King County Real Estate Model

Best Home Features to Predict Real Estate Prices

By Tim McAleer and Crissy Bruce

The problem

Stakeholder

King County Real Estate Agency

Data

Provided by Flatiron

Problem Statement

Currently, an extreme amount of resources is spent on assessing the value of real estate. The Agency needs an efficient solution to pricing that benefits the client and the agency.

Goal

Determine and understand the the features that directly impact the sales price of a home in King County

1. Find most important features that best forecast home value
2. Eliminate those that raise our r^2 while also contributing to accurately impacting sales price

First, a few words on zip codes

A zip code coefficient does not describe the value of a home.

- Zip codes are assigned arbitrarily; a higher numbered zip code says nothing about the home itself.
- Splitting zip codes using git dummies produces 70 variables that say nothing about homes and raises our r^2 without contributing to our actual goal.
- So, we're ditching zip codes.

Scatterplots

- Checking for linear relationships
- Getting a better look at at categoricals.

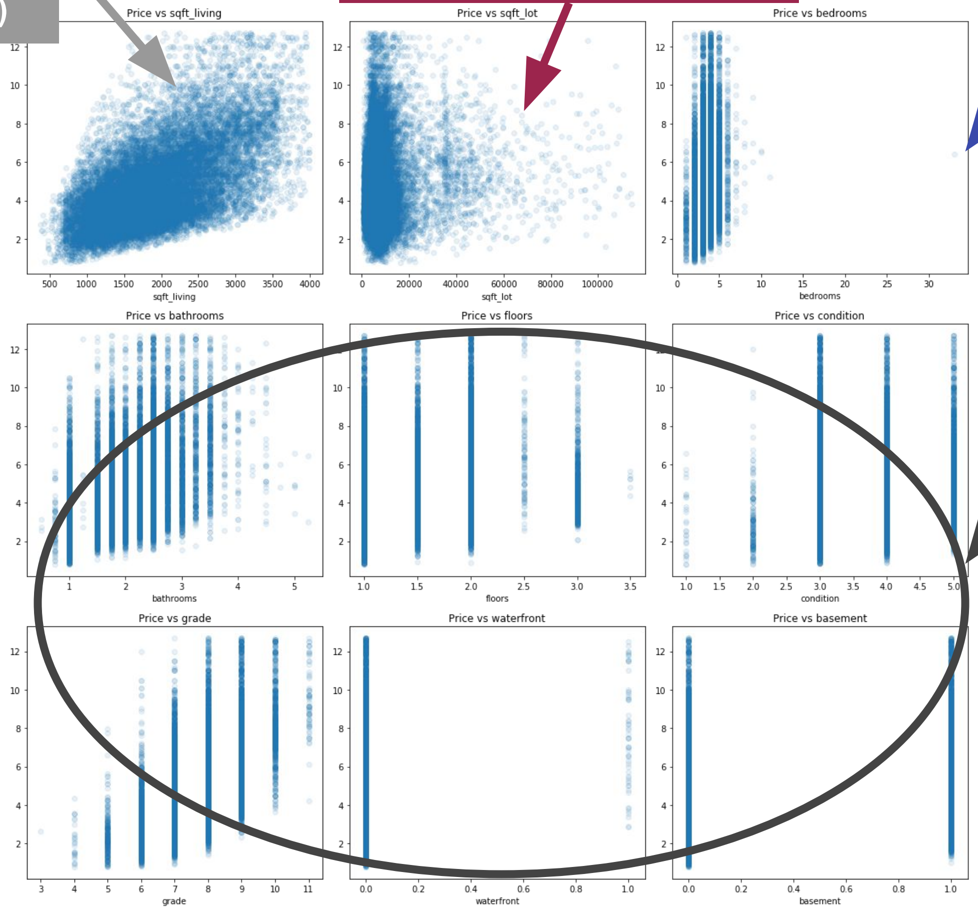
Note: Originally thought to be continuous, grade wasn’t a continuous feature after all.

