The Wine Cellar Ltd

## Project Requirements and Specifications

My client is an independent Wine Shop offering an interesting portfolio with a wide range of quality and rare products and serving both to particulars and the on trade (restaurants, pubs, hotels, etc)

They have presence in Madrid and now want to expand to London, UK, but wonder where will be the best location to make business.

For this they are contacting us so we can make a draft of our possible approach:

#### Requirements

A logical and initial requirement is be to be set up in a commercial area in which there is a reasonable amount of pedestrians walking around and/or in a location in which many people come by to do shopping, but they are not interested to do so in a mall, they want street visibility and that sense of independency and identity.

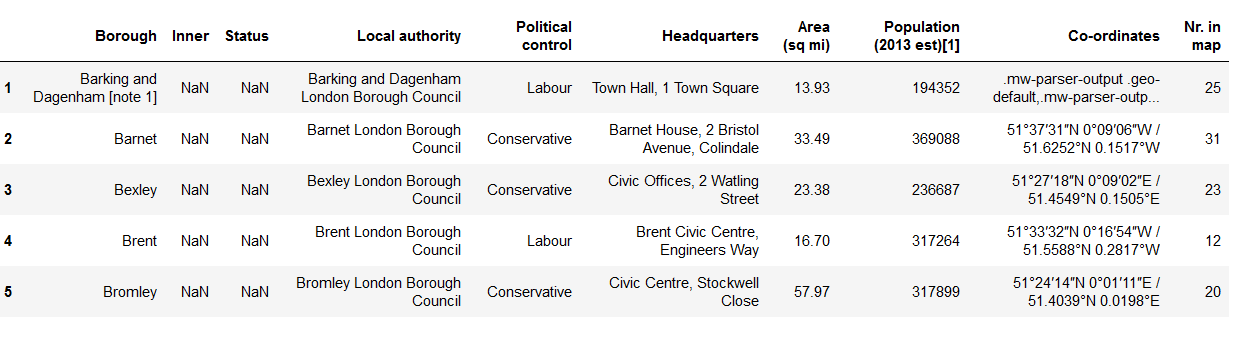
Also, given the nature of their business, there is a need to have a certain amount of possible on trade locals nearby with which they can begin building a name in the city as a hospitality supplier.

Another requirement is that the location has to have a *good to high* economic level, as their portfolio is *mid to high* range, therefore needs customers with a healthy economy to be able to purchase it with certain frequency.

**Data sources and Methodology applied**

To begin the project we have found a source with boroughs and its coordinates

<https://en.wikipedia.org/wiki/List_of_London_boroughs>



Also we have found a source with which Boroughs are considered inner and outer London:

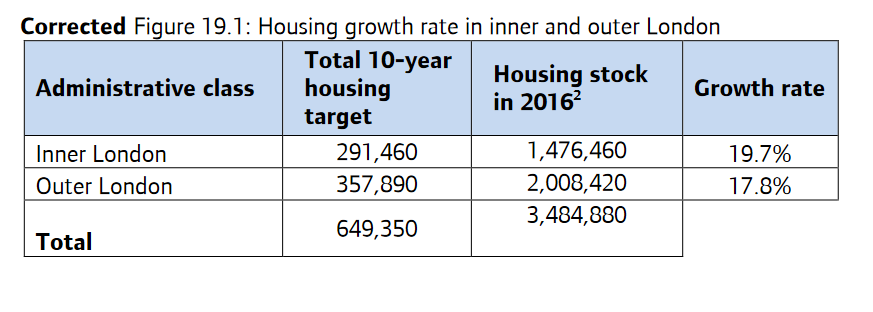
<https://latitudelongitude.org/gb/city-of-london/>

This information is a great find and will be very advantageous for being to establish which boroughs we may want to discard, as we know the client has an interest to be located centrally. This way we can work with a more targeted dataset, making inquiries and classification tasks easier.

