How to select neighbourhoods for your first rental property in Sydney

Introduction

Description & Discussion of the Background

Sydney, capital of New South Wales and one of Australia's largest cities covers 12,367.7 sq km land and is made up of 35 local councils. As of June 2019, Sydney has a population of 5.73 million and spread into 658 suburbs. The metropolitan area of Sydney sprawl about 70 km to the west, 40 km to the north, and 60 to the south. Every year we have new migrater from countries all over the world. Over the last 8 years, Sydney have a very strong population growth rate which reach's it peaks in 2013 of 10.88% with a total 1.7 million growth of population since year 2011 (4.029 Million). [1] The strong population growth causes the high demand of residential property needs. According to afr.com financial review, rental vacancies in Sydney dropped to 2.9 percent which is lowest we have seen in the last 2 years. [2]

I have been living in Sydney almost 13 years now, with all the rental and investment experience I have gain over the years. I'm now consider myself quite familiar of the most popular neighbourhood in Sydney and know which suburb is might be a good choice to live at different life stages. Every now and then, I have some new friends decide to move to Sydney from China. Most of the time, they come to consult me on where should they rent when they first move to this country. Well, this might sound like a simple question, however each new comer actually always has different needs and preference on their dream property and the recommendation needs to vary depend on what feature they care the most.

With all the "consulting experience" I had; I can almost summarise these features in below category.

1. **Price**. (This directly link to the future tenant's budget depend on their financial situation)
2. **Location**. (If they come to Australia to study at Sydney Uni, of course they would prefer rent some place close or convenience to their daily routine)
3. **Quality or feature of the property** (Such as: is this property has a dish washer, security intercom, parking spots etc)
4. **Community** (This will normally link to tenants' life style. They will be asking questions such as is this area safe? Any amenities available around the area is what I usually go? (Restaurants, shops, grocery stores etc. If they have kids, they normally will be looking for neighbourhood with good school as well)

For this paper, we will be mainly analysis 3 factors: **Rental Price**, **neighbourhood safety and available amenities in the area as well** to see how we can use data science to help our new comer to select their dream property when they first move to Sydney.

Data Description:

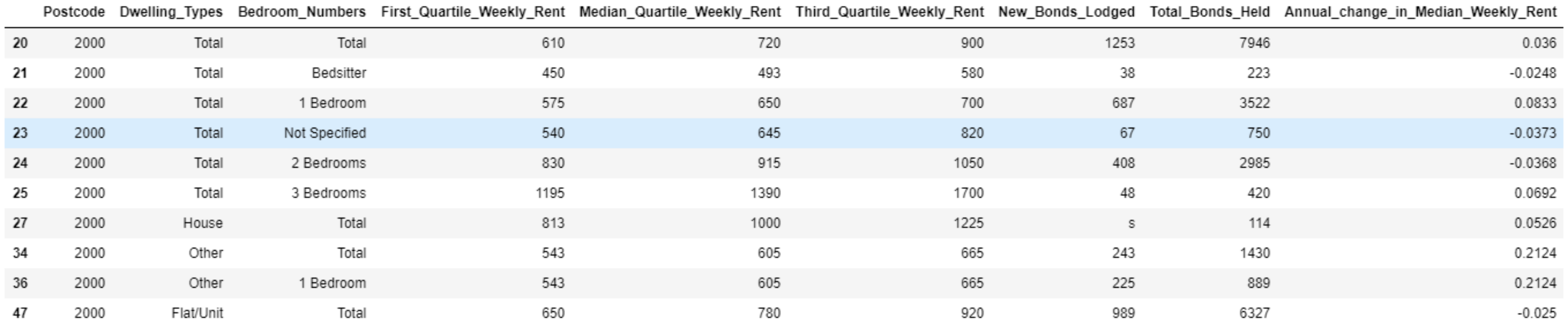
For this analysis, I have collected quite a few data via the internet which most of them are publicly available for download. Detail of these dataset and samples are listed below:

Rental Price Data:

NSW Communities & Justice provide Rent and Sales Report which is consider the sole authoritative source of data on NSW rent movements. This report is published quarterly and has been published on a regular basis since 1987 for rental movement in NSW.

For this analysis, I have downloaded the most recent rent table as of JUN-2019. It is an xlsx file and link to the table is available below. This file contains weekly rent and bond by Quartile and NSW postcode and property type. It also gives you the last Quarter or year rent movement as well. This will be the data we use to identify the rental price for each suburb in Sydney. Please notice the data is on postcode level, not neighbourhood level, hence we will use a proxy in our analysis to assume the rental price for each neighbourhood in one suburb will be around the same. [3]

https://www.facs.nsw.gov.au/resources/statistics/rent-and-sales/dashboard

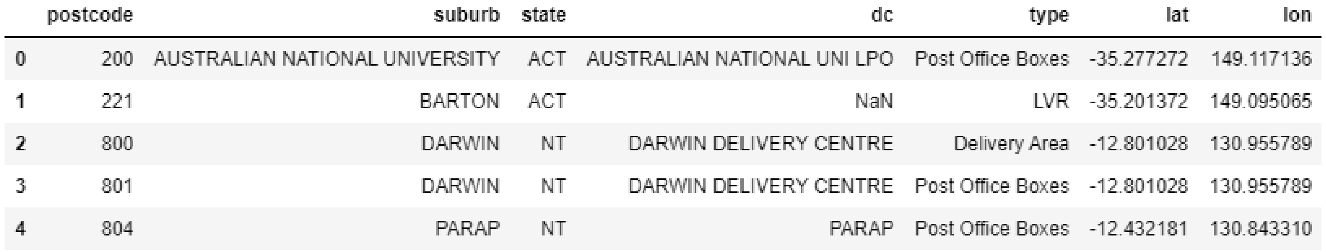


	Postcode	Dwelling_Types	Bedroom_Numbers	First_Quartile_Weekly_Rent	Median_Quartile_Weekly_Rent	Third_Quartile_Weekly_Rent	New_Bonds_Lodged	Total_Bonds_Held	Annual_change_in_Median_Weekly_Rent
20	2000	Total	Total	610	720	900	1253	7946	0.036
21	2000	Total	Bedsitter	450	493	580	38	223	-0.0248
22	2000	Total	1 Bedroom	575	650	700	687	3522	0.0833
23	2000	Total	Not Specified	540	645	820	67	750	-0.0373
24	2000	Total	2 Bedrooms	830	915	1050	408	2985	-0.0368
25	2000	Total	3 Bedrooms	1195	1390	1700	48	420	0.0692
27	2000	House	Total	813	1000	1225	s	114	0.0526
34	2000	Other	Total	543	605	665	243	1430	0.2124
36	2000	Other	1 Bedroom	543	605	665	225	889	0.2124
47	2000	Flat/Unit	Total	650	780	920	989	6327	-0.025

NSW Neighbourhoods Geo Coordinate Data:

Corra website has also supply the full list of Australia neighbourhoods with its longitude and latitude coordinates. Link provided below. Please notice, in NSW, we can have multiple neighbourhoods linked to 1 postcode. For this analysis, we will go down to neighbourhoods' level given some suburbs have a larger size compare to others and neighbourhoods will generally give you a more even split of interest area of Sydney. This will be the Geo data we will feed into Foursquare API to get venue data for each neighbourhood in Sydney. [4]

http://www.corra.com.au/australian-postcode-location-data/

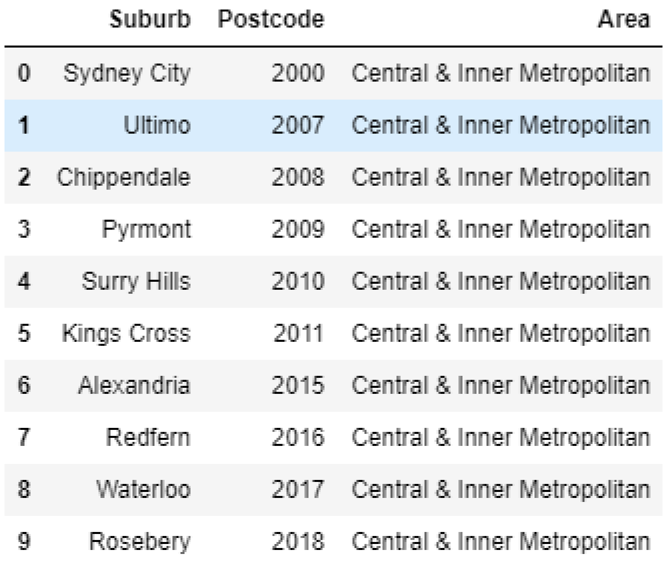


	postcode	suburb	state	dc	type	lat	lon
0	200	AUSTRALIAN NATIONAL UNIVERSITY	ACT	AUSTRALIAN NATIONAL UNI LPO	Post Office Boxes	-35.277272	149.117136
1	221	BARTON	ACT	NaN	LVR	-35.201372	149.095065
2	800	DARWIN	NT	DARWIN DELIVERY CENTRE	Delivery Area	-12.801028	130.955789
3	801	DARWIN	NT	DARWIN DELIVERY CENTRE	Post Office Boxes	-12.801028	130.955789
4	804	PARAP	NT	PARAP	Post Office Boxes	-12.432181	130.843310

Sydney Postcode and Suburb Mapping:

Prospectshop.com website powered by Equifax provided a list of postcodes used by Sydney Metro area only. For our analysis, we will be mainly focus on Sydney not remote area in NSW, as the rental price for rural area can be significantly vary compare to metropolitan area. This list of postcodes will be mainly used for us to filter the data to narrow to Sydney only. [5]

https://www.prospectshop.com.au/resources.aspx

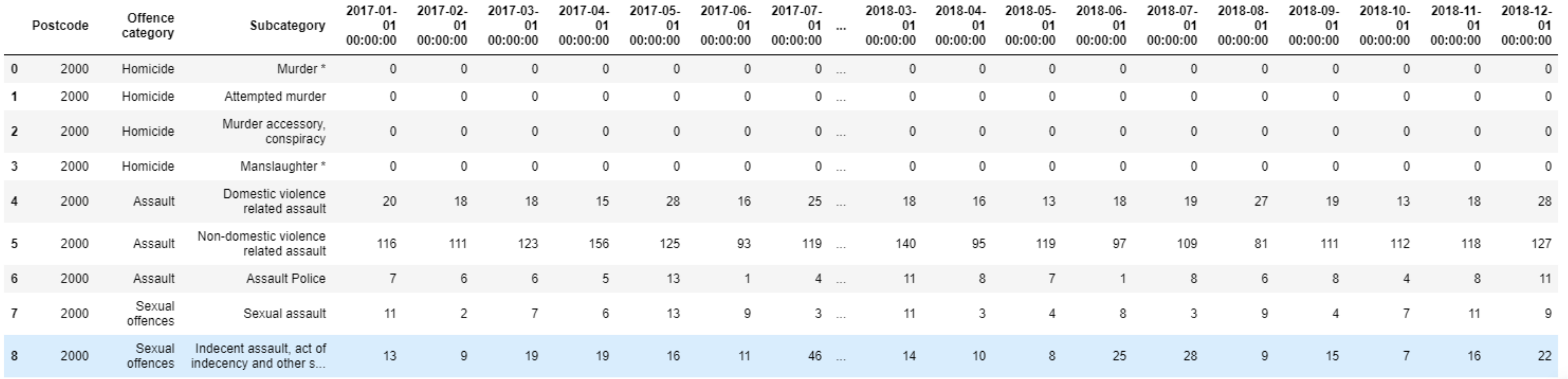


	Suburb	Postcode	Area
0	Sydney City	2000	Central & Inner Metropolitan
1	Ultimo	2007	Central & Inner Metropolitan
2	Chippendale	2008	Central & Inner Metropolitan
3	Pyrmont	2009	Central & Inner Metropolitan
4	Surry Hills	2010	Central & Inner Metropolitan
5	Kings Cross	2011	Central & Inner Metropolitan
6	Alexandria	2015	Central & Inner Metropolitan
7	Redfern	2016	Central & Inner Metropolitan
8	Waterloo	2017	Central & Inner Metropolitan
9	Rosebery	2018	Central & Inner Metropolitan

Monthly Criminal Incidents for NSW:

NSW Bureau of Crime Statistics and Research has collected and provided significant crime, court and custody datasets which updated on annually or quarterly. These data are intended for use by people who want to perform their own analysis of the information. I have downloaded a copy of Monthly data on all criminal incidents recorded by police from 1995 to Sept 2019 from link provided below. It will you details on incident count by postcode by crime type by month for NSW. This data will be mainly used to identify the neighbourhood security in Sydney. Again, this data is on postcode level, not neighbourhood level, hence we will use a proxy in our analysis to assume the neighbourhood security in one suburb will be around the same. [6]

https://www.bocsar.nsw.gov.au/Pages/bocsar_datasets/Datasets-.aspx



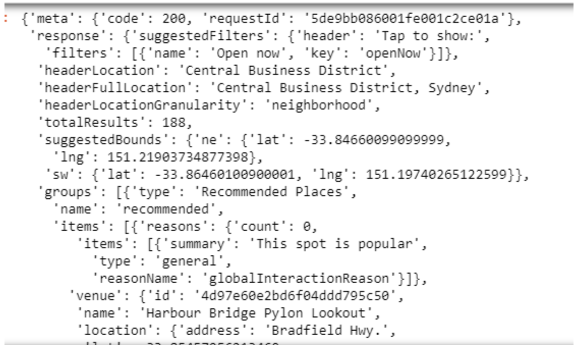
	Postcode	Offence category	Subcategory	2017-01-01 00:00:00	2017-02-01 00:00:00	2017-03-01 00:00:00	2017-04-01 00:00:00	2017-05-01 00:00:00	2017-06-01 00:00:00	2017-07-01 00:00:00	...	2018-03-01 00:00:00	2018-04-01 00:00:00	2018-05-01 00:00:00	2018-06-01 00:00:00	2018-07-01 00:00:00	2018-08-01 00:00:00	2018-09-01 00:00:00	2018-10-01 00:00:00	2018-11-01 00:00:00	2018-12-01 00:00:00
0	2000	Homicide	Murder *	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	2000	Homicide	Attempted murder	0	0	0	0	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	20	18	18	15	28	16	25	...	18	16	13	18	19	27	19	13	18	28
5	2000	Assault	Non-domestic violence related assault	116	111	123	156	125	93	119	...	140	95	119	97	109	81	111	112	118	127
6	2000	Assault	Assault Police	7	6	6	5	13	1	4	...	11	8	7	1	8	6	8	4	8	11
7	2000	Sexual offences	Sexual assault	11	2	7	6	13	9	3	...	11	3	4	8	3	9	4	7	11	9
8	2000	Sexual offences	Indecent assault, act of indecency and other s...	13	9	19	19	16	11	46	...	14	10	8	25	28	9	15	7	16	22

Foursquare API Venues data:

Foursquare API is a social location service provided by Foursquare which allows users to explore the venue by given location details (Typically GEO location coordinates). For this analysis, we will be using the NSW neighbourhoods' geo data as a base to search all venues around the list of neighbourhoods given in the file. Then grouping them into different category to explore how they distribution around Sydney suburbs.

To use Foursquare API, we will require account to be registered with Foursquare before we can call below request (with registered credential) to retrieve the information we need. (Details how to do that will be shared in the Python code supplied)

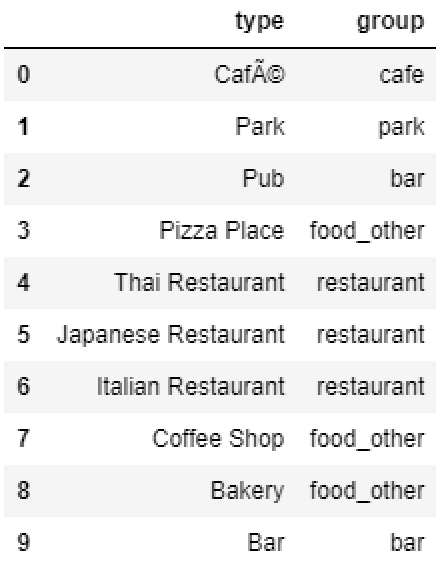
https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}



```
: {'meta': {'code': 200, 'requestId': '5de9bb086001fe001c2ce01a'},
   'response': {'suggestedFilters': {'header': 'Tap to show:',
     'filters': [{'name': 'Open now', 'key': 'openNow'}]},
    'headerLocation': 'Central Business District',
    'headerFullLocation': 'Central Business District, Sydney',
    'headerLocationGranularity': 'neighborhood',
    'totalResults': 188,
    'suggestedBounds': {'ne': {'lat': -33.84660099099999,
      'lng': 151.21903734877398},
     'sw': {'lat': -33.86460100900001, 'lng': 151.19740265122599}},
    'groups': [{'type': 'Recommended Places',
      'name': 'recommended',
      'items': [{'reasons': {'count': 0,
         'items': [{'summary': 'This spot is popular',
           'type': 'general',
           'reasonName': 'globalInteractionReason'}]},
        'venue': {'id': '4d97e60e2bd6f04ddd795c50',
         'name': 'Harbour Bridge Pylon Lookout',
         'location': {'address': 'Bradfield Hwy.',
```

Foursquare API Venues Group:

This is a dataset compiled by myself manually based on the Venue type returned by Foursquare API. Currently the venue type provided by Foursquare API have hundreds of different types. I have provided this mapping file as part of feature engineering process to better group these venues to common type recognized by most of audience such as supermarket, shopping mall, restaurant etc.



	type	group
0	CafÃ©	cafe
1	Park	park
2	Pub	bar
3	Pizza Place	food_other
4	Thai Restaurant	restaurant
5	Japanese Restaurant	restaurant
6	Italian Restaurant	restaurant
7	Coffee Shop	food_other
8	Bakery	food_other
9	Bar	bar

Methodology

Scope:

For this analysis, I will be focus only on 2 bedrooms and touch a bit on 3 bedrooms' apartment in Sydney as these are the most popular rental property types and we have most comprehensive data for most of neighbourhoods in Sydney. An exploratory data analysis will come first and give us an idea of how Sydney rental market look like by area and we will also spend some time on what are the features that might have influence our rental price. The outcome will give us a general idea of which neighbourhood our new comer could consider to rent. Then we will be using unsupervised clustering algorithm to identify other alternative neighbourhoods which have similar venue types and have similar security rating and rental as well to give more options for us to choose.

Data Cleaning:

Most of the data we have downloaded are in either csv or xlsx format which can be easily imported into Python for further processing. I have done a few pre-processing to remove some of the headed in the in the data file and then upload to IBM Watson studio. There are several assumptions I have taken to merge the dataset into one data mart so it will be more fitting for our current analysis.

For rental price we have data as of SEP 2019 and it is on postcode level. Crime incident data has info since 1995 until DEC 2018. For our analysis, given we are looking at recent rental trend, I will remove all crime data before 2017 so we can have the latest trend on postcode level security. Geo data we have is on neighbourhood level and our analysis will be based on neighbourhood eventually. Hence, I have to take the assumption that for all neighbourhood in one postcode, the crime and rental data will be the same.

