Categorisation of Neighborhoods in the Norwich urban area

Capstone Project: IBM Data Science Professional

Anthony Maina

Background & Problem

Each year, many **mid-career professionals relocating from London s**eek smaller, quieter, less expensive cities. The reasons are varied, but typically include some combination of:

- A desire to purchase property at more accessible price
- A desire to raise small children in a safer, cleaner environment
- A desire to change pace (e.g. from a hectic banking career to owning a small local business)
- ...and so on

Rather than move back to the city where they grew up, it is not uncommon for these individuals to consider cities they have never lived in, or have only visited once or twice.

They are likely to want an environment which provides not just calm and quiet, but the variety of local amenities and entertainment venues that they have become used to in London. In short, they want to have their cake and eat it, to the best extent possible.

Norwich is an interesting choice of location, as it is only 2 hours away from London by train, and provides access to a majestic countryside and the scenic seasides of Yarmouth and Cromer.

In this study, we explore the neighborhoods of Norwich, attempting to cluster the neighbourhoods in the city based on the types of venues present. I

The output is a small number of categories, with their defining characteristics (in terms of mix of venues) which can be presented to prospective home-buyers to inform their purchasing decisions.

The key stakeholders are **mid-career professionals relocating from London**, as well as **property agents and property sellers based in Norwich** itself

Data sources

Analysis is based primarily on two datasets:

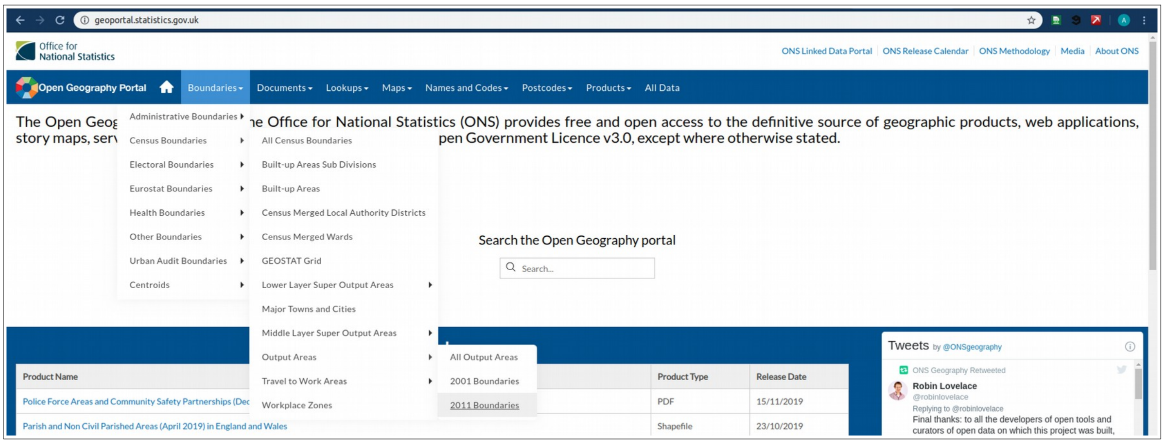
1. **Definition of neighborhood boundaries and centroids.**

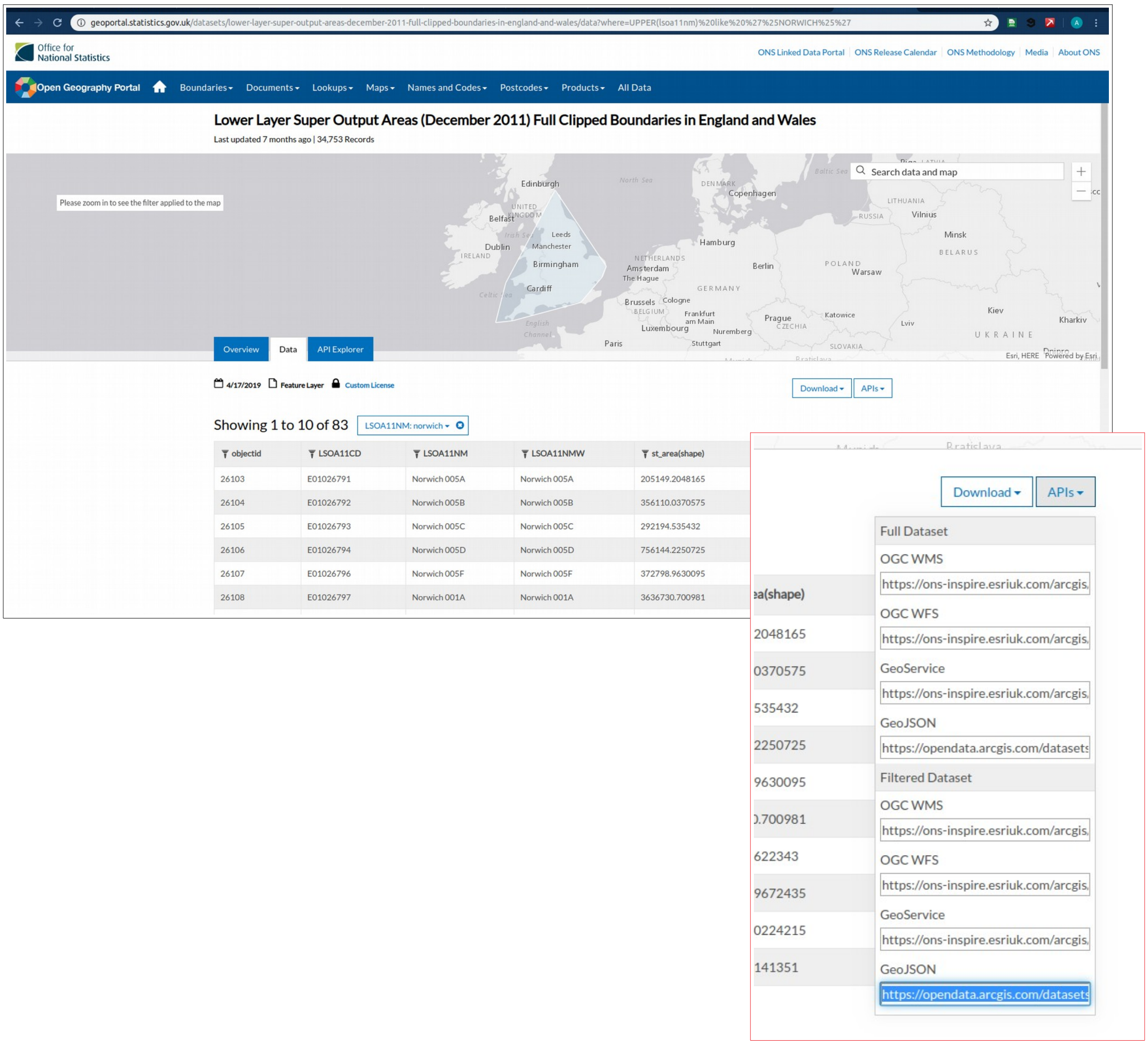
The UK **Office for National Statistics** maintains sets of boundaries used for different purposes, including Census, Electoral, and Local Authority boundaries.

