时空数据分析与挖掘实习

——实习日志三

学 院：遥感信息工程学院

班 级：2006 班（20F10）

姓 名：马文卓

学 号：2020302131249

老 师：田扬戈

时 间：2023 年 5 月 17 日

1. 今日进度

① 确定选题
② 获取 airBnb 数据
③ 对数据进行初步的分析了解

2. 阶段结果与分析

（1） 确定选题

今天，通过前面两天对于 WashingtonDC 的犯罪数据进行分析，我已经对其比较了解了。由于个人比较喜欢旅游，所以就把目光聚焦到旅游必备的租房问题上，目标也就定在全球最大的房屋出租网站 airbnb（爱彼迎）。

最后，经过和老师的几番讨论，确定选题为：为 WashingtonDC 的游客推荐最好的住址（这里的最好是指在保证安全系数的情况下，尽可能的平衡想去景区的距离远近和住房价格之间的矛盾，使游客得到更好的游玩体验）。

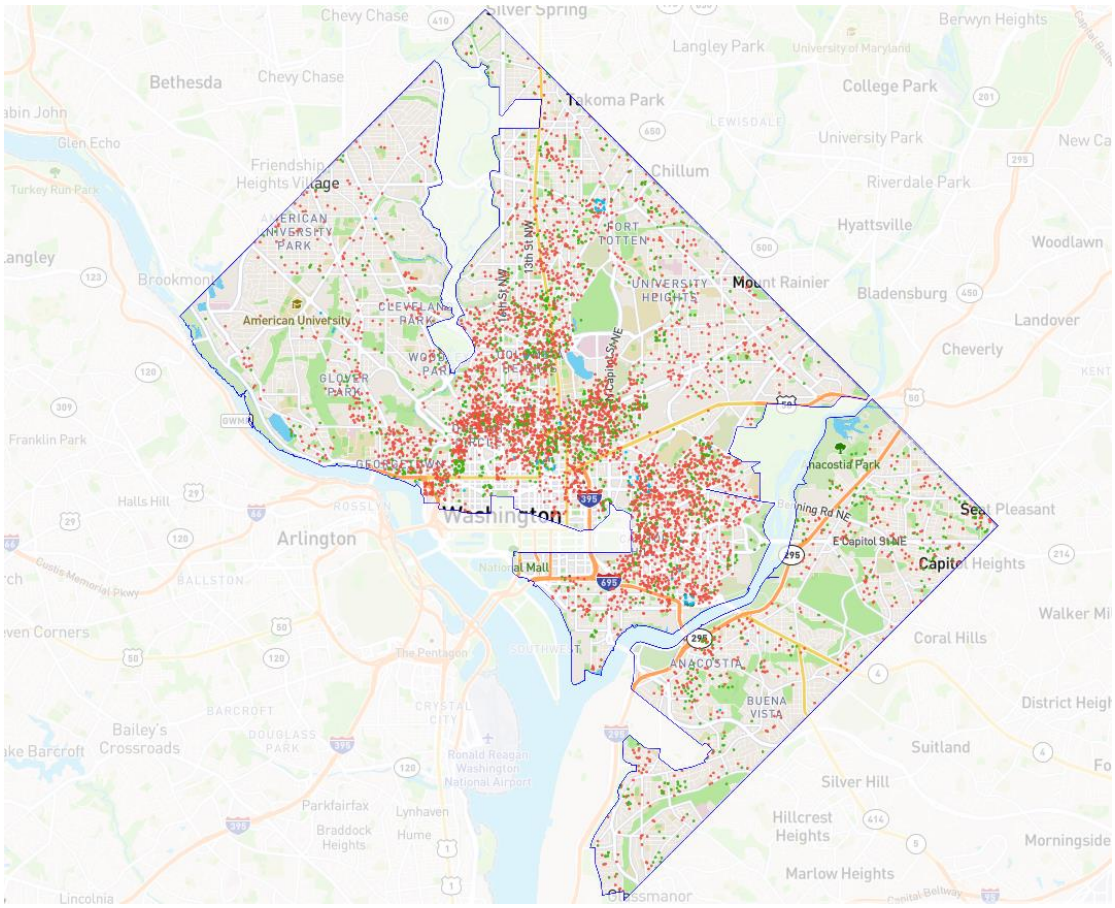
（2） 获取 airBnb 数据，并进行分析

分析上述的选题，可以清晰的知道本题所需要用到的三个数据：犯罪数据、租房数据、区域数据。这里我从 Inside Airbnb: Get the Data 下载了 airbnb2023 年 3 月更新的租房数据，主要包含了房屋出租信息（listings.csv）、评价信息（reviews.csv）、时间信息（calendar.csv）、邻居信息（neighbourhood.csv）。

