Final Report for IBM Capstone Project - Battle of the Neighbourhoods in Scarborough, Toronto

Introduction

The purpose of this project is to use the skills we have learnt on this course to look at the various neighbourhoods in Scarborough, Toronto to see if there is a way we can look at the data available to make more informed choices about where to move to. Hopefully using these skills in this way will mean that I can use these skills to make more informed choices about where to move to in my own Country.

I chose to use Scarborough in Canada due to the high amount of immigration that takes place in this area. Not only does this mean that a lot of the data collection for what we want to look at has already been done in this area, but it also means that more people could be influenced by the work I will cover in this final project. This project will focus primarily on housing prices and school ratings as influencing factors, however we will also look at local amenities and venues as a secondary influencing factor.

Data Selection

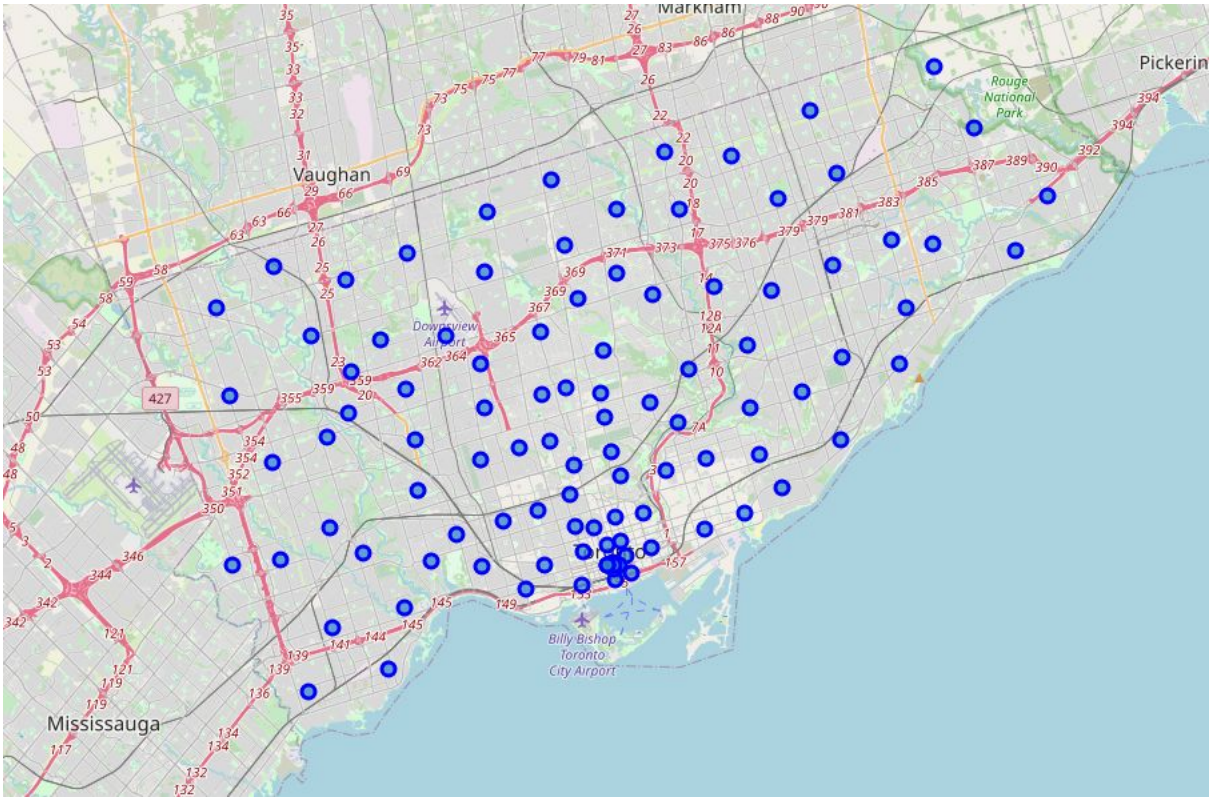
For this project I have gathered Postal Code, Neighbourhood and Borough information from this link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

I will also be using School Rating Data provided by https://www.greatschools.org/

Foursquare API Data

We will be using data provided by the application/website Foursquare. Foursquare is a location data provider that helps to power popular tools such as Google Maps. The location data provided by this will include lots of useful information for us such as venue names, locations, category and if we wanted - menus and photos. As this provides such as wealth of information we will be using the Foursquare API data as our sole source of data for the location based data outside that already mentioned above. For each neighbourhood we have chosen a radius of 100 metres to look around in the Foursquare API to provide venue information so as to not have repeated information too much or an overwhelming amount of information. We want to get a feel for what it's like when stepping out of the door if you were to live here, which I feel a 100 metre radius will provide. The information we will be getting for each venue will be Neighbourhood, Neighbourhood Latitude and Longitude, Venue, Venue Name, Venue Latitude and Longitude as well as Category.

