Neighborhood Analysis of Dessert Venues in SF South Bay Area

Borirak Opasanont

April 8, 2020

IBM Data Science Specialization capstone project on Coursera

Introduction

The San Francisco South Bay Area is a lively area with rich diversity. It is home to a top university and many great tech companies. As such, the area attracts people from all culture, ethnicity, and walks of life. The area also offers rich culinary options ranging from popular chain venues to local single shops with unique offerings. While restaurants have gotten people's attention in defining a neighborhood, dessert places are often overlooked. This could be due to the omnipresence of chain coffee shops that limits our curiosity of this category. Donut, bubble tea, and frozen yogurt are some example dessert categories that are possibly available in a neighborhood. Large fraction of bubble tea shops may correlate with younger population of Asian-descent, while donut shops may be numerous in a neighborhood full of busy working-age individuals. Therefore, **dessert venues could be a good indicator of the type of neighborhoods**.

Objective

This study serves to **provide high-level insights into the neighborhoods** within the South Bay Area **in terms of dessert venue categories** and how they may be **related to the demography** of each neighborhood. This information could help someone looking to move into the area in **deciding on a neighborhood to settle in** based on their sweet-tooth preference. For an investor, this analysis may help in **narrowing the scope of their business decision** on which type of dessert venue to pursue, and could be very powerful when coupled with further demand-supply analysis. Finally, this data could be used **for comparison with other cities** in forming a better understanding of demographic preferences of dessert venues.

Approach

In our analysis, we will use **zip codes as a proxy of neighborhoods** in the South Bay Area which will be scoped to only the densely populated region centered on San Jose. We will use **machine learning to cluster similar neighborhoods** based on their dessert venue profile, and analyze the clusters' similarity and differences. Finally, we will **build a model based on the demographic profile** of each neighborhood with the aim of explaining or correlating the profiles with the neighborhood clusters.

Data Sources

The following datasets will be needed:

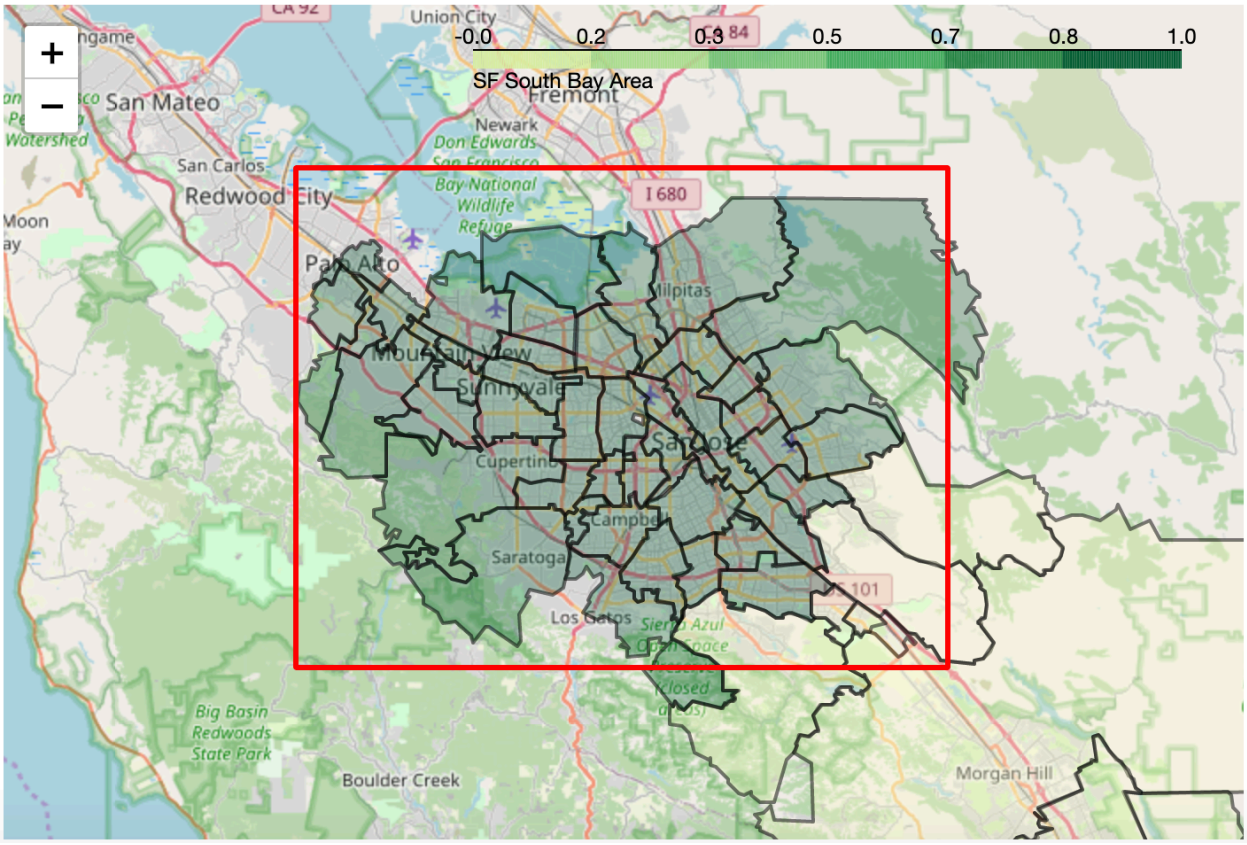
1. Zip code geographical coordinates, and possibly their boundary data for visualization
2. Dessert venues and their category and zip code
3. Demographic data by zip code

