King County Real Estate Analysis

Our findings, our narrative, our *future*

Our questions:

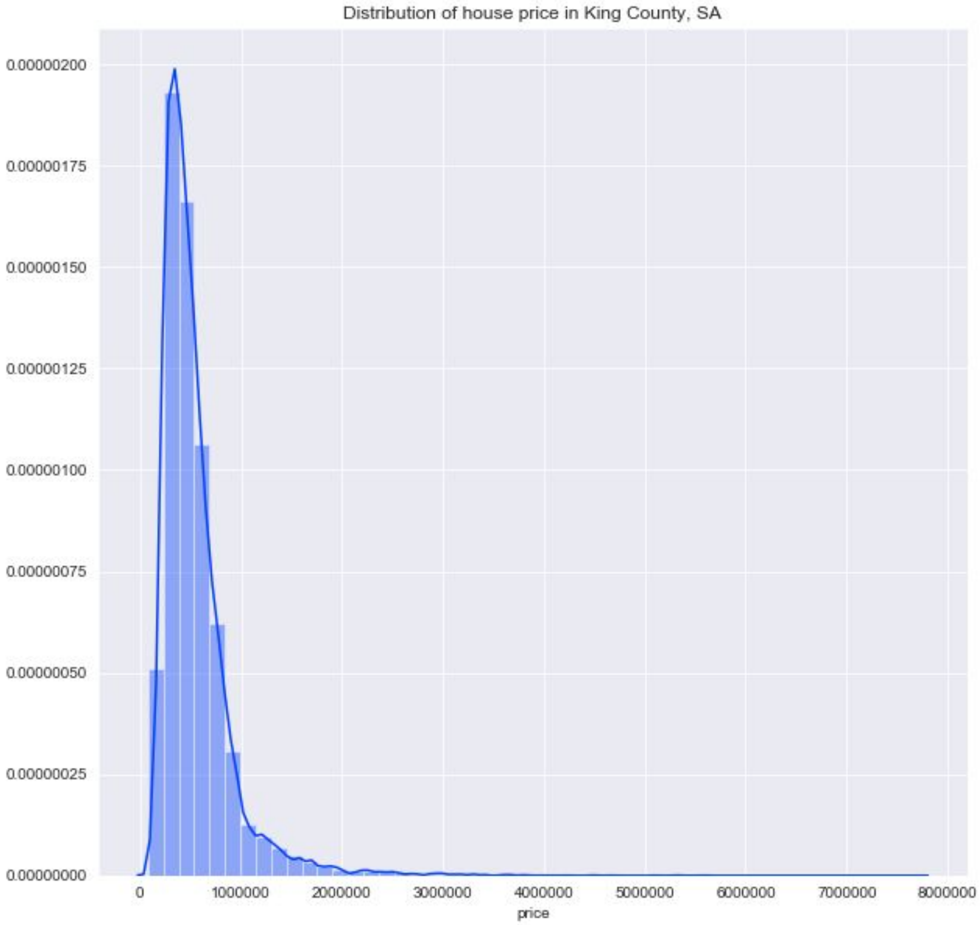
1. How accurate a predictor is the amount of square feet of living space?
2. Do houses more recently modified have higher prices?
3. Are there any clear geographical trends in price?

Stakeholder Overview

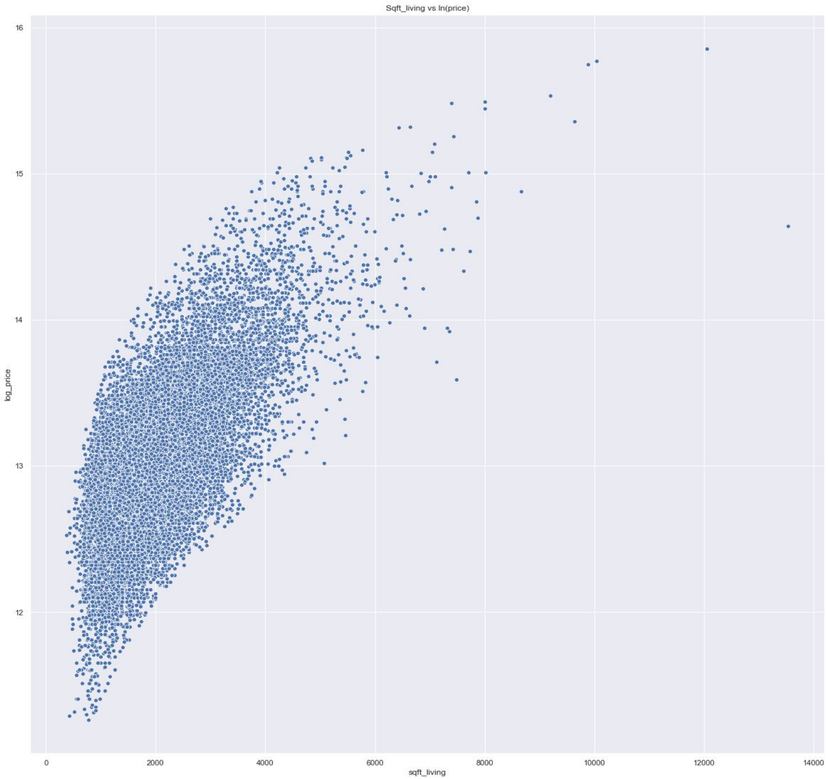
Real Estate Companies savills	Housing Development Firms Taylor Wimpey
Focus on maximising: Sales, Profits	Focussed on maximising sale price after spending on renovations or extensions
Focussed on efficiency in: Costs related to marketing and sales.	Minimising cost when investing in house developemnts

