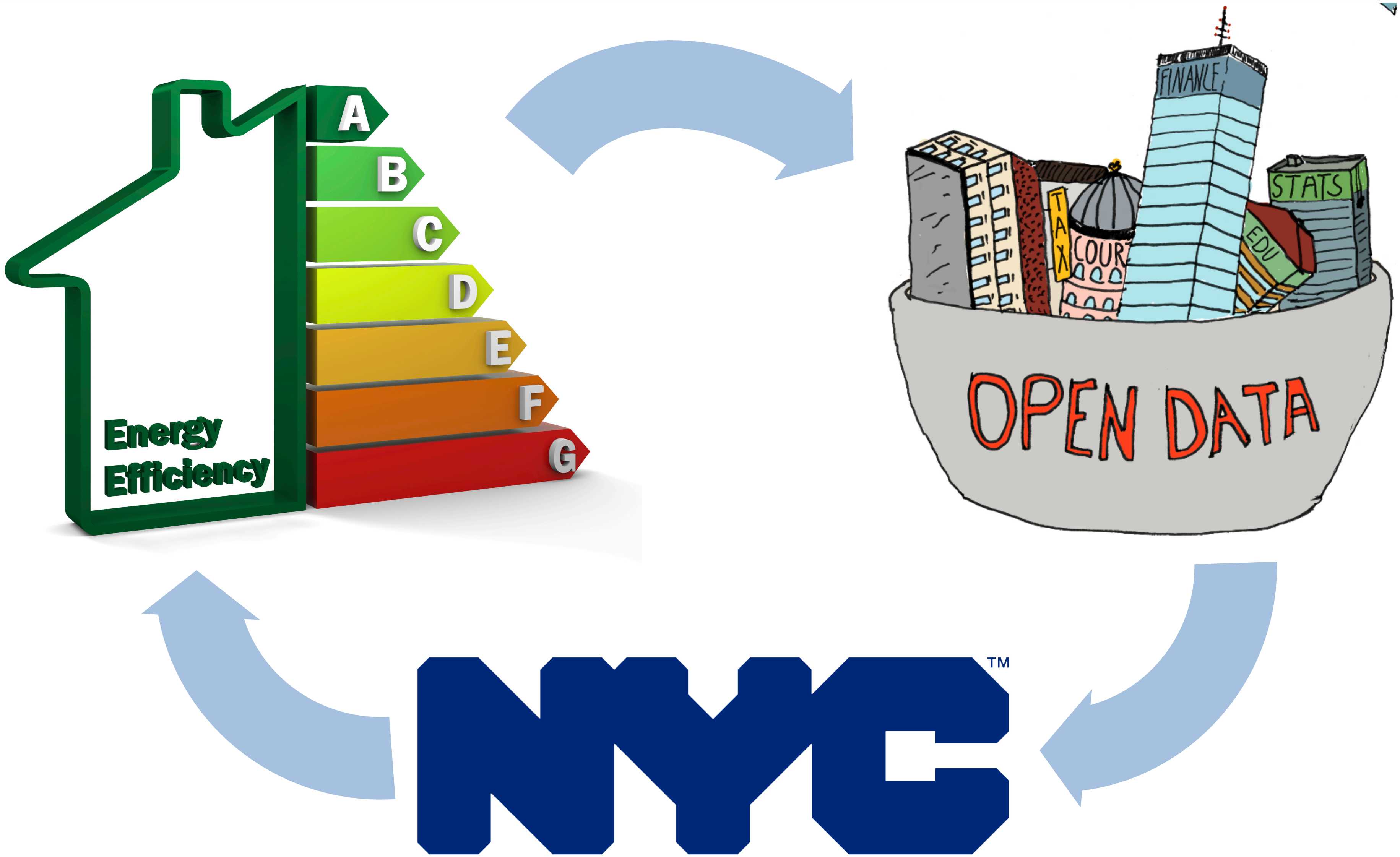
Final Project Presentation GA Data Science 18

