Anna Perelman
IBM Data Science
Professional Certificate
Coursera.com

Applied Data Science Capstone Project

28th May 2020

1. Introduction.

This capstone project concludes a series of data science and machine learning courses provided by IBM company via Coursera.com platform and gives its participants an opportunity to come up with an idea to leverage the Foursquare location data to explore or compare neighborhoods or cities of their choice or to come up with a problem that can use the Foursquare location data to be solved using data analysis techniques presented within the specialization.

I decided to implement the following set up: a well-known big chain of private schools would like to open a new branch in San Francisco, CA. They would like to receive recommendations on the best neighborhood to start with. Their branch usually includes a preschool/daycare unit for ages 6 months to 5 years, and an elementary+middle school unit for grades K-8.

In order to provide this theoretical audience with an appropriate set of recommendations it seems wise to collect some data on the usual set of parameters used by this particular chain and its competitors to decide on the best location for their branches. Let's suppose that after some investigation and discussion with the client it was decided to take into account each neighborhood's population, a number of schools already existing within its borders, its safety level and a number of recreational areas (parks and playgrounds) in the proximity. This layer of information should be then combined with the data on different numbers and categories of venues derived from Foursquare, which will give the audience a wide variety of factors to be taken into account for their decision on where to open a new branch.

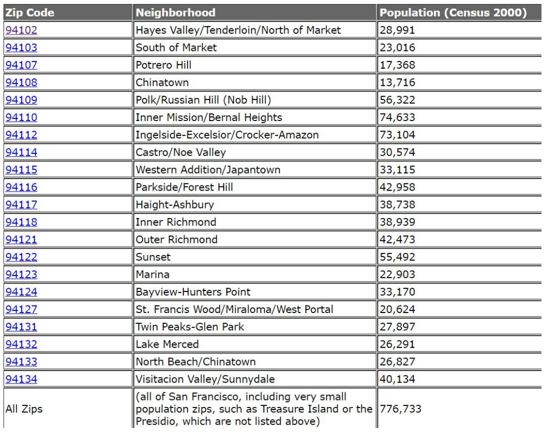
To help my theoretical client with this decision an unsupervised machine learning clustering technique will be used along with colorful maps of neighborhoods and clusters, so they can analyze and compare groups of neighborhoods, and then zoom in into particular ones.

.

2. Data.

This analysis is based on several freely accessible data sources:

a. Information on San Francisco, CA neighborhoods with their zip codes and population can be found at http://www.healthysf.org/bdi/outcomes/zipmap.htm. This simple data set is a table with three columns showing a list of neighborhoods, their zip codes and their population data as can be seen at the following screenshot:



Zip Code	Neighborhood	Population (Census 2000)
94102	Hayes Valley/Tenderloin/North of Market	28,991
94103	South of Market	23,016
94107	Potrero Hill	17,368
94108	Chinatown	13,716
94109	Polk/Russian Hill (Nob Hill)	56,322
94110	Inner Mission/Bernal Heights	74,633
94112	Ingelside-Excelsior/Crocker-Amazon	73,104
94114	Castro/Noe Valley	30,574
94115	Western Addition/Japantown	33,115
94116	Parkside/Forest Hill	42,958
94117	Haight-Ashbury	38,738
94118	Inner Richmond	38,939
94121	Outer Richmond	42,473
94122	Sunset	55,492
94123	Marina	22,903
94124	Bayview-Hunters Point	33,170
94127	St. Francis Wood/Miraloma/West Portal	20,624
94131	Twin Peaks-Glen Park	27,897
94132	Lake Merced	26,291
94133	North Beach/Chinatown	26,827
94134	Visitacion Valley/Sunnydale	40,134
All Zips	(all of San Francisco, including very small population zips, such as Treasure Island or the Presidio, which are not listed above)	776,733

To obtain this table in a format enabling its data analysis using Python language web scraping techniques can be used (BeautifulSoup library is used in this project).

Here is the map of SF neighborhoods:

b. In order to provide geographical maps of the neighborhoods latitude and longitude figures should be added to the above table from Python package *uszipcode*.

c. The information on San Francisco schools with their zip codes, public/private type and grade levels can be found at: https://data.sfgov.org/Economy-and-Community/Schools/tpp3-epx2. This dataset consolidates Infant, Pre-K, and K-14 education points for facilities both public and private, and can be freely downloaded in a variety of formats (CSV format is used in this project).

d. The information on the number of crimes committed in San Francisco along with their latitude and longitude to be converted into zip codes can be found at: https://www.kaggle.com/psmavi104/san-francisco-crime-data#Police_Department_Incident_Reports__2018_to_Present.csv. The data set contains all the crimes listed in the public database from 2018 to present. The data can be downloaded in CSV format for further analysis.

e. The information on parks and playgrounds located in SF neighborhoods can be found at https://www.kaggle.com/san-francisco/sf-recreation-park-department-park-info-dataset. This data set lists all parks, playgrounds, and stadiums within SF borders.

f. The data on different venues located in each neighborhood can be derived from Foursquare location data. Using a free developer account their Places API offers real-time access to Foursquare's global database of rich venue data and user content to power location-based experiences in an app or a website.

3. Methodology

3.1. Exploratory data analysis and descriptive statistics.

Let's suppose that after some discussions and investigation it was decided to take into account the following freely available variables (that were previously mentioned in the introduction section): population, number of schools, crimes and parks/playgrounds for each neighborhood.

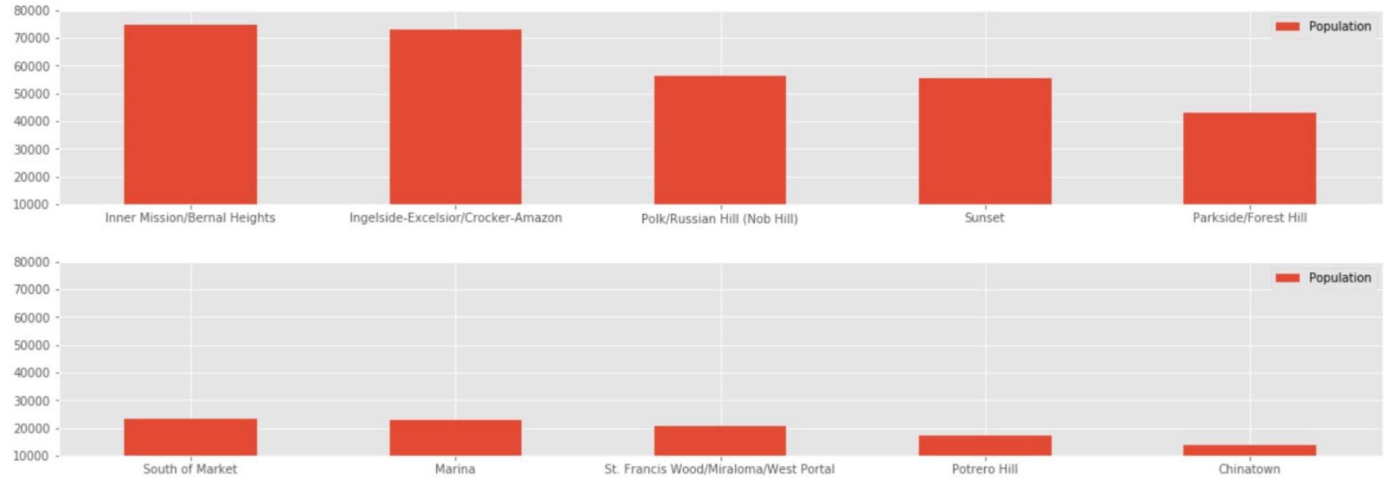
In order to better understand the data some exploratory analysis and descriptive statistics are essential.

