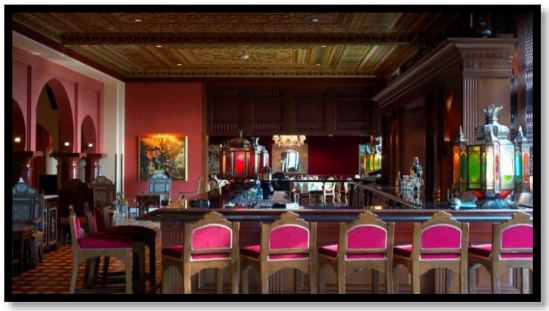
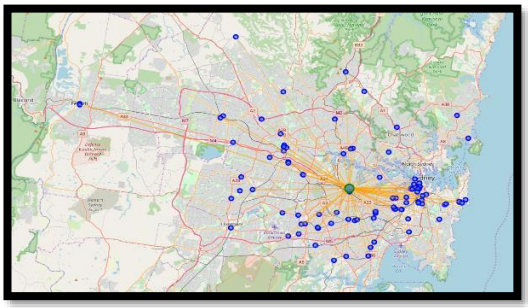
IBM Data Science – Capstone

Choosing the Best Cities for Middle Eastern Eateries in Australia:

An exploratory analysis of Sydney, Melbourne, Canberra, Brisbane and Perth





Farooq Yousaf

March 23, 2020

Table of Contents

1) Introduction and problem Description

In this capstone project, I will use my overall learning from IBM's Professional Data Science Certificate to solve a problem that I often encounter while living in Australia. Moreover, it is also a common issue faced by many local and foreign tourists who to major metropolitans in Australia, however, they are overwhelmed by food choices. Hence, I will use this project to "explore" which major metropolitans in Australia are the best options for visiting restaurants and eateries offering "**Middle Eastern**" cuisine. The cities that I will analyze include; Sydney, Melbourne, Canberra, Perth and Brisbane. Even though Sydney is the densest city by population (ABS, 2020), and normally it would be assumed that it would be easier and more convenient to visit various venues in the city, however, from my personal experience, it takes longer to explore venues in Sydney than in Melbourne or Perth. Hence, this project tries to explore why is that the case, especially when Sydney has more venues than any other city in Australia.

2) Data Description

2.1. Requirements

The data required for this project is mainly Middle Eastern restaurants based in the major cities of Australia; namely Sydney, Melbourne, Perth, Brisbane, and Canberra. For that purpose, I will use the Foursquare API to extract data of Middle Eastern restaurants in the cities under discussion. The major values from that data that are required for this project are the address, latitude and longitude. Moreover, for analysis purposes, Wikipedia and the Australian Bureau of Statistics (ABS) will also be used to expand our analysis. To analyse our data and create data frames, we will use *pandas*, *NumPy* and *folium* (for maps) libraries.

```
[ ] import numpy as np
    import pandas as pd
    pd.set_option('display.max_columns', None)
    pd.set_option('display.max_rows', None)
    import requests
    from pandas.io.json import json_normalize #we will use this normalize our nested JSON file from Foursquare
    import folium # Map library
```

(Imported Libraries)

2.2. Sources

The data for this project will be collected using the Foursquare API using the following link: https://foursquare.com/developers. To extract our data, we will use our Client ID and Client Secret to create our query in Python. On the other hand, the secondary data for Australia will be collected from the following sources:

Wikipedia: https://en.wikipedia.org/wiki/List_of_cities_in_Australia_by_population

ABS: https://www.abs.gov.au/

2.3. Audience

The audience for this report is local and overseas travellers who would like to try Middle Eastern restaurants in Australia. Moreover, this project will help me personally, as it enables me to select and analyse various food districts in Australia, which will further help me in my travelling.

3) Methodology

As John Rollins rightly argues, "lack sufficient understanding of how to go about solving problems using data science techniques" results in failure of adequately addressing the problem at hand (Rollins, 2015). This is the very reason we need a coherent methodology that could not only define our problem and our data sources but also chalks out a plan to solve that problem. In section 1 and 2 of this report, we started with introducing the problem, location and our data sources. In that regard, we need to use a simple exploratory data analysis and map visualizations to find out which major cities in Australia are the best for local and foreign tourists in terms of eating at Middle Eastern restaurants.

