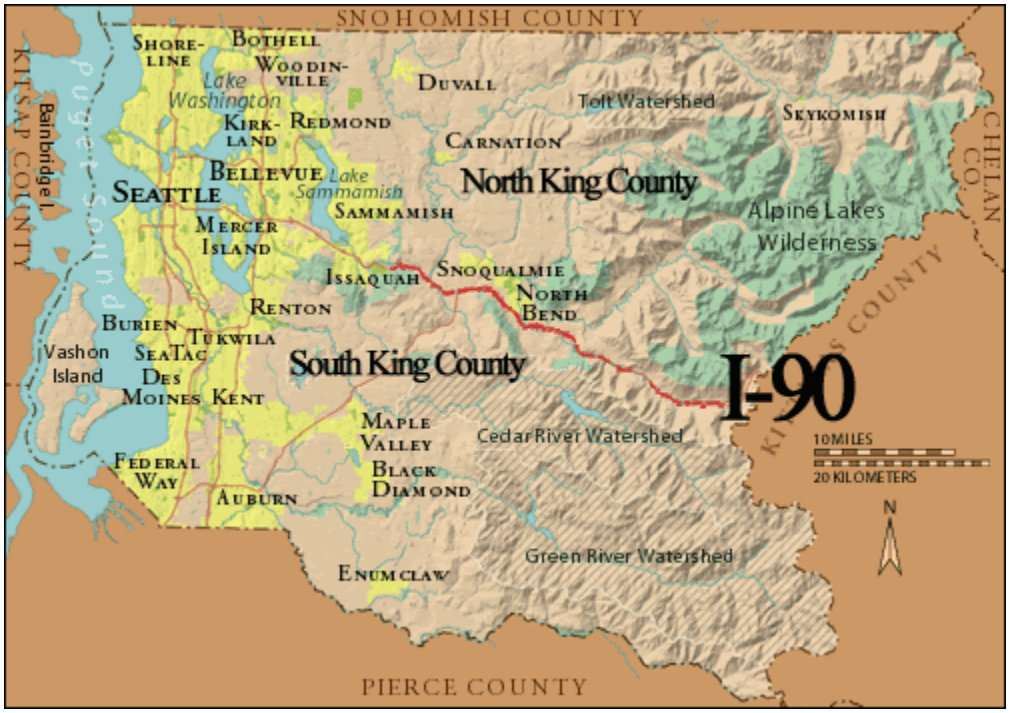
King County Home Prices

Regression Analysis



Main Objective of Analysis

Using various regression models, the goal of this analysis was to forecast prices of homes of King County in Washington state based on various attributes of the homes which included factors such as home size, location, and grading. The project focused on both interpretation and prediction. A detailed data analysis was performed, as well as constructing predictive models to estimate home prices, and after comparing different statistical error scores, determine which was the best model.

Description of Data Set

King County is the most populous center in the state of Washington with a population of 1,931,249 people and 789,232 households according to the 2010 US Census. The data set is a sample of 21,597 homes in King County with the following 21 attributes:

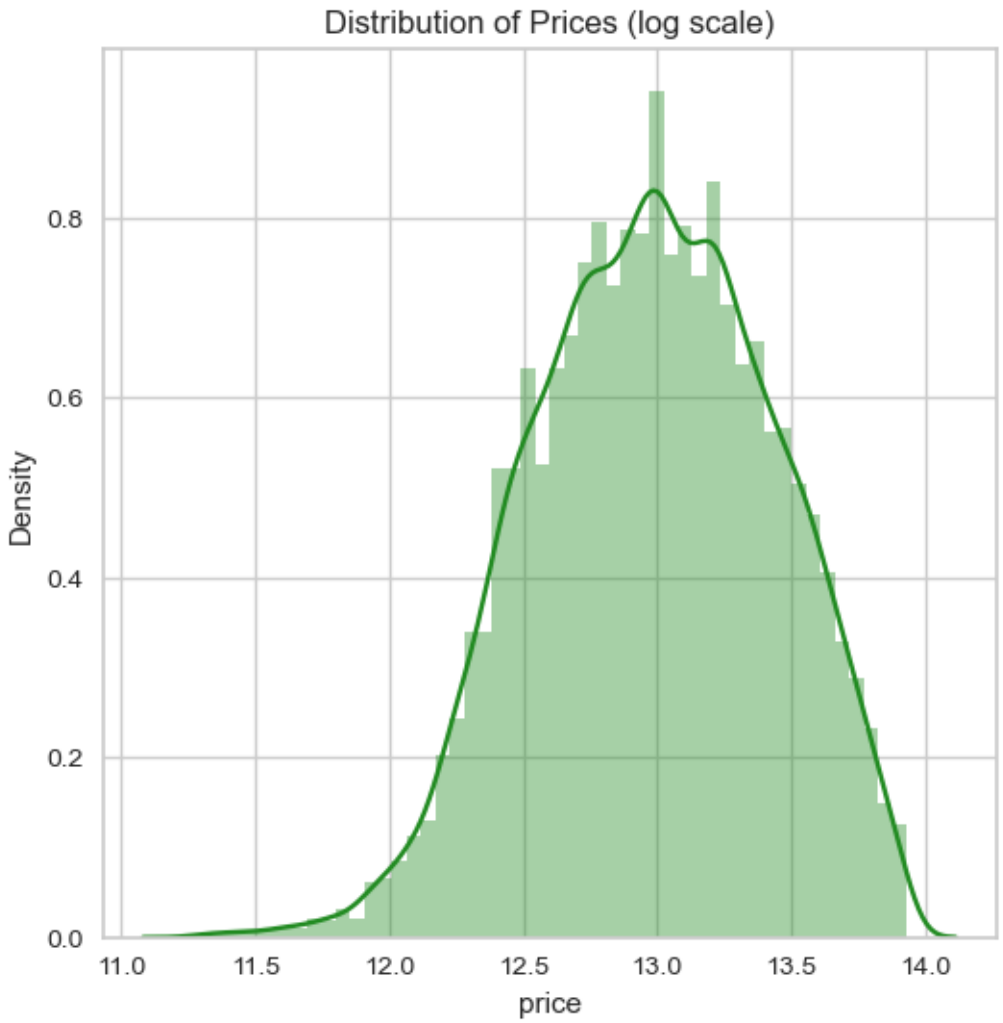
- Identification notation for a house (int64)
- Date house was sold (object)
- Price of house (float64)

