Location Profiling in Crime Analysis

Jianxu Shi

September 20, 2019

1. Introduction

1.1 Background

In 2001, U.S. Department of Justice published a document titled *Introductory Guide to Crime Analysis and Mapping.* In this guide, Dr. Rachel Boba defined *Crime Analysis* as "the qualitative and quantitative study of crime and law enforcement information in combination with socio-demographic and **spatial** factors to apprehend criminals, prevent crime, reduce disorder, and evaluate organizational procedures." The information such as date, time, location and type of crime is quantitative in that statistics can be used to analyze these variables. On the other hand, narratives of crime reports are considered qualitative data. The author pointed out that spatial information is an important factor in the analysis of crime. It is not only important where a crime takes place but also the characteristics of those places and the environment in which the crime occurs. Thus, examination of spatial data such as street networks, parcel information, school locations, business and residential zoning, among others, is imperative for effective crime analysis. Quantitative spatial analysis of crime data is the subject of this study.

1.2 Problem

The relationship between location and crime (incident) is not well understood, making it difficult for law enforcement agencies to prepare for and adapt to the local environment. We can call it *Location Profiling* as opposed to psychological or behavioral profiling. This study aims to determine if correlation exists between characteristics of a location and occurrence of crime incidents. Such correlation can help predict incidents in places with similar characteristics.

1.3 Interest

The target audiences for this project are law enforcement agencies and society in general. The result can help law enforcement agencies prepare for the expected magnitude and nature of incidents according to the location profile. General public can also draw insight from the location profile and plan area development accordingly. Here are examples of questions this study will try to answer:

- Is there correlation between characteristics of location and number of incidents?
- What are the characteristics of a low-incident area?
- What are the characteristics of a high-incident area?
- Does the presence of certain venues correlate with decrease or increase of incidents?

2. Data acquisition and cleaning

2.1 Data sources

We need two types of information to perform spatial analysis of crime data.

Incident Dataset. These are incident reports released by Sheriff's office or Police Department. Many of them are searchable on-line and some are downloadable. For example, Socrata collects and makes incident datasets available from this link. For this project, a copy of "King County Sheriff's Office incident dataset" in .csv format was downloaded. The file has almost 17K rows of recorded incidents for years 2018-2019 in greater Seattle metropolitan area.

Location Data. Foursquare is a location data provider. Information about a location, such as a city, zip code, or neighborhood can be searched from Foursquare.com or extracted by an API call. For this project, location is defined by the pair (city, zip code) and location profile is represented by number of venues and distribution of venue categories for a particular location.

The definition of location by (city, zip code) is the simplest choice to aggregate locations reported in the incident dataset. However, this definition has met a significant challenge in this project. For different pairs of (city, zip code), the Geolocator software may return the same coordinates. It means that a (city, zip code) pair does not uniquely identify a location or a neighborhood. Adjustment was made in this project to work around this challenge.

2.2 Data cleaning

Because this project has multiple stages of data processing, data inspection and cleaning is a continuous process. The incident dataset downloaded has many problems. There were a lot of missing values. If the missing value is in one of the important fields, such as city, zip, or incident type, the record is removed. The remaining records are checked for consistency as follows.

- The valid range of zip codes in King County should be between 98001 and 98288 inclusive.
- County name was mistakenly used in the "city" field.
- Cross check against unitedstateszipcodes.org for validity of (city, zip code) pair. If invalid, use the address reported in the incident dataset to correct the city or zip code.
- Sometimes a re-organization has happened. For example, Cascade-Fairwood merged into Renton.
- A generic name, such as "Seattle" was used. Change to the primary city associated with the zip code.

The output from the Geolocator was less than robust. From the 208 input (city, zip code) pairs, four generated erroneous coordinates and the remaining 204 produced 100 unique sets of coordinates. The (city, zip code) pair does not uniquely identify a location. In view of this, the data is re-grouped by the 100 unique sets of coordinates and incidents re-distributed to the reduced set of locations.

2.3 Spatial data selection

Once the 100 unique locations were identified by the coordinates, it is straightforward to run API calls to Foursquare and extract nearby venue information as the location profile.

3. Methodology

The methodology can be described in five steps:

a. Load incident dataset, clean any incomplete sections, and group incidents by (city, zip).
b. Extract geographical coordinates from Geolocator; visualize incidents on folium map.
c. Use the coordinates to extract venues around each location in Foursquare to build location profile.
d. Examine venue distribution for characteristics of low-incident and high-incident locations.
e. Explore different regression models to predict incidents based on venue count and venue distribution.

