Battle of the Neighbourhoods

Sydney, New South Wales,
Australia.

Gopalrathnam Venkataraman, 07/08/2019.
IBM Coursera Capstone Project.

Contents

List of tables

List of Figures

Summary

The project is aimed at solving the problem of a Sydney local resident who has recently moved to a new suburb and wants to confirm the following

- Is the new suburb similar to the old suburb?
- Is the new suburb conducive for an active life in physical fitness activities?

The old suburb is Parramatta, Sydney, Australia and the new suburb is Hornsby, Sydney, Australia.

The result is based on the most common venues found in both the suburbs and whether the suburbs are classified under the same cluster.

Introduction

Sydney has many suburbs which are unique in their culture. The project is aimed at suggesting whether the suburb the person has moved into is similar to the suburb he was living earlier. This suburb similarity comparison was done by taking 765 suburbs that are within 50 kilometres form the centre of the city

Data Collection

Suburb data:

The suburbs that are within 50 kilometres from the city centre is taken from 'Freemaptools' website. Please see appendix for the link.

	Postcode	Suburb	Distance
0	2061	MILSONS POINT	0.95
1	2000	MILLERS POINT	0.53
2	2000	THE ROCKS	0.41
3	2000	DAWES POINT	0.00
4	2060	NORTH SYDNEY SHOPPINGWORLD	1.98
...	...	...	...
760	2750	PENRITH PLAZA	49.08
761	2563	MENANGLE PARK	49.70
762	2756	SACKVILLE	49.83
763	2570	KIRKHAM	49.69
764	2753	LONDONDERRY	49.45

765 rows × 3 columns

*Figure { SEQ Figure * ARABIC } - Suburb list and distance from the city*

In addition, the geo location data, such as latitude and longitude for all suburb is taken from public website 'matthewproctor'. Please see the appendix for the link

	postcode	locality	State	long	lat	id	dc
0	6532	CARRARANG	WA	115.004595	-28.440886	10861	GERALDTON DC
1	6532	COBURN	WA	115.004595	-28.440886	10862	GERALDTON DC
2	6532	COOLCALALAYA	WA	115.004595	-28.440886	10863	GERALDTON DC
3	6532	DARTMOOR	WA	115.004595	-28.440886	10864	GERALDTON DC
4	6532	DEEPDALE	WA	115.004595	-28.440886	10865	GERALDTON DC
...	...	...	...	...	...	...	...
16833	2823	DANDALOO	NSW	147.838499	-32.076870	5181	TRANGIE LPO
16834	2823	GIN GIN	NSW	147.838499	-32.076870	5182	TRANGIE LPO
16835	2823	TRANGIE	NSW	147.838499	-32.076870	5183	TRANGIE LPO
16836	2824	BEEMUNNEL	NSW	147.785831	-31.373201	5184	WARREN LPO
16837	2824	EENAWEENA	NSW	147.785831	-31.373201	5185	WARREN LPO

*Figure { SEQ Figure * ARABIC } - Geo locations for all suburbs*

Data consolidation and preparation

Both datasets are combined and cleaned to obtain the dataset of all suburbs within 50 kilometres from Sydney city

	Postcode	Suburb	Distance	long	lat
959	2176	ABBOTSBURY	31.53	150.885260	-33.872014
957	2763	ACACIA GARDENS	31.18	150.895470	-33.725619
1108	2560	AIRDS	43.25	150.768408	-34.194216
774	2234	ALFORDS POINT	22.82	151.004412	-34.021938
458	2100	ALLAMBIE HEIGHTS	10.74	151.257781	-33.766409
...	...	...	...	...	...
665	2199	YAGOONA WEST	17.29	151.023171	-33.905586
894	2233	YARRAWARRAH	27.76	151.027919	-34.114626
764	2161	YENNORA	22.06	150.983724	-33.858335
802	2228	YOWIE BAY	23.63	151.101564	-34.037692
327	2017	ZETLAND	6.04	151.206316	-33.903892

*Figure { SEQ Figure * ARABIC } - Cleaned data set containing distance and geo co-ordinates*

Neighbourhood Location Data

Neighbourhood data is sourced to identify the similarities of all the Sydney suburbs that are within 50 kilometres from the city, by using the Foursquare APIs. Neighbourhood data for each suburb was sourced via the Foursquare API, including the most favourite places, such as parks, gyms, cafes, restaurants, stores etc to compare the resident's current suburb to new suburb and consider similarities.

