Housing Sales Prices Vs Venues Data Analysis of Sydney

Shanshan Jin | Coursera Capstone Project | 18-07-2020

1 Table of Contents

2 Introduction

2.1 BACKGROUND

Sydney is the most populous city in Australia and Oceania. Most times, when people talking about Sydney, they actually mean Greater Sydney (Greater Capital City Statistical Area). So does this report. As of June 2019, Greater Sydney's estimated metropolitan population was 5,312,163. Based on a total site area of 12,368.2sq km, the current population density of Greater Sydney area is 430 persons per square kilometer. The built urban area is estimated at 4,196sq km which translates to a density of 1,171 persons per square kilometer. Sydney is made up of 658 suburbs, 40 local government areas and 15 contiguous regions.[1] Being a resident of this city, I decided to use Sydney in my project.

2.2 PROBLEM

Although ranking top 10 in the Global Liveability Ranking list [2], the housing price in Sydney is relatively high and growing fast. When people making decision of purchasing a house, they usually want to buy a house with good growing potential through a lower price. At the same time, they may also want to choose the suburb according to the social places' density. Meanwhile, according to the 2016 census, 18.6% of the population were children aged between 0 to 14 years, those with children in the family may also consider education resource in the suburb.

2.3 INTEREST

When we consider all these problems, we can create a map and information chart where the real estate index is placed on Sydney and each suburb is clustered according to the housing price growing rate, venue density and number of schools with OC.

3 Data acquisition and cleaning

3.1 DATA SOURCES

- I got the Sydney property prices from 2000 to 2019 from Kaggle dataset. [3]
- I got the suburb data of Great Sydney from Corra [4].
- I got the list of Opportunity Classes from education.nsw.gov.au [5].
- I used Foursquare API to get the most common venues of given Borough of Sydney [6].

3.2 DATA CLEANING

- The housing price data is the main fact data for this report. I dropped NA values and extremely low prices in sell price column. Trimmed the suburb column by Pandas strip method. In addition, I added a Year column for further use.
- The OC school list is scraped from the government website using BeautifulSoup. I got the suburb column by splitting the original school location column. Then

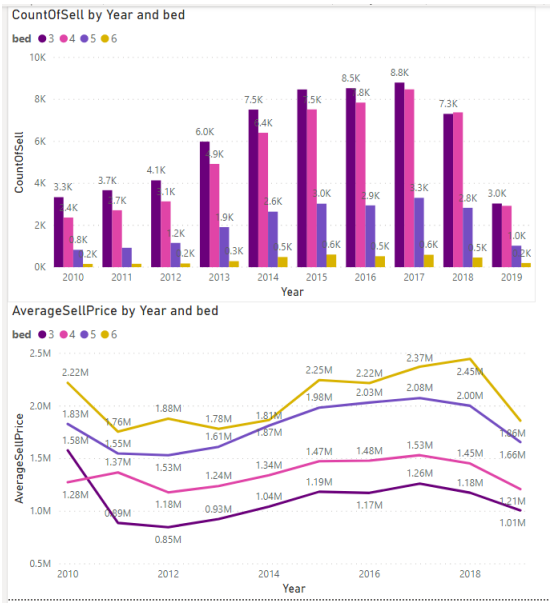
count OC school number by suburb. So, I got the suburb and count of OC data frame.

- The geo data is download from corra.com.au. I kept row with state equals to NSW and type equals to Delivery Area. The latitude and longitude columns are clean and nice. While suburb column in this dataset is all in Upper word, so I used title method to capitalize each word. Thus, it is the same with housing price dataset.
- In order to get venue data of each suburb, I created a function to get venue data according to given latitude and longitude info through Foursquare API.

