Statistical Modeling of House Prices

Tainshu Fan, Jaehee Jeong, Lisa Kaunitz
STAT 412 | Fall 21| Professor Wu

Overview

Executive Summary

The purpose of this study was to find the most significant variables when it comes to pricing a home, and then fit the best model to predict housing prices for the real estate market in King County, WA. We can summarize this with our two Research Questions:

1. What variables are significant in predicting the price of a house?
2. Which model best predicts the price of a home?

We started by exploring the dataset and diagnosing any feature transformations needed. After making the appropriate transformations to our data, we created multiple models and checked the validity and metrics of each one to see which best predicted housing prices. Our final model indicated that the top 5 most important variables in predicting home prices in the King County area are the latitudinal coordinates (lat), the square footage of living space (sqft_living), the quality of construction and design (grade), the square footage of the homes in the surrounding neighborhood (sqft_living15), and the square footage of the home above ground level (sqft_above).

Dataset Background

This dataset contains house sale prices for King County, WA. It includes homes sold between May 2014 and May 2015. 21,613 observations and 21 variables. The variables describe housing features, rather than features about the population. The source of our data was from Kaggle, which delivered a fairly clean dataset with no missing values and was able to be retrieved via a simple download and imported as a csv file. We were interested in this dataset in particular because we thought housing data had a lot of potential and interesting use cases for regression. Typically, there are many different features that go into one home, and it can get very overwhelming for both the buyers to see which features are most important to them, and for the sellers to see which features are the most valuable when selling a home. This is why we decided to explore our two research questions that also allowed us to deploy many advanced regression modeling techniques.

Data Dictionary

Column	Data Type	Description
id	num	Unique ID for each home sold
date	date	Date of the house sale
price	num	Price of each home sold
bedrooms	int	Number of bedrooms
bathrooms	num	Number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living	int	Square footage of the apartments interior living space
sqft_lot	int	Square footage of the land space
floors	num	Number of floors
waterfront	int	A dummy variable for whether the apartment was overlooking the waterfront or not
view	int	An index from 0 to 4 of how good the view of the property was
condition	int	An index from 1 to 5 on the condition of the apartment
grade	int	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design
sqft_above	int	The square footage of the interior housing space that is above ground level
sqft_basement	int	The square footage of the interior housing space that is below ground level
yr_built	int	The year the house was initially built
yr_renovated	int	The year of the house's last renovation
zipcode	int	What zipcode area the house is in
lat	num	Latitude
long	num	Longitude
sqft_living15	int	The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15	int	The square footage of the land lots of the nearest 15 neighbors

Descriptive Analytics - EDA

Exploratory Data Analysis

In the dataset, there are no missing values. We remove some high-end houses from our dataset, detailed explanation following below the table. We create the sqft_basement_yesno column based on the sqft_basement column showing if a house has a basement or not regardless of size.

Table 1

	Minimum	25th Percentile	Median	Mean	75th Percentile	Maximum	Standard Deviation
date	2014-05-02	2014-07-22	2014-10-16	2014-10-29	2015-02-17	2015-05-24	
price	75000	314950	435000	468939	595000	1020000	197841.89
bedrooms	0	3	3	3.321	4	33	0.91
bathrooms	0	1.5	2	2.04	2.5	7.5	0.7
sqft_living	290	1390	1840	1957	2410	7480	757.06
sqft_lot	520	5000	7500	14526	10267	1651359	40000.27
floors	1	1	1	1.472	2	3.5	0.54
waterfront	0	0	0	0.002922	0	1	0.05
view	0	0	0	0.1633	0	4	0.62
condition	1	3	3	3.405	4	5	0.65
grade	1	7	7	7.507	8	12	1.02
sqft_above	290	1170	1510	1694	2080	5710	712.76
sqft_basement	0	0	0	263	500	2720	404.11
sqft_basement_yesno	0	0	0	0.3767	1	1	0.48
yr_built	1900	1951	1974	1971	1996	2015	29.13
yr_renovated	0	0	0	73.21	0	2015	375.12
zipcode	98001	98033	98065	98079	98118	98199	53.31
lat	47.16	47.46	47.57	47.56	47.68	47.78	0.14
long	-122.5	-122.3	-122.2	-122.2	-122.1	-121.3	0.14
sqft_living15	399	1461	1790	1909	2260	4950	601.61
sqft_lot15	651	5040	7528	12386	9840	871200	26447.45

