Real Estate Sale Prices Against Socioeconomic Indicators

Team 5
Aravind Panchanathan - apanchan@gmu.edu
Venkata Lakshmi Parimala Pasupuleti - vpasupu2@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

AIT 582-003 Applications of Metadata in Complex Big Data Problems (Spring
2025)
Professor: Lam Phung
May 7 2025

Abstract

Several social and economic factors influence the real estate market, thus creating a dynamic system. Property values in the market demonstrate sensitivity to three fundamental elements that include criminal statistics along with educational and healthcare access quality, and broader financial security. This research analyzes real estate sale price trends throughout Connecticut from 2001 to 2022 by merging data from home sales with crime reports, school distribution, healthcare facilities, employment levels, and business activity statistics. The objective of this research work is to determine the joint effect of these factors on property value assessment between various towns throughout Connecticut.

The study uses Random Forest Regressor and Linear Regression models to determine which socioeconomic variables drive variances in home prices. The research begins with extensive dataset preparation, which includes data cleaning and merging geographic information and datasets into a single uniform data structure, as well as EDA for analyzing variable correlations and trends.

The analytical results show that property values fall when crime rates rise, but property prices rise when great schools and healthcare facilities, as well as a solid economic base, become available. Stamford and Fairfield distinguished themselves by high-quality public infrastructure and strong economic conditions, resulting in rising property values.

The analytical findings provide important direction for residential purchasers and investment groups, plus public administrators. The findings demonstrate how social indicators impact property costs, thus creating a useful tool for buyers' decisions while supporting official planning frameworks in housing policy and residential market development [1].

Keywords: Real Estate Prices, Crime Rates, Healthcare Access, School Quality, Economic Indicators, Random Forest, Linear Regression, Data Integration, Connecticut, Machine Learning.

1. Introduction:

The real estate market is a dynamic and multifaceted sector influenced by various socioeconomic factors. Factors, such as population density, crime rates, demographics, location, and school districts are only a few elements that can shape real estate trends. According to the U.S. Department of Treasury, housing prices are increasing at a faster rate than incomes, with demand surpassing supply (Feiveson, 2024) [4]. The increasing complexity of the real estate industry has made it harder for customers to make wellinformed decisions, owing to a lack of clear and comprehensive data.

This research aims to address this gap by investigating the relationship between real estate prices and socioeconomic characteristics in Connecticut. Using the Real Estate Sales 2001-2022 GL dataset, we will analyze trends and patterns to offer valuable insights for consumers, real estate professionals, and policymakers. Our objective is to assist stakeholders in better understanding how socioeconomic elements impact property prices, so they can make educated, data-driven decisions in a competitive market. By examining the effects of these variables over time, the project aims to offer a detailed analysis of the factors influencing property price changes in Connecticut between 2001 and 2022.

2. Research Questions and Hypotheses

The team has prepared the following research questions and hypotheses for this study.

Research Questions:

- How do assessed value and sale ratio influence property sale prices across towns?
- How does school grade availability (PreK to 12) affect property sale amounts?
- How do assessed value and year of sale influence property sale amounts?

Hypotheses:

- Higher crime rates are associated with lower property values.
- Regions with more schools will have an increased property value than those with lower school amounts.
- More serious and violent crimes will drive down a property's value more than light misdemeanors.

3. Literature Review:

3.1 Crime Rates and Property Prices

Various research studies demonstrate a direct opposition between the standing of home values and crime amounts within areas. The authors of Freakonomics, Levitt and Dubner (2005) [4], detail the relationship between elevated crime rates and diminished real estate values because property seekers typically choose safer areas instead of ones associated with peril. Glaeser and Gyourko (2008) [2] state that crime reduces housing demand in particular areas which leads to lower property prices. The property value increases in Connecticut at a slower pace within the high-crime rate urban areas such as Bridgeport and Hartford [1].

3.2 School Quality and Property Prices

Real estate values change substantially based on the quality level of schools within the vicinity. The findings of Glaeser and Gyourko (2008) reveal that families choose to reside near superior educational facilities because it increases their home property values. Improved school quality emerges as a main residential attractor for suburban neighborhoods with schools acting as core factors in choice of

residence. Residents in suburban Fairfield County Connecticut experience premium property sale prices because they live in communities with great schools (Feiveson, 2024) [4].

3.3 Population Density and Housing Prices

Real estate values change substantially based on the quality level of schools within the vicinity. The findings of Glaeser and Gyourko (2008) reveal that families choose to reside near superior educational facilities because it increases their home property values. Improved school quality emerges as a main residential attractor for suburban neighborhoods with schools acting as core factors in choice of residence. Residents in suburban Fairfield County Connecticut experience premium property sale prices because they live in communities with great schools (Feiveson, 2024) [3].

3.4 Economic Shifts and Housing Affordability

Economic changes, including broader macroeconomic shifts and demographic trends, have long-term effects on housing prices. Feiveson (2024) [4] discusses the increase in property values over the past two decades in Connecticut, attributing this growth to factors such as an influx of higher-income residents, a limited housing supply, and growing demand.

3.5 Dataset Source and Purpose

In this study, A Housing Area Analysis was conducted within the context of County preference sheets which combined components from both preferences rubrics. The dataset which was used in this study was provided by the Office of Policy and Management (OPM) of the state of Connecticut, specifically by Data and Policy Analytics (DAPA) section. DAPA was created in 2018 and updates the dataset every year, incorporating data from as far back as the early 2000s [5] [8].

3.6 Problem Statement:

The real estate market is currently at a record high, yet consumers often struggle to access comprehensive data for making well-informed decisions. This project aims to bridge that gap by:

- Identifying patterns in real estate prices based on socioeconomic factors.
- Examining the impact of these factors on property values and sale prices [2].

3.7 Dataset

The main dataset, Real Estate Sales 2001-2022 GL, was accessed from the repository of Connecticut government data. It contains key variables such as sale price, assessed value, property type, location, and town. To enhance the analysis, additional datasets on crime rates, school districts, and demographics will be integrated [8].

Key Variables:

Column Name	**NOIR Data Type**	**Description**
Serial Number	Nominal	Unique identifier for each property.
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.

4. Approach

4.1 Data Collection and Preparation

The dataset for this study is sourced from the Real Estate Sales 2001-2022 GL dataset provided by Connecticut's Office of Policy and Management, which contains detailed records of real estate transactions such as sale prices, locations, and assessed values (Office, 2018) [5]. Additional datasets will be included to cover crime rates and school count from publicly accessible sources, including U.S. Census data and local crime reports.

4.2 Data Cleaning and Integration

Data cleaning included addressing missing values, standardizing data formats, and removing outliers that could have potentially mislead the analysis. For instance, missing property sale price data were replaced using the median values, while geographic coding discrepancies will be handled by cross referencing with reliable sources. Data integration combined other datasets relating to other

4.3 Data Preprocessing and Transformation socioeconomic indicators such as crime.

4.3.1 Handling Missing Values

Real-world data contains multiple instances of missing values that affect both the reliability level and the performance quality of predictive models. The Real Estate Sales data involved records showing empty data fields in fundamental features including sale price and assessed property value. When dealing with these missing entries for optimal data preservation the study used median imputation techniques. The real estate pricing data supports better data integrity when using the median as replacement value because the median remains robust when encountering extreme values or outliers. This makes it a more suitable choice for real estate pricing data.

4.3.2 Date Formatting

The Date Recorded and List Year columns contained data as text values before any proper time analysis took place. The process of accurate time-series exploration and modeling received an improvement after proper conversion of these date fields into standardized format. Reformatting the data achieved two important outcomes: it standardized the content while allowing researchers to extract vital information regarding year and month data features. The newly added time-based features enabled teams to discover seasonal patterns while studying sales changes during multiple periods.

```
from pyspark.sql.functions import to_date, date_format
# Convert Year column to integer type
df = df.withColumn("List Year", df["List Year"].cast("integer"))
# Convert Assessed Value column to floating point number
df = df.withColumn("Assessed Value", df["Assessed Value"].cast("double"))
# Convert Sale Amount column to floating point number
df = df.withColumn("Sale Amount", df["Sale Amount"].cast("double"))
# Convert Sales Ratio column to floating point number
df = df.withColumn("Sales Ratio", df["Sales Ratio"].cast("double "))

df.show()
```

```
+-------------+---------+------------+--------------------+--------------+-----------+-----------+----------------+
|Serial Number|List Year|        Town|             Address|Assessed Value|Sale Amount|Sales Ratio|Residential Type|
+-------------+---------+------------+--------------------+--------------+-----------+-----------+----------------+
|      2020177|     2020|     Ansonia|       323 BEAVER ST|      133000.0|   248400.0|     0.5354|   Single Family|
|      2020225|     2020|     Ansonia|      152 JACKSON ST|      110500.0|   239900.0|     0.4606|    Three Family|
|      2020090|     2020|     Ansonia|         57 PLATT ST|      127400.0|   202500.0|     0.6291|      Two Family|
|       200500|     2020|        Avon|       245 NEW ROAD|      217640.0|   400000.0|     0.5441|   Single Family|
|       200121|     2020|        Avon|       63 NORTHGATE|      528490.0|   775000.0|     0.6819|   Single Family|
|        20058|     2020| Barkhamsted|   46 RATLUM MTN RD|      203530.0|   415000.0|0.490433735|   Single Family|
|       200046|     2020|Beacon Falls|       34 LASKY ROAD|      158030.0|   243000.0|     0.6503|   Single Family|
|       200016|     2020|Beacon Falls|        9 AVON COURT|       65590.0|   100000.0|     0.6559|           Condo|
|      2020360|     2020|      Berlin|     94 PERCIVAL AVE|      140600.0|   190790.0|     0.7369|   Single Family|
|        20281|     2020|      Bethel|   16 OXFORD STREET|      170800.0|   307000.0|     0.5563|   Single Family|
|        20364|     2020|      Bethel|1308 LEXINGTON BO...|      195300.0|   365000.0|      0.535|           Condo|
|        20423|     2020|      Bethel|10 CASTLE HILL ROAD|      219870.0|   325000.0|     0.6765|   Single Family|
|        20097|     2020|      Bethel|   8 BLACKMAN AVENUE|      264040.0|   445000.0|     0.5933|      Two Family|
|       200008|     2020|   Bethlehem|    34 HIGHLAND ROAD|       82000.0|   106000.0|     0.7735|   Single Family|
|        20062|     2020|      Bolton|    39 STONEHEDGE LN|      189400.0|   273750.0|     0.6918|   Single Family|
|       200305|     2020|    Branford|49 ROSE ST TOWERS...|       97800.0|   147000.0|     0.6653|           Condo|
|       200400|     2020|    Branford|         460 MAIN ST|      180500.0|   355000.0|     0.5084|      Two Family|
```

4.3.3 Standardizing ZIP Codes

ZIP code data was inconsistent across records, with some entries missing leading zeros (common in Northeastern states like Connecticut). Standardizing ZIP codes was necessary to ensure correct geographical joints and aggregations.

4.3.4 Crime, Healthcare, and School Data Preprocessing

Crime Rates Calculation:

To enable comparison across towns, the research standardized total crime counts by converting them into crime rates according to population per 100,000 people.

Healthcare Facility Counts:

The healthcare facility data was grouped according to the town for identifying the number of hospitals and clinics in the area.

School Counts:

The evaluation determined the school count numbers in each district through aggregation techniques to match real estate town data.

An initial processing method integrated socioeconomic factors into the main real estate data for upcoming analysis steps.

4.3.5 Dataset Merging

After cleaning real estate sales data, it was merged with separate databases containing crime statistics information alongside healthcare facility information and school data as well as economic indicators. A successful combination of datasets depended on the 'Town' field because it served as the primary connection between all information sources.

The left join approach protected the real estate transaction records from getting lost during data combination. The left-hand joint approach maintained all property sales records and added essential

socioeconomic information from the other datasets making the data more suitable for thorough analysis.

```
eco_df = eco_df.withColumnRenamed("Year", "List Year")
df_eco_join_df = final_df.join(eco_df, ["List Year", "Town"], "inner")
```

```
df_eco_join_df = df_eco_join_df.withColumn("BusinessEstablishment", df_eco_join_df["BusinessEstablishment"].cast("integer"))
df_eco_join_df = df_eco_join_df.withColumn("EmploymentData", df_eco_join_df["EmploymentData"].cast("integer"))
df_eco_join_df = df_eco_join_df.withColumn("EmploymentWages", df_eco_join_df["EmploymentWages"].cast("integer"))
df_eco_join_df = df_eco_join_df.withColumn("UnemploymentRates", df_eco_join_df["UnemploymentRates"].cast("double"))
```

```
display(df_eco_join_df)
```

```
df_eco_join_df.groupBy("List Year").agg(count("Town").alias("Number of Towns")).orderBy("List Year").show()
```

```
+---------+---------------+
|List Year|Number of Towns|
+---------+---------------+
|     2006|            111|
|     2007|            107|
|     2008|            111|
|     2009|            130|
|     2010|            130|
|     2011|            115|
|     2012|            114|
|     2013|            106|
```

Spaces: 4

4.4 Statistical Analysis and Modeling

To explore the relationships between real estate prices and socioeconomic factors, we used a decision tree model with multiple regression analysis. This method helped in determining how each factor influenced property prices while accounting for the impact of other variables. The model will analyze crime rates and school quantity to assess their individual effects on property sale prices, allowing us to estimate the relative significance of each factor. We will also perform exploration data analysis (EDA) to visually represent trends in the data, using tools such as heatmaps, scatterplots, and correlation matrices.

