APPLIED DATA SCIENCE CAPSTONE

IBM Data Science Specialization @Coursera

Final Project

Yuhung Chang Jokhai

The Battle of Neighborhoods

Introduction

This report is for the final course of the Data Science Specialization, a 9-courses series created by IBM, hosted on Coursera platform.

The purpose of this project is to utilize all tools and skills equipped from the first 8 courses to solve a given problem about real estate markets between New York and Toronto. The project includes problem discussion, data collection, data analysis, and conclusion.

The approach of this project is to set up a problem and to analyze it with a requirement of leveraging data about a neighborhood. In week 3, the neighborhood was set as Toronto and users were asked to use Foursquare location datasets to answer certain questions. In the final week, the users (i.e., Coursera course taker) can choose a city like Toronto. Therefore, I decided to choose Los Angeles as my target.

The main goal will be exploring the neighborhoods of Los Angeles in order to extract the correlation between the real estate value and its surrounding venues. The idea comes from the process of a normal family finding a place to stay after moving to another city.

To that end, the data sources are targeted from Los Angeles County. Zillow provides API for users to download real estate history record. However, API only allows “one address” per time, which means users cannot download mass data from API based on zip code, city, or state. Luckily, Zillow also provides monthly average home price data based on zip code and property styles. From Zillow, users can download CSV files based on zip code for monthly average prices of 1bedroom, 2bedroom, 3bedroom, 4bedroom, 5bedroom or more, and single-family house styles.

Los Angeles County offers very thorough records for restaurant and market inspection and violation records. We can see different types of violation code or total violation records of a restaurant and its relating zip code. Overall, the project is trying to analyze the data and then provide some ideas or suggestion for people who are interested in buying a new house, investors for real estate or restaurant business in Los Angeles.The results will be plotted on map which needs a nice GeoJson data. LA times happens to provide a free GeoJson about LA area. Overall, the steps of this project can be dissected into the following steps:

1. Clean up data types from Zillow and LA County:
    a. https://www.zillow.com/research/data/
    b. https://data.lacounty.gov/Health/LOS-ANGELES-COUNTY-RESTAURANT-AND-MARKET-VIOLATION/8jyd-4pv9

    c. https://data.lacounty.gov/Health/LOS-ANGELES-COUNTY-RESTAURANT-AND-MARKET-INSPECTIO/6ni6-h5kp

2. Create new features
3. Transform the real estate and violation records and merge with inspection records
4. Find appropriate GeoJSON (http://boundaries.latimes.com/set/zip-code-tabulation-areas-2012/)
5. Visualize some data

Data description

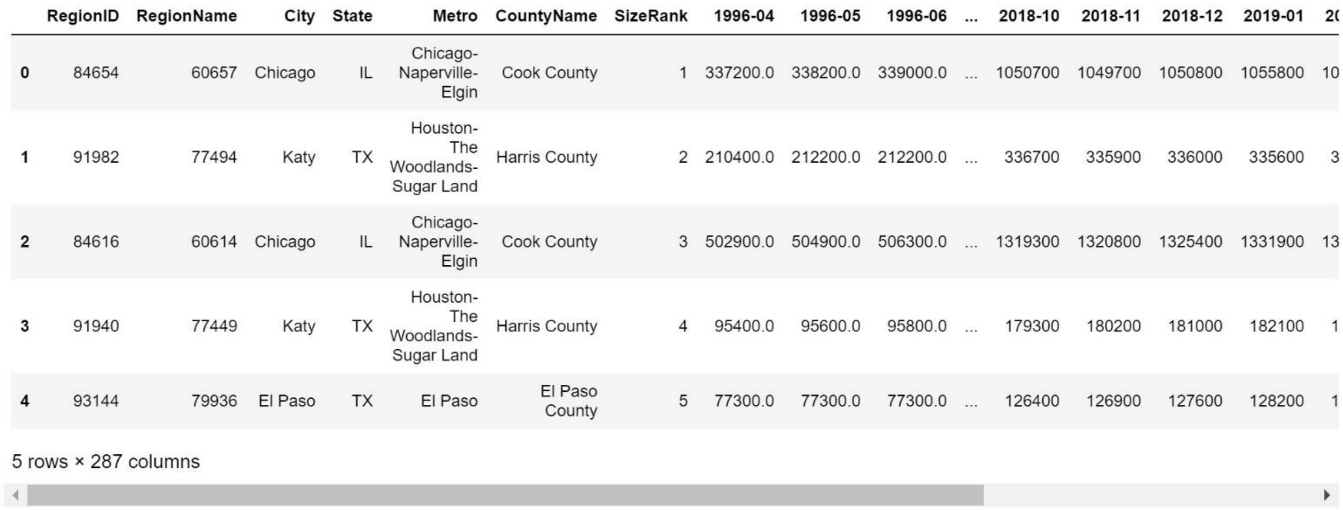
Los Angeles city neighborhoods were chosen as the observation target due to the following reasons:

1. California has the world's 5th largest economy
2. Sign shows that buyers have some negotiating power
3. High demand for rental properties.
4. A good return on investment in Los Angeles

From Zillow (https://www.zillow.com/research/data/), we can download the median estimated home value across a given region and housing type. Here, the datasets are:

1. 1-bedroom
2. 2-bedroom
3. 3-bedroom
4. 4-bedroom
5. 5-bedroom or more
6. Single-family house

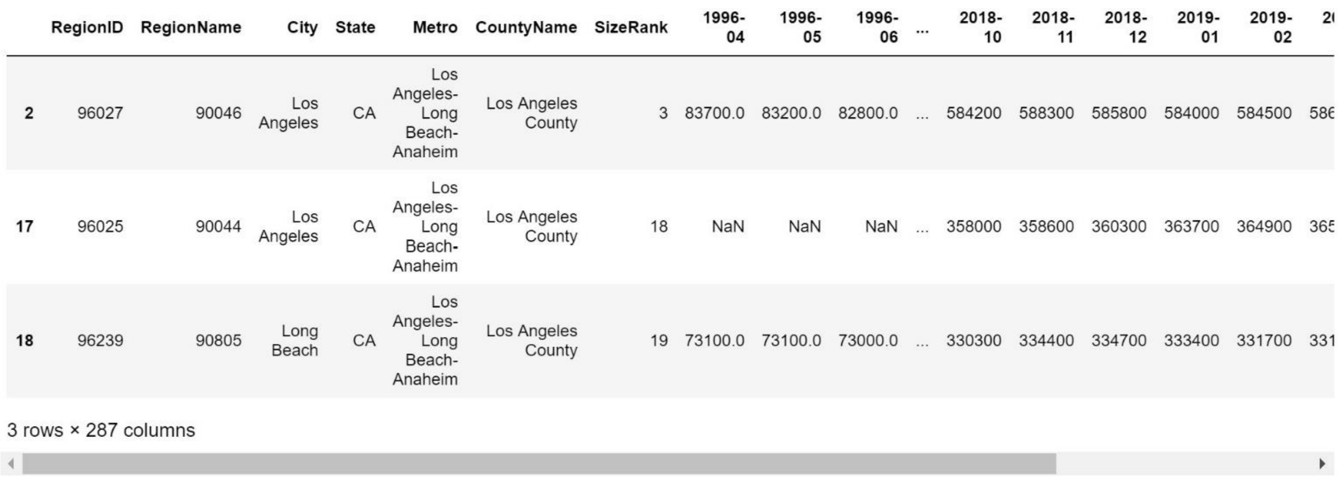
Figure 1 shows how Zillow data looks for single-family house. In this table, ***RegionName*** represents ***zip code***. Also, the data listed from April 1996 to July 2019. The project only focuses on 2015-2019 and Los Angeles. Figure 2 is an example that to filter out cities other than Los Angeles. Figure 3 is an example to show annul average data in 2015 at Los Angeles only.



	RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	1996-05	1996-06	...	2018-10	2018-11	2018-12	2019-01	20
0	84654	60657	Chicago	IL	Chicago-Naperville-Elgin	Cook County	1	337200.0	338200.0	339000.0	...	1050700	1049700	1050800	1055800	10
1	91982	77494	Katy	TX	Houston-The Woodlands-Sugar Land	Harris County	2	210400.0	212200.0	212200.0	...	336700	335900	336000	335600	3
2	84616	60614	Chicago	IL	Chicago-Naperville-Elgin	Cook County	3	502900.0	504900.0	506300.0	...	1319300	1320800	1325400	1331900	13
3	91940	77449	Katy	TX	Houston-The Woodlands-Sugar Land	Harris County	4	95400.0	95600.0	95800.0	...	179300	180200	181000	182100	1
4	93144	79936	El Paso	TX	El Paso	El Paso County	5	77300.0	77300.0	77300.0	...	126400	126900	127600	128200	1

5 rows × 287 columns

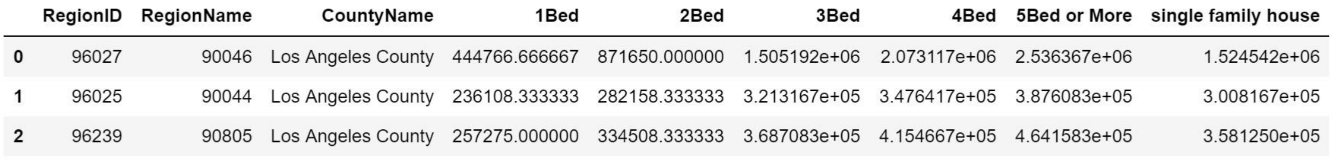
Fig. 1 Zillow data example



	RegionID	RegionName	City	State	Metro	CountyName	SizeRank	1996-04	1996-05	1996-06	...	2018-10	2018-11	2018-12	2019-01	2019-02	20
2	96027	90046	Los Angeles	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	3	83700.0	83200.0	82800.0	...	584200	588300	585800	584000	584500	586
17	96025	90044	Los Angeles	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	18	NaN	NaN	NaN	...	358000	358600	360300	363700	364900	365
18	96239	90805	Long Beach	CA	Los Angeles-Long Beach-Anaheim	Los Angeles County	19	73100.0	73100.0	73000.0	...	330300	334400	334700	333400	331700	331

