探索性数据分析

课程名称：数据可视化导论

学生姓名：姜雨童

专　　业：计算机科学与技术

学　　号：3220103450

浙江大学　*2024.11.30*

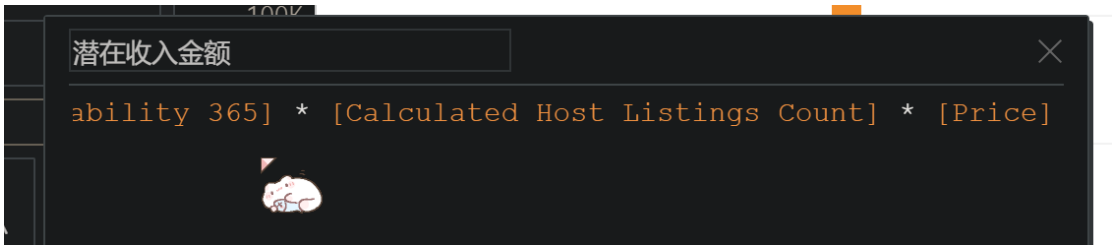
探索性数据分析 1——纽约 Airbnb 数据挖掘

自 2008 年以来，Airbnb 使游客和房东出行更方便，提出更多个性化的体验世界的方式。该数据集包含有关 2019 年纽约出租的信息以及包含其地理信息，价格，评论数量等。

1. 哪些区域生意最好，为什么？

数据集中包含了房源可用性（Availability 365）、出租房源数量（Calculated Host Listings Count）、最小订房天数（Minimum Nights）、评论数量（Number Of Reviews）、价格（Price）等信息，我们将主要从这几个方面来分析。

- 首先，由于这里只给出了评论数量而没有具体区分是好评还是差评，而一般情况下只有在完成交易后顾客才会评论，因此我们简单认为**评论数量**多的房源生意更好。
- 其次，生意好的情况下，租房价格更有可能上涨，而且高房租的情况下，交易次数少仍可以有较高的成交额，因此**价格**高的房源生意也可能更好。
- 另外，数量充足，可订天数多的房源更有概率做成生意，因此我们也可以用 Availability 365 * Calculated Host Listings Count * Price 来粗略估计**潜在交易金额**（创建计算字段，见下面两张图），作为辅助评估指标。