To zoom in at the right level, we use **Lower Layer Super Output Areas (LSOA)** as the basis for boundaries. LSOAs are a geospatial statistical unit used in England and Wales to facilitate the reporting of small area statistics. They are created and maintained by the ONS. They have a minimum population of 1000 with a mean size of 1,500.

Data on LSOA boundaries is published free of charge as either pdf or Feature Layer on the ONS website.

In this study, we use API calls to retrieve the geospatial boundary data. We will collect this data as a **GeoJSON data**, and manipulate it to derive insights

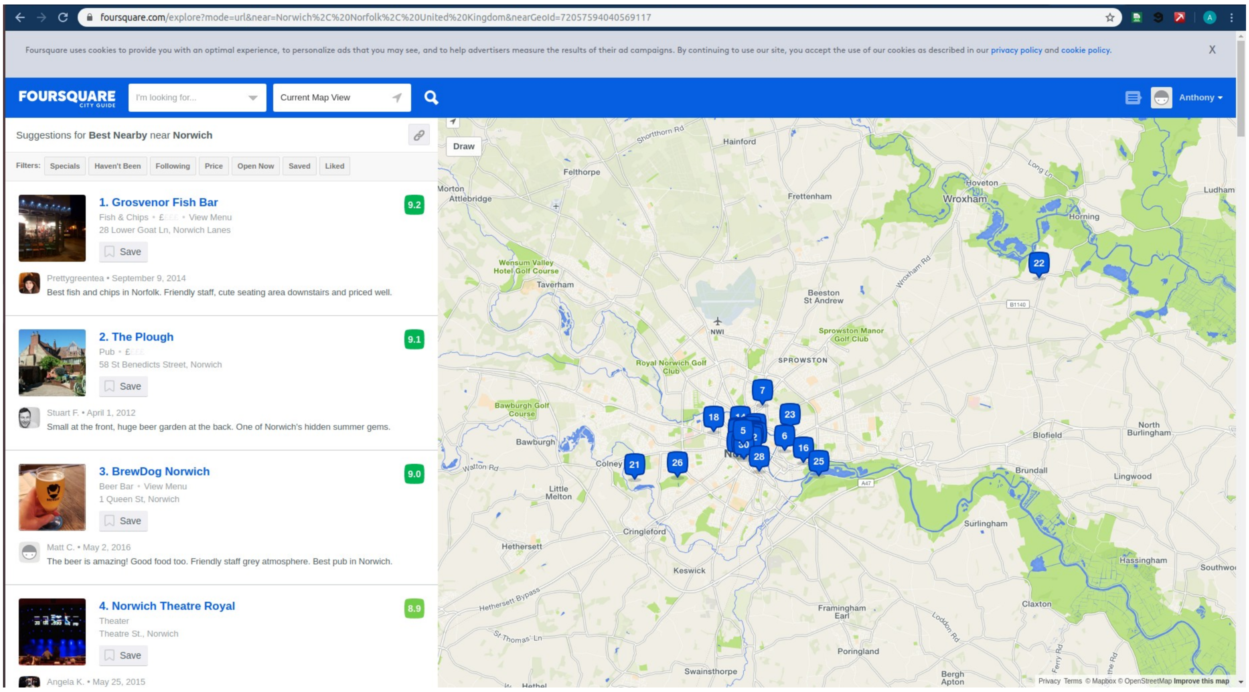




2. Addition of data on local venues: Foursquare

Foursquare provides the leading source of search-and-discovery data on what types of venues exist in a given area, as well as additional information such as usage, ratings, etc.

We use the locations data from Foursquare to map the types of venues which exist in the vicinity of each neighborhood.



