The Battle of the Suburbs in Canberra

Applied Data Science Capstone by IBM/Coursera

Fanpeng Kong

Contents

1 Introduction: Business Problem

Canberra, the capital city of Australia, is often known as "The Bush Capital" to Aussies (and yes, we put that on our vehicle number plates). It is nothing like the usual metropolis like Madrid or Beijing which you would expect from other countries. It is an entirely planned garden-like city and in recent years, Canberra has been ranked among the world's best cities to live. As per 2021, there are about 430 thousands residents living accross the 814 square kilometeres region and the poupulation density of Canberra is only about 528/km2. This figure is not only low compared to other capital cities around the world but also only ranks 6 among all the 8 Australia capitals.

However for any new settlers in Canberra, it could be a bit hard at the beginning to find the right area to live in. Canberra has a total of more than 100 local suburbs and 7 districts (or towncenters as Canberrans refere to) and most of the facities, entertaiments and large shopping centers are located in these towncenters. And if you ask any Canberran for suggestions about the using the public transport system to travel between these suburbs or towncenters, he/she propaly would recommend you to get a car as the first thing after you find an accommadation. The rental cost or property price among these suburbs can also differ to a large extent based on their locations.

Assume that we have a young family who just moved from interstates or overseas and would like to purchase a 3 or 4 bedrooms house in Canberra to settle down. They set a budget for thier property hunting. What would be the best suburbs that they should look into in terms of propety price, travel distance to towncenters etc. In this study, I will try to find the answer for them using the data science skills that I have acquired during the courses.

2 Data

It is necessary to set out the required dataset for clustering and segmenting the suburbs first before attempting any data collection steps. As the choice of the interesting suburb to purchase property is mainly based on several aspects such as real estate price and convenience to towncenters and access to local facilities. After some considerations, I decided to use two different datasets for the suburbs clustering and segmentation analysis. The main dataset includes the median house price with 3 or 4 bedrooms in each suburb, distance of each suburb to its nearest towncenters and the number of venues within 2km from each suburb's center point. The second dataset mainly consists of types or category information of venues close to each suburb within the same distance.

During the data collection state, I exercised different methods from webscraping to open datasets which have been covered in previous courses:

- For the list of towncenters and suburbs in Canberra, webscraping was used to extract relevant fields from the table found in the Wikipedia page List of Canberra Suburbs
- Postcode for each suburb was extracted from an open dataset found on Matthew Proctor's website.
- The latitude and longitude of towncenters and suburbs were retrieved using the geocoder library, whereas the distance from suburb to towncenter was calculated using the Google Distance Matrix API or the Python Haversine library.
- The number and type of venues around each suburb was retrieved by calling the Forsquare APIs.
- The median prices for 3 or 4 bedrooms houses were scraped from Domain's suburb profile page using `requests` and `BeautifulSoup` libraries.

In the following sections, steps to retrieve these data will be explained in more details.

2.1 Get towncenter and suburbs list

The first step for data collection was to get a list of towncenters and suburbs which belong to them. It turned out that this dataset with an easy to use format like .csv or spreadsheet is not easily found online. The Wikipedia page of Suburbs of Canberra includes such information in the expandable tables near the bottom.

A small challenge here was that those foldable tabels are nested and therefore same information may occure at different hierarchy levels. After an cross examination between the webpage and the html inspector, I was able to narrow down the innerest talbe that containing the list of suburbs together with their corresponding towncenters.

