Brooklyn Real Estate:
A Data Analysis

Jay Morrison

GENERAL ASSEMBLY

April 2015

Project Overview & Motivation

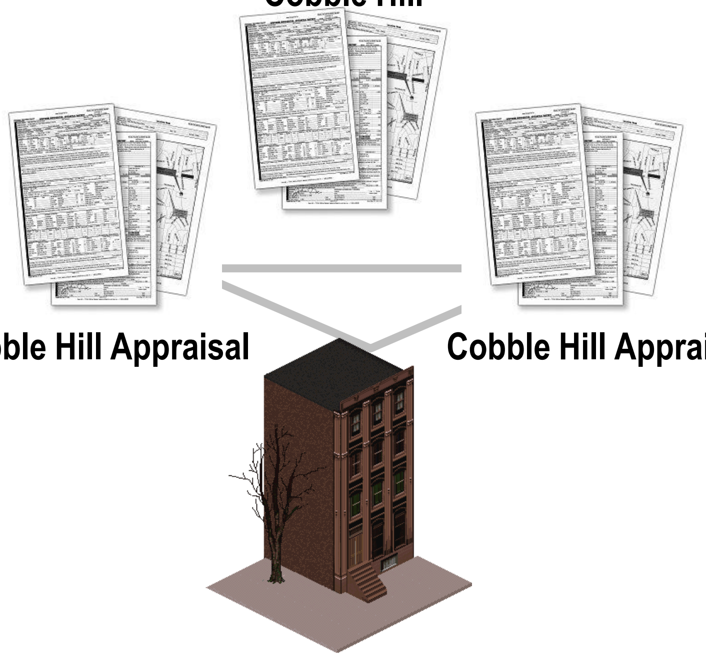
- ***Initial Question:*** Can a model predict the price of Brooklyn-based real estate?
  - ***Refined Question:*** Can the price of a 1, 2 or 3 family home in Brooklyn in 2003 & 2004 be predicted?

- ***Motivation:*** Improve the real estate appraisal process
  - *Significant variation between listing price and appraisal price:*
    - *Assessment commonly based on intuition, hyper-local comparable sales*
    - *Fear of appraising at too high of price, causing price distortions*
    - *Regulation encourages banks to use in-house or affiliated appraisers, creating a conflict of interest*

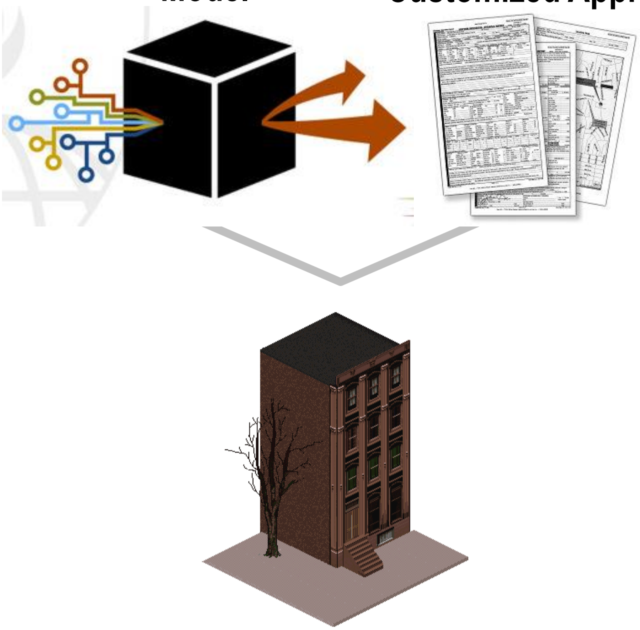
Hypothesis

Real estate appraisals based on a broader data set (e.g. geographical, historical sales, etc.) would improve market transparency for buyers and sellers

Common Appraisal Process



Enhanced Appraisal Process



Data Summary

Selected Data Fields

- Address
- Latitude / Longitude
- Neighborhood
- Median Neighborhood Income
- Building Type
- Square Feet
- Year Built
- Sale Price
- Sale Date

