**Description**

The dataset can be found at <https://www.kaggle.com/new-york-city/nyc-property-sales> and includes all building or building unit sold in a 12-month period between 2016 and 2017. It has been adapted from data originating from New York City’s Department of Finance. Note that the data includes transfers of ownership, so the Sale Price column doesn’t necessarily correspond to the value of the building or unit.

**Motivation**

This dataset seems particularly relevant to working at JPMorgan Chase, since JPMorgan Chase often finances real estate. Many different categories are included along with the all-important sale price. I imagine there are lots of different trends that could be analyzed in this dataset and it’s large enough that traditional methods would be painfully slow. This gives Hadoop a chance to shine.

**Size and Attributes**

There are 84,548 rows. Each row has 22 columns including the following: Borough, Neighborhood, Building Class Category, Tax Class, Block, Lot, Ease-ment, Building Class, Address, Apartment Number, Zip Code, Residential Units, Commercial Units, Total Units, Land Square Feet, Gross Square Feet, Year Built, Tax Class, Building Class at Time of Sale, Sale Price, Sale Date.

Borough is Coded; 1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island