# IBM Data Science Professional Certificate Capstone Project – “Where Shall I Live When I Re-Locate?”

Matthew Calder

June 2020

This report provides an overview of the work undertaken to complete the IBM Data Science Professional Certificate Capstone Project. This will comprise an introduction, explanation of the data used, explanation of the methodology used, presentation of results, a discussion around observations drawn from the results, and a conclusion.

## Introduction

The world is more digitally connected than ever, with many people having the ability to work or study from home (or anywhere for that matter) rather than travel to work, an educational institution or any other place. With this being said, the need to attend events in person from time to time (or all the time) is likely to never completely disappear, and as such people are still re-locating frequently.

The question this work attempts to answer is “Where Shall I live When I Re-Locate?”. Of course, this can be answered by doing your research on the city or area you are due to move to, via talking to the locals, reading anecdotal information, or scouring the internet for information. However this approach would likely be incredibly time consuming and laborious if you were to research all the areas you may be interested in, and you may not receive objective opinions from all these data sources. Due to these limitations, this work presents an objective data-driven approach to categorise postcode districts into discrete categories, to provide the user with a first-pass understanding of what an area may have to offer before then delving into researching the area further.

An example use case of this work may be if a user is moving from an area well known to themselves to a new city, and they wish to move to a similar area. The user can check the category their current home is in, and then find any areas in their future city that have the same category. These areas can then be considered further before the user decides where to move to.

Although the approach here could likely be applied to all areas around the world, due to the required complexities of this project the problem has been bounded to consider all Great British postcode districts only. This work would therefore be of interest to anyone moving within Great Britain, or anyone moving to Great Britain to live or visit.

## Data

This project utilises three separate sources of data. A description of these data sources and how they are used is given below:

1. Data – Venue information including location information, venue name, and venue category for venues around Great Britain.

Use – This information is used to compare different postcode districts by comparing the type and frequency of venues within a given postcode district.

Source – This information is pulled from FourSquare via the available API, by requesting venue information for a given longitude/latitude.[1]

1. Data – A list of longitudes/latitudes for all postcode districts in Great Britain.

Use – This is used to 1) input into the FourSquare API requests to receive the desired venue information, and 2) plot all postcode districts on a map to allow for easy visualisation of the final categories associated with their locations.

Source – roblascelles github[2]

1. Data – A geojson file containing shape data for all the postcode areas within Great Britain.

Use – This is to be used to aid data visualisation only. These areas are plotted on a map of Great Britain to aid the users understanding of the geography, but will also be used to split the map into discrete layers that the user can turn on/off depending on if they are interested in that area or not.

Source – Office for National Statistics licensed under the Open Government Licence v.3.0[3] Contains Ordnance Survey data © Crown copyright and database right 2020

Contains Royal Mail data © Royal Mail copyright and database right 2020

## Methodology

Before beginning it was important to note the difference between a postcode area and a postcode district as both are used in the analysis. A postcode district has a letter\_number combination (i.e AB1) and is related to a smaller area than a postcode area.

A postcode area has a letter combination only (i.e AB) and contains several postcode districts within it.

### Data Retrieval and Exploration

First, the postcode shape data was downloaded and explored. The data was not originally in EPSG 4326 format (required when plotting using folium) and so it had to be converted. This was done by reading the file using geopandas, converting to EPSG 4326 format, and then saving the file again. After the conversion, the shape data was plotted on a folium map to ensure the conversion was completed successfully. Finally with the data source, the file was opened with json and a pandas dataframe containing all the postcode areas included was created. This was used later in the process to ensure the location data for postcode districts was consistent with the shape file data used.

Next, the postcode district location data was downloaded, loaded as a pandas dataframe and explored. The data was cleaned (including dropping an unnecessary “id” column, and deleting any erroneous data such as latitude values equal to 0). This dataframe was then wrangled into the format required by the analysis and compared to the postcode area dataframe to identify any discrepancies. It was found at this stage that the postcode area shape data did not include information for Northern Irish postcodes, Jersey postcode, or Isle of Man postcodes, but the postcode district location data did. These postcodes were therefore deleted at this stage and the problem was bounded to Great Britain, rather than the United Kingdom. These location data points were then added to the map as markers to ensure they were being plotted correctly.

