King County Housing

By Deric Williamson & Eric Cusick

Business Case

A home renovation contract team has employed the services of data scientists to create a model predicting how much a house will sell for in the King County, WA area.

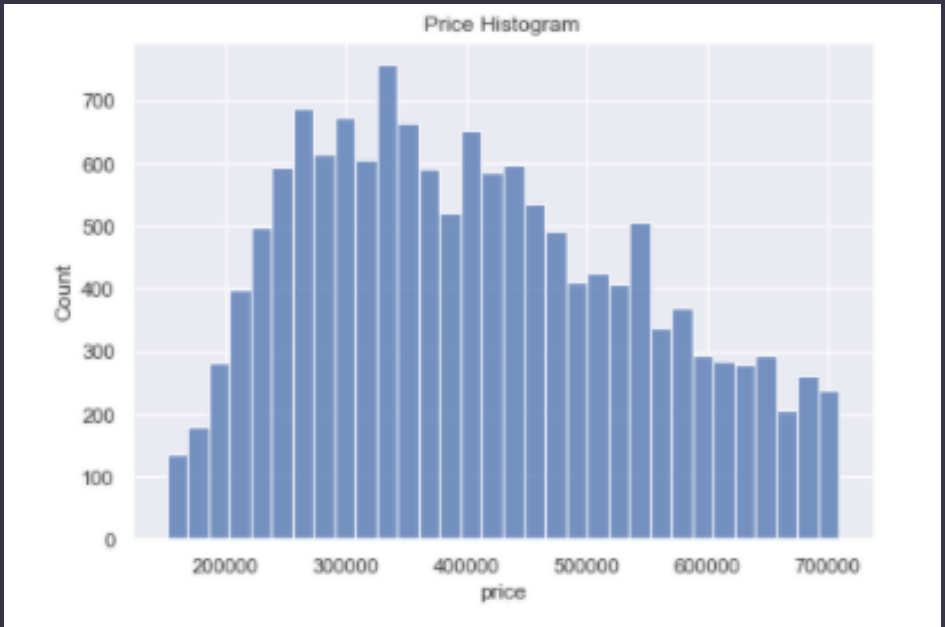
Given the 'kc_house_data.csv' file, the data scientist team is expected to:

- Build an accurate predictive model
- Report some of the primary features of a house that will increase the price so the company will know what to focus their efforts on
- Build a calculator that represents the model

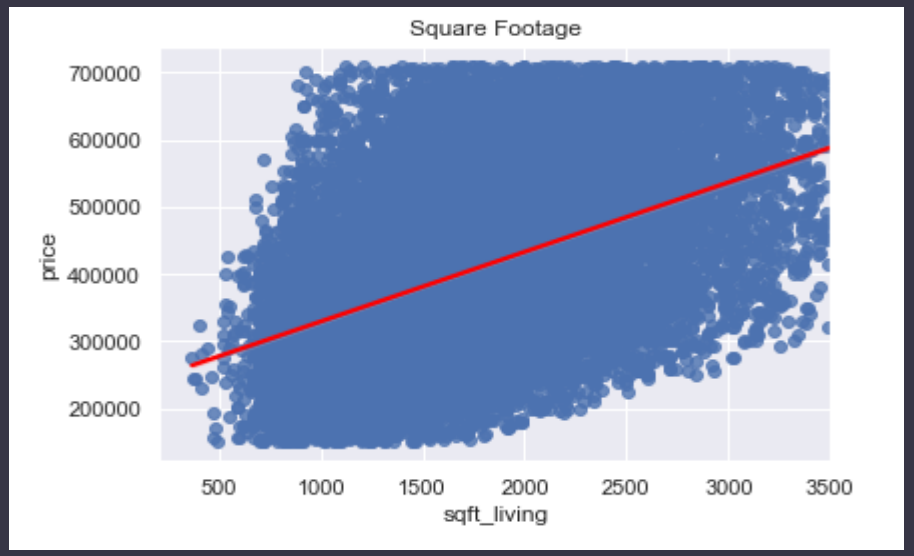
Exploratory Data Analysis (EDA)

