Initial Location Analysis in San Francisco for Restaurant La Tierra De Los Tacos

Daniel Garcia Rodriguez

January 6, 2020

I. Introduction

1.1 Background and Business Problem

My cousin is in the late stages of developing his business plan for his restaurant, Restaurante La Tierra De Tacos (which I will be referring to as 'the company'). After careful planning and consideration, the company has chosen San Francisco (SF) as the location for their business. For the company, SF is a good choice given the city's relatively large population, it's vibrant urban setting, strong transportation infrastructure, among other factors. However, the company is having trouble choosing the best location for their restaurant within San Francisco. The company considers this a vital decision as it may dictate their success (or failure).

The company needs assistance in filtering out the neighborhoods that don't meet their specific requirements in order to conduct detailed location analyses into only a select list of neighborhoods in SF. So, to help find these neighborhoods, I chose the Census tracts designated to the SF county as it most closely resembles the definition of a 'neighborhood' for the company. After careful talk with the company, the following list of requirements that neighborhood must were created:

1. Contain a population of 5,000 + (increase opportunity of traffic and visibility)
2. Near at least one college and surrounded by at least three other types of schools (partnerships with schools and other organizations, and to provide catering/delivery options)
3. Rent at or below the city's average -$3,000
4. Median income of $50,000 or more (safeguard for tough financial times)
5. With minimal competition (mexican restaurants) and similar venues (ie. restaurants) nearby

It is the aim and scope of this report to identify the neighborhoods in SF that meet the aforementioned requirements

1.2 Interest

The information, analyses, in this report will be beneficial to small business owner looking to start a business in San Francisco, especially Mexican restaurants. In addition, it will also be informative and of use to entities looking to get some insights into the city of San Francisco.

II. Data Preprocessing

2.1 Data Requirements

In order to best answer the business question -- finding the neighborhoods in SF that meet aforementioned features -- I have chosen to utilize data from the U.S. Census Bureau and Data SF. Through their several surveys, the U.S. Census Bureau provides one of the best estimates of the current U.S. population demographics at several geographic levels (city, tract, state, etc.).

Similar to the Census, SF's official open data portal, Data SF, collects and makes data easily accessible and open to the public. So, due to both SF's and the U.S. Census Bureau agency's reliability, and accessibility of their data, as well as methods, I have decided to use their data for my analysis. I also chose to use FourSquare API data, a location data platform that provides access to detailed data about venues (ie. nearby venues, trending venues) and user's interactions with them (ie. rating, tips). Their data will be useful in searching for nearby venues around select SF neighborhoods.

2.2 Data Sources

The following is a list of the data with links that direct you to the data used in this report. You can also access the files from my GitHub account: https://github.com/dannygarcia193/IBM_Capstone_Project_Data.

1. ACS 5 Year Estimates (2013-2017) -- Selected Characteristics of the Total and Native Populations in the United States (Census Table ID: 601)
   Filters used: Selected year 2017 only and all the census tracts within SF county
2. ACS 5 Year Estimates (2014-2018) -- Selected Housing Characteristics (Census Table ID: DP04)
   Filters used: Selected year 2018 only and all the census tracts within SF county.
3. Schools
4. Colleges (2011)
5. Analysis Neighborhoods - 2010 Census Tracts Assigned to Neighborhoods
6. Venues Data - For more information regarding the procedures used to retrieve the data please see the section 2.5 of the following link: https://colab.research.google.com/drive/1Mo5-6U3MxpAk05eU9camXflUxNbJgBkj

2.3 Data Cleaning and Feature Selection

In order to analyze and visualize my data, there were many preprocessing steps that I needed to take. For the first dataset (ACS 5 Year Estimates -- 2013-2017), I removed the irrelevant data (ie. geo id, margin of errors) and renamed the relevant columns for reliability purposes. Before preprocessing, the data there were a total of 163 columns(demographic info.) and 197 rows (number of census tracts). However, I filtered the column names by choosing only the columns that started with 'Total' as they were going to be needed for my analysis. This brought down the column count to 22. I then replaced all values in the tract column to merge and compare with other datasets. I did this by stripping all instances of: ", San Francisco County, California" in the tract columns. Lastly, I used the first and second location requirement (see intro) by retrieving only census tracts that had a minimum population of 5,000 people or more, and a median income of $50,000 or more. This reduced the total neighborhoods to 18. So, I was left with 22 columns (demographic data) and 18 rows (in relation to the census tract instances).