3. How the data is used (overview)

The data will be used as follows:

1. Extract and refine boundaries data
   a) Extract the data via an API call and inspect the boundaries using a visualisation package (e.g. Folium)
   b) Once happy with the data, and any outliers or anomalies have been removed, compute the centroids which will form the "centrepoint" of each neighborhoods

2. Add Foursquare Data
   a) For each neighborhood, extract the list of venues within a given radius (e.g. 1km); the radius will be defined based on the closeness of the neighborhoods)
   b) Use "one-hot encoding" to convert the list of venue categories per neighbourhood into a Feature Set
   c) Normalise the featureset by grouping the values by mean
   d) Use **k-clustering** to define clusters based on the similarity of their features (ie mix of venue categories)

3. Perform checks and sensitivity analyses

a) Check the k-cluster score for different values of k, (ie the elbow method)
b) Check the similarity of cluster members to each other (Silhouette method)

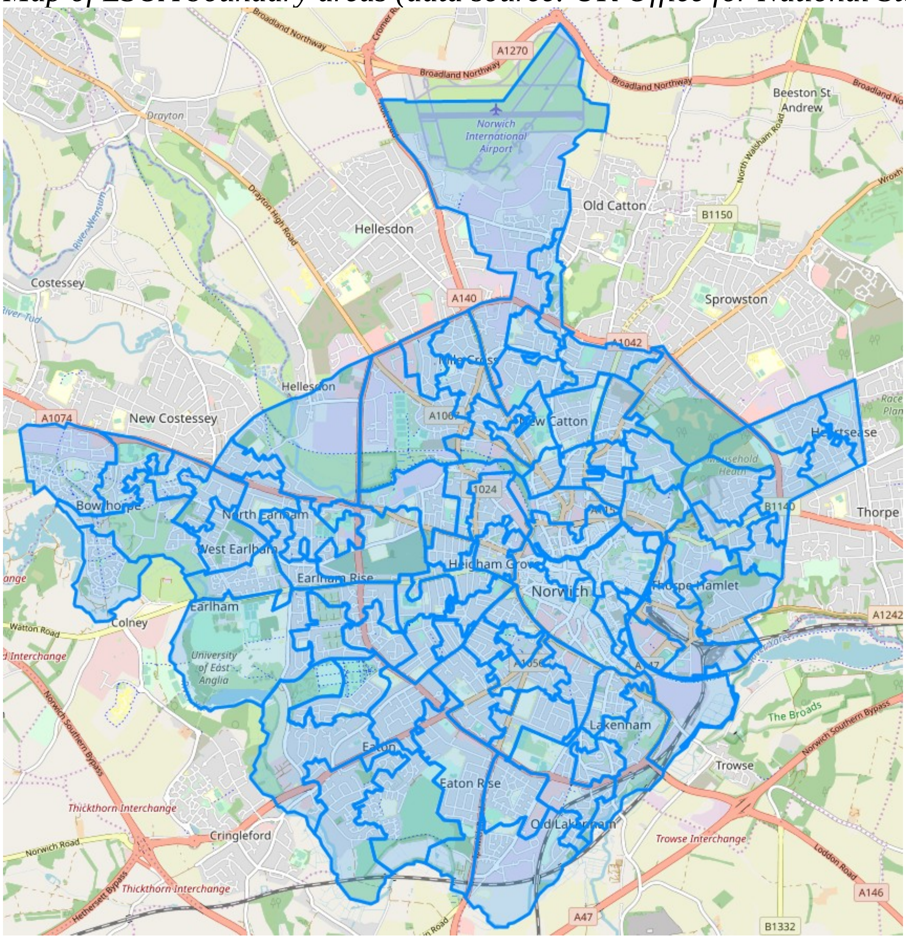
4. Produce results and discussion

Methodology

Extraction and refinement of boundaries data

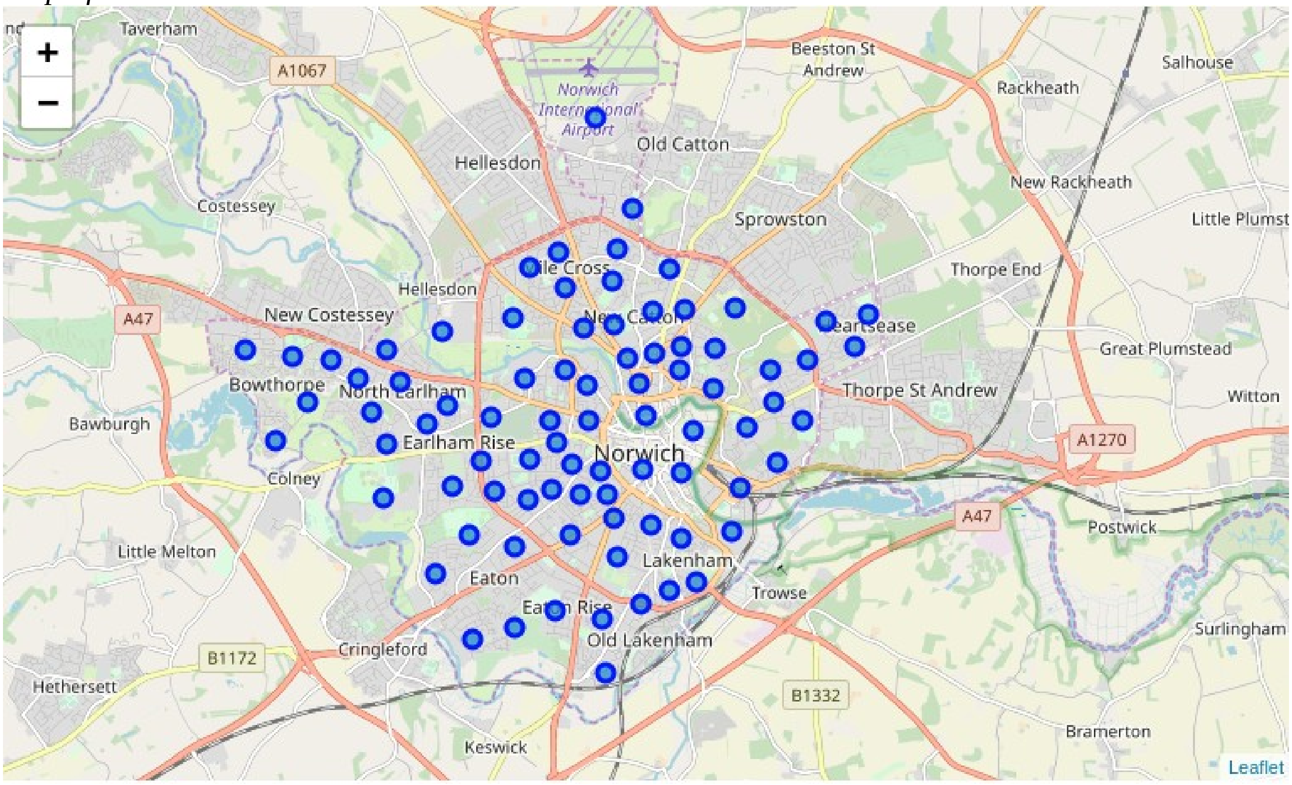
We first extract the boundaries data using an API call to the ONS portal and examine the result. We can see that this is a GeoDataFrame, which in the last column defines a “polygon” representing the boundaries of the LSOAs.

