**Capstone Project – Battle of Neighborhoods**

**Introduction/Business Problem**

A research team wants to analyze New York city properties sale data and visualize the sale activities for each of its 5 boroughs. As part of this project, they also have a requirement to find out the “Borough” that made the most sales.

New York City has diverse culture and point of interests uniquely spread across 5 boroughs and its distinct neighborhoods make the city so special. Therefore, the city’s data set was chosen for this project.

The team would leverage the Kaggle dataset for properties sold in New York city over a 12-month period from September 2016 to September 2017 and Foursquare location data to explore the most common neighborhood venues and details.

The team would identify the borough with most sales based on sale data, then would utilize the Foursquare APIs/location data to identify the recommendations for common/popular venues in the neighborhoods. Using K means, the common venues will be clustered in groups Finally this data will be projected and visualized on a city neighborhood map.

An analytical approach will be used to solve the problem applying advanced machine learning principles along with data transformation & analysis and data visualization techniques.

**Data**

**1. Kaggle data set will be primarily used for this project.**

* <https://www.kaggle.com/new-york-city/nyc-property-sales/data>

This dataset is a record of every building or building unit (apartment, etc.) sold in the New York City property market over a 12-month period, from September 2016 to September 2017

This dataset contains the location, address, type, sale price, and sale date of building units sold. A reference on the trickier fields:

* + **BOROUGH**: A digit code for the borough the property is located in; in order these are Manhattan (1), Bronx (2), Brooklyn (3), Queens (4), and Staten Island (5).
  + **BLOCK**; **LOT**: The combination of borough, block, and lot forms a unique key for property in New York City. Commonly called a BBL
  + **BUILDING CLASS AT PRESENT** and **BUILDING CLASS AT TIME OF SALE**: The type of building at various points in time

**2. NY City Geo spatial data set.**

* Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.
* This dataset exists for free on the web. The link to the dataset: <https://geo.nyu.edu/catalog/nyu_2451_34572>

**3. Top Picks/Common venues in the most transacted borough’s neighborhood of New York city.**

* Foursquare API
  + GET [https://api.foursquare.com/v2/venues/explore/\*](https://api.foursquare.com/v2/venues/explore/*).
  + Response – The following attributes are retrieved from the API response
    - Neighborhood: Name of the neighborhood
    - Neighborhood - latitude & longitude
    - Venue: Venue Name
    - Venue – latitude & longitude
    - Venue Category: category of the venue
* By using the API, we will explore the “top Picks” or recommended venues in the borough

The raw data is first loaded into a **dataframe** using python and transformations applied such as converting the Borough codes to names. The transformed data is then analyzed and grouped to determine the Borough where most sales occurred.

This grouped data is then visualized and plotted in a Bar chart for easy inference. Also, the top borough in terms of sales is identified during the data analysis stage.

NY geospatial data is then loaded into another dataframe in python and used to visualize the neighborhood data in a map visual.

Foursquare API calls are made to get venue details for the neighborhoods and statistical analysis will output the list of venue recommendations for the borough and its neighborhoods.