IBM Applied Data Science Capstone

Gilberto Agostinho

May 2020

1. Introduction

A housing development company wants to identify what is the best area in London to construct a new building for private owners. Their 3-bedroom apartments are targeted to middle-class families with children. With this in mind, they are looking for a region with ample amenities such as supermarkets, parks, restaurants, cafes, etc. They also prefer a more central area.

In this study, we will consider all London boroughs and the areas that make those up, and look for the most suitable candidates for their project.

2. Data

The data used in this project comes from several sources:

1. The list of London areas and their boroughs is obtained from a table from https://en.wikipedia.org/wiki/List_of_areas_of_London
2. The latitude and longitude of these boroughs is obtained using the `geocoder` package. This information will be used to search for venues within those boroughs.
3. Venue data is obtained using the FourSquare API.

3. Methodology

First, a list of London areas and their respective boroughs is scrapped from Wikipedia using the `BeautifulSoup` package. This list will be used as the basis of our search for an ideal area.

Once this list has been scrapped, we need to clean the data. Columns that won’t be used are dropped (such as Dial Code and OS Grid Ref). Next, we use regex (via the `re` package) to clean the list of boroughs. This is necessary because most of them have brackets with numbers that led to footnotes in the Wikipedia page. Using the regex pattern `(.*)\[\d+\]` we are able to select any groups of words that precede bracketed numbers. Next we also convert the Town names from upper case into title case for sake of consistency.

Next, we will obtain each area’s location using `geocoder`. For each area, we look for their latitude and longitude. This will allow us to plot the areas in a map (see Figure 1 in the Section 4 below).

With this information, we can now send requests to FourSquare API’s and find out the types of venues within 500m from the centre of each of these areas. Once the venue data is acquired, we can group them by area and creating a list of all venue categories available. Now that each area is associated to a list of venues, we can then inspect them to filter out any areas that do not conform to our client’s needs.

An ideal area is defined according to our client’s needs:

1. having at least a supermarket, a market, a convenience store, or a grocery store
2. having a park or a public garden
3. access to a gym
4. access to restaurants, cafés, pubs, and museums

4. Results

Using the location data for each area, we can plot their location and colour code them according to their boroughs. For better visualisation, I chose the ‘Stamen Toner’ style of map, given the high number of boroughs (and thus of different colours).

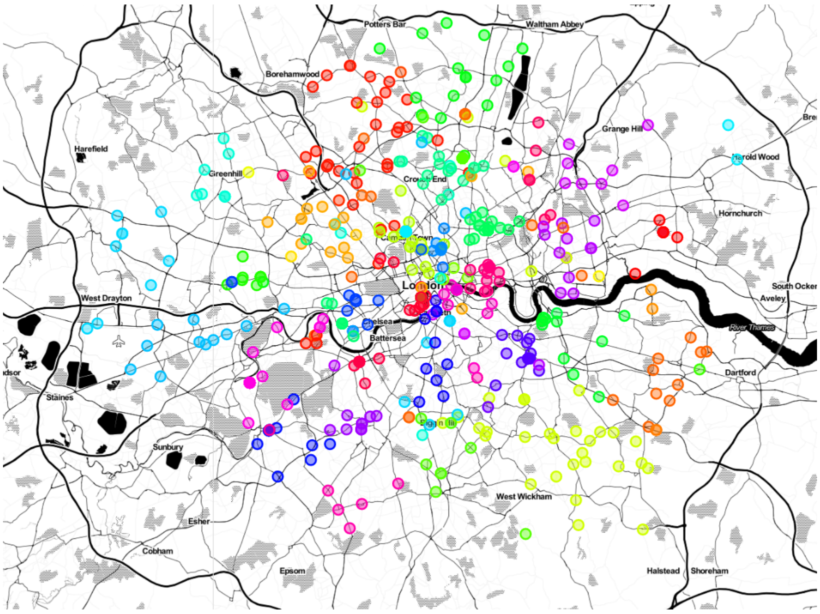
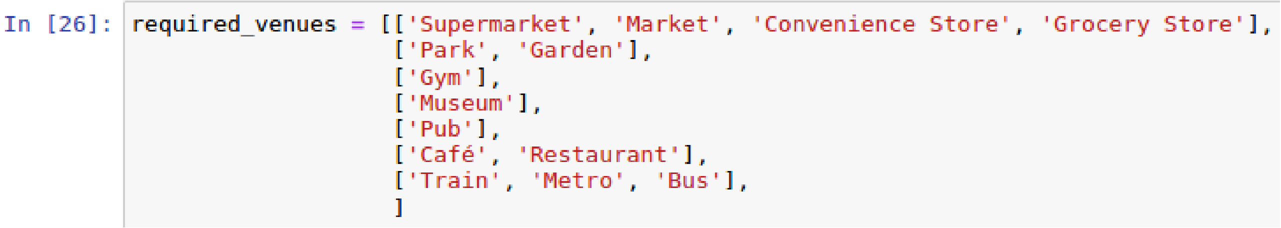


