Capstone Project

Segmenting and Clustering Affordable Housing Projects in San Francisco

1. Introduction: Business Problem

Lots of people in San Francisco still can not afford their own houses. Hundreds of affordable housing pipeline projects were published by Mayor’s Office of Housing and Community Development (MOHCD) and the Office of Community Investment and Infrastructure (OCII). The projects listed are in the process of development--or are anticipated to be developed--in partnership with non-profit or for-profit developers and financed through city funding agreements, ground leases, disposition and participation agreements and conduit bond financing. The Affordable Housing Pipeline also includes housing units produced by private developers through the Inclusionary Affordable Housing Program.

In this project, I will try to find a category of optimal affordable housing projects with better living facilities. This report will be targeted to individuals who want to have their own home but can not afford commodity housing in San Francisco.

Since there are over hundreds of projects in San Francisco, which one will be the right one. 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... So I am going to leverage the Foursquare location data to compare each project to provide reliable suggestions for individuals who need a place to live.

Data science powers we learnt these several weeks will be used to generate a few most promising projects based on this criteria. I will cluster all the affordable housing programs into several categories, advantages of each category will then be clearly expressed to help individuals make their first-step decisions.

2. Data

2.1 Source of Data

Based on the business problem, factors that will influence individuals to make the first-step decision are:

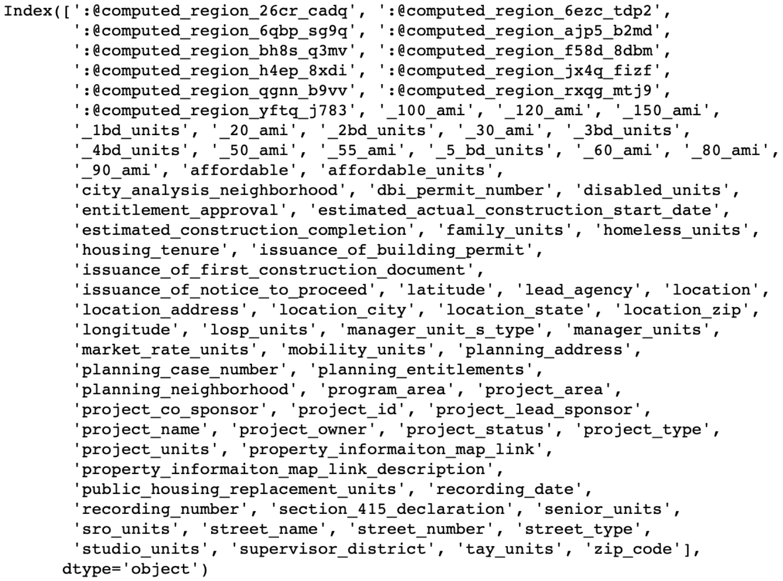
1) Number of existing facilities around each project
2) Type of existing facilities around each project

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

1) Basic information (project name/location/housing tenure) of all the affordable housing programs, 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

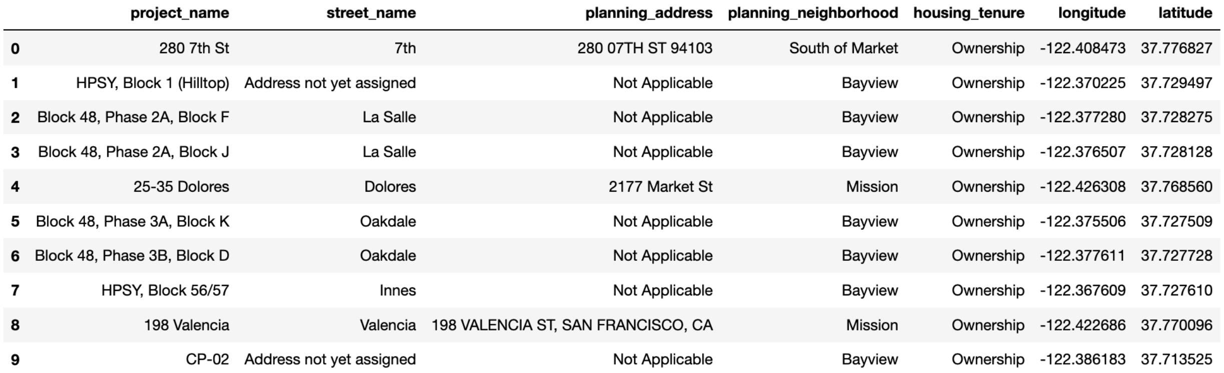
As soon as the business problem is defined, we need to download the dataset and explore it. The dataset can be get on the San Francisco government open data website (https://data.sfgov.org/Housing-and-Buildings/Affordable-Housing-Pipeline/aaxw-2cb8). API is provided for programmatic access to this dataset 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 329 rows and 82 columns, columns as below.



