King County Housing Prices Linear Regression Model

Meredith Newhouse

Overview and Purpose

I want to know what features of a house like location, number of bedrooms, size, etc. impact the price of the house so that I can make the most strategic choices when building a house in King County.

To do this I will create a linear regression model to predict the price increase or decrease of a house in King County, based on specific features.

King County includes the city of Seattle and is the most populous county in Washington State.

The Data

The data used is King County housing data provided by Flatiron School

There are 21,597 houses included in this data

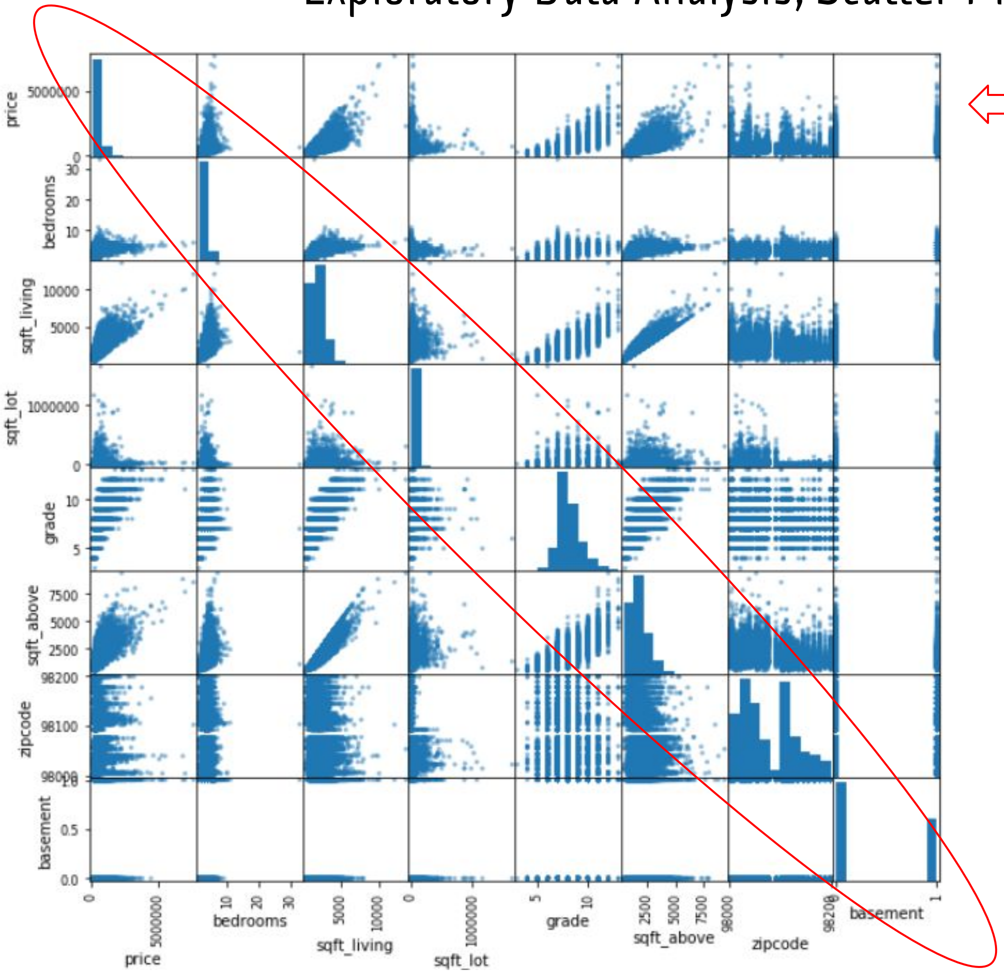
The prices range from $78,000 to $7,700,000

Definitions of key features used

- Date - house was sold
- Price - is prediction target
- bedrooms - of Bedrooms/House
- sqft_living - footage of the home
- sqft_lot - footage of the lot
- waterfront - House which has a view to a waterfront
- view - Has been viewed
- grade - overall grade given to the housing unit, based on King County grading system
- sqft_basement - square footage of the basement
- zipcode - zip
- sqft_lot15 - The square footage of the land lots of the nearest 15 neighbors

★ A 'basement' column was created from the sqft_basement data to indicate whether or not a home had a basement

Exploratory Data Analysis, Scatter Plot Matrix



From this plot I can see which variables, with price on the y axis, may have linear relationships with price.

