Property Sale Amount Prediction in Connecticut

Group 5 :

Venkata Lakshmi Parimala Pasupuleti - vpasupu2@gmu.edu

Aravind Panchanathan - apanchan@gmu.edu

Sanjay Kumar Podishetty - spodishe@gmu.edu

Dinesh Ponnada - dponnada@gmu.edu

Pramath Rajprasad Rao - prajpras@gmu.edu

Lina Saade - lsaade@gmu.edu

Greeshma Priya Pendyala - gpendya@gmu.edu

Professor: Lam Phung

Course: AIT 582 - 003

Date: May 7, 2025

Introduction

1

Economic Impact

Real estate sales directly impact economy and individual wealth.

2

Prediction Benefits

Accurate prediction models improve investment, lending, and planning decisions.

3

Market Focus

Focus on Connecticut real estate market and education-related factors.

4

Technology Approach

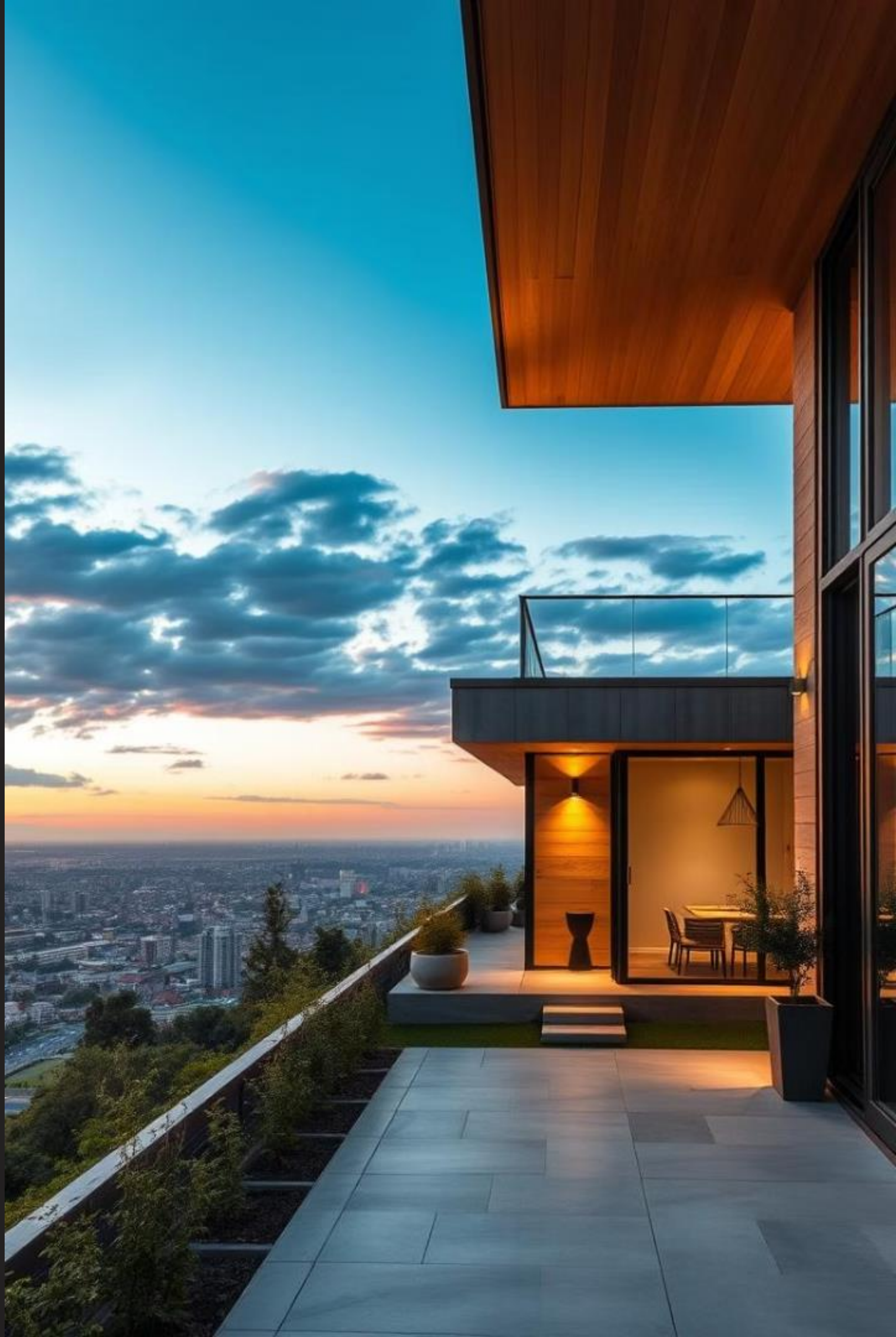
Machine learning used to predict property sale prices.

5

Data Integration

Integration of multiple datasets for better prediction quality.

Pyspark and AWS



Problem Statement

Hidden Patterns

Traditional valuation methods may not capture hidden patterns.

Price Dependencies

Sale prices depend on assessed value, local infrastructure, and socio-economic factors.

Model Requirements

Need for accurate, scalable, and interpretable models.

Data Challenges

Challenge: Handling large datasets with missing values and outliers.

Significance

Stakeholder Benefits

Predictive analytics aid real estate investors, banks, and governments.

Policy Optimization

Accurate forecasts help optimize investments and housing policies.

Value Identification

Early identification of undervalued or overvalued properties.

Infrastructure Insights

Insights into role of town infrastructure (schools) on property prices.

Data Overview

Main Property Dataset

Transactions, Assessments.

Merged on Town and Year fields.

Preprocessing included missing value treatment and standardization.

Housing Sales Data

Sale Amounts.

Schools Data

Number of Schools (PreK–12).

Crime Data

Crime incidents per 100,000 population.

List Year	Interval	The year the property was listed for sale.
Date Recorded	Interval	The date of the sale was recorded locally.
Town	Nominal	The name of the town where the property is located.
Address	Nominal	The physical address of the property.
Assessed Value	Ratio	The value of the property used for local tax assessment.
Sale Amount	Ratio	The amount the property was sold for.
Sales Ratio	Ratio	The ratio of the sale price to the assessed value.
Property Type	Nominal	Types of property include Residential, Commercial, Industrial, Apartments, Vacant, etc.
Residential Type	Nominal	Indicates whether the property is single or multifamily residential
Non-Use Code	Nominal	The sale price is not reliable for use in the determination of a property value
Assessor Remarks	Nominal	Additional remarks or notes from the property assessor.
OPM remarks	Nominal	Remarks from the Office of Policy and Management (OPM).
Location	Ordinal	Latitude and longitude coordinates of the property.

Literature Review

•**Impact of Crime Rates:**
Studies show higher crime rates negatively influence property values by reducing buyer demand and increasing risk perception. *(Gibbons, 2004; Ihlanfeldt & Mayock, 2010)*

•**School Quality and Property Prices:**
Access to high-quality schools is strongly associated with higher home values. Parents often pay a premium to live in better school districts. *(Black, 1999; Nguyen-Hoang & Yinger, 2011)*

•**Healthcare Access and Real Estate:**
Proximity to healthcare facilities improves neighborhood attractiveness and can increase property prices, especially for aging populations. *(Beard et al., 2009)*

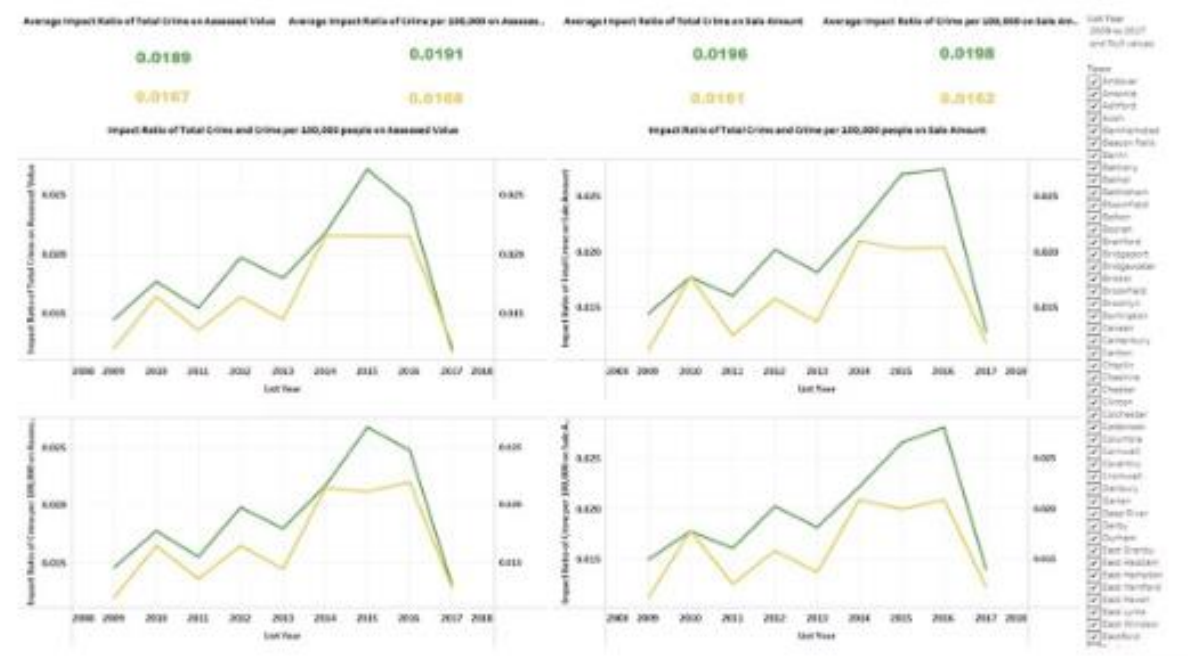
•**Socioeconomic Indicators:**
Higher employment rates, average wages, and business density are positively correlated with higher property sale amounts. *(Glaeser & Gyourko, 2008)*

Research Questions

- Which geographic features (e.g., assessed value and sale ratio) most accurately predict property sale prices in Connecticut?
- How does access to educational institutions (based on available grade levels) influence residential property values?
- What is the impact of assessed value and year of sale on predicting final sale prices of residential properties?