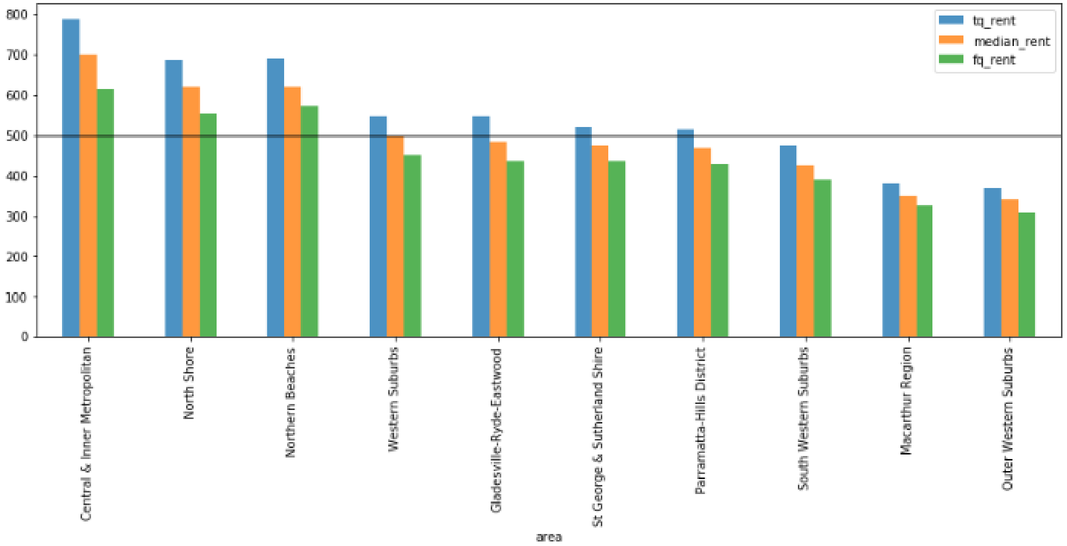
Based on these assumptions, I have constructed a DataMart with average rental price, total rental bond number, total crime count, venue count by venue type (group) in one table. Together with that we can further split the data by neighbourhood, postcode, Sydney region and property type dimension for our further analysis.

Exploratory data analysis:

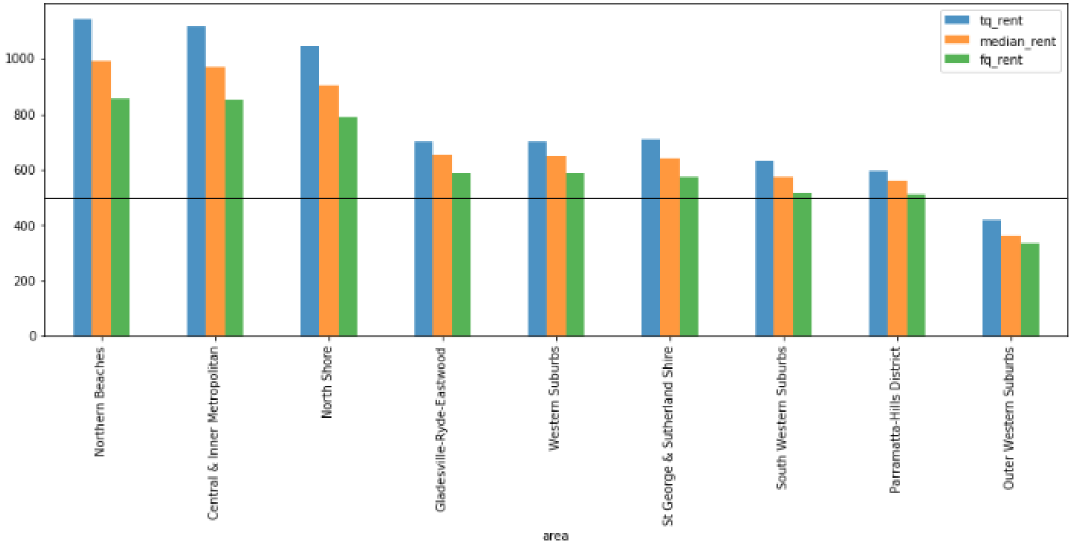
Average rental price by Suburb:

I start with get a general idea of how Sydney rental price looks like for 2/3 bedrooms apartment. By looking at the average rental prices by region, I have created below charts. As we can see, the median rent for 2 bedrooms apartment in Sydney is around 450$ per week. The region has highest average rental price will be CBD central area which around 700$ per week and the lowest will be out Wester Sydney region which is more than 30Km away from CBD and the average medium rental price is around 330$ per week. Similar trend shows in the 3 bedrooms' apartment rental price as well. However instead of 500$ rental per week, for 3 beds the average rental is around 650$ Sydney wide. The highest area in Sydney for 3 beds is Northern beach area which almost approaching to 1000$ per week mark.

Average Rental Price for 2 Bedrooms Apartment by Sydney Region

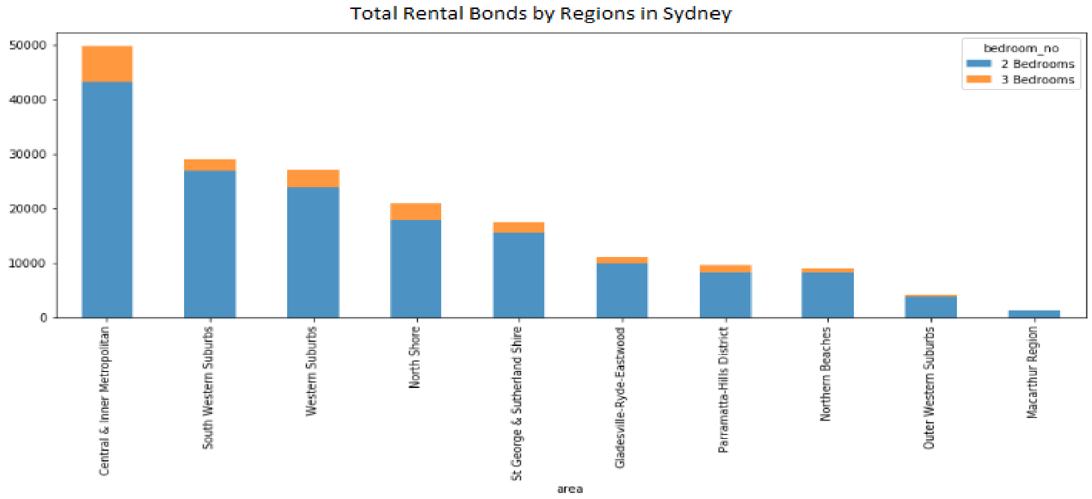


Average Rental Price for 3 Bedrooms Apartment by Sydney Region

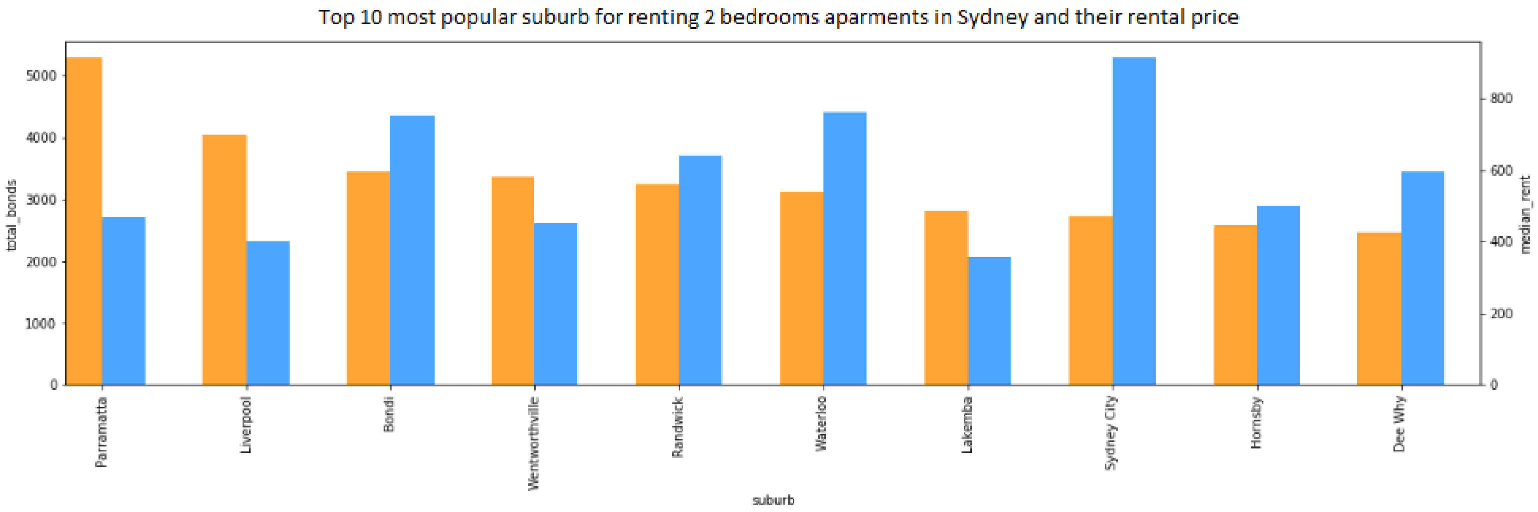


Which are in Sydney are popular among Tenants?

According to the total bonds' statistic, the most popular area in Sydney amount Tenants is Central CBD as expected. Although it has the highest average rental prices, seems to me location is definitely a very important factor when people selecting the place to rent. Also, from below chat, we can see, 2-bedroom apartments are definitely more popular than 3 bedrooms. The total bond of 2 bedrooms as of September 2019 is almost 15 times of 3 bedrooms bond in Sydney.

