Predicting House Prices in King County Washington

Kierra Dangerfield

Table of Contents

Introduction

Buying and selling a home is an exciting and stressful time in someone's life. It's the beginning of a new chapter. There are myriad ways to determine the value of a home whether it is the location of the home, the square footage of a house, or the condition of the house. Our goal is to predict the price of a house based on houses sold between May 2014 to May 2015 in King County, Washington State, USA. King County, Washington is the most populous county in Washington. It is also the 12th most populous in the United States. Washington's most populous city is Seattle. Home buyers and sellers have a difficult time predicting a house's price. I will determine the most important features that will impact the price of the house and make a predictive model to determine a house's price for online property listings and their users.

Data

Our data is from Kaggle called *kc_data_house*. Below is a list of column names with their description and data type.

- **id:** a notation for a house - Numeric
- **date:** Date house was sold - String
- **price:** Price is prediction target - Numeric
- **bedrooms:** Number of Bedrooms/House - Numeric
- **bathrooms:** Number of bathrooms/bedrooms - Numeric
- **sqftliving:** square footage of the home - Numeric
- **sqftlot:** square footage of the lot - Numeric
- **floors:** Total floors (levels) in house - Numeric
- **waterfront:** House which has a view to a waterfront - Numeric
- **view:** Has been viewed - Numeric
- **condition:** How good the condition is (Overall). 1 indicates worn out property and 5 excellent. (http://info.kingcounty.gov/assessor/esales/Glossary.aspx?type=r#g) - Numeric
- **grade:** overall grade given to the housing unit, based on King County grading system. 1 poor ,13 excellent. - Numeric
- **sqftabove:** square footage of house apart from basement - Numeric
- **sqftbasement:** square footage of the basement - Numeric

- **yrbuilt:** Built Year - Numeric
- **yrrenovated:** Year when house was renovated - Numeric
- **zipcode:** zip - Numeric
- **lat:** Latitude coordinate - Numeric
- **long:** Longitude coordinate - Numeric
- **sqftliving15:** Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area - Numeric
- **sqftlot15:** lotSize area in 2015(implies-- some renovations) - Numeric

Analysis

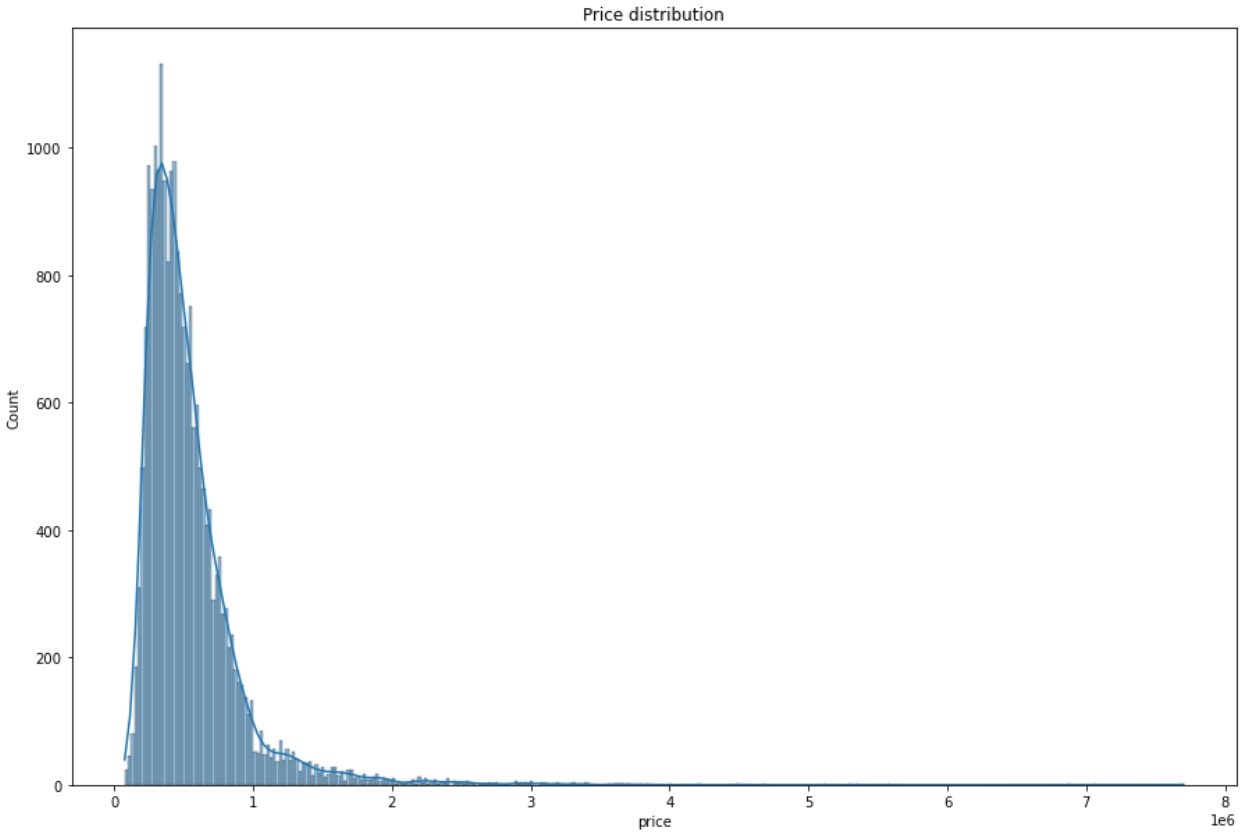

Figure 1 Price Distributions

Figure 1 shows the price distributions of all the house prices in our dataset. Most of the houses are under one million dollars.

