Capstone project of IBM Data Science Professional Certificate
Battle of Neighbourhoods

1. Introduction

This is the capstone project of IBM Data Science Professional Certificate. The aim of this project is to leverage the Foursquare location data to explore or compare neighbourhoods or cities to analyse how these data can solve selected problems.

2. Problem description

Since the beginning of 2020, all people get panic because of the alarming spreading speed of 'Covid-19' pandemic, trigging the biggest global recession. In contrast, the local property market of Canberra, the capital city of Australia, has boomed because of its good performance in fighting with the virus. Both sales price and transaction numbers hit a record high every month since April 2019, attracting more people entering the market and push the price higher correspondingly.

Almost all the people around are either actively seeking in the property market or talking about the market. In this report, we are trying to help Sandy find a Suburb where she will be most comfortable with and at an acceptable price. Currently, she stays in Dickson 2602 and works at Red Hill 2601. Based on her preference, this report will focus on four areas to analyse the comfortable score: health, education, safety and convenience.

3. Data

To find the most promising suburb for Sandy, two main datasets were used for analysis:

a) House price

b) Nearby venues

3.1 House price

Complete suburbs and districts list were extracted from the official GeoJson data provided by the ACT Government GeoHub API (https://actmapi-actgov.opendata.arcgis.com).

Then these lists will be used as an reference to scrap the corresponding useful data from Wikipedia website by using https://en.wikipedia.org/wiki/Canberra_Central as a starter to redirect to each targeting suburb.

Finally, based on the suburbs list, 30,873 transactions in 108 suburbs listed on the real estate website between 2010 and June 2021 will was scrapped for the analysis, among which, 1,473 transactions in 102 suburbs was in the first half year of 2021.

3.2 Nearby venues

In this report, the number of hospitals, community services, walk-in centres and general practises will be taken as health index, which will be scrapped from website https://health.act.gov.au/hospitals-and-health-centres and https://www.whitecoat.com.au.

The number of schools will be used as education index and the number of police stations and crimes of each suburbs will be used as safety index. These data will be get from ACT Open Data Portal https://www.data.act.gov.au.

Finally, as the requirement of this project, the Foursquare API will be used to get the nearby venues of each suburb as an index for the convenience.

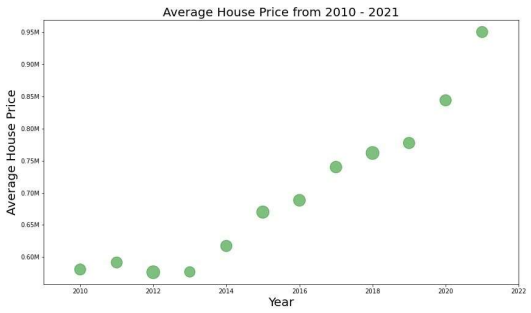
The summary of the dataset is shown as below:

	type	venues	categories
0	Convenience	1696	224
1	Health	45	4
2	Education	153	4
3	Safety	6	1

