King County Housing Prices

By Kathy Lu

Welcome to my presentation on King County Housing Prices! My name is Kathy Lu, and I am studying Data Science at Flatiron School.
King County is a county in Seattle, Washington.

Problems to Solve

Here is a nice waterfront picture of King County's beautiful shoreline. Let's get into what problems we'd like to address with this dataset.

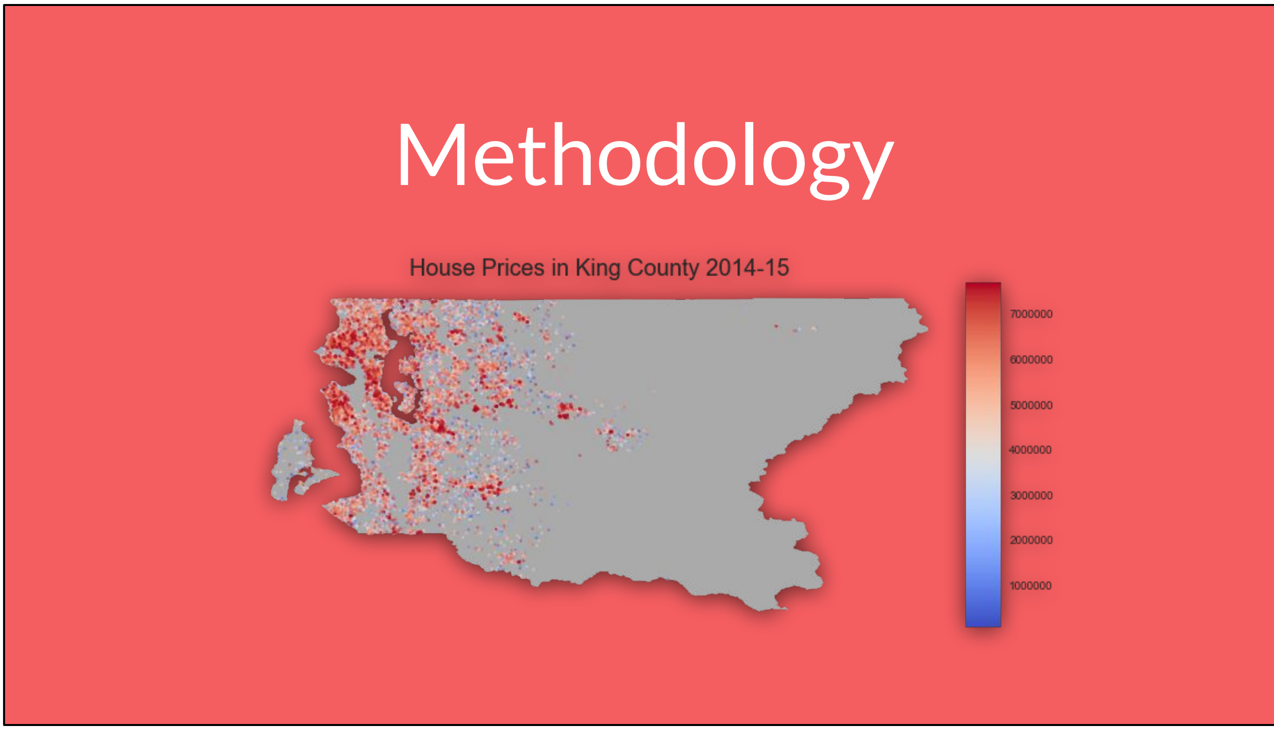
What Problems We Address and Why?

1. Estimating House Prices
   a. Helps realtors and owners price their home
   b. Helps state government tax appropriately
   c. Helps buyers estimate fair market price
2. Maximizing Selling Prices
   a. Helps realtors and owners increase profit
   b. Helps investors understand which renovations correspond to higher selling prices
   c. Helps developers understand what aspects correspond to high selling prices
3. Appreciation of House Value
   a. Helps investors understand which houses might experience appreciation of value over time

First of all, we wanted to try and estimate house prices for houses in King County. This information has business value because a model which predicts house selling prices accurately can help sellers price the house, help governments tax appropriately, and help buyers estimate a fair market price for the house they would like to buy.

We also looked into variables which could help maximize selling prices. Knowing which aspects of a house maximize price can help sellers increase profits from their investments or help developers build high value houses. If the house you're selling does not have these aspects which maximize price, you can have a good idea of which renovations might help you increase your house's selling price.

