King County Housing Prices Analysis

Project for Flatiron's Data Science Bootcamp
by Henry Alpert
October 25, 2020

FLATIRON SCHOOL

Overview

When representing either buyers or sellers, Royalty Real Estate often has difficulty developing internal benchmarks regarding a home’s worth.

How much is a home’s price affected by:

- a recent renovation?
- 10 additional square feet of living space?
- other listings for sale in the area?

This analysis will help Royalty Real Estate develop answers to these and similar questions for its business in King County, Washington.

Data Used

The main data source includes homes sold in King County in 2014 and 2015 and includes info regarding each home's:

- date sold
- price
- bedrooms
- bathrooms
- internal square footage
- lot square footage
- number of floors
- grade
- condition
- waterfront location
- year built
- year renovated

Additional data: Census data from 2015 was mapped onto zip code data.

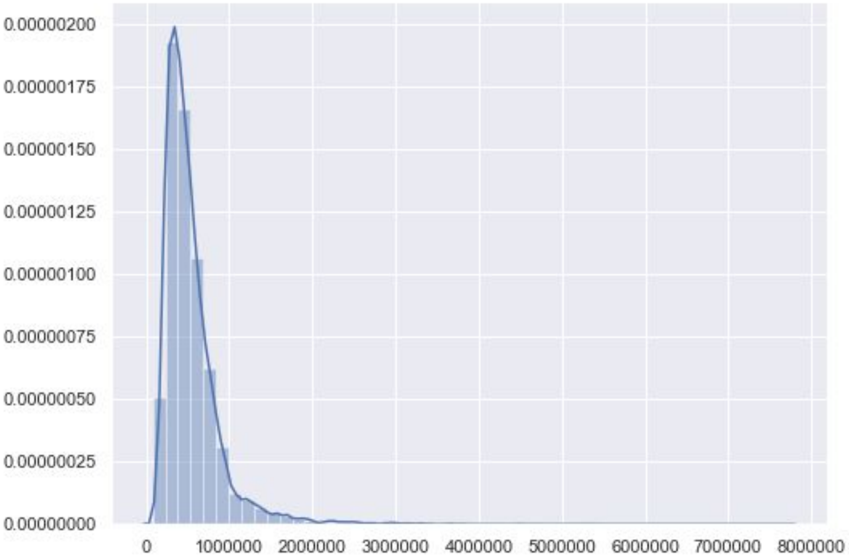
Data Overview

Number of homes sold	21,597
Dates of sale:	2014 - 2015
Price range	$77,000 - $7,700,000
Mean price	$540,200
Year houses built:	1900 - 2015
Living area (mean)	2080 sq. ft
Lot area (mean)	15,100 sq. ft
Population of King County (2015)	1,990,000