Table 1: Variable Summaries

Dealing with Response Variable Price

From our distribution graph of price, we see that our response variable price is not normally distributed and is highly skewed. By taking a closer look, we can see that most of the price ranges below 1 million. While this is something that we expected when looking at housing data, we were not able to go forward with modeling until we tried different options on how to transform and deal with the response variable. It is also important to note that our raw pricing data is unimodal, therefore, any discussions about cutting the data adding a potential split in populations would not be an obstacle, this would be something to look into if our dataset included data and metrics about the population, and if we saw a raw distribution of our response that was bimodal for example.

Figure 1

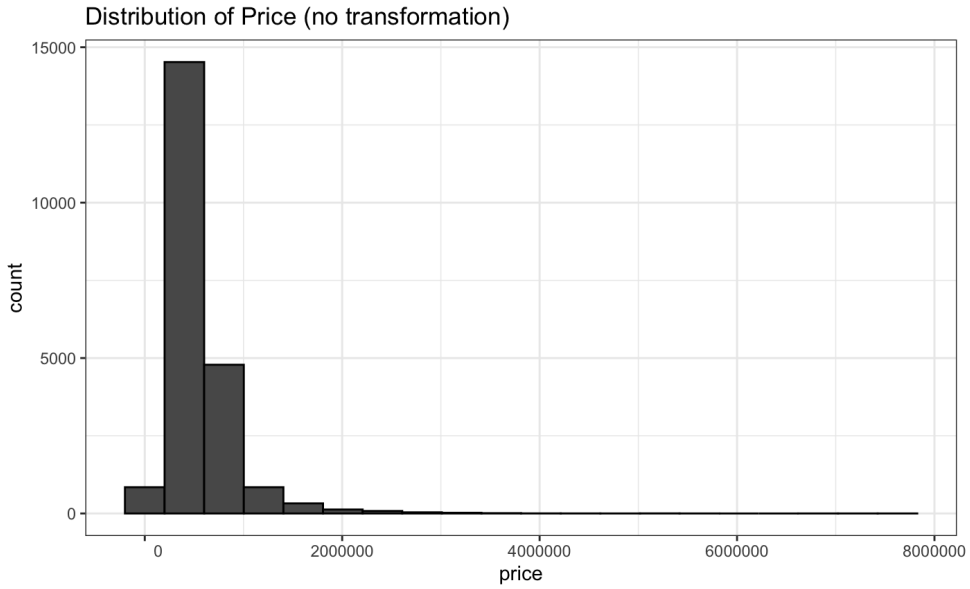


Figure 1: Distribution of Response Variable without Transformation

Originally, we ran a box-cox transformation and thought about log transforming price. However, when we perform log transform, it decreases the explainability of our model. Since the goal of our project is to build models with interpretability in mind instead of black box models, we decided against this approach. Instead, we chose to treat the extreme high-end houses as outliers and remove them from the dataset. We also note that by doing that, our end models will only have a more limited application range where we would only be able to predict houses below 1 million dollars accurately.

