Presented by Group 6
Param Andharia (202203041)
Nakshi Shah (202411063)
Rudra Patel (202201193)

# Housing Price

## IT462: Exploratory Data Analysis

**Presented To: Prof. Gopinath Panda**

# About Dataset

- ***Source:***

  -US Government Open Data Platform

  -(https://catalog.data.gov/dataset/real-estate-sales-2001-2018)

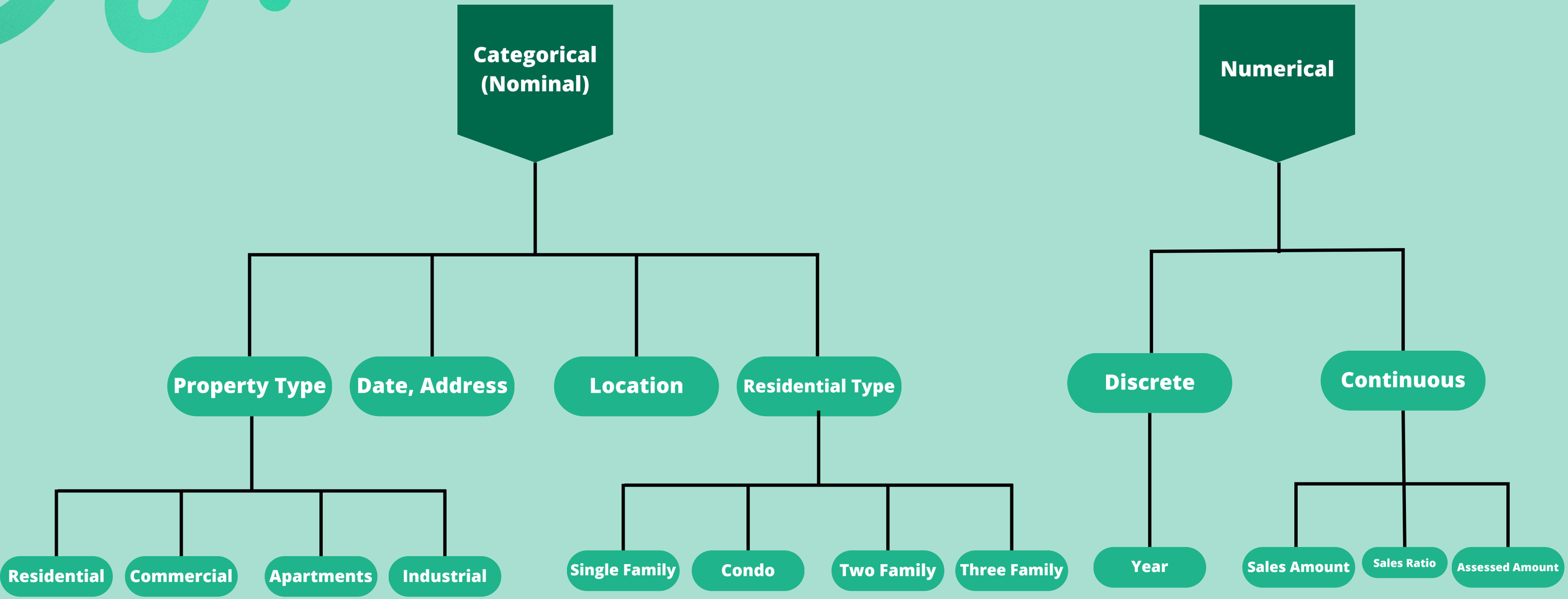
- ***Total Records:***

  -1,048,575

- ***Objective:***

  -Analyze trends and predict sale amounts of real estate properties.

# Dataset Attibutes



# Data Preprocessing

## (i) Hangling Missing Values

- **Deleted Columns:**

Columns with more than 50% missing values (OPM Remarks, Assessor Remarks, Location, Non Use Code) were removed as they were not useful for the analysis.

