Restaurant/Venue Distribution in San Francisco

Christopher Richardson

July 29, 2019

Abstract

The city of San Francisco has long been regarded as a multicultural hub which boasts a wide variety of international restaurants, shops, and neighborhoods. While the casual meander-er may look at the countless shops that dress the streets of San Francisco as nothing more than options for personal entertainment, a deeper look into the string of businesses would reveal the complex and intricate web of business connectivity. This report aims to reveal a methodology that would be of use to any food wholesaler who seeks a more thorough understanding of client (restaurant) distribution and how neighborhoods can be categorized based on their venues.

1 Introduction

In order to support the numerous businesses that spread throughout the city, supply chains must be established tactically in order to maximize efficiency and output. To better tackle the problem, the analysis will be broken into two sections. Section 1 will look at the distribution of various ethnic restaurants within San Francisco. Intuitively, finding pockets of greater client concentration could lead to, among other benefits, an increase in operational efficiency and client satisfaction due to lower transport costs and a closer proximity between business partners. Section 2 will look at the clustering of neighborhoods in San Francisco. This will benefit food distributors as they will gain a better sense of which neighborhoods are more gastronomic as opposed to, say, which neighborhoods are catered towards outdoor activities.

2 Data

The FourSquare Developer will be leveraged in order to acquire all pertinent data for both sections. Section 1 will require data on ethnic restaurants while section 2 will require data on a varied group of venues. The features that will be accessed for each venue (restaurant and other) will be the venue name, the venue coordinates, and the venue category. This will be read into a data frame in order to manipulate the data and to effectuate the required analysis.

It must be noted that FourSquare does not contain data on a venue's neighborhood. Therefore, in order to group by neighborhood, boundaries must first be set for the neighborhood in order to determine the neighborhood in which a venue resides. As such, the city will be divided into its ten primary neighborhoods via a GEOJSON file provided by the Data Visualization course offered by IBM through Coursera. In short, the GEOJSON file will define the boundaries of each respective neighborhood on a map. Then, the coordinates of each restaurant can be tested to determine its neighborhood.

3 Methodology

3.1 Venue Data

The first step was to acquire the relevant venue information. By using the FourSquare API, the coordinates and the name of restaurants of Japanese, Chinese and Italian cuisine were retrieved for section 1 with the help of a query (i.e. a key-word search). For section two, no query was specified in order to retrieve a list of varied venues.

3.2 GEOJSON

Next, the GEOJSON file is read in. In short, the GEOJSON file outlines the boundaries of each respective neighborhood on a folium map by outlining them with polygon-shaped rings. Once completed, it is possible to determine if a set of coordinates fall within a particular ring (i.e. neighborhood). This will allow each venue to be assigned a neighborhood.

3.3 Exploratory Analysis

3.3.1 Section 1: Cuisine Distribution

Initially, a folium map was created where each restaurant was plotted. Japanese restaurants are plotted in red, Chinese restaurants in blue, and Italian restaurants in green. This exploratory stage was to ensure that the 100-restaurant sample (which is the API call limit with FourSquare) of each cuisine qualitatively appears distributed through the city and was not, for example, taken in sequence based on address (and consequently geographically concentrated). Afterwards, the GEOJSON file was read in and plotted to gain a qualitative sense of the distribution of the 3 cuisines in the various neighborhoods. The compiled results are observable in figure 1:

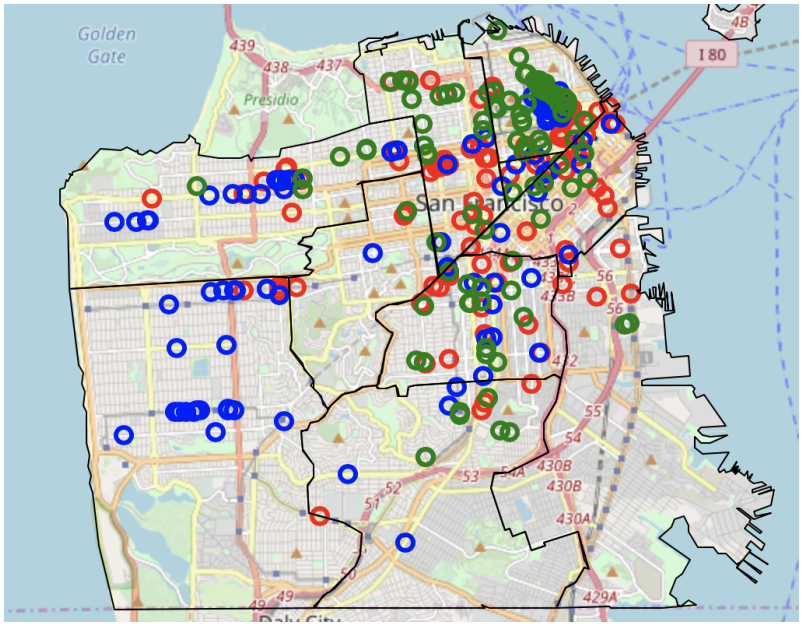


Figure 1: Restaurant Plot with Neighborhood Borders
(Red = Japanese / Blue = Chinese / Green = Italian)

Now that we have established that a certain degree of restaurant distribution exists, it is possible to assign each restaurant with a neighborhood. This is accomplished by passing a restaurant's coordinates through the GEOJSON file. As previously stated in section 2:Data, the coordinates for each neighborhood in the GEOJSON file are linked together in a polygon-shaped ring. Using the Shapely package in Python, the set of coordinates can be transformed into a geo-spatial ring. This allows python to identify whether or not a set of coordinates lie within the shape's boundaries. Therefore, it is possible to determine the neighborhood to which a restaurant belongs. Figure 2: Ethnic Restaurants displays a sample group of restaurants (with their respective neighborhoods) for each cuisine. Each data frame consists of 100 restaurants.

	Name	Category	Longitude	Latitude	Neighborhood
13	Okoze Sushi	Sushi Restaurant	-122.419266	37.799191	CENTRAL
14	Elephant Sushi	Sushi Restaurant	-122.418939	37.798623	CENTRAL
15	Sushirrito	Sushi Restaurant	-122.401675	37.794820	CENTRAL
16	Sushi Taka	Sushi Restaurant	-122.404541	37.793642	CENTRAL
17	Hashiri	Japanese Restaurant	-122.407833	37.782994	SOUTHERN
18	Okane	Japanese Restaurant	-122.403172	37.770727	SOUTHERN

(a) Japanese Restaurants

	Name	Category	Longitude	Latitude	Neighborhood
89	Cheung Hing	Chinese Restaurant	-122.488911	37.753750	TARAVAL
90	Ming's Diner	Chinese Restaurant	-122.489171	37.742459	TARAVAL
91	Bamboo Restaurant	Chinese Restaurant	-122.420675	37.790165	NORTHERN
92	Wonderland Restaurant	Chinese Restaurant	-122.430554	37.772186	NORTHERN
93	San Wang Restaurant	Chinese Restaurant	-122.429396	37.785738	NORTHERN
94	Gourmet Carousel	Chinese Restaurant	-122.423813	37.789148	NORTHERN

(b) Chinese Restaurants

	Name	Category	Longitude	Latitude	Neighborhood
71	Fiorella	Italian Restaurant	-122.484510	37.781887	RICHMOND
72	Osteria	Italian Restaurant	-122.446926	37.788162	RICHMOND
73	Giorgio's Pizzeria	Italian Restaurant	-122.461098	37.783091	RICHMOND
74	La Ciccia	Italian Restaurant	-122.426531	37.742008	INGLESIDE
75	Emmy's Spaghetti Shack	Italian Restaurant	-122.420346	37.745022	INGLESIDE
76	Manzoni	Italian Restaurant	-122.433898	37.734678	INGLESIDE

(c) Italian Restaurants

Figure 2: Ethnic Restaurants

Once each restaurant (separated by cuisine) has its assigned neighborhood, the number of restaurants (for each cuisine) in a given neighborhood can be counted.