Data Sources

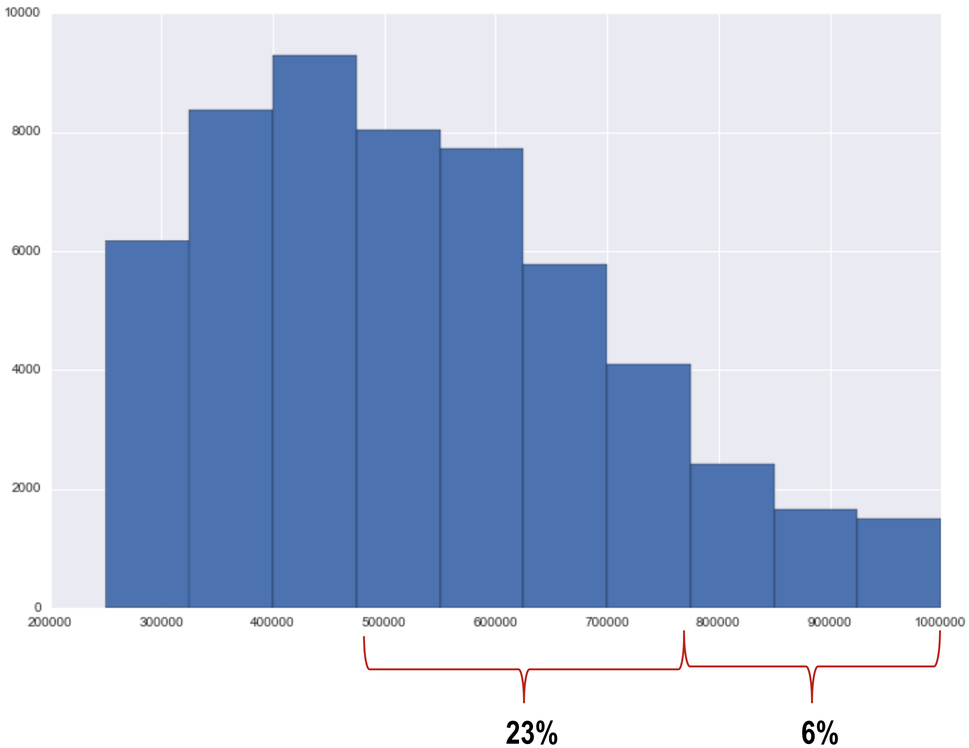
- NYC.gov
- Data Science Tool Kit.org

Data Set Overview

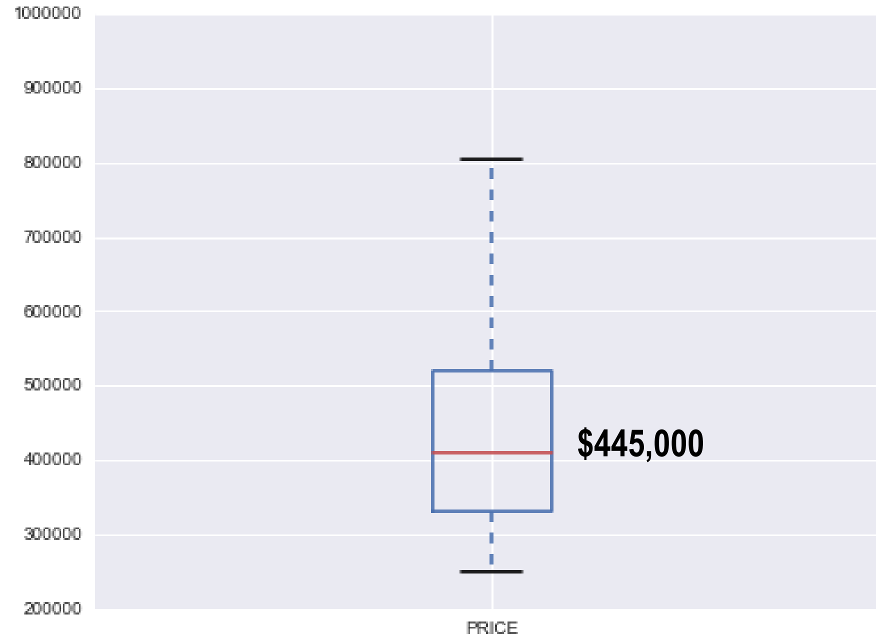
- 100,000 transactions (‘03-’09)
- 16,000 transactions (‘03-’04)
- 59 neighborhoods
- Income range: $27k – $111k
- 41 Types (Res. + Comm.)
- 500sf – 1000000sf
- Built 1800 – 2009
- $250k - $200M
- 2003 – 2009

