Mod 1 Project – King County Housing Data

Thomas O'Gara

Part time data science

Instructor: Victor Geislinger

Problem Statement:

- How do we go about predicting home values in King County?

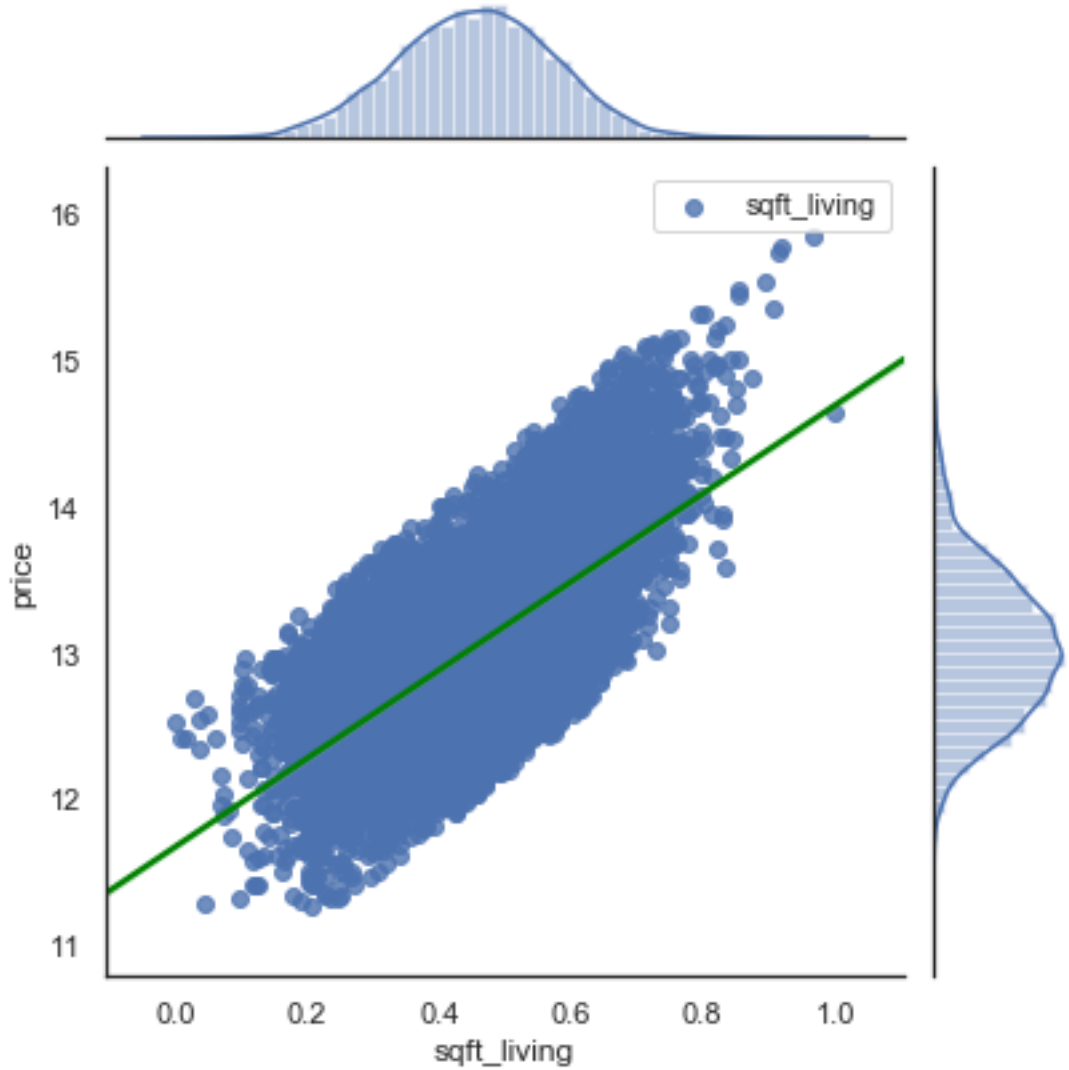
Methodology – “OSEMN”

- **O**btained housing data on 21,000+ King County homes
- **S**crubbed the data for missing values, non-numbers, placeholder values
- **E**xploratory data analysis (EDA)
- **M**odeled the data using multiple linear regression
- I**N**terpreted the data to determine the most important factors in valuing a home in King County