Theodore Love

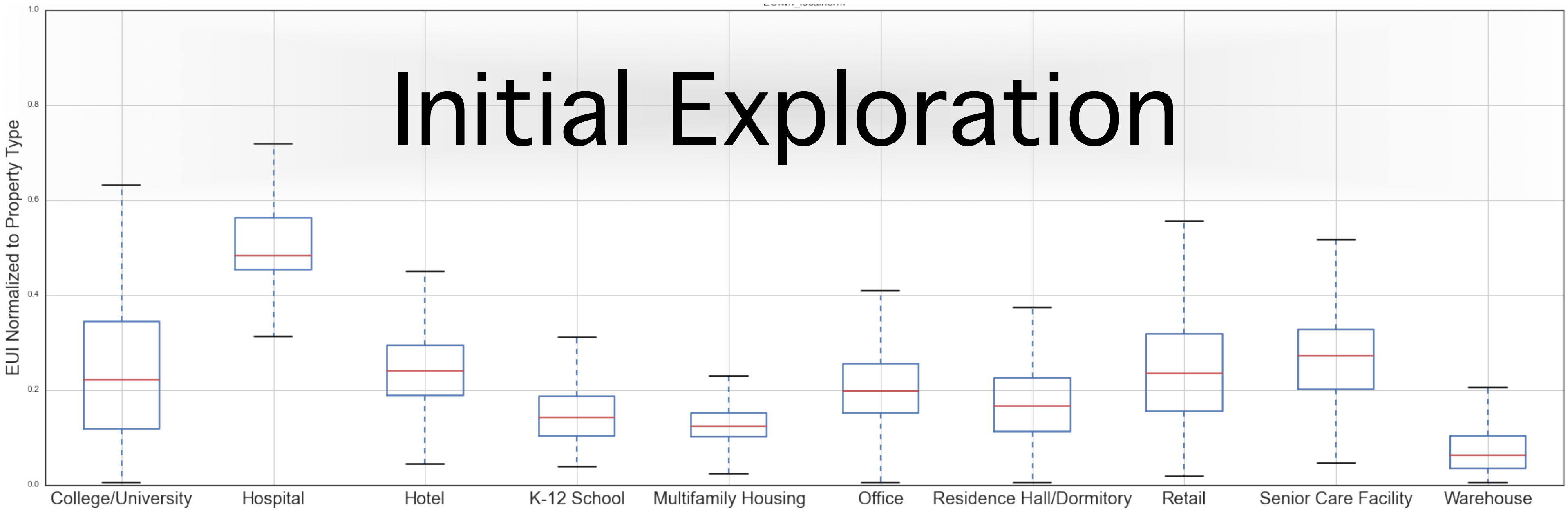
4/20/2015

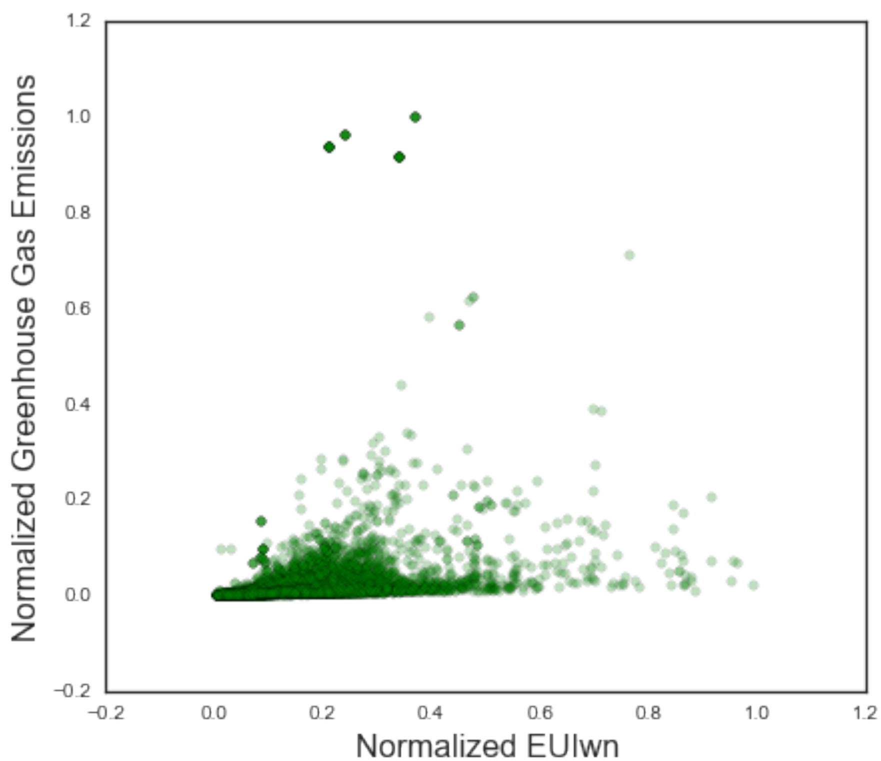
The Inspiration

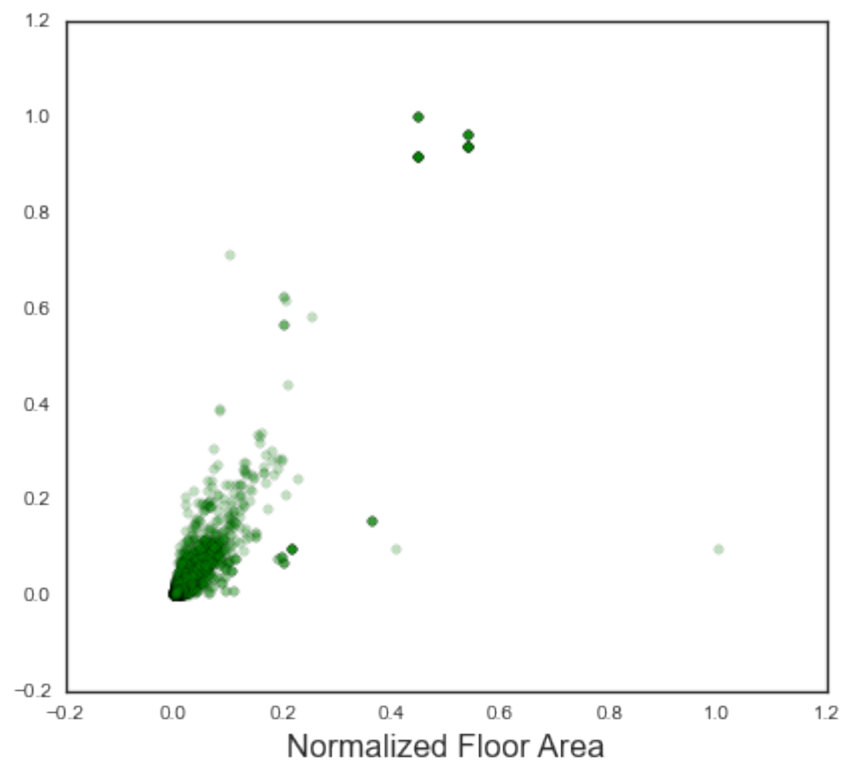


Local Law 84

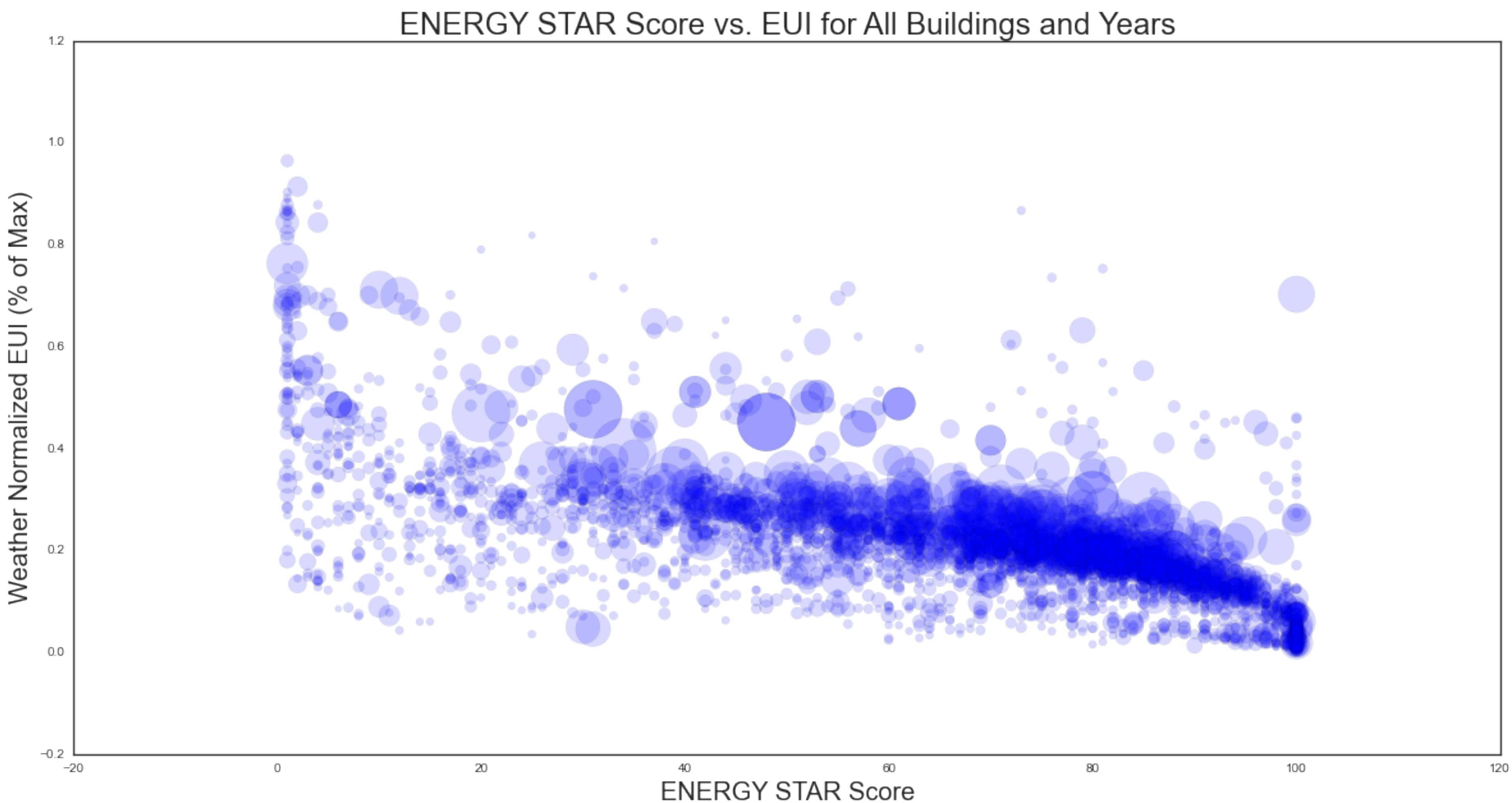
