Analysis Plan for Project 2

1. Data Cleaning and Preprocessing

(1) **load and label the training data**: create 'neighborhood' and 'id' columns to identify the records.
(2) **Combined Variables:** `Exterior` and `LotInfo` are concatenated (e.g., “Inside;Reg;6120;51”). Split these into separate variables (`LotInfo` splits to `LotLocation`, `LotShape`, `LotArea`, `LotFrontage`; `Exterior` splits to `Exterior1st`, `ExterQual`, `ExterCond`).
(3) **one-hot encoding** for categorical features by 'pd.get_dummies'
(4) **Feature Alignment:** Since feature sets differ slightly across neighborhoods, align three datasets to a common set of features.
(5) **Missing Values:** use median imputation for numeric variables, mode for categorical to handle missing values.

2. Feature Engineering

Use linear regression to identify the in-significant variables. Analysis the Correlation matrix to remove multicollinear variables. Feature importance from tree-based models.

3. Dimension Reduction

explore PCA (Principal Component Analysis) if feature space becomes large and models suffer from overfitting or high variance. PCA may be applied after one-hot encoding and standardization.

4. Model Selection and Evaluation

plan to explore and compare the following models:

Model	
Linear Regression	Baseline model;
Ridge & Lasso	To handle multicollinearity and regularization
Random Forest	Handles non-linearities well, feature importance
XGBoost	Strong performance with boosting and fine-tuning
K-Nearest Neighbors	Simple, distance-based, sensitive to scaling

Models will be trained and tuned using **10-fold cross-validation** within training dataset.