Using K-Means Approach to Cluster Boroughs in London.

Pratique Kain
January 3, 2021

Introduction

This report is produced as part of the final assignment for IBM's Data Science Capstone Project. The report provides a clustering of boroughs in London using the K-Means clustering algorithm. Boroughs in London have been categorized into four clusters based on the K-Means analysis.

The primary source for data was Wikipedia's page on the list of boroughs in London. Geocoder was used to identify the coordinates of each borough, which were then used to fetch venue data from the foursquare API. The code for the data analysis and visualization has been provided on my GitHub account.

The project concludes with an analysis of cluster 0 and cluster 3. We conclude that if a tourist in London is looking for din-in restaurants they should go to boroughs in cluster 3. If they are looking for cafes and pubs, their best options are boroughs in cluster 0 (the first cluster)

Problem Definition

Imagine you are visiting London for a few days and want to get to know the city besides the famous tourist destinations that Google recommends to you. You would probably benefit from understanding how the 32 boroughs of London are clustered.

At a very high level clustering is done on the basis of in-sample similarity of entities. As such if you are able to cluster the various boroughs you can efficiently break down your tour of London based on the kind of places you want to visit.

This project clusters the 32 boroughs based on venue data using the Foursquare Places API.

Data & Python Libraries Used

For this project I have used the following data points:

1. List of boroughs in London from Wikipedia[1]
2. Once we have scraped the boroughs from the Wikipedia page we use the geocoder library in python to identify their coordinates.
3. Foursquare Places API[2] to identify venues in the vicinity of each borough.
4. We fetch the category of the venue using a python function.

In addition to the data that was utilized for this project, I am providing a list of the libraries needed to perform this analysis. The underlying code for the analysis is available on my GitHub page.

1. *Pandas* : to analyze the data we scrape in a tabular format. Most of the preprocessing of the data is performed using pandas (in my approach)
2. *Folium* : used for mapping the data.
3. *Requests* : we use requests to fetch data in its raw format from Wikipedia before processing it using the scraper and Pandas.
4. *K Means* : clustering algorithm. There are a lot of free resources to understand the underlying logic for the algorithm. At a very high level we are clustering each observation based on its distance from an assigned mean value.
5. *Geocoder* : to fetch coordinates for each borough in London. We will need coordinates to fetch data on venues from the foursquare API.
6. *Beautiful Soup* : the web scraper needed to parse and organize data from the Wikipedia page on boroughs in London.

All of the analysis is done in a Jupyter notebook.

Methodology and Analysis

The first step in the clustering approach is gathering data on boroughs one wishes to analyze. For that I used the list of boroughs in London from Wikipedia. I fetched the data in its html format before applying the beautiful soup web scraper to parse the table into a pandas format. London is divided in 32 boroughs and you can see a list of the first five below.

