House Prices within the catchment of Outstanding Primary Schools in London, UK

Gaurav Kukreja

March 05, 2020

1. Introduction

1.1 Background

London is the capital and largest city of the United Kingdom, standing on the River Thames in the south-east of England. It's considered financial capital of the world and has been termed as one of the most innovative, investment friendly and popular for working professionals. London exerts a considerable impact upon the arts, commerce, education, entertainment, fashion, finance, healthcare, media, professional services, research & development, tourism and transportation. It is a house of highly ranked institutions for higher education including Imperial College London, London School of Economics, University College London and King's College etc. It attracts people from all around the globe to work and settle. As a result, there is constant influx of migration every year from all age groups, from inside and outside UK.

8.8m

live in London, 13% of the UK's population. London is growing at twice the rate of the UK as a whole.

The Government has a motto of providing free quality school education for every child, rich or poor equally and fairly. There are about 30k state schools all over the country. In order to maintain healthy competition and desirability among those, an assessment is conducted regularly, and Ofsted rating is assigned to each school i.e Outstanding, Good, Requires Improvement and Inadequate. They are rated on the areas of (a) leadership and management; (b) quality of teaching, learning and assessment; (c) personal development, behaviour and welfare and finally (d) outcomes for children and learners.

1.2 Problem

It is a well-known fact that the education is a key determinant of lifetime earnings, poverty risk and prosperity. The low pay and employment status are clearly associated with lower levels of education. There are independent (aka private) schools in operation as well including Eton, Harrow School etc., that are quite prestigious and boasts of delivering leaders to the country. But they come at a hefty cost which most can't afford.

In that case, it is a duty of a parent to ensure that their children get the best possible education at a state school, ideally in an Outstanding Rated School which offers higher probability of success in life ahead, either in further education, employment or entrepreneurship. It all starts with foundation which commences at a Primary School. UK got approx. 20k of those. Out of about 2k outstanding rated primary state schools, 342 of those reside in London where most professionals work and live. As per usual rules, a first child from a household is admitted to a specific school if residence is within the catchment of a school, which makes it critically important to reside within a certain radius of a school

The problem here is (1) It comes at a cost. The house prices within catchment of outstanding schools can cost somewhere between 10-20% premium; (2) Quite difficult to filter out which outstanding school to target out of 342 in London, especially for new immigrants; (3) Which area may be suitable for interest of parents (do they like parks, cafes, pubs, restaurants, theatre etc.) and most important (4) Affordability

1.3 Solution

We will use Data Science methodology to balance all the problem areas above to find a best possible solution for parents depending on their interests and affordability. The aim is to provide a simple and clear options to support well-informed decision making of buying a house in right area for them. After all, this is one of most important investment of life.

1. Data acquisition and cleaning

2.1 Data Sources

We will need below data sources available from different sources:

(a) Ofsted Report web portal

https://reports.ofsted.gov.uk/search?q=&location=&radius=&latest_report_date_start=&latest_report_date_end=&level_1_types=1

This shows details of all the schools across United Kingdom including State Schools, Independent Schools, Special Schools and Pupil Referral Units along with their relevant Ofsted ratings. The info is presented over web as html files on multiple pages. We will use BeautifulSoup Python library to pull it and convert into csv file.

(b) Wikipedia List of Areas in London

https://en.wikipedia.org/wiki/List_of_areas_of_London

This has a web table of Location, Borough, Town and PostCodes mainly. We will use pandas library to read it and convert it into dataframe.

Location		London borough		Post town		Postcode district	Dial code	OS grid ref
Abbey Wood		Bexley, Greenwich [7]		LONDON		SE2	020	TQ465785
Acton		Ealing, Hammersmith and Fulham[8]		LONDON		W3, W4	020	TQ205805
Addington		Croydon[8]		CROYDON		CR0	020	TQ375645
Addiscombe		Croydon[8]		CROYDON		CR0	020	TQ345665
Albany Park		Bexley		BEXLEY, SIDCUP		DA5, DA14	020	TQ478728

(c) Property Data

https://propertydata.co.uk/cities/london

This produces a table with useful property price info at a postcode level. We will use pandas library to read and convert into dataframe.

