### **Use Case Summary**

#### Objective

The goal of this project is to develop a predictive model for Homeowners Association (HOA) fees, helping real estate professionals, buyers, and sellers estimate HOA costs based on property characteristics and location data.

#### Business Impact

* **Homebuyers & Investors:** Helps buyers factor in HOA fees when assessing the affordability of a property.
* **Real Estate Agents:** Allows agents to set better price expectations for listings and guide clients toward cost-effective properties.
* **HOA & Property Managers:** Enables HOAs to compare their fees with similar properties and adjust them based on market trends.

### Prediction of Homeowners Association Fees

Below are some of the data preprocessing steps performed on the housing dataset:

After importing the CSV data as a Pandas DataFrame, I performed data cleaning. I checked for missing values by plotting the null values. I removed 56 missing values from the **Sqft** column. I imputed missing values in the **HOA** column with ‘0’ and filled other numerical columns with ‘0’ values as well.

Next, I carried out feature engineering. I created a correlation matrix to check for multicollinearity among features. The matrix revealed that **Bedrooms** and **Bathrooms** were highly correlated, so I combined these columns into a new feature called **Rooms**. Additionally, I created a **Price/Sqft** column by dividing **Sold Price** by **Sqft**. I also derived new features—**Kitchen\_Features\_Count** and **Floor\_Covering\_Count**—by counting the respective features. Afterward, I dropped the **Bedrooms**, **Bathrooms**, **Sold\_Price**, **Sqft**, **Kitchen\_Features**, and **Floor\_Coverings** columns from the dataset.

I then performed exploratory data analysis (EDA) to examine the distribution of key features, including HOA fees.

To incorporate geospatial classification, I created **HOA\_bins** and **Price/Sqft\_bins** to categorize HOA fees based on location. I assigned **Longitude** and **Latitude** as my X features and **HOA\_bins** as my target variable. The dataset was split into training and test sets, and I trained a classification model on the training data. The model achieved an accuracy of **91%** on the training data and **88%** on the test data.

After obtaining classifications, I merged them with my cleaned dataset. I then performed additional feature engineering, selecting key features for predicting HOA values. I designated these features as **X** and set **HOA** as the target variable **(y)**. The dataset was split into training and test sets, and I applied the **KNN Regressor** to predict HOA fees. The model produced the following results:

* **Mean Absolute Error (MAE):** 606.5173
* **R² Score:** -14.6040

I also applied an **OLS Regressor** to the same dataset, but it yielded a very high MAE, indicating poor performance.

As a result, I selected the **KNN Regressor** as the final model for this dataset.