Applied Data Science Capstone

Andrew Hall

January 2019

1. Introduction

1.1 Background

In the last few months my life has changed dramatically. For many years I worked as a software architect in the finance sector in London but a year ago my wife was offered a contract in Sydney Australia which after some deliberation she decided to accept. Initially she moved to Sydney but my eleven year old daughter and I stayed in London so she could finish her school year. I'm happy to say that our family is now reunited in Australia and I am now a home schooling dad and part time Coursera junkie.

We have now been in Australia for four months and looking to stay longer than originally planned. I have been lucky enough to receive a couple of job offers which I am considering but this means finding somewhere more permanent to live and a good school for our daughter as she transitions from primary to secondary education.

1.2 Problem

The main problem is that we are still relative newcomers to Sydney, We have enjoyed many of the tourist attractions and love the outdoors beach life but we don't have a very good understanding of what different neighbourhoods are like in Sydney, how much it costs to rent in each neighbourhood and what school options are available. We are currently living short term in Manly which we all like for the beautiful beach and restaurants but it would be an expensive long term solution and possibly outside our reach financially.

Whilst there are many good sources of advice on the web these typically reflect a range of opinions rather than facts and as I was completing this course I thought it could be interesting to apply data science principles to identify areas of Sydney most geared to an active, outdoors family life. This might help us make a more informed decision as we begin the next chapter of our journey.

Our key considerations are:
- Availability of good international or secondary girls schools
- Rental prices for a three bedroom apartment
- Low crime rates
- Family friendly amenities such as beach, parks and cinemas
- Commute time to the central business district

The object of this analysis is therefore to group and compare different neighbourhoods in Sydney, using publicly available data to determine where best to focus our property and schooling search efforts.

The code underpinning this document may be found in my GitHub repo at: https://github.com/ah903/cousera-capstone

1.3 Interest

Whilst this is very much a personal project for our family, many families relocate for work and other reasons each year. I would expect the approach, data and research outlined in this paper to be directly applicable to others moving to Sydney.

To give a sense of scale, according to the Australian Government Department of Home Affairs Temporary Resident Skilled Workers Report [1], 10470 people applied for a temporary skilled workers visa in 2019 with approximately 45% of all sponsorships originating in New South Wales. Whilst the data is not granular to city level the population of Sydney accounts for approximately 69% of the population of New South Wales [2]. so it is reasonable to assume the majority of skilled worker visas originate from companies based in and around Sydney.

The analysis in this paper could therefore be applicable to several thousand people each year, particularly if data sources were kept up to date with the latest information.

2. Data Acquisition and Cleaning

Data Sources

To gain some deeper insights into the neighbourhoods of Sydney it will be necessary to combine data from a number of different sources. This is because different government agencies are typically responsible for the publication of data about house rental prices, crime figures and schools. However locality is the common factor which may be distilled down to latitude and longitude or more usefully postal code as a common key to merge data from different sources as not all sources are geocoded with latitude and longitude.

Postal codes in Australia are four digit numbers and are unique to a state but not unique across states. For the most part this distinction is immaterial as the analysis focuses on New South Wales but it would be an important consideration if the analysis were attempted at a national scale. Additionally a single postal code may span several suburbs or neighbourhoods. Some of the datasets used in this analysis report at the more granular postcode and suburb level whilst others are less granular and report simply at the postcode level.