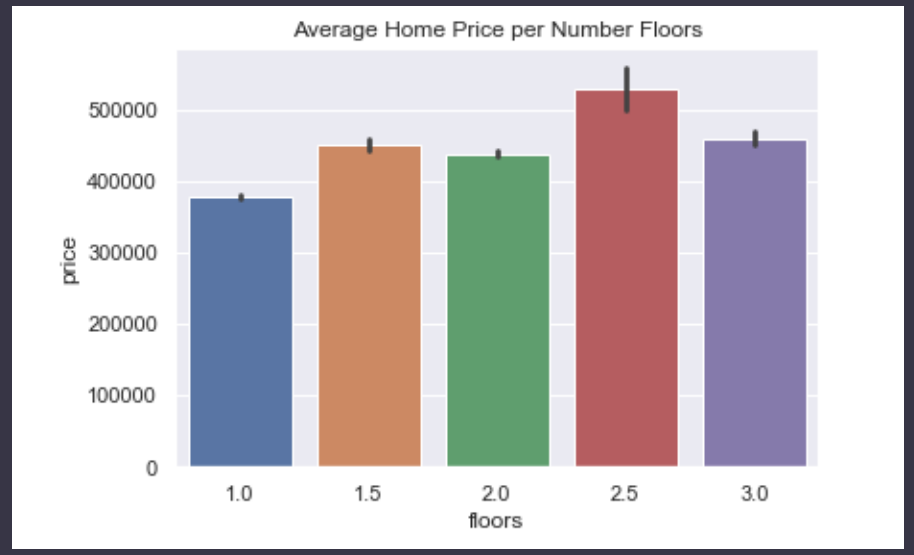
Using the provided data from: kc_house_data.csv

We restricted our data to provide a more meaningful relationship to our model by analyzing data from a house price range from $150,000 to $710,000.



Exploratory Data Analysis (EDA)





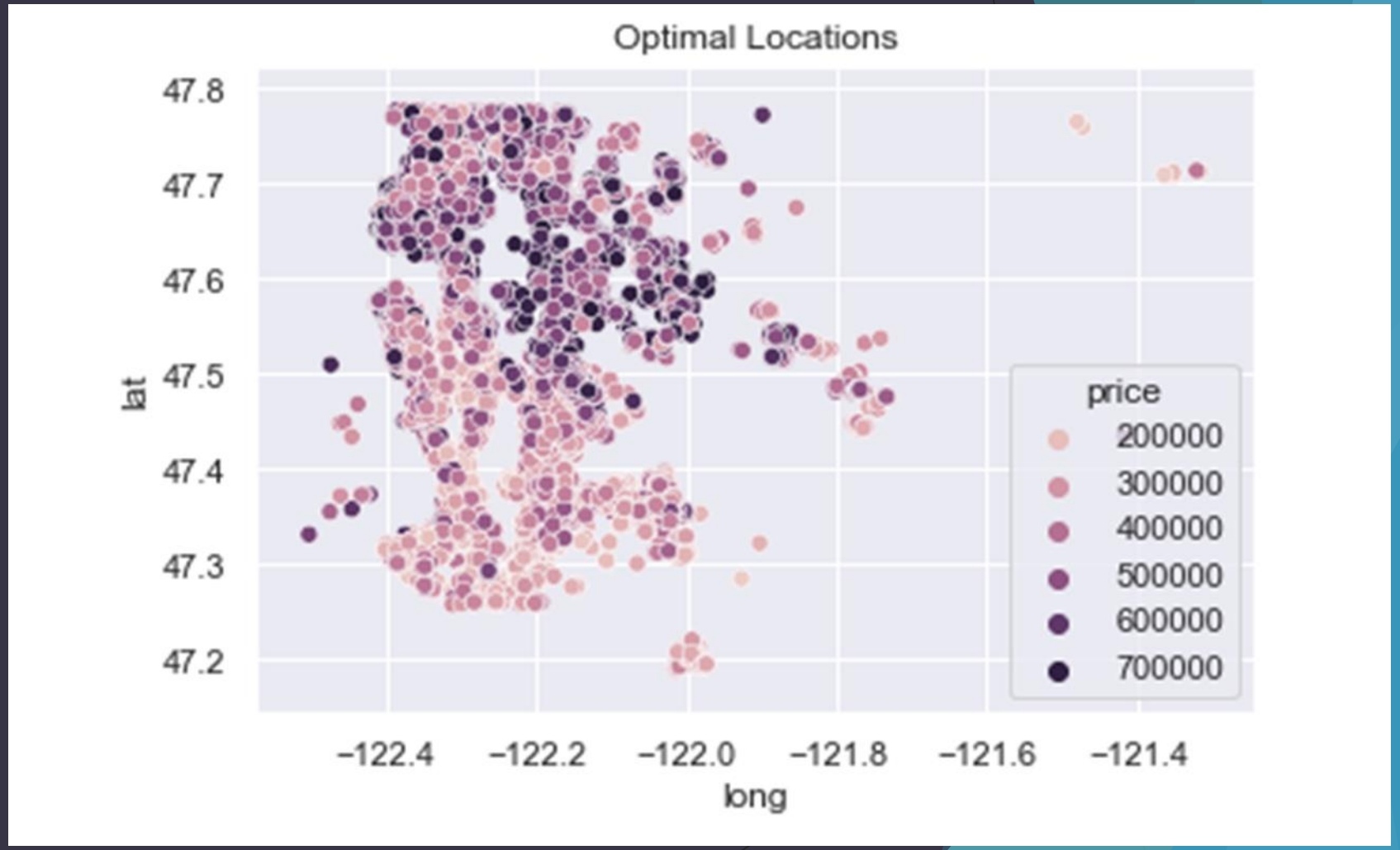
Some other house features were analyzed and corrected from possible outliers, data types, and missing data.