Population:

I presume that the higher the population in a neighborhood, the wider is its clients pool for my theoretical client. Based on the data, here are descriptive statistics for the population in SF neighborhoods:

Statistic name	Statistic value
Count - Number of neighborhoods	19.0
Mean - Average of SF neighborhoods population	36, 736
Std - Standard deviation of SF neighborhoods' population	17, 334
Min - The lowest population between SF neighborhoods	13,716
25% - The 25th percentile of the neighborhoods' population	24, 653
50% - The median of the neighborhoods' population	33, 115
75% - The 75th percentile of the neighborhoods' population	41, 546
Max - The highest population between SF neighborhoods	74, 633

Here are five the most populated SF neighborhoods followed by five the least populated ones:



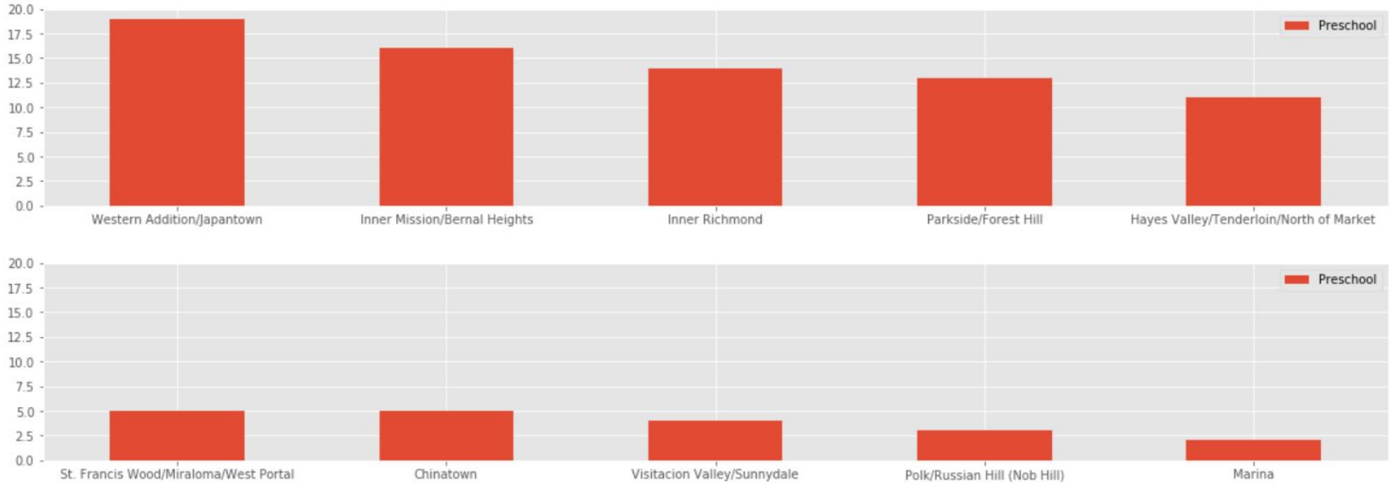
<u>Number of schools:</u>

To enable meaningful comparisons of the proposed branch to its competition, I regrouped the data about SF schools into 3 categories: *Preschool* includes all the institutions for children younger than five years old; *K-8* includes all the institutions providing education for children of appropriate age group; *High_Plus* includes all institutions working with children of 8th grade or higher. My theoretical client needs to know the strength of the potential competition for each neighborhood.

Based on this categorization of the data, here are descriptive statistics for the number of preschools in SF neighborhoods:

Statistic name	Statistic value
Count - Number of neighborhoods	19.0
Mean - Average number of preschools in SF neighborhoods	8.1
Std - Standard deviation of number of preschools in SF neighborhoods	8.1
Min - The lowest number of preschools in SF neighborhoods	2.0
25% - The 25th percentile of the number of preschools in SF neighborhoods	5.5
50% - The median of the number of preschools in SF neighborhoods	6.0
75% - The 75th percentile of the number of preschools in SF neighborhoods	10.0
Max - The highest number of preschools in SF neighborhoods	19.0

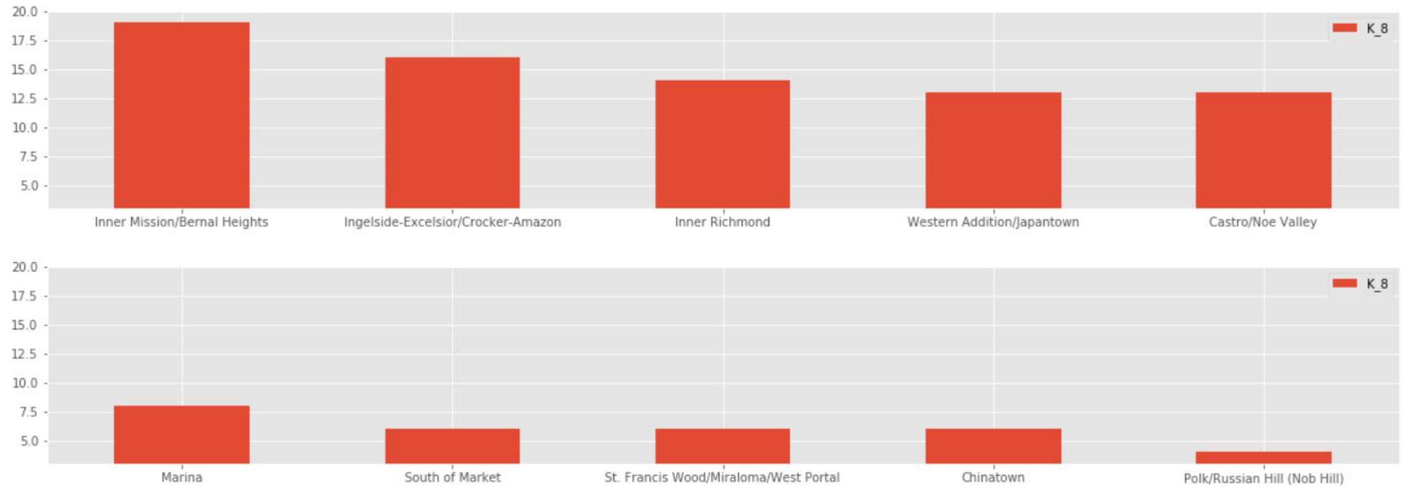
Here are five SF neighborhoods with the highest numbers of preschools in them followed by five the lowests numbers of preschools in them:



The descriptive statistics for the numbers of K-8 schools in SF neighborhoods:

