# Applied Data Science Capstone Project

# Exploring areas and venues in Sydney NSW, Australia

Harry Ngo

## 1. Introduction

New South Wales (NSW) is one of the major states in Australia, being most populous with over 8.1 million residents. There are many cities and towns within NSW, where its capital city Sydney is home to almost two-thirds of the NSW population.

The aim of the project is to explore different areas of Sydney and find optimal locations for a new resident based on the top common venues in each borough. This report will be targeted to people interested in becoming a resident in Sydney NSW, Australia, as well as homebuyers, people staying over a long period and other similar conditions. Others who are interested in learning about different boroughs of Sydney may also benefit from this.

With Australia being a multicultural country, its landscape is constantly evolving, seeing increases in different ethnic communities and venues. Sydney boroughs will be discussed such that stakeholders can choose an optimal living location based on their interests.

## 2. Data

The following data will be used:

1. *Second-level Administrative Divisions, Australia, 2015*[1]
   - Contains second-level administrative and political divisions of Australia. This includes aboriginal councils, boroughs, cities, government and district councils, municipalities, rural cities, shires and territories. The relevant information from the exported JSON file includes the state/territory, borough/council and the multiple coordinates of each division.
2. *Foursquare API*
   - Used to obtain the common venues of each borough within Sydney such as restaurants, stores and parks.

Exploring the JSON file of Australian administrative divisions, the geometry of each location was of type *MultiPoint*. To maintain consistency and handle the data simpler, the first coordinate of each location was taken. This dataset consisted of 11 states or territories and 1395 boroughs, which was subsequently extracted to obtain only data pertaining to NSW; this contained 199 boroughs.

Exploring the data for NSW, there was no quick method to find boroughs in NSW in the Sydney region. Instead, rows were manually removed which were not in the vicinity of Sydney, which was specified to be within an 80 km radius of Sydney.

[1] *Second-level Administrative Divisions, Australia, 2015*, NYU Spatial Data Repository <https://geo.nyu.edu/catalog/stanford-fc944xn1421>

For example, the first five observations of the NSW data frame look like the following:

| | State/Territory | Borough | Latitude | Longitude |
|---|---|---|---|---|
| **0** | New South Wales | Albury | -35.914963 | 146.971710 |
| **1** | New South Wales | Armidale Dumaresq Bal | -30.589890 | 151.632751 |
| **2** | New South Wales | Armidale Dumaresq | -30.483726 | 151.661728 |
| **3** | New South Wales | Ashfield | -33.870983 | 151.119431 |
| **4** | New South Wales | Auburn | -33.850048 | 151.081467 |

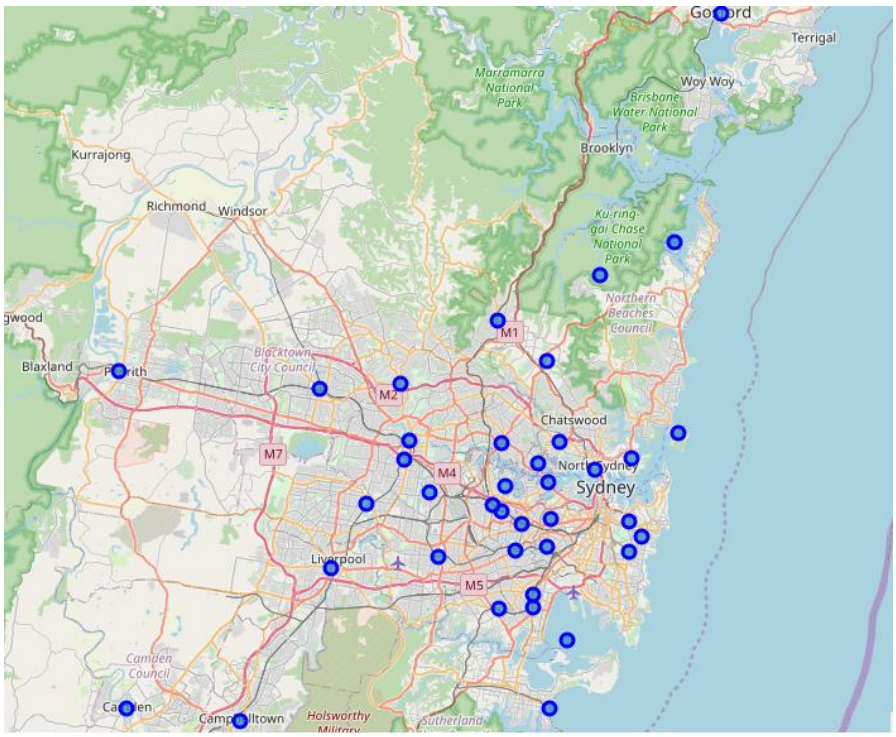
The boroughs of *Albury*, *Armidale Dumaresq Bal* and *Armidale Dumaresq* do not belong within an 80 km radius of Sydney (where the centre is in the city of Sydney), so they were removed from the data frame.