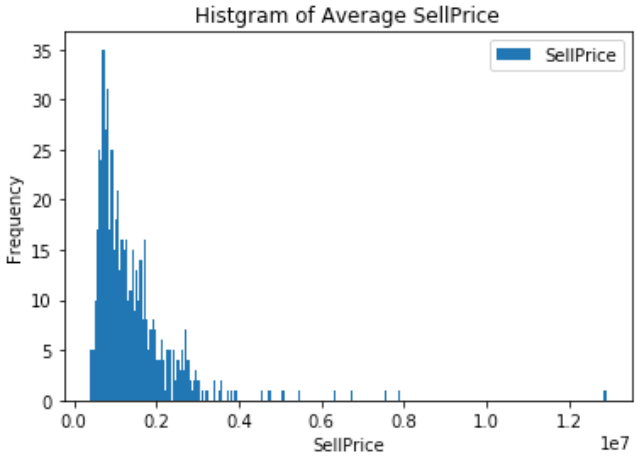
Thus, the data cleaning is almost done.

3.3 FEATURE SELECTION

Firstly, I used the describe method to look into housing price data frame and also use some visualization plots to see what should be included. As can be seen in the plot below. The number of bedrooms lead to differences in price, but the changing trends are similar. So I decided not to use the bed column in this report. Only keep the year, suburb and sell price column.



Secondly, considering that the latest price contains more information indicating the current situation. I calculated average sell price and yoy change rate by year and only selected Year 2019 for analyzing use. Using histogram I found that the housing price is close to normal distribution as followed.



So I normalized the sell price and yoy change rate using scikit learn preprocessing scale method. Then I got a data frame with the following columns.

	suburb	SellPrice	IncreaseRate	SDIR	SDSell
0	Abbotsbury	890091.0	-0.233692	-0.497716	-0.729076
1	Abbotsford	2490000.0	-0.039475	1.074611	0.154194
2	Agnes Banks	745000.0	-0.073095	-0.640306	0.001297
3	Airds	410462.0	-0.338333	-0.969077	-1.204970
4	Alexandria	1541867.0	-0.000998	0.142823	0.329185

Thirdly, I appended OC school number and geo info to the above price data frame. Replaced the nan values with zero.

Fourthly, using foursquare API, I got the venue details within 1000 meters around each suburb. The limitation is 100 venues per suburb. Then I reorganized the venue category data.

	Neighborhood	Venue Category
0	Abbotsbury	Convenience Store
1	Abbotsbury	Athletics & Sports
2	Abbotsbury	Park
3	Abbotsbury	Supermarket
4	Abbotsbury	Shopping Mall

Finally, I merged a data frame that contains normalized sell price, increase rate, number of OC schools, venue numbers by categories. Which is a data frame of 545*377. It could be used for further clustering now.

4 Methodology

For I already have the sell price, yoy increase rate, numbers of OC schools and venue proportion by categories for each suburb in Sydney. I chose to use K-Means clustering to figure out something behind the data. K-Means algorithm is one of the most common cluster methods of unsupervised learning. It clusters data points automatically by their distance. This gives us a chance to look into characteristics of each cluster and provide support for choice.

Meanwhile, as I already have latitude and longitude data for each suburb. I will use folium library to visualize geographic details of each suburb by cluster labels. This would provide a clear visual for story telling.

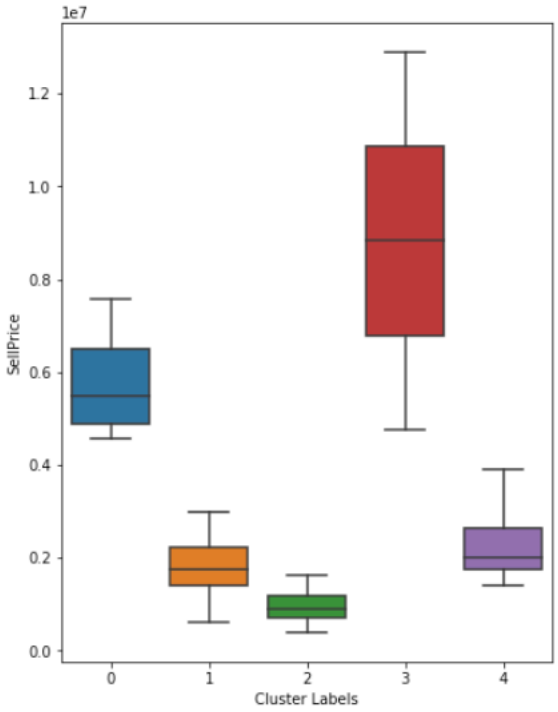
5 Exploring

5.1 GENERAL RESULT

After clustered with K-Means algorism, I appended the cluster label to the data frame. Firstly, explore the description statistics of the data frame and get the following result. The cluster result is significant. 66.3% of the suburbs fell in cluster 2, 18.3% percent in cluster 4, 13.9% in cluster 1 and very few in cluster 0 and 3.

