CAPTSONE PROJECT

Identifying Potential Locations for a New Restaurant Opening

IBM - APPLIED DATA SCIENCE

COURSERA

Table of Contents

1) Introduction

Identifying the best location for a business can be a difficult exercise. This is particularly true for restaurant chains which might not have prior knowledge to local markets in specific areas. Before carrying out more in-depth investigations to find the best suitable place, it would be necessary to already pre-select potential areas of interest. Using some data analysis technique on available data could help in the initial decision-making process.

In this analysis we will use a fictitious restaurant chain wanting to open a new restaurant in Birmingham (UK).

Business Problem
A steakhouse chain would like to identify potential locations to open a new restaurant in Birmingham (UK). Their concept is trendy and attracts a fairly young clientele with their target customers being in age group 21-39.
They have short listed three main requirements:

- They would like locations close from where their target customers (age group 21-39) reside but also near areas with a high volume of working population. The aim is to maximise the catchment area insuring a mix of local residents and workers. Locals would likely be more active during the evenings in a working week and at the weekends whereas workers would be more active during lunchtime and potentially some evenings in a working week.

- The areas around the locations should be ideally where the spending in restaurants by the residents in the target group is high.

- The potential locations should be places with already established restaurants. In the restaurants market, the more the offer the more the customers. Only niche restaurants (high end usually) can afford to be “out of the beaten tracks”.

How can we help this company to short list the best areas in Birmingham?
We will use some data from the ONS (Office of National Statistic) together with data from Foursquare and apply some data science methodologies such as clustering to identify restaurant hotspots in Birmingham and short list locations that appear suitable.

2) Data

a. What data will be used

For this analysis data from the ONS and Foursquare will be used.

ONS Data
The ONS publishes every 10 years since 1981 the results of the UK population census. The data available provides a view on the population demographics, the spending habits, the working population and much more. It is used by companies and public bodies to perform economic forecasts.
The ONS also made available a postcode mapping table linking the census data to geographic areas. Hence it is possible to combine the census data with geographic data to identify the population demographics for defined locations.
For this analysis several datasets from the ONS will be used:

- UK Postcode Directory
- UK weekly expenditures by household
- UK population by OAC (see the UK postcode directory overview for definitions)
- UK LSOA shape files

This data will be used in conjunction with the Foursquare API which will provide the point of interests (venues) in Birmingham.

b. Source of the data

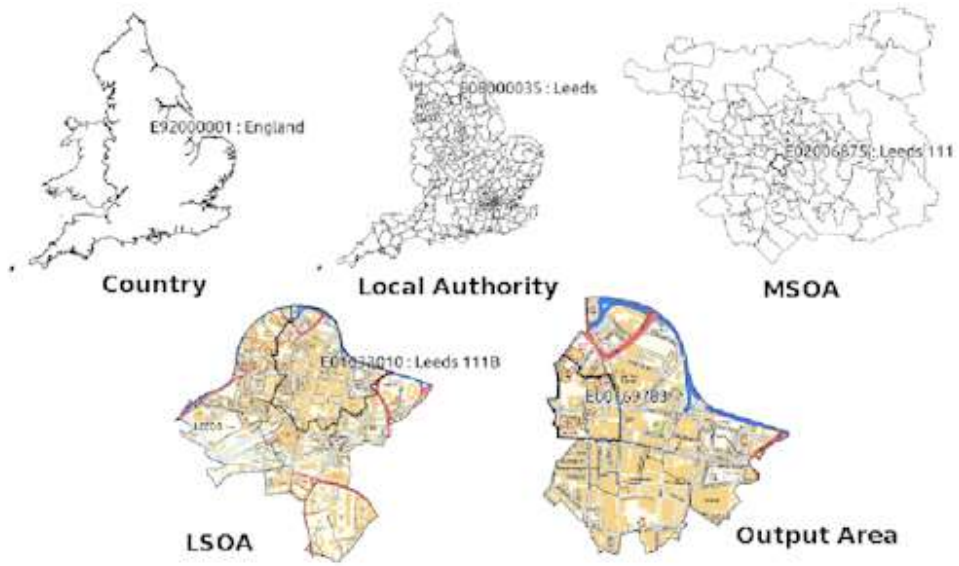
i. UK Postcode Directory

Source
The data is available at this location: https://geoportal.statistics.gov.uk/datasets/ons-postcode-directory-february-2019

Use
This data will be used to map the population of Birmingham by socio economic groups and spending but also to map the working population. The original dataset is at Output Area level which is the most granular data point for the census geography. This level is too small therefore the data will be aggregated at Lower Super Output Areas (see below).

Data overview
It provides a mapping of the UK postcodes to various geographic areas. For the census data, specific geographic boundaries were created. These areas are called Census Output Areas. The most granular level is the Output Area (OA) and the highest is the Country.
Each postcode belongs to an Output Area. For this analysis we will use the LSOA level (Lower Super Output Area.

Overview of the Census boundaries:



The dataset is made of over 30 fields. Here is a snapshot of the main fields:

pcds	oa11	lsoa11	msoa11	oac11	lat	long
B1 1AA	E00175658	E01033625	E02006899	2B2	52.47666	-1.90354

- pcds is the postcode
- oa11 is the Output Area code
- lsoa11 is the Lower Super Output Area code
- msoa11 is the Middle Super Output Area code

ii. Mid-2017 Population Estimates (West Midlands)

Source
https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/censusoutputareaestimatesinthewestmidlandsregionofengland

Use
The data will be use create the age groups for the population of Birmingham.

Data overview
The data provide the population estimate (2017) at oa11 level and by age (from 0 to 90+ by increments of 1 year).

iii. UK weekly spending by household

Source
https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/expenditure/datasets/averageweeklyhouseholdexpenditurebyoutputareaclassificationoacgroupuktablea52

Use
This data will provide an estimate of the spending in restaurants and hotels in an area by calculating the average weekly spending by person in an output area. The ONS only provides spending for restaurants and hotels together hence this will be use as a proxy for restaurant spending.

Data overview
Provides the weekly household expenditure on good and services in the UK by area and by output area category group. The file provides the average number of persons per household.

OAC group	OAC code	Weekly Restaurants and Hotels expenditure by household (in £)
Farming Communities	1A	51.50
Rural Tenants	1B	51.80
Ageing Rural Dwellers	1C	39.10
Students Around Campus	2A	71.70
Inner-City Students	2B	62.30
Comfortable Cosmopolitans	2C	45.50

Each Output Area is assigned an OAC group which broadly describe the socio-economic groups (OAC Group).

iv. UK working population (West Midlands)

Source
The dataset is available on the NOMIS website which is part of the ONS.
https://www.nomisweb.co.uk/query/construct/submit.asp?menuopt=201&subcomp=

Use
The data will provide a view on the working population of Birmingham by LSOA level. The working population is people who are working in the area but are not necessarily living there.

