Assignment 2: EDA report on bonna46/Chess-FEN-and-NL-Format-30K-Dataset

1. Dataset Selection

In this assignment, I selected the GihanPramod99/House_Price dataset from the Hugging Face Datasets Library. I originally thought of working with the bonna46/Chess-FEN-and-NL-Format-30K-Dataset, but in order to start with my first machine learning project, I thought I would give up on the more complicated datasets and stick to a simple one. I would explore the chess dataset for some future work. Price predictions on houses are quite a common and straightforward machine learning application, thus rendering the House_Price dataset an ideal starting point for beginners. It will assure me to gather plenty of examples and documentation in case I encounter challenges. The dataset is also quite easy to load and understand, allowing me to focus on building the groundwork for data exploration and analysis (and I do not need to sample).

2. EDA Process

Data Understanding:

- **Number of samples:** 4140
- **Data structure:** The dataset is a 4140 × 18 matrix, where the rows represent the number of samples, and the columns represent the data features. The key features include:
    - **Target variable:** price (float64)
    - **Metadata:** date (not included in training as it represents when the data was uploaded)
    - **Categorical features:** street, city, statezip, country (all string types)
    - **Numerical features:** bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, view, condition, sqft_above, sqft_basement, yr_built, yr_renovated (all in float64 or int64)

Visualizations - To analyze the dataset, I created multiple graphs and summaries:

- **General overview:** The df.info() summary showed that the dataset contains no null values.
- **Bar graphs:**
    - **City distribution:** There are 43 cities, all located in Washington. The most frequent city is Seattle.
    - **Country distribution:** The dataset only includes properties from the USA, making country a redundant feature.
    - **Year built:** Houses in the dataset were built between 1900 and 2014, with peaks in 2005 and 2006.
- **Histogram of house prices:**
    - The initial histogram contained outliers, including prices of $0 and one extreme value of $26 million, making the distribution less useful.

    - A second histogram, based on df.describe() statistics (mean, std, percentiles), showed a more log normal-like distribution.
- **Correlation matrix:**
    - Strongest correlations:
        - sqft_living and bathrooms (~0.75)
        - sqft_above and sqft_living (~0.8)
        - bathrooms and sqft_above (~0.7)
    - The best predictor for price is sqft_living (~0.45 correlation).

Data Quality Analysis - Using Great Expectations (GX), I identified inconsistencies and data quality issues:

- **Outliers in price:**
    - 19 instances had a price of $0, which is unrealistic.
    - One house had a price of $7,800, which is too low to be credible.
- **Duplicate addresses:**
    - Some properties, like 2500 Mulberry Walk NE, appeared multiple times, indicating duplicate entries.
- **Statezip formatting:**
    - All values start with "WA" followed by five digits (e.g., "WA 98101"). Since all properties are in Washington, we could extract only the zip code for better usability.
- **Land and living area constraints:**
    - sqft_lot is always greater than zero.
    - sqft_living is always greater than or equal to sqft_above, verifying the relationship sqft_living = sqft_above + sqft_basement.
- **Building floors:**
    - Every house has at least one floor.
- **Year built vs. renovated:**
    - If not renovated, yr_built and yr_renovated should be the same, but some inconsistencies exist.
- **Condition rating:**
    - Ratings should be between 1 and 5, which was confirmed.
- **Missing room features:**
    - Some properties had no recorded bedrooms or bathrooms, which is unusual and indicates a data quality issue.

Conclusion

This Exploratory Data Analysis (EDA) provided key insights into the dataset’s structure, feature distributions, correlations, and quality issues. The House_Price dataset is well-structured for a beginner project, though it contains some inconsistencies, including missing or unrealistic values and duplicate records. Addressing these issues will be crucial before any modeling.The use of Great Expectations helped identify and validate data constraints, ensuring a higher-quality dataset for future machine learning applications.