Visualizations

Impact of Crime Rates on Property Sale Amount

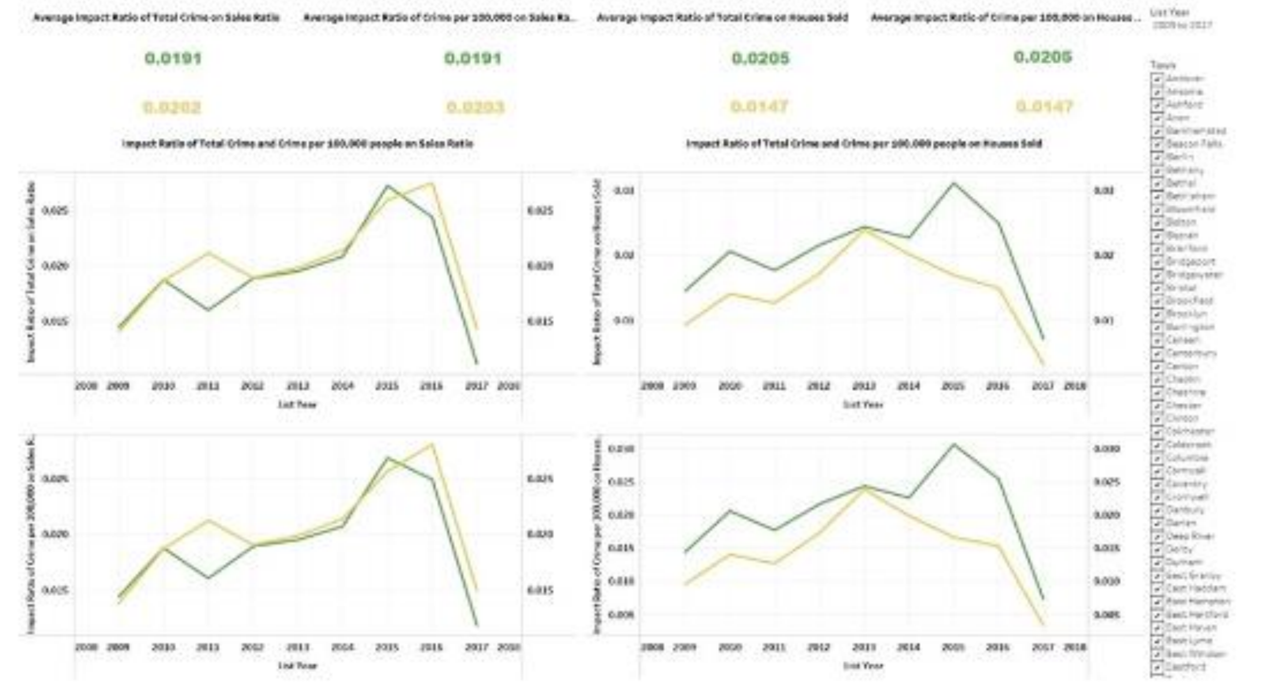


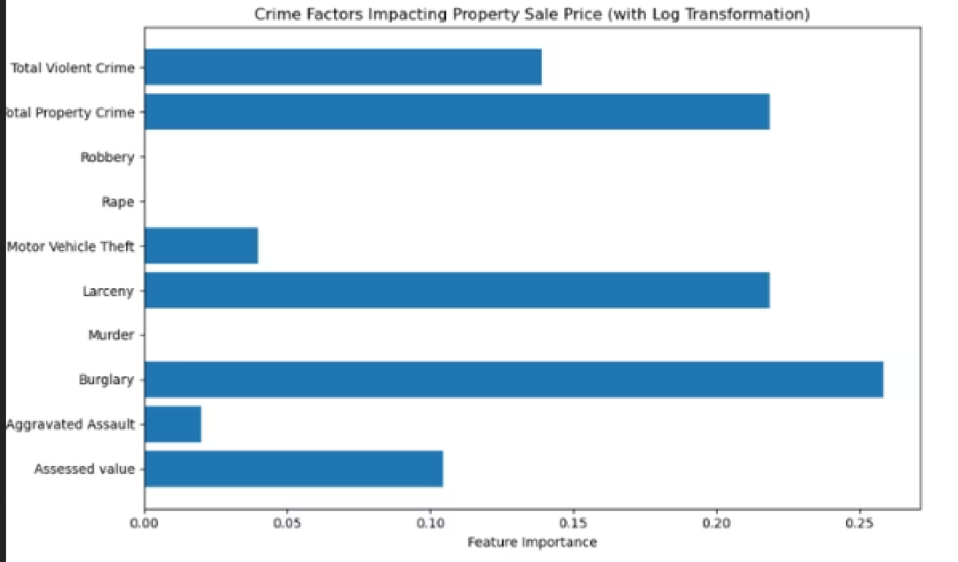
Crime Impact on Sales Activity

- Crime rates showed moderate influence on *Sales Ratio* and *Houses Sold* between 2009–2016.
- After 2016, crime impact declined, indicating increasing l
- Total crime counts had a slightly stronger effect than crir

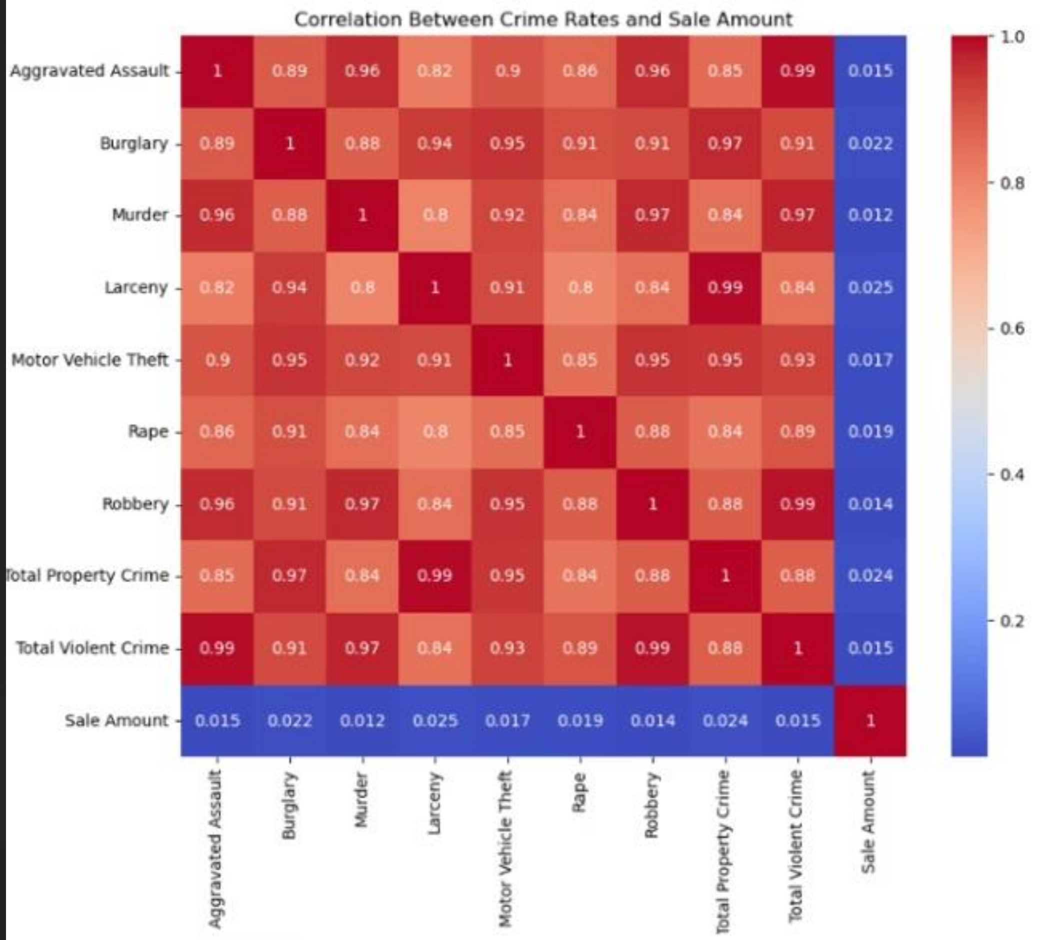
Crime Impact on Property Values

- Higher crime rates correlated with lower *Assessed Values* and *Sale Amounts* until 2015.
- A sharp drop in crime impact after 2016 suggests market stabilization.
- Total crimes influenced property values slightly more than normalized crime rates.





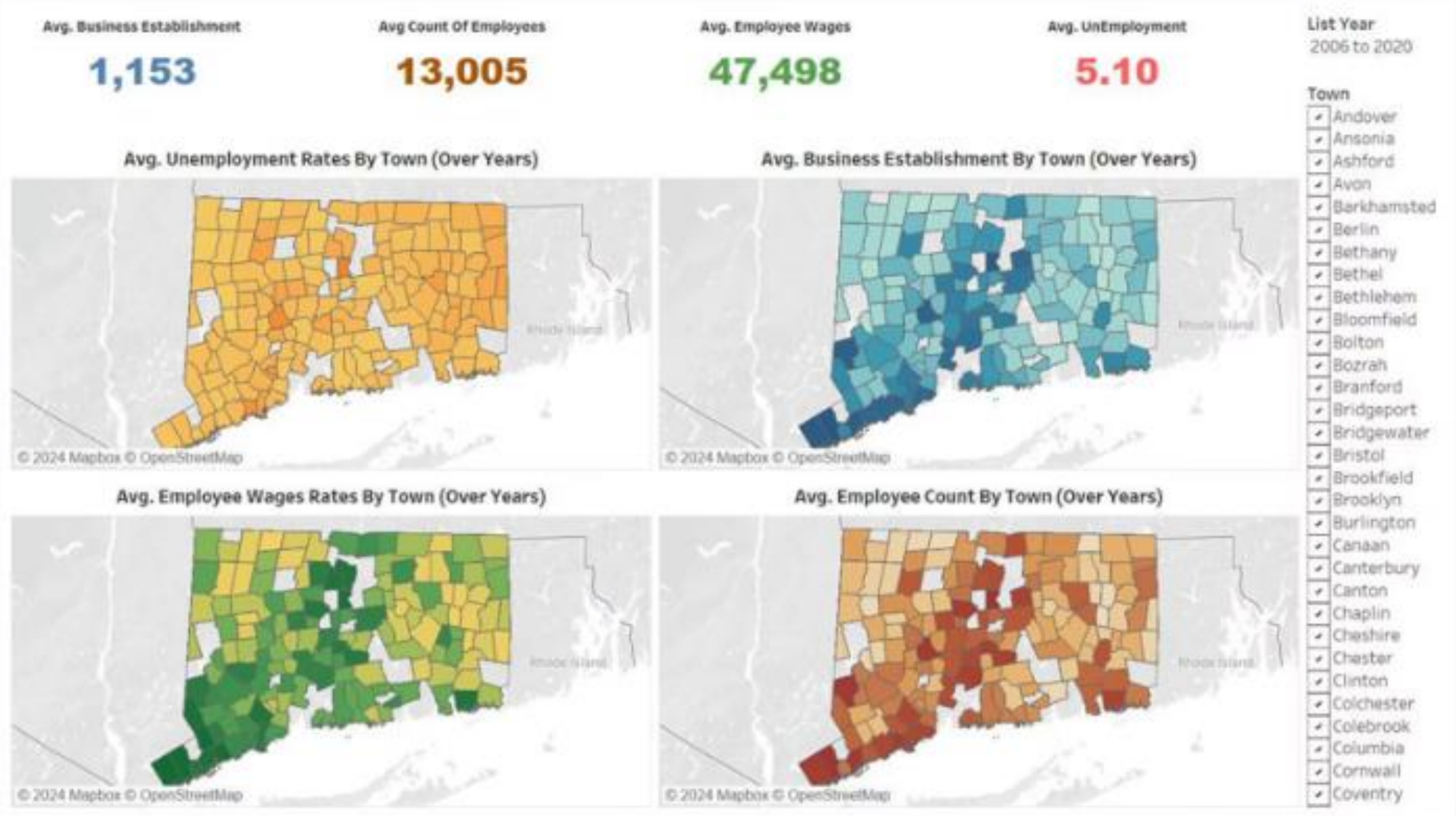
- Correlation analysis shows that crime rates have **very weak direct relationships** with property sale amounts (correlation values around 0.01–0.02).
- Feature importance analysis highlights that **Burglary**, **Total Property Crime**, and **Larceny** have a **greater influence** on property sale prices compared to other crimes.



- **Key Takeaway:** While crime rates individually show weak correlations, certain crimes like Burglary have a **more significant indirect impact** on housing prices when evaluated through machine learning models.

Socioeconomic Factors Across Connecticut Towns

- **Unemployment Rates**:
- **Business Establishments**:
- **Employee Wages**:
- **Employee Counts**:



Avg. Business Establishment	Avg Count Of Employees	Avg. Employee Wages	Avg. UnEmployment
1,153	13,005	47,498	5.10

Avg. Unemployment Rates By Town (Over Years)

Avg. Business Establishment By Town (Over Years)

Avg. Employee Wages Rates By Town (Over Years)

Avg. Employee Count By Town (Over Years)

List Year
2006 to 2020

Town
☑ Andover
☑ Ansonia
☑ Ashford
☑ Avon
☑ Barkhamsted
☑ Berlin
☑ Bethany
☑ Bethel
☑ Bethlehem
☑ Bloomfield
☑ Bolton
☑ Bozrah
☑ Branford
☑ Bridgeport
☑ Bridgewater
☑ Bristol
☑ Brookfield
☑ Brooklyn
☑ Burlington
☑ Canaan
☑ Canterbury
☑ Canton
☑ Chaplin
☑ Cheshire
☑ Chester
☑ Clinton
☑ Colchester
☑ Colebrook
☑ Columbia
☑ Cornwall
☑ Coventry

- Higher unemployment observed in northern and interior regions, possibly reducing property demand.
- Higher concentration of businesses near coastal areas and major cities, boosting local economies.
- Western and southern Connecticut show higher average wages, correlating with higher property values.
- Densely populated urban areas exhibit larger employee counts, supporting stronger housing markets.

Condo Sales Analysis Based on School Availability

Higher Assessed and Sale Values

Towns like Fairfield and Stamford report significantly higher condo assessed values and sale prices.

