Capstone Project – Battle of the Neighborhoods

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 sources, cleaning and feature selection:

Greater London postcodes, coordinates and population:
https://www.doogal.co.uk/UKPostcodesCSV.ashx?area=London

The link above downloads the CSV file that contains a list of all postcodes in the Greater London area along with additional information for each. The relevant columns from the original CSV are:

Postcode – individual postcodes

In Use? – Whether the postcode is still in use or not. This column is used to filtered out postcodes that are no longer in use.

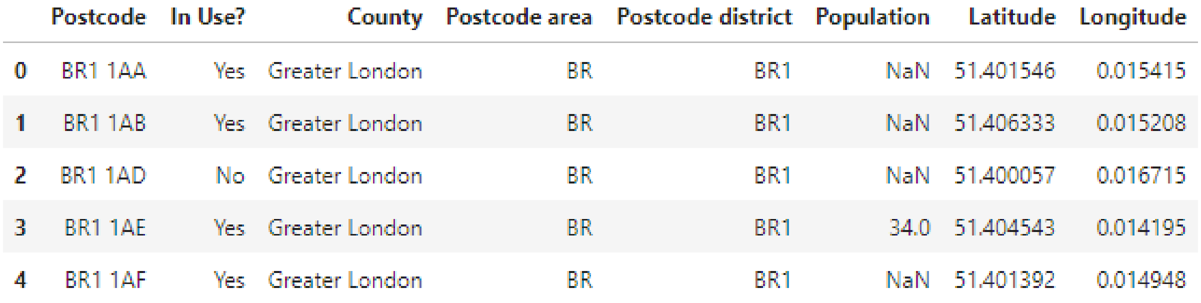
County – Some of the postcodes are under more than one county, so this column will be used to filter out duplicates by taking the wider Greater London county only

Ward – The name of the ward the postcode is in

Postcode area – The area code for the postcode, typically the first few letters in a UK postcode before the first number

Population – residential population recorded for each postcode. As the focus of this project is in potential residential areas, postcodes without population data (i.e. NaN) or less than population of 10 will be assumed to have no or little residential property available and filtered out.

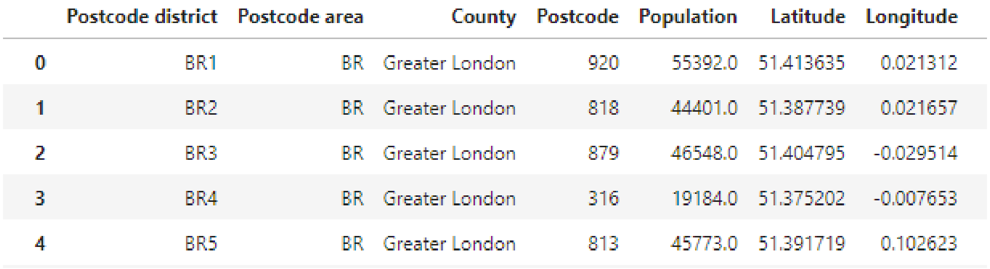
Latitude and Longitude – coordinates for each postcode



	Postcode	In Use?	County	Postcode area	Postcode district	Population	Latitude	Longitude
0	BR1 1AA	Yes	Greater London	BR	BR1	NaN	51.401546	0.015415
1	BR1 1AB	Yes	Greater London	BR	BR1	NaN	51.406333	0.015208
2	BR1 1AD	No	Greater London	BR	BR1	NaN	51.400057	0.016715
3	BR1 1AE	Yes	Greater London	BR	BR1	34.0	51.404543	0.014195
4	BR1 1AF	Yes	Greater London	BR	BR1	NaN	51.401392	0.014948

Fig 1: Example of postcode data extracted from raw CSV

Individual postcodes are then grouped by the ‘Postcode district’ column, which typically is the first section of an individual postcode. The new ‘Postcode’ column will contain the count of postcodes in a postcode district, the new ‘Population’ column will be the sum of all population in each postcodes and the average latitude and longitude were taken as the new coordinates for the postcode district under the new relevant columns. Note that as postcodes with less than 10 population are excluded before grouping, the average coordinates will gravitate less towards areas with little or no residential properties.



