Investigation into the relationship between local businesses and venues and crime in London

IBM Data Science Professional Certificate Capstone Project

Author: Brenda Cundy
Date: 25 June

2021

Table of Contents

1. Disclaimer

This is not a formal analysis. It has been commissioned as a capstone project for IBM Data Science Professional Certificate. No reliance should be placed on conclusions and analysis within. This report has been produced under a number of constraints, including time and requirements of the capstone project and should not be considered fully representative of the quality of work the author can produce under professional conditions.

2. Executive Summary

[placeholder]

3. Introduction

London is the capital city of the United Kingdom, with an estimated population of 9.4 million (source: world population review). Modern London is a diverse city, with varied neighbourhoods. Like all areas of urbanisation crime is a concern to residents, workers and business owners. The consequences of crime can be direct ie for the victim, or indirect, ie loss of business revenue due to avoidance of area.

3.1 Objective

This report will address the following question:

- Within the context of London, is there a relationship between the types of venues/industries and crime?

3.2 Scope

Understanding the nature of any such relationship is out of scope of this report (for example if increased presence of certain businesses is in response to crime rate or a cause of it). The conclusion from this report will include recommendations for additional investigation.

3.3 Stakeholders

Then intended stakeholders for this report are

- local planners (ie for consideration in approving planning applications, and deciding planning strategies)
- local residents and business owners (for understanding potential impact of local facilities and businesses on crime rate)
- prospective residents and business owners (for under potential crime profile on an area based on the local facilities and businesses)

4. Background

London is a diverse and a historic city, first referenced as Londinium in the 2nd Century AD . It is the largest city in the UK, and currently, by population the 32nd largest in the world (source: worldometer). London was subject to rapid growth associated with industrialisation the 19th Century AD, when the population grow fivefold, and immigration in the late 20th century following mid century decline. In addition to population growth the geographical area considered the city has grown to incorporate what were once separate settlement boundaries. Notably in 1965, London was redefined, and what were parts of neighbouring counties became included in London boroughs This history has resulted in a diverse geography, with neighbourhoods not necessarily fitting within the political boundaries by which London is governed. Many London neighbourhoods represent historic parishes, once separate to London., others have developed around transport links, for example the London underground stations.

London is governed by division into 32 local government districts referred to boroughs each governed by a London borough council..These make up the ceremonial county of Greater London. The “City of London”, the historic centre of London, is a separate ceremonial county. However, the two counties together comprise the region of Greater London, all of which is also governed by the Greater London Authority. This is the definition of London used in this report.

Each London borough is subdivided into wards. These wards represent areas of population and seats in the borough council, they are not distinct communities and can be residential with commercial centres and amenities in nearby wards.

Greater London is covered by three separate police forces.

Police Force:	**Responsible for policing of:**
Metropolitan Police	The vast majority of London
City of London Police	City of London
British Transport Police	The national rail network and the London Underground

.

5. Data

Two primary sources of data will be used for this report:

1. **Foursquare**: an independent location data platform, which harvest details and recommendation of venues as input by users of it's mobile application "Foursquare City Guide"
2. **The London Datastore:** a free and open data-sharing portal, provided by the London Assembly and Mayor of London.

This reports utilises three groupings of data:
a. Crime
b. Local businesses/venues
c. Ancillary data to support analysis

5.1 London Crime Data

Data source: The London Datastore. *Recorded Crime Summary: Geographic Breakdown*. Provided by the Metropolitan Police
https://data.london.gov.uk/dataset/recorded_crime_summary

Summary of data:

MPS Borough level crime (most recent 24 months):	Number of offences by month, by category (both major and minor), by borough for 24 months (the most recent that data is available for)
MPS Ward level crime (most recent 24 months):	Number of offences by month, by category (both major and minor), by ward for 24 months (the most recent that data is available for)
Borough level crime (historic):	Number of offences by month, by category (both major and minor), by borough for 2008 to 2018

	MajorText	MinorText	LookUp_BoroughName	201905	201906	201907	201908	201909	201910	201911	...	202007	202008	202009	202010	202011	202012	202101	202102	202103	202104
0	Arson and Criminal Damage	Arson	Barking and Dagenham	11	3	5	3	6	9	8	...	4	6	2	7	4	2	4	6	4	6
1	Arson and Criminal Damage	Criminal Damage	Barking and Dagenham	140	113	134	118	109	109	97	...	122	114	116	120	100	109	100	104	80	100
2	Burglary	Burglary - Business and Community	Barking and Dagenham	21	27	31	35	37	30	30	...	28	23	32	21	18	24	20	18	14	12
3	Burglary	Burglary - Residential	Barking and Dagenham	114	96	71	67	80	97	114	...	72	63	54	68	90	91	69	90	71	75
4	Drug Offences	Drug Trafficking	Barking and Dagenham	9	6	11	8	7	9	14	...	21	9	12	13	17	13	12	9	7	6

Figure 1: MPS Borough Level Crime (most recent 24 month). First 5 records

See appendix for details of crime categories

Data cleansing: None required, data is complete and good quality.