Condo Sale Ratio Trends

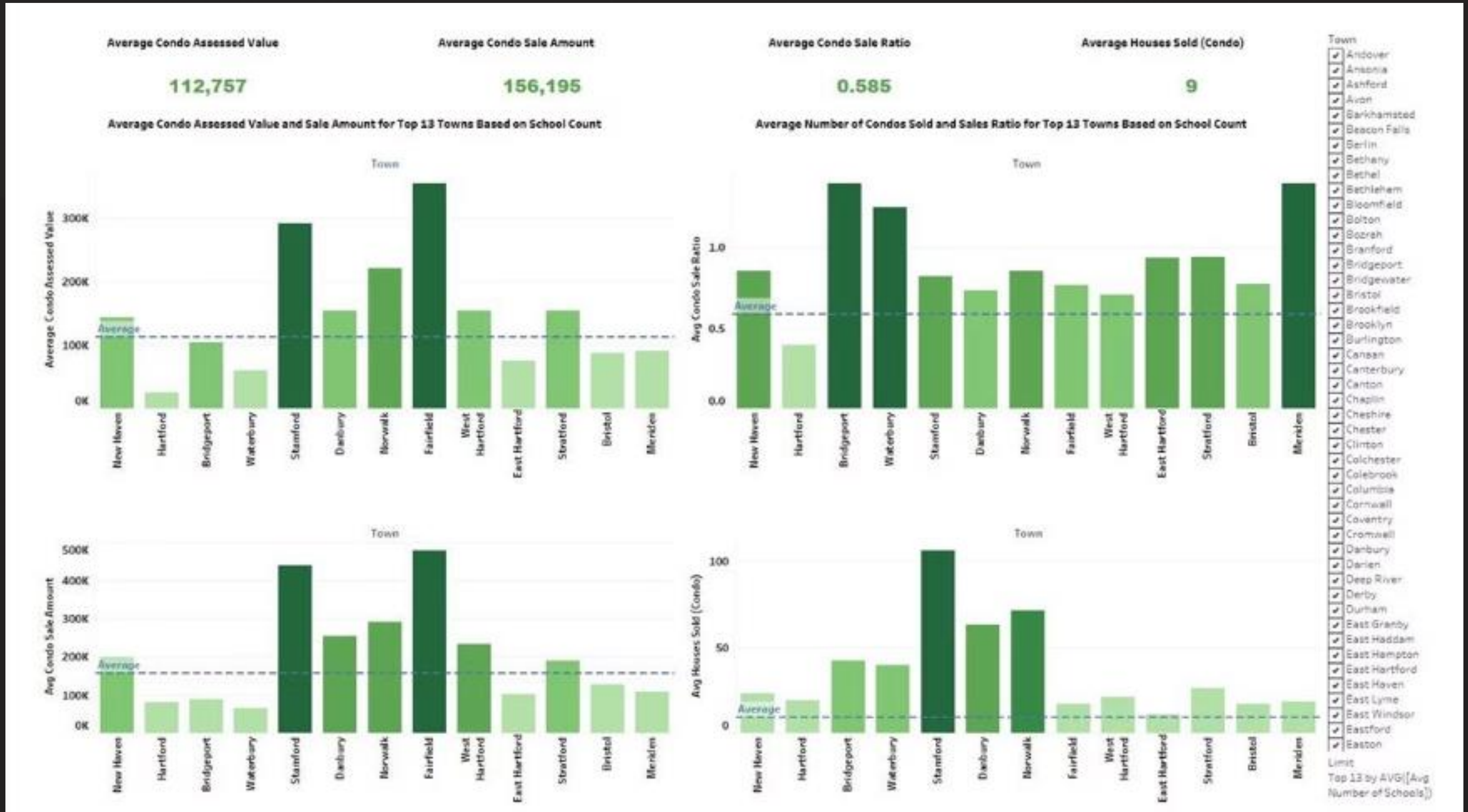
Bridgeport and Waterbury show the highest condo sale ratios, indicating strong buyer demand relative to assessed values.

Units Sold Distribution

Stamford leads in the number of condos sold, reflecting a highly active condo market.

School Count Correlation

Towns with a greater number of schools tend to have more vibrant and higher-value condo markets.



Average Condo Assessed Value: 112,757

Average Condo Sale Amount: 156,195

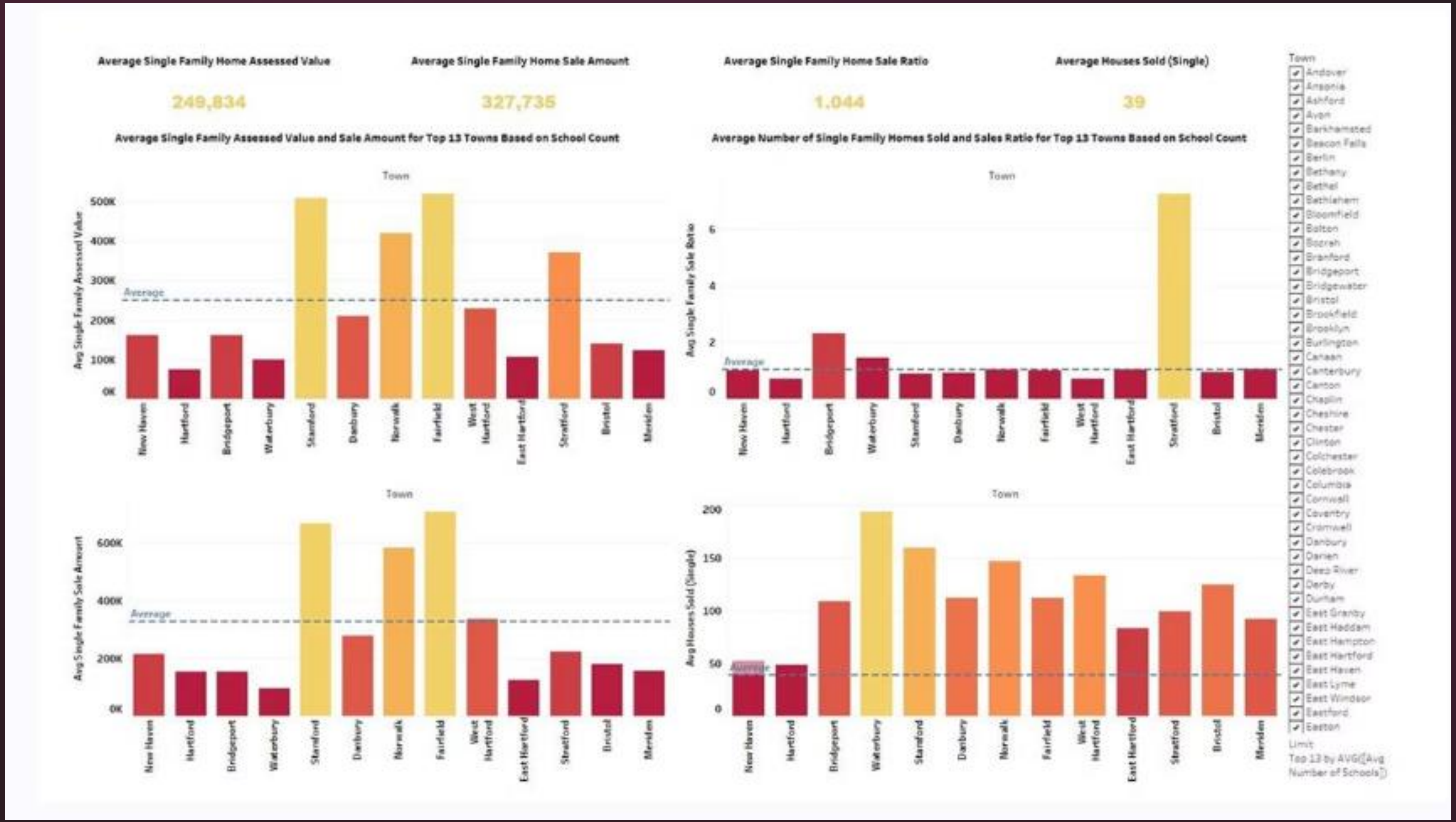
Average Condo Sale Ratio: 0.585

Average Houses Sold (Condo): 9

Average Condo Assessed Value and Sale Amount for Top 13 Towns Based on School Count

Average Number of Condos Sold and Sales Ratio for Top 13 Towns Based on School Count

Single-Family Home Sale Analysis Based on School Availability



Average Single Family Home Assessed Value: 249,834

Average Single Family Home Sale Amount: 327,735

Average Single Family Home Sale Ratio: 1.044

Average Houses Sold (Single): 39

Average Single Family Assessed Value and Sale Amount for Top 13 Towns Based on School Count

Average Number of Single Family Homes Sold and Sales Ratio for Top 13 Towns Based on School Count

Town
☑ Andover
☑ Ansonia
☑ Ashford
☑ Avon
☑ Barkhamsted
☑ Beacon Falls
☑ Berlin
☑ Bethany
☑ Bethel
☑ Bethlehem
☑ Bloomfield
☑ Bolton
☑ Bozrah
☑ Branford
☑ Bridgeport
☑ Bridgewater
☑ Bristol
☑ Brookfield
☑ Brooklyn
☑ Burlington
☑ Canaan
☑ Canterbury
☑ Canton
☑ Chaplin
☑ Cheshire
☑ Chester
☑ Clinton
☑ Colchester
☑ Colebrook
☑ Columbia
☑ Cornwall
☑ Coventry
☑ Cromwell
☑ Danbury
☑ Darien
☑ Deep River
☑ Derby
☑ Durham
☑ East Granby
☑ East Haddam
☑ East Hampton
☑ East Hartford
☑ East Haven
☑ East Lyme
☑ East Windsor
☑ Eastford
☑ Easton

Limit:
Top 13 by AVG([Avg Number of Schools])

Higher Valuations:
Stamford and Fairfield show the highest assessed and sale values for single-family homes among all towns.

Sale Ratios Vary:
Towns like Stratford and Waterbury show higher sale ratios, indicating homes are selling above or close to assessed value.

Units Sold Distribution:
Waterbury and Stamford recorded the highest number of single-family home sales.

School Presence Influence:
Towns with more schools generally have better sales performance and higher average prices, supporting the role of educational infrastructure in boosting real estate markets.

Impact of Healthcare Facilities on Property Sales

Healthcare Facilities Impact

Towns with more healthcare facilities generally had higher assessed values and sale amounts for single-family homes.

Strong Values in Stamford and Bloomfield

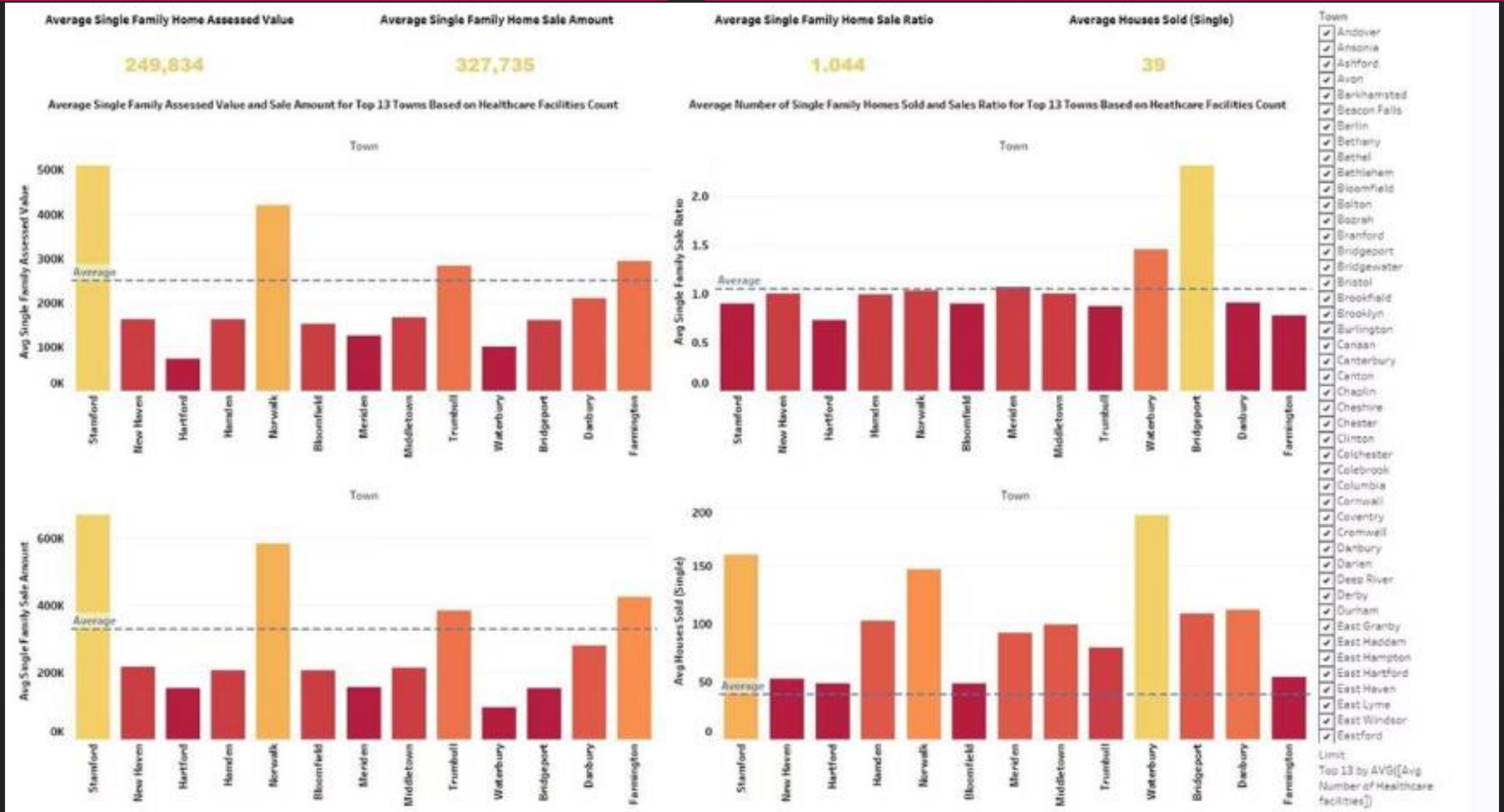
Stamford and Bloomfield showed particularly strong property values linked with better healthcare access.

