Explore Different Boroughs in London

1. Introduction:

1.1 Background

Despite the foreseeable changes Brexit may cause in UK economy, London, the capital city still holds its unique value and remain attractive to all tourists, students and investors around the world, demonstrated by the slow yet steady increase in London housing price trend.

However, among all difference regions in Greater London, what are the characters to consider and how to make sensible choices on where to live, or buy your property, therefore is extremely challenging for non-Londoners who are not familiar with the region.

Therefore, this analysis aims to gather basic understanding of the different boroughs in London, using London house market data available, in conjunction with supporting geographic information such as crime rate, and education data.

People who come to London to stay for a while, or who are keen to buy a property in London but with limited experience may find the analysis helpful.

2. Data

2.1 Data Source

UK Government website (https://www.gov.uk/) provide regular publications on UK House Price Index, which contains property prices for various property types, and the changes in the value of residential properties in England, Scotland, Wales and Northern Ireland. This data will form the base of analysis.

In addition to the above, a list of London boroughs has been sources from Wikipedia, to help narrow down the target areas for analysis.

Further residential information such as volume of reported crime, and number of different types of schools are also sources from UK government data store(link).

2.2 Data Cleansing

Python is utilized for the data analysis; sources data are mainly stored in csv or xlsx format that easily loaded into Python for data cleansing.

For the timeliness of the information, I decided to only use last 24-month data for house price, crime rate, and education information.

It is noted that the spelling of certain Boroughs is slightly different which lead to null values been produced when different data sources are joined together using Borough as a join key. Further data cleansing was performed to align the spelling of Boroughs, such as:" Westminster" is now changed to "City of Westminster", etc.

3. Methodology

The analysis employed Python as programming platform, and used various analysis packages such as Pandas, NumPy, and Matplotlib etc, for different features.

To gain a basic understanding of the regions, firstly we run a few tests to describe the data, using Pandas and demonstrate the result using data virilization technics.

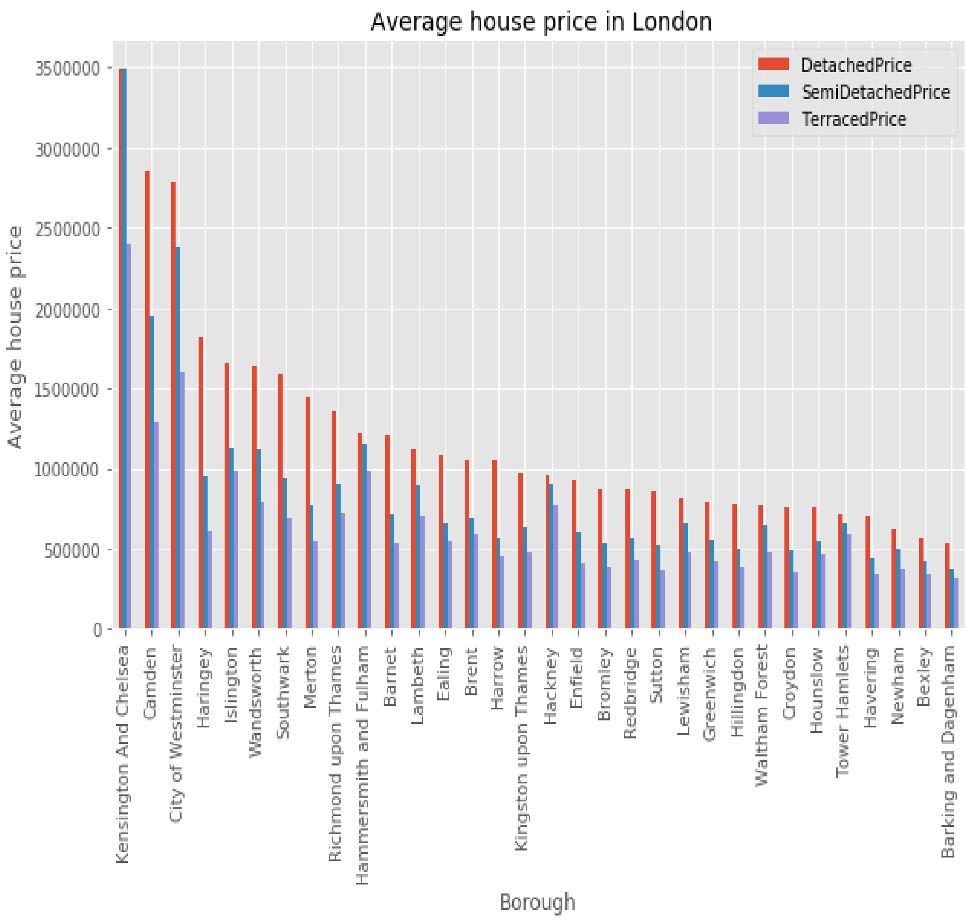
After that, we asked Python to performed unsupervised Machine Learning technicism, to cluster data in to different cluster.

