Flatiron Project 1

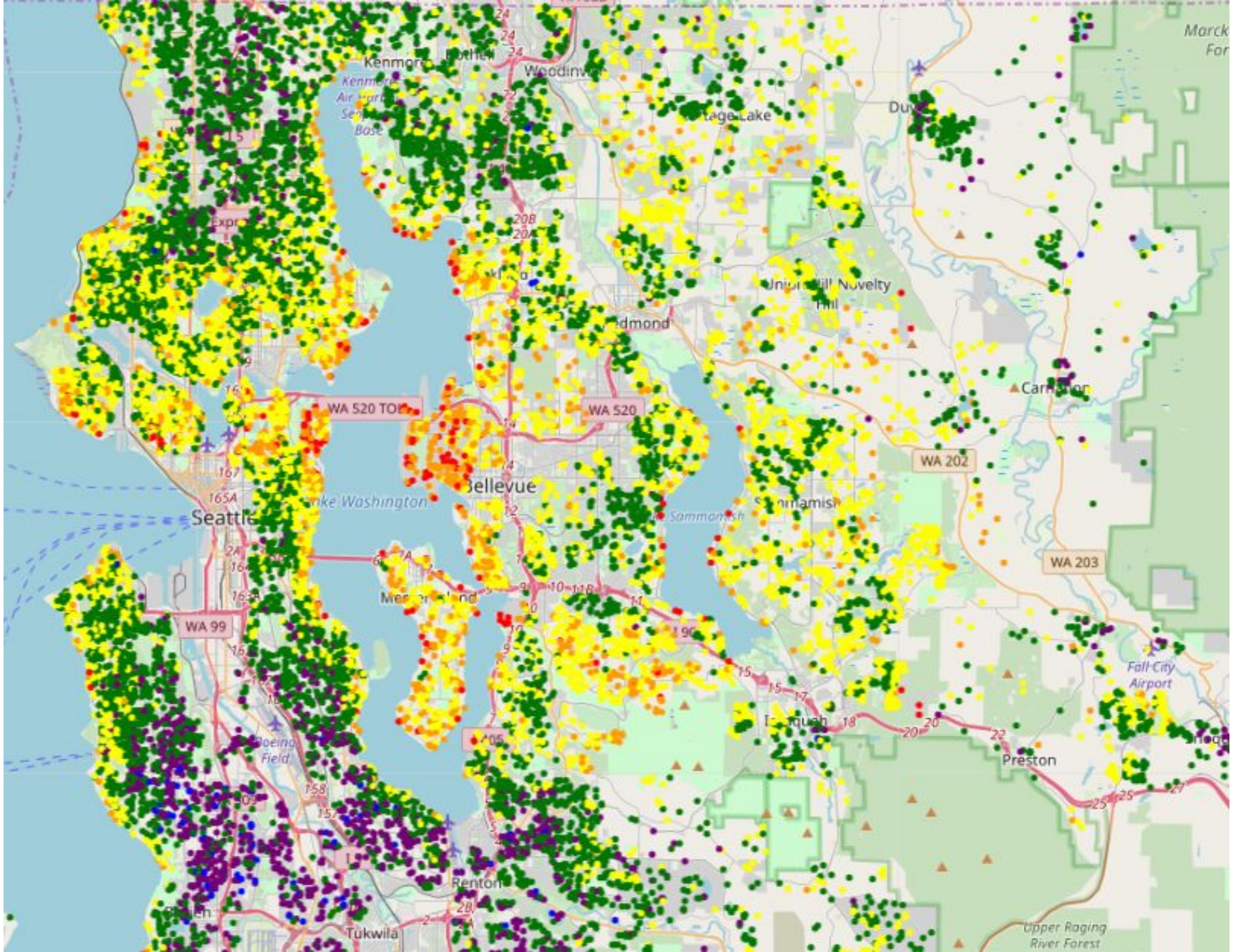
Kings County Housing Data

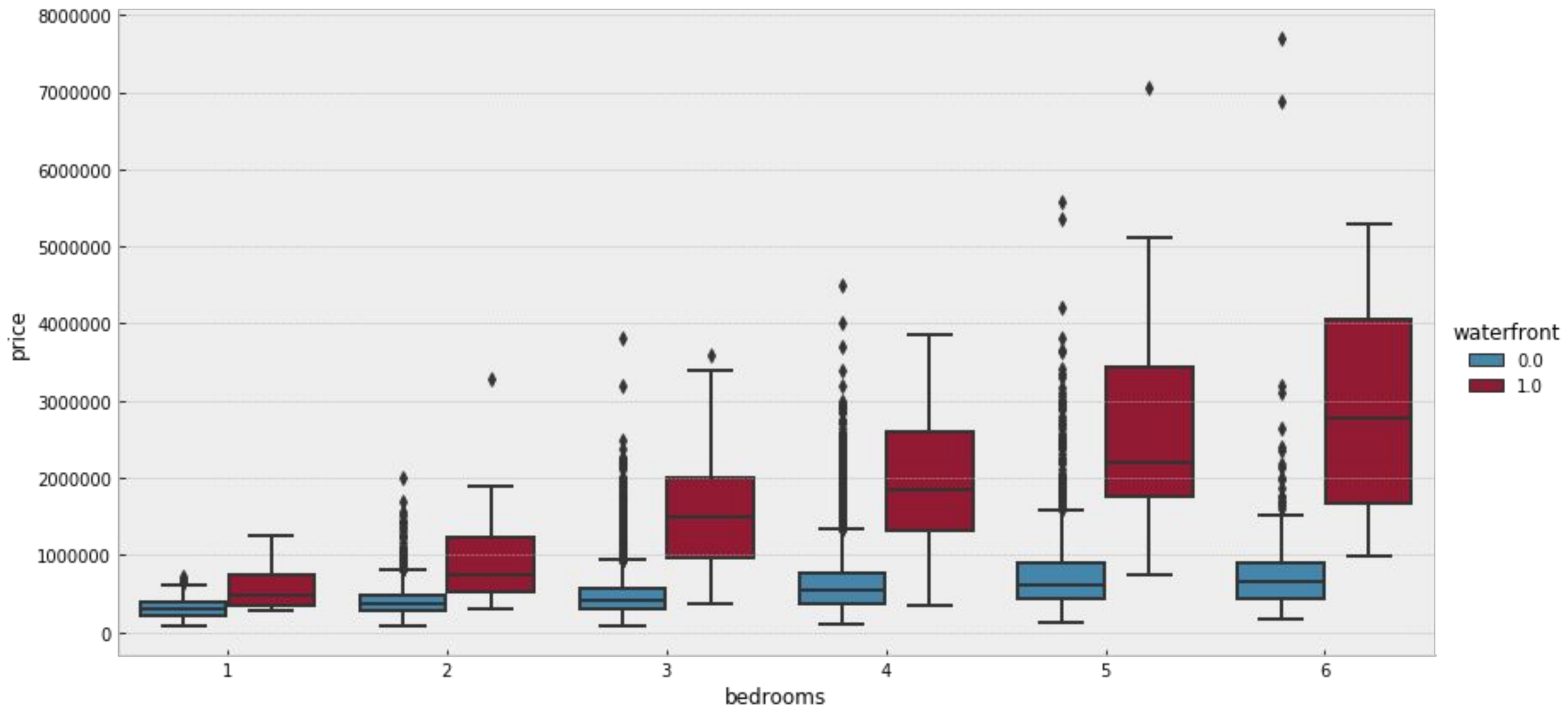
Intuitive Factors Contributing to House Price

- Size, number of bedrooms/bathrooms
- Condition of house, grade, recent renovation
- Location: waterfront view, neighborhood factors
- Year sold

Issues

- How many of these factors can we consider at once before we run into overfit issues?
- Do we expect the magnitude of any of these effects to vary between meaningful subsets of the data?