Higher Sale Ratios

Towns with higher healthcare facility counts also recorded higher houses sold and sale ratios compared to others.

Positive Influence of Healthcare

Suggests that proximity to healthcare positively influences real estate demand and property prices.



Average Single Family Home Assessed Value	Average Single Family Home Sale Amount	Average Single Family Home Sale Ratio	Average Houses Sold (Single)
249,834	327,735	1.044	39

Average Single Family Assessed Value and Sale Amount for Top 13 Towns Based on Healthcare Facilities Count

Average Number of Single Family Homes Sold and Sales Ratio for Top 13 Towns Based on Heathcare Facilities Count

Town
☑ Andover
☑ Ansonia
☑ Ashford
☑ Avon
☑ Barkhamsted
☑ Beacon Falls
☑ Berlin
☑ Bethany
☑ Bethel
☑ Bethlehem
☑ Bloomfield
☑ Bolton
☑ Bozrah
☑ Branford
☑ Bridgeport
☑ Bridgewater
☑ Bristol
☑ Brookfield
☑ Brooklyn
☑ Burlington
☑ Canaan
☑ Canterbury
☑ Canton
☑ Chaplin
☑ Cheshire
☑ Chester
☑ Clinton
☑ Colchester
☑ Colebrook
☑ Columbia
☑ Cornwall
☑ Coventry
☑ Cromwell
☑ Danbury
☑ Darien
☑ Deep River
☑ Derby
☑ Durham
☑ East Granby
☑ East Haddam
☑ East Hampton
☑ East Hartford
☑ East Haven
☑ East Lyme
☑ East Windsor
☑ Eastford

Limit
Top 13 by AVG([Avg Number of Healthcare facilities])

Impact of Healthcare Facilities on Condo Sales

1 Condo Values Soar in Healthcare Hubs

Towns like Stamford and Bloomfield, with abundant healthcare facilities, see sky-high condo assessed values and sale prices.

2 Brisk Condo Sales in Healthy Towns

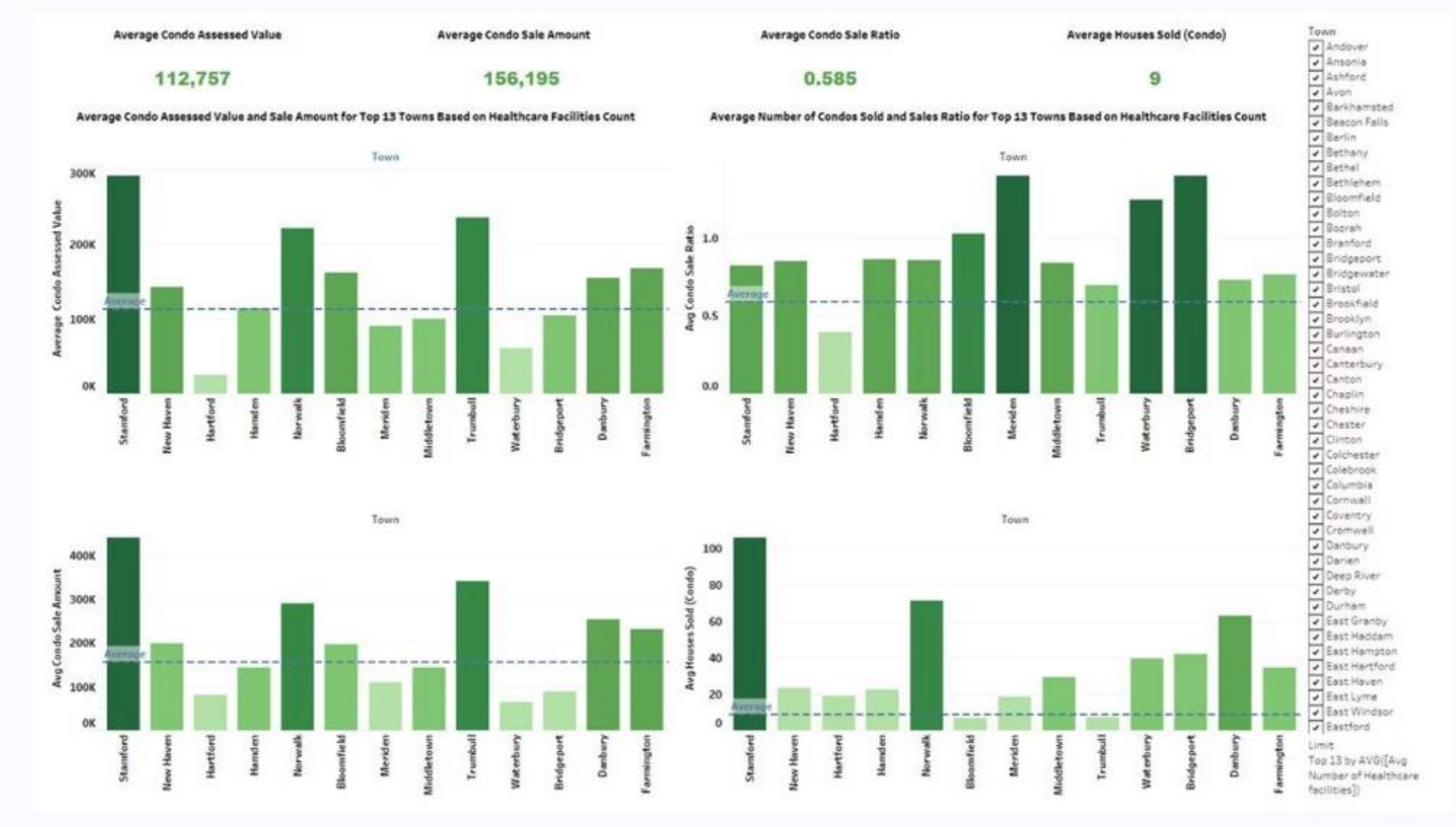
Condo sales ratios and the number of units sold are significantly higher in areas with greater healthcare access.

3 Healthcare Fuels Condo Demand

The presence of robust healthcare infrastructure appears to drive strong real estate activity and property values for condos.

4 Healthcare Matters for Condo Buyers

Convenient access to healthcare is an important socioeconomic factor that shapes condo market trend



Average Condo Assessed Value	Average Condo Sale Amount	Average Condo Sale Ratio	Average Houses Sold (Condo)
112,757	156,195	0.585	9

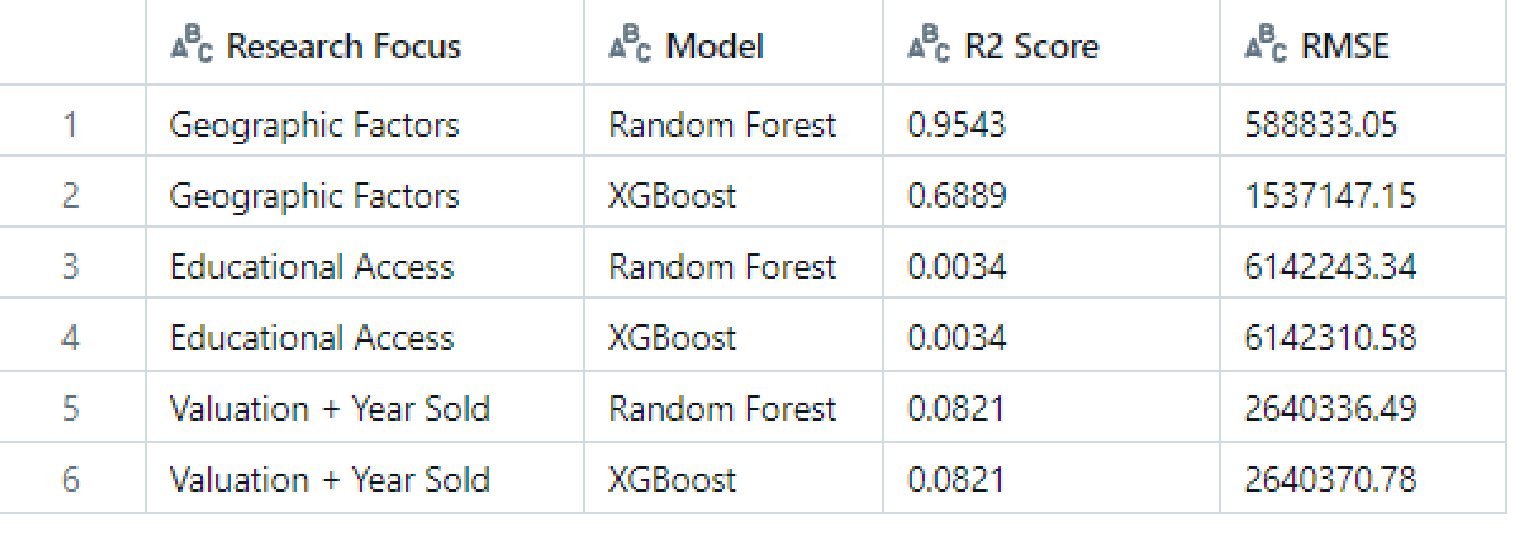
Average Condo Assessed Value and Sale Amount for Top 13 Towns Based on Healthcare Facilities Count

Average Number of Condos Sold and Sales Ratio for Top 13 Towns Based on Healthcare Facilities Count

Town
☑ Andover
☑ Ansonia
☑ Ashford
☑ Avon
☑ Barkhamsted
☑ Beacon Falls
☑ Berlin
☑ Bethany
☑ Bethel
☑ Bethlehem
☑ Bloomfield
☑ Bolton
☑ Bozrah
☑ Branford
☑ Bridgeport
☑ Bridgewater
☑ Bristol
☑ Brookfield
☑ Brooklyn
☑ Burlington
☑ Canaan
☑ Canterbury
☑ Canton
☑ Chaplin
☑ Cheshire
☑ Chester
☑ Clinton
☑ Colchester
☑ Colebrook
☑ Columbia
☑ Cornwall
☑ Coventry
☑ Cromwell
☑ Danbury
☑ Darien
☑ Deep River
☑ Derby
☑ Durham
☑ East Granby
☑ East Haddam
☑ East Hampton
☑ East Hartford
☑ East Haven
☑ East Lyme
☑ East Windsor
☑ Eastford

Limit
Top 13 by AVG([Avg Number of Healthcare facilities])

Modeling Approach



	Research Focus	Model	R2 Score	RMSE
1	Geographic Factors	Random Forest	0.9543	588833.05
2	Geographic Factors	XGBoost	0.6889	1537147.15
3	Educational Access	Random Forest	0.0034	6142243.34
4	Educational Access	XGBoost	0.0034	6142310.58
5	Valuation + Year Sold	Random Forest	0.0821	2640336.49
6	Valuation + Year Sold	XGBoost	0.0821	2640370.78

Model Selection

Used Random Forest Regressor and XGBoost Regressor.

Target Variable

Target variable: Sale Amount.

Feature Selection

Features: Assessed Value, Sale Ratio, Number of Schools.