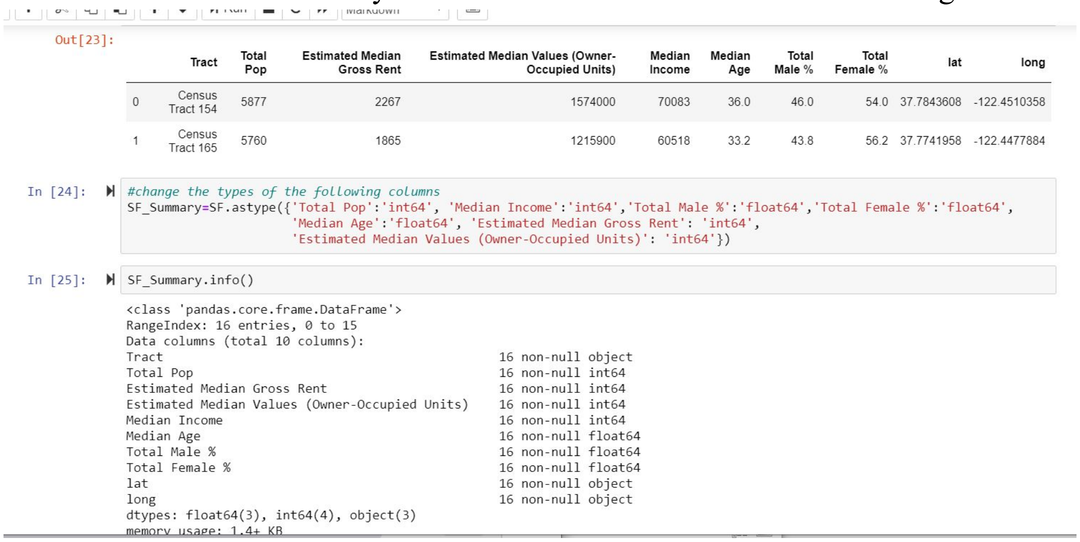
I then went on to preprocess the school and colleges data that each needed several adjustments as they were in geojson format. I made sure to get the necessary details from the geojson files by converting them into dataframes and then extracting the school name, latitude, and longitude columns of each school. This resulted in a dataframe of 445 rows(schools) with 5 columns (type, feature, name, latitude, and longitude) and 46 rows (colleges) with 5 columns (type, feature, name, latitude, and longitude), respectively. From there, since my data contained duplicates, I selected rows that contained the same name (duplicates) and dropped accordingly.

This resulted in a new shape for my dataframe with 440 rows (school instances) with 3 columns (name, latitude, longitude) and 21 rows (college instances) and 3 columns (name, latitude, longitude after dropping two columns that contain irrelevant information.

For my next dataset (ACS 5 year Estimates -- 2014-2018), to get the information for the rent and land value, I loaded the data and saw a lot of data that I had to clean. Before preprocessing my dataframe contained 197 rows (for each census tract) and 574 columns (attributes). I first selected the columns that contained the keywords 'rent' and 'value' which resulted in 124 column names. From that list, I chose the columns that was the total in dollars in regards to the gross rent and median house values. Next, I filtered out the columns that were not in the neighborhood column of the first dataset and that didn't meet the third business requirement rent less than $3,000). Lastly, This gave me a dataframe with 16 rows and 3 columns (median gross rent, median home values, and tract number).

As for the neighborhood boundaries json file, there were a few adjustments that I had to make. I first converted the json file into a dataframe which allowed me to filter based on the relevant census tracts. Once, I retrieved the relevant tracts (16 in total), I added the latitudes and longitudes as separate columns in the main SF dataframe. I then also had to create another json style object to retrieve the select multipolygon coordinates of the 16 tracts which were retrieved by accessing the original dataframe (from first sentence). After these adjustments, I was left with a new json file that contained the boundary coordinates and point coordinates of the selected tracts as well as an updated version of the previous main SF dataframe.

