

House Price Prediction in King County, USA : **A Regression Modeling Approach**

by Intania Rahmadhilla



Table of contents



01

Data Understanding

02

Exploratory Data Analysis

03

Data Preprocessing

04

Modeling

05

Hyperparameter Tuning

06

**Conclusion &
Recommendations**



Case Study :

With the increasing business competition, a customer-centric approach is crucial for improving company profits. Shifting from a product-centric to a customer-centric approach requires a deep understanding of factors influencing house prices in King County, USA. Therefore, this project aims to create a predictive model for house prices.

Objective : To explore the factors influencing house prices, providing valuable insights for buyers, sellers, and other stakeholders.

Goal : Predict housing prices in King Country, USA based on variables contained in the data set. Identify key factors influencing house prices & provide recommendations for model optimization and improvement.

Dataset **Information**

The dataset is accessible at <https://www.kaggle.com/datasets/harlfoxem/housesalesprediction/data>

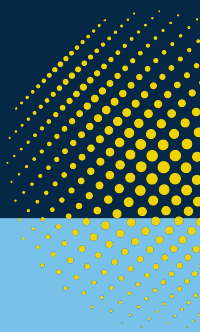
This is a data set containing the sales prices of homes for King County that were sold between May 2014 - May 2015. The dataset consists of :

21.614

rows

21

features



Features

1. id	8.floors	15. yr_built
2. date	9.waterfront	16. yr_renovated
3. price	10.view	17. zipcode
4. bedrooms	11.condition	18. lat
5. bathrooms	12.grade	19. long
6. sqft_living	13.sqft_above	20. sqft_living15
7. sqft_lot	14.sqft_basement	21. sqft_living

Statistical Summary

Numerical Data Type

3

	price	sqft_living	sqft_lot	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
count	2.161300e+04	21613.000000	21613.000e+04	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000	21613.000000
mean	5.400881e+05	2079.899736	1.510697e+04	1788.390691	291.509045	1971.005136	84.402258	98077.939805	47.560053	-122.213896	1986.552492	12768.455652
std	3.671272e+05	918.440897	4.142051e+04	828.090978	442.575043	29.373411	401.679240	53.505026	0.138564	0.140828	685.391304	27304.179631
min	7.500000e+04	290.000000	5.200000e+02	290.000000	0.000000	1900.000000	0.000000	98001.000000	47.155900	-122.519000	399.000000	651.000000
25%	3.219500e+05	1427.000000	5.040000e+03	1190.000000	0.000000	1951.000000	0.000000	98033.000000	47.471000	-122.328000	1490.000000	5100.000000
50%	4.500000e+05	1910.000000	7.618000e+03	1560.000000	0.000000	1975.000000	0.000000	98065.000000	47.571800	-122.230000	1840.000000	7620.000000
75%	6.450000e+05	2550.000000	1.068800e+04	2210.000000	560.000000	1997.000000	0.000000	98118.000000	47.678000	-122.125000	2360.000000	10083.000000
max	7.700000e+06	13540.000000	1.651359e+06	9410.000000	4820.000000	2015.000000	2015.000000	98199.000000	47.777600	-121.315000	6210.000000	871200.000000