		SellPrice	IncreaseRate	Venue Category	CountOfOC
	count	mean	mean	mean	mean
Cluster Labels					
0	7.0	5768744.00	0.33	53.00	0.00
1	79.0	1813349.73	0.13	31.27	0.05
2	377.0	944132.20	-0.12	17.62	0.10
3	2.0	8825000.00	1.81	32.00	0.00
4	104.0	2229687.28	-0.18	32.96	0.06

The boxplot of sell price for each cluster is as follows (without outliers):



5.2 LOOK INTO EACH CLUSTERS

5.2.1 Cluster 0 and 3

The suburbs in cluster 0 and 7 are all with high average sell price and high venue dense. So put them together to have a look. These are the most expensive suburbs in Sydney which are for the real rich people. Kids in these families usually go to private school so they don't need OC schools. This also make sense for normal knowledge. Look into the most popular venues in these areas, we can see that Café, Beach and Park are the most popular.

	Neighborhood	Cluster Labels	Neighborhood	SellPrice	IncreaseRate	SDSell	SDIR	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
43	Bellevue Hill	0	Bellevue Hill	7560000.0	-0.080228	-0.031143	6.057205	Café	Coffee Shop	Pizza Place
148	Darling Point	0	Darling Point	6300000.0	0.091981	0.752039	4.818927	Café	Park	Japanese Restaurant
283	Lavender Bay	0	Lavender Bay	6710000.0	0.409664	2.196820	5.221859	Café	Park	Coffee Shop
400	Point Piper	3	Point Piper	12900000.0	1.345455	6.452674	11.305144	Beach	Café	Harbor / Marina
402	Potts Point	0	Potts Point	4700000.0	1.140011	5.518346	3.246511	Café	Italian Restaurant	Australian Restaurant
420	Rhodes	3	Rhodes	4750000.0	2.275862	10.684046	3.295649	Café	Fast Food Restaurant	Australian Restaurant
429	Rose Bay	0	Rose Bay	5055875.0	0.332773	1.847132	3.596250	Café	Park	Supermarket
494	Vaucluse	0	Vaucluse	5482000.0	-0.183256	-0.499702	4.015029	Beach	Lighthouse	Park
517	Whale Beach	0	Whale Beach	4573333.0	0.633333	3.214039	3.122027	Bakery	Burger Joint	Restaurant

5.2.2 Cluster 1

Cluster 1 contains 79 suburbs with a positive increase rate of 13% and venue density is 31 within 1000 meters. 4 of 79 have OC schools. The average sell price of this cluster is 1.8 million.

5.2.3 Cluster 2

contains 377 suburbs which is the biggest portion of 5 clusters. The increase rate is -12% indicates that the sell price decreased from 2018 to 2019. The venue density is 18 within 1000 meters. 38 of the OC schools are in these suburbs. The average sell price of this cluster is 0.94 million.

5.2.4 Cluster 4

contains 104 suburbs. The increase rate is -18% which is the lowest in all clusters. The venue density is 33 within 1000 meters. 6 of the OC schools are in these suburbs. The average sell price of this cluster is 2.23 million.

5.3 LOOK AT SUBURBS IN MAP

5.3.1 The blue and red points

They are for cluster 0 and 3 and they all located near beaches with ocean or river view.

