Liveable Sydney

Lee Ping Tan

25 March 2019

1. Introduction

1.1 Background

Sydney is the state capital of New South Wales and the most populous city in Australia and Oceania. Sydney is made up of 658 suburbs, 40 local government areas and 15 contiguous regions.

According to Demographia International Housing Affordability Survey, Sydney is one of the world's most expensive city to live, trailing only Hong Kong. There were 1.76 million dwellings in Sydney in 2016 including 925,000 (57%) detached houses, 227,000 (14%) semi-detached terrace houses and 456,000 (28%) units and apartments. The median house price is $1,177,600 whereas the median income is $91,600. However, since the December quarter, Australian property values fell $133.1 billion with Sydney leading the annual residential property price falls of 7.8%. The current downturn is Sydney's worst since the 1980s, according to BIS Oxford Economics. With the forecast of housing price fall to continue, this seems like the best time for Sydneysiders to fulfil their Great Australian Dream (a belief that in Australia, home-ownership can lead to a better life and is an expression of success and security).

1.2 Problem

Although this is a good time to purchase, what is the suburb to buy into? There might be some suburbs that are having similar features but are slightly cheaper than its neighbouring suburbs. Besides, some suburbs might have a price decrease that is slightly more significant than its similar suburbs in the current downturn. How should we determine the similarity of suburbs? The best suburb is not necessarily based only on the housing cost but the liveability of the suburbs. According to domain.com.au, the factors that determine the liveability of a suburb includes access to employment, transportation, culture, education, shopping, café and restaurants, park and recreation, beach access, crime rate, etc. The purpose of this study is to segment Sydney suburbs according to its liveability factors, compare the median house price among similar suburbs and identify a better value suburb, if any.

1.3 Interest

The target audience for this study is potential property buyer, property investor and real estate agents.

2. Data acquisition and cleaning

a. Data access

I have gathered data from different data sources as there are no single point to obtain all the information required for this study. After collecting all the data, I have to merge them into one dataset based on the suburb and postcode from different data sources. The data required with their corresponding data source in this study are:

- List of Sydney suburbs with postcode
  The postcode is required to obtain the property price for each suburb. The best data source I have found for this purpose is from JustWeb site.

- Population & density of each suburb
  I could not find a data source that contains population of all Sydney suburbs. In the end I have to scrape the information from Wikipedia by passing each suburb name into the Wiki url. Example, https://en.wikipedia.org/wiki/Abbotsbury,_New_South_Wales for Abbotsbury, https://en.wikipedia.org/wiki/Warrawee,_New_South_Wales for Warrawee, etc.

- Distance from each suburb to Sydney CBD
  Distance to Sydney CBD is one of our feature selection. This is obtained from digital advocates.

- Geocode for each suburb
  Geocode is required to gather information from FourSquare location data, which is used to calculate our feature selection. This is obtained from geocoder package.

- Property price for each suburb
  This is obtained by using BeautifulSoap package to scrape property pricing data from realestate site based on the suburb name and postcode.

- Crime rate for each suburb
  The crime rate of each suburb is obtained from NSW Bureau of Crime Statistics and Research. Since the population data is from the year 2016. The crime rate has been calculated based on the number of offences in the year of 2016.

- Important venue for each suburb
  The venues used for the suburb segmentation is obtained and analysed from FourSquare location data.

b. Data cleaning

In this study, I am focusing on suburbs which are 30km from Sydney CBD. As distance from CBD is one of our feature selection, it makes sense to omit any data without this feature value. As the population of each suburb is obtained from scraping from Wiki page, some suburbs might have missing or incorrect or zero population value. I have removed these samples from the study. When I was scraping property price data from a real estate website, there are a few suburbs with missing profile page. These suburbs have been removed from my study. There are some suburbs that do not have any property price data even though the suburb profile exist. Suburbs without property

price data usually mean they do not have much activities in buying or selling. Therefore, they should not be included in the study.

The venues in FourSquare location are divided into different categories. Each main category is further divided into many level of sub categories. Usually, the category associated with each venue is a lower level sub category. If I am to use this lower level sub category which has a very big varieties, it will not fit properly into the features that I have chosen for this study. Therefore, I have matched each sub category associated with the venue to its first level of category, which are Food, Shop & Service, Travel & Transport, Outdoors & Recreation, College & University, Professional and Other Places, Arts & Entertainment and Nighlife Spot. The list of FourSquare venue categories is available here.

c. Feature selection

In this study, I am going to use the liveability factors of a suburb as my feature selection. The liveability indicators that are used are:

- **Access to employment**

Sydney central business district (also Sydney CBD, and often referred to simply as "Town" or "the City") is the main commercial centre of Sydney, the state capital of New South Wales and the most populous city in Australia. The city centre employs approximately 13% of the Sydney region's workforce. Based on industry mix and relative occupational wage levels it is estimated that economic activity (GDP) generated in the city in 2015/16 was approximately $118 billion.

Therefore, in this study, I am going to include the proximity of each suburb to Sydney CBD. The smaller the proximity, the lesser the traveling time, and the higher the access to employment.

- **Transportation**

According to Sydney Public Transportation Statistics 2017-18, train service is the most common mode of transportation followed by bus. In this study, I am going to focus on the number of train and bus services within 500 radius of the suburb.

- **Education**

Schools are very important for family with children. The closer they are to your house the more convenient it is. Besides, the more primary and high school options, the better a suburb is.

- **Shopping**

The number of shopping centers determine how convenient a suburb is.