Fig.1 - Map of Scarborough with all of our Postal codes plotted onto it



Methodology

Clustering Approach

In order to effectively compare cities or areas of cities we broke our dataset up into neighbourhoods, segmented them and grouped them into clusters so that we can find similar/desirable neighbourhoods. Clustering data in this manner is a form of unsupervised machine learning known as k-means clustering algorithm.

Table 1. Unclustered results that show most common venues sorted by Neighbourhood

```
[76]: import numpy as np
num_top_venues = 10

indicators = ['st', 'nd', 'rd']

columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = Scarborough_grouped['Neighborhood']

for ind in np.arange(Scarborough_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(Scarborough_grouped.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

[76]:

	Neighborhood	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	Agincourt	Shopping Mall	Pharmacy	Print Shop	Bubble Tea Shop	Skating Rink	Breakfast Spot	Shanghai Restaurant	Sandwich Place	Supermarket	Latin American Restaurant
1	Alderwood, Long Branch	Sandwich Place	Pub	Performing Arts Venue	Gym	Coffee Shop	Pharmacy	Gas Station	Print Shop	Pizza Place	Convenience Store
2	Bathurst Manor, Wilson Heights, Downsview North	Coffee Shop	Sandwich Place	Arcade	Mediterranean Restaurant	Mobile Phone Shop	Fried Chicken Joint	Sushi Restaurant	Park	Restaurant	Middle Eastern Restaurant
3	Bayview Village	Flower Shop	Park	Asian Restaurant	Trail	Gas Station	Yoga Studio	Donut Shop	Eastern European Restaurant	Electronics Store	Elementary School
4	Bedford Park, Lawrence Manor East	Pizza Place	Sandwich Place	Italian Restaurant	Coffee Shop	Butcher	Liquor Store	Juice Bar	Restaurant	Thai Restaurant	Sports Club

Table 2. 5 Clustered results that appear at the top of our table showing most common venues sorted by Postal Code

Now we will have a look at K-Means Clustering Approach

[78]:
```
Scarborough_grouped_clustering = Scarborough_grouped.drop('Neighborhood', 1)
kmeans = KMeans(n_clusters=3, random_state=0).fit(Scarborough_grouped_clustering)
kmeans.labels_
```

[78]:
```
array([1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 2, 1, 1, 1, 0, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
       1, 1, 1, 1, 1, 1, 0, 1, 1, 0], dtype=int32)
```

[79]:
```
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

Scarborough_merged =df_2.iloc[:16,:]

# merge toronto_grouped with toronto_data to add latitude/longitude for each neighborhood
Scarborough_merged = Scarborough_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

