🏡 House Pricing Prediction Project Report

📋 Project Overview

This project focuses on analyzing and predicting house prices using real estate data. The dataset includes various attributes such as location, area, number of rooms, furnishing status, and more. The objective is to understand the key factors influencing property prices and build a machine learning model to predict house prices effectively.

🧩 1. Data Cleaning

Steps Performed:

- Removed duplicate and missing values.
- Standardized inconsistent text entries (e.g., city names, payment options).
- Converted date columns to datetime format.
- Removed irrelevant or redundant columns like Compound and Developer.
- Ensured correct data types for numerical and categorical variables.

Result:
The cleaned dataset contained 25,433 records with important features such as Price, Area, Bedrooms, Bathrooms, City, and more.

🔍 2. Exploratory Data Analysis (EDA)

Key Visualizations and Insights:

- **Price Distribution:** Prices were right-skewed, indicating the presence of high-value luxury properties.
- **Area vs. Price:** A positive correlation — larger properties generally had higher prices.
- **Bedrooms/Bathrooms vs. Price:** More rooms generally increased property price, though with some variability.
- **City-wise Average Prices:** Certain cities had significantly higher average prices than others.
- **Correlation Heatmap:** Strong correlations were observed between Area, Price, and Log_Price.

Summary:
EDA revealed that property size and location are the most influential factors affecting housing prices. Outliers were retained, as they represent valid luxury or large properties in the real estate market.

⚙️ 3. Feature Engineering

Features Added:

- price_per_m2 = Price / Area
- total_rooms = Bedrooms + Bathrooms

- Log transformations: Log_Price, Log_Area
- Label encoding for categorical variables: City, Type, Furnished, Delivery_Term

Features Dropped:

- Compound, Developer (due to high cardinality and missing data)

Result:
The dataset was transformed into a model-ready format with enhanced interpretability and reduced noise.

🤖 4. Machine Learning Modeling

Models Used:

- Linear Regression
- Ridge Regression
- Random Forest Regressor
- Gradient Boosting Regressor

Model Evaluation Metrics:

- **R^2 (Coefficient of Determination):** Model accuracy
- **RMSE (Root Mean Squared Error):** Average model error
- **MAE (Mean Absolute Error):** Mean absolute difference between predicted and actual prices

Model	**R^2**	**RMSE**	**MAE**
Gradient Boosting	0.9998	86,932.14	34,919.53
Random Forest	0.9990	182,126.29	4,338.30
Ridge Regression	0.8375	2,406,109.00	1,498,740.26
Linear Regression	0.8374	2,406,135.00	1,498,759.06

Conclusion:
Gradient Boosting performed the best, achieving near-perfect accuracy ($R^2 \approx 1.0$). Ensemble models significantly outperformed simple linear models.

📈 5. Final Insights

- Property area, location, and total rooms are the strongest predictors of house price.
- Outliers represent valid high-end properties and were therefore retained.
- Gradient Boosting is the optimal choice for predicting house prices in this dataset.
- The model shows exceptional precision in capturing complex non-linear patterns between variables, indicating that ensemble techniques handle real estate pricing data better than simple regressions.

Conclusion

This project demonstrates the full pipeline of a data science workflow — from data cleaning and exploration to feature engineering and model evaluation. The developed Gradient Boosting model can accurately predict housing prices, making it a strong baseline for future real estate predictive analytics.

Through careful consideration and execution of each step in the process, I've created a robust model that not only meets our initial goals but also sets the stage for further improvements and applications in the field of real estate analytics. This comprehensive approach ensures that the insights gained from the data are both meaningful and actionable, paving the way for future advancements in predictive analytics.