4.5 Geospatial Analysis

Geospatial analysis will be performed to determine the geographic distribution of property values and the regions most impacted by socioeconomic factors. This will entail utilizing GIS tools, such as Tableau or QGIS, to create heatmaps and geographical distributions of property prices across Connecticut, allowing us to investigate how spatial patterns relate to crime rates, school quality, and density.

4.6 Data Visualization:

The team will utilize tools such as Tableau, matplotlib, and Seaborn to create visualizations (e.g., heatmaps, scatterplots, bar charts) that highlight trends and patterns.

4.7 Methodology

Model: The XGBoost Decision Tree model was the conduit to allow us to perform regression analysis and other statistical analysis on the data.

Statistical Analysis: Use regression models to measure the influence of socioeconomic factors on property values.
Visualization: Generate heatmaps, scatterplots, and bar charts to effectively showcase trends.
Tools: Leverage Python (Pandas, NumPy, Scikit-learn), Tableau, and SQL for data processing, analysis, and visualization.

4. 7.1 Exploratory Data Analysis (EDA)

Data gathering during the process of merging required connecting the cleaned real estate data with information from crime healthcare school and economic files. The join operation occurred mainly on the 'Town' column since it represented the single matching attribute among all databases.

4. 7.1.1 Correlation Heatmap

EDA started with an analysis of the relationship strength between Sale Amount and its independent socioeconomic variables including crime rates and healthcare access as well as school counts and economic indicators. The relationship between dependent and independent variables became apparent through generation of a correlation heatmap.

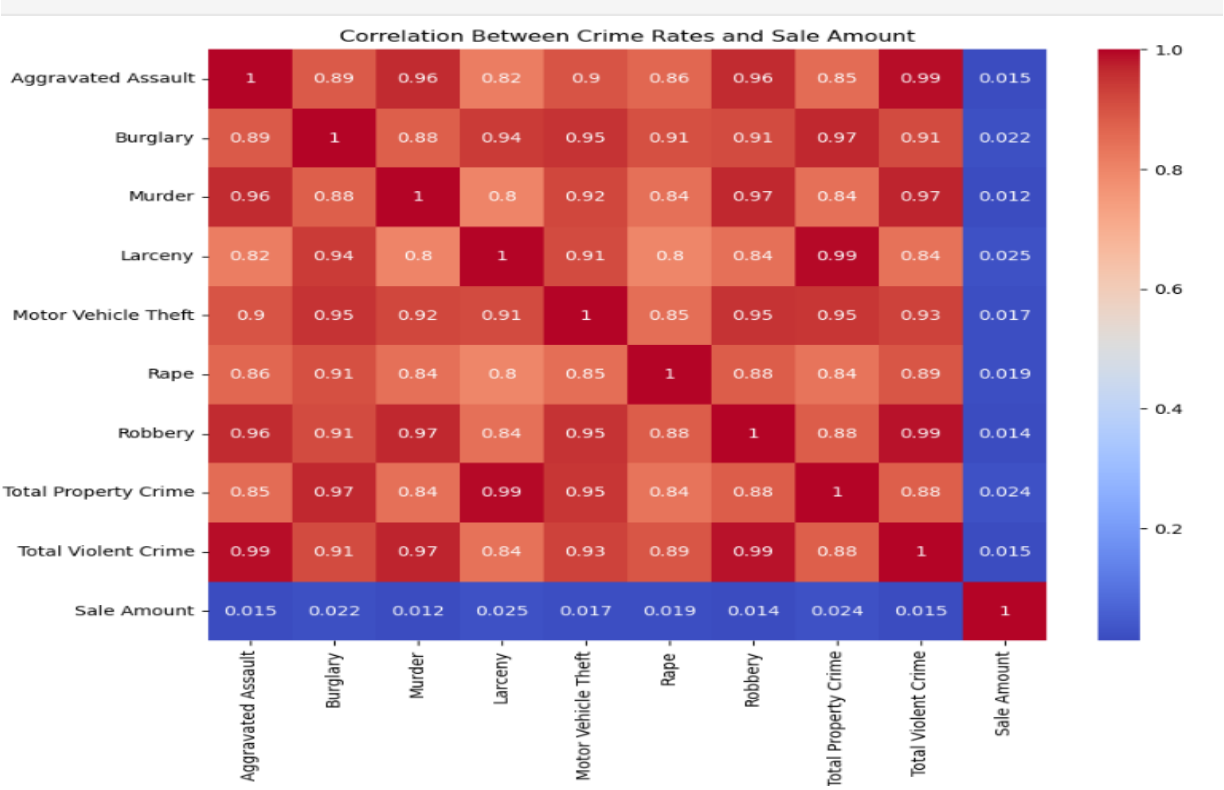

Fig.1. *Relation between Crime Rates and Sales amount*

The research established positive links between house prices and both school establishments and healthcare institutions in the area. Property sale prices decrease when local crime rates increase thus creating a negative correlation between these two measures. The information gathered here functioned as vital data for picking characteristics before conducting modeling procedures.

4. 7.1.2 Distribution of Sale Amounts

The analysis of sale amount distribution allowed researchers to understand the target variable character and determine whether any conversion steps were needed. The data display was created using histogram along with Kernel Density Estimation (KDE).

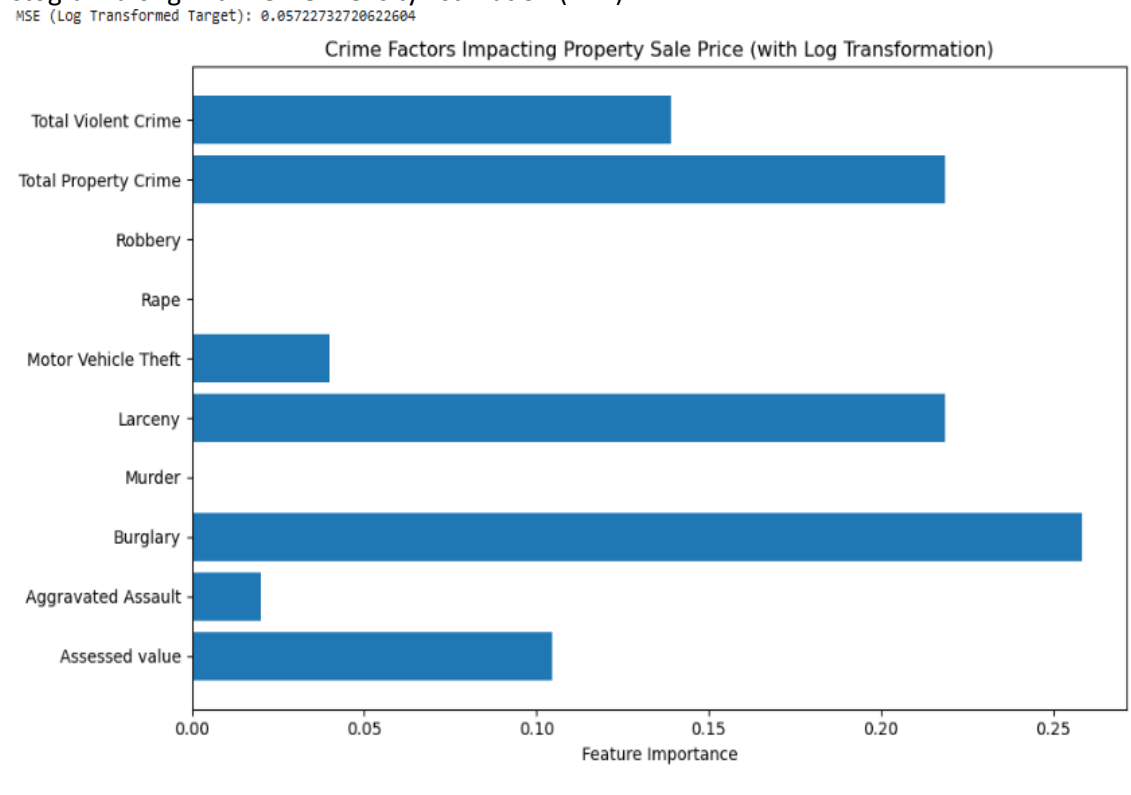

Fig. 2. *crime Feature importance in sales price*.

Distribution data showed right-skewness because lower sales values dominated the dataset but some properties generated high-value exceptions. Studies showed that converting the sale amount data through logarithmic methods were necessary for both data normalization and improved modeling results.

4.7.1.3 Crime Rate Trends vs. Sale Prices

Line visualizations were used to evaluate the impact of yearly crime changes on housing prices through the comparison of crime rates against average property sale figures. The visualization tools mainly relied on Tableau software to develop clear and interactive graphical displays.

Purpose:

The line charts helped to locate episodes during which increased criminal activity occurred simultaneously with reducing property values.

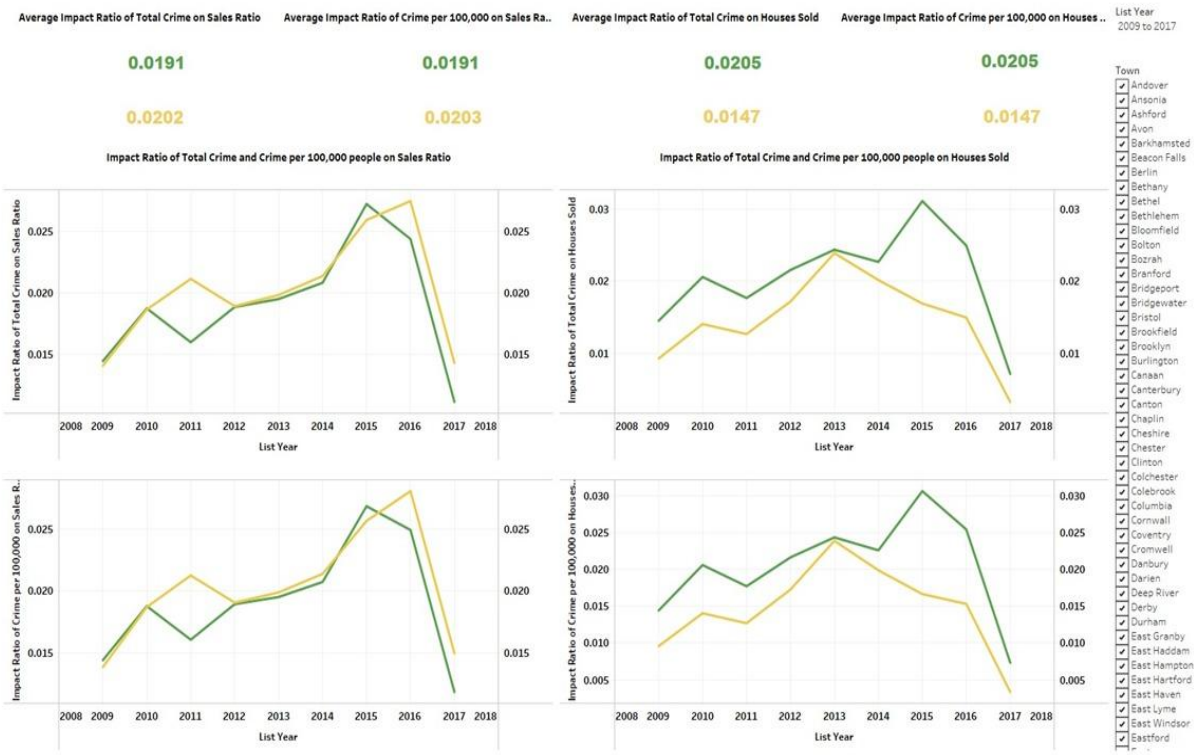

Fig.2. Crime rates vs sales price

A study was conducted to demonstrate if decreasing crime in the area led to improved housing market sales. The home sale dollar value together with the ratio of sales increased during crime reduction intervals beginning in 2013. The collected data confirmed that rising crime statistics produce adverse effects on property market values.

5. Modeling Approach

The analysis employed two regression methods to determine the effects between property market prices and demographic statistics in Connecticut. The prediction models evaluated property sale prices through combinations of crime rates with school access and healthcare facilities and employment metrics and additional economic indicators.

Random Forest vs. XGBoost

For both methods, we employed the modeling setup shown below:

- **Feature Set:** Independent variables included crime data (e.g., burglary, robbery, total property crime, total violent crime) and, in certain models, Assessed Property Value as a socioeconomic variable.
- **Target Variable:** Sale Amount was examined in log-transformed form to decrease extreme values and stabilize variance.
- **Data Scaling:** Standardized features using StandardScaler for consistent distribution.
- **Train-Test Split:** The data was split into 80% training and 20% testing (test_size=0.2, random_state=42).
- **Model Parameters:**
    - **Random Forest Regressor:** n_estimators=100, random_state=42
    - **XGBoost Regressor:** n_estimators=100, learning_rate=0.1, max_depth=6, random_state=42

Model 1: Random Forest Regressor

- Handles nonlinear relationships effectively.
- Performs well on high-dimensional data.
- Quantity Ensemble Averaging (also known as Bagging) functions as an overfitting prevention method.
- Provides feature importance for interpretability.