Scarborough_merged.head()
```

[79]:

	Postalcode	Borough	Neighborhood	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	M1B	Scarborough	Malvern, Rouge	43.81139	-79.19662	1	Zoo Exhibit	Paintball Field	Fast Food Restaurant	Event Space	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Elementary School	Escape Room	Ethiopian Restaurant
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.78574	-79.15875	1	Bar	Construction & Landscaping	Fish & Chips Shop	Falafel Restaurant	Eastern European Restaurant	Electronics Store	Elementary School	Escape Room	Ethiopian Restaurant	Event Space
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.76575	-79.17470	0	Park	Gym / Fitness Center	Athletics & Sports	Gymnastics Gym	Yoga Studio	Doner Restaurant	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Elementary School
3	M1G	Scarborough	Woburn	43.76812	-79.21761	0	Coffee Shop	Chinese Restaurant	Park	Fast Food Restaurant	Event Space	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Elementary School	Escape Room
4	M1H	Scarborough	Cedarbrae	43.76944	-79.23892	1	Thai Restaurant	Bakery	Caribbean Restaurant	Gas Station	Athletics & Sports	Hakka Restaurant	Bank	Playground	Fish & Chips Shop	Dumpling Restaurant

Work Flow

Using the Foursquare API we are able to collect data of nearby places for the neighbourhoods. Due to limitations for the basic developer account limiting the number of http requests we can make we have limited how far it will look around each neighbourhood to 100m.

Results

Fig.2 Map of Scarborough showing clustered neighbourhoods

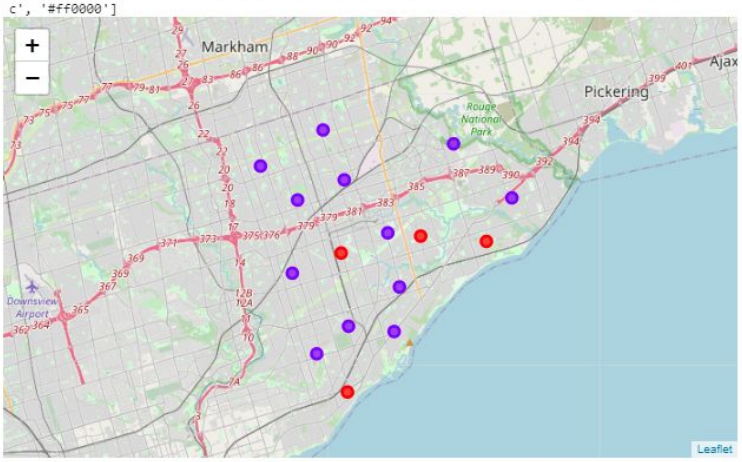
```
x = np.arange(kclusters)
colors_array = cm.rainbow(np.linspace(0, 1, kclusters))
rainbow = [colors.rgb2hex(i) for i in colors_array]
print(rainbow)
# add markers to the map

markers_colors = []
for lat, lon, nei , cluster in zip(Scarborough_merged['Latitude'],
                                   Scarborough_merged['Longitude'],
                                   Scarborough_merged['Neighborhood'],
                                   Scarborough_merged['Cluster Labels']):
    label = folium.Popup(str(nei) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color=rainbow[cluster-1],
        fill=True,
        fill_color=rainbow[cluster-1],
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```

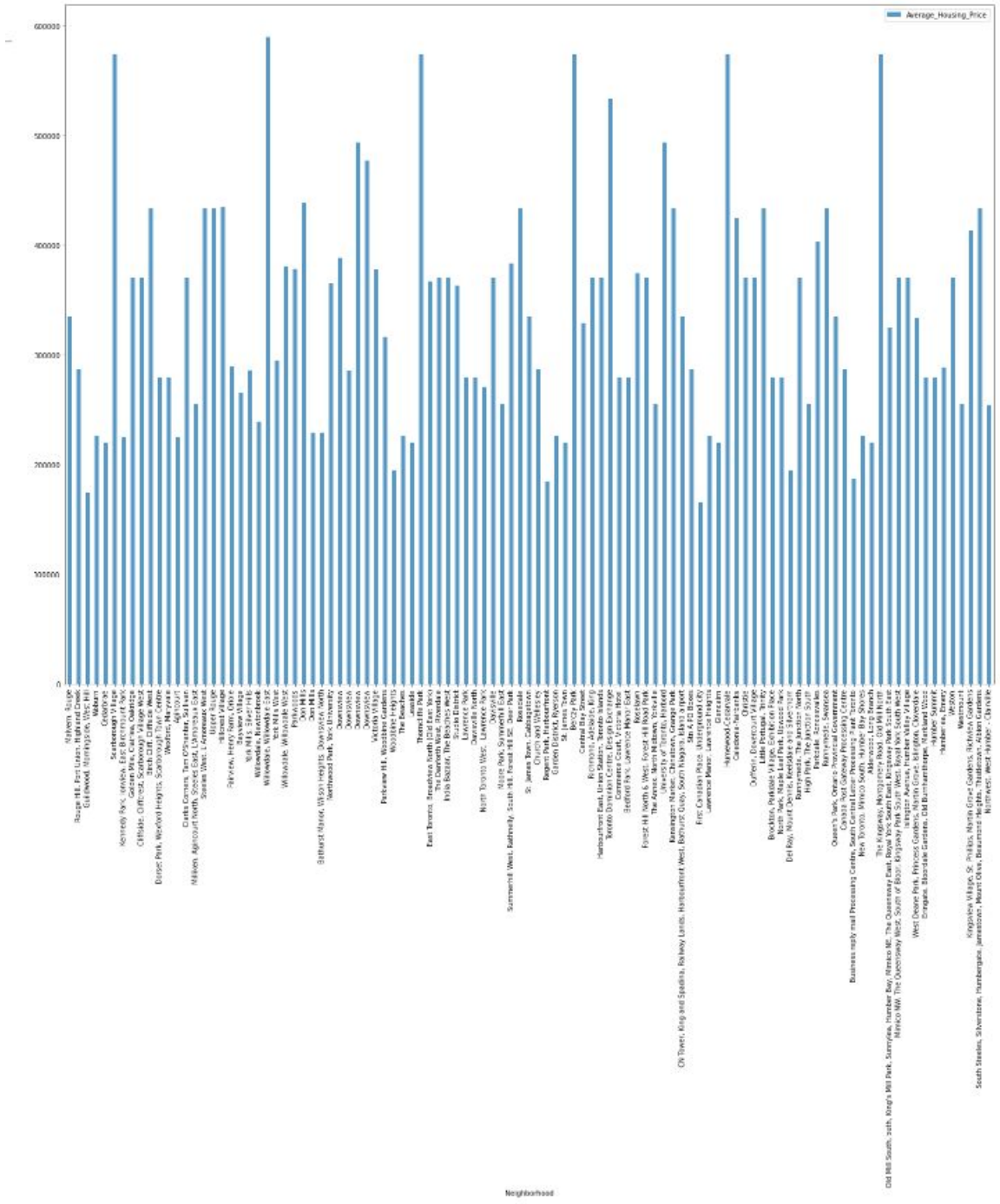
```
['#8000ff', '#4856fb', '#10a2f0', '#2adddd', '#62fbc4', '#9cfba4', '#d4dd80', '#ffa256', '#ff562
c', '#ff0000']
```