- Number of bedrooms (int64)
- Number of bathrooms (float64)
- Square footage of house (int64)
- Square footage of lot (int64)
- Total floors of house (float64)
- If house has view to a waterfront (int64)
- If a house has been viewed (int64)
- Condition of house (int64)
- Grade of house (int64)
- Square footage of house apart from basement (int64)
- Square footage of the basement (int64)
- Year built(int64)
- Year renovated (int64)
- Zip code (int64)
- Latitude (float64)
- Longitude (float64)
- Living area in 2015 (int64)
- Lot area in 2015 (int64)

Exploratory Data Analysis, Data Cleaning, and Feature Engineering

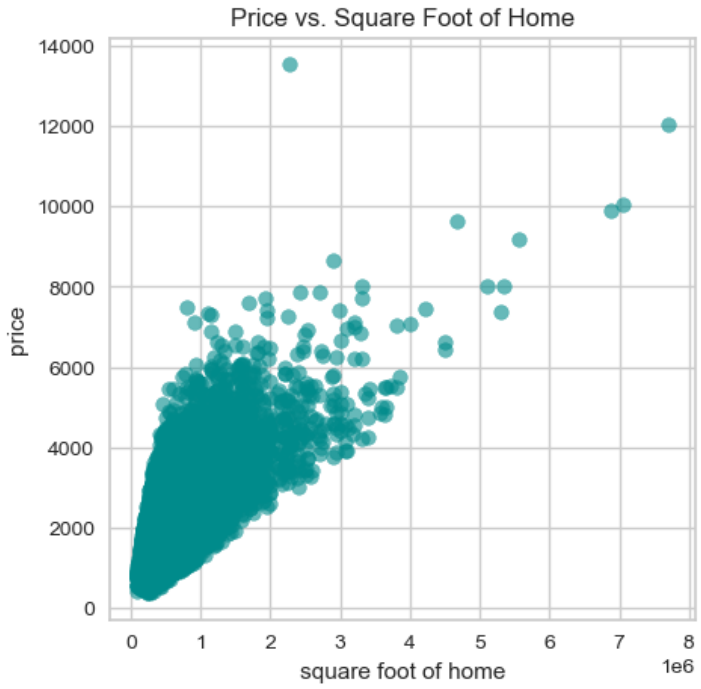
After performing summary statistics of the original data, the median price for homes was $450,000 while the maximum price was $7.7 million and the minimum $78,000. The mean price was $540,000. To examine the spread of the data for prices, a density curve was plotted which indicated that the distribution had low variance and was highly skewed to the right. The most expensive homes in King County are relatively few compared to homes that are more typically priced. To remove outliers, interquartile ranges were used and values outside of the range were removed.

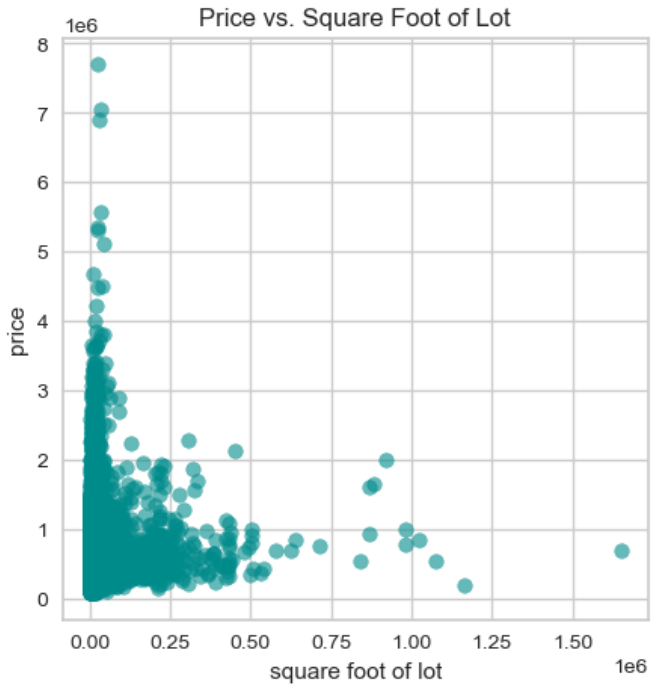
Plotting the density curve again, the spread of the data became more viable for analysis and prediction, but the right skew remained and to facilitate machine learning models, a logarithmic transformation was applied to the values and the density curve was plotted.