	Neighborhood	Japanese Restaurant Count
0	CENTRAL	16
1	SOUTHERN	17
2	BAYVIEW	18
3	MISSION	0
4	PARK	3
5	RICHMOND	12
6	INGLESIDE	4
7	TARAVAL	3
8	NORTHERN	24
9	TENDERLOIN	2

(a) Japanese Restaurant Count by Neighborhood

	Neighborhood	Chinese Restaurant Count
0	CENTRAL	30
1	SOUTHERN	8
2	BAYVIEW	9
3	MISSION	0
4	PARK	1
5	RICHMOND	17
6	INGLESIDE	5
7	TARAVAL	22
8	NORTHERN	6
9	TENDERLOIN	1

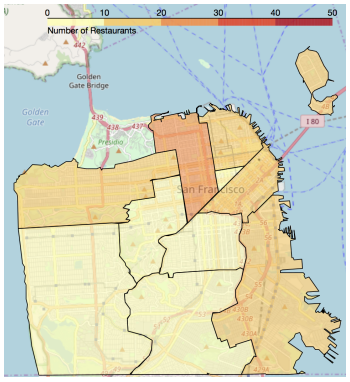
(b) Chinese Restaurant Count by Neighborhood

	Neighborhood	Italian Restaurant Count
0	CENTRAL	43
1	SOUTHERN	7
2	BAYVIEW	16
3	MISSION	0
4	PARK	2
5	RICHMOND	6
6	INGLESIDE	6
7	TARAVAL	0
8	NORTHERN	17
9	TENDERLOIN	1

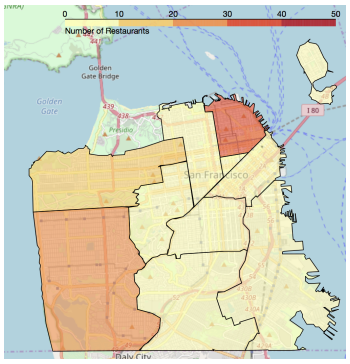
(c) Italian Restaurant Count by Neighborhood

Figure 3: Restaurant Count

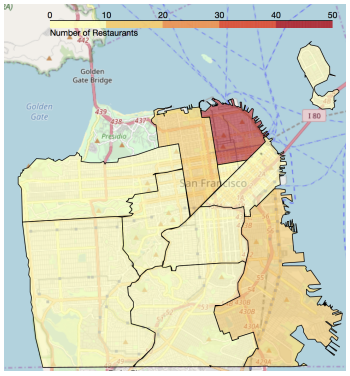
As the frequency of various cuisines in each neighborhood has been calculated, they can now be plotted on a choropleth map to visually depict the distribution using the Folium library.



(a) Japanese Restaurant Distribution



(b) Chinese Restaurant Distribution



(c) Italian Restaurant Distribution

Figure 4: Cuisine Distribution

3.3.2 Section 2: Neighborhood Clustering

This section will explore the distribution of venue types within the neighborhoods of San Francisco. Keeping this goal in mind, it is necessary to derive a model which can categorically cluster the neighborhoods based on similarities. In order to do so, the KMeans machine learning algorithm will be used. KMeans is a vector quantization method whereby a data point is put into a cluster with the nearest Euclidean mean. In other words, it is used to cluster points of greater similarity. Firstly, however, the data must be prepped. Similar to the process in section 3.3.1, venue data was retrieved from the FourSquare API. Again, only 100 venues can be retrieved due to the API call limit. Therefore, this is again an approximation of categorical distribution. Figure 5 below displays the first 5 venues:Venues. Note that the data frame consists of 100 venues.

	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Ina Coolbrith Park	37.798314	-122.413612	Park
1	Collis P. Huntington Park	37.792162	-122.412154	Park
2	City Lights Bookstore	37.797695	-122.406452	Bookstore
3	The House	37.798434	-122.407187	Asian Restaurant
4	Maritime Wine Tasting Studio	37.797292	-122.405652	Wine Bar

Figure 5: Venues

As was done in section 3.3.1, the coordinates of the venues were passed through the GEOJSON file to determine the neighborhood to which each venue belonged. A sample of this data frame is displayed in figure 6:Venues with Neighborhoods.

	Category	Cluster Label	Latitude	Longitude	Name	Neighborhood
16	Coffee Shop	1.0	37.791320	-122.400983	Blue Bottle Coffee	CENTRAL
17	Brazilian Restaurant	1.0	37.806080	-122.418557	Cafe de Casa	CENTRAL
18	Science Museum	1.0	37.800864	-122.398556	Exploratorium	CENTRAL
19	Scenic Lookout	1.0	37.811002	-122.410751	View of Alcatraz	CENTRAL
20	Wine Shop	1.0	37.788039	-122.401466	Flatiron Wine and Spirits	SOUTHERN
21	Grocery Store	1.0	37.785540	-122.405455	Trader Joe's	SOUTHERN
22	Art Museum	1.0	37.785894	-122.400897	San Francisco Museum of Modern Art	SOUTHERN
23	Cycle Studio	1.0	37.790340	-122.397771	SoulCycle SoMa	SOUTHERN

Figure 6: Venues with Neighborhoods

In order to apply the machine learning algorithm KMeans, the data must first be prepared. As stated, KMeans will look for similarities between data points (i.e. the neighborhoods) by looking at their characteristics (i.e. the categories of venues it contains). Then, it will group similar data points together into clusters. Therefore, the data must be one-hot encoded.