4. Exploratory Data Analysis

4.1 Visualization of incident data

First, let's visualize how incidents are distributed across King County. In Figure 1, each circular marker represents a unique neighborhood location. The color of the marker represents incident count for that neighborhood, with red indicating high count and purple indicating low count. Note the "Seattle" location covers a very large area due to the fact that Geolocator does not differentiate many zip codes inside Seattle. It is not surprising to see "Seattle" in red for the area and population it covers. A few other population centers such as Auburn, Kent, Renton and South Seattle also show high incident counts. The further away from the population centers, the lower the incident count.

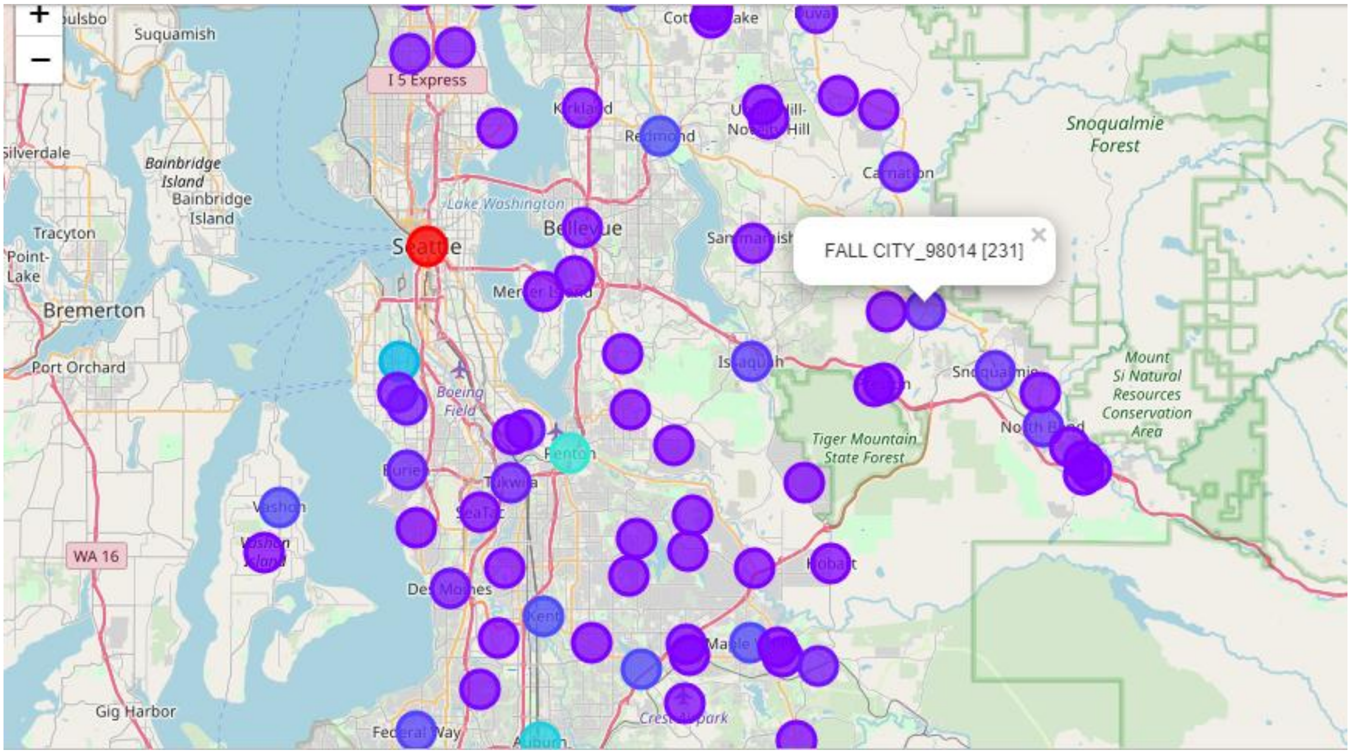


Figure 1 Visualizing incidents in King County (Red = high count, Purple = low count)

4.2 Relationship between number of venues and incidents

Recall in this study, location profile is represented by venues around the location. So let’s look at how venue count correlates with incident count. In Figure 2, as venue count increases so does the incident count but there are many exceptions. It is easy to think both counts increase with the population so a positive correlation is expected. It is clear linear regression does not fit here so polynomial regression will be explored later in this report.

