IBM Data Science Professional Certificate

Applied Data Science Capstone London Businesses Benchmarking

Oussama KIASSI

16 July 2020

I Introduction

I.1 Background

As the capital of the United Kingdom, London attracts around 30M tourist from all over the world every year [1] . Its tourist industry is ever-growing, enlarging the country's economy. However, some tourists face some very hard experiences. They get harassed, smuggled, aggressed and live through other crimes. While it's impossible to completely control the city's safety, they naturally started to look for safe zones to get distracted and to live the English way. So, to fulfill their needs businessmen and small business owners started investing in the safest boroughs of the capital. Thus, our project plays the role of a benchmarking platform that gives the formers an idea on businesses in those safe areas. It primarily aims to cluster common venues of London's safest zones.

I.2 Interest

This project will benefit tourists by displaying the common, yet safest attractions. Consequently, it will allow a greater experience for the formers, while expanding London's tourist industry. Moreover, it will be of a great help to investors. Knowing the frequent venues in the safest boroughs of London, they will start their own businesses with considerable confidence.

II Data requirements

To resolve this problem, I used the following datasets:

- I scrapped London boroughs' population data from June 1981 to June 2019 (10 years spaced) from [2] .
- I found in the metropolitan police of London website a dataset of the city's crimes by borough during the last 24 months [3] .
- I then downloaded London boroughs geographical borders [4] .
- Using BeautifulSoup4, I scraped London areas from [5] .

- I used Foursquare API, “explore request” to retrieve the 200 nearest venues in a radius of 3Km [6] to locations from the areas’ dataset.

III Safe boroughs

In this section I worked on London’s criminality, I acquired data that will allow us to compute crime rate. Next, I visualized the results in form of an interactive map.

III.1 Preprocessing data

Primarily, I wrangled the population dataset.

```
This dataset size is : (34, 7)
```

	Status	PopulationEstimate1981-06-30	PopulationEstimate1991-06-30	PopulationEstimate2001-06-30	PopulationEstimat
Name					
Sutton	Borough	170200	170100	181500	
Tower Hamlets	Borough	144700	166300	201100	
Waltham Forest	Borough	217200	215900	222000	
Wandsworth	Borough	262400	262000	271700	
Greater London	Administrative Area	6805000	6829300	7322400	

Figure 1: Population data frame

I dropped different unnecessary columns containing some old dates (1989, 1999, etc.), then the row of Greater London (this one contains the sum of population).

Borough	Population by June 2019
Southwark	318830
Sutton	206349
Tower Hamlets	324745
Waltham Forest	276983
Wandsworth	329677

Figure 2: Clean population data frame

Secondly, I preprocessed the crime dataset.

```
This dataset size is : (1569, 27)
```

	MajorText	MinorText	LookUp_BoroughName	201807	201808	201809	201810	201811	201812	201901	...	201909	201910	201
0	Arson and Criminal Damage	Arson	Barking and Dagenham	6	5	3	8	5	1	5	...	6	9	
1	Arson and Criminal Damage	Criminal Damage	Barking and Dagenham	127	101	107	132	105	88	97	...	109	109	
2	Burglary	Burglary - Business and Community	Barking and Dagenham	30	18	33	32	39	33	45	...	37	30	
3	Burglary	Burglary - Residential	Barking and Dagenham	94	84	99	94	106	164	114	...	80	97	
4	Drug Offences	Drug Trafficking	Barking and Dagenham	8	7	10	7	7	4	5	...	7	8	

Figure 3: Crime data frame

First and foremost, I printed the violations in the latter, because I want to sort boroughs by crime rate which depends only on crimes. Accordingly, I dropped the rows that do not contain any crime. Moreover, I only left June 2019 values because it is the most recent date in the population dataset. Afterward, I computed the crime rate:

$$\text{Crime rate} = \frac{\text{Crimes} \times 1000}{\text{Population}} \; [7]$$

and I joined it with the final crime dataset.

	Borough	Crimes of June 2019	Crime rate
0	Barking and Dagenham	699	3.283139
1	Barnet	1229	3.104562
2	Bexley	549	2.211151
3	Brent	1096	3.323518
4	Bromley	929	2.795364

Figure 4: Clean crime data frame

III.2 Data visualization

I used for data visualization the folium library that contains functions for interactive map plotting. First, I computed boroughs centroids' coordinates from the GeoJSON file to draw boroughs' labels. Then I visualised London detailed choropleth map.