]:



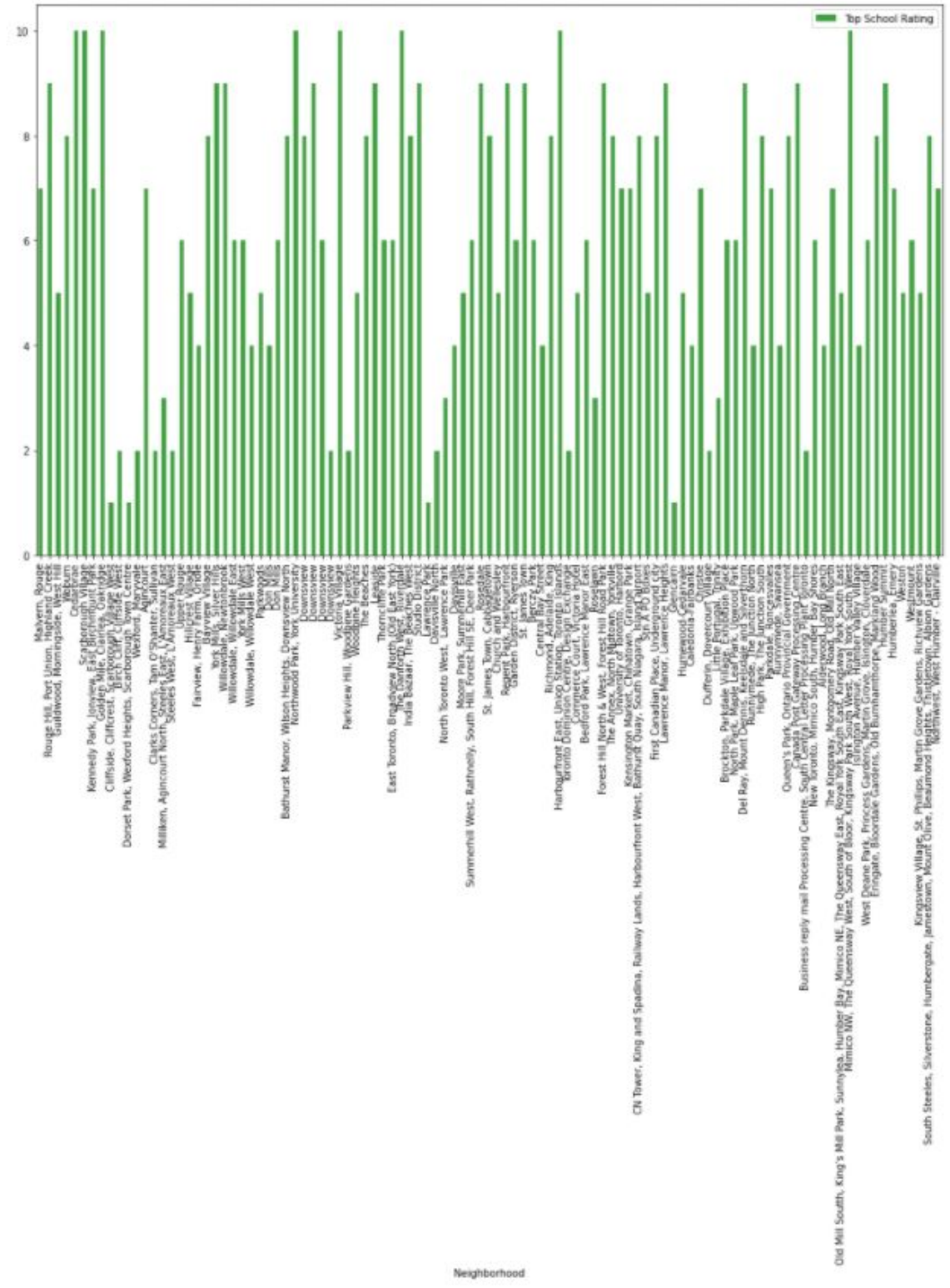
On Figure 2 we can see a total of 16 clusters have been created while using k-means clustering, the colour assigned to each cluster is based on the numerical label given to it during the k-means clustering algorithm.

