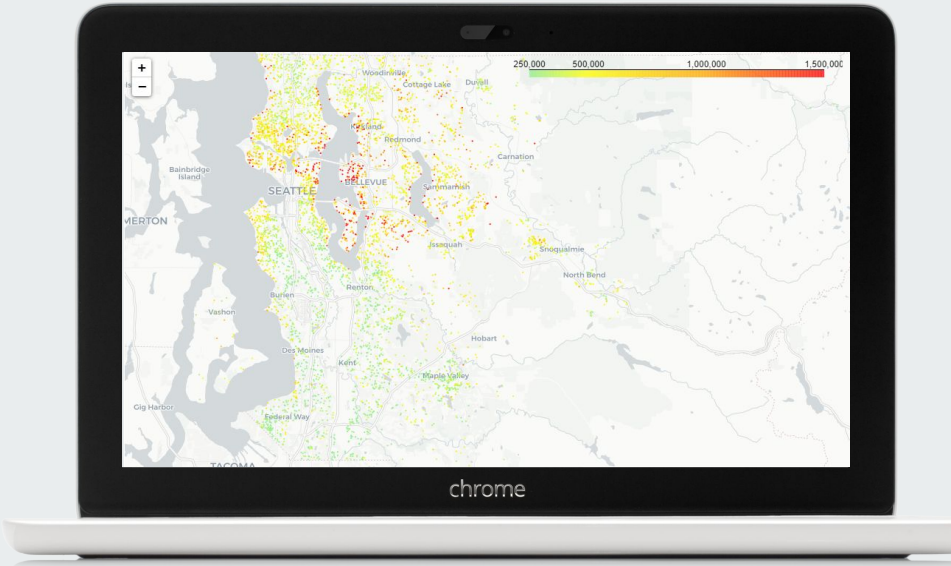
King County Housing Data

Predicting housing prices and informing real-estate investment.