Data Split

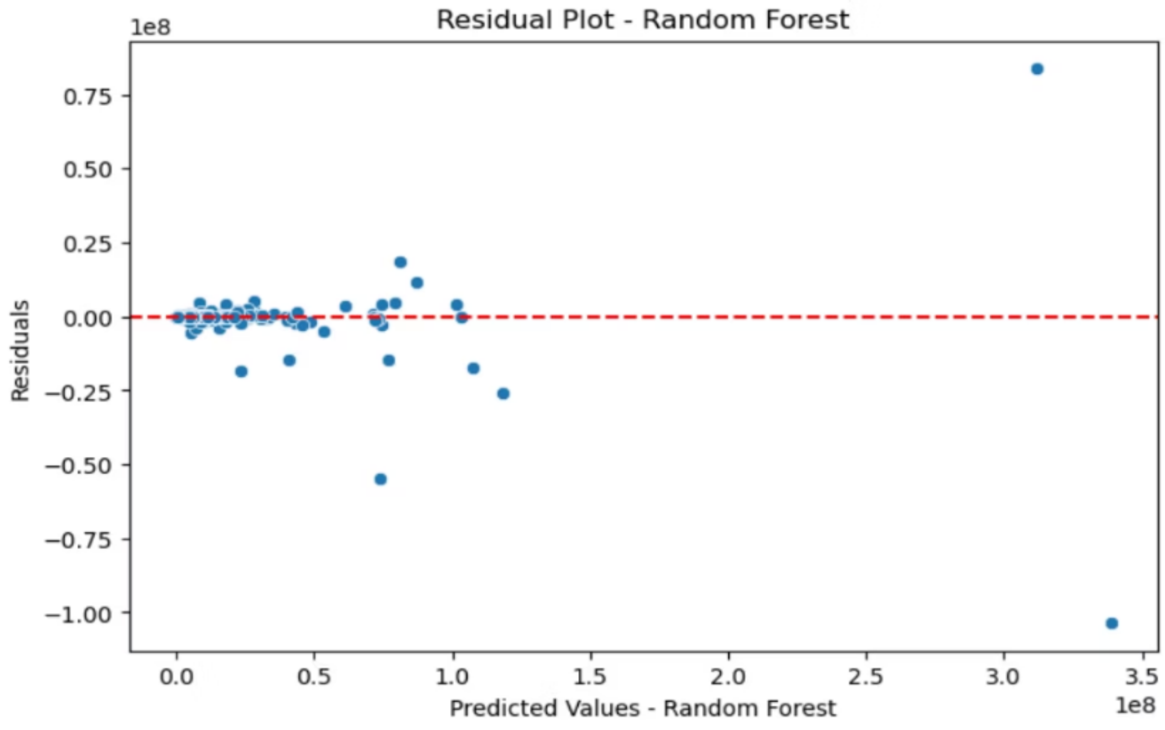
Train/Test split: 80% for training, 20% for testing.

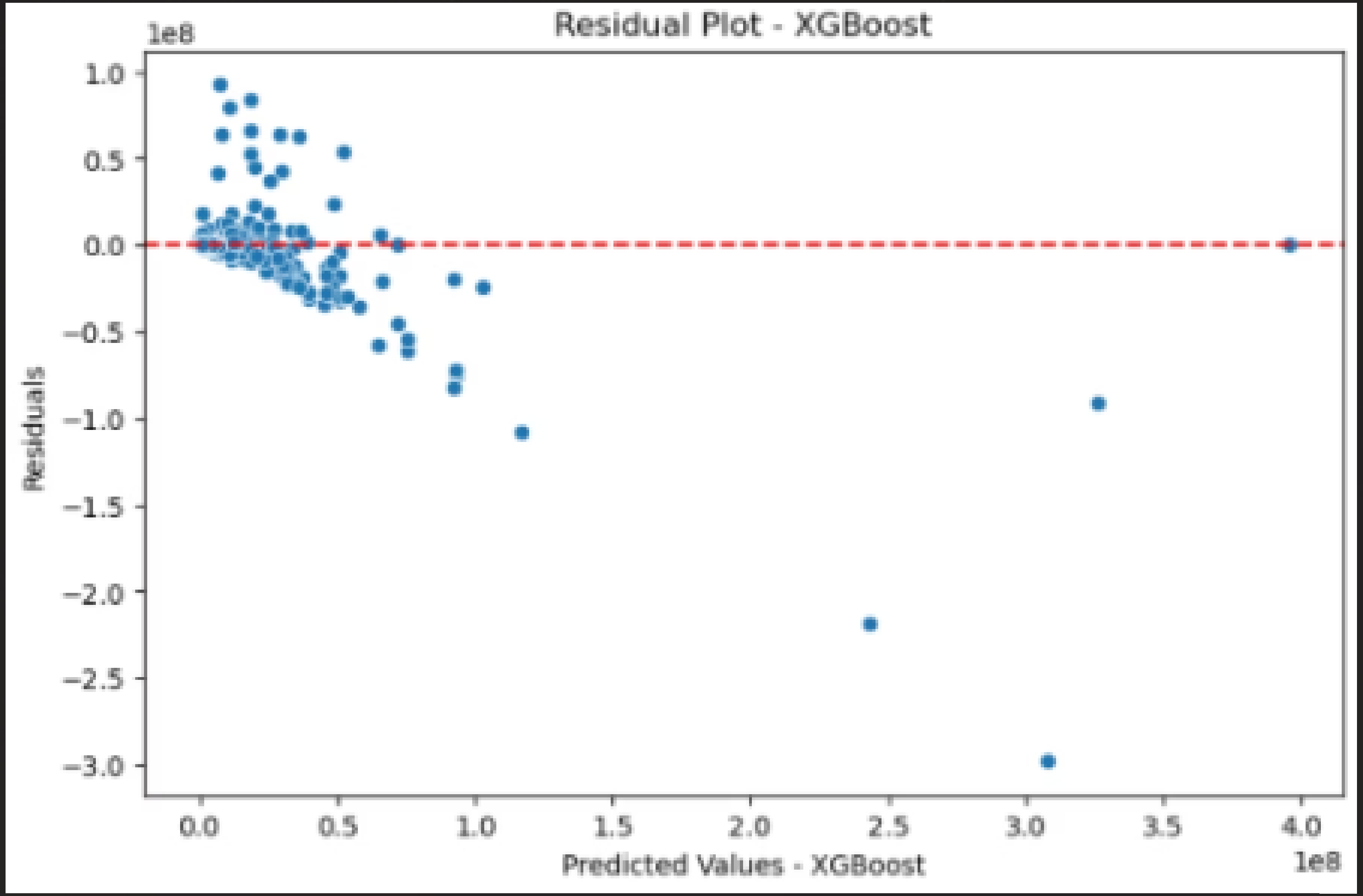
Evaluation Metrics

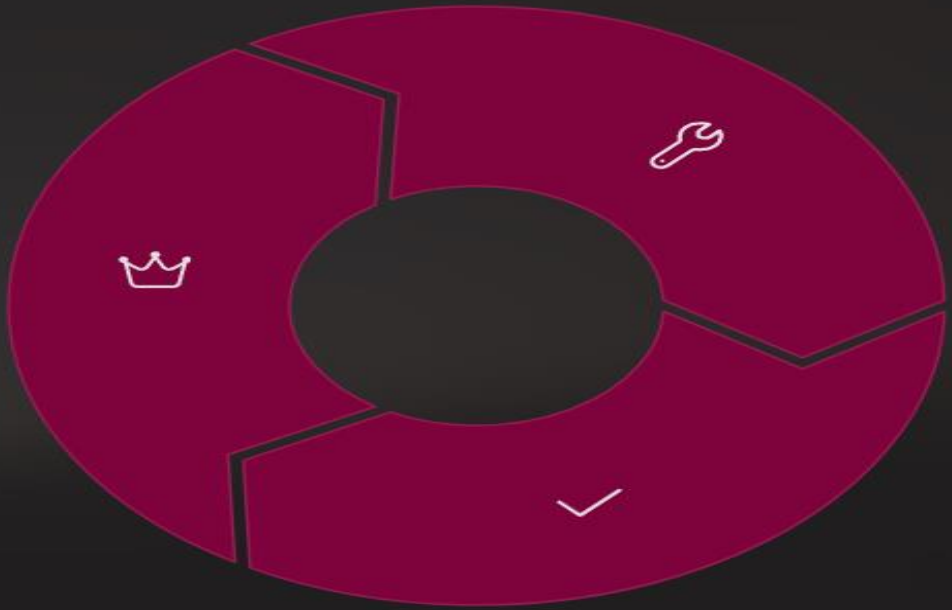
Evaluation metrics: R² Score and RMSE.

Model	R² Score	RMSE
Random Forest	0.9469	$499,925.10
XGBoost	0.3788	$1,710,477.60

Model Result







Best Performance

Random Forest achieved highest R² and lowest RMSE.

XGBoost Limitations

XGBoost struggled without hyperparameter tuning.

Final Selection

Random Forest selected as final model.

