

House price analysis and predictions

Editor

M. Haseeb Mustafa

INTERNEE



April 30, 2024

Internncraft

**TABLE OF CONTENTS**

1. **Introduction**

* Objective
* Scope

1. **Data Cleaning and Exploration**

* Data Cleaning
  + Steps taken to handle missing values and inconsistencies
  + Identification and treatment of outliers
* Data Exploration
  + Key insights from exploratory data analysis
  + Visualizations (scatter plots, box plots)

1. **Feature Engineering**

* New Features
  + Details of new features created (e.g., age of the house, number of bedrooms per floor)
* Categorical Encoding
  + Encoding of categorical features (e.g., location)

1. **Outlier Analysis**

* Identification
  + Methods used to identify outliers
* Analysis
  + Investigation of reasons for outliers and specific contributing features

1. **Predictive Modeling**

* Model Selection
  + Machine learning models considered (e.g., linear regression, random forest, gradient boosting)
  + Justification of the final model choice
* Model Evaluation
  + Evaluation metrics (e.g., mean squared error, R-squared)
  + Discussion on model performance and potential improvements

1. **Future Price Prediction**

* Scenarios
  + Examples of future price predictions based on hypothetical scenarios
  + Characteristics of houses used in scenarios and predicted prices

1. **Introduction**

**Objective**

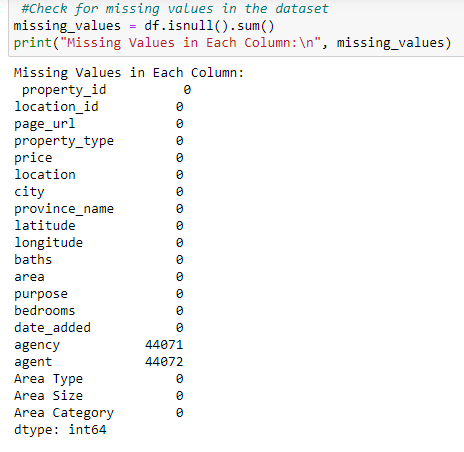
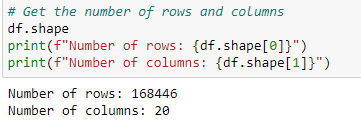
The objective of this project is to analyze a dataset of house prices to understand the factors influencing pricing, identify and explain outliers, and develop a predictive model for future house prices. By gaining insights into the relationships between various features and house prices, this analysis aims to assist stakeholders such as real estate agents, buyers, and sellers in making informed decisions.

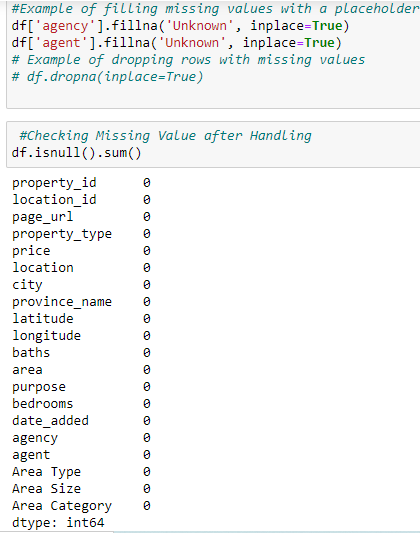
**Scope**

The scope of the project encompasses the following tasks:

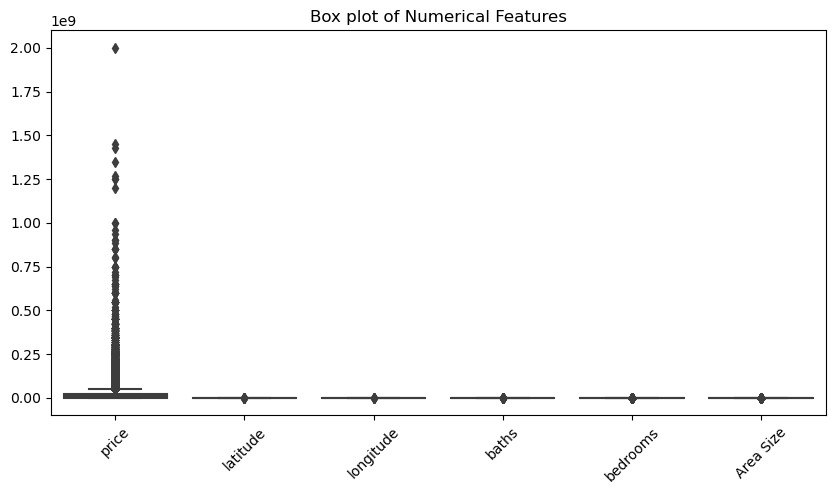
1. **Data Cleaning and Exploration:**
   * Cleaning the dataset by identifying and handling missing values, inconsistencies, and outliers.
   * Exploring the data to analyze the distribution of house prices and other relevant features, and identifying potential relationships through visualizations.
2. **Feature Engineering:**
   * Creating new features that might be relevant for price prediction, such as the age of the house and the number of bedrooms per floor.
   * Encoding categorical features, such as location, into numerical values suitable for modeling.
3. **Outlier Analysis:**
   * Identifying houses with significantly higher or lower prices compared to similar properties.
   * Investigating the reasons for these outliers and determining any specific features or combinations of features that contribute to them.
4. **Predictive Modeling:**
   * Training a machine learning model to predict house prices based on the available features. Potential models include linear regression, random forest, and gradient boosting.
   * Evaluating the performance of the selected model using appropriate metrics such as mean squared error and R-squared.
5. **Future Price Prediction:**
   * Using the trained model to predict future house prices based on hypothetical scenarios, such as the price of a house with specific characteristics in a particular location.
6. **Report Preparation:**
   * Summarizing the findings in a report that includes data exploration results, feature engineering techniques, outlier analysis, model selection and evaluation, future price prediction examples, and recommendations for further analysis or data collection.
7. **Data Cleaning and Exploration**