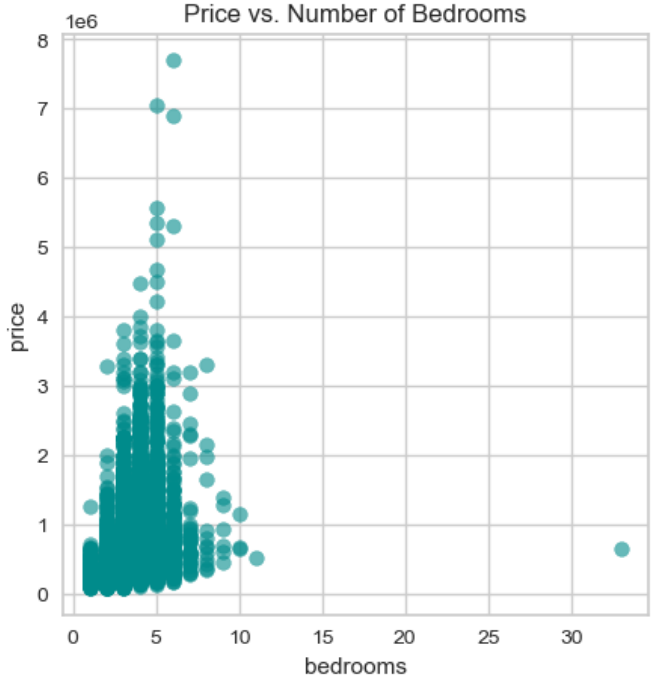


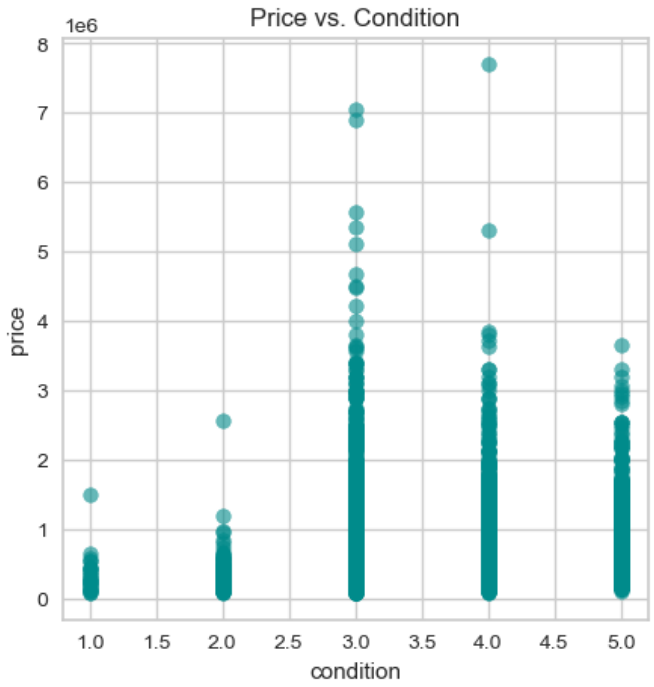
The dataset contained no null values. For exploratory data analysis, the following attributes were removed: 1. Identification Number, 2. Date of Purchase, 3. Number of views, 4. Year Renovated, 5. Latitude, 6. Longitude. With geographic information such as Latitude and Longitude, negative float values could complicate regression so rather than utilizing those factors, Zip Code was used instead, as location can play a significant role in home prices.

Scatter plots were made to observe the relationship between price and other attributes of homes. Upon brief inspection, the attribute that appears to be highly relevant in determining home prices is the square footage of the living area. The square footage of the lot seems more inconsequential. More expensive homes tend to have 4 to 5 bedrooms. Surprisingly, a higher condition score isn't indicative of higher prices and it seems like a condition of 3 tends towards higher prices.

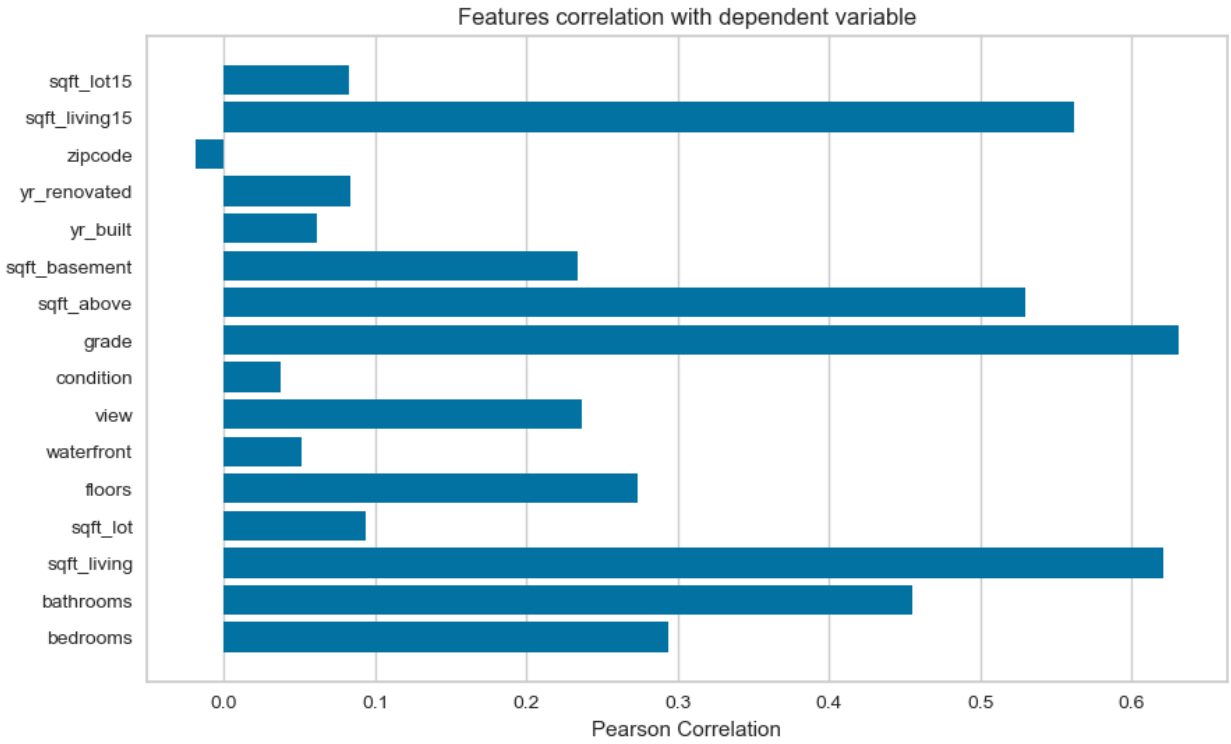



Price vs. Square Foot of Lot
1e6
8
7
6
5
4
3
2
1
0
price
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1e6
square foot of lot

