Real Estate Market Analysis (2001-2022)

1. Introduction

1.1 Motivation

The real estate market has experienced significant fluctuations over the past two decades, shaped by economic cycles, policy interventions, and demographic shifts. Analyzing these trends provides valuable insights into factors influencing property sales, price movements, and investment opportunities. Understanding real estate trends can help potential investors, government agencies, and homeowners make informed decisions about buying, selling, and development.

This study aims to analyze historical real estate sales data from 2001 to 2022, focusing on key economic events such as the housing boom, the 2008 financial crisis, and the subsequent recovery. The goal is to identify price trends, determine factors affecting sales, and predict future market movements using data-driven methodologies.

1.2 Problem Statement

The real estate market is influenced by multiple factors, including economic indicators, interest rates, policy regulations, and market demand. Predicting property prices accurately remains a challenge due to the interplay of these variables. This study seeks to address the following questions:

- How have real estate prices evolved over different economic cycles?
- What are the key drivers affecting real estate sales and pricing?
- How do property types of influence market trends and pricing?
- Can machine learning models effectively predict real estate prices based on historical data?

By answering these questions, this research provides insights into market stability and investment risks, assisting stakeholders in making well-informed decisions.

1.3 Objectives

- *Analyze historical real estate trends:* Identify how the market evolved through different economic phases (boom, crisis, and recovery).
- *Determine the impact of economic indicators:* Assess how factors such as interest rates, tax incentives, and policy changes influence market trends.
- *Examine property type variations:* Understand which types of properties were most sold and how their prices fluctuated.
- *Develop predictive models:* Use machine learning techniques to predict real estate prices and assess model accuracy.

2. Methodology

2.1 Data Collection & Description

The dataset comprises 1,097,629 real estate transactions recorded between 2001 and 2022, sourced from official housing agencies, government economic reports (e.g., Data.gov), and public repositories (USA Real Estate Trends). It includes the following variables:

Dataset Overview

- Size: 1,097,629 entries with 14 columns.
- Geographical Scope: Primarily covers towns in Connecticut (e.g., Ansonia, Avon) with geospatial coordinates (Location field).
- Time Range: Transactions span 2001–2022, as indicated by List Year and Date Recorded.

Key Variables

Property Metadata

- Serial Number: Unique transaction identifier.
- Address, Town: Location details.
- Property Type: Residential, Commercial, etc.
- Residential Type: Single Family, Two Family, etc.

Financial Metrics

- Assessed Value: Government-assessed property value.
- Sale Amount: Actual transaction price.
- Sales Ratio: Ratio of Sale Amount to Assessed Value.

Temporal & Categorical Data

- Date Recorded: Transaction date.
- Non Use Code: Classification for non-primary residences (e.g., rental, vacant).
- Geospatial & Remarks
- Location: Latitude/longitude coordinates.
- Assessor Remarks, OPM Remarks: Supplementary notes (high missingness).

Data Completeness

- High Missingness: OPM Remarks (1.2% filled), Assessor Remarks (15.6% filled).
- Partial Coverage: Non Use Code (28.5% filled) and Residential Type (63.7% filled).

2.2 Data Cleaning & Preprocessing

To ensure the reliability of the data, several preprocessing steps were performed:

- *Handling Missing Values:* Columns with excessive missing data (e.g., Assessor Remarks, OPM Remarks) were dropped. Missing categorical values were filled with the most frequent category, while missing numerical values were imputed using median values.
- *Removing Duplicates:* Identified and removed duplicate records based on property address, sale amount, and recorded date to maintain data integrity.
- *Handling Outliers:* Identified extreme values in sale amount and assessed value using statistical methods such as interquartile range (IQR) and removed erroneous records.

