Venues data analysis – Birmingham vs London & Indian Restaurant Recommendations

Yasin Hassan

November 12th 2019

Part 1: A description of the problem and a discussion of the background.

In my project I will investigate in a general sense similarity between postcodes in Birmingham vs London. The following geospatial analysis will review location data by postcode and aim to form clusters based on location attributes. In this case I will extract data on top venue categories and distinguish which postcodes are similar based on that feature. This could then give insight to a business owner in London to expand on which postcode.

It is known that location is an essential feature for a store or business to perform well. Other types of venues could drive footfall to neighbouring venues. Hence a business owner could use clustering to examine areas that are in favourable clusters. I will observe the venues of each post code and form map visualisations to examine clusters.

The case which I will examine for will be Indian restaurants between both cities and try to recommend which Birmingham postcode an Indian restaurant owner from London would like to consider. Cases of postcodes that we are interested in will be of the following:

- Areas with no Indian restaurants but close to the centre of Birmingham as possible: opportunity to capitalise on an area missing out.
- Areas with Indian restaurants: Established locations with direct rivals may either cannibalise potential sales or let you benefit from footfall that is coming from being near rivals. This idea needs further inspection beyond the scope of this project but potential candidates for this category will be considered

In this notebook I hope to evaluate pros/cons of resulting clusters and give the best recommendation arising from the data.

Part 2: A description of the data and how it will be used to solve the problem.

Firstly, we will need to scrape postal location data of Birmingham and London to start with. Then retrieve the corresponding latitude & longitude coordinates of each postcode via geopy. If not, then perhaps another website to scrape the information.

Wikipedia tables will suffice for postcode data scrape. Their postal district tables will be used as a placeholder for a representative/central full postcode ('B1' as opposed to 'B1 1AA'). The Birmingham postal area (first 1-2 characters in a postcode) is just B while London is a much bigger city with 9 different postal districts (e.g. EC for East Central London).

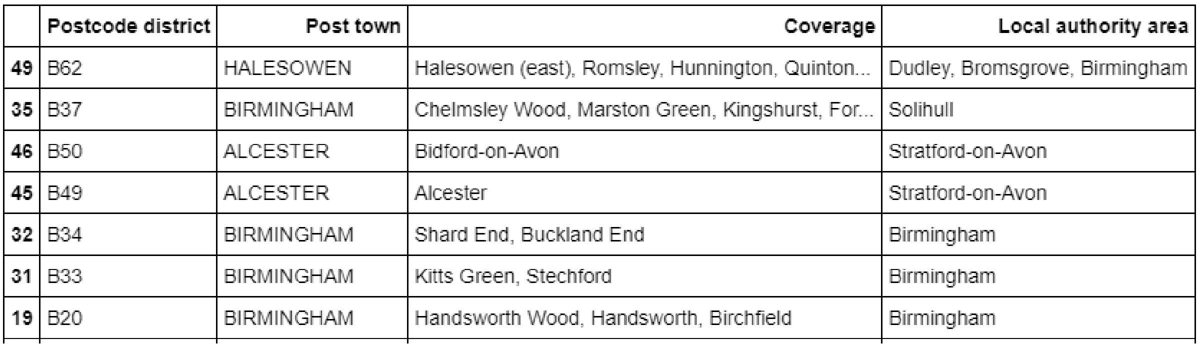
Important Notice: Nominatim within geopy library will be the API to source geographical coordinates from. It is an open-source API which is volunteer run but have its limits on usage. The method that will be implemented uses geopy but despite some efforts to handle failed requests you may come across a time where geocoding won't work. In practice for regular geocoding, it is advisable to store/cache results or use a premium API for stakeholder needs.

We will then need to make calls to the Foursquare API to request venue data for each location. This will then need some pre-processing to extract venue data from the raw json file into something that can be analysed.

After cleaning the data, we can prepare to run K means clustering and form clusters of post codes in Birmingham/London. The output data will try to label each post code to clusters based on venue categories they offer. We then finally discuss findings from each cluster and how they distinguish from each other.