	Neighborhood	Accessories Store	Art Gallery	Art Museum	Asian Restaurant	Bakery	Baseball Stadium	Bath House	Bookstore	Brazilian Restaurant	...	Spiritual Center	Street Food Gathering	Sushi Restaurant	Tea Room	The
0	CENTRAL	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	CENTRAL	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	CENTRAL	0	0	0	0	0	0	0	1	0	...	0	0	0	0	
3	CENTRAL	0	0	0	1	0	0	0	0	0	...	0	0	0	0	
4	CENTRAL	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

5 rows × 64 columns

Figure 7: Venues with Neighborhoods

Next, the data frame can be grouped by neighborhood and the characteristics within each neighborhood can be averaged. This will display the percent composition of each venue category

for each neighborhood. For example, as seen in figure 8:Neighborhood Composition, 2.5 percent of the venues in the neighborhood 'Northern' are sushi restaurants. Keep in mind that these calculations are based on a 100-venue sample and so this 2.5 percent is simply an approximation.

	Neighborhood	Accessories Store	Art Gallery	Art Museum	Asian Restaurant	Bakery	Baseball Stadium	Bath House	Bookstore	Brazilian Restaurant	...	Spiritual Center	Street Food Gathering	Sushi Restaurant	Tea Room
0	BAYVIEW	0.000	0.000	0.000000	0.00	0.000000	0.000000	0.000	0.000	0.00	...	0.000	0.000	0.000	0.000000
1	CENTRAL	0.000	0.000	0.000000	0.05	0.000000	0.000000	0.000	0.050	0.05	...	0.000	0.000	0.000	0.000000
2	NORTHERN	0.025	0.025	0.000000	0.00	0.000000	0.000000	0.025	0.025	0.00	...	0.025	0.025	0.025	0.000000
3	PARK	0.000	0.000	0.000000	0.00	0.000000	0.000000	0.000	0.000	0.00	...	0.000	0.000	0.000	0.000000
4	RICHMOND	0.000	0.000	0.000000	0.00	0.166667	0.000000	0.000	0.000	0.00	...	0.000	0.000	0.000	0.166667
5	SOUTHERN	0.000	0.000	0.074074	0.00	0.000000	0.037037	0.000	0.000	0.00	...	0.000	0.000	0.000	0.000000
6	TENDERLOIN	0.000	0.000	0.500000	0.00	0.000000	0.000000	0.000	0.000	0.00	...	0.000	0.000	0.000	0.000000

Figure 8: Neighborhood Composition

Lastly, to gain a clearer understanding of the meaning behind figure 8, the data will be transformed to display the top 10 most common venues of each neighborhood in figure 9:Most Common Venues per Neighborhood. This will provide a more digestible format of the data.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	BAYVIEW	Grocery Store	Playground	Yoga Studio	Ice Cream Shop	Historic Site	Hawaiian Restaurant	Gym / Fitness Center	Gym	Greek Restaurant	Garden
1	CENTRAL	Wine Bar	Coffee Shop	Park	Italian Restaurant	Bookstore	Greek Restaurant	Men's Store	Gym / Fitness Center	Scenic Lookout	Science Museum
2	NORTHERN	Park	Gym	Liquor Store	Ice Cream Shop	Coffee Shop	Concert Hall	Grocery Store	Opera House	Music Venue	Massage Studio
3	PARK	Park	Rock Club	Yoga Studio	Hotel	Hawaiian Restaurant	Gym / Fitness Center	Gym	Grocery Store	Greek Restaurant	Garden
4	RICHMOND	Yoga Studio	Tea Room	Park	Bakery	Salon / Barbershop	Trail	Grocery Store	Gym	Greek Restaurant	Cocktail Bar
5	SOUTHERN	Coffee Shop	Yoga Studio	Art Museum	Gym	Wine Shop	Farmers Market	Brewery	Clothing Store	Café	Mexican Restaurant
6	TENDERLOIN	Art Museum	Theater	Yoga Studio	Garden	Concert Hall	Cycle Studio	Deli / Bodega	Farmers Market	Food Truck	Grocery Store