```
Index([':@computed_region_26cr_cadq', ':@computed_region_6ezc_tdp2',
       ':@computed_region_6qbp_sg9q', ':@computed_region_ajp5_b2md',
       ':@computed_region_bh8s_q3mv', ':@computed_region_f58d_8dbm',
       ':@computed_region_h4ep_8xdi', ':@computed_region_jx4q_fizf',
       ':@computed_region_qgnn_b9vv', ':@computed_region_rxqg_mtj9',
       ':@computed_region_yftq_j783', '_100_ami', '_120_ami', '_150_ami',
       '_1bd_units', '_20_ami', '_2bd_units', '_30_ami', '_3bd_units',
       '_4bd_units', '_50_ami', '_55_ami', '_5_bd_units', '_60_ami', '_80_ami',
       '_90_ami', 'affordable', 'affordable_units',
       'city_analysis_neighborhood', 'dbi_permit_number', 'disabled_units',
       'entitlement_approval', 'estimated_actual_construction_start_date',
       'estimated_construction_completion', 'family_units', 'homeless_units',
       'housing_tenure', 'issuance_of_building_permit',
       'issuance_of_first_construction_document',
       'issuance_of_notice_to_proceed', 'latitude', 'lead_agency', 'location',
       'location_address', 'location_city', 'location_state', 'location_zip',
       'longitude', 'losp_units', 'manager_unit_s_type', 'manager_units',
       'market_rate_units', 'mobility_units', 'planning_address',
       'planning_case_number', 'planning_entitlements',
       'planning_neighborhood', 'program_area', 'project_area',
       'project_co_sponsor', 'project_id', 'project_lead_sponsor',
       'project_name', 'project_owner', 'project_status', 'project_type',
       'project_units', 'property_informaiton_map_link',
       'property_informaiton_map_link_description',
       'public_housing_replacement_units', 'recording_date',
       'recording_number', 'section_415_declaration', 'senior_units',
       'sro_units', 'street_name', 'street_number', 'street_type',
       'studio_units', 'supervisor_district', 'tay_units', 'zip_code'],
      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 Housing 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 want to buy a house, the housing tenure should be 'Ownership', when the pre-process is done, new dataset with information of ***'project name', 'street name', 'planning address', 'planning neighborhood', 'housing tenure', 'longitude', 'latitude'*** (75 rows and 7 columns) as below.



	project_name	street_name	planning_address	planning_neighborhood	housing_tenure	longitude	latitude
0	280 7th St	7th	280 07TH ST 94103	South of Market	Ownership	-122.408473	37.776827
1	HPSY, Block 1 (Hilltop)	Address not yet assigned	Not Applicable	Bayview	Ownership	-122.370225	37.729497
2	Block 48, Phase 2A, Block F	La Salle	Not Applicable	Bayview	Ownership	-122.377280	37.728275
3	Block 48, Phase 2A, Block J	La Salle	Not Applicable	Bayview	Ownership	-122.376507	37.728128
4	25-35 Dolores	Dolores	2177 Market St	Mission	Ownership	-122.426308	37.768560
