# Predicting King County, Washington House Prices

By: Kierra Dangerfield

# Problem

- Buying and/or selling a home can be a stressful time in someone's life.
- Buyers and sellers need to be able to predict house prices.
- It can be difficult to determine what features of a house will affect a home's price.
- Need to create a predictive model that home sellers, home buyers, and online property listings can use.

# Data: Overview & Cleaning

- Data is from Kaggle
- Data is based on houses sold between May 2014 to May 2015.
- There were no duplicates

# Data Cleaning: Missing Values

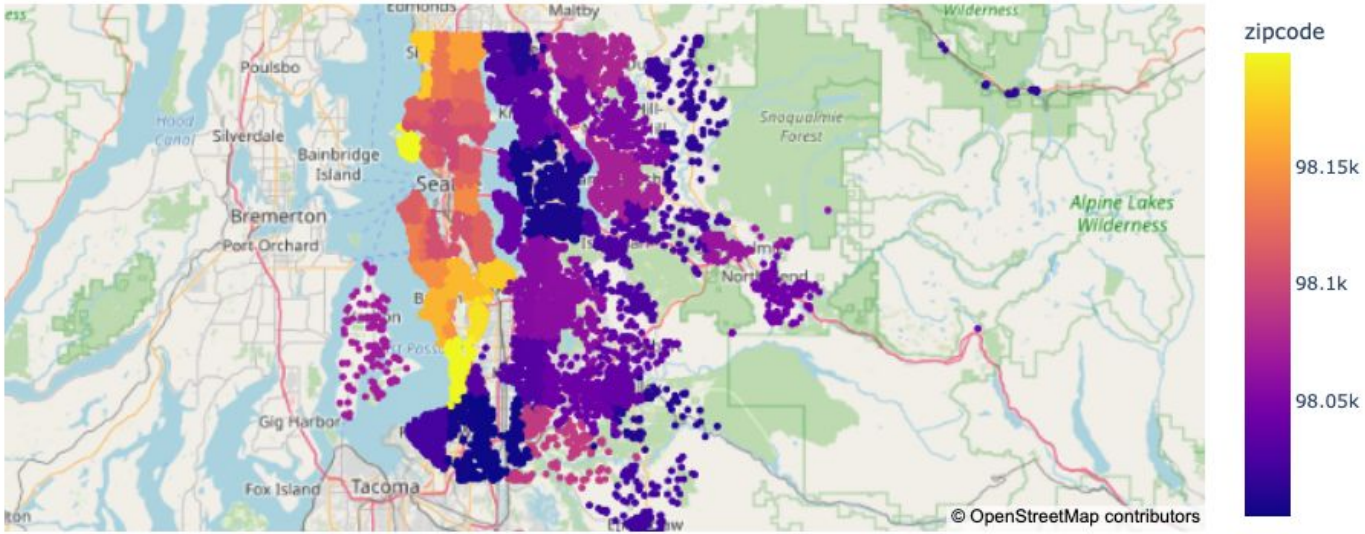
- There were some missing values.
- There were only 2 rows with missing values in the ‘sqft_above'.
- There is a row with the "bedrooms" column as 33.

# Data Cleaning: Outliers

- There are 13 rows with the “bedrooms” column as 0.
- There are outliers in the "sqft_lot" column.
- There are 5 rows with the “sqft_lot” above 75,000 square footage.
- The average went from over 15k to almost 10k square footage.