Lastly, we checked if there was any appreciation of house value over time. This dataset was created by looking at selling prices in King County from 2014 to 2015, so identifying which houses appreciated in value over this year can help investors maximize their return on investment.



Let’s move onto methodology. Here we have a visualization of where houses were sold in King County in 2014-2015, along with their selling prices. Dark red indicates pricier houses, and dark blue corresponds to houses with lowest price.

Methodology

- Data
  - Kaggle
- OSEMN framework
  - Scrubbing, Exploration
- Modelling
  - Minimize predictors
  - Maximize r-squared
  - Lower error
  - Ordinary Least Squares Regression

The open source data used in this study was taken from Kaggle, an online resource which provides open datasets.
I employed the OSEMN framework when investigating this data. After obtaining the data from Kaggle, I performed scrubbing and explorations using Pandas and Matplotlib visualizations. More details can be found in my Jupyter Notebook.
There were 19 features, or predictors, in the dataset, including: square footage, number of bedrooms, condition of house, latitude and longitude, etc.
The target was the selling price of the house.
During modelling, our goal was to minimize the predictors needed to create the final model, maximize r-squared, a measure of how well the model fits the data, and lower error in predictors. We used ordinary least squares regression to create the final model.

Conclusions

Here are our conclusions after investigating and modelling the data.

Predicting House Prices

- Predictors: Bedrooms, Grade, Views, Latitude, Sqft_living
- R-squared: 0.727
- P-value of all predictors: <0.05
- Mean squared error: 0.240 over 10 cross validations

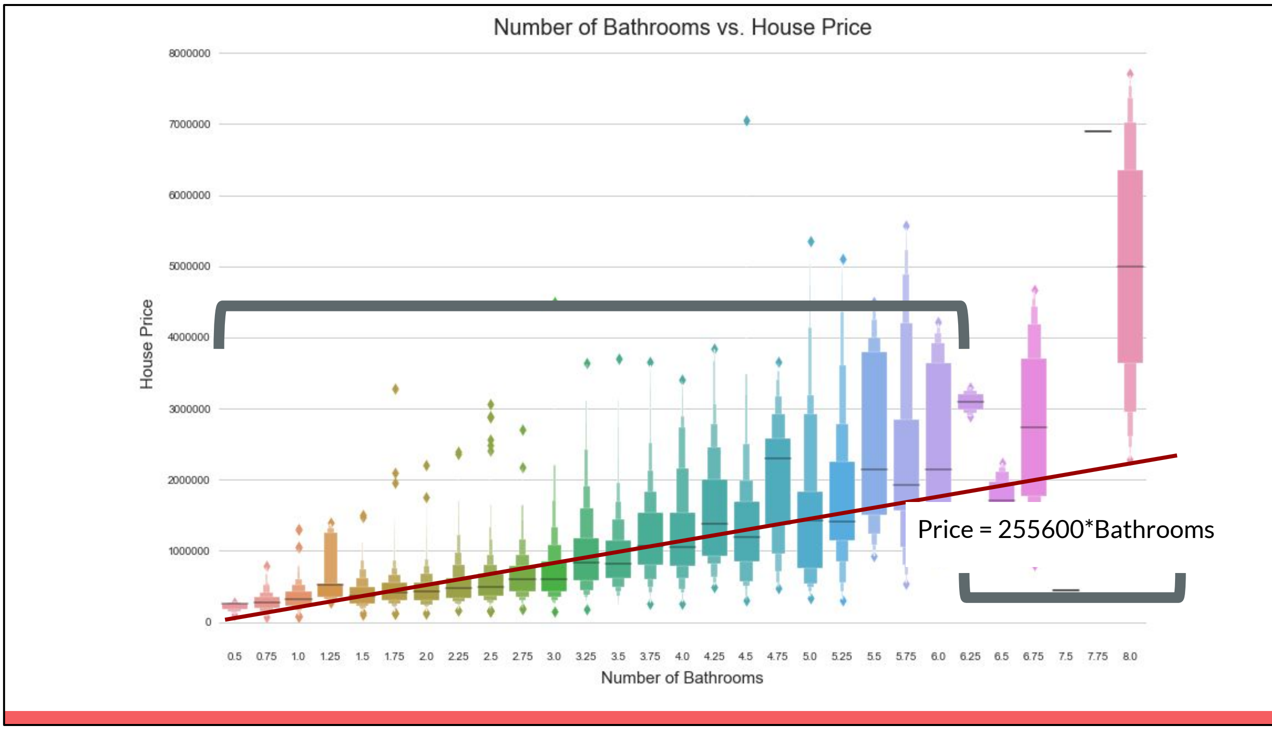
The predictors chosen to predict selling price were: number of bedrooms, grade rating of a house, number of views, latitude, and square footage of the house.
The R-squared was 0.727, and the p-value of all predictors used in analysis was less than 0.05, implying that they were all significant.
After using k-fold validation to test the model, we found it had an average mean squared error of 0.24, which is fairly low. Again, more details can be found in my Jupyter notebook.