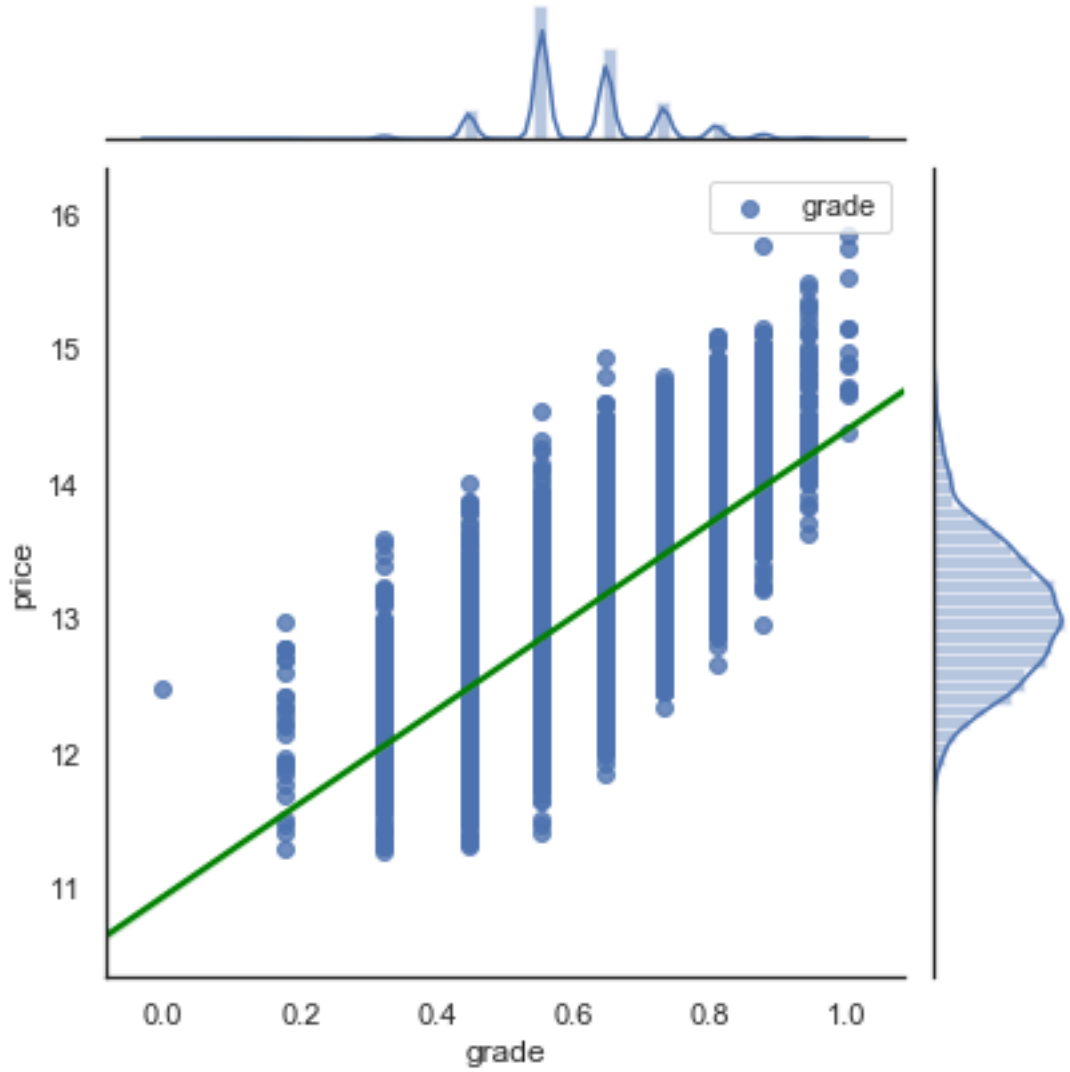
Most important variables to predict home prices

1. Square footage of the home
2. Grade/build quality
3. Square footage of the lot
4. Square footage of neighbor’s homes
5. Location/neighborhood