2.3 Analysis Techniques

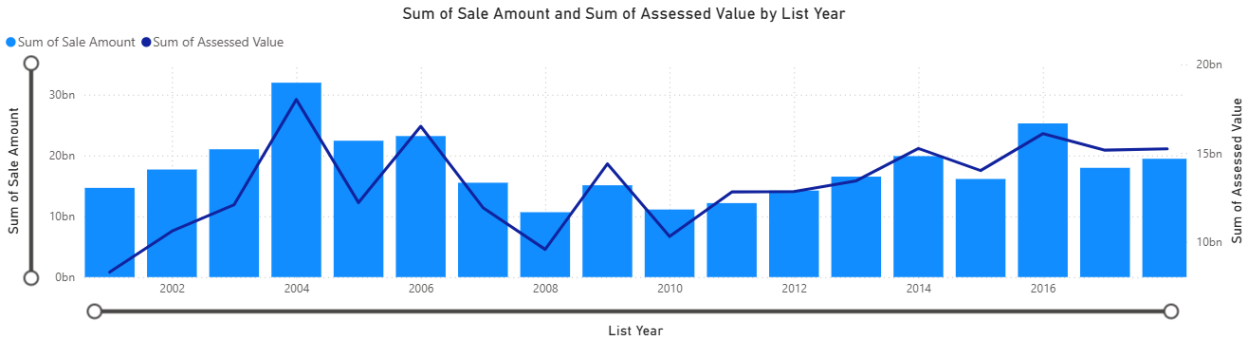
- *Exploratory Data Analysis (EDA):* Visualized trends using histograms, line charts, and heatmaps to identify patterns.
- *Correlation Analysis:* Examined relationships between sale prices, assessed values, and economic indicators using correlation matrices.

- *Machine Learning Models:*
    - *Linear Regression:* Analyzed fundamental relationships between price and features.
    - *Random Forest Regression:* Improved predictive accuracy by handling non-linearity and feature interactions.
    - *Feature Importance Analysis:* Identified key variables affecting real estate prices.