Exploratory Data Analysis

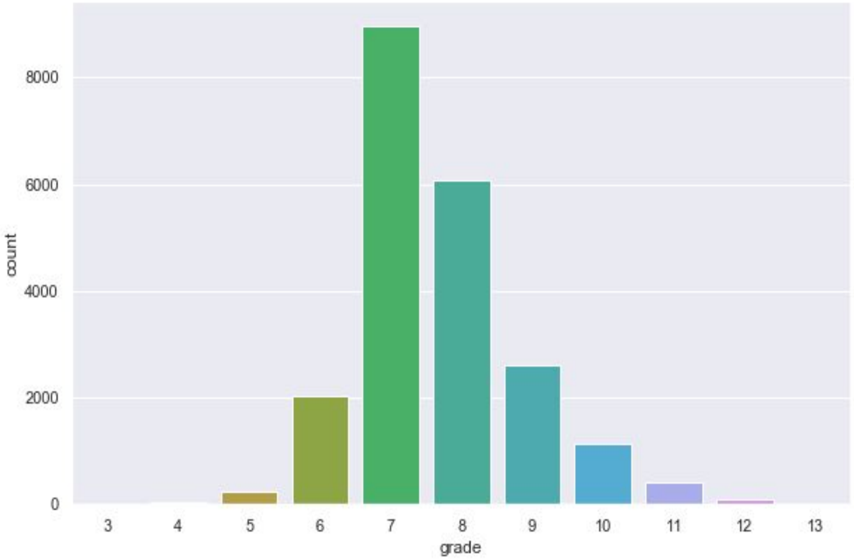
Prices

Most Under 1.5 mil



Grade

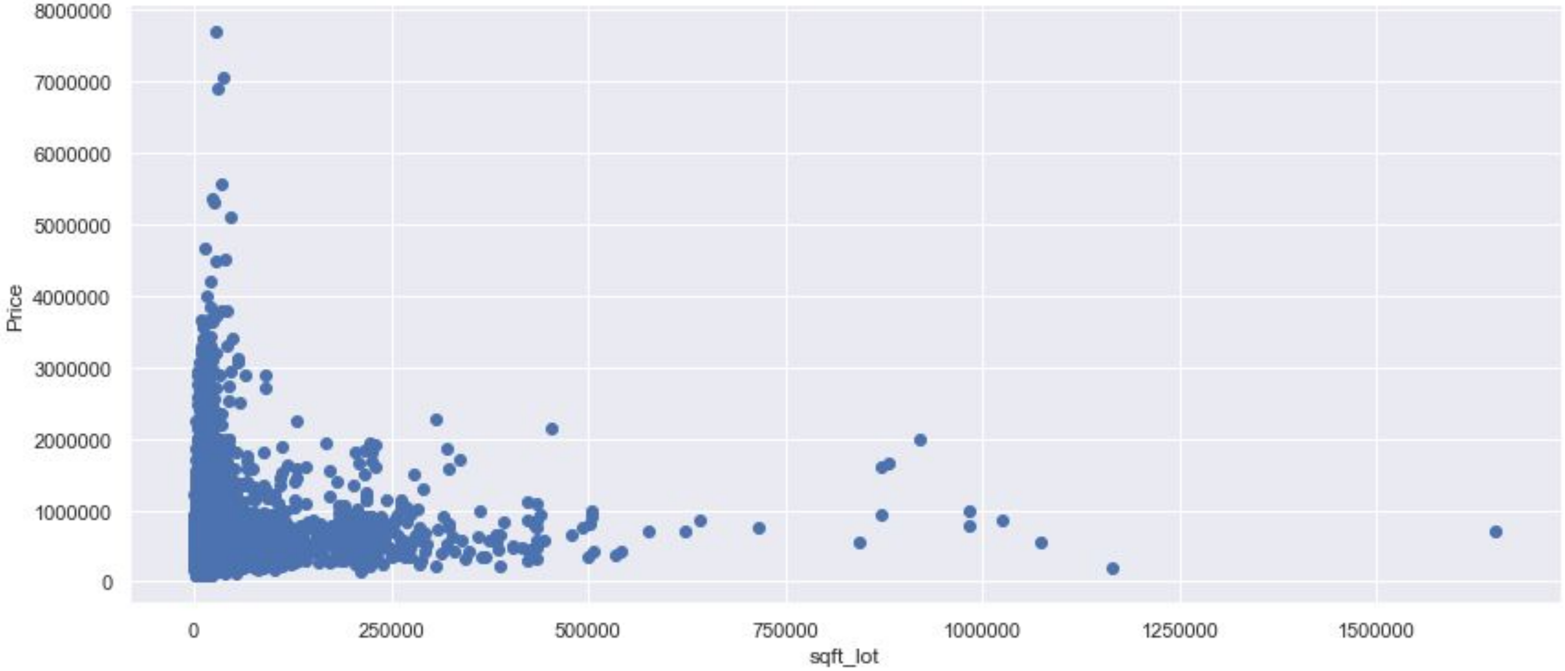
Most 7 or 8



Grade - Classification by construction quality which refers to the types of materials used and the quality of workmanship. Buildings of better quality (higher grade) cost more to build per unit of measure and command higher value.