Initial Exploration



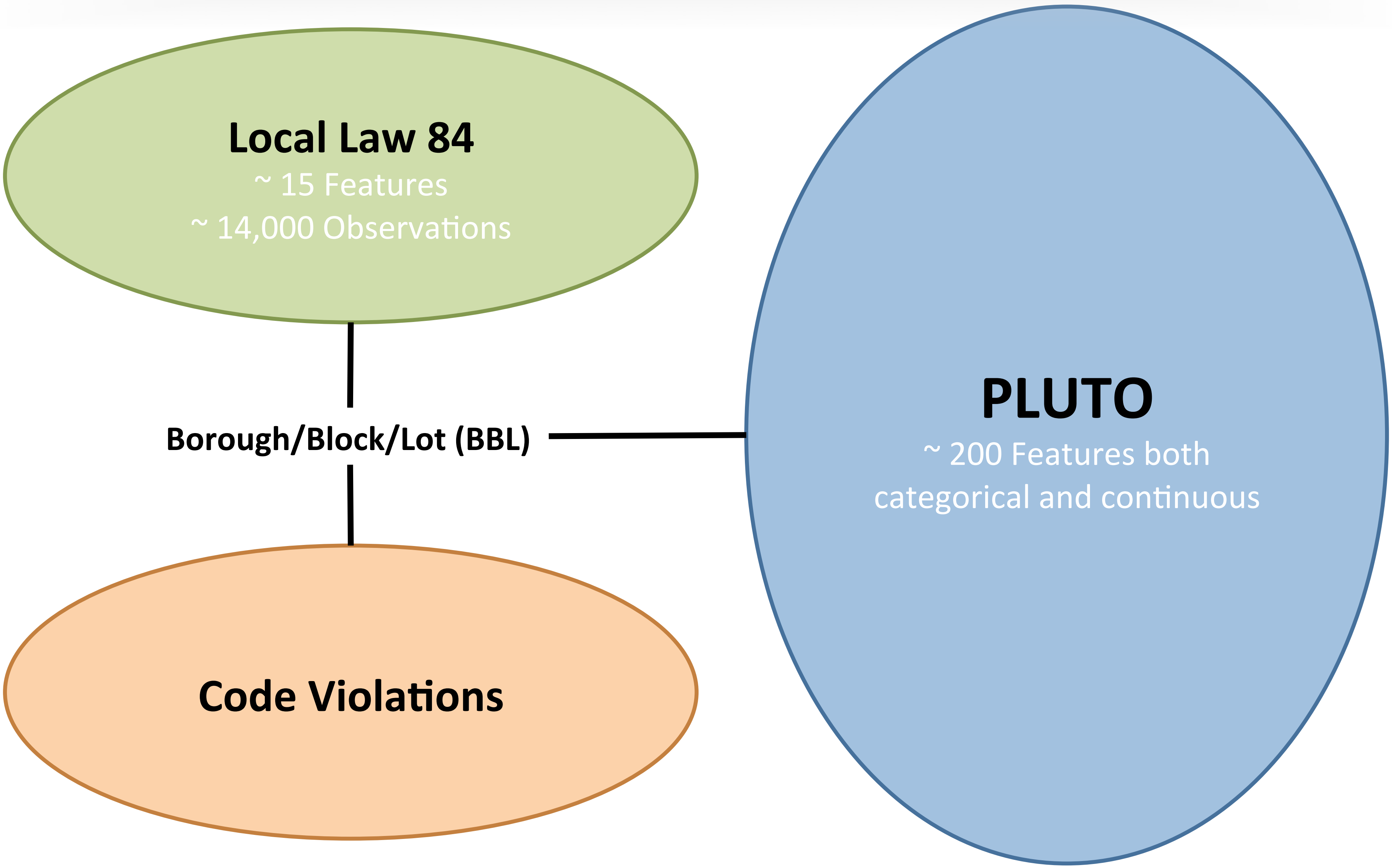




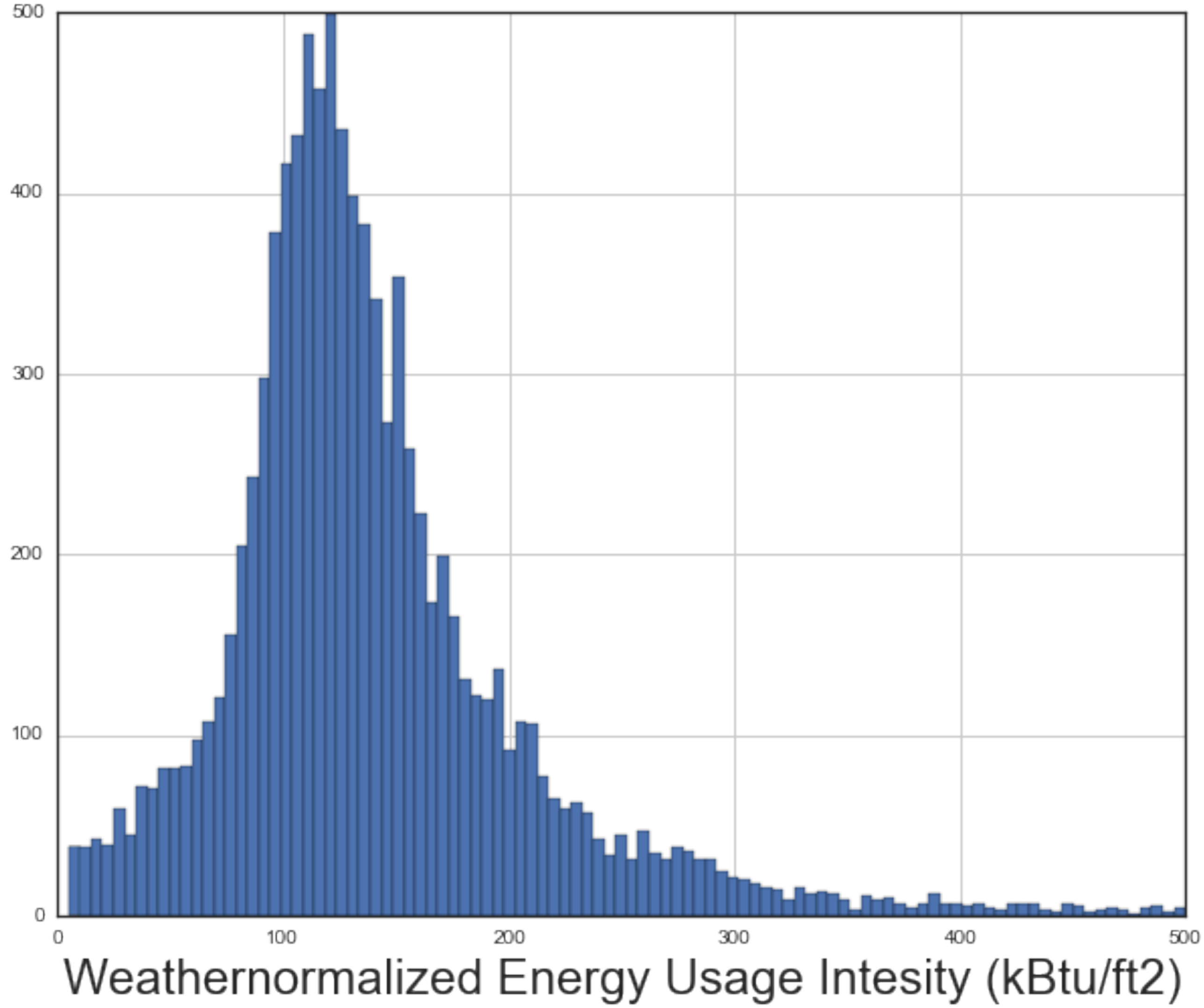
Existing Benchmarks



Going Further with New Data



Let's Try to Predict EUI



