Data Science Capstone Project

Segmenting and Clustering Affordable Rental Projects in San Francisco

1. Introduction: Business Problem

Let assume that your boss request you to have a business trip at San Francisco for 3 months with a limited budget.

There would be a thousand rental projects in San Francisco is waiting for you. There for you have a look to the listed affordable rental project published by Mayor's Office of Housing and Community Development Affordable Rental.

You need to choose the right one which have an affordable payment and nearest the living facility such as hospital/market/gym/park etc… that make you save more money from transportation and movement times.

With the power of Data Science and support from some tools and API, I will try to find a category of optimal affordable rental projects with better living facilities. This report will be targeted to individuals who want to travel to San Francisco for business or holiday with a limit budget.

2. Data

2.1 Source of Data

Based on the business problem, factors that will influence visitor to choose are:

1) Number of existing facilities around each project

2) Type of existing facilities around each project

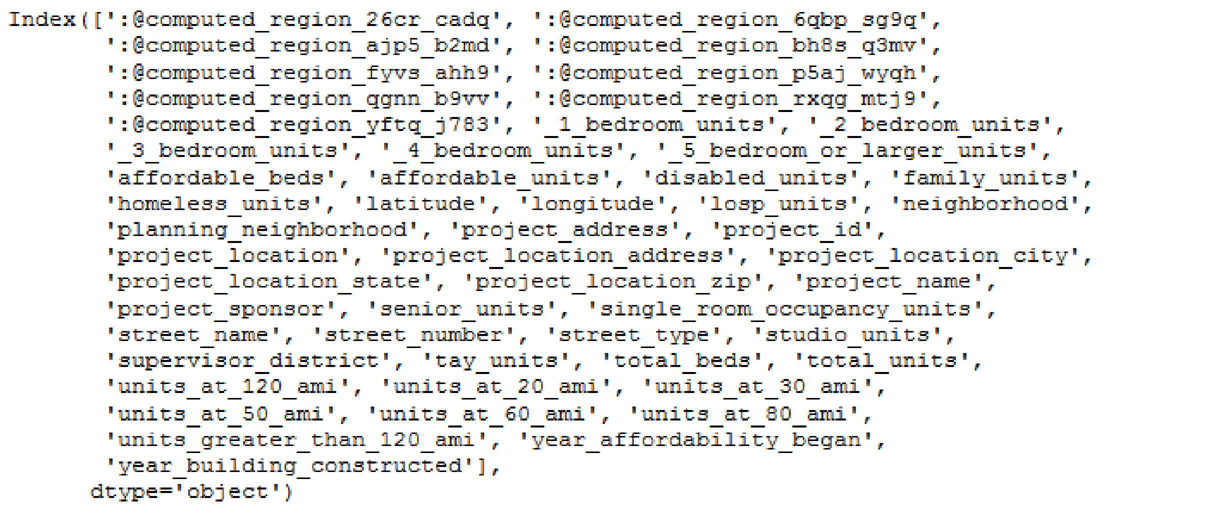
Following data sources will be needed to generate the proper decision:

1) Basic information (project name/location/ Year Affordability Began) of all the affordable rental projects, which can be get from open data website of San Francisco GOV (https://data.sfgov.org/)

2) Number of existing facilities and their type and location in every neighborhood will be obtained using Foursquare API

2.2 Download and Explore Dataset

As soon as the business problem is defined, we need to download the dataset and explore it. The dataset can be get on the San Francisco government open data website (https://data.sfgov.org/resource/yd5s-bd6e.csv). API is provided for programmatic access to this dataset including the ability to filter, query, and aggregate data. After the data is downloaded, read it into a pandas dataframe. Take a quick look at the data, there 354 rows and 52 columns, columns as below.