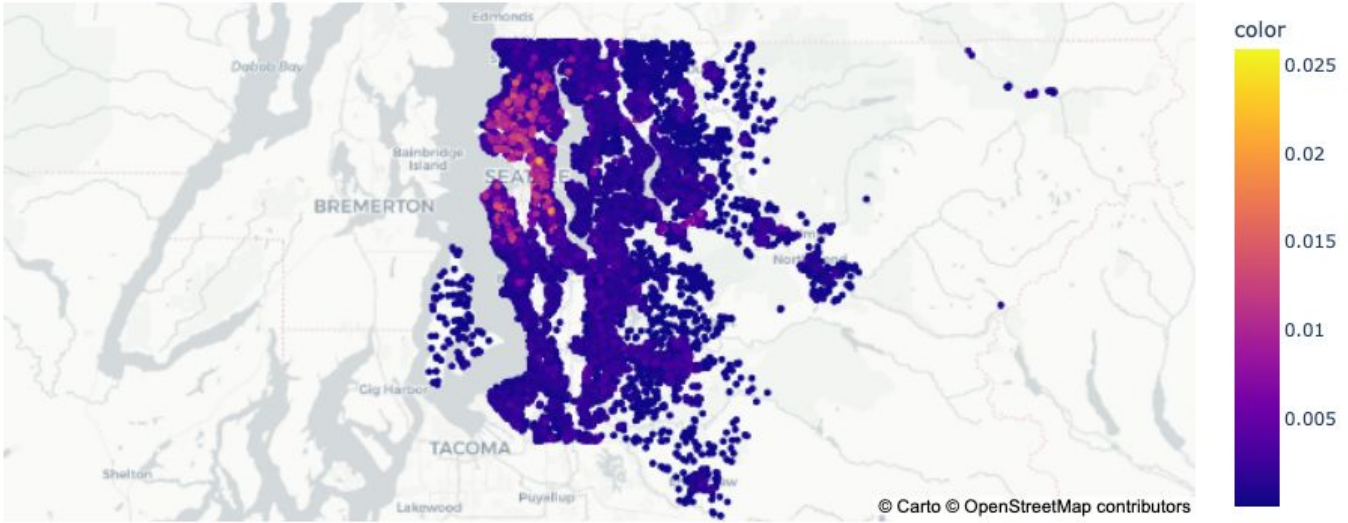
# Analysis: Location

The number of properties of each area in King County, Washington



# Analysis: Location Cont.

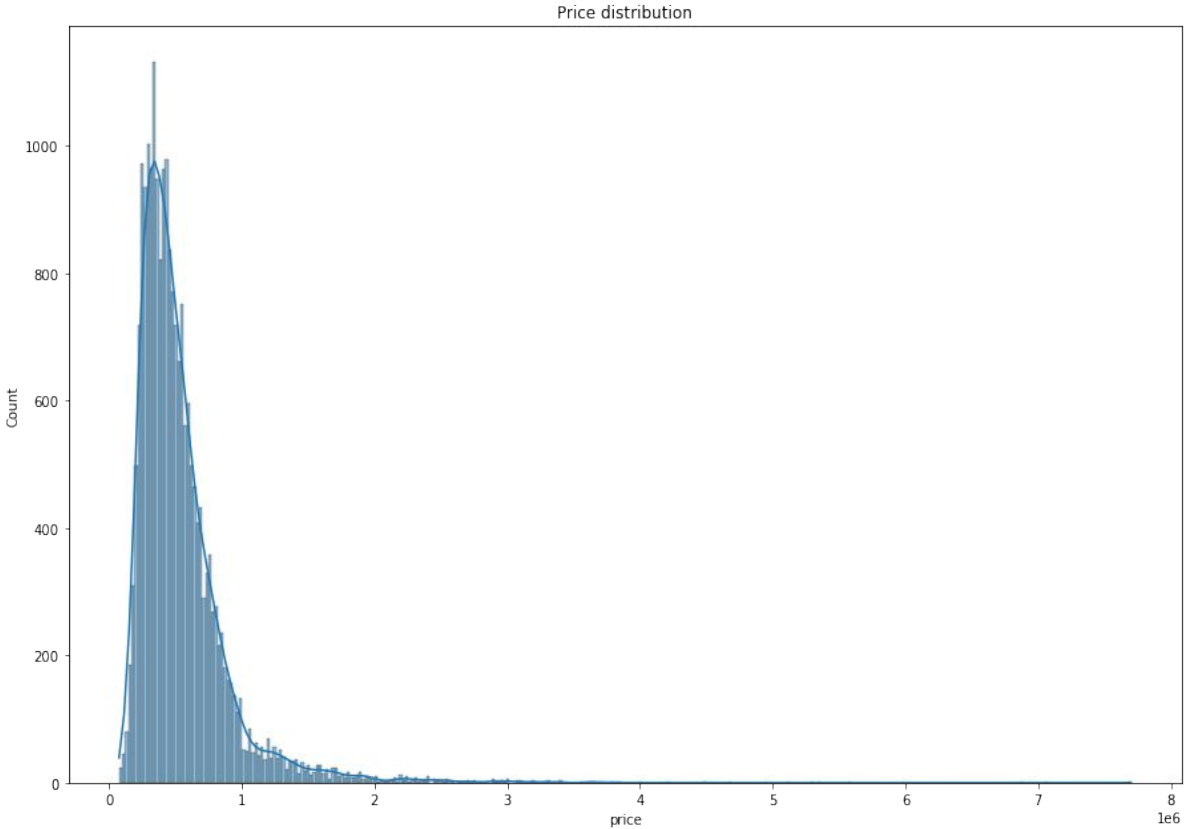
The price per square foot of a house in King County, Washington



