

**探索性数据分析**

课程名称： 数据可视化导论

学生姓名： 周炜

专 业： 计算机科学与技术

学 号： 3210103790

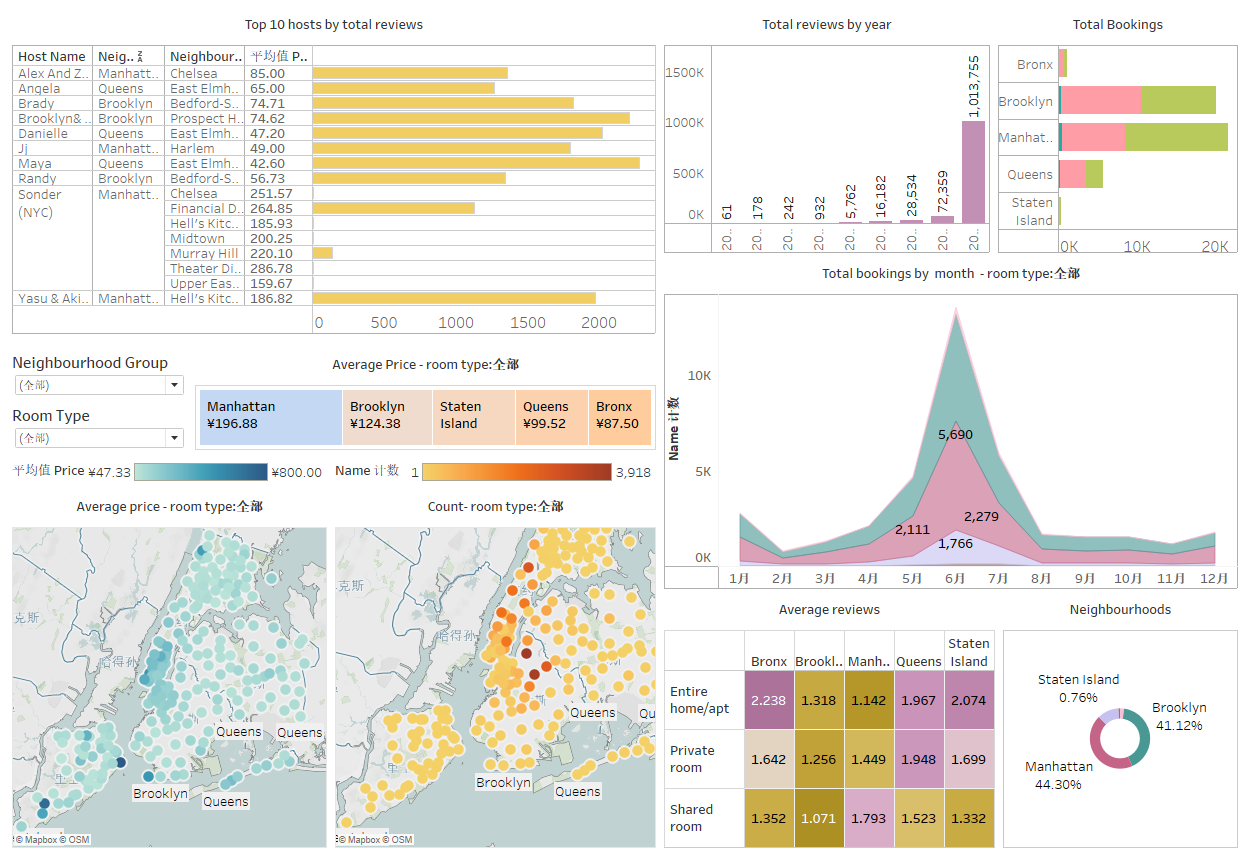
**图片1**

***2023.11.30***

## 探索性数据分析1——纽约Airbnb数据挖掘

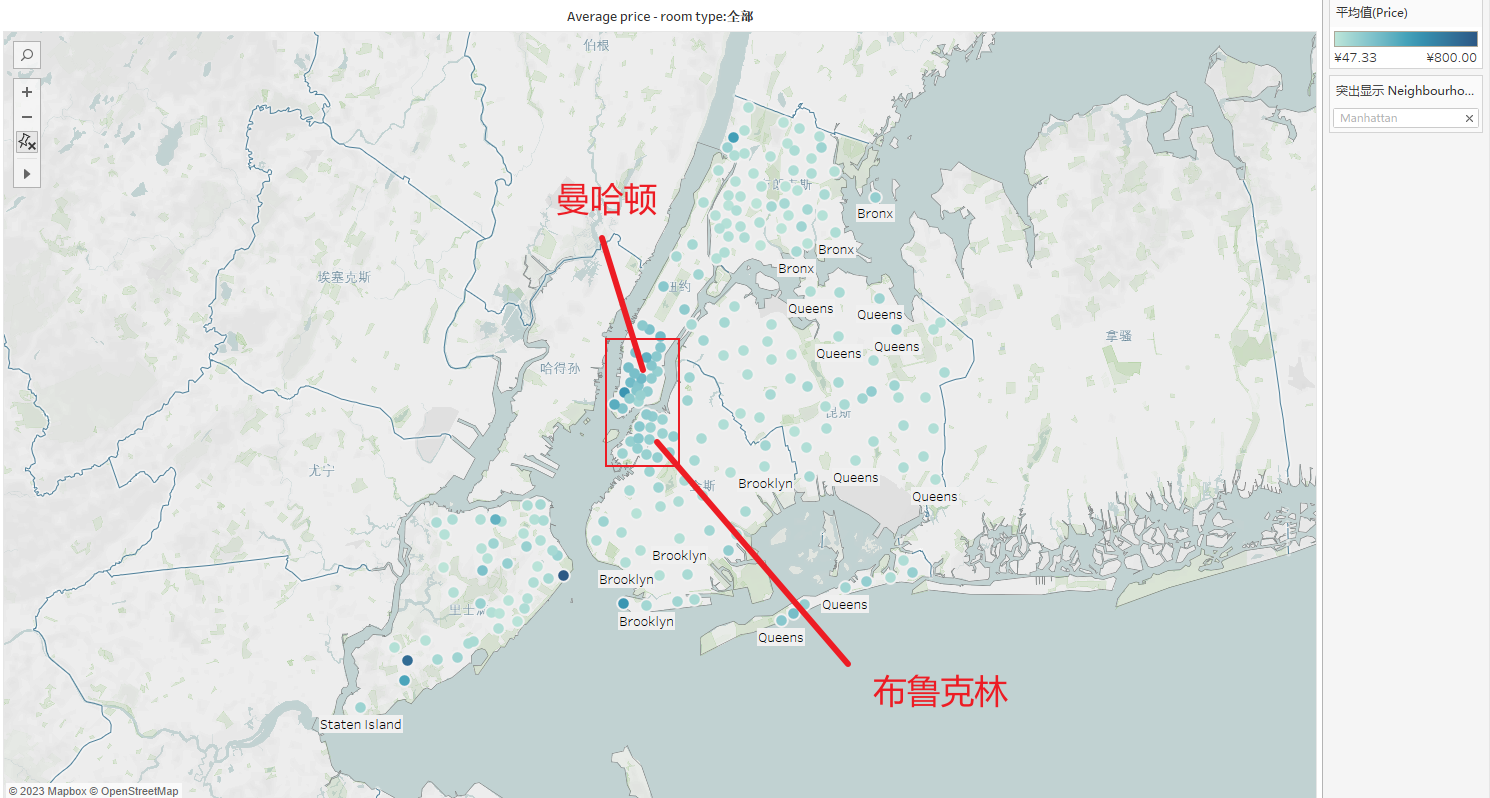
**自2008年以来，Airbnb使游客和房东出行更方便，提出更多个性化的体验世界的方式。该数据集包含有关2019年纽约出租的信息以及包含其地理信息，价格，评论数量等。**

我的作业的可视化程序可见于[**https://public.tableau.com/app/profile/wei.zhou8270/viz/AB-nyc\_17012428488040/Dashboard1?publish=yes**](https://public.tableau.com/app/profile/wei.zhou8270/viz/AB-nyc_17012428488040/Dashboard1?publish=yes)



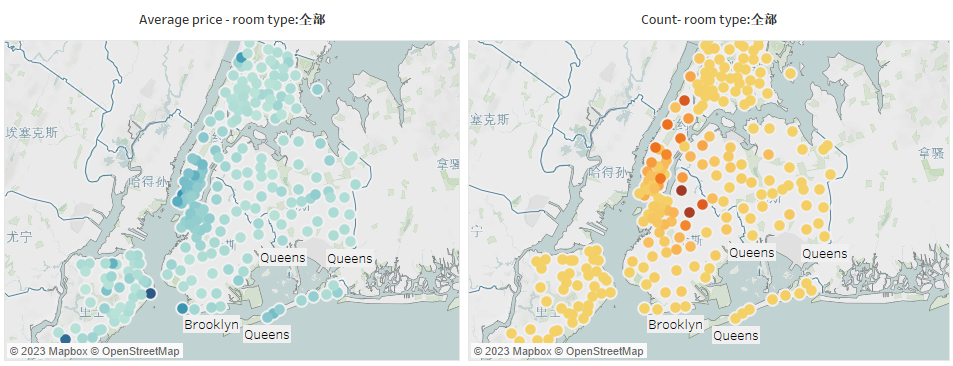
1. **哪些区域生意最好，为什么？**

生意最好的区域大致位于伊斯特河河口两岸的Manhattan和Brooklyn，即大致下列红框的区域，因为这些地方的neighbourhood较为密集，并且均价都较高

