# Dynamic Housing Insights: Unveiling Trends and Predictive Modeling in Real Estate Markets

Akhil Kumar Reddy Bhoomi Reddy - 02080263
Shivakumar Pasem - 02077631
Ben George Samuel - 02078919
Aishwarya Ganesh Bhat - 02097374

1. **The issues**

   The data used in this study was collected from the City of Boston's open data hub. The dataset we used is called "Property Assessment for year 2023". It gives property, or parcel, ownership together with value information, which ensures fair assessment of Boston taxable and non-taxable property of all types and classifications. Within the context of exploring property values, our investigation aims to shed light on specific challenges and nuances embedded in the dataset.

   - Do property values and areas potentially impact the overall representativeness of the dataset?
   - Does value distributions indicate a concentration of higher-value properties, potentially influencing the overall perception of property values?
   - Does the dataset span a wide range of construction years, with significant peaks in certain decades, presenting challenges in capturing historical trends and understanding their impact on property values?

2. **Findings**

   Upon careful examination of Property Assessment for year 2023, our analysis uncovers substantial patterns and trends within the dataset.Location features (city) are crucial in determining property values, emphasizing the importance of spatial factors in the real estate market.Strong correlations exist between property size (living area, gross area) and various value metrics (land, building, and total values).The dataset spans a wide range of construction years, with significant development in the 1900s and a peak in renovations during the 1980s. Year of construction/remodeling has limited correlation with property values, emphasizing the dominance of property size and features in valuation. Skewness in value distributions suggests a concentration of higher-value properties, contributing to a varied real estate landscape.

3. **Discussion**

   The dataset analysis reveals substantial variations in property values across cities, with Readville topping the chart at an average property value of around $5.65 million, signaling a high-end market or the presence of outliers. Boston follows closely at approximately $3.44 million, showcasing its premium real estate status. Histograms and correlation analyses unveil intricate patterns, demonstrating a strong positive correlation between property size and values, while the decision tree regressor model achieves a high explanatory power of approximately 97.62%, with a Mean Squared Error (MSE) of around $3.77 trillion on the test set. The feature importance plot highlights the critical role of location, property size, and specific features such as air conditioning and exterior finish in determining property values.

   Furthermore, the dataset spans construction years from as early as 1700, with a mean construction year of 1926 and a significant portion of buildings erected in the 1900s. The analysis of renovation data peaks in the 1980s, with over 104,000 properties undergoing remodeling. Descriptive statistics provide insights into property characteristics, such as average land, building, and total property values of approximately $376,586, $1,120,689, and $1,500,262, respectively. The decision tree model, while powerful, indicates room for improvement, particularly in addressing outliers, as evidenced by the Root Mean Squared Error (RMSE) of approximately $81,394.79. Overall, this comprehensive analysis not only delves into the nuances of property value determinants but also offers practical insights for stakeholders navigating the diverse and dynamic housing market.

**APPENDIX A: METHOD**

The PROPERTY ASSESSMENT FY2023 dataset, consisting of 180,623 entries, was obtained for in-depth analysis. The dataset exhibits consistent reporting across all examined columns, providing a comprehensive representation of property characteristics for the fiscal year 2023.

Descriptive statistics were computed to gain insights into the properties within the dataset. The average values for land, building, and total property were found to be approximately $376,586, $1,120,689, and $1,500,262, respectively. Notably, the dataset displays a wide range in property values, as evidenced by high standard deviations. The average living and gross areas were observed to be 3,814 and 4,763 square feet, respectively, with a predominant year of construction around 1928.

A significant skewness in the distribution of property values was observed, with median values for land, building, and total property notably lower than their respective means. This skewness indicates a concentration of higher-value properties within a subset of the dataset. Similar skewness was identified in property sizes, emphasising a prevalence of smaller-sized properties.

The Random Forest Regression model was employed to predict property values. This ensemble learning technique, consisting of multiple decision trees, was trained on a subset of the dataset. Hyperparameter tuning was executed to optimise the model's performance, allowing it to capture complex relationships within the data.