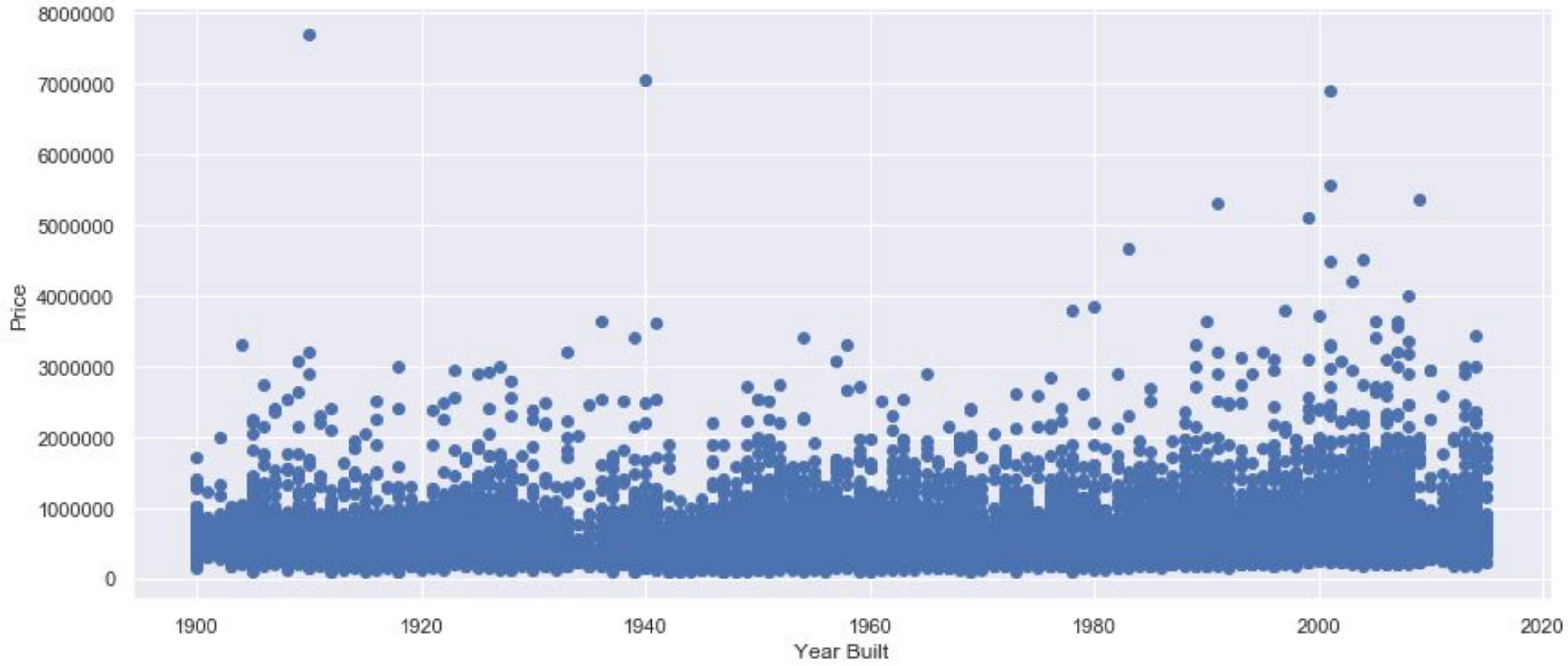
Price vs. Lot Square Ft.