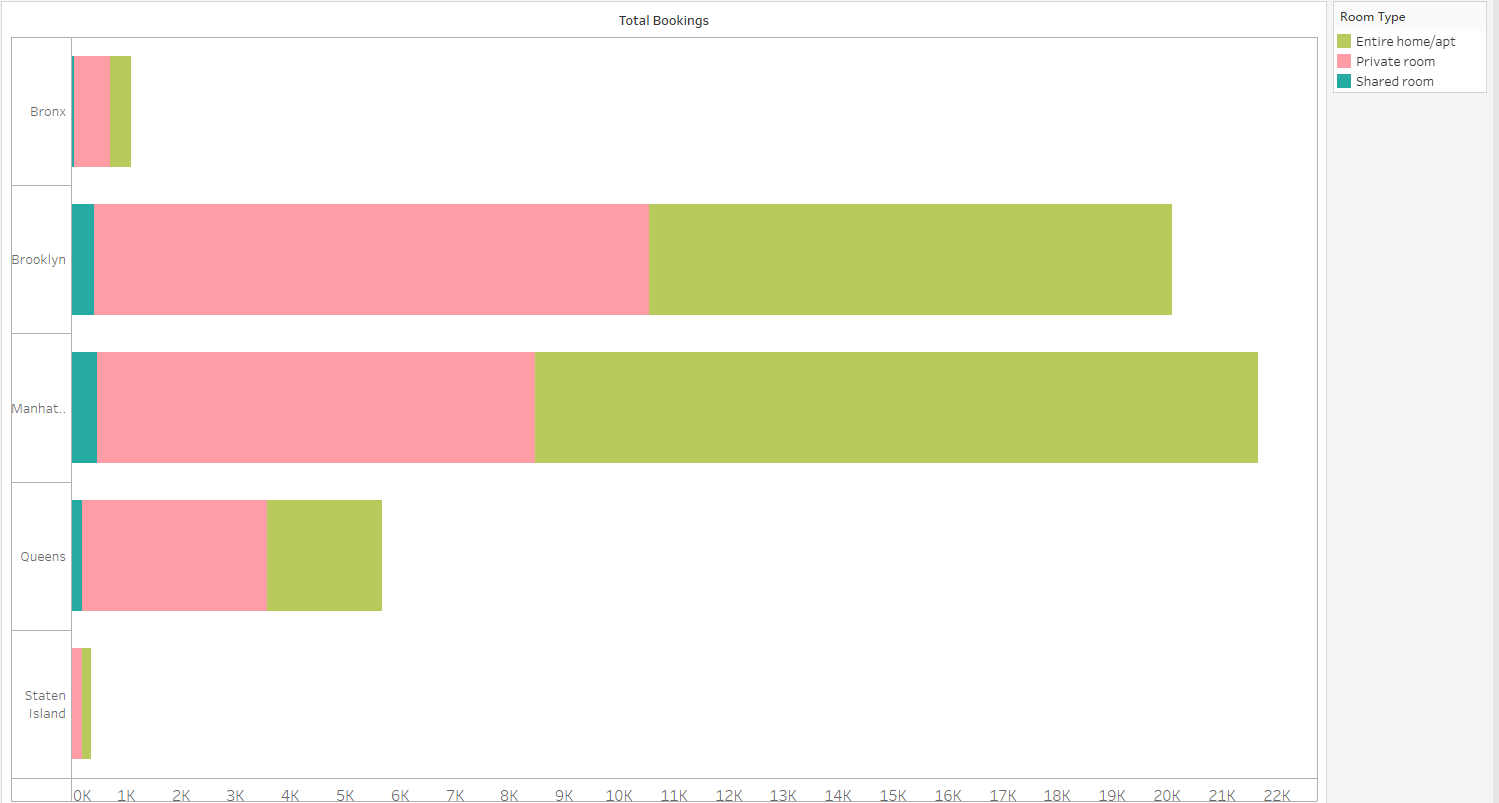


如果五个区来说的话，生意最好的区是Manhattan。Manhattan拥有数据集中53%的bookings，其平均房价是5个区中最高的。

生意的好坏可以通过数量(neighbourhood)和价格进行分析



通过对区域以及room type的计数，可以得到下表



可以发现Manhattan区的交易量最大

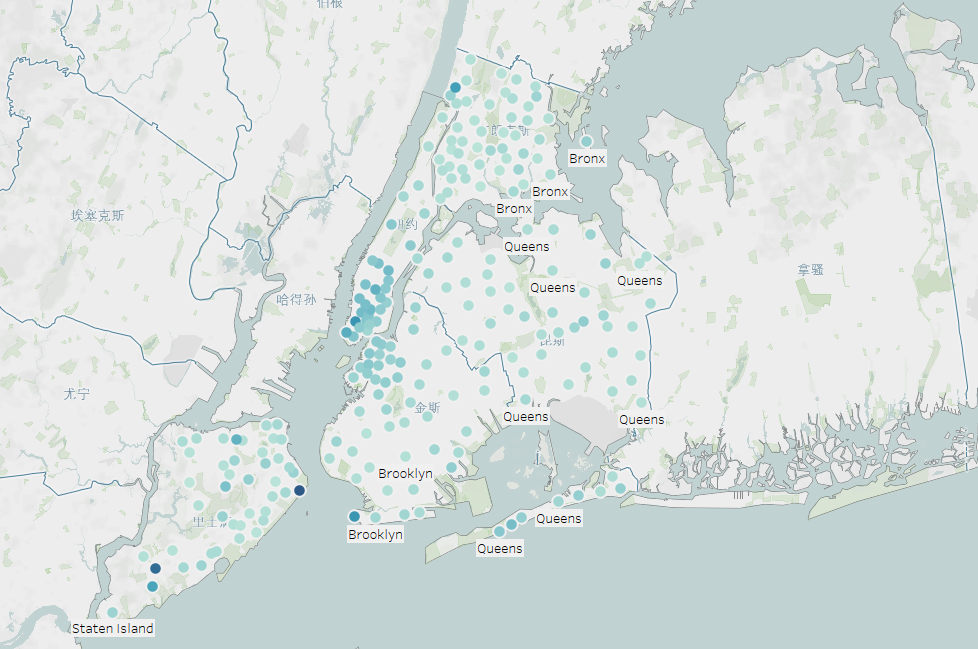
将交易量和价格对比

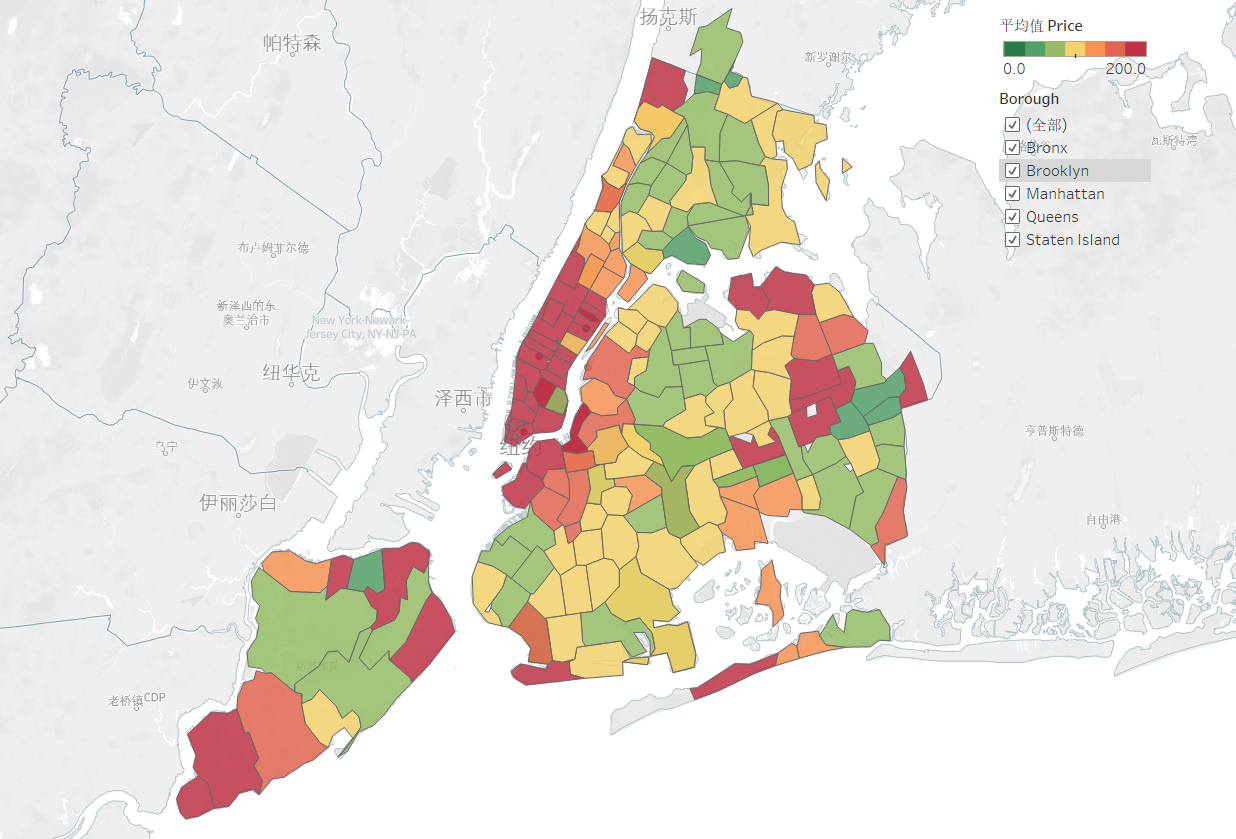
|  |  |
| --- | --- |
|  |  |

并且通过地图的可视化分析，也可以发现Manhattan区的neighbourhood较其他4个区为密集

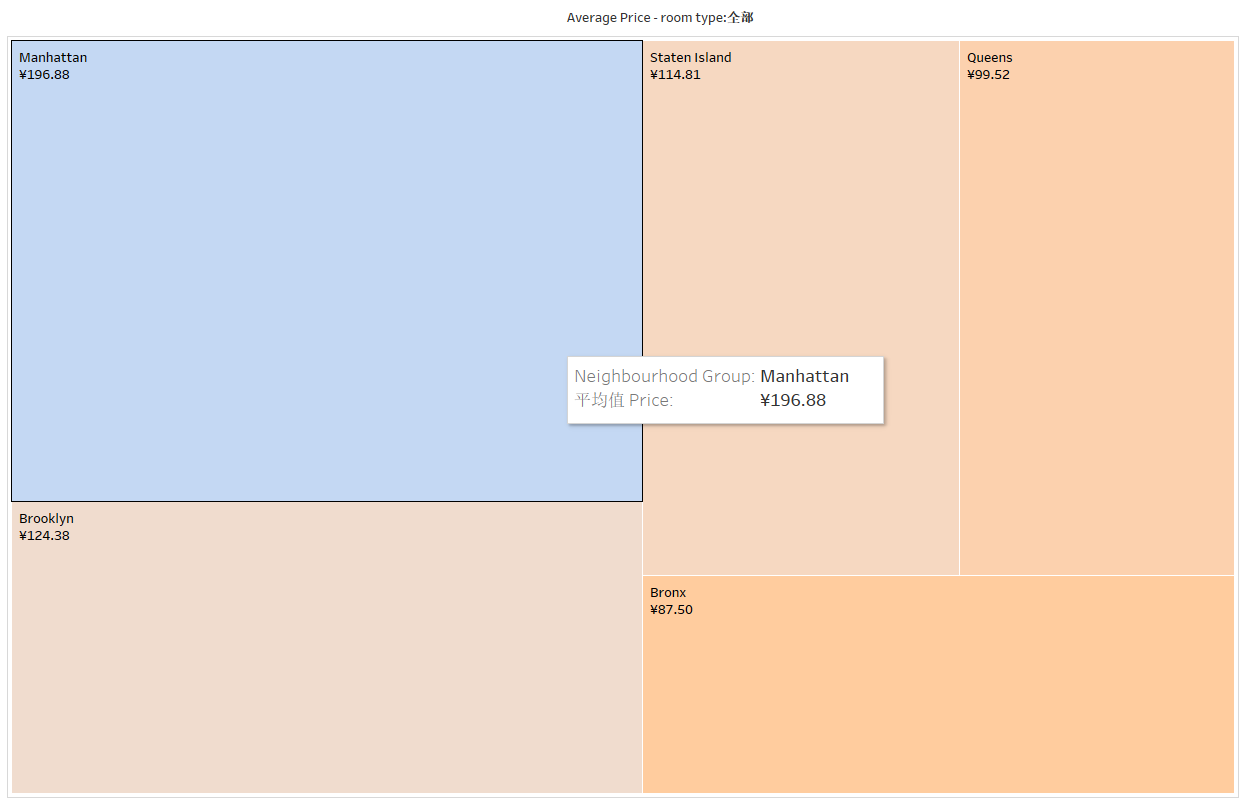


并且均价都比较高



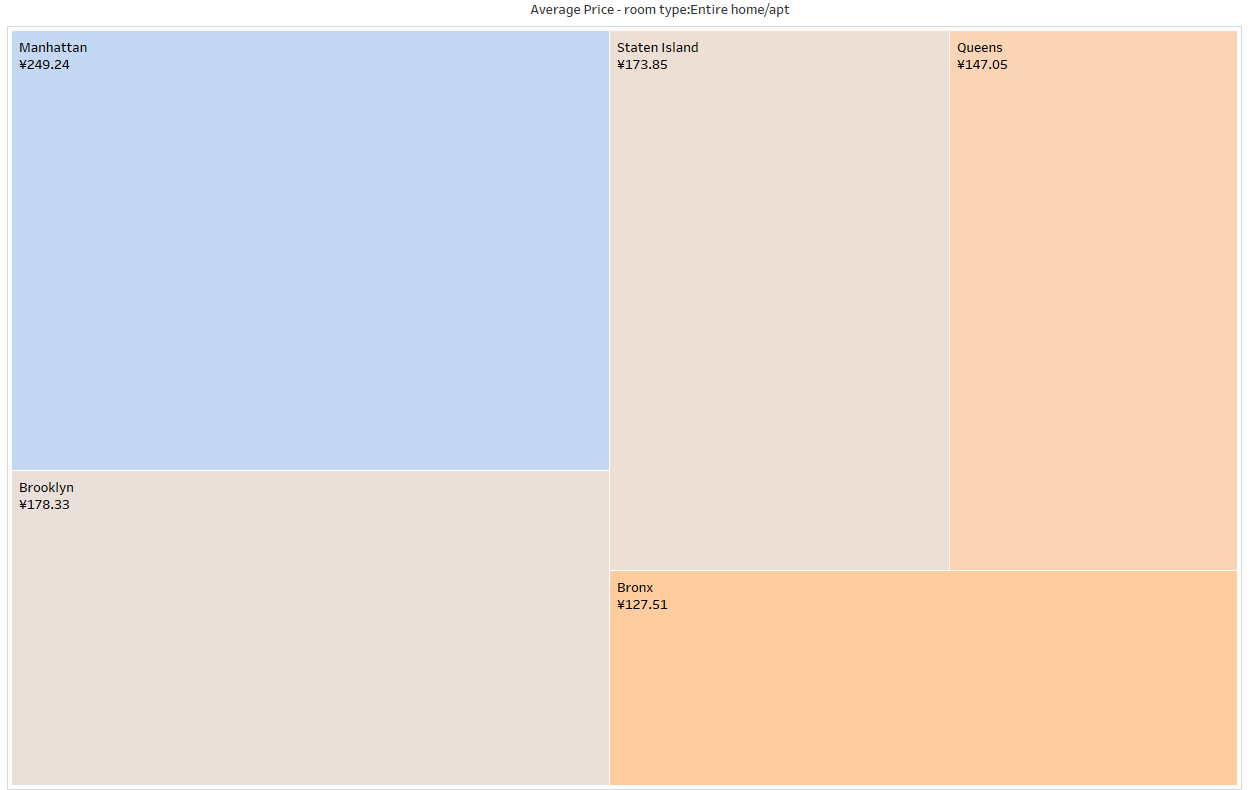