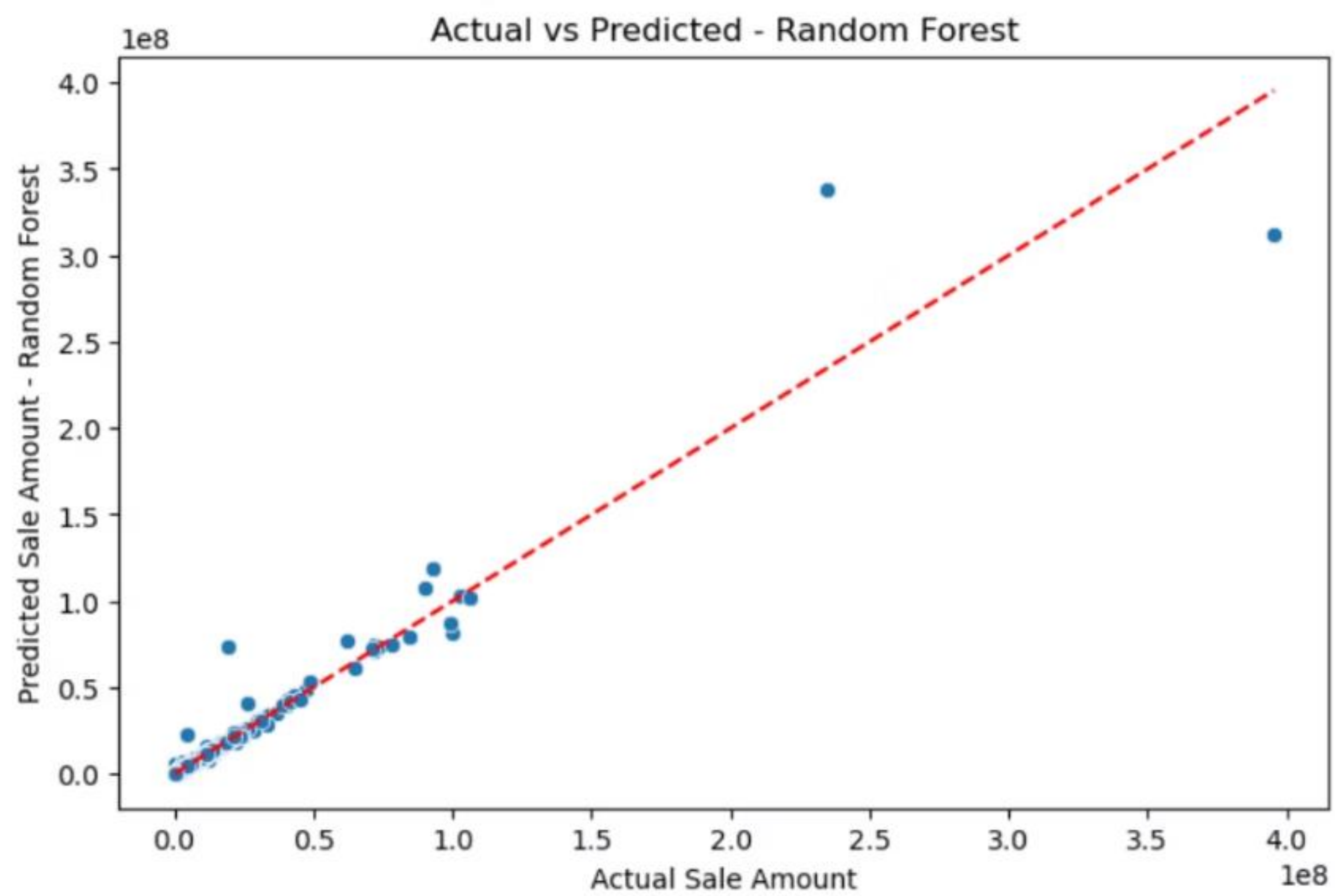
Predicted Analysis

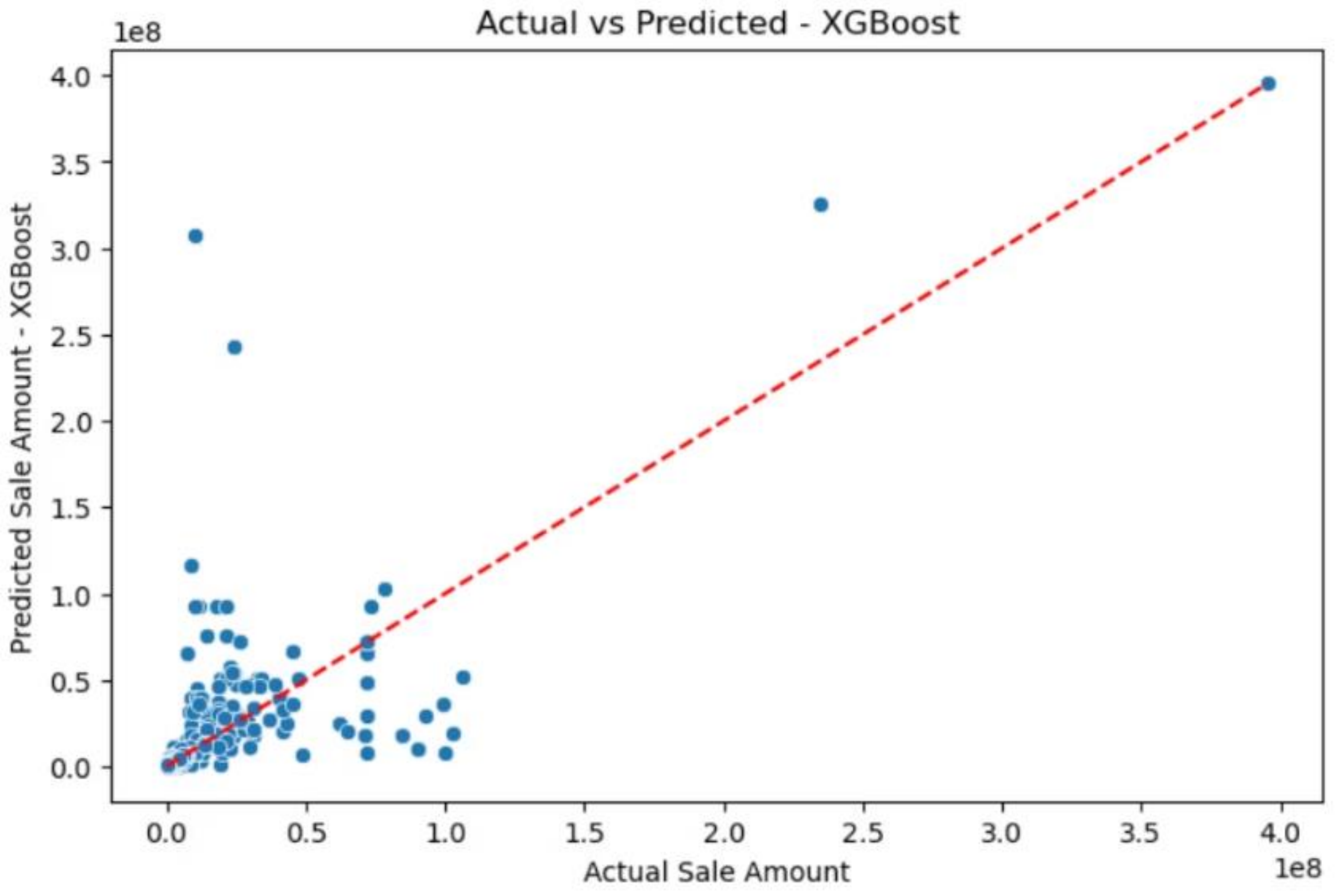
Random Forest

Residuals tightly clustered around zero. Predictions closely matched actual sale prices.

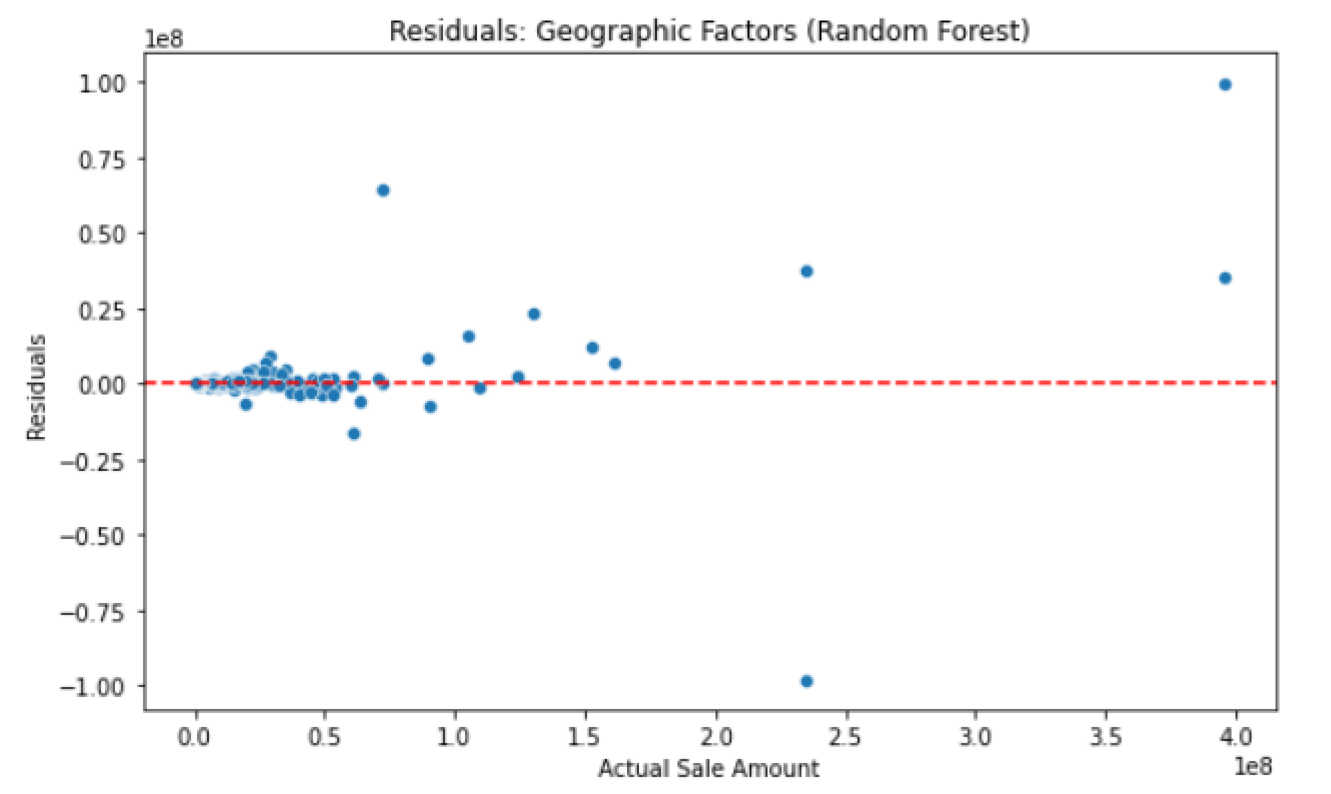
XGBoost

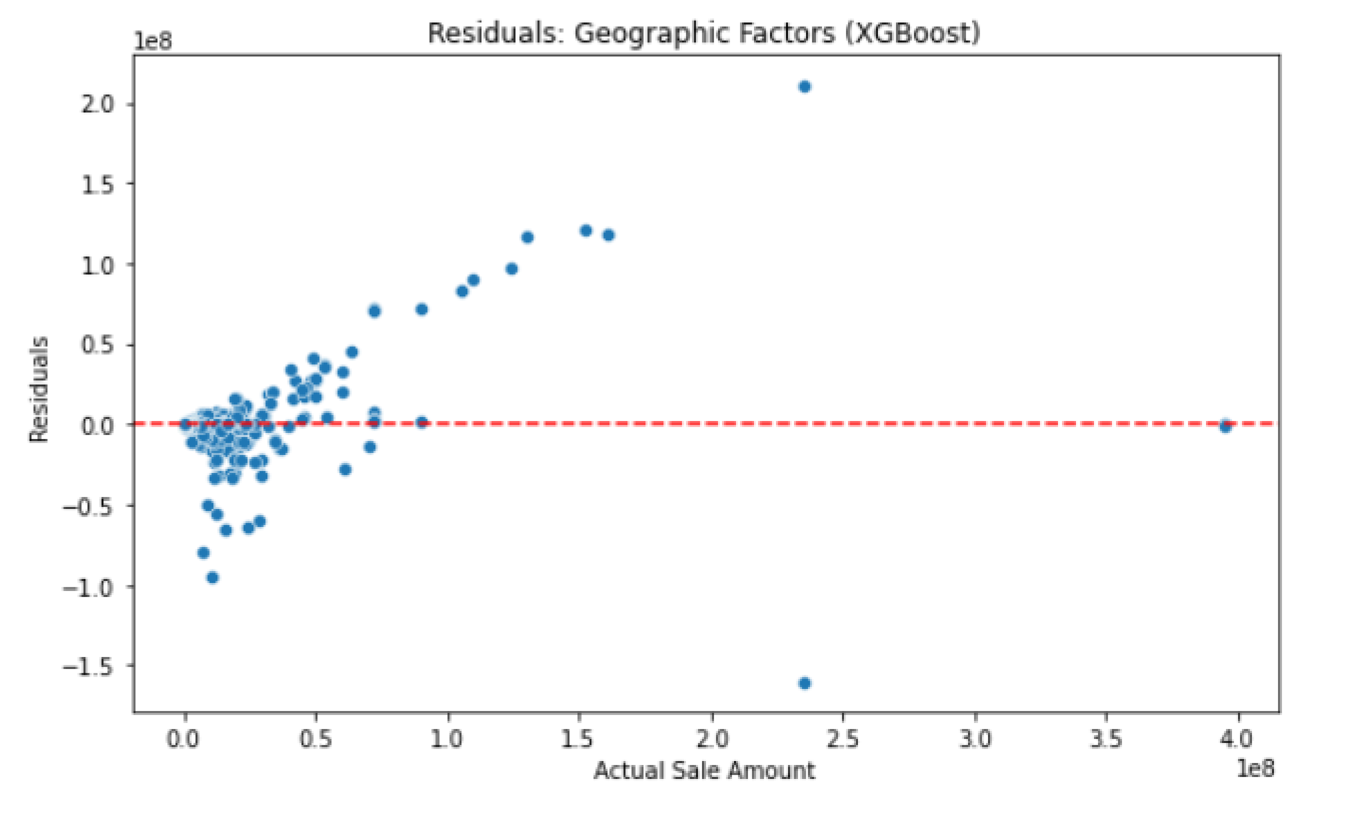
Residuals widely scattered, indicating poor fit. Predictions showed significant underestimation.





- **Geographic Factors**
- **Research Question 1:**
  Which geographic features (e.g., assessed value and sale ratio) most accurately predict property sale prices in Connecticut?
- **Models Used:** Random Forest, XGBoost
- **Best Performing Model:** Random Forest
- **Key Results:**
  - R^2 = **0.9543**
  - RMSE ≈ **$588,833**
- **Insight:**
  - Assessed value and sale ratio are highly predictive.
  - Geographic valuation explains most of the variability in sale price.