	London borough
0	Camden
1	Greenwich
2	Hackney
3	Hammersmith
4	Islington

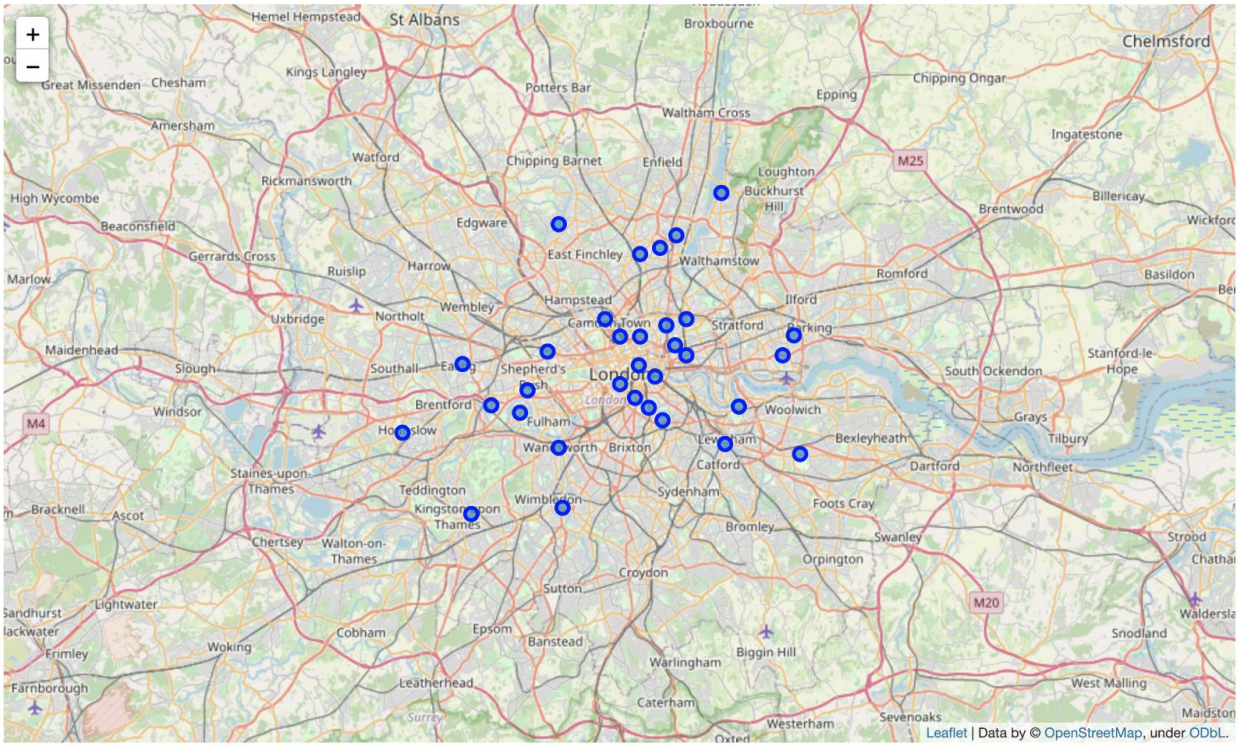
The first five boroughs in London in a pandas dataframe. There are 32 in total.

The next step in our analyses is to get the coordinates for each of the 32 boroughs so we can use the foursquare api to fetch venues data for each location. As before, I am providing a list of the updated dataframe with the first five boroughs and their respective coordinates below.

	London borough	Latitude	Longitude
0	Camden	51.53236	-0.12796
1	Greenwich	51.48454	0.00275
2	Hackney	51.54505	-0.05532
3	Hammersmith	51.49617	-0.22935
4	Islington	51.53279	-0.10614

London boroughs with their coordinates.

The information in the above data frame is used to map the boroughs and fetch venues data from the foursquare API. The foursquare API provides information about venues close to the location (one can define the radius) based on the coordinates the user provides. Using the coordinates above we will use the API to categorize the venues in these boroughs.



We use one-hot encoding to identify the categories for venues in each borough. Below is a compressed list of categories with one-hot encoding for Camden. Please note that these are just the first five rows. The entire data frame comprises all one-hot encoded categories for each borough in London.

	Neighborhood	Afghan Restaurant	African Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	...	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant
0	Camden	0	0	0	0	0	0	0	0	0	...	0	0	0
1	Camden	0	0	0	0	0	0	0	0	0	...	0	0	0
2	Camden	0	0	0	0	0	0	0	0	0	...	0	0	0
3	Camden	0	0	0	0	0	0	0	0	0	...	0	0	0
4	Camden	0	0	0	0	0	0	0	0	0	...	0	0	0

5 rows × 225 columns

Snapshot of venue categories in Camden — one hot encoded.

We proceed to group categories by their mean across the 32 boroughs. This is done since we are using the K Means clustering approach. These mean values will be used to assess similarity of boroughs when clustering them.

	Borough	Afghan Restaurant	African Restaurant	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	...	Vegetarian / Vegan Restaurant	Video Game Store	Vietnamese Restaurant
0	Barking	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
1	Barnet	0.0	0.0	0.0	0.0	0.016667	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
2	Bexley	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
3	Brent	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
4	Bromley	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0

5 rows × 225 columns

Categories grouped by mean values across Boroughs in London.

Based on the data above we can sort the ten most common venues across the different boroughs in London. Below is a snapshot of the first five boroughs and their ten most common venue categories. Once we have this level of analyses, we are ready to implement the K Means algorithm to cluster boroughs.

