San Francisco Housing Price, School and Venues Data

Introduction

San Francisco is considered by many to be the cultural and industrial hub of Northern California. Although the city covers only 46.9 square miles, it's home to nearly 882,000 residents, making it the 16th largest metropolitan statistical area and one of the highest per-capita incomes in the United States, which makes it a very popular area for investors.

Starting in March 2020, economic volatility and shelter-in-place orders caused home sales volume to decline. Housing demand in San Francisco has experienced some ups and downs amid the coronavirus pandemic. Though housing prices in San Francisco are much higher than the national average, they are also currently following a nationwide trend. Low mortgage rates and a large influx of buyers are behind a surge in demand, which has caused housing prices to rise over the last year.

Business Problem: San Francisco has been hit harder by the pandemic than many other cities, small business closures, outmigration and unemployment. It changed neighborhood's business information and maybe started creating new venues to neighborhoods while the economy is rolling to a new normal. As businesses begin to reopen in San Francisco, at some point, a lot of people are going to be returning to work and schools starting in person classes. This project aims to find the best suitable neighborhood with venues and schools in San Francisco, by using data analysis and machine learning.

Target Audience: This data analysis can be useful to realtors, homebuyers, city managers and to people who are planning to open a small business.

Data

To solve the problem, we need the following data:

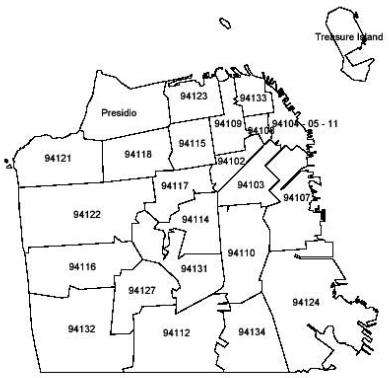
1. List of Neighborhoods in San Francisco, California

   (http://www.healthysf.org/bdi/outcomes/zipmap.htm) to get neighborhood and zip code data.

San Francisco Burden of Disease & Injury Study:
Determinants of Health

About Site | Determinants | Health Outcomes | Web Links | Site Map

San Francisco Neighborhoods as ZIP Codes



Zip Code	Neighborhood
94102	Hayes Valley/Tenderloin/North of Market
94103	South of Market
94107	Potrero Hill
94108	Chinatown
94109	Polk/Russian Hill (Nob Hill)
94110	Inner Mission/Bernal Heights
94112	Ingelside-Excelsior/Crocker-Amazon
94114	Castro/Noe Valley
94115	Western Addition/Japantown
94116	Parkside/Forest Hill
94117	Haight-Ashbury
94118	Inner Richmond
94121	Outer Richmond
94122	Sunset
94123	Marina

2. Latitude and Longitude of the Neighborhoods:

   I used Geocoder to get center coordinates of the neighborhoods.