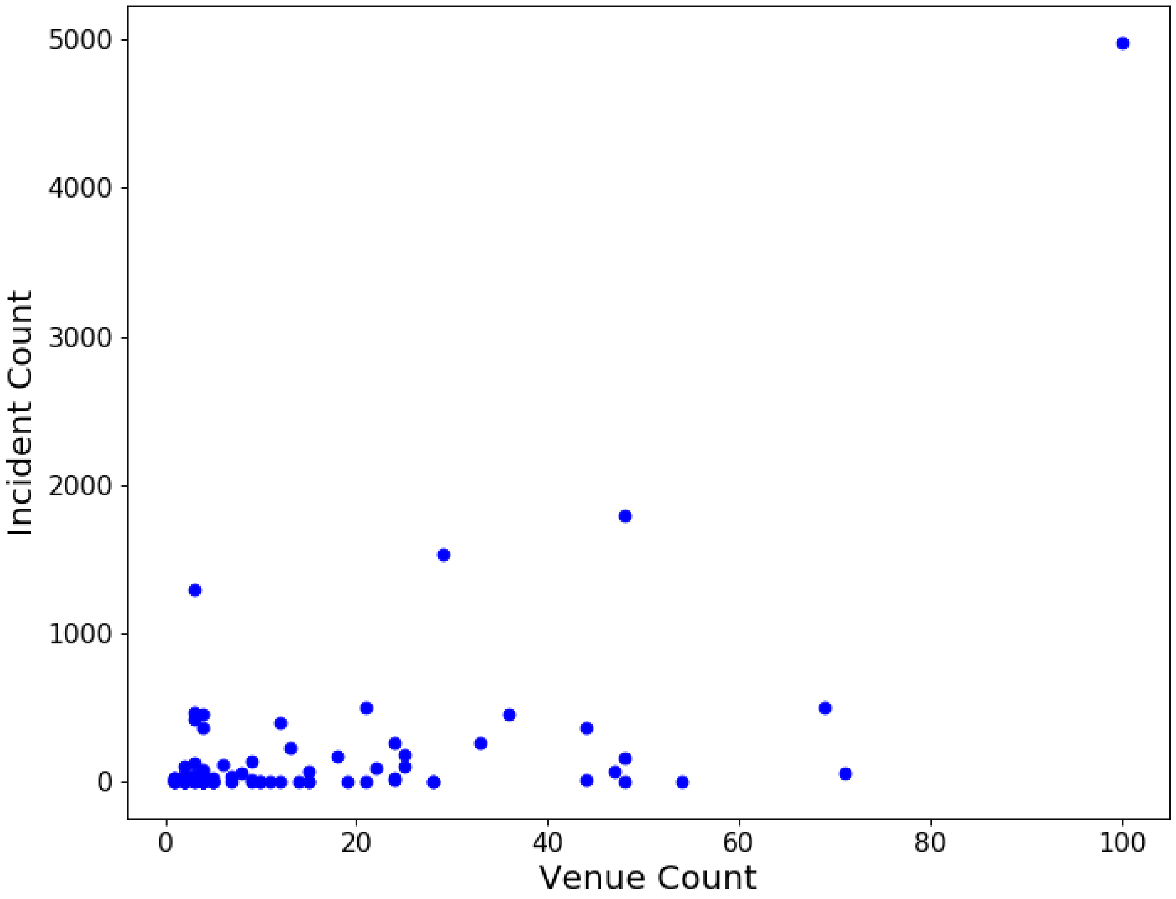