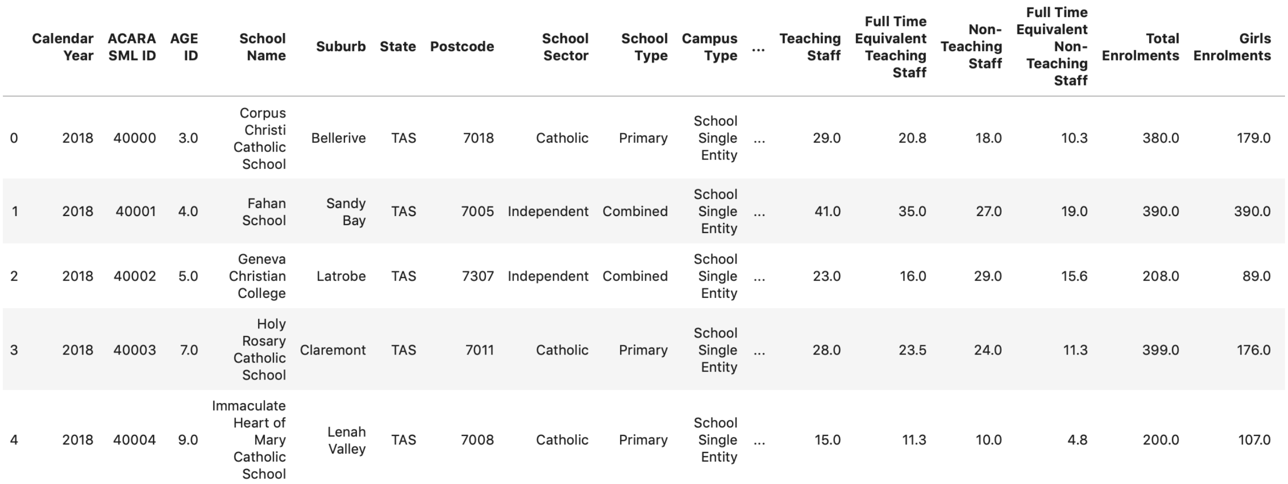
The datasets identified for use in this work described below and sources listed in the references section. Where possible I have used official government data as these are more likely to undergone a degree of quality checking prior to publication. These are complimented by the Foursquare API to obtain additional social data about neighbourhoods.

2.1 Schools Dataset
Schools data is obtained from the Australian Curriculum Assessment and Reporting Authority (ACARA) in two separate datasets. Firstly a school profile dataset [3] that describes the type of school, the socio-economic backgrounds of its pupils, student enrolments and teaching staff and secondly a school locations dataset [4] that includes the address, postcode and geolocation of each school. The big advantage of using these datasets is that firstly they may be joined on a common school identifier and secondly they include both public and private schools so give a full picture of the educational options.

The schools profile and location datasets are summarised in the images overleaf. These datasets are comprehensive, current and well-curated because they are used as sources for several government education websites including the myschool education portal for example [5]. They require little cleaning other than the removal of fields that will not be useful for this assignment.

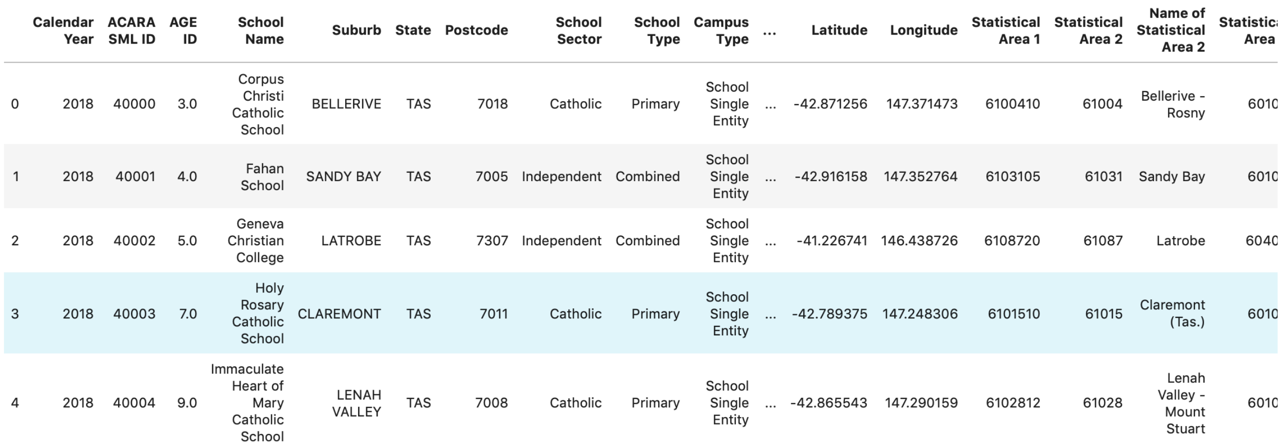
The dataset will be joined and used to create visualisations of the number and location of schools in each postcode filtered by primary co-educational or girls schools.

