Executive Report: Insights from Exploratory Data Analysis and Linear Regression Model

Prepared by: [Your Name]
Date: [Current Date]

1. Introduction

This executive report presents insights gained from the Exploratory Data Analysis (EDA) of a housing dataset and the development of a Linear Regression model for predicting house prices. The analysis provides valuable information about the dataset, addresses missing values, explores relationships between variables, and interprets the results of the Linear Regression model.

2. Data Exploration and Preprocessing

The initial steps involved loading the training and testing datasets and performing data exploration. The shape of the training dataset was checked, and the first few rows were examined. Missing values were identified using the `info()` method, and unique values were counted for the 'loc' and 'title' columns. This exploration laid the foundation for understanding the dataset's structure.

To handle missing values in categorical columns, the `location` and `Title` functions were created. These functions randomly assigned valid values to missing entries in the 'loc' and 'title' columns, respectively. For numerical columns ('bedroom,' 'bathroom,' and 'parking_space'), missing values were filled with rounded mean values. Duplicate rows were also checked in both the training and testing datasets.

3. Exploratory Data Analysis (EDA)

EDA revealed insights into the dataset's characteristics. Visualizations such as count plots for categorical variables ('Title' and 'loc') provided an understanding of the data distribution. The decision to replace missing values in the 'Title' column with random values was supported by comparing count plots of 'Title' and 'loc.'

Correlation analyses were conducted using scatter plots and heatmaps to identify relationships between numerical features and the target variable, 'price.' These visualizations highlighted the relationships between 'bedroom,' 'bathroom,' and 'parking_space' with 'price.'

4. Linear Regression Model

A Linear Regression model was developed to predict house prices based on the numerical features: 'bedroom,' 'bathroom,' 'parking_space,' 'num_loc,' and 'num_title.' The categorical columns 'location' and 'Title' were encoded using Label Encoding. The dataset was then split into training and testing sets.

The Linear Regression model was trained using the training dataset. The model's coefficients provided insights into the relationship between the features and the target variable. Model evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), were calculated to assess the model's performance.

5. Conclusion and Insights

The EDA highlighted the significance of 'bedroom,' 'bathroom,' and 'parking_space' in determining house prices. The Linear Regression model's coefficients further confirmed these insights. As seen from the model's coefficients, each additional bedroom or bathroom is associated with an increase in house price. The provided scatter plots and regression lines visually demonstrated these relationships.

In conclusion, this analysis provides valuable insights into the factors influencing house prices. It underscores the importance of numerical features like 'bedroom,' 'bathroom,' and 'parking_space,' and their correlation with house prices. The developed Linear Regression model serves as a tool for predicting house prices based on these features.

Overall, this executive report summarizes the key findings of the Exploratory Data Analysis and the Linear Regression model, providing actionable insights for stakeholders and decision-makers in the real estate industry.

For further details and a comprehensive overview of the analysis, please refer to the accompanying Jupyter Notebook.