Data overview
All usual people aged 16 to 74 in employment in the area the week before the census.

v. UK lsoa11 shape files

Source
https://geoportal.statistics.gov.uk/datasets/lower-layer-super-output-areas-december-2011-full-clipped-boundaries-in-england-and-wales

Use
This file contains the geographic shapes of the LSOA in the UK.

vi. Foursquare API

The data will be use to get the venues in Birmingham.

c. Data preparation

The datasets were already clean and very little data cleansing was required. To make the process easier some files were combined in Excel (population estimate by age group and latitude/longitude form the postcode directory).

The shapefile from the ONS was reduced to Birmingham using Geopandas. Displaying all the UK shapes created stability issues when rendering in Folium.

i. ONS population data

The ONS population was imported from a csv file. The data only required to add up two age buckets together in order to get the target population (21-39 years old)

ii. Restaurants & Hotels spending

Some preparation was done in Excel before importing the data since the original file formatting would have made the importing too time consuming to code.

iii. Working population

The working population was imported as is from a csv file.

iv. Shape Files

The shape files required more manipulations. Most the data preparation was spent identifying the best way to handle the data.
The shape file was imported using Geopandas to create a geodataframe. Then the projection needed to be changed from Ordonnance Survey Grid to coordinates (latitude/longitude).
Because of issues with the rendering in Folium (multiple crashes), a filter was applied to only return the polygons for Birmingham. From this filtered geodataframe a GeoJSON file was created so it could be used in Folium.

The number of lsoa11 was limited to a 3km radius from the city centre. The area was too wide and there was little gain in trying to pull more data for locations of no interest. Also, this helped reduced the number of calls made to the Foursquare API.

3) Methodology

The first step was to explore the ONS dataset to identify the areas with high spending from the target group (age between 21-29) and with a high-volume working population. This allowed to later evaluate the potential locations.
A second step was to bring all the venues from Foursquare and create a heatmap of restaurants in Birmingham so hotspots (high concentration of restaurants) could be visualised.
A third step was to cluster the restaurants using a geographic clustering algorithm (DBSCAN) based on their locations and then calculate the centroids for each cluster (average latitude longitude) in order to get all the venues from Foursquare. The aim was to reduce the radius to 00m so venues were more specific to each cluster.
Finally, a k-means clustering was run to understand the main characteristics of each location.

a. Explanatory data analysis

The main dataset had 5,895 rows (corresponding to each Output Area in Birmingham) across 22 columns. Once the area was limited to a radius of 3km from the city centre, the number of observations reduced to 1,182 across 7 columns.
The dataset was then aggregated at lsoa11 level using a group by function.

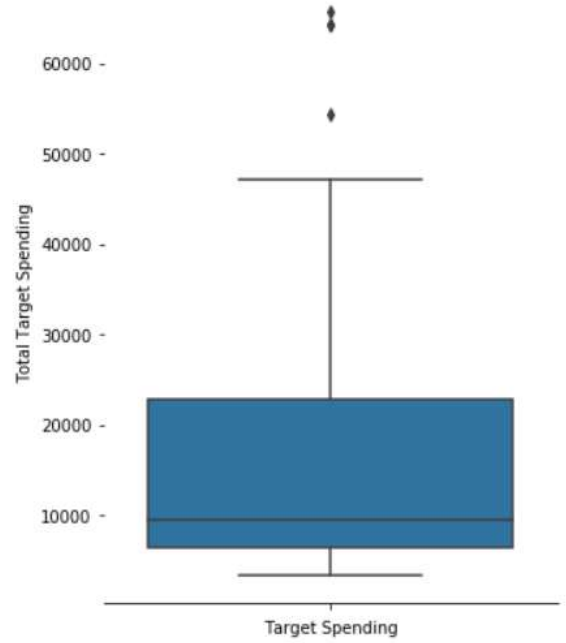
Summary statistic of the LSOA dataset:

	Target_Age	Total Target Spending	Total Spending	Pop_over_15	Total Target Spending £k	Total Spending £k
count	75.000000	75.000000	75.000000	75.000000	75.000000	75.000000
mean	768.866667	17326.081333	33999.561333	1634.573333	17.326667	34.006667
std	382.914264	15995.566048	30262.537397	810.237105	15.995906	30.262707
min	283.000000	3452.600000	10321.200000	839.000000	3.500000	10.300000
25%	483.500000	6380.500000	15250.000000	1229.500000	6.400000	15.250000
50%	639.000000	9568.400000	24946.100000	1365.000000	9.600000	24.900000
75%	971.000000	22845.600000	45236.500000	1774.000000	22.850000	45.250000
max	1932.000000	65851.200000	194016.800000	6521.000000	65.900000	194.000000

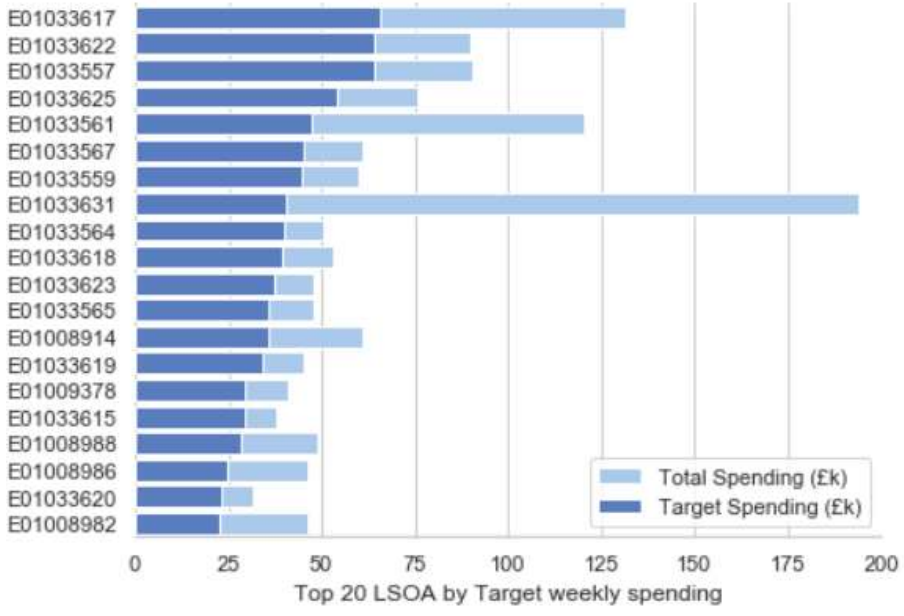
There are 75 observations corresponding to the total LSOA. The mean weekly spending in hotels and restaurants was £17,326 and the mean population of the target group was 768.

