Neighborhood Analysis of Dessert Venues in SF South Bay Area

Borirak Opasanont

April 8, 2020

This report is a part of a submission for IBM Data Science Specialization capstone project on Coursera along with a Jupyter Notebook and a presentation, available on my Github https://github.com/boriopas/ds-capstone-project.

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:

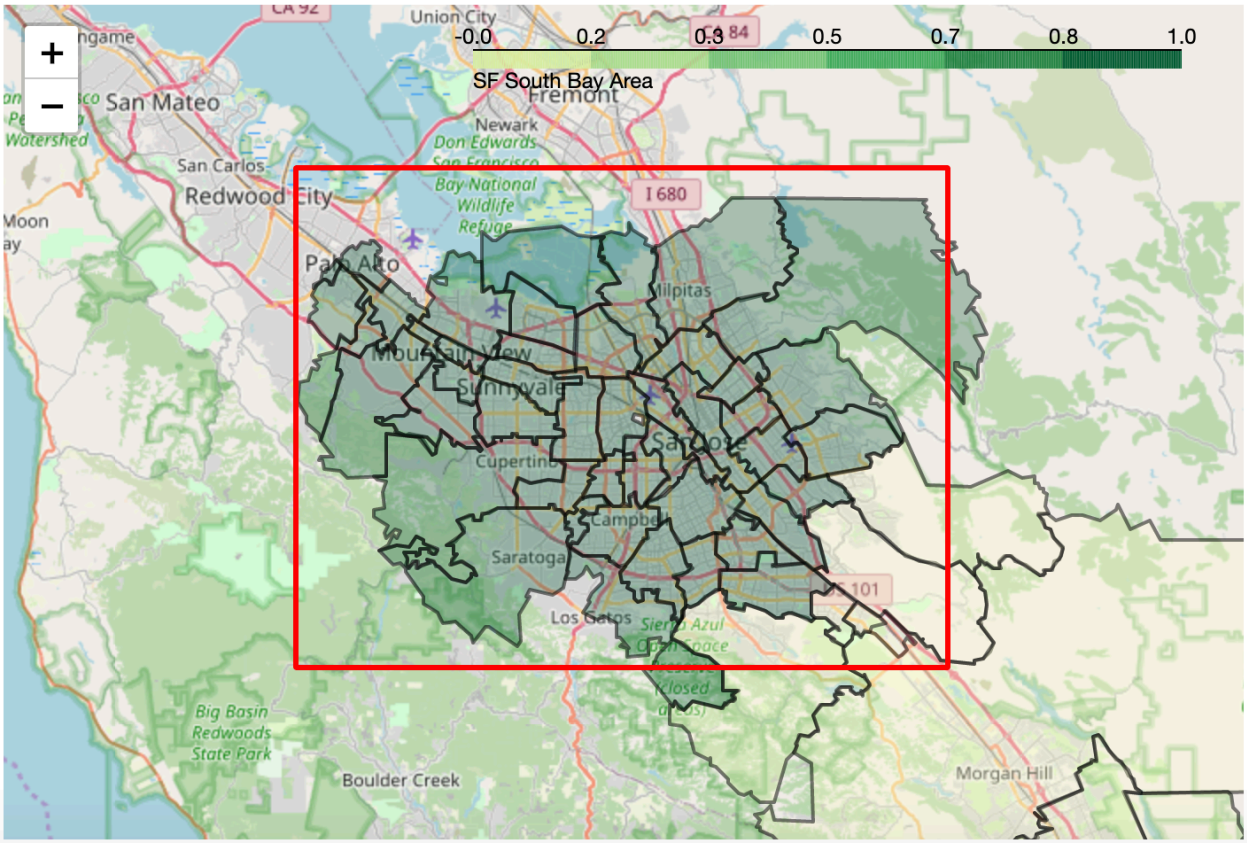
1. Zip code geographical coordinates, and possibly their boundary data for visualization
2. Dessert venues and their category and zip code
3. Demographic data by zip code

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**,[1] which we will use for visualizing the neighborhoods. The neighborhoods **center coordinates are obtained from OpenDataSoft**[2] 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.[3] We will use **Foursquare's dessert venues data** for their category and zip code, obtained using their search API call.[4] 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,[5] 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.[6] 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.