Data limitations: Data used is crime statistics published by the Metropolitan police so will exclude the City of London, and crime recorded on the London underground and national rail network.

5.2 Local businesses/venues data

This report will consider three different sources of data reflecting the facilities and venues. The London datastore contains additional relevant data which could be used for follow-up research.

5.2.1 Foursquare

Data source:
Foursquare API "explore" endpoint returns up to 50 recommended venues. The limit of 50 is as stated in the developer reference guide[1], although experience suggests the true limit is 100. These venues are those from a geographic area either specified by a geocodable location name, or longitude/latitude coordinates provided with a search radius. The foursquare venue database is crowd sourced.
The following approaches can be considered for obtaining venue data from Foursquare:

a) By borough, either:
   i. Using a central point in the borough - either identified from an internet list or using Nominatim to provide coordinates
   ii. Using Foursquare to recognise the borough names as geocodable locations
b) By neighbourhood, either:
   i. Using a central point in the neighbourhood identified by Nominatim
   ii. Using Foursquare to recognise the neighbourhood names as geocodable locations

The approach will be finalised post data exploration.

Summary of data: Foursquare API returns a JSON response, details of which can be found on the developer reference guide. For this exercise the venue name, location and category will be considered. The category provides a description of the type of venue and allows grouping of similar venues for further analysis

Data cleansing: Not required

Data limitations: There are a number of limitations with foursquare data

1. Crowd sourced data is biased by the demographic which uses foursquare
2. The geographic area data is obtained for may not match, or be representative of the geographic area crime data is available for
3. The limit of responses may prevent the returned venues being representative of the area searched.
4. The allocation of category can be subjective, for example coffee shop vs cafe, pub vs gastropub, fried chicken joint vs fast food restaurant.

1 https://developer.foursquare.com/docs/places-api/endpoints/

5.2.2 Venues with 24 Hour Licence

This data allows detailed analysis into a particular venue type.

Data source: The London Data Store *Alcohol and late-night refreshment licensing statistics – licensed premises 24 hour*. Provided by the Home Office.
https://data.london.gov.uk/dataset/alcohol-and-late-night-refreshment-licensing-statistics

Summary of data: Data is provided on an annual basis, with the most recent data being from 2018. Number of licensed premise is provided for each borough, broken down by premises type with totals and subtotals across categories of premises. Propose to use 2018 data with 2017 data used to fill in gaps

	Premises with 24-hour alcohol licences											
			Supermarkets and stores				Hotel bars					
Licensing authority	Total	Pubs, bars and nightclubs	Total	Large supermarkets	Other convenience stores	Supermarket and store type not reported	Total	Open 24 hours to residents and general public	Open 24 hours to residents and their guests only	Hotel bar type not reported	Other premises types	Premises type not reported
Barking and Dagenham	:	:	3	1	2	0	:	0	:	:	0	:
Barnet	:	:	:	:	:	:	:	:	:	:	:	:
Bexley	4	0	4	1	3	0	0	0	0	0	0	0
Brent	46	2	28	4	24	0	3	3	0	0	13	0
Bromley	14	0	11	4	7	0	3	0	3	0	0	0
Camden	:	:	:	:	:	:	:	:	:	:	:	:

Figure 2: Example data on licensed premises (24 hour license)

Data cleansing:
Data is cleansed to enable analysis:

- There are a number of gaps in the data (shown as :), where that data is not available. Use the previous years' value if available.
- Ensure totals and subtotals reflect any data cleansing
- If any data is missing for 2 consecutive years exclude
- Exclude any features (columns) where more than eight records are 0, as there will be insufficient data for meaningful analysis

Borough	Total	Total Supermarkets and stores	Large supermarkets	Other convenience stores	Total Hotel bars	Open 24 hours to residents and their guests only
Barking and Dagenham	3	3	1	2	0	0
Barnet	0	0	0	0	0	0
Bexley	4	4	1	3	0	0
Brent	46	28	4	24	3	0
Bromley	14	11	4	7	3	3

Figure 3: Example cleansed data showing feature reduction

5.2.3 Local Units by industry

Data source:
The London datastore: *Local unity by broad industry group, borough. P*rovided by the Office of National Statistics.
https://data.london.gov.uk/dataset/local-units-broad-industry-group-borough

Summary of data:
Data is available for each year 2003 to 2020. The number of businesses (local units such as a factory or a shop) by Broad Industry Groups, per borough is provided per year.