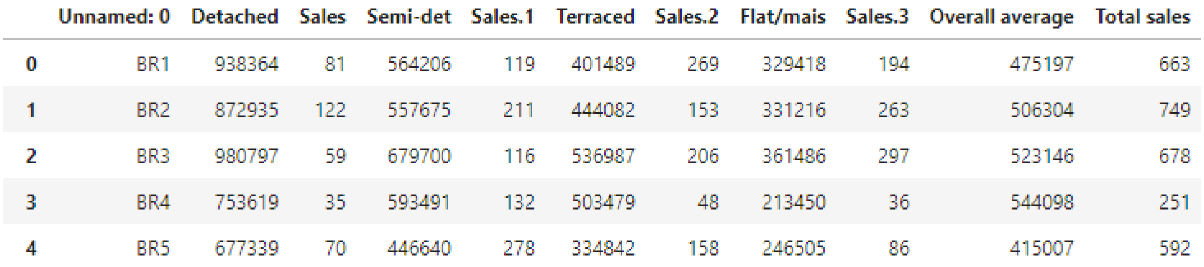
	Postcode district	Postcode area	County	Postcode	Population	Latitude	Longitude
0	BR1	BR	Greater London	920	55392.0	51.413635	0.021312
1	BR2	BR	Greater London	818	44401.0	51.387739	0.021657
2	BR3	BR	Greater London	879	46548.0	51.404795	-0.029514
3	BR4	BR	Greater London	316	19184.0	51.375202	-0.007653
4	BR5	BR	Greater London	813	45773.0	51.391719	0.102623

Fig 2: Example grouping of postcodes by postcode districts

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

Another feature that will be included is the relative cost of moving into a property in one neighborhood versus other neighborhoods. This project will use land registry data on sold house prices as a metric to compare difference in cost to move into a new area. The above link leads to the page where the data on the average sold prices for all properties in each postcode districts in 2019 can be downloaded.

The raw file includes a breakdown of type of property sold, number of sales of that type and the average prices. The property type will be used to determine if properties available in an area is high in flats/apartments, which tend to be smaller, or houses, which will be more suitable for families.

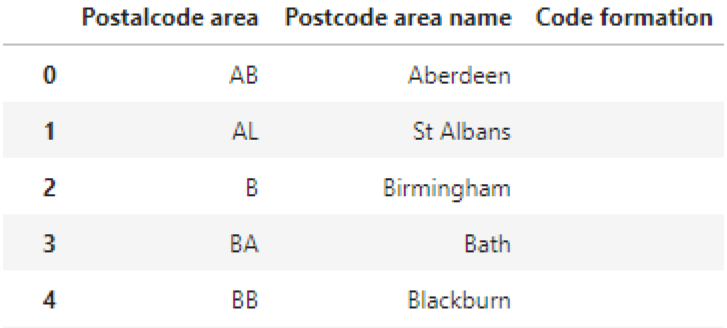


	Unnamed: 0	Detached	Sales	Semi-det	Sales.1	Terraced	Sales.2	Flat/mais	Sales.3	Overall average	Total sales
0	BR1	938364	81	564206	119	401489	269	329418	194	475197	663
1	BR2	872935	122	557675	211	444082	153	331216	263	506304	749
2	BR3	980797	59	679700	116	536987	206	361486	297	523146	678
3	BR4	753619	35	593491	132	503479	48	213450	36	544098	251
4	BR5	677339	70	446640	278	334842	158	246505	86	415007	592

Fig 3: Example of raw data from the average sold house prices by postcode districts

UK postcode area names:
https://en.wikipedia.org/wiki/List_of_postcode_areas_in_the_United_Kingdom

Lastly, to provide some additional information on the wider area a postcode district falls under, the postcode area names will be provided. Here, the area name will be taken from the first letters of the postcode district (i.e. BR from BR1, BR2…), and matched with the list of names in the Wikipedia page link above.



