Capstone Project – The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera.

Problem: Where to start search for relocation in central London and the Home Counties

1. Introduction/ Business Problem

London is a large cosmopolitan city home to a population with very different needs and interests. Many people that live in the British capital will eventually move out of central London to one of the Home Counties, i.e., the counties that surround London:

Surrey, Kent, Sussex, Buckinghamshire, Berkshire, Essex and Hertfordshire.

The decision to move out might be driven by family needs, a wish for a more outdoor lifestyle or many other reasons. Relocation companies sometimes help with this process. For people that were brought up in the area, or know it extremely well, this is an easy decision. For others, especially foreigners, it is a minefield, as it covers a massive physical area.

The purpose of this analysis is to help them in their search. We assume that if they like one area (defined by postcode district in our analysis) in central London, by clustering the data we might find a postcode district in the home counties with the same feel (most common venues). Or they might have been to an area they like in the home counties and are looking for something similar in another county.

If it is impossible to cluster the data to find an exact matching postcode district, at least we can provide a list of most common venues for each postcode district/county. This should give a feel for each postcode district and be a starting point for a physical search.

To aid this analysis we provide the average house prices for each home county and a simple prediction of future house prices, should these continue to grow at recent rates.

2. Data

Following is the data required to solve the above problem and how each type will be used.

- **Postcode Districts in London and Home Counties and respective coordinates**
  This data looks very similar in format to the data used for the Toronto project, i.e., postcode (e.g. KT2), Name, County/London and longitude and latitude coordinates. The data comes from a GitHub csv file, which is based on Ordnance Survey data. This data will be the baseline to connect postcode districts, location coordinates and wider area location (London/County).

- **London coordinates** from geocoder to get the Folium map for London.

- **Foursquare data of venues within 500 yards of each postcode district**.
  The analysis will be very similar to the clustering analysis in the Coursera course to find the top/most common venues in each borough/postcode district. We will use this data to cluster postcode districts and find similar ones.

- **House price data by county**. The Office for National Statistics provides a downloadable file of quarterly average house prices by county - from 1995.
  To aid this analysis we provide the average house prices for each home county and a simple prediction of future house prices, should these continue to grow at recent rates.

3. Methodology

Data Loading, Cleaning and Exploratory Analysis

Postcode Districts in London and Home Counties and respective coordinates

This data is loaded from a csv file. It contains 2856 rows (postcodes) and 10 columns. We start by filtering the data of the UK regions where our counties are: South East of England, East of England and London. Some region data (county) has to be overwritten with the ceremonial county. City of London and Greater London are defined to be the same region – London. The data is filtered again to include only London and our defined home counties. We now have 637 postcodes.

The columns that are of interest to us are postcode, town, region (county), latitude and longitude.

The data is further filtered to provide London postcodes for one part of the analysis.

Please note that in this document where we mention postcode, we always mean postcode district.

London coordinates – latitude and longitude are acquired and stored.

Foursquare data of venues within 500 yards of each postcode district

Given the postcode latitudes and longitudes for London and the Home Counties we acquire the Foursquare data for all venues in a 500-yard radius of the coordinates. The data downloaded provides the name, coordinates and category of each venue. 7378 venues were downloaded.

House price data by county

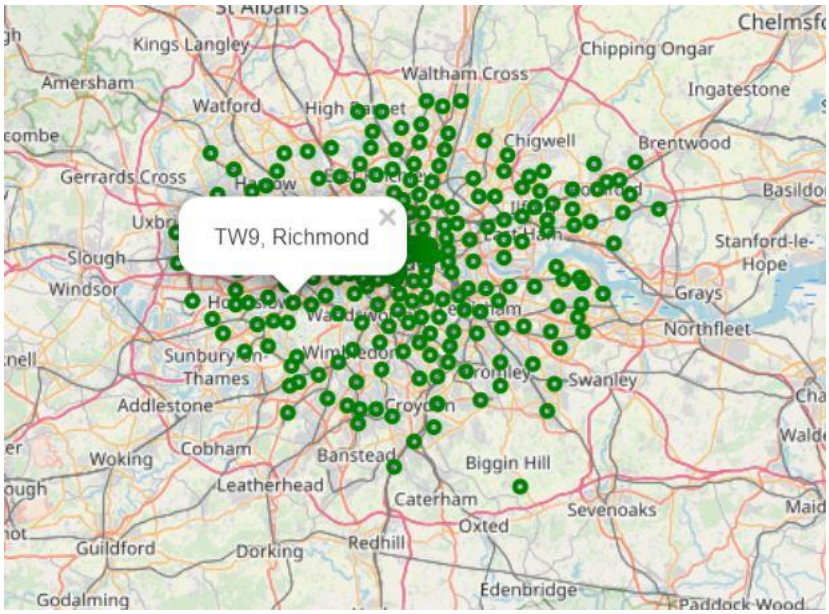
This data was downloaded from a csv file and it includes mean house price data (quarterly) by county between 1995 and 2016. More recent data by county was not available. As above, we filtered the data for the counties we were interest in. Berkshire was missing from the data. A request for this data has been sent to the ONS.

