Capstone Project - Final Assessment

Project: New business location analysis, London, UK

Introduction

In this project we will try to find the the right location for a coffee shop in London (understanding location as Borough or cluster of Boroughs). The criteria to classify and rank the different Boroughs is agreed with the client. Demographic preferences.

1. Growing population
2. Working population
3. Mid-high income
4. Low crime rate

Anyone looking to open a business within the Greater London area could potentially benefit from the results shown.

Data

We are going to search for different sources of data that combined can help us make a better decision. The data we are going to analyze can be grouped into:

- Pure demographics data. Looking at areas where the population is growing.
- Profiling data including unemployment rates, average gross income and average house price and crime indicators.
- Business metrics, number of businesses, two-year business survival rate.
- Competition and other business in the area information.

As for the data sources to extract the data described above, these are the ones we chose.

- London Boroughs Population - UK National Statistics
- London Borough Profiles - London Data Store
- London Boroughs GeoJson file - Carto.com
- FourSquare API - Google Cloud
- Postcode Lookup API - Postcodes.io

Methodology

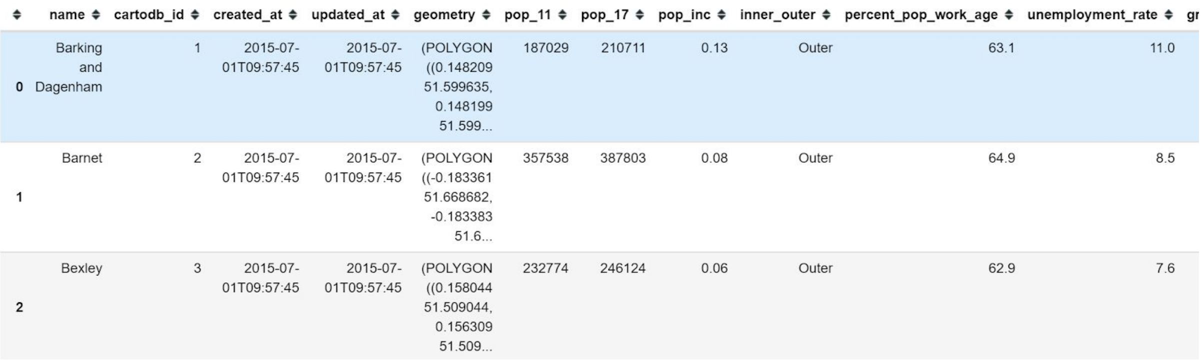
These are the steps taken for the analysis.

Getting/cleaning the data

The first step has been importing the data from the different sources. The data was available in different formats which required different actions depending on the data source.

- London Boroughs Population - UK National Statistics. This data was available in this php webpage. Using requests and Beautifulsoup we extracted the data we needed converting it in a pandas dataframe.
- Profiling data. This data was available in this official website and we simply downloaded the csv file converting it in another pandas dataframe.
- London GeoJson file was also available to download from carto.com. We simply imported the file as a geo dataframe using geopandas.
- The information about nearby businesses was extracted using the FourSquare API

