Capstone Project 2

Milestone Report 2

New York City Property Valuation

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 Collection:

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.

Data Cleaning:

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

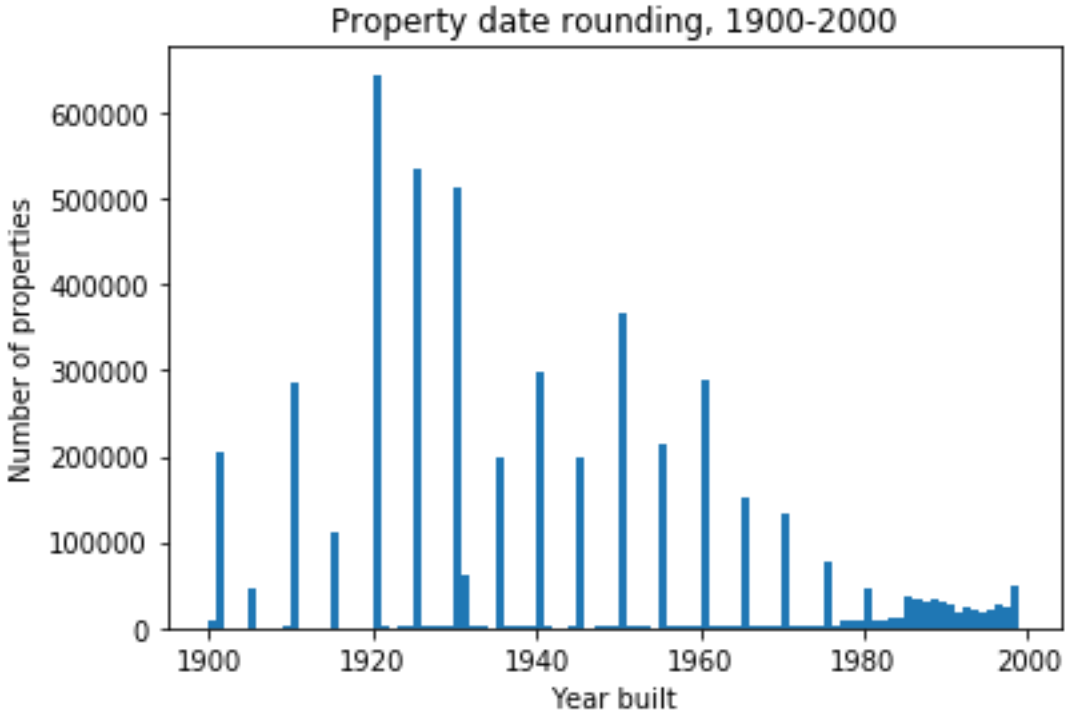

Figure One: Histogram of reported property build dates in the nineteenth century

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. There was no point in making this modification to the modeling data, as this does not improve accuracy.

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

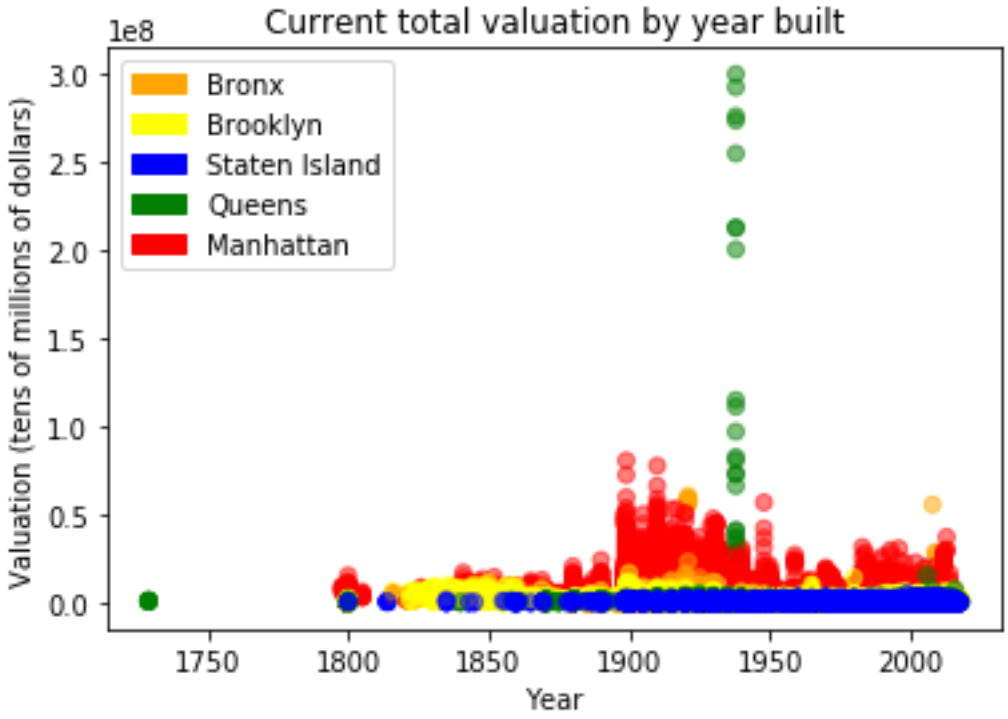

Figure Two: Scatter plot of property valuation by year and borough. Zero-year properties have been omitted from this graph.