Target population Restaurants and Hotels spending

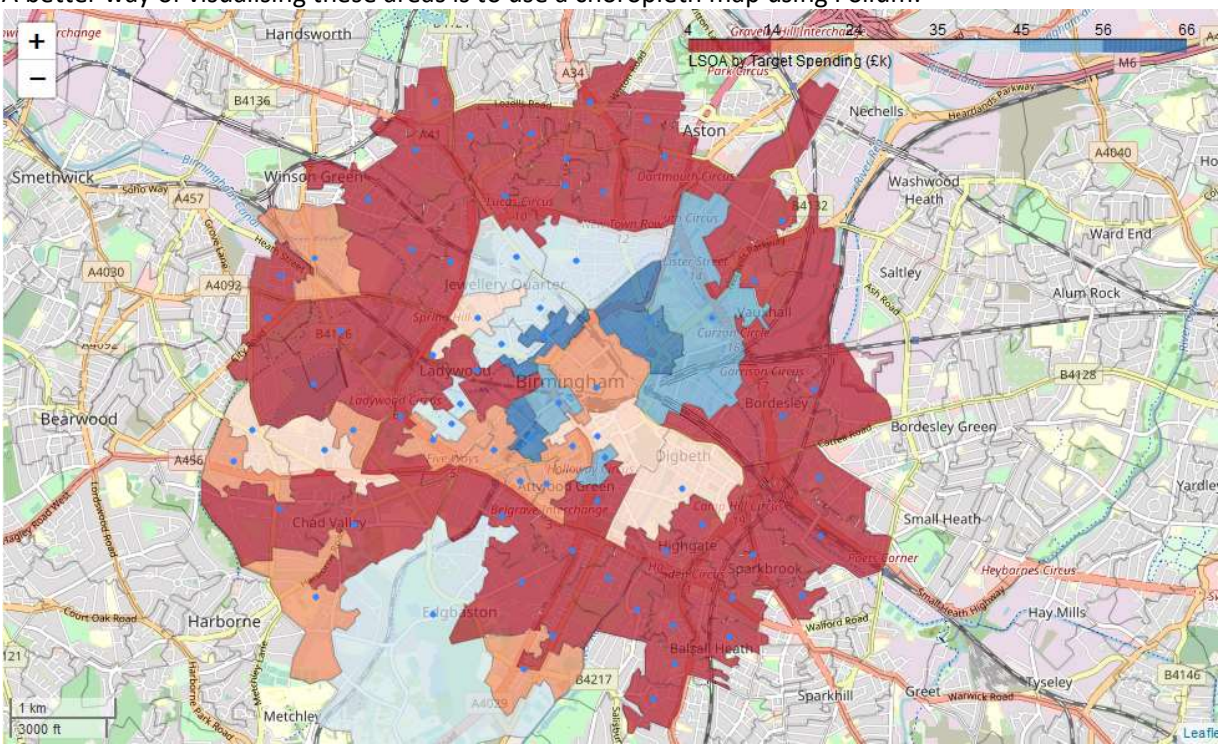
Looking at the spending distribution of the target spending, three LSOA are outliers with much higher spending than the others:



By plotting the data on a bar graph these three lsoa are E01033617, 622 and 557.



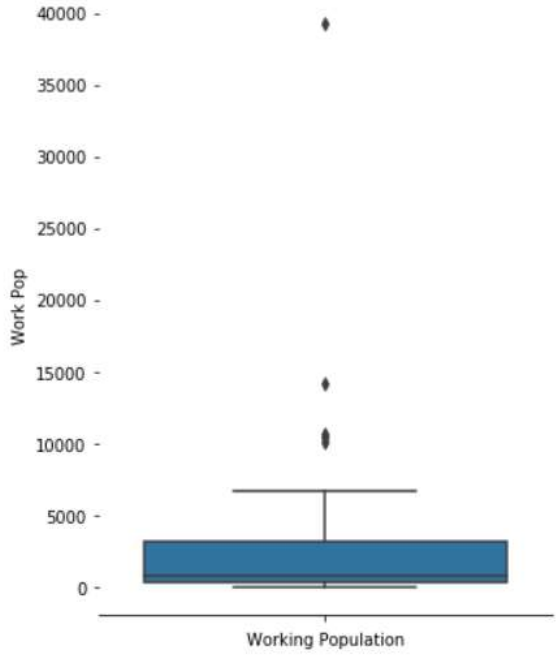
A better way of visualising these areas is to use a choropleth map using Folium:



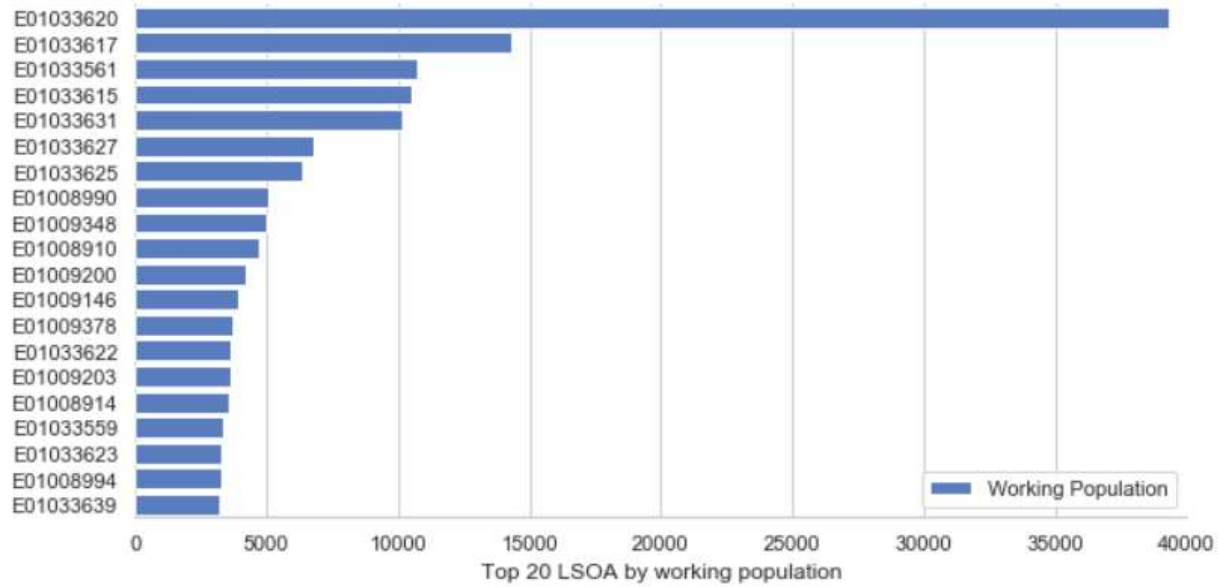
The locations with the highest spending are mainly around the city centre of Birmingham. The three outliers identified are showing in the darker shade of blue.

