Sean Devine
February 2020

Capstone Project

Battle of the Neighborhoods

Report

1. Introduction

A key part of any business is expansion. In Los Angeles, California, stakeholders have decided to open a second branch of their company. Business at their current location in South Park is going well so they would like to open their second branch in a neighborhood that is very similar to their current one. To find a neighborhood that is similar to their current one, the Foursquare API will be utilized in combination with k-means clustering. From the similar neighborhoods, the business stakeholders would like to then select the one with the lowest amount of crime.

2. Data

In order to find the neighborhood best suited for the business, the foursquare API will be used to find the most common venues by neighborhood. This data will then be used to cluster the neighborhoods using k-means. On the cluster map, similar neighborhoods can be easily identified. Then, a second layer of the map will be created which identifies crime per square mile by neighborhood. This data is made available by the US government and spans from 2010 until 2020. With both layers of the map, the stakeholders can easily choose a neighborhood that is both similar to their current location and has a low crime rate.

In summary, the data needed is location and venue data from the foursquare API and crime data with coordinates from the US government.

3. Methodology

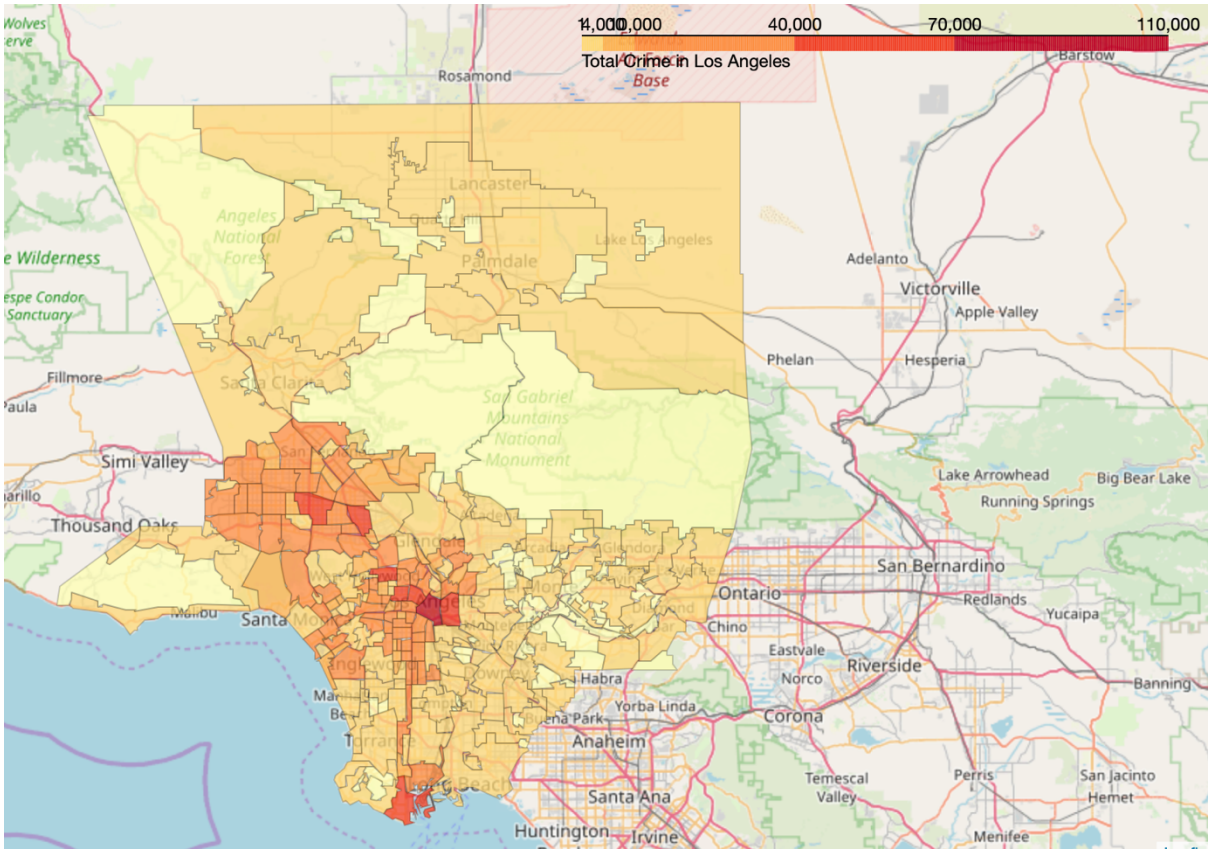
First, data had to be loaded and cleaned which consisted mainly of removing unwanted text and converting strings to floats like the longitude and latitude values. Then, a new dataframe was created that contained the minimum and maximum longitude and latitude value for each neighborhood. The head of the dataframe can be seen below.