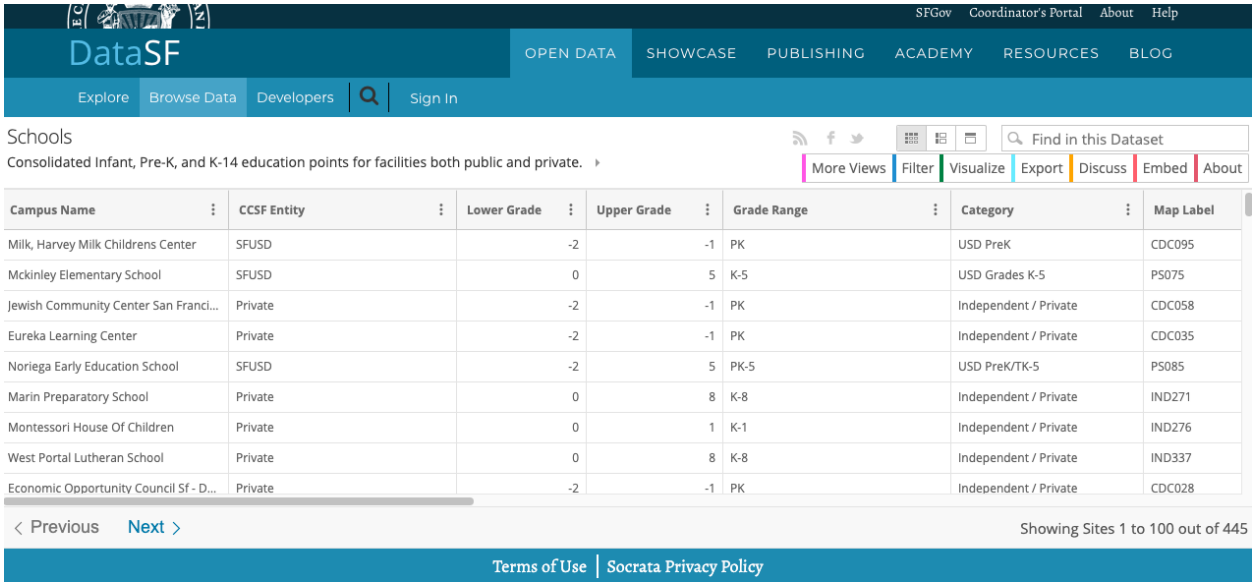
3. Venues Data:

   Foursquare API for most common venues in San Francisco.

4. Schools information:

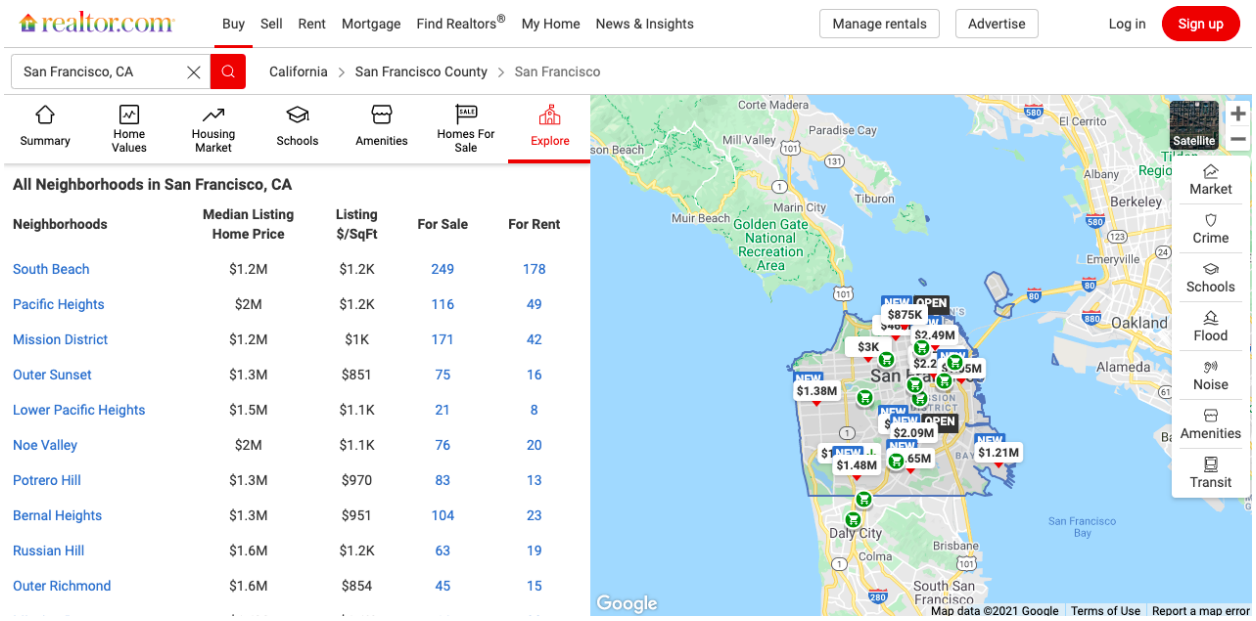
   (https://data.sfgov.org/Economy-and-Community/Schools/tpp3-epx2/data) to get the number of school information for each neighborhood.