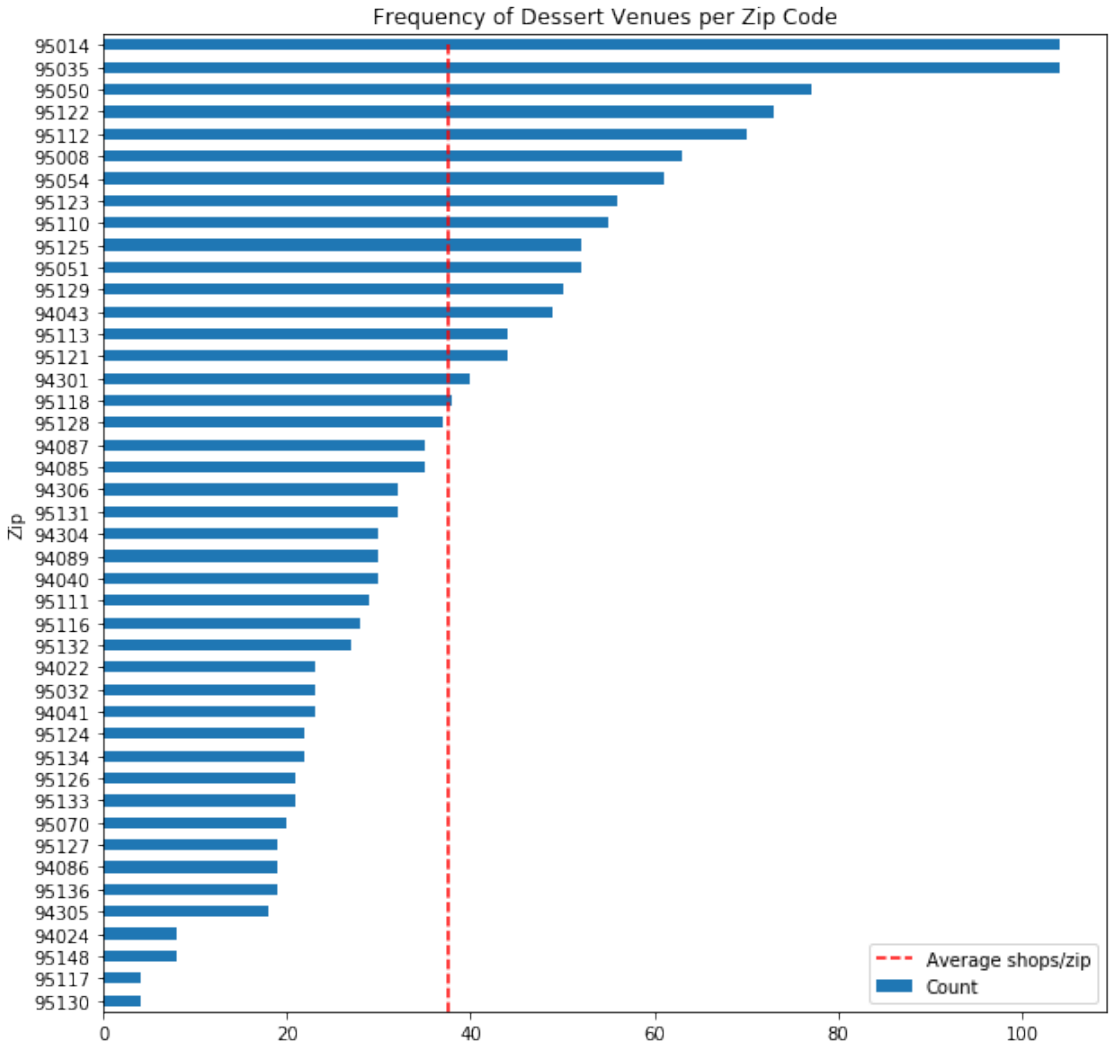
	Zip	African-American	Asian-Pacific Islander	Latino	White	Foreign Born	Other Language	Household with Children	Average Household Size	Age 18-24	Age 25-34	Age 35-44	Age 45-54	Age 55-64	Age 65 Plus	Age 0-17
0	95138	3	47	17	29	39	53.0	53.0	3.35	6	12	19	17	10	6	30
1	95120	1	30	7	59	31	38.0	42.0	3.01	6	5	14	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

