**Restaurants in San Francisco Neighborhoods**

Kenneth Pak

July 26, 2020

**1. Introduction**

**1.1 Background**

One of the key features in any city is its culinary scene. Food is important to people: not only because it provides a thrice-a-day sustenance, but also because of the allure of novel tastes, the shared social experience with friends, and even just the satisfying of a craving. Technology has been revolutionizing the way people discover and share information, and this is even more so with dining. An increasing number of people are flocking to popular or niche food venues as a result of social media posts, food blogs, and business review platforms like Yelp, Google, and Foursquare.

With that in mind, when a company or individual is looking to open a new food venue, one of the main considerations is location. While outstanding products and quality service are an essential part in retaining customers, with an unfavorable location, the venue might not gain enough traction for the business to start. A poorly placed restaurant might be too out of way for people or not properly cater to its intended demographic, resulting in a stunted or declining revenue. In short, the location of a venue determines a significant portion of its visibility, foot traffic, competitors, and complementary enterprises.

**1.2 Problem**

Determining the optimal location of a restaurant is a large and nuanced question; however, there are ways that the research can be narrowed down. This analysis aims to accomplish that by categorizing and clustering neighborhoods of a city by the restaurants that are present in each of the areas.

**1.3 Scope & Interest**

While this analysis could be performed on a larger or smaller scale in any area, for the scope of this analysis, the different neighborhoods of San Francisco, California will be reviewed. These results will be most directly useful and relevant for businesses, or even consumers, looking to better understand the restaurant landscape in San Francisco.

**2. Data**

**2.1 Data Sources**

The datasets used in this analysis comes from a combination of a webpage, Kaggle file, and the Foursquare database. The geolocation, and income data for each postal code in San Francisco come from scraping the HTML file from [this website](http://zipatlas.com/us/ca/san-francisco/zip-code-comparison/median-household-income.htm); however, since the neighborhood names were missing, a dataset of postal codes and neighborhood names in San Francisco was downloaded from [this file](https://www.kaggle.com/savargaonkar/sf-zipcodes-limited/data) found on Kaggle. All the relevant venue information, such as geolocation, venue name, venue type, were all pulled from the Foursquare API. One important note to make about the Foursquare data is that the venues returned was limited to only the food venues, as that is the scope of the analysis.

**2.2 Data Wrangling**

The three datasets that were used in this analysis were eventually combined into one table, but before that, the datasets were individually cleaned and prepared to ensure proper merging. Since the datasets were sourced from different places, there were some problems with the data.

The first problem that was addressed was the irregular formatting of the data. If the dataset does not contain uniformly formatted data, it could lead to errors in the potential uses. The datasets were stripped of whitespaces and special characters. For instance, the population and average household income columns in the scraped HTML had dollar signs and commas that were included, so those erroneous characters were parsed out.

Furthermore, there were some instances of duplicates, primarily from the Kaggle dataset. For some reason, the CSV file downloaded had each entry listed twice, and had to be removed. With the rest of the datasets, they did not contain any, and since the datasets were small, this was easily confirmed.

Another issue was the structure of the data. While the HTML file gave its data in terms of postal codes, for ease of application and interpretation, the data in was set in terms of neighborhoods. In order to accomplish this, the Kaggle dataset, containing postal codes and neighborhoods, was grouped by postal code and added to the base dataset.

Also, there were inconsistencies of data types between datasets that needed to be accounted for. While the postal code in the Kaggle dataset is an integer type, the postal code in in the HTML page is inherently scraped as a string type; these differences were reconciled by changing them both to a string type. Some other columns like latitude and longitude were changed into float and integer types to better match their identity, namely as numbers and not text.

Finally, outliers were checked for in the dataset. After the Foursquare data were pulled and merged to the main neighborhood dataset, each neighborhood was checked for a count. The maximum number of venues to be pulled from each neighborhood was set at 50, but there were some neighborhoods with abnormally low counts such as four or seven venues. If a neighborhood only has a few venues in the area, it would most likely skew the analysis of analyzing the most popular venues; in a more practical view, the area is most likely heavily residential, industrial, or some other area where food venues do not normally operate.