Adjusting the data

- Non-normal distribution
- Data heavily skewed by outliers
- Took the logarithm of our price for final models



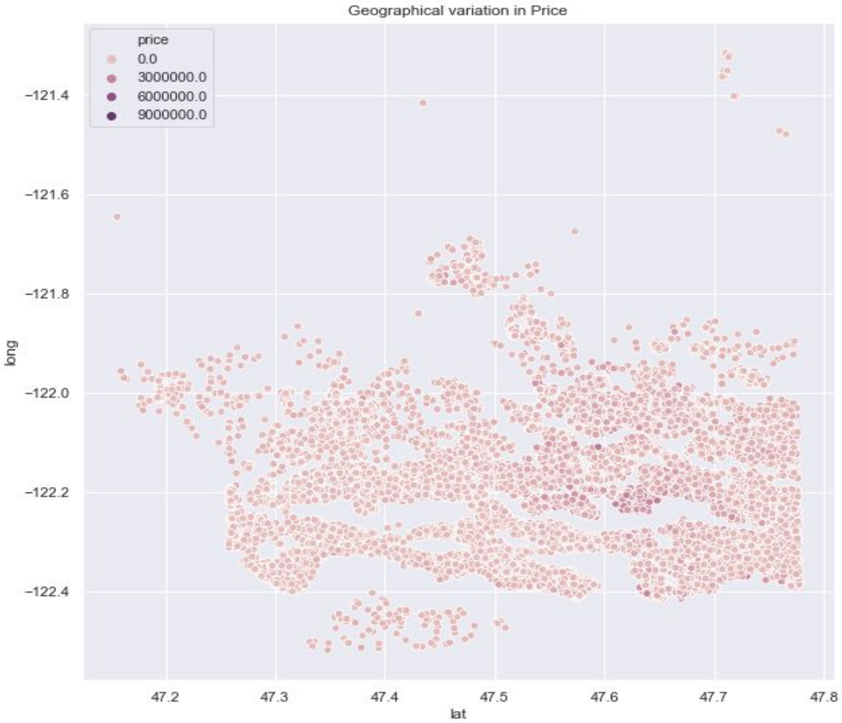
Size: An Efficient Price Predictor?



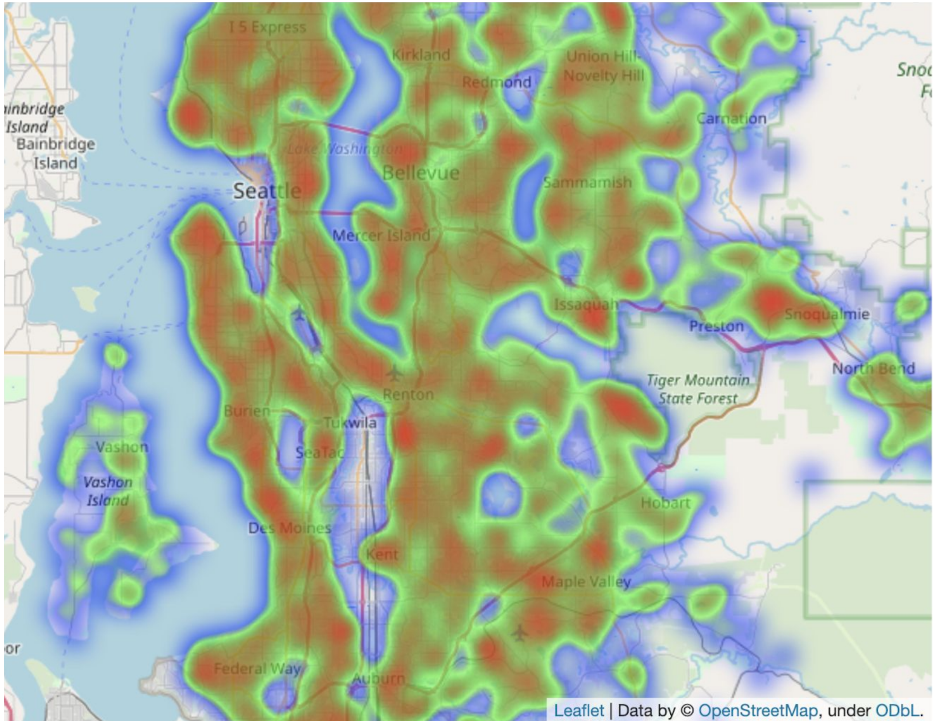
Living Space of a property Vs Price

- Initial data exploration showed a high correlation between living space and price.
- A scatter graph helped to visualize the strength of this.
- Further regression analysis showed it to be the strongest predictor of house price, out of all variables included in our model.

