Coursera Report: Battle of the Neighbourhoods

Business problem:

Cities and towns in England are often berated for having "identikit", bland city centers, so where should a holidaymaker go if they want to have a good time? Which cities are best for "coffee culture"? Which are best for a more traditional feel? Is there some objective data that the general public (or tour operators) can use to make recommendations instead of relying on the often deceitful marketing material issued by the cities themselves or by the cities' tourist boards? Can we use data science to tell us whether it is worth visiting an English city without being disappointed or mislead by advertising?

Data use:

The starting point will be a list of cities and towns in England. English towns and cities have a rather perplexing nomenclature with both small and large localities being called "towns" or "cities", so we will use the website 'https://www.thegeographist.com/uk-cities-population-1000' to scrape the data and then clean it, removing localities in Wales and Scotland and taking the top 500 or so remaining places to analyze.
The data is in the following format:

1	1	London	London	London	8,907,918
2	1	Birmingham	West Midlands	West Midlands	1,153,717
3	1	Glasgow	Glasgow	Scotland	612,040
4	1	Liverpool	Merseyside	North West	579,256
5	1	Bristol	Bristol	South West	571,922
6	2	Manchester	Greater Manchester	North West	554,400

So we will need to clean it, removing unnecessary columns and all the entries which do not relate to England.
We will then use a geocoder service to find the latitude and longitude of the center of these localities. This will provide us with information such as:

```
Birmingham, UK
Birmingham, West Midlands Combined Authority, West Midlands, England, United Kingdom
1 City/Town              Birmingham
Ceremonial county    West Midlands
latitude                 52.4797
longitude                -1.90269
Name: 1, dtype: object
Liverpool, UK
Liverpool, North West England, England, United Kingdom
3 City/Town              Liverpool
Ceremonial county    Merseyside
latitude                 53.4072
```

```
longitude            -2.99166
Name: 3, dtype: object
Bristol, UK
Bristol, City of Bristol, South West England, England, United Kingdom
4 City/Town          Bristol
Ceremonial county    Bristol
latitude             51.4538
longitude            -2.5973
```

Then we will use FourSquare venue explorer to find what venues are available within a reasonable distance to the city center. The information received will be similar to the following:

0	A	52.449601	-1.819154	Costa Coffee	52.446411	-1.822441	Café
1	A	52.449601	-1.819154	Wilko	52.446645	-1.823626	Store
2	A	52.449601	-1.819154	Sainsbury's	52.445749	-1.820319	Supermarket
3	A	52.449601	-1.819154	Subway	52.446418	-1.823159	Restaurant
4	A	52.449601	-1.819154	The Spread Eagle (Wetherspoon)	52.446488	-1.819978	Pub
5	A	52.449601	-1.819154	Quality Hotel Birmingham Airport	52.449284	-1.819689	Hotel
6	A	52.449601	-1.819154	Argos	52.446453	-1.822206	Store

We will use this final data for each location center to cluster the cities and find similar types of cities which will appeal to different type of tourists (or none at all!)

Methodology

Initial experiments looked at how accurate the data was at representing English localities. It was found that a number of localities could not be understood by FourSquare and on further investigation it was found that this was because the names were either misspelled or ambiguous (representing localities in Scotland instead of England). The data was cleaned to correct these errors:

```
"Corsett"="Consett"
"Gerrards CrossChalfont St Peter"="Chalfont St Peter, Gerrards Cross"
"HartleyLongfield"="Hartley Longfield"
"Frampton CotterellWinterbourne"="Frampton Cotterell, Winterbourne"
"Whittlesley"="Whittlesey, Fenland"
"Desbrorough"="Desborough"
"Ferryhill"="Ferryhill, County Durham"
"Stanley"="Stanley, County Durham"
```

The data also contained unnecessary columns such as a graphical representation, population, and county name which were removed.
After this the data from FourSquare was analysed and it was found that for very small localities the FourSquare data was too sparse to be of any use and therefore the data was filtered to take the top 500 localities instead of the top 1000. Furthermore dummy data with the value “no venue” was inserted where no venue was found. Even so, the FourSquare data appeared to be very unreliable and inexplicably scarce, sometimes returning only a handful of venues for very busy city centres.

After this the k-means clustering algorithm was used to attempt to cluster the data in a meaningful way. K-means was used as it is well-suited to this type of data. To verify the

meaningfulness of the clustered data, a subset of the data was analysed using Google maps to see if effectively the locations shared any similarities.

