Predicting Gentrification in London

By Tim de Zwart, April 2020

Link of notebook: https://github.com/tdezwart/IBM-applied-data-science-capstone/blob/master/ads-london-gentrification.ipynb

1. Introduction: Business Problem

Gentrification is a common and controversial topic in politics and urban planning. Although it often increases the value of a neighbourhood, it usually results in a great demographic change, shifting the neighbourhoods composition and average household income, and with this displacing the current population of the neighbourhood (https://en.wikipedia.org/wiki/Gentrification).

Measuring gentrification is a difficult task, however income change and raise in property value over time in a neighbourhood have shown to be good indicators. Furthermore, we all know the stereotypical images of gentrified neighbourhoods: trendy coffee places, boutique shops, sustainable stores, etc. The goal of this report is to analyse the type of venues that can generally be found in gentrified areas, and then show which areas with a low real estate price are similar to these neighbourhoods, in order to predict which parts of the cities are prone to gentrification soon.

The city chosen for this analysis was London. This choice was made because it's a vibrant city with a diverse population (and therefore many different venues) that has been known to suffer from gentrification in the past. The data for London is readily available which makes it a perfect study subject for this project.

The stakeholders for this project are manifold: current inhabitants of areas that want to know what will happen with the value of their property, urban planners that want to predict what will happen to the city, real estate companies wanting to know where to make their next investment, etc.

2. Data

Based on the problem definition, the factors that will help us determine which neighbourhood is prone to gentrification, are:

- The type of venues in a certain area. Areas with many venues that are similar as the venues in areas that are already gentrified are probably prone to gentrify soon.
- The real estate prices in London per area. Areas that currently have a low real estate price are more prone to gentrification than areas that already have a high real estate price.

The data sources that are needed to perform this analysis are:

- A list of areas and boroughs of London, which is found here: https://en.wikipedia.org/wiki/List_of_areas_of_London
- Geographical coordinates (latitude and longitude) of the listed areas in order to plot on maps and find nearby venues. These coordinates are retrieved using Python's geopy library.

- A list of the most gentrified areas over the last years in London in order to compare and cluster the areas, which can be found here: https://londonist.com/london/news/these-are-london-s-most-gentrified-boroughs-but-are-they-really. There boroughs were plotted on a map using a GEOJSON file from https://skgrange.github.io/www/data/london_boroughs.json.
- The number of venues as well as the types of these venues for each area are retrieved using the Foursquare API
- The real estate prices per borough in the UK, which can be found here: http://webarchive.nationalarchives.gov.uk/20170726163612///visual.ons.gov.uk/wp-content/uploads/2017/10/map.csv

Area data

The list of areas and boroughs is scraped of the Wikipedia page, cleaned up (for more information on how this was done, check the notebook), and filtered. For this analysis, only areas within the "Post town" of London itself were used, in order to reduce the area of analysis a bit. Python's geopy library was then used in order to find the coordinates of the corresponding areas. This leaves us with the following Pandas dataframe.

	Area	Borough	Latitude	Longitude
0	Abbey Wood	Bexley	51.491060	0.121216
1	Acton	Ealing	51.508140	-0.273261
2	Aldgate	City	51.514248	-0.075719
3	Aldwych	Westminster	51.513103	-0.114920
4	Anerley	Bromley	51.407599	-0.061939
5	Angel	Islington	51.531946	-0.106106
6	Archway	Islington	51.565437	-0.134998
7	Arnos Grove	Enfield	51.616402	-0.133287
8	Balham	Wandsworth	51.445645	-0.150364
9	Bankside	Southwark	51.508302	-0.096297

Figure 1: Snippet of used dataframe with areas, boroughs and geographical coordinates of London

Most gentrified boroughs

For this project we will analyse the most gentrified boroughs according to londonist.com. The list of gentrified boroughs is quite large, but to be safe we will use the five most gentrified boroughs in the London post town. These boroughs are: Kensington and Chelsea, Richmond upon Thames, Westminster, Hammersmith and Fulham and Merton.

Venue data using Foursquare API

The Foursquare API was used in order to retrieve venue data for venues located in all the areas. For every area a call was made exploring the first 100 venues in a radius of 250 m (however, no areas returned more than 81 venues). The category and locations of all venues are stored in a dataframe.

Real estate prices of London

The final piece of data that was used are the London real estate prices. A list of 2016 UK real estate prices in pounds per square meter was taken from the UK National Archives. This data set includes all boroughs in the UK.