	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	Gym Pool	Pub	Sandwich Place	Chinese Restaurant	Event Space	Food Stand	Food & Drink Shop	Food	Flower Shop	Flea Market
1	Barnet	Coffee Shop	Pub	Café	Park	Restaurant	Bagel Shop	Bakery	Furniture / Home Store	Thrift / Vintage Store	Jewelry Store
2	Bexley	Indian Restaurant	Botanical Garden	Home Service	Park	Farmers Market	Food Truck	Food Stand	Food & Drink Shop	Food	Flower Shop
3	Brent	Golf Course	Playground	Chinese Restaurant	Tennis Court	Metro Station	Yoga Studio	Farmers Market	Food Stand	Food & Drink Shop	Food
4	Bromley	Bar	Stadium	Mediterranean Restaurant	Train Station	Supermarket	Turkish Restaurant	Soccer Stadium	Pub	Sporting Goods Shop	Hostel

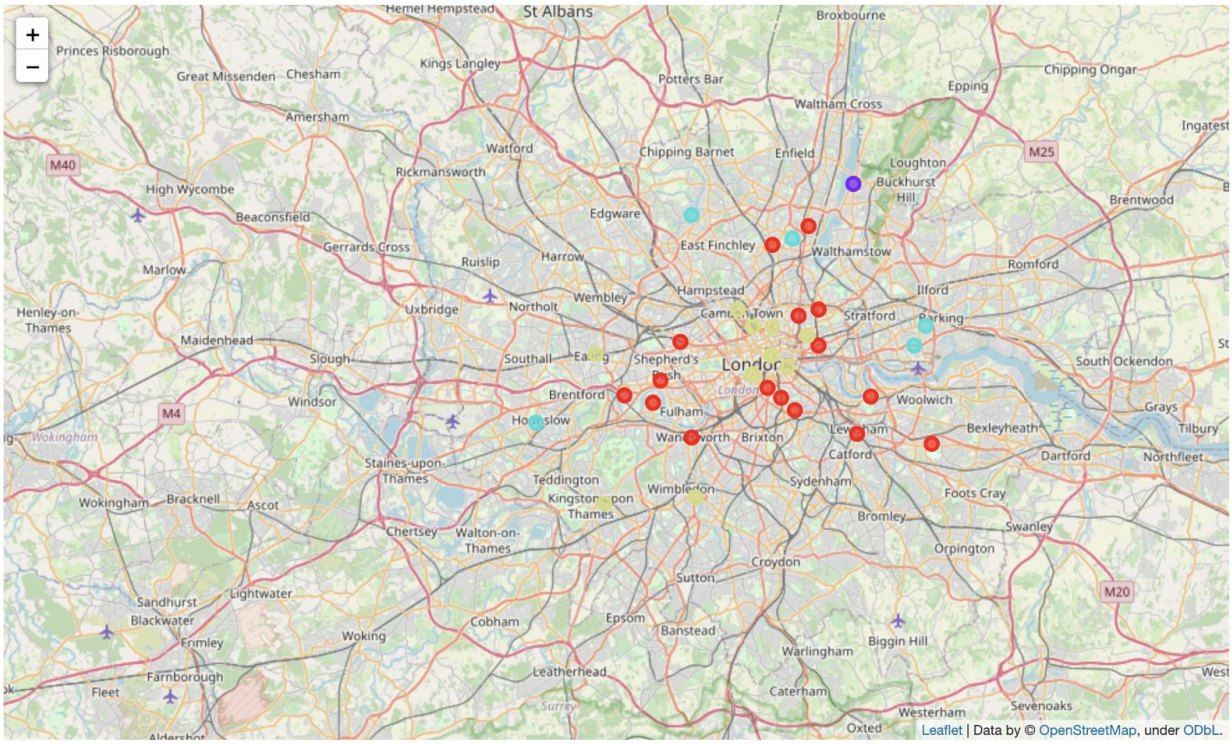
Ten most common venue categories across the 32 boroughs.

Based on the K Means clustering algorithm, we have the following clusters:

London borough	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 Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Camden	51.53236	-0.12796	0	Café	Coffee Shop	Hotel	Breakfast Spot	Park	Burger Joint	Pizza Place	Mexican Restaurant	Escape Room	Plaza
Greenwich	51.48454	0.00275	1	Pub	Grocery Store	Pizza Place	Garden	Café	Coffee Shop	Italian Restaurant	Park	Japanese Restaurant	Burger Joint
Hackney	51.54505	-0.05532	1	Pub	Café	Bakery	Coffee Shop	Brewery	Supermarket	Park	Pizza Place	Italian Restaurant	Vietnamese Restaurant
Hammersmith	51.49617	-0.22935	1	Pub	Café	Coffee Shop	Hotel	Indian Restaurant	Park	Italian Restaurant	Sandwich Place	Gym / Fitness Center	Japanese Restaurant
Islington	51.53279	-0.10614	0	Pub	Coffee Shop	Arts & Crafts Store	Mediterranean Restaurant	Hotel	French Restaurant	Café	Yoga Studio	Restaurant	Pizza Place

This is a snapshot of London Boroughs and the clusters they have been assigned.