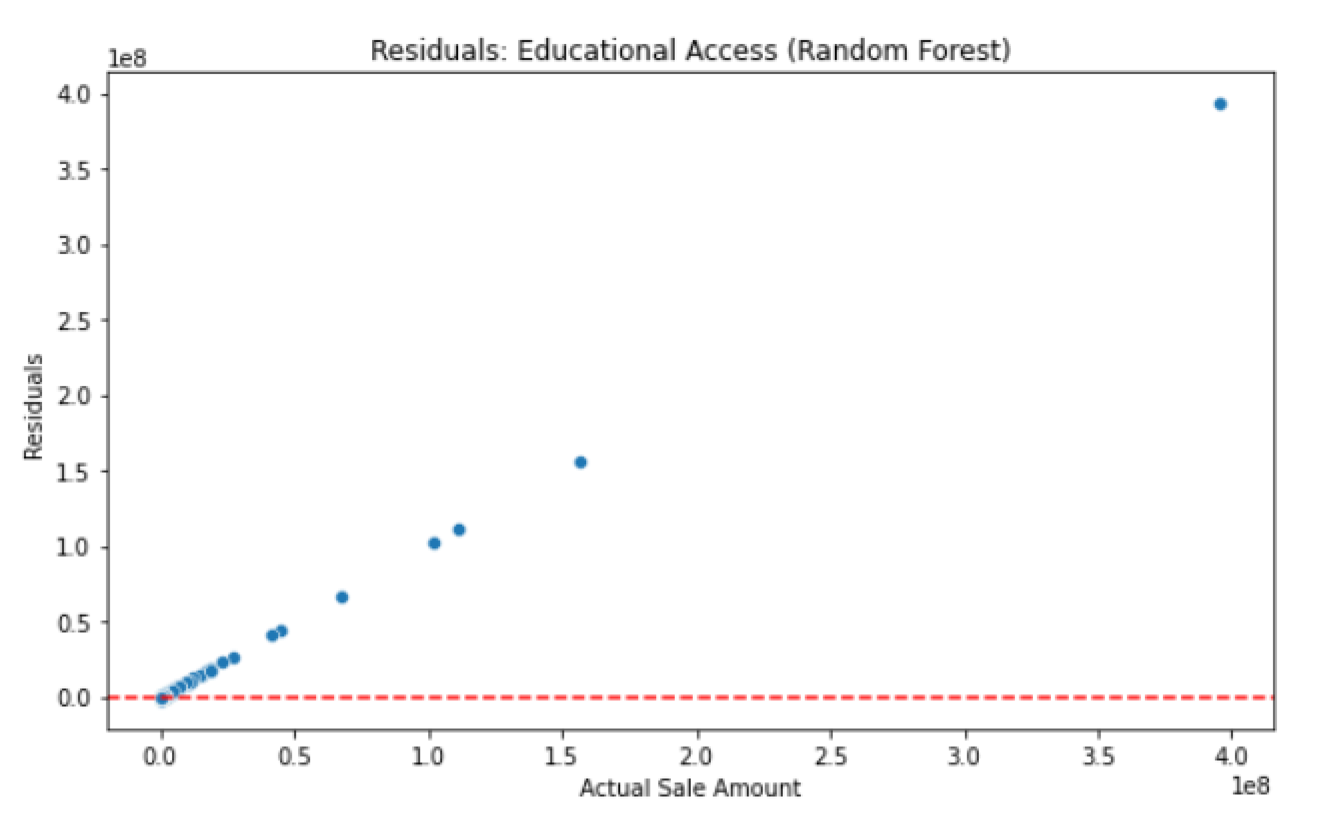


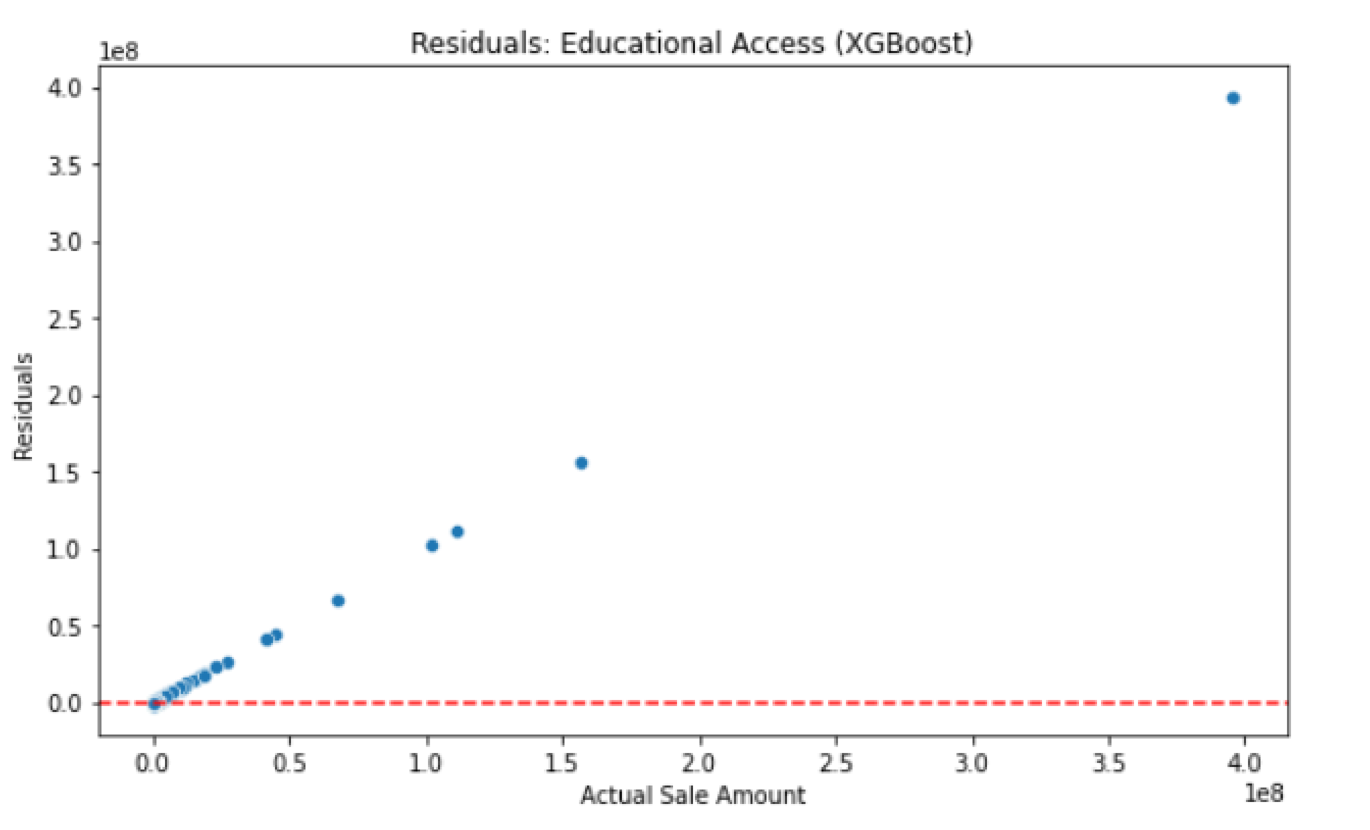


- **Educational Access**
- **Research Question 2:**
  How does access to educational institutions (based on available grade levels) influence residential property values?
- **Models Used:** Random Forest, XGBoost
- **Results for Both Models:**
  - R^2 = **0.0034**
  - RMSE ≈ **$6.14 million**

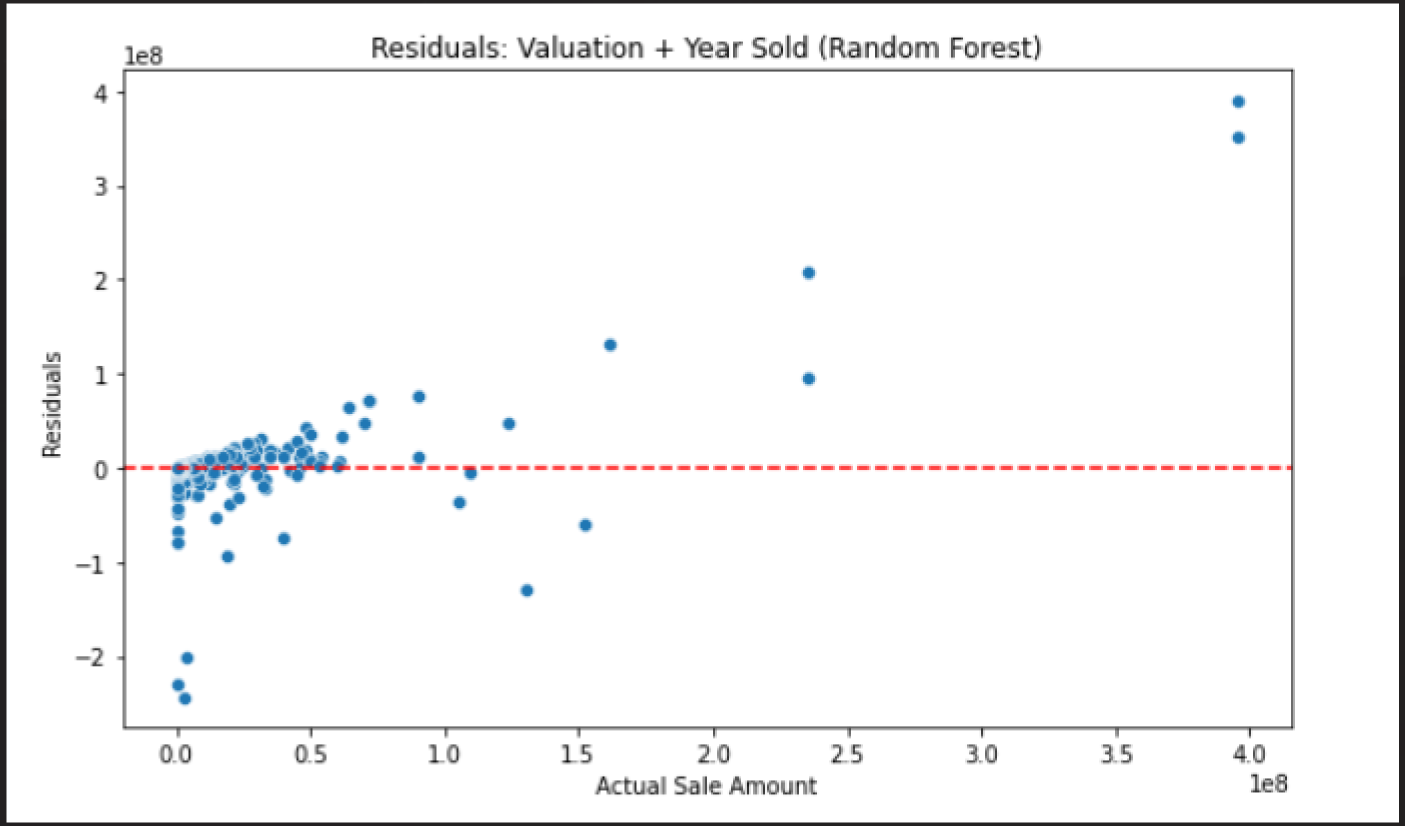
•Insight:

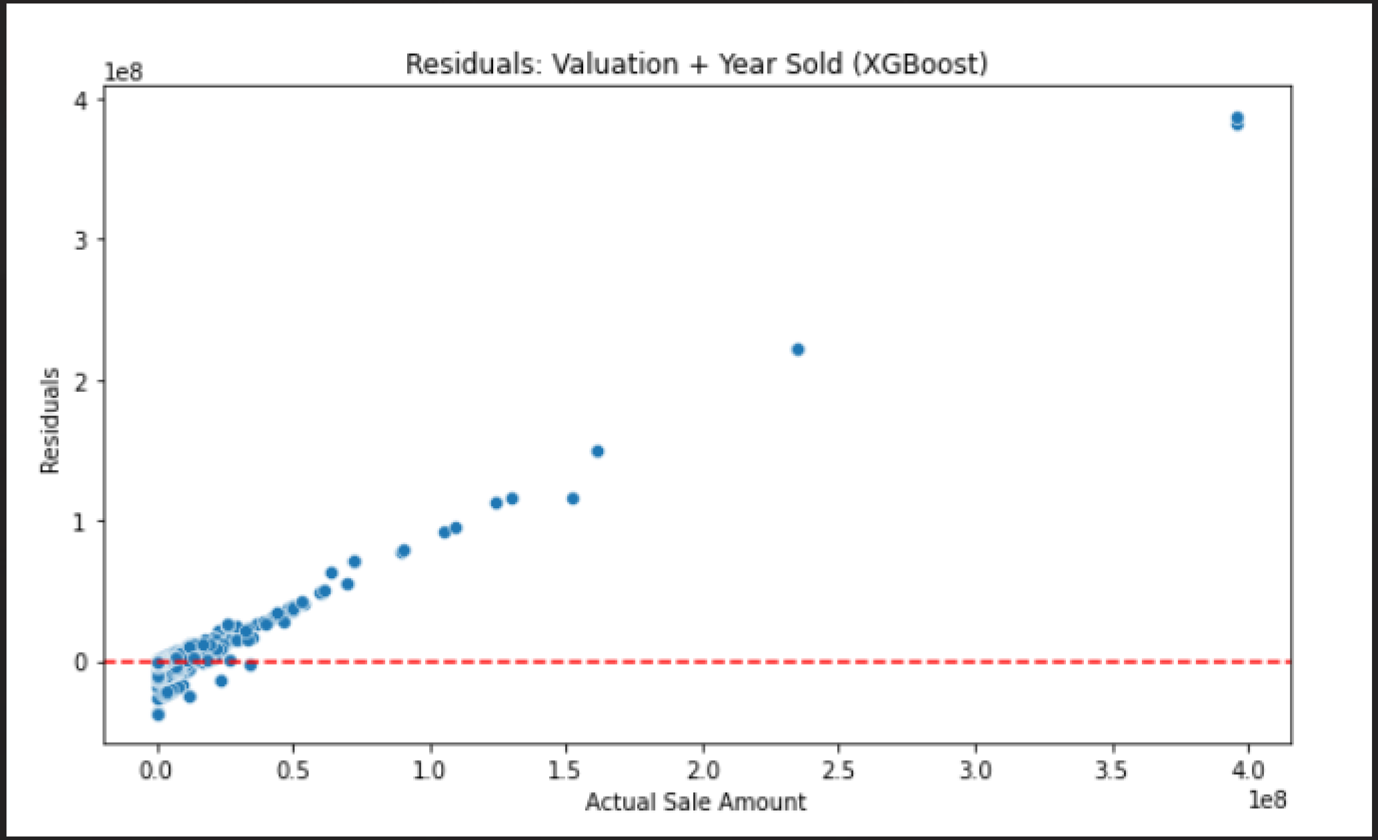
- School grade count alone (PreK–12) does not significantly affect sale price.
- May need to consider quality ratings or proximity in future studies.