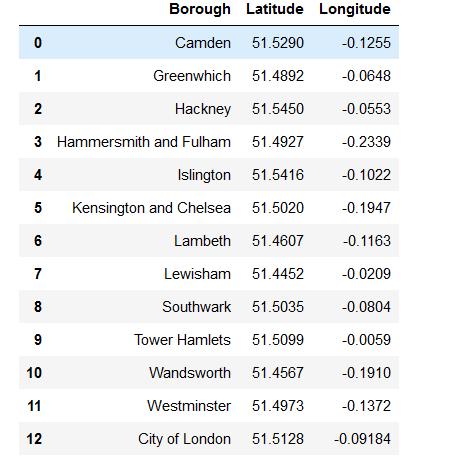
Also in this source we found something very interesting to be able to explain why the client needs to pick inner London areas:

<https://www.london.gov.uk/sites/default/files/nlp_ex_13_gla_response_to_matter_19_-_supplementary_question_-_inner_and_outer_london_boroughs_140119_final.pdf>

There we can find the following table in which we can see the growth rate percentage of 12 Boroughs in inner London greater than the growth of the 20 within the Outer area:



After a good amount of data wrangling and having used our list of inner boroughs, we are left with this reduced and clean data set:

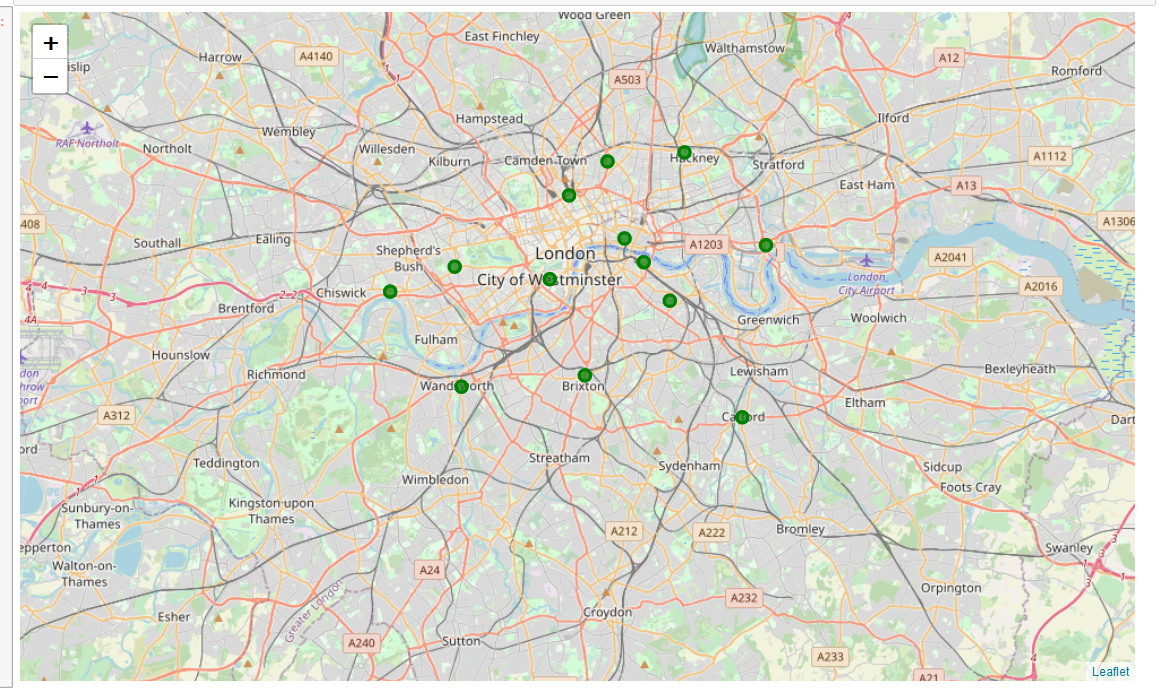


We had also to manually add City of London, for which we found the coordinates in here:

<https://latitudelongitude.org/gb/city-of-london/>

While investigating information about London, we found out that City of London is not considered a Borough, but is indeed considered within the inner area.

To bring a bit of more visual data, we superimposed our selected boroughs over a map of London using Folium:



Then we proceeded to use the Foursquare API, which takes a specific point of the map (via Latitude/Longitude) and find information of the venues around in a radius specified by the user.

This is very handy for a project like this one in which we need to know what kind of commerce is within the boroughs.

Taking this into account is important to know that the usual way of getting Lat/Lon of anywhere (street, ward, district, borough, area) must be always the approximate centre of the location, as you will search by radius. If for example you choose a Lat/Lon that is the “last house of the place”, you will be doing your search partly outside the area you want.

So is important to make sure that your dataframe complies with this requirement.

If unable to find a suitable one, **Geocoder** does that for you.

