Finding the Perfect Location

Norman Lee
November 18, 2019

Introduction

For many budding restaurateurs, one of the most exciting and frightening moments is expanding to new locations. On the surface it might seem like a very natural and organic progression of replicating the initially successful recipe. However, as any established restaurateurs will tell you, there are many more factors. Of all these variables, location often surfaces to the top. People often say, it's all about location location location!
So how does one go about finding that perfect location that would replicate the success of the original restaurant?
Instead of basing on word of mouth, feels, emotion and other eyeballing approaches, the goal is to bring data science to a restaurateur who is looking to expand into other cities or neighborhoods. By using data science the proprietor will be presented with an analysis of the different neighborhoods in a defined area. The analysis will give a non-emotional, objective picture of the area and pave the way toward choosing the right location. Furthermore, the analysis can validate or solidify a potential choice made from emotional or non-scientific methodologies. Lastly, the analysis gives the restaurateur the ability to quantify why a "Unicorn" location is perfect.

The study will identify the different neighborhoods. For each neighborhood, foursquare location data will be utilized to define the characteristics, businesses, interests and etc. Furthermore, the neighborhoods will be grouped by these defining characteristics. Lastly, metrics for each of the cluster will be calculated. Hopefully these clusters will channel a restaurateur to quickly identify one or some neighborhoods that share the very same characteristics that make for a successful restaurant location.

Data

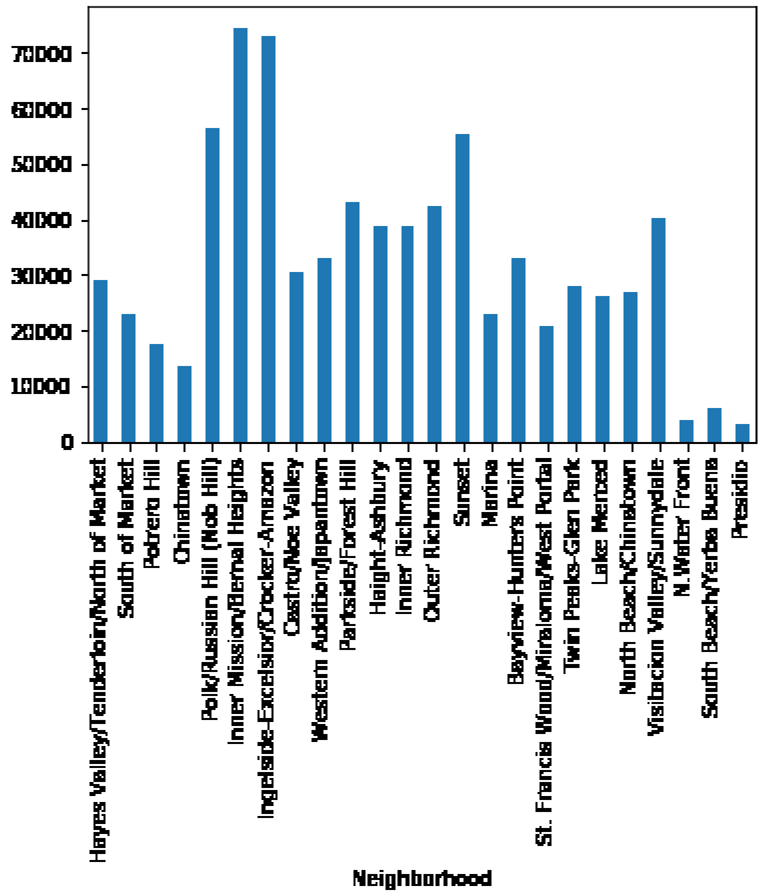
For this example, the city of interest was San Francisco, California. The first step was to scrape zip codes and neighborhood names from http://www.healthysf.org/bdi/outcomes/zipmap.htm. Next, data from http://zipatlas.com/us/ca/san-francisco/zip-code-comparison/population-density.htm was merged to get the 3 missing zip codes. Lastly, the missing neighborhood names were obtained from https://deansereni.com/newsletter/san-francisco-demographics-by-zip-code. Treasure Island and 94104 were discarded due to low population and/or lack of business establishments. At this point the dataset contained zip codes, names and population. The geographic coordinates were obtained using the uszipcode 0.2.4 in Python. A new table with zip code, name and geographic coordinates was created. From here an exploration of top 100 venues within 500 meters in each neighborhood was called from Foursquare by utilizing the geographic coordinates. Using the GET request many response fields (i.e. id, name, contact, location, categories, attributes) were obtained from Foursquare for each venue.