In addition to Random Forest Regression, a Gradient Boosting (GBoost) model was utilized for property value prediction. GBoost sequentially builds decision trees to correct errors from previous iterations. The model underwent training and fine-tuning to enhance its ability to capture intricate patterns within the dataset.

The performance of both the Random Forest Regression and GBoost models was evaluated using standard regression metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. These metrics provided a comprehensive understanding of the models' ability to generalize to unseen data and accurately predict property values.

**APPENDIX B: RESULTS**

Average value per city

```
CITY
READVILLE           5.646300e+06
BOSTON              3.442452e+06
ROXBURY CROSSING    1.955412e+06
ALLSTON             1.401336e+06
BROOKLINE           1.363339e+06
CHARLESTOWN         1.317292e+06
SOUTH BOSTON        1.168539e+06
BRIGHTON            1.134097e+06
EAST BOSTON         1.010138e+06
JAMAICA PLAIN       9.891430e+05
ROXBURY             9.262875e+05
DORCHESTER          8.254037e+05
WEST ROXBURY        7.354985e+05
ROSLINDALE          6.834878e+05
NEWTON              6.295000e+05
MATTAPAN            6.268352e+05
HYDE PARK           6.173737e+05
CHESTNUT HILL       6.004706e+05
DEDHAM              4.827500e+05
Name: TOTAL_VALUE, dtype: float64
```

.Based on the histograms provided, here are some observations regarding the distribution of various variables in this dataset:

NUM_BLDGS: The distribution is highly skewed to the right, with the majority of properties having a low number of buildings, and very few properties having a high number of buildings.

RES_FLOOR & CD_FLOOR (Residential Floors & Commercial Floors?): Both have a majority of the count at the lower end, indicating that most properties have a small number of floors.

RES_UNITS & COM_UNITS (Residential & Commercial Units?): These histograms show a right-skewed distribution, similar to NUM_BLDGS, indicating that most properties have few units, while a small number of properties have many.

RC_UNITS: This variable might represent some type of unit count, and it's heavily skewed, with almost all properties having a value of zero or close to zero.

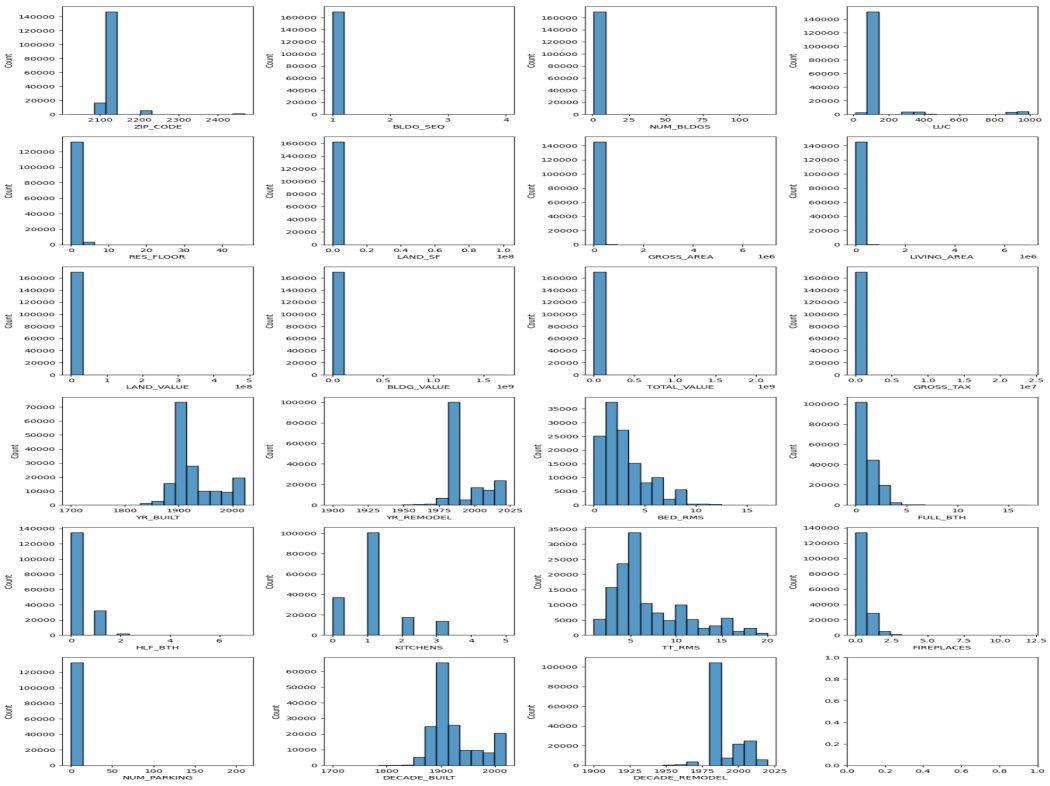
LAND_SF (Land Square Footage?): The distribution is right-skewed, suggesting that most properties have a smaller land area, with a few properties having a very large area.

