IBM Applied Data Science Capstone

Introduction:

Business Problem
Here we are discovering the areas around and within UK neighborhood where there are less or no Asian Restaurant present, and hence opening an Asian Restaurant can be very profitable due to the lack of competitions. Using data science methodology we will identify areas of recommendation.

Discussion of the background
With times, Asian food has become more and more popular across UK.
As Asian population increasing continuously and at the same time, other people are showing interests in Asian cuisine, the market is fit for more varieties in Asian food.
So this project aims to identify areas where this hypothesis can be applied by opening new Asian restaurant.

Data:

To solve the problem, we will need the following data:

- List of Area Code of UK
- Latitude and longitude coordinates of those neighbourhood.
- Venue data, particularly data related to Asian Restaurant presence.

Methodology:

List of Area Code of UK are obtained from below link:

https://en.wikipedia.org/wiki/List_of_postcode_areas_in_the_United_Kingdom'

Then they are parsed using BeautifulSoup.

Then latitude and longitude are extracted using Python Geocoder package.

After that, we will use Foursquare API to get the venue data for those areas.

We then make API calls to Foursquare passing in the geographical coordinates of the areas.

Foursquare will return the venue data in JSON format and we will extract venue latitude and longitude.

With the data, we can check how many venues were returned for each areas and how many areas have asian restaurant already present.

Here is how the data will look like:

	Postcode_area	Postcode_area_name	Latitude	Longitude
0	AB	Aberdeen	57.14548	-2.10272
1	AL	St Albans	51.75360	-0.33730
2	B	Birmingham	52.47891	-1.90592
3	BA	Bath	51.38503	-2.36132
4	BB	Blackburn	53.74859	-2.47992

then, we will perform clustering on the data by using k-means clustering.

K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

We will cluster the areas into 3 clusters based on their frequency of occurrence for Asian restaurant

This is how the clusters data look:

	Postcode_area_name	Asian Restaurant
0	Aberdeen	2
1	Bath	1
2	Belfast	3
3	Birmingham	0
4	Blackburn	0

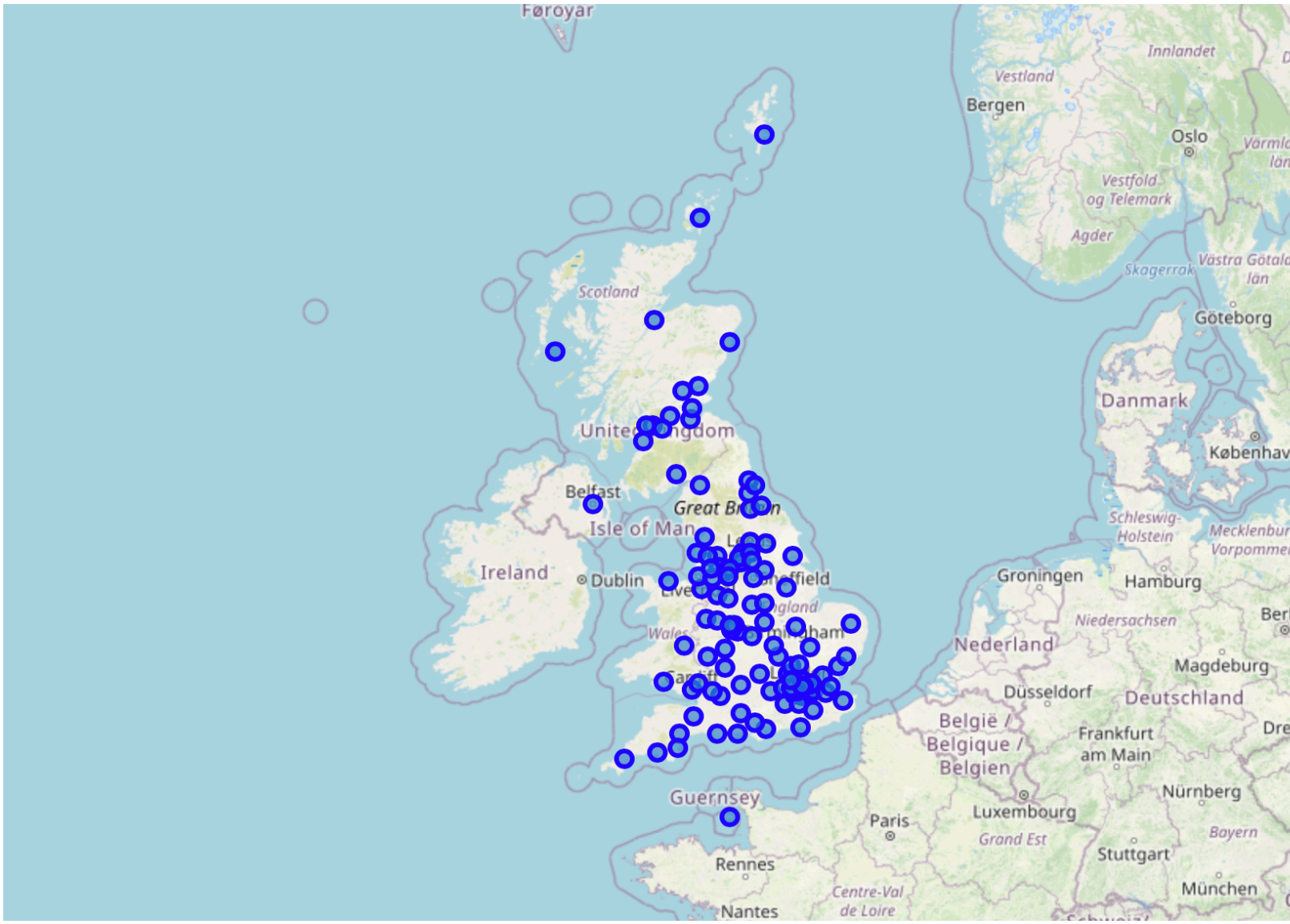
	Postcode_area	Postcode_area_name	Code_formation	Latitude	Longitude	Cluster Labels	Asian Restaurant
0	AB	Aberdeen		57.14548	-2.10272	1.0	2.0
1	AL	St Albans		51.75360	-0.33730	2.0	1.0
2	B	Birmingham		52.47891	-1.90592	0.0	0.0
3	BA	Bath		51.38503	-2.36132	2.0	1.0
4	BB	Blackburn		53.74859	-2.47992	0.0	0.0

