**A project to explore the major regions, localities (Neighbourhood) & cities within the state of New South Wales Australia.**

***Prepared by: Gbenga Afuape***

1. **Introduction of the Business Problem**

A migration agent targeting African Clients called Aus-Migration Pty has required my services in providing major regional distribution of people within the state of New South Wales in Australia. The client would be interested in knowing:

1. What the sizes of those regions are (km2),
2. In seeing the distribution of population for the years 2018 and 2019 respectively and the population density for the year 2019 specifically.
3. Component reasons for the distribution for the year 2019.

Aus-Migration pty requests that in addition to summary tables having some visual maps displaying some of the information would be beneficial for their potential clients.

As The 2019–20 Australian bushfire season, colloquially known as the Black Summer, began with several serious uncontrolled fires in June 2019. Throughout the summer, hundreds of fires burnt, mainly in the southeast of the country. From September 2019 to March 2020, fires heavily impacted various regions of the state of New South Wales. Aus-Migration Pty would like to visualize on a map the said regions and the neighbourhoods around it so as be able to educate their potential clients on some of the hazards and things to be mindful of if and when they eventually migrate to New South Wales.

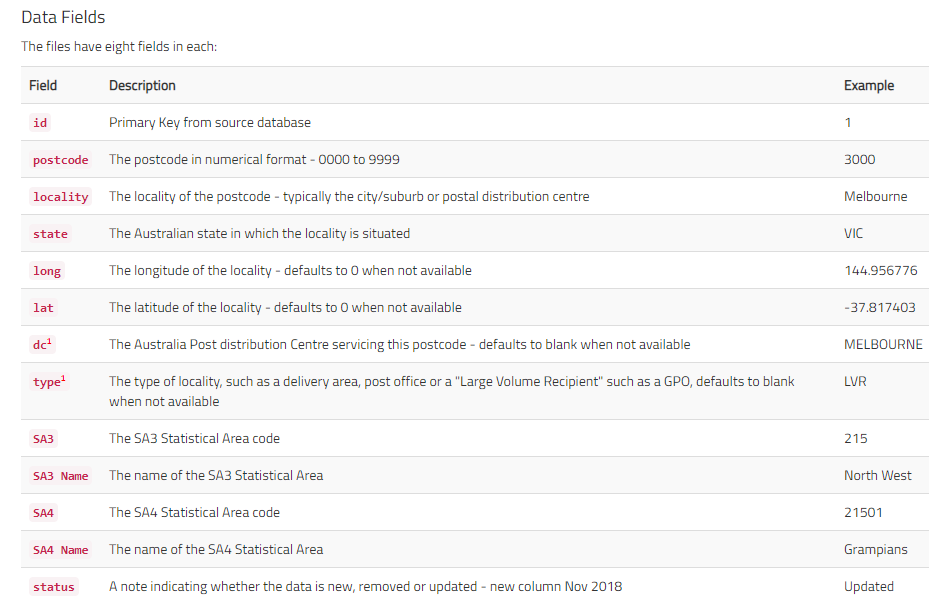
As Sydney is the most populous city in New South Wales and indeed Australia, Aus- Migration Pty would like a map to explore this city in addition to the regions (suburbs and localities) in our reduced data set on New South Wales. Aus- Migration Pty would want to see regions within 5km of these neighbourhoods.

Aus-Migration Pty, is targeting a wide range of customers from Specialist Doctors, Nurses, Researchers, Farmers, Entrepreneurs to students who would be assisted in procuring various types of visas that would enable them emigrate from Africa to New South Wales in Australia to either work (white collar or blue-collar), study or start a business. In support of this they would want a clustering of the neighbourhoods with the Top 10 venues.