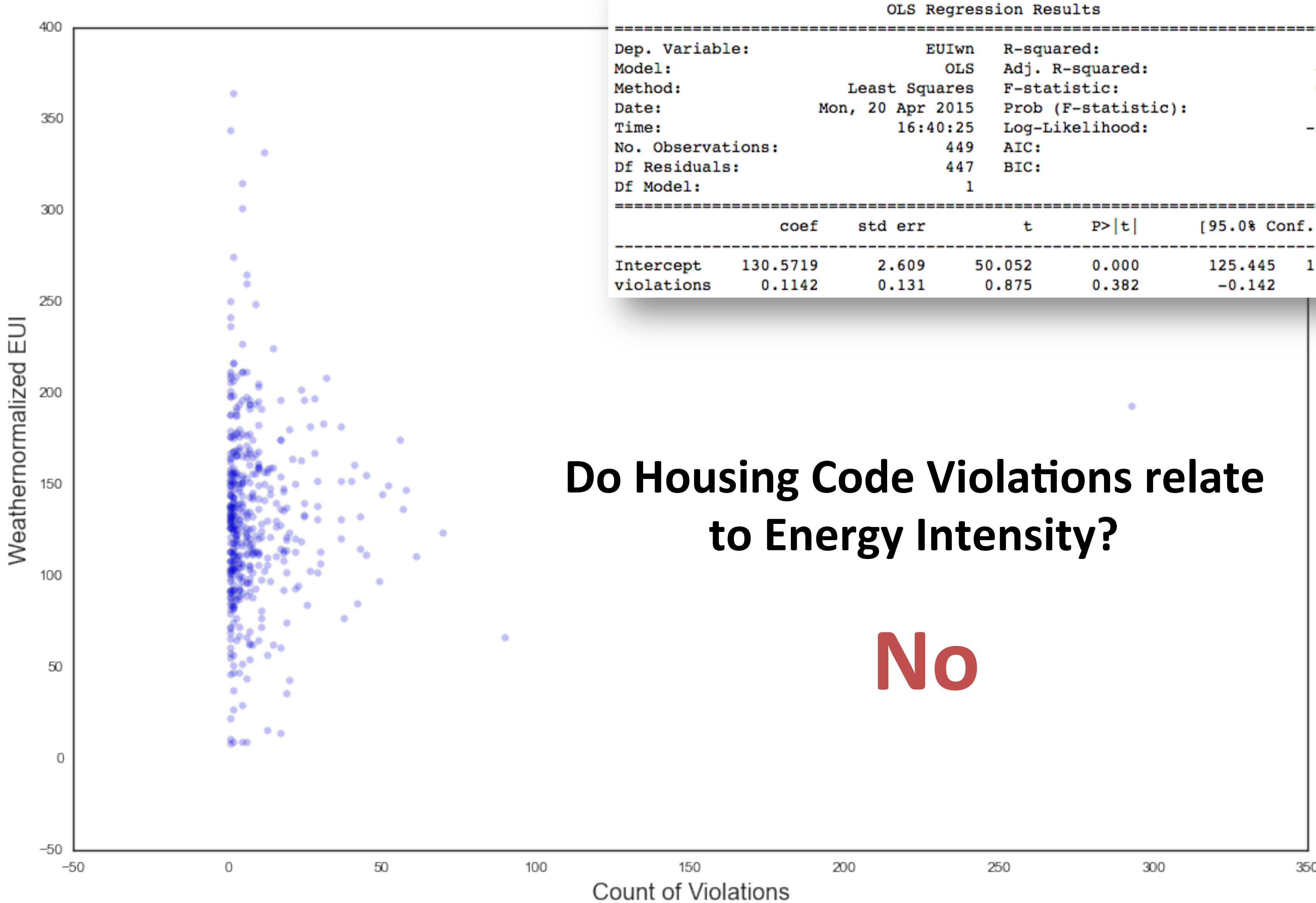
A Dead End

EUIwn ~ violations

OLS Regression Results

Dep. Variable:	EUIwn	R-squared:	0.002
Model:	OLS	Adj. R-squared:	-0.001
Method:	Least Squares	F-statistic:	0.7655
