Xiangyi ZHU

Final Project：Comparing housing prices in LA City and CA

1. Motivation surrounding project topic

California is the most populous state and the largest sub-national economy in the world. Los Angeles City is the largest city located in California. Exploring the housing prices in LA and CA can help us better understand the real estate situation in economically developed region and provide us a first-hand information of the housing prices in this area. The main goal of this project is to compare the housing price per square foot in LA and CA.

2. Brief description of data sources

There are three data sources in this project. The first source is web scraping data which contains the housing prices in LA City (Los Angeles Real Estate - Los Angeles CA Homes For Sale | Zillow). The second data source is from OpenCage Geocoding API (OpenCage - Easy, Open, Worldwide, Affordable Geocoding and Geosearch (opencagedata.com)). The third one is downloaded from Kaggle in a csv format, which contains the housing prices across the United States (Real Estate Data From Trulia | Kaggle).

3. How does the whole combined data system work?

First, the housing prices data in LA City was scraped at Zillow, and several information were scraped, namely, address, floor size (square feet), price. Zillow also provides us coordinates, but it does not provide us the corresponding coordinate system. It is very essential to identify the coordinate system in spatial analysis, so I choose the OpenCage Geocoding API to project addresses information scraped from Zillow. This API returns WGS 84 coordinates. The obtained coordinates data from API combined with the information scraped from Zillow form the dataset, which is used to explore the housing prices in LA City.

The dataset downloaded from Trulia provides us the housing prices across the United States. We need to clean this dataset and get the housing prices information in California, which is used to compare with the housing prices in LA City.

4. Analysis performed

The head of the dataset (200 * 3) scraped from Zillow was shown in Figure 1. After removing all rows with null values and combining with the coordinates data from API, the head of this dataset (175 * 5) was shown in Figure 2.

	name	floorSize	price
0	13691 Gavina Ave UNIT 456, Sylmar, CA 91342	1,800 sqft	$230,000
1	594 S Mapleton Dr, Los Angeles, CA 90024	56,500 sqft	$165,000,000
2	2825 Elm St, Los Angeles, CA 90065	1,551 sqft	$929,000
3	15455 Glenoaks Blvd SPACE 408, Sylmar, CA 91342	2,221 sqft	$399,000
4	9131 Burnet Ave UNIT 13, North Hills, CA 91343	1,130 sqft	$449,000
...	...	...	...

Fig 1. Head of the scraped dataset from Zillow.

	name	floorSize	price	lat	lng
0	13691 Gavina Ave UNIT 456, Sylmar, CA 91342	1,800 sqft	$230,000	34.321620	-118.405085
1	594 S Mapleton Dr, Los Angeles, CA 90024	56,500 sqft	$165,000,000	34.073182	-118.429019
2	2825 Elm St, Los Angeles, CA 90065	1,551 sqft	$929,000	34.098499	-118.232164
3	15455 Glenoaks Blvd SPACE 408, Sylmar, CA 91342	2,221 sqft	$399,000	34.323409	-118.469445
4	9131 Burnet Ave UNIT 13, North Hills, CA 91343	1,130 sqft	$449,000	34.236181	-118.463315
...	...	...	...	...	...

Fig 2. Head of dataset after combining with API coordinates data.

The related columns, Title, Sqr Ft, Price, Latitude, and Longitude, were extracted from the dataset downloaded from Kaggle. Only rows with non-null values from CA State were queried. I renamed this dataframe (2543 * 6) to align with the column names in LA housing prices dataset, and it shows in Figure 3.

	name	floorSize	price	lat	lng
0	6036 Moonstone Peak Dr Bakersfield, CA 93313	1,898 sqft	$238,956	35.275806	-119.069890
1	888 W E St #1803 San Diego, CA 92101	1,276 sqft	$1,449,900	32.715040	-117.170715
2	5535 Ackerfield Ave #26 Long Beach, CA 90805	1,069 sqft	$369,000	33.857777	-118.163666
3	1919 W Coronet Ave #161 Anaheim, CA 92801	1,440 sqft	$150,000	33.852856	-117.948494
4	11186 Berryknoll St San Diego, CA 92126	1,525 sqft	$679,000	32.925060	-117.155810
...	...	...	...	...	...

Fig 3. The dataset extracted from Kaggle, which represents the

Both the price and floorSize columns were in the string type, which means arithmetic calculation cannot be conducted on these two columns. So, I utilized the regular expression library in Python to convert the string format into int format. After changing the data type, a new column, named price_sf, which means the price per square foot, was added to each dataset. The heads of the obtained datasets were shown in Figure 4.

