Ascertaining similarities between Cities using Data Science techniques in Python

By Edward Pritt

Author Note

For completion of the following course:
https://www.coursera.org/professional-certificates/ibm-data-science

14/4/2020

1. Introduction

I have given myself a theoretical problem to solve, this ‘problem’ however is one which will affect many people every year. Whilst the exact specifics of the issue maybe somewhat different for each person’s circumstances; I hope to show that using data science techniques we can apply methods to assist in decision making.

2. Problem

I individual currently live in Camden, located in the City of London in the UK, has recently been offered a relocation package to a branch in the City of Sydney in Sydney Australia.
The problem is simple – “Based on the businesses located in the surrounding neighborhood of their current place of residence – can I try to narrow down locations similar to Camden in the City of Sydney in Australia?”

3. Interest

Anyone looking to move from Camden to Australia, however in practice the analysis can be adapted to compare any two locations.

4. Data acquisition and cleaning

In the interest of efficiency and due to this being theoretical I had to "design" some of my own data. I am sure that the data exists however it was quicker and easier for me under these circumstances to gather some manually myself.

This data lead to the creation of two SCV files which I made using google maps co-ordinates for locations using a list of "neighborhoods" (known as localities in Australia).

I kept the data separate, one for London and one for Sydney and joined these in my analysis rather than creating the "perfect" dataset. This did create an environment where "noise" and missing data was not an issue which in real life would be a likely scenario. Dropping out lines of data with missing values and copying over data was already tested in the previous assessments and therefore has not been repeated here.

Additional data was to be gathered from FOURSQUARE, via their Developer API portal where the list of pre-gathered locations would be combined with local businesses, this would be my final dataset.

5. Exploratory Data Analysis

	Neighbourhood	Venue Category
count	1931	1931
unique	34	245
top	Kings Cross	Café
freq	100	211

I. There we 34 locations in total (which was coincidentally a 50/50 between London and Sydney)
II. Within 500 meters of any of these 34 a total of 245 unique categories of locations were found
III. Kings Cross contained the most locations within 500 meters
IV. Cafés made up the largest number of venue categories (211 in total) or approximately 11% of the total number of venues returned (1931)

6. Data aggregation and clustering

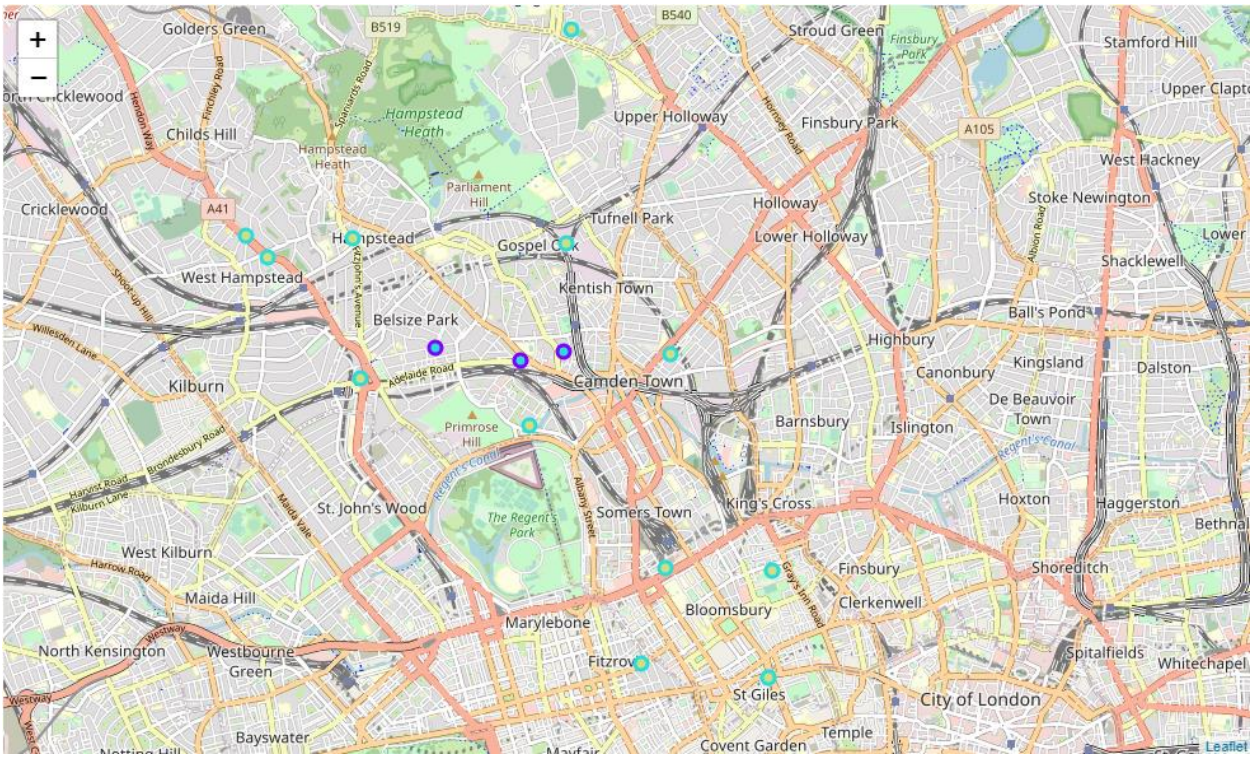
First Cluster (Cluster_0)
14 Neighborhood's in total

