Finding an appropriate place to live in London

Applied Data Science Capstone Project

Mario González Pozo

Contents

Intro

- London is the capital and largest city of England and the United Kingdom.
- It's composed by 32 boroughs plus the City of London.
- Altogether, London has a population of 9 Million (2018) and an area of 1,572 km².

Photo by John Norden / Public domain

Business problem

- The process of selecting an appropriate place to live is tedious and sometimes the decision taken leaving aside many characteristics of a place.
- The model can be considered as a prototype for a generalized model for people who wants to move to other parts of the world (not just London)
- **Is there a way to facilitate relevant information that helps taking a better decision of where to move (within London)?**

Data

- The sources of data are 2:
- 1) Economic and demographic (and miscellaneous) data from each borough of London, obtaining the data from London Borough Profiles, provided by Greater London Authority.
- 2) Foursquare API, which offers the data of nearby venues (up to 100) given a location.

Photo by gren / Public domain

Methodology

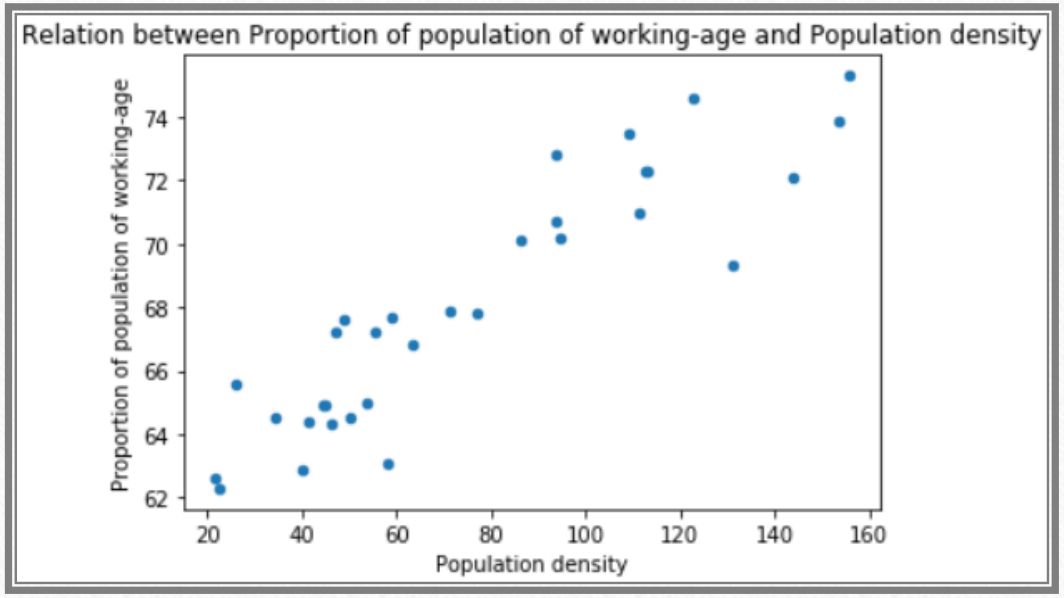
Borough profiles were reviewed, choosing the variables of interest.

Some variables were not considered due to their homogeneity (not needed for cluster analysis) or many missing values.

The City of London was removed because it has too many missing values. Other missing values were replaced by using other sources.

Some outliers were present, but the values were correct. They were kept.

Methodology



- A correlation index was calculated between every pair of variables.
- It was a found a very high correlation between the proportion of population of working-age and the population density. The later was discarded.

Methodology

- The employment rates for male and female variables are used to create a new one which represents the inequality between genders.
- The data is then normalized using standard scaler, to prepare it for cluster analysis.

Area name	Population density	Average age	Percentage of resident population born abroad	Unemployment rate	Gross annual pay	Jobs density	Two-year business survival rates	Crime rate	Median house price	Percentage of Greenspace area	Employment inequality
Barking and Dagenham	-0.465378	-1.506708	0.124633	2.655209	-1.727910	-0.577372	-0.355708	-0.047937	-1.082837	0.115542	1.323052
Barnet	-0.802304	0.525327	-0.127053	1.306531	-0.237380	-0.299539	-0.088666	-0.723545	-0.051266	0.771521	-0.273391
Bexley	-0.921523	1.310432	-1.975980	0.821007	0.005901	-0.438455	-0.188807	-1.079300	-0.921574	-0.046323	0.152327
Brent	0.024459	-0.259777	1.683152	0.767060	-1.211308	-0.438455	0.111615	-0.198072	-0.244525	-0.881205	0.109755
Bromley	-1.400994	1.864624	-1.763014	-0.419776	0.899629	-0.438455	1.513583	-0.677852	-0.409756	2.177189	-0.613965

Methodology

- Cluster analysis was done by using KMeans method from scikit-learn. The values for k were 4 to 7, iterating 500 times each with a random seed, to look for the best model.
- The user is supposed to select a cluster from the group chosen to work with, considering the characteristics of it matches the user preferences. Then Foursquare is called to display the most common venues for each borough inside a cluster.

