Analysis of Residential Real Esatate Data From Connecticut*

Siling Guo, Megan Joseph, John Sawicz

November 18, 2025

1 Introduction

Real estate trends change over the years. The most notable changes have been during the 2008 economic crisis and COVID-19. We decided to investigate real estate in Connecticut, USA to see how it's changed over time and how it differs across cities.

We first clean the data and calculate descriptive statistics. Then, we make graphs of the categorical variables and continuous variables based on the cities we picked. In our advanced analysis, we make heat maps showing the densities of sale price and a multiple linear regression model with sale price as the response and assessed value and town as the predictors.

2 Data

The data we used for this analysis is Real Estate Sales data from 2001-2023 from the State of Connecticut's Office of Policy and Management. The sale price of each property is at least $2,000. Each row is a property which contains information of the town, address, date sold, property type (residential, apartment, commercial, industrial or vacant land), sale price, assessed value, and latitude and longitude coordinates. For the purposes of this analysis, we mainly focus on residential properties and the columns town, property type, sale price, assessed value, and coordinates. Additionally, we picked four cities to focus on: Stamford, Westport, Cheshire, and Sprague. This was done because of the large number of data points and to investigate any differences in towns with varying levels of median income. Westport, CT has the highest median income at $250,001, then we picked Cheshire, CT at $150,787 for upper middle, Stamford, CT for lower middle, and Sprague, CT for the lowest (2025, n.d.). We also wanted to sample towns with different populations and densities.

*Project repository available at: https://github.com/meganajoseph/167r_project.

2.1 Data Cleaning

Many of the data points were missing or were empty characters, so we dropped those rows. Any duplicate rows were also dropped. Since our goal is centered on residential properties, we filtered out rows that were not residential. Additionally, the Sales.Ratio column needed to be transformed into a numeric value.

2.2 Descriptive Statistics

We analyzed the mean, minimum value, maximum value, first quantile, median, and third quantile of the Sale.Amount, Assessed.Value, and Sales.Ratio columns. The information is summarized in the table below.

Column	Mean	Minimum	Maximum	1st Quartile	Median	3rd Quartile
Sale Amount	7.7×10^5	2160	7.2×10^7	3.1×10^5	5.04×10^5	8.2×10^5
Assessed Value	4.6×10^5	0	2.5×10^7	2×10^5	3.3×10^5	5.3×10^5
Sales Ratio	0.8	0	291.94	0.55	0.64	0.76

3 Graphs

3.1 Categorical

We created bar plots to see the number of entries per town of interest and per residential type.

