Applied Data Science Capstone Project

1. Introduction/Business Problem:

Brexit has instilled uncertainty into the UK economy and has inevitably weighed on housing prices since the referendum decision in 2016. The market will be looking towards October's deadline. But what can we expect? Well, whether we get a deal or no-deal, one could argue that risk to real estate prices is skewed to the upside from hereon. In the event of an undesirable no-deal scenario, businesses would still gain clarity and can suitably plan for the near-term; moreover, a no-deal scenario would likely keep GBP depressed against other major G-10 currencies, inherently making property relatively cheap for foreign investors.

So this begs the question – as a real estate investor who is about to anticipate a turn in property prices, how can we quickly identify pockets of land in Central London that are undervalued in order to make an informed investment decision?

Target Audience:

Real Estate investors

Stakeholders:

- Buyers
- Real Estate agents

2. Data Section

This section will detail what data we will be using.

HM Land Registry: Price Paid Data

- Duration: 2019 YTD data
- Description: This dataset includes information on all property in England and Wales that are sold for full market value and lodges with them for registration

Rightmove API

- Duration: Real-time listings
- Description: Rightmove is one of the UK's largest online portals to search properties for sale and to rent in the UK

Foursquare Location Data API

- Duration: Real-time
- Description: To determine proximity of various amenities

3. Methodology

To gather a list of current residential listings, I scraped Rightmove's site in real-time to obtain all the relevant information in a dataframe. Let's take a peek at the first 5 rows:

	price	type	address	url	agent_url	postcode	number_bedrooms	search_date
0	430000.0	2 bedroom apartment for sale	Kingsway, North Finchley, N12	http://www.rightmove.co.uk/property-for-sale/p...	http://www.rightmove.co.uk/estate-agents/agent...	N12	2	2019-06-24 18:14:45.803466
1	550000.0	2 bedroom flat for sale	Bell Street, Marylebone, London, NW1	http://www.rightmove.co.uk/property-for-sale/p...	http://www.rightmove.co.uk/estate-agents/agent...	NW1	2	2019-06-24 18:14:45.803466
2	1075000.0	2 bedroom flat for sale	Hyde Park Square, Hyde Park Estate, London, W2	http://www.rightmove.co.uk/property-for-sale/p...	http://www.rightmove.co.uk/estate-agents/agent...	W2	2	2019-06-24 18:14:45.803466
3	339995.0	2 bedroom terraced house for sale	Sandhurst Road, London, N9	http://www.rightmove.co.uk/property-for-sale/p...	http://www.rightmove.co.uk/estate-agents/agent...	N9	2	2019-06-24 18:14:45.803466
4	365000.0	3 bedroom flat for sale	Raglan Road, Walthamstow, London, E17	http://www.rightmove.co.uk/property-for-sale/p...	http://www.rightmove.co.uk/estate-agents/agent...	E17	3	2019-06-24 18:14:45.803466

The search_date column shows when the data was scraped by python.

I computed the average price for all the properties that fell under one particular postcode, producing a dataframe similar to the below (note only the first 5 rows has been shown):

	postcode	price
0	CR0	475000.000000
1	CR7	540000.000000
2	E1	653262.666667
3	E10	470000.000000
4	E11	647498.750000

Using HM Land Registry's data on price paid for all properties in 2019, computed the mean price sold per postcode and merged this dataset with the one shown above so as to compare current average price per postcode vs paid average price per borough in 2019 YTD – the objective here is to yield some information as to whether properties within a particular postcode are undervalued or overvalued.

Snapshot of the data:

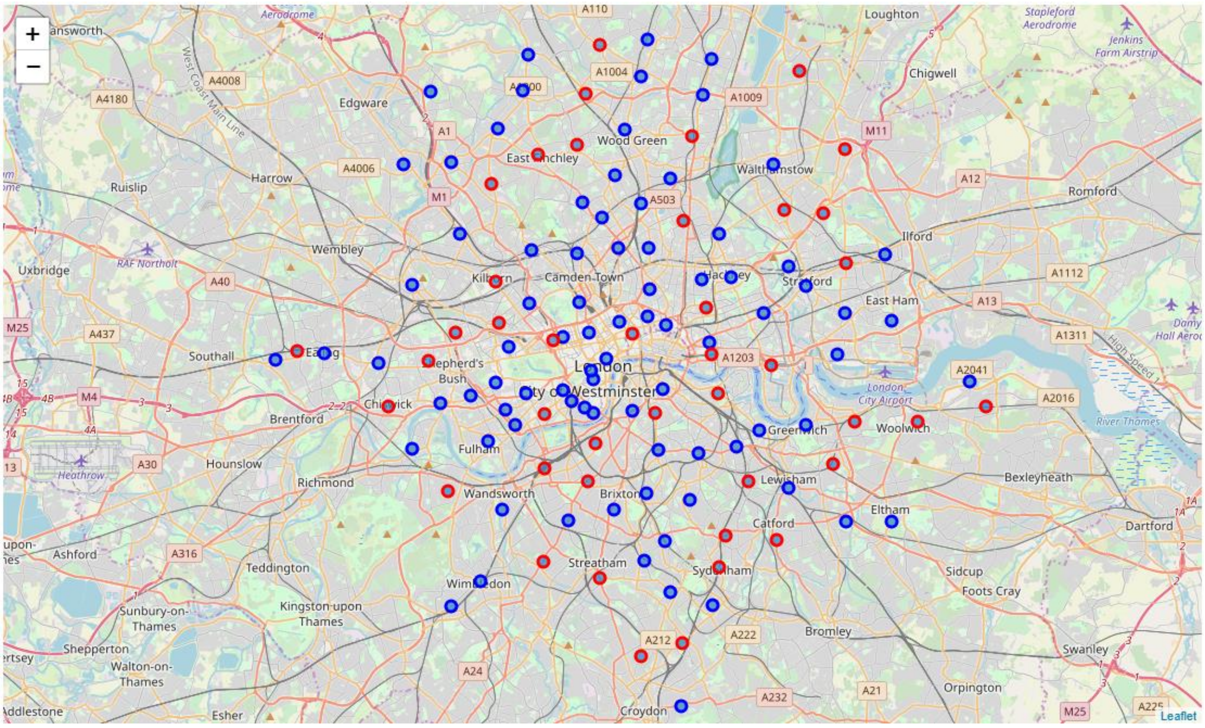
	postcode	price	Average of price_paid	Average of Latitude	Average of Longitude
0	CR0	475000.000000	4.835799e+05	51.373218	-0.078137
1	E1	653262.666667	1.246067e+06	51.516253	-0.060406
2	E12	437497.500000	5.257662e+05	51.551025	0.050848
3	E13	319250.000000	3.690769e+05	51.528171	0.025739
4	E15	390000.000000	3.932639e+05	51.538726	0.000629

I used python's folium library to visualize pockets of undervalued and overvalued land, using latitude and longitude values for each postcode to project this onto an interactive map:

Key:

Blue Circle = Undervalued postcode

Red Circle = Overvalued postcode

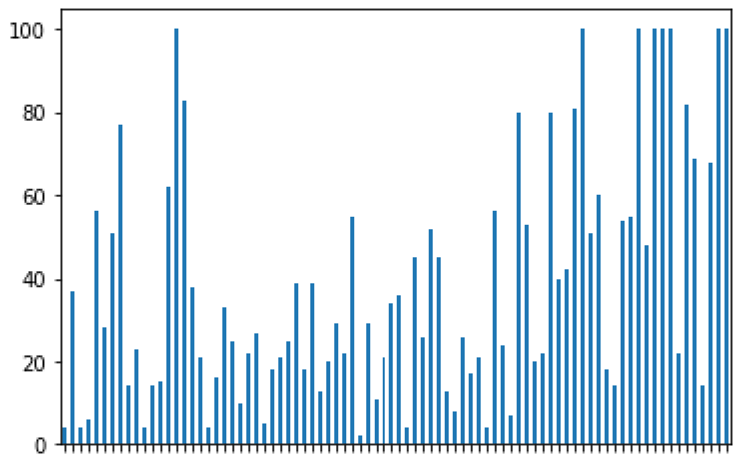


Unsurprisingly in this market, we see many pockets of land which are undervalued. For the purpose of this analysis and as real estate investors, we will only consider the undervalued pockets of land.