并且通过绘制数量（显示为方块的大小）和价格的方块图，可以发现Manhattan区不仅数量最大，而且均价最高

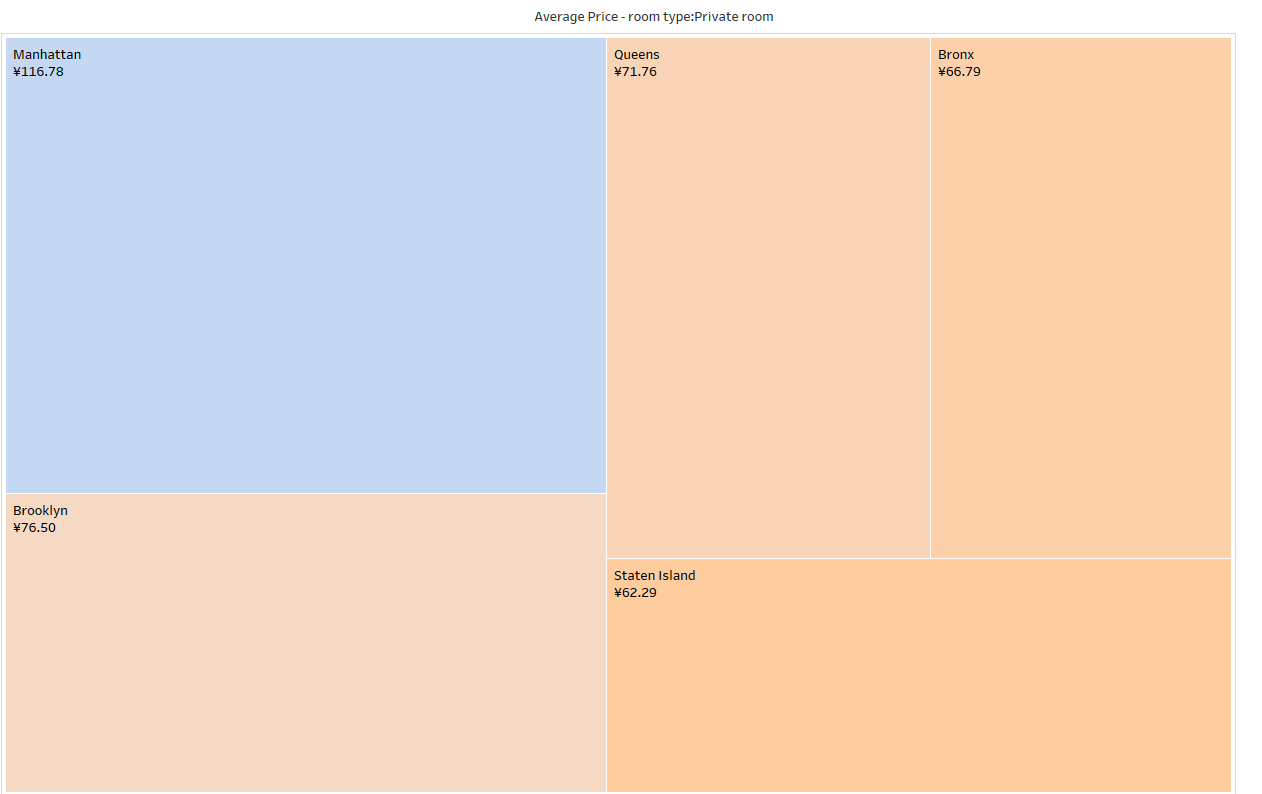


并且在没中房型的表现上都是如此

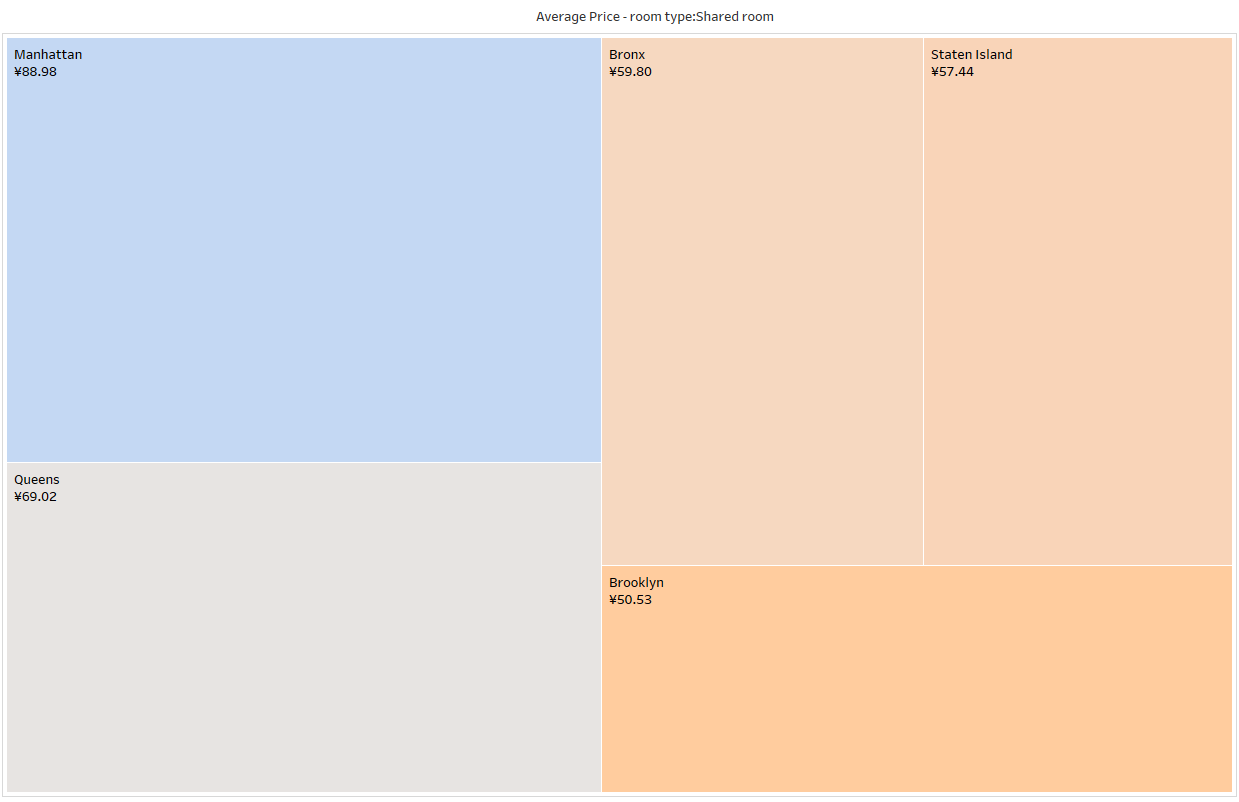
Entire home/apt



Private room

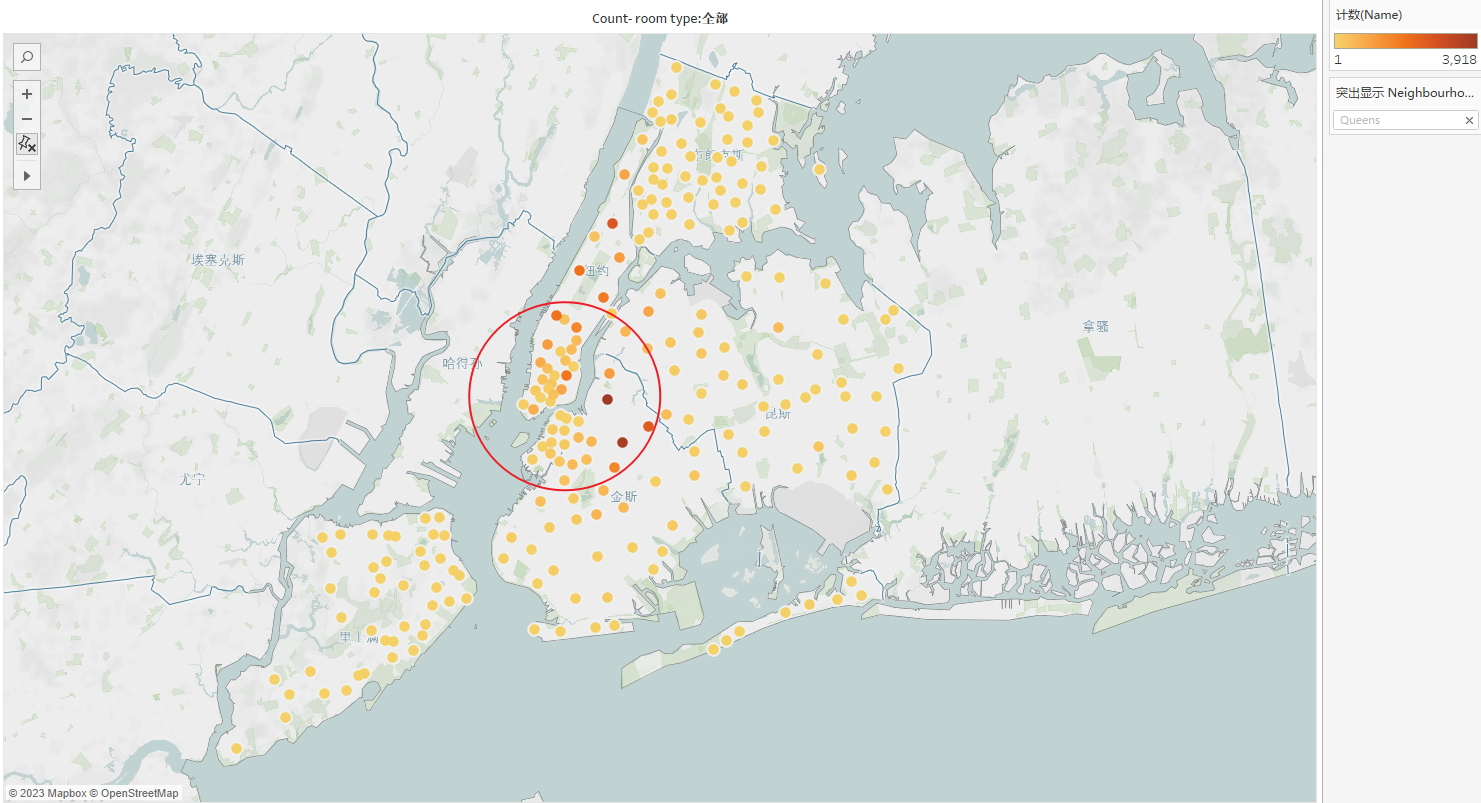


Shared room



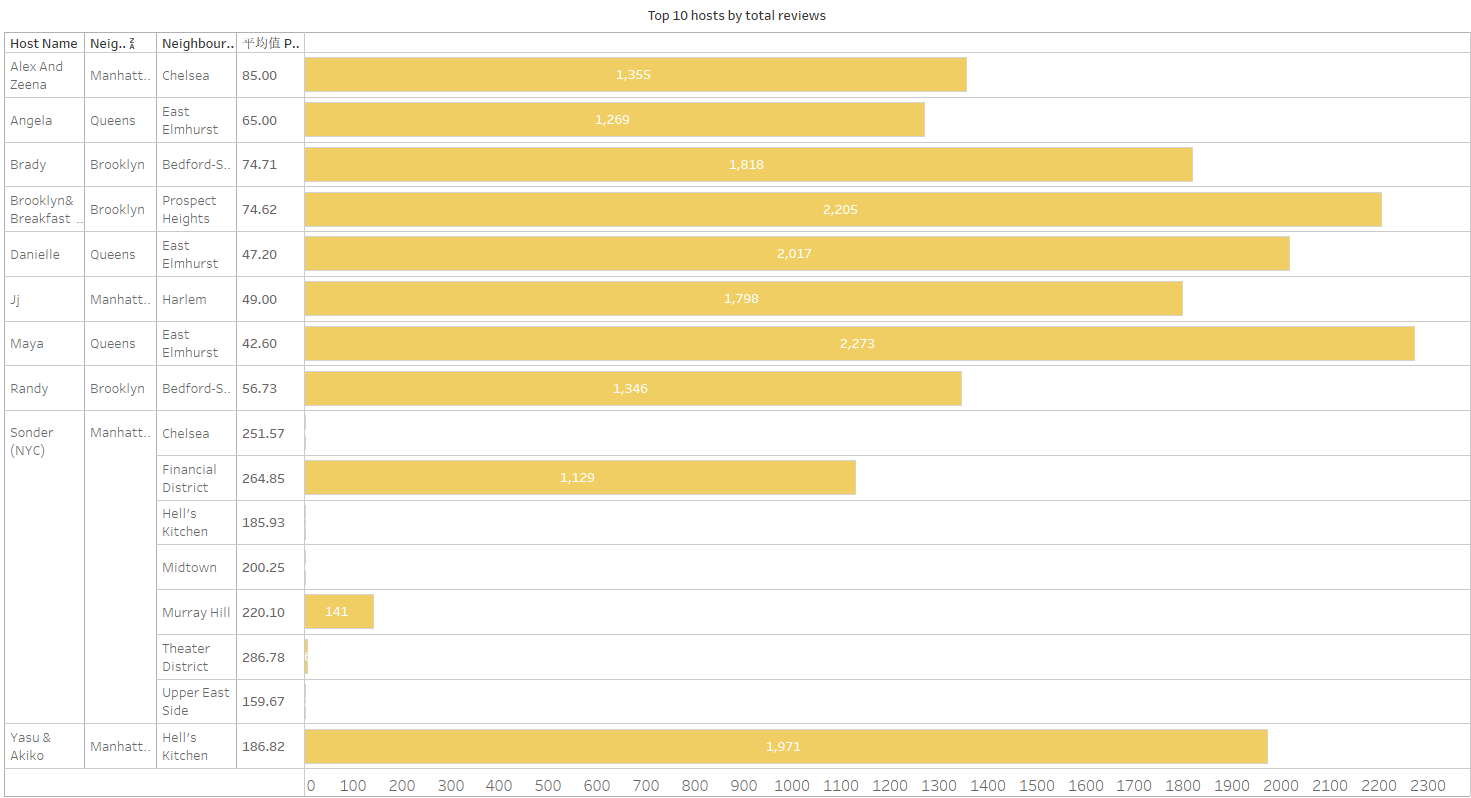
1. **哪些区域的流量比其他区域大，为什么？**

流量大的区域大致如下图所示

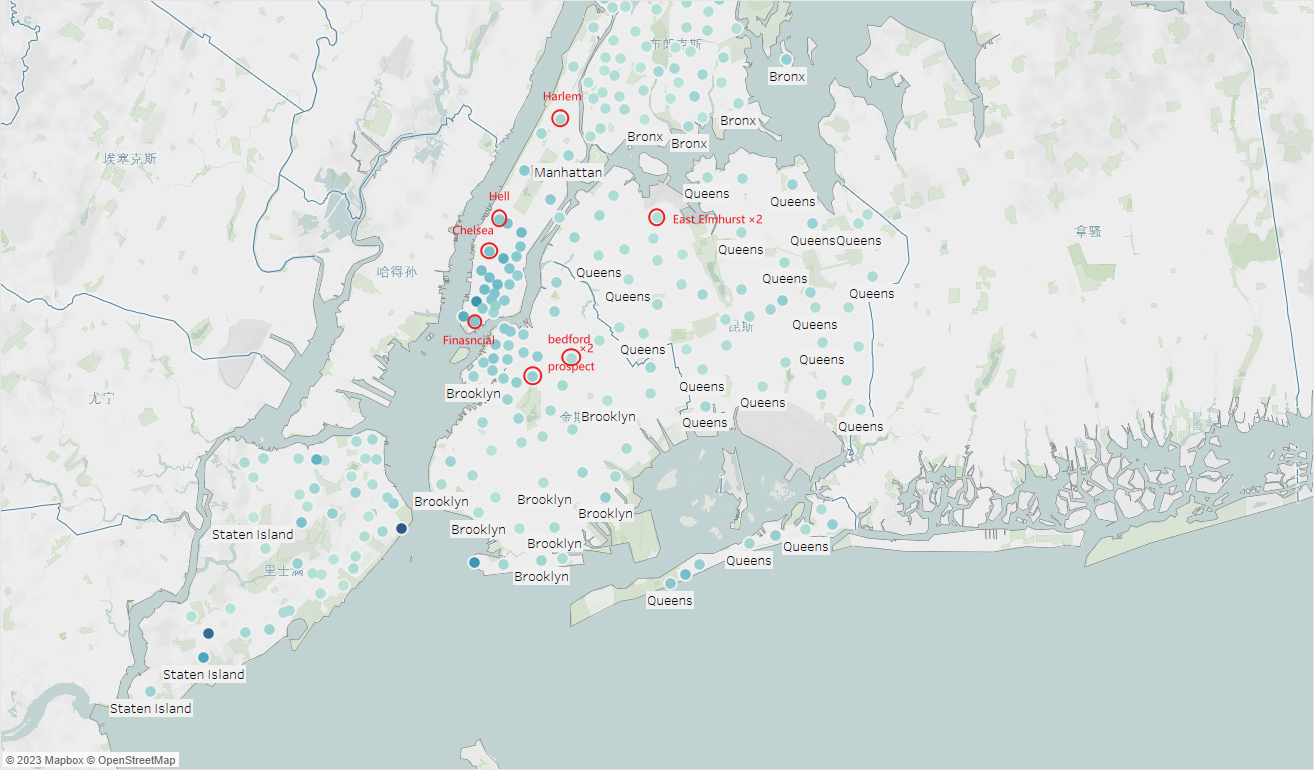


流量可以从neighbourhood的密度和review的数量来观察

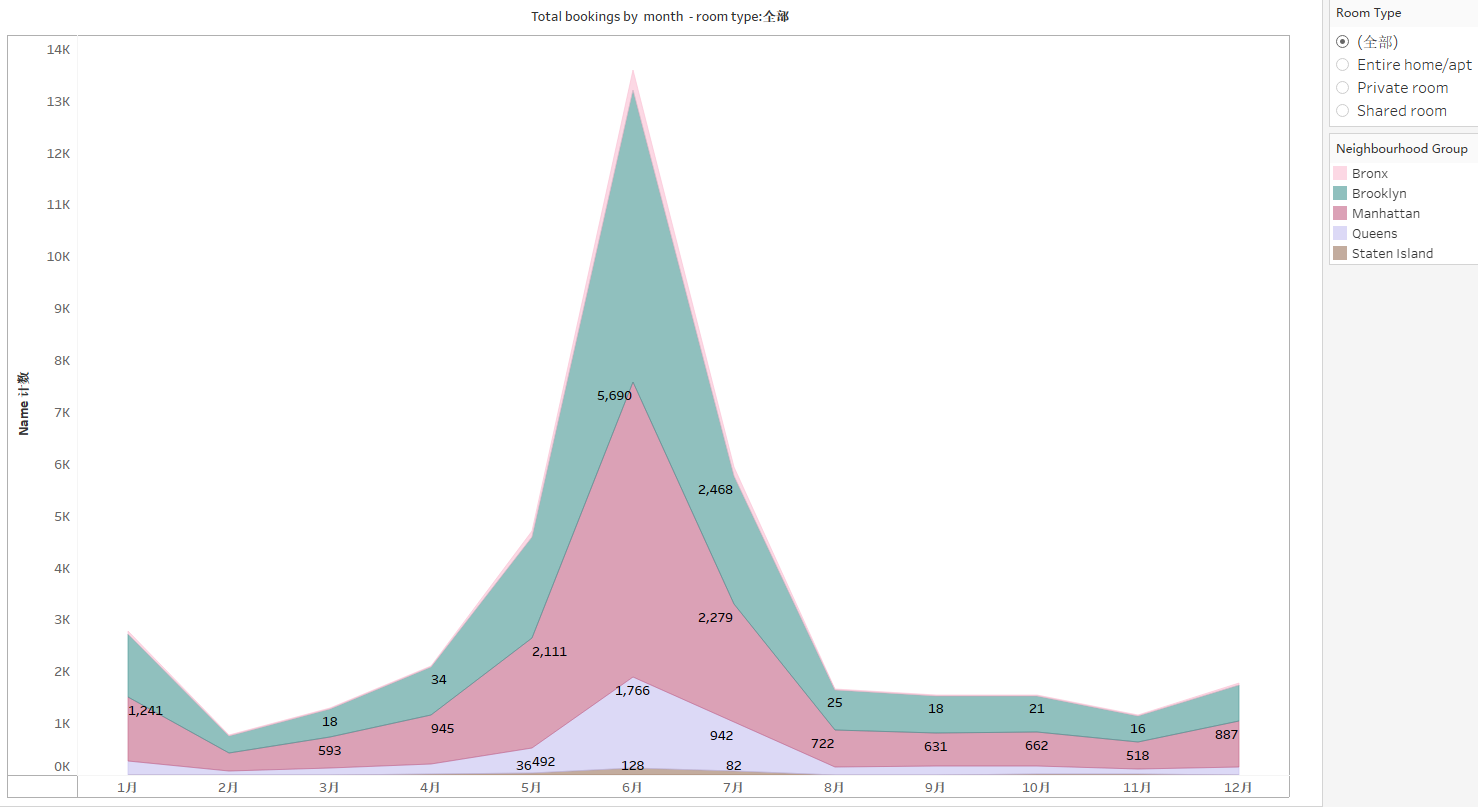