Location, location, location …

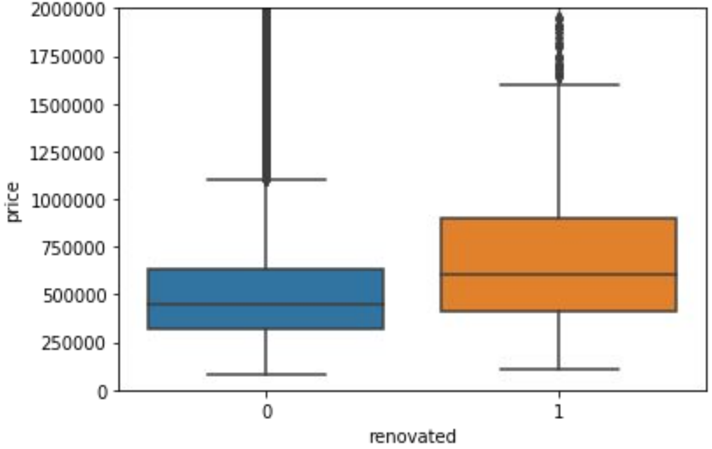


- We spotted clusters of high prices around specific locational points.
- Locational price data impacts both of our stakeholders:
  - Enabling real estate companies to appropriately price their properties
  - Calculation for a housing development profit margin, based on maximum cost per square foot of building houses or extensions
- This led us into further analysis and mapping.



- Our initial visualisation suggested a cluster of high price points around a central area
- Mapping on price per square foot subsets the data, while removing variation based on total property size.
- This sets a benchmark ceiling of spending on cost per square foot in various areas, for a housing development firm to still profit after a build.
- However, more in depth mapping showed there was actually multiple clusters of high priced areas across the dataset.
- Key:
  - Blue: $231.5 / sqft or less
  - Green: $231.5 - $270 /sqft
  - Red: : $405/sqft +

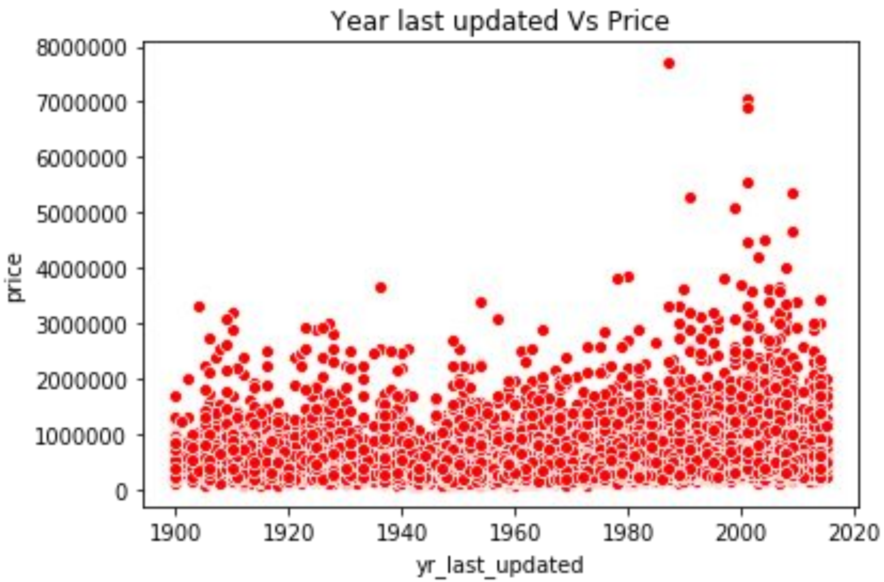
To renovate or not to renovate



Does Renovation affect house price?

- Initial data exploration and visualization showed that the subset of houses that had been renovated in the past were on average a higher price at point of sale.
- This led us to include it in our predictive model, which proved less useful
- In conclusion: house renovations on average increase the price, however did not yield useful predictive qualities at a later stage/

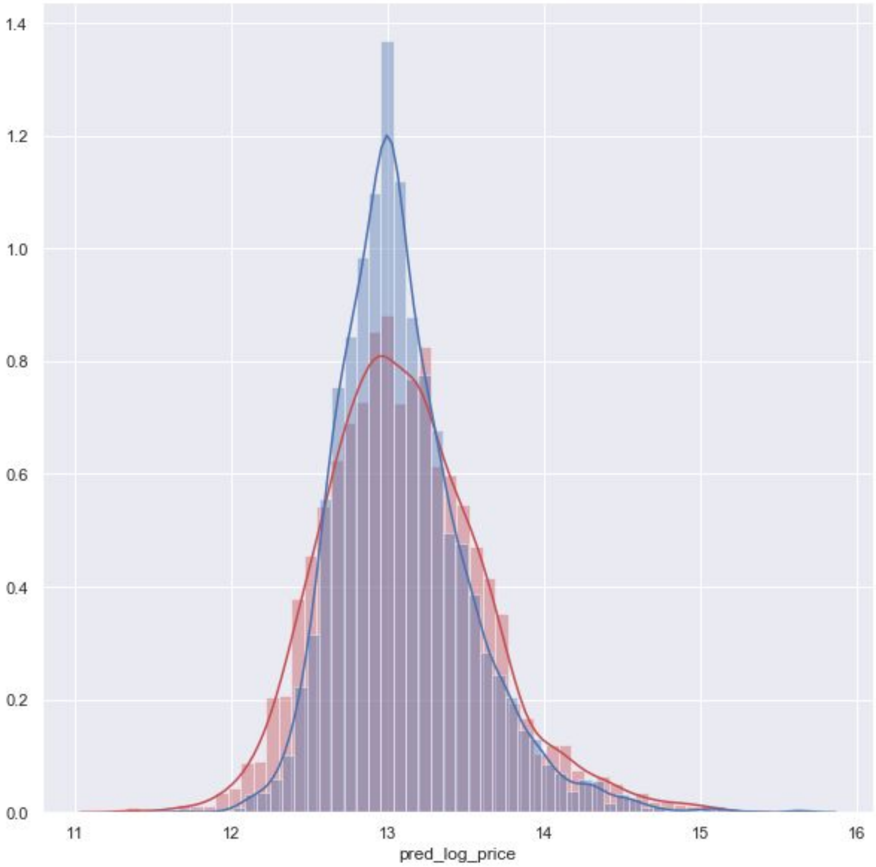
Further exploration around renovations



- Our previous slide led us to further analysis around renovations, and building age.
- We created a new variable to test our hypothesis that buildings that had been recently updated (either through initial build or renovation) would be a higher price.
- If this was the case, it would be useful to include in our predictive model.
- While the scatter plot showed a potential positive correlation, it was very weak.

