SethKauf / King_County_Avocoders_Group_3

Code Issues Pull requests Actions Projects Wiki Security

main

King_County_Avocoders_Group_3 / README.md

SethKauf Update README.md History

1 contributor

```
110 lines (73 sloc) | 4.7 KB
```

King County Real Estate

Image courtesy of iStock

Overview

We looked at the information on homes sold in King County, WA between May 2014 and May 2015 to create a predictive pricing model.

To see our final notebook, click here.

Business Problem

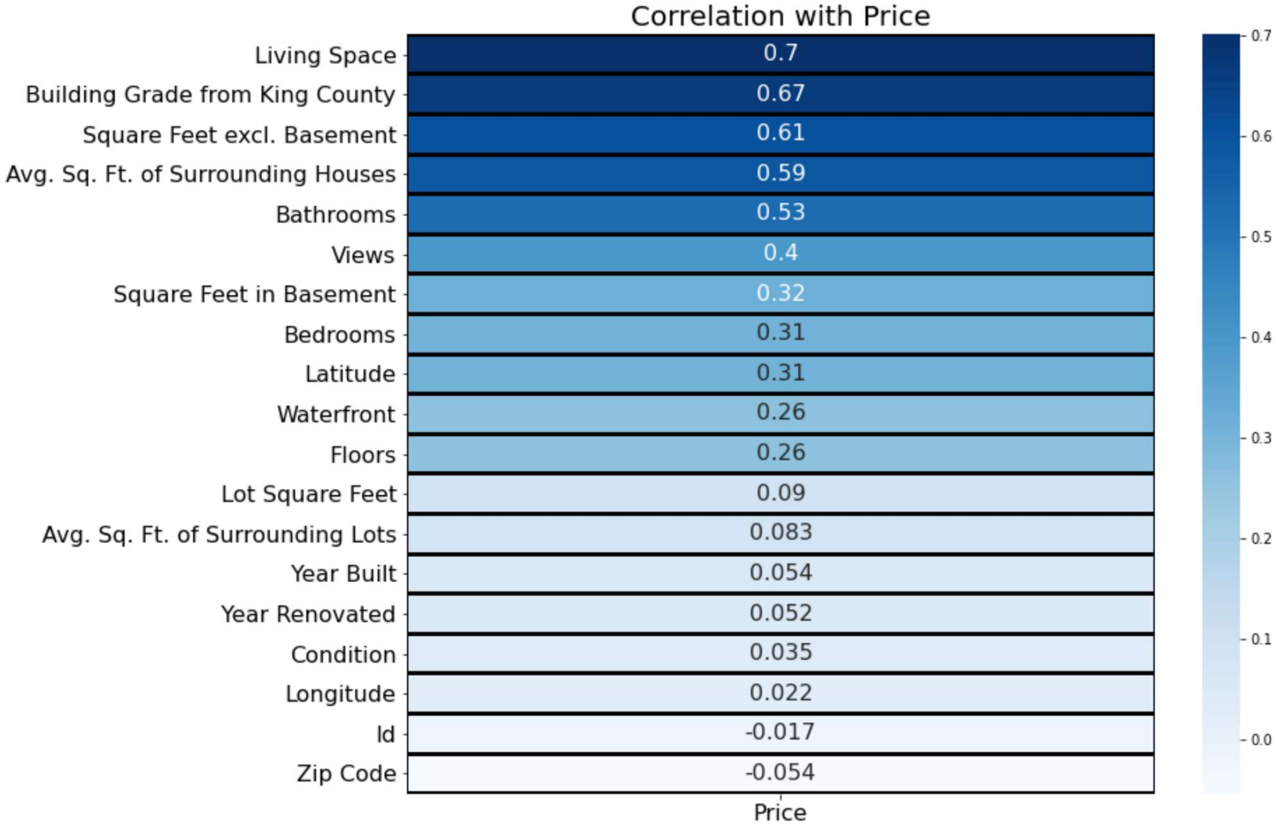
A real estate company in Seattle, WA is listing homes on their website.