Some expensive homes have small lots and vice versa



Price vs. Year Built

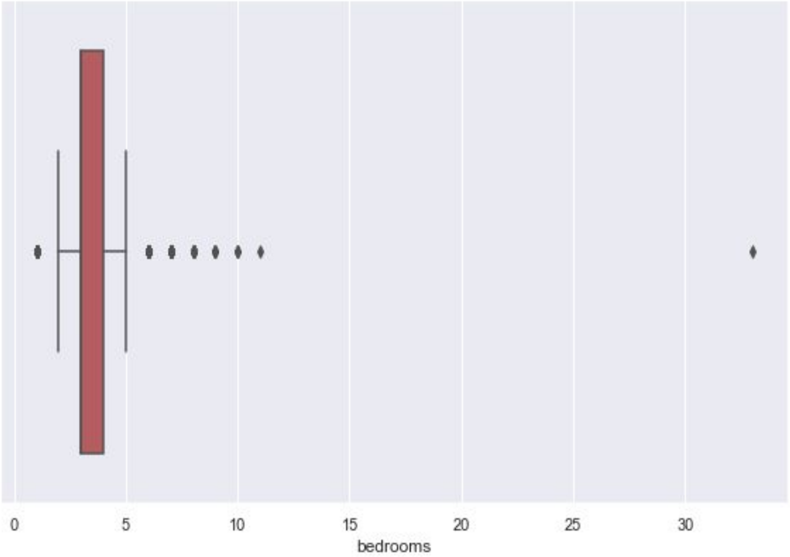
No clear relationship



Eliminating Outliers

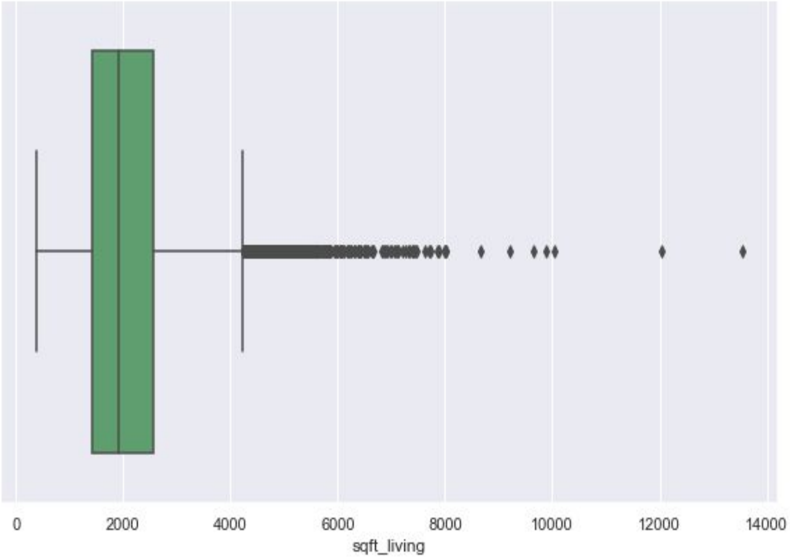
Bedroom

One house has 33 bedrooms



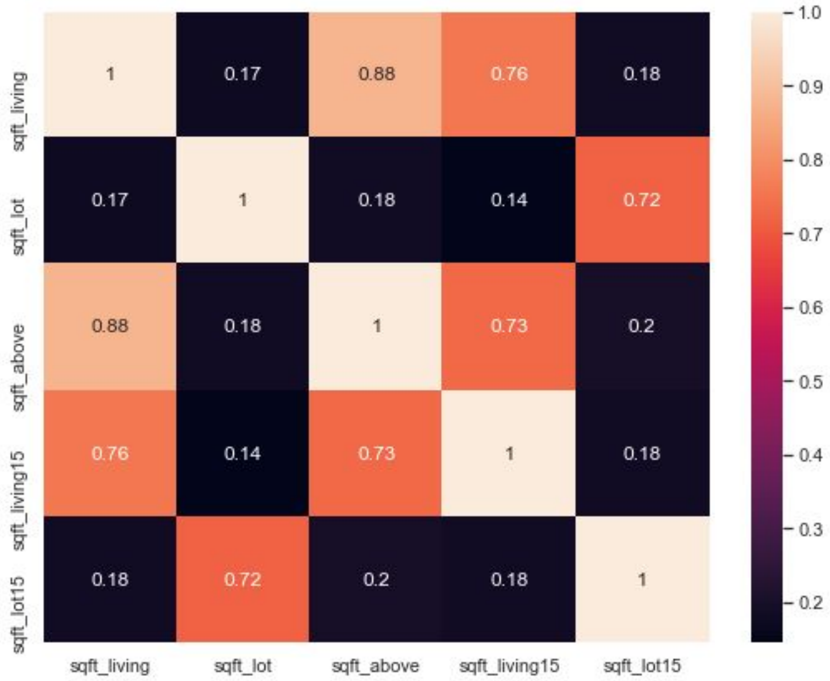
Sq. Ft Living

A handful of homes are over 10,000 sq.ft



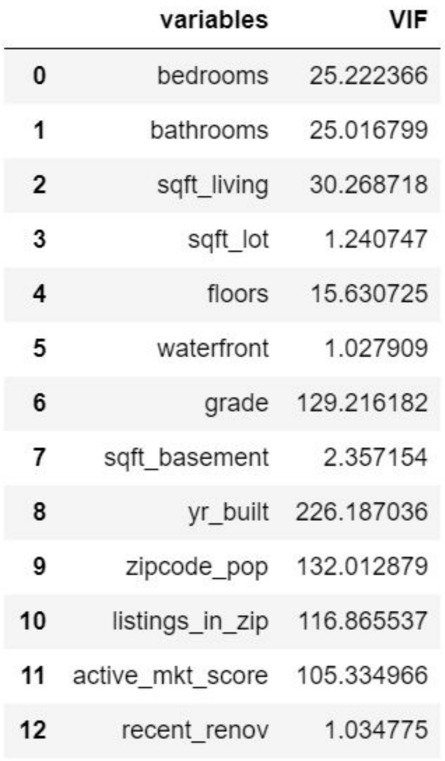
Checking for Multicollinearity

Heat map of square-foot-related categories



Look to eliminate categories > 0.70

Looking to get VIFs under 5.0



	variables	VIF
0	bedrooms	25.222366
1	bathrooms	25.016799
2	sqft_living	30.268718
3	sqft_lot	1.240747
4	floors	15.630725
5	waterfront	1.027909
6	grade	129.216182
7	sqft_basement	2.357154
8	yr_built	226.187036
9	zipcode_pop	132.012879
10	listings_in_zip	116.865537
11	active_mkt_score	105.334966
12	recent_renov	1.034775