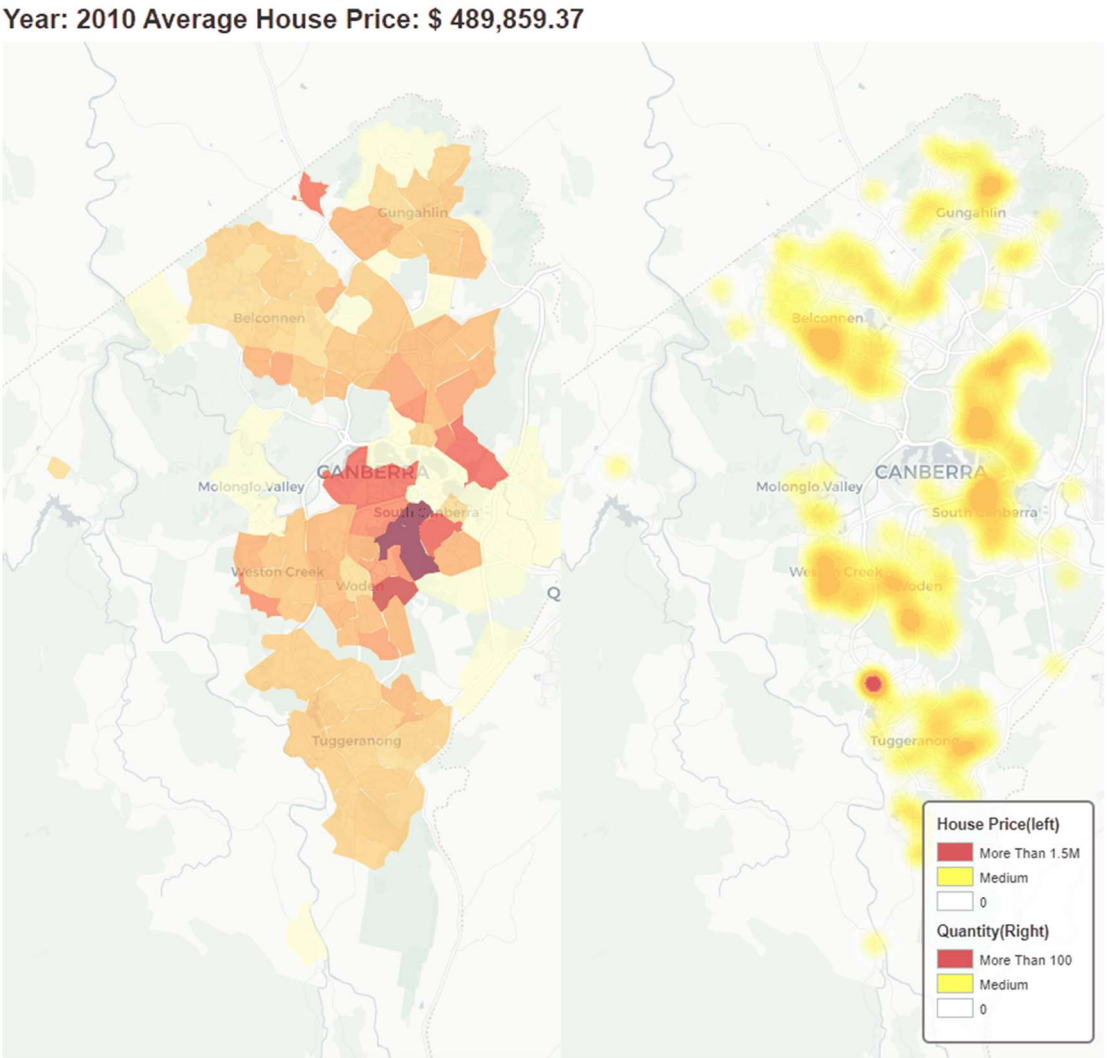
4 Analysis

4.1 House price

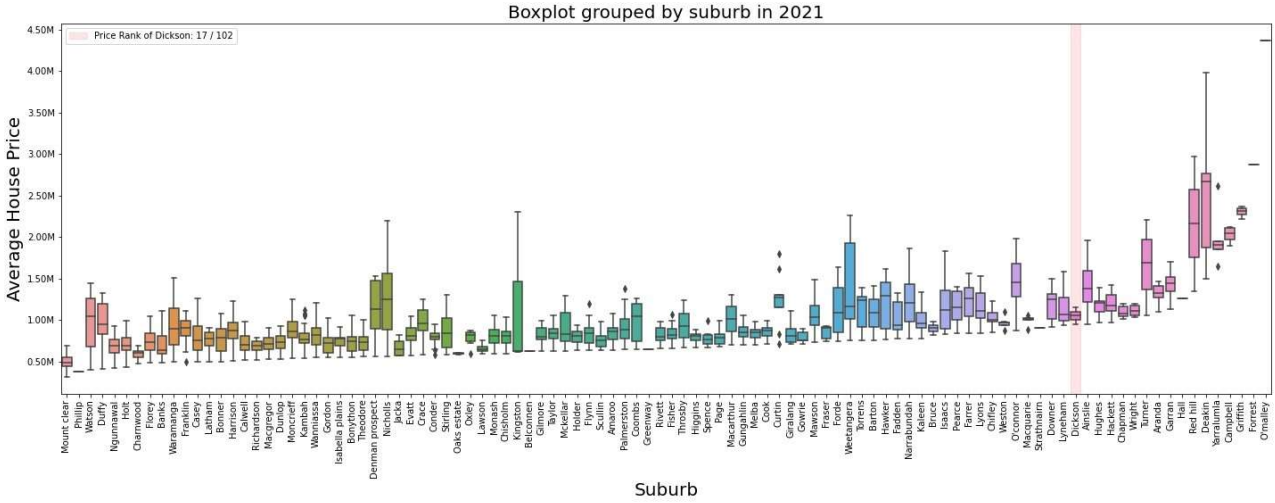
First thing first, exploratory data analysis will be used to analyse the house price of Canberra.



