# A better life across the ditch

**Kevin Jin**

**16th of May, 2019**

## 1. INTRODUCTION

### 1.1 Background

For decades, a steady flow of New Zealanders headed to Australia for better employment and lifestyle options. This trend last peaked in 2012 when the Australian economy was booming. Since then the numbers have declined, flat-lining (below 2000 long-term departures) since 2014. However since the start of 2018 the traditional trend has started to re-emerge. Annual net migration for the year to July 2018 was 63,800, according to new numbers released by Stats NZ. That is now down 8,600 from its peak of 74,200 in July 2017 – but still high by historic levels.

There are number of reasons why New Zealanders choose to migrate to Australia but the most significant ones are housing cost, wage, and job opportunities. But for those who failed to conduct research and gather information on life in Australia prior to moving often find upon arrival that the life style in Australia is vastly different from the life style in New Zealand and this often leads to unsatisfying experience especially given the fact that for majority of migrating population the reason for migration is for financial benefit but not necessarily for change in their living environment.

### 1.2 Interest

Therefore, the aim of this project it to come up with a possible solution for those who wishes to migrate to Australia with minimum amount of change in their living environment. I will try to address the above issue by performing clustering on data collected through the Foursquare API to determine the similarity between major cities of Australia and 3 suburbs of Auckland (Parnell, Henderson, and Rosedale) and make recommendation to people from Auckland who are thinking of moving over to Australia based on the result produced by the clustering.

## 2. DATA ACQUISITION

### 2.1 Data sources

Foursquare allows us to gather a list of recommended venues within a user defined radius of a specified location (e.g. Auckland, Parnell). But in order to have access to these information we must first provide

the latitude and longitude information of the cities. For this project, I obtained the location data from the following site: https://www.latlong.net/

## 2.2 Data pre-processing

Latitude and longitude information of major cities of Australia and 3 suburbs of Auckland on https://www.latlong.net/ had to be scraped from the website using a python module named Beautifulsoup. This information was then re packaged into a data frame format using Pandas to make it easier to work with the Foursquare API:

| | Borough | Neighbourhood | Latitude | Longitude |
|---|---|---|---|---|
| 0 | Queensland | Sunshine Coast | -26.650000 | 153.066666 |
| 1 | Queensland | Gold Coast | -28.016666 | 153.399994 |
| 2 | VIC | Melbourne | -37.840935 | 144.946457 |
| 3 | SA | Adelaide | -34.921230 | 138.599503 |
| 4 | TAS | Launceston | -41.429825 | 147.157135 |
| 5 | SA | North Adelaide | -34.906101 | 138.593903 |
| 6 | QLD | Townsville City | -19.258965 | 146.816956 |
| 7 | QLD | Cairns City | -16.925491 | 145.754120 |
| 8 | WA | Perth | -31.953512 | 115.857048 |
| 9 | VIC | Mildura | -34.206841 | 142.136490 |
| 10 | Greenvale | Ziyou Today | -37.649967 | 144.880600 |
| 11 | Coffs Harbour NSW 2450 | Coffs Harbour NSW 2450 | -30.296276 | 153.114136 |
| 12 | NSW | Orange | -33.283577 | 149.101273 |
| 13 | VIC | Bendigo | -36.757786 | 144.278702 |

Table 1. Location information for major Australian cities.

As stated in the previous section, Foursquare API allows the user to define the maximum amount of venues returned and radius of search for each location and for this project I have decided to limit the maximum number of venues returned to be 100 to prevent bigger cities from dominating the outcome while setting the radius of search to 1 km to enable smaller cities to return as many venues as possible.

