Captsone Project

Identifying Potential Locations for a New Restaurant Opening

Introduction

- In this project we will look at methods that can be used to help identifying potential locations for opening a new restaurant using point of interests (POI) data from Foursquare and data from the UK Office of National Statistics (ONS)
- The approach would be of interest for investors or restaurant chains willing to open a new restaurant in a city/place where they do not know the market

Problem statement:

- A steakhouse chain would like to identify the catchment areas where the spending in restaurants is the highest for residents in their target customer age group (21 to 39 years old). They are also interested to understand any correlations between these areas and the worker population (as people working in or close by)
- The aim is to identify locations with the highest potential customer base being a mix of residents and workers (who are not necessarily living in the area)
- They would be looking at places where there is already competition (hotspots) as it would equate to established locations where customers go.

Data Sources

- For this analysis several datasets from the ONS based on their 2011 census was used:

  - UK Postcode Directory

  It provided a mapping of the UK areas to latitude/longitude as well as a grouping of the census data geographic areas. It also had the economic class for each area (called Output Area Group)

  - UK weekly expenditures by household

  This data contained the Restaurants and Hotels expenditure by area. It was used as a proxy for the total spending in restaurants in Birmingham

  - UK population by Output Areas

  Data about the total population by Output Area with the population age group

  - UK Lower Super Output Areas shape file

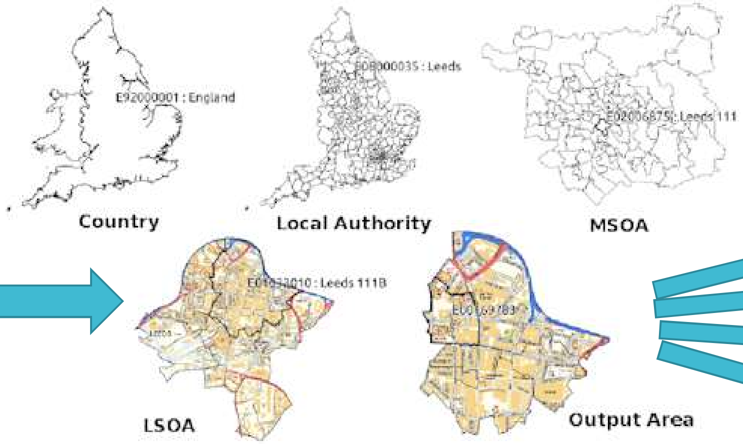
  This file contained the geographic shapes that were used to create choropleth in Folium

- Finally the Foursquare API provided the point of interests (venues) in Birmingham.

Data Sources | ONS 2011 Census Overview

- The Census data contains information about the UK population such as volume of people living in an area, age group, total expenditure, volume of workers etc…
- A census is carried out every 10 years; the latest one was in 2011.
- Although the some data is from 2011, some is updated every year such as the population estimate or the household expenditure
- The most granular level of the census data is the Output Area which group households that have similar socio-demographics. For more information: https://www.ons.gov.uk/census/2001censusandearlier/dataandproducts/outputgeography/outputareas)
- The data will be aggregated at Lower Super Output Areas (LSOA). The LSOA is the second most granular boundary of the Census data:

For the analysis the data will be aggregated at LSOA level



Population volume

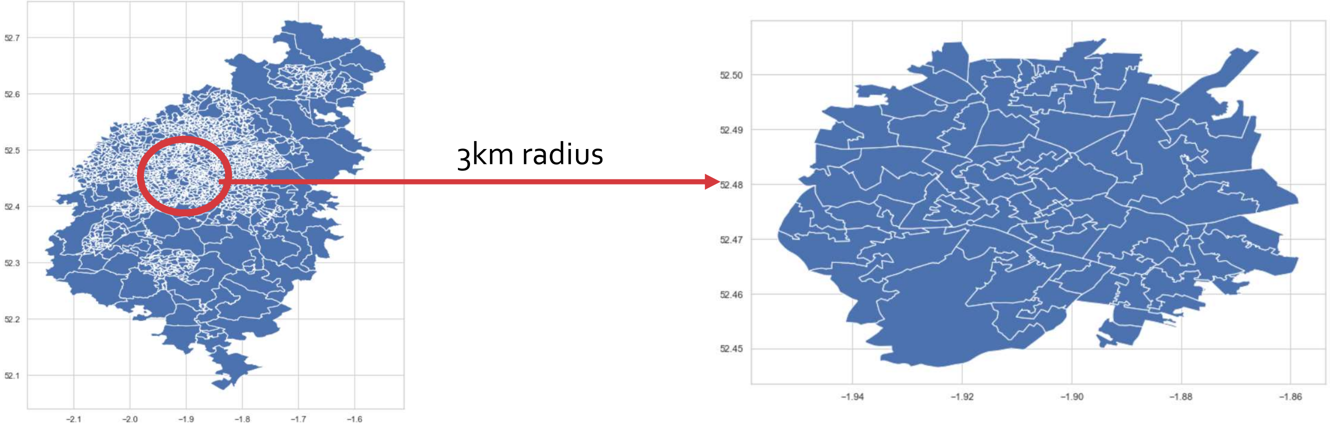
Age group

Economic class

Etc…

Data - Preparation

- All the datasets were imported from csv files apart from the shape files
- All the ONS data was combined into one dataset containing the LSOA codes, the population estimates, the total spending in restaurants and hotels and the working population
- A new columns was created containing the total spending done by residents in the target age group (21-39)
- The shape files were loaded using Geopandas in order to be converted to GeoJSON
- Due to the volume of LSOA geometries, the rendering of the shapes in Folium was creating stability issues with Jupyter Notebook. To limit the number of geometries a 3km radius filter from Birmingham city centre was applied



