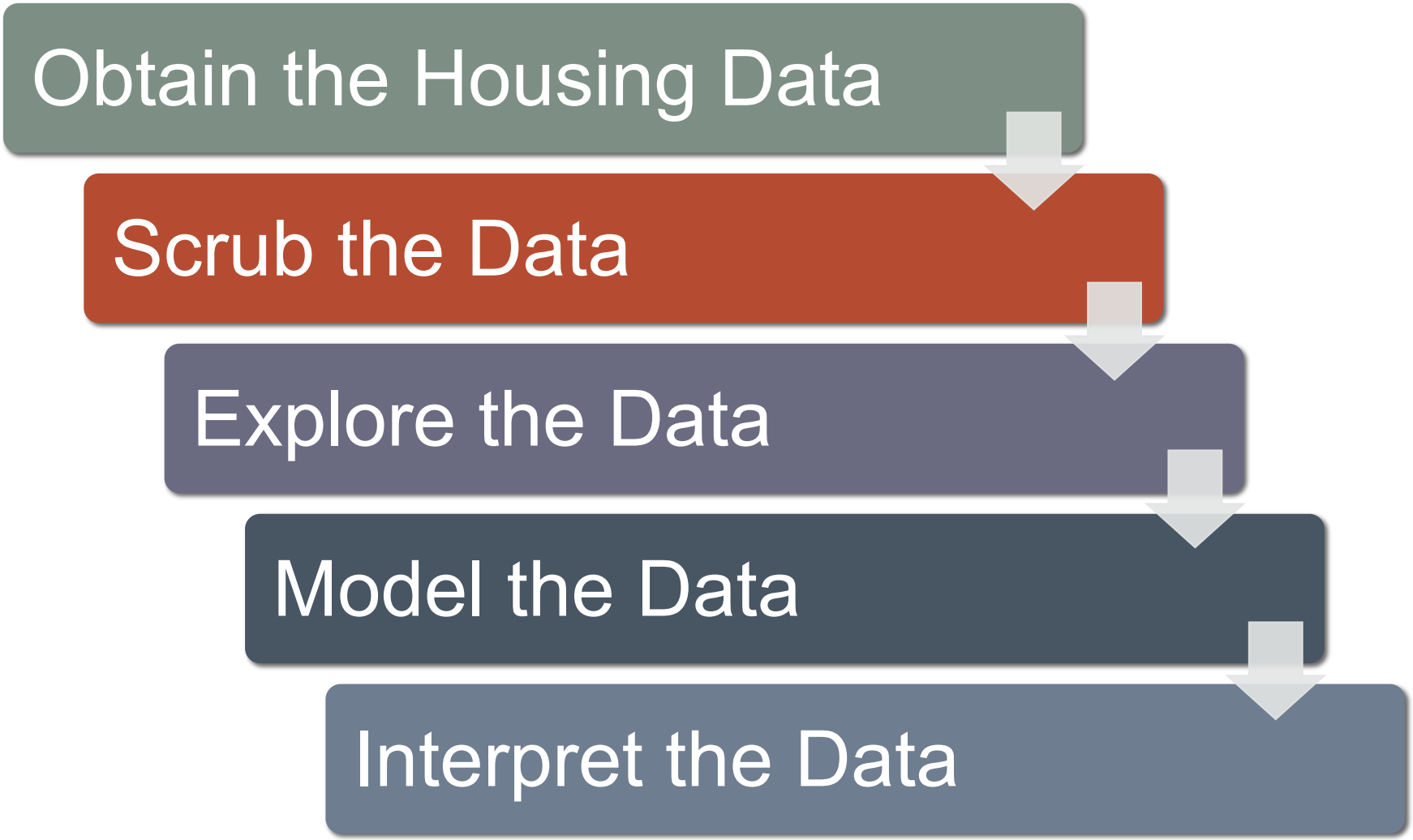
PREDICTING HOUSE PRICES IN KING COUNTY, WA

Presented by Crystal Dugan 05-21-19

Introduction

- I used data from house sales to build a model that can be used to predict house prices.
- Plugging the values of 12 features of a house into the model will generate the predicted price of the house.
- I will briefly go over my methodology and findings in working with the data to develop this model.