Model 2: XGBoost Regressor

- The algorithm applies a method which enhances weak learners one step at a time through boosting gradients.
- The method shows excellent capability when working with missing values and advanced relationships within the data.
- Offers regularization to avoid overfitting.

Model Evaluation Metrics Used:

- **R^2 Score (Coefficient of Determination)**: Measures how well the variance in sale price is explained by the model.
- **RMSE (Root Mean Squared Error)**: Measures the average magnitude of prediction errors.

Model Performance Comparisons

Model	Target Type	**Model**	RMSE	Notes
Random Forest	Raw Sale Amount	~0.96	High	Strong fit, but having outliers
XGBoost	Raw Sale Amount	Lower	High	Underperformed with higher residual spread
Random Forest	Log Sale Amount	0.9977	Lower	Best performance, stable predictions

Model performance comparisons

6. Model Performance Visuals

6.1 Residual Plot Analysis

The residual plots examine how accurately Random Forest and XGBoost models predict and what size of errors they produce.

6.2 Random Forest Residuals:

Random Forest predictions accurately spread across the entire range of sale amounts because the residuals remain connected to zero.

A small number of properties having exceptionally high sale prices produce noticeable outliers in the residual measurements. The distribution exhibits uniform consistency among the entire width.

The analysis pattern demonstrates the Random Forest model properly interprets feature-sale amount relationships through a minimally biased approach.

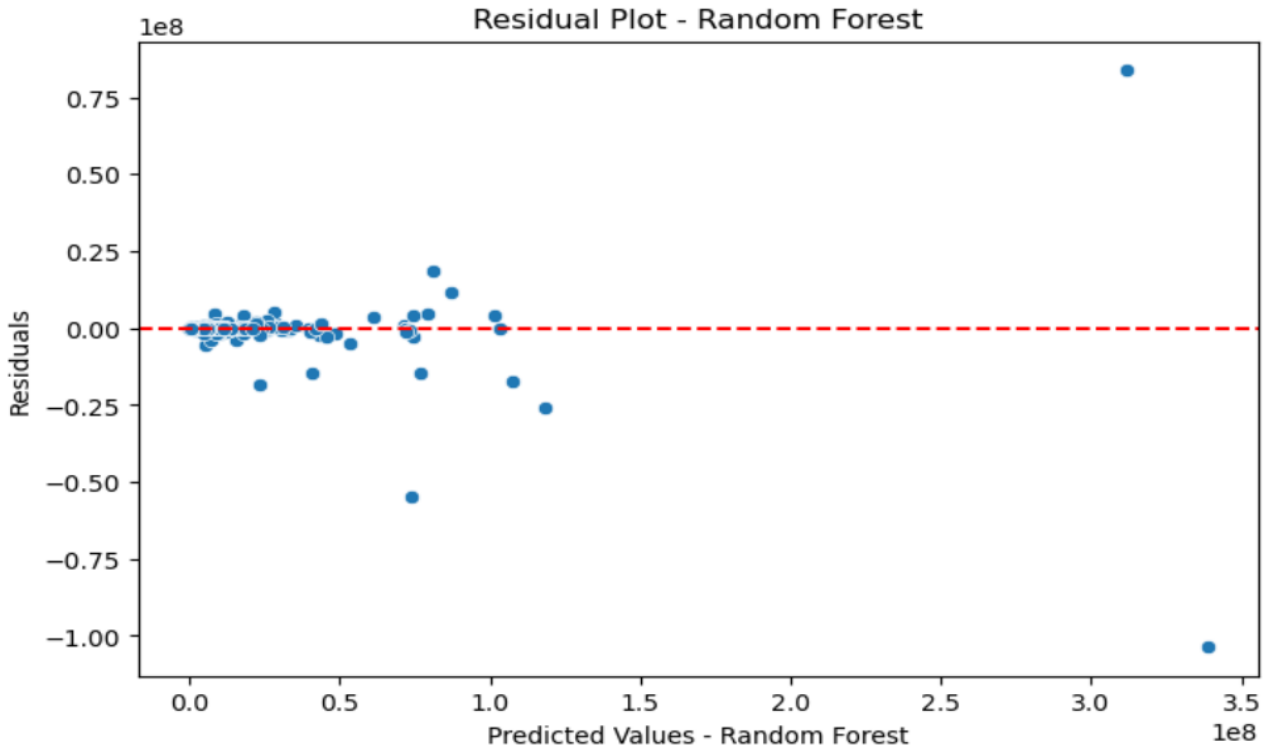

Fig.3. *Random forest prediction analysis-residual plot*.

6.3 XGBoost Residuals:

The residuals have more dispersion and non-symmetrical characteristics than Random Forest.

The distribution of values shows numerous observations situated above or under zero which indicates systematic errors during prediction of specific sale values.

Analysis reveals that XGBoost adopted an improper fitting method since evidence points toward substantial errors in predicting properties at high sale value points.

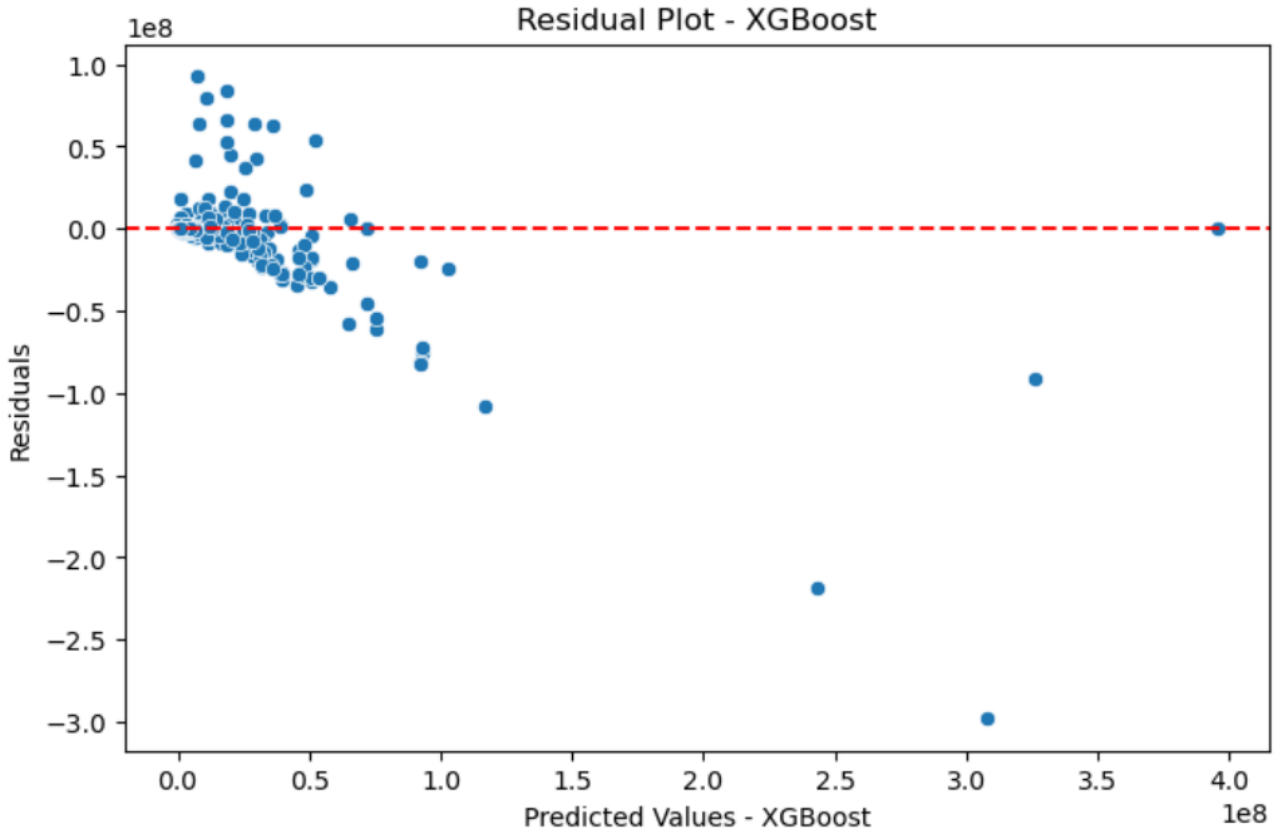

Fig.4. *XG Boost prediction analysis-Residual plot.*

The Random Forest model outperforms XGBoost not only in terms of R^2 and RMSE but also in residual stability.

Random Forest produces better reliable and consistent predictions regarding Connecticut property sale prices when looking at modeling residuals.

6.4 Actual vs Predicted Plot Analysis:

The Actual vs Predicted plots demonstrate the degree of match between model-predicted values and factual sales data.

6.5 Random Forest Actual vs Predicted:

The red diagonal line serves as the best indicator of near-perfect prediction because the scatter points tightly group together. The model correctly describes the price movement patterns for most of the properties on sale. Very expensive property data points exist as outliers while the Random Forest forecasting model represents minimum errors across most of the prediction range.

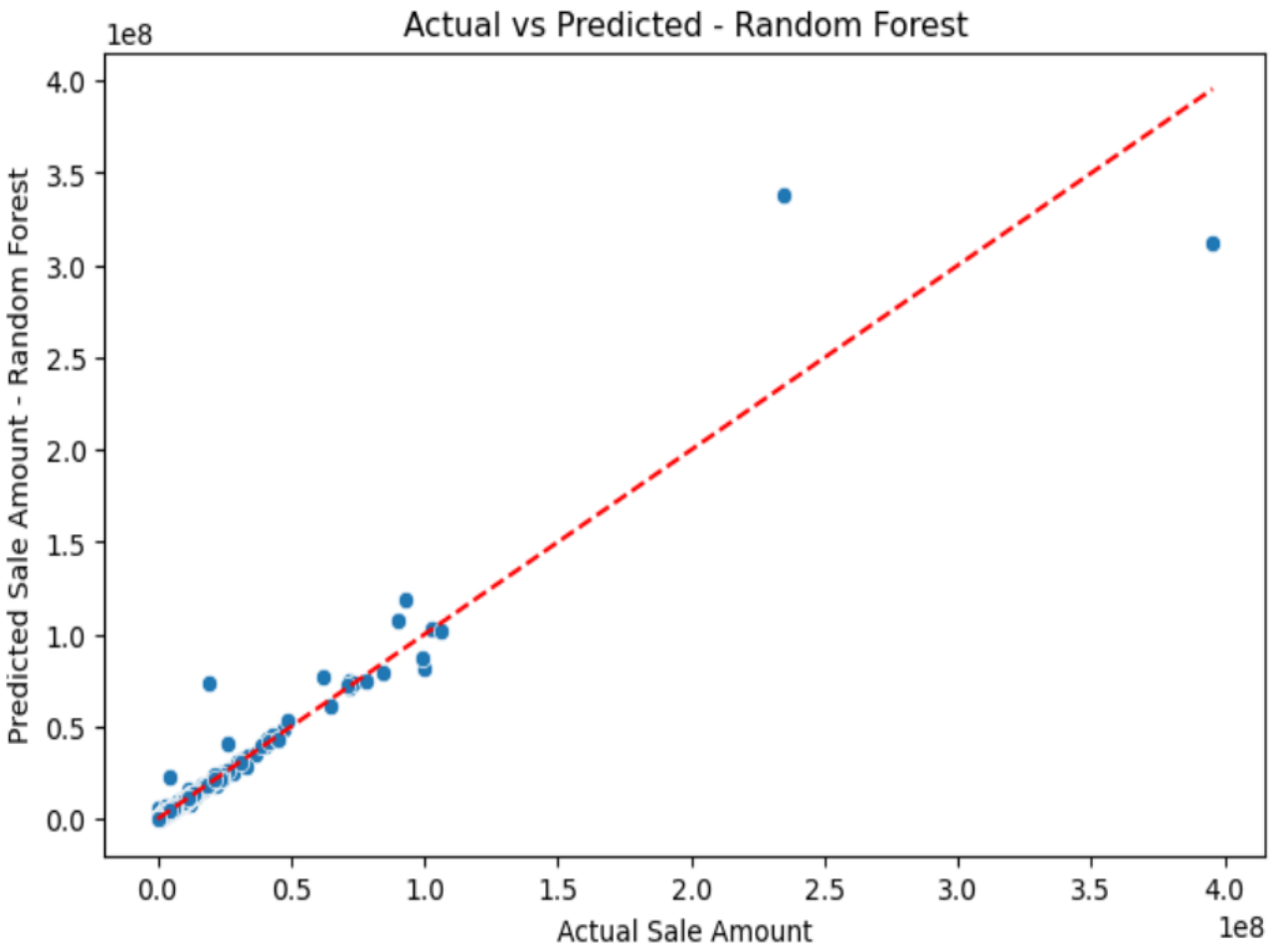

Fig.5. Random Forest Actual vs Predictions of sales amount

6.6 XGBoost Actual vs Predicted:

Many points appear spread throughout the area surrounding the linear line. The model produces lower than actual values when estimating property sale amounts particularly in the case of high-value properties. The wide distribution indicates that XGBoost displays poor performance in connecting features with sale amounts until additional adjustments or data pre-treatment occurs.