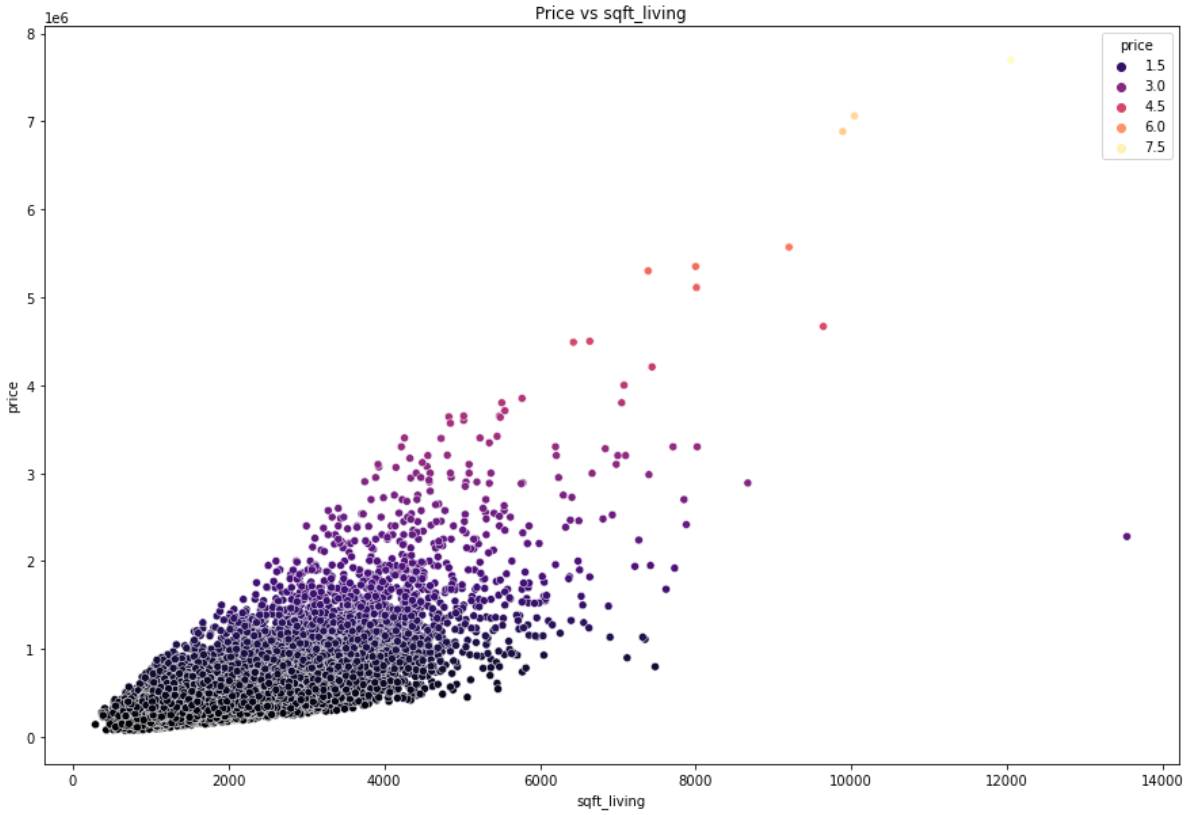

Figure 2 Price vs sqft_living

Figure 2 shows the relationship between *price* and *sqft_living*. The more expensive homes have over 10,000 square footages. Most of the homes are between 2,000 and 6,000 square feet. Based on the correlation heatmap, *price* and *sqft_living* have the highest correlated relationship with 0.70. *Sqft_living* is the most important feature for determining price.