Exploratory Analysis of Zip Codes

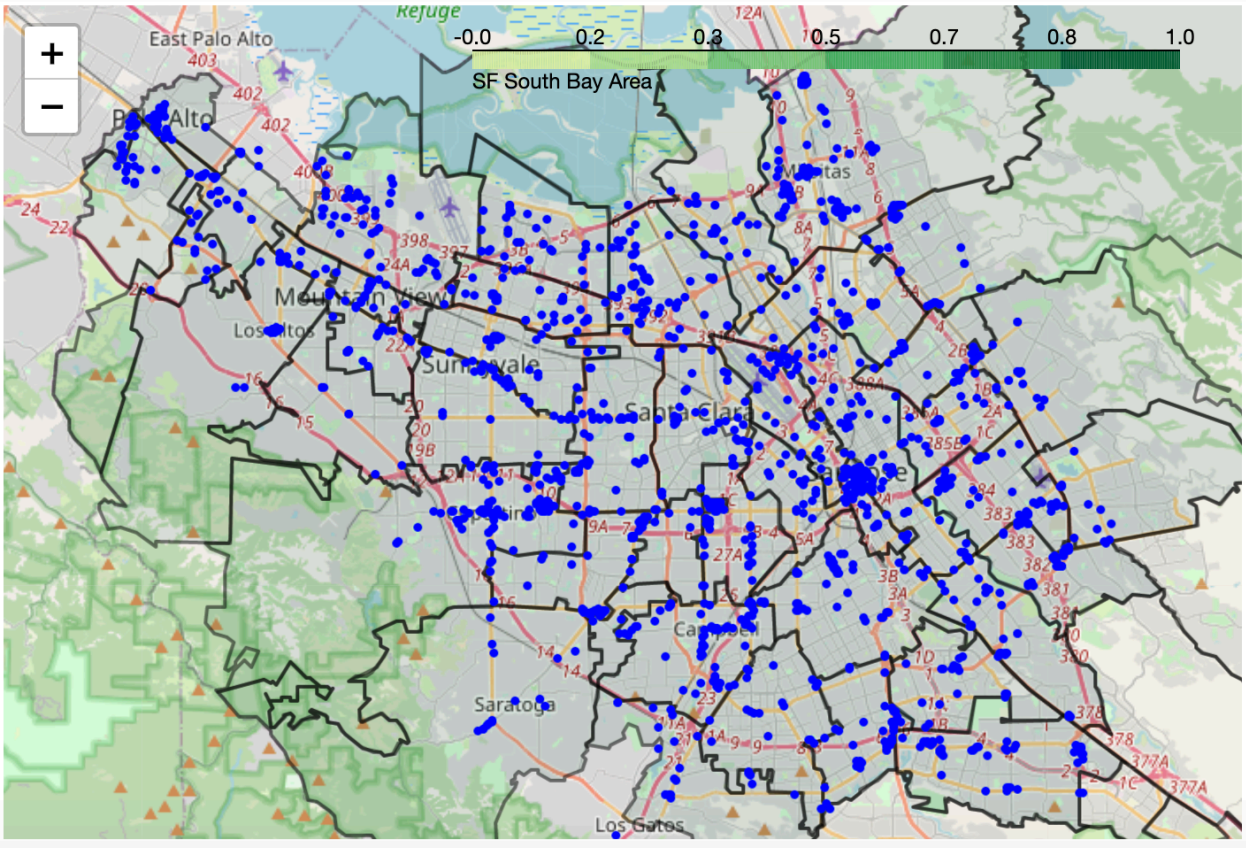
First, we check whether all zip codes contain dessert venue. If a zip code has null venue data, it could be problematic for future data processing. Unfortunately, Alviso with zip code 95002 has no dessert venue and has to be dropped from our zip codes of interest. The other 44 zip codes have frequency of dessert venue distribution as shown below.



Two top zip codes with over 104 dessert venues are 95014 (Cupertino) and 95035 (Milpitas). These two neighborhoods are also known to be densely populated.[7] The bottom 2 are 95117 and 95130 (both in San Jose) with only 4 shops each. These zip codes have moderate number of

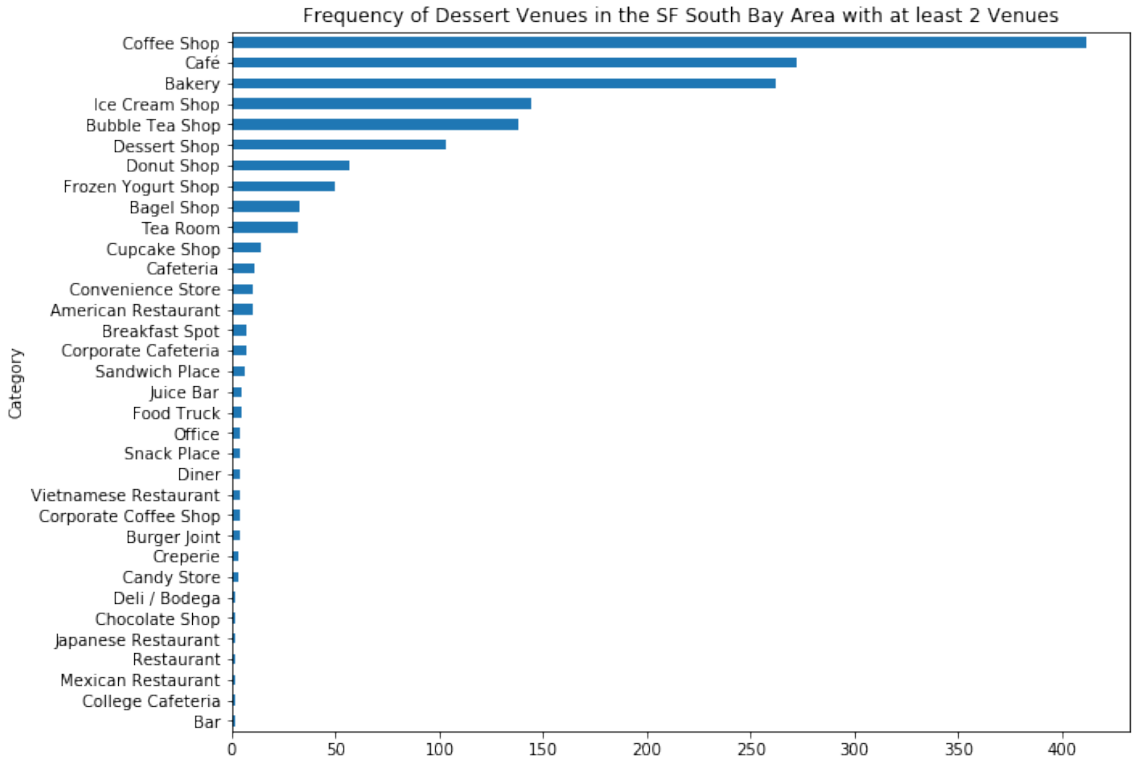
population, so the reason for very small frequency of dessert venues is not known. The average number of shops per zip code is 37.5.