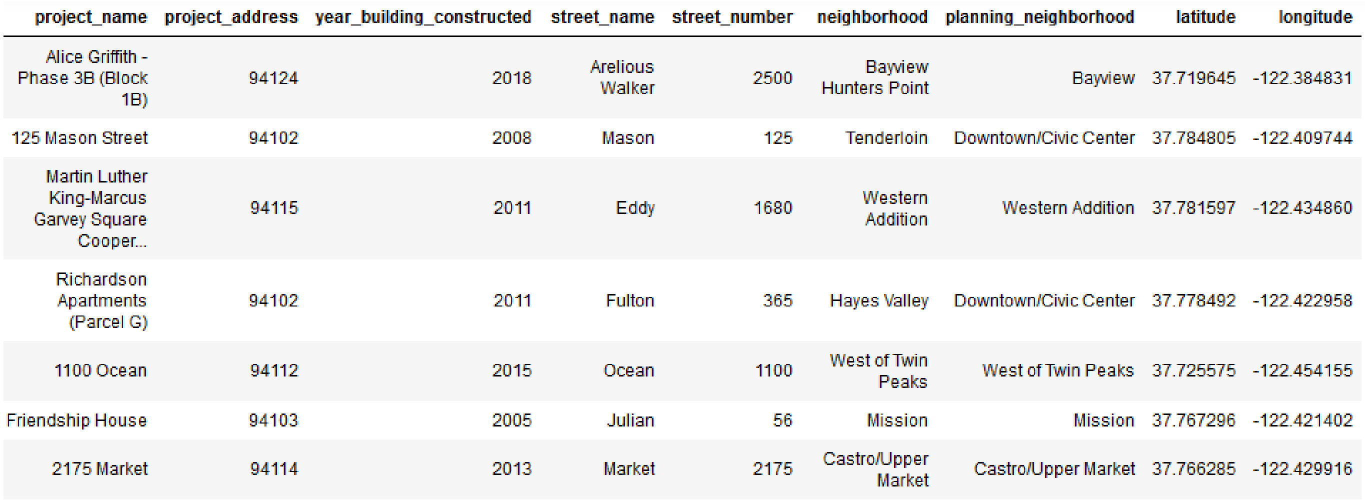


```
Index([':@computed_region_26cr_cadq', ':@computed_region_6qbp_sg9q',
       ':@computed_region_ajp5_b2md', ':@computed_region_bh8s_q3mv',
       ':@computed_region_fyvs_ahh9', ':@computed_region_p5aj_wyqh',
       ':@computed_region_qgnn_b9vv', ':@computed_region_rxqg_mtj9',
       ':@computed_region_yftq_j783', '_1_bedroom_units', '_2_bedroom_units',
       '_3_bedroom_units', '_4_bedroom_units', '_5_bedroom_or_larger_units',
       'affordable_beds', 'affordable_units', 'disabled_units', 'family_units',
       'homeless_units', 'latitude', 'longitude', 'losp_units', 'neighborhood',
       'planning_neighborhood', 'project_address', 'project_id',
       'project_location', 'project_location_address', 'project_location_city',
       'project_location_state', 'project_location_zip', 'project_name',
       'project_sponsor', 'senior_units', 'single_room_occupancy_units',
       'street_name', 'street_number', 'street_type', 'studio_units',
       'supervisor_district', 'tay_units', 'total_beds', 'total_units',
       'units_at_120_ami', 'units_at_20_ami', 'units_at_30_ami',
       'units_at_50_ami', 'units_at_60_ami', 'units_at_80_ami',
       'units_greater_than_120_ami', 'year_affordability_began',
       'year_building_constructed'],
      dtype='object')
```

Fig. 01 Columns of the raw dataset.

Data dictionary is also provided by the website, which can make it easier to understand the data. As this project target at segmenting and clustering ‘Affordable Rental Projects’ in San Francisco base on the living facilities around, the location information will be of great importance.

Also, the project is for individuals who wants to individuals who want to travel to San Francisco for business or holiday with a limit budget. Datasets include some “NaN” value in Longitude and Latitude, we need to drop the rows which not clear Longitude and Latitude. The house too old include many risk and low quality so we only choose the rental house which have constructed year after 2005 when the pre-process is done, new dataset with information of *'project_name','project_address', 'year_building_constructed','street_name','street_number','neighborhood', 'planning_neighborhood', 'latitude','longitude'* (72 rows and 9 columns) as below.



project_name	project_address	year_building_constructed	street_name	street_number	neighborhood	planning_neighborhood	latitude	longitude
Alice Griffith - Phase 3B (Block 1B)	94124	2018	Arelious Walker	2500	Bayview Hunters Point	Bayview	37.719645	-122.384831
125 Mason Street	94102	2008	Mason	125	Tenderloin	Downtown/Civic Center	37.784805	-122.409744
Martin Luther King-Marcus Garvey Square Cooper...	94115	2011	Eddy	1680	Western Addition	Western Addition	37.781597	-122.434860
Richardson Apartments (Parcel G)	94102	2011	Fulton	365	Hayes Valley	Downtown/Civic Center	37.778492	-122.422958
1100 Ocean	94112	2015	Ocean	1100	West of Twin Peaks	West of Twin Peaks	37.725575	-122.454155
Friendship House	94103	2005	Julian	56	Mission	Mission	37.767296	-122.421402
2175 Market	94114	2013	Market	2175	Castro/Upper Market	Castro/Upper Market	37.766285	-122.429916

Fig. 02 Basic information data of ‘Affordable Rental Projects’ after preprocessing

Using Folium library, we can also be used to visualize geographic information of all these projects. And I created a map of San Francisco with ‘Affordable Rental Projects’ base on with the latitude and longitude values get from our data to the visual as below:

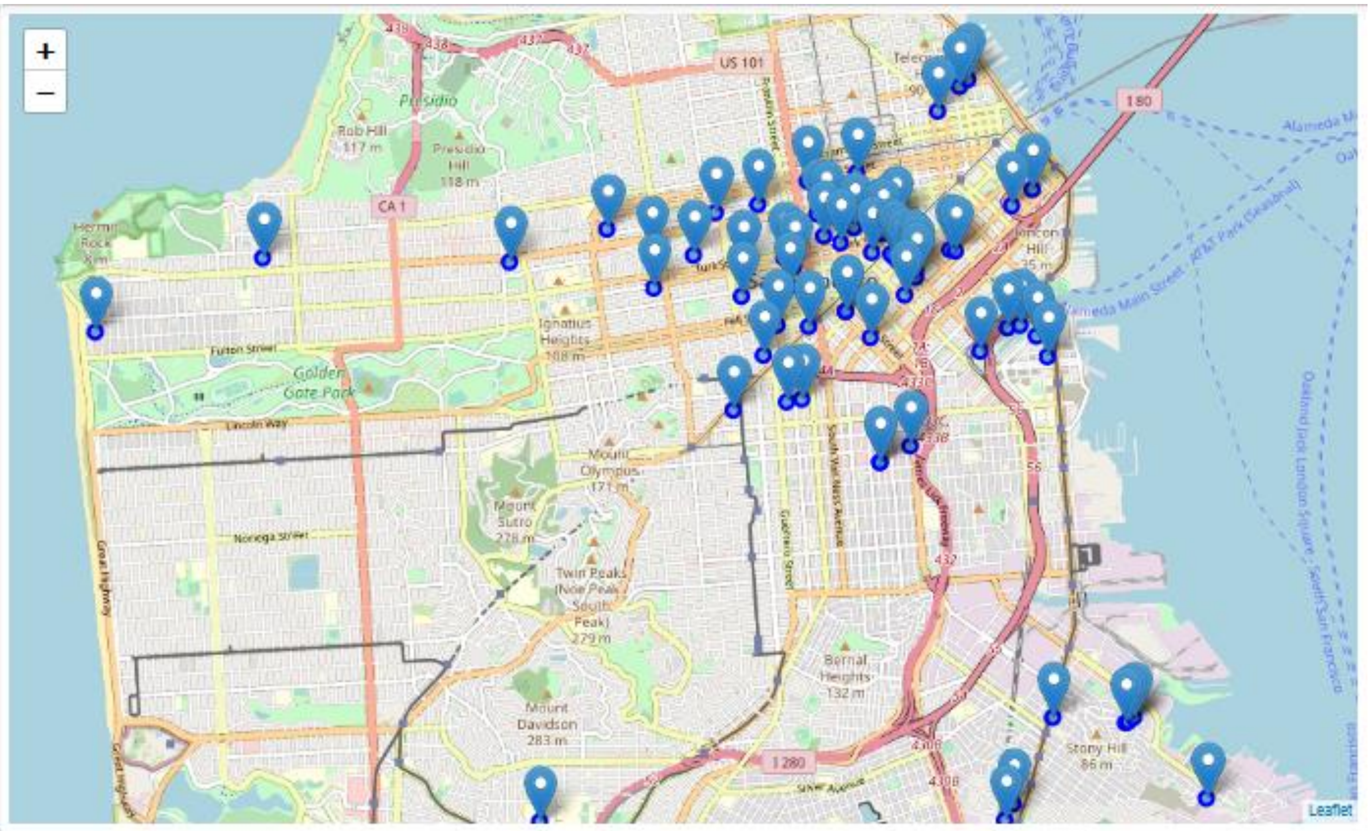


Fig. 03 Geographic visual of ‘Affordable Rental Projects’ using Folium

3. Methodology

In this project, I will try to find a category of optimal affordable rental projects with better living facilities (such as hospital, gym, market, restaurant etc…). This report will be targeted to individuals who want to travel to San Francisco for business or holiday with a limit budget.

1) The first step should be defining the business problem; we already have done that in Introduction

2) The second step should be download the data and explore it, as we have done in Data. My raw data almost has all the information I need for the analysis, such as 'project_name /project_address/year_building_constructed/street_name/street_number/neighborhood /planning_neighborhood/longitude/latitude'. In this step, I pre-processed the data, as the suggestion is year_building_constructed must not be too old because it's include risk of facilities and have a low quality, so 'year_building_constructed' should be before "2005". Also, a map of SanFran with markers was created using latitude and longitude values of the affordable rental projects.

3) The Third step is exploring neighborhoods of each affordable rental projects in San Francisco. We collect all facilities near the rental projects with the support of FourSquare API and visulize them on the map base on the Folium

4) The final step, cluster the all the affordable rental projects with K-means.

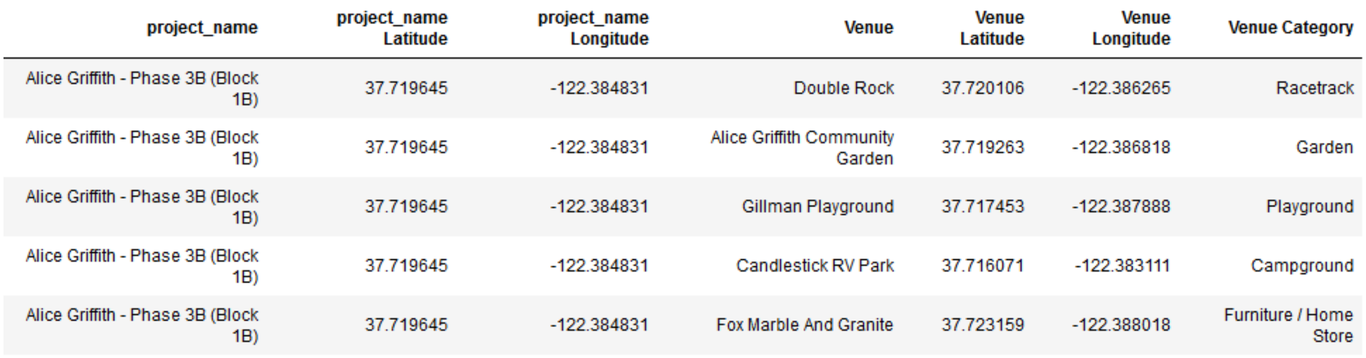
- ✓ According to all the venue data from step 3, I will apply unsupervised learning algorithm K-means to cluster the all the affordable rental projects, and analysis the advantages of each category to help individuals choose the best one they think.
- ✓ I will also visualize geographic details of each cluster, which should be a starting point for individuals to explore and search for optimal affordable rental projects.

4. Analysis

4.1 Analyze Each Project

This project target to explore a category of optimal affordable Rental projects in San Francisco, to help visitor choose their best. As it is widely believed that a mature residential area should be equipped with a range of living facilities, such as restaurants/gyms/markets/hospitals etc....the distance to these living facilities is one of the most important factors that influence to their decision.

