group**（区域组）和 **neighbourhood**（街区）的房源数量
- 统计不同区域的 **number_of_reviews**（评论数量）均值，评论数量可以作为流量的一个指标

4. 流量密集区域的地理分布

- 利用地理数据（**latitude** 和 **longitude**），在地图上显示房源的分布密集度
- 使用不同颜色或大小的点标示高流量（高评论数量）的房源，以识别流量密集区域

5. 流量差异分析：房间类型、价格与最低入住天数

- 分析不同房间类型（**room_type**）在不同区域的分布和流量差异，观察哪些房间类型更受欢迎
- 检查价格（**price**）对流量的影响，是否有价格较低的房源流量更大
- 分析最低入住天数（**minimum_nights**）对流量的影响，尤其是一些热门区域可能有更短的最低入住要求

6. 结论

- 总结各区域的流量情况，分析哪些区域的流量高并尝试解释原因
- 根据不同区域的房源类型、价格和流量特征提出进一步的洞察，例如某些区域的低价短期房源更具吸引力等

三 实验结果与分析（重要可视化结果可在此贴图）

1. 数据加载与初步查看

在此步骤中，我们对数据集进行了初步探索，主要集中在数据的基本结构、缺失值情况以及数值字段的统计特征。

1) 数据结构

数据包含 48895 条记录和 16 个字段。

其中字段包括房源 ID、房源名称、房东 ID、房东名称、所在区域组（如布鲁克林、曼哈顿等）、具体街区、经纬度、房间类型、价格、最少入住天数、评论数量等信息；字段数据类型包括整数、浮点数、字符串。

2) 缺失值

数据集的缺失值主要集中在“last_review（最近一次评论日期）”和“review_per_month（每月评论数量）”，这两个字段均有 10052 个缺失值。由于这两个字段仅影响部分记录和评论信息，因此在分析时可以填充和忽略这些缺失值。除此之外，其他字段比如“房源”和“房东名称”也有少量缺失，但在

分析时一般不构成影响

3) 数值字段的基本统计信息

• 价格(price)：价格分布较为广泛，最小值为 0，最大值为 10000

• 最少入住天数(minimum_nights)：最小值为 1，最大值高达 1250 天，平均为 7 天

• 评论数量(number_of_reviews)：评论数量最高为 629 条，显示出部分房源的流量较高

• 每月评论数(reviews_per_month)：平均为 1.37 条，最大值为 58.5 条，分布较为分散

• 房东房源数量(calculated_host_listings_count)：单个房东房源数量最高可达 327 个，显示出部分房东在平台上拥有大量房源资源

2. 数据清洗与处理

在此步骤中，我们将集中处理数据中的缺失值和极端异常值，以确保后续分析的准确性。主要清洗措施包括：

1) 缺失值处理：

• reviews_per_month：此字段的缺失可能是因为某些房源没有收到评论，因此选择将缺失值填充为 0，表示没有月度评论。

• last_review：此字段缺失的原因同样是由于房源没有评论，因此可以在分析时忽略该字段。

2) 异常值处理：

• 价格(price)：价格的最小值为 0，最大值为 10000，有明显的极值。我们删除价格为 0 的数据，因为这很可能是数据录入错误。

• 最少入住天数(minimum_nights)：最大值为 1250 天，这一值显然不符合一般出租的需求。因此我设定了一个阈值 100 天，将高于 100 天的数据全部过滤，使结果更具普遍性。

