ReDI School of Digital Integration

# WELCOME

We use tech to connect human potential and opportunity with dignity & humility

ReDI School of Digital Integration

# ML House Price Prediction

Predict residential house prices using machine learning models by leveraging historical property and feature data.

We use tech to connect human potential and opportunity with dignity & humility

**Syed Hussnain Haider Kazmi**
**Contact:** hussnain2k13@gmail.com

ReDI School of Digital Integration

# ML House Price Prediction

Motivation: Why this project/ Idea / Dataset?

**Syed Hussnain Haider Kazmi**
**Contact:** hussnain2k13@gmail.com

We use tech to connect human potential and opportunity with dignity & humility

# Topics Covered

The project encompassed the following major topics:

1. **Problem Definition and Dataset**
2. **Methodology**
   - Part 1 – Data Preparation
   - Part 2 – Modeling
3. **Exploratory Data Analysis (EDA) – Correlation Heatmap**
4. **Model Evaluation & Metrics**
5. **Model Comparison & Best Model Selection**
6. **Ethical Considerations and Limitations**
7. **Conclusion and Future Work**

# 1. Problem Definition and Dataset

## Purpose

- Predict house prices from property features (size, location, condition)
- Support buyers, sellers, and investors with data-driven insights

## Problem & Objectives

- Pricing influenced by multiple factors; traditional methods often miss patterns
- **Objectives:**
  - Preprocess and analyze dataset
  - Train & evaluate regression models (Linear, Decision Tree, Random Forest, Gradient Boosting, XGBoost)
  - Compare performance (MAE, RMSE, $R^2$) and select best model

## Dataset

- **Source:** Kaggle House Price dataset (4600 rows)
- **Why:** Rich features, real-world data, suitable for regression

## Key Features

*price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, yr_built, yr_renovated, city, statezip*

# 2. Methodology – Part 1 (Data Preparation)

## Step 1: Data Cleaning & Preprocessing

- Checked for missing values and duplicates → none found
- Handled outliers using IQR method for *'price'* and *'sqft_lot'*
- Separated features *(X)* and target *(price)*

## Step 2: Feature Engineering

- Created **new features**
  - House age = year_sold – yr_built
  - Has_been_renovated (0/1)
- Dropped irrelevant/redundant columns: *city, statezip, condition, view, waterfront*
- Kept *city* for location-based prediction

## Step 3: Encoding & Scaling

- One-Hot Encoding for categorical features: *city, statezip, condition, view, waterfront*
- StandardScaler for numerical features