Maximize Selling Prices

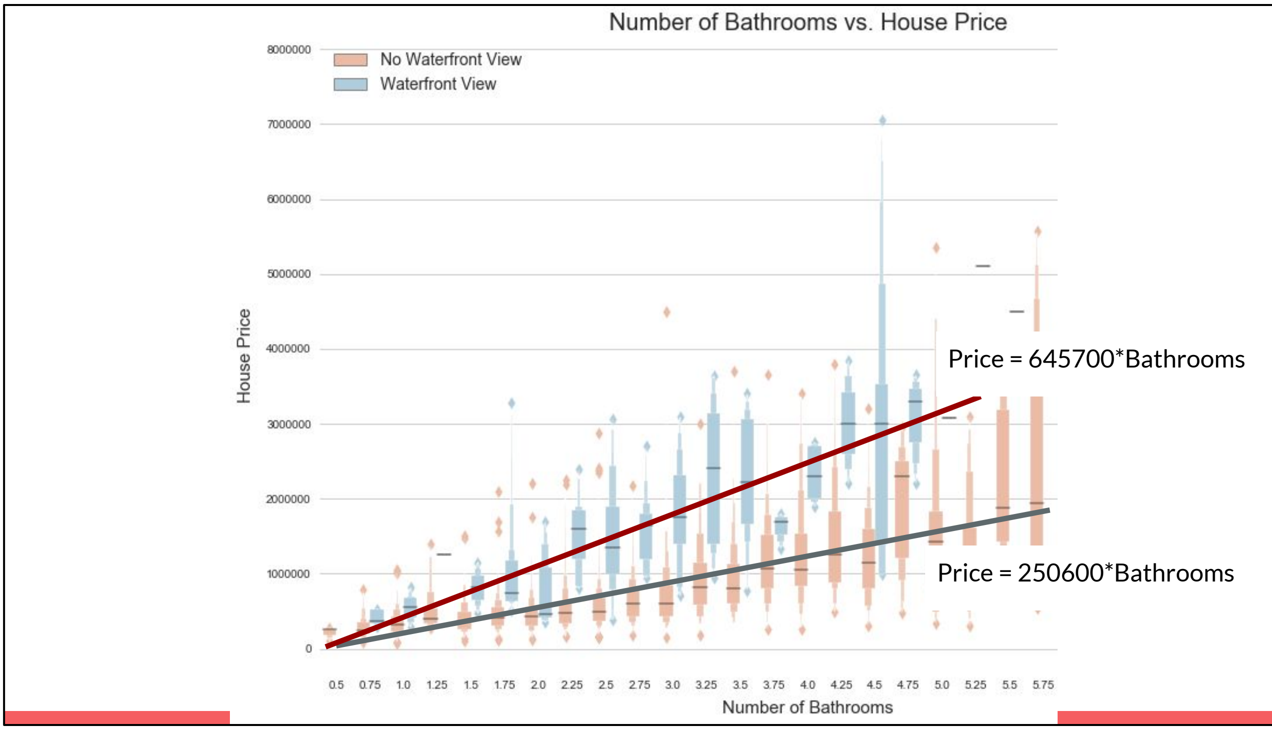
1. Bathrooms
2. Waterfront View
3. Season
4. Past Renovations
5. Condition of House

During the exploration process, I looked at many variables and their correlation with price. The variables I am addressing here are different from those I used to estimate house prices.
This might seem counterintuitive, but remember that estimating houses prices using a linear model and maximizing houses prices are different concepts.
Variables that correspond to higher selling price may not necessarily help when trying to predict low or mid-range house prices. These variables may not have a strictly linear relationship with house price, which was the model used to predict house price. Or, they may have corresponded too closely with other variables (multicollinearity) and been removed from the model. Lastly, there may have been missing data that rendered certain variables less useful in estimation. However, they are still important to visualize and discuss.