Planned Methodology and Workflow

	Project phase	Status	Date	
0	Scope Verification	Completed	04/08/2019	
1	Data collection	Completed	04/08/2019	From public sites
2	Data consolidation	Completed	05/08/2019	Python
3	Data preparation	Completed	05/08/2019	
4	Data Visualization	Completed	06/08/2019	Folium
5	Clustering	Completed	06/08/2019	
6	Recommendations	Completed	07/08/2019	

*Table { SEQ Table * ARABIC } -Methodology and workflow*

Data Visualization

For suburb data visualization, used folium choropleth maps to show all the suburbs in Sydney that are within 50 kilometres from the city.

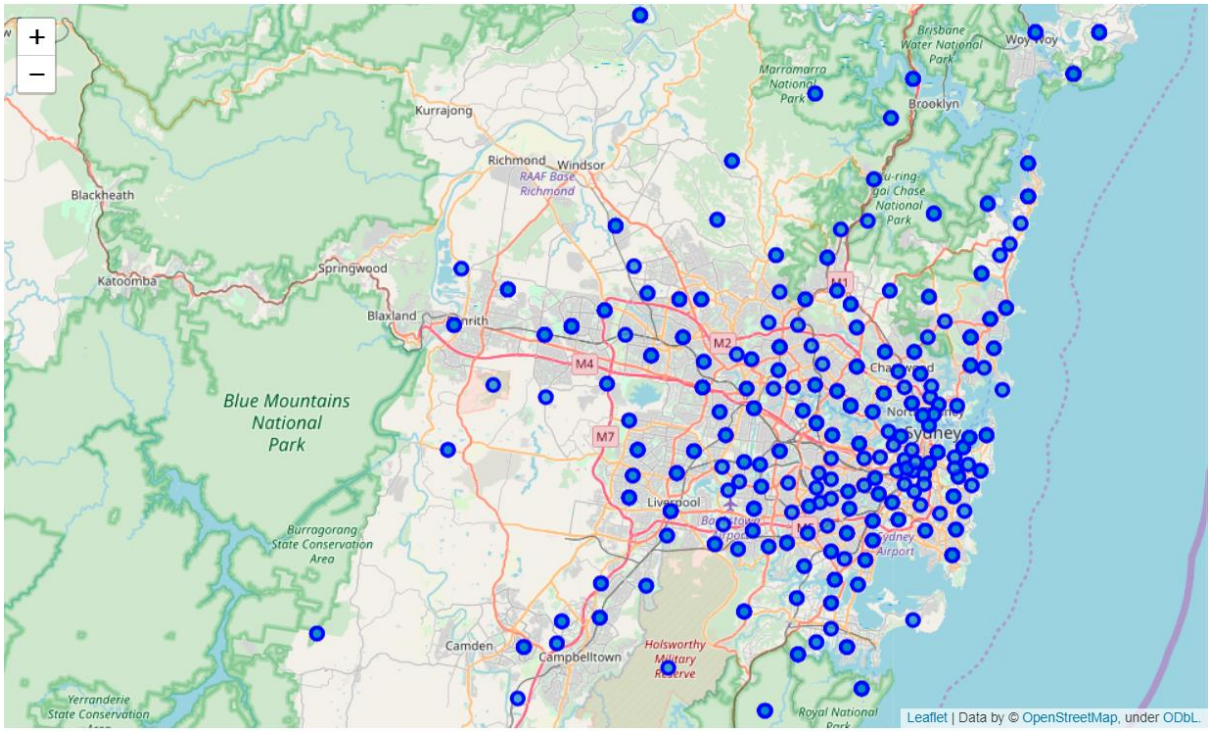

*Figure { SEQ Figure * ARABIC } - Sydney suburbs within 50 kilometres from city centre*

Location data and analysis

Location based data is was sourced from Four Square, with the popular venues for all suburbs listed in a pandas dataframe, for all suburbs