5	Block 48, Phase 3A, Block K	Oakdale	Not Applicable	Bayview	Ownership	-122.375506	37.727509
6	Block 48, Phase 3B, Block D	Oakdale	Not Applicable	Bayview	Ownership	-122.377611	37.727728
7	HPSY, Block 56/57	Innes	Not Applicable	Bayview	Ownership	-122.367609	37.727610
8	198 Valencia	Valencia	198 VALENCIA ST, SAN FRANCISCO, CA	Mission	Ownership	-122.422686	37.770096
9	CP-02	Address not yet assigned	Not Applicable	Bayview	Ownership	-122.386183	37.713525

Fig. 02 Basic information data of 'Affordable Housing Projects'

Python folium library can also be used to visualize geographic information of all these programs. And I created a map of San Francisco with 'Affordable Housing Projects' on top with the latitude and longitude values, the visual as below:

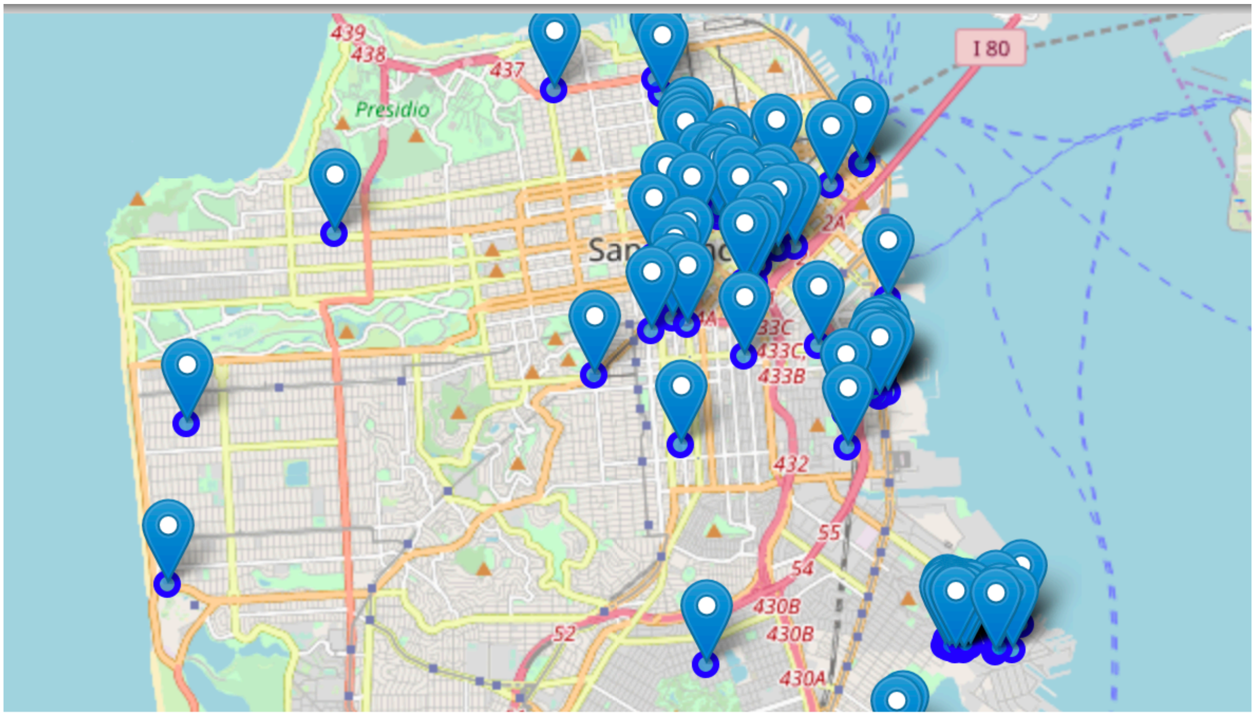


Fig. 03 Geographic visual of 'Affordable Housing Projects'

3. Methodology

In this project, I am going to explore a category of optimal affordable housing projects in San Francisco with better living facilities (such as restaurants/gyms/markets/hospitals etc..), to help individuals make the first-step decision.

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

2) The second step should be downloaded 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 /street name /planning address /planning neighborhood /housing tenure /longitude /latitude', especially the location information, which indeed do me a great favor.
    - In this step, I pre-processed the data, as the suggestion is for individuals who want to buy a house, so 'housing tenure' should be 'Ownership'.
    - Also, a map of San Francisco with markers was created using latitude and longitude values of the affordable housing projects.

3) The Third step is to explore neighborhoods of each affordable housing projects in San Francisco.
    - Obtain number of existing facilities and their type and location in every affordable housing project with Foursquare API.

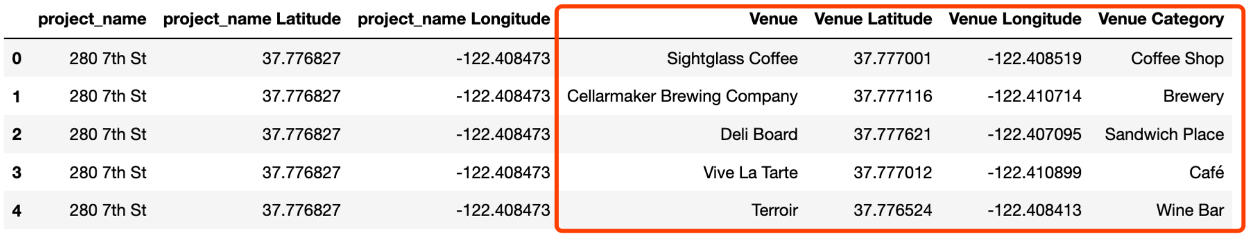
4) The final step, cluster the all the affordable housing projects with K-means.
    - According to all the venue data from step 4, I will focus on using unsupervised learning K-means algorithm to cluster the all the affordable housing projects, and analysis the advantages of each category to help individuals make their first-step decisions.
    - I will also visualize geographic details of each cluster, which should be a starting point for individuals to explore and search for optimal affordable housing projects.

4. Analysis

4.1 Analyze Each Project

This project target to explore a category of optimal affordable housing projects in San Francisco, to help individuals make the first-step decision. 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.... Except Housing Price, living facilities is one of the most important factors that influence the final decision.

