Battle of the Boroughs:

A Beer Garden-based segmentation of London's Boroughs

Pritpal Warner

July 2021

Introduction

Background

The COVID-19 pandemic has hit the UK's hospitality industry hard. Pubs were closed for over a year but are now open with Government COVID-safety restrictions and are eager to make up for lost trade. There is also pent-up demand from customers who want to return to COVID-safe pubs, especially pubs with beer gardens in which to enjoy the great British summer. Given their outdoors nature, a beer garden improves COVID-safety, and they are more important than ever given society's heightened sense of COVID-safety and UK Government guidance. For those unfamiliar with UK pubs, note that not all have beer gardens. Unless otherwise indicated, further references to 'pub' refer to a pub with a beer garden.

Business Problem

Not all Boroughs are equally blessed with beer gardens. The objective of this project is to address the key business problem of "Which Boroughs should businesses target for launching beer gardens?" I'll address this question by segmenting the Boroughs using demographic, geographic and pub (Foursquare) data.

Interest

The hospitality industry and pub-goers should be interested in the output of this project.

Data

There are six main sources of data as described below.

1. Borough geographic and population data from London Datastore

Description: file with columns listing area and population for each Borough.

Purpose: to inform exploratory descriptive analysis and Borough segmentation.

Source: London Datastore

Format: CSV

Link: https://data.london.gov.uk/download/land-area-and-population-density-ward-and-borough/77e9257d-ad9d-47aa-aeed-59a00741f301/housing-density-borough.csv

2. Borough demographic data (from London Datastore

Description: file with breakdown of population by age group for each Borough, including percentage of working age. The data used was a subset of the source.

Purpose: the data was used as a proxy for people of pub-going age which is used to inform exploratory descriptive analysis and Borough segmentation.

Source: London Datastore

Format: CSV

Link: https://data.london.gov.uk/download/london-borough-profiles/c1693b82-68b1-44ee-beb2-3decf17dc1f8/london-borough-profiles.csv

3. Borough demographic data (religious affiliation) from London Datastore

Description: file with breakdown of population by religion. The data used was a subset of the source.

Purpose: to allow calculation of 'Drinkers %' – the percentage of population allowed by their religion to drink alcohol (= 1- Muslim %) to inform exploratory descriptive analysis and Borough segmentation.

Source: London Datastore

Format: CSV

Link: https://data.london.gov.uk/download/percentage-population-religion-borough/abfb6175-f489-4c6e-add2-f4d323183224/population-by-religion-borough.xls

4. List of London Boroughs and neighbourhoods from Wikipedia

Description: table of Boroughs and neighbourhoods in each Borough.

Purpose: to identify complete list of neighbourhoods for each Borough, for later use by Python Geocoder to return neighbourhood latitude and longitude.

Source: Wikipedia

Format: html table

Link: https://en.wikipedia.org/wiki/List_of_areas_of_London

5. Latitude and longitude data for each neighbourhood from Python Geocoder

Description: latitude and longitude coordinates for each neighbourhood.

Purpose: coordinates to be used in Foursquare API search/explore query to identify pubs with beer gardens in each neighbourhood.

Source: Python Geocoder

Format: Python lists

Link: !pip install geocoder

6. Pubs with beer gardens in each neighbourhood and their coordinates

Description: list of pubs with beer gardens in each neighbourhood and their coordinates, using categoryId = 4bf58dd8d48988d11b941735 ('pub') and query='beer garden' to return the required information.

Purpose: to identify beer gardens, informing exploratory descriptive analysis and Borough segmentation through clustering.

Source: Foursquare search/explore query API

Format: .json file

Link: https://api.foursquare.com/v2/venues/explore?client_id={}
&client_secret={}&ll={},{}&v={}&categoryId=4bf58dd8d48988d11b941735&query=beer garden&limit=100'.format(lat, long)

7. Borough boundary polygons

Description: file containing latitude and longitude coordinates specifying the geographical boundaries of each Borough.

Purpose: to enable 1) correct assignment of pubs to Boroughs during processing of the Foursquare query results, 2) visualisation of clusters of Boroughs after clustering.

Source: https://skgrange.github.io/data.html

Format: .json file

Link: https://skgrange.github.io/www/data/london_boroughs.json

Methodology

The key elements of the project comprised an exploratory data analysis, high-level inferential statistical analysis and a machine learning-enabled k-means clustering model. Each will be described below. Data sources are listed in the Data section.