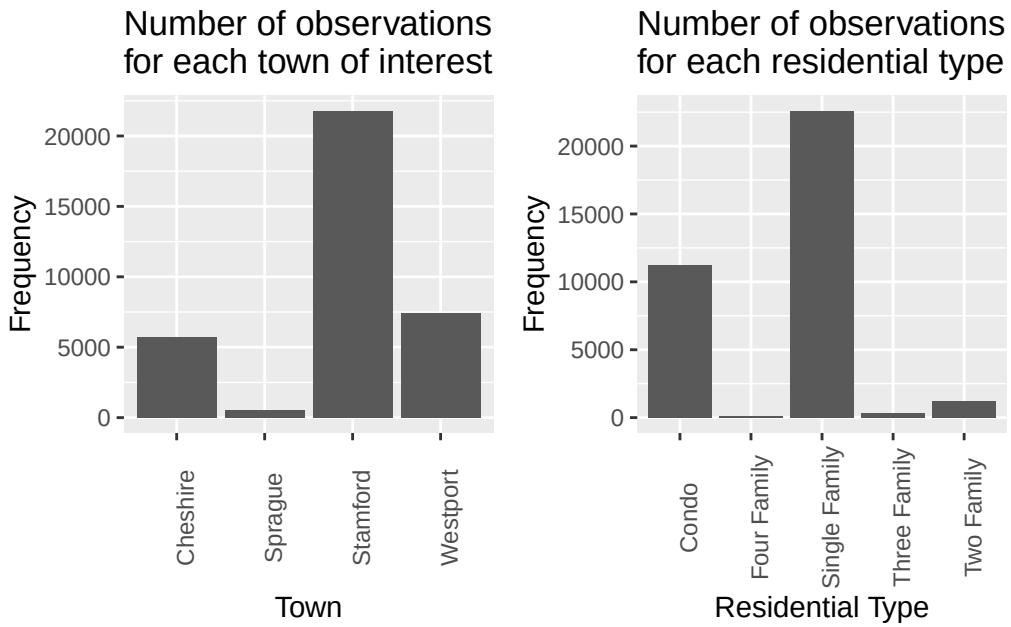


Figure 1: The plot of the left shows that the number of observations is greatest for Stamford and lowest for Sprague. The plot of the right shows that There are more entires for Single Family and Condo properties and very few for Two Family, Three Family, and Four Family.

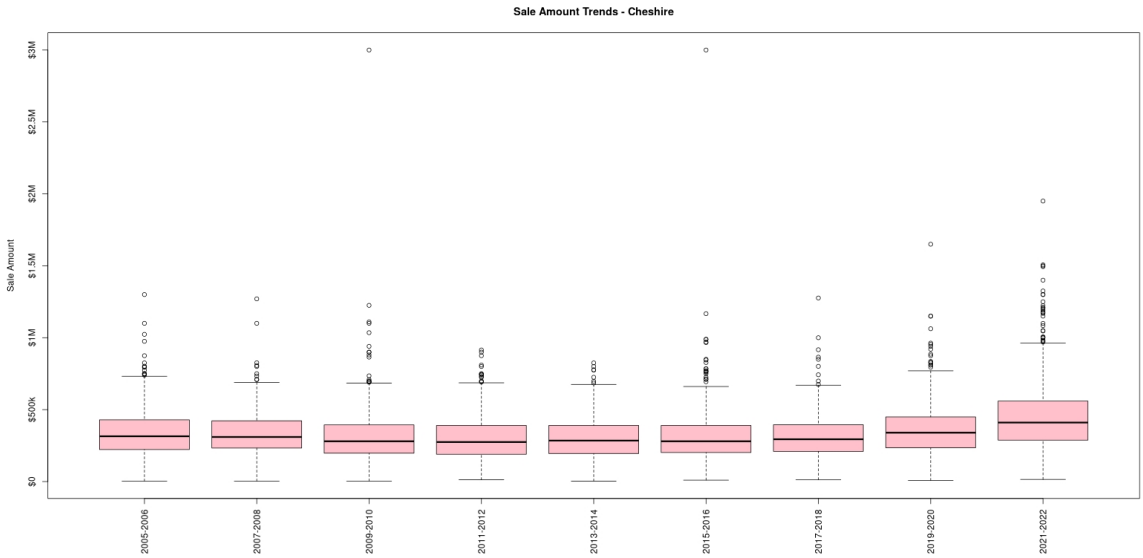
Since Stamford is a big city with a large population, it makes sense for it to have the most entries. On the other hand, Sprague is the opposite as a small town with a small population which accounts for the low amount of sales. Most housing are Single Family or Condos. It is rare to see Two Family and above sized homes being built.

