**Interim Report: Exploratory Data Analysis (EDA) for Real Estate Pricing**

**Project Overview**

This project aims to analyze a dataset of real estate transactions with the goal of uncovering key factors that influence house prices in the residential real estate market. By performing Exploratory Data Analysis (EDA), we aim to identify patterns, trends, and correlations within the dataset that can help optimize pricing strategies and provide actionable insights for the real estate company.

**Key Objectives**

The objectives of the initial EDA phase are as follows:

1. **Data Cleaning**: Ensure the dataset is free of missing values, duplicates, and anomalies.
2. **Univariate Analysis**: Examine the distribution of key features, particularly house prices, to understand their characteristics.
3. **Multivariate Analysis**: Investigate relationships between multiple variables (e.g., LotArea, number of bedrooms) and house prices.
4. **Feature Engineering**: Create new features such as price per square foot and property age to enhance the analysis.
5. **Initial Insights**: Provide preliminary findings on which variables appear to have the most significant impact on house prices.

**Data Cleaning Process**

1. **Missing Values**:
   * The dataset contains some missing values in various columns. These were handled by filling missing numerical values with the median of the respective columns.
   * Categorical variables with missing values were either imputed with the mode or dropped, depending on the number of missing entries.
2. **Duplicates**:
   * Duplicate rows were identified and removed to ensure that the analysis is based on unique records.
3. **Anomalies**:
   * Observations with implausible values (e.g., negative house prices or square footage) were removed.
   * After cleaning, the dataset is now free of missing or erroneous values, making it ready for further analysis.

**Univariate Analysis**

1. **House Prices**:
   * The distribution of house prices shows a skewed right distribution, with a higher concentration of houses priced in the lower-to-mid-range. This suggests that a small number of high-priced properties are skewing the overall average.
   * A histogram and were used to visualize the distribution of house prices:
   * The distribution shows a typical real estate market pattern where most transactions are in the lower price brackets, with fewer high-value properties.
2. **Lot Area**:
   * A similar right-skewed distribution was observed in the lot area variable (sqft\_living), suggesting that most houses are of moderate size, with a few very large properties.
3. **Bedrooms and Bathrooms**:
   * The number of bedrooms (bedrooms) and bathrooms (bathrooms) are normally distributed, with most houses having 2-4 bedrooms and 2-3 bathrooms.
   * Visualizations were created using box plots to identify outliers in both bedrooms and bathrooms.

**Multivariate Analysis**

1. **Correlation Matrix**:
   * A correlation matrix was generated to identify how various features relate to one another, particularly with house prices.
   * **Key Findings from the Correlation Matrix**:
     + Area is highly correlated with price (positive correlation of 0.7+).
     + The number of bedrooms and bathrooms shows moderate correlations with price.
     + Features like waterfront, view, and condition also showed some interesting positive correlations with house prices.
2. **Scatter Plots**:
   * Scatter plots were used to investigate the relationship between variables such as area and sale price, as well as between the number of bedrooms and saleprice. These visualizations further confirmed that lot area and number of rooms are strong predictors of house price.

**Feature Engineering**

1. **Price per Square Foot**:
   * A new feature, price\_per\_sqft, was created to represent the price relative to the size of the property:
   * This feature provides insights into the relative value of properties of different sizes and can be used to identify underpriced or overpriced properties.
2. **Property Age**:
   * A new feature, age, was created by subtracting the year the property was built (yr\_built) from the current year (2024):
   * This feature allows us to assess the impact of a property's age on its value, as newer properties often demand higher prices.

**Challenges and Adjustments**

1. **Handling Missing Data**:
   * Initially, some features had a significant number of missing values, which required careful treatment. In cases where data imputation was not feasible, columns were dropped or replaced with reasonable default values (e.g., using the median for numerical columns).
2. **Outliers**:
   * A few extreme outliers (properties with abnormally high prices or square footage) were observed and removed to ensure that the overall analysis was not skewed.

**Next Steps**

1. **Multivariate Analysis**: Further investigate the relationships between key features such as age, price\_per\_sqft, and amenities (e.g., pool, garage) to refine pricing models.
2. **Time-Series Analysis**: If data includes temporal aspects (e.g., year sold), explore trends in housing prices over time.

**Conclusion**

In the interim phase of the project, significant progress has been made in cleaning and understanding the data. Preliminary insights into the key features influencing house prices, such as square footage, number of bedrooms, and the presence of certain amenities, have been identified. Moving forward, further analysis will refine these insights and provide actionable recommendations for pricing strategies.