We will obtain number of existing facilities and their type and location in every affordable rental project with Foursquare API with a limit as 100 venues and the radius 800 meter for each project from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.

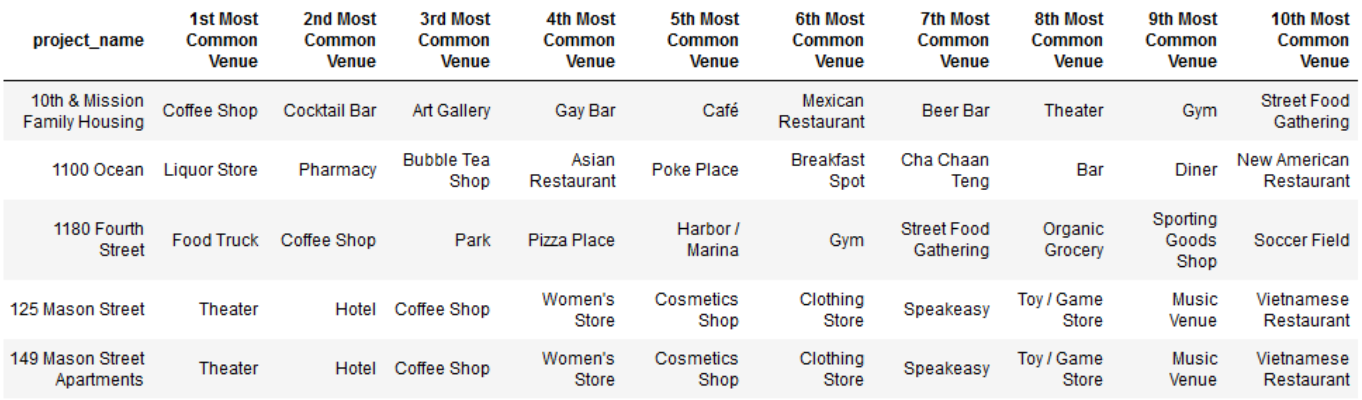


project_name	project_name Latitude	project_name Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Alice Griffith - Phase 3B (Block 1B)	37.719645	-122.384831	Double Rock	37.720106	-122.386265	Racetrack
Alice Griffith - Phase 3B (Block 1B)	37.719645	-122.384831	Alice Griffith Community Garden	37.719263	-122.386818	Garden
Alice Griffith - Phase 3B (Block 1B)	37.719645	-122.384831	Gillman Playground	37.717453	-122.387888	Playground
Alice Griffith - Phase 3B (Block 1B)	37.719645	-122.384831	Candlestick RV Park	37.716071	-122.383111	Campground
Alice Griffith - Phase 3B (Block 1B)	37.719645	-122.384831	Fox Marble And Granite	37.723159	-122.388018	Furniture / Home Store

Fig. 04 Some Venues around a ‘Affordable Rental Projects’

We can also check how many venues were returned for each project and group rows by neighborhood and by taking the mean of the frequency of occurrence of each category.

Print each project along with the top 10 most common venues, and put the data into a new dataframe as below.



project_name	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
10th & Mission Family Housing	Coffee Shop	Cocktail Bar	Art Gallery	Gay Bar	Café	Mexican Restaurant	Beer Bar	Theater	Gym	Street Food Gathering
1100 Ocean	Liquor Store	Pharmacy	Bubble Tea Shop	Asian Restaurant	Poke Place	Breakfast Spot	Cha Chaan Teng	Bar	Diner	New American Restaurant
1180 Fourth Street	Food Truck	Coffee Shop	Park	Pizza Place	Harbor / Marina	Gym	Street Food Gathering	Organic Grocery	Sporting Goods Shop	Soccer Field
125 Mason Street	Theater	Hotel	Coffee Shop	Women's Store	Cosmetics Shop	Clothing Store	Speakeasy	Toy / Game Store	Music Venue	Vietnamese Restaurant
149 Mason Street Apartments	Theater	Hotel	Coffee Shop	Women's Store	Cosmetics Shop	Clothing Store	Speakeasy	Toy / Game Store	Music Venue	Vietnamese Restaurant

Fig. 05 Most common venues around each project

4.2 Cluster Projects

According to all the venue data above, I will focus on using unsupervised learning K-means algorithm to cluster the all the affordable rental projects, and analysis the advantages of each category to help visitor choose their best.
First, I will find the best K with Elbow criterion, and it suggested me the 5 cluster for optimum k of the K-Means.