At present XGBoost delivers inferior predictive results and generates imprecise property value assessments.

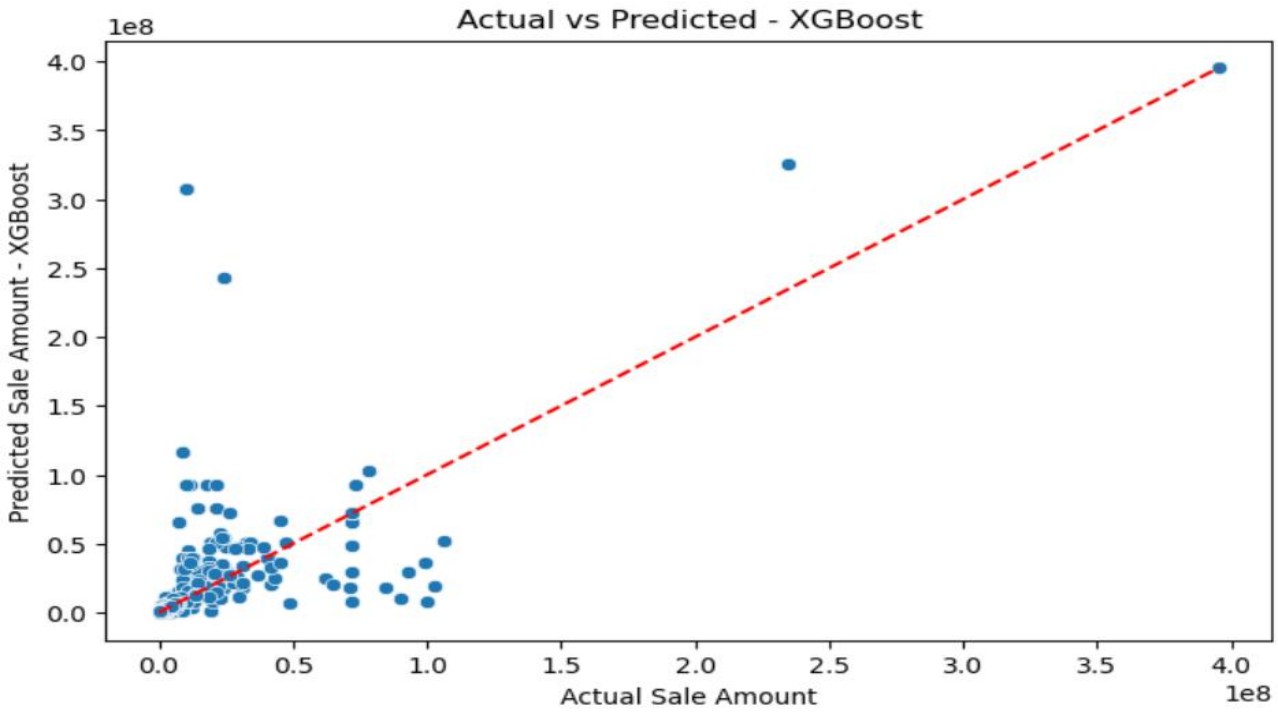

Fig.6. XG Boost Actual vs predictions of sales amount

Random Forest Regressor delivered superior performance than XGBoost with substantially superior R^2 value along with lower RMSE score. The Random Forest model successfully identified all nonlinear functions present between socioeconomic variables and property transaction costs. In order to boost performance in this context the XGBoost model might need better hyperparameter tuning together with additional features.

7. Feature Importance

To understand which variables most impacted the expected sale quantities, we assessed feature importances from the trained Random Forest models in two configurations:

A. Crime-Only Feature Importance

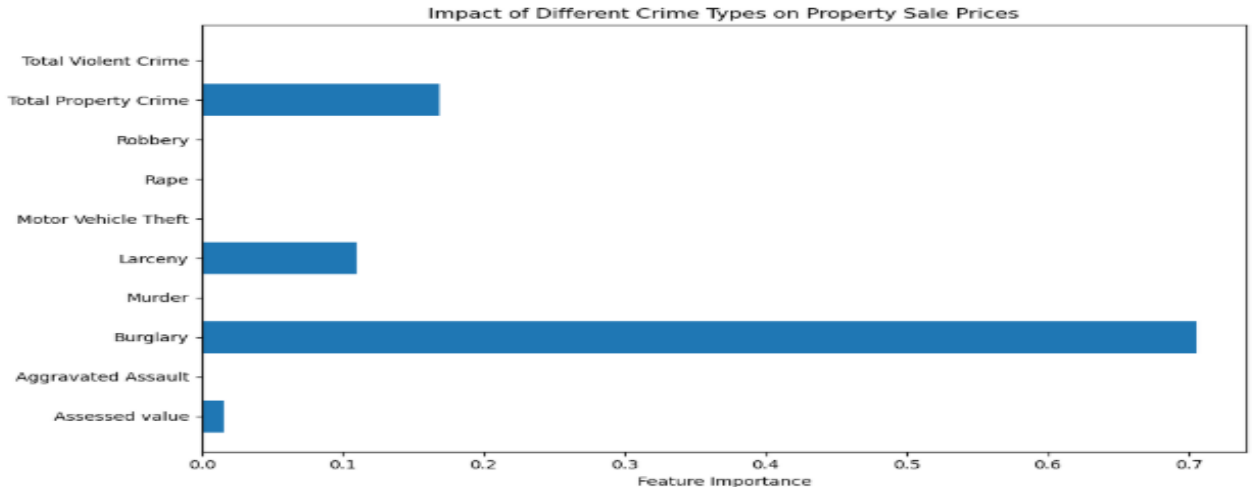

Fig.7. Impact of Different Crime Types on Property Sale Prices

We trained the Random Forest Regressor on solely crime-related variables such burglary, larceny, robbery, and overall crime. Burglary appeared as the most powerful predictor, performing greatly in all crime types. Total Property Crime and Larceny were the next most significant contributions.

Violent crimes such as aggravated assault, rape, and murder had significantly lower significance rankings. Property-related crime may have a higher influence on property value due to buyer considerations and insurance concerns, compared to violent crime.

B. Crime + Socioeconomic Feature Importance

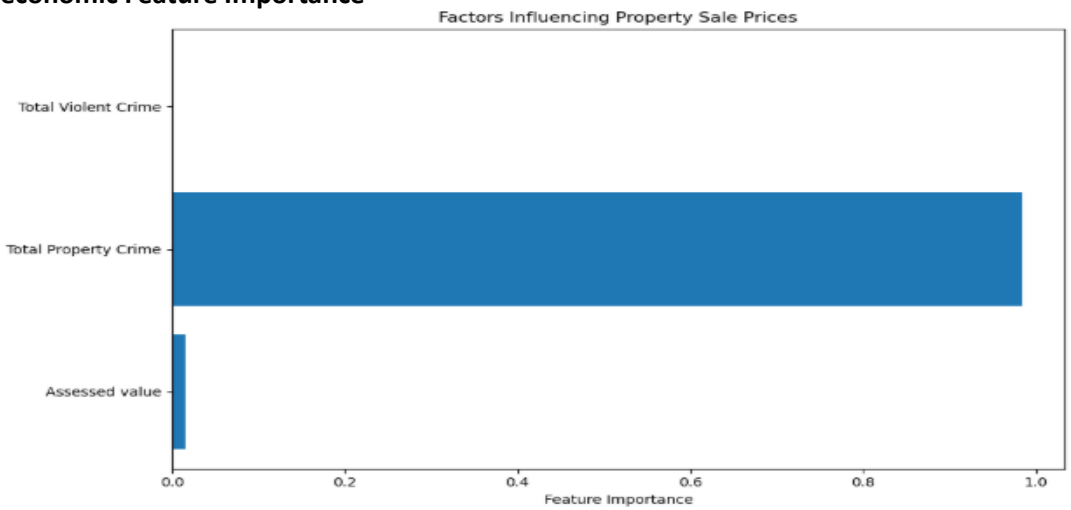

Fig. 8. Factors Influencing Property Sale Prices

This model consists of incorporated criminal factors as well as one essential economic feature that is property's assessed value.

Total Property Crime had the greatest relevance (~0.95), even surpassing assessed value.

Total Violent Crime was also relevant, but to a lesser amount.

Surprisingly, Assessed Value had little influence on the final prediction

Property crime rates may have high impact on buyer behavior than typical valuation indicators, particularly in high-crime or high-variance locales.

8. Research Questions and Findings

The purpose of this study was to use predictive machine learning models and feature analysis to evaluate how different forms of crime data impact real estate selling prices in Connecticut communities.

8.1 Research Question 1. How do assessed value and sale ratio influence property sale prices across towns?

This question examines the link between geographic financial variables and selling prices.

Model	R2	RMSE
Random Forest	**0.9543**	**588,833.05**
XGBoost	0.6889	1,537,147.15

The Random Forest model showed high prediction power, accounting for almost all the variance in real estate sale prices given assessed value and sale ration. This shows the utter effectiveness to capture intricate relations in the data. In comparison, the XGBoost model also did reasonably well although it displayed slightly higher residual errors suggesting a slightly less refined fit compared to the Random Forest.

