Capstone Project – Battle of the Neighborhoods

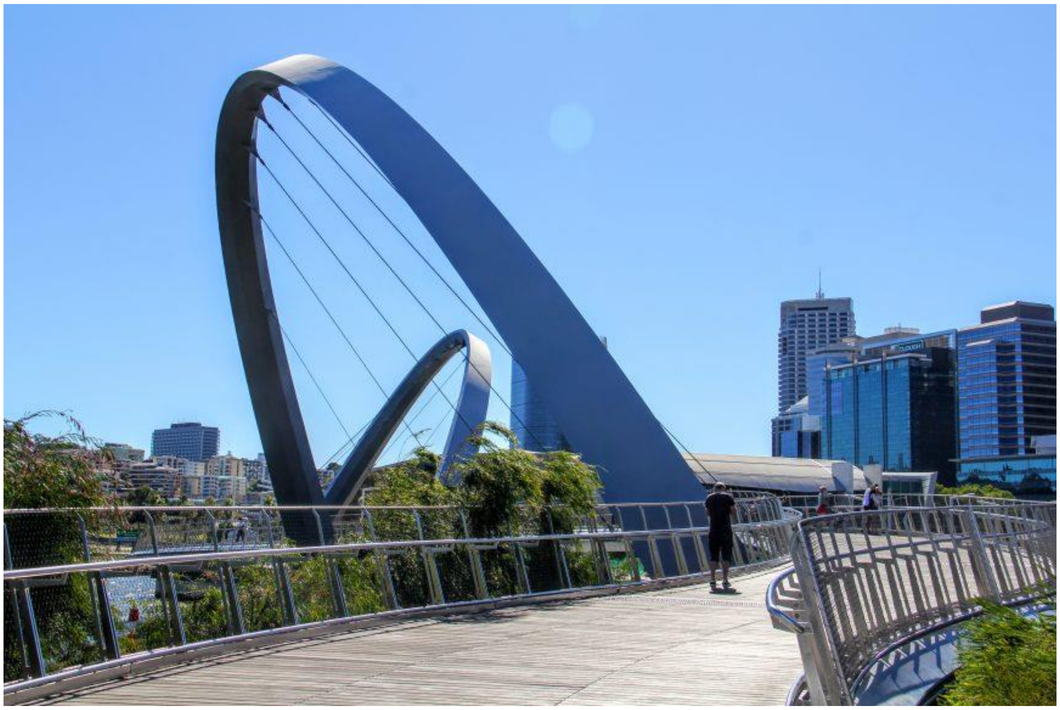
1. Business problem:

to determine where might be the 'best' neighborhood in Perth to start a tourist's food journey to satisfy their gastronomical needs.

1.1. Introduction:

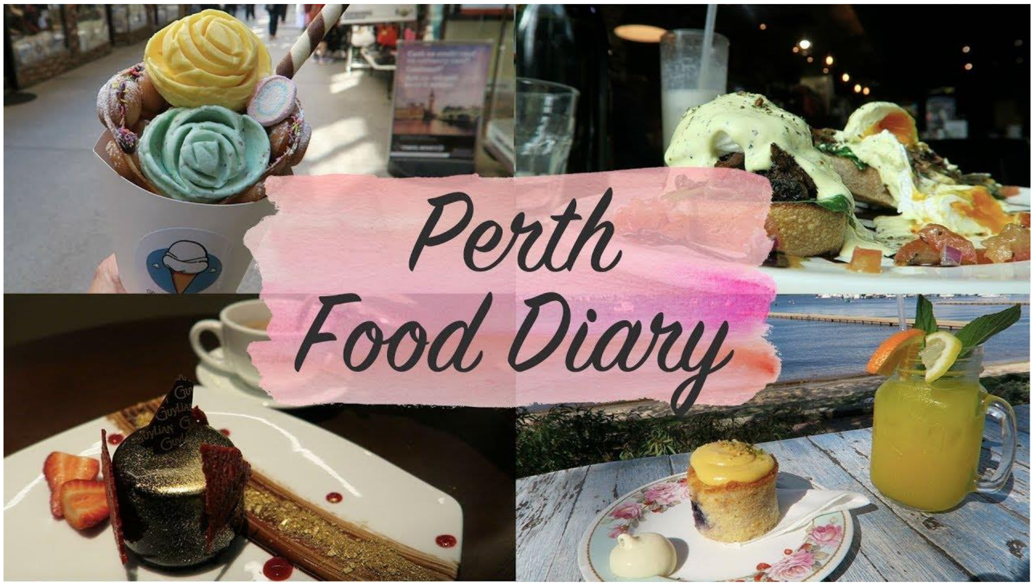
Perth, capital of Western Australia, sits where the Swan River meets the southwest coast. Sandy beaches line its suburbs, and the huge, riverside Kings Park and Botanic Garden on Mount Eliza offer sweeping views of the city. In Perth, you can embrace the best of both worlds, where soft-sand beaches and scenic parks meet a thriving metropolis of small bars, creative restaurants and curated street art.