	Borough	Price per m2
0	Hartlepool	987
1	Middlesbrough	1120
2	Redcar and Cleveland	1182
3	Stockton-on-Tees	1254
4	Darlington	1260
5	Halton	1339
6	Warrington	1750
7	Blackburn with Darwen	1053
8	Blackpool	1107
9	Kingston upon Hull, City of	1127

Figure 2: Real estate prices in pounds per square meter in the UK (2016)

3. Methodology

Now that the data is collected and prepared, some exploratory data analysis is performed. First of all, the area data is used to create a map using Python's folium library. This gives a good overview of the areas that we will be analysing.

Maps of areas and boroughs

Figure 3 shows a map of the geographic location of all the areas that are in our dataframe.

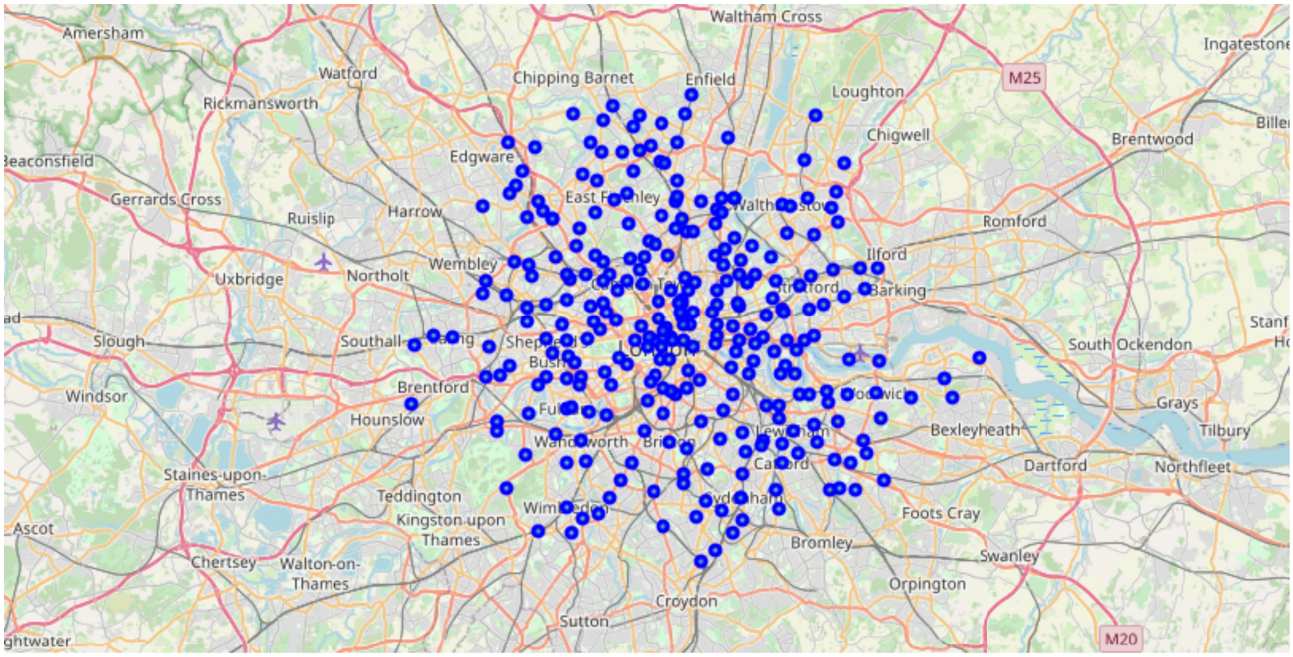

Figure 3: Overview of geographic location in all areas

As can be recalled, our benchmark boroughs to check if an area is similar to a gentrified area are Kensington and Chelsea, Richmond upon Thames, Westminster, Hammersmith and Fulham and Merton. In order to illustrate where these boroughs are located, the GEOJSON file of London boroughs was split up in a JSON file with the five "gentrified" boroughs and a JSON file with the rest of the boroughs. Folium was used to show these boroughs. Figure 4 shows our "gentrified" boroughs in red, and boroughs that are not (yet) gentrified in green.

