IBM Data Science Coursera Capstone Project
Battle of the Neighborhoods - Toronto Cuisine

Introduction

This jupyter notebook will be used for the Applied Data Science Coursera Capstone Project.

A client is looking to open a restaurant in Toronto and is interested in which location they should open it. Because of the ethnic diversity, Torontos's restaurant industry is thriving. Many other factors contributing to the restaurant scene including the ethnic diversity and large population. The large population of Toronto provides restaurants with a large customer base and allows niche restaurants to receive support from the population. The restaurant scene is quite competitive and restaurants of lower quality often go out of the business leaving higher quality.

Toronto has a myriad of neighborhoods and suburbs where our client can open up their restaurant. Our client is a national company looking to expand their restaurant reach into Toronto with of opening up multiple restaurants in Toronto. They have come to us to provide a recommendation on which neighborhood they should open up their first Italian restaurant in.

Data

For this project, I will be utilizing the Foursquare location data to provide information on neighborhoods, venues, restaurant ratings and location data. We can utilize the Foursquare data to pin point areas or clusters with restaurants with large engagement, high ratings and reviews. The Foursquare data can also be used to understand density of population of the neighborhoods and the restaurants within them. For example,

In addition to the Foursquare data, I will also be looking at Toronto neighborhood data. These two data sets can be used in conjunction to pinpoint which neighborhood certain restaurants are located in. This can help us determine which neighborhood quality restaurants are located in and can help us pinpoint a zip code for the new restaurant.

I will use all the data to find restaurants with a lot of reviews and high ratings. These restaurants will be mapped on folium and be grouped by zip code and neighborhoods. The neighborhoods with greater density of these clusters can show us potential locations for this new restaurant. If our client isn't looking to build a location in a neighborhood dense with restaurants rife with competition, we can then look at neighborhood clusters that are less dense for other alternatives on where to build the new restaurant location.

Methodology

To determine which neighborhood the client should build their restaurant, we segmented and clustered the neighborhoods based on their most common venues. In this situation, we are making the assumption that a great density of restaurants is indicative of the fact that the particular neighborhood has a large population that allows these restaurants to survive. Also the greater density means competition and competition encourages growth and innovation. These were the assumptions made. Neighborhoods in Toronto were grouped into clusters using the k-means algorithm.

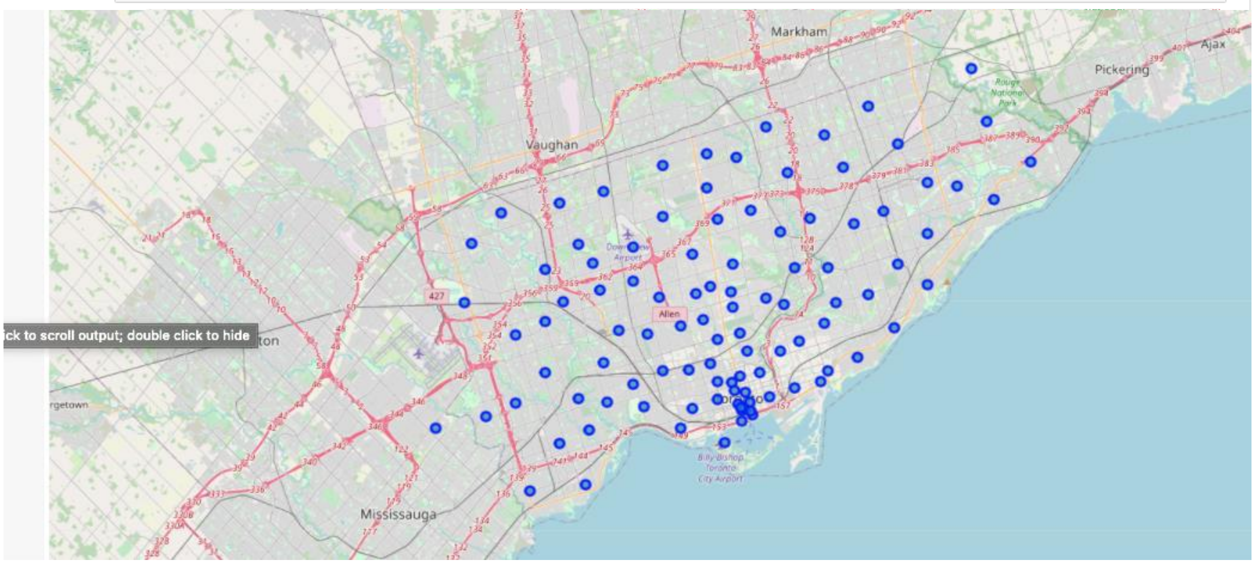
Map of Neighborhoods in Toronto

```
In [329]: import folium

          map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)

          # add markers to map
          for lat, lng, borough, neighborhood in zip(df_toronto['Latitude'], df_toronto['Longitude'], df_toronto['Borough'], df_tor
              label = '{}, {}'.format(neighborhood, borough)
              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_toronto)

          map_toronto
```