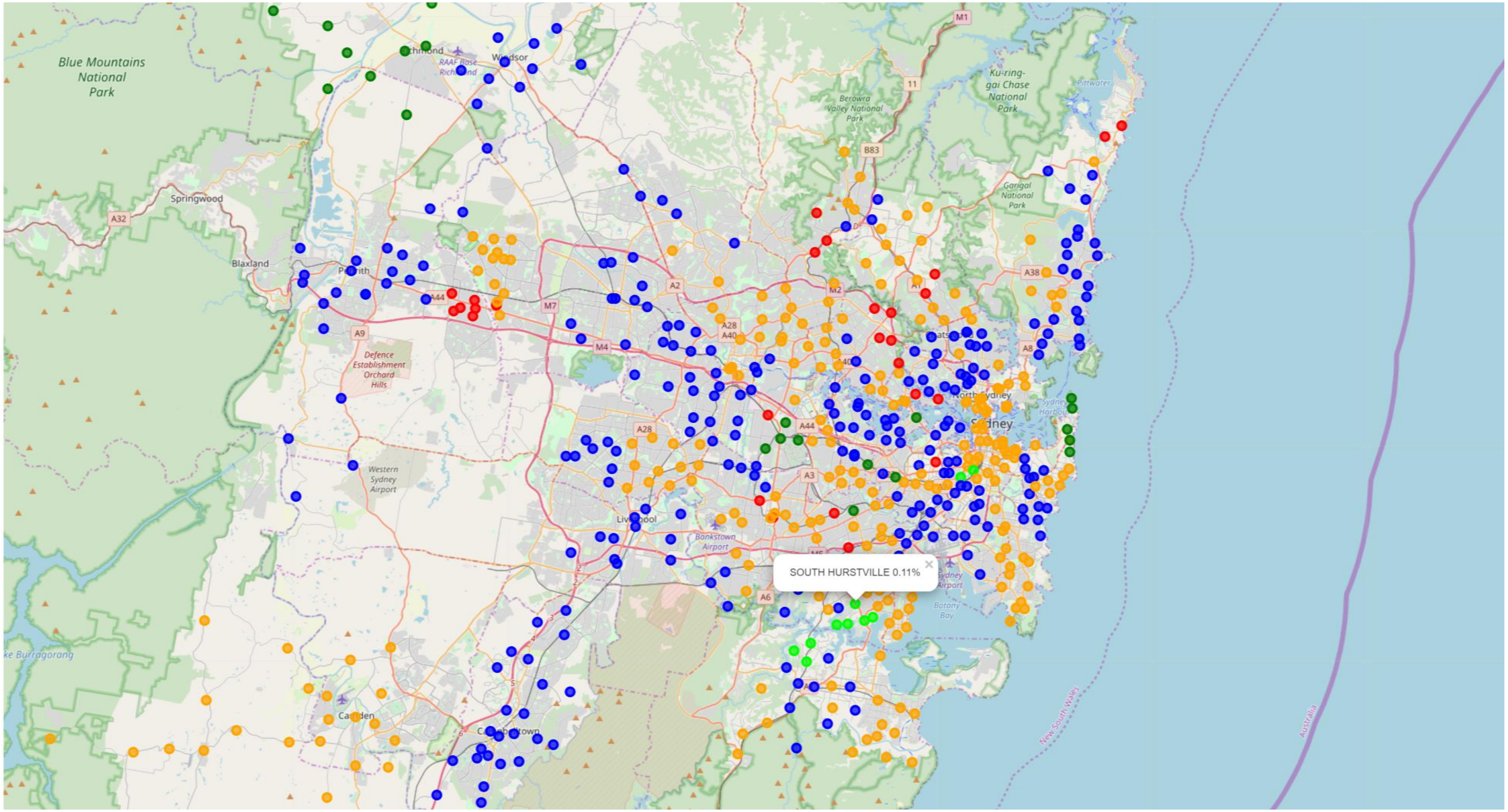


To explore a bit more, I have listed the top 10 most popular suburb by rental price in Sydney. This time we can see the top 2 suburbs are Parramatta and Liverpool (western region and south region) instead of central. Both of them have a relative lower rental price. (For people who never been in Sydney, these 2 areas have very good transportation systems, excellent facilities such as large shopping mall and commercial area.) To my surprise, the third popular suburb is actually Bondi, which is on eastern coastal. Given it close to the most famous Beach in Sydney (Bondi beach). The average rental price in this area for 2 beds is close to 800$ per week.



How is neighbourhood rental price movement look like in 2019?

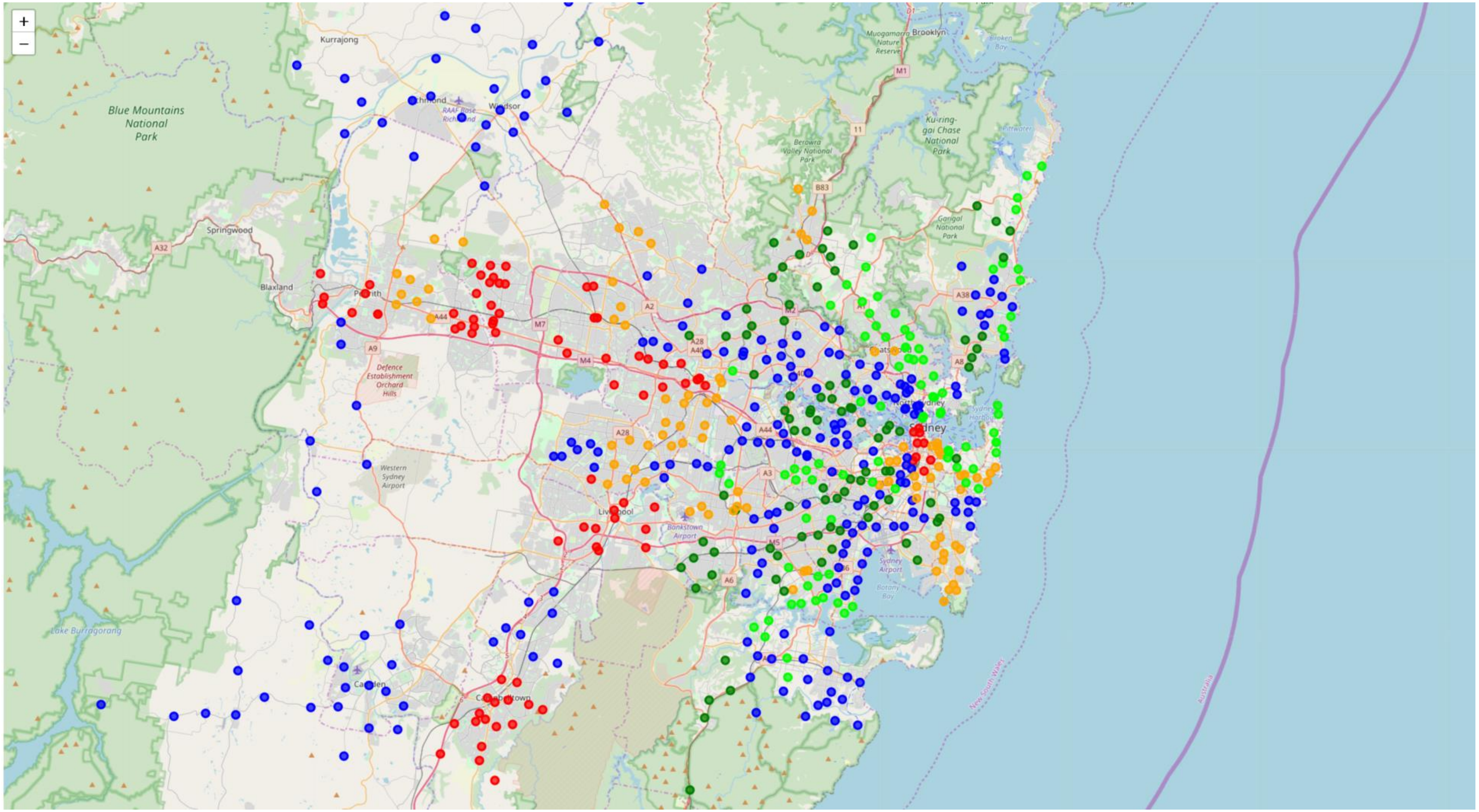
In the rental data I have download, we also have the medium rental price movement 2019 YTD by suburb in Sydney. To let audien ce directly have a visual on how that look like in Sydney, I have plotted the rental movement onto Sydney map using Python Folium. Lime color is more than 10% increase YTD, orange color is for anything decrease and red color is more than 10% decrease YTD. It is quite clear that over all Sydney ren tal price has a decrease trend in 2019 in particular around out west region. The highest growth suburb is South Hurstville which have 11% increase on rent.



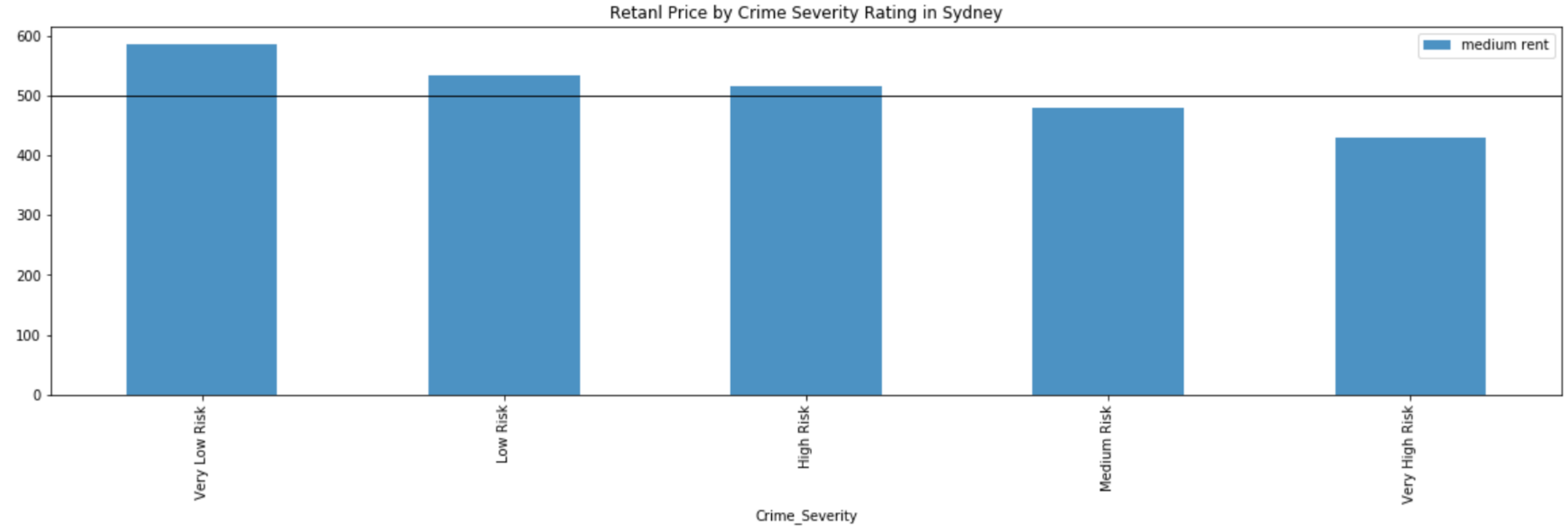
How is neighbourhood safety look like in Sydney?

