

REGRESSION MODELING

To analyze house sales in a northwestern county



© 2019 COURTESY OF MICHAEL BAKER



Bijal Saija



Overview

This data science project aims to address a critical business problem for our stakeholders, a real estate agency specializing in helping homeowners buy and sell homes. The primary focus is on providing valuable insights and advice to homeowners regarding how home renovations may impact the estimated value of their properties. I have used regression modeling to analyze house sales in a northwestern county for this project.



Business Problem

The key business problem identified is the need to empower homeowners with actionable information about potential home renovations. The goal is to guide them in making informed decisions that can enhance the market value of their homes. By understanding the potential impact of specific renovations on the estimated property value, homeowners can prioritize and plan renovations strategically.




The Data

The data is taken from `kc_house_data.csv`

This data is the King County House Sales dataset.

Features used for analysis

Price (target variable)	
Sqft_living (continuous variable)	Condition (categorical variable)
Sqft_lot (continuous variable)	Bedrooms (categorical variable)
Yr_built (categorical variable)	Bathrooms (categorical variable)
Grade (categorical variable)	Sqft_above (categorical variable)



Multi linear regression

EQUATION :

$$Y = a + b_1X_1 + b_2X_2$$

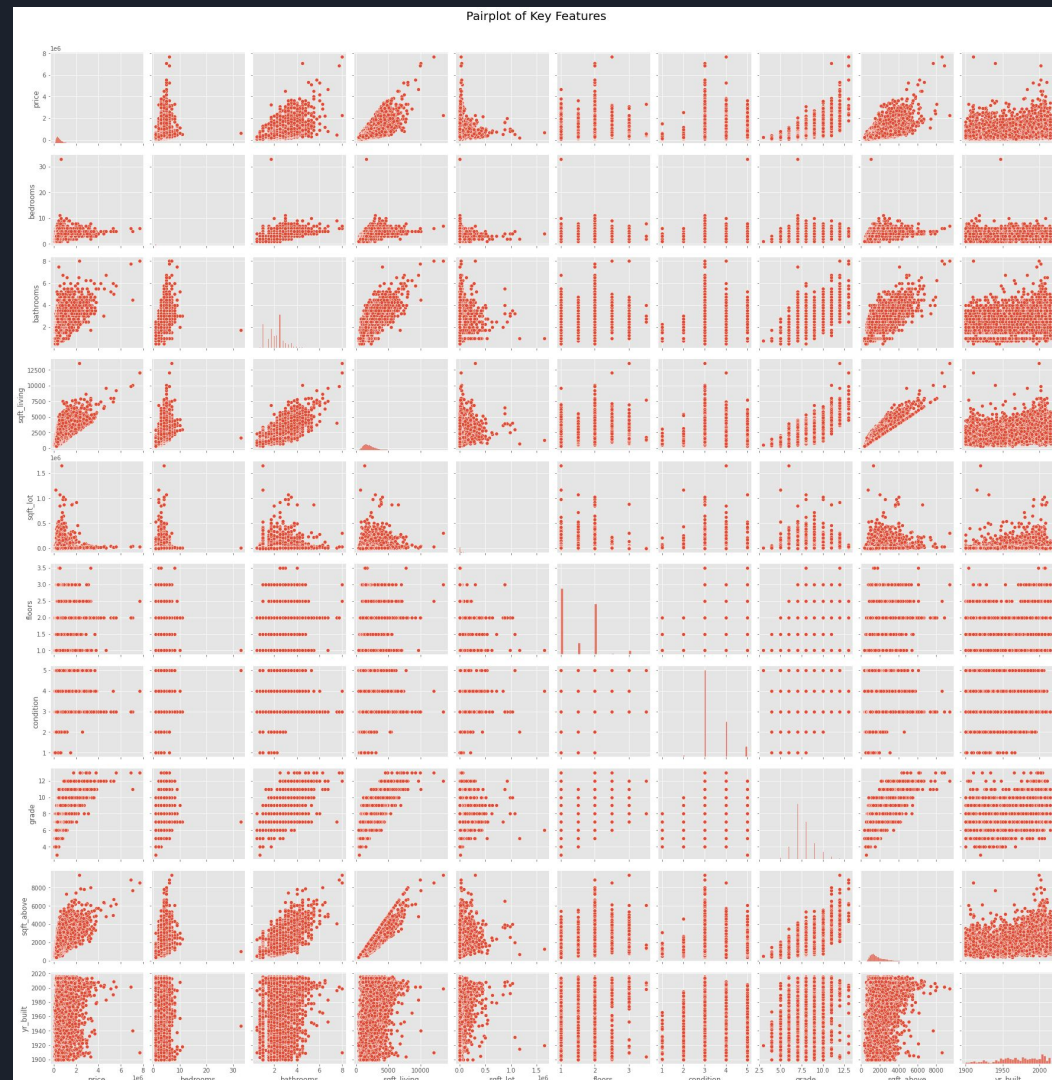
Where,

X_1, X_2, \dots, X_p are independent variables and Y is the target variable and b_1 is intercept and b_2 is Slope for Beta coefficient for X_2 , etc..

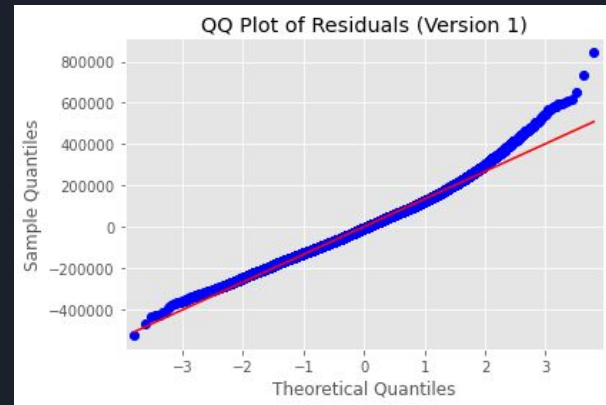
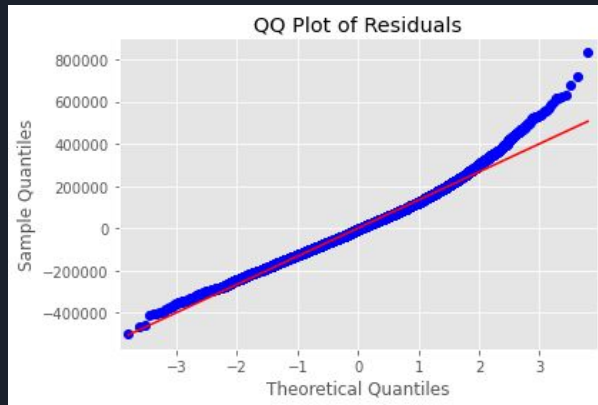
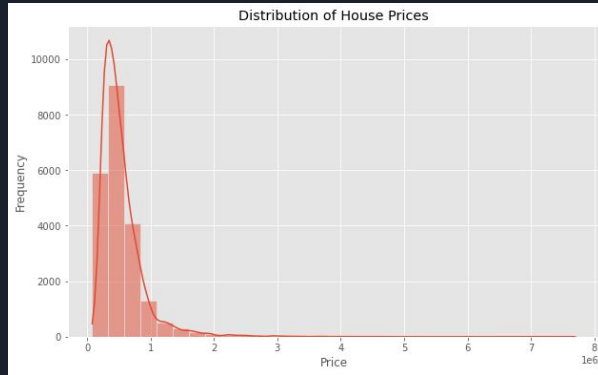


Exploratory Data Analysis

Exploratory Data Analysis involves the scatter plot outputs between price and independent variable.



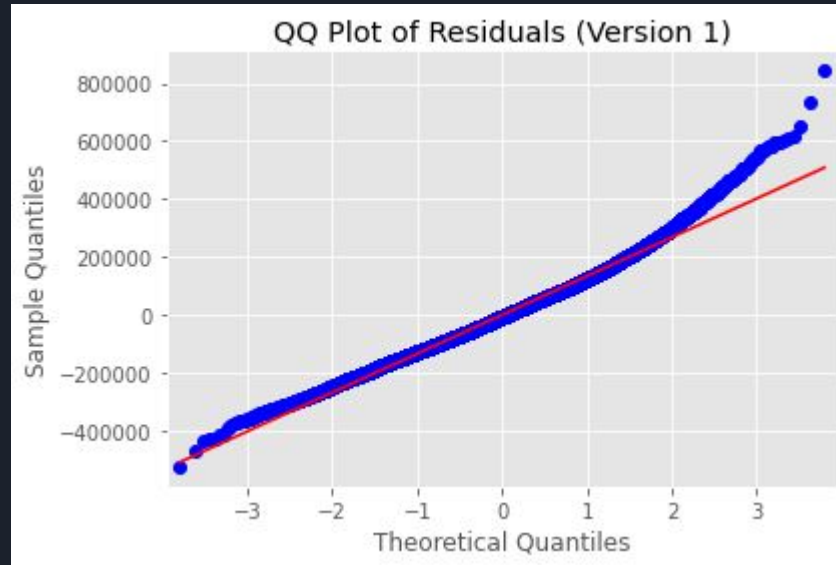
Distribution of price with and without log transformation