- **Café and Restaurants**

Eating out is a popular pastime in Australia and we have a huge choice of fabulous restaurants, cafes, pubs and bars in our cities and towns. Therefore, a suburb with abundant of good cafés and restaurants is definitely more liveable.

- **Park and Recreation**

  Due to Australia's warm climate, many Australian spend large amount of time on outdoor activities like cycling, bush walking, barbecuing, playing cricket or spending time with kids in the playgrounds. Suburbs with easy access to parks and playgrounds are more attractive.
- **Crime Rate**

  By common sense, a safe suburb with low crime rate should have a higher house price compared to suburbs with lots of crimes.

- **House Median Price**

  In Australia, the most common property types are house, and units (which includes apartment, town house and villa). In this study, I am focusing only on houses.

- **Density**

  Density is calculated by dividing the population of a suburb by its area (km).

3. Data Exploratory Analysis

3.1 Descriptive Statistic & Boxplots

First, I explore the data by using the descriptive statistic. Based on the summary below, I noticed that the dataset have some outliers that need further analysis and action. Population has a maximum value of 29,822 whereas its mean is only 10,671.

The next variable to be explored further is House Median Price with a max value of more than 5 million but a mean value of only 1.77 million.

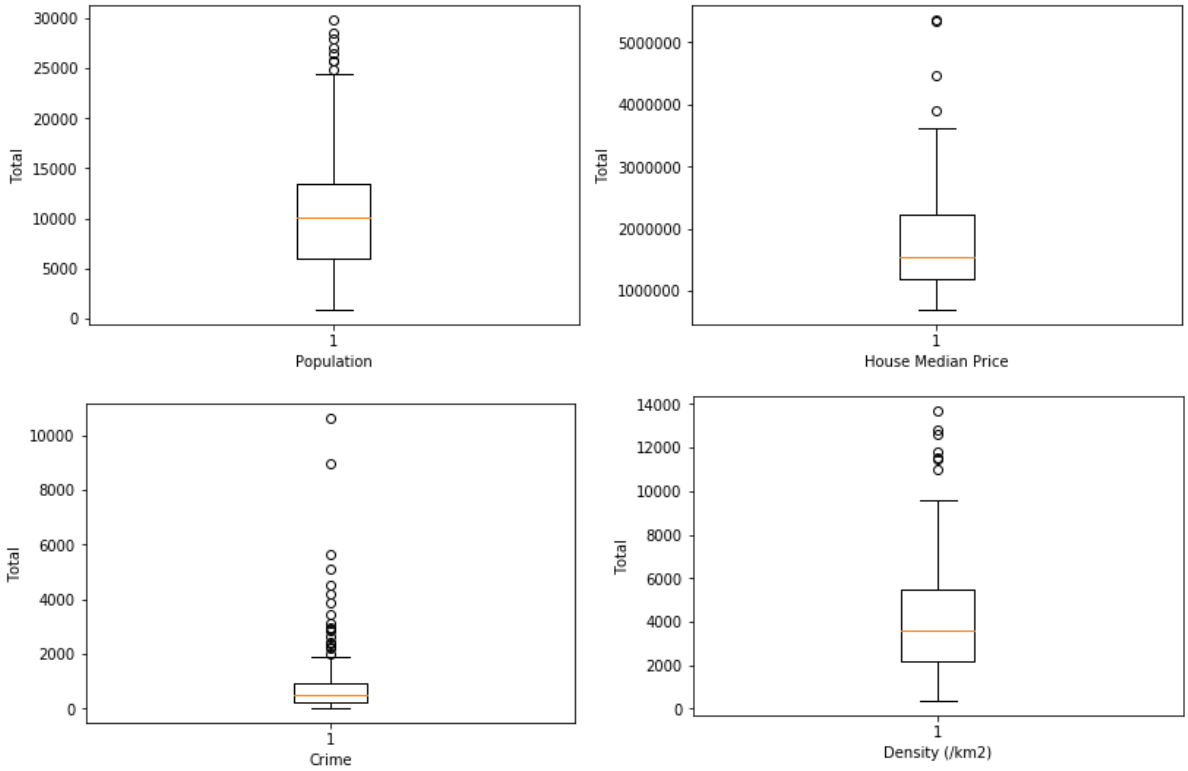
The third variable is Total Crime with a max value of 16,453 and a mean value of only 968.

The forth variable is Density (/km2). Density is calculated from Population divided by Area (km2). In the descriptive summary, it shows a max density of 13,677 and a mean of 4201.

All the other variables that I have retrieve from FourSquare need further analysis too. There are Arts & Entertainment, College & University, Food, Nightlife Spot, Outdoors & Recreation, Professional & Other Places, Shop & Service and Travel & Transport.

	Postcode	Distance (km)	Density (/km2)	Population	Total Crime	House Median Price	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport
count	134.000000	134.000000	134.00000	134.000000	134.000000	1.340000e+02	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000	134.000000
mean	1982.761194	12.077612	4201.12000	10671.477612	968.746269	1.773083e+06	0.447761	0.007463	13.500000	1.619403	1.656716	0.052239	3.104478	0.873134
std	229.576930	7.877852	2792.39196	6566.121473	1513.534765	8.653493e+05	1.073027	0.086387	15.417864	3.045506	1.798514	0.223343	3.414407	1.185258
min	1300.000000	1.000000	363.00000	860.000000	22.000000	6.950000e+05	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2014.500000	5.450000	2195.00000	5999.750000	248.500000	1.196875e+06	0.000000	0.000000	2.250000	0.000000	1.000000	0.000000	1.000000	0.000000
50%	2066.000000	9.400000	3617.00000	10080.000000	487.500000	1.550000e+06	0.000000	0.000000	8.500000	1.000000	1.000000	0.000000	2.000000	1.000000
75%	2119.750000	17.925000	5471.25000	13453.250000	907.000000	2.217500e+06	0.000000	0.000000	17.000000	1.750000	2.000000	0.000000	4.000000	1.000000
max	2234.000000	29.900000	13677.00000	29822.000000	10620.000000	5.360000e+06	7.000000	1.000000	77.000000	18.000000	13.000000	1.000000	18.000000	7.000000

