The Battle of Neighborhoods -Applied Data Science Capstone Project

„What's the best place to search for a flat in London for a Sushi lover"

Introduction

Background:

London is a great multicultural city with plenty of opportunities for everyone. It is a home to great tourist attractions, food and drinks. Being a home to 8.9 million people, there are plenty of places to pick to look for a flat to rent.

With this much variety and opportunity, it's difficult to know where to search.

Problem Description:

Our target user searches for a neighborhood in London to find a flat. He/She is a great Sushi lover, as such their priorities are to find a place that is safe, as well as provides plenty of opportunities to enjoy their favorite food.

I intend to help them with this problem, by finding and filtering to the safest neighborhoods of London. Then for each neighborhood finding what Sushi places are available, and clustering these together to help them decide what would be their best place to live.

Data Acquisition

The data used in this project will consist of the London's Recorded Crime for past 2 years, List of broughs and the FourSuare API.

The list of boroughs & their geographical location will come from Wikipedia:

https://en.wikipedia.org/wiki/List_of_London_boroughs

This data will be used to determine geographical location of each borough.

Crime statistics will come from London database with geographical breakdown:

https://data.london.gov.uk/dataset/recorded_crime_summary

This data will be used to filter the London's boroughs based on safety

This data is saved as a csv file, will need to be extracted and saved

The foursquare API, where a list of sushi restaurants will be requested.

https://api.foursquare.com

This data will be used to search for best sushi restaurants

This data will need to be extracted from the HTML, and will need to be filtered to the relevant information

Methodology

	MajorText	MinorText	LookUp_BoroughName	201903	201904	201905	201906	201907	201908	201909	...	202004	20
0	Arson and Criminal Damage	Arson	Barking and Dagenham	5	5	11	3	5	3	6	...	2	
1	Arson and Criminal Damage	Criminal Damage	Barking and Dagenham	138	130	140	113	134	118	109	...	80	
2	Burglary	Burglary - Business and Community	Barking and Dagenham	29	27	21	27	31	35	37	...	29	

As a first step in the analysis the data containing Boroughs crime levels is downloaded, as this data will provide us two things; the sum of crimes committed in each borough, and the names of those boroughs.

	Borough Name	Sum
0	Barking and Dagenham	37630
1	Barnet	55803
2	Bexley	31822
3	Brent	56196
4	Bromley	44735

This data required grouping, cleaning and sorting in order to make it useable. There were a lot of features available in this dataset, which could've been used for advanced search, however were not in scope of this activity.

	borough_name	coordinates
0	Barking and Dagenham [note 1]	51°33′39″N 0°09′21″E / 51.5607°N 0.1557°E /...
1	Barnet	51°37′31″N 0°09′06″W / 51.6252°N 0.1517°W /...
2	Bexley	51°27′18″N 0°09′02″E / 51.4549°N 0.1505°E /...
3	Brent	51°33′32″N 0°16′54″W / 51.5588°N 0.2817°W /...
4	Bromley	51°24′14″N 0°01′11″E / 51.4039°N 0.0198°E /...

The next step is to download the boroughs data from Wikipedia, this will provide us the geographical location of each borough. There are many other aspects in this data, however we will filter to what is required for us.

Using requests and beautiful soup we can extract the data useful for us.

Once cleaned, we can add sum of crimes to this data. Then we can sort it by and filter it by boroughs where there were less than 45000 crimes, as our target user is interested in safety.

	borough_name	latitude	longitude	crime_sum
0	Kingston upon Thames	51.4085	0.3064	23228.0
1	Southwark	51.5035	0.0804	23905.0
2	Tower Hamlets	51.5099	0.0059	25553.0
3	Newham	51.5077	0.0469	26707.0
4	Harrow	51.5898	0.3346	31514.0
5	Bexley	51.4549	0.1505	31822.0
6	Havering	51.5812	0.1837	34076.0
7	Barking and Dagenham	51.5607	0.1557	37630.0
8	Hammersmith and Fulham	51.4927	0.2339	40623.0
9	Kensington and Chelsea	51.5020	0.1947	40644.0
10	Bromley	51.4039	0.0198	44735.0

For each borough we connect to the FourSquare API to request information about ‚Sushi' restaurants in radius of 6000m of the location of the borough.

This data then needs to be cleaned and filtered to the data important to us. In this case it's the borough name, the name of the restaurant, geographical location and the distance from borough location.

	borough	name	lat	long	distance
1	Kingston upon Thames	YO! Sushi	51.438786	0.268810	4263.0
2	Kingston upon Thames	Umami Sushi Box	51.440078	0.370440	5667.0
3	Southwark	Thames Barrier Sushi	51.500633	0.033207	3285.0
4	Southwark	Sushi Japanese	51.456053	0.010815	7153.0
5	Southwark	Sushi Ya	51.507601	0.022780	4018.0

As part of exploring the data, it was beneficial to make a map of London using folium, and add markers of the safest London's boroughs to this map. We can see that the east side of London seems to be safest based on the sum of crimes committed.

To explore the data it was beneficial to use K-means clustering to see any clusters of boroughs in relations to the amount and names of the sushi places.