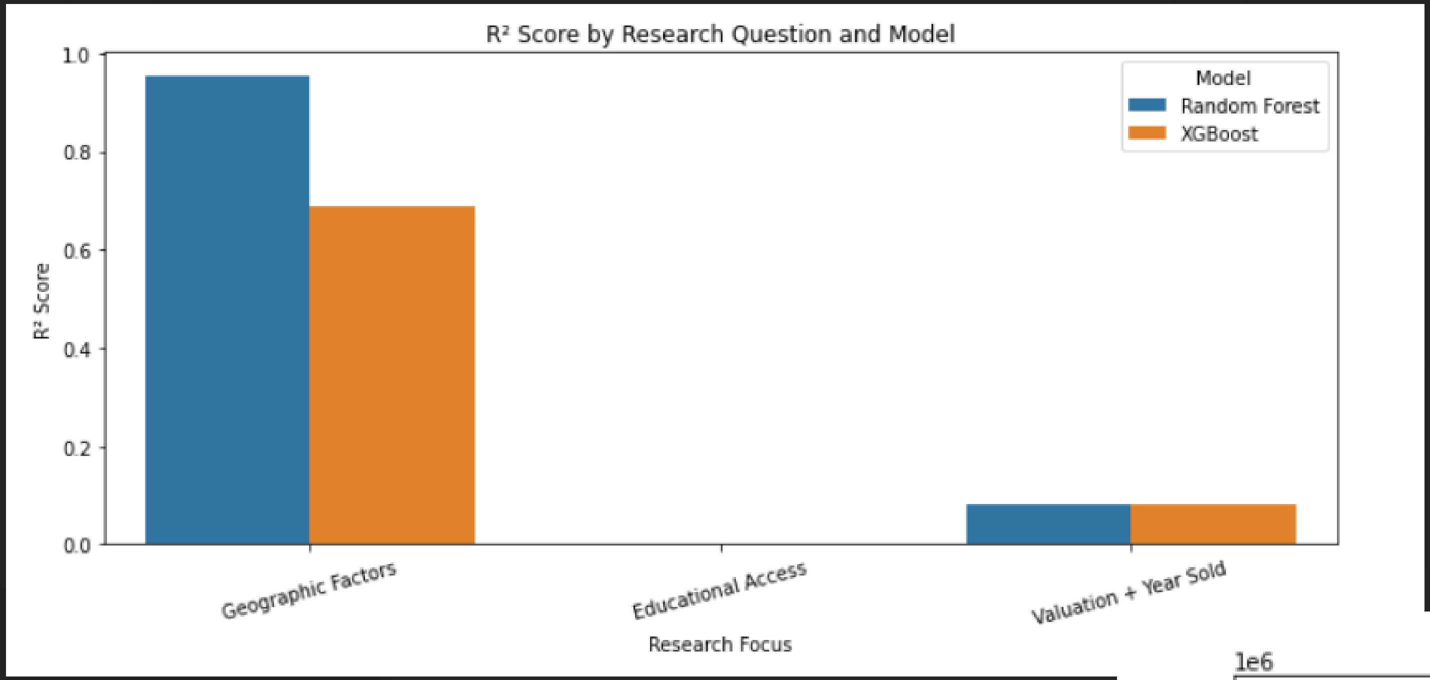


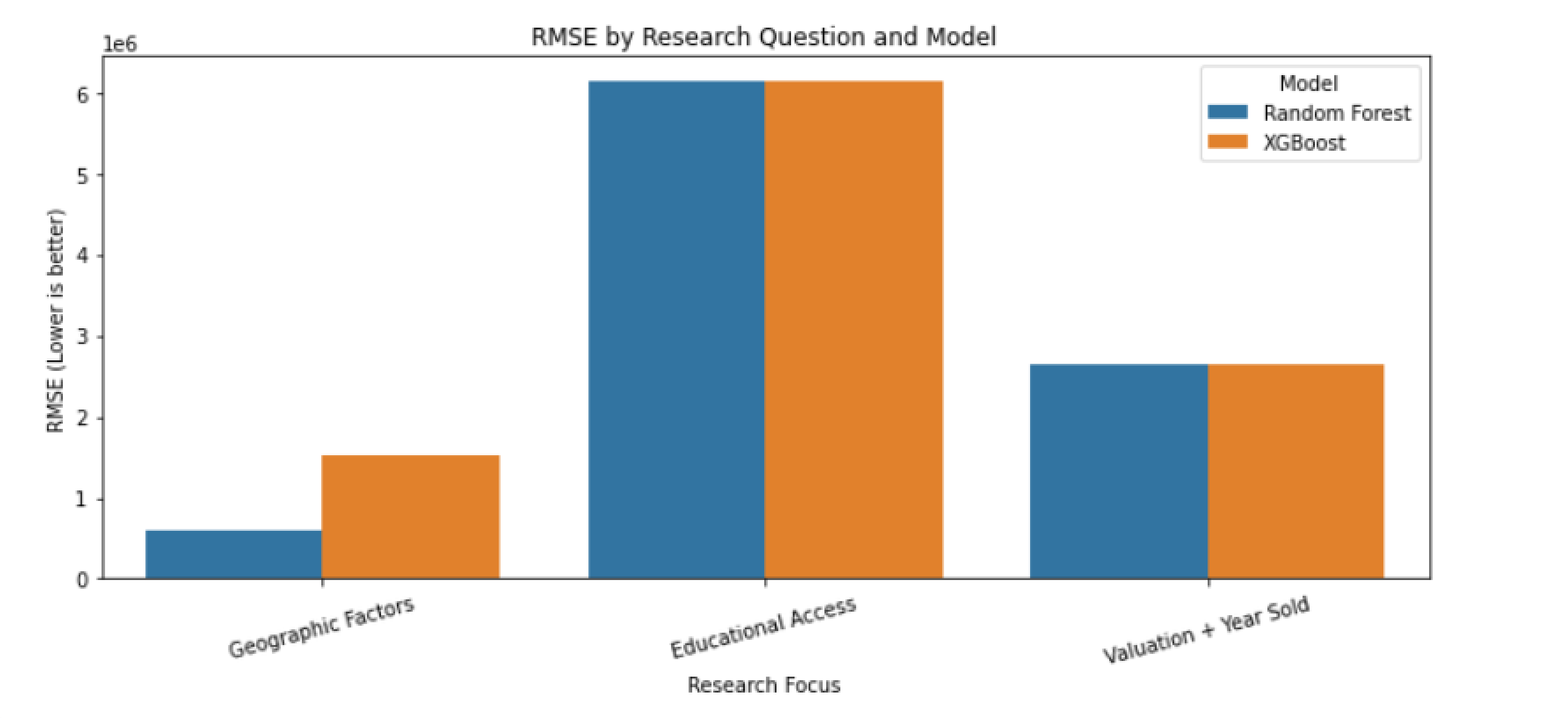
- **Valuation + Year Sold**
- **Research Question 3:**
  What is the impact of assessed value and year of sale on predicting final sale prices of residential properties?
- **Models Used:** Random Forest, XGBoost
- **Key Results:**
  - R^2 = **0.0821**
  - RMSE ≈ **$2.64 million**
- **Insight:**
  - Year of sale adds minimal improvement over assessed value alone.
  - Indicates housing value trends are mostly static or already captured.

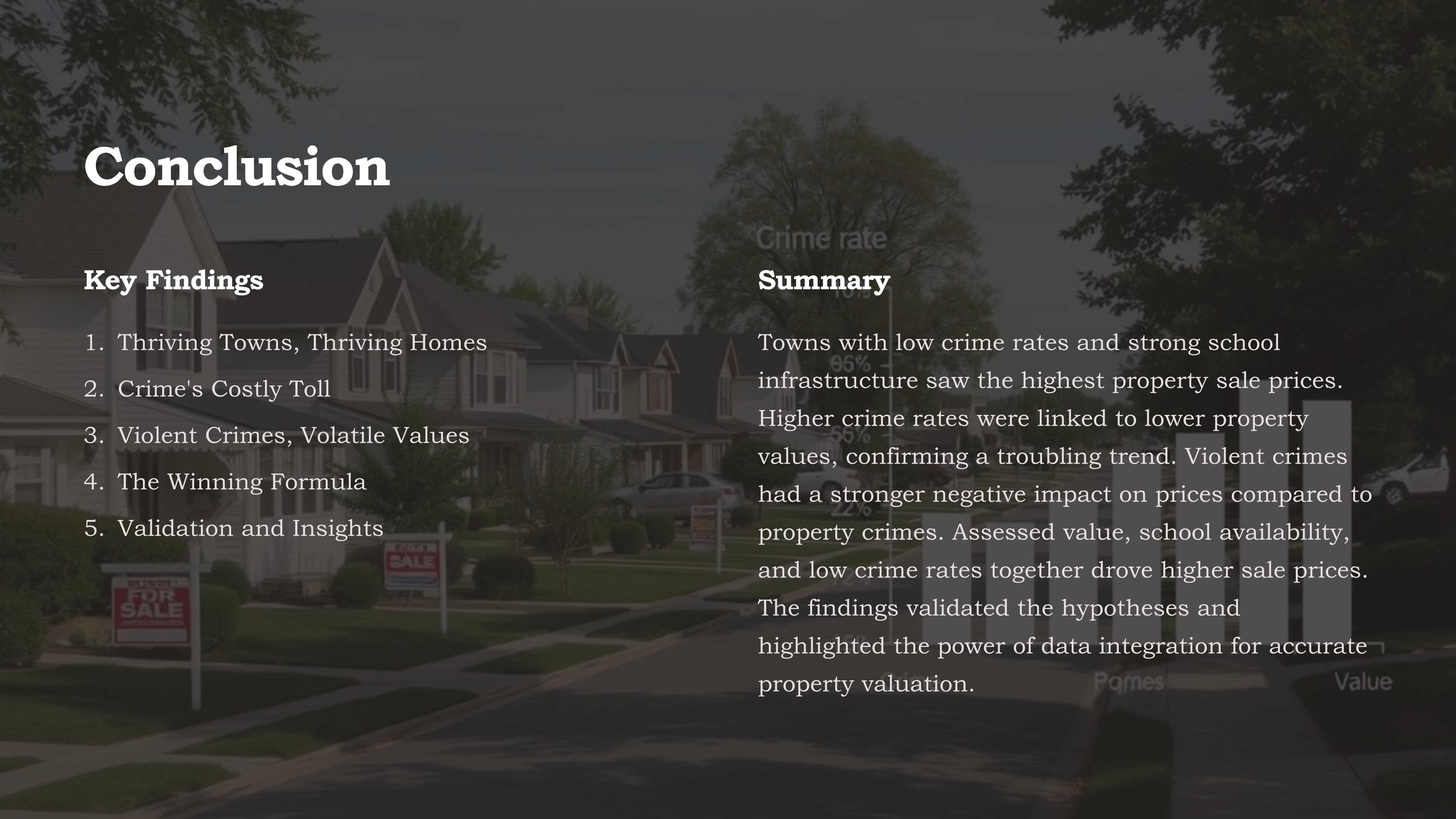




Model performance comparision







Conclusion

Key Findings

1. Thriving Towns, Thriving Homes
2. Crime's Costly Toll
3. Violent Crimes, Volatile Values
4. The Winning Formula
5. Validation and Insights

Summary

Towns with low crime rates and strong school infrastructure saw the highest property sale prices. Higher crime rates were linked to lower property values, confirming a troubling trend. Violent crimes had a stronger negative impact on prices compared to property crimes. Assessed value, school availability, and low crime rates together drove higher sale prices. The findings validated the hypotheses and highlighted the power of data integration for accurate property valuation.

Conclusion

1

Integrate Broader Datasets

Encompass economic indicators, demographic shifts, and housing market trends.

2

Investigate Machine Learning

Explore ensemble stacking and deep learning methodologies.

3

Enhance Data Quality

Implement refined outlier detection and removal strategies.

4

Boost Predictive Accuracy

Develop targeted feature engineering approaches for better model interpretability.

5

Forecast Future Trajectories

Expand analysis to predict future property value trends using time series analysis.

Future Work

1

Expand Data Sources

Incorporate additional external datasets, including economic, demographic, and housing trend data.

Advanced Modeling

Explore advanced machine learning techniques such as ensemble stacking and deep learning models.

Improve Data Quality

Implement more sophisticated outlier detection and removal techniques.

Enhance Features

Conduct feature engineering to enhance predictive performance and interpretability.

Time Series Forecasting

Extend the study to predict future property value trends using time series forecasting.

Thank you

Thank you for your time and attention! We hope this presentation provided valuable insights into property sale amount prediction in Connecticut.