Figure 9: Most Common Venues per Neighborhood

3.3.3 Model Building

The KMeans algorithm clusters data points based on similarity; however, it is necessary to specify the number of clusters to the algorithm. The aim of the KMeans algorithm is to minimize intra-cluster Euclidean distance and to maximize inter-cluster Euclidean distance while alleviating any over-fitting or under-fitting. Therefore, the elbow method can be used to determine the optimal number of clusters. The optimal number of clusters occurs at the kinked 'elbow' in the graph when plotting the sum of squared distances on the y-axis against the number of clusters in the x-axis, where the sum of squared distances is calculated by summing the squared distances of each point and its cluster's epicentre.

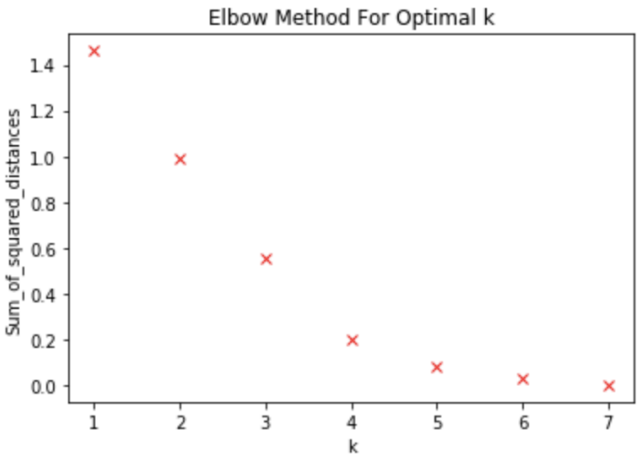


Figure 10: Elbow Method to determine Optimal Cluster (K) Quantity

As seen in figure 10, the optimal number of clusters is 4. With this, we may now construct the KMeans algorithm and fit it using the data from figure 8: Neighborhood Composition. Once this is completed, the algorithm will return the clusters for each neighborhood. We can then append this information to figure 9 to more clearly depict the cluster trends.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	0	BAYVIEW	Grocery Store	Playground	Yoga Studio	Ice Cream Shop	Historic Site	Hawaiian Restaurant	Gym / Fitness Center	Gym	Greek Restaurant	Garden
1	1	CENTRAL	Wine Bar	Coffee Shop	Park	Italian Restaurant	Bookstore	Greek Restaurant	Men's Store	Gym / Fitness Center	Scenic Lookout	Science Museum
2	1	NORTHERN	Park	Gym	Liquor Store	Ice Cream Shop	Coffee Shop	Concert Hall	Grocery Store	Opera House	Music Venue	Massage Studio
3	3	PARK	Park	Rock Club	Yoga Studio	Hotel	Hawaiian Restaurant	Gym / Fitness Center	Gym	Grocery Store	Greek Restaurant	Garden
4	1	RICHMOND	Yoga Studio	Tea Room	Park	Bakery	Salon / Barbershop	Trail	Grocery Store	Gym	Greek Restaurant	Cocktail Bar
5	1	SOUTHERN	Coffee Shop	Yoga Studio	Art Museum	Gym	Wine Shop	Farmers Market	Brewery	Clothing Store	Café	Mexican Restaurant
6	2	TENDERLOIN	Art Museum	Theater	Yoga Studio	Garden	Concert Hall	Cycle Studio	Deli / Bodega	Farmers Market	Food Truck	Grocery Store

Figure 11: Most Common Venues per Neighborhood and Neighborhood Clustering

The neighborhoods can now be plotted based on their clusters in a choropleth-style map in figure 12:Cluster Map. This will render a clearer depiction of the cluster distribution. Note that certain neighborhoods are black as none of the 100 venues are in those neighborhoods.