Kenmore
WA 520
WA 99
WA 202
WA 203
Novelty Hill
Boeing Field
Renton
Tukwila
Preston
Fall City Airport
Upper Raging River Forest


8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
price
1
2
3
4
5
6
bedrooms
waterfront
0.0
1.0

Related Factors Considered Together

- It can be risky piling in extra variables into a regression - particularly if those variables seem to be related/move together
- ‘Grade’ and ‘Condition’ seem to be getting at the same thing, for instance, it probably doesn’t make sense to include both
- Sometimes including related factors can be revealing - if the underlying data is robust, so interesting things can come to light
- Let’s consider ‘Bedrooms’ and ‘Square Feet of Living Space’

Regression comparison

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.1637	0.019	601.991	0.000	11.127	11.200
sqft_living	0.0002	3.98e-06	52.975	0.000	0.000	0.000
flag_2015	0.0331	0.005	6.567	0.000	0.023	0.043
reno_flag	0.2531	0.023	11.168	0.000	0.209	0.298
waterfront	0.6291	0.029	21.766	0.000	0.572	0.686
grade	0.1865	0.003	60.077	0.000	0.180	0.193

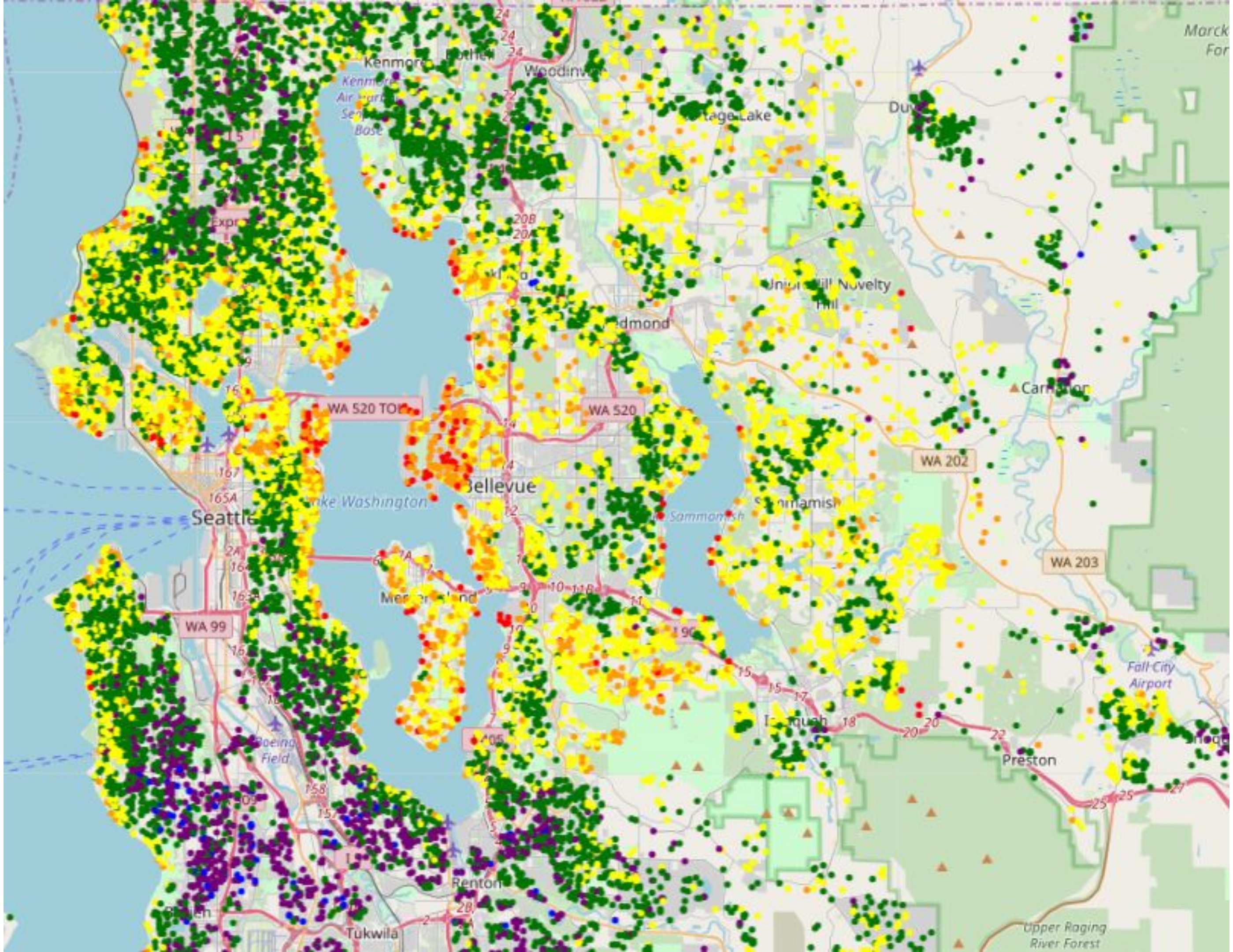
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.6783	0.018	582.144	0.000	10.642	10.714
bedrooms	0.0443	0.003	13.808	0.000	0.038	0.051
bathrooms	0.0720	0.005	15.285	0.000	0.063	0.081
flag_2015	0.0314	0.005	5.953	0.000	0.021	0.042
reno_flag	0.2609	0.024	11.001	0.000	0.214	0.307
waterfront	0.7378	0.030	24.435	0.000	0.679	0.797
grade	0.2677	0.003	94.929	0.000	0.262	0.273