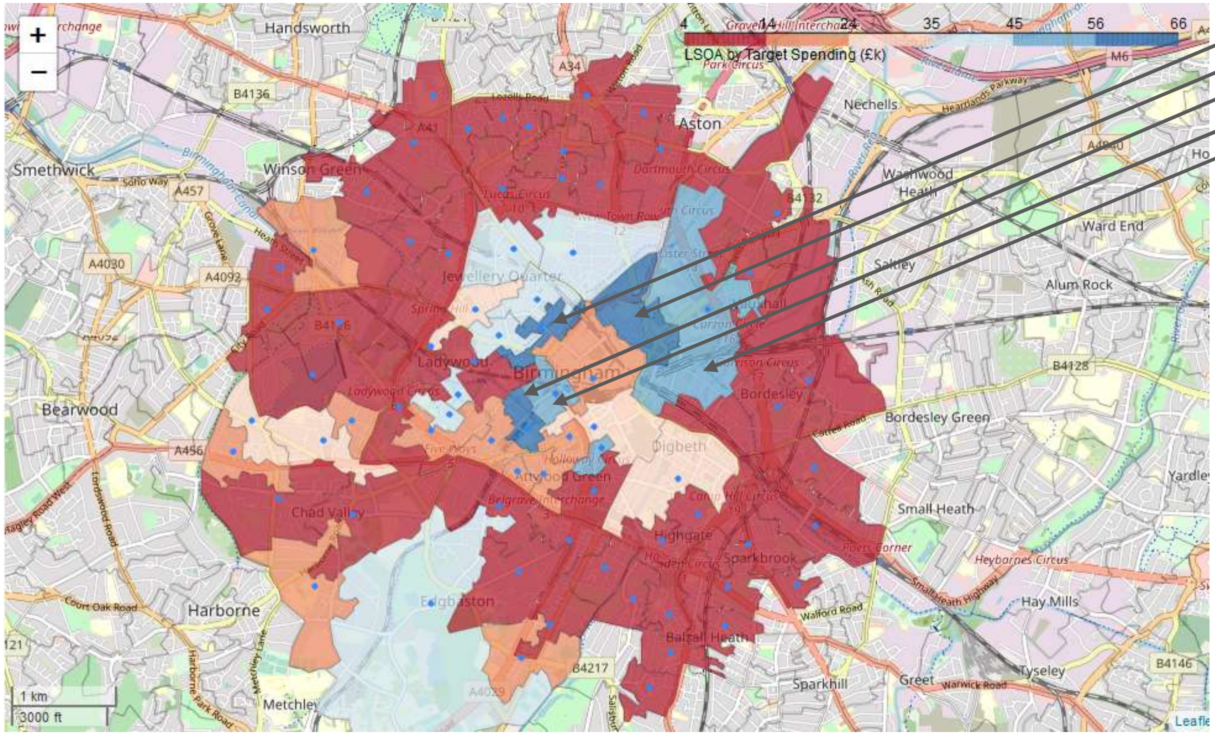
Birmingham LSOAs

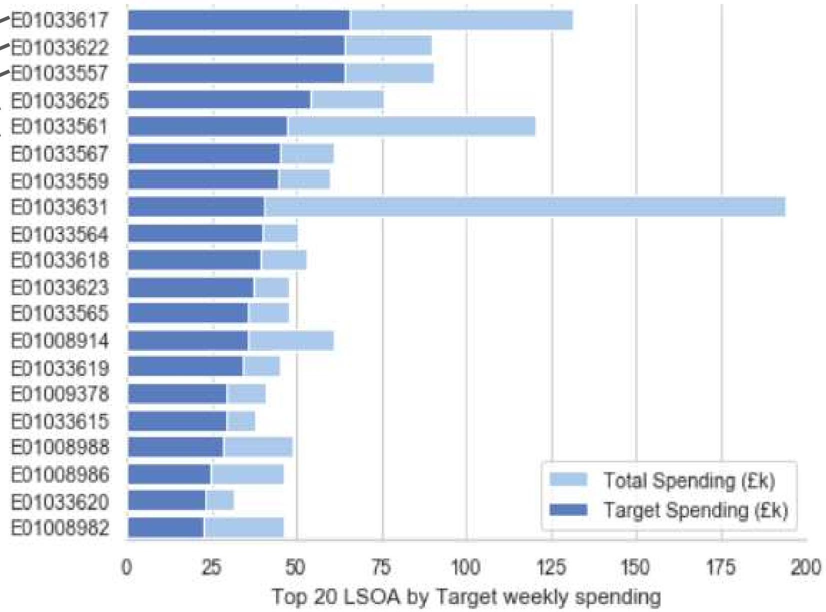
Birmingham LSOAs within 3km radius from the centre

- The ONS data was then merged to the JSON data dataset contained 1,182 rows across 7 columns; this dataset was then aggregated at LSOA level

Methodology | ONS Data Exploration

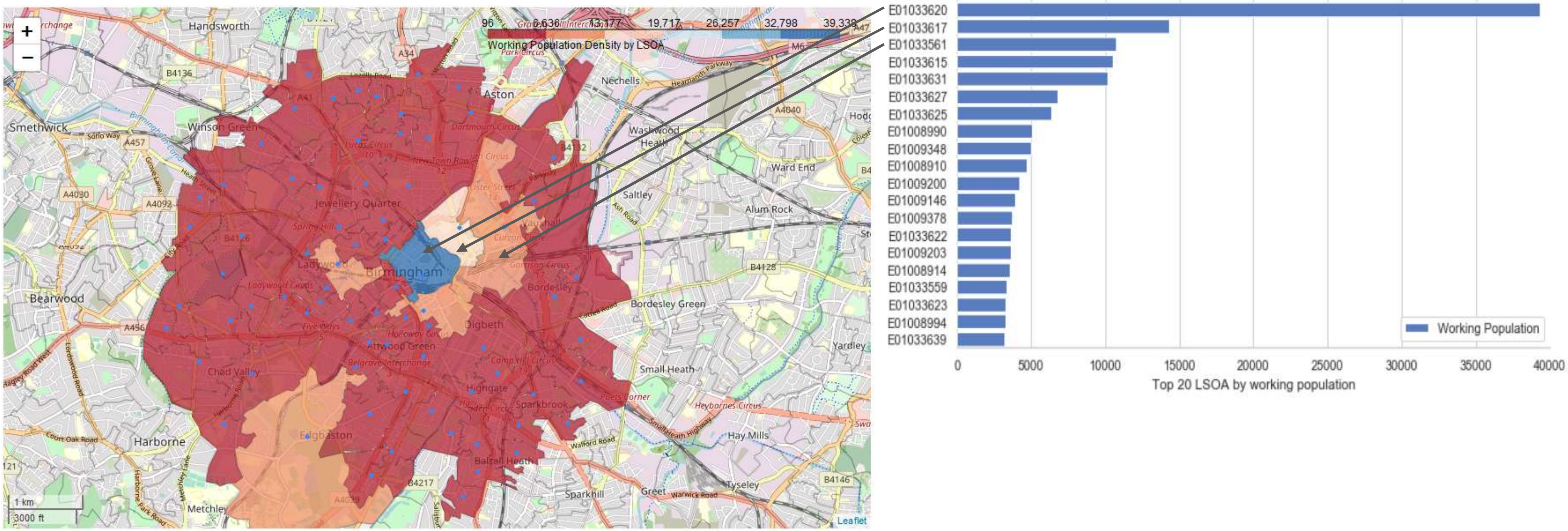
- A first step was to explore the ONS dataset to identify areas with high spending from the residents in the target group
- The locations with the highest spending for the target residents were close to the city centre but not in the city centre:





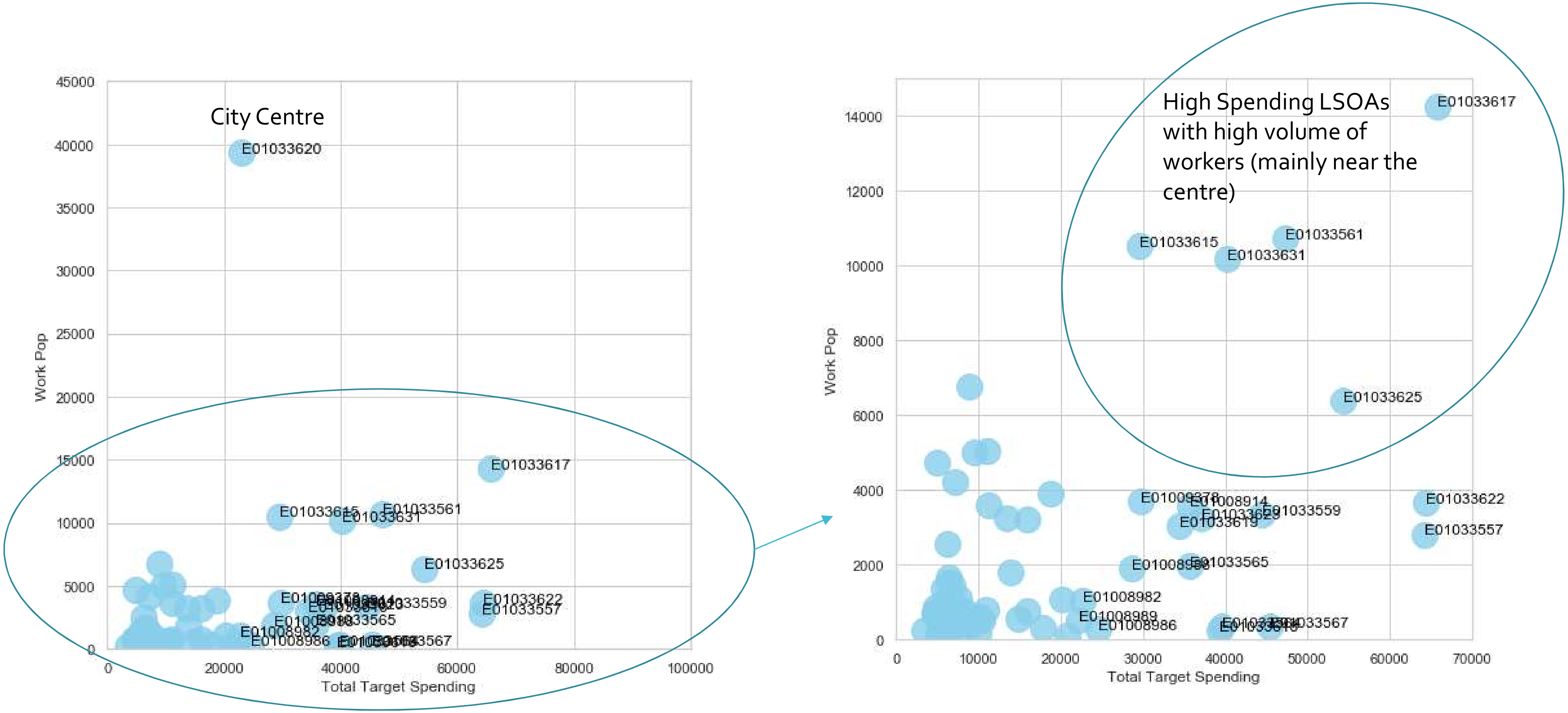
Methodology | ONS Data Exploration

- The working population was checked to understand the LSOA with the higher volume of workers
- The city centre had the highest volume of workers together with LSOAs around it:



Methodology | ONS Data Exploration

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

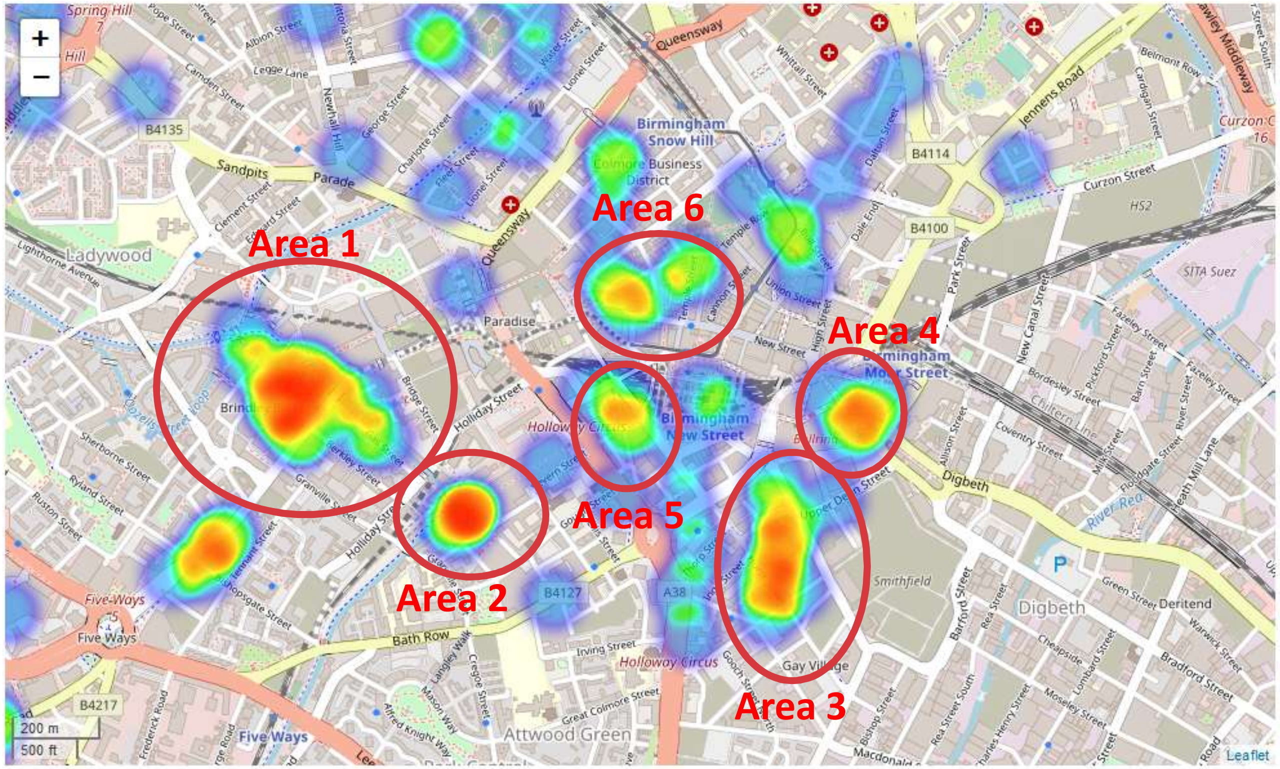


City Centre

High Spending LSOAs with high volume of workers (mainly near the centre)

