Data Science Regression Project: Predicting Home Prices in Bangalore

Introduction:

The objective of this project is to develop a machine learning model that predicts home prices in Bangalore based on various features such as location, total square feet area, number of bathrooms, and number of bedrooms.

Tools Used:

- **Python**: The primary programming language used for data analysis, visualization, and modeling.

- **Pandas** A Python library used for data manipulation and analysis.

- **NumPy**: A fundamental package for numerical computing in Python, used for array manipulation and mathematical operations.

- **Matplotlib and Seaborn**: Python libraries used for data visualization to gain insights and present findings effectively.

- **Scikit-learn**: A powerful machine learning library in Python used for building and evaluating machine learning models.

- Jupyter Notebook: An interactive computing environment that facilitates code execution, visualization, and documentation, making it ideal for exploratory data analysis.

- Kaggle: An online platform where the dataset was obtained, and which provides resources and tools for data science and machine learning projects.

Dataset Description:

The dataset used for this project is obtained from Kaggle and contains information about house prices in Bangalore. It includes features such as area type, availability, location, size, total square feet, number of bathrooms, number of balconies, and price.

Steps Involved:

1. **Data Loading**:

   - Load the Bangalore home prices dataset into a pandas DataFrame using the Pandas library.

2. **Data Exploration**:

   - Examine the shape and columns of the dataset to get an overview.

   - Check unique values and value counts for the 'area_type' feature.

- Drop features that are not required for building the model.

3. **Data Cleaning**:

- Handle missing values by dropping rows with missing values in critical columns.

- Perform feature engineering by adding a new feature 'bhk' (Bedrooms Hall Kitchen).

- Clean the 'total_sqft' feature by converting range values to their average and handling other inconsistencies.

4. **Feature Engineering**:

- Add a new feature called 'price_per_sqft' to represent the price per square foot of each property.

5. **Dimensionality Reduction**:

- Apply dimensionality reduction techniques to reduce the number of unique locations by grouping less frequent locations as 'other'.

6. **Outlier Removal**:

- Identify and remove outliers based on business logic and statistical methods.

- Remove outliers related to total square feet per bedroom and price per square foot.

7. **Model Building**:

- Use linear regression from Scikit-learn to build a machine learning model for predicting home prices.

- Split the dataset into training and testing sets.

- Train the linear regression model and evaluate its performance using the test set.

- Use K-fold cross-validation to measure the accuracy of the model.

Conclusion:

In this project, we successfully developed a machine learning model to predict home prices in Bangalore based on various features. The model achieved a high accuracy score, indicating its effectiveness in predicting home prices. Further improvements and optimizations can be made to enhance the model's performance. The tools mentioned above were instrumental in carrying out the different stages of the project, from data exploration and cleaning to model building and evaluation.