3 rows × 287 columns

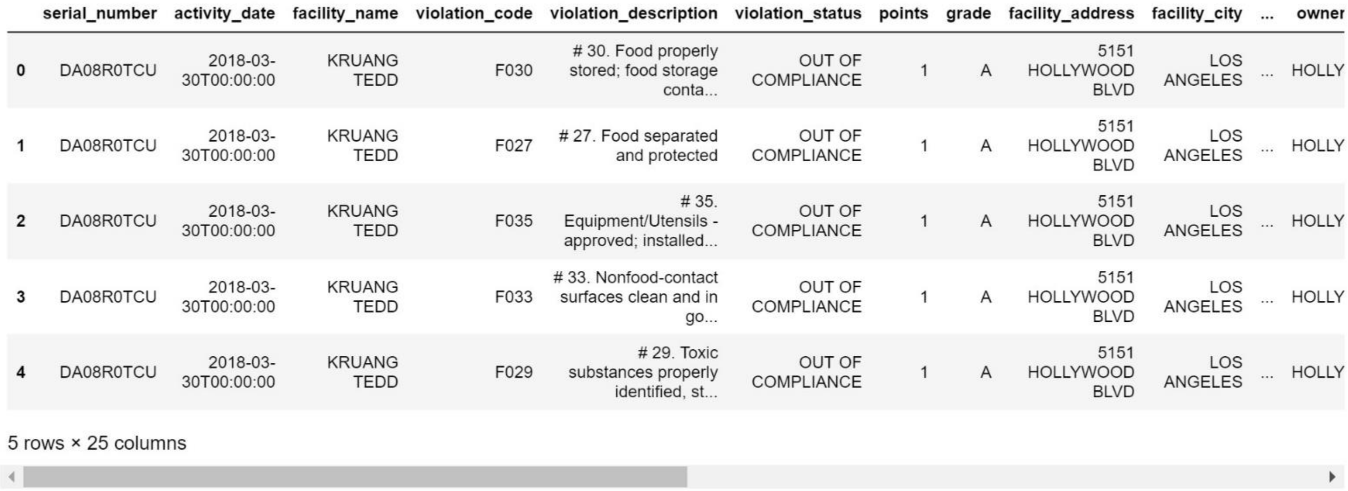
Fig. 2 Zillow data example (Only in Los Angeles County)



	RegionID	RegionName	CountyName	1Bed	2Bed	3Bed	4Bed	5Bed or More	single family house
0	96027	90046	Los Angeles County	444766.666667	871650.000000	1.505192e+06	2.073117e+06	2.536367e+06	1.524542e+06
1	96025	90044	Los Angeles County	236108.333333	282158.333333	3.213167e+05	3.476417e+05	3.876083e+05	3.008167e+05
2	96239	90805	Los Angeles County	257275.000000	334508.333333	3.687083e+05	4.154667e+05	4.641583e+05	3.581250e+05

Fig. 3 Annul average home price in 2015 in Los Angeles.

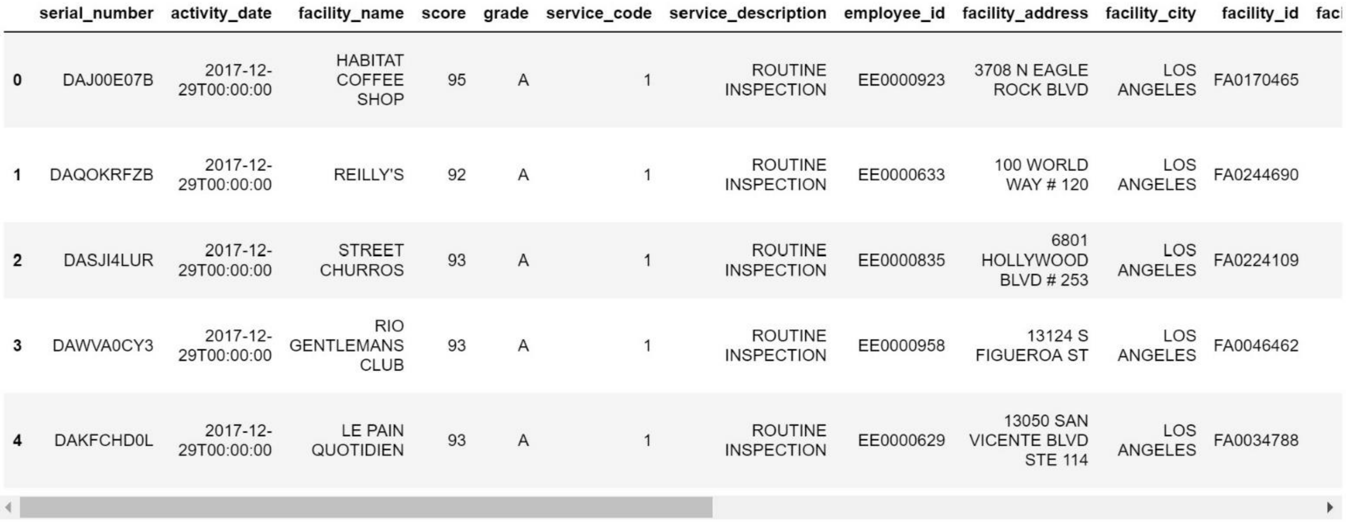
We can do similar things to clean up data of restaurant inspection and violation from Los Angeles government website. Figures 4 and 5 are raw data for Los Angeles food inspection and restaurant violation. Each violation code represents different reasons. We can count all violations for each restaurant, all violations for each zip code, or total average results per zip code (Figures 6 to 11).



	serial_number	activity_date	facility_name	violation_code	violation_description	violation_status	points	grade	facility_address	facility_city	...	owner
0	DA08R0TCU	2018-03-30T00:00:00	KRUANG TEDD	F030	# 30. Food properly stored; food storage conta...	OUT OF COMPLIANCE	1	A	5151 HOLLYWOOD BLVD	LOS ANGELES	...	HOLLY
1	DA08R0TCU	2018-03-30T00:00:00	KRUANG TEDD	F027	# 27. Food separated and protected	OUT OF COMPLIANCE	1	A	5151 HOLLYWOOD BLVD	LOS ANGELES	...	HOLLY
2	DA08R0TCU	2018-03-30T00:00:00	KRUANG TEDD	F035	# 35. Equipment/Utensils - approved; installed...	OUT OF COMPLIANCE	1	A	5151 HOLLYWOOD BLVD	LOS ANGELES	...	HOLLY
3	DA08R0TCU	2018-03-30T00:00:00	KRUANG TEDD	F033	# 33. Nonfood-contact surfaces clean and in go...	OUT OF COMPLIANCE	1	A	5151 HOLLYWOOD BLVD	LOS ANGELES	...	HOLLY
4	DA08R0TCU	2018-03-30T00:00:00	KRUANG TEDD	F029	# 29. Toxic substances properly identified, st...	OUT OF COMPLIANCE	1	A	5151 HOLLYWOOD BLVD	LOS ANGELES	...	HOLLY

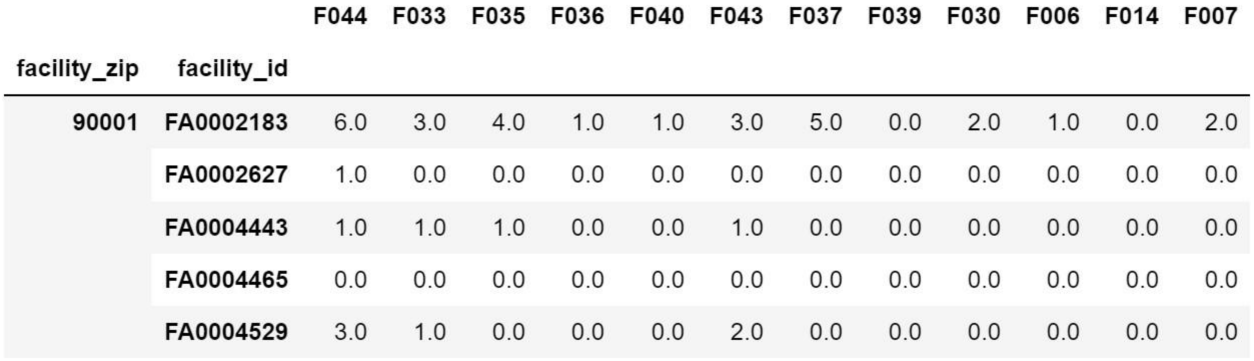
5 rows × 25 columns

Fig. 4 Restaurant violation record at Los Angeles.



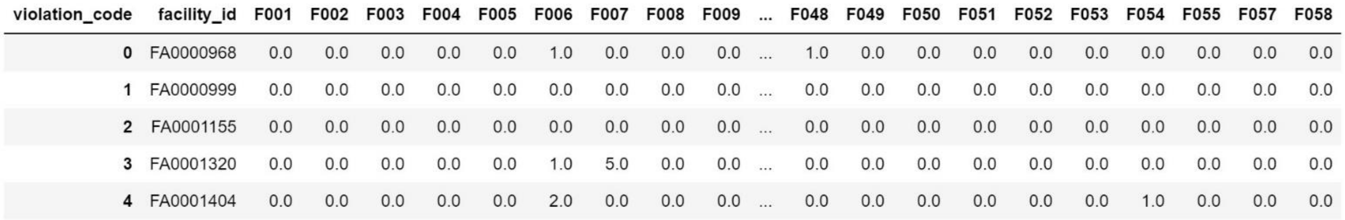
	serial_number	activity_date	facility_name	score	grade	service_code	service_description	employee_id	facility_address	facility_city	facility_id	faci
0	DAJ00E07B	2017-12-29T00:00:00	HABITAT COFFEE SHOP	95	A	1	ROUTINE INSPECTION	EE0000923	3708 N EAGLE ROCK BLVD	LOS ANGELES	FA0170465	
1	DAQOKRFZB	2017-12-29T00:00:00	REILLY'S	92	A	1	ROUTINE INSPECTION	EE0000633	100 WORLD WAY # 120	LOS ANGELES	FA0244690	
2	DASJI4LUR	2017-12-29T00:00:00	STREET CHURROS	93	A	1	ROUTINE INSPECTION	EE0000835	6801 HOLLYWOOD BLVD # 253	LOS ANGELES	FA0224109	
3	DAWVA0CY3	2017-12-29T00:00:00	RIO GENTLEMANS CLUB	93	A	1	ROUTINE INSPECTION	EE0000958	13124 S FIGUEROA ST	LOS ANGELES	FA0046462	
4	DAKFCHD0L	2017-12-29T00:00:00	LE PAIN QUOTIDIEN	93	A	1	ROUTINE INSPECTION	EE0000629	13050 SAN VICENTE BLVD STE 114	LOS ANGELES	FA0034788	

Fig. 5 Restaurant inspection record at Los Angeles.