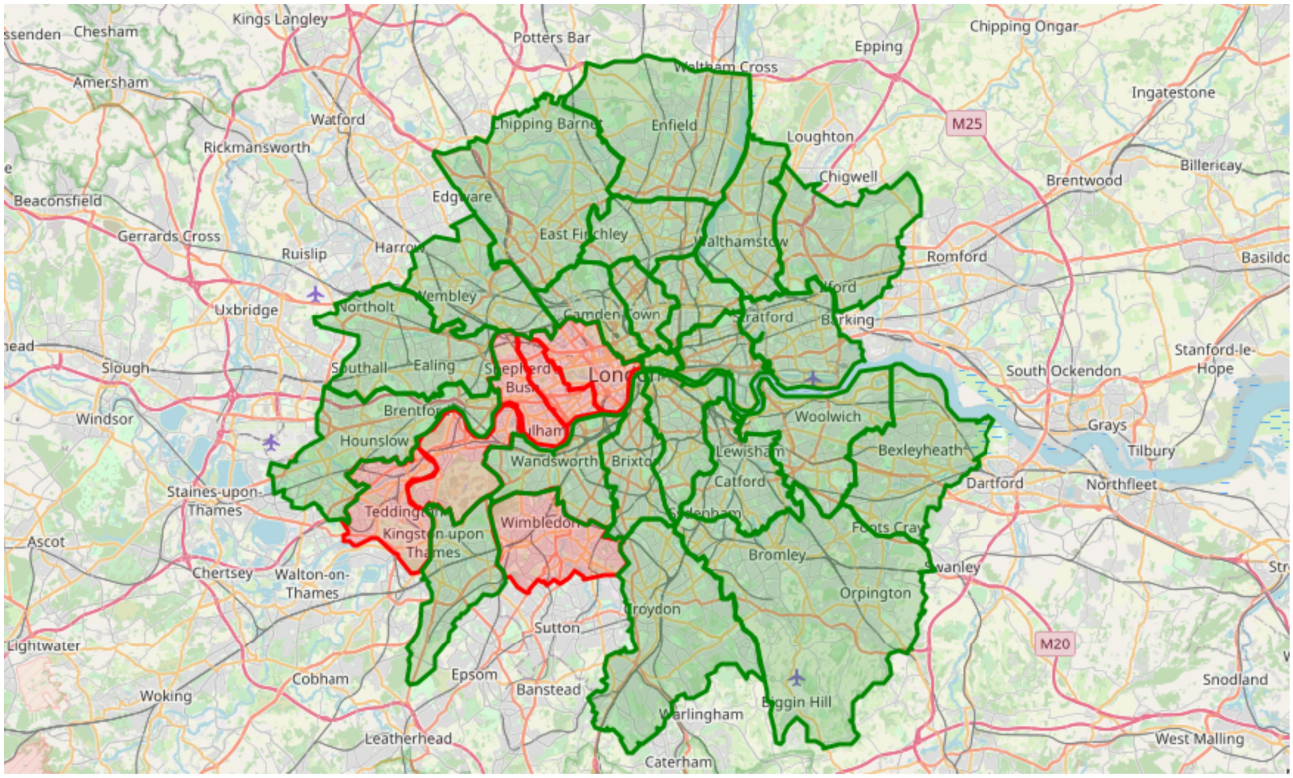

Figure 4: Map of London with most gentrified areas over past years in red

Foursquare API data

Using the Foursquare API, venues in the surroundings of all our areas are found. These are then stored in a dataframe, along with the corresponding area and borough of the venue, as shown in Figure 5. This way it is possible to filter the dataset for only the gentrified or ungentrified boroughs.

	Area	Borough	Latitude	Longitude	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbey Wood	Bexley	51.49106	0.121216	51.49106	0.121216	Sainsbury's	51.492824	0.120724	Supermarket
1	Abbey Wood	Bexley	51.49106	0.121216	51.49106	0.121216	Abbey Wood Railway Station (ABW)	51.490825	0.123432	Train Station
2	Abbey Wood	Bexley	51.49106	0.121216	51.49106	0.121216	The Abbey Arms	51.490693	0.121182	Pub
3	Abbey Wood	Bexley	51.49106	0.121216	51.49106	0.121216	Platform 1	51.491023	0.119491	Platform
4	Abbey Wood	Bexley	51.49106	0.121216	51.49106	0.121216	Costcutter	51.491287	0.120938	Convenience Store
5	Abbey Wood	Bexley	51.49106	0.121216	51.49106	0.121216	Bean @ Work	51.491172	0.120649	Coffee Shop
6	Acton	Ealing	51.50814	-0.273261	51.50814	-0.273261	London Star Hotel	51.509624	-0.272456	Hotel
7	Acton	Ealing	51.50814	-0.273261	51.50814	-0.273261	Dragonfly Brewery at George & Dragon	51.507378	-0.271702	Brewery
8	Acton	Ealing	51.50814	-0.273261	51.50814	-0.273261	The Aeronaut	51.508376	-0.275216	Pub
9	Acton	Ealing	51.50814	-0.273261	51.50814	-0.273261	Amigo's Peri Peri	51.508396	-0.274561	Fast Food Restaurant

Figure 5: Venue data for all venues in analysed areas

We use this data to see which type of venues are mostly present in our five gentrified boroughs, to get an idea about what a typical gentrified borough might look like. As can be seen in Figure 6, venue categories that occur very often in gentrified areas are, as we might have expected, coffee shops, pubs, cafés, but also venues like Italian restaurants, theatres and parks.