GROSS_AREA & LIVING_AREA: Both are right-skewed, indicating that most properties have smaller areas with a few properties having much larger areas.

LAND_VALUE, BLDG_VALUE, TOTAL_VALUE: All three variables show a highly right-skewed distribution, indicating that most properties are on the lower end of value, with a few very high-value properties.

GROSS_TAX: This is also right-skewed, showing that most properties pay a lower amount of tax, with a few properties paying significantly more.

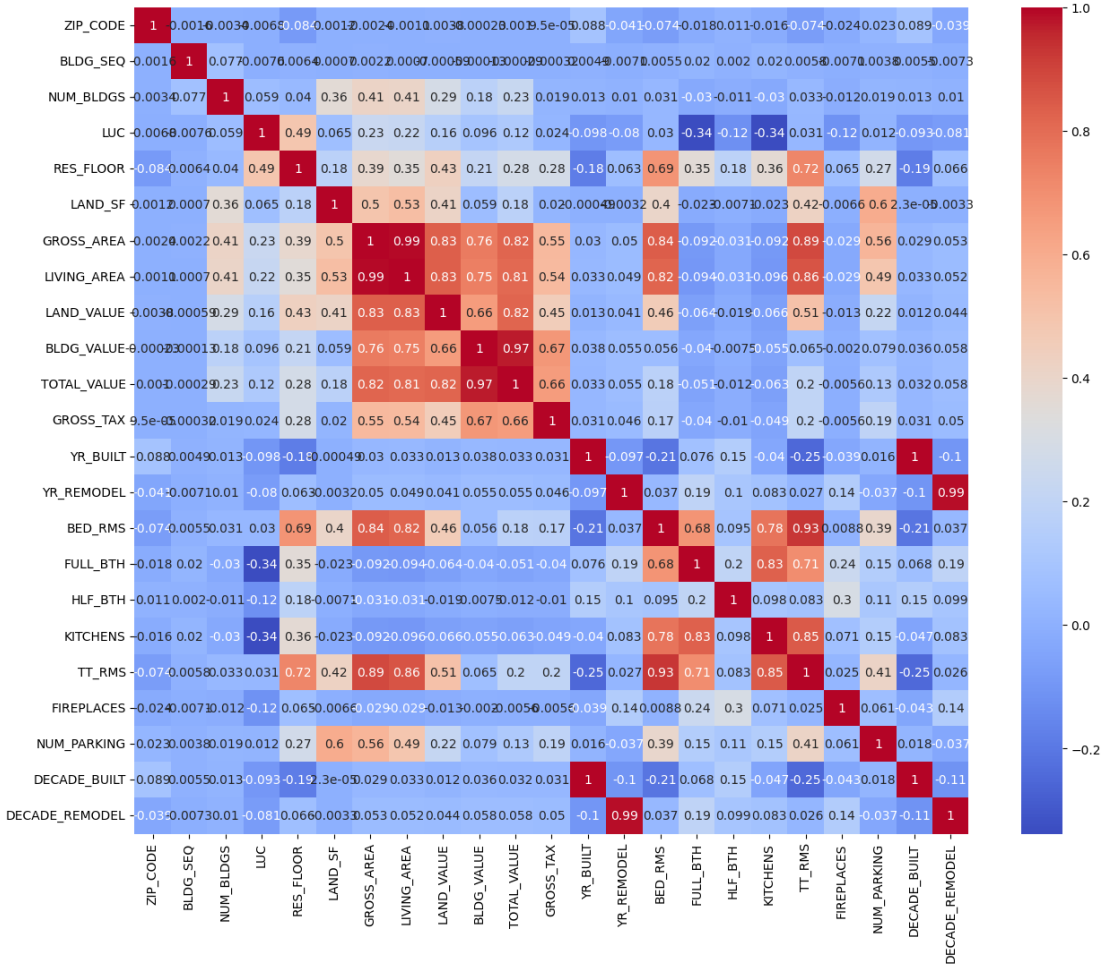
YR_BUILT & YR_REMODEL: These variables show a multimodal distribution, indicating that there are peaks at certain years which could correspond to periods of development or renovation booms. BED_RMS, FULL_BTH, HLF_BTH (Bedrooms, Full Baths, Half Baths): All are slightly right-skewed but show a more uniform distribution for lower counts, indicating a common range within which most property counts fall. KITCHENS: Most properties seem to have one kitchen, with very few having more than one. TT_RMS (Total Rooms?): Appears right-skewed, indicating that most properties have a lower total room count, with fewer properties having a very high count. FIREPLACES: This variable is heavily right-skewed, with most properties having no fireplace and a few having one or more. NUM_PARKING (Number of Parking Spaces): Also right-skewed, with the majority of properties having a low number of parking spaces.