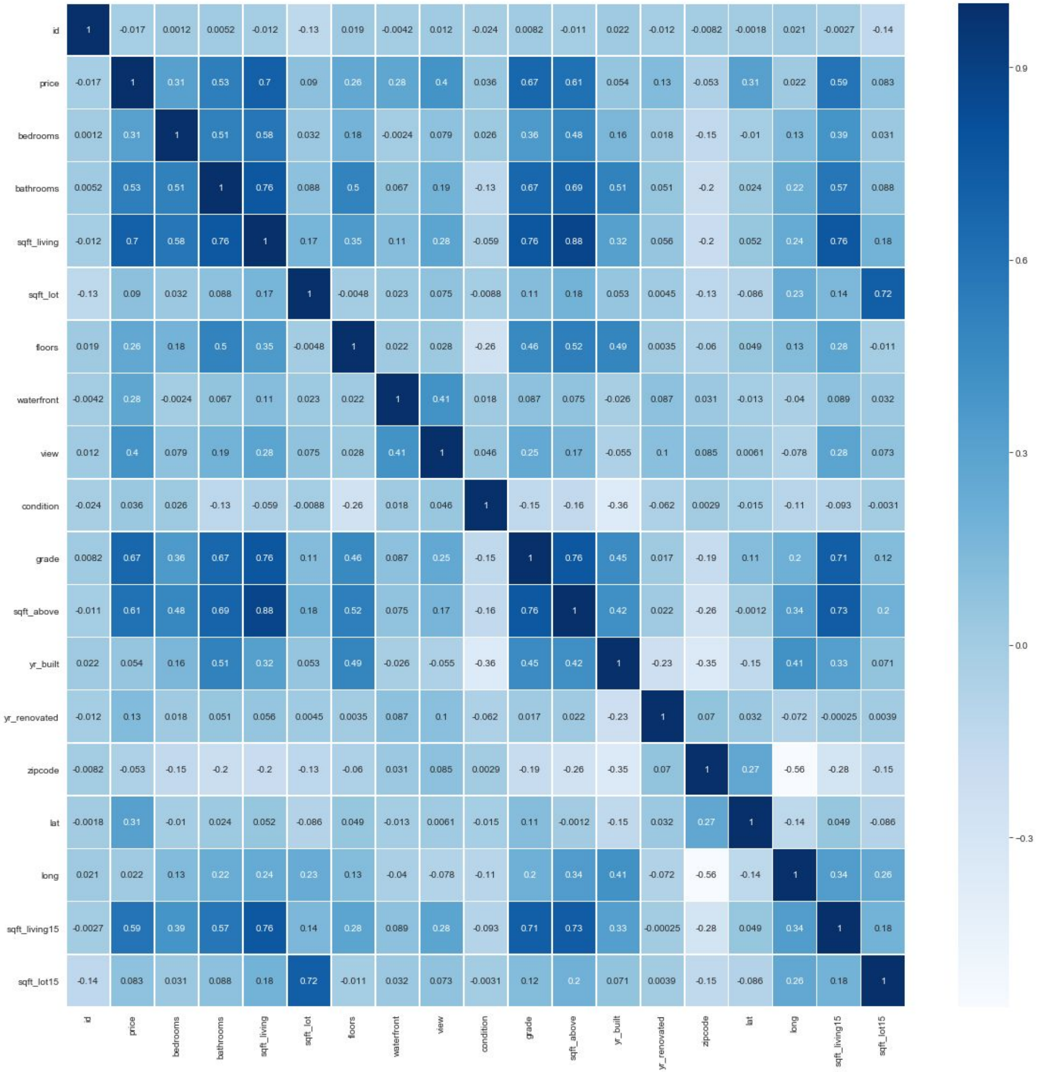
Price Prediction Model

1. Both absolute price and cost per square foot are dependent on location
   a. Average $264/sqft across whole dataset
   b. This relationship is not focussed around a central point
2. Total square foot of a property is the most effective predictor for house price
3. Renovation has some effect on average property price, but limited predictive capabilities
4. We are quite confident in our results, however would have achieved similar using purely square footage - other variables have a limited impact.
5. Separate model for **housing development**

