Estimating Housing Prices

in Kings County using Linear Regression Models

By: TaeJoon Kim and Matthew Andrews

TABLE OF CONTENTS

Business Problem

A real estate firm needs a reliable model to predict house prices based on its features. With an accurate estimate, the firm can quickly identify underpriced houses to invest in and generate the maximum amount of profit upon resale.

Preview of Results

- The variable with the highest effect on house prices was 'waterfront.'
- Proximity to places like schools or government buildings had a notable impact on house prices as well.
- The model was not perfectly linear, due to high range of house prices that were included in the model.

MYHOME
REAL ESTATE

DATA

Some Features Included in the Kings County Data Set:

* price - Price of the house (prediction target)
* bedrooms - Number of bedrooms
* bathrooms - Number of bathrooms
* sqft_living - Square footage of the home
* sqft_lot - Square footage of the lot
* waterfront - Houses with a waterfront view
* condition - How good the condition is (Overall)
* grade - Overall grade of the house, based on King County grading system
* yr_built - Year that the house was built
* lat - Latitude coordinate
* long - Longitude coordinate
* sqft_living15 - Square footage of living space for the nearest 15 neighbors

Initial data contained data on 21,597 houses sold in Kings County in 2014 and 2015

DATA

Kings County GIS Data:

Includes locations of:

*Airports
*Cemeteries
*Commercial Farms
*Places of Culture
*Places of Education
*Fire / Police Station
*Gated Residential Areas
*Public Gathering Spaces
*Utilities

