Analysis of Residential Real Esatate Data From Connecticut*

Siling Guo, Megan Joseph, John Sawicz

November 15, 2025

1 Introduction

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. We also wanted to sample towns with different populations and densities. (2025, n.d.).

2.1 Data Cleaning

Many of the data points were missing or were empty characters, so we dropped those rows. 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.

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

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.

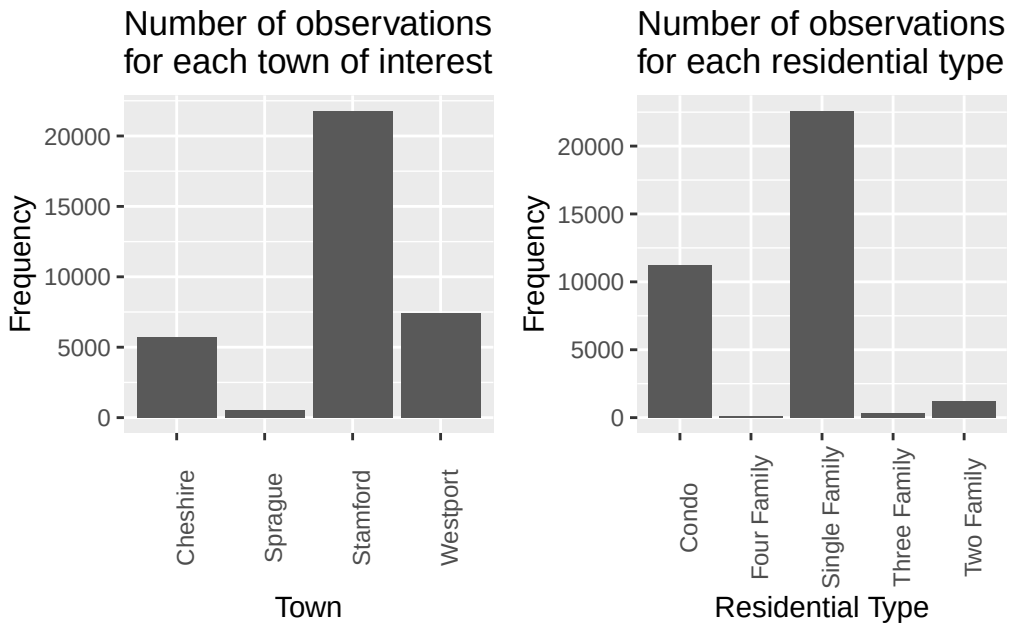


Figure 1: The plot of the left shows that the number of observations is greatest for Stamford and lowest for Sprague. The plot of the right shows that There are more entires for Single Family and Condo properties and very few for Two Family, Three Family, and Four Family.

