Summary of Abdulhafedh (2022): *Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables*

Introduction

This document provides a detailed summary of the article by Abdulhafedh (2022), entitled *"Incorporating Multiple Linear Regression in Predicting the House Prices Using a Big Real Estate Dataset with 80 Independent Variables."* The article presents a comprehensive methodological framework for applying Multiple Linear Regression (MLR) to a large and heterogeneous real-estate dataset, specifically the Ames Housing Dataset. The author's goal is to demonstrate that, when properly implemented, MLR remains a powerful and interpretable statistical tool for house price prediction.

Dataset Description

The study makes use of the Ames Housing Dataset, a rich alternative to the Boston Housing Dataset. It contains 80 explanatory variables describing structural, aesthetic, locational, and functional characteristics of houses sold in Ames, Iowa. These variables include both quantitative and qualitative attributes, such as:

- living area (e.g., GrLivArea),
- basement area (e.g., TotalBsmtSF),
- quality assessments (e.g., OverallQual, KitchenQual, ExterQual),
- year of construction and remodeling (YearBuilt, YearRemodAdd),

- garage characteristics (GarageArea),
- locational indicators (Neighborhood).

The dependent variable is *SalePrice*. The author highlights that the dataset is particularly suitable for academic teaching and applied statistical modelling due to its dimensionality, realistic complexity, and combination of numeric and categorical variables.

Methodological Framework

The article presents a thorough workflow for constructing an MLR model. The steps include:

1. Exploratory Data Analysis (EDA)

The author investigates variable distributions, identifies missing values, and detects potential outliers. Several continuous variables—including SalePrice, basement and living areas—are found to be right-skewed. Correlation analysis reveals substantial multicollinearity among several features.

2. Data Cleaning and Pre-processing

The article emphasises:

- handling missing values via imputation or variable removal,
- converting categorical variables into dummy variables (one-hot encoding),
- standardising or transforming skewed numerical variables when necessary.

3. Feature Reduction

Given the large number of predictors, the author performs:

- **correlation filtering** to remove highly collinear predictors,
- **variance thresholding** to discard near-constant predictors,
- **stepwise selection and backward elimination** to obtain a more parsimonious model.

The feature reduction process significantly improves model stability and interpretability.

4. Fitting the Multiple Linear Regression Model

The OLS model is then estimated:

$$\text{SalePrice} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

with all classical assumptions carefully assessed.

5. Diagnostic Checking

The author evaluates:

- **Linearity**: residual plots indicate mostly linear relationships.
- **Normality of residuals**: assessed using histograms and Q–Q plots.
- **Homoscedasticity**: verified through residual vs. fitted plots.
- **Multicollinearity**: assessed using Variance Inflation Factors (VIF). Predictors with extremely high VIF values are removed from the model.

Despite some mild deviations from homoscedasticity, the model satisfies the assumptions sufficiently well.

Results

After variable reduction and diagnostic refinement, the final OLS model displays strong predictive performance:

$$R^2 \approx 0.928, \quad \text{Adjusted } R^2 \approx 0.926.$$

This indicates that approximately 93% of the variability in house prices is explained by the model—an exceptionally high value in applied econometric modelling.

The most influential predictors identified in the paper include:

- **OverallQual**: the strongest overall predictor of price,
- **GrLivArea**: main above-ground living area,
- **TotalBsmtSF**: basement area,
- **YearBuilt** and **YearRemodAdd**: capturing age and remodeling,

- **GarageArea**: size of the garage,
- **KitchenQual** and **ExterQual**: quality measures,
- **Neighborhood**: locational effects.

These results align with findings from other literature on the Ames dataset and real-estate economics in general.

Model Validation

The model is validated using train–test splitting and cross-validation. The author reports that predictive performance remains consistently high on unseen data, indicating that overfitting was successfully mitigated.

Conclusion

Abdulhafedh (2022) demonstrates that Multiple Linear Regression—when combined with thoughtful feature engineering, multicollinearity control, and rigorous diagnostics—is a robust and interpretable method for predicting house prices in a complex real-estate environment. The Ames Housing Dataset proves to be an ideal benchmark for teaching regression modelling and offers valuable insights into which structural, qualitative, and locational features contribute most to the market value of residential properties.