The summation of this wrangling process was a cleaned dataframe that contained all the nearby food venues for each of the neighborhoods in San Francisco.

**2.3 Feature Selection**

Features for analysis were selected based on their relevancy to segmenting and clustering neighborhoods. For this, the main selected features were the following: neighborhood name, neighborhood geolocation, and venue category. This would allow the neighborhoods to be clustered based on their most common venues. Auxiliary features such as average household income and population were included but not used in the clustering to provide more context behind each neighborhood.

**3. Methodology**

**3.1 Exploratory Analysis**

As not neighborhoods are homogenous in the availability of food venues, the distribution of the venues in each neighborhood were plotted, as shown in Figure 1. Inspecting the plot, the number of food venues nearby a neighborhood can be visualized. While neighborhoods, like North Beach/Chinatown and Marina, have high concentrations of food venues, some neighborhoods, like Lake Merced and Bayview-Hunters Point, have only a few. Since these neighborhoods are eventually be segmented and clustered, having very few data points would skew the clustering interpretation. For instance, Bayview-Hunters Point has one food venue, a Deli/Bodega; it would be erroneous to characterize Bayview-Hunters Point as a Deli/Bodegas-littered neighborhood. Following this reasoning, all neighborhoods with less than or equal to 10 food venues are omitted from the analysis as outliers.

Figure 1: Histogram displaying show all the food venues nearby SF neighborhoods

A screenshot of a cell phone

Description automatically generated

**3.2 One-Hot Encoding**

In order to cluster different neighborhoods based on the frequency of their food venue categories, the categorical data needs to be transformed using one-hot encoding. One-hot encoding is the process of converting categorical data with no ordinal relationship into quantitative data. It allocates a new column for each unique value of the categorical data and displays it in a binary fashion: a value of ‘1’ is placed in the corresponding column, and a value of ‘0’ is placed in all the other columns. In our case, the one-hot encoded dataset can be grouped by the mean of each column to provide a relative measure of frequency, from zero to one, to further analyze. Figure 2 displays a portion of this transformed dataset. There is a total of 75 rows, indicating that there are 75 different types of food venues captured in the SF neighborhoods.

Figure 2: Example of food venue categorical data transformed using one-hot encoding

A screenshot of a cell phone

Description automatically generated

By sorting the dataset by the mean of the one-hot encoded dataset, the dataframe can be easily be reformatted to view the 10 most common categories of venues in each neighborhood, as shown in Figure 3. From the figure, we can get a profile of the venues that exist in each neighborhood. One caveat is that the nearby venues were gathered from a 400-meter radius from the neighborhood’s geolocation; since neighborhoods are not circularly designated, there will be some discrepancies between the actual neighborhood’s venues and the retrieved venues.

Figure 3: A sample of the top 10 most common venues types in SF neighborhoods

A screenshot of a cell phone

Description automatically generated

**3.4 K-Means Clustering**

Since this analysis requires unsupervised categorization, K-Means clustering is a fitting algorithm to apply to create the model. One characteristic of the K-Means clustering algorithm is that the number of clusters need to be specified prior to running the algorithm. To find the optimal number of clusters, a silhouette analysis can be used. A silhouette analysis displays how close different clusters are from each other, and provides a metric, the silhouette score, to determine the optimal number of clusters in a K-Means algorithm. In Figure 4, a silhouette analysis is visualized, and two clusters is seen to have the highest score by far; so for this clustering, two clusters will be used.

Figure 4: A silhouette analysis for the dataframe

A close up of a map

Description automatically generated

Now, the K-Means clustering algorithm can be applied, with a K of two, and each neighborhood is assigned a cluster label: 0 or 1. The results of the analysis will be presented in the next section.

