IBM Data Specialization Capstone Project

Battle of London Boroughs:

New Fitness Centre

Vishal Gupta

7/30/2020

Introduction

This report is for the final course of IBM Data Science specialization hosted on Coursera platform. In this project we need to leverage Foursquare location data for a city of our choice to explore and compare neighbourhoods. We need to come up with a business problem that we would like to solve using Foursquare location data along with any other data, perform data cleansing, exploratory data analysis and opt for a machine learning algorithm that we see will best solve the problem.

Background

The 2019 State of the UK Fitness Industry Report[1] reveals that the UK health and fitness industry is healthier than it has ever been with the number of fitness facilities in the UK up from 7,038 to 7,239 this year. It has more fitness options, more members and a greater market value than ever before. In Europe only Germany has more health and fitness club members than the UK. Several key milestones[2] have been achieved over the last 12 months, the total UK membership grew by 4.7% and has broken the 10 million mark; 1 in every 7 people being a member of some fitness centre and the industry is now worth more than £5.1 billion for the first time. The elusive 15% penetration rate has been exceeded with it now standing at 15.6%.

Along with the industry the customer appetite is also growing both vertically and horizontally. Customers are looking for fitness options clubbed with relaxation and enjoyment. Fitness is no longer restricted to committing few hours daily in gym, it is now becoming a lifestyle change. And to cater to this change in trend the fitness service providers are focussing on other options apart from low/high cost gyms like community driven fitness, dance studios, kickboxing workout, cross training, meditation and all-in-one fitness etc.

In such lucrative market opening a new fitness centre requires serious consideration and is complicated process. Particularly the location is the one of most important decision that will determine whether business will be successful.

Business Problem

If a service provider wants to open a new fitness centre in London, largest city of the UK; then which borough and what category of fitness centre would we recommend. There are a lot of criterion that should be satisfied in order to achieve high revenue, like:

- Population density- residential or working
- Average age and income
- Density of other fitness service providers in different categories- recreational, active sports and wellbeing.
- …. Etc.

The objective of this project is to do basic data analysis and try to understand what current composition of categories of fitness centre; *recreational* like skating rink, dance studio; *active sports* like gym, cricket ground, cycle studio; or *wellbeing* like yoga, indoor play area, meditation etc in London boroughs. Using data science methodology and machine learning techniques like clustering, this project aims to provide shortlisting of boroughs and fitness centre category. A further analysis can be done with additional factors like property rent, target customers, available properties with required facilities like parking; this will help in drilling down to exact location in the shortlisted borough however will not be performed within the scope of this project.

Data description

Considering the business problem in hand we can list the data as below, links shared in Reference section:

1. List of London borough data[3]- This defines the scope of this project, which is confined to London, the capital city of United Kingdom.
2. London borough demographics[3]- This will help in understanding the boroughs better by analysing factors like- total population, population density, average age, employment rate, average household income, number of businesses, employed population, migrant population etc.
3. London Borough coordinates- we used Nominatim function from geocoders library on Python to get central coordinates of each borough.
4. Foursquare API data[4]- To explore venue data for each borough, especially venues related to fitness industry like- gyms, dance studios, cricket stadium, mini golf course, Pilates studio, yoga centre etc.

