
# Predicting Residential Property Values in Boston using Machine Learning

## ALY6140: Python & Analytics Technology

**Professor Daya Rudhramoorthi**

**Prepared by: Group 7 - Sweekruti Narendra Singh & Zaili Gu**

**Northeastern University | ALY6140 | Jun. 23, 2025**

# Background Information

**Why Boston Property Valuation Matters:**

- Residential property values impact taxation, equity, and urban policy.
- Boston's assessed values affect public finance, housing affordability, and development planning.
- The City of Boston provides a rich, open dataset through its FY2024 Property Assessment database.
- This project applies data science and machine learning to uncover patterns and improve value prediction.

# Research Questions

1. What are the key structural and geographic features that influence residential property value in Boston?
2. Can machine learning models accurately predict assessed property value?
3. Are there geographic patterns or inequities in property valuation across neighborhoods or ZIP codes?
4. Do remodeling features (kitchens and bathrooms) correlate with higher property values?

# Dataset Overview

**Dataset – Boston FY2024 Property Assessment.**

- Source: Boston.gov Open Data Portal
- Original Size: 182,242 records × 66 features
- Filtered to ~94,000 residential records based on land use classification

**Includes:**

- Parcel ID, address, ZIP code
- Building type, year built, condition
- Assessed land & building values
- Kitchen/bathroom remodeling indicators

# Data Cleaning & Feature Engineering

- Filtered for residential properties based on LU (land use) field
- Handled missing values using imputation (numeric: median; categorical: mode)
- Removed outliers using IQR filtering (e.g., top 1% in total value and living area)
- Engineered new features:
  - property_age from year_built
  - Binary indicators for remodeled kitchens/bathrooms
  - price_per_sqft = total value ÷ living area
- One-hot encoding for categorical variables (e.g., ZIP code, building style)
- Normalization of numeric features using StandardScaler.

