

IBM Capstone Project

Introduction

As part of the Coursera Data Science Capstone project, I will be using Python and Coursera to analyse the number of Italian restaurants by postcode in the Greater London area. This is to aid in the understanding of the relative density of restaurants to then inform entrepreneurs as to the area with the greatest opportunity for new restaurants.

London has one of the greatest plethora of restaurants in the world, with hundreds of new restaurants opening every month in the capital, however being a highly competitive market, there are also hundreds of restaurants that close every month. Understanding the "right" location to launch a new restaurant is an important decision for any restaurateur. This piece of analysis will be focused on Italian restaurants, to limit the immediate scope and to test the functionality of the tool, however the analysis could be broadened at a late time to cater to a users desired input cuisine.

Data

Throughout this analysis I will be using Python and a range of libraries which have been pre-install or added. The below summarises the required tools:

```
numpy, pandas, ..., imported...
geopy installed...
Nominatim imported...
requests imported...
json_normalize imported...
matplotlib imported...
Kmeans imported...
folium installed...
folium imported...
...Done
```

London data

To filter the data from Foursquare for London, I will need to identify the postcode areas associated with London. I will use Wikipedia for this. The attached [webpage](#) is where I have pulled the information. However this will need cleansing, which is why I installed BeautifulSoup above (a HTML parsing and editing library).

	Location
0	Abbey Wood
1	Acton
2	Aldgate
3	Aldwych
4	Anerley

Location data

This analysis will require data to be accessed from Foursquare, a location data provider. This includes information about the number, location, ratings and many more details regarding venues across the world. I will use this information to extract my data, for which I will be focusing on the data associated with London and Italian restaurants (API ID:

4bf58dd8d48988d110941735). This will enable me to gauge the number of Italian restaurants in the city but then associate it with given postcodes.

The below code looks up on Foursquare the Italian restaurants that are within 1000m of the location, then counts the number of restaurants. If the area is not recognised by Foursquare, then Foursquare will return an error, and the if statement allows for the process to continue and return a 0 value for the number of Italian restaurants.

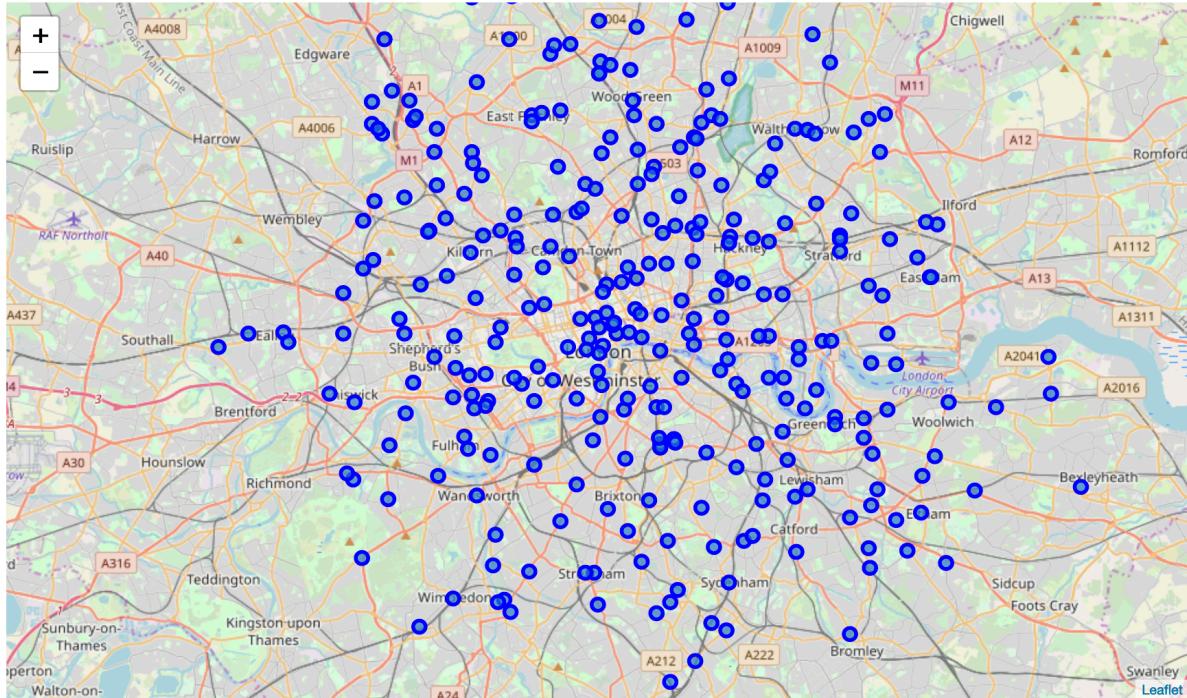
	Location	Count_Italiens
0	Abbey Wood	0
1	Acton	3
2	Aldgate	50
3	Aldwych	50
4	Anerley	0

In order to make sense of the data and start to visualise it, we will need to create latitude and longitude co-ordinates for each location. The below function and code enable me to create the co-ordinates and append them to the dataframe.

