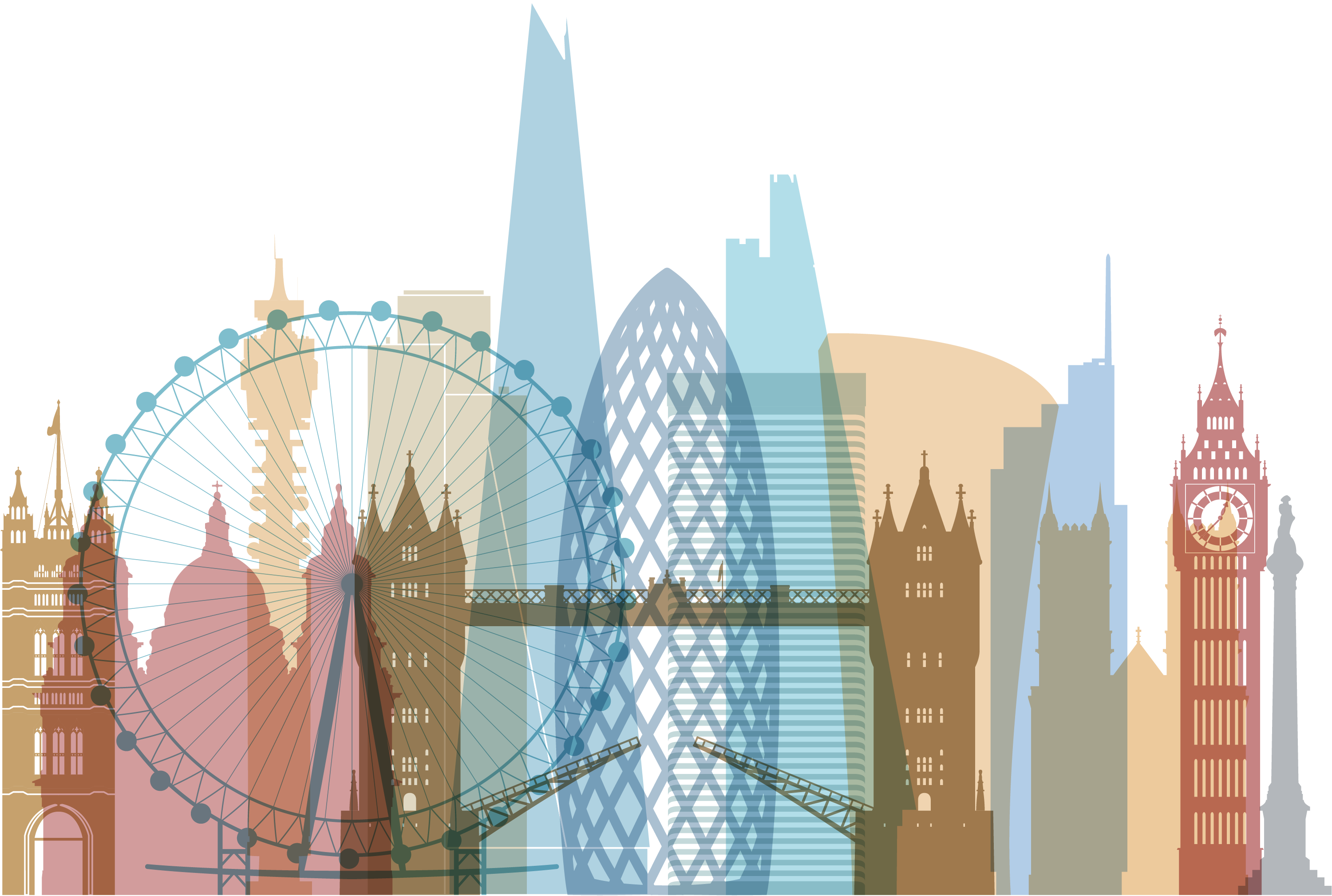
EXPLORING LONDON



IBM DATA SCIENCE PROFESSIONAL CERTIFICATE CAPSTONE PROJECT

KATE REES – AUGUST 2020

**Exploring London**

# Introduction

I live and work in London, UK: currently ranked as the 19th most expensive city in the world to live in. I am interested in exploring whether location data, such as that available with Foursquare, can predict which parts of the city are more expensive, and which are least expensive.

