Summary Of House Price EDA

Detailed Analysis of the Notebook

1. Importing Libraries

- **Libraries Used:**
    - numpy and pandas for data manipulation and analysis.
    - matplotlib.pyplot and seaborn for data visualization.
    - warnings to suppress warning messages during execution.

2. Loading the Dataset

- **Loading Data:**
    - The dataset is read into a pandas DataFrame named df.
    - Initial inspection includes displaying the first few rows and checking the DataFrame's shape.

3. Initial Data Exploration

- **Basic Checks:**
    - The notebook checks for missing values and confirms that the dataset is complete.
    - Basic statistics like minimum and maximum house prices are calculated.

4. Basic Insights

- **Price Range:**
    - House prices range from $75,000 to $7,700,000, indicating a wide variance in the dataset.

5. Exploratory Data Analysis (EDA)

- **Univariate Analysis:**
    - Histograms are plotted for individual features to understand their distributions.
    - Key findings:
        - Most houses have 2 to 4 bedrooms.
        - The majority of houses have 1 or 2 floors.
        - Few houses have a waterfront view.
        - Most houses were built between 2002 and 2005.
- **Bivariate Analysis:**
    - Relationships between features and house prices are explored using line plots.
    - Observations:
        - There are discernible patterns between features like sqft_living, grade, and house prices.
        - Features such as waterfront, view, and condition show clear correlations with house prices.

6. Feature Selection

- **Selected Features:**
  - A subset of features is chosen for modeling, including bedrooms, bathrooms, sqft_living, floors, waterfront, view, condition, grade, yr_built, yr_renovated, and zipcode.

7. Data Visualization

- **Visualizing Relationships:**
  - Various plots illustrate the relationship between selected features and house prices.
  - Histograms and line plots help identify trends and correlations.

8. Model Training and Evaluation

- **Data Preprocessing:**
  - Features and target variable (price) are separated.
  - Data is split into training (80%) and testing (20%) sets.

- **Linear Regression Model:**
  - Trained on the dataset.
  - Performance evaluated using:
    - Mean Squared Error (MSE): Measures average squared difference between actual and predicted values.
    - Mean Absolute Error (MAE): Measures average absolute difference between actual and predicted values.
    - R^2 Score: Indicates the proportion of variance in the dependent variable predictable from the independent variables.
- **Decision Tree Regressor:**
  - Trained and evaluated similarly.
  - Cross-validation is used to ensure robustness and prevent overfitting.

Key Insights

1. **Data Characteristics:**

    - The dataset is clean with no missing values.
    - House prices vary widely, reflecting a diverse real estate market.

2. **Feature Distributions:**

    - **Bedrooms:** Most houses have between 2 to 4 bedrooms.
    - **Floors:** Most houses have 1 or 2 floors.
    - **Waterfront:** Few houses have a waterfront view, suggesting it's a premium feature.
    - **Year Built:** Majority of the houses were built in the early 2000s, indicating a recent housing boom.

3. **Relationships Between Features and Price:**

    - **Waterfront Presence:** Houses with waterfront views are significantly more expensive.
    - **Condition and Grade:** Higher graded and well-maintained houses tend to have higher prices.
    - **Living Area (sqft):** Larger houses (in terms of square footage) are generally more expensive.

4. **Model Performance:**

    - **Linear Regression:**

        - Provides a good baseline for understanding relationships.
        - Metrics: MSE, MAE, and R^2 scores indicate how well the model fits the data.

    - **Decision Tree Regressor:**

        - Captures non-linear relationships better.
        - Cross-validation ensures the model generalizes well to new data.