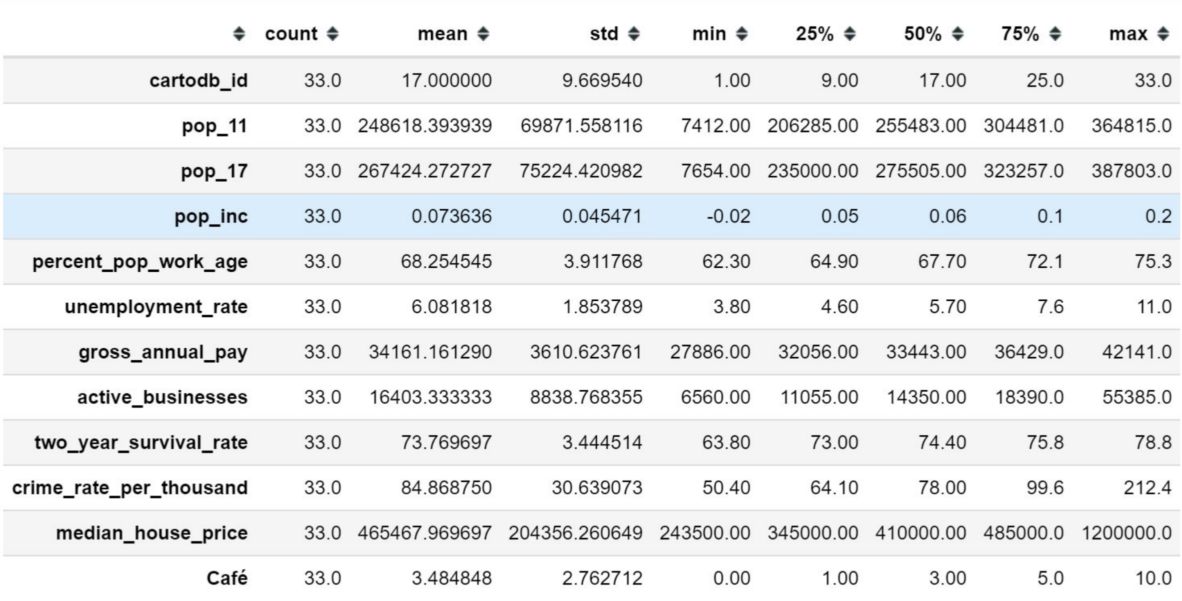
The final step of this section was to merge all data frames into one keeping the borough name as the common column.



⬍	name ⬍	cartodb_id ⬍	created_at ⬍	updated_at ⬍	geometry ⬍	pop_11 ⬍	pop_17 ⬍	pop_inc ⬍	inner_outer ⬍	percent_pop_work_age ⬍	unemployment_rate ⬍	gr
0	Barking and Dagenham	1	2015-07-01T09:57:45	2015-07-01T09:57:45	(POLYGON ((0.148209 51.599635, 0.148199 51.599...	187029	210711	0.13	Outer	63.1	11.0	
1	Barnet	2	2015-07-01T09:57:45	2015-07-01T09:57:45	(POLYGON ((-0.183361 51.668682, -0.183383 51.6...	357538	387803	0.08	Outer	64.9	8.5	
2	Bexley	3	2015-07-01T09:57:45	2015-07-01T09:57:45	(POLYGON ((0.158044 51.509044, 0.156309 51.509...	232774	246124	0.06	Outer	62.9	7.6	

