****

**Course: DATA522 - Solving Big Data Problems-Data Analytics - Spring 2025 (Online)**

**Report: Midterm Progress Report**

**Date: May 15, 2025**

**Name: Meagan O’Briant Student ID: 770047676**

**Predicting House Prices Using Advanced Regression Techniques**

**Team Members:** Meagan O’Briant; Osly Ariel Cabrera Fletes; Orellana Barroso Javier Andres; Akhil Abburi

**Kaggle Competition Chosen:** House Prices - Advanced Regression Techniques (<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques>)

**GitHub URL:** <https://github.com/MeagOBriant/House-Prices-Regression-Final-Project.git>

## Scope of the Project

This project addresses the challenge of accurately predicting residential property prices based on various quantitative and categorical features. The ability to develop a reliable predictive model has practical implications in real estate valuation, investment decision-making, and urban development planning. We aim to create a regression-based machine learning model that minimizes prediction error while maintaining generalizability to unseen housing data.

## Data Description

The dataset is sourced from the Kaggle competition “House Prices: Advanced Regression Techniques.” The dataset contains 1460 training records and 1459 testing records, with 81 features describing various attributes of the houses, including square footage, number of rooms, location features, year built, and quality ratings. The dataset includes both numerical and categorical variables, requiring preprocessing such as encoding, normalization, and imputation of missing values.

## Research Question

Can we accurately predict the sale price of a house based on its physical characteristics and location attributes using machine learning regression models?

## Current Status

As of now, the project is progressing steadily. We have:  
- Selected and downloaded the dataset from Kaggle.  
- Created a GitHub repository and integrated it with RStudio for collaborative version control.  
- Conducted initial exploratory data analysis (EDA), examining feature distributions, identifying missing values, and assessing basic variable relationships.  
- Started data cleaning and preprocessing, including encoding categorical variables and handling NA values.  
- Built and tested basic regression models such as Linear Regression to establish a performance baseline.  
  
We are currently working on more advanced steps, including feature engineering (e.g., log transformation, new variable creation) and model tuning. Our upcoming tasks include evaluating Ridge and Lasso regression models, exploring tree-based models like Gradient Boosting, and applying cross-validation techniques to validate model performance. We also aim to investigate ensemble methods, model interpretability tools like SHAP, and advanced imputation techniques as we continue refining our workflow.

The next phase involves hyperparameter tuning for the final model, improving performance based on cross-validation results. We’re also preparing the final project report and presentation, where we will discuss insights, model performance, and limitations.