Move on Up:
Developing a model to predict housing prices in King County, WA

Joe Nelson

The Scope:

Move on Up is set to disrupt the real estate industry in Seattle by giving homebuyers and sellers an alternative to the 4-6% fees typically charged by traditional real estate agents, using a predictive model for home pricing, and connecting buyers and sellers through an on-line portal.

Our early-stage model was built using the King County, WA real estate data set.

The Data:

- 21,597 residential home sales from 2014-2015
- 70 different zip codes
- Features include livable square feet, basement square feet, lot size, floors, bedrooms, bathrooms, location, condition, grade, whether it abuts the water, and neighboring property information.

We explored each of these features to determine which would be most effective in our model in predicting sale price.

Model Creation

Method:

- Reviewing the Data (geographic distribution, etc.)
- Cleaning the Data (filling in NaNs, placeholders, removing outliers, etc.)
- Reviewing for Normality and Ensuring minimal multicollinearity
- Training the Model
- Validating the Model

King County, Washington

Located on the Puget Sound, extending eastward from Seattle

Price per Foot by Zip Code

- Price-per-square-foot is a common metric, which we derived from the data by creating a new feature by dividing the sales price by livable square feet.
- The following slide shows sales spread over the county geographically, but there was no clear, categorical method to show how lat/long data affected price.
- The slide after shows how price per square foot varied by zip code, which made a better categorical predictor.