Data from: https://gis-kingcounty.opendata.arcgis.com/

DATA ANALYSIS

DATA CLEANING

Remove unnecessary columns, Take care of unusual values or null values

DATA SELECTION

Outliers (+/- 3 stdev) and multicollinear features were removed

DATA INSPECTION

Visualizations and correlation matrices were created to understand data

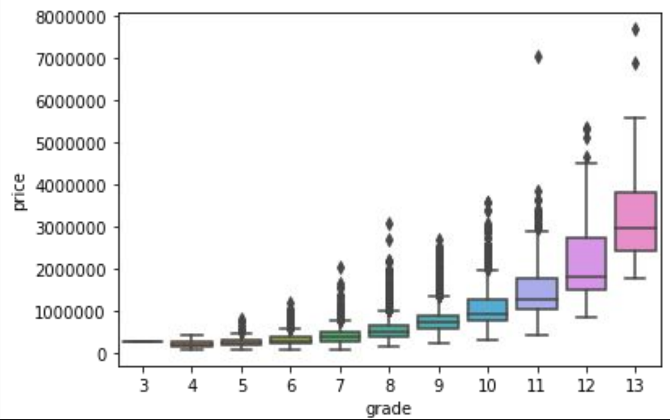
DATA TRANSFORM

After inspecting data visualizations, transformed and scaled some variables

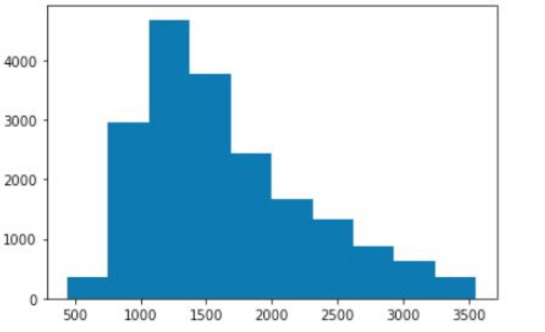
REPEAT AS NECESSARY

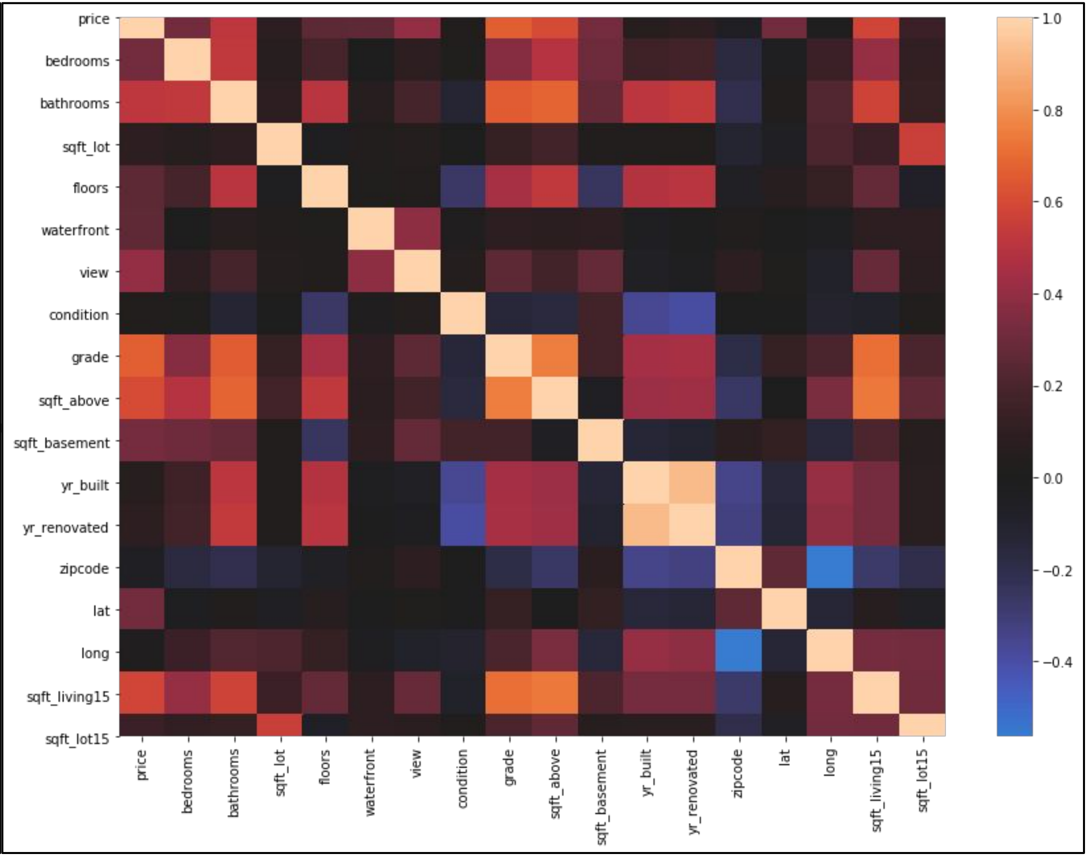
Some steps were repeated until sufficient results were generated

DATA ANALYSIS: VISUALIZATIONS




```
In [176]: plt.hist('sqft_above', data=data);
          # log tranform
```