The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the 'best' neighborhood in Perth to start your food journey. Through this project, we will find the most suitable location for a "foodie" to satisfy their gastronomical needs.



1.2. Target audience:

This project is aimed toward "foodies" who want to explore Perth. By using concepts in data science, we will help them use this information to make informed decision where to start their food journey, so they can optimize their visit in the capital city of Western Australia.



2. Data requirements:

- The data required will be a combination of CSV files from https://www.matthewproctor.com/full_australian_postcodes_wa which will provide the list of 'localities' including its geographical coordinates.
- In addition, we will also use the Foursquare data to find where restaurants are located within a certain neighborhood.

3. Methodology:

3.1. Data preparation

After collecting all the data and putting it into data frames, cleaning and merging of the data was required before starting analysis. Data was acquired from (https://www.matthewproctor.com/full_australian_postcodes_wa); there were localities that has duplicate values, in addition some localities have no coordinates. Therefore, the following assumptions were made:

- Only localities and postcodes that has no duplicates will be processed.
- Only localities and postcodes that has value will be processed.

After the implementation of the following assumptions, the following data frame was produced.

Out[63]:

	Postcode	Locality	State	Longitude	Latitude
1	0872	GIBSON DESERT NORTH	WA	131.298809	-21.949513
10	6000	CITY DELIVERY CENTRE	WA	115.859912	-31.948762
13	6001	PERTH	WA	115.763228	-31.99212
14	6003	HIGHGATE	WA	115.869136	-31.939272
16	6004	EAST PERTH	WA	115.874601	-31.956931
17	6005	KINGS PARK	WA	115.836896	-31.95707
19	6006	NORTH PERTH	WA	115.852913	-31.92934
20	6007	LEEDERVILLE	WA	115.834335	-31.935675
22	6008	DAGLISH	WA	115.811432	-31.956599
26	6009	BROADWAY NEDLANDS	WA	115.804692	-31.985791
31	6010	CLAREMONT	WA	115.77639	-31.971647

3.2. Data Exploration

Now after cleansing the data, the next step was to analyze it. We then created a map using Folium and color-coded each locality.

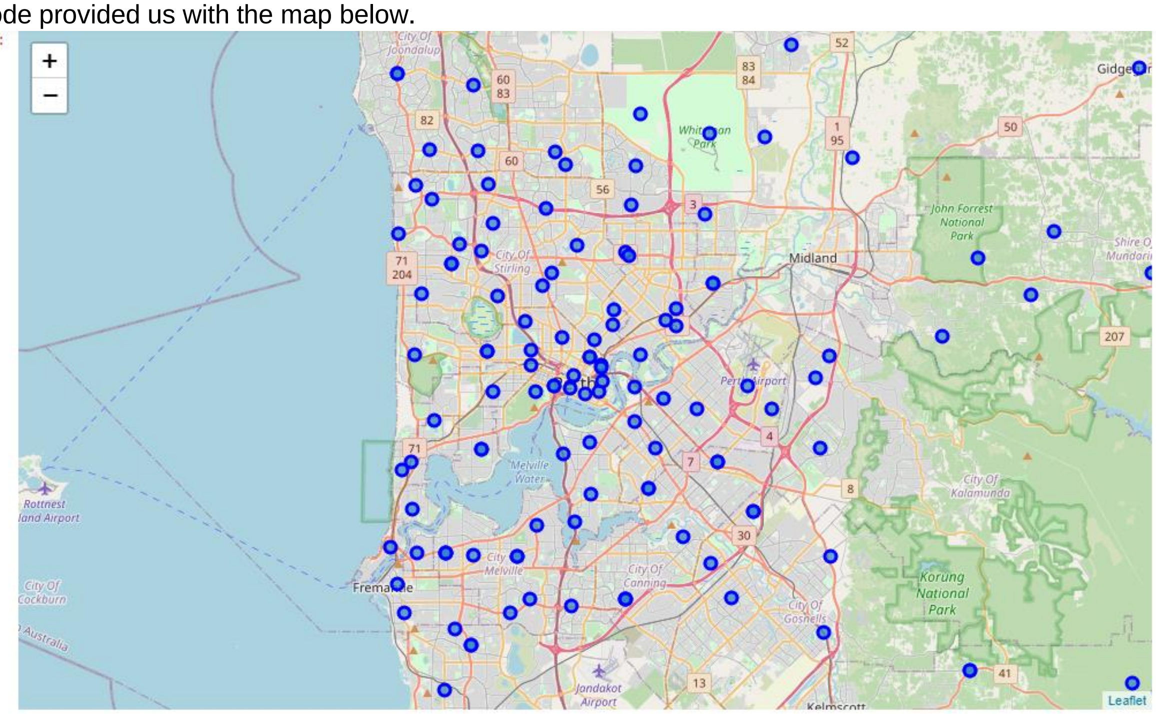
In [78]:
```python
# create map of Perth using latitude and longitude values
map_perth = folium.Map(location=[latitude, longitude], zoom_start=11)

 # add markers to map
for lat, lng, label in zip(df_perth['Latitude'], df_perth['Longitude'], df_perth['Locality']):
    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(map_perth)

map_perth
```