   I exported the .csv file to github, and used that data on Notebook.



5. Average housing price:

(https://www.bayareamarketreports.com/trend/san-francisco-neighborhood-map) and (realtor.com) for Neighborhood Average home price



Methodology

To get Neighborhood and Zip Codes info I used web scraping by utilizing the pandas HTML table scraping method directly from a web page into a data frame. And added the median housing price column, cleaned, merged the Map_of_Schools.csv file that downloaded from the web. Finally I created a new data frame that only contained the columns I was interested in. The Schools column number includes all private and public schools from PreK to community colleges.

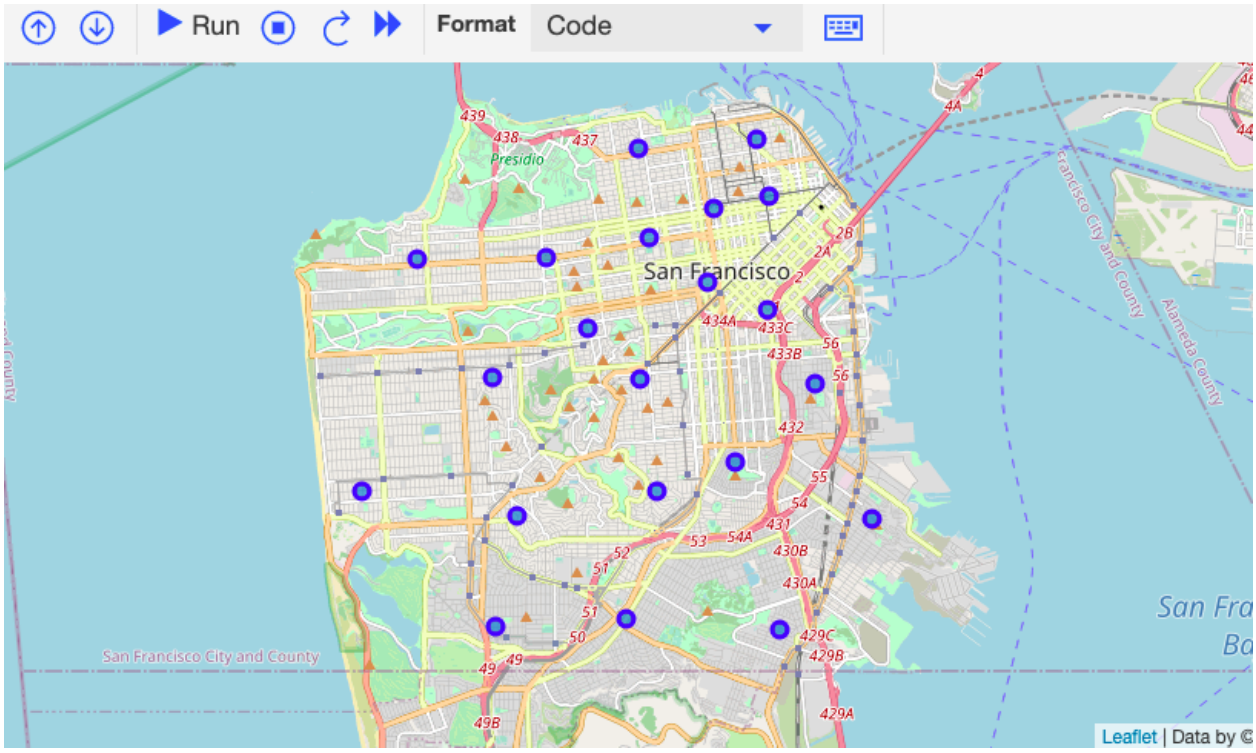
```
df=pd.merge(d_sf_price, sf_school, on='Zip Code', how='left')
df
```