3.2 Continuous

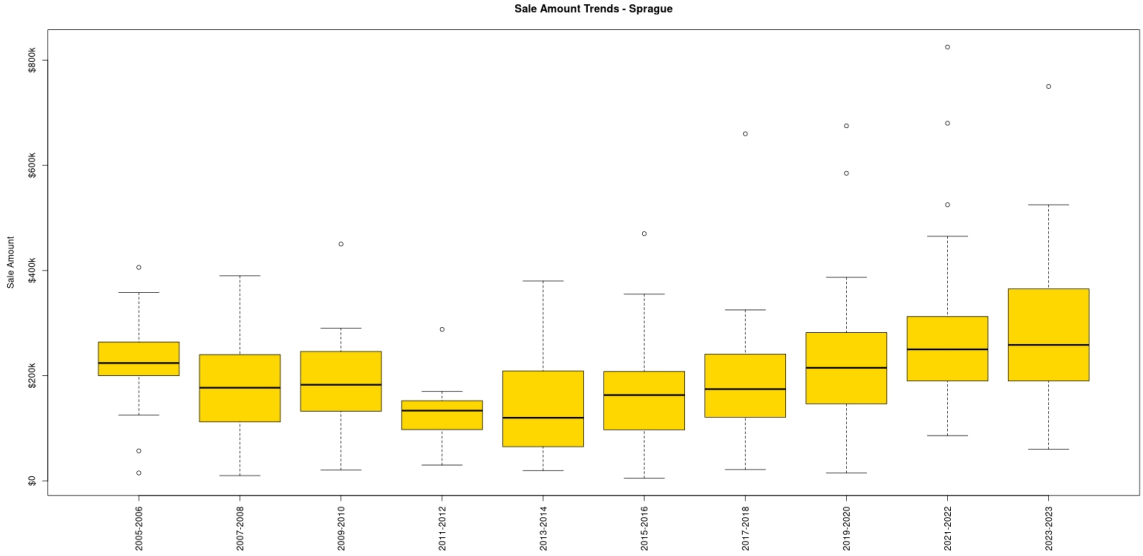
3.2.1 Box Plots By Town

We constructed a series of boxplots and scatterplots for each of our continuous variables (sale price, assessed value, and sales ratio), and then repeated this step for each of the four towns selected (Stamford, Westport, Cheshire, and Sprague). We decided that aligning biannual boxplots vertically and then sorting them in time order would demonstrate trends over time in a more effective manner.

Sales Amounts:



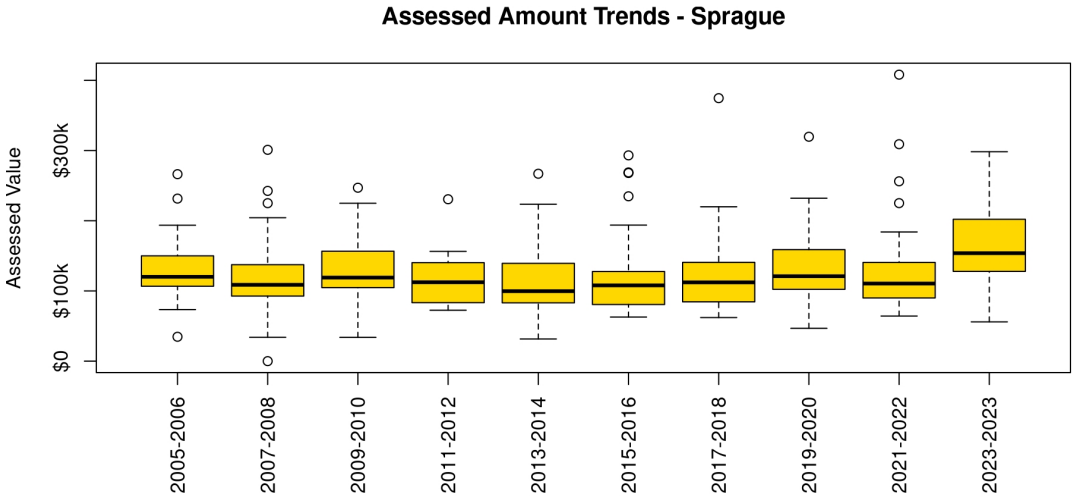
Prices in Cheshire seemed to slump for longer after 2008, both in the outlier spread and the main quartiles, not really rebounding until the COVID-19 pandemic. It is farther away from major metropolitan areas and the coast with a smaller population. It appeals to buyers with more modest means.



Sprague, the smallest of the towns we chose, displayed more year to year variability and a great deal more market driven variability than larger towns. Part of that is likely due to lower sales volume, and perhaps since it is less prosperous it was a relative bargain and accessible to more buyers.

Assessed Values:

The assessments tend to display less year on year variability than the actual sales prices. The IQR remains remarkably constant, with the most visible variability in the outlier field. This suggests that for all but the most expensive homes, the assessor is simply following a universal formula to calculate the value of a house. With the most expensive, it could be a different metric that takes in unique features that only wealthier people have, or it could be that fewer of these sales cause more variation by year.



Sprague is an interesting case, because it's a relatively poor area with fewer residents. The high assessment variability here can probably be chalked up to fewer units assessed every year. There could also be the opposite issue to the high outliers. The low outliers may be in a uniquely bad condition, each requiring its own consideration.