Clear linear relationship (Price vs. sqft_living)

Needs some work to get more normally distributed (Price vs. sqft_lot)

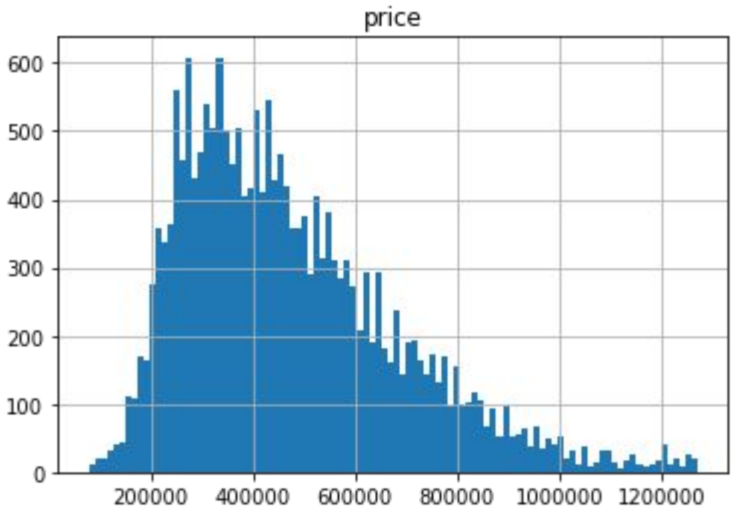
Outliers, especially right here on the edge (33)