High Correlation between Area and Value Metrics: There's a strong positive correlation between GROSS_AREA, LIVING_AREA, LAND_VALUE, BLDG_VALUE, and TOTAL_VALUE. This suggests that as the size of a property increases, its assessed land and building value, as well as the total value, tend to increase. This is expected as larger properties usually have higher values.

Number of Units and Value: There are notable positive correlations between the number of residential units (RES_UNITS) and commercial units (COM_UNITS) with TOTAL_VALUE, LAND_VALUE, and BLDG_VALUE, which indicates that properties with more units are likely to be valued higher.
Negligible Correlation with Year Built and Remodel: YR_BUILT and YR_REMODEL show very low correlations with other variables, suggesting that the year a building was constructed or remodeled has little linear association with its size or assessed value in this dataset.

Tax and Value Correlation: GROSS_TAX is strongly correlated with TOTAL_VALUE, LAND_VALUE, and BLDG_VALUE, which is expected since tax is often calculated based on the assessed value of the property.
Number of Parking Spaces: NUM_PARKING has a moderate correlation with LIVING_AREA and GROSS_AREA, indicating that properties with larger living and total areas tend to have more parking spaces.
Fireplaces: FIREPLACES has a very low to negligible correlation with most variables, suggesting that the number of fireplaces in a property doesn't have a strong linear relationship with its size, value, or the number of rooms.

