



# King County Real Estate Analysis



Our findings, our narrative, our *future*



Our questions:

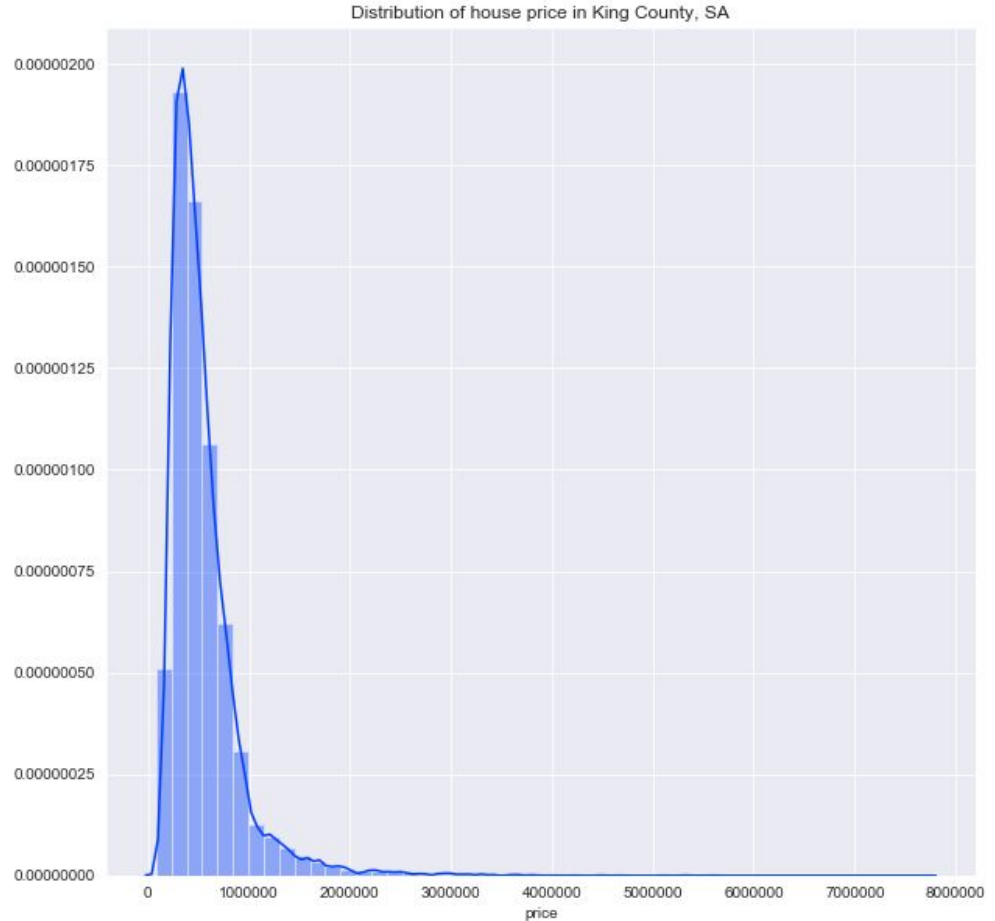
1. How accurate a predictor is the amount of square feet of living space?
2. Do more recently modified houses have higher prices?
3. Are there any clear geographical trends in price?