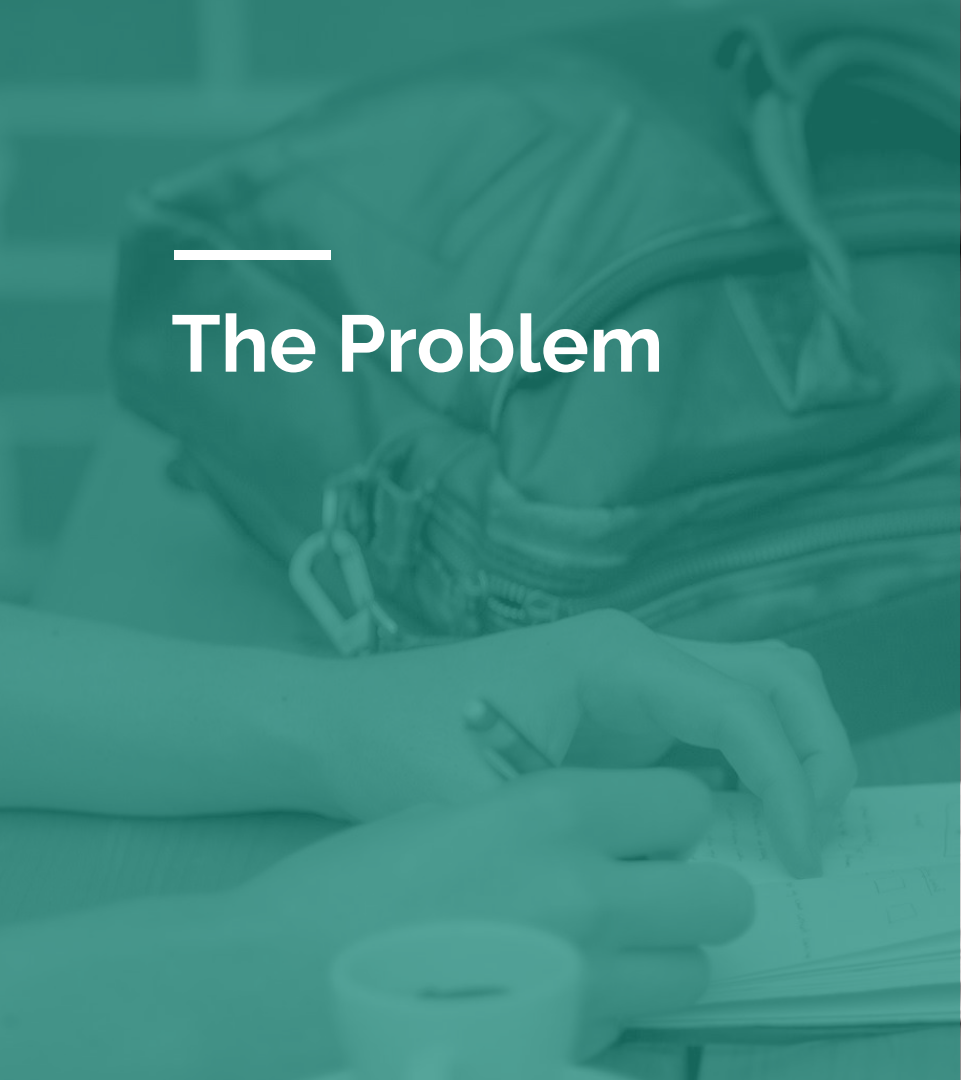


Outline

Context

It can be difficult to make real-estate investments.

People tend to be comfortable with aesthetic and geographic decisions about where to live, but can be a little lost when it comes to valuation and investment decisions. Maybe our model can be a tool to fill in some of these gaps?



The Problem

Buying a house is an overwhelming decision.

There are tons and tons of factors to consider. What if we had a model that interpreted large numbers of these factors to filter down the number of choices to just the most compelling options?

