Coursera Applied Data Science Capstone Project Week 4 – Nottingham Battle of the Neighbourhoods

Richard Garner May 2020

Contents

[Introduction : Business Problem 1](#_Toc40812798)

[Data 2](#_Toc40812799)

[Crime in Nottingham 2](#_Toc40812800)

[Records to be Discarded 3](#_Toc40812801)

[LSOA and Choice of Geographic Grouping 3](#_Toc40812802)

[Results 4](#_Toc40812803)

[Discussion 4](#_Toc40812804)

[Data 4](#_Toc40812805)

[Results 4](#_Toc40812806)

[Conclusion 6](#_Toc40812807)

## Introduction : Business Problem

Nottingham is a medium sized city in the East Midland of the England in the UK. It has a thriving and lively nightlife with a wide and eclectic range of restaurants, cafes bars and clubs catering to many different categories of people including two universities. It is a chosen weekend destination for groups of young people from all over the UK.

Nottingham also has, like many cities in the UK, crime issues, with parts of the city and its environs exceeding the national crime rate by in excess of 100% (within particular crime categories).

<https://www.plumplot.co.uk/Nottingham-violent-crime-statistics.html>

The requirement in this case is to assess the optimum placement of restaurants / cafes / bars in areas that are more popular for dining and there is regular footfall and to take account of the crime statistics for that location. The data search should happen for the boroughs of Nottingham and the connected urban boroughs of Broxtowe, Gedling and Rushcliffe. It is intended that this data should be used by people who are looking to set up a hospitality business within the environs of Nottingham to help them decide which factors that they will take into account when choosing restaurant location and from that be able to suggest potential sites.

## Data

Based on the definition of the problem, we need to breakdown the research into two separate areas - the crime in a particular area and corresponding restaurant activity in that area.

### Crime in Nottingham

Based on the definition of the problem, we need to know the following type of information:-:-

* The number of crimes in a given location
* The type of crimes in a given location
* The frequency of crimes in a given location.

Data about recorded crimes is held for each Police force in the UK in the [Police Data Portal](https://data.police.uk/). The portal does provide an [API](https://data.police.uk/docs/) to access the data but this will only provide records for the specified month – if no month is specified then data for the previous month is returned. Since it is part of this analysis will include a view on the frequency of the crimes then we need to collect data on more than one time period. The Police Data portal also allows the selection and download of CSV files – selecting the Police Force and the Month that the crime was recorded in - to download. The data is down loaded in a zip file containing CSV files in separate directories for each selected month. Since this data cannot be accessed directly from Python, it was decided to download data relevant to the greater Nottingham area for the whole of the year 2019 and hold this in DB2 for subsequent querying using python. It was decided that only crime data should be downloaded, and that outcomes and stop and search data would not have any significance recording the crime rate.

The format of the CSV is described in the in the table below. From examination of this data it is clear that various fields will not be required in the analysis and as such the last column of the table shows which of the fields should be retained or discarded for use in this analysis.

| **Field** | **Meaning** | **Discard** |
| --- | --- | --- |
| Crime Id | Identifier for the individual crime | Discard |
| Month | The Year and Month in which the crime was committed. | Retain |
| Reported By | The force that provided the data about the crime. | Discard |
| Falls Within | At present, also the force that provided the data about the crime. This is currently being looked into and is likely to change in the near future. | Retain |
| Latitude and Longitude | The anonymised coordinates of the crime. See [location-anonymisation](https://data.police.uk/about/#location-anonymisation) for more information. | Retain |
| LSOA code and Name | References to the Lower Layer Super Output Area that the anonymised point falls into, according to the provided by the Office for National Statistics. | Retain |
| Crime Type | Categorisation of the crime | Retain |
| Last Outcome Category | reference to whichever of the outcomes associated with the crime occurred most recently. | Discard |
| Context | A field provided for forces to provide additional human-readable data about individual crimes. Currently, for newly added CSVs, this is always empty. | Discard |

Fields that are not to be retained are generally those used to identify the individual crime and as such are of no use in the aggregation of the crime information.

The LSOA information uses a standard method (according the UK Office of National Statistics) of grouping the crimes into a recognised geographical location. In particular the naming format allows human readable information regarding which Borough the crime was committed in - see below.