Methodology

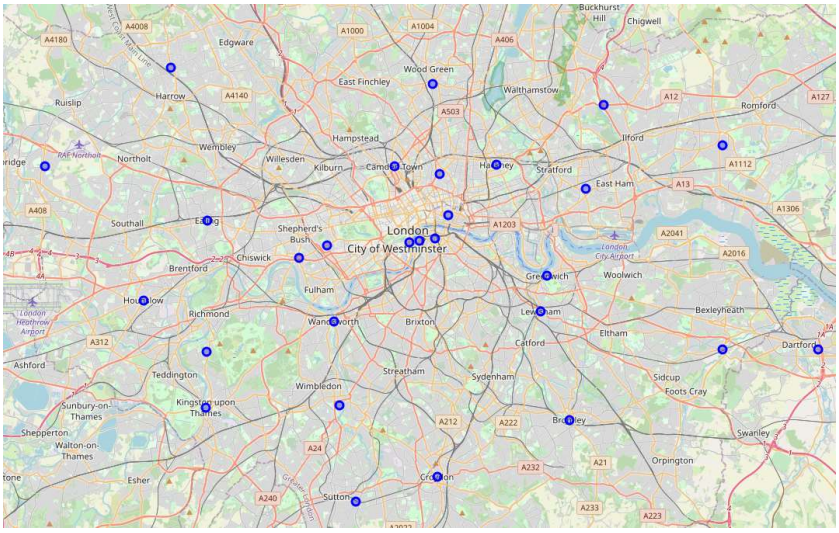
Data Processing

- Firstly, we used public data source [1] for list London boroughs and the demographics like total population, average age, median household income etc. We processed this data – replaced missing values with average of rest of the values in respective columns e.g. '% population born abroad'; removed all irrelevant columns e.g. 'New Homes', 'Rented from private landlord'; renamed some columns; to make variables comparable we converted percentages into absolute values and then normalised the data to bring values between 0 and 1 without distorting the difference in the range of the values in the respective column. This gave us final demographic DataFrame:

Code	Borough	Total_Population	Number_Of_Households	Population_Density(per hectare)	Average_Age	Population_Aged(0-15)	Population_WorkingAge	Population_Aged(65+)	Migrant_Population	BAME_Population
E09000001	City of London	0.022587	0.033494	0.194728	1.000000	0.011798	0.025441	0.023770	0.016796	0.009649
E09000002	Barking and Dagenham	0.536448	0.491720	0.371945	0.761574	0.668564	0.521569	0.353296	0.425866	0.413239
E09000003	Barnet	1.000000	0.952285	0.288597	0.863426	0.966783	1.000000	0.950534	0.739259	0.601304
E09000004	Bexley	0.627053	0.614652	0.259133	0.902778	0.591859	0.607730	0.706728	0.212024	0.208754
E09000005	Brent	0.852413	0.761262	0.493617	0.824074	0.816287	0.890502	0.653985	0.964924	0.859791

- Next we used Nominatim function from geocoders library is used to extract central longitude and latitude for each borough, merged that information in DataFrame and visualized it using folium library:

ıe_Age	Population_Aged(0-15)	Population_WorkingAge	Population_Aged(65+)	Migrant_Population	BAME_Population	Employed_Population	Median_Household_Income	Number_Active_Businesses	Latitude	Longitude
)00000	0.011798	0.025441	0.023770	0.016796	0.009649	0.019507	1.000000	0.471788	51.515618	-0.091998
761574	0.668564	0.521569	0.353296	0.425866	0.413239	0.471901	0.462433	0.118444	51.554117	0.150504
863426	0.966783	1.000000	0.950534	0.739259	0.601304	0.915775	0.637064	0.472872	51.653090	-0.200226
)02778	0.591859	0.607730	0.706728	0.212024	0.208754	0.629568	0.581421	0.163853	51.441679	0.150488
324074	0.816287	0.890502	0.653985	0.964924	0.859791	0.792014	0.505187	0.284283	51.441635	0.234519



- Used Foursquare API to explore boroughs and fetch venues within 2.5Km radius of central coordinate of each borough with the limit of 500 venues, a total 2611 venues were returned. Before analysis we categorised all venues in high level 'Parent_Category'- Food, Fashion, Bar, Leisure, Supermarket, Active Sports, Recreational Sports, Wellbeing and Others. Since we were interested only in Fitness related venues, we exported rows with 'Parent_Category' as 'Active Sports', 'Recreational Sports' and 'Wellbeing' to a new DataFrame.

