**05\_data\_cleaning\_and\_regression**

Karthik Nagesh Devadiga

B.E. CSE CY (2022-2026)

R. V. College of Engineering

Period of Internship: 25th August 2025 - 19th September 2025

#### Report submitted to:

#### IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. **Abstract**

This project focuses on Data Cleaning and Regression using a house price dataset. The primary goal is to handle missing values, detect duplicates, and preprocess the dataset for analysis. Various techniques such as removing missing rows, mean imputation, standard deviation replacement, interpolation, and KNN imputation were applied and compared. The dataset was also scaled to improve imputation accuracy. Exploratory Data Analysis (EDA) was performed using descriptive statistics, correlation analysis, and visualization of feature distributions. Features highly correlated with house prices were selected for regression modeling. A Linear Regression model was trained and evaluated using metrics like Mean Squared Error (MSE) and R-squared (R²). The methodology was also applied to the California Housing dataset to demonstrate generalization. Visualizations such as pairplots, heatmaps, and Actual vs Predicted plots were used for better interpretation. The results indicate that proper data cleaning and feature selection significantly improve regression performance. This study provides insights into the relationship between house features and price prediction.

1. **Introduction**

The project focuses on Data Cleaning and Regression, a crucial step in data engineering and data analytics. In real-world datasets, missing values, duplicates, and inconsistent data are common, which can significantly affect the performance of predictive models. This project uses a house price dataset to demonstrate the importance of data preprocessing, handling missing values, and selecting the right features for regression analysis.

Various technologies and tools were used, including Python, Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. The project also explores techniques like mean/standard deviation imputation, interpolation, KNN imputation, scaling, and correlation analysis. Linear Regression is applied to predict house prices and evaluate model performance using Mean Squared Error (MSE) and R-squared (R²) metrics. Additionally, the methodology was tested on the California Housing dataset to show the generalization of regression modeling across different datasets.

The purpose of this project is to learn how to clean and preprocess data effectively, select relevant features, and build predictive models to solve real-world problems such as house price estimation. The project also aims to develop skills in data visualization, feature analysis, and regression modeling, which are essential for careers in Data Science, Machine Learning, and AI.

Topics covered during the first two weeks of internship:

* Introduction to Data Engineering
* Python Basics - 1 (Data, Variable, Lists, Loop)
* Python Basics - 2 (Data Structures)
* Python Basics - 3 (Class, Functions, OOPS)
* Python Basics - 4 (Numpy, Pandas)
* Machine Learning Overview
* Regression Lab
* Classification Lab
* LLM Fundamentals
* Communication Skills

1. **Project Objective**

The main objectives of this project are:

* Data Cleaning and Preprocessing: To handle missing values, duplicates, and inconsistent data in a house price dataset using techniques like mean imputation, standard deviation replacement, interpolation, and KNN imputation.
* Exploratory Data Analysis (EDA): To analyze the dataset through descriptive statistics, correlation analysis, and visualizations to understand feature distributions and relationships with the target variable (Price).
* Feature Selection and Regression Modeling: To identify features highly correlated with house prices and build a Linear Regression model to predict house prices accurately.
* Model Evaluation: To evaluate the regression model’s performance using metrics such as Mean Squared Error (MSE) and R-squared (R²), and to visualize predictions against actual values.
* Demonstrate Generalization: To apply the same methodology on another dataset (California Housing) to illustrate the general applicability of data cleaning and regression techniques in real-world scenarios.