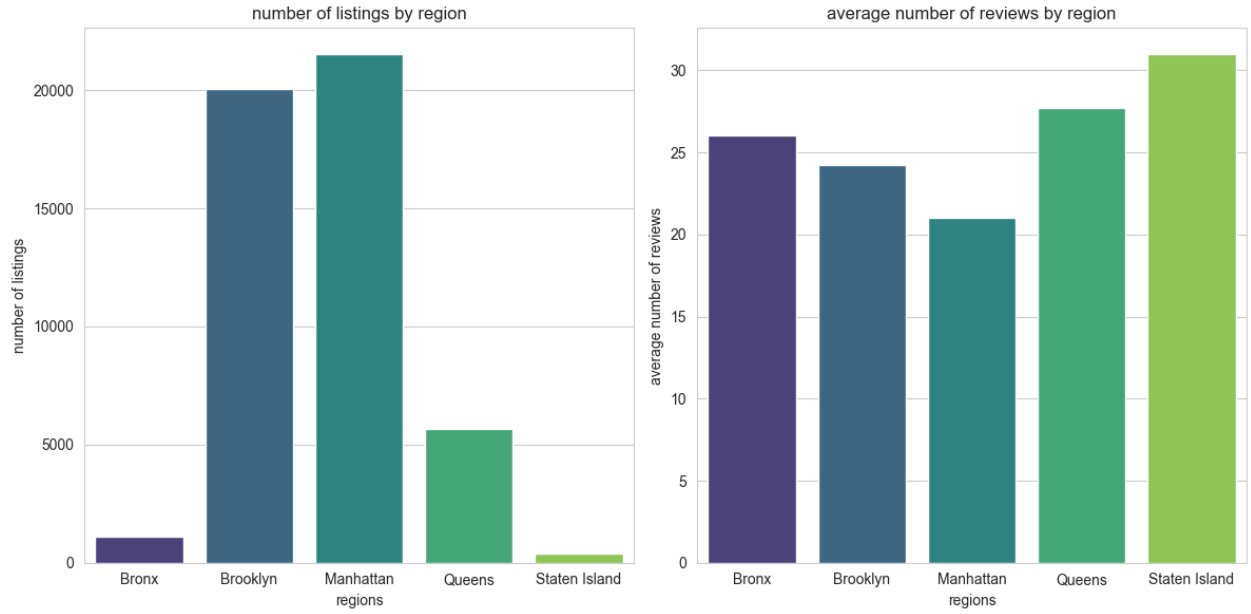
3. 各区域房源数与评论数分析

此步骤中，我分析了不同区域的房源数量和评论数，以便识别流量较高的区域。具体来说：

1) 房源数量：统计每个区域组(neighbourhood_group)中房源数量，展示不同区域的房源集中度。通过该分布，可以识别出曼哈顿和布鲁克林拥有最多的房源资源，分别为 21660 和 20095.

2) 评论数量：计算每个区域中的平均评论数量，以此衡量房源的受欢迎程度或流量高低。区域平均评论数量的对比可以帮助确定哪些区域的房源更受欢迎，也代表了更高的流量。可以看出斯塔滕岛尽管房源数较少（373），但评论均值最高（30.94），标明其可能拥有较高的流量；布朗克斯和皇后区的评论均值也较高，分别为 25.98 和 27.70。

