House Sales in King County

We Push to Master
Jaspreet Kang, Junyan Zheng, Melissa Duffus-Santos, Minsu Jeon

Overview

Background

- King County, particularly Seattle, is one of the fastest growing cities in America.

-As a result, housing market in King County is of interest.

-We will be looking at house sales in King County in 2014/2015.

Dataset

-The dataset contains 21,613 houses, and 21 variables.

-The response variable (variable being predicted) is Price (price the house was sold).

-Dataset contains no NA values. However, outliers are present.

Predictors

-Originally there are 20 predictors (such as bedrooms, bathrooms, sqft_living, etc).

-We remove 2 predictors (House ID, date house was sold).

-Added one predictor (city).

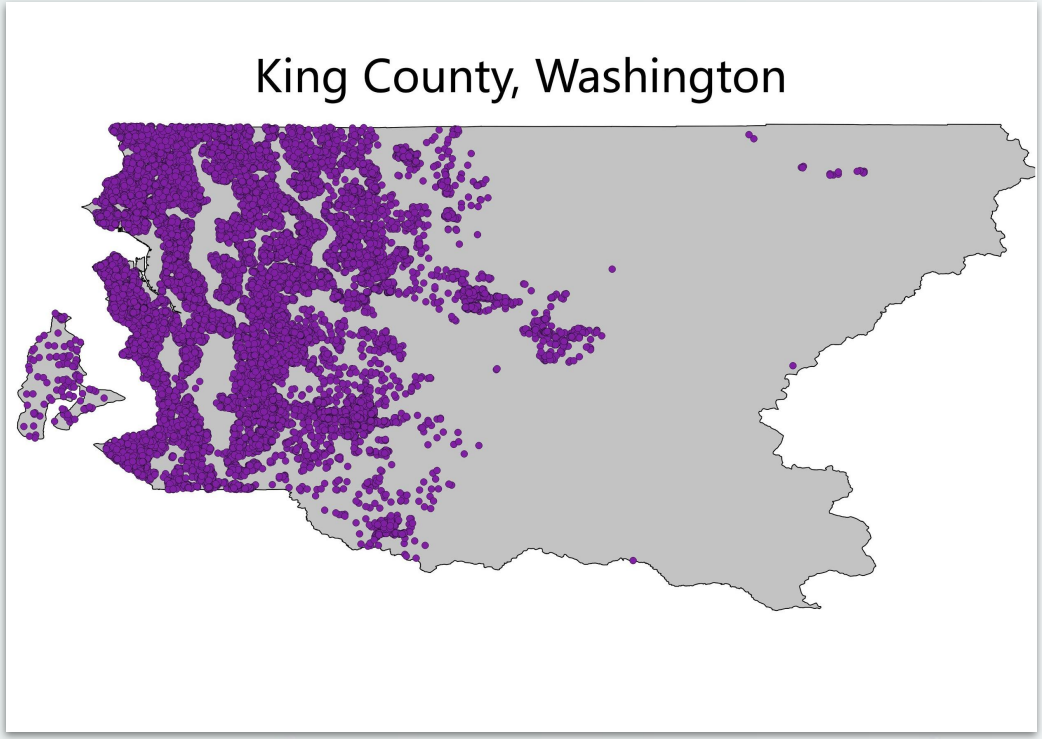
Correlation

-The highest correlation predictors to Price includes: sqft_living, grade, sqft_above, sqft_living15

-Many predictors highly correlated with each other.

-Ex: sqft_living and sqft_living15

Map Visualization



-There are 70 unique zipcodes, and 24 unique cities.

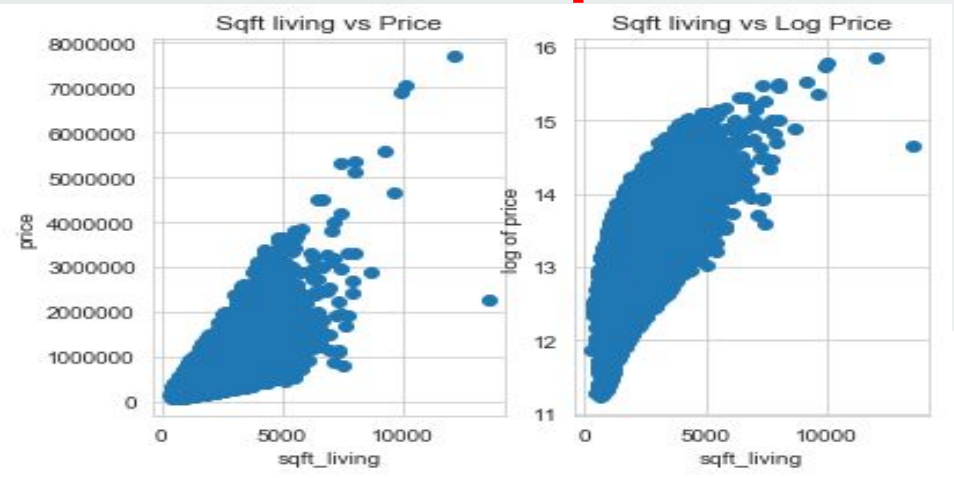
-These cities include: Seattle, Bellevue, Kent, Renton, Federal Way, and more.