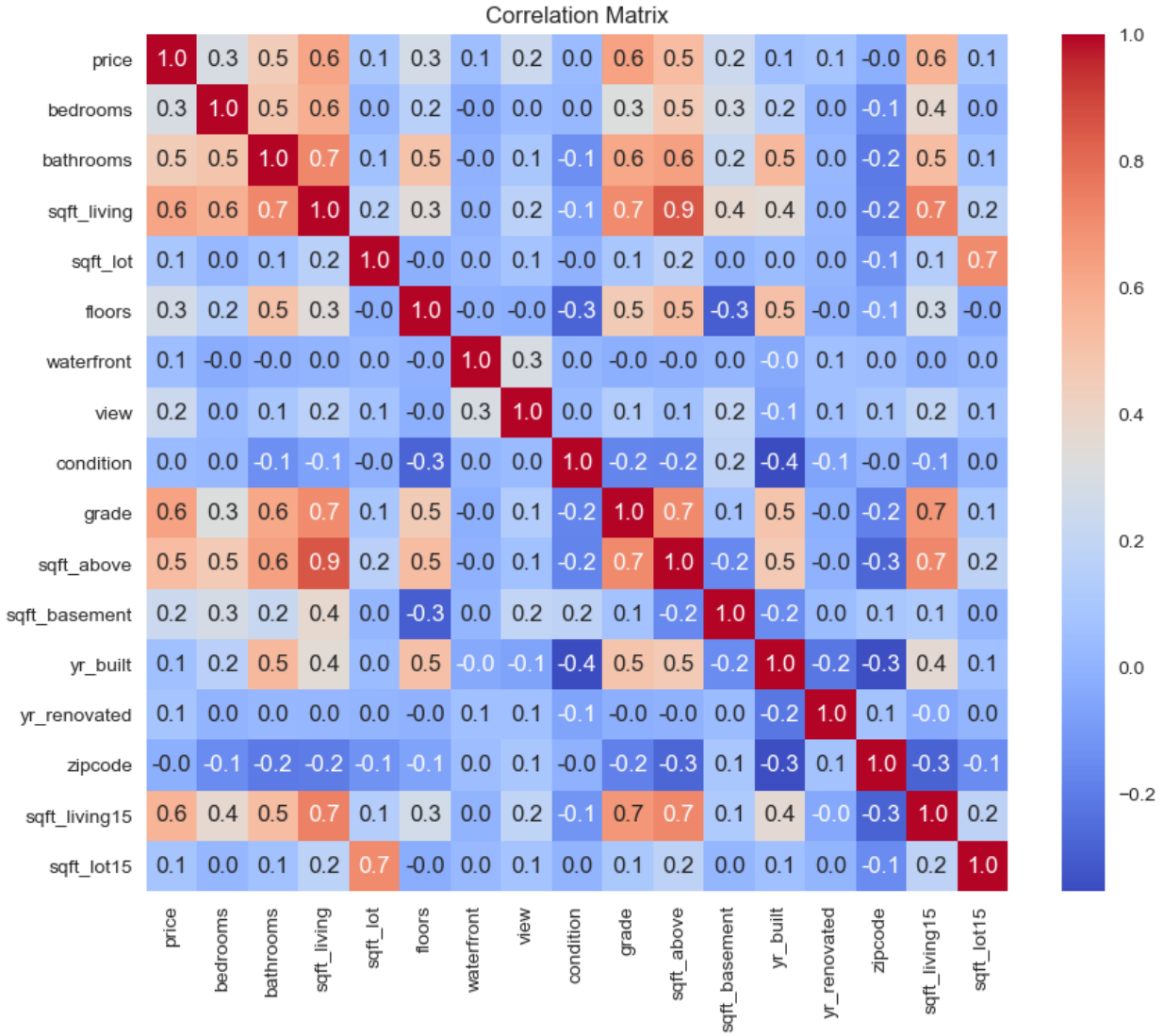

Price vs. Number of Bedrooms
1e6
8
7
6
5
4
3
2
1
0
price
0
5
10
15
20
25
30
bedrooms


Price vs. Condition
1e6
8
7
6
5
4
3
2
1
0
price
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
condition

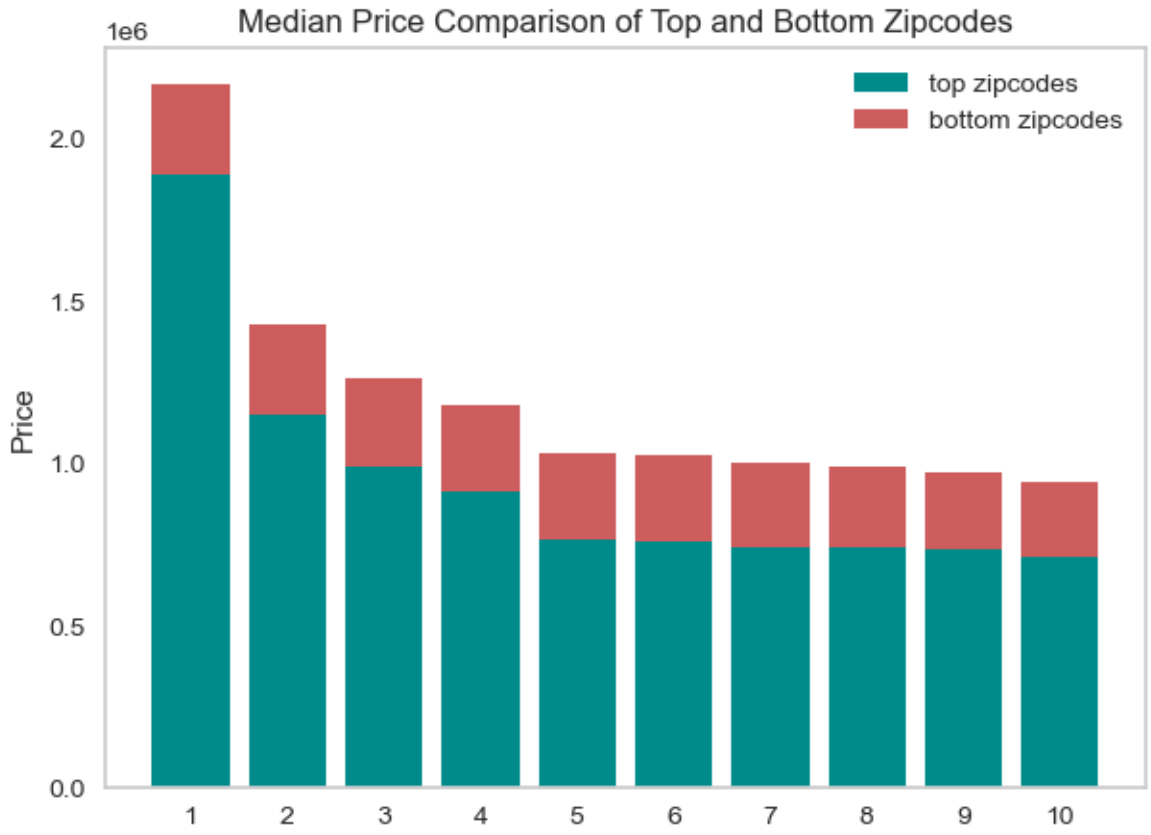
A more quantitative visualization comparing correlations with the target variable of price reveals more. Zip code as a numerical type rather than a categorical type seems like an irrelevant attribute, but later on, regression models performed better including this attribute.