	Postalcode area	Postcode area name	Code formation
0	AB	Aberdeen	
1	AL	St Albans	
2	B	Birmingham	
3	BA	Bath	
4	BB	Blackburn	

Fig 4: Example of list of Postcode area names from the Wikipedia page

Methodology

Initial Exploratory Analysis

After extracting, filtering and grouping the Postcode district dataset, the calculated coordinates of each postcode district was plotted on a map of London using Folium. The postcode district code and the area code were used as labels. The map showed areas where there is a high concentration of postcode districts, such as central London, and less populated areas with lower number of postcode districts in the same area size.

Note that much of the southeast of central London (e.g. City of London) was empty compared to the west side of central London (e.g. Covent Garden), this was due to filtering out postcodes with less than 10 population, which reflected business districts with little or no residential properties within a postcode district group.

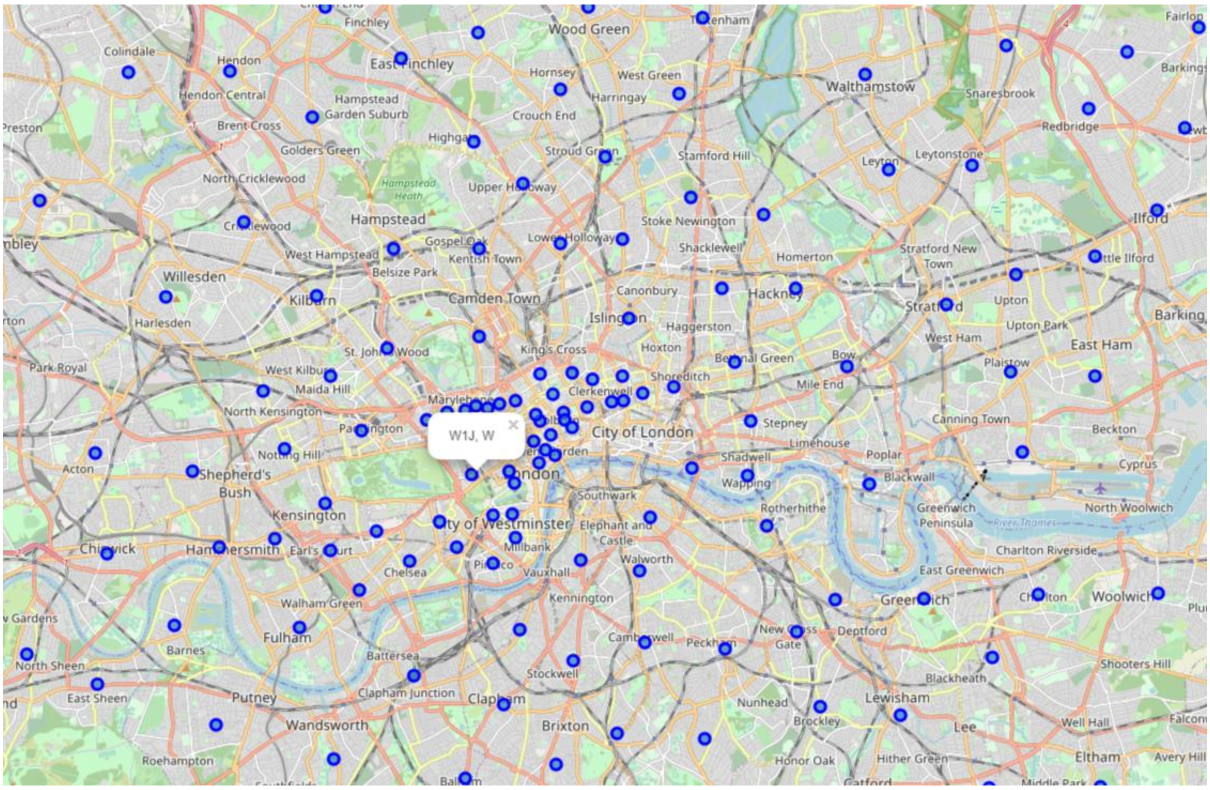
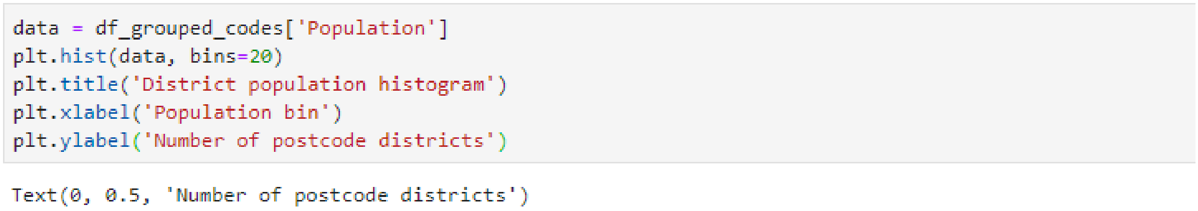


