Coursera Capstone – IBM Applied Data Science – Initial Plan - Omar Imran – July 2020

Project title: Best neighbourhood to open an Indian restaurant in Leeds

Description of problem and background

A relative wants to open an Indian restaurant in Leeds, England. More specifically a South Indian restaurant. The relative believes there to be a niche in the market for this particular cuisine in Leeds. Leeds is a diverse city with restaurants of many cuisines, but perhaps a gap in the market for South Indian. The relative is also not very familiar with all the neighbourhoods of Leeds.

Thus, the relative wants to understand which neighbourhoods of Leeds are most appropriate for setting up a South Indian restaurant. This may be done in several ways, but the relative would like to see patterns in the types of competing restaurants in each neighbourhood area. The target audience is my relative or more broadly entrepreneurs seeking to develop a restaurant in Leeds in the most appropriate place for maximised profits.

Description of data and how it will be used to solve problem

There is an existing online wikipedia article which lists different places in Leeds as well as their postcode (https://en.wikipedia.org/wiki/List_of_places_in_Leeds). The data from this will be accessed and manipulated using BeautifulSoup (python package). Geocoder (python library) will be used to determine the geographical coordinates of these places in Leeds. Foursquare API will be used to get the data of the venues at each of these places. Notable venue types of interest will be the different types of restaurants, particularly (South) Indian restaurants.

Machine learning in the form of k-means clustering combined with visualisation using folium will allow the analysis of the data. Thus, we will be able to draw conclusions to address the problem statement of this project.

Methodology

The methodology can be split into the below steps:

1. Importing libraries – numpy for data handling, pandas for dataframe work, geopy for geocoding locations, matplotlib for data visualisations, folium for map visualisation and finally sklearn for machine learning work.
2. Scraping Leeds neighbourhoods from Wikipedia – using beautifulsoup to parse data from the Wikipedia article mentioned in Data. Only the relevant columns from the article were appended and stored in a dataframe – i.e. PlaceName and PostTown. PostTown is needed as some places, while technically considered Leeds, are in practice too far from Leeds centre to be considered for the client. So PostTown was used to filter only the PlaceNames registered to a Leeds postcode. After this point only PlaceNames was kept. There are 109 places at this point.
3. Geocoding the places in Leeds – the places are geocoded one by one and corresponding latitude and longitude are stored in the dataframe. This dataframe is saved for reference.
4. Superimpose places on map – having geolocated Leeds and centred a folium map at these coordinates, markers are placed for each place in the primary Leeds dataframe. This gives a good visualisation that most places are centred near Leeds centre and become sparser in density towards outer Leeds. This map is saved for reference.
5. Explore venues with Foursquare – having setup a client account on Foursquare API we are able to call for the venue data for each of the places. Initially we called for the top 100 places within 500m radius from each place. But after some trial and error, and contextual understanding of the Leeds area. It was determined to settle with 2km radius. This leads to 4599 venues, 192 of which are of a unique type. There is a venue for each place (this did not occur when looking at only a 500m radius). From venue analysis there are plenty of different unique cuisines of restaurants.
6. **Clustering analysis – by one hot encoding our data and filtering out venues which have the word 'Restaurant' in them we arrive at an updated dataframe. We also group venues by place so we return to a 109 row dataframe. We also set up a dataframe which takes the sum of Indian restaurants and total restaurants per neighbourhood. This will be useful for presentation purposes later on. For the same**

presentation purpose we also set up a different dataframe to list the most common venue per neighbourhood. We assign 5 clusters in the clustering analysis and once this analysis is performed using KMeans we display the results. Results are displayed by merging all the aforementioned tables to get a nice table which list the coordinates, cluster label, sum of Indian restaurant, sum off all restaurants and most common cuisine for each place! A map is output using folium, colour coded for each cluster and saved for reference. The data is segmented to understand more about each cluster