Now having 24 columns, I decided to drop 2 columns that contained similar information regarding each census tract. From there, I split the main SF dataframe into three parts (age, race, and summary). I dropped the age and race split data frames since the company decided that it was not going to be relevant for this analysis. With that, I merged the rent dataframe with the main SF dataframe and converted certain column items into integers and float to be able to visualize and analyze them later. After doing some checks, I concluded that there were no missing or null values in the summary data frame. Below is a view of the resulting dataframe.



Out[23]:

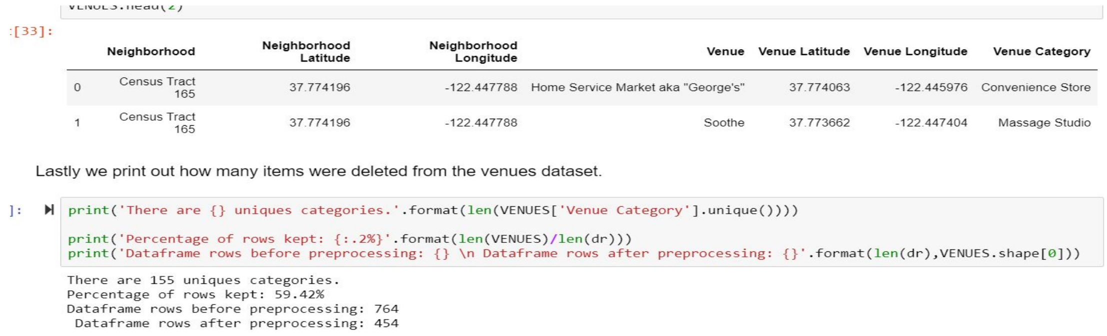
	Tract	Total Pop	Estimated Median Gross Rent	Estimated Median Values (Owner-Occupied Units)	Median Income	Median Age	Total Male %	Total Female %	lat	long
0	Census Tract 154	5877	2267	1574000	70083	36.0	46.0	54.0	37.7843608	-122.4510358
1	Census Tract 165	5760	1865	1215900	60518	33.2	43.8	56.2	37.7741958	-122.4477884

In [24]:
```
#change the types of the following columns
SF_Summary=SF.astype({'Total Pop':'int64', 'Median Income':'int64','Total Male %':'float64','Total Female %':'float64',
                      'Median Age':'float64', 'Estimated Median Gross Rent': 'int64',
                      'Estimated Median Values (Owner-Occupied Units)': 'int64'})
```

In [25]:
```
SF_Summary.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16 entries, 0 to 15
Data columns (total 10 columns):
Tract                                             16 non-null object
Total Pop                                         16 non-null int64
Estimated Median Gross Rent                       16 non-null int64
Estimated Median Values (Owner-Occupied Units)    16 non-null int64
Median Income                                     16 non-null int64
Median Age                                        16 non-null float64
Total Male %                                      16 non-null float64
Total Female %                                    16 non-null float64
lat                                               16 non-null object
long                                              16 non-null object
dtypes: float64(3), int64(4), object(3)
memory usage: 1.4+ KB
```

Lastly, after retrieving the nearby venues located near each census tract (which was stored as a dataframe) from the FourSquare API call, I saw there were a couple of adjustments that had to be made. The first adjustment was getting rid of all the duplicates data given that some tracts were near each other and thus shared the same ‘nearby venues.' In the end, I had to drop 310 rows from the 764 rows that I had retrieved, leaving me with 454 distinct venues and 155 unique venue categories. I was able to achieve this by filtering out the venues that shared the same name and geographic coordinates.Lastly, in order to associate the venues with their respective venues that they fell under, I had to compare each venues coordinate points with each multipolygon coordinate boundaries (neighborhood geojson). If the points were in the multipolygon then they were assigned to said neighborhood. After a quick analysis no null values or duplicates were found in this dataframe.



```
VENUES.head(2)
```

:[33]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Census Tract 165	37.774196	-122.447788	Home Service Market aka "George's"	37.774063	-122.445976	Convenience Store
1	Census Tract 165	37.774196	-122.447788	Soothe	37.773662	-122.447404	Massage Studio

Lastly we print out how many items were deleted from the venues dataset.

