时空数据分析与挖掘实习

——实习报告

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 20 日

目录

一、 实习概述

1. 实习目标

本次实习的目的主要如下：
① 完整体验时空数据的获取及处理流程
② 利用数据挖掘、分析等手段分析数据，并从数据中挖掘有用的知识
③ 利用 python 等工具和课程学习的知识完成某个较为有意义的选题

2. 实习要求

本次实习的要求如下：
① 每天提交实习日志记录当天完成内容和遇到的问题及其解决方案
② 撰写实习报告描述实习中的具体做法、相关数据、实验结果等
③ 代码完整、数据完整、格式规范

3. 实习选题

随着旅游业的日益发展，美国首都华盛顿特区（Washington，DC）的游客数量逐年递增。但是不得不提的是，华盛顿特区也是犯罪大都市，其犯罪率一直居高不下，这也导致了华盛顿特区旅游存在不安全的因素。

作为喜欢旅游的大学生，又学习了时空数据的挖掘与分析这门课程，兴趣使然决定使用华盛顿特区的犯罪数据和爱彼迎官网的华盛顿特区租房数据进行数据挖掘与分析，来为前往华盛顿特区旅游的游客推荐较好的房源住址。这里的较好是指：较为安全、较为便宜、距离目标景点较近、房间的评价较好等多维度考虑。

综上所述，本次实习的选题为：华盛顿特区旅游租房选择问题。

二、 数据介绍

1. Washington DC Crime 数据介绍

本次实习我们使用的是从 https://opendata.dc.gov 开源数据网站中下载的华盛顿特区（WashingtonDC）犯罪事件数据。值得注意的是，我从该网站上下载了十年的（2014-2023）华盛顿特区犯罪数据（两种形式：csv 和 shapefile），总共 308766 个犯罪事件作为本次实习的原始数据集。

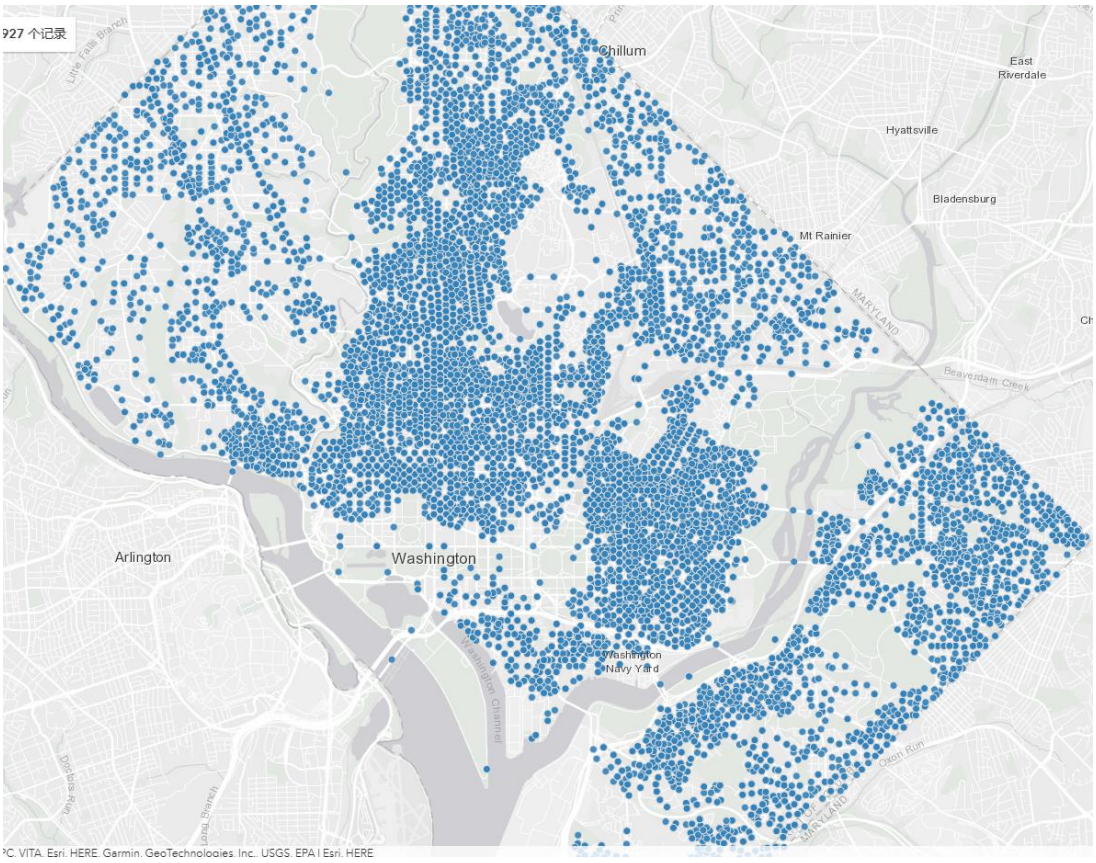


图 1 Washington DC 2022 Crime Incidents 数据示例

此数据集每天更新，覆盖了整个华盛顿地区的大部分犯罪类型和犯罪的详细信息。关于数据集的其他详细信息可见官网：

Theme: 主题

Theme Keyword Thesaurus: None
Theme Keyword: Crime
Theme Keyword: Feeds
Theme Keyword: DC GIS
Theme Keyword: cdw
Theme Keyword: MPD
Theme Keyword: safety
Theme Keyword: robbery
Theme Keyword: homicide
Theme Keyword: assault
Theme Keyword: burglary
Theme Keyword: theft

Status: 状态

Progress: Complete
Maintenance and Update Frequency: Daily

Spatial Domain: 边界

Bounding Coordinates:

West Bounding Coordinate: -77.114200
East Bounding Coordinate: -76.909900
North Bounding Coordinate: 38.994900
South Bounding Coordinate: 38.813400

图 2 Washington DC 2022 Crime Incidents 数据信息示例

在本数据集中，每个犯罪案件包含么 24 个字段，如下表所示（其中标红的为关键字段）。字段的详细信息见官网。

表 1 Washington DC Crime Incidents 数据字段含义表

Key	Definition	Description
CCN	Criminal Complaint Number	刑事诉讼编号（警方给每个案件分配的唯一编号）
REPORT_DAT	Crime Report Date	案件被警方报道的时间（可能晚于发生时间）
		警局成员在案件报到时的值班。白班一般

SHIFT	MPD Shift	在早上 7 点到下午 15 点之间(军事时间)；晚上 15 点到 23 点，午夜 23 点到 7 点。如果是未知的，则该字段将显示“UNK”
METHOD	Type of weapon used to commit crime	执行犯罪时用的武器
OFFENSE	Crime Offense	犯罪行为的定义（种类）
BLOCK	Block Name	街道名
XBLOCK	Block X Coordinate	犯罪事件街区的 X 坐标(质心)，参考马里兰州平面 NAD 1983 米
YBLOCK	Block Y Coordinate	犯罪事件街区的 Y 坐标(质心)，参考马里兰州平面 NAD 1983 米
WARD	Distric Ward Identifier	地区分区编号
ANC	Advisory Neighborhood Commission Identifier	咨询居委会编号
DISTRICT	Police District	警区号
PSA	Police Service Areas	警察服务区号
NEIGHBORHOOD_CLUSTER	Neighborhood Cluster	社区聚类号
BLOCK_GROUP	Census Block Group	普查组号
CENSUS_TRACT	Census Track	普查区号
VOTING_PRECINCT	Voting Precinct	投票选区
X	X Coordinate	犯罪事件的 X 坐标,参考马里兰州平面 NAD 1983 米
Y	Y Coordinate	犯罪事件的 Y 坐标,参考马里兰州平面 NAD 1983 米
LATITUDE	Lattitude	犯罪事件发生的纬度
LONGITUDE	Longitude	犯罪事件发生的经度
BID	Business Improvement Districts	商业区号
START_DATE	Crime Start Date	犯罪开始日期
END_DATE	Crime End Date	犯罪结束日期
OBJECTID	Internal feature number	自动生成的连续唯一整数

其中值得我们注意的是，统计者将犯罪类型分为两个主要的犯罪类别:暴力犯罪和财产犯罪。暴力犯罪包括杀人、性虐待、使用危险武器攻击和抢劫。暴力犯罪可以通过使用的武器进一步搜查:枪支，非枪支，或两者兼而有之。财产犯罪包括入室盗窃、机动车盗窃、车辆盗窃、盗窃(其他)和纵火。如下表：

表 2 Washington DC Crime Incidents 犯罪类型表

Category	Offense	Description
Property Crime	THEFT/OTHER	盗窃（其他）
	THEFT F/AUTO	车辆盗窃
	MOTOR VEHICLE THEFT	机动车盗窃
	BURGLARY	入室盗窃
	ARSON	纵火
Violent Crime	ASSAULT W/DANGEROUS WEAPON	危险武器攻击
	HOMICIDE	杀人

	SEX ABUSE	性虐待
	ROBBERY	抢劫

以 2022 年的数据为例，我们可以看到盗窃罪的数量是最多的，纵火罪相对较少。整体而言财产犯罪远多于暴力犯罪。

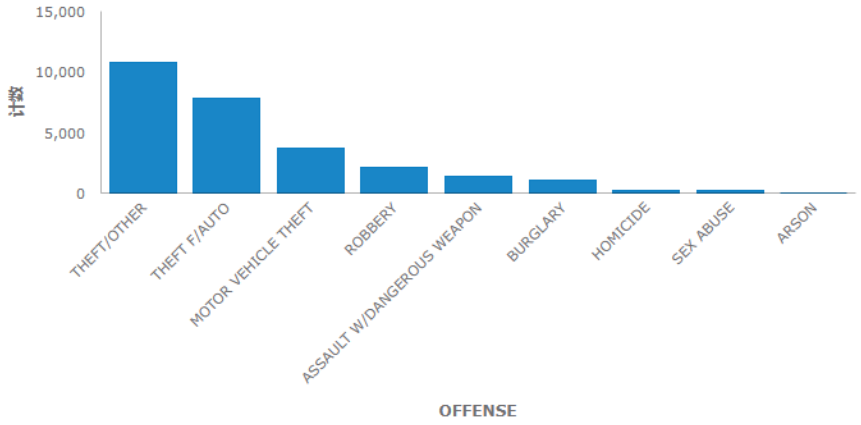


图 3 Washington DC 2022 Crime Incidents 犯罪类别数量统计图

从犯罪手段来看，使用枪支和刀的还是少数，更多的还是利用其他手段。

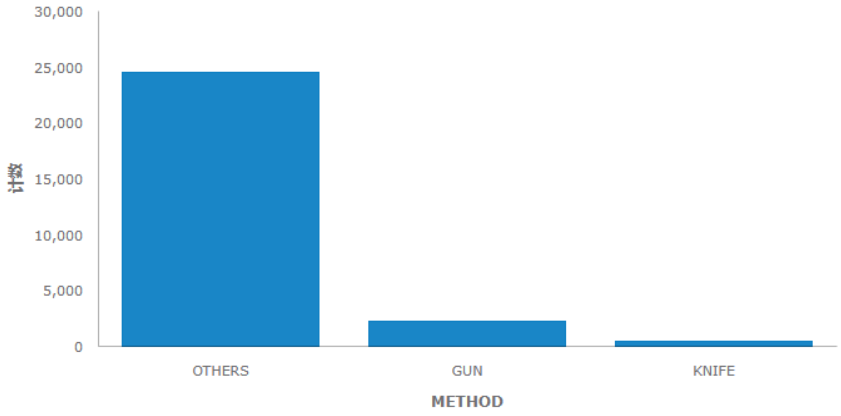


图 4 Washington DC 2022 Crime Incidents 犯罪手段数量统计图

2. Washington DC AirBnb 租房数据介绍

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.