Data Preparation and Modelling

The cities that we selected from the analysis included Sydney, Melbourne, Canberra, Perth and Brisbane. To analyse the Middle Eastern restaurants in these cities, we required the data for venues from Foursquare API. Hence, we used the Foursquare API to search for Middle Eastern venues in our sample set of cities using the following code:

```
[ ] LIMIT = 100 #limiting queries to a 100
    cities = ['Sydney, NSW', 'Melbourne, VIC', 'Perth, WA', 'Canberra, ACT', 'Brisbane, QLD'] #our five target cities and thier states
    results = {}
    for city in cities:
        url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&near={}&limit={}&categoryId={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            city,
            LIMIT,
            "4bf58dd8d48988d115941735") # Foursquare Code for Middle Eastern Eateries
        results[city] = requests.get(url).json()
```

(Foursquare API Code)

As our data was in a highly nested JSON format, we turned it into a Pandas Dataframe using *json_normalize*. After normalizing the data, we had the data in our required format of "Name", "Address", "Lat" and "Lng".

```
[ ] df_venues={}
    for city in cities:
        venues = json_normalize(results[city]['response']['groups'][0]['items'])
        df_venues[city] = venues[['venue.name', 'venue.location.address', 'venue.location.lat', 'venue.location.lng']]
        df_venues[city].columns = ['Name', 'Address', 'Lat', 'Lng']
```

After normalizing our data, we get the Name, Address, Lat and Long coordinates

```
[ ] df_venues
```

```
{'Brisbane, QLD':                                          Name  \
0                                        Naïm
1                          Gad's: Charcoal Chicken
2                                        Byblós
3   Arabella's Charcoal and Middle Eastern Cuisine
4                         Sunshine Kebabs Underwood
5   Arabella's Charcoal and Middle Eastern Cuisine
6                                    Baba Ganouj
7                                 Watany Manoushi
8                                 Sinbad's Kebabs
9                          Baalbek Lebanese Cuisine
10                                       Naïm
11                                   1001 Nights
12                                   ISPA Kebabs
13                         Farah Persian Restaurant
14                                   Baba Ganouj
15                          Rockys Bakehouse & Cafe
16                         King Ahiram Lebanese Food
17                              NoNo's Lebanese Food

                                      Address        Lat         Lng
0                          14 Collingwood St -27.458005  152.994333
1                                        NaN -27.560947  153.083769
2                Shop 7.13, Portside Wharf -27.441268  153.069326
3                      149-151 Boundary St -27.479606  153.012340
4                                        NaN -27.609529  153.112593
5                           1932 Logan Rd -27.557887  153.080813
6                                 Grey St -27.480853  153.022939
7                                        NaN -27.560835  153.083558
```

(Normalizing our Json Data)

Data Analysis

Once we have normalized our data, we will use a two-step approach to analyze our data. In the first step, we will plot the data normally as clusters of different venues in our sample cities. In our second step, we will calculate the geographical centres of the city and the calculate the mean distance of venues from that centre. The analysis, even though not directly in our scope, will also tell us how Middle Eastern restaurants are clustered in each of the five cities, indicating the diversity of the cities.

4) Results

The results of this project were divided into two parts. The first part of maps consists of data without the mean distance of locations from the geographical distance of the city. Those values just visualize the locations on the map. The second results are locations and their average mean distance from the geographical centre of the city. The following are our results:

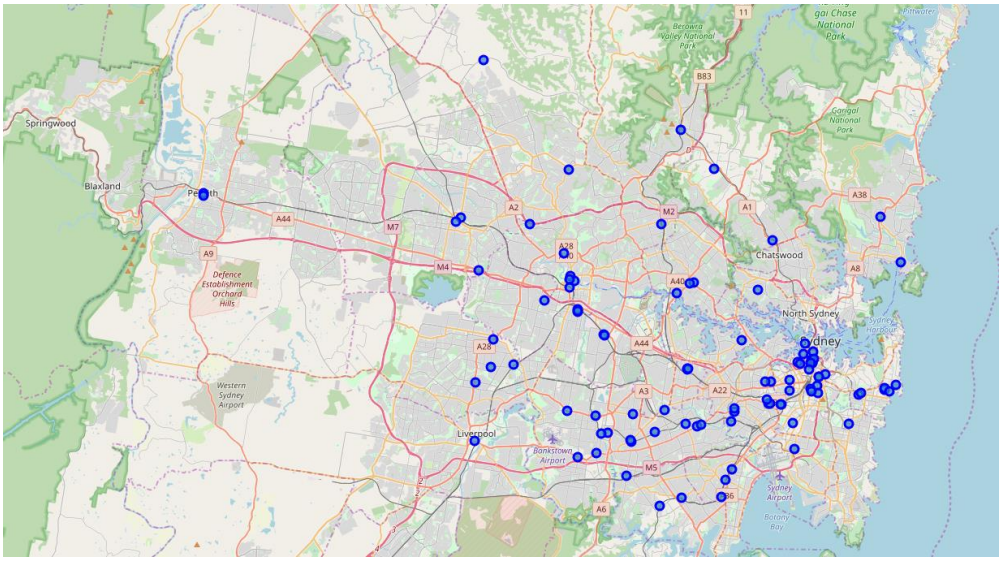
4.1. Results with normal plotting

While plotting our data for various cities, we first normally plotted the venues gathered from Foursquare API, along with calculating the central locations of all the cities. However, as we do not need to calculate the distance of venues from the centre, we first normally plotted the data.

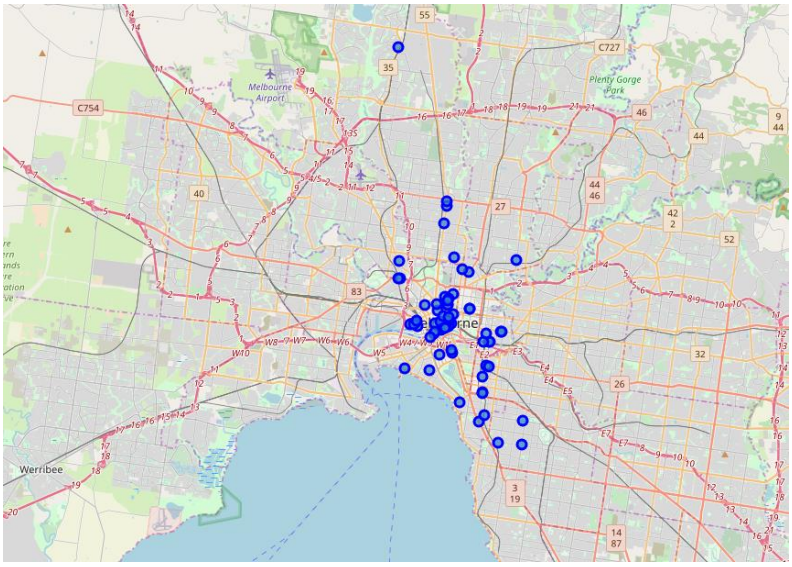
```
[ ]  maps = {}
     for city in cities:
         city_lat = np.mean([results[city]['response']['geocode']['geometry']['bounds']['ne']['lat'],
                             results[city]['response']['geocode']['geometry']['bounds']['sw']['lat']])
         city_lng = np.mean([results[city]['response']['geocode']['geometry']['bounds']['ne']['lng'],
                             results[city]['response']['geocode']['geometry']['bounds']['sw']['lng']])
         maps[city] = folium.Map(location=[city_lat, city_lng], zoom_start=11)

         # add markers to map
         for lat, lng, label in zip(df_venues[city]['Lat'], df_venues[city]['Lng'], df_venues[city]['Name']):
             label = folium.Popup(label, parse_html=True)
             folium.CircleMarker(
                 [lat, lng],
                 radius=5,
                 popup=label,
                 color='blue',
                 fill=True,
                 fill_color='#3186cc',
                 fill_opacity=0.7,
                 parse_html=False).add_to(maps[city])
```