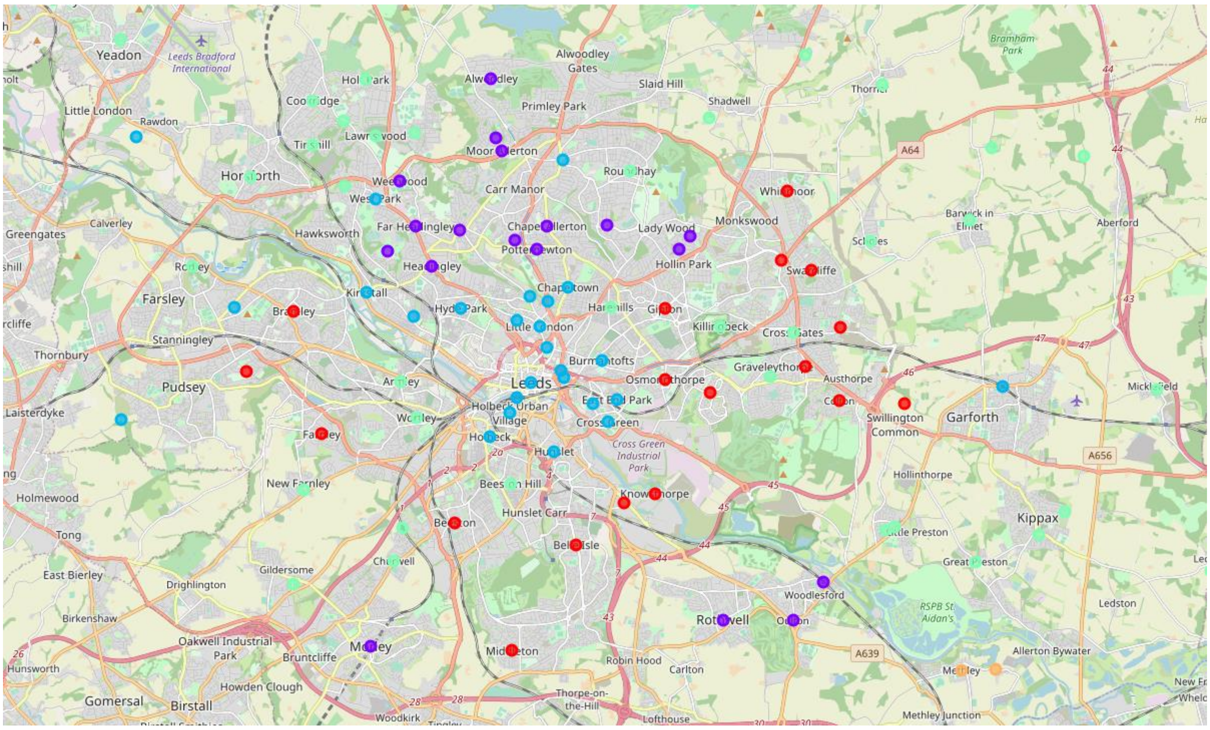
7. Clustering analysis repeated for custom created feature – whereas before machine learning trained only on the different types of restaurant, we create a new feature which sums all the restaurants together and include this in the training. This is in the hope to account for total market competition for each cluster. The same visualisation of a folium-based map and cluster segmentation is performed, and observations are made.
8. Clustering analysis repeated by normalising all features – the previous step lead to overweighting by the new custom feature. Attempts to address this are made by normalising data. Observations are then made again upon mapping and cluster segmentation of data.

Results

There are three separate clustering analyses done in this project. Cluster maps and initial observations are separated into each analysis.