Date:	Mon, 20 Apr 2015	Prob (F-statistic):	0.382
Time:	16:40:25	Log-Likelihood:	-2382.5
No. Observations:	449	AIC:	4769.
Df Residuals:	447	BIC:	4777.
Df Model:	1		

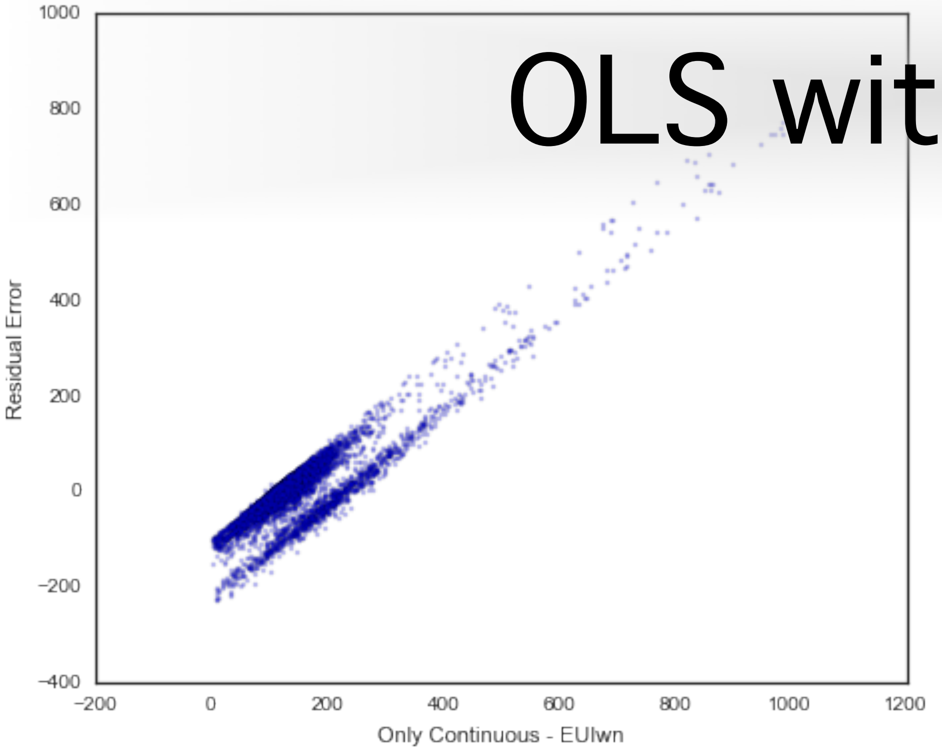
