New York Condominiums and Complaints

Can complaints have an affect on the price of Condominiums?

Dr. James R. Gatewood
Junior Data Scientist, Mathematician
Urban Planner/designer Enthusiast

Presentation Outline

1. Motivation
2. Analysis
3. Results
4. Additional Questions
5. Applications
6. Conclusion

Motivation

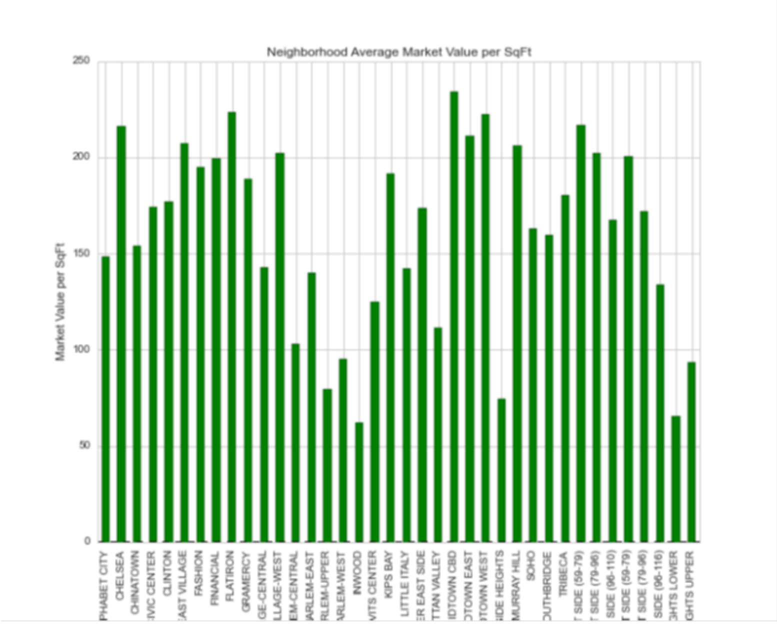
- Since the industrial revolution, people have been moving to cities for opportunities and stimulation
- Cities are dense clusters of vibrant centers of human activity: research, commerce, tourism, culture, etc
- Use mathematical models to study cities which will aid in the in-depth understanding of their workings and how they evolve over time.
- Its an interesting science, but also may be useful in a planning scenario

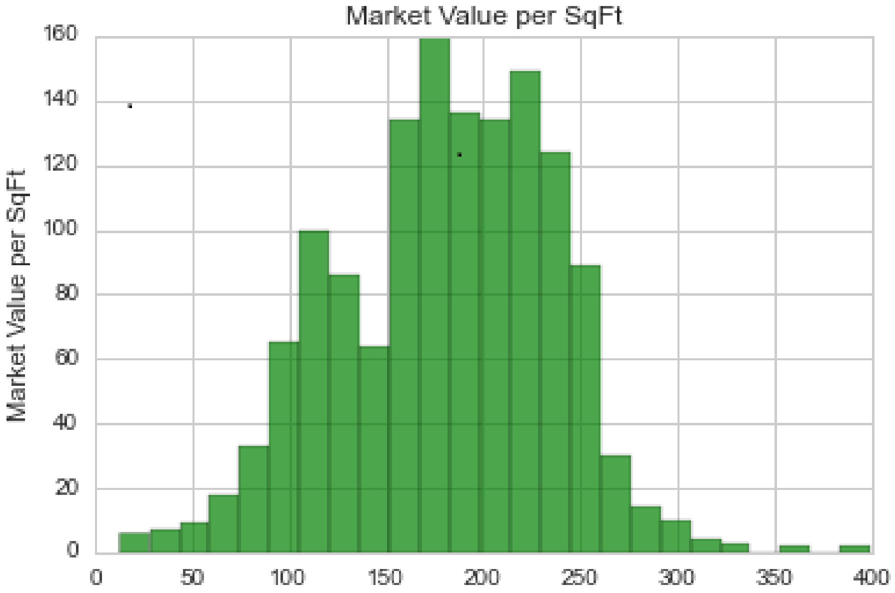
Motivation

- Do certain characteristics and complaints have an effect on the prices of Manhattan Condominiums?
  - Only explored Manhattan Neighborhoods
  - Choose a few complaints out of dozens
- Hypothesis: *Yes, complaints have some effect on the overall cost of the condominiums.*
- Data Sources (files from www.nycopendata.socrata.com)
  - DOF Condominiums comparable rental incomes
  - 311 Service Requests

Data Wrangling: Clean and Transform

- Target Variable: Market Value per Sqft
  - A calculation of the value of each square foot of an area of a house, condo or any building. It is a simple, but useful calculation that is mostly used to compare similar properties.