Overview of the Modeling Process

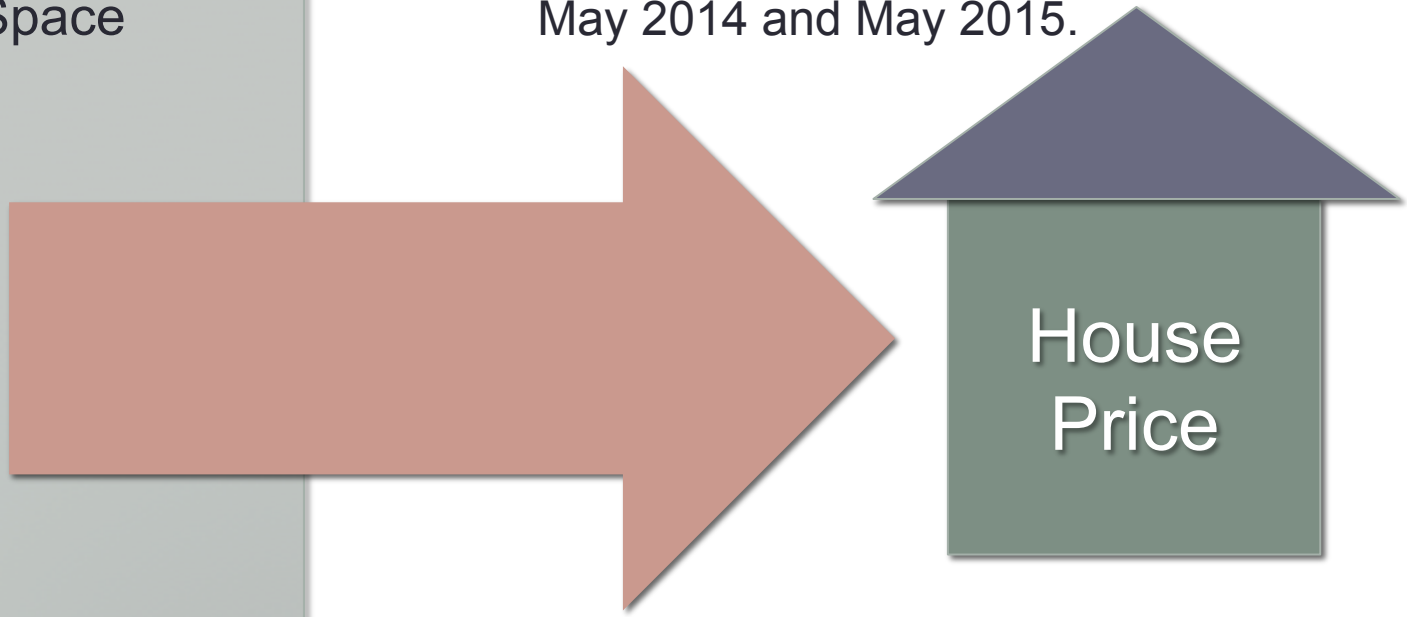


The Data Set

Features Used

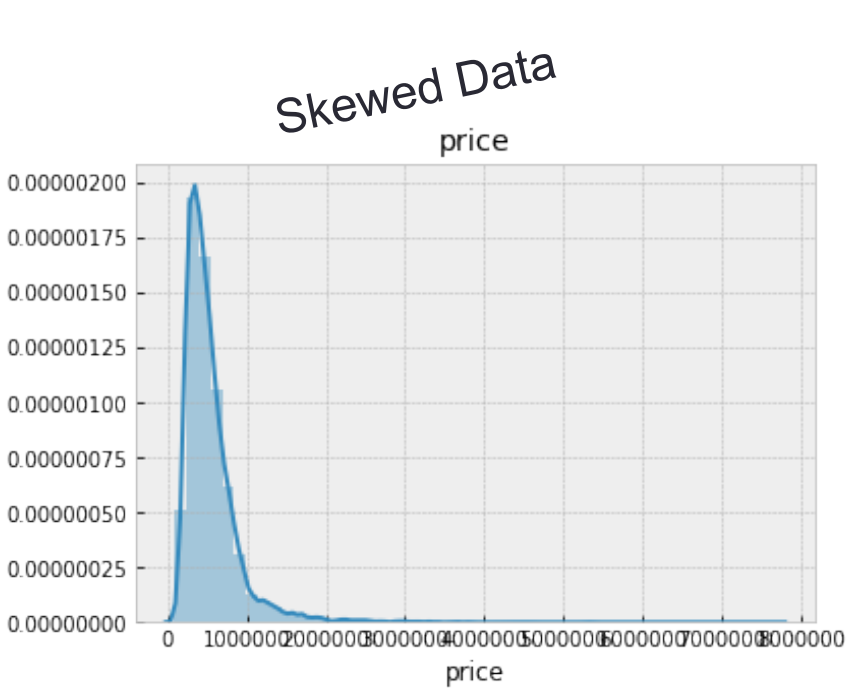
- Zip code
- Sqft Living Space
- Waterfront
- Grade
- Condition
- Sqft Lot
- Bedrooms
- Bathrooms
- Floors
- Year Built
- Sqft Living of nearest 15
- Sqft Lot of nearest 15