Exploratory Data Analysis

The objective of the exploratory data analysis ('EDA') was to compile a relevant dataset which would be used to solve the problem. This involved steps or Feature Selection and Analysis.

Feature selection

I decided to use:

- Borough population
- Borough area
- % of population allowed to drink alcohol (non-Muslim % used as proxy)
- % of population allowed to visit a pub (% of population of working age used as a proxy)
- Pubs with beer gardens at a neighbourhood level
- Foursquare API call as described in the Data section. Elimination of duplicates. Removal of out-of-London results.
- Calculation of derived metrics: Drinkers/km2, Drinkers/Pop %, Pubs/km2, Population/Pub and Drinkers/Pub
- Average distance to nearest beer garden for all pubs in Borough.

Analysis steps

1. Import London Boroughs data from csv in webpage and develop dataframe
2. Add demographic data re population of working age and alcohol drinkers
3. Scrape Wikipedia page for neighbourhood-level info (for Foursquare API call) using BeautifulSoup
4. Add latitude and longitude to neighbourhoods using Geocoder Python package
5. Get neighbourhood pubs using Foursquare venues/explore API request using neighbourhood lat/long coordinates¶- Auto radius (not specifying a radius variable) was used to prevent under reporting of pubs
6. Remove duplicates by calculating distance of pub to duplicated neighbourhood centres and assigning the pub to the closest neighbourhood, and then delete instances of the same pub in other neighbourhoods
7. Correct for any inaccuracies in the Foursquare data and for out-of-London results using spatial join of pub coordinates and Borough polygons from the dataframe
8. Find the distance to the nearest venue (intra-Borough) for exploratory analysis and clustering¶
9. Combining three separate dataframes (geographic, demographic and pub data for each Borough)
10. Removing redundant variables – examples include Drinkers/km2 and Drinkers/Pub, due to the inclusion of Population/km2 and % of population allowed to drink alcohol.

The analysis resulted in 2230 pubs across the 33 Boroughs[1].

[1] City of London isn't technically a Borough but is considered as such here

Inferential Statistical Analysis

Inferential statistical testing – the correlations between variables - was conducted for reference only, as it wasn't needed because:

- The objective is to identify the differences between Boroughs based on the variables used - the cluster analysis in the next section will highlight the important features.
- I'm not trying to predict the future, so don't need to identify the most relevant variables for that purpose.

Machine Learning Modelling

A 'k-means' cluster analysis model was used to cluster and segment the Boroughs. The steps for this were:

1. Data pre-processing/normalisation
2. Determining the optimal k using the 'elbow method', involving a visual inspection of the plot of mean squared error at different k values.
3. Running the model with the optimal k to assign each Borough to one of k clusters
4. Consolidation of EDA results and clustering results into one dataframe
5. Visualisation of clustering results – using a Choropleth map plot to illustrate which cluster each Borough had been assigned.

