Capstone Project

Capstone project for the IBM Data Science Professional Certificate

Introduction

London is the most populous city in the UK, as well as a technological, financial, and social hub. As such, there is always a huge number of people looking to move into London, whether for career or other reasons.

Due to the high demand, house prices and rents in London are higher than anywhere else in the country. There is fierce competition for living space, and an endless turnover of new arrivals and Londoners jostling for the best places to live.

Problem

London is a large city, with many distinct areas, each with their own character and available amenities. Some of these areas are more suitable or desirable for residents, but newcomers to London have only limited information on the various options, which can lead to them making suboptimal decisions; more information would give them a better basis to search for housing on. This project is focused on creating accessible visualisations to help these newcomers quickly understand the differences between the various boroughs of London.

The guiding question of this project is "which boroughs of London are the best to live in?" This is, of course, a subjective question, with the answer heavily depending on the priorities of the questioner, but any individual's answer can only be improved by access to more information in an easily-understandable way.

Intended audience

This project is intended to benefit newcomers to London - people from outside the UK or simply elsewhere in the country who are looking to move into the capital. Being able to see, in an immediate and visual manner, key information about different areas of London would enable these people to make more informed and data-driven decisions about where to look for places to rent/buy.

Data

The data used during this project came from several sources:

The GLA's London Borough Profiles

This dataset contains key indicators on a variety of topics for each London borough. The data includes information on the unemployment rate, election results, and home ownership, along with many other factors. Data from this source will form the majority of the information used in this project, but there are some missing or out-of-date columns (such as average income) in the dataset.

HMRC's dataset on the average income of taxpayers in London Boroughs

This dataset contains information about the average income of tax-payers in each London borough, collected between 1999 and 2017. This data can be combined with the Borough Profiles dataset in order to give more complete and up-to-date information on each borough.

GeoJSON for London Boroughs

In order to visualise information about London Boroughs, it's necessary to have detailed information on the boundaries between each one. This data was provided on GitHub by Emil Culic.

The Foursquare API

This API allows the retrieval of information about social and commercial venues, such as restaurants, in particular areas. This will be used to add more information to overall figures about each borough, and give people a clear idea of the social/dining scene in a given area.

Data collation & intentions

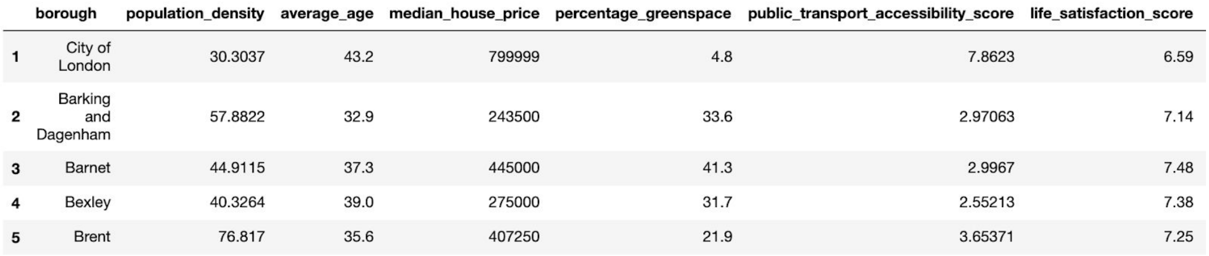
All of the data sources listed above needed significant cleaning and shaping to be used together for the purposes of this project. This processing is described in the methodology section.

Once the data had been cleaned and processed, it was collated together in order to give a high-level overview of the differences between London Boroughs. This overview was represented on various maps, so that people looking to rent in London can quickly see the differences between different areas.

