IBM Data Science Certification

# CapStone Project

Geospatial Analysis of Sheffield, United Kingdom

## 1. Introduction

I wish to create a model/tool which enable people who are thinking of moving to and or opening a business in Sheffield, UK.

Sheffield is a city in the north east of the England, in the county of south Yorkshire. It is famously a heavy industrial city but has begun recently changing into a city of multi economic sources. There have been significant changes to areas across the city in recent years. Specifically, with the adoption of small enterprises being favoured by residents over large brands.

In this particular example of the report a person is wanting to open a coffee shop in Sheffield, they will need to now the areas that are already saturated with coffee shops and the areas that are in place for a new coffee shop.

What determines a good location to open a business? **Not an area which is already saturated**

This will have to be an iterative approach to help refine the model as I add more specifics to it

## 2. Data

What data sources will I need access to and what packages will I be using

**Geospatial** data to identify where Sheffield is, and where its areas/postcodes are. **Postcode.csv** is from the royal mail data portal it contains the high-level postcodes for Sheffield with the corresponding geo coordinates

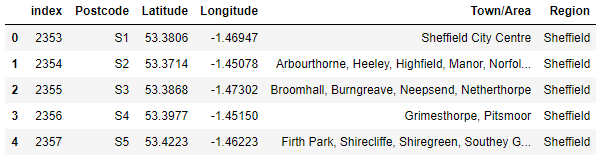
**Foursquare API** in order to identify what kind of facilities are in each area.

**Pandas, NumPy, Sklearn, Geopy, Folium**

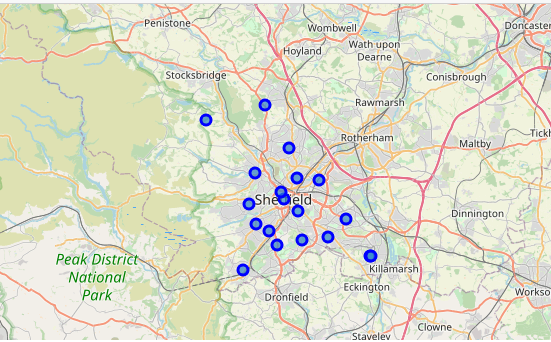
I will be using Pandas to import the geodata and the cleanse it into a usable format. Ready to be analysed using Sklearn and the plotting using Numpy.

## 3. Methodology

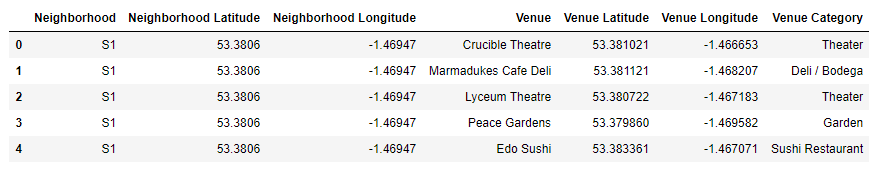
Data is imported using pandas, then it will be filtered down to show only Sheffield postcodes (see below for a snapshot of the data table). Also all unesscary columns have been removed.



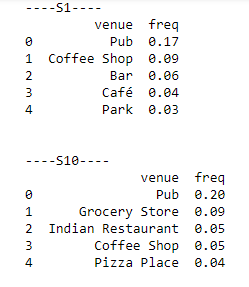
Now that I have a list of the high-level Sheffield postcodes with their corresponding geo coordinates, they will be plotted over a map of Sheffield to give you an idea where he post codes are and how they are spread out.



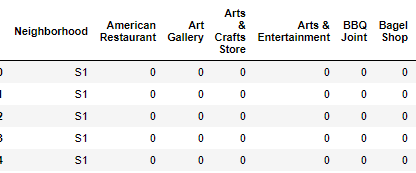
Now we have our geodata point, I am going to download the top 100 items within 1500 radius of each post code centre; utilising the Foursquare api. All of these venues are added to a single dataframe a function is applied to get the venues category. This data frame also includes the corresponding postcode.



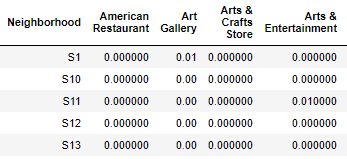
With the new data frame of venues and their corresponding the post codes I used a one hot encoding to each venue to enable to associated with each area as a frequency %. This is then used in the kmeans algorithm to generate a list of which areas are similar. I utilized a k value of 9 due the emerging nature of the different areas of Sheffield from personal experience. Below is a sample output of the top 4 venues by frequency by area.



We now the have all the information we need to run a kmeans algorithm over the data. Firstly we apply a one hot encoding to out data to get a data table with where we have columns for each venue type. For each venue the corresponding value of 1 is filled in for there column and 0 in the others. This table is then grouped in to a new dataframe, grouped by postcode and the values of 1 and 0 across the venue category columns are then averaged to create the % frequency of each venue type in each post code.