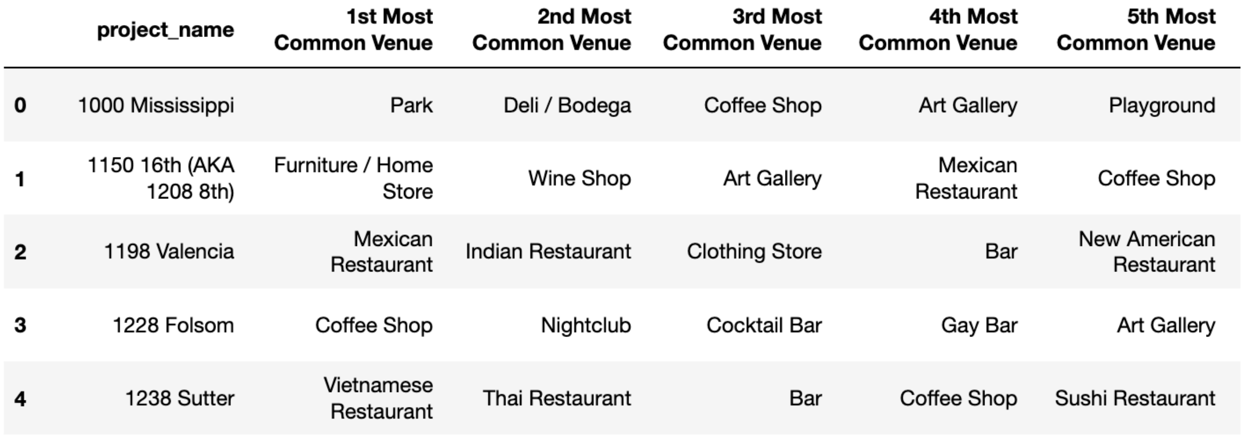
We will obtain number of existing facilities and their type and location in every affordable housing project with Foursquare API with a limit as 100 venue and the radius 500 meter for each program from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.



	project_name	project_name Latitude	project_name Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	280 7th St	37.776827	-122.408473	Sightglass Coffee	37.777001	-122.408519	Coffee Shop
1	280 7th St	37.776827	-122.408473	Cellarmaker Brewing Company	37.777116	-122.410714	Brewery
2	280 7th St	37.776827	-122.408473	Deli Board	37.777621	-122.407095	Sandwich Place
3	280 7th St	37.776827	-122.408473	Vive La Tarte	37.777012	-122.410899	Café
4	280 7th St	37.776827	-122.408473	Terroir	37.776524	-122.408413	Wine Bar

Fig. 04 Venues around 'Affordable Housing 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
0	1000 Mississippi	Park	Deli / Bodega	Coffee Shop	Art Gallery	Playground
1	1150 16th (AKA 1208 8th)	Furniture / Home Store	Wine Shop	Art Gallery	Mexican Restaurant	Coffee Shop
2	1198 Valencia	Mexican Restaurant	Indian Restaurant	Clothing Store	Bar	New American Restaurant
3	1228 Folsom	Coffee Shop	Nightclub	Cocktail Bar	Gay Bar	Art Gallery
4	1238 Sutter	Vietnamese Restaurant	Thai Restaurant	Bar	Coffee Shop	Sushi Restaurant

Fig. 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 housing projects, and analysis the advantages of each category to help individuals make their first-step decisions.

First, I will find the best K with Elbow criterion, and it suggested me the 4 degree for optimum k of the K-Means.

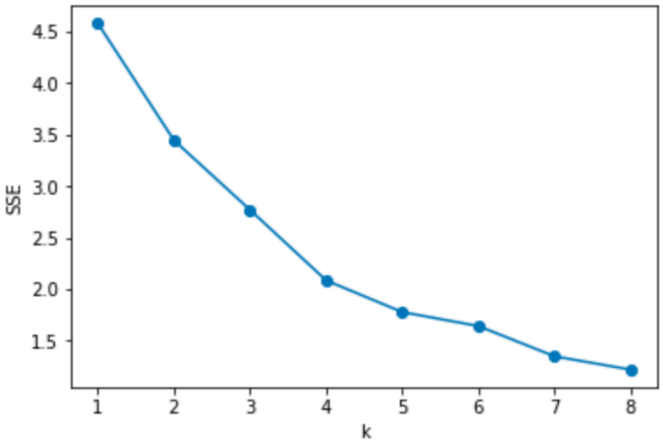
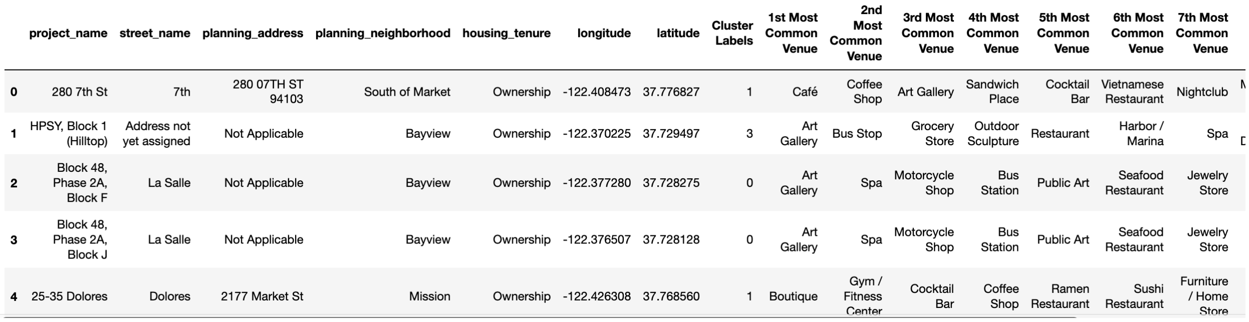


Fig. 06 Best K with Elbow criterion

Below is the merged table with cluster labels for each program.