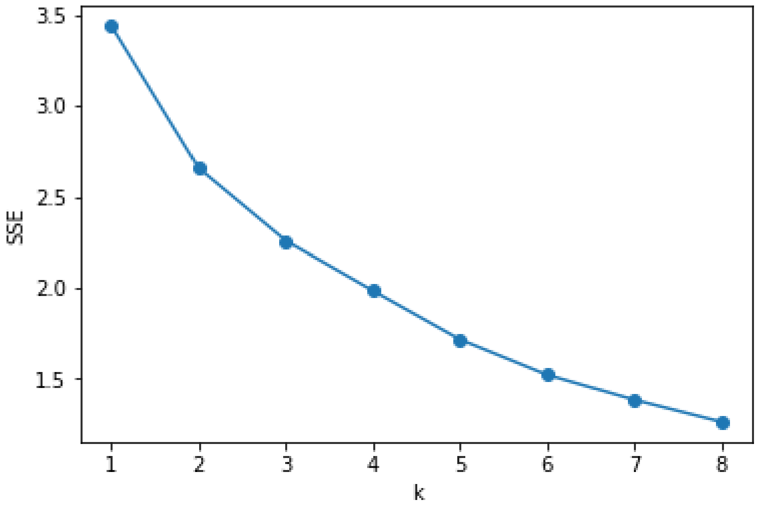
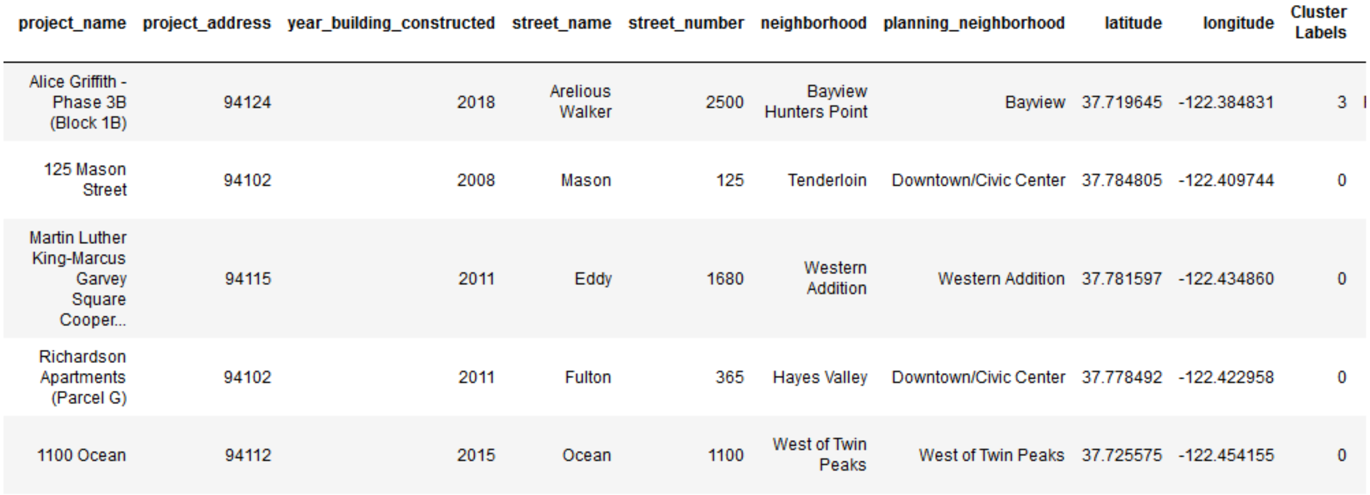


Fig. 06 Apply Elbow criterion to find the best K for our K-Means

Base on our chart, we will confuse between K = 4 or K = 5 so I choose K = 5 and attached the label of each rental project to our data table as below:



project_name	project_address	year_building_constructed	street_name	street_number	neighborhood	planning_neighborhood	latitude	longitude	Cluster Labels
Alice Griffith - Phase 3B (Block 1B)	94124	2018	Arelious Walker	2500	Bayview Hunters Point	Bayview	37.719645	-122.384831	3
125 Mason Street	94102	2008	Mason	125	Tenderloin	Downtown/Civic Center	37.784805	-122.409744	0
Martin Luther King-Marcus Garvey Square Cooper...	94115	2011	Eddy	1680	Western Addition	Western Addition	37.781597	-122.434860	0
Richardson Apartments (Parcel G)	94102	2011	Fulton	365	Hayes Valley	Downtown/Civic Center	37.778492	-122.422958	0
1100 Ocean	94112	2015	Ocean	1100	West of Twin Peaks	West of Twin Peaks	37.725575	-122.454155	0

Fig. 07 Attached cluster labels to the table project

Again, with the support of Folium, I visualized geographic details of each cluster, which should be a starting point for individuals to explore and search for optimal affordable rental projects.

