CAPSTONE PROJECT

The idea of this project is to provide the best location for a restaurant based on external sources of data. What I will try during this notebook is to show different sources of data to identify the best location.

DATASET USED

1.- Foursquare info from previous week

2.- Neiborhoud boundaries from (https://open.toronto.ca/dataset/neighbourhoods/ (https://open.toronto.ca/dataset/neighbourhoods/))

3.- Business Improvement areas (https://open.toronto.ca/dataset/business-improvement-areas/ (https://open.toronto.ca/dataset/business-improvement-areas/))

During this notebook I will try to link the situation of the main food related placed in the city of toronto with the biggest business development area. This will lead us to find which is the % of restaurantes in each area and the proportion compared to the rest. Based on this if we want to place a restaurant it should be done in the best business area with the lowest restaurant rate

METHODOLOGY

I provide an study where I evaluate the realtionship between the number of elements in each area compared with the number of food related ones. Lowest ratio is the indicator to place the restaurant

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.cm as cm
from scipy.spatial import distance_matrix
import matplotlib.colors as colors
import folium # plotting library
from sklearn.cluster import KMeans
from geopy.geocoders import Nominatim # convert an address into latitude and longit
ude values
from math import cos, asin, sqrt
%matplotlib inline
```

READ Datasets

In [2]:
```
# Foursquare data:
df_square=pd.read_csv('Toronto_data.csv')
df_square.head()
```

Out[2]:

	Unnamed: 0	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	0	Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	1	Malvern	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
2	2	Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
3	3	Highland Creek	43.784535	-79.160497	Affordable Toronto Movers	43.787919	-79.162977	Moving Target
4	4	Rouge Hill	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar

In [3]:
```
df_areas=pd.read_csv('Business Improvement Areas Data.csv')
df_areas.head()
```

Out[3]:

	_id	AREA_ID	DATE_EFFECTIVE	AREA_ATTR_ID	PARENT_AREA_ID	AREA_SHORT_CODE	AREA_LONG_
0	739	2478937	2019-05-28T21:47:59	26004921	NaN	020-01	
1	740	2478936	2019-05-28T21:47:59	26004920	NaN	042-01	
2	741	2478935	2019-05-28T21:47:59	26004919	NaN	093-01	
3	742	2478934	2019-05-28T21:47:59	26004918	NaN	033-00	
4	743	2478933	2019-05-28T21:47:59	26004917	NaN	002-00	

In the code below we will assign the closest business area based on distance to the center of the area. This way we can calculate the total number of places by AREA, which will give us a size of it.

In [4]:
```
def distance(lat1, lon1, lat2, lon2):
    p = 0.017453292519943295
    a = 0.5 - cos((lat2-lat1)*p)/2 + cos(lat1*p)*cos(lat2*p) * (1-cos((lon2-lon1)*
p)) / 2
    return 12742 * asin(sqrt(a))

def closest(data, v):
    return min(data, key=lambda p: distance(v['LATITUDE'],v['LONGITUDE'],p['LATITUD
E'],p['LONGITUDE']))['AREA_NAME']




def find_area():
    tempData = []
    for index, row in df_areas.iterrows():
        tempDict = {}
        tempDict['LATITUDE']=row['LATITUDE']
        tempDict['LONGITUDE']=row['LONGITUDE']
        tempDict['AREA_NAME']=row['AREA_NAME']
        tempData.append(tempDict)
    return_value=[]
    for index, row in df_square.iterrows():
        temp_results = {}
        tempRow = {'LATITUDE': row['Venue Latitude'], 'LONGITUDE': row['Venue Longi
tude']}
        temp_results['AREA']=closest(tempData,tempRow)
        temp_results['Venue Category']=row['Venue Category']
        temp_results['Venue Latitude']=row['Venue Latitude']
        temp_results['Venue Longitude']=row['Venue Longitude']
        return_value.append(temp_results)
    return return_value

df_square['AREA']=""
df_temp_rest=find_area()


df = pd.DataFrame(df_temp_rest, columns =['AREA', 'Venue Category' ,'Venue Latitude
','Venue Longitude' ])

data_grouped=df.groupby("AREA")["AREA"].count()

df_n = pd.DataFrame(data_grouped, columns=['AREA'])

df_n.rename(columns={'AREA':'Total'},inplace=True)