For the analysis, we transposed the data to make the counties the columns and created a year column and an index.

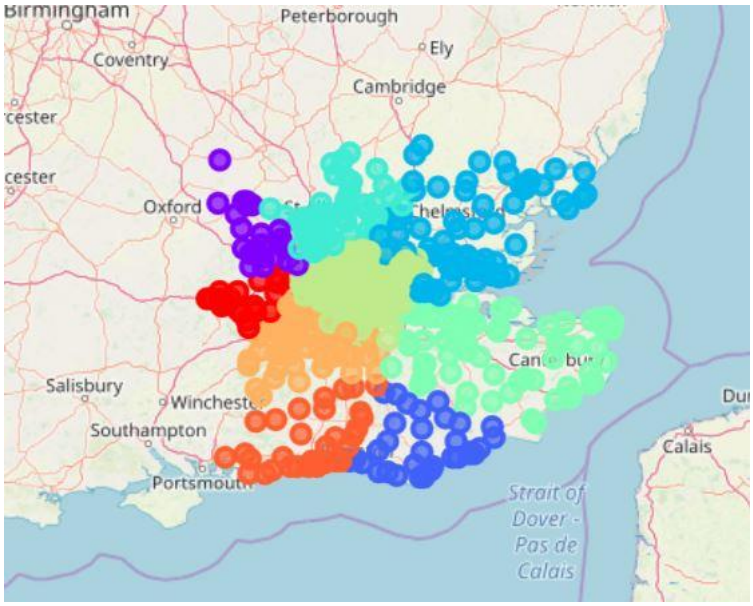
Postcode Methodology

The first step was to use Folium Maps to plot the London postcodes and London and Home Counties. This is to give us a feel for the areas:

London Postcodes



London and the Home Counties Postcodes



Once the data was downloaded, we needed to understand what it meant for each postcode. So, we defined the top 10 venues for each postcode. We created a table of the number of venues by postcode and category. We hope that by defining the top 10 most common venues in each postcode, we can capture a bit of its essence. The output is a table of the top 10 category of venues by postcode. These are a few lines from the table:

	Postcode	Top 1st Venue	Top 2nd Venue	Top 3rd Venue	Top 4th Venue	Top 5th Venue	Top 6th Venue	Top 7th Venue	Top 8th Venue	Top 9th Venue	Top 10th Venue
0	AL1	Platform	Bookstore	Fried Chicken Joint	Café	Grocery Store	Coffee Shop	Lebanese Restaurant	Mediterranean Restaurant	Breakfast Spot	Chinese Restaurant
1	AL10	Chinese Restaurant	Noodle House	Sandwich Place	Grocery Store	Supermarket	Pharmacy	Pool	Japanese Restaurant	Food Court	Fish & Chips Shop
2	AL2	Chinese Restaurant	Home Service	Park	Furniture / Home Store	Department Store	Flower Shop	Farmers Market	Fast Food Restaurant	Field	Film Studio
3	AL4	Pub	Plaza	Grocery Store	Yoga Studio	Flea Market	Farm	Farmers Market	Fast Food Restaurant	Field	Film Studio
4	AL5	Pub	Italian Restaurant	Coffee Shop	Supermarket	Pizza Place	Mediterranean Restaurant	Bar	Bookstore	Fish & Chips Shop	Park

Clustering or Grouping

Now we are ready to cluster the data. This is where we believe the analysis adds value by being able to identify similar areas. The data is clustered into different groups using k-means clustering. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

By clustering the data into 3, 5 and 10 clusters/groups, we felt we could capture the granularity/similarity we need for each part of our analysis.