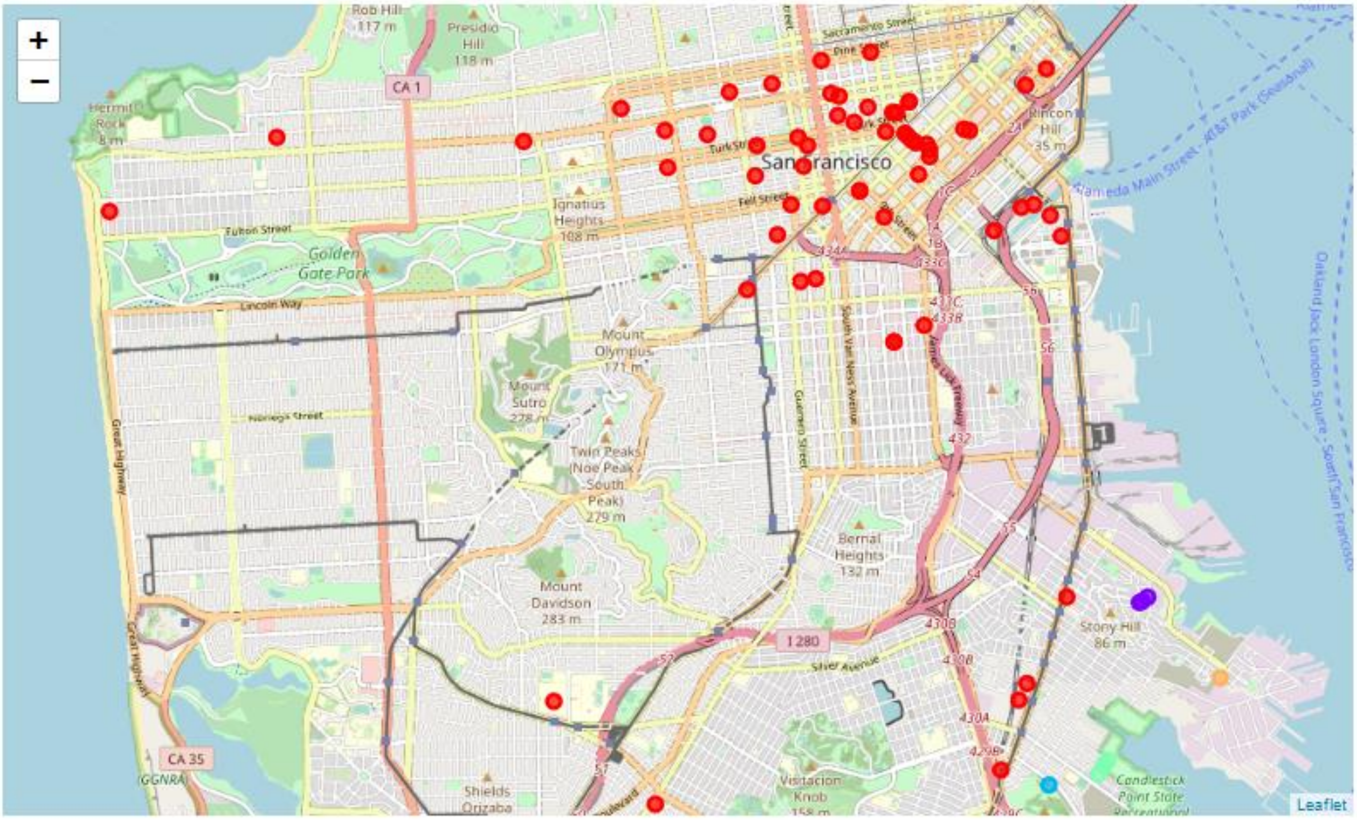
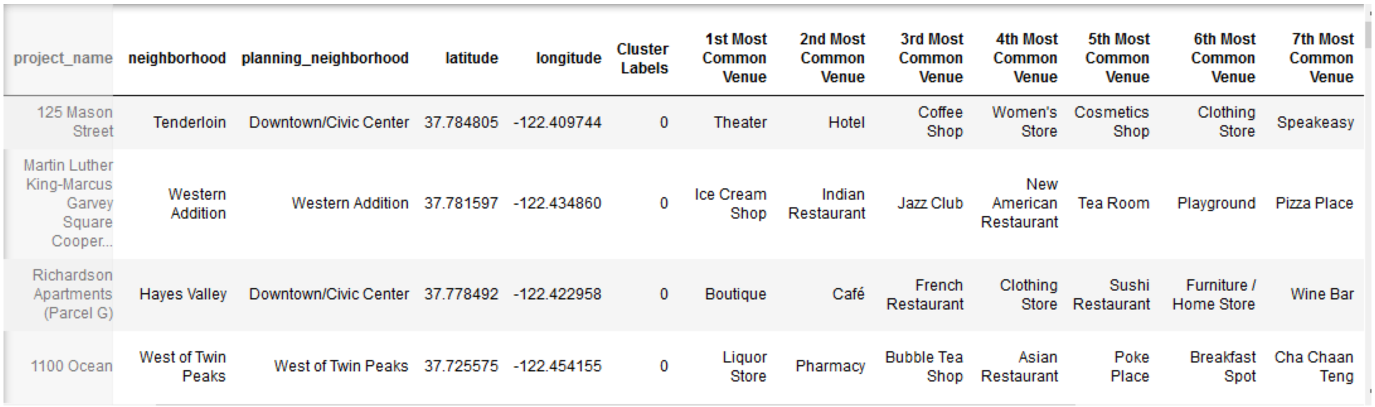


Fig. 08 Visualization of Rental project after attached cluster labels using Folium

4.3 Examine Clusters

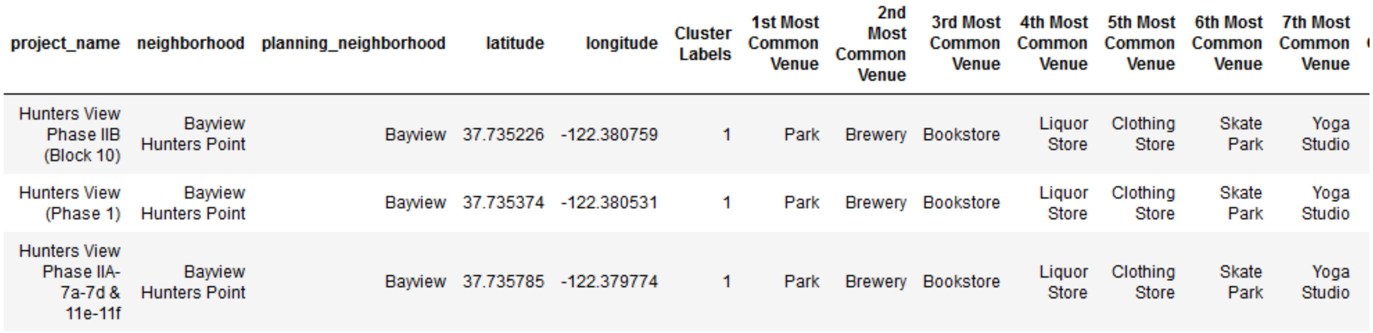
After the K-means algorithm was applied, all the affordable rental projects were divided into 5 clusters:

1) Cluster 1 contains 64 affordable rental projects, top 10 Most Common Venue mainly contains Restaurant/Coffee Shop/Tea Room/ Wine Bar…, it looks like a Restaurants areas suitable for visitor who love to tried new food and dishes or teenager to explore new culture.



