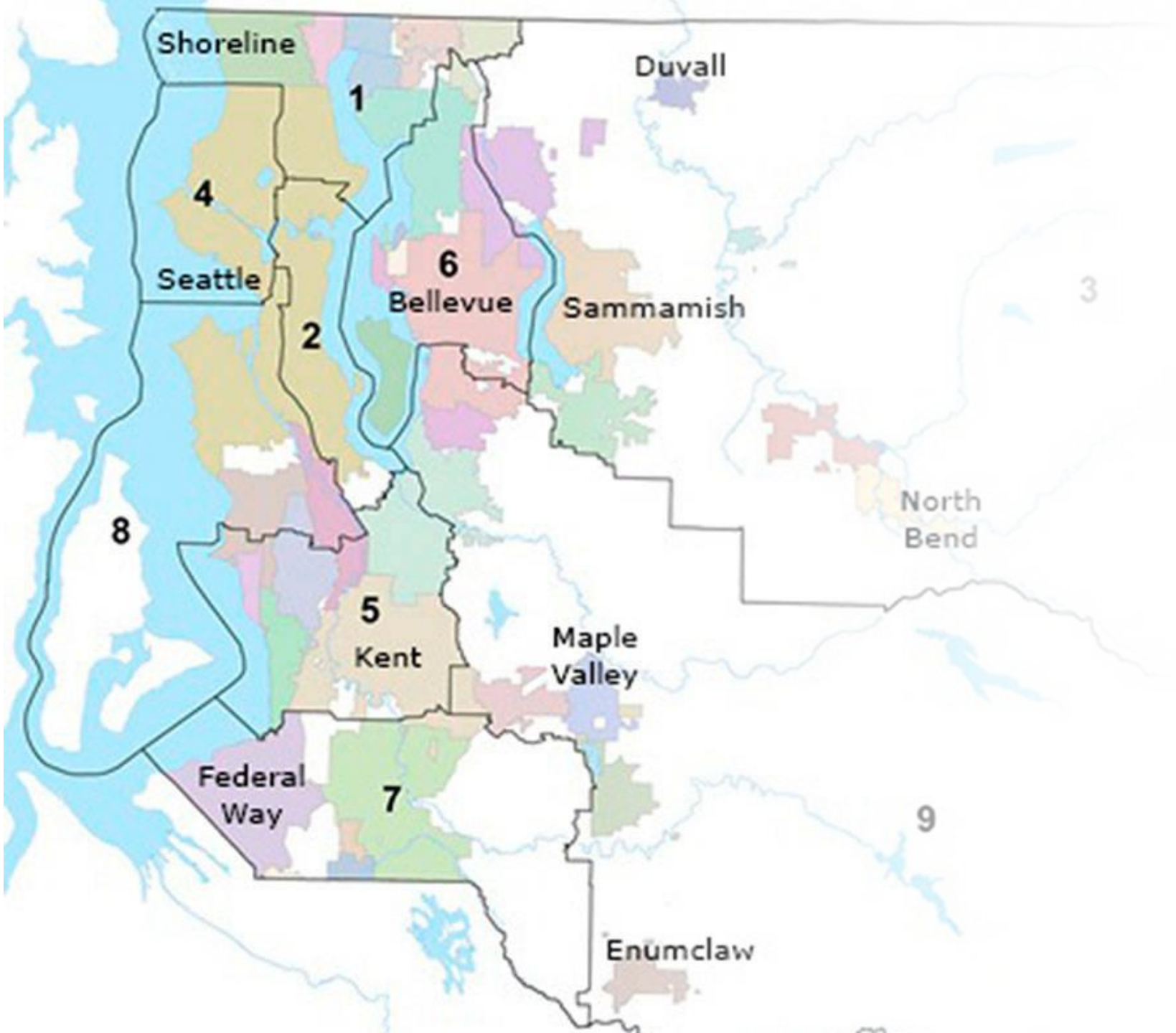


Predicting Housing Prices

Data Analysis of King County House Prices Data

Elissa Lee

July, 27 2020

Problem Statement

- Customer has recently put their waterfront property on the market. Management has asked the analytics team to **predict the house price** based on previous housing data.