3. Results and Analysis

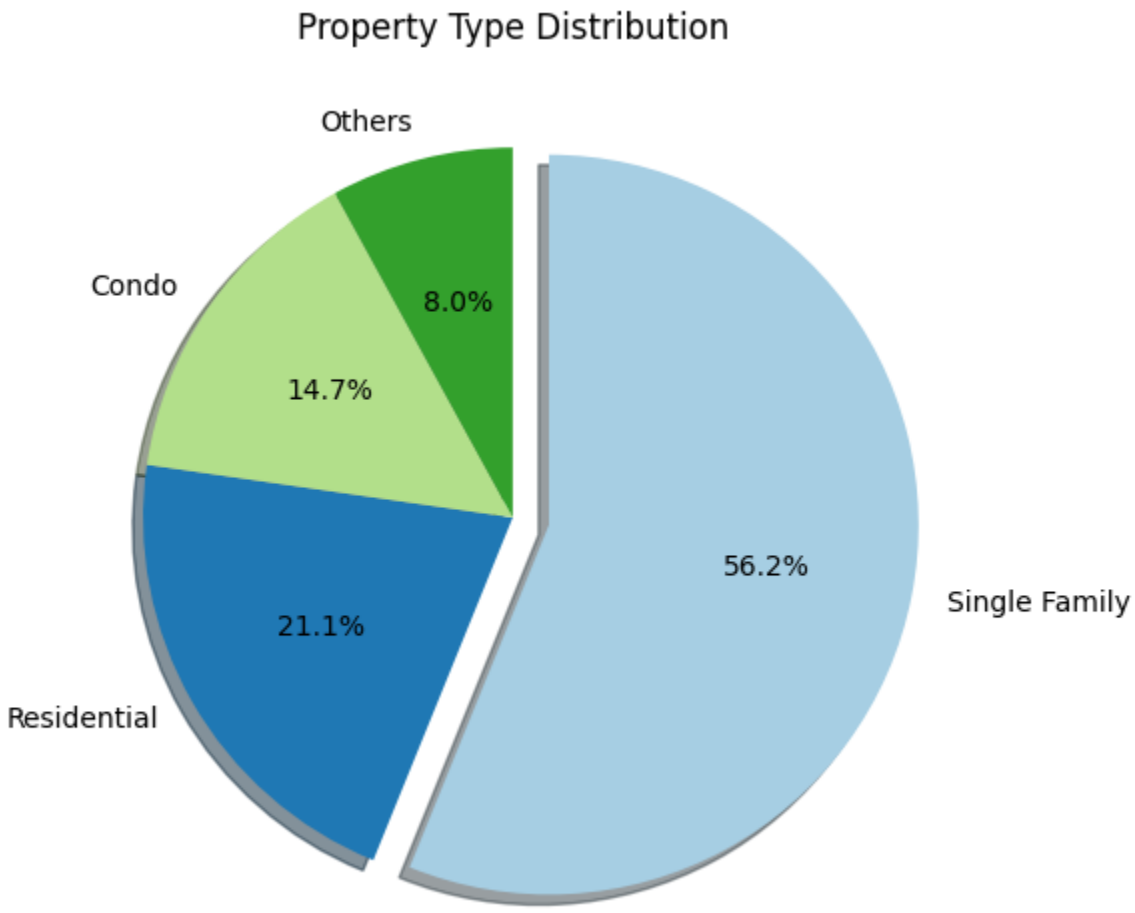
3.1 Real Estate Market Trends

1. *Boom Period (2001-2007):*
    a. Rapid increase in housing demand driven by easy credit access and low interest rates.
    b. Property prices peaked between 2004-2005 before showing early signs of instability.
    c. Increased investment in high-growth areas led to speculative buying.
2. *Crash & Financial Crisis (2008-2010):*
    a. The housing bubble burst, leading to a steep decline in property values.
    b. Many homeowners faced foreclosure due to mortgage defaults.
    c. Significant drop in real estate transactions due to tighter lending restrictions.
3. *Recovery Phase (2010-2022):*
    a. Gradual stabilization with government stimulus packages and low-interest policies.
    b. Increased sales volume observed from 2014-2017, with new investment trends in suburban areas.
    c. Post-pandemic shifts in housing demand due to remote work and urban-to-suburban migration.



3.2 Property Types and Sales Distribution

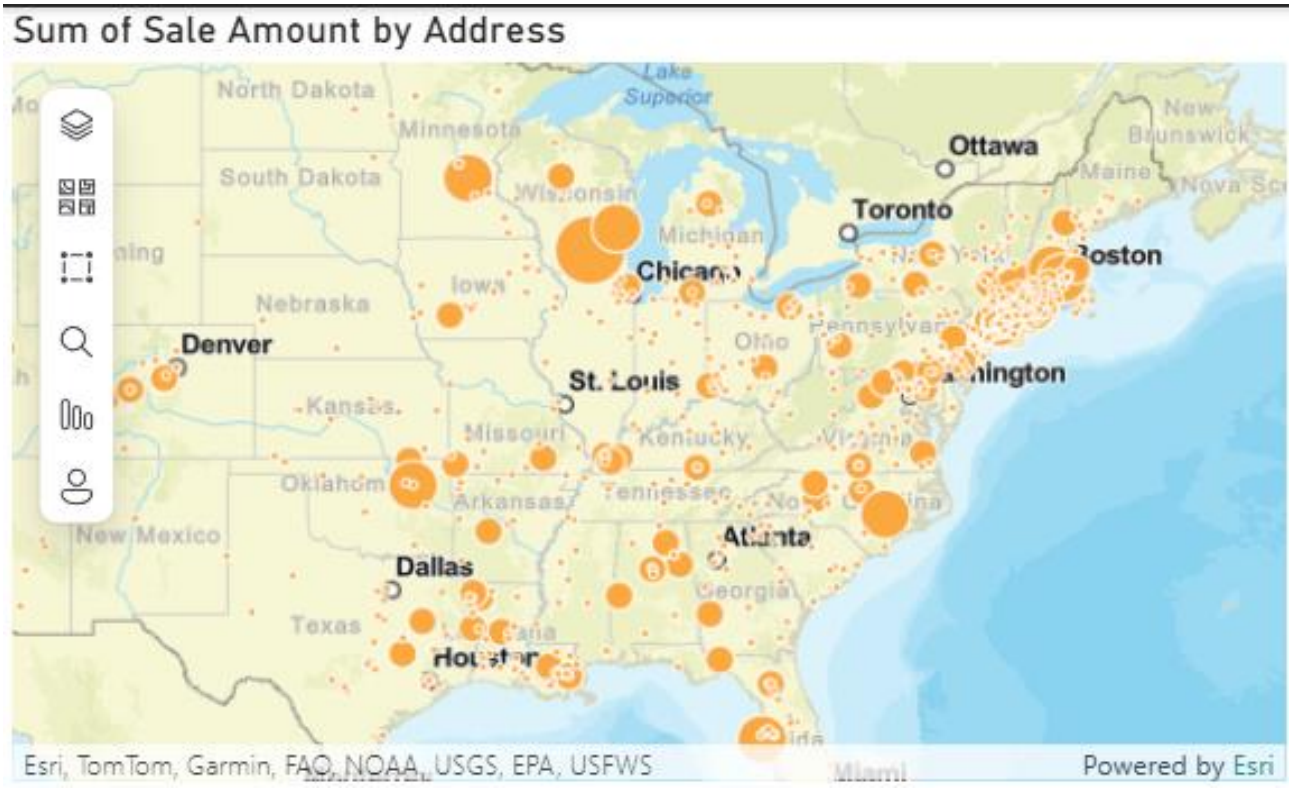
- *Single-Family Homes (56.2%)*: Dominated the market due to high demand for personal residences.
- *Condos (14.7%)*: Gained popularity in urban centers and among first-time homebuyers.
- *Commercial & Industrial Properties (8.0%)*: Sales varied with economic growth cycles.
- *Other Property Types (21.1%)*: Included mixed-use developments, investment properties, and rental units.