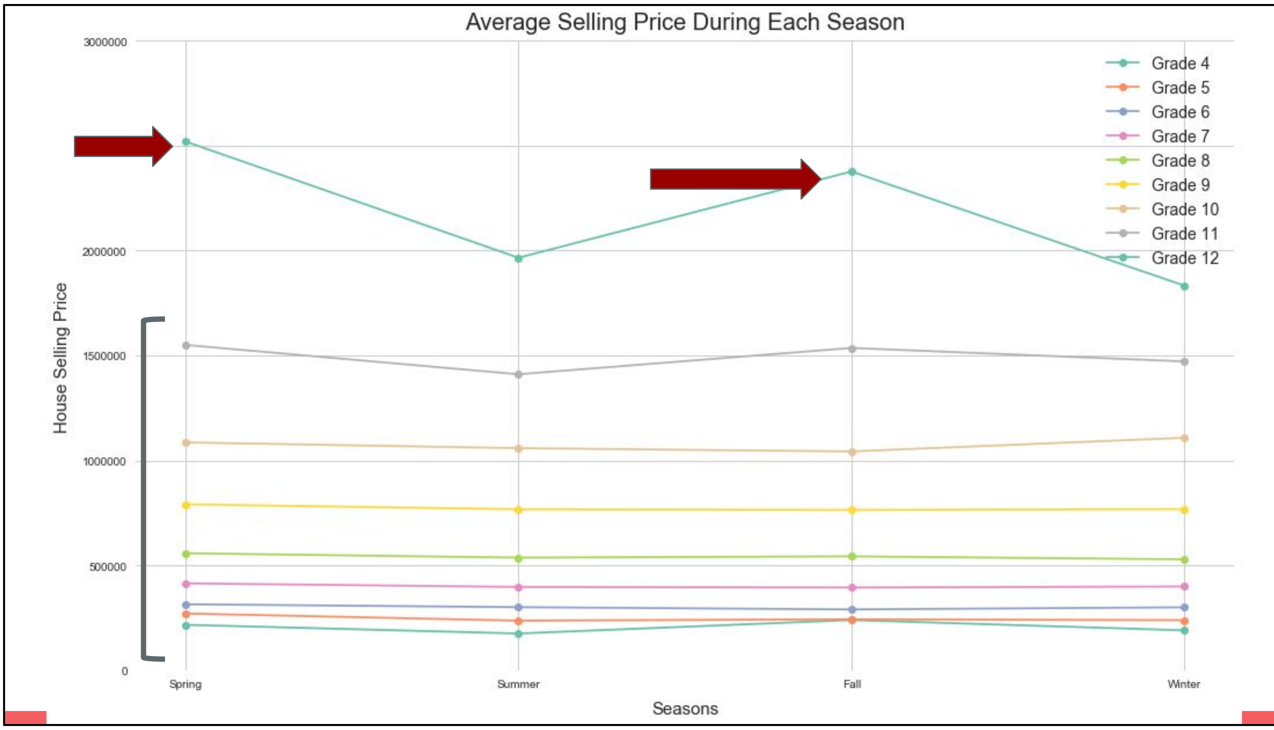
With that said, let's get into it!



Here we have a graph of the number of bathrooms in a house and its selling price. We can see <click> that there is a very clear increase in average selling price as the number of bathrooms increases. However, this pattern is less clear <click> in houses which have the most bathrooms, likely since there is less data for houses with more than 6 bathrooms. The linear model fitted <click> here (with p value <0.01) implies that there is a $255,600 increase in housing price corresponding with each extra full bathroom. It may be important to consider that adding a full bathroom to certain houses corresponds to an increase in house price of over 200 thousand dollars.



Here I have separated the houses into those with a waterfront view and those without. You can clearly tell that there is a difference in price between houses with a waterfront view and houses without. Furthermore, you can see that houses without a waterfront view <click> experience less increase in price with each extra bathroom compared to houses built with a waterfront view <click>. The linear model fit to houses without a waterfront view implies that there is a $250600 increase in selling price corresponding to each addition of a full bathroom. For houses with a waterfront view, there is a $645700 increase in selling price corresponding to each extra full bathroom. In other words, building a house with more bathrooms corresponds to a greater return on investment if that house has a waterfront view.
Both linear models had p values less than 0.01, implying that this relationship is significant.