They want to develop a model that will give a good ball-park estimate of the house's price before listing.
Using the information we have from the King County database, what would be an accurate predictor of pricing for these homes?

Selecting the Target, determining our methods

- Because the model's goal is to predict price, that will be the target
- We will use simple linear regression to test multiple linear models

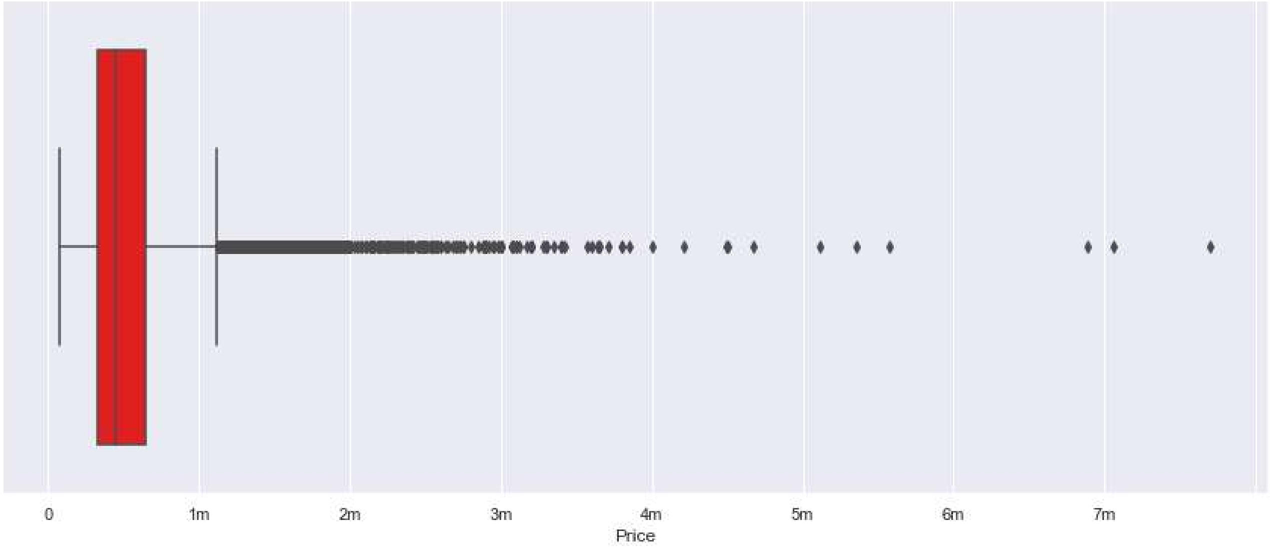
First we looked at a Heatmap for correlations between Price and all features



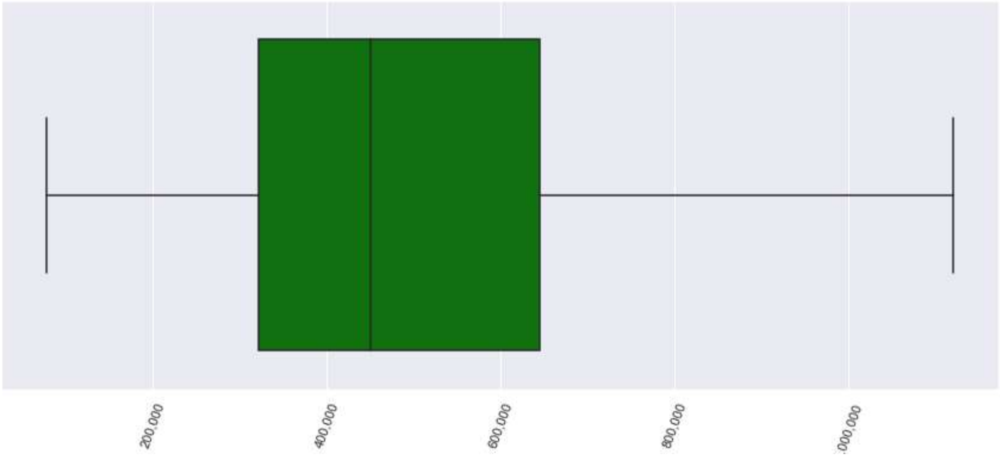
We noticed the Living Space feature (squarefootage of the homes) has the highest correlation to our target, followed by the Grade feature.