Modeling

Changed Categories

New Categories

- ***active market score*** - a ratio of number of sales in a zip code per the zip code population. Score ranges from 3.4 to 22.0
- ***recently renovated*** - any house renovated since 1990 (25 years ago in 2015)
- ***year built 1 - yr built 6*** - six categories of a roughly 20 year-range per category (1900-1919, 1920-1939, etc.)

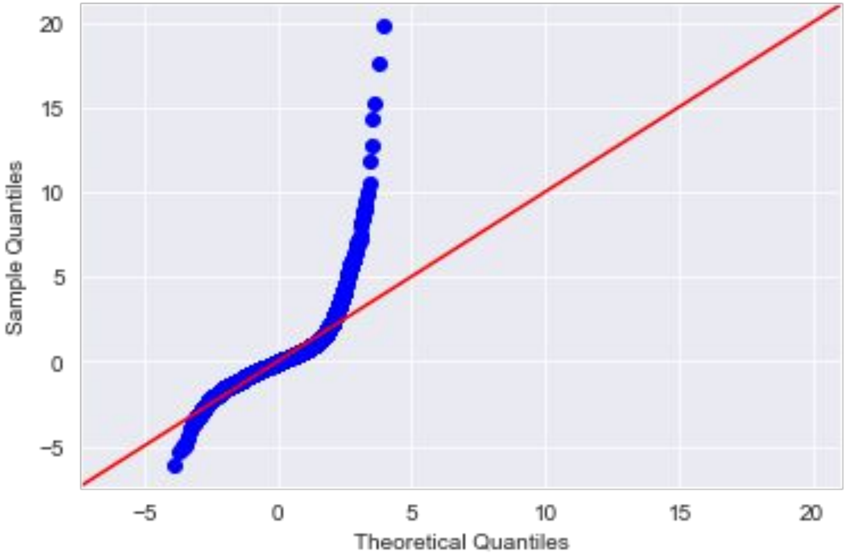
Some Removed Categories

- view
- latitude
- longitude
- condition
- sqft_living15
- sqft_lot15

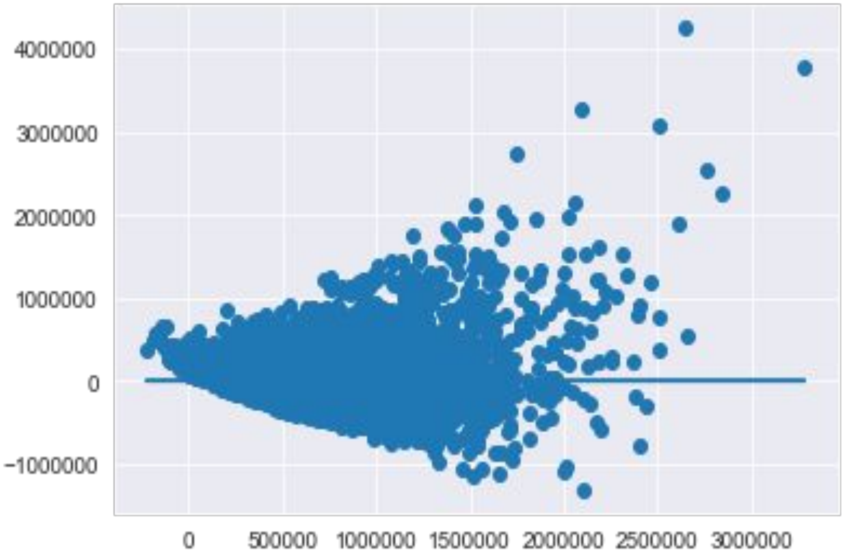
First model didn’t fit well

Note: Continuous categories were normalized

Q-Q Plot



Residual Scatter Plot



Final model was limited to houses with a price less than or equal to $900,000.