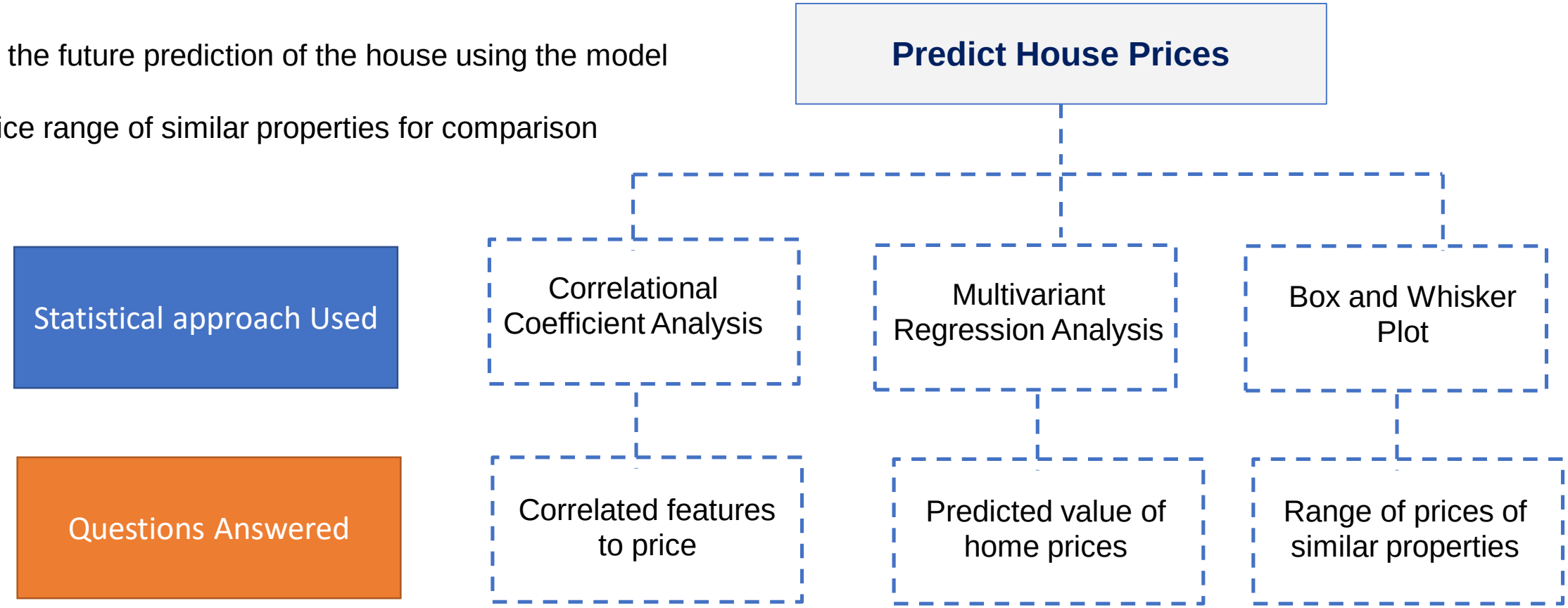
Features

Bedrooms: 4
Baths: 3.5
Sq Ft Living: 3603
Sqft Lot: 2253.5
Sqft Above: 3603
Waterfront: Yes
View: 4
Lat: 47.7419242
Long= 122.2842920
Built: 1977
Remodeled: 2004
Floor: 3

Price: ????