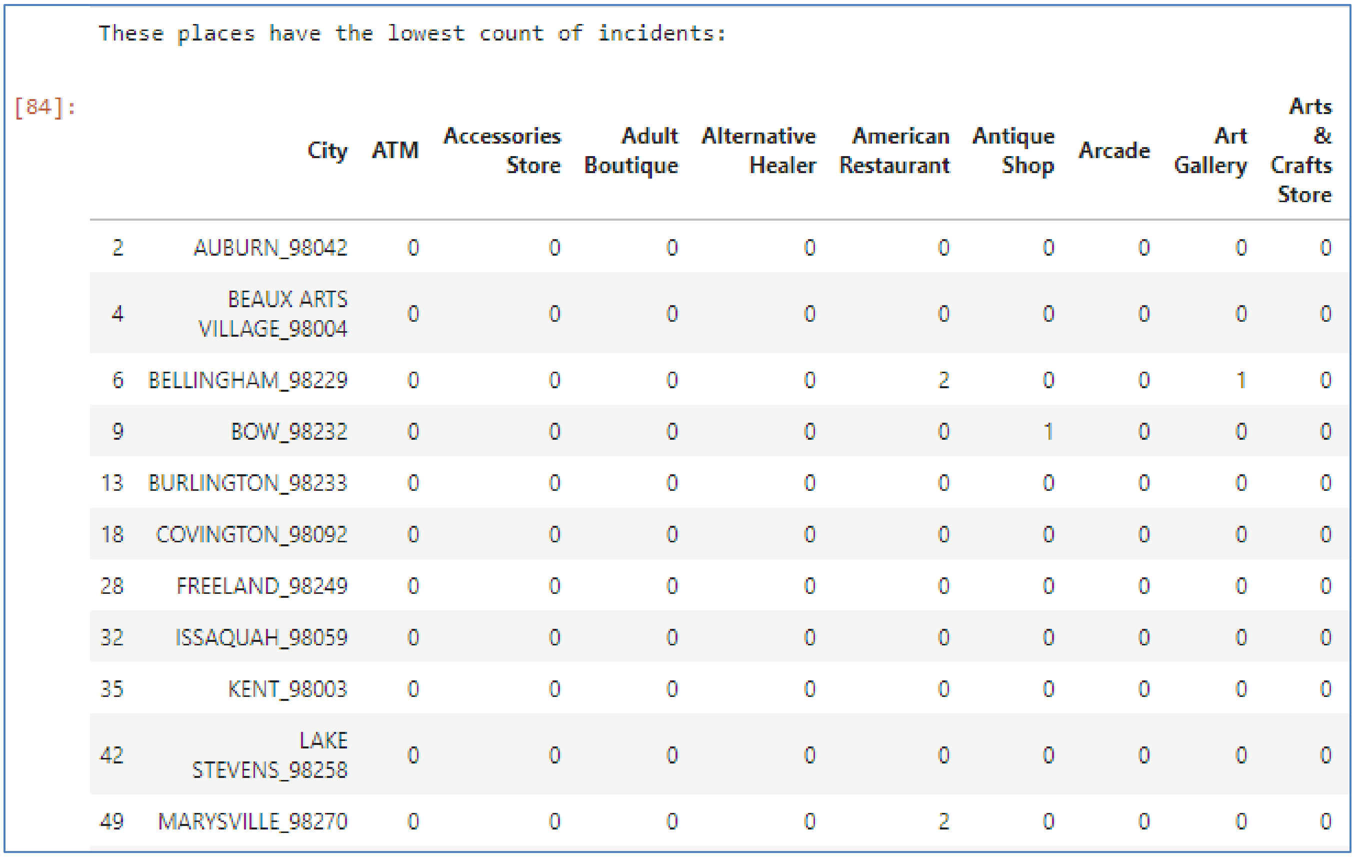