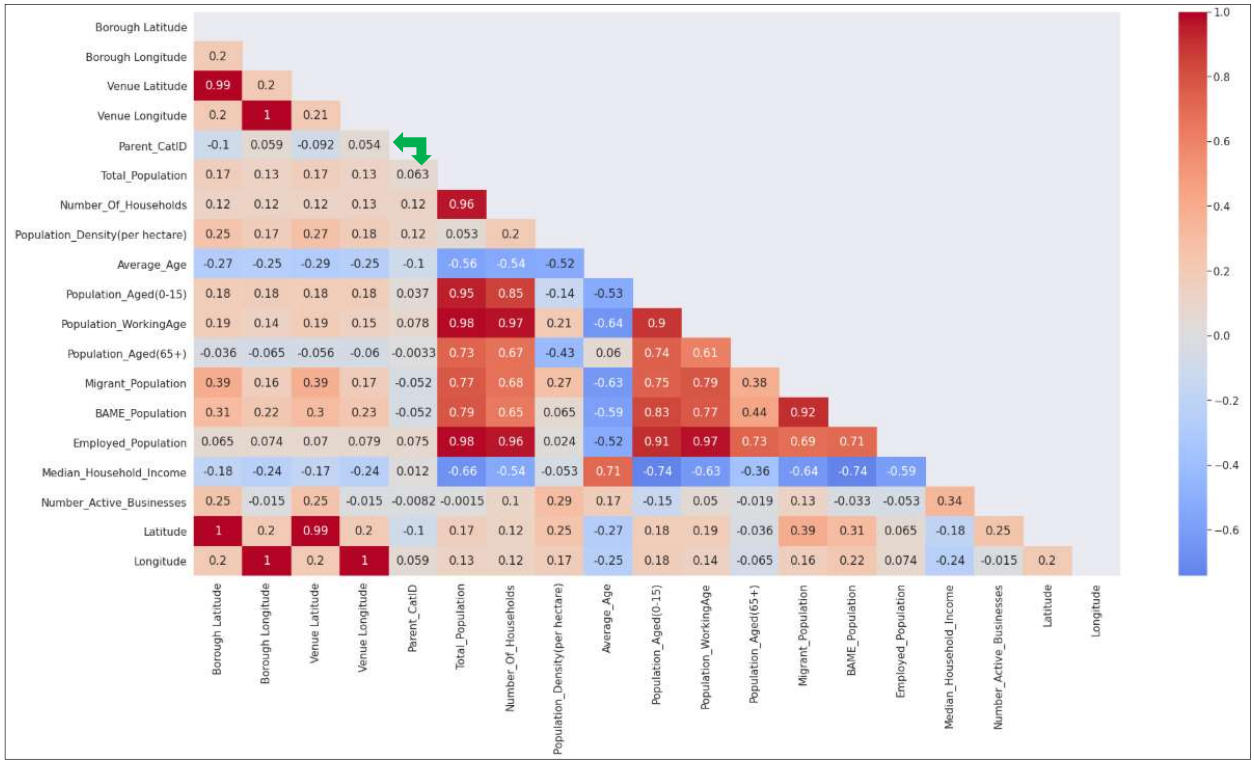
	Borough	Borough Latitude	Borough Longitude	Venue_id	Venue	Venue Latitude	Venue Longitude	Venue Category	Parent_Category
0	City of London	51.515618	-0.091998	4fc31eede4b05b8503be268b	Virgin Active	51.517952	-0.097651	Gym / Fitness Center	Active Sports
1	City of London	51.515618	-0.091998	4bf7edc04a67c928825c24cf	Barbican Conservatory	51.519859	-0.093202	Botanical Garden	Wellbeing
2	City of London	51.515618	-0.091998	55e5df82498e9f0b8a9b9606	1Rebel	51.518378	-0.083861	Boxing Gym	Active Sports
3	City of London	51.515618	-0.091998	51797f6be4b06c63fd263c8c	Cyclebeat	51.511686	-0.086461	Gym / Fitness Center	Active Sports
4	City of London	51.515618	-0.091998	53749f5c498e46fef6b4c193	1Rebel	51.515569	-0.080040	Gym / Fitness Center	Active Sports

- Merged demographics and venue data into a single DataFrame:

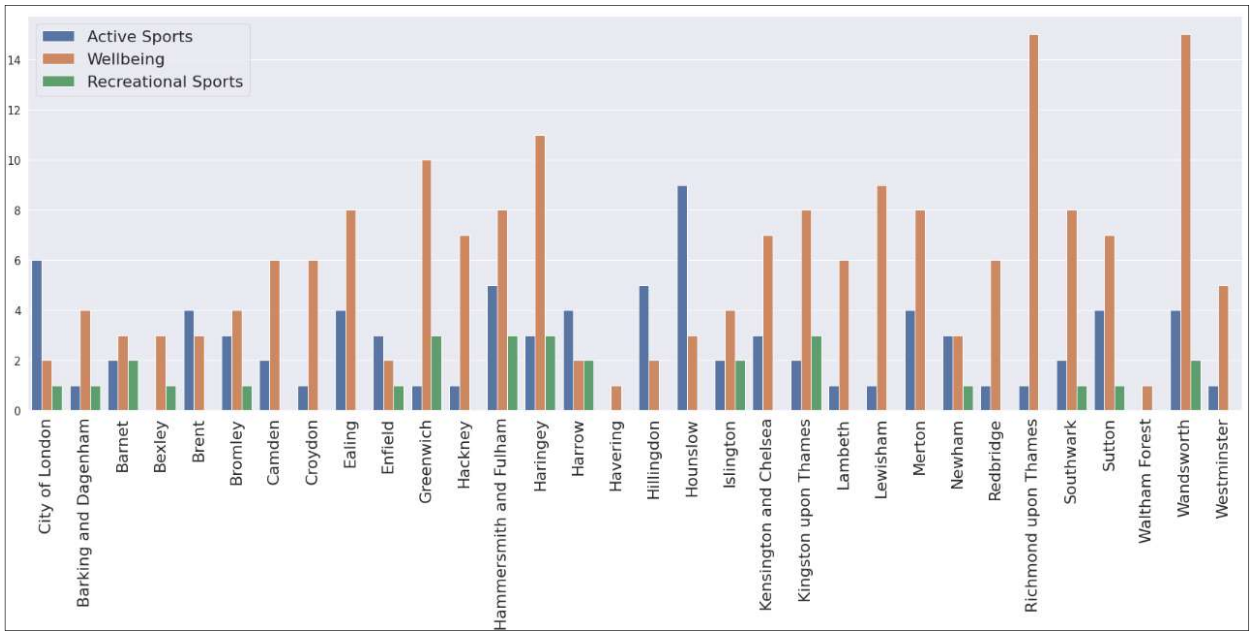
Borough	Borough Latitude	Borough Longitude	Venue_id	Venue	Venue Latitude	Venue Longitude	Venue Category	Parent_Category	Parent_CatID	Code	Total_Population	Number_Of_Households	P
City of London	51.515618	-0.091998	4fc31eede4b05b8503be268b	Virgin Active	51.517952	-0.097651	Gym / Fitness Center	Active Sports	1	E09000001	0.022587	0.033494	
City of London	51.515618	-0.091998	4bf7edc04a67c928825c24cf	Barbican Conservatory	51.519859	-0.093202	Botanical Garden	Wellbeing	3	E09000001	0.022587	0.033494	
City of London	51.515618	-0.091998	55e5df82498e9f0b8a9b9606	1Rebel	51.518378	-0.083861	Boxing Gym	Active Sports	1	E09000001	0.022587	0.033494	
City of London	51.515618	-0.091998	51797f6be4b06c63fd263c8c	Cyclebeat	51.511686	-0.086461	Gym / Fitness Center	Active Sports	1	E09000001	0.022587	0.033494	

Exploratory Data Analysis

- Explored if there is any correlation between different variables. There are some strong positive/negative correlation between certain variables like 'Number of households' and 'Total Population' OR 'BAME population' and 'Median Household income'. However, couldn't see any strong correlation between 'Parent Category' of the venue and all other variable. So, based on given data it seems no single variable is directly influencing the category of the venue (⮌).

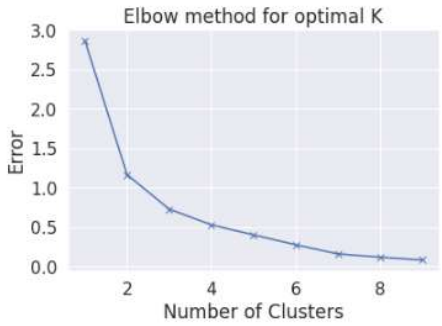


Below figure shows the number of total fitness venues per borough. As we can see some boroughs have high concentration of venues like 'Wandsworth' as compared to 'Bexley'