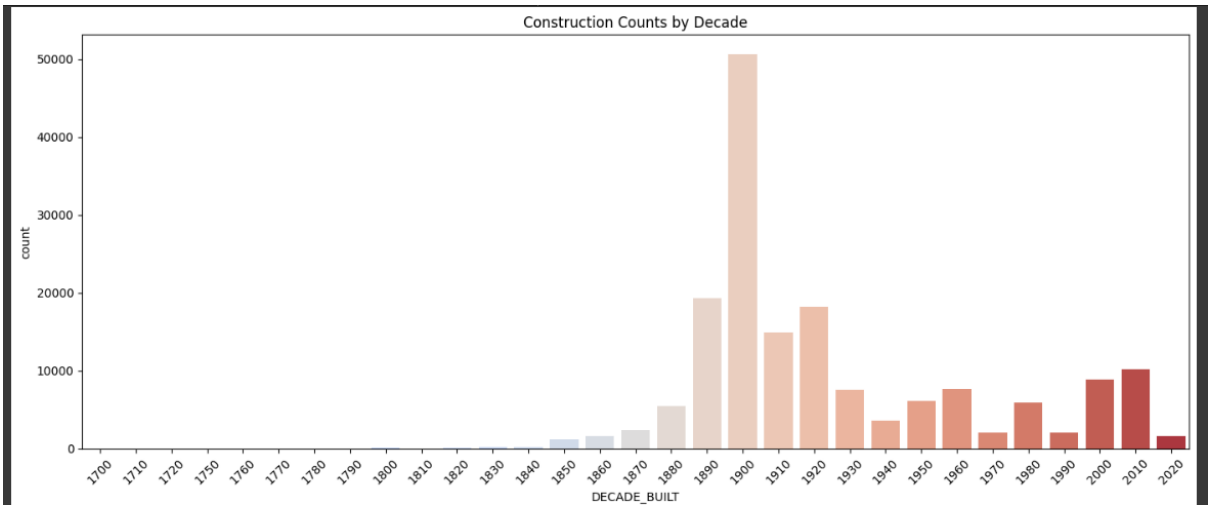


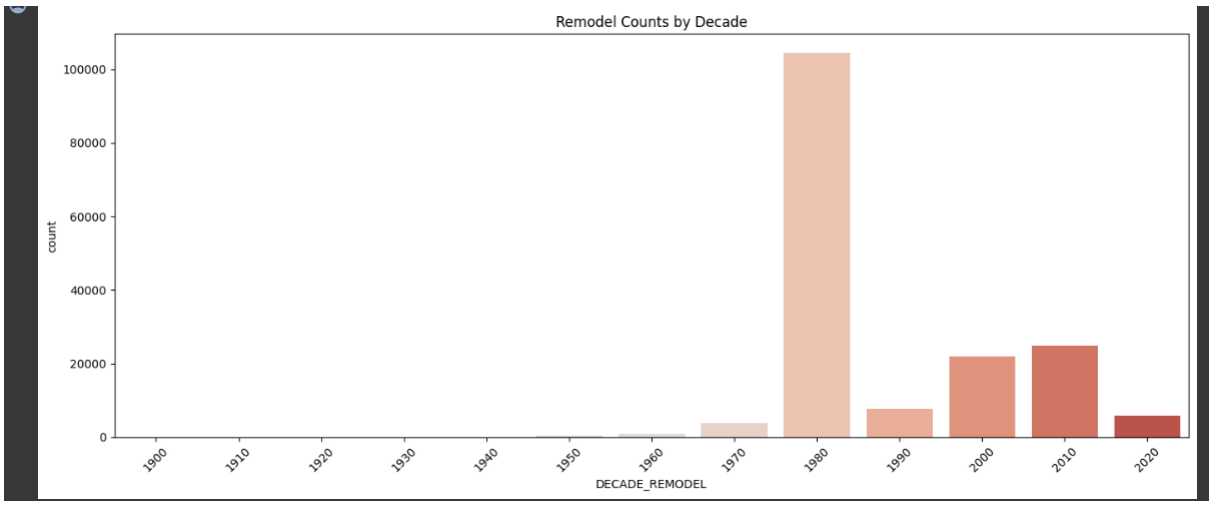
```
Descriptive Statistics for YR_BUILT:
count    170072.000000
mean       1926.912996
std          41.674859
min        1700.000000
25%        1900.000000
50%        1910.000000
75%        1950.000000
max        2022.000000
Name: YR_BUILT, dtype: float64

Descriptive Statistics for YR_REMODEL:
count    170072.000000
mean       1993.110465
std          13.407092
min        1900.000000
25%        1985.000000
50%        1985.000000
75%        2004.000000
max        2022.000000
Name: YR_REMODEL, dtype: float64

Skewness and Kurtosis for YR_BUILT:
Skewness: 0.9144669902491912
Kurtosis: -0.13604427294246202

Skewness and Kurtosis for YR_REMODEL:
Skewness: 0.6530503971368538
Kurtosis: -0.12532935705133275
```






The Decision Tree Regressor model has been successfully trained and evaluated. Here are the results:

The Mean Squared Error (MSE) of the model on the test set is approximately 3,765,824,030,907.86. This value indicates the average squared difference between the predicted values and the actual values. In this context, this number is quite large, which could suggest that there are some predictions with significant errors. However, without the context of the scale of TOTAL_VALUE, it's hard to assess the severity of this MSE value.