When all shop locations are overlayed onto the map (as blue dots below), we noticed an interesting pattern where shops are more frequent along major roads and highways. Large number of shops are in Cupertino (bottom left) and Milpitas (top right) as previously discussed.



High Level Statistics of Venue Category

Next, we look at dessert venue categories as shown in the bar chart below. As we have casually suspected, coffee shops are the most numerous at 412 shops total, followed by café and bakery. Ice cream, bubble tea, frozen yogurt and donut shops are next, along with the generic category of dessert shop. Overall, there are 63 unique categories, but only 34 categories with at least 2 venues are shown on the chart.



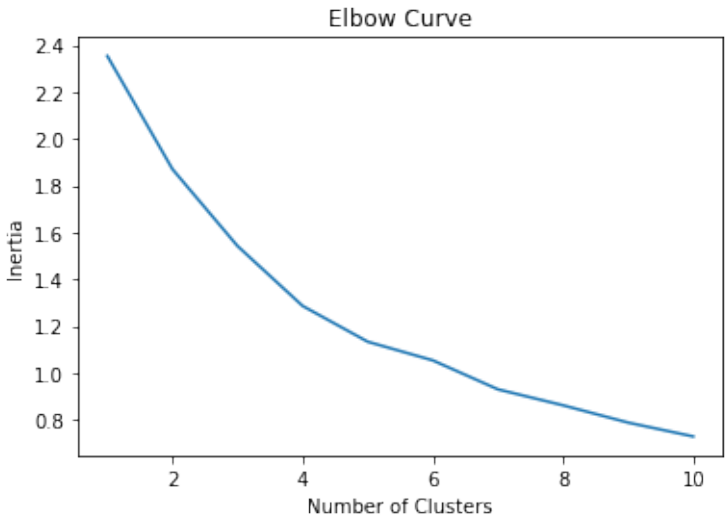
Category Clustering

One of our goals is to cluster the neighborhoods by dessert venue category. To do so, **one-hot encoding is done**, creating a table with 44 rows of zip codes, and 63 columns of venue category plus one column for zip code. This table is used to sum and analyze the top 10 most frequent categories in each zip code. The goal of this exercise is **to decide on the number of top categories to be used** in clustering. This is important because if too many top levels are used, the zip codes with less venues will eventually have insufficient data, leading to null and possibly incorrect category binding.

```
----95124----
                venue  freq
0         Coffee Shop  0.36
1          Donut Shop  0.18
2              Bakery  0.18
3               Café  0.14
4  Frozen Yogurt Shop  0.09
5     Bubble Tea Shop  0.05
6       Grocery Store  0.00
7      Ice Cream Shop  0.00
8    Indoor Play Area  0.00
9     Other Nightlife  0.00
```

Take for example zip code 95124 shown above, where the top 6 categories are populated (rows 0-5), but 7th to 10th (rows 6-9) returned zero frequency with automatically suggested categories. If used without care, this code can provide misleading results. After looking through results of all 44 zip codes, we found 2 zip codes (95117 and 95130) which have only top 3 categories. These were the ones we identified earlier as having only 4 shops in each zip code. There are also other 11 zip codes that have at least top 5 but not all 10 top categories. Therefore, we decide to **use top 5 categories as a basis for clustering**, and keep an eye on the two zip codes.