# EDA – Value Distribution

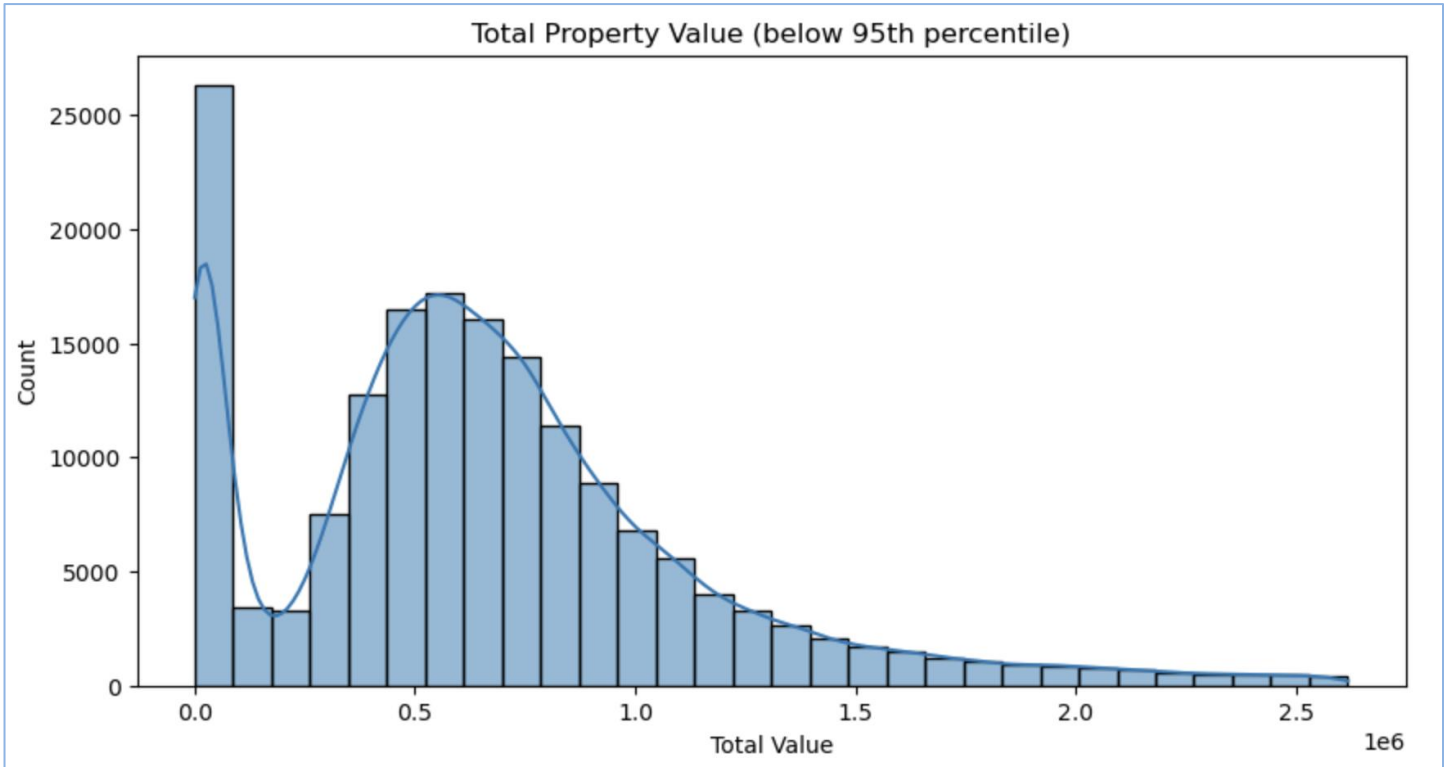
**Distribution of Property Values:**

• The total assessed value is right-skewed, with most homes valued under $1 million

• A large spike near $0 likely reflects tax-exempt or anomalously coded properties

• Data was filtered to the 95th percentile for clearer analysis

• This distribution supports using non-linear models and log transformation in modeling



# EDA – Land Use & ZIP Code Impact

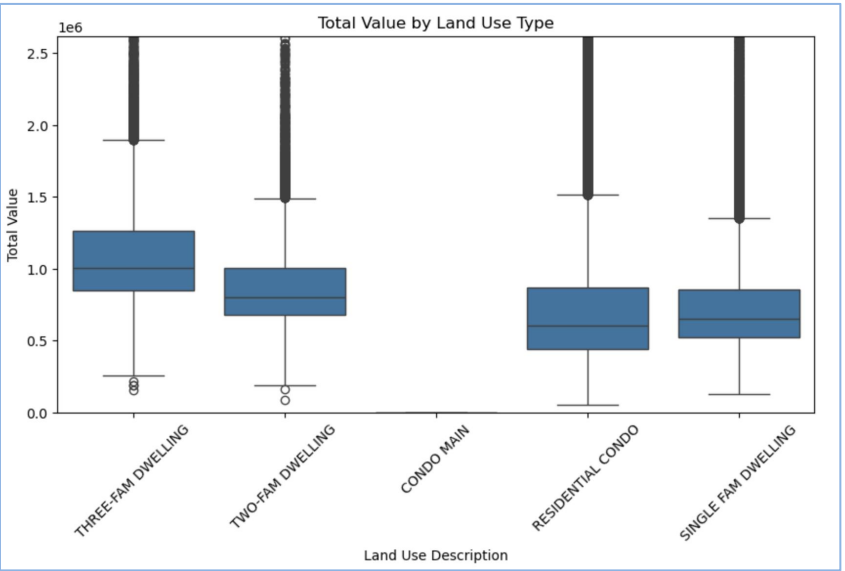
**Property Value by Land Use & ZIP Code:**

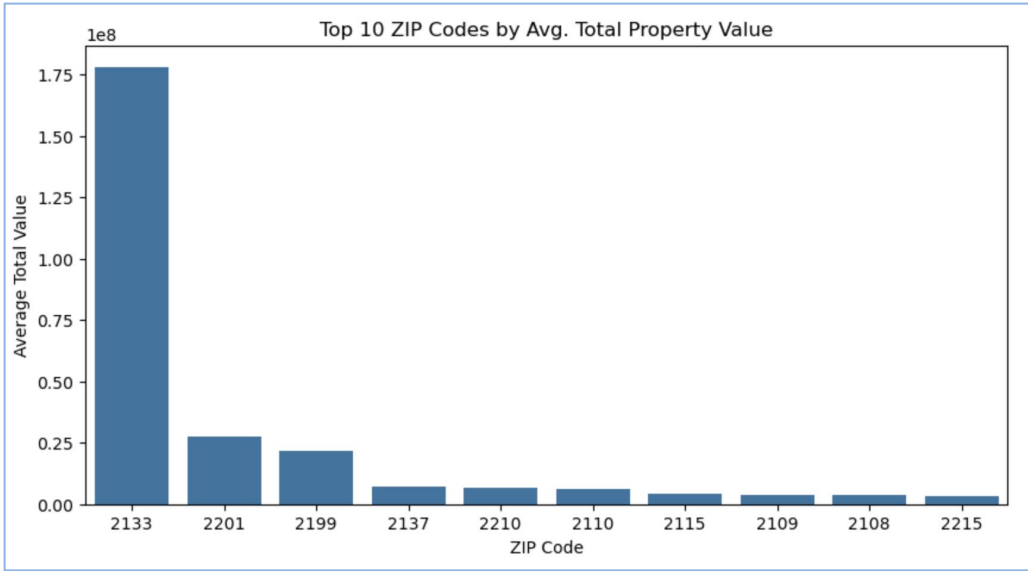
Boxplot – Total Value by Land Use:

- Multi-family homes (2- & 3-family) have higher median values than single-family or condos
- High number of outliers reflects variation in property size and condition
- CONDO MAIN category was removed due to flat/invalid values

Bar Chart – Top 10 ZIP Codes by Avg. Value:

- ZIP code 02133 stands out with the highest average value
- Indicates geographic disparities in Boston's property valuation



# EDA – Kitchen Style & Feature Correlation

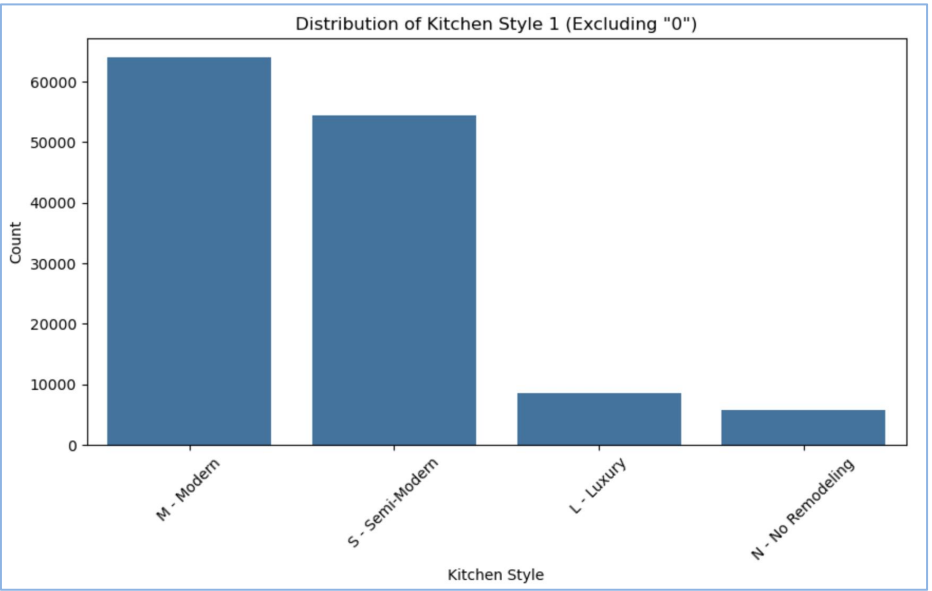
**Remodeling Trends & Feature Correlation:**

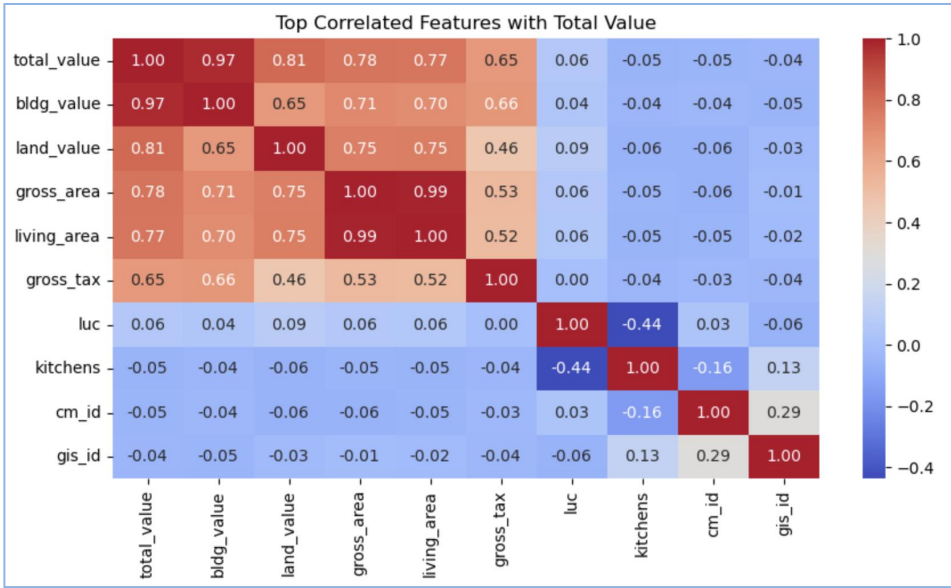
Kitchen Style Distribution:

- Majority of homes feature modern or semi-modern kitchens
- Entries labeled "0" were removed to eliminate invalid data
- Supports the idea that remodeling influences property value

Correlation Heatmap:

- Strong positive correlations with total_value:
  - bldg_value, land_value, gross_area, living_area (r > 0.75)
- Weak or negative correlation from features like:
  - Number of kitchens, internal wall type, ZIP in some cases

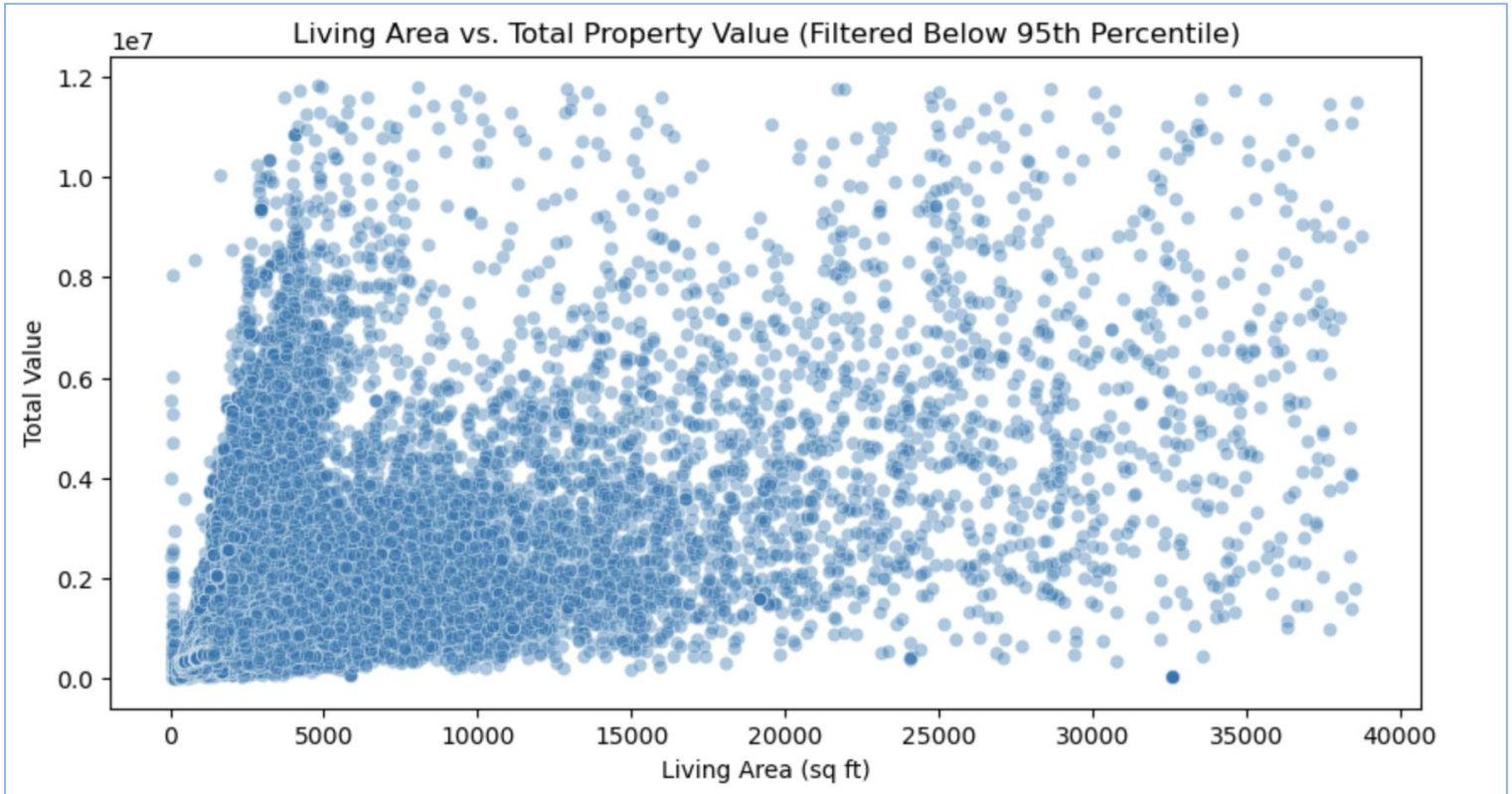


# EDA – Living Area vs. Property Value

**Living Area vs. Property Value:**

- Scatter plot (filtered below 95th percentile) shows a positive, non-linear relationship
- Larger homes generally have higher assessed values, but the gain tapers off
- Pattern suggests diminishing returns with increasing square footage
- Reinforces the need for non-linear models like Random Forest or XGBoost