The R-squared ($R^2$) value is 0.9762, which is close to 1. This indicates a very high level of fit to the data, meaning that the model explains a large portion of the variability in the target variable. An $R^2$ value of 0.9762 implies that approximately 97.62% of the variation in TOTAL_VALUE can be explained by the model's inputs.

```
mse : 3765824030907.8647
r2 : 0.9762394427390666
```

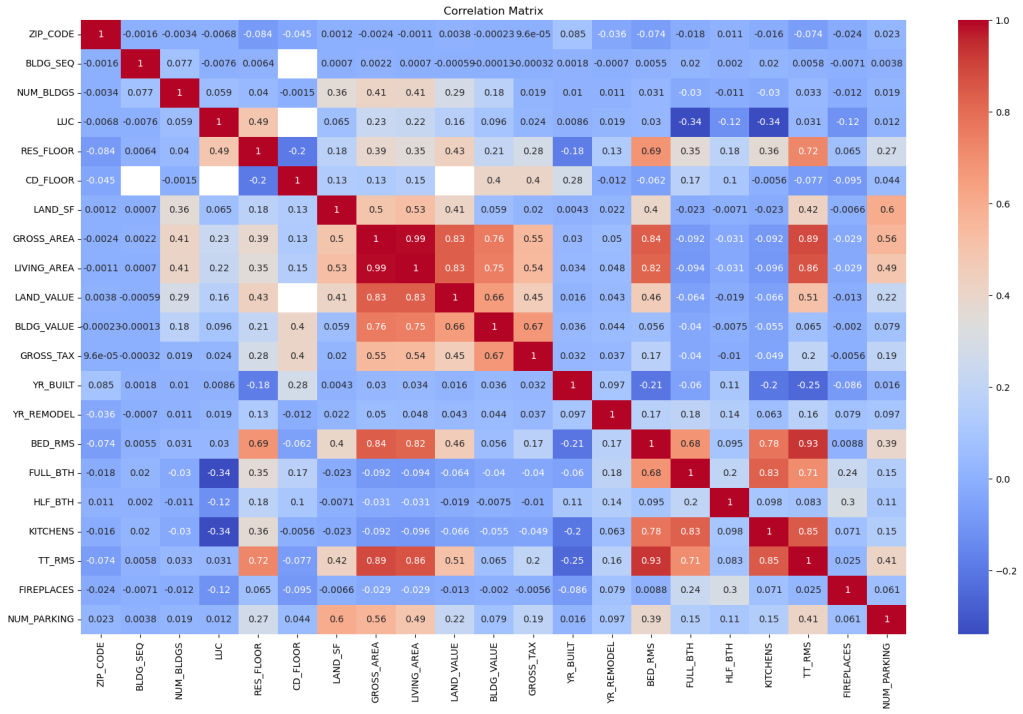
Mean Squared Error (MSE): This value represents the average squared difference between the actual and predicted values. A lower MSE indicates better model performance. In this case, the MSE seems quite large, which might be influenced by the scale of the TOTAL_VALUE variable. If TOTAL_VALUE is typically a large number (for instance, in a dataset where values represent property prices), a high MSE can occur. It's important to interpret MSE in the context of the scale of this data. R-squared ($R^2$): This is a statistical measure that represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. An $R^2$ of 0.9701 means that approximately 97.01% of the variance in this target variable can be explained by the model. This is generally considered a very high $R^2$ value, indicating that this model fits the data well.

```
MSE (cleaned): 8727514004492.355
R-squared (cleaned): 0.9691378537760184
```