31]:

	Zip Code	Neighborhood	MedianHomePrice	Schools
0	94102	Hayes Valley/Tenderloin/North of Market	1320000	23
1	94103	South of Market	1200000	14
2	94107	Potrero Hill	1898000	20
3	94108	Chinatown	1280000	12
4	94109	Polk/Russian Hill (Nob Hill)	1550000	10
5	94110	Inner Mission/Bernal Heights	1700000	41
6	94112	Ingelside-Excelsior/Crocker-Amazon	1200000	28
7	94114	Castro/Noe Valley	2000000	20
8	94115	Western Addition/Japantown	2150000	37

To find their coordinates I used Geocoder. After gathering all these coordinates, I visualized the map of San Francisco using the Folium package.

	Zip Code	Neighborhood	MedianHomePrice	Schools	Latitude	Longitude
0	94102	Hayes Valley/Tenderloin/North of Market	1320000	23	37.777015	-122.421875
1	94103	South of Market	1200000	14	37.772000	-122.408735
2	94107	Potrero Hill	1898000	20	37.759050	-122.398155
3	94108	Chinatown	1280000	12	37.792160	-122.408220
4	94109	Polk/Russian Hill (Nob Hill)	1550000	10	37.790105	-122.420590
5	94110	Inner Mission/Bernal Heights	1700000	41	37.745185	-122.415905
6	94112	Ingelside-Excelsior/Crocker-Amazon	1200000	28	37.717485	-122.440255
7	94114	Castro/Noe Valley	2000000	20	37.759975	-122.437105
8	94115	Western Addition/Japantown	2150000	37	37.784895	-122.435125
9	94116	Parkside/Forest Hill	1500000	25	37.740140	-122.499415
10	94117	Haight-Ashbury	3250000	17	37.768785	-122.448920



Next, I used the Foursquare API to pull the list of top 100 venues within 500 meters radius with the Foursquare developer account ID and API key to pull the data.

```
print(sf_venues.shape)
sf_venues.head()
```

```
(1182, 7)
```

0]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Hayes Valley/Tenderloin/North of Market	37.777015	-122.421875	SFJazz Center	37.776350	-122.421539	Jazz Club
1	Hayes Valley/Tenderloin/North of Market	37.777015	-122.421875	Dumpling Home	37.776050	-122.422969	Dumpling Restaurant
2	Hayes Valley/Tenderloin/North of Market	37.777015	-122.421875	Blue Bottle Coffee	37.776430	-122.423224	Coffee Shop
3	Hayes Valley/Tenderloin/North of Market	37.777015	-122.421875	Louise M. Davies Symphony Hall	37.777976	-122.420157	Concert Hall
4	Hayes Valley/Tenderloin/North of Market	37.777015	-122.421875	Fig & Thistle Wine Bar	37.777256	-122.423365	Wine Bar

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

```
There are 244 uniques categories.
```

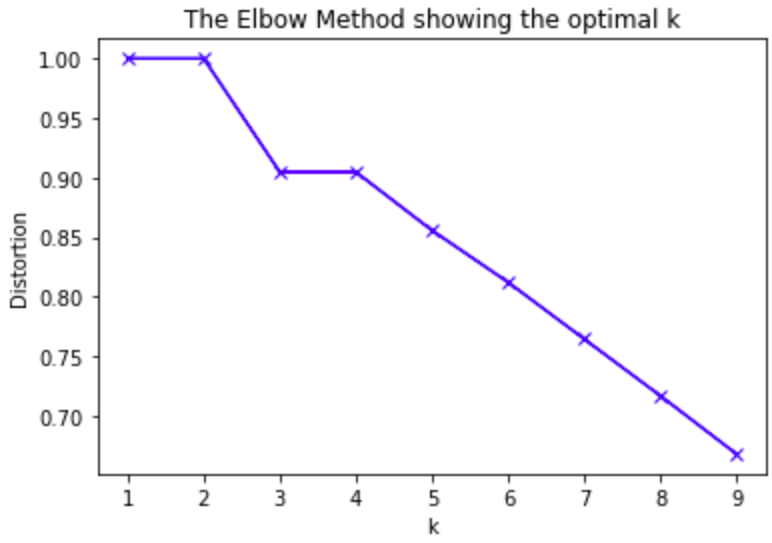
There are 244 unique categories.