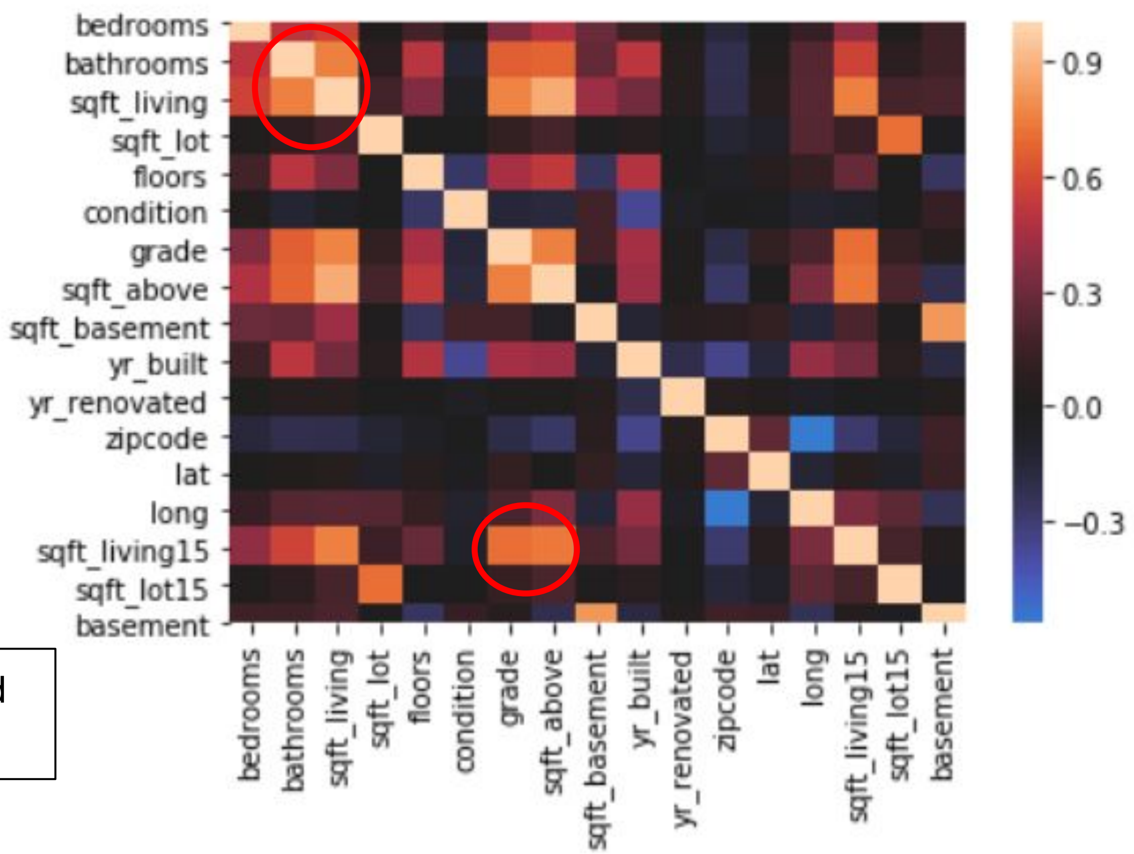
- On initial look, the variables with potential linear relationships with price are, bedrooms, sqft_living, sqft_living15, grade

I can also look on the diagonal and see histograms showing the distribution for each of the variables.

- I can see that variables like grade are very close to normally distributed
- Price, bedrooms, bathrooms, sqft_living all look to be right skewed.

