Mapping Crime Data in Los Angeles to determine good locations for housing.

Crime and Housing

Janich, Karl

Table of Contents

Introduction

A problem often encountered by prospective homeowners is where to purchase a home. Safety is of paramount importance in determining a home location. Los Angeles provides a public database of information, similar to most other major cities. By finding clusters in this data in general or by demonstrating where violent crimes tend to be clustered, one can see where safer neighborhoods may lie to move into.

It would be interesting to overlay other data, such as housing prices, homelessness, or other social welfare indicators, with this data, but that would fall outside the scope of the project.

Data Description

The data used is a database of all crimes in Los Angeles from 2010 to the present, updated weekly. This database is readily downloaded for free at https://data.lacity.org/A-Safe-City/Crime-Data-from-2010-to-2019/63jg-8b9z. It was subject to a Kaggle competition in 2017, owing to its structure amenable to analysis. The columns are noted below (bolded items are to be used in our review):

- DR_NO – Record number
- Date Rptd – Date the crime was reported
- DATE OCC – Date the crime occurred
- TIME OCC – Time the crime occurred
- AREA – Area in which the crime occurred, as defined by the 21 areas serviced by LAPD community police stations
- AREA NAME – The name of the area defined by the community police station
- Rpd Dist No – Four-digit code representing a sub-area within an AREA
- **Part 1-2** – Part 1 or 2 crime as defined by the Uniform Crime Reporting Statistics portion of the US Department of Justice
- Crm Cd – Crime code
- **Crm Cd Desc** – A short description of the crime
- Mocodes – Codes denoting the modus operandi, explaining how the crime was committed
- Vict Age – Victim Age
- Vict Sex – Victim Sex
- Vict Descent – Race/ethnicity of victim
- Premis Cd – Type of location where the crime occurred
- Weapon Used Cd – What type of weapon was used in the crime
- Weapon Desc – More thorough description of the weapon if one was used
- Status – Status of the case in terms of the investigation
- Status Desc – More thorough description of the investigation status
- Crm Cd 1 – Primary crime code
- Crm Cd 2-4 – Secondary crime codes if more than one crime occurred. These are secondary to the primary crime.
- LOCATION – Street address of the crime

- Cross Street – Nearest major intersection of the crime
- **LAT** – Latitude of the crime location
- **LON** – Longitude of the crime location

Methods

For our analysis, we will use the bolded columns above to define locations on a map of Los Angeles to see if there is are spatial clusters that would be beneficial to avoid. Within the areas that have crimes reported, it would be helpful to determine whether there are areas that may be classified based on part 1 or part 2 crimes or the crime descriptions. There are about 2 million crimes listed in the LAPD database, which presents a significant challenge with regard to memory management and algorithm run times. Therefore, we will only look over the past 2 years, which gives a manageable dataset of roughly 200,000.

We performed an initial exploration of the data by using a heatmap over Los Angeles using the Folium library (version 0.10.1) to get an overall view of potential clusters of crime or areas without it. We then applied a variety of clustering methods after data normalization (K-Means, Mini Batch K-Means, DBSCAN using Scikit-Learn 0.15.0) to try to isolate specific areas for further analysis. Finally, a chloropleth map defined over reporting districts was used to identify areas of very high crime concentration and define hard borders of geographic regions.