Figure 2

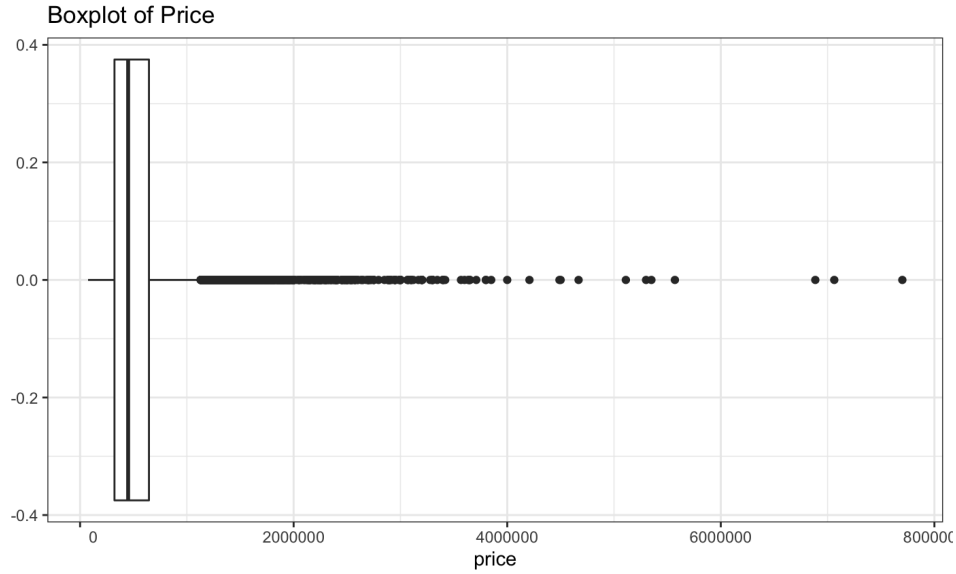


Figure 2: Boxplot of Untransformed Response Variable

To remove these high-end houses from our dataset, we decided to use a method of 1.5*IQR, with IQR being the interquartile range of our original price data (3rd quartile - 1st quartile). After removing data that falls outside of this range, we observe that our final dataframe has 20194 rows instead of the original 21613 row, which is a less than 7% data loss. This is considered acceptable by us and we proceed with our modeling.

Figure 3

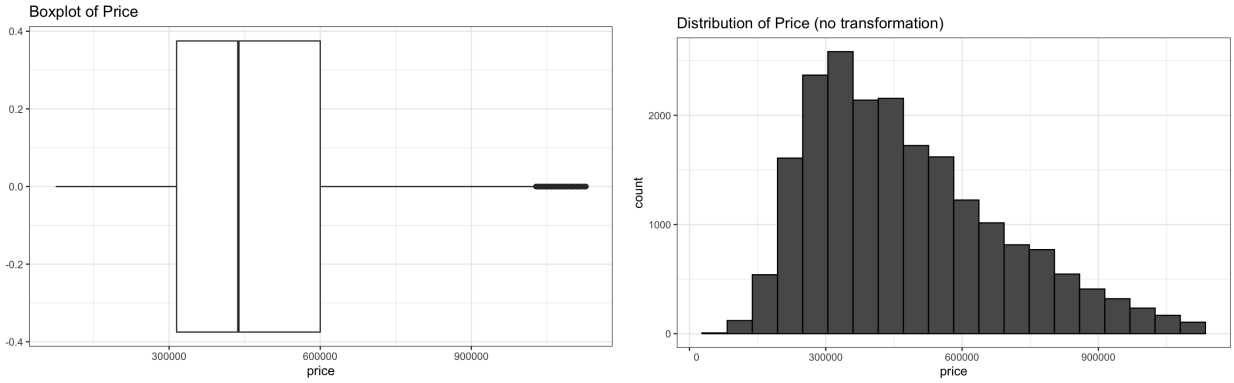


Figure 3: Descriptive Plots of Transformed Response Variable

Before we do feature selection, we dropped some features first. From the graph below, we can see that there is no trend between the price variable and date. Therefore, we will not use this variable to create a model.