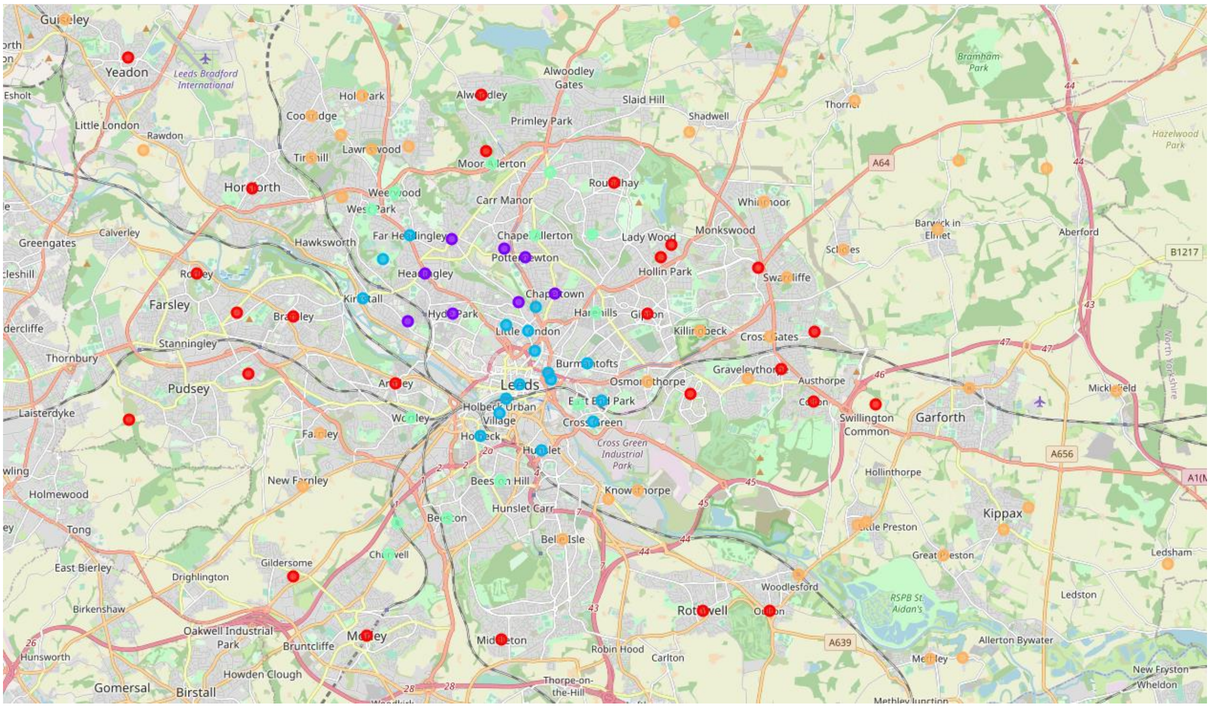
1. Cluster analysis #1

Cluster 0 seems fast food dominant, Cluster 1 seems Italian restaurant dominant, Cluster 2 seems Thai/Indian dominant, Cluster 3 seems to be Vietnamese dominant which in this case due to ordering convention means there is a lack of restaurants in this area, Cluster 4 have only one unspecified cuisine restaurant.