This code provided us with the map below.

Out[78]:



Next, we used the Foursquare API to get a list of all the Venues in Perth which included Parks, Schools, Café Shops, Asian Restaurants etc. Getting this data was crucial in analyzing the number of Restaurants all over Perth. We then merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods.

Out[92]:

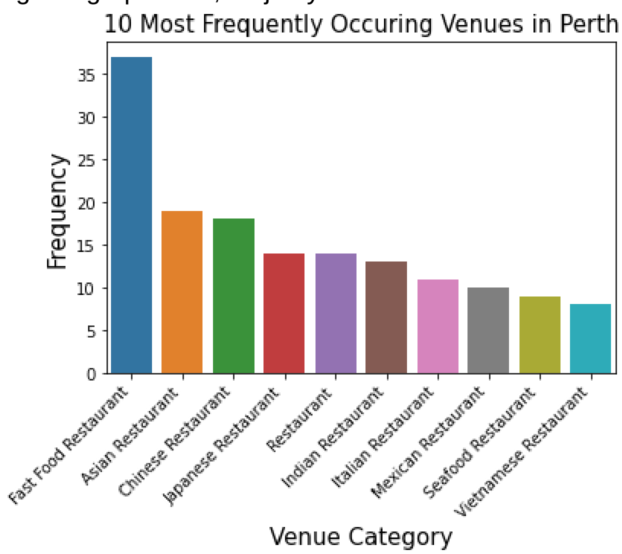
	name	categories	lat	lng
0	Toastface Grillah	Sandwich Place	-31.952441	115.860964
1	Le Vietnam	Vietnamese Restaurant	-31.954256	115.860512
2	Twilight Hawkers Markets	Food Truck	-31.953079	115.859452
3	Max + Sons	Coffee Shop	-31.951645	115.858371
4	Ribs & Burgers	Burger Joint	-31.951655	115.858375
5	Palsaik	Korean Restaurant	-31.954315	115.860515
6	Alfred's Pizzeria & Smallbar	Pizza Place	-31.954890	115.859901
7	lululemon	Clothing Store	-31.952688	115.860083

We found out that there were 74 venues returned by Foursquare.

Out[101]:

	Venue_Category	Frequency
0	Fast Food Restaurant	37
1	Asian Restaurant	19
2	Chinese Restaurant	18
3	Japanese Restaurant	14
4	Restaurant	14
5	Indian Restaurant	13
6	Italian Restaurant	11
7	Mexican Restaurant	10
8	Seafood Restaurant	9
9	Vietnamese Restaurant	8

Putting it in graph form, majority of the restaurant are Fast Food.



3.3. Data Analysis

Let's us analyze the data more and analyze the top 5 venues of each neighborhood.
By using pandas groupby on the neighborhood column, we will calculate the mean of the frequency of occurrence of each venue category.