I also look into the neighbourbood safety in Sydney by plot the total number of major crime incidents from Jan 2017 to end of 2018 onto Sydney map. I have split the number of crime incident for each suburb into 5 percentiles.

Light green color represents the lowest 0 – 15%; Dark green color represents 15% - 30%; Blue color represents 30% - 70%; Orange color is for 70%-85% and the rest are high risk area shows as red color on the map. As shown on below chart. The areas have most of major crime incident spread among out west, west south and CBD Area and the safest area in Sydney will be the northern suburbs, inner west, southern suburbs or among the east or east north coastline.

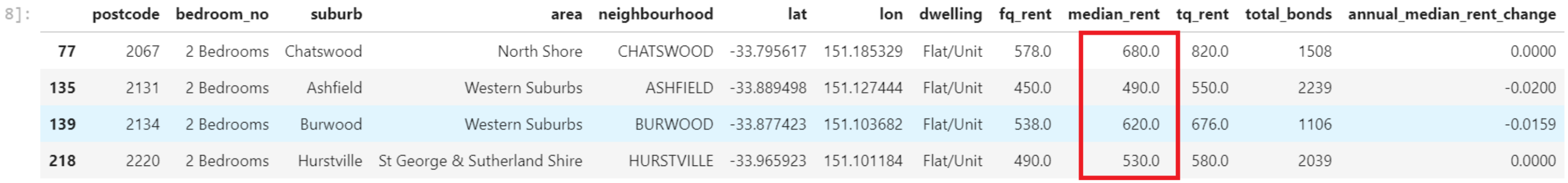


Will rental price affect by the neighbourhood safety factor? I'm sure it should have some influence. If we simply average the medium rent for all suburbs by neighbourhood safety, we can see at least the rental price is in reverse order compare to the risk factor.



Design an unsupervised model to solve our problem:

As most of my friend are Asian background. When they first come to Australia, I normally recommend them to rent at a suburb surround by Asian community such as, Burwood, Hurstville, Chatswood, Ashfield etc. However, these suburbs normally have a higher rental price. As shown below: you can see these 4 suburbs all have a medium rental price above the Sydney average $450. Especially Chatswood, which reach to 680$ per week for a 2-bedroom apartment. For students come to Sydney to study, some time they prefer place cheaper and have the similar community as these listed 4. Hence, I plan to use Kmeans algorithm as part of this clustering study to see if we can identify any other neighbourhood in Sydney that have similar venues and facilities like these 4 Asian suburb I have listed above with cheaper rental price which can be afford by an Asian oversea student.



8]:

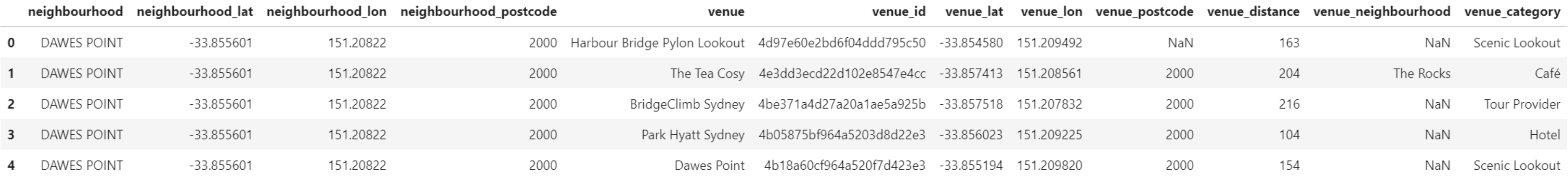
	postcode	bedroom_no	suburb	area	neighbourhood	lat	lon	dwelling	fq_rent	median_rent	tq_rent	total_bonds	annual_median_rent_change
77	2067	2 Bedrooms	Chatswood	North Shore	CHATSWOOD	-33.795617	151.185329	Flat/Unit	578.0	680.0	820.0	1508	0.0000
135	2131	2 Bedrooms	Ashfield	Western Suburbs	ASHFIELD	-33.889498	151.127444	Flat/Unit	450.0	490.0	550.0	2239	-0.0200
139	2134	2 Bedrooms	Burwood	Western Suburbs	BURWOOD	-33.877423	151.103682	Flat/Unit	538.0	620.0	676.0	1106	-0.0159
218	2220	2 Bedrooms	Hurstville	St George & Sutherland Shire	HURSTVILLE	-33.965923	151.101184	Flat/Unit	490.0	530.0	580.0	2039	0.0000

PUBLIC

Collect additional venue data using Foursquare API

To start our model, we need additional venue info to be collected which I have utilized the Foursquare API to explore the neighbourhood and segment them. To get as much info I can, I did not put a limit on the number of venues return from API and marked the radius 600 meter for each neighbourhood from their given latitude and longitude which normally the town centre (where the post office located). After I call the API, I store the rerun JSON object in a data frame and keep information as below:

The most important info we get is the venue and venue category. We have total 9151 venues return from Foursquare API for whole Sydney. Please notice, some of these venues could be duplicated as they might belong to 2 neighbourhoods. We will not remove these duplicates as these neighbourhood should be considered share the same venue in this case.



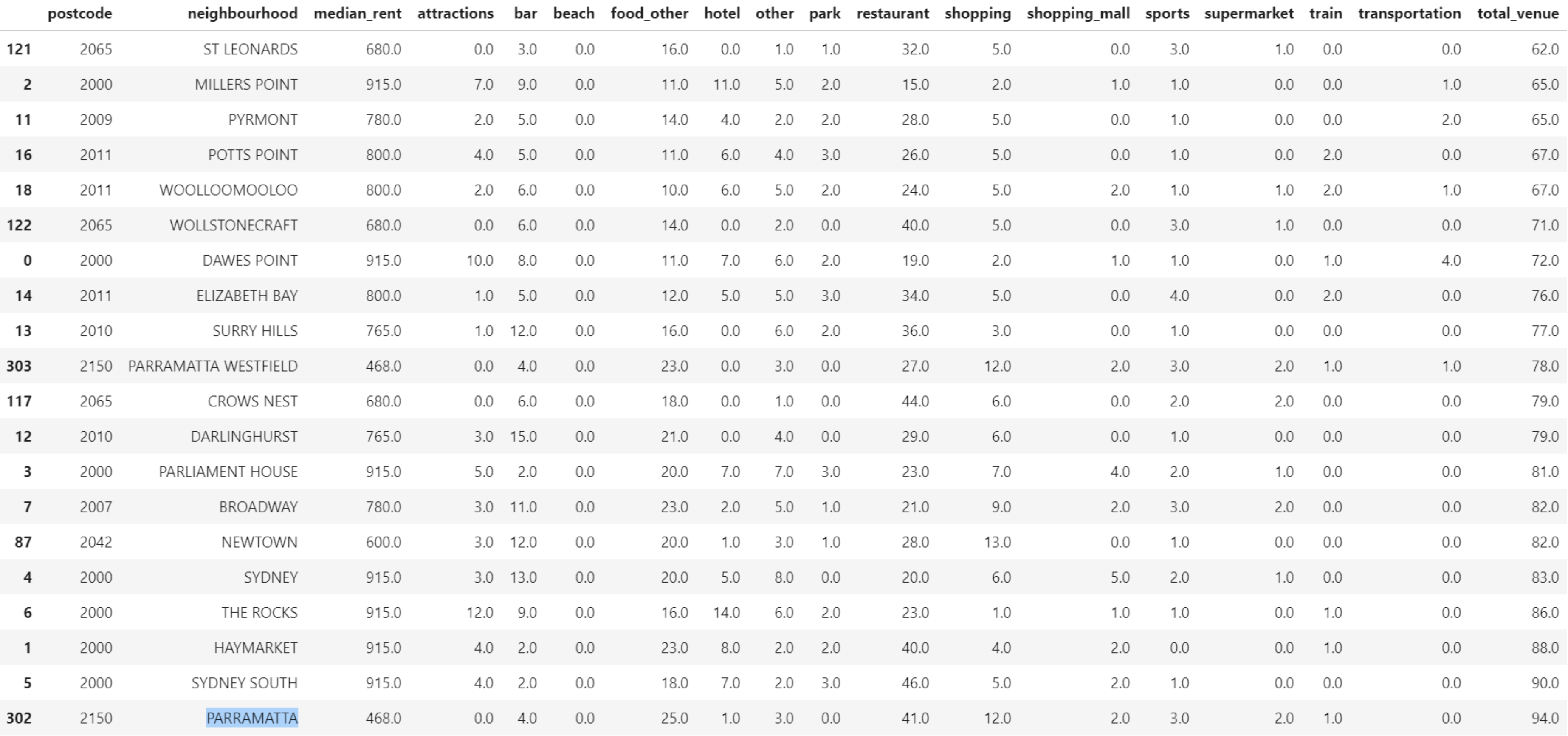
	neighbourhood	neighbourhood_lat	neighbourhood_lon	neighbourhood_postcode	venue	venue_id	venue_lat	venue_lon	venue_postcode	venue_distance	venue_neighbourhood	venue_category
0	DAWES POINT	-33.855601	151.20822	2000	Harbour Bridge Pylon Lookout	4d97e60e2bd6f04ddd795c50	-33.854580	151.209492	NaN	163	NaN	Scenic Lookout
1	DAWES POINT	-33.855601	151.20822	2000	The Tea Cosy	4e3dd3ecd22d102e8547e4cc	-33.857413	151.208561	2000	204	The Rocks	Café
2	DAWES POINT	-33.855601	151.20822	2000	BridgeClimb Sydney	4be371a4d27a20a1ae5a925b	-33.857518	151.207832	2000	216	NaN	Tour Provider
3	DAWES POINT	-33.855601	151.20822	2000	Park Hyatt Sydney	4b05875bf964a5203d8d22e3	-33.856023	151.209225	2000	104	NaN	Hotel
4	DAWES POINT	-33.855601	151.20822	2000	Dawes Point	4b18a60cf964a520f7d423e3	-33.855194	151.209820	2000	154	NaN	Scenic Lookout

Before I go straight into the model, I also created a mapping table to group these venues by types to get some general understanding on venue data myself. One the high level, I can see these venues can be grouped into below major groups. This is good for general understanding of the venue data.

bar	cafe	hotel	other	park	restaurant	shopping	shopping mall	sports	supermarket	train	transportation	attraction

If we just list the top 20 neighbourhoods with most of venues around as well as the least 20 neighbourhoods with at least 5 venues (some rural area in Sydney do not have venues around listed in 4 Square API. (Listed below). You can see clearly, for these neighbourhoods with less venues have lower average medium rent compare to these neighbourhoods with a lot of venues around. However, this is not always the case. Parramatta, the most popular suburb in Sydney we find out previously has the highest venue count, however for 2-bedroom apartment rental price in this area is just marginally above the Sydney average.