Categorical Features

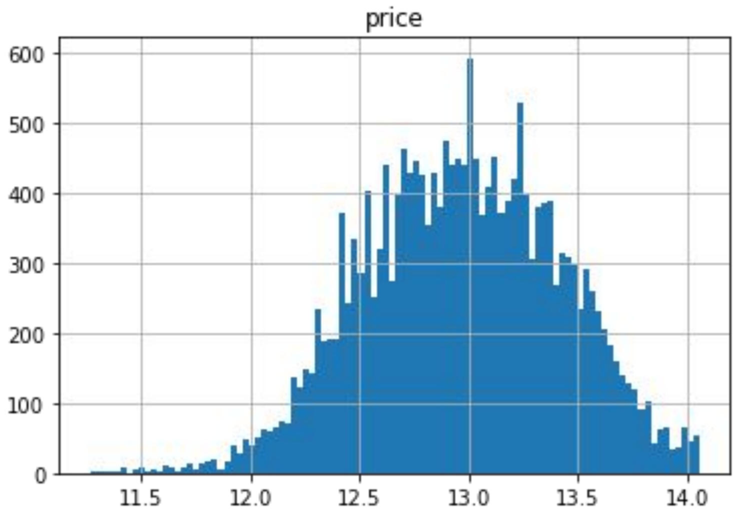


Log Transformations on Price

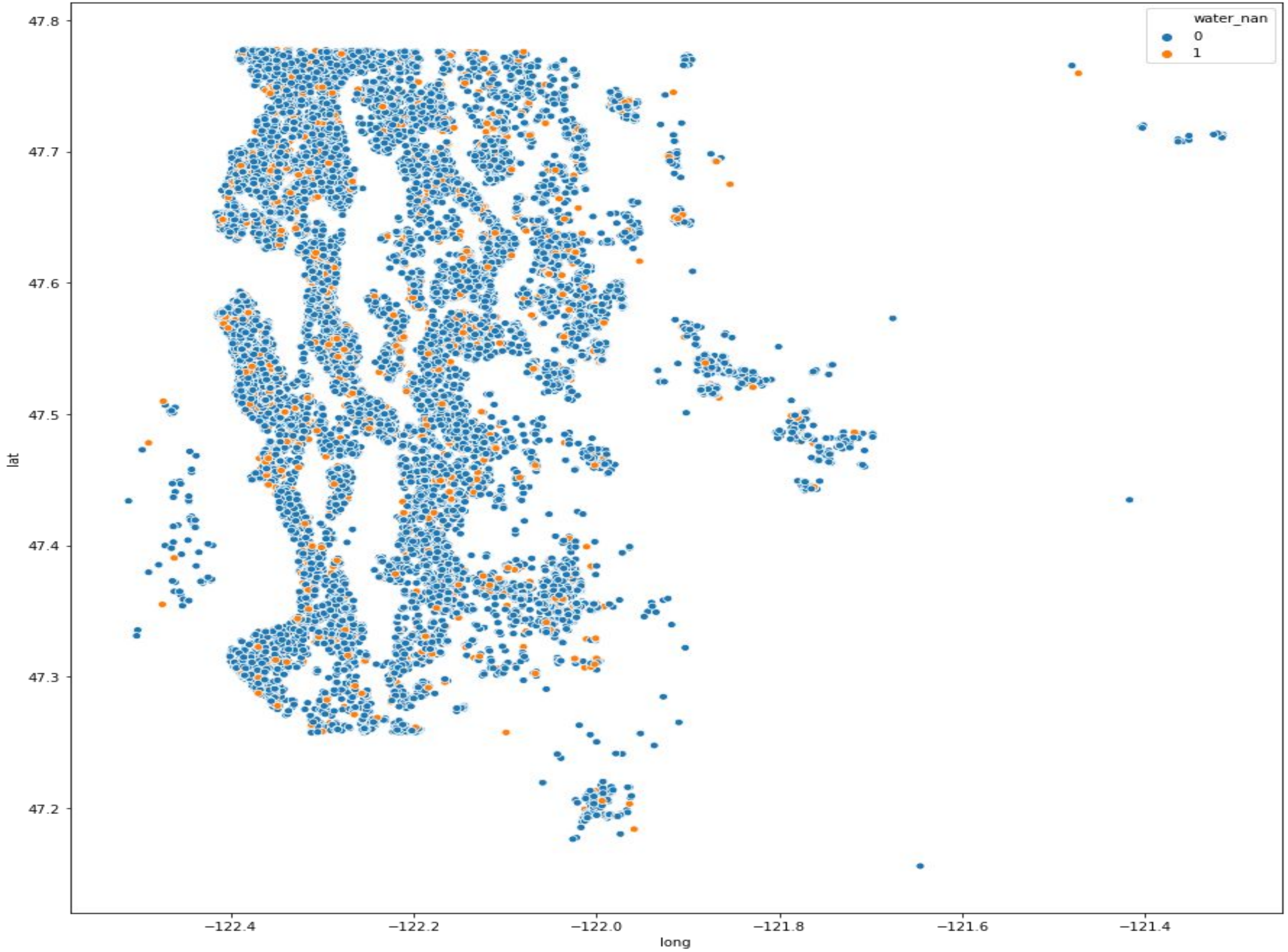
Before Log Transformation



After Log Transformation



NaN



- Looks to be <u>fairly distributed</u> across King County.
- Small percentage of homes are waterfront properties, while a large majority are <u>not</u>.
- Change all NaN values to 0, as this results in a <u>relatively small</u> percentage of <u>error</u>. It's outside the scope of this project.
- Future iterations could include utilizing latitude and longitude of properties to better ascertain its waterfront status.

Baseline Model

Dep. Variable:	price	R-squared:	0.472
Model:	OLS	Adj. R-squared:	0.472
Method:	Least Squares	F-statistic:	1760.
Date:	Mon, 23 Nov 2020	Prob (F-statistic):	0.00
Time:	13:24:44	Log-Likelihood:	-5897.1
No. Observations:	19688	AIC:	1.182e+04
Df Residuals:	19677	BIC:	1.190e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	10.8452	0.027	407.352	0.000	10.793	10.897
sqft_living	0.0002	1.15e-05	19.005	0.000	0.000	0.000
sqft_lot	-1.456e-06	2.36e-07	-6.155	0.000	-1.92e-06	-9.92e-07
sqft_above	1.24e-07	1.25e-05	0.010	0.992	-2.45e-05	2.47e-05
bedrooms	-0.0233	0.003	-7.028	0.000	-0.030	-0.017
bathrooms	-0.0433	0.005	-8.009	0.000	-0.054	-0.033
floors	0.0720	0.006	11.793	0.000	0.060	0.084
condition	0.0937	0.004	24.890	0.000	0.086	0.101
grade	0.1876	0.004	52.960	0.000	0.181	0.195
waterfront	0.5021	0.043	11.555	0.000	0.417	0.587
basement	0.1347	0.009	15.058	0.000	0.117	0.152