Once the data is clustered, we join it to the existing postcode data. Each postcode has a Cluster label assigned to it:

Postcode	Top 1st Venue	Top 2nd Venue	Top 3rd Venue	Top 4th Venue	Top 5th Venue	Top 6th Venue	Top 7th Venue	Top 8th Venue	Top 9th Venue	Top 10th Venue	Cluster label	latitude	longitude	town	region
AL1	Platform	Bookstore	Fried Chicken Joint	Café	Grocery Store	Coffee Shop	Lebanese Restaurant	Mediterranean Restaurant	Breakfast Spot	Chinese Restaurant	0	51.74836	-0.32237	St Albans	Hertfordshire
AL10	Chinese Restaurant	Noodle House	Sandwich Place	Grocery Store	Supermarket	Pharmacy	Pool	Japanese Restaurant	Food Court	Fish & Chips Shop	0	51.75958	-0.22920	Hatfield	Hertfordshire
AL2	Chinese Restaurant	Home Service	Park	Furniture / Home Store	Department Store	Flower Shop	Farmers Market	Fast Food Restaurant	Field	Film Studio	0	51.72064	-0.33353	St Albans	Hertfordshire
AL4	Pub	Plaza	Grocery Store	Yoga Studio	Flea Market	Farm	Farmers Market	Fast Food Restaurant	Field	Film Studio	1	51.77133	-0.29398	Sandridge	Hertfordshire
AL5	Pub	Italian Restaurant	Coffee Shop	Supermarket	Pizza Place	Mediterranean Restaurant	Bar	Bookstore	Fish & Chips Shop	Park	0	51.81622	-0.35177	Harpenden	Hertfordshire

The data is now ready to be plotted and analysed in case studies 1, 2 and 3 for 3, 5 and 10 clusters respectively. Or for postcodes being grouped in 3, 5 and 10 types of neighborhoods.

Please note that these are all starting points for searching new places. Different starting points or modelling data could change the endpoints. Data is also not static and can change outcomes.

House Prices Methodology

The house price analysis is supposed to give an overall view of house prices in each county. We plot the existing data and use Linear Regression to extend (predict) the available data to further years.

In no way is this a ‘real prediction’ of house prices, just a feel for what the prices would be should they continue to rise at a similar rate. To create a proper house price prediction model, more factors would be needed, different ways of dealing with time series and different models.

4. Results

Postcode Results

We split the results into 3 case studies:

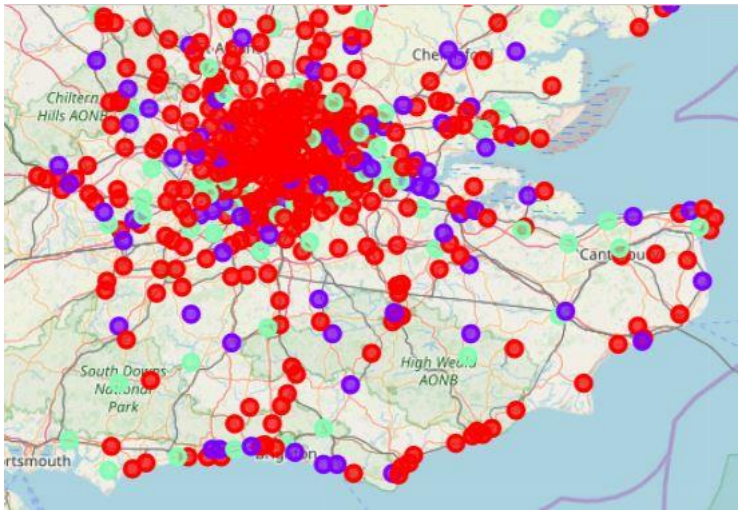
1. Postcodes split into 3 clusters or similar groups – High Level Overview
2. Postcodes split into 5 clusters or similar groups – relocation from Eltham to Home Counties
3. Postcodes split into 10 clusters or similar groups – relocation from Horsham to Kent

Case Study 1

Postcodes split into 3 clusters or similar groups – High Level Overview

The user of this information wishes to have a generic feel for London and the Home Counties. There is a predominant cluster (red group) which covers most of Central London, the M23 corridor and so on. This is hardly surprising as, at a high level, we expect those areas to be reasonably homogeneous.

