Capstone Project 2 Milestone Report:

Goal:

In 2020, the total market value of all properties in New York City was assessed at around 1.4 trillion dollars. That's the total GDP of the United States in 1973, or the total GDP of the United Kingdom in 1996, or China in 2002, or Australia today. Anyone who operates in or adjacent to the property market, whether they are homeowners, real estate developers, or stock market traders, has a stake in knowing how much a property is worth.

The issue is that properties, especially residential properties, are difficult to price because the market caters to a bewildering variety of needs. A three-bedroom, five-bathroom house would be perfect for a large family but a detriment to a bachelor. For this reason, a more consistent metric of property price is its assessed value, an appraisal done for the purpose of calculating property taxes. The Department of Finance values properties once per year, with the option to challenge the assessment if it is believed to be incorrect. This metric provides a more consistent means of tracking property value than the market price.

This project focuses on predicting New York City property assessed value from factors such as property size, presence of easements, year constructed, and borough of the city. The target audience includes anyone interested in property valuation, both homeowners and companies.

Data:

Initial NYC property valuation data was taken from the NYC OpenData website, here. The provided spreadsheet tracks 5.74 million Class 1 property valuations from 2010 to 2017, with 117 columns of associated data such as house number, year of any alterations made, tentative and final valuations of land and property, and more. Tax class 1 includes 1-3 unit residential properties. Class 2 covers larger residentials, such as apartments, 3 denotes utility company property, and 4 is everything else. This project was restricted to Class 1 to remain at a manageable size and because valuation metrics may vary between property classes.

Where necessary, the data were verified from other property – focused websites such as Zillow. Additional information on the mechanisms of property valuation and the New York City property market were taken from sites such as the New York City Department of Finance website, here.

It was immediately clear that the YRB column, indicating the year a property was built, was only an estimate.

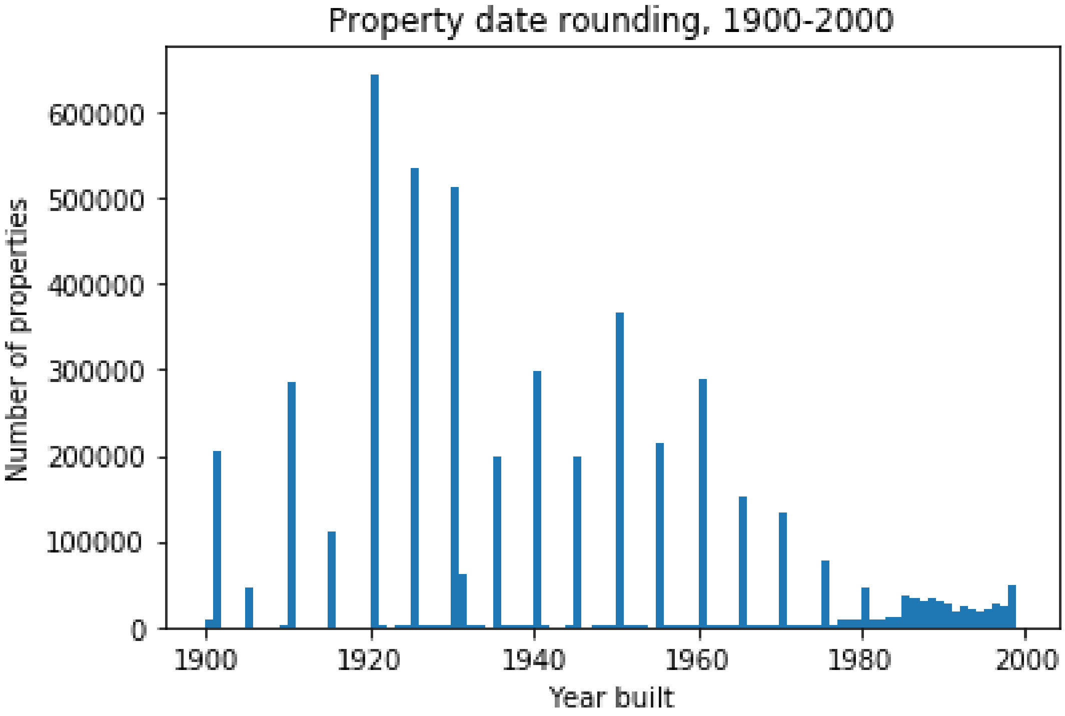

Figure One: Histogram of reported property build dates in the nineteenth century

Properties have been resoundingly reported as being built in years divisible by five up until the late 1980s, when the pattern stops. Until recently, the year-built column was only an estimate, not a precise number. Not only does this introduce error into the data, it also obfuscates other trends in valuation over time. For this reason, for visualization purposes only, we wrote a program to subtract a random number, 0-4, from the provided year, introducing more variation and spreading out the data.

