# Soostone Data Science Assignment

Attached, you will find a small toy dataset we've obtained from a public source, containing real estate sales records in NYC. Using this dataset, we ask that you perform a few tasks and share your work with us. Please don't feel you need to spend a large amount of time on each task; we understand your submission will be a first iteration work product, but it will help us understand your skillset. Overall, we expect this assignment would take 3-6 hours to complete, depending on your preferences and familiarity with the underlying concepts.

1. Perform exploratory analysis on this dataset and produce a showcase/storyline of a few interesting patterns and your observations. You will walk us through your findings during our interview. You may use any tool you like, but a Jupyter notebook using Python is a common choice.
2. Write a SQL query (using only standard SQL that can run on any SQL-supporting database; subqueries and common table expressions are allowed), or optionally a chain of SQL queries that depend on each other, that computes the following items at the granularity of a single sale. There must be no hard-wired numbers - everything must be computed solely through the SQL query.
    a. A column to be called "sale_price_zscore" that represents, for each sale/row, the Z-Score of "SALE_PRICE" of that row as normalized against the entirety of the dataset

    b. A column to be called "sale_price_zscore_neighborhood" that represents, for each sale/row, the Z-Score of "SALE_PRICE" but as normalized based on the NEIGHBORHOOD and BUILDING_CLASS segment to which that row belongs

    c. Columns that compute "square_ft_per_unit" and "price_per_unit"
3. Build a simple model (e.g. regression or any type of ML model) that can predict SALE_PRICE of a given case given all the other columns available in the dataset. Feel free to use any techniques and intermediate steps you deem useful in this effort. Analyze your model, its performance/characteristics/diagnostics, any interesting findings (for example, what factors affect prices the most) and be prepared to discuss the approach in more detail and some ways in which your initial attempt can be further enhanced.