| | Neighborhood | American Restaurant | Aquarium | Argentinian Restaurant | Art Gallery | Art Museum | Arts & Crafts Store | Asian Restaurant | Athletics & Sports | Australian Restaurant | BBQ Joint | Bagel Shop | Bakery | Bar | Basketball Court | Beach | Beer Bar | Beer Garden | Big Box Store |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Adelaide | 0.01 | 0.0 | 0.01 | 0.010000 | 0.0 | 0.0 | 0.010000 | 0.0 | 0.030000 | 0.00 | 0.0 | 0.020000 | 0.060000 | 0.0 | 0.00 | 0.00 | 0.01 | 0.000000 |
| 1 | Albury | 0.00 | 0.0 | 0.00 | 0.000000 | 0.0 | 0.0 | 0.021739 | 0.0 | 0.000000 | 0.00 | 0.0 | 0.000000 | 0.065217 | 0.0 | 0.00 | 0.00 | 0.00 | 0.021739 |
| 2 | Bankstown NSW | 0.00 | 0.0 | 0.00 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.00 | 0.0 | 0.018182 | 0.018182 | 0.0 | 0.00 | 0.00 | 0.00 | 0.000000 |
| 3 | Bendigo | 0.00 | 0.0 | 0.00 | 0.018519 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.018519 | 0.00 | 0.0 | 0.018519 | 0.018519 | 0.0 | 0.00 | 0.00 | 0.00 | 0.018519 |
| 4 | Brisbane | 0.00 | 0.0 | 0.00 | 0.020000 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.020000 | 0.01 | 0.0 | 0.000000 | 0.010000 | 0.0 | 0.01 | 0.01 | 0.00 | 0.000000 |

Table 1. One hot encoded data set.

Because the venue category information returned by Foursquare API are in string format (e.g. Café), the entire dataset had to be one hot encoded in order to make it compatible with the k means clustering algorithm.

# 3. Methodology

## 3.1 Data analysis

Upon closer inspection, I have noticed that some of the smaller cities such as Glenore Grove or Launceston made very little contribution to the dataset compared to bigger cities like Melbourne and at this stage I suspected that these would appear as outliers after clustering:

| | Neighborhood | 1st Most Common Venue | 2nd Most Common Venue | 3rd Most Common Venue | 4th Most Common Venue | 5th Most Common Venue | 6th Most Common Venue | 7th Most Common Venue | 8th Most Common Venue | 9th Most Common Venue | 10th Most Common Venue |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Adelaide | Hotel | Bar | Café | Coffee Shop | Italian Restaurant | Australian Restaurant | Wine Bar | Japanese Restaurant | Sculpture Garden | Pub |
| 1 | Albury | Café | Hotel | Bar | Motel | Thai Restaurant | Department Store | Shopping Mall | Italian Restaurant | Pharmacy | Garden |
| 2 | Bankstown NSW | Café | Vietnamese Restaurant | Fast Food Restaurant | Gym | Sports Bar | Supermarket | Buffet | Chinese Restaurant | Coffee Shop | Convenience Store |
| 3 | Bendigo | Café | Pizza Place | Restaurant | Pub | Fast Food Restaurant | Theater | Thai Restaurant | Liquor Store | Shopping Mall | Supermarket |
| 4 | Brisbane | Café | Coffee Shop | Hotel | Korean Restaurant | Burger Joint | Bookstore | Pub | Australian Restaurant | Art Gallery | Dumpling Restaurant |
| 5 | Cairns City | Café | Sporting Goods Shop | Toy / Game Store | Supermarket | Shopping Mall | Burger Joint | Sandwich Place | Basketball Court | Hotel | Home Service |
| 6 | Coffs Harbour NSW 2450 | Chinese Restaurant | Supermarket | Pub | Electronics Store | Shopping Mall | Garden | Fishing Store | Sandwich Place | Café | Pizza Place |
| 7 | Darwin | Hotel | Café | Australian Restaurant | Pub | Bar | Hostel | Vietnamese Restaurant | Asian Restaurant | Coffee Shop | Plaza |
| 8 | Gladstone QLD | Pub | Pier | Fast Food Restaurant | Beach | Park | Sports Bar | Department Store | Deli / Bodega | Flower Shop | Fishing Store |
| 9 | Glenore Grove | Convenience Store | Dumpling Restaurant | Food & Drink Shop | Food | Flower Shop | Fishing Store | Fish & Chips Shop | Fast Food Restaurant | Farmers Market | Event Space |
| 10 | Gold Coast | Italian Restaurant | Skating Rink | Thai Restaurant | Gas Station | Zoo | Dumpling Restaurant | Flower Shop | Fishing Store | Fish & Chips Shop | Fast Food Restaurant |
| 11 | Gosford | Pizza Place | Fast Food Restaurant | Food | Bistro | Bus Stop | Café | Sandwich Place | Football Stadium | Thrift / Vintage Store | Church |