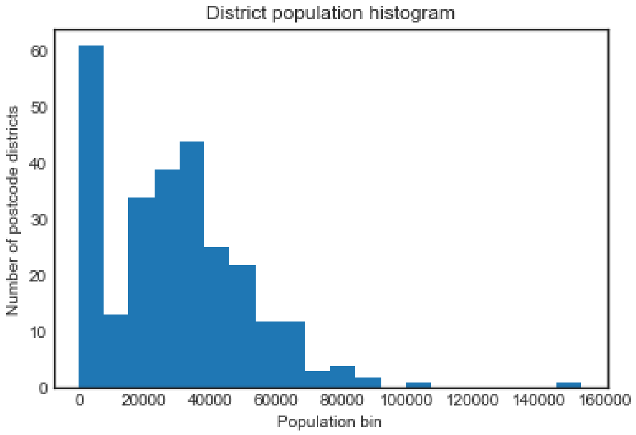
Fig 5: Snapshot of London area map with postcode districts plotted

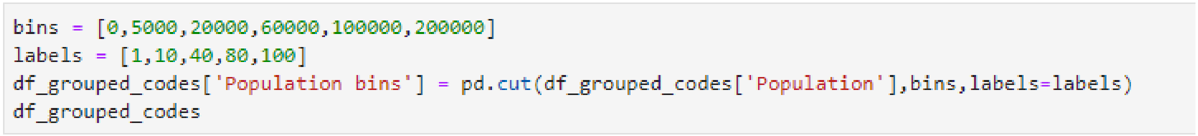
To consider the influence of population level for each district, the population of each postcode district is classed into five bins with labels that broadly reflect the scale of population difference for each district. This will be used as weights when classifying the districts based on venue data from Foursquare. Initial histogram plot showed many postcode districts had less than 5,000 in population, and a few had more than 100,000.



```
data = df_grouped_codes['Population']
plt.hist(data, bins=20)
plt.title('District population histogram')
plt.xlabel('Population bin')
plt.ylabel('Number of postcode districts')
```

```
Text(0, 0.5, 'Number of postcode districts')
```





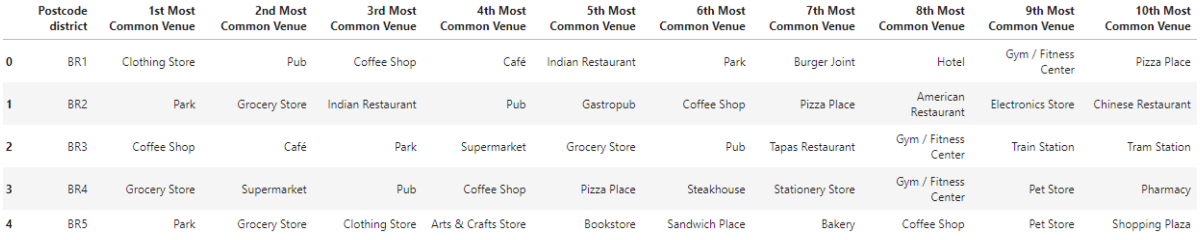
```
bins = [0,5000,20000,60000,100000,200000]
labels = [1,10,40,80,100]
df_grouped_codes['Population bins'] = pd.cut(df_grouped_codes['Population'],bins,labels=labels)
df_grouped_codes
```

Fig 6: Assigning each postcode district into 5 population bins

Venue data

Using the Foursquare API's search feature, a list of venues within a 1.5km radius was extracted for each postcode district using their coordinates. A total of 389 unique venue categories were returned. A quick review of the venue type that was extracted for all districts showed 'Pub' as the most popular venue, this is followed by 'Coffee Shop', Café', 'Grocery Store', and 'Park'.