The only significant relationship between seasons and selling price was found in houses of grade 12 <click> for fall and spring, where the season had a positive impact on price. Houses with grades 3 and 13 were removed due to lack of data. We can infer that for the vast majority of houses, there is not much significant variation in price due to the change in seasons<click>. The exception is houses of grade 12 <click>, where there is a significant increase in selling price during the fall and spring ($p<0.05$).

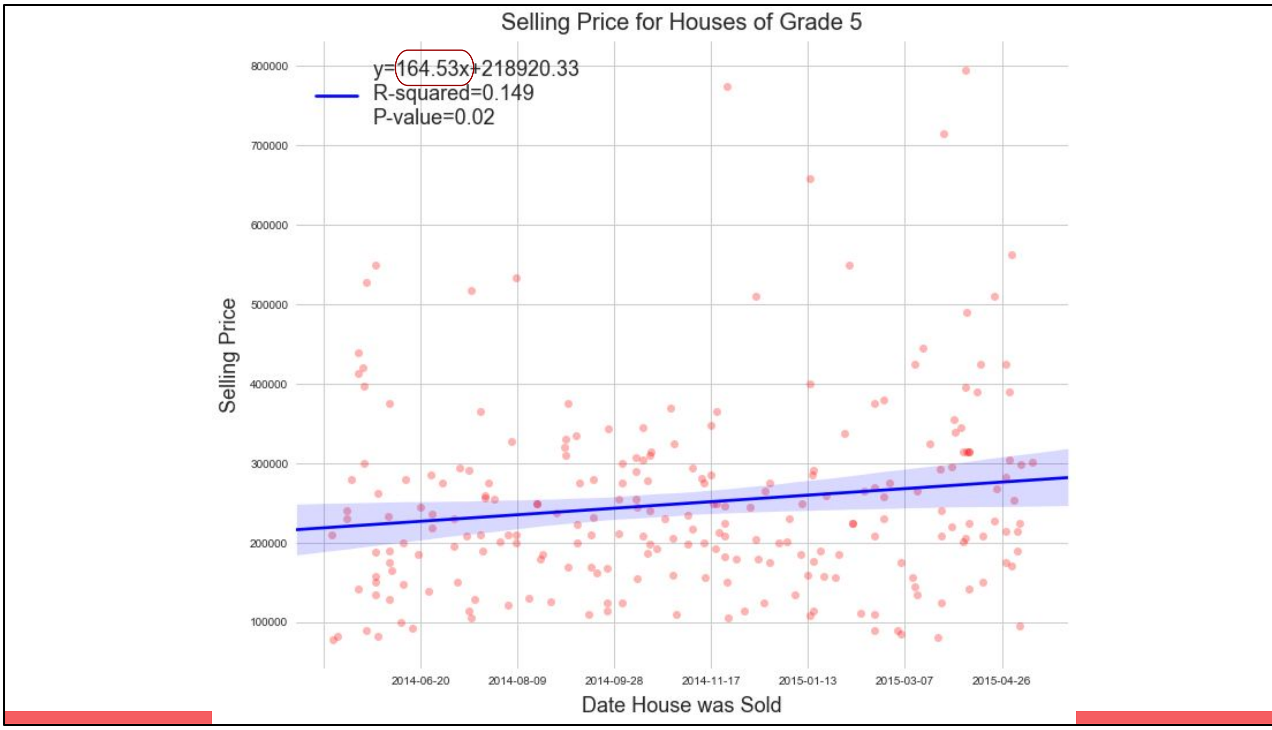
Therefore, we would recommend sellers with houses of grade 12 to consider selling during the fall and spring seasons to maximize selling price. However, sellers with houses of grades 4-11 should not be overly concerned with a loss in house price due to changes in season.

The average housing price is approximately $538,513, $530,776, and $767,198, for the categories of missing, no renovations, and past renovations, respectively.
It does seem that renovations correspond to an improved selling price ($p<0.01$), as we can see a massive increase in selling price (200k+) <click>when comparing houses which were not renovated previously, or have unknown renovation status, with houses which were previously renovated. Therefore, if a house has been previously renovated, this information should be researched, recorded, and presented to buyers in order to sell a house for more money. I would recommend sellers spend time researching their houses' previous renovations in order to improve their selling price.

