Venues effect on SF housing price

1. Introduction

Housing price has been increased a lot in 2020 in US. There are many reasons of price increasing, including high inflation, low inventory, and so on. As an international city, housing price in San Francisco is always a hot topic in economic domain. In this project, I focus on one potential reason of high housing price issue in San Francisco. This work will show the relationship between housing price and location of venues using public geographic information.

2. Data

2.1 Data Source

In this work, we will leverage multiple public data resources to support our analysis. First, I use Foursquare API to query geospatial data, including venues categories, latitude, longitude in SF. In addition, I use all zip code in SF to query all these data successfully. To query all zip code in SF, I refer the information in public website: https://www.zip-codes.com/city/ca-san-francisco.asp which contains 51 zip code. However, it would be very challenge to get the data including zip code combined with latitude and longitude. Thus, I refer the database in Federal Census Center to get the 2020 geolocation information(https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html). In this database, it contains zip code combined with latitude and longitude information so that it would be easier for me to merge geolocation data with zip code data in SF. Finally, the most challenge part to get data is to get averaged housing price for each zip code of SF. To lift this blocker, I get all the averaged housing price up to June 2021 from Redfin by typing zip code one by one. Then I record all of the prices in the project for future use.

2.2 Data Preprocess

The most of data preprocess part is to combine the multiple data sources together. As shown previously, we have 4 data sources: foursquare venues data, SF zip code data, geolocation data, and housing price data. Firstly, I combine SF zip code data with geolocation data together by the key of zip code. Then the data with latitude, longitude,

and SF zip code would be ready. With this effort, we use all these information to get all SF venues information in SF from foursquare web server. Then I grouped SF venues data by zip code and then combine grouped data with averaged housing price by zip code, called completed data.

3. Methodology

To generate the further analysis, I use kmeans algorithm to cluster completed data by latitude and longitude. Then considered our housing price information, we can make analysis about if housing price have correlation with location information. The dataset we used to make analysis is shown in Figure 1. As we can see, this data contains cluster labels, zip code, latitude, longitude, and venue category for our analysis.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Cluster Labels
0	94102	37.779583	-122.41934	Louise M. Davies Symphony Hall	37.777976	-122.420157	Concert Hall	0
1	94102	37.779583	-122.41934	War Memorial Opera House	37.778601	-122.420816	Opera House	0
2	94102	37.779583	-122.41934	Herbst Theater	37.779548	-122.420953	Concert Hall	0
3	94102	37.779583	-122.41934	San Francisco Ballet	37.778580	-122.420798	Dance Studio	0
4	94102	37.779583	-122.41934	Asian Art Museum	37.780178	-122.416505	Art Museum	0

Figure 1. clustered data for analysis

However, how to decide k value in k-means algorithm would be challenge. I use elbow method to calculate elbow information with the sum square of error. As shown in the Figure 2, the best k value is 5 in our case.

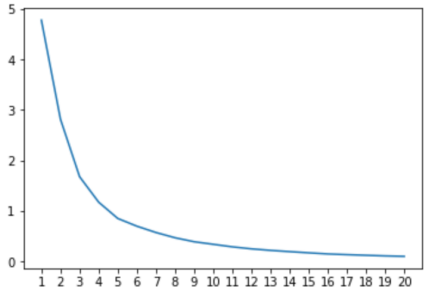

Figure 2. SSE to decide best K in kmeans algorithm

To analysis which venue will have the largest effect of housing price, I also try to generate top 10 venue for each zip code. Combined with housing price information, I will make analysis using the table shown in the Figure 3.

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	housing_price(K)	Cluster Labels
94102	Beer Bar	Theater	Coffee Shop	Cocktail Bar	Vegetarian / Vegan Restaurant	Marijuana Dispensary	Wine Bar	Vietnamese Restaurant	Sushi Restaurant	Dance Studio	1285.0	0.0
94103	Coffee Shop	Nightclub	Beer Bar	Theater	Gay Bar	Cocktail Bar	Dance Studio	Marijuana Dispensary	Gym / Fitness Center	Gym	905.0	0.0
94104	Men's Store	Coffee Shop	Japanese Restaurant	Gym	Hotel	Salad Place	Mediterranean Restaurant	Sushi Restaurant	Food Truck	New American Restaurant	645.0	0.0
94105	Coffee Shop	Food Truck	Seafood Restaurant	Japanese Restaurant	Gym	Salad Place	Dessert Shop	Bookstore	Art Museum	Mediterranean Restaurant	1000.0	0.0
94107	Café	Coffee Shop	Mexican Restaurant	Sushi Restaurant	Park	Brewery	Sandwich Place	Bubble Tea Shop	Breakfast Spot	Art Gallery	1146.0	0.0

Figure 3. housing price data with clustering and venue information

4. Results and Discussion

After generating all of the data, I will make analysis based on it. Firstly, I try to find if location information has effect on housing price. As we can see box plot in Figure 4. In cluster 3, the price is 0, we will deep dive why it happen later, it may because there is no new sold housing information in that area or other reason. Compared with other clusters, the cluster 0 has the lowest housing price and the distribution of all 4 clusters are significantly different. Thus we can make conclusion here is that location information does matter on housing price. The range is from less than 1M to greater 2.5M.