Zip Code Data

The Santa Clara County, for which the SF South Bay Area resides, Public Health Department (SCCPHD) provides open source data on the **zip code geo-boundary as a GeoJSON file**,[i] which we will use for visualizing the neighborhoods. The neighborhoods **center coordinates are obtained from OpenDataSoft**[ii] and are useful for displaying the clusters. I utilize Google Maps to refine the area of interest, and narrowed down to **45 zip codes of interest**.

Zip	City	Latitude	Longitude	Zip	City	Latitude	Longitude
94022	Los Altos	37.37714	-122.124	**95111**	San Jose	37.28404	-121.827
94024	Los Altos	37.35374	-122.087	**95112**	San Jose	37.34854	-121.886
94040	Mountain View	37.38021	-122.088	**95113**	San Jose	37.33394	-121.892
94041	Mountain View	37.38949	-122.078	**95116**	San Jose	37.34964	-121.854
94043	Mountain View	37.40679	-122.075	**95117**	San Jose	37.31104	-121.962
94085	Sunnyvale	37.38894	-122.018	**95118**	San Jose	37.25764	-121.891
94086	Sunnyvale	37.37834	-122.024	**95121**	San Jose	37.30571	-121.811
94087	Sunnyvale	37.35009	-122.036	**95122**	San Jose	37.32964	-121.834
94089	Sunnyvale	37.40629	-122.008	**95123**	San Jose	37.24443	-121.832
94301	Palo Alto	37.44369	-122.151	**95124**	San Jose	37.25674	-121.923
94304	Palo Alto	37.39782	-122.166	**95125**	San Jose	37.29509	-121.896
94305	Stanford	37.42704	-122.165	**95126**	San Jose	37.32634	-121.918
94306	Palo Alto	37.41939	-122.133	**95127**	San Jose	37.36947	-121.821
95002	Alviso	37.42744	-121.975	**95128**	San Jose	37.31698	-121.936
95008	Campbell	37.27884	-121.954	**95129**	San Jose	37.30774	-122
95014	Cupertino	37.31791	-122.048	**95130**	San Jose	37.28964	-121.983
95032	Los Gatos	37.24119	-121.953	**95131**	San Jose	37.38631	-121.89
95035	Milpitas	37.43645	-121.894	**95132**	San Jose	37.40599	-121.848
95050	Santa Clara	37.34779	-121.951	**95133**	San Jose	37.37354	-121.858
95051	Santa Clara	37.34624	-121.985	**95134**	San Jose	37.41254	-121.945
95054	Santa Clara	37.39324	-121.961	**95136**	San Jose	37.26934	-121.849
95070	Saratoga	37.27054	-122.023	**95148**	San Jose	37.3305	-121.791
95110	San Jose	37.33555	-121.899				

When plotted on a map, the **neighborhoods of interest are shown in green along with a bounding box for venue search in red**. The bounding box is decided to be as small as necessary to conserve search efforts, which left out parts of a neighborhood to the east and the south. These parts are in the mountainous area with very low probability of a dessert venue.