Working population Restaurants and Hotels spending

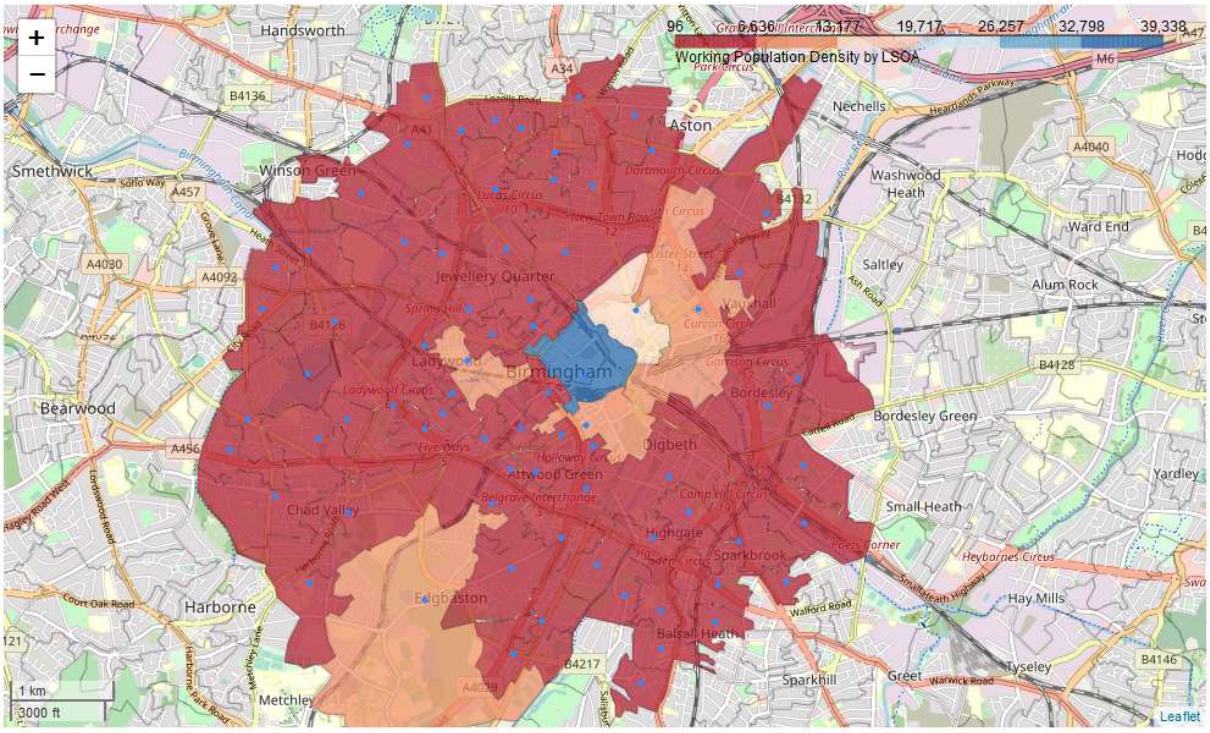
There are outliers but this is expected since some locations will have a high concentration of working places (city centre for example).



E01033620 has got the highest volume of workers:

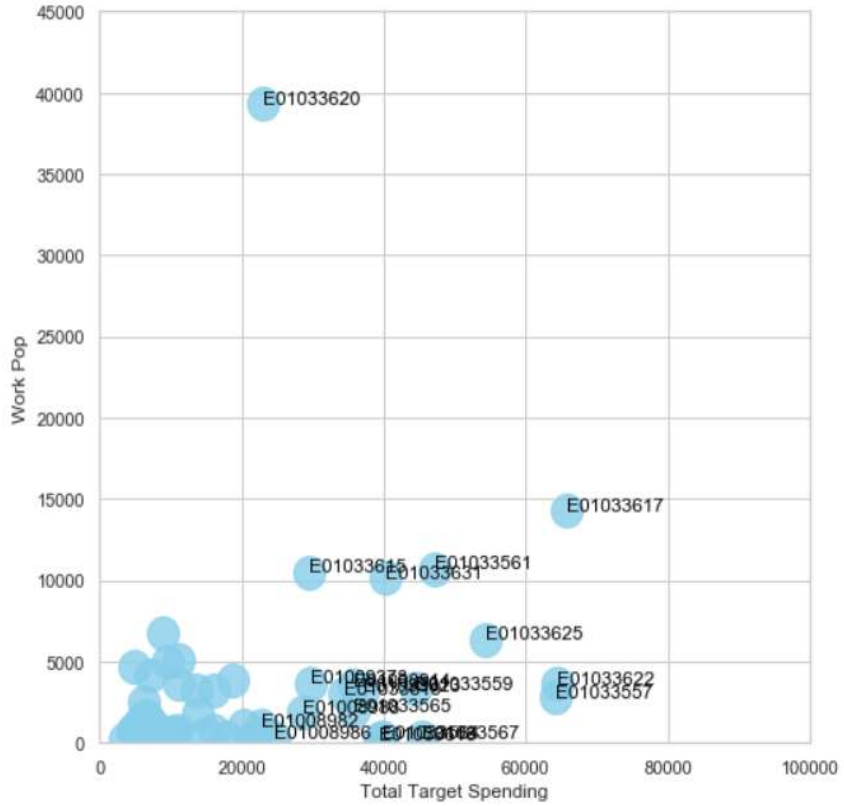


To visualise the distribution, the working population was plotted on a map:



The city centre and lsoa around the city centre have the highest working population.

To identify areas with high spending and high working population a scatter plot was created:



A few locations seemed to have high spending and also a fairly high working population (E01033617 in particular).

Foursquare data

Venues
Using the API, 200 venues in a radius of 500m of each centroid of the 75 LSOAs were extracted.
1,465 records were returned:

```
print(Birmingham_venues.shape)
Birmingham_venues.head()
```

```
(1465, 7)
```

	lsoa11cd	LSOA Latitude	LSOA Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	E01033617	52.484849	-1.891885	Boston Tea Party	52.483028	-1.893479	Café
1	E01033617	52.484849	-1.891885	The Jekyll & Hyde	52.483560	-1.895151	Pub
2	E01033617	52.484849	-1.891885	Scruffy Murphy's	52.482146	-1.892448	Rock Club
3	E01033617	52.484849	-1.891885	The Sir Doug Ellis Woodcock Sports Centre	52.486192	-1.886339	Gym / Fitness Center
4	E01033617	52.484849	-1.891885	Philpotts	52.483104	-1.896114	Sandwich Place

Restaurants venues
In order to identify restaurants only, the venue categories were grouped (so only unique categories were returned) and parent categories were created. This step was done manually in Excel. Essentially each

category was assigned a main category. For example, Italian Restaurant was classified as Restaurant. This allowed to create a new dataframe including restaurants only.

```
# Restaurant category dataframe
cat_rest_filter =  cat['Main Category']=='Restaurant'
cat_rest=cat[cat_rest_filter]
print(cat_rest.shape)
cat_rest.head()
```