我对top10的host name的review的数量进行了统计，最终得到了下表



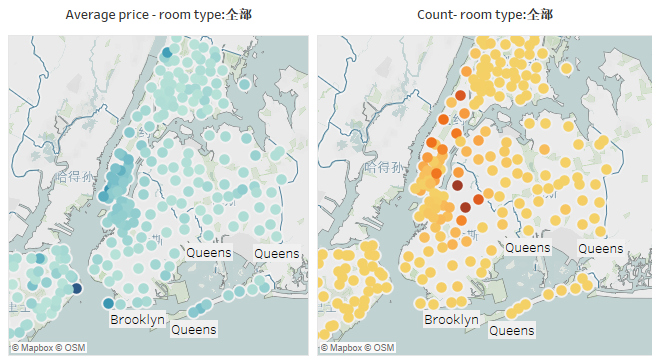
在地图上的分布的区域如下，可以发现大致与题目一中的生意好的区域大致吻合



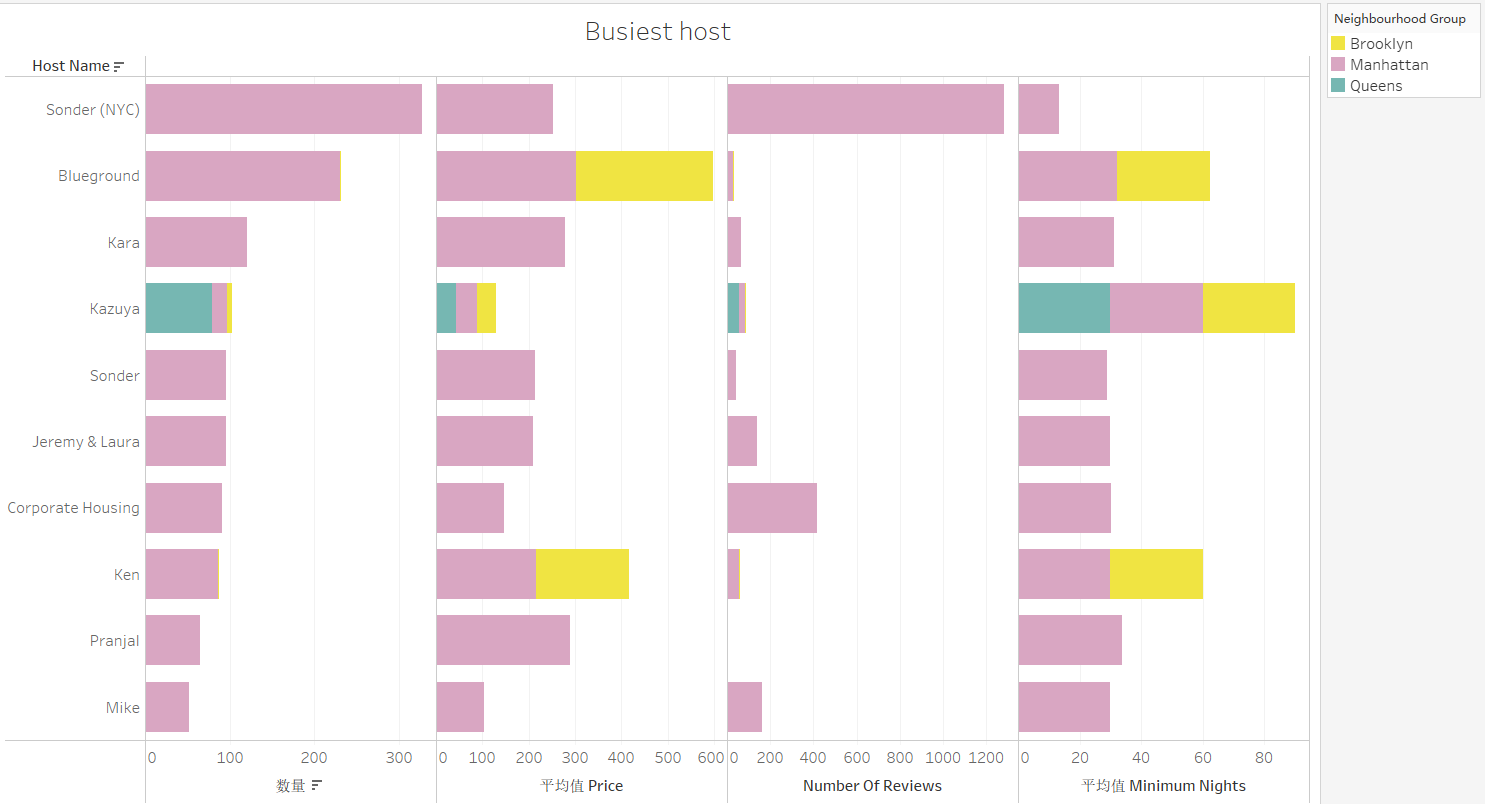
并且我对booking的数量绘制了流量图，仍然可以发现Manhattan的预定数量也比其他区域大很多



显示在地图上可以发现booking数量大的点也基本上和第一题中生意好的区域重合



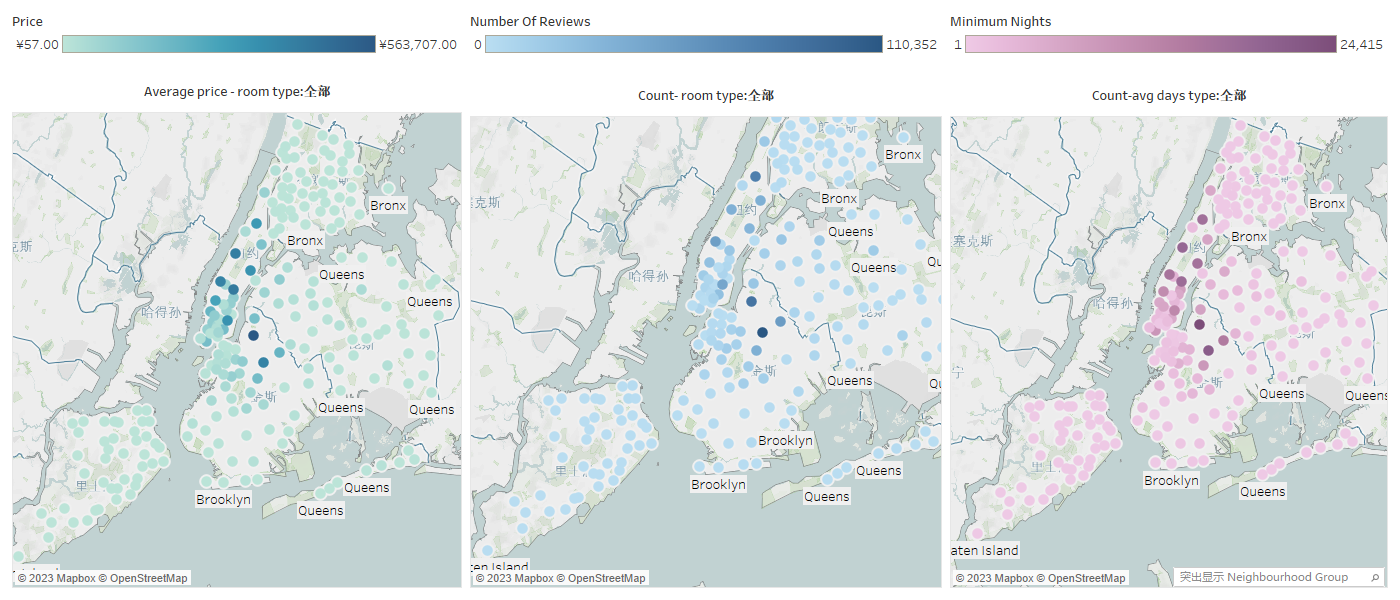
把Hostname中的按数量排序，找出前10后，我绘制了下表， 发现最繁忙的host大部分都位于Manhattan，很少有出现在Brooklyn或者Queens



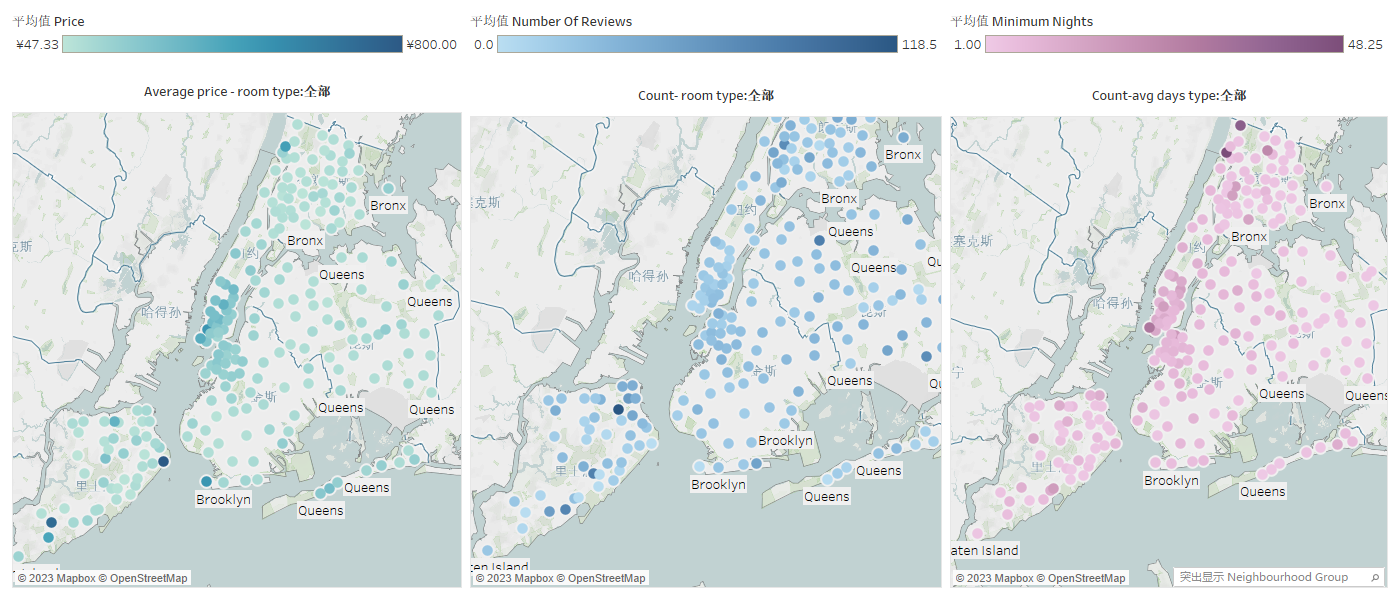
1. **价格，评论数量和预订天数[[1]](#footnote-1)之间是否存在一些关系？**