Results

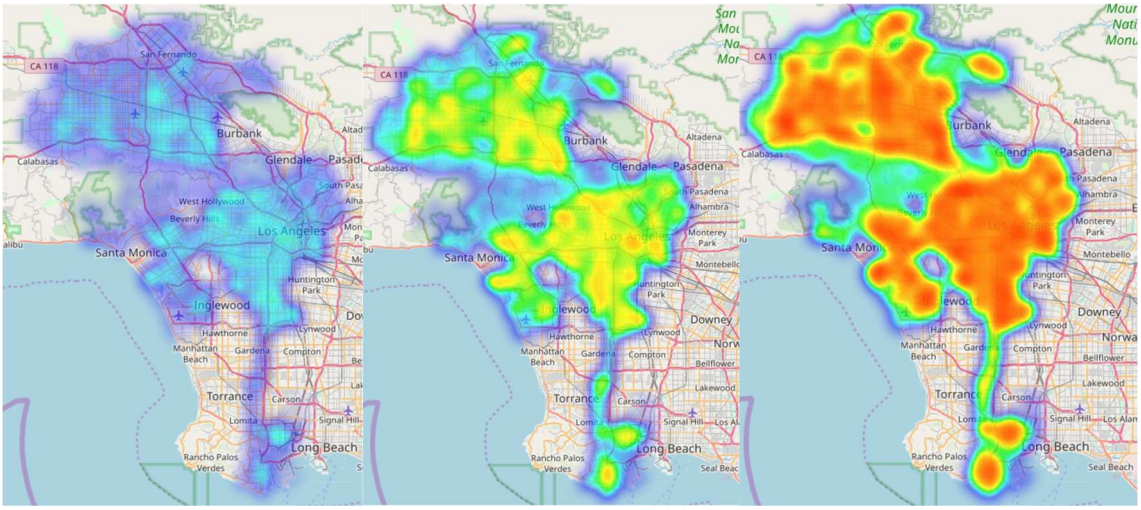

Figure 1 - Heat maps with radius 5 (left), 7.5 (center), and 10 (right).

Initial Heatmap

Our initial heatmap shows some larger defined clusters when appropriately tuned (Figure 1). However, when the radius is adjusted significantly higher or lower, these clusters become less defined. We can see some regions of the city where the heatmap is more intense, but the borders are ill-defined, which encompasses a vast area of the map. Therefore, clustering was employed to see if there were specific locations to include or exclude in terms of crime pockets.

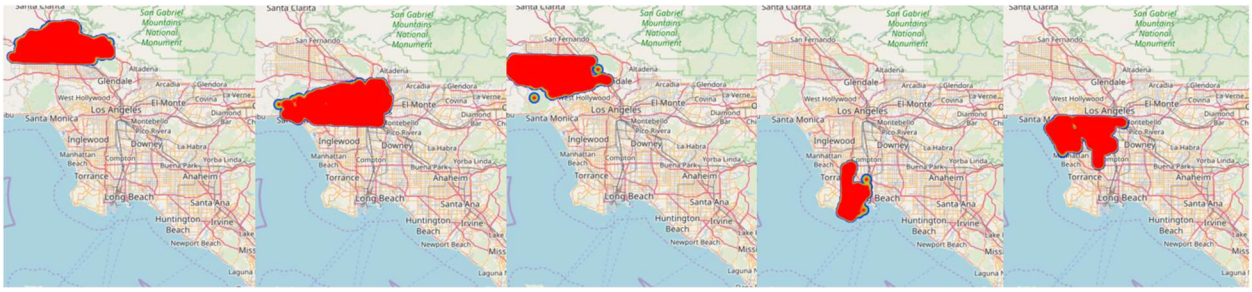

Figure 2 - K-Means clusters with the number of clusters set to 5.

Clustering with various algorithms

To better define spatial clusters of crimes and determine if that would provide useful information concerning where a family may settle to avoid crime, K-Means and DBSCAN algorithms were applied to the data to see if any pockets in the city may be between or outside of clusters. As Figure 2 demonstrates, the K-means clustering algorithm does divide the crime areas effectively, even with only 5 clusters. However, it does not add information to the heat map demonstrated above, where there are only soft boundaries.

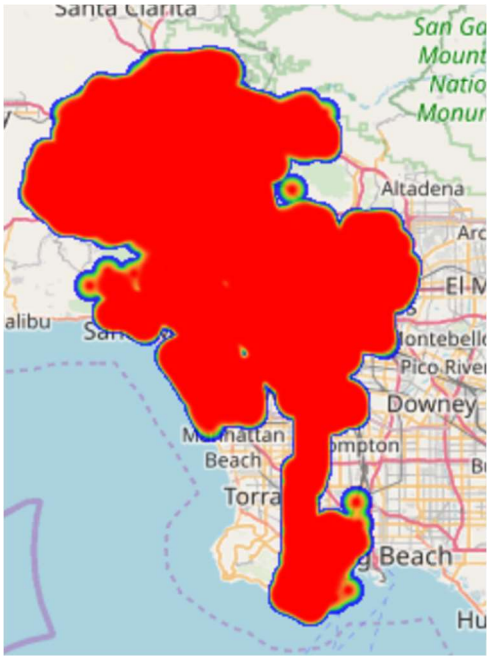

Figure 3 - DBSCAN only yields a single cluster with a variety of EPS trials.