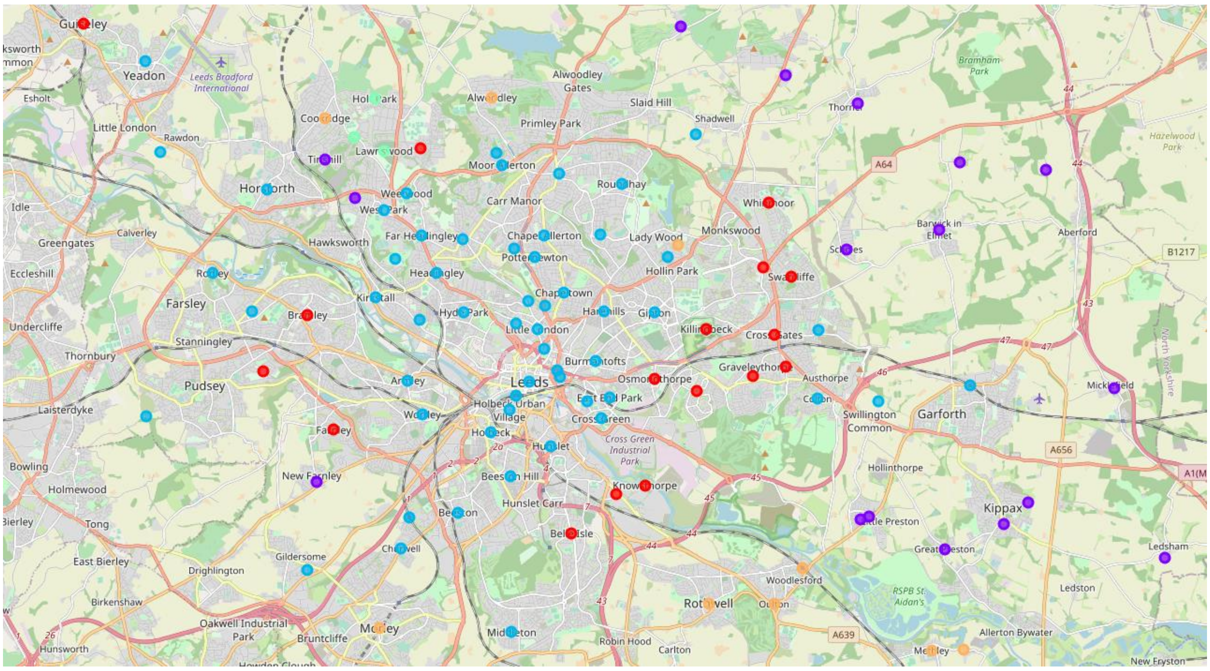
2. Cluster Analysis #2

Cluster 0 has 3-7 total restaurants in each place of varying cuisine, cluster 1 has 18-21 restaurants in each place of mainly Thai/Italian cuisine, cluster 2 has 12-18 restaurants in each place of Thai/Indian cuisine, cluster 3 has 8-13 restaurants of varying cuisine and cluster 4 has mainly Jewish restaurants which by the numbering convention actually means it has 0-1 total restaurants per place.



3. <u>Cluster Analysis #3</u>

Cluster 0 has 1-3 restaurants per place, primarily fast food. Cluster 1 has 0 restaurants per place. Cluster 2 has variety of types of cuisines and is hard to define as a cluster. Cluster 3 has only one Italian restaurant per place. Cluster 4 is dominated by unspecified 'restaurant' type.



Discussion

Firstly, there is not enough data on restaurants. Even when increasing the venue reach to 2km (quite a large allowance for a town centre), some places only have 3 or 4 restaurants. The effect of this means that some clusters only have 1 or 2 restaurants, so it is hard to derive strong insights. The radius could have increased but increasing it further would lead to the overlap of some neighbourhoods, especially near the centre.

From the first analysis - Cluster 2 seems of most interest for an Indian restaurant owner, and cluster 2 is based centrally on the map. Thus, there may be a lot of competition in central Leeds for an Indian restaurant and perhaps more of a market niche in other areas (particularly cluster 3 as there are no restaurants here. On the other hand, this could mean that there is only a taste for Indian food in central Leeds.

From the second analysis - the overweighting of the total number of restaurants category in clustering analysis (due to high unnormalized value range) means that the clusters mainly depend on the new feature only. The bulk of Indian dominated places lie in cluster 1 and 2. These areas lie in centre Leeds. Which again reiterate the conclusions drawn from the first analysis.
From the third analysis - the clusters didn't give much more useful insight into Indian Restaurants as such. Thus, the new feature didn’t seem to contribute much to useful analysis. It was rather expected that there would be more restaurant in the centre of Leeds. The new feature didn’t seem to highlight how Indian restaurants varied in high population areas, which was the initial aim.

Conclusion

If more restaurants existed within a 2km radius of the centres of the different neighbourhoods in Leeds, then clustering could be more insightful. What is clear is that there is a lack of restaurants in many areas in outer Leeds or perhaps just a low number of restaurants registering their location close to the town centre. So, there is indeed **a market gap in outer-Leeds town centres**. Also, the dominant restaurant cuisine in central Leeds seems to be Thai food, followed by Indian. Thus, there is **a taste for India in central Leeds which could be profited upon**.

In summary, I would advise to open a restaurant just outside the centre of Leeds, to capture the market interest in Indian food, but to stay further away from dense market competition.