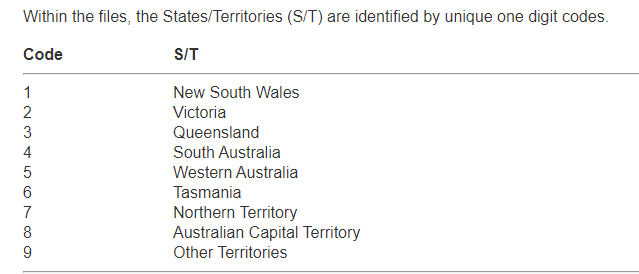
1. **Data Section**

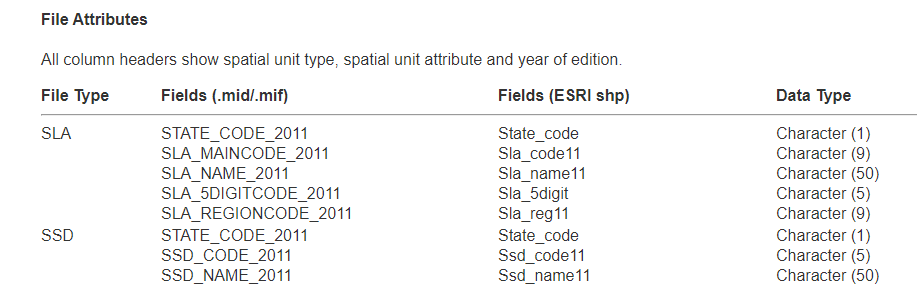
To assist Aus-Migration Pty I would be leveraging 3 Data Sources or types:

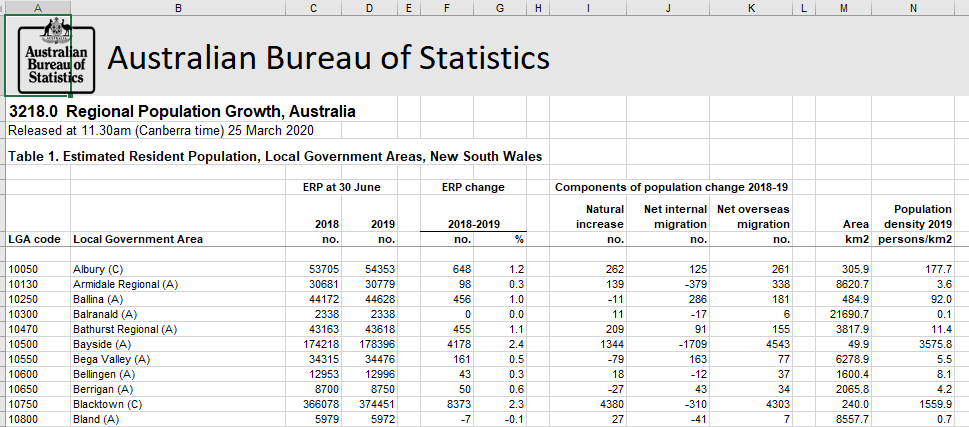
1. A raw data file from Australia Post with an up-to-date downloadable database of Australian Postcodes and Localities, including accurate longitude and latitude values. This data set is in a CSV format filtered to New South Wales. The description of each field is as below:



1. Australia Bureau of Statistics (ABS) provides the digital boundary file as shape file which would be used for making some of the maps in addition to that from Foursquare. The ABS also provided the population distribution data with description of the examples shown below.







1. Foursquare location data. I will be using the Foursquare API. With the Foursquare API. I would first construct a URL to send a request to the API to search for a specific type of venues. I then search for specific type of venues around a given location eg. Sydney or Illawara. We can also learn more about a specific venue or store or shop, like their full address, their working hours, and their menu if they have one, and so on. Furthermore, we can explore a given location by finding what popular spots exist in the vicinity of the location, and for this data a regular call to the API would be made. I will make the call to the database and get a JSON file of the trending venues that are nearby. In the JSON file, for each trending venue, we get mostly its name, unique ID, location, and category. From the JSON file I will then extract the categories for a Dataframe. See sample JSON output and dataframe below:

{'meta': {'code': 200, 'requestId': '5e91415f542890001b018226'},

'response': {'suggestedFilters': {'header': 'Tap to show:',

'filters': [{'name': 'Open now', 'key': 'openNow'}]},

'headerLocation': 'Sydney',

'headerFullLocation': 'Sydney',

'headerLocationGranularity': 'city',

'totalResults': 128,

'suggestedBounds': {'ne': {'lat': -33.97070495499995,

'lng': 151.11638853595687},

'sw': {'lat': -34.06070504500004, 'lng': 151.0080114640431}},

'groups': [{'type': 'Recommended Places',

'name': 'recommended',

'items': [{'reasons': {'count': 0,

'items': [{'summary': 'This spot is popular',

'type': 'general',

'reasonName': 'globalInteractionReason'}]},

'venue': {'id': '4dafa82e4b22d9b3bcc7031b',

'name': 'Smiley Thai',

'location': {'address': '560 Box Rd',

'lat': -34.016321853913176,

'lng': 151.06603632638232,

'labeledLatLngs': [{'label': 'display',

'lat': -34.016321853913176,

'lng': 151.06603632638232}],

'distance': 360,

'postalCode': '2226',

'cc': 'AU',

'city': 'Jannali',

'state': 'NSW',

'country': 'Australia',

'formattedAddress': ['560 Box Rd', 'Jannali NSW 2226', 'Australia']},

'categories': [{'id': '4bf58dd8d48988d149941735',

'name': 'Thai Restaurant',

'pluralName': 'Thai Restaurants',

'shortName': 'Thai',

'icon': {'prefix': 'https://ss3.4sqi.net/img/categories\_v2/food/thai\_',

'suffix': '.png'},

'primary': True}],

'photos': {'count': 0, 'groups': []}},

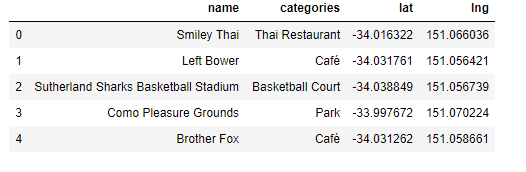
'referralId': 'e-0-4dafa82e4b22d9b3bcc7031b-0'},

{'reasons': {'count': 0,

'items': [{'summary': 'This spot is popular',

'type': 'general',

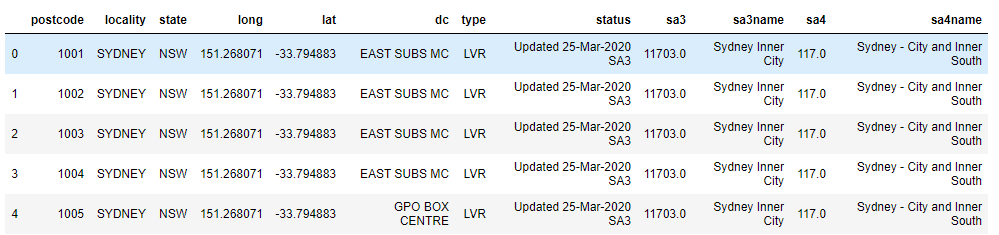
'reasonName': 'globalInteractionReason'}]},



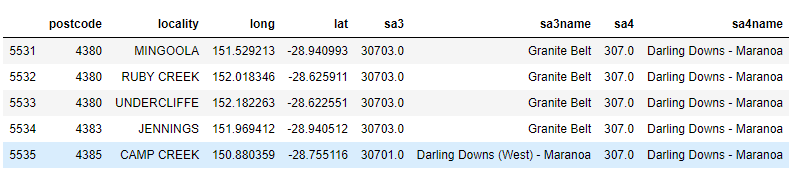
1. **Methodology**

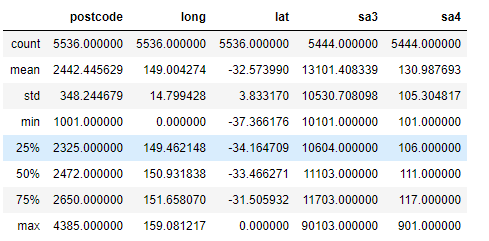
**3.1. Data Wrangling and Exploratory Data Analysis**

As we received a raw data set from the Australia Post office we will first have a peek at what this data contains. The summary table of the features and the top 5 rows is shown below:

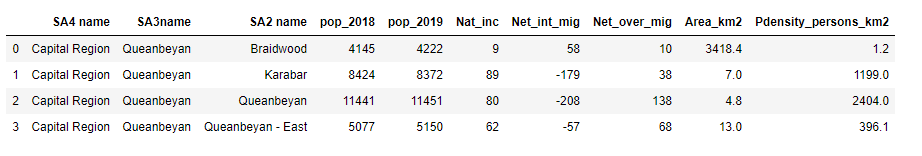


We immediately see a couple of columns which would not be meaningful to anyone but Australia post. We can proceed to drop those columns and this time display the tail of the data set and the 5 number summary statistics of this data set:

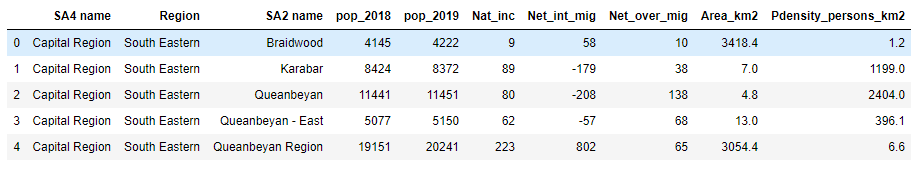


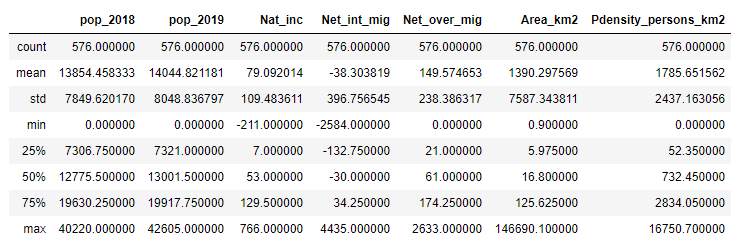


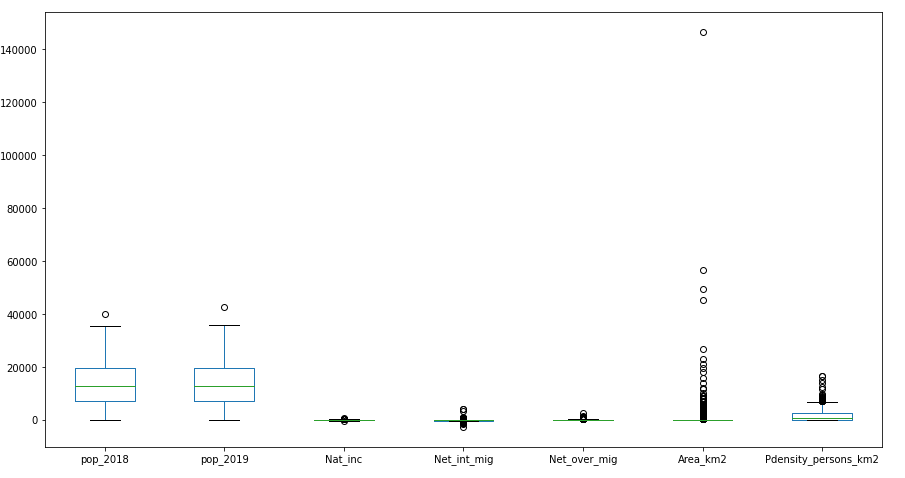
We see that there are over 5000 post codes in NSW, however, our client is not interested in this level of details. Furthermore, We do have population distribution data for New South Wales as at Mar 2020 from Australia Bureau of Statistics. Let us take a peek at this data set:

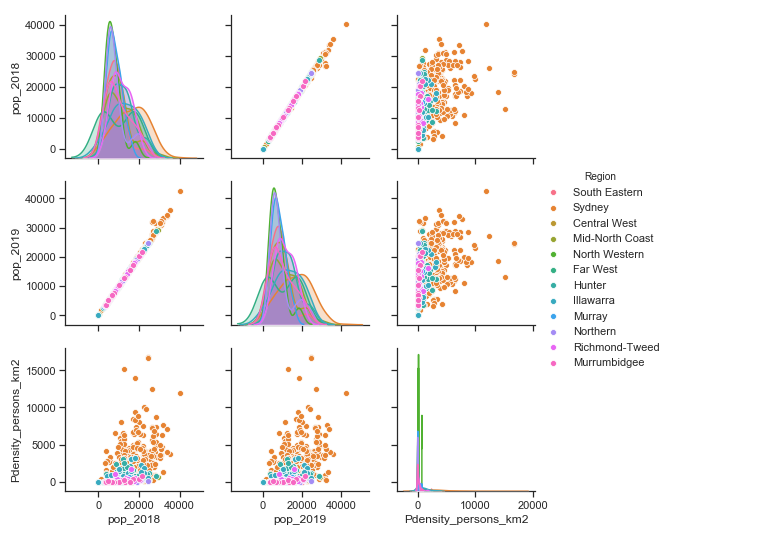


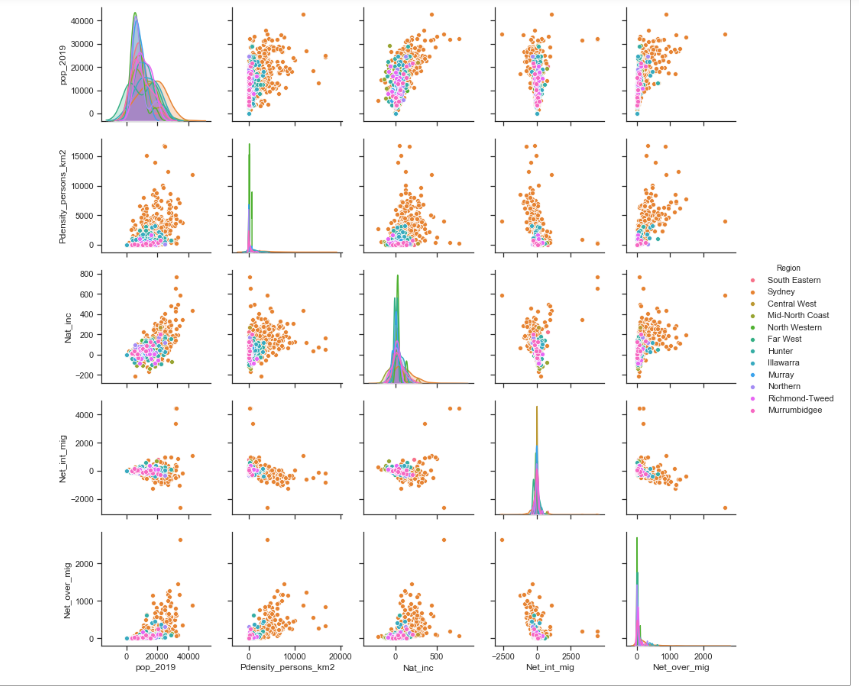
To be able to provide some of the population statistics and maps required by Aus-Migration pty, we need to create a Region column based on the Statistical Area level: we create that below and determine the 5 number summary statistics, the box plot and scatter matrix of a much high level grouping below.:











By grouping into 12 major regions we can satisfy some of the requests of Aus-Migration pty, We will discuss some of the insights (identify potential outliers in the data set and also if there is a noticeable trend) of the EDA (Boxplot and scatter plots) in the next section.

* 1. **Exploring Neighbourhood, Segmentation and Cluster Analysis**

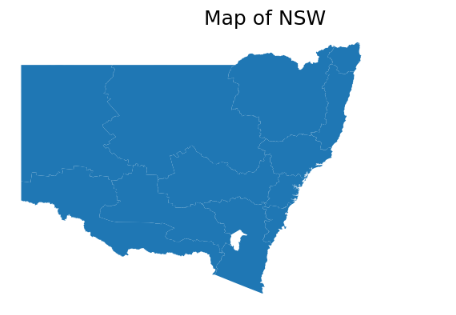
As we have our Fours API developer account set-up, we can begin accessing the location data. Here we would do some exploration of Sydney using 105 neighbourhoods, before reducing this to 45 ensuring most of the neighbourhoods affected by the 2019-2020 fire are captured.

Since our client has diverse group of potential clients, we want to be able to see what neighbourhoods are similar and what makes them so.

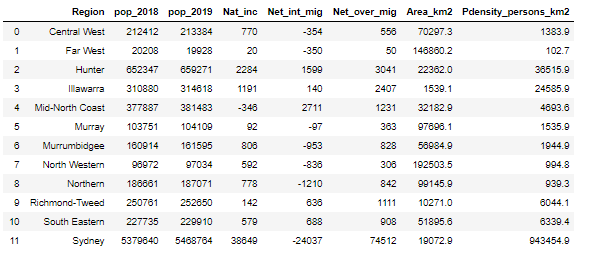
We will be using the Unsupervised Clustering algorithm. K-mean Clustering as it is one of the most common. We will ultimately create a data frame from these location data. We will see if it is sensitive to 5 or 6 clusters. We will also look for sensitivities to the radius of 5km and 2.5km. We LIMIT searches to 100, as we observe some cities like Sydney do have a high population density.