2.1.1. School Profile Dataset



	Calendar Year	ACARA SML ID	AGE ID	School Name	Suburb	State	Postcode	School Sector	School Type	Campus Type	...	Teaching Staff	Full Time Equivalent Teaching Staff	Non-Teaching Staff	Full Time Equivalent Non-Teaching Staff	Total Enrolments	Girls Enrolments
0	2018	40000	3.0	Corpus Christi Catholic School	Bellerive	TAS	7018	Catholic	Primary	School Single Entity	...	29.0	20.8	18.0	10.3	380.0	179.0
1	2018	40001	4.0	Fahan School	Sandy Bay	TAS	7005	Independent	Combined	School Single Entity	...	41.0	35.0	27.0	19.0	390.0	390.0
2	2018	40002	5.0	Geneva Christian College	Latrobe	TAS	7307	Independent	Combined	School Single Entity	...	23.0	16.0	29.0	15.6	208.0	89.0
3	2018	40003	7.0	Holy Rosary Catholic School	Claremont	TAS	7011	Catholic	Primary	School Single Entity	...	28.0	23.5	24.0	11.3	399.0	176.0
4	2018	40004	9.0	Immaculate Heart of Mary Catholic School	Lenah Valley	TAS	7008	Catholic	Primary	School Single Entity	...	15.0	11.3	10.0	4.8	200.0	107.0

2.1.2. School Location Dataset



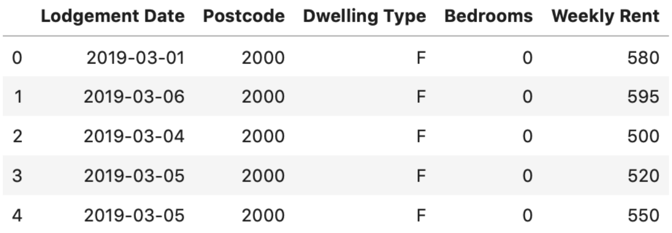
	Calendar Year	ACARA SML ID	AGE ID	School Name	Suburb	State	Postcode	School Sector	School Type	Campus Type	...	Latitude	Longitude	Statistical Area 1	Statistical Area 2	Name of Statistical Area 2	Statistica Area
0	2018	40000	3.0	Corpus Christi Catholic School	BELLERIVE	TAS	7018	Catholic	Primary	School Single Entity	...	-42.871256	147.371473	6100410	61004	Bellerive - Rosny	6010
1	2018	40001	4.0	Fahan School	SANDY BAY	TAS	7005	Independent	Combined	School Single Entity	...	-42.916158	147.352764	6103105	61031	Sandy Bay	6010
2	2018	40002	5.0	Geneva Christian College	LATROBE	TAS	7307	Independent	Combined	School Single Entity	...	-41.226741	146.438726	6108720	61087	Latrobe	6040
3	2018	40003	7.0	Holy Rosary Catholic School	CLAREMONT	TAS	7011	Catholic	Primary	School Single Entity	...	-42.789375	147.248306	6101510	61015	Claremont (Tas.)	6010
4	2018	40004	9.0	Immaculate Heart of Mary Catholic School	LENAH VALLEY	TAS	7008	Catholic	Primary	School Single Entity	...	-42.865543	147.290159	6102812	61028	Lenah Valley - Mount Stuart	6010

2.1.3 School Performance Measures

School performance in Australia is also measured on a number of key metrics known as the National Assessment Program for Literacy and Numeracy (NAPLAN). This would have been a useful complimentary dataset for my analysis but at the present time this data is not available in a packaged form although it is possible to search for results on the web school by school. Inclusion of NAPLAN ratings in this analysis is deferred as a possible future enhancement.