Using onehot encoding and grouping methods, the mean of the frequency of occurrence of each category was calculated for each postcode district. Then, a function was created to display the top 10 venues of each district.



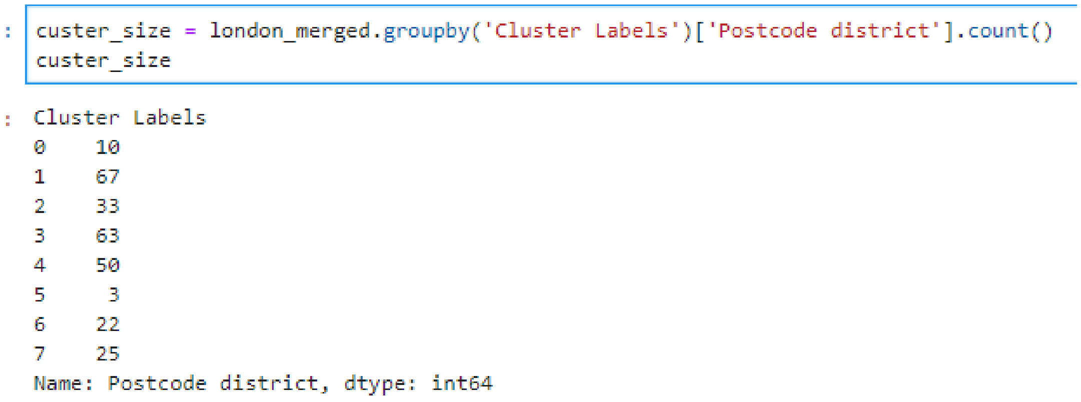
	Postcode district	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	BR1	Clothing Store	Pub	Coffee Shop	Café	Indian Restaurant	Park	Burger Joint	Hotel	Gym / Fitness Center	Pizza Place
1	BR2	Park	Grocery Store	Indian Restaurant	Pub	Gastropub	Coffee Shop	Pizza Place	American Restaurant	Electronics Store	Chinese Restaurant
2	BR3	Coffee Shop	Café	Park	Supermarket	Grocery Store	Pub	Tapas Restaurant	Gym / Fitness Center	Train Station	Tram Station
3	BR4	Grocery Store	Supermarket	Pub	Coffee Shop	Pizza Place	Steakhouse	Stationery Store	Gym / Fitness Center	Pet Store	Pharmacy
4	BR5	Park	Grocery Store	Clothing Store	Arts & Crafts Store	Bookstore	Sandwich Place	Bakery	Coffee Shop	Pet Store	Shopping Plaza

Fig 7: Sample of postcode districts with top 10 most common venues listed across

Cluster Analysis

K-means clustering was used to group the postcode districts by the type and frequency of venues, weighted by population bins. Districts in the same cluster should ‘feel’ more similar than those in different clusters. To help determine a good number of clusters k, the inertia attribute was used to identify the sum of squared distances to the nearest cluster center, for k ranges from 1-20. Then, based on the Elbow Method, the value 8 was chosen as the optimal value for k in this analysis.

After running K-means clustering on the data, where districts were grouped into 8 different clusters based on their similarity. Each cluster also had different numbers of districts, with cluster 1 containing 67 districts and cluster 5 only 3 districts.



```
custer_size = london_merged.groupby('Cluster Labels')['Postcode district'].count()
custer_size
```

```
Cluster Labels
0    10
1    67
2    33
3    63
4    50
5     3
6    22
7    25
Name: Postcode district, dtype: int64
```

Fig 8: Number of postcode districts in under each cluster

The London map then was updated with each color representing different clusters. At glance, west and east central London districts are in two different clusters, and some of the outer London districts, such as the one in Wimbledon shares the same cluster as west central London districts (in florescent blue), indicating more similarity to those areas than other nearby districts such as one in Wandsworth (in slightly darker blue hue).