Figure 2 Scatter plot of venue count vs. incident count

4.3 Characteristics of low-incident neighborhoods

Figure 3 shows some of the neighborhoods with the lowest count of incidents. Most of these neighborhoods have very few venues. Those with some venues tend to concentrate in a few categories such as restaurants.



These places have the lowest count of incidents:

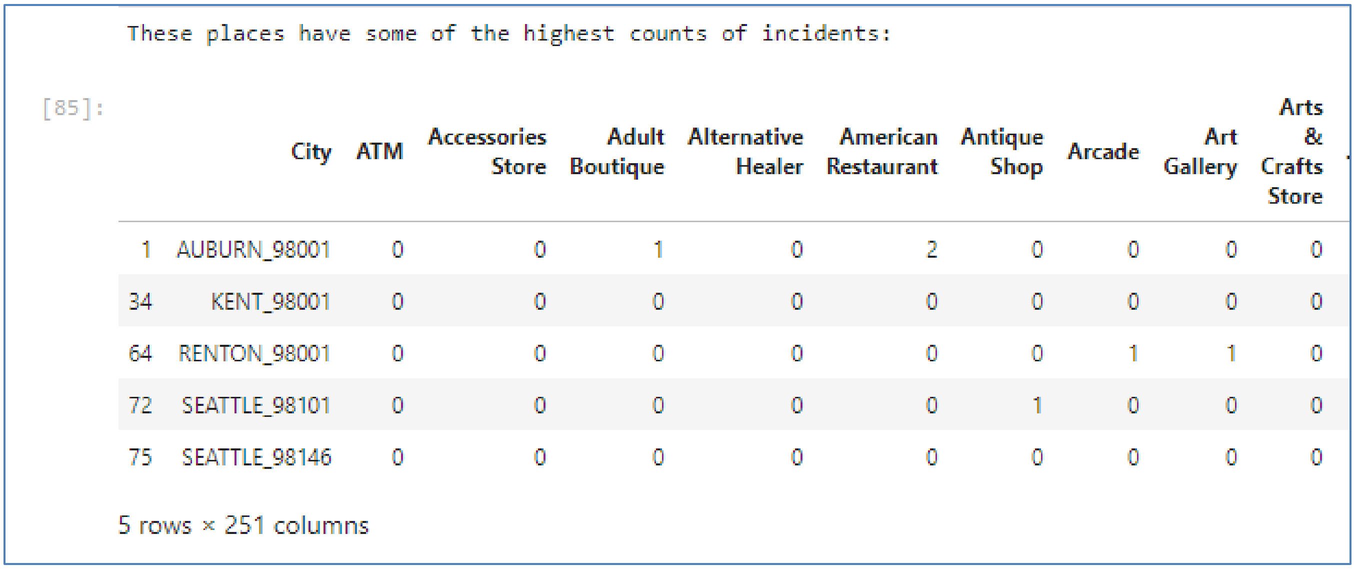
[84]:

	City	ATM	Accessories Store	Adult Boutique	Alternative Healer	American Restaurant	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store
2	AUBURN_98042	0	0	0	0	0	0	0	0	0
4	BEAUX ARTS VILLAGE_98004	0	0	0	0	0	0	0	0	0
6	BELLINGHAM_98229	0	0	0	0	2	0	0	1	0
9	BOW_98232	0	0	0	0	0	1	0	0	0
13	BURLINGTON_98233	0	0	0	0	0	0	0	0	0
18	COVINGTON_98092	0	0	0	0	0	0	0	0	0
28	FREELAND_98249	0	0	0	0	0	0	0	0	0
32	ISSAQUAH_98059	0	0	0	0	0	0	0	0	0
35	KENT_98003	0	0	0	0	0	0	0	0	0
42	LAKE STEVENS_98258	0	0	0	0	0	0	0	0	0
49	MARYSVILLE_98270	0	0	0	0	2	0	0	0	0

Figure 3 Neighborhoods with low count of incidents

4.4 Characteristics of high-incident neighborhoods

Figure 4 shows top five neighborhoods with the highest number of incidents. Not surprisingly, they are large population centers and tend to have a variety of venues.



These places have some of the highest counts of incidents:

[85]:

	City	ATM	Accessories Store	Adult Boutique	Alternative Healer	American Restaurant	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store
1	AUBURN_98001	0	0	1	0	2	0	0	0	0
34	KENT_98001	0	0	0	0	0	0	0	0	0
64	RENTON_98001	0	0	0	0	0	0	1	1	0
72	SEATTLE_98101	0	0	0	0	0	1	0	0	0
75	SEATTLE_98146	0	0	0	0	0	0	0	0	0

5 rows × 251 columns

Figure 4 Neighborhoods with high count of incidents

5. Results from Predictive Modeling

5.1 Polynomial regression on number of venues

The relationship between venue count and incident count is complex. Polynomial regression model is attempted to predict incident count based on the venue count. The result is shown in Figure 5. While the model shows the correct trend, the variance is too large to call the model accurate.