Figure 4

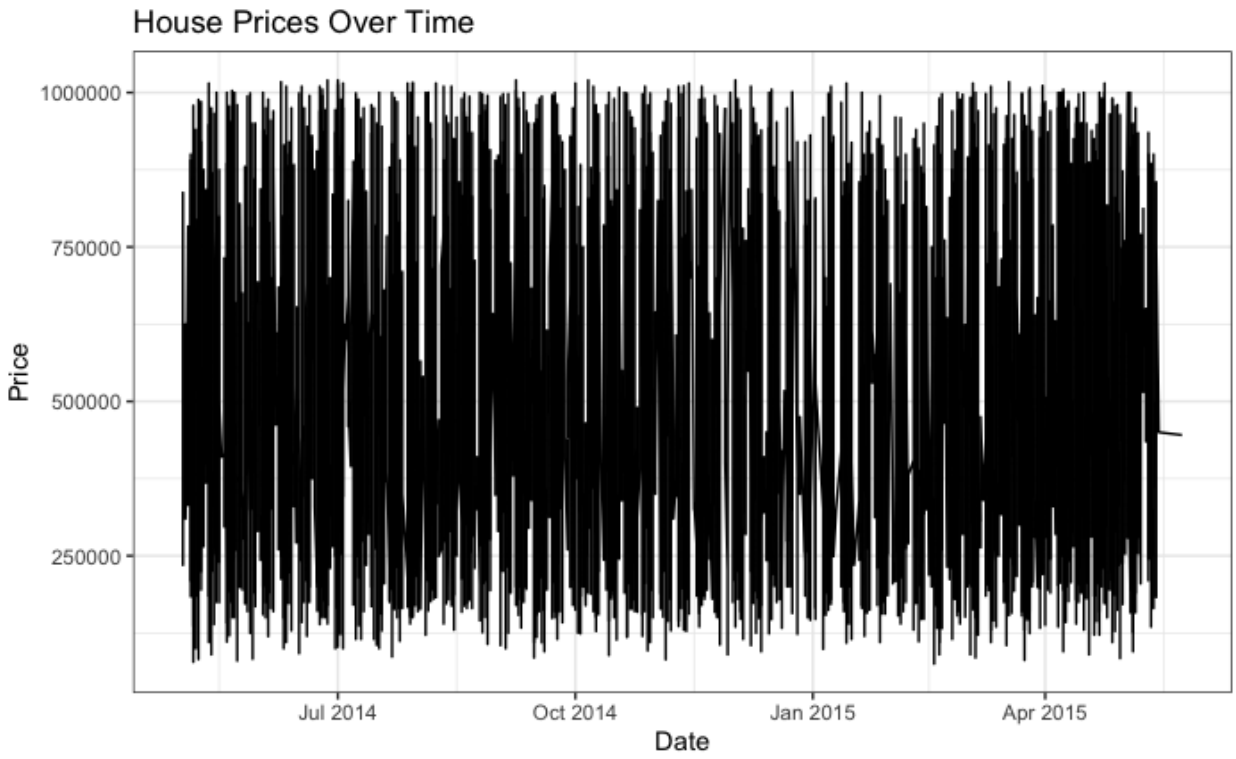


Figure 4: Plot of Housing Prices over the Full Timeline of our Data

We immediately dropped the ID variable because it does not have any meaning within our context, and we dropped zip code as well because this can be represented by the latitude and longitude features in our data. Also, we dropped sqft_basement because we created a variable which is sqft_basement_yesno since more than half of the houses do not have a basement. We decided to use an indicator factor instead of using the actual size of the basement, as that would have been more representative of our data, as well as being a better feature for modeling.

Feature Selection

Since our original dataset contained a lot of different variables, we wanted to look at only the most important variables to include in our models. We thought of 3 ways to do feature importance analysis, first looking at Pearson's correlation with prices, then running a random forest regression model, and finally using penalty drive models such as the ridge and lasso regression models.