3.1 Descriptive Statistics

Firstly, we chose to look at the prices of 3 house types (Detached, Semi-Detached, Terraced) in London over last 12 months for each Borough.

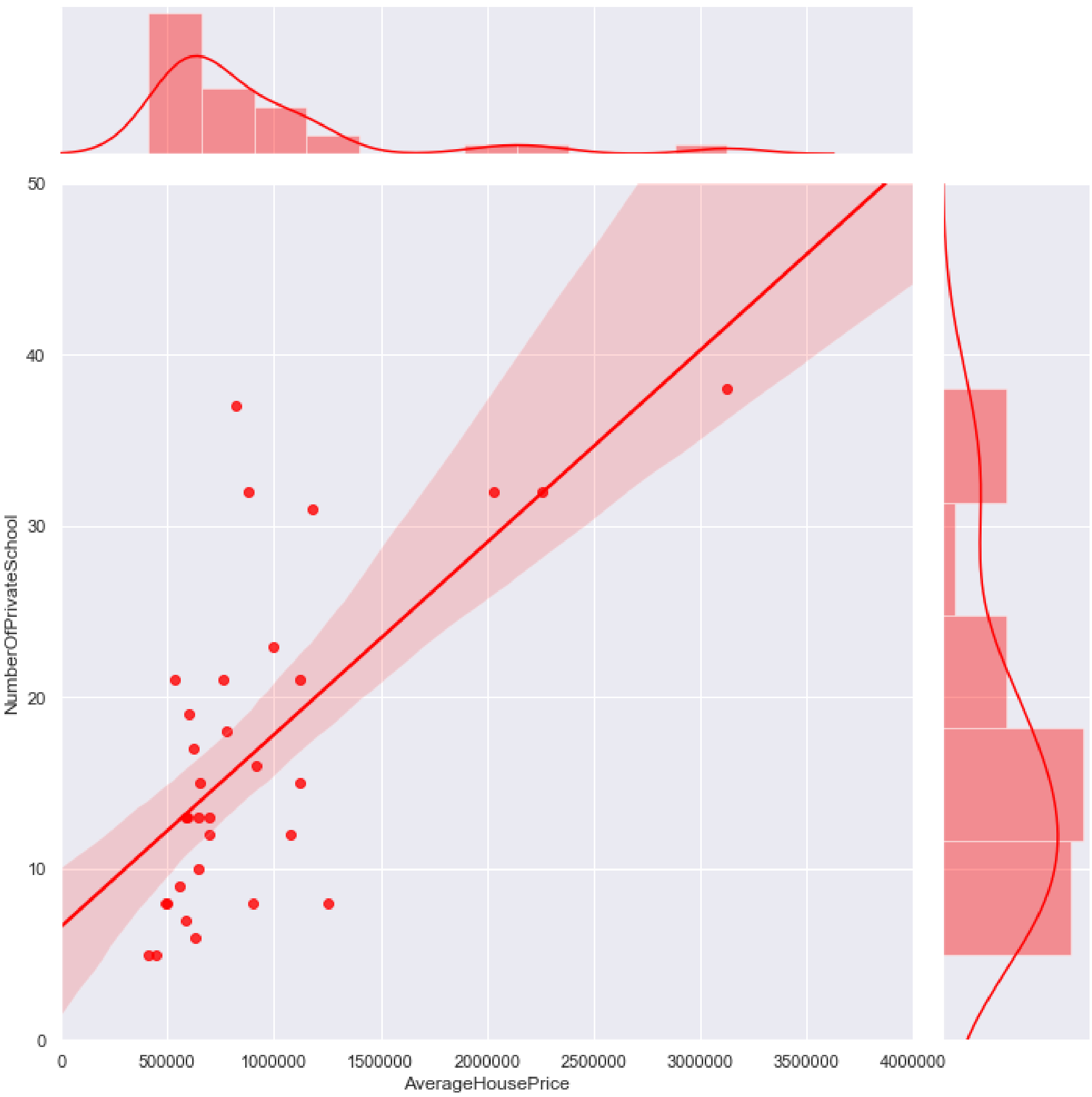
It can be observed from the bar chart below that, Kensington and Chelsea, Camden, City of Westminster have the highest three average prices among all London Boroughs, the Detached house price ranges from around 2.5 million pounds to 3.5 million pounds. The rest of 30 Boroughs have the average of detached houses ranges from around 0.5 million to 1.5 million pounds, with an observed average price around 1 million.

