King County House Sales

Andrew Muller
Asher Khan

Business Case

Predicting how much a house should be sold for in order to determine whether a house on the market is being underpriced or overpriced. Our clients are homeowners looking to sell their house, but do not know how much to sell their house for.

Data Used

- King County data provided:
  - 21,597 Homes
  - Houses ranging from $78,000 to $7,700,000
  - Year Built from 1900 to 2015

	price	yr_built
count	2.159700e+04	21597.000000
mean	5.402966e+05	1970.999676
std	3.673681e+05	29.375234
min	7.800000e+04	1900.000000
25%	3.220000e+05	1951.000000
50%	4.500000e+05	1975.000000
75%	6.450000e+05	1997.000000
max	7.700000e+06	2015.000000

EDA- Correlations

Correlated to Price:

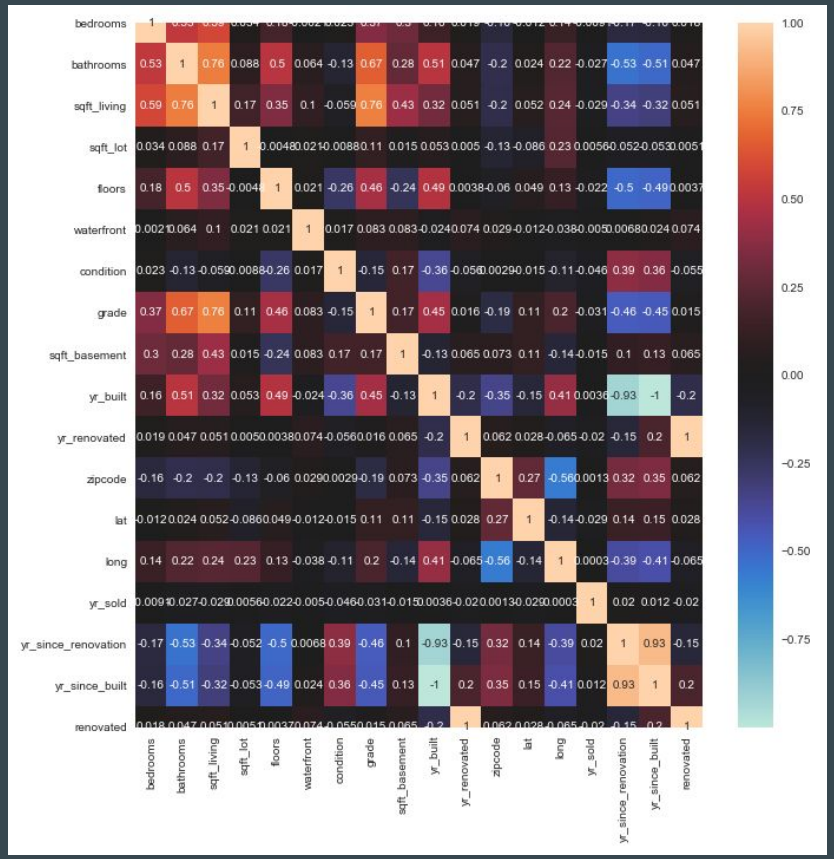
- bedrooms
- bathrooms
- sqft_living
- sqft_lot
- sqft_basement
- yr_built

Created New Columns:

- yr_since_built
- yr_since_renovated
- renovated

Sqft_living- Footage of home
Sqft_lot- Footage of entire lot
Sqft_basement- Footage of basement

EDA- Collinearity (Before)