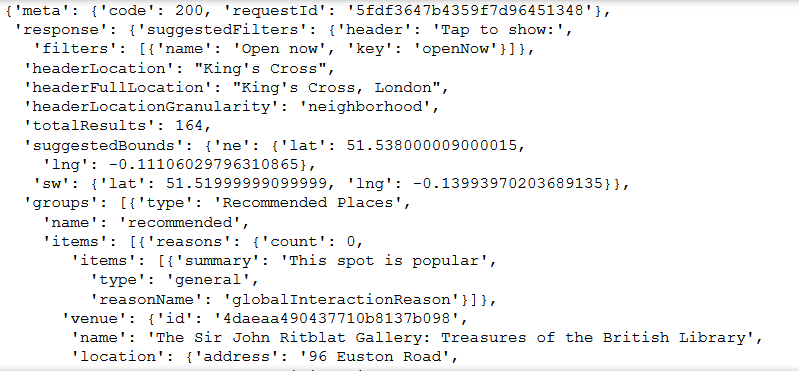
**Foursquare analysis**

First we explored one borough to warm a bit with the API.

Then we proceeded to use the full dataframe to get the venues info.

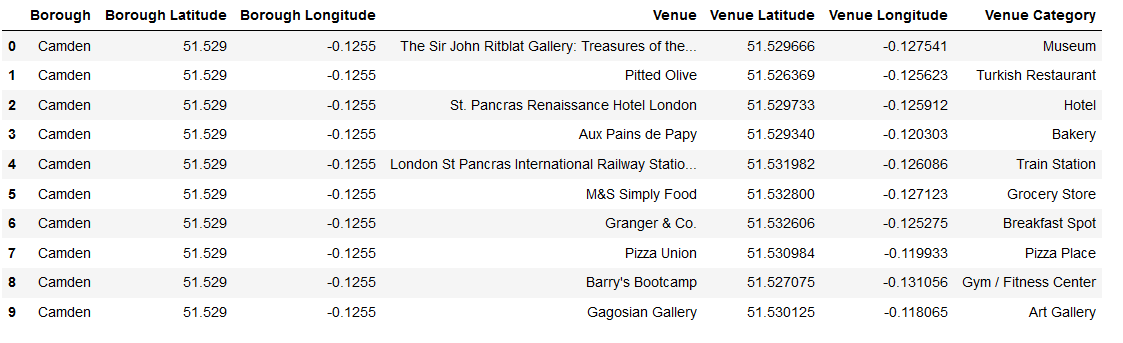
This is done via JSON, that we had to normalize and bring into a dataframe, for which you need to understand where the information within the JSON is contained

*JSON output:*

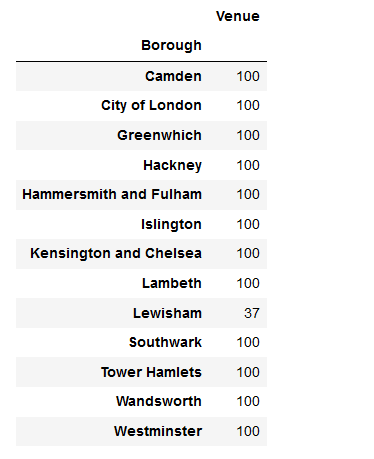


*Brought into a DataFrame:*

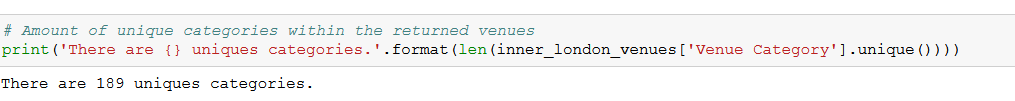
This is done by merging the data with our previous dataframe so we can have visibility of the venues by of our selected Boroughs. Let´s take a peak to the head() of the new dataframe:



The default venue retrieve limit for Foursquare API is 100, by making a count of our venues per Borough, we can see we got 100 in almost all of them (thing to have into account is the time of the day in which you do the query, as at different times will show you different venues, as are the active ones)



By making a simple line of code we know we got a total of 189 unique venue types:



**One-Hot Encoding vs Dummy coded variable**

Then we used pandas get dummies to convert categorical variable into dummy/indicator variables which returns a DataFrame (Dummy-coded data)

