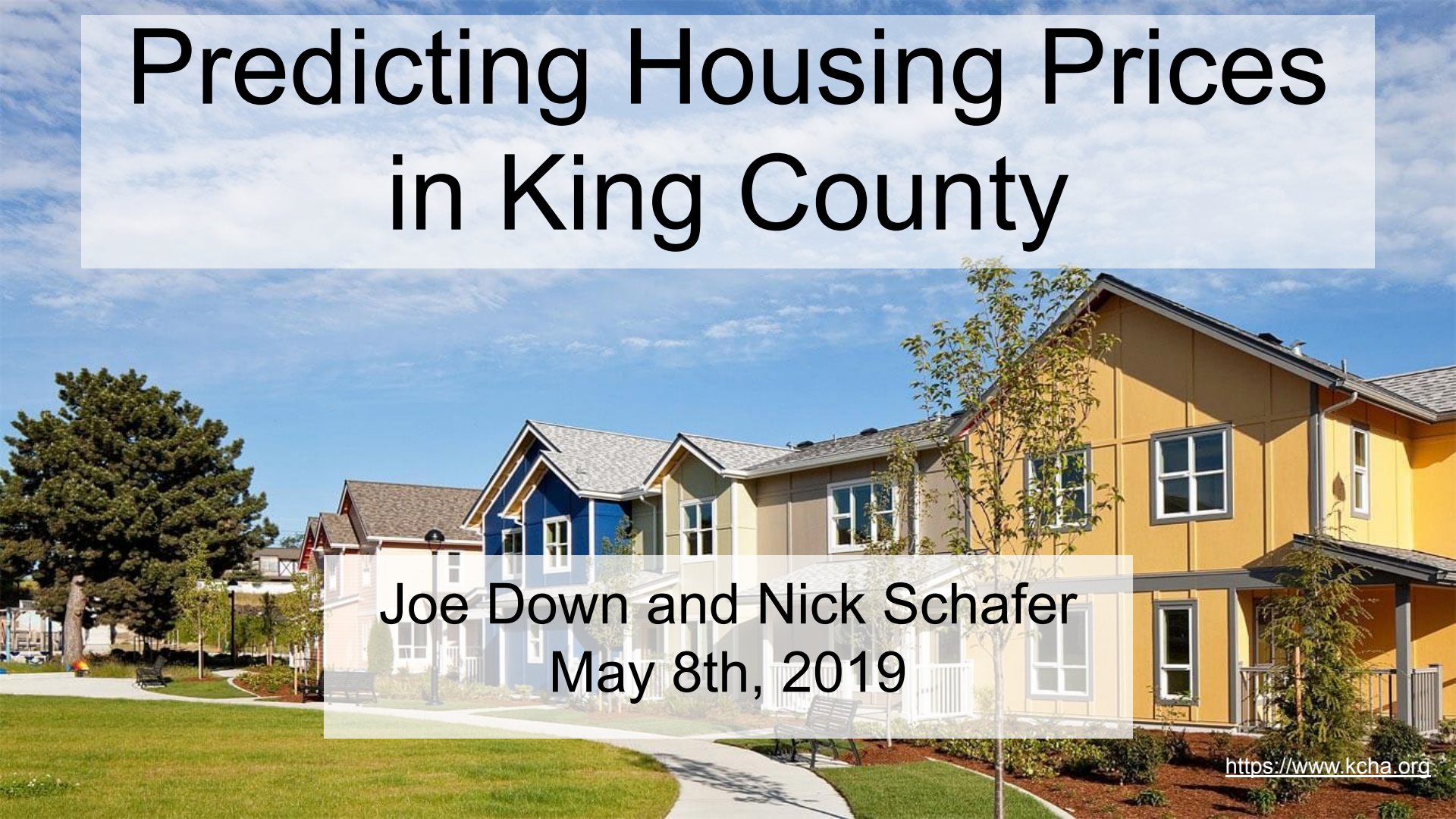
Predicting Housing Prices in King County



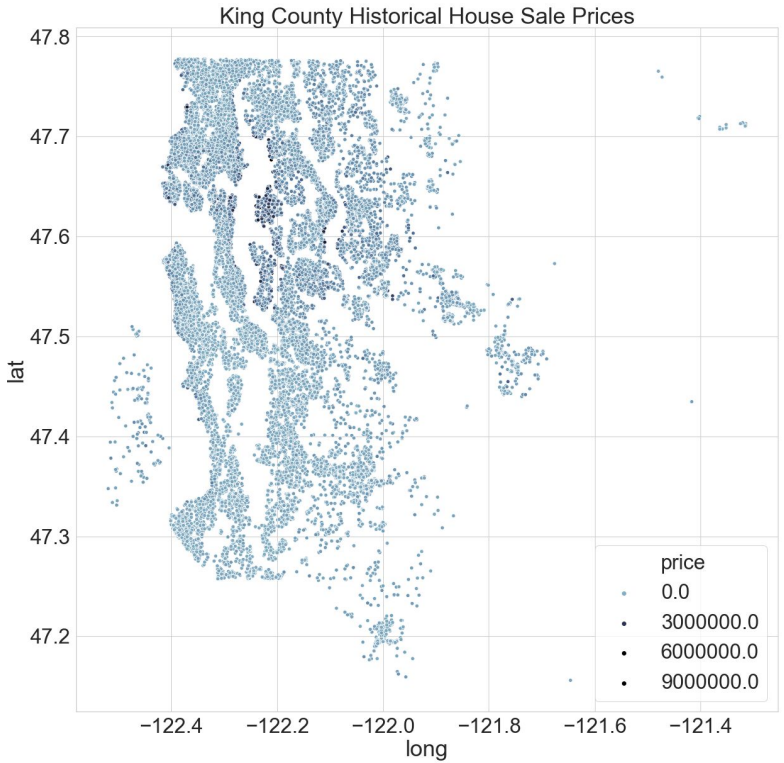
Joe Down and Nick Schafer

May 8th, 2019

https://www.kcha.org

Predicting housing prices in King County

1. How do log transformations of both the regressors and the target affect our model?
2. How does normalization of the data using standardization affect both our R squared and root mean squared error?
3. How does one hot encoding of zipcode affect our overall ability to predict the price of a home in the King County area?