	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	130.5719	2.609	50.052	0.000	125.445	135.699
violations	0.1142	0.131	0.875	0.382	-0.142	0.371

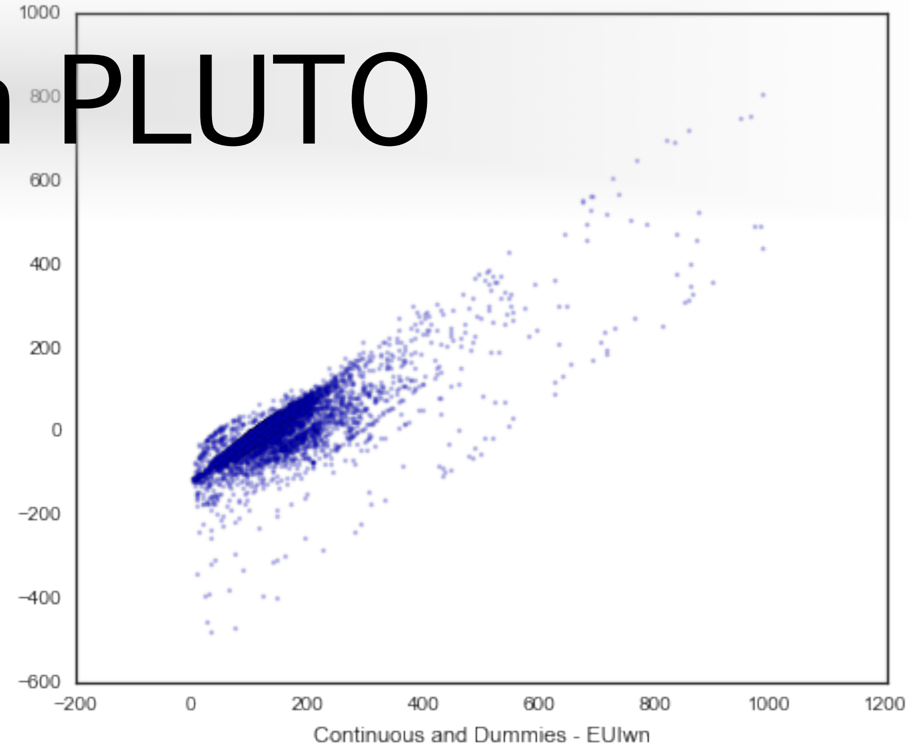


Do Housing Code Violations relate to Energy Intensity?

