Leveraging economic and foursquare data to find least competitive and most promising location for opening a retail place (bar/coffee/restaurant)

Capstone project - IBM Data Science Specialization

by Misha Obolonskyi

25 Jan 2020

Index

1. Description of the problem, its background and potential stakeholders

Today many investors and cafe owners seek different opportunities to increase revenues. In retail location might be a single most important factor that ultimately decides the success of the business. With this exercice I would like to address the problem of finding the best location for potential place for a coffee bar in London.

2. Description of the data and how it will be used to solve the problem

Data that I would use:

- Foursquare data on different locations, their success, number of reviews, visitors etc
- Public data on London about borough economics, demographics etc (https://data.london.gov.uk/)
- Miscellaneous data sources to complement core data sets

3. Methodology

Methodology consists of loading needed libraries and loading and preparing data for analysis.

3.1 Loading libraries

Libraries in use:

- numpy
- pandas
- json
- geopy
- requests
- matplotlib
- k-means
- beautifulsoup4
- re
- rdflib

- BytesIO
- ZipFile
- Graph
- Folium

3.2 Loading and preparing data for analysis

In order to get socio-economic data about London as well as names of areas, I will work with LSOA level data. LSOA stands for Lower Layer Super Output Areas.

For London I am getting data from London data bank - https://data.london.gov.uk/download/lsoa-atlas/0193f884-2ccd-49c2-968e-28aa3b1c480d/lsoa-data.csv

LSOA Atlas that plots all the relevant information is here - https://londondatastore-upload.s3.amazonaws.com/instant-atlas/lsoa-atlas/atlas.html

Based on head and shape we see that file contains vast amoung of information for each LSOA. For further analysis, I'll focus only on LSOAs in central London and key socio-economic data

Data cleaning and transformation

As regards LSOA, I'll narrow analysis to areas in Central London that organized in boroughs:

- City of London
- Westminster
- Camden
- Islington
- Tower Hamlets
- Hackney
- Kensington and Chelsea

Data I'll keep for further analysis:

- Population Density;Persons per hectare;2013

- Households;All households;2011
- House Prices;Median Price (£);2009
- House Prices;Median Price (£);2014
- House Prices;Sales;2009
- House Prices;Sales;2014
- Economic Activity;Economically active: Total;2011
- Economic Activity;Unemployment Rate;2011
- Household Income, 2011/12;Mean Annual Household Income estimate (£)

Example:

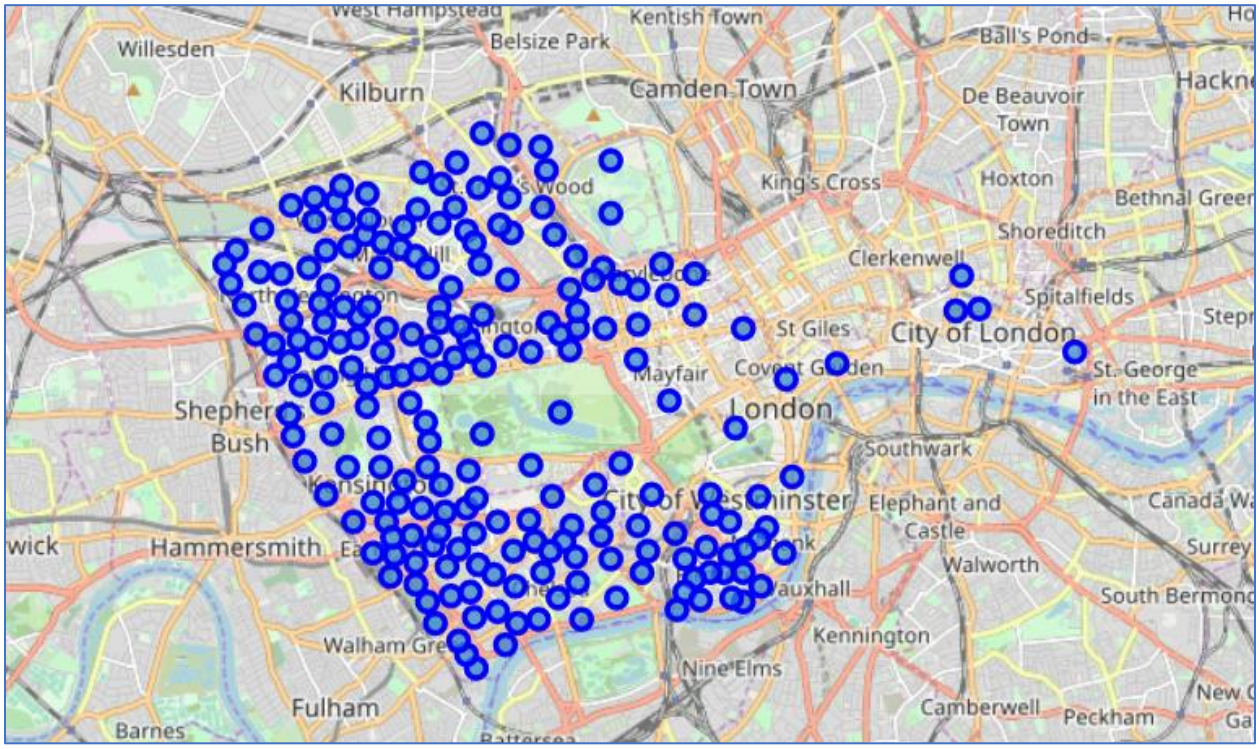
	Lower Super Output Area	Names	Mid-year Population Estimates;All Ages;2001	Mid-year Population Estimates;All Ages;2002	Mid-year Population Estimates;All Ages;2003	Mid-year Population Estimates;All Ages;2004	Mid-year Population Estimates;All Ages;2005	Mid-year Population Estimates;All Ages;2006	Mid-year Population Estimates;All Ages;2007
0	E01000907	Camden 001A	1519.0	1538.0	1528.0	1547.0	1593.0	1492.0	1455.0
1	E01000908	Camden 001B	1547.0	1577.0	1605.0	1637.0	1629.0	1598.0	1605.0
2	E01000909	Camden 001C	1600.0	1598.0	1618.0	1634.0	1610.0	1635.0	1615.0
3	E01000912	Camden 001D	1738.0	1728.0	1678.0	1707.0	1710.0	1669.0	1645.0