5.3.2 The purple points

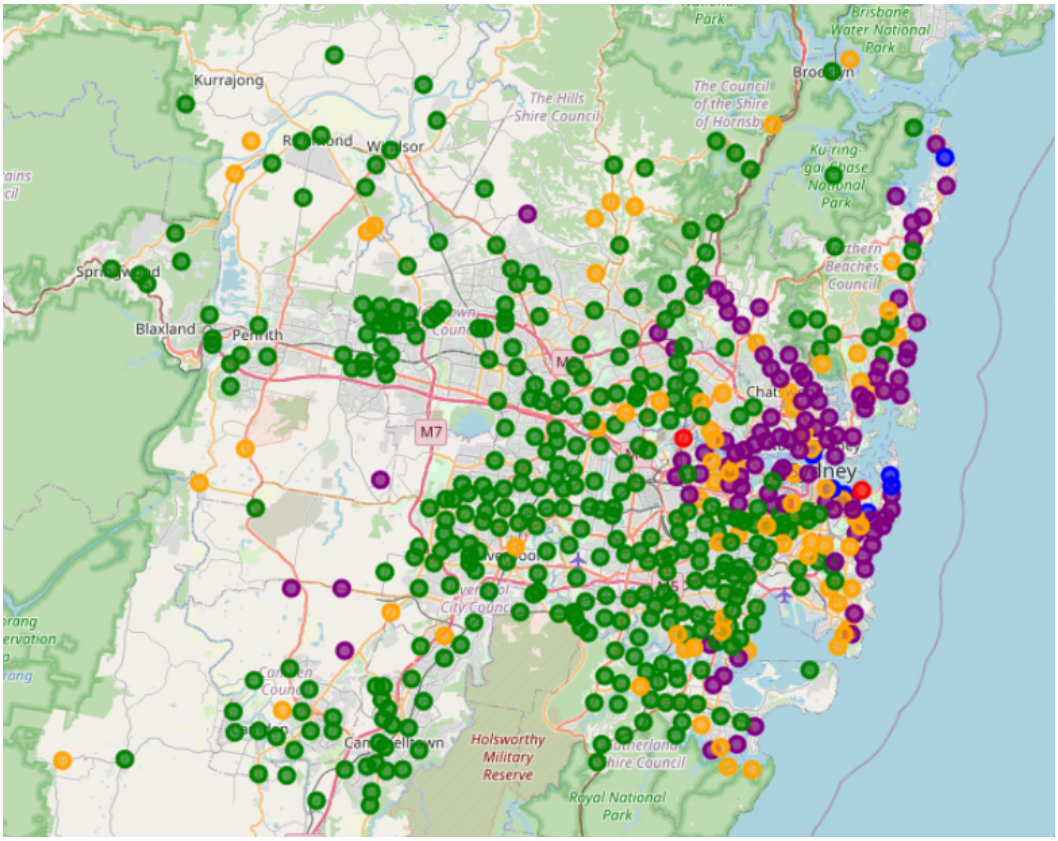
They are for cluster 4. They located near CBD of Sydney and the housing price is relatively high. But the increase rate is the lowest. This indicates that they are not as popular as years before. People are moving away from city center. Maybe the transportation getting more and more convent is a reason for this. But will not discuss in this report.

5.3.3 The orange points

They are for cluster 1. With a medium sell price and positive increase rate. Also have good venue density and mostly locate not far away from CBD.

5.3.4 The green points

They are for cluster 2. The average sell price is the lowest. There are less venues and far away from CBD. So these may contributes to the low sell price.



6 Conclusion

Personally, I feel this report is helpful in making decision of buying a house. For me, suburbs in cluster is I is the most suitable. It is convenient with good venue density and not far away from CBD. It is not very expensive compared with other clusters except cluster 2. It has a good increase rate for investment.

As a result, people decided to buy a house or want to start a venue in Sydney can have a look at planforms providing such information as sell price, increase rate, venue density and categories. People can achieve better outcomes through their access to the platforms where such information is provided.

7 Appendix

[1] City of Sydney(www.cityofsydney.nsw.gov.au)

[2] Global Liveability Ranking-Wikipedia

[3] Sydney property prices from 2000 to 2019(https://www.kaggle.com/mihirhalai/sydney-house-prices/data?select=SydneyHousePrices.csv) cleaned by Mihir Halai.

[4] Database of Australian Postcodes(https://www.corra.com.au/australian-postcode-location-data/).

[5]Opportunity Class list (https://education.nsw.gov.au/public-schools/selective-high-schools-and-opportunity-classes/year-5/what-are-opportunity-classes/list-of-opportunity-classes)

[6] Forsquare API