# Stakeholder Overview

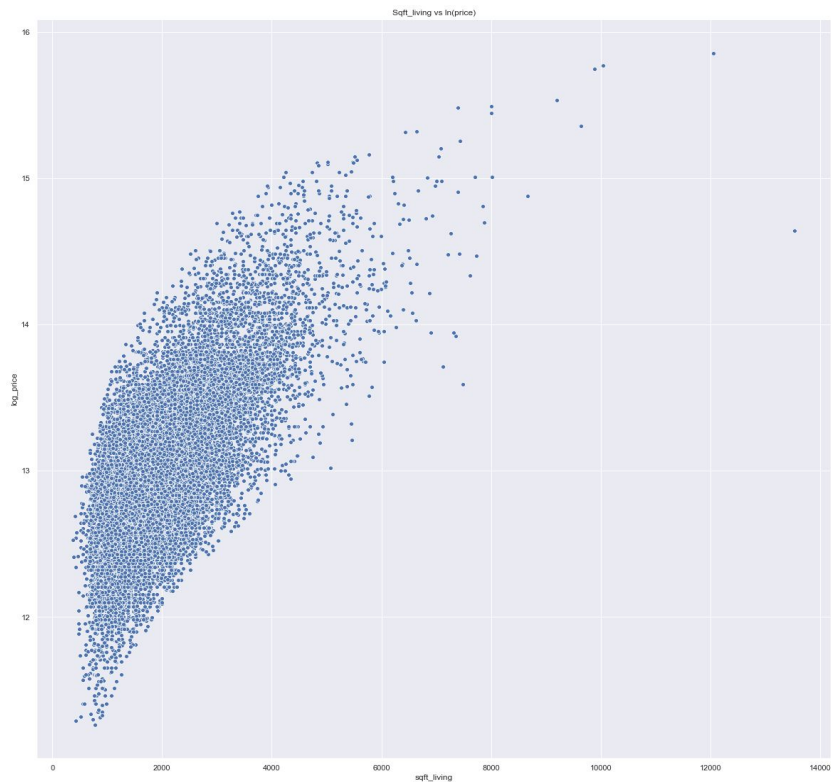
<h2>Real Estate Companies</h2> 	<h2>Housing Development Firms</h2> 
Focus on maximising: Sales, Profits	Focussed on maximising sale price after spending on renovations or extensions
Focussed on efficiency in: Costs related to marketing and sales	Minimising cost when investing in house developments

# Adjusting the data

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



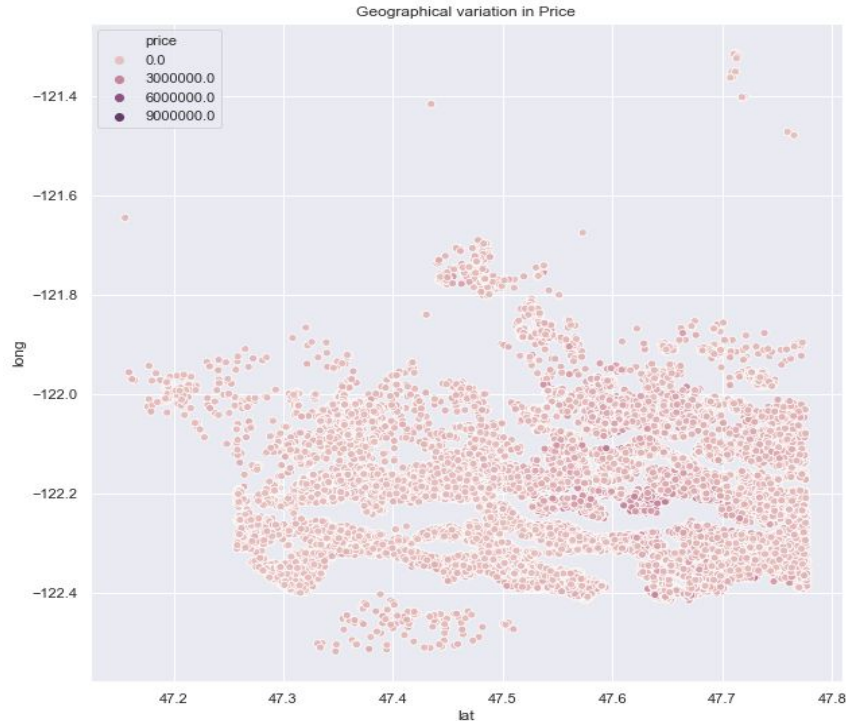
# Size: An Efficient Price Predictor?