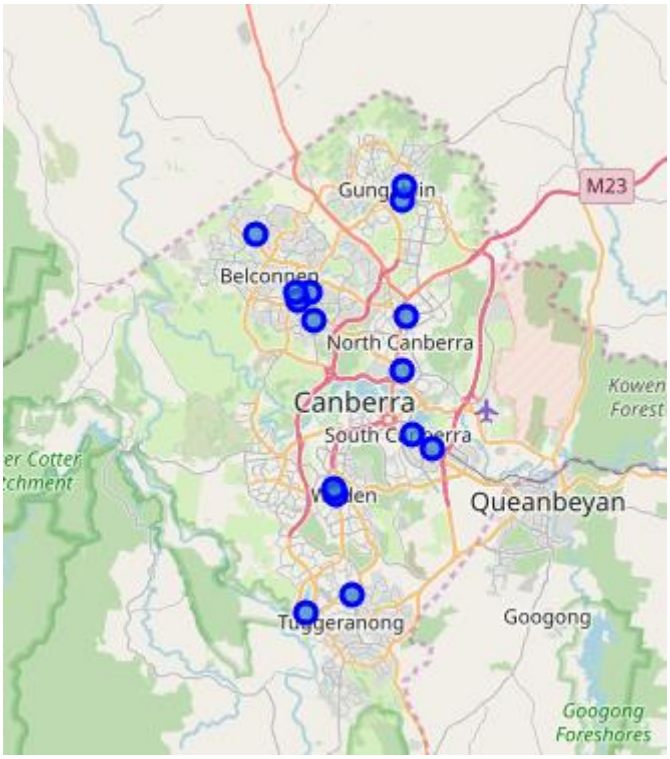
Sydney



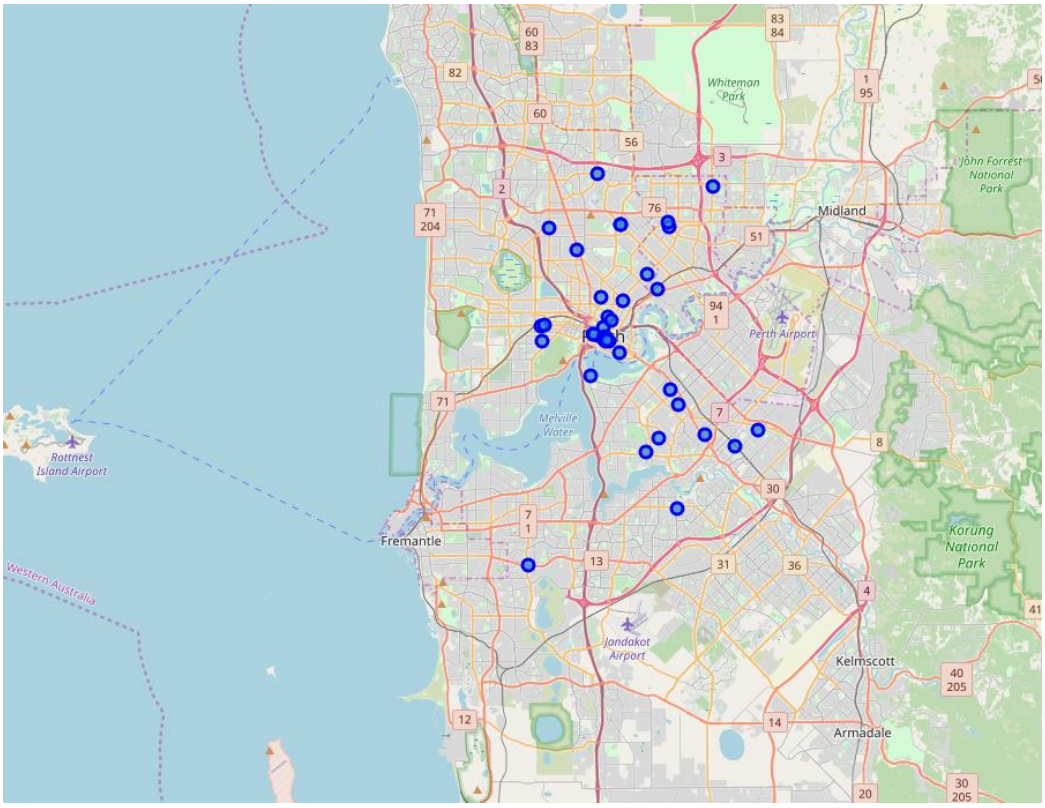
Melbourne



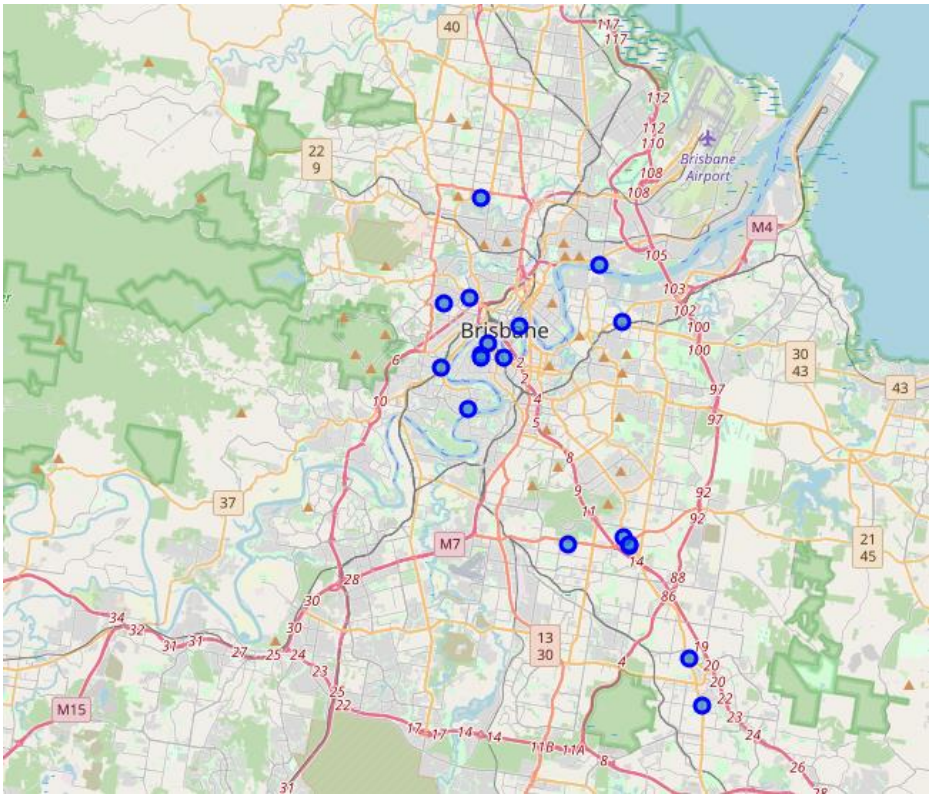
Canberra



Perth



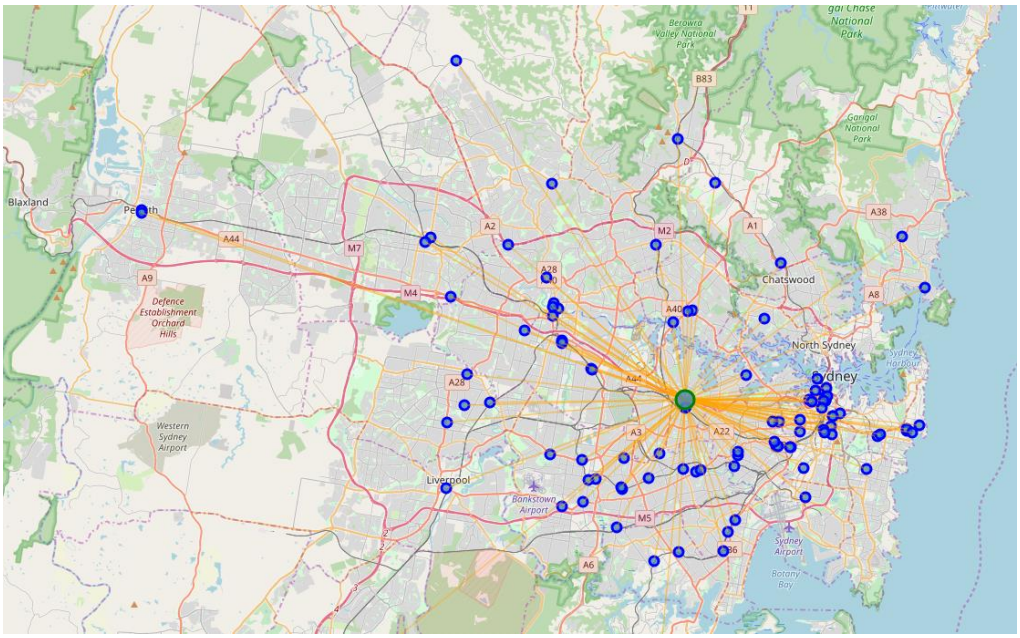