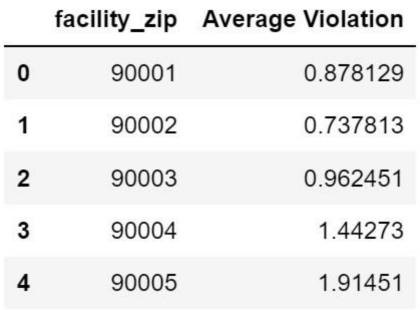
		F044	F033	F035	F036	F040	F043	F037	F039	F030	F006	F014	F007
facility_zip	**facility_id**												
90001	**FA0002183**	6.0	3.0	4.0	1.0	1.0	3.0	5.0	0.0	2.0	1.0	0.0	2.0
	FA0002627	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FA0004443	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
	FA0004465	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	FA0004529	3.0	1.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0

Fig6. Group table based on zip code and facility id (i.e., restaurant ID).



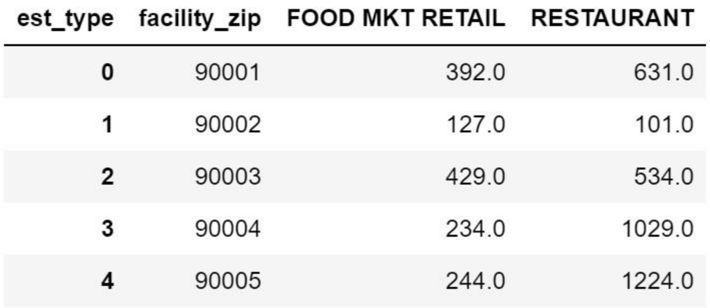
violation_code	facility_id	F001	F002	F003	F004	F005	F006	F007	F008	F009	...	F048	F049	F050	F051	F052	F053	F054	F055	F057	F058
0	FA0000968	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	FA0000999	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	FA0001155	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	FA0001320	0.0	0.0	0.0	0.0	0.0	1.0	5.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	FA0001404	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0

Fig7. Violation for each facility id (i.e., restaurant).



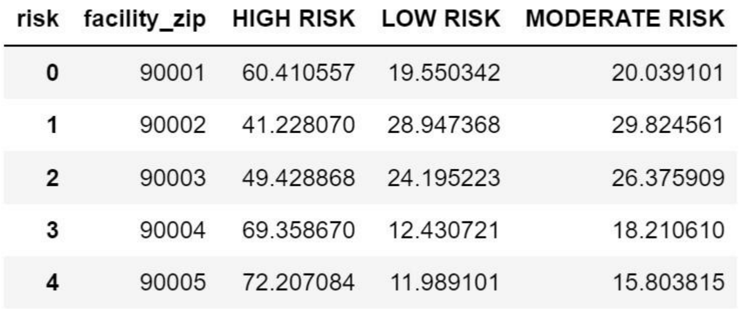
	facility_zip	Average Violation
0	90001	0.878129
1	90002	0.737813
2	90003	0.962451
3	90004	1.44273
4	90005	1.91451

Fig8. Total violation for each zip code in LA.



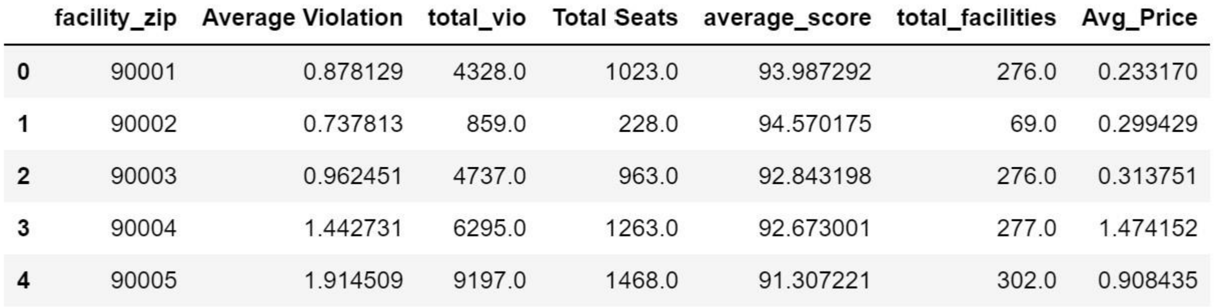
est_type	facility_zip	FOOD MKT RETAIL	RESTAURANT
0	90001	392.0	631.0
1	90002	127.0	101.0
2	90003	429.0	534.0
3	90004	234.0	1029.0
4	90005	244.0	1224.0

Fig9. Total restaurant per zip code (i.e., restaurant).



risk	facility_zip	HIGH RISK	LOW RISK	MODERATE RISK
0	90001	60.410557	19.550342	20.039101
1	90002	41.228070	28.947368	29.824561
2	90003	49.428868	24.195223	26.375909
3	90004	69.358670	12.430721	18.210610
4	90005	72.207084	11.989101	15.803815

Fig10. Percentage of risks of restaurants per zip code in LA



	facility_zip	Average Violation	total_vio	Total Seats	average_score	total_facilities	Avg_Price
0	90001	0.878129	4328.0	1023.0	93.987292	276.0	0.233170
1	90002	0.737813	859.0	228.0	94.570175	69.0	0.299429
2	90003	0.962451	4737.0	963.0	92.843198	276.0	0.313751
3	90004	1.442731	6295.0	1263.0	92.673001	277.0	1.474152
4	90005	1.914509	9197.0	1468.0	91.307221	302.0	0.908435

Fig11. Total average information in LA.

Data Visualization

Let us see the map of real estate value in Los Angeles for 2015 and 2018. Black means Zillow doesn't have data in that zip code; red means the highest value zip code; and blue means the lowest value zip code. The unit of price is million US dollars. It has to be noted that the color represents "mean" value in each zip code.

From the map, we can tell that

1. 2-bedroom and 3-bedroom properties (condos) are the most popular choices for people in LA.
2. For family, single-family-house is more popular than "5-bedroom or more room" type of properties.
3. Home price is always high for any kid of properties nearby Santo Monica or Long Beach areas.
4. From 2015 to 2018, the real estate market has been raised up 1.5-2.0 times for most areas in LA.
5. LA downtown and Inglewood areas have the highest restaurants densities.
6. Restaurants in outskirt LA areas have higher food inspection scores than other LA areas.

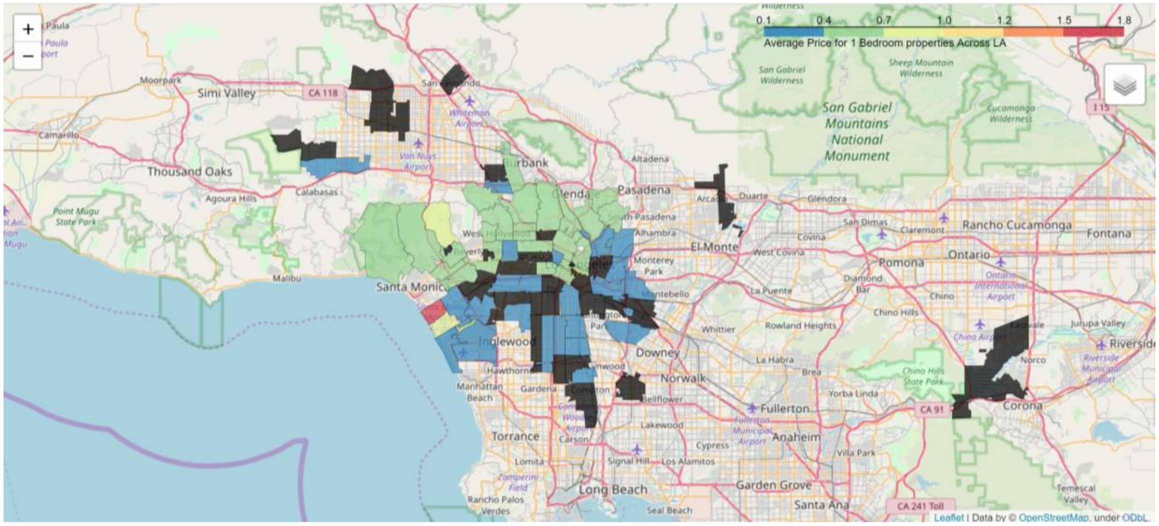


Fig. 12 2015 1-bedrrom property (unit: million US dollars):

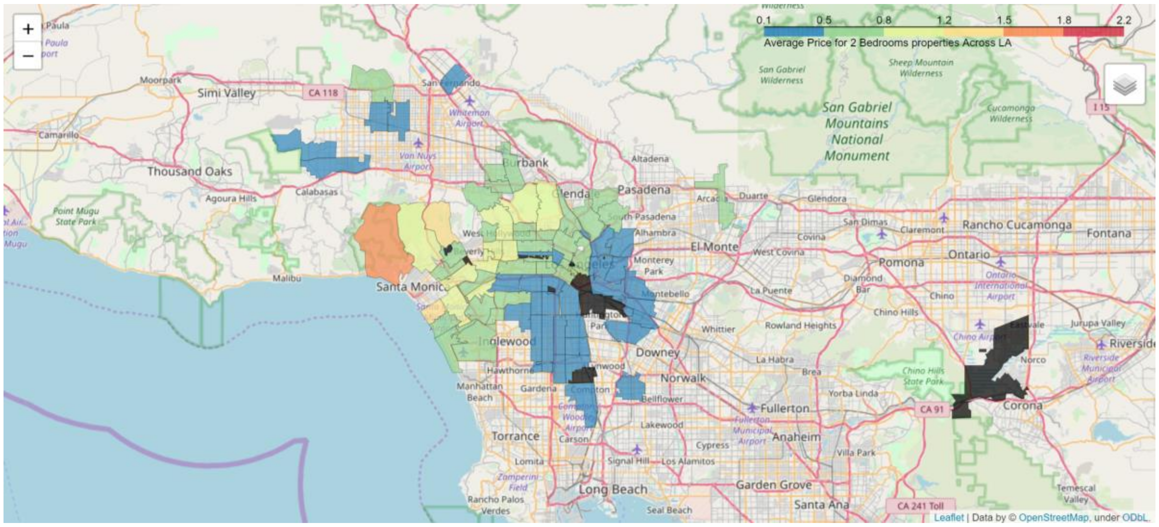

0.1 0.5 0.8 1.2 1.5 1.8 2.2
Average Price for 2 Bedrooms properties Across LA
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig. 13 2015 2-bedrrom property (unit: million US dollars)