No

OLS with PLUTO





OLS Regression Results

Dep. Variable:	EUIwn	R-squared:	0.220
Model:	OLS	Adj. R-squared:	0.219
Method:	Least Squares	F-statistic:	212.6
Date:	Mon, 20 Apr 2015	Prob (F-statistic):	0.00
Time:	17:01:21	Log-Likelihood:	-39550.
No. Observations:	6811	AIC:	7.912e+04
Df Residuals:	6801	BIC:	7.919e+04
Df Model:	9		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	114.9118	1.916	59.969	0.000	111.156	118.668
PerComArea	-18.1909	4.605	-3.950	0.000	-27.219	-9.163
PerOfficeArea	120.2945	5.997	20.058	0.000	108.538	132.051
PerOtherArea	129.8724	6.077	21.372	0.000	117.960	141.784
PerRetailArea	154.0383	9.447	16.305	0.000	135.519	172.558
PerGarageArea	125.3753	14.756	8.497	0.000	96.450	154.301
Easements	19.8924	6.129	3.245	0.001	7.877	31.908
NumBldgs	0.8138	0.160	5.079	0.000	0.500	1.128
NumFloors	1.2586	0.146	8.595	0.000	0.972	1.546
MaxAllwFAR	-1.3232	0.367	-3.601	0.000	-2.044	-0.603

Omnibus:	5182.231	Durbin-Watson:	1.994
Prob(Omnibus):	0.000	Jarque-Bera (JB):	164207.190
Skew:	3.351	Prob(JB):	0.00
Kurtosis:	26.102	Cond. No.	237.

OLS Regression Results

Dep. Variable:	EUIwn	R-squared:	0.359
Model:	OLS	Adj. R-squared:	0.356
Method:	Least Squares	F-statistic:	111.5
Date:	Wed, 18 Mar 2015	Prob (F-statistic):	0.00
Time:	23:14:04	Log-Likelihood:	-38881.
No. Observations:	6811	AIC:	7.783e+04
Df Residuals:	6776	BIC:	7.807e+04
Df Model:	34		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
Intercept	119.5677	2.098	56.999	0.000	115.455	123.680
PerOfficeArea	66.9291	3.782	17.697	0.000	59.515	74.343
PerOtherArea	39.5388	5.277	7.493	0.000	29.194	49.883
PerRetailArea	48.5308	10.524	4.612	0.000	27.901	69.161
PerGarageArea	46.2411	16.258	2.844	0.004	14.371	78.111
Easements	15.8513	5.595	2.833	0.005	4.884	26.818
NumFloors	1.2730	0.113	11.266	0.000	1.051	1.494
I1	350.1366	13.971	25.061	0.000	322.748	377.525
			•••			
O6	143.0251	51.820	2.760	0.006	41.443	244.608
G1	97.2183	25.589	3.799	0.000	47.056	147.381
F9	-56.6767	14.451	-3.922	0.000	-85.006	-28.347

Omnibus:	4195.869	Durbin-Watson:	1.994
Prob(Omnibus):	0.000	Jarque-Bera (JB):	129991.633
Skew:	2.436	Prob(JB):	0.00
Kurtosis:	23.840	Cond. No.	826.

Tree Regressions with Full PLUTO Data

	features	importance
147	LandUse_05	0.129002
77	BldgClass_I1	0.080447
25	YearBuilt	0.066685
19	AssessLand	0.038579
20	AssessTot	0.037660
26	YearLastWork	0.037520
150	LandUse_08	0.036264
3	ComArea	0.030964
1	LotArea	0.030778
15	LotFront	0.029329
16	LotDepth	0.027269
18	BldgDepth	0.027257
23	BuiltFAR	0.026796
12	NumFloors	0.025325
22	ExemptTot	0.024598
17	BldgFront	0.024171
14	UnitsTotal	0.023993
2	BldgArea	0.019196
5	OfficeArea	0.017562
30	PerRetailArea	0.016927

