# Final Project – The Battle of Neighborhoods

## Introduction

This project is the final project in the IBM Data Science Professional Certificate specialization course in Coursera. The goal is to define a problem statement and solve the problem using Machine Learning Methodology. Additionally, FourSquare data should be used to solve the problem.

## Business Problem

A friend of mine is moving to New York and he is interested to know the sales price of houses / properties in and around New York City, so he can decide which area he can move to. He is interested to know the price variation based on the locality (Neighborhood, the different venues around the Neighborhood, etc).

## Methodology

**Analytic approach**: Use predictive analysis to determine if different venues in a location will have an influence of the sales price of the properties in a Neighborhood.

**Data Requirements**: We need the history of sales price of properties sold in New York city, we need data to identify the different venues in and around New York.

**Data Collection**: I will be using couple of data sets to solve the problem:

* Average sales price of properties in New York City, to get the average Sales Price of properties sold in NY city. This data is openly available [here](https://opendata.cityofnewyork.us/).
* Geo location data to get the list of Neighborhoods
* Data from FourSquare that will provide the different venues available in that location.

**Data Understanding & Preparation**: Once I get all the data from the different sources, I will analyze the data first – what information is available. Then data should be cleaned to give effective results.

**Modeling:** Once the data is prepared, I will use regression modelling to create a model. Data will be spilt into Training and Test data. Training data will be used to train the model and Test data will be used to predict from the model.

**Model evaluation**: I used Ridge regression to evaluate the accuracy of the model.

## Results:

After 5-degree Polynomial Regression and Ridge Regression, we have a higher R Squared value that indicates the accuracy of the model. We do see with the Regression plots, while Gross Square Feet has effect on the price, the other parameters, like Playground or Year built didnt have an impact on the price of the properties

## Conclusion

Though the developed model can predict sales price of the properties in and around New York, it also shows the current data on venues doesn’t impact the sales price. We will need more data to make this feasible. With this model, I can give my friend a predicted Sales Price of the properties in different Neighbourhood based on the size of the property.