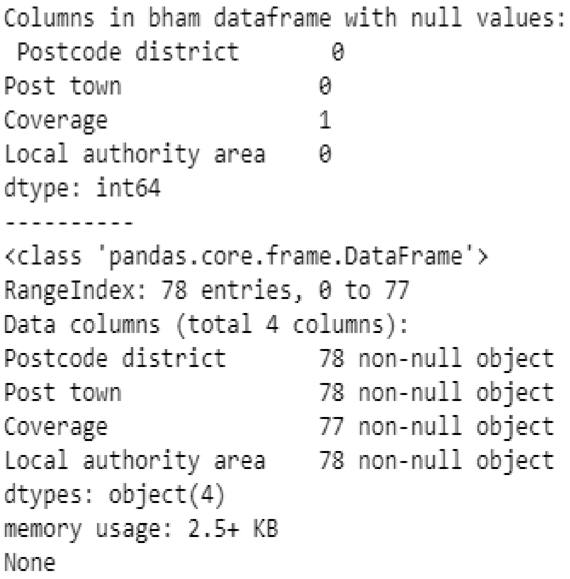
Part 3: Methodology.

The preliminary task for the analysis was to acquire the location data. We utilise the BeautifulSoup library to parse HTML documents of the target website that contained the postcode districts. The first URL used was from the B postcode area Wikipedia page. By inspecting the HTML document, we find the section containing the table and use panda’s read_html method to convert it to a dataframe. The initial data had the following fields: *Postcode district, Post town, Coverage, Local authority.*

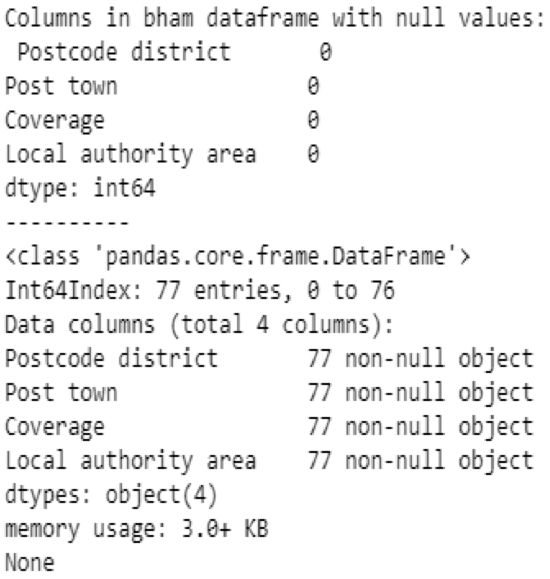


	Postcode district	Post town	Coverage	Local authority area
49	B62	HALESOWEN	Halesowen (east), Romsley, Hunnington, Quinton...	Dudley, Bromsgrove, Birmingham
35	B37	BIRMINGHAM	Chelmsley Wood, Marston Green, Kingshurst, For...	Solihull
46	B50	ALCESTER	Bidford-on-Avon	Stratford-on-Avon
45	B49	ALCESTER	Alcester	Stratford-on-Avon
32	B34	BIRMINGHAM	Shard End, Buckland End	Birmingham
31	B33	BIRMINGHAM	Kitts Green, Stechford	Birmingham
19	B20	BIRMINGHAM	Handsworth Wood, Handsworth, Birchfield	Birmingham

Now we perform some basic EDA and check for null or missing values. The LHS figure below shows that 1 null value exists in column ‘Coverage’. We proceed to clean the dataset and prepare to acquire Latitude & Longitude information on postal districts.

