Capstone Project Report: Clustering San Francisco Neighborhoods by Restaurant Price

Introduction / Business Problem

Suppose you are looking to open a restaurant in San Francisco. San Francisco is huge, so you might be looking to narrow down which general neighborhoods might be suitable. To find a suitable neighborhood, we will mainly look at general restaurant prices in all neighborhoods of San Francisco, with the idea that if you want to establish a fine dining restaurant, you may not want to do so in a neighborhood where restaurants tend to be very cheap, or vice versa.

As an extension, suppose you own a restaurant chain with existing locations in certain neighborhoods. What neighborhoods might be best to expand to?

We will help to provide an answer to this business case by clustering neighborhoods in San Francisco by restaurant pricing. By doing so, we can see what cluster your new restaurant would belong to and thus find suitable neighborhoods for it.

Data

First, we will need to identify all coordinates of neighborhoods in San Francisco. Using those coordinates, we will use FourSquare's Places API, namely the Venue group and Explore endpoint to find all restaurants/food locations within a certain radius of each neighborhood that are classified in each of FourSquare's four price buckets: $ to $$$$. We will then process this data, analyze it, looking at the characteristics of restaurants within each neighborhood, and finally provide a k-means clustering of the neighborhoods by restaurant price.

Methodology

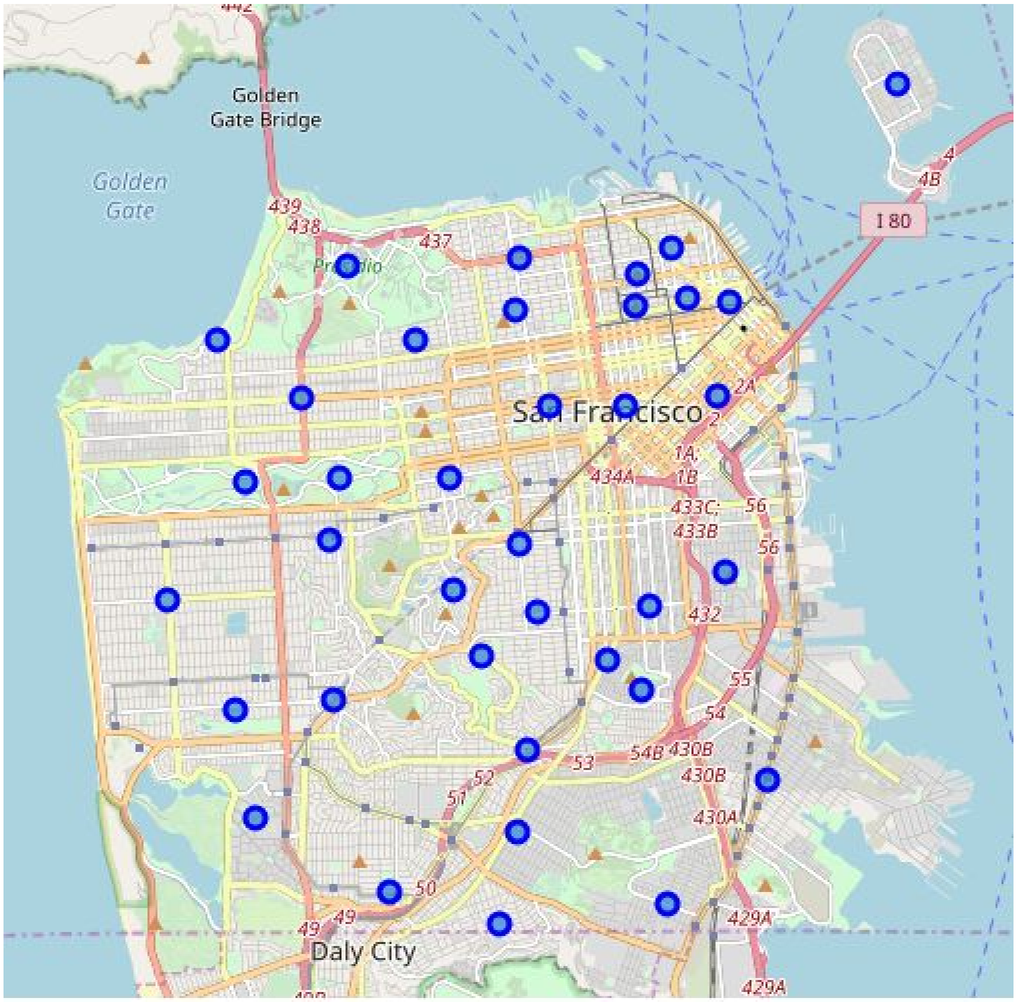
The first part of my data exploration and analysis was finding all neighborhoods of San Francisco and their coordinates. I tried to use Nominatim as much as possible to automate the process, but I had to map out the coordinates and see which neighborhoods were not mapped particularly well by Nominatim and manually correct them.