This chart provided readers with a basic understanding of house housing price to fit individual's budget consideration.

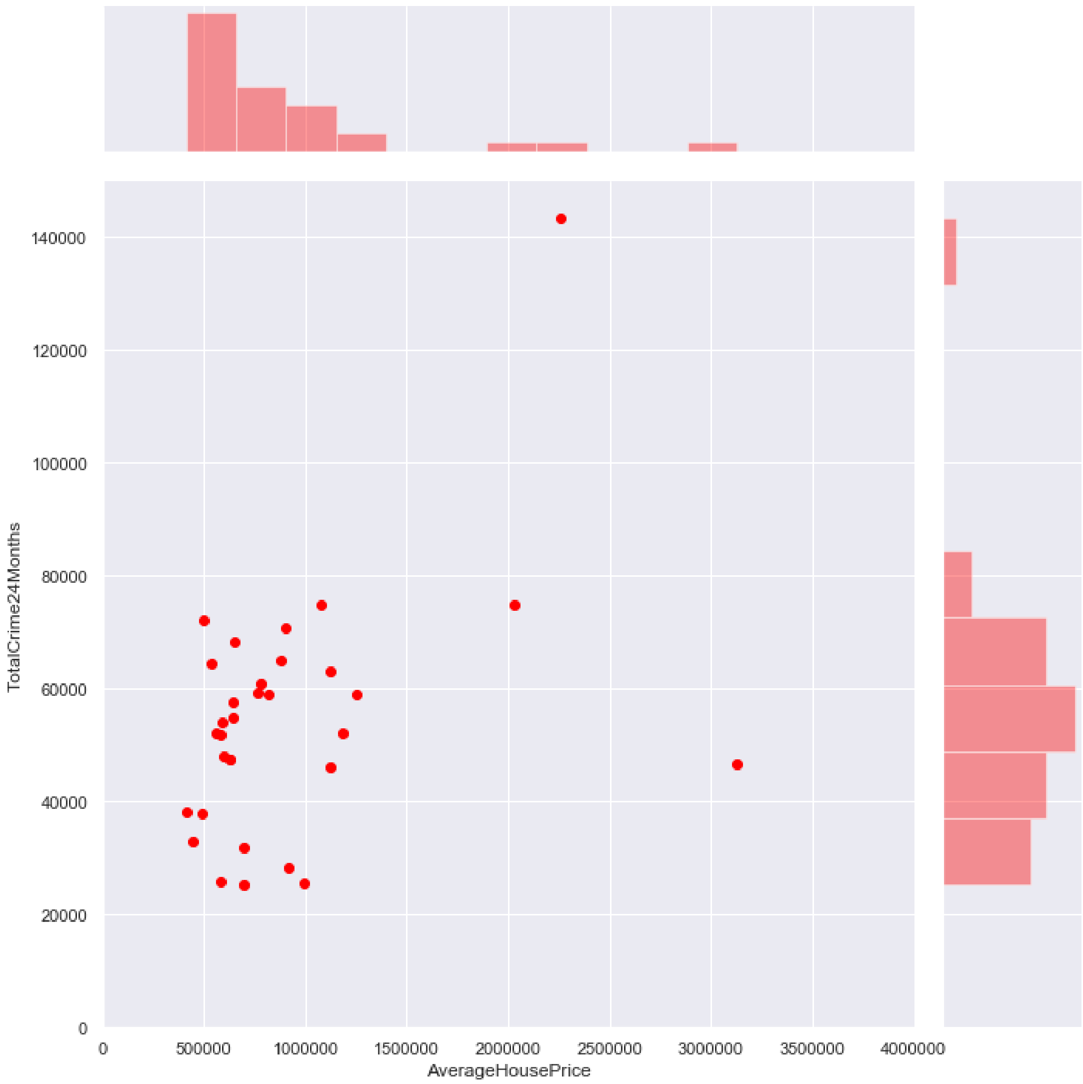


We then look at the value of information provided by 2 supporting data: number of crimes over last 24 months, and number of schools in the area.

It can be observed that a relatively clear positive correlation between average house price and number of private school: the lower average house price areas tend to have private school around 10 to 20 schools per area (cluster at the left bottom corner), whist when average house price increases, the number of private school goes up to around 20 to 30 school per Borough.



However, from the chart below, when plotting the number of reported crimes, there is no clear sign of corresponding relationships between average house price and volume of crime. We can see an extreme value at a relatively high house price area (2 to 2.5 million house price), with the highest crime volume of over 140k crimes reported over last 24 months. Therefore, we cannot draw the conclusion that the more expensive neighbourhood are definitely safer, sometime in real life, the high-income area may encourage more crime such as stealing or robbery.