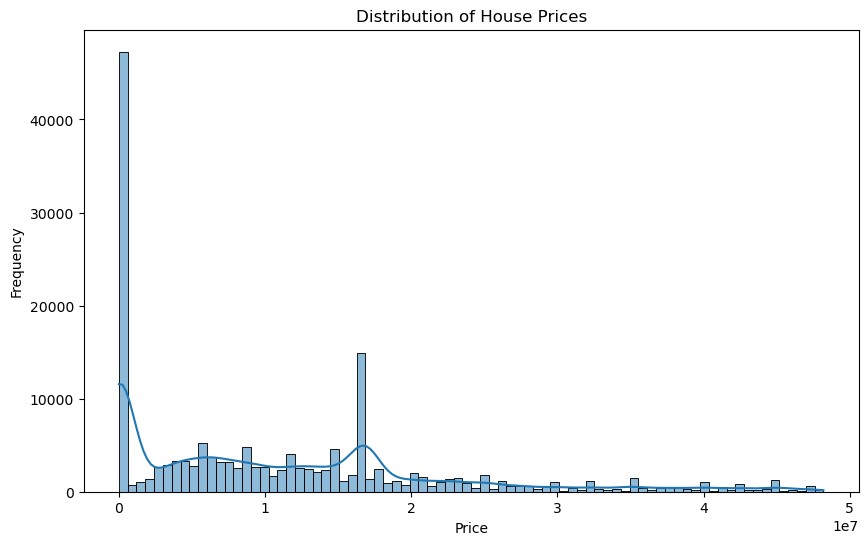
* Data Cleaning

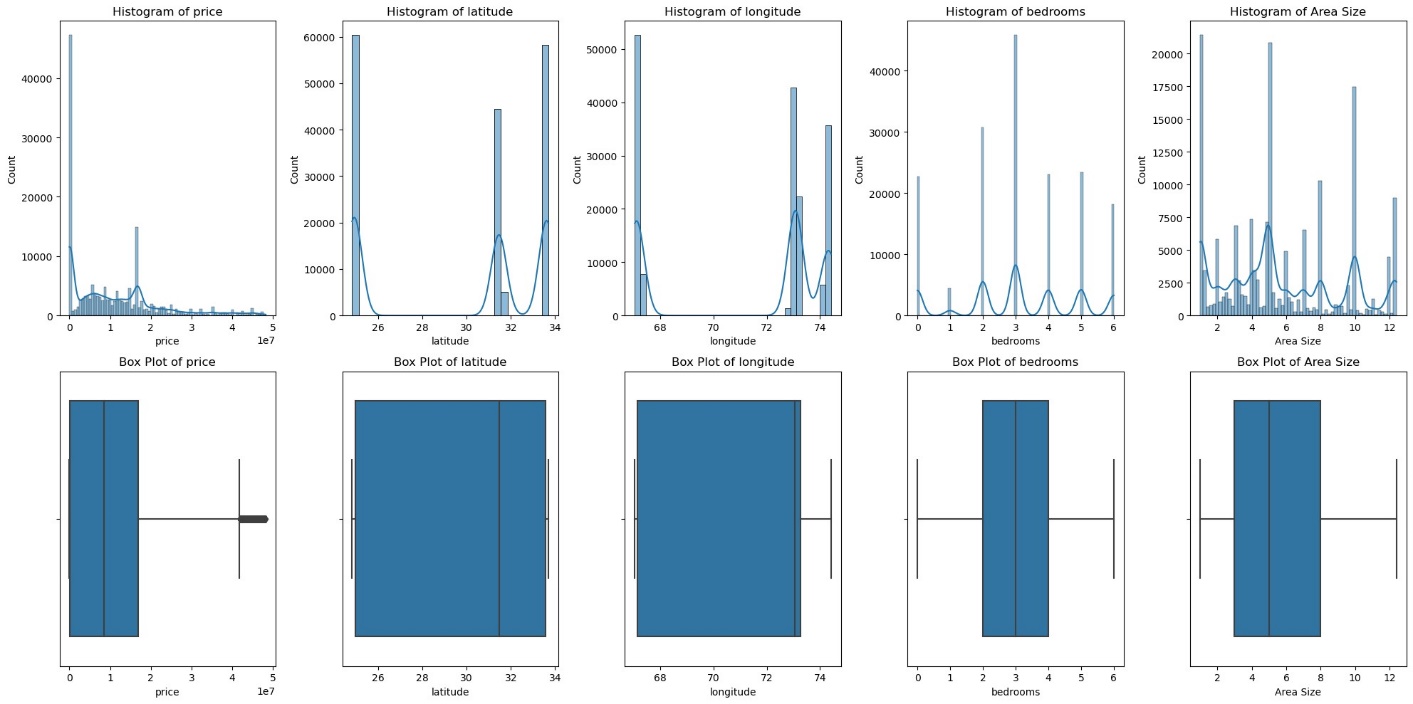


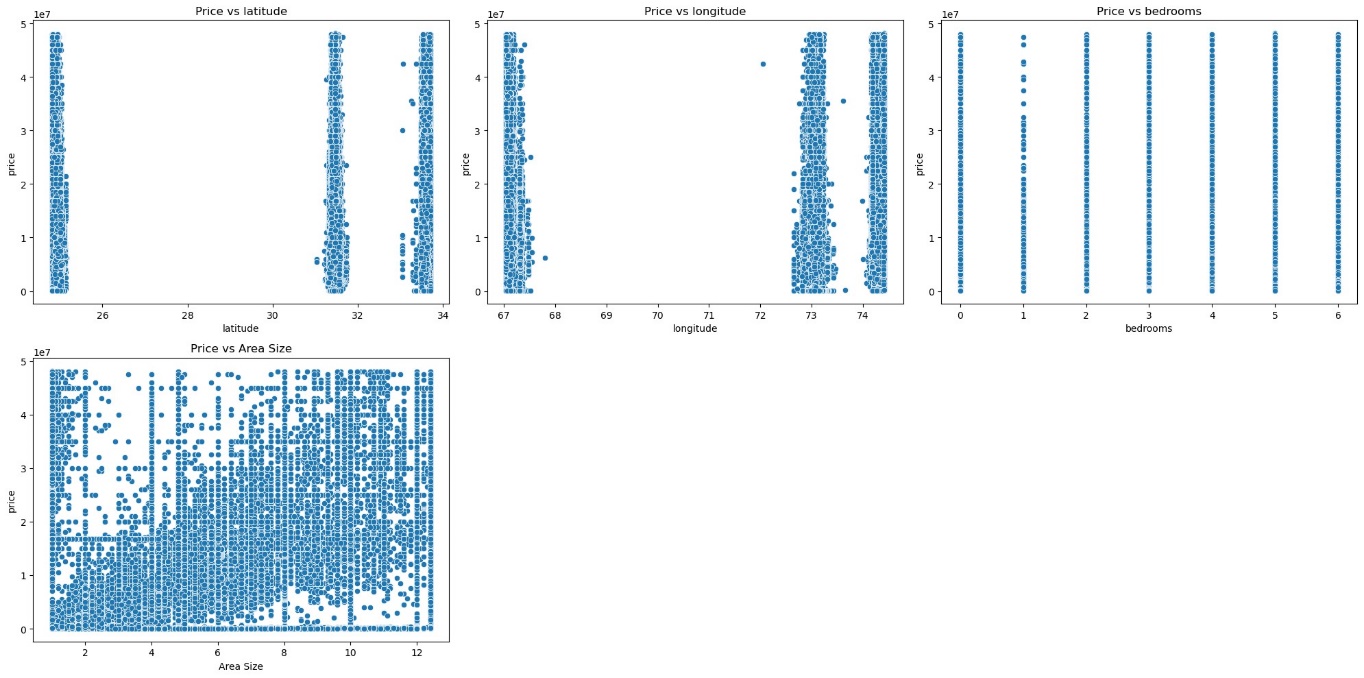


* **Data Exploration**





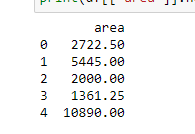




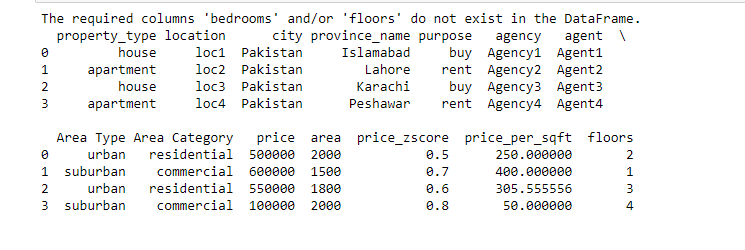