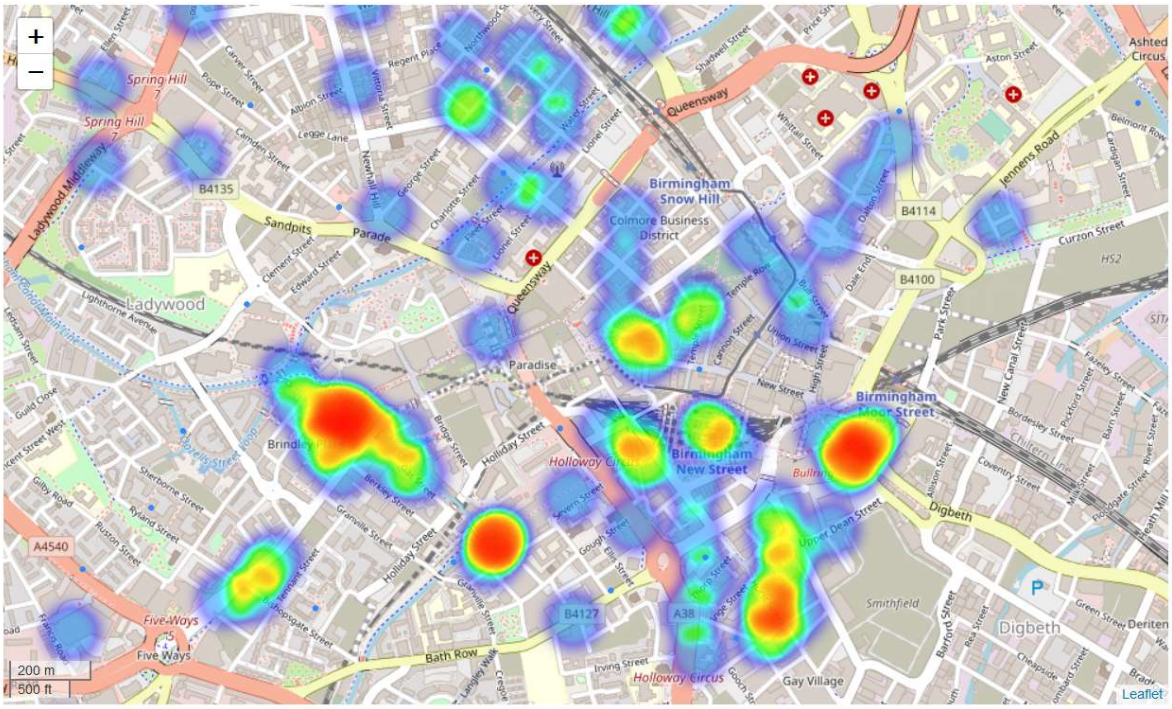
```
(45, 2)
```

	Venue Category	Main Category
0	American Restaurant	Restaurant
2	Argentinian Restaurant	Restaurant
4	Asian Restaurant	Restaurant
6	BBQ Joint	Restaurant
15	Brazilian Restaurant	Restaurant

The data had to be deduplicated since each LSOA centroid pulled 200 venues located in a 500m radius hence a venue could appear in multiple LSOA if in a radius of less than 500m from multiple LSOA. Example below:

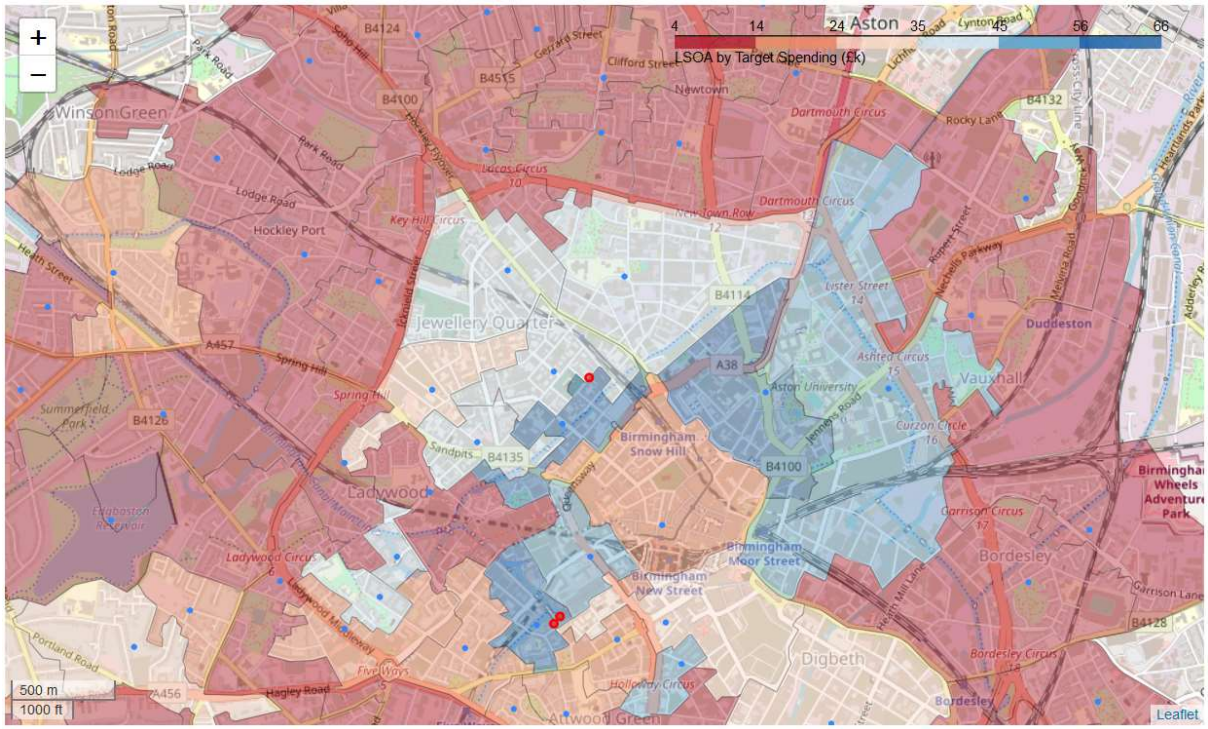
	lsoa11cd	LSOA Latitude	LSOA Longitude	Venue	latitude	longitude	Venue Category
0	E01033617	52.484849	-1.891885	GDK German Doner Kebab	52.481257	-1.895923	Doner Restaurant
1	E01033620	52.479151	-1.899140	GDK German Doner Kebab	52.481257	-1.895923	Doner Restaurant
2	E01033617	52.484849	-1.891885	Big John's	52.482091	-1.894325	Fast Food Restaurant
3	E01033617	52.484849	-1.891885	Caspian Pizza	52.484160	-1.892128	Fast Food Restaurant
4	E01033557	52.474878	-1.908052	Pitstop	52.474977	-1.914555	Fast Food Restaurant

This resulted in 169 unique records. To visualise the distribution and identify the restaurants hotspots, a heatmap was created using Folium:



The map shows distinct clusters mostly on the south of the city centre (if using the train line as x axis). One area in particular on the left seems to have a very high concentration of restaurants.