The data frame consisted of 58 boroughs, with some similar boroughs such as Liverpool – East and Liverpool – West. A further investigation when plotting the borough coordinates on a map of Sydney shows that picking out the first *MultiPoint* was not accurate. As I cannot find any other data sets online which have correct single coordinates, I cleaned up the data frame manually. I removed all other duplicates of the same borough (e.g. east, west) as they overlapped other boroughs. The final data frame now consisted of 38 boroughs.

## 3. Methodology

After retrieving the boroughs in our specified Sydney region as well as the longitude and latitude for each borough, *Foursquare API* will be used for exploratory analysis to find venues near our neighbourhoods. Next, K-means clustering algorithm will be used to create clusters of neighbourhoods which have similar characteristics.

First, using the *geopy* and *folium* library we can create a map of Sydney with the 38 boroughs superimposed on top to check if their locations are correct.




Next, using *Foursquare API*, nearby venues are discovered from the given coordinates. Information of the venues such as its name, category and coordinates are extracted from a defined radius of 1 kilometre. This data is then merged into a new data frame containing neighbourhood and venues information.

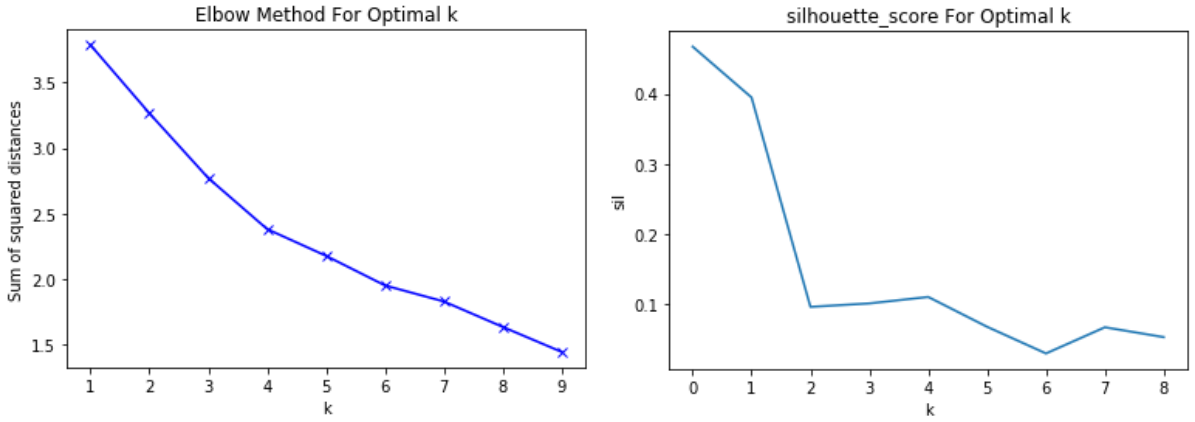
For example, the first five venues of Ashfield found were:

| | Neighbourhood | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Ashfield | -33.8889 | 151.1249 | Manmaruya | -33.889603 | 151.126183 | Japanese Restaurant |
| 1 | Ashfield | -33.8889 | 151.1249 | Choo Choo Train | -33.888275 | 151.124196 | Chinese Restaurant |
| 2 | Ashfield | -33.8889 | 151.1249 | Thai Number 1 | -33.887142 | 151.125685 | Thai Restaurant |
| 3 | Ashfield | -33.8889 | 151.1249 | New Shanghai Night 新夜上海 | -33.888461 | 151.124393 | Shanghai Restaurant |
| 4 | Ashfield | -33.8889 | 151.1249 | Taste of Shanghai | -33.888279 | 151.123992 | Dumpling Restaurant |