EDA cont. Multicollinearity

pairs	cc
(sqft_living, sqft_above)	0.876448
(basement, sqft_basement)	0.820893
(sqft_living, grade)	0.762779
(sqft_living, sqft_living15)	0.756402
(sqft_above, grade)	0.756073
(sqft_living, bathrooms)	0.755758

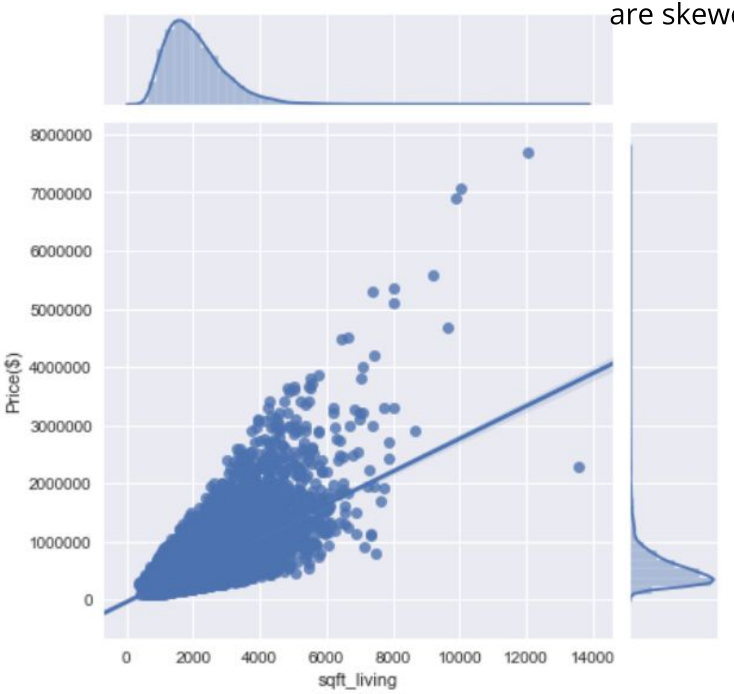


It looks like sqft_living, sqft_above, and grade are the most correlated

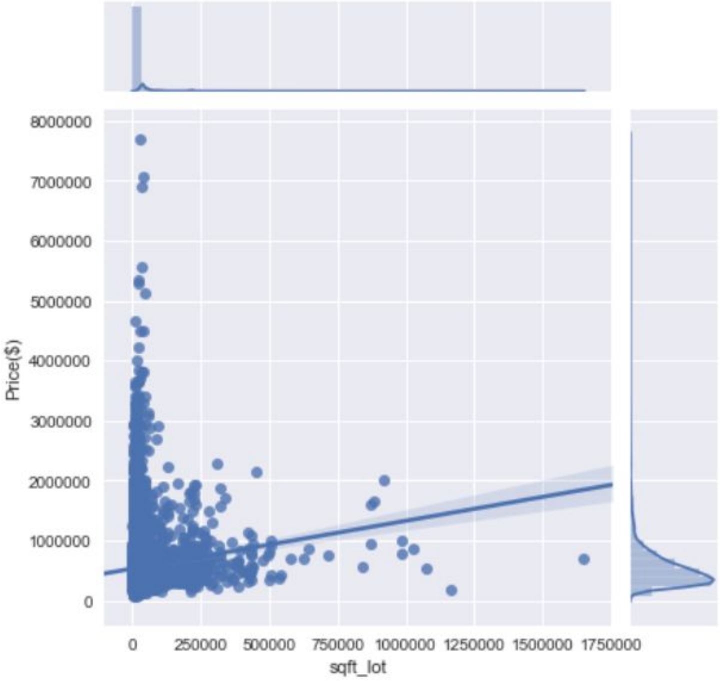
EDA cont. Continuous Variables and Linear Relationships

These two graphs indicate potential linear relationships between price and home size and price and lot size. I can also see that for both variables their distributions are skewed to the right