Using a one-hot encoding and the K clusters of 8, dataframe was adjusted to add cluster label for further use.

	borough_name	latitude	longitude	crime_sum	no of restaurants	avg restaurant dist	Cluster Labels	Atomic Sushi
0	Kingston upon Thames	51.4085	0.3064	23228.0	2.0	4965.000000	5	0.000000
1	Southwark	51.5035	0.0804	23905.0	12.0	6092.416667	1	0.000000
2	Tower Hamlets	51.5099	0.0059	25553.0	46.0	4752.347826	1	0.021739
3	Newham	51.5077	0.0469	26707.0	20.0	5022.800000	1	0.050000
4	Harrow	51.5898	0.3346	31514.0	1.0	3706.000000	4	0.000000

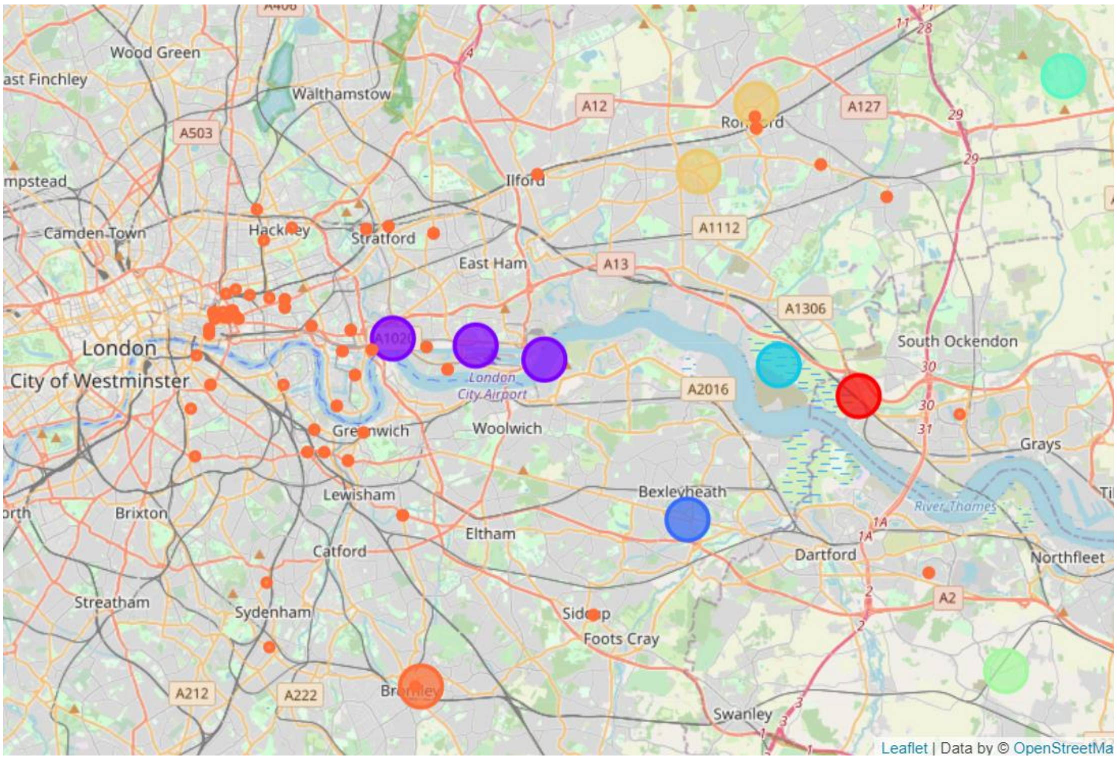
5 rows × 60 columns

Discussion

To explore the data further, the clustered boroughs were placed on a folium map. Locations of each Sushi places were also added for convenience.

We can see that there is a clear winner in terms of clusters for our target user. The Purple boroughs have plenty of Sushi places in close distance and are safe.

A good alternative would be the yellow cluster as well, as it has couple of Sushi places in close distance to each borough.



Recommendation

	borough_name	latitude	longitude	crime_sum	no of restaurants	avg restaurant dist	Cluster Labels	Atomic Sushi	Dumo Sushi	Gourmet Sushi	...	Sushinoen	Takeshi sushi	Thames Barrier Sushi
1	Southwark	51.5035	0.0804	23905.0	12.0	6092.416667	1	No	No	No	...	No	No	Yes
2	Tower Hamlets	51.5099	0.0059	25553.0	46.0	4752.347826	1	Yes	Yes	Yes	...	Yes	Yes	Yes
3	Newham	51.5077	0.0469	26707.0	20.0	5022.800000	1	Yes	No	No	...	No	Yes	Yes

Based on our analysis it is clear that the Purple cluster provides the best boroughs for our target user.

To support their choice the dataframe with information of the geographical location, crime, number of restaurants, average distance and specification of what Sushi places are available.

Conclusion

The analysis focused on finding the best borough in London for a user who values safety and wants to enjoy their favorite food, Sushi.

In this analysis a Wikipedia data, London 's Crime data and FourSquare API was used to try and recommend the best boroughs for our target user.

Many methods of data manipulation, and even machine learning was used (K-means clustering) to achieve the recommendation.

There were three boroughs recommended, based on safety, distance of Sushi restaurants, and amount of those restaurants; South Wark, Tower Hamlets, Newham.

I believe this to be a good recommendation based on the limited input of priorities for the target user. This could be expanded to using far more data for the recommendation, such as; rent prices, type of crime, other amenities etc.