**Understanding the market in Canberra, Australia**

*Jacob Siddall*

April 2019

## **Introduction**

1. *Background*

Canberra is the capital city of Australia and has a population of approximately 400,000. By 2058, the city’s population is projected by local government officials to reach 700,000. As a result, business interest in the region is growing in the form of thriving apartment developments and a growing food and entertainment scene.

1. *Problem*

I have been approached by a group of businesses, from overseas, interested in understanding (1) where younger adult/medium-high income Canberrans cluster and (2) existing competition in the area. Analysing these two issues together will provide business intelligence to inform the selection of optimal locations for establishing new business enterprises in the Canberra area.

1. *Interest*

Interested parties in this project are the hypothetical business group, but also anyone interested in learning about how to use data to improve business outcomes.

## **Data collection and cleaning**

1. *Data sources*

Two data sources were used for this analysis:

1. [2011 Australian Bureau of Statistics (ABS) census data](http://www.abs.gov.au/)

The ABS conducts a national census of the Australian population approximately every five years. Unfortunately, the ABS does not yet have a functioning Automated Programming Interface (API) for the dataset of interest. Instead, data was outputted from the ABS website into two Excel CSV files using the ABS ‘TableBuilder’ product – namely ‘proportion (%) of population by Canberra postcode and income bracket’ and ‘proportion (%) of population by Canberra postcode and age bracket’.

1. [Foursquare location data](https://developer.foursquare.com/)

In order to identify already existing local business competition within the food and entertainment industry, the Foursquare API was used to retrieve relevant location data by Canberra postcode.

1. *Data cleaning and feature selection*

Throughout the notebook for this project, there is many variations of the same data described above. Data cleaning and manipulation was performed on both income and age ABS datasets to prepare them for various forms of analysis. This included:

* Transforming data types
* Pivoting/stacking data tables
* Renaming and reordering columns
* Applying basic calculations across cases (e.g. sum)
* Dropping redundant features
* Computing new variables required for the analysis

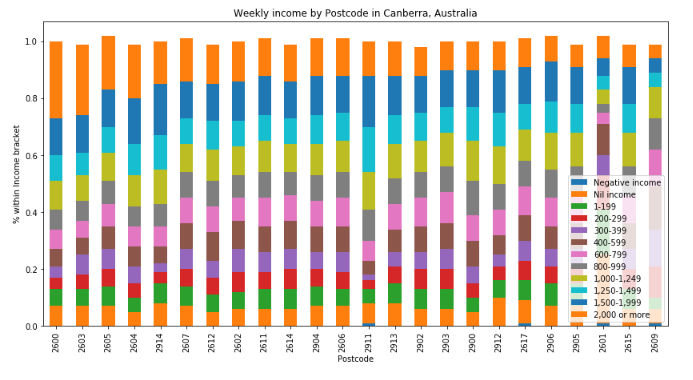
## **Methodology**

For this project, segmentation and clustering, using the k-means clustering algorithm, was selected as the primary method of analysis. Prior to using k-means clustering, the datasets were explored using graph representation, frequency distributions, and data descriptions.

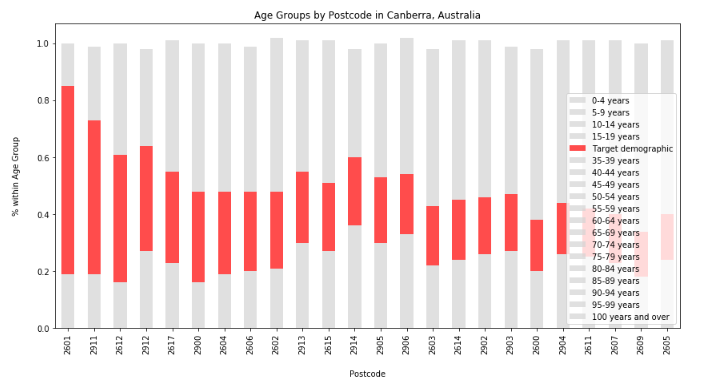
Segmentations and clustering was chosen for this project because of the nature of the business question and the availability of data. Ideally, other machine learning methods such as regression could be applied to predict the geographic areas with the most sales, however sales data this granular was not available publicly.

In the first graph displayed below, we can see all postcodes in the Canberra region (columns) broken down by weekly household income. This graph has then been sorted to show the postcodes with the most to least households earning ‘2,000 or more’ weekly. The postcode with the most of these earners is 2600, which is the Canberra city centre.