Methodology

Data preparation

I began by importing and cleaning the borough profiles; this data was mostly complete (with a minimum of missing values), but did have some empty rows and a large amount of data that was not of interest. I subset the dataframe to focus on only those rows and columns which were relevant to the analysis; this step also handled the missing data. Following this, I renamed all columns to be in a standard format (lowercase, no spaces) and as clear as possible. A snapshot of the dataframe is below:



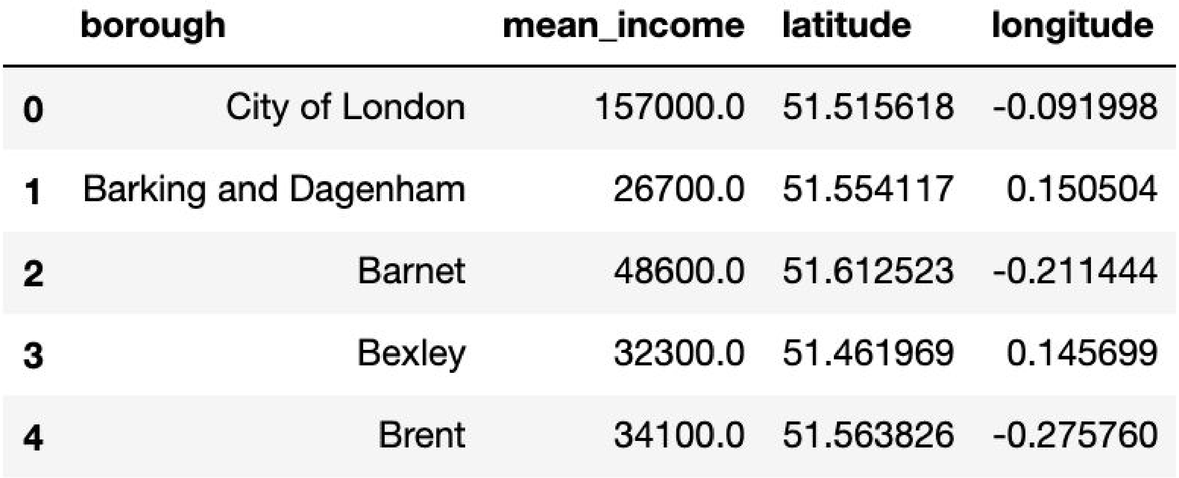
	borough	population_density	average_age	median_house_price	percentage_greenspace	public_transport_accessibility_score	life_satisfaction_score
1	City of London	30.3037	43.2	799999	4.8	7.8623	6.59
2	Barking and Dagenham	57.8822	32.9	243500	33.6	2.97063	7.14
3	Barnet	44.9115	37.3	445000	41.3	2.9967	7.48
4	Bexley	40.3264	39.0	275000	31.7	2.55213	7.38
5	Brent	76.817	35.6	407250	21.9	3.65371	7.25

The next stage was to combine the income information from the second data source with the borough dataframe. I imported the second dataset with the name income_df and performed the necessary cleaning/shaping: reducing it to only two columns, with names in the standard format. Both dataframes now contained a “borough” column.

Before I could merge the two datasets, I had to ensure that the “borough” column was consistent between dataframes, so it could be used as a key to merge on. To this end, I replaced all dashes in the income dataframe with spaces, to match the borough dataframe. After that, I merged the two dataframes together into one dataset.

The final stage of data preparation was to add latitude and longitude information to each borough, so I could later attach map markers in the appropriate places and pull data from FourSquare. I used Geopy to access the latitude and longitude for each location, attaching them as columns to the dataframe; the City of London was the only borough which broke the standard pattern and needed to have its latitude and longitude added in directly.

