**Project 1, PSL Fall 2024**

Olivia Dalglish and Arindam Saha

The goal of this project is to predict the sale price from homes in Ames, Iowa. The dataset contains 81 variables on the homes in the dataset, which will be further described below. This report outlines the key steps taken in preprocessing the dataset and implementing predictive models.

**1. Data Preprocessing**

**1.2 Handling Missing Values**

* **Categorical Features:** Missing values in categorical variables are imputed with the most frequent value (mode) in the respective columns. This strategy minimizes the impact of missing data by substituting it with the most common category, reducing the risk of introducing bias into the model.
* **Numerical Features:** Missing values in numerical columns are imputed with the median value for each column. The median is preferred over the mean to avoid sensitivity to outliers, which could skew the results if the data distribution is not symmetric.

**1.3 Feature Engineering**

New features are created to better capture relationships between variables:

* **Total\_SF:** This feature represents the total square footage, computed by summing the square footage of the first and second floors and the basement (First\_Flr\_SF, Second\_Flr\_SF, Total\_Bsmt\_SF). This helps to incorporate the size of the house as a key predictor.
* **Total\_Bath:** This feature aggregates the number of full and half bathrooms, with half bathrooms weighted at 0.5 to reflect their smaller size and value compared to full bathrooms. This feature provides a more accurate representation of the home's total bathroom capacity.

**1.4 Dropping Irrelevant Features**

Several features were excluded from the analysis that had low correlation with the final :

* **Target and Identifier:** The target variable Sale\_Price and the identifier column PID were excluded from the feature set. Sale\_Price is the variable being predicted, while PID does not contribute to the prediction.
* **Columns with High Missing Data:** Features with a high proportion of missing values, such as Mas\_Vnr\_Type, Garage\_Yr\_Blt, and Misc\_Feature, were removed from both the training and test datasets. This decision prevents the introduction of noise or bias caused by excessive missing data in these columns.

**1.5 Categorical Variable Encoding**

Categorical variables were identified and transformed using one-hot encoding. The encoder was set to handle unknown categories in the test set gracefully to prevent errors during the transformation.

**1.6 Standardization of Numerical Features**

All numerical features were standardized using the StandardScaler, which scales the features to have a mean of 0 and a standard deviation of 1. Standardization is essential when applying models like Ridge regression, which are sensitive to the scale of input features. This ensures that features with larger numerical ranges do not disproportionately influence the model’s performance.

**1.7 Target Variable Transformation**

The target variable, Sale\_Price, was transformed using a logarithmic transformation. This technique was applied to reduce skewness and stabilize variance in the target variable. Log-transforming the target also helps models better predict relationships in cases where prices are not linearly related to the features.

**2. Model Implementation**

**2.1 Ridge Regression Model**

A Ridge regression model was implemented to predict house prices. Ridge regression was chosen because it adds a regularization term to the linear regression model, which helps prevent overfitting, especially when the dataset contains multicollinearity (i.e., highly correlated features).

The model was trained on the preprocessed training dataset and applied to the test dataset. The regularization strength (alpha) was set to 1.0, a standard choice that balances bias and variance without over-penalizing large coefficients.

**2.2 Decision Tree Regressor**

**Conclusion**