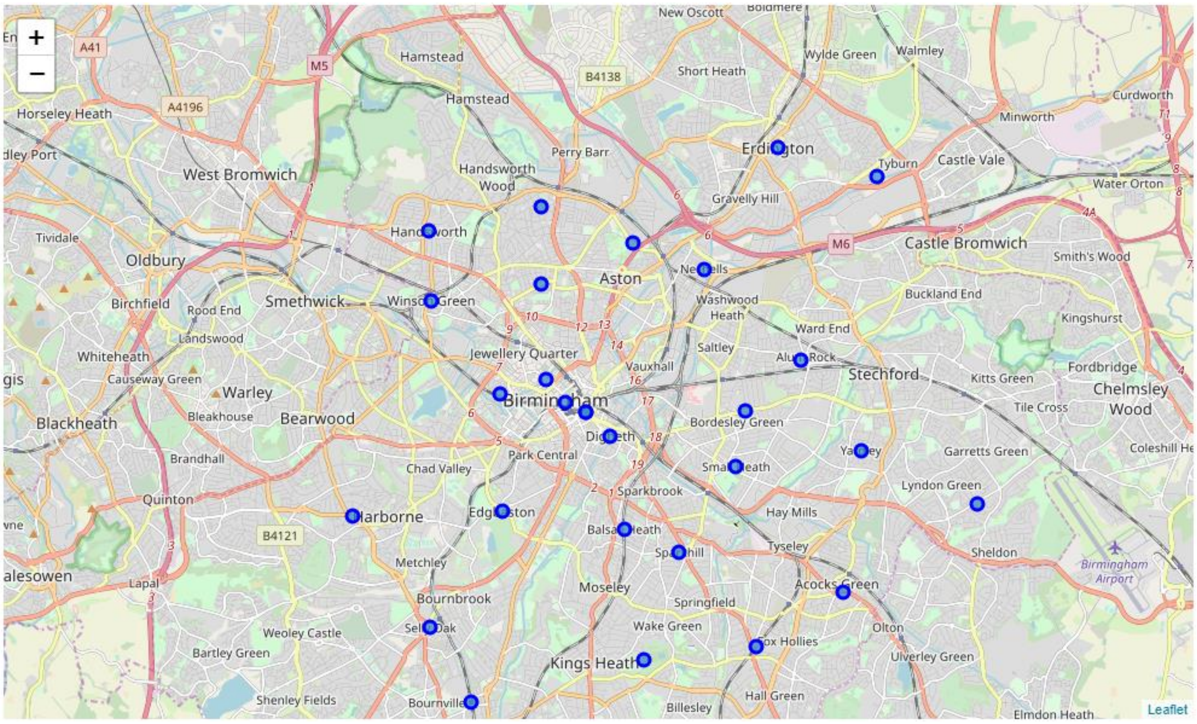


```
Columns in bham dataframe with null values:
 Postcode district      0
Post town              0
Coverage               1
Local authority area   0
dtype: int64
----------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78 entries, 0 to 77
Data columns (total 4 columns):
Postcode district      78 non-null object
Post town              78 non-null object
Coverage               77 non-null object
Local authority area   78 non-null object
dtypes: object(4)
memory usage: 2.5+ KB
None
```



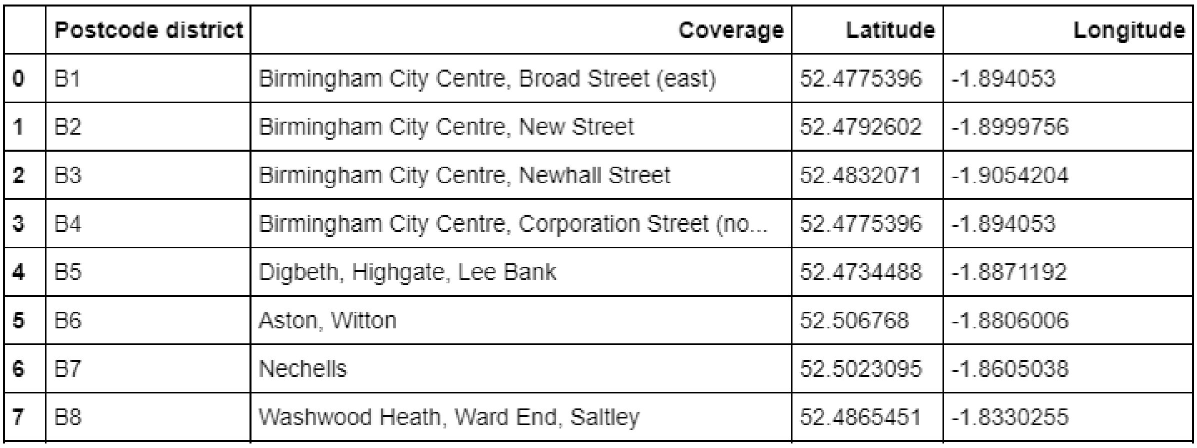
```
Columns in bham dataframe with null values:
 Postcode district      0
Post town              0
Coverage               0
Local authority area   0
dtype: int64
----------
<class 'pandas.core.frame.DataFrame'>
Int64Index: 77 entries, 0 to 76
Data columns (total 4 columns):
Postcode district      77 non-null object
Post town              77 non-null object
Coverage               77 non-null object
Local authority area   77 non-null object
dtypes: object(4)
memory usage: 3.0+ KB
None
```

In order to get coordinates for our postcode areas we create a recursive function to handle exceptions when running geocoder. This method itself can fail depending on how busy Nominatim service is. However, it is recommended to cache results or export it as a csv file going forward. Final touches are made to our postcode dataframe adding a new feature called ‘Address’. It is a column that concatenates coverage and postcode district (separated by comma) so that we may use it for retrieving coordinates by geocoding. After running geocoder through all values of address then we check the output with a folium map plot.