Utilizing FourSquare's API to explore venues nearby the undervalued postcodes, I set a limit of 100 venues and an exploration radius of 500 meter from their respective latitude and longitude values. Looking at a snapshot of the output for the first postcode (CR0):

	Postcode	Postcode Latitude	Postcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	CR0	51.373218	-0.078137	Doughmasters	51.376033	-0.074297	Sandwich Place
1	CR0	51.373218	-0.078137	Sandilands London Tramlink Stop	51.375094	-0.077862	Tram Station
2	CR0	51.373218	-0.078137	The Cricketers	51.375100	-0.083706	Pub
3	CR0	51.373218	-0.078137	Kiosk	51.370674	-0.083797	Candy Store

The bar chart below shows that for some postcodes, our FourSquare API returned up to 100 venues:



295 unique venue categories were returned by FourSquare.

We then used FourSquare's API to gather the top 10 venues in each postcode, example below:

	Postcode	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
0	CR0	Tram Station	Pub	Candy Store	Sandwich Place	Yoga Studio	Fish Market	Farm	Farmers Market	Fast Food Restaurant	Film Studio
1	E1	Hotel	Indian Restaurant	Pub	Coffee Shop	Grocery Store	Steakhouse	Turkish Restaurant	Sandwich Place	North Indian Restaurant	Flower Shop
2	E12	Train Station	Gym / Fitness Center	Restaurant	Event Space	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Film Studio	Fish & Chips Shop
3	E13	Bus Station	Pub	Café	Gym	Yoga Studio	Flower Shop	Fast Food Restaurant	Film Studio	Fish & Chips Shop	Fish Market
4	E15	Platform	Hotel	Pub	Café	Sandwich Place	Coffee Shop	Bookstore	General Entertainment	Bar	Supermarket

Given that we have some common categories across different postcodes, we can use an unsupervised machine learning method called K-Means to cluster the postcodes together by similarity.

We will run a K-Means test to cluster the boroughs into 5 clusters, producing a merged table with cluster labels similar to the below:

	Postcode	Postcode Latitude	Postcode Longitude	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
0	CR0	51.373218	-0.078137	0	Tram Station	Pub	Candy Store	Sandwich Place	Yoga Studio	Fish Market	Farm	Farmers Market	Fast Food Restaurant	Film Studio
1	CR0	51.373218	-0.078137	0	Tram Station	Pub	Candy Store	Sandwich Place	Yoga Studio	Fish Market	Farm	Farmers Market	Fast Food Restaurant	Film Studio
2	CR0	51.373218	-0.078137	0	Tram Station	Pub	Candy Store	Sandwich Place	Yoga Studio	Fish Market	Farm	Farmers Market	Fast Food Restaurant	Film Studio
3	CR0	51.373218	-0.078137	0	Tram Station	Pub	Candy Store	Sandwich Place	Yoga Studio	Fish Market	Farm	Farmers Market	Fast Food Restaurant	Film Studio
4	E1	51.516253	-0.060406	1	Hotel	Indian Restaurant	Pub	Coffee Shop	Grocery Store	Steakhouse	Turkish Restaurant	Sandwich Place	North Indian Restaurant	Flower Shop

4. Results

Upon running exploratory data analysis, our algorithm has identified 84 investable postcodes and clustered them into 5 clusters for the investor to choose what type of amenities they would like their investments to be near to.

It is clear that most instances fall under Cluster 1:

	Cluster Labels
1	3074
0	144
3	6
4	4
2	4

Let’s take a look at the type of venue categories within Cluster 1:

	1st Most Common Venue
Hotel	648
Coffee Shop	562
Pub	478
Café	375
Grocery Store	171
Italian Restaurant	148
Clothing Store	122
Exhibit	100
Theater	100
Fast Food Restaurant	71
Platform	56
Art Gallery	42
Pizza Place	39
Cricket Ground	29
Bookstore	22
Chinese Restaurant	22
Gym / Fitness Center	18
Bus Stop	14
Supermarket	13
Asian Restaurant	11
Bar	10
French Restaurant	8
Historic Site	7
Laundromat	4
Furniture / Home Store	4

This cluster could be labelled as hospitality given the number of hotels within it.

5. Discussion

Based upon the findings in the results section, the investor can now make a conscious decision to decide which ‘undervalued’ cluster would fall into his/her investable universe given the amenities within each cluster.

We could take this analysis further by building a linear regression model to individually value houses and measure that output vs the current listing price to determine if each property is undervalued

6. Conclusion

The following conclusions can be made:

- Knowledge about real-time market prices can be very helpful for the investor
- Knowledge about differing cluster segments can help the investor expand his/her investable universe