3.3 Impact of Economic and Policy Factors

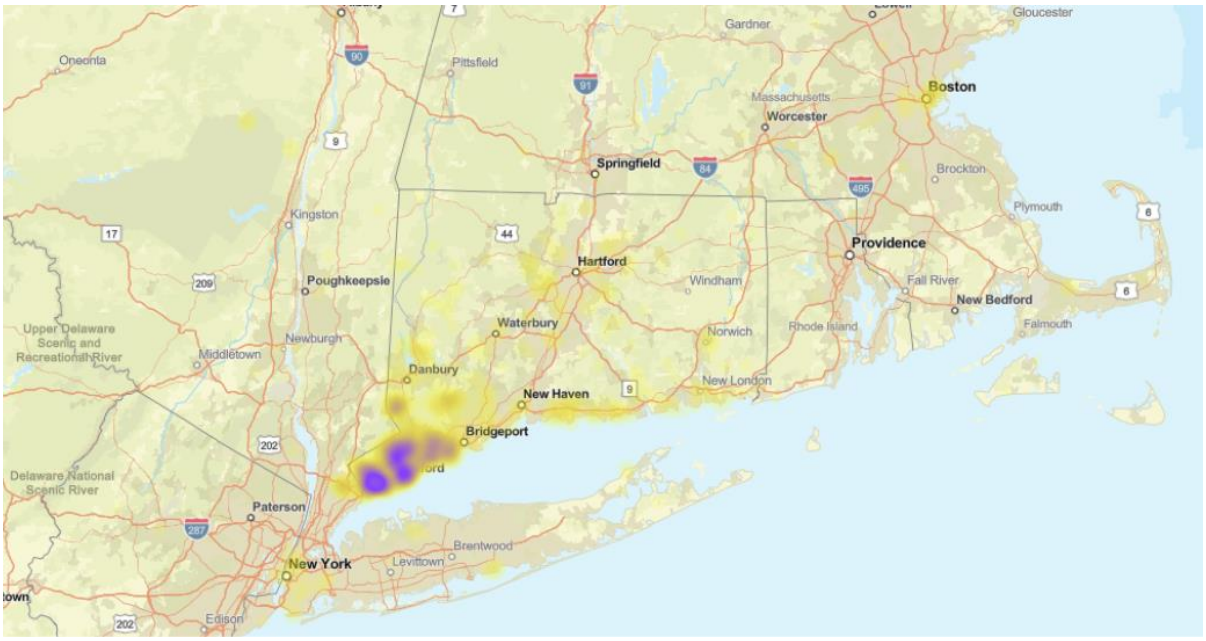
- *Interest Rates & Mortgage Policies:* Higher rates correlated with reduced housing affordability.