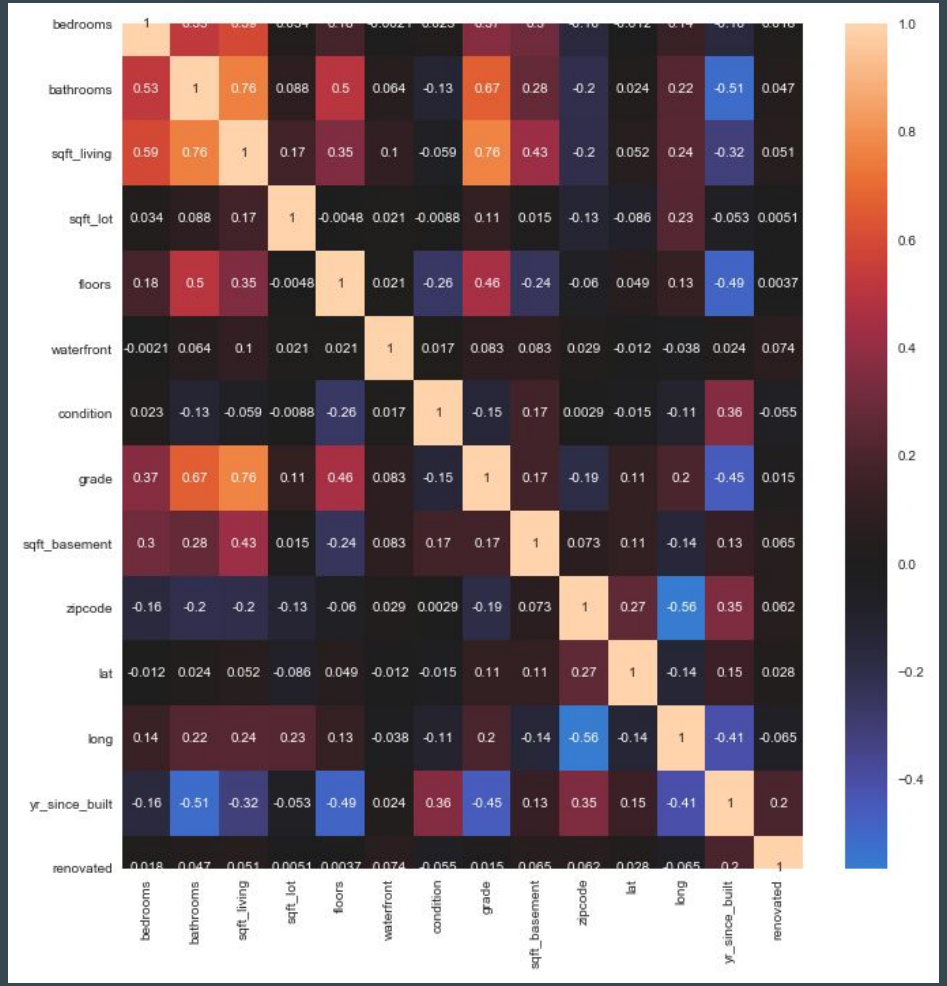


pairs	
(yr_renovated, renovated)	0.999968
(yr_built, yr_since_built)	0.999873
(yr_since_renovation, yr_since_built)	0.926424
(yr_since_renovation, yr_built)	0.926173

EDA- Collinearity (After)

In order to remove collinearity, we dropped the yr_since_renovation, yr_built and the yr_renovated columns.

These are also incorporated into other features, renovated and yr_since_built, so they were safe to remove.



EDA-Features to Transform

Transformed all continuous variables:

- bedrooms
- bathrooms
- sqft_living
- sqft_lot
- sqft_basement
- lat
- long
- yr_since_built