图 5 Washington DC AirBnb 租房数据

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。

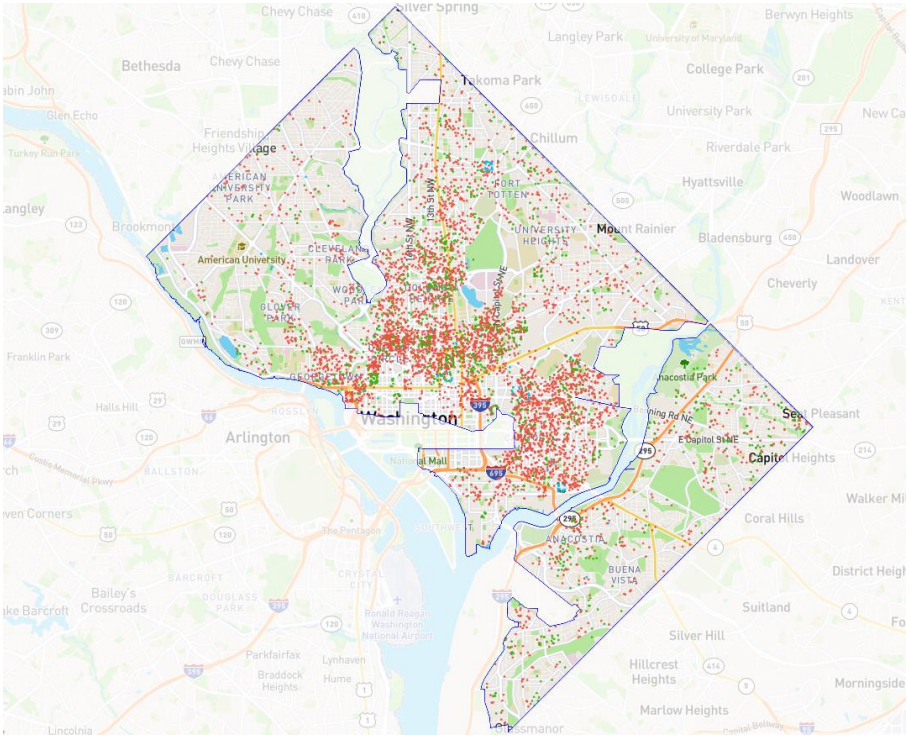


图 6 Washington DC AirBnb 租房位置分布图

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。

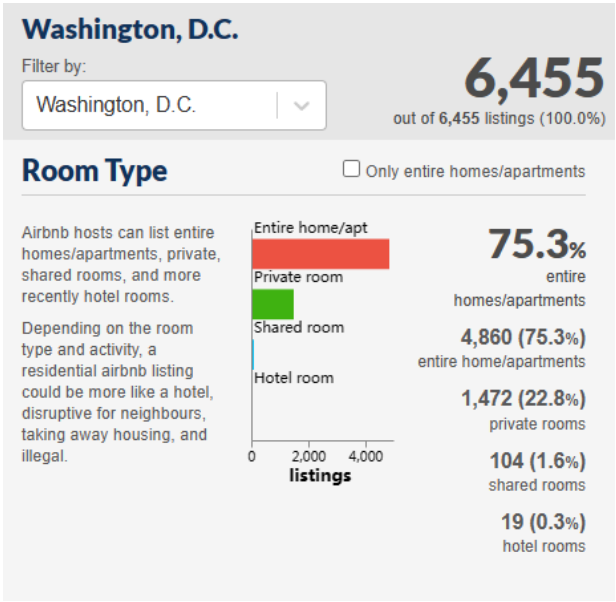


图 7 Washington DC AirBnb 租房类型统计图

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。

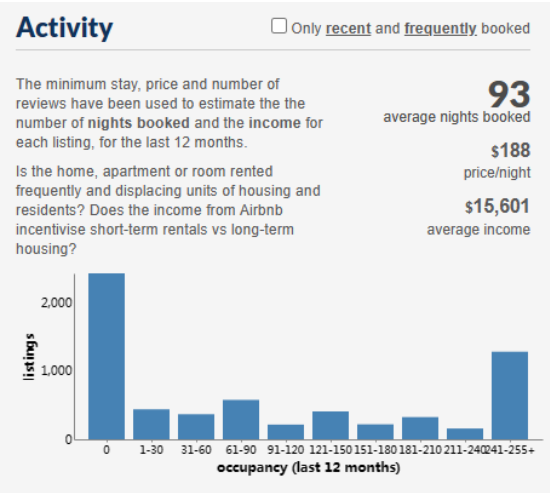


图 8 Washington DC AirBnb 房源租期分布图

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。

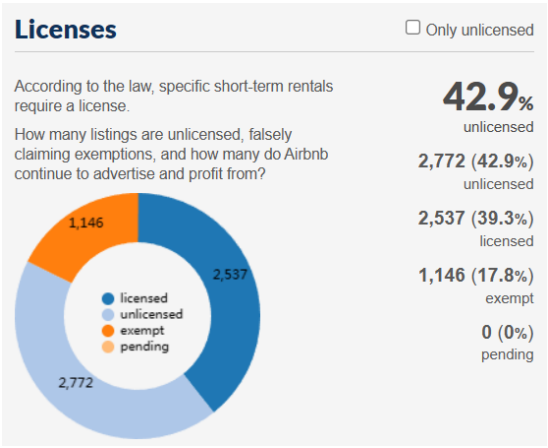


图 9 Washington DC AirBnb 房源许可证饼图

所有房源当中，短期出租占据了大部分（53.3%）。

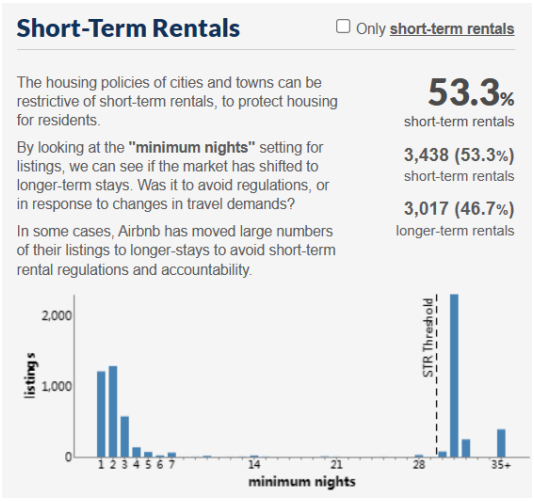
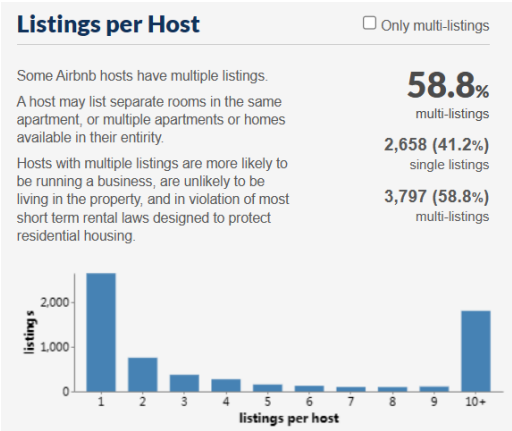
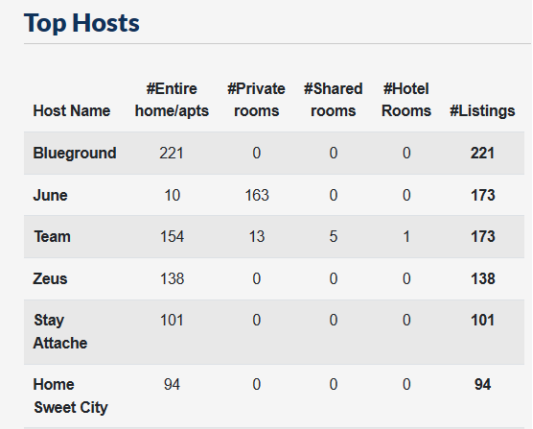


图 10 Washington DC AirBnb 房源租期柱状图

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

图 11 Washington DC AirBnb 房东拥有房源数量统计图

房源数据（listing.csv）主要有如下关键字段：

表 3 Washington DC AirBnb 房源字段表

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。

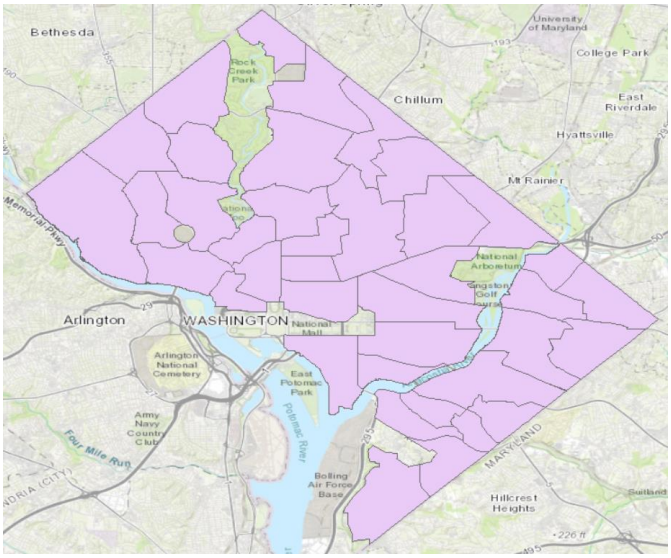


图 12 Washington DC 区域划分图

三、 数据分析

1. Washington DC Crime 数据空间分析

对于 Washington DC 的犯罪数据，我做了如下空间分析。

在 arcgis 当中，通过合并工具，合并 2014-2023 年十年的犯罪数据，并且可视化在地图上，形成点位分布图。同时，我也展示了每一年的犯罪案件点位的分布图。可以不难看出，每年其实犯罪点位的分布都相差不大，基本集中在国会大厦、白宫等中心地带。