King County Housing Data - Geographical Distribution Across King County, WA
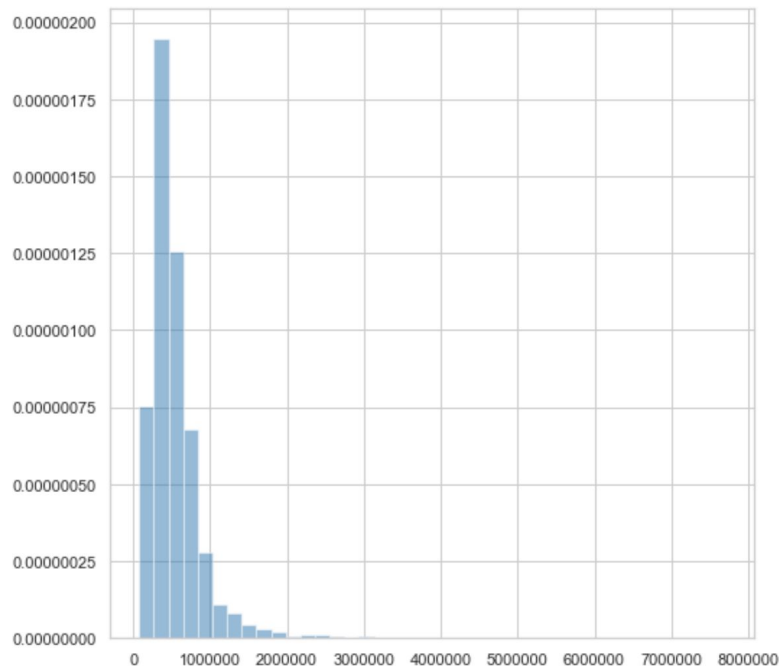By Price per Square Foot
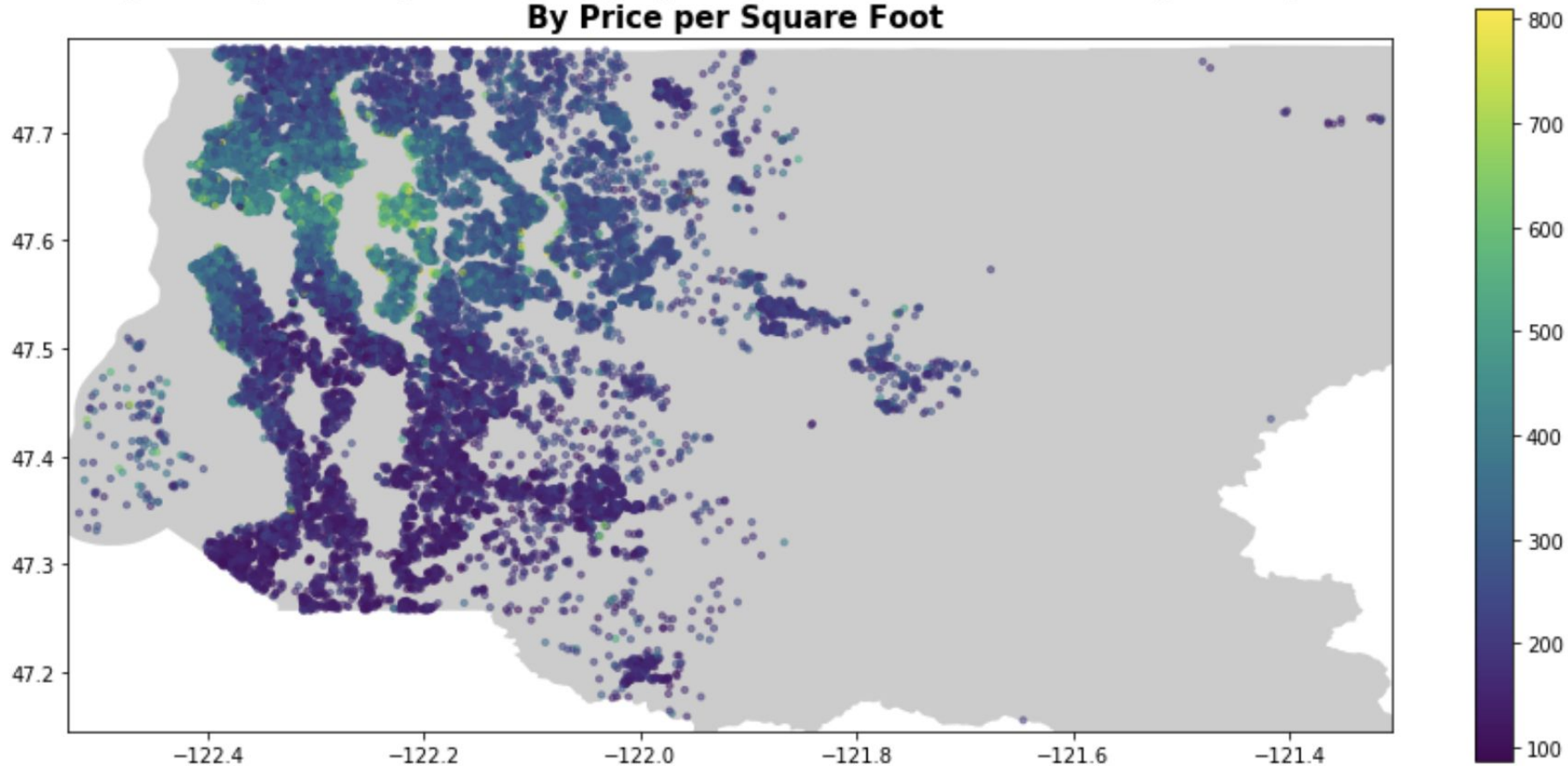
Price Per Square Foot by Zip Code
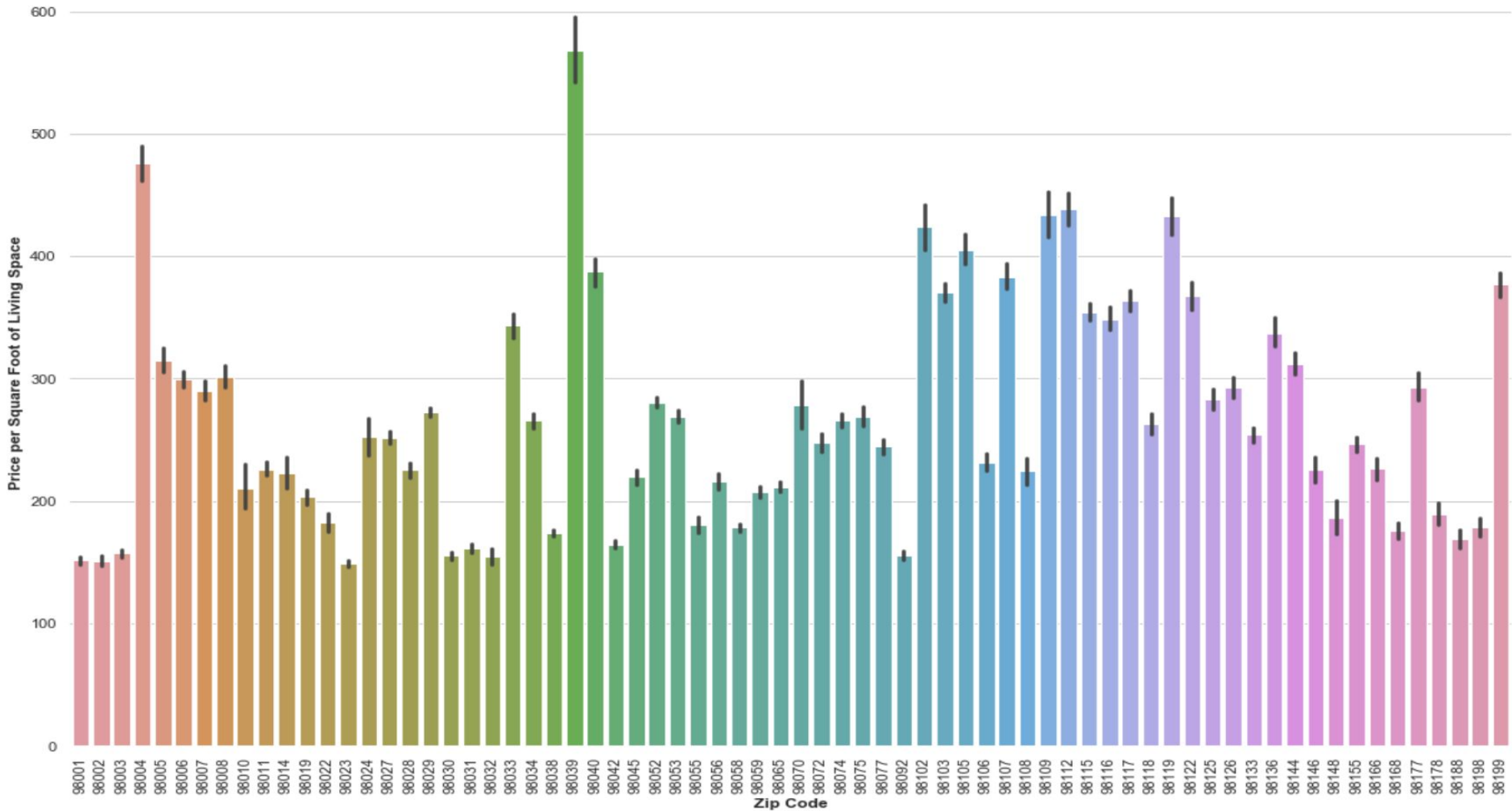
Price Varied Greatly