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.2202	0.021	533.014	0.000	11.179	11.261
sqft_living	0.0002	4.67e-06	48.089	0.000	0.000	0.000
bedrooms	-0.0186	0.003	-5.658	0.000	-0.025	-0.012
flag_2015	0.0333	0.005	6.616	0.000	0.023	0.043
reno_flag	0.2542	0.023	11.221	0.000	0.210	0.299
waterfront	0.6159	0.029	21.256	0.000	0.559	0.673
grade	0.1836	0.003	58.352	0.000	0.177	0.190

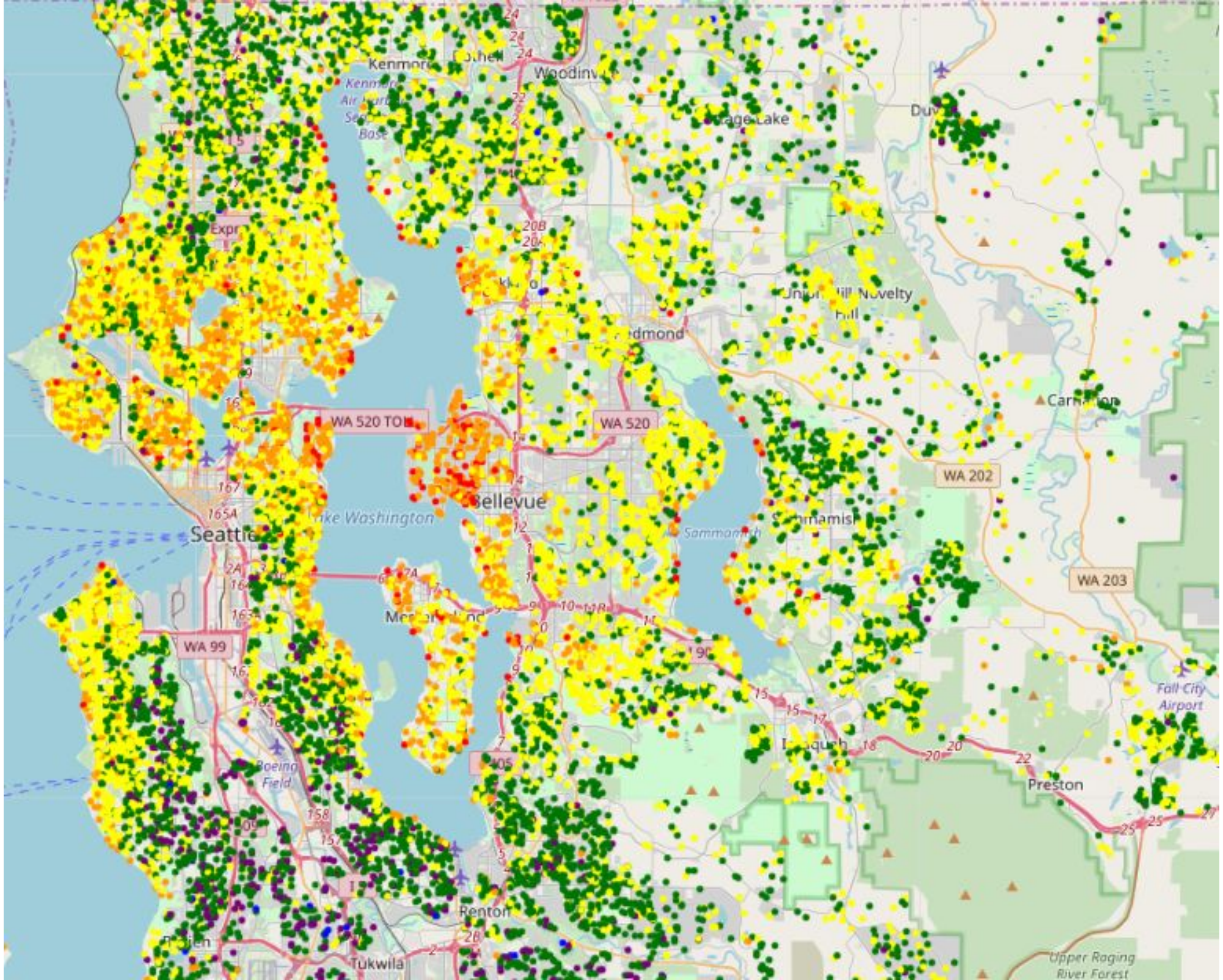
Incorporating Neighborhood

- Intuitively, neighborhood and location seem like they should matter a great deal
- Metrics might include distance/accessibility to some central area/downtown or neighborhood specific flags
- Zipcodes might serve as a useful way to demarcate neighborhoods, with two possible issues:
    - There are 70 zipcodes! Adding 69 dummy variables will be a burden on our regression. When do we run into overfit issues?
    - Some of the zipcodes in this county are big and might themselves contain some real variance within them
- Let’s consider the price map and another with an adjustment for illustration