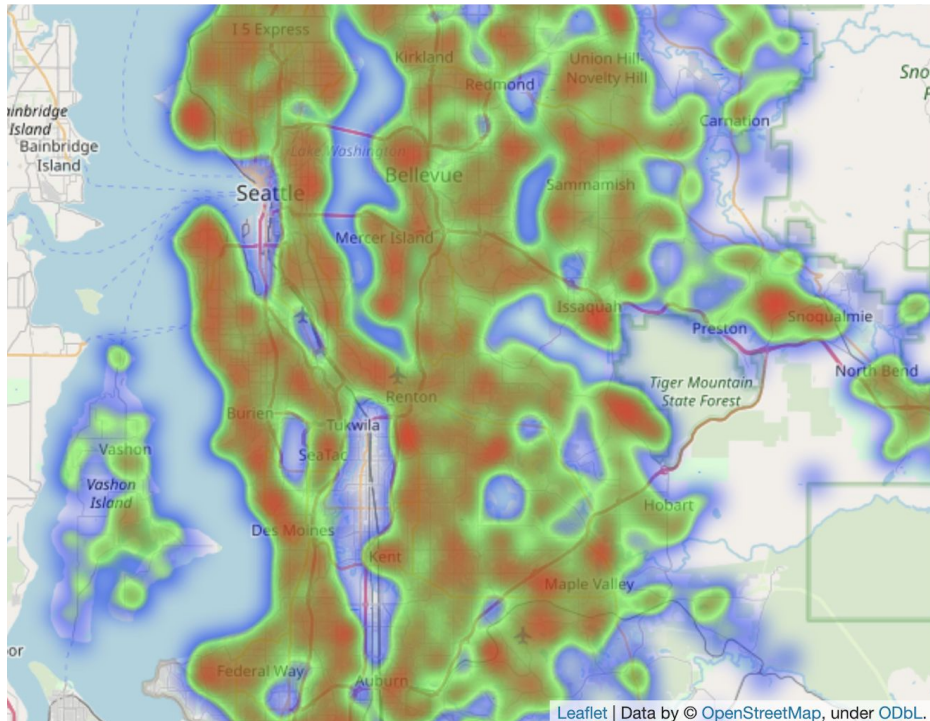
## Living Space of a property Vs Price

- Initial data exploration showed a high correlation between living space and price
- Non-normality of price distribution led us to take the log
- Further regression analysis showed it to be the strongest predictor of house price, out of all variables included in our model

# Location, location, location ...

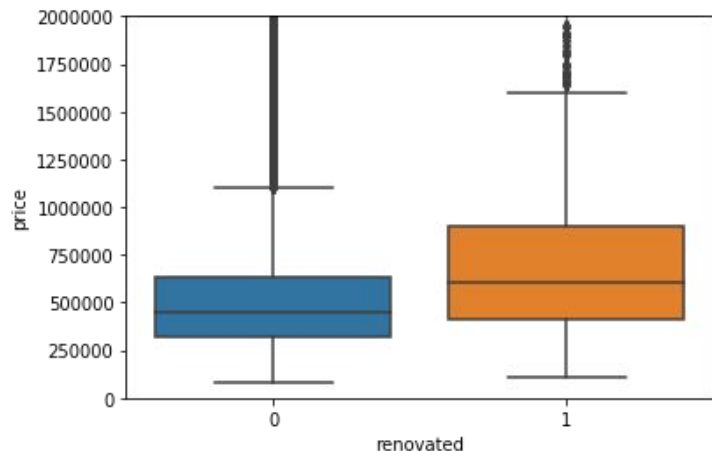


- We spotted clusters of high prices around specific locational points
- Locational price data impacts both of our stakeholders:
  - Enabling real estate companies to appropriately price their properties
  - Calculation for a housing development profit margin, based on maximum cost per square foot of 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 profit after a build
- However, more in depth mapping showed there were 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