# 2. Methodology – Part 2 (Modeling)

## Step 4: Train-Test Split

- 80% training, 20% testing
- Ensured reproducibility with *random_state=42*

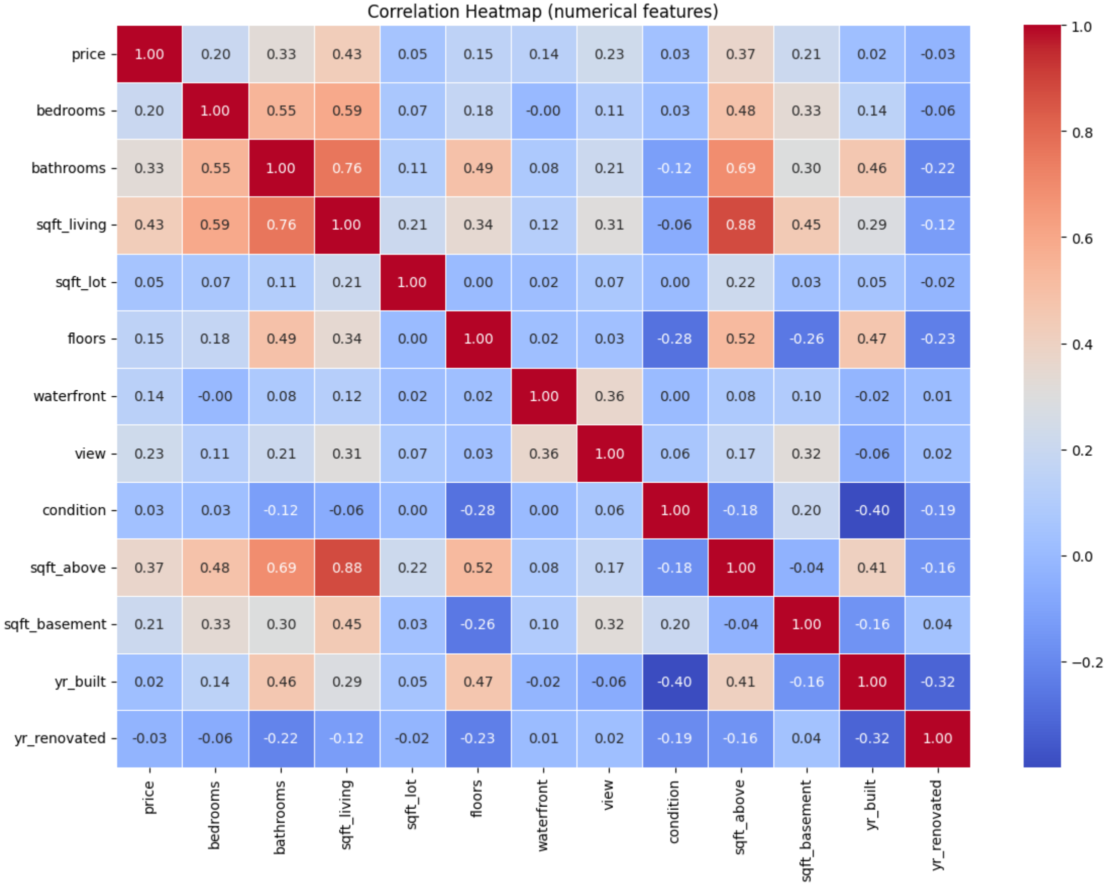
## Step 5: Model Development

- Trained 5 regression models:
    - Linear Regression
    - Decision Tree
    - Random Forest
    - Gradient Boosting
    - XGBoost
- Used pipelines for consistent preprocessing

## Step 6: Evaluation and Selection

- Metrics: **MAE**, **RMSE**, **$R^2$**
- Compared models using evaluation metrics and $R^2$ scores
- XGBoost selected as **best-performing model**

# 3. Exploratory Data Analysis *[Correlation Heatmap]*



# 4. Model Evaluation & Metrics

## Evaluation Metrics Used

- **MAE (Mean Absolute Error):** Average absolute difference between predicted & actual prices
- **RMSE (Root Mean Squared Error):** Penalizes larger errors more than MAE
- **$R^2$ Score:** Explains variance captured by the model (closer to 1 → better)