Another issue is that around 5% of the properties report 0 as their year of construction. As New York City was first settled in 1624, we may infer that these properties do not have a known data of construction. To correct this, we replace each zero value with the mean of the borough. It is possible that the properties with missing values are on average older than other properties – note that properties built after 1985 have not had their dates rounded, indicating better record-keeping in the modern era. For this reason, we will also examine models where the zero values have been replaced with years significantly below the mean.

Although it is not viable to manually examine every column of the data, one cluster of outliers was immediately, obviously apparent:

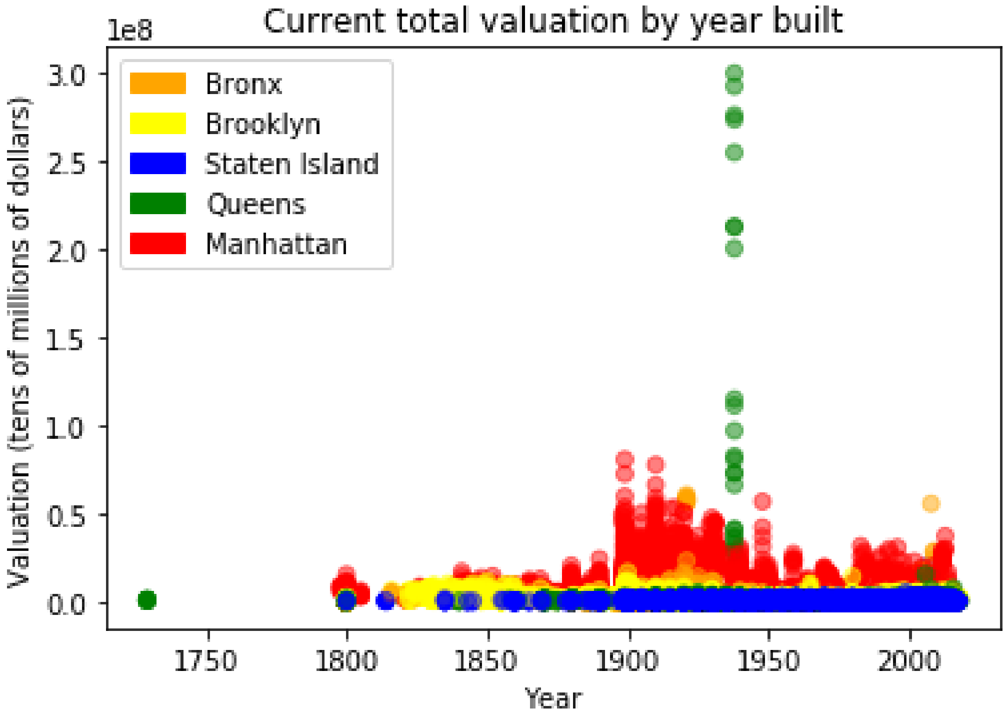

Figure Two: Scatter plot of property valuation by year and borough. Zero-year properties have been omitted from this graph.

The 1930 Queens properties, the large green spike in the middle of the graph, are anomalous in two respects: First, every other property in Queens is valued at a fraction of these nineteen properties. Second, every one of them was ostensibly built in 1930. A quick examination of property sites such as Zillow falsifies this: 604 Shore Road is listed as being built in 1919. 22 Cherry Ave: 1860. 44 Fleet street? 1969. Although several of these properties are listed as being built in 1930, several have been shifted by decades. I cannot explain why properties both older and newer than 1930 would be edited in this manner, but I have manually corrected this especially egregious error.

Finally, several categorical variables such as type of easement (A, B, E-M, etc.) needed to be broken out into separate columns in order to prevent modeling from creating inadvertent correlation between them. TODO – better phrasing than ‘inadvertent correlation?’

Statistical Data Analysis:

One especially interesting valuation pattern that emerged when examining the data was the number of stories the building had.

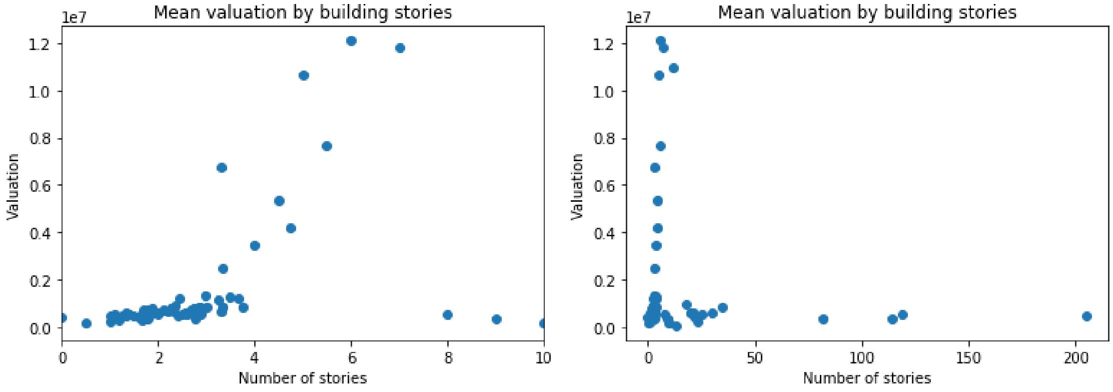

Figure Three: Mean valuation by number of stories, 0-10 and 0-200.