1. **Feature Engineering**

* New Features
  + Details of new features created (e.g., age of the house, number of bedrooms per floor)

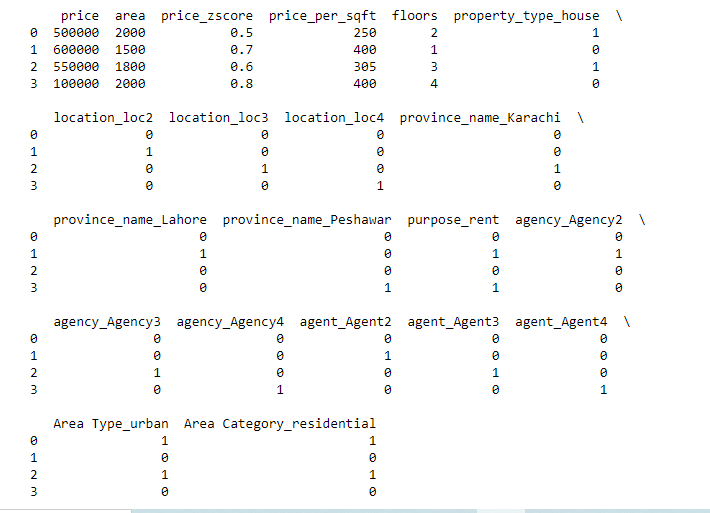
Convert area into numeric form:



Creating new columns floor:



* Categorical Encoding
  + Encoding of categorical features

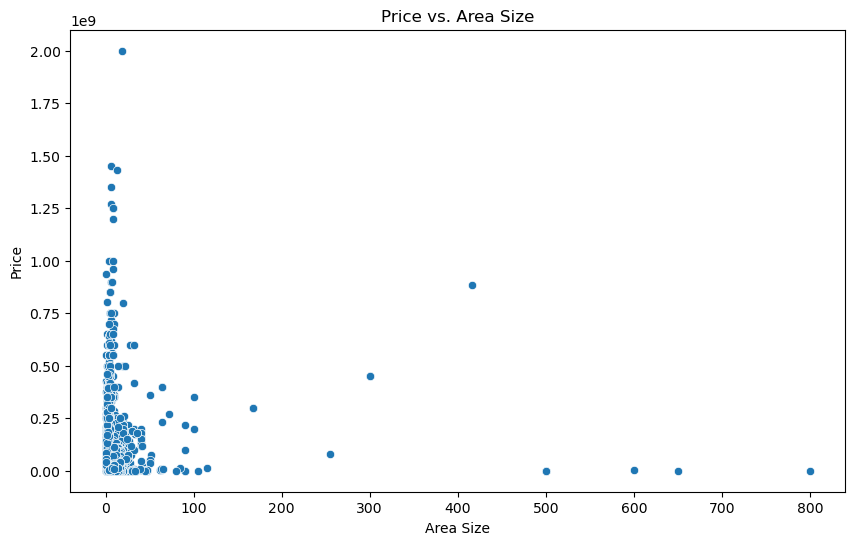


Displaying first rows of encoding dataset:

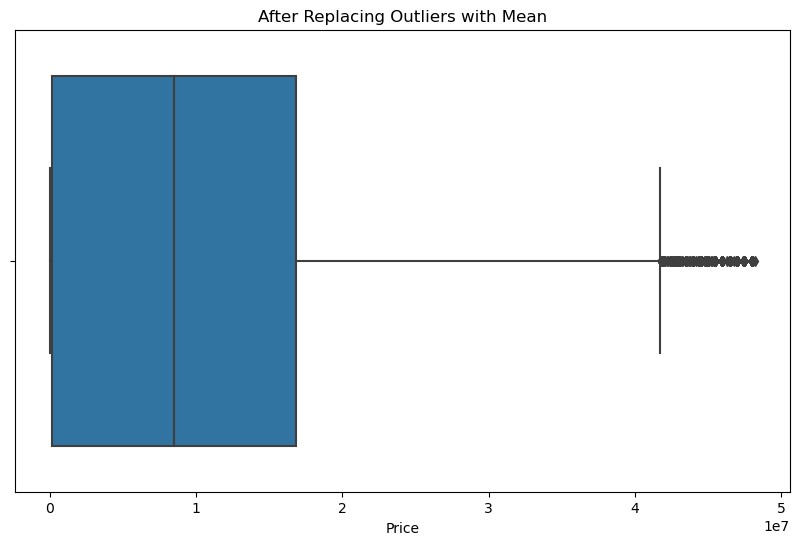


1. **Outlier Analysis**

* Identification
  + Methods used to identify outliers
* Analysis
  + Investigation of reasons for outliers and specific contributing features