### API Interaction

Next, FourSquare API credentials were defined, and a function to request the nearest 100 venues of interest for a circular radius of 1km for a given longitude/latitude was defined. This function was then called for all 2,000+ postcode districts, and the results stored in a data frame.

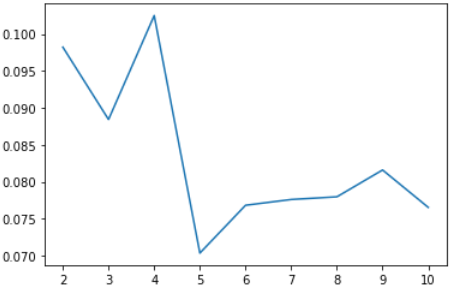
It should be noted here that FourSquare caps requests to 950/day for the most basic account, and so to re-create this step a user would need to submit creditcard details to boost this cap to 99,500 requests/day.

### Clustering

### Before clustering, the dataframe created from FourSquare data had to be wrangled into the correct format. To do this several steps were conducted as follows:

1. Reduce number of categories to ensure a comparison is possible (the number of categories was reduced from >550 to <250 by grouping all like-categories. E.g. grouping all categories with the string “Restaurant” into a single “Restaurant” category.
2. The dataframe was then one-hot encoded to create a new column for each of these categories
3. All of the individual rows were then grouped by postcode district to ensue one ow for each district was used that comprised the top 100 venues related to it

### The Silhouette method was used to determine how many clusters to use. The plot created can be seen below:



From this the optimum number of clusters is 4, as this was the highest silhouette scoring number. The dataframe was therefore clustered into 4 discrete categories.

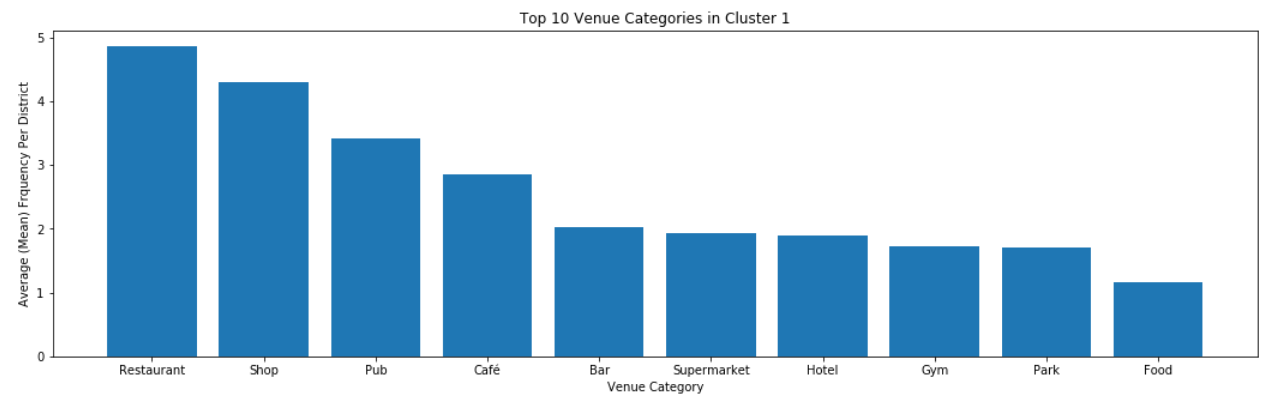
### Cluster Analysis

Finally, the defined clusters were analysed in an attempt to identify a meaningful categorical name for each of them. To do this, the occurrence of categories within each cluster was summed, and divided by the amount of districts that lay within that cluster.; this gave an average amount of a given category for a given postcode district for the top10 categories. These were then plotted I bar charts and viewed to aid in deciding names for the clusters. The average number of venues was also found for each cluster to give an idea of how populated an area was with venues.