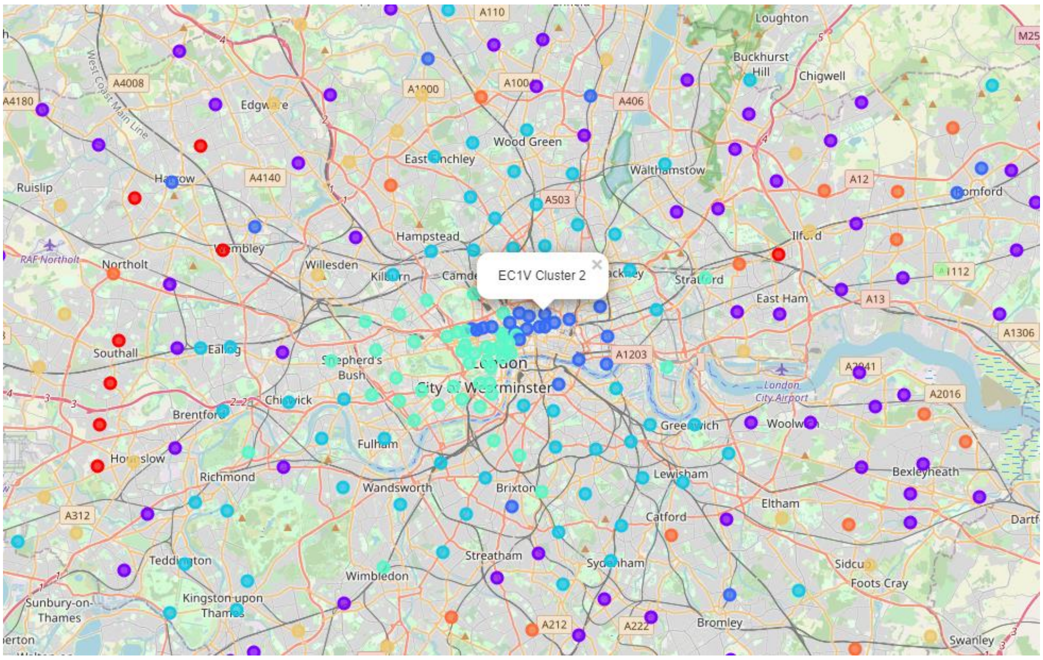
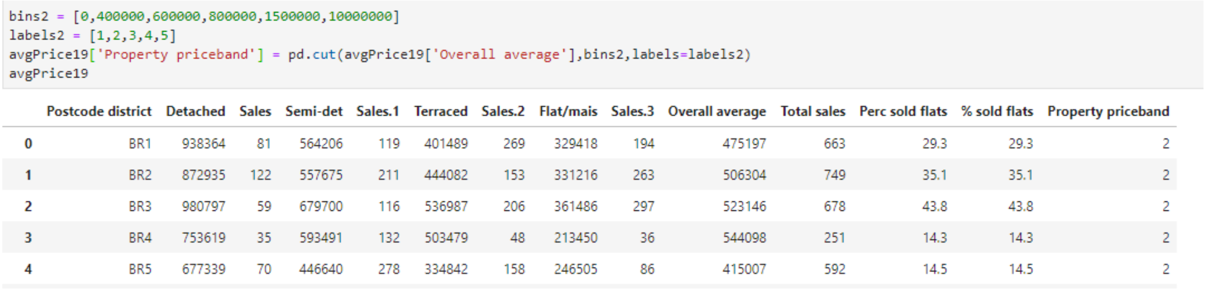


Fig 9: London map with cluster-based color coded Postcode districts

Additional data for labels

Like population bins earlier, the average house price for each postcode district was also grouped into bins into the property price bands. Furthermore, to provide more context to the prices of each district, the proportion of flats that made up the average price was also calculated using the house price dataset. So that an area that had a high percentage of flats will indicate a lack of residential houses available, and a higher cost per sqm compare to an area of similar average price but lower percentage of flats.