Map of LSOA boundary areas (data source: UK Office for National Statistics)



We compute the centroid of each shape and re-run the map to show the centroids rather than the polygons.

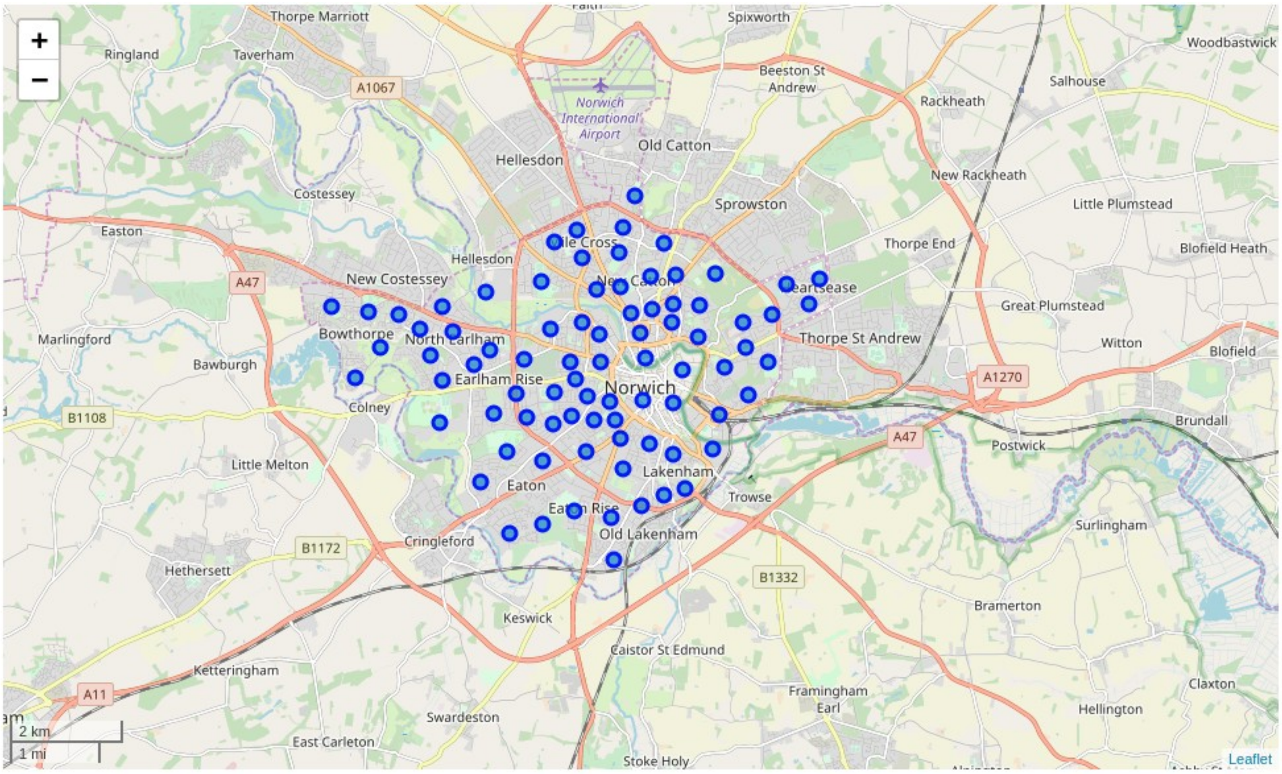
Map of LSOA centroids



We notice inspect the resulting map, and note the distribution of the centroids, looking for any outliers which could adversely impact the result. We notice for example that there is a centroid located very near Norwich International Airport.

Not only is an International Airport unlikely to be near the first choice of areas to live, due to noise, we also know that an airport is likely to have significantly different features (in terms of venues) compared to a typical suburb or neighborhood. To be safe, we drop this centroid and focus on the remainder.

Map of LSOA centroids excluding airport



Extraction of Foursquare data

We extract the Foursquare data showing all the venues within a 1km (0.6 mile) radius of each centroid, up to a limit of 50 venues. Since Norwich is a small provincial city, dominated by suburbs rather than the town centre, this limit feels more appropriate than say 100 or higher.