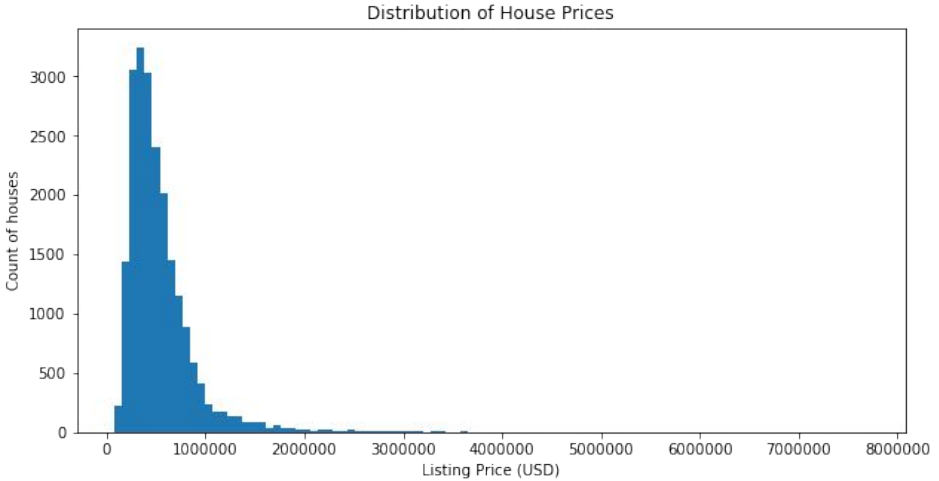
Our Data

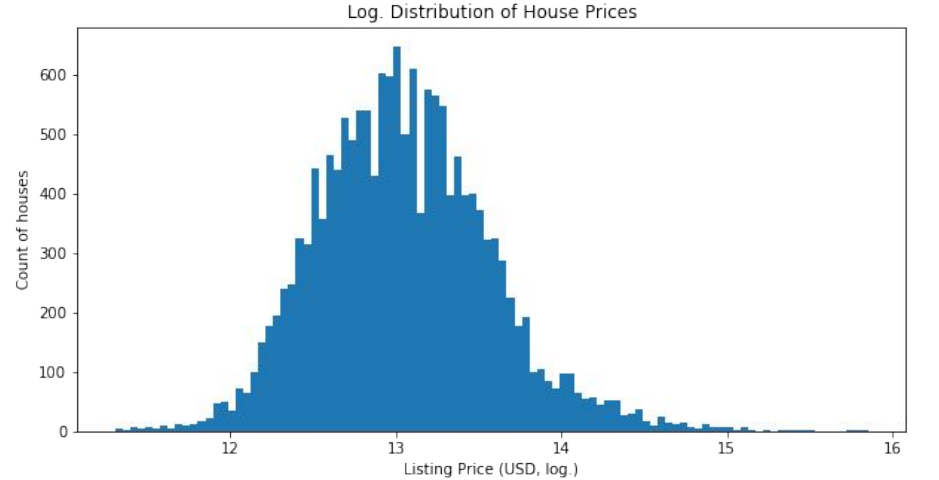
The data being used for this analysis is from the KC housing data, which contains 21597 rows and 21 columns of information on houses and prices from 2014 and 2015.

Some cleanup was performed to remove duplicates and fill empty values and process numeric, categoric, and spatial data.

Our Data

Here we can see the standard (top) and log (bottom) distributions of list price for houses. To increase the accuracy of the model, our numeric data was transformed to log numeric before being fed in. In log data, when we see a deviation of one, that is a 100% increase or decrease.





Building The Model

01

To build our predictive model, we'll perform a 75-25 split into train and test sets and run a multiple linear regression using ordinary-least squares to estimate prices.

When we run this model on all of our data, it becomes clear that some of the p-values (which measure confidence) and correlation coefficients (impact) of some columns are less useful, so a few iterations are made to clean up these inputs.

Building The Model

02

In the final model, the predictions we made were derived from Zipcode, Home condition, Home grade, Waterfront, View, Bedrooms, and Square Footage columns.

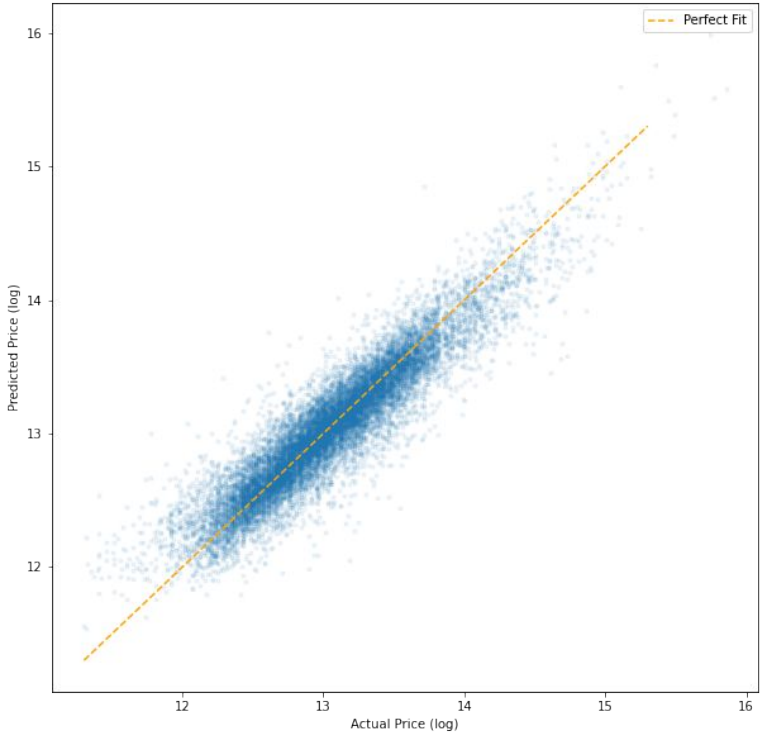
Some information (homes graded as “average”) had minimal predictive power and was dropped in later iterations of the model. Other columns in the dataset (bathrooms) were not used as they either contained redundant information or an insignificant improvement in predictive power.

