****

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

**Report: Final Project Abstract**

**Date: April 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>

**Abstract**

This project focuses on developing a predictive model for estimating housing sale prices using the Ames Housing dataset provided by Kaggle. The dataset contains 79 explanatory variables describing various attributes of residential homes, including lot size, year built, quality ratings, neighborhood, and more. Predicting housing prices is a highly relevant problem in real estate analytics and serves as an excellent opportunity to explore regression modeling and data visualization in RStudio. The objective is to apply statistical methods to accurately predict house prices and evaluate model performance using Root Mean Square Error (RMSE), which serves as the competition’s scoring metric.

**Approach/Strategy**

The project will be conducted in RStudio using a set of packages tailored for data analysis and machine learning. The initial steps involve data exploration and cleaning using dplyr and tidyverse, which will help identify and address missing values, skewed distributions, and potential outliers. Feature engineering will follow, including transformations and categorical encoding to optimize model input. Visualizations will be created using ggplot2 to uncover patterns and relationships in the data.  
  
 For modeling, the project will begin with basic linear regression to establish a performance baseline. More advanced models like random forest and gradient boosting will then be applied using packages such as randomForest, xgboost, and caret. Model performance will be assessed using cross-validation and RMSE. All work will be documented and version-controlled via GitHub, with the final deliverable structured in PowerPoint format per course guidelines.

**References:**

Balbaa, M. E., Tashtemirovich, A. O., Ismailova, N., & Batirova, N. (2023). Real-time analytics in financial market forecasting: A big data approach. In The International Conference on Future Networks and Distributed Systems (ICFNDS ’23), December 21–22, Dubai, United Arab Emirates. ACM. (UoNA e-library).

Liu, Y., & Tang, A. (2022).

Loshin, D. (2013). *Big data analytics : from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph*. Elsevier/Morgan Kaufmann.

‌  
Prediction method of government economic situation based on big data analysis. Digital Government: Research and Practice, 3(4), Article 27. (UoNA e-library).