London Boroughs Clustering

IBM Data Science Professional Certificate

Capstone Project

Tomás Bertoglia

Munich, 25. Jun 2020

Table of contents

I. Executive Summary

This reports shows an example of how data science can solve the problem of combining different techniques in order to satisfy the often hybrid objectives of our Stakeholders.

In this case, two clustering methods will be applied over the London‘s Boroughs, one on a demographic level and one in regard to the venues‘ configuration that characterize each of the Boroughs.

After proving that both clustering methodologies are not correlated, an intersection of their results would ultimately allow us to fulfill our stakeholder‘s requirements.

II. Introduction

The Business Problem:

RightPlacer Inc. is a young Consulting Firm, specialized in helping investors setting up their business in the best location according to their needs.

A team of Business consultants from RightPlacer Inc. has been requested to advise Mr. Johnson, an important client interested in setting-up an exclusive chain of Coffee Shops in London.

Particularly, Mr. Johnson would like to know in which Boroughs should he place his Coffee Shops.

Yet, the task is not that simple as he is pursuing two objectives simultaneously: guarantee business success of his chain and create a strong Brand, Top-of-mind within the high-end segment of customers.

He claims that even though he wants to have a profitable business, he is also very concerned about building a strong brand reputation and having their coffee shops regarded as an exclusive product. An asset that has been proven very valuable in the long run.

The three consultants are having problems reaching a consensus:

-One of them suggests a demographic segmentation to focus on upscale neighborhoods with similar characteristics between them.

-The second one proposes that they would do much better implementing segmentation by venues, to locate the coffee shops in those neighborhoods where people most often go for a Coffee.

-The third one claims that any of these segmenting modalities would lead to a similar result and is indifferent to any of them.

Finally, they agree to join efforts and try different strategies to analyze their results.

How should they proceed? How could they conciliate their approaches to fulfill Mr. Johnson's requirements?

III. The Data

Since the team of Business consultants agreed to try different strategies, we will use 3 sources of data:

a) Boroughs & Coordinates:

We will start by creating a base dataset containing the different boroughs of London, along with their geographic coordinates.

To do this, we will scrap the corresponding website from Wikipedia using the *BeautifulSoup* package and prepare the data according to the format needed for this dataset.

This dataset will be later used in the Clustering phase for the geographic location of the different Boroughs. This means, we will have to make sure that the format of the coordinates is set as Float, in order to be properly parsed by the folium package.

The final dataset should have this format:

	Borough	Latitude	Longitude
0	Barking and Dagenham	51.5607	0.1557
1	Barnet	51.6252	-0.1517
2	Bexley	51.4549	0.1505
3	Brent	51.5588	-0.2817
4	Bromley	51.4039	0.0198

Datasource url: 'https://en.wikipedia.org/wiki/List_of_London_boroughs'

b) Demographic Clustering:

Next, we will need to prepare a dataset appropriate to the clustering based on demographic profiling of London's boroughs.
For this purpose, we will download a csv file from the official source (Greater London Authority (GLA)) and select the demographic features we will use for this clustering.

The goal of this phase is to identify the Boroughs' clusters that better fit with the customer's profile that Mr. Johnson is trying to target.

From the 81 available features available in the source, we will initially choose the following as the most relevant for the demographic clustering required by our Stakeholder:

-Population: GLA_Population_Estimate_2017
-Population density: Population*density*(per_hectare)_2017
-Average Age: Average_Age,_2017
-Employment rate (%): Employment*rate*(%)_(2015)
-Crime Rate: Crime_rates_per_thousand_population_2014/15
-House Price: Median_House_Price,_2015
-Median income: Modelled_Household_median_income_estimates_2012/13

The dataset should look like this:

	Borough	Population	Population Density	Average Age	Employment Rate (%)	Crime Rate	House Price	Median Income
1	Barking and Dagenham	209000	57.9	32.9	65.8	83.4	243500	£29,420
2	Barnet	389600	44.9	37.3	68.5	62.7	445000	£40,530
3	Bexley	244300	40.3	39.0	75.1	51.8	275000	£36,990
4	Brent	332100	76.8	35.6	69.5	78.8	407250	£32,140
5	Bromley	327900	21.8	40.2	75.3	64.1	374975	£43,060

As we are working with Features of different dimensions, the *StandardScaler* package will be used to normalize the data prior to the Clustering phase.