Statistical Summary

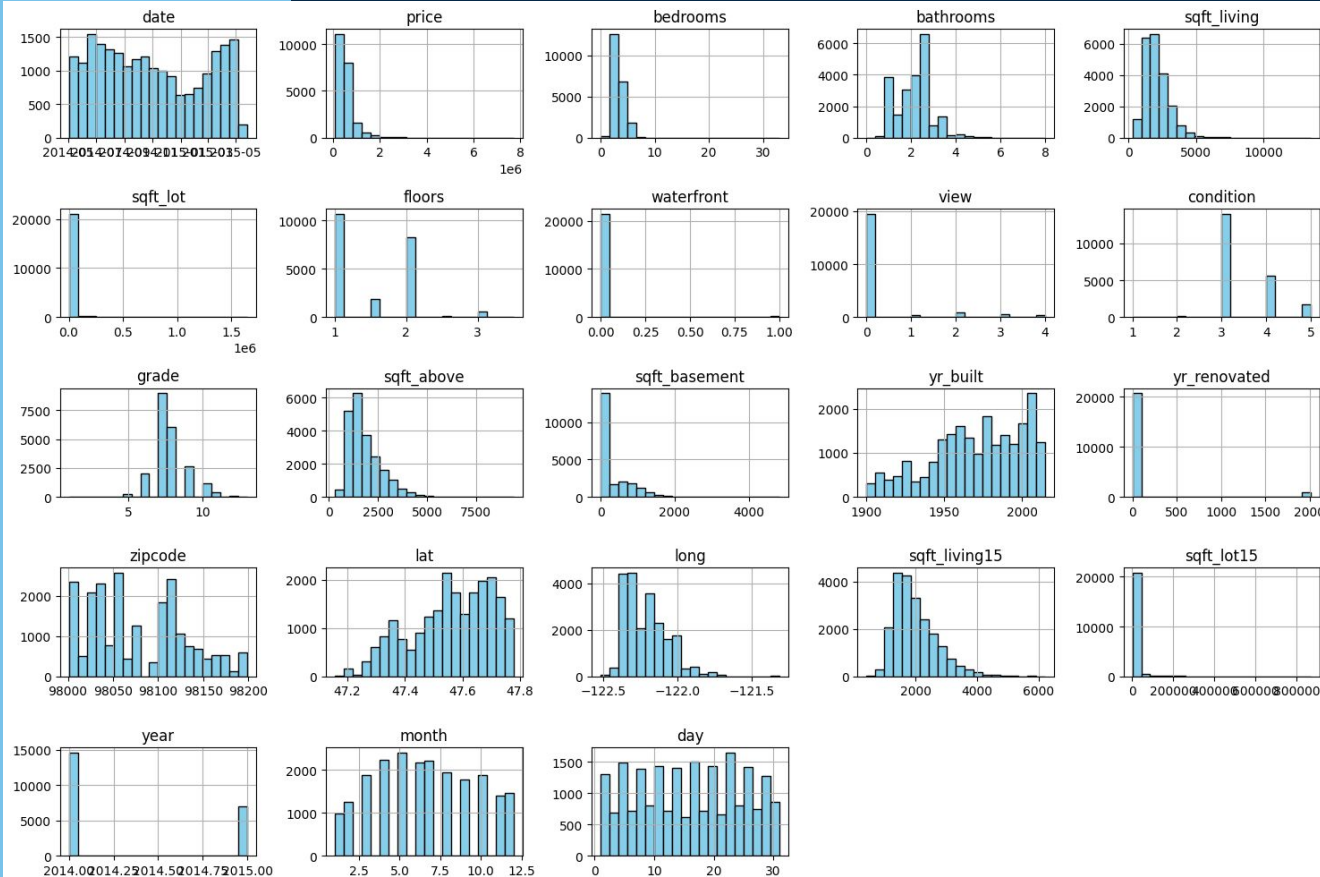
Categorical Data Type

	bedrooms	bathrooms	floors	waterfront	view	condition	grade
count	21613	21613.0	21613.0	21613	21613	21613	21613
unique	13	30.0	6.0	2	5	5	12
top	3	2.5	1.0	0	0	3	7
freq	9824	5380.0	10680.0	21450	19489	14031	8981