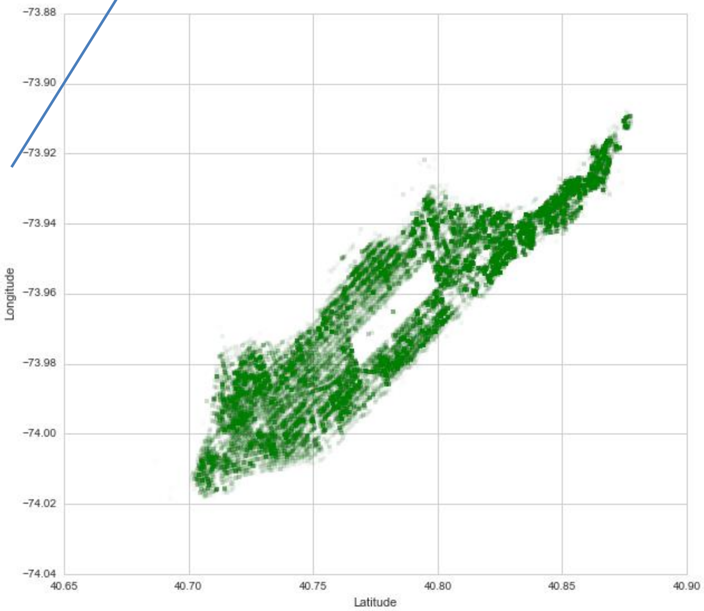




Data Wrangling: Clean and Transform

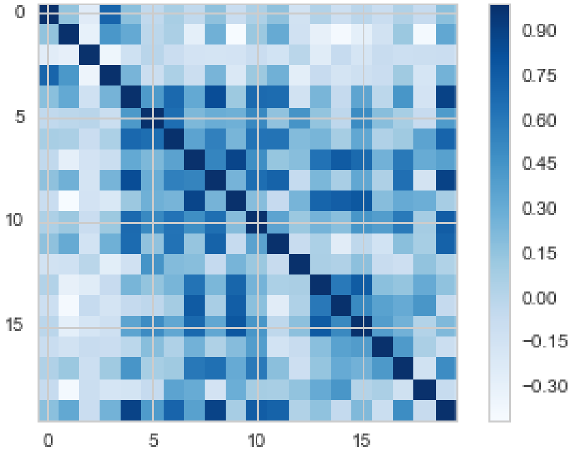
- Features:
  1. Total Number of Units in Condominiums
  2. Year Built (Changed to age of building)
  3. Estimated Gross Income (Applied the Log function)
  4. Taxi Complaint
  5. Noise - Commercial
  6. Food Establishment
  7. Noise - Vehicle
  8. Street Condition
  9. Noise - Street/Sidewalk
  10. Sidewalk Condition Traffic
  11. Graffiti
  12. Elevator
  13. School Maintenance
  14. DOF Property - Reduction Issue
  15. Root/Sewer/Sidewalk Condition
  16. Overgrown Tree/Branches
  17. Construction
  18. Noise