The second part was to use FourSquare's API to get all restaurants within certain price buckets within a 500 meter radius of the neighborhood coordinates.

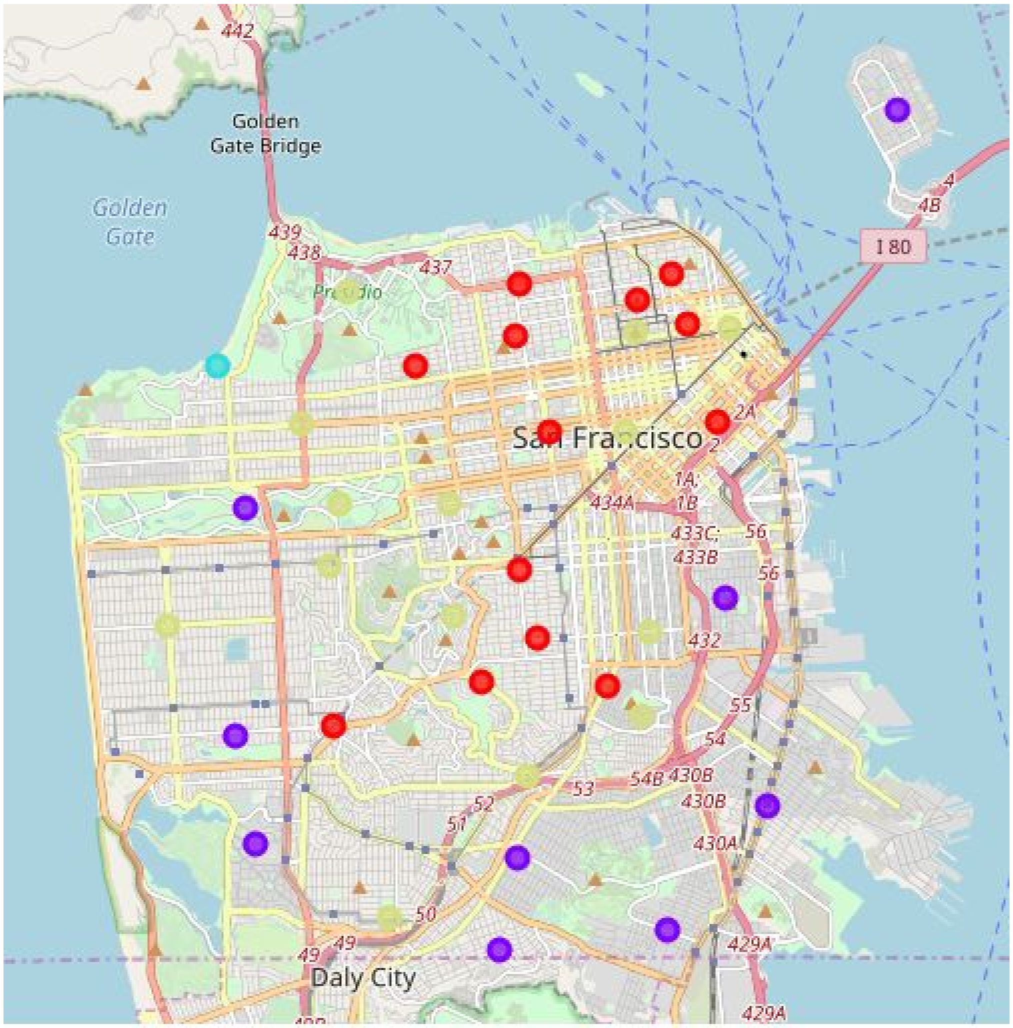
The third part was processing the FourSquare data and clustering the neighborhoods using the frequency of restaurants in each price bucket as features. Finally, I used Folium to display the clustering.

Results

The following map was the result of the first step of analysis: identifying all of San Francisco’s neighborhoods and appropriate coordinates.