Most Common Venues Near Neighborhoods

In [370]:
```python
#function to return most common venues in descending order

def return_most_common_venues(row, num_top_venues):
    row_categories = row.iloc[1:]
    row_categories_sorted = row_categories.sort_values(ascending=False)

    return row_categories_sorted.index.values[0:num_top_venues]
```

In [371]:
```python
num_top_venues = 10

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

# create columns according to number of top venues
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))

# create a new dataframe
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = toronto_grouped['Neighborhood']

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

neighborhoods_venues_sorted.head()
```

Out[371]:

	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	Chinese Restaurant	Shopping Mall	Bakery	Coffee Shop	Sandwich Place	Caribbean Restaurant	Japanese Restaurant	Sri Lankan Restaurant	Bank	Restaurant
1	Alderwood, Long Branch	Discount Store	Pizza Place	Pharmacy	Park	Gas Station	Skating Rink	Sandwich Place	Donut Shop	Garden Center	Bagel Shop
2	Bathurst Manor, Wilson Heights, Downsview North	Pizza Place	Park	Coffee Shop	Bank	Mediterranean Restaurant	Ski Chalet	Fried Chicken Joint	Sushi Restaurant	Supermarket	Men's Store
3	Bayview Village	Intersection	Bank	Japanese Restaurant	Grocery Store	Gas Station	Trail	Skating Rink	Chinese Restaurant	Restaurant	Café
4	click to scroll output; double click to hide ce Manor East	Italian Restaurant	Coffee Shop	Fast Food Restaurant	Bank	Sandwich Place	Juice Bar	Baby Store	Pub	Bagel Shop	Bakery

K-Means Approach

In [372]:
```python
from sklearn.cluster import KMeans

# set number of clusters
kclusters = 10

toronto_grouped_clustering = toronto_grouped.drop('Neighborhood', 1)

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

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

Out[372]: array([0, 5, 5, 5, 4, 4, 5, 4, 4, 5], dtype=int32)

In [373]:
```python
# add clustering labels
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