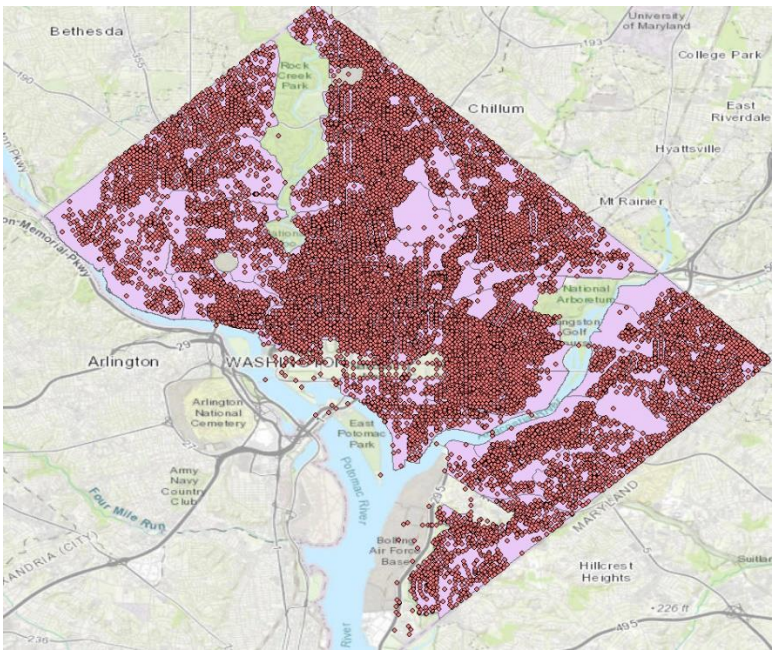


图 13 Washington DC 2014-2023 年犯罪点位分布汇总图

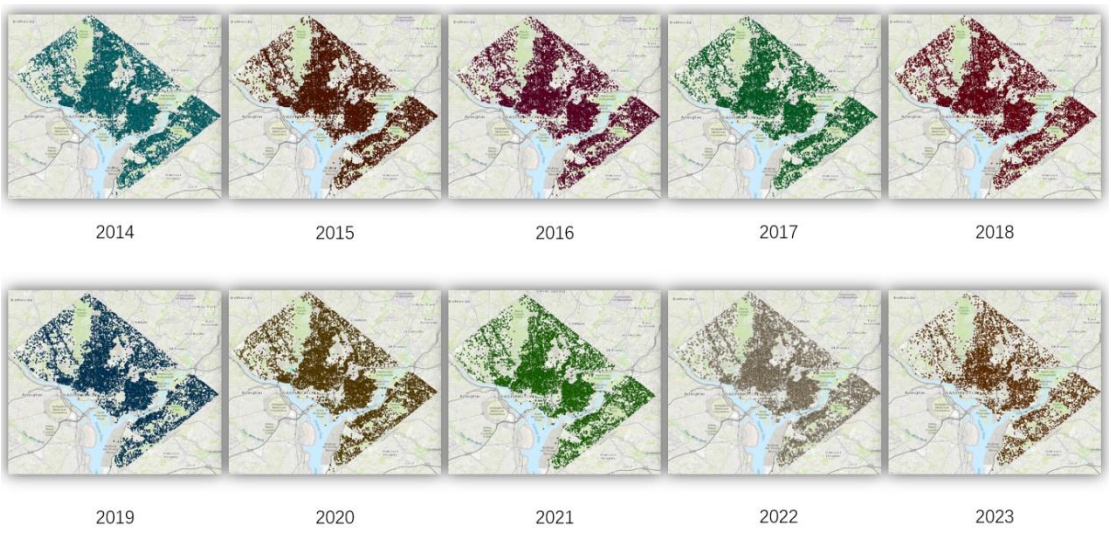


图 14 Washington DC 2014-2023 年犯罪点位分布对比图

为了更好的得到各个地方犯罪数量的一个分布，热力图无疑是一个很好的分析方式。下图为 2014-2023 年华盛顿特区所有犯罪数据绘制而成的热力图。从热力图中可以看到，犯罪案件的地点位置在市中心较为密集，在华盛顿的边缘地区较为稀疏。但是较为例外的是，右下角地区的边缘，犯罪事件还是较为密集的。

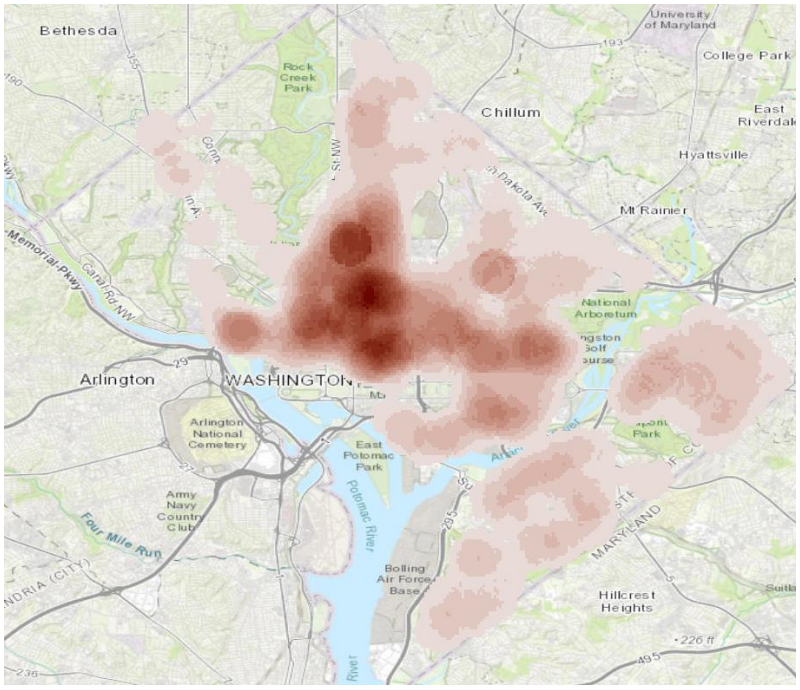


图 15 Washington DC 2014-2023 年犯罪点位热力图

同时，我还根据余弦距离绘制了热点图，如下图所示。根据热点图，其实我们可以看到和上面一致的分布规律，印证了我们上述的分析正确性。

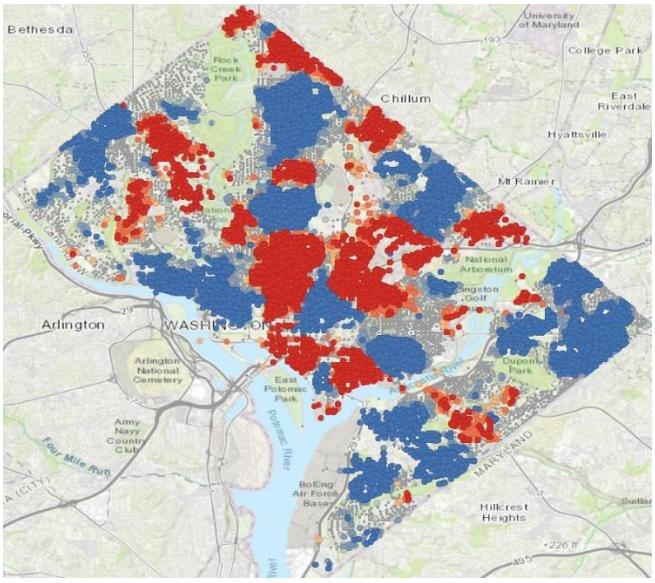


图 15 Washington DC 2014-2023 年犯罪点位热点图

2. Washington DC Crime 数据时序分析

想要了解这 10 年来华盛顿特区犯罪数据时序规律，我做了如下时序分析。首先按照年份统计每年发生犯罪案件的数量，绘制成柱状图，如下图所示。其中 2023 年由于只统计到 5 月份，所以数量较少。整体上犯罪数据数量呈下降趋势。值得注意的是在 2016-2017 和 2019-2020 有两次较为明显的下降，后者可能是由于疫情导致，前者的原因有待考究。下面的新闻其实可以验证我们分析的正确性。

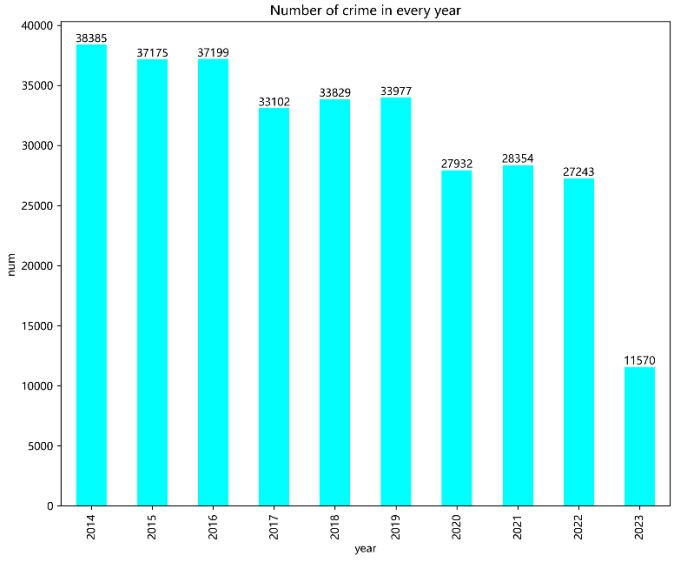


图 16 Washington DC 2014-2023 年犯罪数量年份统计柱状图

根据美国联邦调查局FBI报道，2015和2016连续两年出现犯罪率上涨升的情况，到了**2017年美国各地报告了125万起暴力犯罪，每10万美国居民中有383起暴力犯罪**，比2016年略有下降。

图 17 新闻截图

下图为统计的十年间每个月的犯罪数量，绘制成柱状图。我们可以看到，按照月份进行统计时，可以看出明显的时序规律：每年犯罪数量都是先增后减，形成一个拱门形，并且大概在 7、8、9 月份达到犯罪数量高峰。我认为可能是由于夏季人们外出整体流量较多，导致犯罪数量较多。

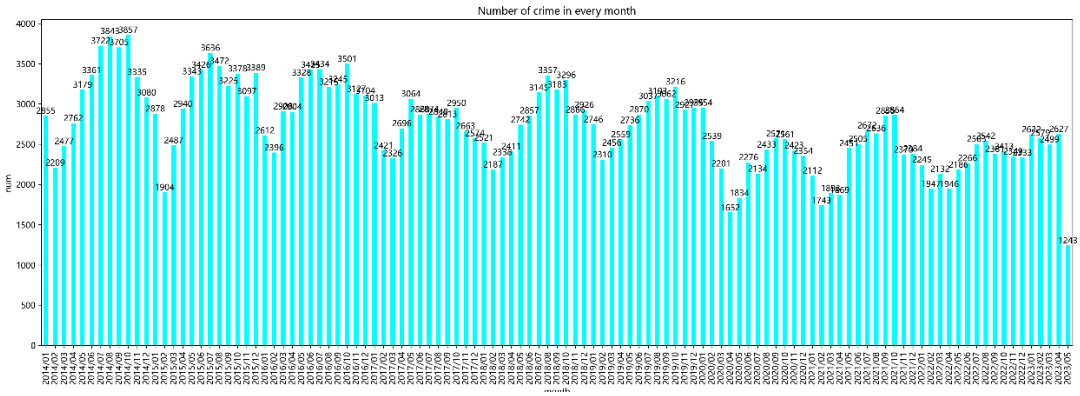


图 18 Washington DC 2014-2023 年犯罪数量月份统计柱状图

以天为单位进行统计的时候，我们可以看到犯罪数量整体以一种波的形式进行波动，可以印证的是，整个犯罪数据确实具有某种时序规律。

图 19 Washington DC 2014-2023 年犯罪数量日期统计柱状图

根据上述粗略的分析，我对每一年的犯罪数据以天为单位进行了时序分解，以得到它们的趋势分量、季节分量、残差分量，有助于我更好的分析犯罪数据的时序规律。如下图所示，以 2022 年数据为例（我对每一年的数据都做了分解和预测，如果想了解更多的实验结果可见 result 文件夹），时序分解的结果如下图所示。

