Venues Data Analysis for London SE

Introduction

Problem and background

- London has 8 main postcode areas, namely the N, NW, SW, SE, W, WC, E and EC postcode areas. In our analysis, we want to focus on the SE postcode area (South Eastern part of London). It loosely corresponds to the Boroughs named after Southwark, Lewisham and Greenwich plus indicated parts of those named after Croydon (north), Lambeth (east), Bexley (west) and Bromley (its northwest corner).
- In this analysis, we want to try and cluster the districts in the London SE postcode area (South East) in a meaningful way.

Data Description

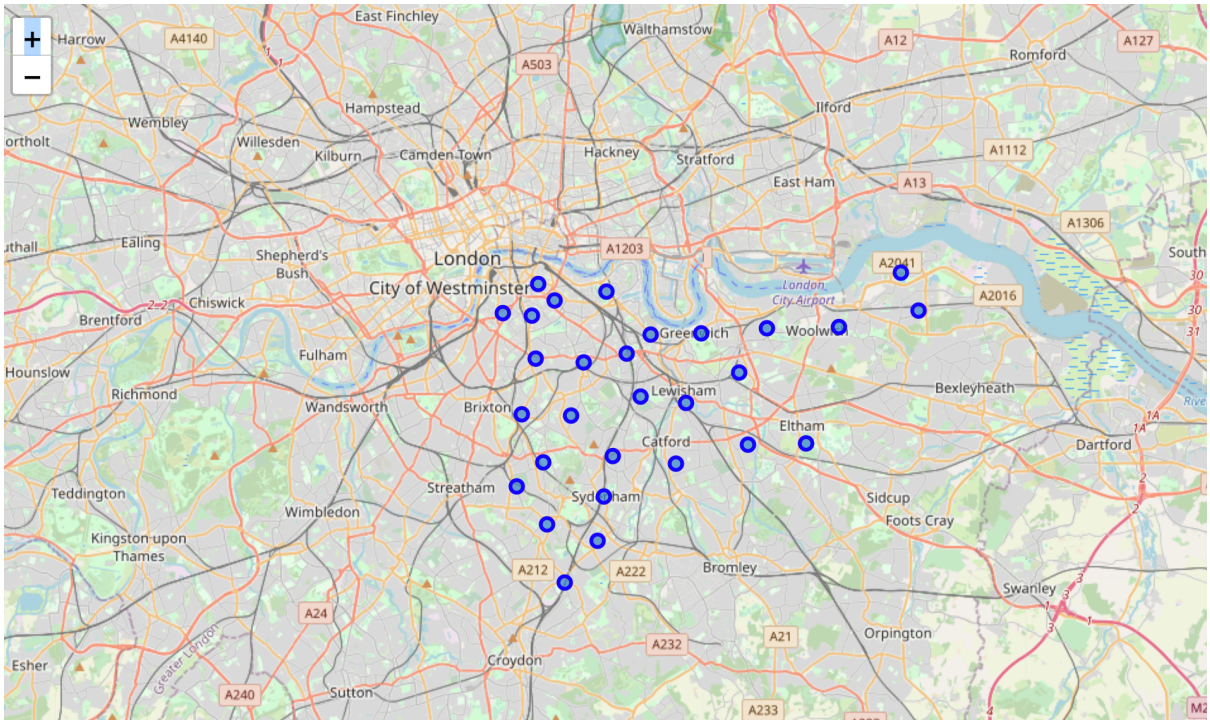
- I downloaded the 'Outcode Area Postcodes' from the FreeMapTools website. The csv file provides all postcode areas with their corresponding latitude and longitude. It is also possible to download the full list of UK postcodes with their latitude and longitude on this website.
- The digit(s) following the first two letters 'SE' correspond to a district within that area. This is followed by a space and then a number denoting a sector within said district, and finally by two letters which are allocated to streets or sides of a street. SE has 29 postcode districts, and 129 postcode sectors.
- The hierarchy is as follows: postcode area > postcode district > sector within district > streets within sector.
- I used the Foursquare API to get the most common venues for each of the 29 postcode districts of London SE.

Methodology

- First, we want to find the corresponding latitude and longitude for each of the 29 postcode districts located in London SE. We can clean the data downloaded from the FreeMapTools website, and reduce it to London postcode SE.
- Then we will use the Foursquare API to explore these districts. We will use the explore function to obtain the most common venue categories in each district.
- Then we will use this feature to group the districts into clusters. We will use the k-means clustering algorithm to complete this task.
- Finally, we will use the Folium library to visualise the results, i.e. the districts in London SE and their emerging clusters.

Analysis

Using the Python folium library, I first visualised London SE and its postcode districts, using latitude and longitude values.



I then used the Foursquare API to explore the postcode districts and segment them. I designed the limit as 100 venue and the radius 500 meter for each district from their given latitude and longitude. Below is a snap of the first five rows of the information retrieved from the Foursquare API, namely the venues' name, categories, latitude and longitude.