Data overview

- 21597 Rows
  - 1 row = 1 house sale
- 21 Columns
- Format: Raw CSV File

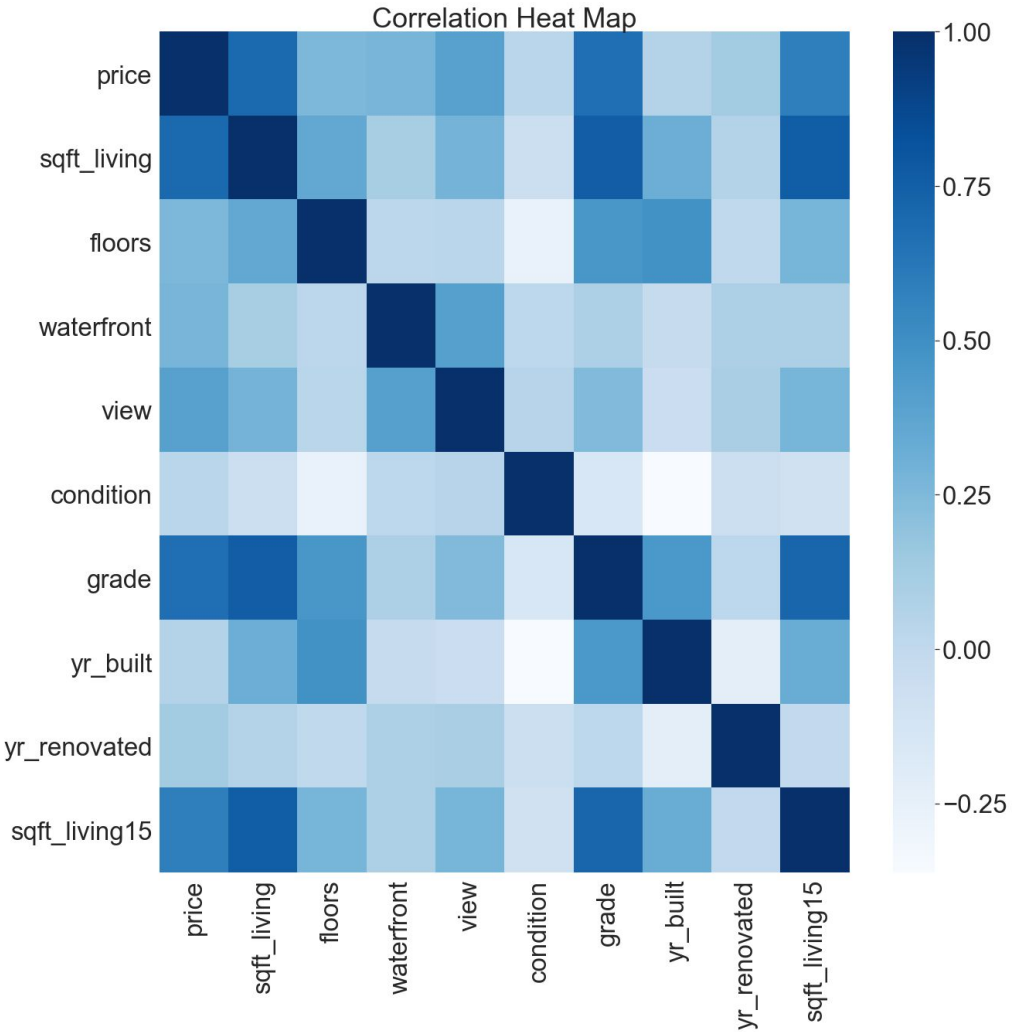
Column	Description
id	Unique identifier for each house
date	Date house was sold
price	Price the house was sold for
bedrooms	Number of bedrooms in the house
bathrooms	Number of bathrooms in the house
sqft_living	Square footage of the house
sqft_lot	Square footage of the lot
floors	Number of floors in the house
waterfront	Has a view to a waterfront
view	How many times the house was viewed
condition	Overall condition of the house from 1 (worst) to 5 (best)
grade	Grade from King County from 3 (worst) -13 (best)
sqft_above	Square footage of the house apart from the basement
sqft_basement	Square footage of the basement
yr_built	Year the home was built
yr_renovated	Year when the home was renovated
zipcode	Zipcode of the home
lat	Lattitude coordinate
long	Longitude coordinate
sqft_living15	Square footage of living space for the nearest 15 neighbors
sqft_lot15	Square footage of the land lots of the nearest 15 neighbors

Data cleaning

- Columns that are not useful for predicting prices were dropped
  - id
- Rows of data that information that may be critical for prediction were dropped
  - waterfront
  - view
- Similar types of data were converted to compatible types
  - date
  - yr_built
  - yr_renovated
- Missing data was filled in using reasonable replacements
  - yr_built → yr_renovated
  - sqft_basement ? → 0.0