BASELINE MODEL

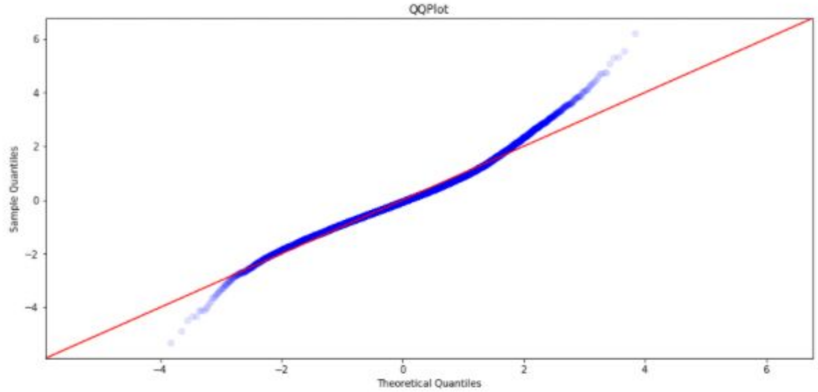
Began with a model that has

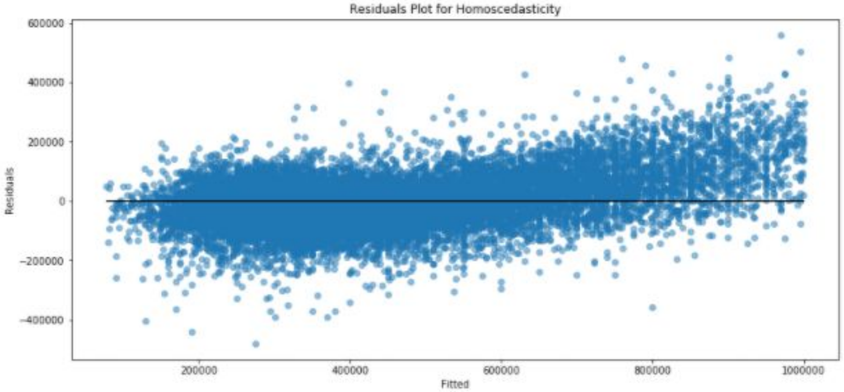
- no transformations
- no feature selection
- no removal of similar features

This resulted in:

Adj. R^2: 0.785

And the following plots:

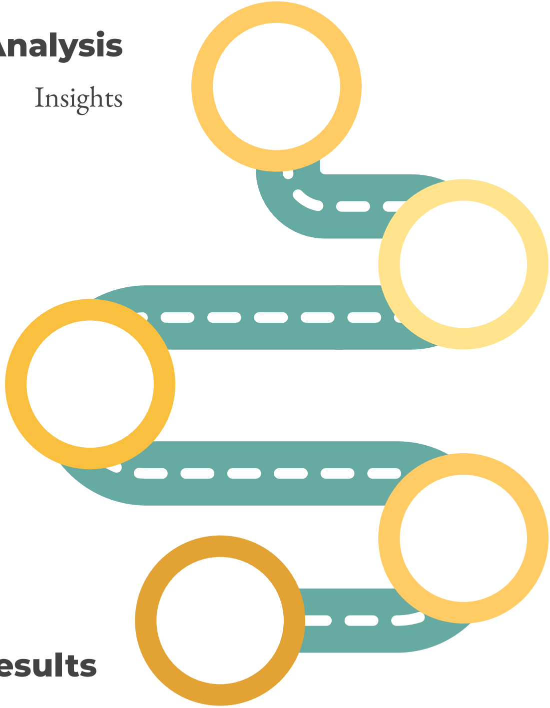




Creating a Model

Data Analysis

Insights



Feature Selection

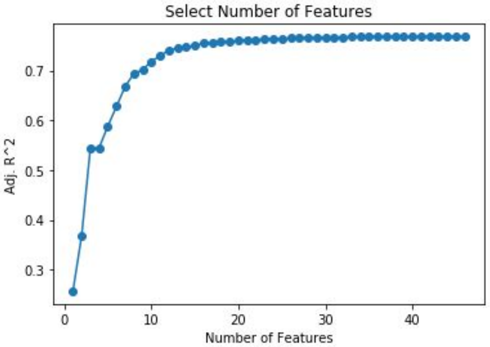
- Removing high p-values
- Remove multicollinear features
- Stepwise: add low / remove high
- Recursive Feature Elimination