We used the unsupervised **k-means as our technique for clustering**, using the top 5 categories from each zip code. In order to determine the best number of clusters, the **elbow method** is adopted using inertia as the metric. From the elbow plot shown below, we decided on **k=5 as the appropriate number of clusters**.



We will discuss the results of clustering with demographic data in the Results and Discussion section.

Further Data Cleaning of Demographic Data

There are 16 features of demographic data that we can explore, consisting of 7 age ranges, 5 race/ethnicity sub-populations, percentage of population who speaks other language at home, household with children, household with single parent, and average household size. We found that the household with single parent data has 4 missing values, and has to be dropped. We also found zip code 95113 to be missing the other language data and the household with children data, and also has to be dropped. **At the end, we have 43 zip codes and 15 features to use for modeling**.

By looking at the order of magnitude of data, we deemed that it was **unnecessary to scale the data**. This is because the data such as age and ethnicity were already expressed as percentage of population with no decimals. Also, the Average Household Size which varies within 1.82 to

4.45. Therefore, the data set values are between 0 to 100, which is reasonable order of magnitude for mathematical iterations. Hence, we used the data without scaling.

Modeling Neighborhood Clusters with Decision Tree

Our other goal is to see if the neighborhood clustering can be explained by demographic make-up. Decision tree is a supervised learning technique that can be used to build a model for labelled data. In our case, the **cluster labels are the classes, and we can use the demographic features for classifying the data**.

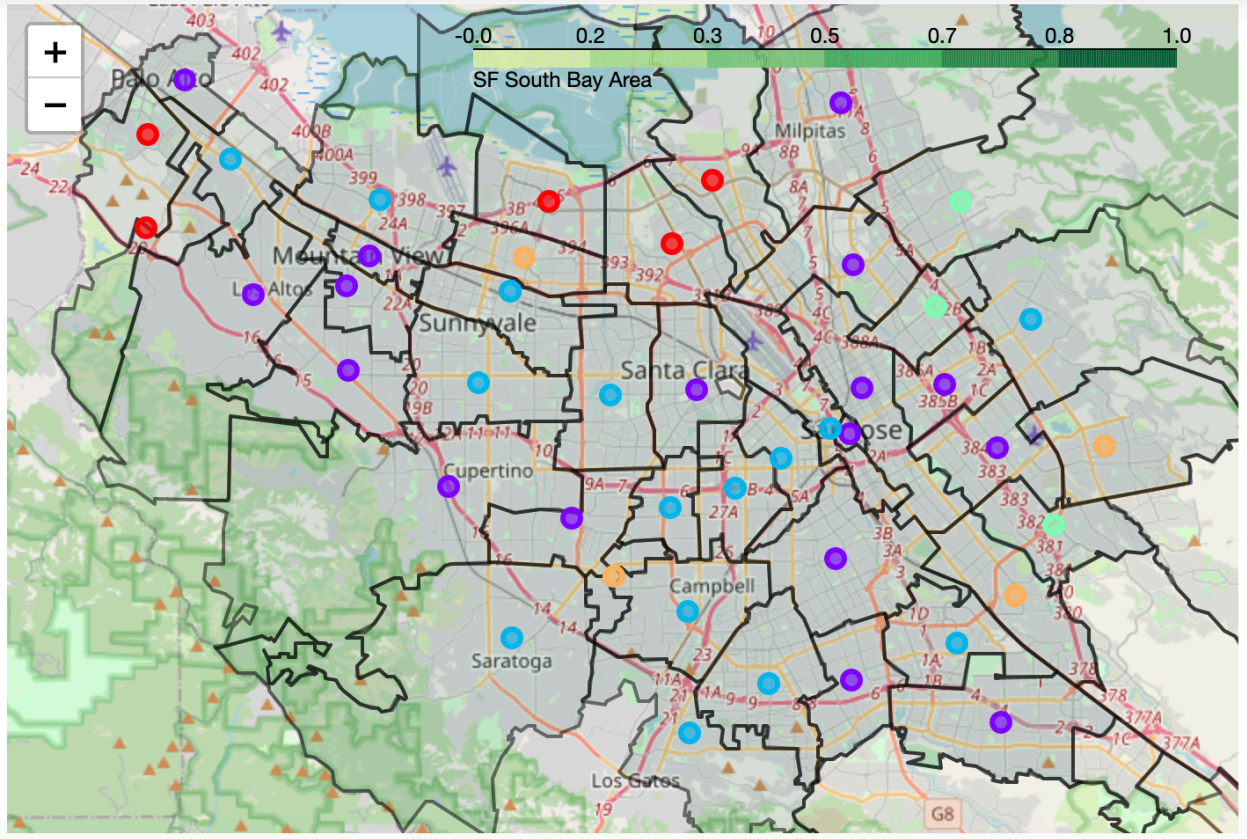
For initial model optimization, we chose to **split the dataset into training and testing sets**, with test set taking 15% of the dataset. This results in 7 rows of data being assigned for testing, and 36 other rows sufficiently large for training the model. We use **entropy as the criterion**, as we found no significant difference when using the Gini Index. This finding is similarly discussed in a helpful blog about tuning decision trees.[8]