Exploring Price Data

Let's first look at the full price data in a boxplot.



This is no-good. Let's remove outliers.



A lot of the data falls from about $70,000 to about $1,200,000, we will stratify our data on this parameter.

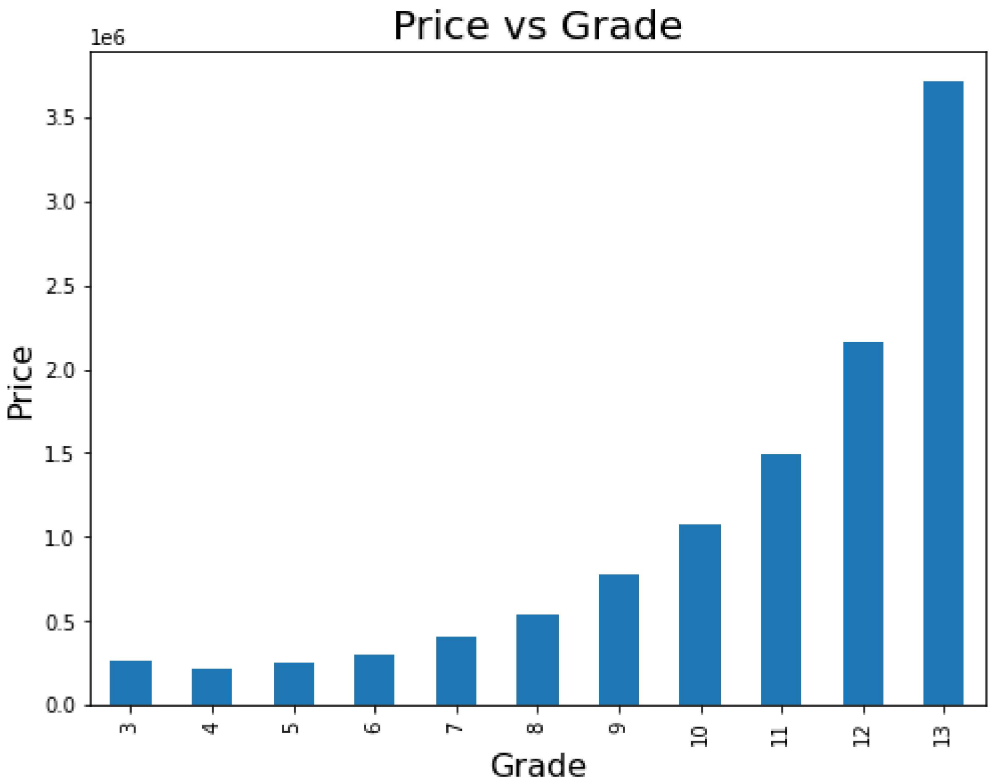
Exploring Living Space Data

The highest correlated feature was living space.
Our first model tested directly tested this onto Price, but it had nearly no effect.

Exploring Grade Feature

Grade is a feature that helped us better stratify the housing prices in our dataset.

It has a clear upward trend as seen below.

