Final project proposal - group 10

Dataset

- source & overview
  - This dataset is from the Kaggle competition: *London House Price Prediction – Advanced Techniques*.
  - It has 266324 samples for training set, and 16547 samples for testing set.
  - It contains records describing residential property transactions in London, including structural, geographic, and temporal characteristics. The goal is to predict property sale prices based on these attributes.
- Feature Description
  - The dataset consists of **15 features**, which can be further categorized to:
    - **Numerical features:** bathrooms, bedrooms, livingRooms, floorAreaSqM, currentEnergyRating.
    - **Categorical features:** postcode, outcode, country, tenure, propertyType.
    - **Temporal/spatial features:** latitude, longitude, sale_month, sale_year.
    - **Textual feature:** fullAddress.
  - Some columns contain missing values and skewed distributions; preprocessing will be needed to reduce skewness.

Task

- This project is formulated as a **supervised regression** problem:
  - **Input (X):** Property-level features (numerical, categorical, temporal, spatial).
  - **Output (Y):** Continuous variable — *sale price*.
  - **Evaluation metrics:** MSE, RMSE, MAE, and R^2 to measure both error magnitude and explanatory fit.
  - **Challenges:** Missing data, spatial heterogeneity, and multicollinearity between related variables (e.g., bedrooms vs floor area).
  - **Objective:** Accurately predict London property prices while maintaining model interpretability and computational efficiency.
- Here are the methods we plan to apply. According to our progress and findings, these choices may be adjusted.

category	model	purpose
Baseline	Linear Regression, Ridge/Lasso	Simple, interpretable baseline
Tree-based	Decision Tree, Random Forest	Capture non-linear relationships
Advanced	XGBoost, LightGBM	Advanced approach
Optional	Simple MLP	Comparison to neural network

Analysis and Discussion (Tentative)

If time permits, we plan to conduct the following analyses:

1. **Exploratory Data Analysis (EDA):**
   - Correlation heatmap between features and price
   - Price distribution by property type, bedrooms, and region
   - Missing value and outlier visualization
2. **Model Results & Interpretation:**
   - Compare MAE and RMSE across models
   - Discuss which features influence price the most (via feature importance, SHAP)
   - Discuss key hyperparameters and their effects (e.g., learning rate, depth)