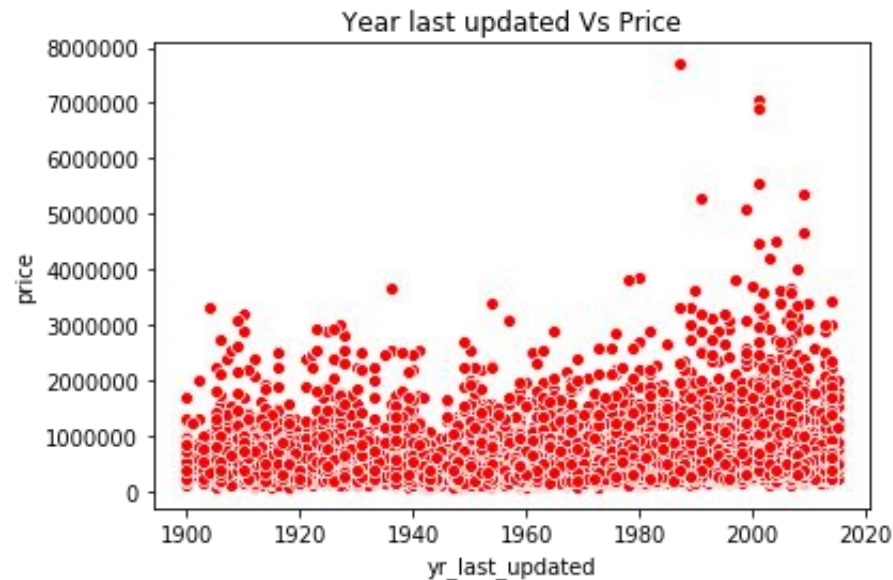


## Does Renovation affect house price?

- 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



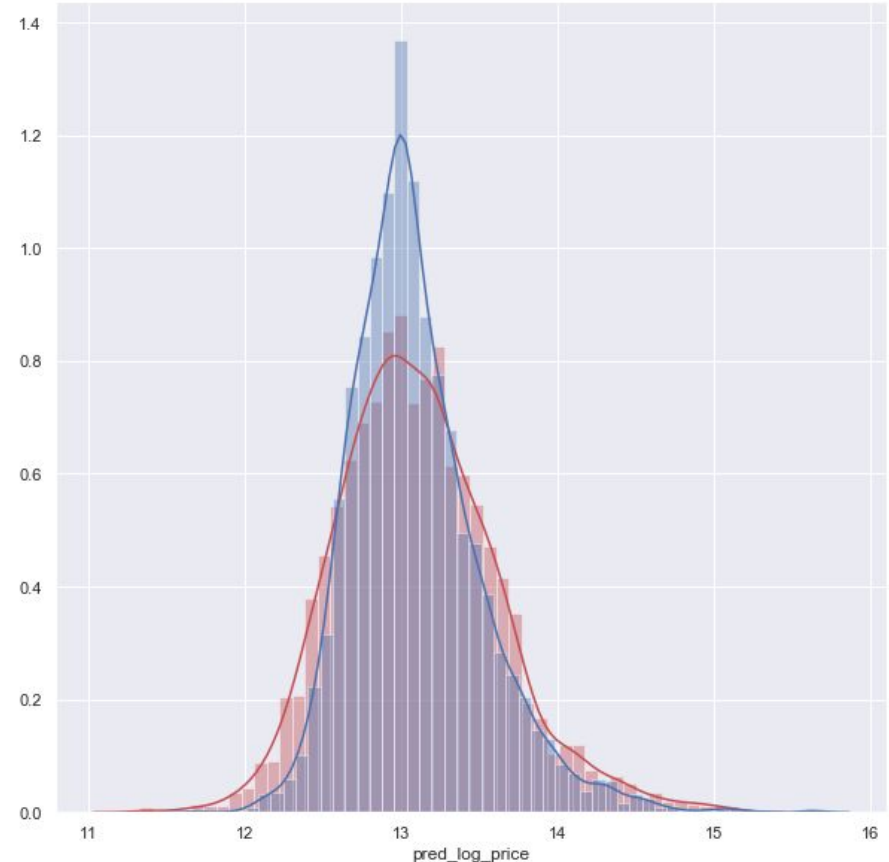
# Further exploration around renovations



- 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. Doubling square footage from 1500 to 3000 results in 82.2% increase in price
3. Renovation increases price of 40-year old house by 0.8%; a 90-year old house by 1.8%.
4. The difference between a house 11km away from the highest priced per sqft house to 22km away was 14.3%.
5. Separate model for **housing development**