toronto_merged = df_toronto

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

toronto_merged.head() # check the last columns
```

Out[373]:

	Postal Code	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 Mos Commo Venue
0	M3A	North York	Parkwoods	43.753259	-79.329656	5.0	Park	Pharmacy	Bus Stop	Shopping Mall	Fish & Chips Shop	Shop & Service	Superma
1	M4A	North York	Victoria Village	43.725882	-79.315572	6.0	Coffee Shop	Sporting Goods Shop	Gym / Fitness Center	Pizza Place	Men's Store	French Restaurant	Golf Cou
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	4.0	Coffee Shop	Café	Park	Theater	Restaurant	Pub	Breakfas Spot
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	4.0	Clothing Store	Coffee Shop	Fast Food Restaurant	Restaurant	Sushi Restaurant	Seafood Restaurant	Furniture Home S
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494	4.0	Coffee Shop	Japanese Restaurant	Sushi Restaurant	Park	Thai Restaurant	Café	Italian Restaura

Let's look at the 10 clusters made to determine which clusters our client might like to open a restaurant.

Clustering Approach

Cluster 0

In [358]:
```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 0, toronto_merged.columns[[1] + list(range(5, toronto_merged
.shape[1]))]]
```

Out[358]:

	Borough	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 M Comm Venue
6	Scarborough	0.0	Trail	Fast Food Restaurant	Coffee Shop	Spa	Construction & Landscaping	Martial Arts School	Supermarket	Caribbean Restaurant	Bank	Bakery
26	Scarborough	0.0	Bank	Gas Station	Coffee Shop	Bakery	Indian Restaurant	Yoga Studio	Thai Restaurant	Restaurant	Caribbean Restaurant	Athleti Sports
34	North York	0.0	Coffee Shop	Restaurant	Pizza Place	Furniture / Home Store	Sandwich Place	Chinese Restaurant	Japanese Restaurant	Market	Bar	Massa Studio
44	Scarborough	0.0	Coffee Shop	Intersection	Bakery	Pizza Place	Trail	Fast Food Restaurant	Bus Line	Mexican Restaurant	Metro Station	Beer S
49	North York	0.0	Coffee Shop	Intersection	Gas Station	Chinese Restaurant	Park	Athletics & Sports	Dim Sum Restaurant	Bakery	Convenience Store	Medite Restau
56	York	0.0	Furniture / Home Store	Intersection	Grocery Store	Discount Store	Sandwich Place	Dessert Shop	Shopping Mall	Italian Restaurant	Gas Station	Fast F Restau
65	Scarborough	0.0	Coffee Shop	Furniture / Home Store	Pharmacy	Restaurant	Electronics Store	Asian Restaurant	Intersection	Fast Food Restaurant	Indian Restaurant	Bakery
76	Mississauga	0.0	Coffee	Hotel	Chinese	Middle Eastern	Fried Chicken	Burrito	Bus Station	Bakery	Mexican	Asian

Cluster 1

In [359]:
```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 1, toronto_merged.columns[[1] + list(range(5, toronto_merged
.shape[1]))]]
```

Out[359]:

	Borough	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
46	North York	1.0	Park	Bank	Shopping Mall	Pizza Place	Moving Target	Vietnamese Restaurant	Grocery Store	Yoga Studio	Escape Room	Dry Cleaner
101	Etobicoke	1.0	Park	Bus Stop	Shopping Mall	Eastern European Restaurant	Ice Cream Shop	Italian Restaurant	Event Space	Dumpling Restaurant	Electronics Store	Escape Room

Cluster 2

In [360]:
```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 2, toronto_merged.columns[[1] + list(range(5, toronto_merged
.shape[1]))]]
```

Out[360]:

	Borough	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
45	North York	2.0	Park	Pool	Yoga Studio	Falafel Restaurant	Dry Cleaner	Dumpling Restaurant	Eastern European Restaurant	Electronics Store	Escape Room	Ethiopian Restaurant

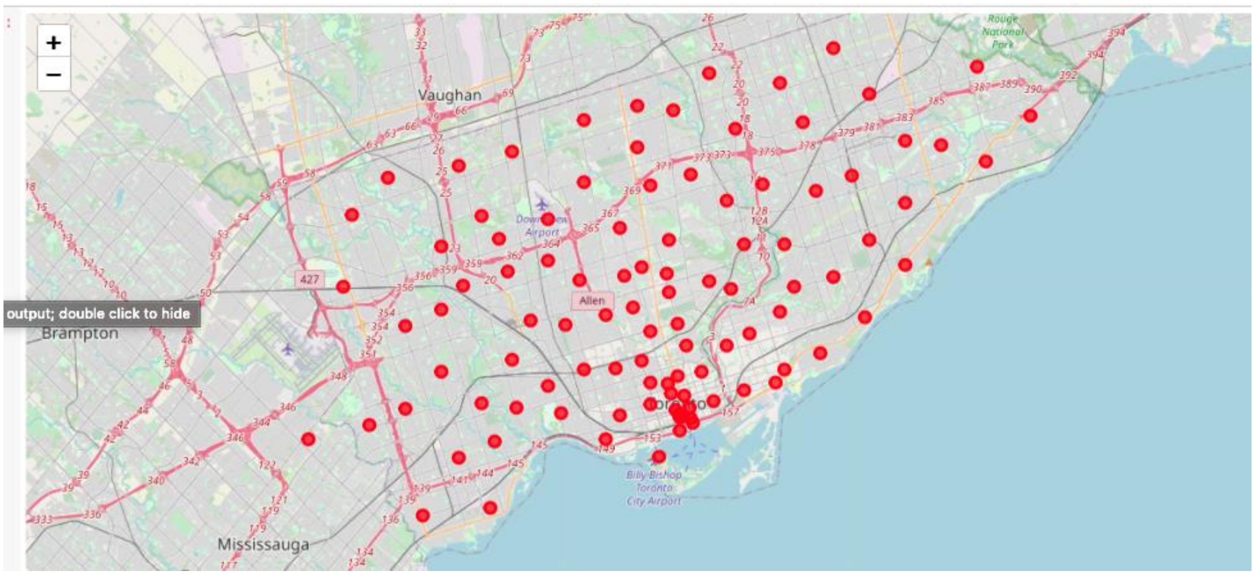
RESULTS

In [381]:
```
# create map
map_clusters = folium.Map(location=[latitude, longitude], zoom_start=11)