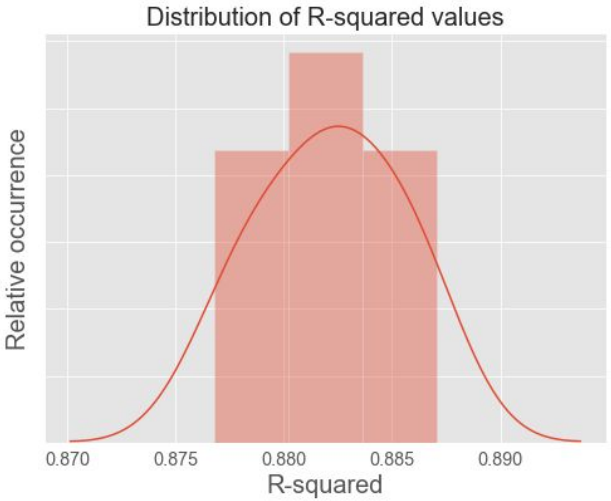
Initial Observations

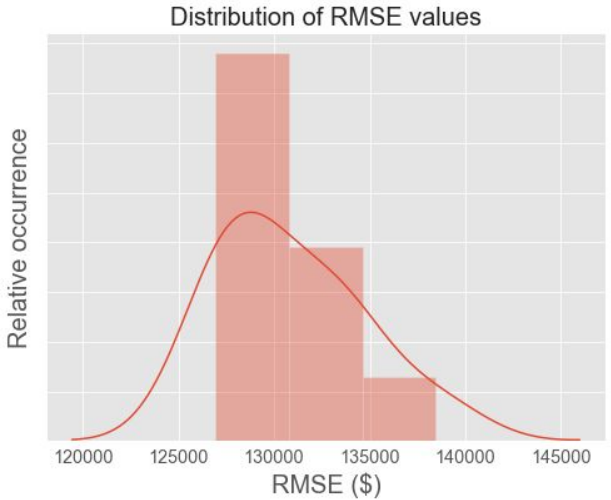
- Observing measures of central tendency
- Correlation heat map
- Looking at scatter plots and distribution



Model evaluation approach

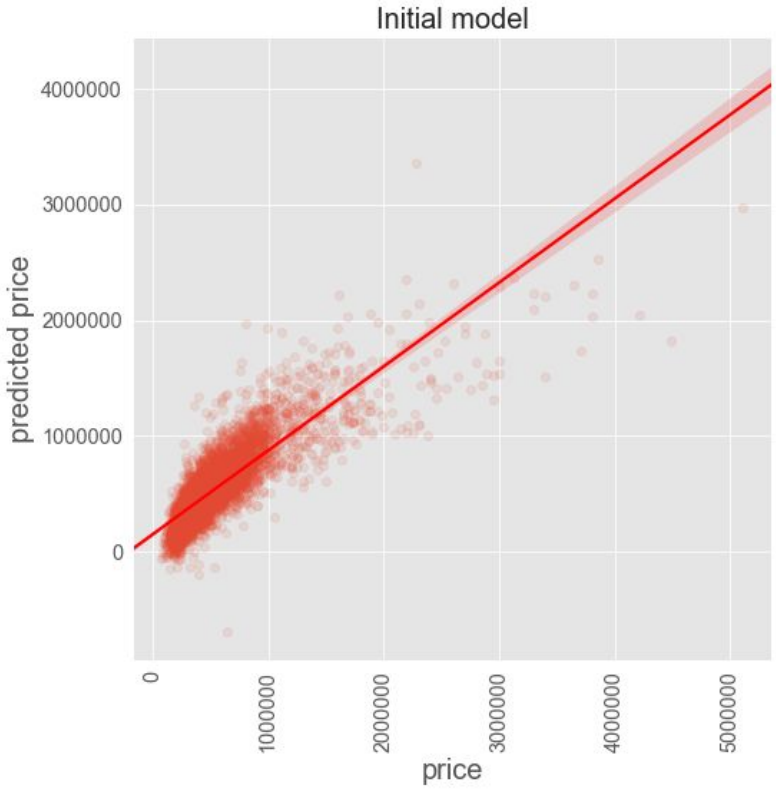
- Perform 10 independent test-train splits
- Train on ⅔ of the data, test on ⅓ of the data
- R-squared and root-mean-squared error values presented hereafter are averages over the 10 test-train splits
- The R-squareds and RMSEs tend to have a relatively narrow distribution and are therefore well-represented by the mean





Initial model

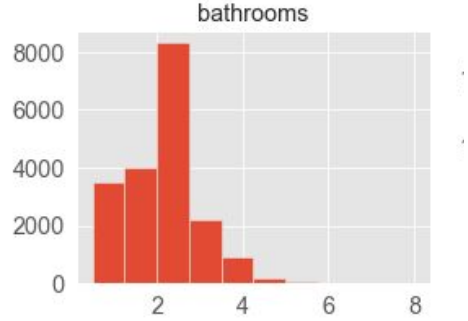
- We used Scikit-Learn to build an ordinary least squares linear regression model
- The target variable for prediction was “price”
- Initially, we test a model without log transformations, normalization, or one-hot encoding of zipcodes

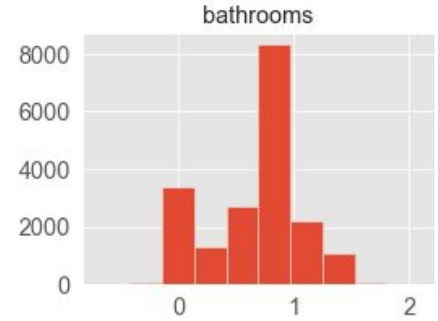


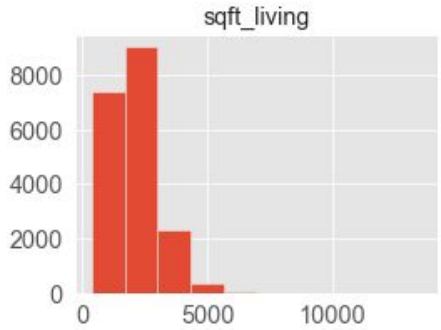
R-squared: 0.70, RMSE: 200785

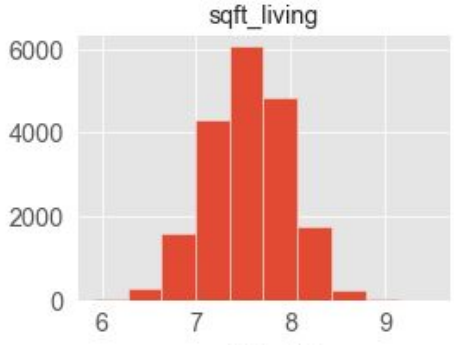
Log transformation of data

We used log transformations to make our distributions more normal.