首先直接把三者都呈现在地图上，

总和的角度来说[[2]](#footnote-2)，三者的总和大呈现正相关

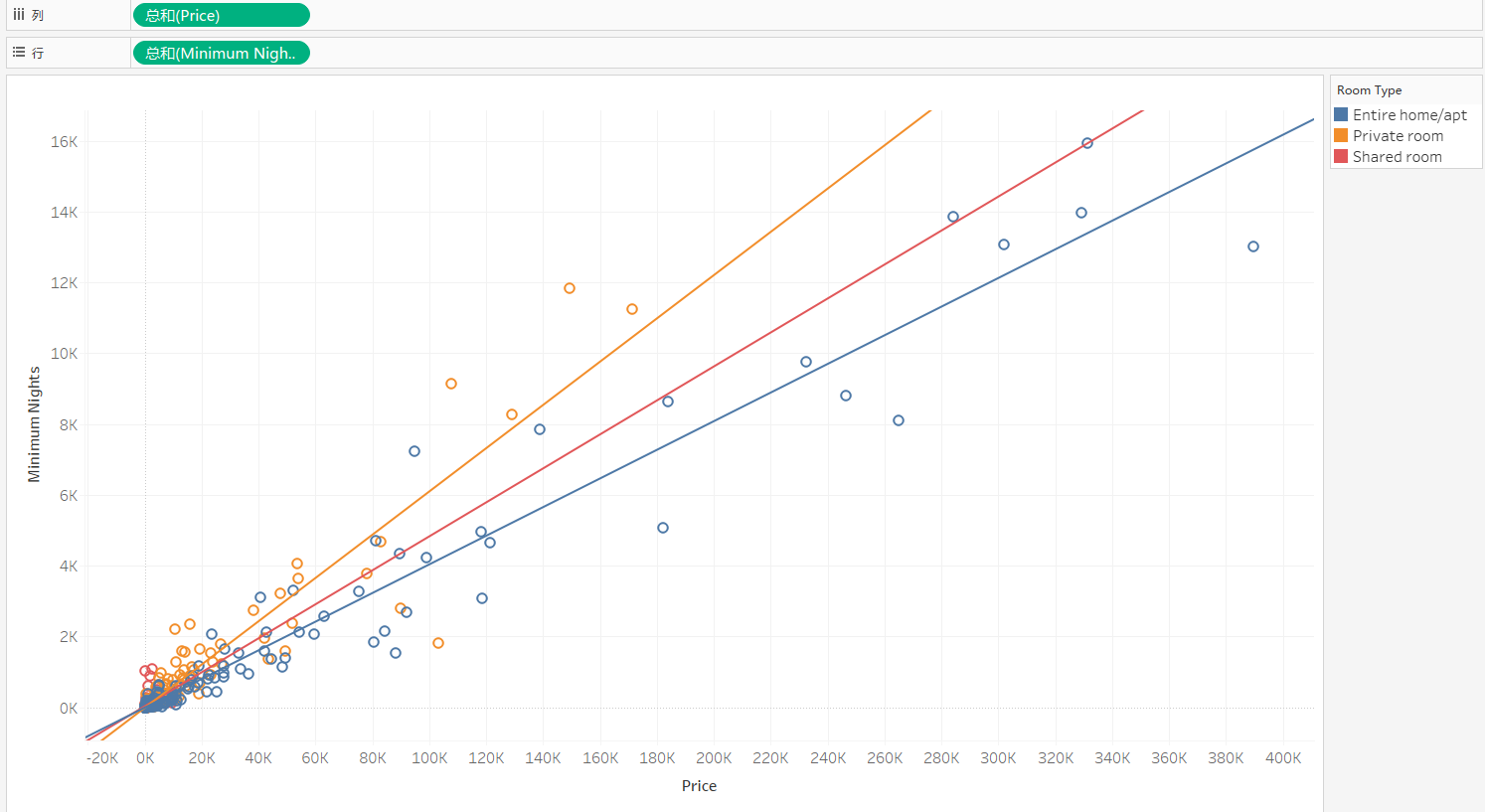


均值的角度来说，可以大致发现，三者在极值的区域存在一定的正相关性，但是中间区域几乎无关：评论数最高的一些地区价格也基本高，并且预定天数也会多一些；评论数最低的一些地区价格也基本低，并且预定天数也会少一些；但是其他的地方这三者几乎没有关系。

