Capstone Project - London Borough Crime Rates and Venues

Applied Data Science Capstone by IBM/Coursera

Table of Contents

1. Introduction

1.1 Background

Rising crime rates in some areas have been linked to a lack of amenities and 'things to do' for young people in those areas.

For example, access to green spaces such as parks may help reduce crime[1], and reducing youth clubs may increase crime rates[2].

If there is a link between types of venue and crime rate, then this can help with planning in the future for which venues to build and fund in order to prevent crime, and which to avoid. It might also suggest areas that require more police attention and patrols if there is a concentration of certain types of venues in those areas.

1.2 Problem

This report will attempt to investigate whether crime rates in London boroughs are related to the number or type of venues in those boroughs. It will do this by trying to answer two questions.

1.2.1 Question 1:

Are crime rates lower or higher in areas with more venues? Are they higher or lower depending on the number of venues of different types?

---

[1] https://theconversation.com/can-parks-help-cities-fight-crime-118322
[2] https://www.barnardos.org.uk/news/new-research-draws-link-between-youth-service-cuts-and-rising-knife-crime

1.2.2 Question 2:

Are crime rates affected by the types of venues in an area? Does the combination of types of venues in an area have an impact on crime rates?

1.3 Audience

Stakeholders are local and UK government, police, town planners, community and volunteer groups.

2 Data Sources and Cleaning

2.1 Data Sources

The data on crimes in all London boroughs for the past 24 months was gathered by the Metropolitan Police and is available through the London Datastore here. I used the data from December 2018 through November 2020.

The estimated population for each London borough was scraped from Wikipedia.

Geometric data for London boroughs came from two sources, their central coordinates from the Office for National Statistics' (ONS) Geoportal, and a GeoJson file of their polygons from a repository by martinjc on Github.

The data on venues and venue types in each borough came from Foursquare, via the API.

2.2 Data Cleaning

The City of London was excluded as it is not technically a London Borough, and it has a very different population and venue profile to the rest of London.

I discovered the Wikipedia table had coordinates for each London borough, but when I plotted these on a map, they were not central to each borough, and so would not be ideal for a search. This is why I used the coordinates from the ONS instead.

In order to determine the proper search radius for the call to the API, I plotted a circle on a map of the largest borough by area in London, Bromley. The radius needed to cover the whole of Bromley was 12500 m, so that was the radius used.

Figure 1 Map of Bromley with search circle

With a radius that big, my searches inevitably overlapped with other boroughs, especially with the smaller boroughs. I wrote a function to check each venue was in the borough being searched before it was added to the list.

It was decided to restrict the venues examined to four categories, offering various options of 'things to do', rather than categories that offered only 'places to be' like restaurants or bars. This was to reduce the numbers of calls and results from the API, as gathering *all* venues per borough was clearly not feasible.

I used the categories available through the API, to give rough estimations.

The four categories are as follows:

1) Arts & Entertainment
    - for example, cinemas, music venues, bowling alleys, museums
2) Outdoor Recreation
    - for example, parks, tennis courts, football fields
3) Study
    - Universities, colleges and libraries (not schools)
4) Social and Religious
    - Community centres, social clubs and religious centres

The venue limit for this type of search through the API is 50, and I discovered I was hitting the cap for most of my searches. I therefore added 5 extra random smaller search circles per borough, to pick up extras and get more of a flavour of venue types in each borough. This added some duplicates to my dataset as well as new venues, and so I removed all duplicates.

As a side note, as part of investigating duplicates, I discovered some 'unique ids' for venues seem to be repeated in the dataset, or the data is not very reliable. I found one unique venue id was repeated in my data 20 times, in 20 different boroughs, all with different coordinates. This id was assigned to London Zoo, which is in fact on the border between Camden and Westminster, and not in the other 18 boroughs! I am therefore unsure as to the total reliability of the Foursquare data, while believing it still gives a good general idea.

