Assignment: Exploratory Data Analysis and Price Prediction

Dataset: Download the dataset **Bengaluru_House_Data**. You can find the dataset through an online search or from your instructor.

Objective: Perform Exploratory Data Analysis (EDA) and implement Machine Learning models for price prediction.

Part 1: Data Exploration and Cleaning

1. **Describe the dataset**:
    - Use the describe() function to generate statistical insights.
    - Use the info() function to understand the dataset structure.
2. **Shape of the Dataset**:
    - Determine the number of rows and columns using the shape attribute.
3. **Univariate Analysis**:
    - Identify **outliers** for numerical features using visualizations (boxplots).
    - Find the total number of **missing values** for each feature.
    - Identify the **cardinality** of categorical features (number of unique values).
4. **Handling Missing Values**:
    - Replace missing/null values in **numerical features** with the **mean** of the respective columns.
5. **Feature Reduction**:
    - Drop the **two features** with the highest number of missing values.
6. **Categorical vs Numerical Features**:
    - Identify and list all the **categorical** and **numerical** features of the dataset.

Part 2: Price Prediction

Task 7(a): Implement Machine Learning Models

1. Perform a train-test split on the dataset (e.g., 80% for training, 20% for testing).
2. Use the following models to predict the house prices:

- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest Classifier

3. Compare the accuracy of the models using evaluation metrics.

Task 7(b): Model Evaluation Using K-Fold Cross-Validation and Hyperparameter Tuning

1. Implement **K-Fold Cross-Validation** (k = 5) for the same models:
   - KNN
   - Decision Tree
   - Random Forest Classifier
2. Perform **Grid Search** to optimize hyperparameters for each model.
   - Example hyperparameters to tune:
     - KNN: Number of neighbors (n_neighbors)
     - Decision Tree: Maximum depth (max_depth), Criterion (gini/entropy)
     - Random Forest: Number of estimators (n_estimators), Maximum depth (max_depth)
3. Evaluate the models after hyperparameter tuning.

Task 7(c): Compare Performance

1. Compare the **Confusion Matrices** and **accuracy metrics** of:
   - Models without K-Fold Cross-Validation (from Task 7a).
   - Models with K-Fold Cross-Validation and Grid Search (from Task 7b).
2. Analyze how hyperparameter tuning and cross-validation improve the model performance.