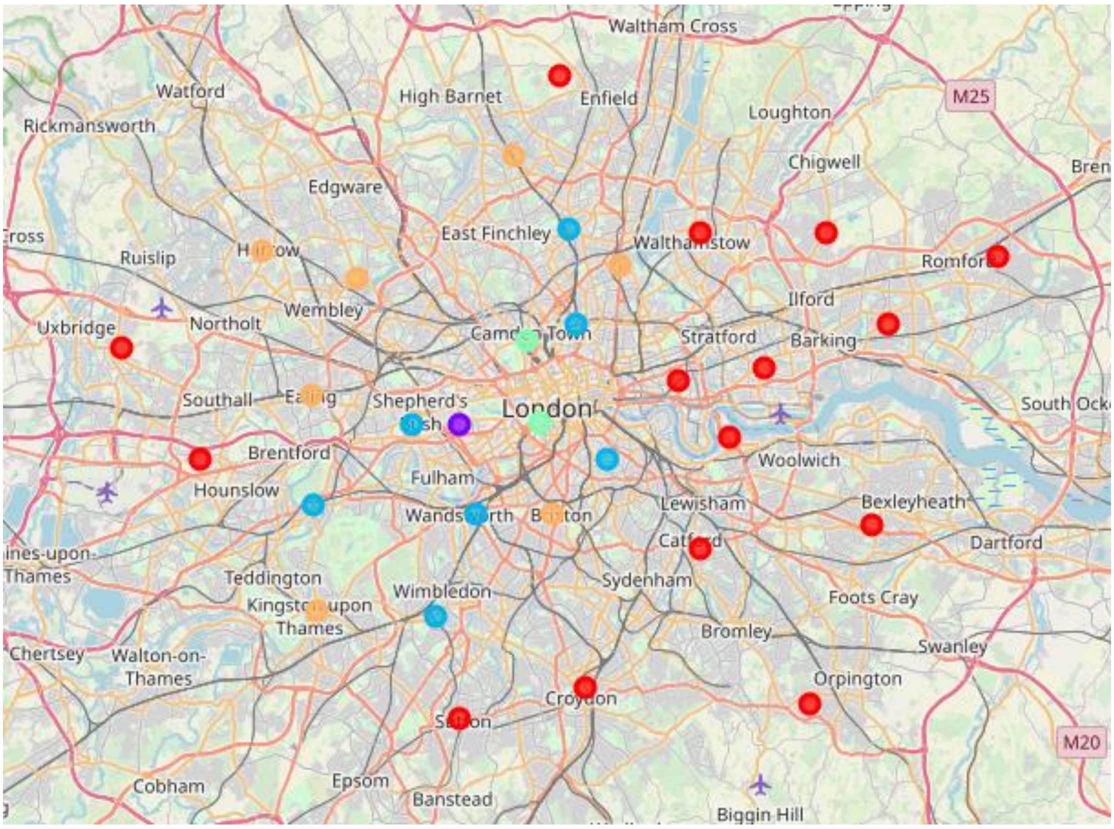


3.2 Data Clustering

Now we have a basic understanding on the average price range, distribution of crime reported in London, and numbers of schools in London, our next interest is to see how we can group all the Boroughs into a few buckets where each Borough will be evaluated against all other neighbors with regards to their similarities.

K- means clustering is utilized, with the below features considered:

Features introduced to data clustering						
Price of Detached house	Price of Semi-Detached house	Price of Terraced house	Number of public funded primary school	Number of public funded secondary school	Number of private schools	Total crimes reported within last 24 months



Each of the Boroughs in the analysis are grouped into 5 buckets, indicated by 5 different colors from the graphs shown above.

4. Results

Post K-means clustering analysis, it is noticed that the outsider Brough's of London are largely grouped together (Red dots). If we take a closer look at the result, each of the Borough have very similar housing price and numbers of schools, and as a group , this cluster has a quite different features when comparing to another group, for example the blue dots, which has a much significant higher price of housing, and more numbers of schools with lower number of reported crime.

As K-means is a way of unsupervised machine learning technique, we are not expecting the result to provide any opinion on where to buy or to live. It is a way to differentiate each region based on the features we feed into model. For example, a red dot from west London (Hounslow), share similar characteristics from a red dot from east London (Romford).

5. Discussion

Although data clustering provided an un-biased view of the grouping of regions, there are a few limitations of the analysis that worth to call out:

Firstly, the features selected to feed into modal are limited to only three categories: price, education and crime. Therefore, if we would like to bring in more features to better mimic real-life situation, such as region's population and diversity figure, public transportation information etc., the result can be different from what we see so far.

Secondly, we asked the machine to provide only 5 clusters in the outcome, by changing the requested number of clusters and feed more features as mentioned above, we can observe how machine can refine the result.

And lastly, there are still room to enhance the quality of the data feed. For example, the total number of crimes provides the frequency of the reported incidents, but not the severity of the incidents. If we compare an area with high crime rate but mainly robbery, to another area with lower crime rate but mainly murder, relying on pure volume will not provide sufficient insight.

6. Conclusion and further direction

To conclude, the analysis is helpful for any reader who have no or limited knowledge of living in London. The analysis grouped 33 regions in London, and provided 5 un-biased view of final outcome.

The analysis provided sensible directions on what are the similar boroughs based on the features selected, however, it is an unsupervised technique that the result only provides facts presented in data, rather than any opinions to readers.