Hospitals and Health

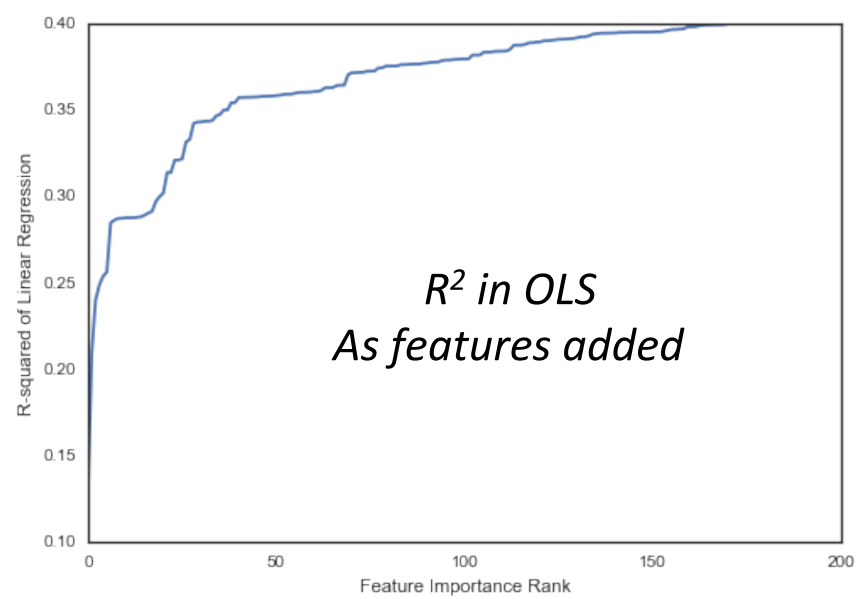
Commercial & Office Buildings

Age of Building

Value of building/land

Public Facilities & Institutions

Size/usage percent of building



Optimization of Decision Tree Regression

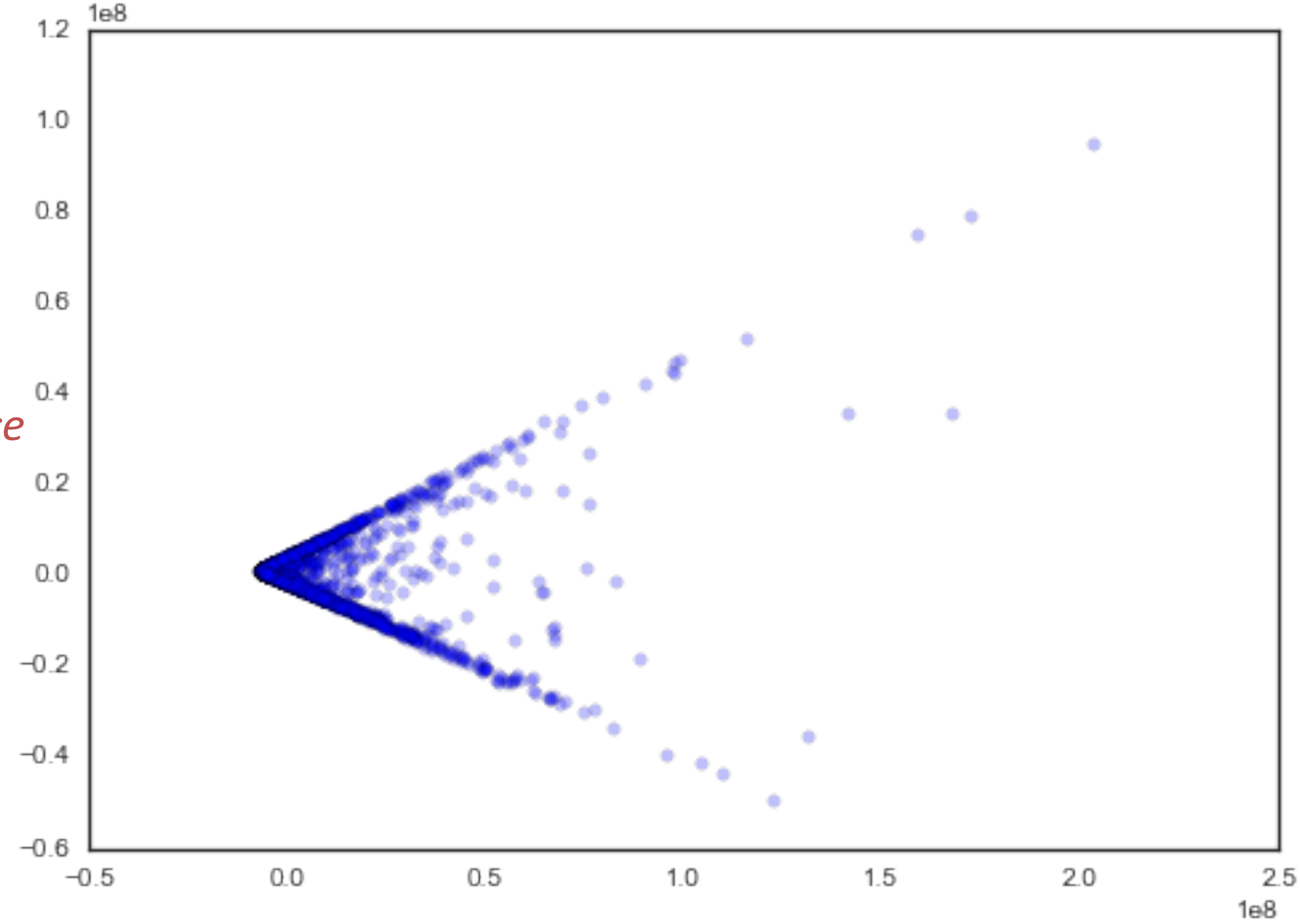
Cross-validation found over fitting issues across sub-samples even with low depths.

Random Forest Regression

R^2 from 0.23 to 0.29 with std dev ~0.03

Quick PCA

PCA 2
12% of variance



PCA 1
83% of variance

Next Steps

- Further examination of PCA
- Exploration of residuals and classification of high users
- See how regressions works on new data
- Prepare white paper (BECC Conference)

theolove@gmail.com