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 Distribution of Sale Price

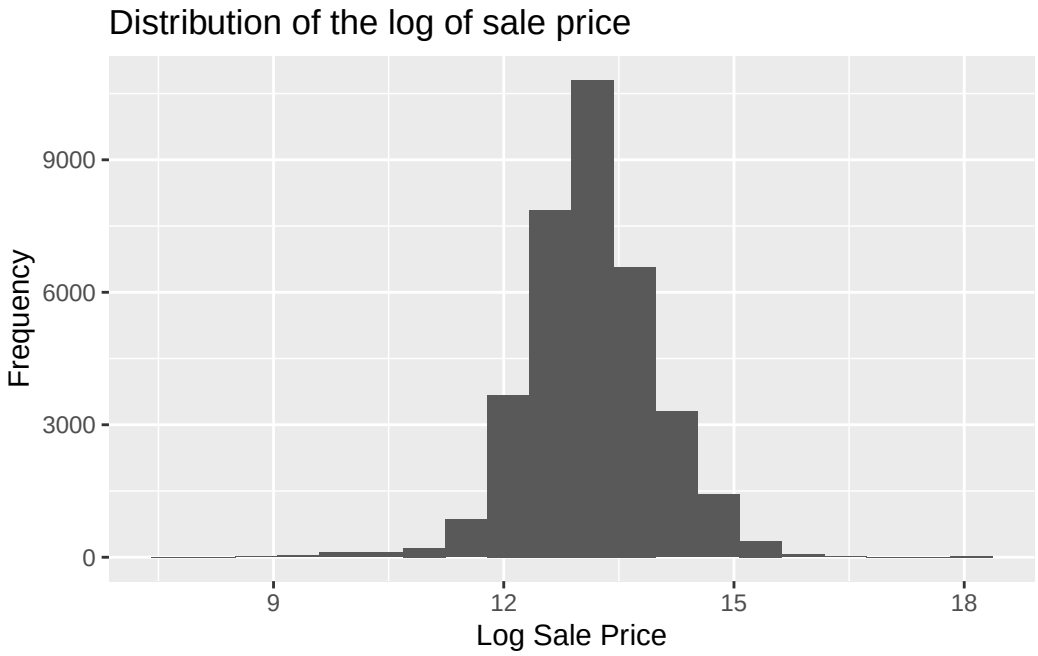


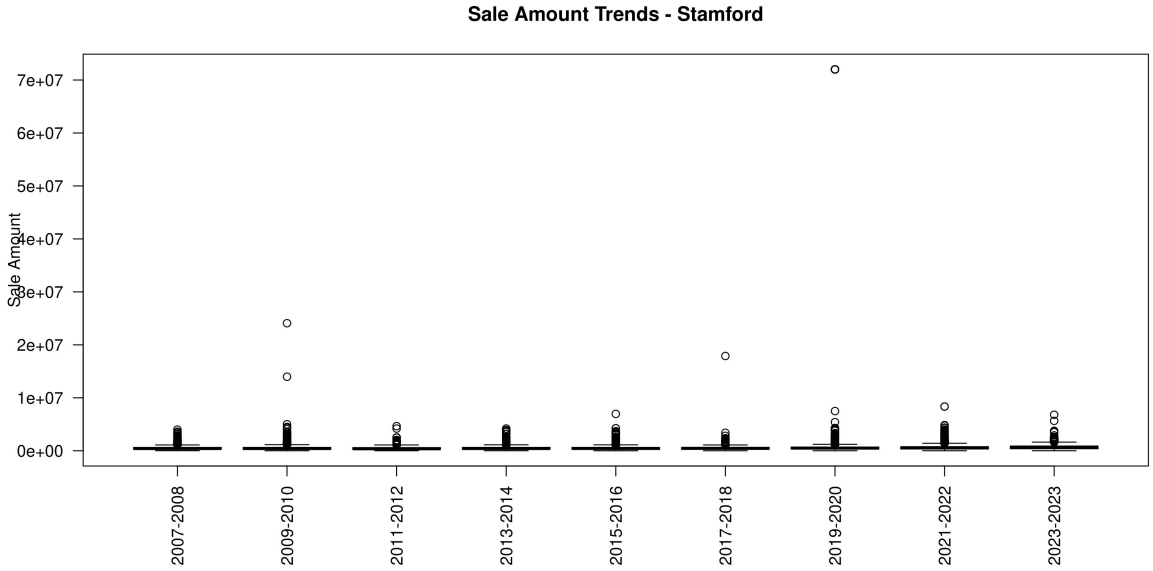
Figure 2: Sale price seems to be centered at around $\$e^{13}$.

We took the log of the sale price because it is extremely right skewed otherwise. We see that the data is centered at around $\$e^{13} \approx \$442,413$.

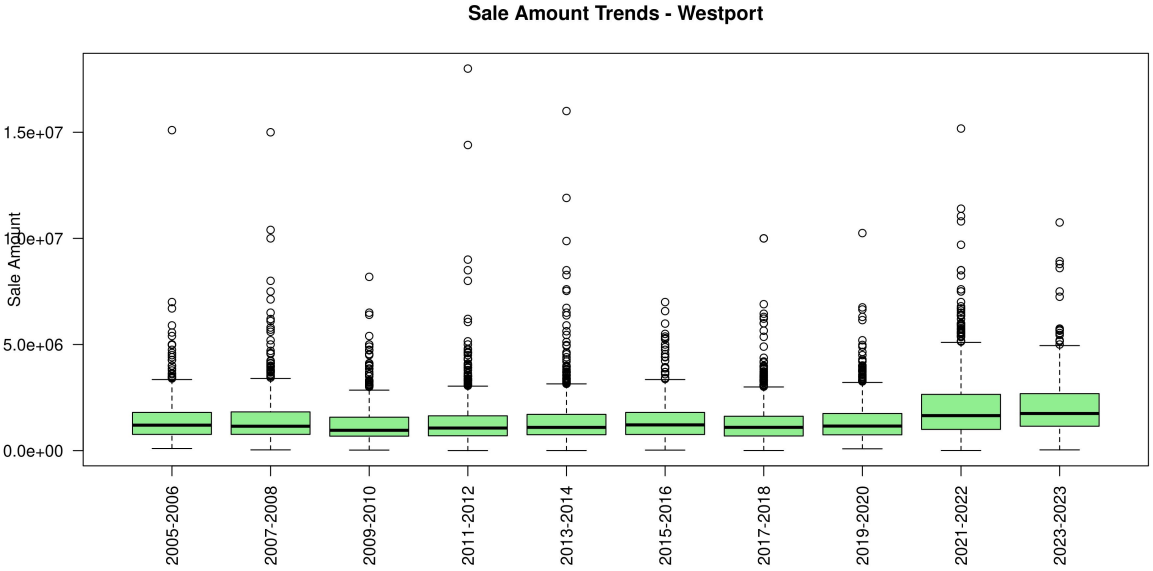
3.2.2 Box Plots By Town

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

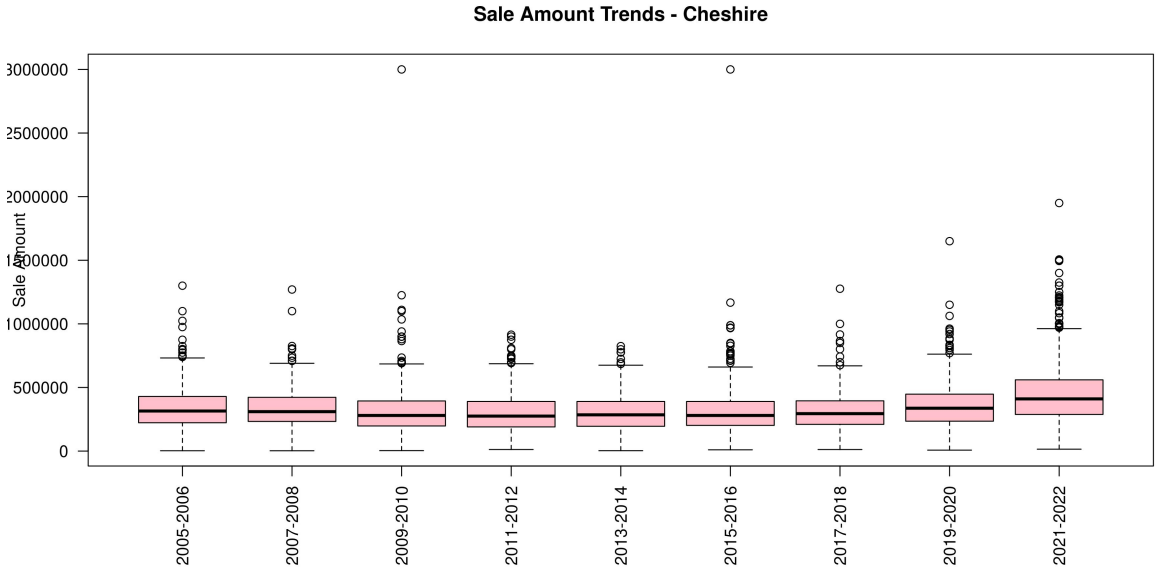
Sales Amounts:



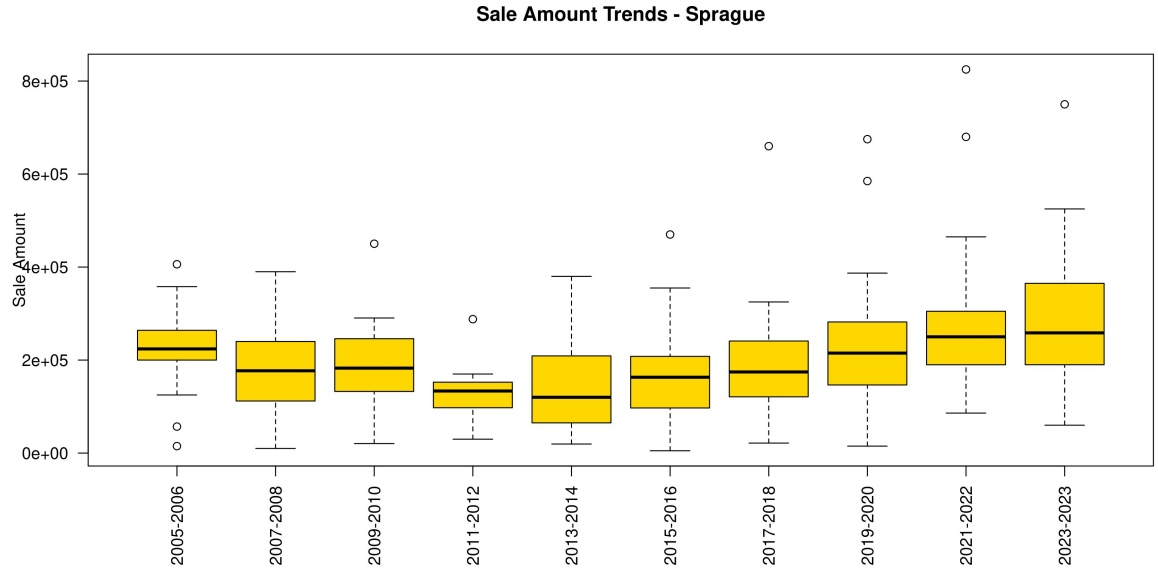
We are still combing through the upper end of the Sale Amount column for errors. In a dataset this large, there are bound to be mistakes which lead to unreasonable figures which deserve to be deleted.



Here we see some evidence of national and regional property market trends in a very right skewed town. The large number of outliers, some of which are priced rather high, speaks to a desirable coastal community within easy commuting distance of New York City. Notice the rapid outlier rebound after the 2008 financial crisis.