Figure 5

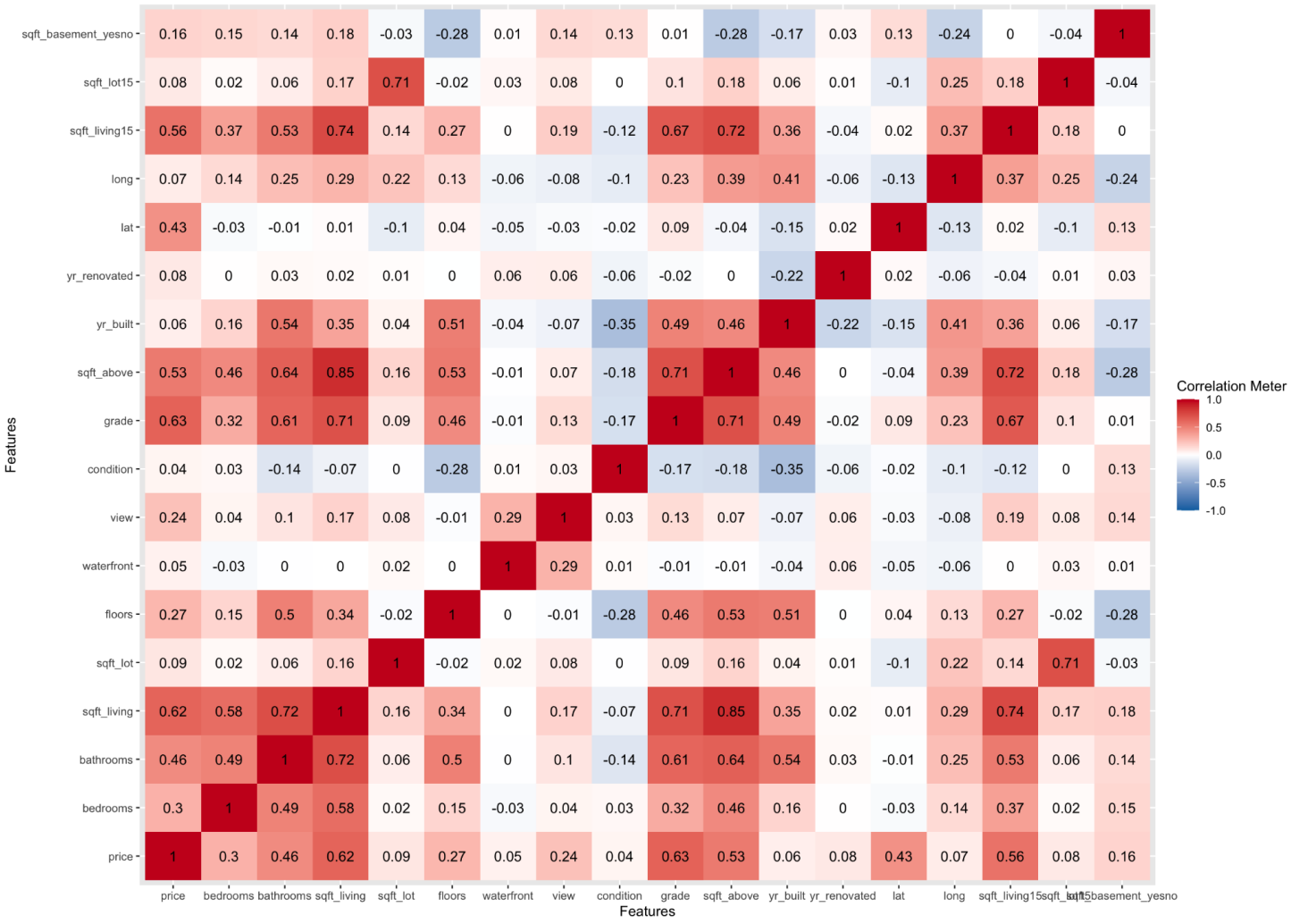
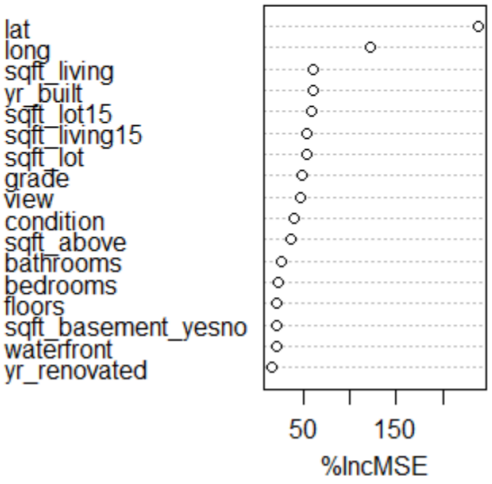


Figure 5: Output of Pearson's Correlation Matrix for Each Variable in the Data

The first way was to directly look at Pearson's correlation between our various variables and price. The correlation matrix is a great way to visualize high correlations with the response variable, as well as getting a first look into any multicollinearity between the variables. The figure above is representing highly correlated variables with a dark red hue, and the lower correlated variables with a blue hue, those with little to no correlation are going to be represented with light or white colors. We are specifically looking for variables with high correlations in the first column of the matrix (price), and are similarly looking at independent variables that may be highly correlated as well as an indicator of collinearity. From the correlation graph above, we see that grade, sqft_living, sqft_living15, sqft_above, bathrooms, lat, bedrooms, floors, view, sqft_basement_yesno have the highest correlation with price.