Second Cluster (Cluster_1)
1 Neighborhood in total

Third Cluster (Cluster_2)
8 Neighborhood's in total

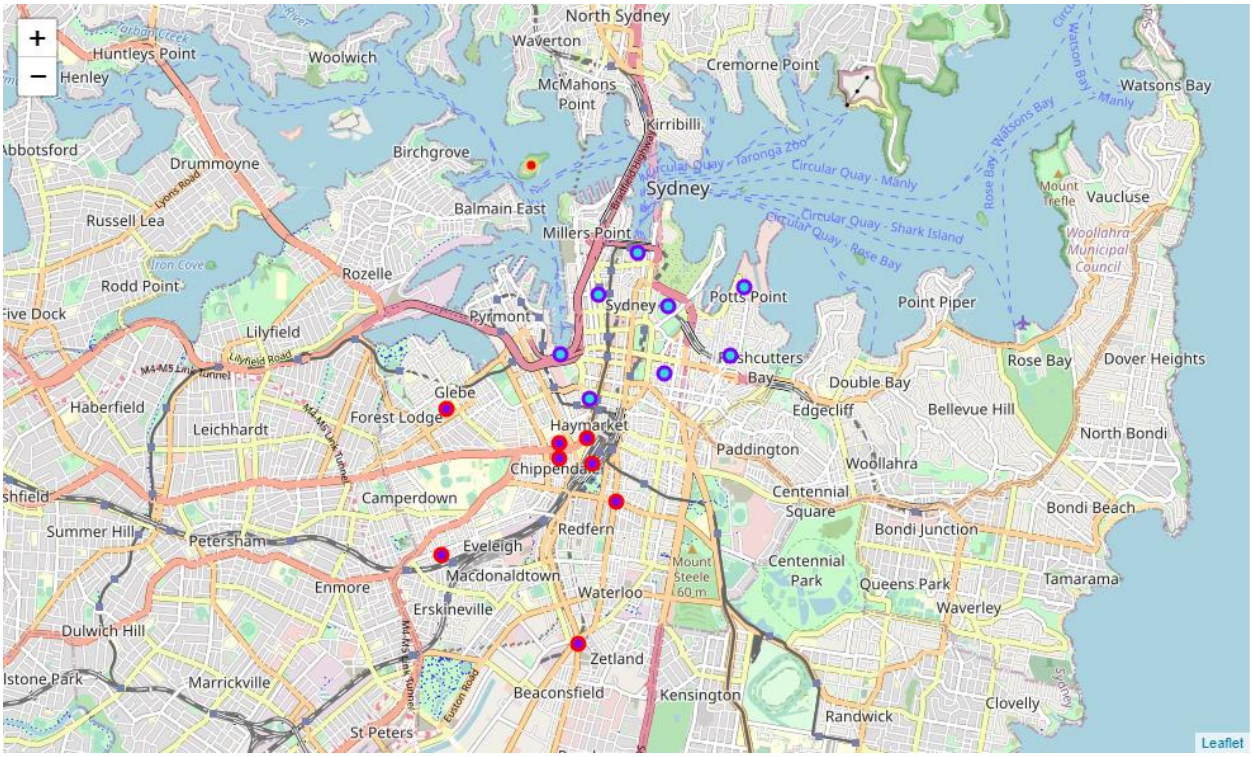
Fourth Cluster (Cluster_3)
11 Neighborhood's in total

Fig.1 London Visualization of clusters



Note: light blue and yellow color scheme is difficult to see in this report.

Fig 2. Sydney Visualization of clusters



7. Conclusion

In its simplest form we can conclude that the 1st cluster is too bespoke to London and that its mix of venues is too niche and therefore offers us no insight into potential alternative locations for our relocation

The 2nd cluster appears to be an outlier of sorts; this area or neighborhood in Sydney is like no other and forms its own cluster, again therefore in isolation is not suitable.

The 3rd cluster is much like our 1st cluster however it is this time situated in Sydney rather than London.

Our 4th and final cluster is the only cluster that contains neighborhoods from both London (3) and Sydney (11) based on the information and analysis of local businesses this is therefore our only directly comparable neighborhoods and should be the 1st area to be explored for potential relocation.

8. Closing comments and future considerations

This model was simple and only factored in one real world consideration (the locality of businesses). In reality there will be other considerations such as local transport and affordability which would affect the decision making process.

Now that some areas appear to share similarities in their economic environment these could be further investigated. More careful analysis of the classification of businesses to create more effective clusters or using different metrics which are considered important to the user should be considered using different metrics for example number of check-ins to indicate how “busy” a location might be.