### Records to be Discarded

Given that the download selections are for the Month and Police force, the downloaded CSV files contain data regarding crimes committed outside the Greater Nottingham area. In this case it was decided that records with LSOA indicating that the crime was committed outside of the urban boroughs in or contiguous to Nottingham (Nottingham, Broxtowe, Gedling and Rushcliffe) should be discarded in order to cut down on Storage requirements.

### LSOA and Choice of Geographic Grouping

Lower Layer Super Output Area is a geospatial statistical unit used in England and Wales to facilitate the reporting of small area statistics. They are part of the coding system created by the Office for National Statistics (ONS). They have a minimum population of 1000 with a mean size of 1,500.

As has been noted each crime has been recorded with the Lower Layer Super Output Area (LSOA) which is a UK standard format for statistical reporting – further information can be found [here](https://geoportal.statistics.gov.uk/datasets/lower-layer-super-output-areas-december-2011-population-weighted-centroids?geometry=-1.607%2C52.888%2C-0.564%2C53.032). Each LSOA is available with population based geographical centroids.

Following the selection of the appropriate crime records it was found that there are 398 different LSOAs selected for the greater Nottingham Urban Area. As an area for geographic grouping, it was felt that the LSOA is somewhat small and as such a larger area geographical grouping may be required. It was decided to access the MSOA (Middle Super Output areas) which are hierarchically superior to – i.e. larger than - corresponding LSOAs. Data is available from the ONS joining the LSOA to the corresponding MSOA. This can be a straight forward file download which was then loaded into DB2 from where it could be queried in SQL with the join of the MSOAs to the corresponding LSOAs and therefore the crimes could be assigned to a MSOA as well as an LSOA.

The crimes could be summarised by various factors, which are shown in the result section. As part of this the crime types were reviewed and it was decided that all types should be retained as it is identified that all have a potential impact on the property or its customers.

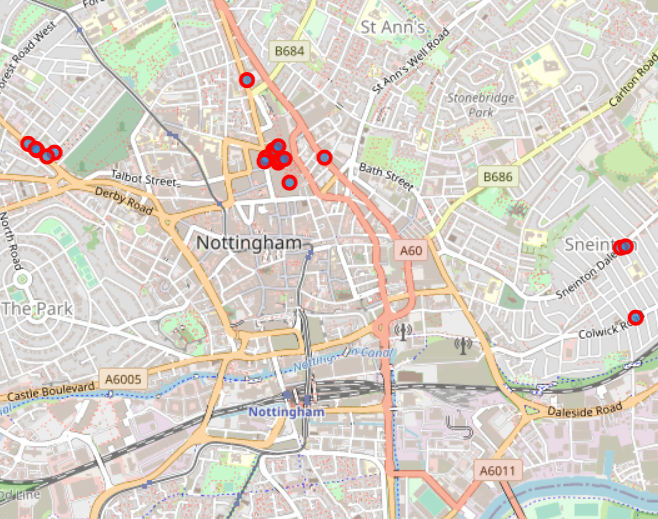
Population weighted centroids for the MSOAs can be accessed from the ONS website using an API to return a JSON file with the Longitude and Latitude for the MSOA. The data can be selected using a square to select the MSOAs within. The square used for selection for Nottingham is from (approximately) Longitude -1.350 Latitude 52.700 (SW corner) to Long. -1.000 Lat 53.200 (NE corner).  The JSON file can be joined to a table of the crimes summarised by MSOA. This can then be the basis of Pulling restaurant data from Foursquare. This provides 76 MSOAs for further analysis.

## Foursquare Data

Once the Centroids had been established, they can be used to select data from Foursquare using the API. Foursquare allows the interrogation of data concerning venues adjacent to a given geographic area. It also provides 'premium' data to collect more information regarding the venues, such as ratings, tips and further information. Given that this analysis is for the general area it was decided to start with the non premium service and select the venues only.

Venue data was selected from the Foursquare API by selecting venues within 400 meters of each of the MSOA centroids for the Nottingham Urban Area. When the selection was made a selection was made to select restaurants by using the categoryid as described on the Foursquare website [here](https://developer.foursquare.com/docs/api-reference/venues/search/#request).

The Venue data was then loaded into a map and reviewed. It was at this point that it was noticed that not as many venues as were expected had been selected – only 156 for the four boroughs - particularly in the city centre. This is probably because the MSOAs are too large an area, and the centroids provided by the UK Office of National Statistics are weighted by population. Since people do not generally live In UK central business districts the population weighting in a large area would be moved away from the city centre – towards the residential suburbs. See the excerpt from the map in the figure below:-



The venues are shown in red – as can be seen there are only a few on the edge of the city centre.

