Housing and Population in London

Data Science Capstone Project

November 2020

By: Ludwig Forsberg

Introduction

I lived and worked in London for about 5 years, from 2013 to 2018 just after having graduated university. During that time I read a lot about the lack of housing, and moreover, affordable housing in London and in Greater London. The politicians of the day on both sides were going back and forth on when, if, how and where concerning the issue.

Given Greater London's historic wide appeal for immigrants, with a population of 8.9 million as of 2018, this issue is not likely to go away soon, even after Brexit I believe. Local politicians will, in my opinion, therefore need to address the issue sooner rather than later.

Problem

Therefore, the following project will be directed to the office of the Mayor of London and look at the supply of housing and population by borough or Local Authority (LA), as it is called, in Greater London and try to determine if there is a shortage of housing given the local population. If so, around which underground station would there be an opportunity to build more housing if money was not an object given access to leisure activities, offices, cultural institutions etc.

Data

In order to complete the above analysis I will require different sets of data sources. As I have identified them, I will require the following:

1. London Underground station coordinates
2. LA (Local Authority) coordinates
3. Housing supply statistics from Greater London by LA
4. Population statistics of Greater London by LA
5. Foursquare API venue data for London by LA

London Underground Station Coordinates

The coordinates I managed to find online in a downloadable csv format, here https://www.doogal.co.uk/london_stations.php. The file provides station name, coordinates as well as zone of the station which could be useful for further analysis.

LA Coordinates

The LA coordninates were found here:
https://geoportal.statistics.gov.uk/datasets/ae90afc385c04d869bc8cf8890bd1bcd_1
This was the latest I could find, from 2017, however, for the purposes of this paper the fact that it is 3 years old is of little consequence as the boundaries tend to seldom change.

Housing Supply Statistics by LA

The data on housing supply was found on the government's website, gov.uk on https://www.gov.uk/government/statistical-data-sets/live-tables-on-dwelling-stock-including-vacants.
This information will be used in order to see how much supply there is compared to the population. Some assumptions will be made here as to the nature of the supply; that each dwelling can house min 2 people and they all fulfill the minimum standard of living requirements. This data is for 2019, and therefore, for the purpose of this paper, is not too old for usage.

Population Statistics by LA

This data is from the government's official statistics office, the ONS, Office for National Statistics, for 2015, https://data.london.gov.uk/dataset/office-national-statistics-ons-population-estimates-borough. It is the latest data I could find that included boroughs and LAs, with a total population for London of 8.7 million. I will use the percentage split of the boroughs in the data I found and use those percentages to determine the current population by borough as of 2019, 8.9 million.

Foursquare API Venues

The venue data has been extracted using the Foursquare API. This data contains venue recommendations for all boroughs in Greater London and is used to study the popular venues of different boroughs.

Data Cleaning

The sources of data needed minor cleaning and wrangling. A column or two was dropped in the Underground csv to make it easier to work with.
The biggest was combining the housing and population statistics with the LA coordinates, which I did in an excel before as the population and housing supply statistics contained multiple tables on each tab/sheet.

Methodology

The target variable that I was after, namely, the difference in the percentage split of housing supply and population by borough. Since this was a matter of percentages I decided to add 3 columns, one for percentage of housing supply by borough, one for percentage share of population by borough and the final column with the difference between the two.

At this point I was wondering what exploratory analysis could be undertaken to help me identify which boroughs to continue with. As I wasn't looking for a relationship between the two, since my problem or question wasn't about the relationship between housing supply and population, I decided to skip correlation analysis altogether.

Rather, since my problem was in which boroughs is there a greater share of population than the supply of housing, I decided to choose the top boroughs that had the biggest difference between share of population and share of housing supply, i.e. had a greater share of the population than they had of the housing supply:

	Borough	Longitude	Latitude	Housing Supply	Population	% Housing	% Population	% Diff
0	Redbridge	0.070085	51.58589	104688	296793	2.91	3.42	0.18
1	Newham	0.027369	51.53132	116979	332817	3.26	3.84	0.18
2	Brent	-0.275680	51.56441	120448	324012	3.35	3.74	0.12
3	Hillingdon	-0.441820	51.53663	110734	297735	3.08	3.43	0.11
4	Harrow	-0.336030	51.59467	91909	247130	2.56	2.85	0.11
5	Hounslow	-0.378440	51.46243	101838	268770	2.83	3.10	0.10
6	Barking and Dagenham	0.129506	51.54552	75829	201979	2.11	2.33	0.10

Figure 1. Table showing in the red square the shares of housing and population and the difference between the two

After the boroughs with the greatest difference, and therefore the biggest opportunity existed for an increase in housing supply, had been established I proceeded to add the Foursquare API venue data to the selected boroughs.

Once the venue data had been appended to the data frame, I proceeded with the k-Means clustering to cluster the various neighborhoods. After running some tests of the number of clusters I settled on 5 clusters.
Finally, I added the underground stations to the cluster map to see if there were any regions far from underground stations.