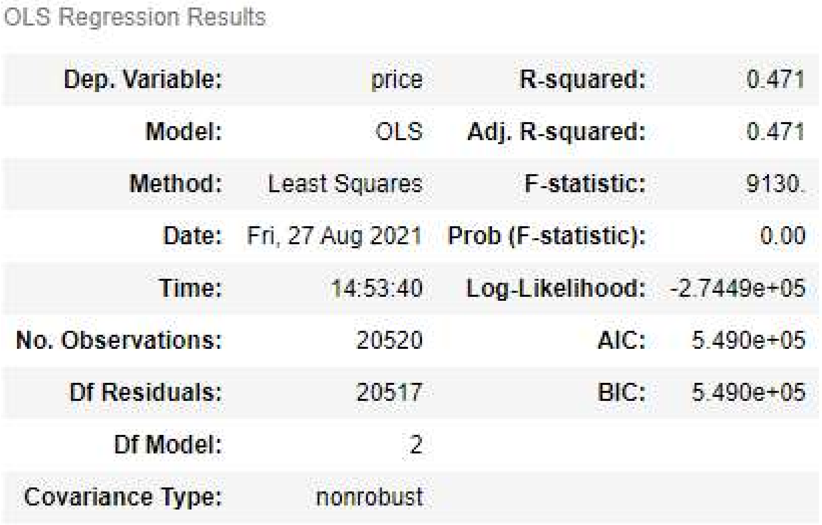


We then tested it with the similar train-test-split model from Living Space.

Unfortunately, it also did not return anything substantial.

First Model, Target ~ Two Highest

We created an OLS model using the top two features.



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.471
Model:	OLS	**Adj. R-squared:**	0.471
Method:	Least Squares	**F-statistic:**	9130.
Date:	Fri, 27 Aug 2021	**Prob (F-statistic):**	0.00
Time:	14:53:40	**Log-Likelihood:**	-2.7449e+05
No. Observations:	20520	**AIC:**	5.490e+05
Df Residuals:	20517	**BIC:**	5.490e+05
Df Model:	2		
Covariance Type:	nonrobust		

Although the r^2 is still low, it's already a little better.

Feature Engineering

Grade

For Grade, we created dummy variables as numeric stand-ins, so we can add Grade to our upcoming model.

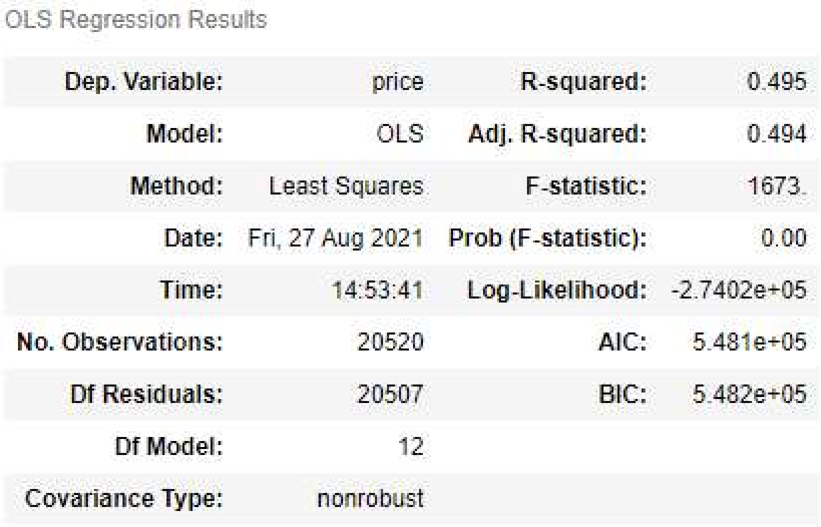
Zip Code

We grouped the zip codes together and used their average price per zip code in place of the zip code itself.

Second Model, train_test_split on First Few Features.

We ran our train_test_split on several features, notably Living Space, Waterfront, Zip Code, and Grades as dummy variables.

It only improved slightly as seen below, but we seemed to be on the right track.