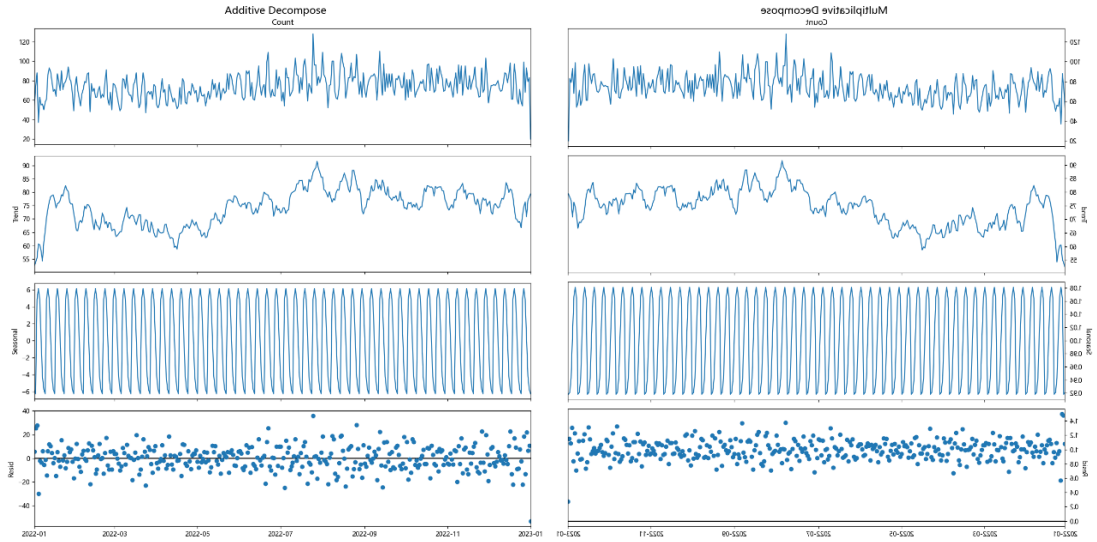


图 19 Washington DC 2022 时序分解图（左为加性分解、右为多项式分解）

分解结果中：

- 季节性分量：显示时间序列数据在每个季节周期中的重复模式。
- 趋势分量：显示时间序列数据中的长期趋势或变化方向。
- 残差分量：表示无法由季节性和趋势性解释的剩余部分，可以视为随机波动或噪声。

可以较为明显的看出，其季节性分量是具有很好的重复模式的。在此基础上，我也对每一年的数据进行了 Arima 时序预测，得到后面 20 天的犯罪数量，这里还是以 2022 年的数据作为示例，进行分析。从预测结果来看，整体上还是比较符合时序规律的。非常值得注意的是 2022 年 8 月犯罪数据有一个非常明显的剧增，可以通过下面的新闻看到，我们分析的正确性。

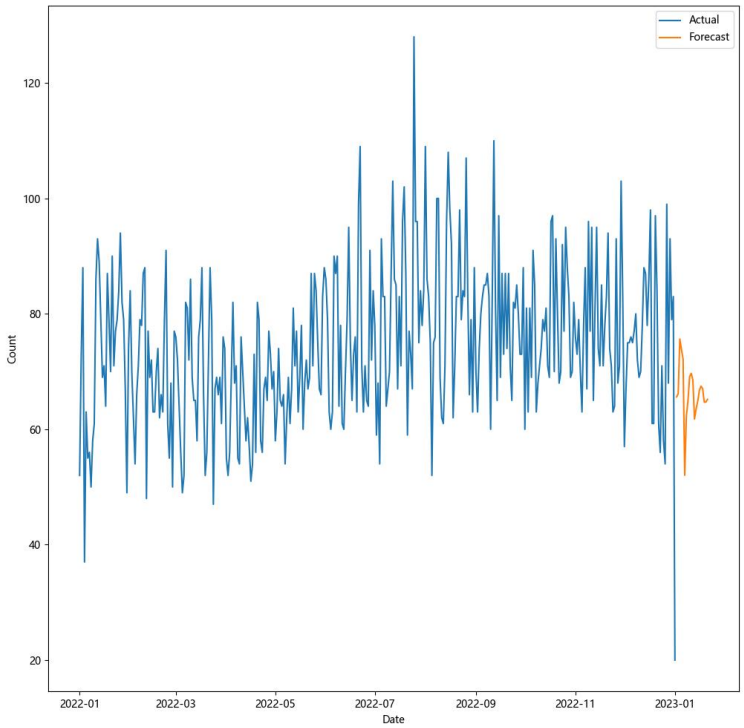
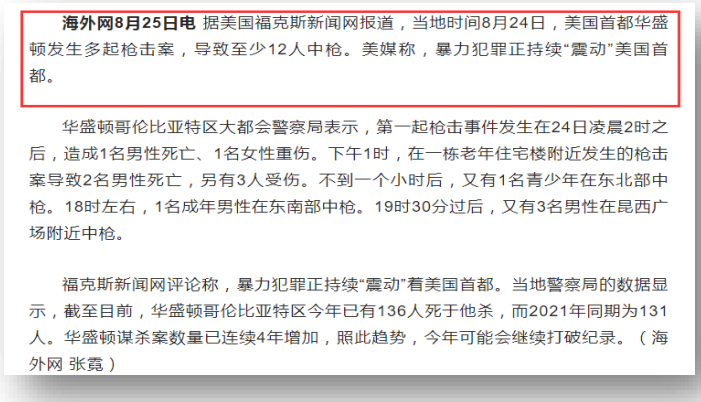


图 20 Washington DC 2022 时序预测图



海外网8月25日电 据美国福克斯新闻网报道，当地时间8月24日，美国首都华盛顿发生多起枪击案，导致至少12人中枪。美媒称，暴力犯罪正持续"震动"美国首都。

华盛顿哥伦比亚特区大都会警察局表示，第一起枪击事件发生在24日凌晨2时之后，造成1名男性死亡、1名女性重伤。下午1时，在一栋老年住宅楼附近发生的枪击案导致2名男性死亡，另有3人受伤。不到一个小时后，又有1名青少年在东北部中枪。18时左右，1名成年男性在东南部中枪。19时30分过后，又有3名男性在昆西广场附近中枪。

福克斯新闻网评论称，暴力犯罪正持续"震动"着美国首都。当地警察局的数据显示，截至目前，华盛顿哥伦比亚特区今年已有136人死于他杀，而2021年同期为131人。华盛顿谋杀案数量已连续4年增加，照此趋势，今年可能会继续打破纪录。（海外网 张霓）

图 21 新闻截图

3. Washington DC AirBnb 数据空间分析

首先为了方便我们后续推荐住房的分析，根据爱彼迎数据中的区域信息，我将 Washington DC 分为了 39 个区域，并且分别编号，如下图所示。

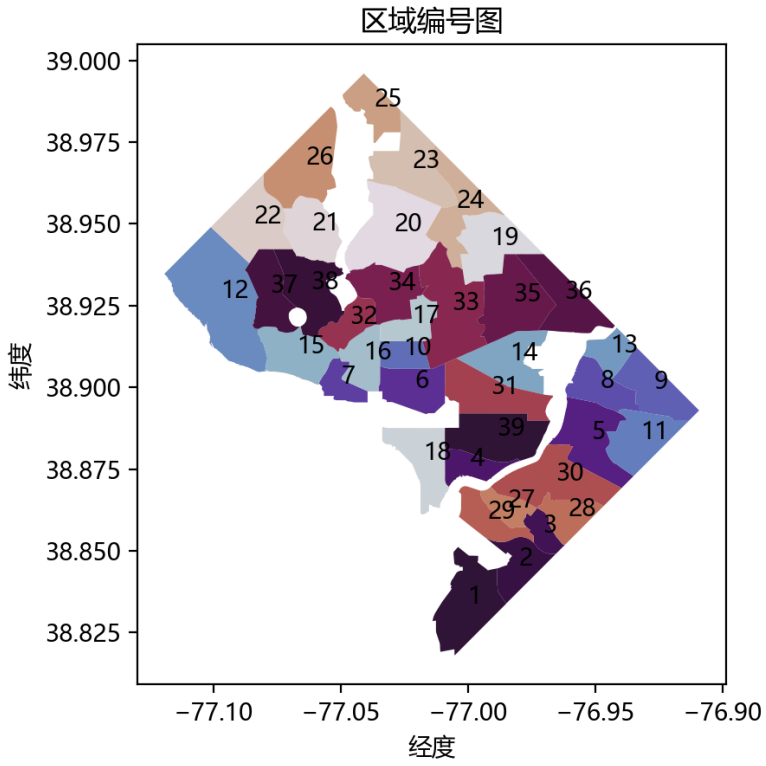


图 22 Washtington DC 区域编号图

然后，我使用 21-23 年的犯罪数据（近几年的犯罪数据对于后续的选址比较有说服力）和 23 年更新的 AirBnb 租房数据与区域信息一起可视化。可以看到，发现犯罪事件和房源的集中区域都在相同地方(国会大厦、博物馆、购物中心等)。

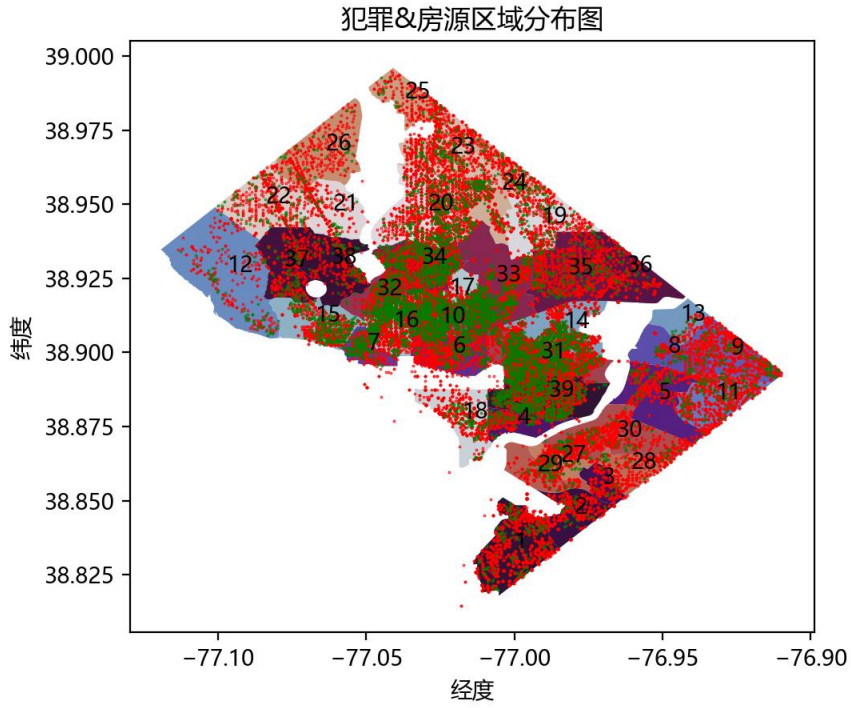


图 23 Washington DC 房源、犯罪事件区域分布图

为了方便后续的计算每个区域的危险值和危险率，需要统计每个区域的犯罪事件和房源数量，这里我将其绘制成三维柱状图，并且输出结果表格。通过结果

可以看到，犯罪数量最多的区域时 31 号区域（4870 起），即为美国国会大厦所在区域，说明我们上面的犯罪数据空间分析是正确的。而房源最多的区域也是 31 号区域（720 个），说明确实人群越密集的地方房源越多也越容易发生犯罪。