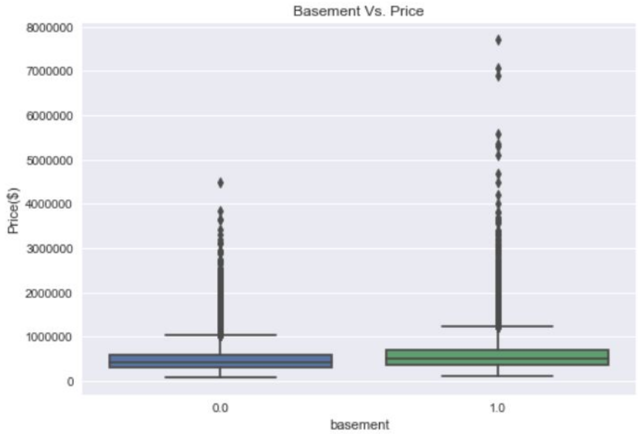
Home Size Vs. Price

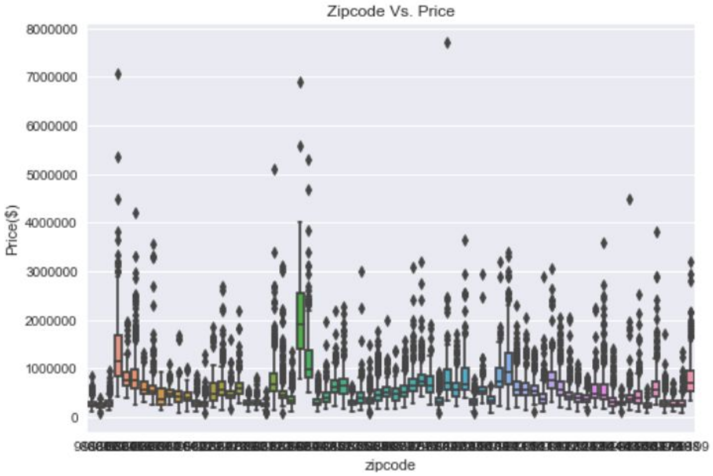


Lot Size vs. Price

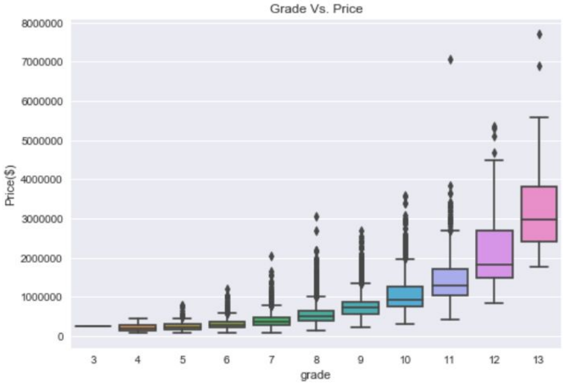


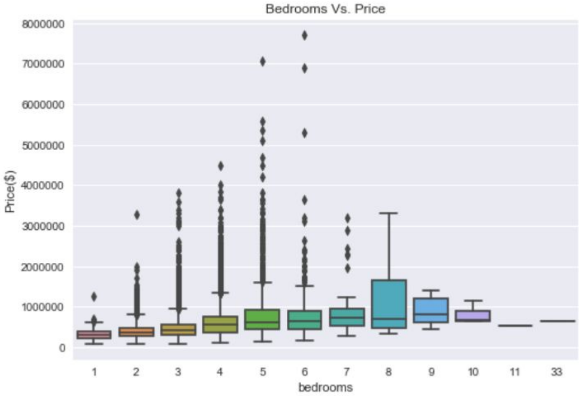
EDA cont. Categorical Variables





With these four graphs I can see the distributions of each variable against price. All the variables look to be at least moderately impacted by price.





Baseline Model

The first run of the model resulted in an R^2 value of .83 and an RMSE of 151374.63. It is clear that despite the high R-squared value, the corresponding high RMSE and the features with high p-values make the model not a good fit.