# Distribution of Target Variable: Price

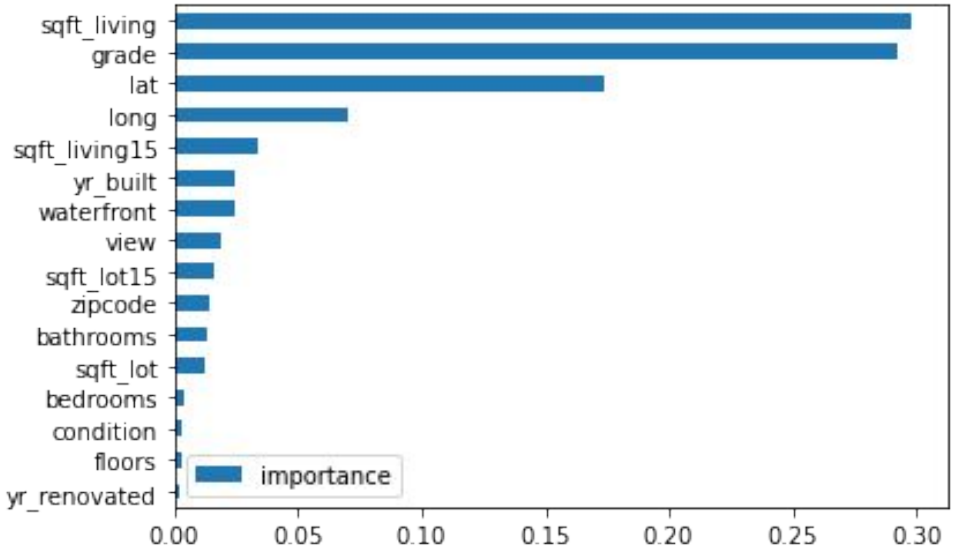
- Based on the distribution of price

- There are two modes: **$350,000** and **$450,000**.



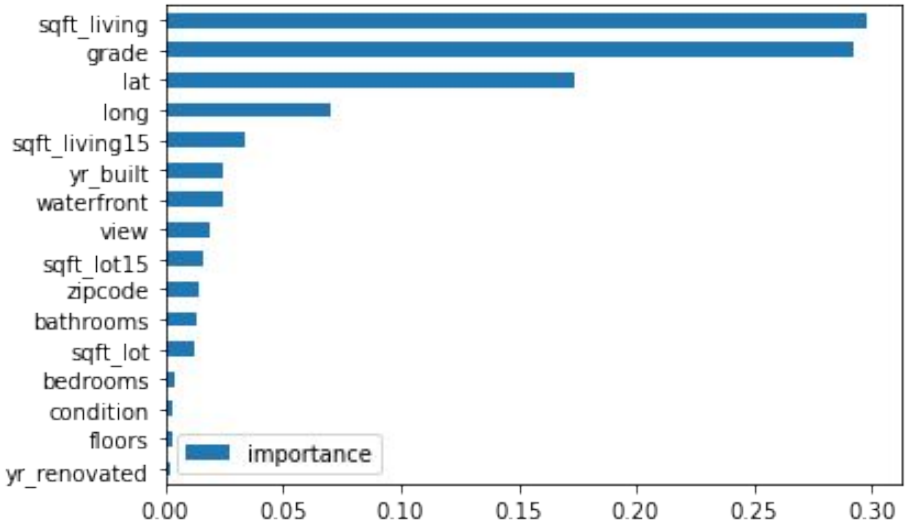
# Random Forest Features Importance

- The feature importance chart tells us the most important features in determining the price of the house.
- “Sqft_living” is the square footage of a house.
- It is the most important feature in determining the price of a home.