Canberra Central	**North Canberra**	Acton · Ainslie · Braddon · Campbell · City · Dickson · Downer · Hackett · Lyneham · O'Connor · Parkes · Reid · Russell · Turner · Watson
	South Canberra	Barton · Capital Hill · Deakin · Forrest · Fyshwick · Griffith · Kingston · Narrabundah · Parkes · Red Hill · Yarralumla
Belconnen		Aranda · Belconnen · Bruce · Charnwood · Cook · Dunlop · Evatt · Florey · Flynn · Fraser · Giralang · Hawker · Higgins · Holt · Kaleen · Latham · Lawson · Macgregor · Macnamara · Macquarie · McKellar · Melba · Page · Scullin · Spence · Strathnairn · Weetangera
Gungahlin		Amaroo · Bonner · Casey · Crace · Forde · Franklin · Gungahlin · Harrison · Jacka · Mitchell · Moncrieff · Ngunnawal · Nicholls · Palmerston · Taylor · Throsby
Molonglo Valley		Coombs · Denman Prospect · Molonglo · Whitlam · Wright
Tuggeranong		Banks · Bonython · Calwell · Chisholm · Conder · Fadden · Gilmore · Gordon · Gowrie · Greenway · Hume · Isabella Plains · Kambah · Macarthur · Monash · Oxley · Richardson · Theodore · Wanniassa
Weston Creek		Chapman · Duffy · Fisher · Holder · Rivett · Stirling · Waramanga · Weston
Woden Valley		Chifley · Curtin · Garran · Hughes · Farrer · Isaacs · Lyons · Mawson · O'Malley · Pearce · Phillip · Torrens

I decided to separate the *Canberra Central* district into *North Canberra* and *South Canberra* for the analysis. The html source for each district row corresponds to a `tr` tag, within which are a `th` tag representing the *Towncenter* name and a `td` tag consisting of an unordered list denoting the suburbs (`ul` tag).

```
<tr>
  <th class="navbox-group" scope="row" style="width:1%">
    <a href="/wiki/Belconnen" title="">Belconnen</a> event
  </th>
  <td class="navbox-list navbox-odd" style="text-align:left;border-left-width:2px;border-left-style:solid;width:100%;padding:0px">
    <div style="padding:0em 0.25em">
      <ul>
        <li>
          <a href="/wiki/Aranda,_Australian_Capital_Territory" title="Aranda, Australian Capital Territory">Aranda</a> event
          ::after
        </li>
```

After retrieved this table information, I separated each suburb to individual rows and create a Pandas DataFrame `df_can`. This master DataFrame hosting the main datasets for the suburbs was updated in the following steps to include postcode, geo location, median house price, distance to towncenters and number of venues. Hereafter I stored the `df_can` to individual .csv files at different stages which allowed me to read the datasets from stored files instead of webscraping again for future runs should we wish (for speed or data consistency concerns).

2.2 Add postcode

Although the postcode for each suburb was not directly used in this analysis, it was necessary to collect them as I needed them later when scraping the house price information from Domain website. A FreeDatabase of Australian Postcodes have been found on Matthew Proctor's website. After downloading the dataset, I filtered the postcode for Canberra suburbs and merged them into previous `df_can` DataFrame. Apart from the postcode fields, this open dataset also includes other information such as geo location and zone data which might be useful for future study. I decided to use `geopy` library to get the latitude and longitude by myself instead of using the data from this dataset as an exercise in next step.