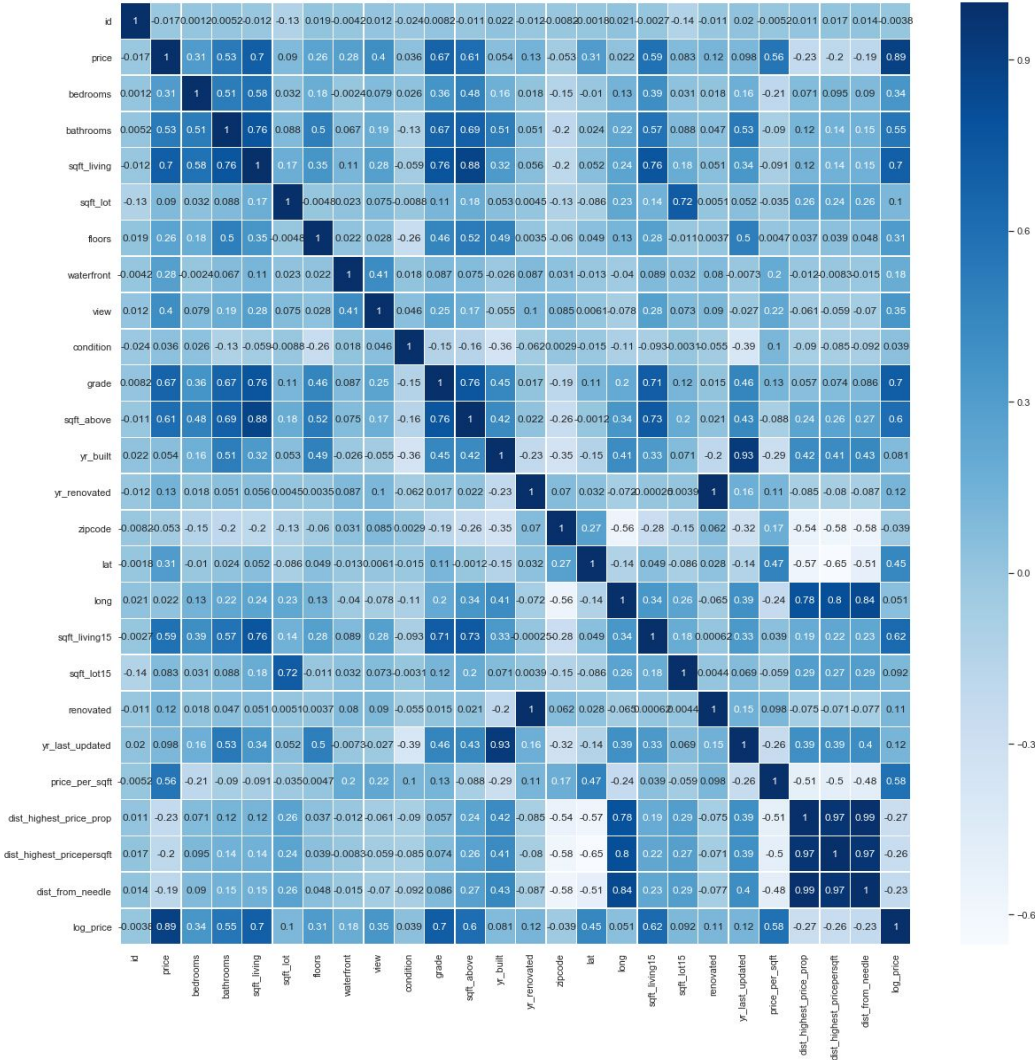


# Thank you

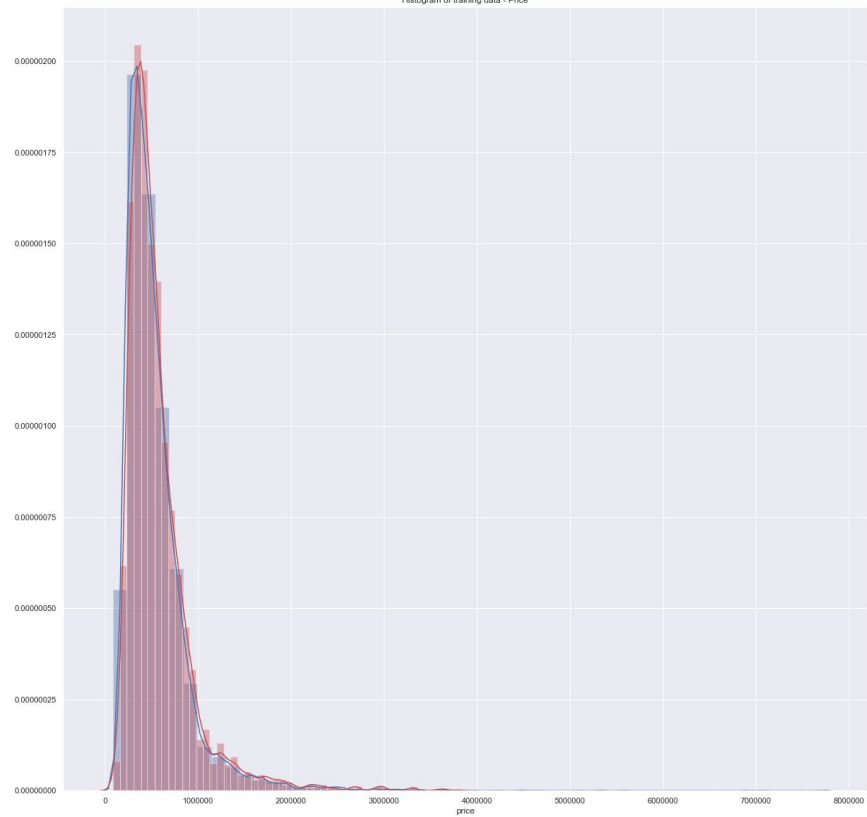
Questions, please

---