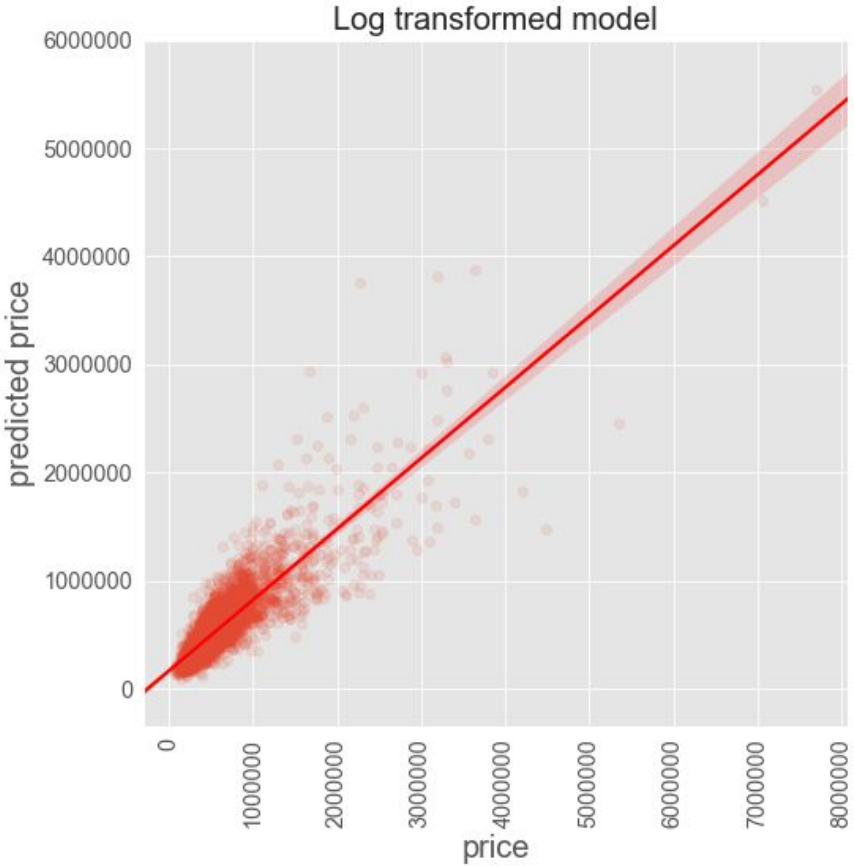




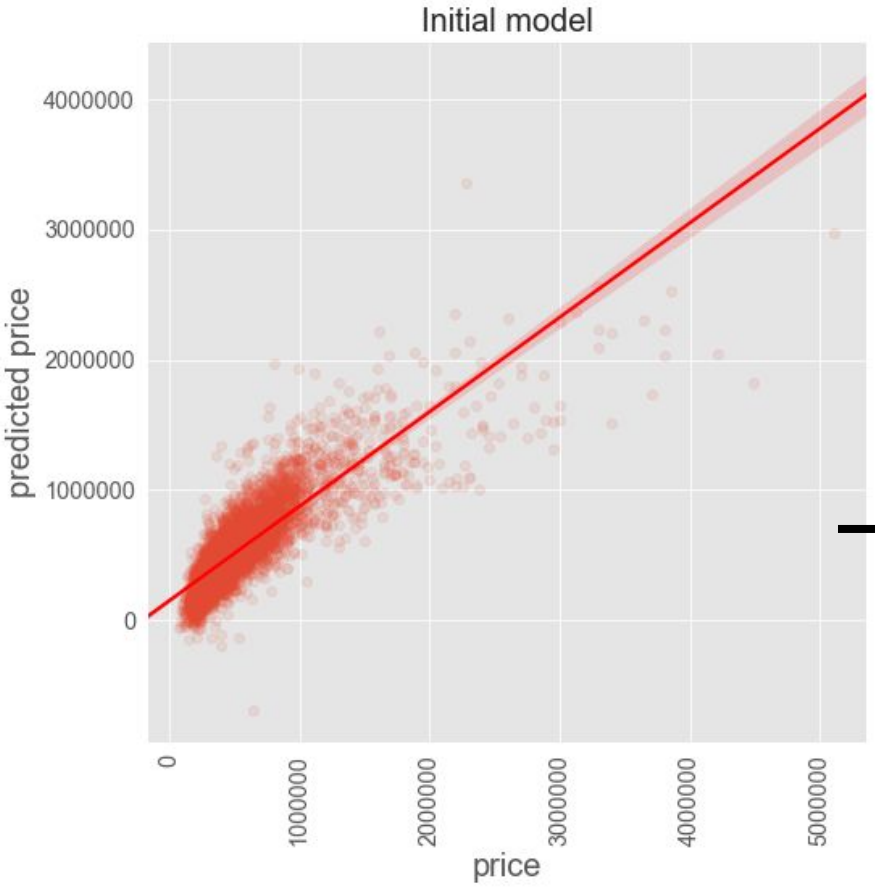


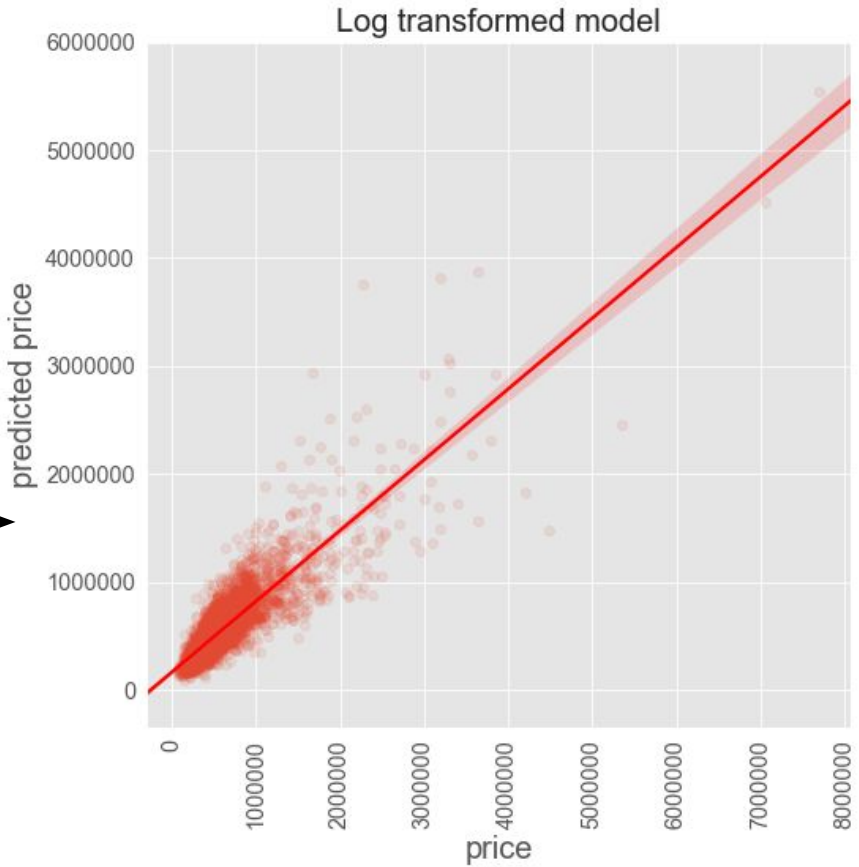
Log transformation of data

- Performing the log transformation of the data improves both the R-squared value and the root-mean-square error.



R-squared: 0.77, RMSE: 195287