Left figure shows the trend of house price from 2010 to 2021. The average house price started from only 0.5 million in 2010 and then increasing dramatically from 2013. Another turning point was on 2019 with a more rapid increasing rate, pushing the average house price to almost 1 million, which can be shown in the DualMap shown below as well.



By further analysis of average house price in 2021 using the boxplot as below, the house price of suburb where Sandy currently lives range from 0.9 million to 1.4 million with average price at about 1.1 million. It ranks at 17th highest out of all suburbs with data recorded, which means that the suburb she currently lives in is much expensive than most of other suburbs. Therefore, we will find the similar suburbs with lower price for sandy where she is most comfortable with.



4.2 Nearby venues

Two methods will be used to analyse the nearby venues. The first one is to find the most common venues of each suburb and ranks it based on the number of different venue types. The second method is to divide the suburbs into different groups, based on the nearby venues, by using k-mean clustering algorithm.

4.2.1 Most common venues

1) Based on Sandy's preference(Restaurant, Café, Grocery Shop, Pub & Bar, Gym, Park, Health, Education, Safety): Dickson ranks as the 14th with 6 out of 9 types of venues chosen by Sandy, and there are 23 suburbs have equal or higher scores than Dickson.

	suburb	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	Numbe Of Venues Type
0	Gungahlin	Restaurant	Health	Café	Grocery Store	Education	Pub&Bar	Gym	Park	Safety	9
1	Belconnen	Restaurant	Café	Health	Grocery Store	Pub&Bar	Gym	Safety	Education	Park	9
2	Phillip	Restaurant	Café	Grocery Store	Gym	Health	Education	Park	Pub&Bar	Safety	9
3	Greenway	Restaurant	Café	Health	Grocery Store	Gym	Education	Safety	Park	None	8
4	Barton	Café	Park	Pub&Bar	Restaurant	Education	Grocery Store	Gym	Health	None	8
13	Dickson	Restaurant	Café	Health	Education	Grocery Store	Pub&Bar	None	None	None	6

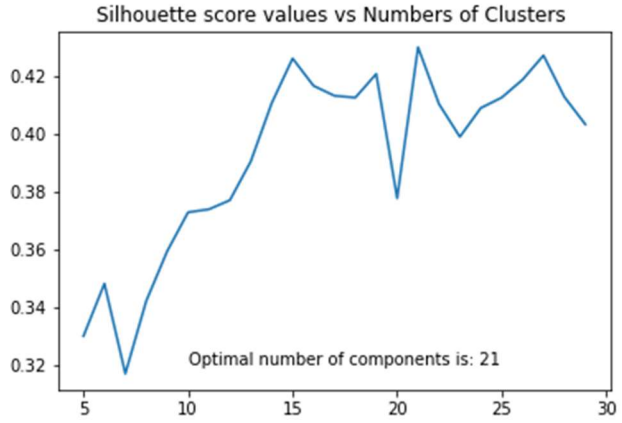
2) Based on all the available venue categories (233 unique categories in total): Dickson ranks as the 12th with 32 out of 233 types of venues chosen by Sandy, and there are 12 suburbs have equal or higher score than Dickson