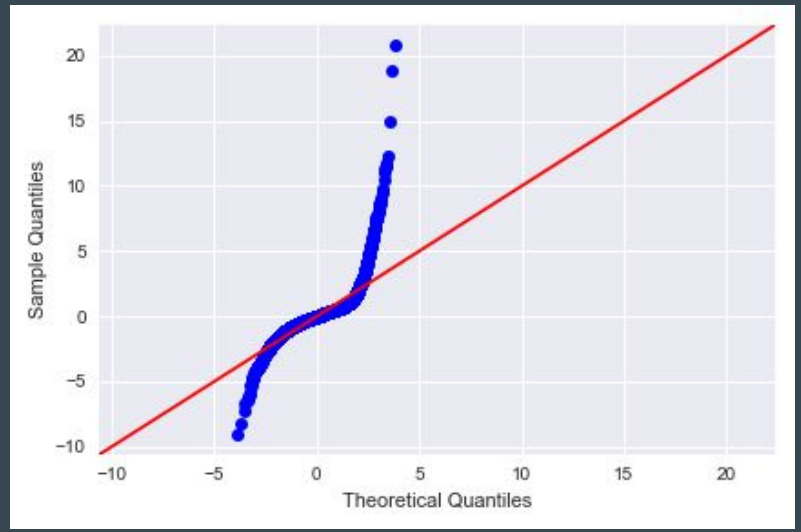
Get Dummies: Floors, condition, grade, zipcode.

Waterfront was already in indicador variable form.

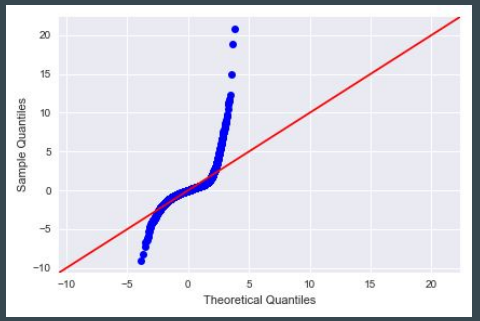
Baseline Model

- R2 of 0.821
- adjusted R2 of 0.820
- Train RMSE of 157156
- Test RMSE of 139700
- 81 significant features (p-value < 0.05)
- 103 features total

Dep. Variable:	price	**R-squared:**	0.821
Model:	OLS	**Adj. R-squared:**	0.820
Method:	Least Squares	**F-statistic:**	730.0
Date:	Tue, 24 Nov 2020	**Prob (F-statistic):**	0.00
Time:	14:03:19	**Log-Likelihood:**	-2.1678e+05
No. Observations:	16197	**AIC:**	4.338e+05
Df Residuals:	16095	**BIC:**	4.345e+05
Df Model:	101		
Covariance Type:	nonrobust		



Iterative Modeling Process

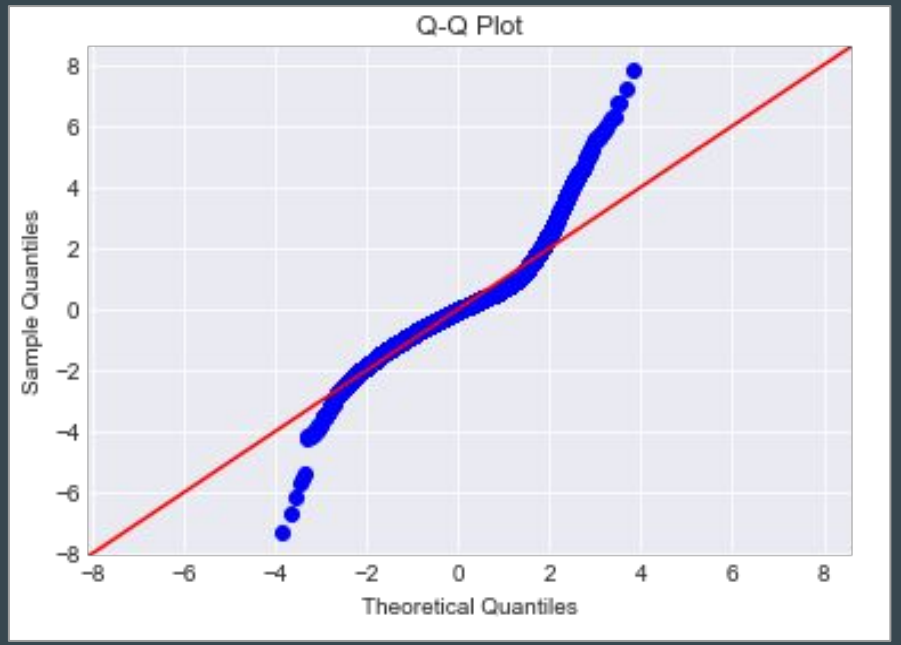


Dropping Collinear Features:

- yr_since_renovation, yr_built and the yr_renovated

Removing Outliers:

- Corrected the typo of 33 bedrooms
- Price outlighers higher than three times the standard deviation away from the mean



- ~~R2 = 0.8264414496035375~~
- R2 adjusted = 0.8253755271400415
- RMSE (train) = 107801.01535312463
- RMSE (test) = 105098.35236657722
- number of significant features = 82

Final Model

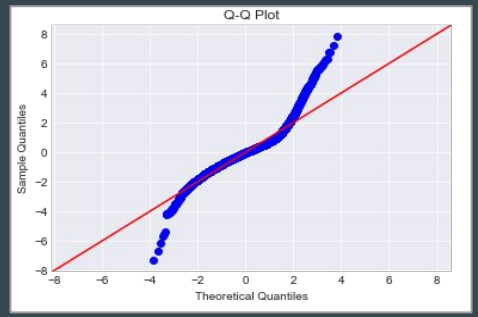
Log Transforming:

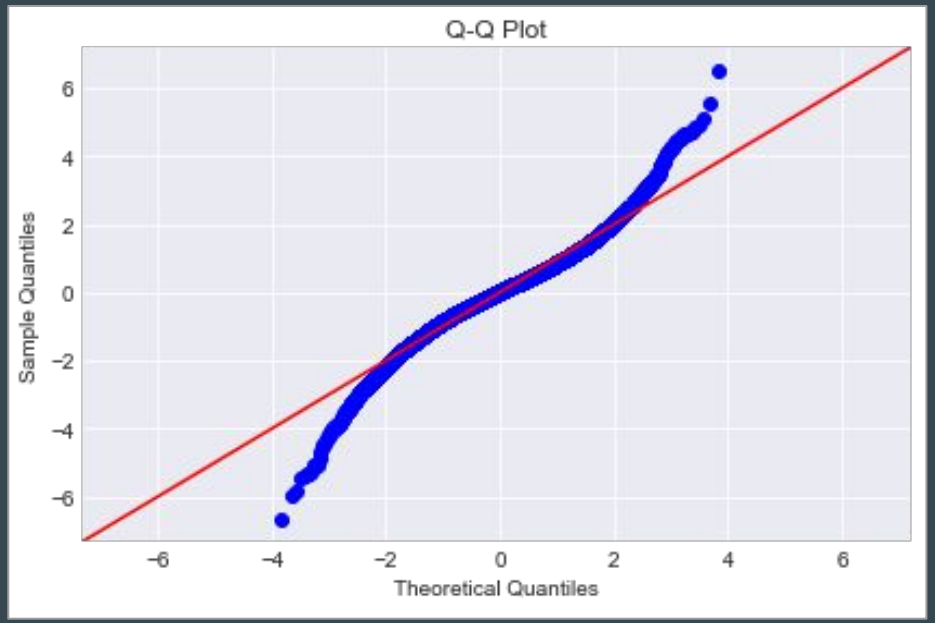
- Log transformed price, bedrooms, bathrooms, sqft_living, sqft_lot.

Removing Insignificant Features:

- Looping through our model we removed features with p-values higher than 0.05.
- Removed 9 insignificant features.