]:

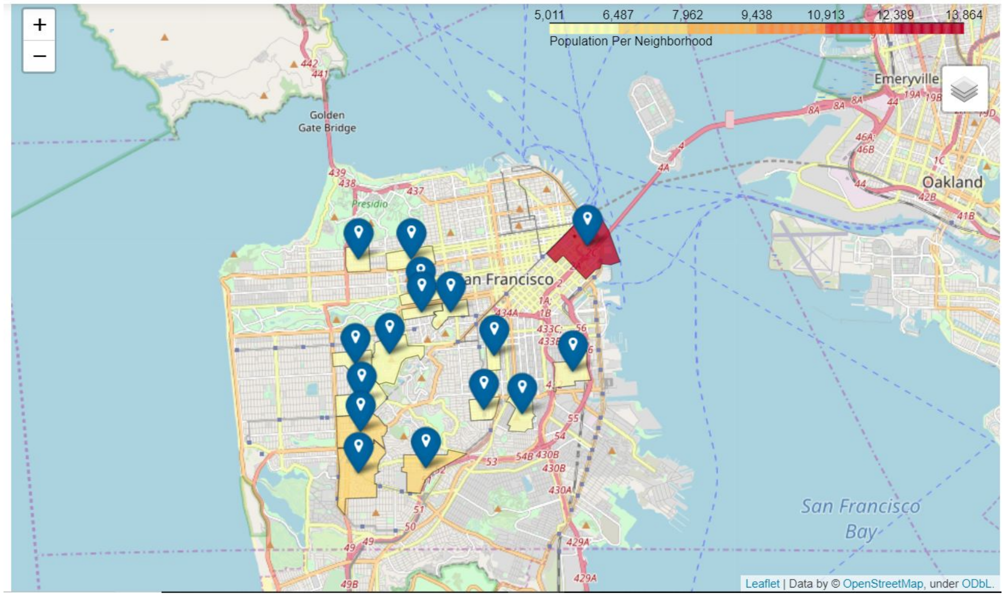
```
print('There are {} uniques categories.'.format(len(VENUES['Venue Category'].unique())))

print('Percentage of rows kept: {:.2%}'.format(len(VENUES)/len(dr)))
print('Dataframe rows before preprocessing: {} \n Dataframe rows after preprocessing: {}'.format(len(dr),VENUES.shape[0]))
```

```
There are 155 uniques categories.
Percentage of rows kept: 59.42%
Dataframe rows before preprocessing: 764
 Dataframe rows after preprocessing: 454
```

III. Exploratory and Modeling Data Analyses

3.1 Initial Neighborhood Visualization

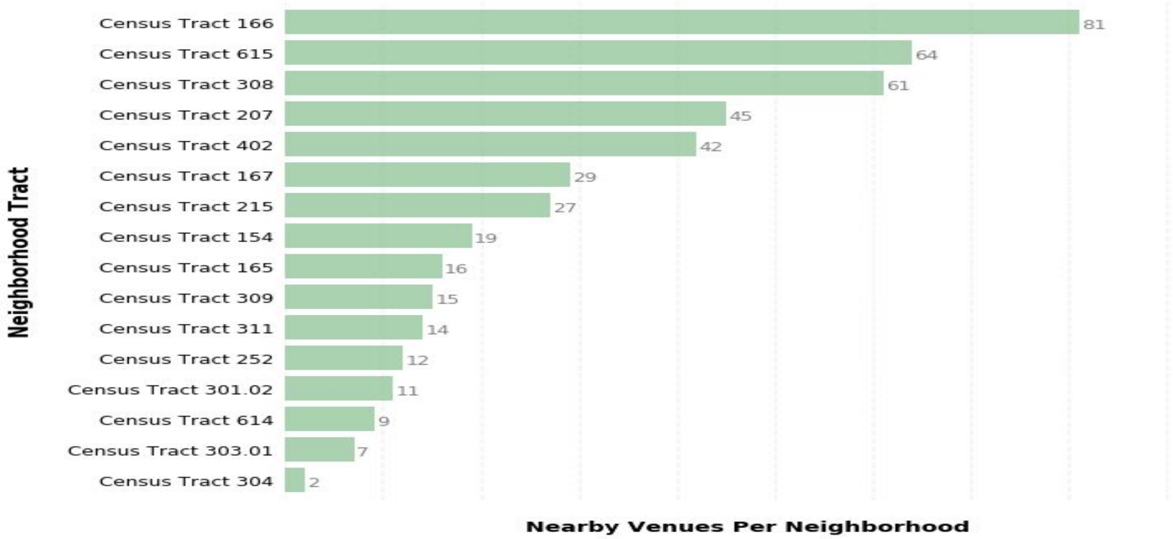
I decided to first visualize the tracts along with their respective total population and median income by using the folium library in Python. The Folium library allows for creation of interactive leaflet map using coordinate data. I took the coordinates from the main SD dataframe to visualize the boundaries of the tract. I also used the total population column as the choropleth layer for my map and added some markers with popups to both identify and show information (median income and total population) regarding each tract.



