# Sydney - Food and Travel

Surya Soujanya Kodavalla

March 2020

## 1 Introduction

The purpose of this project is to identify the trending food ventures in each suburb of Sydney, Australia with accessible public transportation. Sydney has around 600 plus suburbs in the city with an excellent metro system and an amazing variety of cuisines available to gorge on. By calculating the distance from the nearest train stations and clustering restaurants based on this distance, anyone new to the beautiful city of Sydney can enjoy the food there without much of a hassle about transport.

The above problem has been addressed by finding the trending food places in each suburb of Sydney which are closest to train stations. Specifically, this report will be targeted to travelers interested in visiting the top food places which are are convenient to travel to and from in Sydney. We are particularly focusing on the train stations in Sydney as their metro network is very reliable with services running from 4am to around midnight on most train lines.

To do the analysis as mentioned above, initially a list of all the suburbs, all the train stations and the trending restaurants in Sydney are obtained. From this data, the nearest train station to each restaurant is found based on their locations and the distances calculated. After clustering the restaurants based on the distances previously calculated, we can come to a conclusion about the most accessible restaurants, etc.

Data science methodologies and the K-means clustering algorithm have been used to get useful data and generate different clusters of food places all over Sydney based on their distances from the nearest train stations.

## 2 Data

Based on definition of the problem, data needed for analysis is:

- list of suburbs in Sydney and their locations
- list of train stations in Sydney and their locations
- top food venues in different suburbs
- distance between food venues and the nearest train station

## 2.1 List of Suburbs in Sydney and their locations

- Data for list of suburbs in Sydney was scraped from the Wikipedia page and it was cleaned by removing extra parts of the code from in between the data and other extra characters through multiple iterations. The scraped and cleaned list of suburbs was then made into a dataframe and stored as a csv file.
- While adding the latitudes and longitudes of each of the suburbs to the dataset, using the geopy package, the location 'Sydney, Australia' was added to the address to get the accurate location as there can possibly be many places with the same name in different cities.
- The link to the data set is available here.
- The link to the code used for scraping is available here.

| | Suburb | Latitude | Longitude |
|---|---|---|---|
| **0** | Abbotsbury | -33.869285 | 150.866703 |
| **1** | Abbotsford | -33.850553 | 151.129759 |
| **2** | Acacia Gardens | -33.732459 | 150.912532 |
| **3** | Agnes Banks | -33.614508 | 150.711448 |
| **4** | Airds | -34.090000 | 150.826111 |
| **...** | ... | ... | ... |
| **673** | Yennora | -33.862008 | 150.968610 |
| **674** | Yowie Bay | -34.050278 | 151.103333 |
| **675** | Zetland | -33.907662 | 151.208218 |
| **676** | Jordan Springs | -33.725022 | 150.727127 |
| **677** | South Granville | -33.857778 | 151.010556 |

Figure 1: Dataset of suburbs in Sydney

## 2.2 List of Train Stations in Sydney and their locations

- Similar to the dataset for the suburbs, the dataset for the railway stations was scraped using beautifulsoup from the Wikipedia page. It was cleaned

by removing extra parts of the code from in between the data and other extra characters through multiple iterations. The words 'Railway Station' were added to each of the names of the stations to make locating them easier. The scraped and cleaned list of railway stations was then made into a dataframe and stored as a csv file.

- While adding the latitudes and longitudes of each of the stations to the dataset, using the geopy package, the location 'Sydney, Australia' was added to the address to get the accurate location as there can possibly be many places with the same name in different cities.
- The link to the data set is available here.
- The link to the code used for scraping is available here.

| | Station | Latitude | Longitude |
|---|---|---|---|
| **0** | Mount Kuring-gai Railway Station | -33.653064 | 151.136886 |
| **1** | Ingleburn Railway Station | -33.998333 | 150.851111 |
| **2** | Yennora Railway Station | -33.862008 | 150.968610 |
| **3** | Epping Railway Station | -33.772863 | 151.082090 |
| **4** | Rooty Hill Railway Station | -33.768865 | 150.841180 |
| **...** | ... | ... | ... |
| **163** | Mulgrave Railway Station | -33.627219 | 150.831660 |
| **164** | Croydon Railway Station | -33.877794 | 151.115650 |
| **165** | Punchbowl Railway Station | -33.928717 | 151.052259 |
| **166** | Mascot Railway Station | -33.928956 | 151.195178 |
| **167** | Summer Hill Railway Station | -33.893395 | 151.136873 |