	name	max_lat	min_lat	max_lng	min_lng	sqmi
0	Acton	34.542751	34.451958	-118.079704	-118.259918	39.339109
1	Adams-Normandie	34.037411	34.025513	-118.291405	-118.309006	0.805350
2	Agoura Hills	34.168514	34.124958	-118.719414	-118.800362	8.146760
3	Agua Dulce	34.558304	34.451550	-118.254677	-118.379532	31.462632
4	Alhambra	34.111145	34.059933	-118.108192	-118.164833	7.623814

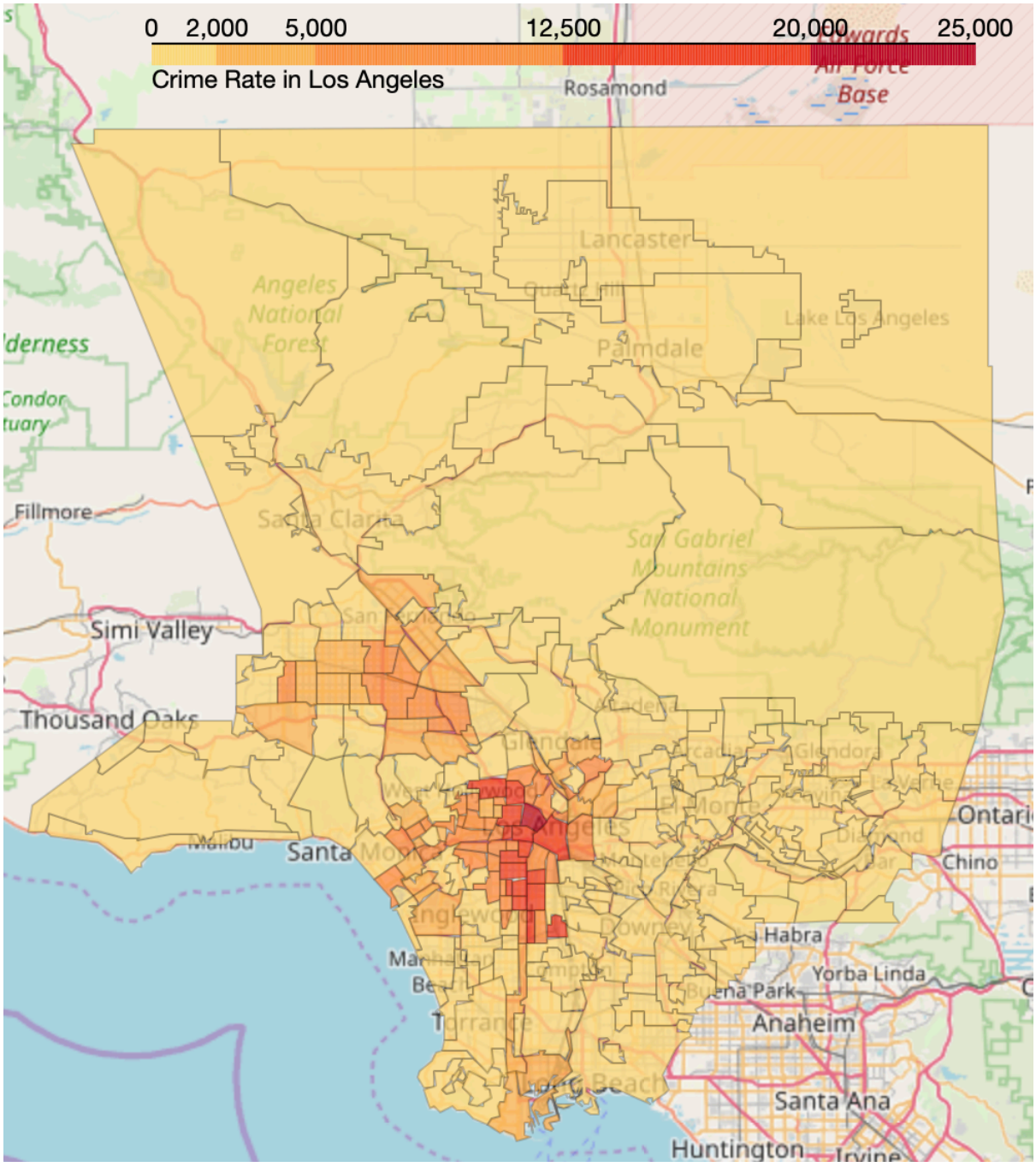
After this, all geometry elements of the geographical data of LA were checked to ensure that they were valid and would not cause any errors. Two neighborhoods had invalid geometry attributes so they were fixed using the ‘buffer()’ method.

