Capstone Project - The Battle of Neighborhoods

- **By Arun Sridharan**

1. Introduction

1.1. Business Problem

A start-up idea that serves high school students with all the essentials that they require to perform better in their academics and other related fields. Let's name it '**SkoolestHub**'. This will be a retail store consisting of all essentials with buy and rent options.

SkoolestHub decides to setup its retail operations in New York City (NYC). As a first step, SkoolestHub need the repository of high schools operating in NYC. Based on this list, they need to setup the optimal number of stores in NYC that cater to the needs of all the schools in this area.

1.2. Interest

This is a type of greenfield analysis problem that can be used by organisations interested in finding out the optimal number of new locations that can be found to perform it's business operations.

2. Data

2.1. Requirements

Data requirements for solving this business problem are as follows:

2.1.1. List of High Schools in New York City

2.1.2. Latitude and Longitude information of all schools in the list

2.2. Sources

The information on the list of high schools in NYC is obtained from Wikipedia page (https://en.wikipedia.org/wiki/List_of_high_schools_in_New_York_City). To find the Latitude and Longitude information of all schools, we use geopy package through which foursquare API is used to retrieve the information.

3. Methodology

3.1. Data import and processing

Using pandas, the data from excel is imported to python. A new column 'Search Address' is also created by concatenating the columns of 'School', 'Borough' and 'Region'.

1. Import the list of schools in NYC

In [1]:
```
import pandas as pd
```

In [3]:
```
df = pd.read_excel("C:/Users/aruns/Desktop/NYC_School_List.xlsx")
df["Search Address"] = df["School"].astype(str)+', '+df["Borough"]+', '+df["Region"]
df.head()
```

Out[3]:

	Region	Borough	School	P.S. Number	Type	Religious Affiliation	Search Address
0	New York City	The Bronx	Academy for Language and Technology	X365	Public	NaN	Academy for Language and Technology, The Bronx...
1	New York City	The Bronx	Academy for Scholarship and Entrepreneurship: ...	X270	Public	NaN	Academy for Scholarship and Entrepreneurship: ...
2	New York City	The Bronx	Academy of Mount Saint Ursula	NaN	Private, girls	Roman Catholic, Ursuline	Academy of Mount Saint Ursula, The Bronx, New ...
3	New York City	The Bronx	Adlai E. Stevenson Educational Campus	NaN	Public	NaN	Adlai E. Stevenson Educational Campus, The Bro...
4	New York City	The Bronx	Alfred E. Smith Career and Technical Education...	X600	Public	NaN	Alfred E. Smith Career and Technical Education...

3.2. Get latitude and longitude information

Enter the foursquare credentials including Client_ID and Client_Secret. Using geopy, use the 'Search Address' column to get the latitude and longitude values of the schools

2. Get the latitude and the longitude information of the schools

In [4]:
```
CLIENT_ID = 'ETGQH3VGLAASMD020YM5WLTSZTQVWZZIGTQFFMEN4FFWO1BB' # your Foursquare ID
CLIENT_SECRET = 'BSL2E0MOO30JHSNHNFQ5VDBEOMWJK3R1ZSYMGHVMPST0B2MN' # your Foursquare Secret
VERSION = '20180605' # Foursquare API version

print('Your credentials:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentials:
CLIENT_ID: ETGQH3VGLAASMD020YM5WLTSZTQVWZZIGTQFFMEN4FFWO1BB
CLIENT_SECRET:BSL2E0MOO30JHSNHNFQ5VDBEOMWJK3R1ZSYMGHVMPST0B2MN
```

In [5]:
```
from geopy.geocoders import Nominatim # module to convert an address into latitude and longitude values
from geopy.extra.rate_limiter import RateLimiter

geolocator = Nominatim(user_agent="foursquare_agent")
# 1 - conveneint function to delay between geocoding calls
geocode = RateLimiter(geolocator.geocode, min_delay_seconds=1)
# 2- - create location column
df["location"] = df["Search Address"].apply(geocode)
# 3 - create longitude, laatitude and altitude from location column (returns tuple)
df["point"] = df["location"].apply(lambda loc: tuple(loc.point) if loc else None)
# 4 - split point column into latitude, longitude and altitude columns
df[["latitude", "longitude","altitude"]] = pd.DataFrame(df["point"].tolist(), index=df.index)
df.head()
```

3.3. Data cleaning

In this step, the rows where the latitude and longitude are not found is removed from the dataset. The index of the new dataset is reset.