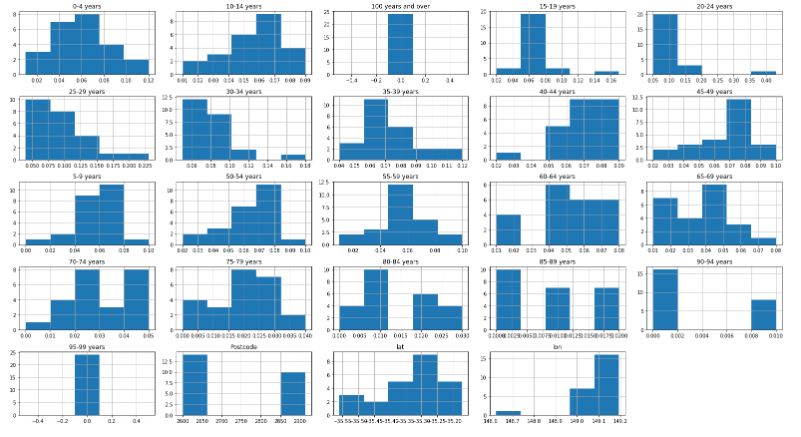
1. *Exploratory data analysis*



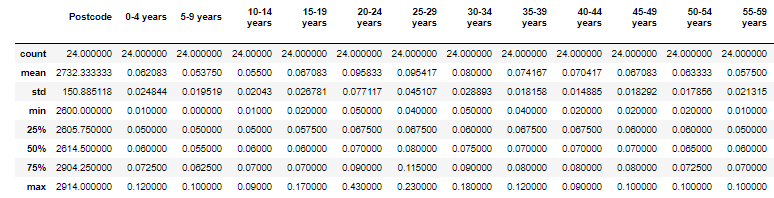
Following this is another stacked column graph, this time representing all of the postcodes of Canberra by age. Instead of breaking age groups down into their brackets, three age groups aged 20-34 have been folded into a single ‘Target demographic’ (identified by red). From this we can observe that the first three postcodes have an abundance of 20-34 year olds.

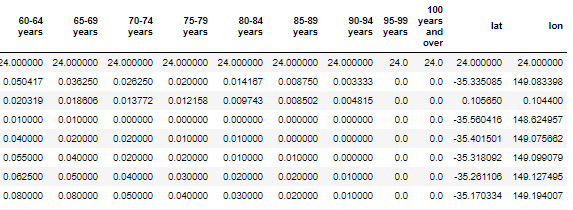


Next, we take an alternative look at our age groups, displayed as a set of frequency distributions. This gives us an idea of the shape of our data. For example, those aged ‘100 years and over’ are centred on 0, as there are no people of this age in our population. Further, in the first square we can see that there are eight postcodes where 0-4 year olds represent approximately 6% of the population.



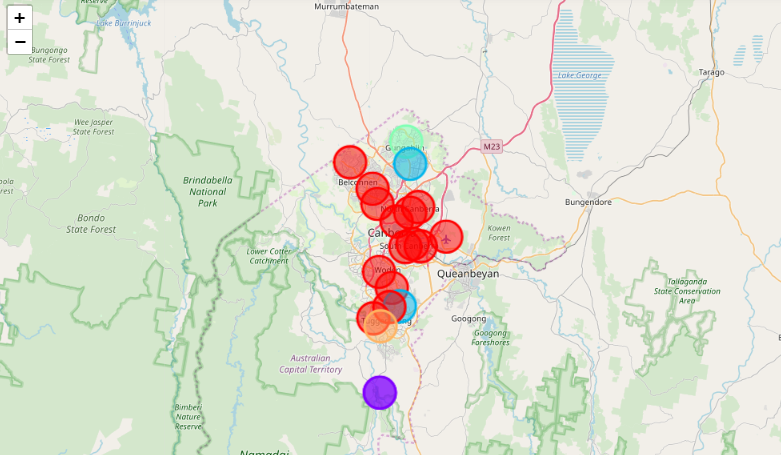
Further to this, we can learn about our data through the python describe() function. Below, we can see that we are dealing with 24 postcodes, in which the mean proportion of 20-24 year olds is 9.5 per cent and the postcode with the highest (max) proportion of 20-24 year olds is 43% (a university is located in this postcode).