## (ii) Grouping And Cleaning

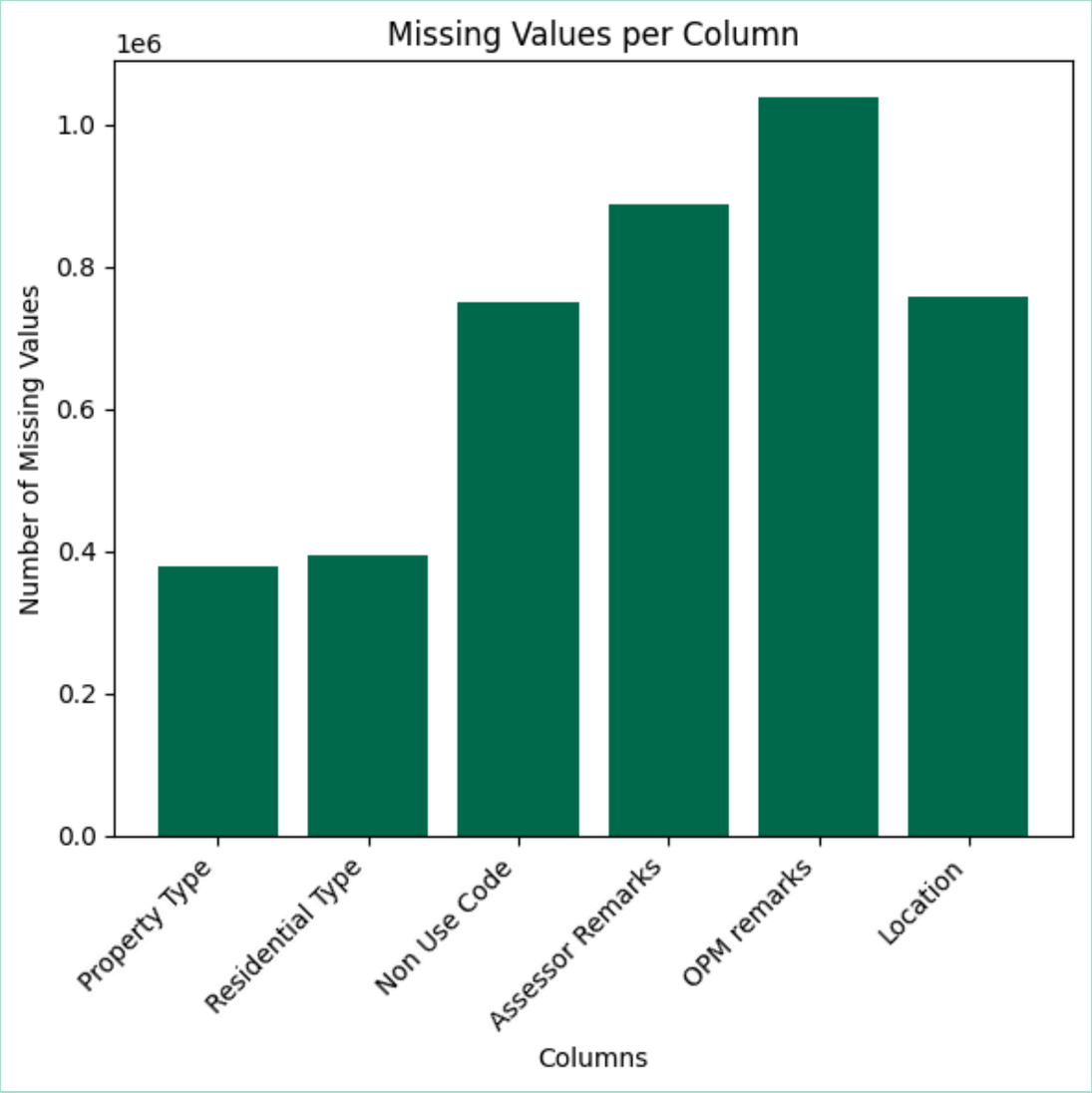
- **Dividing the Dataset:**

Residential Type had missing values when Property Type was non-residential. So deleting those would result in deleting all non-residential data, which is not useful.

- **Solution:**

a. Residential Housing: Contains data with Property Type as residential.

b. Non-Residential Housing: Contains data with Property Type as non-residential.



# Data Preprocessing

## (iii) Cleaning Categorical Columns

- **Problem:**

  - The Address column had a large number of unique values (521,395), making encoding impractical for such a large dataset.

- **Solution:**

-Simplified addresses by extracting only the street name (e.g., 1963 MAIN ST → MAIN ST).

- **Result:**

-Reduced unique values by approximately 80%.