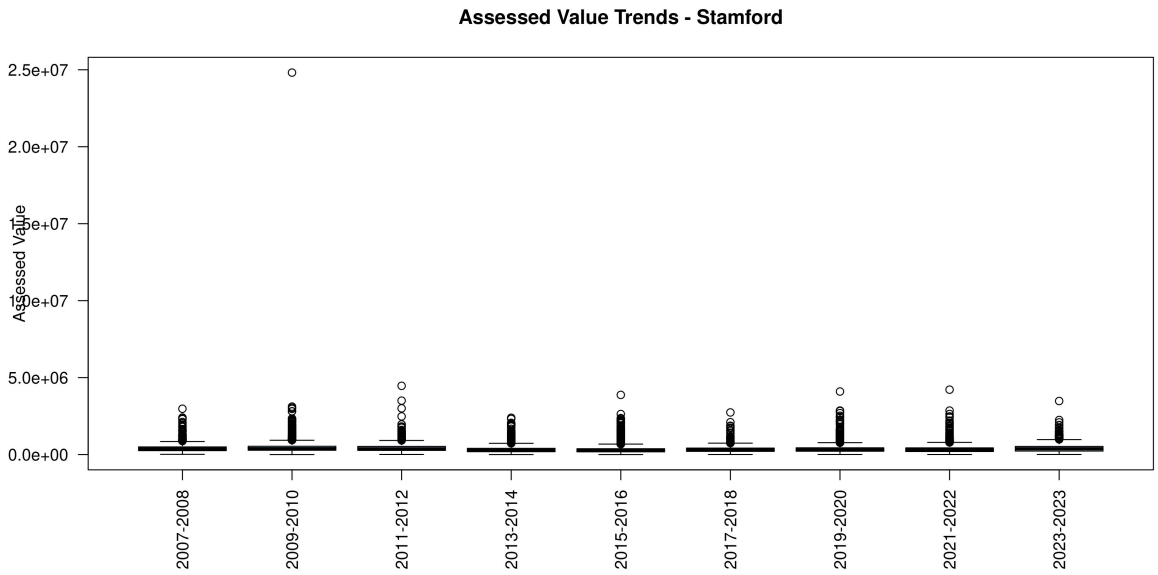


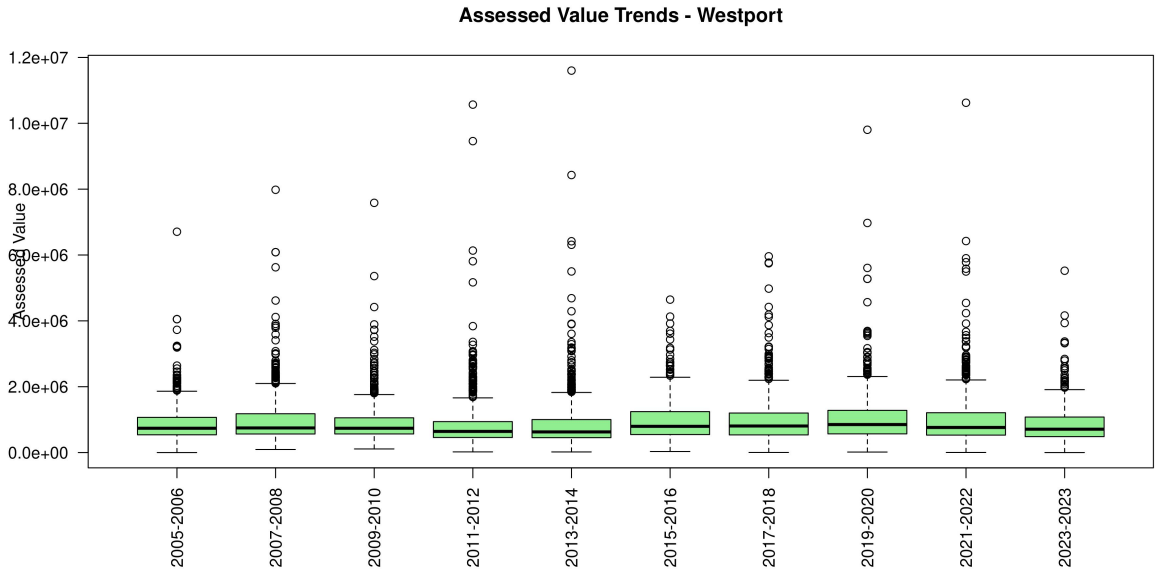
Prices in Cheshire seemed to slump for longer after 2008, both in the outlier spread and the man quartiles, not really rebounding until the Covid-19 pandemic. It is farther away from major metropolitan areas and the coast, with a smaller population. It appeals to buyers with more modest means.