	Unnamed: 0	name	floorSize	price	lat	lng	price_sf
0	0	13691 Gavina Ave UNIT 456, Sylmar, CA 91342	1800	230000	34.321620	-118.405085	127.777778
1	1	594 S Mapleton Dr, Los Angeles, CA 90024	56500	165000000	34.073182	-118.429019	2920.353982
2	2	2825 Elm St, Los Angeles, CA 90065	1551	929000	34.098499	-118.232164	598.968407
3	3	15455 Glenoaks Blvd SPACE 408, Sylmar, CA 91342	2221	399000	34.323409	-118.469445	179.648807
4	4	9131 Burnet Ave UNIT 13, North Hills, CA 91343	1130	449000	34.236181	-118.463315	397.345133
...	...	...	...	...	...	...	...

	name	floorSize	price	lat	lng	price_sf
0	6036 Moonstone Peak Dr Bakersfield, CA 93313	1898	238956	35.275806	-119.069890	125.898841
1	888 W E St #1803 San Diego, CA 92101	1276	1449900	32.715040	-117.170715	1136.285266
2	5535 Ackerfield Ave #26 Long Beach, CA 90805	1069	369000	33.857777	-118.163666	345.182413
3	1919 W Coronet Ave #161 Anaheim, CA 92801	1440	150000	33.852856	-117.948494	104.166667
4	11186 Berryknoll St San Diego, CA 92126	1525	679000	32.925060	-117.155810	445.245902
...	...	...	...	...	...	...

Fig 4. Heads of datasets with floorSize and price columns in the int data type.

After getting the price per square foot column in each dataset, two boxplots were plotted based on this column. Figure 5 shows the boxplot of housing price per square foot in LA City dataset.

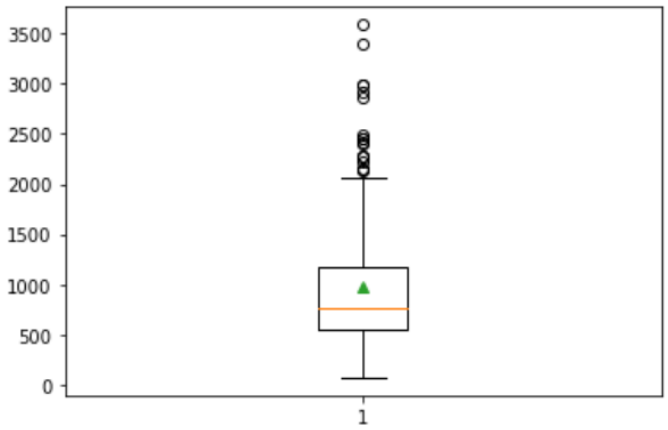

Fig 5. Boxplot of price per square foot column in LA dataset. The orange line represents the median and the green dot represents the mean value. The outliers were shown in circles.

Figure 6 shows the boxplot of per square foot column in CA dataset.

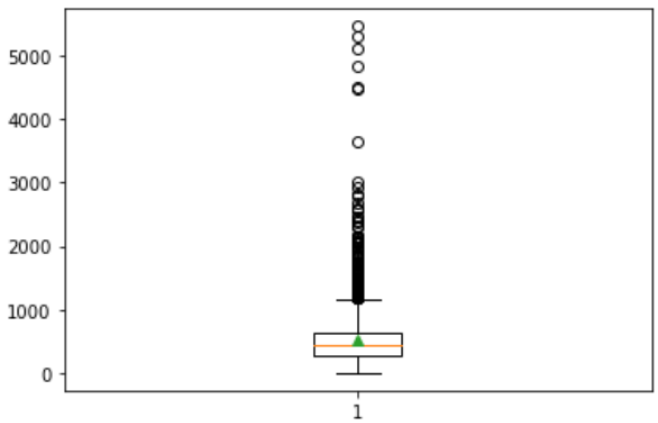

Fig 6. Boxplot of price per square foot column in CA dataset. The orange line represents the median and the green dot represents the mean value. The outliers were shown in circles.

These two plotted boxplots provide us an intuitive information about the first quartile, median,

mean, and third quartile of data in price per square foot column. Figure 7 shows these two boxplots together.