1. *Segmentation and Clustering*

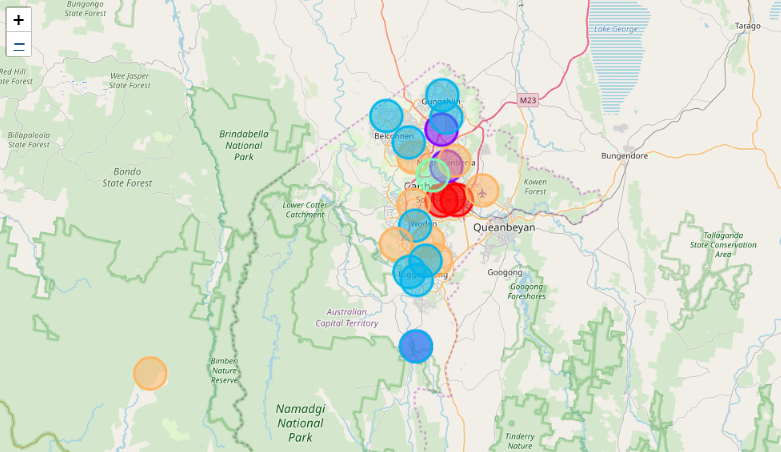
Following the exploration of the datasets, segmentation and k-means clustering was applied to cluster the postcodes of Canberra into alike groups based on the frequency of *venue types* obtained from the Foursquare API.



From this analysis, five clusters that emerged:

* + Red seemed to indicate a proximity (within 2,000 metres) to, and high frequency of, places like hotels, cafes and restaurants.
  + Light blue indicated proximity to lakes and fast food restaurants.
  + Green indicated a single unique postcode more proximate to places such as athletics and sports venues, rather than cafes/hotels.
  + Purple indicated an outlier in the facilities available, probably due to remoteness (being located far away from the city’s centre).
  + Orange was an oddity, with the most frequent venue being cupcake shops and wine bars. As a local who is familiar with the southern Tuggeranong area, this makes sense.

Then, the demographic data collected from the ABS website was run through the clustering algorithm and mapped separately. This time, my interest was to see which suburbs were high income with a high proportion of younger adults.



Of interest in this map is that:

* + Red indicated a high frequency of high incomes with an older population.
  + Light blue indicated middle income areas with mixed younger and older populations.
  + Green indicated a concentration of poor young students (the Australian National University).
  + Purple indicated a relatively mixed population.
  + Orange broadly indicated lower income areas.

## **Results**

Together, the results from my analysis area as follows:

|  |  |
| --- | --- |
| **Market Competition** | **Age and Income** |
| Red seemed to indicate a proximity (within 2,000 metres) to, and high frequency of, places like hotels, cafes and restaurants. | Red indicated a high frequency of high incomes with an older population. |
| Light blue indicated proximity to lakes and fast food restaurants. | Light blue indicated middle income areas with mixed younger and older populations. |
| Green indicated a single unique postcode more proximate to places such as athletics and sports venues, rather than cafes/hotels. | Green indicated a concentration of poor young students (the Australian National University). |
| Purple indicated an outlier in the facilities available, probably due to remoteness (being located far away from the city’s centre). | Purple indicated a relatively mixed population. |
| Orange was an oddity, with the most frequent venue being cupcake shops and wine bars. As a local who is familiar with the southern Tuggeranong area, this makes sense. | Orange broadly indicated lower income areas. |

## **Discussion**

The discussion I would like to have revolves around a set of recommendations that have been drawn from discovered insights. My recommendations to the interested business, based on the above analysis (more details about which can be found throughout the Jupyter Notebook) are to:

1. ***Explore opportunities in the Gungahlin region of Canberra.***

Gungahlin (northern suburbs) market competition in the food and entertainment industry if comparatively lower than other Canberra regions. At the same time, the demographic makeup of Gungahlin is like Belconnen, in that it is mostly middle income older and younger adults.

1. ***Avoid the market saturated areas of Civic and Kingston.***

Although these areas attract high-income cohorts there is already an established food and beverage industry here. People residing in these suburbs are also older, on average, than our target demographic.

1. ***Avoid areas marked in orange on the second map, if your target market is high-income earners.***

While business opportunities can be found across income brackets, if your target group is high-income earners the orange clusters have identified low‑middle income areas.

## **Conclusion**

This report has presented an analysis using ABS data on the age and income characteristics, and Foursquare data on the market competition characteristics, of the postcodes of Canberra, Australia.

Segmentation and clustering was applied to reveal those areas of Canberra that will likely provide optimal locations for establishing a new food and entertainment business. Namely, businesses were encouraged to look at the Gungahlin (northern) region of Canberra, as this contained little industry competition and a high frequency of middle-high income younger adults.