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

**Project Objective:**

The objective of the project is to do exploratory data analysis and data preprocessing of the Housing dataset.

The dataset contains 5000 records and 16 features including the key features of Sqft, Year\_Built, Zipcodes, No. of Bedrooms and No. of Bathrooms. There are 14 numerical features and 2 categorical features in the dataset. I have created data dictionary.

I used Python’s libraries such as Pandas for importing the dataset, and Matplotlib/Seaborn for visualising the analysis. I began by conducting EDA by focusing on the datatypes and null values and found that 8 features including HOA (>10%) Sqft (>1) have null values. Some attributes were not in correct datatype.

Then I started the univariate analysis of numerical features.

I used box plot for checking the distribution of the values in Lot\_Acres. I found that

* Lot\_Acres is right skewed with a handful of values above 3.5 acres, extending up to 2154 acres.
* Around 90% of the data is roughly normally distributed with median at 0.99 and 90th percentile value at 3.77.
* Mean value is 4.66

I used histograms for analysing the distribution of the features ‘Sold\_Price’, ‘Sqft’, ‘Bedrooms’, ‘Bathrooms’. I found that

* 99% of the properties have Sold\_price between 169000 and 2M USD. Max. value is 5.3M USD.
* There were 5 records with 0 values in Year\_Built column. Omitting 0 values, the distribution is left-skewed, with median value of 1997. Max. value is 2019.
* Transformed variable with “>8” as value for bedrooms with value > 8. Median & Mode = 4. Max. value is 36.
* 99% of the values in Bathrooms column are between 1 and 7. Median is 4. Max. value is 36.

**Data Preprocessing**

* HOA has >10% missing values.
* To check if it is a significant attribute, its correlation with Sold\_Price was measured, which is 0.17 (very mild correlation).
* To assess if zipcode-based imputation approach would be reasonable for replacing missing HOA, US zipcodes were reviewed in an online Zip repository. Each zipcode has 1000s of residential units, and are >100 sq.miles in area, as shown below.
* Hence, HOA has been removed from the dataframe due to large missing values.

For handling missing data in Sqft column, I imputed the missing data by finding the mean value by using No. of Bedrooms and Sold\_Price. As 99% of the data in Bedrooms columns has values <7, I used those values to >8 which are greater than 8 bedrooms. And I created 4 bins in Sold\_Price column. I used those two features for finding the missing values in Sqft column.

I changed the datatypes of some of the features including Year\_Built, Bedrooms, Bathrooms, Sold\_Price, and Garage.

As the attributes ‘Kitchen\_Features’ and ‘Floor\_Covering’ have list values. As a data scientist, I can not use those for modelling. So I converted those columns to binary values using one-hot encoding. After I did this total attributes are 743.

Next I need to analyse 0 values in Lot\_Acres and Year\_Built and understand data quality issue. Devise suitable approach for imputation of the above 0 values. Some City and State values are showing as None. Need to assess reason, and update zipcode package.