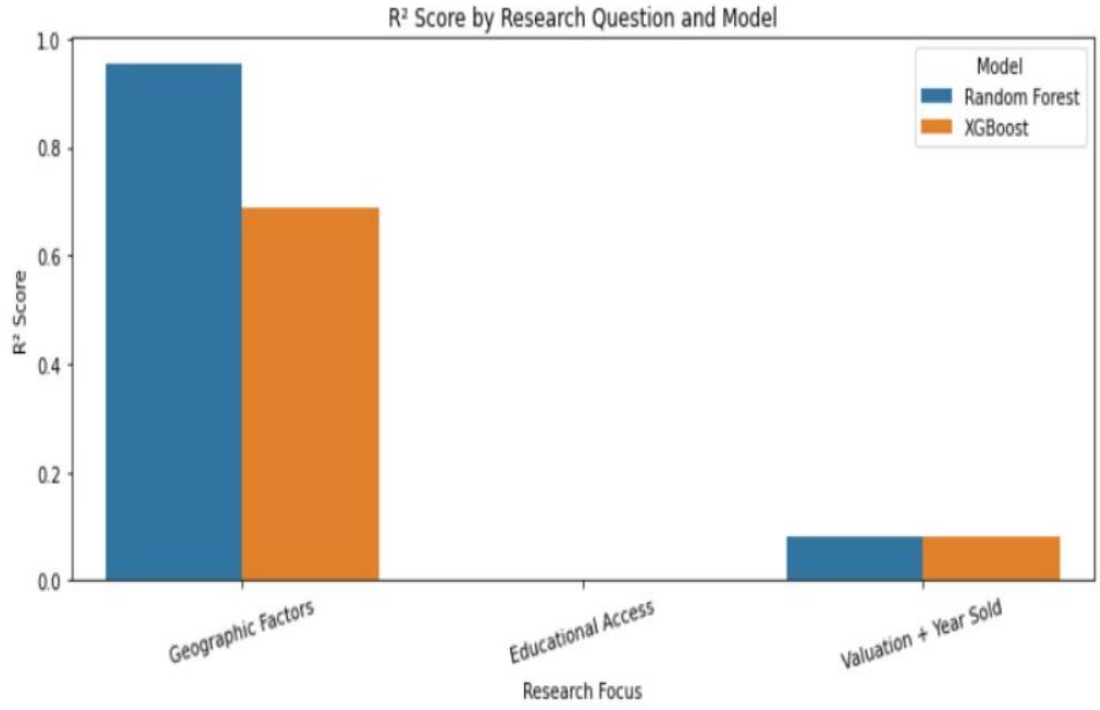

Fig.9. R2 Score vs Research Focus

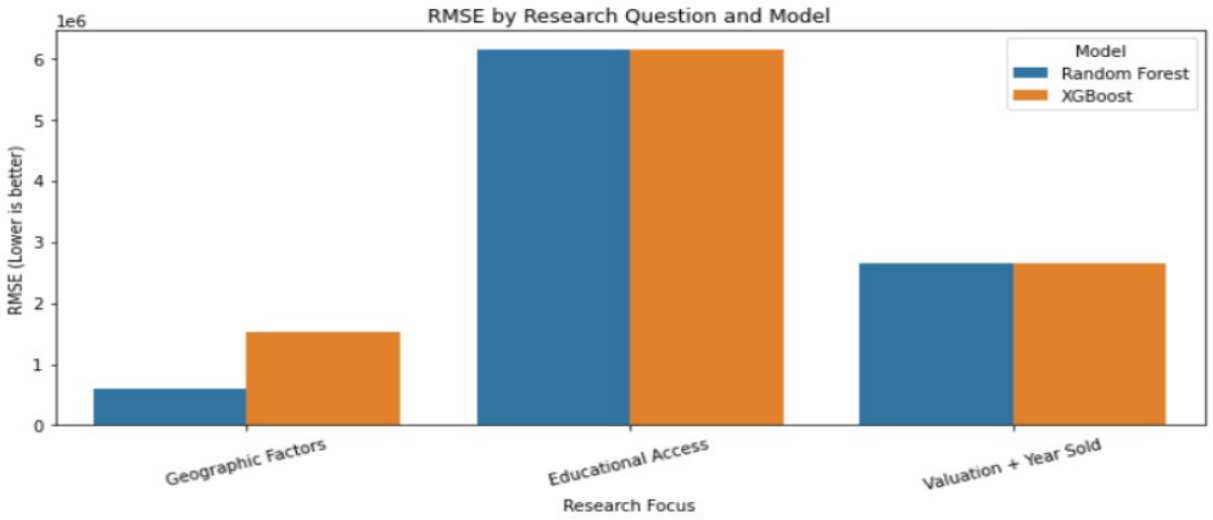

Fig.10. RMS vs Research Focus

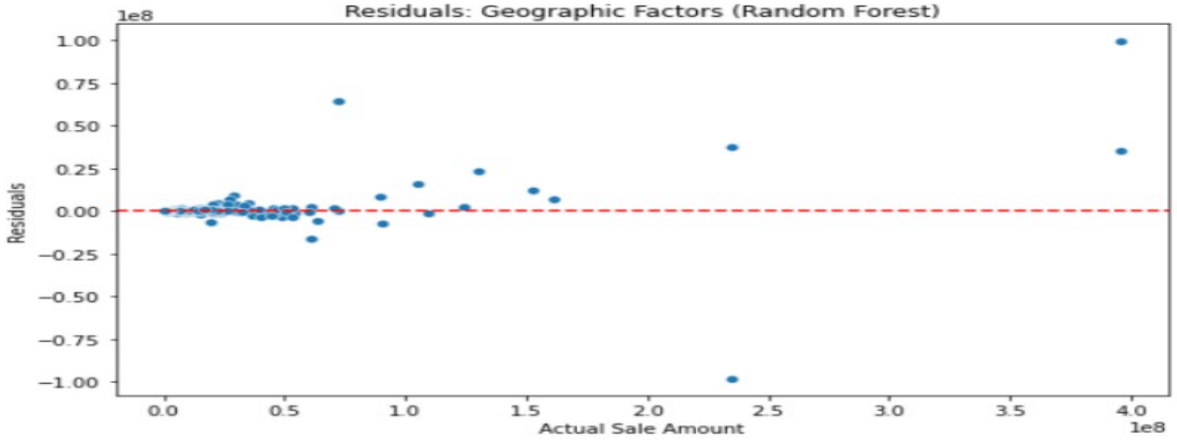

Fig 11: Actual Sale Amount vs Residuals(Random Forest)

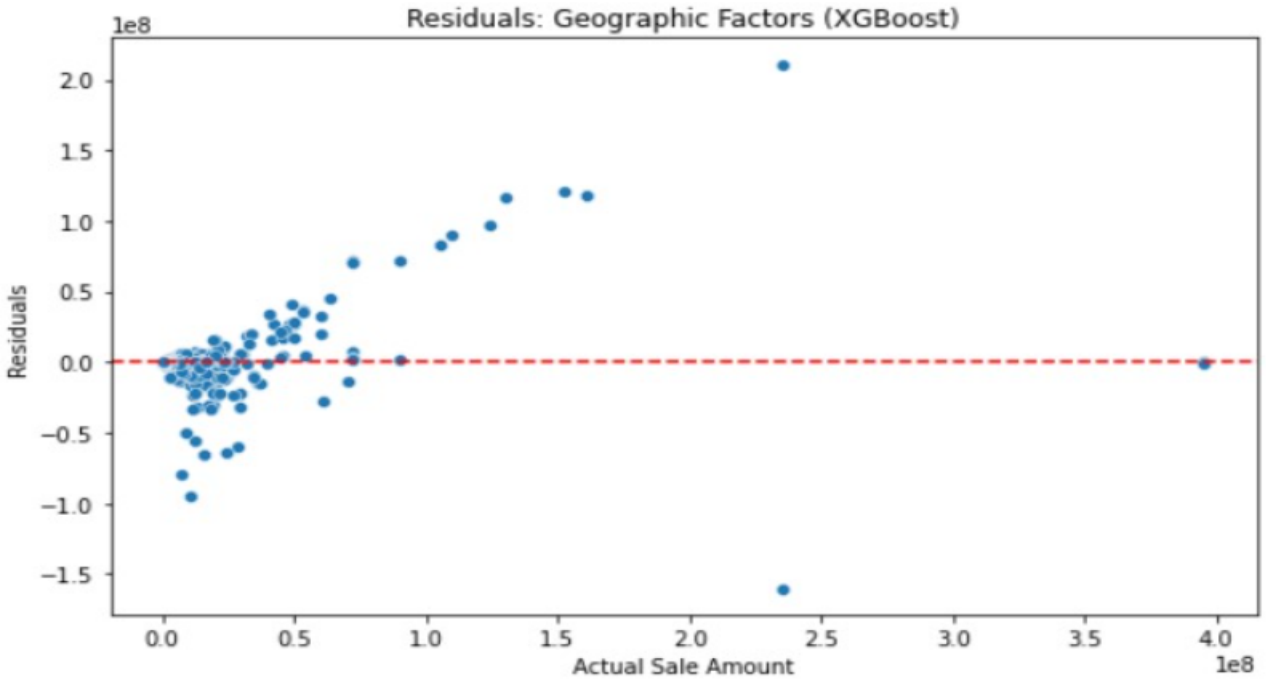

Fig. 12. Actual Sale Amount vs Residuals (XGBoost)

8.2 Research Question 2: How does educational access (PreK–12 school availability) affect sale prices?

This analysis measures the effect of the number of institutions of learning on the demands of housing or pricing.

Model	R2	RMSE ($)
Random Forest	0.0034	6,142,243.34
XGBoost	0.0034	6,142,310.58

Both models had scores close to zero R 2, implying that school counts are not predictive of the property value. High values of RMSE provide additional evidence of the lack of independence.

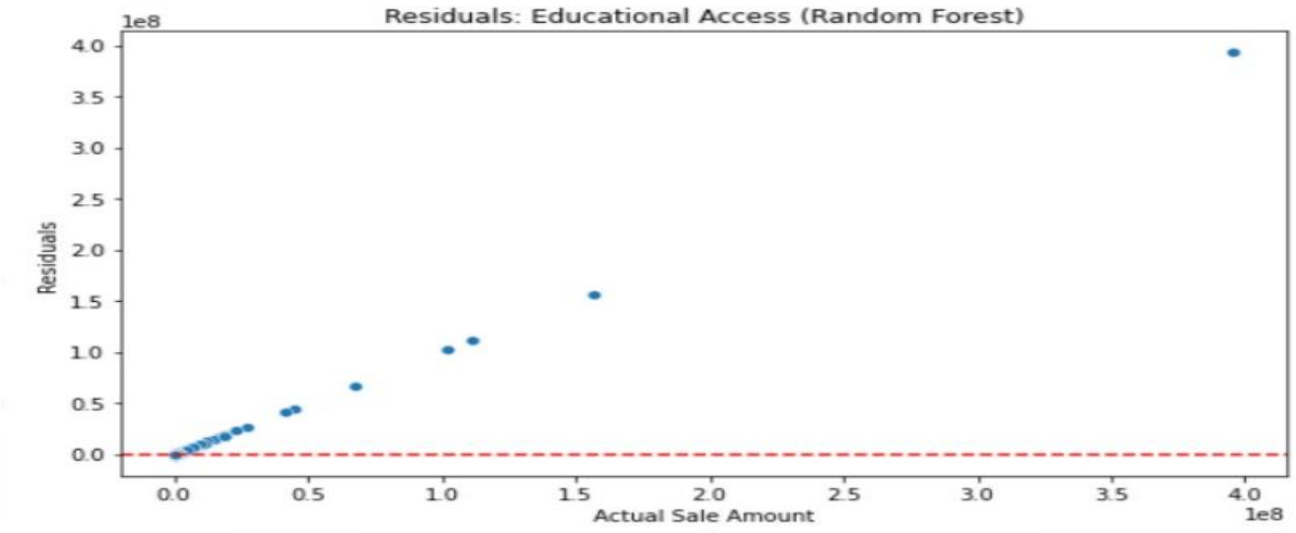

Fig. 13. Actual Sale Amount vs Residuals (Random Forest)

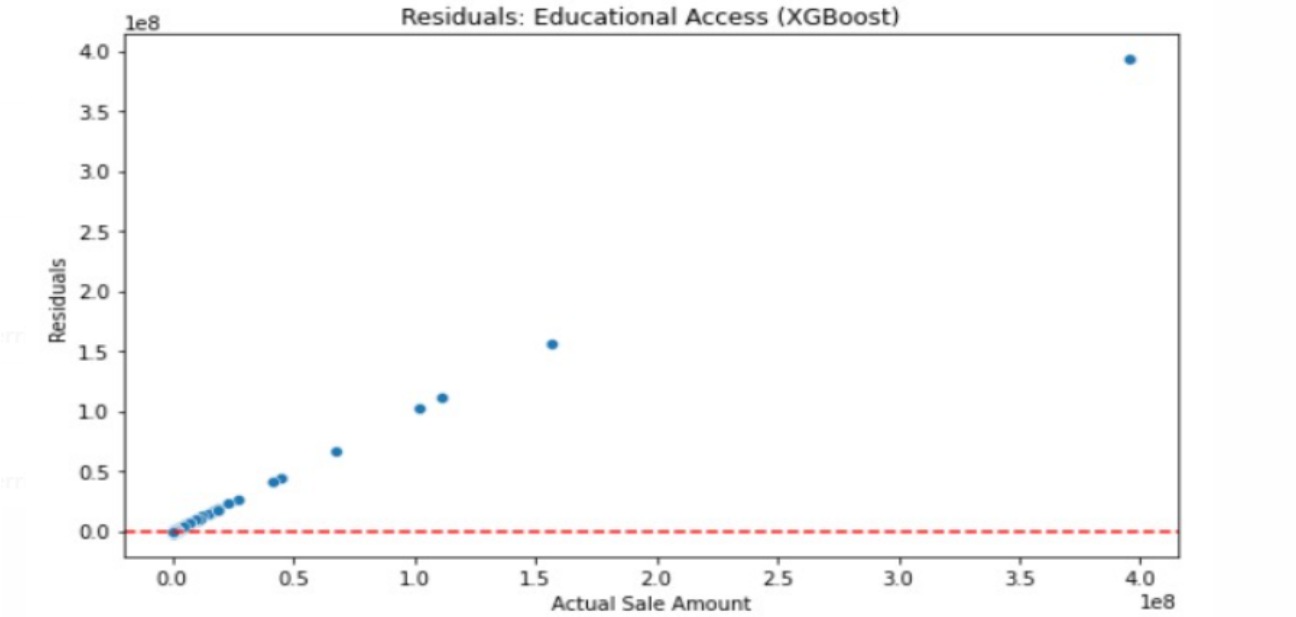

Fig. 14. Actual Sale Amount vs Residuals (XGBoost)

8.3 Research Question 3: How do assessed value and year of sale influence property sale prices?

The explained variance was weak by the year variable, and the models accounted for only 8% of the variance.

Relatively high RMSE values indicate that the assessed value already contains a lot of the property's market trajectory.

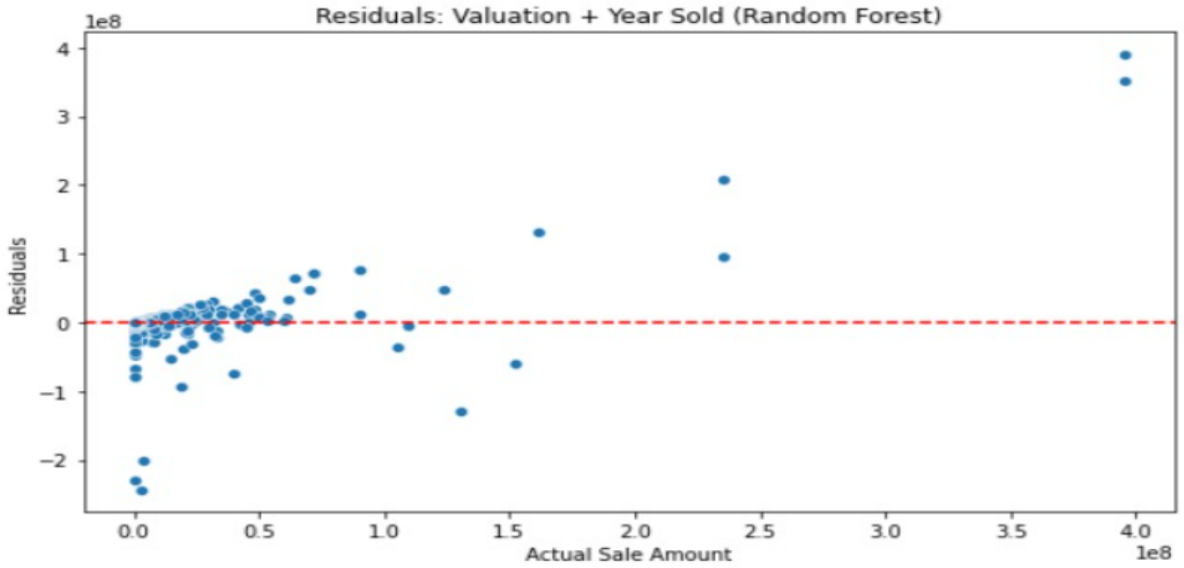

Fig. 15. Actual Sale Amount vs Residuals (Random Forest)