	Neighborhood	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	ABBOTSBURY	Pub	Café	Grocery Store	Women's Store	Diner	Fish & Chips Shop	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant
1	ACACIA GARDENS	Convenience Store	Bus Station	Women's Store	Discount Store	Fish Market	Fish & Chips Shop	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant
2	ALLAMBIE HEIGHTS	Bus Station	Women's Store	Diner	Fish Market	Fish & Chips Shop	Field	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Electronics Store
3	ALLAWAH	Thai Restaurant	BBQ Joint	Pub	Train Station	Electronics Store	Dog Run	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Women's Store
4	ARNCLIFFE	Middle Eastern Restaurant	Intersection	Paper / Office Supplies Store	Café	Grocery Store	Tunnel	Australian Restaurant	Falafel Restaurant	Electronics Store	Eastern European Restaurant

*Figure { SEQ Figure * ARABIC } - Most popular venues for a suburb*

Apply Clustering algorithm

To find similar suburbs across all suburbs in Sydney, a Kmeans algorithm was applied taking into consideration all the common venues found in each of the suburb.

In order to find the optimum K, an elbow method is used and arrived at an optimum K values to be 5, to cluster around 600 suburbs. The output is not so much of an elbow, but an optimum value is chosen

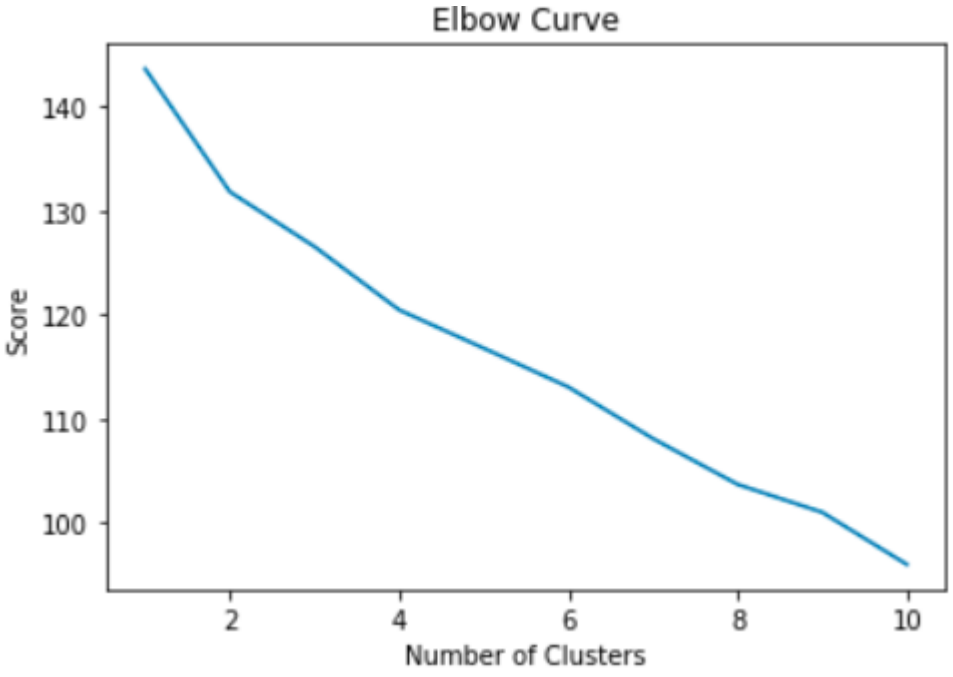

*Figure { SEQ Figure * ARABIC } - Elbow curve*

With the K value, the output of the suburb cluster is shown below

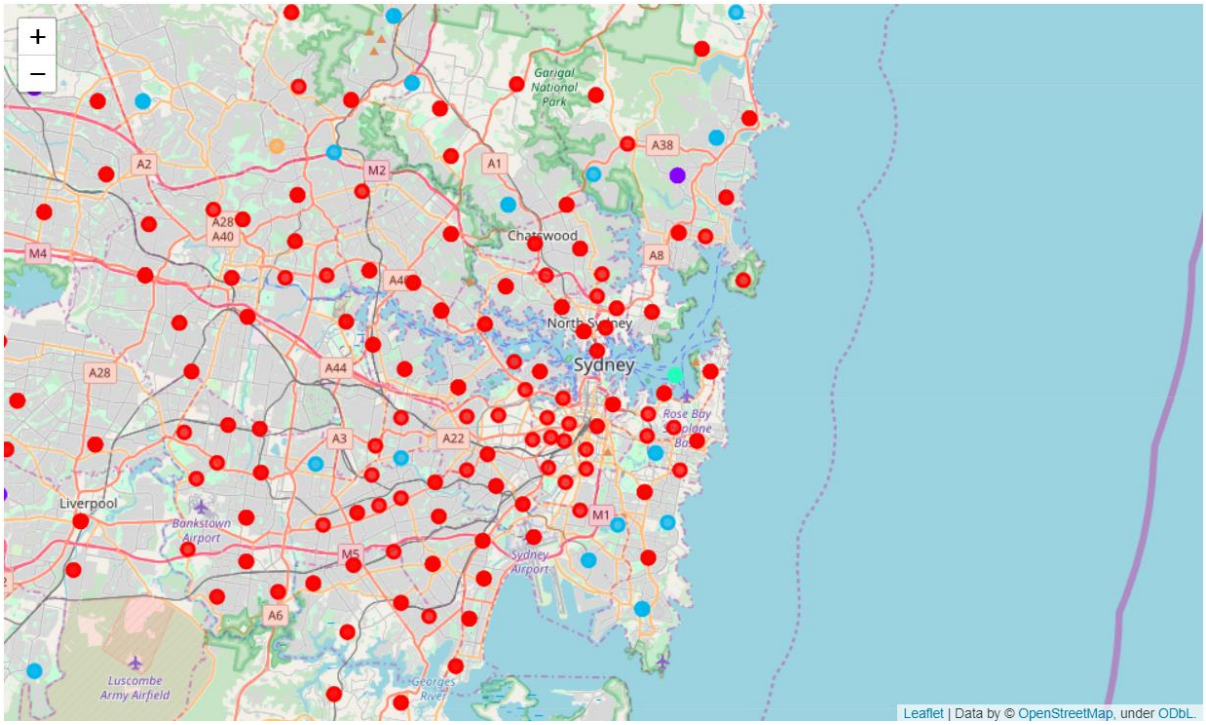

*Figure { SEQ Figure * ARABIC } - Sydney suburb clustering*

Clustering analysis:

Looking at the cluster, we find the following

- Most of the Sydney clusters are similar in nature, as this is evident from plethora of red dots
- Parramatta and Hornsby are a different suburb, as they belong to a different clusters.
- Looking closely into the common venues for Parramatta and Hornsby, it is evident that the culture and lifestyle of the suburb are different. Parramatta is largely into Food and eating out whereas Hornsby is about fitness and outdoor activities.