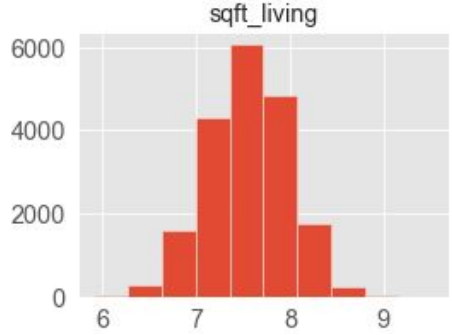


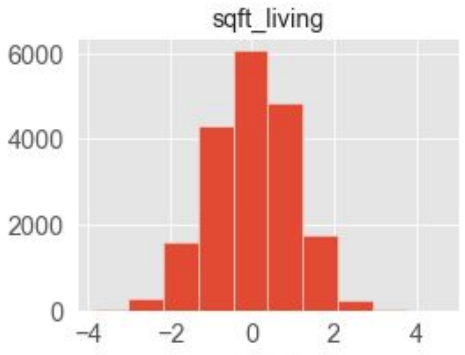


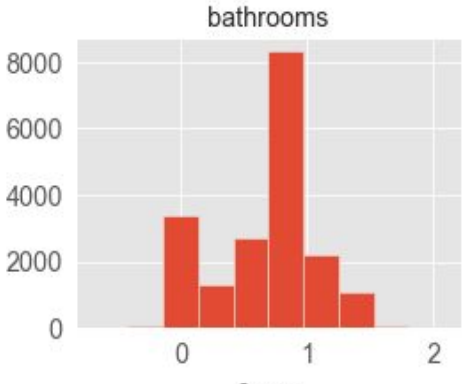
R-squared: 0.77, RMSE: 195287

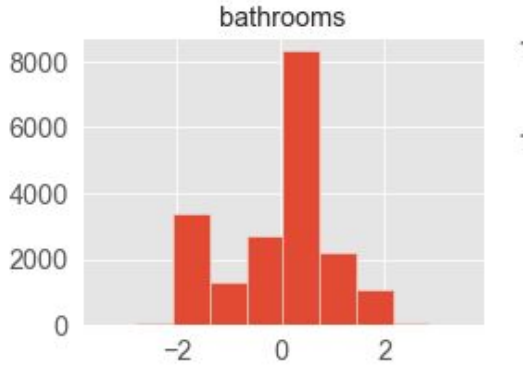
Normalization of the data

All of the variables that we had previously log transformed, we now normalize using Z-score normalization.