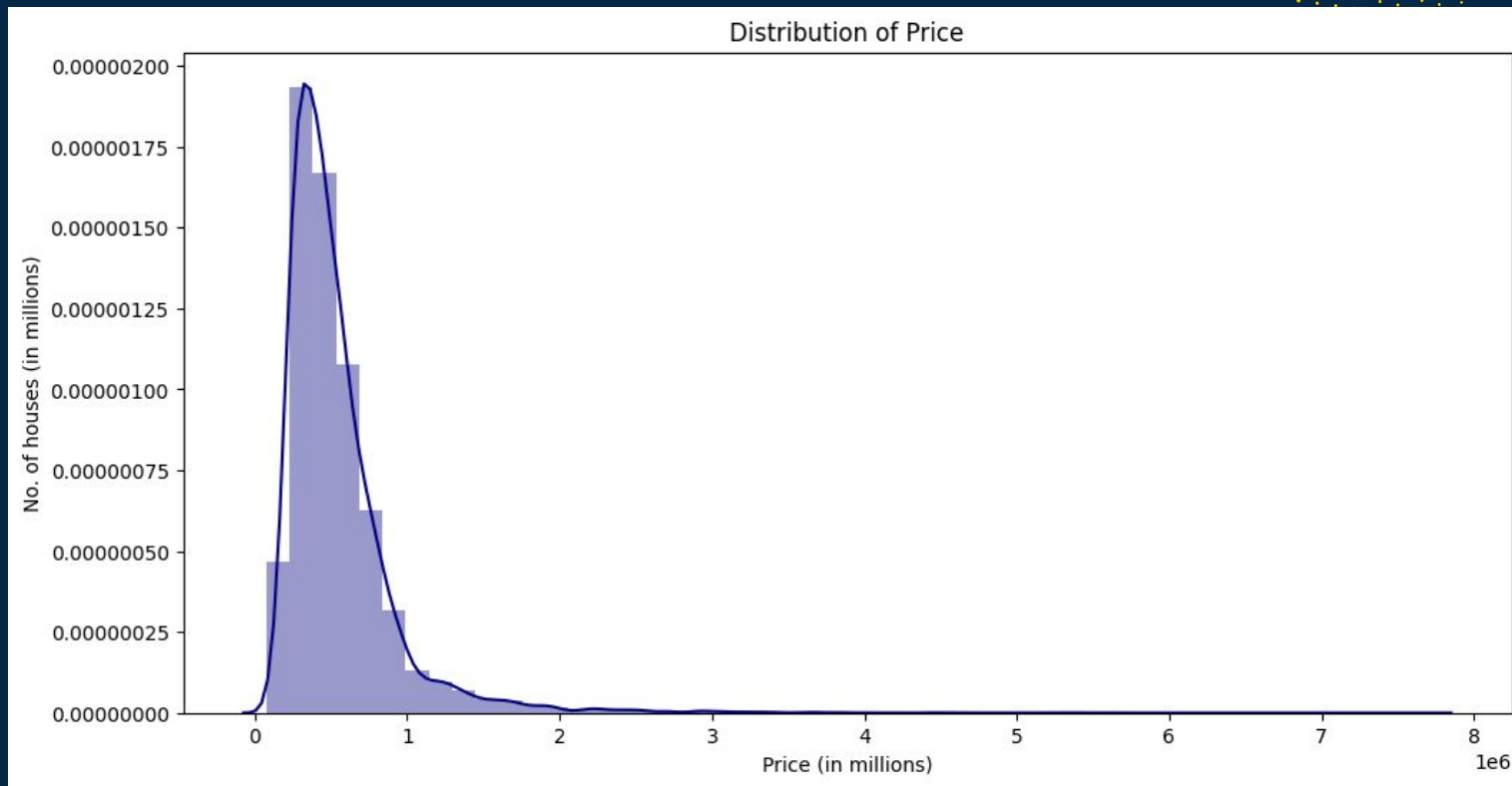
Exploratory Data Analysis

(Distribution data)