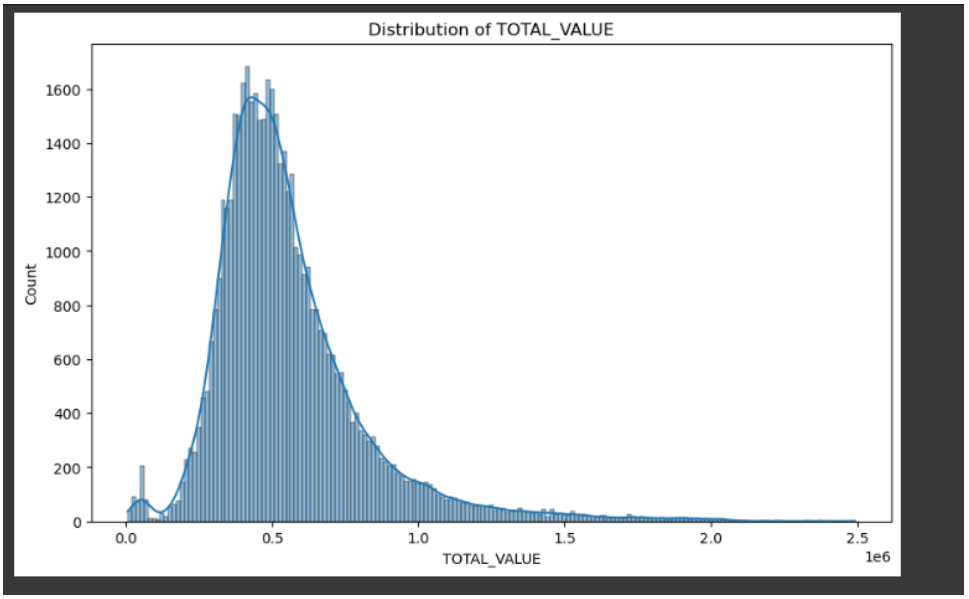
High Positive Correlation (Red Squares): There are several variables that show a strong positive correlation with each other. For example, GROSS_AREA, LIVING_AREA, LAND_SF (land square footage), and BLDG_VALUE (building value) all have strong positive correlations with each other. This suggests that as one of these variables increases, the others tend to also increase, which makes intuitive sense since larger areas would generally have higher values.

Negative Correlation (Blue Squares): There are fewer examples of strong negative correlations in this dataset. However, there are some variables with moderate negative correlations, such as RES_FLOOR with ZIP_CODE and CD_FLOOR with ZIP_CODE.

Weak or No Correlation (White Squares): Many variable pairs have correlations close to zero, indicating no linear relationship between them. For instance, NUM_BLDS (number of buildings) shows little to no correlation with ZIP_CODE




Distribution of Total value



MAE (Mean Absolute Error): On average, the model's predictions are about $50,650.80 off from the actual property values. This gives you an idea of the average error in the predictions.

RMSE (Root Mean Squared Error): The average magnitude of the error is $81,394.79. Since RMSE is more sensitive to outliers than MAE, this higher value suggests that there may be some larger errors in predictions skewing the results.

R-squared: An $R^2$ of 0.9024 means that approximately 90.24% of the variance in the property value is predictable from the features. This is a high $R^2$ value, suggesting that the model explains a large portion of the variability in the target variable.

```
MAE: 50650.79607820274
RMSE: 81394.79297791979
R-squared: 0.9023758663869607
```