Unfortunately, some locations did not return any venues (Botany Bay, Warringah) which may have been due to the placement of coordinates, so they were removed from the analysis. Overall, there were 143 unique categories and 890 venues in total for our boroughs. The next image displays the top 10 venue categories that were returned; however, some were inaccurate as they did not capture at least 10 different venues.

| | Neighbourhood | 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 | Ashfield | Dumpling Restaurant | Chinese Restaurant | Platform | Shanghai Restaurant | Café | Supermarket | Electronics Store | Japanese Restaurant | Asian Restaurant | Coffee Shop |
| 1 | Auburn | Playground | Garden | Tennis Court | Café | Wine Bar | Discount Store | Electronics Store | Dumpling Restaurant | Dry Cleaner | Donut Shop |
| 2 | Bankstown | Vietnamese Restaurant | Café | Sports Bar | Department Store | Fast Food Restaurant | Convenience Store | Coffee Shop | Chinese Restaurant | Steakhouse | Grocery Store |
| 3 | Baulkham Hills | Café | Pizza Place | Supermarket | Pharmacy | Dessert Shop | Dumpling Restaurant | Shopping Mall | Burger Joint | Food Court | Fast Food Restaurant |
| 4 | Blacktown | Fast Food Restaurant | Juice Bar | Sandwich Place | Coffee Shop | Supermarket | Convenience Store | Park | Café | Portuguese Restaurant | Chinese Restaurant |

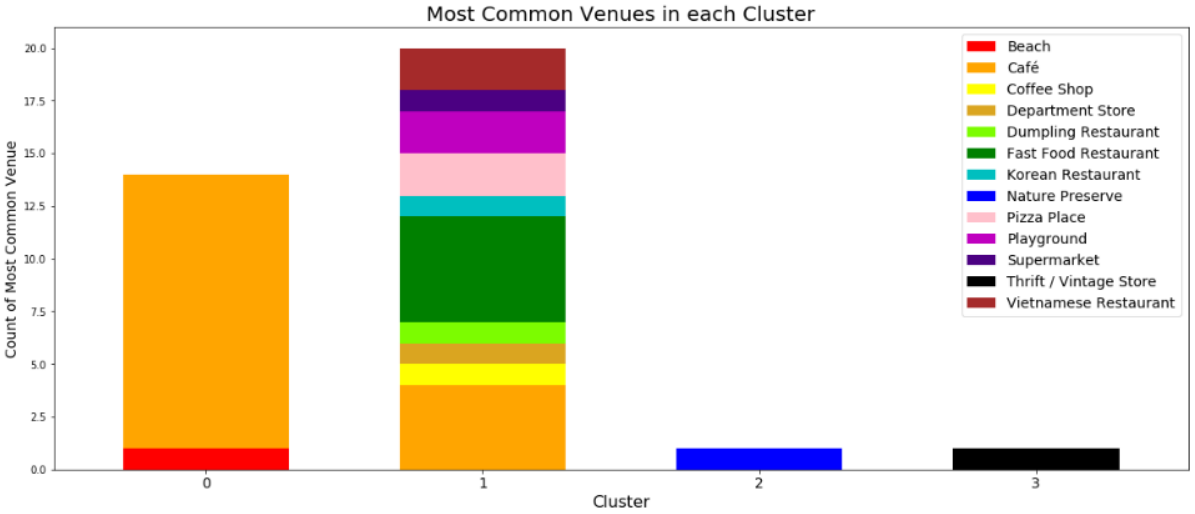
After using one hot encoding and taking the mean of the frequency for each venue category, *K*-means clustering is used, an unsupervised learning algorithm used to create *K* clusters of data points based on feature similarity. Observing the elbow method and silhouette score, we find that the optimal number of clusters to use is $K = 4$. From the silhouette score, the highest score after $K = 2$ attained is at $K = 4$.