Omnibus:	6.474	Durbin-Watson:	1.977
Prob(Omnibus):	0.039	Jarque-Bera (JB):	6.170
Skew:	-0.019	Prob(JB):	0.0457
Kurtosis:	2.922	Cond. No.	2.69e+05

Results

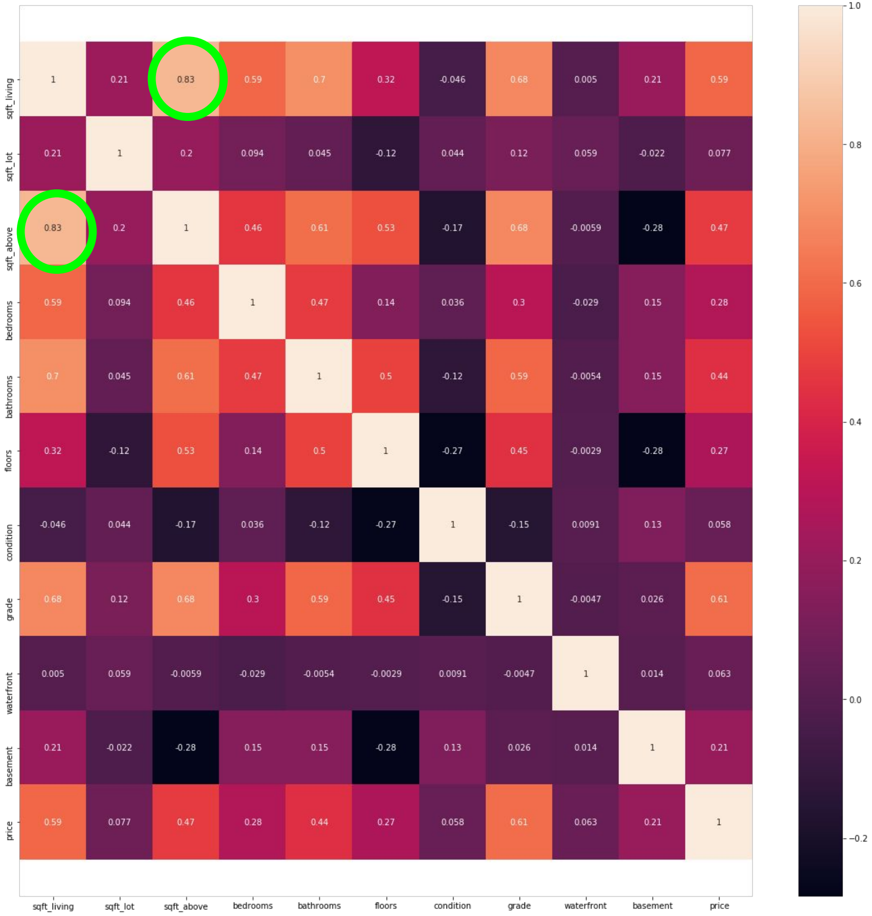
R-squared is **0.472** RMSE is **110963.47**

It is obvious that the model is not a good fit with such a low R-squared and such a high RMSE.

P-values

- Sqft_above feature p-value is high, so more work to do on that one
- Low values of the others, however, indicate strong evidence that there is a relationship between those features and price (aka strong evidence against the likeliness of no relationship-null hypothesis)

Iterative Process Continues



Heatmap

- Multicollinearity: <u>Sqft_liv & Sqft_abv</u>.
  - Dropped Sqft abv because Sqft_liv had a stronger linear relationship to price than Sqft_abv.
  - Remember that high p-value in our first model for Sqft_abv...feature=Gone!
- Categorized grades to **low, med, high**
- Categorized floors to single level or multi-story
- Drop condition after p-value of dummies show it is an inaccurate category
- More Log Transforming - Sqft_liv and Sqft_lot

