Data Science Capstone

Project Report

Where to live in Canberra, Australia

Using the Goldilocks principle to select suburbs

Thomas Farrington

A report submitted in partial fulfilment of the requirements
for the Applied Data Science Capstone Project course in the
IBM Data Science specialization
Coursera

November 2019

Where to live in Canberra, Australia:
Using the Goldilocks principle to select suburbs

This report was typeset in 11pt Minion Pro using the KOMA-Script book class and XeLaTeX.

Contents

Figures

Chapter 1

Introduction

Moving to a new city, for whatever reason, is a stressful experience. Getting the choice of where to live right has a significant bearing on your happiness, not least because you are likely going to be stuck with that choice for seven to ten years. Get it wrong and your experiences will likely be negative and you will want to move earlier than you might have planned. Get it right, or as close to it a possible and the chance of an enjoyable and successful experience is greatly increased. How can we then use the power of data science to improve the chances of success?

Firstly we can determine the factors that are important to a potential buyer, which vary buyer to buyer, such as price, amenities, connectivity to neighbouring suburbs and city centre, availability and accessibility of local and wider transport links, work locations. Ranking the factors in terms of importance we can use the top three to five to produce a coarse profile for each suburb and buyer. By matching profiles to the buyer we can then eliminate many of the potential suburbs from consideration. For the remaining suburbs micro-profiles can be produced using the rest of the factors.

1.1 The Goldilocks principle

The Goldilocks principle[1] is a reference to the British fairy story *Goldilocks and the Three Bears*. It refers to things being "just right". So in finding the Goldilocks suburb we are looking for the suburb whose combination of factors are "just right", or optimal for the buyer. Taking distance to work as a factor, the buyer almost certainly does not want to be living in the same suburb as their employment location — in many cases in Canberra this would be almost impossible — but equally they do not want to have too long a commute journey. For example the buyer accepts that a commute is inevitable, but does not want to travel for more than 30 minutes for each journey. In this case we might assume the Goldilocks zone for this factor and buyer is 10–30 minutes. Extending this for all factors under consideration, we can determine suburbs that fall into the Goldilocks zone for each. Recommended suburbs are those which have the highest number of factors (not forgetting importance) in the Goldilocks zone. All factors in the Goldilocks zone is "just right" and has the highest recommendation.

[1] https://en.wikipedia.org/wiki/Goldilocks_principle

1.2 What factors could be considered?

The following, in no particular order, is a list of potential factors for consideration:

1. Median house price
2. 5 year median house price growth rate
3. Journey to work (time and available modes of transport)
4. Proximity and quality of schools
5. Shopping Centres
6. Medical Facilites
7. Fitness Centres
8. Other amenities (entertainment, bars, cafés, restaurants etc.)
9. Transport links
10. Demographics:
    a) Age groups
    b) Country of birth
    c) Religion
    d) Occupation

1.3 Objective

To recommend suburb(s) to house buyers in Canberra that match their requirements.

1.4 Project scope

Due to the compressed timescale this project has been completed in, only a few factors have been considered. These have also been ranked in order of importance as shown in Table 1.1

Table 1.1 Factors under consideration.

Importance	Factor
1	Median house price
2	5 year median house price growth rate
3	Local amenities (cafés, restaurants, gyms etc.)

Analysis has also been limited to the suburb centroid level. More refined spatial resolution is out of scope for this project. Furthermore this project does not provide absolute recommendations. There may be other factors not communicated by the buyer that would heavily influence the outcomes.

1.4.1 The notional house buyer

For the purposes of this report a notional house buyer has been used as an example who has the following requirements:

- Prepared to pay ± 20 % of current Canberra median house price
- Suburb must exhibit five-year growth stronger than the Canberra average
- Suburb amenity profile must have cafés, bars, restaurants and supermarkets in the top 10 venues per suburb

Chapter 2

Data: sources and use

The data sources required and used in this report are summarised in Table 2.1 and described in Section 2.1.

Table 2.1 Data sources.

Dataset	Source	Purpose
Canberra Suburbs	ACT Government website	Provides list of Suburb names and spatial location
Domain address API	Domain	Suburb IDs
Domain suburb performance API	Domain	House prices
4square venues API	4square	Shopping centre, bar, café locations

2.1 Data sources

2.1.1 Suburbs

The suburb data is available as a feature class (Division) in the Land Administration file geodatabase available daily from the ACT Goverment's ACTmapi data portal.[1] This feature class was pre-processed using an FME® workspace (Figure 2.1) to transform the dataset from spatial polygon features into a CSV file containing suburb name and centroid coordinates, in the LL-WGS84 coordinate reference system. A sample of the resulting CSV is shown in Figure 2.2

2.1.2 Domain: housing

This data can be retrieved using Domain's API Properties & Locations. Using a free account up to 500 calls per day can be made. The following endpoints are required for the project.