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.495
Model:	OLS	**Adj. R-squared:**	0.494
Method:	Least Squares	**F-statistic:**	1673.
Date:	Fri, 27 Aug 2021	**Prob (F-statistic):**	0.00
Time:	14:53:41	**Log-Likelihood:**	-2.7402e+05
No. Observations:	20520	**AIC:**	5.481e+05
Df Residuals:	20507	**BIC:**	5.482e+05
Df Model:	12		
Covariance Type:	nonrobust		

Final Model

First, we stratified price to select houses at below $1,200,000, as mentioned earlier.

We kept some of our previously engineered features such as Zip Code price means and the Grade columns.

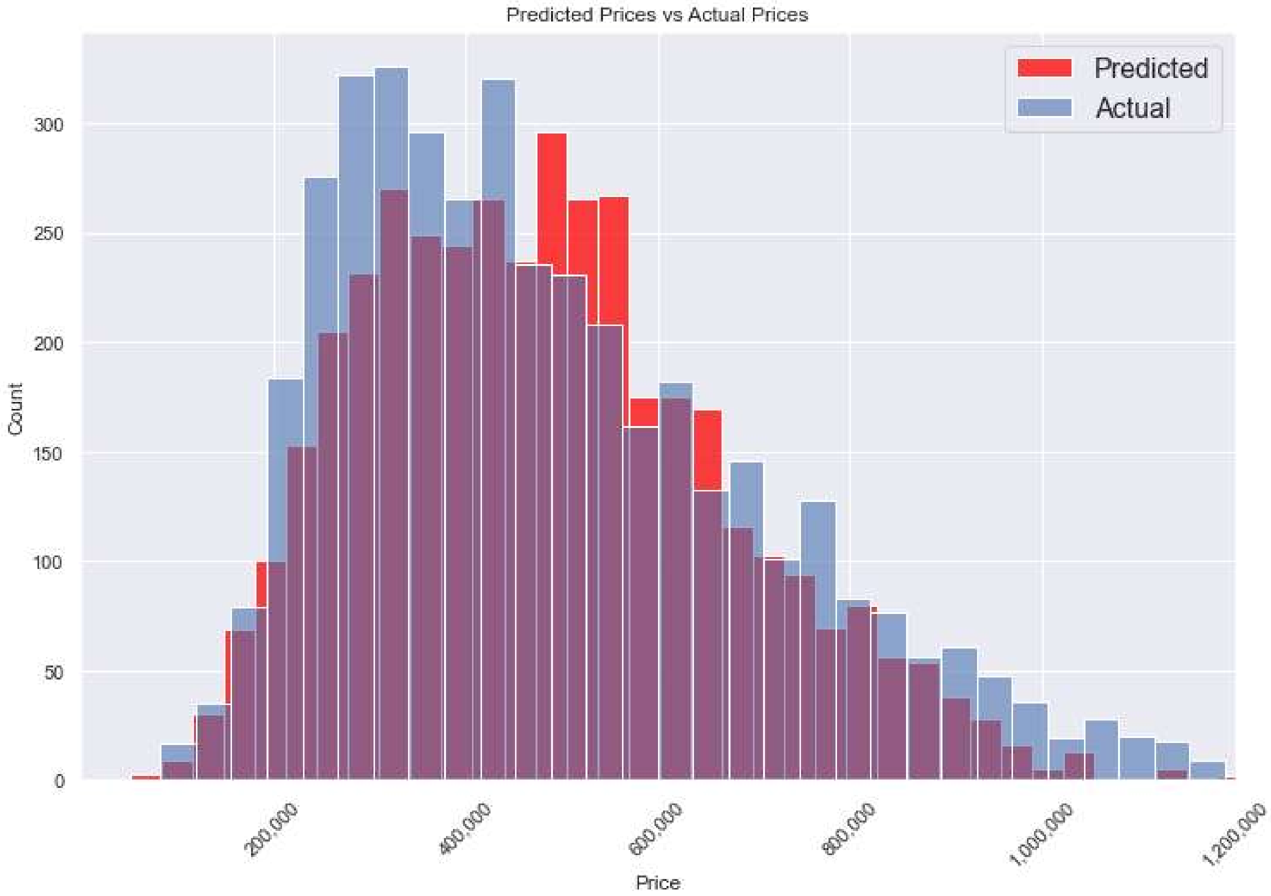
Here are the features we ended up using in our final model:
* Bathrooms
* Bedrooms
* Grade (4-11, excl. 7)
* House Age
* Latitude
* Living Space
* Longitude
* Waterfront property