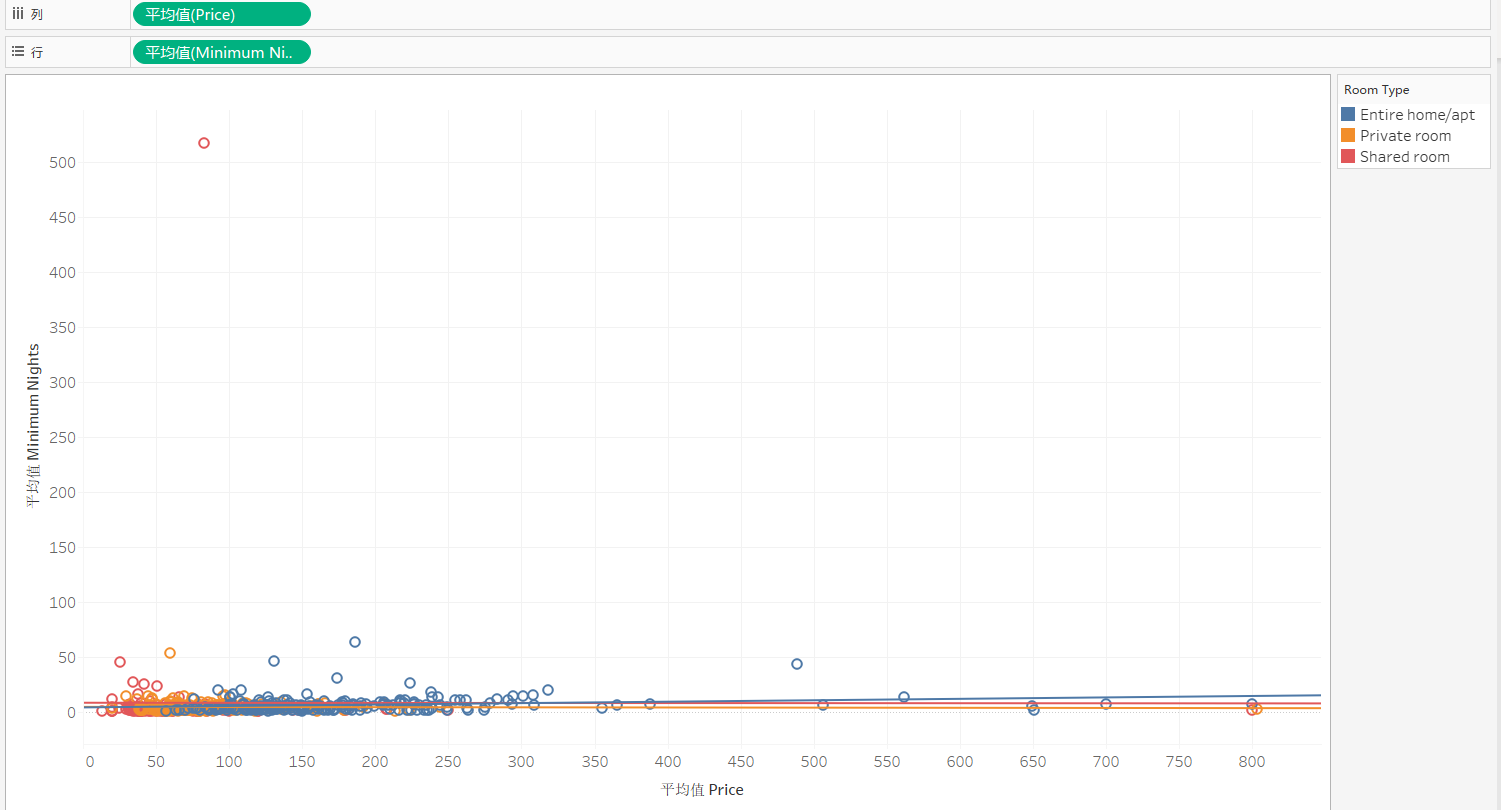


绘制二元关系图

总和上可以发现预定天数和价格大致呈现正相关



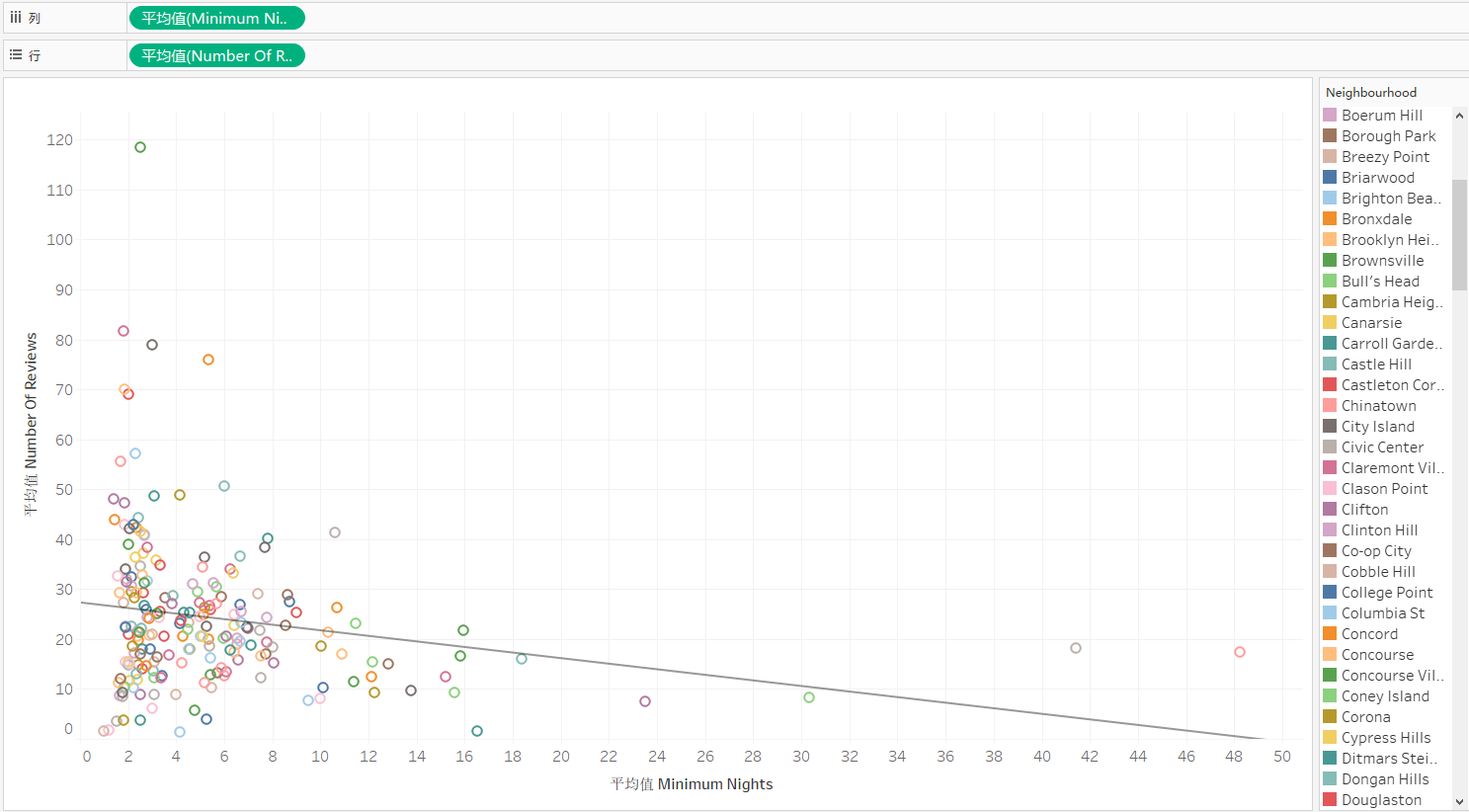
均值上可以发现预定天数和价格和关系不大



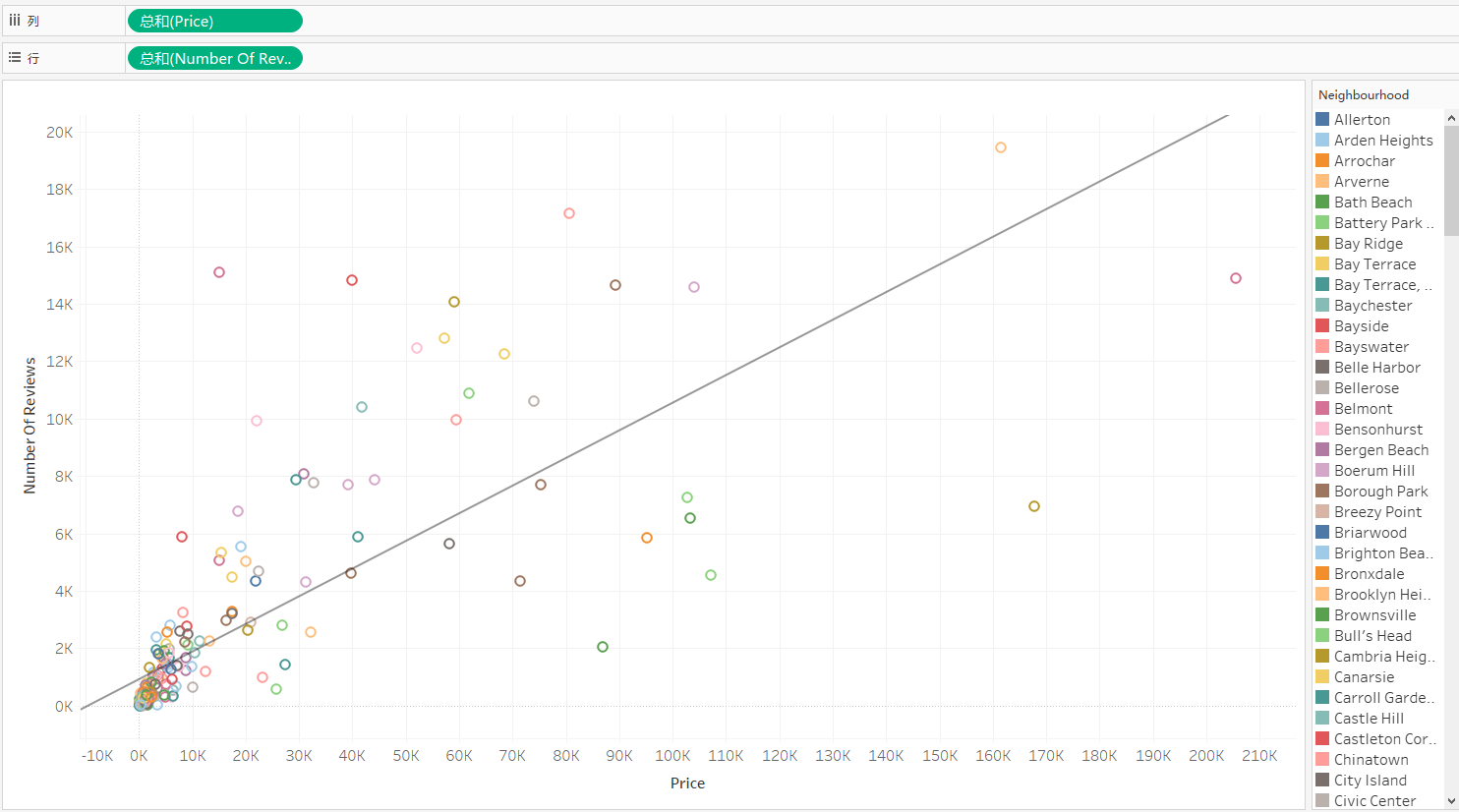
但是预定的天数和评论数在总和上大致呈现正相关



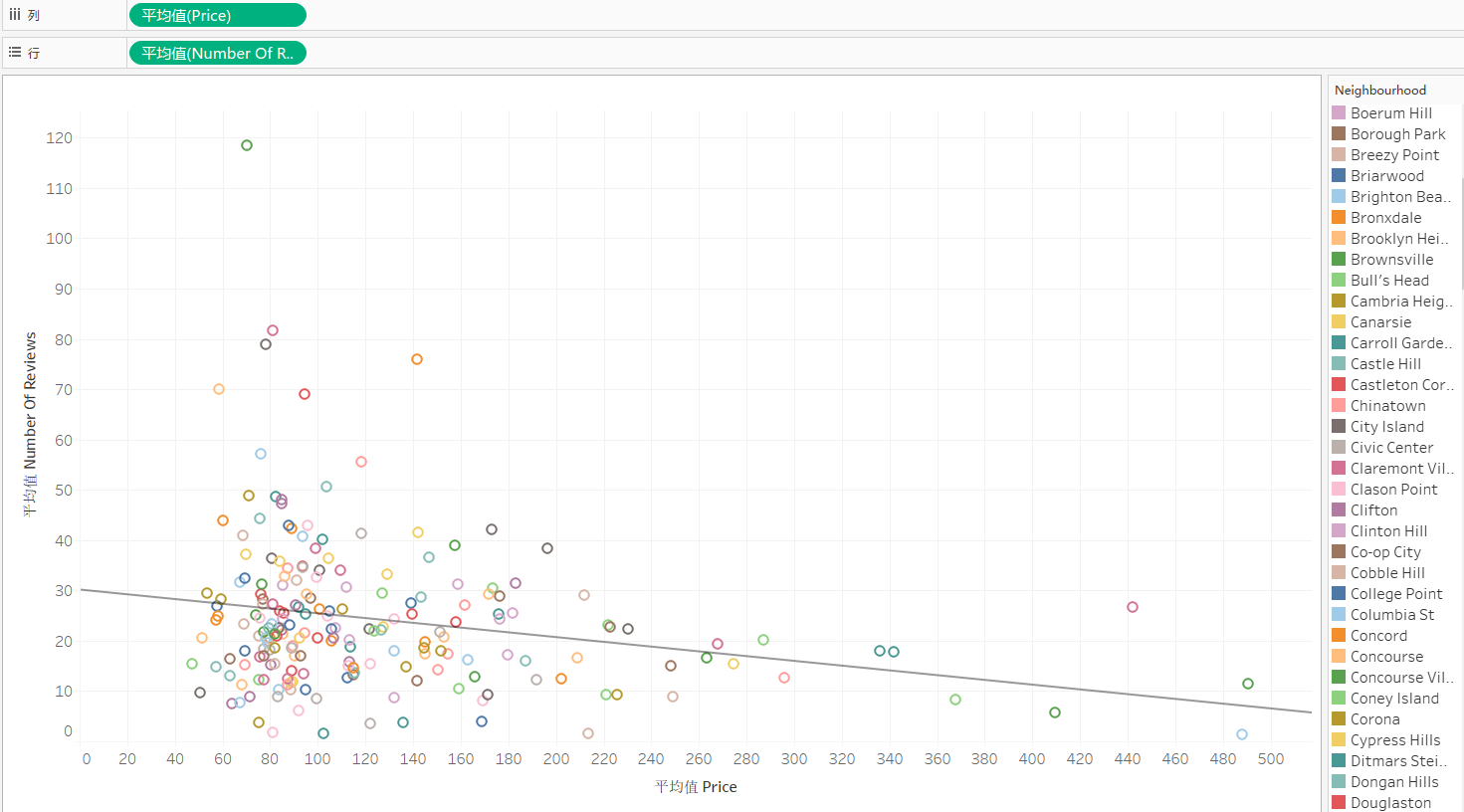
但是预定的天数和评论数在均值上几乎不相关



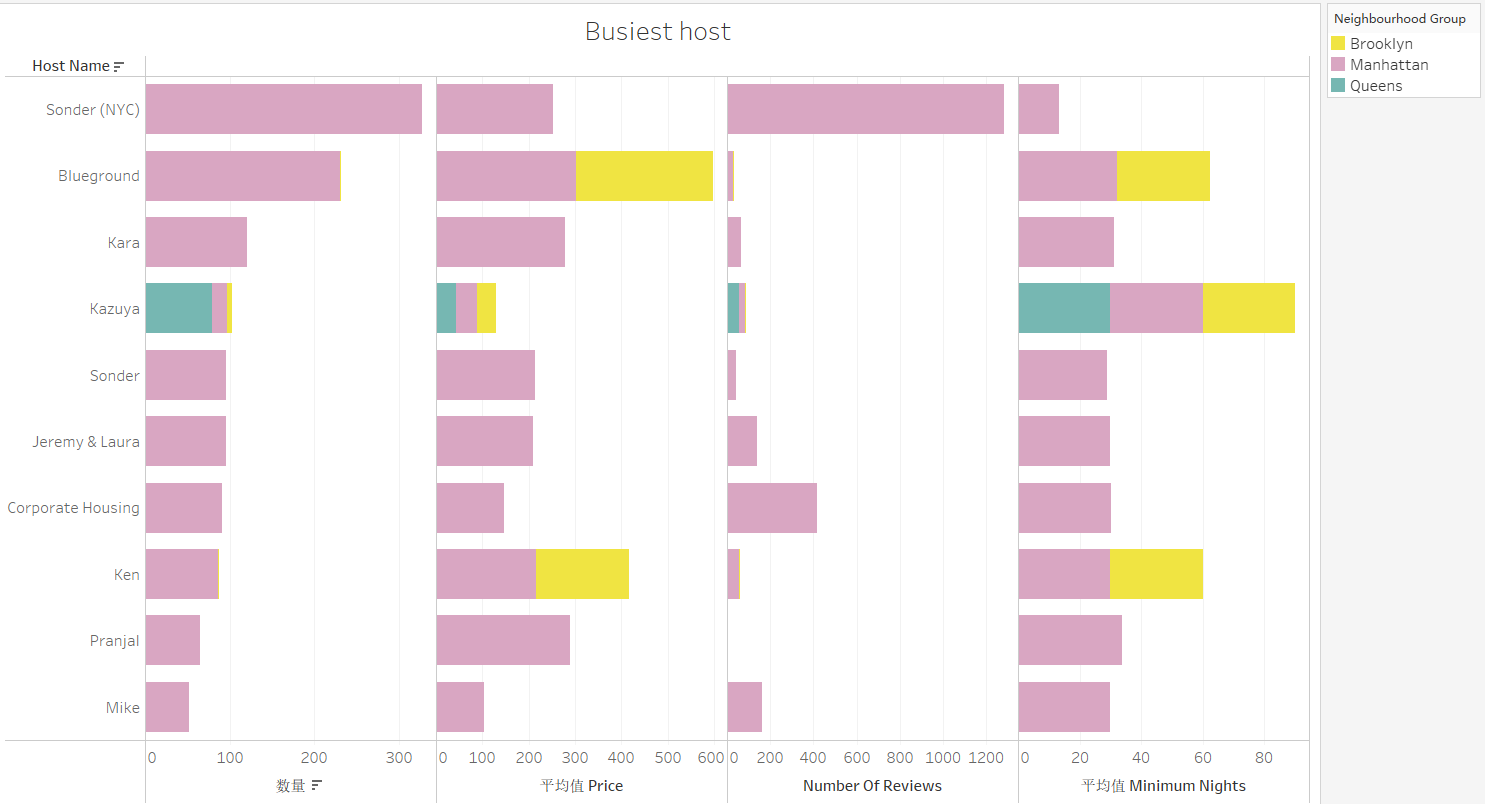
而某个区域总价格高的，一般来说该区域总review数量也大



但是对于平均值来说，也不存在太强的相关性



在最繁忙的10个host中，可以发现，三者之间似乎关系并不大