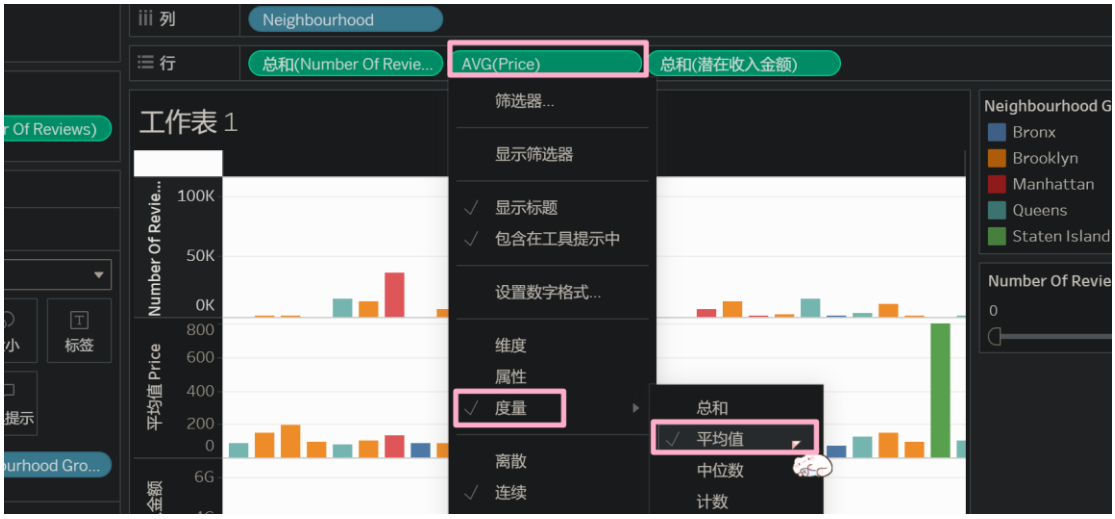


下面是根据区域划分大小进行的具体分析。

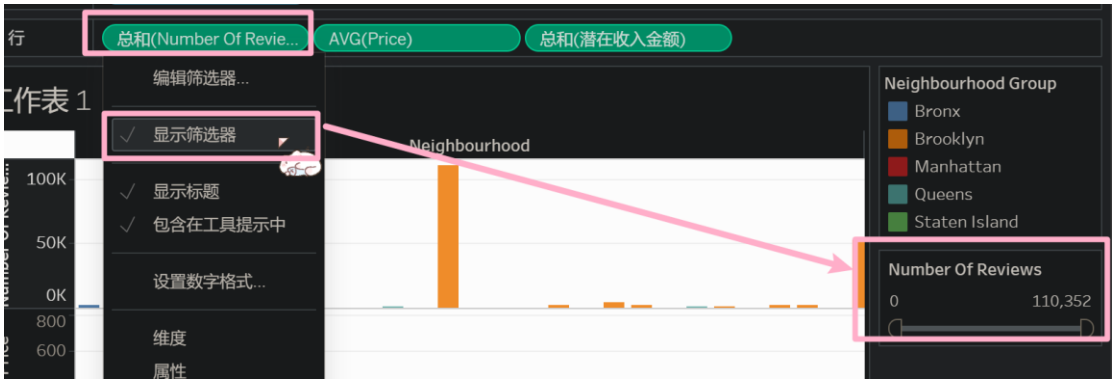
从较小的区域划分（Neighbourhood）来看：

通过拖动相应的属性块来布置视图，比如每一列表示一个 Neighborhood 的数据（粉色框），三个柱状图分别对应评论数量、价格和潜在交易金额（蓝色框），Neighborhood Group 则以颜色来区分（橙色框），使结果更清晰。

某一区域的评论数就是各房源评论的总数，因此可以采用默认的“总和”指标。但是不同区域的房源数量不一定相同，因此衡量房源价格时，不能简单用总和，而是使用“平均数”这一指标（或采用“中位数”）：



最后，由于小区域的数量较多，难以在视图内完整呈现，需要使用筛选器过滤掉生意较差的房源。这里我们以评论数为主要分析指标，房源平均价格为次要分析指标，潜在交易金额为辅助分析指标（更多用于验证），因此只对上述两个指标做筛选。



适当过滤掉评论过少/房源价格特别低的区域后，我们得到如下结果：

其中粉框内为推测生意**最好**的区域（Bedford-Stuyvesant，Chelsea，Harlem，Hell's Kitchen，Midtown，Upper East Side，Upper West Side，Williamsburg），绿框内为推测生意较好的区域（Crown Heights，East Harlem，East Village）。

从较大的区域划分（Neighborhood Group）来看：

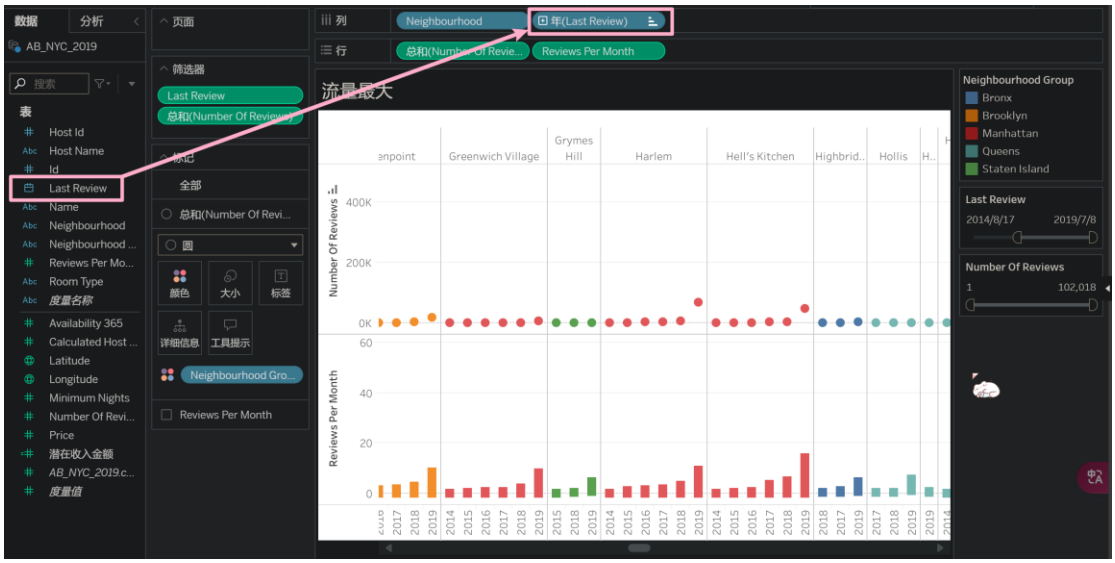
使用类似的方法构建视图（横轴为 Neighborhood Group，Neighborhood 则以大小表示），可以看到 Brooklyn 和 Manhattan 的生意最好。