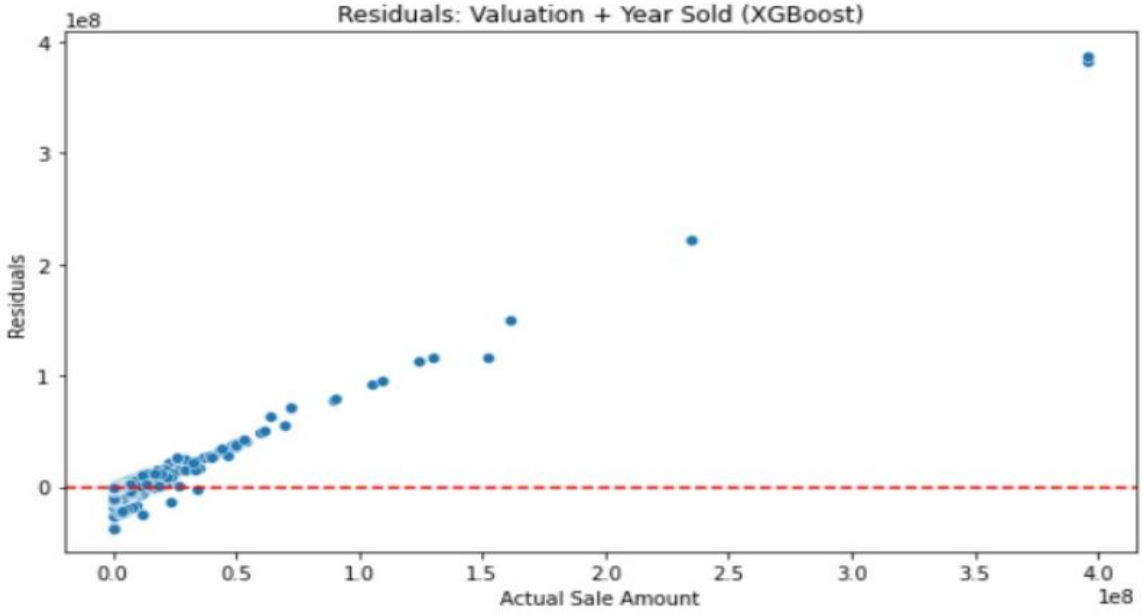

Fig.16. Actual Sale Amount vs Residuals (XGBoost)

9. Preliminary Results

9.1 Crime Rates vs Property Value

```
Year: 5.1643
Month: 8701.11
Aggravated Assault: 1311.17
Arson: -597.733
Burglary: -731.446
Larceny: -183.748
Motor Vehicle Theft: -502.752
Murder: -5815.39
Rape: -3241.8
Robbery: 1791.2
Total Crime: -173.435
Total Property Crime: 338.065
Total Violent Crime: -750.593
Condo_OHE_0: 29389.8
Four Family_OHE_1: 77791.8
Single Family_OHE_2: 93814.5
Three Family_OHE_3: 68866.1
Two Family_OHE_4: 16527.7
```

Fig. 17. Preliminary Results

The model's preliminary results include coefficients next to each crime/property type. The typical findings confirmed our hypotheses, but there were a few outliers that the team will need to conduct exploratory data analysis on to delve deeper. For example, most of the crime in the model had a negative output, but aggravated assault and total property crime produced a positive coefficient.

9.2 Condo vs. Single-Family Housing Trends

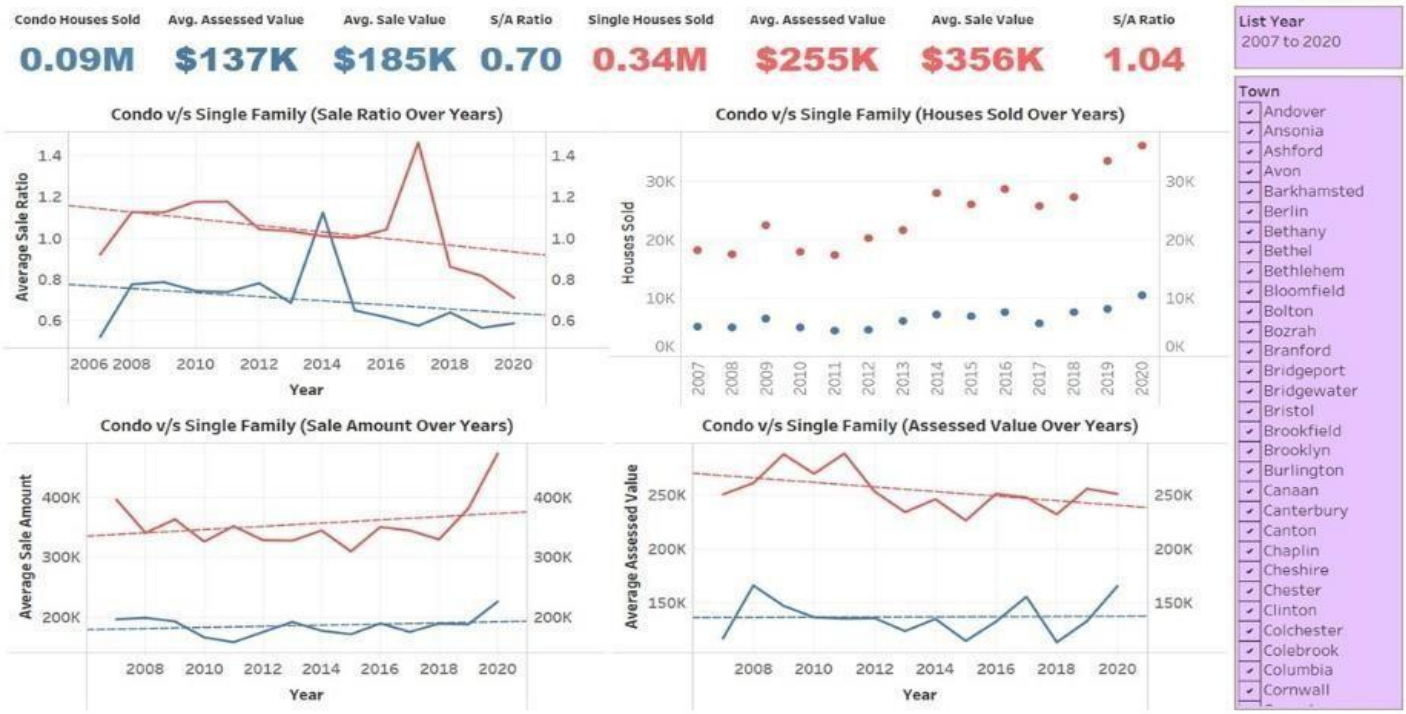

Fig. 18. Condo vs. Single-Family Housing Trends

This visualization illustrates a comparison between condos and single-family homes' empirical research. The data shows that single-family homes regularly have higher assessed and selling values than condominiums, as seen in the top bar measures. The four graphs also show trends in sale ratios, houses sold, sale amounts, and assessed values, demonstrating that single-family homes dominate the market with greater sales and prices. The condo market appears to be relatively stable but with fluctuations in sale ratios and assessed values.

9.3 Impact of Employment Factors on Assessed Property Values

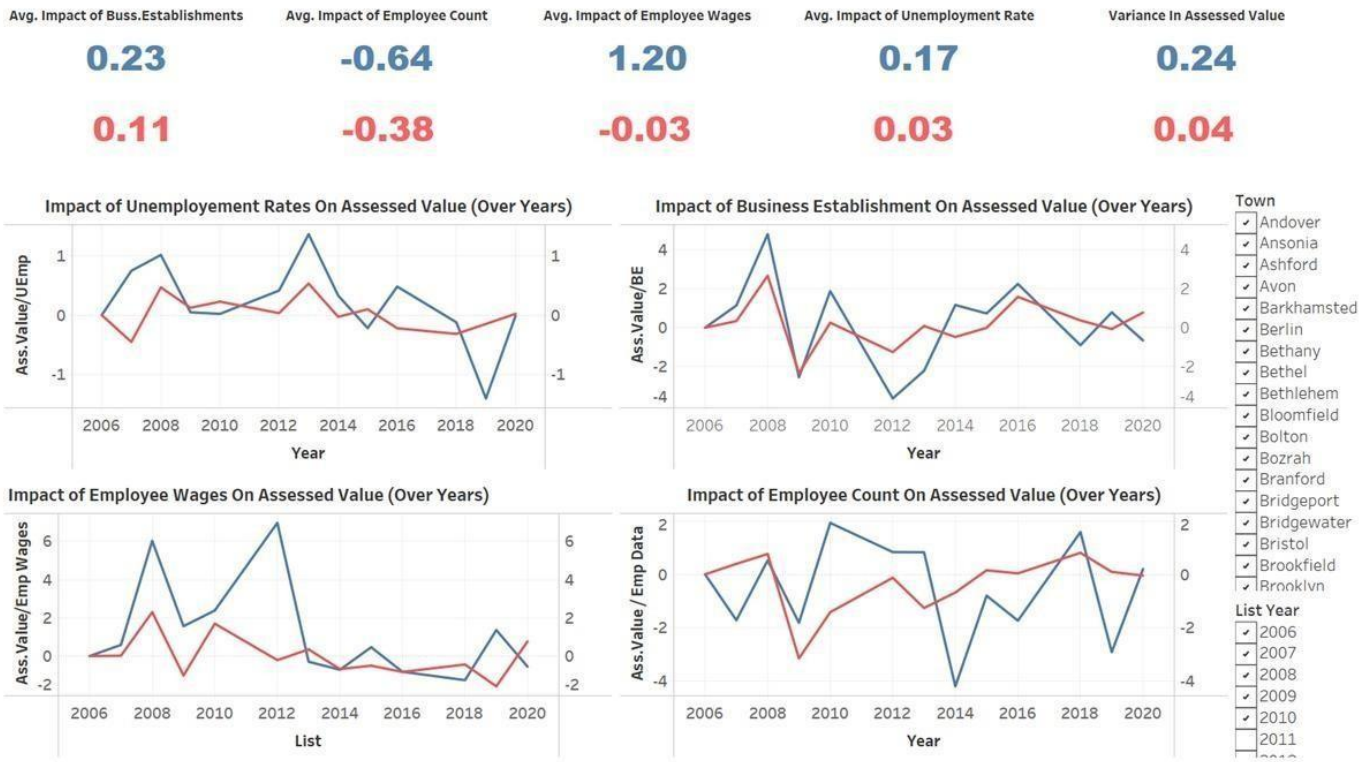

Fig. 19. Fluctuations in Assessed Property Values

This visualization examines how different economic factors influenced assessed property values over time, considering company premises, employee count, earnings, and unemployment rates. The top part provides numerical summaries of the average impact of these factors, with positive and negative values indicating their respective influence. The four graphs illustrate trends in the relationship between assessed values and key economic indicators, showing fluctuations over time. The data suggests that employee wages have a significant positive impact, while employee count has a negative correlation with assessed values.

9.4 Impact of Employment on Sale Amount

Fig. 20. Influence of Economic Indicators on Real Estate Sale Amounts

This picture depicts a data visualization analysis of the influence of socioeconomic variables on real estate sales volumes from 2006 to 2020. The top section displays crucial variables, demonstrating that company premises and employee pay have a positive correlation with sales, whilst unemployment rates have a negative influence. The charts illustrate how these factors fluctuate over time across different towns. Notably, places with more businesses and higher wages tend to see increased property values, whereas higher unemployment corresponds with lower selling amounts. The variation in sale amounts suggests market instability in specific regions.