df_population=df_n.sort_values(by=['Total'],ascending=False)
```

First let's find out where are the actual restaurants placed

In [5]:
```
df.loc[df['Venue Category'].str.contains("Restaurant"), 'food_related'] = True
df.loc[df['Venue Category'].str.contains("Gastropub"), 'food_related'] = True
#df_square.loc[df_square['Venue Category'].str.contains("Bar"), 'food_related'] = T
rue


df_area_food=df[df['food_related'] == True]
len(df_area_food)

df_area_grouped=df_area_food.groupby("AREA")["AREA"].count()
df_area_grouped = pd.DataFrame(df_area_grouped, columns =['AREA'])
df_area_grouped.rename(columns={'AREA':'Total'},inplace=True)

df_restaurants=df_area_grouped.sort_values(by=['Total'],ascending=False)
```

Let's find out which is the best business area based on the propotion of restaurants

In [6]:
```
# Number of Items
df_population

# Number of Restaurants
df_restaurants

df_test=df_population.join(df_restaurants, lsuffix='_caller', rsuffix='_other')

df_test['Ratio']=(df_test['Total_other']*100)/df_test['Total_caller']

df_test=df_test.reset_index()

df_test.head()
```

Out[6]:

	AREA	Total_caller	Total_other	Ratio
0	Financial District	976	262.0	26.844262
1	Downtown Yonge	305	65.0	21.311475
2	Toronto Entertainment District	249	38.0	15.261044
3	Kennedy Road	204	59.0	28.921569
4	Kensington Market	199	60.0	30.150754

In [7]:
```
address = 'Toronto'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of Toronto are {}, {}.'.format(latitud
e, longitude))
map_toronto = folium.Map(location=[latitude, longitude], zoom_start=10)
```

```
The geograpical coordinate of Toronto are 43.653963, -79.387207.
```

In [8]:
```
for lat, lng, venue_type in zip(df['Venue Latitude'], df['Venue Longitude'], df['Ve
nue Category']):
    label = '{}'.format(venue_type)
    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)

for lat, lng, area_name in zip(df_areas['LATITUDE'], df_areas['LONGITUDE'], df_area
s['AREA_NAME']):
    label = '{}'.format(area_name)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='red',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_toronto)

map_toronto
```

Out[8]:

In [12]:
```
# Print the investment area with less ratio restaurant / rest