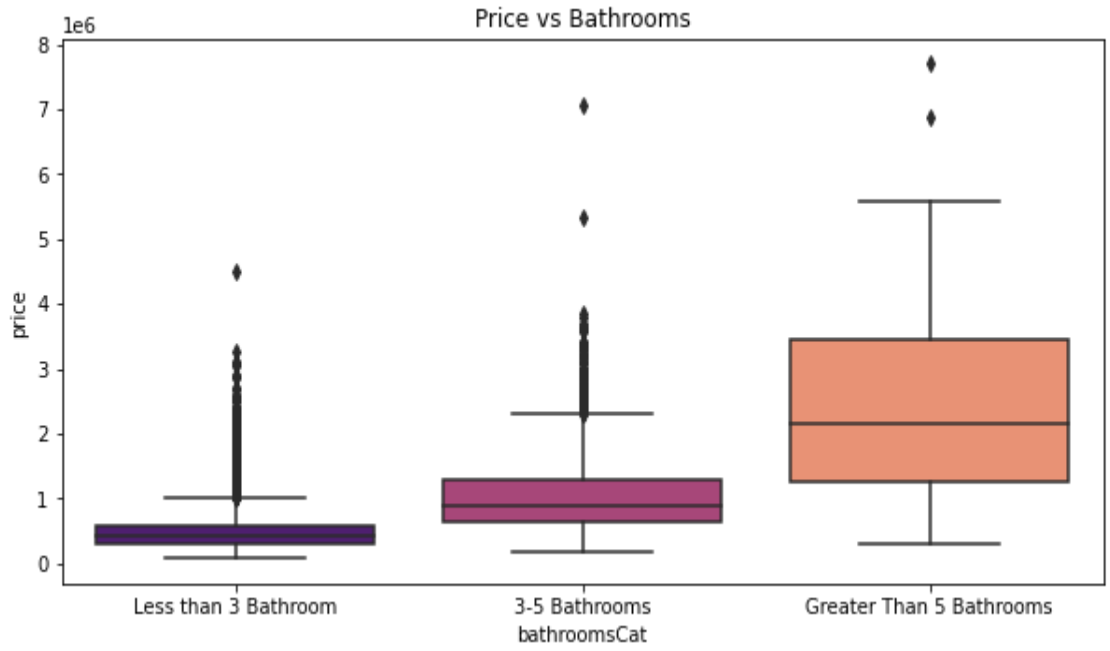

Figure 3 Price vs Bathrooms

Figure 3 above shows the relationship between price in bathrooms. *Price* and *bathrooms* have a positive correlation according to our heatmap. Typically, the houses with more bathrooms have a higher price. They also have a higher median. A home seller could sell their home for a higher price if they have a home with more than five bathrooms. A home buyer can expect to pay more if the home has more than five bathrooms.

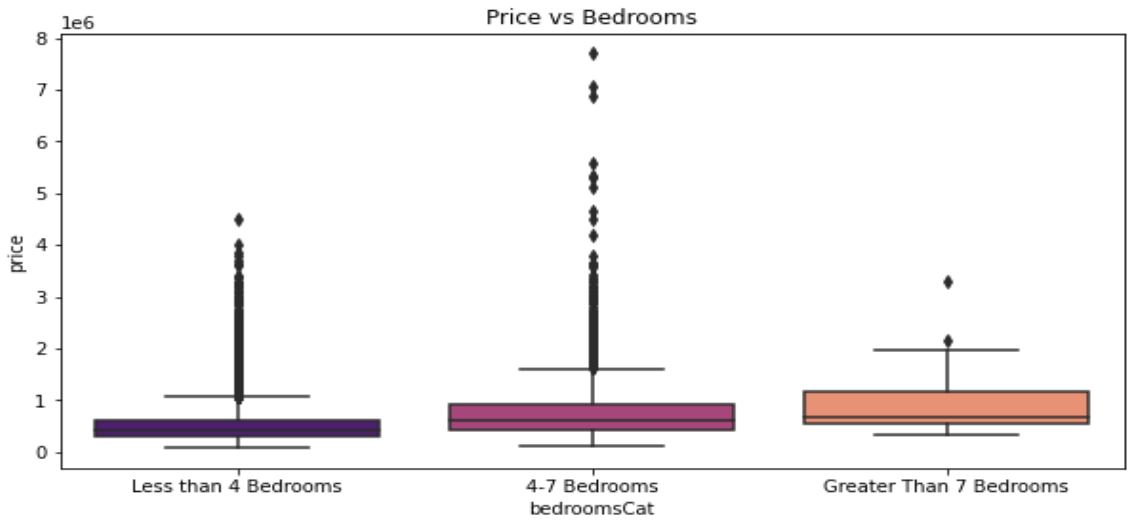

Figure 4 Price vs Bedrooms

Figure 4 shows the relationship between *price* and the number of *bedrooms*. The houses with 4-7 bedrooms have the highest house price. The houses with 4-7 bedrooms have the highest range in the dataset. A home seller's house doesn't have to have more than seven bedrooms to have a higher price. Based on the correlation heatmap, *price* and *bedrooms* have a correlation of 0.31. Homes with 4-7 bedrooms and homes with more than 7 bedrooms have similar medians.