2.2 Rental Dataset

Rental agreement in Australia require the tenant to pay a bond known as a lodgement as a contingency against damage to the property. Typically a lodgement equates to 4 weeks rent. Lodgements are not held by the landlord but by a branch the New South Wales state government who also publish data each year on every lodgement which includes the weekly rental price. The dataset is available from New South Wales Government Rental Bond Lodgement Data website [6] but is recorded only at post code granularity without an accompanying suburb breakdown.

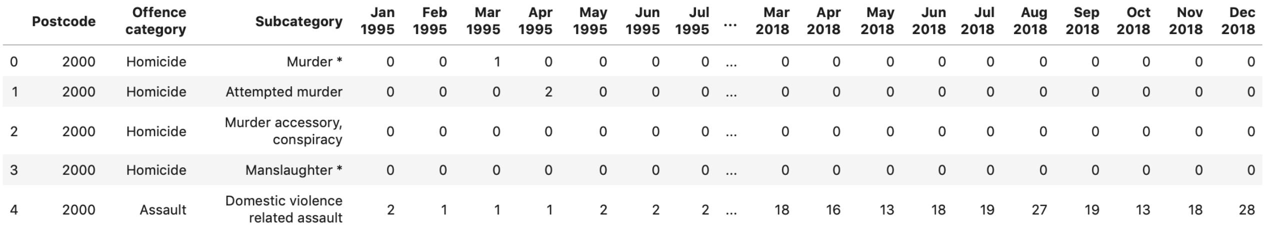


	Lodgement Date	Postcode	Dwelling Type	Bedrooms	Weekly Rent
0	2019-03-01	2000	F	0	580
1	2019-03-06	2000	F	0	595
2	2019-03-04	2000	F	0	500
3	2019-03-05	2000	F	0	520
4	2019-03-05	2000	F	0	550

The data also requires some degree of cleaning to remove uncategorised entires, rows with missing rental values and properties listed with zero rooms which essentially equate to garage or parking space rentals. A sample of the dataset is shown below.

2.3 Crime Dataset

Data on reported crimes across New South Wales is collated and published by the Bureau of Crime Statistics and Research (BOSCAR) which describes a monthly time series of crimes by postcode from 1995 to 2018 [7]. This is quite a large dataset and is well suited to trend analysis across time. However for the purposes of this assignment the data will need to be aggregated and summarised because I am predominantly interested in the absolute figures for the most recent year in the dataset and the postal areas that correspond to inner Sydney rather than the outlying areas of New South Wales. Summarising the data using measures such as mean or median may also offer a useful insight



	Postcode	Offence category	Subcategory	Jan 1995	Feb 1995	Mar 1995	Apr 1995	May 1995	Jun 1995	Jul 1995	...	Mar 2018	Apr 2018	May 2018	Jun 2018	Jul 2018	Aug 2018	Sep 2018	Oct 2018	Nov 2018	Dec 2018
0	2000	Homicide	Murder *	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	2000	Homicide	Attempted murder	0	0	0	2	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	2000	Homicide	Murder accessory, conspiracy	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	2000	Homicide	Manslaughter *	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	2000	Assault	Domestic violence related assault	2	1	1	1	2	2	2	...	18	16	13	18	19	27	19	13	18	28

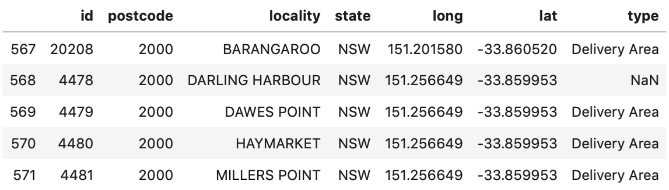
The main data manipulations required to work with this dataset will be firstly to aggregate the different categories of crime to obtain values at a postcode level and secondly to rollup months to years. It may be useful to calculate further metrics such as mean reported crime figures across the entire time period. This dataset is rich and we could do many other interesting forms of analysis such as introducing a breakdown of serious crime versus misdemeanours and public order offences.