Examining the distribution of data to understand patterns and identify outliers

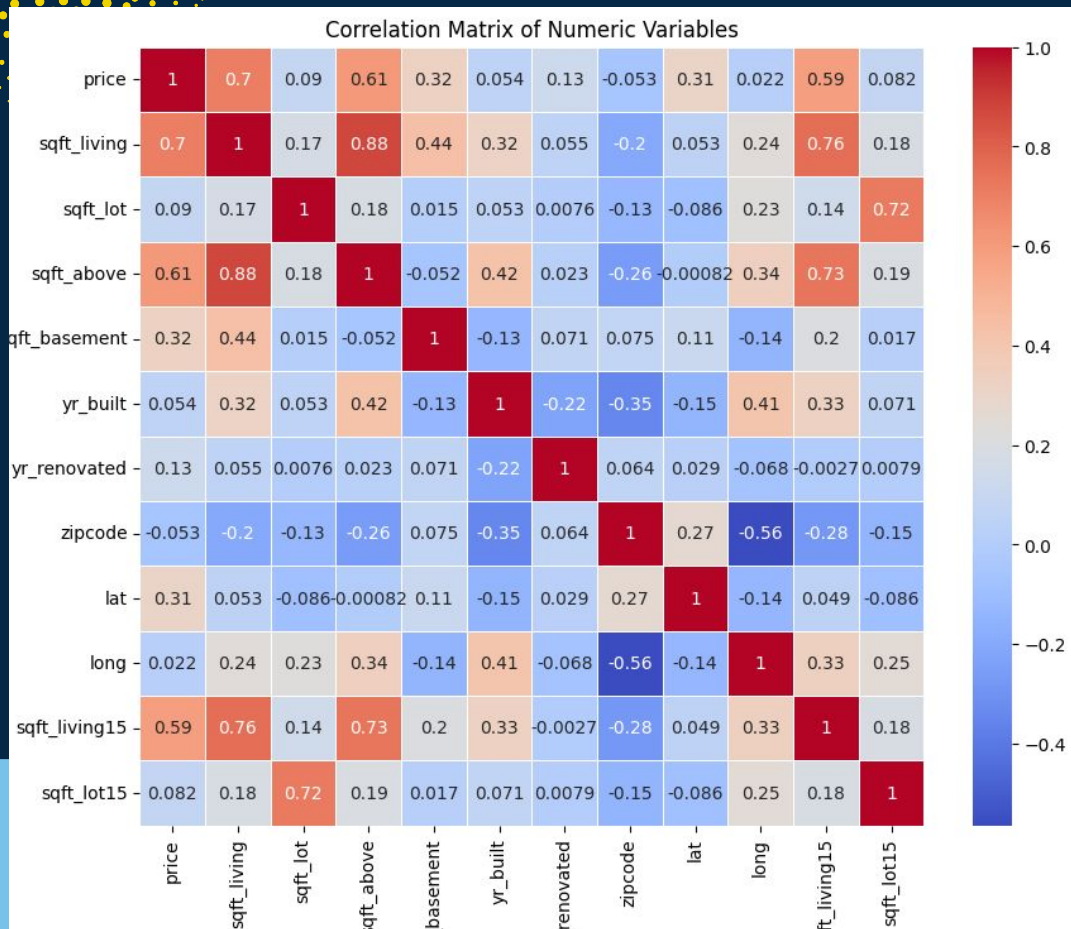


Distribution of Target Variable



This graph provides an overview of the distribution of property prices in the dataset.

