The Battle of the Neighbourhoods Capstone Project

2. Data

2.1 Data Sources

In this project, data was used from three sources: London DATASTORE, Foursquare API, and LocationIQ.

2.1.1 London DATASTORE datasets

London DATASTORE is a free and open data-sharing portal to access data related to London. The London DATASTORE has numerous data resources. From this source, the following datasets were utilised:

1. Ward profiles and atlas dataset: this dataset presents 63 key summary measures on London wards and the populations within them. The profiles were created using the most up-to-date information available at the time. The most recent version was published in 2015. Some measures may be outdated, but should still provide an adequate overview of the wards. For example, the GCSE data available is related to the old system, however more recent data is only available at the Borough level and given the relative importance families often place on the quality of secondary education, it is important to get a picture of performance at the ward level.

Fig 2.1 Head of ward profiles and atlas dataset

	Ward name	Old code	New code	Population - 2015	Children aged 0-15 - 2015	Working-age (16-64) - 2015	Older people aged 65+ - 2015	% All Children aged 0-15 - 2015	% All Working-age (16-64) - 2015	% All Older people aged 65+ - 2015	Mean Age - 2013	Median Age - 2013	Area - Square Kilometres	Population density (persons per sq km) - 2013	% BAME - 2011	% Not Born in UK - 2011
0	City of London	00AA	E09000001	8100	650	6250	1250	7.96206	76.8868	15.1512	41.3039	39	3.15148	2538.06	21.4	36.7
1	Barking and Dagenham - Abbey	00ABFX	E05000026	14750	3850	10150	750	25.9583	69.0142	5.02748	29.5	29	1.3	10500	71.9	57.3
2	Barking and Dagenham - Alibon	00ABFY	E05000027	10600	2700	6800	1050	25.7013	64.2769	10.0217	33.8	33	1.4	7428.6	29.9	24.7

2. Key stage 1 and Key stage 2 results by borough for 2018: these datasets provide percentage measures of the children achieving the expected levels in reading, writing, maths and science at the Borough level. This is considered an adequate proxy for primary education level quality within Boroughs. There was no data available for primary achievement at the ward level.

Fig 2.2 Head of key stage 2 results by Borough dataset

	Unnamed: 0	Unnamed: 1	All	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unna
0	NaN	NaN	Number of eligible pupils	NaN	NaN	NaN	Percentage reaching the expected standard	NaN	NaN	NaN	Percentage reaching the higher standard	NaN	
1		NaN	English Reading	English Writing	Mathematics	All (reading,writing and maths)	English Reading	English Writing	Mathematics	All (reading,writing and maths)	English Reading	English Writing	Mather
2	Code	Area	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	E09000001	City of London	29	29	29	29	76	86	90	72	45	62	
4	E09000002	Barking and Dagenham	3411	3411	3411	3410	75	83	80	67	27	43	

3. Land registry average house prices by ward dataset: this provided average (mean) house prices at the ward level for years up to and including 2017.

Fig 2.3 Head of land registry mean house prices by ward dataset

	New code	Ward name	Borough name	Year ending Dec 1995	Year ending Mar 1996	Year ending Jun 1996	Year ending Sep 1996	Year ending Dec 1996	Year ending Mar 1997	Year ending Jun 1997	Year ending Sep 1997	Year ending Dec 1997	Year ending Mar 1998	Year ending Jun 1998	Year ending Sep 1998	Yea endin De 199
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Na
1	E09000001	City of London	City of London	-	-	-	-	-	-	-	-	-	-	-	-	
2	E05000026	Abbey	Barking and Dagenham	51077.6	49868.8	49901.6	51935.1	50766.5	50189.5	47807.5	48004.6	51318.4	53046	56650.8	62273.5	62804.
3	E05000027	Alibon	Barking and Dagenham	45490.4	44701.5	44486	45894.1	46145.1	47019.2	48608.8	50142.5	51232.4	51968.8	52586.1	54027.2	56402.