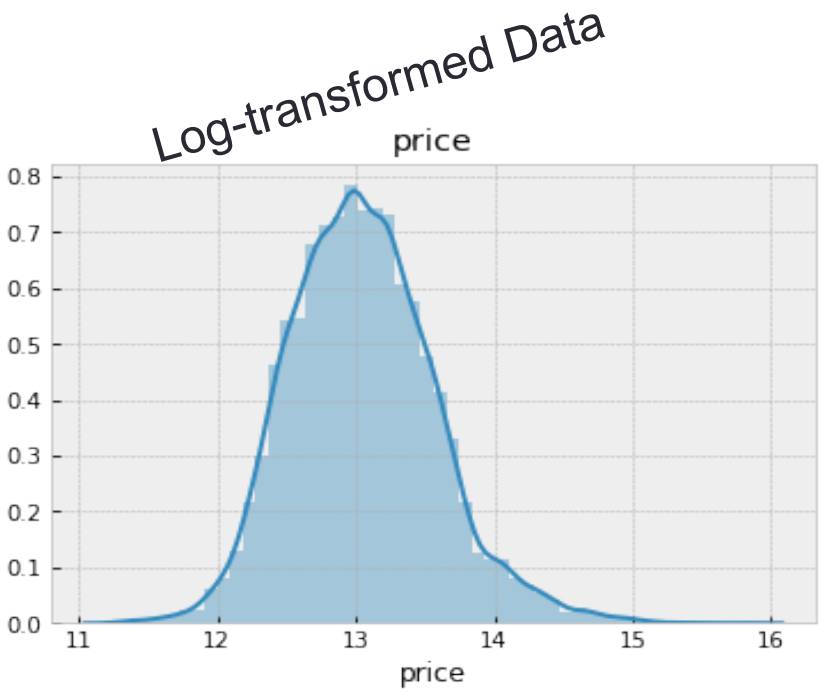
The data comes from a dataset in Kaggle called House Sales in King County, USA. This dataset includes houses sold between May 2014 and May 2015.