Sales Ratio:

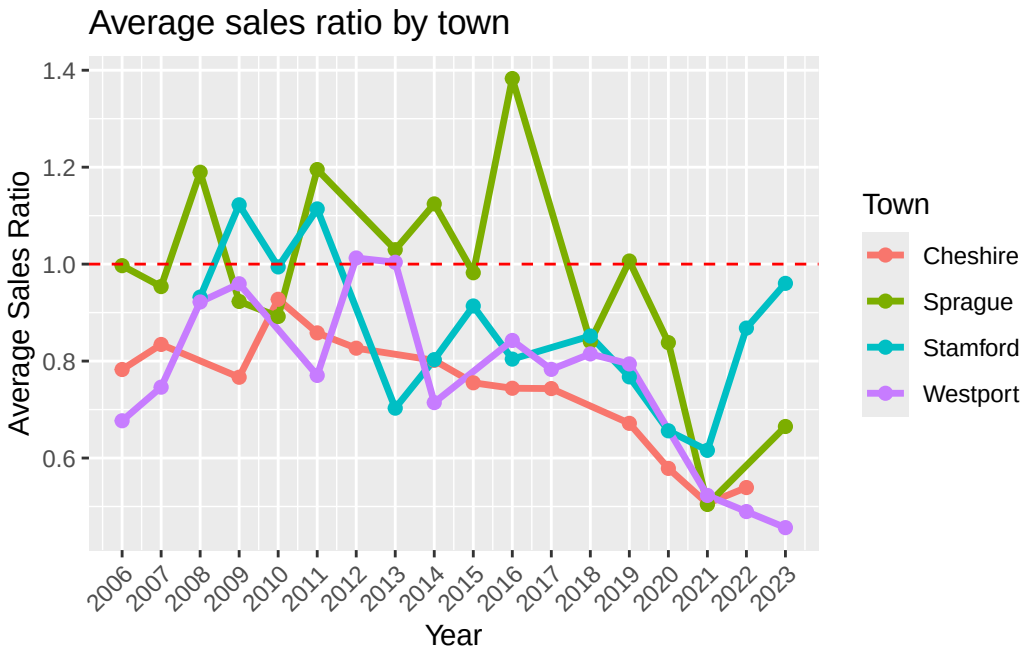


Figure 2: Average sales ratio by town:Average Sales Ratio (ASR) = Average Assessed Value (AAV) / Average Sales Amount (ASA). ASR > 1 means the Average Assessed Value is higher than the Average Sales Amount ; ASR $= 1$ means the AAV is equal to the ASA; ASR < 1 means the AAV is lower than the ASA.

Westport is a wealthy town with high property values, and Cheshire is an upper-middle-class town. Stamford is a urban city near NYC and its economy is closely related to NYC. Sprague is a small town with lower-middle-income. In general, Figure 2 shows that the well-off towns (Westport/Cheshire/Stamford) experience smaller sales ratios during the 2008 financial crisis. The less economically developed town (Sprague) and the NYC-dependent-economic-structure town (Stamford) experience greater market volatility. Considering the time-lag factor, during the 2007–2008 subprime mortgage crisis, housing prices plummeted, causing the ASR in many towns to rise rapidly, even faster than the speed of assessment adjustments. After 2009, the ASR in most towns declined because the real estate market began to recover, and housing prices increased faster than the assessment values. During the global pandemic in 2020, the ASR of all four towns showed a significant downward trend, reflecting a surge in housing prices caused by the extremely low mortgage interest rates and the increase in housing demand driven by

the work-from-home trend.

4 Advanced Analysis

We decided to make a heat map of sale price, assessed value, and sale ratio for the entirety of Connecticut and the cities we selected.