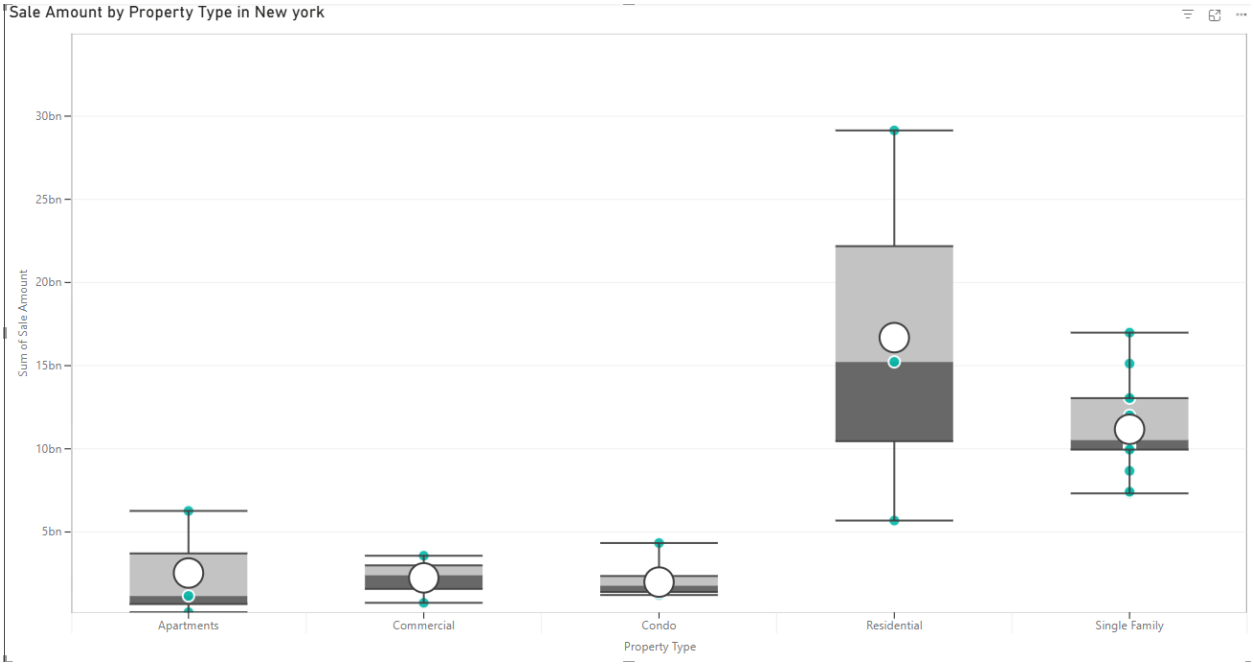
- *Tax Incentives & Government Regulations:* Stimulated real estate investments during recovery periods.
- *Demographic Shifts:* Increased millennial homeownership and migration trends influenced demand.



3.4 Regional Variations

- *New York City & Surrounding Areas:* Highest real estate transactions recorded.
- *Boston & Connecticut Corridor:* Moderate activity with steady demand for luxury and commercial properties.