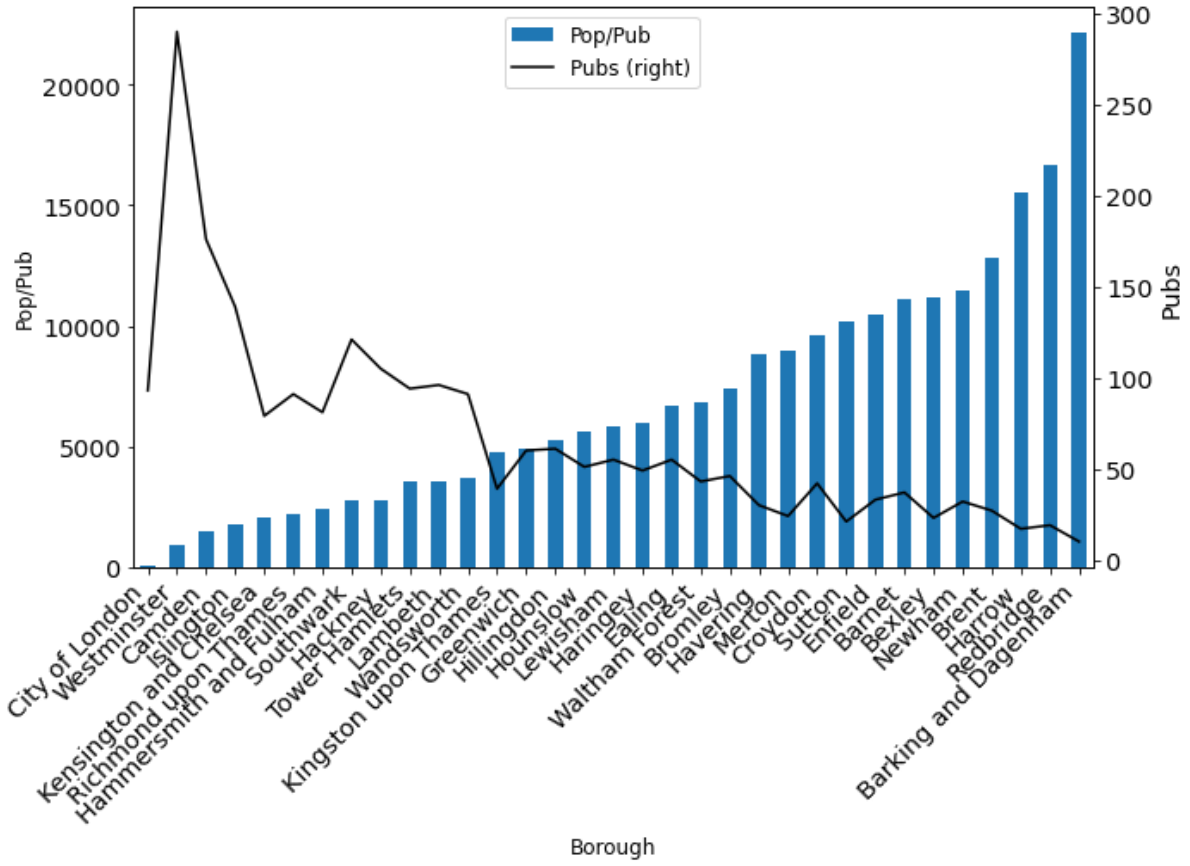
Results

Exploratory Data Analysis

The final dataframe used in the clustering model is presented below. There are 33 rows representing the 33 Boroughs and 11 columns for the Borough and ten measurement metrics (model variables).

	Borough	Population	Area km2	Pop/km2	Working_ Age_%	Drinkers_ %	Pubs	NN	Pubs/km2	Pop/Pub
0	Barking and Dagenham	221495	36.1	6135	63	75	10	745	0.3	22149
1	Barnet	411275	86.7	4743	65	88	37	720	0.4	11115
2	Bexley	256845	60.6	4238	63	95	23	765	0.4	11167
3	Brent	346437	43.2	8019	68	70	27	558	0.6	12831
4	Bromley	339466	150.1	2261	63	97	46	589	0.3	7379
5	Camden	259344	21.8	11896	71	84	176	144	8.1	1473
6	City of London	8164	2.9	2815	73	86	93	84	32.1	87
7	Croydon	403461	86.5	4664	65	91	42	549	0.5	9606
8	Ealing	369685	55.5	6660	67	87	55	399	1	6721
9	Enfield	346635	80.8	4290	64	85	33	729	0.4	10504
10	Greenwich	294837	47.3	6233	68	94	60	329	1.3	4913
11	Hackney	292023	19	15369	72	87	105	212	5.5	2781
12	Hammersmith and Fulham	195981	16.4	11950	72	89	81	199	4.9	2419
13	Haringey	291330	29.6	9842	71	88	49	286	1.7	5945
14	Harrow	263484	50.5	5217	65	88	17	654	0.3	15499
15	Havering	265930	112.4	2365	62	98	30	857	0.3	8864
16	Hillingdon	319467	115.7	2761	66	87	61	530	0.5	5237
17	Hounslow	286947	56	5124	68	85	51	375	0.9	5626
18	Islington	244372	14.9	16400	75	92	139	170	9.3	1758
19	Kensington and Chelsea	161552	12.1	13351	69	90	79	179	6.5	2044
20	Kingston upon Thames	184660	37.3	4950	67	89	39	411	1	4734
21	Lambeth	342250	26.8	12770	75	89	96	216	3.6	3565
22	Lewisham	320574	35.1	9133	70	93	55	307	1.6	5828
23	Merton	214740	37.6	5711	67	94	24	339	0.6	8947
24	Newham	366943	36.2	10136	70	58	32	391	0.9	11466
25	Redbridge	316288	56.4	5607	65	72	19	822	0.3	16646
26	Richmond upon Thames	203312	57.4	3542	65	96	91	252	1.6	2234
27	Southwark	332679	28.9	11511	74	94	121	226	4.2	2749
28	Sutton	213340	43.8	4870	64	93	21	461	0.5	10159
29	Tower Hamlets	331620	19.8	16748	74	61	94	195	4.7	3527
30	Waltham Forest	292788	38.8	7546	68	77	43	466	1.1	6809
31	Wandsworth	337783	34.3	9847	73	94	91	254	2.7	3711
32	Westminster	262317	21.5	12200	72	82	290	106	13.5	904