Table 2. Unbalanced data contribution from each city.

For this particular dataset there were 184 unique categories which can be interpreted as the features of the dataset that will be used for clustering. Prior to transforming dataset suitable for clustering algorithm, a data frame containing information about top 10 most common venues for each city was created to check if there is any visible pattern:

| Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|
| **Adelaide** | 100 | 100 | 100 | 100 | 100 | 100 |
| **Albury** | 46 | 46 | 46 | 46 | 46 | 46 |
| **Bankstown NSW** | 55 | 55 | 55 | 55 | 55 | 55 |
| **Bendigo** | 54 | 54 | 54 | 54 | 54 | 54 |
| **Brisbane** | 100 | 100 | 100 | 100 | 100 | 100 |
| **Cairns City** | 14 | 14 | 14 | 14 | 14 | 14 |
| **Coffs Harbour NSW 2450** | 12 | 12 | 12 | 12 | 12 | 12 |
| **Darwin** | 70 | 70 | 70 | 70 | 70 | 70 |
| **Gladstone QLD** | 6 | 6 | 6 | 6 | 6 | 6 |
| **Glenore Grove** | 1 | 1 | 1 | 1 | 1 | 1 |
| **Gold Coast** | 4 | 4 | 4 | 4 | 4 | 4 |
| **Gosford** | 18 | 18 | 18 | 18 | 18 | 18 |
| **Henderson** | 23 | 23 | 23 | 23 | 23 | 23 |
| **Launceston** | 2 | 2 | 2 | 2 | 2 | 2 |
| **Melbourne** | 100 | 100 | 100 | 100 | 100 | 100 |
| **Mildura** | 20 | 20 | 20 | 20 | 20 | 20 |
| **Mount Gambier** | 17 | 17 | 17 | 17 | 17 | 17 |
| **North Adelaide** | 49 | 49 | 49 | 49 | 49 | 49 |

Table 3. 10 most common venues for each city.

Although there seemed to be a pattern among major cities whose contribution to the data set was high (high level of occurrences in café, coffee shop, and restaurants) clustering them into distinct groups was not an easy task.

## 3.2 k-means clustering

Although there are many different models for clustering, for this project I have decided to use k means clustering because it's fast, simple and often used as a base line method for data segmentation for its performance to computational resource cost. When using k-means clustering algorithm, determining the best number of clusters in a dataset is an important issue and must be chosen with care. However, there is no hard fast rule to this question and the selection of optimal number of clusters is somewhat subjective. Generally, as the number of k increase the loss will also decrease but this is not a linear relationship and there is a point where increasing k will start to affect the result in a negative way. Therefore, I have decided to employ a method called an elbow method where the sum of squared error (SSE) for each k is plotted on a graph and take the location of a bend in the plot as an indicator of the appropriate number of k:

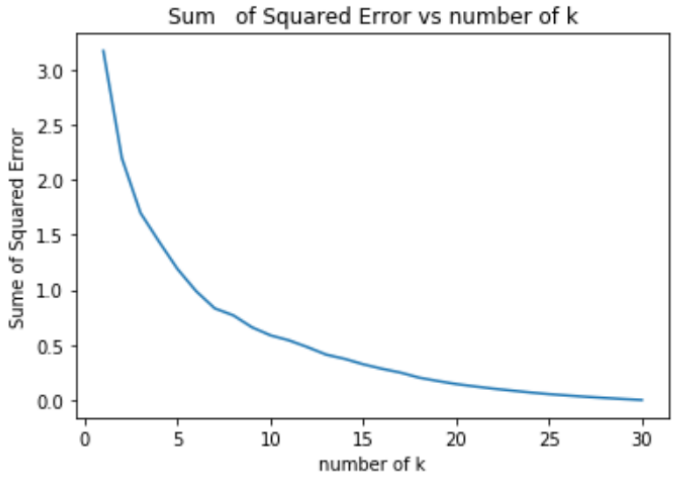


Fig 1. Sum of squared error vs number of k plot.