3 Methodology

3.1 Calculations

Some basic calculations were needed to prepare the data for analysis

I totalled the number of crimes, and (using the data from Wikipedia) calculated the crime rate per 100,000 people as follows:

crime rate = (number of crimes / population) x 100000

	Borough	Population	longitude	latitude	Total_Crime	Crime_Rate
0	Barking and Dagenham	212906	0.129506	51.5455	39456	18532.12
1	Barnet	395896	-0.218210	51.6111	59167	14945.09
2	Bexley	248287	0.146212	51.4582	33929	13665.23
3	Brent	329771	-0.275680	51.5644	59257	17969.14
4	Bromley	332336	0.039246	51.3727	47499	14292.46

Figure 2 Dataframe showing crime rate calculated

I created two dataframes, one for each question.

The first contained the number of venues per borough per 100,000 people, calculated in the same way as crime rate.

	Borough	Population	longitude	latitude	Total_Crime	Crime_Rate	Arts	Outdoors	Study	Social and Religious	Total_Venues
0	Barking and Dagenham	212906	0.129506	51.5455	39456	18532.12	32.88	45.09	31.47	35.70	145.13
1	Barnet	395896	-0.218210	51.6111	59167	14945.09	18.94	26.02	16.92	17.93	79.82
2	Bexley	248287	0.146212	51.4582	33929	13665.23	26.58	36.25	26.98	26.18	115.99
3	Brent	329771	-0.275680	51.5644	59257	17969.14	36.09	47.61	32.45	30.02	146.16
4	Bromley	332336	0.039246	51.3727	47499	14292.46	19.86	33.70	18.96	20.46	92.98

Figure 3 Dataframe showing number of venues per capita calculated

The second contained all the venues I had found, ready to be used for K-means clustering.

	Borough	Latitude	Longitude	Venue	id	Venue Latitude	Venue Longitude	Venue Category
15406	Westminster	51.486248	-0.136422	Цитадель Зла	51335a52e4b08f70f97b92f5	51.491922	-0.131127	Temple
15407	Westminster	51.486248	-0.136422	Calvary Chapel Westminster	4bf8fd854a67c928ba3a26cf	51.494386	-0.144643	Church
15408	Westminster	51.486248	-0.136422	Church Without Walls Westminster	53be9614498ecfab3dd7efcf	51.496195	-0.140322	Church
15409	Westminster	51.486248	-0.136422	Trinity Church London	5cbe101f1987ec00391da72a	51.494026	-0.143262	Church
15410	Westminster	51.486248	-0.136422	Tate Britain Members Room	4d457345bbb1a1433e5f5272	51.490793	-0.127343	Café

Figure 4 Dataframe showing some of the collected venues

3.2 Exploratory Data Analysis

3.2.1 Question 1

I first used df.corr() to check for correlations, and found possible correlations between numbers of venues per 100,000 people and crime rate.

	Population	longitude	latitude	Total_Crime	Crime_Rate	Arts	Outdoors	Study	Social and Religious	Total_Venues
Population	1.000000	0.101452	0.205335	0.462454	-0.083029	-0.536884	-0.654368	-0.616275	-0.568212	-0.612635
longitude	0.101452	1.000000	0.058903	0.023742	-0.036664	-0.163587	-0.256997	-0.246945	-0.126620	-0.207623
latitude	0.205335	0.058903	1.000000	0.221116	0.150148	-0.034221	-0.206123	-0.095017	-0.014435	-0.091863
Total_Crime	0.462454	0.023742	0.221116	1.000000	0.836521	0.213974	-0.008658	0.111516	0.218465	0.137362
Crime_Rate	-0.083029	-0.036664	0.150148	0.836521	1.000000	0.607689	0.405440	0.539596	0.631287	0.562504
Arts	-0.536884	-0.163587	-0.034221	0.213974	0.607689	1.000000	0.887690	0.950977	0.961246	0.983443
Outdoors	-0.654368	-0.256997	-0.206123	-0.008658	0.405440	0.887690	1.000000	0.880012	0.865974	0.940002
Study	-0.616275	-0.246945	-0.095017	0.111516	0.539596	0.950977	0.880012	1.000000	0.937188	0.974450
Social and Religious	-0.568212	-0.126620	-0.014435	0.218465	0.631287	0.961246	0.865974	0.937188	1.000000	0.970326
Total_Venues	-0.612635	-0.207623	-0.091863	0.137362	0.562504	0.983443	0.940002	0.974450	0.970326	1.000000

Figure 5 Dataframe correlation

I plotted these, to explore them further.