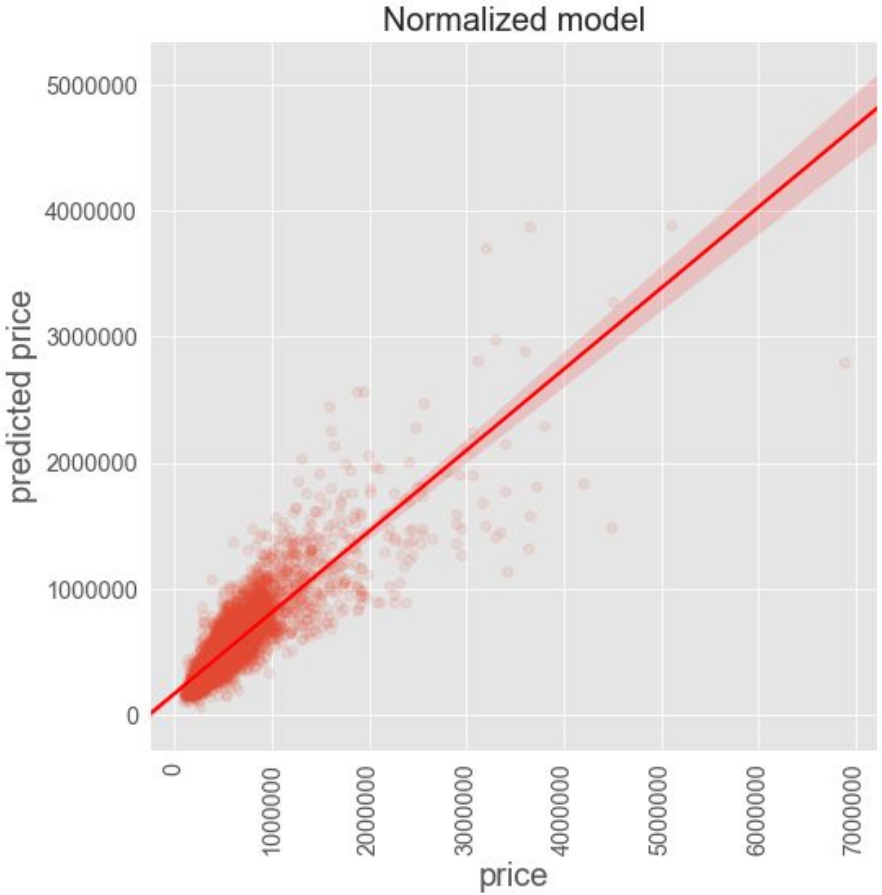




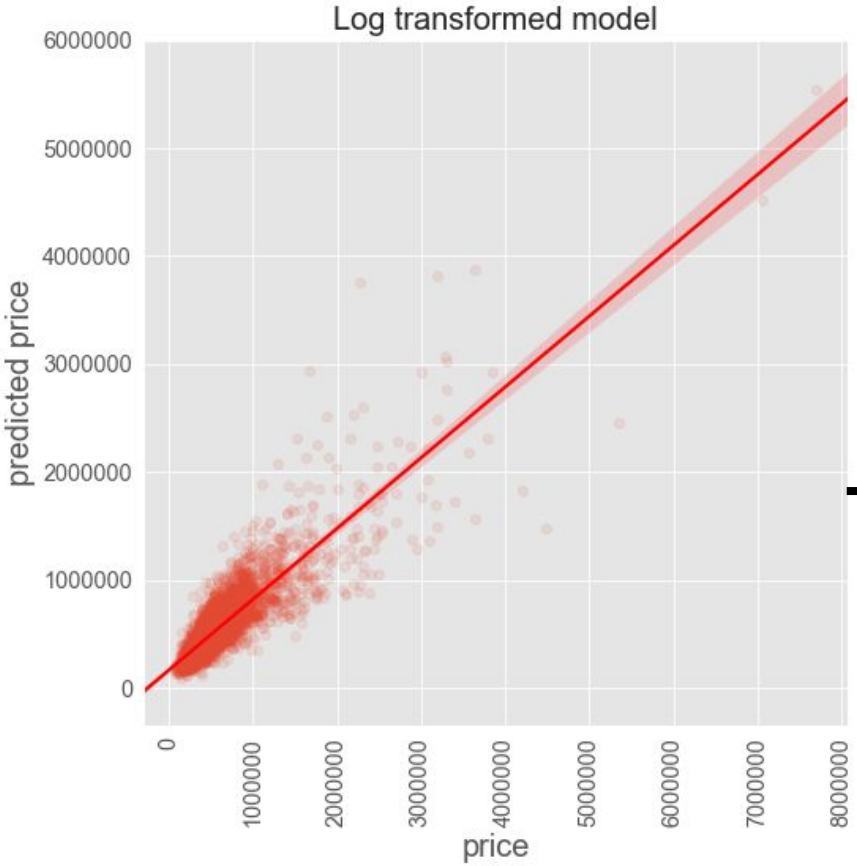


Normalization of the data

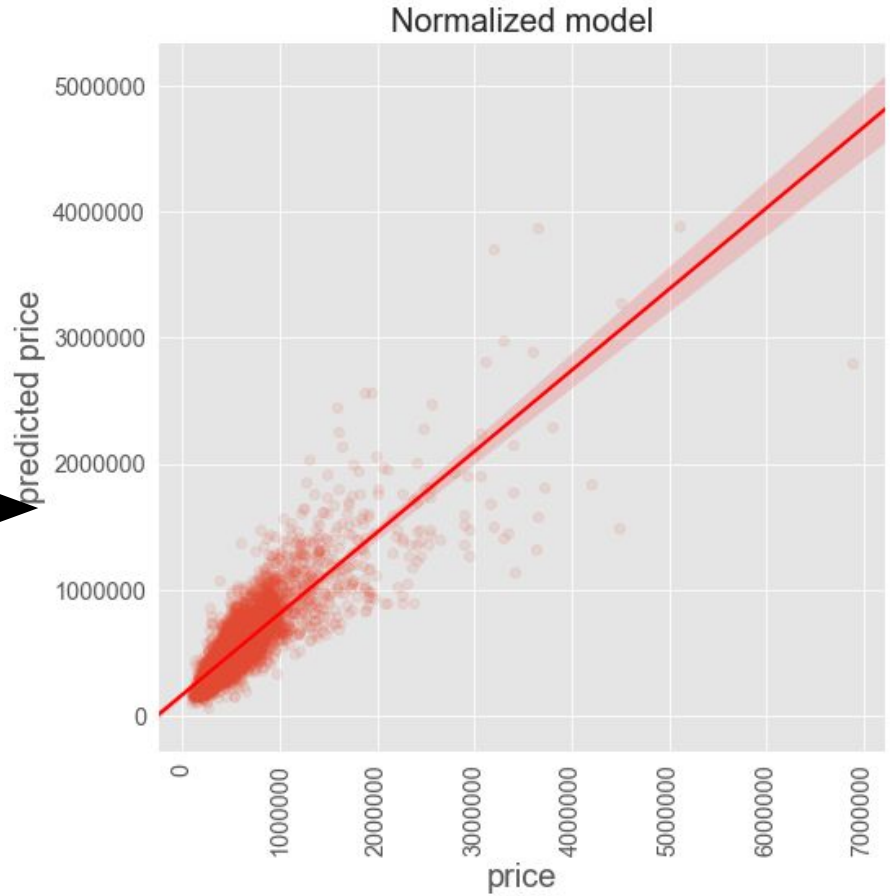
- Normalization of the input data did not significantly improve the quality of the predictions



R-squared: 0.77, RMSE: 192740



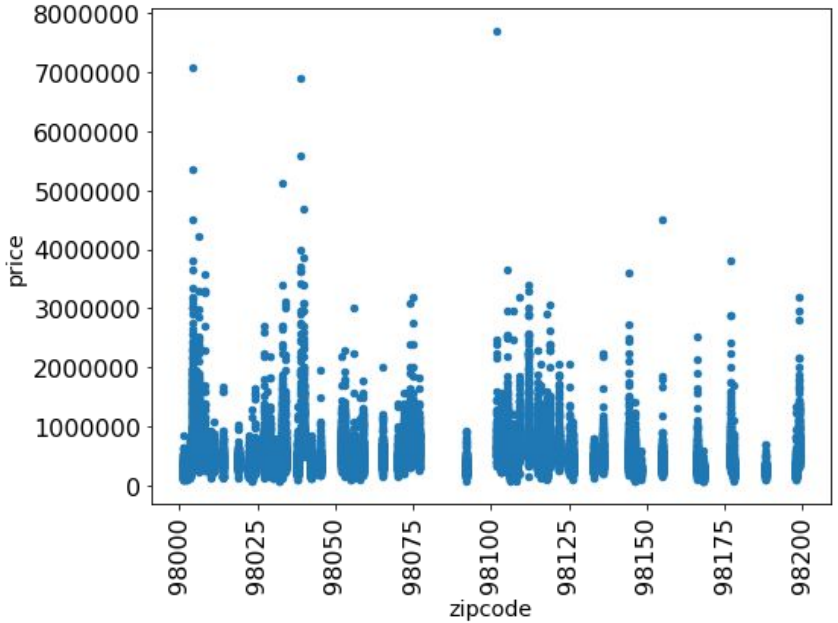
R-squared: 0.77, RMSE: 195287