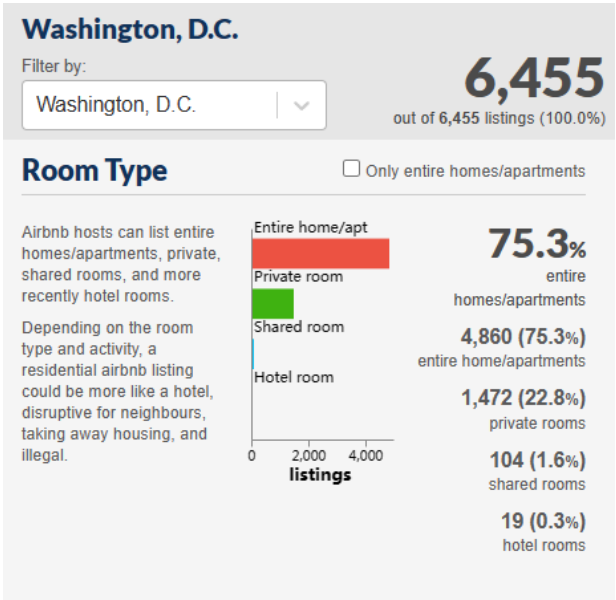
19 March, 2023 (Explore)

Country/City	File Name	Description
Washington, D.C.	listings.csv.gz	Detailed Listings data
Washington, D.C.	calendar.csv.gz	Detailed Calendar Data
Washington, D.C.	reviews.csv.gz	Detailed Review Data
Washington, D.C.	listings.csv	Summary information and metrics for listings in Washington, D.C. (good for visualisations).
Washington, D.C.	reviews.csv	Summary Review data and Listing ID (to facilitate time based analytics and visualisations linked to a listing).
Washington, D.C.	neighbourhoods.csv	Neighbourhood list for geo filter. Sourced from city or open source GIS files.
Washington, D.C.	neighbourhoods.geojson	GeoJSON file of neighbourhoods of the city.

拿到 airbnb 的数据指后，对其进行分析。首先，住房的分布如下图所示，和之前分析的犯罪数据分布基本一致，都聚集在美国国会大厦、博物馆、购物中心、唐人街等华盛顿最繁华的地段。其实，这个规律和常识是相符合的，在繁华的地段房源肯定密集，同时由于人流太混杂，小偷小盗也经常发生。

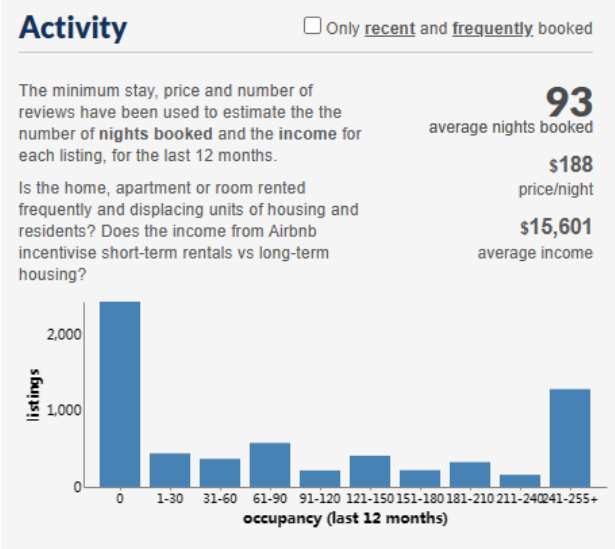


通过分析，发现 20223 年这段时间一共有 6455 个可租房源。其中房屋累心分为 4 类：整个公寓（entire home/apt）、私房（private room）、合租房（shared room）、酒店（hotel room）。其中 entire home 占据了大多数（75.3%），然后私人房间占据 22.8%，合租房和酒店都是极少数。