Out[125]:

	Neighborhood	Afghan Restaurant	African Restaurant	Asian Restaurant	Australian Restaurant	Chinese Restaurant	Dim Sum Restaurant	Dumpling Restaurant	Eastern European Restaurant	Fast Food Restaurant
0	ALFRED COVE	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
1	ALKIMOS	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	APPLECROSS	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	ATWELL	0.0	0.000000	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.200000
4	AUSTRALIND	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
5	BALGA	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	BALLAJURA	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000
7	BATEMAN	0.0	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	BAYSWATER	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
9	BECKENHAM	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

	Neighborhood	French Restaurant	Greek Restaurant	Halal Restaurant	Indian Restaurant	Indonesian Restaurant	Italian Restaurant	Japanese Restaurant	Korean BBQ Restaurant	Korean Restaurant	Malay Restaurant	Mexican Restaurant	Middle Eastern Restaurant
0	ALFRED COVE	0.0	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1	ALKIMOS	0.0	0.000000	0.000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	APPLECROSS	0.0	0.000000	0.000	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000
3	ATWELL	0.0	0.000000	0.000	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.000000	0.200000	0.000000
4	AUSTRALIND	0.0	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	BALGA	0.0	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	BALLAJURA	0.0	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	BATEMAN	0.0	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000
8	BAYSWATER	0.0	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
9	BECKENHAM	0.0	0.000000	0.000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Output of each neighborhood with the top 5 most common venues.

```python
num_top_venues = 5

for hood in perth_grouped['Neighborhood']:
    print("----"+hood+"----")
    temp = perth_grouped[perth_grouped['Neighborhood'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_top_venues))
    print('\n')
```

```
----ALFRED COVE----
                         venue  freq
0          Fast Food Restaurant   1.0
1             Afghan Restaurant   0.0
2             Korean Restaurant   0.0
3  Vegetarian / Vegan Restaurant   0.0
4            Turkish Restaurant   0.0


----ALKIMOS----
                         venue  freq
0             Indian Restaurant   1.0
1             Afghan Restaurant   0.0
2             Korean Restaurant   0.0
3  Vegetarian / Vegan Restaurant   0.0
4            Turkish Restaurant   0.0
```

We then use the k-means to cluster the neighborhoods into 6 clusters.

Run k-means to cluster the neighborhood into 6 clusters.