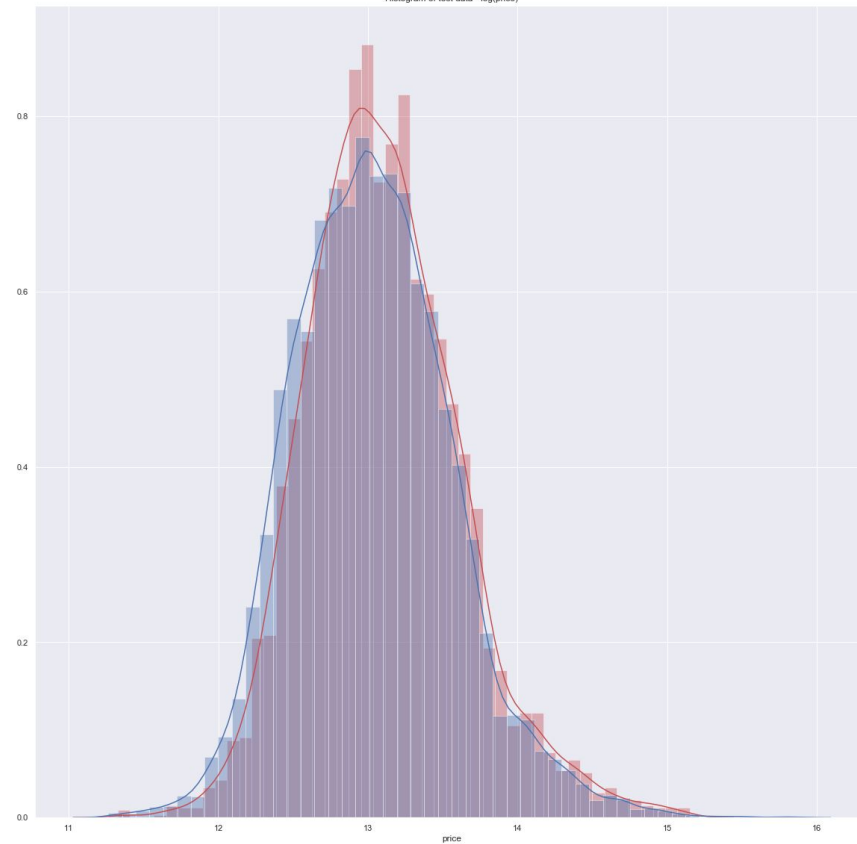
Heatmap of correlation coefficients for all variables including new ones