Large differences between the Boroughs are presented in the plot below, where two important metrics (Population/Pub and Pubs) for the clustering model are plotted on a combined bar and line plot. Referencing the above table, it can be seen City of London has the best (lowest) Population/Pub of 87 while Barking and Dagenham has the worst at 22,149. From the line plot, City of London has 93 pubs, Westminster is the best with 290, while Barking and Dagenham is also the worst for this metric with only 10 pubs.



Inferential Statistical Analysis

The Pandas method .corr() was used to produce the below table of correlations between each variable. Working Age %, Pubs and NN (average nearest neighbour in metres) were moderately/strong correlated to Pop/Pub. As discussed previously, no variables were dropped after this because the clustering model would highlight key differences between the Boroughs.

	Population	Area km2	Pop/km2	Working_Age_%	Drinkers_%	Pubs	NN	Pubs/km2	Pop/Pub
Population	1	0.44	0.07	-0.09	-0.2	-0.1	0.34	-0.63	0.27
Area km2	0.44	1	-0.69	-0.72	0.27	-0.4	0.69	-0.49	0.34
Pop/km2	0.07	-0.69	1	0.81	-0.3	0.56	-0.7	0.2	-0.44
Working_Age_%	-0.09	-0.72	0.81	1	-0.17	0.63	-0.9	0.53	-0.68
Drinkers_%	-0.2	0.27	-0.3	-0.17	1	0.04	-0	-0.02	-0.31
Pubs	-0.13	-0.4	0.56	0.63	0.04	1	-0.7	0.55	-0.69
NN	0.34	0.69	-0.65	-0.85	-0.03	-0.7	1	-0.58	0.84
Pubs/km2	-0.63	-0.49	0.2	0.53	-0.02	0.55	-0.6	1	-0.53
Pop/Pub	0.27	0.34	-0.44	-0.68	-0.31	-0.7	0.84	-0.53	1

Machine Learning Modelling – Results of sklearn k-means Clustering Model

The below table shows the cluster number, renamed ‘Zone’ for familiarity with Londoners (due to the public transport charging zones), added to the dataframe. City of London is on its own in Zone 1, there are ten Boroughs in each of Zones 2 and 3, seven Boroughs in Zone 4, and five Boroughs in

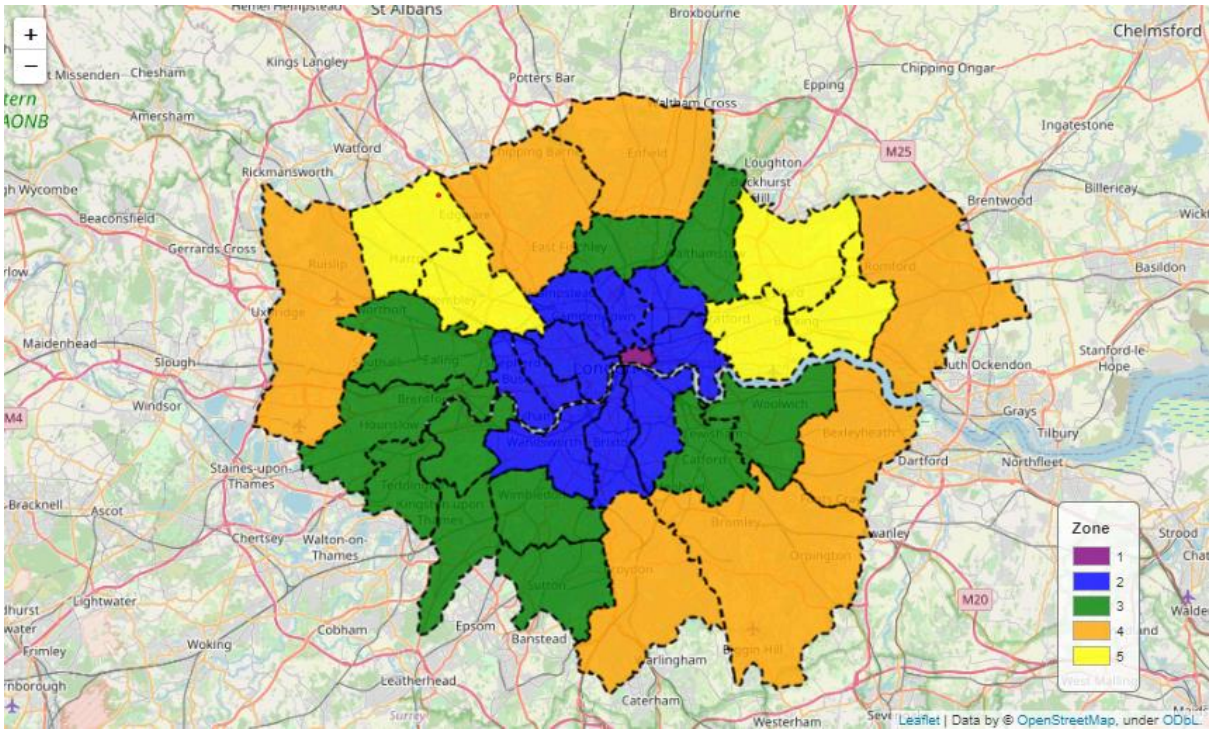
Zone 5. Boroughs and their zones are visualised on a Folium map of Great London and immediate surroundings.

