King County Real Estate

David Rasmussen

October 2020

Contents

1. The Data

The Data Provided

Notes:

- Outliers and extraneous data were removed.
- Null values were replaced with the mode values were appropriate.

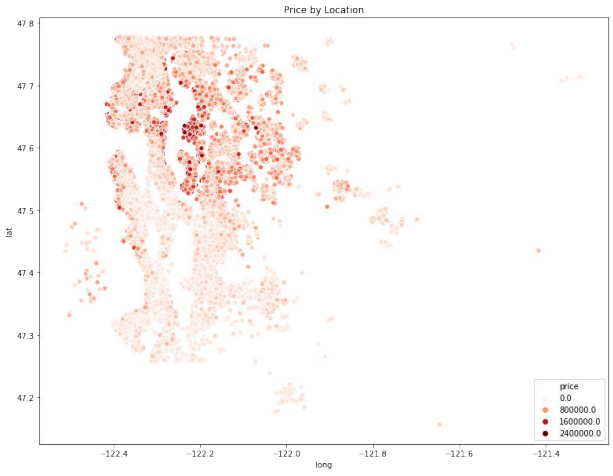
King County Real Estate Data

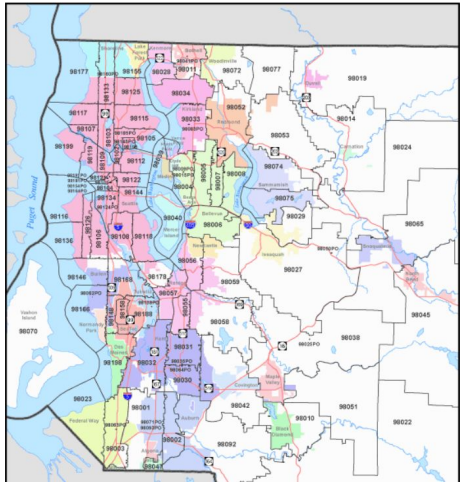
- Date the property sold
- The price the property sold at
- The number of bedrooms and bathrooms
- Square footage information
- If the property has a view of the waterfront
- What the condition and grade the property is in
- When the property was built and if/when it had been renovated
- The latitude and longitude of each property

2. Insights and Regression Model

What are my findings?

Location, Location, Location



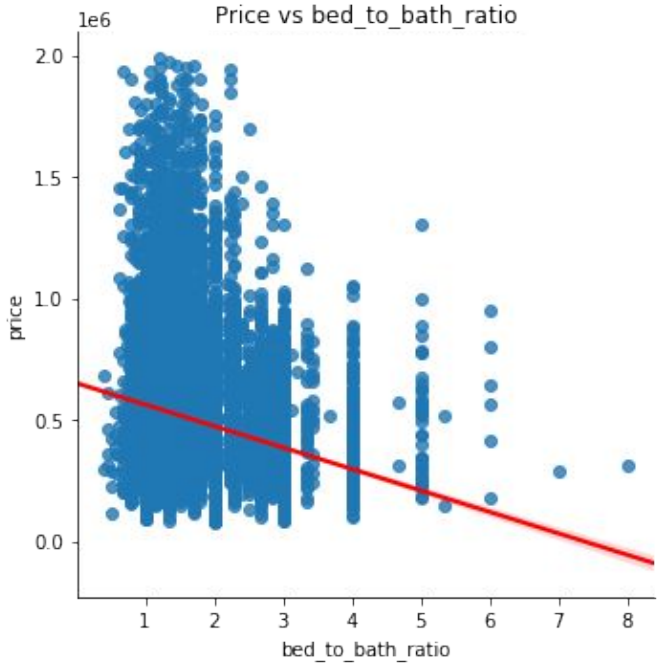


Which locations in King Country are the most expensive?

I used the latitude and longitude data in concert with price highlight the most expensive locations. Findings

- The most expensive real estate is clustered around Lake Washington and the Seattle proper
- The specific zip codes which the most expensive real estate are: 98033, 98039, 98004 and 98040

Bedrooms per Bathroom ratio vs Price



Graph depicts the relationship between number of bedrooms that share a bathroom versus price.

Insights:

- As expected there is a discernible negative relationships. More bedrooms per bathroom equates to lower price
- The correlation between the two variables: -0.23

Which Variables Most Impact Price?

See Appendix A for correlation matrix of all the variables with at least a correlation absolute value of 0.20.

Insights:

- The ‘grade’ variable has the highest correlation at 0.63. Grade is categorical and based on a scale between 1-13. The higher the grade, the higher the price
- The matrix highlights the how the variables are inter-correlated which is important when considering multicollinearity.

Predictive Regression Model

After removing variables to reduce multicollinearity and model noise I chose the following independent variables to predict price:

- The quality ‘grade’ of the property
- Bedroom to bathroom ratio
- Dummy variable for the following zip codes: 98033, 98039, 98004 and 98040.