Figure 2: Dataset of train stations in Sydney

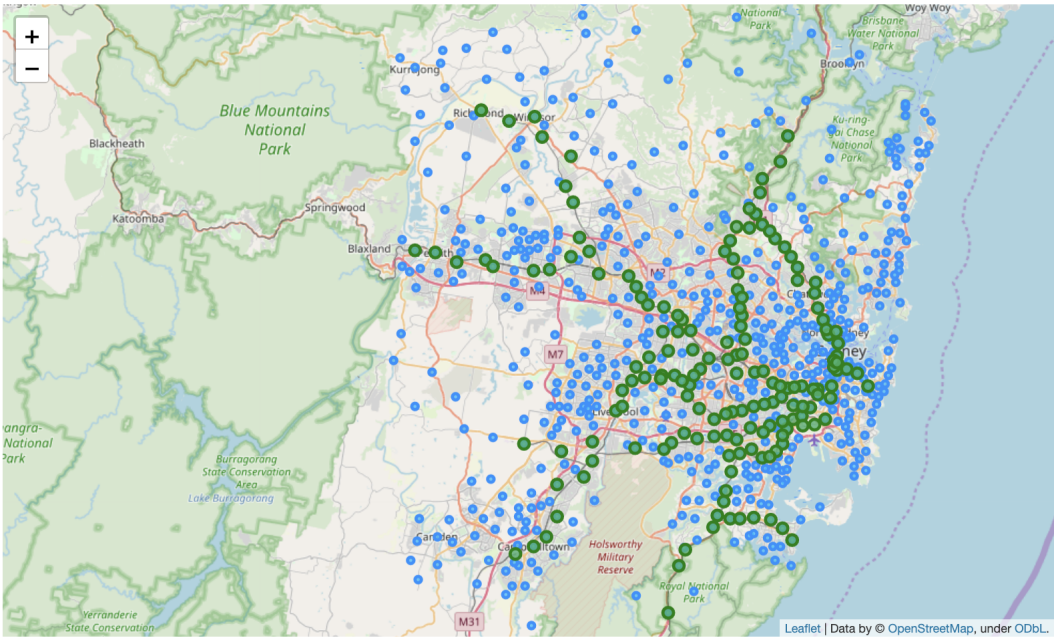


Figure 3: Visualization of the stations and suburbs in Sydney

## 2.3 Top food places and their locations

- The trending food places in each suburb are obtained by using the Foursquare API on the suburbs dataset.
- The category ID for food places (4d4b7105d754a06374d81259) is used to return only the food places in the specified areas.
- The required details of each venue such as latitude and longitude are obtained from the json file generated.

| | Suburb | Suburb Latitude | Suburb Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Abbotsbury | -33.869285 | 150.866703 | Star Buffet (Club Marconi) | -33.864698 | 150.879435 | Buffet |
| 1 | Abbotsbury | -33.869285 | 150.866703 | Hero Sushi | -33.858125 | 150.898700 | Sushi Restaurant |
| 2 | Abbotsbury | -33.869285 | 150.866703 | Soju | -33.857248 | 150.898864 | Korean Restaurant |
| 3 | Abbotsbury | -33.869285 | 150.866703 | McDonald's | -33.888236 | 150.881752 | Fast Food Restaurant |
| 4 | Abbotsbury | -33.869285 | 150.866703 | Maison Coffee | -33.871530 | 150.890440 | Café |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 12419 | South Granville | -33.857778 | 151.010556 | Subway | -33.863021 | 151.044586 | Sandwich Place |
| 12420 | South Granville | -33.857778 | 151.010556 | Suliya Gogiya | -33.864365 | 151.044980 | BBQ Joint |
| 12421 | South Granville | -33.857778 | 151.010556 | I Love Manoush | -33.864015 | 151.045103 | Pizza Place |
| 12422 | South Granville | -33.857778 | 151.010556 | Darband Persian Restaraunt | -33.849675 | 151.034835 | Persian Restaurant |
| 12423 | South Granville | -33.857778 | 151.010556 | Sangak Bakery & Restaurant | -33.853684 | 150.984464 | Kebab Restaurant |

Figure 4: Dataset obtained from Foursquare API

# 3 Methodology

In this project, the plan is to cluster the food places based on their distances from the nearest railway stations.