* Year Renovated

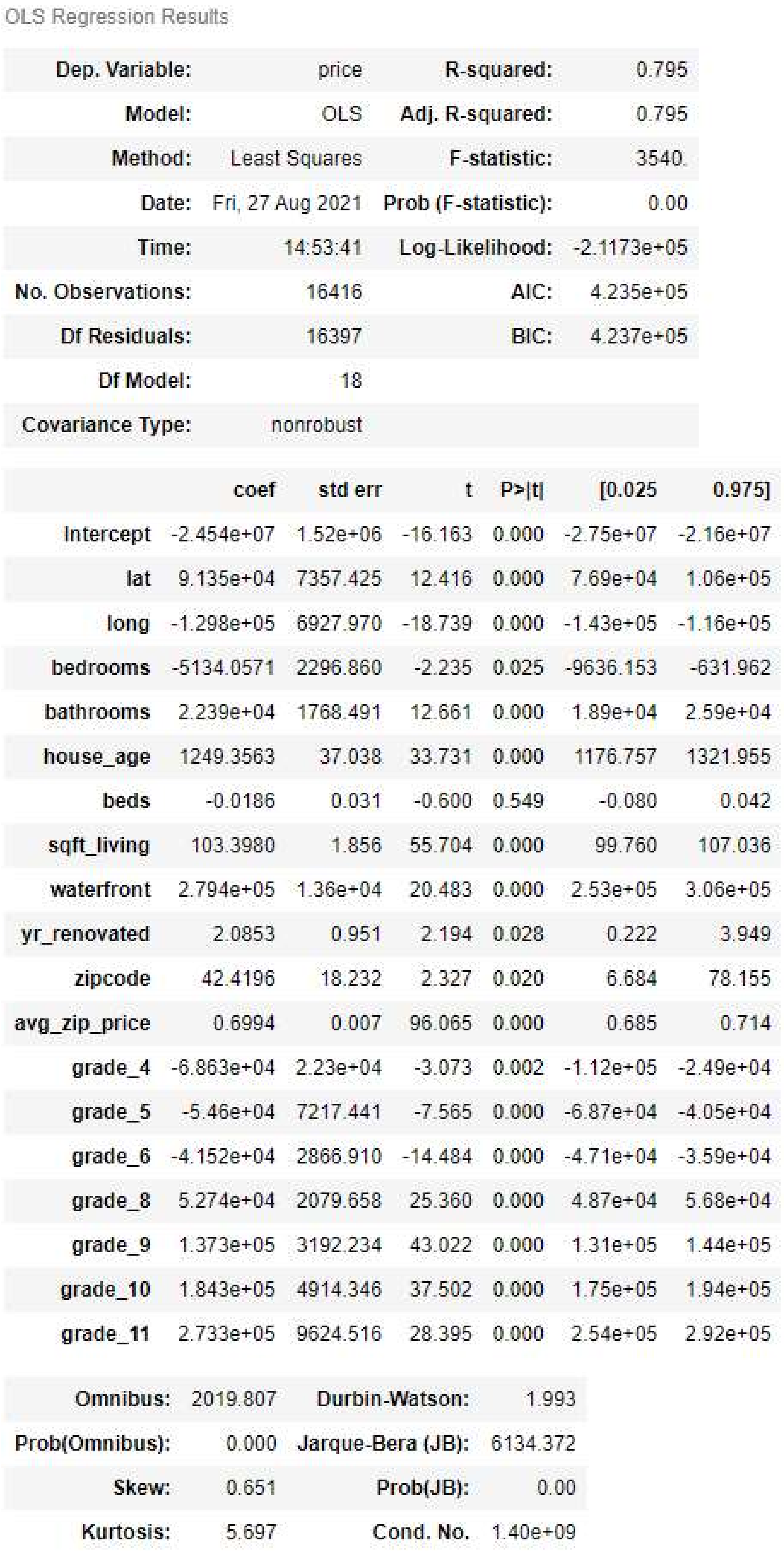
* Zip Code
* Zip Code price means

It returned an 80% effective model with a validation score of 80%.

Below is a graph of the actual prices transposed on the predicted prices.



Below is also a screengrab of the final model scores in OLS (non-scaled). The full model breakdown is in the final notebook.



OLS Regression Results