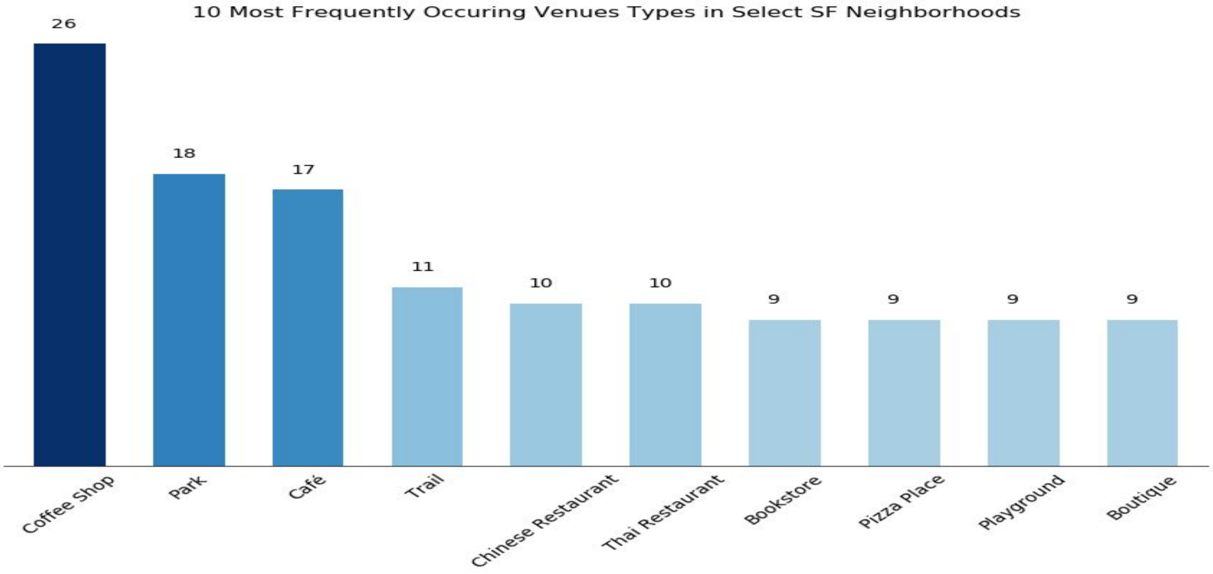
Looking at the map, you can see that most of the districts that meet the first three criterias are mostly roughly located around the center and upper east side of SF.

3.2 Foursquare Venues Analysis

To get a better understanding of the FourSquare venues data, I first grouped the data by neighborhood and created the horizontal bar chart below. Since I caped the nearby venues at 100, I expected a good chunk of them to be close to 100 given that SF is known to have many restaurants. To my surprise, only a few of the neighborhoods that matched the company's first three criterias had close to 81 nearby venues with roughly half of the tracts having a nearby venue count that is less than 19.