	Borough	Latitude	Longitude
0	Kingston upon Thames	51.3867	-0.2836
1	Croydon	51.3572	-0.0843892
2	Bromley	51.3675	0.0555781
3	Hounslow	51.4686	-0.348958
4	Ealing	51.5223	-0.33116

Figure 5: Centroids of London boroughs

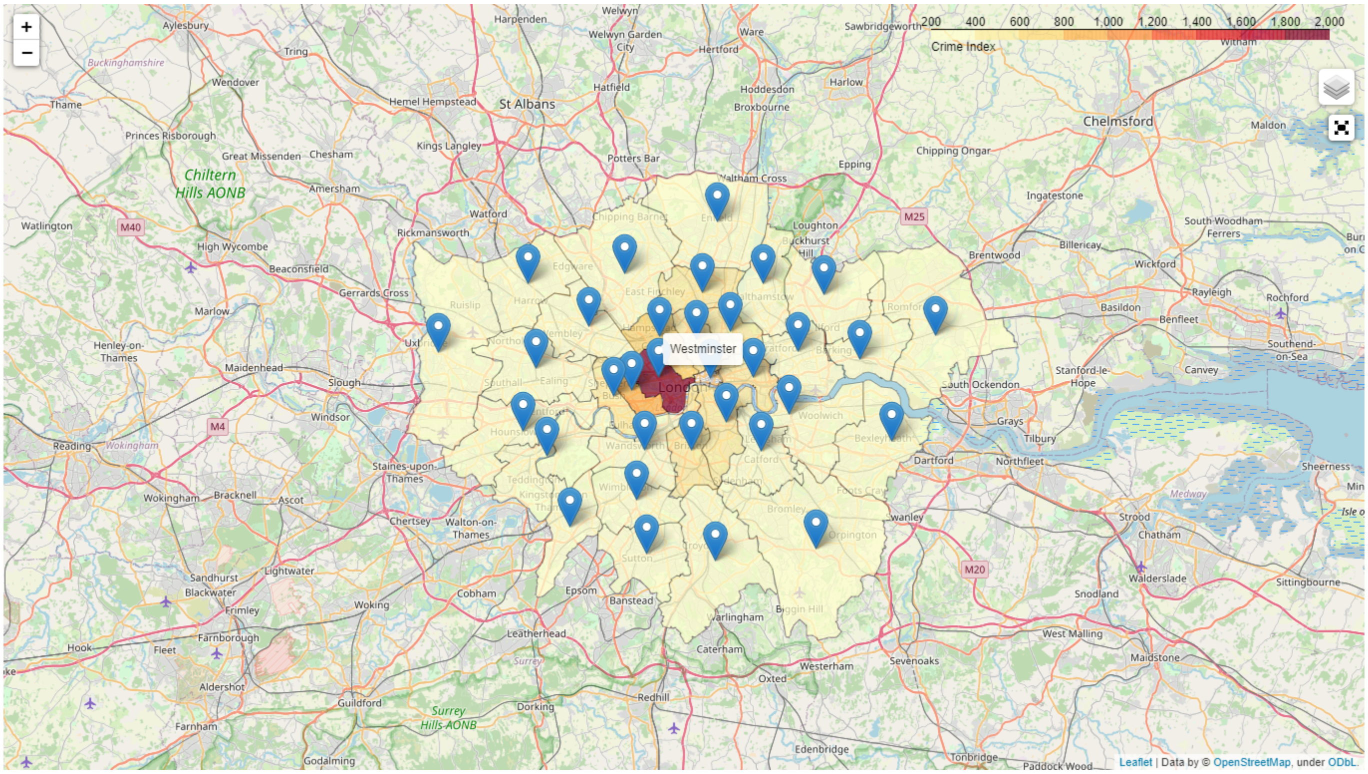


Figure 6: Choropleth map of London with markers

IV Common secure venues

IV.1 Preprocessing

Firstly, I chose the threshold of safety as 4 to retrieve the most secure boroughs from the previous crime dataset.

	Borough	Crimes of June 2019	Crime rate
0	Barking and Dagenham	699	3.283139
1	Barnet	1229	3.104562
2	Bexley	549	2.211151
3	Brent	1096	3.323518
4	Bromley	929	2.795364

Figure 7: Secure boroughs data frame

Secondly, I worked on the areas' dataset.

```
This dataset size is : (533, 6)
```

	Location	London borough	Post town	Postcode district	Dial code	OS grid ref
0	Abbey Wood	Bexley, Greenwich [7]	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham[8]	LONDON	W3, W4	020	TQ205805
2	Addington	Croydon[8]	CROYDON	CR0	020	TQ375645
3	Addiscombe	Croydon[8]	CROYDON	CR0	020	TQ345665
4	Albany Park	Bexley	BEXLEY, SIDCUP	DA5, DA14	020	TQ478728

Figure 8: Areas data frame

I only kept areas where town was equivalent to London, then I removed any number from the borough column, also the unnecessary columns.