Analysis Procedure

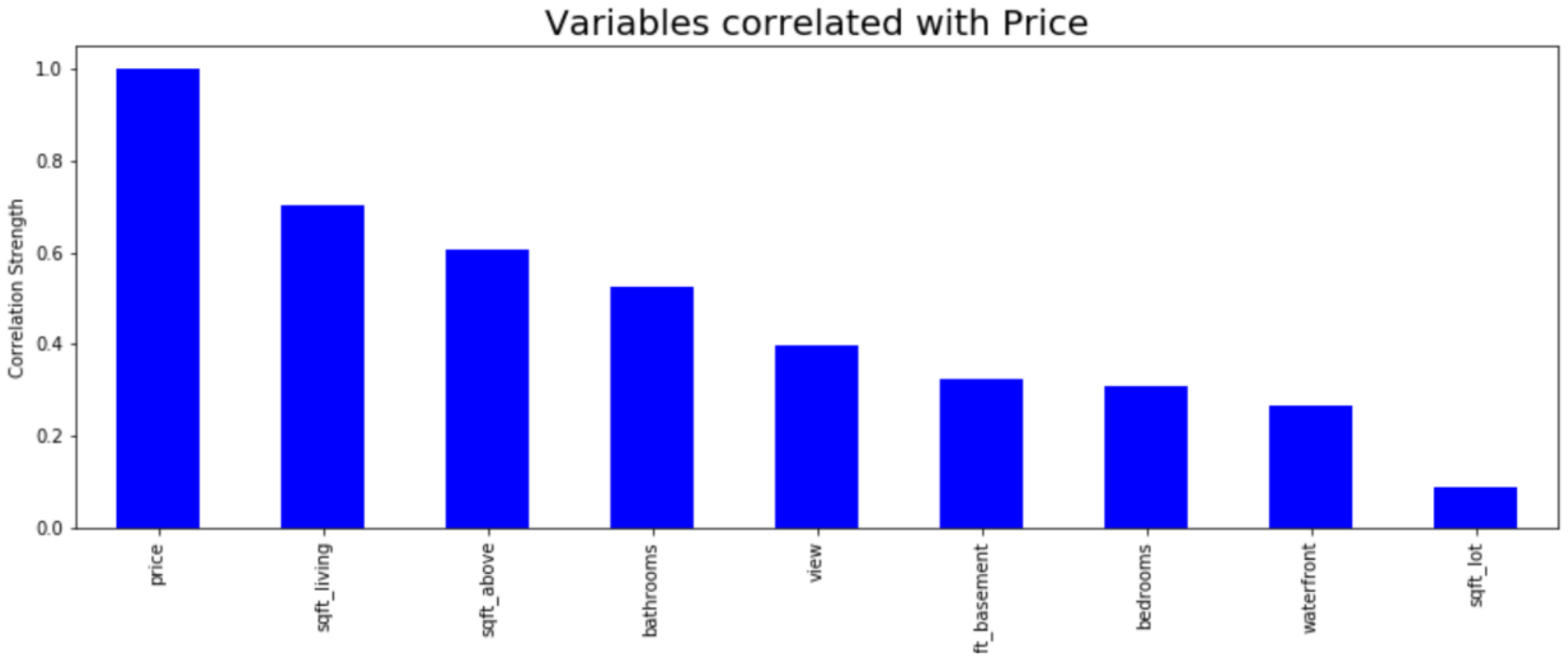
1. Identify attributes that strongly correlated to price
2. Create a model to predict housing prices using multivariant regression equation.
3. Assess the accuracy of the model
4. Making the future prediction of the house using the model
5. Find price range of similar properties for comparison