In the first step, we collected the required data as shown above by data scraping, cleaning and using the geopy package and the Foursquare API.

The second step is the calculation and exploration of the nearest railway station to each food place and the distance between them. This data is then combined with the dataset of top food places that was previously obtained. To find the closest railway station to each restaurant, the distance between each restaurant returned by the call to the Foursquare API and train station is calculated using the geopy package. Then the station with the least distance is concluded to be the nearest station to each food place.

The third and final step will be focusing on clustering these food places based on their distances from the nearest railway stations and exploring the clusters made. A visualization of the clusters obtained will be made by plotting a few points of each cluster on a map along with the railway stations. The clusters obtained can be used to explore the restaurants and suburbs as per the requirement of the traveler.

# 4 Analysis

After exploring the details of each cluster obtained by using the describe function, it can be seen as to which cluster covers what range of distances and which restaurant falls in which cluster.

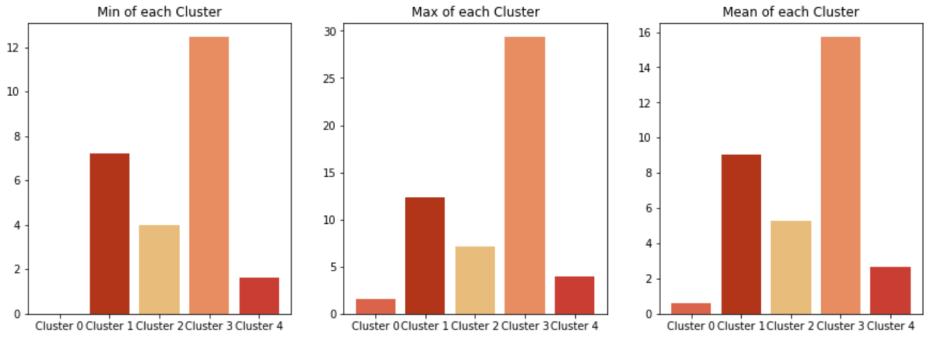


Figure 5: Visual representation of the details obtained using the describe function

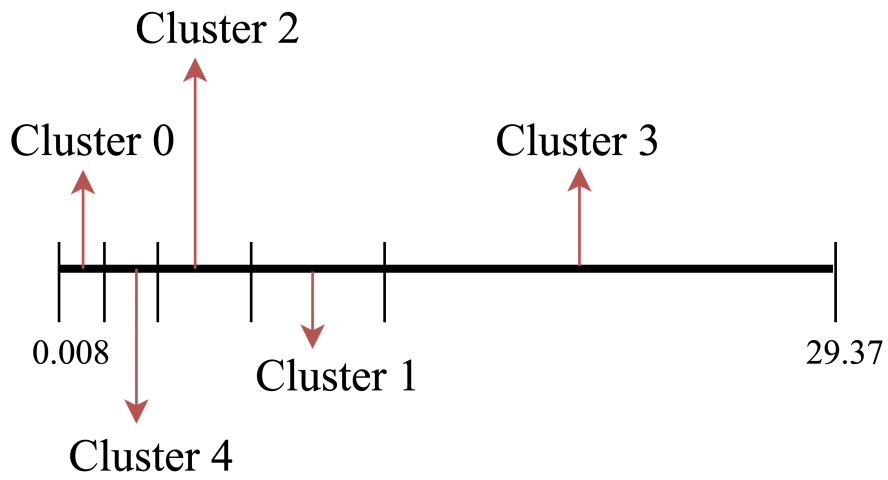


Figure 6: Division of distances and ranges covered by each cluster

From the above graphs, we can see that cluster 0 has the restaurants in walk-able distance to a train station. To go to a restaurant in any of the other clusters, especially in cluster 3, it is better to arrange for another mode of transport.