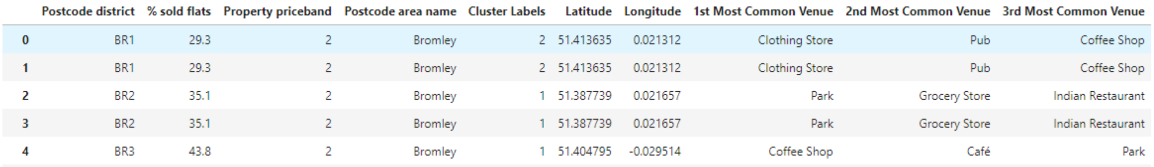


```
bins2 = [0,400000,600000,800000,1500000,10000000]
labels2 = [1,2,3,4,5]
avgPrice19['Property priceband'] = pd.cut(avgPrice19['Overall average'],bins2,labels=labels2)
avgPrice19
```

	Postcode district	Detached	Sales	Semi-det	Sales.1	Terraced	Sales.2	Flat/mais	Sales.3	Overall average	Total sales	Perc sold flats	% sold flats	Property priceband
0	BR1	938364	81	564206	119	401489	269	329418	194	475197	663	29.3	29.3	2
1	BR2	872935	122	557675	211	444082	153	331216	263	506304	749	35.1	35.1	2
2	BR3	980797	59	679700	116	536987	206	361486	297	523146	678	43.8	43.8	2
3	BR4	753619	35	593491	132	503479	48	213450	36	544098	251	14.3	14.3	2
4	BR5	677339	70	446640	278	334842	158	246505	86	415007	592	14.5	14.5	2

Fig 10: Assigning property price bands and proportion sold that were flats

Getting postcode area name from Wikipedia, the names were joined to the final table along with the new property columns, as well as other postcode district information to provide more data for labels in the final map.



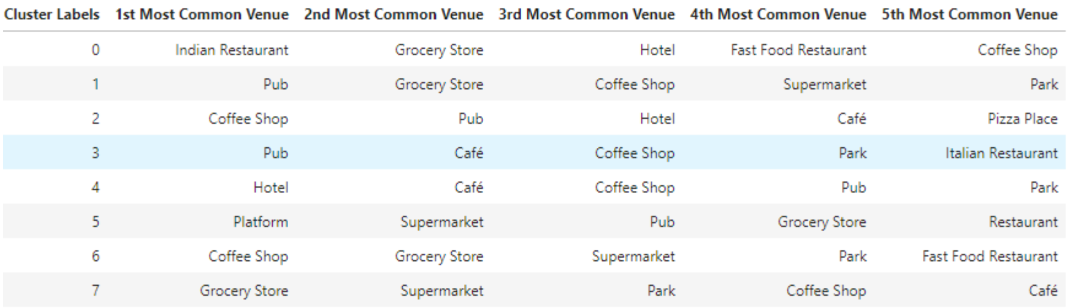
	Postcode district	% sold flats	Property priceband	Postcode area name	Cluster Labels	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	BR1	29.3	2	Bromley	2	51.413635	0.021312	Clothing Store	Pub	Coffee Shop
1	BR1	29.3	2	Bromley	2	51.413635	0.021312	Clothing Store	Pub	Coffee Shop
2	BR2	35.1	2	Bromley	1	51.387739	0.021657	Park	Grocery Store	Indian Restaurant
3	BR2	35.1	2	Bromley	1	51.387739	0.021657	Park	Grocery Store	Indian Restaurant
4	BR3	43.8	2	Bromley	1	51.404795	-0.029514	Coffee Shop	Café	Park

Fig 11: Final table for mapping and labelling the postcode districts

To provide more information how what makes each cluster different, like grouping by postcode district earlier, the venues were grouped by the cluster labels after merging the new cluster column with earlier tables using postcode district as key.