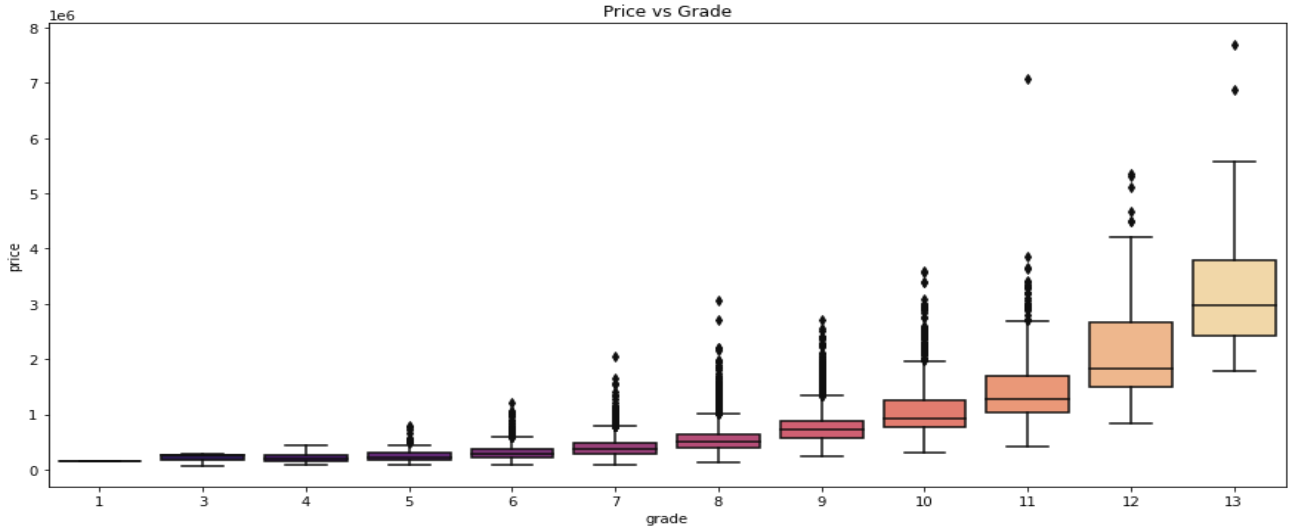

Figure 5 Price vs Grade

Figure 5 shows the relationship between *price* and *grade*. Based on the correlation heatmap, it has a positive correlation of 0.67. Typically, the houses with a higher grade have a higher house price. They also have a higher median. A home seller can sell their house for a higher price if their home has a higher grade.

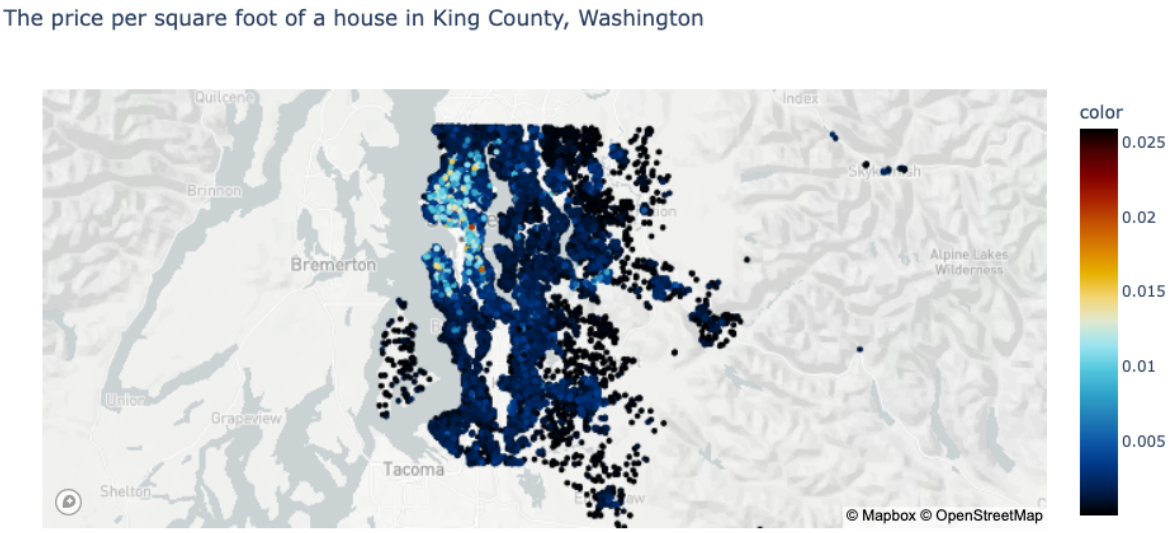

Figure 6 King County Map

First we need to take a closer look at the area. Figure 6 is a map of the King County, Washington area. It shows the price per square foot of a house. Looking at the region of King County, Washington, the closer the house is to Seattle, Washington, the more expensive it is. The color key represents the house price per sqft. The map shows that the closer the house is to Seattle, Washington, the more expensive the house per square foot is.

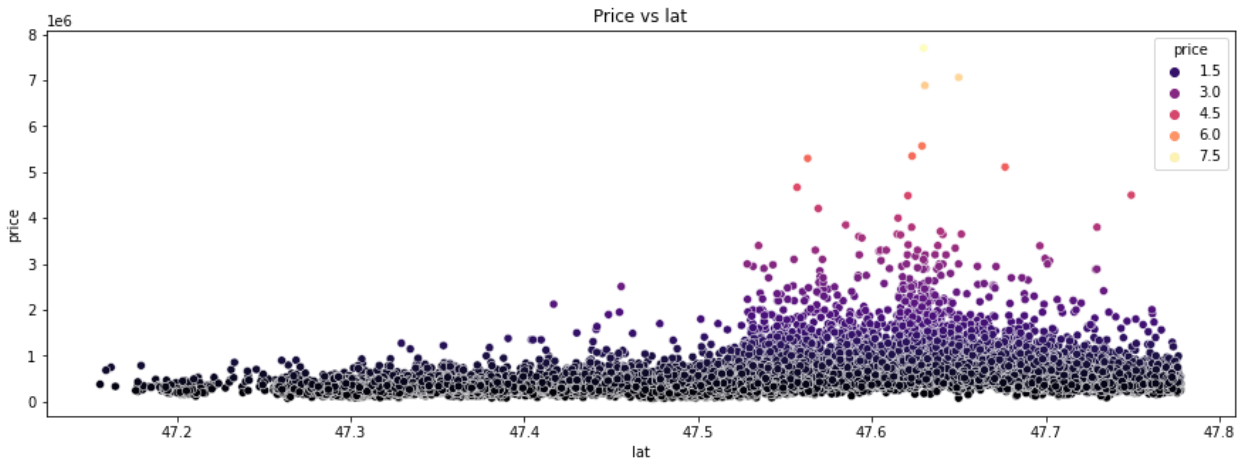

Figure 7 Price vs Latitude

Figure 7 shows the relationship between *price* and *latitude*. The houses with latitudes between 47.6 and 47.7 have some of the highest house prices. This is also where most of the outliers lay. The latitude for Seattle, Washington is 47.6062 N. This makes sense that the houses between 47.6 and 47.7 are some of the highest priced houses. A home buyer can expect to pay more if the home's latitude is between 47.6 and 47.7.