Exploratory Approach: Making Sense of Brooklyn Real Estate

Histograms



Box Plots



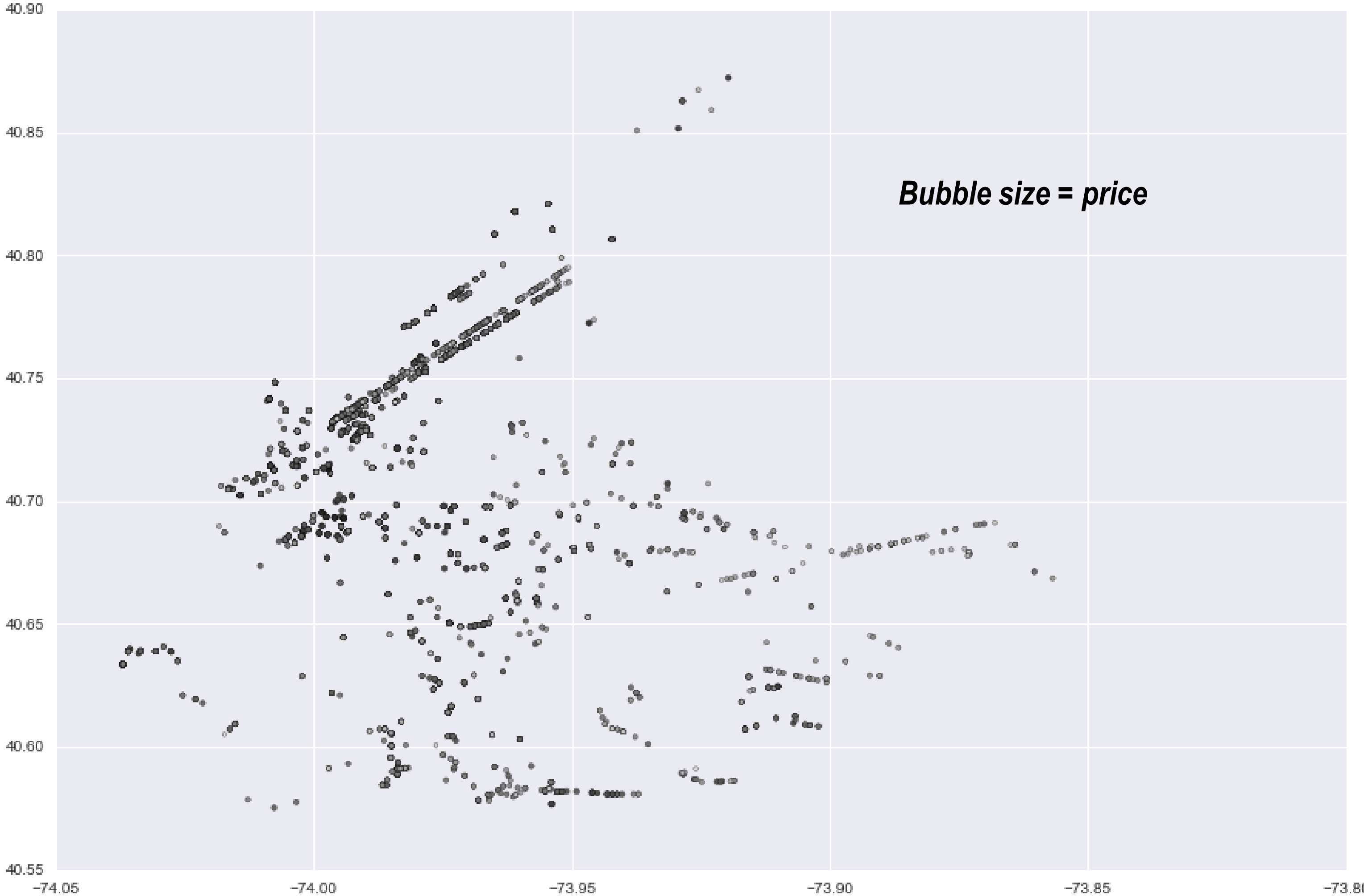
- Residential Brooklyn real estate* is characterized by a few high priced outliers

*One, Two and Three Family Homes

Exploratory Approach: Making Sense of Brooklyn Real Estate

Visualizations

Latitude



Longitude

Predictive Analysis Approach

1) **Feature Selection**: Identify features were *p* value < .05

Tested Features

- # of Commercial Units
- # of Residential Units
- Latitude / Longitude
- Square Footage
- Year Built