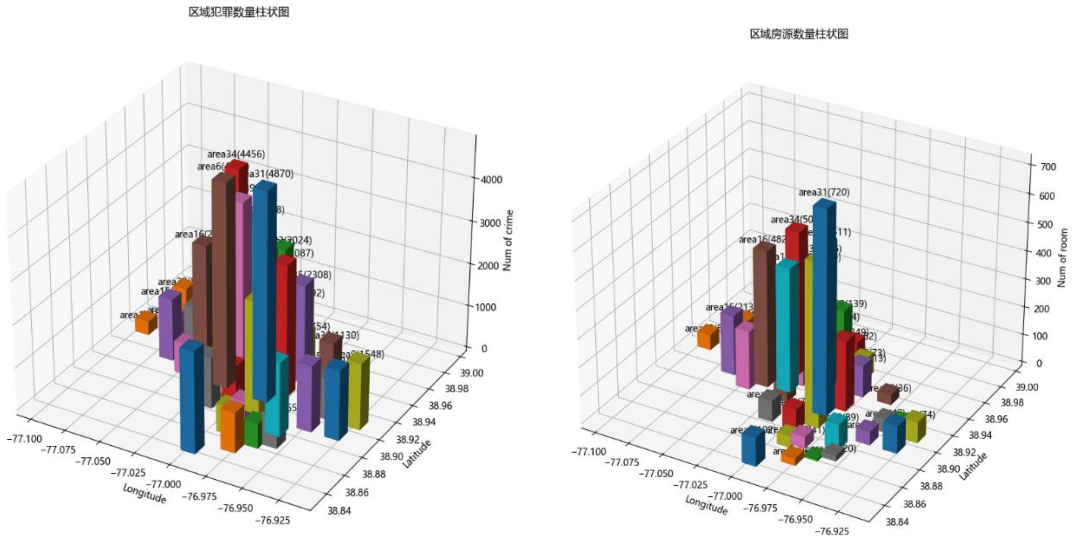


图 24 Washington DC 各区域犯罪（左）、房源（右）数量统计三维柱状图

Area_ID	Num of Crime	Num of Room
1	2365	102
2	943	26
3	582	7
4	1209	77
5	1545	48
6	4754	258
7	729	207
8	947	47
9	1548	74
10	2812	440
11	1633	96
12	320	55
13	101	12
14	3087	249
15	1451	213
16	2991	482
17	3963	239
18	1196	74
19	754	73
20	2978	339
21	522	32
22	810	62
23	2024	139

图 25 Washington DC 各区域犯罪、房源数量统计图

4. Washington DC AirBnb 数据主题挖掘

在爱彼迎的租房数据当中，包含了历史租客对房源的评价信息，一共近 33w 条评论，由此借助上课所学习的 lda 模型，来对这些评论数据进行主题挖掘，来分析分析这些评论的主要观点是什么。

具体构建方法：对于每条评论数据，进行数据清理等操作（小写化、去符号、去停用词等，在此就不过多赘述），然后提取其中的所有非中性词组成一个文档，然后进行字典的生成、词袋构建、lda 模型的生成和训练、最后绘制词云打印主题、并且进行 pcoa 主题评价。这里判断是否为中性词主要用到了 nltk 库中的情感分析，在后续计算评价得分时详细描述。

根据上述的步骤，得到如下主题。我们可以很清晰的看到，挖掘出的这 6 个主题中基本都是 good、great、perfect、nice 等褒义词，基本可以断定整体来

说爱彼迎上的用户大部分评价都是好评，说明爱彼迎房源质量整体还不错。

```
[(0, '0.078*"beautiful" + 0.069*"amazing" + 0.055*"lovely" + 0.051*"clean" + 0.050*"enjoyed" + 0.049*"comfortable" + 0.047*"recommend" + 0.039*"definitely" + 0.034*"fantastic" + 0.029*"well"'),
 (1, '0.225*"perfect" + 0.085*"wonderful" + 0.054*"clean" + 0.046*"comfortable" + 0.039*"recommend" + 0.037*"definitely" + 0.034*"love" + 0.024*"well" + 0.019*"perfectly" + 0.019*"loved"'),
 (2, '0.266*"nice" + 0.161*"good" + 0.076*"clean" + 0.041*"comfortable" + 0.030*"well" + 0.022*"pretty" + 0.020*"recommend" + 0.018*"enjoyed" + 0.017*"like" + 0.013*"responsive"'),
 (3, '0.166*"easy" + 0.149*"great" + 0.096*"clean" + 0.087*"super" + 0.079*"comfortable" + 0.037*"awesome" + 0.035*"well" + 0.034*"loved" + 0.033*"cute" + 0.033*"thanks"'),
 (4, '0.534*"great" + 0.077*"clean" + 0.059*"recommend" + 0.053*"definitely" + 0.051*"excellent" + 0.043*"responsive" + 0.042*"helpful" + 0.023*"value" + 0.023*"friendly" + 0.018*"recommended"'),
 (5, '0.092*"no" + 0.064*"like" + 0.045*"want" + 0.034*"best" + 0.025*"help" + 0.022*"problem" + 0.019*"sure" + 0.018*"free" + 0.018*"happy" + 0.017*"fine"')]
```

图 26 Washington DC 房源评论数据挖掘

图 27 主题词云

然后我对训练的 lda 模型进行主坐标分析(PCOA)，得到的效果如下图所示。可以看到几个主题之间除了 5 和 2 之外都有较好的区分性。

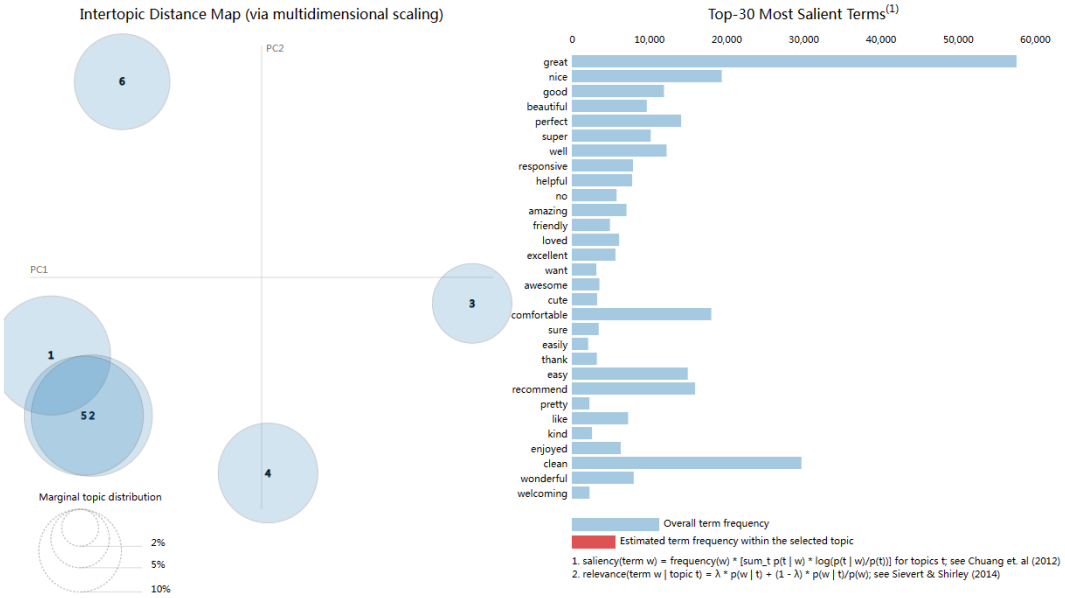


图 28 lda 模型 pcoa 分析图

四、 选址算法

1. 算法流程图

本次的选址算法旨在保证游客安全、租房成本、距离成本在可接受范围内的同时，尽可能的衡量房源的危险性、价格、距离目的地远近、评价得分等四个要素来为游客推荐几所优秀房源。同时考虑到游客的个性化安全需求，本算法还设定个性化参数以保证对大多数游客适用。本次算法的流程图如下：

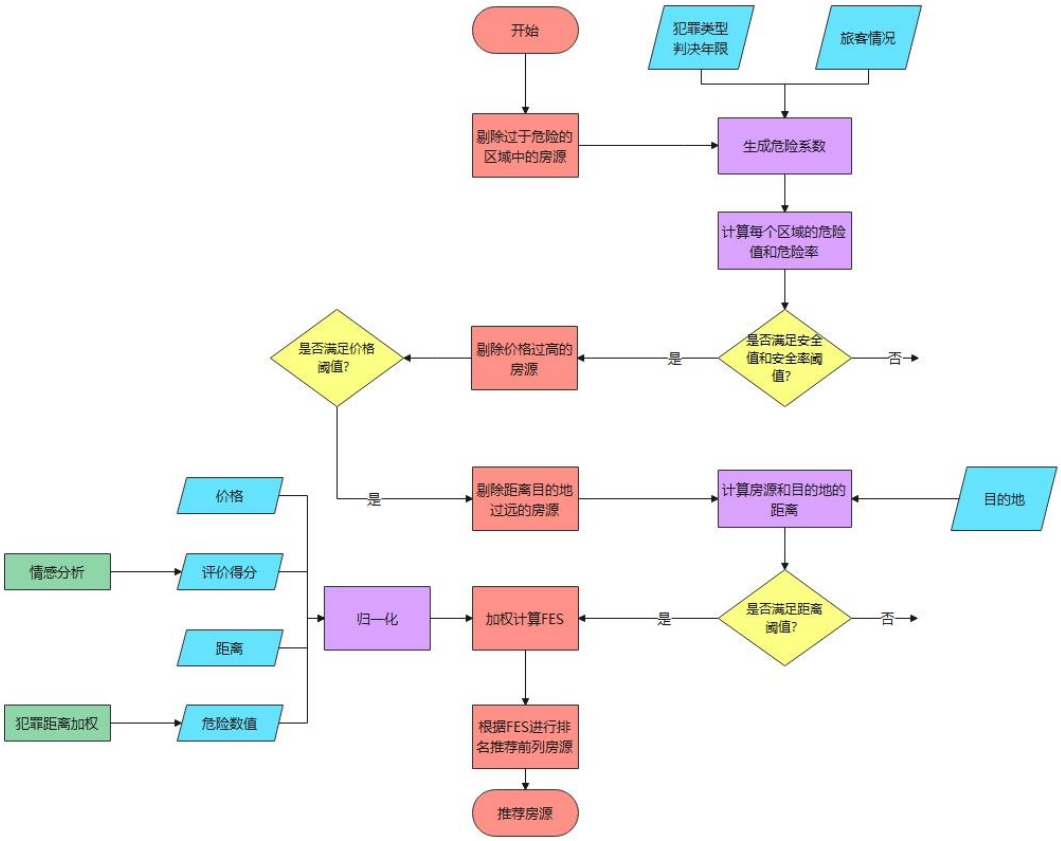


图 29 选址算法流程图

2. 剔除危险区域

安全是所有的根本，犯罪事件发生次数太多、重大犯罪案件发生的区域游客都应该尽可能的避免在附近居住。本着为游客安全着想的原则，本算法第一个步骤就是剔除太过于危险的区域。