9.5 Impact of Employment on Houses Sold

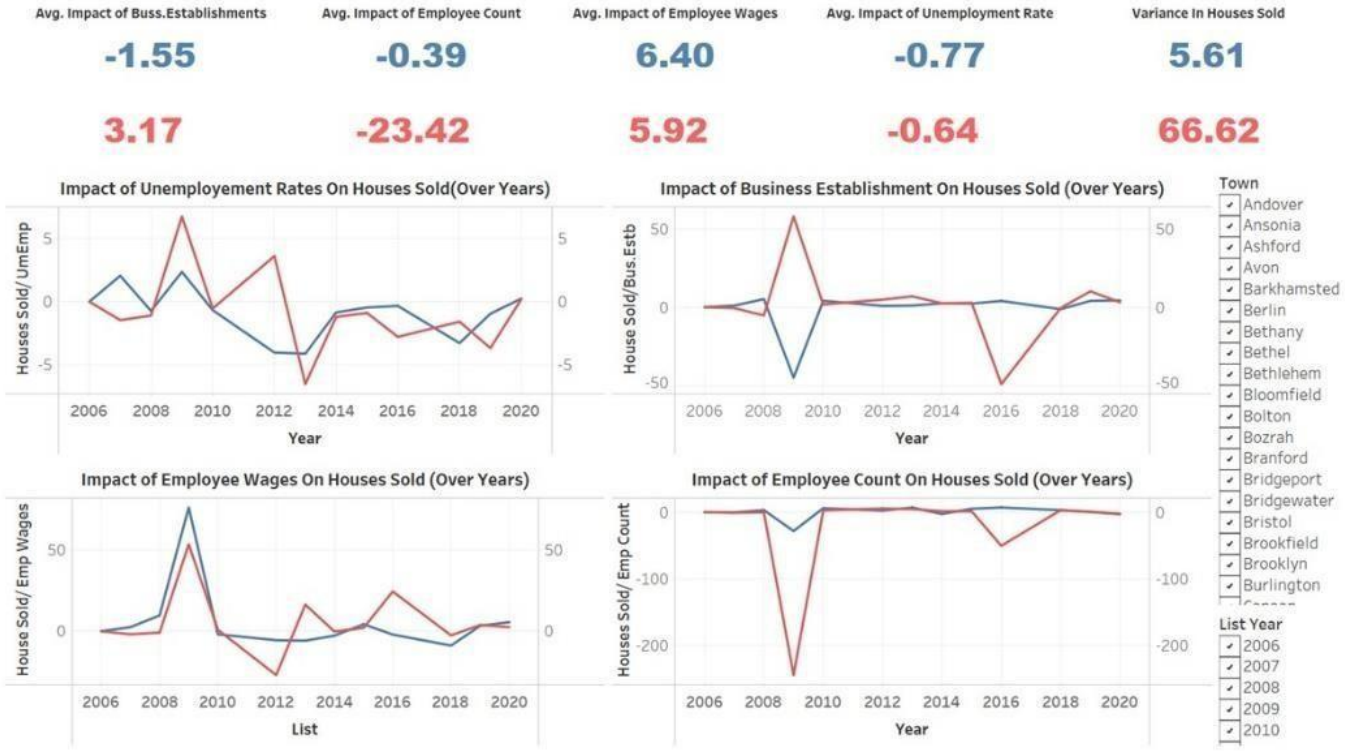

Fig. 21. Impact of Economic Indicators on Number of Houses Sold

This image showcases an analysis of how various socioeconomic factors have influenced the number of houses sold between 2006 and 2020. The top section indicates that company establishments and employee earnings have a positive correlation with house sales, while unemployment rates and employee count have a negative impact. The graphs illustrate trends over time across different towns, showing fluctuations in home sales as these socioeconomic factors shift. Notably, a higher number of businesses and rising wages tend to drive home sales upward, while unemployment appears to have a dampening effect. Employee wages appear to have a strong positive impact, while employee count and business establishments show mixed effects. The variation in houses sold indicates severe volatility in some areas, emphasizing the influence of socioeconomic conditions on the real estate market.

9.6 Public Infrastructure and Crime Trends (2008–2018)

Healthcare facilities and schools decreased in numbers by 288 and 593 respectively throughout Connecticut towns during the 2008 to 2018 period. Total crime numbers showed variation, yet the crime rate dropped from above 400 to about 350 per 100,000 people which indicates rising public safety levels which could potentially sway housing preferences.

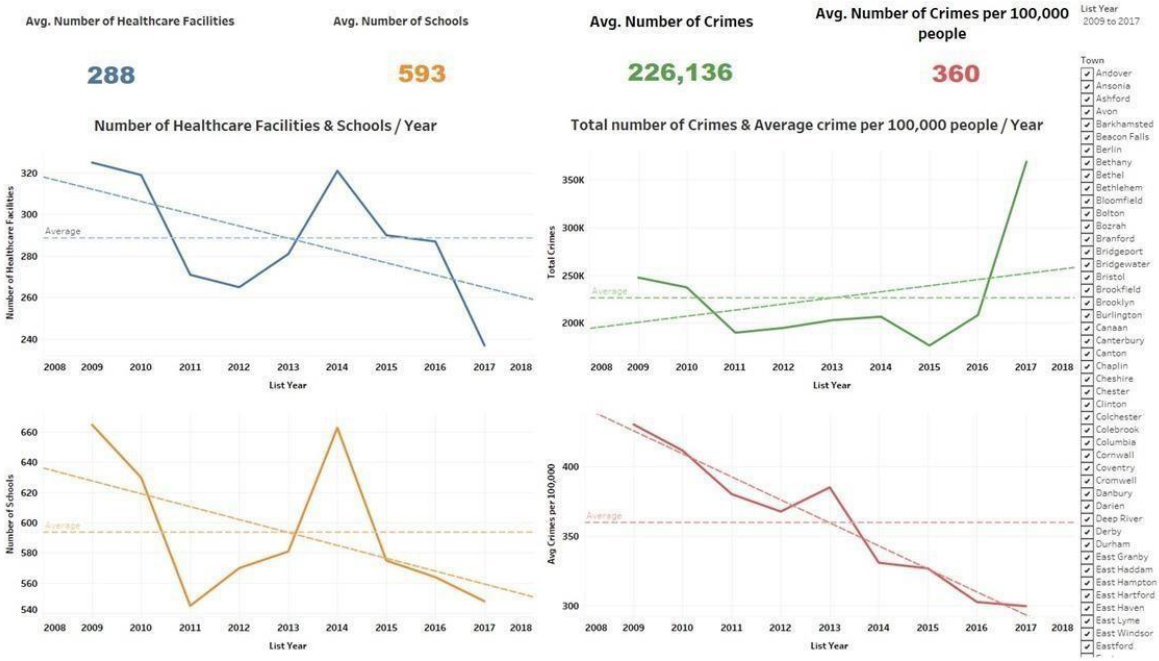

Fig. 22. Trends in Healthcare, Schools, and Crime

9.7 Housing Prices and Healthcare Access

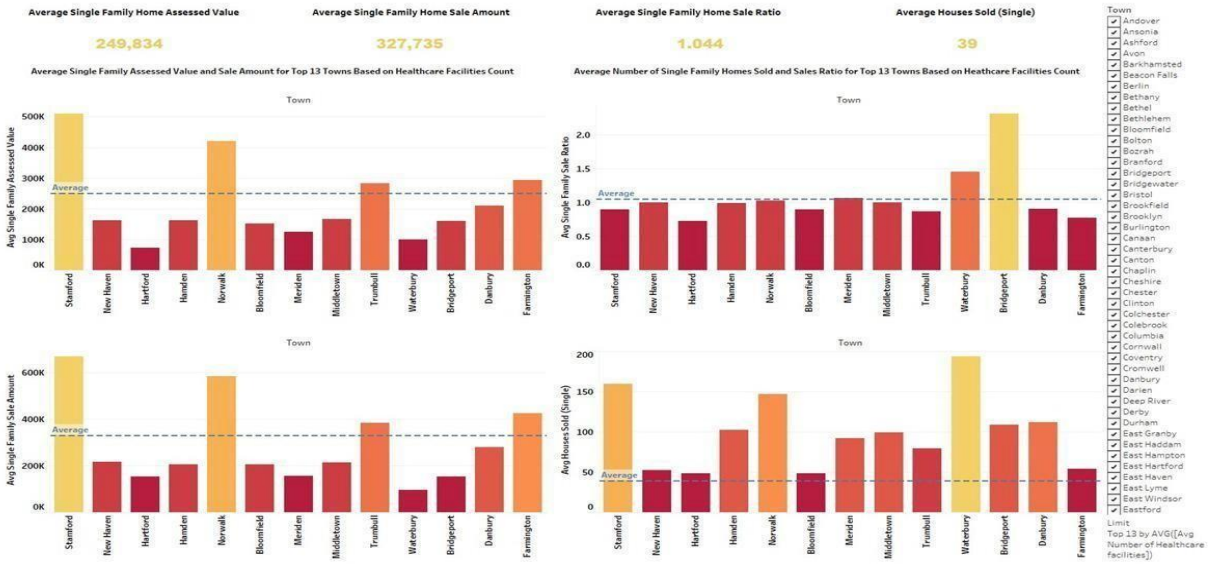

Fig. 23. Single-Family Home Prices by Healthcare Facility Count

The research evaluates healthcare facility totals against single-family home values across the top 13 towns with maximum facility presence. Both Stamford and Norwalk routinely maintained higher single-family home assessed and sale values despite Norwalk and Waterbury and Bridgeport leading the sales unit counts. The data analyzed reveals two different effects on real estate by healthcare accessibility and affordability. Access benefits housing values in certain areas yet different factors drive increased housing market transactions in other regions.

9.8. Housing Prices and School Access

Fig. 24: Condo Prices by Healthcare Facility Count

Figure 8, shifts focus to condominiums. The assessed prices alongside purchase amounts from condo properties remained below those of single-family homes. Stamford kept its position as a top-performing town even within the condo market. The City of Waterbury demonstrated exceptional condo sale volumes because it offered an affordable housing choice that maintained essential services including healthcare services for its residents.

9.9. Regional Economic Indicators and Housing Activity

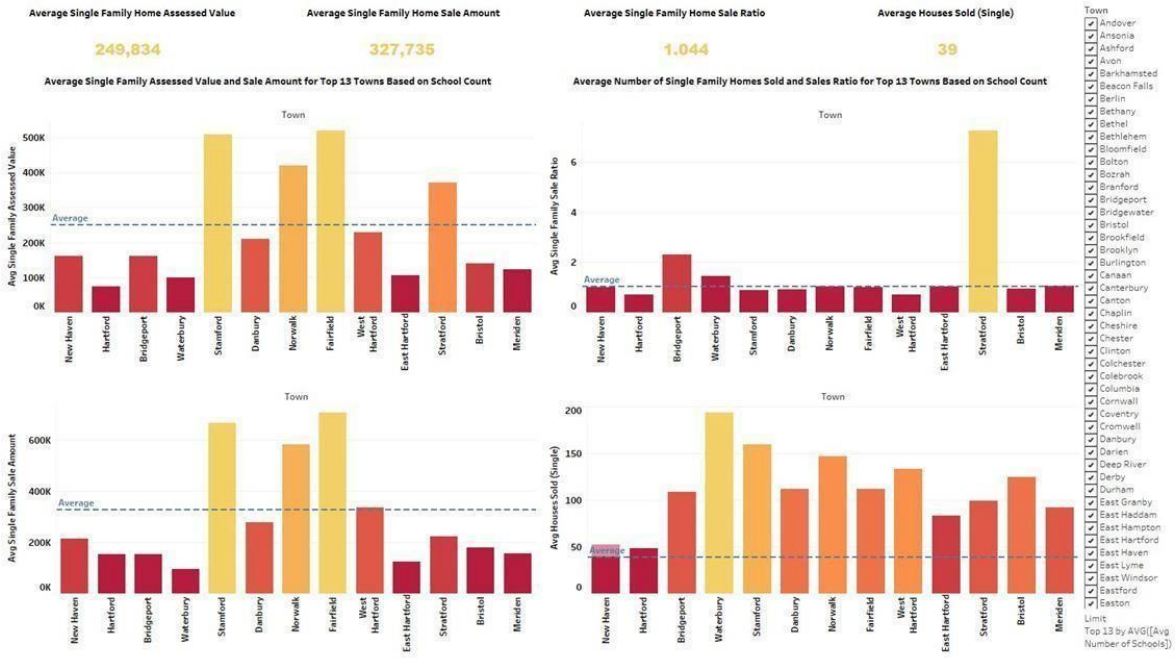

Fig. 25: Single-Family Home Prices by School Count

Single-family housing values demonstrate patterns with respect to the number of available schools as Figure 9 suggests. The presence of additional schools in Fairfield, Stamford and Norwalk positively influenced home prices through higher sales and assessed values. The town of Stratford demonstrated the greatest home sales despite its high number of homes which implies that top-performing districts push property values upwards, but education-focused communities achieve solid market demand through cost effective opportunities for students.

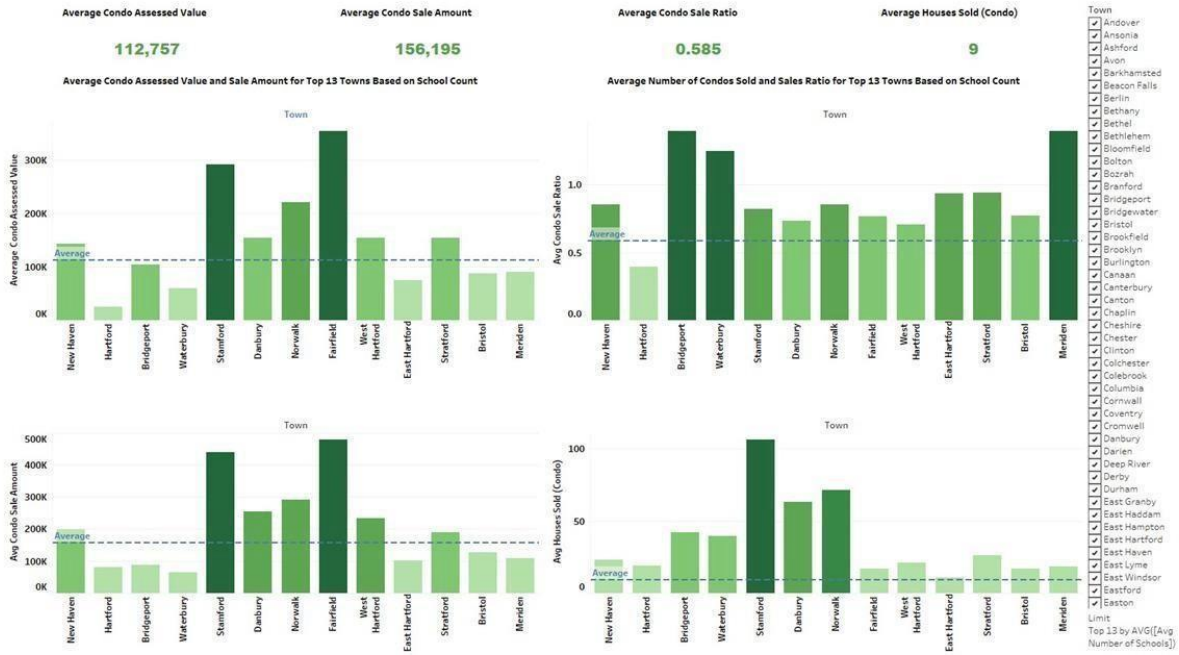

Fig. 26: Condo Prices by School Count

Figure 10 presents an assessment of the relationship between real estate prices and school districts for condos. The estimated values of condos and number of sales in Fairfield and Stamford outperformed those recorded in the other locations. Property transactions for condos were more prevalent in the towns of Bridgeport and Waterbury. The repeating pattern confirms that educational facilities increase home prices, yet affordability continues to drive urban real estate markets.