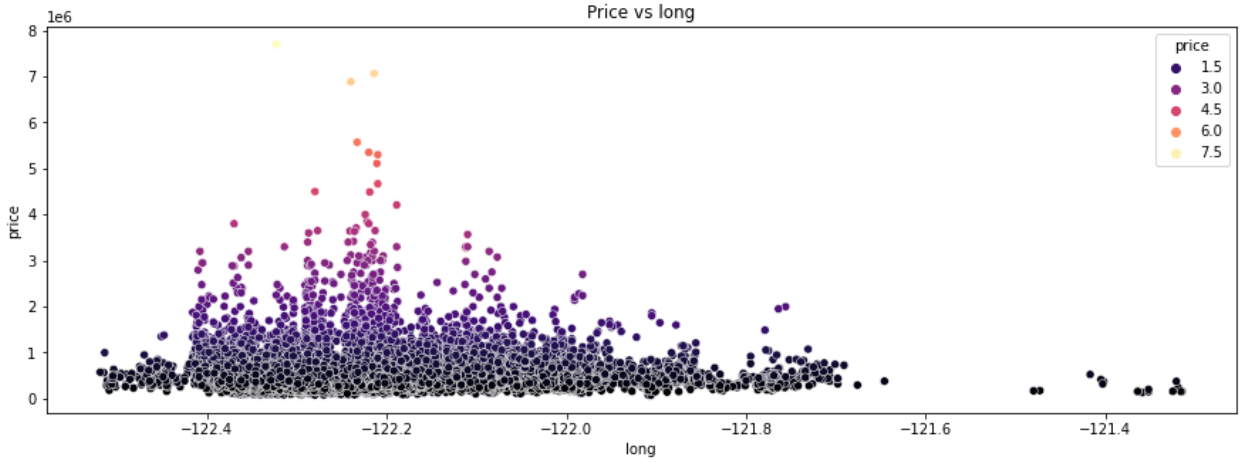

Figure 8 Price vs Longitude

Figure 8 shows the relationship between *price* and *longitude*. The more expensive homes have a longitude close to -122.2. Seattle, Washington longitude is 122.3321 W. Based on our correlation heatmap, longitude does not have a significant relationship with price. It has a correlation of 0.022. A home buyer can expect to pay more if the home's longitude is closer to Seattle, Washington.

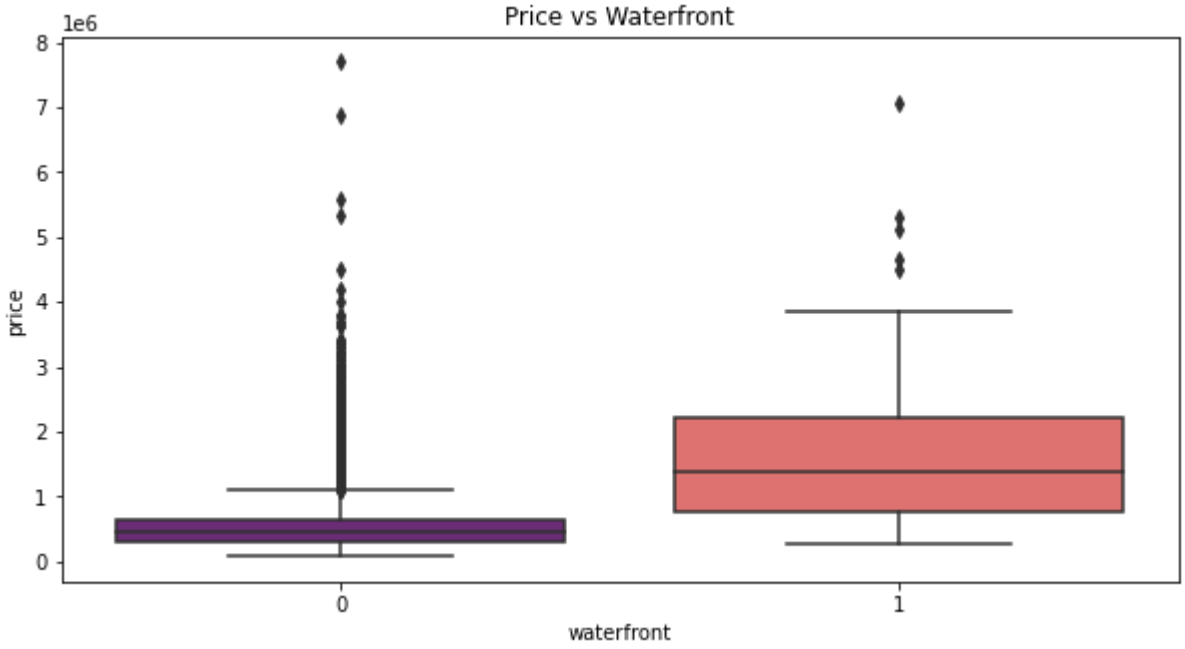

Figure 9 Price vs waterfront

Figure 9 shows the relationship between *price* and *waterfront*. Typically, houses with a waterfront view have a higher median house price, but the houses without a waterfront have higher outliers. A home seller doesn't have to have a waterfront to sell their home for a higher price.