Statistic name	Statistic value
Count - Number of neighborhoods	19.0
Mean - Average number of K-8 schools in SF neighborhoods	9.9
Std - Standard deviation of number of K-8 schools in SF neighborhoods	3.8
Min - The lowest number of K-8 schools in SF neighborhoods	4.0
25% - The 25th percentile of the number of K-8 schools in SF neighborhoods	8.0
50% - The median of the number of K-8 schools in SF neighborhoods	9.0
75% - The 75th percentile of the number of K-8 schools in SF neighborhoods	12.5
Max - The highest number of K-8 schools in SF neighborhoods	19.0

Here are five SF neighborhoods with the highest numbers of preschools in them followed by five the lowests numbers of preschools in them:



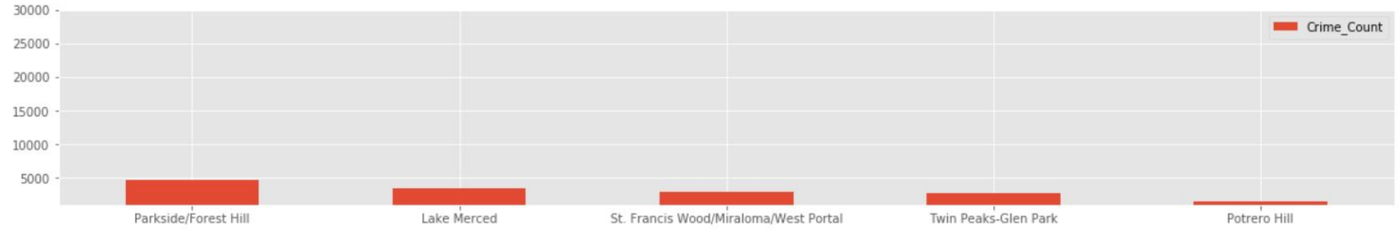
Number of crimes

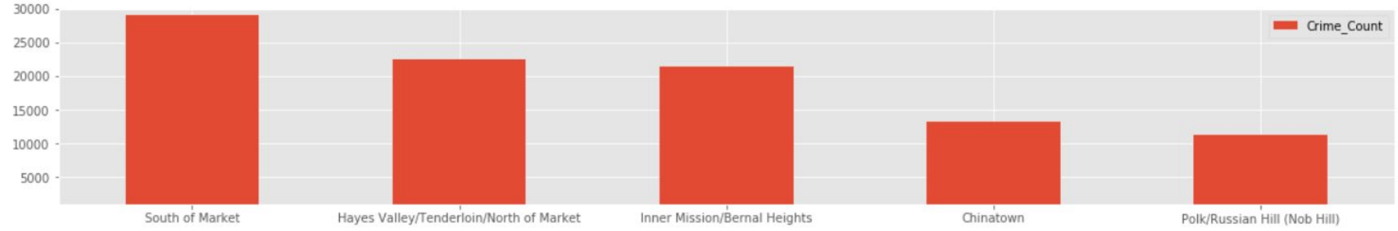
Safety is no doubt the most important criteria for every organization working with children, therefore we have to find a metric of the safety of each neighborhood. I assume that a number of crimes committed in SF neighborhoods between 01/01/2018 and 01/01/2019 may well serve as a metric for this purpose.

The descriptive statistics of this variable are as follows:

Statistic name	Statistic value
Count - Number of neighborhoods	19.0
Mean - Average number of crimes committed in SF neighborhoods	9351.1
Std - Standard deviation of number of crimes committed in SF neighborhoods	7406.7
Min - The lowest number of crimes committed in SF neighborhoods	1564.0
25% - The 25th percentile of the number of crimes committed in SF neighborhoods	4804.5
50% - The median of the number of crimes committed in SF neighborhoods	7250.0
75% - The 75th percentile of the number of crimes committed in SF neighborhoods	10455.5
Max - The highest number of crimes committed in SF neighborhoods	29087.0

Here are five SF neighborhoods with the highest numbers of K-8 schools in them followed by five the lowests numbers of K-8 schools in them:



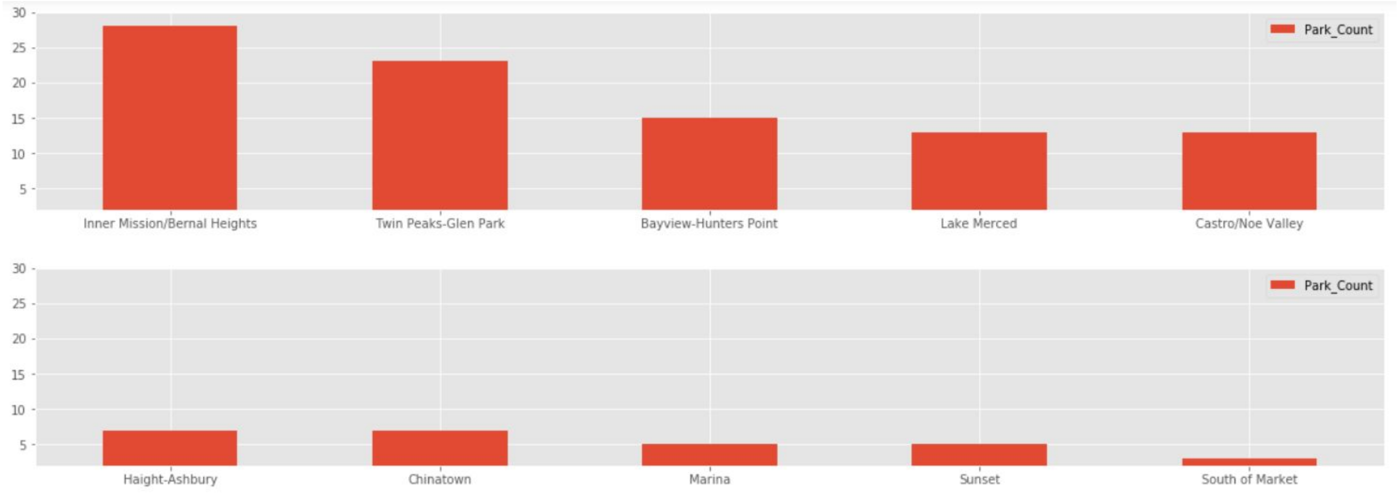


Number of parks and playgrounds:

As a lot of families like to spend some time outside right after the school dismissal for socialization and quality time, a recreational area in the proximity of a school seems to be an advantage. Here are the descriptive statistics for this variable:

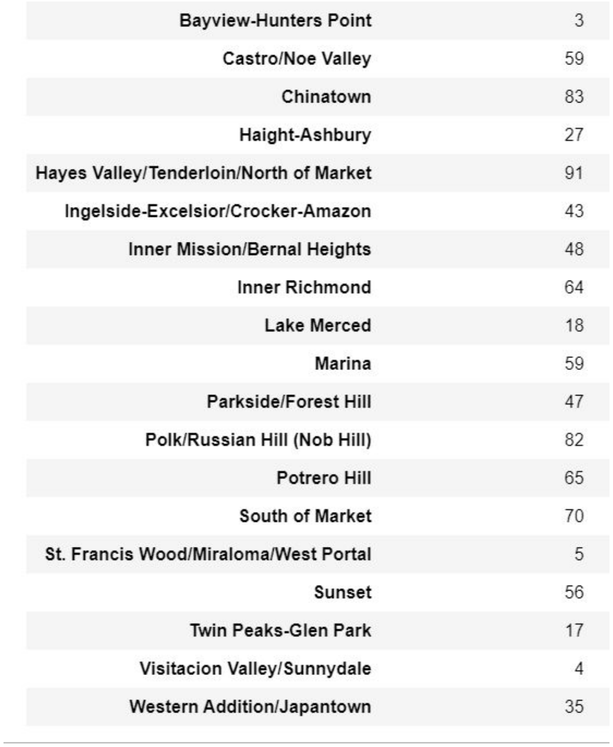
Statistic name	**Statistic value**
Count - Number of neighborhoods	19.0
Mean - Average number of parks and playgrounds in SF neighborhoods	4.7
Std - Standard deviation of number of parks and playgrounds in SF neighborhoods	2.8
Min - The lowest number of parks and playgrounds in SF neighborhoods	1.3
25% - The 25th percentile of the number of parks and playgrounds in SF neighborhoods	3.1
50% - The median of the number of parks and playgrounds in SF neighborhoods	3.6
75% - The 75th percentile of the number of parks and playgrounds in SF neighborhoods	5.4
Max - The highest number of parks and playgrounds in SF neighborhoods	12.6

Here are five SF neighborhoods with the highest numbers of parks and playgroundsin them followed by five the lowests numbers of parks and playgrounds in them:



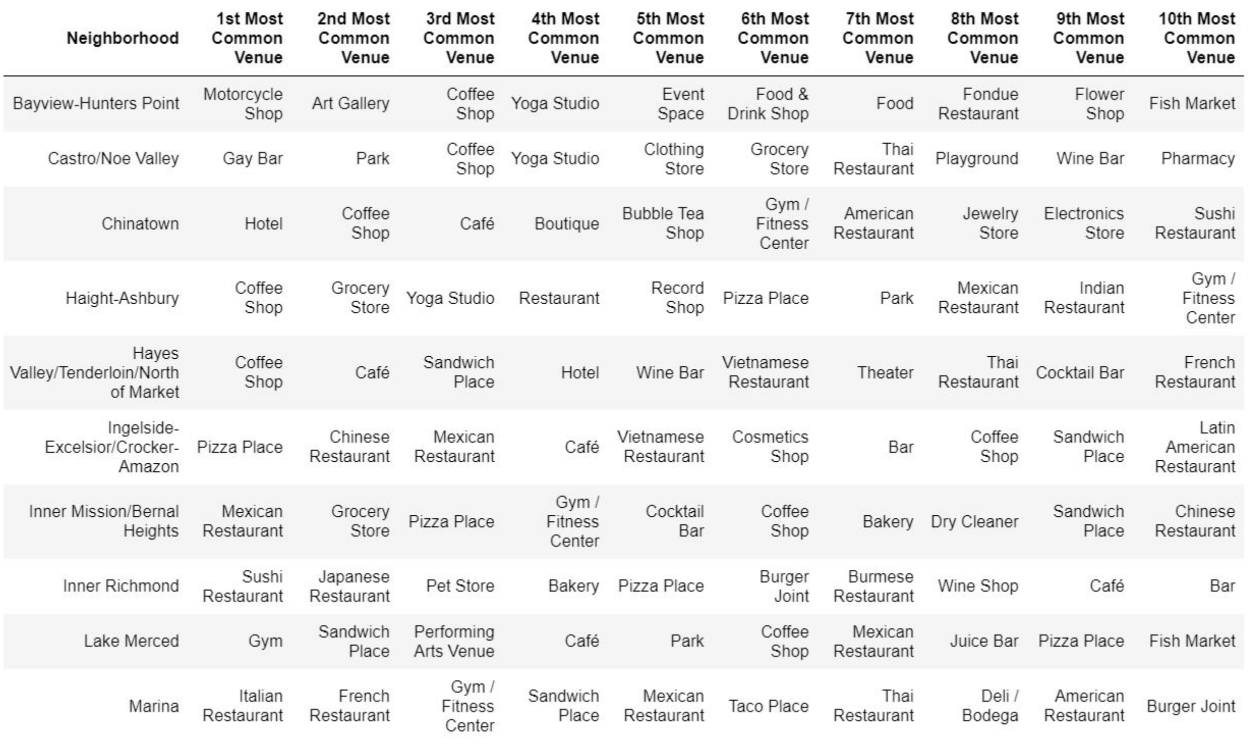
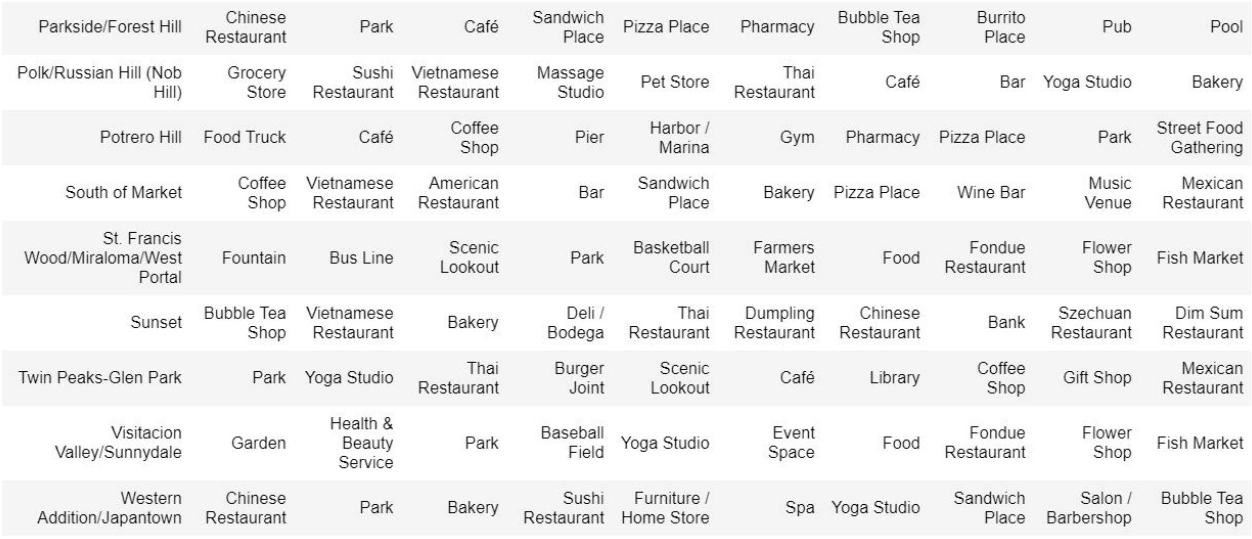
Different venues within each neighborhood:

Using Foursquare location data we can check the venues located within each neighborhood. Clearly places like bookstores and toyshopes, family friendly cafes, convenient grocery stores can make the neighborhood advantageous for opening a new branch since our potential customers can find additional benefits in visiting it daily. On the other hand, some categories can make it disadvantageous (night bars or adult shops, for example). Here presented number of unique venue categories located within each neighborhood:



Bayview-Hunters Point	3
Castro/Noe Valley	59
Chinatown	83
Haight-Ashbury	27
Hayes Valley/Tenderloin/North of Market	91
Ingelside-Excelsior/Crocker-Amazon	43
Inner Mission/Bernal Heights	48
Inner Richmond	64
Lake Merced	18
Marina	59
Parkside/Forest Hill	47
Polk/Russian Hill (Nob Hill)	82
Potrero Hill	65
South of Market	70
St. Francis Wood/Miraloma/West Portal	5
Sunset	56
Twin Peaks-Glen Park	17
Visitacion Valley/Sunnydale	4
Western Addition/Japantown	35

If we check 10 the most frequent venue categories for each neighborhood the following picture can be observed:

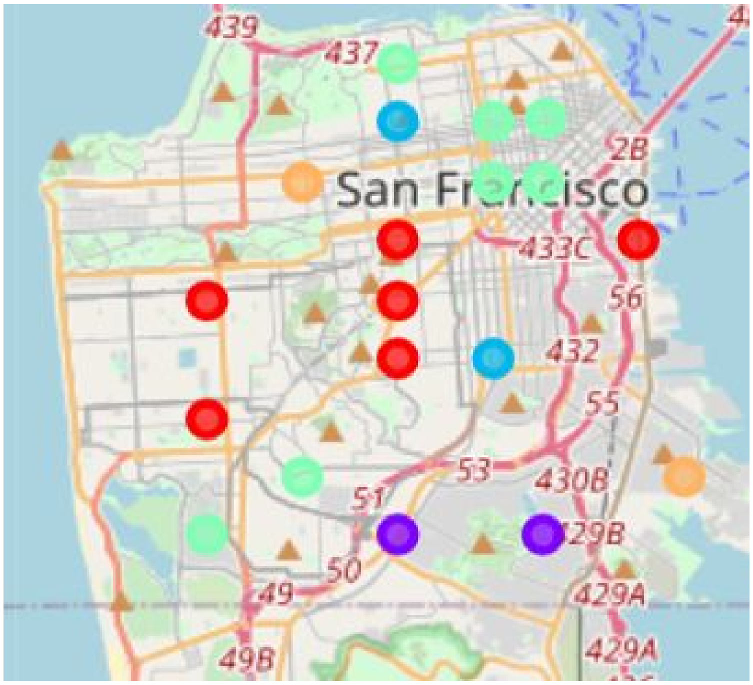
Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Bayview-Hunters Point	Motorcycle Shop	Art Gallery	Coffee Shop	Yoga Studio	Event Space	Food & Drink Shop	Food	Fondue Restaurant	Flower Shop	Fish Market
Castro/Noe Valley	Gay Bar	Park	Coffee Shop	Yoga Studio	Clothing Store	Grocery Store	Thai Restaurant	Playground	Wine Bar	Pharmacy
Chinatown	Hotel	Coffee Shop	Café	Boutique	Bubble Tea Shop	Gym / Fitness Center	American Restaurant	Jewelry Store	Electronics Store	Sushi Restaurant
Haight-Ashbury	Coffee Shop	Grocery Store	Yoga Studio	Restaurant	Record Shop	Pizza Place	Park	Mexican Restaurant	Indian Restaurant	Gym / Fitness Center
Hayes Valley/Tenderloin/North of Market	Coffee Shop	Café	Sandwich Place	Hotel	Wine Bar	Vietnamese Restaurant	Theater	Thai Restaurant	Cocktail Bar	French Restaurant
Ingelside-Excelsior/Crocker-Amazon	Pizza Place	Chinese Restaurant	Mexican Restaurant	Café	Vietnamese Restaurant	Cosmetics Shop	Bar	Coffee Shop	Sandwich Place	Latin American Restaurant
Inner Mission/Bernal Heights	Mexican Restaurant	Grocery Store	Pizza Place	Gym / Fitness Center	Cocktail Bar	Coffee Shop	Bakery	Dry Cleaner	Sandwich Place	Chinese Restaurant
Inner Richmond	Sushi Restaurant	Japanese Restaurant	Pet Store	Bakery	Pizza Place	Burger Joint	Burmese Restaurant	Wine Shop	Café	Bar
Lake Merced	Gym	Sandwich Place	Performing Arts Venue	Café	Park	Coffee Shop	Mexican Restaurant	Juice Bar	Pizza Place	Fish Market
Marina	Italian Restaurant	French Restaurant	Gym / Fitness Center	Sandwich Place	Mexican Restaurant	Taco Place	Thai Restaurant	Deli / Bodega	American Restaurant	Burger Joint
Parkside/Forest Hill	Chinese Restaurant	Park	Café	Sandwich Place	Pizza Place	Pharmacy	Bubble Tea Shop	Burrito Place	Pub	Pool
Polk/Russian Hill (Nob Hill)	Grocery Store	Sushi Restaurant	Vietnamese Restaurant	Massage Studio	Pet Store	Thai Restaurant	Café	Bar	Yoga Studio	Bakery
Potrero Hill	Food Truck	Café	Coffee Shop	Pier	Harbor / Marina	Gym	Pharmacy	Pizza Place	Park	Street Food Gathering
South of Market	Coffee Shop	Vietnamese Restaurant	American Restaurant	Bar	Sandwich Place	Bakery	Pizza Place	Wine Bar	Music Venue	Mexican Restaurant
St. Francis Wood/Miraloma/West Portal	Fountain	Bus Line	Scenic Lookout	Park	Basketball Court	Farmers Market	Food	Fondue Restaurant	Flower Shop	Fish Market
Sunset	Bubble Tea Shop	Vietnamese Restaurant	Bakery	Deli / Bodega	Thai Restaurant	Dumpling Restaurant	Chinese Restaurant	Bank	Szechuan Restaurant	Dim Sum Restaurant
Twin Peaks-Glen Park	Park	Yoga Studio	Thai Restaurant	Burger Joint	Scenic Lookout	Café	Library	Coffee Shop	Gift Shop	Mexican Restaurant
Visitacion Valley/Sunnydale	Garden	Health & Beauty Service	Park	Baseball Field	Yoga Studio	Event Space	Food	Fondue Restaurant	Flower Shop	Fish Market
Western Addition/Japantown	Chinese Restaurant	Park	Bakery	Sushi Restaurant	Furniture / Home Store	Spa	Yoga Studio	Sandwich Place	Salon / Barbershop	Bubble Tea Shop

3.2 Clustering analysis with K-Means (unsupervised machine learning).

It can be seen that each SF neighborhood included in this analysis is characterised by its population, number of different schools, crime level, availability of parks/playgrounds, and existence of different venues providing goods and services that contribute positively or negatively to the neighborhood’s suitability for my client’s purposes. The complexity of analyzing these different quantitative characteristics may be decreased by creation of several clusters or groups of similar neighborhoods. The next steps then can be an understanding of the nature of their similarities, selecting the most promising cluster based on these similar characteristics, and then zooming into this cluster in order to find the best candidates to be the new branch’s location.