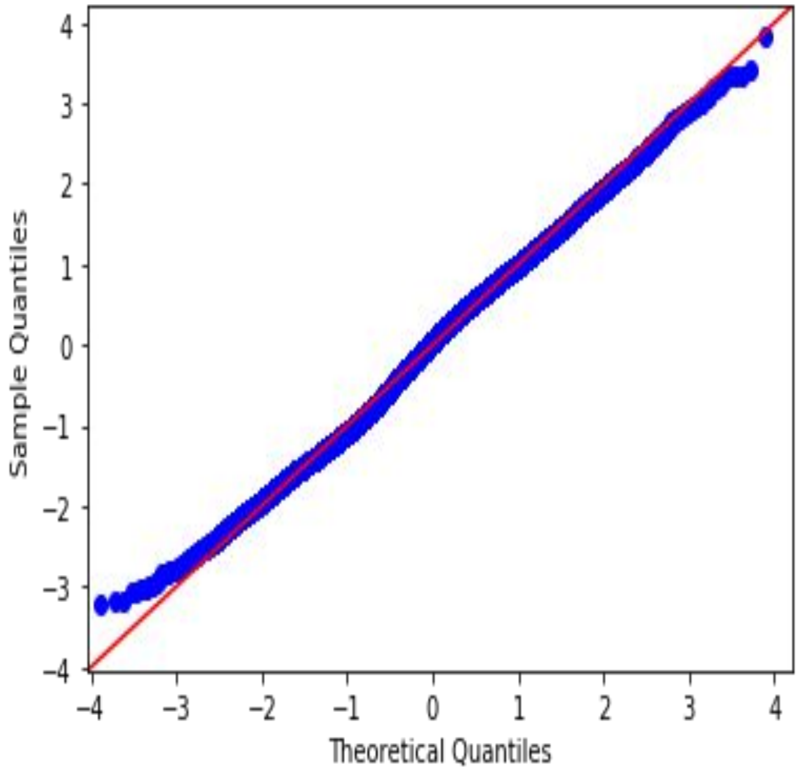
Final Model

OLS Regression Results

Dep. Variable:	price	R-squared:	0.404
Model:	OLS	Adj. R-squared:	0.403
Method:	Least Squares	F-statistic:	1479.
Date:	Mon, 23 Nov 2020	Prob (F-statistic):	0.00
Time:	13:25:58	Log-Likelihood:	-7098.0
No. Observations:	19687	AIC:	1.422e+04
Df Residuals:	19677	BIC:	1.429e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.5094	0.072	117.465	0.000	8.367	8.651
sqft_living	0.6688	0.013	53.233	0.000	0.644	0.693
sqft_lot	-0.0620	0.004	-15.134	0.000	-0.070	-0.054
bedrooms	-0.0494	0.004	-13.361	0.000	-0.057	-0.042
bathrooms	-0.0255	0.006	-4.553	0.000	-0.037	-0.015
waterfront	0.5391	0.046	11.661	0.000	0.448	0.630
basement	0.0898	0.006	14.835	0.000	0.078	0.102
mid	0.1544	0.009	17.226	0.000	0.137	0.172
high	0.4815	0.018	27.416	0.000	0.447	0.516
multistory	0.0685	0.007	10.086	0.000	0.055	0.082

Omnibus:	140.044	Durbin-Watson:	1.983
Prob(Omnibus):	0.000	Jarque-Bera (JB):	96.999
Skew:	-0.033	Prob(JB):	8.65e-22
Kurtosis:	2.662	Cond. No.	368.



Final Results

R-squared is **0.404** RMSE is **167133.77**

The final model is still not a good fit with such a low R-squared and such a high RMSE.

- R-squared of .404 indicates that **only 40%** of the data can be explained by our model. Our model continues to decrease in terms of the linear fit as as we make iterate through different models It appears on the QQ plot that the residuals on the edges seem to be the culprit

- All p-values are good so there is strong evidence that there is a relationship between those features and price (aka strong evidence against the likeliness of no relationship-null hypothesis)

Next Steps

- Further investigate how to increase r^2 while lowering RMSE
    - Change order of iterations?
- Binning zip codes based on geography (ex. NW King County vs. SW King County, etc.)
- Look at other types of models
- API (link)

Questions?

Project Repo: https://github.com/tcmcaleer/SeattleHousing

Thanks to Yish and all Class Members for all the help!