The Battle of Neighborhoods

Capstone Project – July 2020

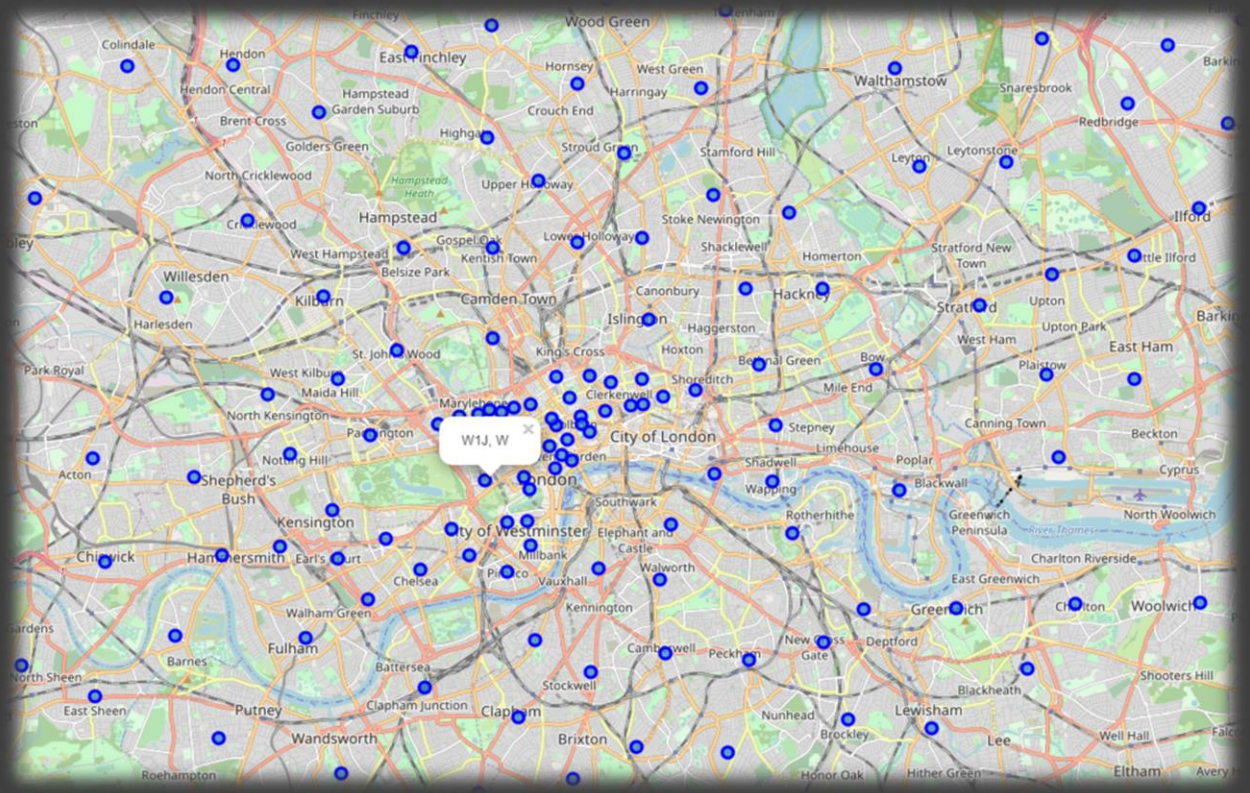
RY

Contents

Introduction

The HR department of a large multinational company is helping employees who are unfamiliar with London to find an area to live in. These employees might be transferring from another office outside of London/UK, or who lived in one area of London for a few months and now looking for a more long-term place.

The aim of this project is to provide insight into different districts in London based on similarities and dissimilarities, which an employee can use to focus their accommodation search on. This project will analyze the districts in the Greater London area, UK, and take in key considerations employees might have when looking for areas to live in.



Data

- The data used for this project will be acquired from publicly available sources, such as Wikipedia and/or government sites, as well as venue information through Foursquare API's search feature. The dataset for location points consists of postcodes in the Greater London area, postcode districts and area for each postcode and coordinate information for each postcode.

- The focus will be on unique postcode district codes, they consist of groups of postcodes representing population centers and neighborhoods in each wider district within the Greater London area. Additional information for each postcode district, such as population, average housing price in the 2019, will also be used alongside local venue data to provide insight into the qualities and affordability of a neighborhood based on postcode district codes.

Data source

- Greater London postcodes, coordinates and population:
  https://www.doogal.co.uk/UKPostcodesCSV.ashx?area=London
- Average sold house prices by postcode districts in 2019:
  https://landregistry.data.gov.uk/app/standard-reports/download-report?utf8=%E2%9C%93&report=avgPrice&areaType=county&area=GREATER+LONDON&aggregate=pcDistrict&period%5B%5D=2019&age=any
- UK postcode area names:
  https://en.wikipedia.org/wiki/List_of_postcode_areas_in_the_United_Kingdom
- Foursquare venue data

Methodology

- Postcode district data were compiled by grouping postcodes that shares the same districts together, filtering out postcodes that had little/no population, with coordinates taken from the average of the group

- Venue data for each district was extracted using the Foursquare API's search feature, searching within a 1.5km radius

- K-means clustering was then used to group districts into 8 groups based on nearby venues and weighted by population bins

- Additional data on house prices and percentage of houses that were flats were added to the final labels for the districts, alongside cluster information to provide a better guide to end-user