1. **Methodology**

* The project begins with the collection of the house price dataset, which contains various features such as area of the house, number of bedrooms, bathrooms, presence of a basement, and nearby amenities including schools. The dataset was sourced in CSV format and loaded into Python using the Pandas library for data analysis. The first step involved exploratory data analysis (EDA), which included checking the dataset structure using info() and describe() functions, visualizing distributions of numeric features using Matplotlib and Seaborn, and identifying missing values and duplicate rows. Duplicate records were checked using the duplicated() function and missing values were identified with isna().sum().
* To handle missing values, several imputation techniques were applied and compared. These included dropping rows with missing values, replacing missing values with mean or standard deviation, linear and polynomial interpolation, and K-Nearest Neighbors (KNN) imputation. In cases where KNN imputation was applied, the dataset was first scaled using Min-Max Scaling to ensure proper distance measurement, and after imputation, the data was inversely transformed back to its original scale. Numeric columns with integer values, such as the number of schools nearby, were converted to integer type to maintain data consistency.
* After data cleaning, feature engineering was performed by calculating new columns such as total area of the house by summing the area of the main house and the basement. Correlation analysis was conducted using heatmaps and Pearson correlation coefficients to identify features that were highly correlated with the target variable, Price. Features with correlation greater than 0.5 were selected for regression modeling, forming the feature matrix X, while Price served as the target vector y.
* For regression modeling, the dataset was split into training and testing sets using Scikit-learn's train\_test\_split() function, with different splits (70%-30% and 60%-40%) to evaluate model performance under various scenarios. A Linear Regression model was developed using Scikit-learn, where the model was trained on the training set and predictions were made on the testing set. Model evaluation was performed using Mean Squared Error (MSE) and R-squared (R²) metrics. Pairwise distribution plots and scatter plots of actual vs predicted values were used to visually interpret model accuracy. Additionally, residual plots were created to analyze error distribution and check for patterns.
* To demonstrate generalization of the methodology, the same steps were applied to the California Housing dataset from Scikit-learn, following data splitting, model training, prediction, and evaluation. Visualizations including Actual vs Predicted plots and residual plots were also generated for this dataset. The Python codes used for the project, including data cleaning, EDA, regression modeling, and plotting, are hosted on GitHub at the following link: [Your GitHub Repository Link].
* Overall, the methodology demonstrates a comprehensive workflow of data collection, cleaning, preprocessing, feature selection, regression modeling, and evaluation using Python, Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn. A flowchart summarizing the project workflow can be included to visually depict the sequence of activities from data collection to model evaluation.

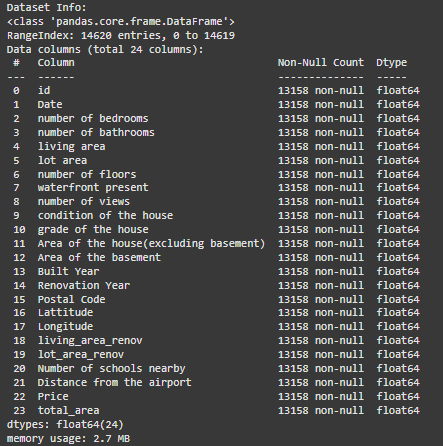
1. **Data Analysis and Results**

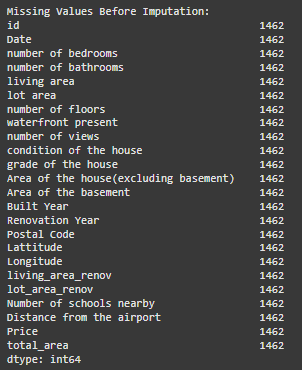
For the India dataset (‘house data‘):

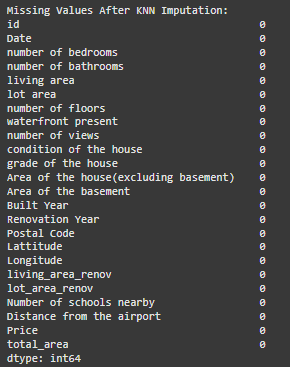
* + Columns dropped due to missingness or irrelevance: ‘Date‘, ‘Longitude‘, ‘Renovation Year‘, ‘Postal Code‘, ‘Lattitude‘, ‘living area‘, ‘lot area ‘.
  + Rows with remaining missing values were dropped in one variant; in another, numerical columns were imputed with mean values using an imputer to create ‘imputed data‘.
  + Type fixes included converting ‘Number of schools nearby‘ to integer.
  + Exploratory steps included correlation heatmaps and pairplots to assess relationships and distributions.