这一步骤揭示了不同区域的房源几种情况及其流量分布，为后续的地理和房源属性分析奠定基础。



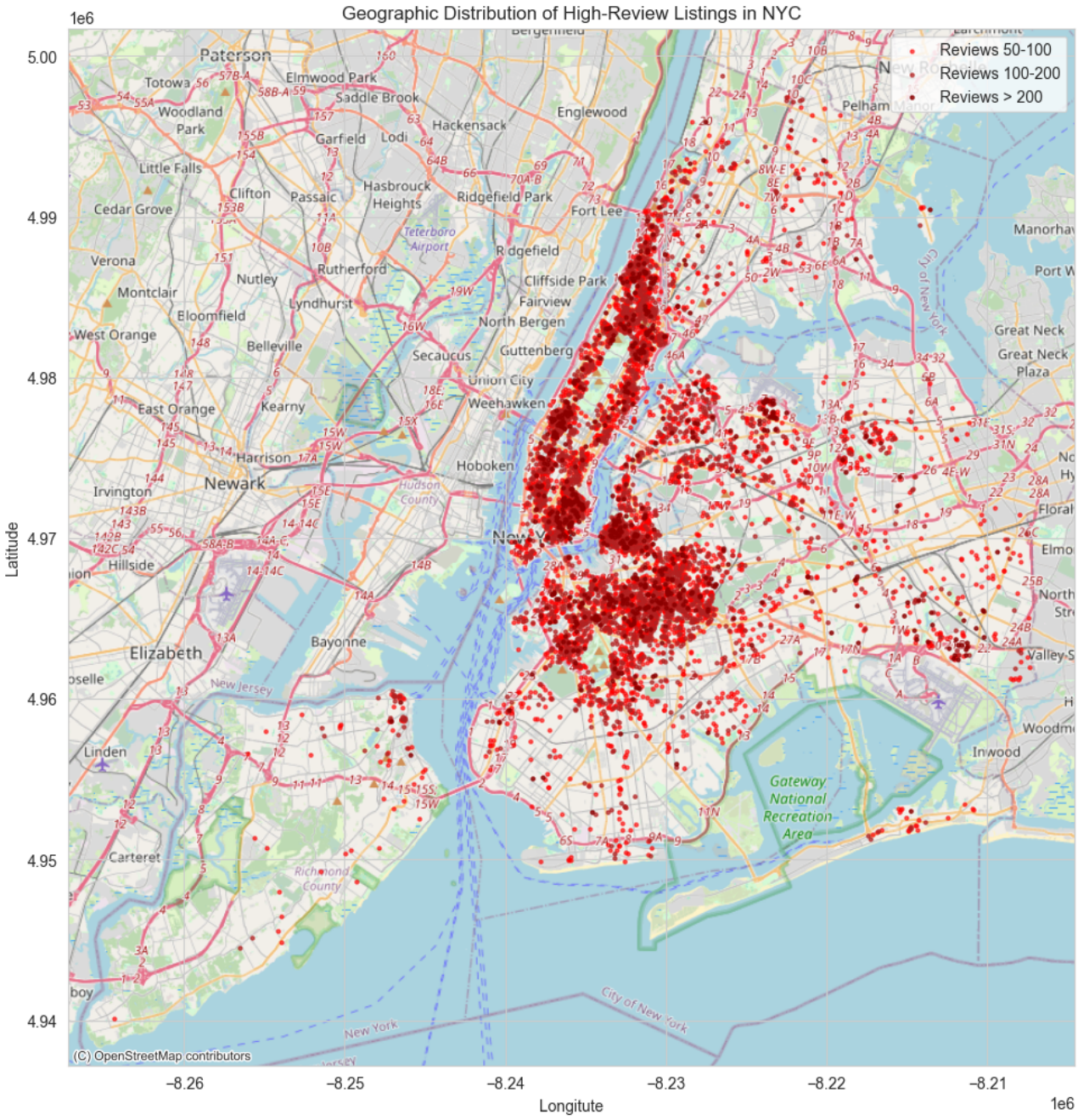
4. 流量密集区域的地理分布

在此步骤中，我通过分析流量密集区域的地理分布，展示纽约市各区域房源流量的空间分布情况，具体如下：

1) 流量等级划分：将房源评论数量划分为三个等级：50-100(低流量)、100-200(中流量)、200 以上(高流量)。其中 50 以下的房源不做考虑。评论数量是衡量房源流量的一个重要指标，评论越多，通常意味着房源的访问量和关注度越高

2) 地理分布展示：通过 python 的 geopandas 和 contextily 库，在实际地图上通过散点图展示流量密集的区域

从地理分布图中，曼哈顿和布鲁克林的房源分布密集，且大部分是高流量房源，标明这些区域的流量需求较高。皇后区和布朗克斯的房源流量相对较低，但在一些特定区域仍有中等流量房源，可能是局部的热点位置。



5. 流量差异分析：房间类型、价格与最低入住天数

在此步骤中，我们分析房源的 房间类型、价格 和 最低入住天数 对流量（评论数量）的影响，以确定哪些因素能够提升房源的关注度。具体分析内容如下：

1) 房间类型与流量的关系

• 不同房间类型可能吸引不同的客户群体，房间类型的差异可能直接影响房源流量

• 结果：在不同区域内，“整套公寓”和“私人空间”的流量差异较大。布鲁克林和布朗克斯的私人房间平均评论数量更多，标明这些区域内私人房间需求较

高，而在曼卡顿，整套公寓的流量则更为集中

2) 价格对流量的影响

• 价格是选择房源的重要因素，较低的价格通常能吸引更多预定。

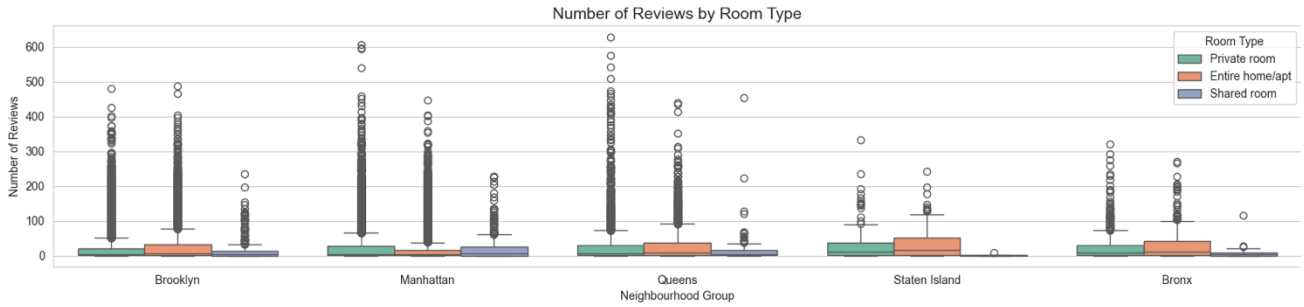
•结果：价格与评论数量之间并没有明显的线性关系，但在价格较低区间（0-500 美元），房源的评论数量较多，尤其是布鲁克林和皇后区。价格超过 500 美元的房客评论数量逐渐减少，标明高价位房源较难吸引高流量。