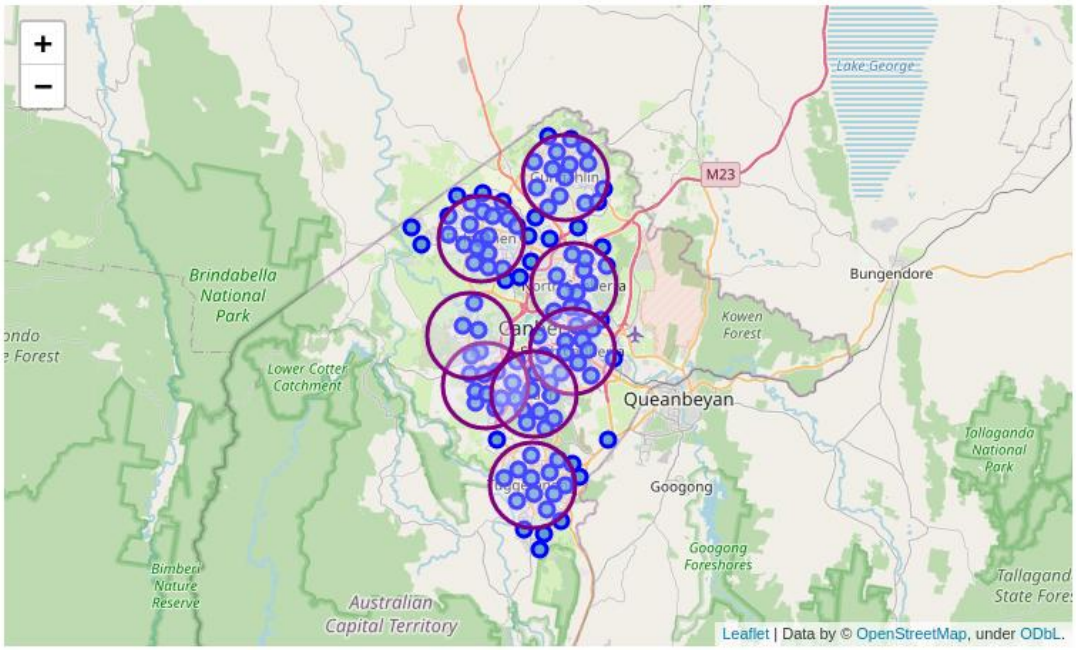
2.3 Add geo location

To get the latitude and longitude information for each suburb, the `geopy` library demonstrated in previous labs is used.

A separate DataFrame `df_town` was created to include the latitude and longitude information for towncenters. After merging the geo location into main DataFrame `df_can`, it looks like below:

	Towncenter	Suburb	Postcode	Latitude	Longitude
0	North Canberra	Acton	2601	-35.285232	149.112968
1	North Canberra	Ainslie	2602	-35.262195	149.147880
2	Gungahlin	Amaroo	2914	-35.169587	149.128021
3	Belconnen	Aranda	2614	-35.258055	149.080426
4	Tuggeranong	Banks	2906	-35.471889	149.099657

Based on the geo information of all the suburbs and towncenters, I created the following Folium map to visualize the towncenters and suburbs:



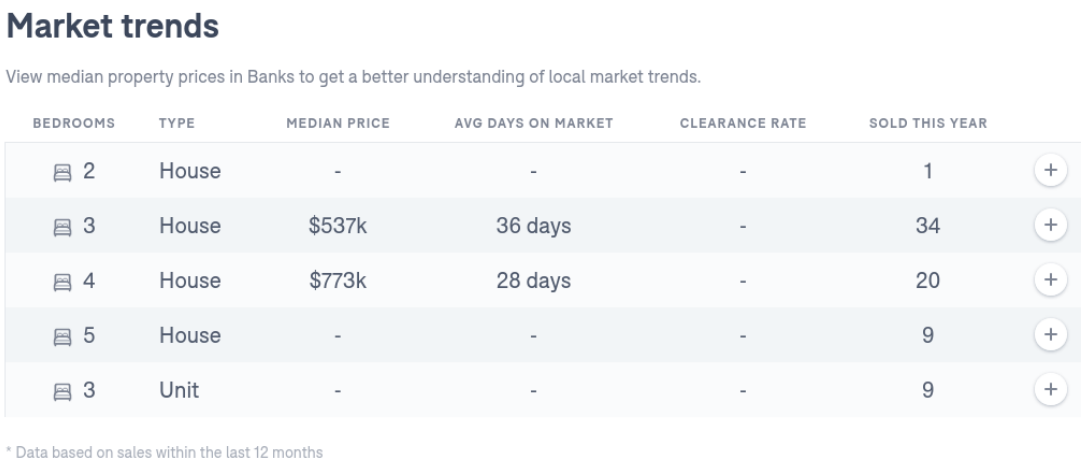
2.4 Driving distance to nearest Towncenter

My first attempt to calculate the driving distance between suburb to towncenter was to use Google Distance Matrix API following this blog post. Unfortunately at the time of this study, there is no longer free options to use the Google Distance Matrix API. As an alternative, I explored the Haversine Distance as an alternative to the driving distance.

By iterating each suburb and towncenter and calling the `haversine` function, a distance matrix representing the Haversine Distance between suburbs and towncenters were created and stored into a DataFrame. An extra column `Nearest` hold the distance of a suburb to its closest towncenter was appended, although this nearest towncenter may not necessarily be the demographic one that a suburb belongs to. This nearest distance to towncenter information was merged into the master DataFraem `df_can`.