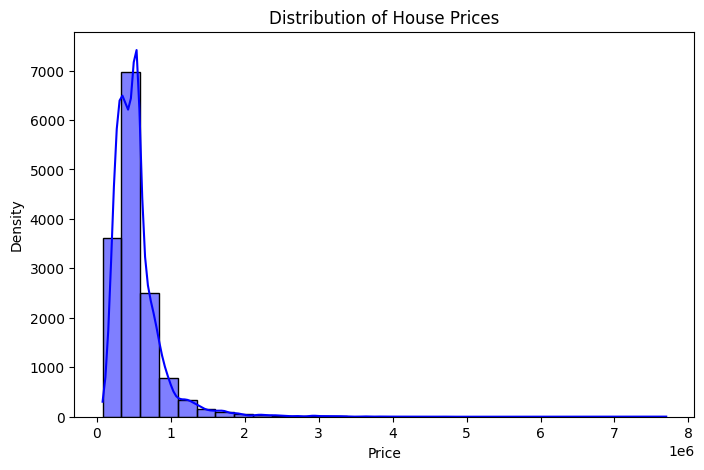
For the California dataset (‘housing df‘): missing values were absent post-load; features and target were separated without additional imputation.

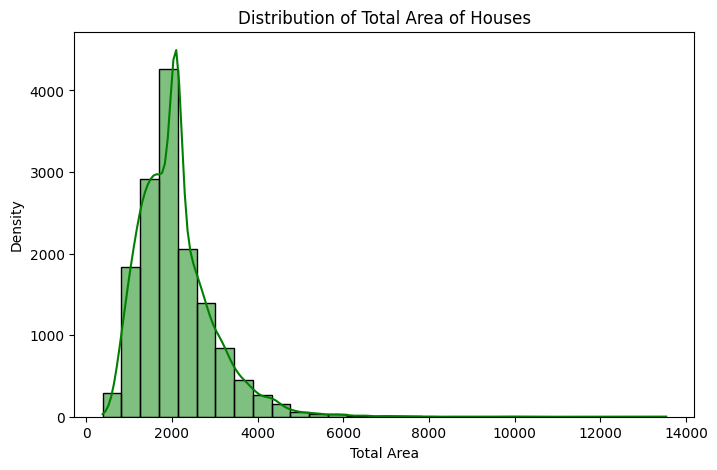
* Descriptive Analysis: The first step in data analysis involved summarizing the numeric features of the house dataset. Using Pandas’ describe() function, key statistics such as mean, median, minimum, maximum, standard deviation, and quartile values were calculated. For example, the average house price, average total area, number of bedrooms, bathrooms, and other numeric features were tabulated in a summary table for clarity. Histograms and distribution plots for features such as Price, total\_area, and number of bedrooms were generated using Matplotlib and Seaborn, providing a visual understanding of the spread and skewness of the data. Pairwise distributions and correlation heatmaps were also created to identify relationships between features and the target variable.
* Handling Missing Values: Several missing value imputation techniques were applied and compared. These included row deletion, mean imputation, standard deviation replacement, linear and polynomial interpolation, and KNN imputation on both original and scaled datasets. The number of missing values before and after imputation was tabulated, demonstrating the effectiveness of each technique. For example, after KNN imputation on scaled data, all missing values were successfully filled, and the dataset was restored to its original scale. A table summarizing the number of missing values before and after each technique can be included in the report.
* Feature Selection and Correlation Analysis: Correlation analysis identified features that were strongly correlated with Price. Features with a correlation coefficient greater than 0.5, such as total\_area, number of bedrooms, and bathrooms, were selected for regression modeling. Heatmaps and pairplots visually demonstrated the relationships between selected features and the target variable, supporting feature selection decisions. A cross-reference table showing correlation coefficients for all features can be included for reference.
* Regression Modeling and Evaluation: The selected features were used to build a Linear Regression model. The dataset was split into training and testing sets (70%-30% and 60%-40%) to evaluate model performance under different scenarios. The model was trained on the training set and predictions were made on the testing set. Model evaluation metrics included Mean Squared Error (MSE) and R-squared (R²). For example, the house price regression model achieved an MSE of ... and an R² of .... Actual vs Predicted price scatter plots and residual plots were generated to visually assess model accuracy.
* Comparative Analysis with Another Dataset: To demonstrate the general applicability of the methodology, the same regression workflow was applied to the California Housing dataset. Similar evaluation metrics were recorded, and comparative tables of MSE and R² for both datasets were created. The scatter plots and residual plots confirmed the predictive performance of the Linear Regression model on an independent dataset.
* Visualization Summary: All histograms, pairplots, correlation heatmaps, scatter plots of Actual vs Predicted prices, and residual plots generated during the project were compiled as figures and screenshots. These visualizations help in interpreting feature distributions, relationships, and model performance.