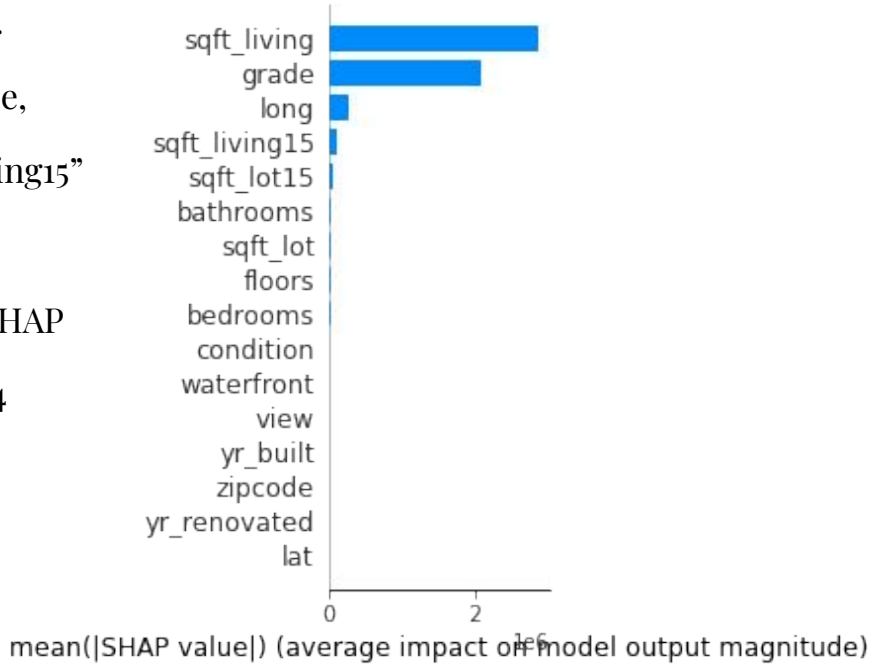
# Random Forest Features Importance

- "Grade" represents the overall grade given to the housing unit, based on King County grading system. 1 poor ,13 excellent
- "lat" and "long" represents the latitude and longitude of a house.
- We will focus on these top 4 features in determining the price.