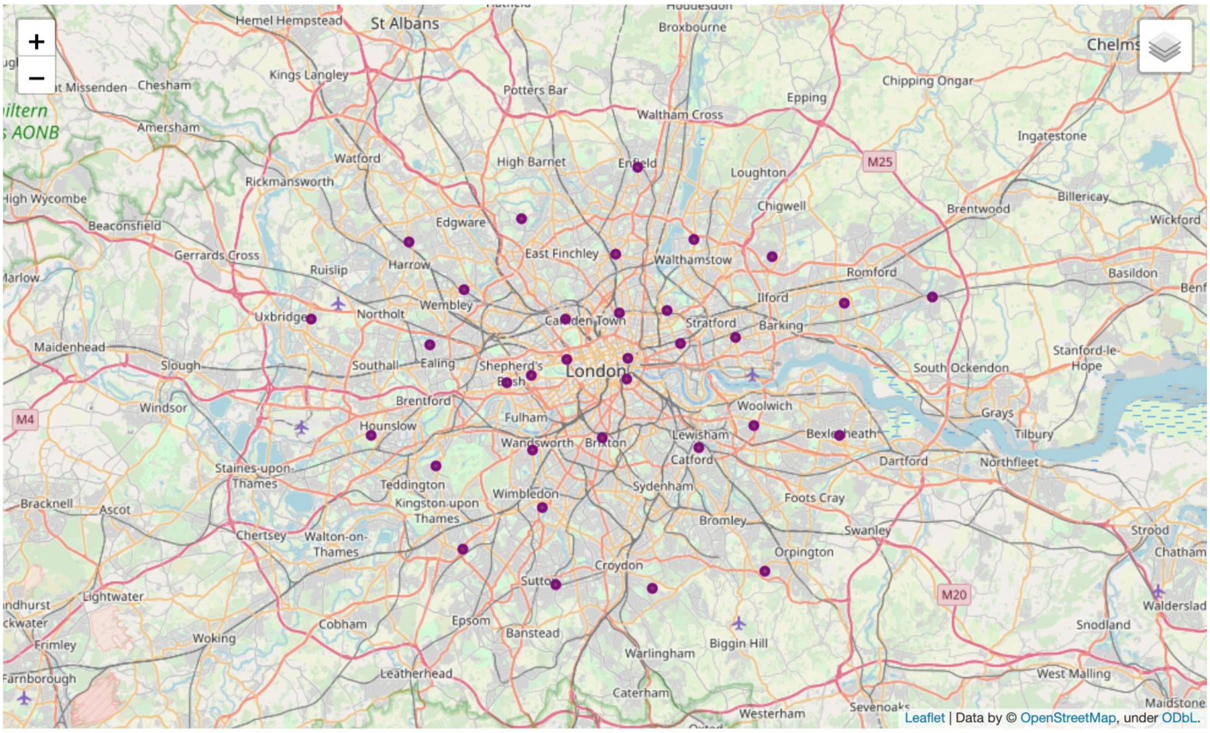
A subset of the final dataframe is shown below, demonstrating the successful merge and location additions.



	borough	mean_income	latitude	longitude
0	City of London	157000.0	51.515618	-0.091998
1	Barking and Dagenham	26700.0	51.554117	0.150504
2	Barnet	48600.0	51.612523	-0.211444
3	Bexley	32300.0	51.461969	0.145699
4	Brent	34100.0	51.563826	-0.275760

Visualising data

With the data cleaned and prepared, I was able to start building the visualisations. I began with a simple map of London with each borough marked upon it; clicking the borough marker produced a popup with the borough name. I created this map using the Folium library.



I then created two more visualisations, each using a subset of columns from the borough dataframe to show choropleths of key statistics for the boroughs.

Each visualisation contained - in addition to the borough markers - four optional layers of choropleth; these could be toggled using Folium's LayerControl object, and only appeared one at a time.

The first visualisation focused on practical demographic information; the choropleths used a yellow-to-red spectrum to represent values for four different variables:

- Average age of residents
- Mean income of residents
- Median house price
- Population density

The second visualisation focused on quality-of-life information; these choropleths used a blue-to-purple spectrum to represent values. The four variables chosen were as follows:

- Happiness score (out of 10)
- Life satisfaction score (out of 10)
- Percentage of green-space
- Transport accessibility score (out of 10)

Images of, and links to, these visualisations can be found in the "results" section of this report.

Collecting venue information

Based on the previous visualisation, I determined that the final stages of my analysis would focus on the borough of Haringey, as the choropleths suggested that this was both an affordable and pleasant place to live. Although Haringey is the only borough for which I produced the final visualisation, it would be trivial to extend the analysis to other boroughs.

I accessed the FourSquare API and pulled details of 100 venues (the maximum permitted) in the vicinity of Haringey. For each venue, I extracted the name, the latitude and longitude, and the “venue category” (as determined by FourSquare).