	suburb	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	...	225th Most Common Venue	226th Most Common Venue	227th Most Common Venue	228th Most Common Venue	229th Most Common Venue	230th Most Common Venue	231th Most Common Venue	232th Most Common Venue	233th Most Common Venue	Numbe Of Venues Type ALL
0	Braddon	Noodle House	Burger Joint	Fast Food Restaurant	Restaurant	Food & Drink Shop	Comic Shop	Ice Cream Shop	Pizza Place	Dry Cleaner	...	None	None	None	None	None	None	None	None	None	64
1	City	Burger Joint	Mexican Restaurant	Noodle House	Shopping Plaza	Department Store	Molecular Gastronomy Restaurant	Music Venue	Indian Restaurant	Gym	...	None	None	None	None	None	None	None	None	None	64
2	Belconnen	Thai Restaurant	Liquor Store	Paper / Office Supplies Store	Hardware Store	Market	Pub	Post Office	Bar	Middle Eastern Restaurant	...	None	None	None	None	None	None	None	None	None	50
3	Reid	Hotel	Sushi Restaurant	Noodle House	Sandwich Place	Plaza	Molecular Gastronomy Restaurant	Japanese Restaurant	Theme Park Ride / Attraction	Museum	...	None	None	None	None	None	None	None	None	None	46
4	Phillip	Steakhouse	College	Community Health Centres	General Practise	Ice Cream Shop	Bakery	Juice Bar	Italian Restaurant	Liquor Store	...	None	None	None	None	None	None	None	None	None	42
11	Dickson	Vegetarian / Vegan Restaurant	Sports Club	Memorial Site	Mountain	Pharmacy	Malay Restaurant	Locksmith	Nature Preserve	Mexican Restaurant	...	None	None	None	None	None	None	None	None	None	32

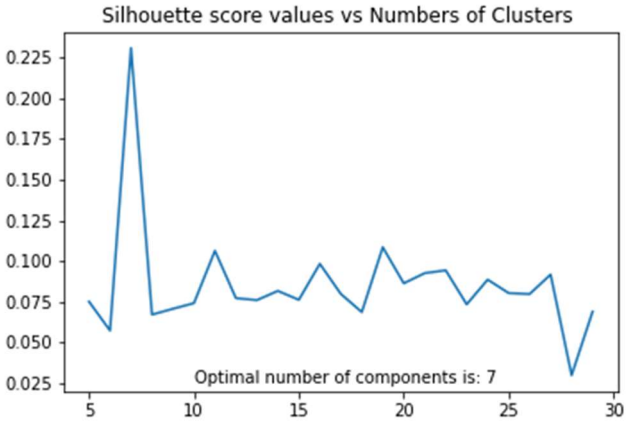
4.2.2 K-means clustering

In this part, Silhouette score was used to determine the optimal k size, and the same technique was used as Most common venue, the data was analysed in two ways:

1) Based on Sandy's preference: the optimal k size based on Sandy's preference is 21, and 16 suburbs were identified in the same cluster as Dickson.



2) Based on all the available venue categories: the optimal k size based on Sandy's preference is 7, and 23 suburbs were identified in the same cluster as Dickson.