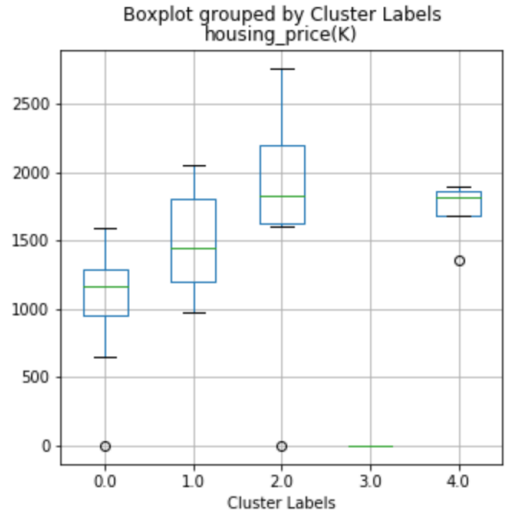

Figure 4, box plot of housing price in different clusters

To visualize the clustering information to better understand which area has the highest housing price, I use folium package to visualize the clusters in the map, as shown in

Figure 5. In the map, the colors from the cluster 0 to cluster 4 is purple, blue, green, orange, and red. Then back to our previous 0 housing price issue in cluster 3, we can see that cluster is from airport which means little housing are available there. Then the housing around the bridge and business center are lowest because low demands in covid 19 period. The price around the park and living place are much higher than other places which make sense to us.

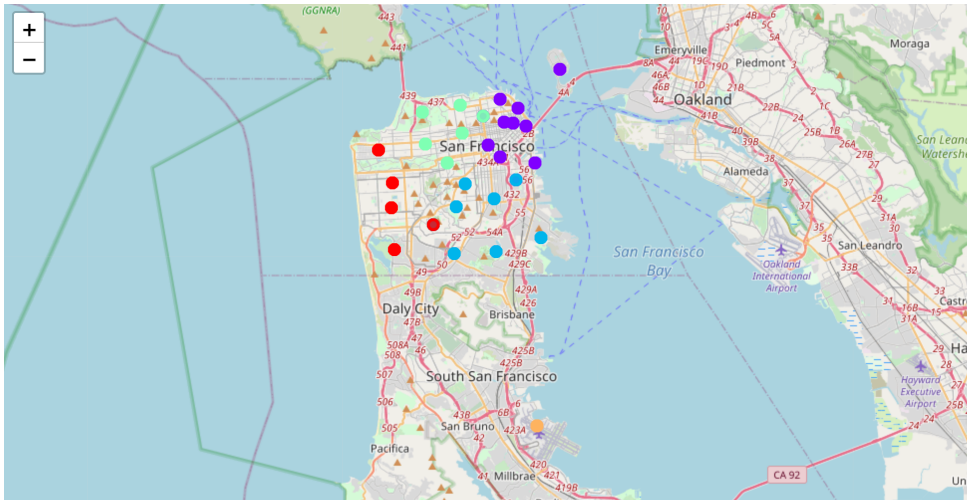

Figure 5. cluster visualization on the map

To analyze the correlation between venues and housing price, I select top most venue for each zip code and then check the top venue for each cluster. As shown in the Figure 6, the top venue of each cluster from 0 to 3 is coffee shop, bakery and park, park, nothing, and Chinese restaurant, as shown in the Figure 6. Based on this graph, we can see park is very important to housing price. That makes sense to this special time which most people prefer to live in more peaceful place, instead of crowded place like business center.

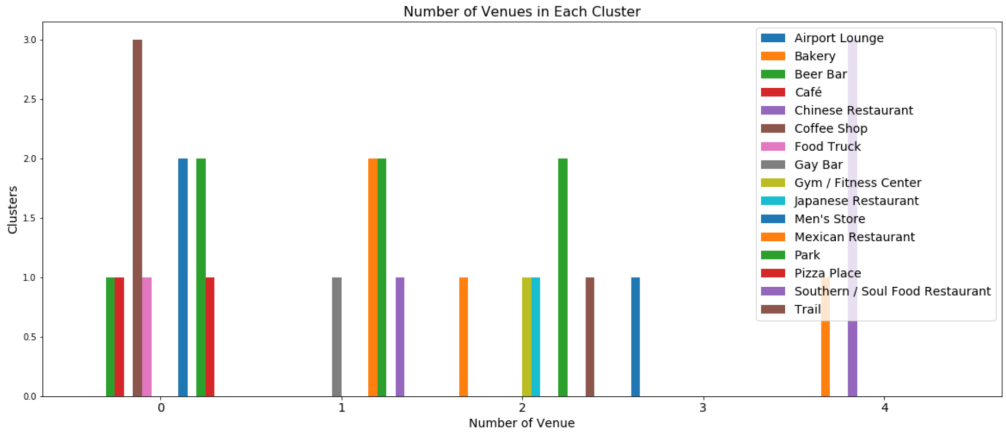

Figure 6. Number of top venues in each cluster

5. Conclusion

In this work, I try to find the reason why housing price increase and what is the main effect on housing price in San Francisco using public dataset and python packages. Based on our analysis, I find the location is a very important effect on SF housing price. In this special time of covid 19, more and more people would love to live the places with more park and have much more peaceful life, instead of living in the crowded business center.