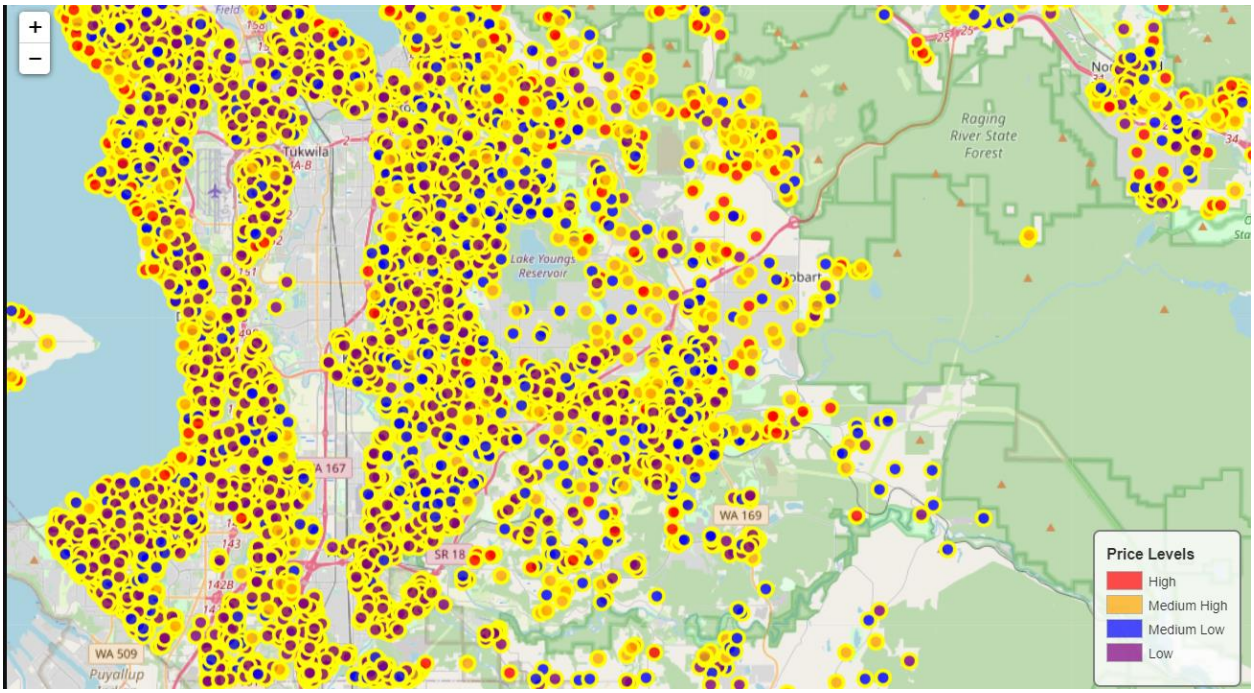
A correlation matrix was also created to visualize relationships between various attributes and price.



Homes in different zip codes were compared as well, the top 10 most and least median prices. The difference in location in relationship to price was significant and wide, indicating variability between home prices based on neighborhood.



Finally, a map was constructed to visualize geographically home prices with an interactive popup showing the price, price rank, and zip code of each home in the dataset. The data used to make the interactive map included the attributes of price, longitude, latitude, and zip code.