# set color scheme for the clusters
x = np.arange(kclusters)
ys = [i + x + (i*x)**2 for i in range(kclusters)]
colors_array = cm.rainbow(np.linspace(0, 1, len(ys)))
rainbow = [colors.rgb2hex(i) for i in colors_array]

# add markers to the map
markers_colors = []
for lat, lon, poi, cluster in zip(toronto_merged['Latitude'], toronto_merged['Longitude'], toronto_merged['Neighborhood'], toronto_merged['Clust
    label = folium.Popup(str(poi) + ' Cluster ' + str(cluster), parse_html=True)
    folium.CircleMarker(
        [lat, lon],
        radius=5,
        popup=label,
        color='red',
        fill=True,
        fill_color='red',
        fill_opacity=0.7).add_to(map_clusters)

map_clusters
```



DISCUSSION

Based on the clusters created, we can see a greater density of restaurants in clusters 0, 4, 5, and 6. We can see that in cluster 0, amongst the 1st most common venues, the most prevalent are restaurants and coffee shops. This is a good indicator that this area is bustling with traffic. When we further look into the 2nd most common venues for cluster 0, we see that intersections are common further enhancing our statement. Looking at cluster 4, we can see that amongst the 1st most common venues, coffee shops are prevalent while amongst the 2nd and 3rd most common venues are restaurants and cafes. The presence of many restaurants and cafes can give us some understanding this area must be competitive but that there is also a large population that is able to service these many restaurants.

While it seems that cluster 5 has a lot of restaurants in the 4th, 5th and 6th most common venues, in the top most common venues we find pharmacies, pizza places, and general stores. If our client were looking to build a pizza restaurant, cluster 5 gives us an indication that the population in the area often services pizza places. In its top most common venues, cluster 6 seems to have a pizza places and fast food restaurants. If fast food was our clients interests, this would be a good location to look into.

CONCLUSION

From the data and results, we can recommend the client to look into building a restaurant in cluster 4 due to popularity and density of that area. If our client were looking to build a pizza restaurant, cluster 5 would be a good location to look into.