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