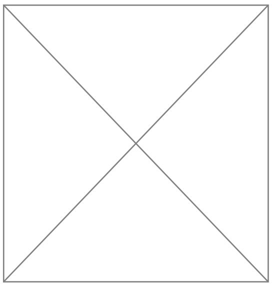
Data Science Capstone

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 11 pt Minion Pro using the KOMA-Script book class and XeLaTeX.

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 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	Shopping Centres
4	Other amenities (cafés, restaurants)

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.

Chapter 2

Data: sources and use

To solve the questions raised in Chapter 1, access to several datasets must be obtained.

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 preprocessed 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 and demographics

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.

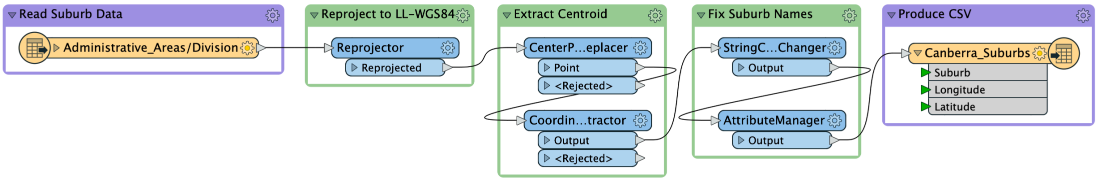
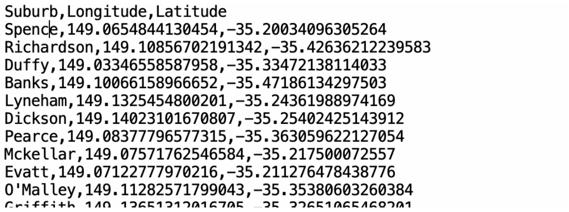


Figure 2.1 FME workspace to extract Canberra suburb centroid data.



```
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
Griffith,149.13651312016705,-35.32651065468201
```

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:

$$g_5 = \frac{\bar{p}_{2019} - \bar{p}_{2015}}{\bar{p}_{2015}} \times 100\,\% \tag{2.1}$$

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 500 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>