🏗️ Project Summary: Nairobi Real Estate Price Prediction

Objective: To build a high-performance machine learning model capable of predicting residential property prices in Nairobi, utilizing categorical signals and numerical features.

📊 Performance Metrics

- **Best Model:** Random Forest Regressor
- **Mean Absolute Percentage Error (MAPE): 22.66%**
- **Robustness:** The model achieved an **Average Cross-Validation R^2 of 0.82**, proving it is stable across different subsets of data.

🔍 Key Drivers of Price (Updated Findings)

Based on our latest **Feature Importance** analysis, the model identifies value through these primary signals:

1. **Property Type (Townhouse):** The strongest single predictor of price. In the Nairobi market, a "Townhouse" designation acts as a major price "multiplier" compared to other types.
2. **Bedroom Count:** Surpassed House Size in importance, indicating that the utility and "room count" are cleaner price signals for the model than raw square footage.
3. **Specific Location Premiums:** Locations such as **Runda**, **Kiambu Road**, and **Lavington** emerged as top 10 drivers, showing that "Location" isn't just a label, but a direct contributor to millions in added value.

📈 Diagnostic Insights

- **Model Bias:** Our **Distribution of Errors** shows a high density centered at **0**, confirming that the model is unbiased and highly accurate for the majority of mid-range properties.
- **The "Luxury Gap":** The presence of "Fat Tails" in the error distribution (misses exceeding 40M KES) highlights that ultra-luxury properties in premium zones follow unique pricing logic that standard features cannot fully capture yet.

🛠️ Technical Implementation Highlights

- **Modular Preprocessing:** Used a ColumnTransformer to handle standardized scaling for numbers and One-Hot Encoding for 50+ unique Nairobi neighborhoods.
- **Outlier Management:** Applied np.log1p transformation to the target variable. This successfully compressed the price scale, allowing the model to learn from both modest apartments and multi-million KES estates simultaneously.
- **Frequency Thresholding:** By removing locations with fewer than 3 listings, we improved the model's reliability and prevented it from "hallucinating" prices for areas with insufficient data.

🚀 Future Recommendations

To drop the error rate below 15% and address the luxury outliers:

- **Cluster Analysis:** Group locations by "economic zones" rather than just names.
- **Advanced Boosting:** Implement **XGBoost** or **LightGBM** to better capture the non-linear relationship between "Runda" and "Townhouse."
- **Feature Expansion:** Scrape data on "Year of Construction" to account for the depreciation of older builds versus the premium of new developments.
-