To give ourselves a quick flavour of the result, we inspect a summary of the most common and least common venue categories

Summary of most common and least common venue categories

```
115 unique venue categories found in Norwich

The most commonly occuring venue categories in the vicinity of a neighbourhood are (by count of occurence):

                      Venue Category
Pub                              401
Grocery Store                    123
Coffee Shop                       98
Café                              97
Hotel                             79
Bar                               67
Fast Food Restaurant              65
Pizza Place                       64
Supermarket                       63
Chinese Restaurant                62
Park                              56
Restaurant                        50
Indian Restaurant                 41
Italian Restaurant                38
Sandwich Place                    35
Shopping Mall                     32
Theater                           32
Clothing Store                    31
American Restaurant               29
Tea Room                          28

The least commonly occuring venue categories in the vicinity of a neighbourhood are (by count of occurence:

                      Venue Category
Post Office                        2
Bus Station                        2
Roller Rink                        2
Shoe Store                         1
Hockey Field                       1
Caribbean Restaurant               1
Stationery Store                   1
Arts & Crafts Store                1
Insurance Office                   1
Art Museum                         1
Ski Trail                          1
Medical Supply Store               1
Video Game Store                   1
Kitchen Supply Store               1
Buffet                             1
Storage Facility                   1
Museum                             1
Forest                             1
Lake                               1
Seafood Restaurant                 1
```

Within the most common categories, we can see that Norwich has a great number of pubs, (not surprising, given the city's national reputation for a large number of independent breweries). We also see other features we would expect to be common, such as grocery stores, supermarkets and a large number of restaurants. We note that there is a certain inconsistency with Foursquare's categories; some Restaurants are marked by type of cuisine, others are simply marked "Restaurant". If we tried to count the number of Restaurants there are in Norwich, we would need to be careful to include both groups, using for example a wildcard search on “Restaurant”. This is a separate topic, for now we are satisfied the data looks normal

Next we look at the least common categories. Some of these instantly make sense – Norwich is a small city, so two bus stations seems about appropriate. Similarly for the Roller Rink, the Art Museum and the Ski trail.

Others are slightly more confusing. For instance, there must be more than one Shoe Store in Norwich, however we know it is common for such stores to be inside a Shopping Mall rather than directly on the high street. Similarly for Video Game Store and the Statonery Store.

And then there is “Buffet” whose significance is not apparent at all.

These concerns are not of immediate importance to our study, so for now we are satisfied with the result.

Data analysis

Preparation of Data for analysis

The first step in the analysis is to create a Onehot encoding on the Venue Categories, to make them ready for the cleaning analysis.

This results in the below dataframe, whose length is equal to the total number of venues found (1910) and whose width is equal to the number of venue categories (108, excluding the Neighborhood column)

Snapshot of norwich_onehot dataframe

Out[20]:

	Neighborhood	American Restaurant	Antique Shop	Art Museum	Arts & Crafts Store	Asian Restaurant	Auto Garage	Automotive Shop	Bakery	Bar	...	Storage Facility	Supermarket	Tapas Restaurant	Tea Room	Resta
0	Norwich 005A	0	0	0	0	0	0	0	0	0	...	0	1	0	0	
1	Norwich 005A	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	Norwich 005A	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	Norwich 005A	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	Norwich 005A	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

5 rows × 109 columns

In [21]:
```
norwich_onehot.shape
```
Out[21]: (1910, 109)

We group this by mean and by Neighborhood, resulting in a dataframe whose length is now equal to the number of neighborhoods. Each row represents a “score” of how frequently that Venue category appears in the population of venues in that neighborhood.

This dataset is ready for analysis by a k-clustering algorithm.

Snapshot of norwich_grouped dataframe

```
In [28]: norwich_grouped = norwich_onehot.groupby('Neighborhood').mean().reset_index()
         norwich_grouped.head()
```

Out[28]:

	Neighborhood	American Restaurant	Antique Shop	Art Museum	Arts & Crafts Store	Asian Restaurant	Auto Garage	Automotive Shop	Bakery	Bar	...	Supermarket	Tapas Restaurant	Tea Room	Tha Restauran
0	Norwich 001B	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.100000	0.0	0.0	0.0
1	Norwich 001C	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.000000	0.0	0.0	0.0
2	Norwich 001D	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.062500	...	0.187500	0.0	0.0	0.0
3	Norwich 001E	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.071429	...	0.071429	0.0	0.0	0.0
4	Norwich 001F	0.071429	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	...	0.071429	0.0	0.0	0.0

5 rows × 116 columns

Selection of k: Elbow analysis

We run elbow analysis to see whether there is a clear elbow point pointing to the best value,k, for number of clusters.

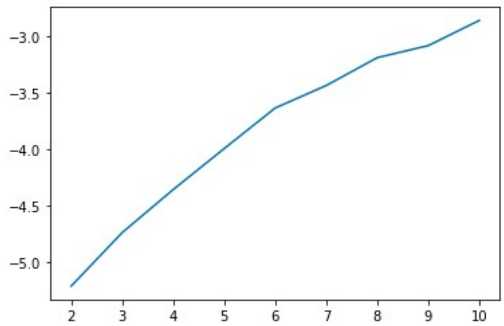
Rather than an elbow, we note that the curve approaches zero almost monotonically, so it is difficult to assess the best k. 6 could be a good guess, but we need to go further.