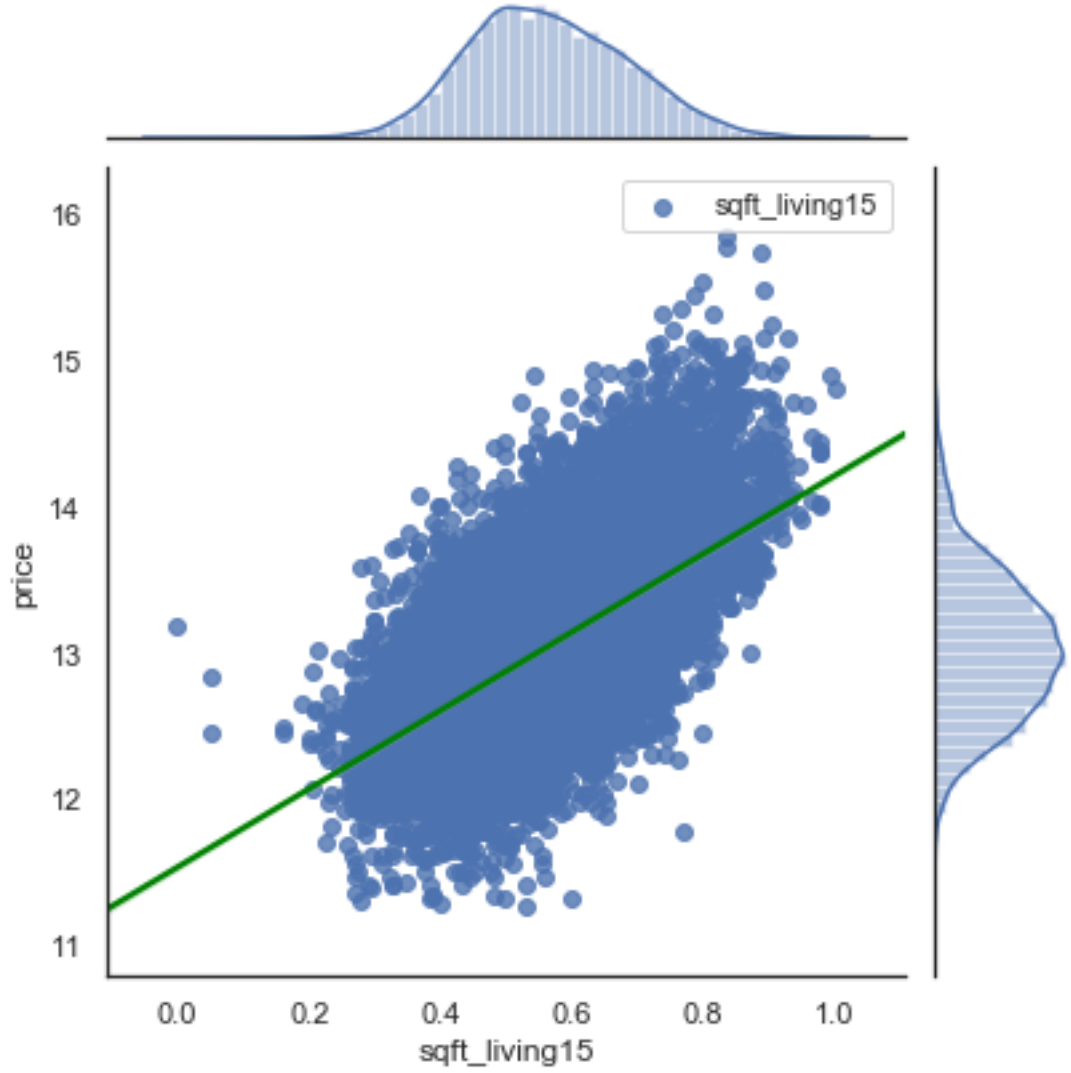
Square footage of the home



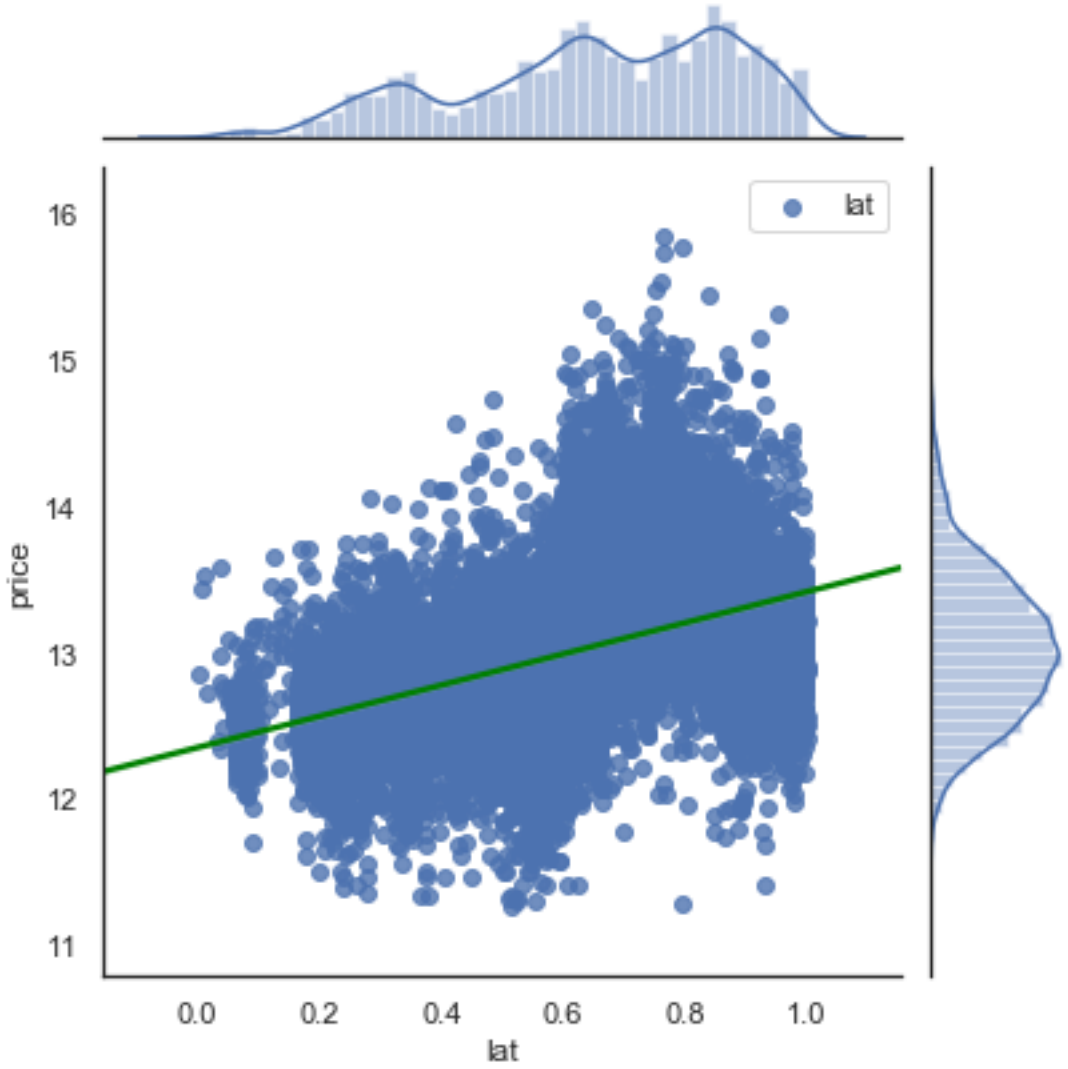
Grade/build quality

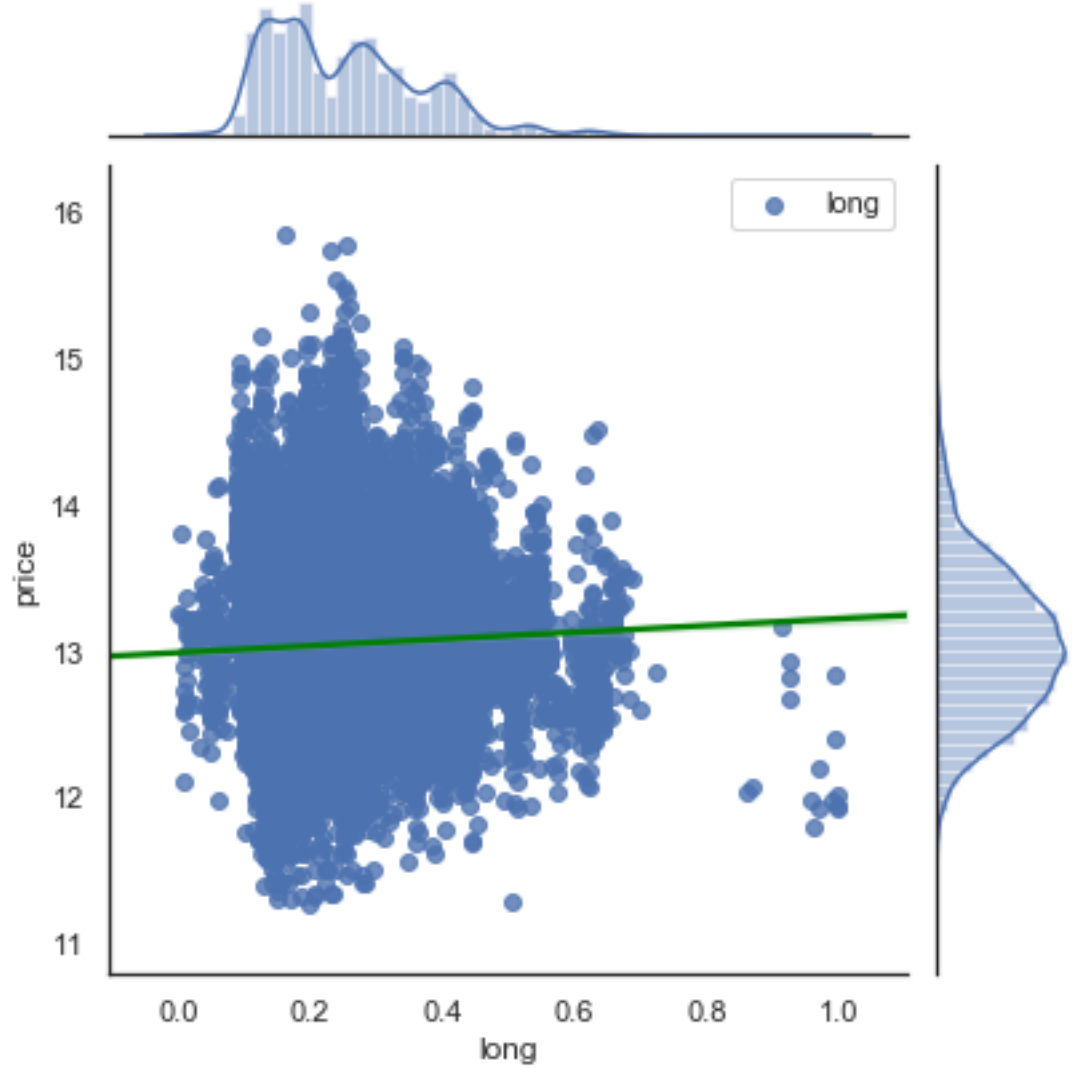


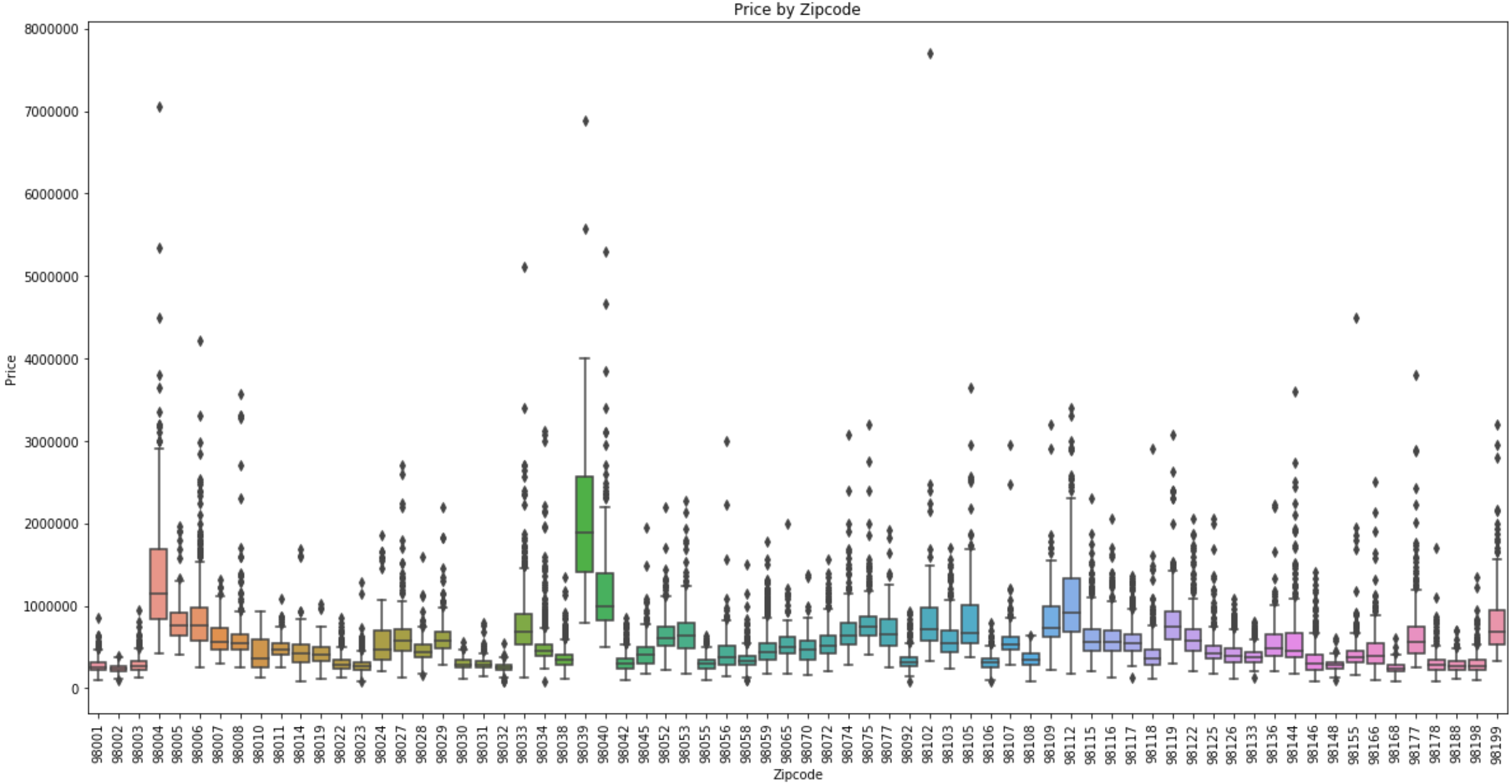
Square footage of neighbor’s homes



Location - where do you want to be?