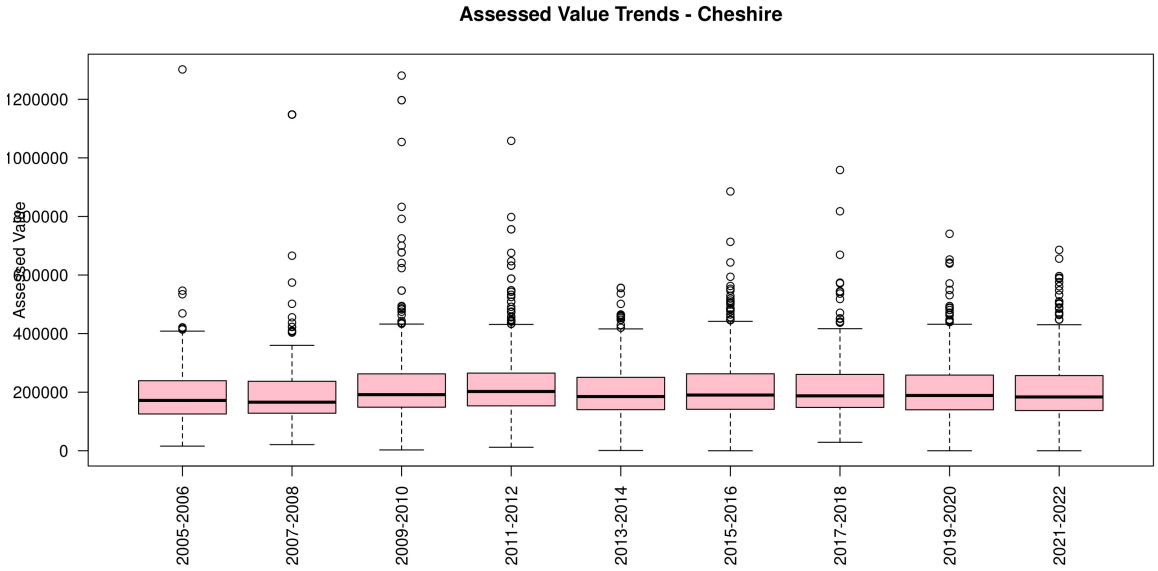


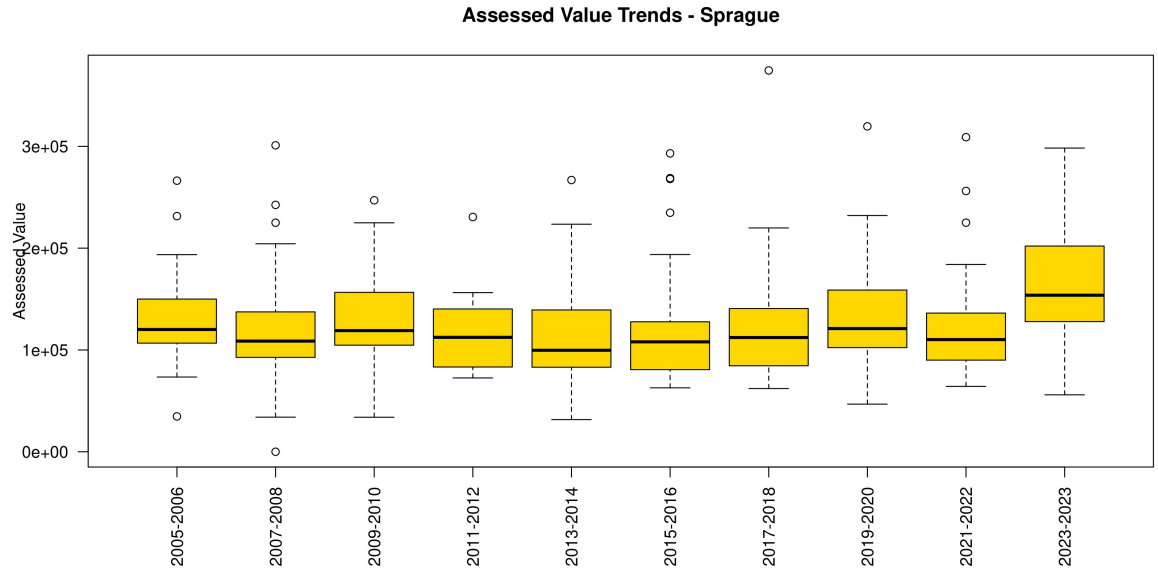
Sprague, the smallest of the towns we chose, displayed more year to year variability and a great deal more market driven variability than larger towns. Part of that is likely due to lower sales volume, and perhaps since it is less prosperous it was a relative bargain and accessible to more buyers.

Assessed Values:






Assessed Value Trends - Cheshire
Assessed Value
1200000
1000000
800000
600000
400000
200000
0
2005-2006
2007-2008
2009-2010
2011-2012
2013-2014
2015-2016
2017-2018
2019-2020
2021-2022


Assessed Value Trends - Sprague
Assessed Value
3e+05
2e+05
1e+05
0e+00
2005-2006
2007-2008
2009-2010
2011-2012
2013-2014
2015-2016
2017-2018
2019-2020
2021-2022
2023-2023

Sales Ratio:

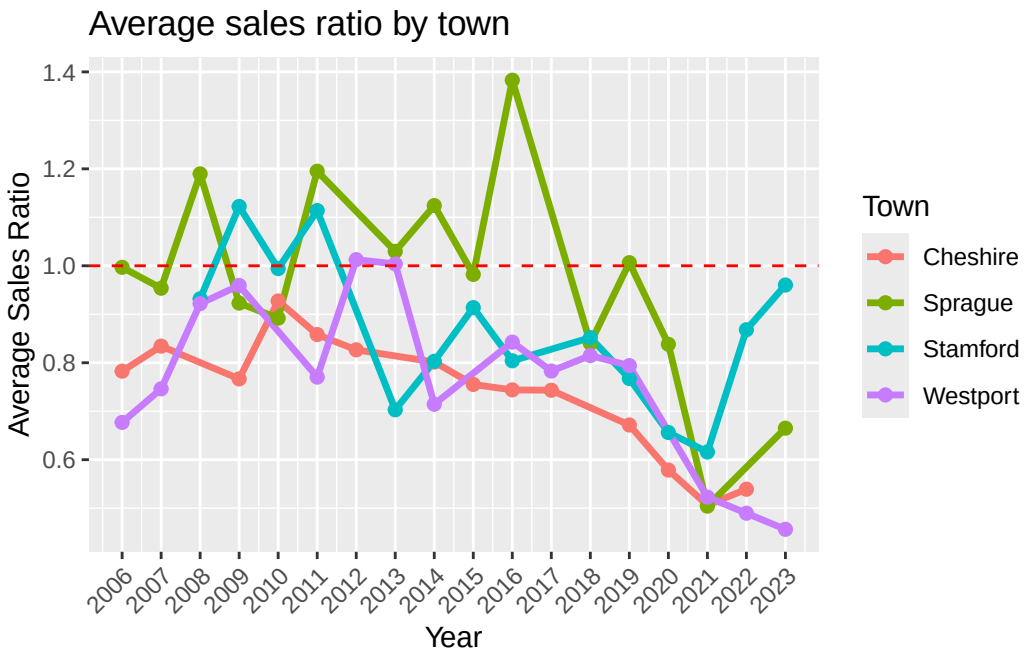


Figure 3: Average sales ratio by town

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 4 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 5 shows the densities of sale price in the four cities we picked. Stamford and Westport seem to have the most data points, with Sprague being the most sparse. Sprague, Westport, and Stamford are more uniform whereas Cheshire has more variability in prices.