Identified Features

- # of Residential Units
- Latitude
- Square Footage

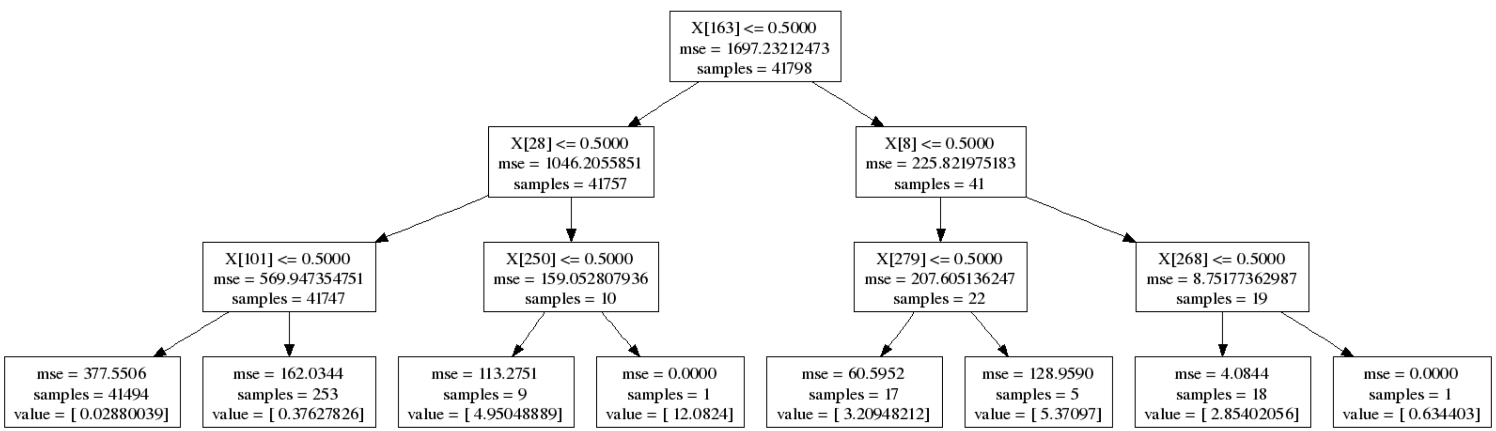
Predictive Analysis Approach

2) Linear regression with identified features – *Price as dependent variable*

```
NORMAL FIT SUMMARY
                            OLS Regression Results
==============================================================================
Dep. Variable:                  PRICE   R-squared:                       0.078
Model:                            OLS   Adj. R-squared:                  0.078
Method:                 Least Squares   F-statistic:                     414.2
Date:                Sat, 18 Apr 2015   Prob (F-statistic):          2.81e-258
Time:                        14:32:11   Log-Likelihood:            -2.0234e+05
No. Observations:               14643   AIC:                         4.047e+05
Df Residuals:                   14639   BIC:                         4.047e+05
Df Model:                           3
====================================================================================
                       coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------
Intercept        -7.602e+06   1.48e+06     -5.152      0.000     -1.05e+07 -4.71e+06
GROSSSIZE           53.0075      1.605     33.018      0.000        49.861    56.154
RESIDENTIALUNITS -2.697e+04   2672.907    -10.089      0.000     -3.22e+04 -2.17e+04
latitude          1.962e+05   3.62e+04      5.414      0.000      1.25e+05  2.67e+05
==============================================================================
Omnibus:                    21780.243   Durbin-Watson:                   0.948
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         34638917.554
Skew:                           8.577   Prob(JB):                         0.00
Kurtosis:                     240.654   Cond. No.                     2.17e+06
==============================================================================
```

Predictive Analysis Approach

3) **Generate Decision Tree**
   - a) Feature Selection: # of Residential Units, Latitude, Square Footage
   - b) Split into Test / Train
   - c) Calculate Mean Square Errors (large!!! – yikes!)
   - d) Run cross-validation to find optimal depth
   - e) Convert .DOT file into .PNG and use GraphViz to visualize



Lessons Learned

1) Linear regression offered little explanation of price (dependent variable)

2) Decision trees provided a more structured approach to segmenting the data
   - However, MSE was large

3) Segmenting data into neighborhoods based on median income did not improve decision tree

Next Steps

1) Conduct time series analysis of price fluctuations using real estate transactions from 2003 – 2014

2) Visualize price / square footage changes

3) Broaden analysis by including additional demographic information
   - Income
   - Families
   - Commercial development (e.g. how does introducing a Starbucks or Whole Foods into the neighborhood impact price?)