Dep. Variable:	price	**R-squared:**	0.830
Model:	OLS	**Adj. R-squared:**	0.829
Method:	Least Squares	**F-statistic:**	726.6
Date:	Tue, 20 Oct 2020	**Prob (F-statistic):**	0.00
Time:	15:09:58	**Log-Likelihood:**	-2.1617e+05
No. Observations:	16197	**AIC:**	4.326e+05
Df Residuals:	16088	**BIC:**	4.334e+05
Df Model:	108		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.77e+05	2.06e+05	3.282	0.001	2.73e+05	1.08e+06
bathrooms	2.8e+04	2927.760	9.563	0.000	2.23e+04	3.37e+04
sqft_living	81.3683	15.739	5.170	0.000	50.518	112.218
sqft_lot	0.2424	0.043	5.673	0.000	0.159	0.326
floors	-3.149e+04	3490.581	-9.023	0.000	-3.83e+04	-2.47e+04
condition	2.925e+04	2103.736	13.904	0.000	2.51e+04	3.34e+04
sqft_above	86.6973	15.740	5.508	0.000	55.845	117.550
sqft_basement	40.9229	16.298	2.511	0.012	8.978	72.868
yr_built	-342.0894	71.012	-4.817	0.000	-481.280	-202.898
yr_renovated	34.1960	3.494	9.788	0.000	27.348	41.044
sqft_living15	17.2628	3.204	5.387	0.000	10.982	23.544
sqft_lot15	-0.1770	0.068	-2.611	0.009	-0.310	-0.044
bedrooms_2	9839.5991	1.33e+04	0.741	0.458	-1.62e+04	3.59e+04
bedrooms_3	8297.8915	1.33e+04	0.626	0.531	-1.77e+04	3.43e+04
bedrooms_4	-1.6e+04	1.35e+04	-1.181	0.238	-4.26e+04	1.06e+04
bedrooms_5	-3.062e+04	1.43e+04	-2.135	0.033	-5.87e+04	-2504.076
bedrooms_6	-4.239e+04	1.75e+04	-2.422	0.015	-7.67e+04	-8089.585
bedrooms_7	-1.162e+05	3.23e+04	-3.597	0.000	-1.8e+05	-5.29e+04
bedrooms_8	-6.486e+04	5.58e+04	-1.163	0.245	-1.74e+05	4.45e+04

Iterative Process

- I began my modeling process by removing all features with a p-value greater than .05
- Then I log transformed Price to help normalize the data
- I did more rounds of removing features with high p-values
- Then I dealt with multicollinearity and removed highly correlated features.
- Many features were correlated which led me to remove a lot of features, which ultimately dropped the R-squared value significantly
- I log transformed my continuous variable features
- Finally, I did one more round of dropping features with high p-values

Final Model

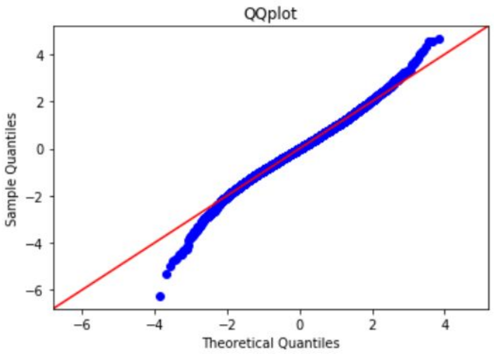
The final (log-transformed) RMSE of the model is .302
The final R-squared is .670. Though the R-squared is lower than the original model, all the feature p-values are 0 and the RMSE is relatively low.

Dep. Variable:	price	**R-squared:**	0.670
Model:	OLS	**Adj. R-squared:**	0.669
Method:	Least Squares	**F-statistic:**	475.4
Date:	Wed, 21 Oct 2020	**Prob (F-statistic):**	0.00
Time:	13:26:00	**Log-Likelihood:**	-3601.6
No. Observations:	16197	**AIC:**	7343.
Df Residuals:	16127	**BIC:**	7882.
Df Model:	69		
Covariance Type:	nonrobust		