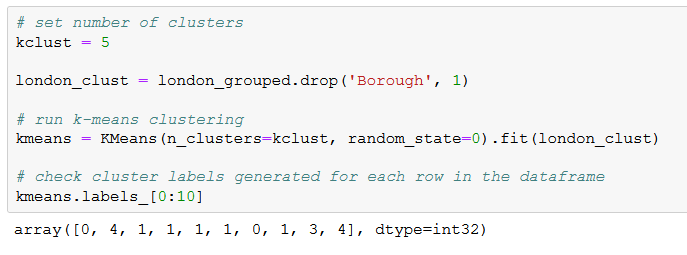
To differenciate one hot encoding of dummy coded variable here´s a bit of info:

[...] there is some redundancy in One-Hot encoding. For instance, if we know that a passenger’s flight ticket is not First Class and not Economy Class, then it must be Business Class. So we only need to use two of these three dummy-coded variables as a predictor. More generally, the number of dummy-coded variables needed is one less than the number of possible values, which is K-1. In statistics, this is called a dummy encoding variable, or dummy variable. Dummy encoding variable is a standard advice in statistics to avoid the dummy variable trap, however, in the world of machine learning, One-Hot encoding is more recommended because dummy variable trap is not really a problem when applying regularization

After using get dummies and doing a series of operations to group rows by borough and take the mean of the frequency of occurrence of each category, we printed the 10 most common venues per Borough and brought it into a DataFrame:



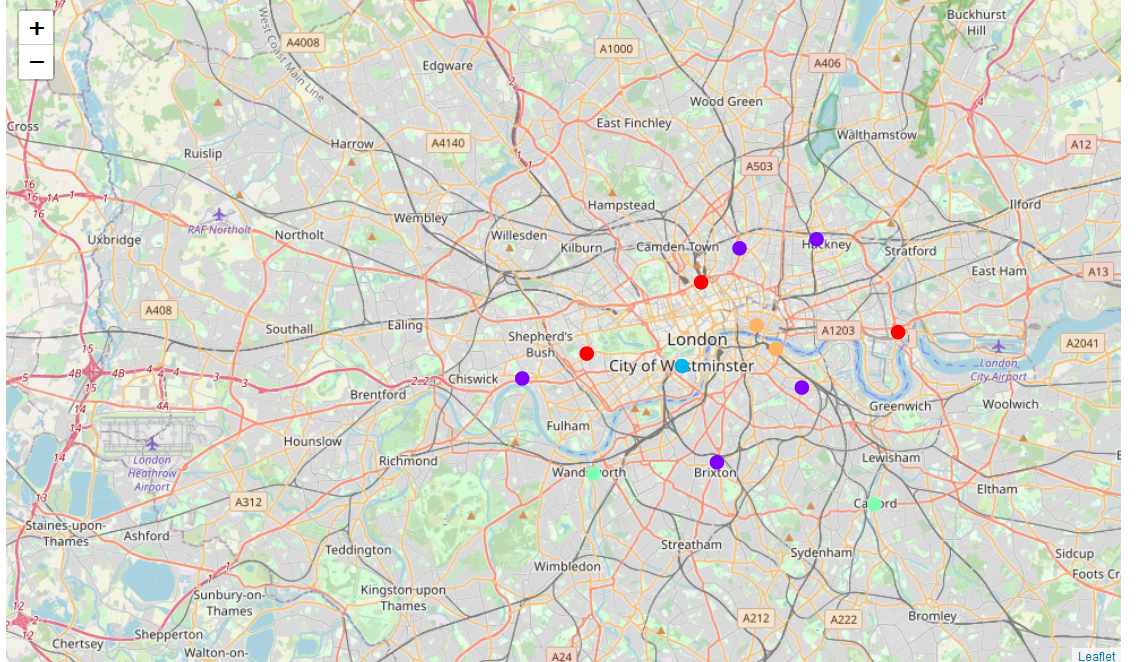
Then we clustered the venues:



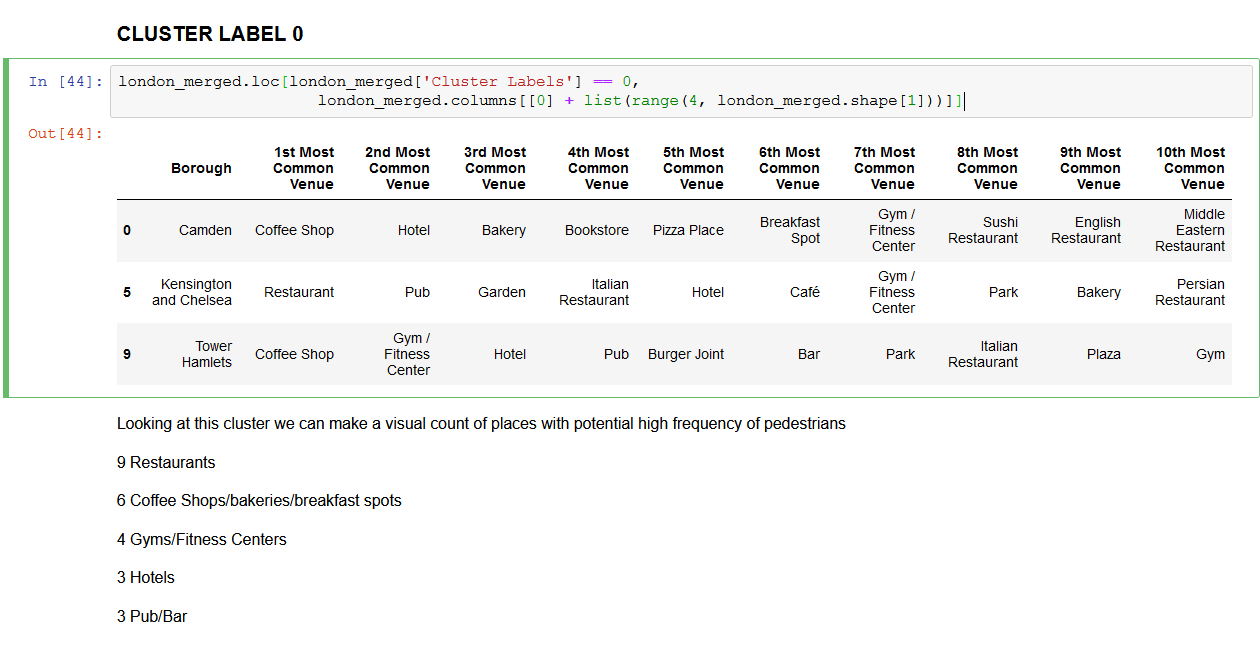
And merged it with our inner London Boroughs dataframe:



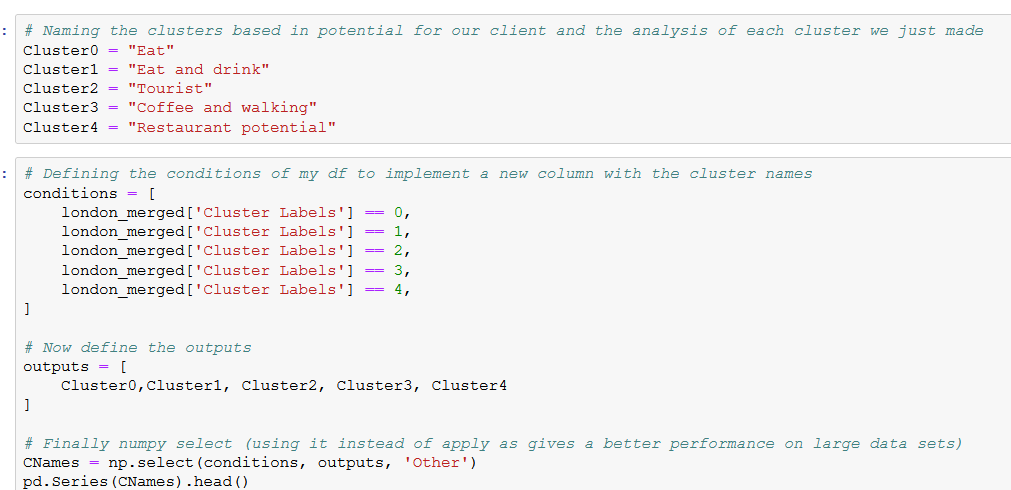
This gives us the possibility to visualize the clustering in a Folium map:

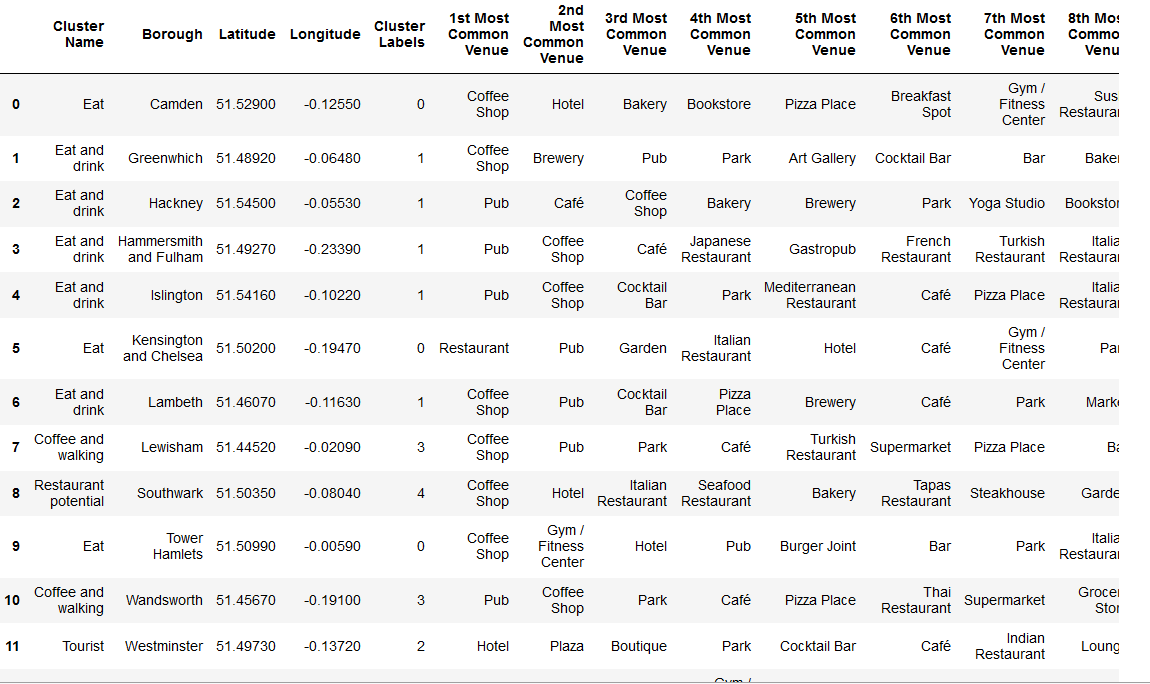


Now we can analyse each cluster on its own. Example:



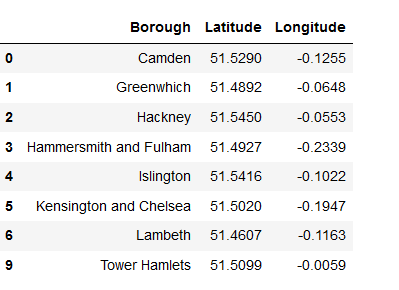
This way we can name our clusters with more meaningful names, at least for our objectives, and include it as well in our dataframe:

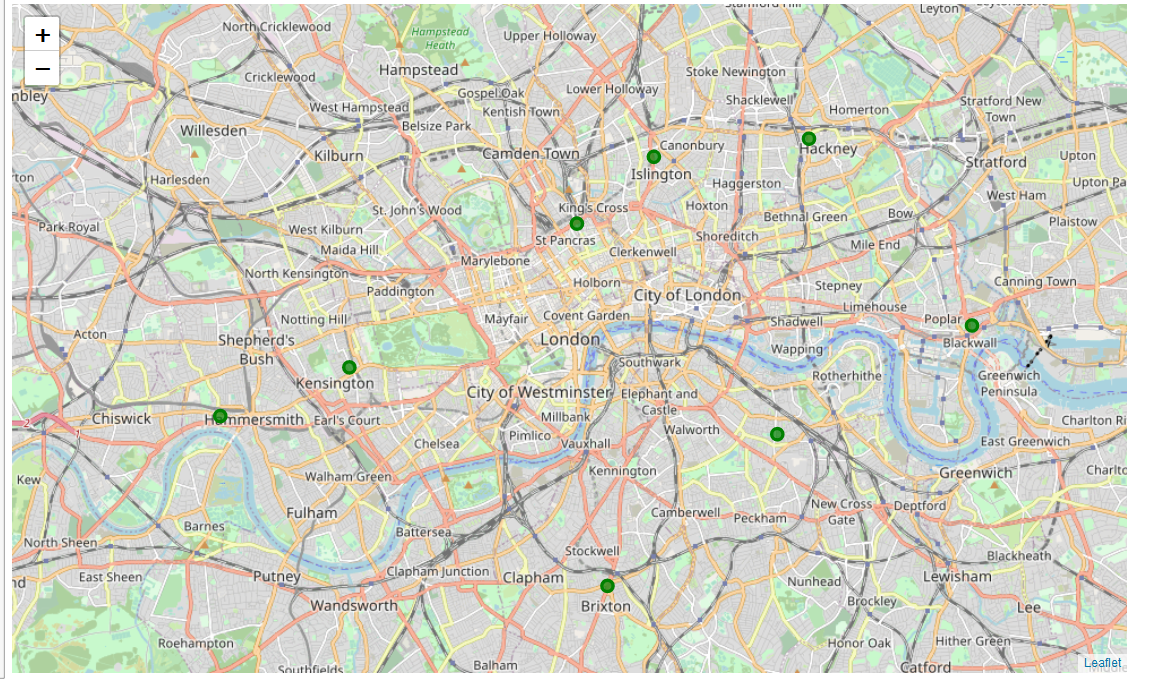




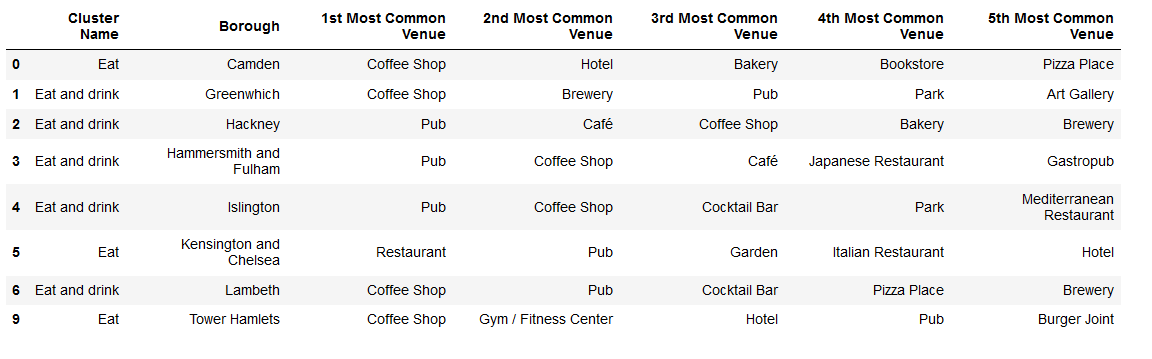
## Getting near to the conclusions

The above gave us the opportunity to reduce our dataframe again to the more suitable boroughs, so now instead of 12, we have 8:





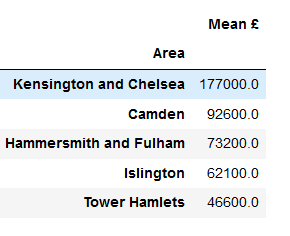
To have a clearer picture of the selected Boroughs we have reduced the venues info to the 5 most common ones:

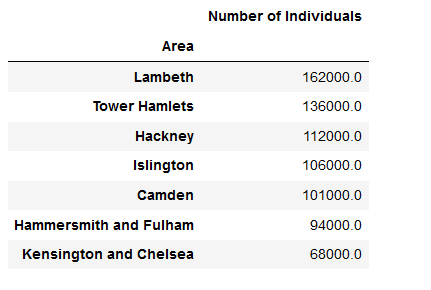


Then we also brought bit of more info to comply with the client requirements, average income and population (as in 2017-2018) for each of the selected boroughs:

Source: <https://data.london.gov.uk/dataset/average-income-tax-payers-borough>

After wrangling the data we have the below dataframes:





## Conclusion

Our client wanted Boroughs with a lot of people and commerce and also a high income, and a good number of On Trade business as well.

Our selection is already based not only in commercial areas (inner London) but also in the ones with good amount of commerce within the on trade.

Based in population we could choose the 3 first Boroughs in the last table, but since client needs good to high income the 3 chosen boroughs are finally:

* Kengsington and Chelsea
* Camden
* Hammersmith and Fulham

This is an approximation to a broad project that could be done in a much more granular level selecting a different data set, for example by diving in the internet, we found as well that the boroughs have further subdivisions like Districts and Wards.

This way we could get clusters with many more areas on them giving a much more precise picture of each borough and its venues

Also we could use street names for getting the coordinates and explore specific streets within certain Boroughs to be able to point to an exact location to recommend to our client.

So if our client is happy with what we can do and is willing to go ahead, we will dive deep into London to give them a precise answer to their problem.

We have shown as well, the leverage that provides an API like Foursquare to answer real business questions.