London and the Home Counties – 3 Groups

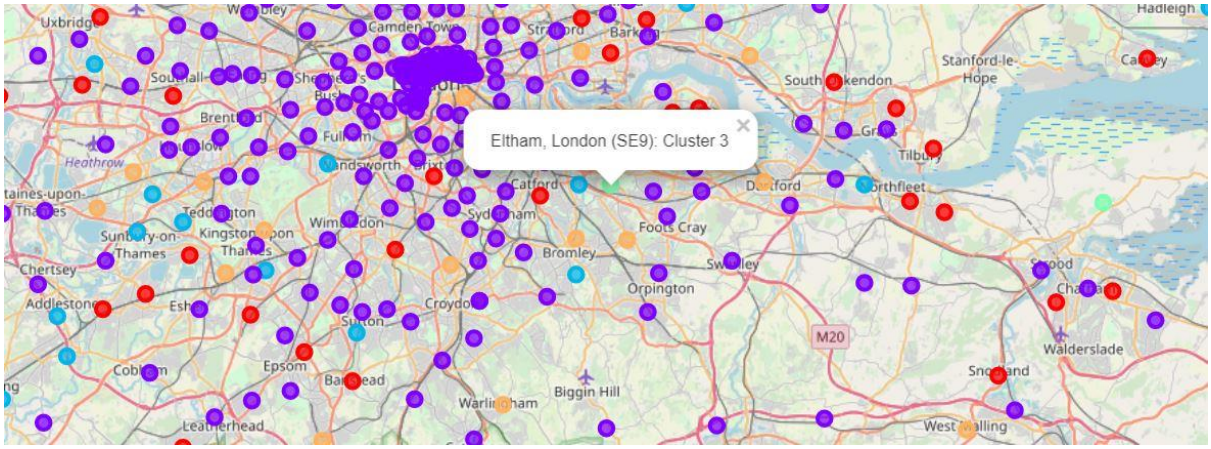


Case Study 2

Postcodes split into 5 clusters or similar groups – relocation from Eltham to Home Counties

Ms Smith who lives in Eltham wants to move to the Home Counties. She wants the new area to have a similar feel but not necessarily exactly the same. Grouping the data into 5 groups feels just right.

London and the Home Counties – 5 Groups



Eltham turns out to be in a small cluster of only five postcodes for this analysis. By looking at the top venues we can see a recurring theme of golf courses, yoga and other venues which we can understand how these neighborhoods probably ended up grouped together.

	Postcode	Top 1st Venue	Top 2nd Venue	Top 3rd Venue	Top 4th Venue	Top 5th Venue	Top 6th Venue	Top 7th Venue	Top 8th Venue	Top 9th Venue	Top 10th Venue	Cluster label	latitude	longitude	town	region
30	BN6	Gas Station	Golf Course	Yoga Studio	Flower Shop	Farmers Market	Fast Food Restaurant	Field	Film Studio	Fish & Chips Shop	Fish Market	3	50.92918	-0.15276	Clayton	West Sussex
194	GU4	Medical Supply Store	Golf Course	Yoga Studio	Food	Fast Food Restaurant	Field	Film Studio	Fish & Chips Shop	Fish Market	Fishing Store	3	51.24315	-0.53999	Guildford	Surrey
276	ME3	Golf Course	Yoga Studio	Flower Shop	Farmers Market	Fast Food Restaurant	Field	Film Studio	Fish & Chips Shop	Fish Market	Fishing Store	3	51.43322	0.54668	Hoo St Werburgh	Kent
320	PO22	Kids Store	Golf Course	Yoga Studio	Food	Fast Food Restaurant	Field	Film Studio	Fish & Chips Shop	Fish Market	Fishing Store	3	50.79835	-0.64756	Felpham	West Sussex
392	SE9	Golf Course	Yoga Studio	Flower Shop	Farmers Market	Fast Food Restaurant	Field	Film Studio	Fish & Chips Shop	Fish Market	Fishing Store	3	51.44465	0.05651	Eltham	London