[1] This is released under Creative Commons 4.0 (CCBY v4.0).
ACT Divisions from http://www.actmapi.act.gov.au. © ACT Government.

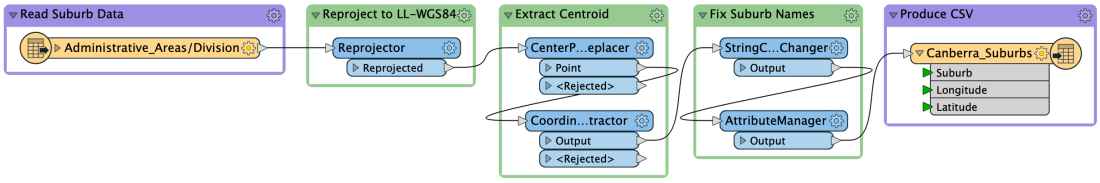


Figure 2.1 FME workspace to extract Canberra suburb centroid data.

```
Suburb,Longitude,Latitude
Spence,149.0654844130454,-35.20034096305264
Richardson,149.10856702191342,-35.42636212239583
Duffy,149.03346558587958,-35.33472138114033
Banks,149.10066158966652,-35.47186134297503
Lyneham,149.1325454800201,-35.24361988974169
Dickson,149.14023101670807,-35.25402425143912
Pearce,149.08377796577315,-35.363059622127054
Mckellar,149.07571762546584,-35.217500072557
Evatt,149.07122777970216,-35.211276478438776
O'Malley,149.11282571799043,-35.35380603260384
```

Figure 2.2 Canberra suburbs CSV sample.

2.1.2.1 addressLocators

The addressLocators API returns suburb, address and postcode IDs. This HTTP GET request will be used to add suburb IDs to the Canberra Suburbs data for retrieving further data.

Request template

https://api.domain.com.au/v1/addressLocators?searchLevel=Suburb&suburb=<suburb>&state=<state>

where <suburb> and <state> are parameters relating to each suburb.

Attributes required

- response ["addressComponents"] ["shortName"]
- response ["ids"] ["id"]

Sample HTTPS GET request

Get the Suburb ID for Parramatta, NSW:

https://api.domain.com.au/v1/addressLocators?searchLevel=Suburb&suburb=Parramatta&state=NSW

Sample response

```
[{
    "types": ["Suburb"],
    "addressComponents": [
        {"component": "Suburb", "shortName": "Parramatta"},
        {"component": "Postcode", "shortName": "2150"},
        {"component": "State", "shortName": "NSW"}],
    "ids": [{"level": "Suburb", "id": 35042}]
}]
```

2.1.2.2 suburbPerformanceStatistics

Retrieves property prices for the last n periods aggregated at the suburb level and uses the suburbId discovered from the first call.

Request template

https://api.domain.com.au/v1/suburbPerformanceStatistics?state=<state>&suburbId=<suburbid>&propertyCategory=house&chronologicalSpan=12&tPlusFrom=1&tPlusTo=5&values=MedianSoldPrice

Retrieves annual results for the last five years.

Attributes required

- response ["series"] ["seriesInfo"] ["year"]
- response ["series"] ["seriesInfo"] ["values"] ["medianSoldPrice"]

From the response, the 5 year growth rate, g_5, of median price, $\bar{p}$, can be calculated thus:

$$g_5 = \frac{\bar{p}_{2019} - \bar{p}_{2014}}{\bar{p}_{2014}} \times 100\,\% \tag{2.1}$$

2.1.3 FourSquare

The FourSquare API includes many endpoints grouped into categories. The most useful to this project is the venues group within which is the search endpoint which returns all venues within a set radius of a point. Here the suburb centroid coordinates will be used and 1 000 m as a starting search radius. The search will also be limited to certain categories of venue suitable for the analysis, i.e. cafés, restaurants. As with the Domain requests, the FourSquare API returns JSON by default.