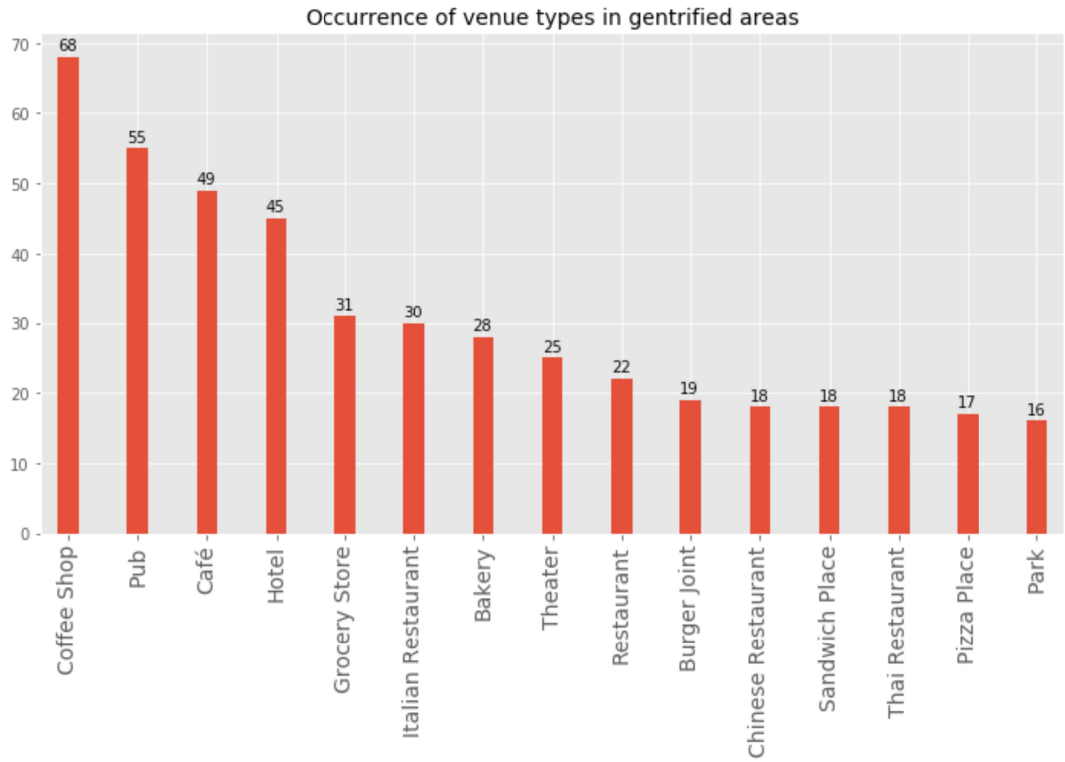

Figure 6: Overview of amount of venues per category in gentrified areas

Now that we have an idea of what a typical gentrified area looks like in terms of venues, an overview is made that shows the frequency of venue categories for each neighbourhood. This is done by one-hot-encoding the venue list and then taking the mean per area. This gives an overview of how similar areas are to each other.

	Area	African Restaurant	American Restaurant	Antique Shop	Aquarium	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	...	Video Game Store	Vietnamese Restaurant	Volleyball Court	Warehouse Store
0	Abbey Wood	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.0	0.000000	0.0	0.0
1	Acton	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.0	0.000000	0.0	0.0
2	Aldgate	0.0	0.027778	0.0	0.0	0.0	0.0	0.027778	0.000000	0.000000	...	0.0	0.027778	0.0	0.0
3	Aldwych	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.031250	0.031250	...	0.0	0.000000	0.0	0.0
4	Anerley	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.0	0.000000	0.0	0.0
5	Angel	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.023256	0.000000	...	0.0	0.000000	0.0	0.0
6	Archway	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.0	0.000000	0.0	0.0
7	Arnos Grove	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.0	0.000000	0.0	0.0
8	Balham	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.0	0.000000	0.0	0.0
9	Bankside	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000	0.041667	0.083333	...	0.0	0.000000	0.0	0.0

Figure 7: Frequency of given venue category in area

Clustering the areas

The dataframe given in Figure 7 is used to cluster areas together that are similar. Sci-kits K-Means method was used, an unsupervised machine learning algorithm that clusters areas together based on their similarity. In this case, we hope that the algorithm will recognize that areas with certain venue types are similar to areas with the same types of venues. For this project, it was chosen to look for 10 clusters in order to create multiple clusters with a significant size. Looking for either more or less clusters might give more accuracy. An overview of the different clusters is created with a folium map.