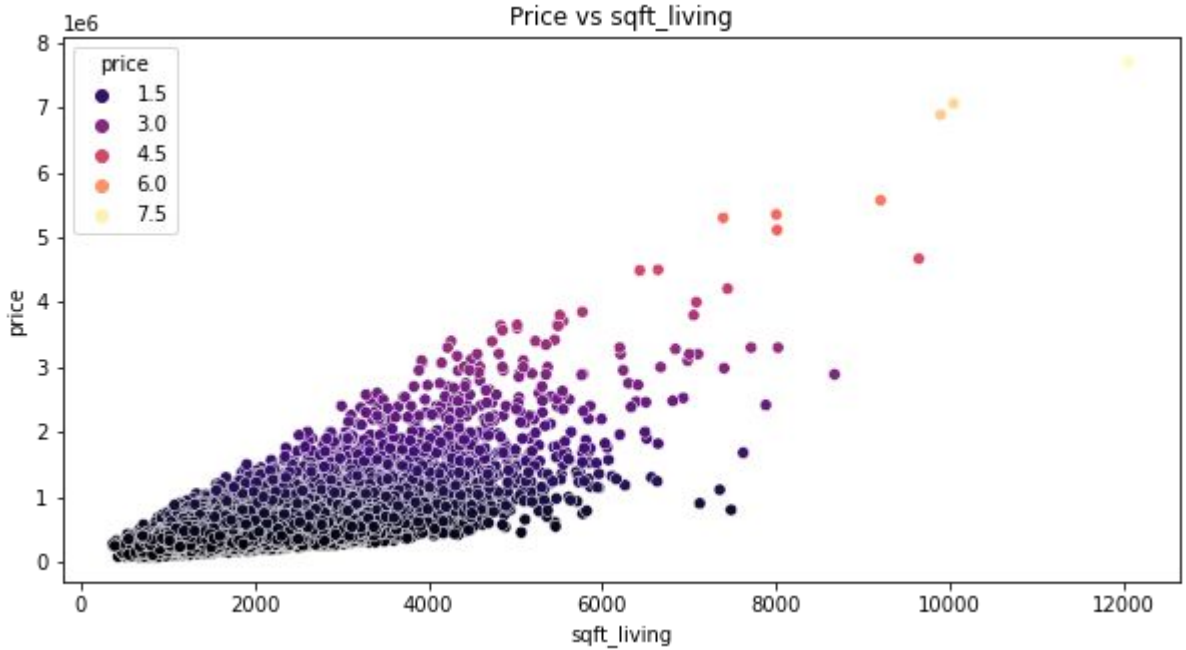
# SHAP Feature Importance

- SHAP is based on feature attributions.
- Based on the SHAP feature importance, “sqft_living”, “grade”, “long”, “sqft_living15” are the top four features.
- The RF Feature Importance and the SHAP Feature Importance have similar top 4 important features.



