IBM Data Science Professional Cirtificate Capstone

Categorisation of Birmingham areas by prevalence of restaurants and takeaways

1 Introduction and business problem

This report has been written to detail the work carried out as part of the Data Science Capstone model of the IBM Professional Cirtificate in Data science. The report will detail the business problem being explored, the data being utilised, the method appled, the results of the project, and a summary conclusion.

1.1. Background

It is well undertstood in the UK water industry that sewer blockages are a bad thing. They cause flooding and poillution while costing. The industry spends £$100m cleaning sewers annually. So to be able to target high risk areas is a clear benefit. There is a known correllation between sewer blockage incident rates and the prevalence of restauyrants and takeaways in an area. This is due to the proportion of fats oils and greases in their wastewater effluent.

1.2. Problem

For this project we will assume that an asset manager is looking to trial a fleet of sewer monitors in the city of Birmingham, UK. The asset manager is looking to install the monitors in areas not just with a blockage history but in areas with a high theroretical risk. To solve this problem we can use foursquare venue data to categorise area's by ammenities to select the top 5 areas to target with monitor installations to detect blockages. The targetting of location will help create a more efficient trial.

2 Data collection

In order to sovle this problem we will focus on the spatial data required. Using the template methods in the Toronto project we will utilise a list of post code centroids along with the data from the foursquare API.

2.1. Spatial Unit Data

For spatial unit data we will utilise the postcode-outcode dataset from the website https://www.freemaptools.com/download-uk-postcode-lat-lng.html. An example of the table is show in Table 1 below. The data is made available in a a comma delimeted text file. The text file has been represented in a tabular manner here for legibility.

id	postcode	latitude	longitude

2	AB10	57.1400	-2.1173
3	AB11	57.1388	-2.0909
4	AB12	57.1010	-2.1106
5	AB13	57.1080	-2.2378

Table 1: Sample postcode centroid data

This postcode area data set will need to be narrowed down to the birmingham area before it is utilised. Non-geographic postcodes do exist in the UK, however that are a lower level of granularity than this dataset. The granularity of this dataset has been selected as the individual postcodes would not contain sufficient numbers of properties to categorise.

2.2. Foursquare API data

The geographic ammenity data which we will use in combination with the postcode table for classification will come from the Foursquare API. Foursquare provide data via their API free for use in this acedmeic purpose. Foursquare holds data on different venues including name, location, type, popularity, reviews, and images. For the purpose of this assessment we are meerly concerned with type and location data. An example of the data returned by the api for a GET request for is below


```
"meta": {
  "code": 200,
  "requestId": "5ac51d7e6a607143d811cecb"
},
"response": {
  "venues": [
    {
      "id": "5642aef9498e51025cf4a7a5",
      "name": "Mr. Purple",
      "location": {
        "address": "180 Orchard St",
        "crossStreet": "btwn Houston & Stanton St",
        "lat": 40.72173744277209,
        "lng": -73.98800687282996,
        "labeledLatLngs": [
          {
            "label": "display",
            "lat": 40.72173744277209,
            "lng": -73.98800687282996
          }
        ],
        "distance": 8,
        "postalCode": "10002",
        "cc": "US",
        "city": "New York",
        "state": "NY",
        "country": "United States",
        "formattedAddress": [
          "180 Orchard St (btwn Houston & Stanton St)",
          "New York, NY 10002",
          "United States"
        ]
      },
```

```
        "categories": [
          {
            "id": "4bf58dd8d48988d1d5941735",
            "name": "Hotel Bar",
            "pluralName": "Hotel Bars",
            "shortName": "Hotel Bar",
            "icon": {
              "prefix": "https://ss3.4sqi.net/img/categories_v2/travel/hotel_bar_",
              "suffix": ".png"
            },
            "primary": true
          }
        ],
        "venuePage": {
          "id": "150747252"
        }
      }
    ]
  }
}
```


The data returned is in .json format but we can use pandas to create a dataframe of the key information we need.

3 Methodology

The outline methodology for this project covers the following stages. Data preparation, data exploration, data analyis, and results visualisation.

3.1. Data preparation

In order to work on the data using python the collected data was required to be imported into pandas as dataframes. The input data took two forms for the spatial areas. The first postcodes, and the second ward boundaries. The two datasets were then merged on postcodes so that the centroid coordinates and labels were shown within the same dataframe.

Venue data was collected using the Foursquare Api using a call of the following format

```
get('https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},
{}&radius={}&limit={}')
```

A for loop was used to carry out a fetch all venues with 1km of each ward centroid. This was then loade d into a datframe to process further.

3.2. Data processing

In order to create the datasets required for clustering it was required to create a dataframe of the top 5 venue types within each ward. The proportion of each venue was needed to be normalised.

This was carried out using a technique called onehot encoding. The process then creates a table such as the one below

	ward	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Acocks Green	Supermarket	Pub	Furniture / Home Store	Bowling Alley	Grocery Store
1	Alcester	Indian Restaurant	Bar	Supermarket	Tea Room	Grocery Store
2	Aston	Soccer Field	Photography Studio	Soccer Stadium	Advertising Agency	Movie Theater
3	Batchley & Brockhill	Construction & Landscaping	Golf Course	Park	Movie Theater	Playground
4	Belle Vale	Pub	Indian Restaurant	Pizza Place	Grocery Store	Park