To visualise the data on a map we will need to geocode each postal area so need a source of reference data.

2.4 Postal districts of New South Wales

From the previous data samples it should be clear that we are dealing with some datasets that are geocoded and granular to a suburb level and other datasets that are only reported at a postcode level and do not include geocoding. The common denominator across all the datasets is the postcode and this therefore provides a common key to join different datasets. Ideally we require a reference dataset of postal regions

Unfortunately the Australian postal service has recently decided to make postcode data a commercial arm of its business which means paying for access via either an API or a download. There are a number of free data sources available on the web that vary in terms of coverage and

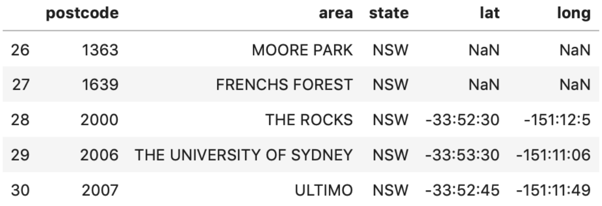


	id	postcode	locality	state	long	lat	type
567	20208	2000	BARANGAROO	NSW	151.201580	-33.860520	Delivery Area
568	4478	2000	DARLING HARBOUR	NSW	151.256649	-33.859953	NaN
569	4479	2000	DAWES POINT	NSW	151.256649	-33.859953	Delivery Area
570	4480	2000	HAYMARKET	NSW	151.256649	-33.859953	Delivery Area
571	4481	2000	MILLERS POINT	NSW	151.256649	-33.859953	Delivery Area

currency. This project has selected two sources of data that geocode postcodes by suburb [8] and postcodes without suburb [9] to match up with the other datasets.

2.4.1 Geocoded Postcodes By Suburb

2.4.1 Geocoded Postcodes



	postcode	area	state	lat	long
26	1363	MOORE PARK	NSW	NaN	NaN
27	1639	FRENCHS FOREST	NSW	NaN	NaN
28	2000	THE ROCKS	NSW	-33:52:30	-151:12:5
29	2006	THE UNIVERSITY OF SYDNEY	NSW	-33:53:30	-151:11:06
30	2007	ULTIMO	NSW	-33:52:45	-151:11:49

In this latter dataset some conversion of longitude and latitude data will be required to convert the data demo degrees, minutes and seconds to decimal which will be both easier to work with and consistent with the geocoded suburb and postcode dataset.

In addition to these two postcode reference datasets we will also use a geojson file that defines postcode areas [10] as a series of latitude longitude polygons. Using a geojson defined mapping layer will enable a more accurate visualisation of our data sources through heat maps and choropleth representations on postcode boundaries.

2.5 Neighbourhood Dataset
In addition to the factual and geographical data sources listed above we will also use socially crowdsourced data about different neighbourhoods to understand more about the key amenities and character of different parts of Sydney. As in earlier parts of the course we will use the Foursquare API to interrogate a large and rich dataset of information to identify and categorise family friendly neighbourhood with good access to beach, park and cinema facilities.

3. Approach

In the previous section I described several sources of data that could help build a rounded, quantifiable view of the areas of Sydney that would suit our family. The datasets are all keyed on New South Wales postcode and provide information about schools, crime, rental availability and prices augmented with social data.

This is a lot of data to work with which in turn implies a lot of data cleansing and standardisation work, some of which has been identified. The general approach to understanding the data will be to use Folium to perform a lot of the initial data exploration and analysis looking at each of these individual features in turn. By applying a geocoding to the data and using a postcode geojson file it will be possible to create heat map visualisations of crime rates, primary school distribution and rental prices in different parts of Sydney. Furthermore we should be able to summarise these datasets to create numeric rankings.