Then I created a top 10 venue category table for each neighbor.

7]:

	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	Bayview-Hunters Point	Park	Southern / Soul Food Restaurant	Latin American Restaurant	Pizza Place	Playground	Non-Profit	Fried Chicken Joint	Bus Station	Skate Park	Mexican Restaurant
1	Castro/Noe Valley	Gay Bar	Thai Restaurant	Coffee Shop	Park	Playground	Seafood Restaurant	Cosmetics Shop	Convenience Store	Mediterranean Restaurant	Deli / Bodega
2	Chinatown	Coffee Shop	Hotel	Bakery	Hotel Bar	Spa	Bubble Tea Shop	Bar	Men's Store	Steakhouse	Dim Sum Restaurant
3	Haight-Ashbury	Boutique	Coffee Shop	Thrift / Vintage Store	Clothing Store	Shoe Store	Convenience Store	Bookstore	Breakfast Spot	Café	Ice Cream Shop
4	Hayes Valley/Tenderloin/North of Market	Wine Bar	Sushi Restaurant	Pizza Place	French Restaurant	Boutique	Cocktail Bar	Bakery	Coffee Shop	New American Restaurant	Mexican Restaurant
5	Ingelside-Excelsior/Crocker-	Mexican	Pizza Place	Bar	Filipino	Restaurant	Vietnamese	Sandwich	Burrito Place	Bus Station	Bakery

I performed the clustering with k-means clustering method. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster. I have clustered the neighborhoods in San Francisco into 3 clusters based on elbow method with jaccard distance which gave a clearer picture than euclidean and canberra.