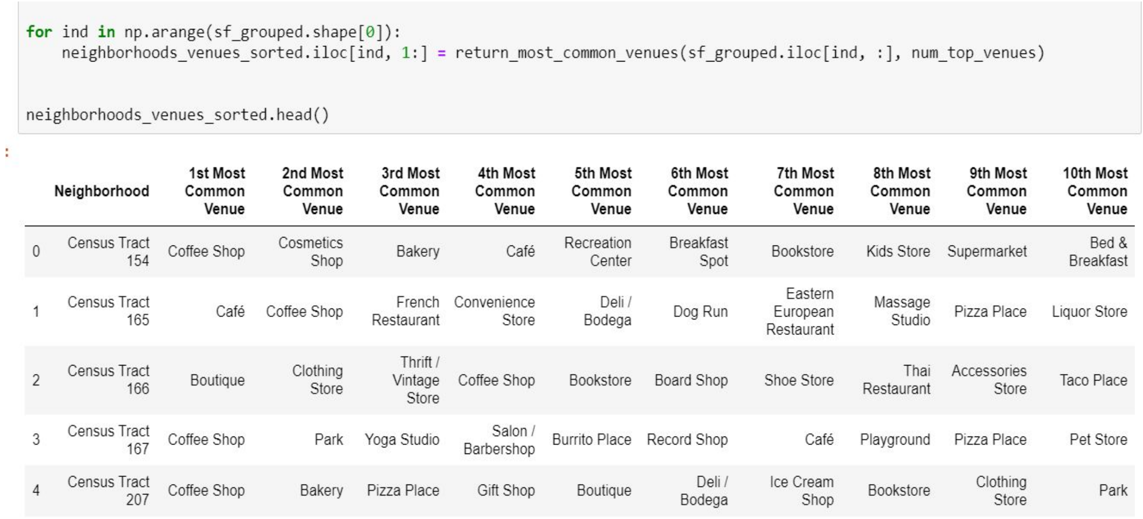


For my next graph, since I had a total of 155 unique venue categories, I decided to visualize the 10 most common venue category from the census tracts. As the figure shows, coffee shops (in combination to cafes) were by far the most frequently occurring venues. Parks were also quite common. These amount of parks and trails were not too surprising given that around three of the tracts were nearby or at a rural area (see map above). The coffee shops did however come at a surprise, it seems like the residents in these tracts love their cup of coffee.



In order to further filter out to assess the tracts by the fourth business requirement (see intro), I have decided to cluster them based on the venue categories. By clustering the tracts, I hoped to find further insights into the types of venues that each tract has. With that insight, I can further filter down the venues that meet the company's requirements.

In order to segment the tracts according to their venue categories, I created another dataframe by applying the one-hot encoding technique for the venue categories. I then grouped the dataframe by tract and took the mean of each one-hot encoded venue categories. I also created a dataframe of the top 10 most common venue categories for each tract, see table below for comparison reasons.