3. Data cleaning

In [50]:
```
df_clean = df.dropna(subset=["latitude","longitude"])
df_latlong = df_clean.reset_index()
df_latlong.head()
```

Out[50]:

	index	Region	Borough	School	P.S. Number	Type	Religious Affiliation	Search Address	location	point	latitude	longitude	altitude
0	0	New York City	The Bronx	Academy for Language and Technology	X365	Public	NaN	Academy for Language and Technology, The Bronx...	(Academy For Language And Technology, Harrison...	(40.84933, -73.91516, 0.0)	40.849330	-73.915160	0.0
1	6	New York City	The Bronx	Aquinas High School	NaN	Private, girls	Roman Catholic	Aquinas High School, The Bronx, New York City	(Aquinas High School, 685, East 182nd Street, ...	(40.8510979, -73.887549, 0.0)	40.851098	-73.887549	0.0
2	9	New York City	The Bronx	Belmont Preparatory High School	X434	Public	NaN	Belmont Preparatory High School, The Bronx, Ne...	(Belmont Preparatory High School, East 189th S...	(40.8593835, -73.8891737, 0.0)	40.859383	-73.889174	0.0
3	12	New York City	The Bronx	Bronx Academy of Letters	X551	Public	NaN	Bronx Academy of Letters, The Bronx, New York ...	(Bronx Academy of Letters, 339, Morris Avenue,...	(40.8136235, -73.9260636, 0.0)	40.813623	-73.926064	0.0
4	13	New York City	The Bronx	Bronx Aerospace High School	X545	Public	NaN	Bronx Aerospace High School, The Bronx, New Yo...	(Bronx Aerospace High School, East Gun Hill Ro...	(40.8752925, -73.8615858, 0.0)	40.875293	-73.861586	0.0

3.4. Visualization of schools on a map

Using 'folium' map feature, the coordinates are plotted on the map. A feature group called 'schools' are created to mark the coordinates.

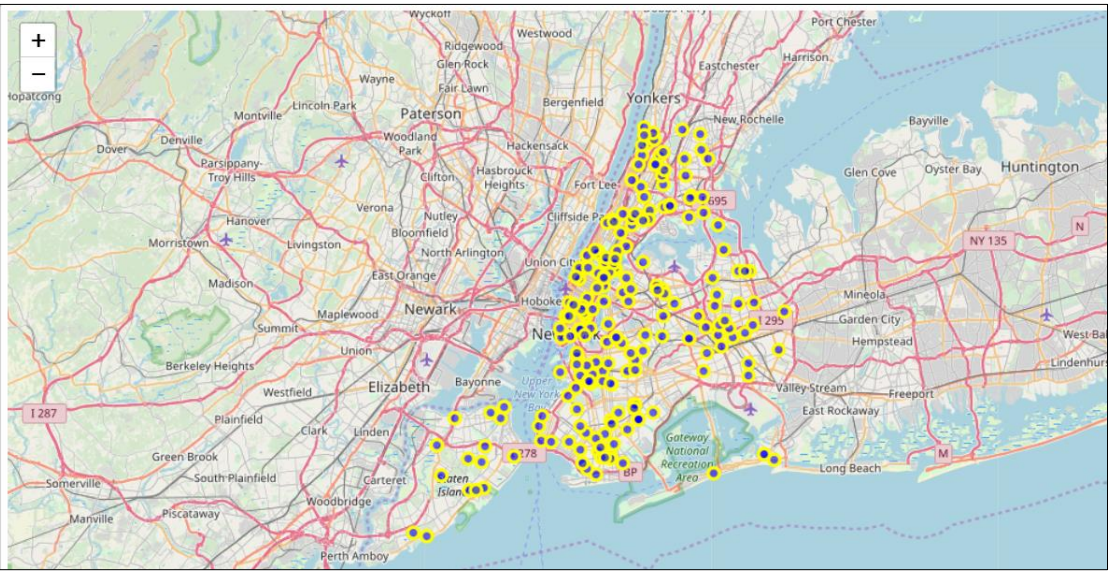
4. Visualization of schools on a map

In [8]:
```
import folium
```

In [133]:
```
# create map and display it
ny_map = folium.Map(location=[40.7053837,-74.0157489], zoom_start=10)

# instantiate a feature group for the schools in the dataframe
schools = folium.map.FeatureGroup()

# loop through the schools and add each to the schools feature group
for lat, lng, in zip(df_latlong.latitude, df_latlong.longitude):
    schools.add_child(
        folium.CircleMarker(
            [lat, lng],
            radius=5, # define how big you want the circle markers to be
            color='yellow',
            fill=True,
            fill_color='blue',
            fill_opacity=0.6
        )
    )

# add incidents to map
ny_map.add_child(schools)
```

The schools are plotted in the map.

Out[133]:



3.5. Segmentation of locations using clustering

Using k-means clustering technique, the latitude and longitude is used as input. The optimal number of clusters are iteratively found.