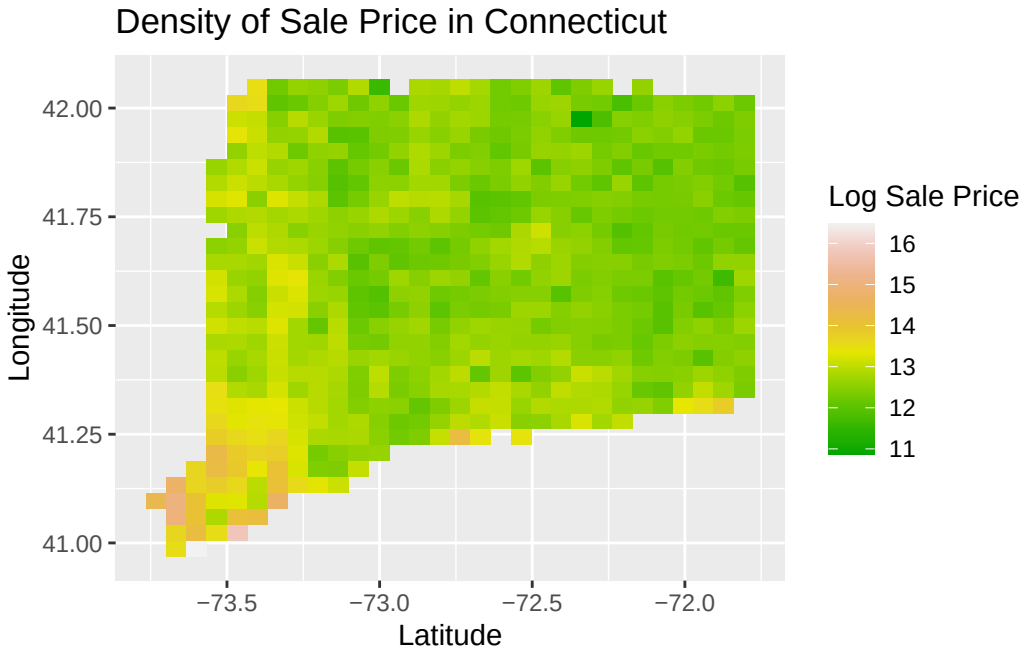


Figure 3: Heat map showing the densities of sale price throughout Connecticut, USA

Figure 3 shows the densities of sale price for Connecticut. For the majority of Connecticut, the sale price is uniform. However, the closer you get to New York, the higher the sale price is.

Figure 4 shows the densities of sale price in two of the four cities we picked. Stamford has the most data points and Cheshire has more variability in prices.

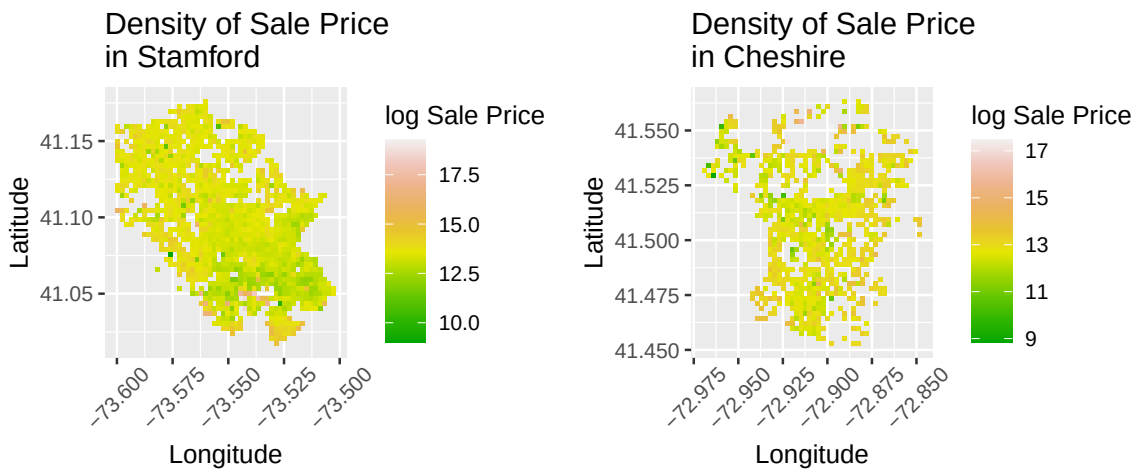


Figure 4: Heat maps of sale price for each city of choice: Stamford, CT on the left, Cheshire, CT on the right.

To understand whether the relationship between assessed value and sale amount differs across towns, we applied a transformed multiple regression model and t-test.