After running *K*-means clustering, the cluster labels generated were added to a new data frame, merged from our original data frame and their corresponding top 10 venues.
Note that some venues are inaccurate as neighbourhoods may have captured less than 10 venues. This could have been improved by having a wider radius to capture more information, or a better coordinate for the location.

| | State/Territory | Neighbourhood | Latitude | Longitude | Cluster Labels | 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 | New South Wales | Ashfield | -33.8889 | 151.1249 | 1.0 | Dumpling Restaurant | Chinese Restaurant | Platform | Shanghai Restaurant | Café | Supermarket | Electronics Store | Japanese Restaurant | Asian Restaurant | Coffee Shop |
| 1 | New South Wales | Auburn | -33.8606 | 151.0247 | 1.0 | Playground | Garden | Tennis Court | Café | Wine Bar | Discount Store | Electronics Store | Dumpling Restaurant | Dry Cleaner | Donut Shop |
| 2 | New South Wales | Bankstown | -33.9182 | 151.0350 | 1.0 | Vietnamese Restaurant | Café | Sports Bar | Department Store | Fast Food Restaurant | Convenience Store | Coffee Shop | Chinese Restaurant | Steakhouse | Grocery Store |
| 3 | New South Wales | Baulkham Hills | -33.7619 | 150.9929 | 0.0 | Café | Pizza Place | Supermarket | Pharmacy | Dessert Shop | Dumpling Restaurant | Shopping Mall | Burger Joint | Food Court | Fast Food Restaurant |
| 4 | New South Wales | Blacktown | -33.7668 | 150.9053 | 1.0 | Fast Food Restaurant | Juice Bar | Sandwich Place | Coffee Shop | Supermarket | Convenience Store | Park | Café | Portuguese Restaurant | Chinese Restaurant |

As having cluster labels as 0, 1, 2 or 3 are unintuitive to interpret, a data visualisation of the 1st most common venue for each neighbourhood in each cluster can be formed. This would help in creating labels for each cluster. The following stacked bar chart for each cluster is shown below.




From analysing the bar chart and the top ten venues data frame, the clusters can be generalised and labelled as the following:

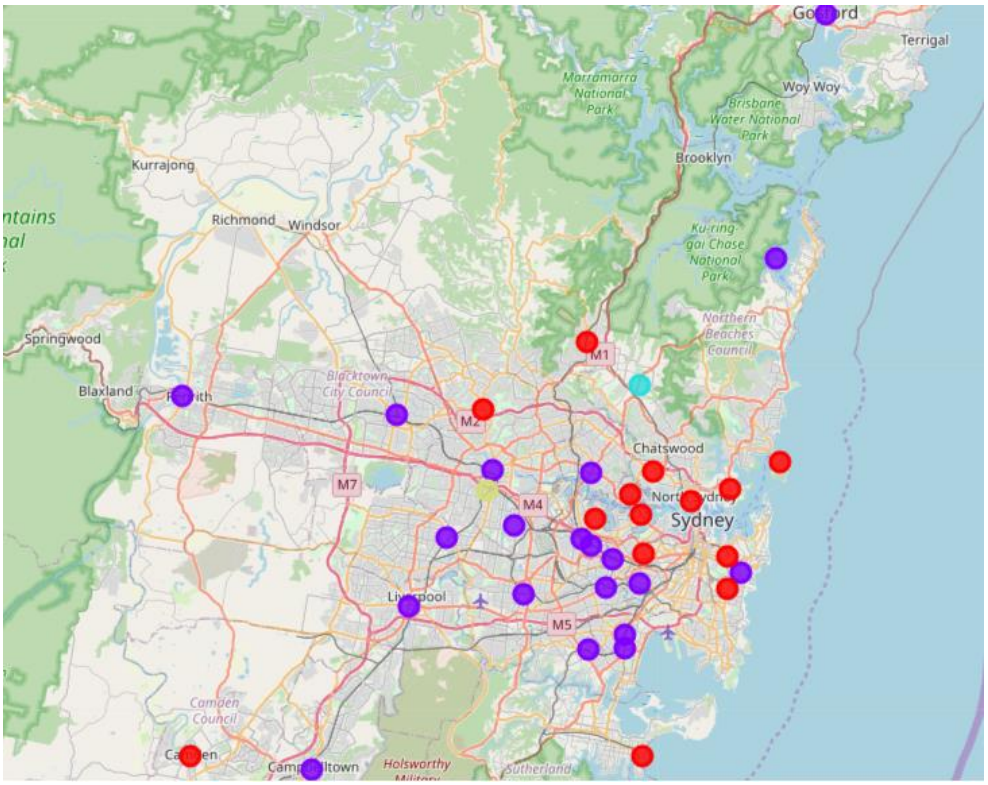
- *Cluster 0*: Numerous Café Venues
- *Cluster 1*: Multiple Restaurant/Café Venues and Shops
- *Cluster 2*: Nature Reserve Venues
- *Cluster 3*: Thrift Stores and Park Venues