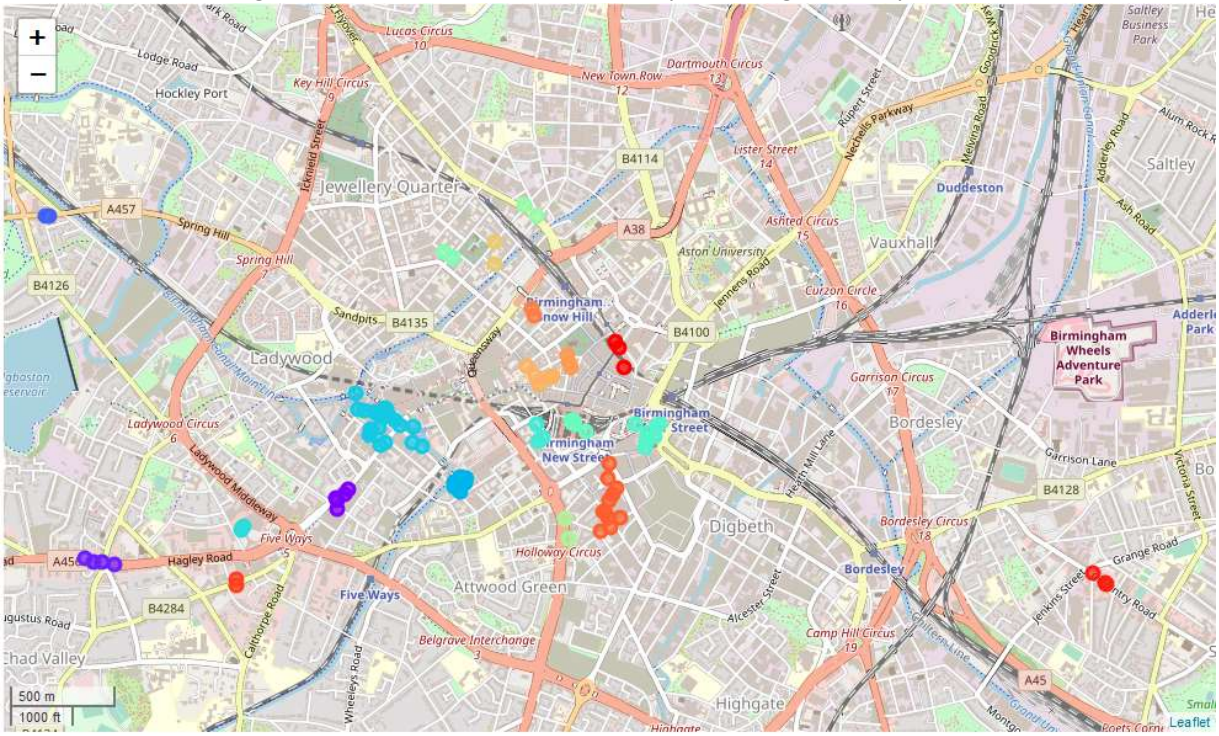
Direct competitors (Steakhouse restaurants) were plotted on a map to identify their locations:



E01033557 has already 2 steakhouse restaurants whereas E01033627 has got none.

b. Restaurant Geographic Clustering

In order to do a k-means clustering it was necessary to create geographic clusters of the restaurant's hotspots.
As we have seen previously, because of the way the data was pulled from the API, venues are duplicated across multiple locations. A k-means clustering would not provide much insights as the clusters would be too similar.
DBSCAN clustering (Density-based spatial clustering of applications with noise) was used. From these clusters the centroids (avg lat/long) were taken and the Foursquare API was queried again but limiting the radius to 200m.
The DBSCAN clustering created 25 distinct clusters broadly matching the hotspots:



c. K-means clustering

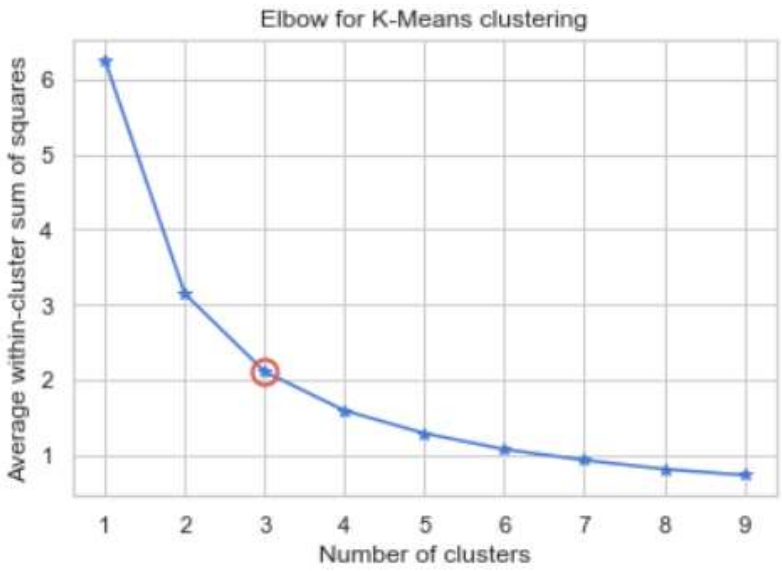
In order to run the K-means clustering, it was necessary to return the centroid of each clusters.

Cluster Labels	latitude	longitude
0	52.480805	-1.895652
1	52.474614	-1.915063
2	52.471929	-1.931923
3	52.502859	-1.906620
4	52.486599	-1.935536

These centroids were used to return the Foursquare venues within a 200m radius. 109 categories were returned.

	lsoa11cd	LSOA Latitude	LSOA Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Main Category
0	0	52.480805	-1.895652	Loki Wine Merchant & Tasting House	52.482015	-1.897398	Wine Shop	Wine Shop
1	0	52.480805	-1.895652	Tilt	52.479715	-1.896372	Gaming Cafe	Gaming Cafe
2	0	52.480805	-1.895652	Monty's Sandwich Bar	52.482102	-1.896516	Sandwich Place	Sandwich Place
3	0	52.480805	-1.895652	Bistro 1847	52.481777	-1.896972	Bistro	Bar
4	0	52.480805	-1.895652	GDK German Doner Kebab	52.481257	-1.895923	Doner Restaurant	Restaurant

The restaurant category was removed so they wouldn't be categorised by the K-means algorithm.
An Elbow method was used to identify the optimum number of clusters:



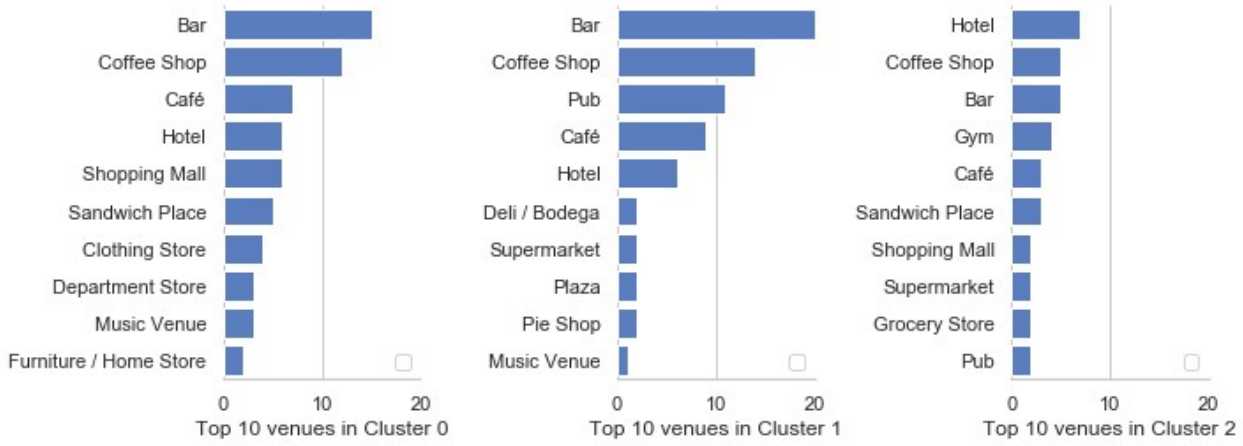
The optimum number of clusters was 3. Using this value, the K-Mean clustering was run.