Because the population sizes and income levels differ across the four selected towns, we applied a log transformation to both Sale Amount and Assessed Value to make the data fit better and look clearer in the plots. This is the our model:

$$log(Sale.Amount) = \beta_0 + \beta_1 * log(Assessed.Value) + \beta_2 * Town + \beta_3 * [log(Assessed.Value) \times Town] + \varepsilon$$

Here, β_0, the intercept, is the expected log sale price in the reference town (Cheshire) when the log(assessed value) is 0, holding all other variables constant. β_1, the coefficient of log(Assessed.Value), represents the average percent increase in sale price in Cheshire for each additional 1% increase in assessed value, holding all other variables constant. β_2, the set of coefficients, represents the intercept difference between each individual town (Sprague, Stamford, Westport) and Cheshire. β_3, the set of coefficients, represents the difference in effect of assessed value on sale price across individual towns (Sprague, Stamford, Westport) compared to Cheshire. ε is the error term.

To ensure valid inference (Figure 5), we need to meet these four assumptions: **Linearity**: The original scatter plots are messy and hard to get useful information from, so we applied a log transformation in the regression model. After transformation, the plot shows that the variables are linearly correlated; the sale price increases as the assessed value increases in general. **Independence of error**: The Residual vs Time plot shows that most of the residuals are aligned with the reference line. It indicates that most of the errors are independent. **Equal Variance of error**: The Residual vs fitted plot shows that as the fitted value increases, the residuals form a cone shape, and the spread decreases with the fitted value. The equal variance

assumption is violated. Because the four towns we chose differ in the estate market, and the sale amount and the assessed value vary, the variances are not constant.Therefore, the validity of t-tests may be affected. **Normality of error**: The QQ plot indicates that half of the residuals closely align with the reference line, and the left tail falls below the expected quantiles, and the right tail falls above the expected quantiles. This indicates residual non-normality. The Residual histogram shows that the residual distribution is approximately normal, and there is a strong peak around 0, and both tails are spread. Together with the QQ plot, the pattern indicates that the normality assumption is not fully satisfied.

