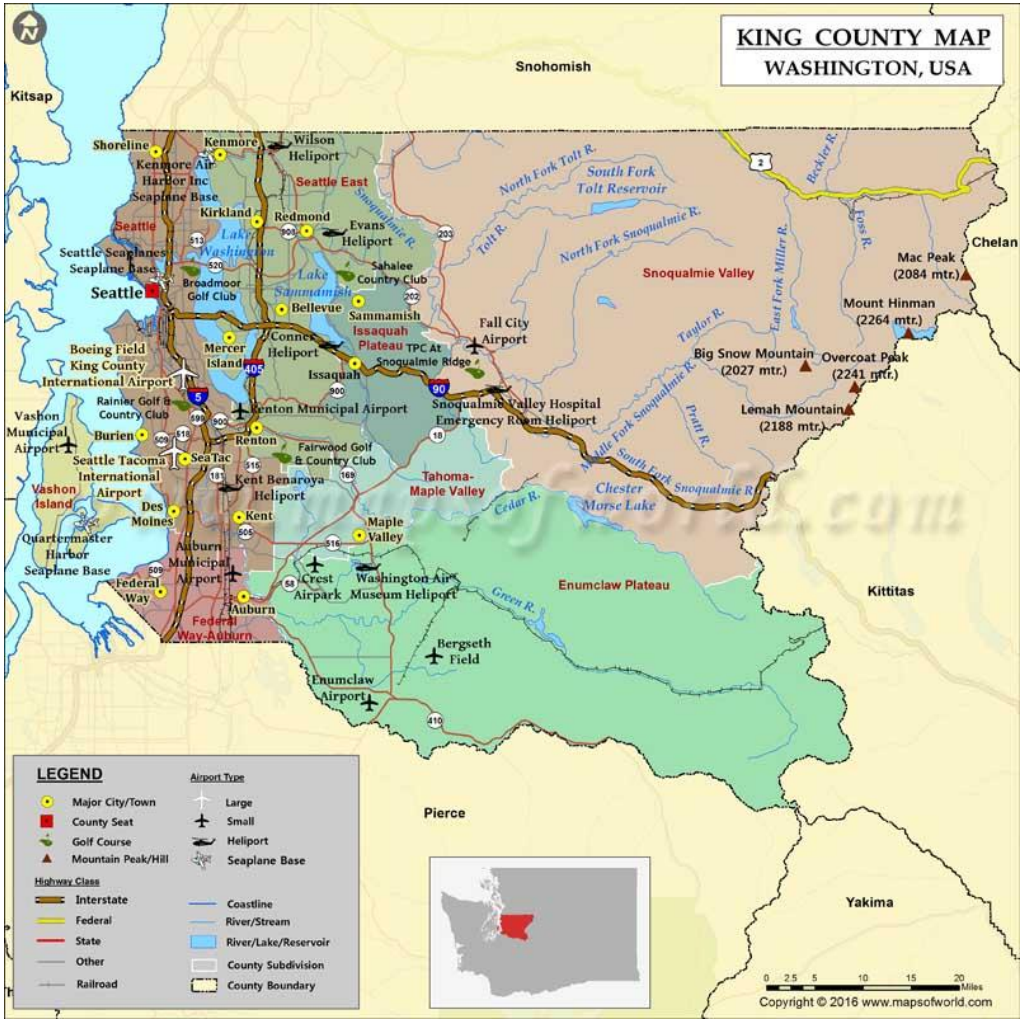
Analysing Home Prices in King County

By Dylan Garsee

Goals

- What are the most important features to consider when looking for a home?
- What features are most correlated with the price of a home?
- Finally, how much house can one expect to be able to purchase with their budget?