由于不同的犯罪类型对游客的影响有很大的差异性，因此本算法设置危险系数来表示不同犯罪类型对于游客的影响程度不同。在生成危险系数的时候，为了让危险值更有说服力，我查询了数据中 9 种犯罪类型的大致刑期，以其作为危险系数。同时考虑到每个案件哪怕类型相同但是程度也不一，所以采用随机数指定系数范围的形式来模拟这一变化。为了使系数更具说服力，我选择计算 10 次后的危险系数平均值为后续计算使用。值得注意的是，由于谋杀罪非常严重，所我将其系数设置为 1000（很大）。

更进一步考虑，游客不同的个人情况也会导致危险系数的不同。这里为了游

客的个性化安全考虑，设置了三个个性化参数：性别、是否自驾游、是否携带贵重物品。这三个参数默认为 1，既没有影响。但是，

- 当游客为女性时，性虐待系数变为原来 5 倍
- 当游客为自驾游时，汽车盗窃相关犯罪危险系数变为原来 2 倍
- 当游客携带贵重物品时，盗窃、抢劫相关罪行危险系数变为原来 1.5 倍

```python
def generate_dangerous_coefficient(gender=0,car=0,rich=0):
    '''
        根据犯罪的轻重(判期长短)，生成危险系数(会根据游客的特殊性进行变化)
        女性性虐待系数变大
        自驾游车辆盗窃系数变大
        带有贵重物品的盗窃抢劫等系数变大
    '''
    #个性化参数
    additional_gender=max(5*gender,1)
    additional_car=max(2*car,1)
    additional_rich=max(1.5*rich,1)
    # 危险系数
    THEFT_OTHER=random.uniform(0.1,2)*additional_rich
    THEFT_AUTO=random.uniform(10,15)*additional_car
    MOTOR_VEHICLE_THEFT=random.uniform(5,15)*additional_car
    BURGLARY=random.uniform(5,15)*additional_rich
    ARSON=random.uniform(10,30)
    ASSAULT_DANGEROUS_WEAPON=random.uniform(10,30)
    HOMICIDE=1000
    SEX_ABUSE=random.uniform(5,15)*additional_gender
    ROBBERY= random.uniform(5,15)*additional_car
    dangerous_coefficient={
        'THEFT/OTHER':THEFT_OTHER,
        'THEFT F/AUTO':THEFT_AUTO,
        'MOTOR VEHICLE THEFT':MOTOR_VEHICLE_THEFT,
        'BURGLARY':BURGLARY,
        'ARSON':ARSON,
        'ASSAULT W/DANGEROUS WEAPON':ASSAULT_DANGEROUS_WEAPON,
        'HOMICIDE':HOMICIDE,
        'SEX ABUSE':SEX_ABUSE,
        'ROBBERY':ROBBERY
    }
    return dangerous_coefficient
```

图 30 生成危险系数代码

对于区域危险性本算法提出两个指标进行衡量：危险值和危险率。

- 危险值（dangerous value）：一个区域所有犯罪案件的危险系数总和。
- 危险率（dangerous rate）：一个区域所有犯罪案件的危险系数总和除以该区域的犯罪总数，即该区域的平均危险系数。

$$dangerous\_value[area] = \sum_{crime}^{crimesInArea} dangerous\_coefficient[crime[type]]$$

$$dangerous\_rate[area] = \frac{\sum_{crime}^{crimesInArea} dangerous\_coefficient[crime[type]]}{num(crimesInArea)}$$

按照上述要求进行计算，最后得区域的危险值、危险率的柱状图。危险值高说明这个区域发生的犯罪事件多或者发生的犯罪事件较为严重。如果仅仅危险值低，也不一定真的安全。例如，13 号区域危险值很低，但是危险率很高，说明这里发生的案件都是危险系数很高的案件。所以必须危险值和危险率都低于安全阈值我才认为这个区域是适合居住的。如下图所示，红色柱子代表每个区域的危险值，黄色柱子代表每个区域的危险率，蓝色虚线代表安全值阈值，绿色虚线代表安全率阈值。

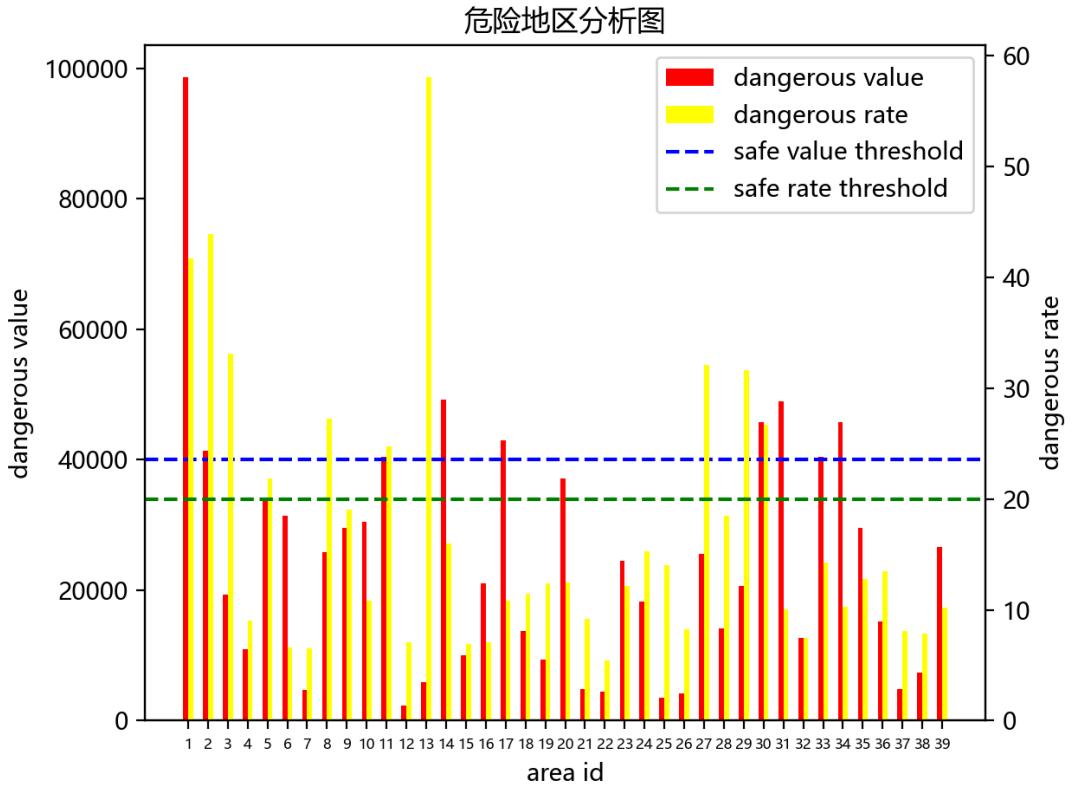


图 31 区域危险值&率柱状图

将那些不满足阈值条件的区域中的房源剔除（不考虑），在下图中标注出安全区域（绿色）和危险区域（红色）。可以看到，基本上市中心和右下角边缘地区都是危险性较高的区域。

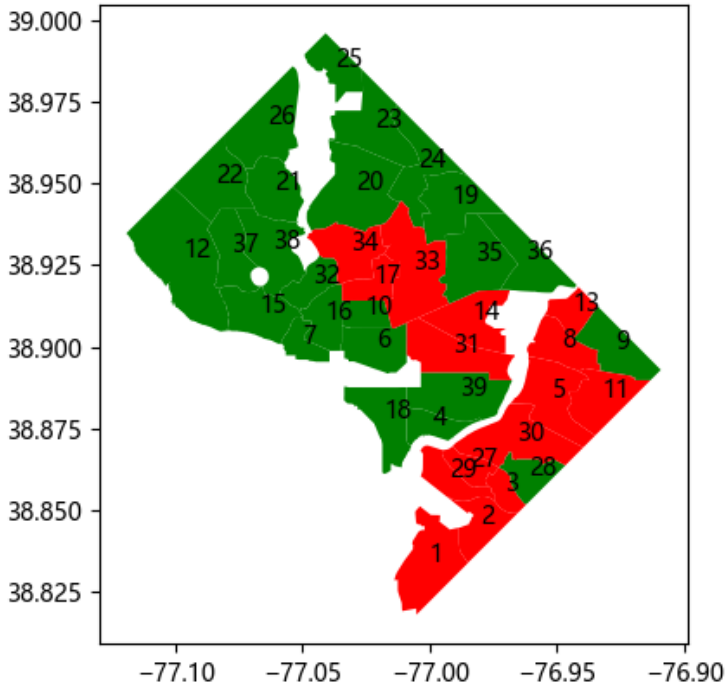


图 32 安全区域标注图

为了体现个性化参数的差异，下面我做了这样的实验。上述的实验内容弄都是在没有设置个性化参数的情况下做的，下面我做了一组当游客为女性且自驾游且带有贵重物品的实验，得到的安全地区标注图如下。我们可以看到，很明显安全区域相较于没有特殊安全需求的图 32 要少很多。