Example of the One Hot Encoding



Example of the one hot encoding frequency % grouped by Post Code

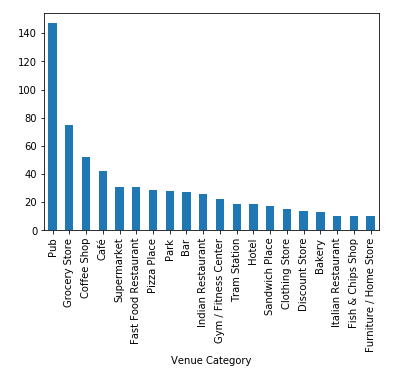
With this new dataframe we will run it through a kmeans object which outputs the clustering labels results. These results are then appended to the data so we can output a coloured table and a coloured map of the results



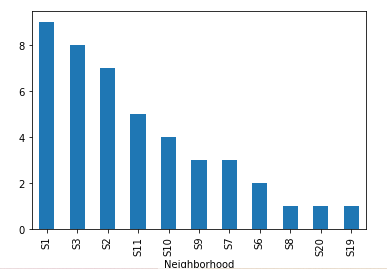
Example of the kmeans output joined in to the data table

## 4. Results

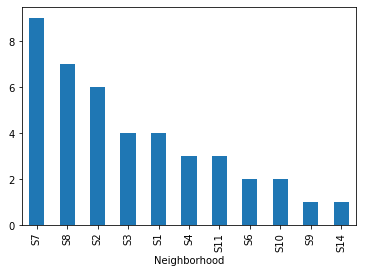
Initial analysis quickly shows which are top 20 venue categories in the Sheffield. Here we can see that Pubs are the biggest number with ~145 of them. Coffee shops are in third with ~50 of them.



By taking this data then filtering it to show only coffee shops and then grouped by post code, we can see which areas have coffee shops and how many in each. We can see that S1, S2 and S3 have the most coffee shops

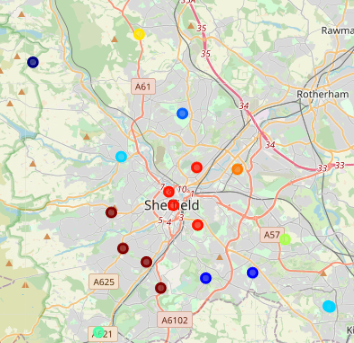
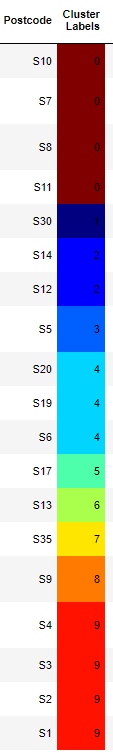


A similar analysis looking at cafes shows; that S7, S8, S2 have the most cafes. This is important as in the UK there is often very little distinction between cafes and coffee shops.

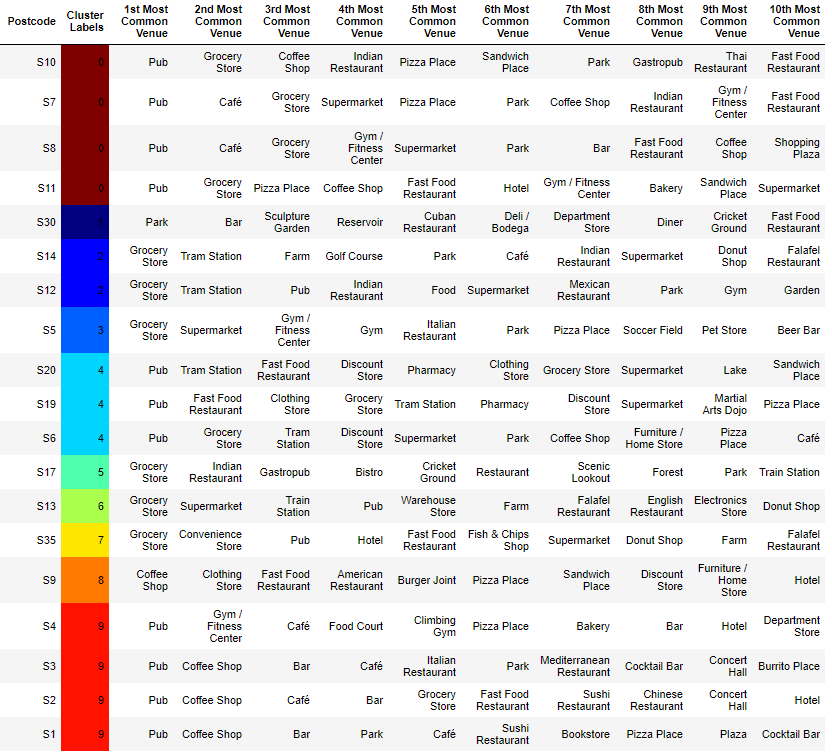


So out initial analysis suggests that S1, S2, S3, S7 and S8 would not by an ideal area to start a coffee shops as there would be a lot of competition.

Now we apply the kmeans to the data to see how the group together. The table and map below show the post codes, and which cluster the apply to.



In the next table we can see the areas with the top 10 venue types.



From this list combined with the previous charts we will be able to recommend an area to where to open a coffee shop

## 5. Discussion

As mentioned, before we are looking for a place to open a coffee shop in Sheffield. It is important to remember that in the UK there is little difference between cafes and coffee shops. From the previous graphs we could see that S1, S2, S3, S7 and S8 have the most cafes and coffee shops. There fore these areas would not the best place to open a coffee shop as they are already saturated.

From previous knowledge know that areas in the 0 cluster have the heights amount of disposable income in the city. So we should look at these areas for opening the new coffee shop. Using the table above, with k-means results in we van choose the area in cluster 0 with coffee shop /café low ranking. To this end I recommend opening the shop in post code S10.

## 6. Conclusion & recommendations

So we would recommend that the client opens there coffee shop in the post code S10.

Let’s talk about some caveats, this analysis is single dimensional only looking density of venues in each area. In future iterations of this model I would use social economic data to measure the distribution of wealth and available income in the city. Plus the being able to see age demographics will help with the design of coffee shop and its offerings. I would also look at the residential/commercial or industrial ratio in each area. This would help to estimate the potential footfall and when the shop would be busy.