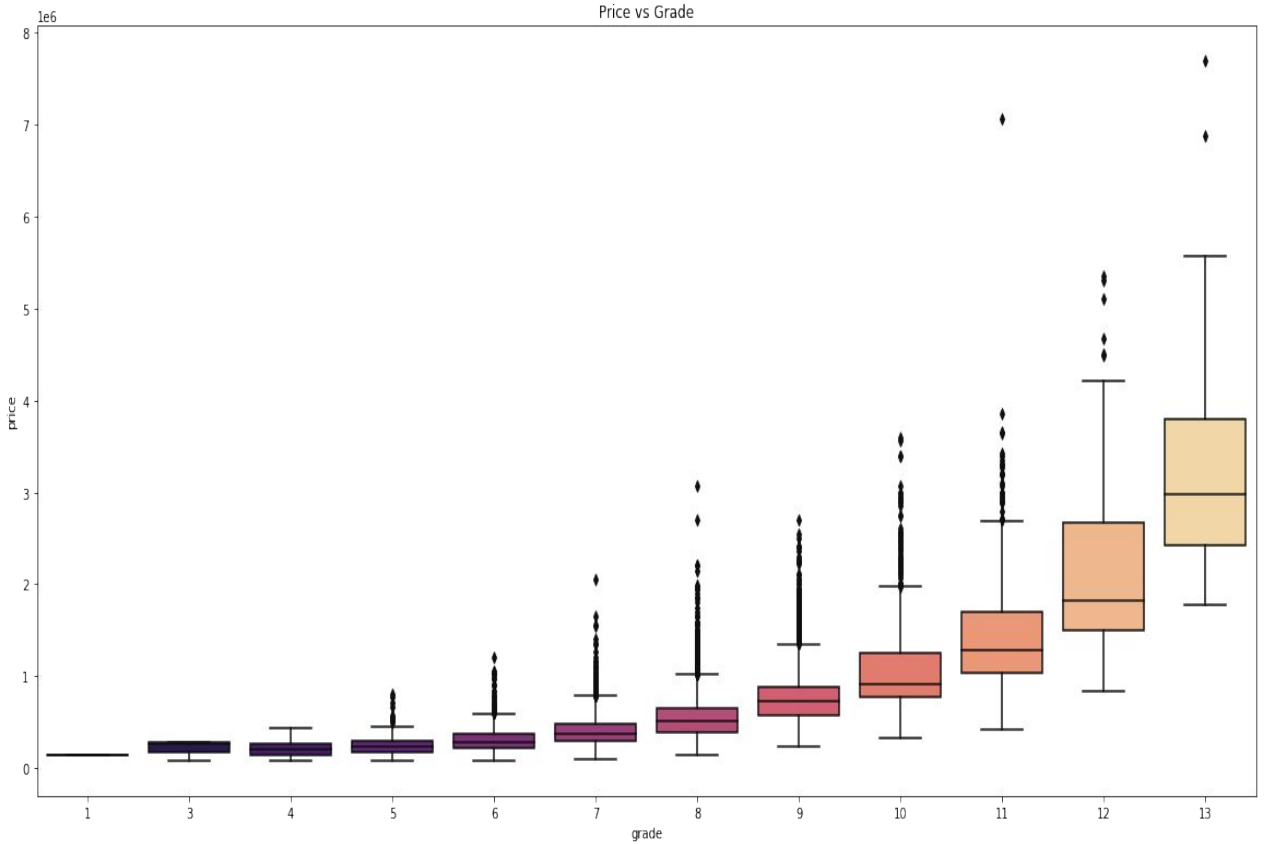
# Analysis: Square Footage vs Price

- This chart is showing that the higher the square footage of a house, typically the higher the price will be.
- The most expensive home has 12,000 square feet.
- Most of the homes are between 2,000 and 6,000 square feet.



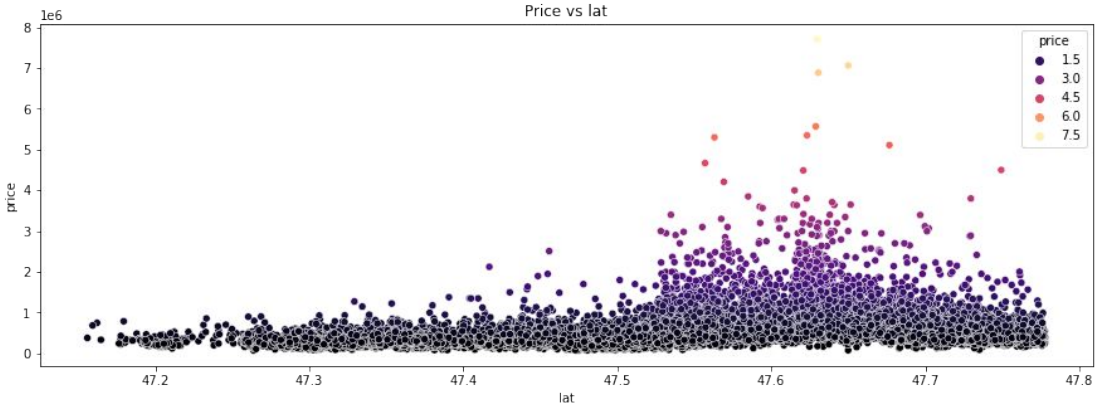
# Analysis: Grade vs Price

- The grade is based off of King County's grading system.
- You should expect a higher house price if you have a better grade.
- Typically, the houses with a higher grade have a higher house price.



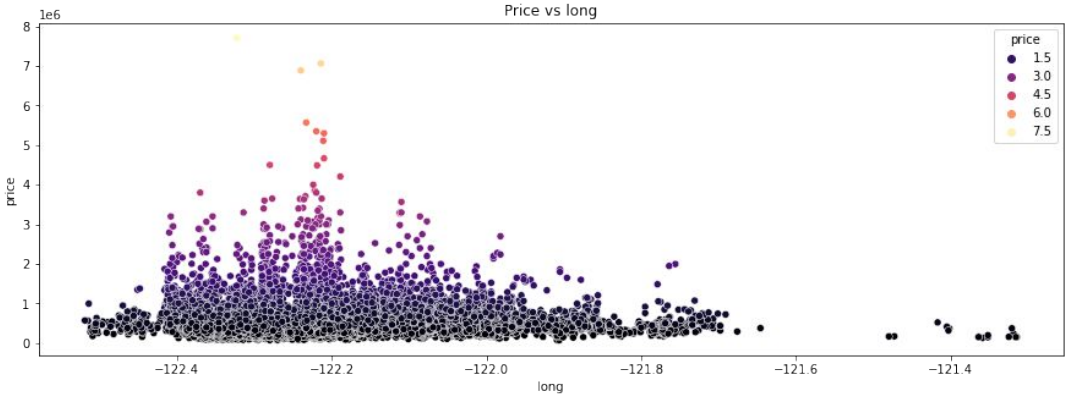
# Analysis: Latitude vs Price

- Seattle, Washington has a latitude of 47.608013.
- The closer a house is near Seattle, it's a possibility of it will have a higher house price.
- The higher priced homes are in between 47.5 and 47.7 latitude.