Area	Avg yield	Avg price	£/sqft	5yr +/-	Explore data
BR1	3.7%	£447,159	£466	+25%	Explore data
BR2	3.1%	£513,115	£469	+27%	Explore data
BR3	3.7%	£456,224	£488	+26%	Explore data
BR4	-	£601,225	£460	+24%	Explore data

(d) Foursquare Venue Details

https://developer.foursquare.com/

This is an excellent developer portal to search or explore venues (restaurants, bars, hotels, parks etc.) near any location within certain radius of a location (by postcode, town, borough etc.). It also dispenses category of venues which is quite useful for our analysis. We will use developer API to extract the relevant info for our project.
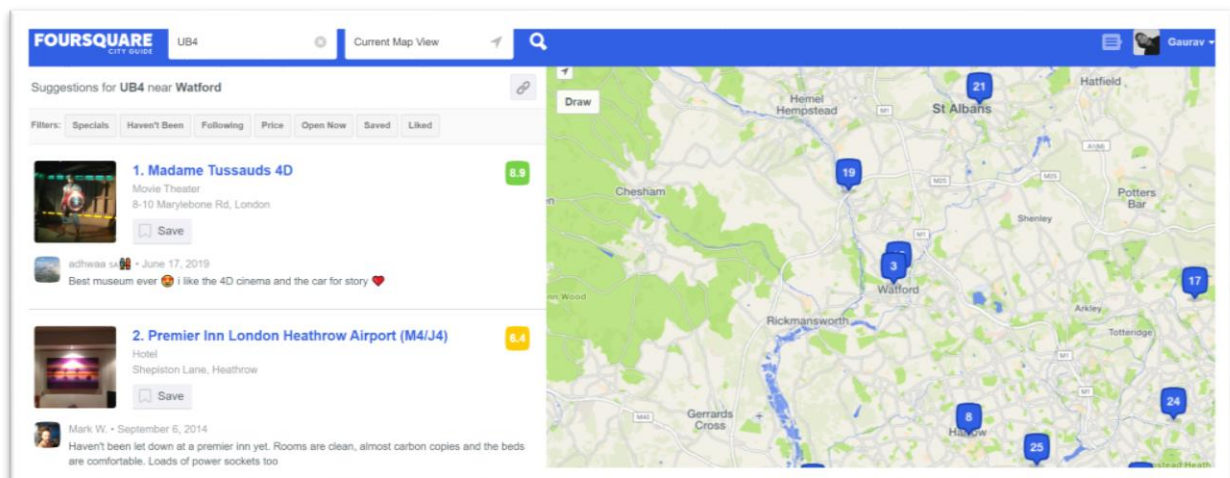
2.2 Data Cleaning

Data downloaded or scraped from first 3 sources above were combined into 1 table after many adjustments to each set.

In Ofsted School Report data, there were several irrelevant information. For example, we are only aiming for State Primary schools thus we removed Independent & Special Schools as well as Pupils Referral Units. We also like to see Ofsted assessment for recent years (2018 or onwards) as assessment conducted many years ago may not sustain today. We also filtered down on Outstanding Schools only. Since other set of data has PostCode as common denominator, we combined multiple rows at that level, picking info from School addresses.

In property pricing data, we only need Postcode and Price, thus removed rest of the information. We had to remove commas and currency sign from numbers to make it work for further analysis.

In Wikipedia table for London areas, we had to clean up a lot to make it usable. We removed the reference #s given next to Boroughs in square brackets. Some postcodes had unnecessary spaces or no space, when needed, so adjusted that. Many postcodes were linked to same location, so we created multiple lines to join it with other set of data tables easily.

2.3 How will that be used

After combining 3 tables into 1, we would come down to required lines of data only for unique Postcodes/areas which we will search to get location coordinates using geocoder library and arcgis maps provider.

We will use those coordinates to get venues within a certain radius (1000m) using Foursquare API which will be used to cluster those locations into different labels based on category of venues nearby (Hotel, Bars, Gyms, Theatres etc.). We intend to use k-means algorithm which is most common cluster method of unsupervised learning. We will then display those clusters on a map using folium library.

We will also use pyplot chart to represent average house prices (adding 20% on top as premium for the school catchments) of each area within each cluster labels which will help parents to view different price points within their preferred cluster label (which is basically their interest types like Social Venues vs. Pub & Restaurants vs. Golf vs. Theatres or combination of those depending on what is produced by cluster labels using this algorithm). An example visualization of that is below.