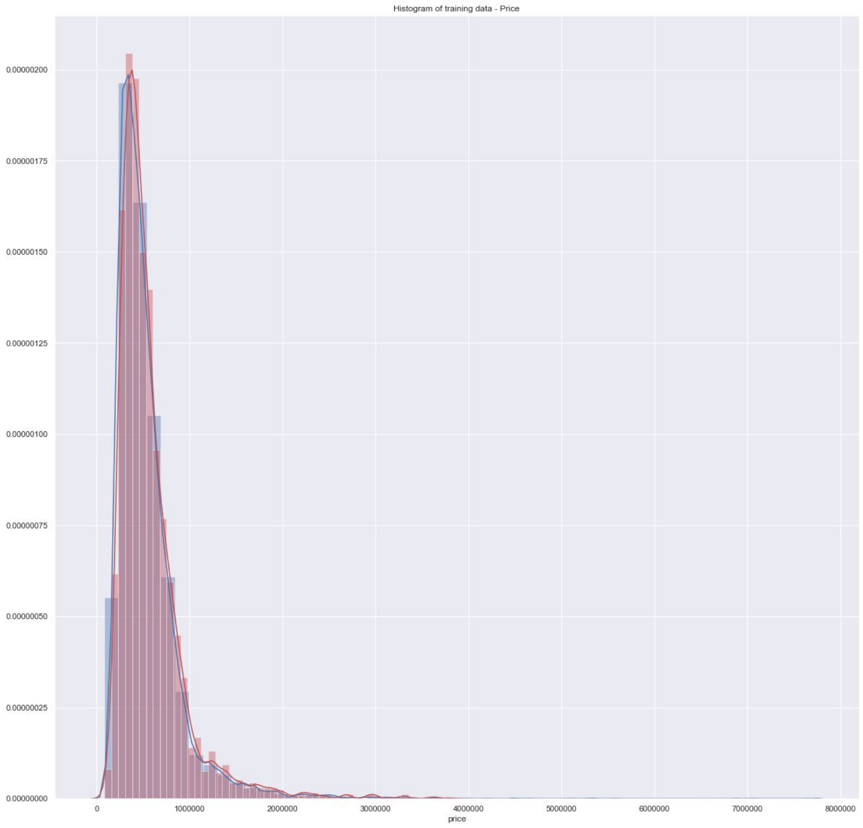


Thank you

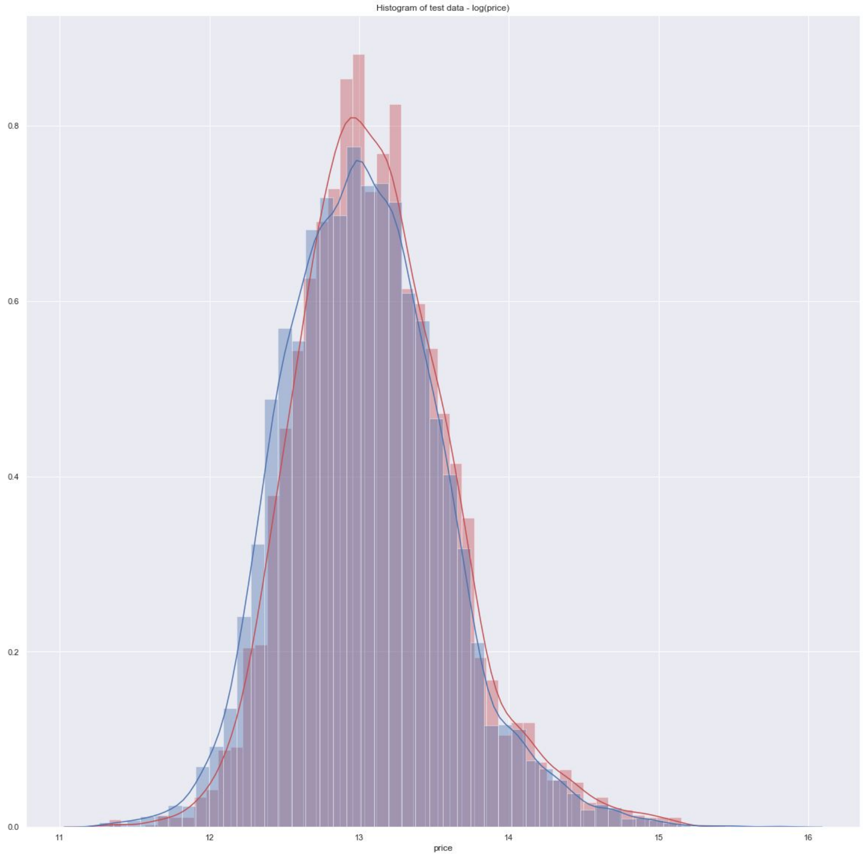
Questions, please



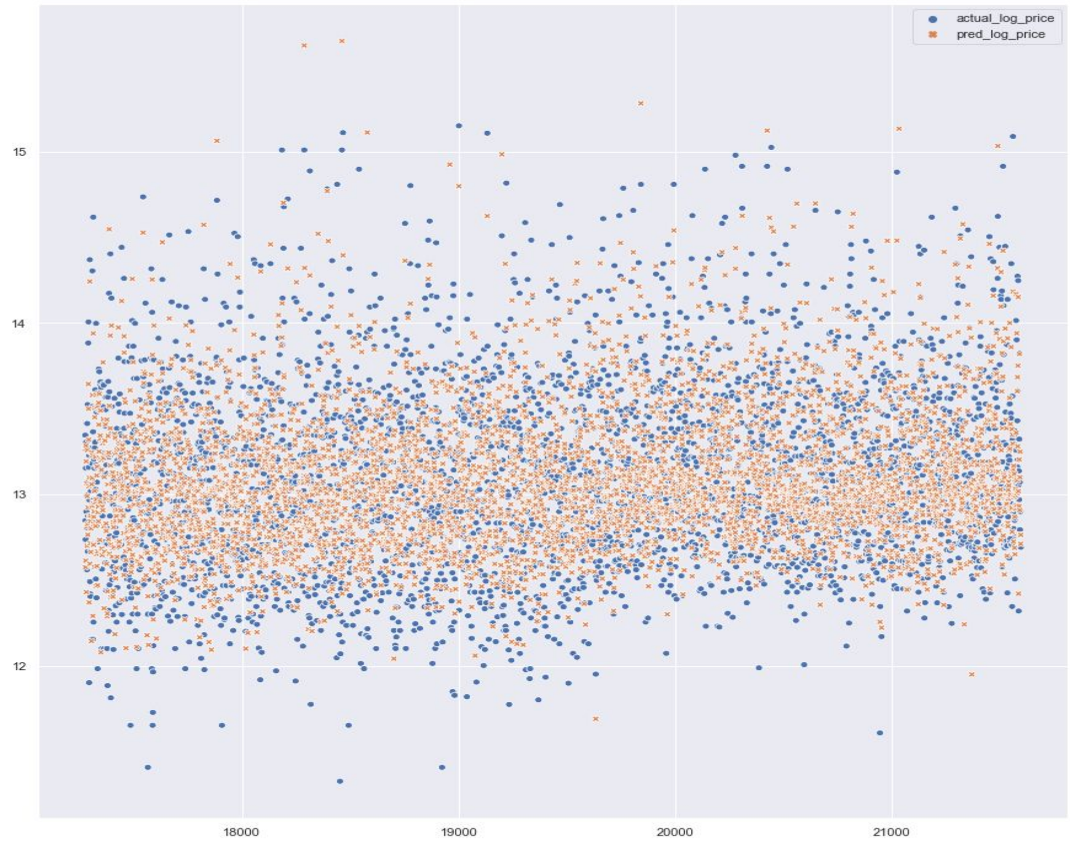
	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
id	1	-0.017	0.0012	0.0052	-0.012	-0.13	0.019	-0.0042	0.012	-0.024	0.0082	-0.011	0.022	-0.012	-0.0082	-0.0018	0.021	-0.0027	-0.14
price	-0.017	1	0.31	0.53	0.7	0.09	0.26	0.28	0.4	0.036	0.67	0.61	0.054	0.13	-0.053	0.31	0.022	0.59	0.083
bedrooms	0.0012	0.31	1	0.51	0.58	0.032	0.18	-0.0024	0.079	0.026	0.36	0.48	0.16	0.018	-0.15	-0.01	0.13	0.39	0.031
bathrooms	0.0052	0.53	0.51	1	0.76	0.088	0.5	0.067	0.19	-0.13	0.67	0.69	0.51	0.051	-0.2	0.024	0.22	0.57	0.088
sqft_living	-0.012	0.7	0.58	0.76	1	0.17	0.35	0.11	0.28	-0.059	0.76	0.88	0.32	0.056	-0.2	0.052	0.24	0.76	0.18
sqft_lot	-0.13	0.09	0.032	0.088	0.17	1	-0.0048	0.023	0.075	-0.0088	0.11	0.18	0.053	0.0045	-0.13	-0.086	0.23	0.14	0.72
floors	0.019	0.26	0.18	0.5	0.35	-0.0048	1	0.022	0.028	-0.26	0.46	0.52	0.49	0.0035	-0.06	0.049	0.13	0.28	-0.011
waterfront	-0.0042	0.28	-0.0024	0.067	0.11	0.023	0.022	1	0.41	0.018	0.087	0.075	-0.026	0.087	0.031	-0.013	-0.04	0.089	0.032
view	0.012	0.4	0.079	0.19	0.28	0.075	0.028	0.41	1	0.046	0.25	0.17	-0.055	0.1	0.085	0.0061	-0.078	0.28	0.073
