IBM Data Science – Capstone Project

Opening a new Vietnamese Restaurant in the UK

Introduction

For my capstone project, I am taking the role of a data scientist that is working for a local entrepreneur that would like to open a restaurant in the UK. I will be most focusing on the Midlands as this is where the entrepreneur is based. The reason for a Vietnamese restaurant is the cuisine is a great mix of flavours, however this is not a common cuisine in the UK and more people should try it!

This project therefore focuses on where the most appropriate place to open a Vietnamese restaurant would be and why. It is likely that there are not many Vietnamese restaurants already so using clusters of similar restaurants should help the entrepreneur make their decision. This location will be key to the success of the entrepreneur's business so this will need to be a robust solution

I will be using the techniques learned in the IBM Data Science course to find the appropriate location to set up the restaurant. This will utilise unsupervised learning and Foursquare API calls

Data

I will be using external data for this course and this is referenced below:

- Download of all UK postcodes from https://www.doogal.co.uk/PostcodeDistricts.php. This needs to be cleaned and processed ready for me to use it
- Handily the co-ordinates are given in the above csv file so no need for a geocoder package in Python to find the co-ordinates
- I will require the geocoder package to find the co-ordinates of the midpoints of the maps that I want to create in Folium
- Using the Foursquare API to find venue data local to the Midland where the entrepreneur lives

Methodology

Initially, I downloaded the UK Postcodes file from https://www.doogal.co.uk/PostcodeDistricts.php as a CSV file. This made the data easy to work with an I didn't need to worry about a html parser like Beautiful Soup (however this would have worked just the same). The file that I downloaded contained the following features:

- Postcode – first section of the UK postcode
- Latitude
- Longitude
- Easting
- Northing
- Grid Reference
- Town/Area
- Region
- Postcodes – number of individual postcodes contained in the first section of the UK postcode
- Active postcodes – number of postcodes that are used
- Population – number of people that live in the postcode

- Households – number of households in the postcode
- Nearby districts – postcodes that are close to this one

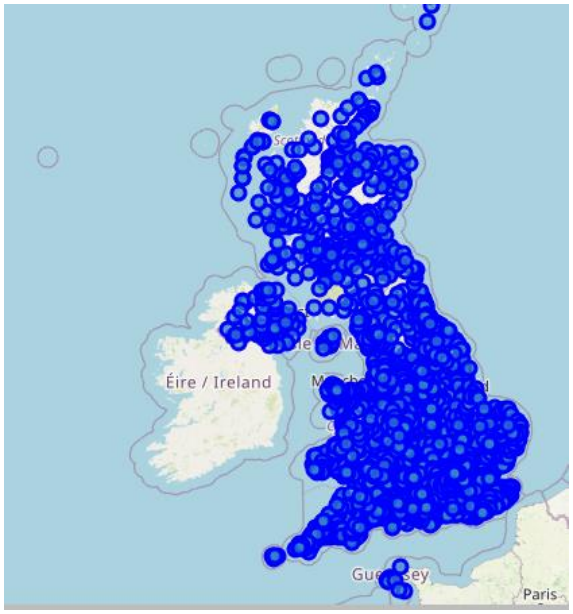
Next, I dropped a few of the column from the CSV file as they were not all necessary to the modelling. The dropped columns were:

- Easting
- Northing
- Grid Reference
- Nearby Districts

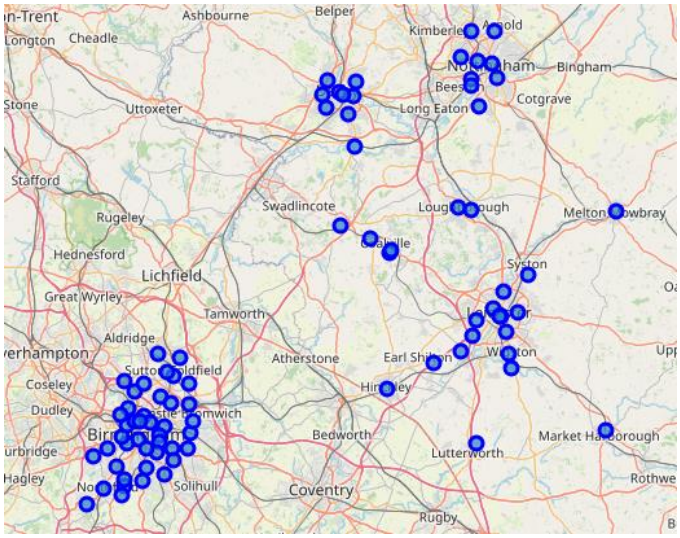
I also needed to change one of the column headings to make sure that it could be used as a variable:

- Changed ‘Town/Area to ‘Area’

