Choosing a suburb in Canberra, Australia

Using data science and the Goldilocks principle to find suburbs that best fit user's criteria

Tom Farrington
Coursera
IBM Data Science Capstone Project

Introduction

- Choosing a place to live is a difficult decision to get right
- Particularly if the buyer is not familiar with the city
- Using data science the choice of suburb can be narrowed significantly to get the ones that satisfy the Golidlocks' principle, i.e. those that are 'just right'
- This may not be the one that is exceptional in one factor …
- but the one that is 'good' in all factors.

The notional house buyer

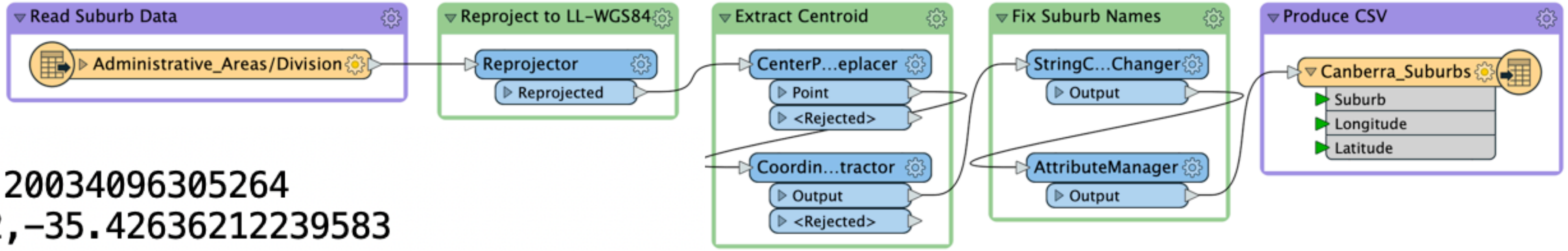
To test the analytical process we use this test case.

A house buyer would like to buy where:

- They can pay ±20% of Canberra median house price
- The suburb 5-yr growth should be stronger than Canberra average
- The suburb must have cafés, restaurants, bars and supermarkets in its top 10 list of venue types

Canberra suburbs

- Australia's capital city, Canberra has 116 suburbs
- Some are primarily commercial, others are residential
- Suburb data collected from ACT Government and processed into a CSV file containing suburb centroid Long/Lat and suburb name via FME workspace




```
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
```

Other Data

- Domain API – addresslocators & suburbPerformanceStatistics
  - suburb ID
  - house price data

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

- FourSquare API – venues/explore
  - venues within radius of suburb centroid

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>

Methodology

1. Get data

For each suburb fetch:

- Domain API
  - Suburb ID
  - 2014 Median House Price
  - 2019 Median House Price
- FourSquare API
  - Up to 100 venues of all categories within 1km of suburb centroid

Methodology

2. Process housing data

- Calculate 5-yr growth
- Select suburbs where 2019 median in range and strong growth