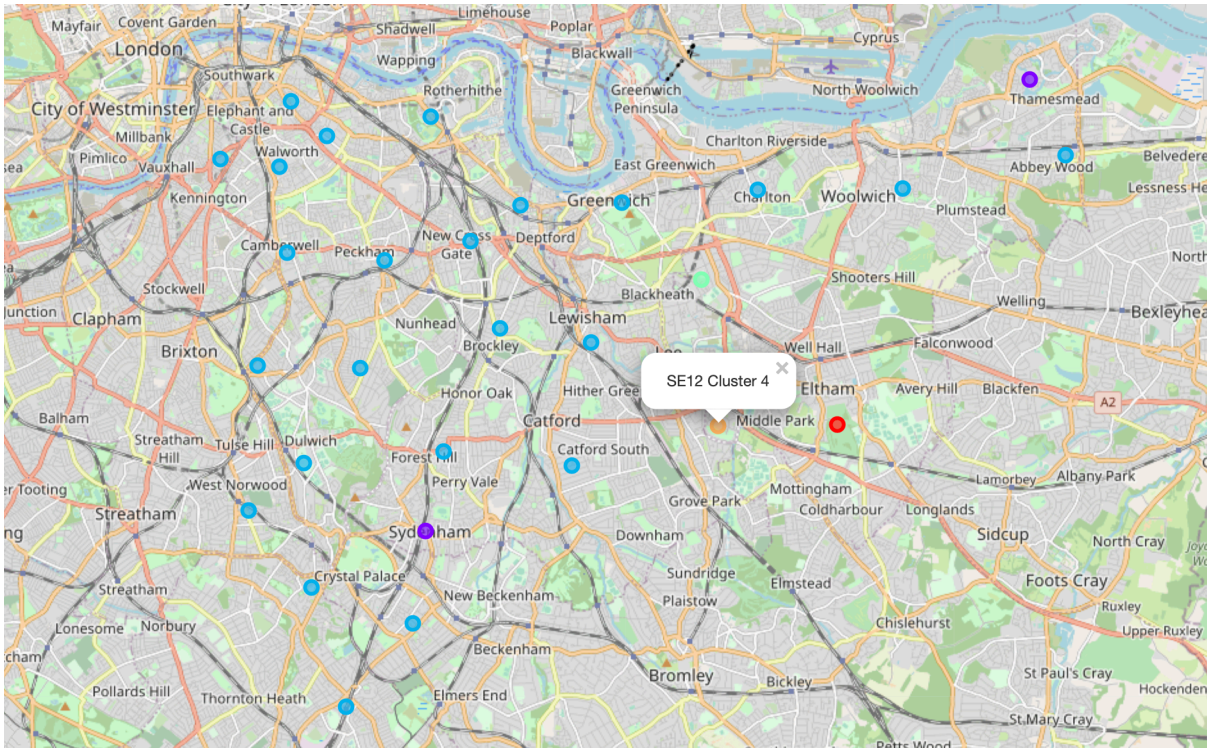
Some venues' categories are common to different districts. To cluster districts, I used the unsupervised learning k-means algorithm to cluster the boroughs together. K-means is one of the most common cluster methods of unsupervised learning.

```
[46]: se_merged.head()
```

[46]:

	district	latitude	longitude	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	SE1	51.49838	-0.08949	2	Pub	Italian Restaurant	Garden	Park	Residential Building (Apartment / Condo)	Lebanese Restaurant	Coffee Shop	Theater	Café	Fast Food Restaurant
1	SE10	51.48162	-0.00089	2	Pub	Garden	Coffee Shop	Grocery Store	Café	Turkish Restaurant	Historic Site	Science Museum	Indian Restaurant	Pier
2	SE11	51.48880	-0.10862	2	Pub	Café	Coffee Shop	Gastropub	Pizza Place	Indian Restaurant	Italian Restaurant	Fish & Chips Shop	Kebab Restaurant	Museum
3	SE12	51.44430	0.02483	4	Park	Laundromat	Yoga Studio	Gaming Cafe	Fried Chicken Joint	French Restaurant	Forest	Food Truck	Food & Drink Shop	Flower Shop
4	SE13	51.45837	-0.00910	2	Pub	Clothing Store	Fast Food Restaurant	Coffee Shop	Café	Gym	Grocery Store	Video Game Store	Restaurant	Portuguese Restaurant

Visualise the clusters:



Discussion

We examine each cluster and determine the venue categories that distinguish each cluster. Based on the defining categories, we can characterise each cluster.

- Cluster 1 corresponds to SE9 (Eltham) only. The most common venues for this cluster are hardware stores, followed by golf courses, so we can deduce that it is a residential area.
- Cluster 2 corresponds to SE26 (Sydenham) and SE28 (Thamesmead). The most common venues for these districts are supermarkets and fast food restaurants.
- Cluster 3 is the largest cluster, as it corresponds to 28 districts, with the most common venues being pubs, restaurants and cafés.
- Cluster 4 corresponds to SE3 (Charlton), with the most common venues being photography studios, followed by yoga studios.
- Cluster 5 corresponds to SE12 (Catford), with the most common venues being parks, followed by laundromats and yoga studios. We can deduce that it a residential area as well.

Conclusion

We managed to cluster the districts within the SE postcode area. It can be useful for individuals looking to buy or rent a house, or set-up a business. As a next step, it would be useful to:

- Create maps and information charts showing the housing prices and where each district is clustered according to the venue density.
- Extend the model to the other postcode areas in London, i.e. N, NW, SW, W, WC, E and EC.
- Zoom into a district to cluster the individual postcodes.
- Use another clustering algorithm. Different approaches can be attempted to cluster London districts. Not every classification method can yield the same high quality results for this

metropole. I chose to use the k-means algorithm, but it would be interesting to try and use another algorithm.

- Try and access the data dynamically from specific platforms or packages.