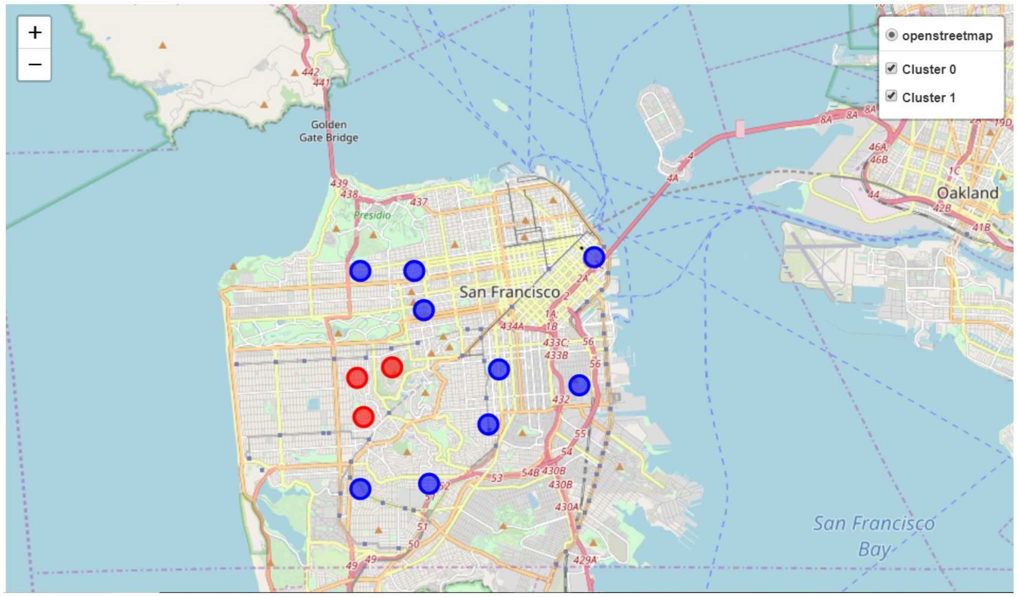


```
for ind in np.arange(sf_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(sf_grouped.iloc[ind, :], num_top_venues)


neighborhoods_venues_sorted.head()
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Census Tract 154	Coffee Shop	Cosmetics Shop	Bakery	Café	Recreation Center	Breakfast Spot	Bookstore	Kids Store	Supermarket	Bed & Breakfast
1	Census Tract 165	Café	Coffee Shop	French Restaurant	Convenience Store	Deli / Bodega	Dog Run	Eastern European Restaurant	Massage Studio	Pizza Place	Liquor Store
2	Census Tract 166	Boutique	Clothing Store	Thrift / Vintage Store	Coffee Shop	Bookstore	Board Shop	Shoe Store	Thai Restaurant	Accessories Store	Taco Place
3	Census Tract 167	Coffee Shop	Park	Yoga Studio	Salon / Barbershop	Burrito Place	Record Shop	Café	Playground	Pizza Place	Pet Store
4	Census Tract 207	Coffee Shop	Bakery	Pizza Place	Gift Shop	Boutique	Deli / Bodega	Ice Cream Shop	Bookstore	Clothing Store	Park

3.4 Clustering By Venue Categories

I have decided to cluster the tracts based on the venue categories by using the K-Means clustering algorithm in the scikit learn library. Since there are only 16 tracts, I chose 3 as the clusters that I wanted to create. I was able to get the following clusters.

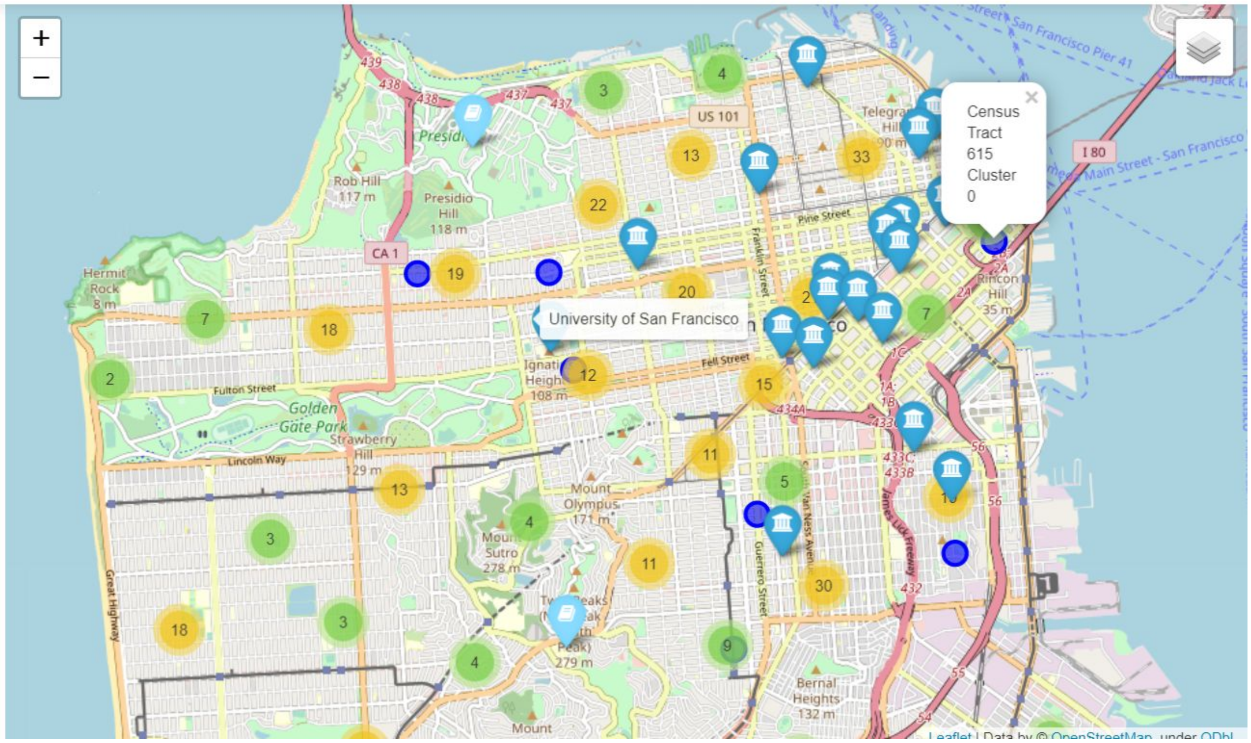


From looking at the results, the cluster labeled as 1 seems to contain outdoor venues as their top 5 most common venue category (park,trails, and mountain). This is followed by dessert-related venues (donut shops and dumpling restaurants). On the other hand, cluster label 0 is dominated by coffee shops and restaurants as their most common venues. However, there are some tracts in cluster 1 with some outdoor venues. This makes sense though, as those tracts are very near the other tracts in cluster 0 which as stated previously, is characterized by its outdoor venues. Despite this, I have chosen cluster 0 as the cluster of focus, as most of the top venues are are similar to the company's business type (food service).

After that, I filtered out the tracts from cluster 0 that contained at least one Mexican-food related venue which were the following: 166, 167, 252 , 308 (see github notebook for reference).Since the company required the least competition, it seemed fair to drop these tracts given that their most common venues include a Mexican-related food venue.

3.5 Visualizing San Francisco's Schools

To find the tracts that fulfilled the last requirements (see intro), I decided to map the schools,colleges, and tracts. Since there were only 11 colleges, I added them to the folium map as markers. However, for the schools, since there were around 400, I decided to show them on the map as cluster markers. These cluster markers cluster the coordinate points (of the schools in this case) that are close to each other and show them as an aggregate total (clusters differ at different zoom level). However it starts showing the individual cluster markers, the more zoomed in, you get to the map. Below is the map that I created. I split the picture in two (same zoom level) to make it easier to view here.



After a quick analysis, we can observe that tracts 402 and 311 didn't meet the last requirement.

IV. Discussion