The data scrape procedure for London was similar, with the exception that each postal district table was situated in different Wikipedia pages. A simple method to loop through all URL's and append all data together yielded a dataframe like the initial Birmingham table.

The table below was the output from using geopy and removing some redundant columns from the dataframe. It contains a subset of both city's postcodes with latitude and longitude coordinates.



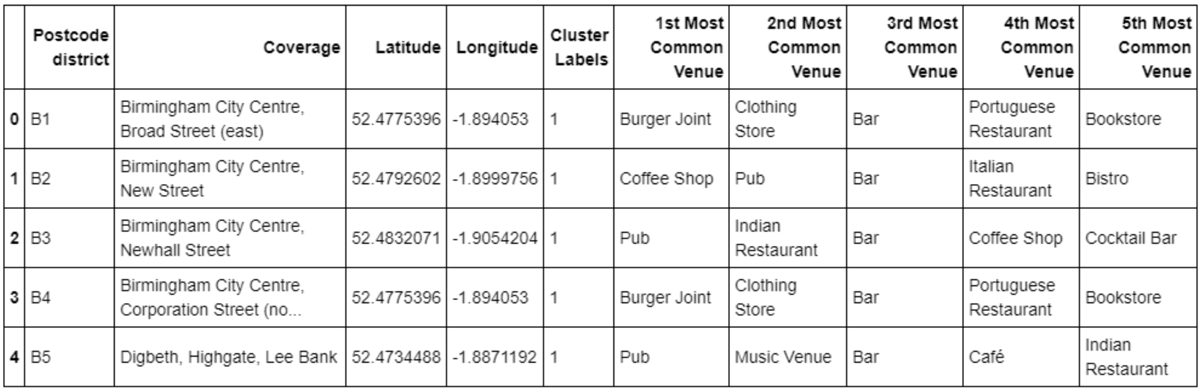
	Postcode district	Coverage	Latitude	Longitude
0	B1	Birmingham City Centre, Broad Street (east)	52.4775396	-1.894053
1	B2	Birmingham City Centre, New Street	52.4792602	-1.8999756
2	B3	Birmingham City Centre, Newhall Street	52.4832071	-1.9054204
3	B4	Birmingham City Centre, Corporation Street (no...	52.4775396	-1.894053
4	B5	Digbeth, Highgate, Lee Bank	52.4734488	-1.8871192
5	B6	Aston, Witton	52.506768	-1.8806006
6	B7	Nechells	52.5023095	-1.8605038
7	B8	Washwood Heath, Ward End, Saltley	52.4865451	-1.8330255

Foursquare API was then utilised to pair each location of interest with information on venues. The nearby venue information contained each venue name, its venue category, venue latitude, venue longitude. The constraints placed on the search are a 1km radius from a location with a limit of 50 venues.

We will need to do some exploratory data analysis and to proceed with preparing our input for clustering. In this project we have allowed all types of categories so that the stakeholder can consider the venue landscape on a location and decide if certain amenities could be great to be near to. The dataframe we have will turn into a table containing fractions of existing venue categories in each post code. This is done alongside some EDA.

The last part of analysis will be using the sklearn K Means algorithm to segment postcodes into 6 clusters. The results are then visualised, and we discuss the results of each cluster. Each cluster should be able to offer a distinction from each other and I will then recommend areas to consider.