	project_name	street_name	planning_address	planning_neighborhood	housing_tenure	longitude	latitude	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
0	280 7th St	7th	280 07TH ST 94103	South of Market	Ownership	-122.408473	37.776827	1	Café	Coffee Shop	Art Gallery	Sandwich Place	Cocktail Bar	Vietnamese Restaurant	Nightclub
1	HPSY, Block 1 (Hilltop)	Address not yet assigned	Not Applicable	Bayview	Ownership	-122.370225	37.729497	3	Art Gallery	Bus Stop	Grocery Store	Outdoor Sculpture	Restaurant	Harbor / Marina	Spa
2	Block 48, Phase 2A, Block F	La Salle	Not Applicable	Bayview	Ownership	-122.377280	37.728275	0	Art Gallery	Spa	Motorcycle Shop	Bus Station	Public Art	Seafood Restaurant	Jewelry Store
3	Block 48, Phase 2A, Block J	La Salle	Not Applicable	Bayview	Ownership	-122.376507	37.728128	0	Art Gallery	Spa	Motorcycle Shop	Bus Station	Public Art	Seafood Restaurant	Jewelry Store
4	25-35 Dolores	Dolores	2177 Market St	Mission	Ownership	-122.426308	37.768560	1	Boutique	Gym / Fitness Center	Cocktail Bar	Coffee Shop	Ramen Restaurant	Sushi Restaurant	Furniture / Home Store

Fig. 07 Merged table with cluster labels

I also visualized geographic details of each cluster, which should be a starting point for individuals to explore and search for optimal affordable housing projects.

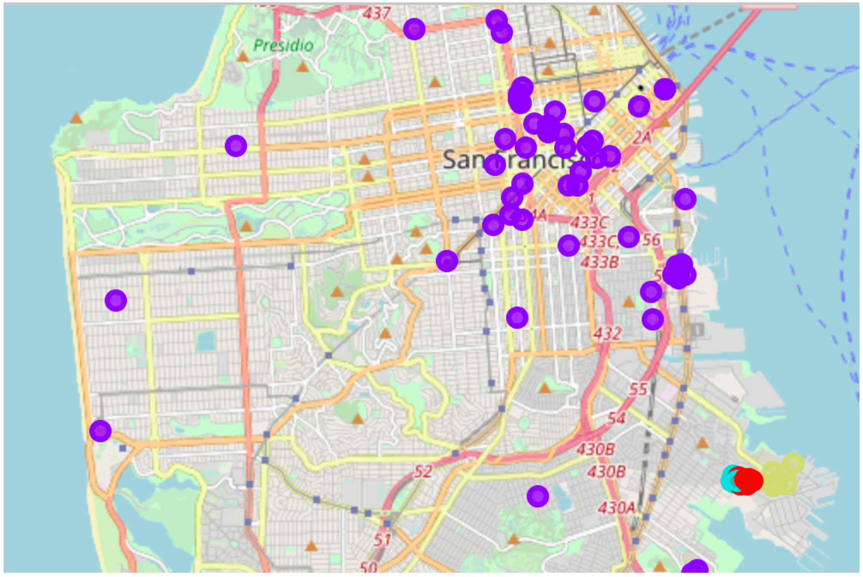
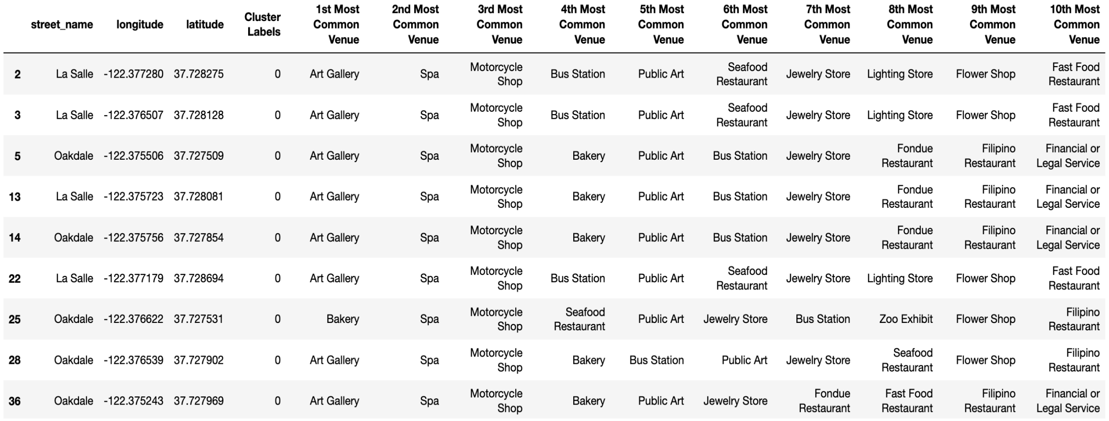


Fig. 08 Geographic details of each cluster

4.3 Examine Clusters

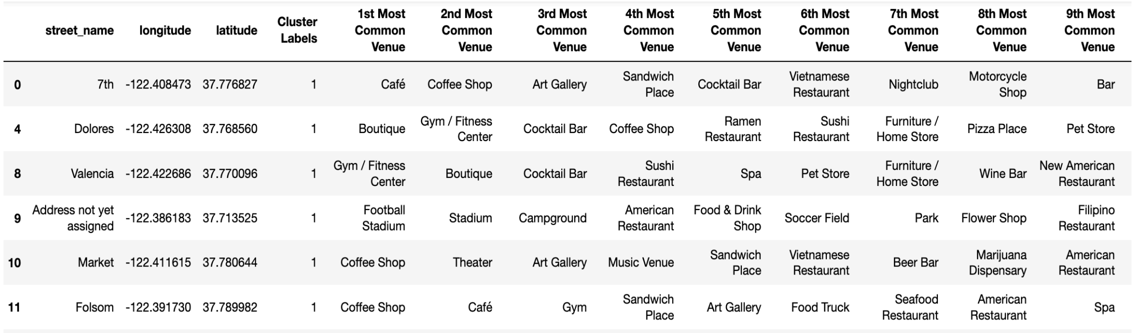
After the K-means algorithm was applied, all the affordable housing pipeline projects were divided into 4 clusters:

1) Cluster 1 contains 12 affordable housing pipeline projects, top 10 Most Common Venue mainly contains Art Gallery /Spa /Bus Station /Public Art /Jewelry Store /Flower Shop, if the individual is an artist or have great interest in art, these areas will be quite good.