This highlighted another problem with FourSquare: the inconsistency of nomenclature, with similar venues being given different names, e.g. “art gallery” and “art museum”, “coffe shop” and “cafe” being different categories. Furthermore, things like airports and airport lounges were separated, which didn’t help classification, as were different types of restaurant. To tackle these inconsistencies the data was once again manipulated to give meaningful characteristics using the function:

```
def convert_the_name(the_name):
    if(the_name in ["Sandwich Place", "Pizza Place", "Stakhouse", "Burger Joint"]):
        return "Fast Food"
    if(the_name=="Coffee Shop"):
        return "Café"
    if("Store" in the_name):
        return "Store"
    if("Restaurant" in the_name):
        return "Restaurant"
    if("Art" in the_name):
        return "Art"
    if("Museum" in the_name):
        return "Art"
    if("Bar" in the_name):
        return "Bar"
    if("Auto" in the_name):
        return "Auto"
....
```

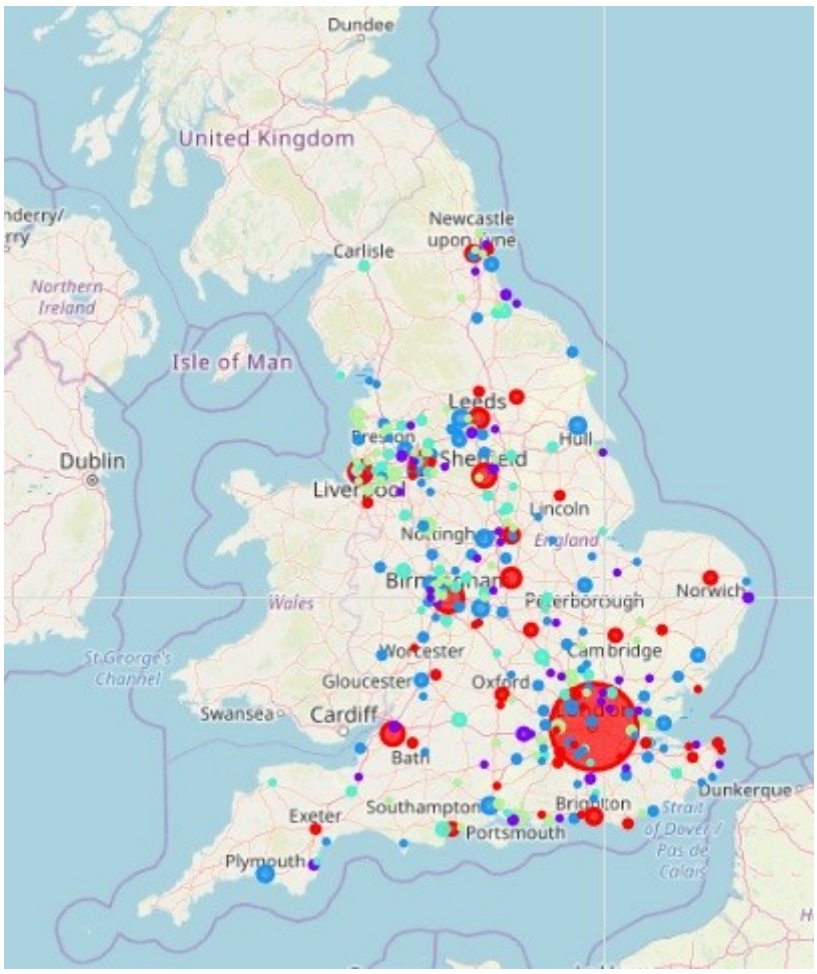
K-means was then used again to classify the data, using different numbers of clusters. The Elbow Method was used to get an idea of the optimal number of clusters. This gave k=3 as the optimal number, but as this seemed very low experiments were carried out using both k=3 and k=5 too see which gave better results.

Results

Venue data was gathered from FourSquare. While there was an abundance of information for particular locations it was evident that the venue information was incomplete and that a large number of venues were not contained within the FourSquare data.

First the data was grouped with k=5. The data was clustered into five groups and mapped onto a map of England. The algorithm divided English Cities and Towns into 5 distinct categories. No city was left without any category.