Snapshot of elbow analysis result

```
In [24]: #Define procedure to iteratively compute k-means score for k values from 1 to 10

         Ks = range(2, 11)
         km = [KMeans(n_clusters=i, random_state=0) for i in Ks]
         score = [km[i].fit(norwich_grouped_clustering).score(norwich_grouped_clustering) for i in range(len(km))]
         plt.plot(Ks, score)
```

```
Out[24]: [<matplotlib.lines.Line2D at 0x7ff24c591f90>]
```



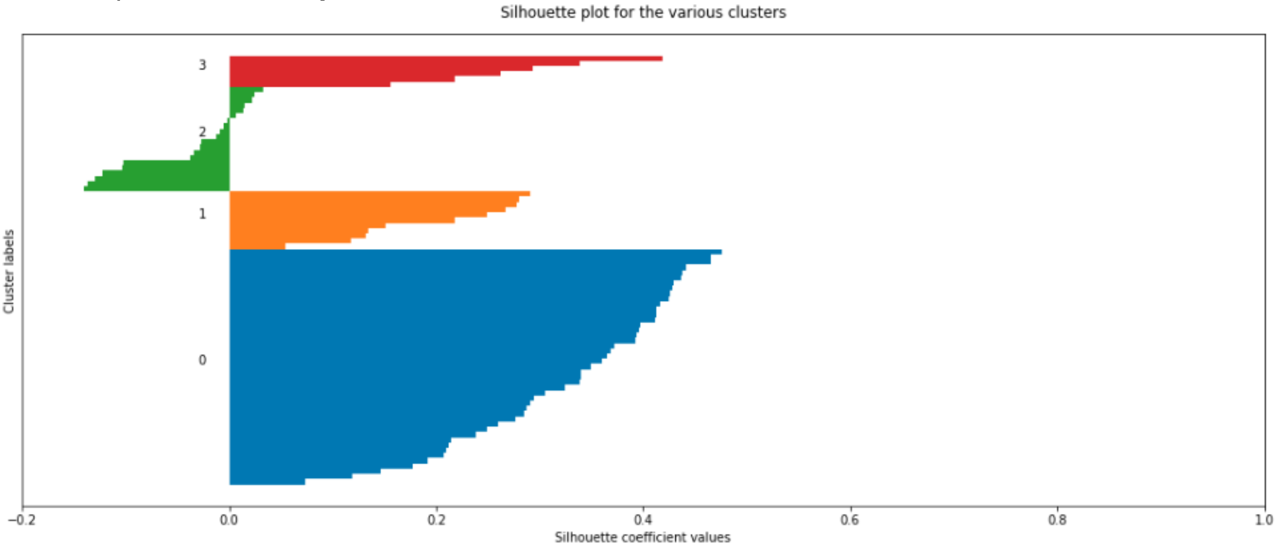
Selection of k: Silhouette analysis

Silhouette is a technique which provides a succinct graphical representation of how well each object in a cluster has been classified

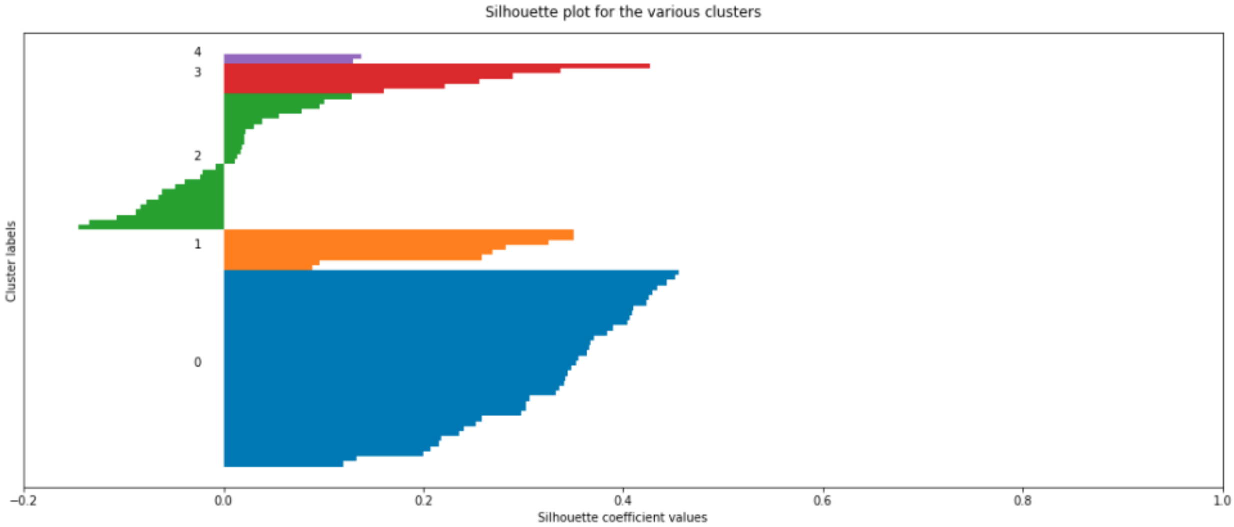
The silhouette value (or score) per object effectively measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. A negative value indicates that the object is more closely matched to another cluster than it is to its current cluster.

