Preprocessing and Model Performance Report

1. Preprocessing Summary

The preprocessing pipeline included the following steps:

1. Handling Missing Values:
- Filled missing compound names with 'No compound'.

2. Encoding Categorical Variables:
- Binary encoding for Yes/No columns.
- Mapping finishing types from ordinal categories.
- Mapping view types into numerical categories.

3. Removing Unwanted Columns:
- Dropped listing_date and days_on_market.

4. Outlier Removal:
- Applied IQR filtering per compound to remove price, area, and distance outliers.

5. Removing Duplicate Rows:
- Ensured dataset contains unique apartment listings.

2. Model Performance Comparison

Linear Regression

Train R^2: 0.7324

Train MAE: 395958.41

Test R^2: 0.7448

Test MAE: 409566.5

LightGBM

Train R^2: 0.8226

Train MAE: 320,372

Test R^2: 0.7371

Test MAE: 408,382

Random Forest

Train R^2: 0.8475

Train MAE: 295,927

Test R^2: 0.7100

Test MAE: 431,286

XGBoost

Train R^2: 0.7876

Train MAE: 351,902

Test R^2: 0.7436

Test MAE: 407,062

CatBoost

Train R^2: 0.8751

Train MAE: 260120.62

Test R^2: 0.8702

Test MAE: 284065.07