## Results

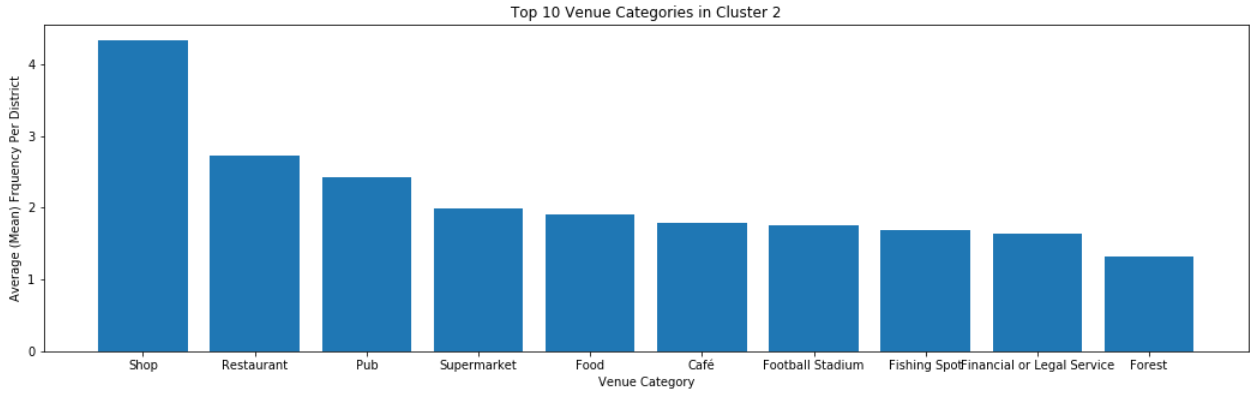
### Cluster Analysis Results

The results from the cluster 1 analysis can be seen in the bar chart below:



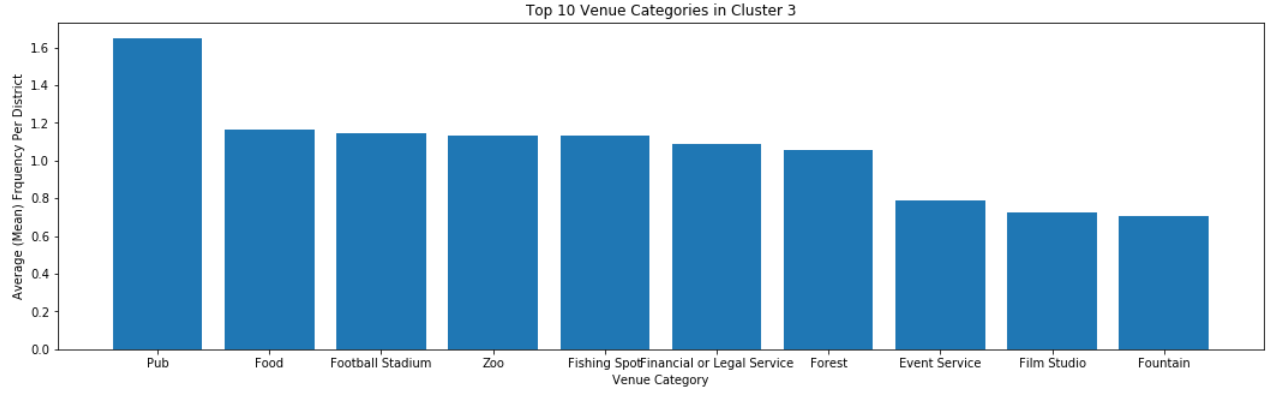
This cluster has the highest density of venues and most relate to eating and shopping. The name given to this cluster was therefore “High No. Eateries and Shopping”

The results from the cluster 2 analysis can be seen in the bar chart below:



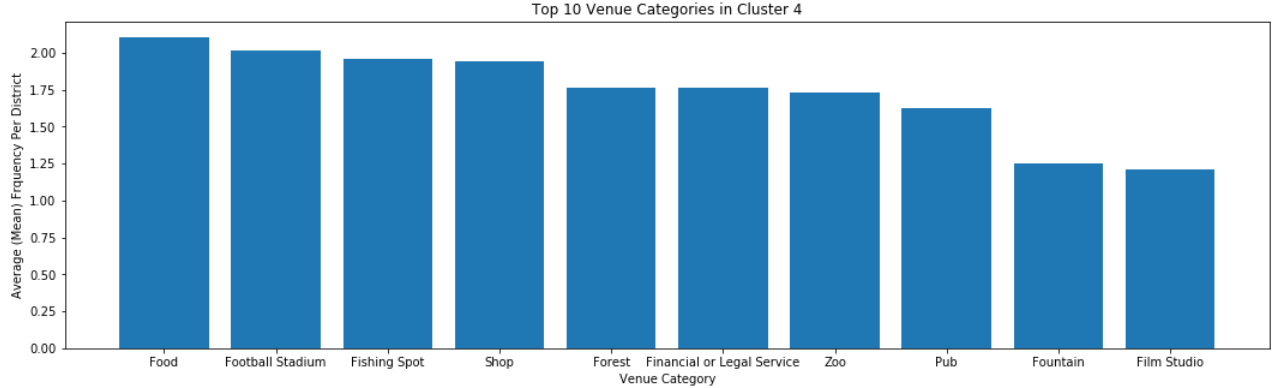
This cluster has a similar makeup to cluster one but with less venues for a given cluster, and more of these relating to shopping than eating. The name given to this cluster was therefore “Medium No. Shopping and Eateries”

The results from the cluster 3 analysis can be seen in the bar chart below:



Locations relating to this cluster returned the least amount of venues, and the majority were pubs. The name given to this cluster was therefore “Least Amount of Venues (Mainly Pubs)”

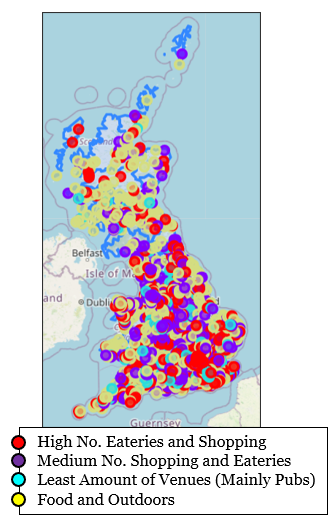
The results from the cluster 4 analysis can be seen in the bar chart below:



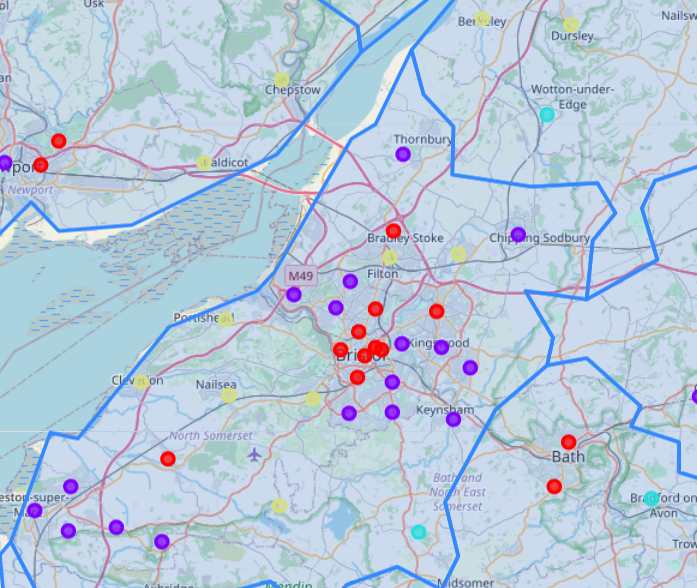
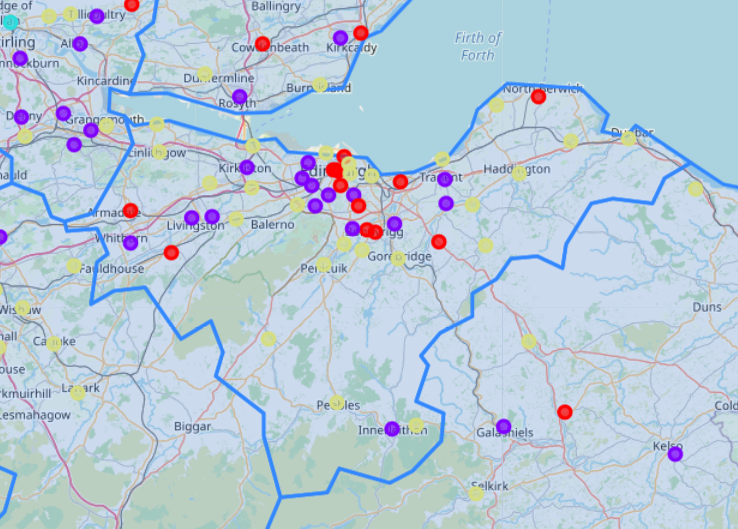
This shows the main venue in these locations are related to food, and there are more outdoor venues than the other clusters. The name given to this cluster was therefore 'Food and Outdoors'.

### Data Visualisation Results

The clusters associated with all the postcode districts in Great Britain can be seen in the image on the left-hand side below. Alongside this (on the right) is an example of the filtering available with these results. It can be seen from this second image that if the user is interested in certain districts only, they can interact with the map to view and compare these areas.



**Edinburgh Vs. Bristol Example**



**Edinburgh Clusters**

**Bristol Clusters**

## Discussion

When interrogating the results visually we can infer that a fairly accurate model has been created. This can be seen as the centre areas of big cities largely fall into the same category which includes a lot of shops/restaurants/bars and cafes, and rural areas largely fall into another category. However, the

There are two major limitations that have been identified when creating and reviewing this model. These are:

1. A 1km circle around the centre of the postcode is used for all districts when identifying associated venues. In some cases this may be a good fit however in some larger districts this may not be sufficient. If this model was to be improved in the future it is recommended that a method of finding venues from all around the district (rather than just the centre) in a non-circular fashion is implemented.
2. Venue location only is used to categorise districts. The inclusion of housing prices, school information, infrastructure or other data-sources would increase the usefulness of this analysis for different stakeholders. Again, if this model is to be improved, future work is suggested to identify the major data sources that would affect stakeholders’ decision making the most, and implement these when creating the clusters.

It is suggested that the model in its current state can be used to inform interested stakeholders on what districts may be of interest to them when considering where to re-locate to, however it is recommended this information is used to filter out districts that are not of interest, and then further research is used in conjunction with this first pass to help inform a complete decision.

## Conclusion

A technique to compare all postcode districts within Great Britain has been presented in which these districts are categorised into 4 discrete categories. After analysis, these categories have been defined as 1) High No. Eateries and Shopping, 2) Medium No. Shopping and Eateries, 3) Least Amount of Venues (Mainly Pubs), and 4) Food and Outdoors. On inspecting the results it appears a fairly good model has been created, with some limitations identified. These limitations are 1) the way in which venue data is identified, and 2) the lack of additional datasets used in conjunction with venue data. Based upon the current limitations, future work is suggested to create a different way of venue identification (rather than a circle for all areas), and to include further datasets alongside venue data to improve the model.

The current model can be used to help inform a first pass decision for interested parties on where they should consider further in an attempt to answer their question of “Where Shall I live When I Re-Locate?”.

## References

[1] <https://developer.foursquare.com/>

[2] <https://github.com/roblascelles/uk-postcode-map/blob/master/data/uk-postcode-area.json>

[3] <https://www.freemaptools.com/download-uk-postcode-lat-lng.htm>