Analysing London Real Estate Data in Real-time using K-means Clustering

IBM Applied Data Science Capstone Project

Adam Clark

10/05/2020

1 Introduction

1.1 Background

London is an ever changing beast. Over the 33 boroughs there are around 250 unique areas. On a daily basis these neighbourhoods change in demographics, atmosphere and appeal. Certain areas undergo gentrification and turn from neighbourhoods full of pubs and offices to family friendly areas with cafes, art galleries and schools. Other areas retain cultural appeal with an influx of bars and restaurants whilst some even delineate along ethnic lines. It would take decades of living in the city to become acquainted with all of it's many communities. Even then, some of the communities will have changed completely over that time.

At the same time London house prices are some of the most expensive in the world, total sales last year accounted for over 10 billion. So this is a huge market with lots of competition. Running a real estate agency in London you need to be able to meet your client's needs accurately and quickly to stay in the game.

1.2 Problem

In a day you may have several clients in to discuss purchasing property in London. One may be a wealthy individual looking to be close to theaters and bars. The next may be a family looking for somewhere quiet but commutable. So how do you know where to start with the scale of a city like London? How can you pinpoint areas that reflect their needs? How can you scour a city

of 250 areas quickly and efficiently? How can you do it in real-time so that your London map doesn't become outdated as areas change?

1.3 Interest

The main factor of interest will probably be real estate agencies looking to increase efficiency in how they approach the categorisation of London into market areas. Other interest may come from personal individuals looking to move around London based upon their own needs.

2 Data Acquisition and Cleaning

2.1 Data Sources

We compile several 'live' data sources from across the internet to maintain an up-to-date picture of London:

- The UK wide postcode data, to allow us to accurate determine Lat/Long positions for each area in london. Data is here.
- London Borough data with postcodes, to allow us to group the Lat/Long Positions of each area. Data is here.
- All of each area's categorical data i.e. shops, restaurants, bars. We get this by passing the Lat/Long positions to the Foursquare API. This returned 100 venues within a 500m radius of the area centre.
- For the house price layer we can use the UK government's published figures. Data is here.
- For the borough boundaries we used a GEOJson file containing all the boundaries. Data is here.

2.2 Data Cleaning

The UK wide postcode data needed to be cut down as it represented all the possible postcodes in the country and was about 9000 rows long. This is where the postcode to borough data from wikipedia came in. By scraping this data from the wikipedia page we can extract all the postcode rows that are in the London Boroughs list by pairing on 'region' or 'Borough.

The Borough data did need some cleaning as some of the strings for 'Borough' contained "[Note 1]" and "&" instead of "and". A simple str -

replace function sorted this. The cleaned data served as a mask to put on the UK wide data to give us data just pertaining to London Boroughs.

Using the Lat/Long data from the postcodes we could pull all the local 'venues' from the foursquare API. This was done by defining a function that pull the relevan columns from the JSON data including the venue name and category along with the neighbourhood Lat/Long/Name. This was put into a new data frame.

This data frame was converted into a 'one hot' data frame grouped by the area to clean and format the data ready for K-Means clustering.

The county boundary GEOJson needed no cleaning.

The UK house price data were already clean, it was just a matter of only pulling out the latest london borough data and transposing.

3 Data Analysis

3.1 Clustering Parameters

The one hot data for each area was fed into a K-Means clustering algorithm to apply cluster labesl to each area. The SciKitLearn KMens algorithm was used with random cluster centers to start the model.

As an unsupervised data set the cluster counts were adjusted until the data appeared to cluster correctly on the map. This was clear by using the house price overlay to check the cluster numbers.

It was determined in this fashion that 9 clusters managed to accurately separate central London from the suburbs of London.

3.2 Cluster label features

By doing analysis on the clusters and their top venues for each location the following table was devised to explain how the data was labelled by the model.

This would be used to help identify and segment each client into one of the 0 categories below. A family may prefer cluster 5, a house share of students may prefer cluster 4 and a wealthy city worker may prefer cluster 1.