Having taken care of any error-causing inconsistencies, both the dataframe shown above and the one containing the geographical data of LA were sorted by neighborhood name. This enabled quicker processing of the function that assigns each crime to a neighborhood. Assigning crimes to neighborhoods was a very lengthy process at first due to the use of large nested loops. However, by slicing the dataframe before looping, this was sped up significantly.

Having assigned crimes to neighborhoods, the following choropleth map of total crime in Los Angeles was created:



To get a more accurate representation of the crime rate, the map was altered to portray crime per square mile.



This improved the overall quality of the map because now areas with more area like Long Beach were not misrepresented in terms of crime.

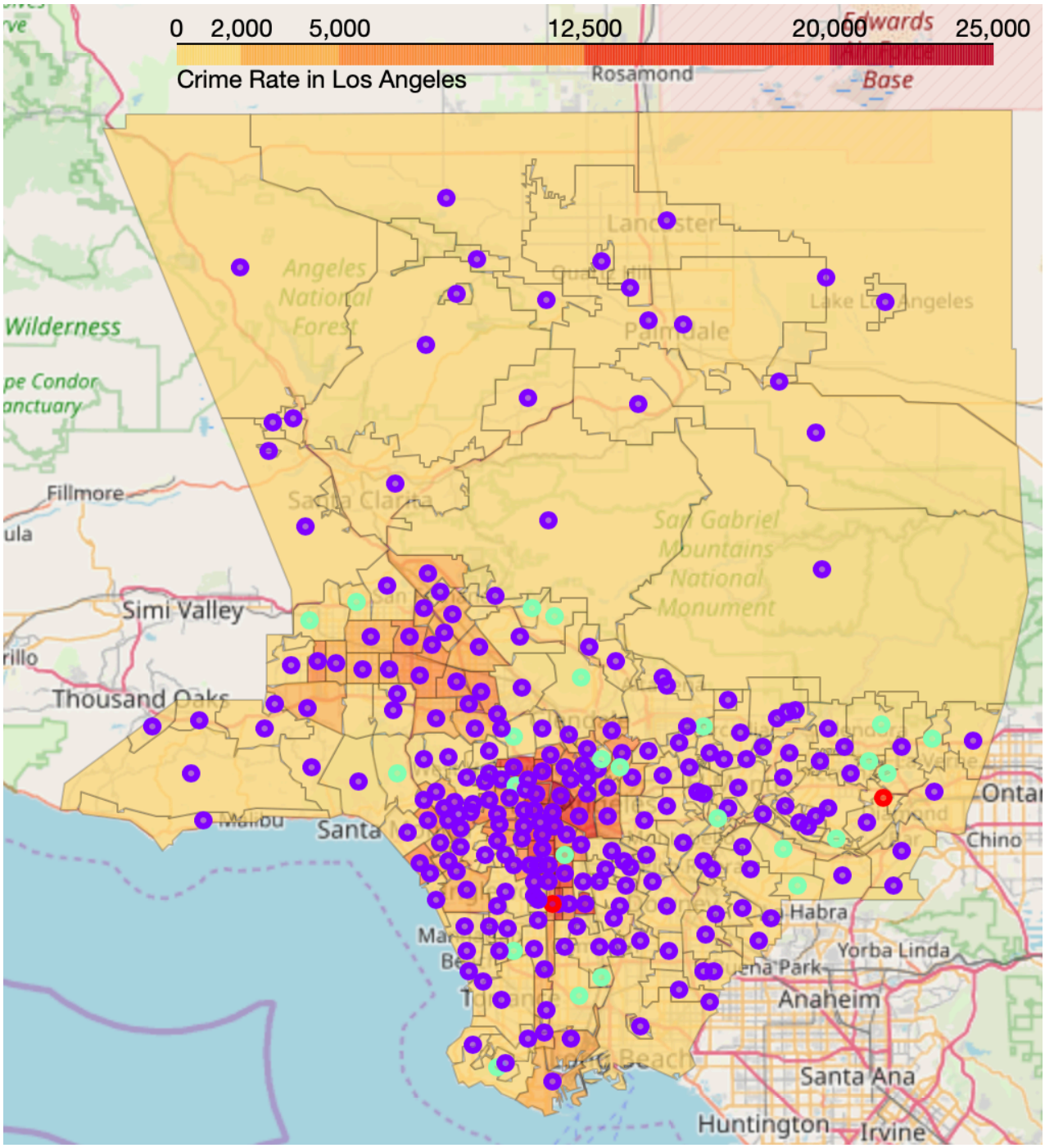
Following the analysis and visualization of the crime rate, it was time to cluster the neighborhoods by venues using the Foursquare API. Having retrieved all nearby venues for each neighborhood, a one hot encoding approach was used to prepare the data for analysis. The table below shows the frequency of venues for each neighborhood.