Figure 6

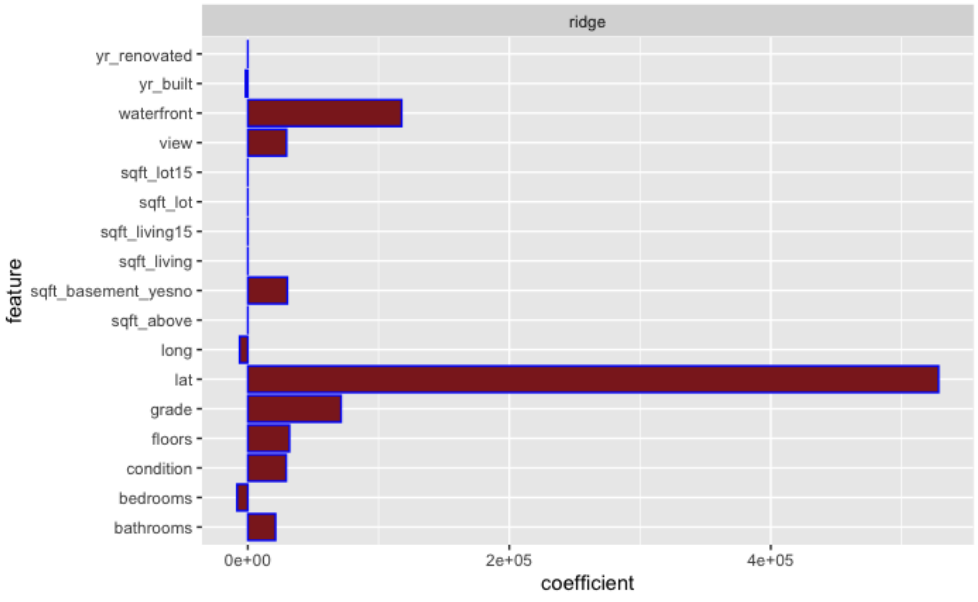


	%IncMSE	IncNodePurity
bedrooms	368468179	7.338490e+12
bathrooms	1048709070	1.831961e+13
sqft_living	8786870563	1.260677e+14
sqft_lot	2189603954	2.333135e+13
floors	639748085	6.198054e+12
waterfront	54723966	1.566108e+12
view	529390170	9.344959e+12
condition	480587411	6.776946e+12
grade	7595461717	1.092432e+14
sqft_above	4101935841	5.276618e+13
yr_built	3517422801	3.235424e+13
yr_renovated	50080695	2.441799e+12
lat	27582305547	2.536420e+14
long	4177259121	3.505328e+13
sqft_living15	3889542418	6.332456e+13
sqft_lot15	2030748467	2.502583e+13
sqft_basement_yesno	480926482	5.004984e+12

Figure 6: Random Forest Regression Feature Importances

Another method we considered is to use a random forest model in the beginning and utilize the feature importance function built in. After building the model, we decided to look at the %IncMSE as we believe that is the more robust and informative measure. %IncMSE looks at the increase in MSE of predictions (estimated with out-of-bag-CV) as a result of variable X being permuted. In the end, our top 10 features are the following: lat, sqft_living, grade, sqft_living15, sqft_above, long, yr_built, sqft_lot15, sqft_lot, and bathrooms.

Figure 7



The third method that we used for the feature selection is ridge and lasso models. Even though the lambda values were different for each model, the coefficients were very similar. For this reason, we only attached the coefficient plot of the ridge model. When we look at the plot above, only half of the features have a meaningful coefficient which are lat, waterfront, grade, floors, condition, view, sqft_basement_yesno, bathrooms, long, bedrooms. For yr_built, long, and bedrooms features have a negative coefficient. For yr_built and long features, this could be understandable, but for the bedrooms feature, it could be questionable. Having a negative coefficient indicates that if a house has more bedrooms, the price will decrease which does not follow intuition or typical market standards.