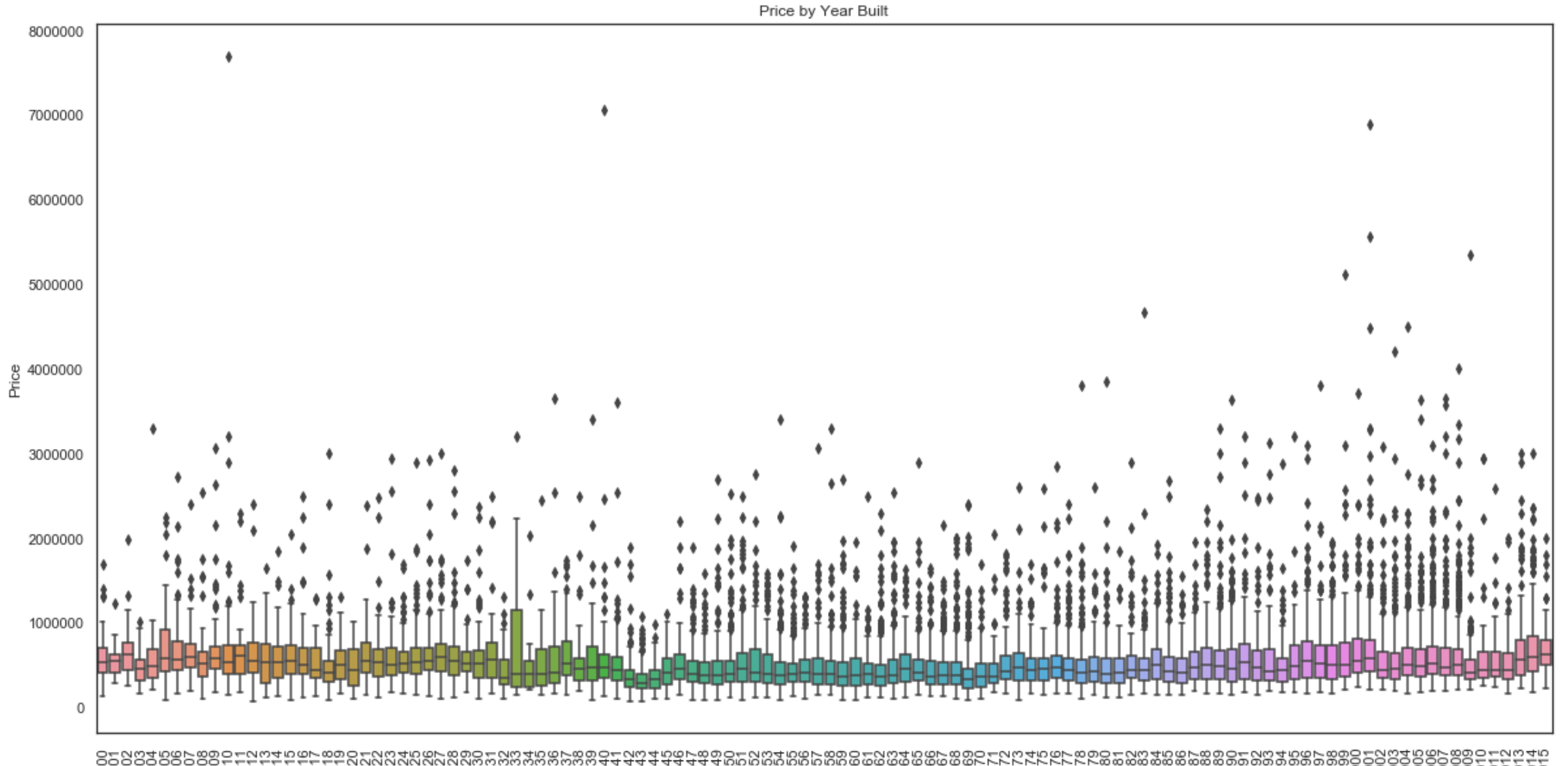

Price by Zipcode
Price
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
Zipcode
98001 98002 98003 98004 98005 98006 98007 98008 98010 98011 98014 98019 98022 98023 98024 98027 98028 98029 98030 98031 98032 98033 98034 98038 98039 98040 98042 98045 98052 98053 98055 98056 98058 98059 98065 98070 98072 98074 98075 98077 98092 98102 98103 98105 98106 98107 98108 98109 98112 98115 98116 98117 98118 98119 98122 98125 98126 98133 98136 98144 98146 98148 98155 98166 98168 98177 98178 98188 98198 98199

Location continued:

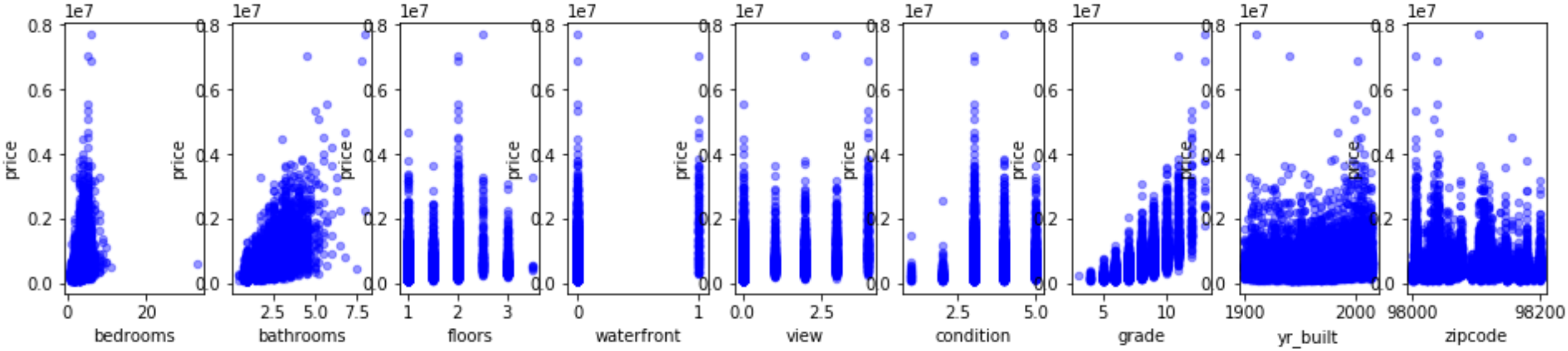
Most expensive zip codes:

- 98004: Medina
- 98039: Bellevue
- 98040: Mercer Island
- 98112: Northeast Seattle
- 98199: Northwest Seattle

Is a newer home more valuable?



What about other variables?