- The home prices varied greatly, from $78,000 to $7.7 million.

- The distribution of price has a long tail into the higher home prices. We may achieve a more normal distribution by selecting a more limited scope from the data set.
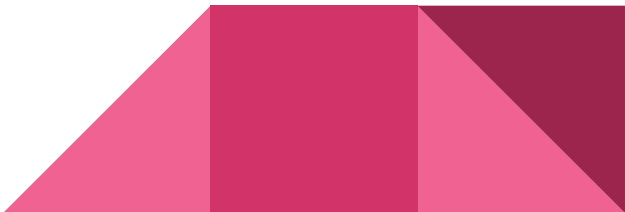
Price Varied Greatly

- We ultimately decided on focusing on the middle-class, including home sales between **$200,000 - $600,000**.
- This brought the mean and median of the usable data set closer (mean of 389,270, and median of 382,500).
- This left us with just 14,370 data points to use in training the model.
- The histogram to the right shows the more-normal distribution.
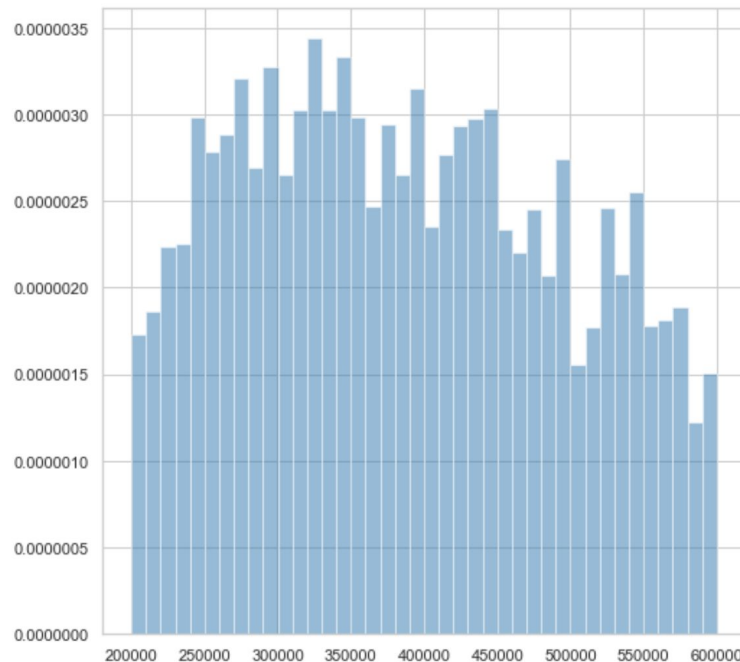
Model Outcomes

- The model shows a positive linear relationship between price and livable square feet, grade, condition, the number of bathrooms, lot size and whether they abutted water.
- Generically, houses started at just under $10,000 plus $75.41 per square foot; other predictors (bathrooms, bedrooms, lot size, basement size, etc.) then adjusted this number.
- Waterfront properties enjoyed a nearly $100,000 premium, based purely upon whether they abutted the water.
- Newer homes were generally valued *less*; possible explanations include a bias in the model (where newer homes were smaller, etc.).

Model Outcomes

- Grade and condition also provided a significant premium.
- Condition ranged from 1-5, with an approximately $30,000 premium for each step up.
- Grade had many more levels, ranging from 1-13, and had a lower premium for each step ($11,840 per step).
- Curiously, bedrooms have an inverse relationship with price. I'm not sure how to explain this one; it seems counterintuitive.

Future Work

- The model is not quite ready; it needs to be refined further.
- There may be bias in the model related to year built.

Thank you!