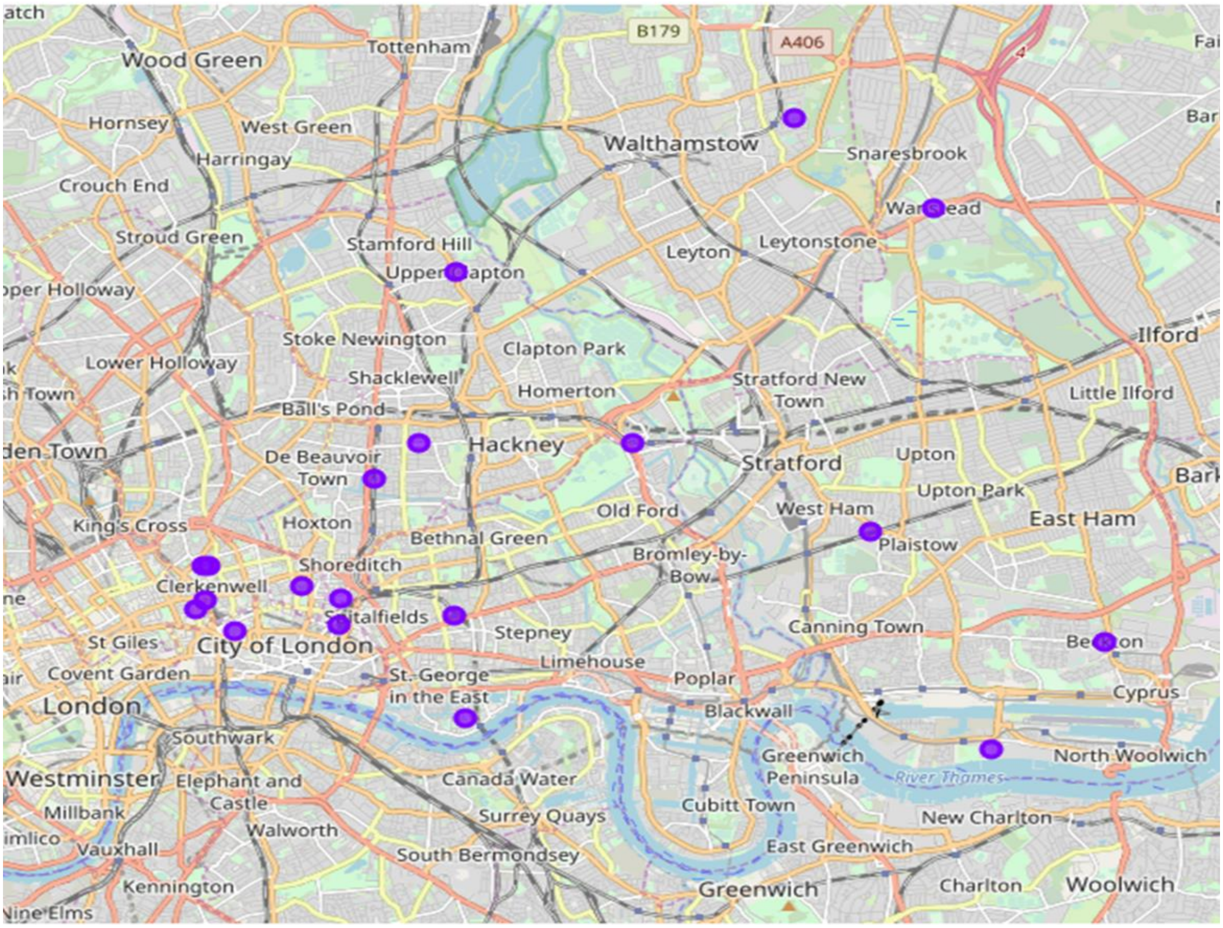
Part 4: Results and Discussion

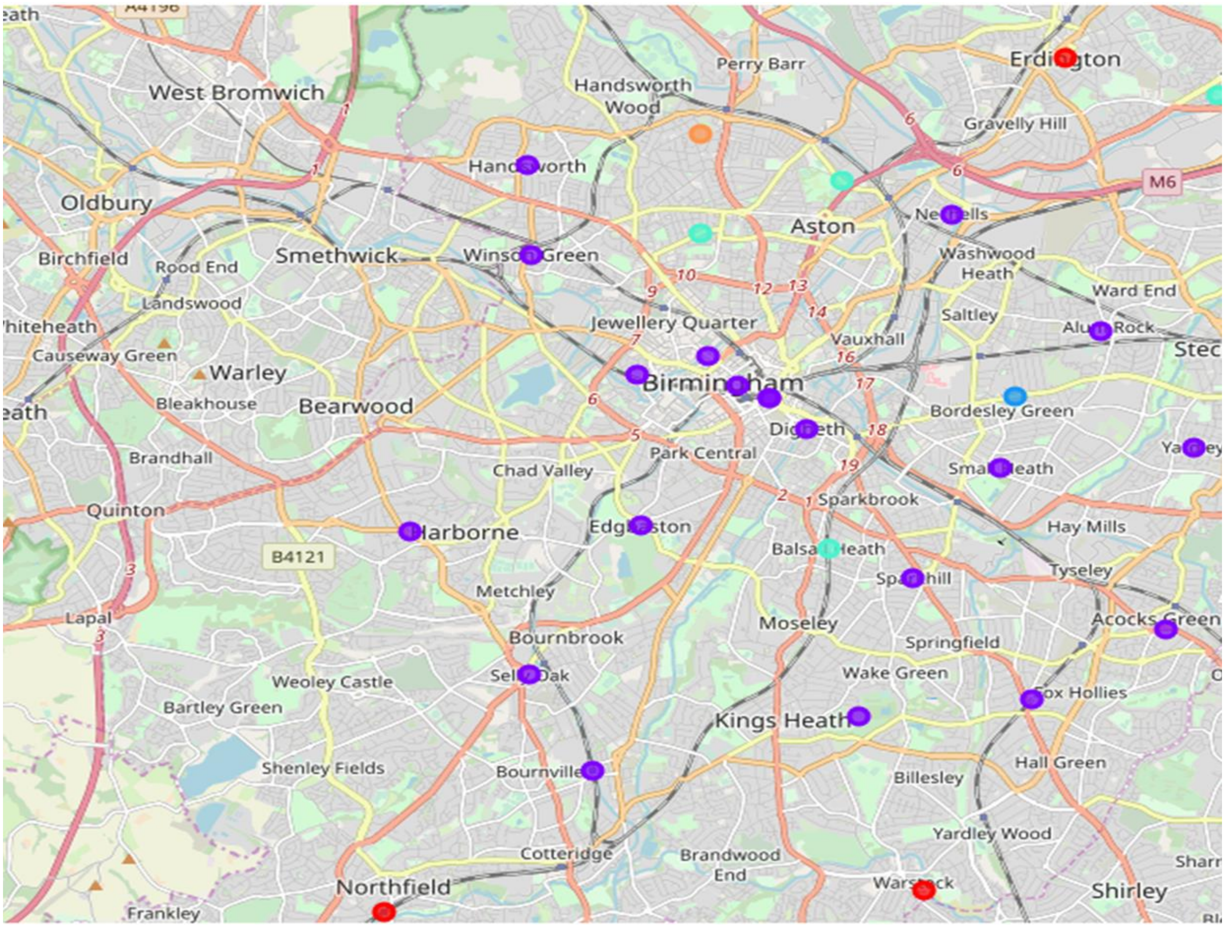


	Postcode district	Coverage	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	B1	Birmingham City Centre, Broad Street (east)	52.4775396	-1.894053	1	Burger Joint	Clothing Store	Bar	Portuguese Restaurant	Bookstore
1	B2	Birmingham City Centre, New Street	52.4792602	-1.8999756	1	Coffee Shop	Pub	Bar	Italian Restaurant	Bistro
2	B3	Birmingham City Centre, Newhall Street	52.4832071	-1.9054204	1	Pub	Indian Restaurant	Bar	Coffee Shop	Cocktail Bar
3	B4	Birmingham City Centre, Corporation Street (no...	52.4775396	-1.894053	1	Burger Joint	Clothing Store	Bar	Portuguese Restaurant	Bookstore
4	B5	Digbeth, Highgate, Lee Bank	52.4734488	-1.8871192	1	Pub	Music Venue	Bar	Café	Indian Restaurant

The table above contains the merged table of cluster labels for each postcode district.

On the page below are the folium maps marked with assigned coloured markers that denote each cluster generated.