	Location	Borough	Town	Postcode	Dial code	OS grid ref
0	Abbey Wood	Bexley, Greenwich	LONDON	SE2	020	TQ465785
1	Acton	Ealing, Hammersmith and Fulham	LONDON	W3, W4	020	TQ205805
2	Aldgate	City	LONDON	EC3	020	TQ334813
3	Aldwych	Westminster	LONDON	WC2	020	TQ307810
4	Anerley	Bromley	LONDON	SE20	020	TQ345695

Figure 9: Areas data frame 2nd version

Since some areas are located in two boroughs, I split them to rows where each contains only one. Then I dropped the first rows keeping only the ones with one borough.

(332, 3)

	Location	Borough	Postcode
0	Abbey Wood	Bexley	SE2
1	Abbey Wood	Greenwich	SE2
2	Acton	Ealing	W3, W4
3	Acton	Hammersmith and Fulham	W3, W4
4	Aldgate	City	EC3

Figure 10: Areas data frame 3rd version

Likewise, some areas seem to have two postcodes. So, I appended rows containing just the first one and dropped the unsplit ones.

(332, 3)

	Location	Borough	Postcode
0	Abbey Wood	Bexley	SE2
1	Abbey Wood	Greenwich	SE2
2	Acton	Ealing	W3, W4
3	Acton	Hammersmith and Fulham	W3, W4
4	Aldgate	City	EC3

Figure 11: Areas data frame 4th version

Besides, I combined the previous knowledge with this latter to result with the safest areas in London. Furthermore, I used the Coords function to get each area coordinates.

(165, 5)

	Location	Borough	Postcode	Latitude	Longitude
0	Abbey Wood	Bexley	SE2	51.492450	0.121270
1	Acton	Ealing	W3	51.513240	-0.267460
2	Anerley	Bromley	SE20	51.410090	-0.056830
3	Arkley	Barnet	EN5	51.644415	-0.179183
4	Arnos Grove	Enfield	N11	51.616310	-0.138390

Figure 12: Areas data frame 5th version

Following this further, I used the Foursquare API to request the nearest 200 venues in a radius of 3Km to the locations' positions.

This dataset size is : (15533, 7)

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbey Wood	51.49245	0.12127	The Plumstead Pantry	51.481712	0.083707	Café
1	Abbey Wood	51.49245	0.12127	Lesnes Abbey	51.489526	0.125839	Historic Site
2	Abbey Wood	51.49245	0.12127	Lidl	51.496152	0.118417	Supermarket
3	Abbey Wood	51.49245	0.12127	Dagenham Sunday Market	51.517026	0.111949	Flea Market
4	Abbey Wood	51.49245	0.12127	Sainsbury's	51.492826	0.120524	Supermarket

Figure 13: Foursquare API request

Then I computed and sorted venues in each neighborhood by frequency.

```
Abbey Wood :
                   Venue  Frequency
0          Grocery Store      0.167
1            Supermarket      0.139
2                   Park      0.083
3                    Pub      0.056
4   Fast Food Restaurant      0.056


Acton :
                       Venue  Frequency
0                Coffee Shop       0.11
1                        Pub       0.08
2                       Café       0.07
3                       Park       0.06
4   Middle Eastern Restaurant      0.05
```

Figure 14: Venues by frequency in neighborhood

In addition, I assembled the data of each neighborhood into one data frame.

```
This dataset size is : (152, 11)
```

	Neighbourhood	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	Abbey Wood	Grocery Store	Supermarket	Park	Fast Food Restaurant	Train Station	Pub	Bakery	Flea Market	Coffee Shop	Clothing Store
1	Acton	Coffee Shop	Pub	Café	Park	Middle Eastern Restaurant	Gastropub	Bakery	Gym / Fitness Center	Grocery Store	Pizza Place
2	Anerley	Pub	Park	Italian Restaurant	Coffee Shop	Café	Gastropub	Gym / Fitness Center	Indian Restaurant	Pizza Place	Grocery Store
3	Arkley	Coffee Shop	Pub	Café	Grocery Store	Italian Restaurant	Park	Turkish Restaurant	Supermarket	Restaurant	Pharmacy
4	Arnos Grove	Café	Park	Bakery	Grocery Store	Coffee Shop	Greek Restaurant	Turkish Restaurant	Pub	Gym / Fitness Center	Portuguese Restaurant

Figure 15: Venues data frame

Finally, I only kept the numerical data so that I can proceed with clustering.