Figure 10.1 Top House Prices and their zip code

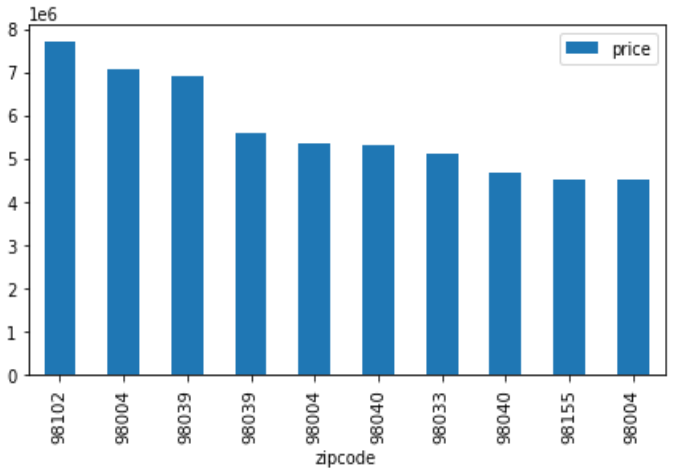

Figure 10.2 Bottom House Prices and their zip code

Figure 10.1 and 10.2 shows the relationships between *price* and the *zipcode* of the house location. Figure 10.1 shows the highest ten priced houses with their zip code, and figure 10.2 shows the lowest ten priced homes and their zip code. Based on the correlation heatmap, *price* and *zipcode* have a correlation of -0.053. A future home buyer/seller can keep in mind that the zip code will not be a major factor in the price.

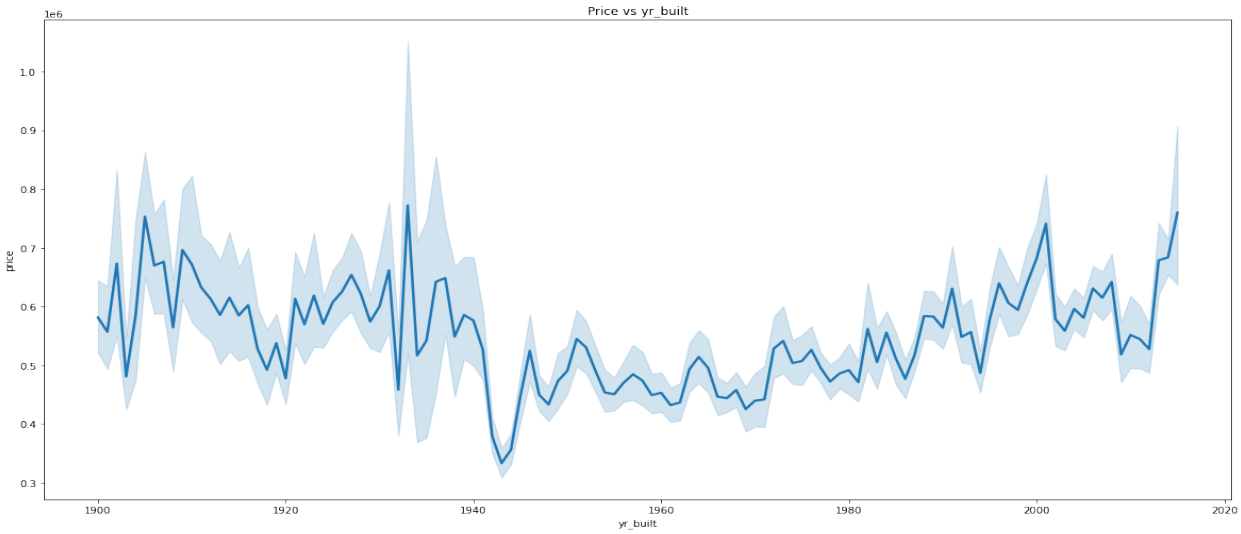

Figure 11 Price vs yr_built

Figure 11 shows the relationship between *price* and the year the house was built. You can see the trend of house prices over the years. Based on the correlation heatmap, *price* and *yr_built* does not have a high correlation. It has a correlation of 0.054. The lowest house price was between 1940 and 1960.

