# Identifying Optimal Business Locations

## David Simon

# Introduction/Business Problem

## Background

Location of customer-facing businesses are crucial to success. This is evident by companies investing several thousands of dollars in footfall studies before deciding to open a store in a specific location. While marketing research highlights that the strength of a retail brand might exceed the advantages provided by location,[[1]](#footnote-1) it is seldom disputed that a new business has a substantially better chance of survival if its initial location is well-chosen.

## Business Problem

Where should one build a new coffee bar, a sandwich shop or a restaurant? The goal of this project is two-fold: (1) to identify locations of already successful food retail establishments in the City of San Francisco and (2) to attempt to identify locations in the City of San Francisco that possess similar characteristics and may be promising targets in which to build such establishments. The results of this project will be of interest to new entrepreneurs in the food business as well as existing restaurant chains exploring new locations.

# Data Sources

This analysis will crucially rely on data from FourSquare’s API as well as the geospatial data on the boundaries of San Francisco provided to us in an earlier lab. The two initial choices that the modeler has to confront are the following: (1) what constitutes a successful business?, and (2) how does one characterize the vicinity of businesses that makes them successful? I discuss these two questions next.

## Identifying Successful Businesses

Two potential measures of success for a business that can be assessed through the Foursquare API are (1) the rating associated with the business and (2) the number of tips associated with a business. It is beneficial to use these two measures in combination with one another. For example, a high rating with a low number of tips might indicate a place that has very few visitors and a high rating might be a statistical aberration. In a similar fashion, high number of tips but a low rating might mean a bad business that brings about a lot of intense negative reactions from visitors.

A candidate measure for the quality of the business would be to convert all ratings and number of tips into a z-score. That is, every from every rating and number of tips I subtract the mean for a comparable groups (for example, if a Chinese restaurant has a rating of 3.8 and the average Chinese restaurant has a 2.5 rating, this will lead to a score of 3.8-2.5 = 1.3) and then divide this difference with the standard deviation of the scores of all Chinese restaurants (in the above example, if the standard deviation of all Chinese restaurants’ score is 2, I arrive at a z-score of 1.3/2 = 0.65). This way, both the business ratings and the number of tips will have a zero-mean and standard deviation of one distribution. After this standardization, I can add up the z-scores associated with ratings and tips for a given restaurant to find the highest composite scores.

Alternatively, I can just use the number of tips as a measure for potential of the business. Some of the businesses are good, some are bad and that is determined by the quality of people who run the place, the quality of the business concept, etc. However, the number of tips, as far as they serve as a proxy for the number of visits might be a better indicator of exogenous factors to a restaurant’s success – that is, factors that are independent of the quality and the business concept of a place. I would expect that even a bland and boring sandwich place located among bustling office buildings would have a high number of tips.

This methodology allows me to identify the top 5-10 successful business in San Francisco within each category.

## Assessing the Neighborhood of Successful Businesses

## 

For purposes of this project, the assumption will be that the businesses found in the vicinity of one type of successful business will be very similar to each other. For example, as noted above, one would expect a popular lunch spot to be surrounded by office buildings. In the first step, I test the similarity of the neighborhoods of popular places in one category to each other. Unless they are drastically different, the “average” neighborhood of a successful business seems to be a good target for a search.

Once this target is identified, I can define a “coordinate grid” for the city of San Francisco and iterate through coordinates to find locations that are similar to the target. These locations will be suggested as potentially promising locations for a new business of a certain type.

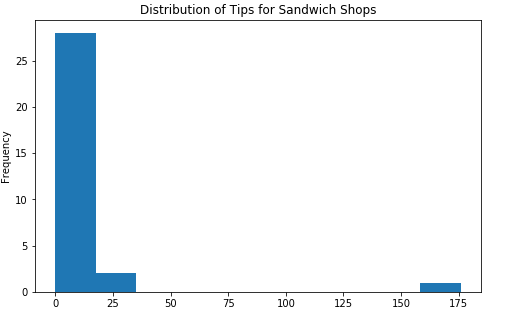
Since this project relies heavily on Foursquare “Premium” calls, which are limited in number, I chose to perform this analysis only on sandwich shops. A simple change of the search term allows to perform this analysis for other types of retail establishments as well.

