DATA SCIENCE CAPSTONE – Week 1

1. INTRODUCTION

1.1 BACKGROUND

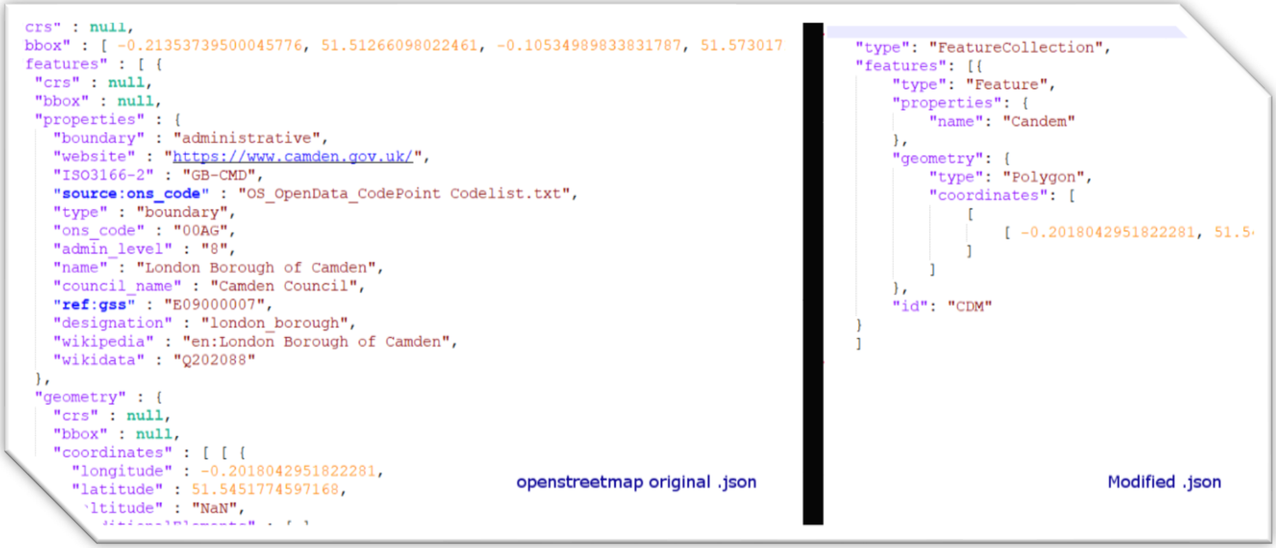
For this capstone project I decided to study a problem similar to the ones we already saw in this data science course. I decided to focus on a particular industry in the city of London with the intention of identifying the best area where to open a new venue. The field I want to analyse for this project is the fitness and wellness centers industry. Fitness centers have a market value of 2 billion pounds and employs almost 50000 people across the UK. Business analysis on this market have estimate that this industry is expected to grow due to a general increase in health awareness. Therefore it is reasonable that new investors in this sector and already established brands desire to know which area is the best for investments. I will try to explain to the stakeholders of a fitness company which are the best locations for growing their business.

1.2 BUNSINESS PROBLEM

In order to identify the best location where to open a new fitness centers in London we have to define which are the most important factors that can influence the industry under investigation. Fitness and welless centers is a type of industry that rely on people willing to dedicate part of their time, effort and money to their physical health/aesthetic. Therefore it is reasonable that demographic data such as, age, income, population density etc. might gives us some clues and help us resolve this problem. Another important data we should consider is the amount of gyms/fitness centers and their location. Moreover, for the purpose of this project, it is advantageous to find a way to subdivide the vast London expanse in smaller areas in a way that is intuitive and direct. This will facilitate our work and ultimately will help us explain the data the our shareholder