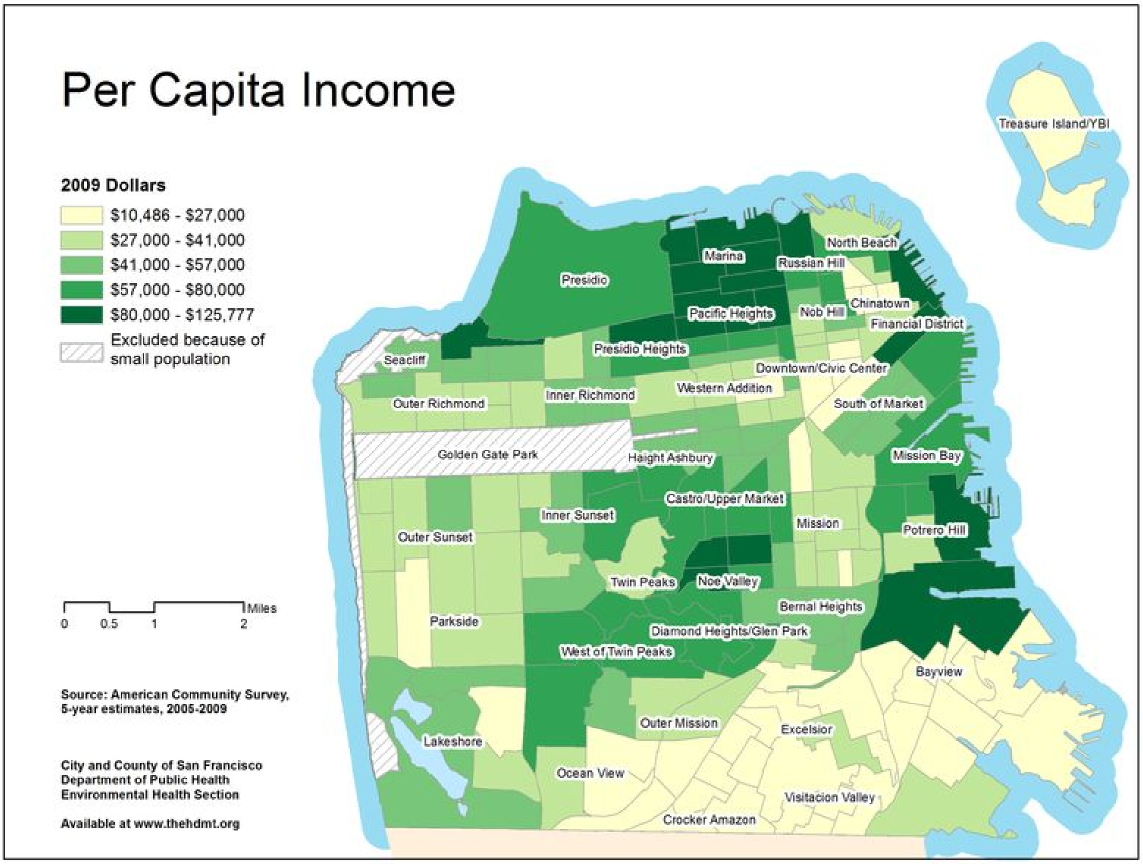
The following was my final result after clustering neighborhoods by nearby restaurant price:



I chose a total of 4 clusters. One cluster, at the top left in light blue, stood by itself as it had no nearby restaurants. The green, purple, and red clusters are somewhat geographically clustered even though I did not factor geographical coordinates into the k-means clustering.

Discussion

An interesting comparison might be to a Per Capita Income in San Francisco.



Given this map, an interpretation could be made that the Red cluster might be good for a more expensive restaurant, the Green cluster might be good for perhaps $$ -- $$$ restaurants, the Purple cluster is not especially good for restaurants as it is on the outskirts of San Francisco, where people will be more spread out and further from Downtown, and the Blue cluster is inconclusive, but perhaps a good place for a restaurant given there is not much competition around the area.

The Red and Green clusters are relatively similar, but an interesting observation is the triangle of green cluster neighborhoods in the northeast of San Francisco: at Chinatown, Nob Hill, and Downtown/Civic Center. Looking at the Per Capita Income map, these tend to be lower-income areas (Nob Hill slightly less so).

If the restaurateur has a restaurant already in a red cluster area, say in Marina, I would suggest another red cluster area like Noe Valley to put a similar restaurant.

Of course, there are many more factors to a restaurant's success that are not included in this study, but this clustering map may be helpful to narrow down areas of interest.

Conclusion

In this report, we gathered locational data for all of San Francisco’s official neighborhoods, used FourSquare to find the price characteristics of restaurants near each neighborhood, clustered them using k-means on their price points, and provided a map that may help suggest to a restaurateur where might be a good location for a restaurant. Of course, there are parallels between this study and a per-capita income study, which may be interesting to research more in depth. For example, even if a certain area is very wealthy, perhaps it is mostly residential and the residents head downtown for fancy restaurants. This type of mismatch between income in the area and restaurant pricing in the area might signal an interesting business opportunity.