Results

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, it can be seen that 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

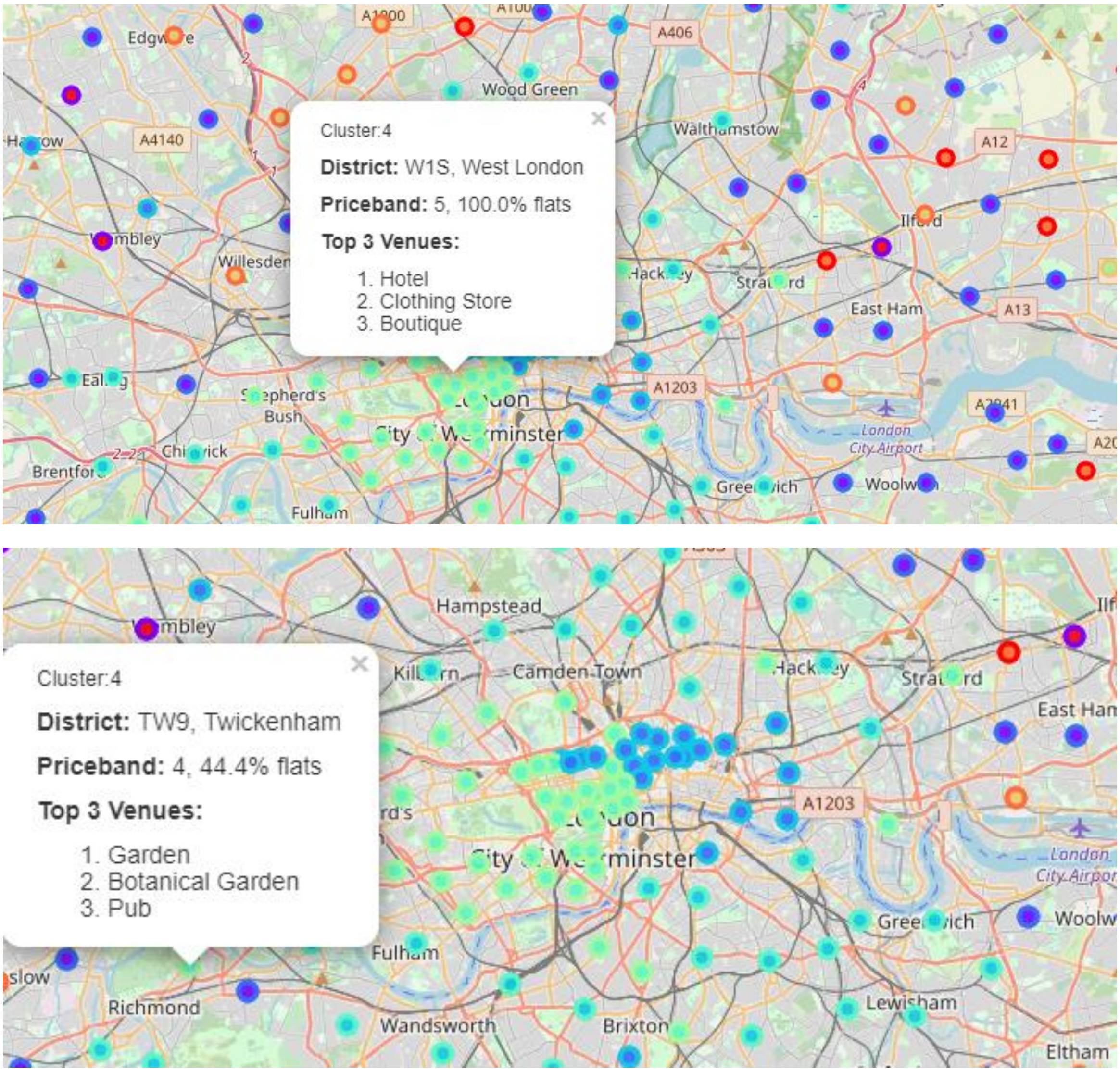
Fig 12: Top 5 most common venues by cluster labels

With the additional information for each postcode district, clicking on the postcode district dots on the London map will now provide information on house price bands, proportion of sold houses that were flats, and the top three venues in the area.

Discussion

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.



Cluster:4

District: W1S, West London

Priceband: 5, 100.0% flats

Top 3 Venues:

1. Hotel
2. Clothing Store
3. Boutique

Cluster:4

District: TW9, Twickenham

Priceband: 4, 44.4% flats

Top 3 Venues:

1. Garden
2. Botanical Garden
3. Pub

Fig 13: Comparing postcode districts on the map

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, i.e. 2019 house prices, which will need updating if price changes significantly, and will not reflect rental price exactly.