When a boxplot is plotted for all the above variables, I noticed that there are quite a lot of outliers. Actually, not all outliers should be removed from the dataset. Some domain knowledge is required to determine whether an outliers should be kept or removed. Since, I do not have any domain knowledge in this area, I have removed all outliers to keep things simple.



Using IQR to calculate the outliers, I was able to retrieve the outlier data and have them removed from the samples.

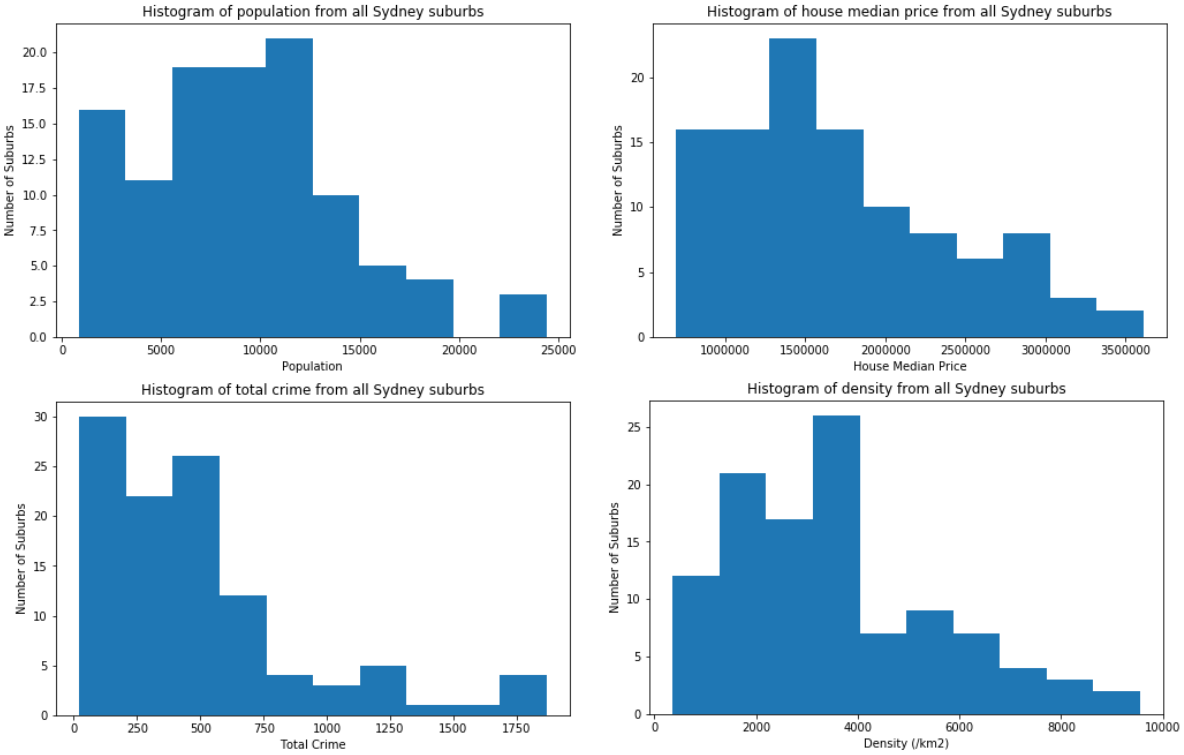
```
population_Q1 = df_sydney['Population'].quantile(0.25)
population_Q3 = df_sydney['Population'].quantile(0.75)
population_outliers = (df_sydney['Population'] < (population_Q1 - 1.5 * IQR)) | (df_sydney['Population'] > (population_Q3 + 1.5 * IQR))

house_price_Q1 = df_sydney['House Median Price'].quantile(0.25)
house_price_Q3 = df_sydney['House Median Price'].quantile(0.75)
house_price_outliers = (df_sydney['House Median Price'] < (house_price_Q1 - 1.5 * IQR)) | (df_sydney['House Median Price'] > (house_price_Q3 + 1.5 * IQR))
```