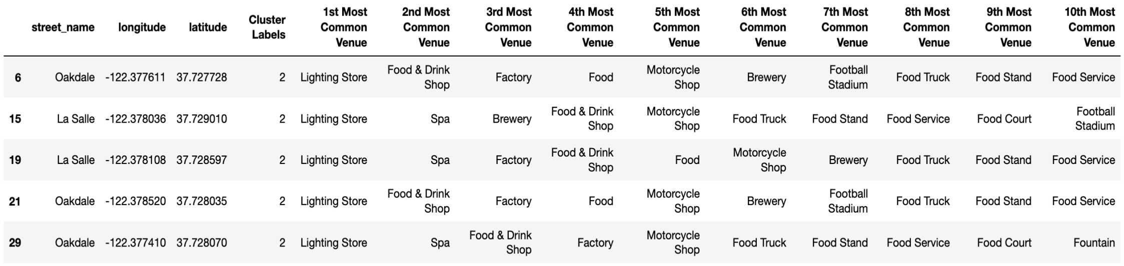
	street_name	longitude	latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	La Salle	-122.377280	37.728275	0	Art Gallery	Spa	Motorcycle Shop	Bus Station	Public Art	Seafood Restaurant	Jewelry Store	Lighting Store	Flower Shop	Fast Food Restaurant
3	La Salle	-122.376507	37.728128	0	Art Gallery	Spa	Motorcycle Shop	Bus Station	Public Art	Seafood Restaurant	Jewelry Store	Lighting Store	Flower Shop	Fast Food Restaurant
5	Oakdale	-122.375506	37.727509	0	Art Gallery	Spa	Motorcycle Shop	Bakery	Public Art	Bus Station	Jewelry Store	Fondue Restaurant	Filipino Restaurant	Financial or Legal Service
13	La Salle	-122.375723	37.728081	0	Art Gallery	Spa	Motorcycle Shop	Bakery	Public Art	Bus Station	Jewelry Store	Fondue Restaurant	Filipino Restaurant	Financial or Legal Service
14	Oakdale	-122.375756	37.727854	0	Art Gallery	Spa	Motorcycle Shop	Bakery	Public Art	Bus Station	Jewelry Store	Fondue Restaurant	Filipino Restaurant	Financial or Legal Service
22	La Salle	-122.377179	37.728694	0	Art Gallery	Spa	Motorcycle Shop	Bus Station	Public Art	Seafood Restaurant	Jewelry Store	Lighting Store	Flower Shop	Fast Food Restaurant
25	Oakdale	-122.376622	37.727531	0	Bakery	Spa	Motorcycle Shop	Seafood Restaurant	Public Art	Jewelry Store	Bus Station	Zoo Exhibit	Flower Shop	Filipino Restaurant
28	Oakdale	-122.376539	37.727902	0	Art Gallery	Spa	Motorcycle Shop	Bakery	Bus Station	Public Art	Jewelry Store	Seafood Restaurant	Flower Shop	Filipino Restaurant
36	Oakdale	-122.375243	37.727969	0	Art Gallery	Spa	Motorcycle Shop	Bakery	Public Art	Jewelry Store	Fondue Restaurant	Fast Food Restaurant	Filipino Restaurant	Financial or Legal Service