From the graph, I have determined that there is an elbow around k = 10.

## 4. Results

After clustering the data set, following result was obtained:

| Cluster Labels | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 15.000000 | 1.000000 | 1.000000 | 1.000000 | 7.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |

Table 3. Cluster labels frequency table.

From table 3 we can see that apart from 8 outliers, most cities can be grouped into 2 distinct clusters. In order to confirm my initial speculation on cities with the least amount of contribution to the data set becoming outliers, I have made a table that displays each element (city) belonging to each cluster as can be seen below:

| Cluster Labels | Neighborhood |
|---|---|
| 0 | Adelaide |
| | Albury |
| | Bendigo |
| | Brisbane |
| | Darwin |
| | Gosford |
| | Henderson |
| | Mildura |
| | Mount Gambier |
| | North Adelaide |
| | Orange |
| | Perth |
| | Phillip ACT |
| | Townsville City |
| | Wollongong |
| 1 | Launceston |
| 2 | Glenore Grove |
| 3 | Terrey Hills |
| 4 | Bankstown NSW |
| | Cairns City |
| | Melbourne |
| | Parnell |
| | Rosedale |
| | Sydney |
| | Westmead |
| 5 | Gladstone QLD |
| 6 | Sunshine Coast |
| 7 | Ziyou Today |
| 8 | Coffs Harbour NSW 2450 |
| 9 | Gold Coast |

Table 4. Elements belonging to each cluster.

As expected, cities which made little contribution to the data set were considered as outliers. In order to have a better understanding of the characteristics of 2 major clusters (0 and 4), a table listing top 10 most common venues with most frequent venue for each row was made:

| | value | | | |
|---|---|---|---|---|
| variable | count | unique | top | freq |
| 1st Most Common Venue | 7 | 1 | Café | 7 |
| 2nd Most Common Venue | 7 | 6 | Hotel | 2 |
| 3rd Most Common Venue | 7 | 6 | Coffee Shop | 2 |
| 4th Most Common Venue | 7 | 7 | Bus Station | 1 |
| 5th Most Common Venue | 7 | 7 | Sports Bar | 1 |
| 6th Most Common Venue | 7 | 6 | Park | 2 |
| 7th Most Common Venue | 7 | 6 | Sandwich Place | 2 |
| 8th Most Common Venue | 7 | 7 | Asian Restaurant | 1 |
| 9th Most Common Venue | 7 | 7 | Steakhouse | 1 |
| 10th Most Common Venue | 7 | 7 | Steakhouse | 1 |

| | value | | | |
|---|---|---|---|---|
| variable | count | unique | top | freq |
| 1st Most Common Venue | 15 | 6 | Café | 8 |
| 2nd Most Common Venue | 15 | 12 | Café | 2 |
| 3rd Most Common Venue | 15 | 11 | Café | 2 |
| 4th Most Common Venue | 15 | 9 | Coffee Shop | 3 |
| 5th Most Common Venue | 15 | 12 | Mexican Restaurant | 2 |
| 6th Most Common Venue | 15 | 14 | Indian Restaurant | 2 |
| 7th Most Common Venue | 15 | 14 | Thai Restaurant | 2 |
| 8th Most Common Venue | 15 | 15 | Japanese Restaurant | 1 |
| 9th Most Common Venue | 15 | 13 | Shopping Mall | 2 |
| 10th Most Common Venue | 15 | 14 | Supermarket | 2 |

Table 5. Characteristics table for Cluster 4 (left) and Cluster 0 (right).