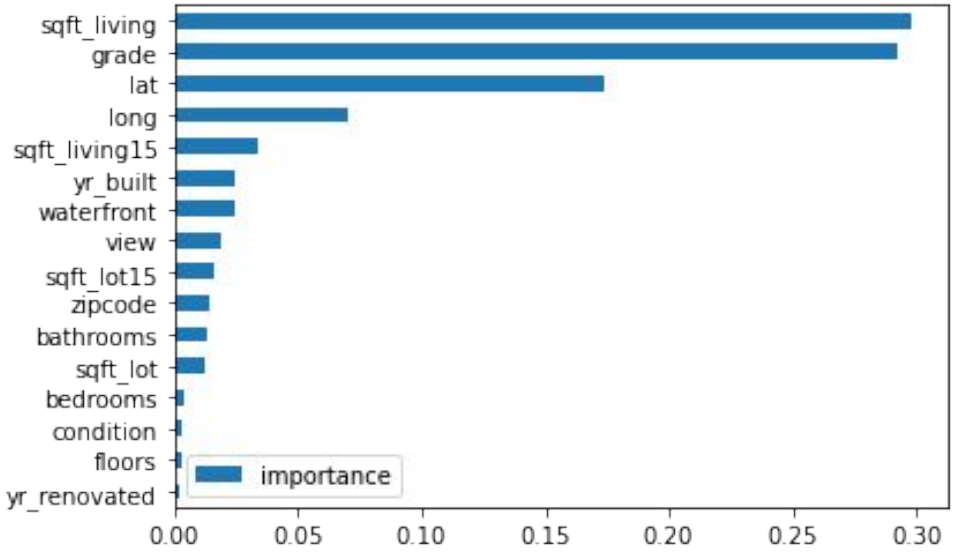
# Analysis: Longitude vs Price

- Seattle, Washington has a longitude of -122.335167.
- The closer a house's longitude is near Seattle, it's a possibility of it will have a higher house price.
- The higher priced homes are in between -122.4 and -122.2 longitude.



# Results: Top Features

- ‘grade’, ‘sqft_living’, ‘lat’, and ‘long’ are the top features in considering a house’s price.
- “Bedrooms”, “condition”, “floors”, “yr_renovated” are insignificant variables based on the feature importance.
- I used 3 models to determine the best model to use



# Results: Three Models

- **Linear Regression** shows the relationship between 2 variables.
- **Ridge Regression** is a L2 regularization that helps with overfitting.
- **Random Forest Regressor** is a supervised learning algorithm that uses ensemble learning method for regression to make a more accurate prediction than a single model.

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

# Results: Three Models - cont.

- I used a Linear Regression, Ridge Regression, and a Random Forest Regression model.
- Used GridSearchCV to determine the best parameters
- Random Forest Regressor performed the best based on the RMSE (root mean squared error)

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

# Results: RMSE AND MAPE

- Root Mean Squared Error (RMSE) is the square root of the average squared error.
- RMSE is used to evaluate the quality of predictions by showing how far predictions fall from measured true values.
- Mean Absolute Percentage Error (MAPE) is used to measure forecast accuracy.
- MAPE is the average of the percentage errors

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

# Results: Random Forest Models

- The parameters for the model is the following: max_depth=30, max_features='auto', n_estimators=400)
- I created 4 different Random Forest Models with different independent variables.
- 'Rf_model' is the model with the original independent variables.

| 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 | 0.70 | $213,361 | 29% |