The technique usually used is unsupervised machine learning technique K-Means that has been proven to be fairly effective. Wikipedia defines this method as vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster (Wikipedia). Let’s assume that domain knowledge and expertise, it was decided that the most reasonable number of clusters in this case is five.

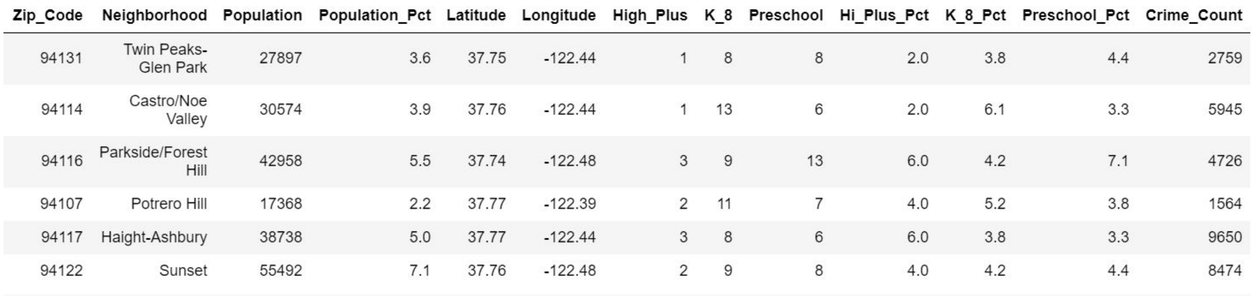
Here are these clusters on SF map:



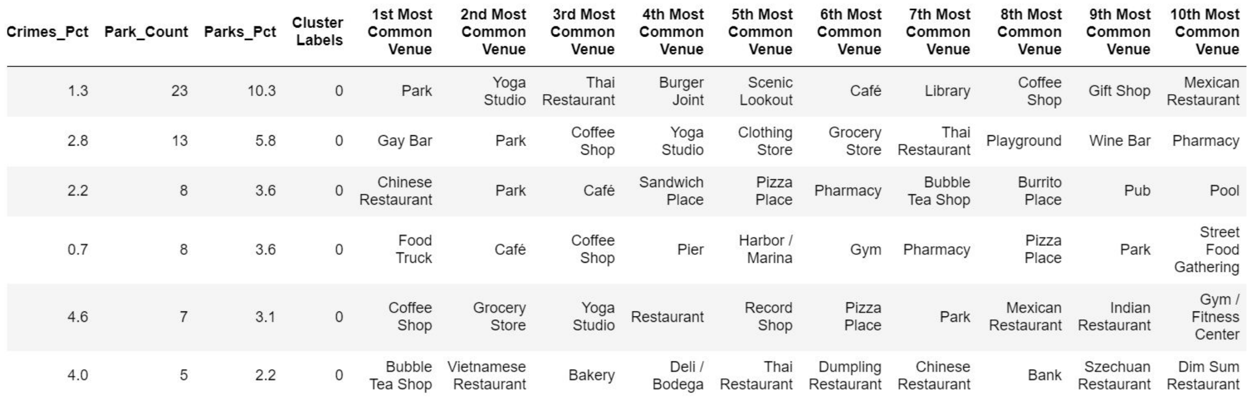
4. Results

As a result of data processing and investigation followed by running a K-Means algorithm, the following clusters have been created:

Cluster 1:



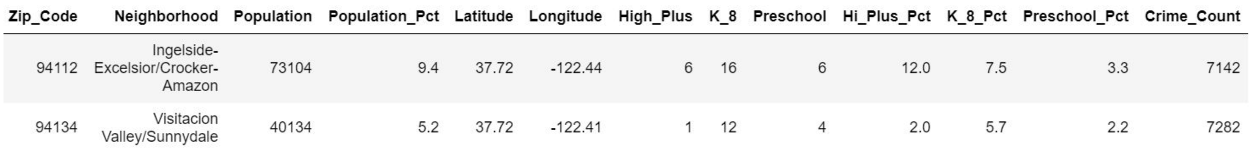
Zip_Code	Neighborhood	Population	Population_Pct	Latitude	Longitude	High_Plus	K_8	Preschool	Hi_Plus_Pct	K_8_Pct	Preschool_Pct	Crime_Count
94131	Twin Peaks-Glen Park	27897	3.6	37.75	-122.44	1	8	8	2.0	3.8	4.4	2759
94114	Castro/Noe Valley	30574	3.9	37.76	-122.44	1	13	6	2.0	6.1	3.3	5945
94116	Parkside/Forest Hill	42958	5.5	37.74	-122.48	3	9	13	6.0	4.2	7.1	4726
94107	Potrero Hill	17368	2.2	37.77	-122.39	2	11	7	4.0	5.2	3.8	1564
94117	Haight-Ashbury	38738	5.0	37.77	-122.44	3	8	6	6.0	3.8	3.3	9650
94122	Sunset	55492	7.1	37.76	-122.48	2	9	8	4.0	4.2	4.4	8474



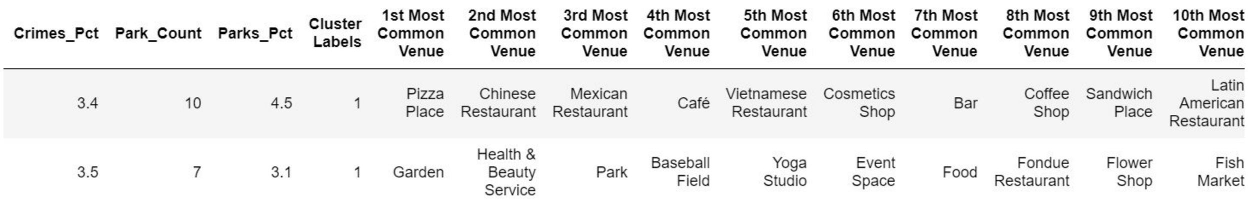
Crimes_Pct	Park_Count	Parks_Pct	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1.3	23	10.3	0	Park	Yoga Studio	Thai Restaurant	Burger Joint	Scenic Lookout	Café	Library	Coffee Shop	Gift Shop	Mexican Restaurant
2.8	13	5.8	0	Gay Bar	Park	Coffee Shop	Yoga Studio	Clothing Store	Grocery Store	Thai Restaurant	Playground	Wine Bar	Pharmacy
2.2	8	3.6	0	Chinese Restaurant	Park	Café	Sandwich Place	Pizza Place	Pharmacy	Bubble Tea Shop	Burrito Place	Pub	Pool
0.7	8	3.6	0	Food Truck	Café	Coffee Shop	Pier	Harbor / Marina	Gym	Pharmacy	Pizza Place	Park	Street Food Gathering
4.6	7	3.1	0	Coffee Shop	Grocery Store	Yoga Studio	Restaurant	Record Shop	Pizza Place	Park	Mexican Restaurant	Indian Restaurant	Gym / Fitness Center
4.0	5	2.2	0	Bubble Tea Shop	Vietnamese Restaurant	Bakery	Deli / Bodega	Thai Restaurant	Dumpling Restaurant	Chinese Restaurant	Bank	Szechuan Restaurant	Dim Sum Restaurant

We can characterize this cluster as being fairly populated (between 25th and 75th percentiles), highly competitive (the number of preschools and K-8 schools being above 50th percentile), safe (the number of crimes being below 50th percentile), with lots of coffee shops and restaurants, and parks.