Once all datatypes are in the appropriate format, I can calculate CAGRs for House median Prices and House Sales.

Enriching LSOA London data with coordinates

https://opendatacommunities.org/downloads/graph?uri=http://opendatacommunities.org/graph/lower-layer-super-output-areas

Centroids of each LSOA mapped:



Based on the scrapping of all venues we see that number of places is quite significant. For the purpose of analysis we need to narrow the list of eating places such as Cafe, Restaurant, Bar, Pub etc. List of venue categories - https://developer.foursquare.com/docs/resources/categories

Data from Foursquare:

	LSOA name	LSOA Latitude	LSOA Longitude	Venue ID	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	City of London 001A	51.51801	-0.09677	4fc31eede4b05b8503be268b	Virgin Active	51.517952	-0.097651	Gym / Fitness Center
1	City of London 001A	51.51801	-0.09677	4ad3be62f964a52012e620e3	Postman's Park	51.516860	-0.097643	Park
2	City of London 001A	51.51801	-0.09677	4ac518d2f964a5203ca720e3	Museum of London	51.518019	-0.096060	History Museum
3	City of London 001A	51.51801	-0.09677	4ad4f4c0f964a520e70021e3	St Bartholomew the Great (St Bartholomew-the-G...	51.518631	-0.099890	Church
4	City of London 001A	51.51801	-0.09677	4c1cd32bb306c928426b64b7	Barbican Art Gallery	51.519800	-0.093969	Art Gallery

	Venue ID	Venue name	Tip Count	Rating	Price tier
0	4ad7a8ddf964a520650d21e3	Dose Espresso	69	8.4	2
1	5384e5ed498e11317d174b6d	Ask For Janice	65	8.4	2
2	4ada5cf0f964a520e32121e3	The Old Red Cow	59	8.2	2
3	4dff32bae4cdf7246077a9aa	Pilpel	36	9.1	2
4	4ac518d6f964a5201ea820e3	Club Gascon	21	8	3

Considering recent changes for Foursquare API, it is possible to get only number of tips for venue. Although number of users and checkins would show better picture about trendiness of the place, I believe that number of tips is also a good proxy of that.

Aggregated data before modeling:

	LSOA	Venue type	Tip Count	Rating	Price tier	Population Density;Persons per hectare;2013	Households;All households;2011	House Prices;Median Price (£);2009	House Prices;Median Price (£);2014	House Prices;Sales;2009	Hous Prices;Sales;201
0	City of London 001A	Bar	68.0	9.10	3.0	114.0	876.0	480000.0	720000	41.0	67.0
1	City of London 001A	Beer Bar	59.0	8.20	2.0	114.0	876.0	480000.0	720000	41.0	67.0
2	City of London 001A	Coffee Shop	35.0	8.15	1.5	114.0	876.0	480000.0	720000	41.0	67.0

4. Results

We have done modelling part and clustered all venue types within particular LSOA by appropriate cluster. Here are the results

	Cluster Labels	LSOA	Venue type	Tip Count	Rating	Price tier	Population Density;Persons per hectare;2013	Households;All households;2011	House Prices;Median Price (£);2009	House Prices;Median Price (£);2014	House Prices;Sales;2009	Pr
219	1	Kensington and Chelsea 006A	Bakery	81	8	1	110	703	790000	1175000	16	17
220	1	Kensington and Chelsea 006A	Bar	3	8	2	110	703	790000	1175000	16	17
221	1	Kensington and Chelsea 006A	Breakfast Spot	30	8	1	110	703	790000	1175000	16	17
222	1	Kensington and Chelsea 006A	Burger Joint	59	8	2	110	703	790000	1175000	16	17

5. Discussion of the results

As a result of clustering analysis, we see that arranged clusters of neighborhoods by places (and their relative popularity) as well as welfare in that neighborhood. Even though clustering was appropriate, algorithm could be significantly improved if additional data points about particular venue are added to the dataset.

6. Conclusions

Based on the analysis there are many most promising locations. City of London and several smaller neighborhoods in Kensington are among the top places based mainly on the welfare and population.