Transformations

- One-Hot-Encoding / Binning
- Normalising / Scaling
- Logging
- Polynomials

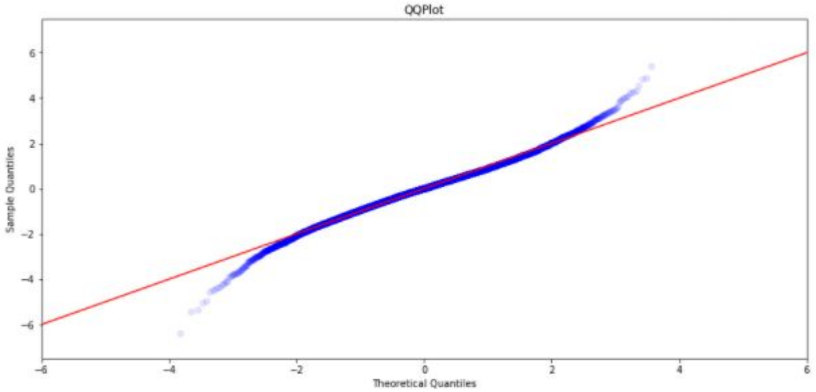
Train Model

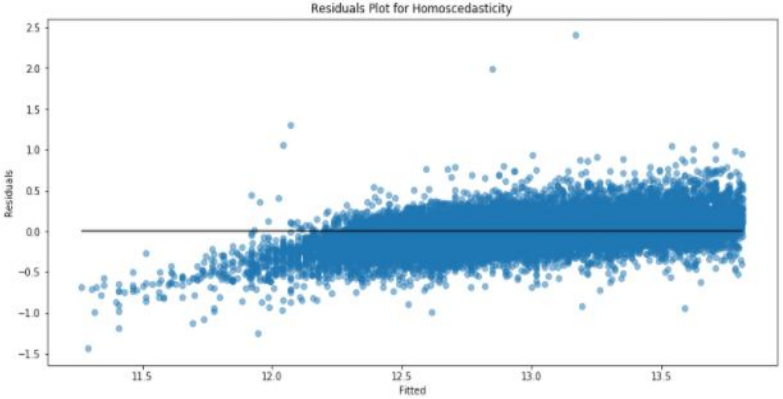
- Ordinary Least Squares



Results

FINAL MODEL





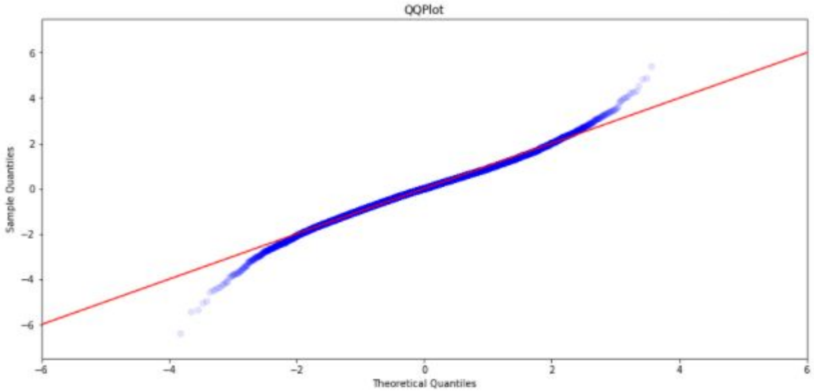
Final Model:

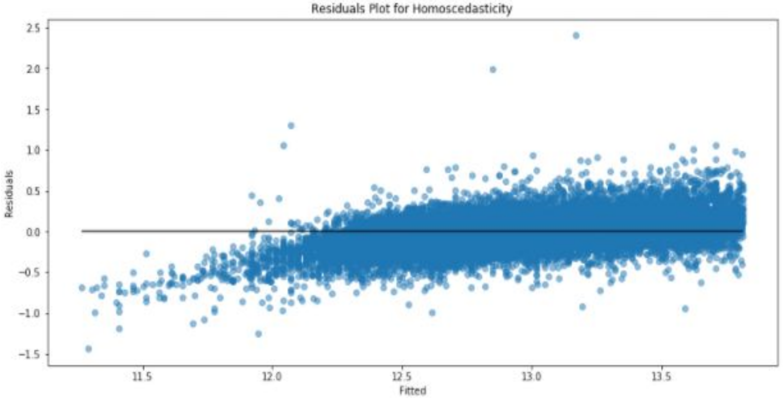
Adj. R^2: 0.741

RMSE (train): $104127

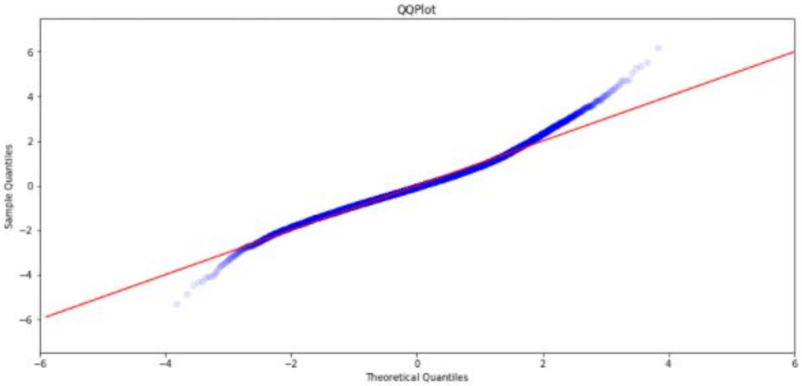
RMSE (test): $118296

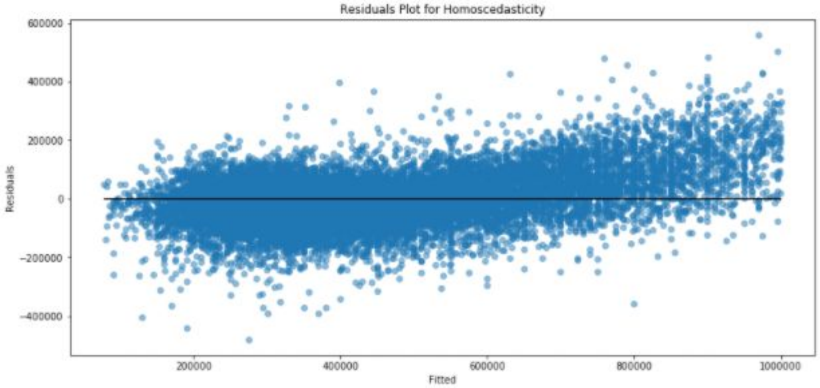
FINAL MODEL



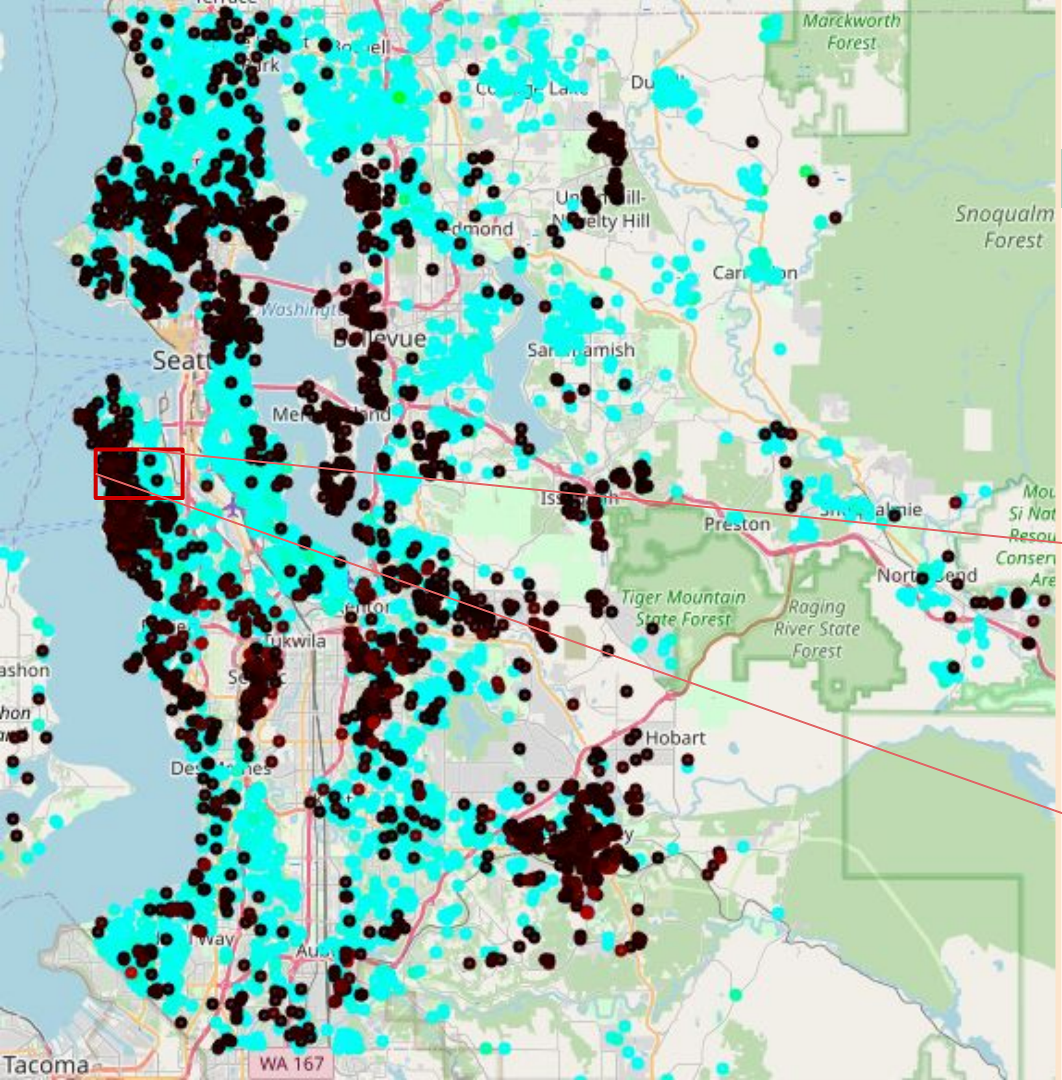


Final Model:
Adj. R^2: 0.741
RMSE (training): $104127
RMSE (test): $118296





Base Model:
Adj. R^2: 0.785
RMSE (training): $90503
RMSE (test): $92445

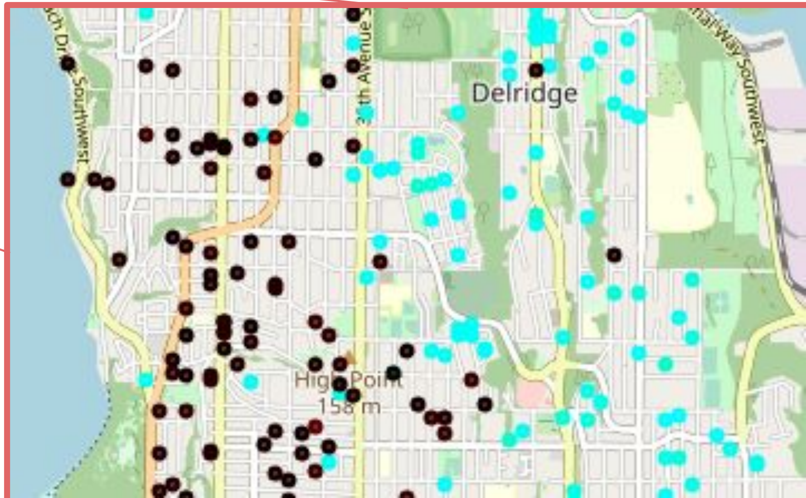


Percentage Error Map

Black - **Red** Largest **Under**estimates -Increasing

Cyan - **Green** Largest **Over**estimates -Increasing

Note: Under/Overestimates as percentage of actual price, not absolute



INSIGHTS - POSITIVE COEFFICIENTS

Waterfront

Coeff. = $168,200

Grade

Coeff. = $46,320

Condition

Coeff. = $27,250

Educational

Coeff. = $17,220

INSIGHTS - NEGATIVE COEFFICIENTS

Government

Coeff. =
- $10,290

Access Point

Coeff. =
- $6,364

Bedrooms

Coeff. =
- $5,965

Seasonal Homes

Coeff. =
- $4,568

CONCLUSIONS / RECOMMENDATIONS

Quick Features

- Having a waterfront.
- Having a high grade and condition rating.

Square Footage

Sqft_living = $55.14
Sqft_lot = $0.32
=> around 170x. Larger lots may not be all that favorable.

Other Features

- Proximity to educational, government, access point buildings.
- Fewer bedrooms per sq. footage.

Future Studies

- Incorporating more variables into the model (i.e. taxes, garage)
- Using a nonlinear model or narrowing the range of prices

THANK YOU!

Contact us:
TaeJoon Kim (tjkim614@gmail.com)
Matthew Andrews (2maltanno@gmail.com)

Github repository:
https://github.com/Maltanno/Phase2_Project/tree/main