2.1.2 Geocoder and LocationIQ

To obtain latitude and longitude coordinates of all London wards, the provider LocationIQ was used alongside geocoder. LocationIQ provides geocoding based on OpenStreetMap's Nominatim.

Fig 2.4 Head of ward and coordinates dataframe generated by geocoder and LocationIQ

	Ward loc	Latitude	Longitude
0	City of London, City of London	51.515618	-0.091998
1	Abbey, Barking and Dagenham	51.535688	0.075530
2	Alibon, Barking and Dagenham	51.547549	0.153114
3	Becontree, Barking and Dagenham	51.549265	0.127538
4	Chadwell Heath, Barking and Dagenham	51.567904	0.128041

2.1.3 Foursquare

Using the latitude and longitude coordinates, Foursquare located the wards and returned Foursquare location venue data for all wards.

The explore endpoint was used to ensure that the Foursquare API explored and returned popular spots around each location: by learning about the popular spots we can get a sense of the character of each area – what there is to do, the facilities available and what the current residents and visitors enjoy in each area.

With the average area of London wards being 2.6km, the radius for the search within each ward was set to 1300m, and venues limited to 100 to ensure a sufficient number of venues for each ward that would assist in drawing conclusions on the character of the area. Foursquare returned 429 unique venue categories for the 571 London wards.

Fig 2.5 Head of dataframe containing ward venue data generated by Foursquare API

	Ward loc	Ward Latitude	Ward Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	City of London, City of London	51.515618	-0.091998	Goodman Steak House Restaurant	51.514398	-0.090745	Steakhouse
1	City of London, City of London	51.515618	-0.091998	The Ned Hotel	51.513565	-0.090166	Hotel
2	City of London, City of London	51.515618	-0.091998	The Merchant House	51.513264	-0.093039	Cocktail Bar
3	City of London, City of London	51.515618	-0.091998	Daunt Books	51.513982	-0.092995	Bookstore
4	City of London, City of London	51.515618	-0.091998	Hawksmoor Guildhall	51.515647	-0.090997	Steakhouse

2.2 Data cleaning and feature selection

2.2.1 Government datasets

On examining the individual datasets, there was a lot of data that was not relevant to the problem at hand.

The ward profiles and atlas dataset contained many measures relating to the health of the existing population and other such measures, which would not be particularly relevant to a family looking to move into the area. The focus when selecting measures to retain was on those variables that property buyers find important when looking for a property. The measures retained for the purposes of this project were: new ward code (for initial merging with other datasets); Crime rate - 2014/15; Average GCSE capped point scores – 2014; % area that is open space – 2014; Average Public Transport Accessibility score – 2014; % detached houses – 2011; % semi-detached houses – 2011; % terraced houses – 2011; % Flat, maisonette or apartment – 2011.

The Key stage 1 and Key stage 2 datasets contained multiple years. The data for the most recent year (2018) was selected, and the average percentage across all subjects (reading, writing, maths and science) was retained. The overall average of Key stage 1 and Key stage 2 combined was calculated as to provide a single overall measure of the quality of primary education.

The house price dataset contained data from 1995. Given the fluctuating nature of property prices in London, and the need to know prices as close to current reality as possible to make an informed decision, only the most recent data was retained – year ending 2017.

The Ward name format on ward profile dataset was not conducive to a latitude-longitude coordinate search, and so the ward name combined with Borough name from the house prices dataset was used instead.

Broader level summary data, such as UK level Key stage 1 data, was removed from all datasets to ensure they did not skew the data and results.

2.2.2 Geocoder and Location IQ

Once the datasets were merged, the wards were converted into a list and run through the geocoder to obtain latitude and longitude coordinates.

Running the geocoder revealed some of the ward names as being ambiguous to the provider and so producing incorrect coordinates. To combat this issue, the problematic ward names were changed to a location central to the ward e.g. Queen's Park, Westminster (the geocoder was returning a similar named location in New Zealand) was changed to Queen's Park Library. Some continued to be returned as incorrect, and so were corrected

The resulting coordinates were added to the ward dataframe. The resulting dataframe contained a total of 571 wards.