# Modeling Overview

## Predictive Modeling Approach:

We tested three supervised regression models to predict total_value:

1. Linear Regression – simple, interpretable baseline
2. Random Forest Regressor – robust ensemble model for non-linear patterns
3. XGBoost Regressor – advanced boosting model tested on top numeric features

## All models were evaluated using:

- $R^2$ Score
- RMSE (Root Mean Squared Error)
- MAE (Mean Absolute Error)

## Note:

Each model used an 80/20 train-test split, and preprocessing included encoding and scaling as appropriate.

# Model 1 – Linear Regression

## Rationale:

Linear Regression is used as a baseline model. It assumes a linear relationship between predictors and the target variable, providing interpretable coefficients.

## Performance Results:

- $R^2$ Score: 0.9947
- RMSE: $370,163.90
- MAE: $11,049.16

## Interpretation:

The model achieved exceptionally high accuracy, suggesting a strong linear relationship between features like gross area and number of rooms with property value. However, the model may overfit, especially at the extremes of the value distribution. Residual analysis would be useful for further validation.

# Model 2 – Random Forest Regressor

## Rationale:

Random Forest is an ensemble method that reduces overfitting and improves generalization by combining predictions from multiple decision trees.

## Performance Results:

- $R^2$ Score: 0.9522
- RMSE: $1,108,968.61
- MAE: $64,340.47

## Interpretation:

The model provided strong predictive performance with reduced assumptions compared to linear regression. However, it struggled with outliers and extreme values, as shown by a higher RMSE. Still, it offers robustness against multicollinearity and is useful for understanding non-linear relationships..

# Model 3: XGBoost Regressor

## Rationale:

XGBoost is a gradient boosting technique optimized for speed and performance. It handles complex non-linear patterns and is robust to overfitting with proper tuning.

## Performance Results:

- $R^2$ Score: 0.2531
- RMSE: $4,385,129.73
- MAE: $199,211.74

## Interpretation:

Despite XGBoost's reputation for high performance, this model underperformed in our case. Possible reasons include:

- High variance in numeric features
- Lack of optimal hyperparameter tuning
- Complexity exceeding the benefit for this dataset

This result highlights the importance of model selection and feature preparation in real-world scenarios.

# Model Comparison Overview

| Model | $R^2$ Score | RMSE | MAE |
|---|---|---|---|
| Linear Regression | 0.9947 | ~$370K | ~$11K |
| Random Forest | 0.9522 | ~$1.1M | ~$64K |
| XGBoost | 0.2531 | ~$4.4M | ~$199K |

## Key Takeaways:

- **Linear Regression** performed best with extremely high $R^2$ and lowest error, likely due to the linear nature of key variables.
- **Random Forest** provided good performance but showed signs of overfitting in complex cases.
- **XGBoost** underperformed, suggesting a mismatch between model complexity and dataset characteristics.

## Final Recommendation:

**Linear Regression** is most suitable for this dataset, offering both simplicity and strong predictive power.

# Conclusion & Key Insights

The Boston FY2024 property dataset showed that building area, land value, and property age are strong predictors of total assessed value.

- Linear Regression outperformed ensemble models in both accuracy and interpretability.
- Exploratory analysis confirmed expected patterns:
  - Larger living area → higher value
  - Remodels (especially kitchens) associated with increased value
  - Geographic disparities observed across ZIP codes
- Machine learning can enhance property assessments but must be aligned with feature behavior and data quality.

# Recommendations & Future Work

**Recommendations:**

• Use linear regression as a reliable method for city-level valuation tools
• Prioritize data quality: cleaning and standardization significantly impact model success
• Consider remodel indicators (e.g., kitchens, bathrooms) in appraisal evaluations

**Future Work:**

• Apply log transformation to reduce skewness in total_value
• Explore time-series trends if historical assessments are available
• Conduct residual analysis and apply cross-validation to enhance model robustness
• Expand to include neighborhood-level variables like crime rates, school ratings, and transit access

# Thank You!