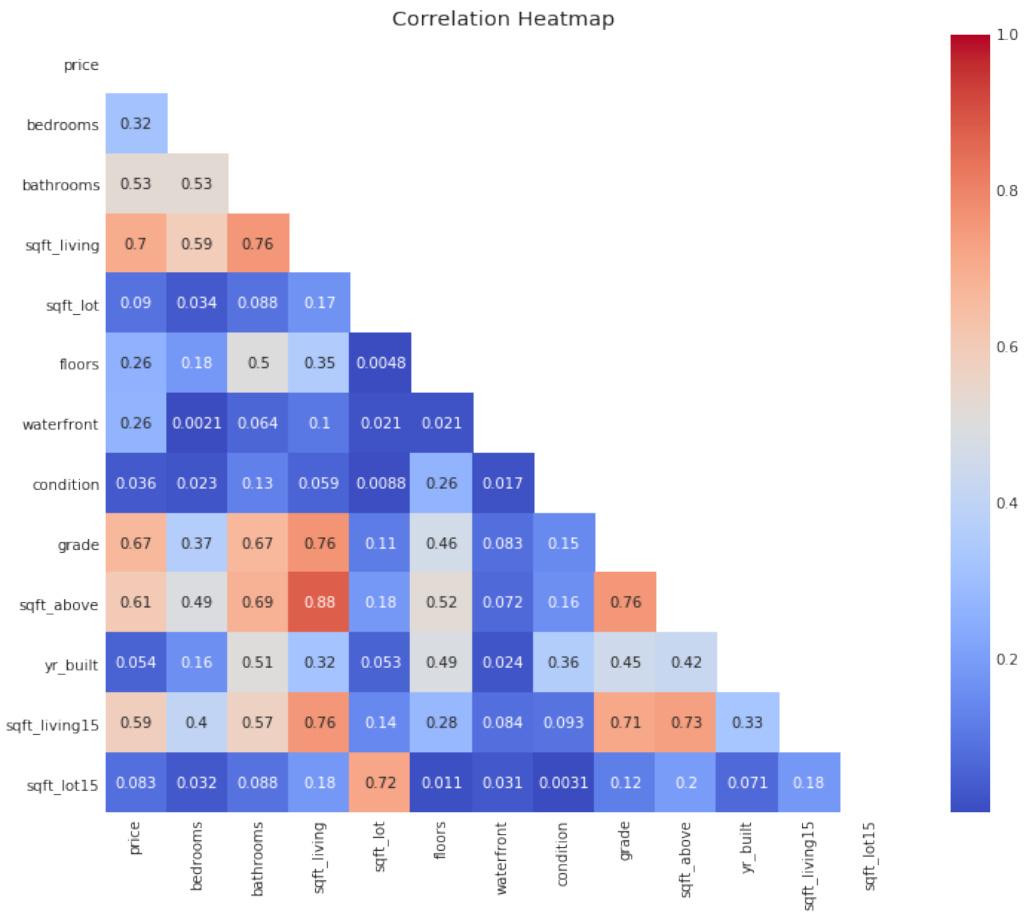
Scrubbing the Data

This intails getting fixing or deleting poor data (ex: missing a lot of data points or formatted incorrectly) and transforming data that is skewed so it is more normally distributed. This improves model performance.





Scrubbing/Exploring the Data

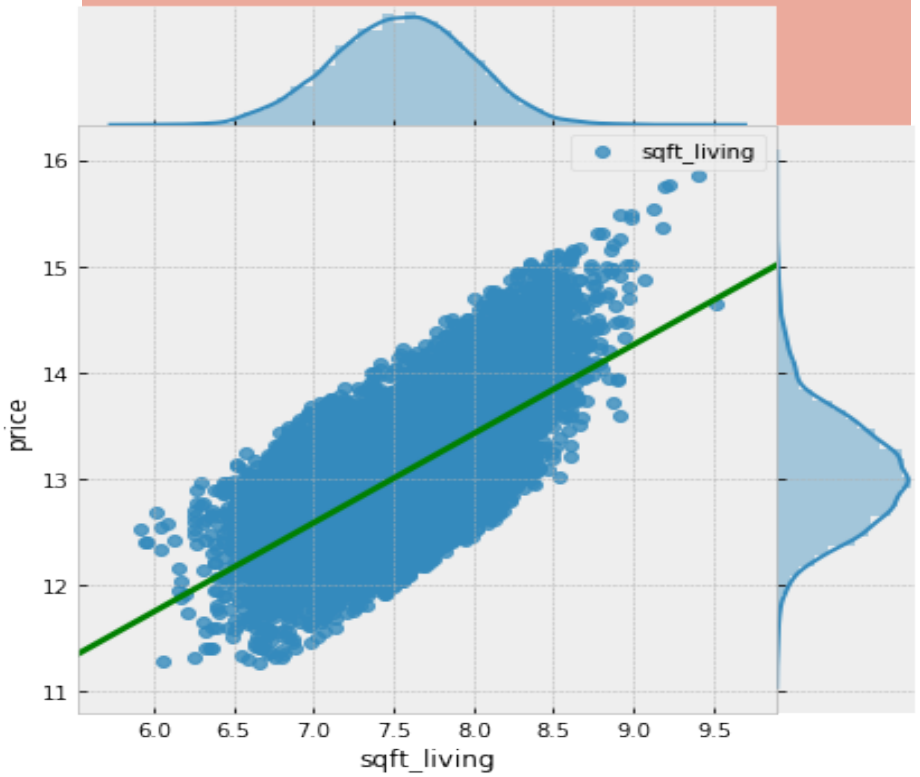


- You can see from this graph that sqft_living, grade, and sqft_above are all strongly correlated to price.
- sqft_above is also correlated strongly with sqft_living so it was dropped from the model (bias).

Exploring the Data

- Look for correlations between features and target (price).
- This graph shows that more living space correlates with higher price.
- Features with no correlation to price get dropped.