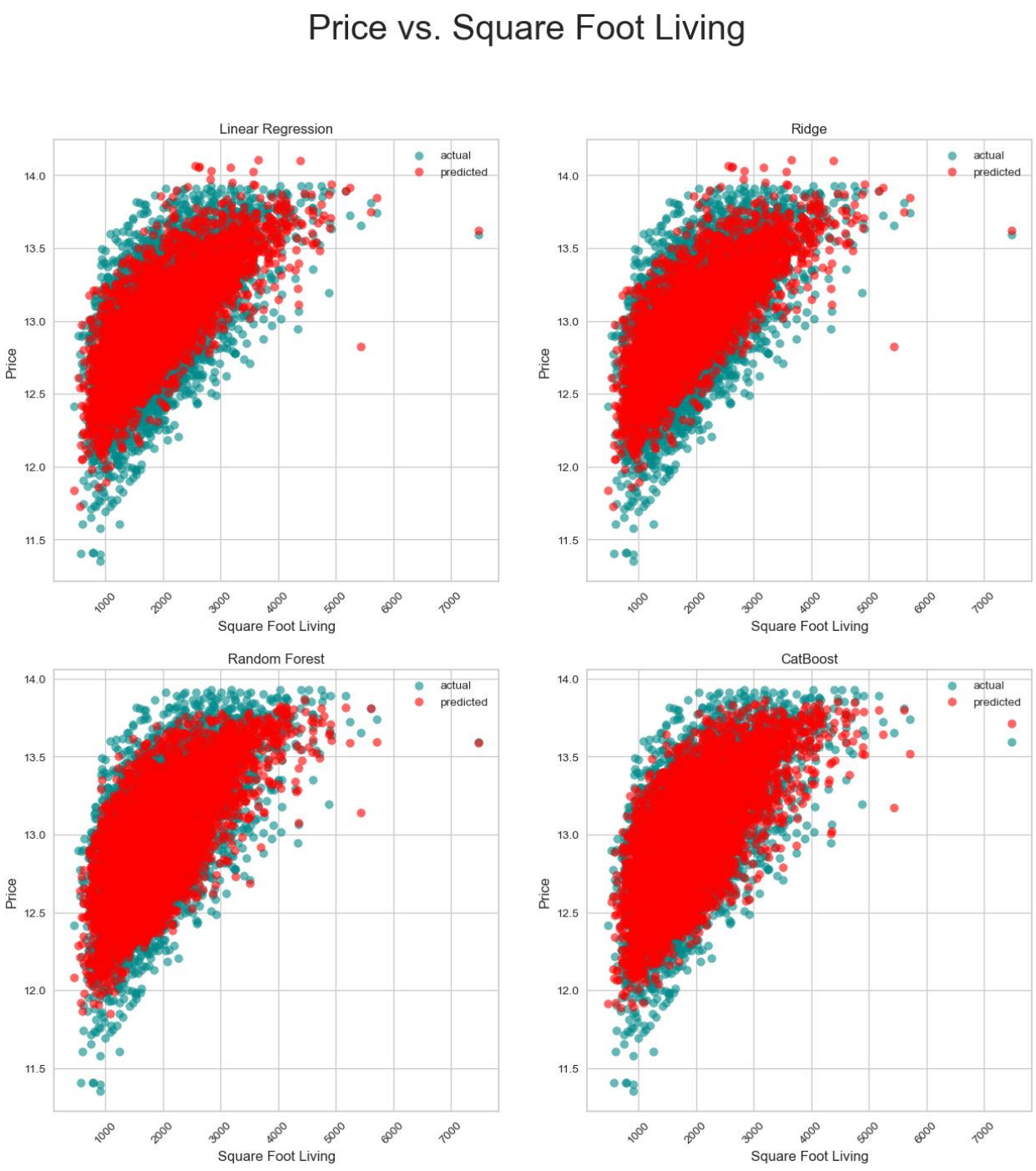


Regression Models

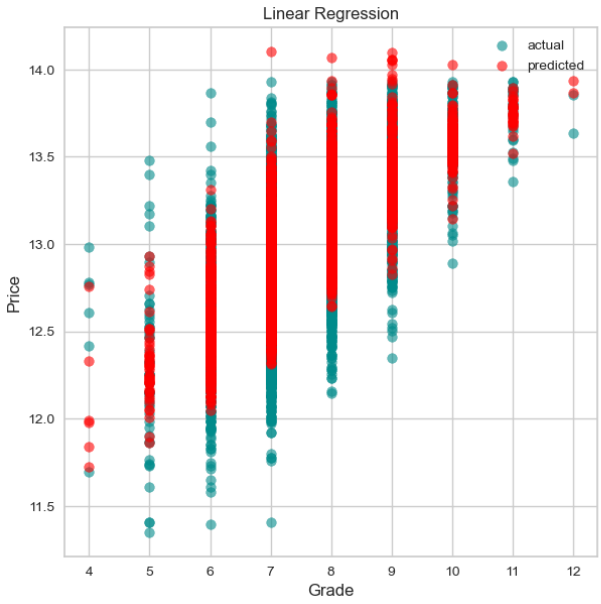
For the purposes of this analysis, four regression models were used, standard linear regression, ridge, random forest, and CatBoost. For linear regression and its variant ridge, `Polynomial Features` of degree 2 and `StandardScalar` were used to transform the data to train the models. Random forest and CatBoost are tree-based models and required no such transformations. `GridSearchCV` was used for ridge and CatBoost to determine the most optimal set of hyperparameters. `RandomizedSearchCV` was used for the random forest regressor to reduce training time, as the grid of hyperparameters was bigger and the algorithm takes longer to train. These were compared to determine the best R^2, mean absolute error, and root mean square error scores for a 80/20 train test split. CatBoost performed the best with lowest root mean squared error and mean absolute error scores as well as an R^2 value close to 1.0. Linear regression and ridge had identical scores, as `GridSearchCV` determined that the two shared same hyperparameters. Tree-based regression models like random forest and CatBoost provided the best metrics.

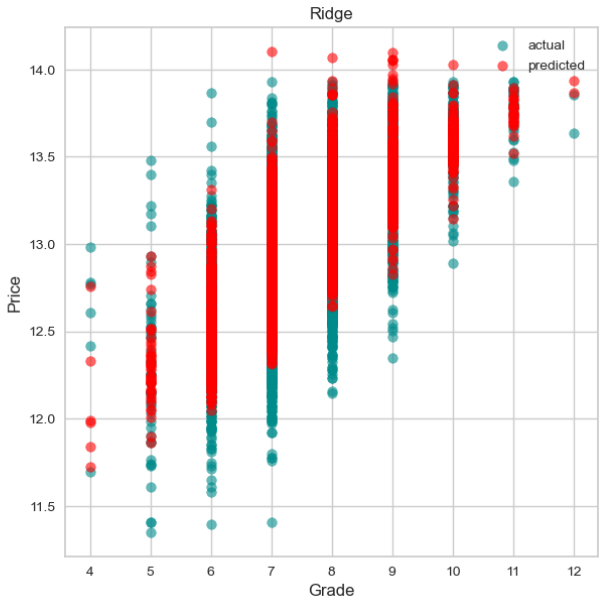
	model	r2	rmse	mae
0	Linear Regression	0.361785	0.279039	0.219690
1	Ridge	0.361784	0.279039	0.219690
2	Random Forest	0.738476	0.199002	0.141216
3	CatBoost	0.797312	0.176281	0.128588

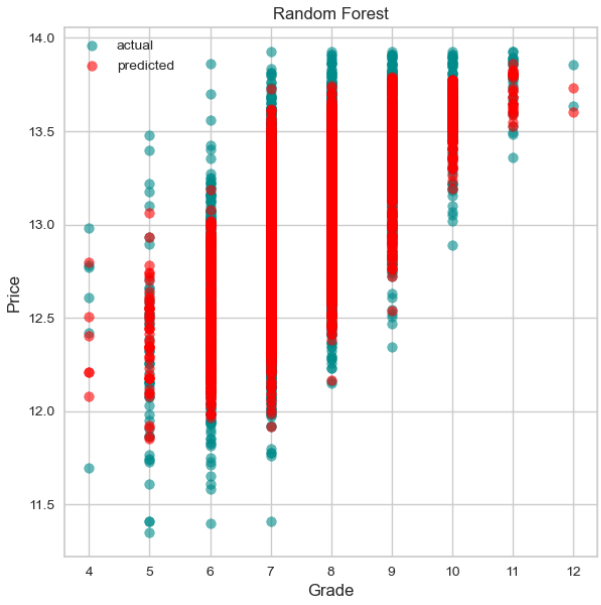
The following graphs visualize predictions and true data together using scatter plots for Price versus Square Foot Living, Grade, Square Foot Above, and Bathrooms. As indicated by the scores, CatBoost predictions most tightly align with true values for price.