Model Overview

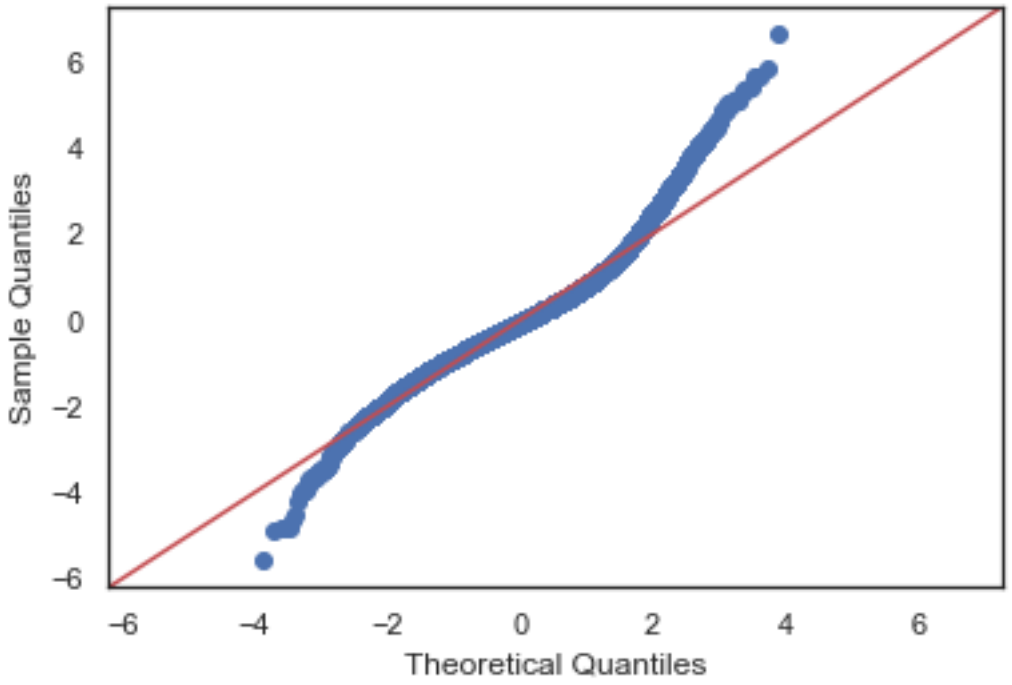
OLS Regression Results

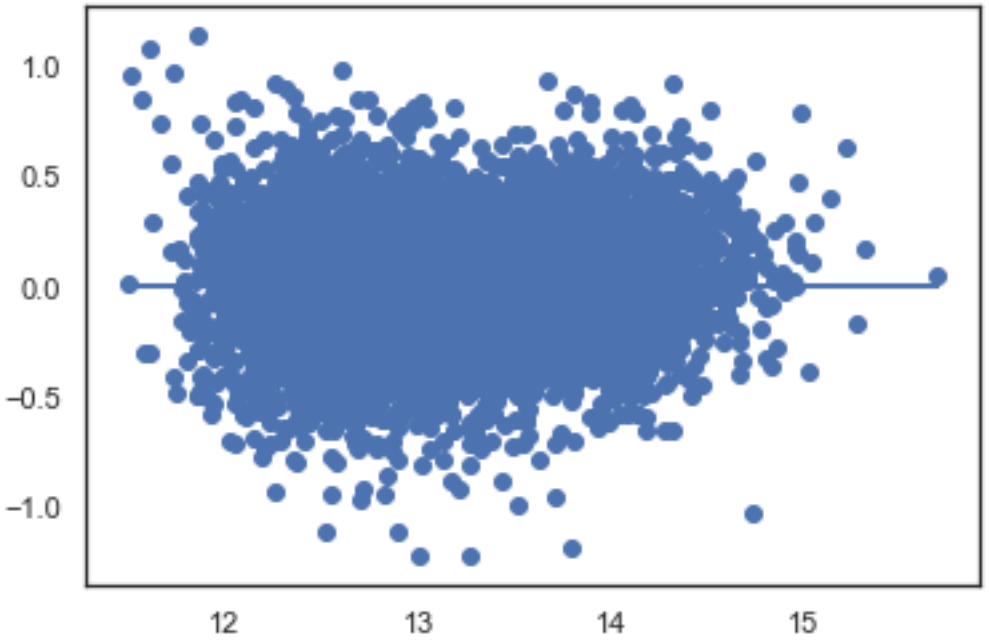
Dep. Variable:	price	**R-squared:**	0.878
Model:	OLS	**Adj. R-squared:**	0.878
Method:	Least Squares	**F-statistic:**	1646.
Date:	Wed, 12 Jun 2019	**Prob (F-statistic):**	0.00
Time:	10:38:44	**Log-Likelihood:**	5811.6
No. Observations:	21082	**AIC:**	-1.144e+04
Df Residuals:	20989	**BIC:**	-1.070e+04
Df Model:	92		
Covariance Type:	nonrobust		

Omnibus:	1302.533	**Durbin-Watson:**	2.001
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	5939.522
Skew:	-0.055	**Prob(JB):**	0.00
Kurtosis:	5.598	**Cond. No.**	281.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.9549	0.059	186.655	0.000	10.840	11.070
bedrooms	-0.2552	0.024	-10.684	0.000	-0.302	-0.208
bathrooms	0.1798	0.016	10.950	0.000	0.148	0.212
sqft_living	1.5118	0.027	55.100	0.000	1.458	1.566
sqft_lot	0.6256	0.030	20.897	0.000	0.567	0.684
grade	1.0027	0.022	46.034	0.000	0.960	1.045
sqft_basement	-0.0624	0.005	-13.359	0.000	-0.072	-0.053
lat	0.3064	0.046	6.703	0.000	0.217	0.396
long	-0.5190	0.064	-8.059	0.000	-0.645	-0.393
sqft_living15	0.4074	0.019	21.551	0.000	0.370	0.444
sqft_lot15	-0.0980	0.030	-3.290	0.001	-0.156	-0.040
floor[0]	0.0244	0.005	4.806	0.000	0.014	0.034
floor[1]	0.0085	0.004	2.014	0.044	0.000	0.017
floor[2]	0.0504	0.016	3.235	0.001	0.020	0.081
floor[3]	-0.0703	0.009	-7.446	0.000	-0.089	-0.052
floor[4]	-0.0362	0.070	-0.518	0.605	-0.173	0.101
view[0]	0.1188	0.011	11.285	0.000	0.098	0.139
view[1]	0.1151	0.006	17.798	0.000	0.102	0.128
view[2]	0.2029	0.009	23.130	0.000	0.186	0.220
view[3]	0.4891	0.011	44.462	0.000	0.468	0.511
condition_2	0.1269	0.038	3.360	0.001	0.053	0.201
condition_3	0.2417	0.035	6.880	0.000	0.173	0.311
condition_4	0.2755	0.035	7.838	0.000	0.207	0.344
condition_5	0.3416	0.035	9.658	0.000	0.272	0.411