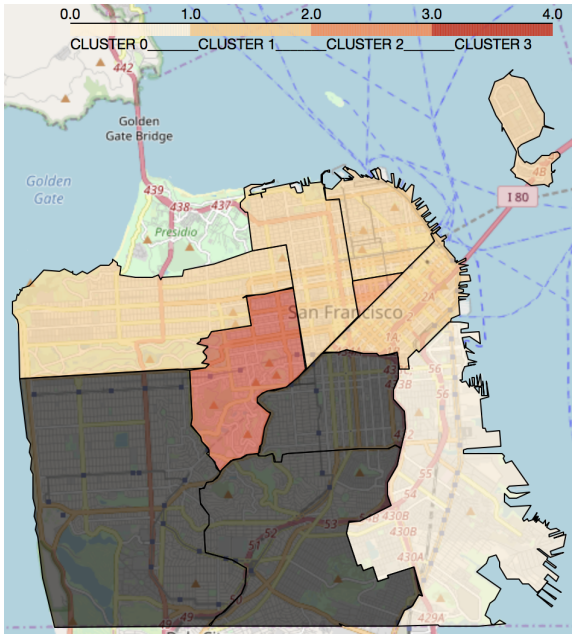
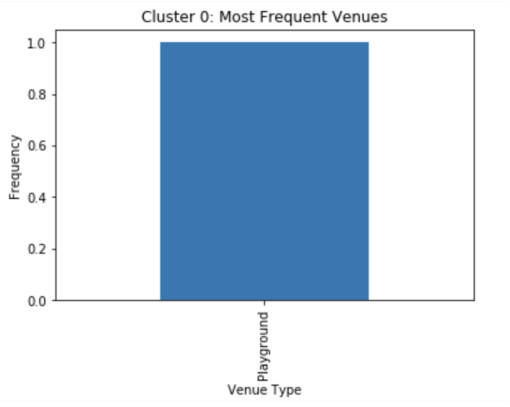
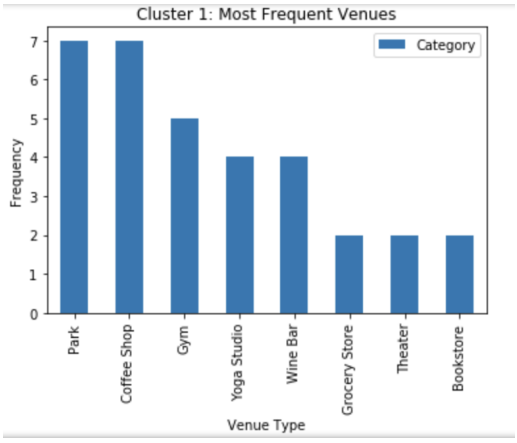


Figure 12: Cluster Map

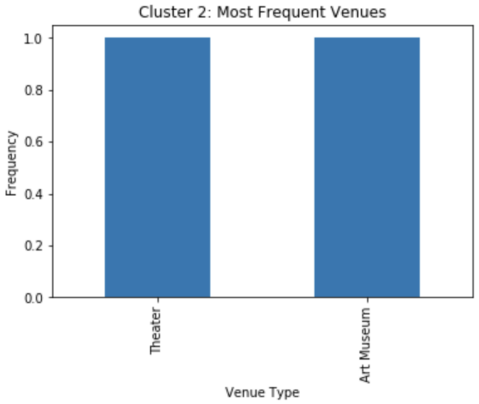
Histograms will reveal the composition of each cluster by displaying the most common venues in each cluster; this will allow us to decipher how the clusters were grouped by the KMeans algorithm.



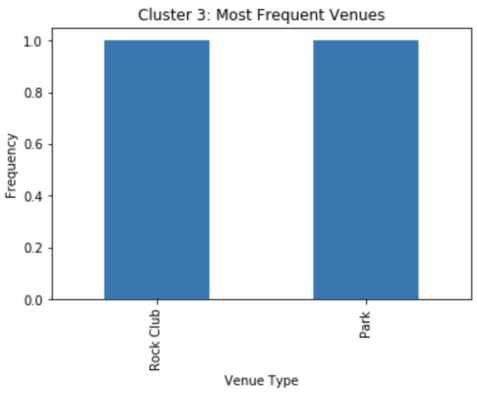
(a) Cluster 0



(b) Cluster 1



(c) Cluster 2



(d) Cluster 3

Figure 13: Cuisine Distribution

4 Results

4.1 Section 1: Cuisine Distribution

To recapitulate, section 3.3.1 aimed to display the distribution of Japanese, Chinese, and Italian restaurants in San Francisco so that wholesale distributors could determine which neighborhoods best suited their produce. As can be seen from figure 3:Restaurant Count, both Chinese and Italian cuisine have an overwhelming concentration in the Central neighborhood of San Francisco. Conversely, Japanese cuisine is more evenly distributed throughout the city. These statements are confirmed by referring to figure 4:Cuisine Distribution. Italian cuisine appears to be highly concentrated in the Central neighborhood, Chinese cuisine is also highly concentrated in the Central neighborhood yet possesses a strong present in the Taraval neighborhood, and Japanese cuisine is relatively distributed among the coastal neighborhoods evenly with a particular concentration in the Northern neighborhood. Therefore, a wholesaler can determine an optimal location, geographically speaking, based on these findings.