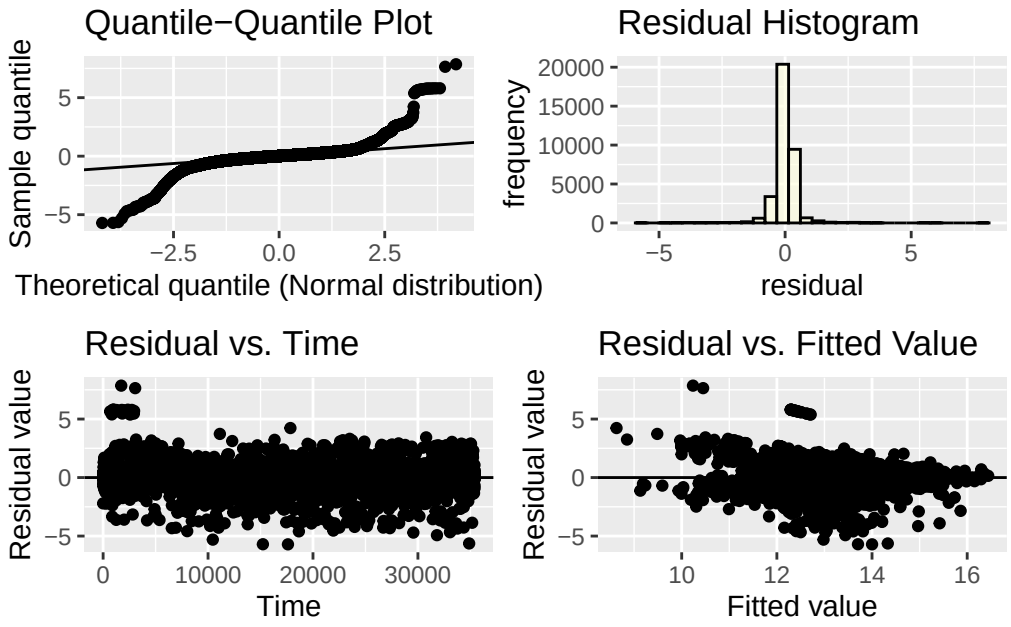


Figure 5: Assumptions

Figure 6 shows that the assessed value is positively correlated with the sale amount in all four towns. In Cheshire, for each additional percentage increase in assessed value, the expected sale amount will increase by 0.745%. The interaction term shows that, compared with Cheshire, the slope of the assessed value and the expected sale amount in Sprague (β =0.047 p = 0.384 $>$ 0.05) and Stamford (β = -0.004, p = 0.742 $>$ 0.05) is similar. The relationship is stronger in Westport (β = 0.153, p $<$ 0.001), showing that the expected sale amount in Westport tends to change more than in the other towns when assessed value changes. In general, the model fits the data well (R^2 = 0.68). Because the equal variance assumption is violated, its explanatory power is limited. The assessed value and the town can only explain part of the difference in sale amount, and other factors (policy, house area, etc.) may have an influence.

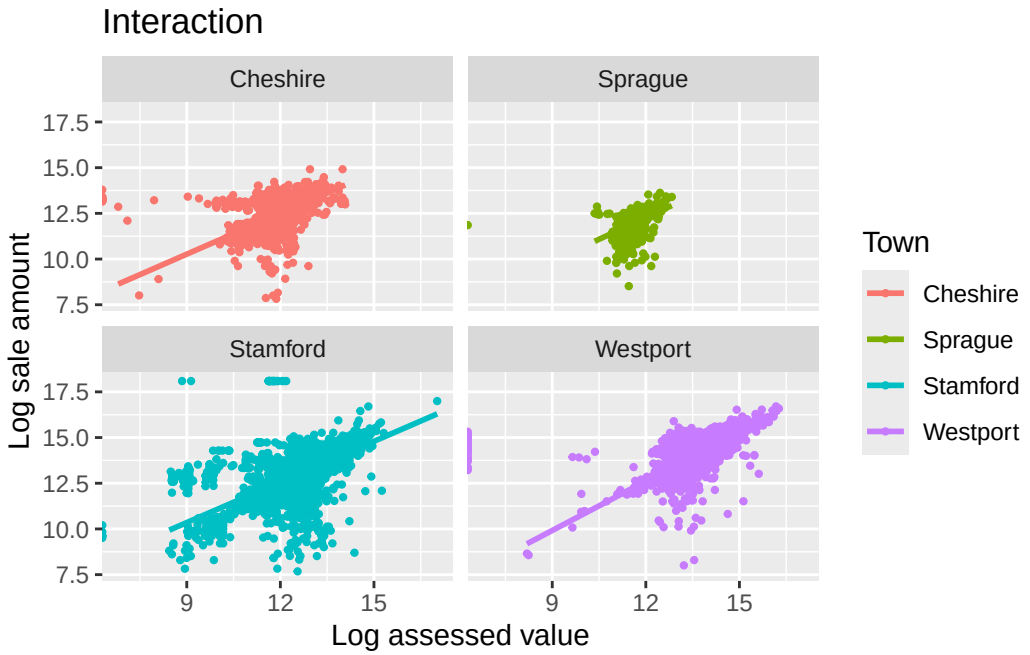


Figure 6: Scatter plot of regression model

5 Conclusion

Overall, our analysis of four towns in Connecticut shows that the real estate markets don't change randomly, but are strongly dependent on the overall economic circumstances. The differences in geographic location, city size, and level of economic development among cities also led to different market responses during the 2008 economic crisis and COVID-19. Our plots and models highlight these patterns and help clarify how the market evolves over the long run. In the future, we could try regrouping different areas by similar patterns using a different approach. This might reveal relationships across regions that don't show up in our basic summary statistics. We can also consider other socioeconomic variables, such as education levels and median household income, to better explain differences in the real estate market. A more thorough study of outlier data may also be interesting for several reasons. For instance, it could perhaps reveal unfair or inconsistent assessment values.

References

2025, Data Commons. n.d. "Place Rankings - Data Commons." *Data Commons*. https://datacommons.org/ranking/Median_Income_Household/CensusCountyDivision/geoId/09?h=geoId%2F0915021860&unit=%24.