Methodology | Foursquare Restaurant Venues

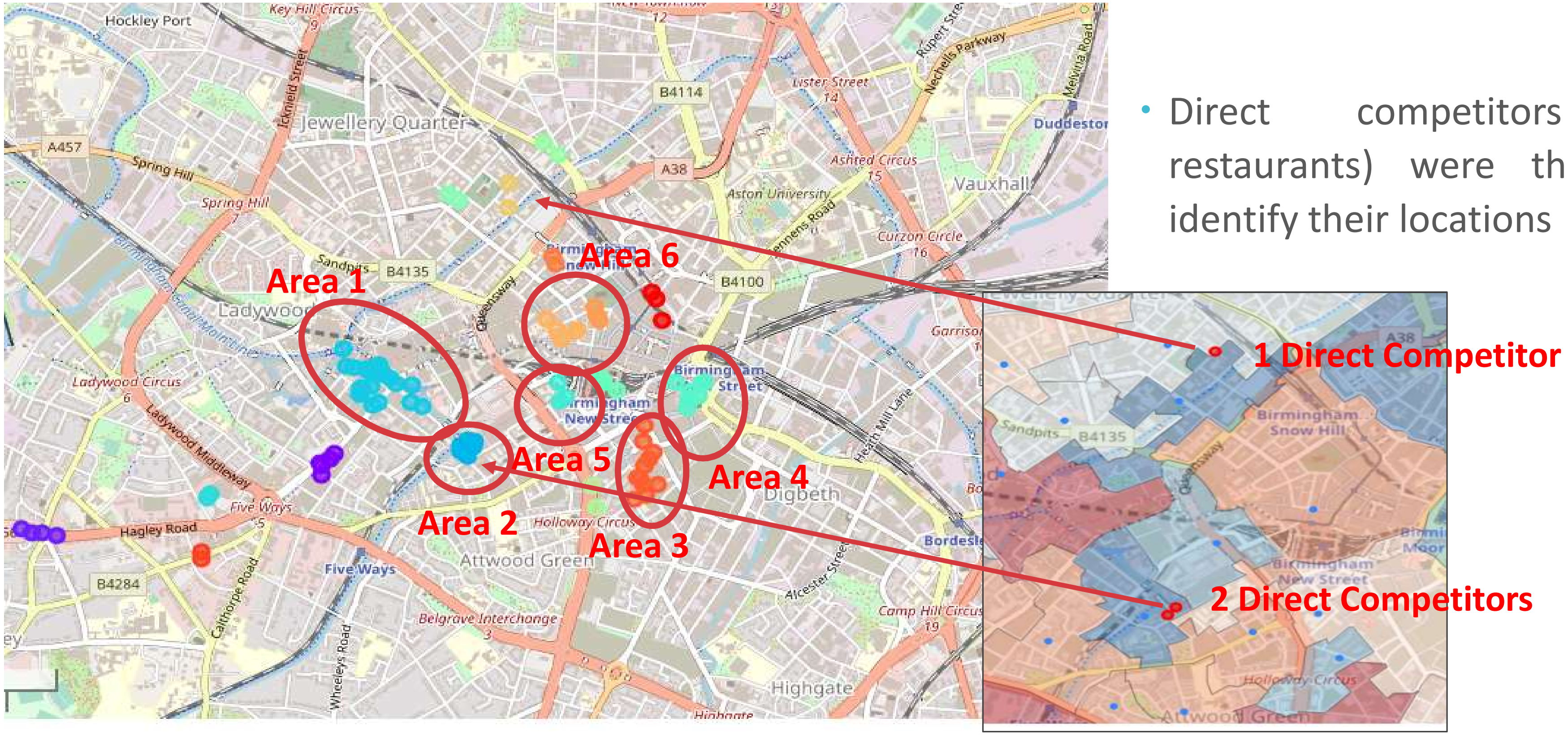
- Using the Foursquare API, venues in Birmingham were extracted using a radius of 500m from the LSOA centroids
- To identify the restaurant hotspots, a dataframe was created with unique restaurant venues and was displayed on a Folium heatmap (the density representing the volume of restaurants)



- The map shows distinct clusters mostly on the south of the city centre (if using the train line as x axis)
- In particular area 1 and 2 seem to have the highest concentration of restaurants

Methodology | Restaurant Geographic Clustering

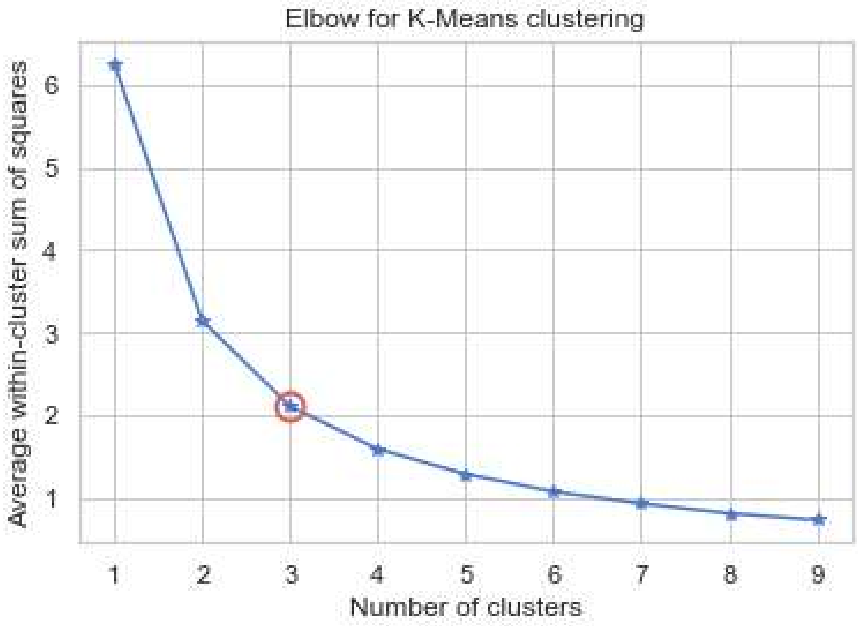
- With a better understanding of the hotspots, it was necessary to represent the hotspots as clusters in order to later on retrieve the venues specific to each of these clusters
- A DBSCAN clustering was run and broadly matched the hotspots:



- Direct competitors (Steakhouse restaurants) were then mapped to identify their locations

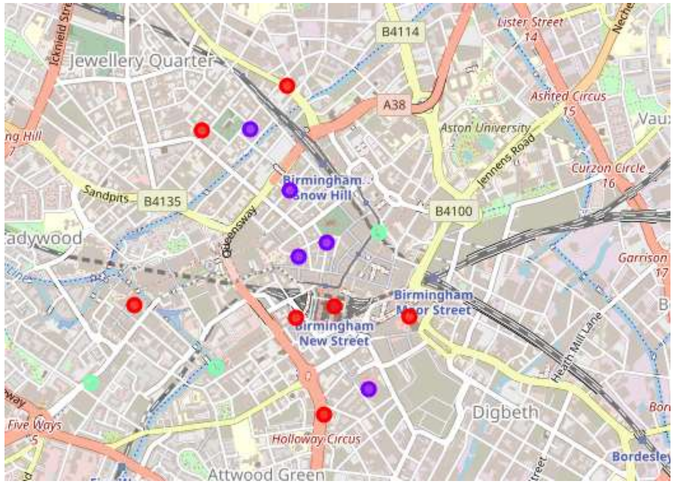
Methodology | K-Means Clustering

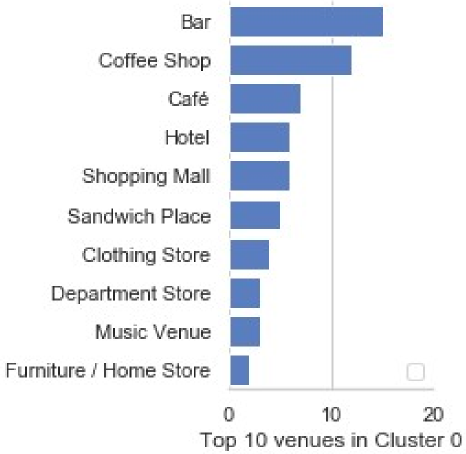
- Another query was sent using the Foursquare API to return venues within 200m from each hotspots centroids
- From the data, restaurants were excluded and the remaining venues were used to run a K-Means clustering
- To identify the best number of clusters an elbow method was used. The optimum number of clusters was 3 (k=3)

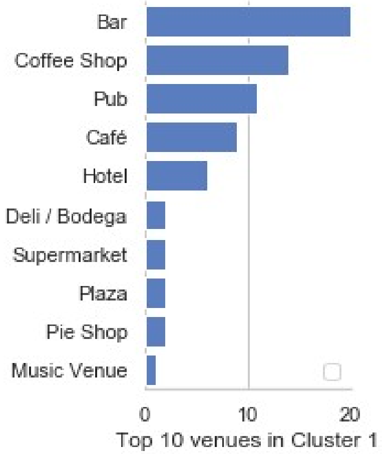


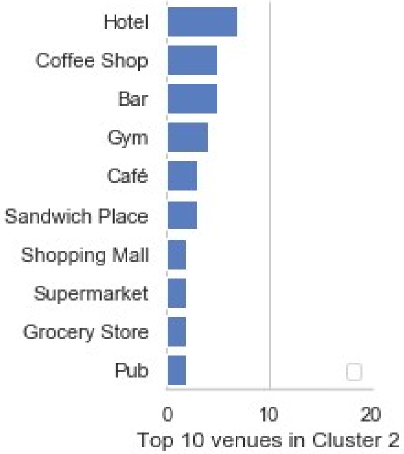
Results | Clusters' Characteristics

- Once the clustering was run, the clusters were displayed on a map and each cluster was analysed to understand its characteristics









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 ●

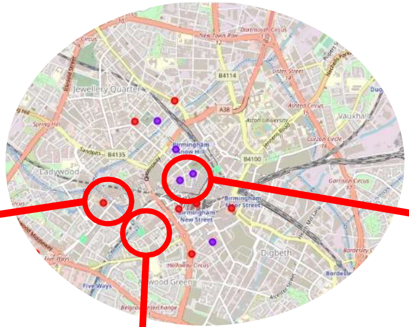
High volume of bars followed by coffee shops. Pubs come in third place.