Datasource url: 'https://data.london.gov.uk/dataset/london-borough-profiles'

c) Clustering by Venues:

In this section we will use the API of Foursquare, particularly the *Explore Endpoint* and prepare a parallel clustering.
This time, based on the venue's configuration and their relevance in the different boroughs of London. The idea is to cluster the different Boroughs in regard to the Venues that are more often found in each of them. Ultimately, our intention here is to focus on Boroughs were customers normally go when looking for a Coffee Shop, as to guarantee the business success challenge presented by the Stakeholder.

This means, we will put special attention to these Boroughs with high concentration of *Cafés* or *Coffee Shops*.

Our Goal is to create a dataset like this:

	Borough	African Restaurant	Airport	Airport Lounge	Airport Service	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Australian Restaurant	BBQ Joint	Bakery	Bar	Bed & Breakfast	Beer Bar	Beer Garden	Beer Store
0	Barking and Dagenham	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0
1	Barnet	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0
2	Bexley	0.0	0.0	0.0	0.0	0.034483	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.034483	0.000000	0.0	0.0	0.0	0.0
3	Brent	0.0	0.0	0.0	0.0	0.027027	0.0	0.0	0.0	0.0	0.0	0.013514	0.0	0.0	0.0	0.013514	0.067568	0.0	0.0	0.0	0.0
4	Bromley	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.025641	0.0	0.0	0.0	0.025641	0.025641	0.0	0.0	0.0	0.0

Datasource url (API): 'https://api.foursquare.com/v2/venues/explore'

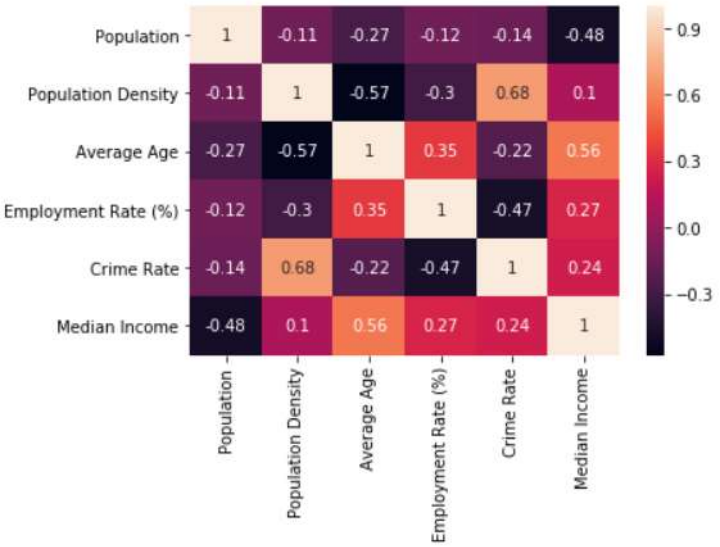
IV. Methodology

a) Exploratory Analysis of Demographic Dataset:

After our Demographic Dataset is ready, we prepared a correlation analysis to observe how our future clustering features behave between them.

There seems to be a high correlation between ‚House Price' and 3 other Features (Median Income, Crime Rate and Population Density).

Accordingly, the Feature will be excluded and the resulting Dataset tested again for correlations.



The new dataset seems much more consistent with only 1 big correlation between 2 of their variables.

We will keep it as it is for now, since removing this feature from the dataset could restrict us from relevant information for the incoming clustering phase.

b) Demographic Clustering:

Having the required dataset prepared, we are ready to implement a demographic clustering.

Before setting the model and considering that we are dealing with features of different magnitudes, we apply the normalization provided by the *StandardScaler* package.

Next, we run the run k-means with a cluster number = 5, in order to cluster the different London's Boroughs and obtain our labels. Later, we will attach the results to our ‚Boroughs & Coordinates' Dataset for mapping purposes.

As said before, the main objective of this section is to identify the cluster we could regard as high-end profiled, one of the requirements posed by Mr. Johnson for branding purposes.

c) Clustering by Venues:

Following a similar procedure as the one used along our curse we will use the ‚Clustering by Venues' Dataset, previously obtained through the Foursquare API, to perform a clustering of the different London's Boroughs in regard of the venues' configuration that characterize each of them.

The goal would be to associate the Boroughs around clusters defined by their venues' profile. Later, we will identify the idoneous profile to guarantee an appropriate flow of customers to Mr. Johnson's coffee shop chain.