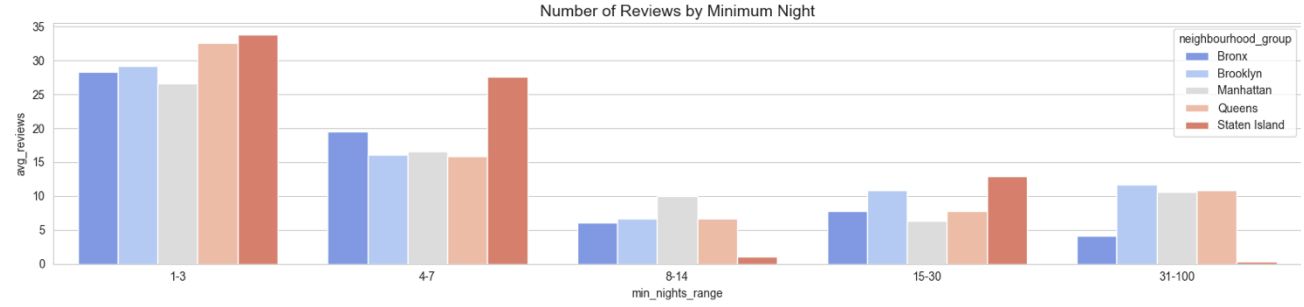
3) 最低入住天数与流量的关系

• 最低入住天数直接影响房源的订购灵活性

• 方法：将最低入住天数分为 5 个区间（1-3、4-7、8-14、15-30、31-100），通过对比平均评论的方法，分析不同入住要求对流量的影响

• 结果：低入住天数（1-3 天）的房源拥有较高的平均评论数量，尤其在曼哈顿和布鲁克林等热门地区。随着最低入住天数增加，平均评论数量逐渐减少，这表明较低的最低入住天数有助于提升流量





6. 结论

1) 区域流量特征

• **曼哈顿和布鲁克林**是流量最高的区域，特别是曼哈顿的中城区和布鲁克林的某些街区，他们的房源的平均评论数量显著高于其他区域。

• **皇后区和布朗克斯**的房源数量较少，但一些特定地点仍有较高流量。

2) 各种因素对流量的影响

房间类型：私人房间在布鲁克林和布朗克斯的流量较高，表明这些区域更受单人游客和短期旅客欢迎。整套公寓在曼哈顿更为热门。

价格：中低价位（0-500 美元）的房源在各区域普遍流量较高。价格较低的房源更易吸引游客，尤其是在布鲁克林和皇后区。高价位房源（超过 500 美元）的评论数量明显较少，这可能是由于价格较高，客群更加小众，预订频率相对降低。

最低入住天数：较短的最低入住要求（1-3 天）显著提升了房源流量，尤其在游客密集的曼哈顿和布鲁克林区域更为显著。较长的入住要求（如 30 天以上）房源的评论数量偏低。

实验总结：

在画图时，使用 matplotlib 会出现图片的中文乱码。查阅资料以后，发现可以通过设定 matplotlib 库的默认字体可以改变。但我试了以后还是不行，解决不了问题我选择解决问题，于是图片全都使用了英文注释。

使用地图库时，发现老教程中给的 Stamen.TonerLite 的地图源不可用，因此换成了 OpenStreetMap 的地图源，成功在地图上实现了标点。

参考文献：

1. Seaborn Documentation: https://seaborn.pydata.org/
2. GeoPandas Documentation: https://geopandas.org/
3. Contextily Documentation: https://contextily.readthedocs.io/
4. 知乎用户. (2020). 彻底解决 Python 里 matplotlib 不显示中文的问题. 知乎专栏. https://zhuanlan.zhihu.com/p/104081310