## Background Knowledge

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.

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. See the glossary linked to below.

For further reference on individual fields see <https://www1.nyc.gov/assets/finance/downloads/pdf/07pdf/glossary_rsf071607.pdf>

For the building classification codes see <https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html>

Note that because this is a financial transaction dataset, there are some points that need to be kept in mind:

* Many sales occur with a nonsensically small dollar amount: $0 most commonly. These sales are actually transfers of deeds between parties: for example, parents transferring ownership to their home to a child after moving out for retirement.
* 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.

## To-do:

* **What is the problem you will solve with the EDA? Or What questions you will answer? [Define at least 5-10 questions]**
* Find out all the data entry errors
* Convert them to missing values
* Deal with all missing values by using concepts taught in the class (and their relevant charts)
* Remove any unnecessary data
* Detect outliers
* Conduct EDA (by using scatterplots, bar, correlation heatmaps, histograms, t-tests, anova etc.) and describe at each step what you understand about the data
* Use Google’s Data Studio to find out more insights.

Submit the GDS insights as well as the notebook on Saturday 6th Feb, 2021. If you do not submit, then it will be marked 0.