Moreover, we would also look for the presence of other sort of venues that could help us further segment our clusters in regard to the profiling our stakeholder is demanding from us.

d) Adjusted Rand Index (ARI):

After we conclude the Demographic Clustering and the Clustering by Venues, we will test the hypothesis of one of the Business Consultants, which stated: ‚both segmenting modalities would lead to a similar result'.

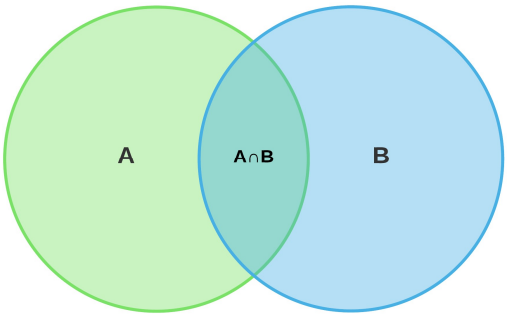
For this purpose, we will use the Adjusted Rand Index (ARI), a tool that helps comparing clustering results, delivering a score between 0 and 1. This means, equivalent clustering methodologies will hold a 1 as Adjusted Rand Index (ARI) score, whereas a 0 means that there is no correlation between clustering methods.

Moreover, a result of 1 would practically mean that we can keep working with the results of just one of the clustering modalities, while a result closer to 0 means that the clustering techniques deliver different approaches which should be evaluated independently.

e) Intersection & mapping:

Most probably, the Adjusted Rand Index (ARI) will not lead to a value equal to 1.

This means, it would make sense to use the information provided by bot our segmenting methodologies in order to reach our Mr. Johnson's goal.



For this purpose, we would apply an intersection (inner join) between the results of the 2 implemented clustering processes.

The idea of this intersection would be to identify those Boroughs which belong to both the clusters we identified as idoneous for our stakeholder: Those demographically characterized by high-profiled customers and those with a high business potential due to their venue's' configuration.

V. Results

a) Demographic Clustering:

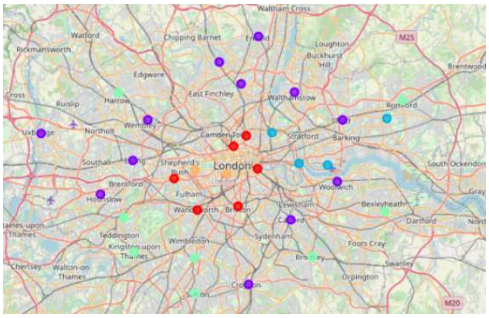
After applying our demographic clustering, we identified 2 clusters as relevant for Mr. Johnson, namely clusters 0 and 4.

Those clusters share both a relatively high House Price and also high Median Income, in accordance with the customer profile requirements the team of consultants should fulfill.

	Borough	Cluster Labels	Population	Population Density	Average Age	Employment Rate (%)	Crime Rate	House Price	Median Income
5	Camden	0	242500	111.3	36.4	69.2	123.5	700000	£43,750
11	Hammersmith and Fulham	0	185300	113	35.7	77.5	113.2	730000	£43,820
17	Islington	0	231200	155.6	34.8	72.6	121.2	583000	£39,790
20	Lambeth	0	328900	122.7	34.5	78.5	104.6	450000	£38,490
26	Southwark	0	314300	108.9	34.4	74.2	100.6	475000	£37,100
30	Wandsworth	0	321000	93.7	35.0	78.8	72.6	557000	£47,480

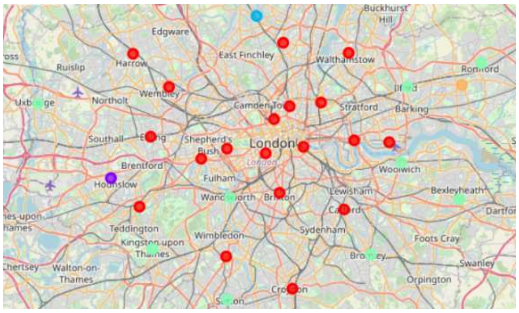
	Borough	Cluster Labels	Population	Population Density	Average Age	Employment Rate (%)	Crime Rate	House Price	Median Income
18	Kensington and Chelsea	4	159000	131.1	39.3	68.2	120.9	1200000	£55,620
31	Westminster	4	242100	112.7	37.7	65.6	212.4	920000	£47,510