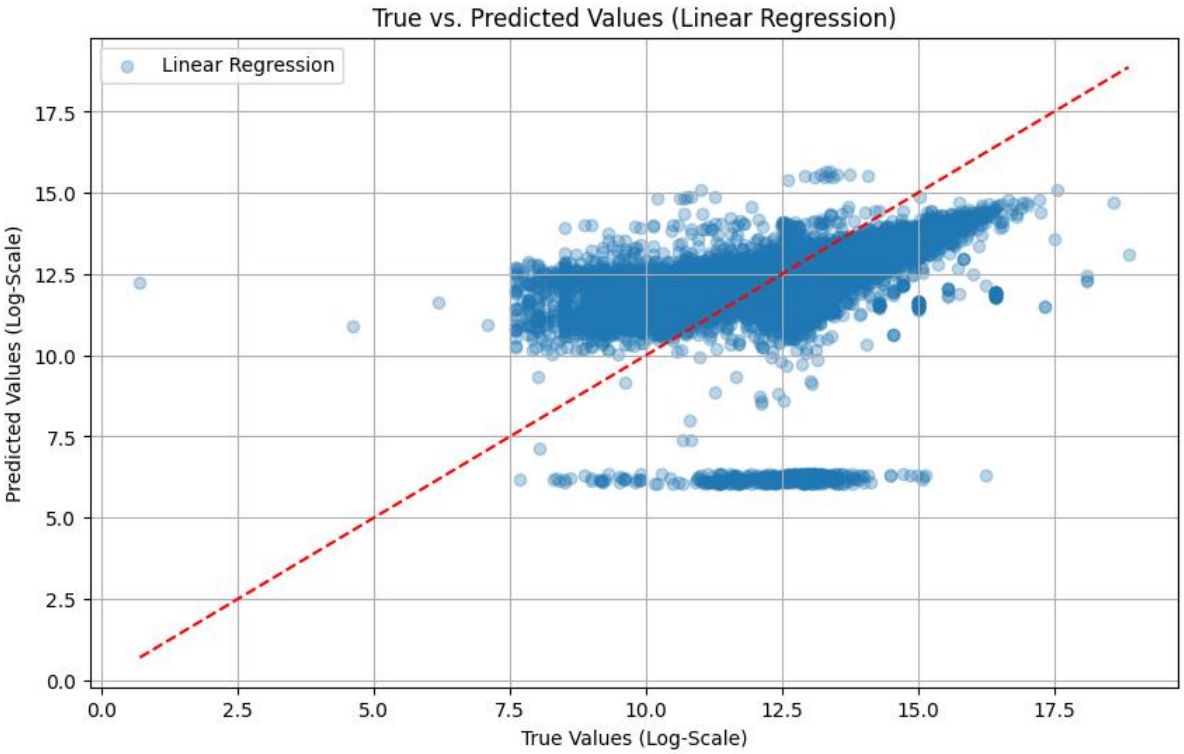
4. Predictive Modeling & Insights

4.1 Model Performance

- Accuracy (R^2): 36.17%
- Mean Squared Error (MSE): 0.509

Interpretation:

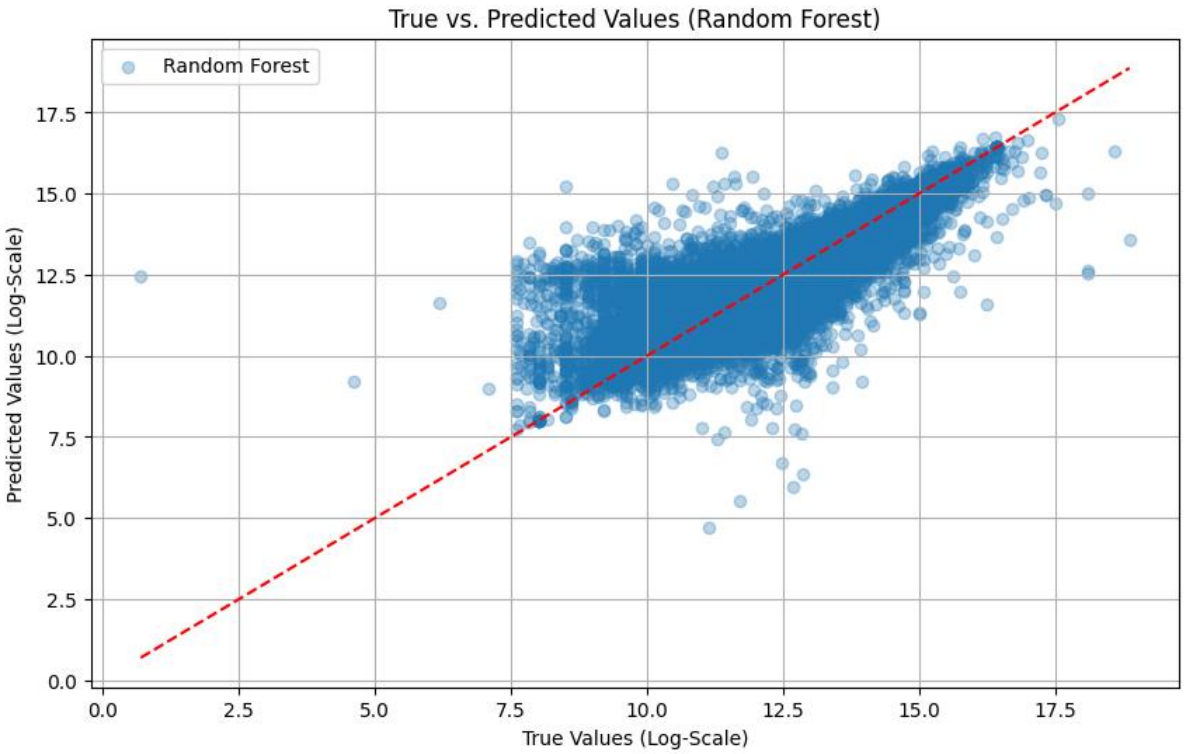
- The model explains 36% of the variance in sale prices, indicating limited predictive power.
- High MSE suggests significant errors, likely due to non-linear trends (e.g., extreme price fluctuations during the 2008 crisis) and outliers.
- Struggles with capturing complex interactions (e.g., interest rates × property type).