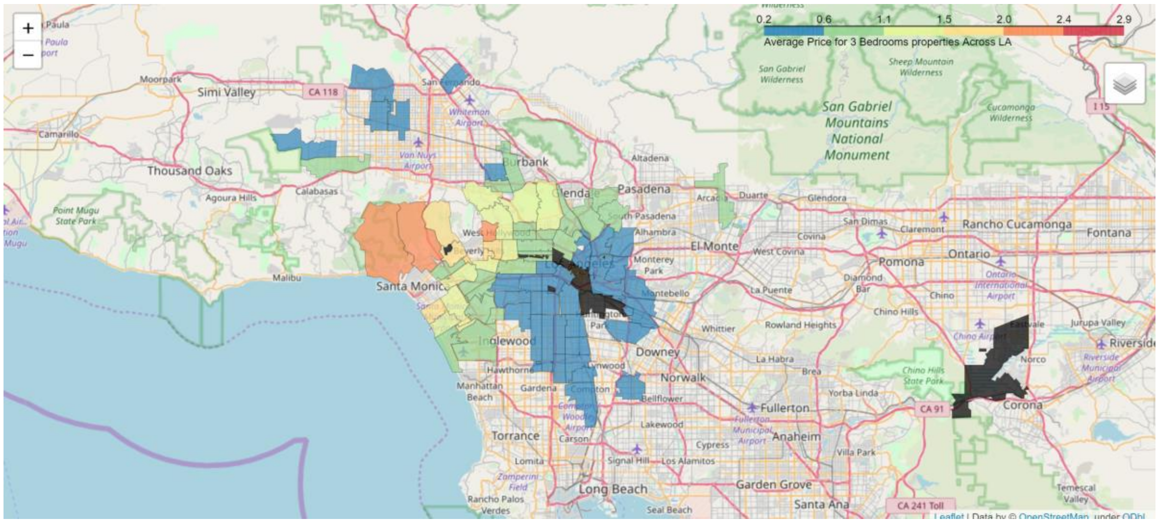

0.2 0.6 1.1 1.5 2.0 2.4 2.9
Average Price for 3 Bedrooms properties Across LA
Leaflet | Data by © OpenStreetMap, under ODbL

Fig. 14 2015 3-bedrrom property (unit: million US dollars):

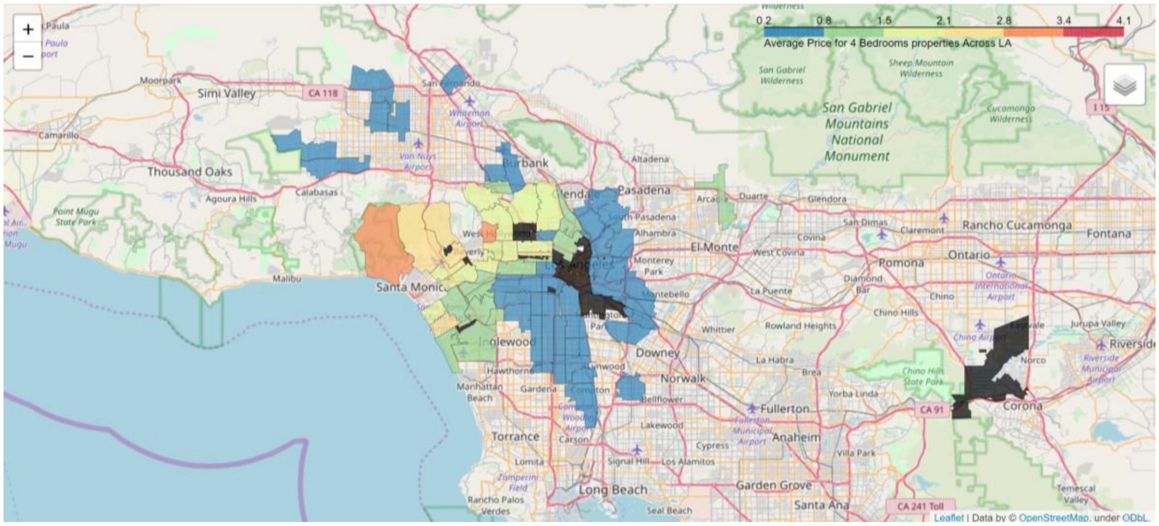

0.2 0.8 1.5 2.1 2.8 3.4 4.1
Average Price for 4 Bedrooms properties Across LA
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig. 15 2015 4-bedrrom property (unit: million US dollars)

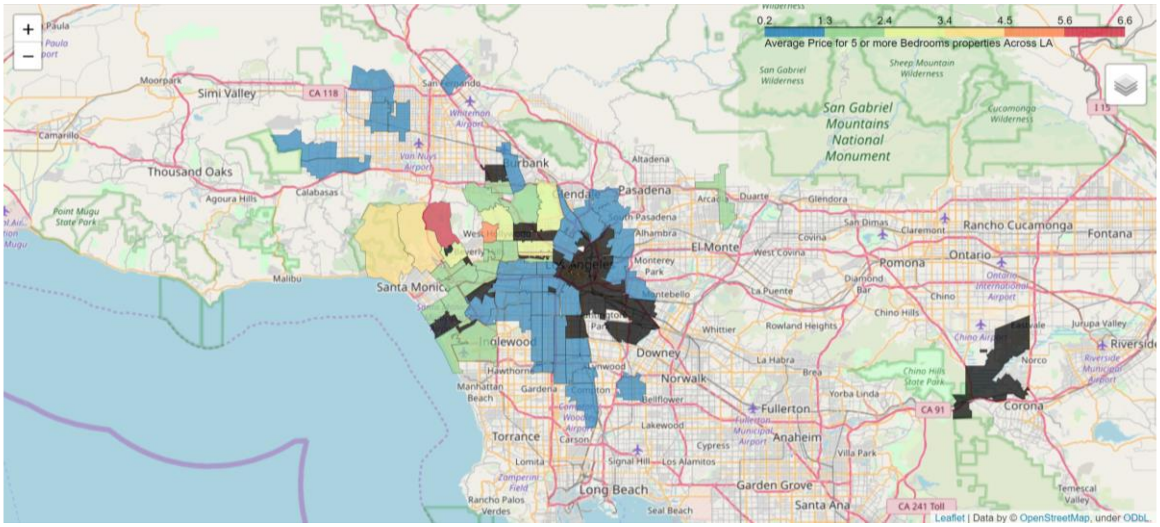


Fig. 15 2015 5-bedrrom or more property (unit: million US dollars)

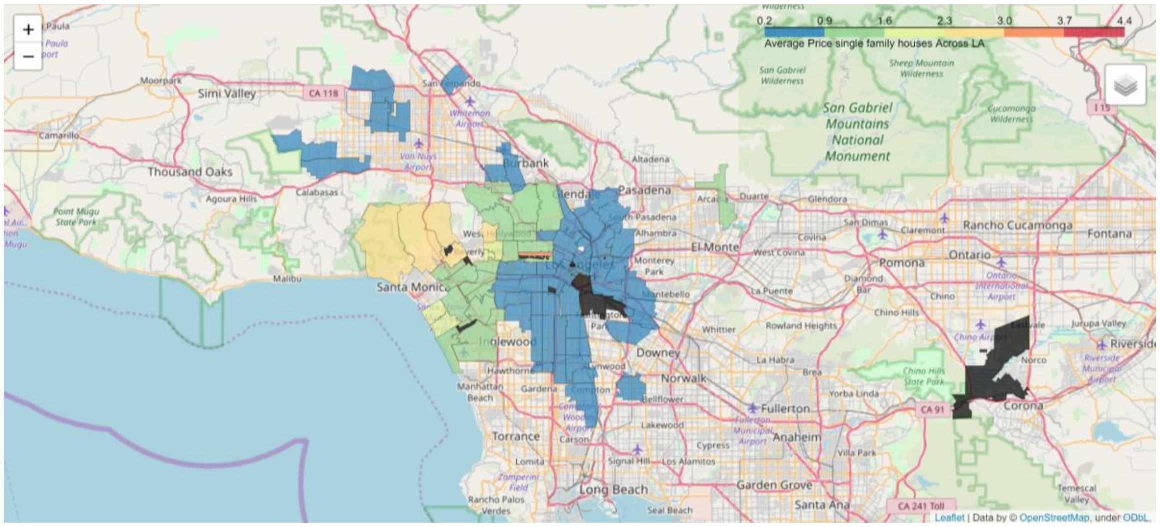


Fig. 16 2015 Single family house (unit: million US dollars)