condition	-0.024	0.036	0.026	-0.13	-0.059	-0.0088	-0.26	0.018	0.046	1	-0.15	-0.16	-0.36	-0.062	0.0029	-0.015	-0.11	-0.093	-0.0031
grade	0.0082	0.67	0.36	0.67	0.76	0.11	0.46	0.087	0.25	-0.15	1	0.76	0.45	0.017	-0.19	0.11	0.2	0.71	0.12
sqft_above	-0.011	0.61	0.48	0.69	0.88	0.18	0.52	0.075	0.17	-0.16	0.76	1	0.42	0.022	-0.26	-0.0012	0.34	0.73	0.2
yr_built	0.022	0.054	0.16	0.51	0.32	0.053	0.49	-0.026	-0.055	-0.36	0.45	0.42	1	-0.23	-0.35	-0.15	0.41	0.33	0.071
yr_renovated	-0.012	0.13	0.018	0.051	0.056	0.0045	0.0035	0.087	0.1	-0.062	0.017	0.022	-0.23	1	0.07	0.032	-0.072	-0.00025	0.0039
zipcode	-0.0082	-0.053	-0.15	-0.2	-0.2	-0.13	-0.06	0.031	0.085	0.0029	-0.19	-0.26	-0.35	0.07	1	0.27	-0.56	-0.28	-0.15
lat	-0.0018	0.31	-0.01	0.024	0.052	-0.086	0.049	-0.013	0.0061	-0.015	0.11	-0.0012	-0.15	0.032	0.27	1	-0.14	0.049	-0.086
long	0.021	0.022	0.13	0.22	0.24	0.23	0.13	-0.04	-0.078	-0.11	0.2	0.34	0.41	-0.072	-0.56	-0.14	1	0.34	0.26
sqft_living15	-0.0027	0.59	0.39	0.57	0.76	0.14	0.28	0.089	0.28	-0.093	0.71	0.73	0.33	-0.00025	-0.28	0.049	0.34	1	0.18
sqft_lot15	-0.14	0.083	0.031	0.088	0.18	0.72	-0.011	0.032	0.073	-0.0031	0.12	0.2	0.071	0.0039	-0.15	-0.086	0.26	0.18	1