4) Results

The main objective was to create meaningful clusters using k-means to get a better understanding of each restaurant hotspots characteristics and combine this information with the ONS data investigation findings.

First, we established that areas with high spending and high volume of workers were close to the city centre. This enabled to understand better the type of potential customers in restaurant hotspots.
The second step was to locate the restaurant hotspots. This was done using the DBSCAN algorithm.
The last step was to run a k-means on these hotspots.

To understand the characteristics of each cluster the volume of venues per cluster was returned:

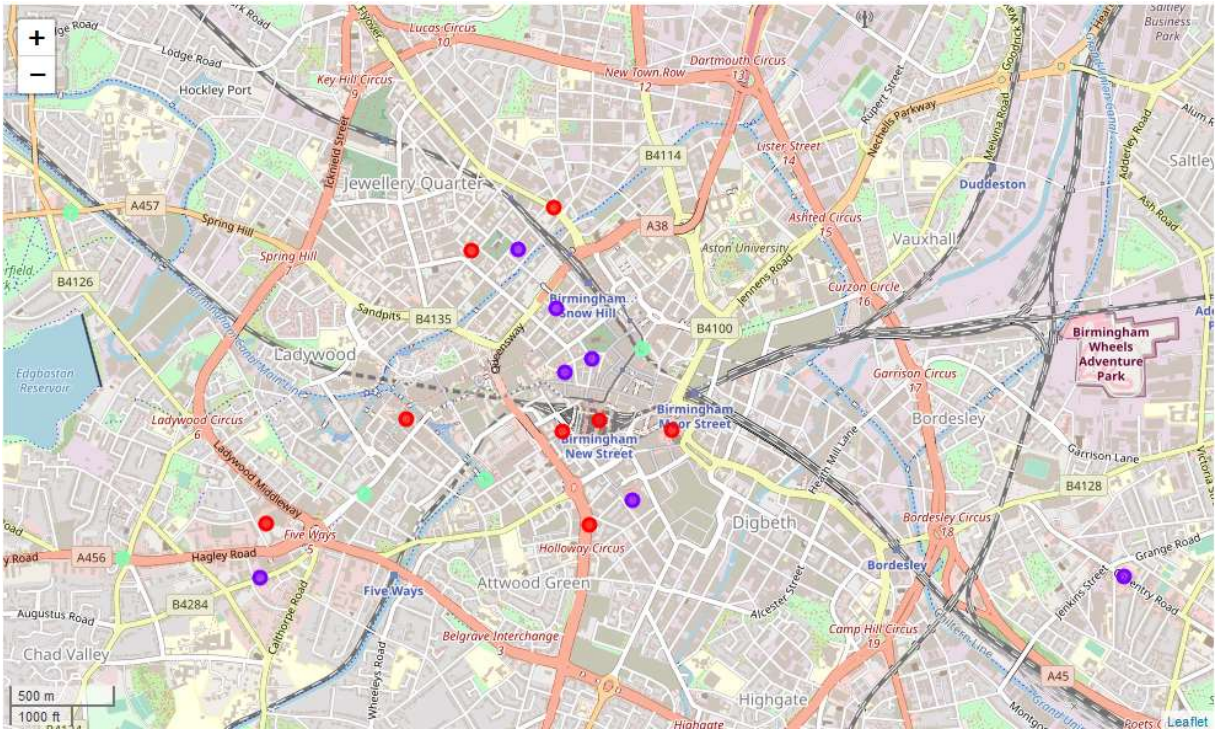


Cluster 0
High number of bars, coffee shops and cafés. It has hotels on fourth place followed by places for shopping with a few shopping centres and other stores. There are also a few music venues

Cluster 1
High volume of bars followed by coffee shops. Pubs come in third place.

Cluster 2
This cluster has got hotels in first place then coffee shops and bars. Gyms are in fourth position. This cluster has a few shopping centres but not as many as cluster 0. It has sandwich places, supermarkets and grocery stores with pubs coming last.

Then the clusters were displayed on a map.



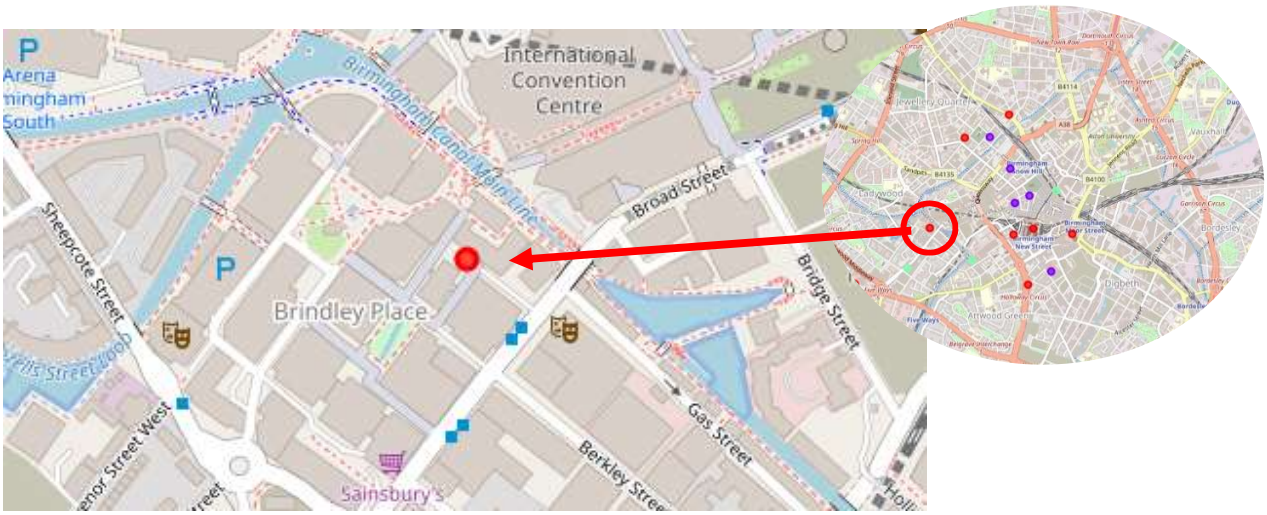
By combining the various pieces of information, recommendations of potential locations were made (see next section).

5) Discussion

a. Recommendations

Based on these clusters and on the analysis done on the ONS data three potential locations were identified.

