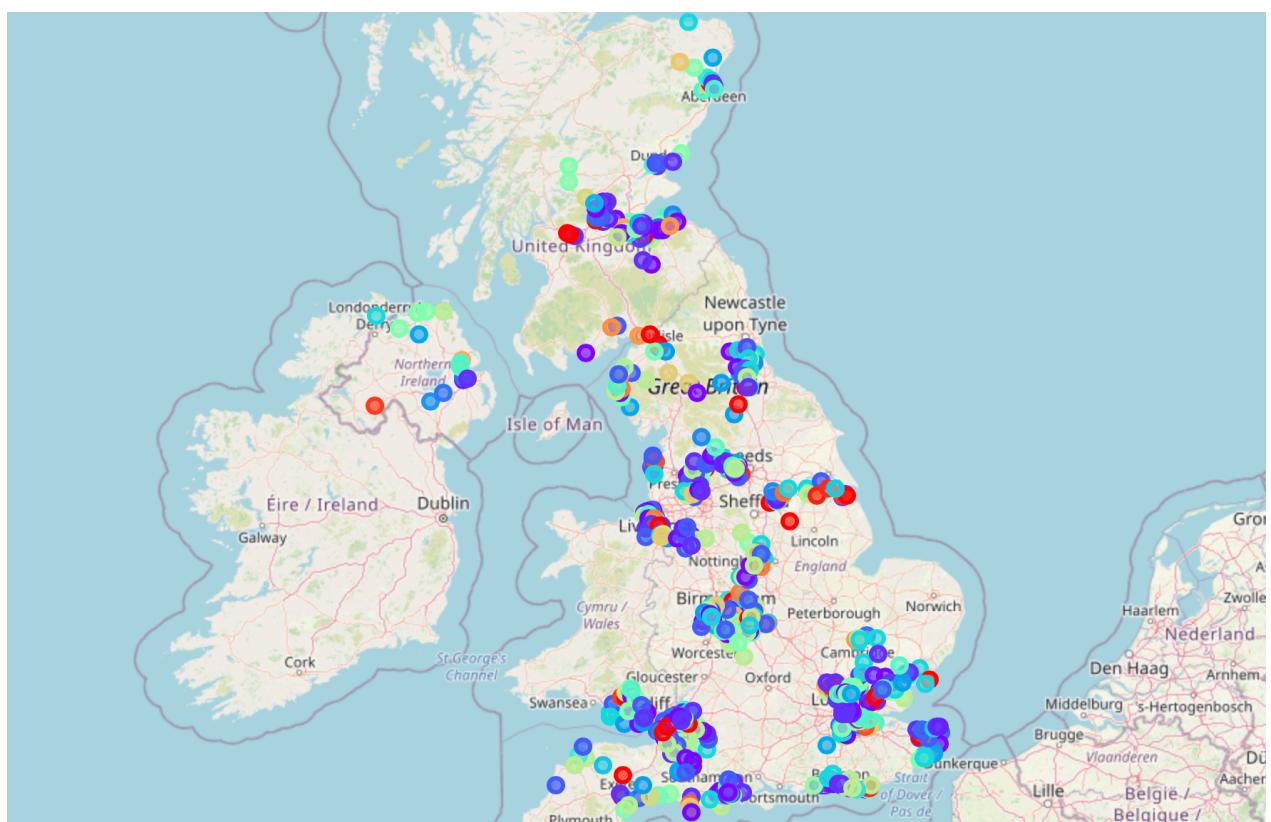


Capstone Project

UK Area Similarity

The current pandemic has changed our lifestyle in many ways. One newly adopted norm is working from home. This means that employees don't have to be close to their workplace anymore. Residents in the UK may want to explore other parts of the country that are similar to their current area. This project aims to help them to identify similar areas in the UK based on what amenities are available in the area.



The code for this work can be found here: https://github.com/vthawfeek/Capstone_Project/blob/main/UK_Area_Similarity_Capstone_Project.ipynb.

Introduction

This project aims to group different areas in the UK into similar clusters based on their neighbourhood venues. The generic results of this work can have applications beyond just deciding relocation areas including business prospects and buying property.