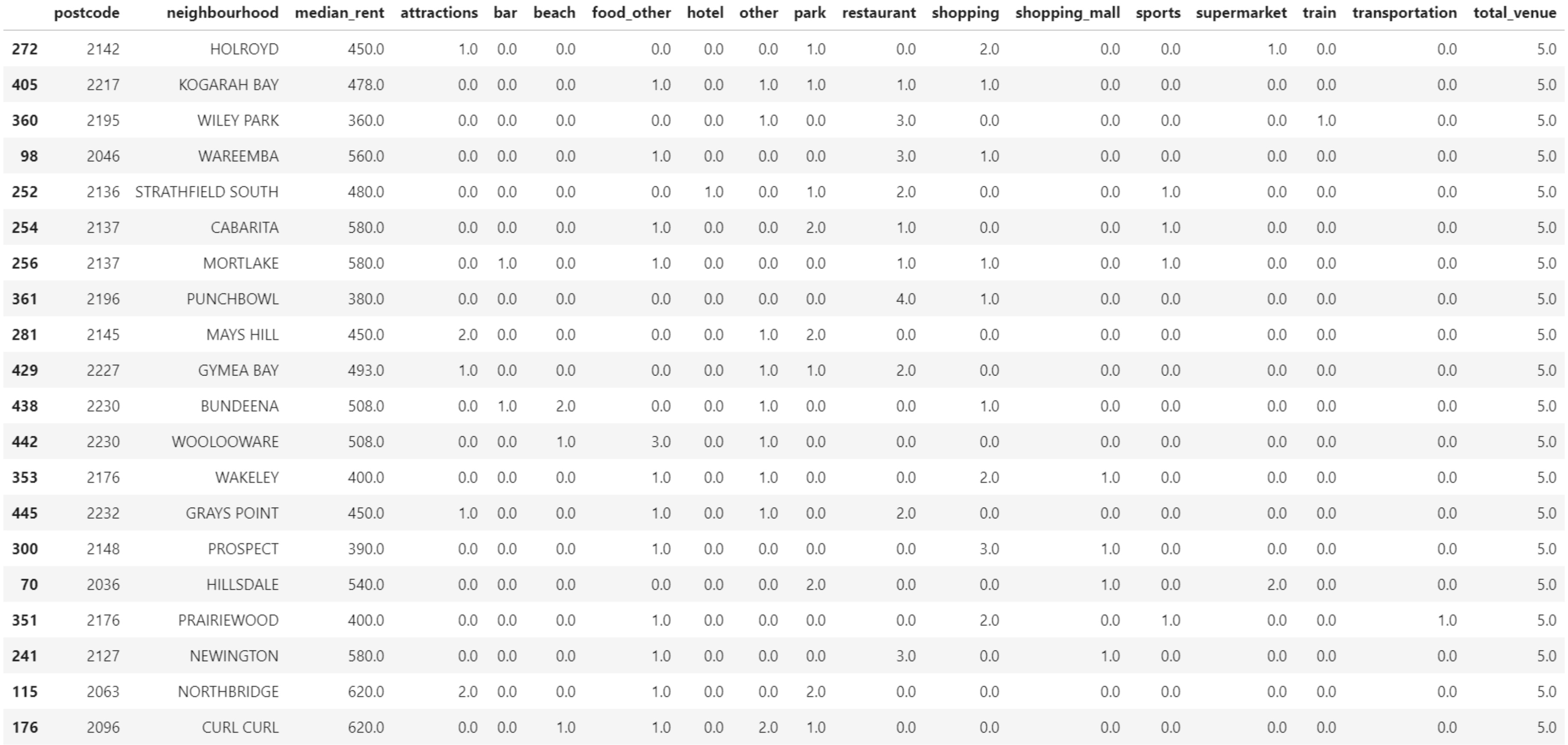
Top 20



	postcode	neighbourhood	median_rent	attractions	bar	beach	food_other	hotel	other	park	restaurant	shopping	shopping_mall	sports	supermarket	train	transportation	total_venue
121	2065	ST LEONARDS	680.0	0.0	3.0	0.0	16.0	0.0	1.0	1.0	32.0	5.0	0.0	3.0	1.0	0.0	0.0	62.0
2	2000	MILLERS POINT	915.0	7.0	9.0	0.0	11.0	11.0	5.0	2.0	15.0	2.0	1.0	1.0	0.0	0.0	1.0	65.0
11	2009	PYRMONT	780.0	2.0	5.0	0.0	14.0	4.0	2.0	2.0	28.0	5.0	0.0	1.0	0.0	0.0	2.0	65.0
16	2011	POTTS POINT	800.0	4.0	5.0	0.0	11.0	6.0	4.0	3.0	26.0	5.0	0.0	1.0	0.0	2.0	0.0	67.0
18	2011	WOOLLOOMOOLOO	800.0	2.0	6.0	0.0	10.0	6.0	5.0	2.0	24.0	5.0	2.0	1.0	1.0	2.0	1.0	67.0
122	2065	WOLLSTONECRAFT	680.0	0.0	6.0	0.0	14.0	0.0	2.0	0.0	40.0	5.0	0.0	3.0	1.0	0.0	0.0	71.0
0	2000	DAWES POINT	915.0	10.0	8.0	0.0	11.0	7.0	6.0	2.0	19.0	2.0	1.0	1.0	0.0	1.0	4.0	72.0
14	2011	ELIZABETH BAY	800.0	1.0	5.0	0.0	12.0	5.0	5.0	3.0	34.0	5.0	0.0	4.0	0.0	2.0	0.0	76.0
13	2010	SURRY HILLS	765.0	1.0	12.0	0.0	16.0	0.0	6.0	2.0	36.0	3.0	0.0	1.0	0.0	0.0	0.0	77.0
303	2150	PARRAMATTA WESTFIELD	468.0	0.0	4.0	0.0	23.0	0.0	3.0	0.0	27.0	12.0	2.0	3.0	2.0	1.0	1.0	78.0
117	2065	CROWS NEST	680.0	0.0	6.0	0.0	18.0	0.0	1.0	0.0	44.0	6.0	0.0	2.0	2.0	0.0	0.0	79.0
12	2010	DARLINGHURST	765.0	3.0	15.0	0.0	21.0	0.0	4.0	0.0	29.0	6.0	0.0	1.0	0.0	0.0	0.0	79.0
3	2000	PARLIAMENT HOUSE	915.0	5.0	2.0	0.0	20.0	7.0	7.0	3.0	23.0	7.0	4.0	2.0	1.0	0.0	0.0	81.0
7	2007	BROADWAY	780.0	3.0	11.0	0.0	23.0	2.0	5.0	1.0	21.0	9.0	2.0	3.0	2.0	0.0	0.0	82.0
87	2042	NEWTOWN	600.0	3.0	12.0	0.0	20.0	1.0	3.0	1.0	28.0	13.0	0.0	1.0	0.0	0.0	0.0	82.0
4	2000	SYDNEY	915.0	3.0	13.0	0.0	20.0	5.0	8.0	0.0	20.0	6.0	5.0	2.0	1.0	0.0	0.0	83.0
6	2000	THE ROCKS	915.0	12.0	9.0	0.0	16.0	14.0	6.0	2.0	23.0	1.0	1.0	1.0	0.0	1.0	0.0	86.0
1	2000	HAYMARKET	915.0	4.0	2.0	0.0	23.0	8.0	2.0	2.0	40.0	4.0	2.0	0.0	0.0	1.0	0.0	88.0
5	2000	SYDNEY SOUTH	915.0	4.0	2.0	0.0	18.0	7.0	2.0	3.0	46.0	5.0	2.0	1.0	0.0	0.0	0.0	90.0
302	2150	PARRAMATTA	468.0	0.0	4.0	0.0	25.0	1.0	3.0	0.0	41.0	12.0	2.0	3.0	2.0	1.0	0.0	94.0