SEE APPENDIX B FOR REGRESSION RESULTS

Insights:

- Low p-values indicate that each predictive variable is significant
- Coefficient of determination (R-squared) of 87% indicates high explanatory value

Notes:

- A Durbin-Watson of 2 means that there is no autocorrelation amongst the residuals.
- The high Jarque-Bera indicates that the residuals are not normally distributed which is an issue.
- Based on q-q plot, significant heteroskedasticity exists. The residuals vary with price. There is more noise in the model as price increases.
- I would like to do further work to address the non-linearity of the model. Based on the q-q plot and a graph of the residuals (both in appendix C) it appears the the independent variables have a non-linear impact on price.

3. Further Analysis

I list a few topics which I would like to explore if given the time and data.

Further Analysis

Time Series Data

To determine pricing trends and how changes in variable impact price

Demographics

Where is the most attractive location for young professionals, families and retirees.

Economic Data

How sensitive are prices to the business cycle

School Districts

How each school district impacts the relative value of real estate

College Campuses

The impact college campuses have on local pricing

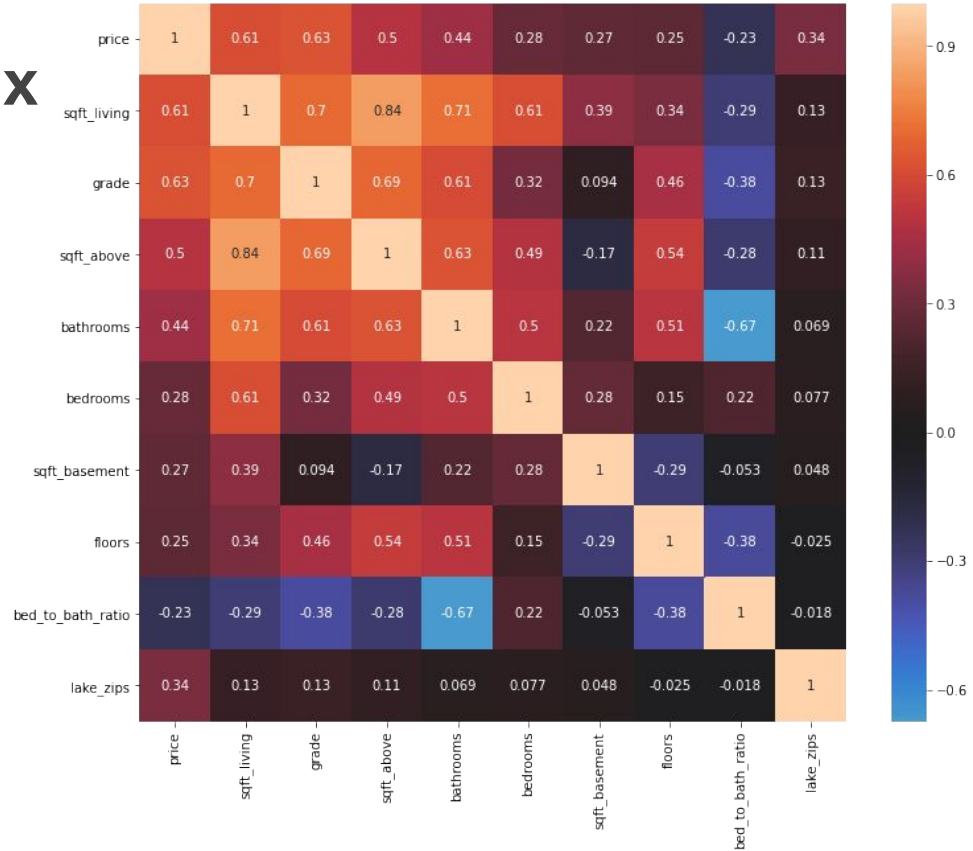
Stock Price Impact

Local large tech and the impact on housing prices

Thanks!

Does anyone have any questions?

Appendix A: Correlation Matrix



Appendix B: Regression Result

OLS Regression Results

Dep. Variable:	price	R-squared (uncentered):	0.870
Model:	OLS	Adj. R-squared (uncentered):	0.870
Method:	Least Squares	F-statistic:	4.290e+04
Date:	Thu, 29 Oct 2020	Prob (F-statistic):	0.00
Time:	08:54:53	Log-Likelihood:	-2.6306e+05
No. Observations:	19301	AIC:	5.261e+05
Df Residuals:	19298	BIC:	5.261e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
grade	8.042e+04	466.968	172.208	0.000	7.95e+04	8.13e+04
bed_to_bath_ratio	-6.779e+04	1850.122	-36.642	0.000	-7.14e+04	-6.42e+04
lake_zips	3.614e+05	6971.702	51.839	0.000	3.48e+05	3.75e+05

Omnibus:	6188.946	Durbin-Watson:	1.968
Prob(Omnibus):	0.000	Jarque-Bera (JB):	24366.702
Skew:	1.560	Prob(JB):	0.00
Kurtosis:	7.535	Cond. No.	37.5

Appendix C: Q-Q plot and graph of residuals