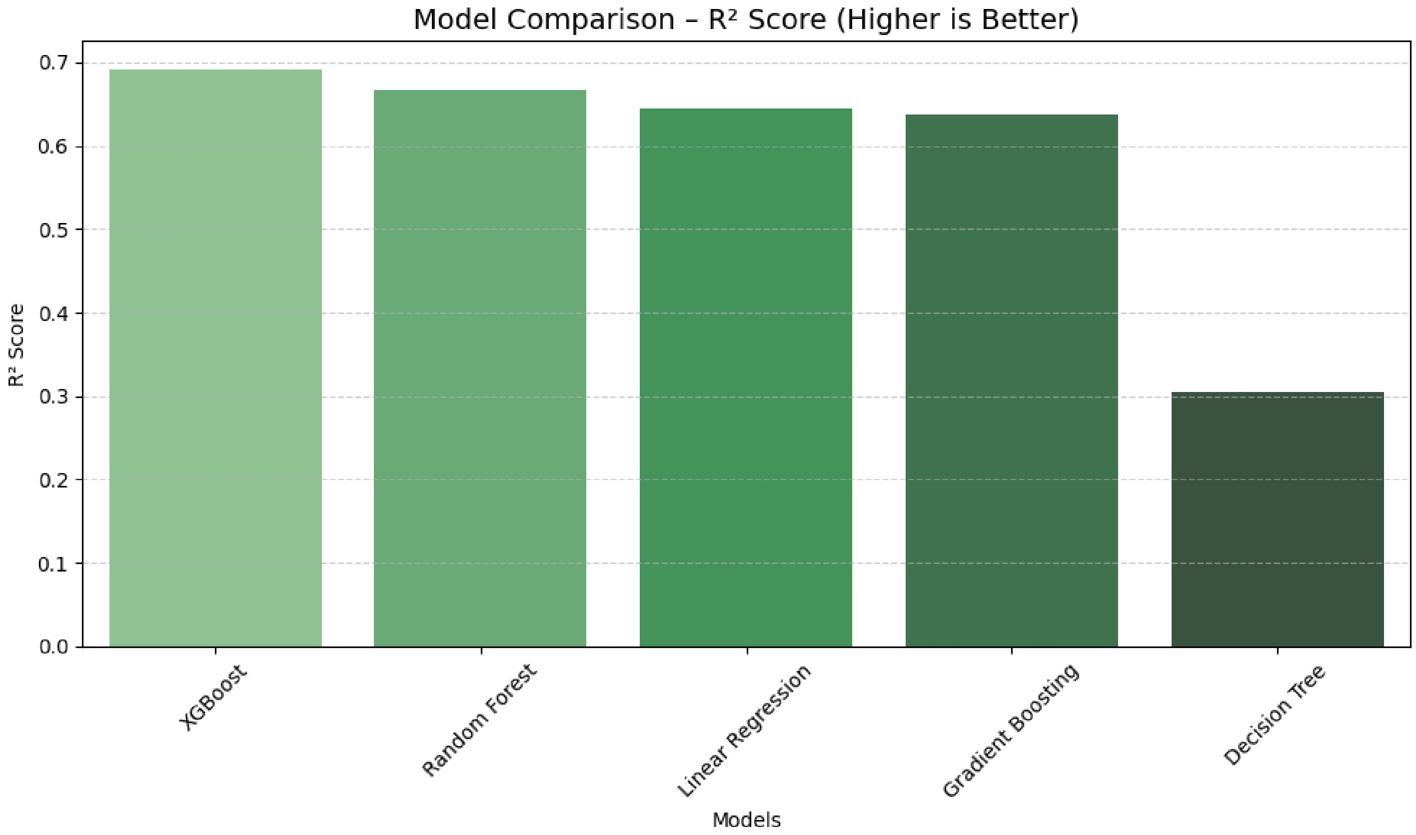
## Model Performance Overview

- Trained 5 regression models on pre-processed data
- Predicted prices on test set and evaluated metrics
- Recorded all results in a summary table for comparison

# 4. Model Evaluation & Metrics

| MODEL | MAE | RMSE | $R^2$ |
|---|---|---|---|
| XGBOOST | 76622.215760 | 120005.577660 | 0.690844 |
| RANDOM FOREST | 80753.549907 | 124429.664012 | 0.667630 |
| LINEAR REGRESSION | 77733.731146 | 128540.436437 | 0.645306 |
| GRADIENT BOOSTING | 89127.342755 | 129807.674638 | 0.638278 |
| DECISION TREE | 110962.330760 | 180009.877922 | 0.304388 |

# 5. Model Comparison & Best Model Selection



# 5. Model Comparison & Best Model Selection

## Model Comparison Highlights

- XGBoost: **Best $R^2$ (0.69)**, lowest MAE & RMSE → most accurate predictions
- Random Forest: Good performance, slightly lower $R^2$
- Linear Regression: Moderate performance
- Decision Tree: Lowest $R^2$ → underfitting observed

## Best Model Selection

- **XGBoost chosen** as final model for predictions
- Reasons:
  - Handles numerical & categorical features efficiently
  - Robust to outliers and complex patterns
  - Consistently highest $R^2$ & lowest errors

# 6. Ethical Considerations and Key Limitations

## Ethical Considerations

- **Location Bias:** Model may favor high-income areas.
- **Historical Inequities:** Old building/renovation patterns influence prices.
- **Sampling Bias:** Data from one region only.
- **Transparent Processing:** No personal data; clear documentation.

## Key Limitations

- **Limited Scope:** Single region, single time period.
- **Missing Factors:** School quality, crime, interior condition not included.
- **Market Variability:** Static data cannot capture rapid trends.

# 7. Conclusion & Future Work

## Conclusion

- **Strong Workflow:** Completed full ML pipeline from cleaning to modeling.
- **Best Model:** XGBoost delivered highest accuracy and lowest errors.
- **Useful Features:** Engineered features improved prediction quality significantly.

## Future Improvements

- **Tuning Models:** Apply grid/random search to boost performance.
- **More Features:** Add neighbourhood, school, or macroeconomic indicators.
- **Explainability:** Use SHAP to understand feature impact clearly.

# Bibliography


- ❑ House Price Prediction Dataset
  https://www.kaggle.com/datasets/shree1992/housedata

- ❑ GitHub link to the project repository
  https://github.com/SyedHussnainHaiderKazmi/ML-House-Price-Prediction

# Thanks a lot!


**Contact**

Syed Hussnain Haider Kazmi
Student (Machine Learning – Online) ReDI School Munich

hussnain2k13@gmail.com