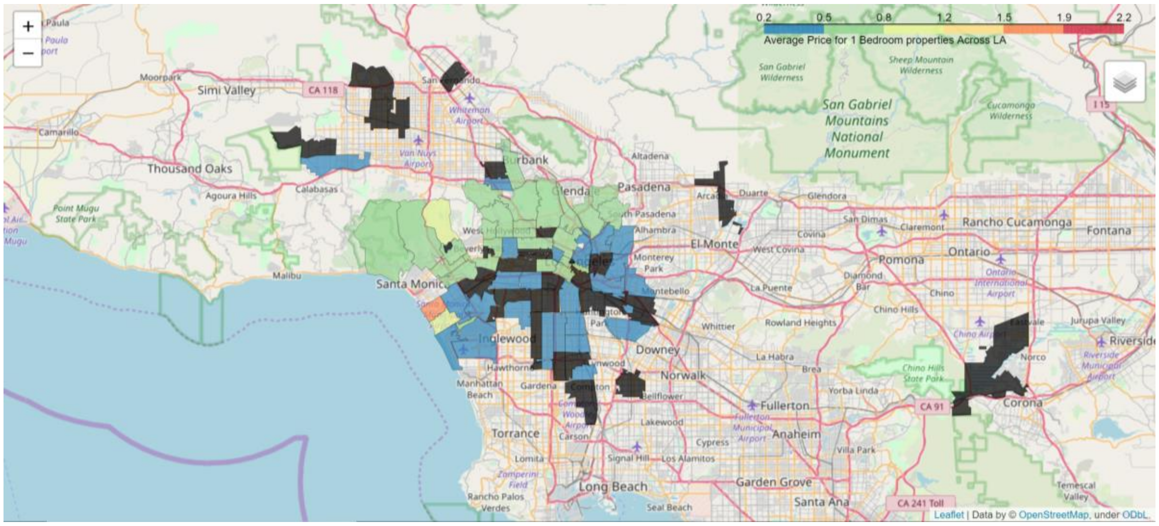


Fig. 17 2018 1-bedrrom property (unit: million US dollars)

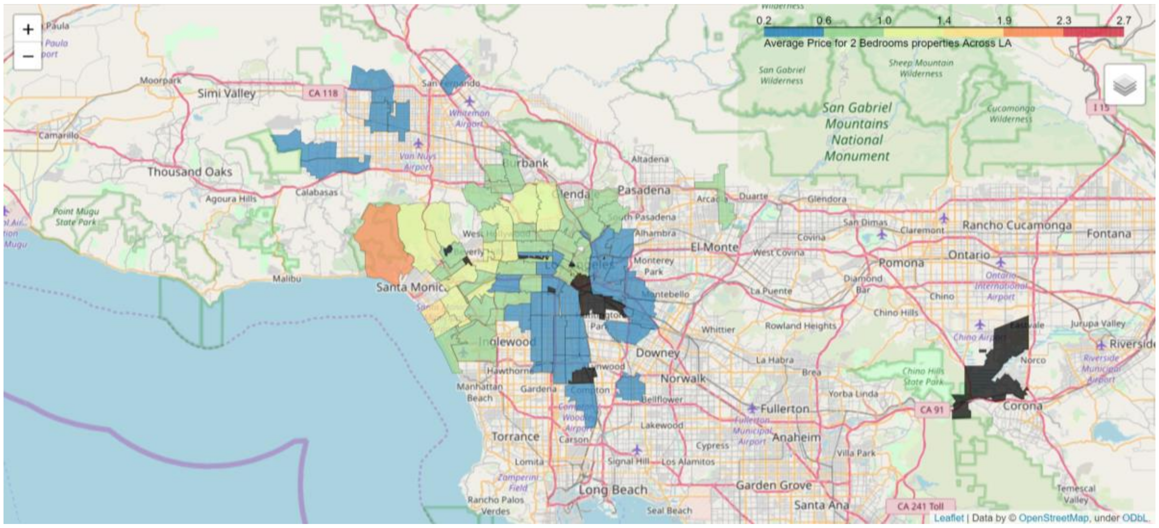

0.2
0.6
1.0
1.4
1.9
2.3
2.7
Average Price for 2 Bedrooms properties Across LA
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig. 18 2018 2-bedrrom property (unit: million US dollars)

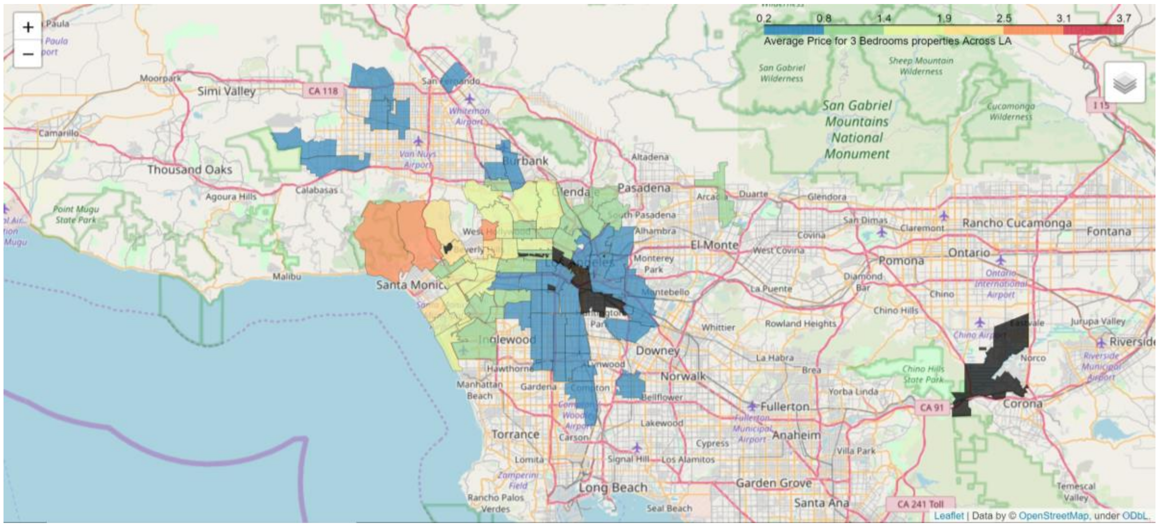

0.2
0.8
1.4
1.9
2.5
3.1
3.7
Average Price for 3 Bedrooms properties Across LA
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig. 19 2018 3-bedrrom property (unit: million US dollars)

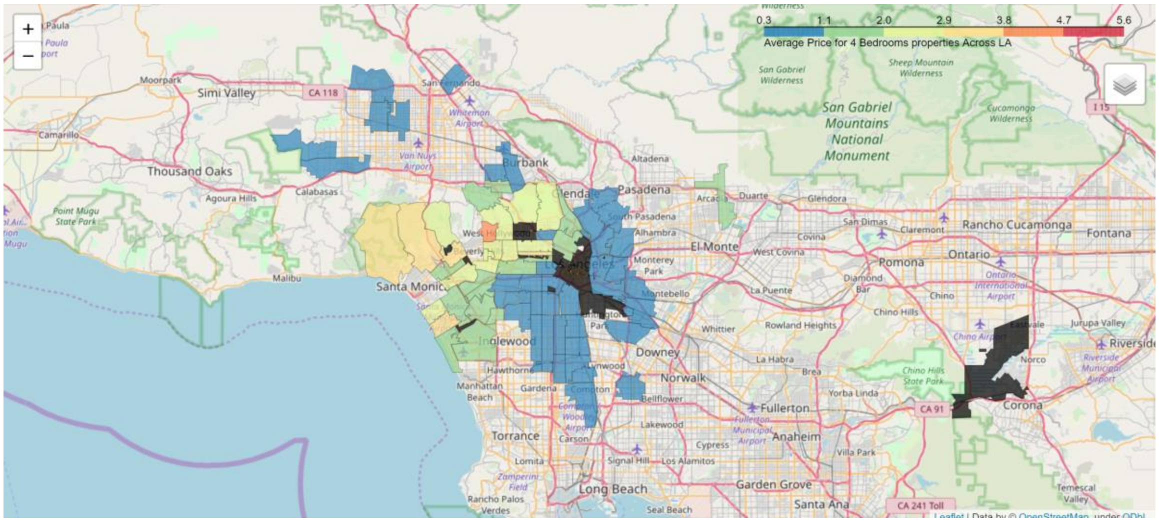


Fig. 12 2018 4-bedrrom property (unit: million US dollars)

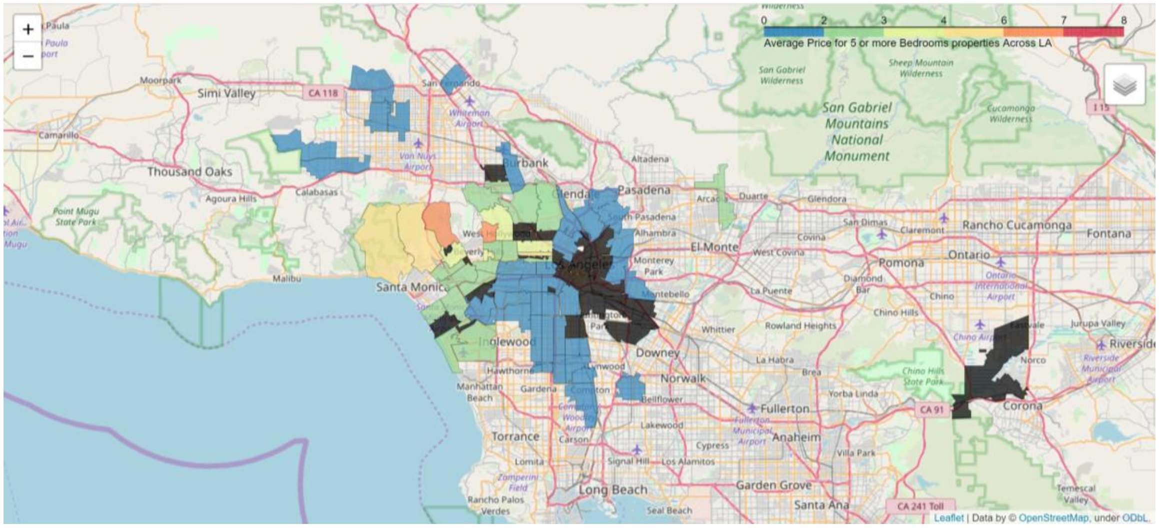

0
2
3
4
6
7
8
Average Price for 5 or more Bedrooms properties Across LA
+
−
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig. 12 2018 5-bedrrom or more property (unit: million US dollars)