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. Checking Zillow's listings for these properties indicates that several of them were not, in fact, built in 1930. This anomaly was difficult to explain until we encountered another error in the data.

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.

Modeling this data made the Queens spike clearer:

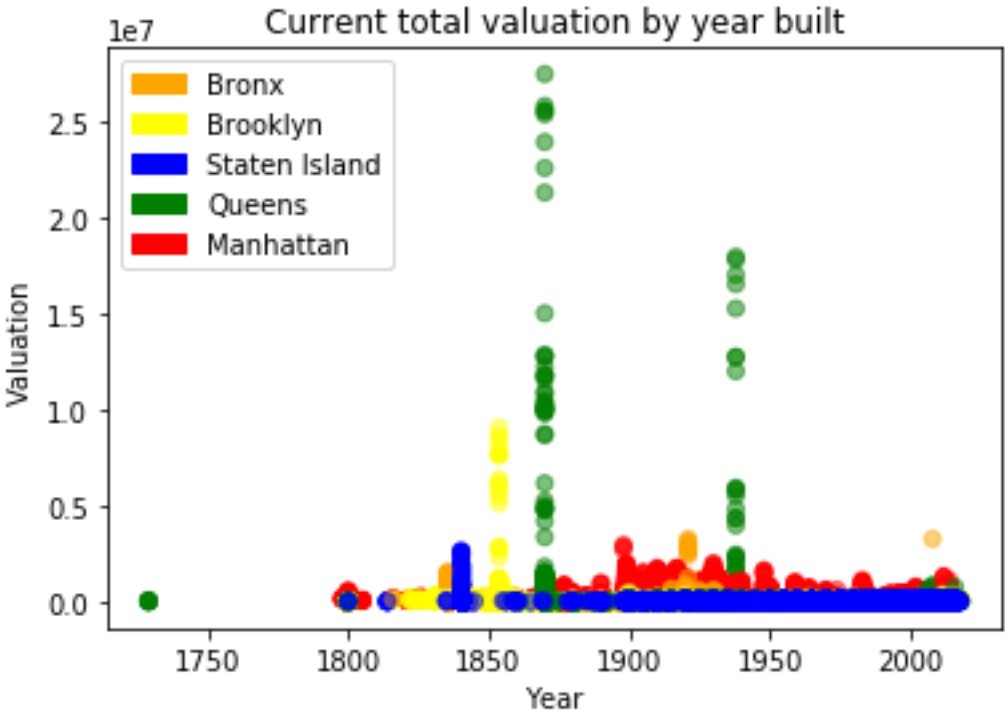

Figure Two: Scatter plot of property valuation by year and borough. Zero-year properties have been listed as being built in the mean year of their borough.

We can see that each borough has a number of valuable properties that were listed as being built in 0. It is likely that the 1930 Queens spike was due to someone entering 1930 as the year for a number of properties with unknown years of construction. It is unclear why only a few properties from one borough would be altered in this manner, but this does seem to be the most likely explanation.

Finally, several categorical variables such as type of easement listed as class (A, B, E-M, etc.) needed to be broken out into separate columns in order to prevent modeling from creating inadvertent correlation between them.

Statistical Data Analysis:

The single most important factor in the dataset is the borough.