In [145]:
```python
# set number of clusters
kclusters = 6

perth_grouped_clustering = perth_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(perth_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

Out[145]: `array([0, 4, 4, 4, 0, 4, 0, 4, 1, 1])`

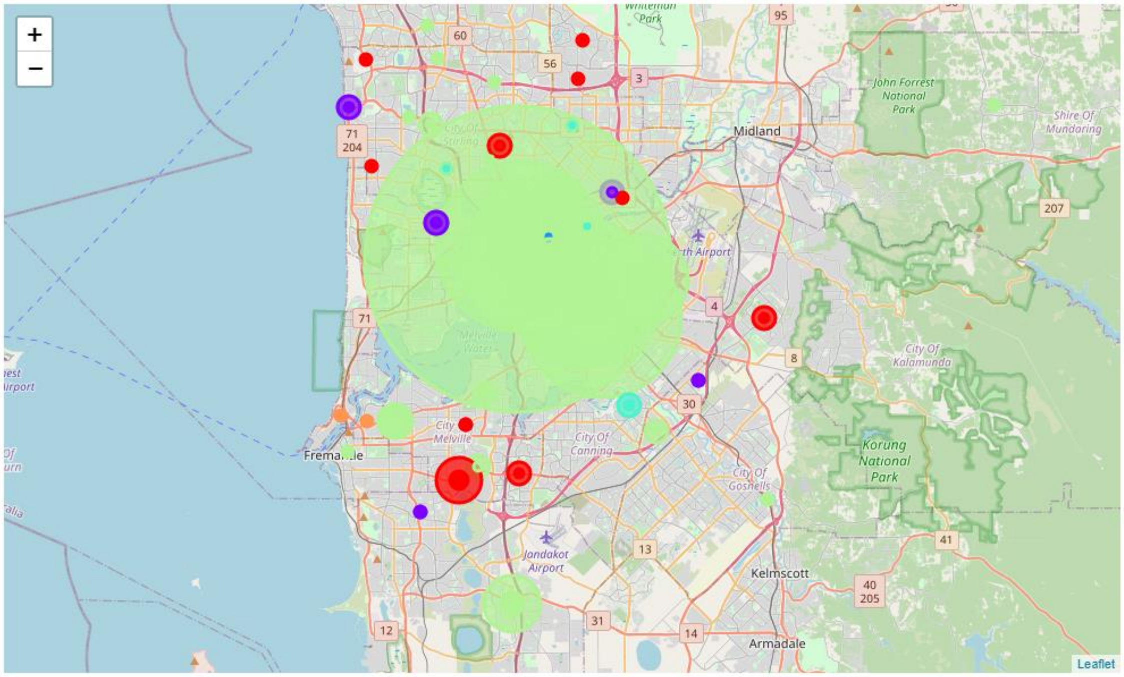
In addition, we will also display the top 10 most common venues that is visited by tourist in each neighborhood.

Out[159]:

	Postcode	Neighborhood	State	Longitude	Latitude	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
10	6000	CITY DELIVERY CENTRE	WA	115.859912	-31.948762	4.0	Chinese Restaurant	Vietnamese Restaurant	Ramen Restaurant	Asian Restaurant	Indian Restaurant	Vegetarian / Vegan Restaurant	Mexican Restaurant	Middle Eastern Restaurant	Portuguese Restaurant	Japanese Restaurant
14	6003	HIGHGATE	WA	115.869136	-31.939272	4.0	Mexican Restaurant	Eastern European Restaurant	Vegetarian / Vegan Restaurant	Tapas Restaurant	Fast Food Restaurant	Restaurant	Indian Restaurant	Portuguese Restaurant	Vietnamese Restaurant	Greek Restaurant
16	6004	EAST PERTH	WA	115.874601	-31.956931	4.0	Indian Restaurant	Japanese Restaurant	Italian Restaurant	Asian Restaurant	Chinese Restaurant	Restaurant	Greek Restaurant	African Restaurant	Australian Restaurant	Dim Sum Restaurant
19	6006	NORTH PERTH	WA	115.852913	-31.929340	4.0	Italian Restaurant	Asian Restaurant	Fast Food Restaurant	Portuguese Restaurant	Indian Restaurant	Vietnamese Restaurant	African Restaurant	Australian Restaurant	Chinese Restaurant	Dim Sum Restaurant
22	6008	DAGLISH	WA	115.811432	-31.956599	4.0	French Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	African Restaurant	Asian Restaurant	Australian Restaurant	Chinese Restaurant	Dim Sum Restaurant	Dumpling Restaurant	Eastern European Restaurant
39	6014	FLOREAT	WA	115.808169	-31.936389	1.0	Malay Restaurant	Seafood Restaurant	Vietnamese Restaurant	Eastern European Restaurant	Indian Restaurant	Halal Restaurant	Greek Restaurant	French Restaurant	Fast Food Restaurant	Dumpling Restaurant
46	6017	HERDSMAN	WA	115.814562	-31.908357	3.0	Australian Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	African Restaurant	Asian Restaurant	Chinese Restaurant	Dim Sum Restaurant	Dumpling Restaurant	Eastern European Restaurant	Fast Food Restaurant

Let us then visualize the clusters.

Out[181]:



4. Discussion and recommendations:

Most of the restaurants are in cluster 4 followed by cluster 0. Even though there is a huge number of Neighborhoods in cluster 0, majority of them are Fast Food Restaurant. We see that in cluster 1 and cluster 5, there are only less than 5 restaurants to satisfy your food journey. Looking at the nearby venues, the optimum place to start your food journey is in cluster 4, where there are many restaurants in the area that includes Chinese, French, Middle Eastern and Mexican to name a few. The second-best cluster would be cluster 1, where they offer seafood, Thai and Malay cuisine.

One must take into consideration that the clustering is completely based on data obtained from the Foursquare API. Also, the analysis does not take into consideration of the quality of the food, price and parking, etc.

5. Conclusion

In summary, this project is aimed for tourists that want to optimize their food journey when visiting Perth. We made use of numerous Python libraries like Pandas, Matplotlib, and Seaborn just to name a few. In addition, we made use of sklearn to help us cluster the restaurant via k-means algorithm and visualize the resulting data with folium.

Final decision on the optimal cluster to start a tourist's food journey will solely be dependent on the tourist preference. Some may like to eat fast food, while other may opt for seafood. This project only recommends the cluster with the most restaurant not taking into consideration several factors like quality, price, parking, ease of access and if it is kosher or halal.