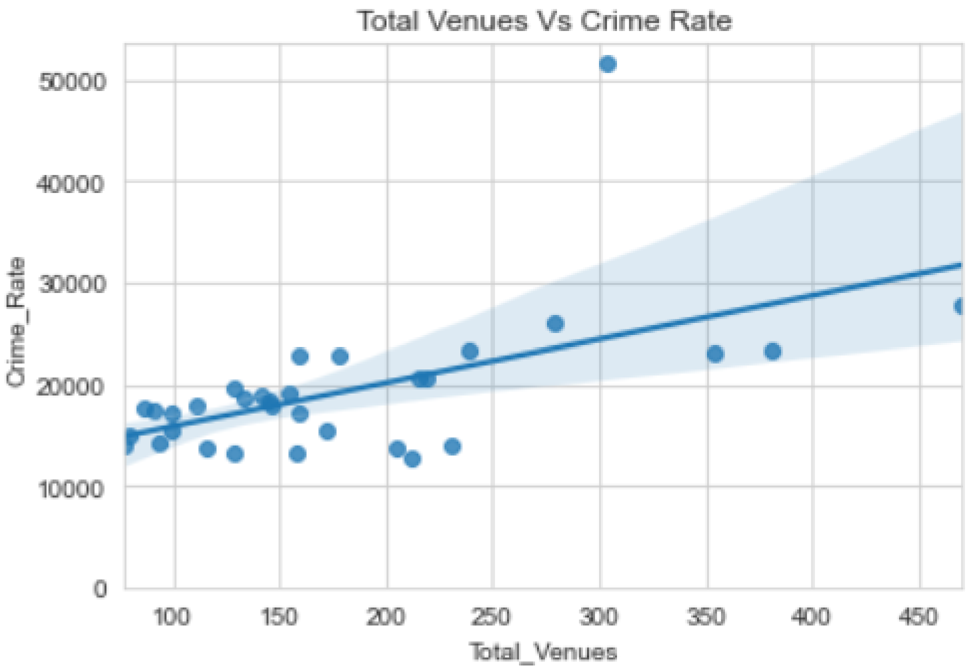

Figure 6 Plot of Total Venues vs Crime Rate

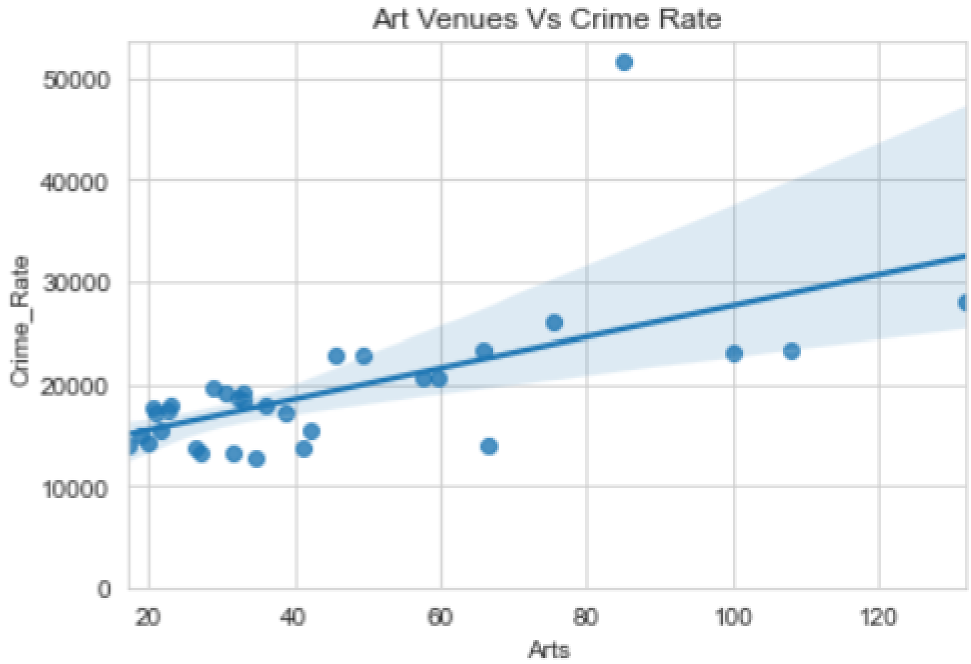

Figure 7 Plot of Art Venues vs Crime Rate

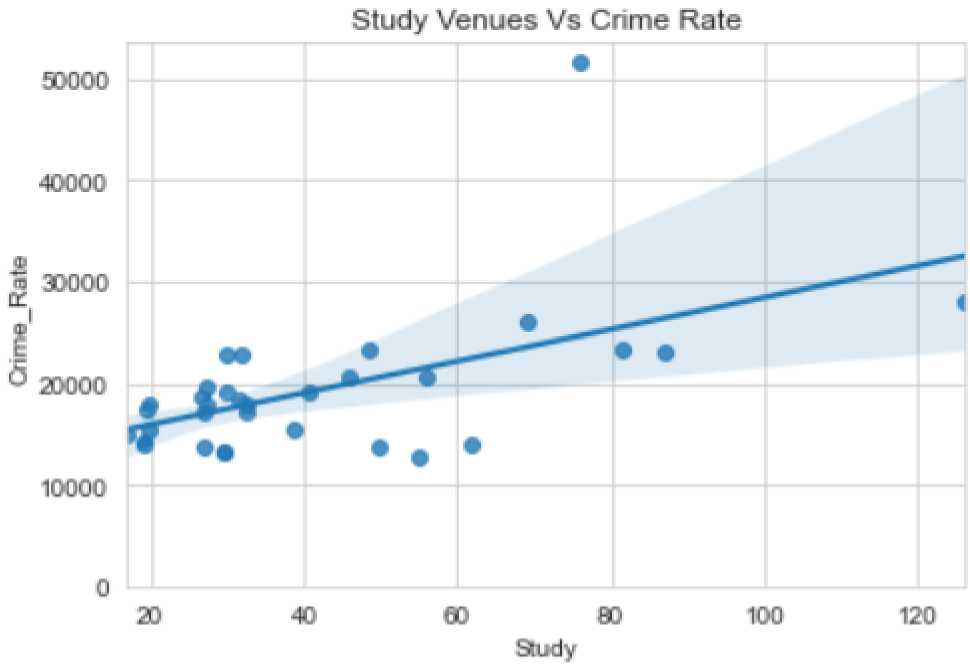

Figure 8 Plot of Study Venues vs Crime Rate

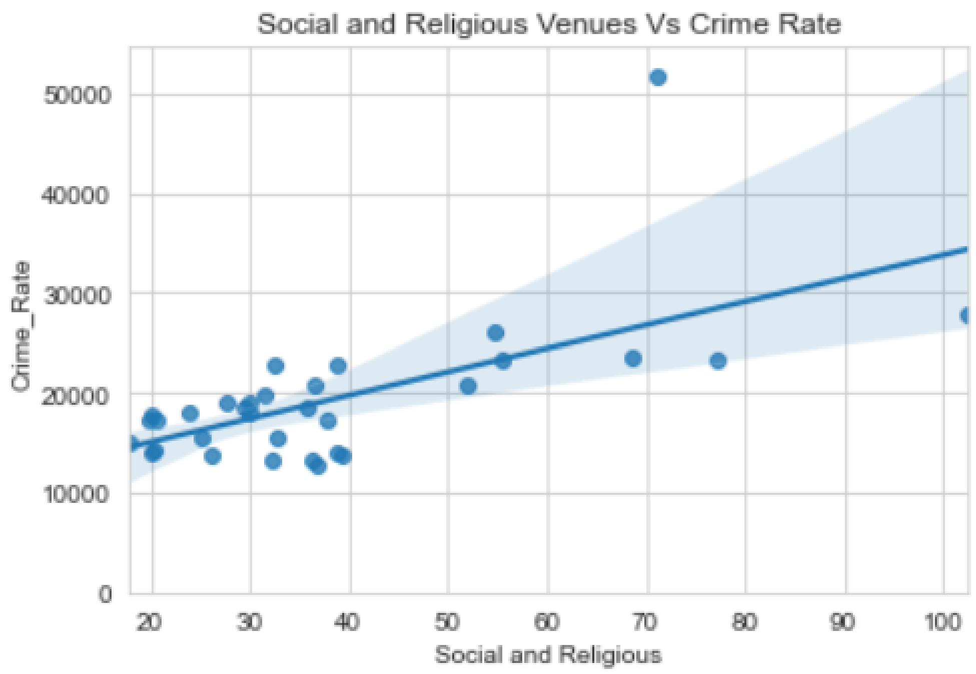

Figure 9 Plot of Social and Religious Venues vs Crime Rate

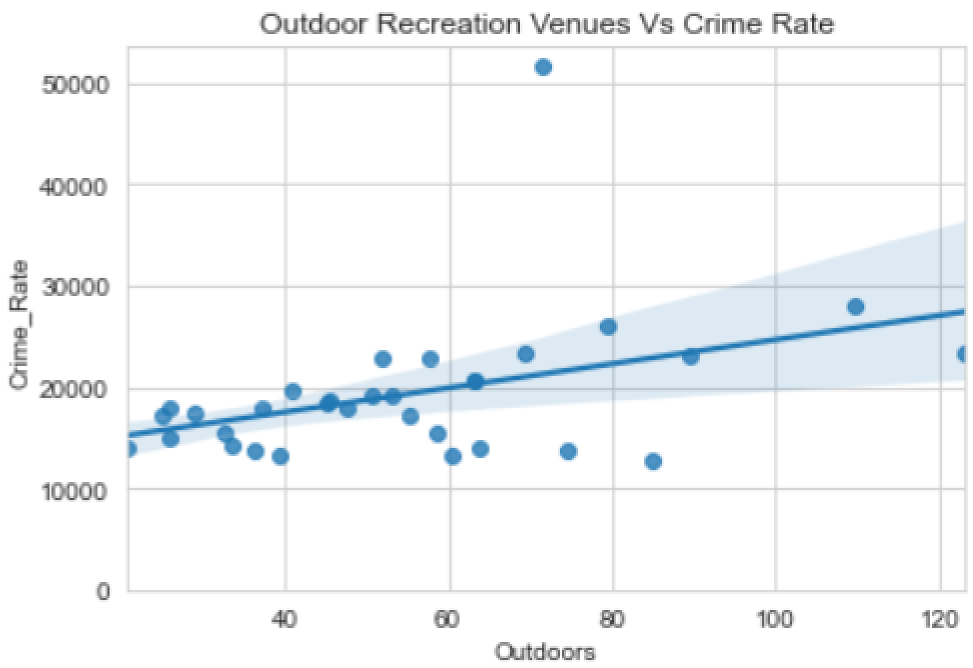

Figure 10 Plot of Outdoor Recreation Venues vs Crime Rate

All of these seem to have a positive correlation with crime rate in that, as the number of each venue type increases, so does the crime rate.

In order to discover if these correlations were statistically significant, I performed Pearson correlation tests. The results of this are given in the results section.

3.2.2 Question 2

3.2.2.1 K-means clustering

As I was going to use K-means clustering to divide the boroughs into categories, I first had to prepare the data for that by replacing categorical data with numbers and calculating means.

I then calculated the top 10 most common venues per borough.

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Barking and Dagenham	Church	Park	Social Club	College Academic Building	Library	General College & University	Gym / Fitness Center	University	Art Gallery	Community Center
1	Barnet	Church	Park	Library	General College & University	College Academic Building	Gym / Fitness Center	Plaza	Social Club	Garden	University
2	Bexley	Church	Park	Library	Historic Site	College Academic Building	Social Club	General College & University	Community Center	University	Gym / Fitness Center
3	Brent	Church	Park	Library	University	College Academic Building	College Classroom	Plaza	Golf Course	Historic Site	General Entertainment
4	Bromley	Church	Park	Historic Site	Library	College Academic Building	University	Community Center	Social Club	General College & University	Outdoor Sculpture

Figure 11 Dataframe with top 10 venue types

Before performing k-means clustering, I used the elbow method to determine the optimal value for k.