Well-documented in the UK press is the so-called ‘Waitrose Effect[[1]](#footnote-2)’: house prices near a Waitrose store – a high-end supermarket chain in the UK – are higher. I am interested in exploring this idea further. I intend to use Foursquare’s API to identify concentrations of different categories of amenities in each area, and combine it with house price data to look for correlations, and use data science methodology to explore whether we can use Foursquare’s data to predict house prices in a given London ward.

The results of this project would be of interest to anyone wanting to invest in property in London. House buyers could use the results spot ‘up and coming’ areas before house prices rise, and property companies could use the results to decide where to invest in new building projects - and to estimate how much apartments might sell for in that area.

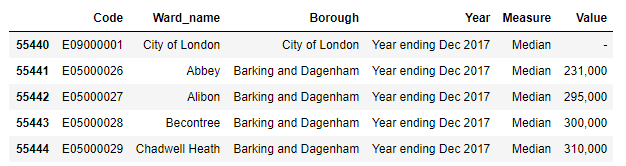
# Data

### Data Source 1: London House Price Data

The Mayor of London’s office publishes statistics on house prices for each London Borough, and subdivided into Wards. [It is available to download here](https://data.london.gov.uk/download/average-house-prices/59be940c-ffb8-426d-a833-6146ea77de5c/land-registry-house-prices-ward.csv) from the London Datastore.

This house price data contains values (in GBP) of both Mean and Median house prices in each ward. To decide which metric to use, I compared the distribution of both, and found that the median was was closer to a normal distribution and therefore less affected by outliers – e.g. a few very expensive properties.

A sample of the data is shown below:



**Missing data:**

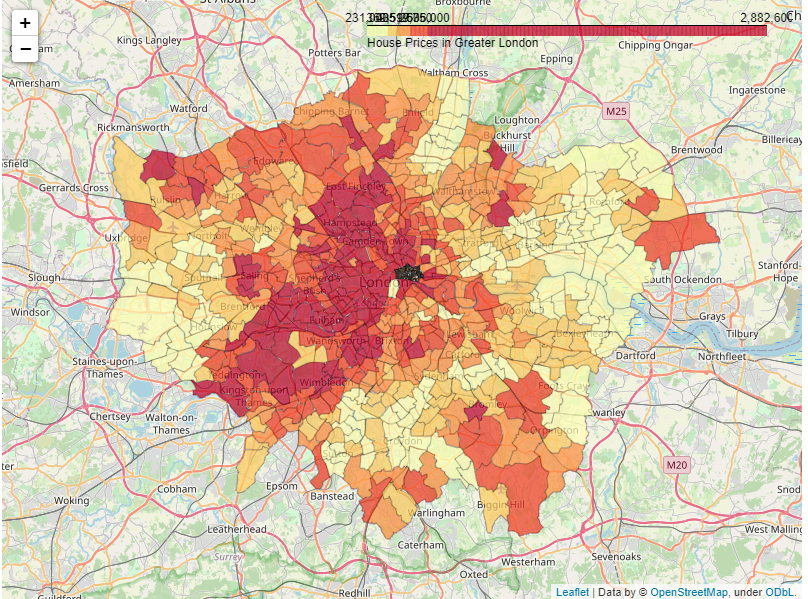
Notice there is no value for house prices in the City of London – this is the central business / financial district, and falls under a different jurisdiction to the rest of Greater London. I decided to ignore the City of London for the purposes of this project.

### **Data Source 2:** London geographic data

The house price data does not contain details of the geographic position of the Wards and Boroughs, so we will need a separate data source for this.

I found a geoJSON file with boundary data for London wards on the Financial Times’ visual and data journalism Github repository at [https://github.com/ft-interactive](https://raw.githubusercontent.com/ft-interactive/geo-data/master/uk/london/london-wards-2014.geojson)

This will allow us to define the boundaries of each London ward. The Choropleth Map below shows the London Wards coloured by their house price value – with Red being the Highest value areas, and pale yellow being the Lowest. The black region is the City of London, for which we have no data.



**Data Source 3:**  Foursquare Location Data

Foursquare is best known for it’s ‘City Guide’ social media application that enables users to discover places of interest near their location. It also provides an API, which allows developers access to a rich source of location data.

We will use the GeoJSON obtained above to calculate a ‘centroid’ location within each Ward, and use the Foursquare API to discover venues within a certain radius of each Ward centroid.

An example of the data obtained for each Ward is shown below:



1. <https://www.propertywire.com/news/uk/research-reveals-waitrose-effect-uk-house-prices/> [↑](#footnote-ref-2)