The clusters, displayed with yellow and red dots in the map below, do also encompass Boroughs with a very central location within London, another relevant factor to guarantee business success.



b) Clustering by Venues:

The clustering by venues revealed the presence of one big and central cluster, with a high presence of Coffee Shops, Cafés, Hotels and even Art Galleries, what would guarantee a high affluence of target customers to our Stakeholders' premises.
This cluster is depicted with red dots in the image below:



c) Adjusted Rand Index (ARI):

A very low ARI Score evidences that both clustering methodologies deliver totally uncorrelated and therefore inequivalent results, contrary to what one of the firm consultants hypothesized.

```
from sklearn.metrics.cluster import adjusted_rand_score
```

```
adjusted_rand_score(kmeans_venues.labels_, kmeans_dem.labels_)
```

```
-0.0217877094972067
```

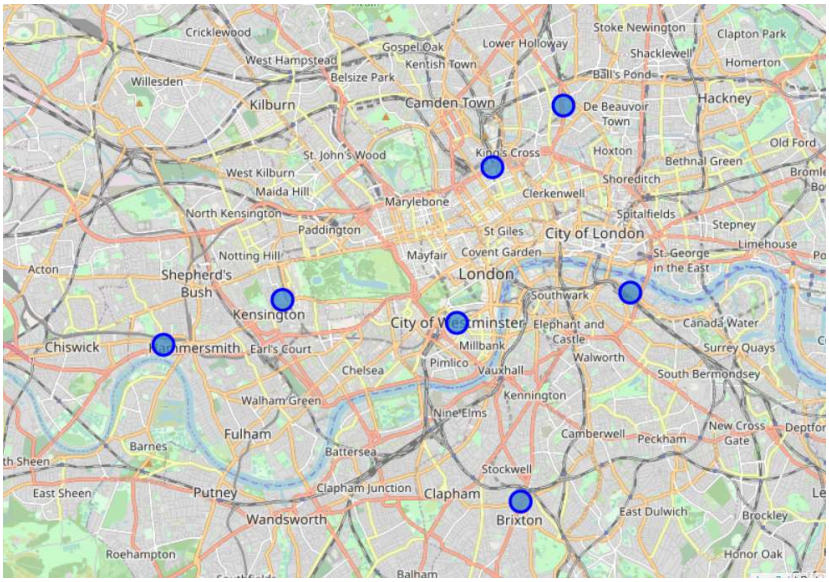
For us, this means that both methodologies of clustering deliver relevant information which should be used in conjunction as to guarantee an optimal borough selection, able to fulfill both requirements posed by Mr. Johnson.

d) Intersection & mapping:

Accordingly, we proceed to intersect the clusters we selected from the different clustering methodologies and identify the elements (boroughs) both have in common.

As a result, we obtained a short list of 7 Boroughs, which will be the ones we will ultimately recommend to Mr. Johnson for setting up his Coffee Shop chain:

- Camden
- Hammersmith and Fulham
- Islington
- Kensington and Chelsea
- Lambeth
- Southwark
- Westminster



VI. Discussion

After finishing the data processing, modelling and obtaining the results it was particularly interesting to notice how different clustering methodologies, although equally valid and consistent can lead to very different results, as evidenced by the ARI Score.

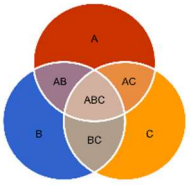
Moreover, the demographic clustering exposed the key role of the data scientist in regard to the features selection.

There may be an underlying temptation to include as much features as possible to better ‚fit' the model, yet we should always analyze the nature of the features, what do they actually represent and above all how they correlate with each other.

In this regard, the correlation matrix stands out as an excellent tool, whose visual properties help to easily spot the presence of correlations within the variables we will finally choose for the clustering phase.

VII. Conclusion

The report intends to provide an example on how combined challenges require often combined strategies.



Accordingly, the intersection of clustering methodologies should optimally solve the different challenges posed by our stakeholder and it is flexible enough to tolerate additional requirements that may arise after we present the results.

Particularly, if Mr. Johnson requires to further limit the subset of Boroughs he would like to focus on.

Finally, and referring to the initial learnings we had on the data science methodology section, in a real-life scenario it would be recommendable to incorporate our stakeholder in a more active role.

A good example of this could be a discussion about his concept he has of high-end customers, which could provide us further input in the feature's selection phase of the demographic dataset construction.