-The three cheapest cities were: Federal Way, Auburn, and Enumclaw (median price came around $270,000)

-The three most expensive cities were: Bellevue($749,000), Mercer Island($993,750), and Medina($1.9 million).

-Seattle fell in the middle (median: $453,000)

Exploratory Analysis



Sqft Living vs (Log) Price

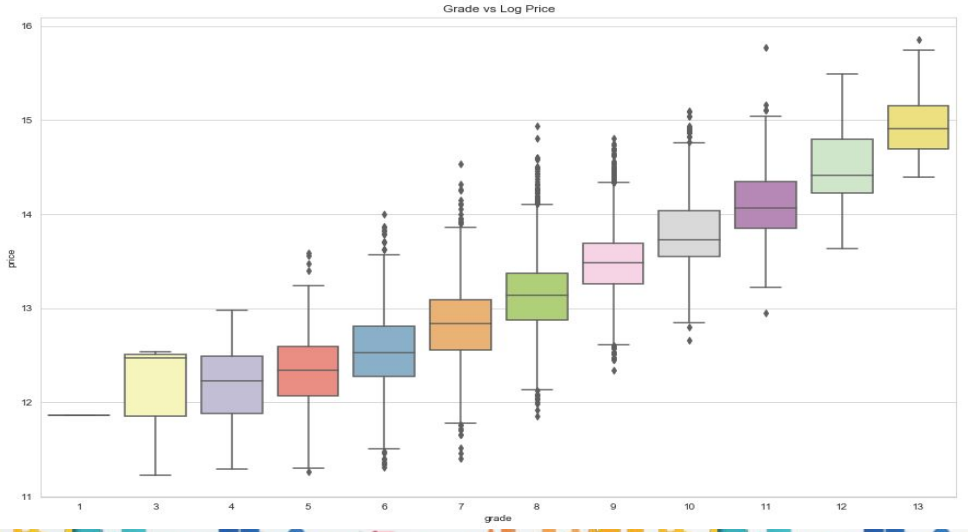
- A positive relationship is shown between House Price and Sqft living.

- Correlation between these two comes out to 0.70.

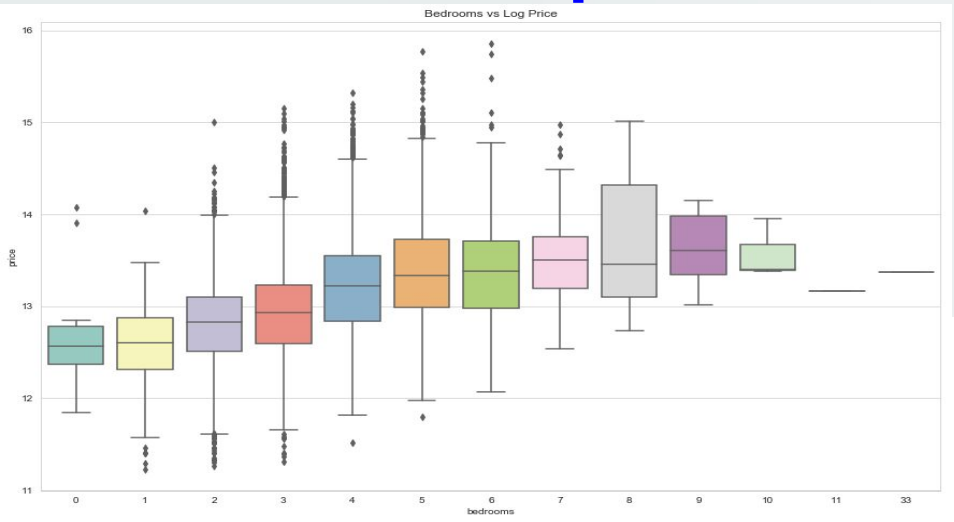
Grade vs Log Price

-With the exception of Grade 3 to Grade 4, as Grade of the House goes up, the House Price goes up.

-Correlation between these two comes out to 0.67.