Next, we consider the relevant max_depth value. The upper bound of max_depth is initially set at 15, as there are 15 features. If the tree is allowed to fit beyond this value, there is a high chance that the model will use features more than once. It does not mean that the model cannot use a feature more than once, but it is likely that the model would overfit the data. Results from performing a number of initial fits with different random_state values show no deep than depth of 6. Hence, an **upper bound for max_depth of 6** will be used. The **lower bound of max_depth should be 3**. This is because at depth of 2, the model can only provide maximum of $2^2 = 4$ leaves which is less than our number of classes of 5.

With this range of max_depth in mind, we perform **grid search with cross validation** in order to find best max_depth. In choosing the number of cross validation folds (cv), we initially set 7, which would result in 14% of data being assign to each test set in a fold. However, the sk-learn library gave a warning that the number of test sample should be larger than the number of class times three, which is 5 x 3 = 15, representing 34% of data. Therefore, we set cv = 3, i.e. **3 folds validation of the model**. Results from a few attempts shows that the **max_depth with highest mean_test_score is often 5**. Note that the mean_test_score hovers around 0.35 to 0.57. We fit the model again using test set made from 33% of the data, and max_depth = 5 to generate our final model.

Results and Discussion

In this section we discuss results from k-means clustering and the demographic modeling from decision tree technique. The **five clusters are represented by five different colors**, as shown on the map below as well as the series of bar charts for top dessert venue categories in the following page.



Clusters can be described as follows:

- **Cluster 0 (Red) – "Offices and College"** – Consists of area around Stanford University on upper left corner of the map, and neighborhoods near the bay in North San Jose. This cluster has large proportion of **coffee shop and cafe**, which makes sense for **busy office workers and students**. The presence of cafeteria and college cafeteria also agrees well with our location analysis. There are 5 zip codes in this this cluster.

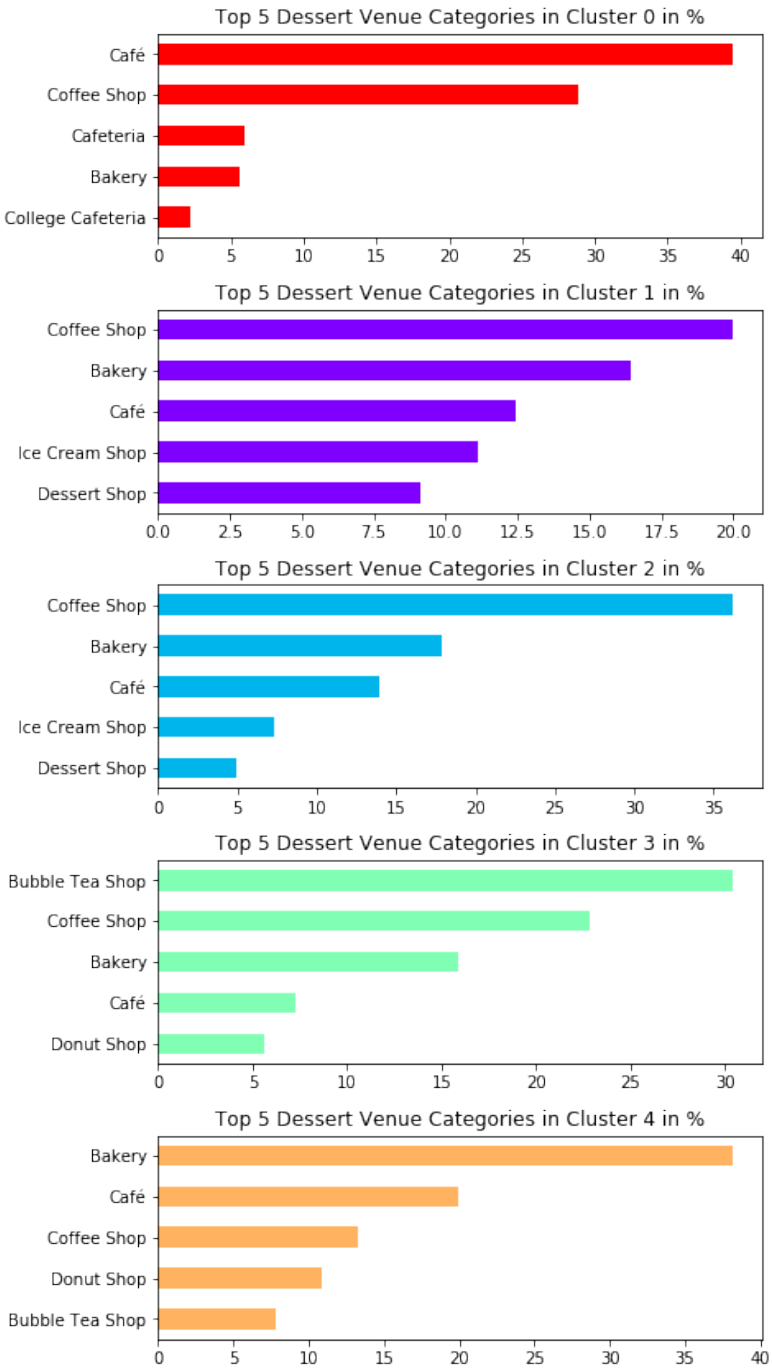
- **Cluster 1 (Purple) – "Mixed Use"** – Mostly offices and industrial as well as residential neighborhoods scattered within the South Bay Area. This cluster has a broad variety of dessert venues with coffee shop taking the top spot at only about 20%. This cluster is probably best explained as a **mix bag of dessert venues as well as mixed use areas**. This is the most numerous type of cluster with 16 zip codes.