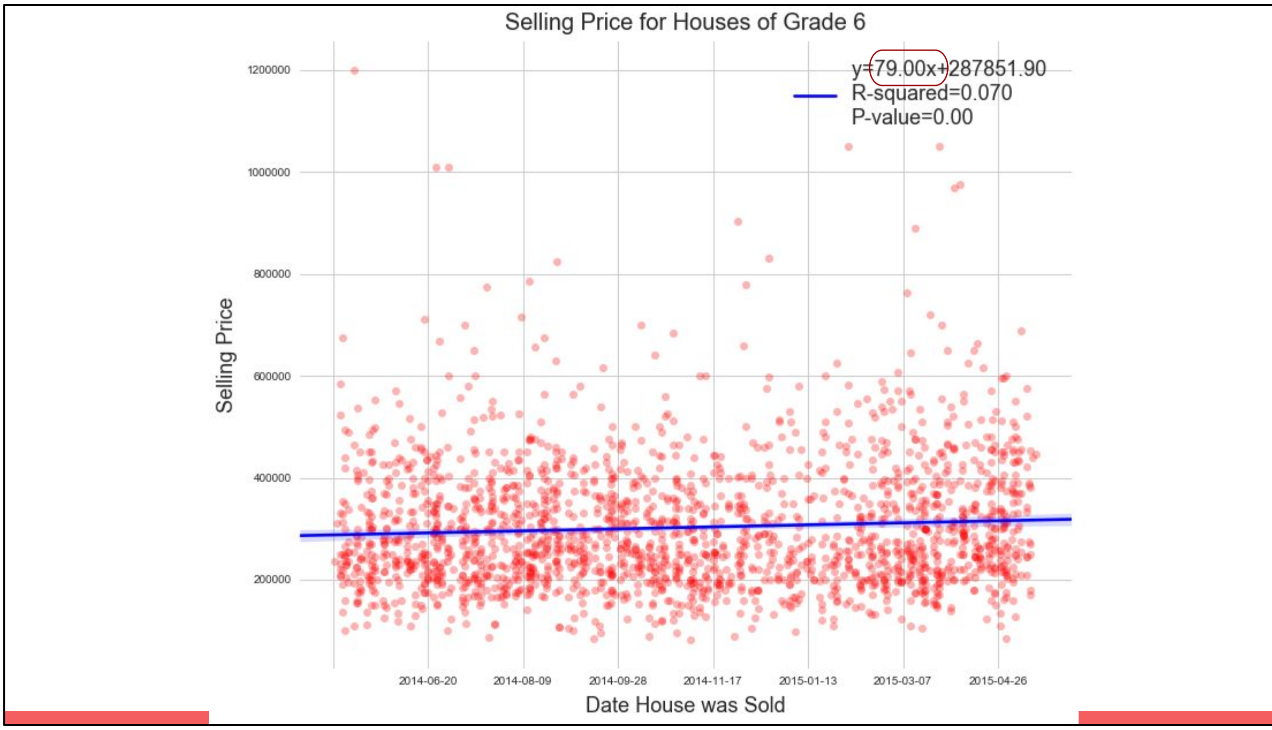
Appreciation of House Value

- Data over 1 year (2014-2015)
- Significant only for grades 5,6,7
- Nonsignificant if not parsed into grades

As mentioned previously, the data used in this presentation was collected over one year, from 2014-2015. We initially could not find evidence of appreciation in value over time. However, once the data was organized by house grade, we found that there was an increase in selling price for houses of grades 5, 6 and 7.



For houses of grade 5, there was an overall positive trend over time, with house gaining approximately 165$ <click> in value daily over the year.



Houses of grade 4 gained approximately 79$ <click> in value daily over the year.

Houses of grade 7 gained approximately 59$ in value daily. In summary, if an investors' interest is in short term real estate gain, it would be best to invest in houses of grades 5-7, with houses of grade 5 having the greatest daily return on investment.

Future Work

That ends my presentation of my data exploration on King County House prices. However, I do still have some future projects in mind.

- **Zipcodes with highest and lowest house prices**
- **Appreciation of House prices over a decade**
- **Effect of Renovations on House Price**
- **Random Forest Model**

Some of the projects I may take on in the future include: investigating locations of highest and lowest house prices in King County, collecting and investigating data on house prices over a decade, and not just a year, partnering with realtor agencies to see if there is a causal relationship between renovations and house price, and fitting a random forest model to better predict house prices.

Thank you for your time. Do you have any other questions for me?

Image Credits

https://www.racialequityalliance.org/jurisdictions/king-county-washington/ (shoreline)

https://www.movoto.com/for-sale/bothell-wa/westhill/open-houses/ (house)

https://twitter.com/kccouncil (logo)