## LSOA Revisited

At this point it was appropriate to review the approach to finding the geographic data points to review the information in order to get better coverage of Nottingham. Going back to original crime details held in data extracted from the police portal, it can be seen can see that each crime has an anonymised latitude and longitude associated with it. It should therefor be possible to generate a crime weighted centroid for an LSOA (or indeed the linked MSOA) by finding the average of latitude and longitude for all crimes committed in the LSOA (or MSOA). It was decided to attempt to work this through for an LSOA – in order to provide more detailed coverage of the city.

So once the LSOA centroids were determined these we used to select Restaurant data from Foursquare in the same manner as was carried out for the MSOAs – although the radius for selection was set to 300 meters. This time 842 venues were selected from Foursquare across 61 different categories in 219 LSOAs. A screen print of the cnere of Nottingham with Venues in yellow and LSOAs in blue can be seen below.

The venue data was then aggregated for each LSOA with a count of the number of venues found in the LSOA .

## Methodology

The summary table for the LSOA crime data (for all LSOAs) had the LSOAs with venue data added to it. This was then standardised – which is to say an average number of crimes and aver age number of venues was calculated for all the LSOAs and these figures where then used to divide the corresponding count figure in each record. This meant that the standardised crime count for an LSOA gives a multiple of the Nottingham Urban average crime count for that LSOA. Similarly, the standardised Venue count gives a multiple of the Nottingham area (restaurant) venue count for the LSOA.

The standardised LSOA data was then fed into a K-means model to so that the LSOA could be grouped together based on the characteristics of Standardised Crime Rates and Venue frequency. The grouping was then summarised and displayed on a map.

The exercise was then repeated by joining the LSOA Venues table to the LSAO crime data. For this merged table there was only records for those LSOAs that had a venue. A new adjusted average was found for the counts of venues in the LSAO (Venue frequency) and that was then used to create the standardised venue frequency for each of the LSOA Venue records. A standardised value for crimes was found by using the existing average crime count. The adjusted figure was used here to

Edit Force

LSOA- Select Boroughs in Close to Nottingham

Crime Types

The format of that data is defined at the following link

## Results

Here we will review the results of the analysis of the data that we have found. The results will be separately discussed in the following discussion section. Each result will be alphabetically referenced for subsequent discussion.

1. Crime in Nottingham by month, by Crime Type.
2. Crime in Nottingham by month by Borough
3. Crime in Nottingham by Crime Type by Borough
4. Clustering all LSOAs by Crime rate and venue frequency

In this case the average number of crimes per LSOA in 2019 was found to be 208.35

The average number of restaurants found for all Nottingham LSOAs was found to be 2.12

1. Clustering LSOAs with Restaurants by Crime Rate and venue frequency.

The average number of restaurants found Nottingham LSOAs was with a restaurant was found to be 3.84

## Discussion

In this section the various parts of the project are discussed.

## Data

The use of the Lower Super Output Areas (LSOAs) was driven by the use of these units in the Crime data provided by the UK Police portal. Initial doubts regarding the large number number of LSOAs proved incorrect, as the use of MSOAs did not provide the required level of detail in the granularity of the geographical units.

The Foursquare data does appear to be incomplete for Nottingham. For example, the data does not include Nottingham’s premier restaurant, Sat Bains. This would be an interesting restaurant to include in the grouping as it is situated away from other venues in a relatively isolated site. Indeed, it can be found on the maps provided, but the site does not appear as a data point – unlike an adjacent venue which appeared before the data was filtered for restaurant venues. However, Sat Bains would have been an interesting outlier for the K-means cluster modelling but given that this is a “destination venue” it may well be the case that in practice, the restaurant could be placed anywhere in Nottingham with easy access to the main roads. Nevertheless, other venues that exist in Nottingham are missing, particularly in the more outlying areas.

However, the omission is concerning and further work could be required to understand if this is a factor of how the Foursquare data is gathered and the demographic of the people doing the gathering or any other related reasons.

## Results

The first three parts of the results have been put together in order to review the overall nature of crime and how it is distributed in the greater Nottingham urban area. The last two points discuss the results of the groupings of the various LSOAs based on crime and venue frequency in the greater Nottingham area.

1. Crime in Nottingham by Month, by Crime Type.

The variation in the monthly crime figures is proportionally small – meaning that the figures are quite regular with small variation. Looking at the breakdown of the results by crime type, these too appear to have only small variation within them, appearing to have practically the same proportions each month.