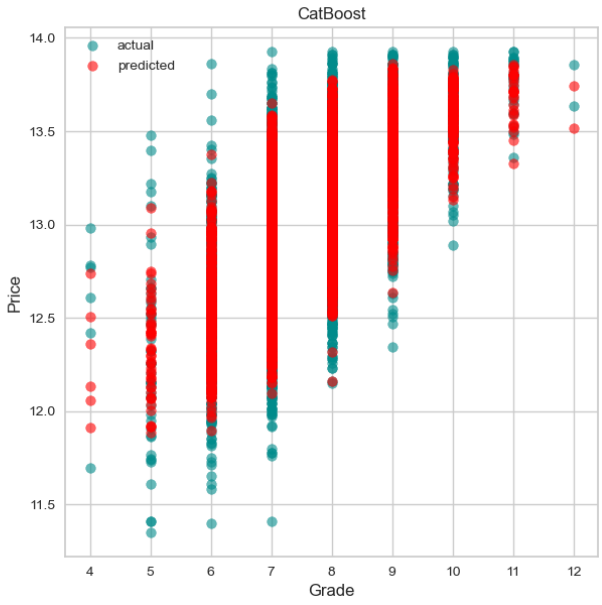


Price vs. Grade

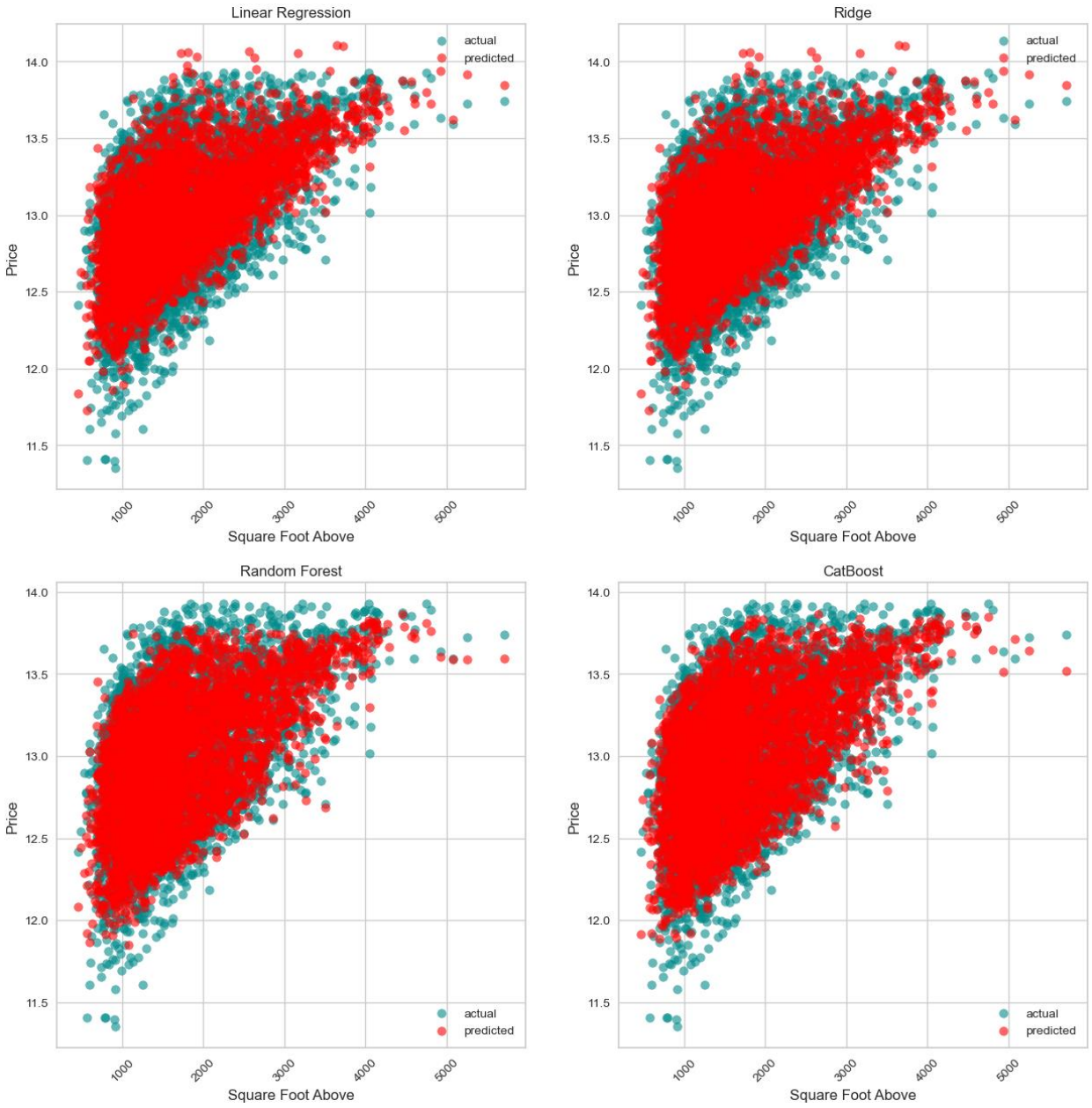




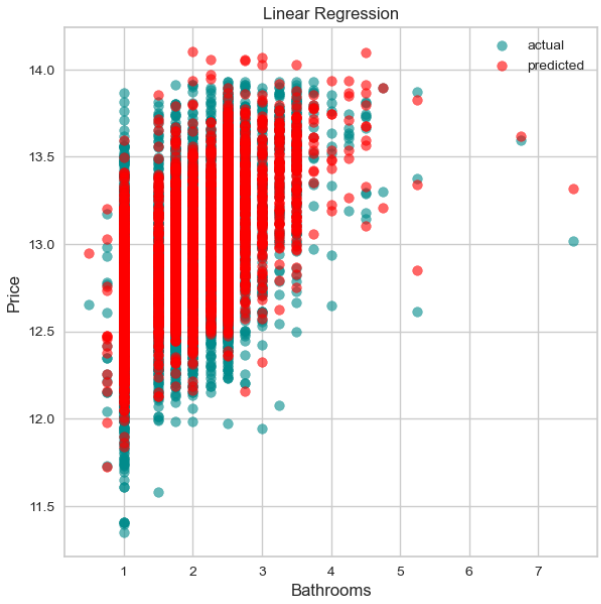


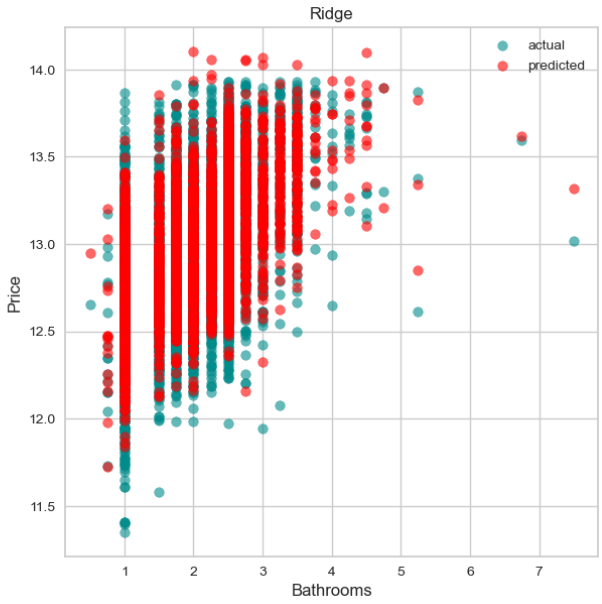


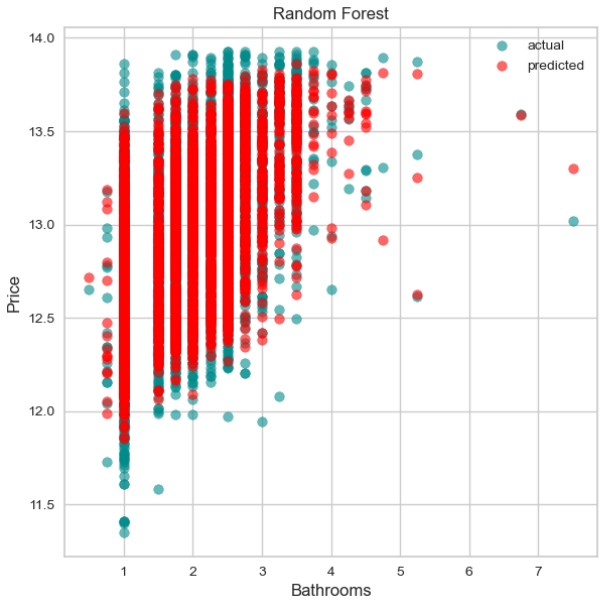
Price vs. Square Foot Above

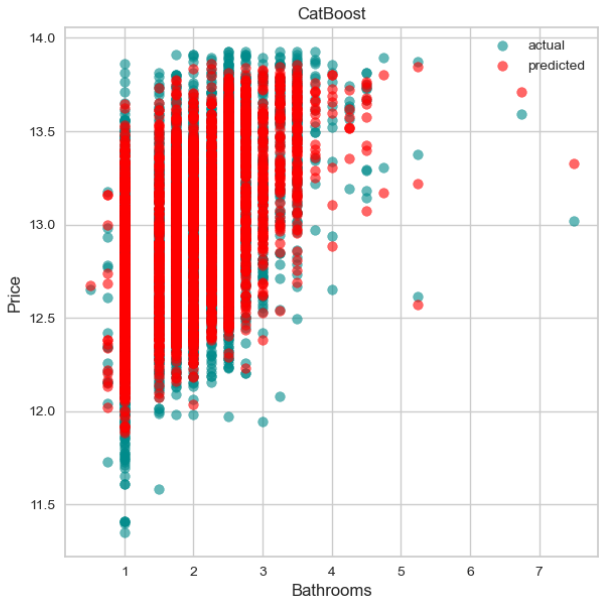


Price vs. Bathrooms









Key Findings and Insights

This present analysis was both interpretive and predictive. All of the values in the cleaned dataset were numerical, either integers or floats. Categorical attributes were completely absent which facilitated good regression modeling, although zip code seemed more like a categorical attribute. Scoring varied between the linear regression-based models and the tree-based models. Between random forest and CatBoost, on top of the better score, a boosting model like CatBoost trains faster, facilitating experimenting with a larger grid to set the optimal hyperparameters. Given that and the superior prediction metrics, the CatBoost model is clearly the best choice to utilize in this project.

Next Steps

Using regression to predict house prices in this dataset was fairly straightforward. A potential future analysis would involve classification algorithms to determine the grade of a house based on various attributes in the dataset, as the grades are integers in a small fixed range and can be converted into categories. Also, a larger and more recent data set could be used for analysis and prediction. What was lacking in this current project in feature engineering was feature selection, eliminating attributes that are either useless or even hamper predictive power. A method such as SHAP can be applied to achieve feature selection. And conversely, data with other relevant attributes not included in this dataset could be utilized to enhance analysis and further improve prediction models. Experimenting with other models such as XGBoost can be done as well.