Coursera Capstone Project

Neighborhood Analysis by Venue Ratings

Dwayne Thaele

January 20, 2021

Table of Contents

GitHub Repository

The Github repository for this project can be found at:

https://github.com/dthaele-coursera/Final_Project.git

Introduction

Background and Problem Description

Applications like Yelp and Google Maps are used to identify venues in a city and assess their suitability by ratings submitted by past customers. These applications have proven themselves to be valuable. However, for application users not familiar with a city, they may not obtain full context of the venues and their locations. For example, someone may wish to visit a nice restaurant in a nice area of a city, or avoid bad restaurants in bad parts of a city.

Data Description

Foursquare is a US business that uses detailed location data, along with business, and customer input to "tap into this intelligence to create better customer experiences and smarter business outcomes." See: https://foursquare.com.

Foursquare provides a user-based rating system where Foursquare used feedback from customers, combined with a wealth of other data and artificial intelligence techniques to calculate overall ratings for a venue. Rating are numeric from 0 to 10, 10 being the highest rating; a 0 rating generally used for venues that have not received a rating. For a detailed discussion about Foursquare's rating system, see: https://medium.com/foursquare-direct/finding-the-perfect-10-how-we-developed-the-foursquare-venue-rating-system-c76b08f7b9b3

For this project I'll be using Foursquare's venue ratings and location data that is provided by their API. For a reference about the API, see: https://developer.foursquare.com/docs/apireference/venues/search/

At present, my project will only focus on venues that are restaurants. Depending on complexity I may add additional venues.

Additionally, my project will only focus on the city of San Francisco. However depending on complexity, I may include the cities of Chicago and New York.

While beyond the scope of this project, additional supplemental data of crime incidents per neighborhood could be added. Safety conscious users may wish to use crime data in determining the neighborhood they wish to patron.

Problem Solution

The foundation of my project, *Neighborhood Analysis by Venue Ratings*, is an assumption that venues with high ratings are generally in more desirable areas, where venues with lower ratings are generally in lower desirable areas. This serves the primary use case where a visitor, unfamiliar with an area or local venues would like to visit the are and decide on a venue after they arrive and *check-out* the neighborhood. The visitor may even decide to spend more time in the neighborhood if its appealing to them.

I will group and color code venues based on their ratings:

Venue Rating Range	Color Designation
9 - 10	Green
7 - 8	Blue
5 - 6	Yellow
3 - 4	Orange
1 - 2	Red
0	White

I will cluster the results and depict them in an overlay of a city map.

The outcome will be a map that reveals the more and least desirable areas based on rating. This provides users with a simple way to determine which area are likely to have the most positive (dining) experience.

This study only focuses on restaurant venues. However, the approach can be easily applied to and venue type.

Methodologies

There are two methodologies I used in Neighborhood Analysis by Venue Ratings:

Methodology 1: Clustering by Neighborhood

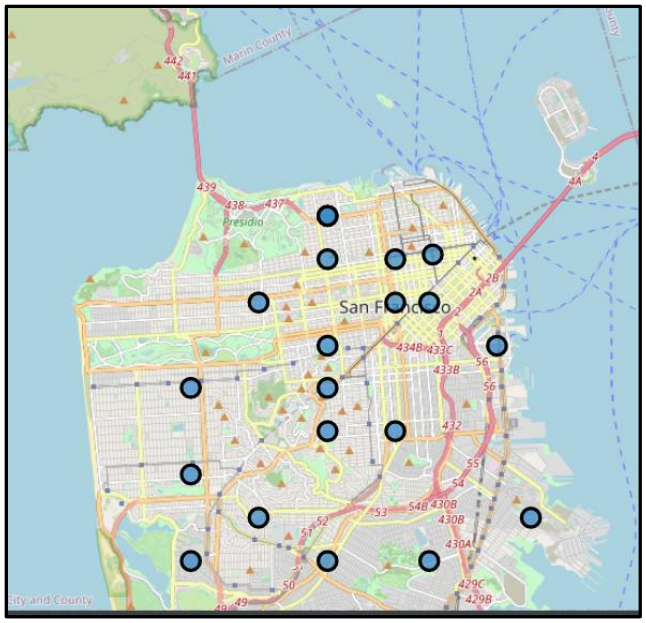
With the first methodology I divided the city of San Francisco into neighborhood using zip codes from the webpage http://www.healthysf.org/bdi/outcomes/zipmap.htm:

Zip Code	Neighborhood
94102	Hayes Valley/Tenderloin/North of Market
94103	South of Market
94107	Potrero Hill
94108	Chinatown
94109	Polk/Russian Hill (Nob Hill)
94110	Inner Mission/Bernal Heights
94112	Ingelside-Excelsior/Crocker-Amazon
94114	Castro/Noe Valley
94115	Western Addition/Japantown
94116	Parkside/Forest Hill
94117	Haight-Ashbury
94118	Inner Richmond
94121	Outer Richmond
94122	Sunset
94123	Marina
94124	Bayview-Hunters Point
94127	St. Francis Wood/Miraloma/West Portal
94131	Twin Peaks-Glen Park
94132	Lake Merced
94133	North Beach/Chinatown
94134	Visitacion Valley/Sunnydale
All Zips	(all of San Francisco, including very small population zips, such as Treasure Island or the Presidio, which are not listed above)

I then used

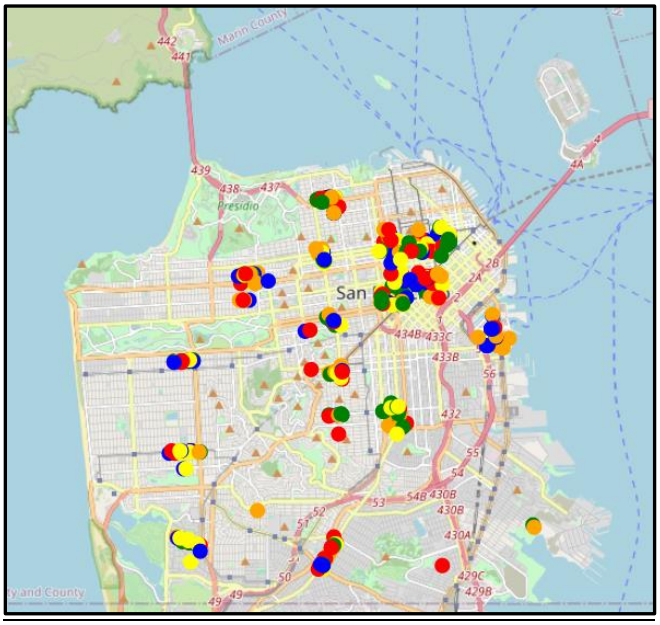
then San Francisco's latitude and longitude coordinates to generate a folium map of San Francisco. I applied markers to the map to identify central locations of each neighborhood.

Sample Plot of San Francisco Neighborhood Markers



I then calculate a color code for the venues based on their Foursquare Rating and plot the venues onto the map of San Francisco:

Venue Color	Venue Rating
Green	9 – 10
Blue	7 – 8
Yellow	5 – 6
Orange	3 – 4
Red	1 - 2



Inspecting the dataframe “neighborhood_master_sorted” we can see the neighborhoods with their color codes identifying the highest to lowest average venue ratings along with other important data.