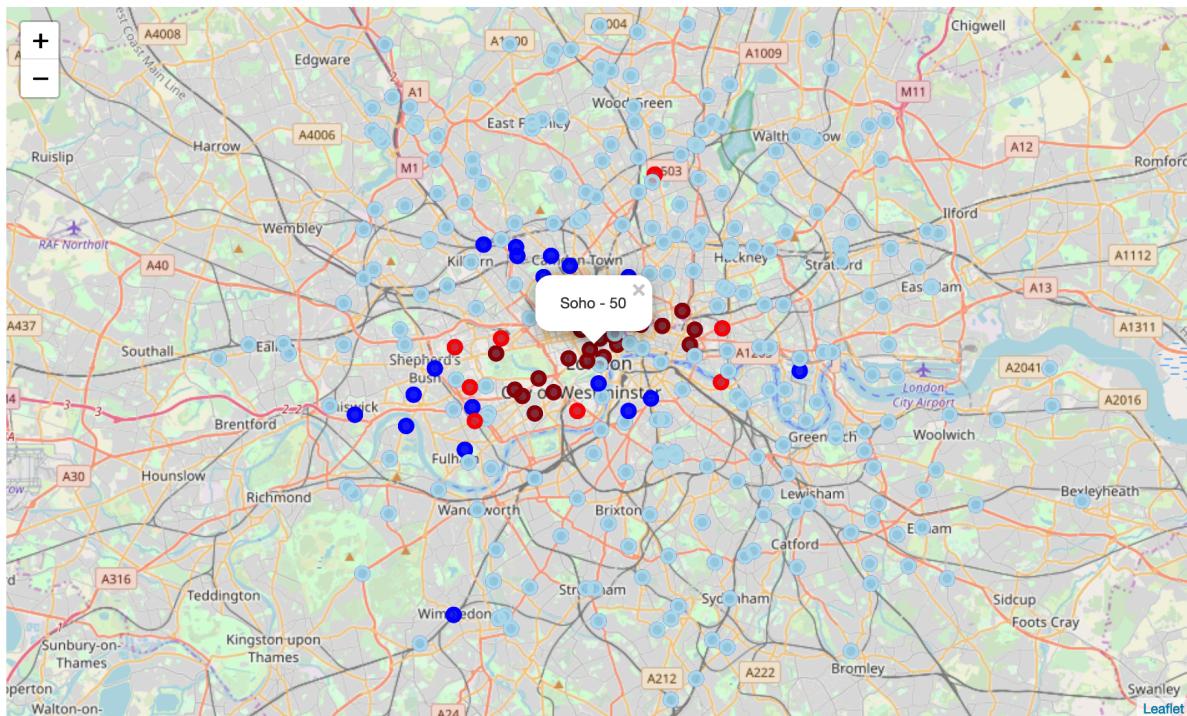
	Location	Count_Italiens	Latitude	Longitude
0	Abbey Wood	0	51.492450	0.121270
1	Acton	3	51.513240	-0.267460
2	Aldgate	50	51.513304	-0.077771
3	Aldwych	50	51.513291	-0.117093
4	Anerley	0	51.412330	-0.065390

Visualising the data

Now that I've got a dataset with the locations and the total number of restaurants, I can start to visualise where the high density areas may be located. I use the Folium module to create a map of London, using the geocoder to centre the map on the city.



Having visualised the various locations across London, it is not immediately obvious where the high and low density areas are using a single colour scheme - you need to click on the icons to find out how many Italian restaurants there are. The next piece of code then uses the pd.cut method to bucket the locations into 4 groups based on their count of restaurants, with the low ones being coloured light blue, and the highest density areas in dark red.



Results

As you can see from the map above, it is clear that the highest density areas are in Central London, around Mayfair, Chelsea, Covent Garden, Soho, Aldgate, Farringdon and a number of other areas. However there are other areas also in Central London which have a low density, e.g. Temple, Blackfriars.

Discussion

Based on the trends shown in the map above, the opportunity for a new restaurant would be best placed in Central London, but away from the high density areas already. This would look to minimise competition, but enable access to the high density footfall. This could be in Marylebone or Westminster areas, as these are both large tourist hubs with people looking to find quality restaurants.

There are limitations to the analysis, with some geocoding not accurately identifying the correct locations, or FourSquare no able to search for some locations - potentially causing 0 values incorrectly - this would need to be worked through, and further datasets brought in to corroborate the outcomes.

Conclusion

In conclusion there are clear trends that Central London areas have higher densities of restaurants - but there are areas which have lower comparative density to the neighbours. This would pose a potential opportunity for aspiring Italian restauranteurs as you can understand the demand levels from the sustainability of a large number of Italian restaurants.