## 附录(Tableau操作相关过程)

**下载软件**

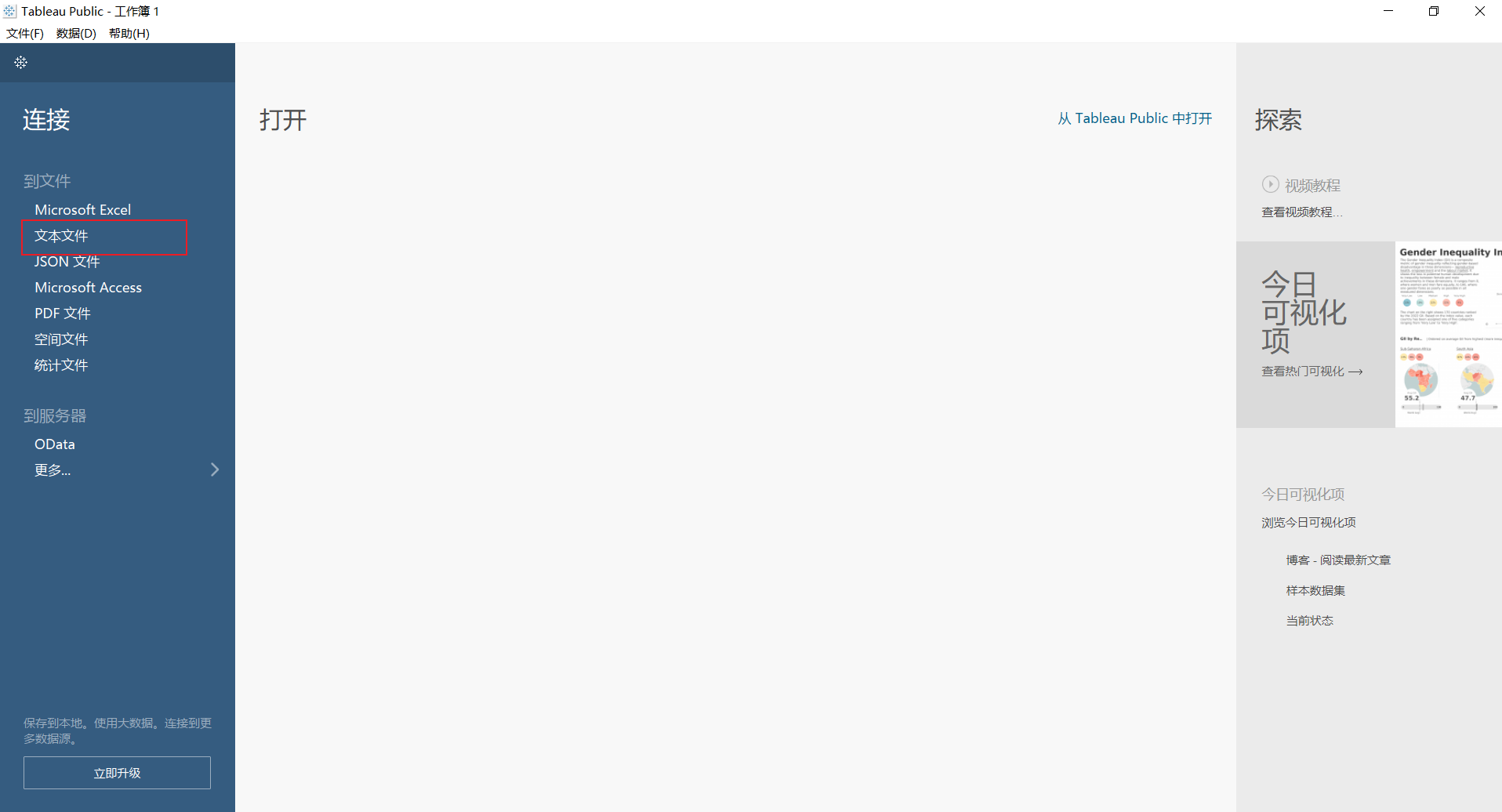
我使用的是桌面版的Tableau，

在<https://www.tableau.com/zh-cn/products/public/download>进行了下载

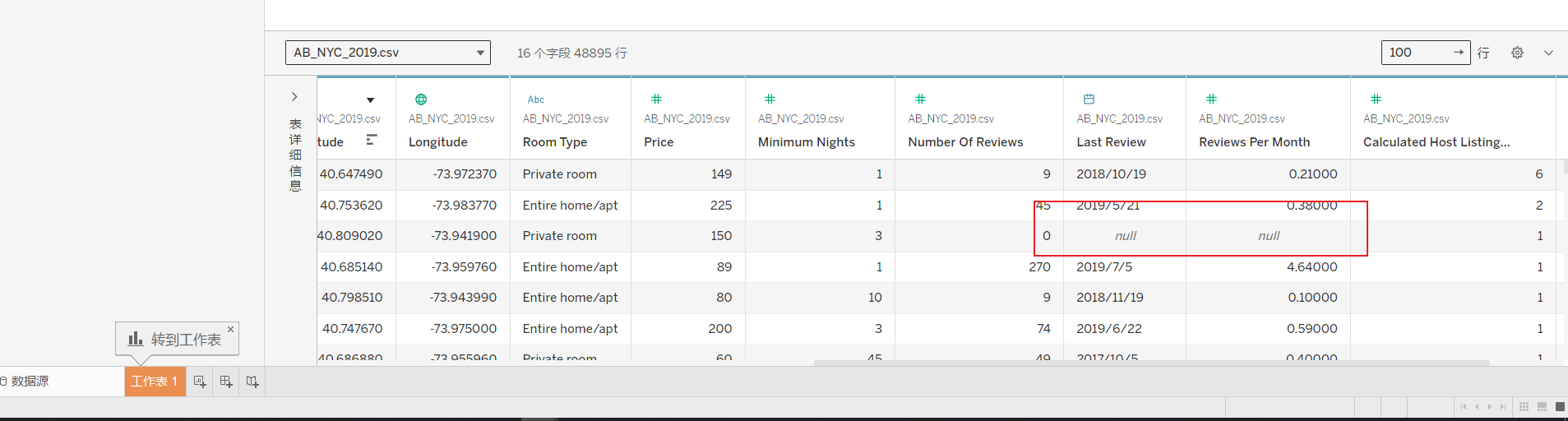
**导入数据**

点击文本文件，导入AB\_NYC\_2019.csv文件

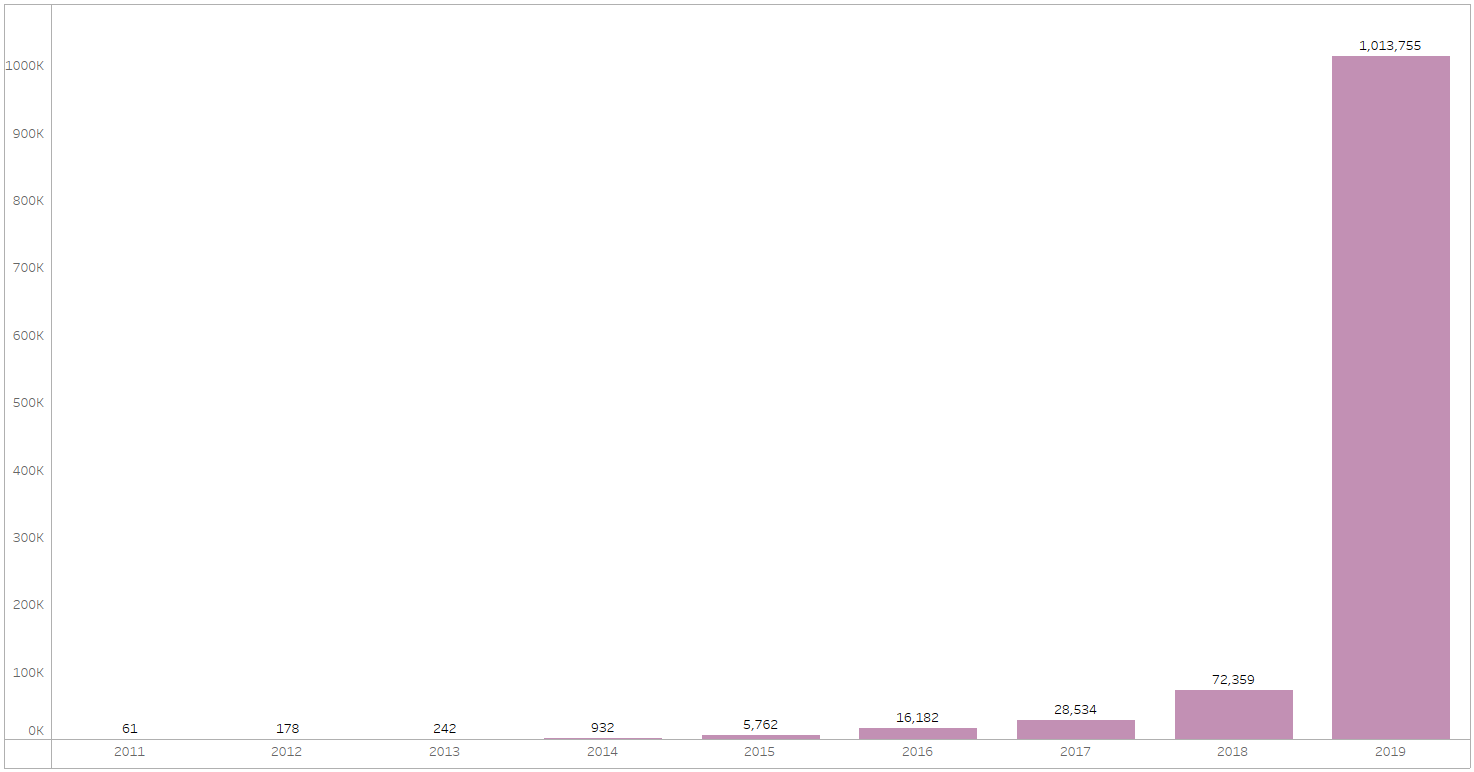
该数据集描述了 2019 年纽约州纽约市的交易记录和指标(<https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data>)



导入后发现存在一些为空的值，这在后续的操作中需要额外的注意（因为Tableau可以自动处理缺值的情况(拖动数据的时候可以选择包含null或者不包含)，因此我不再对数据进行预处理了，而选择直接使用）