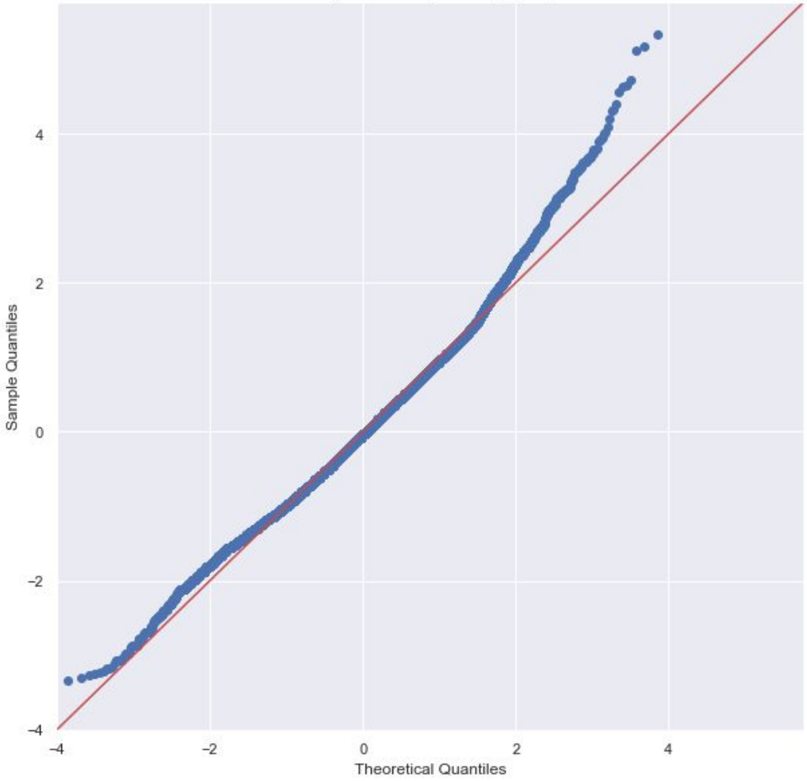


Histogram of training data - Price

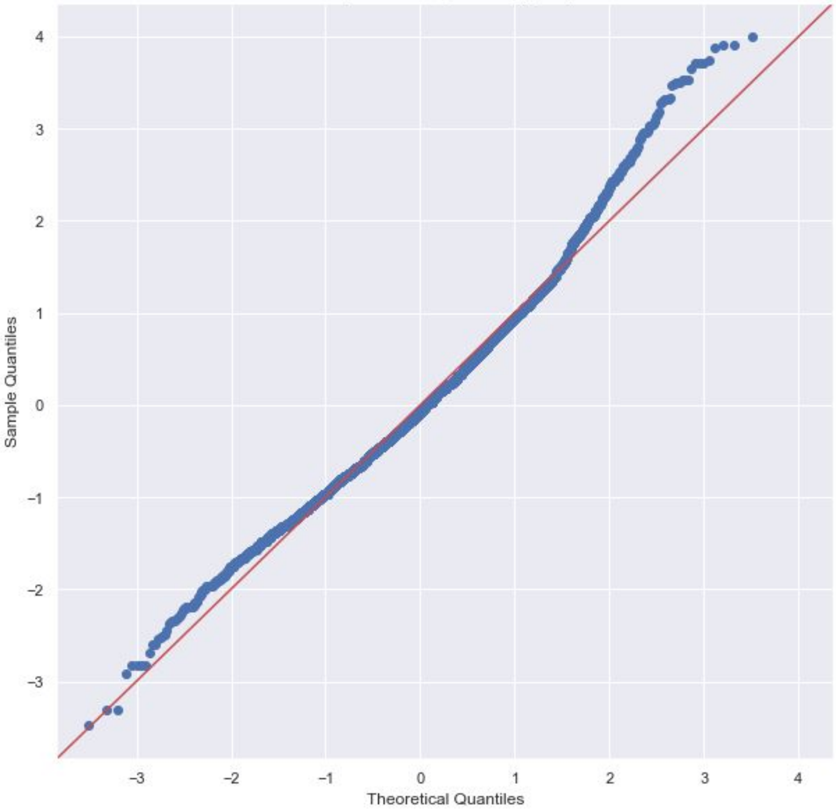


Histogram of test data - log(price)


actual_log_price
pred_log_price
15
14
13
12
18000
19000
20000
21000



QQ plot of training data - log(price)



QQ plot of testing data - log(price)