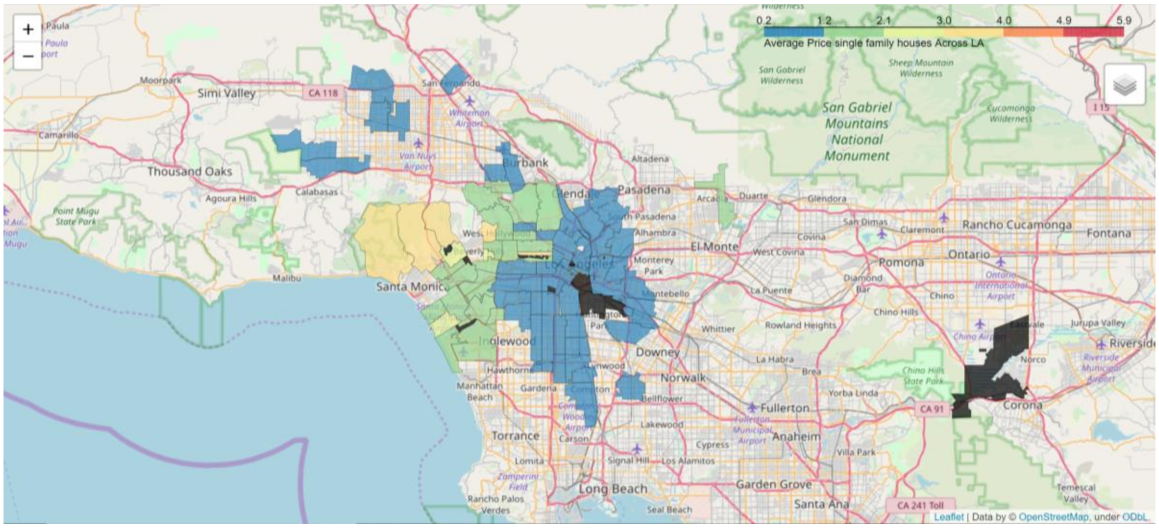

0.2
1.2
2.1
3.0
4.0
4.9
5.9
Average Price single family houses Across LA
+
−
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig. 22 2018 single-family house property (unit: million US dollars)

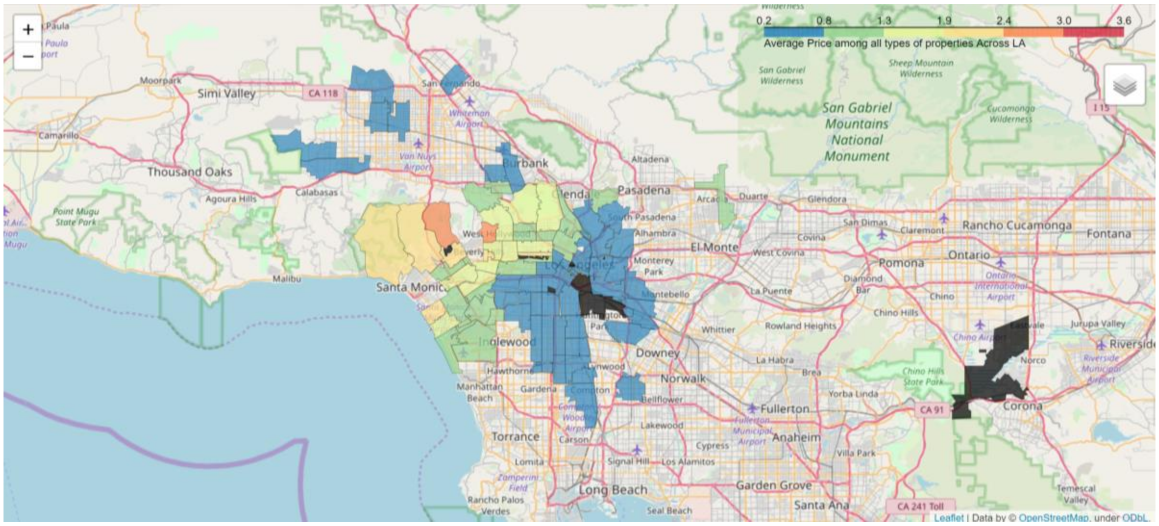

0.2
0.8
1.3
1.9
2.4
3.0
3.6
Average Price among all types of properties Across LA
+
−
Leaflet | Data by © OpenStreetMap, under ODbL.

Fig. 23 Average price among all types of properties (unit: million US dollars)

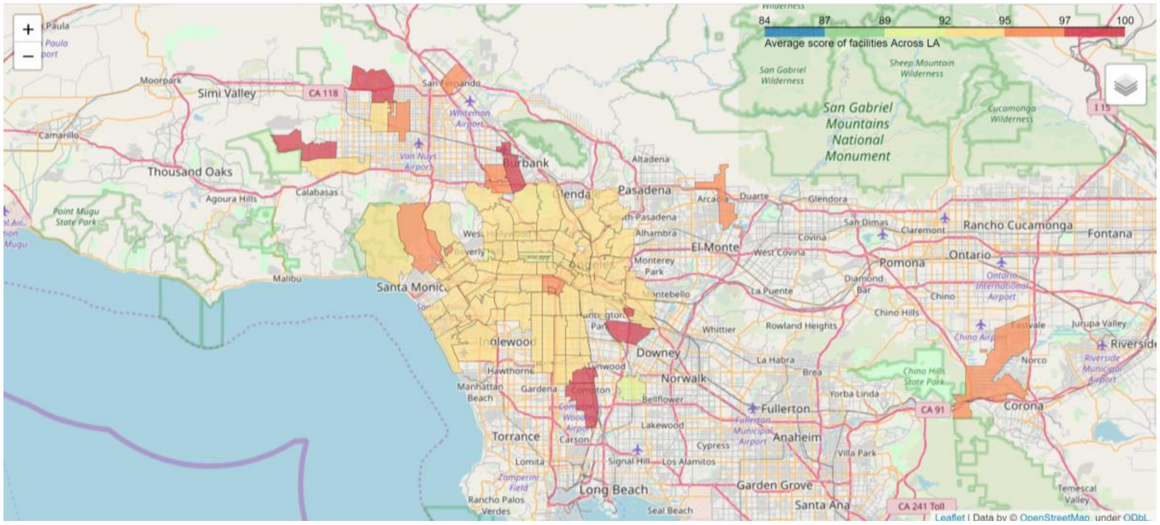


Fig. 24 **Average inspection score of restaurants in LA**

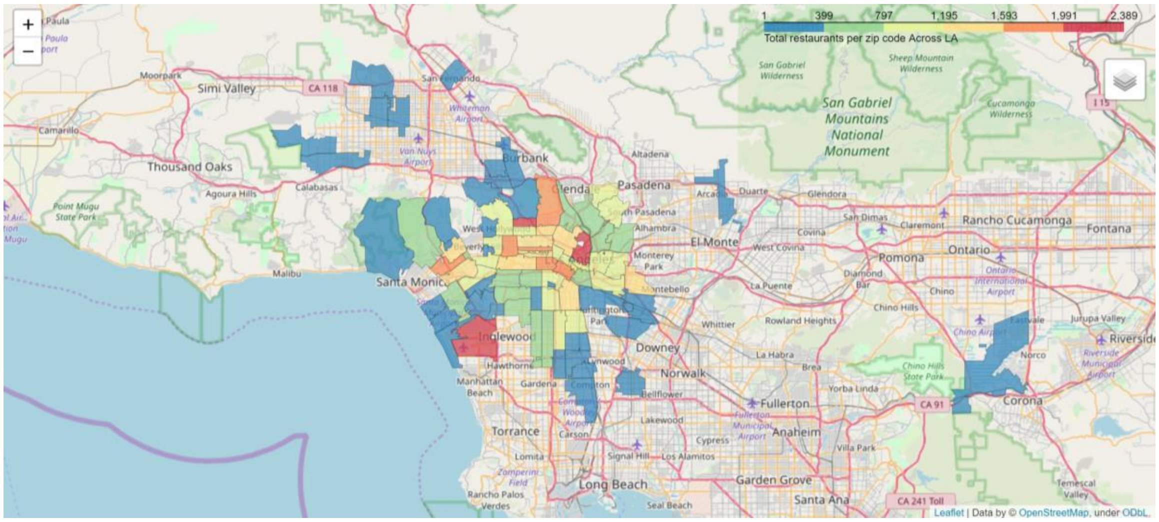


Fig. 25 Total restaurants in LA

Linear Regression Analysis

After viewing data on choropleth maps, let's do some simple analysis with linear regression models. Fig. 26 is a scatter plot matrix for all important factors per zip code in LA, including total number of restaurants, total seats, total violations, average scores, total restaurants with violations, and average home/house price. From Fig. 26, we can see that the condition (e.g., violations or seats) has no influence on home price in LA.