		UK SIC 2007																
Code	Area	SIC07: 01-03 : Agriculture , forestry & fishing	SIC07: 05-39 : Production	SIC07: 41-43 : Constructio n	SIC07: 45 : Motor trades	SIC07: 46 : Wholesale	SIC07: 47 : Retail	SIC07: 49-53 : Transport & Storage (inc. postal)	SIC07: 55-56 : Accommodatio n & food services	SIC07: 58-63 : Information & communicatio n	SIC07: 64-66 : Finance & insurance	SIC07: 68 : Property	SIC07: 69-75 : Professional , scientific & technical	SIC07: 77-82 : Business administratio n & support services	SIC07: 84 : Public administratio n & defence	SIC07: 85 : Educatio n	SIC07: 86-88 : Health	SIC07: 90-99 : Arts, entertainment , recreation & other services	SIC07: Total
E09000001	City of London	20	730	725	25	605	995	285	1,280	2,185	3,700	1,055	8,005	6,105	55	215	330	1,050	27,365
E09000002	Barking and Dagenham	10	330	1,460	225	410	645	530	410	610	90	145	810	780	170	200	605	360	7,790
E09000003	Barnet	30	625	3,200	350	1,155	2,165	600	1,000	2,570	530	2,055	4,910	2,205	55	535	1,250	1,585	24,820
E09000004	Bexley	10	445	1,905	260	335	805	380	570	1,065	180	225	1,510	840	25	225	555	560	9,895
E09000005	Brent	5	585	2,255	410	955	1,600	670	915	1,870	270	755	2,630	1,355	60	320	790	1,015	16,460
E09000006	Bromley	60	495	2,355	290	525	1,665	310	855	2,080	390	510	3,465	1,570	25	380	900	1,195	17,070
E09000007	Camden	30	1,005	1,515	150	1,240	2,450	465	2,140	4,610	855	1,525	10,385	3,530	115	680	1,205	2,845	34,745
E09000008	Croydon	15	495	2,285	390	540	1,540	500	1,020	2,030	335	525	2,930	1,445	50	420	1,095	1,000	16,615

Figure 4: Example local units data as available from London Data Store

Data cleansing:
Data is cleansed to enable analysis:

- Additional UK regional data is included in the dataset, records removed as only London Borough data is required
- No gaps identified in data

Data limitations:
Number of units provides a single snapshot, in reality businesses are not static, opening and closing through the year.

5.3 Ancillary Data

The following data is utilised to support analysis

5.3.1 London Population Data

Population data enables the primary data to be population adjusted – eg crime rate per 10,000 capita to be calculated.

Data source:
London Datastore: *London Borough Profiles*. Provided by the Greater London Authority
https://data.london.gov.uk/dataset/london-borough-profiles
London Datastore: *London Ward Profiles*: Provided by the Greater London Authority
https://data.london.gov.uk/dataset/ward-profiles-and-atlas

Summary of data:
Extensive indicators are provided about each London ward and borough. Of interest to this analysis is population.
The population figure provided is:
2017 Estimate : Borough level data
2015 Estimate: Ward level data

	Code	Area_name	Inner/_Outer_London	GLA_Population_Estimate_2017	GLA_Household_Estimate_2017	Inland_Area_(Hectares)	Population_density_(per_hectare)_2017	Average_Age,_2017	Proportion_of_population_ag 15,_
0	E09000001	City of London	Inner London	8800	5326	290	30.3	43.2	
1	E09000002	Barking and Dagenham	Outer London	209000	78188	3,611	57.9	32.9	
2	E09000003	Barnet	Outer London	389600	151423	8,675	44.9	37.3	
3	E09000004	Bexley	Outer London	244300	97736	6,058	40.3	39.0	
4	E09000005	Brent	Outer London	332100	121048	4,323	76.8	35.6	

Figure 5: Header of Borough profile data

Data limitations:
It is assumed that population growth has been consistent across boroughs/wards since the date of the population estimates, therefore the population number is a reasonable proxy to use for current population.