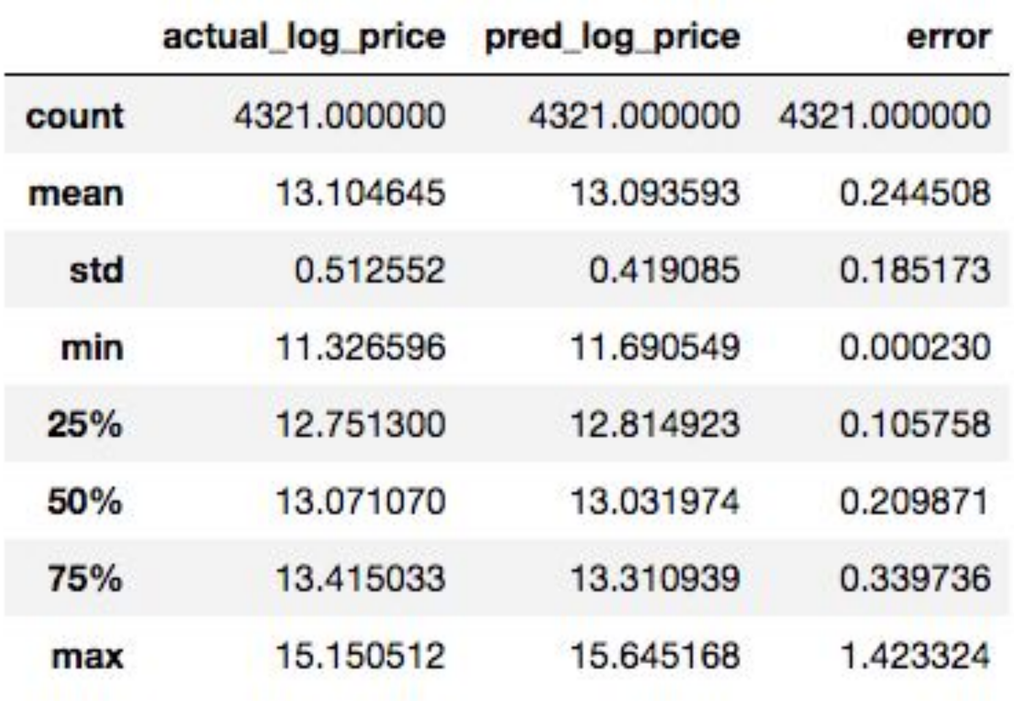
Multivariate model summary

Dep. Variable:	price	**R-squared:**	0.594
Model:	OLS	**Adj. R-squared:**	0.594
Method:	Least Squares	**F-statistic:**	1.262e+04
Date:	Wed, 23 Oct 2019	**Prob (F-statistic):**	0.00
Time:	09:33:30	**Log-Likelihood:**	-5734.0
No. Observations:	17276	**AIC:**	1.147e+04
Df Residuals:	17273	**BIC:**	1.150e+04
Df Model:	2		
Covariance Type:	nonrobust		

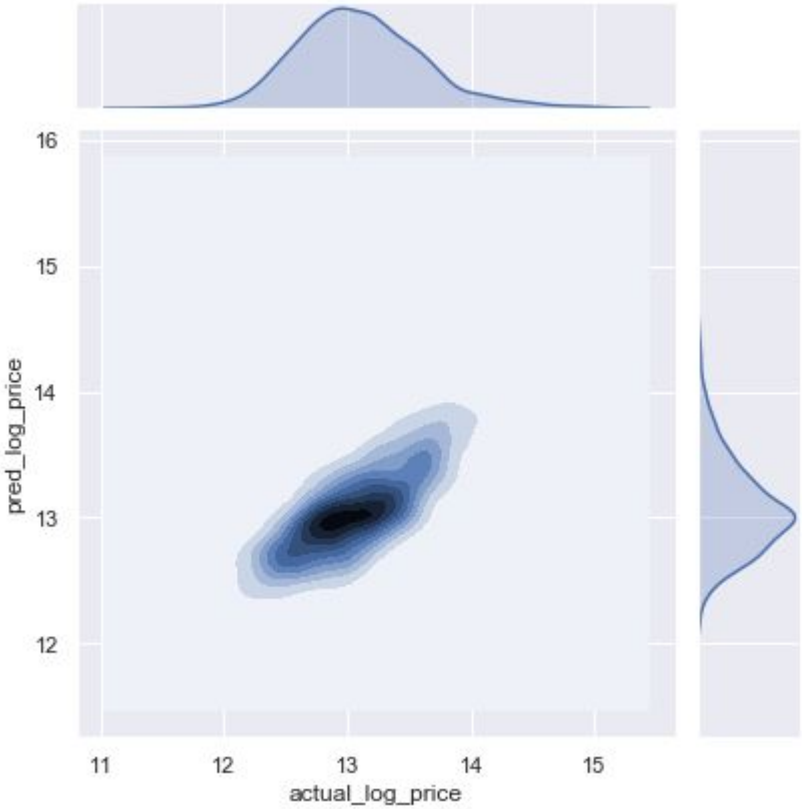
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	6.5085	0.046	141.241	0.000	6.418	6.599
log_sqft_living	0.9112	0.006	147.999	0.000	0.899	0.923
dist_highest_pricepersqft	-1.3670	0.018	-77.939	0.000	-1.401	-1.333

Omnibus:	48.480	**Durbin-Watson:**	2.001
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	49.760
Skew:	0.116	**Prob(JB):**	1.57e-11
Kurtosis:	3.124	**Cond. No.**	138.

Measures of dispersion for model on test data



	actual_log_price	pred_log_price	error
count	4321.000000	4321.000000	4321.000000
mean	13.104645	13.093593	0.244508
std	0.512552	0.419085	0.185173
min	11.326596	11.690549	0.000230
25%	12.751300	12.814923	0.105758
50%	13.071070	13.031974	0.209871
75%	13.415033	13.310939	0.339736
max	15.150512	15.645168	1.423324


16
15
14
13
12
pred_log_price
11
12
13
14
15
actual_log_price