Figure 1: Areas of London, colour-coded according to the Boroughs they belong to.

We can now start filtering these areas according to their amenities categories. For this, we use a list of subslists as shown below:

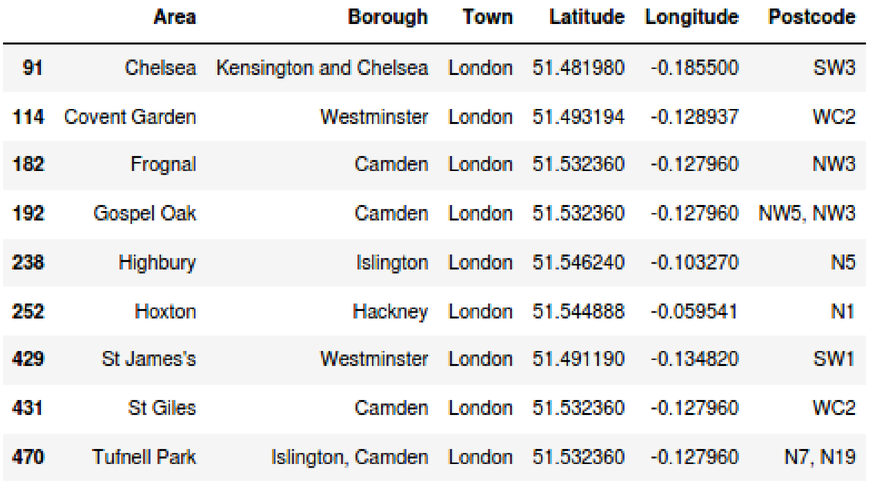


```
In [26]: required_venues = [['Supermarket', 'Market', 'Convenience Store', 'Grocery Store'],
                            ['Park', 'Garden'],
                            ['Gym'],
                            ['Museum'],
                            ['Pub'],
                            ['Café', 'Restaurant'],
                            ['Train', 'Metro', 'Bus'],
                            ]
```

Figure 2: list of required amenities.

Each sublist above contain equivalent venues. For instance, our algorithm search for areas that contain either a ‘Supermarket’, or a ‘Market’, or a ‘Convenience Store’, or a ‘Grocery Store’. It also search for keywords inside strings, so when searching for ‘Gym’ it finds both ‘Gym’ as well as ‘Gym / Fitness Center‘.

After iterating over all areas, there are nine matches containing the required venues. These are displayed in the figure below:



	Area	Borough	Town	Latitude	Longitude	Postcode
91	Chelsea	Kensington and Chelsea	London	51.481980	-0.185500	SW3
114	Covent Garden	Westminster	London	51.493194	-0.128937	WC2
182	Frognal	Camden	London	51.532360	-0.127960	NW3
192	Gospel Oak	Camden	London	51.532360	-0.127960	NW5, NW3
238	Highbury	Islington	London	51.546240	-0.103270	N5
252	Hoxton	Hackney	London	51.544888	-0.059541	N1
429	St James's	Westminster	London	51.491190	-0.134820	SW1
431	St Giles	Camden	London	51.532360	-0.127960	WC2
470	Tufnell Park	Islington, Camden	London	51.532360	-0.127960	N7, N19

Figure 3: Areas of London matching the required venues.

Let’s now plot these four matches in the map, using the same colour code as in the previous map in Figure 1.