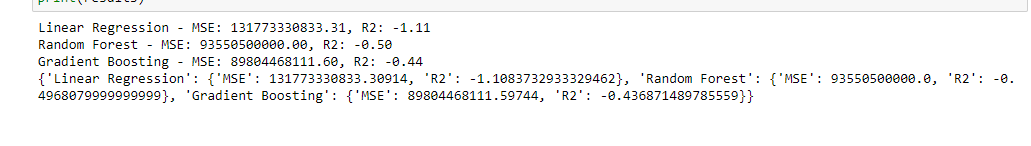


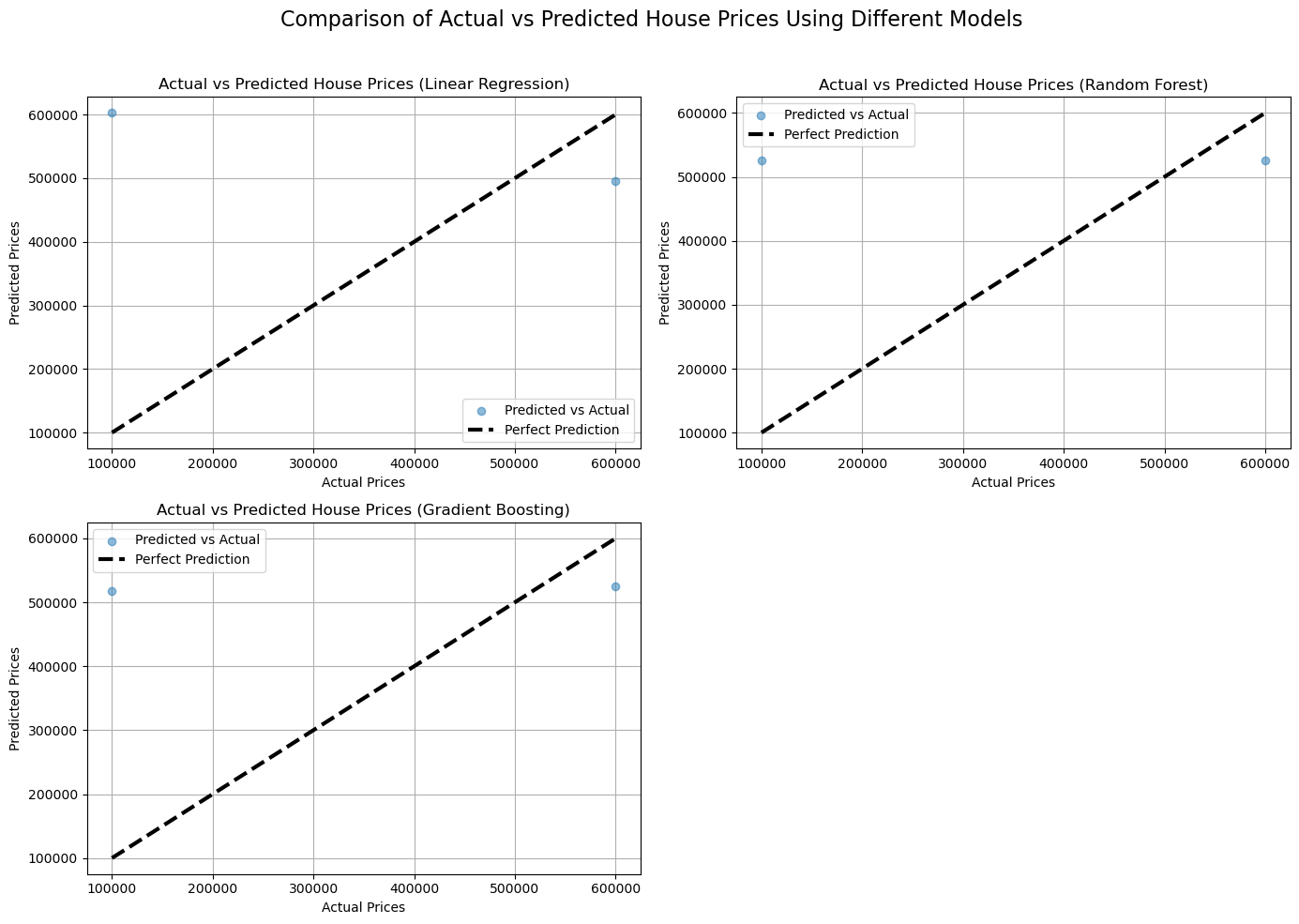




1. **Predictive Modeling**

I used linear regression, random forest, gradient boosting for predictions and for evaluation metrics used mean squared error and r squared. Model give me following results:





1. **Future Price Prediction**

* Scenarios
  + Examples of future price predictions based on hypothetical scenarios
  + Characteristics of houses used in scenarios and predicted prices