（作为交叉验证，小区域划分视图时，生意最好的区域均来自橙色标识的 Brooklyn 和红色标识的 Manhattan。）

2. 哪些区域的流量比其他区域大，为什么？

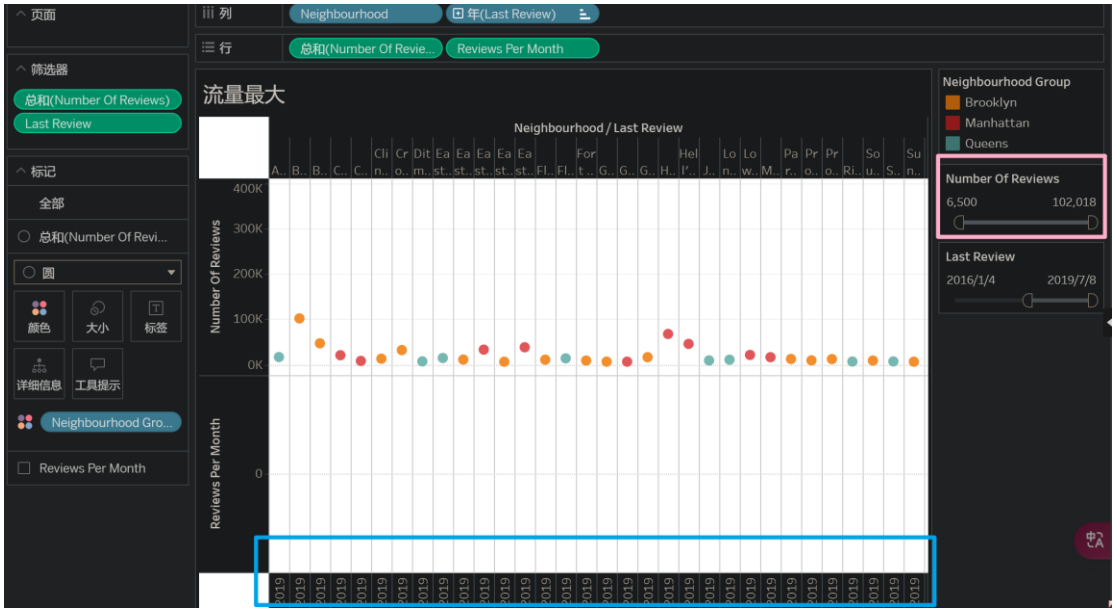
我们主要根据评论总数来估计流量，同时把最新评论时间纳入评估指标，排除近期内没有流量的区域。