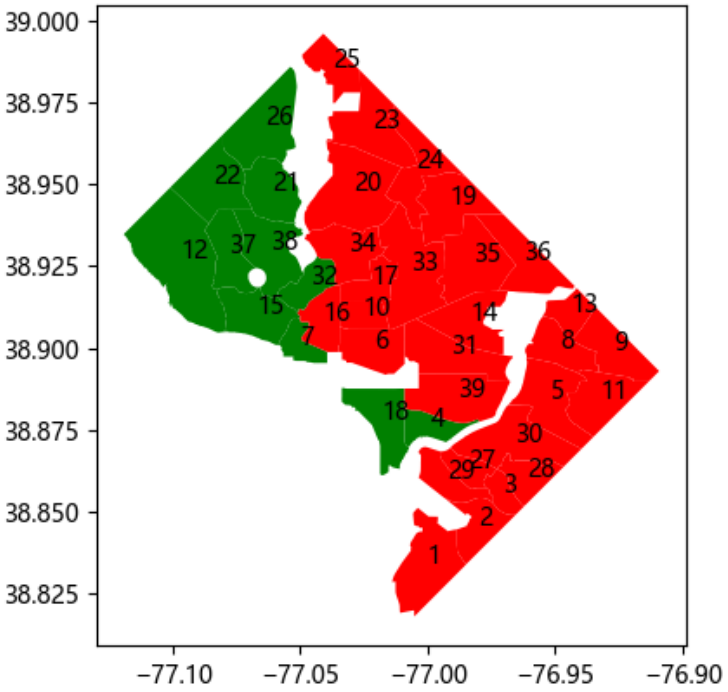


图 33 女性、自驾游、带有贵重物品游客安全区域标注

3. 剔除较贵区域

由于价格也是游客考虑的重要因素之一，过于贵的租房价格有些游客承担不起。因此，在本算法中也让游客自己设置价格阈值，如果高于价格阈值的房源我们暂不考虑。

首先对房源的不同种类和总体的价格进行简单的分析，绘制如下箱型图和表格。整体来说房源的价格范围较大，从 0-$7500 不等。就房源种类而言，整房出租平均价格最高且波动最大，私房出租其次；酒店平均价格最低且波动较小。这可能是由于酒店还是属于管制房源，而其他房源都是私人房源的缘故。

Room Type	Average	Min	Max	Variance
Entire home/apt	207.90617283950618	10	7500	66830.16362444748
Private room	132.9891304347826	20	2000	34020.546077504725
Shared room	54.99038461538461	25	140	648.009522928994
Hotel room	68.10526315789474	0	489	12118.094182825487
SUM	187.9468628969791	0	7500	59436.694465371605

图 34 房源价格五数一览图

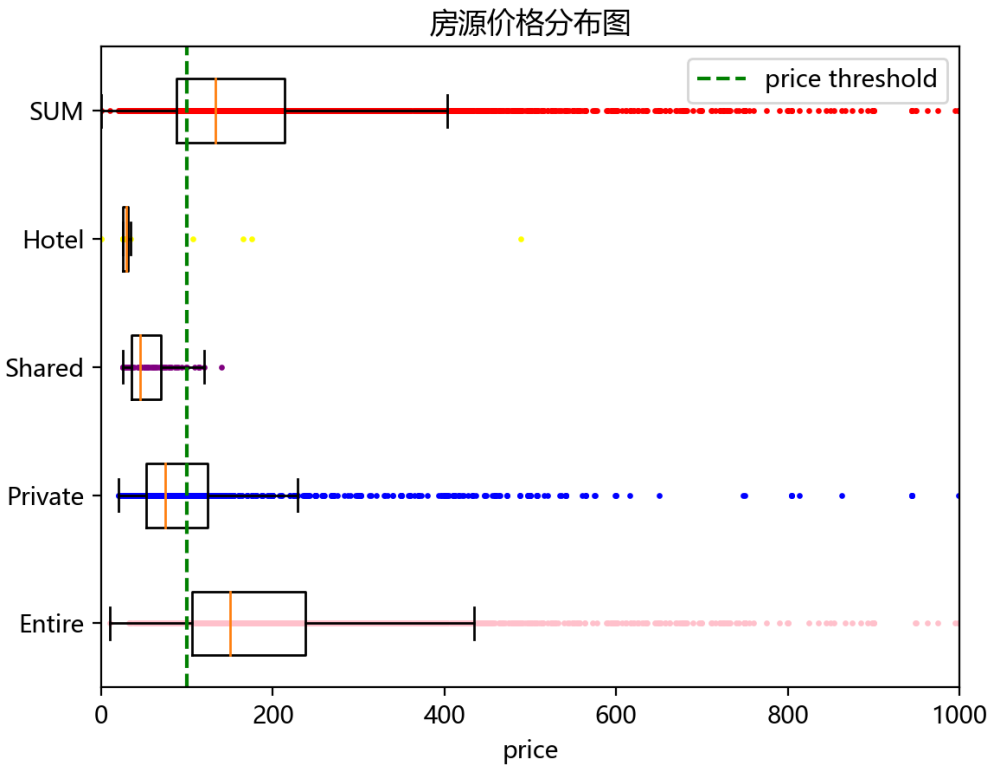


图 35 房源价格分类箱型图

剔除掉上述价格阈值之上的房源之后，在图中进行标注，绿色为可选房源，红色为过贵房源（这里使用$100 作为价格阈值）。可以看到，过贵房源集中在市中心，郊区的房子普遍便宜。

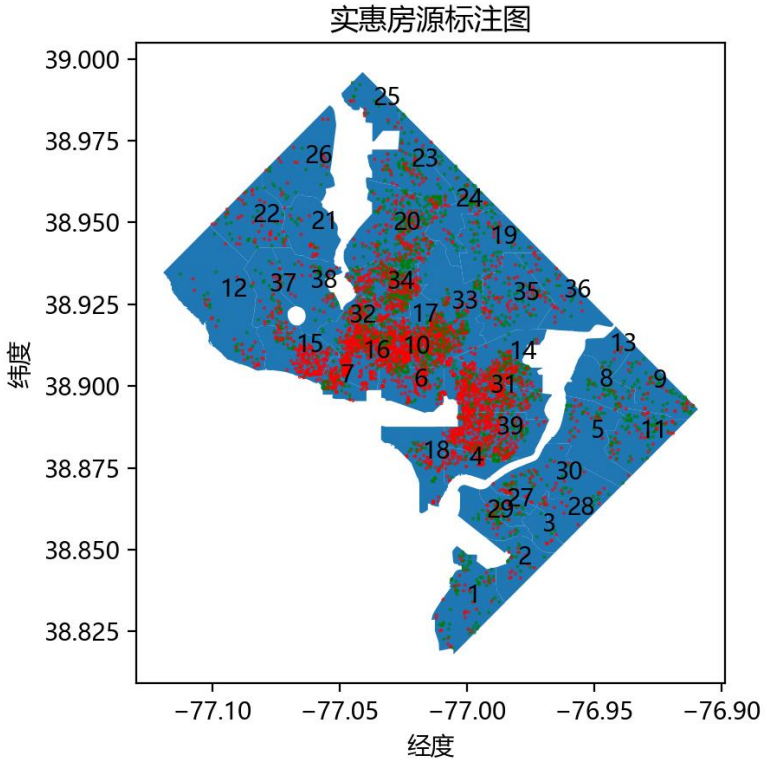


图 36 实惠房源标注图

4. 剔除较远区域

对于每个旅客来说都不想自己的住处距离想要去的景区太过遥远，不然大好时光就浪费在路途当中。本算法中也让旅客自己选择距离阈值，剔除掉那些距离目的地太过遥远的房源。

这里使用经纬度计算房源和目的地的直线距离（这里以“美国国会大厦”作为目的地，距离阈值为 5 公里）。将距离较近的房（绿色）和剔除的较远房源（红色）标注在图中。

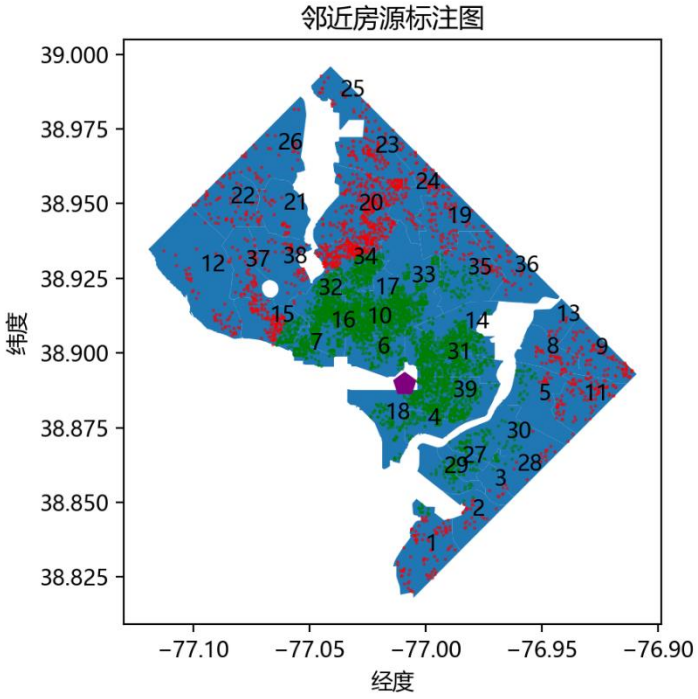


图 37 邻近房源标注图

5. 计算评价得分

对于大多数房源来说，历史租客的评价在很大程度上能反映该租房的整体质量，对于旅客来说，一个房源的历史评价也是重要的参考依据之一。

本算法通过 nltk 的情感分析功能，来计算每条评论的情感得分，最后将每个房源的所有评论得分进行累加，得到房源的评价得分。值得注意的是，对于一个词而言，compound 得分大于 0 则其为积极的，得分小于 0 则为消极的，得分等于 0 则为中性的。

```
def get_sentiment_score(text):
    '''
    计算compound"情感得分
    >0: 积极的
    <0: 消极的
    =0: 中性的
    '''
    sid = SentimentIntensityAnalyzer()
    sentiment_scores = sid.polarity_scores(text)
    return sentiment_scores['compound']
```

图 38 获取情感得分代码

按照上述步骤得到每个房源的评价得分，直方图、饼图和五数概括如下。从下面的结果可以知道，绝大多数的房源的正面评价都是多于负面评价的，说明爱彼迎的房源质量整体较好。从饼图来看，大多数房源的评价得分还是在 0-50 的区间，做的极好的房源还是占少数。

Name	Average	Min	Max	Variance
Room Scores	44.3061859448554	-1.505	645.9660000000034	3782.50594285203