Data

To explore the similarities between areas in the UK, I use the data provided by freemaptools [here](#). This downloaded csv file contains a list of all the postcodes and corresponding latitudes and longitudes of areas in the UK. I also use Wikipedia for information on the names of the areas from [here](#).

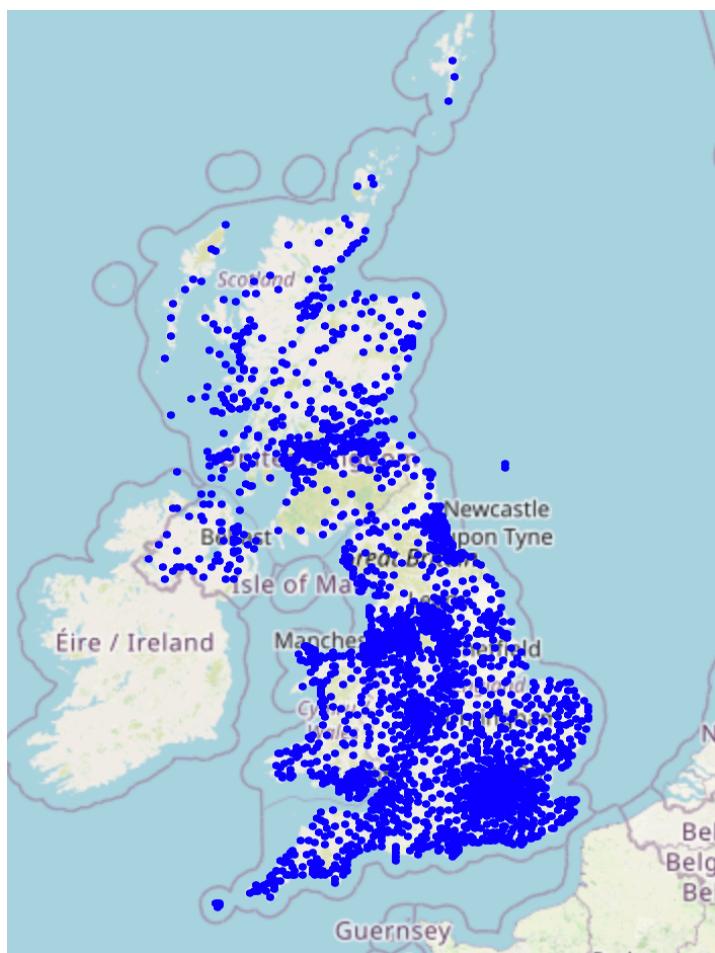
Methodology

1. I start by collecting data from the different sources using relevant Python libraries.
2. The second stage is to preprocess this data for analysis. Location names from the Wikipedia dataset is mapped to postcode-latitude-longitude information in freemaptools. Only the first section of the postcodes are selected for each location. For example, area ‘Cambridge’ will have postcodes ‘CB1’, ‘CB5’, ‘CB8’, etc. This data is then visualised on the UK map using geopy and folium libraries.
3. Next, I use Foursquare API to get venues in the neighbourhood. The top 100 venues within a 500 meter radius are acquired for each latitude-longitude. To make this process efficient, a function is defined to retrieve the information from Foursquare. The daily limit for a free account with Foursquare should taken into consideration. Fortunately, the number of queries in this case was within this limit. The venue category information was transformed into one-hot key code for the next step.
4. A k-means clustering is performed on the data. It is important to know the optimal number of clusters for this. I use the elbow approach to identify the cluster counts that place the centroids equidistant from all the instances in the cluster.
5. Cluster numbers is assigned to each area and this information is visualised on the UK map. Each area can have more than one cluster assigned to it because of the different postcodes that it can have.

6. To figure out area similarities, I build a clustered heat map by grouping postcodes to areas. This map plots a dendrogram showing how similar different areas are to each other.