Wood Green
Tottenham
B179
A406
Hornsey
West Green
Walthamstow
Snaresbrook
Harringay
Crouch End
Wanstead
Stroud Green
Stamford Hill
Leyton
Leytonstone
Upper Clapton
Upper Holloway
Stoke Newington
Clapton Park
Ilford
Lower Holloway
Shacklewell
Stratford New Town
Little Ilford
Ball's Pond
Homerton
Camden Town
Hackney
De Beauvoir Town
Stratford
Upton
Upton Park
King's Cross
Hoxton
Old Ford
West Ham
East Ham
Bethnal Green
Plaistow
Shoreditch
Bromley-by-Bow
Clerkenwell
Spitalfields
Stepney
Canning Town
St Giles
City of London
Beckton
Covent Garden
Limehouse
St. George in the East
Poplar
London
Cyprus
Blackwall
Southwark
Greenwich Peninsula
North Woolwich
Westminster
Elephant and Castle
Canada Water
River Thames
Millbank
Cubitt Town
New Charlton
Surrey Quays
Walworth
Vauxhall
East Greenwich
South Bermondsey
Kennington
Greenwich
Charlton
Woolwich
Nine Elms


A4196
Perry Barr
Erdington
West Bromwich
Handsworth Wood
Gravelly Hill
Handsworth
M6
Oldbury
Aston
Nechells
Smethwick
Winson Green
Birchfield
Rood End
Washwood Heath
Landswood
Ward End
Jewellery Quarter
Saltley
Whiteheath
Vauxhall
Alum Rock
Causeway Green
Warley
Birmingham
Bearwood
Bordesley Green
Bleakhouse
Digbeth
Brandhall
Park Central
Small Heath
Yardley
Chad Valley
Sparkbrook
Quinton
Hay Mills
Harborne
Edgbaston
B4121
Balsall Heath
Tyseley
Sparkhill
Metchley
Lapal
Moseley
Acocks Green
Bournbrook
Springfield
Weoley Castle
Selly Oak
Wake Green
Fox Hollies
Bartley Green
Kings Heath
Hall Green
Shenley Fields
Bournville
Billesley
Yardley Wood
Cotteridge
Brandwood End
Northfield
Warstock
Shirley
Frankley

We then label each of the clusters as follows:

- Cluster 0: "Groceries Store" dominant category with odd few venues not located close to Birmingham town centre. Significantly <50 venues based on search criteria used. **Count = 3**.
- Cluster 1: Typical high streets, town centres with bars and variety of restaurants too. This cluster was the only cluster to return at most 50 venues for a postcode. **Count = 40.**
- Cluster 2 (and 5): Outliers. **Count = 1** for each label.
- Cluster 3: 'Indian Restaurant' dominant. Elements of football/soccer. **Count = 4**
- Cluster 4: Parks & Natural reserves with nearby Pubs. Significantly <50 venues per postcode. **Count = 2**

The analysis segments each postcode to 6 clusters that try to distinguish them apart. This showed 40 out of 51 postcodes used were likely to be shopping/leisure destinations residing somewhat near the centre of their respective cities. This was cluster 2 and this would contain the optimal location to consider for a stakeholder growing their Indian restaurant business just from the variety of amenities. This all points to the fact that these locations are venue dense and could help to consider which parts in Birmingham. Interesting picks from cluster 1:

- B10, B11, B28 are areas with several Indian restaurants. This insight could be followed up with a deep dive to see if you would benefit from locating near rivals. This idea is not recommended without further analysis.
- B4, B16 are at an excellent location with hardly any Indian restaurants to compete with at the nearest vicinity.

Cluster 3 is interesting because they're not areas that are 'venue dense' based on our used criteria. However, there are Indian restaurants featured quite heavily in the data of this cluster. For instance, we have B12 with 5 Indian restaurants within 1km radius (~36% of all venues returned). This suggests that the areas are likely to be saturated with the % of Indian restaurants present within the area. But there is a small chance that there's room for more and to really establish certain postcodes as a hub for Indian restaurants. I would imagine cluster 4 to not cost too much (assuming from the lack of nearby amenities) and to therefore operate on a smaller budget. But this requires more of a deep dive and knowing more about stakeholder constraints.

Part 5: Conclusion

Purpose of this project was to identify B postcode areas close to centre with either hardly any Indian restaurants or with direct competitors in order to aid stakeholders in narrowing down the search for optimal location for a new branch from London. By calculating venue density distribution from Foursquare data, we can see types of venues that are of interest to stakeholders and rank the top 5. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) for final exploration by stakeholders.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighbourhood etc.