	Neighborhood	ATM	Accessories Store	Airport	Alternative Healer	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant
0	Acton	0.0	0.0	0.0	0.0	0.040000	0.000000	0.0	0.0
1	Adams-Normandie	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0
2	Agoura Hills	0.0	0.0	0.0	0.0	0.022222	0.022222	0.0	0.0
3	Agua Dulce	0.0	0.0	1.0	0.0	0.000000	0.000000	0.0	0.0
4	Alhambra	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0

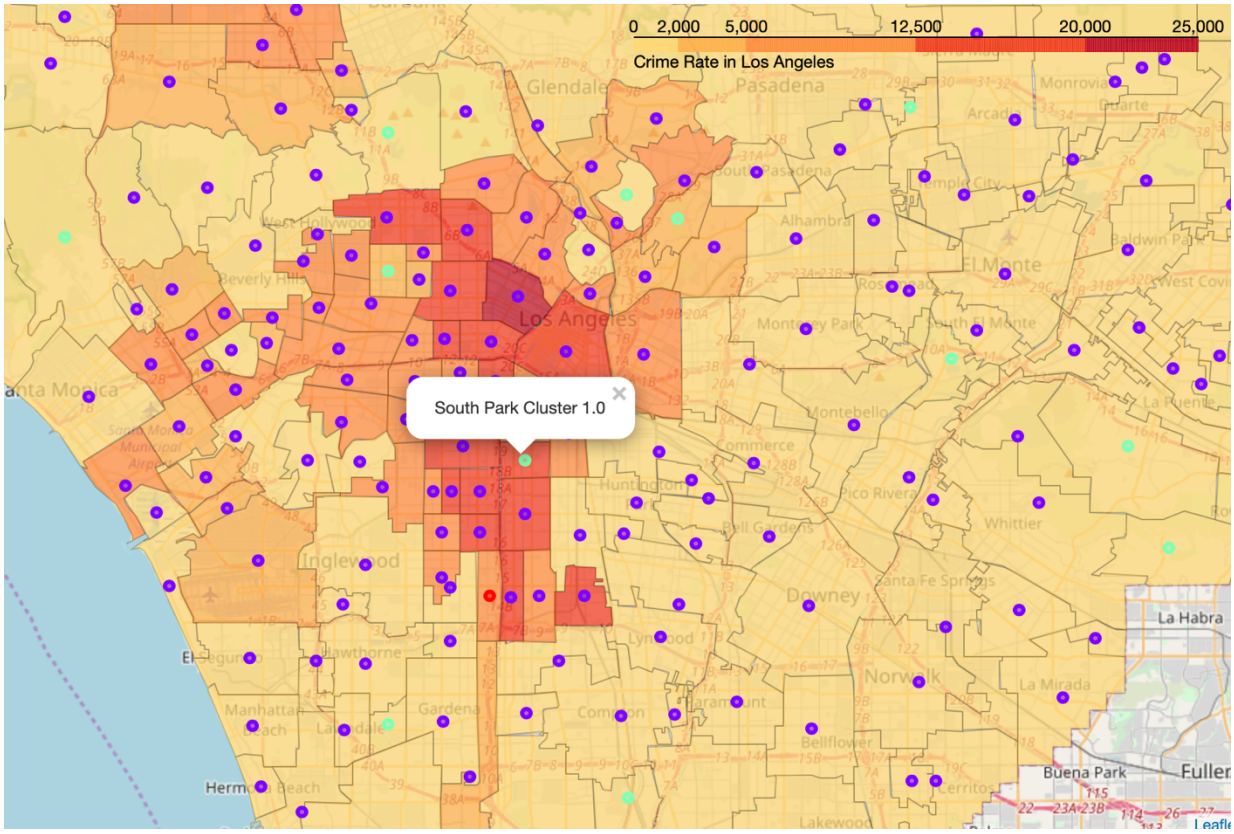
With the data prepared, the silhouette score for each k-means clustering was calculated with k values ranging from 3 to 10. The highest silhouette score was 0.27 with a k value of 3. The distribution of neighborhoods is as follows: 238 neighborhoods in cluster 0, 25 in cluster 1 and only 2 in cluster 2.