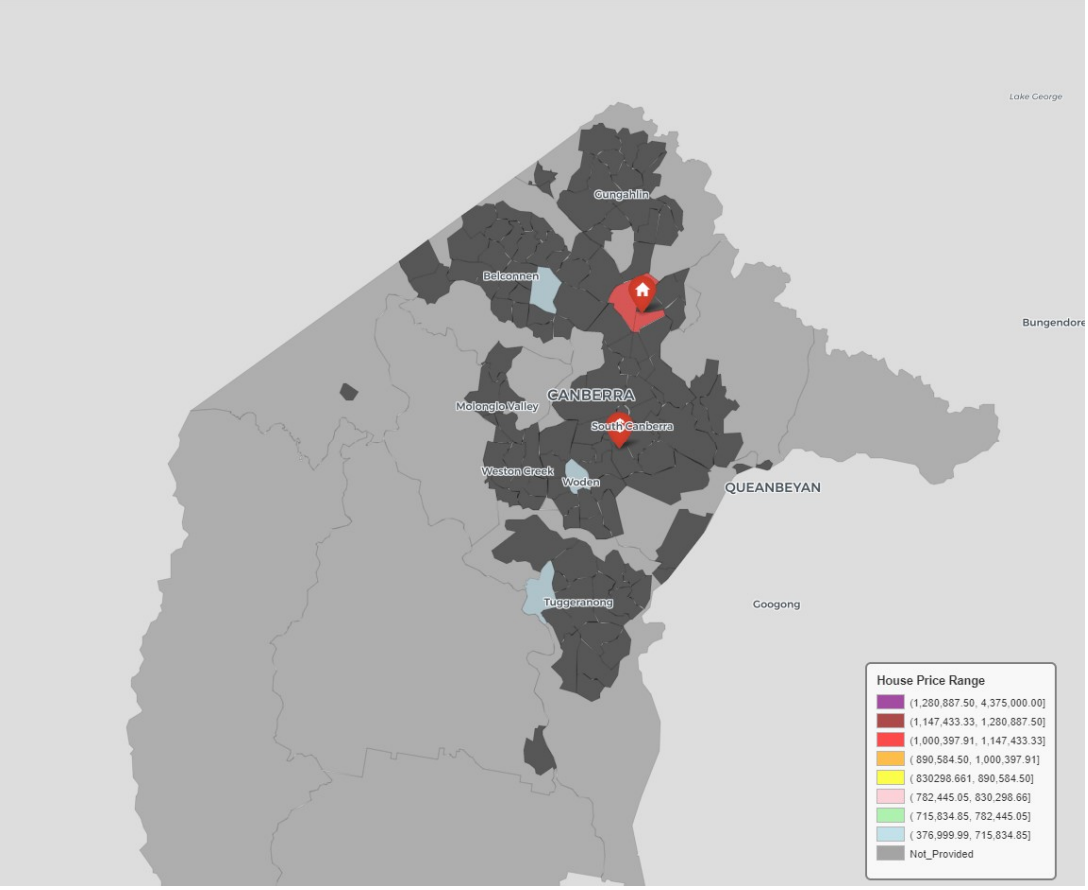


5. Results and Discussion

Different analyses have been performed to find the suburbs for sandy. Based on the assumptions: 1)the suburb should within the same group as Dickson in both preferred venues and all venues method; 2)the suburb should have equal or higher venue types as Dickson in both preferred venues and all venues method; 3)the average house price should lower than her budget which is 1.5 million. The suitable suburbs have been narrowed down to only 5 suburbs: Dickson, Lyneham, Belconnen, Greenway and Phillip.

	suburb	district	population	density/km	avg_price	safety_score	Numbe Of Venues Type	Numbe Of Venues Type ALL
0	Dickson	Canberra Central	2,149	1,340	1,051,250.0000	0.0009	6	32
1	Lyneham	Canberra Central	5,112	929	1,146,333.3333	0.0008	6	33
2	Belconnen	Belconnen	6,657	1,513	620,000.0000	0.0000	9	50
3	Greenway	Tuggeranong	1,894	357	651,000.0000	0.0004	8	34
4	Phillip	Woden Valley	2,936	1,129	377,000.0000	0.0003	9	42

Finally taking her current workplace into consideration, among 5 suburbs selected as the best choice for Sandy, Greenway should be popped out because of the long commute time.



6. Limitations

For simplified purpose, the data scrapped from the website have not been validated. And there are some assumptions made in this project which may not fit for all situations.

7. Conclusion

Choosing a suitable suburb to purchase a dream house is a challenging task, so the help from the data analysis can make this decision more educated and visualized. Similarly, data analysis can also be used to solve many other real-life problems.