Segmenting and Clustering the School District Offices in the San Mateo County, California

Yangxue Dong

June 25, 2020

1. Introduction

1.1 Background

The San Mateo County Office of Education (SMCOE) supports 23 school districts in the county. Although the county office building is often used to host events and meetings for the districts, it is not always the best location for specific purposes. For example, there is only one small cafeteria near the county office building, so it is hard for people who attend large events such as Educator Recruitment Fair to find places to eat and chat. The location of the county office building is also not ideal for events about outdoor education and arts. Therefore, the county office may seek opportunities to partner with the districts and hosts the events and meetings in the district offices.

1.2 Problem and Audience

The purpose of this project is to leverage the Foursquare location data to help the county office leadership decide which district offices are most appropriate for hosting certain events. The audience are the county office superintendents, the administrators and directors, and the district superintendents. The results can give them a better picture about the locations of the district offices and make event proposals accordingly.

2. Data Acquisition and Cleaning

2.1 Data sources

There were two main data sources for this project:

The Foursquare location data: The Foursquare API was used to explore school districts in San Mateo County. The explore function was applied to get the most common venue categories near each school district, and then this feature was used to group the districts into clusters.

District list and location: This data was be retrieved from the official website of California Department of Education. The public districts data files are posted in both Excel and tab delimited text formats. They are dynamically driven and reflect real-time data. The data files include the location (street address, Latitude, and Longitude) and contact information for all the school districts in California. The file structure is also available on the same web page https://www.cde.ca.gov/ds/si/ds/pubschls.asp, which includes the descriptions of the elements in the downloadable public districts files.

2.2 Data Cleaning

The public districts data file has 22 fields, including CD Code, County, District, Street, City, Zip, State, Mail Street, Mail City, Mail Zip, Mail State, Phone, Ext, AdmFName, AdmLName, AdmEmail, Latitude, Longitude, DOC, DOC Type, Status Type, and Last Update. For the purpose of this project, only three fields were kept, which are District, Latitude and Longitude. In addition, the districts in other counties have been dropped so that the final data frame only keeps the districts in San Mateo County.

3. Methodology and Results

3.1 Exploratory Data Analysis

After creating the API request URL and making the GET request, the top 100 venues that are within a radius of 500 meters of each district office were retrieved. Specifically, the returned information includes the venue name, location (latitude and longitude), and category. The size of the resulting data frame is (394, 7) and the venues returned for each school district are between 1 (La Honda-Pescadero Unified and Menlo Park City Elementary) and 84 (Redwood City Elementary). There are 139 unique categories curated from all the returned venues. See Table 1 for details on the number of venues returned for each district.

Table 1. Number of Venues Within a Radius of 500 Meters of Each District Office District