9.10. Regional Economic Indicators and Housing Activity

Analysis shows how average employee wages relate to business establishment count as well as unemployment levels across geographic locations. Higher employment rates and denser business activities along with more steady economic indicators marked southwestern Connecticut towns particularly Stamford Norwalk and Fairfield. The towns demonstrated superior results in housing metrics because economic strength creates both rising property values and constant market demand for real estate.

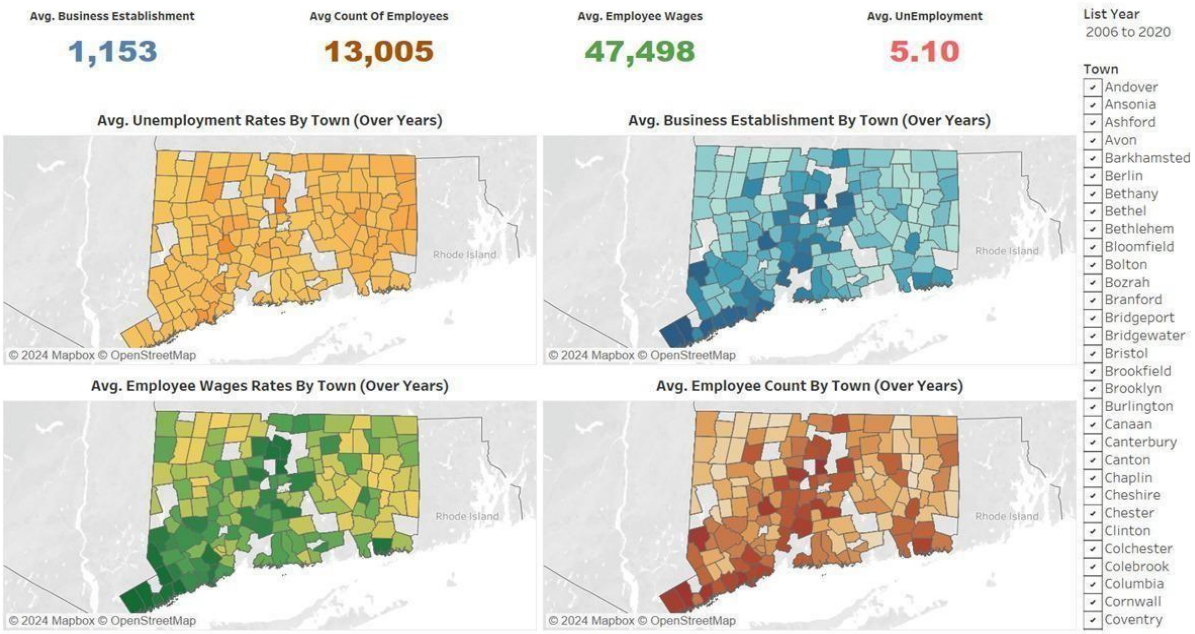

Fig. 27: Economic Indicators by Town (2006–2020)

9.11. Crime Impact on Real Estate Trends

Public property value data reveals the causal link depicted in Figure 12. Areas with elevated crime frequencies experienced simultaneous reductions in actual property sale prices alongside evaluated residential property values. Property valuation suffered its most serious suppressive effects from crime throughout the 2014–2016 period. Public safety issues produce substantial negative impacts on real estate market operations in mid-density urban areas.

The analysis in Figure 13 shows how criminal activities affect the movement of transactions. The reduction in crime rates starting from 2013 correlated with improved numbers for home sales ratio and volume. The improvement of public safety creates more favorable conditions for buyers which translates into better real estate transaction activity.

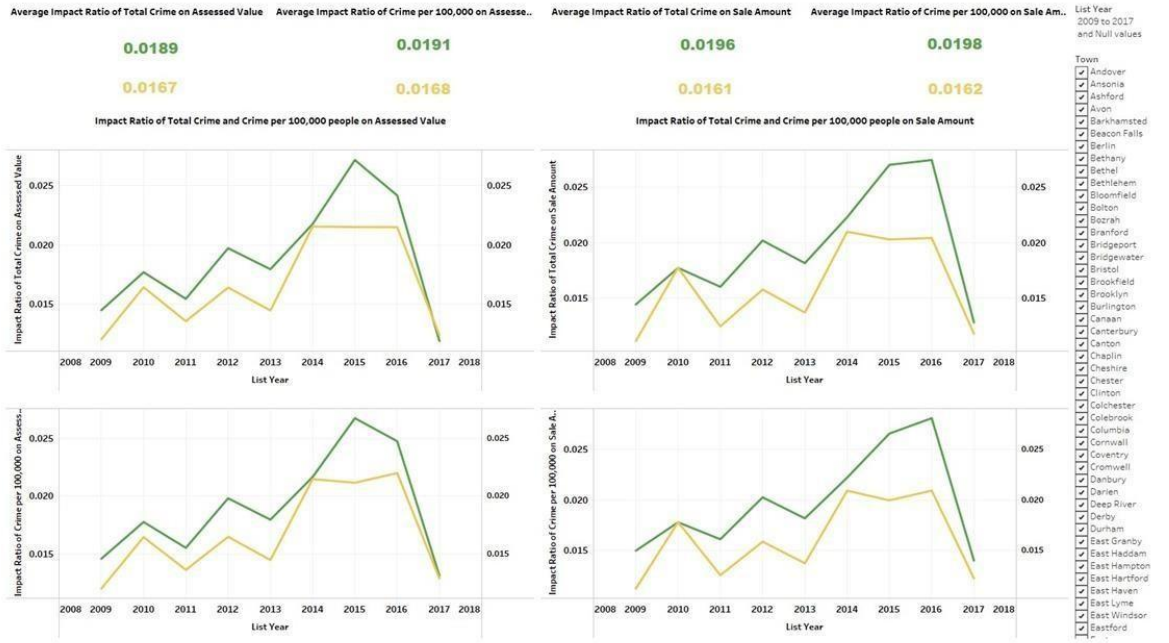

Fig. 28: Crime Impact on Assessed Value and Sale Amount

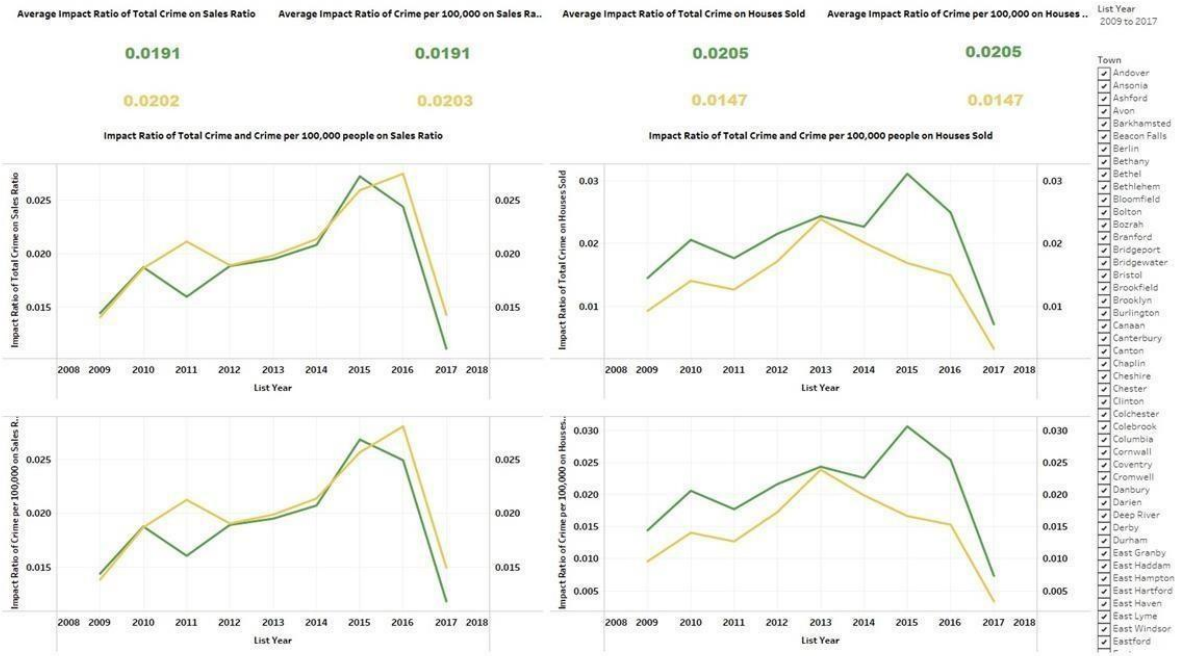

Fig.29: Crime Impact on Sales Ratio and Homes Sold

Conclusion

This study examines Real estate sale prices in Connecticut between 2001 and 2022 are analyzed with respect to socioeconomic indicators in this study. This study used a combination of crime rate data with measures of school accessibility and healthcare clinics and economic metrics to find clear properties value trends.

A positive relationship existed between school quality and local employment opportunities and housing market strength. The combination of elevated crime rates and unemployment creates negative impacts on local market value and real estate sales performance. The communities of Stamford and Fairfield established themselves as top areas with valuable properties mainly because of their reliable public services and steady economic climate.

The project depended on regression analysis in combination with visualizations and mapping software to analyze the data before we present the final interpretation to readers. Such data point correlations may be studied to help homebuyers, realtors, and government officials make better decisions.

This analysis shows that real estate goes beyond geographical location because it encompasses elements like community life, protective standards and scholarly institutions and economic strength. Our analysis connects the various influencing elements to create better insights about property value determinants in modern market conditions.

Timeline:

Project Timeline	
Task Name	**Completion NLT**
Project Proposal	February 10th, 2025
Project Milestone 1	March 2nd, 2025
Data prep and start	March 5th, 2025
Test regression and other analytical methods	March 12th, 2025
Code Review	March 25th, 2025
Implement regression methodology	April 1st, 2025
Analysis and trend comparison	April 5th, 2025
Review Result	April 6th, 2025
Project Milestone 2	April 14th, 2025
PPT slides	April 20th, 2025
Report	April 28th, 2025
Final Project Presentation	April 28th, 2025

References

[1] Feiveson, L. (2024, June 24). Rent, house prices, and demographics. *U.S. Department of the Treasury.* https://home.treasury.gov/news/featured-stories/rent-house-prices-and-demographics

[2] Glaeser, E. L., & Gyourko, J. (2008). The impact of building restrictions on housing affordability. *Economic Policy Review, 9*(2), 21-39. https://www.newyorkfed.org/medialibrary/media/research/epr/03v09n2/0306glae.pdf

[3] Gyourko, J., & Molloy, R. (2015). Regulation and housing supply. *Handbook of Regional and Urban Economics, 5*, 1289-1337. https://doi.org/10.1016/b978-0-444-59531-7.00019-3

[4] Levitt, S. D., & Dubner, S. J. (2005). *Freakonomics: A rogue economist explores the hidden side of everything.* HarperCollins.

[5] Office of Policy and Management. (2018). Real estate sales 2001-2022 GL. *Connecticut Data Portal.* https://data.ct.gov/Housing-and-Development/Real-Estate-Sales-2001-2022-GL/5mzw-sjtu/about_data

[6] Zhang, J., & Deng, X. (2022). Real estate tax, housing price, and housing wealth effect: An empirical research on China housing market. *Discrete Dynamics in Nature and Society, 2022*, 1-9. https://doi.org/10.1155/2022/4809499

[7] Jin, S., Zheng, H., Marantz, N., & Roy, A. (2024). Understanding the effects of socioeconomic factors on housing price appreciation using explainable AI. *Applied Geography*, *169*, 103339. https://doi.org/10.1016/j.apgeog.2024.103339

[8] *SBond. (2024, November 11). Real estate sales by town Connecticut 1999 to 2023 [Data visualization]. RPubs. https://rpubs.com/SBond/1244153*