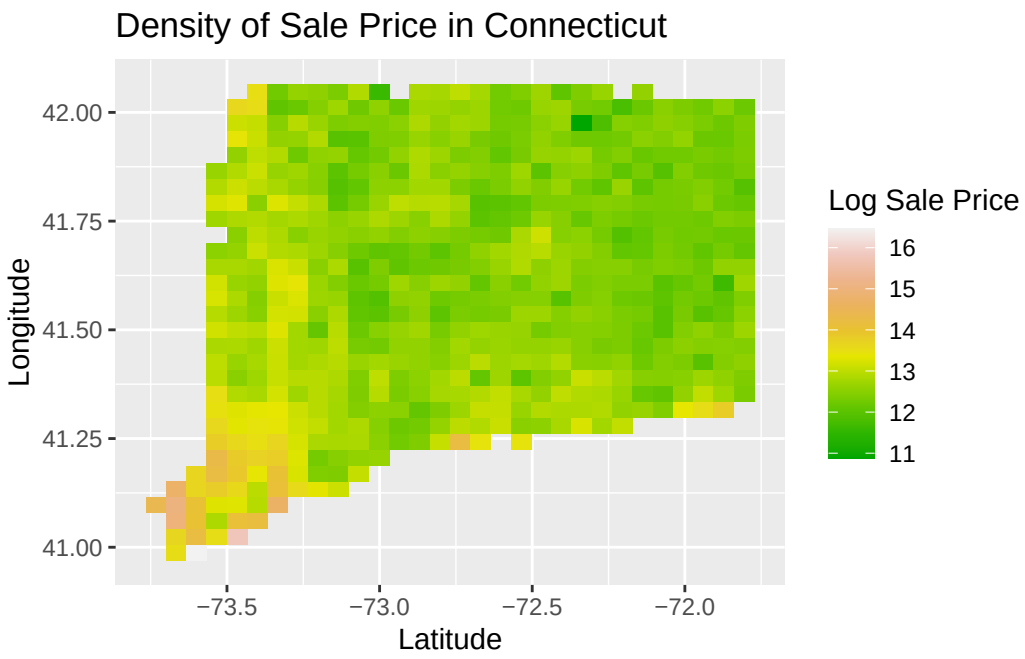


Figure 4: Heat map showing the densities of sale price throughout Connecticut, USA