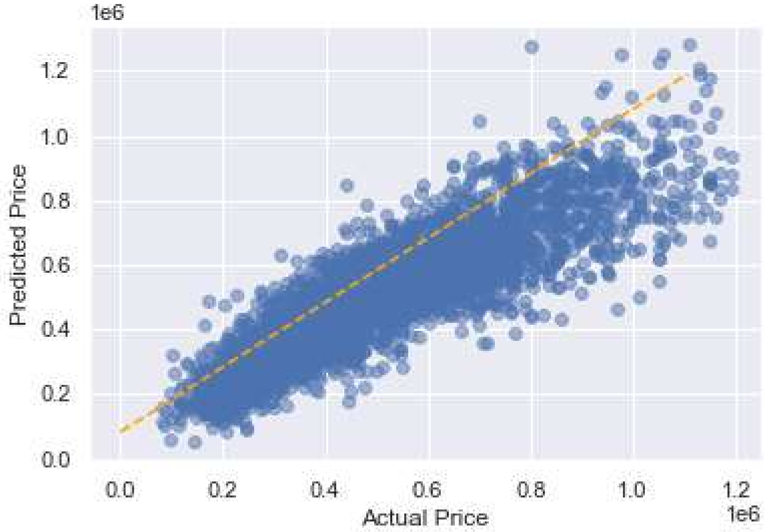
Dep. Variable:	price	**R-squared:**	0.795
Model:	OLS	**Adj. R-squared:**	0.795
Method:	Least Squares	**F-statistic:**	3540.
Date:	Fri, 27 Aug 2021	**Prob (F-statistic):**	0.00
Time:	14:53:41	**Log-Likelihood:**	-2.1173e+05
No. Observations:	16416	**AIC:**	4.235e+05
Df Residuals:	16397	**BIC:**	4.237e+05
Df Model:	18		
Covariance Type:	nonrobust		