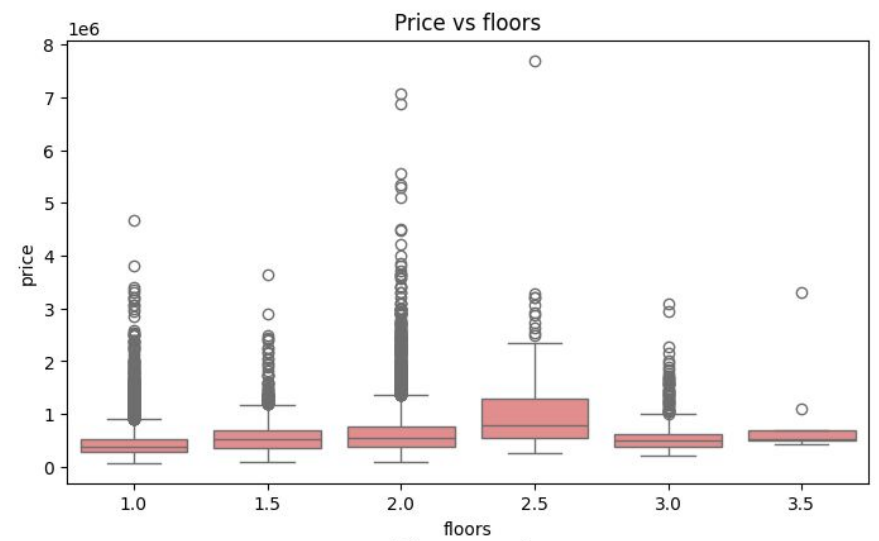
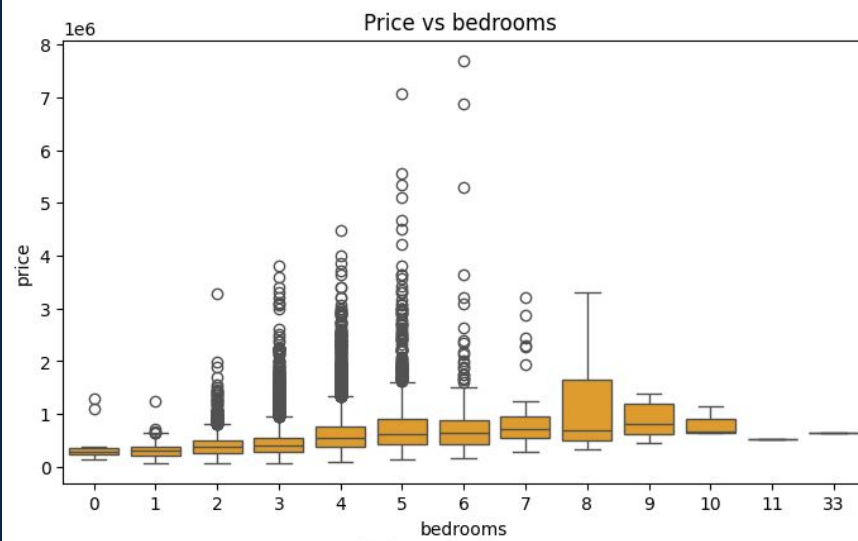
Correlation Matrix



We can see the relationship between variables quickly by looking at the correlation matrix on the side.

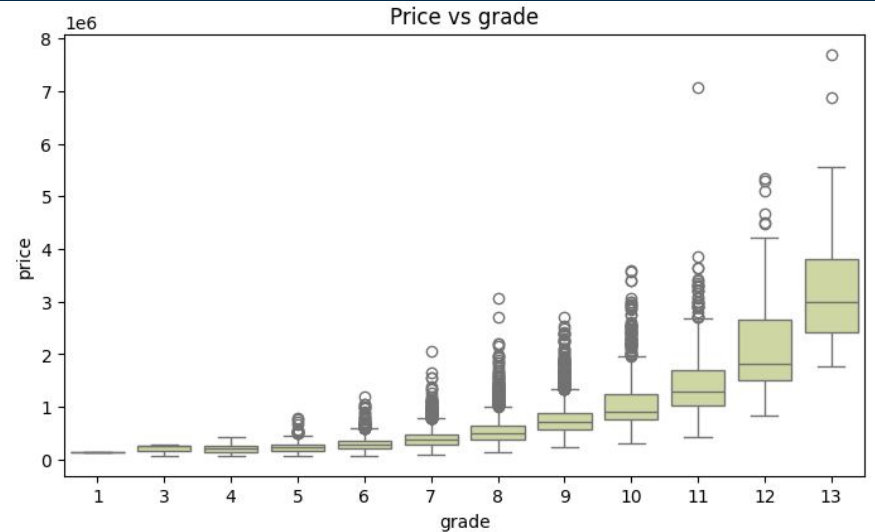
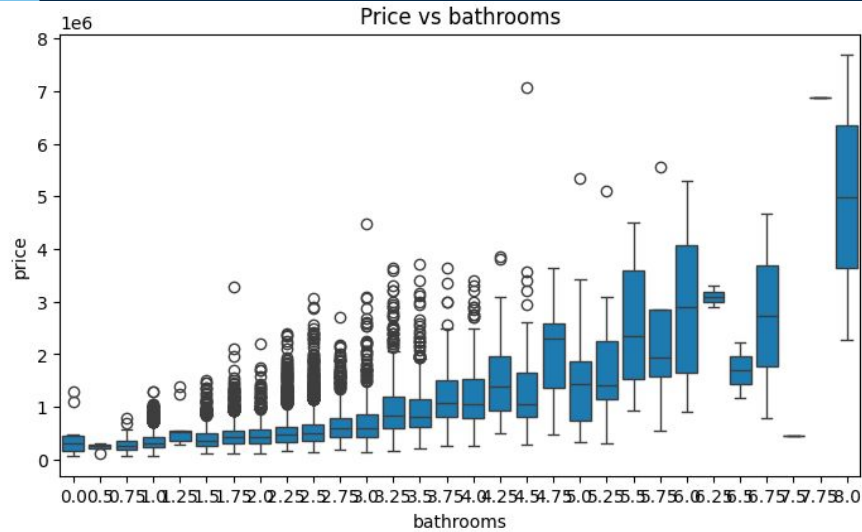
- "sqft_living" has the highest correlation to "price" with 0.7
- There is a small positive correlation between "lat" and "price" with 0.31, indicating that prices tend to increase slightly.