R-squared: 0.77, RMSE: 192740

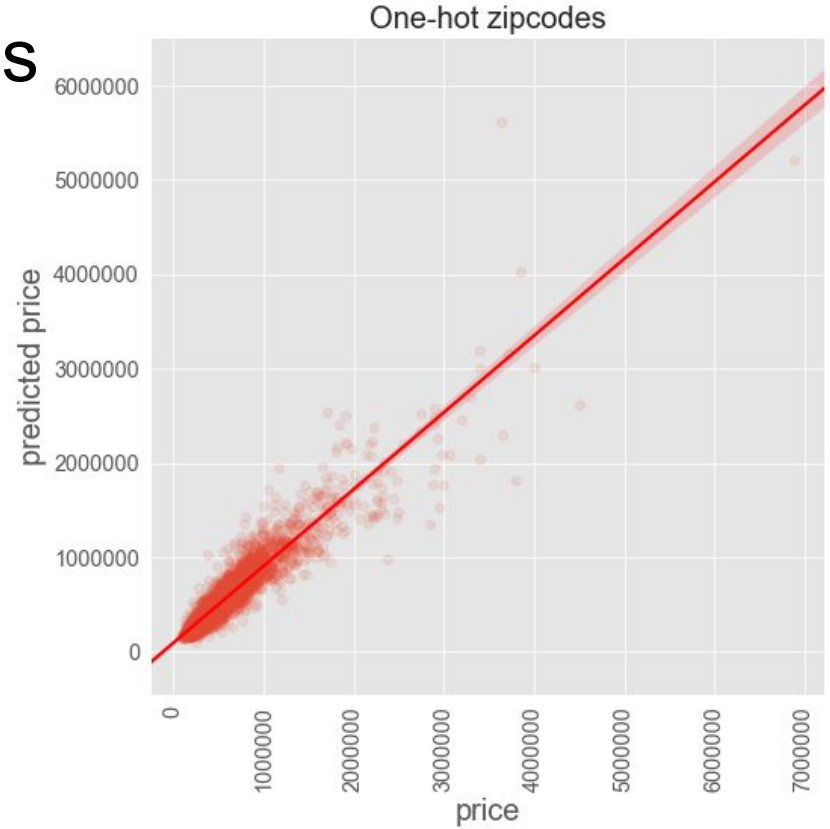
One-hot encoding of zipcodes

- Housing prices may vary significantly by zipcode, but the raw zipcode data is not expected to be correlated to housing prices
- We used one-hot encoding to add 70 new feature columns to our data set
- The original zipcode column was then dropped from the data set