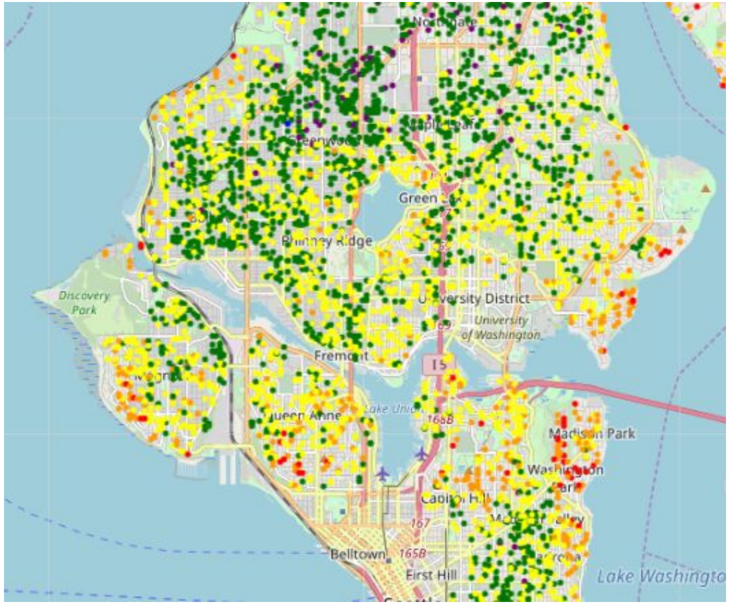
Original map



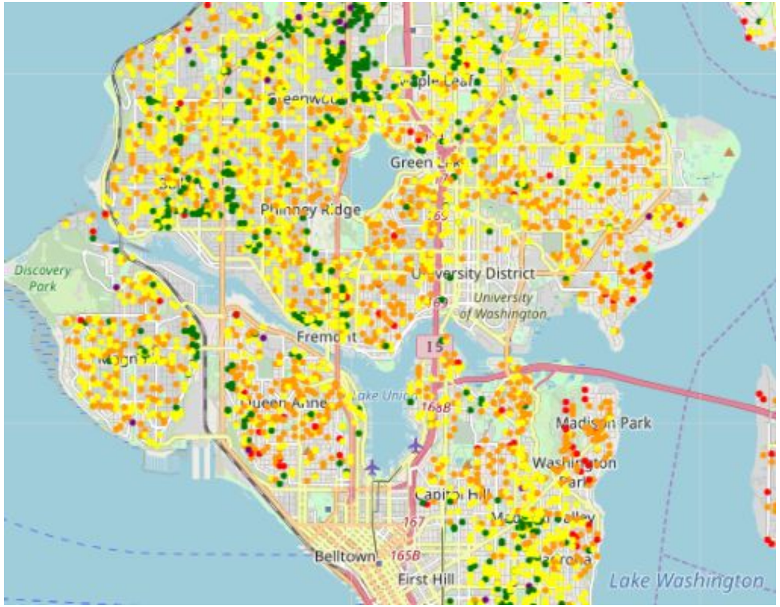
Map with ‘naive model’ errors



Clear Neighborhood Effects