- Accuracy (R^2): 72.28%
- Mean Squared Error (MSE): 0.221

Interpretation:

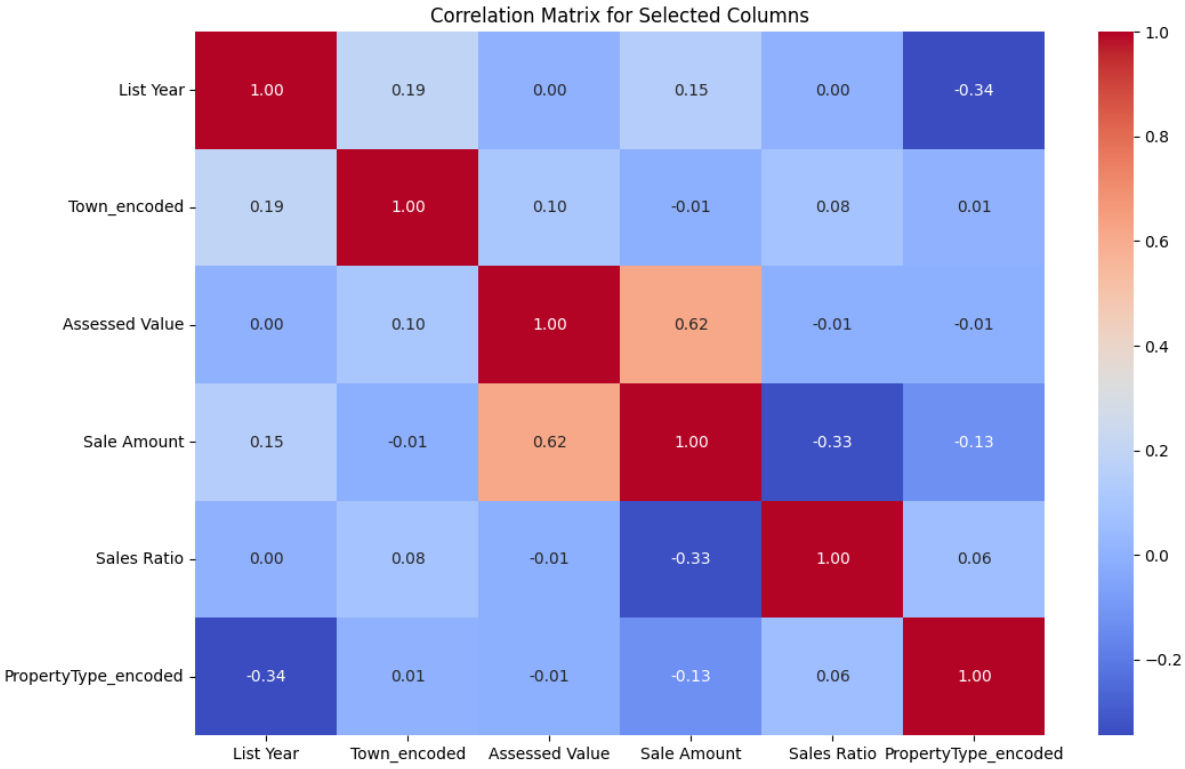
- Explains 72% of the variance, a substantial improvement over linear regression.
- Lower MSE indicates better handling of non-linear relationships and outliers.
- Effectiveness stems from capturing feature interactions (e.g., Assessed Value × Property Type) and market seasonality.



4.2 Key Predictors of Sale Prices

- *Assessed Value:* Strong correlation (*0.62*) with sale price.
- *Property Type:* Negative correlation (*-0.34*) with sale year, indicating shifting market preferences.

- *Sales Ratio:* Weak correlation with price fluctuations, suggesting influence from external factors.



5. Conclusion & Future Work

5.1 Summary of Findings

- Real estate prices and sales followed economic cycles, influenced by major financial events.
- Single-family homes remained the most sought-after property type.
- Government policies, interest rates, and demographics played a significant role in market trends.
- Random Forest Regression provided the most accurate predictions among the models tested.

5.2 Future Work

- *Feature Engineering Enhancements:* Include employment rates and inflation data.
- *Advanced Modeling:* Explore Gradient Boosting and Neural Networks for improved accuracy.

- *Policy Impact Analysis:* Investigate how zoning laws and tax incentives shape market dynamics.

References

https://catalog.data.gov/dataset/real-estate-sales-2001-2018
https://github.com/Raju-jrDev/USA-Real-Estate-Trends