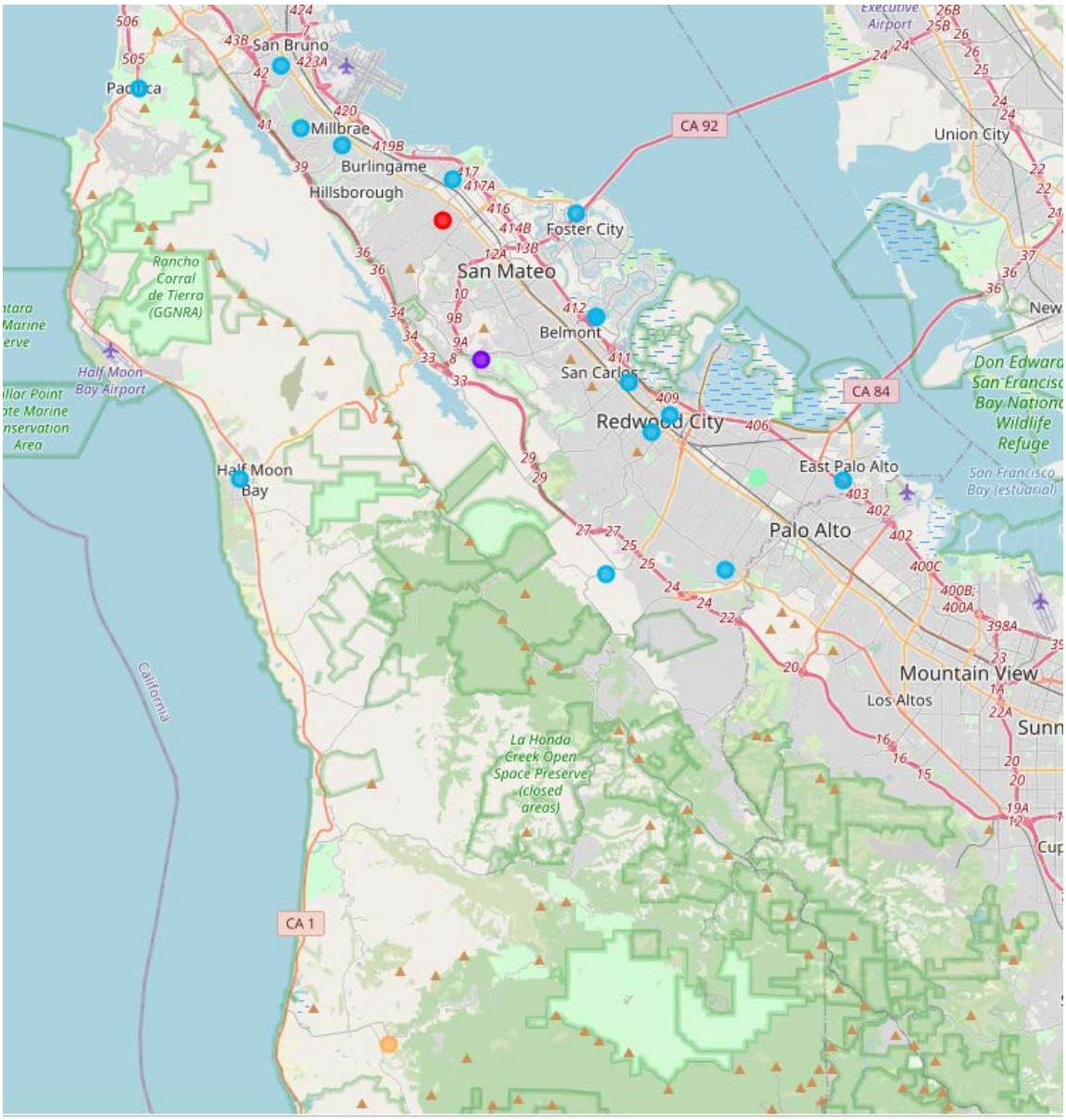
District	Venues
Bayshore Elementary	20
Belmont-Redwood Shores Elementary	3
Brisbane Elementary	8
Burlingame Elementary	24
Cabrillo Unified	21
Hillsborough City Elementary	2
Jefferson Elementary	13
Jefferson Union High	8
La Honda-Pescadero Unified	1
Las Lomitas Elementary	12
Menlo Park City Elementary	1
Millbrae Elementary	4
Pacifica	9
Ravenswood City Elementary	17
Redwood City Elementary	84
San Bruno Park Elementary	26
San Carlos Elementary	39
San Mateo County Office of Education	16
San Mateo County ROP	16
San Mateo Union High	20
San Mateo-Foster City	15
Sequoia Union High	9
South San Francisco Unified	21
Woodside Elementary	5

In order to analyze each neighborhood, the venue categories were dummy coded and the new data frame size is (394, 140). The rows were grouped by district and by taking the mean of the frequency of occurrence of each category. Then the new data frame size is (24, 140). The venues were sorted in descending order and the top 10 venues for each district were displayed. For instance, the top 10 venues for Bayshore Elementary are pizza place, fast food restaurant, theater, fried chicken joint, plaza, pet store, discount store, dive bar, sandwich place, and music venue.

3.2. Machine Learning: k-Means Clustering

A k-means clustering was run to cluster the districts into 5 clusters. Figure 1 is the map that visualize the resulting clusters, with each color represents a cluster.

Figure 1. School District Offices Clustering



Further examining the clusters shows the districts within each cluster.

Cluster 1: Belmont-Redwood Shores Elementary. This school district office is near scenic lookout and trail.

Cluster 2: Bayshore Elementary, Brisbane Elementary, Burlingame Elementary, Cabrillo Unified, Jefferson Elementary, Jefferson Union High, Pacifica, Las Lomitas Elementary, Millbrae Elementary, Ravenswood City Elementary, Redwood City Elementary, San Bruno Park Elementary, San Carlos Elementary, San Mateo County Office of Education, San Mateo-Foster City, San Mateo Union High, Sequoia Union High, South San Francisco Unified, Woodside Elementary, and San Mateo County ROP. These school district offices are close to restaurants and stores.

Cluster 3: Menlo Park City Elementary, which is next to an art gallery.

Cluster 4: La Honda-Pescadero Unified, which is close to a racetrack.

Cluster 5: Hillsborough City Elementary. There are some sport facilities nearby.

4. Discussion and Conclusion

Combining the results of the exploratory data analysis and the k-means clustering, it appears that if the county office needs to host a large gathering and the participants will have meals during the events, it would be best to partner with the districts in Cluster 2. It will be easier for the participants to find the food they like, especially if they cannot eat the food provided in the events due to religion, culture, and health issues. It will also be more convenient for food delivery. In addition, the shops and stores nearby can be helpful if the participants need something urgently. After reviewing the most common venues near these district offices, we can further decide which districts are most ideal for some cultural events. For example, Cabrillo Unified and Redwood City Elementary may be the best choices for community events about Mexican culture, whereas Jefferson Elementary and South San Francisco Unified may work better for the Chinese culture events, given the restaurants within the walking distance. In contrast, the leadership may consider Hillsborough City Elementary School District for sport events and Belmont-Redwood Shores Elementary School District for outdoor education.