**APPENDIX C: CODE**

```
# Load the dataset
data = pd.read_csv('/content/fy2023-property-assessment-data.csv')
# Basic information
print("Dataset Shape:", data.shape)
print("\nData Types:\n", data.dtypes)
# Descriptive Statistics
print("Descriptive Statistics for YR_BUILT:")
print(data['YR_BUILT'].describe())
print("\nDescriptive Statistics for YR_REMODEL:")
print(data['YR_REMODEL'].describe())
# Distribution Analysis
print("\nSkewness and Kurtosis for YR_BUILT:")
print(f"Skewness: {data['YR_BUILT'].skew()}")
print(f"Kurtosis: {data['YR_BUILT'].kurt()}")
print("\nSkewness and Kurtosis for YR_REMODEL:")
print(f"Skewness: {data['YR_REMODEL'].skew()}")
print(f"Kurtosis: {data['YR_REMODEL'].kurt()}")
# Decade or Period Analysis
data['DECADE_BUILT'] = (data['YR_BUILT'] // 10) * 10
data['DECADE_REMODEL'] = (data['YR_REMODEL'] // 10) * 10
decade_built_counts = data['DECADE_BUILT'].value_counts().sort_index()
decade_remodel_counts = data['DECADE_REMODEL'].value_counts().sort_index()
print("\nConstruction Counts by Decade:")
print(decade_built_counts)
print("\nRemodel Counts by Decade:")
print(decade_remodel_counts)
# Time Series Analysis
plt.figure(figsize=(14, 6))
sns.countplot(x='DECADE_BUILT', data=data, palette="coolwarm")

# Correlation with Value
value_correlation_built = data['YR_BUILT'].corr(data['TOTAL_VALUE'])
value_correlation_remodel = data['YR_REMODEL'].corr(data['TOTAL_VALUE'])
print(f"\nCorrelation between YR_BUILT and TOTAL_VALUE: {value_correlation_built}")
```

```
print(f"Correlation between YR_REMODEL and TOTAL_VALUE: {value_correlation_remodel}")
# Outliers Detection
sns.boxplot(x=data['YR_BUILT'])
plt.title('Boxplot for Year Built')
plt.show()
# Comparison of Distributions
plt.figure(figsize=(14, 6))
sns.histplot(data['YR_BUILT'], color="skyblue", label='YR_BUILT', kde=True)
sns.histplot(data['YR_REMODEL'], color="red", label='YR_REMODEL', kde=True)
plt.legend()
plt.title('Comparison of Distributions for Year Built and Remodeled')
plt.show()
# Renovation Rate
renovation_rate = data['YR_REMODEL'].notnull().sum() / len(data)
print(f"\nRenovation Rate: {renovation_rate:.2%}")
# Calculate the correlation matrix for the entire DataFrame
correlation_matrix = data.corr()
# For LAND_VALUE
land_value_correlations = correlation_matrix['LAND_VALUE'].drop(labels=['LAND_VALUE',
'BLDG_VALUE', 'TOTAL_VALUE','GROSS_TAX']).sort_values(ascending=False)
# For BLDG_VALUE
bldg_value_correlations = correlation_matrix['BLDG_VALUE'].drop(labels=['LAND_VALUE',
'BLDG_VALUE', 'TOTAL_VALUE','GROSS_TAX']).sort_values(ascending=False)
# For TOTAL_VALUE
total_value_correlations = correlation_matrix['TOTAL_VALUE'].drop(labels=['LAND_VALUE',
'BLDG_VALUE', 'TOTAL_VALUE','GROSS_TAX']).sort_values(ascending=False)
# Display the most correlated fields with each of the three variables
print("Most correlated fields with LAND_VALUE:")
print(land_value_correlations)
print("\nMost correlated fields with BLDG_VALUE:")
print(bldg_value_correlations)
print("\nMost correlated fields with TOTAL_VALUE:")
print(total_value_correlations)
from sklearn.metrics import mean_absolute_error, r2_score
true_values = np.expm1(y_test)
# predictions = np.expm1(xgb_log_predictions)
# Calculate MAE and RMSE
mae = mean_absolute_error(true_values, predictions)
rmse = np.sqrt(mse)  # where mse is mean_squared_error(true_values, predictions)
# Calculate R-squared
r2 = r2_score(true_values, predictions)
print(f'MAE: {mae}')
print(f'RMSE: {rmse}')
print(f'R-squared: {r2}')
```

**Contributions:**
All contributed equally

**Google Collab:**
[AMS_Project3.ipynb - Colaboratory (google.com)]