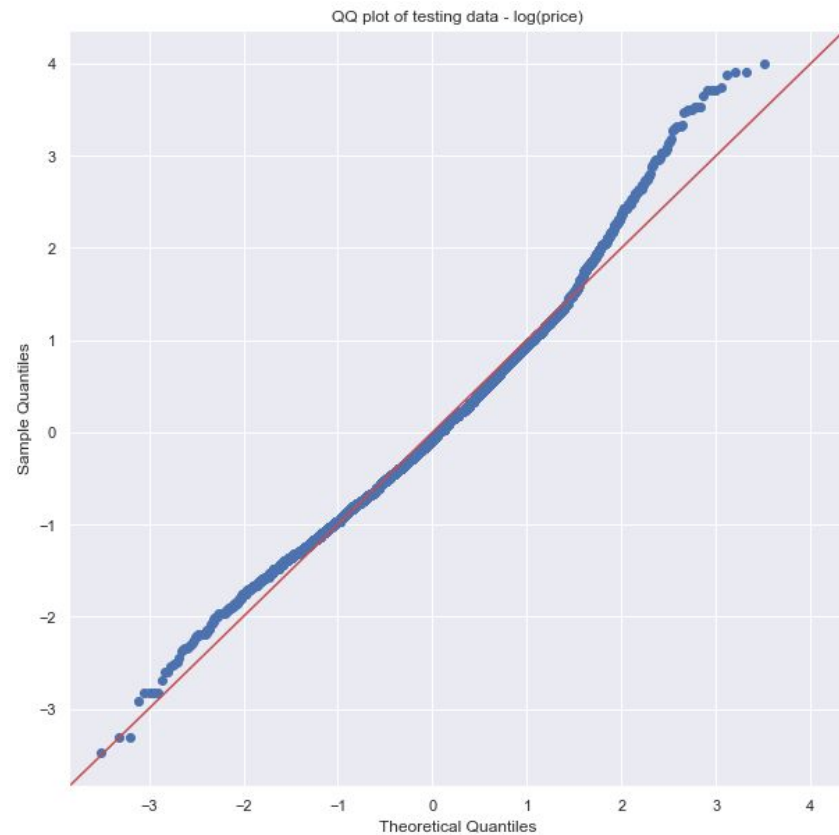
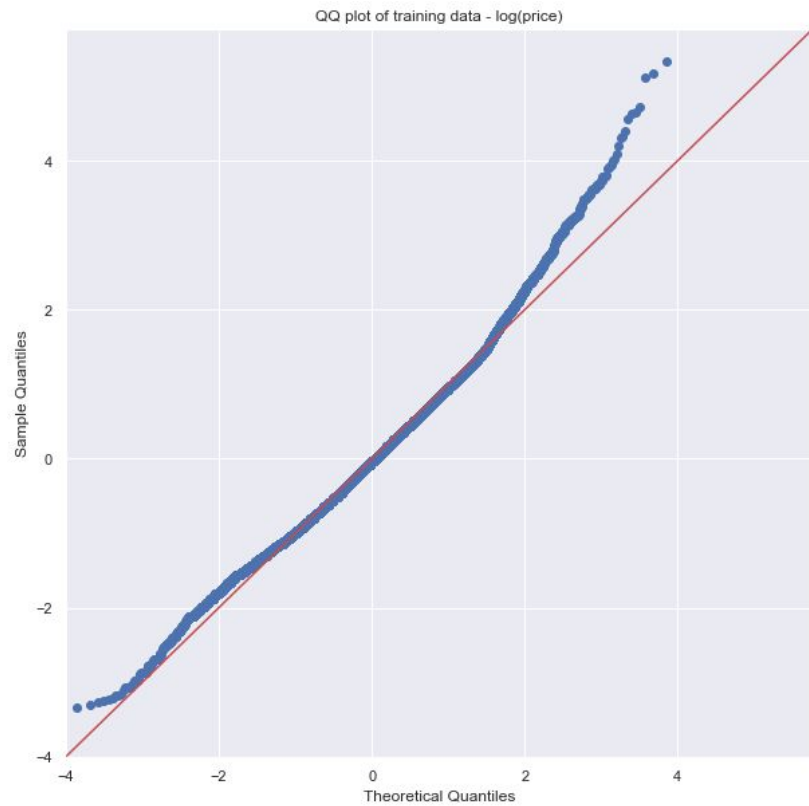
Histogram of training data - Price



Histogram of test data - log(price)