另外，还可以发现，实际上并不是只有2019年的数据，而是包含了2011-2019的数据，但是前几年的数据量并不是很大，我这里并不没有去掉这些数据



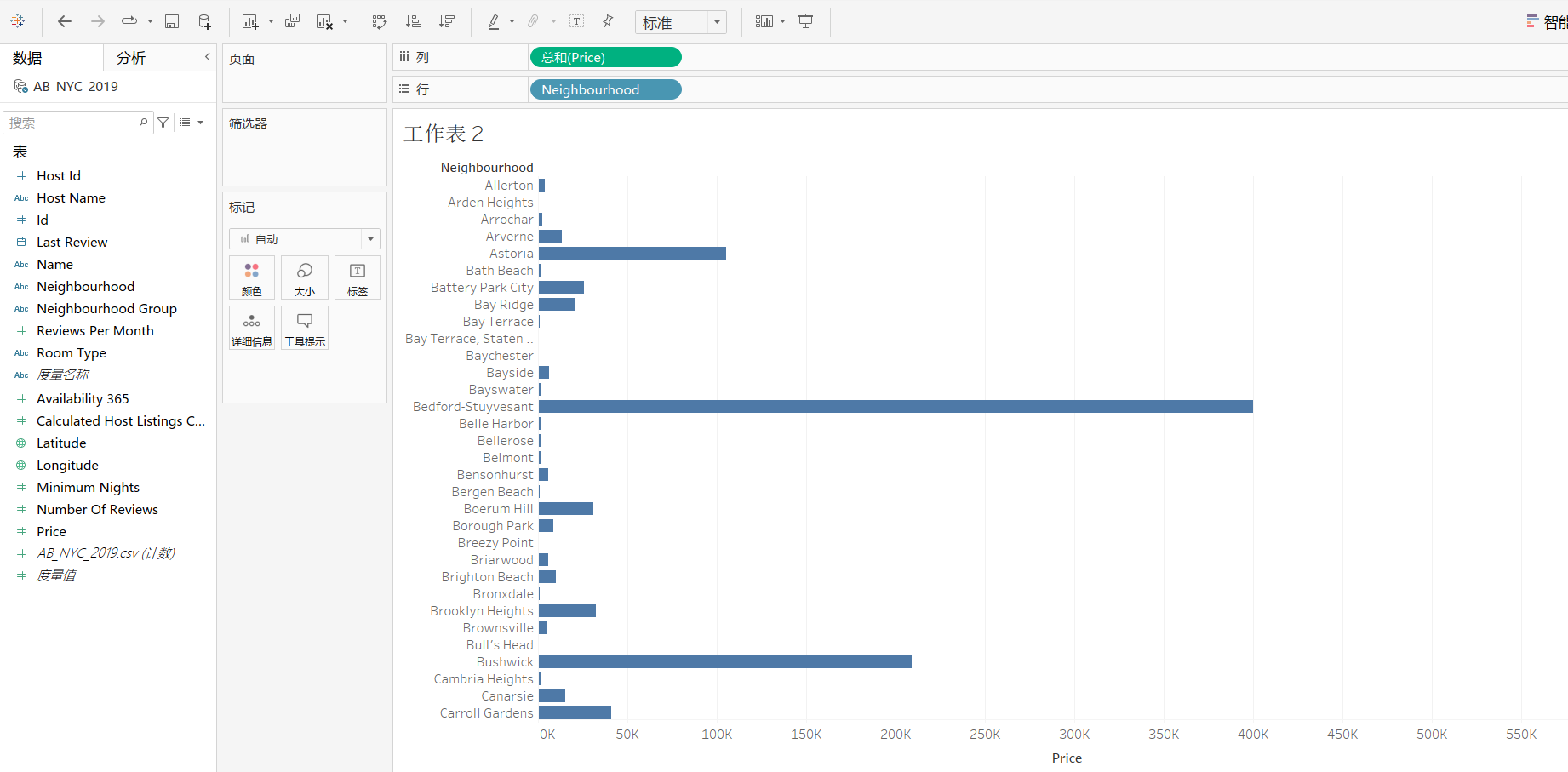
**建立数据表**



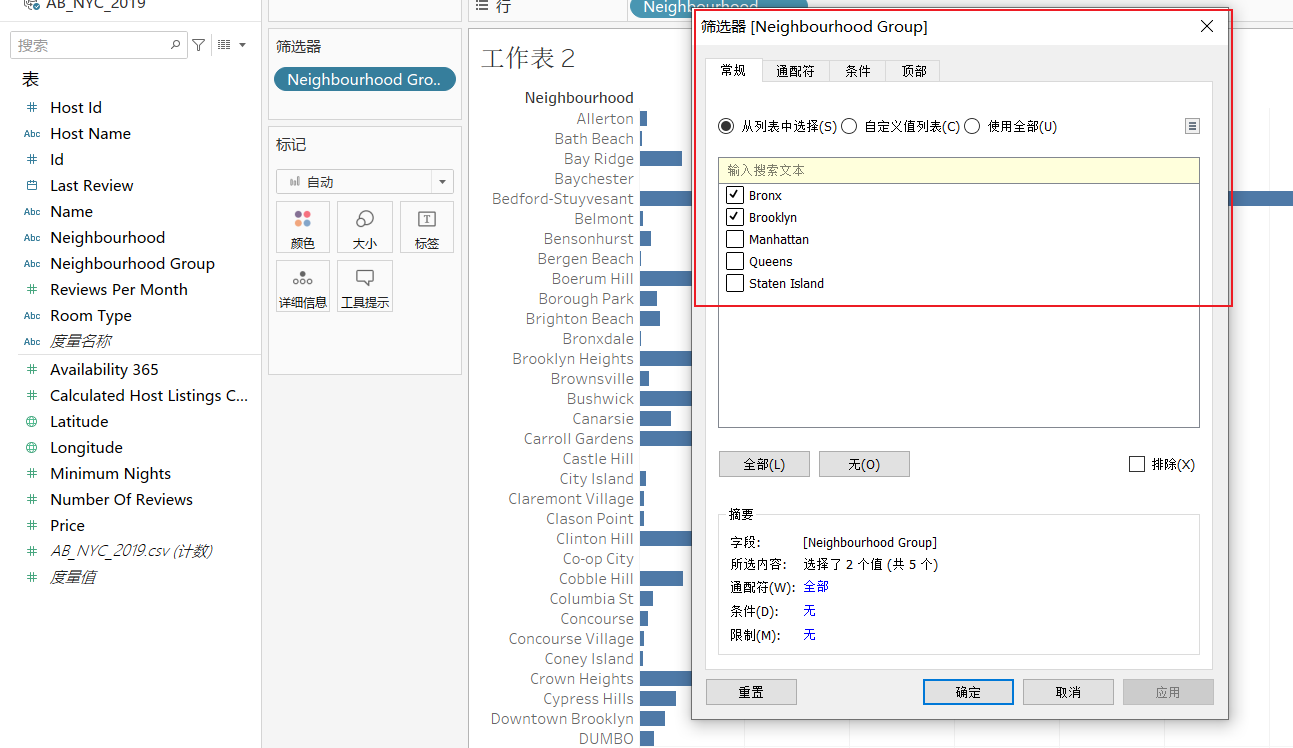
此时可以发现，之前导入的数据已经能够在侧边栏目中看到了



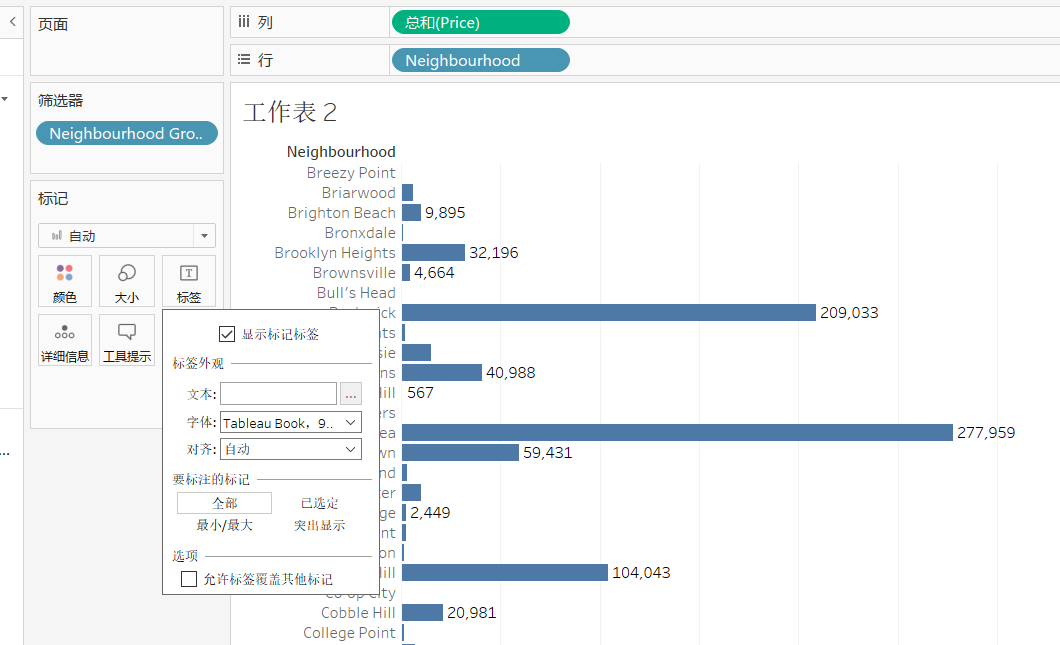
直接把数据项拖入到行列的方框中，可以自动生成一张



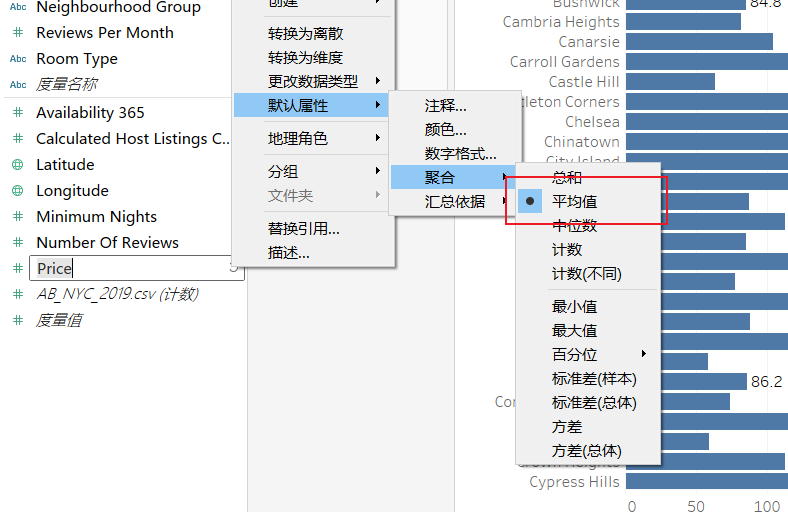
我们还可以通过增加筛选器来选择性的呈现数据



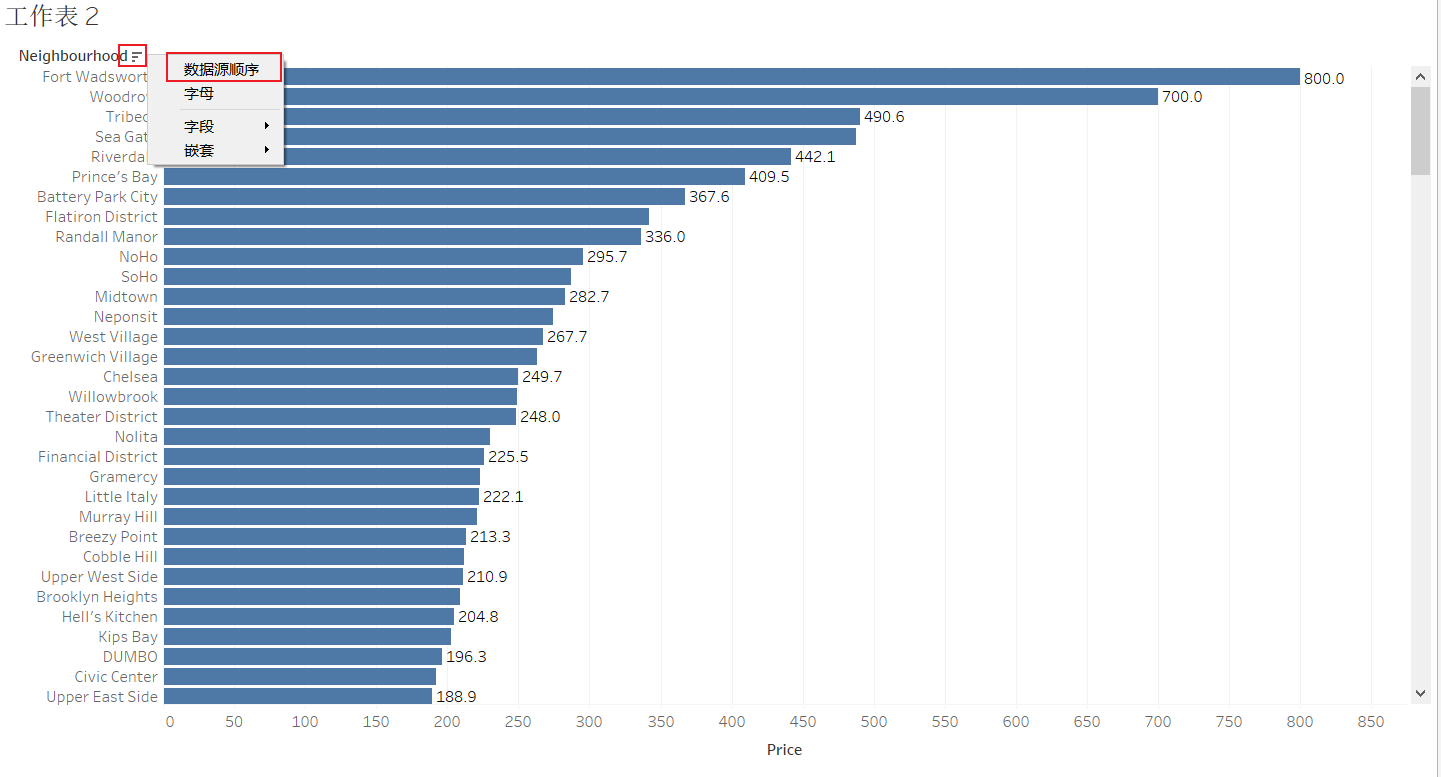
添加数据值的显示



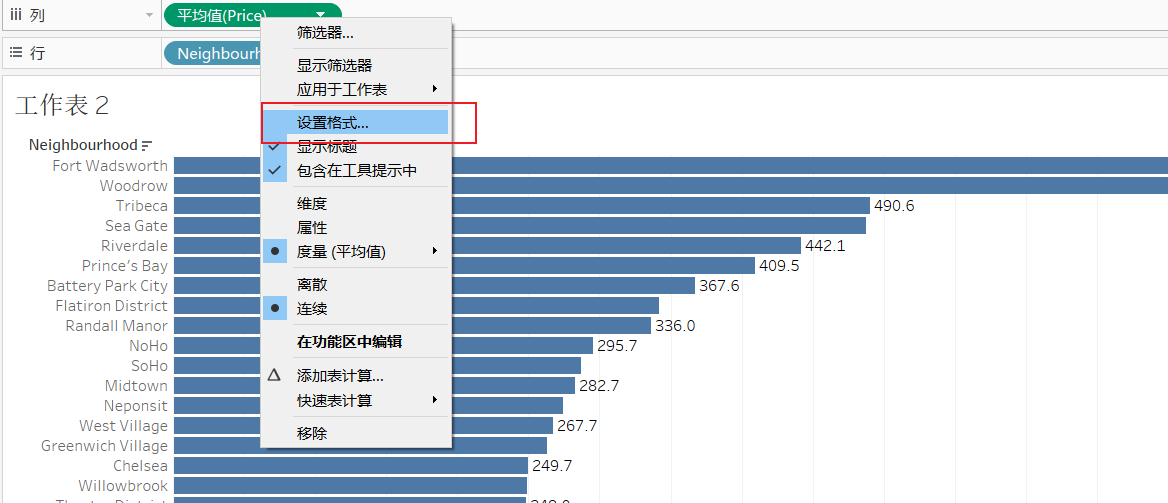
但是此时显示的是总价格，平均值才会更加合理，我们可以通过之间编辑数据进行聚合得到



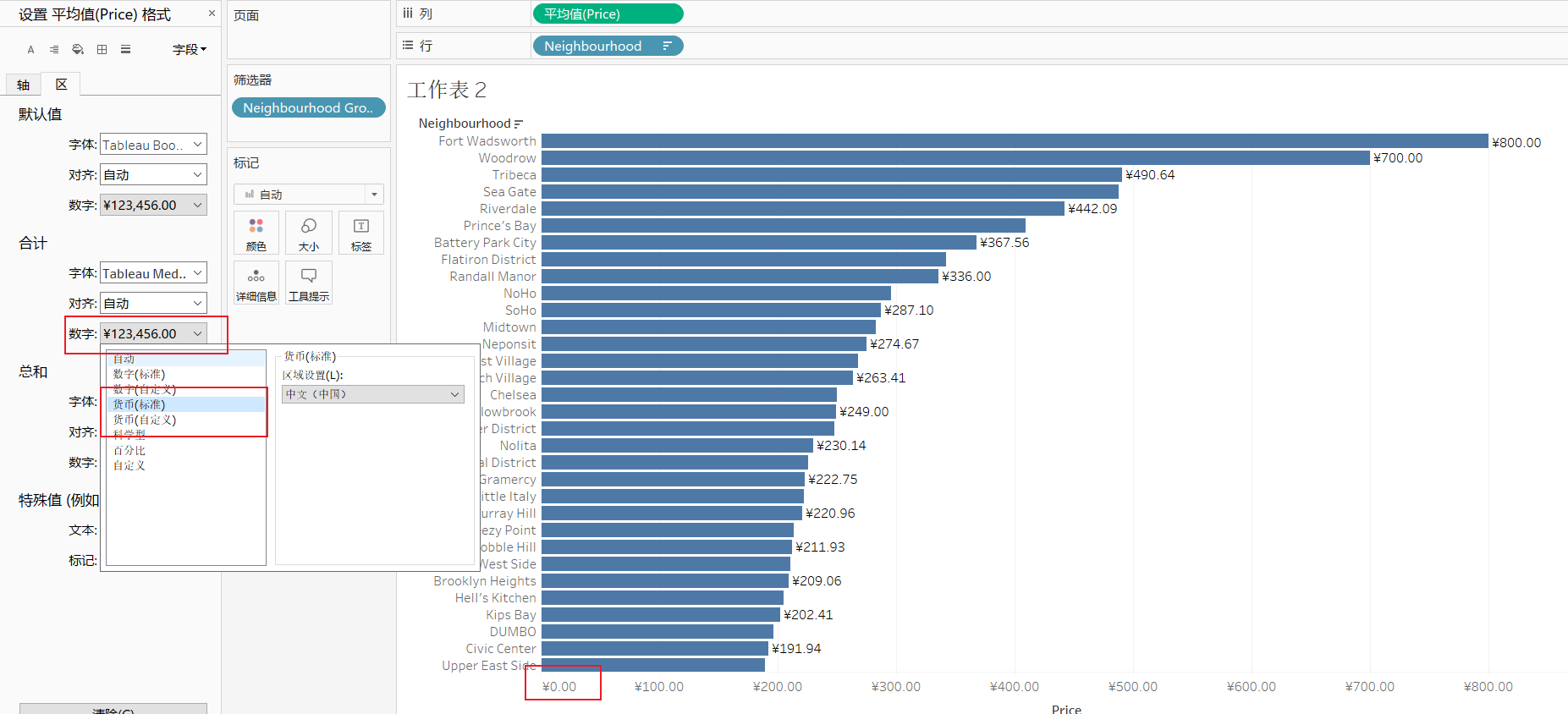
并且还可以直接进行排序



注意到现在的数字没有货币单位，可以通过设置格式来进行更改



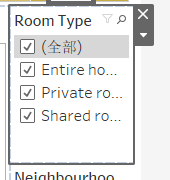
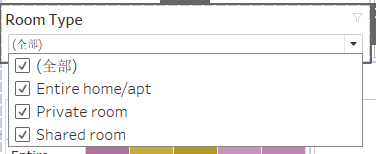
更改数据类型后即可显示货币单位



添加筛选器，并且改进一下样式之后，最终效果如下

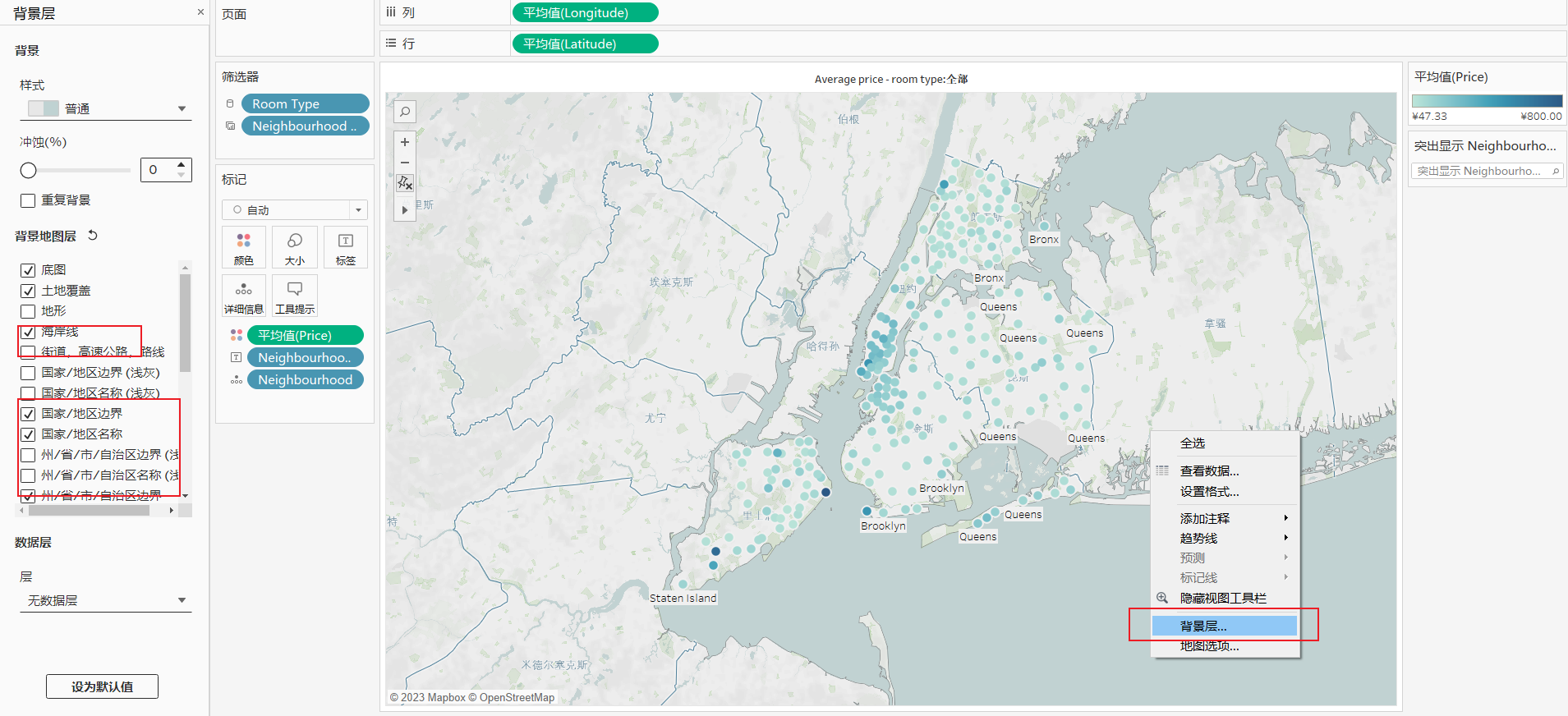


并且筛选器的样式能够更改

可以用类似的操作建立起地图

地图的操作中，选择背景层可以设置背景层的样式



其余的图表全都同理，不再赘述

1. 由于数据中并没有直接的预定天数，因此我把minimum\_nights视为了预定天数 [↑](#footnote-ref-1)
2. 即三者的统计度量都是总和，下同 [↑](#footnote-ref-2)