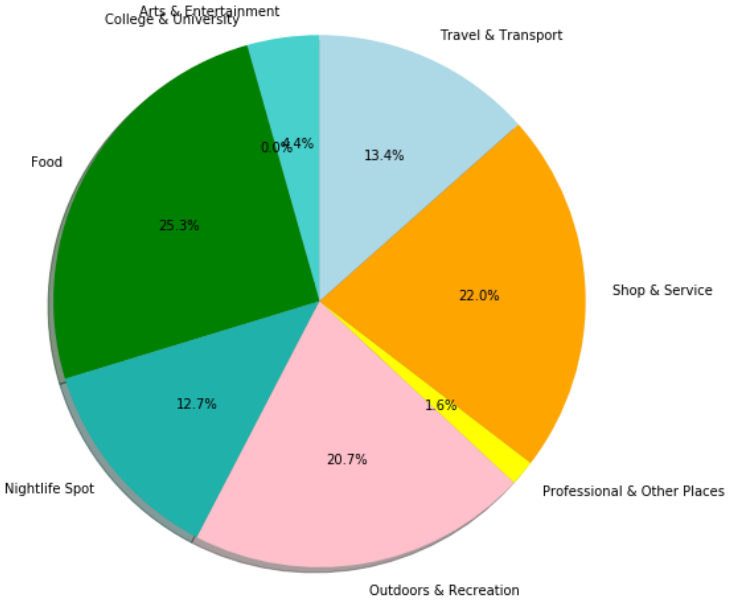
```
crime_Q1 = df_sydney['Total Crime'].quantile(0.25)
crime_Q3 = df_sydney['Total Crime'].quantile(0.75)
crime_outliers = (df_sydney['Total Crime'] < (crime_Q1 - 1.5 * IQR)) | (df_sydney['Total Crime'] > (crime_Q3 + 1.5 * IQR))

density_Q1 = df_sydney['Density (/km2)'].quantile(0.25)
density_Q3 = df_sydney['Density (/km2)'].quantile(0.75)
density_outliers = (df_sydney['Density (/km2)'] < (density_Q1 - 1.5 * IQR)) | (df_sydney['Density (/km2)s'] > (density_Q3 + 1.5 * IQR))
```

After the outliers are removed, the population variable shows a more even distribution as following histogram:



Then the total number of venues from FourSquare are counted as the table below. Out of 134 samples, there are many suburbs that have zero count for more than one venues. Zero count could mean 2 possibilities, either the venue does not exist in the suburb or there are no data available for that particular venue for the suburb. The first possibility is what I was looking. For example, if a suburb does not have any nightlife spot, this suburb should be more family friendly, and so forth. However, if it falls into the 2nd possibilities, then these samples should not be included in the study.



From the above pie chart, noticed that Food, Shop & Service and Outdoors & Recreation are 3 main venues that have the highest percentage of having venue count more than zero. For other venues, they either does not have any data or they have zero appearance in most suburbs.

Next, I have to determine whether to include these venues into my feature selection. If the venue is included into the feature selection, how should I treat the zero count. Should I treat it as missing data or should I treat it as true zero count?

Since food is the most important aspect of life, I have used food venue as a cut off point. The assumption that I have made is all suburb must have a least one food venue. Those with zero count are treated as missing data. Once, food venue with missing data (zero count) are removed, all zero count for other venues are treated as true zero count.

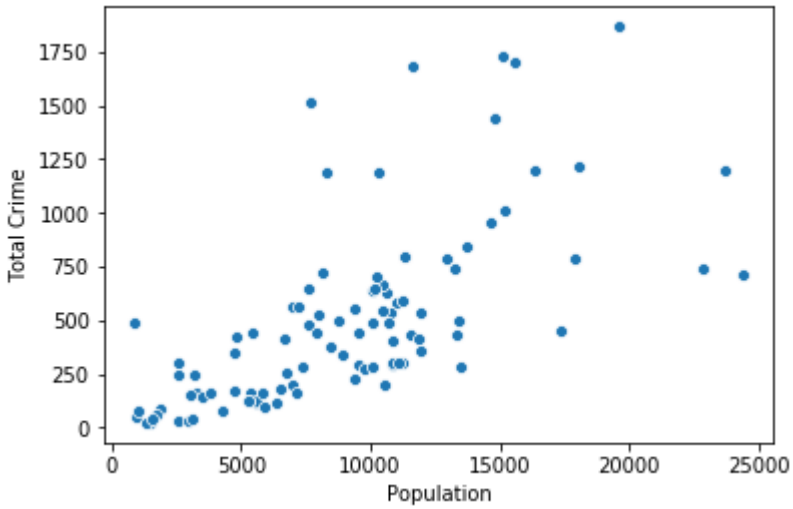
Besides, due to insufficient data, the following venues have been removed from my feature selection:

- College & University
- Arts & Entertainment
- Professional & Other Places
- Nightlife Spot
- Travel & Transport

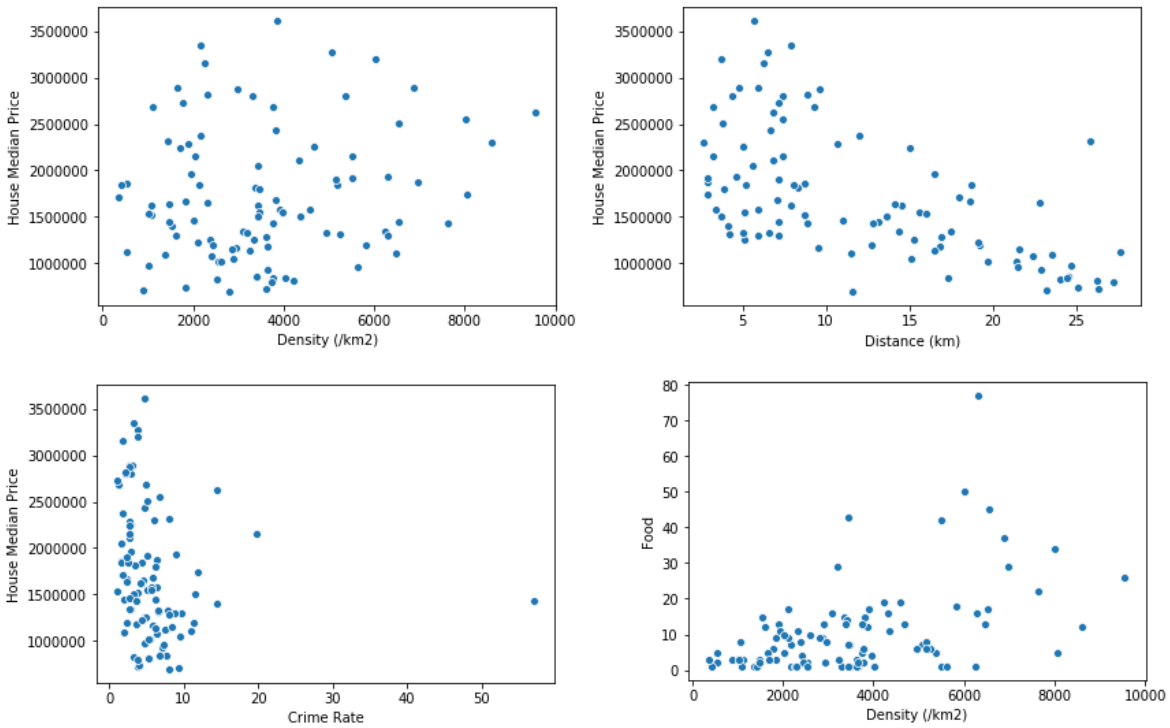
3.2 Relationships between population and total crime