Results

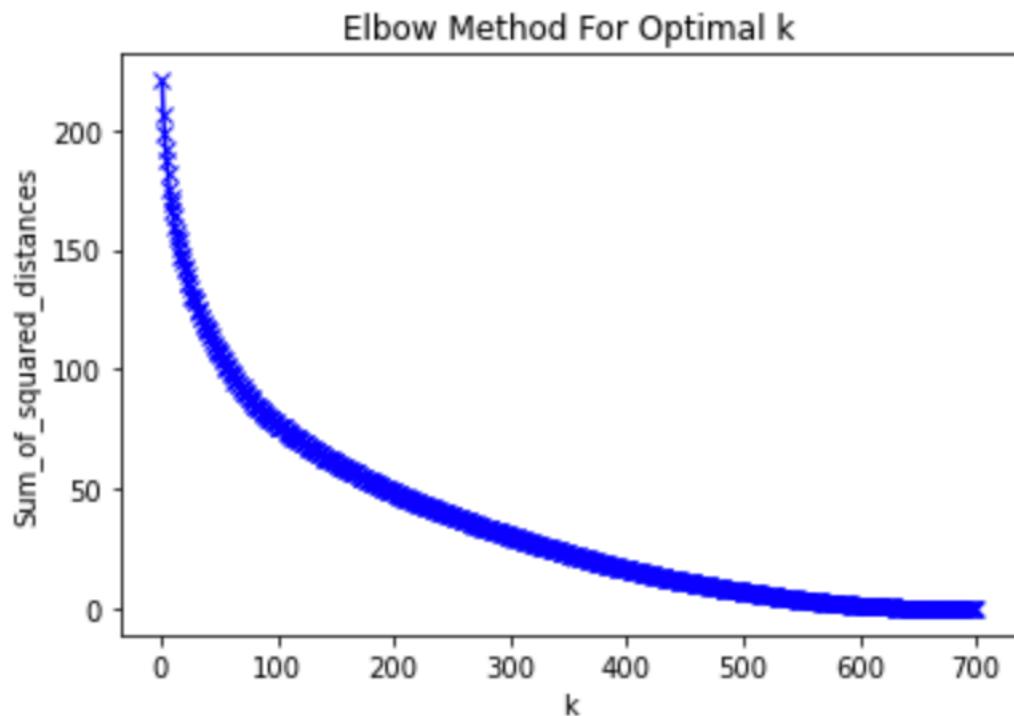
Data is collected from freemaptools and Wikipedia and processed for further analysis. This dataset contains information on the area, postcode, latitude and longitude. In total there were 2977 instances.



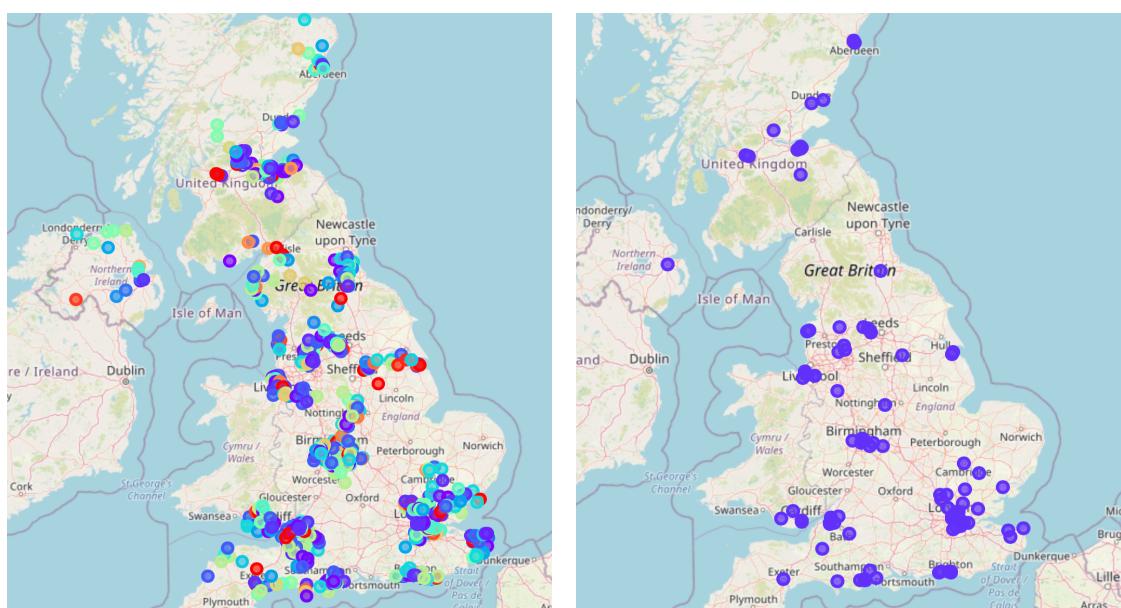
Neighbourhood venue information for all the postcodes was obtained from Foursquare API. Some of the postcodes did not have this data and were ignored for this analysis. The final processed file has 8070 instances and contains information on the area and its neighbourhood venues. There were a total of 381 unique venue categories.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0 AB10	57.135204	-2.120402	Hass's Fish And Chips	57.138142	-2.114179	Fish & Chips Shop
1 AB10	57.135204	-2.120402	Maggie's Grill	57.137788	-2.114346	American Restaurant
2 AB10	57.135204	-2.120402	Atlantis Restaurant at The Mariner Hotel	57.134499	-2.127158	Hotel Bar
3 AB10	57.135204	-2.120402	The Holburn	57.137526	-2.113954	Bar
4 AB10	57.135204	-2.120402	Tesco Express	57.138017	-2.113941	Grocery Store