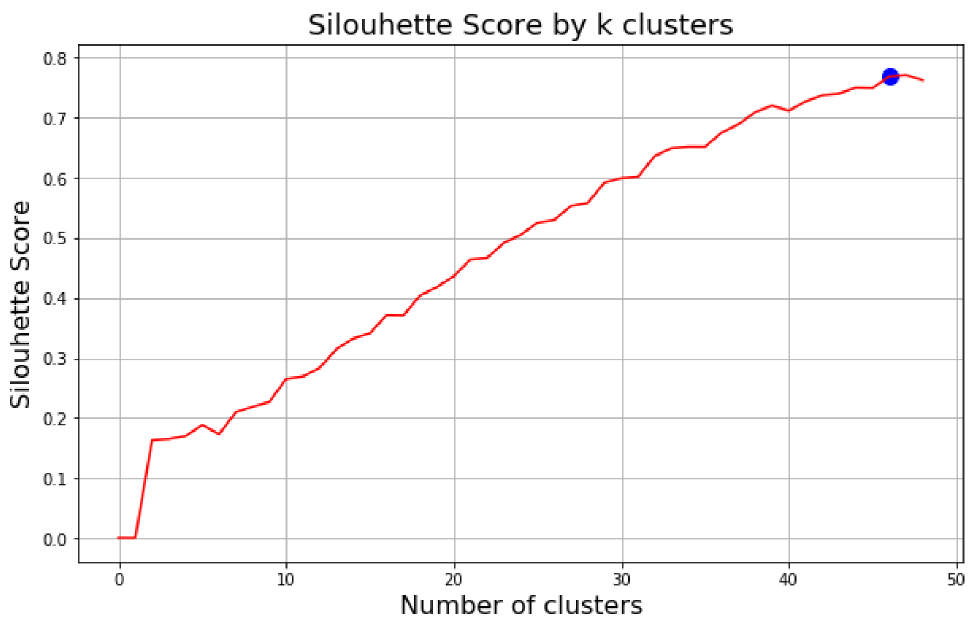
```
This dataset size is : (152, 284)
```

	Accessories Store	Afghan Restaurant	African Restaurant	Airport	American Restaurant	Antique Shop	Aquarium	Arcade	Argentinian Restaurant	Art Gallery	...	Vietnamese Restaurant	Warehouse Store
0	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	0.00	0.0	...	0.00	0.027778
1	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	0.01	0.0	...	0.00	0.000000
2	0.0	0.0	0.0	0.0	0.00	0.01	0.0	0.0	0.00	0.0	...	0.01	0.000000
3	0.0	0.0	0.0	0.0	0.00	0.00	0.0	0.0	0.00	0.0	...	0.00	0.000000
4	0.0	0.0	0.0	0.0	0.01	0.00	0.0	0.0	0.00	0.0	...	0.00	0.000000

Figure 16: Venues data frame (dummies)

IV.2 K-Means Clustering

I looped through 100 iteration to determine the optimal number of clusters which corresponds to a maximal silhouette score. In fact, I didn't iterate until 100, since I used a heuristic approach. I broke the loop when the algorithm didn't get information after two iterations.




```
best silouhette score is with 46 clusters.
```

Figure 17: Silhouette plot

It seems like the optimal number of clusters is 46. After, I integrated the clustering in the dataset.

	Location	Borough	Postcode	Latitude	Longitude	Cluster Labels	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 M Comr Ve
0	Abbey Wood	Bexley	SE2	51.492450	0.121270	7	Grocery Store	Supermarket	Park	Fast Food Restaurant	Train Station	Pub	Ba
1	Acton	Ealing	W3	51.513240	-0.267460	39	Coffee Shop	Pub	Café	Park	Middle Eastern Restaurant	Gastropub	Ba
2	Anerley	Bromley	SE20	51.410090	-0.056830	43	Pub	Park	Italian Restaurant	Coffee Shop	Café	Gastropub	Gy Fit Ce
3	Arkley	Barnet	EN5	51.644415	-0.179183	29	Coffee Shop	Pub	Café	Grocery Store	Italian Restaurant	Park	Tur Restau
4	Arnos Grove	Enfield	N11	51.616310	-0.138390	9	Café	Park	Bakery	Grocery Store	Coffee Shop	Greek Restaurant	Tur Restau

Figure 18: Secure venues clustered data frame

Lastly, I displayed the clusters.