# Methodology

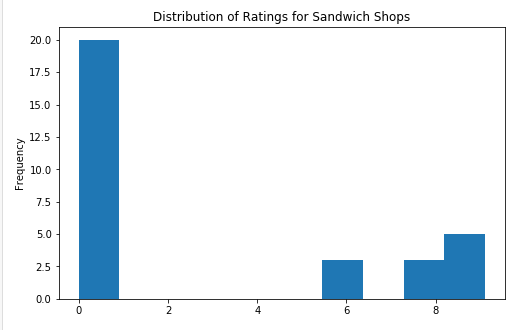
The first task is to find a list of sandwich shops in San Francisco that are covered by the Foursquare database. At first glance, this seems like a straightforward task, since the Foursquare search option allows to query a specific search string (in this case, “sandwich shop”), in the vicinity of a given city (in this case, “San Francisco, CA”). However, the shortcoming of this method is that such a query only returns 50 results. From the Foursquare API documentation, it remains unclear whether these 50 results are selected randomly or by a given criterion.

An alternative avenue I chose to collect data was to set up a grid of evenly spaced points in San Francisco and collect all data on all sandwich shops within 500 feet of them. This method results in potential overlaps that can be eliminated but ensures that all relevant data points are collected. I set up the boundary coordinates of San Francisco using a tool from <https://boundingbox.klokantech.com/> and created equally spaced both longitude and latitude coordinates using the *linspace* method from Numpy. I later verified by a simple plot in Folium that the points identified through this method indeed span most of the City of San Francisco. Through Foursquare’s search function, I identified the (1) latitude, (2) longitude, (3) name and (4) unique id of the sandwich shops associated with each of these grid points. As expected, these searches resulted in a substantial number of overlaps. I eliminated entries where all four of the fields downloaded were identical to avoid having duplicate records in the data.

Since the search query in Foursquare is a basic search, a venue query was necessary to download the tips and ratings associated with each of the venues. This was relatively straightforward, since I could iterate through the venue ids identified in the previous step and pull the associated ratings and tips. An unexpected challenge in this step was the high number of venues with few to no tips and few to no ratings. The relative paucity of ratings and tips might limit the statistical power of the results of this analysis. **Table 1** and **Table 2** below presents the distribution of tips and ratings for each of the sandwich shops, respectively.

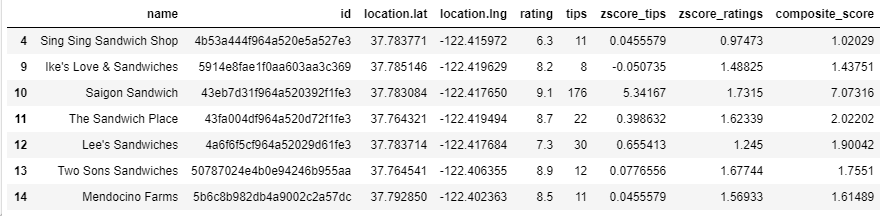


**Table 1.** Distribution of tips for sandwich shops



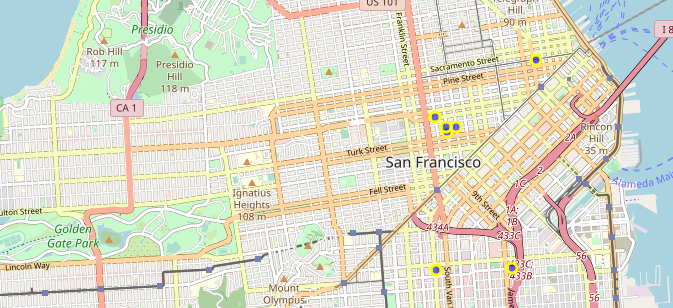
**Table 2.** Distribution of ratings for sandwich shops[[2]](#footnote-2)

After identifying these points, I identified three potential groups for high-achieving sandwich shops: (1) the five sandwich shops with the highest number of tips, (2) the five sandwich shops with the highest average rating, and (3) the five sandwich shop with the highest average combined score of tips and ratings. In order to make an appropriate summation of the tips and ratings for the composite measure, I converted each tips and ratings into a z-score (zero mean and unit standard deviation), by subtracting the sample mean and dividing with the sample standard deviation for each observation of ratings and tips, respectively. As expected, there was a substantial overlap between these three groups and altogether seven sandwich shops were identified as top venues.



**Table 3.** The seven successful sandwich shops

**Table 4.** below shows the geographical distribution of these sandwich shops. It is telling that these shops are clustered in a relatively small area of San Francisco.