Venues Data

Foursquare provides a convenient access to its rich venues database through an open API with a free developer personal account.[iii] We will use **Foursquare's dessert venues data** for their category and zip code, obtained using their search API call.[iv] A formatted sample of data looks like this:

	id	Name	Category	City	Zip	Latitude	Longitude
0	54cd6c94498e0db58a66f52b	Tous Les Jours	Bakery	Fremont	nan	37.489026	-121.929387
1	527f0187498e4f765ccdeb74	La More cafe, Inc.	Bakery	Fremont	94539	37.492503	-121.930102
2	4c9fb0262fb1a1432e1ef640	Sogo Bakery	Bakery	Fremont	94539	37.493435	-121.930059
3	4e9ca981b803b7506d57ceb2	Baker's Goods	Bakery	NaN	nan	37.501542	-121.929292
4	5934469293bd6364b4f23b41	Sheng Kee Bakery	Bakery	Fremont	94539	37.493475	-121.930135

Note: Foursquare imposes certain limits on their API calls. First, the search call limits maximum of 30 returned venues (not 50 as advertised, as we have experimentally found). This forces us

to **sub-divide the search area into grids**, and perform an API call for each grid to obtain all venues. Second, the free personal account allows up to 99,500 regular calls per day. While this is seemingly an abundant number, we should be mindful not to exceed this number and potentially miss out on some data.

Foursquare conveniently provides a list of venue categories,[v] of which we **consider the following as dessert places**:

Category	Category ID
Bakery	4bf58dd8d48988d16a941735
Coffee Shop	4bf58dd8d48988d1e0931735
Dessert Shop	4bf58dd8d48988d1d0941735
Bubble Tea Shop	52e81612bcbc57f1066b7a0c
Bagel Shop	4bf58dd8d48988d179941735
Café	4bf58dd8d48988d16d941735
Donut Shop	4bf58dd8d48988d148941735
Tea Room	4bf58dd8d48988d1dc931735

In making our API calls, we will **iterate through these categories and geographical grids** for obtaining our venues data. The data will be filtered accordingly to remove duplicate venues, missing zip codes, and to scope it within the zip codes of interest.

Demographic data

The SCCPHD also provides **demographic data per zip code**, available as a csv file.[vi] The data, updated on 6/29/2018, contains percentage of race/ethnicity, age distribution, as well as household information such as average size, whether it has children, or single parent. A sample data is shown below.

	Zip	African-American	Asian-Pacific Islander	Latino	White	Foreign Born	Other Language	Household with Children	Average Household Size	Age 18-24	Age 25-34	Age 35-44	Age 45-54	Age 55-64	Age 65 Plus	Age 0-17
0	95138	3	47	17	29	39	53.0	53.0	3.35	6	12	19	17	10	6	30
1	95120	1	30	7	59	31	38.0	42.0	3.01	6	5	14	20	14	15	27
2	95136	5	28	28	35	32	45.0	38.0	2.79	8	18	16	13	10	10	24
3	95123	4	19	30	43	27	42.0	38.0	2.96	9	15	16	15	10	10	25
4	94304	2	28	5	62	33	37.0	20.0	1.82	5	20	17	9	7	27	16

We will filter the data down to the zip codes of interest and remove any missing data before using it for our demographic modeling.

Methodology

To be added later

Results

To be added later

Discussion

To be added later

Conclusion

To be added later

References

[i] https://data-sccphd.opendata.arcgis.com/datasets/demographic-statistics-zip-code, updated on 6/29/2020, accessed 4/8/2020.
[ii] https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/export/?refine.state=CA, accessed 4/8/2020.
[iii] https://developer.foursquare.com/
[iv] https://developer.foursquare.com/docs/api-reference/venues/search/, accessed 4/8/2020.
[v] https://developer.foursquare.com/docs/build-with-foursquare/categories/, accessed 4/8/2020.
[vi] https://data-sccphd.opendata.arcgis.com/datasets/demographic-statistics-zip-code, updated on 6/29/2020, accessed 4/8/2020.