Cluster Label	Cluster Colour	Key features
0	Light grey	Out of town retail centres: Furniture Stores, Supermarkets
1	Black	Commercial centres with Hotels, Theatres, Restaurants, Coffee shops, Offices and Bars
2	Purple	Residential areas with Coffee Shops, Restaurants, Supermarkets & Mid-budget Families
3	Dark Green/BLue	Out of own Parks, Yoga Studio and Falafel Restaurants (Mill Hill)
4	Light Green/Blue	Park areas with Transport Links and Restaurants & Wealthy Families
5	Dark Green	Residential Areas with Grocery stores/supermarkets (quieter suburbs)
6	Light Gren	Residential Areas with a high density of pubs
7	Yellow	Ethinically diverse areas, predominantly Eastern European (Chingford)
8	Red	Ethinically diverse areas, predominantly Indian

3.3 Results

We can see the clustering results with the house prices overlay underneath in the sequence of images that follows.

This ouotput is an interactive folium map that can be used to explore greater London which similar neighbourhoods and their associated house prices.

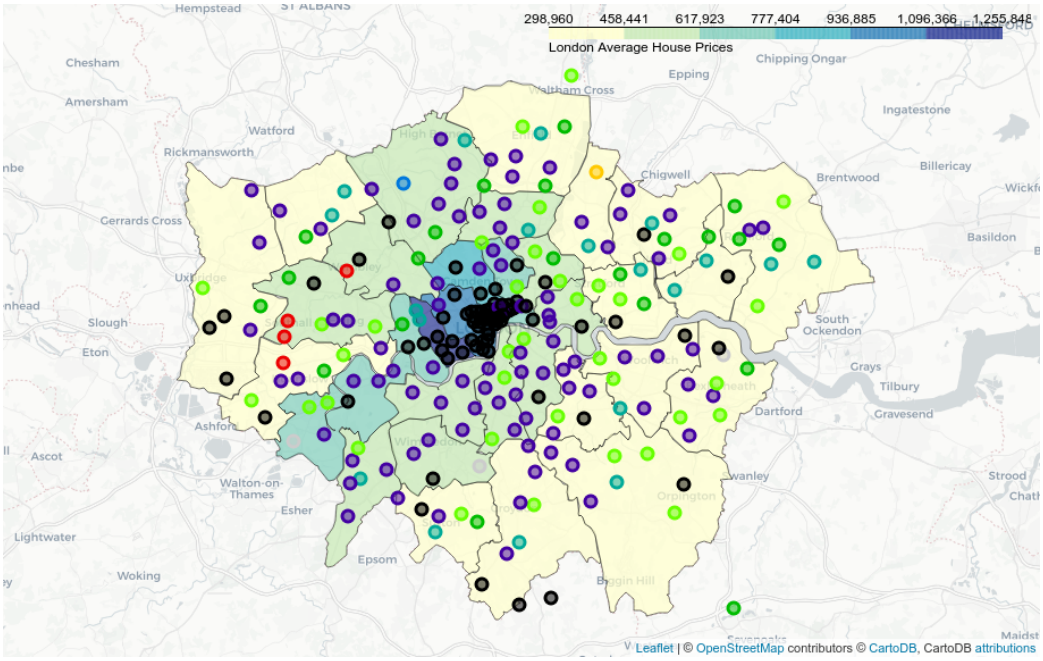

Figure 1: A Map of LDN Clusters Zoom = 9.