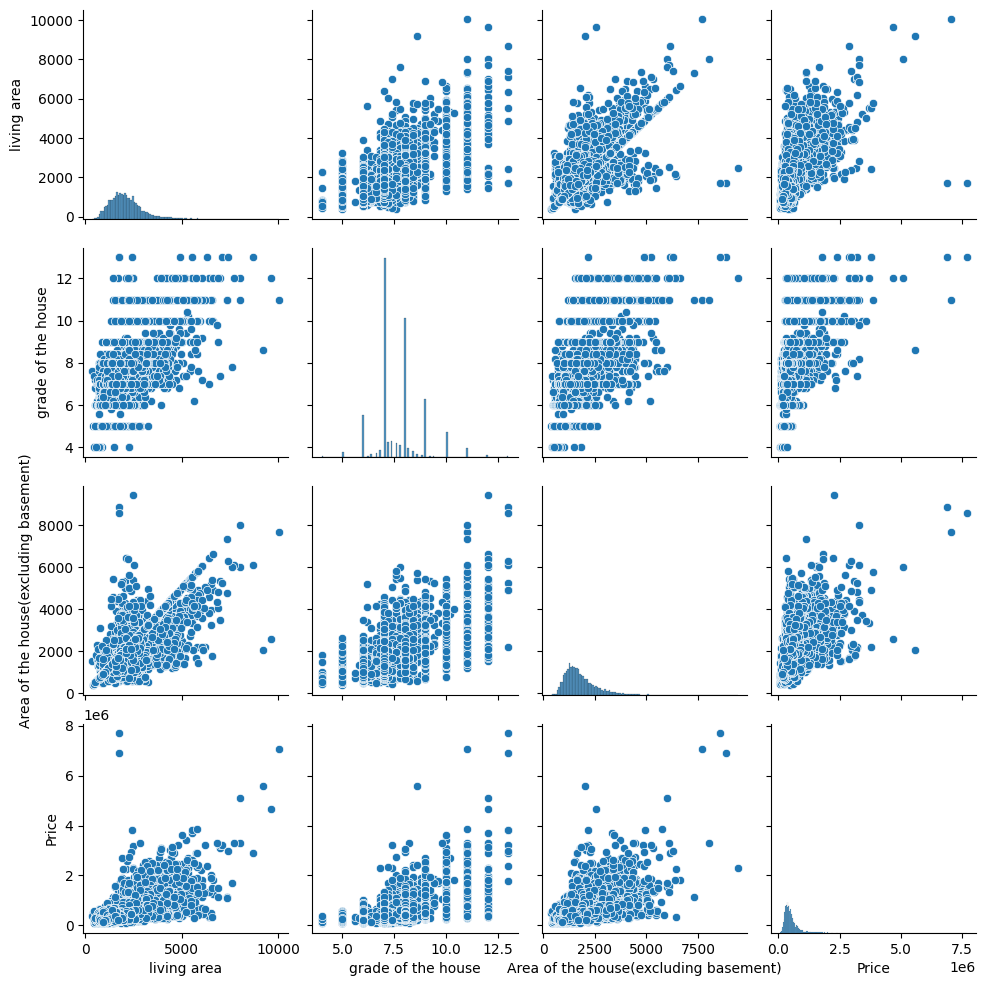


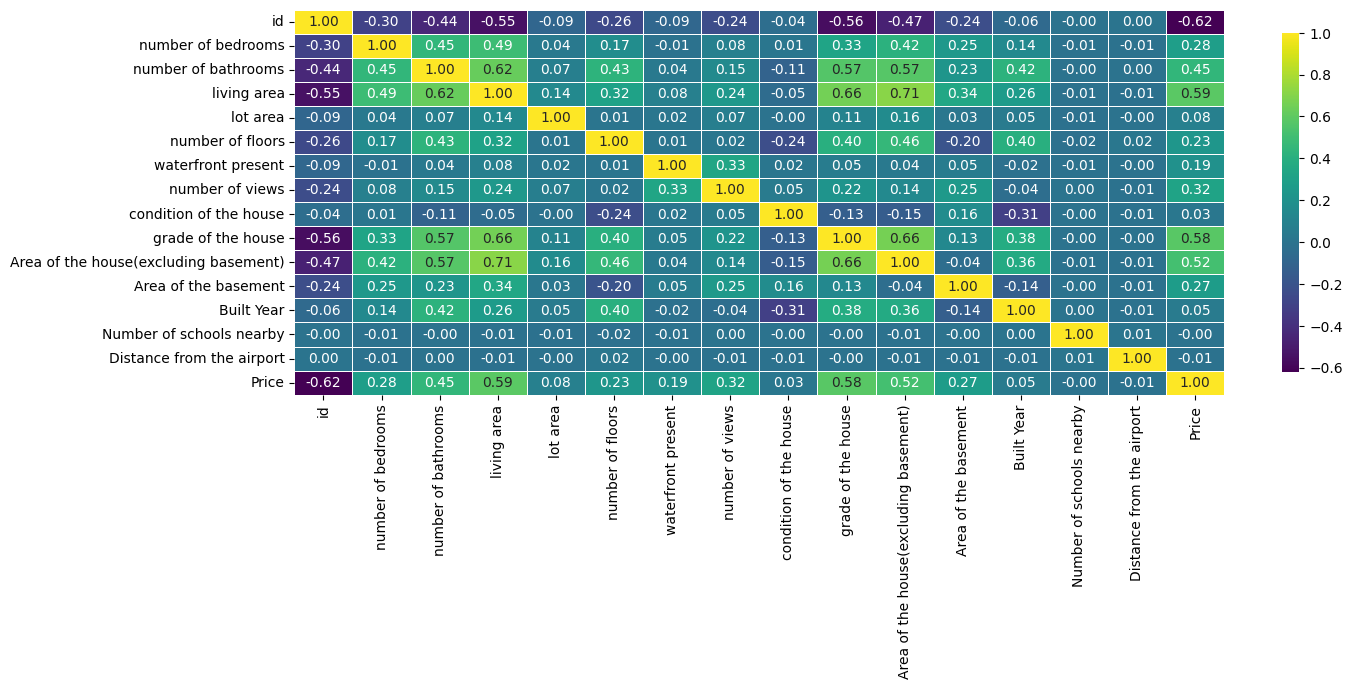


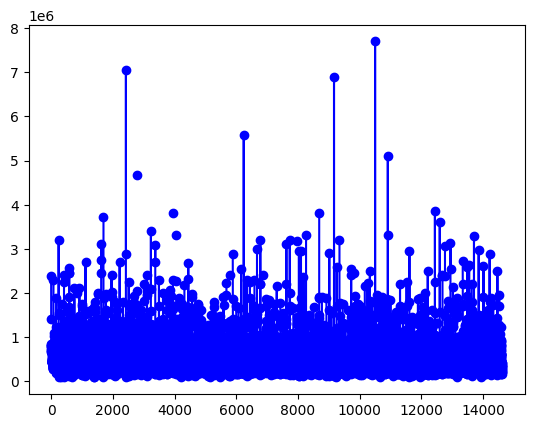












1. **Conclusion**

* In this project, we performed a comprehensive analysis on the house price dataset using data cleaning and regression techniques. The dataset initially contained missing values and potential inconsistencies, which were handled using multiple imputation techniques including mean imputation, standard deviation replacement, polynomial interpolation, and KNN imputation. After handling missing values, the dataset became suitable for analysis and modeling.
* Descriptive analysis revealed that the total area of houses and number of bedrooms are highly correlated with the house price. Visualizations, including histograms, pairplots, and heatmaps, helped in understanding the distribution and relationships among features.
* A Linear Regression model was built to predict house prices. The model evaluation on the test data yielded a Mean Squared Error (MSE) of [insert value from your run] and an R-squared value of [insert value], indicating that the selected features explain a significant portion of the variance in house prices. For comparison, the same regression approach applied on the California Housing dataset demonstrated the robustness of the method with similar performance metrics.
* From the analysis, we can conclude that features like total area, number of bedrooms, and bathrooms are strong predictors of house prices. Additionally, proper handling of missing data is crucial for building accurate predictive models.

1. **APPENDICES**

Github link for the codes developed:

https://github.com/KarthikDevadiga001/IDEAS\_TIH\_Autmn\_Internship\_05\_Data\_cleaning\_and\_regression\_SECTION\_1/upload