Validating the Model

01

We want to make sure the model is accurate before we use it for anything. This can be measured by R-squared and mean-squared error.

The final predictive power here had an R-squared of 0.85 and a mean-squared error of 0.04 on unseen log data.

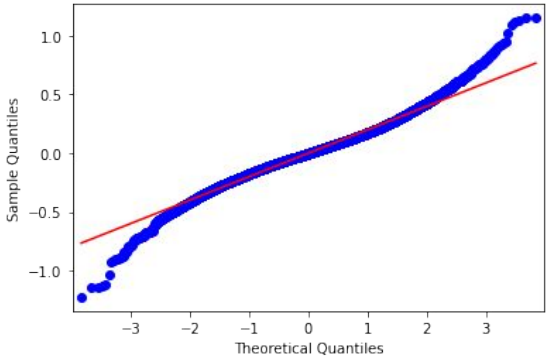


Validating the Model

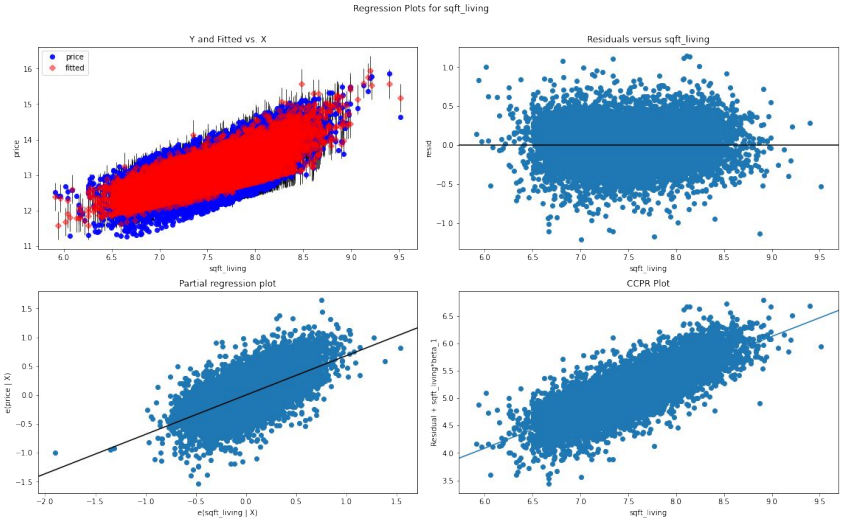
02

The main assumptions we want to check in linear regressions are linearity, normality, and heteroscedasticity.

The normality is best seen in the Q-Q plot to the top right, while the others are specific to every feature. More information can be seen in the Notebook!



Tons of plots!



Conclusions

Conclusions

01 - Location, location, location

In building our model, a few predictors stood out as much more important than others. Location data like certain Zip Codes predicted the most about a property's price, for example having a house in 98039 (in Medina) correlated with more than a doubling in price.

Waterfront properties correlated with an almost 50% increase in price, while homes categorized as having excellent views were predicted to be about 35% more valuable. In making investments, each of these location considerations should naturally be considered.

Conclusions

02 - Structural Details

Square Footage of living space was perhaps the most stable numeric predictor, with a doubling of square footage representing a 68% increase in predicted property value. Bedrooms, bathrooms, and floors were surprisingly negligible in predicting price, most were later removed from the model to reduce redundant information (multicollinearity).