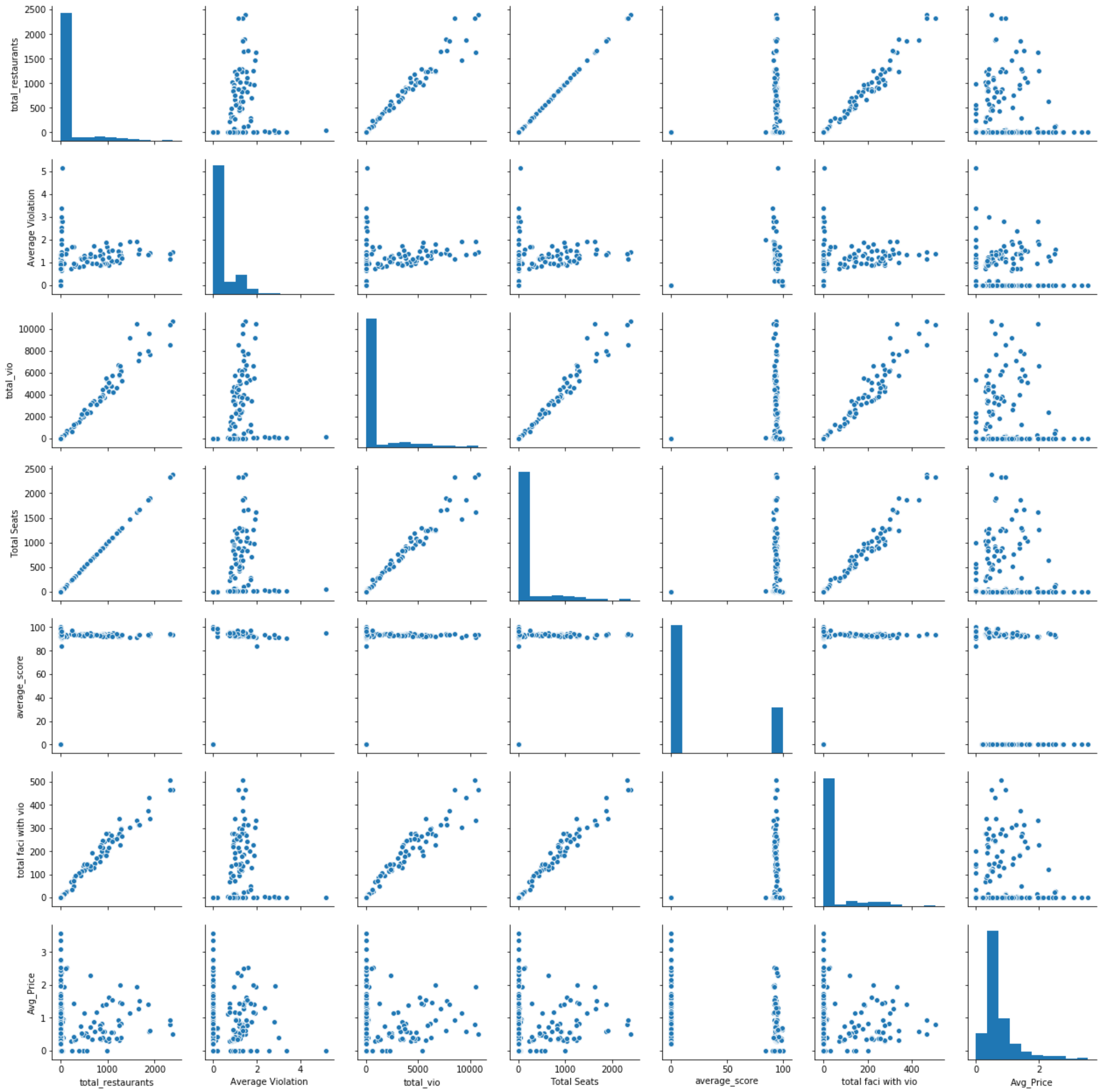
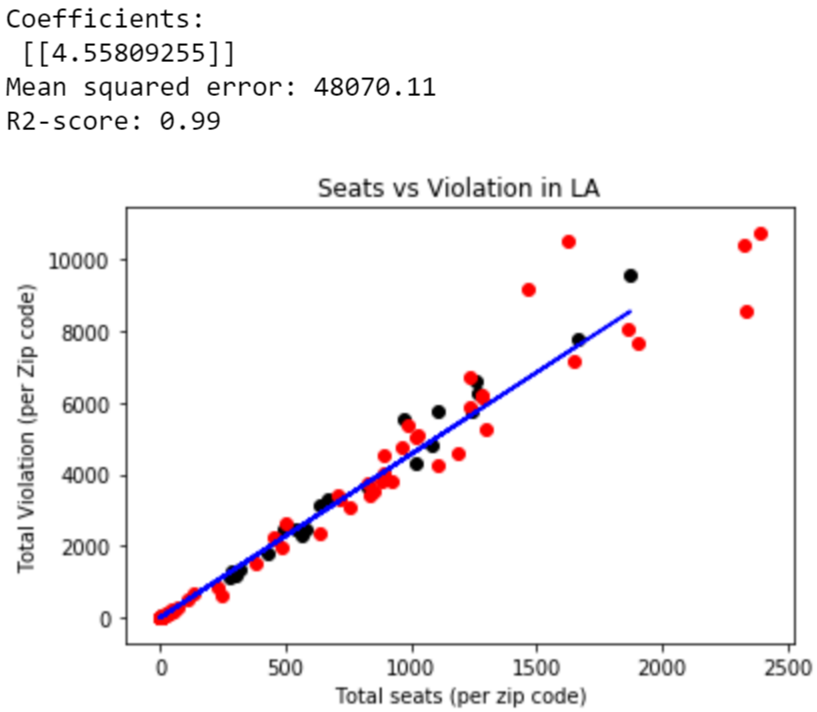


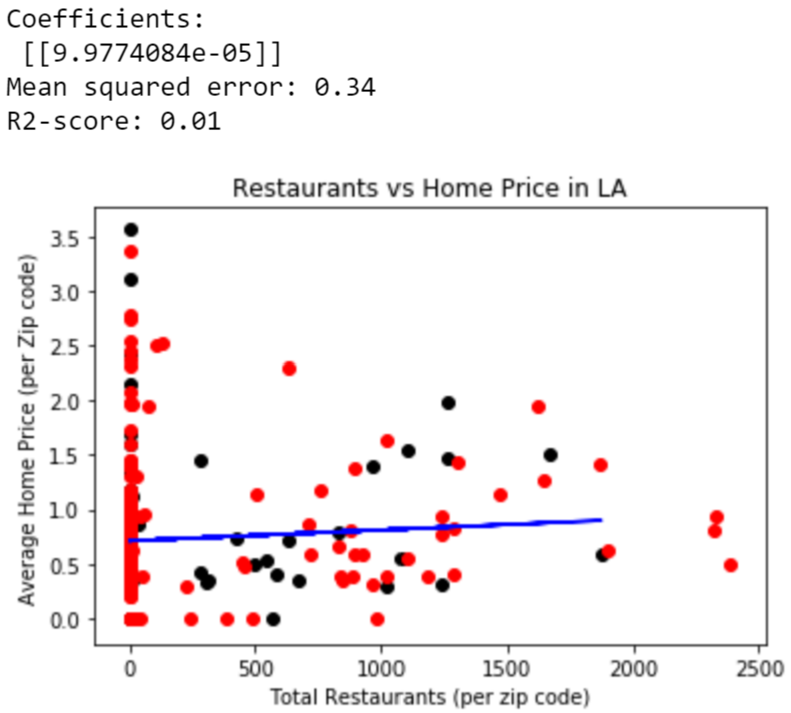
Fig. 26 Scatter plot matrix for important factors

Figures 27 to 32 show more details about the simple linear regression analysis between two selected factors in LA. It shows an area with more seats (for restaurants) comes with more food/restaurant violations. It does make sense because it usually harder for a big facility to manage hygiene conditions than a smaller facility. However, the restaurant hygiene conditions are not relevant to the home price in LA.



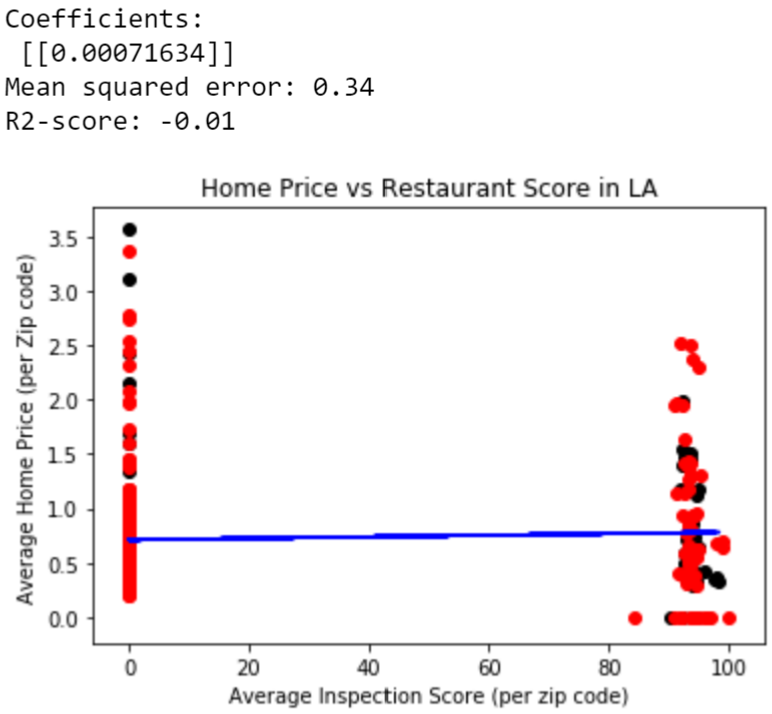
```
Coefficients:
 [[4.55809255]]
Mean squared error: 48070.11
R2-score: 0.99
```

Fig. 27 Total seats vs Average violations in LA (per zip code)



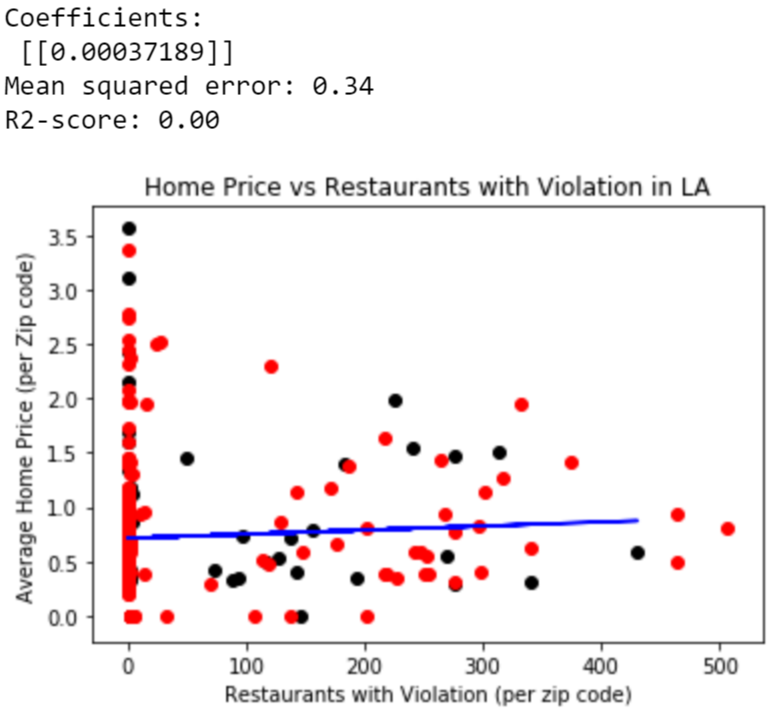
```
Coefficients:
 [[9.9774084e-05]]
Mean squared error: 0.34
R2-score: 0.01
```

Fig. 28 Total Restaurants vs Average Home Price in LA (per zip code)



```
Coefficients:
 [[0.00071634]]
Mean squared error: 0.34
R2-score: -0.01
```

Fig. 29 Average Home Price vs Restaurant Inspection Score in LA (per zip code)



```
Coefficients:
 [[0.00037189]]
Mean squared error: 0.34
R2-score: 0.00
```

