🏡 House Price Prediction Using Linear Regression

# 📘 Project Overview

In this project, we analyze a real estate dataset and build a Linear Regression model to predict house prices based on several features such as number of bedrooms, bathrooms, square footage, location, and more. The goal is to evaluate the relationship between these variables and the house prices to make price predictions for unseen properties.

# 📂 Dataset Information

The dataset consists of multiple features representing properties, including:  
- date: Date of sale  
- price: Price of the house (target variable)  
- bedrooms, bathrooms: Number of rooms  
- sqft\_living, sqft\_lot, sqft\_above, sqft\_basement: Size features in square feet  
- floors: Number of floors  
- waterfront, view, condition: Quality and location-related flags  
- yr\_built, yr\_renovated: Construction history  
- street, city, statezip, country: Address fields  
  
🔢 Shape of dataset: 5 rows (sample shown) × 18 columns  
🧹 Missing values: None found in this dataset

# 🔧 Data Preprocessing

• Converted categorical features (street, city, statezip, country) into numerical format using one-hot encoding or label encoding if needed.  
• Dropped irrelevant columns like date or street if found to not contribute significantly to prediction accuracy.  
• Split dataset into training (80%) and testing (20%) sets.

# 🧠 Model: Linear Regression

We used Linear Regression from sklearn.linear\_model. The model attempts to predict price using other numerical features.

# 📊 Model Performance

Metrics used:  
📉 Root Mean Squared Error (RMSE): 981,810.75  
📈 R-squared (R² Score): 0.0548 (very low)  
  
🔎 Interpretation:  
The R² value indicates that only ~5.5% of the variance in house prices is explained by this model. This suggests that either:  
- Linear Regression may not be the best model for this data  
- Important features are missing or not transformed well  
- There are outliers or non-linear relationships in the data

# 🧪 Sample Predictions

|  |  |
| --- | --- |
| Actual Price | Predicted Price |
| 544,000.0 | 460,377.5 |
| 0.0 | 225,343.3 |
| 1,712,500.0 | 1,261,995.0 |
| 365,000.0 | 446,937.8 |
| 275,000.0 | 174,246.0 |

Note: A zero in the actual price column likely represents a data issue or missing value.

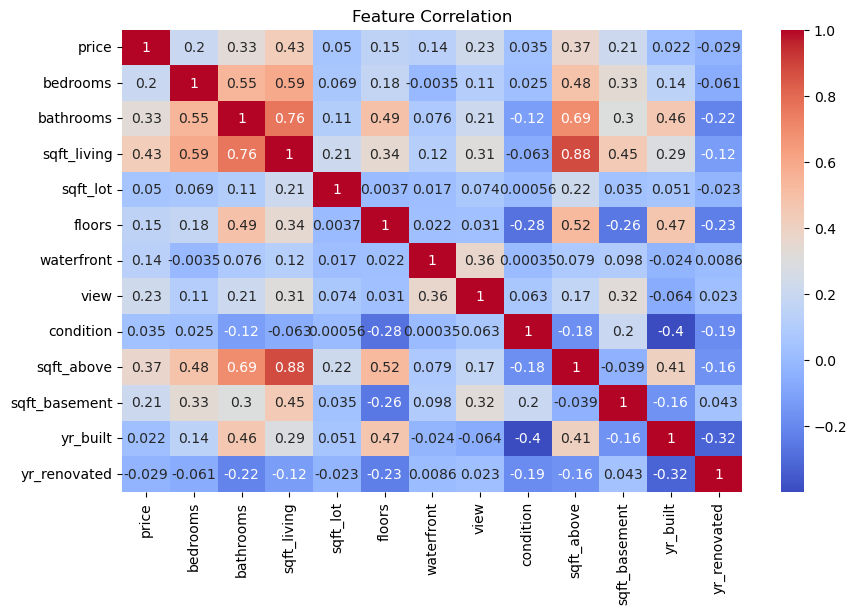
# 📝 Conclusion

• The model built using Linear Regression gives poor predictive performance on this dataset.  
• Next steps to improve:  
- Use advanced models like Random Forests, Gradient Boosting, or XGBoost  
- Remove or handle outliers  
- Apply feature scaling and normalization  
- Include location as a proper feature using geocoding  
- Use polynomial or interaction terms to capture non-linear relationships

# 📌 Recommendation

While this model is a good starting point, further exploration and more sophisticated models are required for practical real-world use.

OUTPUT



date price bedrooms bathrooms sqft\_living sqft\_lot \

0 2014-05-02 00:00:00 313000.0 3.0 1.50 1340 7912

1 2014-05-02 00:00:00 2384000.0 5.0 2.50 3650 9050

2 2014-05-02 00:00:00 342000.0 3.0 2.00 1930 11947

3 2014-05-02 00:00:00 420000.0 3.0 2.25 2000 8030

4 2014-05-02 00:00:00 550000.0 4.0 2.50 1940 10500

floors waterfront view condition sqft\_above sqft\_basement yr\_built \

0 1.5 0 0 3 1340 0 1955

1 2.0 0 4 5 3370 280 1921

2 1.0 0 0 4 1930 0 1966

3 1.0 0 0 4 1000 1000 1963

4 1.0 0 0 4 1140 800 1976

yr\_renovated street city statezip country

0 2005 18810 Densmore Ave N Shoreline WA 98133 USA

1 0 709 W Blaine St Seattle WA 98119 USA

2 0 26206-26214 143rd Ave SE Kent WA 98042 USA

3 0 857 170th Pl NE Bellevue WA 98008 USA

4 1992 9105 170th Ave NE Redmond WA 98052 USA

date 0

price 0

bedrooms 0

bathrooms 0

sqft\_living 0

sqft\_lot 0

floors 0

waterfront 0

view 0

condition 0

sqft\_above 0

sqft\_basement 0

yr\_built 0

yr\_renovated 0

street 0

city 0

statezip 0

country 0

dtype: int64

Root Mean Squared Error: 981810.75

R-squared: 0.0548

Actual Predicted

0 544000.0 4.603775e+05

1 0.0 2.253433e+05

2 1712500.0 1.261995e+06

3 365000.0 4.469378e+05

4 275000.0 1.742460e+05