As we can see from the visualization below, the more the number of bathrooms, the higher the price of the house.



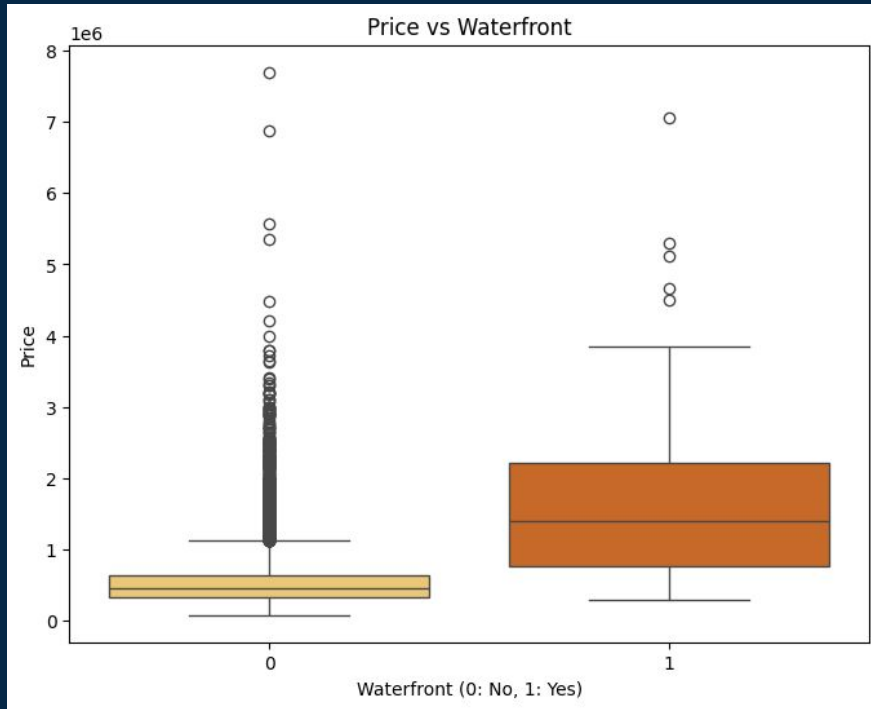
The visualization above shows that houses with 2.5 floors have a higher price.

The visualization below shows that the higher the grade of the house, the higher the price.