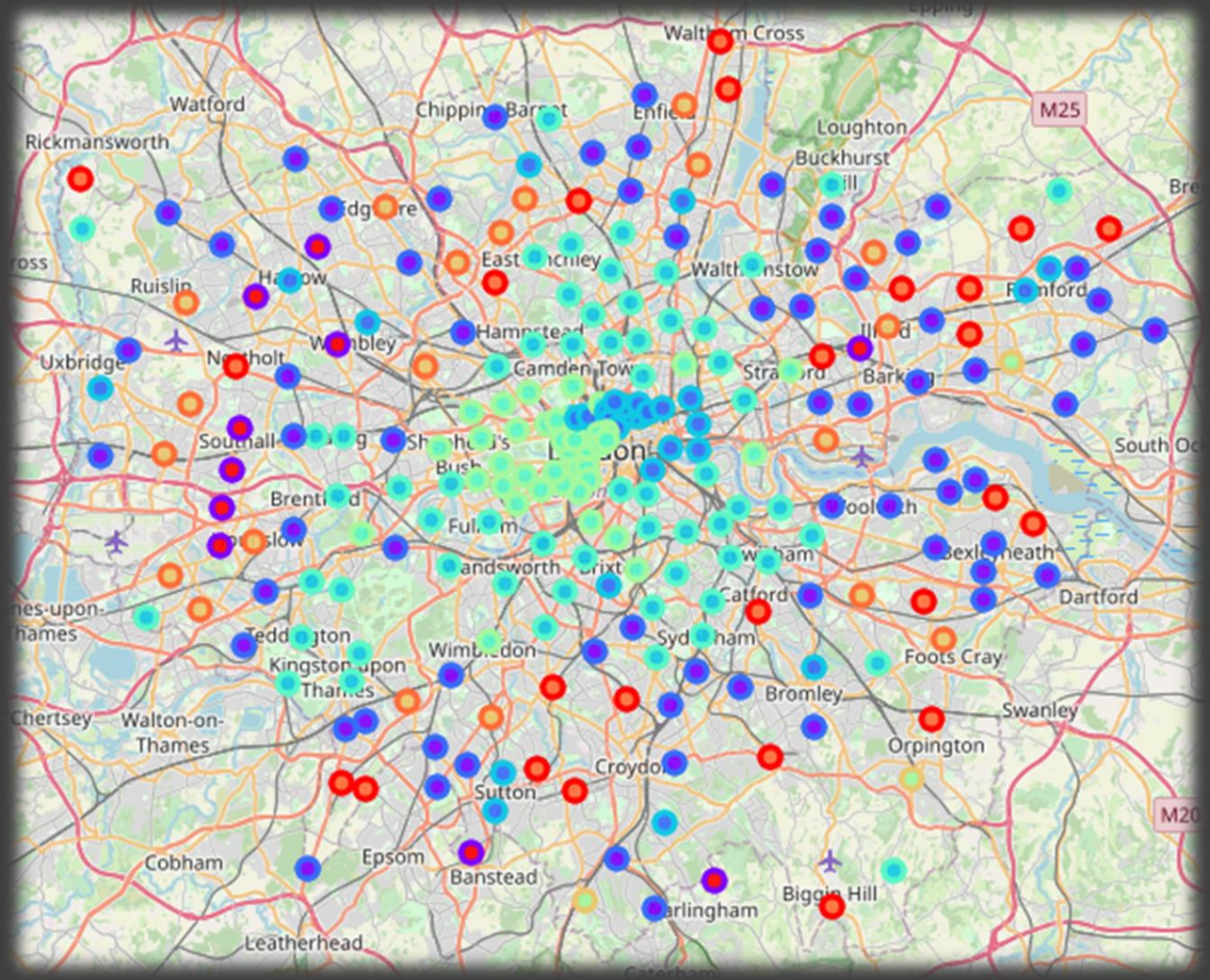
Clustering Analysis

Looking at the most common venues by clusters, we can get a sense the type of venues that make a cluster different to another. For example, cluster 0 has a high frequency of ‘Indian Restaurant’ in the area, and although cluster 3 and cluster 1 have ‘Pub’ as 1st most common venue, cluster 3 might have more upmarket feel with ‘Café’ and ‘Park’ high on the list too

Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Indian Restaurant	Grocery Store	Hotel	Fast Food Restaurant	Coffee Shop
1	Pub	Grocery Store	Coffee Shop	Supermarket	Park
2	Coffee Shop	Pub	Hotel	Café	Pizza Place
3	Pub	Café	Coffee Shop	Park	Italian Restaurant
4	Hotel	Café	Coffee Shop	Pub	Park
5	Platform	Supermarket	Pub	Grocery Store	Restaurant
6	Coffee Shop	Grocery Store	Supermarket	Park	Fast Food Restaurant
7	Grocery Store	Supermarket	Park	Coffee Shop	Café

Mapping the Clusters

Postcode districts were grouped into 8 clusters based on similarity of venues found in the vicinity and relative population size, this meant that districts that shared similar population size are more likely to be compared against each other for differences in venue types



Comparing Districts Within a Cluster

- The map view provides a general layout of how similar districts are spread across London, providing a quick view on where a district of same cluster outside of the usual concentration may lie
- This can be useful if someone is familiar of their own area already and looking to move to somewhere similar but in a different part of London
- For instance, if an employee currently living in a cluster 4 district (e.g. W1S) and looking to move to a similar but more suburban district where they can live in a house instead of flats, TW9 will be a good area to start their search; as it's less expensive but also likely to have more garden venues

Conclusion

- Using a combination of venue data and population, postcode districts across the Greater London area were mapped and classified into 8 clusters, with additional house price data to help aid someone narrowing down their search further
- However, the data does not include other area information that could be important for the end user, such as local school quality for families, local crime stats and commute distance to workplaces
- Another limitation is the use of historical data, such as 2019 house prices, which will need updating to stay relevant to current situation