Request template

https://api.foursquare.com/v2/venues/explore?&client_id=<client_id>&client_secret=<client_secret>&v=<version>&ll=<suburb_lat,>, <suburb_long>&radius=<radius>&limit=<limit>&categoryId=<list of category IDs>

Chapter 3

Methodology

3.1 House prices

Beginning with the `Canberra_Suburbs.csv` the data is augmented; firstly by adding the suburb id from the Domain properties API; and, secondly by the median house price in 2014 and 2019 agin from the Domain API. Lastly the percentage 5-year growth in median house price is calculated except where either median value is NaN or zero. In these cases the growth is set to NaN. Figure 4.1 shows the head of the Canberra dataframe.

	Suburb	Domain ID	Longitude	Latitude	Median Price 2014	Median Price 2019	5 yr Growth (%)
0	Acton	61	149.112771	-35.281319	NaN	NaN	NaN
1	Ainslie	71	149.148232	-35.263600	695000.0	975000.0	140.29
2	Amaroo	91	149.127417	-35.168831	544000.0	687000.0	126.29
3	Aranda	101	149.080925	-35.257771	662000.0	960000.0	145.02
4	Banks	131	149.100662	-35.471861	470000.0	485000.0	103.19
5	Barton	141	149.137112	-35.307921	NaN	NaN	NaN
6	Beard	6451	149.211188	-35.342198	NaN	NaN	NaN
7	Belconnen	171	149.068356	-35.235251	NaN	420000.0	NaN
8	Bonner	3041	149.142229	-35.157298	533000.0	661000.0	124.02
9	Bonython	191	149.077786	-35.434445	528000.0	643000.0	121.78
10	Braddon	221	149.136733	-35.271422	NaN	1250000.0	NaN
11	Bruce	241	149.093361	-35.245587	730000.0	750000.0	102.74
12	Calwell	271	149.106573	-35.442859	465000.0	589000.0	126.67
13	Campbell	291	149.154939	-35.288914	843000.0	1255000.0	148.87
14	Canberra Airport	6461	149.194880	-35.303439	NaN	NaN	NaN
15	Capital Hill	321	149.124435	-35.308179	NaN	NaN	NaN
16	Casey	3291	149.094391	-35.166508	433000.0	600000.0	138.57
17	Chapman	351	149.036772	-35.354587	715000.0	908000.0	126.99
18	Charnwood	361	149.033784	-35.200478	402000.0	478000.0	118.91
19	Chifley	371	149.078003	-35.353877	585000.0	791000.0	135.21

Figure 3.1 Canberra suburb house price dataframe (head).

This allows the analysis of house prices per suburb to drive our decision-making process.

3.2 Suburb clustering

To determine the characteristic amenity profile of each suburb, K-Means clustering was chosen. This method is ideal for grouping data into clusters based on the similarity in their features.

Beginning with the resulting dataframe from Section 3.1, the centroid Longitude and Latitude values were used to query the FourSquare venues API. Here an http request was made for each suburb requesting up to 100 venues of all categories within 1 km of the suburb centroid. The (Figure 3.2).

	Suburb	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Acton	-35.281319	149.112771	Australian National Botanic Gardens	-35.278050	149.109428	Botanical Garden
1	Acton	-35.281319	149.112771	Llewellyn Hall	-35.280604	149.123442	Concert Hall
2	Acton	-35.281319	149.112771	National Film & Sound Archive	-35.283131	149.121143	Museum
3	Acton	-35.281319	149.112771	Monster Kitchen and Bar	-35.285122	149.122547	Hotel Bar
4	Acton	-35.281319	149.112771	BrodDogs	-35.278428	149.122443	Food Truck

Figure 3.2 Canberra venues dataframe.

This dataframe was then grouped by Suburb to produce the "Top 10" venue categories per Surburb (Figure 3.3)

	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	10th Most Common Venue
0	Acton	Café	Food Truck	Bakery	Plaza	Museum	Botanical Garden	Coffee Shop	Concert Hall	Sandwich Place	Hotel
1	Ainslie	Hotel	Bakery	Fish & Chips Shop	Shopping Plaza	Café	Business Service	Sports Club	Grocery Store	Pub	Australian Restaurant
2	Amaroo	Supermarket	Indian Restaurant	Shopping Plaza	Lake	Grocery Store	Yoga Studio	Filipino Restaurant	Fountain	Football Stadium	Food Truck
3	Aranda	Café	Middle Eastern Restaurant	Chinese Restaurant	Thrift / Vintage Store	Nature Preserve	Yoga Studio	Fried Chicken Joint	Fountain	Football Stadium	Food Truck
4	Banks	Pizza Place	Construction & Landscaping	Gym / Fitness Center	Grocery Store	Yoga Studio	Filipino Restaurant	Fountain	Football Stadium	Food Truck	Food Court

Figure 3.3 Venue categories per Suburb.