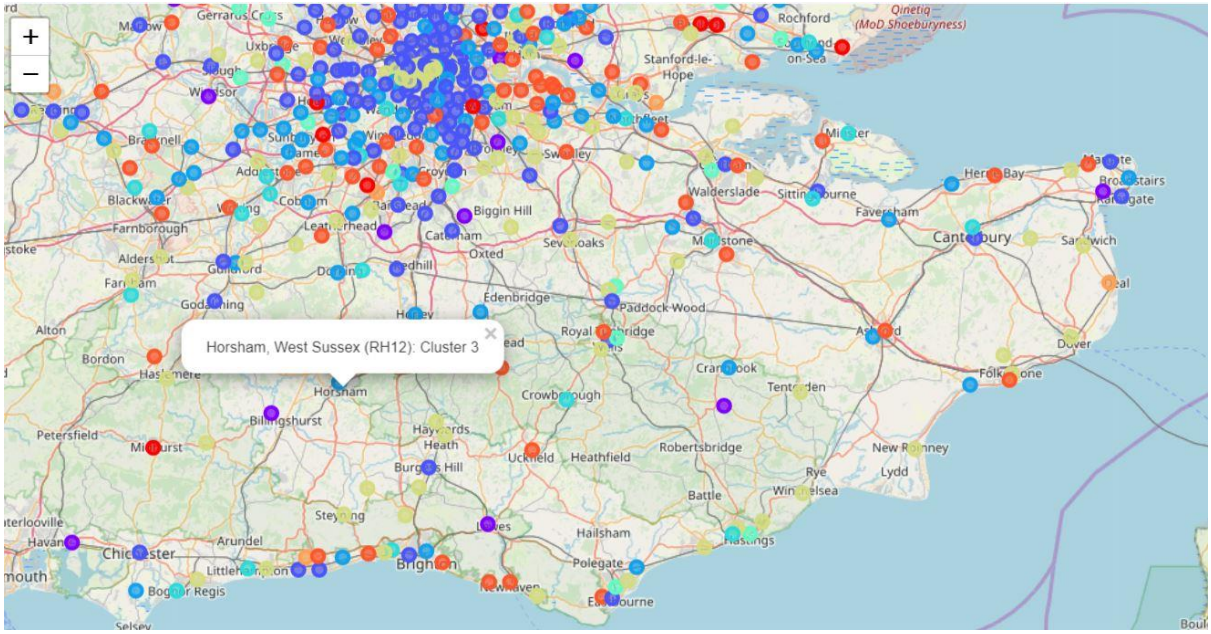
We suggest Ms Smith has a look at Clayton (West Sussex), Guildford (Surrey), Hoo St Werburgh (Kent) and Felpham (West Sussex).

Case Study 3

Postcodes split into 10 clusters or similar groups – relocation from Horsham to Kent

A couple wants to move from Horsham to Kent for work reasons. They quite like were they live and would like some town quite similar, so we split the data in 10 groups.

London and the Home Counties – 10 Groups



Immediately in the map we can see some Kent postcodes in the same cluster. We suggest they start researching the following 11 towns in the same cluster, such as Faversham, Broadstairs, Canterbury and so on.