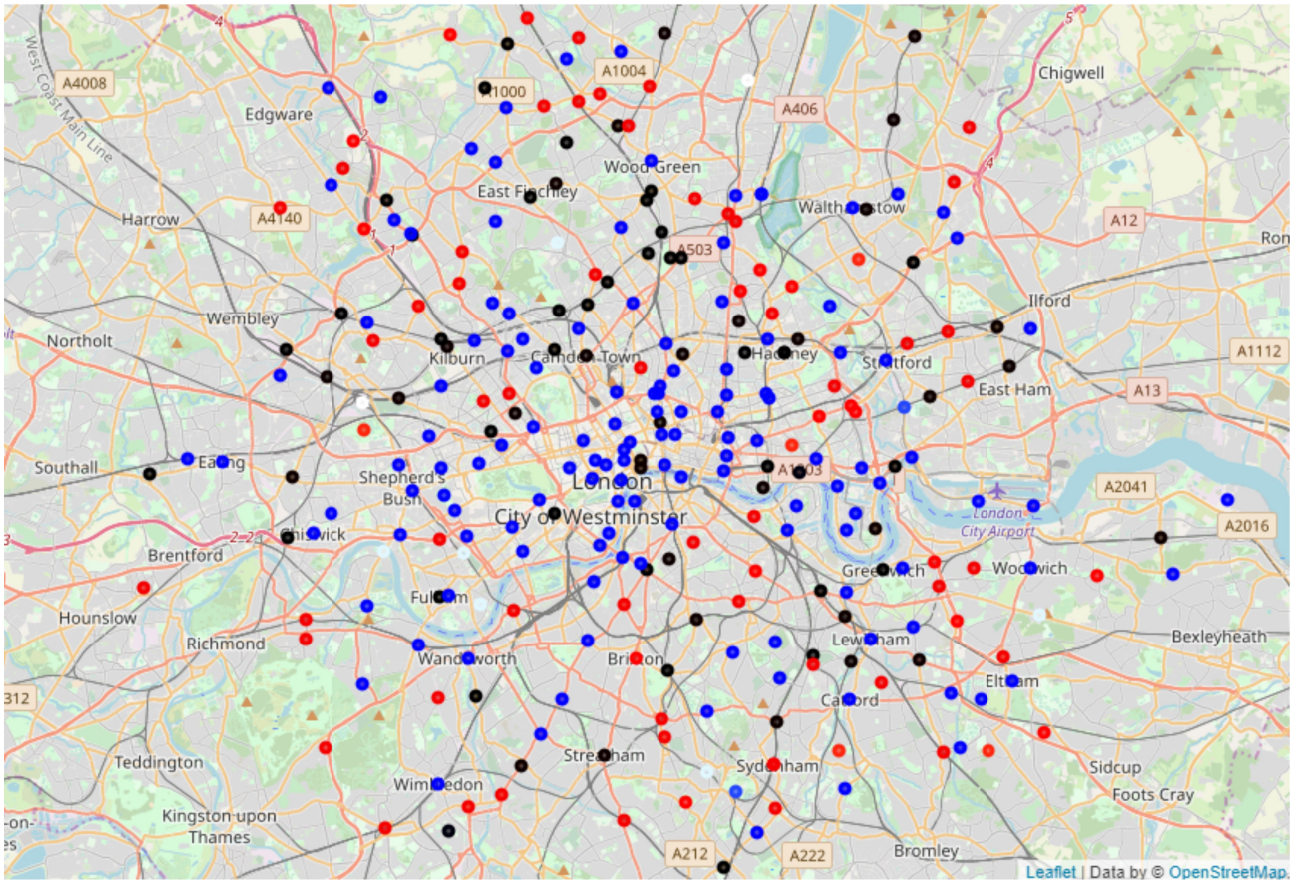

Figure 8: Areas in London clustered based on their similarity in venue types

As can be seen from Figure 8 most areas fall into either the blue, red or black cluster, with some areas falling into the (less visible) white and light green clusters as well. What interests us are which is the most common area in the gentrified areas. Therefore, a bar chart is made of only areas in the gentrified boroughs, showing which cluster label is prevalent.