Preparing the Data

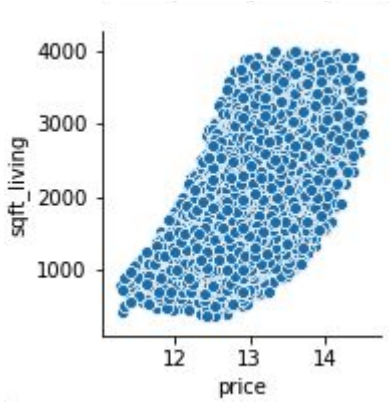
The dataset provided was very thorough, perhaps too thorough in the scope of this particular analysis. Some dropped data points include:

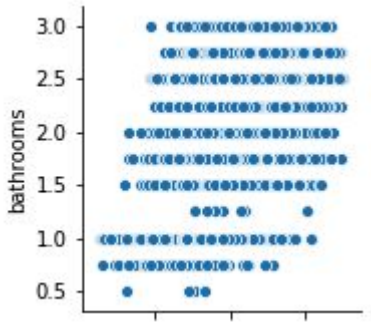
- Waterfront
- View
- Latitude/Longitude
- Square footage of the nearest 15 neighbors.

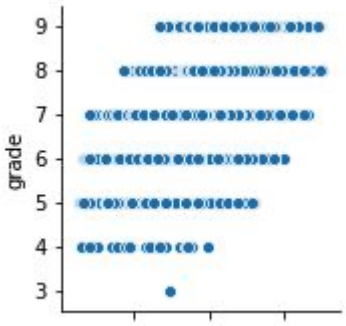
Checking for correlation

When building a model, correlated data works best so that it can be as accurate as possible. After processing the data, the top three datasets correlated with Price were:

1. Square footage of home
2. Grade given to home by King County
3. Number of Bathrooms

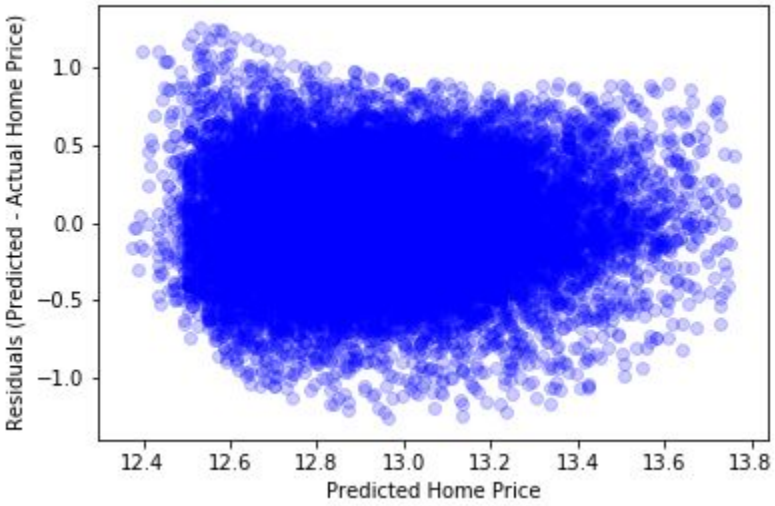






R-squared:	0.312

	coef
Intercept	12.2338
sqft_living	0.0004



First Model

With the data ready to go, the first model has indicated that we can explain 31% of the variance in the model, and that with every 1% increase in the budget, the square footage increases by 12.35.

Also testing for heteroscedasticity, it seems the model over-predicts at lower ends of the price.

Second and Third Models

After adding the second and third correlated features, the model became too unpredictable, with Square Footage all but disappearing and Number of Bathrooms trending negative.

Upon further analysis, the issue is mostly likely multicollinearity, meaning two or more features are linear with each other, thus throwing off the model. The VIF (variance inflation factor) has a threshold of 5, and any number higher than that strongly suggests multicollinearity.

With 18.65, 19, and 22, multicollinearity is almost certain.

	VIF	feature
0	18.650771	sqft_living
1	19.079247	grade
2	22.043320	bathrooms

Conclusion/Next Steps

The ultimate downfall of the model was multicollinearity, so upon revisiting, I would better address that.

I would also adjust my hypothesis. While the ultimate question came from a good place and promising early data, sometimes the numbers aren't there to support the hypothesis.

Thank you!