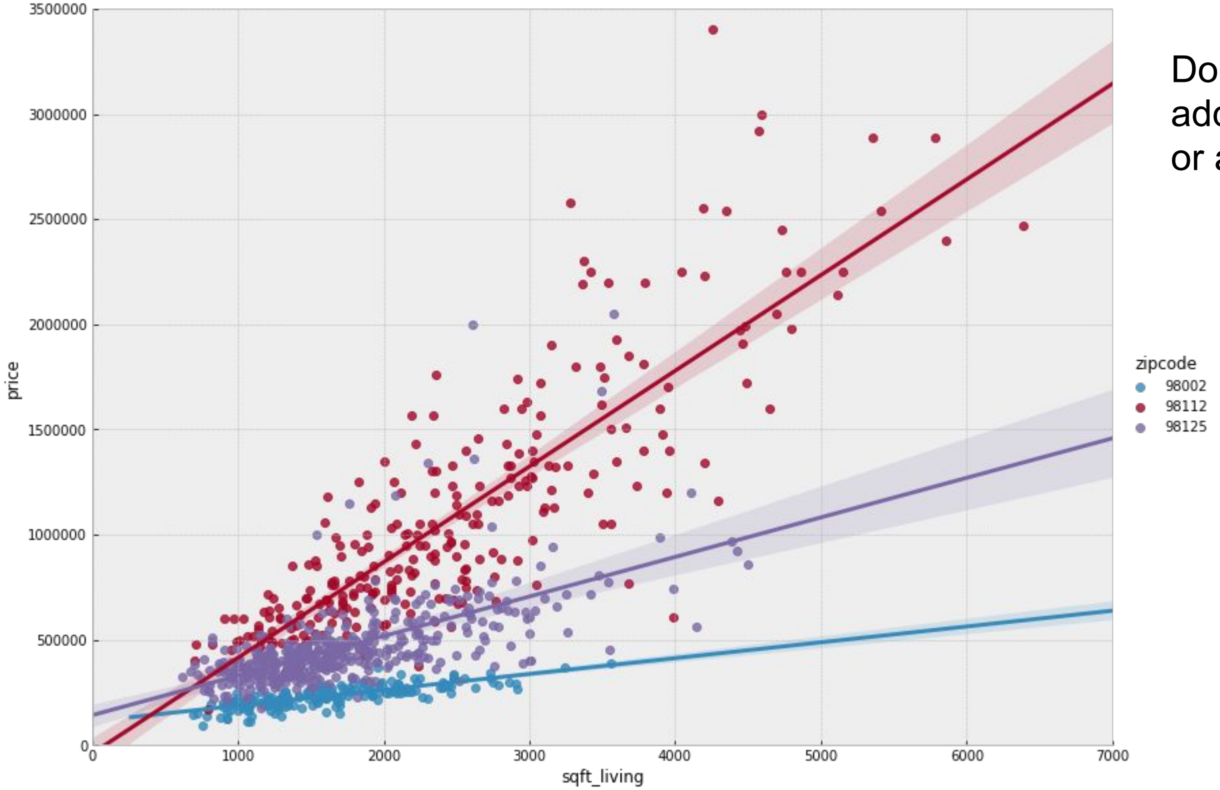


Unadjusted



Adjusted

Possible Variance in Effect by Zipcode



Do we have enough data to address this with interaction terms or a multi-level regression?