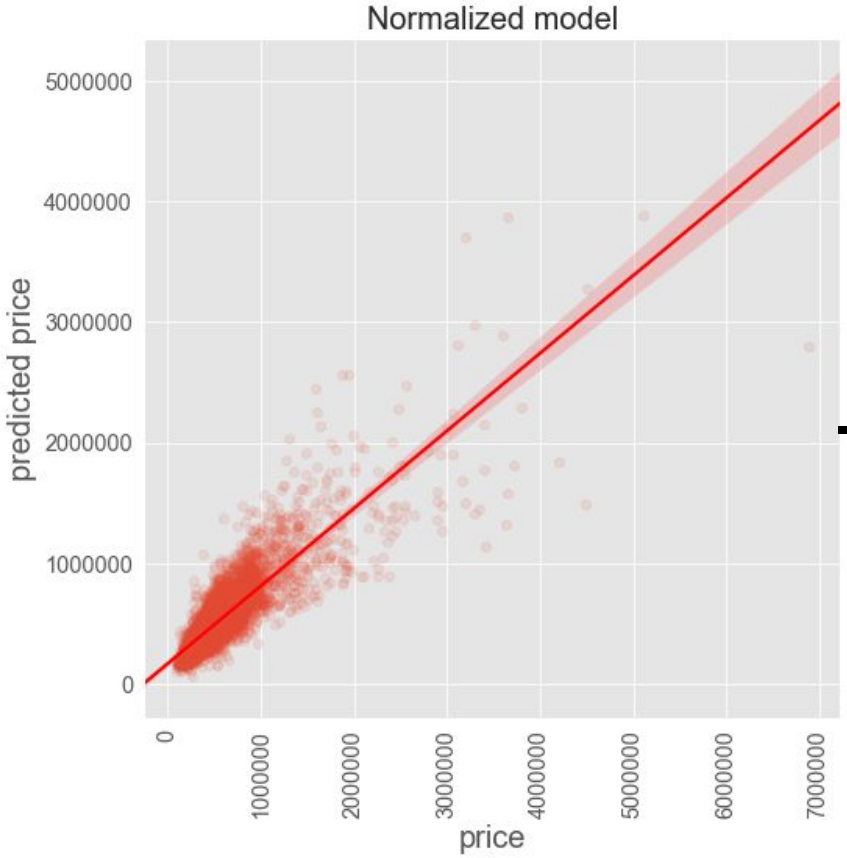


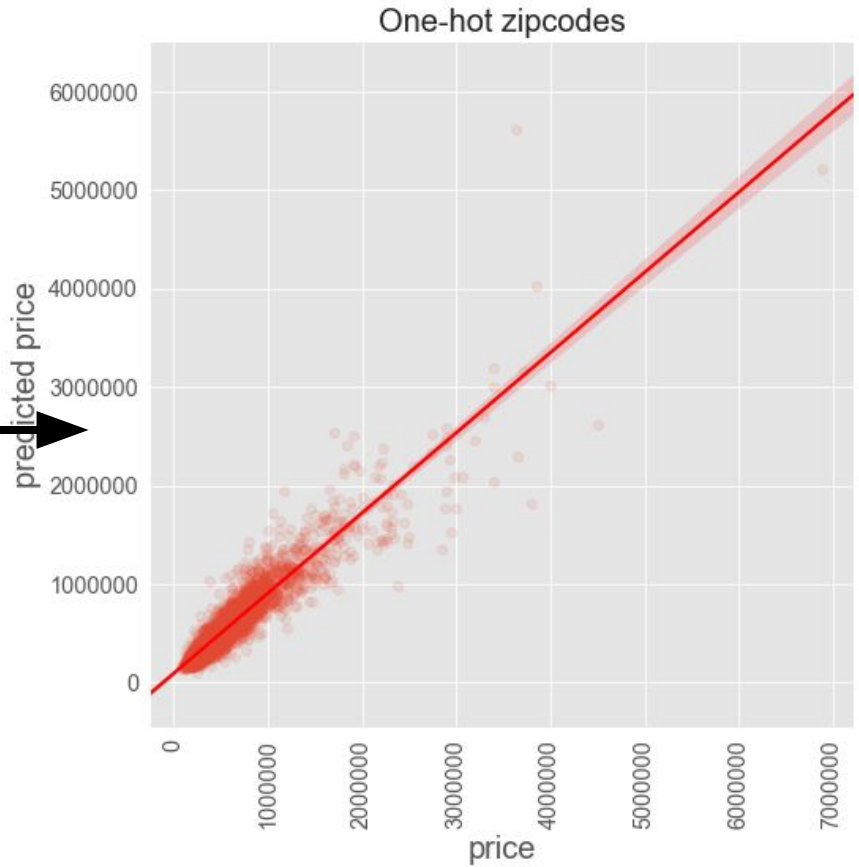
One-hot encoding of zipcodes

- Adding additional features using one-hot encoding of the zipcodes significantly improves both the R-squared value and the root-mean-squared error
- The root-mean-squared error of ~$130,000 still seems quite high, despite having an R-squared value of almost 0.9



R-squared: 0.88, RMSE: 130864





R-squared: 0.77, RMSE: 192740

R-squared: 0.88, RMSE: 130864

Further considerations

- Deriving new features by non-linear transformations of the given data
  - Price per square foot of living space by zipcode
  - Zipcode ranking feature
  - Long, lat mean price ranking
- Other types of data transformations, normalization, and checking for outliers
- Linear regression assumption checking
- Feature selection
- Interpretation of coefficients in regression
- Different models for different zipcodes
- Clustering based on proximity

The End

- Thanks for listening!