df_ratio=df_test.sort_values(by=['Ratio'],ascending=True)
df_ratio
```

Out[12]:

	AREA	Total_caller	Total_other	Ratio
23	Village of Islington	57	3.0	5.263158
45	Pape Village	17	1.0	5.882353
37	Weston Village	32	2.0	6.250000
40	MarkeTO District	29	2.0	6.896552
30	Queen Street West	42	3.0	7.142857
20	shoptheQueensway.com	65	5.0	7.692308
49	Emery Village	13	1.0	7.692308
6	Wexford Heights	106	9.0	8.490566
41	Uptown Yonge	29	3.0	10.344828
15	Albion Islington Square	76	8.0	10.526316
16	Historic Queen East	74	8.0	10.810811
46	DuKe Heights	17	2.0	11.764706
26	Crossroads of the Danforth	47	6.0	12.765957
2	Toronto Entertainment District	249	38.0	15.261044
22	Liberty Village	57	9.0	15.789474
28	Gerrard India Bazaar	44	7.0	15.909091
31	Riverside District	42	7.0	16.666667
17	Dupont by the Castle	72	12.0	16.666667
35	Dovercourt Village	34	6.0	17.647059
24	Sheppard East Village	53	10.0	18.867925
19	Wilson Village	71	14.0	19.718310
57	Fairbank Village	5	1.0	20.000000
1	Downtown Yonge	305	65.0	21.311475
42	Danforth Village	27	6.0	22.222222
38	Leslieville	30	7.0	23.333333
53	Bloor Annex	8	2.0	25.000000
14	Bloor West Village	78	20.0	25.641026
13	Bayview Leaside	81	21.0	25.925926
9	Cabbagetown	92	24.0	26.086957
5	St. Lawrence Market Neighbourhood	183	48.0	26.229508
...	...	...	...	...
29	Junction Gardens	44	16.0	36.363636
11	Yonge Lawrence Village	90	35.0	38.888889
43	Eglinton Hill	20	8.0	40.000000
25	Greektown on the Danforth	51	22.0	43.137255
8	Chinatown	103	51.0	49.514563
61	Baby Point Gates	4	2.0	50.000000
34	Bloor Street	35	21.0	60.000000
7	CityPlace and Fort York	105	NaN	NaN
44	Long Branch	20	NaN	NaN

RESULTS

In this section we can see that Village of Islington is the best area for a Restaurant

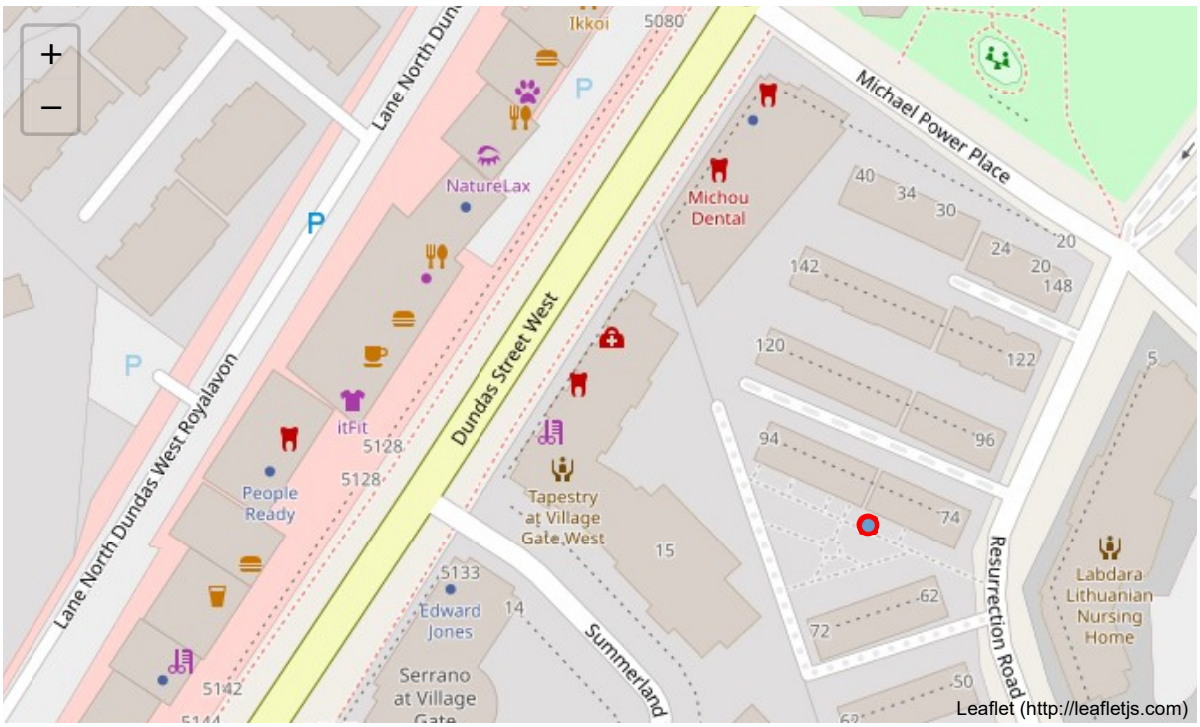
In [22]:
```
df_winner = df_areas[df_areas['AREA_NAME'] == 'Village of Islington' ]
df_winner
```

Out[22]:

	_id	AREA_ID	DATE_EFFECTIVE	AREA_ATTR_ID	PARENT_AREA_ID	AREA_SHORT_CODE	AREA_LONG
59	798	2478878	2019-05-28T21:47:59	26004862	NaN	026-01	

In [23]:
```
map_winner = folium.Map(location=[df_winner.iloc[0]['LATITUDE'], df_winner.iloc
[0]['LONGITUDE']], zoom_start=20)
label = '{}'.format(df_winner.iloc[0]['AREA_NAME'])
label = folium.Popup(label, parse_html=True)
folium.CircleMarker(
    [df_winner.iloc[0]['LATITUDE'], df_winner.iloc[0]['LONGITUDE']],
    radius=5,
    popup=label,
    color='red',
    fill=True,
    fill_color='#3186cc',
    fill_opacity=0.7,
    parse_html=False).add_to(map_winner)
map_winner
```

Out[23]:



CONCLUSION

As a summary of this exercise we can evaluate which are the areas where Toronto has been investing in business development. This situation generates a great ecosystem to generate business development in the area, in our analysis we have established the ratio between different business that are in the area. Based on both ideas, we can summarize that if an area is growing fast and the restaurant ratio is smaller than the rest we can ensure that this area will be a good investment point for a food place.