Categorisation of Neighborhoods in the Norwich Urban Area

IBM Data Science Professional – Capstone Project

November 2019

Introduction / Business Problem

Each year, many **mid-career professionals relocating from London** seek smaller cities. Reasons include

- More accessible price of housing
- Desire to raise small children in a safer, cleaner environment
- Desire to change pace (e.g. from a hectic banking career to owning a small local business)

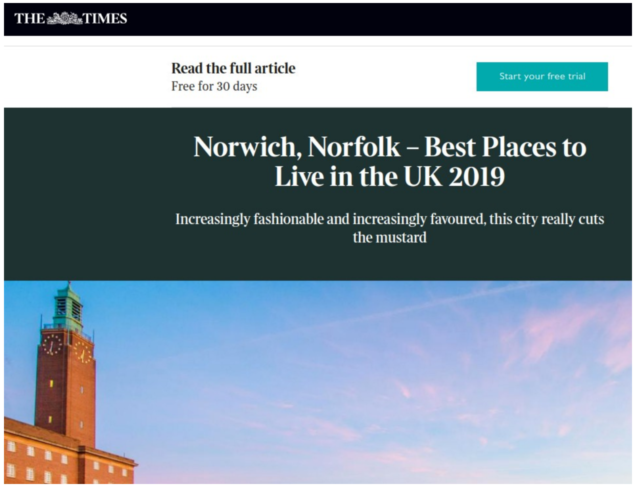
...and so on

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.

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.

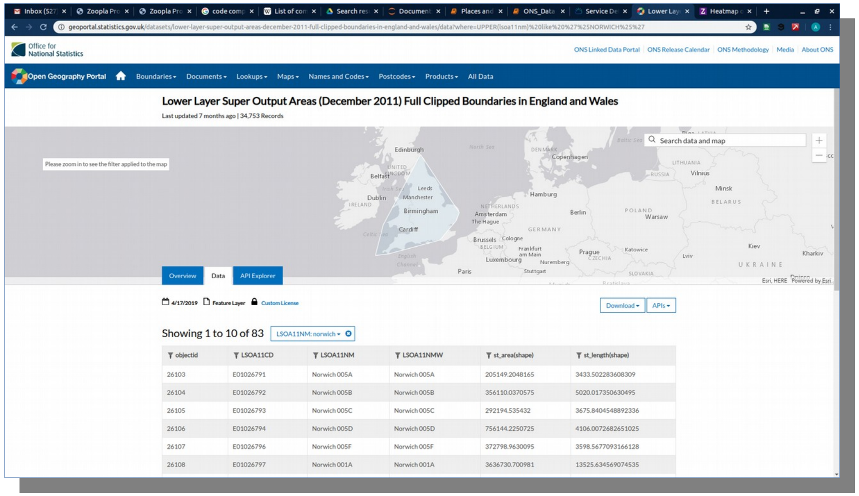
The **key stakeholders are mid-career professionals relocating from London**, as well as **property agents and property sellers based in Norwich** itself



Data

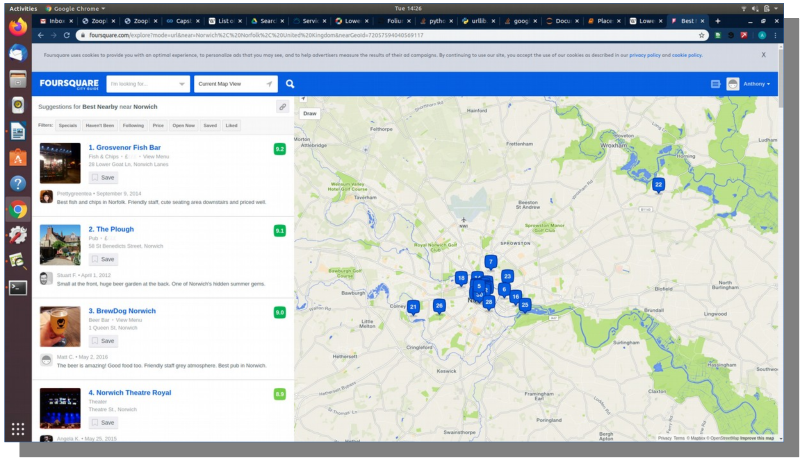
<u>1. Definition of neighborhood boundaries and centroids</u>.