Cluster 2:



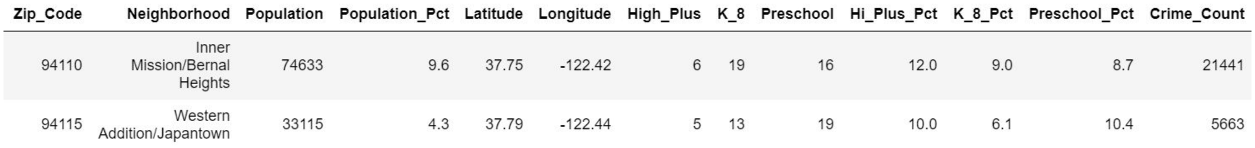
Zip_Code	Neighborhood	Population	Population_Pct	Latitude	Longitude	High_Plus	K_8	Preschool	Hi_Plus_Pct	K_8_Pct	Preschool_Pct	Crime_Count
94112	Ingelside-Excelsior/Crocker-Amazon	73104	9.4	37.72	-122.44	6	16	6	12.0	7.5	3.3	7142
94134	Visitacion Valley/Sunnydale	40134	5.2	37.72	-122.41	1	12	4	2.0	5.7	2.2	7282



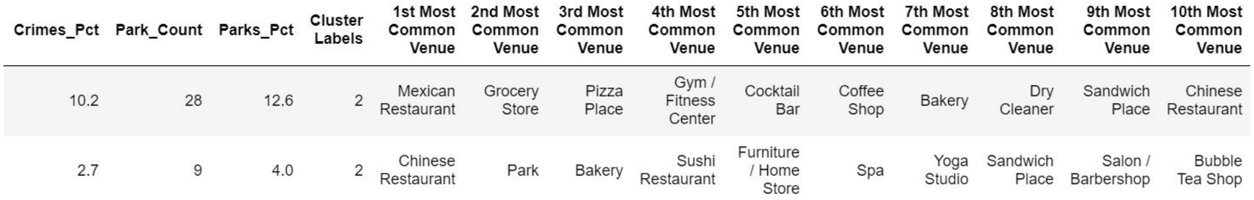
Crimes_Pct	Park_Count	Parks_Pct	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3.4	10	4.5	1	Pizza Place	Chinese Restaurant	Mexican Restaurant	Café	Vietnamese Restaurant	Cosmetics Shop	Bar	Coffee Shop	Sandwich Place	Latin American Restaurant
3.5	7	3.1	1	Garden	Health & Beauty Service	Park	Baseball Field	Yoga Studio	Event Space	Food	Fondue Restaurant	Flower Shop	Fish Market

This cluster with only two neighborhoods in it can be characterized with high population numbers, big quantity of K-8 schools and low number of preschools, low safety level (number of crimes committed is higher than 50th percentile), minimal number of parks, and lots of restaurants.

Cluster 3:



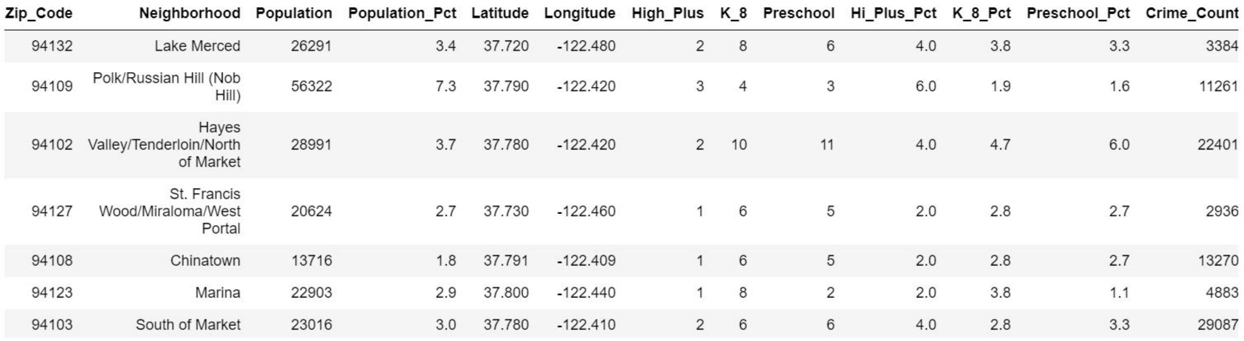
Zip_Code	Neighborhood	Population	Population_Pct	Latitude	Longitude	High_Plus	K_8	Preschool	Hi_Plus_Pct	K_8_Pct	Preschool_Pct	Crime_Count
94110	Inner Mission/Bernal Heights	74633	9.6	37.75	-122.42	6	19	16	12.0	9.0	8.7	21441
94115	Western Addition/Japantown	33115	4.3	37.79	-122.44	5	13	19	10.0	6.1	10.4	5663



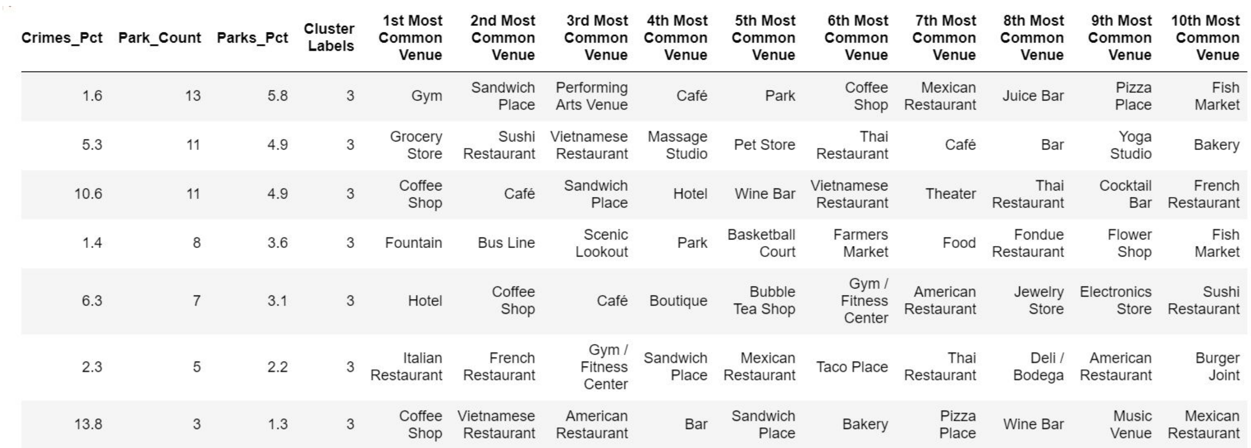
Crimes_Pct	Park_Count	Parks_Pct	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
10.2	28	12.6	2	Mexican Restaurant	Grocery Store	Pizza Place	Gym / Fitness Center	Cocktail Bar	Coffee Shop	Bakery	Dry Cleaner	Sandwich Place	Chinese Restaurant
2.7	9	4.0	2	Chinese Restaurant	Park	Bakery	Sushi Restaurant	Furniture / Home Store	Spa	Yoga Studio	Sandwich Place	Salon / Barbershop	Bubble Tea Shop

Similarly to the previous cluster this one also contains only two neighborhoods with pretty different characteristics of population, safety and number of parks parameters, but they both have high amount of schools and preschools along with restaurants and coffee shops.