I noticed that there is a strong relationship between population and total crime as shown in scatter plot below. It makes sense that if a suburb have move people, it should has more number of offences. As such, total crime variable is removed from the feature selection and

replaced by crime rate, a percentage which is calculated by dividing total crime with population.



3.3 Scatter plots that show relationship between different variables



4. Modelling

There are few methods for clustering, namely K-Means, Hierarchical and DBSCAN. K-Means is among the most popular clustering methods whereas DBSCAN is popular because it could detect outliers. In this study, I have used both K-Means and DBSCAN methods.

4.1 Principal Component Analysis (PCA)

Before, K-Means/DBSCAN clustering is applied, I used PCA to scatterplot my samples in a 2 dimensional view. Principal Component Analysis or PCA is a linear feature extraction technique. It performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. PCA combines your input features in a specific way that you can drop the least important feature while still retaining the most valuable parts of all of the features. As an added benefit, each of the new features or components created after PCA are all independent of one another.

4.2 K-Means Clustering

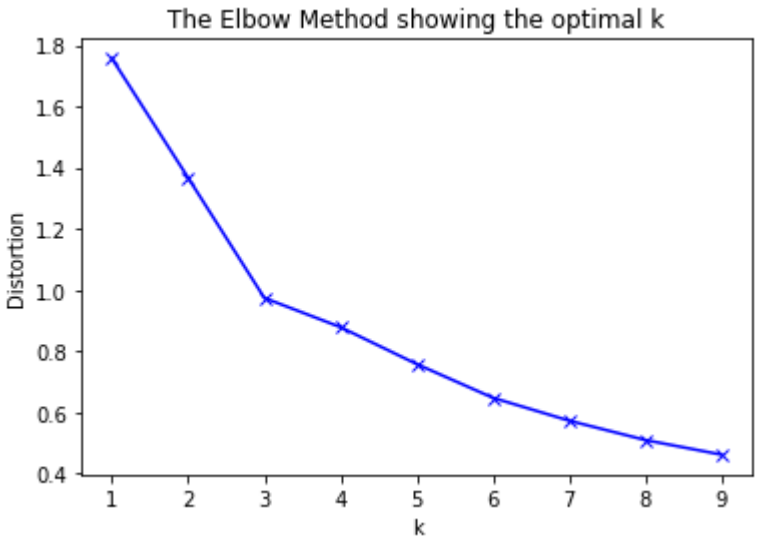
I have tested the model with different features combination and found out the following combination yield a better results.

- House Price Level
- Distance (km)
- Density (/km2)
- Food
- Outdoors & Recreation
- Shop & Service
- Crime Rate

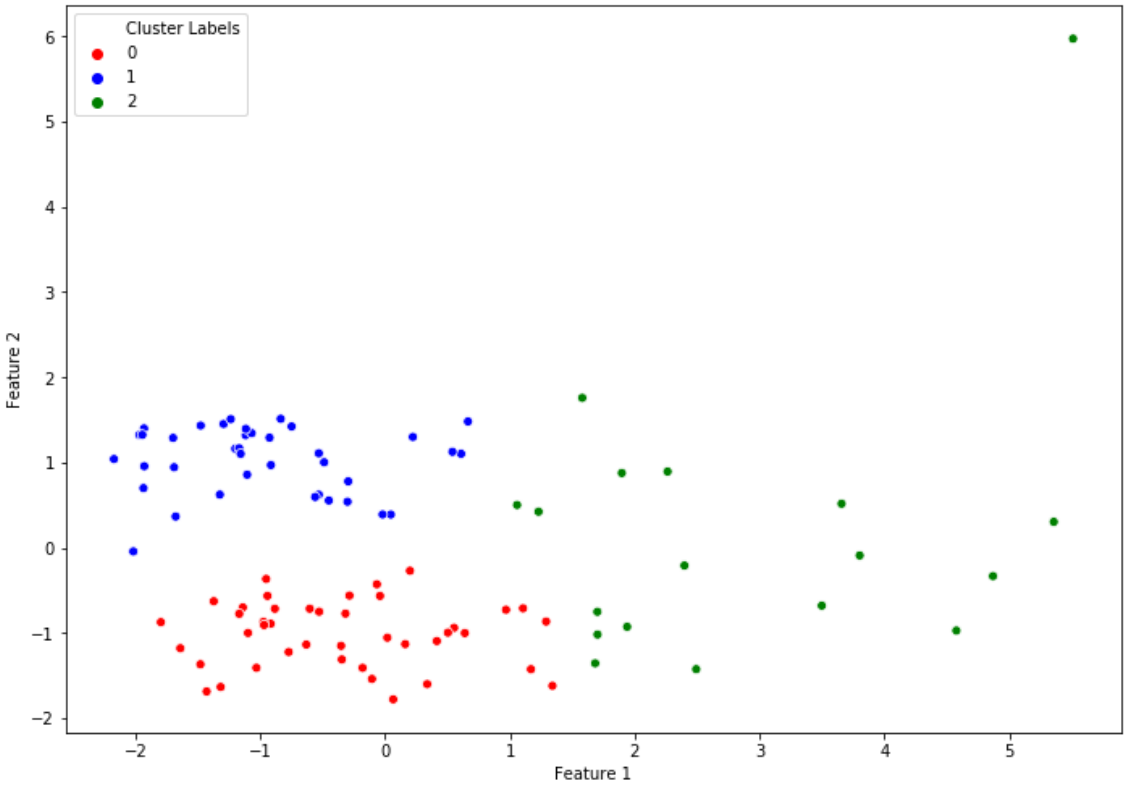
House Price Level is derived from House Median Price. House Price Level has been categorized into 3 levels as follows:

House Price Level	Min	Mean	Max
0	1,502,500	1,838,938	2,315,000
1	2,375,000	2,857,316	3,612,500
2	695,000	1,118,721	1,460,000

I have used elbow method to determine the optimum k for my model. Elbow method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. However, using my dataset, I could not really find a very optimum k. I have chosen the 3 as my k value where the curve form an elbow and flatten out.



The following scatter plot shows how the dataset is clustered after k-means clustering.

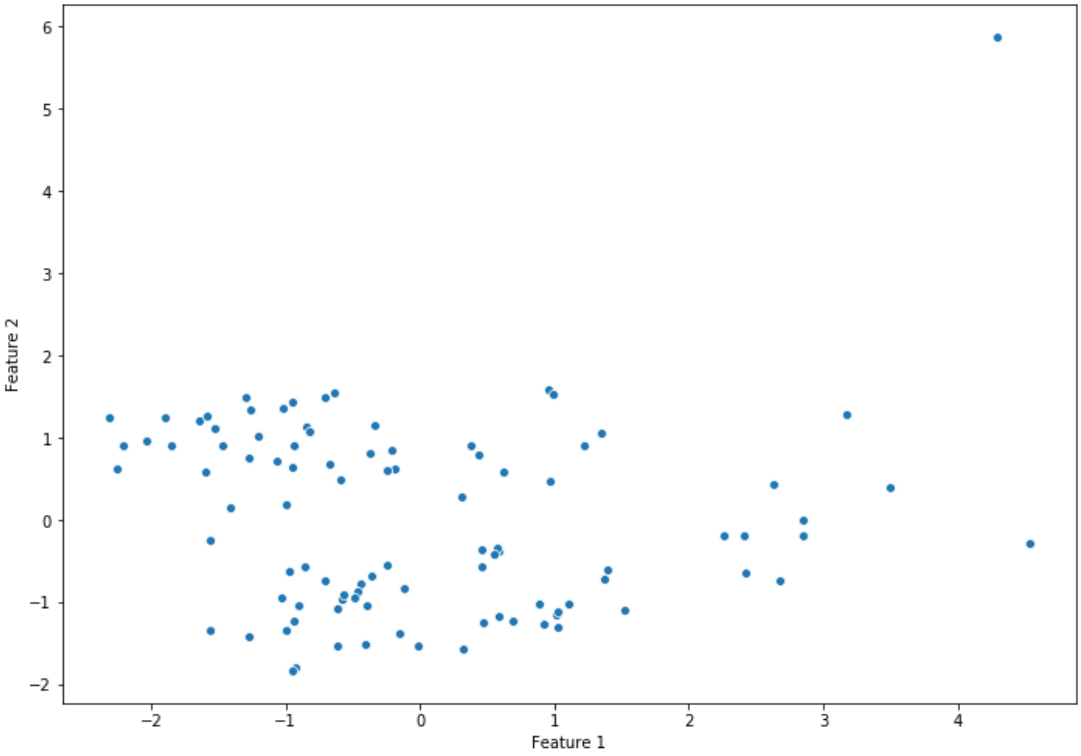


4.3 DBSCAN

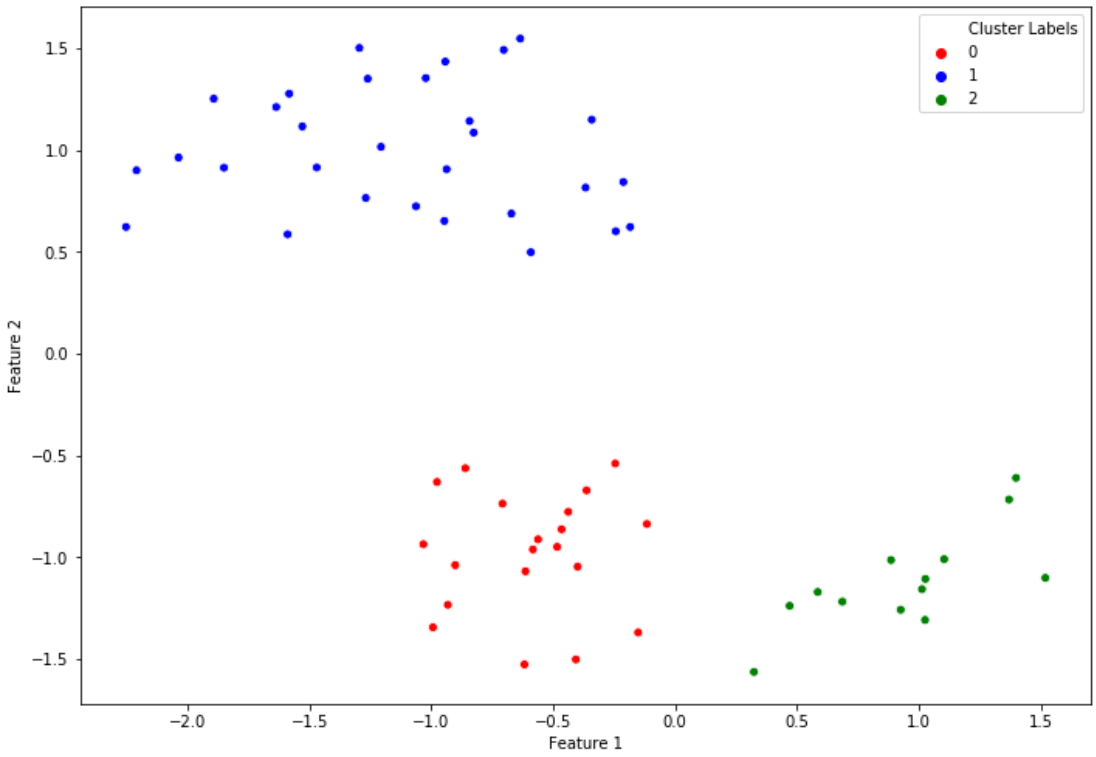
The features used for DBSCAN are:

- House Price Level
- Distance (km)
- Density (/km2)

- Food
- Crime Rate



The above scatter plot shows how the data points before clustering compared to the scatter plot below after DBSCAN clustering. Noticed that some outliers have been removed.



5. Results

Both K-Means and DBSCAN are showing similar results, where the suburbs have been grouped into 3 clusters. However, DBSCAN is showing less suburbs in the end results as suburbs that are considered outliers have been removed during the clustering process.

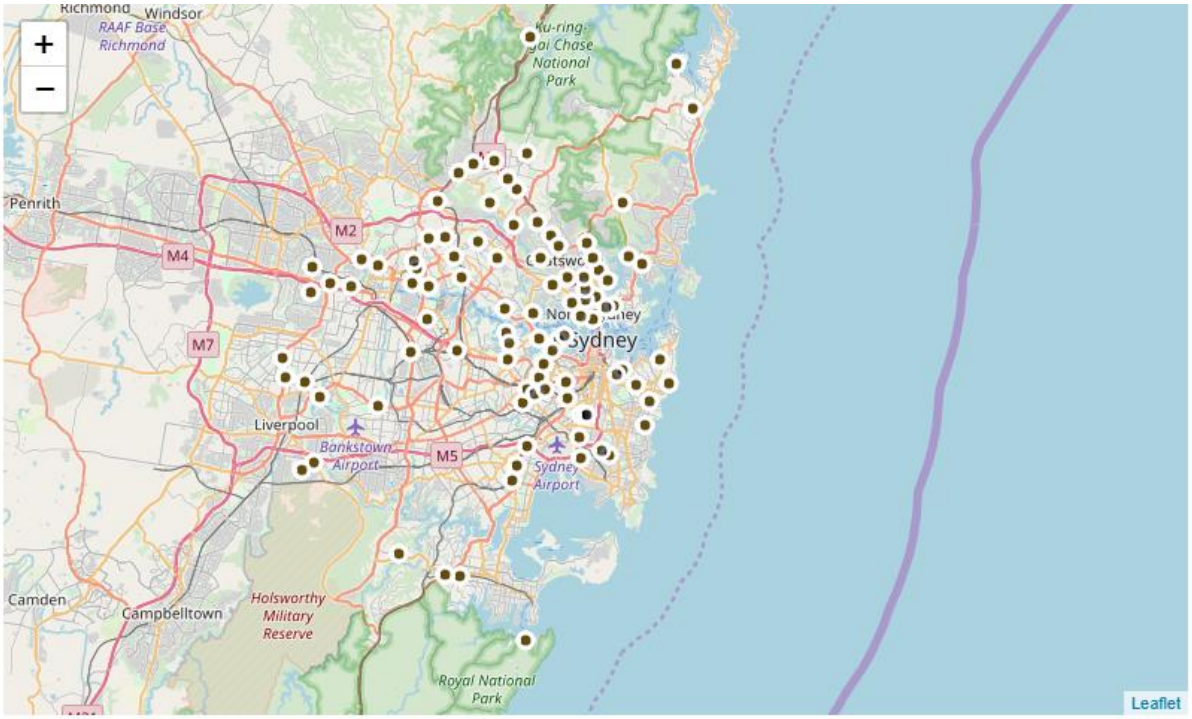
The comparison among the clusters' criteria from K-Means modelling is shown below.