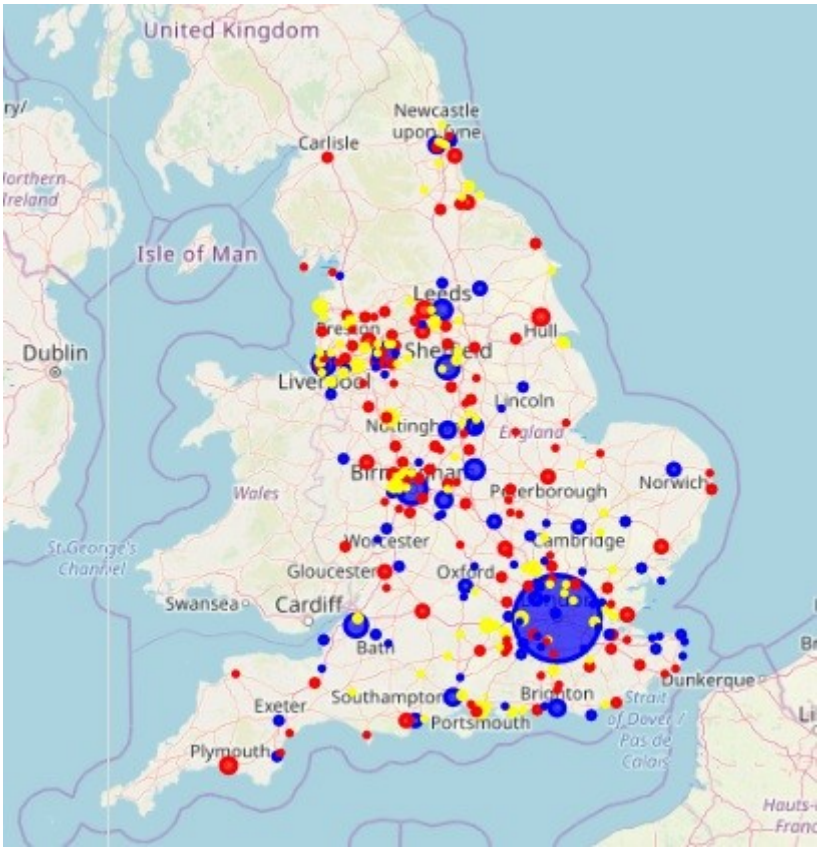
This gave the following map:



Examination of the characteristics of the data clusters:

1. The first cluster appears to contain big cities with a vibrant restaurant and coffee culture, perhaps with Art and Concert venues as well as some smaller towns with the same vibe.
2. The second cluster appears to be smaller "downmarket" towns with a mix of various shops, pubs and cafes and for which FourSquare has incomplete data
3. The third category is made of slightly more upmarket towns with an interesting mix of various types of shops, restaurants and cafes.
4. The fourth category appears to be towns where the main attraction is an abundance of mainstream shops, complemented by fast food places, pubs and cafes
5. The final category encompasses towns where inexplicably there is very little FourSquare data. This inclues large towns such as Portsmouth, Luton and Blackpool which certainly don't lack in venues!

The second experiment looked at clustering with k=3, resulting in the following map:



Examination of the characteristics of the data clusters:

1. The first category appears to be towns where the main attraction is an abundance of mainstream shops, complemented by fast food places, pubs and cafes
2. The second cluster appears to contain big cities with a vibrant restaurant and coffee culture, perhaps with Art and Concert venues as well as some smaller towns with the same vibe.
3. The final category encompasses towns where inexplicably there is very little FourSquare data. This inclues large towns such as Portsmouth, Luton and Blackpool which certainly don't lack in venues!

Discussion

The results confirmed to a certain extent the view that cities and towns in England have an "identikit" feel, with the same sort of amenities available throughout the country. Clustering in 5 or in 3 groups didn't really result in any noticeable difference. In the case of the 5 clusters, three of the clusters didn't seem to really be distinguishable so the division in 3 clusters really did make sense, as predicted by the Elbow Method.

However, one of the clusters was of towns for which there was very little FourSquare data, which was quite surprising, given the size of some of the localities. A close examination of the data revealed that the FourSquare data was very unreliable, with a very large number of important venues missing even in big cities where there appeared to be a good amount of data.

It therefore appears that FourSquare may not be the best place to gather venue data for this sort of analysis.

Conclusion

We started this report asking if cities and towns in England are really "identikit", bland city centres. The answer to this seems to be yes, as far as venues such as shops, cafes etc. is concerned. However the FourSquare data was hopelessly incomplete, with very little in terms of landmarks and very incomplete data as regards shops and other amenities. Indeed for a number of cities there inexplicably appeared to be very little data indeed.

So where should a holidaymaker go if they want to have a good time? If you want to discover something new perhaps you should try one of the many cities for which FourSquare has no data. Which cities are best for "coffee culture"? The big ones such as London, Birmingham, Manchester, Leeds. Which are best for a more traditional feel? Smaller ones such as Bradford and Plymouth.

Is there some objective data that the general public (or tour operators) can use to make recommendations instead of relying on the often deceitful marketing material issued by the cities themselves or by the cities' tourist boards? Can we use data science to tell us whether it is worth visiting an English city without being disappointed or mislead by advertising? Only partially. Unfortunately the dataset we used (FourSquare) is woefully incomplete and cannot be relied upon fully for this task.