Table 2

Correlation	Random forest	Ridge
grade	lat	lat
sqft_living	sqft_living	waterfront
sqft_living15	grade	grade
sqft_above	sqft_living15	floors
bathrooms	sqft_above	condition
lat	long	view
bedrooms	yr_built	sqft_basement_yesno
floors	sqft_lot15	bathrooms
view	sqft_lot	long
sqft_basement_yesno	bathrooms	bedrooms

Table 2: Significant Variables from each Importance Method

The table above has the top 10 features from different analyses. The red font means the feature is important in all three methods, and the orange font means the feature is important in two of the methods. Overall, it seems that the location of the house is important. Also, how well a house is designed and the number of bathrooms seem to affect the house prices a lot. The grade of the house, likely correlated with year built is also important. Besides that, the amount of sqft a house has along with the neighboring houses' sqft are also important. We decided to use all the features from the result of the random forest model because based on common knowledge of housing price factors, it seems that random forest aligns with expectations the most, and is the best performing model to answer our research question.

Predictive Analytics - Modeling

Simple Linear Regression

In order to find the best model to predict housing prices, we implemented and compared six total regression modeling techniques. We started by using the simplest and most interpretable model as a baseline, this was the simple linear regression model. This model included the ten variables found from the random forest feature importance algorithm and our IQR threshold price variable as the response. The diagnostic plots showed that our model was valid and met all assumptions. The summary of our model showed that all but one variable from our feature importance were statistically significant at an alpha level of 0.001, leaving out sqft_lot15 (the square footage of the land lots of the closest 15 homes). The model also yielded an adjusted R-squared of 0.6738, meaning we are able to explain 67% of the variability of the data with our model. The final results from the linear regression model were that of the interpretability of variables in relation to the housing price. Most significantly, we found that in regards to location, home further west will have a higher value. With all else held constant, a one unit increase in the quality or "grade" of the home results in a price increase of $77,660, which goes to show the scale at which the variable has been objectively set. The model showed that for each increase of 100 sqft average of the homes in your neighborhood, we expect a price increase of about $5,000. Results like these are what make our simple linear regression model our second best in terms of performance, and our best in terms of interpretability.