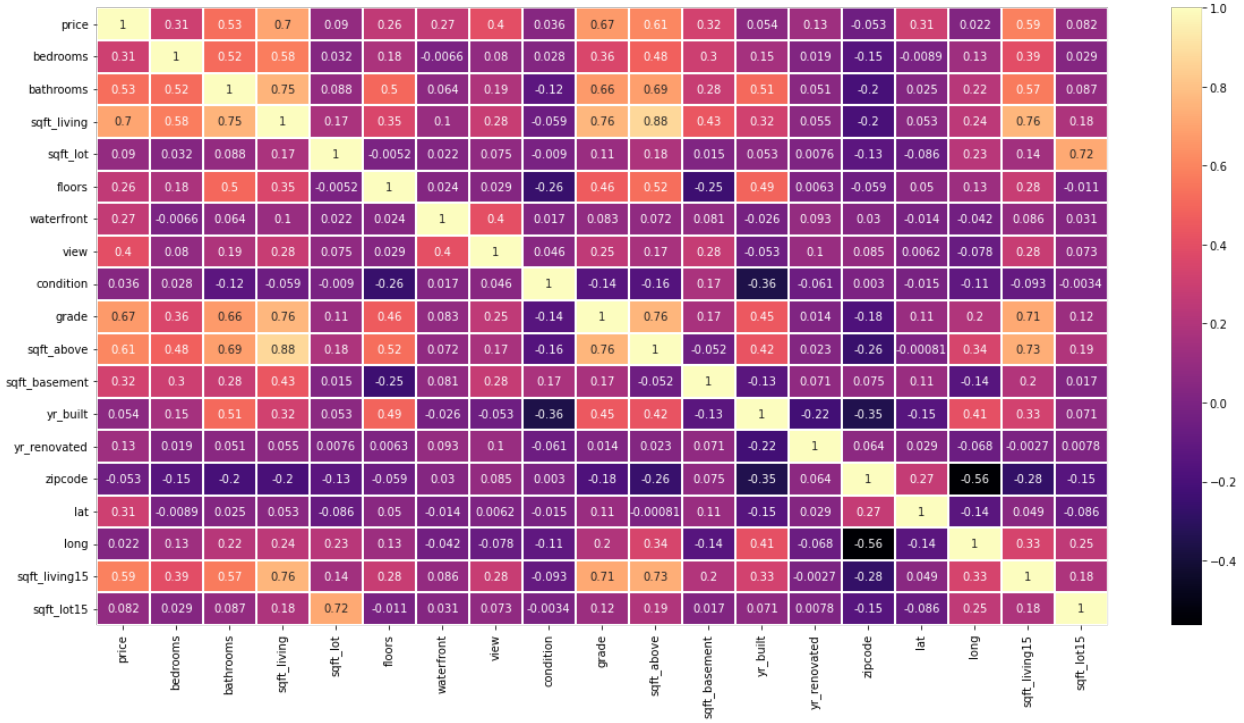

Figure 12 Correlation Heatmap

In Figure 12, a correlation heatmap depicts the relationships between the features of the DataFrame. The heatmap lets us know that there's a strong positive correlation between 'price' and 'sqft_living'(0.70), 'grade'(0.67), and 'sqft_above'(0.61). These features will affect the price of the house. The heatmap also shows if there are any multicollinear variables. There is a high correlation between two independent variables that is above 0.80. Those variables are ‘sqft_living’ and ‘sqft_above’. They have a correlation of 0.88. ‘Sqft_living’ is the square footage of the entire house

including the basement. ‘Sqft_above’ is the square footage above the basement. As the ‘sqft_living’ increases, there is a strong possibility that the ‘sqft_above’ will also be increased. ‘sqft_living’ and ‘sqft_above’ are considered multicollinear.

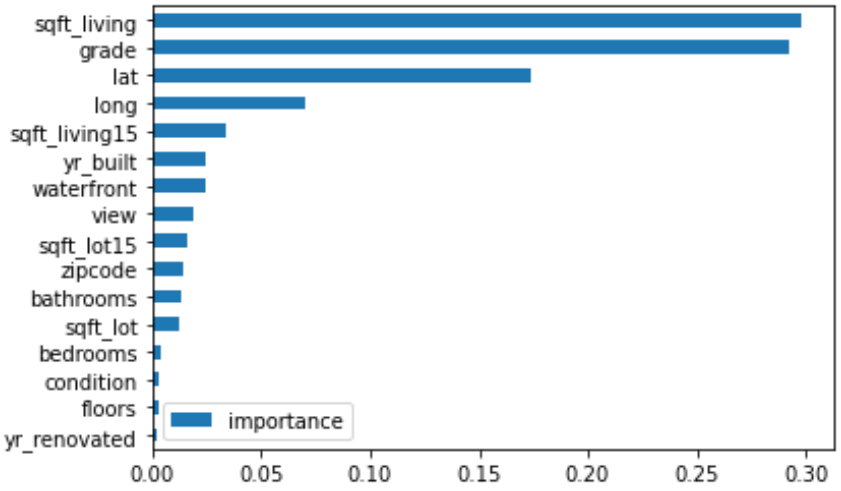

Figure 13.1 Random Forest Regressor Feature Importance

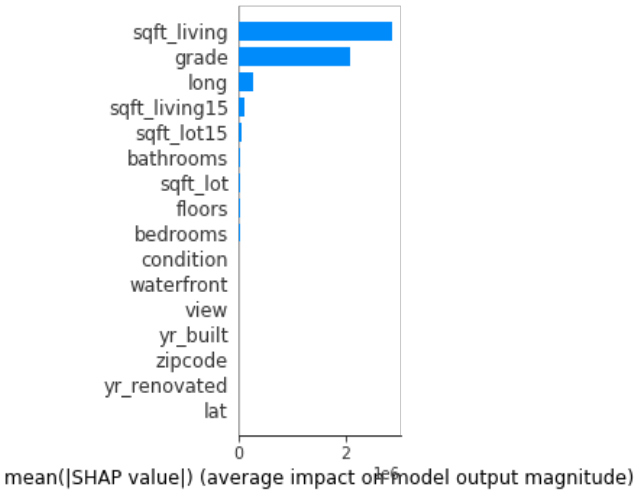

Figure 13.2 SHAP Feature of Importance

Figure 13.1 shows that the top four features are “sqft_living”, “grade”, “lat”, “long” are the top features that will affect the price of the house. It is based on the Random Forest Regressor Feature Importance. Figure 13.2 shows that the top four features are “sqft_living”, “grade”,”long”, and “sqft_living15” are the top features that will affect the price of the house. It is based on the SHAP feature importance. Feature importance describes which features are relevant. The Scikit-learn’s feature importance tells us how much a feature is used in each tree of the forest. The SHAP feature importance is based on the prediction of an instance by computing the contributions of each feature to the prediction.