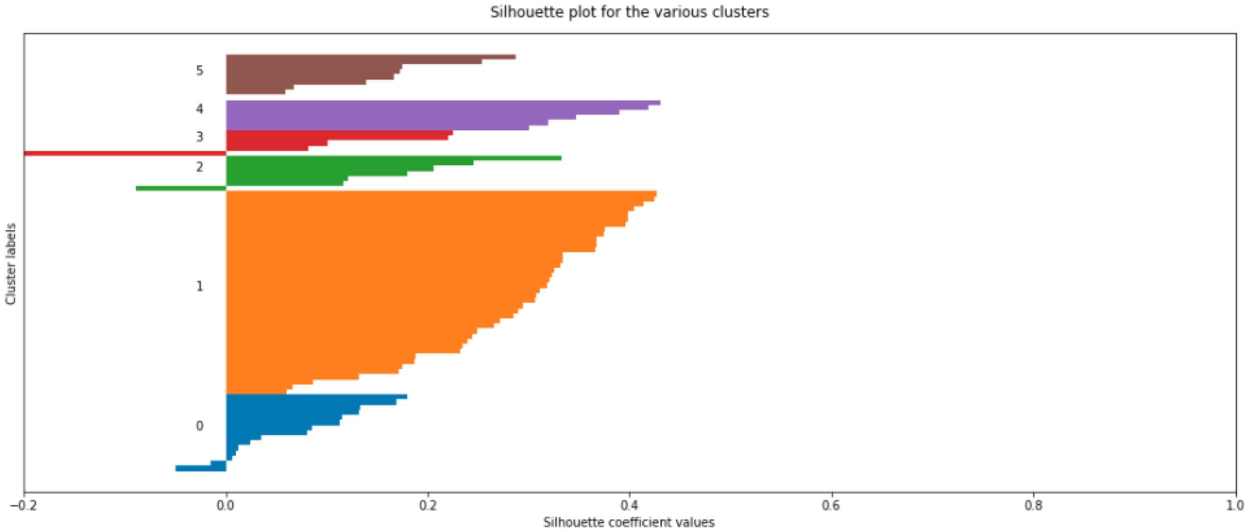
Results of silhouette analysis: k=4



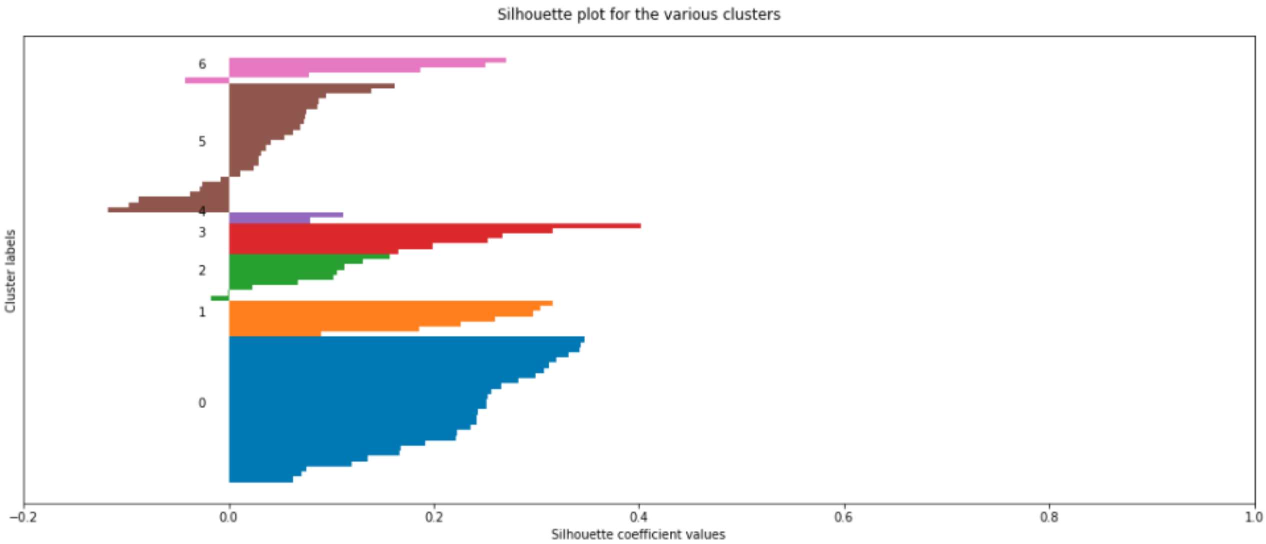
Results of silhouette analysis: k=5



Results of silhouette analysis: k=6



Results of silhouette analysis: k=7



We notice two interesting features regardless of the choice of k:

- All result in one large cluster, and a varying number of smaller clusters.
- All result in some “misclassification” (represented by negative scores) in one or more clusters

We infer from this that

- There are a large number of neighborhoods which are more “obviously” groupable (that is, containing higher feature similarity).
- There are other neighborhoods which are more difficult to group into any cluster. This could be for several reasons, for example if the types of categories showing up in these neighborhoods are not showing up in any other neighborhoods (e.g the Park, the Lake, the Caribbean restaurant)