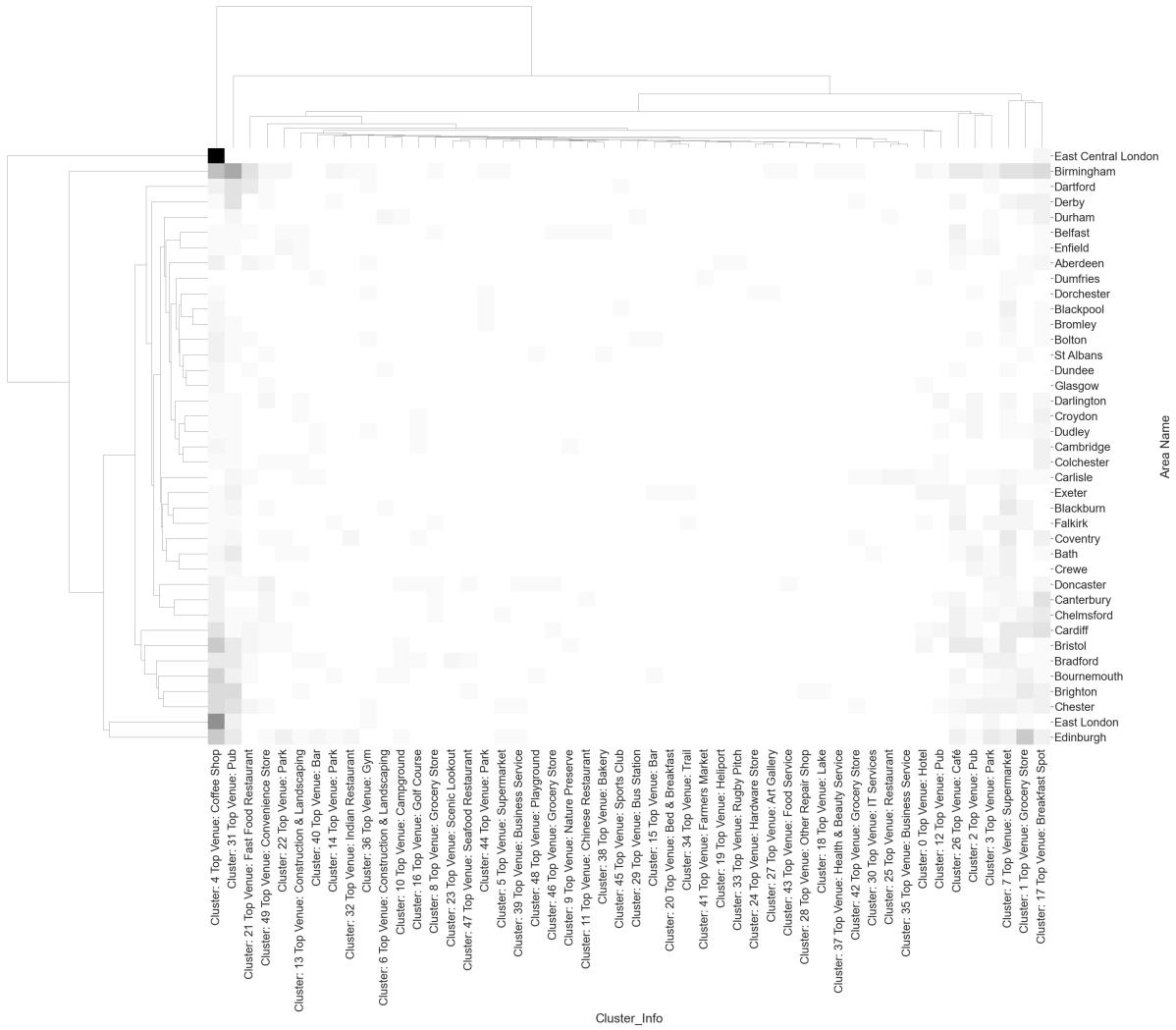
Subsequently, a k-means test was performed to cluster the different postcodes. The k-mean elbow method was used to decide an optimum cluster number. In this case it turned out to be 50.