Results

Model	Best Parameters	RMSE Score	MAPE
Linear Regression	**n/a**	$ 202,853	26.7%
Ridge Regression	**alpha**: 1	$ 202,870	26.7%
Random Forest Regressor	**max_depth**: 30, **max_features**: 'auto', **n_estimators**: 400	$ 127,996	13.1%

Table 1 Model Comparisons

Table 1 shows the comparisons of the three models I created: Linear Regression, Ridge Regression, and Random Forest Regressor. I used GridSearchCV to find the best parameters to use for each model. The “Best Parameters” column is the result of the GridSearchCV. The “RMSE Score” column represents the root mean squared error of each model. The “MAPE” column represents the mean absolute percentage error. Random Forest Regressor has the lowest root mean squared error and mean absolute percentage error.

Model	Test R^2 score	RMSE Score	MAPE
rf_model	0.88	$126,597	13%
rf_model2	0.87	$142,506	13.8%

rf_model3	0.83	$159,335	15%
rf-model4	.70	$213,361	29%

Table 2 Random Forest Regressor Model Comparisons

There is overfitting in the modeling. I created four models with different independent variables to determine if I can lessen the overfitting. ‘Rf_model’ is the model with the original independent variables. ‘Rf_model2’ is without the insignificant variables based on the Random Forest Regressor Feature Importance. ‘Rf_model3’ is the top four features based on the Random Forest Regressor feature importance. “rf_model4” is based on the top four features based on the SHAP feature importance. The “rf_model” model with the original independent variables has the highest R squared score. The “rf_model” has the lowest root mean squared error. The “rf_model” has the lowest mean absolute percentage error.

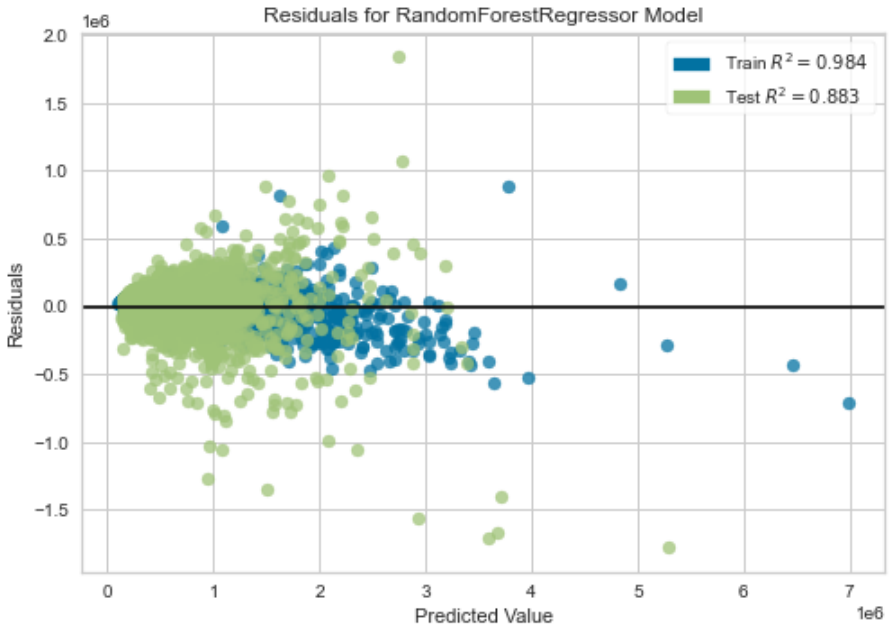

Figure 14 Residuals Plot

Figure 14 is a residual plot of the results of the Random Forest Regressor model.

Conclusion

Buying and selling a house is an exciting chapter in someone's life. There are factors that determine a house's price based on its features like bathrooms, bedrooms, floors, etc. Based on our data on the King County, Washington house data from 2014-2015, 'grade', 'sqft_living', 'lat', and 'long' are the top features based on the Random Forest features' importance. I used GridSearchCV to determine the best parameters to use and created a Random Forest Regressor model with an R-squared score of 0.88. It has a root mean square error of $126,597. The next step is to improve the model and correct the overfitting in the model. The "rf_model" is the one with the best R-squared score and has the most independent variables. The "rf_model" has the lowest root mean squared error. I believe this dataset needs more information on the houses sold to better predict the house's price. This is an older dataset. I would be interested to explore an updated dataset of housing prices. It would also be interesting to view the year-to-year and month-to-month differences or trends of the houses sold. If the dataset needs more data to correct the overfitting, I suggest getting data on all the houses in King County, Washington.