Brisbane



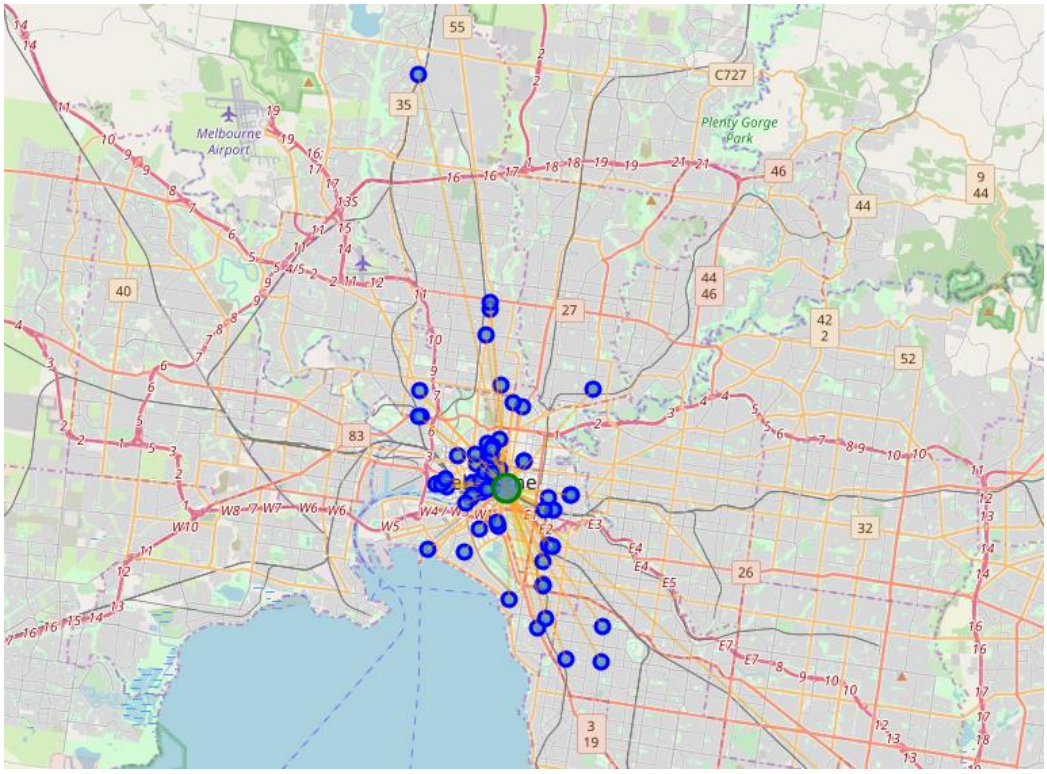
4.2. Results with mean distance

Now that we have seen the clusters plotted on the map, we calculate the mean distance (average distance) of venues from the geographical centre of the city. The plot gives us a clearer picture of the true distance for a tourist or a traveller to visit various venues.