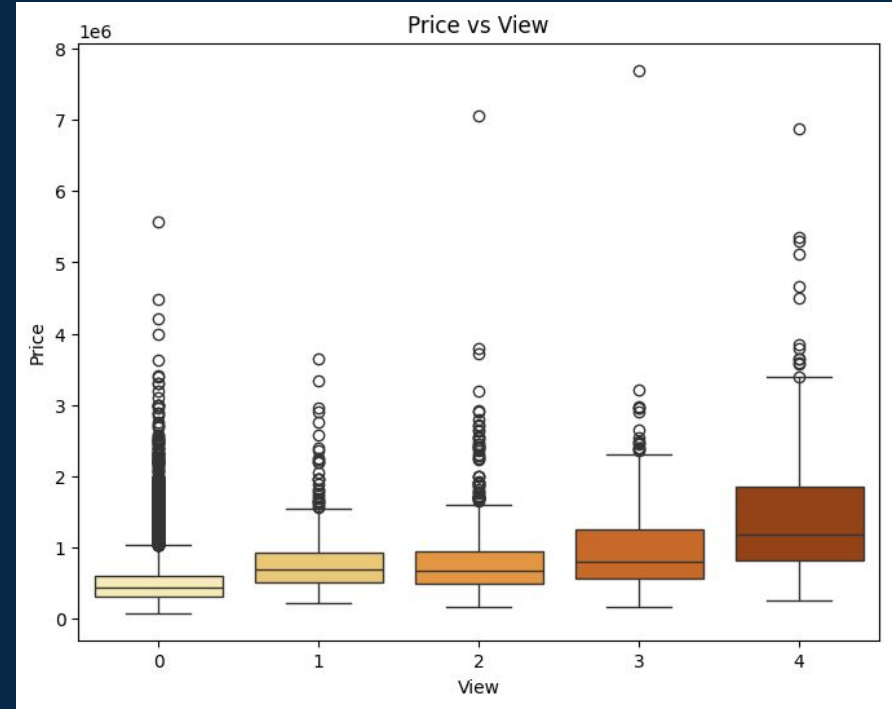


It can be seen above that the more bathrooms a house has, the higher the price of the house.

From the "Price vs View" boxplot, it is clear that properties with a better view tend to have higher prices



The comparison between properties with a waterfront view (Waterfront = 1) and those without a waterfront view (Waterfront = 0)



This scatter plot visualizes the association between the above-ground living area (sqft_above) and house prices



The scatter plot illustrates the relationship between the land area (sqft_lot) and house prices.

Missing Value and Duplicated Data



0

Missing Values



0

Duplicate Data

Data Type Conversion (type casting)

'bedrooms',

'bathrooms',

'floors',

'waterfront',

'view',

'condition',

'grade'



integer



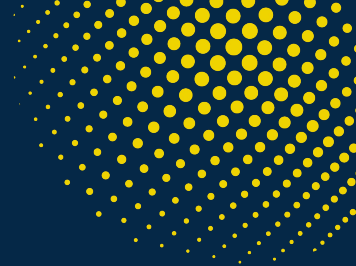
category

Modeling


01 *Linear Regression*

02 *Random Forest*


03 *Decision Tree*



Base Model Comparison – Regression



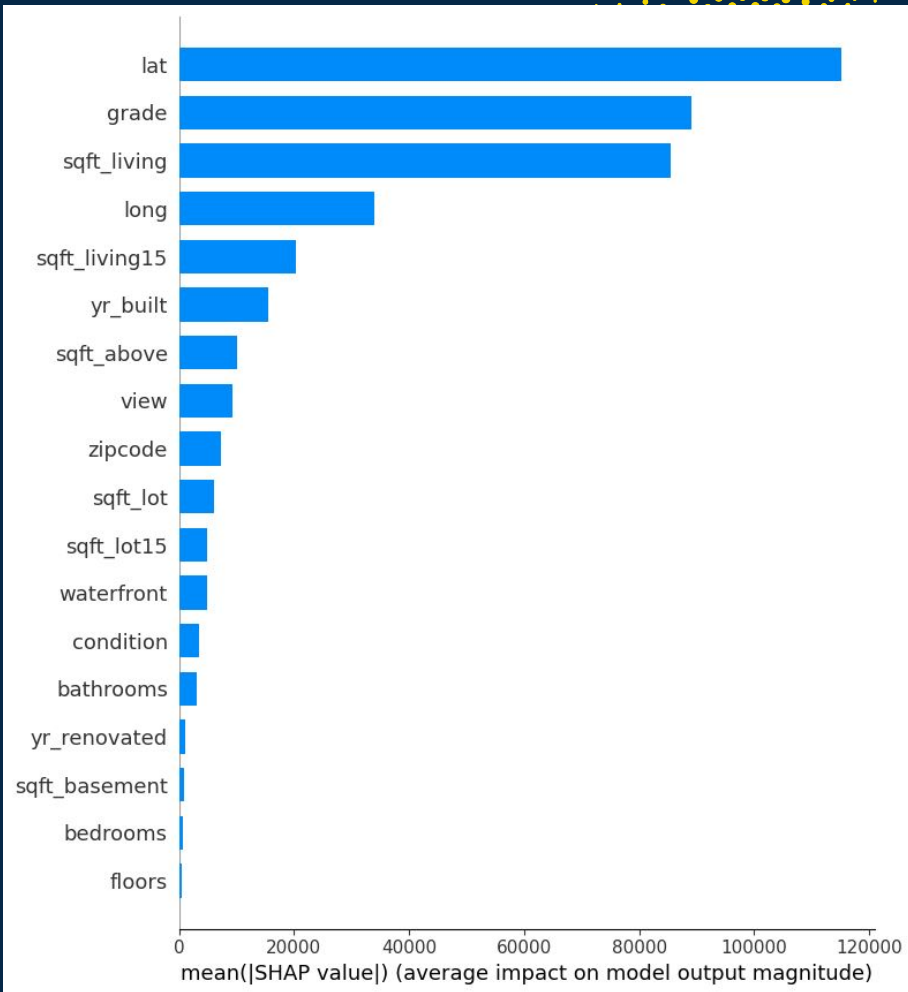
	MSE	MAE	RMSE	R-squared
Linear Regression	43306014067.04	128033.34	198876.059	0.7101177881517419
Random Forest	16658919793.42	69439.09	129069.43	0.8779034510579915
Decision Tree	38856338482.64	103711.80	197120.11	0.7152141380063048