Key Correlated Features

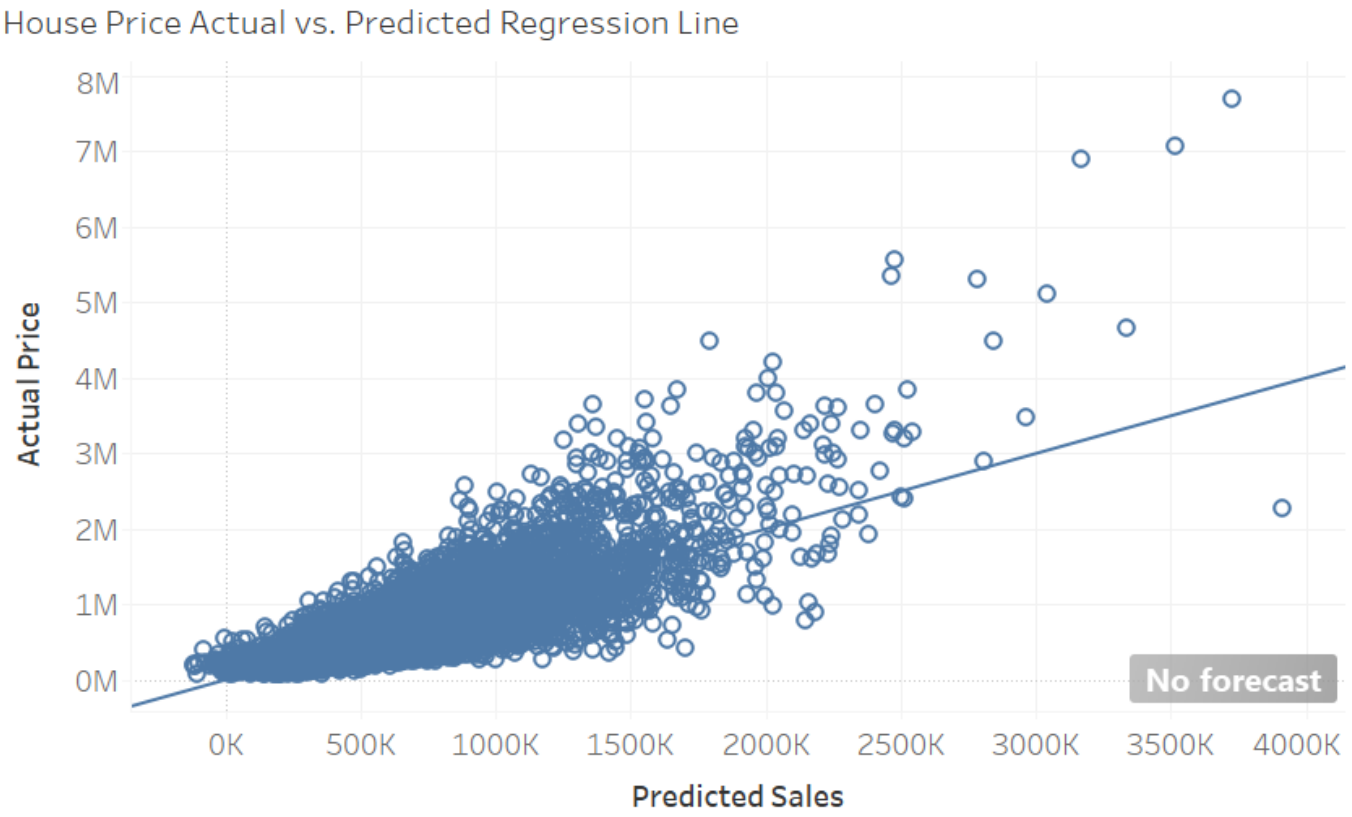
- Through correlational coefficient analysis, we were able to infer that:

    1. sqft_living (r=0.7), sqft_lot(r=0.61), and sqft_above(r=0.53) had a high correlation with housing prices
    2. sqft_basement (r=0.4), waterfront(r=0.32), and view(0.31) had a medium correlation with housing prices
    3. Suprisingly bedrooms (r=0.27) and bathrooms(r=0.09) had a small or no correlation to housing prices.



Create a Model

- 13 variables were used to create a multivariant regression equation. We used the 80-20 train-test rule; 80% of the data to train the model and 20% of the data to test the model. (Train the model means create the model, and Test the model means test the accuracy of the model.)



Result of the Model

- 66% of predictions can be made through this statistical model, with **R^2 (coefficient of determination) of 0.66**

- The model predicts the home prices of King County with an **Root Mean Squared Error: 206758.90**, indicating that our model misses actual house prices by about $206K. This can be interpreted as a fairly small error given that the data's price ranges from 75K to about 7.7 (Million)

```
Mean Absolute Error(MAE): 134046.72123452649
Mean Squared Error(MSE): 42749245476.358826
Root Mean Squared Error: 206758.90664336283
R^2 (coefficient of determination) regression score: 0.6600979058732752
Intercept:  -49937583.93393242
Coefficient: [ 1.67185184e+02 -7.51123058e-02  1.20767548e+02  4.64176359e+01
  5.83071625e+05  6.59912624e+04 -4.45231003e+04  5.70894366e+04
  6.44565192e+05 -1.56718437e+05  2.08338151e+03 -8.63325664e+02
  8.38695586e+03]
```

Predict the House Price

- The multivariant regression equation was applied to predict the future price of a house, located in Lake Forest Park, WA.
- The actual market price was $2,038,860 in 2015 which shows that the model was a close prediction but had an error of $556,140




Features

Bedrooms: 4
Baths: 3.5
Sq Ft Living: 3603
Sqft Lot: 2253.5
Sqft Above: 3603
Waterfront: Yes
View: 4
Lat: 47.7419242
Long= 122.2842920
Built: 1977
Remodeled: 2004
Floor: 3