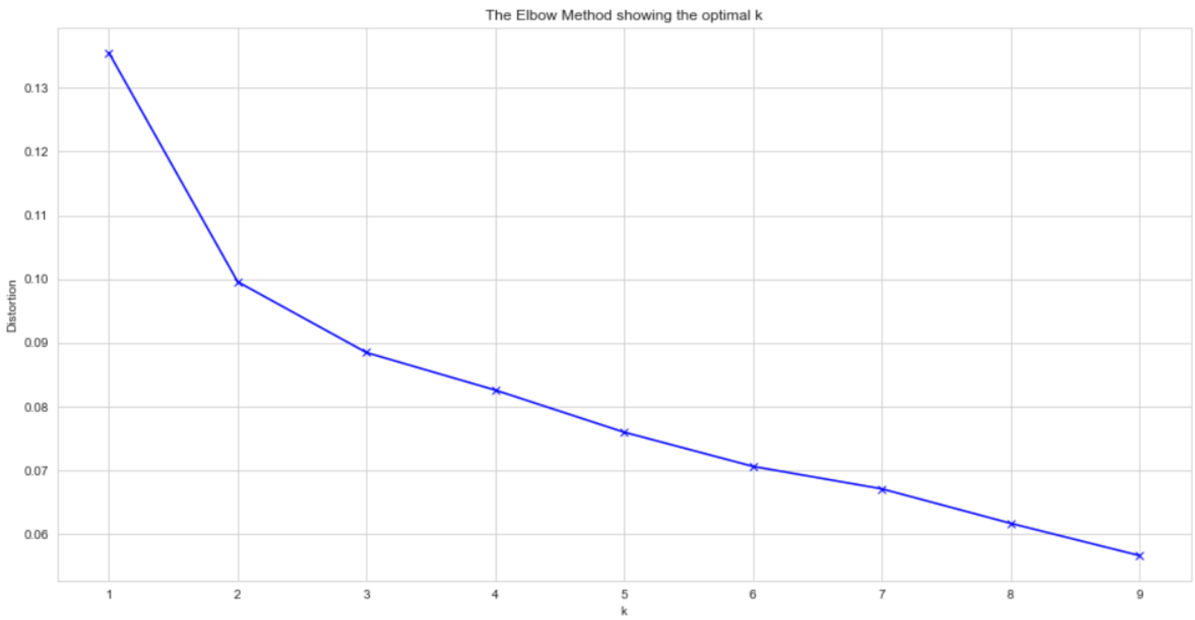

Figure 12 The Elbow Method showing the optimal k

As this was not entirely clear, I also used the silhouette score, which confirmed an optimal value for k as 2.

```
['2 : 0.2061669945491676',
 '3 : 0.17707182043569528',
 '4 : 0.1420152226189853',
 '5 : 0.09159348012545228',
 '6 : 0.11747048615985217',
 '7 : 0.062171409635601735',
 '8 : 0.0675178491109012',
 '9 : 0.07201417143295899',
 '10 : 0.06773546795493024']
```

Figure 13 Values for k vs silhouette scores

I performed the k-means clustering using a value of 2 for k, and examined the clusters formed.

I named these clusters ‘Fun London’ and ‘Sensible London’.

Both have churches as their most common venue, and most boroughs have a park as their second most common venue, but after that they diverge.

Fun London has more art galleries, music venues, theatres and outdoor spaces like gardens and playgrounds among its top 10 venues.

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	Camden	Church	Park	Plaza	University	Art Gallery	Library	Theater	Social Club	Gym / Fitness Center	Music Venue
9	Greenwich	Church	Park	College Academic Building	Library	University	Art Gallery	Gym / Fitness Center	Plaza	General College & University	Community Center
10	Hackney	Church	Park	Art Gallery	Gym / Fitness Center	Community Center	University	Library	Music Venue	Garden	Performing Arts Venue
11	Hammersmith and Fulham	Church	Park	Gym / Fitness Center	Art Gallery	General Entertainment	University	College Classroom	Garden	Gym	Historic Site
12	Haringey	Church	Park	Gym / Fitness Center	Music Venue	Historic Site	Library	University	Art Gallery	General Entertainment	Playground
17	Islington	Church	Park	Art Gallery	Music Venue	Library	General Entertainment	Gym / Fitness Center	Playground	General College & University	Garden
18	Kensington and Chelsea	Church	Art Gallery	Park	Gym / Fitness Center	Garden	College Academic Building	University	General College & University	Gym	Historic Site

Figure 14 Sample of Fun London

Sensible London has fewer of these and more colleges, libraries and sports fields.