Column	Zone	Borough	Population	Area km2	Pop/km2	Working_Age_%	Drinkers_%	Pubs	NN	Pubs/km2	Pop/Pub
0	1	City of London	8164	2.9	2815	73	86	93	84	32.1	87
1	2	Camden	259344	21.8	11896	71	84	176	144	8.1	1473
2	2	Hackney	292023	19	15369	72	87	105	212	5.5	2781
3	2	Hammersmith and Fulham	195981	16.4	11950	72	89	81	199	4.9	2419
4	2	Islington	244372	14.9	16400	75	92	139	170	9.3	1758
5	2	Kensington and Chelsea	161552	12.1	13351	69	90	79	179	6.5	2044
6	2	Lambeth	342250	26.8	12770	75	89	96	216	3.6	3565
7	2	Southwark	332679	28.9	11511	74	94	121	226	4.2	2749
8	2	Tower Hamlets	331620	19.8	16748	74	61	94	195	4.7	3527
9	2	Wandsworth	337783	34.3	9847	73	94	91	254	2.7	3711
10	2	Westminster	262317	21.5	12200	72	82	290	106	13.5	904
11	3	Ealing	369685	55.5	6660	67	87	55	399	1	6721
12	3	Greenwich	294837	47.3	6233	68	94	60	329	1.3	4913
13	3	Haringey	291330	29.6	9842	71	88	49	286	1.7	5945
14	3	Hounslow	286947	56	5124	68	85	51	375	0.9	5626
15	3	Kingston upon Thames	184660	37.3	4950	67	89	39	411	1	4734
16	3	Lewisham	320574	35.1	9133	70	93	55	307	1.6	5828
17	3	Merton	214740	37.6	5711	67	94	24	339	0.6	8947
18	3	Richmond upon Thames	203312	57.4	3542	65	96	91	252	1.6	2234
19	3	Sutton	213340	43.8	4870	64	93	21	461	0.5	10159
20	3	Waltham Forest	292788	38.8	7546	68	77	43	466	1.1	6809
21	4	Barnet	411275	86.7	4743	65	88	37	720	0.4	11115
22	4	Bexley	256845	60.6	4238	63	95	23	765	0.4	11167
23	4	Bromley	339466	150.1	2261	63	97	46	589	0.3	7379
24	4	Croydon	403461	86.5	4664	65	91	42	549	0.5	9606
25	4	Enfield	346635	80.8	4290	64	85	33	729	0.4	10504
26	4	Havering	265930	112.4	2365	62	98	30	857	0.3	8864
27	4	Hillingdon	319467	115.7	2761	66	87	61	530	0.5	5237
28	5	Barking and Dagenham	221495	36.1	6135	63	75	10	745	0.3	22149
29	5	Brent	346437	43.2	8019	68	70	27	558	0.6	12831
30	5	Harrow	263484	50.5	5217	65	88	17	654	0.3	15499
31	5	Newham	366943	36.2	10136	70	58	32	391	0.9	11466
32	5	Redbridge	316288	56.4	5607	65	72	19	822	0.3	16646

Summary tables of Zones, Boroughs and mean value of clustering variables per Zone