类似地，构建视图，并将最新评论时间也作为横轴，第一个视图（上方）为评论总数，第二个视图（下方）为每月评论数（且二者高度相关联）。

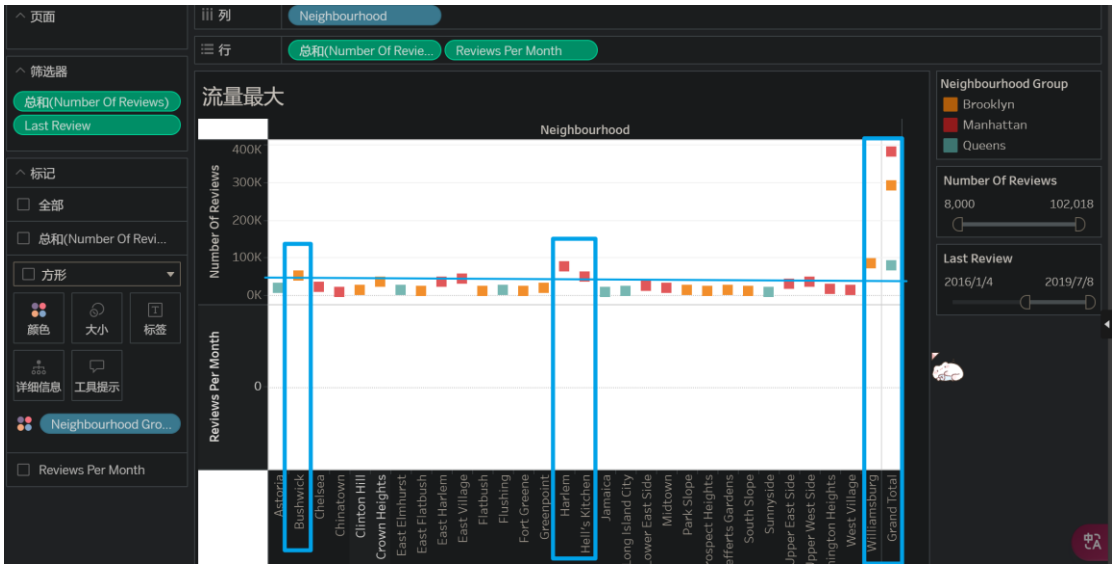


对 Last Riview 使用筛选器，选出最新评论时间在 16 年 1 月以后的房源，过滤那些近期内没有流量的房源：

对 Number of Review 使用筛选器，过滤掉评论数较低的区域（粉框）。可以看到当评论数高于一定程度时，已经自动把最新评论在比较久之前的房源过滤掉了（蓝框）。