图 39 房源评价得分房源一览

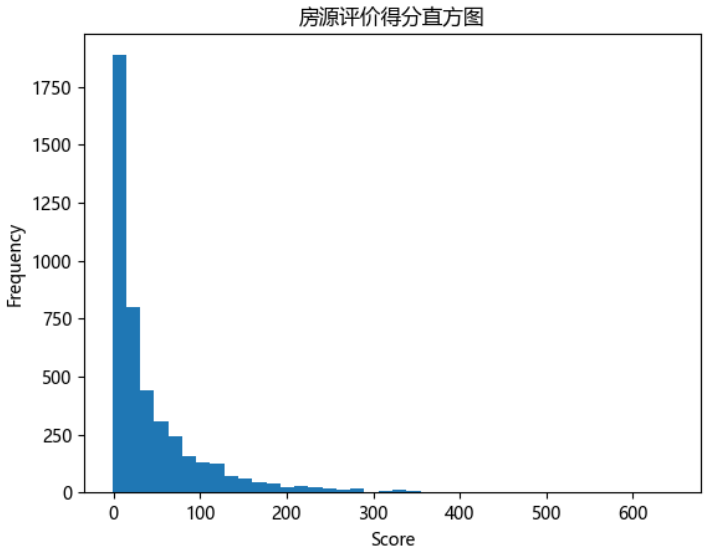


图 40 房源评价得分分布直方图

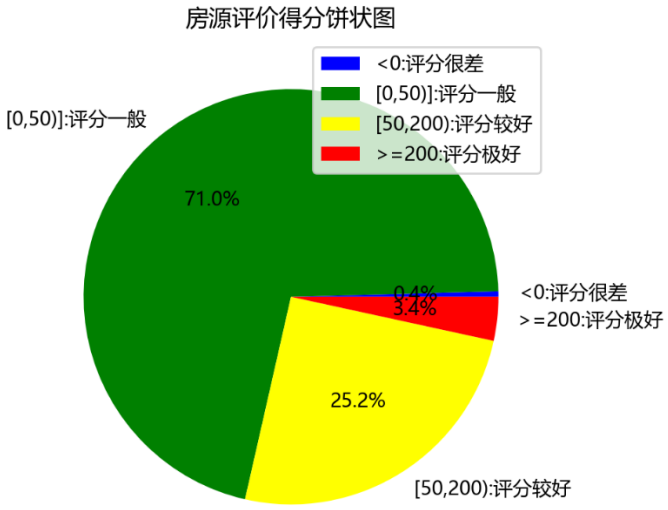


图 41 房源评价得分分布饼图

6. 计算最终得分

经过上述的筛选，我们提出了太过危险的、太贵的、太远的房源，在剩下的房源当中如何寻找最好的房源进行推荐呢。本算法定义一个指标 FES（Final Evalution Score）来衡量房源的好坏。FES 综合考虑危险值、危险率、价格、距离目的地远近、评价得分等五个因素（其中有两个危险性因素，这样即便是均

匀权重危险性也会占据大比重，符合我们安全为重的理念）。具体的计算方式为，先将上述五个指标归一化到[0,1]的范围内，然后按照下面的公式计算 FES。

$$dangerous\_value = normalize(dangerous\_value)$$
$$dangerous\_rate = normalize(dangerous\_rate)$$
$$price = normalize(price)$$
$$distance = normalize(distance)$$
$$score = normalize(score)$$
$$FES = -a * dangerous\_value - b * dangerous\_rate - c * price - d * distance + e * score$$

得到每个房源的 FES 之后，绘制各种类和所有房源的 FES 箱型图五数一览图。

Room Type	Average	Min	Max	Variance
Entire home/apt	62.353957735831706	0.0	100.0	130.77189556733623
Private room	58.3395272401216	1.2114080843652286	99.09424482110612	186.26012788402255
Shared room	61.862273483018186	42.02007442066928	81.23818915612704	70.24320521265288
Hotel room	73.01859217689758	69.58628667966416	83.37614621361692	15.85451575049802
SUM	61.461974963159726	0.0	100.0	145.3300346451137

图 42 房源 FES 五数一览图

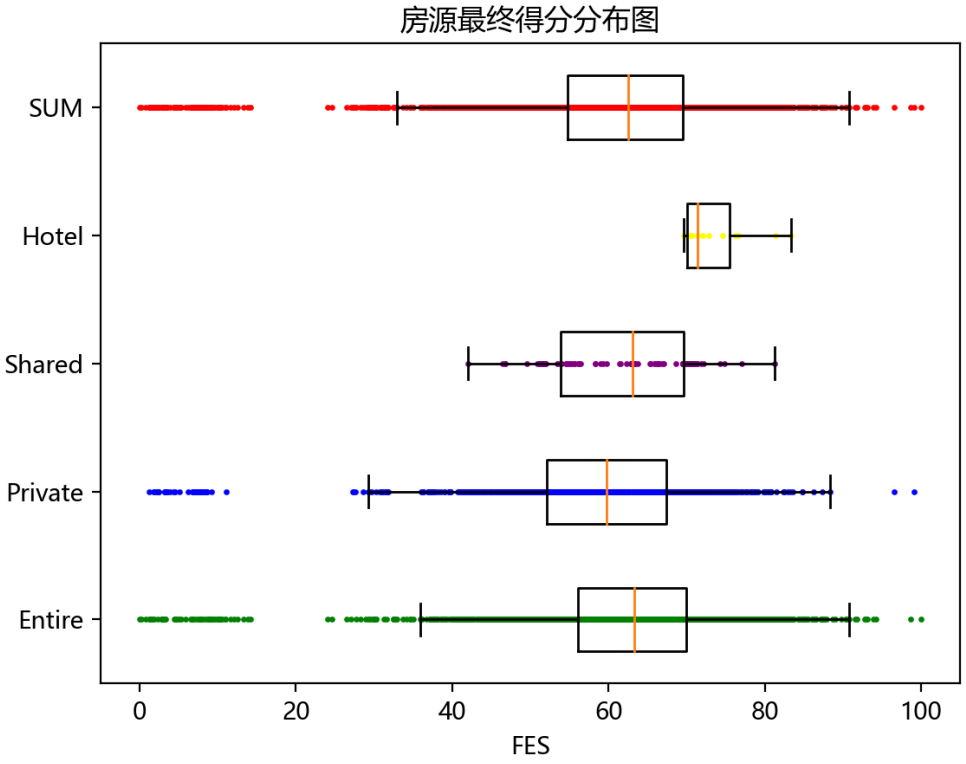


图 43 各类房源 FES 箱型图

可以从上述结果看出：

- 酒店的平均得分最高且最稳定，一方面是由于其数据较少，另外还可能和其为聚集型居住、安保较好，安全性便较高有关系。再加上之前价格分析种提到的，酒店均价最低，所以最后得分最好。
- 合租的公寓得分也比较高，可能是因为多人合住的缘故安全性有保障。
- 相比之下私房和整房的波动较大，而且平均得分较低。一方面是这两者价格较贵，并且单人居住安全性得不到保障。
- 总体而言，百分制的 FES（Final Evaluation Score），所有房源的均分在及

格线之上，还是说明爱彼迎的成功性的。

7. 根据最终得分推荐房源

讲过上述所有的步骤，首先我们筛选出不符合安全阈值、价格阈值、距离阈值的房源，然后在剩下的房源当中按照 FES 从大到小的顺序推荐 FES 最大的 10 所房源给旅客。

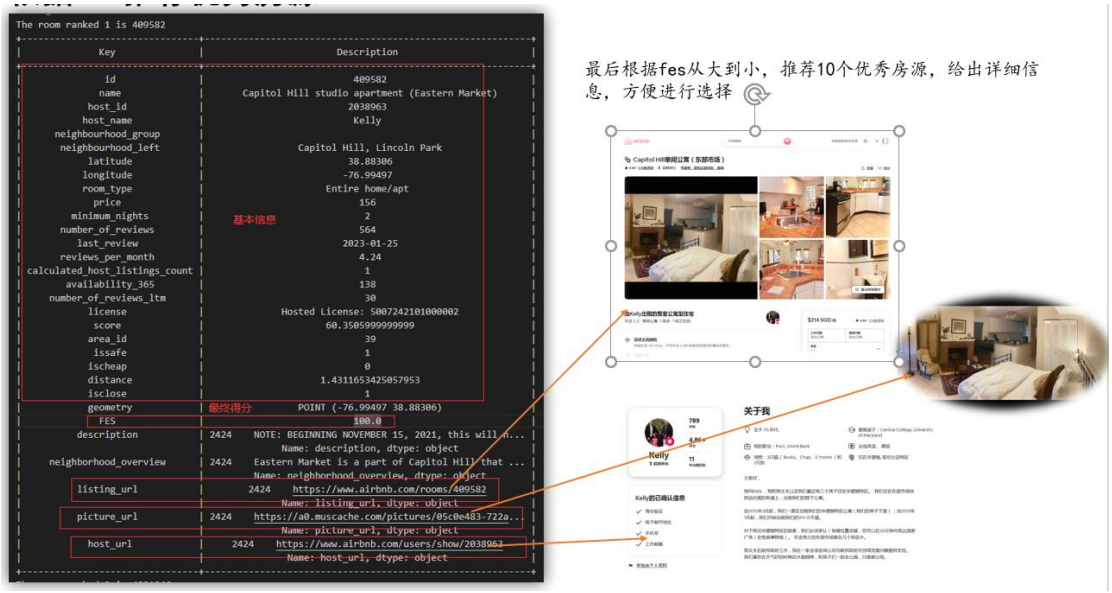


图 44 房源推荐示意图

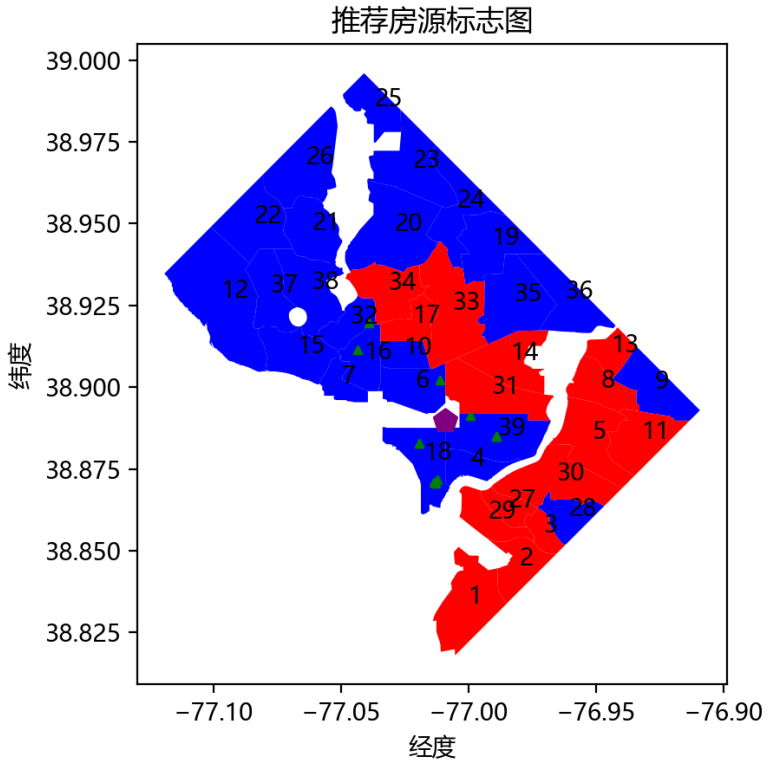


图 45 推荐房源标注图

五、 总结与展望

本次实习经过一周的实践，我更加熟悉了数据挖掘分析方法，也知道了数据处理、挖掘、分析、可视化的具体流程。更加令人感到欣慰的是，我所选择的主题具有一定的实际意义，并且基本顺利完成。当然，在这个选题上还有一些不足，后续可进行改进：

- 在计算距离阈值的时候江直线距离换为通勤距离，这样更符合要求规范
- 整个系统规范化，添加前端，实现软件的外包装

本次实习，我使用数据挖掘、时空数据分析等多种手段，从多种因素来考虑华盛顿地区旅游租房选址问题，这是我第一次使用课程知识来解决一个较为有趣的和生活极其相关的实际问题。这也更让我相信那句话，“在如今这个行业，有人负责提供娱乐消遣，有人负责便捷社会，而也有人负责改变时代！”。我始终坚信，数据是时代的前进的源动力，数据挖掘与分析是时代的先行者！