Results

- The following clusters were obtained:
- **Cluster A1**: The average age from these regions is relatively high and have a low population density of which just a few are immigrants. The job density is low, but the unemployment rate is also low, being similar for both genders. Description for the cluster would be tranquil boroughs which have a stable but small economy, is inhabited by older people, and having many natural and outdoor areas.
- **Cluster A2** seems to be another small and stable economy with low population density, having a great difference between employment rate between genders. Other features are average.
- **Cluster B1** is characterized for having relatively young people and a relatively low income along with the lowest job density. This could be residential areas too that have small businesses.

Results

- **Cluster B2** has many immigrants and young people and very high unemployment. House prices are low, as well as income and job density. Overall, this cluster has boroughs with a relatively bad economy in global terms, i.e., in a sense of inequality where poverty coexist with a minority of people who has a good status. Given there is a low greenspace percentage, the area is probably associated with residential and commercial areas.
- **Cluster C** has a high density of people of working age and a considerable percentage of immigrants. The unemployment is low and there isn't much difference between male and female employment rates. The jobs density is high and the income too, having high house prices. The crime rate is considerably higher. This cluster could be seen as an important economic region, probably an industrial or commercial area with many business/workplaces going on and large workforce.
- **Cluster D** corresponds to only 1 borough which is clearly different from the others, it has a very high job density and gross annual pay, has the highest migrant resident population, extremely high crime rate and highest median house price, with a great percentage of green areas.

Results

- The top 10 most common venues are obtained for every borough:

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Barking and Dagenham	Park	Coffee Shop	Pub	Café	Supermarket	Clothing Store	Portuguese Restaurant	Grocery Store	Gym	Gym / Fitness Center
1	Barnet	Café	Coffee Shop	Pub	Park	Supermarket	Turkish Restaurant	Bakery	Golf Course	Pizza Place	Bar
2	Bexley	Pub	Park	Grocery Store	Coffee Shop	Gym / Fitness Center	Café	Furniture / Home Store	Supermarket	Italian Restaurant	Golf Course
3	Brent	Park	Pub	Pizza Place	Café	Coffee Shop	Middle Eastern Restaurant	Bakery	Gym / Fitness Center	Portuguese Restaurant	Indian Restaurant
4	Bromley	Pub	Park	Pizza Place	Gym / Fitness Center	Coffee Shop	Indian Restaurant	Italian Restaurant	Gastropub	Café	Bar

Results

- An example was tested, assuming one is interested in a tranquil cluster to live and not interested in work aspects, basically like cluster A1, and also it's of interest venues related to food. The following table shows the alternatives that match the preferences. The best alternative would likely be Bromley since it has many food related venues.

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
4	Bromley	Pub	Park	Pizza Place	Gym / Fitness Center	Coffee Shop	Indian Restaurant	Italian Restaurant	Gastropub	Café	Bar
14	Havering	Hotel	Park	Cocktail Bar	Coffee Shop	Pizza Place	Bookstore	Ice Cream Shop	Clothing Store	Gym / Fitness Center	Movie Theater
25	Richmond upon Thames	Pub	Garden	Park	Café	Coffee Shop	Botanical Garden	Gastropub	Italian Restaurant	Rugby Stadium	Hotel

Discussion

- A solution to automate the cluster making and, in this way, generalize the model to another places, is by using machine learning and certain criteria which leads to well defined clusters. This doesn't form part of the developed model considering its simplicity versus the complexity of training a machine learning model to think and discern which cluster is good or bad characterized (and the number of clusters to be used).
- More data about the boroughs could've been useful too, to get a broader representation of a borough, like knowing what kind of economic activities are primarily performed, if it's commercially focused, residential area, a tourism area, etc.
- The limit of getting 100 venues per borough at a time could also affect the model accuracy, since we are trying to encompass a wide area (whole borough), and possibly the lost of information is important to characterize each borough.

Conclusion

- A prescriptive model was developed to help decision making of a person in need of readily available information about certain places. This said, a person who wants to move to London, probably would like research further about the place which looks better by the use of the model.
- Also, the model is very limited since it just includes London (target audience is very bounded). The generalization of the model to other places is a must if we would like it to be more valuable.