**Exploratory Data Analysis & Data Preprocessing - Housing Data**

Project Objective:

The objective of this project is to perform exploratory data analysis (EDA) and data preprocessing on the Housing dataset. The dataset contains 5,000 records and 16 features, including key attributes such as Sqft, Year\_Built, Zipcodes, No. of Bedrooms, and No. of Bathrooms. The dataset comprises 14 numerical features and 2 categorical features. A data dictionary was also created as part of the exploratory analysis.

Tools and Libraries:

I used Python libraries such as Pandas for importing and manipulating the dataset, and Matplotlib/Seaborn for visualizing the analysis.

Exploratory Data Analysis:

Initial Steps:

* Focused on the data types and null values.
* Identified 8 features with missing values, including HOA (>10%) and Sqft (>1%).
* Found that some attributes had incorrect data types.

Univariate Analysis of Numerical Features:

* Conducted a detailed analysis of individual features using box plots and histograms.

1. Lot\_Acres:
   * Right-skewed distribution with a few extreme values above 3.5 acres, extending up to 2,154 acres.
   * Approximately 90% of the data is roughly normally distributed, with a median of 0.99 and a 90th percentile value of 3.77.
   * Mean value: 4.66.
2. Sold\_Price, Sqft, Bedrooms, and Bathrooms:
   * Sold\_Price: 99% of properties are priced between $169,000 and $2M, with a maximum value of $5.3M.
   * Year\_Built: Found 5 records with 0 values. After omitting these, the distribution was left-skewed, with a median of 1997 and a maximum value of 2019.
   * Bedrooms: Transformed values greater than 8 into a new category (“>8”). Median and mode values were 4, with a maximum of 36.
   * Bathrooms: 99% of values ranged between 1 and 7. Median value: 4, maximum: 36.

Data Preprocessing:

Handling Missing Values:

1. HOA:
   * With over 10% missing values, the correlation of HOA with Sold\_Price was measured and found to be 0.17 (very mild correlation).
   * To evaluate zipcode-based imputation for HOA, US zip codes were reviewed from an online repository. It was observed that each zip code covers thousands of residential units and spans areas greater than 100 sq. miles.
   * Based on this analysis, HOA was removed due to its large proportion of missing values.
2. Sqft:
   * Missing values were imputed using the mean value, calculated by grouping data based on No. of Bedrooms and Sold\_Price.
   * Since 99% of the data in the Bedrooms column had values less than 7, I handled values greater than 8 separately. Additionally, the Sold\_Price column was binned into 4 categories to aid imputation.

Handling Incorrect Data Types:

* Corrected the data types of features including Year\_Built, Bedrooms, Bathrooms, Sold\_Price, and Garage.

Transforming Categorical Features:

* Features like Kitchen\_Features and Floor\_Covering contained lists of values, which are unsuitable for modeling. These were converted into binary columns using one-hot encoding, resulting in a total of 743 attributes.

Geo-tagging the zipcodes with City and State values

Used pyzipcode library and ZipCodeDatabase method within it to add City and State values for each record, based on the zipcode.

Addressing Data Quality Issues:

* Reviewed 0 values in Lot\_Acres and Year\_Built, which require further analysis to understand the underlying data quality issues. A suitable imputation approach will be devised.
* Some City and State values were missing (None). These will be further investigated by updating the zip code package.