# House Price Prediction

# Introduction & Motivation

The real estate market is influenced by various factors such as location, size, and market trends. Accurate house price prediction is essential for stakeholders like homeowners, buyers, and real estate agents.

Real estate valuation is often unpredictable. This project aims to simplify the valuation process using machine learning (ML) to deliver accurate, data-driven predictions and uncover market insights.

# Methodology

## Data Collection

- Dataset sourced from Kaggle.
- Features include house size, location, number of rooms, garage availability, and condition.

## Data Preprocessing

- **Handling Missing Values**: Imputation techniques applied.
- **Feature Encoding**: Categorical variables transformed into numerical values.
- **Outlier Detection**: Boxplots and Z-scores used to handle outliers.
- **Feature Scaling**: Standardized numerical features.

## Pipeline Steps

- **EDA**: Correlation and distribution analysis performed.
- **Feature Engineering**: Created derived features such as Location Ratio.
- **Model Training**: Linear regression with an 80-20 train-test split.
- **Evaluation**: $R^2$ score and MAE assessed using Scikit-learn.

## Tools Used

- Python libraries: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.
- IDE: Google Colab

# Results, and Discussion

## Experimental Setup

- Dataset includes factors like Price, Location, Area (sqft), Bedrooms, and Condition.
- Preprocessing ensures high-quality inputs for modeling.

## Results

- **R² Score**: 0.85 (85% variance explained).
- **MAE**: $15,000 (average prediction error).
- Strong positive correlation observed between Area and Price.

## Visual Insights

- Price distribution analysis shows skewness due to high-priced outliers.
- Residual plot validates model's predictive strength.
- Feature importance highlights Area, Bedrooms, and Location as key predictors.

## Discussion

- Linear regression provides reliable results but has limitations.
- Advanced models like Gradient Boosting could improve performance.
- Missing features like economic trends and neighborhood indicators limit accuracy.

# Related Work

**Machine Learning in Real Estate Pricing**: Research highlights the effectiveness of decision tree regression and linear regression.

**Feature Selection with Lasso Regression**: Lasso regression simplifies models by eliminating irrelevant features.

**Deep Learning Applications**: Multimodal ML methods incorporate textual, visual, and numerical data for advanced prediction accuracy.

# Future Work & Conclusion

## Future Work

- Employ advanced models (e.g., Random Forest, Gradient Boosting).
- Incorporate additional features like economic indicators and crime rates.
- Develop interactive dashboards for real-time predictions.
- Explore time-series analysis to capture market trends.

## Conclusion

- Linear regression achieved strong predictive performance ($R^2$ = 0.85, MAE = $15,000).
- Key predictors: Area (sqft), Bedrooms, and Location.
- Improvements needed for handling outliers and capturing non-linear relationships.