- **Cluster 2 (Blue) – "Offices and Residential"** – Residential and businesses scattered largely towards South San Jose. This cluster has the same top five as Cluster 1, but with **stronger proportion of coffee shop**. This could be due to the cluster having **larger fraction of offices, particularly in Sunnyvale and Santa Clara area**. There are 15 zip codes in this cluster.

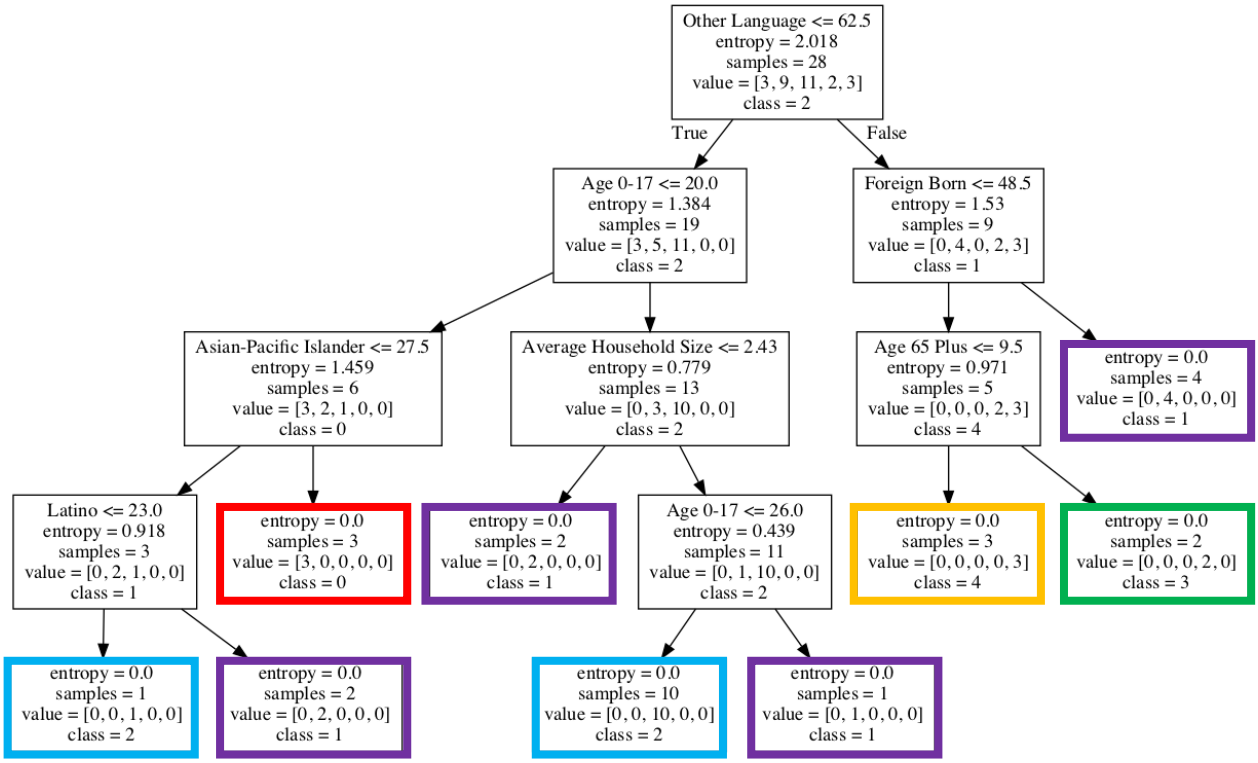
- **Cluster 3 (Green) – "Residential"** – Residential neighborhoods on the east San Jose side. Interestingly, **bubble tea shop takes top spot** in this cluster. While bubble tea is

largely enjoyed by many, it is often associated with people of Asian-descent, as it is originated from Taiwan.[9] It would be interesting to see the demographic profile of this cluster. There are only 3 zip codes in this cluster.