5.3.2 Geocoding

Nominatim API is used to obtained coordinates of locations, based on a place name (ie London boroughs and neighbourhoods.
https://wiki.openstreetmap.org/wiki/Nominatim
This is used for data visualisations, and is investigated for utilising the Foursquare API to obtain venue data for a specific geography.

5.3.3 London maps

The python Folium library is used for visualisations
http://python-visualization.github.io/folium/

London boroughs are overlaid using ***london_boroughs_proper.geojson*** sourced from:
https://joshuaboyd1.carto.com/tables/london_boroughs_proper/public/map
London wards are overlaid using ***london-wards-2014.geojson*** sourced from:
https://github.com/ft-interactive/geo-data

5.3.4 List of London Neighbourhoods

Wikipedia provide a list of London neighbourhoods is scrapped from
'https://en.wikipedia.org/wiki/List_of_areas_of_London'

Location	London borough	Post town	Postcode district	Dial code	OS grid ref
Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785
Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805
Addington	Croydon[8]	CROYDON	CR0	020	TQ375645
Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665
Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728
Aldborough Hatch	Redbridge[9]	ILFORD	IG2	020	TQ455895
Aldgate	City[10]	LONDON	EC3	020	TQ334813
Aldwych	Westminster[10]	LONDON	WC2	020	TQ307810
Alperton	Brent[11]	WEMBLEY	HA0	020	TQ185835
Anerley	Bromley[11]	LONDON	SE20	020	TQ345695
Angel	Islington[8]	LONDON	EC1, N1	020	TQ345665
Aperfield	Bromley[11]	WESTERHAM	TN16	01959	TQ425585

Figure 6: List of London neighbourhoods as shown on wikipedia

5.3.5 List of London borough locations

The wikipedia list of London boroughs provides coordinates for each https://en.wikipedia.org/wiki/List_of_London_boroughs

Borough	Inner	Status	Local authority	Political control	Headquarters	Area (sq mi)	Population (2019 est)[1]	Co-ordinates	Nr. in map
Barking and Dagenham[note 1]			Barking and Dagenham London Borough Council	Labour	Town Hall, 1 Town Square	13.93	212,906	51.5607°N 0.1557°E	25
Barnet			Barnet London Borough Council	Conservative	Barnet House, 2 Bristol Avenue, Colindale	33.49	395,896	51.6252°N 0.1517°W	31
Bexley			Bexley London Borough Council	Conservative	Civic Offices, 2 Watling Street	23.38	248,287	51.4549°N 0.1505°E	23
Brent			Brent London Borough Council	Labour	Brent Civic Centre, Engineers Way	16.70	329,771	51.5588°N 0.2817°W	12
Bromley			Bromley London Borough Council	Conservative	Civic Centre, Stockwell Close	57.97	332,336	51.4039°N 0.0198°E	20
Camden	✓		Camden London Borough Council	Labour	Camden Town Hall, Judd Street	8.40	270,029	51.5290°N 0.1255°W	11

Figure 7: List of London boroughs (in part) as shown on wikipedia

6. Outline approach

Exploratory data analysis will confirm the data transformations required which may include:
- data aggregation (ie crime data over a time period)
- adjusting data for population (ie so crime/venues per 10,000 capita is considered)
- exclusion of outliers

Exploratory data analysis will also consider trends in different levels of crime data categorisation classification (total, major category, minor category) to determine at which level analysis should be conducted.

Foursquare data will be used to form clusters of geography (either neighbourhood or borough) based on similar venue profiles. Venue profiles will be determined using normalisation of number of venues of different categories in a geography. K-means clustering will be used to identify clusters of similar venue profiles. Map based visualisations will be used to enable manual inspection for correlations. Ie cluster labels superimposed on a choropleth map of crime data. Data exploration will confirm the geography used in this analysis.

Where crime and business data exists for the same geography, pairwise pearson correlation analysis will be conducted to identify where correlation exists.

This analysis will be conducted in python, libraries used will include:

BeautifulSoup	for web scraping
numpy	to handle data in a vectorized manner
pandas	for data analysis
geopy.geocoders Nominatim	to convert an address into latitude and longitude values
sklearn.cluster	for k-means clustering
folium	for map rendering
matplotlib	for visualisations
seaborn	for visualisations