- 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 thereporting 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.**

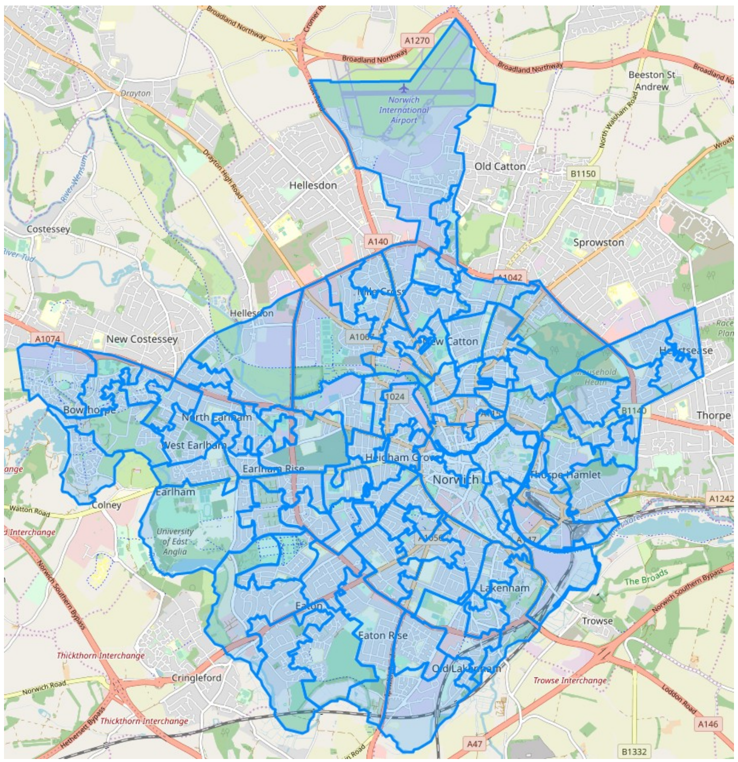


2. Addition of data on local venues: Foursquare

- 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 ofeach neighborhood.

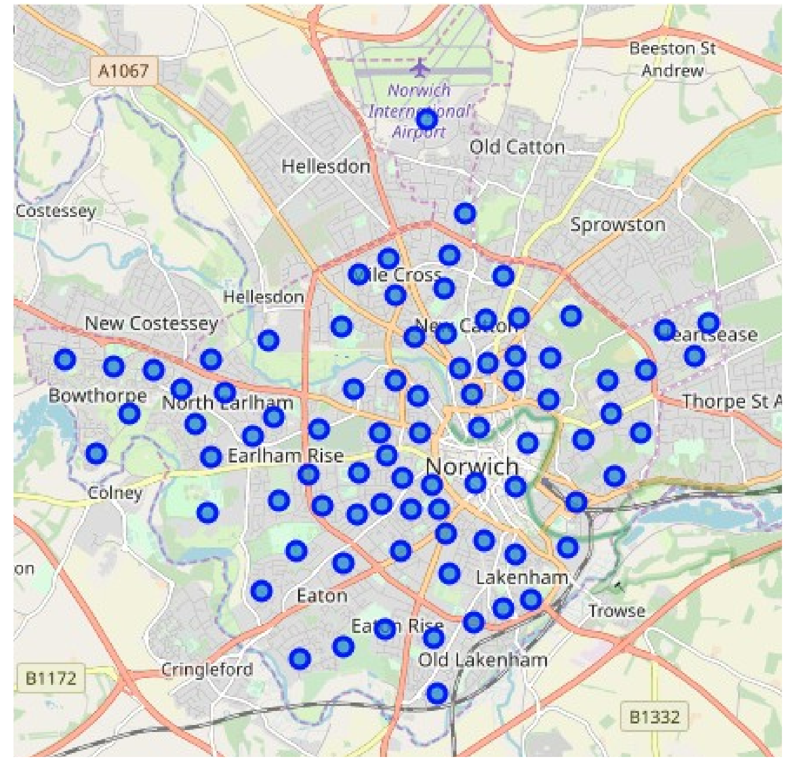


Methodology: ONS location boundaries and centroids



1. Raw Data

- **API call to ONS portal to** extract GeoJSON file of Norwich LSOAs
- ONS data in the form of **geoJSON polygons,** displayed here in Folium map