Methodology

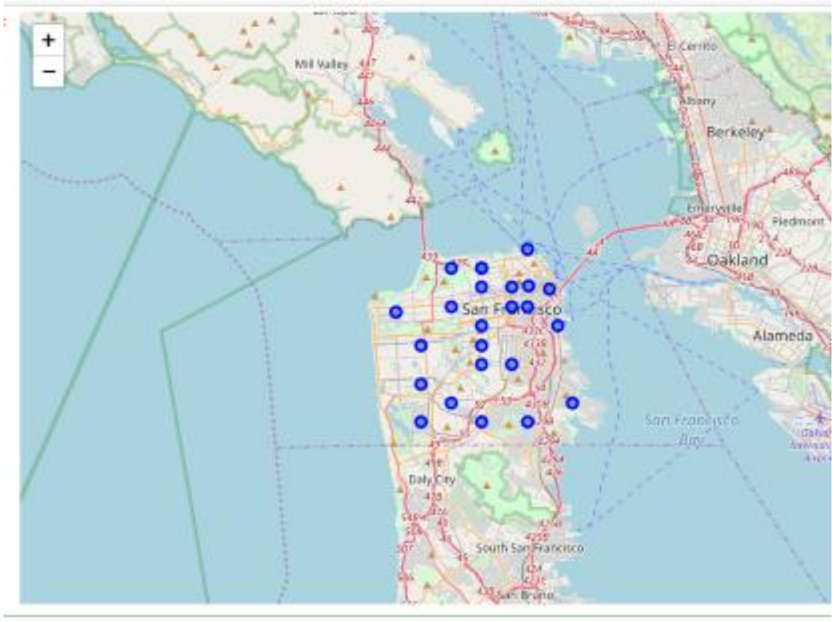
Since the objective was to find the perfect location, it would be wise to make sure the neighborhoods identified all fall within the geographical area of interest. A visual exploratory data analysis was completed by using geopy library and Folium to create a map of San Francisco with neighborhoods superimposed on top. This allowed for quick visual scan for outliers. The data compiled from Foursquare for each neighborhood was organized by categories and the mean of the frequency of occurrence of each category was reported. For each neighborhood the top 5 most common venues were determined. To help define similar neighborhoods based on the venues, k-means was ran to cluster the neighborhoods into 8 groups. Finally each cluster was examined to determine the distinguishing characteristics. This will paint a picture of the type of venues within a neighborhood. For example, the most common venue for each neighborhood can be shown. In addition, the number of restaurants or other types of establishment within certain neighborhood can be determined. All these data points will help the restaurateur decide on the best location.

Results and Discussion

From the initial data collection, it’s apparent that San Francisco is not as large and diverse as cities such as New York or Toronto. Only 24 neighborhoods were identified and even then, three had extremely small population.



After settling on the number of neighborhoods and getting the zip codes, one data point was way off when the visual exploratory data analysis was conducted.