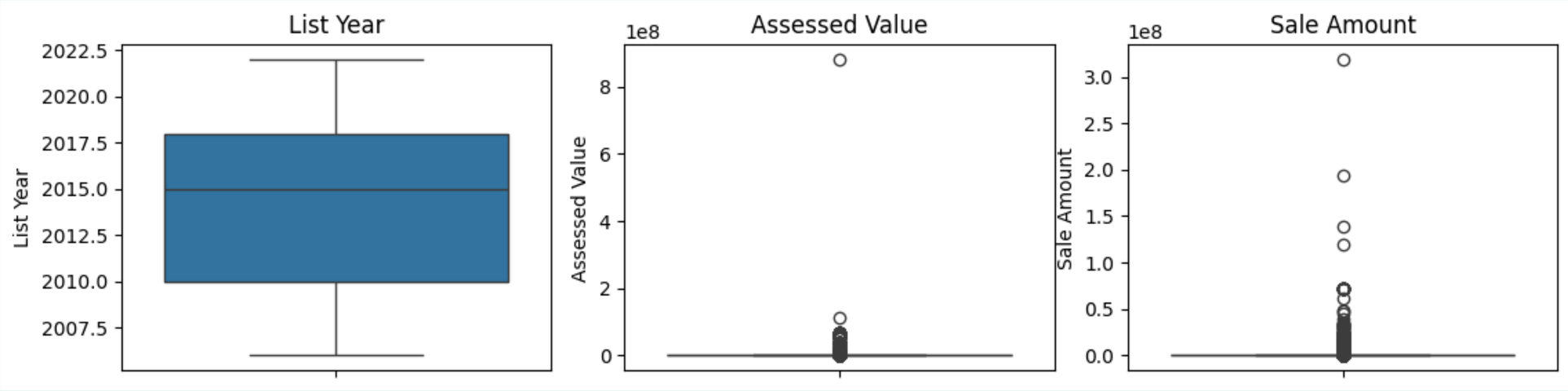
-Unique values in Address after cleaning:

  - Residential Data: 91,807
  - Non-residential Data: 7,948

# (iv) Outlier Detection and Removal

As we have now splitted our data into two datasets, we will have to apply the same preprocessing techniques to both the datasets to maintain consistency.

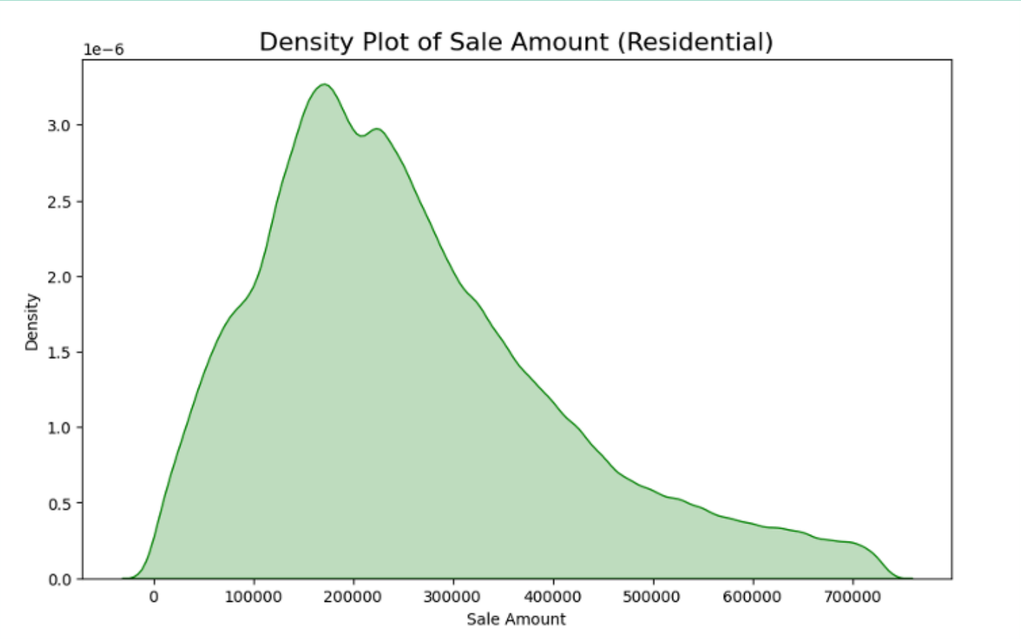
By Using IQR method for Outlier Detection, we found 54506 outliers among the 653902 rows in the residential dataset. So only 8.33% of rows were removed.