Data Exploration

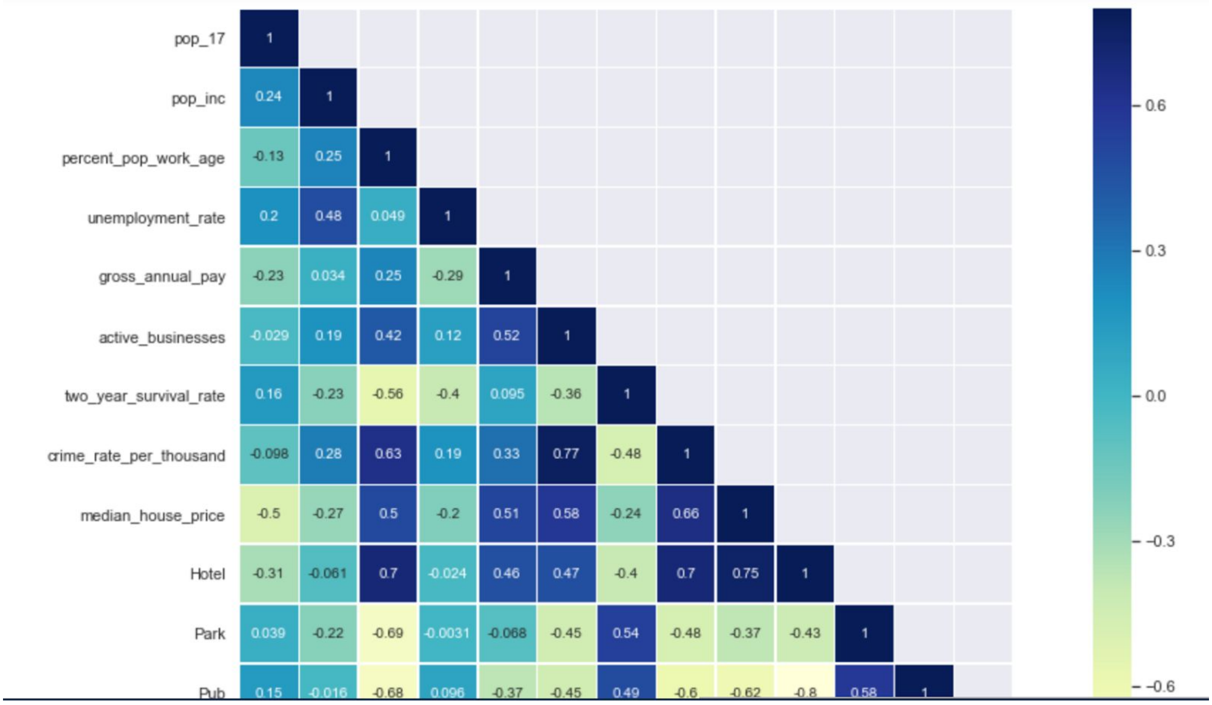
In this part of the analysis we explored different views to understand the data.
Using the describe method we could explore all the numeric variables.



⬍	count ⬍	mean ⬍	std ⬍	min ⬍	25% ⬍	50% ⬍	75% ⬍	max ⬍
cartodb_id	33.0	17.000000	9.669540	1.00	9.00	17.00	25.0	33.0
pop_11	33.0	248618.393939	69871.558116	7412.00	206285.00	255483.00	304481.0	364815.0
pop_17	33.0	267424.272727	75224.420982	7654.00	235000.00	275505.00	323257.0	387803.0
pop_inc	33.0	0.073636	0.045471	-0.02	0.05	0.06	0.1	0.2
percent_pop_work_age	33.0	68.254545	3.911768	62.30	64.90	67.70	72.1	75.3
unemployment_rate	33.0	6.081818	1.853789	3.80	4.60	5.70	7.6	11.0
gross_annual_pay	33.0	34161.161290	3610.623761	27886.00	32056.00	33443.00	36429.0	42141.0
active_businesses	33.0	16403.333333	8838.768355	6560.00	11055.00	14350.00	18390.0	55385.0
two_year_survival_rate	33.0	73.769697	3.444514	63.80	73.00	74.40	75.8	78.8
crime_rate_per_thousand	33.0	84.868750	30.639073	50.40	64.10	78.00	99.6	212.4
median_house_price	33.0	465467.969697	204356.260649	243500.00	345000.00	410000.00	485000.0	1200000.0
Café	33.0	3.484848	2.762712	0.00	1.00	3.00	5.0	10.0

We also checked if there were any null / missing values in our data set.

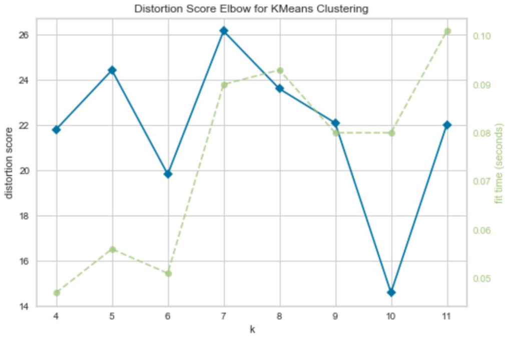
To understand the relationship between the different variables, we performed a correlation analysis which we plotted using Seaborn heatmaps.