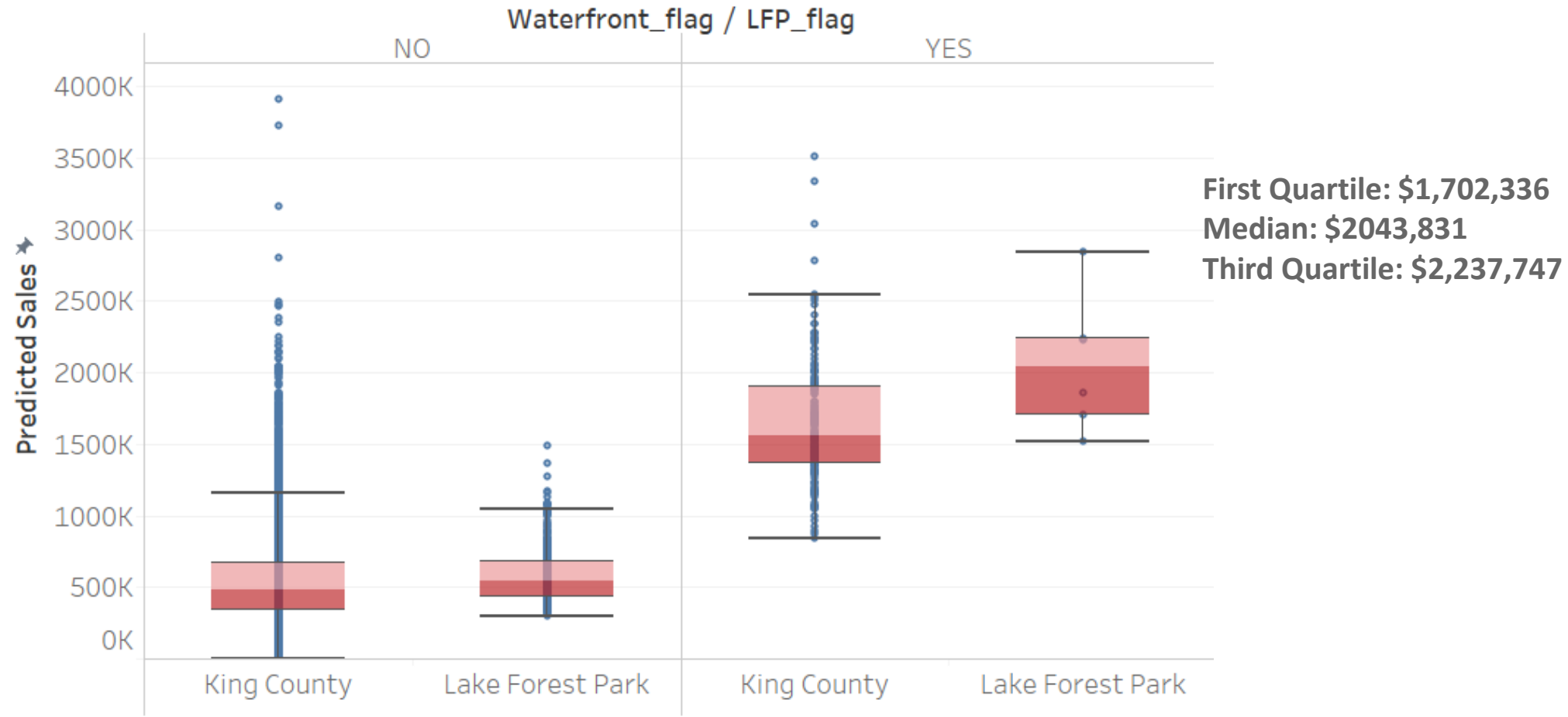
Price:

$2,038,860

Find Price Range

- The predicted price range of waterfront properties in Lake Forest Park are commonly between $1.7 ~ $2.2M with a median price of $2.0M
- The predicted price from the linear model perfectly fits near the median price range from this boxplot.

Waterfront Price Range Boxplot



First Quartile: $1,702,336
Median: $2043,831
Third Quartile: $2,237,747

Conclusion

- The predicted house price was **$2,038,860**, with **RMSE ~206K**

- This model was a fairly good prediction with a **R^2 of 0.66, but could be better**

Key Learnings

- More data on the waterfront property would improve the price prediction of waterfront houses.
  - Normally, attributes such as number of garages, waterfront size (water_depth and water_width), and bike trail access, are consideration factors when pricing a waterfront house but were not available for this model.

- Next time, I will split the test into two separate regression equation by classifying houses into waterfront (Yes/No) and compare the accuracies of each model.
  - More attributes for each classified group would be required to generate an high quality prediction.