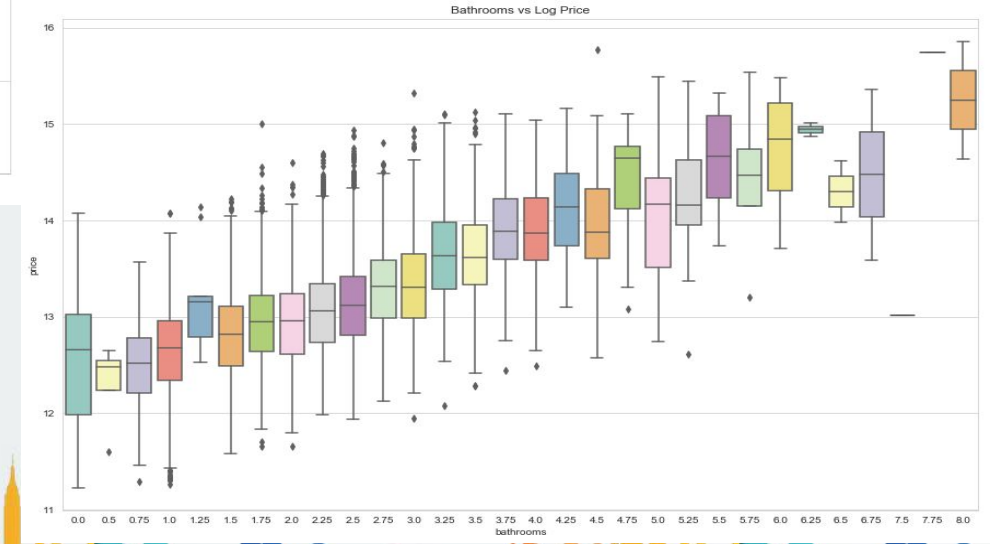


Exploratory Analysis



Bedrooms vs Log Price

- There a slight upward trend in House Prices as Bedrooms go up. Furthermore, the variability in house prices is higher in 2-6 bedrooms is higher than in 0-1 and 7-11 bedroom houses. With the higher variability, more outliers exist
- Correlation between these two comes out to 0.31.

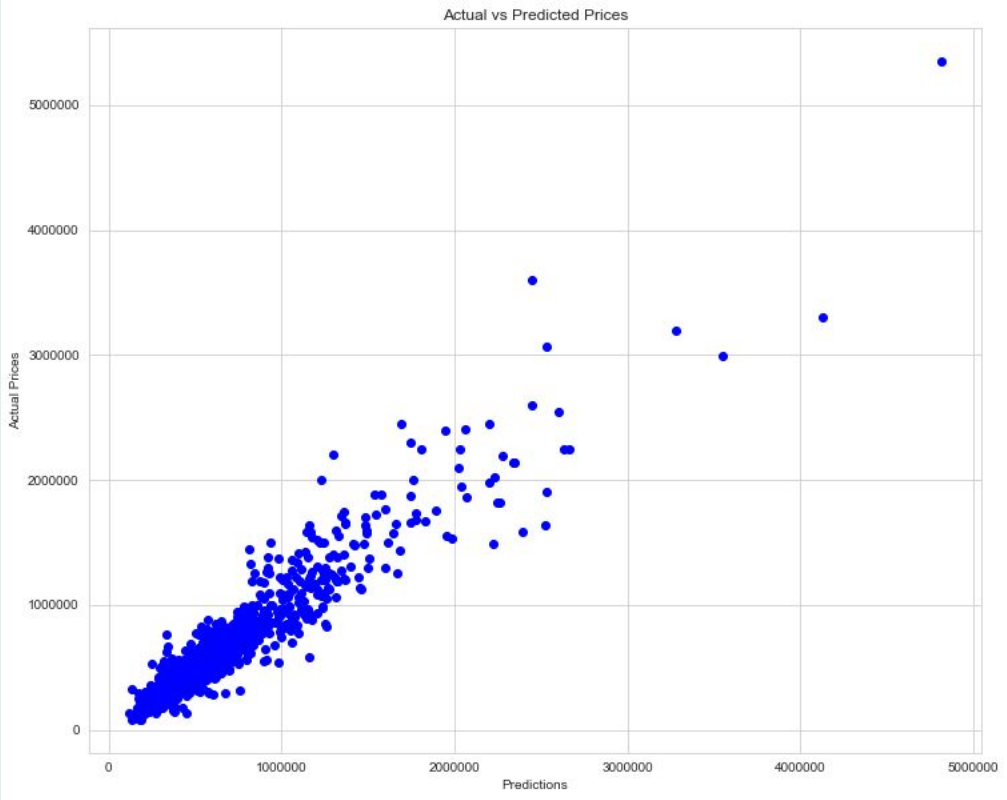


Bathrooms vs Log Price

-Generally, as number of bathrooms increases, House Prices increase.

-Correlation between these two comes out to 0.53.

Modeling the Data



-The Statistical Modeling method used for our final model was Gradient Boosting for regression.

-First split the data into two sets: 90% training, 10% testing. Further performed 20-fold cross-validation to ensure the model was not overfitting.

-Results came out to an average testing R-squared value of around 89%.

-Plot on the left shows a strong linear relationship between our predictions and the observations in our testing set. Therefore, our model does well in predicting House Prices.

Stats 131
Final Project

Thank you

For listening, and thank you for teaching us Python