Cluster 4:



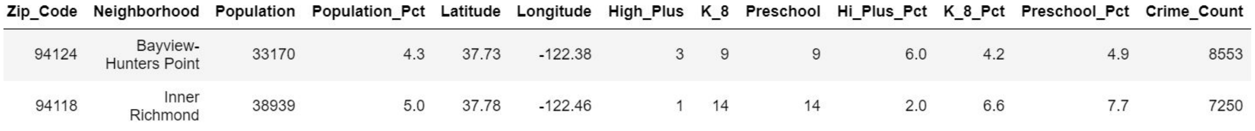
Zip_Code	Neighborhood	Population	Population_Pct	Latitude	Longitude	High_Plus	K_8	Preschool	Hi_Plus_Pct	K_8_Pct	Preschool_Pct	Crime_Count
94132	Lake Merced	26291	3.4	37.720	-122.480	2	8	6	4.0	3.8	3.3	3384
94109	Polk/Russian Hill (Nob Hill)	56322	7.3	37.790	-122.420	3	4	3	6.0	1.9	1.6	11261
94102	Hayes Valley/Tenderloin/North of Market	28991	3.7	37.780	-122.420	2	10	11	4.0	4.7	6.0	22401
94127	St. Francis Wood/Miraloma/West Portal	20624	2.7	37.730	-122.460	1	6	5	2.0	2.8	2.7	2936
94108	Chinatown	13716	1.8	37.791	-122.409	1	6	5	2.0	2.8	2.7	13270
94123	Marina	22903	2.9	37.800	-122.440	1	8	2	2.0	3.8	1.1	4883
94103	South of Market	23016	3.0	37.780	-122.410	2	6	6	4.0	2.8	3.3	29087



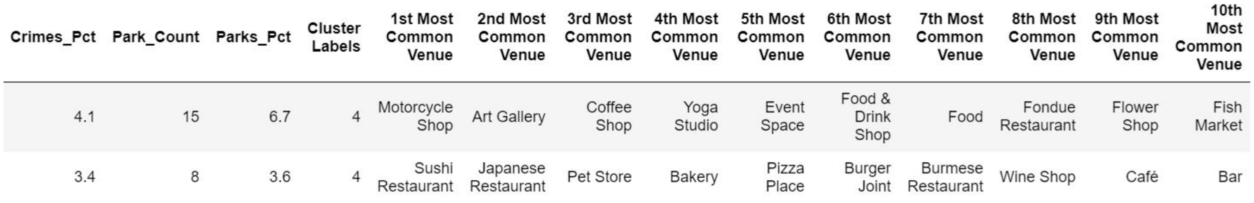
Crimes_Pct	Park_Count	Parks_Pct	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1.6	13	5.8	3	Gym	Sandwich Place	Performing Arts Venue	Café	Park	Coffee Shop	Mexican Restaurant	Juice Bar	Pizza Place	Fish Market
5.3	11	4.9	3	Grocery Store	Sushi Restaurant	Vietnamese Restaurant	Massage Studio	Pet Store	Thai Restaurant	Café	Bar	Yoga Studio	Bakery
10.6	11	4.9	3	Coffee Shop	Café	Sandwich Place	Hotel	Wine Bar	Vietnamese Restaurant	Theater	Thai Restaurant	Cocktail Bar	French Restaurant
1.4	8	3.6	3	Fountain	Bus Line	Scenic Lookout	Park	Basketball Court	Farmers Market	Food	Fondue Restaurant	Flower Shop	Fish Market
6.3	7	3.1	3	Hotel	Coffee Shop	Café	Boutique	Bubble Tea Shop	Gym / Fitness Center	American Restaurant	Jewelry Store	Electronics Store	Sushi Restaurant
2.3	5	2.2	3	Italian Restaurant	French Restaurant	Gym / Fitness Center	Sandwich Place	Mexican Restaurant	Taco Place	Thai Restaurant	Deli / Bodega	American Restaurant	Burger Joint
13.8	3	1.3	3	Coffee Shop	Vietnamese Restaurant	American Restaurant	Bar	Sandwich Place	Bakery	Pizza Place	Wine Bar	Music Venue	Mexican Restaurant

This cluster is characterized by low population, average number of schools and preschools, pretty high crime levels, lots of parks and restaurants.

Cluster 5:



Zip_Code	Neighborhood	Population	Population_Pct	Latitude	Longitude	High_Plus	K_8	Preschool	Hi_Plus_Pct	K_8_Pct	Preschool_Pct	Crime_Count
94124	Bayview-Hunters Point	33170	4.3	37.73	-122.38	3	9	9	6.0	4.2	4.9	8553
94118	Inner Richmond	38939	5.0	37.78	-122.46	1	14	14	2.0	6.6	7.7	7250



Crimes_Pct	Park_Count	Parks_Pct	Cluster Labels	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
4.1	15	6.7	4	Motorcycle Shop	Art Gallery	Coffee Shop	Yoga Studio	Event Space	Food & Drink Shop	Food	Fondue Restaurant	Flower Shop	Fish Market
3.4	8	3.6	4	Sushi Restaurant	Japanese Restaurant	Pet Store	Bakery	Pizza Place	Burger Joint	Burmese Restaurant	Wine Shop	Café	Bar

This cluster with only two neighborhoods again can be characterized by an average population level, number of schools and preschools, relatively high level of crime, great amount of parks and restaurants.

5. Discussion

Based on the above results the first cluster seems to be a good option for the new branch. Although a wide variety of schools and preschools is already presented in its neighborhoods, it still shows enough potential for a new player. Within this cluster Twin Peaks/Glen Parks looks like an optimal combination of venues (coffee shops, library) and population and competition factors.

Within fourth cluster the neighborhood St. Francis Wood/Miraloma/West Portal is worth attention for the same reasons.

6. Conclusion

The goal of this project was to give a hypothetical owner of a big preschool and school chain an advice on where to open a new branch in San Francisco, CA. In order to answer this question a database of characterics of San Francisco, CA neighborhoods was built. The list of factors taken into consideration included population, number of schools and preschools, number of crimes committed during the year of 2018, number of parks existing in the neighborhoods, number and category of surrounding venues. After a preliminary exploratory analysis an unsupervised machine learning algorithm K-Means was run on the database, as a result of which five clusters were built. Zooming into the first and the fourth clusters, two candidates to be the new branch's location were suggested: Twin Peaks/Glen Parks and St. Francis Wood/Miraloma/West Portal. Both of these neighborhoods are fairly populated and safem with reasonable competition leve, and are surrounded by a nice combination of venues and parks. As St. Francis Wood/Miraloma/West Portal being a good option by itself is a part of a fourth cluster that doesn't appear to serve as an appropriate location, I'd recommend to my theoretical client to start with Twin Peaks/Glen Parks.