# Multivariate model summary

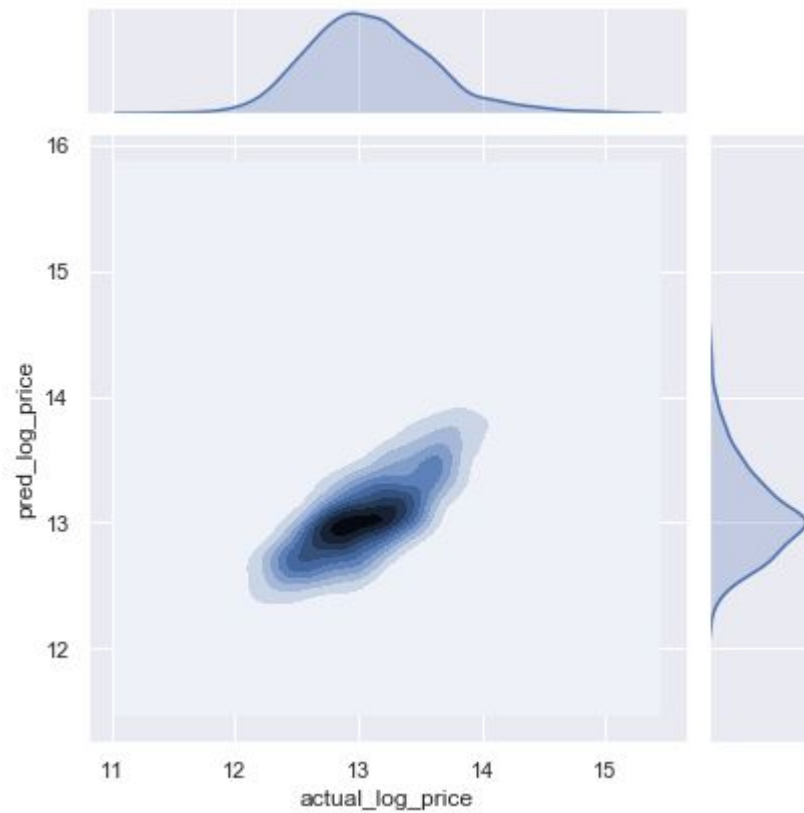
Dep. Variable:	price	R-squared:	0.610
Model:	OLS	Adj. R-squared:	0.610
Method:	Least Squares	F-statistic:	9019.
Date:	Wed, 23 Oct 2019	Prob (F-statistic):	0.00
Time:	14:26:01	Log-Likelihood:	-5372.8
No. Observations:	17276	AIC:	1.075e+04
Df Residuals:	17272	BIC:	1.078e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	12.0262	0.207	58.151	0.000	11.621	12.432
sqft_living	0.0004	2.97e-06	144.129	0.000	0.000	0.000
dist_highest_pricepersqft	-1.3372	0.019	-70.648	0.000	-1.374	-1.300
yr_last_updated	0.0002	0.000	2.212	0.027	2.69e-05	0.000
Omnibus:	220.331	Durbin-Watson:	1.995			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	390.197			
Skew:	-0.055	Prob(JB):	1.86e-85			
Kurtosis:	3.728	Cond. No.	2.41e+05			



# 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



# Multivariate model (log(sqft\_living) summary

Dep. Variable:	price	R-squared:	0.594						
				coef	std err	t	P> t	[0.025	0.975]
Model:	OLS	Adj. R-squared:	0.594	const	6.5085	0.046	141.241	0.000	6.418 6.599
Method:	Least Squares	F-statistic:	1.262e+04	log_sqft_living	0.9112	0.006	147.999	0.000	0.899 0.923
Date:	Wed, 23 Oct 2019	Prob (F-statistic):	0.00	dist_highest_pricepersqft	-1.3670	0.018	-77.939	0.000	-1.401 -1.333
Time:	14:26:02	Log-Likelihood:	-5734.0						
No. Observations:	17276	AIC:	1.147e+04	Omnibus:	48.480	Durbin-Watson:	2.001		
Df Residuals:	17273	BIC:	1.150e+04	Prob(Omnibus)	0.000	Jarque-Bera (JB):	49.760		
Df Model:	2			:					
Covariance Type:	nonrobust			Skew:	0.116	Prob(JB):	1.57e-11		
							138.		
				Kurtosis:	3.124	Cond. No.			