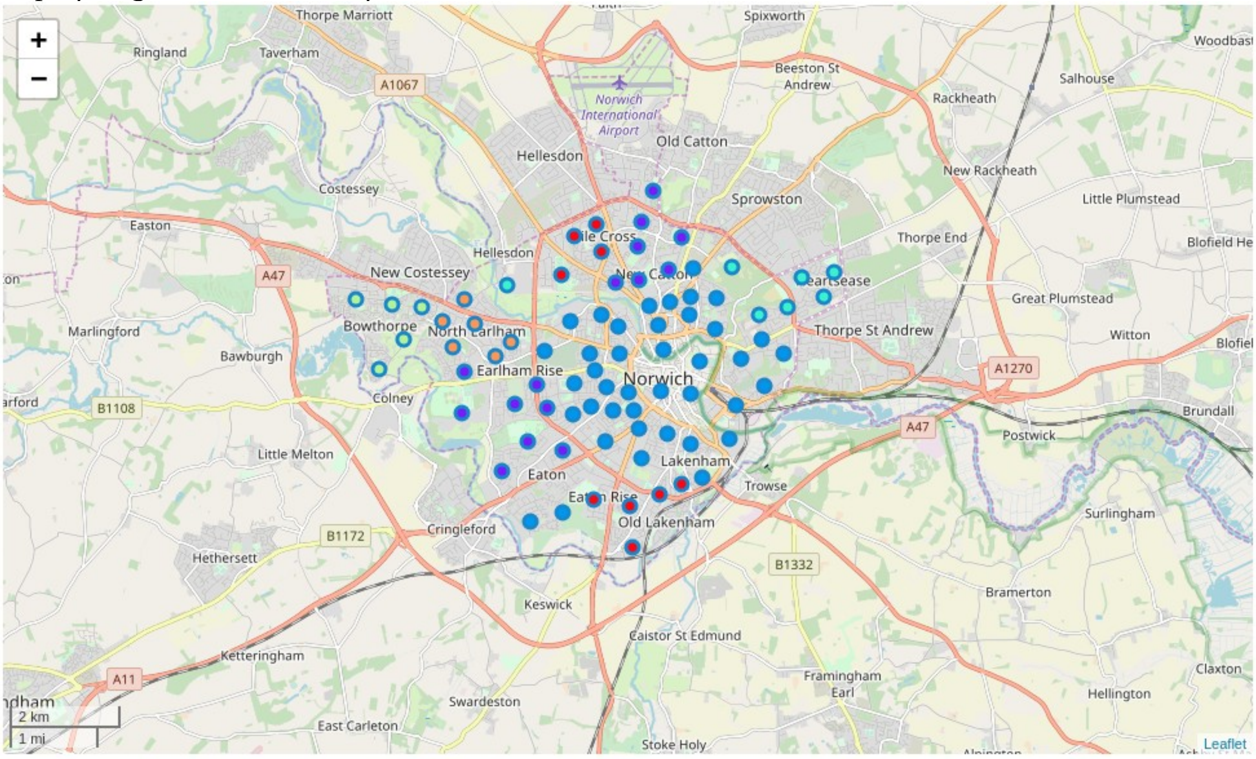
This could be an interesting venue for further study

For the purposes of this study, we will select k=6. This selection seems to minimise the amount of apparent “misclassification”

Results

We are able to produce a map of the different clusters, and can observe from the map the distribution of the 6 clusters. We note that 2 are non-contiguous (ie they are spread over more than one geographical region. These are clusters 0, 2 and 5.

Map of neighborhoods classified into k-clusters



We are also able to produce a table showing the most common venue types showing up in each cluster.

Summary table of most common venues found per cluster

Out[36]:

Cluster	Number of neighborhoods	1st most common venue	2nd most common venue	3rd most common venue	4th most common venue	5th most common venue
0.0	15.0	[Grocery Store]	[Park]	[Convenience Store, Pub]	[Grocery Store]	[Breakfast Spot, Event Service, Park, Theater]
1.0	40.0	[Pub]	[Coffee Shop]	[Café]	[Café, Grocery Store]	[Café, Pizza Place]
2.0	7.0	[Furniture / Home Store]	[Fast Food Restaurant]	[American Restaurant, Supermarket]	[Automotive Shop, Bar, Café, Convenience Store...	[Pub]
3.0	5.0	[Hotel]	[Bowling Alley, Department Store, Hotel, Resta...	[Shopping Mall]	[Soccer Field]	[Concert Hall, Department Store, Restaurant, S...
4.0	6.0	[Grocery Store]	[Convenience Store, Fast Food Restaurant]	[Fast Food Restaurant]	[Chinese Restaurant]	[Pub]
5.0	9.0	[Supermarket]	[Fast Food Restaurant, Supermarket]	[Electronics Store]	[Hotel]	[Electronics Store]

For the mid-career professional planning to move, we could interpret the option space as follows:

- **Cluster 0** is good for the professional who wants to be outside the city centre, and enjoy access to open spaces and parks, but would still like to have the convenience of easy shopping, restaurant and cafe options.
- **Cluster 1** is for the professional who wants to stay relatively close to the city and enjoy the convenience of multiple cafes and bars
- **Cluster 2** is for the professional who loves DIY, and doesnt mind being away from the social centre of the city. She can still reach the odd restaurant or bar, but she has better access to Furniture and Home supply stores
- **Clusters 3, 4, and 5** are more remote, and would suit the professional who is far less interested in the social scene, and rather places a premium on peace and quiet.

Discussion

The problem of selecting optimum neighborhoods was an interesting use case for k-means clustering.

A few experiments could be carried out as part of a more detailed study to tune the model parameters, for instance

- radius
- number of venues limit
- Using the k-means++ algorithm to choose different initial values (or "seeds") for the k-means clustering algorithm.

While the feature set was restricted to presence of a venue only, a few further steps would likely be necessary to further enrich the analysis. In addition to the list of venues, the following features could also be considered:

- Popularity of venues (e.g. by Foursquare trending data)
- Average income in the area (e.g. from the ONS census statistics data)
- Average house prices in the area (e.g. from a website like Zoopla.co.uk, which aggregates and provides such data for UK addresses)
- Average transit time from neighborhood centroid to Norwich train station, for those who will need to commute daily to work (data from Google Maps)

Addition of these layers of information could unlock further exciting insights and potentially create more explicit clustering.

Conclusion

Machine Learning methods were used to classify neighborhoods in Norwich based on similarity of venues in their vicinity.

Definition of location boundaries was obtained from the UK Office for National Statistics, based on the Lower Layer Super Output Area (LSOA).

Data on venues was obtained from Foursquare, the world’s leading provider of location data.

This information could be useful for mid-career professionals relocating from London to Norwich, or to property agents/sellers looking to more effectively advertise homes for sale.