Fig 3. Housing Price by Neighbourhoods in each Cluster



In figure 3 we can see the average housing prices for each of the neighbourhoods within the clusters. We can see that within the Scarborough area there is a lot of fluctuation on average housing price based on neighbourhood.

Fig 4. School Ratings by Neighbourhood within each Cluster



In Figure 4 we can see the school ratings for the nearest school within each neighbourhood. This will mean that some of the schools for each neighbourhood may be shared by another neighbourhood.

Discussion

The purpose for this project was to gather information in a digestible manner so that people can make more informed decisions around what neighbourhood may be most suitable for them. With this in mind we have several useful figures and tables to consider, namely Table 2, Figure 3 and Figure 4. The combination of these will show the average housing price and school rating for each neighbourhood, as well as the local amenities within the cluster the neighbourhood is found within. The idea of desirability for each neighbourhood would be largely based on our own personal circumstances, and so interpretation of these graphs could be taken in a variety of ways.

For my own personal circumstances, if I were moving to Scarborough I would want to consider school ratings, with a low housing price as I am still young. As I do not currently have children I would also strongly consider local amenities. One such neighbourhood that jumps out to me would be Cedarbrae as it has an average housing price of just over $200,000, a school rating of 10/10 and most importantly has its most common venue as “Thai Restaurant” and second most common as “Bakery”. As a lover of food I would strongly consider Cederbrae given it these factors, as well as having a nearby athletics and sports venue to work off the food I would have eaten at the common bakeries and Thai restaurants.

Conclusion

For this project we have used k-means clustering to group together neighbourhoods to more easily look at local amenities when comparing neighbourhoods. For each neighbourhood we have considered the average housing price as well as school rating. Using the results provided above we have created tools that would allow anyone to more easily compare these neighbourhoods to make more informed decisions about suitable areas to move to within the Scarborough region. I feel that the tools that this practical has had me use will also be applicable in my own personal circumstance, allowing me to use these skills to make more informed choices about any future moves I have.

Libraries used within this project

- **Pandas** for creating and manipulating our dataframes
- **Folium** this will act as our visualisation library to visualse the neighbourhood cluster distribution
- **SciKit Learn** for importing and using k-means clustering
- **JSON** to handle our JSON files
- **XML** to separate data from presentation and XML to store data in a plain text format
- **Geocoder** To retrieve location based data
- **Beautiful Soup and Requests** to scrape and create libraries to handle our http requests
- **MatPlotLib** as our Python based plotting module