First of all, there are a surprising number of properties with non-integer stories: over a million total. Half-stories like 2.5 and 1.5 are the most common, then 2.75, 1.67, 1.75, etc. It's unclear who would mark down a house as having 2.85 stories, or 1.60, 1.99, or various other improbable decimals, but we will assume that these are unusually specific and punctilious housing records. Second, property valuation increases exponentially until around 6 stories, then abruptly falls back to a consistent level. It's our hope that a simple random forest model will be able to accurately model this.

In about 30,000 properties, the valuation was protested. Some of these, most notably protest code 6, 1, and 6E were valued significantly higher than average. Others, like codes 9 and 5, were valued below the mean on average.

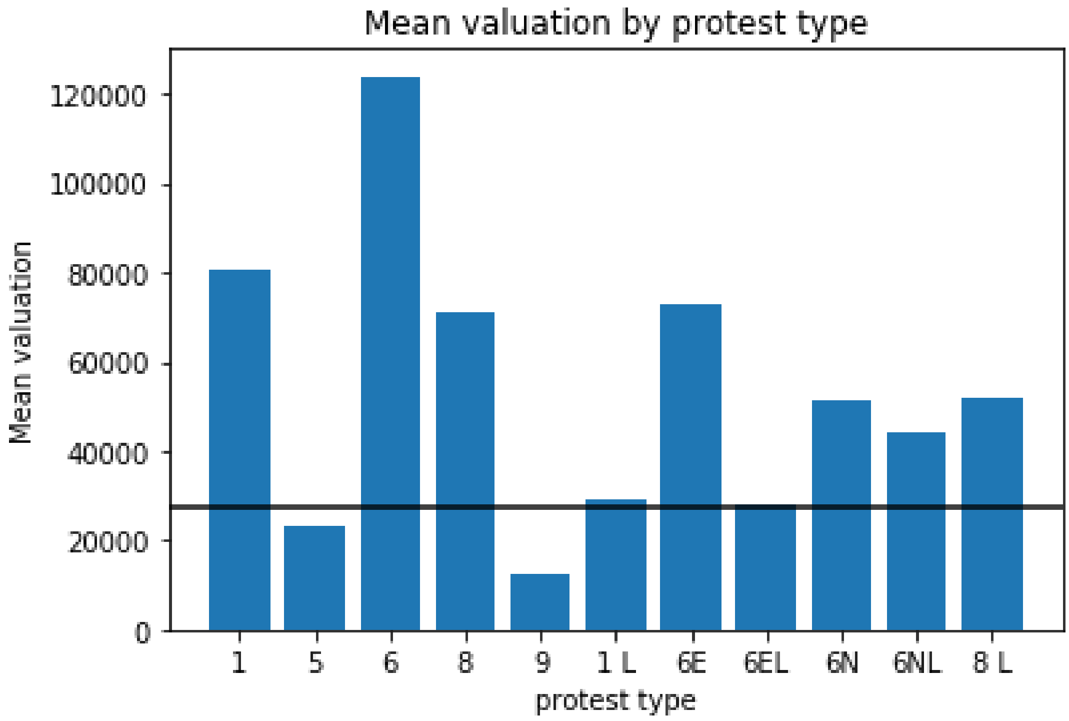

Figure Four: Mean valuation by protest type with bar for overall mean. TODO: Clean this up

Property area saw no such boundary, with prices consistently positively correlated to land area. The last factor that we expected to be strongly correlated with price was the presence of easements, but there were almost no Class 1 properties with easements (< 2000), so this was mostly irrelevant.

New York City caps the amount that a property can increase in price in a year to 6%, and in five years to 20%, so tracking the price of a property in 2010, where known, was be a simple way to add this factor to the model. Where not known, the mean value for the borough was used. This could be improved upon by tracking the property value in the most recent year known, but for now we will use a simpler approach.

Next steps:

In another example of punctiliousness, the dataset includes columns for current/transitional/final, transitional/actual, assessed/exempt, and land/total value, for a total of 24 permutations of value assessment. For simplicity, we will attempt to model Final Actual Assessed Total Value. For our next steps, we will begin working with random forest modeling to predict ‘Final Actual Assessed Total Value’ from all other factors. It was unclear how long it took to publish a transitional assessment versus a final assessment, so we will test models including and withholding this factor. The transitional value is likely close to the final value, but if it is only generated two days in advance, it will be of minimal real-life application.