Therefore, to make a sensible choice on where to live, readers can choose any particular cluster from the result and perform future analysis, to refine the model by adding additional supporting data.

In addition, it is recommended that the analysis can be further developed using Google API or other similar location data to reveal more insights on what the Borough like, for example bring in more features on what are the shops on high street etc.

Appendix

Cluster 0

	Borough	DetachedPrice	SemiDetachedPrice	TerracedPrice	total crime last 24 month	State-funded nursery	State-funded primary	State-funded secondary	Special schools	Independent	All schools	latitude	longitude	Cluster Labels
0	Barking and Dagenham	538728.815971	377073.153252	319039.078555	38231.0	0.0	44.0	13.0	2.0	5.0	65.0	51.546483	0.129350	0
2	Bexley	571584.413457	426410.955665	340698.307180	33099.0	0.0	57.0	16.0	5.0	5.0	84.0	51.451902	0.117179	0
4	Bromley	873546.185963	529320.575549	392514.148248	47973.0	0.0	77.0	19.0	4.0	19.0	122.0	51.367970	0.070062	0
8	Croydon	759048.276654	487158.067247	356658.193863	64392.0	5.0	86.0	25.0	6.0	21.0	144.0	51.376165	-0.098234	0
10	Enfield	930159.555061	596659.126418	408701.544361	57762.0	0.0	68.0	21.0	6.0	13.0	109.0	51.662291	-0.118065	0
11	Greenwich	793966.198834	560556.586051	421058.084906	54167.0	4.0	63.0	17.0	4.0	13.0	102.0	51.493367	0.009821	0
16	Havering	697922.745983	439848.911651	344801.333918	37853.0	0.0	61.0	18.0	3.0	8.0	91.0	51.577924	0.212083	0
17	Hillingdon	785044.786247	499075.210711	388245.840768	52096.0	1.0	69.0	22.0	9.0	9.0	111.0	51.535183	-0.448138	0
18	Hounslow	755847.867062	540223.426334	461322.268024	51863.0	0.0	53.0	20.0	5.0	13.0	92.0	51.482836	-0.388206	0
23	Lewisham	809687.729888	653849.275829	475823.287461	54853.0	2.0	63.0	14.0	5.0	10.0	95.0	51.441458	-0.011701	0
25	Newham	620601.512866	496778.365221	378831.650913	72057.0	7.0	66.0	23.0	2.0	8.0	109.0	51.525516	0.035216	0
26	Redbridge	868358.845533	570406.537589	437624.651014	47511.0	0.0	52.0	19.0	4.0	17.0	95.0	51.588612	0.082398	0
29	Sutton	859262.749785	525741.783393	367987.496668	25766.0	2.0	40.0	15.0	5.0	7.0	71.0	51.361428	-0.193961	0
30	Tower Hamlets	709958.925230	657957.325652	589825.633891	68210.0	6.0	70.0	20.0	6.0	15.0	119.0	51.520261	-0.029340	0
31	Waltham Forest	765047.365607	649242.913465	476564.785896	47579.0	3.0	51.0	18.0	4.0	6.0	85.0	51.588638	-0.011763	0

Cluster 1

	Borough	DetachedPrice	SemiDetachedPrice	TerracedPrice	total crime last 24 month	State-funded nursery	State-funded primary	State-funded secondary	Special schools	Independent	All schools	latitude	longitude	Cluster Labels
20	Kensington And Chelsea	3.494937e+06	3.489462e+06	2.396601e+06	46637.0	4.0	27.0	6.0	2.0	38.0	78.0	51.49908	-0.193825	1