**Table 4.** The blue dots denote the selected sandwich places

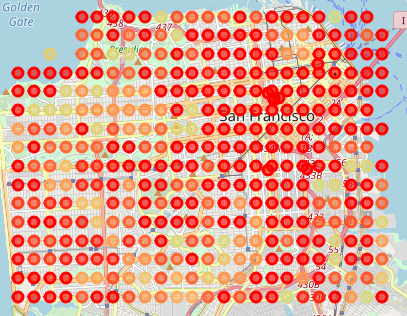
Having identified these locations, the exercise of identifying similar points is relatively simple. I identify neighboring establishments in the 500-yard vicinity of these sandwich shops from the Foursquare API in the same fashion as already performed earlier in this class. Then I set up another grid of San Francisco and identified the neighboring businesses. I then ran the k-means clustering using the relative frequency of the types of shops nearby with a varying number of clusters. In order to strictly delineate visually between the points that are assigned to the same group as the 7 selected sandwich shop, I set the cluster indicator of the clusters that the sandwich shops were assigned to zero, while I added 50 to the cluster number of all other clusters.

# Results

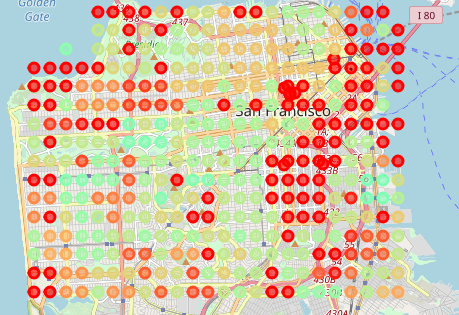
**Table 5, Table 6, Table 7** and **Table 8** show the distribution of clusters with 10, 25, 50, and 100 clusters selected, respectively.



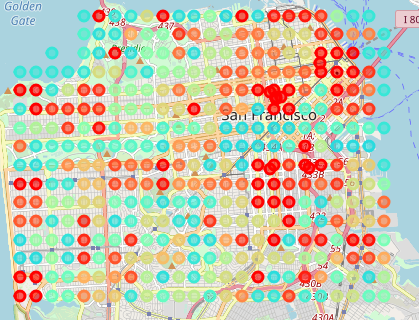
**Table 5.** Clustering results using 10 clusters – Red dots are most comparable to the selected sandwich shops



**Table 6.** Clustering results using 25 clusters – Red dots are most comparable to the selected sandwich shops



**Table 7.** Clustering results using 50 clusters – Red dots are most comparable to the selected sandwich shops



**Table 8.** Clustering results using 100 clusters – Red dots are most comparable to the selected sandwich shops

# Discussion

From the plots it is clear that increasing the number of clusters severely decreases the number of points that are deemed similar by the k-means algorithm to the selected sandwich shops. This is understandable as one would expect a major city to have a number of distinct neighborhoods with different demand patterns. A further piece of external validation is that a number of data points deemed similar to the location of sandwich shops is in downtown San Francisco where large office buildings are located – one would expect higher demand for lunch foods around this area. It is also notable that there is a cluster of similar points around Van Ness, where two of the sandwich shops are located.

# Conclusion

This project identified and documented a method using k-means clustering to find locations similar to a given successful business location. While this study identified such locations, this method is only as useful as the accuracy of the Euclidean distance measurements associated with the neighboring businesses. It is possible, for example, that due to the relatively small number of neighboring businesses and the relatively high number of *types* of businesses associated with the Foursquare database, spurious correlations arise – for example, two locations are placed in similar clusters due to NOT having any given sort of businesses in their vicinity. The possibility for these spurious correlations is further highlighted by the presence of numerous datapoints classified in the same cluster as popular sandwich stores while being on the water in the San Francisco Bay. A possible refinement of this analysis would be to identify the businesses that actually influence success of sandwich shops, whether through a more detailed analysis of the individual store types, rolling up multiple store types into one, or through a more sophisticated quantitative method that allows for identifying relevant factors, such as Principal Component Analysis.

1. See, for example, Swoboda, Bernhard, et al. "The importance of retail brand equity and store accessibility for store loyalty in local competition." Journal of Retailing and Consumer Services 20.3 (2013): 251-262. [↑](#footnote-ref-1)
2. In Tables 1 and 2, a rating or tip number of zero signifies missing values. [↑](#footnote-ref-2)