However we also want to incorporate FourSquare data by using the API to query each postcode the number of venues that cater to the family activities. This will help to build a profile for each neighbourhood. By merging all of these sources together we can create a single data frame of features that describes individual amenities, rental price, crime rates and schools.

This dataset will be used to classify neighbourhoods using a simple K-means algorithm to group suburbs according to how well they fit our our families needs across a range of dimensions. To achieve best results it’s likely that some data standardisation will need to be performed so large values in the dataset do not crowd out the influence of small values for example, number of crimes versus number of schools.

The reason for choosing a k-means clustering algorithm is two fold. Firstly our data is unlabelled so we don't at this stage know what the groups are, although we could potentially think about potential labels along the lines of strongly matched, matched and not suitable giving three clusters. However we will need to explore this further as part of the analysis. The second reason for choosing the k-means algorithm is that it is simple to understand and computationally fast to perform meaning it is low cost to experiment with the number of clusters and randomised start positions for each.

Feature Selection

The features that will be used from our data are:

- Postcode
- Number of primary schools that are secular, co-educational or girls only. It would be interesting to explore further the distribution between government and private schools within this data
- Average Rental Price (3 bedroom accommodation)
- Average reported crime incidents per year, or the number reported in the most recent year
- Number of beaches, surfing or surf club facilities in the postcode
- Number of cinemas in the postcode
- Number of restaurants in the postcode
- Number of parks in the postcode

3. Exploratory Data Analysis

To add in week 5

4. Predictive Modelling

To add in week 5

Classification Models

To add in week 5

Performance of different classification models

To add in week 5

5. Conclusions

To add in week 5

6. Future Direction

To add in week 5

Glossary

ACARA	Australian Curriculum Assessment and Reporting Authority
BOCSAR	Bureau of Crime Statistics and Research
NAPLAN	National Assessment Programme for Literacy and Numeracy

References

[1] Australian Government Department of Home Affairs (2019) Temporary Resident Skilled Report as at 300919 Available at: https://www.homeaffairs.gov.au/research-and-stats/files/temp-res-skilled-quarterly-report-30092019.pdf

[2] Australian Bureau of Statistics (2019) Australian Demographic Statistics, Jun 2019 Available at: https://www.abs.gov.au/AUSSTATS/abs@.nsf/mf/3101.0

[3] School Profile (2018) Australian Curriculum Assessment and Reporting Authority, Available at: https://www.acara.edu.au/docs/default-source/default-document-library/school-profile-2018.xlsx?sfvrsn=0

[4] School Locations (2018) Australian Curriculum Assessment and Reporting Authority. Available at: https://www.acara.edu.au/docs/default-source/default-document-library/school-locations-20189cf512404c94637ead88ff00003e0139.xlsx?sfvrsn=0

[5] MySchool Education Portal (2019) Australian Curriculum Assessment and Reporting Authorit. Available at: https://myschool.edu.au/

[6] New South Wales Government, Rental Bond Lodgement Data (2019) Available at: https://www.fairtrading.nsw.gov.au/about-fair-trading/data-and-statistics/rental-bond-data

[7] New South Wales Bureau of Crime Statistics and Research (2019), Recorded crime by offence (2019), Available at: https://www.bocsar.nsw.gov.au/Pages/bocsar_datasets/Datasets-.aspx

[8] Proctor, M. (2020) Download Australian Post Codes Available at: https://www.matthewproctor.com/australian_postcodes

[9] Radioactive Networks (2020), Australian Latitude and Longitude from Postcode. Available at: http://www.radio-active.net.au/web3/APRS/Resources/AusLatLon

[10] Singham, L. (2019)GeoJson file for Australia Available at: https://github.com/ucg8j/aus_cloropleth/blob/master/au-postcodes-Visvalingam-0.1-density.js