# Plots and Distribution After Preprocessing

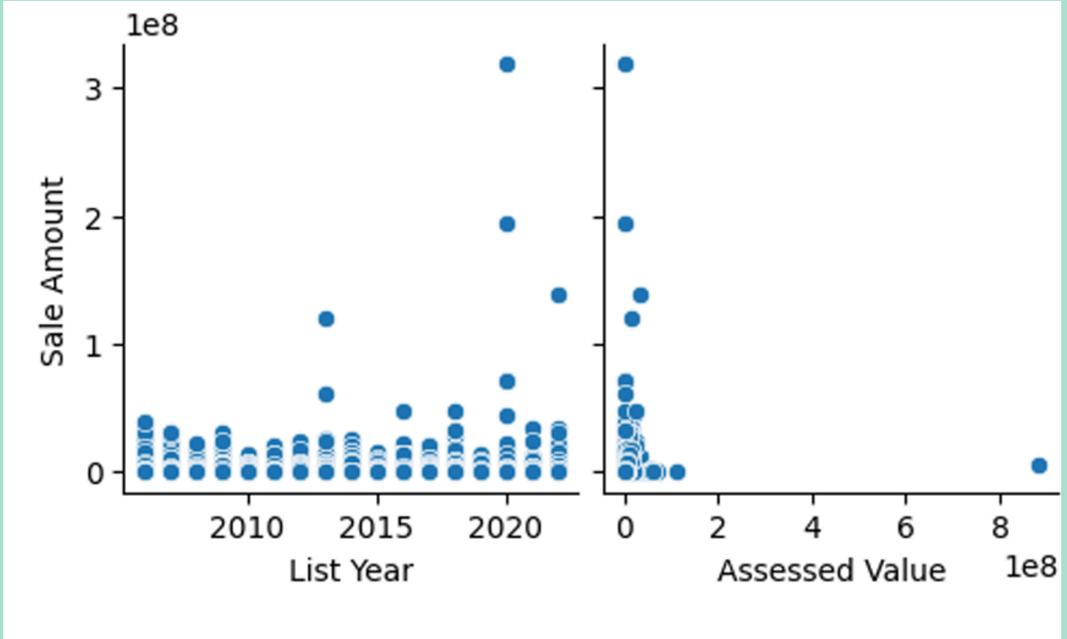
## (i) Estimated PDF of the Target Variable

- Right Skewed, most sales concentrated in range of 100,000$ to 300,000$
- Even after Outlier removal by IQR, some seem to be present beyond $700,000.




## (ii) Pair Plot of Target variable with other Numerical Features

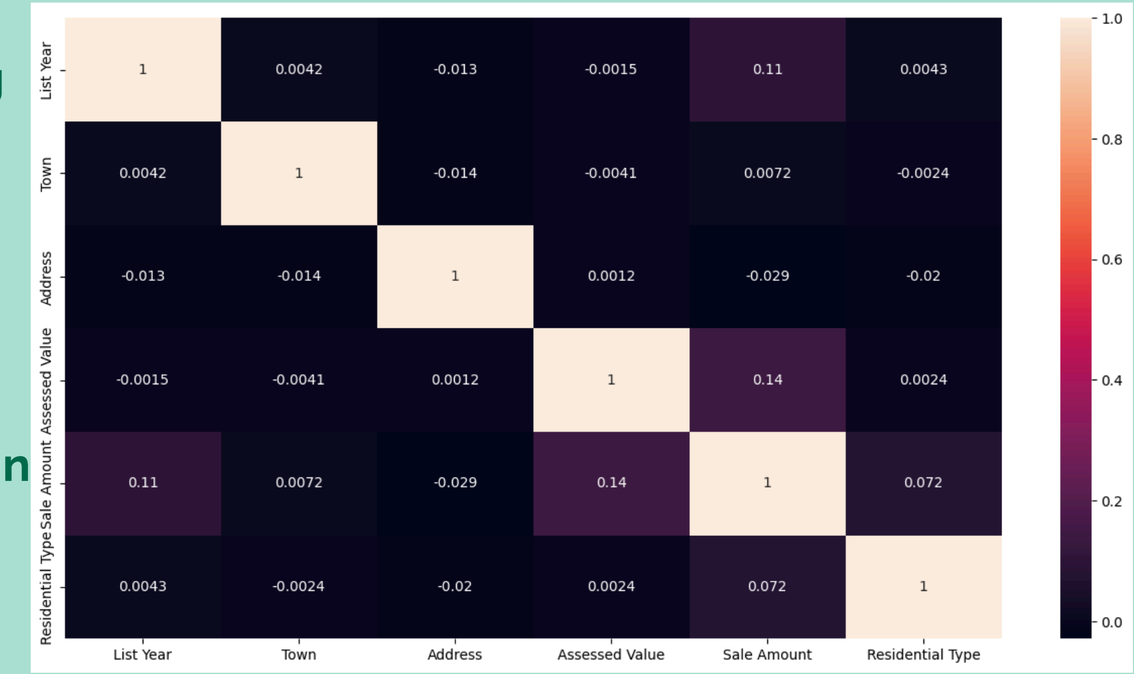
Most sale amounts are concentrated at lower values, indicating the majority of transactions are within a specific range over all years.