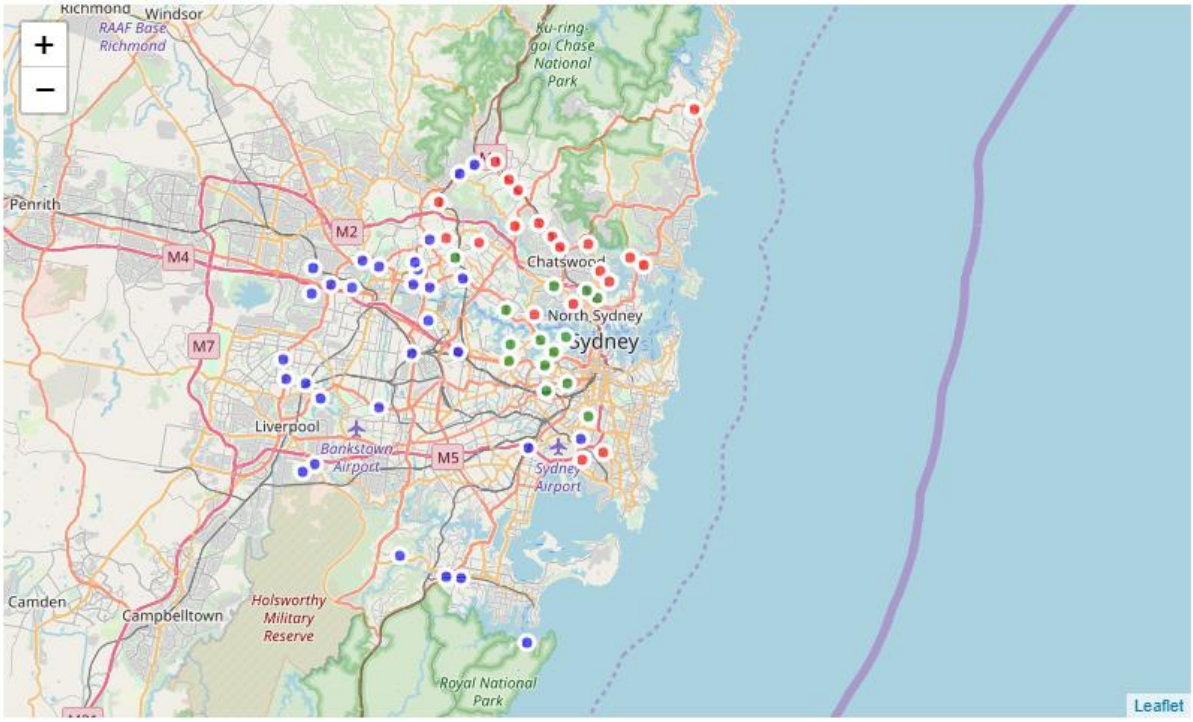
Cluster 1	Cluster 2	Cluster 3
House Price Level 0.0 Distance (km) 9.0 Density (/km2) 2804.0 Food 8.0 Outdoors & Recreation 1.0 Shop & Service 2.0 Crime Rate 3.0	House Price Level 2.0 Distance (km) 18.0 Density (/km2) 2955.0 Food 6.0 Outdoors & Recreation 1.0 Shop & Service 2.0 Crime Rate 6.0	House Price Level 1.0 Distance (km) 5.0 Density (/km2) 6518.0 Food 27.0 Outdoors & Recreation 4.0 Shop & Service 6.0 Crime Rate 11.0
▪ Medium house price ▪ 2nd closest to city ▪ Low crime rate	▪ Low house price ▪ Farthest from city ▪ Medium crime rate	▪ High house price ▪ Closest to city ▪ High density ▪ High number of food venues ▪ High number of shopping venues ▪ High number of recreations

The comparison among the clusters' criteria from K-Means modelling is shown below.

Cluster 1	Cluster 2	Cluster 3
House Median Price 2261408.0 Density (/km2) 2104.0 Distance (km) 12.0 Crime Rate 3.0 Food 6.0	House Median Price 1032350.0 Density (/km2) 3029.0 Distance (km) 20.0 Crime Rate 6.0 Food 6.0	House Median Price 1803500.0 Density (/km2) 4229.0 Distance (km) 6.0 Crime Rate 5.0 Food 13.0
▪ High house price ▪ Low density ▪ 2nd closest to city ▪ Lowest crime rate	▪ Low house price ▪ Medium density ▪ Farthest from city	▪ Medium house price ▪ High density ▪ Closest to city

Using the results from DBSCAN, I have grouped the suburbs in folium map. The first map shows all the Sydney suburbs in my study and the 2nd map shows the grouping of Sydney suburbs into 3 clusters. Noticed that some suburbs are missing in the 2nd map as explained earlier in the report.





6. Conclusion

The most time consuming task in this study is data collection and cleaning. As there is no single source available, I have to look to different sources with data in a variety of format. Besides, there are a lot of missing and incorrect data. I do not have a domain expert to guide me in the

study and have to make a lot of assumption in the process. Another main issue in this study is the use of FourSquare location data to obtain the numbers of top venues for each suburb. The venue usually only exist if it is popular and has been recommended by FourSquare users. For example, it is more likely to find venues like café and restaurants than to find schools in FourSquare location data. However, one of the main objective of our study is to calculate the liveability of a suburb which is highly dependent on important venues like café & restaurant, transportation and good schools, to name a few. Due to insufficient data available for this category of venue, I have omit them from my feature selection which indeed have a significant impact to my end results. Although I have managed to segment Sydney suburbs after different combination of feature selection and comparison of different clustering models, there is definitely a big room for improvement.

7. Future Direction

The model for this study could be better improved by sourcing a more reliable data. Secondly, there are more features that we could include in the study. For example, one interesting feature that I could think of is the look and feel of the suburbs. One idea is to gather as many photos as possible for each suburb and perform include them as one of the feature for clustering.