1. **Results**

Using Geopandas and the shapefile from the Australian Bureau of Statistics we obtain the majore 12 statistical divisions for the State of New South Wales.

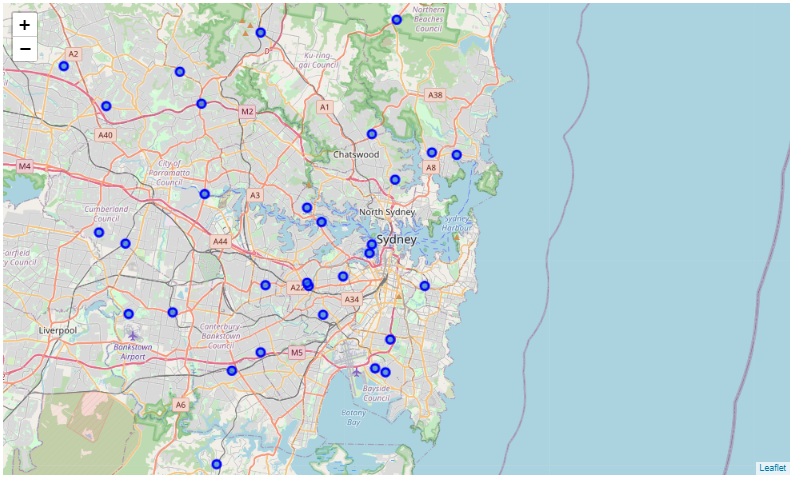


We obtain the corresponding population at the group as shown below:

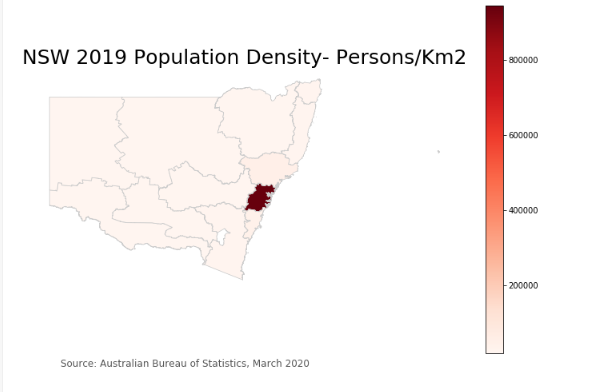


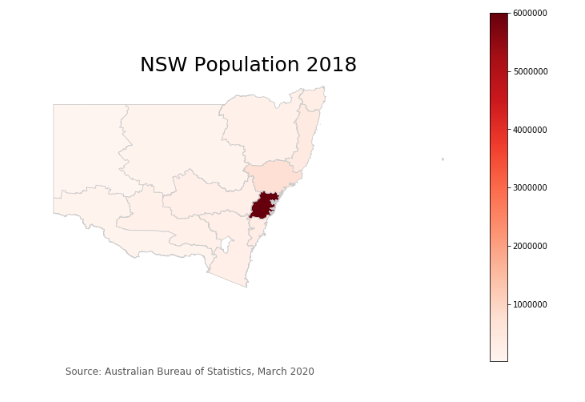
This table summarizes the total population at the major regions for 2018 and 2019. It also describes the population density for 2019 based on the regional area. It further breaks down the basis for the 2019 population. We immediately see that Sydney is densely populated coming with 943,454 persons/km^2. Hunter and Illawarra coming a distant 2nd and 3rd place respectively. The far west just does not seem to be inhabited. Lastly, we can see that most of the increase is from overseas migration to Sydney. We also note that people did move to other cities from Sydney.

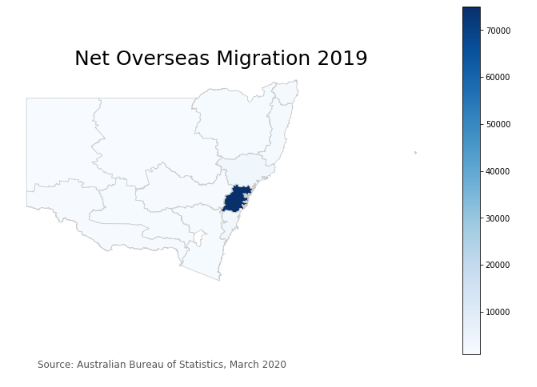
We obtain the Folium Map with latitude and longitude centred on Sydney to see the surrounding neighbourhoods. The next set of maps were requested by Aus Migration Pty. Essentially, we hung the population distribution on the NSW shape map.