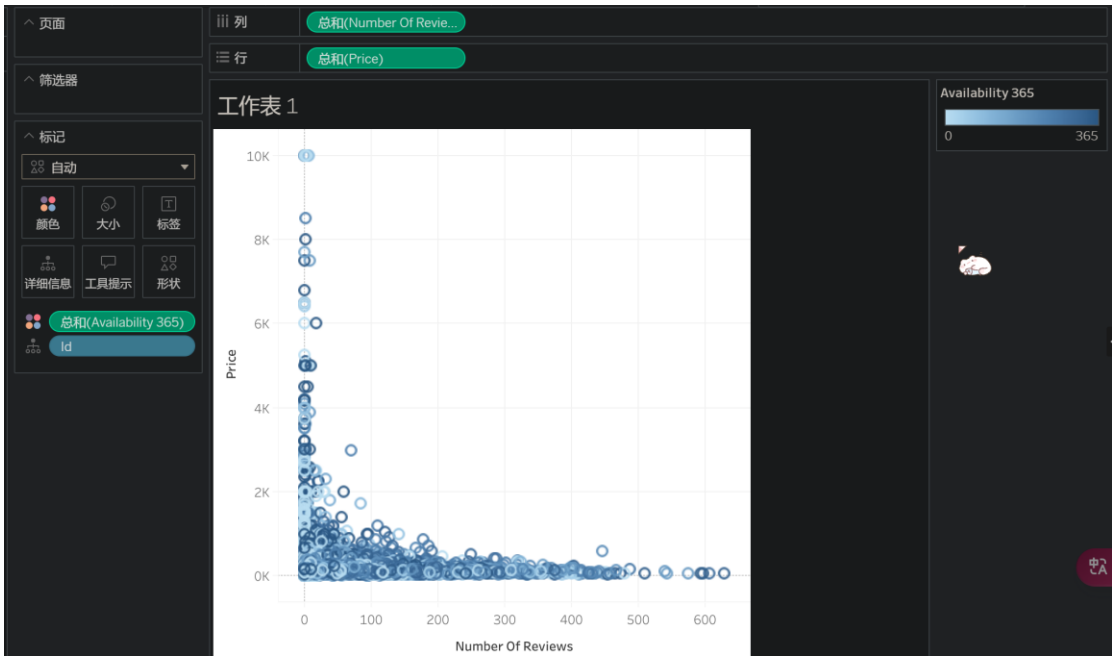


最终得到**流量较大**的区域为：Bushwick，Harlem，Hell's Kitchen，Williamsburg。视图最右一栏为大区域（Neighborhood Group）的数据，可以看到 Brooklyn（橙色）和 Manhattan（红色）的评论数要高于其他区域，因此推测这两个大区域的流量更高。



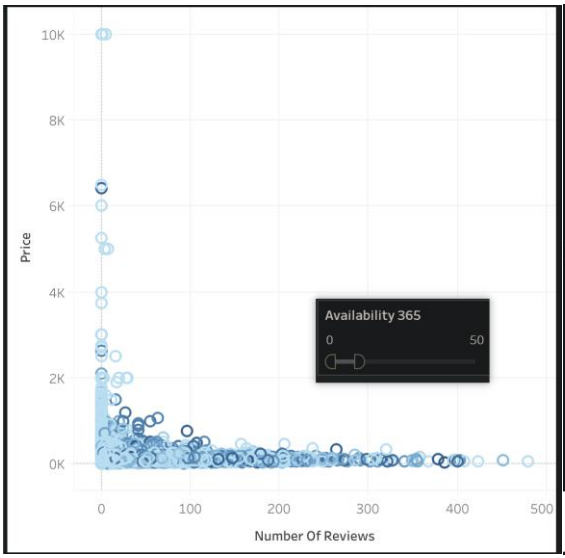
3. 价格，评论数量和预订天数之间是否存在一些关系？

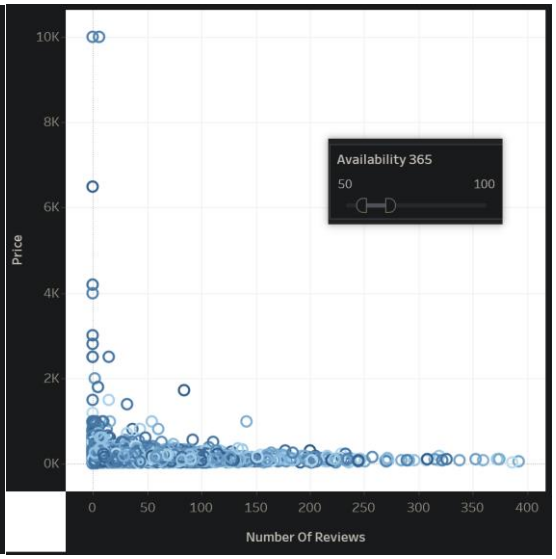
因为要探究几者的关系，我们尽量在一张图中呈现数据。选择构建散点图，每个房间（id）为一个点，x 轴为评论数量，纵轴为房间价格，利用颜色来编码预订天数（深色代表预定天数多）。