2.5 Add median price

Again, the dataset for the real estate price for each suburb is not readily (or at least freely) available online. Domain provides a suburb profile page where you can enter the name of a suburb and serch its profile which happens to include a Market trends table. For example, following figure shows the property sales information under a section named *Market trends* found in suburb *Banks*:



Market trends

View median property prices in Banks to get a better understanding of local market trends.

BEDROOMS	TYPE	MEDIAN PRICE	AVG DAYS ON MARKET	CLEARANCE RATE	SOLD THIS YEAR
2	House	-	-	-	1
3	House	$537k	36 days	-	34
4	House	$773k	28 days	-	20
5	House	-	-	-	9
3	Unit	-	-	-	9

* Data based on sales within the last 12 months

The price of properties under interest are 3 or 4 bedrooms house for the young family who want to settle down in Canberra. A little bit of exploration on the URLs for each suburb revealed that the suffix part has a pattern of `suburb-name/act/postcoe` with special character in the suburb name like single quote or space being replaced by a - character. Using the webscraping techniques (`requests` and `BeautifulSoup`), I was able to extract the relevant field data from this table.

It is worth to note that not all suburbs have this Market trends table as it is based on the sales data in the past 12 months. Additionally even the table exists, it is not always to have the price for 3/4 bedroom house for the very same reason. I took care of these exceptions during the webscraping process and below I will explain how to handle the missing information for some suburbs.

2.5.1 Fixing missing prices

After a close examination of the price data for 3 or 4 bedrooms house, I found the missing rates for only 3 or 4 bedrooms are below 10% whereas missing rate for both 3 and 4 bedrooms are even higher 28%. Instead of dumping the rows with missing values or using the column mean to fill in the missing values directly which is likely to insert many same price values, I used the following strategy to fix the missing prices:

1. Assume the ratio for 4-bedroom and 3-bedroom price is similar in all the suburbs, calculate this ratio based on the average of suburbs which contain both values.
2. For the suburbs mising only one price, let's use the other price and the above average ratio to estimate the other missing price.
3. Finally for suburbs missing both prices, use the column mean respectively.

The average price for 3 and 4 bedrooms house was calculated and merged into the master DataFrame `df_can`.

2.6 Use Foursquare to check venues

Similar to previous labs covered in the course, I used the Foursquare API to explore the venues around each suburb and categorize them. As mentioned in the beginning, Canberra has a quite extended geography and a rather low population density, the radius to count the venues for each suburb was set to 2km. The returned venues were grouped by each suburb and the total number of venues for each suburb was added into the main DataFrame. This colcludes the end of main data collection and the master DataFrame `df_can` now looks like below:

	Suburb	Towncenter	Postcode	Latitude	Longitude	Nearest	Median Price	venue count
0	Acton	North Canberra	2601	-35.285232	149.112968	2.921057	3044.733334	100.0
1	Ainslie	North Canberra	2602	-35.262195	149.147880	1.413531	14295.000000	61.0
2	Amaroo	Gungahlin	2914	-35.169587	149.128021	1.106799	726.000000	38.0
3	Aranda	Belconnen	2614	-35.258055	149.080426	4.783948	3044.733334	29.0
4	Banks	Tuggeranong	2906	-35.471889	149.099657	5.702031	655.000000	9.0

2.6.1 Venue category analysis

After grouping the venues by each suburb, I sorted the type of venues for each suburb and found the top 10 most common venues in each suburb. This dataset was added into an additional DataFrame `df_venues` and was used later on to cluster and segment the suburbs from the perspective of venues. An example of the top 3 venues for each suburb is shown below and unsurprisingly Café and Hotels are found to be the most common ones in the suburbs close to CBD like Acton or Ainslie:

	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Acton	Café	Hotel	Coffee Shop
1	Ainslie	Café	Hotel	Chinese Restaurant
2	Amaroo	Café	Fast Food Restaurant	Supermarket
3	Aranda	Gym	Café	Supermarket
4	Banks	Supermarket	Sandwich Place	Fried Chicken Joint

3 Methodology

In this study, I planned to categrize suburbs into different groups with similar profiles based on different metrics. In the previous data collection stage, I have assembed two Pandas DataFrames: one consists of the median price for 3 to 4 bedroom houses, distance to closest towncenter and total number of venues in each suburb within 2km. Whereas in the second DataFrame,the top 10 most common venues in each suburb are included.