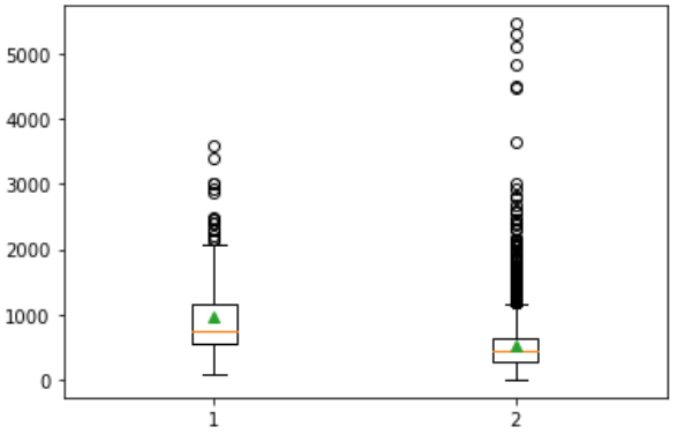

Fig 7. Show these two boxplots in one canvas. The left one shows the price per square foot in LA, and the right one shows the price per square foot in CA.

Figure 8 shows the geographic location of each entity for sale in LA and the size of the circle represents the value of price per square foot. The higher the housing price per square in LA, the bigger the size of the circle is used to represent it. The entities with higher housing price per square are clustered around 34.1 N, -118.4 E.

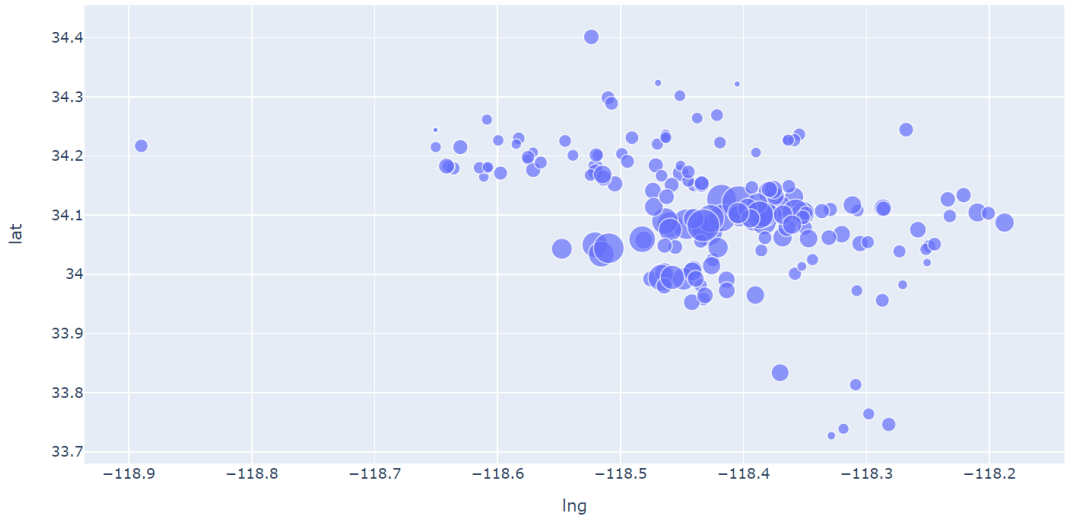

Fig 8. The location of houses for sale in LA.

5. Conclusion drawn

The maximum value, mean, median, minimum, and standard deviation of price per square foot in LA are 3588. 539, 976.517, 757.429, 71.92, and 669.351, respectively. The maximum value, mean, median, minimum, and standard deviation of price per square foot in CA are 5452.882, 530.872, 445.223, 3.961, and 428.152, respectively. Although, the price per square foot of house in CA experiences bigger range compared to the one in LA, LA still have higher mean and median of house price per square foot, which are 976.517 and 757.429 dollars. The higher housing price tends to be cluster in one specific area.

6. Maintainability/extensibility of project

The code used in web scraping step in this project can be reused for scraping any area at Zillow for housing prices when changing some parameters. We can further analyze the relationship between house pricing and its coordinates, in other words, the location. Also, there are some limitations in this project. First, when scraping Zillow, at each page only ten entities can be explored. I double checked my code and I guess that is because Zillow blocked the remaining data, but I am still not sure. Second, the dataset downloaded from Kaggle is from 2019 and more up to date dataset should be further explored. Given that all returned coordinates from the OpenCage Geocoding API use WGS 84 (also known as EPSG: 4326), the dataset scraped from Zillow after coordinating can be further imported to ArcGIS software to conduct further spatial analysis.