Figure 8

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.548e+07  1.101e+06 -23.150  < 2e-16 ***
bathrooms      3.302e+04  2.264e+03  14.587  < 2e-16 ***
sqft_living    6.087e+01  3.055e+00  19.922  < 2e-16 ***
sqft_lot       1.946e-01  3.747e-02   5.194 2.09e-07 ***
grade          7.766e+04  1.584e+03  49.025  < 2e-16 ***
sqft_above     8.392e+00  2.933e+00   2.862  0.00422 **
yr_built      -2.002e+03  4.729e+01 -42.339  < 2e-16 ***
lat            5.132e+05  7.343e+03  69.889  < 2e-16 ***
long          -3.771e+04  8.383e+03  -4.499 6.89e-06 ***
sqft_living15  5.353e+01  2.645e+00  20.241  < 2e-16 ***
sqft_lot15    -9.237e-02  5.567e-02  -1.659  0.09712 .
```

Figure 8: Output from Linear Regression Model

Random Forest Regression

The model that best predicted home prices for our data was the random forest regression model. To understand why the random forest algorithm is so effective, one must first know what a decision tree is. A decision tree is rather intuitive, it is essentially a bunch of if else statements where if a condition is satisfied, we get certain results. For our housing dataset, it could be something like if a house has more than 3000 sqft, one specific decision will say that house is

worth 500,000. Now, if we think of a forest, it is essentially just a bunch of trees. A random forest is no different, it is an ensemble of simple decision trees. Each individual decision tree gets a vote and for our regression purposes, we then average across all the predicted y values from each of the decision trees. Some pros of random forest include its high accuracy and its ability to deal with non-linear relationships or collinearities. In our case, the pricing data has a bunch of variables relating to house square feet which will likely cause collinearity issues. By using a random forest, we won't have to worry too much about that. However, random forests do have the possibility of overfitting which is why we split our data into 70% training and 30% testing. By doing so, we would test the accuracy of our model from data that we have not seen before, thereby solving the potential overfitting, though based on the size of our dataset, overfitting will not be likely. The other con of the random forest model is its lack of interpretability. This is why we included a simple linear regression.

Honorable Mention Regression Models

The other models that received an honorable mention and are very close to the performance of the linear regression were the stepwise regression, lasso, ridge, and principal component regression. The stepwise regression model yielded similar results to that of the linear regression which was intuitive because it is essentially many iterations of linear regression models. We did not specify forward or backward, instead, we used the step function with "both" as the direction. This model uses the AIC statistic for selecting the models. Our final result for this model was that of a few tenths worse performance than the linear regression, and it showed the same variables as statistically significant. The next model we looked at was PCR and was hypothesised to do really well. However, we were surprised to see our result because it is typical for principal component analysis and regression to be directly applicable to variables representing real estate to minimize the dimensionality of the data, however, not only did this not improve our model, PCR was our worst performing model. Ridge and lasso models are not any better than the linear models. This result was expected since we dropped some features that were important in the original dataset from the ridge model while we were doing feature selection.

Model Summary

We used the root mean squared error (RMSE) as our performance metric for each of our models because residuals are a good measure of how far the fitted data is from the actual. An advantage of using RMSE over looking at the R-squared value is because it keeps the integrity of the complexity of our model. When we add more variables to our model the R-squared will increase and say that we are able to explain more variance, but we would be adding more complexity and not be taking into account the tradeoff between bias and variance. RMSE is also more interpretable for our regression purposes, looking at our results table, we can see that our random forest model will on average give a prediction that is $84,023 off of the actual price. Figure x shows a table comparing our performance metric for each of our models.

Table 3

Model	RMSE
Simple Linear Regression	**120,550.69**
Step Regression	120,550.70
PCR	234,244.40
Lasso	120,550.78
Ridge	120,550.89
Random Forest	**84,023.75**

Table 3: Model Result Outputs

Conclusion

Results and Findings

Coincidentally, our two best models consisted of both the most complex, and the simplest of the six. The random forest regression model had the best performance by far with the lowest RMSE. This model did the best in answering our second research question for predicting housing prices. The next best model, linear regression, performed slightly better than step, lasso, and ridge; nonetheless, it is great for interpreting the coefficients and telling a clear story for how our inputs affect the response variable, which is why it is the best for answering our first research question. In regards to our research questions, we are able to reference the linear regression model to answer the first research question, and we are able to reference the random forest regression model to answer our second research question.

Shortcomings

Although we were successful in answering our research questions and exploring the dataset, there were three obstacles within the data. First, it is difficult to interpret our best performing model, the random forest. While this is the most robust model, and can be conceptually explained and understood, it is still a “black-box” when it comes to inputting data and hyperparameters into a function. One way that we were able to get around this was by exploring the results of the simple linear regression model for interpretability. Next, we believe our models do not perform overall, regardless of their performance in relation to one another. Random forest yielded the lowest RMSE of 84,023 which is indicative that the model itself does not do a great

job at predicting housing prices, it only is the best out of the others. We typically want to see an RMSE close to zero if we want to present a strong model, however, because of the intricacies of the data and the distribution of the response variable, we were left with a high RMSE. Our final limitation is that we had a limited range of house prices. Having more data in general will yield better results, and this is something that may help for future use.

Future Recommendations

There are two ways to improve this project for future study and they all fall under the same umbrella of data acquisition. Like we mentioned in the beginning, the data only consisted of variables specific to the home and did not include any features regarding the population. Adding features about the population such as Income, Education, Socio-economic status, Family Size, and more, would allow us to extend our scope to look at different effects these variables have on the home prices, look at different subsets of the data, and allow us to give insights on how to best market home to certain areas of a population to help sell the homes. Lastly, this study only pertains to the time frame of 2014 to 2015, and is outdated in regards to the housing market today. Having more than this one year time span would allow us to no only get more data to work with and improve modeling, but it would also allow us to create more meaningful insights that are consistent with the current market.