- **Cluster 4 (Orange) – "Bakery Towns" – Bakery takes up 40% of dessert venues in this cluster of small neighborhoods** scattered in the South Bay Area. It would also be interesting to see the demographic correlation with this group. There are 4 zip codes in this highly unique cluster.



With demographic data, the **best fitted decision tree model has depth of 4**, instead of the specified max depth of 5, as shown in figure below. The tree uses 7 out of 15 features, with Age 0-17 being a repeated node once.



In terms of demography, the clusters can be described as follows:

- **Cluster 0 (Red) – "Offices and College"** – Consists of larger population of English as a primary language, lower population under the age of 17, and lower Asian-Pacific Islander population. In other words, this area makes up of **larger fraction of English-speaking adults**. This description ties well with the cluster being a primarily a college and office area with **coffee shops, cafes and cafeteria**.

- **Cluster 1 (Purple) – "Mixed Use"** – there are **4 paths to get to this label**, and there is no clear logic. From the rightmost branch, a neighborhood may reach this label by having larger proportion of Non-English as primary language as well as larger foreign-born population. Or from the leftmost branch, by having smaller percentage of Other Language population, less minor-age population, less Asian-Pacific but more Latino population than other areas. **This cluster is still best described as "mixed" even from demographic perspective**.

- **Cluster 2 (Blue) – "Offices and Residential"** – There are two ways to reach this label and both require **smaller proportion of Other Language as primary language, as well as smaller minor-age population**. However, one requires Average Household Size of larger than 2.43, while the other requires smaller population of Asian-Pacific Islander as well as Latino. The common demographic features fit well with the office and residential description of these neighborhoods, given that businesses often use English as a primary language.

- **Cluster 3 (Green) – "Residential"** – To reach this label, a neighborhood needs to have larger fraction of Other Language, but smaller fraction of Foreign-Born population, which may indicate post-first-generation immigrants. However, the neighborhood needs to have larger population of age > 65. Perhaps, there is **a community, young and old, of many generations of immigrants in this cluster of neighborhoods, where people really enjoy bubble tea**. Interestingly, percentage of Asian-Pacific Islander population is not a feature that defines this cluster label, opposite to what was initially hypothesized.

- **Cluster 4 (Orange) – "Bakery Towns"** – Almost similar to Cluster 3, this cluster is reached when a neighborhood has larger fraction of Other Language, and smaller fraction of Foreign-Born population. This unique cluster of small-area neighborhoods rich with bakeries is **quite likely similar to Cluster 3 in demography, except with smaller population of age 65 plus**.

It should be noted that the model has an accuracy score of 0.467, which is one of the best we could achieve out of multiple random attempts. The low and fluctuating accuracy score highlights the fact that we have very limited data (43) for fitting a model with 5 class labels. Nevertheless, there are 3 out of 5 clusters that are fairly well explained by their demographic features, namely Clusters 0, 1, and 2.

Conclusion

Dessert venue categories can provide interesting insights into neighborhoods in the SF South Bay Area. Five clusters are derived based on their top dessert venue categories and a demographic profile model is presented with accuracy of 0.467. The information in this study could serve as an initial guide for someone who prioritizes dessert options as the plan to move into the South Bay Area, or an investor looking for sweet ideas. The model's accuracy is likely to be further improved if the more data is used, possibly from the whole Bay Area or California.

References

[1] https://data-sccphd.opendata.arcgis.com/datasets/demographic-statistics-zip-code, updated on 6/29/2020, accessed 4/8/2020.
[2] https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/export/?refine.state=CA, accessed 4/8/2020.
[3] https://developer.foursquare.com/
[4] https://developer.foursquare.com/docs/api-reference/venues/search/, accessed 4/8/2020.
[5] https://developer.foursquare.com/docs/build-with-foursquare/categories/, accessed 4/8/2020.
[6] https://data-sccphd.opendata.arcgis.com/datasets/demographic-statistics-zip-code, updated on 6/29/2020, accessed 4/8/2020.
[7] https://data-sccphd.opendata.arcgis.com/datasets/demographic-statistics-zip-code
[8] https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680, accessed 4/8/2020.
[9] https://en.wikipedia.org/wiki/Bubble_tea, accessed 4/8/2020.