Our analysis shows that our selected tracts venue categories are characterized by coffee shops, outdoor venues such as parks, and asian restaurants. This makes sense since most of the tracts that these venues belong too are located in areas near several colleges and are centered around parks or trails. We can also assume that these neighborhoods have a relatively large Asian population given the popularity of Asian restaurants observed. In regards to their spatial location in SF, the tracts observed seems to be scattered around the mid-center of the city.

After digging further into the selected tracts, we found seven tracts/neighborhood candidates for the business to choose from. These tracts were found to be affluent with a median income mean of 77,701 dollars, and an average gross rent of 2,229 dollars. In particular, tract 615 stood out with their population size nearly doubling the others (13,864), relatively high median income (103,451 dollars), and proximity to several colleges. Special consideration should be given to this tract.

It should be noted that these results provide only a direction at which the company can conduct further and more detailed analysis of the suggested neighborhood locations. Given the methods used in finding the answer to the business question, I can only conclude that the candidate tracts are potential good locations for the company to look further into as the requirements outlined in the introduction of this report. Despite the limitations mentioned, this report is still useful for the company as it condenses the number of locations that they have to analyze.

VI. Conclusion

Restaurante Los Tacos Del Mar had trouble finding a good location within SF to start their business given the size of SF and large quantity of neighborhoods. In order to find a location that aligns with their business strategy, they came up with five requirements that a neighborhood needed to have: a median income of more than 50,000 dollars, a total population of at least 5,000 people, the median rent less than 3,000 dollars, have nearby venues that are similar in terms of their business type but contain the least amount of competition, and lastly, be nearby at least one college and five other types of schools. With the requirements in place, I retrieved relevant data from the U.S. Census, Data SF, and FourSquare's API with the aims of finding tracts that met the requirements set out by the company.

Several cleaning steps were needed after retrieving the data. After some filtering of the data, the original 197 tracts in SF were condensed to 18 that met the first three requirements previously mentioned. To filter by the last two requirements, I used the K-Means clustering technique and selected cluster 0 as the best representative of the company's business type. In order to fulfill the fourth requirement (see intro), I looked through the dataframe and was able to filter out four tracts that contained Mexican restaurants as one of their common venue categories. The remaining tracts were then mapped along with the schools and colleges in order to select the ones that met the last requirement (see intro). The tracts that met the last requirements were then shown to the company as potential tracts for the business to look further into. This report proved to be beneficial as the company was able to conduct detailed analyses into the resulting seven tracts and ultimately, find the best tract to settle their business in.