From table 5, we can see that there are recognizable differences between two different clusters. Cities belonging to cluster 4 looks more like a busy urban area where there is a good balance in the ratio of entertainment facilities, transportation, and eateries present in the area whereas cities belonging to cluster 0 there seems to be an overwhelming number of cafés and eateries present in the area with occasional presence of facilities like shopping mall or supermarket.

This result seems to agree with reality to a certain degree given that Sydney, Melbourne and Auckland are the busiest and most urbanized cities in Australia and New Zealand and they all share similar characteristics. Another point to notice is that a lot of major cities in Australia belong to cluster 0 and this might explain why a lot of kiwis from Auckland migrating to Australia often find themselves having to adapt to the new environment.

Using Folium, we can represent our final result on a map:

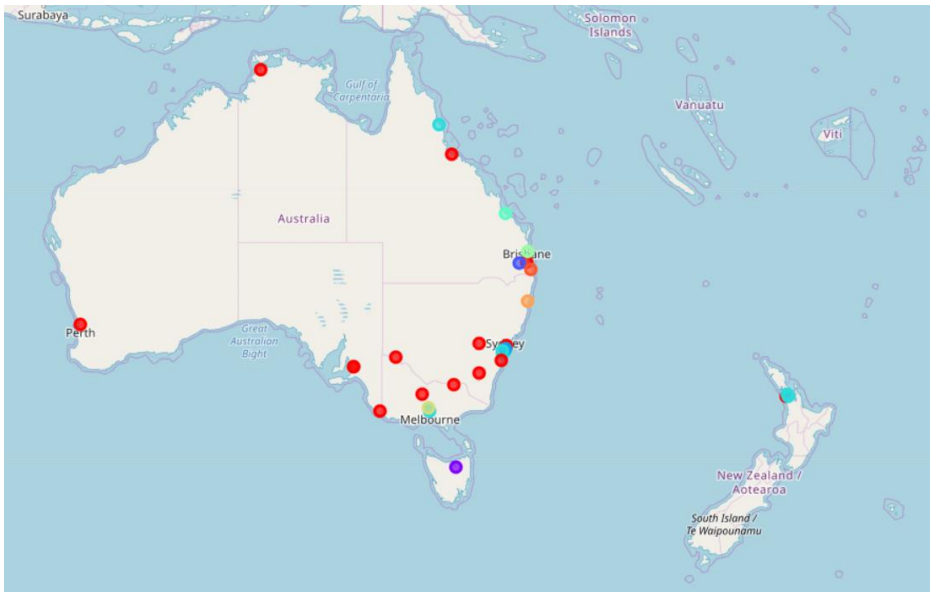


Fig 2. Cities of Australia and New Zealand clustered into 10 distinct groups.

## 5. Discussion

Based on the observations we have made in the results section, it seems that we will be able to give recommendation on list of cities that will be suitable for the people living in Auckland who are planning on migrating to Australia for financial related reasons (e.g. employment, housing issues) but wants to find a city that has similar living conditions as Auckland. Of course, the recommendation given based on this experiment should not be taken as a be all end all solution as the limitation of this project is pretty clear given that the data set I have used for my model is limited in terms of both quantity (limited to 100 venues per city) and quality (lack of variety as I have only used frequency of venues provided by a single provider) and there are many more factors to be considered in determining the similarities between cities. However, the results that I have obtained from this experiment may be used in conjunction with other studies to draw more complete conclusion in the future.

## 6. Conclusion

In this project, I have explored the possibility of grouping major cities of Australia and Auckland into number of clusters where each member of the same cluster share similar characteristics. For this experiment, I have used the Foursquare API to collect the information on frequency of trending venues for each city and used k-means clustering algorithm with optimum number of k to divide samples into 10 distinct clusters. The result I have obtained from this experiment can be useful for kiwis from Auckland who are thinking of migrating to Australia and wants to know which Australian city would have the living environment that is most similar to Auckland.