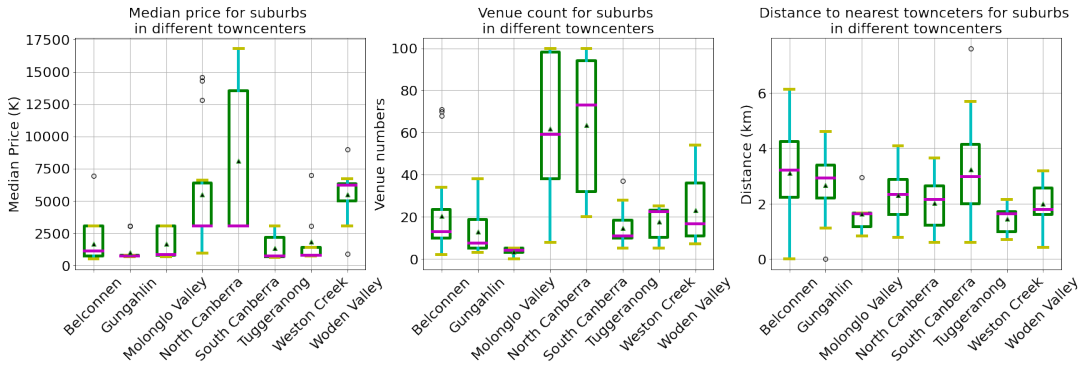
In the next analysis step, I will first **check some statistics** of our dataset, for example, histogram of house prices and cloest distance to towncenter. Then I perform unsupervised **k-means clustering** on these two datasets to cluster and segment suburbs. After clustering, I will visiualize the clustered suburbs using Folium.

By the end of the study, I should have well segmented suburbs and be able to give recommendations to the young family seeking for properties according to different criteria.

4 Analysis

4.1 Get some statitics

Before proceeding to the clustering and segmentation on the suburbs, I performed some basic explanatory data analysis on the main DataFrame. I explored statitics of median house price, number of venues and distance to nearest towncenters for each suburb grouped by their governing towncenters.



From the above 3 plots, it is evident that there are more venues in North and South Canberras and the house price there are also much higher than suburbs in other towncenters. The distance for suburbs to closed towncenter

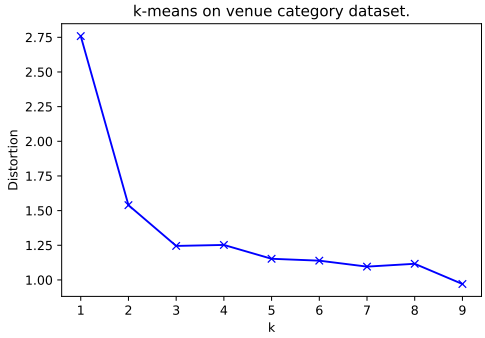
however does not present a significant difference which indicate the well planned geo structure of Canberra suburbs.

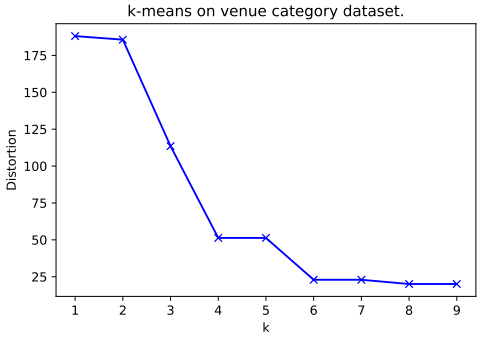
4.2 k-means clustering on master and venue category dataset

K-means clustering is one of the most commonly used clustering techniques to perform unsupervised learning on datasets. It aims to partition the dataset into k clusters in which each sample belongs to the cluster with the nearest mean. The k-means clustering minimises the data sample variance within clusters and therefor is useful for this study to group suburbs with similar profiles like median house price or distance to towncenters.