Once I had this data, I grouped the venues into seven distinct categories, combining those suggested by FourSquare into more holistic groups. The groups selected are below.

- Cafés
- Restaurants
- Retail
- Arts
- Bars/Pubs
- Nature
- Fitness

Finally, I created one more Folium map, this one focused on Haringey itself, rather than London as a whole. This map contained no borough markers or choropleth layers. Instead, each category of venue was represented by a different coloured set of markers. These could be independently toggled on and off, allowing users both to see the general distribution of venues in the area and to narrow their exploration to the type they were most interested in.

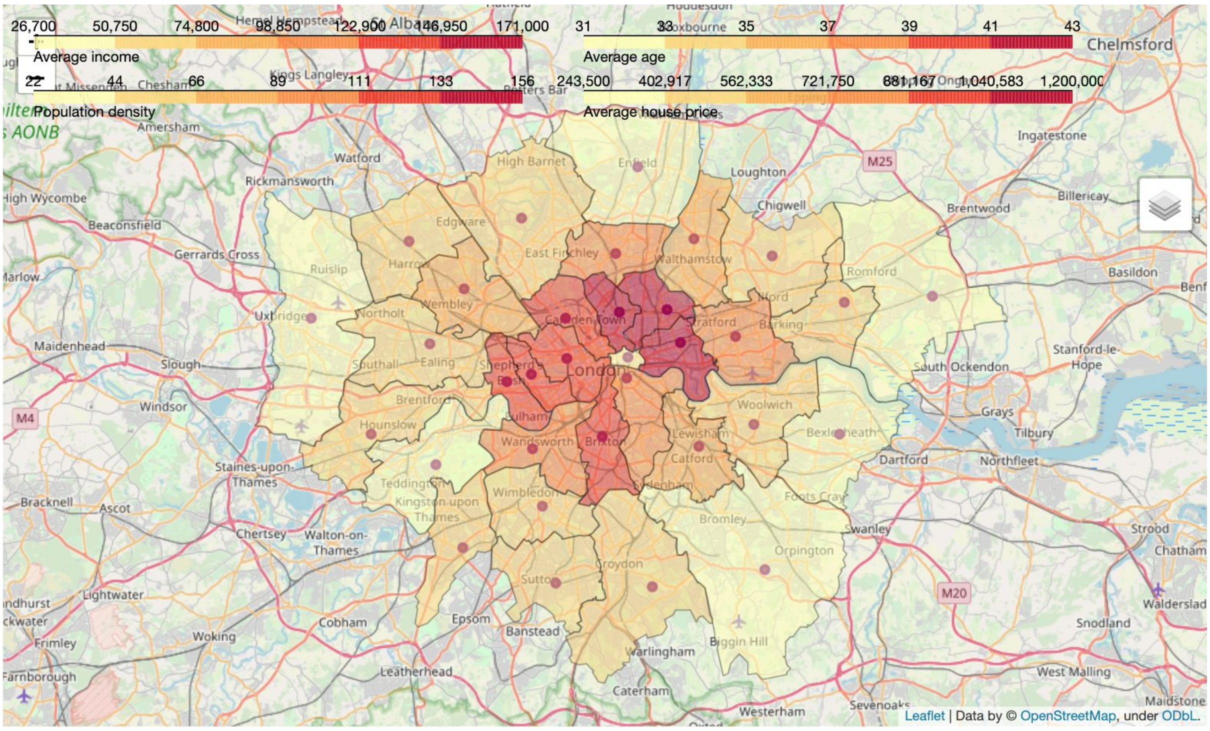
As before, a link to the final visualisation can be found in the “results” section.

Results

The final visualisations are, by design, interactive, and so cannot be shown in a standard report. However, static images of, and links to, those visualisations (hosted online) can be found below. Each separate visualisation provides key information about London boroughs in an accessible and immediately understandable way.