Bottom 20

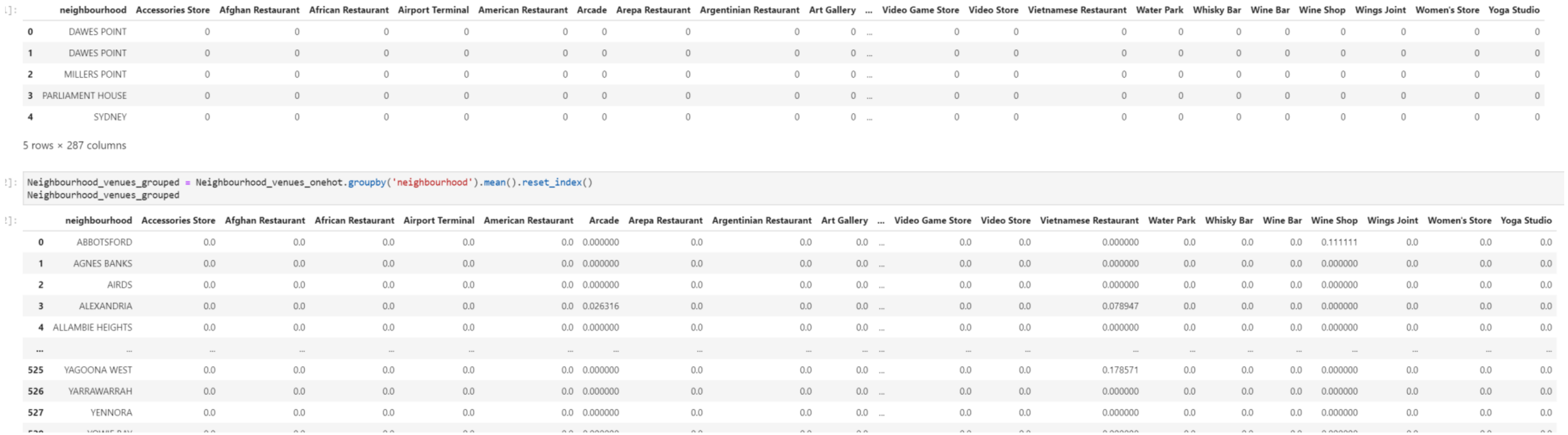
PUBLIC



	postcode	neighbourhood	median_rent	attractions	bar	beach	food_other	hotel	other	park	restaurant	shopping	shopping_mall	sports	supermarket	train	transportation	total_venue
272	2142	HOLROYD	450.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	2.0	0.0	0.0	1.0	0.0	0.0	5.0
405	2217	KOGARAH BAY	478.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	5.0
360	2195	WILEY PARK	360.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0	0.0	0.0	0.0	0.0	1.0	0.0	5.0
98	2046	WAREEMBA	560.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	3.0	1.0	0.0	0.0	0.0	0.0	0.0	5.0
252	2136	STRATHFIELD SOUTH	480.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	2.0	0.0	0.0	1.0	0.0	0.0	0.0	5.0
254	2137	CABARITA	580.0	0.0	0.0	0.0	1.0	0.0	0.0	2.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	5.0
256	2137	MORTLAKE	580.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	5.0
361	2196	PUNCHBOWL	380.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	1.0	0.0	0.0	0.0	0.0	0.0	5.0
281	2145	MAYS HILL	450.0	2.0	0.0	0.0	0.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0
429	2227	GYMEA BAY	493.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0
438	2230	BUNDEENA	508.0	0.0	1.0	2.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	5.0
442	2230	WOOLOOWARE	508.0	0.0	0.0	1.0	3.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0
353	2176	WAKELEY	400.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	2.0	1.0	0.0	0.0	0.0	0.0	5.0
445	2232	GRAYS POINT	450.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0
300	2148	PROSPECT	390.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	3.0	1.0	0.0	0.0	0.0	0.0	5.0
70	2036	HILLSDALE	540.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	1.0	0.0	2.0	0.0	0.0	5.0
351	2176	PRAIRIEWOOD	400.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	0.0	1.0	0.0	0.0	1.0	5.0
241	2127	NEWINGTON	580.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	3.0	0.0	1.0	0.0	0.0	0.0	0.0	5.0
115	2063	NORTHBRIDGE	620.0	2.0	0.0	0.0	1.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0
176	2096	CURL CURL	620.0	0.0	0.0	1.0	1.0	0.0	2.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0

Data Prepare for our Kmeans algorithm

First 1 hot encoding has been performed on the dataset by neighbourhood by venue type. Then we calculate the appearance of each venue type in each neighbourhood so we can sort and create a new table to show what are the most common venues in each neighbourhood.



	neighbourhood	Accessories Store	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Video Game Store	Video Store	Vietnamese Restaurant	Water Park	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio
0	DAWES POINT	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	DAWES POINT	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	MILLERS POINT	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	PARLIAMENT HOUSE	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	SYDNEY	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

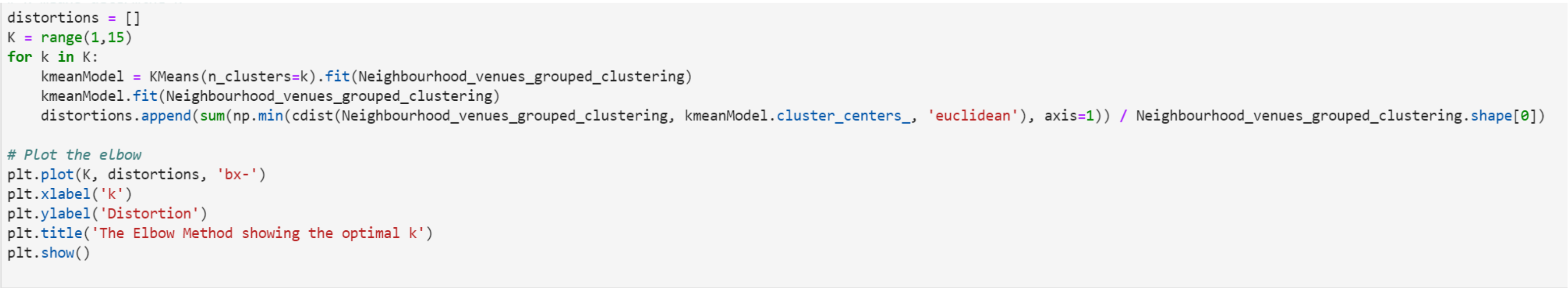
5 rows × 287 columns

```
Neighbourhood_venues_grouped = Neighbourhood_venues_onehot.groupby('neighbourhood').mean().reset_index()
Neighbourhood_venues_grouped
```

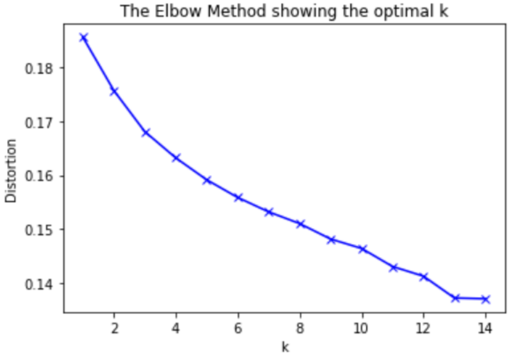
	neighbourhood	Accessories Store	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Video Game Store	Video Store	Vietnamese Restaurant	Water Park	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio
0	ABBOTSFORD	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.111111	0.0	0.0	0.0
1	AGNES BANKS	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
2	AIRDS	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
3	ALEXANDRIA	0.0	0.0	0.0	0.0	0.0	0.026316	0.0	0.0	0.0	...	0.0	0.0	0.078947	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
4	ALLAMBIE HEIGHTS	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
525	YAGOONA WEST	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.178571	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
526	YARRAWARRAH	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
527	YENNORA	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0

I have removed these suburbs with 20 venues or less as they are lack of info to be cluster with other neighbourhoods together. Also, I have notice we have quite a few venue types in Sydney that only appear very few times eg: Argentinian Restaurant, Tunnel, Trade school. I have removed these columns as well, although they will have very few influences on the model, I think take away these noises is good practice. The final table I have constructed is called Neighbourhood_venues_grouped_clustering

I have used the K-Means with elbow method to identified that 14 clusters is optimum k of the K-Means as shown below. Euclidean distance has been chosen by me to check on the result for each level of clustering.



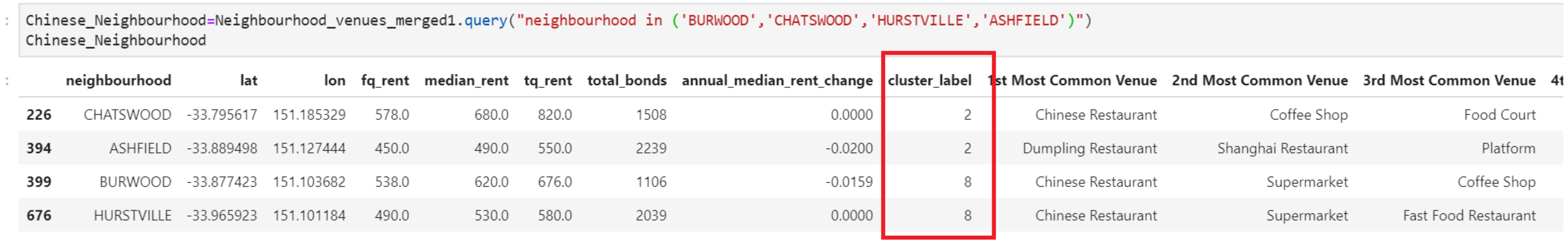
```
distortions = []
K = range(1,15)
for k in K:
    kmeanModel = KMeans(n_clusters=k).fit(Neighbourhood_venues_grouped_clustering)
    kmeanModel.fit(Neighbourhood_venues_grouped_clustering)
    distortions.append(sum(np.min(cdist(Neighbourhood_venues_grouped_clustering, kmeanModel.cluster_centers_, 'euclidean'), axis=1)) / Neighbourhood_venues_grouped_clustering.shape[0])

# Plot the elbow
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```



A table with first 10 most common venues in each suburb has also been created later on to for us to view what are the common venues in these neighbourhood that has been clustered together. After the clustering get done, a new field called cluster label has been created in the dataset shows which neighbourhood has been grouped together.

Result

Lets have a look how our clustering perform. I have specifically pull out these 4 Chinese suburbs I have mentioned previously. As show on below screen shot, Chatswood, Ashfield has been grouped into cluster 2 and Burwood, Hurstville has been clustered into Group 8. All four of them have the most common venue as Chinese Restaurants, Shanghai Restaurant etc.