A Robust Model (With Zipcode Dummies)

Dep. Variable:	log_price	**R-squared:**	0.839
Model:	OLS	**Adj. R-squared:**	0.839
Method:	Least Squares	**F-statistic:**	1478.
Date:	Tue, 07 May 2019	**Prob (F-statistic):**	0.00
Time:	10:22:54	**Log-Likelihood:**	2945.5
No. Observations:	21597	**AIC:**	-5737.
Df Residuals:	21520	**BIC:**	-5122.
Df Model:	76		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.6810	0.018	665.965	0.000	11.647	11.715
bedrooms	-0.0233	0.002	-11.356	0.000	-0.027	-0.019
bathrooms	0.0393	0.003	13.258	0.000	0.034	0.045
sqft_living	0.0003	2.71e-06	119.616	0.000	0.000	0.000
flag_2015	0.0498	0.003	16.121	0.000	0.044	0.056
reno_flag	0.0913	0.014	6.557	0.000	0.064	0.119
waterfront	0.6736	0.018	37.265	0.000	0.638	0.709
grade_dummy	0.1831	0.013	13.797	0.000	0.157	0.209
zip_98002	-0.0541	0.019	-2.894	0.004	-0.091	-0.017
zip_98003	0.0314	0.017	1.865	0.062	-0.002	0.064
zip_98004	1.1718	0.016	71.533	0.000	1.140	1.204

- Dummy variables (all 69 of them…) are all statistically significant
- Test/train split suggests that the model performs very well with minimal overfitting:
  - For naive model
    - Train error: 0.1185
    - Test error: 0.1226
  - For model with zipcode dummies:
    - Train error: 0.0402
    - Test error: 0.0381

Adding Interaction Terms

Dep. Variable:	log_price	**R-squared:**	0.863
Model:	OLS	**Adj. R-squared:**	0.862
Method:	Least Squares	**F-statistic:**	930.6
Date:	Tue, 07 May 2019	**Prob (F-statistic):**	0.00
Time:	16:32:59	**Log-Likelihood:**	4660.1
No. Observations:	21597	**AIC:**	-9028.
Df Residuals:	21451	**BIC:**	-7863.
Df Model:	145		
Covariance Type:	nonrobust		

- Interaction terms for sqft_living offer marginal improvement in predictive power
- Test/train split suggests that we continue to avoid over-fit:
    - Train error: 0.0381
    - Test error: 0.03809

Do We Have Sufficient Data for a Multi-Level Reg.?

- No.
- Replaced each of my key metrics (sqft, beds, baths, recently renovated flag and grade) with a set of interaction terms with the dummies for zipcodes
- To illustrate high risk of over fit, used a loop to run test/train splits multiple times
- High risk for over fit depending on test/train split

train: 0.03751 test: 0.03947
train: 0.03742 test: 0.03985
train: 0.03765 test: 0.03873
train: 0.03727 test: 0.04048
train: 0.0369 test: 0.042
train: 0.03758 test: 0.03942
train: 0.03745 test: 0.03973
train: 0.03781 test: 5877896237.35734
train: 0.03752 test: 376938786.34866
train: 0.03749 test: 0.0397
train: 0.03747 test: 0.03959
train: 0.03748 test: 0.03954
train: 0.03704 test: 0.04148
train: 0.03708 test: 0.04138
train: 0.03666 test: 348532781857.2288
train: 0.0382 test: 29492675.57095
train: 0.03703 test: 0.04141
train: 0.03782 test: 0.0384
train: 0.03667 test: 267494933561.2978
train: 0.03737 test: 3995021625.19367