DOCUMENTATION FOR REAL ESTATE PRICE PREDICTION

SUMMARY

This document provides a comprehensive analysis of a real estate dataset. The primary objective is to predict the house price per unit area based on various features such as transaction date, house age, distance to the nearest MRT station, number of convenience stores, latitude and longitude. The analysis involves data cleaning, handling outliers, resolving skewness, and building a predictive model using machine learning techniques.

METHODOLOGY

1. Data Loading: The dataset was loaded into a pandas Data Frame for analysis.
2. Data Cleaning: The dataset was checked for missing values and data types. No missing values were found, and all columns were of numerical data types.
3. Handling Outliers: Outliers were detected and resolved using the Interquartile Range (IQR) method. A custom function was created to replace outliers with the upper and lower bounds.
4. Handling Skewness: Skewness in the data was detected and resolved using appropriate transformations such as Box-Cox and Yeo-Johnson transformations.
5. Date Format Handling: The transaction date was converted from a decimal year format to a standard date format. Year, month, and day were extracted into separate columns.
6. Renaming and Deleting Columns: Columns were renamed for better readability, and unnecessary columns were dropped.
7. Bivariate Analysis: Relationships between key features and the target variable were visualized using scatter plots and correlation coefficients.
8. Correlation Matrix: A correlation matrix was plotted to understand the relationships between all features.

9. Model Training and Testing: The dataset was split into training and testing sets. Various regression models were trained and evaluated to determine the best model for prediction.
10. GUI for Prediction: A graphical user interface (GUI) was created to allow users to input data and get predictions for house prices.

Problem Statement

The primary objective of this analysis is to predict the house price per unit area based on various features available in the dataset. The features include transaction date, house age, distance to the nearest MRT station, number of convenience stores, latitude, and longitude. Accurate prediction of house prices is crucial for buyers, sellers, and real estate agents to make informed decisions. The challenge lies in handling outliers, skewness, and ensuring the model's accuracy in predicting unseen data.

Problem Resolution

1. Data Cleaning: Ensured no missing values and correct data types.
2. Handling Outliers: Used the IQR method to detect and replace outliers.
3. Handling Skewness: Applied Box-Cox and Yeo-Johnson transformations to resolve skewness.
4. Date Format Handling: Converted transaction date to standard format and extracted year, month, and day.
5. Renaming and Deleting Columns: Renamed columns for readability and dropped unnecessary columns.
6. Bivariate Analysis: Visualized relationships between key features and the target variable.
7. Correlation Matrix: Plotted a correlation matrix to understand feature relationships.
8. Model Training and Testing: Split data into training and testing sets, trained various regression models, and selected the best model based on R^2 score.
9. GUI for Prediction: Created a GUI for users to input data and get house price predictions.

Conclusion

The analysis successfully addressed the problem of predicting house prices per unit area using various features. By employing data cleaning techniques, handling outliers and skewness, and using machine learning models, we achieved a high R^2 score with the Random Forest Regressor. The GUI provides an easy-to-use interface for making predictions based on user input. The findings indicate that location and accessibility to essential facilities significantly influence house prices. This analysis can aid buyers, sellers, and real estate agents in making informed decisions.