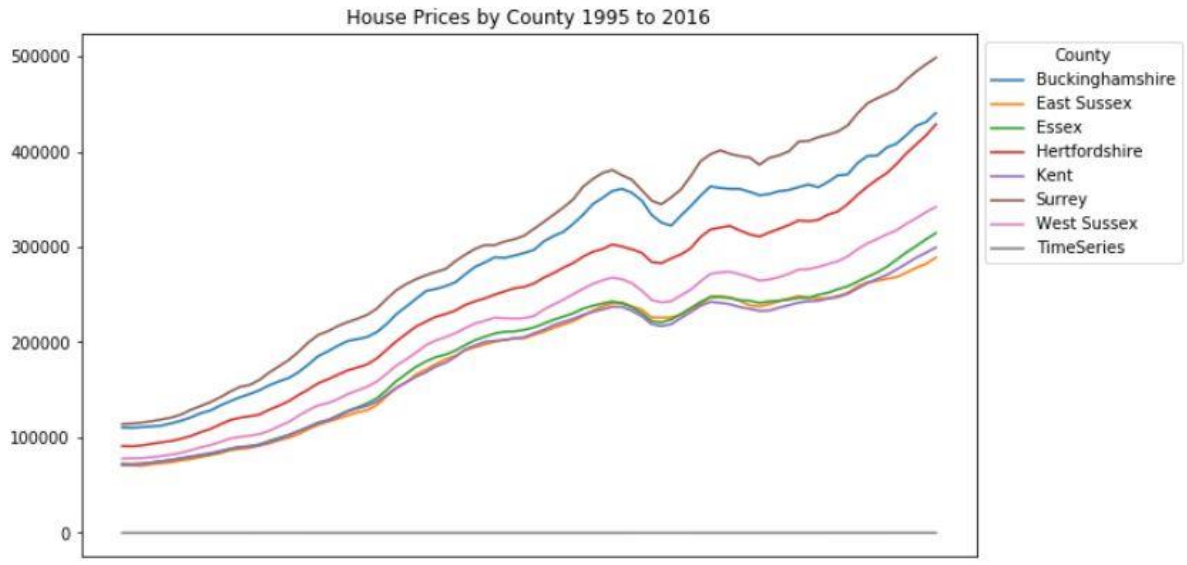
Below is the information for Horsham and the 11 cities on our search list:

	Postcode	Top 1st Venue	Top 2nd Venue	Top 3rd Venue	Top 4th Venue	Top 5th Venue	Top 6th Venue	Top 7th Venue	Top 8th Venue	Top 9th Venue	Top 10th Venue	Cluster label	latitude	longitude	town	region	region_id
336	RH12	Pub	Nature Preserve	American Restaurant	Grocery Store	Yoga Studio	Flea Market	Farmers Market	Fast Food Restaurant	Field	Film Studio	3	51.07579	-0.33254	Horsham	West Sussex	8

	Postcode	Top 1st Venue	Top 2nd Venue	Top 3rd Venue	Top 4th Venue	Top 5th Venue	Top 6th Venue	Top 7th Venue	Top 8th Venue	Top 9th Venue	Top 10th Venue	Cluster label	latitude	longitude	town	region
84	CT10	Café	Fast Food Restaurant	Pub	Convenience Store	Train Station	Turkish Restaurant	Flea Market	Farmers Market	Field	Film Studio	3	51.36208	1.43073	Broadstairs	Kent
94	CT21	Café	Pub	Supermarket	Seafood Restaurant	Indian Restaurant	Light Rail Station	Coffee Shop	Yoga Studio	Fishing Store	Farmers Market	3	51.07233	1.07795	Hythe	Kent
96	CT5	Flea Market	Pub	Indian Restaurant	Gym	Yoga Studio	Farm	Farmers Market	Fast Food Restaurant	Field	Film Studio	3	51.35318	1.03641	Canterbury	Kent
101	DA1	Pub	Pizza Place	Clothing Store	Nightclub	Performing Arts Venue	Coffee Shop	Yoga Studio	Fishing Store	Farmers Market	Fast Food Restaurant	3	51.44637	0.20915	Dartford	Kent
103	DA11	Pub	Grocery Store	Park	Fast Food Restaurant	Fish & Chips Shop	Yoga Studio	Fishing Store	Farm	Farmers Market	Field	3	51.43455	0.35392	Gravesend	Kent
105	DA13	Pub	Pizza Place	Train Station	Restaurant	Yoga Studio	Fishing Store	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	3	51.38235	0.35555	Meopham Station	Kent
117	DA9	Pub	Fast Food Restaurant	Train Station	Bus Stop	Yoga Studio	Flea Market	Farm	Farmers Market	Field	Film Studio	3	51.44771	0.27975	Stone	Kent
269	ME13	Restaurant	Deli / Bodega	Pub	Soccer Field	Yoga Studio	Farmers Market	Fast Food Restaurant	Field	Film Studio	Fish & Chips Shop	3	51.30275	0.89675	Faversham	Kent
273	ME19	Pub	Gastropub	Chinese Restaurant	Grocery Store	Coffee Shop	Flea Market	Farmers Market	Fast Food Restaurant	Field	Film Studio	3	51.29307	0.41123	West Malling	Kent
468	TN17	Pub	Grocery Store	Coffee Shop	Café	Yoga Studio	Flea Market	Farmers Market	Fast Food Restaurant	Field	Film Studio	3	51.09610	0.53810	Cranbrook	Kent
472	TN23	Pub	Gas Station	Fish & Chips Shop	Yoga Studio	Flower Shop	Farmers Market	Fast Food Restaurant	Field	Film Studio	Fish Market	3	51.13902	0.86075	Ashford	Kent