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Barking and Dagenham	Church	Park	Social Club	College Academic Building	Library	General College & University	Gym / Fitness Center	University	Art Gallery	Community Center
1	Barnet	Church	Park	Library	General College & University	College Academic Building	Gym / Fitness Center	Plaza	Social Club	Garden	University
2	Bexley	Church	Park	Library	Historic Site	College Academic Building	Social Club	General College & University	Community Center	University	Gym / Fitness Center
3	Brent	Church	Park	Library	University	College Academic Building	College Classroom	Plaza	Golf Course	Historic Site	General Entertainment
4	Bromley	Church	Park	Historic Site	Library	College Academic Building	University	Community Center	Social Club	General College & University	Outdoor Sculpture
6	Croydon	Church	Park	Library	College Classroom	Theater	College Academic Building	Art Gallery	University	Community Center	Plaza
7	Ealing	Church	Park	Library	College Academic Building	University	Bridge	General College & University	Soccer Field	Gym / Fitness Center	General Entertainment
8	Enfield	Church	Park	Library	College Academic Building	General College & University	Art Gallery	Garden	University	Gym / Fitness Center	Tennis Court

Figure 15 Sample of Sensible London

I also examined the total number of each type of venue per cluster.

Cluster Names	Population	Total_Crime	Crime_Rate	Arts	Outdoors	Study	Social and Religious	Total_Venues
Fun London	4109075	941148	348433.91	2427	2736	2122	1931	9216
Sensible London	4843220	776814	269614.81	1333	1902	1319	1283	5837

Figure 16 Total numbers of venues per cluster

3.2.2.2 Plotting

I plotted the crime rates, divided into three categories, low, medium and high, on a map of London. On top of that I plotted the clusters as dots.

The red dots are Fun London, the purple dots are Sensible London.

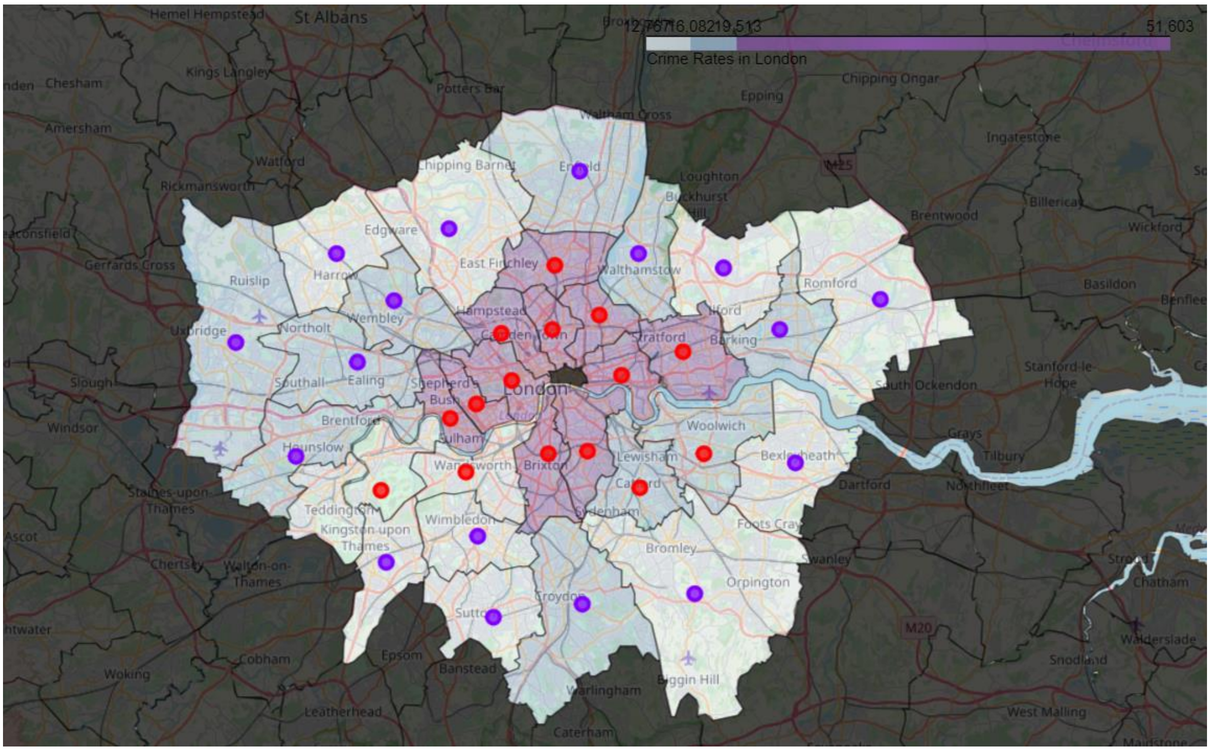

Figure 17 Map of London crime rate and clusters

As can be seen, there does appear to be some correlation between the clusters and crime rate, with Fun London being mostly in higher crime rate areas.