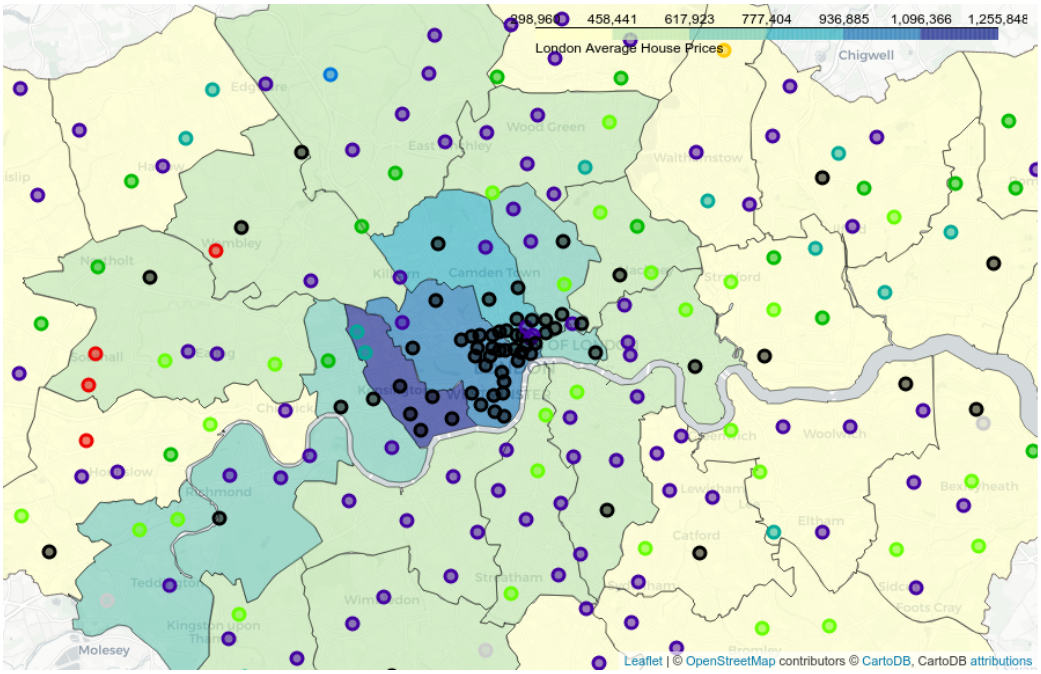

Figure 2: A Map of LDN Clusters Zoom = 10.

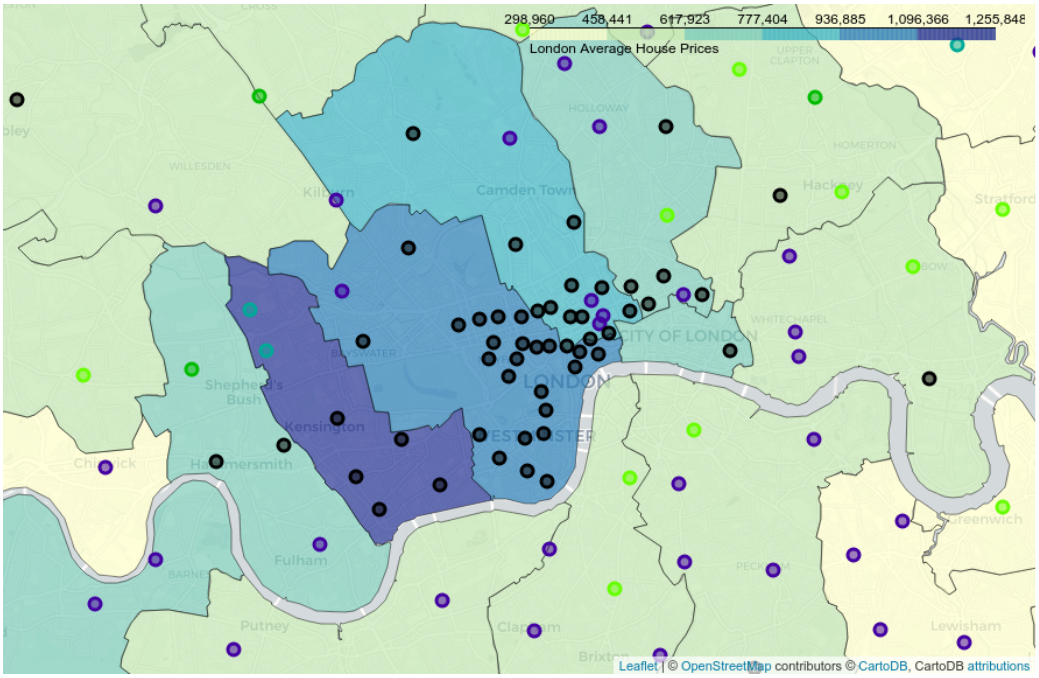

Figure 3: A Map of LDN Clusters Zoom = 11.

4 Conclusions

In this study we have succesfully used K-Means to cluster similar neighbourhoods and overlaid price data with the clusters onto an interactive map. This map would prove invaluable to increasing the efficiency of a real estate agency in London to identify target areas for clients to begin house hunting.

5 Further Improvement

I feel it would be good to incorporate the pricing data into the K-Means algorithm.

Then a function could be applied where a client gives a series of criteria they are looking for in an area - e.g. 'close to pubs', 'in the price range x to y', 'lots of galleries nearby'. Then areas can be provided to the client based upon their likes. Essentially adding a recommender function.

This could be done with the onehot encoded data set.