	Zip Code	Neighborhood	Latitude	Longitude	Venue Count	Avg Rating	Rating Sum
16	94124	Bayview-Hunters Point	37.730	-122.380	2	6	12
4	94108	Chinatown	37.791	-122.409	30	6	180
11	94117	Haight-Ashbury	37.770	-122.440	12	6	82
18	94131	Twin Peaks-Glen Park	37.750	-122.440	7	6	44
14	94122	Sunset	37.760	-122.480	30	6	196
5	94109	Polk/Russian Hill (Nob Hill)	37.790	-122.420	30	6	181
15	94123	Marina	37.800	-122.440	30	6	180
8	94114	Castro/Noe Valley	37.760	-122.440	29	5	157
3	94107	Potrero Hill	37.770	-122.390	30	5	160
10	94116	Parkside/Forest Hill	37.740	-122.480	30	5	157
9	94115	Western Addition/Japantown	37.790	-122.440	16	5	91
19	94132	Lake Merced	37.720	-122.480	17	5	94
12	94118	Inner Richmond	37.780	-122.460	30	5	163
6	94110	Inner Mission/Bernal Heights	37.750	-122.420	30	5	178
1	94102	Hayes Valley/Tenderloin/North of Market	37.780	-122.420	30	5	169
2	94103	South of Market	37.780	-122.410	30	4	138
17	94127	St. Francis Wood/Miraloma/West Portal	37.730	-122.460	1	4	4
7	94112	Ingelside-Excelsior/Crocker-Amazon	37.720	-122.440	30	4	139
20	94133	North Beach/Chinatown	37.800	-122.440	30	4	149
13	94121	Outer Richmond	37.800	-122.700	1	1	1
21	94134	Visitacion Valley/Sunnydale	37.720	-122.410	1	1	2

Methodology 2: K-means Clustering

k-means clustering is a method of vector quantization, that aims to partition observations into (k) clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid). For the K-means analysis, the dataframe “new-pdf” with the columns ['Venue Latitude', 'Venue Longitude', 'Venue Rating'] is used to create the clusters. The column[“Cluster Label] is used to identify the cluster that a venue belongs to.

From several trial analysis, it was determined the optimal number of clusters to define is 5.

Detailed information on all clusters, venues, and neighborhoods

All Cluster and Venue Data

	Cluster Label	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Rating	Venue Color
50	0	South of Market	37.780	-122.410	Tín	37.780840	-122.405770	Vietnamese Restaurant	1	Red
307	0	Inner Richmond	37.780	-122.460	Uncle Boy's	37.777327	-122.461631	Burger Joint	1	Red
42	0	South of Market	37.780	-122.410	Arsicault Bakery	37.780789	-122.413433	Bakery	1	Red
297	0	Inner Richmond	37.780	-122.460	Wako	37.783032	-122.461576	Japanese Restaurant	1	Red
283	0	Parkside/Forest Hill	37.740	-122.480	Szechuan Taste Restaurant	37.742889	-122.476096	Chinese Restaurant	1	Red
53	0	South of Market	37.780	-122.410	The Cavalier	37.783252	-122.406884	English Restaurant	1	Red

Results

Results are determined by metrics collected for the approach used:

Approach	Metrics	Results
Neighborhood Clustering	• Most Venues in Neighborhood	• Neighborhood with highest venue count is preferred
K-means Clustering	• Highest Avg Venue Rating per cluster • Most Venues in Cluster	• Cluster with highest Avg venue rating is preferred • Cluster with highest Venue Count is secondary

Neighborhood Clustering Results

Detailed information can be viewed in the "pdf" dataframe within the program:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Rating	Venue Color
0	Hayes Valley/Tenderloin /North of Market	37.780	-122.420	PLÄJ	37.778733	-122.422106	Scandinavian Restaurant	9	Green
1	Hayes Valley/Tenderloin /North of Market	37.780	-122.420	Urban Bowls	37.778139	-122.422168	Poke Place	3	Orange
2	Hayes Valley/Tenderloin /North of Market	37.780	-122.420	Brenda's French Soul Food	37.782896	-122.418897	Southern / Soul Food Restaurant	8	Blue
3	Hayes Valley/Tenderloin /North of Market	37.780	-122.420	Robin	37.779127	-122.423378	Sushi Restaurant	2	Red
4	Hayes Valley/Tenderloin /North of Market	37.780	-122.420	DragonEats	37.778289	-122.423266	Vietnamese Restaurant	7	Blue