Normality and Homoscedasticity assumptions

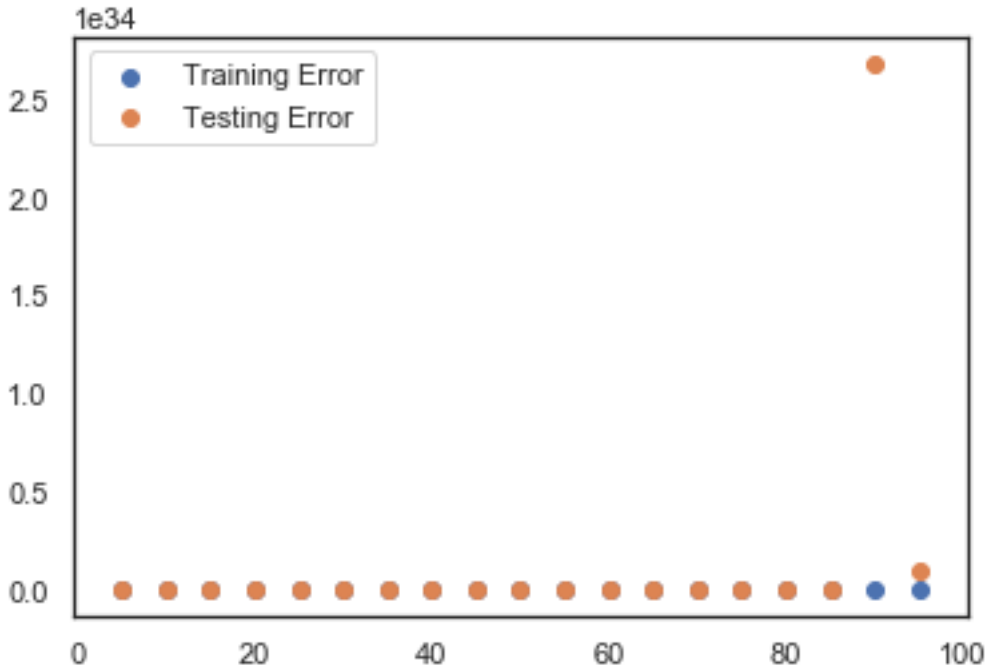




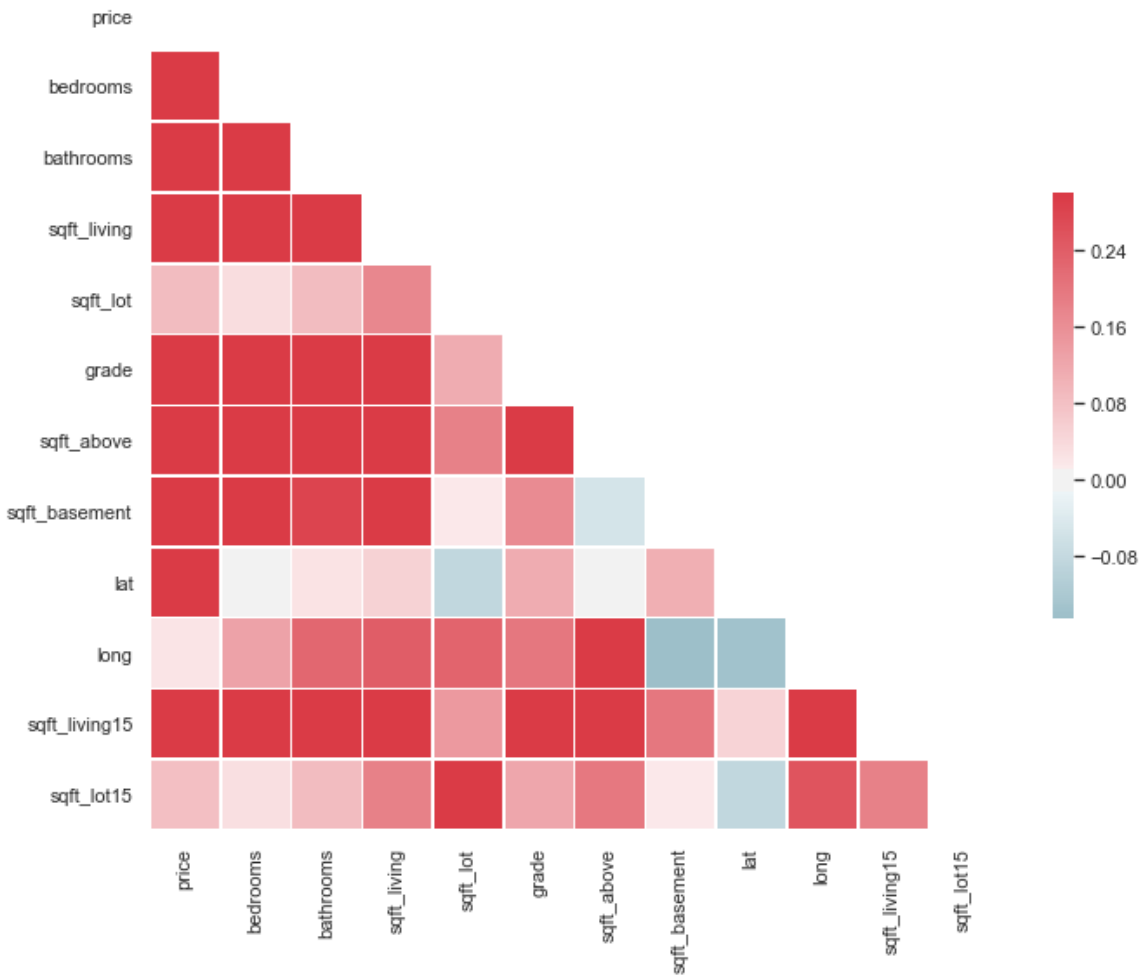
Is this model a good predictor of future home prices?

```
mse_train = np.sum((y_train-y_hat_train)**2)/len(y_train)
mse_test =np.sum((y_test-y_hat_test)**2)/len(y_test)
print('Train Mean Squarred Error:', mse_train)
print('Test Mean Squarred Error:', mse_test)
```

```
Train Mean Squarred Error: 0.034076607061633214
Test Mean Squarred Error: 0.032997092011984586
```



What are potential problems with this model?




```
#Checking for Multicollinearity using variance inflation factor.
#this accounts for multicollinearity with a relation of 3 or more variables v
from statsmodels.stats.outliers_influence import variance_inflation_factor

X = subset.drop(['price'], axis=1)
vif = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
list(zip(X.columns, vif))
```

```
[('bedrooms', 27.43936322112261),
 ('bathrooms', 28.899237707052393),
 ('sqft_living', 102.81077627409047),
 ('sqft_lot', 22.845588831113602),
 ('grade', 112.24690739331463),
 ('sqft_basement', 3.1025613421357163),
 ('lat', 273.2524962797792),
 ('long', 193.36742543893328),
 ('sqft_living15', 72.38936688954992),
 ('floor', 1.4059878707820597),
 ('floor', 4.266375251366034),
 ('floor', 1.1034480605684647),
 ('floor', 1.5573844297385897),
 ('floor', 1.0107382278913626),
 ('view', 1.0703526808346429),
 ('view', 1.1481992787171456),
 ('view', 1.1259385931881711),
 ('view', 1.1098055988207716),
 ('condition_2', 4.837279868308839),
 ('condition_3', 321.7072792016948),
 ('condition_4', 129.5277054533042),
 ('condition_5', 39.66082642763849),
```

Business recommendation for home values

- Square footage of living space is key
- High grade building
  - Doesn’t need to be a new build – just needs to be high quality workmanship
- Pick a neighborhood with large homes nearby
  - Better to be the worst house in a good neighborhood
- Northwest area of King County is best
  - Focus on top 5 zip codes
  - Close to the water

Thank you!

Questions?