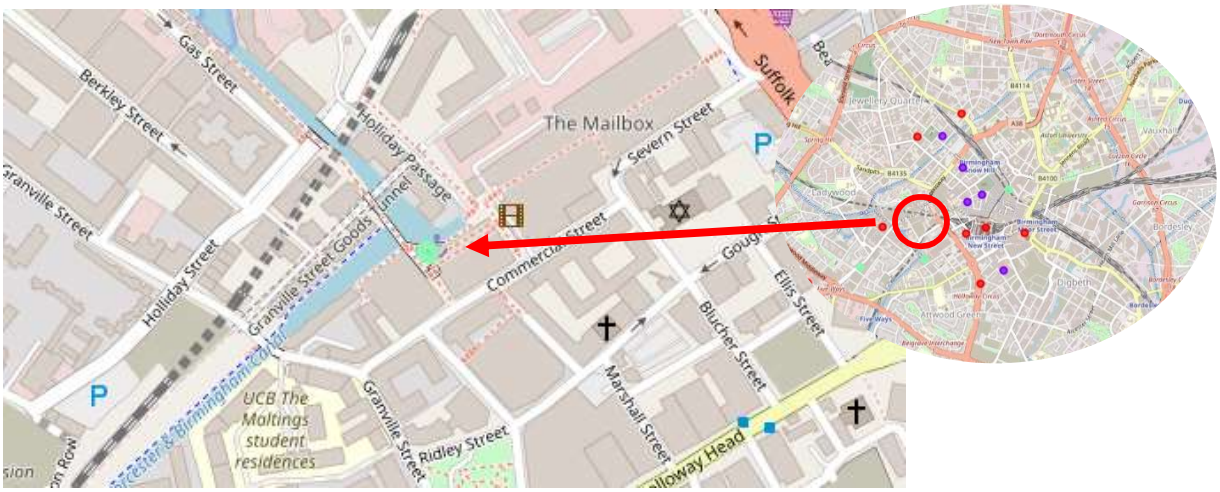
Option A: Around Brindley place (DBSCAN cluster 8 and k-means cluster 0)



The Pros:
This location has got a high volume of restaurants, bars, cafés and stores that are likely to attract the type of customers our fictitious steakhouse chain would like to target. Combined with the presence of stores and hotels in the area, foot traffic must be fairly high. This location is a close to a good catchment area made of residents in the target age group (21-29) having a fairly high weekly spending in restaurants and hotels.

The Cons:
However, it would need to be considered the high concentration of restaurants in the area. Is the local market saturated? How well competitors are doing? The potential high foot traffic must also equate to higher rents but this is likely to be the case with any locations near the city centre. Despite being close to good catchment areas it is at the boundaries of an area with lower spending. The workplace population is also lower than other locations

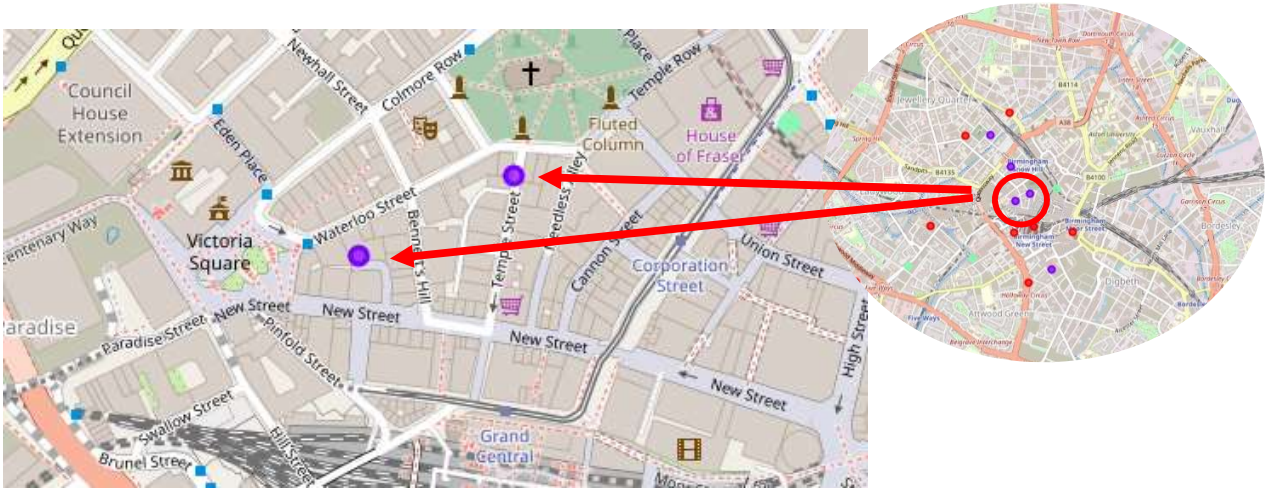
Option B: Near the Mailbox (DBSCAN cluster 7 and k-means cluster 2)



The Pros:
This appears to be the second place in terms of restaurants density therefore foot traffic should be high. This location is in an area where residents in the target group have a high spending in restaurants and hotels. It seems to be situated close to shopping centres as well.

The Cons:
One of the main issues with this location is that there are already two steakhouse restaurants. Would a third one be viable? Also, this area does not seem to have as much going on as in Option A (and C that will be covered below).

Option C: Near Victoria Square (DBSCAN cluster 19, 20 and k-means cluster 1)



The Pros:
High number of bars, pubs, cafés and hotels. Because right in the middle of the City Centre, very likely to have high foot traffic. The city centre is surrounded by areas with residents in the target group age spending the most in restaurants and hotels. It also has the highest volume of workers in the area. There is some competition already there but it does not seem as concentrated than in option A and B.

The Cons:
Despite having a high volume of workers, there are very few residents. Are customers drawn into the city centre at the weekends? Also, what is the activity in the evening once workers are back home?

b. Other considerations

This analysis has some limitations: the data used from the ONS is 8 years old (it is refreshed every 10 years) and the venue data is only collected by users of Foursquare. We do not know the customer penetration Foursquare has got in the UK hence the data could be biased towards a certain category of consumers therefore some venues might not show. This approach could also be improved by using additional data such as the foot traffic which could be sourced from Foursquare.
More accurate spending data would be very valuable together with more detailed population demographics. Some companies do provide such data but at a (high) cost. Usually these are companies that have loyalty programmes that allow them to collect consumer data. Other data providers could be banks who are already starting to investigate the potential of their spending data. Having such data would allow to use other technique to predict the spending in a location for example using a geographic regression.
Finally, one aspect to consider is data privacy. The multiplication of data providers through APIs and readily available data could be in breach of data protection laws (GDPR in Europe). Using such data with a commercial lens would require care in terms on how and what to process.

6) Conclusion

As we have seen a combination of various techniques can be used to identify potential locations for a new restaurant. They allow without a deep knowledge of an area to pre-select what could be the best places in terms of competition and catchment.
However as very often answers are not necessarily obvious and there is not one perfect solution but a few options to consider. This is down to the data analyst/data scientist to understand the business problem and requirements and translate their findings in a meaningful way for the stakeholders.