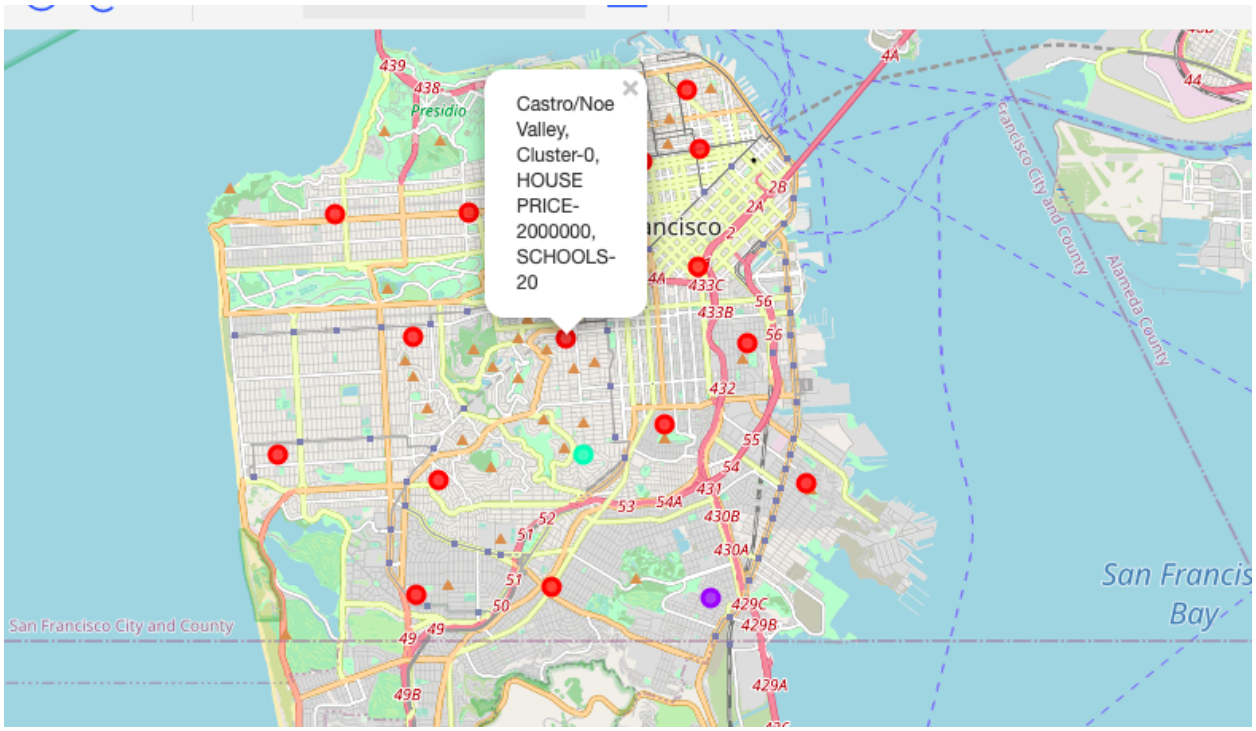


Results

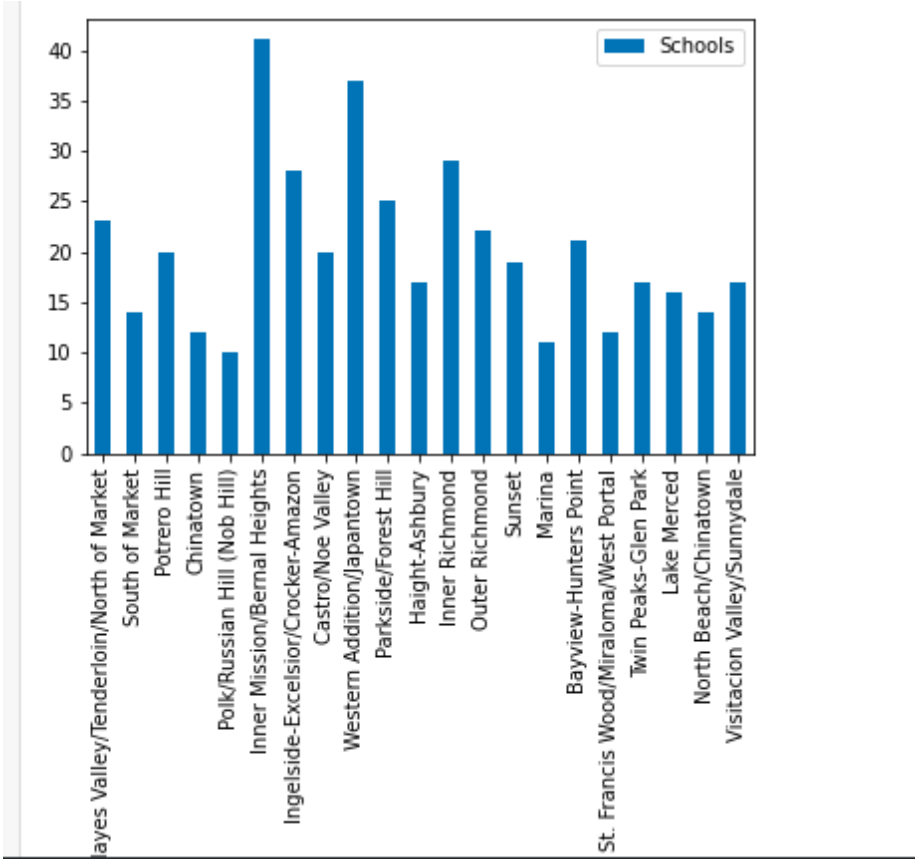
Let’s create a clustered map from our main table.