Figure 9: Cluster labels of areas in gentrified areas

As can be seen from Figure 9, most of the areas in the gentrified boroughs belong to the cluster with label 4, which happens to be the blue cluster. This means that the areas that are most similar to gentrified areas are in cluster 4. Therefore, we can assume that areas with cluster label 4 that are currently not in gentrified areas but have a low real estate price are prone to gentrification.

London real estate prices

The last piece of exploratory data analysis, is analysing the real estate prices of our boroughs that are currently not gentrified (the green areas in Figure 4). As you might recall, we are mostly interested in areas that currently have a low real estate price. The London real estate prices per borough are combined with the created JSON for the not gentrified areas to make a choropleth map, using folium.

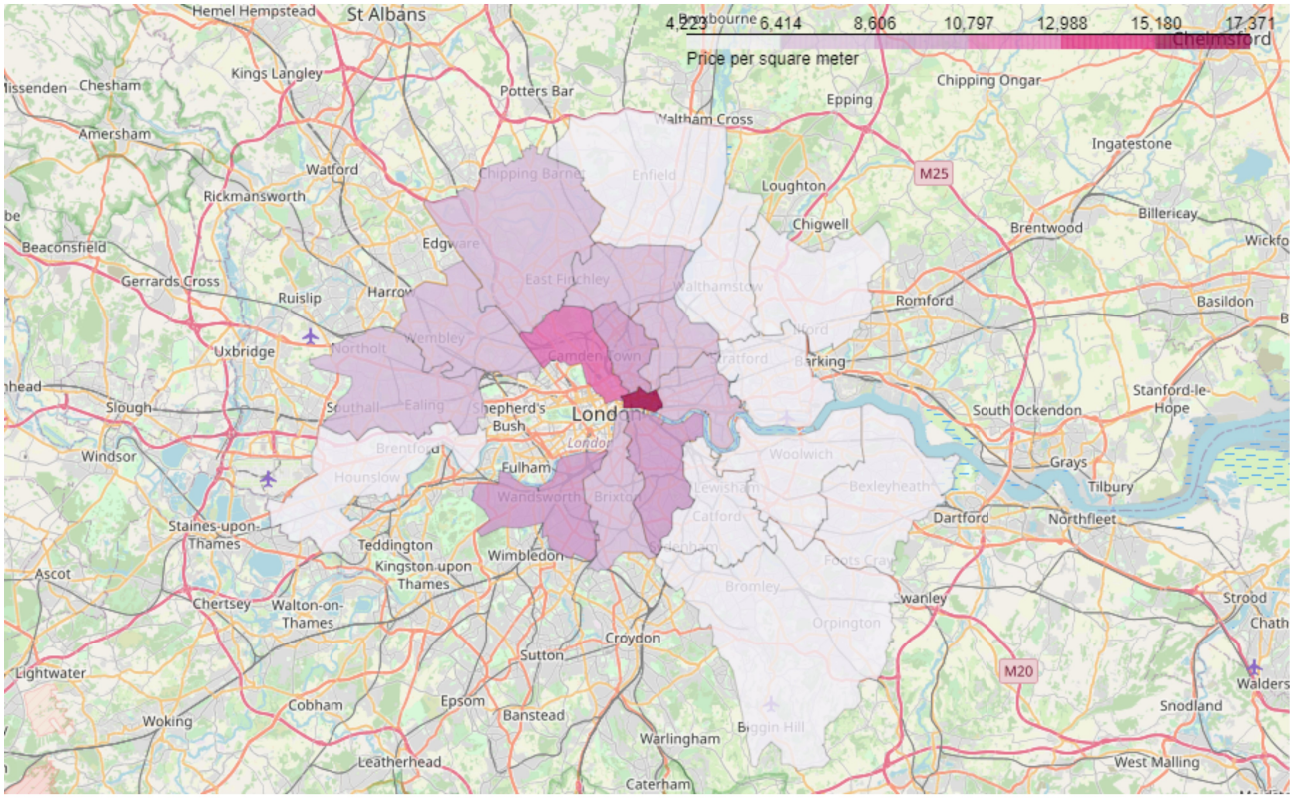

Figure 10: Choropleth map of London real estate prices in pounds per square meter for boroughs of interest

As expected, the highest real estate prices are found in the centre (and more specifically The City). However, in the outer areas there are still some boroughs with relatively low real estate prices.

4. Results

In order to predict which areas in London are prone to gentrify, we can combine the choropleth map of real estate prices and the areas that belong in cluster 4 (these are the areas that are most similar to gentrified areas in terms of venues). Recall the main assumption:

An area is prone to gentrification if it is similar to areas that are already gentrified and the real estate prices are still low.