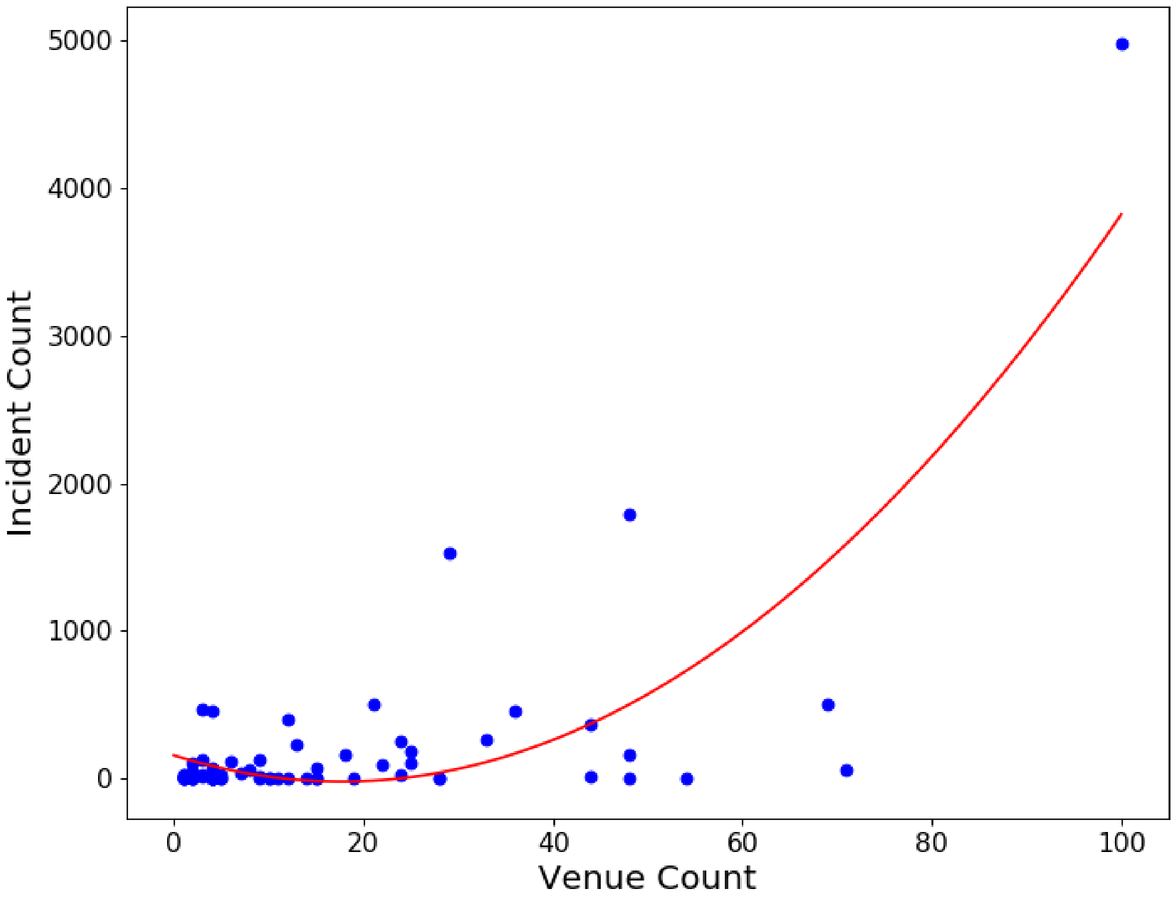


Figure 5 Polynomial regression of incident count vs. venue count

5.2 Multiple linear regression on distribution of venues

Even as the single variable regression correctly predicted the trend, it did so with poor accuracy. In this study, multiple-variable linear regression has also been considered. Here the variables are venue counts in all the venue categories. Unfortunately, the result is not reliable because the number of variables is larger than the sample size.

6. Conclusions

Location profiling, defined as the quantitative spatial analysis of crime data, is introduced to crime analysis. A method to perform location profiling is illustrated in this project. First, incident reports were downloaded, cleaned and grouped by a location identifier initially set to (city, zip code) pair. Second, the location identifier was passed through the Geolocator to obtain geographic coordinates. Third, the geographic coordinates were used to create location profile. Finally, the incident data was analyzed in combination with the location profile. The goal is to use location profile to predict incidents. Data visualization has been used to visualize incidents on map. Characteristics of low-incident neighborhoods were compared with those of high-incident neighborhoods. Predictive modeling has been used to analyze the relationship between incident count and location profile.

7. Discussion and Future Directions

As seen in previous sections, data aggregation to (city, zip code) is not an optimal solution. A better solution is to aggregate to a pre-defined set of neighborhoods. The incident dataset should be processed into a more homogeneous distribution of neighborhoods, instead of variation from a small village to the city of Seattle in this project. It makes the incident more comparable between neighborhoods.

Another weakness in current study is the definition of location profile. As it stands, the list of nearby venues can be arbitrary, independent of the environment they are in.

For future study, one can explore several possible enhancements to the method used in current study:

- Use a pre-defined set of neighborhoods.
- Build a function to look up a physical address within the boundaries of neighborhoods.
- Define a set of standard attributes to characterize a neighborhood.
- Use a more refined Geolocator which can differentiate multiple zip codes in the same city.
- Consider classifying or clustering neighborhoods instead of predicting them.