	Zip Code	Neighborhood	MedianHomePrice	Schools	Latitude	Longitude	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
0	94102	Hayes Valley/Tenderloin/North of Market	1320000	23	37.777015	-122.421875	0	Wine Bar	Sushi Restaurant	Pizza Place	French Restaurant	Boutique	Cocktail Bar
1	94103	South of Market	1200000	14	37.772000	-122.408735	0	Nightclub	Gay Bar	Café	Furniture / Home Store	Cocktail Bar	Motorcycle Shop
2	94107	Potrero Hill	1898000	20	37.759050	-122.398155	0	Café	Grocery Store	Breakfast Spot	Coffee Shop	Indie Theater	Liquor Store
3	94108	Chinatown	1280000	12	37.792160	-122.408220	0	Coffee Shop	Hotel	Bakery	Hotel Bar	Spa	Bubble Tea Shop
4	94109	Polk/Russian Hill (Nob Hill)	1550000	10	37.790105	-122.420590	0	Thai Restaurant	Grocery Store	Café	Massage Studio	Sushi Restaurant	Bar
5	94110	Inner Mission/Bernal Heights	1700000	41	37.745185	-122.415905	0	Playground	Mexican Restaurant	Park	Yoga Studio	Sandwich Place	Brewery
6	94112	Ingelside-Excelsior/Crocker-Amazon	1200000	28	37.717485	-122.440255	0	Mexican Restaurant	Pizza Place	Bar	Filipino Restaurant	Restaurant	Vietnamese Restaurant

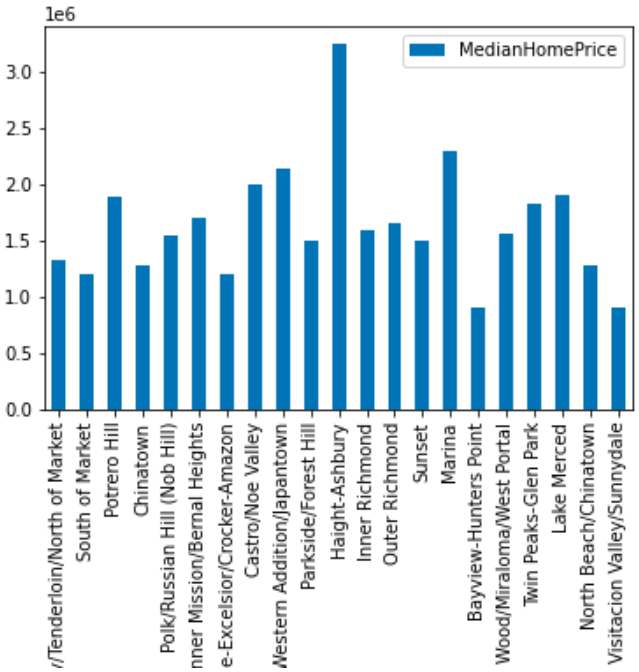
On the map we can see Neighborhood name, Cluster number, Median Housing Price and Number of Schools in that neighborhood.



I used a Plot diagram to see Housing price and school number in each neighbor




```
]: import matplotlib.pyplot as plt
   df.plot(x='Neighborhood', y='MedianHomePrice', kind='bar')
   plt.show()
```



Discussion

As you can see on the clustered map, San Francisco venues spreaded almost evenly through the neighborhoods, only one neighbor on Cluster 1 and Cluster 2.

Because of the limited number of API calls and there are only a few neighborhoods in San Francisco, we are not able to see the whole picture of results. For business opportunities and real estate decisions we must consider other information like housing price, crime rate, school info, geographical information etc. A larger and more varied data set would give more realistic information.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilizing k-means clustering and providing a recommendation to the audencies. These analytical tools open a window of possibilities for decision making across the various businesses when combined with accurate and wide data sets.