```
Chinese_Neighbourhood=Neighbourhood_venues_merged1.query("neighbourhood in ('BURWOOD','CHATSWOOD','HURSTVILLE','ASHFIELD')")
Chinese_Neighbourhood
```

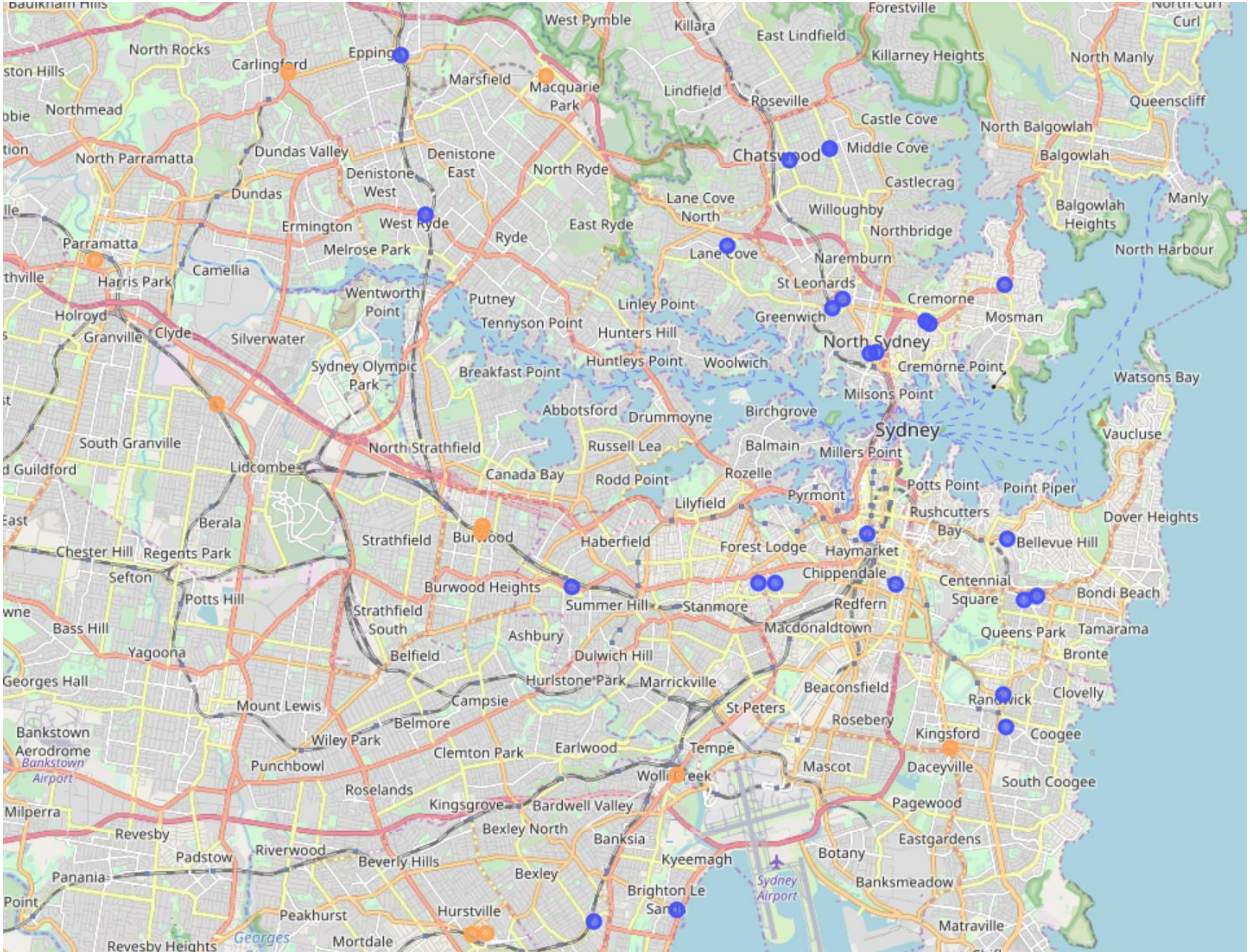
	neighbourhood	lat	lon	fq_rent	median_rent	tq_rent	total_bonds	annual_median_rent_change	cluster_label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4t
226	CHATSWOOD	-33.795617	151.185329	578.0	680.0	820.0	1508	0.0000	2	Chinese Restaurant	Coffee Shop	Food Court	
394	ASHFIELD	-33.889498	151.127444	450.0	490.0	550.0	2239	-0.0200	2	Dumpling Restaurant	Shanghai Restaurant	Platform	
399	BURWOOD	-33.877423	151.103682	538.0	620.0	676.0	1106	-0.0159	8	Chinese Restaurant	Supermarket	Coffee Shop	
676	HURSTVILLE	-33.965923	151.101184	490.0	530.0	580.0	2039	0.0000	8	Chinese Restaurant	Supermarket	Fast Food Restaurant	

So do we have any other similar neighbourhood like these 2 cluster in Sydney? Lets plot them on the map. It seems to me that we have a few alternative neightbourhood that similar to these 4 I normally recommend to my friend. Also, a few of these neighbourhood has a bit lower medium re ntal price for 2-bedroom apartment while the safety still remains relatively okay.

Blue shows on the map is cluster 2 and Orange for cluster 8.

If the budget is enough, they can consider neighbourhood in cluster 2 which are similar like Chatswood and Ashfield. e.g.: N orth Sydney, Camperdown Crows Next, West Ryde, Epping etc.

The cheaper option for neighbourhoods like Burwood and Hurstville will be in cluster 8. e.g. : Kingsford, Carlingford Court, Parramatta, Auburn and Wolli Creek, Macquarie Park.


Baulkham Hills
West Pymble
Killara
Forestville
North Curl Curl
East Lindfield
North Rocks
Epping
Killarney Heights
North Manly
Carlingford
Marsfield
Macquarie Park
Baulkham Hills
Lindfield
Roseville
Queenscliff
Northmead
Castle Cove
North Balgowlah
Middle Cove
Dundas Valley
Denistone East
Chatswood
Balgowlah
North Parramatta
Denistone West
North Ryde
Castlecrag
Dundas
Lane Cove North
Willoughby
Balgowlah Heights
Manly
Ermington
West Ryde
Ryde
East Ryde
Northbridge
Parramatta
Melrose Park
Lane Cove
Naremburn
North Harbour
Camellia
Harris Park
St Leonards
Wentworth Point
Putney
Linley Point
Cremorne
Holroyd
Greenwich
Mosman
Tennyson Point
Hunters Hill
Granville
Clyde
Silverwater
North Sydney
Sydney Olympic Park
Huntleys Point
Woolwich
Cremorne Point
Breakfast Point
Watsons Bay
Milsons Point
Abbotsford
Drummoyne
Birchgrove
Sydney
Vaucluse
South Granville
North Strathfield
Russell Lea
Balmain
Millers Point
Guildford
Lidcombe
Canada Bay
Rodd Point
Rozelle
Potts Point
Point Piper
Pyrmont
Lilyfield
Rushcutters Bay
Dover Heights
Berala
Strathfield
Burwood
Haberfield
Forest Lodge
Haymarket
Bellevue Hill
Chester Hill
Regents Park
Chippendale
Centennial Square
Bondi Beach
Sefton
Burwood Heights
Summer Hill
Stanmore
Redfern
Potts Hill
Strathfield South
Macdonaldtown
Queens Park
Tamarama
Bass Hill
Ashbury
Bronte
Yagoona
Belfield
Dulwich Hill
Hurlstone Park
Marrickville
Beaconsfield
Georges Hall
Clovelly
Campsie
Randwick
Mount Lewis
St Peters
Rosebery
Bankstown Aerodrome
Belmore
Kingsford
Coogee
Bankstown Airport
Wiley Park
Clemton Park
Earlwood
Tempe
Punchbowl
Mascot
Daceyville
Wolli Creek
South Coogee
Roselands
Milperra
Kingsgrove
Bardwell Valley
Pagewood
Bexley North
Revesby
Banksia
Eastgardens
Padstow
Riverwood
Kyeemagh
Botany
Beverly Hills
Sydney Airport
Panania
Bexley
Brighton Le Sands
Banksmeadow
Peakhurst
Hurstville
Matraville
Revesby Heights
Georges
Mortdale

Discussion

Sydney is one of the largest cities in the world and also one of the most multicultural cities as well. The type of venues exists in Sydney shows a very complex mix and the size of each neighbourhood vary significantly. With all the 9000+ venues I have extracted from Foursquare API, just restaurant itself I can see 50+ different types. A few different approaches have been tried on to clustering these neighbourhood together and finally I have selected K means. Unfortunately, these has no perfect method to find out all the details on Sydney neighbourbood with the limited data we have, and not all customer will yield the best quality results however, based on my 10+ year experience living in Sydney, I'm quite happy with what I have found using above method. A few "wow" interesting finding also give me new knowledge about this city I have never knew before.

The important part is, I have used most of the knowledge I have gain during this data scientist cause. Such as using Kmeans algorithm as part of this clustering study. Understand how to use the Elbow method to identify the optimum k value. If more data is available, we can certainly drill down to more details and expanded our analysis in to more granular blocks.

For this analysis, I focus on Chinese communities in Sydney, which can be easily changed to the other type of communities as well by looking at the data from different angle. All data and python code of this project are saved on my Github page, which can be used in the future if have time. I have also tested feature selection during my analysis. However due to this is more of an unsupervised model, hence I did not include the result in my analysis. The code exists as comments at the end of my notebook for reference.

I hope this analysis and all the visualization created during this project will help new comer to Sydney to find their first idea rental property and for future studies as well.

Conclusion

There are a lot of community hidden in Sydney and most of them have very similar culture and facilities. However, dig in deeper you would be surprised to see the rental price for these areas could vary significantly. If you would like to save money while also enjoy a high quality life style that fits your need. Be patient and do some research yourself and you might be surprised to what you can find.

To data science!

Terence Sun (Tianrui Sun)

References:

[1] “Greater Sydney”

https://www.cityofsydney.nsw.gov.au/learn/research-and-statistics/the-city-at-a-glance/greater-sydney

[2] “Sydney rental vacancy rate drop”. Su-Lin Tan Reporter

https://www.afr.com/property/residential/sydney-rental-vacancy-rate-drops-20191011-p52ztk”

[3] “NSW Rent and Sales Report”. NSW Communities & Justice

https://www.facs.nsw.gov.au/resources/statistics/rent-and-sales/dashboard

[4] NSW Postal Geo Data

http://www.corra.com.au/australian-postcode-location-data/

[5] “Sydney metropolitan Postcode”. Equifax

https://www.prospectshop.com.au/resources.aspx

[6] “NSW all criminal incidents recorded”. NSW Bureau of Crime Statistics and Research

https://www.bocsar.nsw.gov.au/Pages/bocsar_datasets/Datasets-.aspx