Conclusions

03 - Home Condition and Luxury

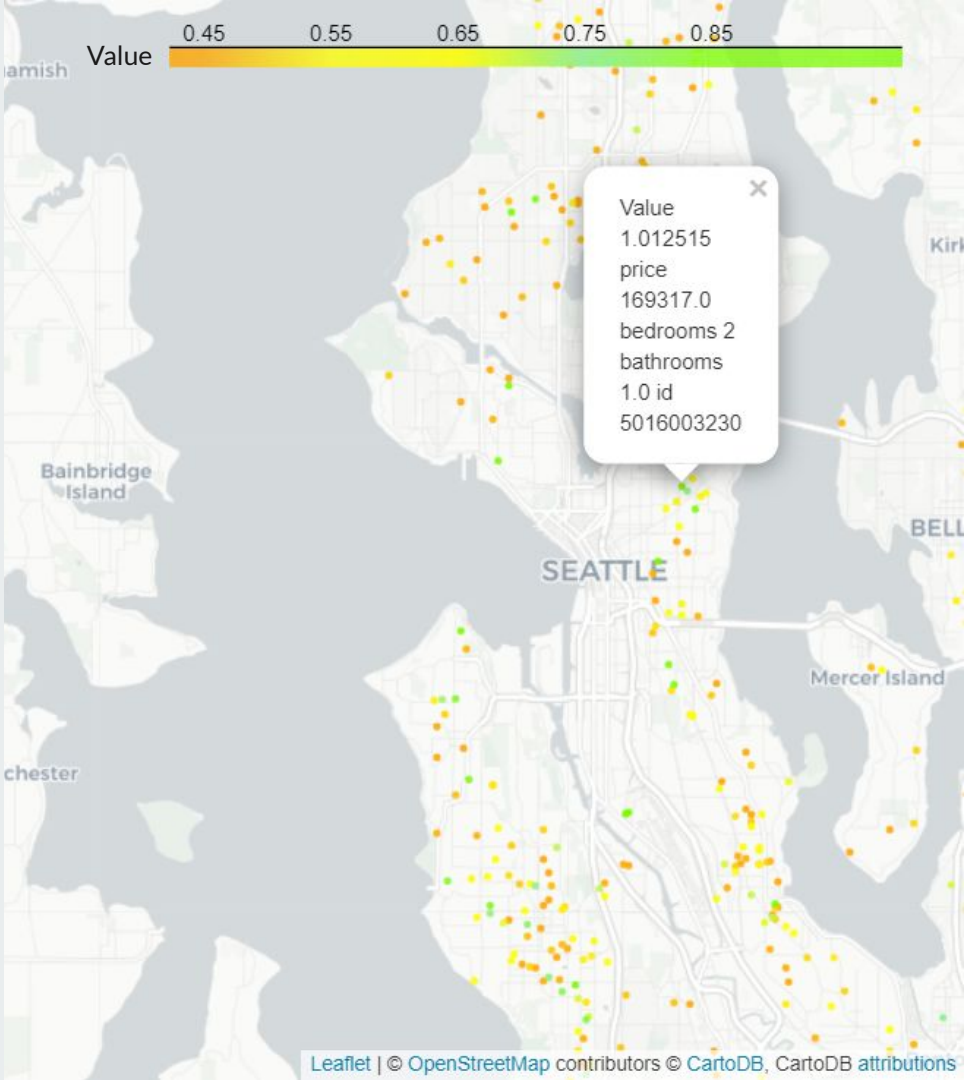
Homes described as “Excellent” (+30%), “Luxury” (+43%), or “Mansion” (+70%) naturally correlated with price increases, while poor home condition was reflected by a 36% decrease in home value. Positive descriptions of home condition were not as impacting as negative ones.

Conclusions

04

To make it easier to act on these findings, we can compare estimated and actual prices to show homes which seem like good deals.

The 500 estimated “best deals” were colored by this value and displayed on an interactive map which could be shared or built into applications.



Next Steps

Actions:

- ➔ More validation should be done to see to what degree this sort of prediction can realistically steer investments
- ➔ More feedback should be collected from those within the industry on how they currently make decisions
- ➔ Giving Tools such as the map to existing developers, home-buyers, and real-estate offices

Model Improvements:

- ➔ More geographic data (schools, hospitals, parks, etc) could improve the model accuracy
- ➔ Data across time is particularly interesting
- ➔ More tools could be built and integrated into the workflow of existing decision-makers

Thank you! Questions?

Presentation by Blake McMeekin
blakemcme@gmail.com
@the_grand_blooms
www.linkedin.com/in/blakemcme