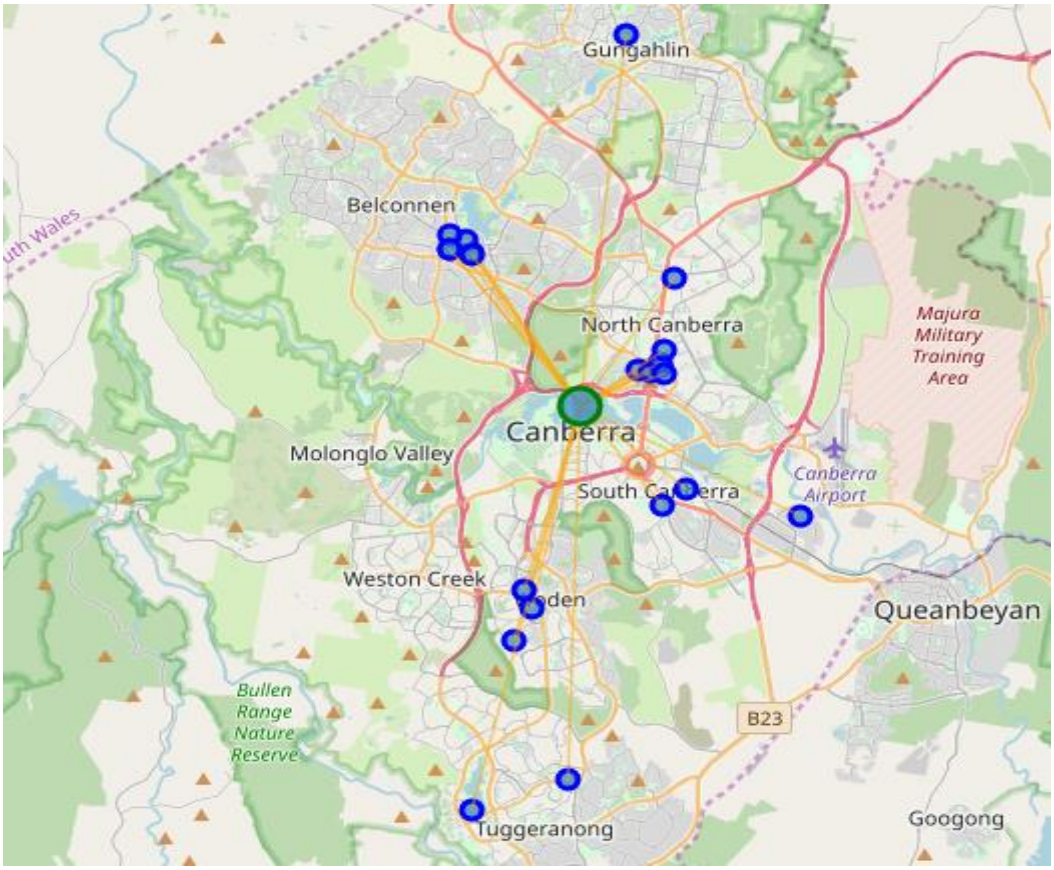
Sydney



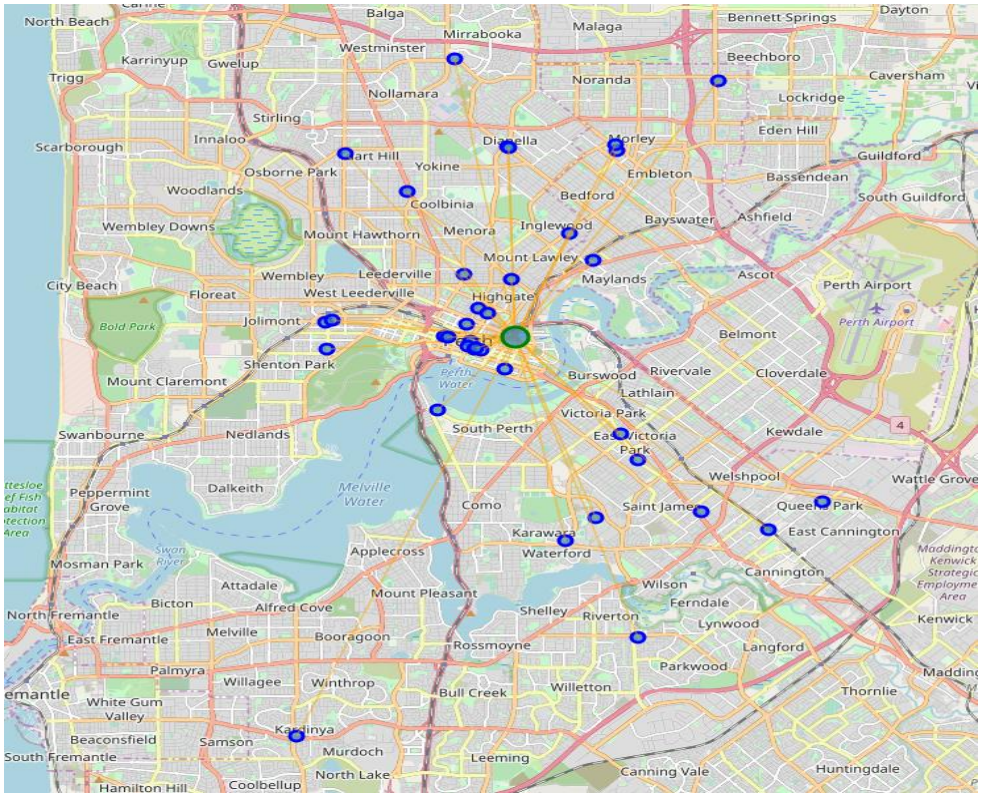
Melbourne



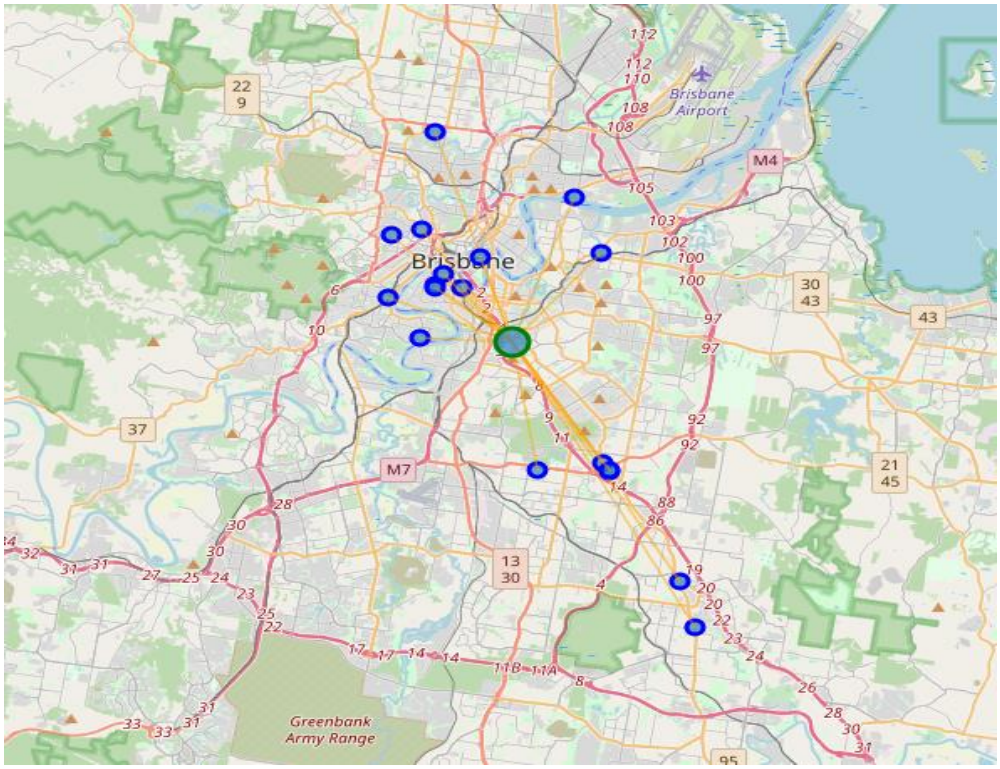
Canberra



Perth



Brisbane



4.3. Values of Mean Distance of each city

Even though the mean distance value gives us a clear picture, we now calculate the mean distance by excluding the maximum value of the “outliers”, to get a

homogenous value. For this purpose, we use the *NumPy* mean function. We use the following code for each city:

```
[ ] city = 'Sydney, NSW'
    venues_mean_coor = [df_venues[city]['Lat'].mean(), df_venues[city]['Lng'].mean()]

    print(city)
    print("Mean Distance from Mean coordinates")
    dists = np.apply_along_axis(lambda x: np.linalg.norm(x - venues_mean_coor),1,df_venues[city][['Lat','Lng']].values)
    dists.sort()
    print(np.mean(dists[:-1]))
```

(Code to normalise the mean values without outliers)

Sydney, NSW

Mean Distance from Mean coordinates: 0.11287612594187721

Canberra, ACT

Mean Distance from Mean coordinates: 0.07047281616415198

Melbourne, VIC

Mean Distance from Mean coordinates: 0.03318223070418223

Perth, WA

Mean Distance from Mean coordinates: 0.04176312230297355

Brisbane, QLD

Mean Distance from Mean coordinates: 0.05986191980294174

5) Discussion

The findings from our analysis present an elaborate picture. As seen from our findings, when the venues were plotted without the mean distance, the map, for a layman, might have shown that both Melbourne and Sydney might seem like good clusters with higher density. However, our mean distance shows that even though Melbourne is indeed good, Sydney does not present a convenient picture as the mean distance for Sydney is the highest compared to all other cities. This is because Sydney has a total area of 12,368 km^2, whereas other cities have the following areas: Brisbane 15,826 km^2, Melbourne 9,990 km^2, Perth 6,418 km^2, and Canberra 814.2 km^2 (Wikipedia, 2020). However, even though Brisbane, area-wise, is