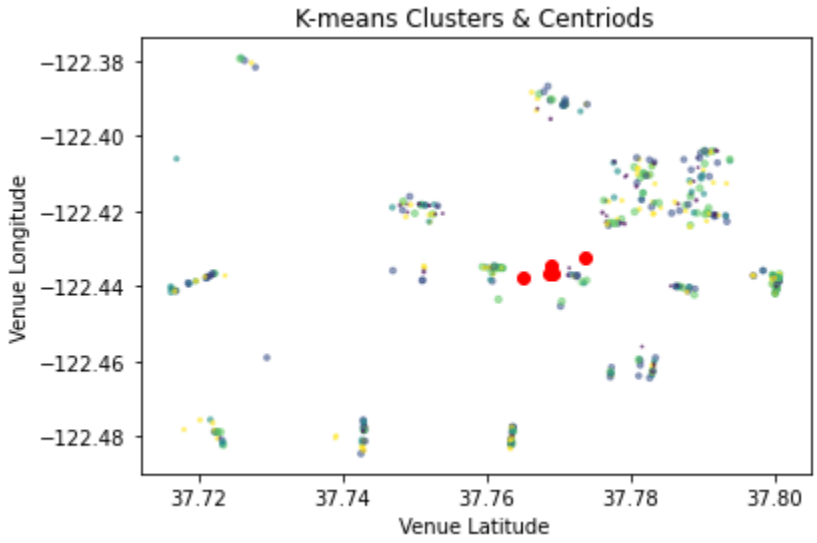
Here is a display of the results from the neighborhood clustering which specifies the highest neighborhood Avg venue Rating and lists all neighborhoods that have the highest rating:

```
**NEIGHBORHOOD CLUSTERING RESULTS**
The highest Average Venue Rating is 6
The venues with the highest Average Rating are listed below:
```

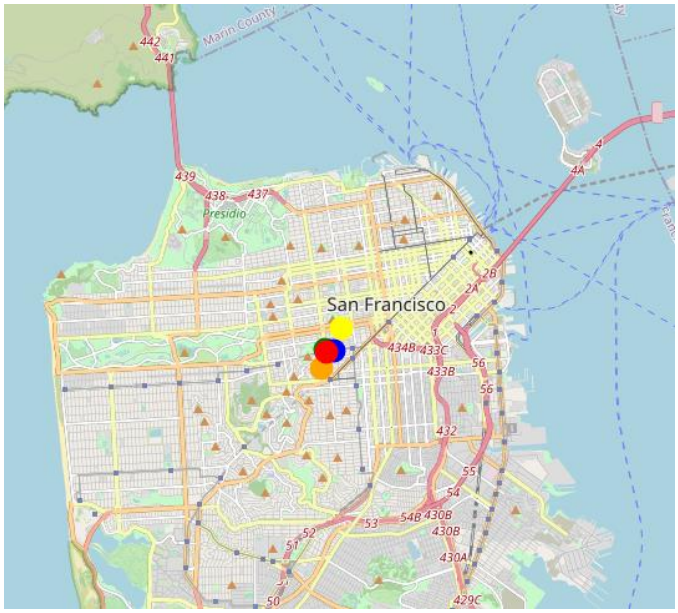
	Neighborhood	Venue Count	Avg Rating
16	Bayview-Hunters Point	2	6
4	Chinatown	30	6
11	Haight-Ashbury	12	6
18	Twin Peaks-Glen Park	7	6
14	Sunset	30	6
5	Polk/Russian Hill (Nob Hill)	30	6
15	Marina	30	6

K-means Clustering Results

Below is a plot of the K-means clustering. In this program execution this clusters are close together due to closeness in ratings across the venues. The distance separating the clusters varies, especially if Rate Limiting is in effect and random ratings are used.



Using the geographical coordinates of each cluster's centroids , they are plotted on a map of San Francisco below:



Here is First-part sample results:

Highest Avg Venue Rating per cluster

```
**K-means RESULTS Part 1**
K-Means Cluster 3 has the Highest Average Venue Rating
K-means Clusters Ranked by Highest Average Venue Rating
```

	Cluster	AVG Rating	Venue Count	Avg Rating
3	3	9.55	103.0	9.55
1	1	7.44	78.0	7.44
2	2	5.43	89.0	5.43
0	0	3.45	83.0	3.45
4	4	1.55	92.0	1.55

And Second-part sample results:

Most Venues in Cluster

```
**K-means RESULTS Part 2**
K-Means Cluster 3 has the Highest Venue Count
K-means Clusters Ranked by Highest Venue Count
```