To test this for certain, I performed a T test. The results are in the results section.

4 Results

4.1 Question 1: Results of Pearson correlation tests

The null hypothesis is that there is no correlation between the number of total venues, or the number of each type of venues, and crime rate.

I have chosen a p value of 0.05, so p must be below that for the result to be statistically significant.

Venues	**Correlation coefficient**	**p**
Arts	0.608	0.0
Outdoors	0.405	0.021
Study	0.54	0.001
Social and Religious	0.631	0.0
Total Venues	0.563	0.001

From this we can **reject the null hypothesis.**

There is a correlation between numbers of venues, and numbers of types of venues and crime rate.

This correlation is **positive**, the more of each type of venue, the higher the crime rate.

It seems to be a fairly strong correlation, especially for arts venues and social and religious venues, with a correlation coefficient of approximately 0.6.

The coefficient for study venues is lower at 0.54, and lower still for outdoor venues at 0.405.

4.2 Question 2: Results of T test

The null hypothesis is that there is no difference between the mean crime rates for the two clusters.

The T test statistic was 3.377, with a p value of 0.002.

As p is smaller than 0.05, the null hypothesis is **rejected**.

There is a difference between the mean crime rates of the two clusters.

From the map, Fun London has a higher crime rate than Sensible London.

4.3 Results Overview

I have shown a positive correlation between the total number of venues and crime rates, and between the numbers of all types of venues and crime rates.

I have also shown a positive correlation between the types of venues in Fun London (music venues, theatres and art galleries) and crime rates.

However, it is very important to remember that correlation is not the same as causation, as is discussed below.

5 Discussion

So, having shown a correlation between the numbers of venues and crime rate, does this mean to reduce crime we should demolish amenities?

Clearly that sounds ridiculous.

Just because I have shown that two things are correlated, it does not mean that one caused the other.

In this case, I would say that it is highly unlikely to be true.

While it was interesting to try this approach, there are many other factors that impact on crime rates, such as poverty or drug use.

It may also be my list of venues was too broad, and I really needed to be able to focus on places intended for young people such as community centres or youth clubs, which was not possible at this time. Perhaps I also needed to focus more specifically on the types of crime committed by younger people. Were stakeholders interested in repeating this study, that is how I would set it up in the future.

I also found a correlation between the types of venues in Fun London (more music venues, theatres and art galleries) and a higher crime rate, and between the types of venues in Sensible London (more gyms, sports fields and colleges) and lower crime rates.

In this case correlation may indeed be causation! But perhaps not quite in the way this study meant.

It can be seen that, actually, Fun London has even more Study type venues than Sensible London does, just in different proportions to Art type venues, so adding more study venues to make Fun London more 'sensible' isn't the answer to reducing crime rates.

I suspect that the venues in Sensible London are not keeping crime rates lower by providing 'things to do' for young people who might otherwise be out committing crimes, but that the venues in Fun London are instead providing more opportunities for crime. Those 'fun' venues are likely attracting visitors from outside the area.

From the map, it can be seen that Fun London is really Central London, which has a large number of tourists and visitors, bars, clubs, pubs etc.

Tourists are more likely to be victims of crimes, as are people slightly the worse for wear at music venues and theatres!

Perhaps to reduce crime, without tearing down venues providing vital income, patrols in these areas need to be increased.

6 Conclusion and Future Direction

The main lesson from this is that causes behind crime rates are more complicated than can be easily discovered by simply looking at one factor.

Should stakeholders wish to investigate this particular factor further, I would suggest refining the study to specific types of venues most suitable for young people, and those types of crime committed mostly by young people. I would also suggest using a different data source for the venue data. If the study was looking only at community and youth clubs, perhaps the data could come from the local authorities.