the largest "proper" city in Australia, it has a lower population than Sydney and Melbourne. Hence, this also suggests why Brisbane has fewer venues than Sydney and Melbourne because the latter cities have a higher population than Brisbane. Hence, our findings suggest that the best location for any tourist to try Middle Eastern food in Melbourne, as they will have to travel less to explore different venues in the city. After Melbourne, the two best options, in terms of less travel, are Perth and Canberra. The reason both

Sydney and Brisbane, even being major metropolitans in Australia, have a higher mean distance in terms of locations is that both the cities are area wise the biggest cities in Australia.

6) Conclusion

The project initially started with defining the problem statement, which questioned which major cities in Australia were the best for local and foreign tourists to visit Middle Eastern restaurants. Initially, from the normal clusters, it would seem that Melbourne and Sydney were indeed the best locations to visit the Middle Eastern venues. However, the mean distance suggested that Sydney, with a larger area, had a higher mean distance between the geographic centre of the city and all the venues. Similarly, Melbourne was the best city in this regard as the mean distance was the lowest, suggesting that tourists would have to travel less to explore different venues. This project also explained how data science can solve even the simplest of problems in a more attractive manner.

References

ABS. (2020). Australian Demographic Statistics, Sep 2019. Retrieved from https://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0

Rollins, J. (2015). Why we need a methodology for data science. *Big Data and Analytics Hub*. Retrieved from https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science

Wikipedia. (2020). List of cities in Australia by population. Retrieved from https://en.wikipedia.org/wiki/List_of_cities_in_Australia_by_population