4. Results

With cluster and crime per square mile values for each neighborhood, the clusters were added to the crime choropleth map:

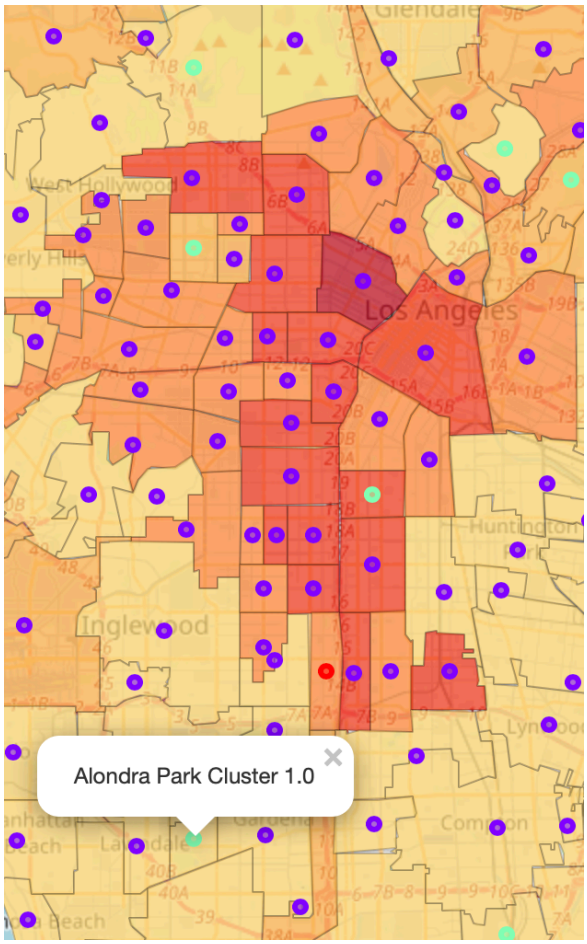


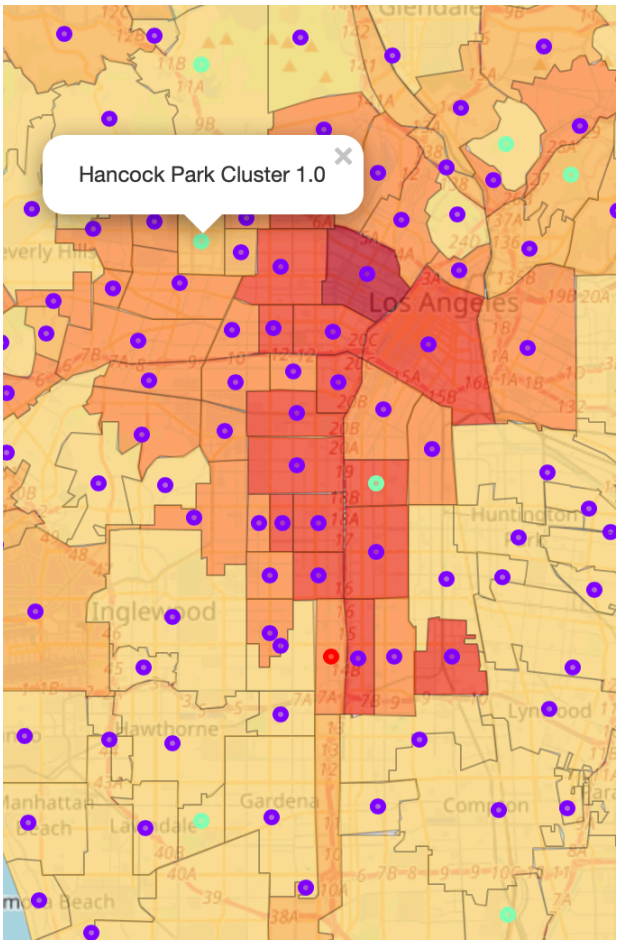
Zooming in on the center of the map around the clients current location, South Park, we can see that the business is currently situated in a relatively high crime area.



5. Discussion

Using the map as guidance, I suggest that the new branch should be opened either in Hancock Park or Alondra Park. Both of these neighborhoods are very similar to the current one and pose a lesser risk of crime. If being as close to the center of LA as possible is desirable, then Hancock Park is the obvious choice. However, it is surrounded by high crime areas so maybe that crime could soon spread to Hancock Park in which case Alondra Park would be the safer choice.





6. Conclusion

In conclusion, there are certainly viable options for expansion. Having done this initial filtering, now I suggest additional analysis pertaining to possible competition, rent prices, available capital etc. If the clients are satisfied with this analysis, this project could certainly be expanded to encompass additional variables and provide an even more complete suggestion. If all variables like rent prices are factored in, this project could ensure that the second branch will be a success.