Cleaned Dataset Summary

I began by loading and exploring the raw dataset to understand the structure, column names, data types, and missing values. During the cleaning process, I focused on handling outliers, fixing invalid entries, and managing missing values. Using histograms and boxplots, I identified and removed rows with extreme or unrealistic values in columns such as bedrooms, bathrooms, garage, lot_acres, sqrt_ft, and sold_price. I also replaced invalid values in the year_built column (e.g., 0) with the median year to ensure accuracy. For the categorical columns with missing values — kitchen_features and HOA — I filled them with 'None' to indicate that the information wasn't available. Additionally, I converted the HOA column to numeric format, handling any non-numeric entries by converting them to NaN and then filling those with 0. To wrap up, I verified that all columns had appropriate data types and compared the number of rows before and after cleaning. A total of 180 rows were removed. These steps made the dataset cleaner and more suitable for future analysis and modeling.

7 Bullet Points – Tools and Steps I Used

- I used Pandas to load the CSV file, explore the dataset, inspect column types, and perform data cleaning.
- I used Matplotlib and Seaborn to create histograms and boxplots, which helped visualize the data distribution and detect outliers.
- I removed rows with extreme values in columns like sold_price, bedrooms, bathrooms, garage, lot_acres, and sqrt_ft.
- I corrected invalid values in year_built by replacing any 0 values with the median year from valid entries.
- I filled missing values in the kitchen_features column with 'None' to indicate no feature was listed.
- I converted the HOA column to numeric using pd.to_numeric() and filled any invalid or missing values with 0.
- I checked that the data types were consistent and compared the row count before and after cleaning to confirm that 180 rows were removed.