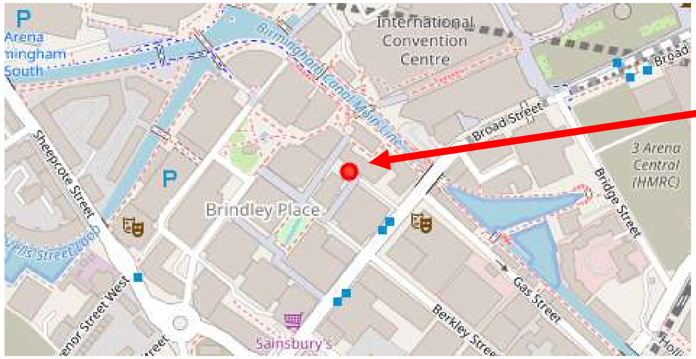
Cluster 2 ●

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.

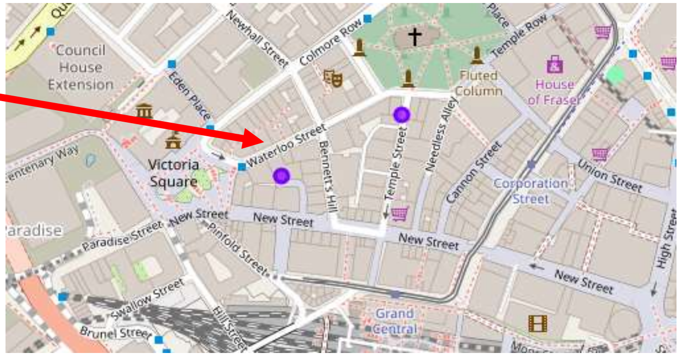
Results | Recommendations

- Based on these clusters and on the ONS data three potential locations were identified:

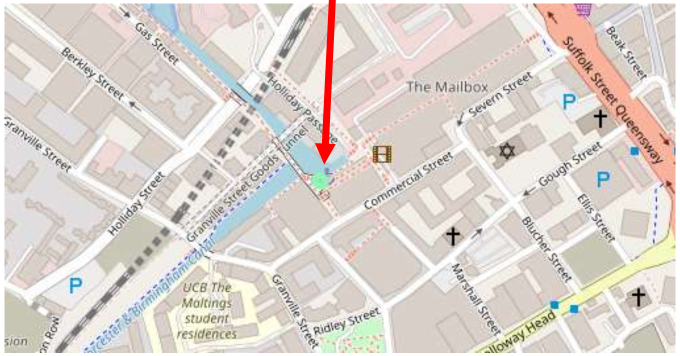




Option A: Around Brindley place (DBSCAN cluster 8 and k-means cluster 0)



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



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

Results | Recommendations

Option A: Around Brindley place	Option B: Near the Mailbox	Option C: Near Victoria Square
The Pros: - high volume of restaurants, bars, cafés and stores likely to attract the target customers - presence of stores and hotels in the area - close to a good catchment area made of residents in the target age group (21-29)	The Pros: - second place in terms of restaurants density, foot traffic should be high - area where residents in the target group have a high spending - situated close to shopping centres as well.	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 - surrounded by areas with residents in the target group age - highest volume of workers in the City - some competition but not as concentrated as option A and B.
The Cons: - high concentration of restaurants in the area - higher rents? - workplace population is also lower than other locations	The Cons: - already two steakhouse restaurants - not as many venues like bars as the other areas	The Cons: - very few residents. - what would be the activity in the evenings once workers are back home?

Discussion

- 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.
- Having such data would allow to use other techniques to predict the spending in a location for example using a geographic regression.
- 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.

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….
- … there is not one perfect solution but a few options to consider.
- The data analyst/data scientist needs to understand the business problem and requirements to translate their findings in a meaningful way to the stakeholders.