We then proceed to map these clusters.



Cluster map of boroughs in London using K Means Algorithm

The boroughs have been clustered into four categories. For simplicity let us look at one of the clusters and see how that information could benefit a tourist. Below is a table of the venues available at cluster 0 (the first cluster).

	London 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	Camden	Café	Coffee Shop	Hotel	Breakfast Spot	Park	Burger Joint	Pizza Place	Mexican Restaurant	Escape Room	Plaza
4	Islington	Pub	Coffee Shop	Arts & Crafts Store	Mediterranean Restaurant	Hotel	French Restaurant	Café	Yoga Studio	Restaurant	Pizza Place
8	Southwark	Coffee Shop	Pub	Seafood Restaurant	Restaurant	Italian Restaurant	Hotel	Bakery	Scenic Lookout	Asian Restaurant	Garden
11	Westminster	Hotel	Café	Coffee Shop	Plaza	Garden	Monument / Landmark	Outdoor Sculpture	Park	Indian Restaurant	Historic Site
13	Barnet	Pub	Café	Coffee Shop	Yoga Studio	Restaurant	Wine Bar	Beer Bar	Italian Restaurant	Cocktail Bar	Pilates Studio
18	Ealing	Coffee Shop	Pub	Park	Hotel	Italian Restaurant	Café	Burger Joint	Pizza Place	Thai Restaurant	Portuguese Restaurant
21	Harrow	Pub	Coffee Shop	Theater	Hotel	Falafel Restaurant	Gym / Fitness Center	Art Museum	Park	Scenic Lookout	Event Space
22	Havering	Coffee Shop	Pub	Café	Yoga Studio	Pizza Place	Grocery Store	Music Venue	Market	Middle Eastern Restaurant	Burger Joint
25	Kingston upon Thames	Pub	Coffee Shop	Clothing Store	Café	Gym / Fitness Center	Indian Restaurant	Sushi Restaurant	Italian Restaurant	Thai Restaurant	Japanese Restaurant
26	Merton	Grocery Store	Pub	Coffee Shop	Hotel	Italian Restaurant	Thai Restaurant	Bar	Park	Tram Station	Sushi Restaurant

The boroughs above have been clustered together based on the venue categories. For instance if a tourist in London is looking for a good coffee shop, their best bet would be to visit one of the boroughs in the first cluster (above). However, if the tourist is interested in grabbing some groceries from a supermarket or is interested in sit-in dining, they would probably benefit more from visiting one of the boroughs in cluster 3 (below).

	London 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
12	Barking	Hotel	Fast Food Restaurant	Supermarket	Park	Coffee Shop	Indian Restaurant	Grocery Store	Sandwich Place	Clothing Store	Bakery
15	Brent	Café	Grocery Store	Indian Restaurant	Japanese Restaurant	Metro Station	Supermarket	Turkish Restaurant	Coffee Shop	Fast Food Restaurant	Persian Restaurant
17	Croydon	Grocery Store	Pub	Bus Stop	Fast Food Restaurant	Park	Turkish Restaurant	Coffee Shop	Café	Chinese Restaurant	Museum
24	Hounslow	Indian Restaurant	Hotel	Coffee Shop	Clothing Store	Grocery Store	Fast Food Restaurant	Bakery	Pizza Place	Pub	Pharmacy
27	Newham	Grocery Store	Park	Supermarket	Gym / Fitness Center	Light Rail Station	Shopping Plaza	Farm	Coffee Shop	Furniture / Home Store	Soccer Field

Conclusion

In this project I have attempted to cluster the boroughs in London using venues data from the foursquare API. This has been done using the K Means clustering algorithm. It is important to note that K Means is not the only clustering algorithm out there but one I have chosen because of my familiarity with it. Neither is it necessary to use four as the number of clusters you might want to divide your boroughs into. I ran the elbow method to analyze the optimal distribution of the number of clusters and arrived at two. I used four because of a slightly lower distortion score (score of how dissimilar members within a cluster are).

Sources

[1] https://en.wikipedia.org/wiki/London_boroughs

[2] https://foursquare.com/city-guide