Choropleth of practical demographic information

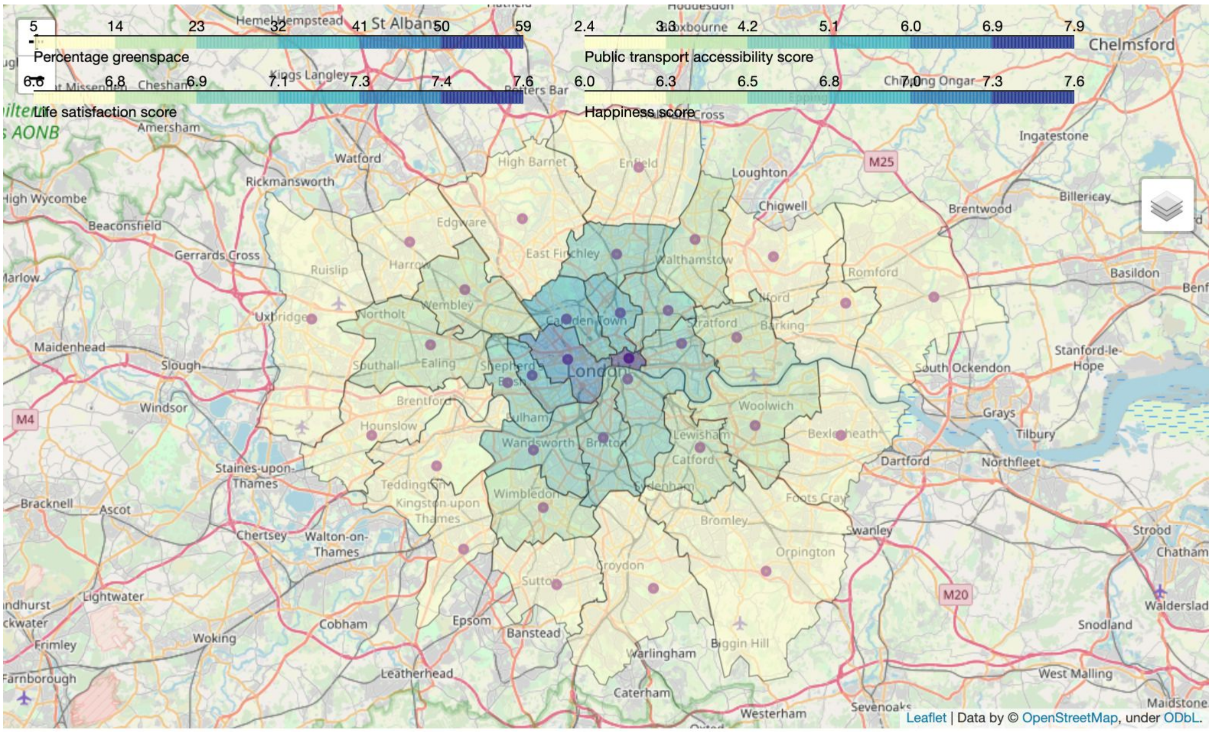
average age, income, house price, and population density



[Link to live visualisation]()