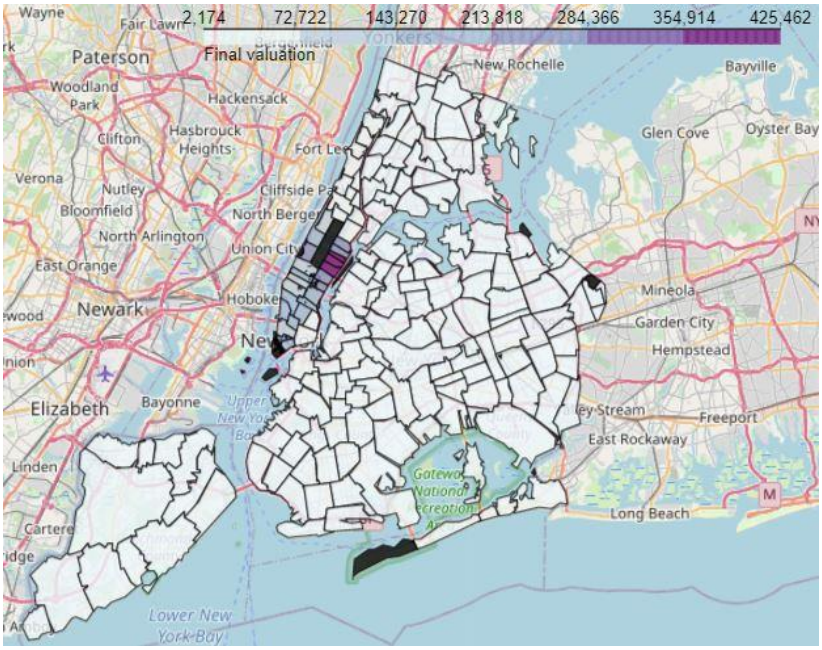

Figure Three: Choropleth map of New York City mean valuations by zip code.

In this unweighted map of New York property valuations, Manhattan's zip codes are valued so far above the rest of the city that they aren't colored in at all.

With some wrangling, it is possible to produce interesting maps from this data.

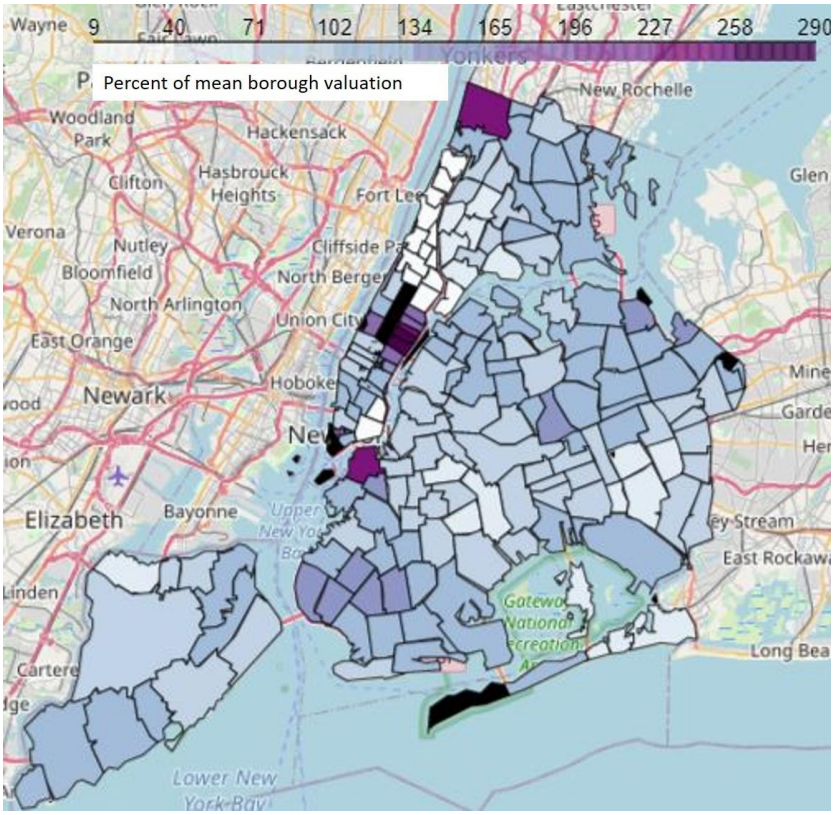

Figure Four: Choropleth map of New York City zip codes, percent of mean borough valuation.

This choropleth map shows how far each zip code is from the mean valuation of its borough. We will experiment with other valuations as we continue working with this mapping module, Folium. Already, though, you can see that the Upper East Side is the wealthiest part of New York City, North Staten Island properties are valued less than south Staten Island properties, North Riverdale is highly valued, et cetera.

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

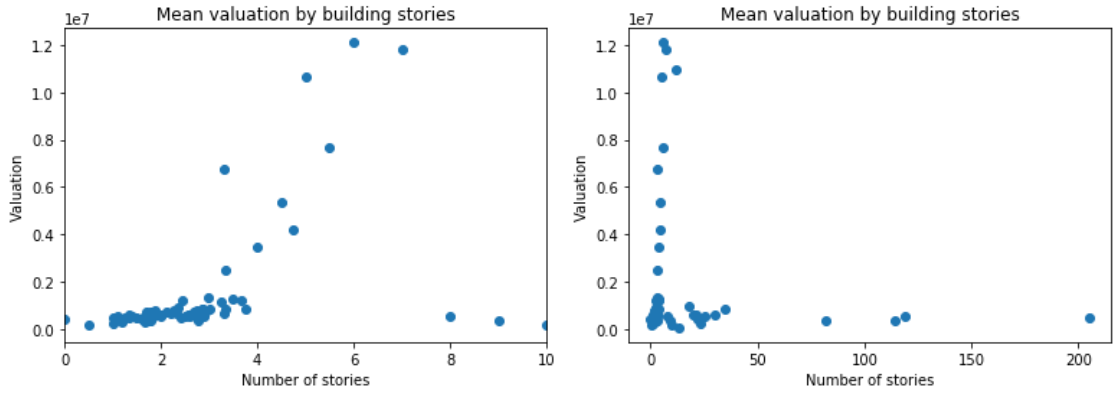

Figure Five: Mean valuation by number of stories, 0-10 and 0-200.