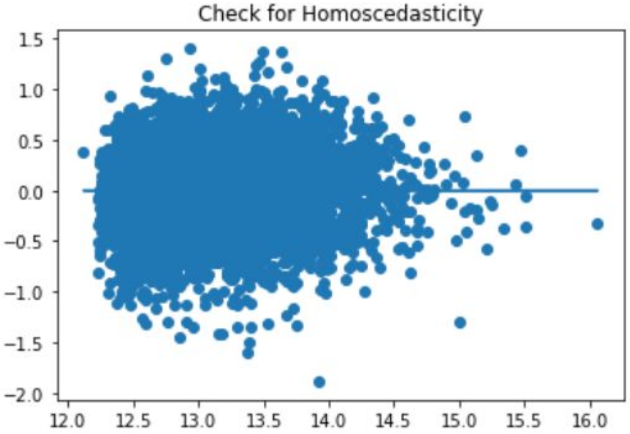
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.3202	0.031	364.984	0.000	11.259	11.381
sqft_lot	0.1274	0.003	38.383	0.000	0.121	0.134
yr_renovated	4.275e-05	6.6e-06	6.476	0.000	2.98e-05	5.57e-05
bedrooms_5	0.1786	0.009	19.069	0.000	0.160	0.197
bedrooms_6	0.1676	0.021	7.819	0.000	0.126	0.210
bedrooms_7	0.1984	0.057	3.451	0.001	0.086	0.311
waterfront_1	0.4282	0.036	11.998	0.000	0.358	0.498
view_1	0.2538	0.020	12.925	0.000	0.215	0.292
view_2	0.2704	0.012	22.722	0.000	0.247	0.294
view_3	0.4384	0.016	27.433	0.000	0.407	0.470
view_4	0.5885	0.024	24.175	0.000	0.541	0.636
view_unknown	0.1026	0.045	2.264	0.024	0.014	0.192
grade_12	0.6378	0.038	16.902	0.000	0.564	0.712
grade_13	0.9681	0.097	9.990	0.000	0.778	1.158
zipcode_98004	1.3528	0.021	65.545	0.000	1.312	1.393
zipcode_98005	0.9136	0.028	32.932	0.000	0.859	0.968
zipcode_98006	0.8615	0.017	51.126	0.000	0.828	0.895
zipcode_98007	0.7780	0.031	25.481	0.000	0.718	0.838
zipcode_98008	0.6502	0.022	30.200	0.000	0.608	0.692
zipcode_98010	0.1691	0.034	4.996	0.000	0.103	0.236
zipcode_98011	0.5105	0.025	20.317	0.000	0.461	0.560
zipcode_98014	0.1913	0.032	5.919	0.000	0.128	0.255
zipcode_98019	0.3121	0.027	11.656	0.000	0.260	0.365

Interpretation

Linear Model



The data in the model follow the normality line relatively well, although not perfectly



Homoscedasticity is not perfectly met with this data. This indicates that a linear regression model may not be the best model for the data

R-squared - .670, higher R-squared values represent smaller differences in the observed data and the fitted values created by the model

RMSE - The average error of the model is about 1.35 dollars of log price.

Key Coefficients:

As lot size increases by 1%, price increases .1274%. For every 20% increase in lot size price increases 2.35%

Having a waterfront property increases log price by $1.53

Having a basement increases log price by $1.07

Having 5 bedrooms increases log price by $1.20 and 7 bedrooms increases log price by $1.22

Having a high grade of 13 increases log price by $2.63

The zipcodes with the highest log price increase are 98004 at $3.87 , 98039 at $5.12, 98112 at $3.60

Conclusions

A linear model may not be the best model for this data.

However, I can see that waterfront properties and properties in certain areas increase the price of the house. The zipcodes that increased price the most included Seattle, a major city, and waterfront neighborhoods.

More bedrooms increase price, but past 5 bedrooms the increase is small.

Having a high grade, based on the King County grading system, is very important. Looking into the grading system will be key when building a house.

Perhaps building a house with a larger lot size and including a basement, in a less expensive zipcode may be beneficial to increase value of the house without the extreme cost certain neighborhoods entail.

Next Steps

Further analysis in exploring interactions between data such as lot size and the land lots of the nearest 15 neighbors could be helpful.

Integrating polynomial regression may help create a model with a better fit.

Further understanding of what features affect the grading system in King County.

Thank You

Check out my github repo here:
https://github.com/newhousem/Linear-Regression-Project

Contact me: meredithnewhouse@gmail.com

Thank you to Flatiron for providing the data sets used in this analysis and Yish for helping to answer all of my questions.