1. Crime in Nottingham by Month by Borough

Given that the data here is shown by borough and not detailed by LSOA this graph still imparts some information – the Borough of Nottingham has disproportionately more crime reported than the outlying boroughs. Whilst population density may be a factor (the figures are not used here) it is also worth noting that Nottingham has a busy city centre and people travel in. Clearly some of those people commit crimes when they travel in.

1. Crime in Nottingham by Crime Type by Borough

Again, this graph is only displayed at the level of Borough and as such the detail of the individual LSOAs is not available. Once again, the Borough of Nottingham dominates the number of crimes but the proportionality between the boroughs appears to be retained, with a slight proportional increase for shoplifting in Broxtowe and a proportional drop for vehicle crime in Nottingham.

1. Clustering all LSOAs by Crime rate and venue frequency

The K-means clustering gives the following groups:-

**Group 0** – This group is characterised by having a slightly higher than average crime rate and a higher than average frequency of restaurants. There are 61 LSOAs in this group and reviewing this on the map shown in the results (LSOAs in this group are red) these LSOAs represent mains streets in some of the sub-urban town centres – as opposed to the main city centre.

**Group 1** – This group of LSOAs is characterised by a lower than average crime rate with practically zero restaurant frequency. On the map for this, the purple LSOAs are the mainly residential outer suburbs.

**Group 2** – This group is characterised by a very high crime rate and also a very high frequency of restaurants. On the map, the 2 LSOAs in this group are in the City Centre.

**Group 3** – There is one LSOA in this group – shown in green on the map very close to the city centre. LSOA has a very high crime rate and also shows a significantly higher than average restaurant frequency.

**Group 4** – This group is characterised by a relatively high crime rate but also a high frequency of restaurants. On the map these LSOAs are shown in yellow and are very close to the city centre, on the eastern side.

1. Clustering LSOAs with Restaurants by Crime Rate and venue frequency.

The grouping of those LSOAs with a restaurant venue represents a more focussed approach to examining and characterising the areas of Nottingham that have a venue present. As stated before, the number here represent (i) for Crimes the 1 = the average crime rate across all LSOAs in the Nottingham Urban Area. The number given to the LSOA represents a multiple of that aver age rate. (ii) for Venues 1 = the average frequency of restaurants calculated across all LSOAs that have a restaurant in them. The number assigned to the LSOA in this case is a multiple of that average.

This grouping would be useful analysis for stakeholder who want to site their location in proximity to existing locations. The K-means clustering has given the following groups:-

**Group 0** – This group is characterised by a lower than average crime rate and venue frequency. Looking at the map the red LSOAs are distributed over the town centres of the Nottingham suburbs.

**Group 1** – This group of LSOAs is characterised by very high crime rate and very high venue frequency. On the map the 2 LSOAs in the group are the green ones in the city centre.

**Group 2** – This group is characterised by a relatively high multiples of crime rate and a higher venue frequency. On the map these LSOAs are the blue ones mainly distributed around the city centre.

**Group 3** – These LSOAs have double the average crime rate and a roughly average venue frequency. On the map these LSOAs show as green and are distributed around the suburban area.

**Group 4** – There is one LSOA in this groups – characterised by a very high crime rate with double the average venue frequency. On the map the yellow LSOA appears quite close to the city centre.

## Conclusion

The clustering of the LSOAs using crime rate and grouping of the various LSOAs would show that there is a huge increase in the crime rate in Nottingham City Centre and at the same time this is where there is a the greatest density of restaurants. This shows that many restaurants exist in areas with high crime rates. It is not suggested that there is a causal link – more a case that restaurants are more successful where there are more people, and where there is more people, there is more crime. Undoubtedly Nottingham City centre attracts a great many people.

The use of the first grouping of LSOAs is supported by the second – which represents a refinement of the first model.

However there is no clear recommendation of which area that a stakeholder would want to site a potential business in. The model would give them a chance to understand the groups and then examine the potential of an LSOA based on crime rate and locally existing restaurants. The approach used works best if the stakeholder want to be co-located with other restaurants.

The model could be improved by incorporating the following:-

Foursquare premium services need to be incorporated – particularly the number of check-ins and customer rating.

Cost information such as Business rate, Rents, Insurance rates etc.

Day of the week that the Crime was committed.

The addition of such information would give a user more information in order to optimise the choice of LSOA for the location of a potential .