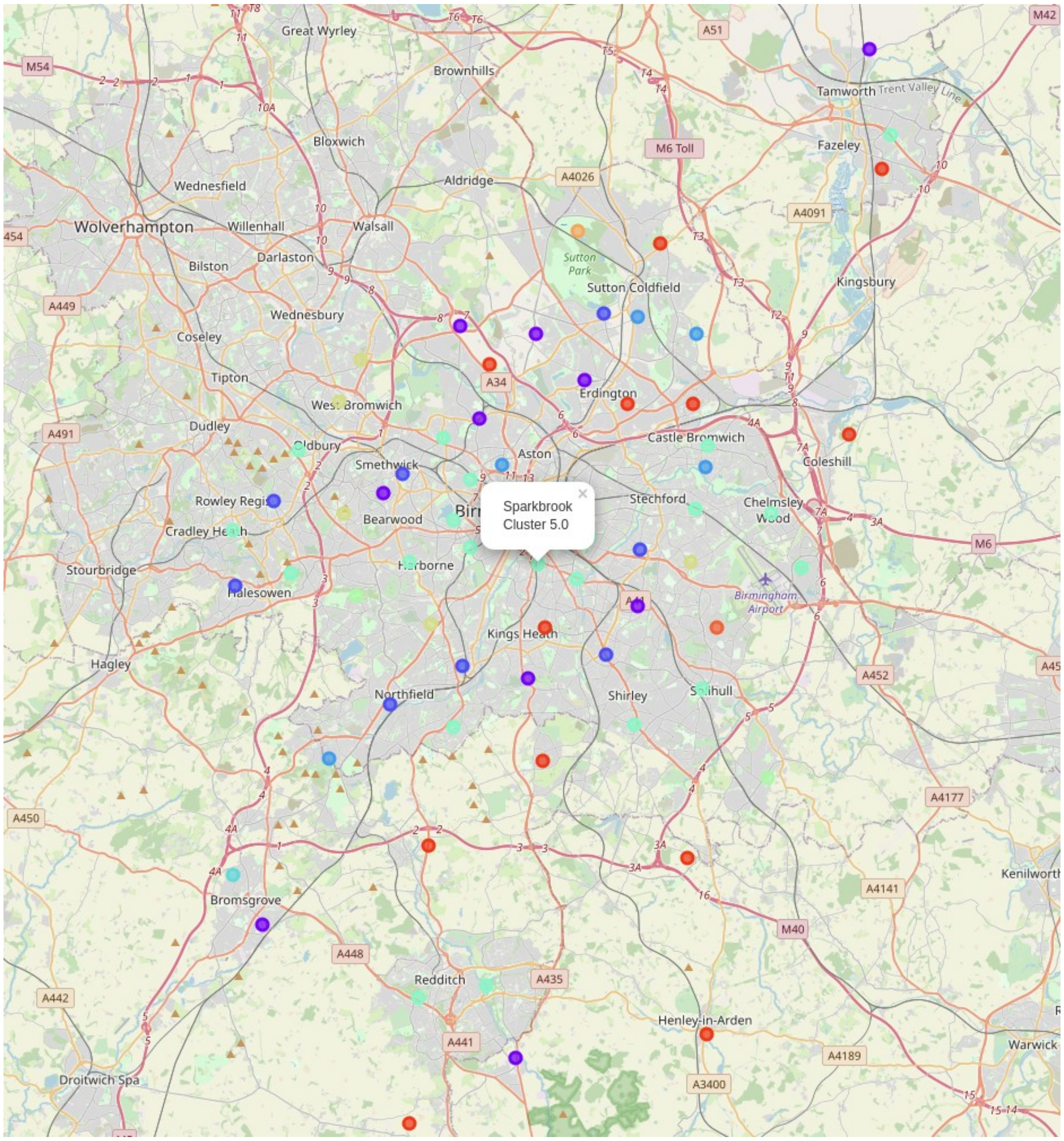
3.3. Data analysis

The clustering method selected was K- means. This was used as it is a reasonablly efficient and simple algorithm to apply. The clustering on a single attribute has been used to simplify the project.

The number of clusteres created was adjusted based on the output to provide a best fit to the known distribution of the wards in Birmingha, without appearing to be overfit.

4 Results

The results from the clustering shows that areas known to be high in restaurants and takeways tending to exist in cluster 5.0. A test case for this was the ward of Sparkbrook which is bart of Birmingham’s famous Balti Triangle. The more rural anmmd suburban areas were also clustered togther as red and purple. The image below shows the categorisation of wards across the birmingham area.


Sparkbrook
Cluster 5.0
Great Wyrley
Brownhills
Tamworth
Trent Valley Line
Fazeley
Bloxwich
Aldridge
Wednesfield
Wolverhampton
Willenhall
Walsall
Bilston
Darlaston
Sutton Park
Sutton Coldfield
Kingsbury
Wednesbury
Coseley
Tipton
West Bromwich
Erdington
Dudley
Oldbury
Castle Bromwich
Aston
Smethwick
Coleshill
Rowley Regis
Stechford
Chelmsley Wood
Bearwood
Cradley Heath
Stourbridge
Harborne
Halesowen
Birmingham Airport
Kings Heath
Hagley
Shirley
Solihull
Northfield
Bromsgrove
Redditch
Kenilworth
Henley-in-Arden
Warwick
Droitwich Spa
M54
M6 Toll
M6
M40
M42
A51
A4026
A4091
A449
A34
A491
A452
A4177
A450
A4141
A448
A435
A442
A441
A4189
A3400
A45

5 Discussion

The spread of results shows that the likely locations for blockage monitoring could be the inner city urban areas denoted by the cyan catgegory. It is likely that the red and purple categories would make poor choices for monitoring.

6 Conclusion

This execise while providing simple insight into the clustering of birmingham of wards. Further work to be carried out on this line of investigation would include the mapping of sewer blockage data by ward and also detailed analysis of the demographic and social factors in each ward. This could lead to a more precise and insightful blockage model.