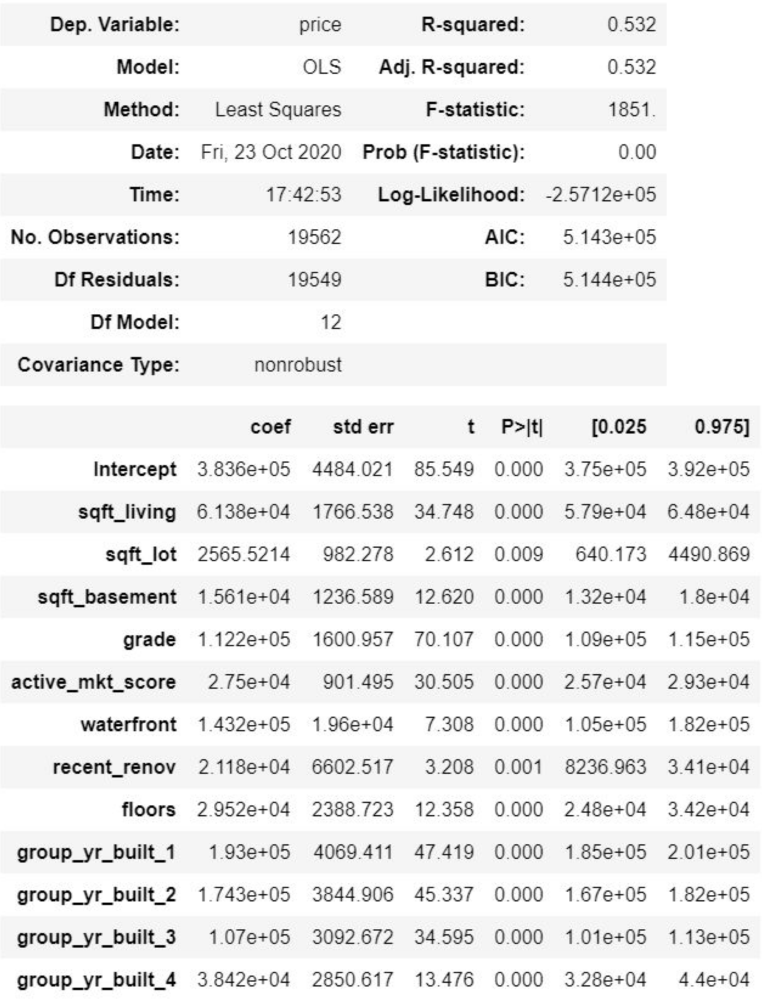
Final Model

- R-squared - 0.532
- p-values lower than 0.05 threshold

Selected coefficient interpretations

Avg. expectations if all other variables stay the same:

- For each 910.9 increase in sqft_living, home's price increases $61,380
- For each grade increase, price increase is $112,000
- House on a waterfront is expected to be worth $143,200 more
- Houses in active markets command higher prices
- Each additional floor added to a house expected to increase worth $29,520
- For homes built before 1980, older houses expected to be worth more than newer houses

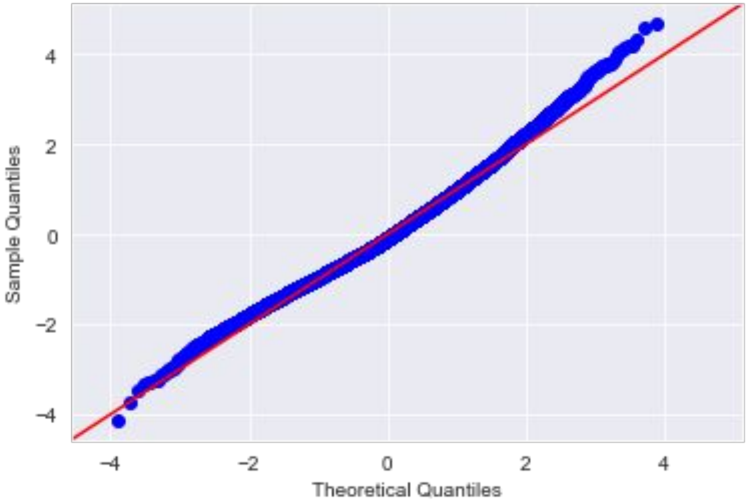


Dep. Variable:	price	**R-squared:**	0.532
Model:	OLS	**Adj. R-squared:**	0.532
Method:	Least Squares	**F-statistic:**	1851.
Date:	Fri, 23 Oct 2020	**Prob (F-statistic):**	0.00
Time:	17:42:53	**Log-Likelihood:**	-2.5712e+05
No. Observations:	19562	**AIC:**	5.143e+05
Df Residuals:	19549	**BIC:**	5.144e+05
Df Model:	12		
Covariance Type:	nonrobust		