Zone	Boroughs	B count	Population	Area km2	Pop/km2	Working_Age_%	Drinkers_%	Pubs	NN	Pubs/km2	Pop/Pub
1	City of London	1	8164	2.9	2815	73	86	93	84	32.1	87
2	Camden, Hackney, Hammersmith and Fulham, Islington, Kensington and Chelsea, Lambeth, Southwark, Tower Hamlets, Wandsworth, Westminster	10	275992	21.6	13204	72	86	127	190	6.3	2493
3	Ealing, Greenwich, Haringey, Hounslow, Kingston upon Thames, Lewisham, Merton, Richmond upon Thames, Sutton, Waltham Forest	10	267221	43.8	6361	67	89	48	362	1.1	6191
4	Barnet, Bexley, Bromley, Croydon, Enfield, Havering, Hillingdon	7	334725	99	3617	64	91	38	677	0.4	9124
5	Barking and Dagenham, Brent, Harrow, Newham, Redbridge	5	302929	44.5	7022	66	72	21	634	0.5	15718

Borough segmentation



Discussion of Results

Each of the five Zones is discussed below.

Zone 1 is a special case as it only contains one Borough – the City of London - London's financial heartland. It has the best Pubs/km^2 (pub density), Population/Pub (87) and NN (distance in metres to nearest pub), driven by a large number of pubs (93) in an area less than 3km^2.

The Population/Pub figure is artificially low and not a good like-with-like comparison with other Zones because 1) Population it understated because doesn't include the sizeable commuting workforce who don't live in the City of London, 2) the number of Pubs reflects demand from this non-resident commuting workforce.

The 93 Pubs figure feels high (based on my local knowledge) – it is likely that the Foursquare definition of 'beer garden' covers anything from a temporary standing-room-only outdoor space to a permanent expansive garden with seating. It is also possible, though unlikely, that the Foursquare search algorithm is returning pubs stated as having 'no beer garden' in reviews.

Zone 2's ten Boroughs surround Zone 1 and have the highest average Pop/km^2 (population density) and Pubs. These Boroughs cover London's creative, cultural, entertainment and tourism industries which have synergies with and so drive demand for pubs.

Zone 3 also has ten Boroughs but differs from Zone 2 due to a lower higher Pop/km^2 which is driven by a lower average number of Pubs per Borough. It is middle of the pack for all measurement metrics. My hypothesis is that Zone 3 Boroughs are the most prosperous of the outer London Boroughs – I'd test this is future projects using household income data.

Zone 4 is characterised by the most populous and biggest Boroughs – seven Boroughs with an average area of 99km^2. This is more than twice that of any other Zone, which largely drives the second-worst Pop/Pub. All of the Boroughs in the Zone are on London's outskirts.

Zone 5 has five Boroughs in two blocks in the north-east and north-west. It is characterised by the lowest Pubs and Drinkers % (defined as 1 – Muslim % in the Data section) figures and highest Pop/Pub figure. It isn't clear if there is any link between Drinkers % and Pubs, and I don't need this information to make a recommendation. However, if I was to repeat this analysis I would refine the Drinkers % methodology by identifying migrant populations and their propensity to visit pubs.

Recommendations

"Which Boroughs should businesses target for launching beer gardens?"

With the current information, my recommendation would be for businesses to explore Zones 4 and 5 for development on the basis that:

1. Zones 4 and 5 have the most pent-up demand due to the lowest Pub and poor Pop/Pub figures.
2. The cost of launching in business in Zones 4 and 5 are likely significantly lower than in Zones 1, 2 and 3
3. Zones 1, 2 and 3 are already well-served with pubs.

Conclusions

There is pent-up demand from customers who want to return to COVID-safe pubs, especially pubs with beer gardens in which to enjoy the great British summer. Given their outdoors nature, a beer garden improves COVID-safety, and they are more important than ever given society's heightened sense of COVID-safety and UK Government guidance.

However, not all Boroughs are equally blessed with beer gardens, so an analysis was undertaken to identify Boroughs which businesses should target for launching beer gardens.

My recommendation based on the results of the analysis is that the following Boroughs be targeted as they have the most pent-up demand due to the lowest Pub and poor Population/Pub figures:

- Zone 5: Barking and Dagenham, Brent, Harrow, Newham, Redbridge
- Zone 4: Barnet, Bexley, Bromley, Croydon, Enfield, Havering, Hillingdon.

The analysis could be refined further by expanding the dataset to include:

- household income
- availability of public transport
- business costs
- local pricing/customer price sensitivity
- migrant populations and their propensity to visit pubs.