Let's look at a map of London showing the real estate prices and the areas that are similar to gentrified areas.

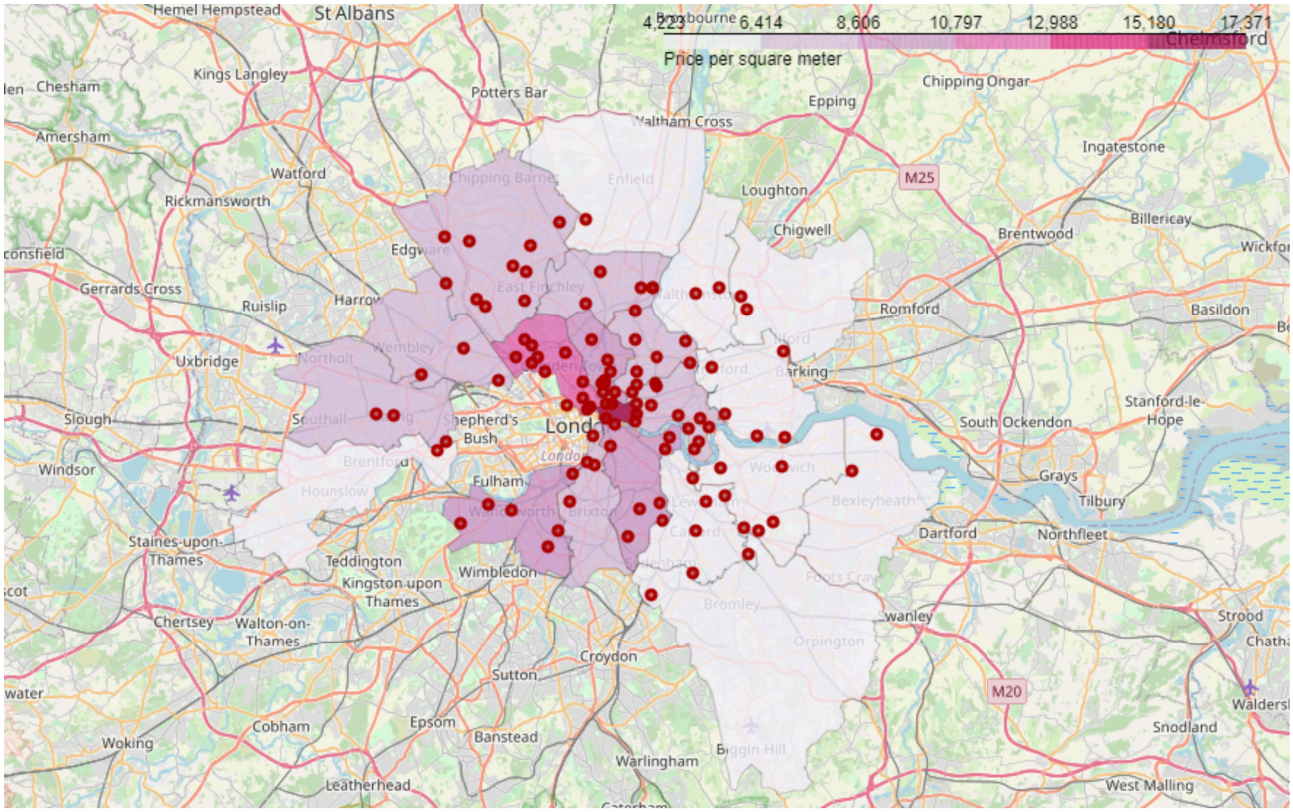

Figure 11: Map of London showing areas similar to gentrified areas and real estate prices

As can be seen in Figure 11, many areas that are similar to gentrified areas can be found in and around the center, in boroughs that already have high real estate prices. This is no surprise; these might be areas that have already gentrified in the past or areas that have always been popular. However, it can also be observed that there are some areas in boroughs with lower real estate prices that have similar venues as areas that are currently popular. Most prominently, there is a pocket in the southeastern part where the real estate prices areas are still low.