# Feature Selection

Due to a very low correlation with the target variable, “Sales Amount,” we needed to handle this step manually.

- The “Address” column was preprocessed and transformed into “Street Name” for improved encoding and enhanced correlation.
- The “Sales Ratio” column contained data regarding the ratio of “Assessed Value” to “Sales Amount.” Including this in our final dataset, alongside “Assessed Value,” would simplify predictions for “Sales Amount,” eliminating the need for a model. Therefore, this column was removed from the features used in predictions.
- - In the Residential dataset, the “Property Type” column was unnecessary since all entries were Residential, so it was discarded. Similarly, in the Non-Residential dataset, all values in the “Residential Type” column were null, leading to its removal as well.



# Model Fitting

- **Linear Regression:**

***Purpose:*** Used as a baseline model to establish a simple relationship between predictor variables and the target variable (Sale Amount).

***How it Works:*** Assumes a linear relationship, fitting a line that minimizes the sum of squared errors between predicted and actual Sale Amount values.

***Target Variable:*** Sale Amount, predicted based on features like Assessed Value and Property Type.

- **Gradient Boosting:**

***Purpose:*** An ensemble technique that builds multiple decision trees sequentially, with each tree correcting the errors of the previous one.

***How it Works***: Combines weak models (trees) into a strong model by focusing on areas where previous models made errors.

***Target Variable:*** Sale Amount, predicted by combining the corrections from each tree.

# Model Fitting

- **XGBoost:**

***Purpose:*** An optimized version of Gradient Boosting that improves speed, accuracy, and overfitting handling.

***How it Works:*** Utilizes regularization and parallel processing to reduce overfitting and improve performance, particularly in large datasets.

***Target Variable:*** Sale Amount, predicted with high accuracy using decision trees built iteratively with enhanced processing.

- **Random Forest:**

***Purpose:*** An ensemble model that averages the results of multiple decision trees to reduce overfitting.

***How it Works***: Builds decision trees using random subsets of the data and averages their predictions to make the final result more robust.

***Target Variable:*** Sale Amount, predicted by aggregating the outputs of many trees to make the final prediction more stable and accurate.

# Model Comparison & Analysis

**Linear Regression:** Simple but performs poorly with complex datasets, as indicated by the low R-squared value.

**Gradient Boosting:** Performs better than Linear Regression by iteratively correcting errors, but still has limitations in capturing all patterns in the data.

**XGBoost:** Performs well with high accuracy, handling large datasets efficiently. It improves upon Gradient Boosting with faster computation and regularization.

**Random Forest:** Performs the best, capturing the most variance in the target variable. It minimizes overfitting and effectively handles large datasets with diverse feature types.

| Model | 80-20 Split $R^2$ | 80-20 Split MSE | 70-30 Split $R^2$ | 70-30 Split MSE |
|---|---|---|---|---|
| Linear Regression | 0.0381 | 21,737,995,103.50 | 0.0405 | 21,622,571,538.02 |
| Gradient Boosting | 0.7310 | 6,080,061,233.61 | 0.7303 | 6,078,494,772.18 |
| XGBoost | 0.7662 | 5,282,967,511.96 | 0.7646 | 5,304,788,282.50 |
| Random Forest | 0.7685 | 5,231,274,192.41 | 0.7646 | 5,306,053,404.58 |

# Main challenges identified

01.

***Separating the Dataset and performing each step twice:***
**Because of dividing the dataset into Residential and Non-Residential data, we had to perform each preprocessing and cleaning and model fitting step twice, on each of them.**

02.

***Low Inter-variable Correlation:***
Weak correlation between features made it challenging for models to form strong relationships with the target variable (Sale Amount).

03.

***Model Complexity:***
The dataset's complexity, including diverse property types, required advanced ensemble models like Random Forest and XGBoost, which better handled non-linear relationships.

# Conclusion

- **The study analyzed real estate transactions to identify factors influencing sale prices and predict them effectively.**
- **Key predictors included assessed value, property type, and square footage.**
- **Random Forest was the best-performing model, capturing the most variance, while Linear Regression struggled with the dataset's complexity.**
- **Challenges such as low inter-variable correlation and dataset complexity were addressed using advanced ensemble models like Gradient Boosting and XGBoost.**
- **Future work involves incorporating more features (e.g., proximity to amenities) and conducting localized analyses for better insights.**

Presented by Group - 6

# Thank you very much!