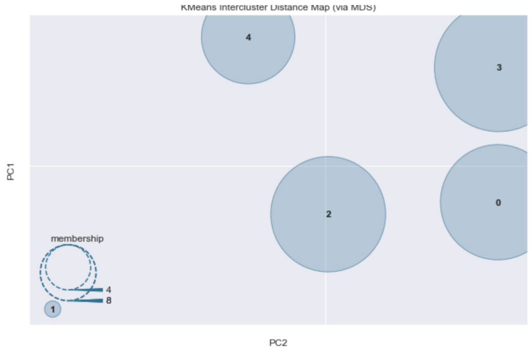


Clustering

Since the goal of the project was to group the different boroughs into clusters, we used KMeans to do it. Steps:

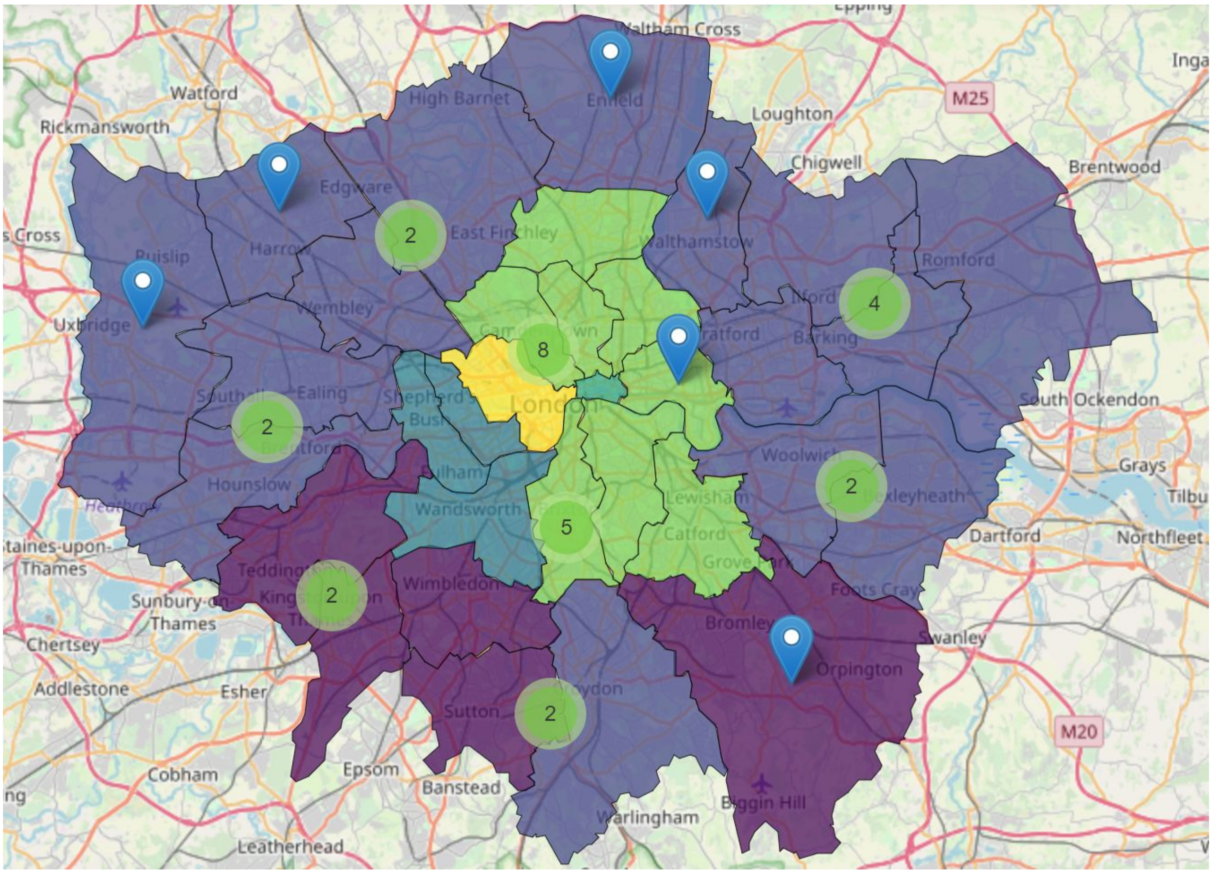
1. We imported the relevant libraries
2. We converted categorical variables into numeric using get_dummies
3. We standardize the data using StandarScaler
4. Train the model
5. Explore the labels
6. To see if we had correctly chosen the number of clusters we looked at the distortion score elbow chart (shown below)
7. To understand the distance between clusters we used the intercluster distance map from yellowbrick