First of all, there are a surprising number of properties with non-integer stories: over a million total, or almost 20% of the dataset. 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. Due to the very small total number of protested valuations, there was no reason to include this data in the random forest model.

Property area was 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, would be a

simple way to add this factor to the model. However, this factor would actually be too good for our model to handle. Adding a variable that is always so closely correlated to the final valuation would eclipse all other variables in the model. It is for this reason that we will proceed assuming that no records on previous valuations are present.

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. We considered including tentative valuations as a variable in the model, but a close reading of New York City property valuation law indicates that tentative valuations become final unless challenged, and as discussed above, very few property owners challenge their valuation. For this reason, the feature was discarded.

Machine Learning

The classification best suited to the dataset was determined to be 'Is this property valued above the mean valuation for its borough?' From borough to borough, this covered between 47-33% of the dataset. Classifying by borough ensured that properties in each borough would be both positive and negative, preventing the borough data from dominating the model.

The dataset was too large to realistically model in its entirety. A simple three-variable random forest classification model took over six hours to train on the entire dataset. Initially, this was solved by splitting the dataset up by borough, but this had several issues. Firstly, a model trained on one borough performed significantly worse on the other boroughs, and secondly, it created issues when trying to draw conclusions about the New York City housing market as a whole. For this reason, the second round of models were trained on data from the entire dataset, randomly sampled to only take 1% of the total data.

These models were run ten times with different samplings to verify that key features of the dataset were preserved by the sampling. The standard deviations of the feature weights, from .008 to .0008, indicated that there was almost no difference in the models produced by this sampling and all key features were preserved.

We strove to avoid any data correlation when training the models and included only one column for each factor. Geographical data was tracked by zip code, a higher-resolution option than tracking by borough that also had geojson files available for mapping. We considered tracking by block, but block data had partial correlation with borough: Manhattan block numbering only runs to around two thousand, while all other boroughs hit at least six thousand, and most eight. For this reason, and for better mapping compatibility, zip code was chosen to track geographical data. Land area, year built, current year, and total units in the property were also factored into the model.

This produced a model with train score 0.856, test score 0.841, indicating almost no model overfit. Precision and recall were calculated to be 0.817 and 0.759. We then calculated the best depth for the model and found the following pattern:

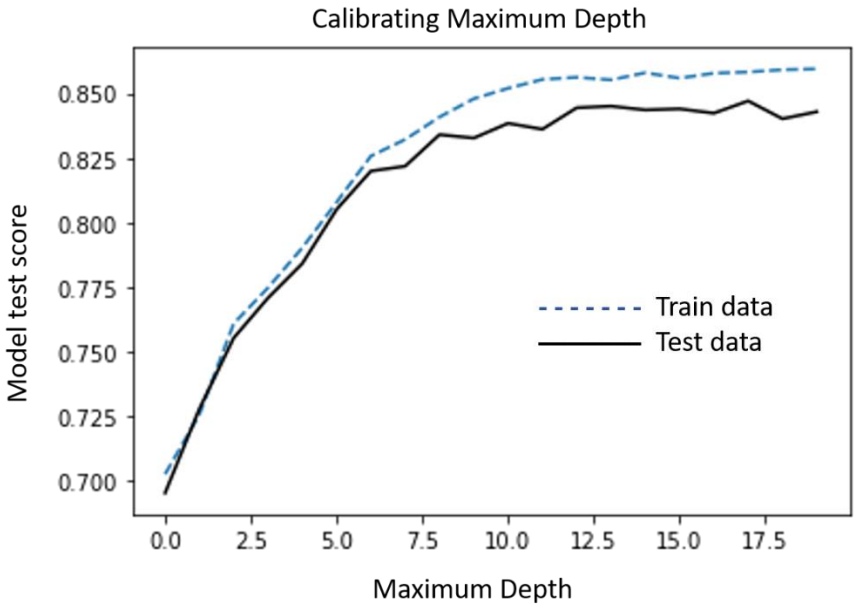

Figure Six: Model performance at varying depths, train and test score.

Maximum depth of around 8 appears to produce the best results with minimal overfit.

Next Steps

Having constructed a robust model that can predict valuation anywhere in New York City, our next steps are to experiment with alternative data inputs discussed above, such as setting the zero-year YRB data to be significantly older than the mean of the borough. We will

also experiment with different methods of mapping the data, as it is difficult to graph the range of valuations between the Manhattan data and the other boroughs.