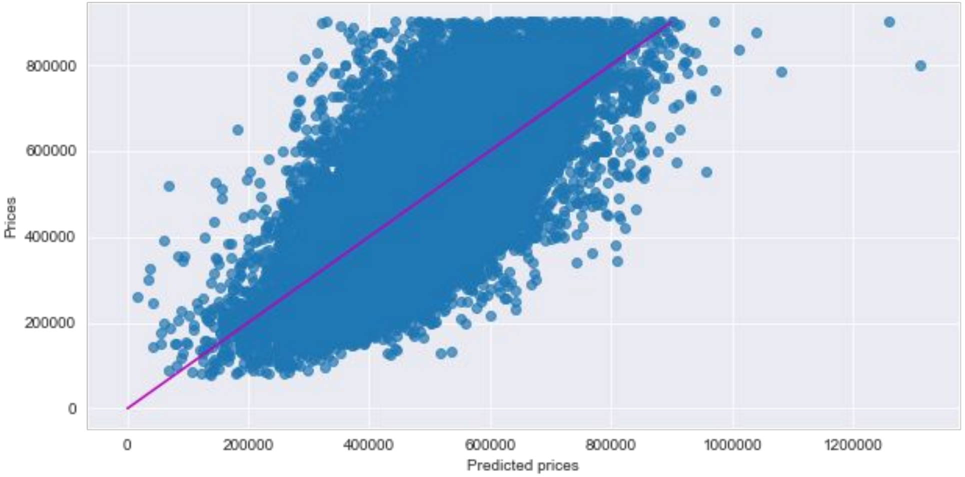
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.836e+05	4484.021	85.549	0.000	3.75e+05	3.92e+05
sqft_living	6.138e+04	1766.538	34.748	0.000	5.79e+04	6.48e+04
sqft_lot	2565.5214	982.278	2.612	0.009	640.173	4490.869
sqft_basement	1.561e+04	1236.589	12.620	0.000	1.32e+04	1.8e+04
grade	1.122e+05	1600.957	70.107	0.000	1.09e+05	1.15e+05
active_mkt_score	2.75e+04	901.495	30.505	0.000	2.57e+04	2.93e+04
waterfront	1.432e+05	1.96e+04	7.308	0.000	1.05e+05	1.82e+05
recent_renov	2.118e+04	6602.517	3.208	0.001	8236.963	3.41e+04
floors	2.952e+04	2388.723	12.358	0.000	2.48e+04	3.42e+04
group_yr_built_1	1.93e+05	4069.411	47.419	0.000	1.85e+05	2.01e+05
group_yr_built_2	1.743e+05	3844.906	45.337	0.000	1.67e+05	1.82e+05
group_yr_built_3	1.07e+05	3092.672	34.595	0.000	1.01e+05	1.13e+05
group_yr_built_4	3.842e+04	2850.617	13.476	0.000	3.28e+04	4.4e+04

Graphs to Show Fitting of Final Model

Q-Q Plot



Price vs. Predicted Price Scatter Plot



Note: Scatter plot shows model seems less predictively accurate at higher prices

Possible Additional Steps

- For area-related categories such as living square footage, basement square footage, and lot square footage, the data is likely influenced as to whether a home is in an urban, suburban, or rural area. An apartment in Seattle is different than a rural farmhouse. I would develop more useful models for different zip codes.
- See if older houses being worth more is correlated to zip codes. I suspect that this finding is related to urban areas with older housing stock.
- Dive into the engineered feature of “active market score” and investigate more how these scores affect prices.

Sources

- King County housing dataset can be found:
  - https://www.kaggle.com/harlfoxem/housesalesprediction
- Data from the US Census can be found:
  - https://data.census.gov

Contact

Henry Alpert
halpert3@gmail.com

LinkedIn: https://www.linkedin.com/in/henryalpert/

GitHub Project Repo:
https://github.com/halpert3/flatiron-mod2-project