**4. Results**

**4.1 Tabular Visualization**

In Figures 6 and 7 is the summary of all the clustering results, displayed in a table format. This is particularly useful in focusing on the venues common in each neighborhood. In the first cluster (Figure 6), there is a strong similarity, as both neighborhoods present have identical top three venues: food truck, café, and Mexican restaurant. In the second cluster (Figure 7), however, there is less of a distinct characteristic; all the other neighborhoods, 13 out of 15, are present in the second cluster. On inspection, the neighborhoods in the second cluster do not seem to share any strong commonalities.

Figure 6: Cluster 1 results

A screenshot of a cell phone

Description automatically generated

Figure 7: Cluster 2 results

A screenshot of a cell phone

Description automatically generated

**4.2 Geolocation Visualization**

Another form of visualization that was employed was geolocation plotting. Since this analysis is based on geolocation of data points, plotting the neighborhoods, along with their respective clusters, on a map would give the clearest perspective of the location of the results. Using the Folium package in Python, a map with the scope of San Francisco was generated, and the results of the analysis were plotted on the map (Figure 8). Each color corresponds to a cluster, and each point is a neighborhood in San Francisco. The first cluster, colored in red, seems to be adjacent to each other, and located on the middle-right of San Francisco. The other cluster populates the rest of the neighborhoods.

Figure 8: Folium map of the SF neighborhoods, clustered

A close up of a map

Description automatically generated

**5. Discussion**

When viewing the analysis, the most notable point in the results was the overall distribution of the clusters. 13 out of 15 neighborhoods, or roughly 87 percent, were placed in one cluster, while the other two neighborhoods, 13 percent, were placed in the other.

One way to explain this distribution is with the differences in neighborhoods. While some neighborhoods were removed earlier in the exploratory analysis for having too few venues nearby, most of the other neighborhoods might have venues that are too different from each other, which would create a generally distant relationship among them. As a result, the two neighborhoods that were closely related would be clustered together, and the other ones would be lumped together.

With this train of thought, the neighborhoods in the largest cluster should not be interpreted as all homogenous, but instead, less heterogenous from the two neighborhoods in the other cluster. This view is supported when looking at the tabular visualization of the results in Figure 7: there is no visually discernible pattern among the neighborhoods.

Overall, the biggest inference that can be made from this analysis is that the different neighborhoods of San Francisco are generally very different from each other, except for the two that are clustered separately: Potrero Hill and South of Market. This means that if someone was looking to open a new food venue, they would have to look more closely at the profiles of the neighborhoods (Figure 6 and 7) to narrow down their choices. For instance, if that person was looking to open a new Indian restaurant that is catering to the Indian demographic, that person might find the most success checking out the Castro and Noe Valley neighborhoods, as Indian food venues are most common there. If, however, that same person was looking to introduce an Indian food venue but catering to non-Indians, the analysis results show that they might want to stay away from those neighborhoods. This suggestion is made because the frequency of a certain ethnic cuisine in an area suggests that people from that ethnic group are more concentrated in that area. With that mind, consumers could also take these neighborhood segmentations as guidelines on what area to find certain ethnic foods.

**6. Conclusion**

Throughout this study, the food venues of San Francisco neighborhoods were analyzed. The neighborhoods were segmented based on the frequency of different categories of food venues present, and clustered through a K-Means clustering algorithm. This analysis and subsequent visualization proved useful in allowing restauranteurs to have a better preliminary understanding of the profiles of San Francisco neighborhoods. Also, this similarly is shown to be useful to the general public in understanding what kinds of restaurants are present in each neighborhood.

If this analysis was to be taken further, different cities’ neighborhoods could be added to expand the scope of the analysis. This would allow relationships to be drawn among different cities: someone could potentially be recommended a similar neighborhood in a different city, characterized by its food venues. For this to happen, a lot more data would have to be aggregated, but if executed properly, this could result in a deeper level of tangible applications. For instance, a successful restaurant owner in one city could be able to identify other areas in the world that might perform similarly, allowing the owner to expand optimally.