2) Cluster 2 contains 53 affordable housing pipeline projects, this cluster contains most projects, top 10 Most Common Venue mainly contains Coffee Shop /Gym /Bar /Sandwich Place /Art Gallery /all kinds of Restaurants /Spa /Park /Theater etc.., almost cover every aspect of our daily life, obviously, it will be very convenient to live in these areas.



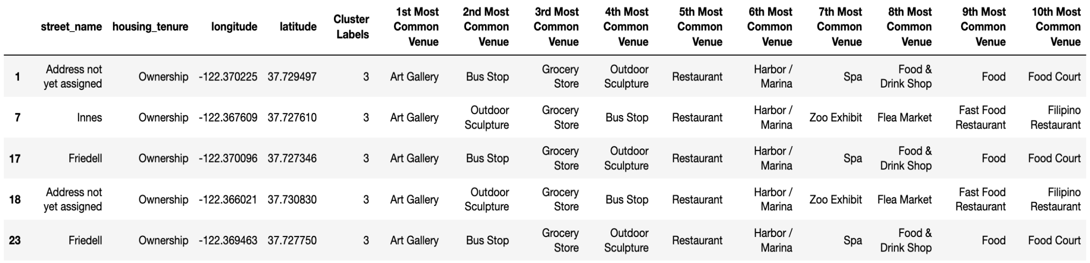
	street_name	longitude	latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	7th	-122.408473	37.776827	1	Café	Coffee Shop	Art Gallery	Sandwich Place	Cocktail Bar	Vietnamese Restaurant	Nightclub	Motorcycle Shop	Bar
4	Dolores	-122.426308	37.768560	1	Boutique	Gym / Fitness Center	Cocktail Bar	Coffee Shop	Ramen Restaurant	Sushi Restaurant	Furniture / Home Store	Pizza Place	Pet Store
8	Valencia	-122.422686	37.770096	1	Gym / Fitness Center	Boutique	Cocktail Bar	Sushi Restaurant	Spa	Pet Store	Furniture / Home Store	Wine Bar	New American Restaurant
9	Address not yet assigned	-122.386183	37.713525	1	Football Stadium	Stadium	Campground	American Restaurant	Food & Drink Shop	Soccer Field	Park	Flower Shop	Filipino Restaurant
10	Market	-122.411615	37.780644	1	Coffee Shop	Theater	Art Gallery	Music Venue	Sandwich Place	Vietnamese Restaurant	Beer Bar	Marijuana Dispensary	American Restaurant
11	Folsom	-122.391730	37.789982	1	Coffee Shop	Café	Gym	Sandwich Place	Art Gallery	Food Truck	Seafood Restaurant	American Restaurant	Spa

3) Cluster 3 contains 5 affordable housing pipeline projects, top 10 Most Common Venue mainly contains Spa /Factory /Food & Drink Shop /Motorcycle Shop /Brewery /Food Truck /Food Stand /Food Service, it seems like these are food producing areas, living facilities are too simple here, but it may be good for those who work here.