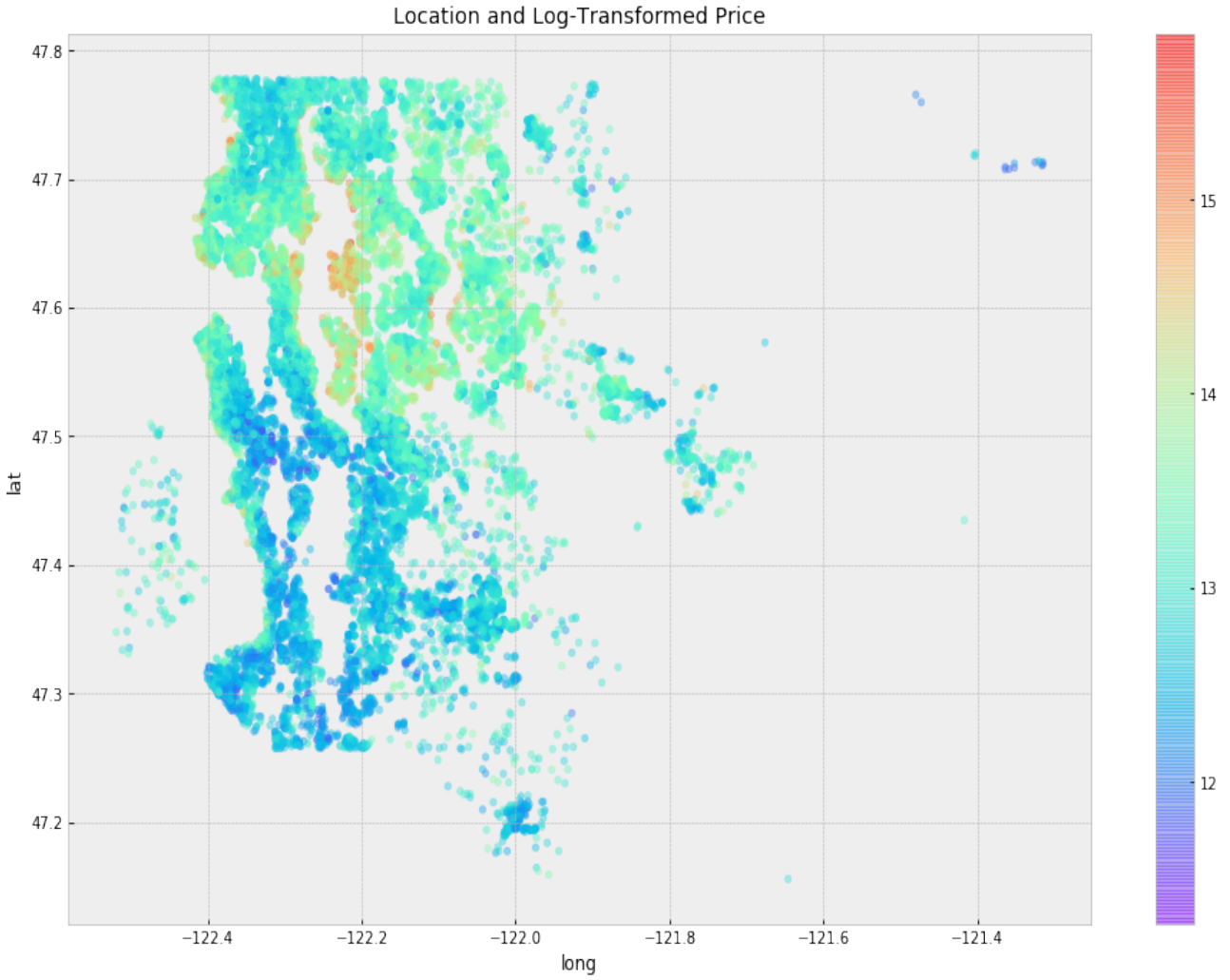
- You can see “sqft_living” has a strong positive correlation with the price.



Exploring the Data - Location

One of the biggest factors affecting price is (not surprisingly) location. Using the zip code lets me factor location into my model (this is an example of categorical data).

This graph shows longitude and latitude of each house, color indicating relative price.



Interpreting the Data

- The predictive model I developed from this data is able to account for 87.6% of the variation in price. This gives me a fairly high level of confidence in my model to accurately predict housing prices.
- The remaining 12.4% could be influence of factors like sampling error, market fluctuations, or features not included in the dataset.
- The features that have the most effect on the sales price are zip code, grade, square feet living space, and number of bathrooms.