After one-hot encoding the unlabelled "Top 10" data, SciKit-Learn's K-means clustering method was used, initially for five clusters and then for ten. The resulting labels are then attached to each suburb in the dataframe.

Chapter 4

Results and discussion

4.1 House prices

	Suburb	Domain ID	Longitude	Latitude	Median Price 2014	Median Price 2019	5 yr Growth (%)
0	Acton	61	149.112771	-35.281319	NaN	NaN	NaN
1	Ainslie	71	149.148232	-35.263600	695000.0	975000.0	140.29
2	Amaroo	91	149.127417	-35.168831	544000.0	687000.0	126.29
3	Aranda	101	149.080925	-35.257771	662000.0	960000.0	145.02
4	Banks	131	149.100662	-35.471861	470000.0	485000.0	103.19
5	Barton	141	149.137112	-35.307921	NaN	NaN	NaN
6	Beard	6451	149.211188	-35.342198	NaN	NaN	NaN
7	Belconnen	171	149.068356	-35.235251	NaN	420000.0	NaN
8	Bonner	3041	149.142229	-35.157298	533000.0	661000.0	124.02
9	Bonython	191	149.077786	-35.434445	528000.0	643000.0	121.78
10	Braddon	221	149.136733	-35.271422	NaN	1250000.0	NaN
11	Bruce	241	149.093361	-35.245587	730000.0	750000.0	102.74
12	Calwell	271	149.106573	-35.442859	465000.0	589000.0	126.67
13	Campbell	291	149.154939	-35.288914	843000.0	1255000.0	148.87
14	Canberra Airport	6461	149.194880	-35.303439	NaN	NaN	NaN
15	Capital Hill	321	149.124435	-35.308179	NaN	NaN	NaN
16	Casey	3291	149.094391	-35.166508	433000.0	600000.0	138.57
17	Chapman	351	149.036772	-35.354587	715000.0	908000.0	126.99
18	Charnwood	361	149.033784	-35.200478	402000.0	478000.0	118.91
19	Chifley	371	149.078003	-35.353877	585000.0	791000.0	135.21

Figure 4.1 Canberra suburb house price dataframe (head).

It is expected that for some suburbs the median house prices returned will be NaN. This is likely due to the suburb's properties being primarily commercial rather than residential. Furthermore, due to the recent suburb developments in Canberra, there are several new suburbs, which may also account for missing 2014 data.

Taking the summary statistics for the house price data (Figure 4.2) it can be seen that the average 2019 median house price is $ 787 918 and in 2014 it was $ 605 058. Canberra's average five-year growth was 128.64 %.

	Domain ID	Longitude	Latitude	Median Price 2014	Median Price 2019	5 yr Growth (%)
count	116.000000	116.000000	116.000000	8.700000e+01	9.600000e+01	86.000000
mean	1575.568966	149.097883	-35.296502	6.050575e+05	7.879167e+05	128.451047
std	1905.070759	0.047611	0.083825	2.042642e+05	2.865158e+05	10.593759
min	61.000000	148.925092	-35.508979	4.020000e+05	4.200000e+05	101.540000
25%	618.500000	149.064579	-35.352567	4.815000e+05	6.000000e+05	121.817500
50%	1036.000000	149.100564	-35.297153	5.440000e+05	7.140000e+05	128.640000
75%	1531.000000	149.133244	-35.226085	6.550000e+05	8.697500e+05	134.620000
max	7841.000000	149.227164	-35.153357	1.900000e+06	2.210000e+06	154.390000

Figure 4.2 House price summary statistics.