To perform k-means clustering on the master dataset `df_can`, several irrelevant columns such as *Towncenter*, *Postcode*, *Latitude*, *Longitude* are dropped from the DataFrame. Then the relevant data are normalized first before the clustering. For k-means on the second venue category dataset, the number of different venue categories in each suburb are averaged first as a one-hot pre-processing.

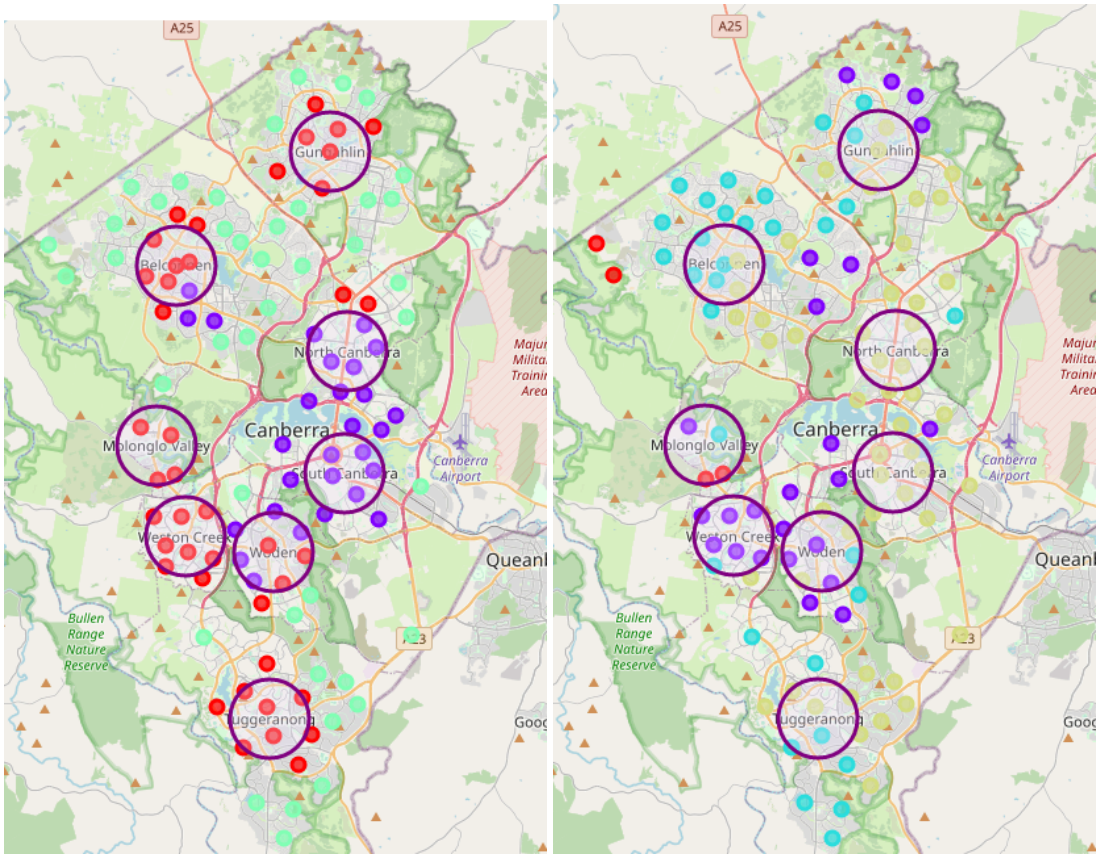
To find the optimal number of clusters, the elbow method will be used where the optimal k value is decided when the quality of the clustering see a sharp drop. The tow figures below shows the distortion of the clusters on the two datasets and the optimal number of clusters are found to be 3 and 4 for the main and venue category datasets respectively.