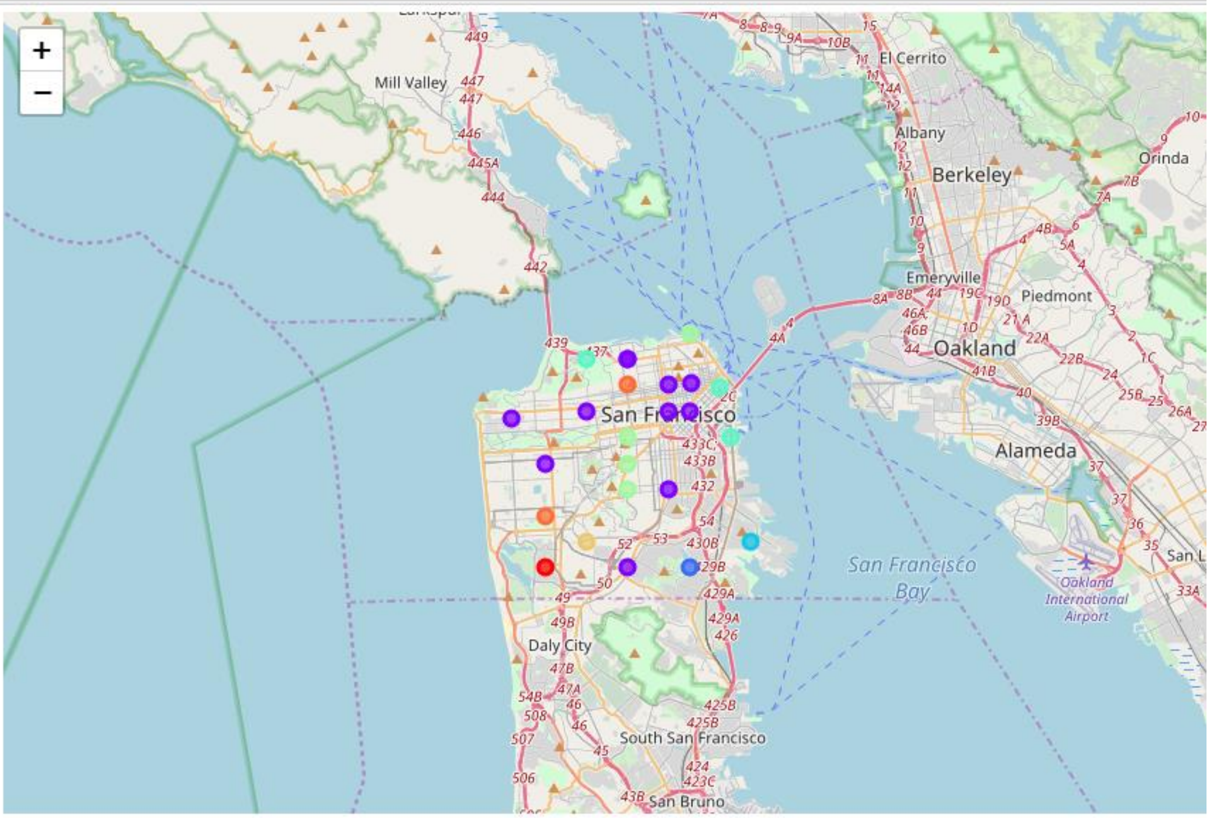


The error was corrected and the venues data from Foursquare was pulled.
A total of 240 unique categories were extracted. However, upon closer examination several neighborhoods had extremely low venue counts. These neighborhoods were deemed to be bad potentials for new location due to lack of other businesses. These areas could potentially be strict residential zones or parks. To get a better understanding, additional analysis was done. Here’s an example. Visitacion Valley/Sunnydale only had 5 venues. Upon further review, it was found that the only interests were baseball field, garden, park, scenic lookout, and a restaurant. Either due to regulations or high barrier to entry/low returns, no other businesses were established. This neighborhood could potentially be a dark horse if there were no other food venue. However, someone had entered the market and it could be saturated due to environment.

	Monument / Landmark	37.76	-122.44	Pink Triangle Park & Memorial	37.762383	-122.436252
	Nail Salon	37.76	-122.44	Hand Job Nails & Spa	37.759934	-122.434908
	Optical Shop	37.76	-122.44	Eye Gotcha Optometric	37.759651	-122.434967
...	...	...	...	...	...	...
Visitacion Valley/Sunnydale	Baseball Field	37.72	-122.41	Louis Sutter Playground	37.722388	-122.413928
	Garden	37.72	-122.41	Visitacion Valley Greenway	37.717687	-122.407316
	Park	37.72	-122.41	John McLaren Park Lookout Point	37.717758	-122.407291
	Scenic Lookout	37.72	-122.41	Wilde Overlook	37.718066	-122.412379
Western Addition/Japantown	American Restaurant	37.79	-122.44	I Forgot It's Wednesday	37.787963	-122.441275
	Arts & Crafts Store	37.79	-122.44	Atelier Yarns	37.787802	-122.440452
	Bakery	37.79	-122.44	B. Patisserie	37.787945	-122.440804
	Bubble Tea Shop	37.79	-122.44	Tea Hut	37.788031	-122.440775
	Burrito Place	37.79	-122.44	El Burrito Express	37.786365	-122.440108
	Chinese Restaurant	37.79	-122.44	Eliza's Restaurant	37.787790	-122.441574

Lastly, the neighborhoods were clustered via K-means. Given the relatively small size, only 8 clusters were designated.

7]:




```
#Examining the Clusters
#Cluster 1
```

```
sf_merged.loc[sf_merged['Cluster Labels'] == 0, sf_merged.columns[[1] + list(range(5, sf_merged.shape[1]))]]
```

For each cluster the top 10 most common venues were shown. Therefore, a restaurateur can quickly get a feel for the businesses within that cluster. Furthermore, one could quickly gauge the number of food related venues and the popularity. From the study, one could quickly see that cluster two contained the most vibrant, commercial neighborhoods. It also pointed to the abundance of food related businesses and their ranking within the neighborhood. As such, this data would point the proprietor to enter the cluster given the high traffic volume. However, it also gave a quick indication as to what genre of restaurant had reached saturation.

Conclusion

The study allows a restaurateur to quantify the desirability of a location. Furthermore, by clustering, one can quickly find related neighborhoods that share similar characteristics as the original location. The study is not meant to be the end of the process but the beginning. It helps to channel and narrow the scope for the proprietor in hope of going down the right direction. It allows the owner to establish the characteristics of the current location as the baseline and quickly find similar potential locations. Lastly, the study can be conducted in different cities and the data easily compared. The ability to describe what “perfect” location means will be an awesome tool for a successful restaurateur who’s about to build an empire.