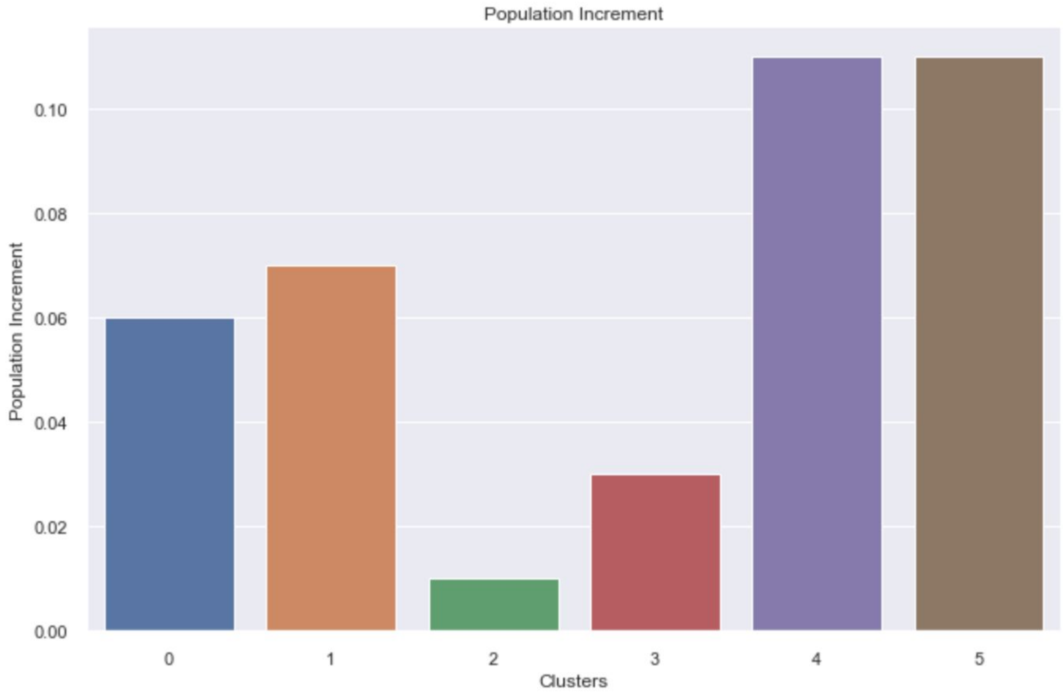


Exploring the clusters

Our next step was to explore the clusters and see the differences attending to the criteria we had defined in order to choose a winner for our project.
We also explored the clusters on a map to visually get an idea of the associations.



Then, we compared how each cluster performed against the criteria.



This comparison favoured cluster 4 above the rest. For that reason, we chose it as the ideal cluster for a new business location (according to the specs)

Building visualizations

The final section is a number of maps showing the information. This section helps our analysis and adds information visually.

Results

Cluster 4 (Camden, Hackney, Haringey, Islington, Lambeth, Lewisham, Southwark, Tower Hamlets) is declared as best option to open the business.

Recommendations

Further analysis within cluster 4 would be recommended to identify the ideal borough for the business.