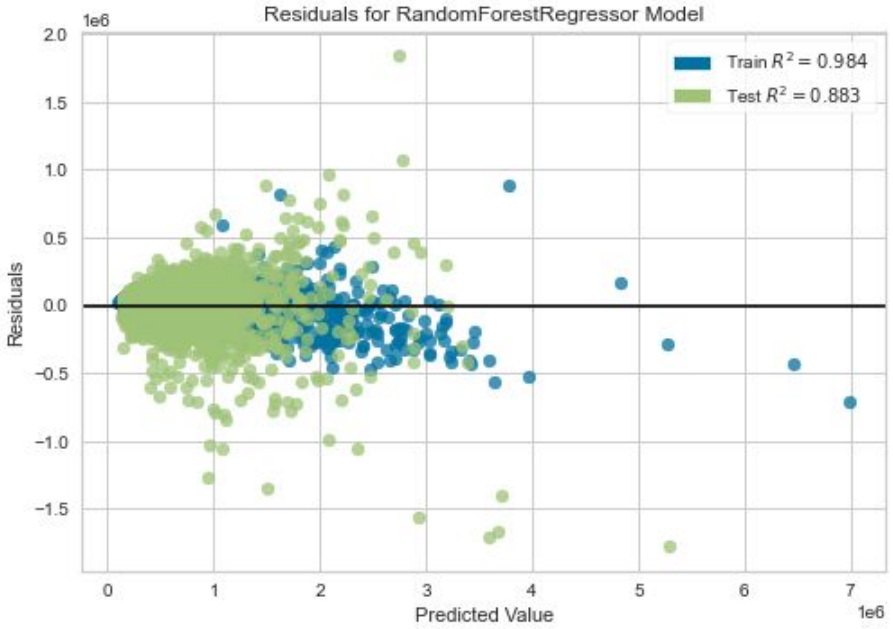
# Results: Random Forest Models Cont.

- “rf_model2” is without the insignificant variables based on the Random Forest Feature Importance.
- "rf_model3” is the top features based on the Random Forest feature importance.
- “rf_model4” is the top features based in the SHAP feature importance.
- “rf_model” is the model that has performed the best.

| 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 | 0.70 | $213,361 | 29% |

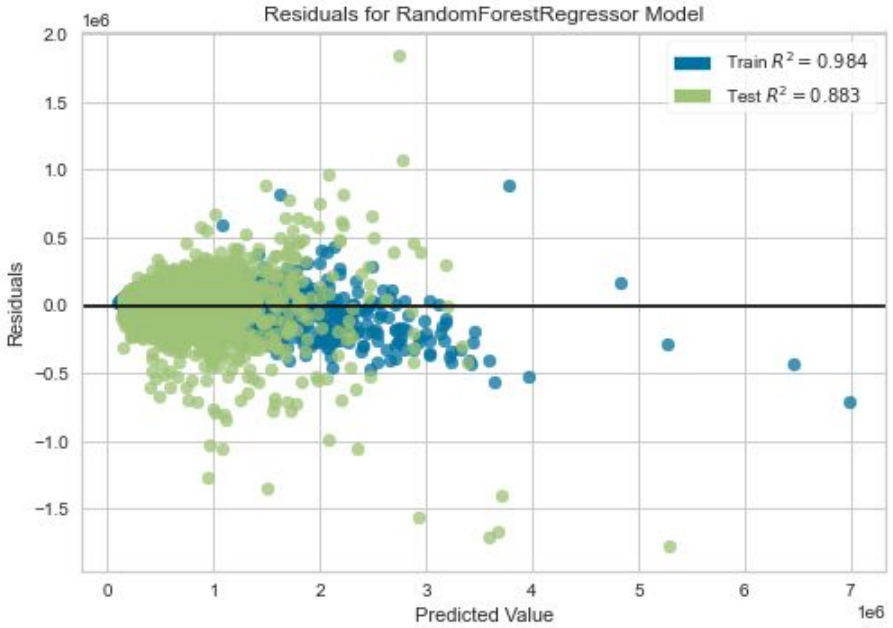
# Conclusion

- Before a home buyer, home seller, or online property listing can use this model, there will need to be improvements.
- Based on the residuals plot for the "rf_model", it shows that there is overfitting.
- The residuals plot shows the errors of the prediction between the observed value and the predicted value.



# Conclusion: Cont.

- The next step is to correct the overfitting.
- One way to correct the overfitting is to get more data on the houses sold.
- Hopefully, after making those improvements, this model can be used for a home buyer/seller or an online property listing.



# End