The house buyer wants to pay 80–120 % of the Canberra median and buy in a suburb that has better than average price growth. The Canberra house price dataframe can be filtered and ordered (Figure 4.3) to show that for the house price factors the house buyer should be looking, in the first instance, at the following suburbs:

1. Wright
2. Cook
3. Fraser
4. Mawson
5. Narrabundah

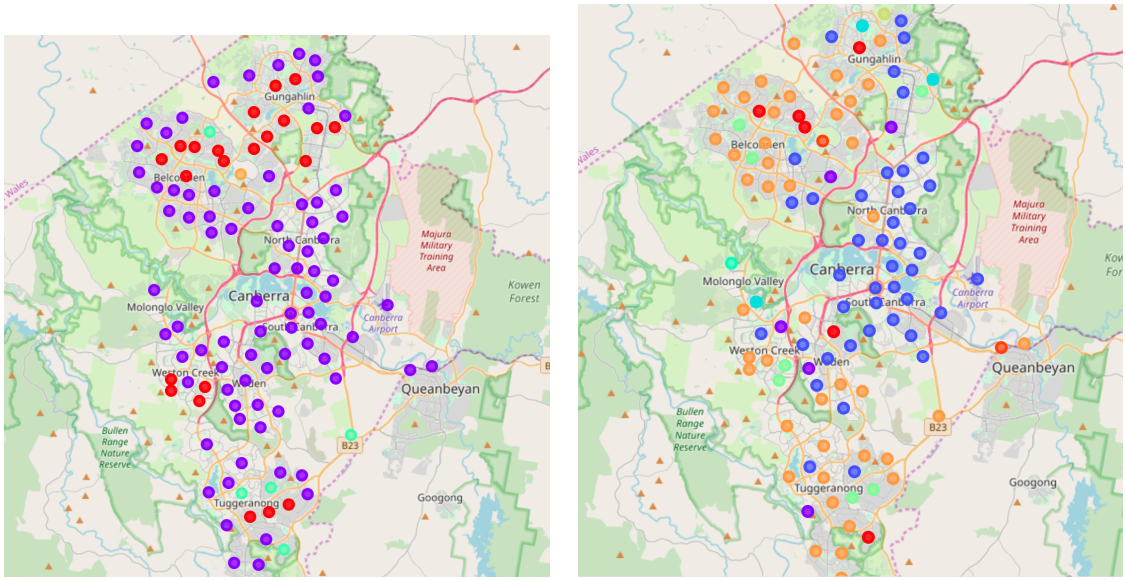
	Suburb	Domain ID	Longitude	Latitude	Median Price 2014	Median Price 2019	5 yr Growth (%)
114	Wright	7121	149.033242	-35.320674	580000.0	880000.0	151.72
23	Cook	441	149.066321	-35.260416	529000.0	765000.0	144.61
42	Fraser	661	149.045276	-35.191903	500000.0	723000.0	144.60
74	Mawson	1141	149.100467	-35.363007	545000.0	760000.0	139.45
81	Narrabundah	1211	149.148882	-35.335116	660000.0	920000.0	139.39
31	Duffy	551	149.033466	-35.334721	545000.0	759000.0	139.27
87	Oxley	1331	149.078932	-35.409159	476000.0	655000.0	137.61
41	Franklin	2301	149.143494	-35.197892	530000.0	725000.0	136.79
64	Kaleen	951	149.108439	-35.226296	530000.0	725000.0	136.79
52	Hackett	761	149.162325	-35.250558	663000.0	899000.0	135.60

Figure 4.3 Top 10 suburbs by house price factors.

4.2 Suburb clustering

From the data returned for each suburb by the FourSquare API, suburbs can be clustered into groups the share similar characteristics. Where no data or corrupted data was returned (unfortunately a frequent occurrence with FourSquare) the suburb was excluded from the final analysis.

Displaying the results of the K-Means clustering method spatially means that it is much easier to see the cluster membership and distribution (Figures 4.4a and 4.4b).



(a) $k = 5$. (b) $k = 10$.

Figure 4.4 Suburb clustering. Map data by © OpenStreetMap, under ODbL.

Given the similarity of many of the suburbs, a low k is not providing significant distinction between suburbs. As can be seen in Figure 4.4b raising k to 10 allows suburbs to be differentiated. By matching the cluster profile to the house buyer's preferences a number of suburbs can be recommended.

Conflating the house price data with the suburb cluster data leads to a more refined recommendation. Joining the house price filtered dataframe with the $k = 10$ clustering results produces the dataframe and map in Figures 4.5 and 4.6 respectively. As the house buyer prefers suburbs with the profile of cluster 2, this analysis recommends the house buyer looks at the following suburbs first as they fit their requirements the best:

1. Cook
2. Narranbundah
3. Oxley
4. Franklin
5. Hackett

	Suburb	Domain ID	Longitude	Latitude	Median Price 2014	Median Price 2019	5 yr Growth (%)	Cluster Labels
0	Wright	7121	149.033242	-35.320674	580000.0	880000.0	151.72	8
1	Cook	441	149.066321	-35.260416	529000.0	765000.0	144.61	2
2	Fraser	661	149.045276	-35.191903	500000.0	723000.0	144.60	8
3	Mawson	1141	149.100467	-35.363007	545000.0	760000.0	139.45	8
4	Narrabundah	1211	149.148882	-35.335116	660000.0	920000.0	139.39	2
5	Oxley	1331	149.078932	-35.409159	476000.0	655000.0	137.61	2
6	Franklin	2301	149.143494	-35.197892	530000.0	725000.0	136.79	2
7	Kaleen	951	149.108439	-35.226296	530000.0	725000.0	136.79	8
8	Hackett	761	149.162325	-35.250558	663000.0	899000.0	135.60	2
9	Waramanga	1751	149.061993	-35.352154	510000.0	691000.0	135.49	6

Figure 4.5 The clustered suburbs by house price factors.

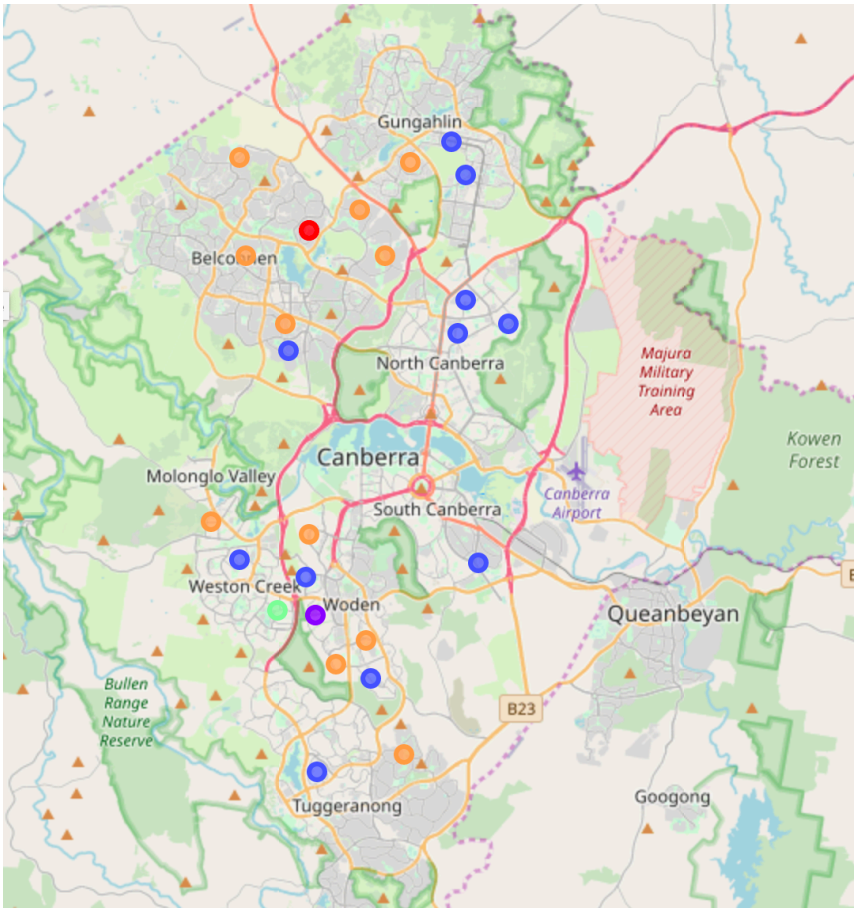


Figure 4.6 Clustered suburbs by house price factors (Cluster key — 0 : Red; 1 : Purple; 2 : Blue; 8 : Orange).

Chapter 5

Conclusions

For the notional house buyer used in this analysis, there are a number of clear recommendations to be made. However, given that only a few factors were examined the list still remains quite large and may contain obvious errors if further information about the house buyers requirements were known. Taking just the house price factors the list of suburbs is still about 25 % of the total number of suburbs and in combination with the very broad suburb venue profiling carried out and that some suburbs did not return a profile form the FourSquare API and are therefore missing, the list is still quite large.

In the future this analysis should be extended to use more reliable venue data to build more accurate suburb profiles. More detailed questions should then be asked of the house buyer. Furthermore more factors should be used, such as commuting time/distance, distance to nearest large shopping/town centre, e.g. City, Belconnen, Woden.

Putting more useful information into the analysis should then permit the recommendation system to target the 'best' suburb for the house buyer more accurately.

In this report we find that the suburb that is 'just right' for the notional house buyer is Cook.