2. Processed data

- **GeoPandas library** used to compute centroids, and overlaid onto Folium map

Methodology: Foursquare data

```
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
```

1. Most common venue categories

- Pubs are highest frequency, (not surprising, given the city's national reputation for a large number of independent breweries).
- Others: grocery stores, supermarkets and a large number of restaurants.
- Slight inconsistency with Foursquare's categories; some Restaurants are marked by type of cuisine, others are simply marked "Restaurant".
- **Overall conclusion**: **OK for our purposes**

2. Least common venue categories

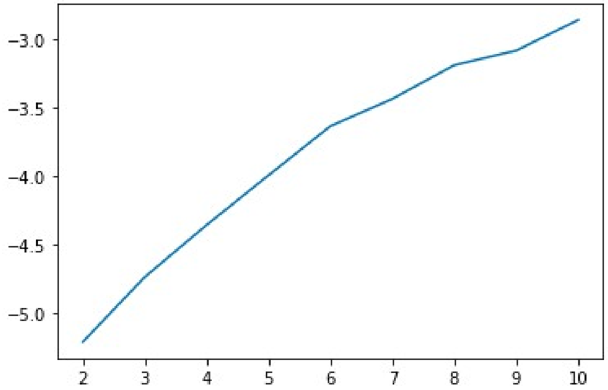
- Some instantly make sense (e.g. given a small city, two bus stations seems about appropriate).
- Others slightly misleading. e.g. only one Shoe Store in Norwich - However we know it is common for such to be inside a Shopping Mall
- Rare case ("Buffet") whose significance is not apparent at all.
- **Overall conclusion**: **OK for our purposes**

Selection of k: Elbow method

```
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>]
```

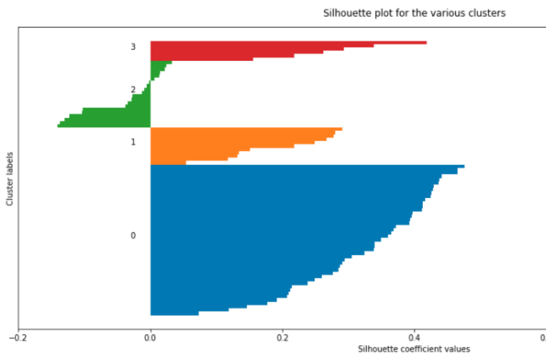


<u>**No obvious elbow, curve almost monotonically tends to zero as k is increased**</u>

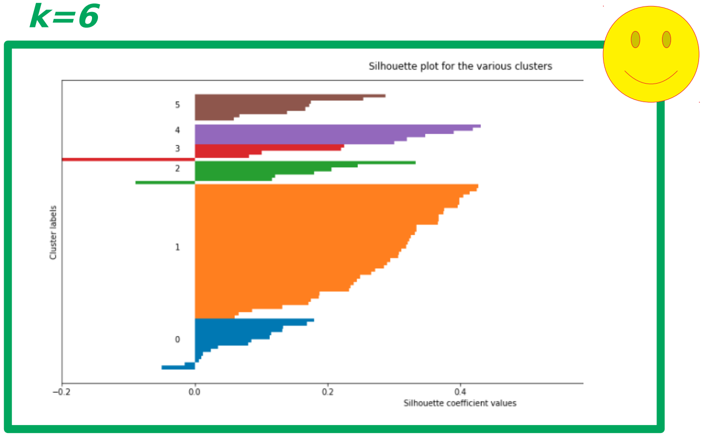
- Although Elbow method gives good indication of overall k-means score, it is **vague on the cleanness of classification per cluste**r for each choice of k
- **For a more explicit answer, we use the Silhouette method (next page)**