	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

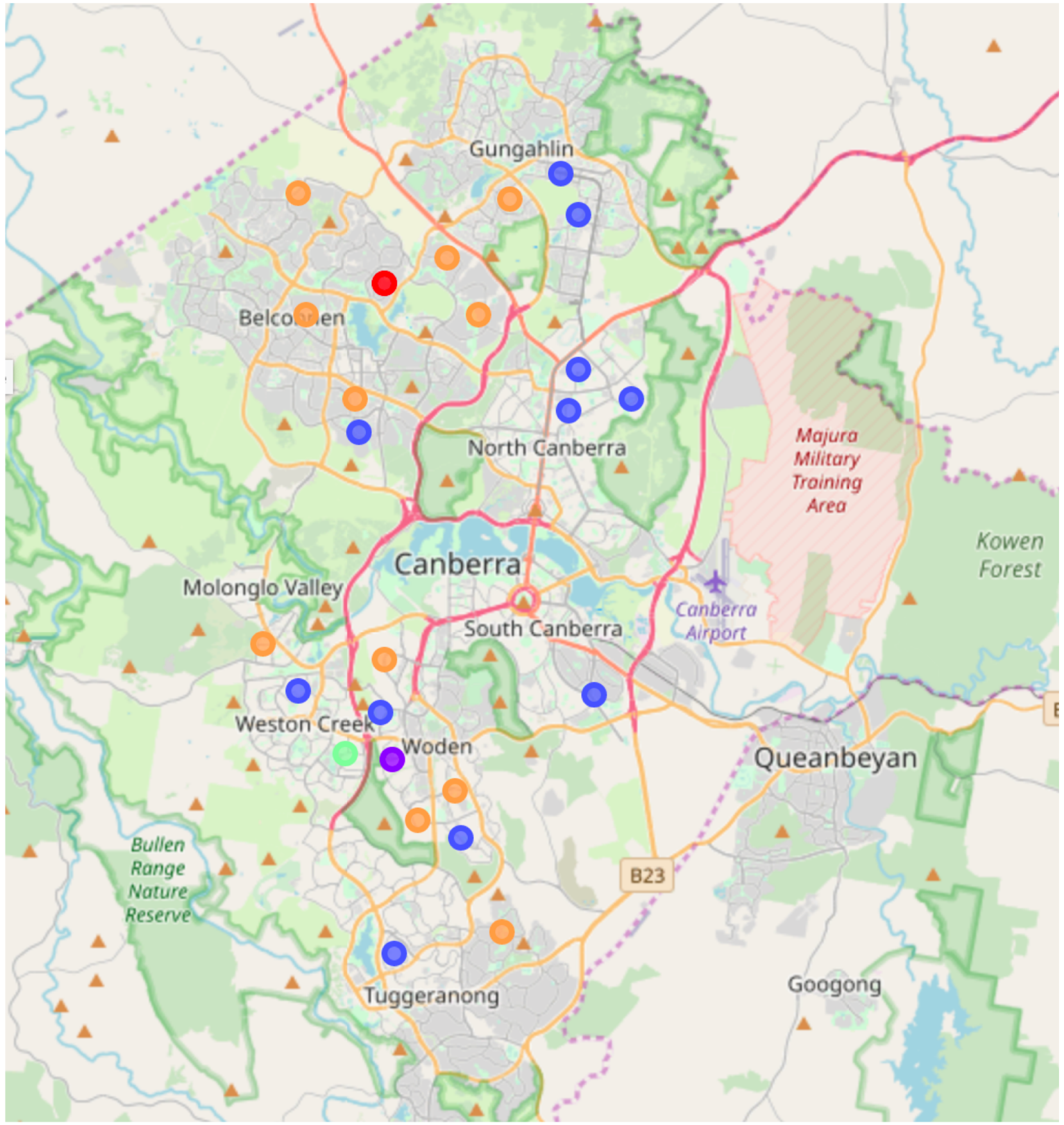
Methodology

3. Process suburb venue data

- Summarise venues by category and keep top 10.
- Use K-Means clustering (k = 10) to find similar suburbs
- Determine which cluster(s) fit buyers profile
- Join to filtered data from step 2

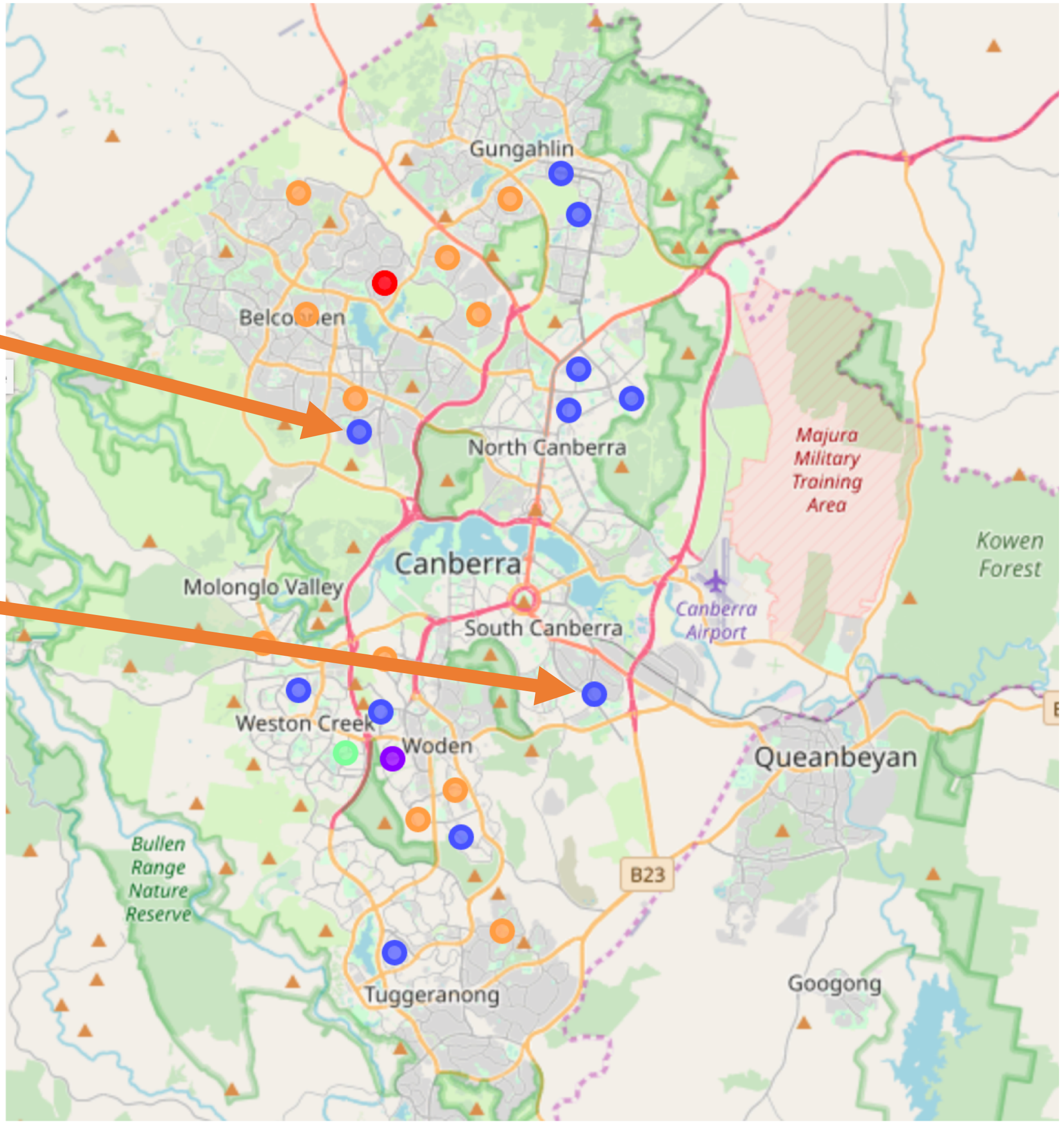
Results

	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



Results

	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



Further work

- Add further factors to decision-making process:
  - Commuting distance/time
  - Distance to nearest town centre
  - Suburb demographics
  - Location and quality of nearby schools
- Additional data sources:
  - Australian Bureau of Statistics (census data)
  - Better quality venue data
  - Travel time data (e.g. Google, HERE)

Questions

?