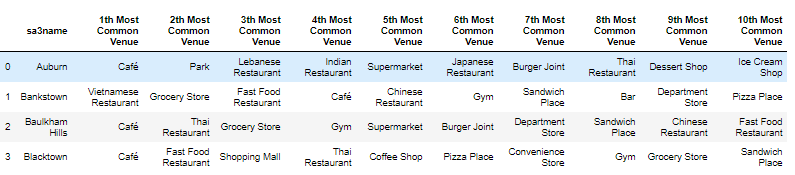
Folium Map showing Sydney and the neighbourhoods

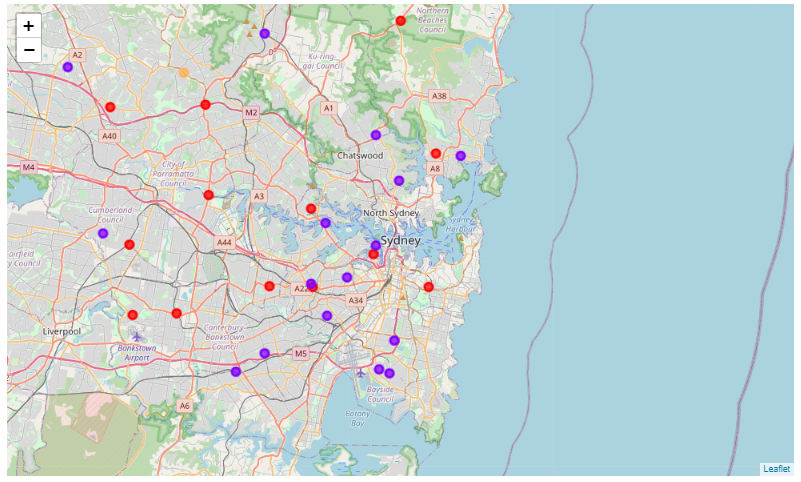




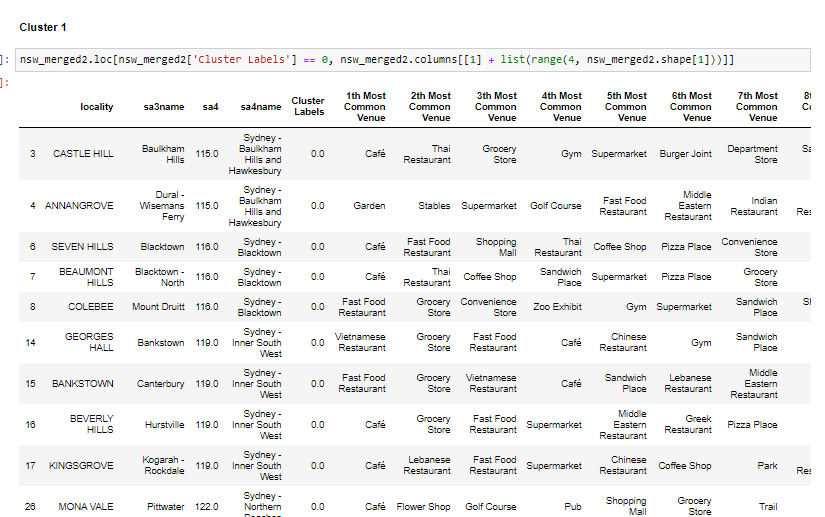


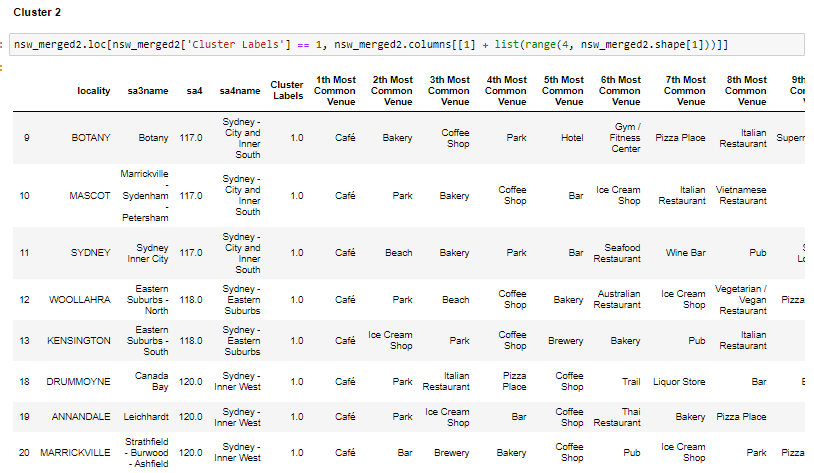
We then proceed to the segmentation and clustering stage. We use the Four Square API to make calls to the location database, obtain and group the various venues categories, carry out one-hot encoding and then generate a new dataframe of the 10 most common venue for the K means clustering algorithm. We show the top 4 rows below and the map of the resulting clusters:

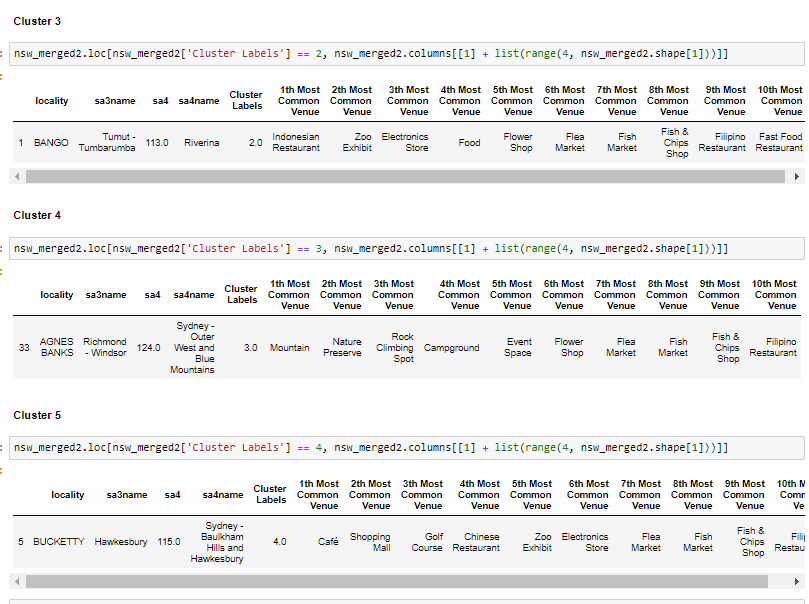




**Folium Map showing Clusters of the neighbourhood around Sydney and the 11 major regions.**







1. **Discussion**

It is noteworthy that most migrants wanting to come to Australia and indeed the state of New South Wales always head to Sydney even though Sydney is densely populated and whilst having very developed and thriving Neighbourhood does share similar characteristics as other neighbourhoods.

To the entrepreneur clients of Aus Migration Pty, Clusters 1 and 2 provide ample neighbourhoods to explore Café or catering business. The clients should not limit themselves to getting one from the ground running but might identify chain or franchise business in those industries. As not one African restaurant stood out exploring the possibility of opening one is advised.

To the students, there is ample opportunity to get part-time work to supplement their finances within the proviso of their student visas. Exploring the neighbourhoods particularly Cluster 2 where the 2 prestigious universities are located.

Cluster 4 and 5 are particularly interesting in that 2 specific neighbourhoods are flagged here in Sydney outer west. This neighbourhood would support outdoor related businesses and farming. This neighbourhood should be explored by specialist doctors who may be state sponsored and required to locate their practice in regional areas.

1. **Conclusion**

The New South Wales population data for 12 major regions have been put through an unsupervised clustering. We have been able to successfully segment the NSW into 5 clusters. Clusters 1 has about 20 localities (neighbourhoods) which are mainly in the Sydney City, inner west, inner south and the northern beaches.

Cluster 2 bounds another 21 localities (or neighbourhoods) and interestingly some of these are in the inner west of Sydney, north Sydney and South west.

Clusters 3,4 and 5 have 1 neighbourhood each.

The unsupervised Clustering algorithm (k-means) for this work was not sensitive to the radius (5km vs 2km) used in making calls to the Four Square Database nor increasing the clusters to 6.

We have successfully made use of the population data from the Australian Bureau of Statistics and the Four Square API to determine the 5 clusters of neighbourhoods in New South Wales or within the 12 major statistical divisions.