	street_name	longitude	latitude	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
6	Oakdale	-122.377611	37.727728	2	Lighting Store	Food & Drink Shop	Factory	Food	Motorcycle Shop	Brewery	Football Stadium	Food Truck	Food Stand	Food Service
15	La Salle	-122.378036	37.729010	2	Lighting Store	Spa	Brewery	Food & Drink Shop	Motorcycle Shop	Food Truck	Food Stand	Food Service	Food Court	Football Stadium
19	La Salle	-122.378108	37.728597	2	Lighting Store	Spa	Factory	Food & Drink Shop	Food	Motorcycle Shop	Brewery	Food Truck	Food Stand	Food Service
21	Oakdale	-122.378520	37.728035	2	Lighting Store	Food & Drink Shop	Factory	Food	Motorcycle Shop	Brewery	Football Stadium	Food Truck	Food Stand	Food Service
29	Oakdale	-122.377410	37.728070	2	Lighting Store	Spa	Food & Drink Shop	Factory	Motorcycle Shop	Food Truck	Food Stand	Food Service	Food Court	Fountain

4) Cluster 4 contains 5 affordable housing pipeline projects, top 10 Most Common Venue mainly contains Art Gallery /Bus Stop /Grocery Store /Outdoor Sculpture /Restaurant /Harbor / Marina /Spa, another cluster closely relate to art, as we can see from the map, this cluster is not far from cluster 1, people who are interest in art can also take these projects into consideration.



	street_name	housing_tenure	longitude	latitude	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	Address not yet assigned	Ownership	-122.370225	37.729497	3	Art Gallery	Bus Stop	Grocery Store	Outdoor Sculpture	Restaurant	Harbor / Marina	Spa	Food & Drink Shop	Food	Food Court
7	Innes	Ownership	-122.367609	37.727610	3	Art Gallery	Outdoor Sculpture	Grocery Store	Bus Stop	Restaurant	Harbor / Marina	Zoo Exhibit	Flea Market	Fast Food Restaurant	Filipino Restaurant
17	Friedell	Ownership	-122.370096	37.727346	3	Art Gallery	Bus Stop	Grocery Store	Outdoor Sculpture	Restaurant	Harbor / Marina	Spa	Food & Drink Shop	Food	Food Court
18	Address not yet assigned	Ownership	-122.366021	37.730830	3	Art Gallery	Outdoor Sculpture	Grocery Store	Bus Stop	Restaurant	Harbor / Marina	Zoo Exhibit	Flea Market	Fast Food Restaurant	Filipino Restaurant
23	Friedell	Ownership	-122.369463	37.727750	3	Art Gallery	Bus Stop	Grocery Store	Outdoor Sculpture	Restaurant	Harbor / Marina	Spa	Food & Drink Shop	Food	Food Court

5. Results and Discussion

Although there are 372 affordable housing pipeline projects in San Francisco, only 75 projects provide ownership housing. As it is widely believed that a mature residential areas should be equipped with a range of living facilities, such as restaurants/gyms/markets/hospitals etc... Except Housing Price, living facilities is one of the most important factors that influence the final decision.

This analysis shows that there are 300 unique venue categories around all the pipeline 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 housing pipeline projects were divided into 4 clusters:

5) Cluster 1 contains 12 affordable housing pipeline projects, top 10 Most Common Venue mainly contains Art Gallery /Spa /Bus Station /Public Art /Jewelry Store /Flower Shop, if the individual is an artist or have great interest in art, these areas will be quite good.

6) Cluster 2 contains 53 affordable housing pipeline projects, this cluster contains most projects, top 10 Most Common Venue mainly contains Coffee Shop /Gym /Bar /Sandwich

Place /Art Gallery /all kinds of Restaurants /Spa /Park /Theater etc.., almost cover every aspect of our daily life, obviously, it will be very convenient to live in these areas.

7) Cluster 3 contains 5 affordable housing pipeline projects, top 10 Most Common Venue mainly contains Spa /Factory /Food & Drink Shop /Motorcycle Shop /Brewery /Food Truck /Food Stand /Food Service, it seems like these are food producing areas, living facilities are too simple here, but it may be good for those who work here.

8) Cluster 4 contains 5 affordable housing pipeline projects, top 10 Most Common Venue mainly contains Art Gallery /Bus Stop /Grocery Store /Outdoor Sculpture /Restaurant /Harbor / Marina /Spa, another cluster closely relate to art, as we can see from the map, this cluster is not far from cluster 1, people who are interest in art can also take these projects into consideration.

This Project simply processed the ownership affordable housing programs data, and cluster them into four categories based one the living facilities data, the results can only help individuals make their first-step decisions. Further analysis can be done base on these four clusters, which can help provide more detail information to clarify the advantages of each category.

6. Conclusion

Purpose of this project is trying to find a category of optimal affordable housing projects with better living facilities. Target to provide first-step suggestions to individuals who want to have their own home but can not afford commodity housing in San Francisco. 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 Foursquare location data was leveraged to compare each project to provide reliable suggestions for individuals who need a place to live. With unsupervised learning K-means algorithm, all the affordable housing projects were clustered in to 4 categories, the advantages of each category were expressed to help individuals make their first-step decisions.

This Project simply processed the ownership affordable housing programs data, and cluster them into four categories based one the living facilities data, the results can only help individuals make their first-step decisions. Further analysis can be done base on these four clusters, which can help provide more detail information to clarify the advantages of each category.