House Price Prediction Project — Concepts & Line■by■Line Explanation

1. New Concepts Learned

- Missing value handling (median imputation for numeric, placeholder for categorical)
- Rare category grouping (<1% frequency mapped to "other")
- Ordinal encoding for quality-related features
- OneHotEncoding with handle_unknown="ignore"
- ColumnTransformer for applying different transformations to numeric vs categorical columns
- Pipelines to chain preprocessing + model
- Cross■validation (5■fold RMSE)
- Debugging data leakage and mismatched encodings
- Model persistence using joblib
- Inference pipeline for test data

2. Techniques Used

- Feature engineering (mapping quality scores, rare-category grouping)
- Scaling numeric features using StandardScaler
- Model comparison: Linear Regression, Random Forest, Gradient Boosting
- RMSE computation and variance analysis
- Avoiding train-test leakage by fitting preprocessing only on training data
- Consistent preprocessing for inference

3. Project Code Explanation (High-Level Summary)

- Loading training data and analyzing missing values
- Converting quality strings using ordinal maps
- Grouping rare categories in all categorical features
- Separating numeric and categorical columns by dtype
- Creating numeric pipeline (median imputer + scaler)
- Creating categorical pipeline (constant imputer + one-hot encoder)

- Combining them with ColumnTransformer

- Training multiple models inside pipelines

- Running cross-validation and computing RMSE

- Debugging massive RMSE errors due to incorrect scaling and encoder issues

- Training final Gradient Boosting model and saving with joblib

- Loading test data and applying the same preprocessing

- Generating predictions and exporting submission.csv

This PDF summarizes the concepts, techniques, and logic behind each step of the House Price Prediction ML project.