Cluster 2

	Borough	DetachedPrice	SemiDetachedPrice	TerracedPrice	total crime last 24 month	State-funded nursery	State-funded primary	State-funded secondary	Special schools	Independent	All schools	latitude	longitude	Cluster Labels
13	Hammersmith and Fulham	1.215054e+06	1.154776e+06	988332.396183	46023.0	4.0	37.0	13.0	5.0	21.0	83.0	51.499016	-0.229150	2
14	Haringey	1.815304e+06	9.448097e+05	615613.119689	63013.0	3.0	64.0	14.0	6.0	15.0	104.0	51.590611	-0.110971	2
19	Islington	1.656366e+06	1.126603e+06	980549.474928	58947.0	3.0	46.0	11.0	6.0	8.0	78.0	51.546506	-0.105806	2
24	Merton	1.441685e+06	7.743666e+05	539612.555855	28389.0	0.0	44.0	9.0	3.0	16.0	73.0	51.409774	-0.210808	2
27	Richmond upon Thames	1.352620e+06	9.068283e+05	726477.845063	25471.0	1.0	45.0	11.0	2.0	23.0	82.0	51.461305	-0.303771	2
28	Southwark	1.586562e+06	9.438556e+05	692751.561823	74980.0	5.0	74.0	19.0	9.0	12.0	120.0	51.483448	-0.082088	2
32	Wandsworth	1.633030e+06	1.122656e+06	793280.378173	52162.0	3.0	61.0	11.0	7.0	31.0	116.0	51.457072	-0.181782	2

Cluster 3

	Borough	DetachedPrice	SemiDetachedPrice	TerracedPrice	total crime last 24 month	State-funded nursery	State-funded primary	State-funded secondary	Special schools	Independent	All schools	latitude	longitude	Cluster Labels
5	Camden	2.850226e+06	1.952578e+06	1.286210e+06	74864.0	1.0	42.0	10.0	5.0	32.0	94.0	51.539026	-0.142552	3
7	City of Westminster	2.783676e+06	2.384244e+06	1.603275e+06	143349.0	4.0	40.0	13.0	3.0	32.0	93.0	51.500175	-0.133233	3

Cluster 4

	Borough	DetachedPrice	SemiDetachedPrice	TerracedPrice	total crime last 24 month	State-funded nursery	State-funded primary	State-funded secondary	Special schools	Independent	All schools	latitude	longitude	Cluster Labels
1	Barnet	1.213280e+06	714684.582475	536867.328748	59112.0	4.0	91.0	26.0	6.0	37.0	166.0	51.625149	-0.152936	4
3	Brent	1.055258e+06	695558.102353	587059.275344	60983.0	4.0	60.0	15.0	4.0	18.0	103.0	51.567281	-0.271057	4
9	Ealing	1.087245e+06	661041.885124	541934.840918	59413.0	4.0	68.0	16.0	6.0	21.0	117.0	51.513254	-0.304314	4
12	Hackney	9.603845e+05	909784.311727	772467.509688	65138.0	2.0	58.0	16.0	3.0	32.0	113.0	51.573445	-0.072438	4
15	Harrow	1.052568e+06	570707.666496	455674.838746	31820.0	1.0	42.0	13.0	4.0	13.0	75.0	51.580559	-0.341995	4
21	Kingston upon Thames	9.722135e+05	639571.847528	472129.336597	25390.0	1.0	36.0	11.0	3.0	12.0	64.0	51.412330	-0.300689	4
22	Lambeth	1.120551e+06	895116.223354	698312.693072	70894.0	5.0	60.0	19.0	5.0	8.0	99.0	51.457148	-0.123068	4

Data Source

List of London boroughs	https://en.wikipedia.org/wiki/List_of_London_boroughs
UK House Price Index	https://www.gov.uk/government/publications/about-the-uk-house-price-index
Data Dictionary	https://www.gov.uk/government/publications/about-the-uk-house-price-index/about-the-uk-house-price-index
Recorded Crime: Geographic Breakdown	https://data.london.gov.uk/dataset/recorded_crime_summary
Schools and Pupils by Type of School, Borough	https://data.london.gov.uk/dataset/schools-and-pupils-type-school-borough

Python Notebook

https://github.com/linyang1987/Coursera_Capstone/blob/master/Data%20Science%20Final%20Project.ipynb