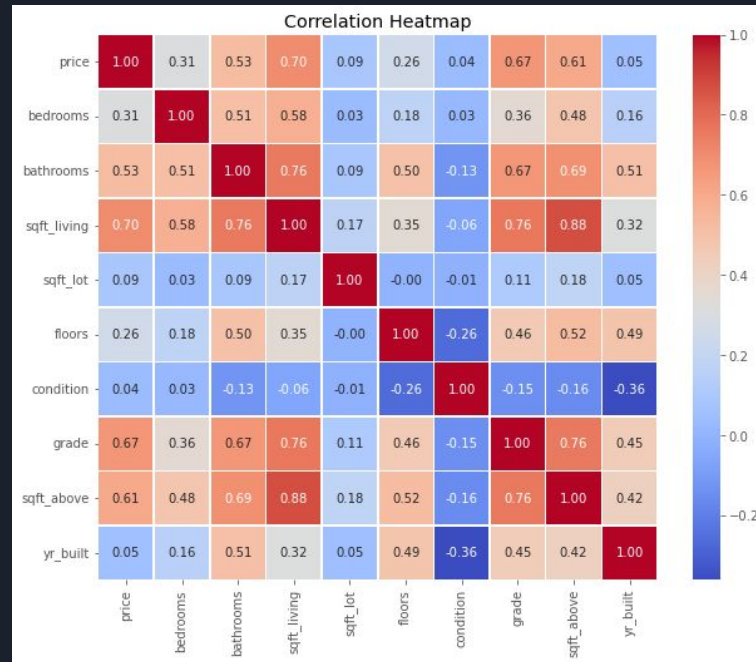
The housing price is transformed using natural log and appears very close to normal distribution.


This ensures linearity relationship b/w price and independent variables.

The distribution is not too much skewed after removing outliers.



In Hit map, we can clearly see that multicollinearity in between 2 independent continuous variable pair (sqft_living-sqft_above) is more than 0.7 that is why removing sqft_above column from the dataset.





R-squared: 0.509 This indicates the proportion of the variance in the dependent variable (price) that is predictable from the independent variables. In this case, approximately 50.9% of the variability in the home prices is explained by the model. Adjusted R-squared: 0.509 Similar to R-squared but adjusted for the number of predictors in the model. It provides a more accurate measure in the presence of multiple predictors. Observations: 13,604

Residuals: 13,553 Model: 18 Total: 13,572 F-statistic: 780.3

A measure of how well the overall model fits the data. A higher F-statistic suggests a better fit. Prob (F-statistic): 0.00

The p-value is associated with the F-statistic. A low p-value indicates that the model is statistically significant. Coefficients: Intercept (const): 4.545e+05 Coefficients for Predictors: sqft_living, sqft_lot, bedrooms, bathrooms, floors, grade, condition, yr_built Each coefficient represents the change in the dependent variable (price) per one-unit change in the respective independent variable, holding other variables constant. Statistical Significance: t-statistic and $P > |t|$: Indicates the statistical significance of each coefficient. The lower the p-value, the more significant the predictor. In this case, all predictors seem to be statistically significant (p-value < 0.05). Model Fit: AIC (Akaike Information Criterion): 3.e+05 BIC (Bayesian Information Criterion): 3.592e+05 Information criteria that penalize models for complexity. Lower values indicate a better fit. Residuals: Omnibus: 713.538 Durbin-Watson: 1.998 Jarque-Bera (JB): 1027.241 Tests for normality and autocorrelation of residuals. A low Durbin-Watson suggests potential autocorrelation. Notes: Standard Errors: Assume that the covariance matrix of the errors is correctly specified. This summary provides a comprehensive overview of the linear regression model, its fit to the data, and the statistical significance of each predictor. It suggests that the model explains a substantial portion of the variance in home prices, and the included predictors are statistically significant.



Top 3 features to predict the price

Sqft_living, grade and yr_built

yr_built_group_1971-1990

sqft_living

yr_built_group_1991-2010

grade_8

grade_9

Because F-statistics of this features are more than 0.05