Other information was added were the top two venues for each borough. This aids in finding quick features of each location after creating the new map containing the clusters. The new columns are then merged to create the final data frame containing all information.

| | State/Territory | Neighbourhood | Latitude | Longitude | Cluster Labels | 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 | Top Venues | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | New South Wales | Ashfield | -33.8889 | 151.1249 | 1 | Dumpling Restaurant | Chinese Restaurant | Platform | Shanghai Restaurant | Café | Supermarket | Electronics Store | Japanese Restaurant | Asian Restaurant | Coffee Shop | 3 Chinese Restaurant, 3 Dumpling Restaurant | Multiple Restaurant/Cafe Venues and Shops |
| 1 | New South Wales | Auburn | -33.8606 | 151.0247 | 1 | Playground | Garden | Tennis Court | Café | Wine Bar | Discount Store | Electronics Store | Dumpling Restaurant | Dry Cleaner | Donut Shop | 2 Playground, 1 Café | Multiple Restaurant/Cafe Venues and Shops |
| 2 | New South Wales | Bankstown | -33.9182 | 151.0350 | 1 | Vietnamese Restaurant | Café | Sports Bar | Department Store | Fast Food Restaurant | Convenience Store | Coffee Shop | Chinese Restaurant | Steakhouse | Grocery Store | 9 Vietnamese Restaurant, 5 Café | Multiple Restaurant/Cafe Venues and Shops |
| 3 | New South Wales | Blacktown | -33.7668 | 150.9053 | 1 | Fast Food Restaurant | Juice Bar | Sandwich Place | Coffee Shop | Supermarket | Convenience Store | Park | Café | Portuguese Restaurant | Chinese Restaurant | 6 Fast Food Restaurant, 2 Coffee Shop | Multiple Restaurant/Cafe Venues and Shops |
| 4 | New South Wales | Burwood | -33.8774 | 151.1037 | 1 | Café | Chinese Restaurant | Coffee Shop | Noodle House | Supermarket | Juice Bar | Department Store | Pharmacy | Fast Food Restaurant | Sandwich Place | 5 Café, 4 Chinese Restaurant | Multiple Restaurant/Cafe Venues and Shops |

## 4. Results

The final map displaying the clusters created by K-means clustering is shown below.




The results of K-means clustering of a total of 38 neighbourhoods depict 2 major clusters and 2 single neighbourhood clusters. From examining the clusters, we found that:

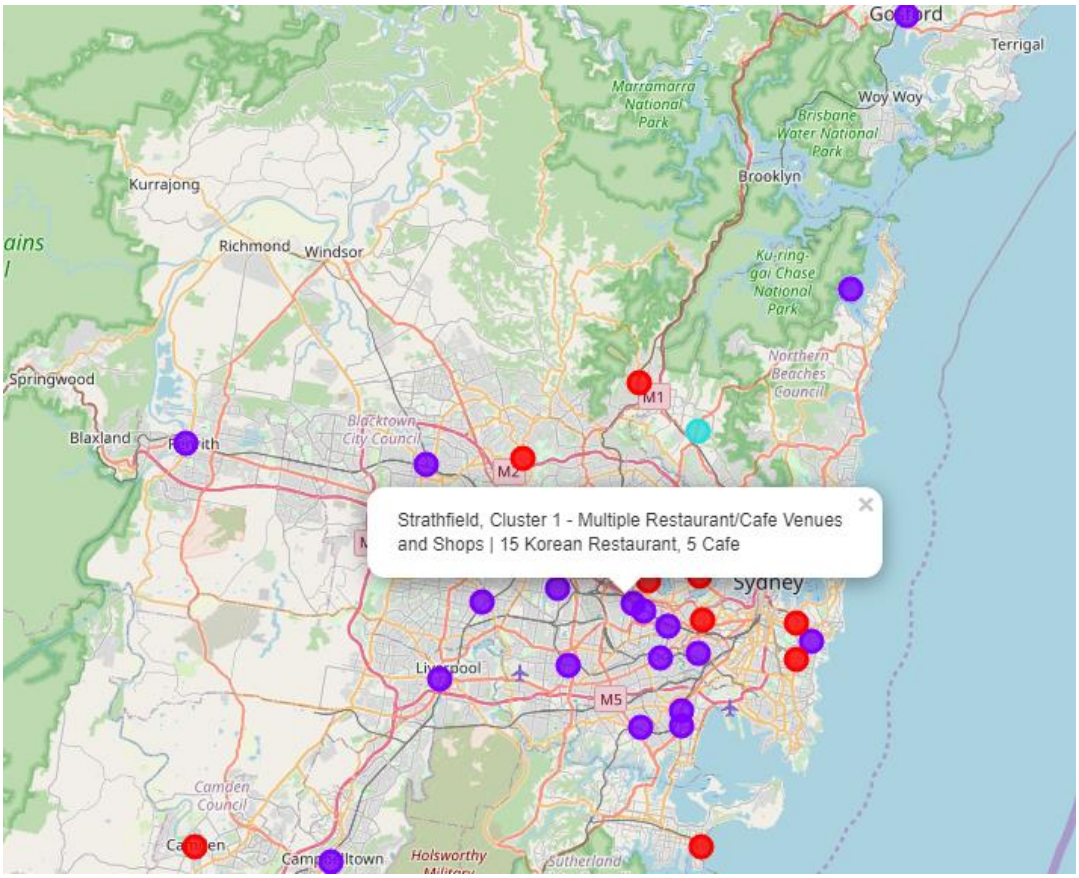
- **Cluster 0** (red) - abundance of cafés can be found at these locations, with 2 unique locations (Manly and Sutherland Shire, in this case, Cronulla) having beaches
- **Cluster 1** (purple) - wide variety of social areas, shopping stores and restaurants/cafés
- **Cluster 2** (cyan) - surrounded by nature reserves (Ku-ring-gai)
- **Cluster 3** (pale yellow) - consists of a park and thrift/vintage store (Holroyd)

## 5. Discussion

These clusters could have been more unique with more characteristics by including more neighbourhoods/council areas. A one km radius around the coordinates for each neighbourhood may not have been large enough to capture information about the area. For example, only 2 venues were found around the locations for clusters 2 and 3. This could be improved by choosing better coordinates such that more venues would be enclosed. Increasing the radius of the circles would also capture more information about other venues within the areas.

It is recommended for people looking to visit or live in Sydney NSW to research more about the neighbourhoods to investigate whether the areas suit their lifestyle or culture. The red dots (Cluster 0) are mostly closer to the CBD area of Sydney, with the areas being characterised by café venues. The purple dots (Cluster 1) are further out from Sydney CBD,

with the areas being characterised by their variety of food venues. By closer inspection, some areas are uniquely identified by the number of restaurants from a specific cuisine. For example, our analysis captured Italian restaurants (7) at Leichhardt, Vietnamese restaurants (8) at Marrickville and Korean restaurants (15) at Strathfield.




From analysing the top venues at different neighbourhoods, this resulted in exploring many areas characterised by cuisines and cafés. Choosing an area to reside in based on this factor alone is not optimal but could help specific groups of people looking for different areas for social gatherings and food.

## 6. Conclusion

People looking to become residents and other stakeholders could choose an optimal location in Sydney NSW based on their interests such as the cuisine of food at restaurants, the number of café venues or being near beaches and parks. However, this is not recommended as there are a wide range of factors that come into play when looking for an area to visit or live in. Clustering by top venues could be useful for people looking to open a restaurant in areas where the cuisine is distinctive from other restaurants in the area. The analysis performed in the report has showcased some unique venues for each cluster.

Overall, the final decision to find an optimal location for a new resident will be made by the individuals themselves. As well as finding about how the neighbourhoods are characterised by the number of venues, they should also consider other factors such as transport, housing prices and access to different necessities.