	Cluster	AVG Rating	Venue Count	Avg Rating
3	3	9.55	103.0	9.55
4	4	1.55	92.0	1.55
2	2	5.43	89.0	5.43
0	0	3.45	83.0	3.45
1	1	7.44	78.0	7.44

Detailed information on all clusters and venues included (only Cluster 1 shown)

```
Summary of Venues in cluster 1
Total Venues in Cluster 1 = 88
Average Venue Rating = 9.6
```

	Cluster Label	Neighborhood	Venue	Venue Category	Venue Rating
389	1	Bayview-Hunters Point	Day Darnet Catering	Food	10
154	1	Inner Mission/Bernal Heights	Al's Place	New American Restaurant	10
246	1	Western Addition/Japantown	My Ivy	Thai Restaurant	10
79	1	Potrero Hill	Poke Delish	Food Truck	10
113	1	Chinatown	Le Colonial	Vietnamese Restaurant	10
121	1	Polk/Russian Hill (Nob Hill)	Bob's Donuts	Donut Shop	10
127	1	Polk/Russian Hill (Nob Hill)	Cordon Bleu	Vietnamese Restaurant	10
94	1	Chinatown	Big 4 Restaurant	American Restaurant	10

Discussions and Observations

- Results will vary over time and iterations
- Results when using randomized Venue Rating vary greatly when actual Venue Ratings are used.
- All venues had actual Foursquare ratings: Actual range from 7.4 to 9.6
- Of the 21 San Francisco neighborhoods evaluated, on Outer Richmond did not have any restaurants identified or rated by Foursquare.
- For K-means clustering, I found that through many trials, creating 5 clusters generally provided the best result and avoided outliner venues.
- The stochastic nature of K-mean embeds some level of randomness to its results i.e. running the program in identical data can lead to different results.
- Because of the narrow range of actual Venue Ratings, clustering results (Neighborhood and K-means) were generally concentrated. However, when random values are used as Venue Ratings, the clusters were more spread-out.
- Calls to the Foursquare API were very limiting due to quotas and rating limiting associated with the Personal-Developer account. The impact being only 1 API call per day would be successful.
- It has been observed that the Foursquare API is not entirely stable in-that it may yield different data. For example, sometimes neighborhoods won't have any venues associated with them. Other times neighborhoods are not passed. These issues may cause the program to generate an error, re-running the program and making another call to the API corrects this.
- Neighborhood Clustering deliver narrower geographical coverage because the venues are confined to 1 neighborhood, but broader rating ranges
- K-means clustering provide broader geographical coverage, but weighs venue ratings higher than proximity

Conclusions

General Conclusions

San Francisco is a city with a richness for diversity. This diversity is not only revealed by its population, but also the tremendous variety of restaurants and neighborhoods.

Neighborhood Analysis by Venue Ratings did not yield an actionable conclusion. While the methodology and applied analytical techniques were sound, the venue ratings data was not diverse enough to distinguish venues or their neighborhoods from each other. The Venue Rating range only varied between 7.4 and 9.6 – (respectively blue and green in color). This lead to all venues were clustered fairly closely together. This result was not actionable because using the actual data did not produce a result where a user could pick a venue / location based on rating. *Like Rome, all roads lead to the same location*.

When Foursquare's rating limiting was in effect, random numbers were applied to the venue ratings and the results where much more actionable because the venue rating varied between 1 and 10, this driving diversity in the locations.

Therefore, it is concluded that without increased rating diversity for the venues, using venue ratings to identify ideal locations and venue cannot be achieved.

Lastly, there is a substantial more analysis and results that could be performed for this project. However time and resource limits dictate the Capstone Project "Neighborhood Analysis by Venue Ratings " be concluded in its current state.

Specific Conclusions

Analysis of results did provide some interesting conclusions based on the methods used:

- If the desired goal is to focus on a central location with acceptance of a mix of venue ratings, then Neighborhood Cluster should be used
- If the desired goal is to focus on highly rated venues with acceptance of greater distances, then K-Means clustering should be used