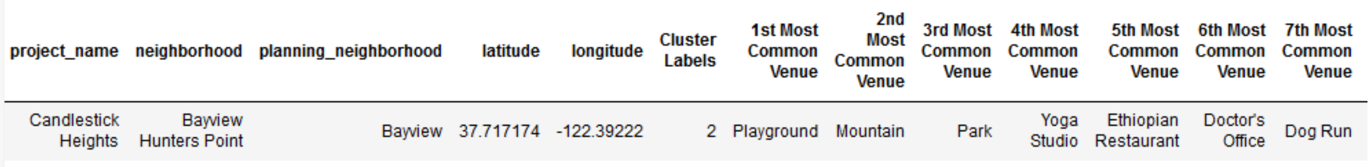
project_name	neighborhood	planning_neighborhood	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
125 Mason Street	Tenderloin	Downtown/Civic Center	37.784805	-122.409744	0	Theater	Hotel	Coffee Shop	Women's Store	Cosmetics Shop	Clothing Store	Speakeasy
Martin Luther King-Marcus Garvey Square Cooper...	Western Addition	Western Addition	37.781597	-122.434860	0	Ice Cream Shop	Indian Restaurant	Jazz Club	New American Restaurant	Tea Room	Playground	Pizza Place
Richardson Apartments (Parcel G)	Hayes Valley	Downtown/Civic Center	37.778492	-122.422958	0	Boutique	Café	French Restaurant	Clothing Store	Sushi Restaurant	Furniture / Home Store	Wine Bar
1100 Ocean	West of Twin Peaks	West of Twin Peaks	37.725575	-122.454155	0	Liquor Store	Pharmacy	Bubble Tea Shop	Asian Restaurant	Poke Place	Breakfast Spot	Cha Chaan Teng

2) Cluster 2 contains 3 affordable rental projects, this cluster contains most projects, top 10 Most Common Venue mainly contains Park /Brewery /Bookstore /Liquor Store /Clothing Store /Skate Park /Yoga Studio /Donut Shop /Dumpling Restaurant /Electronics Store etc.., living facilities are common for daily life, but it may be good for those who work here.



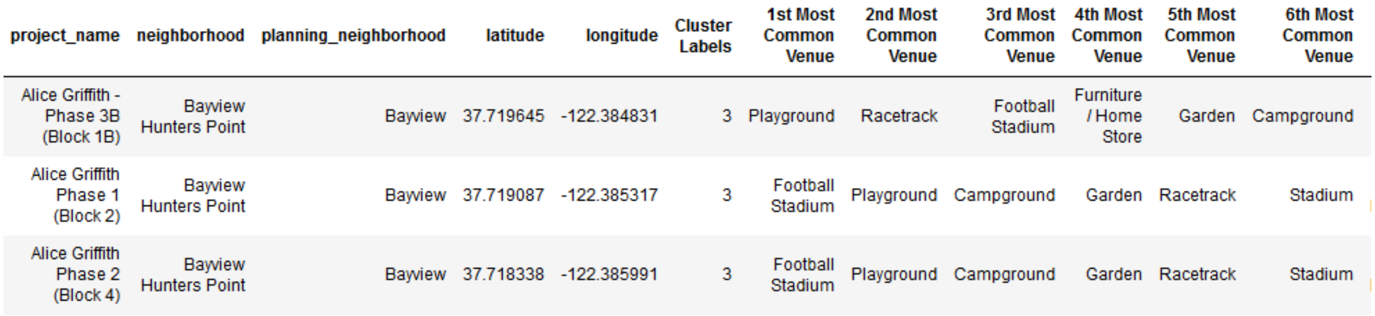
project_name	neighborhood	planning_neighborhood	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
Hunters View Phase IIB (Block 10)	Bayview Hunters Point	Bayview	37.735226	-122.380759	1	Park	Brewery	Bookstore	Liquor Store	Clothing Store	Skate Park	Yoga Studio
Hunters View (Phase 1)	Bayview Hunters Point	Bayview	37.735374	-122.380531	1	Park	Brewery	Bookstore	Liquor Store	Clothing Store	Skate Park	Yoga Studio
Hunters View Phase IIA-7a-7d & 11e-11f	Bayview Hunters Point	Bayview	37.735785	-122.379774	1	Park	Brewery	Bookstore	Liquor Store	Clothing Store	Skate Park	Yoga Studio

3) Cluster 3 contains 1 affordable rental projects, top 10 Most Common Venue mainly contains Playground /Mountain /Park /Yoga Studio /Ethiopian Restaurant /Doctor's Office /Dog Run /Donut Shop /Restaurant /Electronics Store, it seems like this place quite far from the city and suitable for someone like peacful and fresh air.