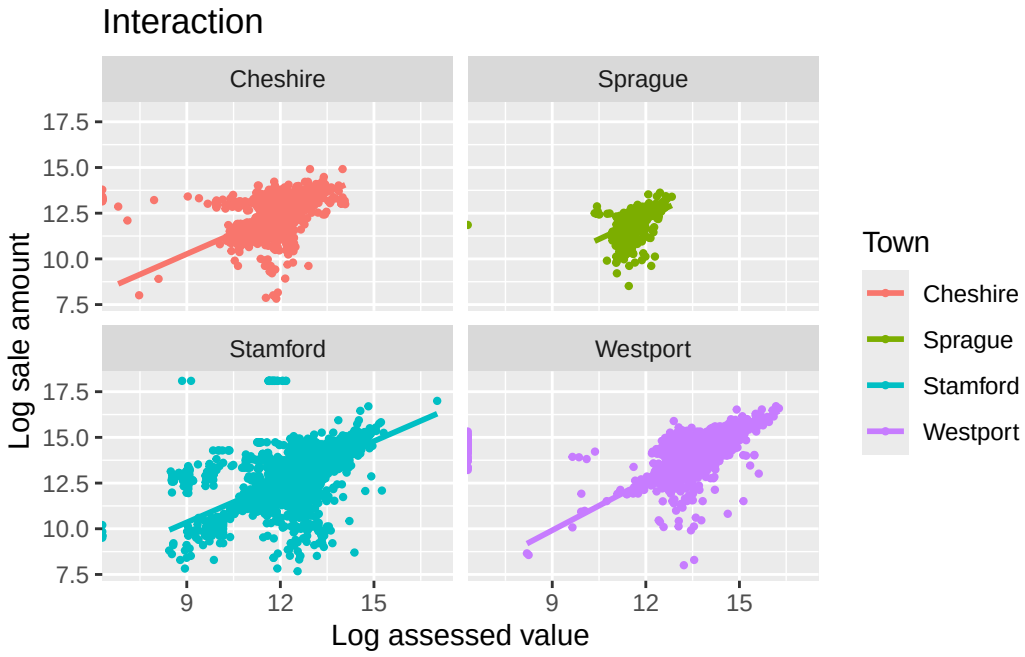


Figure 6: Scatter plot of regression model

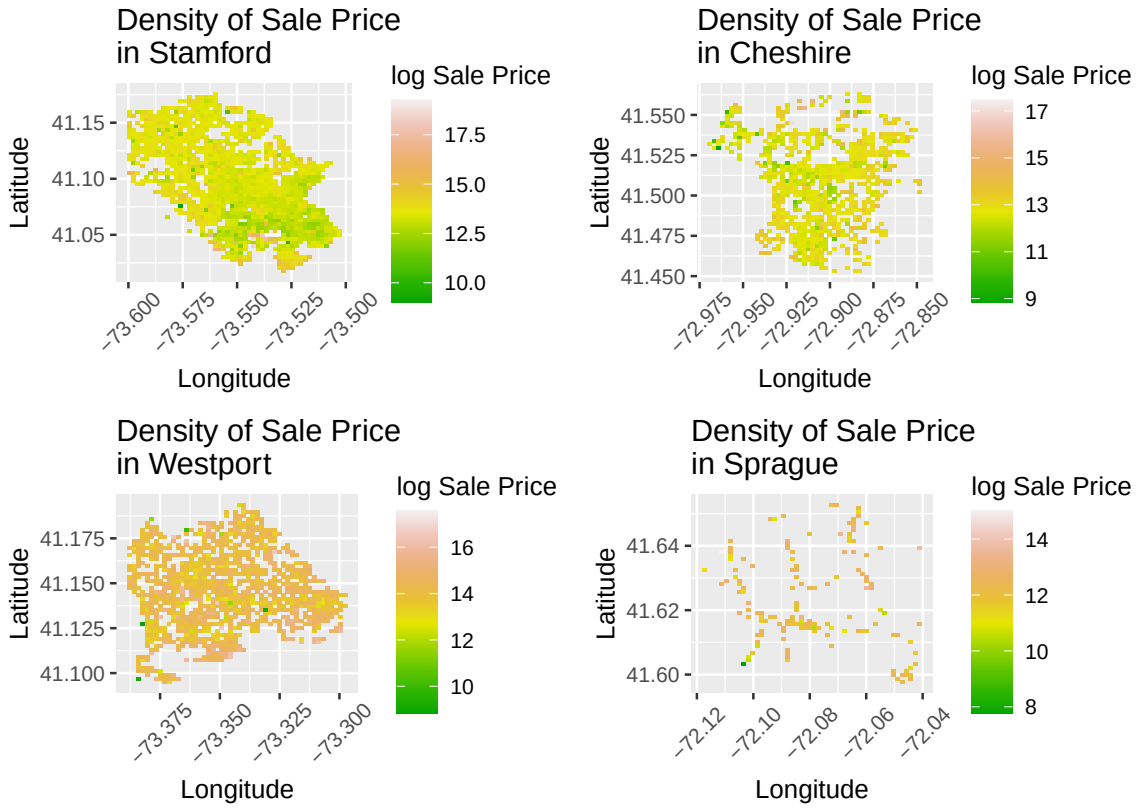


Figure 5: Heat maps of sale price for each city of choice: Stamford, CT on the top left, Cheshire, CT on the top right, Westport, CT on the bottom left, and Sprague, CT on the bottom right.