For regression accuracy purposes, we decided to drop house square footage data that was above 3,500 square feet. The graph reflects linearity.

On average, a house with more than one floor will sell at a higher price than a home with only a single floor.

Exploratory Data Analysis (EDA)

In King County, the location of the home can have a large impact on how expensive the home sells for. The graph shows the more expensive homes in the county are towards the northside, and lower priced homes are in the south.



Modeling Process

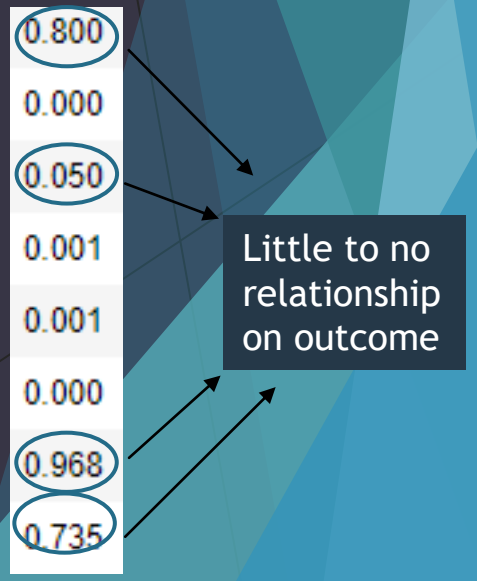
During the modeling process, we:

- Removed collinear variables
  - Highly correlated features might alter after changing one feature
- Detecting high P-values
  - High p-values assume there is little to no relationship in outcome
- Log transformed house square foot
  - Will help with model accuracy
- Categorized zipcode into four categories
  - Zipcodes were categorized by price and frequencies found in raw data

High Collinear Examples:

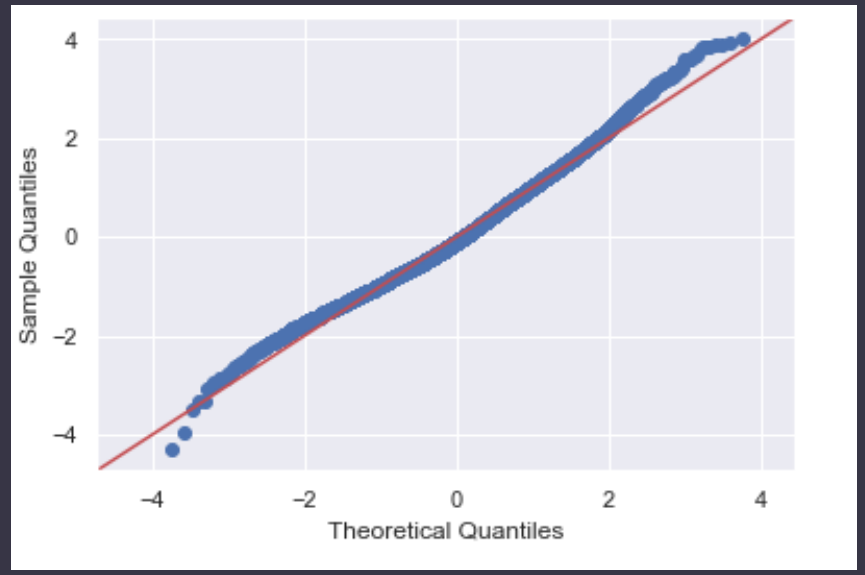
	pairs
(sqft_lot15, sqft_lot)	0.866612
(sqft_above, sqft_living)	0.819687
(sqft_living15, sqft_living)	0.703033

P-Values



Final Model

- R-Squared is a statistical measure between 0 and 1 which calculates how well the regression line fits in our data set.
- A coefficient is an indicator for each unit of “x” that is equal to the difference in outcome



Next our recommendations…

Dep. Variable:	price	R-squared:	0.598
Model:	OLS	Adj. R-squared:	0.598
Method:	Least Squares	F-statistic:	1885.
Date:	Fri, 27 Nov 2020	Prob (F-statistic):	0.00
Time:	23:27:53	Log-Likelihood:	-1.4608e+05
No. Observations:	11416	AIC:	2.922e+05
Df Residuals:	11406	BIC:	2.923e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.448e+05	7.78e+04	9.568	0.000	5.92e+05	8.97e+05
bedrooms	-1.242e+04	1324.159	-9.379	0.000	-1.5e+04	-9823.469
bathrooms	1.603e+04	2167.048	7.399	0.000	1.18e+04	2.03e+04
floors	2.141e+04	2078.515	10.301	0.000	1.73e+04	2.55e+04
yr_built	-892.1104	38.789	-22.999	0.000	-968.143	-816.077
has_basement	9442.8296	1959.058	4.820	0.000	5602.738	1.33e+04
zipcode_type_cheap_low_volume	-3.168e+04	2083.131	-15.207	0.000	-3.58e+04	-2.76e+04
zipcode_type_expensive_high_volume	1.509e+05	2999.175	50.307	0.000	1.45e+05	1.57e+05
zipcode_type_expensive_low_volume	1.494e+05	2265.757	65.935	0.000	1.45e+05	1.54e+05
sqft_living_log	1.821e+05	3845.515	47.359	0.000	1.75e+05	1.9e+05

Omnibus:	297.419	Durbin-Watson:	2.019
Prob(Omnibus):	0.000	Jarque-Bera (JB):	322.498
Skew:	0.394	Prob(JB):	9.34e-71
Kurtosis:	3.239	Cond. No.	1.88e+05

Recommendations

Looking at our model, we recommend:

- Size, location and the number of floors have the biggest influences on house prices
  - More expensive homes are located on the northside of the county, while less expensive homes are in the south
- It's better to have an additional bathroom over an additional bedroom
- Year Built has the lowest impact on resale price

We believe that the renovation team will be able to utilize this information to make sound decisions, to potentially increase their profits and work with efficiency when fixing homes.