I then cleaned the data frame by dropping na values. Once this was complete, I plotted all the postcodes in the data frame on a map using Folium



I then made a new map that would focus on the area that the entrepreneur is based so that I could visualise the area and the locations available to them

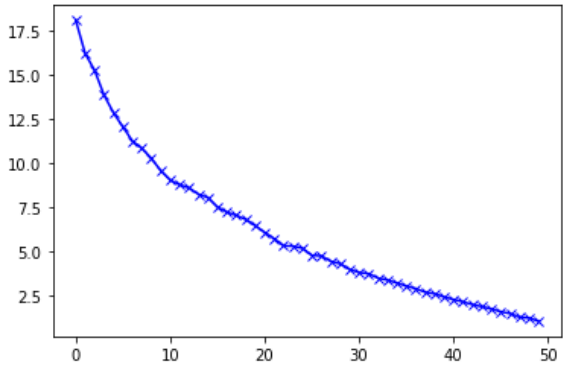


I then used the Foursquare API to pull the nearest 100 venues within 500 meters like the tutorials. This was able to pull in the names, categories and latitude/longitude of the venues into a data frame. I then used one hot encoding to make each unique category into a column to make a matrix and then calculated the mean values to show how frequent the categories were within an area

I checked to see if there were any Vietnamese restaurants on the list and there were 2. I wanted to focus on these restaurants as currently Vietnamese restaurants are very few and far between so making sure that there was the demand was the greater issue. I then performed k-means clustering to group similar locations together by assigning them to clusters. This is a very popular unsupervised machine learning algorithm and is perfectly suited to this task.

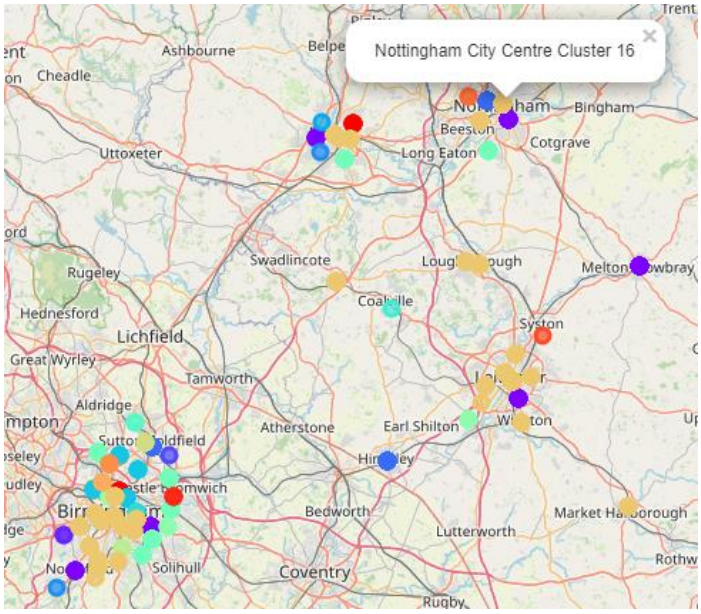
Results

I performed several different k-means values and plotted on the graph below:



Using the elbow method, I found that the optimal k value was 22 so I chose 22 clusters for the analysis which seemed very high. I think this is due to each location being chosen being very different to the others which might have caused a few issues with accuracy.

From the 22 clusters, cluster 16 was the biggest and contained most of the restaurants and contained both Vietnamese restaurants. Cluster 16 was in the town centre of each of the major towns/cities, so this tells me that the local busy town centres will be the important locations rather than smaller towns or on the outskirts. This is represented below in the mustard coloured dots:



Recommendations

I believe that any location within the main town centre of the major towns and cities would really benefit from a Vietnamese restaurant. As shown in the map above, the outskirts of the major towns and cities are different clusters and not where other popular restaurants are based showing this is where the entrepreneur should focus on. The main considerations should then be the other costs relative to the areas they would look to put their restaurant, including the rental cost and the number of people/affluences within the close area.

Limitations

This was a short project and as such there are a few limitations to the analysis:

- I would have liked to expand the search to all locations in the UK to see if there are any specific locations that would be better than others, rather than focusing solely at postcode level
- Within the postcodes dataset there were features about population and number of households and it would be good to expand the analysis to include these features. The success of a restaurant is dependent on the number of customers that come along to the restaurant, tell their friends and family and the repeat visits
- I could change the target to be any Asian based restaurant instead of only Vietnamese, however this assumes that all Asian cuisine is the same or at least similar in taste which isn't necessarily the case

Conclusion

Having performed the analysis for the entrepreneur, I am confident that any location in the Midlands areas would be appropriate, however focusing on the centres of the major towns and cities would be the best way to ensure success as there are then lots of other restaurants around that are busy that will help bring in traffic to their proposed restaurant