**NYC Property Sales Data analysis**

Proposal of final project

Chaoying Lyu ([clyu12@syr.edu](mailto:clyu12@syr.edu))

Hong Zhu (hzhu25@syr.edu)

Ruiwei Zhang (rzhang72@syr.edu)

Ziyan Zhou (zzhou56@syr.edu)



# Objective

The NYC Property Sales dataset contains the house sale prices and relative information about the houses, like address and land square feet. Several models will be applied to the dataset to find the influential factors on housing prices. After the data analysis, the main goal of the project is to find ways to calculate property sale price with property features as a reference for buyers and sellers when they are trading property.

## Data Set Description:

The NYC Property Sales dataset is a record of every building or building unit (apartment, etc.) sold in New York City over a 12-month period from September 2016 to September 2017. There are 84548 rows and 22 columns in total, containing the location, address, type, sales price and sales date. The dependent variable in the data is sale price. The link to the data set is as follow: https://www.kaggle.com/new-york-city/nyc-property-sales?select=nyc-rolling-sales.csv. There are two interesting thing that we found about the dataset: a) $0 are common in the sale price column of the dataset which represent actually transferring of deeds between parties (i.e. parents transfer ownership to their home to a child after moving out for retirement); b) This dataset uses the financial definition of a building/building unit, for tax purposes. In case a single entity owns the building in question, a sale covers the value of the entire building. In case a building is owned piecemeal by its residents (a condominium), a sale refers to a single apartment (or group of apartments) owned by some individual.

## Preliminary Data Exploration:

1. **Choose the columns we want to use**
   * Drop the empty and iterator-like columns:'EASE-MENT', '\_c0'
   * Extract month and year columns from SALE DATE
2. **Information of each column**

We have 22 columns of data. 12 columns of data are numeric columns. 9 columns of them are string data.

1. **Deal with NA value**

**Chart, bar chart

Description automatically generated**

We fill 'LAND SQUARE FEET' and 'GROSS SQUARE FEET' with average. In addition, We delete the rows have NA sale price

1. **Summary statistics**
   * The graph below shows the statistics for numeric attributes. For example, the mean of sales is 1.28 million dollars. The third quarter of gross square feet is 95,000 square feet. The median of the total unit is 1. The first quarter oldest property was built in 1920.Graphical user interface

     Description automatically generated
   * The graph below shows the correlation between attributes. For example, we can find the correlation between price and gross square feet is near 0.4.
   * After removing outliers, we can find that the sale price of the property in New York City between 2016 and 2017 shows right-skewed distribution.

**Chart

Description automatically generatedChart, histogram

Description automatically generated**

## Predictions:

1. Spot trends in the property market, in association with affairs in 2016 and 2017 which may have effect on property sales.
2. Predict the locations where the property sales price or price per sq.ft. is highest or lowest.
3. Predict which neighborhood may have relatively new or old property for reference for buyers who have preference on property age.
4. Predict sale price of different building class.
5. Build different models, like Linear Regression, Random Forest, Ridge Regression .etc, to predict sale price and compare the models using appropriate metrics to find the best model.

## Inference:

1. Find the most important factor on the sale price for all house sales.
2. Distinguish the differences in the most influential factor on house prices in different places
3. Verify if there’s a gap between small-size houses and large-size houses in sale price’s factors
4. Find the differences in factors on sale prices between residential and commercial units
5. Check if the building class may affect the sale price factors.

## Non Spark Packages

Pandas, numpy, matplotlib.pyplot, seaborn, nltk, sklearn, scipy