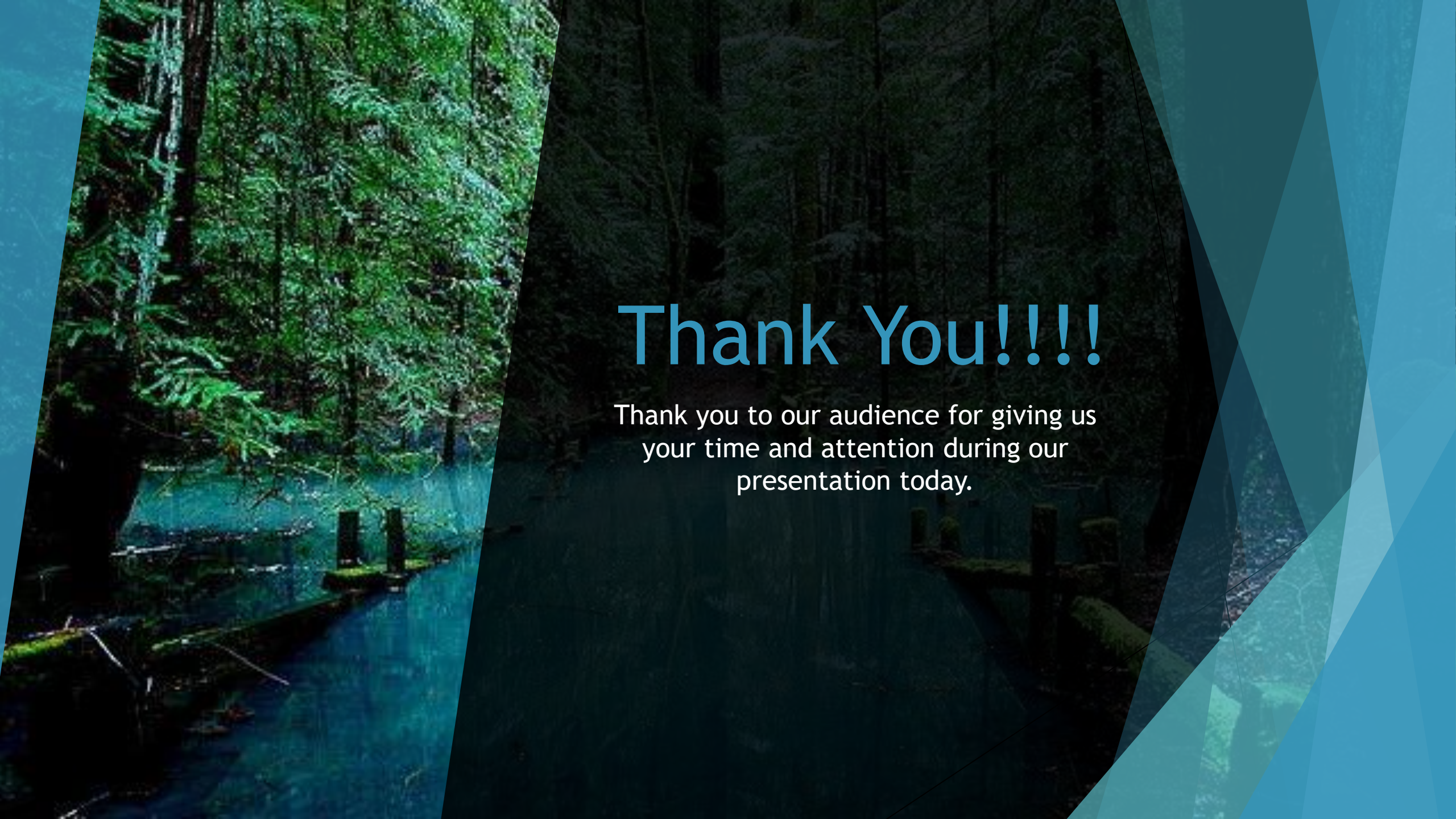
Predictive House Sales Calculator

Bedrooms	3
Bathrooms	2
Floors	1
Year Built	2005
Has Basement	No
Zipcode	98178
House Sqft	1200
Estimated Price	$231,927

Based off our model, we created a calculator that inserts the housing features to output the predictive housing prices.

Future Work

- Add to the dataset to explore the price difference between rural and urban homes
  - Possibly restrict the model further by choosing only urban or rural locations
- Create more features from web scraping to increase the accuracy of the model
- Distinguish a correlation between housing prices and school district placement



Thank You!!!!

Thank you to our audience for giving us your time and attention during our presentation today.