Results

The map shows the boroughs that were selected were all outside central London, and either to the west or east of the city:

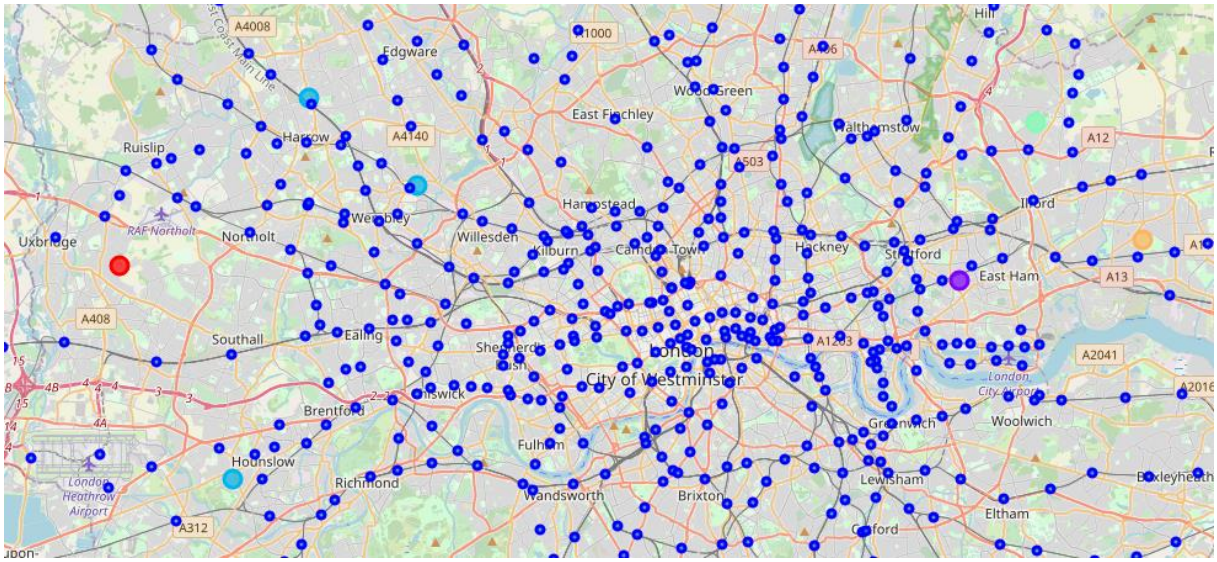

Figure 2. Underground stations in blue, and clusters in bigger multi-colored circles

Reviewing the clusters and top venues, only cluster 2, with 3 boroughs included, did not have any form of transportation in top common venues, however, comparing with the map of with Underground stations as well it is clear to see that there are transportation possibilities for all of the 7 boroughs. However, there seems to be fewest for cluster 2.

The only borough that stands out from the map in figure 2 is the red circle on the left-hand side, which is Hillingdon. It has the smallest concentration of underground stations around itself.

Furthermore, the results showed that the 7 boroughs all have a varied and vibrant selection of different venues among their most common venues. Based on those results, there isn't one cluster one could disregard for lack of variety;

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Sporting Goods Shop	Soccer Field	Playground	Construction & Landscaping	Transportation Service	Grocery Store	Asian Restaurant	Auto Garage	Boutique	Burger Joint
1	Pub	Boutique	Transportation Service	Hotel	Asian Restaurant	Auto Garage	Burger Joint	Bus Stop	Business Service	Café
2	Coffee Shop	Hotel	Electronics Store	Park	Pedestrian Plaza	Café	Indian Restaurant	Sandwich Place	Supermarket	Burger Joint
2	Indian Restaurant	Platform	Bus Stop	Coffee Shop	Grocery Store	Thai Restaurant	Afghan Restaurant	Supermarket	Sandwich Place	Pub
2	Indian Restaurant	Pizza Place	Hotel	Asian Restaurant	Auto Garage	Burger Joint	Restaurant	Business Service	Diner	Grocery Store
3	Transportation Service	Pharmacy	Hotel	Asian Restaurant	Auto Garage	Boutique	Burger Joint	Bus Stop	Business Service	Café
4	Park	Lake	Transportation Service	Hotel	Asian Restaurant	Auto Garage	Boutique	Burger Joint	Bus Stop	Business Service

Figure 3. Each cluster and borough show a clear variation of venues present

Discussion

Given the above mentioned results, the fact that all boroughs and their clusters have a wide variety of venues and have all relatively close to underground stations, except 1, Hillingdon, I would say that based on the above all boroughs except Hillingdon show the greatest opportunity for more housing, as they

are have the greatest difference between share of population and share of housing supply, they have a great variety of venues in their vicinity and they are all close to transportation and the underground.

Conclusion

To conclude, in order to get the highest return on investment based on the variables chosen and analyzed in this paper, I would recommend increasing the supply of housing in the boroughs; Newham, Brent, Harrow, Hounslow, Redbridge, and Barking and Dagenham.

While a high-level recommendation has been able to have been reached from the research and analysis, there are a few other attributes that could be added to further help the Office of the Mayor of London.

Data such as income, education and age can be added to make sure housing is built where it is most needed. For instance, if there is a lack of housing in an area with low relative income and education then council housing might be the best move forward instead of high rise expensive apartments.

Other limitations of this study is that the data collected were from different years, although it was what I could find online that looked reliable.

References:

https://www.doogal.co.uk/london_stations.php

https://geoportal.statistics.gov.uk/datasets/ae90afc385c04d869bc8cf8890bd1bcd_1

https://www.gov.uk/government/statistical-data-sets/live-tables-on-dwelling-stock-including-vacants

https://data.london.gov.uk/dataset/office-national-statistics-ons-population-estimates-borough