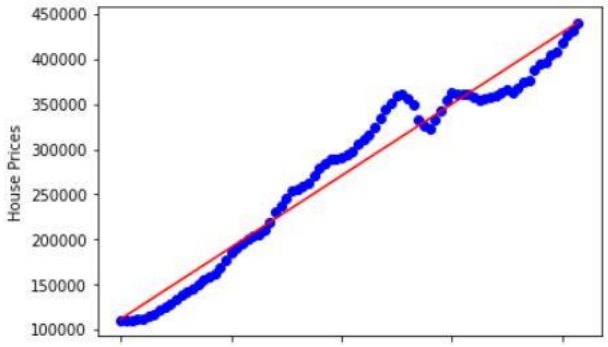
House Prices Results

In order to aid the above analysis, we looked at mean house prices by county. The plot below shows the data for 1995 to 2016 (quarterly). The dramatic rise in house prices in every county is evident with Surrey topping the chart and East Sussex at the bottom. This could give an indication of a good counties to search if price sensitive. For instance, in case study 3, moving from Horsham to Kent might mean is affordable for them as West Sussex and Kent price means are similar. Here is the chart:

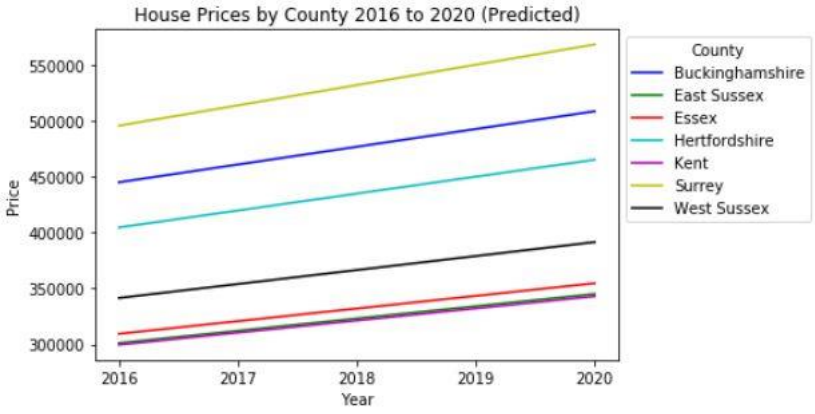


Almost for a bit of fun, we extended the house price data to 2020 to see what would happen in each county if house prices continued to grow at similar rates. We run a linear regression model and below is the regression plot vs. real data for Buckinghamshire (between 1995 and 2016):

Buckinghamshire Regression Plot



Finally, we use linear regression to predict house prices between 2016 and 2020:



If house prices continue to rise at this rate not many will be able to afford them!

5. Discussion

The results provided by this analysis can be much further extended. We find that by clustering in different ways we get different types of analysis.

This project code could be connected to a simple app where people type a postcode and other variables and get similar postcodes with a higher and lower of similarity (on a map).

The analysis is simply an initial analysis of the feel of areas/postcodes. Different starting points or modelling data could change the endpoints. Data is also not static and can change outcomes.

The extension of this analysis is immense, for instance, instead of top venues we could plot demographics data or political data and use it to target political campaigns.

The house price analysis was just meant to give a flavor of the counties' house prices for guidance. Further analysis on house prices by postcode could also be extended, by including other factors in the model and testing different models.

6. Conclusion

Overall, we can observe that London neighborhoods can be easily grouped.

This analysis fulfils its promise of finding a similar postcode for someone wishing to relocate in London and the Home counties.

The project can provide the baseline for an app or a different type of research.