Selection of k: Silhouette analysis

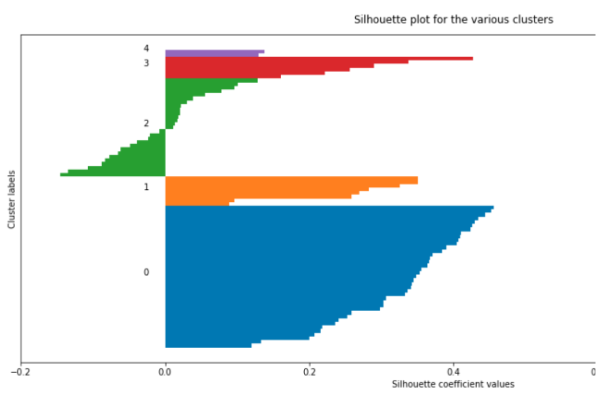
k=4



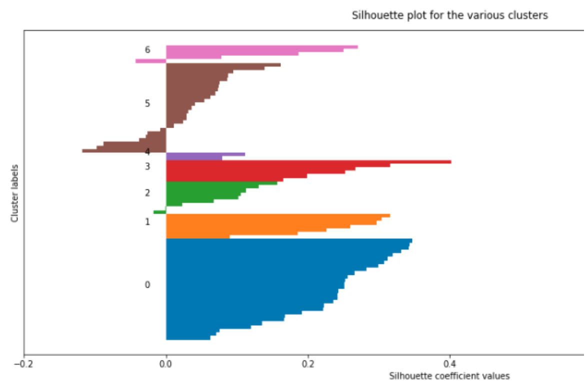
k=6



k=5



k=7



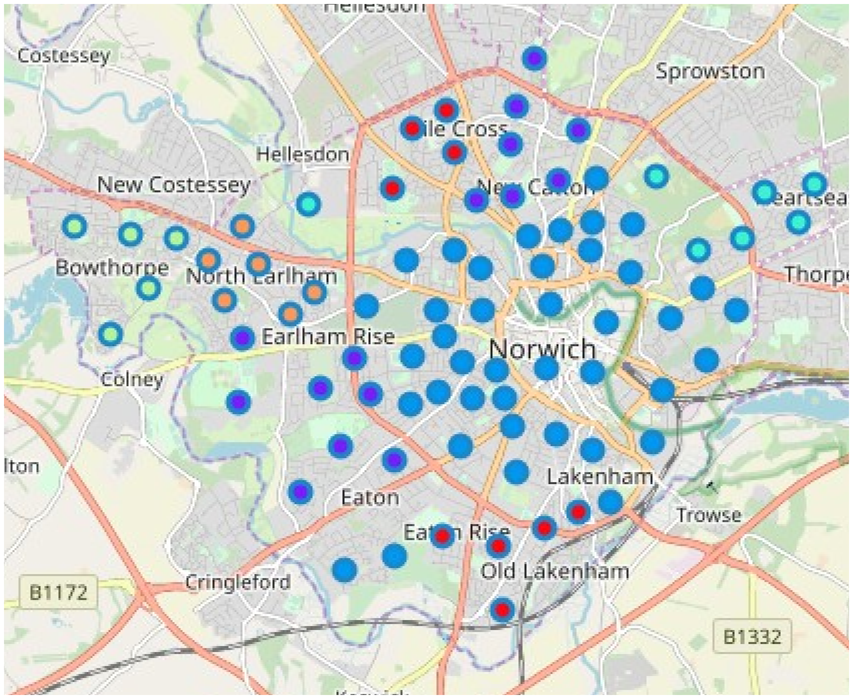
- Silhouette is a technique which provides a succinct graphical representation of how well **each object** in a cluster has been classified

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

- Value 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**

- 4 k values tested (4,5,6,7); **k= 6 is chosen** since it gives the fewest number of negative k-score values

Results & Analysis



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]

- 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.

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 results.