Cross Validation & Hyperparameter Tuning (Random Forest)

```
'max_depth': 20, 'min_samples_leaf': 1,  
'min_samples_split': 5, 'n_estimators':  
    200
```

Model	MSE	MAE	RMSE	R-squared
Base Model	16658919793.42	69439.09	129069.43	0.8779034510579915
Tuned Model	17586082278.07	69474.10	129345.47	0.8773806393838899



SHAP VALUES

insight into how each feature (variable) contributes to the model's predictive results.

Conclusion

After going through the EDA phase, training multiple regression models, and performing hyperparameter tuning, the results indicate that the Random Forest model outperforms Linear Regression and Decision Tree. This is evidenced by the high R-squared value and other evaluation metrics.



The base Random Forest model showed promising results with a Mean Squared Error (MSE) of 16,658,919,793.42, Mean Absolute Error (MAE) of 69,439.09, Root Mean Squared Error (RMSE) of 129,069.43, and R-squared (R^2) of 0.8779. These metrics indicate a strong predictive performance

After hyperparameter tuning, the Random Forest model's performance slightly changed, resulting in a MSE of 17,586,082,278.07, MAE of 69,474.10, RMSE of 129,345.47, and R^2 of 0.8774. While there is a slight increase in MSE and RMSE, the R^2 remains high, indicating robust predictive capabilities

Recommendations

1. Model deployment

Considering the marginal change in performance after tuning, the base Random Forest model remains a solid choice for deployment due to its strong predictive capabilities

2. Feature Importance

Based on SHAP values, features such as lat, sqft_living, number of bedrooms, and grade rating have significant contributions to predicting house prices. Focus on these features in marketing or property improvements.

By following these recommendations, it is expected to maximize the value of this house price prediction model and provide maximum benefit to homeowners and stakeholders.



Recommendations with Business Impact Potential

1. The Random Forest model boasts high accuracy and predictive power, aiding businesses in making more precise estimations of property prices. This can reduce the risk of pricing errors and enhance customer confidence
2. Using insights from the model, businesses can create more personalized and tailored marketing strategies based on customer preferences. For example, focusing on properties with exclusive views or proximity to essential facilities.
3. The model can assist in determining more accurate pricing for new or renovated homes, enhancing competitiveness and minimizing the risk of customer loss due to mispricing
4. The model identifies key features influencing property prices, guiding businesses to strategically invest in renovations or upgrades that provide the highest return on investment. This ensures resources are allocated efficiently.



Thanks!

Check out the notebook on my GitHub!

<https://github.com/intaniarr/king-county-house>