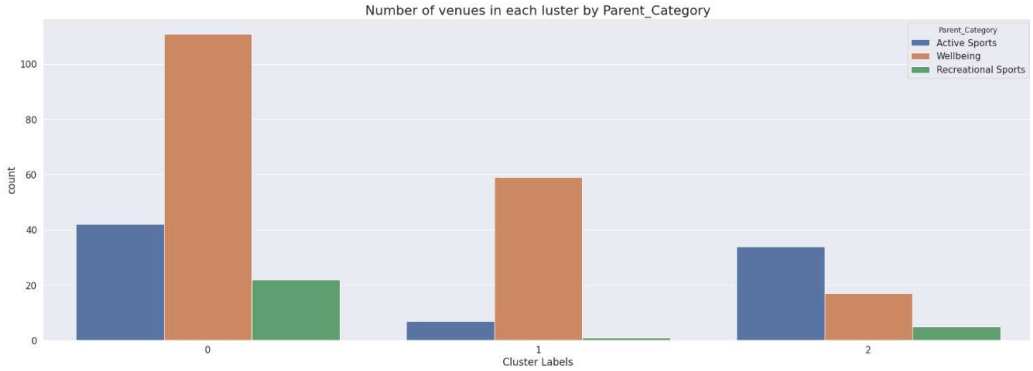
Cluster Analysis

- Before performing cluster analysis, we used elbow method to identify optimum number of clusters.

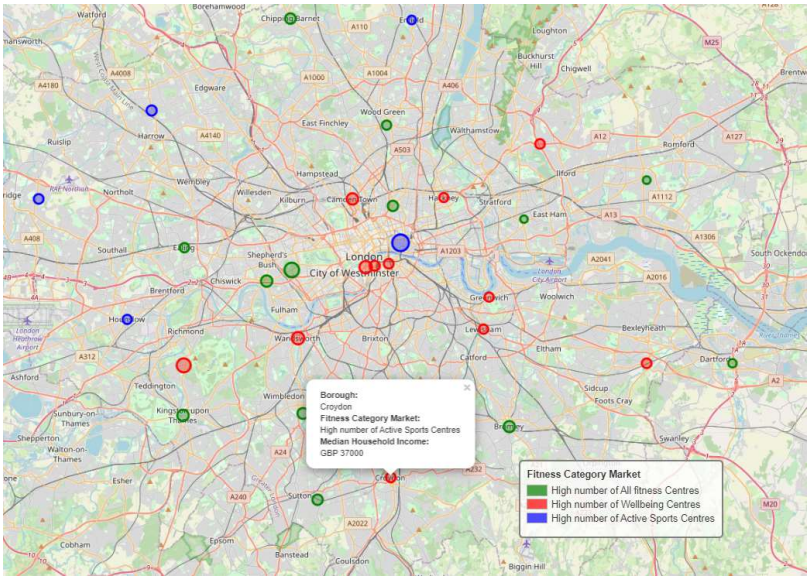


In the plot above elbow is at K = 3 indicating the optimal clusters for this dataset is 3

- Lastly, we used K-means algorithm assigned clusters to each borough. The results from k-means clustering shows that we can categorize boroughs into 3 clusters based on frequency of occurrence of each fitness centre category:
  - Cluster 0- Boroughs with 'High number of all fitness Centres'
  - Cluster 1- Boroughs with 'High number of Wellbeing Centres'
  - Cluster 2- Boroughs with 'High number of Active Sports Centres'



The results of the clustering are visualized in map below, the size of the bubble is based on 'Median household income':



Results & Discussion

Based on initial data analysis we were able to segregate London boroughs into three clusters with clear indication of current market on fitness centres. This will help fitness service provider to focus on specific boroughs based on fitness service they are offering.

- ✓ 'Cluster 0' clearly shows high density of all types of fitness centres hence it will be good to not prefer these boroughs or perform extensive research before finalizing any of cluster 0 boroughs:

	Borough	Active Sports	Recreational Sports	Wellbeing	Fitness Category Market	Total Population	Population Density	Median Household income
1	Barnet	2	2	3	High number of all fitness Centres	0.536448	0.371945	0.462433
2	Bexley	0	1	4	High number of all fitness Centres	1.000000	0.288597	0.637064
4	Bromley	3	1	4	High number of all fitness Centres	0.852413	0.493617	0.505187
5	Camden	2	0	6	High number of all fitness Centres	0.841632	0.140344	0.676831
8	Ealing	4	0	8	High number of all fitness Centres	0.902464	0.406764	0.566960
12	Hammersmith and Fulham	5	3	8	High number of all fitness Centres	0.475616	0.726160	0.688777
13	Haringey	3	3	11	High number of all fitness Centres	0.713552	0.603545	0.556743
15	Havering	1	0	1	High number of all fitness Centres	0.652721	0.145448	0.576391
18	Islington	2	2	4	High number of all fitness Centres	0.593429	1.000000	0.625432
19	Kensington and Chelsea	3	0	7	High number of all fitness Centres	0.408111	0.842721	0.874253
20	Kingston upon Thames	2	3	8	High number of all fitness Centres	0.450205	0.302487	0.690663
23	Merton	4	0	8	High number of all fitness Centres	0.534138	0.355412	0.659541
24	Newham	3	1	4	High number of all fitness Centres	0.880133	0.608711	0.452373
28	Sutton	4	1	7	High number of all fitness Centres	0.520021	0.296916	0.627790
29	Tower Hamlets	1	0	1	High number of all fitness Centres	0.780287	0.987535	0.549041

- ✓ Although at a very high level but we also observed certain opportunities that are worth exploring further. For e.g. when we compare Cluster 1- 'High number of Wellbeing Centres' & Cluster 2 – 'High number of Active Sports Centres' boroughs; there are multiple boroughs in Cluster 2 like 'Croydon', 'Greenwich', 'Lewisham' with good potential for 'Active Sports Centre':

Cluster 1 - 'High number of Wellbeing Centres'

	Borough	Active Sports	Recreational Sports	Wellbeing	Fitness Category Market	Total Population	Population Density	Median Household income
3	Brent	3	0	3	High number of Wellbeing Centres	0.627053	0.259133	0.581421
6	City of London	6	1	2	High number of Wellbeing Centres	0.622433	0.715158	0.687677
7	Croydon	1	0	8	High number of Wellbeing Centres	0.992043	0.287110	0.581578
10	Greenwich	1	3	10	High number of Wellbeing Centres	0.718943	0.380254	0.555643
11	Hackney	1	0	7	High number of Wellbeing Centres	0.704055	0.925308	0.552342
21	Lambeth	1	0	6	High number of Wellbeing Centres	0.844199	0.788315	0.604998
22	Lewisham	1	0	9	High number of Wellbeing Centres	0.778747	0.554667	0.564288
25	Redbridge	1	0	6	High number of Wellbeing Centres	0.780801	0.346471	0.579378

Cluster 2 – 'High number of Active Sports Centres'

	Borough	Active Sports	Recreational Sports	Wellbeing	Fitness Category Market	Total Population	Population Density	Median Household income
0	Barking and Dagenham	2	0	4	High number of Active Sports Centres	0.022587	0.194728	1.000000
9	Enfield	4	2	2	High number of Active Sports Centres	0.854723	0.264725	0.520434
14	Harrow	4	2	2	High number of Active Sports Centres	0.647587	0.321277	0.611129
16	Hillingdon	5	0	2	High number of Active Sports Centres	0.772587	0.167171	0.582207
17	Hounslow	9	0	3	High number of Active Sports Centres	0.703799	0.314763	0.555329

Limitations & Future Research

In this project we considered very limited factors and unfortunately couldn't establish very strong influence of any one factor on the deciding the fitness category of centres. Future research could device a methodology to further enhance this research with more factors and more data to determine preferred borough (or even postal code) to open a specific fitness centre. We also used free Sandbox Tier account of Foursquare API that came with limitation as to number of API calls and results we could fetch. Future research could make use of paid account to bypass this limitation and get better information like price bracket, social presence etc.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data. We performed machine learning algorithm and clustered the data into 3 clusters based on their similarities. We also provided high level recommendation to relevant stakeholders i.e. fitness service providers regarding locations worth exploring further using relevant variable like income, population density, average age, commercial property price etc. This will help them in identifying high potential locations where they can capitalize the opportunities.

References

[1]- https://www.sportsthinktank.com/news/2019/05/the-2019-state-of-the-uk-fitness-industry-report

[2]- https://cilconsultants.com/base/assets/The-UK-health-and-fitness-industry-in-2019.pdf

[3]- https://data.london.gov.uk/dataset/london-borough-profiles

[4]- https://foursquare.com/developers/apps