	Postcode	Suburb	Distance	long	lat	Cluster Labels	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
714	2077	HORNSBY	19.59	151.097508	-33.691922	2	Playground	Coffee Shop	Sports Club	Field	Discount Store	Fish Market	Fish & Chips Shop

	Postcode	Suburb	Distance	long	lat	Cluster Labels	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
702	2150	PARRAMATTA WESTFIELD	19.56	151.006506	-33.814463	0	Thai Restaurant	Chinese Restaurant	Café	Coffee Shop	Asian Restaurant	Burger Joint	Dessert Shop

*Figure { SEQ Figure * ARABIC } - Comparison of suburbs*

Summary and Recommendations

We started the exercise with two questions to be answered. Let us revisit the questions and check whether the clustering algorithm has answered these.

1. Is the new suburb similar to the old suburb?

From the clustering algorithm, we find that the Parramatta and Hornsby are not similar. Parramatta's common venues are related to cafes and restaurants, whereas Hornsby is all about playgrounds, sports clubs and fields.

2. Is the new suburb conducive for an active life in physical fitness activities?

From the clustering algorithm, we can see that Hornsby will satisfy the person's appetite for outdoor and fitness activities as three of the top four common venues are related to fitness activities.

Python packages and dependencies

Python package	Dependency
Pandas	Library for data dictionary and analysis
Numpy	Library to handle arrays
Geopy	Library to retrieve location data
Requests	Library to handle http requests
Matplotlib	Library for data visualization
Folium	Library for rendering maps
Sklearn	Library for Kmeans algorithm

*Table { SEQ Table * ARABIC } - Packages and dependencies*

Appendix:

project Data	Data source
Australian postcodes with geo co-ordinates	https://www.matthewproctor.com/australian_postcodes
Australian postcodes with distance from city centre	https://www.freemaptools.com/find-australian-postcodesinside-radius.htm

*Table { SEQ Table * ARABIC } - Links to data sources*