4.2 Section 2: Neighborhood Clustering

As was discussed in section 3.3.2, the section's objective was to group the neighborhoods into clusters based on their similarities and to identify the common intra-cluster attributes. This led to identifying a 'theme' for each cluster. Examples of possible classification labels can be identified as 'outdoor activities' or 'urban gastronomy'.

By referring to figure 13(a), the only venue found in cluster 0 is a playground. Therefore, we will label cluster 0 as 'Playgrounds'.

Figure 13(b), which represents cluster 1, possesses areas with a large number of gyms, parks, yoga studios and coffee shops. There is an evident tendency towards a 'healthy living' due to the outdoor and exercise-driven nature of these areas, as well as some traits of 'urban living' due to the high number of coffee shops. Therefore, cluster 1 will be labeled as 'healthy, urban living'.

In figure 13(c), which represents cluster 2, there is a pattern of artistic venues as represented by theaters and art museums. As such, cluster 2 will be labeled as 'artsy areas'.

For figure 13(d), which represents cluster 3, there is no obvious association between a rock clubs and parks. Therefore, we may attribute this clustering to a coincidental proximity between the data points. This will be further elaborated on in the 'Discussion' section.

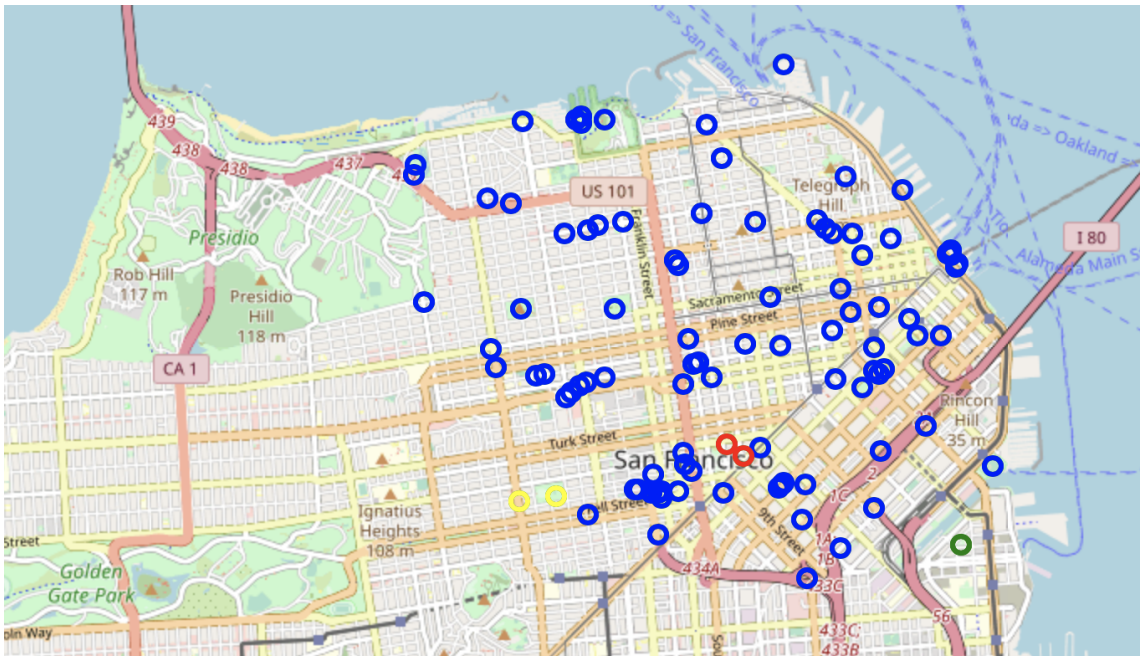


Figure 14: Clustered Restaurants by Color
Green = Cluster 0 (Playgrounds) // Blue = Cluster 1 (Healthy, Urban Living) // Red = Cluster 2 (Artsy Areas) // Yellow = Cluster 3 (Undetermined)

5 Discussion

5.1 Limitations

Certain limitations arose while developing the report, namely pertaining to data. While FourSquare provides users with access to large amounts of detailed information, the 100-venue limit per call became problematic. The goal of the report was to depict a clear image of the distribution of different types of cuisine and venues within San Francisco; however, as FourSquare only returns recommended/popular venues when using the 'explore' feature, the resulting information became biased towards popularity.