The cluster numbers were assigned to each postcode and visualised on the UK map. We observe the some areas like Oxford and Norwich have no venue information while big cities like London and Birmingham have several. The predominant group in big cities is ‘Cluster 4’ (right).



This clustering result will help to understand similar postcodes in the UK. We might also be interested in broader questions such as which areas have similar neighbourhoods. A dendrogram map of different areas based on postcode clustering provides this information. We find that ‘East Central London’ is quite distinct from other areas and forms a separate class. Some examples of similar areas are the following: ‘East London’ & ‘Edinburgh’, ‘Dundee’ & ‘Glasgow’ and ‘Cambridge’ & ‘Colchester’.



On closer observation of ‘Cluster 4’, we find that the most common venues are coffee shops and pubs. This makes sense because most places around big cities indeed have these venues.

Cluster Labels	1st Most Common Venue	Counts
4.0	Bar	5
	Café	4
	Clothing Store	8
	Coffee Shop	34
	Convenience Store	4
	Discount Store	3
	Fireworks Store	1
	Grocery Store	12
	Gym / Fitness Center	3
	History Museum	1
	Hotel	8
	Hotel Pool	1
	IT Services	1
	Indian Restaurant	5
	Italian Restaurant	3
	Market	2
	Movie Theater	1
	Park	2
	Pharmacy	1
	Pizza Place	2
	Platform	2
	Pub	28
	Sandwich Place	1
	Stationery Store	1
	Supermarket	4

Discussion

From the cluster information overlaid on the UK map, we can notice that data is missing for a few areas like Oxford and Norwich. This has to be taken into consideration when interpreting the results. As expected there are more postcodes in big cities like London, Birmingham, Liverpool and Manchester.

Majority of areas in cities fall under cluster 4. The most popular venue in this cluster are coffee shops and pubs.

If you like a big city atmosphere but want to relocate to a different area, results show that there are some towns which also are under cluster 4. These places can be good choices for you!

Conclusion

1. Data from Foursquare, freemaptools and Wikipedia were used to investigate the similarity of areas in the UK based on their neighbourhood.
2. Information distribution was heterogenous with more data for big cities like London and Birmingham while few areas like Oxford, Peterborough and Norwich had no data.
3. Most areas in cities were grouped under cluster 4 where the most popular venues were coffee shops.
4. Results show that there are other towns outside big cities that also were categorised as cluster 4 indicating a similar surrounding. For example, we can observe from the cluster map that East London is similar to Edinburgh.
5. This map can be used to evaluate the similarity of areas for various purposes such as relocation or starting a profitable business or buying properties.