Choropleth of quality-of-life information

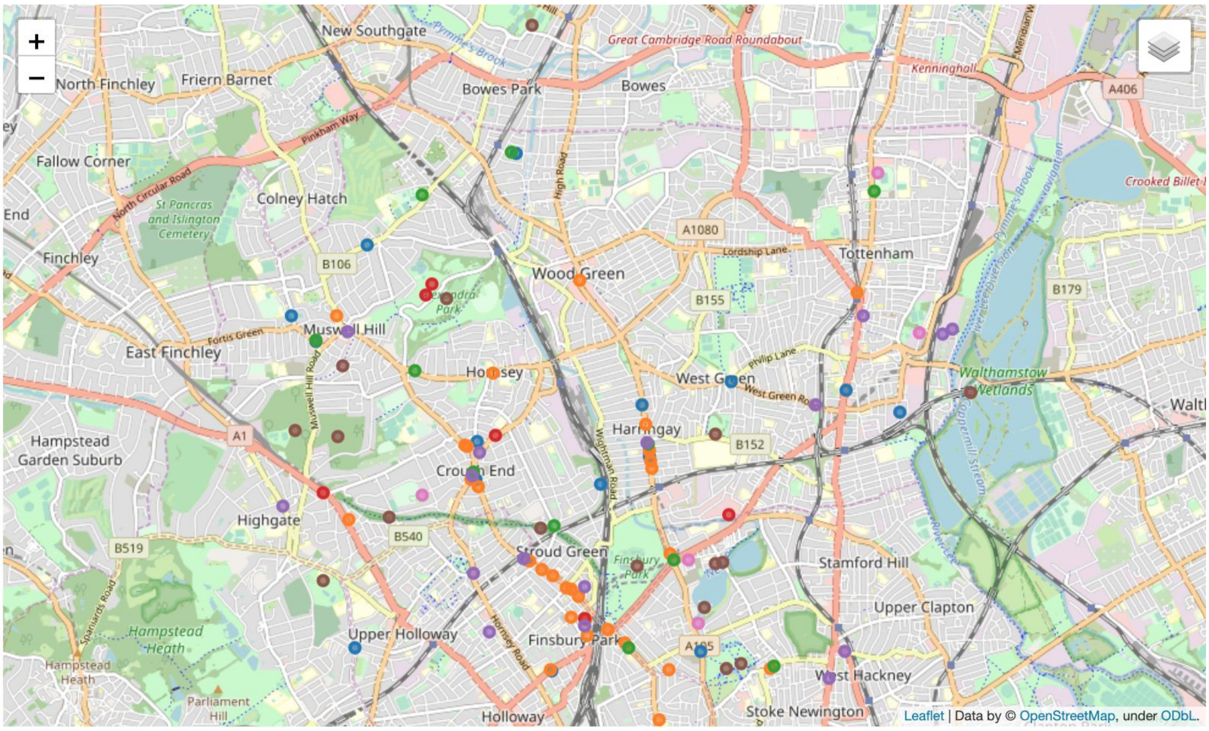
happiness, life satisfaction, percentage green-space, access to transport



[Link to live visualisation]()

Map of the Haringey borough, with a sample of venues in the area

different venue types are independently toggleable



Link to live visualisation

Discussion

I found these visualisations to be very informative. While to some extent, they confirmed what I already knew - property gets cheaper the further from the center you are - being able to quantify those insights is valuable. In addition, the visualisations enable a granular view of the city, broken down into boroughs rather than broad compass directions, which is useful for people deciding between areas.

Based on the first two visualisations, several inferences can be made:

1. The City of London is probably not where most people would like to live; it is extremely expensive, and - while it does have exceptional transport links - people in this borough report some of the lowest happiness and life satisfaction scores across the whole city.
2. Hackney and Tower Hamlets contain young professionals: people living in close proximity in the cheapest areas close to the center of the city.
3. Kensington & Chelsea and the City of London are the richest and also the most expensive areas to live in.
4. Although transport is harder to find on the outskirts of the city, the people living there are more likely to be happy & satisfied.
5. Money doesn't buy happiness, unless you are rich enough to live in Kensington and Chelsea, which reports not only high incomes and prices, but also high happiness scores.

A lot more information can be extracted from these visualisations, as they provide information for each borough on eight different variables; it's quick and easy to make comparisons not just between boroughs but between the positives and negatives of each separate borough.

The Haringey visualisation gives even more granular information. Based on the venues shown, it is clear that Haringey is a good place to live for those interested in dining and nightlife, but perhaps not as ideal for those interested in arts and culture. As above, this analysis could be extended to every borough, pulling out the key elements of that area and giving users an understanding of the sort of diversions available there.

Conclusions

The initial aim of this project was to produce visualisations that would help newcomers to London understand the different areas of the city, and guide them in their housing search; I believe that this aim has been accomplished. The visualisations are live and accessible online, and they do provide that high-level overview of the different areas and - to some extent - what it would be like to live there.

In the future, I’d like to explore a greater range of variables, looking at factors such as crime rates and the quality of schools. I’d also like to access other APIs than the FourSquare one, adding information about transport and a greater range of retail options to the maps. More information being accessible to users would only further support this project’s aims and assist people wanting to understand the different areas of London.