In addition, the distribution of venue types throughout the city were not as varied as desired. For example, referring to figure 12:Cluster Map, 3 of the 10 neighborhoods had no venues in within their borders. Additionally, by observing figure 14:Clustered Restaurants by Color, it is evident that the majority of venues were categorized into the same cluster. A possible solution would be to access a larger and more detailed data set in order to account for less popular venues (which is determined by FourSquare) but are still pertinent in describing/defining the different areas of San Francisco. In addition to the high concentration in cluster 1, data was also lacking among the other clusters. Therefore, although all clusters were labeled based on their composition in section 4:Results, this was done warily as the lack of supporting data has rendered the results unconvincing.

5.2 Improvements and Further Investigation

In order to answer the questions of this report more thoroughly, larger data sets are required. As stated in 5.1:Limitations, the bias of the data and the lack of data led to shallow and possibly inaccurate generalizations of the composition of the various neighborhoods. A larger and more varied data set would ideally alleviate the 'popular/recommended' bias as it would depict a more realistic picture of San Francisco's venue distribution.

While it is beneficial to have an understanding of venue and restaurant distribution, a wholesaler would require additional information in order to make educated and tactical business decisions. A relevant question that a wholesaler could consider is which neighborhoods are experiencing new entrants (and possibly clients) into the restaurant market. By posing such a question, a wholesaler could gain a first-mover advantage by foreseeing potential client entrance. Therefore, this report suggests that further research be fulfilled regarding restaurant openings within the city. By coupling the results of such a research paper with the methodology of this report, a wholesaler will not only gain a better sense of where potential clients are establishing themselves, but also how the composition of the various neighborhoods are being affected by said entrants. This could influence future business in a variety of ways: over-saturation of various ethnic cuisines in certain areas, increase in food traffic due to the opening of new restaurants, etc. One method would be to aggregate information from both reports to help foresee an opening business' success. The future report would help identify the opening restaurants while the methodology of this report could reveal if the potential client restaurants 'fit' thematically in their neighborhood. While a new restaurant could potentially satisfy all of the on-paper requirements of a wholesaler (price schedule, delivery rate, etc.), it might not be consistent with the gastronomic trend of the area and so customer volume could suffer. In other words, the future report would provide a foundation while this report would help filter the results.

6 Conclusion

To recapitulate, the purpose of this report was to provide food wholesalers in San Francisco with two key pieces of information: Firstly, it depicted the distribution of various ethnic cuisines within the city. Secondly, it clarified the general groupings of venues in the various neighborhoods of San Francisco.

By utilizing the FourSquare database, we were able to retrieve valuable information which led to further insight on venues, and more specifically restaurants, in the San Francisco area. The results were as follows: Italian cuisine was concentrated in the Northern neighborhood, Chinese cuisine was concentrated in both the Northern and Taraval neighborhoods, and Japanese cuisine was well distributed along the coastal neighborhoods. Regarding the clustering of neighborhoods, cluster

0 was categorized as 'playgrounds', cluster 1 was categorized as 'urban, healthy living', cluster 2 was categorized as 'artsy areas' and cluster 3 could not be identified due to lacking patterns in its composition. This is accredited to a spurious proximity between data points as rock clubs and parks have no evident thematic correlation.

The conclusions drawn from the limitations and subsequent improvements/further investigation section can be summarized as follows: the data sets were too small to paint a truly accurate picture of neighborhood clustering, but the research done in this report has developed the methodology required to complete such a task in addition to depicting cuisine distribution. To build on this report, it is recommended that further research is done on entrants into the restaurant market in San Francisco to identify possible future clients. By pairing this research with the methodology of this report, a wholesaler will gain a better sense of which restaurants represent feasible business partnerships.