The results will allow us to identify which areas have higher Asian restaurant, which have no Asian restaurant and which areas have very few restaurant.

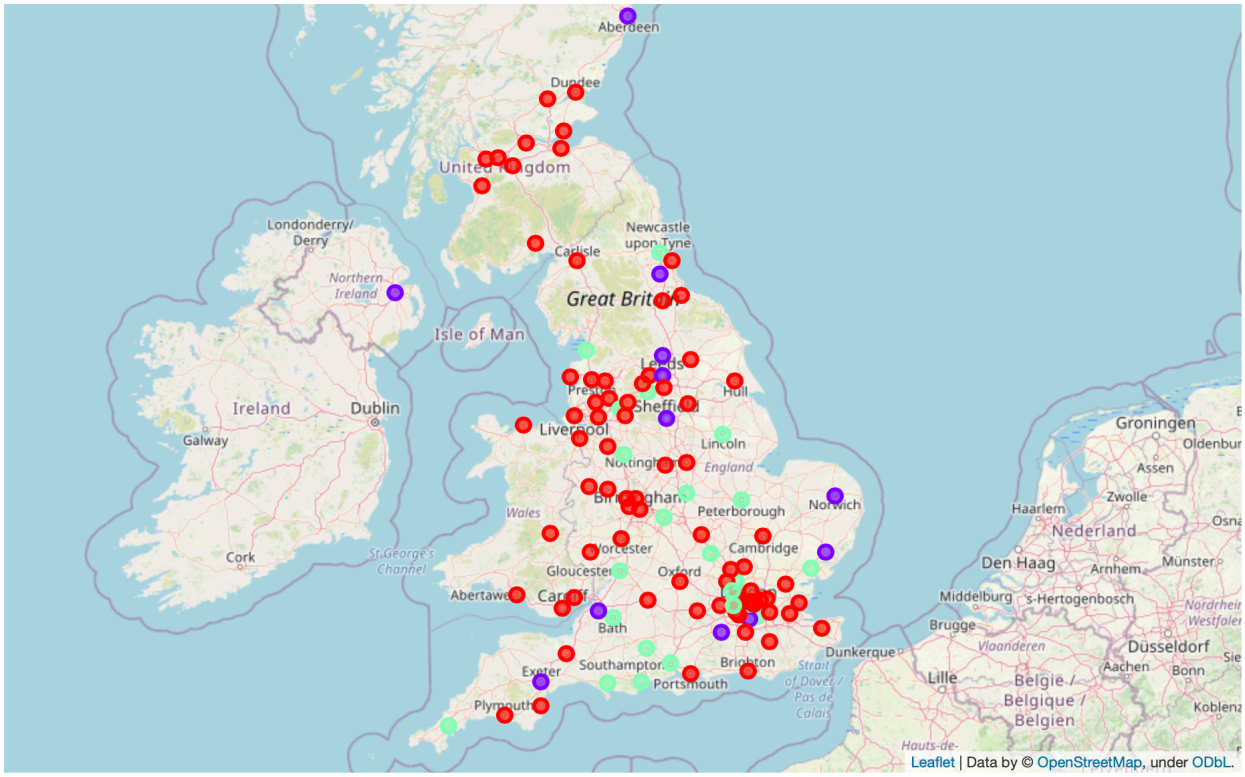
Areas with both less and no asian restaurant will be the target of this study.

Results:

Before clustering this is how the areas look like:



After clustering this is how they are distributed:



Cluster 1 : 83 areas with no Asian restaurant
Cluster 2 : 12 areas with 23 Asian restaurant
Cluster 3 : 25 areas with 26 Asian restaurant

Discussion:

We identified cluster 1 and 2 where there are no/very less number of Asian restaurant and hence is suitable for an investment for a new Asian restaurant in the vicinity of those areas.

Recommendations:

Cluster 3 areas can be left as they are already having multiple Asian restaurant per area.

Conclusions:

Now that we have identified the areas that we would like to focus on, now another study regarding the demography of the regions will help us to further narrow down our recommendations.