On plotting a few points from each cluster and the train stations on the map of Sydney, this is how it looks:

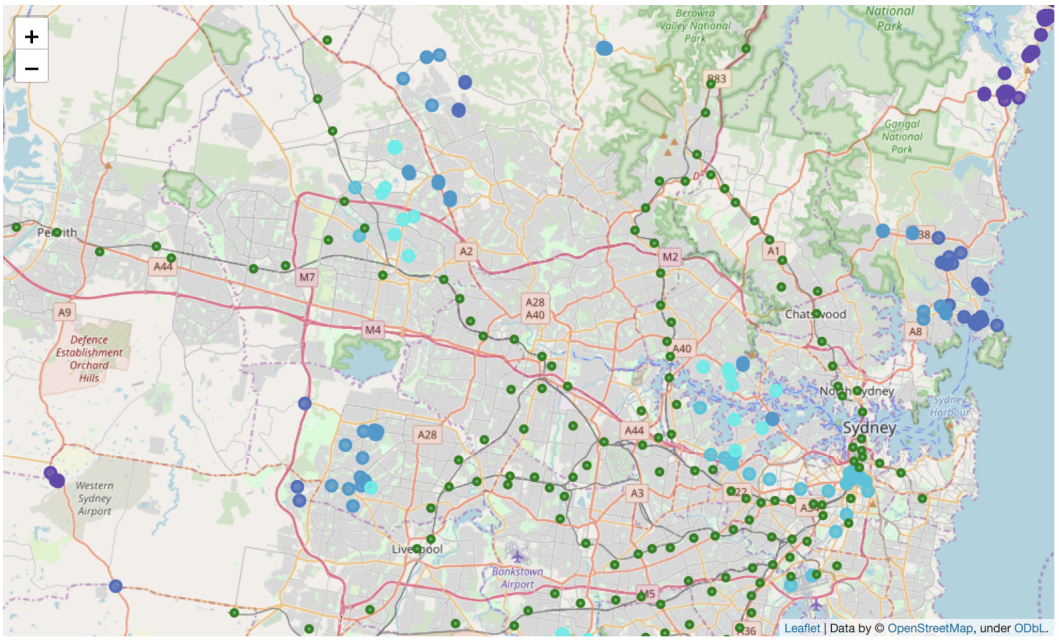


Figure 7: Visualization of the stations and a few restaurants from each cluster

# 5 Results and Discussion

On close inspection of the details of the clusters, we see that cluster 0 has the places within walk-able distance of train stations with a maximum distance of 1.6 kilometers.

Suburbs like Strathfield, Lavender Bay, Darling Point, Ashfield and a few other suburbs have a majority of food places which are nearest to train stations.

The Chatswood Railway Station has the most food places within walk-able distance.

Sushi Bay is a Japanese restaurant which is most closest to a train station from the list of trending restaurants obtained. Hungry Jack's, Brick Pit Espresso, Gelato Messina and The Shop Picnic Point are a few of the many other food places located within walk-able distance of a train station.

Multiple other inferences can be made from the analysis done according to the needs and interests of the traveler.

*Please note that it is only for this specific run of the code and the restaurants given for this specific call of the Foursquare API that cluster 0 has the restaurants within walk-able distance of the train stations with the measures given above. It is possible that if the code is run again, the call to the Foursquare API may give different results depending on the trending venues at that particular point of time which may change the details of each cluster obtained in cells 80 to 84. It is not necessary that cluster 0 always has the minimum distance restau-*

*rants. It is important to understand how the clustering algorithm works on the distances array given as the input - the algorithm divides the different distances into different clusters so that each cluster covers a particular range of distances. On studying the details of each cluster given by the describe function as shown above, it can be understood as to which range is covered by which cluster.*

## 6 Future Direction

This project can be further extended by creating analysis of food places closest to both train stations and the iconic places to visit in Sydney like the Opera House and its famous beaches.

The same analysis can be done on other cities as well.

# Bon Voyage and Bon Appetit!