Transformed into dummy variables and fill in the neighborhoods



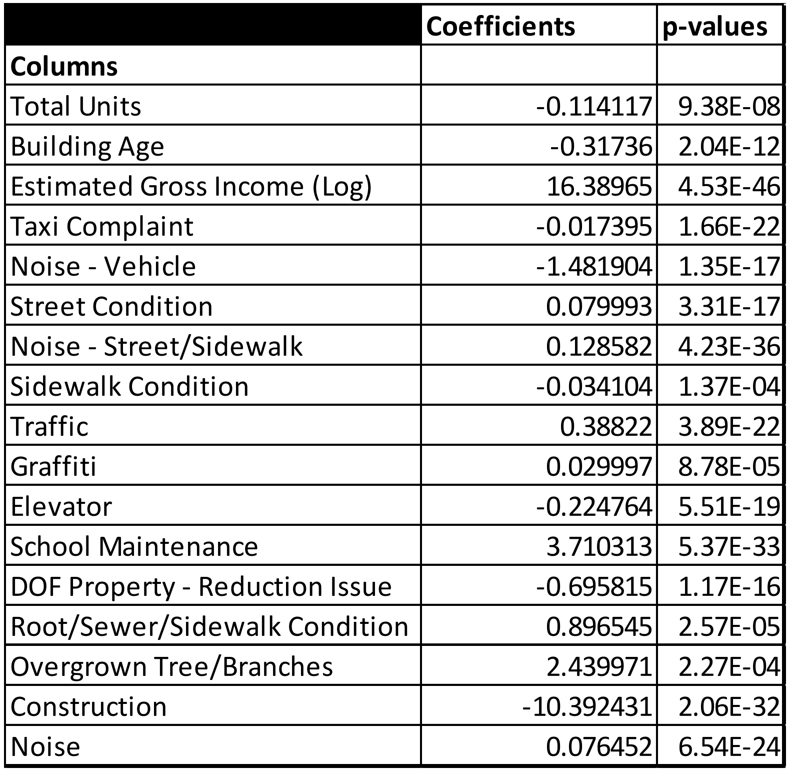
Analysis

- Two Machine Learning Algorithms:
  - Ridge Regression and Decision Tree Regression
    - – Use ridge regression: when too many independent variables have a near linear relationship, multicollinearity occurs
    - – Ridge regression adds a degree of bias to regression estimates
  - Decision Tree Regression
    - – Builds a regression model in the form of a tree structure (since my output is continuous)



Analysis

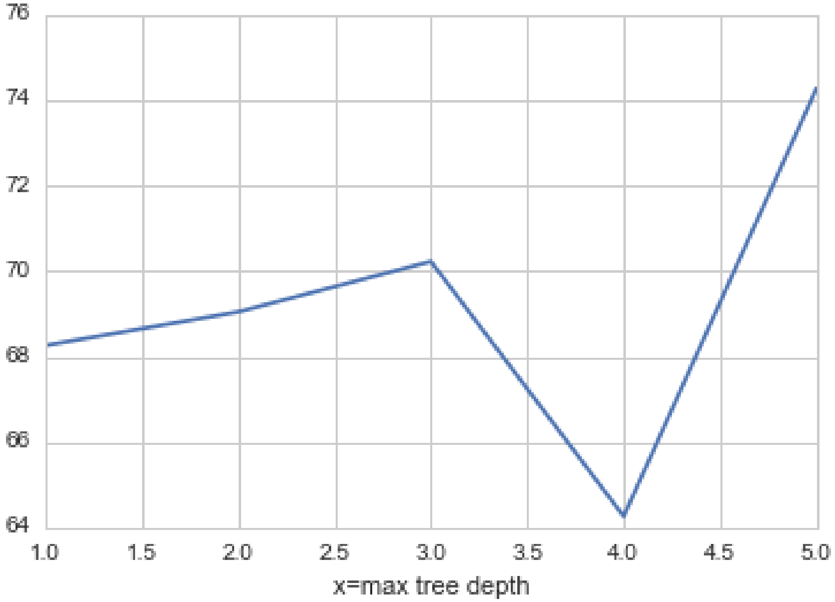
- Ridge Regression results
- R^2 = 0.6650934
- Complaints are not the best predictors of Market Value Sqrt



	Coefficients	**p-values**
Columns		
Total Units	-0.114117	9.38E-08
Building Age	-0.31736	2.04E-12
Estimated Gross Income (Log)	16.38965	4.53E-46
Taxi Complaint	-0.017395	1.66E-22
Noise - Vehicle	-1.481904	1.35E-17
Street Condition	0.079993	3.31E-17
Noise - Street/Sidewalk	0.128582	4.23E-36
Sidewalk Condition	-0.034104	1.37E-04
Traffic	0.38822	3.89E-22
Graffiti	0.029997	8.78E-05
Elevator	-0.224764	5.51E-19
School Maintenance	3.710313	5.37E-33
DOF Property - Reduction Issue	-0.695815	1.17E-16
Root/Sewer/Sidewalk Condition	0.896545	2.57E-05
Overgrown Tree/Branches	2.439971	2.27E-04
Construction	-10.392431	2.06E-32
Noise	0.076452	6.54E-24

Analysis

- **Decision Tree Regression**
- Best score: 64.2668180897
- Best depth: 4



Feature	Importance
Total Units	0.01349
Building Age	0.13960
Estimated Gross Income Log	0.19982
Taxi Complaint	0.00000
Noise - Commercial	0.00000
Food Establishment	0.00000
Noise - Vehicle	0.06276
Street Condition	0.08383
Noise - Street/Sidewalk	0.44266
Sidewalk Condition	0.05384
Traffic	0.00000
Graffiti	0.00000
Elevator	0.00399
School Maintenance	0.00000
DOF Property - Reduction Issue	0.00000
Root/Sewer/Sidewalk Condition	0.00000
Overgrown Tree/Branches	0.00000
Construction	0.00000
Noise	0.00000

Challenges

- Multi-collinearity
- Which machine learning techniques to use
- Which features to explores and which to omit.
- Truly understanding what the values of the results represent.

Additional Questions

- What would the effect be if I included all complaints from the file?
- How do the complaints vary (seasonally)?
- Would the results be different if I explored zip codes instead of neighborhoods?
- Additional machine learning algorithms.

Conclusions

- Questions?????

- Thank you
  - Ed
  - Julia
  - Pooja