5. Segmentation of locations using clustering (k-means)

In [85]:
```
import numpy as np
import matplotlib.pyplot as plt
from scipy.cluster.vq import kmeans2, whiten
```

In [117]:
```
coordinates = df_latlong[["latitude", "longitude"]].values
x, y = kmeans2(whiten(coordinates), 8, iter = 20)
df_cluster = pd.DataFrame({"Cluster":y})
```

In [122]:
```
df_join = df_latlong.join(df_cluster)
df_final = df_join.drop(columns=["index", "Search Address", "location", "point", "altitude"])
df_final.head()
```

Out[122]:

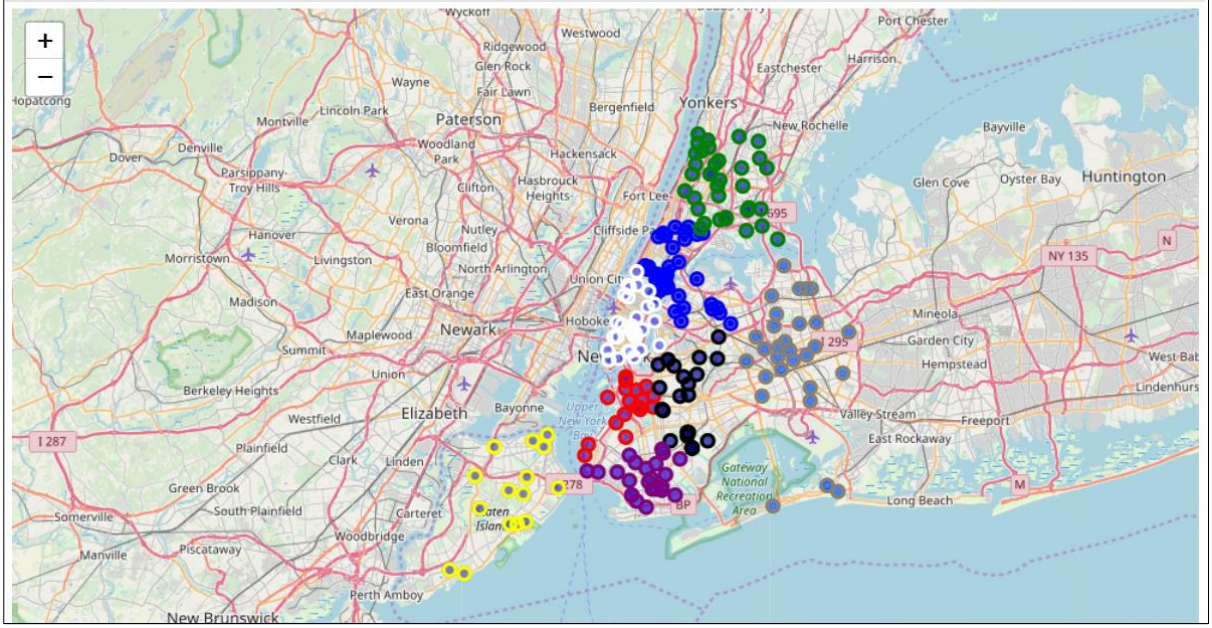
	Region	Borough	School	P.S. Number	Type	Religious Affiliation	latitude	longitude	Cluster
0	New York City	The Bronx	Academy for Language and Technology	X365	Public	NaN	40.849330	-73.915160	1
1	New York City	The Bronx	Aquinas High School	NaN	Private, girls	Roman Catholic	40.851098	-73.887549	1
2	New York City	The Bronx	Belmont Preparatory High School	X434	Public	NaN	40.859383	-73.889174	1
3	New York City	The Bronx	Bronx Academy of Letters	X551	Public	NaN	40.813623	-73.926064	0
4	New York City	The Bronx	Bronx Aerospace High School	X545	Public	NaN	40.875293	-73.861586	1

4. Results

Based on k-means clustering technique, the optimal number of stores to be setup in NYC are is 8. Each 8 stores will cater to the needs of all the schools in the specific cluster. Map displays the 8 clusters.

6. Results - Clustered locations of Schools

```
ny_map
```



5. Discussion

The results found in this exercise is based on the distance between each location of the school. Any new school in the vicinity is opened after the stores is setup, then the new school needs to assign to the closest store. Not all factors like the numbers of students in a school. This will become a capacity planning exercise where the number of students can be included as weights in the model for each school.

6. Conclusion

This project has successfully clustered the high schools in NYC area. Based on the clustering results, the number of stores to be setup in NYC area is determined. The project used machine learning technique called 'k-means clustering' for segmentation. The geocodes of the high schools was determined by foursquare API. The visualization of the high schools on a map is also available.