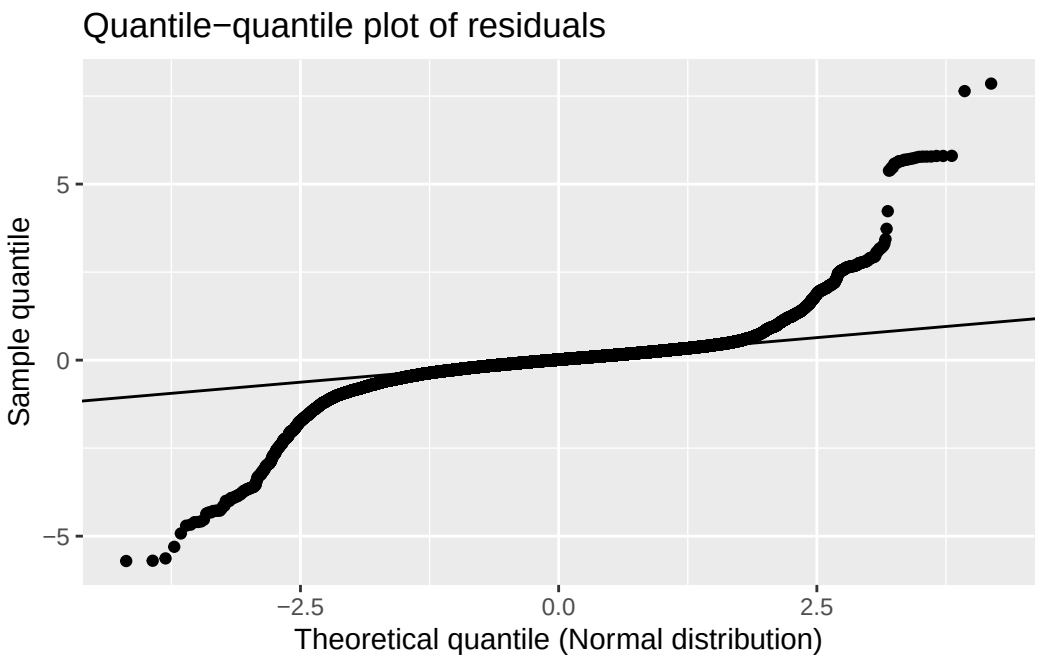


Figure 7: Quantile-quantile plot of residuals

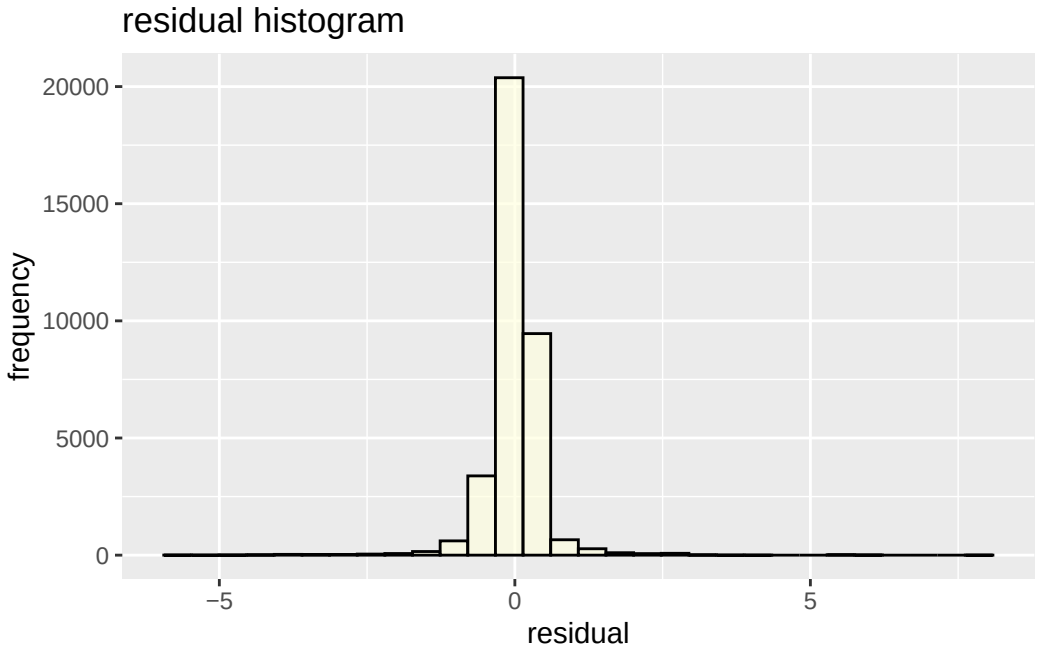


Figure 8: residual histogram

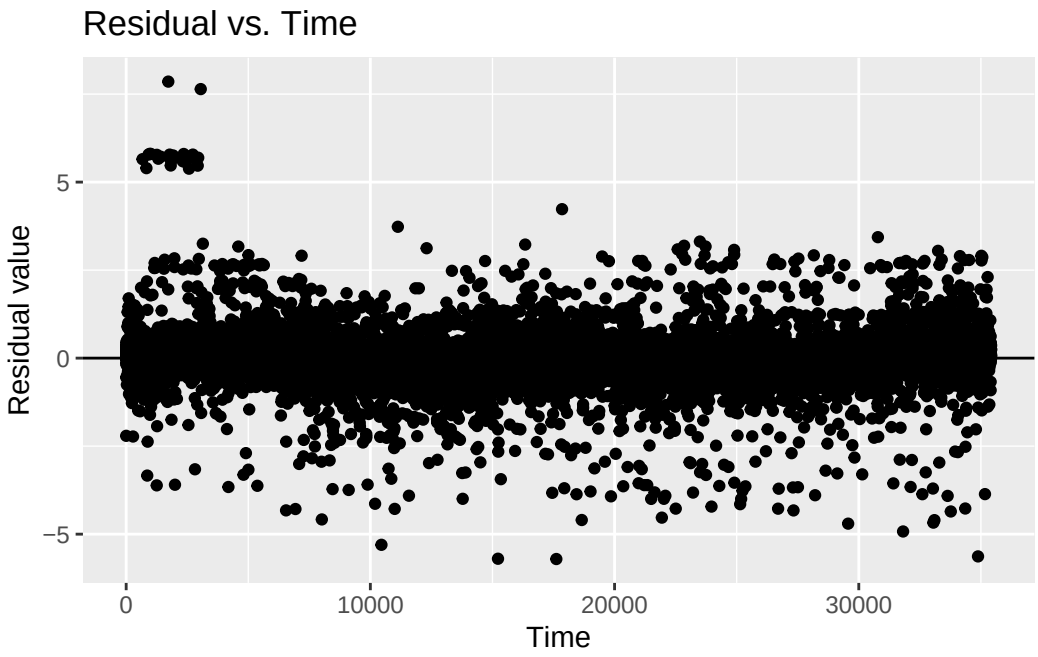


Figure 9: Residual vs. Time

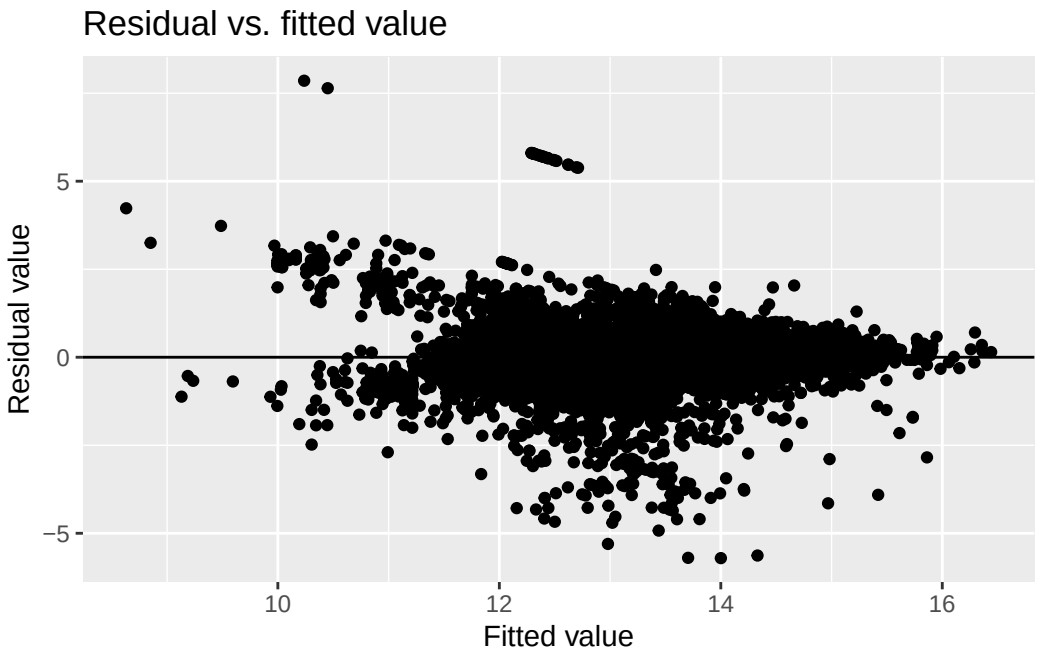


Figure 10: Residual vs. fitted value

5 Conclusion

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.