IBM Data Specialization Capstone Project

Battle of London Boroughs:

New Fitness Centre

Vishal Gupta

7/24/2020

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 find most optimal borough in London, along with the recommendation of category 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. 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

To consider the problem 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.

The process of collecting and cleaning the data:

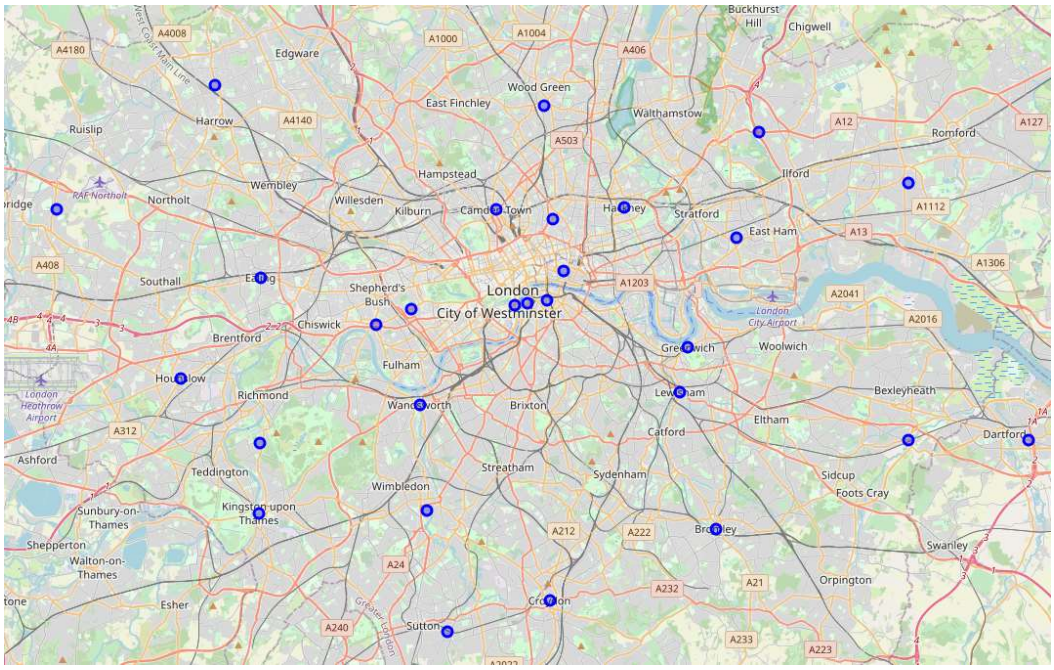
London Boroughs and London Demographic DataFrame:

- Download excel from the weblink and load data from excel into a DataFrame
- Data Cleaning-
  - Drop columns that aren't required and give relevant names to some of the columns.
  - Data conversion- convert data of columns like '% population' in absolute number for each borough for example employed, migrant population etc to bring all features on same scale.
  - Missing values- replace with average of remaining values in the column for e.g. % population BAME etc. for 'City of London' borough.
  - Normalise data- Since features like age, population, income, number of business are on different scale we normalise data to bring values between 0 and 1 without distorting the difference in the range of the values in the respective column.
- Below is the screenshot of final London Borough 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	En
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	

Longitude and Latitude of each Borough

- Nominatim function from geocoders library is used to extract central longitude and latitude for each borough and view them on map using folium library:



Venue DataFrame

- Extract venues within 2.5KM radius of central coordinate of each borough, filter data for relevant categories related to fitness. Final DataFrame of Venue

Borough	Borough Latitude	Borough Longitude	Venue_id	Venue	Venue Latitude	Venue Longitude	Venue Category	Parent_Category
City of London	51.515618	-0.091998	4fc31eede4b05b8503be268b	Virgin Active	51.517952	-0.097651	Gym / Fitness Center	Active Sports
City of London	51.515618	-0.091998	55e5df82498e9f0b8a9b9606	1Rebel	51.518378	-0.083861	Boxing Gym	Active Sports
City of London	51.515618	-0.091998	51797f6be4b06c63fd263c8c	Cyclebeat	51.511686	-0.086461	Gym / Fitness Center	Active Sports
City of London	51.515618	-0.091998	53749f5c498e46fef6b4c193	1Rebel	51.515569	-0.080040	Gym / Fitness Center	Active Sports
City of London	51.515618	-0.091998	4bc88fbf8b7c9c74915938cf	Thames Path	51.508810	-0.086496	Trail	Recreational Sports

Next Action

We will submit final report in next week with more details on the analysis- Methodology we followed, Result and the Conclusion.

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