Fig. 30 Average Home Price vs Restaurant with Violation in LA (per zip code)

```
Coefficients:
 [[0.00264363]]
Mean squared error: 0.35
R2-score: -0.03
```

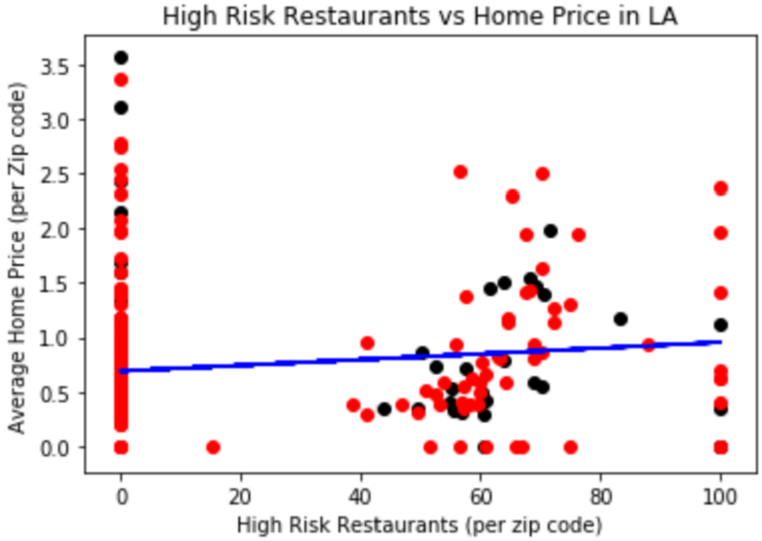


Fig. 31 Average Home Price vs High Risk Restaurants in LA (per zip code)

```
Coefficients:
 [[-0.0032043]]
Mean squared error: 0.34
R2-score: 0.01
```

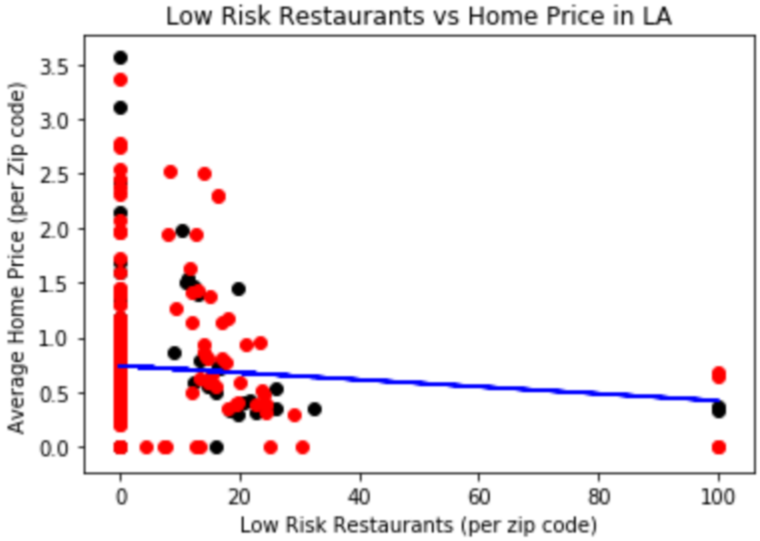


Fig. 32 Average Home Price vs Low Risk Restaurants in LA (per zip code)

Summary

Overall, the restaurant hygiene performance or the density of restaurants in LA has nothing to do with the real estate market. Tables 1 to 3 are summaries of the main factors and real estate market in 2015 and 2018. The average home price among 4 years (2015-2018) is 700,000. However, if we compare different types of properties, we can easily find that the values of 2-bedroom and 4-bedroom properties have increase the most in the past 3 to 4 years.

Although, the restaurants performance has nothing to do with the real estate market, we can still make the following suggestions for potential home buyers or investors:

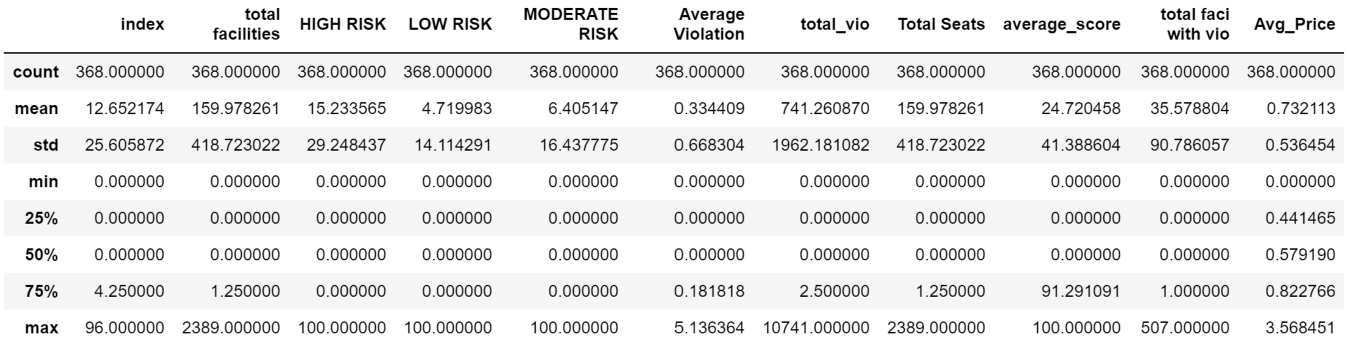
A. If you have lower budgets for housing, you can choose areas other than downtown or areas close to Santa Monica or Long Beach.

B. If you are a restaurant investor, you probably want to run a business with smaller scale which you can do better maintenance for your restaurant hygiene performance.

C. If you are a restaurant investor, you may want to avoid “high risk” area. High risk area may have higher frequency of visiting by government inspectors.

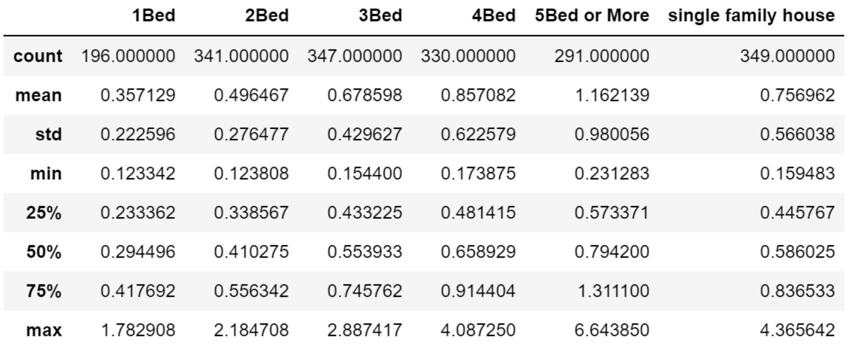
D. If you just move to LA and try to find a good place to eat, you probably want to choose a restaurant locating in a “low risk” area to prevent any possible issues.

Table 1. Statistics summary of the main factors



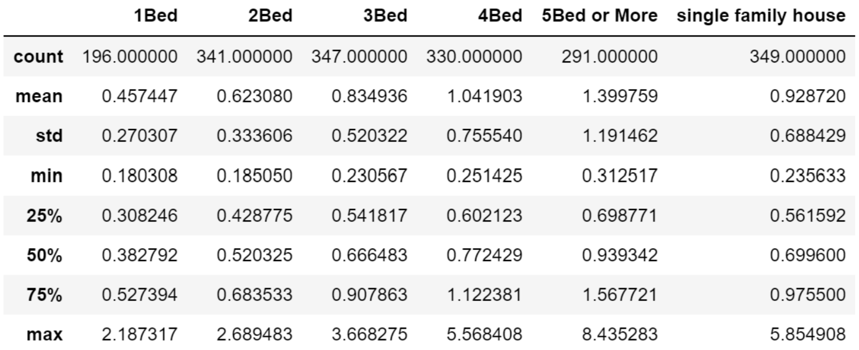
	index	total facilities	HIGH RISK	LOW RISK	MODERATE RISK	Average Violation	total_vio	Total Seats	average_score	total faci with vio	Avg_Price
count	368.000000	368.000000	368.000000	368.000000	368.000000	368.000000	368.000000	368.000000	368.000000	368.000000	368.000000
mean	12.652174	159.978261	15.233565	4.719983	6.405147	0.334409	741.260870	159.978261	24.720458	35.578804	0.732113
std	25.605872	418.723022	29.248437	14.114291	16.437775	0.668304	1962.181082	418.723022	41.388604	90.786057	0.536454
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.441465
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.579190
75%	4.250000	1.250000	0.000000	0.000000	0.000000	0.181818	2.500000	1.250000	91.291091	1.000000	0.822766
max	96.000000	2389.000000	100.000000	100.000000	100.000000	5.136364	10741.000000	2389.000000	100.000000	507.000000	3.568451

Table 2. Statistics summary of 2015 real estate market



	1Bed	2Bed	3Bed	4Bed	5Bed or More	single family house
count	196.000000	341.000000	347.000000	330.000000	291.000000	349.000000
mean	0.357129	0.496467	0.678598	0.857082	1.162139	0.756962
std	0.222596	0.276477	0.429627	0.622579	0.980056	0.566038
min	0.123342	0.123808	0.154400	0.173875	0.231283	0.159483
25%	0.233362	0.338567	0.433225	0.481415	0.573371	0.445767
50%	0.294496	0.410275	0.553933	0.658929	0.794200	0.586025
75%	0.417692	0.556342	0.745762	0.914404	1.311100	0.836533
max	1.782908	2.184708	2.887417	4.087250	6.643850	4.365642

Table 3. Statistics summary of 2018 real estate market



	1Bed	2Bed	3Bed	4Bed	5Bed or More	single family house
count	196.000000	341.000000	347.000000	330.000000	291.000000	349.000000
mean	0.457447	0.623080	0.834936	1.041903	1.399759	0.928720
std	0.270307	0.333606	0.520322	0.755540	1.191462	0.688429
min	0.180308	0.185050	0.230567	0.251425	0.312517	0.235633
25%	0.308246	0.428775	0.541817	0.602123	0.698771	0.561592
50%	0.382792	0.520325	0.666483	0.772429	0.939342	0.699600
75%	0.527394	0.683533	0.907863	1.122381	1.567721	0.975500
max	2.187317	2.689483	3.668275	5.568408	8.435283	5.854908