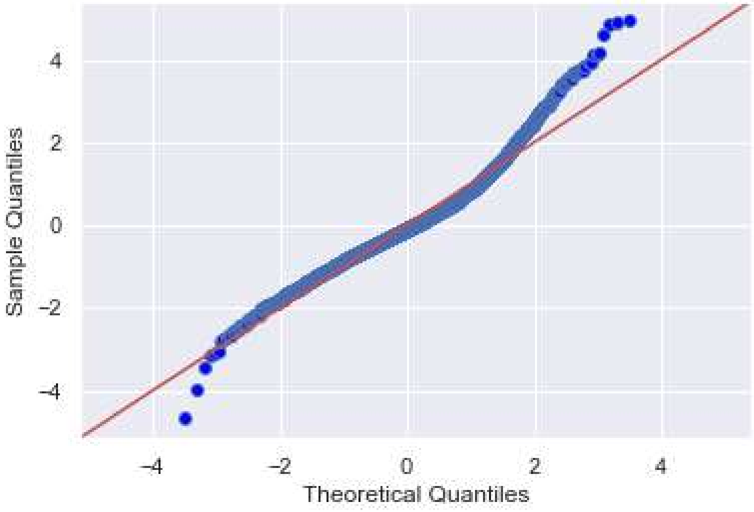
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-2.454e+07	1.52e+06	-16.163	0.000	-2.75e+07	-2.16e+07
lat	9.135e+04	7357.425	12.416	0.000	7.69e+04	1.06e+05
long	-1.298e+05	6927.970	-18.739	0.000	-1.43e+05	-1.16e+05
bedrooms	-5134.0571	2296.860	-2.235	0.025	-9636.153	-631.962
bathrooms	2.239e+04	1768.491	12.661	0.000	1.89e+04	2.59e+04
house_age	1249.3563	37.038	33.731	0.000	1176.757	1321.955
beds	-0.0186	0.031	-0.600	0.549	-0.080	0.042
sqft_living	103.3980	1.856	55.704	0.000	99.760	107.036
waterfront	2.794e+05	1.36e+04	20.483	0.000	2.53e+05	3.06e+05
yr_renovated	2.0853	0.951	2.194	0.028	0.222	3.949
zipcode	42.4196	18.232	2.327	0.020	6.684	78.155
avg_zip_price	0.6994	0.007	96.065	0.000	0.685	0.714
grade_4	-6.863e+04	2.23e+04	-3.073	0.002	-1.12e+05	-2.49e+04
grade_5	-5.46e+04	7217.441	-7.565	0.000	-6.87e+04	-4.05e+04
grade_6	-4.152e+04	2866.910	-14.484	0.000	-4.71e+04	-3.59e+04
grade_8	5.274e+04	2079.658	25.360	0.000	4.87e+04	5.68e+04
grade_9	1.373e+05	3192.234	43.022	0.000	1.31e+05	1.44e+05
grade_10	1.843e+05	4914.346	37.502	0.000	1.75e+05	1.94e+05
grade_11	2.733e+05	9624.516	28.395	0.000	2.54e+05	2.92e+05

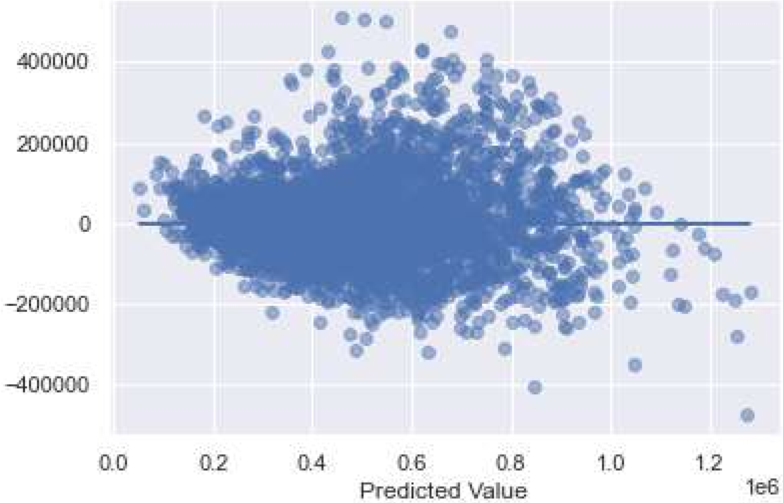
Omnibus:	2019.807	**Durbin-Watson:**	1.993
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	6134.372
Skew:	0.651	**Prob(JB):**	0.00
Kurtosis:	5.697	**Cond. No.**	1.40e+09

Regression Tests

Finally, our tests for Linearity, Normality, and Homoscedasticity







Our tests show our model, although still somewhat affected by outliers, is mostly good for a linear regression model.

Next Steps

We recommend that future iterations of the model look into other features from the original dataset, such as Latitude and Longitude, or the Year Renovated and Year Built, unscaled.

We also recommend adding other features, such as number of houses on the market in the area, the "Walking score", proximity to school(s), etc.