DBSCAN was subsequently applied to the dataset, but only defined one cluster (Figure 3) unless unreasonably small radii were chosen for EPS, which is not unexpected given the diffuse nature of the original heatmap. Outliers are seen, but the same issue as K-Means is apparent, where there is not a well-defined boundary outside of which provides good housing candidates. It should be noted, as well, that DBSCAN would only run with Scikit-Learn 0.15 because versions before and after precompute the distance matrix (with a size in memory proportional to n^2, where n is the number of data points: $n = 235226, n^2 = 55{,}331{,}271{,}076 \cong 51.5\ GB$), causing memory errors.

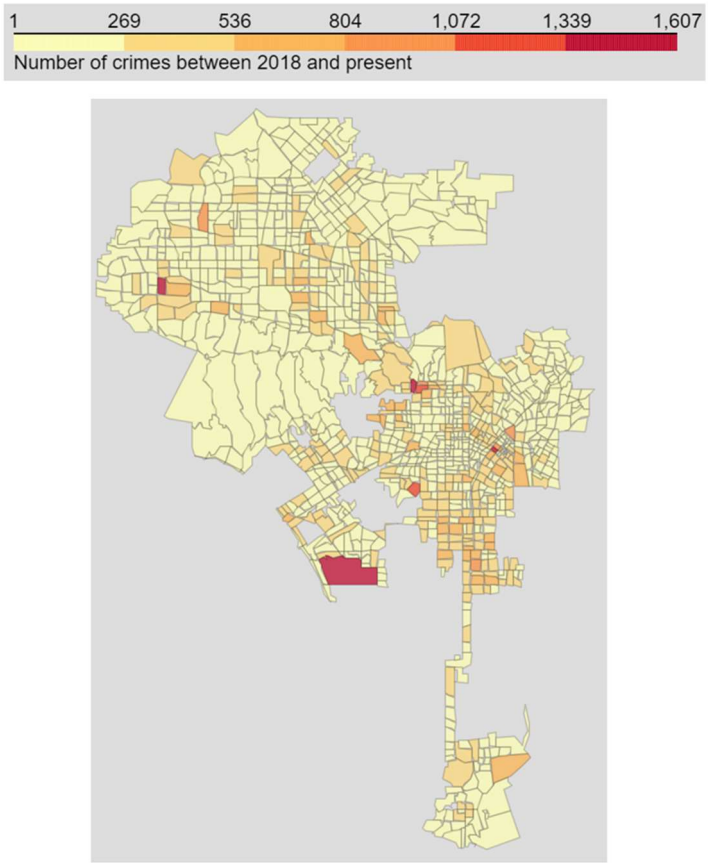

Figure 4 - Chloropleth map by reporting district. Isolated from the base map for better delineation of regions

Chloropleth Mapping

By using reporting districts defined by the LAPD, we generated a map with harder boundaries, as this is what we are seeking in the problem statement. The chloropleth map in Figure 4 shows us there are a few areas of very concentrated crimes but that outside of these areas, crime is less prevalent. These less-prevalent areas may provide good locations for selecting a house for a family with a low probability of experiencing a Part 1 crime.

Discussion

The chloropleth map gives us a very good sense of where crime is highly concentrated. Using DBSCAN, K-Means, and Mini Batch K-Means, we can see various visualizations of clusters over Los Angeles, though they do not do much more than our initial heatmap to inform our neighborhood choices. The chloropleth map provided a very beneficial representation, showing where very high concentrations were in defined geographic zones, giving us hard borders to work with.

As homeowners, these hard borders may be useful to select a home in which to raise a family. However, there are multiple other considerations for purchasing a home, including price, schools, and nearby facilities.

A controversy that has been raised with the use of such crime maps is that other entities, such as banks and insurance companies, have used such maps to deny funding or insurance based on neighborhood, a practice known as "redlining." This practice, unfortunately, has disproportionately affected minorities in the U.S. and brings the above analysis into question from an ethical standpoint when used to assess financial risk. Thankfully, the US Department of Justice and Department of Housing and Urban Development have been penalizing institutions that perform this practice to decrease the threat of institutional racism.

Conclusion

The areas in which crime is concentrated can be quantified with publicly available data. The use of this data is one factor of many that involve moving to a new location, and by using data visualizations, we may quantify where a family may desire to focus a search for a new home.