	Location	Borough	Cluster Labels	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
105	North Finchley	Barnet	0	Coffee Shop	Pub	Grocery Store	Turkish Restaurant	Supermarket	Café	Park	Italian Restaurant	Japanese Restaurant	Gym / Fitness Center
163	Woodside Park	Barnet	0	Coffee Shop	Pub	Grocery Store	Turkish Restaurant	Supermarket	Café	Park	Italian Restaurant	Japanese Restaurant	Gym / Fitness Center

Figure 19: Clusters

IV.3 Results

This part is the most interesting. We are going to give data proven answers to users' questions.

- What could I do like a business in this specific borough?

```
Would you like to know the common businesses in a specific borough (1)? or the places with high interest of your business
(2)? or businesses that go along with yours (3)?  1

 In which borough do you want to start your business?  newham
Indian Restaurant
Hotel
Pub
Grocery Store
```

Figure 20: First question

- Where should I start this type of business?

```
Would you like to know the common businesses in a specific borough (1)? or the places with high interest of your business
(2)? or businesses that go along with yours (3)?  2

 What business are you interested in?  pub
Bromley
Barnet
Ealing
Lewisham
Greenwich
Brent
Waltham Forest
Hounslow
Merton
Newham
Redbridge
Croydon
```

Figure 21: Second question

- What businesses go along with mine in this location?

```
Would you like to know the common businesses in a specific borough (1)? or the places with high interest of your business
(2)? or businesses that go along with yours (3)?  3

 What is your business?  pub

 Where is it located?  newham
```

	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
56	Café	Park	Indian Restaurant	Grocery Store	Mediterranean Restaurant	Restaurant	Bar	Toy / Game Store	Ice Cream Shop
92	Café	Park	Grocery Store	Restaurant	Bar	Coffee Shop	Bistro	Pizza Place	Mediterranean Restaurant
134	Café	Park	Grocery Store	Restaurant	Bar	Coffee Shop	Bistro	Pizza Place	Mediterranean Restaurant

Figure 22: Third question

- As for the error messages, herewith one of the examples:

```
Would you like to know the common businesses in a specific borough (1)? or the places with high interest of your business
(2)? or businesses that go along with yours (3)?  0

 Please respond by the number allocated to the question!
Would you like to know the common businesses in a specific borough (1)? or the places with high interest of your business
(2)? or businesses that go along with yours (3)?  1

 In which borough do you want to start your business?  j

 We think that the borough you have chosen isn't safe or is not a part of Greater London. You might look up another one.

 In which borough do you want to start your business?
```

Figure 23: Error message

V Conclusion

In this project, I gave some really interesting information about businesses in London. In fact, I made a choropleth map visualizing danger zones with high crime rate for tourists. Since they are usually interested by safe attractions, I tried to convert this disposition into a business need. Where I made a business benchmarking platform in favor of businessmen and small business owners, that gives critical information about those safest boroughs.

VI Discussion

Truth be told, this project will be more precise if we had access to recent data about population. However, it will shift the tourist industry mindset into a more need-oriented one which will be advantageous for the different stakeholders.

"Data science for need-oriented businesses."

References

[1] London's tourism industry, http://www.uncsbrp.org/tourism.htm.

[2] London boroughs' populations from june 1991 to june 2019, https://www.citypopulation.de/en/uk/greaterlondon/.

[3] London crimes by borough in the last 24 months (metropolitan data), https://data.london.gov.uk/download/recorded_crime_summary/d2e9ccfc-a054-41e3-89fb-53c2bc3ed87a/MPS%20Borough%20Level%20Crime%20%28most%20recent%2024%20months%29.csv.

[4] London boroughs geographical borders, https://skgrange.github.io/www/data/london_boroughs.json.

[5] London areas, https://en.wikipedia.org/wiki/List_of_areas_of_London.

[6] Foursquare api, https://developer.foursquare.com/.

[7] What exactly does "crime rate" mean and how do you calculate it?, https://ukcrimestats.com/blog/faqs/what-exactly-does-crime-rate-mean-and-how-do-you-calculate-it/, 20 mai 2020.