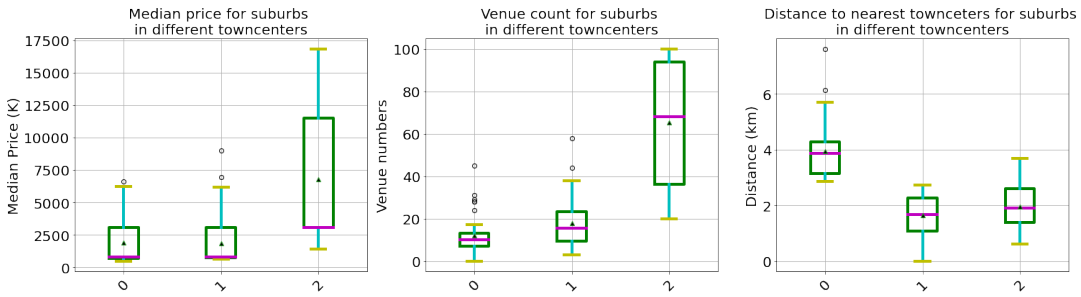
The two figures below showed Folium maps and the clustered suburbs based on the main dataset the venue category dataset, together with towncenters. From the left figure, it is clear that most suburbs in North and South Canberra have been segmented into one category. The other cluster include those suburbs which are very close to towncenters but do not belong to the inner north or south Canberra. The rest suburbs which are located near the edge of each local towncenter make up the last category.

The right figure showing the suburbs clusters based on the venue category information tells a different story though. While suburbs in North and South Canberra still fall in one category, they are joined by some suburbs from the Tuggeranong area. The other obvious trend is that suburbs belonging to north and south towncenters seem to differ from the type of venues they host. This might be related to the development and poupulation growth shift in Canberra happened in recent years but will need some further analysis.



4.3 Revisit statistics on clustered suburbs

After clustering and segmenting all the Canberra suburbs based on the main data set, i.e. median price for 3 or 4 bedrooms houses, distance to nearest towncenters and number of venues within 2km radius, the same statitics were performed again on the suburbs grouped by clusters instead of towncenters. Cluster 2 includes most of the inner north and south suburbs in Canberra. These suburbs cover the parliamentary triangle of Canberra and many of the national galleries and institutions can be found here too. Therefore the number of venues are much higher than the other two clusters. The downside of these suburbs are the much higher median house prices as shown in the first figure. Cluster 1 suburbs mainly represent those close to local towncenters and in my opinion have the best trade-offs: surly the do not have as many venues as the inner ones but their house prices are also much lower. The last cluster 0 is made up by the suburbs near the edge of each local towncenters evidenced by the nearest distance to towncenters shown in the right most figure. They do not show many advantages compared to suburbs in the other two groups: for example, their distance to nearest towncenter almost doubled than those in cluster 1; the number of venues are much less while the median price for 3 or 4 bedrooms house showed a very similar profile to cluster 1 suburbs.



5 Results and Discussion

Using the main dataset, most cluster 1 suburbs are located at the inner north or south Canberra. While these suburbs are well established with various venues, their price are also much higher than other suburbs. Suburbs

in cluster 2 have the lowest price, but meanwhile they also do not have as many venues and the travel distance to towncenters are higher. Cluster 0 suburbs seem to have the best balance for all three factors and not superisingly they are located around other towncenters apart from the North and South Canberra.

Clustering results using venue category information are even more interesting: suburbs in North and South Canberra (together with Tuggeranoon) again are clustered into one segment while suburbs in other north towns (Belconnen and Gungahlin) are segmented together. Suburbs in Woden and Westen Creek share some similarities unlike other towns. These difference might very well becasue of the developmnet shift from South Canberra to North Canberras. Due to the scope of the study, I will not furth explore the behind reasons.

Based on the above analysis, hopefully our young family now have more information to refere to when they finnaly decide which suburb to settle in.

6 Conclusion

In this study, various techniques varing from webscraping to open dataset have been used to collect data related to Canberra suburbs. Unsupervised k-means clustering algorithms have been applied to two assembed datasets: one master dataset consisting of median house price, number of venues and closest distance to nearby towncenters; while the second dataset includes venue category information around each suburb.

As per the decision to choose which suburb to live in, certainly more factors should be considered, e.g. local schools, developing trend. They are not included in this study due to the limited time and the already lengthy contet, but definitely will be helpful to provide an even accurate and detailed suburb segmentation.

Complete source of this capstone project can be found on my Github.