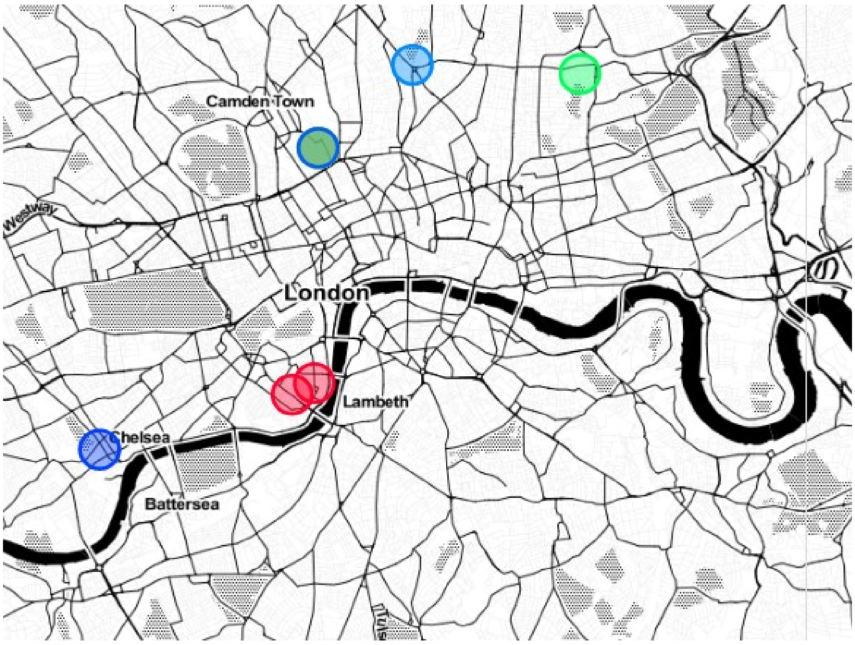


Figure 4: Locations of the areas of London matching the required venues.

Finally, we use the central location of London (found using `geopy.geocoders.Nominatim`) to calculate the distances to each of these areas. The function `haversine_distance` from the package `mpu` allows us to calculate distances in kilometres between two pairs of latitudes and longitudes.

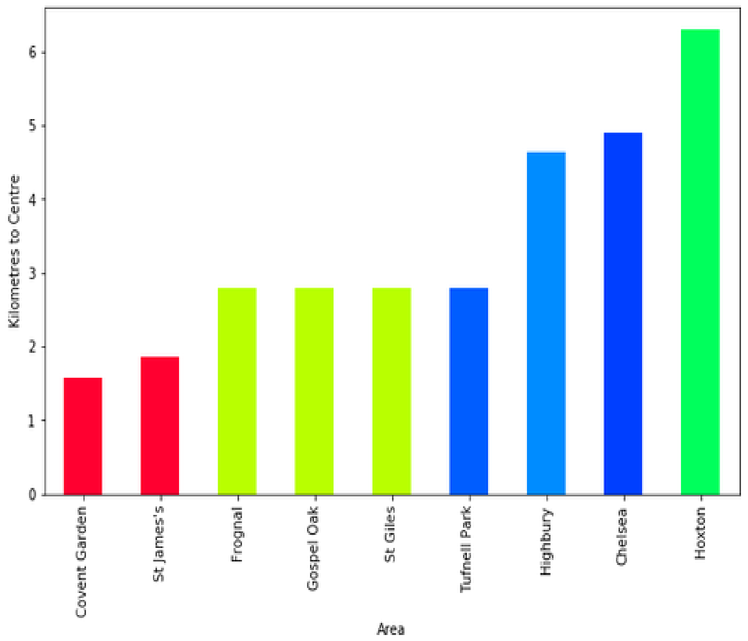


Figure 5: Areas of London, colour-coded according to the Boroughs they belong to

Figure above shows the distances to the centre for all nine selected areas.

5. Discussion

From our results above, we found nine areas in London which conform to our client's needs: Convent Garden, St James's, Frognal, Gospel Oak, St Giles, Tufnell Park, Highbury, Chelsea, and Hoxton. Using the graph above (Figure 5), we can see that either Convent Garden or St James's (both in the borough of Westminster) are areas that our client should consider since they also match the requirement of being close to the centre of London (both are within 2km of the centre), and both have all the required amenities. The three locations in the borrow of Camden (shown in light green) should also be considered given that the .

Unfortunately the FourSquare does not include primary schools in the list of venues, which would have been useful for filtering the locations further given the target customer of our client.

6. Conclusion

Given our findings above, we can strongly recommend our client to consider the areas of Convent Garden or St James's. Both are very central and have all basic amenities required within a radius of 500m.

Areas in the borough of Camden could also be good choices but are a little further away from the centre of the city than the selected two.