House Price Prediction Project Documentation

1. Importing Libraries and loading the Data

The first step is to import the necessary libraries and load the training and testing datasets.

The first step was to load the training and test datasets using Pandas. This involved reading the CSV files into DataFrame objects and inspecting the data to understand its structure and identify any missing values or inconsistencies.

2. Data Exploration

I perform an initial exploration of the data to understand its structure and summary statistics. EDA was conducted to understand the distribution of the data, identify any patterns or anomalies, and visualize relationships between variables.

3. Data Visualization

Next, I visualize the distributions of key numerical features in the dataset which includes, distribution of Bedrooms, distribution of Bathrooms, distribution of Parking Spaces, distribution of Prices and correlation matrix

4. Data Cleaning and Feature Engineering

This section involves handling missing values, encoding categorical variables, and transforming features. Rows with missing values in the 'loc' column were dropped. For the 'title' column, missing values were filled using distribution-based imputation. The IQR values of other numerical columns ('bathroom', 'bedroom', 'parking_space') were used to fill missing values based on their corresponding 'title' category.

Categorical columns ('loc' and 'title') were encoded using `LabelEncoder` to convert them into numerical values suitable for machine learning algorithms.'

5. Model Training and Evaluation

We split the data into training and validation sets, then train and evaluate multiple regression models.

The best trained model was used for to evaluate our test data was selected. The best regression model(XGBoost Regressor) for this project was select based on its Mean Square Error:

324035627834.29736 & R2 Score: 0.738. The model achieved the least MSE and highest R2 score making it the best trained regression model2

6. Prediction on Test Set

Finally, we predict house prices on the test dataset using the XGBoost Regressor model and the predicted price values on the test set was saved to a new csv file.