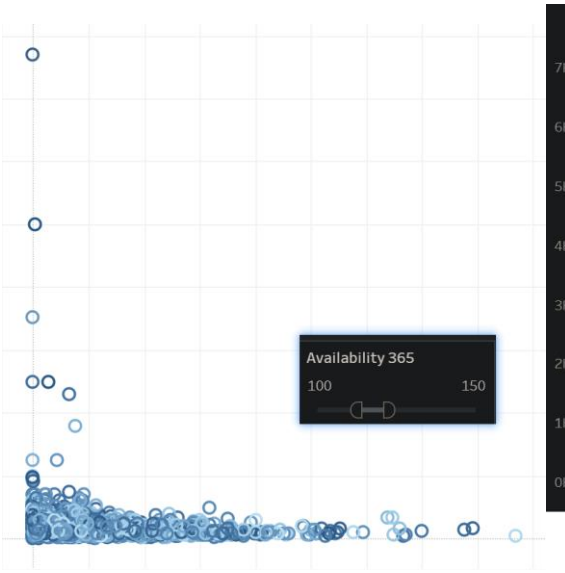


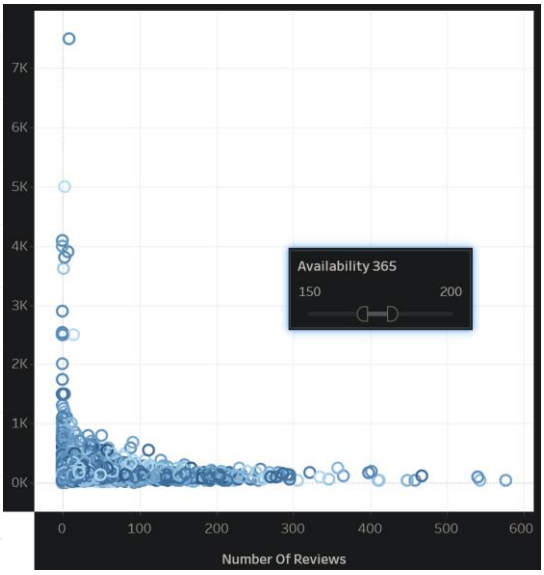
由于数据量较大，这里针对预定天数（Minimum Nights）进行筛选，给出了以 50、100、150 等为限制边界的几段数据展示（下六图）。

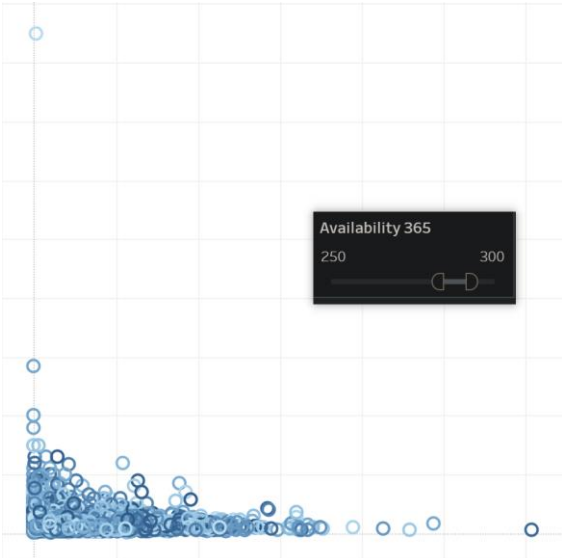
可以看到，房源价格和评论数量近似成反比，价格高则评论相对少。预定天数低时，房源价格普遍不高，且评论数量多；预定天数较高时，有部分高价房源出现，且评论数量相对较少。

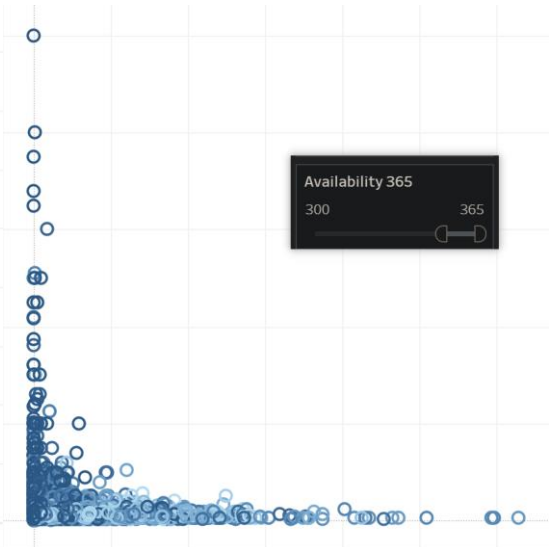

10K
8K
6K
4K
2K
0K
Price
0
100
200
300
400
500
Number Of Reviews
Availability 365
0
50


10K
8K
6K
4K
2K
0K
Price
0
50
100
150
200
250
300
350
400
Number Of Reviews
Availability 365
50
100


Availability 365
100
150


7K
6K
5K
4K
3K
2K
1K
0K
0
100
200
300
400
500
600
Number Of Reviews
Availability 365
150
200


Availability 365
250
300


Availability 365
300
365