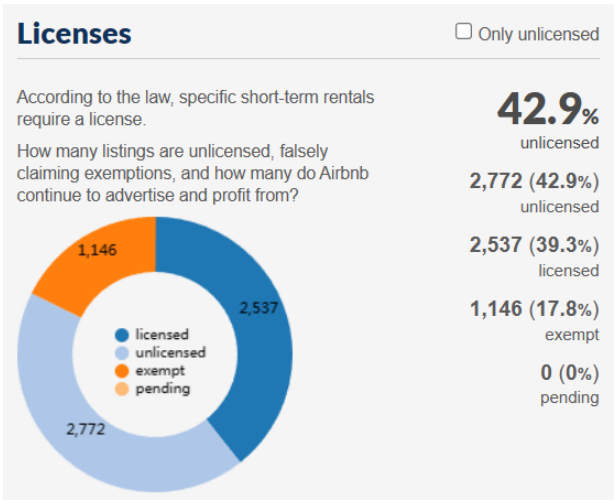


这里使用房屋的最短住宿天数、价格、评论的数量来估算每间房子的平均租出天数和收入。可以看到，华盛顿特区的房源平均租期在 93 天左右，但是看图可知，短租和超长租的较多。华盛顿特区的房源平均价格在$188 一晚，房主的

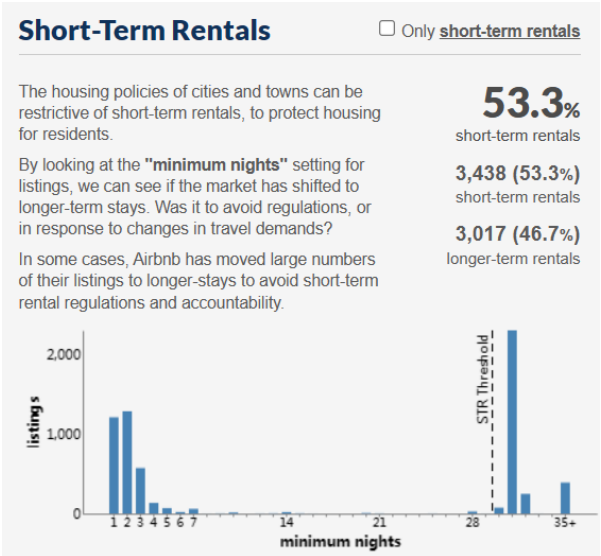
平均收入在$15,601。



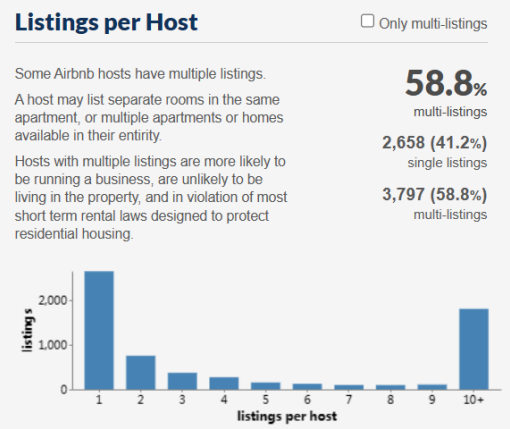
对于房源的许可证而言，有 42.9%的房子是未经许可的，只有 39.3%的房源是经过许可的。