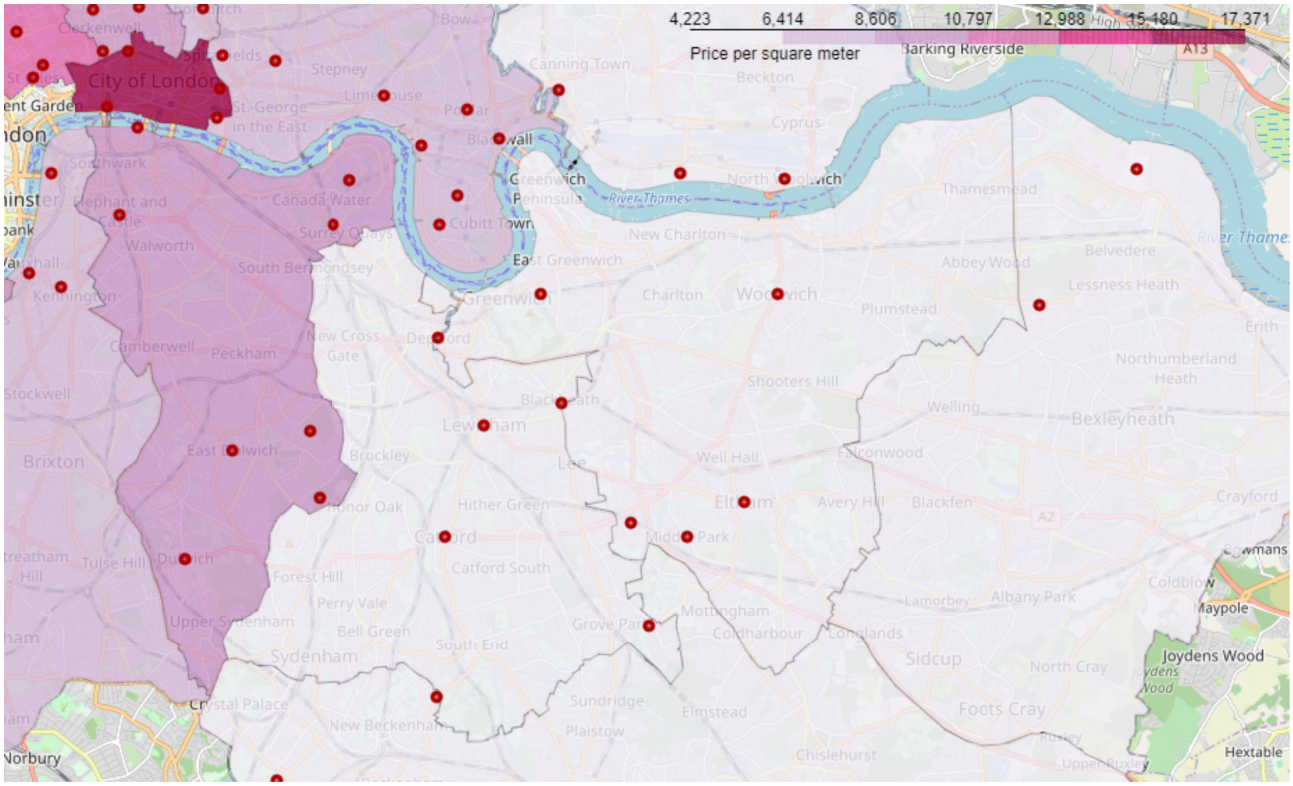

Figure 12: Southeastern London with low real estate prices and many areas that are similar to gentrified areas

Therefore, from Figure 12 it can be concluded that the areas most prone to gentrification in the near future are in southeastern London. Boroughs that seem the most at risk are for example Lewisham or Greenwich.

5. Discussion

Gentrification, of course, is a much more complicated problem than was assumed in this project. In reality determining whether an area is prone to gentrification depends on much more factors than just the venues in the neighborhood (although research shows that it plays a big role).

Therefore, in order to make this research more valuable, the similarity of neighborhoods should be compared using multiple aspects such as demographic data, migration data, etc. Furthermore, more effort could be put in to find the optimal amount of clusters, to make the clustering process more accurate and the real estate price could be analysed more locally and with more recent data. Finally, the term “gentrification” and its indicators or measurements should be defined more clearly, which was not in the scope of this assignment.

Other interesting things that could be done with this project would be for example predicting the real estate prices based on the venues in an area using linear regression, to predict the “real value” of an area (as opposed to the current real estate price). Areas that have a big gap between the predicted “real value” of real estate and the current real estate price, might be even more prone to gentrification.

6. Conclusion

This project was done in order to fulfill the requirements for the IBM Data Science Capstone course. Although the data analysis, methodology and machine learning techniques were all done according to standard rules and conventions, no real value can be assigned to this analysis. However, for me it was a very valuable experience, practicing the wrangling of data, using packages as Pandas and Folium, learning to work with API’s, learning to handle JSON files, practicing with machine learning techniques in a real world problem and in general getting more hands-on experience with data science, and I enjoyed it sincerely.