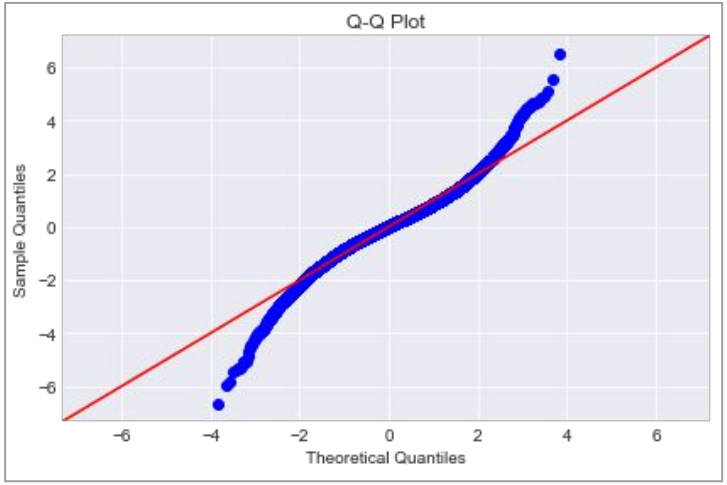
- R2 = 0.8566083994927415
- R2 adjusted = 0.8558099143415272
- RMSE (train) = 103086.41637517781
- RMSE (test) = 99701.98273007278
- number of significant features = 90

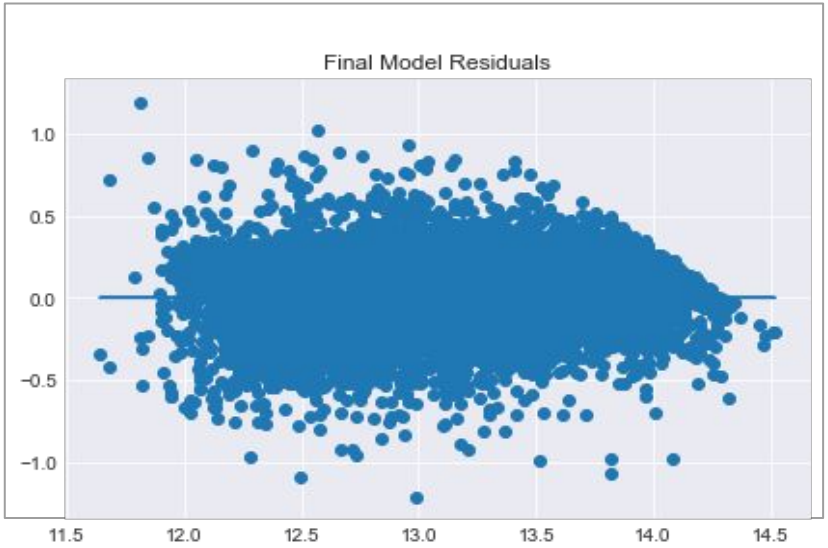




Interpreting Results

- Each bedroom decreases the sale price of a house by 5%
- Each bathroom increases the sale price of a house by 6%
- A 1% change in square footage living area increases the sale price of a house by .48%
- A 1% change in square footage lot area increases the sale price of a house by .07%
- If the house is on the waterfront, the sale price of a house increases by 60%
- A 1% change in square footage basement area decreases the sale price of a house by .00005%
- If you move north, a 1 degree increase in latitude increases the sale price of a house by 65%
- If you move east, a 1 degree increase in longitude decreases the sale price of a house by 53%
- A 1-year increase in the age of a house increases its sale price by .04%
- A house that has been renovated has its sale price increased by 6%





Recommendations and Future Work

- We attempted to add interaction features to our model, but our results indicated that they only decreased the accuracy of our model. With more time, we could take a deeper look at these and find out why that is the case, and see if other interactions could help our model.
- Similarly, adding polynomial features could make our predictions more accurate. Trial-and-error would be needed to determine which features could be changed in this way to improve our model.
- Using a mapping library could turn the longitude and latitude into more directly beneficial information, like distance to a school or grocery store. With more time, we could create new features using this information to add to our model.

Thank You!

Repo: https://github.com/MullerAC/king-county-house-sales