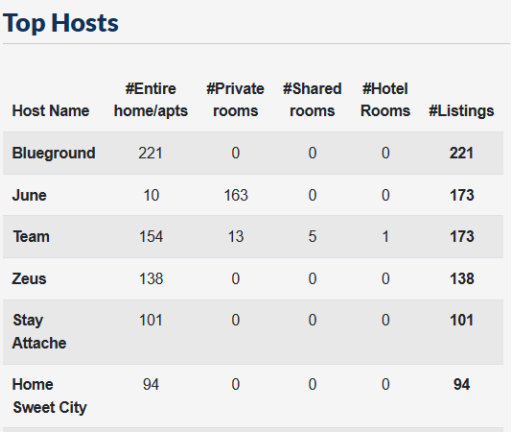


所有房源当中，短期出租占据了大部分（53.3%）。



在华盛顿特区的房主中有58.8%的人有多套房源，其中拥有10套以上的房主占据大多数，可以得知这些人要么是富豪，要么就是专门从事与租房行业，隶属于某些房地产公司。





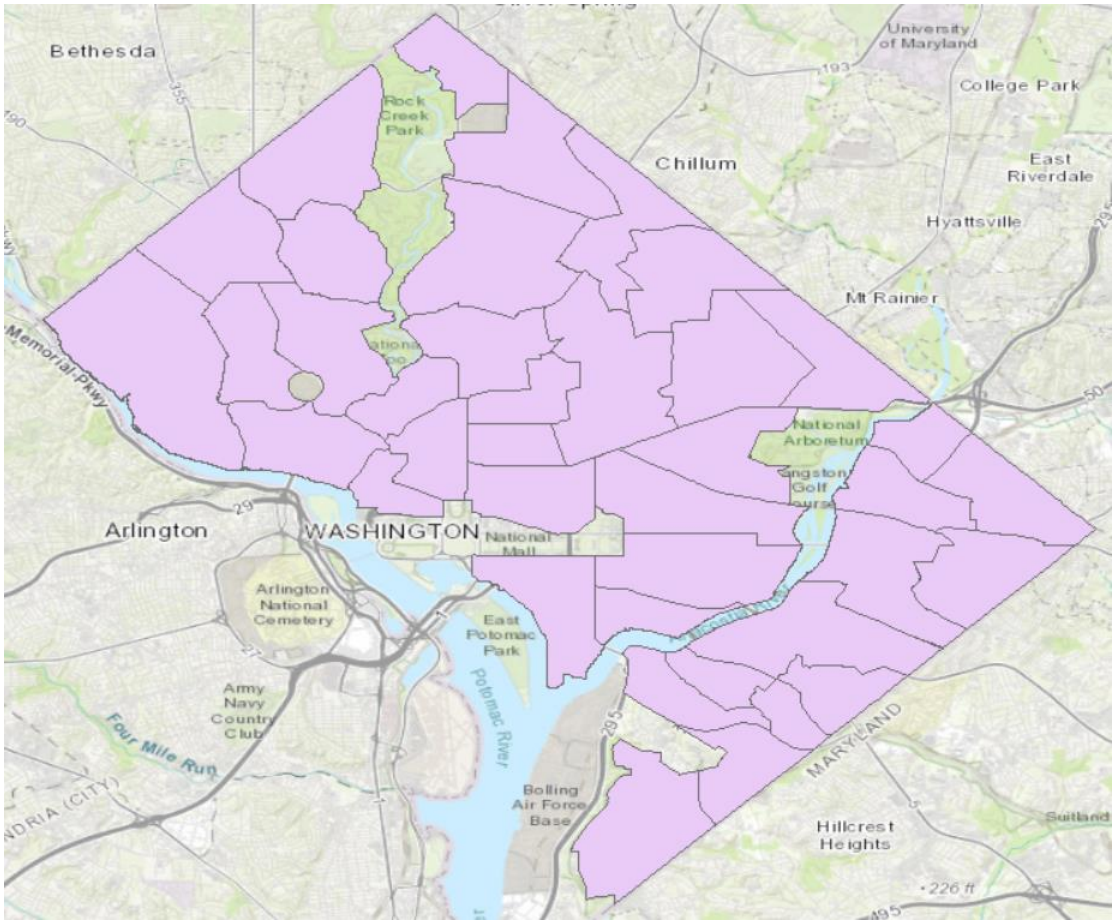
Top Hosts

Host Name	#Entire home/apts	#Private rooms	#Shared rooms	#Hotel Rooms	#Listings
Blueground	221	0	0	0	221
June	10	163	0	0	173
Team	154	13	5	1	173
Zeus	138	0	0	0	138
Stay Attache	101	0	0	0	101
Home Sweet City	94	0	0	0	94

房源数据（listing.csv）主要有如下关键字段：

Key	Description
id	房屋编号
name	房屋名称
host_id	房主编号
hostname	房主姓名
neighbourhood_group	邻居组别
neighbourhood	邻居
latitude	纬度
longitude	经度
room_type	房屋类别
price	单价
minimum_nights	最少租房天数
number_of_reviews	评论数量
last_review	最近一次评论时间
reviews_per_month	每个月评论数量
calculated_host_listings_count	该房主拥有房源数量
availability_365	该房源一年内空余天数
number_of_reviews_ltm	评论的项目个数
license	许可证

airbnb 的 neighbourhood 数据将 WashingtonDC 划分成了 39 个区域，本次选题中我们就以这 39 个区域为基础进行分析。



3. 遇见问题与解决方案

今天所遇到的最大的问题是获取 airbnb 的数据，最开始我使用 python 爬虫在官网爬取，但是由于网络等限制，爬取出来的表格是空的。最后，所幸在上述网站上找到了数据源。

然后就是区域数据，最开始下载下来是 geojson 的格式，想要导入 arcgis 需要 shapefile 的格式，最后通过一个在线 geojson 编辑网站完成这个转换。