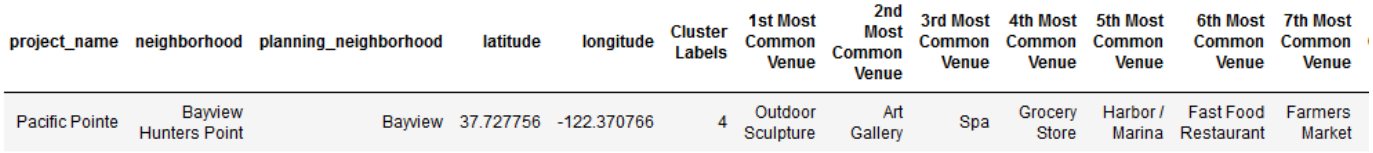
project_name	neighborhood	planning_neighborhood	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
Candlestick Heights	Bayview Hunters Point	Bayview	37.717174	-122.39222	2	Playground	Mountain	Park	Yoga Studio	Ethiopian Restaurant	Doctor's Office	Dog Run

4) Cluster 4 contains 3 affordable rental projects, top 10 Most Common Venue mainly contains Playground /Racetrack /Football Stadium /Furniture / Home Store /Garden /Campground /Coworking Space /Field etc… this cluster is community areas, maybe someone who love sport or out door activities will choose this cluster.



project_name	neighborhood	planning_neighborhood	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
Alice Griffith - Phase 3B (Block 1B)	Bayview Hunters Point	Bayview	37.719645	-122.384831	3	Playground	Racetrack	Football Stadium	Furniture / Home Store	Garden	Campground
Alice Griffith Phase 1 (Block 2)	Bayview Hunters Point	Bayview	37.719087	-122.385317	3	Football Stadium	Playground	Campground	Garden	Racetrack	Stadium
Alice Griffith Phase 2 (Block 4)	Bayview Hunters Point	Bayview	37.718338	-122.385991	3	Football Stadium	Playground	Campground	Garden	Racetrack	Stadium

5) Cluster 5 contains 1 affordable rental projects, this cluster contains most projects, top 10 Most Common Venue mainly contains Spa /Grocery Store /Harbor Marina /Fast Food Restaurant /Farmers Market /Exhibit /Event Space /Ethiopian Restaurant/ Outdoor Sculpture /Art Gallery etc.., this place maybe suitable for artist, who loves the art or some exhibit.



project_name	neighborhood	planning_neighborhood	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	7th Most Common Venue
Pacific Pointe	Bayview Hunters Point	Bayview	37.727756	-122.370766	4	Outdoor Sculpture	Art Gallery	Spa	Grocery Store	Harbor / Marina	Fast Food Restaurant	Farmers Market

5. Results and Discussion

Although there are 354 affordable Rental projects in San Francisco, only 72 projects which have the constructed year after 2005. As it is widely believed that some mature residential areas should be equipped with a range of living facilities, such as restaurants/gyms/markets/hospitals etc... Except Rental Price, living facilities is one of the most important factors that influence the finally decision.

This analysis shows that there are 301 uniques venue categories around all the projects, Top 10 Most Common Venue list above mainly relate to Food/Sports/Art/Leisure/Public Transport/Market etc...

As we can see from 4.3 Examine Clusters, after the K-means algorithm was applied, all the affordable rental projects were divided into 5 clusters:

1) Cluster 1 contains 64 affordable rental projects, top 10 Most Common Venue mainly contains Restaurant/Coffee Shop/Tea Room/ Wine Bar…, it looks like a Restaurants areas suitable for visitor who love to tried new food and dishes or teenager to explore new culture.
2) Cluster 2 contains 3 affordable rental projects, this cluster contains most projects, top 10 Most Common Venue mainly contains Park /Brewery /Bookstore /Liquor Store /Clothing Store /Skate Park /Yoga Studio /Donut Shop /Dumpling Restaurant /Electronics Store etc.., living facilities are common for daily life, but it may be good for those who work here.

3) Cluster 3 contains 1 affordable rental projects, top 10 Most Common Venue mainly contains Playground /Mountain /Park /Yoga Studio /Ethiopian Restaurant /Doctor's Office /Dog Run /Donut Shop /Restaurant /Electronics Store, it seems like this place quite far from the city and suitable for someone like peacful and fresh air.
4) Cluster 4 contains 3 affordable rental projects, top 10 Most Common Venue mainly contains Playground /Racetrack /Football Stadium /Furniture / Home Store /Garden /Campground /Coworking Space /Field etc… this cluster is community areas, maybe someone who love sport or out door activities will choose this cluster.
5) Cluster 5 contains 1 affordable rental projects, this cluster contains most projects, top 10 Most Common Venue mainly contains Spa /Grocery Store /Harbor Marina /Fast Food Restaurant /Farmers Market /Exhibit /Event Space /Ethiopian Restaurant/ Outdoor Sculpture /Art Gallery etc.., this place maybe suitable for artist, who loves the art or some exhibit.

6. Conclusion

Purpose of this project is try to find a category of optimal affordable Rental projects with better living facilities. Target to individuals who want to travel to San Francisco for business or holiday with a limit budget. As it is widely believed that a most of residential areas should be equipped with a range of living facilities, such as restaurants/gyms/markets/hospitals etc...

The Foursquare location data was leveraged to compare each project to provide reliable suggestions for individuals who want to choose some place with better living facilities. With unsupervised learning K-means algorithm, all the affordable Rental projects were clustered in to 5 categories, the advantages of each category was expressed to help individuals choose their best one.

This Project simply processed the Rental affordable rental projects data, and cluster them into 5 categories based one the living facilities data, the results can only help individuals choose the place they want to rent with affordable price and better living facilities. Further analysis can be done base on these 5 clusters, which can help provide more detail information to clarify the advantages of each category.