2. DATA

Completing this project means, first of all, finding data related to the problem of choice. I will use fundamentally three sources of information: fourthsquare for venues locations, openstreetmap.org for borough coordinates and the ONS of England (Office for National Statistics) for demographic data. Openstreetmap.org is a valuable opensource collection of the world maps. I will use this resource to find and download all London borough borders coordinates point by point. The .json file provided by openstreetmap.org is going to be little dense of information so I will code a small python script on my local machine to transform all these .json file to a more readable and essential version. Lastly, I will push them to my github repository so they can be easily accessed from my notebook.



```
crs" : null,
bbox" : [ -0.21353739500045776, 51.51266098022461, -0.10534989833831787, 51.573017
features" : [ {
 "crs" : null,
 "bbox" : null,
 "properties" : {
   "boundary" : "administrative",
   "website" : "https://www.camden.gov.uk/",
   "ISO3166-2" : "GB-CMD",
   "source:ons_code" : "OS_OpenData_CodePoint Codelist.txt",
   "type" : "boundary",
   "ons_code" : "00AG",
   "admin_level" : "8",
   "name" : "London Borough of Camden",
   "council_name" : "Camden Council",
   "ref:gss" : "E09000007",
   "designation" : "london_borough",
   "wikipedia" : "en:London Borough of Camden",
   "wikidata" : "Q202088"
 },
 "geometry" : {
   "crs" : null,
   "bbox" : null,
   "coordinates" : [ [ {
     "longitude" : -0.2018042951822281,
     "latitude" : 51.5451774597168,
     ltitude" : "NaN",
```

openstreetmap original .json

```
"type": "FeatureCollection",
"features": [{
    "type": "Feature",
    "properties": {
        "name": "Candem"
    },
    "geometry": {
        "type": "Polygon",
        "coordinates": [
            [
                [ -0.2018042951822281, 51.5
            ]
        ]
    },
    "id": "CDM"
}
]
```

Modified .json

The demographic data I'm going to use will be gathered from the Office for National Statistics (ONS) of England which is an accredited source. I will consider various aspects of the population and I will consider the data per borough. In this way every area will have its own characteristics and will be easy to link them to the fitness industry for comparison. Firstly, the basic information that comes to my mind is to consider the total population per borough. Higher population is reasonably correlated to a higher number of facilities. We could argue that not all people that lives in a borough might want to go to the gym closest to home which might be especially true for those individuals that live close to the borough borders. However, this particular situation should not undermine the usefulness of this data since it is most likely a phenomenon that happens equally across all boroughs hence it's a constant, invariant to our problem.

The data of the total population per borough on its own its not enough. Another interesting parameter to consider is the average age per borough. We can expect that a younger population is more prone to engage in physical activities. Which means that a lower average age indicates a well suited borough for our business. Moreover, the average age data will also integrate productively with the total population data. For example, we could have a borough with a high number of people but with a high average age which will make this borough a little less appealing for our purposes. These two data can work together in defining more correctly boroughs of interest. Another parameter that we could consider is the average income per borough. A borough with a wealthy population is indicative of a more expensive area which can be a double edge sword. Properties and rents will be more expensive meaning that the investment required is going to be higher. However, it is reasonable to think that a wealthier population can afford to pay for more services, therefore making fitness centers a better investment. The population density (population per square km) is another parameter I will consider for this project. This data can be useful in further differentiate the areas. Most of London borough have similar sizes besides some bigger ones that are further towards the outskirts. Population density gives us the relation between population and borough extension. This ratio can be synonym of venues availability and possibly lower costs.

Lastly I will consider fourthsquare data. Fourthsquare is a reliable source of location data from which we can obtain very valuable information. I will make use of this platform to obtain various data from the business venues of interest. Fourthsquare allows us to search for venue type in a given area and can gives us also other venue insights such as: unique ids, geo-coordinates, names, photos, ratings etc. Solving my problem will require the use of only some of them. First of all, latitude and longitude are fundamental for the scope of my project. Mapping the location of all the venues of interest in London will give me one of the most important information; venue density and distribution. I will leverage this data by intertwining it with all the information gathered. Unique id will be useful to differentiate between venues with identical names which will be useful when dealing with bunsiness chains (e.g. Virgin Active, PureGym, The Gym, etc.).